# ΠΜΣ Μεγάλα Δεδομένα και Αναλυτική

Τμήμα Ψηφιακών Συστημάτων

Πανεπιστήμιο Πειραιώς

# Ανάλυση Οδηγικής Συμπεριφοράς με τη Χρήση Αλγορίθμων Μηχανικής Μάθησης

Χρήστος Θεοδωρόπουλος

ΑΜ: ΜΕ1808

**Υπεύθυνος Καθηγητής: Χρήστος Δουλκερίδης**


**Αθήνα, 2021**

# Ευχαριστίες

Με την ολοκλήρωση της μεταπτυχιακής διπλωματικής μου εργασίας, θα ήθελα να ευχαριστήσω θερμά όλους όσους συνέβαλλαν άμεσα ή έμμεσα στην εκπόνηση της, και ειδικότερα τον επιβλέπων καθηγητή μου, κ. Χρήστο Δουλκερίδη για την επιστημονική του καθοδήγηση και τις πολύτιμες υποδείξεις του.

Τέλος θα ήθελα να δώσω θερμές ευχαριστίες στην οικογένεια μου για την κατανόηση και τη συμπαράσταση που μου προσέφεραν.

# Περίληψη

Παγκοσμίως ο αυξημένος αριθμός τροχαίων ατυχημάτων αποτελεί πρόβλημα υψίστης σημασίας έχοντας επιπτώσεις στην οικονομία των κρατών αλλά και στην ίδια την κοινωνία. Η πρόβλεψη και η βελτίωση της οδηγικής συμπεριφοράς με τη χρήση αλγορίθμων μηχανικής μάθησης μπορεί να θεωρηθεί ίσως ένα από τα σημαντικότερα εργαλεία που θα συνεισφέρει στον τομέα της οδικής ασφάλειας. Η αναγνώριση του τρόπου οδήγησης ενός οχήματος μπορεί να αποτελέσει ιδιαίτερα πολύτιμη γνώση για τις αυτοκινητοβιομηχανίες, την αυτόνομη οδήγηση, τις ασφαλιστικές εταιρίες αλλά και πληθώρα νέων εφαρμογών.

Τα τελευταία χρόνια σημειώνεται σημαντική ανάπτυξη στο επιστημονικό πεδίο της μηχανικής μάθησης κυρίως λόγω της αύξησης της υπολογιστικής ισχύος και του αποθηκευτικού χώρου των υπολογιστικών συστημάτων. Η χρήση της μηχανικής μάθησης δίνει την δυνατότητα στους υπολογιστές να κάνουν προβλέψεις χρησιμοποιώντας πειραματικά δεδομένα χωρίς να έχουν προγραμματιστεί προηγουμένως ρητά. Η χρήση αυτών των σύγχρονων και καινοτόμων μεθόδων τείνει να ξεφεύγει από τα στενά όρια των βιομηχανικών και επαγγελματικών διαδικασιών. Πλέον με τη ανάπτυξη των Internet of Things (IoT) και Internet of Vehicles (IoV) η μηχανική μάθηση χρησιμοποιείται εξίσου και σε καθημερινές δραστηριότητες όπως είναι η οδήγηση ενός οχήματος.

Στόχος της παρούσα διπλωματικής εργασία είναι να εξετάσουμε και να διακρίνουμε τους βασικούς διαφορετικούς τρόπους οδήγησης κάνοντας χρήση αλγορίθμων μηχανικής μάθησης. Η παρούσα μελέτη περιλαμβάνει μια προσέγγιση για την αναγνώριση της οδηγικής συμπεριφοράς διάφορων οδηγών σε διαφορετικές διαδρομές χρησιμοποιώντας δεδομένα χρονοσειρών που παράχθηκαν από τους αισθητήρες του CANBUS συστήματος των οχημάτων. Συγκριμένα στην παρούσα έρευνα εφαρμόστηκαν διάφορες τεχνικές συσταδοποίησης πάνω στο σύνολο των δεδομένων που αποτελείτο από χρονολογικές σειρές.

Στα αποτελέσματα της έρευνας θα επιχειρήσω να εκτιμήσω την οδηγική συμπεριφορά ενός συνόλου οδηγών χρησιμοποιώντας δεδομένα που συλλέχθηκαν κατά την διάρκεια της οδήγησης σε διάφορες διαδρομές. Κάνοντας χρήση των μεθόδων συσταδοποίησης θα ομαδοποιήσω τα δεδομένα των χρονολογικών σειρών σε συστάδες. Κατόπιν θα διερευνηθούν τα δεδομένα εντός των συστάδων ώστε να κατηγοριοποιηθούν οι ομάδες αναλόγως του τρόπου οδήγησης.



Λέξεις κλειδιά: Οδηγική συμπεριφορά, Μηχανική μάθησης, Συσταδοποίηση, Χρονοσειρά, Ατυχήματα

## Abstract

Worldwide, the increasing number of road accidents is one of the most serious issues and has an impact, not only on the state economy but also on the society itself. Nowadays, prediction and improvement of driving behaviour are keys to contribute to road safety by using machine learning algorithms. The taken information about driving behaviour can be important data for automotive industries, insurance companies and other new applications.

In recent years, due to the increase in computing power and computer systems storage spaces, significant growth has happened in the scientific field of machine learning. The use of machine learning enables computers to make predictions using experimental data without the need for human intervention beforehand. The use of these modern and innovative methods tends to go beyond the narrow confines of industrial and professional processes. Now with the use of Internet of Things (IoT) and Internet of Vehicles (IoV) it is used equally in daily activities such as driving a vehicle.

The dissertation aims to examine and distinguish the different ways of driving using machine learning algorithms. This research includes an approach by recognizing the driving behaviour of different drivers and a variety of routes.

The time-series data has generated by vehicle sensors. Specifically, in the present study, various clustering techniques applied to the data set.

In the end, in the research results, I will attempt to assess the behavioural driving of a set of drivers using collected data, while driving on various routes. I will group their data using the clustering method and then categorize them according to the driving style.



Keywords: Driving behaviour, Machine learning engineering, Clustering, Time series, Accidents

# *Περιεχόμενα*

## Κατάλογος Πινάκων



## Κατάλογος Εικόνων

# Κεφάλαιο 1

## 1.1 Εισαγωγή

Από τα πιο σοβαρά προβλήματα σε παγκόσμια κλίμακα αποτελούν τα τροχαία ατυχήματα με ανυπολόγιστες επιπτώσεις τόσο στον οικονομικό τομέα όσο και στο κοινωνικό σύνολο. Οι αιτίες πρόκλησής τους ποικίλουν και η αντιμετώπιση τους είναι μείζονος σημασίας ζήτημα. Κοινός παρονομαστής στην πλειοψηφία των περιπτώσεων είναι ο ανθρώπινος παράγοντας. Συνεπώς, η εκτίμηση του προφίλ οδηγικής συμπεριφοράς αποτελεί ίσως το σημαντικότερο πρόβλημα στο πεδίο της οδικής ασφάλειας. Κάθε χρόνο χιλιάδες άτομα χάνουν τη ζωή τους ή τραυματίζονται σοβαρά σε ατυχήματα που συμβαίνουν στους ευρωπαϊκούς δρόμους (European Transport Safety Council, 2019)

Ένα μεγάλο ποσοστό των ατυχημάτων οφείλεται στην αποκλίνουσα συμπεριφορά των οδηγών σε σχέση με τους κανόνες οδικής συμπεριφοράς (NHTSA, 2016). Εκτιμάται ότι η συστηματική χρήση της ζώνης ασφαλείας, καθώς και η τήρηση των ορίων ταχύτητας και των κανόνων οδικής κυκλοφορίας θα επέφερε μείωση των ανθρώπινων απωλειών κατά 12.000 ετησίως. Αποκλειστικά και μόνο το 2018 στην Ευρώπη, έχασαν την ζωή τους 25.100 άνθρωποι και 135.000 τραυματίστηκαν (European Commission, 2019a). Περίπου το 14% των ατόμων που έχασαν την ζωή τους βρίσκονταν στην ηλικιακή ομάδα 18-24 χρονών (European Parliament, 2019).

Το 2010 η Ευρωπαϊκή Ένωση αναθεώρησε τη δέσμευση της για βελτίωση της οδικής ασφάλειας, θέτοντας ένα νέο στόχο. Μέχρι το 2020 να έχει μειωθεί κατά 50% ο αριθμός των δυστυχημάτων συγκριτικά με τα επίπεδα του 2010. Μία ετήσια μείωση της τάξης του 6,7% μεταξύ της περιόδου 2010-2020 θα ήταν επαρκής για την επίτευξη του στόχου. Δυστυχώς όμως, σύμφωνα με τα επίσημα καταγεγραμμένα στοιχεία από το 2010 μέχρι και το 2017 η ετήσια μείωση κυμάνθηκε στο 3,4%. Συνεπώς, ο ευρωπαϊκός στόχος για μείωση των δυστυχημάτων μέχρι το 2020 δεν επιτευχθεί (European Commission, 2017). Εντούτοις, σε σχέση με παλαιότερα έτη 2001-2017 έχει υπάρξει σημαντική μείωση της τάξεως του 57% στο ποσοστό θνησιμότητας λόγω τροχαίων (European Parliament, 2019). Σημαντικό ρόλο διαδραμάτισε η χρήση νέων ανθεκτικότερων υλικών σε συνδυασμό με τη νέα δομική σχεδίαση των οχημάτων αυξάνοντας την παθητική ασφάλεια. Παράλληλα η χρήση σύγχρονων συστημάτων ενεργητικής ασφάλειας όπως τα ABS, ESP, Traction control, lane keep assist κ.α. συμβάλουν στην ασφαλέστερη μετακίνηση (Jarašūniene & Jakubauskas, 2007). Ακολουθώντας την κατευθυντήρια γραμμή για ασφαλέστερες μετακινήσεις, η μοντέρνα προσέγγιση που υιοθετούν οι αυτοκινητοβιομηχανίες εστιάζει στην σχεδίαση και ανάπτυξη νέων ‘έξυπνων’ οχημάτων χρησιμοποιώντας δεδομένα από τους αισθητήρες των οχημάτων.

## 1.2 Ορισμός προβλήματος

Ο τρόπος οδήγησης κάθε ατόμου είναι προσωποποιημένος και μοναδικός όπως το δακτυλικό αποτύπωμα. Συνεπώς, η ταυτοποίηση του θα αποτελέσει ένα νέο εργαλείο με ευρεία χρήση στο πεδίο της οδικής ασφάλειας, της οικολογικής οδήγησης, της διαχείρισης ενέργειας των νέων ηλεκτρικών οχημάτων αλλά και σε πολλά ακόμα πεδία. Αξιοποιώντας τα διαγνωστικά συστήματα και τα συστήματα καταγραφής δεδομένων του αυτοκινήτου (CAN Bus system, IoV) θα παρέχεται η δυνατότητα να αναγνωρίσουμε την οδηγική συμπεριφορά μέσω της ομαδοποίησης σε διαφορετικές ομάδες των οδηγών και των διαδρομών τους και ταυτόχρονα διερευνώντας τα χαρακτηριστικά των ομάδα αυτών . Είναι γεγονός ότι οι οδηγοί που οδηγούν τηρώντας τον Κώδικα Οδικής Κυκλοφορίας (ΚΟΚ) τείνουν να εμπλέκονται σε λιγότερα ατυχήματα. Συνεπώς κρίνεται επιτακτική η ανάγκη ανάπτυξης μίας μεθόδου η οποία θα αποκωδικοποιεί, θα αναγνωρίζει και θα κατηγοριοποιεί τα διαφορετικά πρότυπα οδήγησης. Την ανάγκη αυτή καλείται να καλύψει η επιστήμη των δεδομένων και η μηχανική μάθηση.

## 1.3 Αντικείμενο διπλωματικής

Η παρούσα διπλωματική εργασία έχει ως στόχο την αναγνώριση του τρόπου οδήγησης ενός οχήματος κάνοντας χρήση τεχνικών εξόρυξης δεδομένων και αλγορίθμων μηχανικής μάθησης. Μία από τις κυριότερες εφαρμογές μηχανικής μάθησης είναι η ομαδοποίηση (συσταδοποίηση) των δεδομένων. Χρησιμοποιώντας ένα σύνολο δεδομένων χρονοσειρών προερχόμενο από το CAN Bus σύστημα ενός οχήματος, εξετάστηκε η υλοποίηση συγκεκριμένων αλγορίθμων και η σύγκρισή τους σχετικά με τη δημιουργία ενός καλού σχήματος συσταδοποίησης. Απώτερος σκοπός της συσταδοποίησης των δεδομένων είναι η διάκριση των οδηγών και των διαδρομών τους σε συστάδες, η διερεύνηση των στοιχείων εντός των συστάδων και τέλος η κατηγοριοποίηση τους με κριτήριο τον τρόπο οδήγησης.

## 1.4 Δομή διπλωματικής εργασίας

Παρακάτω συνοψίζεται η δομή της διπλωματικής εργασίας, όπως θα παρουσιαστεί στα επόμενα κεφάλαια:

- Στο κεφάλαιο 2 παρουσιάζονται προηγούμενες έρευνες σχετικές με το αντικείμενο μελέτης της διπλωματικής
- Στο κεφάλαιο 3 αναγράφεται το θεωρητικό υπόβαθρο που είναι απαραίτητο για την ανάγκες της εργασίας
- Στο κεφάλαιο 4 παρουσιάζονται οι αλγόριθμοι που υλοποιήθηκαν καθώς και τα δεδομένα που χρησιμοποιήθηκαν

- Στο κεφάλαιο 5 παραθέτουμε την πειραματική μελέτη και τα τελικά αποτελέσματα.
- Το κεφάλαιο 6 περιέχει τα συμπεράσματα της έρευνας.
- Στο κεφάλαιο 7 παραθέτουμε την βιβλιογραφία που χρησιμοποιήθηκε για την εκπόνηση της εργασίας.

# Κεφάλαιο 2

## 2.1 Σχετικές έρευνες

Πολλές έρευνες έχουν ασχοληθεί με την εφαρμογή μεθόδων μηχανικής μάθησης στο πεδίο της αυτοκινητοβιομηχανίας και οδηγικής συμπεριφοράς. Σχετική έρευνα από τους Minh Van Ly et al. (2013) αναλύει και κατηγοριοποιεί τον τρόπο οδήγησης για διαφορετικούς οδηγούς χρησιμοποιώντας δεδομένα προερχόμενα από αισθητήρες αδράνειας ενός οχήματος σε πραγματικές συνθήκες οδήγησης. Το σύνολο των δεδομένων τους αποτελούνταν από διαμήκεις και πλευρικές μετρήσεις επιταχύνσεων και επιβραδύνσεων αλλά και μετρήσεις γυροσκοπίων για καταγραφή της αλλαγής κατεύθυνσης του οχήματος συναρτήσει του χρόνου. Η μέθοδος που υλοποιήθηκε βασίζεται σε semi-supervised learning συνδυάζοντας τις μεθόδους unsupervised learning (clustering K-means) και supervised learning (SVM). Τέλος, οι συγγραφείς προτείνουν μελλοντικά την εξέταση του παραπάνω θέματος με περισσότερους οδηγούς, εκτελώντας το μοντέλο ενδεχομένως σε κάποιο Smartphone και εξετάζοντας τη διαφοροποίηση του τρόπου οδήγησης ημερησίως και ενημερώνοντας τον οδηγό ανάλογα.

Παράλληλα, η μελέτη των Naiwala et al. (2016) πραγματεύεται την κατηγοριοποίηση του τρόπους οδήγησης σε μια κλειστή διαδρομή με στροφές με τη χρήση αλγορίθμων μηχανικής μάθησης. Τα δεδομένα προέρχονται από ένα προσομοιωτή οδήγησης όπου συμμετείχαν 16 άτομα. Επίσης, περιλαμβάνουν γωνία κλήσης τιμονιού, διαμήκεις και εγκάρσιες επιταχύνσεις και μετατοπίσεις, δεδομένα από το πεντάλ επιτάχυνσης και επιβράδυνσης με ρυθμό στα 60HZ. Για τις ανάγκες της έρευνας στα δεδομένα καταργήθηκε η διάσταση του χρόνου και χρησιμοποιήθηκε ως index η χιλιομετρική απόσταση στην κλειστή διαδρομή της προσομοίωσης. Για να μειωθούν οι διαστάσεις των δεδομένων χρησιμοποιήθηκε η τεχνική του Principal Component Analysis (PCA), και εφαρμόστηκαν τεχνικές κανονικοποίησης (normalization). Τέλος, επιλέχθηκε δειγματοληπτικά ένας αριθμός των δοκιμών ανά οδηγό και κατηγοριοποιήθηκαν από έναν ειδικό σε θέματα οδηγικής ικανότητας διαχωρίζοντας τα σε 2 κατηγορίες (low/average-skilled, high-skilled). Ως μοντέλα εφαρμόστηκαν το k-Nearest Neighbor (k-NN) χρησιμοποιώντας την ευκλείδεια απόσταση και το Support Vector Machine (SVM) και συγκρίθηκαν με βάση την ακρίβεια (accuracy) ως μέτρο απόδοσης. Καταλήγοντας οι ερευνητές, προτείνουν μελλοντικά περαιτέρω έρευνα την προσθήκη επιπλέον οχημάτων στην προσομοίωση και την προσθήκη διαδρομών με διαφορετική κλίση.

Το 2012 οι Ahmad Aljaafreh et al. παρουσίασαν ορισμένες μεθόδους κατηγοριοποίησης της οδηγικής συμπεριφοράς σε 'Υπό της κανονικής', 'Κανονική', 'Επιθετική', 'Πολύ επιθετική'. Χρησιμοποιώντας επιταχυνσιόμετρα και καταγράφοντας τις εγκάρσιες και διαμήκεις επιταχύνσεις καθώς και την ταχύτητα του οχήματος, οι ερευνητές εξέτασαν τη χρήση ενός συστήματος fuzzy logic για τη κατηγοριοποίηση των δεδομένων. Ως μελλοντικό πεδίο έρευνας προτείνουν την ενσωμάτωση και άλλων οδηγικών γεγονότων.

Επιπροσθέτως σχετική έρευνα είναι και των Yadav et al. (2017) που προτείνει ένα νέο σύστημα για κατηγοριοποίηση σε ασφαλής και οικονομική οδήγηση χρησιμοποιώντας δεδομένα από το σύστημα αυτοδιάγνωσης (OBD II) των οχημάτων και βίντεο από mini-dash κάμερες. Το σύνολο των δεδομένων περιλάμβαναν 51 διαφορετικές μεταβλητές. Στο μοντέλο που ανέπτυξαν εφαρμόστηκαν τεχνικές κανονικοποίησης και γραμμικής συσχέτισης ώστε να καταλήξουν σε 5 βασικές μεταβλητές που χρησιμοποίησαν στο γραμμικό τους μοντέλο. Ως μελλοντική έρευνα επάνω στο θέμα προτείνουν την εξέταση της βελτιστοποίησης των υπερ-παραμέτρων του μοντέλου καθώς και την προσθήκη περισσοτέρων μεταβλητών.

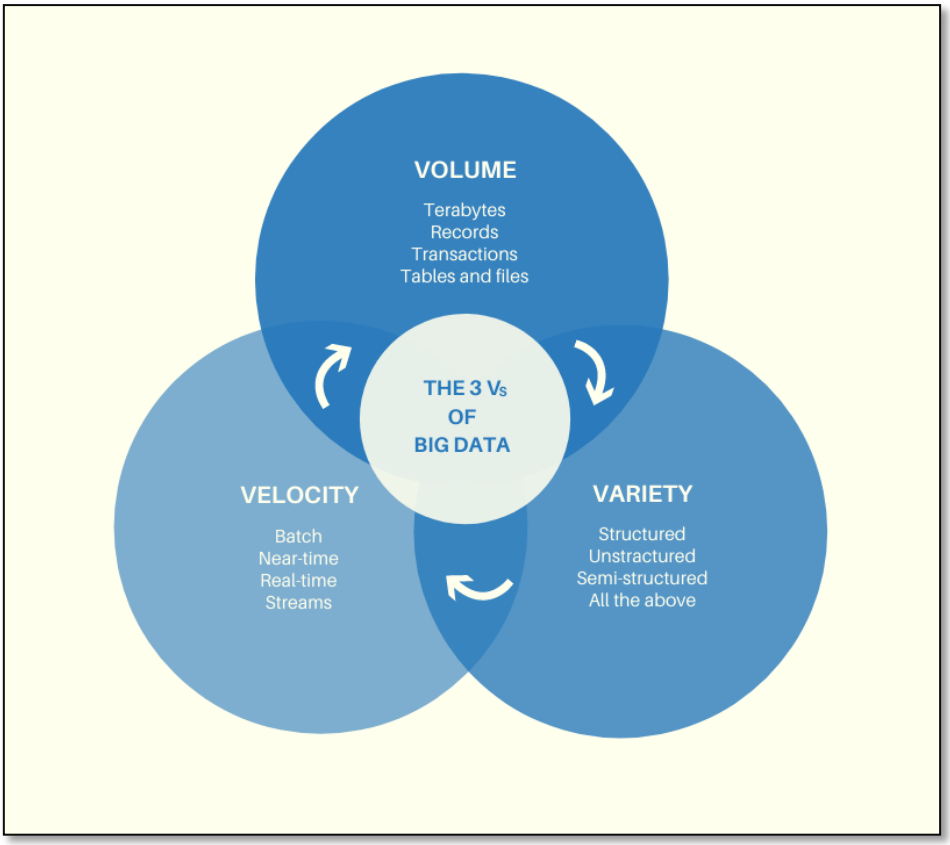
Το 2018 σε έρευνα των Zardosht et al. παρουσιάστηκε μία ακόμα μέθοδο αναγνώρισης του τρόπου οδήγησης βασισμένη στην ιεραρχική συσταδοποίηση δεδομένων από το σύστημα CAN Bus του οχήματος λαμβάνοντας υπόψη την ταχύτητα του οχήματος, την πίεση που ασκείται στα πεντάλ επιτάχυνσης-επιβράδυνσης, και στη γωνία του τιμονιού. Τα δεδομένα προέρχονταν από συνθήκες πραγματικής οδήγησης σε αστικό περιβάλλον από 12 οδηγούς. Έπειτα από επεξεργασία των δεδομένων εφαρμόστηκε ιεραρχική συσταδοποίηση, καταλήγοντας σε διαχωρισμό των οδηγών σε 2 συστάδες. Τέλος, παρατηρώντας τις μέσες τιμές των τιμών των συστάδων και συγκρίνοντας τες διαπίστωσαν ότι η πρώτη συστάδα περιλάμβανέ μεγαλύτερες ταχύτητες – επιταχύνσεις συνεπώς περιείχε οδηγούς που οδηγούσαν επιθετικά σε συγκρίσει με τη δεύτερη συστάδα. Συνεπώς κατέληξαν στην κατηγοριοποίηση των οδηγών.

# Κεφάλαιο 3

## 3.1 Επιστήμη των Δεδομένων

Ο 21ος αιώνας έχει χαρακτηριστεί από πολλούς ως ο αιώνας τους πληροφορίας. Τα τελευταία χρόνια έχει αναπτυχθεί ένα νέο πεδίο γνώσης κάνοντας εκτεταμένη χρήση τεχνικών και θεωριών από διάφορους τομείς όπως της Στατιστικής, των Μαθηματικών και της Πληροφορικής. Το πεδίο αυτό φέρει το όνομα Επιστήμη των Δεδομένων και είναι το αποτέλεσμα τους ραγδαίας εξέλιξης στην επεξεργασία νέων πληροφοριών. Αυτή η γρήγορη ανάπτυξη έχει σημειωθεί στη συλλογή, την μεταφορά και την αποθήκευση όλο και μεγαλύτερου όγκων δεδομένων από διάφορες πηγές (Big Data) καθώς και στον τομέα την ανάλυσης δεδομένων. Τα Big Data δημιουργούνται καθημερινά γύρω τους με καταιγιστικό ρυθμό από διάφορες πηγές τους τα Smartphone, συσκευές IoT, αισθητήρες, ψηφιακές συναλλαγές έχοντας απλά πρόσβαση στο ιντερνέτ. Ως εκ τούτου, οι συμβατικές σχεσιακές βάσεις δεδομένων καθίστανται ανεπαρκείς στη διαχείριση τους και δίνει χώρο στην ανάπτυξη νέων μεθόδων και τεχνικών.

Με βάση το άρθρο του Marous (2012) τα 3 κύρια χαρακτηριστικά των Big Data συνοψίζονται στο παρακάτω σχήμα που δημιουργήθηκε (Εικόνα 1):




*Εικόνα 1. Οι τρείς πυλώνες των μεγάλων δεδομένων*

Το πρώτο χαρακτηριστικό αναφέρεται στην ταχύτητα δημιουργίας των δεδομένων (Velocity), το δεύτερο χαρακτηριστικό στην ποικιλομορφία των δεδομένων (Variety) και τέλος στον όγκο τους (Volume). Συνεπώς, αποκαλύπτεται η τεράστια μελλοντική πρόκληση την οποία καλούμαστε να αντιμετωπίσουμε και αφορά στην εύρεση μίας ρεαλιστικής και αποδοτικής προσέγγισης για την επεξεργασία των big data.

Παράλληλα με την εκθετική αύξηση που έχει παρουσιάσει ο όγκος των δεδομένων, έχει αυξηθεί η πολυπλοκότητα τους με αποτέλεσμα τη χρήση όλο και περισσότερων υπολογιστικών συστημάτων για την ανάλυση τους. Η ανάγκη αυτή για ανάλυση και εξόρυξη πληροφορίας οδηγεί στην ανάπτυξη μεθόδων και αλγορίθμων που θα παράγουν δημιουργώντας έναν νέο κλάδο γνωστός ως Μηχανική Μάθηση.

## 3.2 Μηχανική μάθηση

Με τον όρο μηχανική μάθηση εννοούμε την ικανότητα ενός υπολογιστικού συστήματος να δημιουργεί μοντέλα και πρότυπα από ένα σύνολο δεδομένων χωρίς να έχει προηγουμένως προγραμματιστεί ρητά (Samuel, 1959). Με άλλα λόγια είναι μία μέθοδος ανάλυσης των δεδομένων που αυτοματοποιεί την δημιουργία αναλυτικών μοντέλων. Αποτελεί υποκατηγορία του πεδίου της Τεχνητής Νοημοσύνης αλλά χρησιμοποιεί και εμπλέκεται και με άλλα πεδία όπως της Στατιστικής, των αλγορίθμων και των βάσεων δεδομένων. Τα συστήματα μηχανικής μάθησης μπορούν αφού προηγουμένως εκπαιδευτούν, να μάθουν από τα δεδομένα και στη συνέχεια με την εμπειρία και το χρόνο να βελτιωθούν. Συνεπώς είναι σε θέση:

- να μεταβάλλονται και να βελτιώνονται διαρκώς, αναφορικά με τις λειτουργίες που εκτελούν,
- να μεταβάλλουν τη γνώση τους είτε μετασχηματίζοντας την εσωτερική δομή τους είτε αποκτώντας επιπλέον γνώση,
- και να εκτελούν γενικεύσεις.

Η μηχανική μάθηση δεν είναι καινούργια επιστήμη. Τα πρώτα ερευνητικά βήματα πραγματοποιήθηκαν ήδη από τα μέσα του 19 αιώνα και πιο συγκεκριμένα από τον Alan Turing στην εργασία του ‘Computing Machinery and Intelligence’ όπου περιγράφει την λειτουργία μίας μηχανής ικανή να μαθαίνει (Turing, 1950). Στα τέλη όμως τις δεκαετίας του ’90 απέκτησε νέα δυναμική η οποία προήλθε από την εμφάνιση του διαδικτύου, την ταυτόχρονη συσσώρευση όλο και περισσότερων δεδομένων καθώς και την ανάπτυξη ισχυρότερων υπολογιστικών συστημάτων.

## 3.3 Εφαρμογές μηχανικής μάθησης

Όπως αναφέρθηκε και νωρίτερα σκοπός της μηχανικής μάθησης είναι η διερεύνηση των υπολογιστικών διαδικασιών ώστε να καταστήσει δυνατή την οργάνωση και την εξαγωγή γνώσης μέσα από την υπάρχουσα εμπειρία. Η ταχεία ανάπτυξη των υπολογιστικών συστημάτων και ο συνεχώς αυξανόμενος όγκος των δεδομένων την καθιστά ευρέως διαδεδομένη στη χρήση της με πληθώρα εφαρμογών σε διάφορους τομείς μεταξύ των οποίων συγκαταλέγονται και οι:

1. Μηχανές αναζήτησης
2. Φίλτρα spam
3. Self Driving
4. Ιατρική
5. Οικονομική επιστήμη
6. Ρομποτική
7. Συναλλαγές
8. Γεωργία
9. Αεροπορία
10. Τηλεπικοινωνίες

## 3.4 Κατηγορίες Μηχανικής Μάθησης


*Εικόνα 2. Κατηγορίες Μηχανικής Μάθησης*

Η Μηχανική Μάθηση αποτελεί το πιο αναπτυσσόμενο τομέα της Τεχνητής Νοημοσύνης. Τα τελευταία χρόνια έχουν αναπτυχθεί διάφορες τεχνικές μηχανικής μάθησης οι οποίες

επιλέγονται και ανταποκρίνονται ανάλογα με τη φύση του προβλήματος. Οι τεχνικές αυτές μπορούν να κατηγοριοποιηθούν σε 3 κατηγορίες (Εικόνα 2)(Ayodele, 2010; Nzfaruqui, 2019):

- **Επιβλεπόμενη Μάθηση, supervised learning:** Η επιβλεπόμενη μάθηση αποτελεί από τους βασικότερους τρόπους μάθησης. Σε αυτή τη κατηγορία, τα δεδομένα φέρουν ετικέτες "labels" και χωρίζονται σε δύο υποομάδες το σύνολο δεδομένων εκπαίδευσης και το σύνολο δεδομένων ελέγχου. Για την εκπαίδευση του αλγορίθμου χρησιμοποιείται η μία υποομάδα όπου είναι ένα μικρό μέρος του συνόλου δεδομένων γνωστό ως training set, και στη συγκεκριμένη φάση βρίσκει τη σχέση μεταξύ των δεδομένων και των ετικετών τις "labels" κατασκευάζοντας μια συνάρτηση τις γνωστές εξόδους. Στόχος είναι η γενίκευση τις συνάρτησης ώστε εν τέλη ο αλγόριθμος έχοντας εκπαιδευτεί να τροφοδοτηθεί με το υπόλοιπο κομμάτι του συνόλου των δεδομένων γνωστό ως test setκαι να υπολογίσει τις αντίστοιχες εξόδους. Στο τέλος το μοντέλο μάθησης αξιολογείται για την πιστότητα των προβλέψεων του. Ένα σημαντικό θέμα όπου πρέπει κάποιος να επιστήσει την προσοχή του είναι το Bias–variance tradeoff και αναφέρεται στην κατάλληλη αναλογία training set και test set ώστε το μοντέλο να επιτύχει την μεγαλύτερη ακρίβεια στην προβλέψεις του.

  Τα δύο βασικά προβλήματα στα οποία βρίσκουν εφαρμογή οι αλγόριθμοι επιβλεπόμενης μάθησης είναι σε προβλήματα ταξινόμησης και προβλήματα παρεμβολής. Ενδεικτικά αναφέρονται ορισμένοι αλγόριθμοι επιβλεπόμενης μάθησης είναι οι εξής:
  - Δέντρα απόφασης
  - Support Vector Machine
  - Neural Network
  - Linear Regression
- **Μη Επιβλεπόμενη Μάθηση, unsupervised learning:** Στην διαδικασία της μη επιβλεπόμενης μάθησης το σύστημα καλείται να ανακαλύψει τη δομή των δεδομένων εισόδου, τυχόν συσχετίσεις και κρυμμένα μοτίβα χωρίς τις να γνωρίζει τις τιμές εξόδου χωρίς δηλαδή να του παρέχεται κάποια εμπειρία μιας και τα δεδομένα που χρησιμοποιεί δεν φέρουν ετικέτες "labels". Συνεπώς στο τέλος δεν υπάρχει διαδικασία αξιολόγησης του μοντέλου με την μορφή που υφίσταται στην επιβλεπόμενη μάθηση (Ayodele, 2010; Figueirêdo, 2020). Ενδεικτικοί αλγόριθμοι είναι οι παρακάτω:
  - K-means
  - Hierarchical clustering
  - Neural Networks

- **Ενισχυτική Μάθηση:** Τα μοντέλα ενισχυτικής μάθησης αλληλεπιδρούν δυναμικά με το περιβάλλον και εκπαιδεύεται μέσα από μία διαδικασία επιβράβευσης και ποινών. Χρησιμοποιούνται κυρίως στη ρομποτική καθώς και στα ηλεκτρονικά παιχνίδια.

## 3.5 Χρονοσειρές

Μία χρονοσειρά είναι το σύνολο των ποσοτικών τιμών μίας μεταβλητής, οι οποίες συλλέγονται διαδοχικά και διαχρονικά και εκφράζουν την εξέλιξη των τιμών της μεταβλητής κατά την διάρκεια ίσων χρονικών περιόδων. Ειδικότερα δοθέντος ενός χαρακτηριστικού Α, μια χρονοσειρά αποτελείται από ένα σύνολο Ν παρατηρήσεων, της οποίας οι τιμές ελήφθησαν σε ίσες χρονικές στιγμές. Για παράδειγμα, η καταγραφή των τιμών ενός αισθητήρα στη διάρκεια ενός κύκλου εργασίας σε μία παραγωγική διαδικασία αποτελεί μία χρονοσειρά. Στην ανάλυση των χρονοσειρών η ιδέα βασίζεται στην υπόθεση ότι είναι εφικτή η προεκβολή των παρελθοντικών τιμών αναλύοντας τα ιδιαίτερα χαρακτηριστικά τους (Μαργιά, 2009). Δηλαδή χρησιμοποιώντας και αναλύοντας τις ιστορικές τιμές μίας μεταβλητής, μπορούμε να εξάγουμε συμπεράσματα για την εξέλιξη (πρόβλεψη) της μεταβλητής σε μελλοντικά χρονικά διαστήματα (Κουγιουμτζής, 2011). Τα ιδιαίτερα χαρακτηριστικά μια χρονοσειράς μπορεί να είναι η εποχικότητα, ανοδικές ή καθοδικές τάσεις κλπ (Μπεγκόμ, 2013; Καλαμβόκη, 2017).

Από μαθηματικής σκοπιάς, μια χρονοσειρά ορίζεται από ένα σύνολο τιμών της μορφής $y_1$, $y_2$, $y_3$,…, $y_N$ όπου ο δείκτης Ν παριστάνει τις ίσες χρονικές στιγμές t, και αποτελούν μία άπειρη ακολουθία τιμών της μεταβλητής Υ. Οι χρονοσειρές διακρίνονται σε συνεχείς και διακριτές. Στη συνεχείς χρονοσειρές η καταγραφή των τιμών της εξεταζόμενης μεταβλητής γίνεται σε συνεχείς χρονική διάρκεια πχ ημερήσια καταγραφή θερμοκρασίας, ενώ στις διακριτές σε συγκεκριμένα χρονικά διαστήματα πχ μηνιαίες πωλήσεις αυτοκινήτων (Μαργιά, 2009; Καλαμβόκη, 2017).

Οι αναλύσεις χρονοσειρών βρίσκουν ποικίλες εφαρμογές σε διάφορα πεδία όπως:

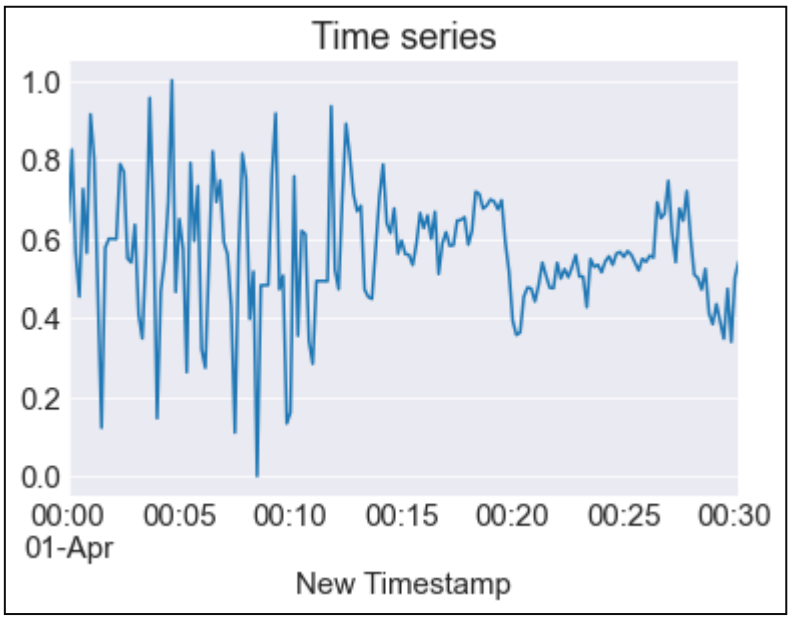
- Χρηματοοικονομικά
- Οικονομικά
- Ιατρικής
- Βιομηχανικής παραγωγής
- Μάρκετινγκ

## 3.6 Βασικά χαρακτηριστικά χρονοσειράς

Το αρχικό βήμα για τη σωστή μελέτη και ανάλυση μια χρονοσειράς είναι η δισδιάστατη απεικόνισή των τιμών των δεδομένων συναρτήσει του χρόνου t. Η απεικόνιση είναι μείζονος σημασίας τόσο για την ανάλυση όσο και για την πρόβλεψη των μελλοντικών τιμών της. Για

την απεικόνιση των δεδομένων το κυριότερο γράφημα που χρησιμοποιείται είναι το διάγραμμα χρόνου.

Το διάγραμμα χρόνου απεικονίζει τις παρελθοντικές τιμές συναρτήσει του χρόνου όπως φαίνεται στην παρακάτω εικόνα (Εικόνα 3).




*Εικόνα 3. Απεικόνιση χρονοσειράς*

Μέσω του διαγράμματος χρόνου αποβλέπουμε στην κλασσική ανάλυση της χρονοσειράς στα επιμέρους μέρη της δηλαδή τα βασικά της χαρακτηριστικά όπως συνοψίζονται παρακάτω (Μαργιά, 2009; Καλαμβόκη, 2017):

- **Τάση (Trend):** Ως τάση ορίζεται η μακροπρόθεσμη μεταβολή του μέσου επιπέδου των τιμών μιας χρονοσειράς στη μονάδα του χρόνου. Κατ' επέκταση η τάση μπορεί να είναι ανοδική, πτωτική ή σταθερή σε ένα συγκεκριμένο χρονικό διάστημα, όπως απεικονίζεται και στο παρακάτω σχήμα, και μπορεί να εκτιμηθεί διαγραμματικά.
- **Εποχικότητα (Seasonal):** Εκφράζεται ως μια περιοδική διακύμανση στις τιμές μιας χρονοσειράς λόγω εποχιακών παραγόντων, σε συγκεκριμένες χρονικές στιγμές και διάρκεια μικρότερη του έτους.
- **Κυκλικότητα (Cyclical):** Η κυκλικότητα είμαι μία εποχικότητα αλλά όχι σε σταθερές περιόδους, και διάρκεια μεγαλύτερη του έτους.
- **Ακραίες τιμές (Outliers):** Είναι τιμές μίας εξεταζόμενης μεταβλητής οι οποίες διαφέρουν σημαντικά από τη μέση τιμή των υπολοίπων.
- **Τυχαίο κομμάτι:** Το τυχαίο κομμάτι της χρονοσειράς, όπου παρουσιάζονται διακυμάνσεις λόγω τυχαίων γεγονότων χωρίς δυνατότητα ερμηνείας.
- **Στασιμότητα:** Ένα επιπλέον χαρακτηριστικό το οποίο αποτελεί βασική προϋπόθεση στην ανάλυση των χρονοσειρών είναι η ύπαρξη στασιμότητας. Με τον όρο στασιμότητα εννοούμε ότι ο μέσος όρος και οι διακυμάνσεις των τιμών μια

χρονοσειράς δε διαφοροποιούνται με το χρόνο. Μία μη στάση χρονοσειρά μπορεί να παρουσιάζει αλλαγές στη μέση τιμή, αυξητικές ή πτωτικές τάσεις και εποχικότητα.

Ο τρόπος σύνδεσης των παραπάνω βασικών χαρακτηριστικών μιας χρονοσειράς ορίζεται με τις δύο παρακάτω προσεγγίσεις (Πίνακας 1):

*Πίνακας 1. Προσεγγίσεις Χρονοσειρών*

| | |
|---|---|
| Προσθετικό μοντέλο : | $Yt = Τάση t + Εποχικότητα_t + Κυκλικότητα_t + Τυχαίο κομμάτι_t$ |
| Πολλαπλασιαστικό μοντέλο: | $Y_t = Τάση_t * Εποχικότητα_t * Κυκλικότητα_t * Τυχαίο κομμάτι_t$ |

# Κεφάλαιο 4

## 4.1 Εισαγωγή

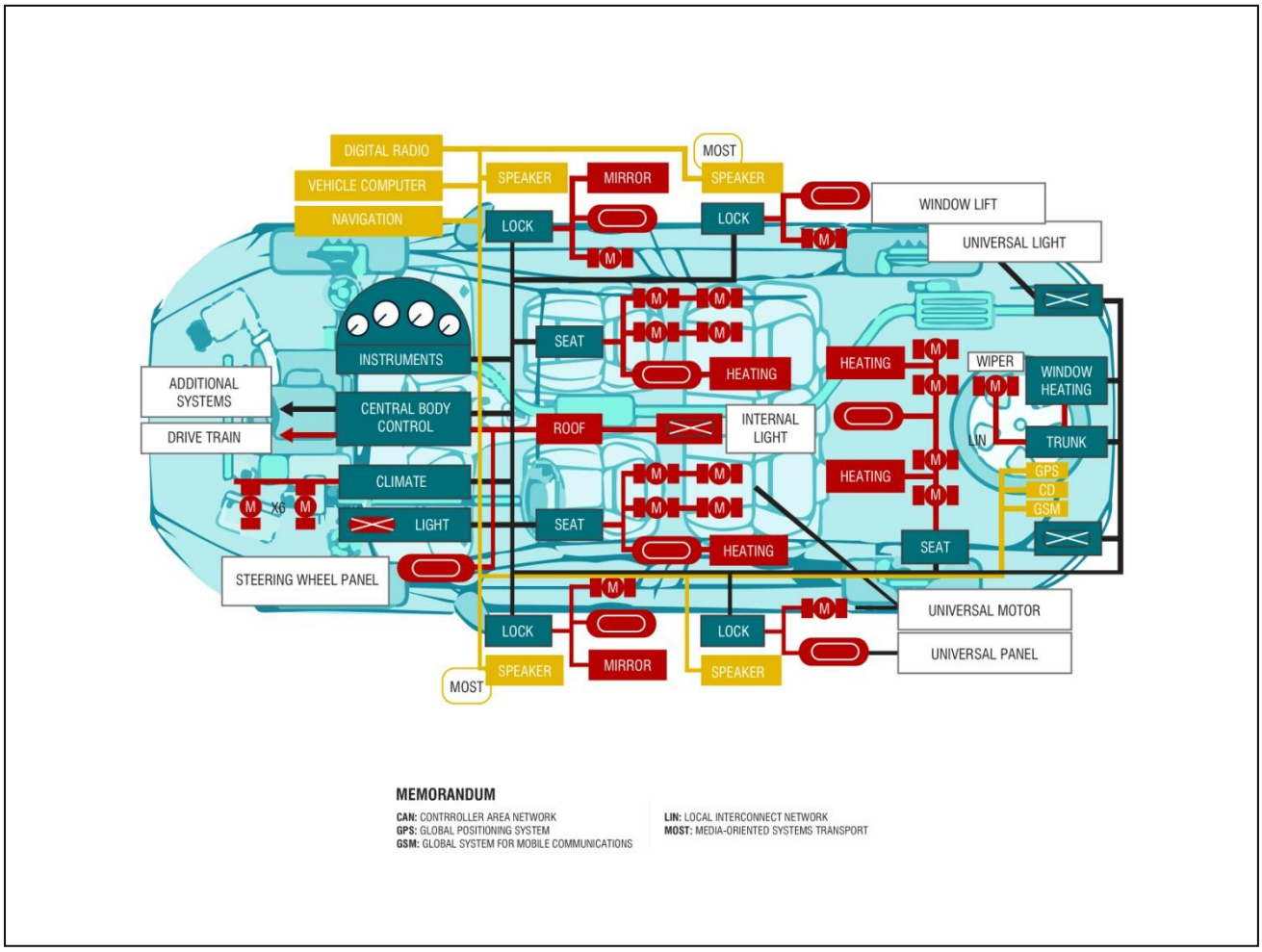
Ο πρωταρχικός στόχος της συγκεκριμένης ποσοτικής έρευνας είναι να υλοποιήσουμε μεθόδους και τεχνικές συσταδοποίησης χρησιμοποιώντας δεδομένα προερχόμενα από τους αισθητήρες λειτουργίας του CAN Bus συστήματος του οχήματος. Τα δεδομένα εκφράζουν την εξέλιξη διαφόρων μεταβλητών συναρτήσει του χρόνο συνεπώς χαρακτηρίζονται ως χρονοσειρές. Τελικός στόχος είναι οι συσταδοποίηση των χρονοσειρών αυτών σε ομάδες συγκρίνοντας μεταξύ τους και διακρίνοντας τους οδηγούς και τις διαδρομές τους και κατηγοριοποιώντας τους ανάλογα με τον τρόπο οδήγησης τους. Όπως αναφέρθηκε και στην βιβλιογραφική αναφορά, έχουν υλοποιηθεί αρκετές ερευνητικές εργασίες στο συγκεκριμένο πεδίο. Ως επί το πλείστον χρησιμοποιήθηκαν δεδομένα από τα CAN Bus συστήματα των οχημάτων και σε ορισμένες περιπτώσεις τροχιοδεικτικά δεδομένα καθώς και εικόνες-βίντεο καταγεγραμμένα κατά τη διάρκεια της οδήγησης (Csselectronics, n.d.). Όμως οι βασικές διαφορές είναι ότι τα δεδομένα εξετάστηκαν, είτε αγνοώντας τελείως τη διάσταση του χρόνου, είτε έχοντας γίνει pre-labeled σε κλειστές διαδρομές με τη χρήση προσημειώσεων και χρησιμοποιώντας supervised αλγορίθμους μηχανικής μάθησης. Για παράδειγμα, στην έρευνα Naiwala et al. (2016) εφαρμόστηκαν supervised αλγόριθμοι μηχανικής μάθησης. Τα δεδομένα είχαν κατηγοριοποιηθεί και περιείχαν labels διακρίνοντας τα σε low/average-skilled, high-skilled εκ των προτέρων. Αντίθετα, στην παρούσα έρευνα θα χρησιμοποιηθούν unlabeled δεδομένα πολλαπλών χρονοσειρών ώστε να εφαρμόσουμε τεχνικές whole time series clustering με τη χρήση lock-step και elastic μετρικών απόστασης. Τέλος, κατόπιν ερευνητικής ανάλυσης των clusters θα καταλήξουμε στη διάκριση των χρονοσειρών και κατ' επέκταση των οδηγών και των διαδρομών με διάκριση την οδηγική συμπεριφοράς τους.

## 4.2 Δεδομένα εισαγωγής

Η δημόσια και η ιδιωτική ασφάλεια αναφέρεται στην προστασία του πληθυσμού καθώς και των μεμονωμένων ατόμων μέσω της πρόληψης και της προστασίας από κινδύνους όπως ατυχήματα, εγκλήματα και καταστροφές. Συχνά αποτελεί κυβερνητική ευθύνη η τήρηση των παραπάνω, μέσω των σωμάτων ασφαλείας, των ιατρικών υπηρεσιών κ.α. αλλά σημαντικό ρόλο διαδραματίζουν πλέον ανεξάρτητες αρχές και οργανισμοί (Civilprotection, 2020). Κοινός παράγοντας αυτής της προσπάθειας αποτελεί η συλλογή και η πρόσβαση σε δεδομένα τα οποία προέρχονται όμως από πληθώρα διαφορετικών πηγών. Αυτό το πρόβλημα καλείται να επιλύσει το "AEGIS - Advanced Big Data Value Chain for Public Safety and Personal Security" big data project το οποίο αποτελεί μια καινοτόμα ιδέα όπου συνδυάζοντας

τεχνολογίες Linked Data και Big Data παρέχει μία ολοκληρωμένη πλατφόρμα συλλογής και διαχείρισης δεδομένων δίνοντας στο κοινό τη δυνατότητα πρόσβασης σε αυτά (Biliri, 2017; European Commission, 2019b).

Τα ερευνητικά δεδομένα που υπάρχουν διαθέσιμα, προήλθαν από το " AEGIS - Advanced Big Data Value Chain for Public Safety and Personal Security" big data project το οποίο χρηματοδοτήθηκε από το European Union's Horizon 2020 (Zenodo, 2019). Αποτελούνται από δεδομένα χρονολογικών σειρών τα οποίες συλλέχτηκαν κατά τη διάρκεια 6 διαδρομών διάρκειας από 00:30:00 λεπτά έως και 2:12:12 ώρες:λεπτά από τρείς διαφορετικούς οδηγούς (2 διαδρομές / οδηγό), στο οδικό δίκτυο της Αυστρία μέσω του συστήματος CAN BUS συστήματος του οχήματος (Εικόνα 4). Τι είναι όμως το CAN Bus σύστημα;




*Εικόνα 4. Controller Area Network Systems*

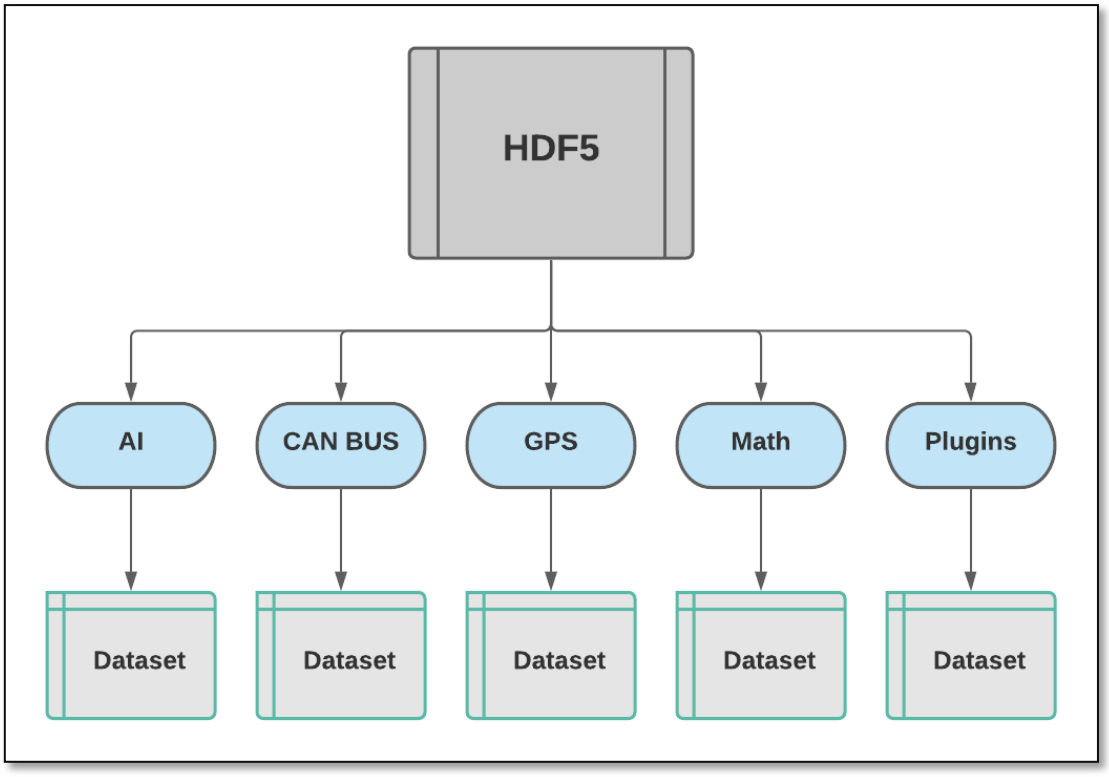
Τα σύγχρονα αυτοκίνητα πέραν του εγκεφάλου ECU περιέχουν και επιμέρους μονάδες ελέγχου όπως είναι οι αερόσακοι, ο κλιματισμός καθώς και διάφοροι περιφερειακοί αισθητήρες απαραίτητοι για την λειτουργία του οχήματος. Το CAN Bus (Controller Area Network) είναι ένα σύστημα σειριακού διαύλου το όποιο επιτρέπει την επικοινωνία μεταξύ όλων των επιμέρους συστημάτων και την ανταλλαγή πληροφοριών μεταξύ αυτών

(Bozdal et al., 2020; Csselectronics, n.d.; Copperhilltech, n.d.). Το σύνολο των δεδομένων προήλθε από το CAN Bus σύστημα ενός οχήματος αγνώστων στοιχείων.

## 4.3 Δεδομένα του πειράματος

Τα δεδομένα που χρησιμοποιήθηκαν στο πείραμα περιέχουν μετρήσεις από τους αισθητήρες του CAN Bus συστήματος ενός οχήματος. Το σύνολο δεδομένων περιέχει μετρήσεις προερχόμενα από το παραπάνω σύστημα με ρυθμό καταγραφής στα 20HZ, αποθηκευμένα σε αρχείο τύπου hdf. Ένα hdf file έχει ιεραρχική δομή όπου περιέχει 2 βασικές οντότητες όπως αναπαρίστανται στη παρακάτω εικόνα 5 :

- τα groups
- τα datasets




*Εικόνα 5. Μέθοδος ανάλυσης δεδομένων*

Το κάθε αρχείο περιέχει συνολικά 5 group πινάκων όπως απεικονίζονται στα παραπάνω σχήμα (Εικόνα 5). Έπειτα από διερευνητική ανάλυση των δεδομένων και για την εξαγωγή ποιοτικότερων αποτελεσμάτων κρίθηκε σκόπιμο η εξαίρεση των group με ονομασία AI, Math, GPS κυριώς διότι περιέχουν μεταβλητές με Null values που προήλθαν πιθανόν από σφάλματα κατά την καταγραφή των αισθητήρων. Αντίστοιχα διατηρούμε και εισάγουμε τα δεδομένα των CAN Bus και Plugins σε δομές δεδομένων τύπου data frames. Επειδή οι τιμές των παρακάτω μεταβλητών δίνονται σε κάποιο χρόνο δειγματοληψίας αποτελούν χρονοσειρές. Ακολουθεί μια σύνοψη και σύντομη περιγραφή των εξεταζόμενων μεταβλητών:

- **Timestamp:** Απεικονίζει το χρόνο δειγματοληψίας t.

- **Accelerometer_X- Y_Z:** Η επιτάχυνση είναι ένα διανυσματικό μέγεθος και εκφράζει τη μεταβολή της ταχύτητας ως προς στη μονάδα του χρόνου. Ως διανυσματικό μέγεθος χαρακτηρίζεται από το μέτρο (μέγεθος) καθώς και την κατεύθυνση (διεύθυνση και φορά). Το σύνολο των δεδομένων περιλαμβάνει τιμές επιτάχυνσης και στους τρεις άξονες x, y, z.
- **Vehicle speed:** Ως ταχύτητα ενός σώματος ορίζεται ο ρυθμός μεταβολής της θέσης του ως προς το χρόνο. Είναι αντίστοιχα διανυσματικό μέγεθος, συνεπώς χαρακτηρίζεται τόσο από το μέτρο (μέγεθος) της, όσο και από τη φορά (κατεύθυνση) της.

Για τους σκοπούς της παρακάτω έρευνας χρησιμοποιήθηκαν οι τιμές της επιτάχυνσης στον διαμήκη άξονα καθώς και η ταχύτητα του οχήματος συναρτήσει του χρόνου. Συνολικά το σύνολο των δεδομένων περιέχει δεδομένα επιτάχυνσης και ταχύτητας του οχήματος για τρείς διαφορετικούς οδηγούς με δύο διαφορετικές διαδρομές ανά οδηγό. Τα δεδομένα εφόσον περιέχουν την διάσταση του χρόνου αποτελούν χρονοσειρές.

## 4.4 Μεθοδολογία

Η συσταδοποίηση αποτελεί μια από τις πιο σύνηθες τεχνικές μη επιβλεπόμενης μάθησης. Χρησιμοποιείται για την εύρεση μοτίβων μέσα από την τμηματοποίηση ενός συνόλου δεδομένων σε συστάδες. Ως συστάδα ορίζεται μια συλλογή ομοιογενών ομαδοποιημένων αντικειμένων προερχόμενα από ένα σύνολο δεδομένων. Η συσταδοποίησης αποτελεί μία πολύ σημαντική διαδικασία στην διερευνητική ανάλυση των δεδομένων καθώς καταλήγει στην αναγνώριση εσωτερικών δομών σε αυτά χωρίς προηγουμένως να έχει οριστεί σαφώς κάποια κατηγοριοποίηση τους. Τελικός στόχος της συσταδοποίησης είναι τα στοιχεία που ανήκουν σε μία συστάδα να είναι περισσότερο όμοια μεταξύ τους εν συγκρίσει με τα στοιχεία των άλλων συστάδων. Ειδικότερα τα τελευταία χρόνια με τις ολοένα και αυξανόμενες δυνατότητες αποθήκευσης και επεξεργασίας, τα δεδομένα των περισσότερων εφαρμογών αποθηκεύονται σε μορφές χρονοσειρών. Επομένως η συσταδοποίηση είναι μια προσφιλή τεχνική με ευρεία χρήση σε πολλά πεδία εφαρμογών όπως του μάρκετινγκ, των οικονομικών, της ιατρικής κλπ (Aghabozorgi et al., 2015).

Οι χρονοσειρές μπορούν να χαρακτηριστούν ως δυναμικά δεδομένα καθώς μεταβάλλονται συναρτήσει του χρόνου, καταλαμβάνουν μεγάλο χώρο στην αποθήκευσης τους και συνήθων είναι πολυδιάστατα. Επομένως η εφαρμογή μεθόδων συσταδοποίησης σε αυτό το είδος των δεδομένων αποτελεί πρόκληση καθώς διαφοροποιείται από τις σύνηθες πρακτικές. Ειδικότερα η συσταδοποίηση χρονοσειρών μπορεί να αποτυπωθεί με τον παρακάτω ορισμό:

Έστω ένα σύνολο δεδομένων με τις παρακάτω δοθέντες n χρονοσειρές

$$T = \{T1, T2, T3, ..., Tn\}$$

Μέσω της συσταδοποίησης δύναται να επιτευχθεί διαμέριση του αρχικού συνόλου σε k ομάδες

$$C = \{C1, C2, C3, ..., CK\},$$

με τέτοιο τρόπο ώστε να αποτελούν υποσύνολο του αρχικού συνόλου και μάλιστα χωρίς να είναι επικαλυπτόμενες. Δηλαδή:

$$T = \cup k\, i=1\ \ Ci \quad και\ Ci \cap Cj = \emptyset \quad με\ i \neq j$$

Επιπροσθέτως, δομικό στοιχείο στη διαδικασία της συσταδοποίησης είναι ο υπολογισμός των μέτρων ομοιότητας και των μέτρων απόστασης όπως θα αναλυθεί στην επόμενη ενότητα. Το στοιχείο αυτό συνεπάγεται τη διάκριση της συσταδοποίησης σε δύο κατηγορίες όπως αναλύονται παρακάτω (Εικόνα 6) (Keogh & Lin, 2005):

- **Whole (Sequence) time-series clustering:** Το συγκεκριμένο είδος συσταδοποίησης, ειδικεύεται στη σύγκριση και ομαδοποίηση με βάση την ομοιότητα τους ενός πλήθους n ανεξάρτητων χρονοσειρών. Ο υπολογισμός των μέτρων ομοιότητας γίνεται σε ολόκληρο το μήκος των χρονοσειρών
- **Subsequence time series clustering:** Σε αυτό το είδος συσταδοποίησης, αρχικά πραγματοποιείται αποκλειστικά σε μία μόνο χρονοσειρά. Ένα κινούμενο χρονικό παράθυρο διατρέχει το μήκος της χρονοσειράς χωρίζοντας την σε μικρότερα τμήματα τα οποία διακρίνει σε συστάδες χρησιμοποιώντας τα μέτρα ομοιότητας.

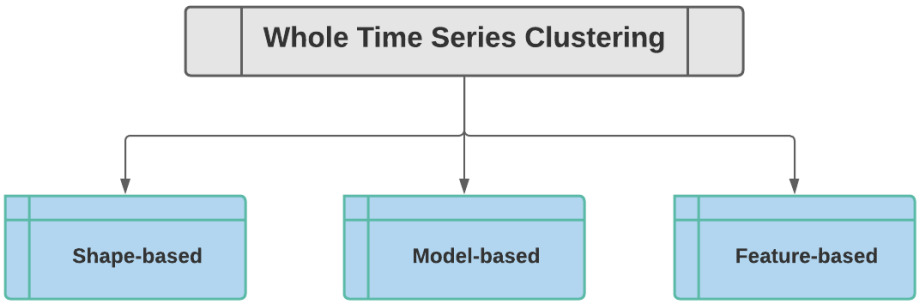

*Εικόνα 6. Προσεγγίσεις στην συσταδοποίηση χρονοσειρών*

Εκτός από την παραπάνω κατηγοριοποίηση, ο τρόπος που υλοποιείται η συσταδοποίηση διακρίνεται σε 4 κατηγορίες (Aghabozorgi et al., 2015):

- **Feature-based:** Στη συγκεκριμένη προσέγγιση, οι πολυδιάστατες χρονοσειρές μετατρέπονται σε διανύσματα μικρότερων διαστάσεων. Στη συνέχεια εφαρμόζεται ένας οποιοσδήποτε σύνηθες αλγόριθμος κατηγοριοποίησης.

- **Model-based:** Υλοποιείται εφαρμόζοντας μιας μετατροπή στα δεδομένα της χρονοσειράς. Συγκεκριμένα αυτή η προσέγγιση υποθέτει ότι τα δεδομένα προήλθαν από κάποιο μοντέλο, συνεπώς προσπαθεί να κατασκευάσει το μοντέλο μέσα από αυτά. Εφόσον κατασκευαστεί θα καθορίσει και τις ομάδες τους.

- **Shaped-based:** Στη συγκεκριμένη προσέγγιση συσταδοποίησης, επιχειρούν να φέρουν στη ίδια συστάδα χρονοσειρές, λαμβάνοντας ως κριτήριο τη μικρότερη δυνατή απόσταση μεταξύ τους. Για να επιτευχθεί αυτό χρησιμοποιούνται μέτρα ομοιότητας κατάλληλα για δεδομένα χρονοσειρών.




*Εικόνα 7. Μέθοδοι συσταδοποίησης*

Όπως αναφέρθηκε στην εισαγωγή του κεφαλαίου 4, στόχος της εργασίας είναι να συγκριθούν οι χρονοσειρές με δεδομένα διαφορετικών οδηγών σε διαφορετικές διαδρομές και να επιτευχθεί ο διαχωρισμός τους σε συστάδες. Επομένως η μέθοδος που υλοποιήθηκε στην πειραματική διαδικασία είναι η whole time series clustering και ειδικότερα η προσέγγιση της έρευνας βασίστηκε σε shape-based clustering (Εικόνα 7).

### 4.4.1 Μέτρα ομοιότητας – Μέτρα απόστασης

Βασικός στόχος της συσταδοποίησης είναι η τμηματοποίηση ενός συνόλου δεδομένων σε συστάδες ώστε τα στοιχεία του συνόλου των δεδομένων που ανήκουν σε μία συστάδα να είναι περισσότερο όμοια μεταξύ τους εν συγκρίσει με τα στοιχεία των άλλων συστάδων. Τα

μέτρα ομοιότητας και τα μέτρα απόστασης χρησιμοποιούνται για να περιγράψουν ποσοτικά την ομοιότητα δύο σημειακών δεδομένων. Συνεπώς κάθε αλγόριθμος συσταδοποίησης χρονοσειρών στηρίζεται στα μέτρα ομοιότητας ή απόστασης ώστε να είναι εφικτή μια ανάλυση σε συστάδες. Ειδικότερα τα μέτρα ομοιότητας ή απόστασης μπορούν να αποτυπωθούν με τον παρακάτω ορισμό:

Έστω δύο χρονοσειρές T, S με μήκος T:

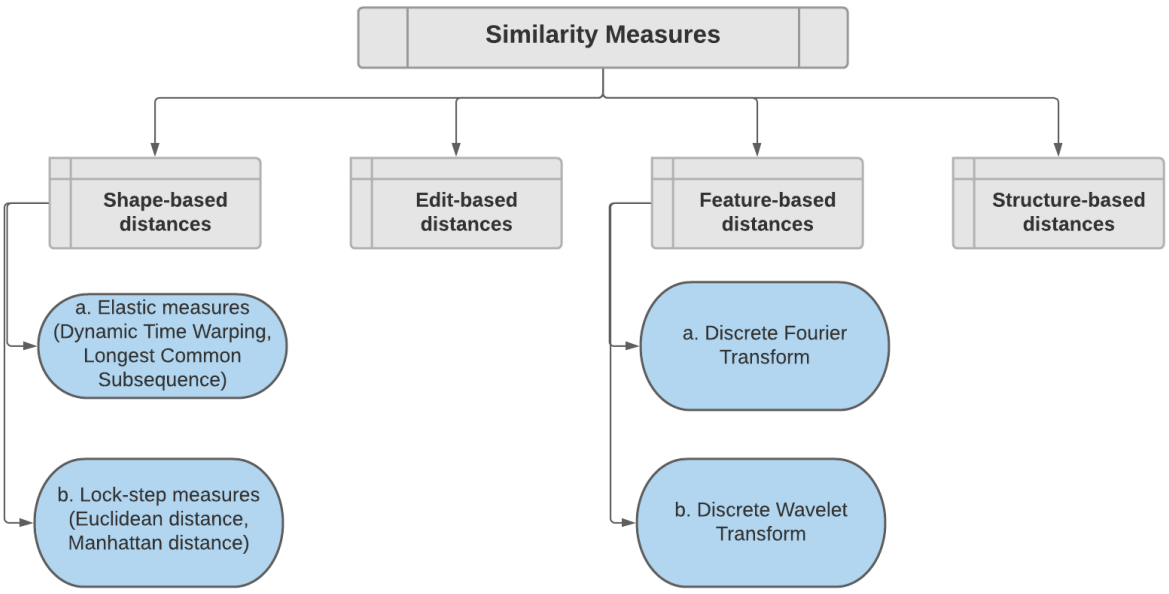
$$T = \{T_1, T_2, T_3, \ldots, T_T\} \text{ και } S = \{S_1, S_2, S_3, \ldots, S_T\}$$

Η απόσταση μεταξύ των δύο χρονοσειρών και όλων των σημείων τους ορίζεται ως εξής:

$$Dist\,(T,S) = \sum_{t=1}^{T} dist(Tt, St)$$

Υπάρχουν διάφορές μέτρα ομοιότητας που προσεγγίζουν το ζήτημα του υπολογισμού της απόστασης μεταξύ δύο χρονοσειρών και μπορούν να κατηγοριοποιηθούν σε 4 κατηγορίες (Εικόνα 8) (Liao, 2005; Esling & Agon, 2012; Wang et al., 2013):

1. Shape-based distances
   a. Elastic measures (Dynamic Time Warping, Edit Sequence on Real Sequence)
   b. Lock-step measures (Euclidean distance, Manhattan distance)
2. Edit-based distances
3. Feature-based distances
   a. Discrete Fourier Transform
   b. Discrete Wavelet Transform
4. Structure-based distances




*Εικόνα 8. Μέτρα ομοιότητας- Μέθοδοι συσταδοποίησης*

**Shape-based distances**

Στην πρώτη κατηγορία μέτρων ομοιότητας - απόστασης (Shape-based distances), όπως δηλώνει και το όνομα τους εξετάζουν τη συνολική δομή (shape) και τη ομοιότητα μεταξύ των εξεταζόμενων χρονοσειρών μέσω της τοπικής τους σύγκρισης. Διακρίνουμε 2 υποκατηγορίες τα elastic measure και τα lock-step measure. Τα lock-step measure εξετάζουν τα χρονικά σημεία t των εξεταζόμενων χρονοσειρών ένα προς ένα με βασική προϋπόθεση οι χρονοσειρές να είναι ίδιου μήκους, ενώ στα elastic measure γίνεται σύγκριση μεταξύ των σημείων των εξεταζόμενων χρονοσειρών χωρίς απαραίτητα να είναι ένα προς ένα αλλά ένα προς πολλά (Aghabozorgi et al., 2015; Özkoç, 2020).

Η δεύτερη κατηγορία μετρικών ομοιότητας (Edit-based distances) βασίζεται στην ανομοιομορφία μεταξύ δύο χρονοσειρών και συγκεκριμένα στον ελάχιστο αριθμό παρατηρήσεων που απαιτούνται ώστε η μία χρονοσειρά να μετατραπεί στην άλλη.

Η τρίτη κατηγορία (Feature-based distances) συχνά χρησιμοποιούνται για να μειωθούν οι διαστάσεις ενός πολυδιάστατου συνόλου δεδομένων, εξάγοντας χαρακτηριστικά από τις χρονοσειρές και μετρώντας την απόσταση μεταξύ αυτών.

Τέλος, η τέταρτη κατηγορία (Structure-based distances) αφορά συγκρίσεις μεταξύ μεγάλων διαστάσεων δομών χρονοσειρών.

Στην παρούσα πειραματική μελέτη χρησιμοποιήθηκαν μετρικές της πρώτης και τρίτης κατηγορίας οι οποίες αναλύονται στην επόμενη ενότητα.

**Lock-step measures:**

Σε αυτή την υποενότητα θα εξετάσουμε τα Lock step μέτρα που χρησιμοποιήθηκαν και συγκεκριμένα την Ευκλείδεια απόσταση, την Μανχάταν απόσταση. Έστω οι χρονοσειρές T, S μήκους N, M αντίστοιχα. Απαραίτητη προϋπόθεση για τη χρήση των lock step μετρικών είναι η ομοιογένεια του μήκους μεταξύ των χρονοσειρών (N = M), και η σύγκριση του χρονικού σημείου i της T χρονοσειράς με το αντίστοιχο ίδιο χρονικό σημείο i της δεύτερης χρονοσειράς S. Γίνεται αμέσως αντιληπτό ότι παραπάνω κατηγορία μετρικών είναι ευαίσθητη σε τυχόν ύπαρξη υστέρησης μεταξύ των εξεταζόμενων χρονοσειρών.

1. **Ευκλείδεια Απόσταση:**

Η πιο συνηθισμένη μετρική απόστασης για αριθμητικά δεδομένα είναι η Ευκλείδεια απόσταση. Έστω 2 διανύσματα ίσου μήκους x, y σε ένα n-διάστατο χώρο, η Ευκλείδεια απόσταση τους δίνεται από τον τύπο:

$$Dist\ (x, y) = (\sum_{i=1}^{n} |xi - yi|)\ 1/2$$

2. **Manhattan Απόσταση:**

Η απόσταση Manhattan ονομάζεται επίσης και "City block Distance" και ορίζεται ως το άθροισμα των αποστάσεων όλων των στοιχείων (Brownlee ,2020). Δηλαδή, για δύο σημεία $I_1$ και $I_2$ σε ένα n-διάστατο χώρο. Ουσιαστικά αποτελεί υποπερίπτωση της Ευκλείδειας απόσταση και δίνεται από τον τύπο:

$$Dist\ (x, y) = (\sum_{i=1}^{n} |xi - yi|)$$

3. **Συσχέτιση Pearson:**

Η μετρική αυτή χρησιμοποιεί μια περισσότερο πολύπλοκη σχέση για να υπολογίσει την ομοιότητα μεταξύ δύο διανυσμάτων λαμβάνοντας υπόψη την γραμμικής τους συσχέτιση (Borgatti, n.d.). Παίρνει τιμές στο κλειστό διάστημα p =[-1,1] όπου:

- το p = -1 δείχνει ισχυρή αρνητική γραμμική συσχέτιση
- το p = 0 δείχνει καμία γραμμική συσχέτιση
- το p = 1 δείχνει ισχυρή θετική γραμμική συσχέτιση

Η συσχέτιση κατά Pearson υπολογίζεται από τον ακόλουθο τύπο

$$\rho(x, y) = \frac{Cov(x,y)}{\sigma x \sigma y} = \frac{E\ [(x - \mu x)\ (y - \mu y)]}{\sigma x \sigma y} = \frac{\sum_{i=1}^{n}(xi - x^{-})(yi - y^{-})}{\sqrt{\sum_{i=1}^{n}(xi - x^{-})^{2}}\sqrt{\sum_{i=1}^{n}(yi - y^{-})^{2}}}$$

όπου το $\mu_x$,$\mu_y$ είναι οι μέσες τιμές των x, y και το $\sigma_x$,$\sigma_y$ οι διακυμάνσεις τους αντίστοιχα. Η απόσταση υπολογίζεται από τον τύπο:

$$dcor(x, y) = 1-\rho$$

**Elastic-step measures:**

Σε αυτή την υποενότητα θα εξετάσουμε τα elastic step μέτρα που εφαρμόσαμε στην πειραματική διαδικασία και συγκεκριμένα τη Dynamic Time Warping. Σε αντίθεση με την lock step κατηγορία, η elastic step επιτρέπει τη σύγκριση σημείων ένα προς πολλά ή ένα προς κανένα συνεπώς είναι πιο ανεκτική σε διαφορετικού μήκους καθώς και μη συγχρονισμένες χρονοσειρές.

1. **Δυναμική Χρονική Διαστρέβλωση (Dynamic Time Warping)**

Η μέθοδος της δυναμικής χρονικής διαστρέβλωσης χρησιμοποιείται για τον υπολογισμό της ομοιότητας μεταξύ δύο χρονοσειρών που διαφέρουν σε μήκος ή σε ταχύτητα.

Το βασικό πλεονέκτημα είναι ότι επιτρέπει τη μη γραμμική στρέβλωση (επιμήκυνση ή συρρίκνωση) μιας αλληλουχίας τιμών ώστε να ταιριάξει με μία άλλη αλληλουχία ακόμα και εάν παρουσιάζουν χρονική υστέρηση. Έστω ότι 2 χρονοσειρές είναι παρόμοιες αλλά με διαφορά χρονικής φάσης, τότε οποιαδήποτε lock step μετρική θα εμφάνιζε μεγάλη απόσταση σε σύγκριση ένα προς ένα σημείων, ενώ η DTW όχι. Βασικό της μειονέκτημα είναι η μεγάλη υπολογιστική της πολυπλοκότητα καθώς και η αντιστοίχηση όλων των σημείων των χρονοσειρών ακόμα και των ακραίων τιμών. Ο τρόπος υπολογισμού συνοψίζεται στα παρακάτω βήματα:

1. Αρχικά οι χρονοσειρές χωρίζονται σε ίσα σε πλήθος σημεία
2. Υπολογίζεται η Ευκλείδεια απόσταση μεταξύ του πρώτου σημείου της πρώτης χρονοσειράς και όλων των σημείων της δεύτερης χρονοσειράς και διατηρούμε την μικρότερη απόσταση.
3. Πηγαίνουμε στο δεύτερο σημείο της χρονοσειράς όπου πραγματοποιείται ξανά το δεύτερο βήμα επαναληπτικά για όλα τα σημεία.
4. Στο τέταρτο βήμα χρησιμοποιείται ως σημείο αναφοράς η δεύτερη χρονοσειράς και επαναλαμβάνονται τα βήματα 2, 3
5. Τελικώς έχουμε τις ελάχιστες αποστάσεις των πιθανών συνδυασμών μεταξύ των τιμών των χρονοσειρών.

Έστω δύο ακολουθίες αριθμών x, y μήκους n, m αντίστοιχα. Οι ακολουθίες αυτές αποτυπώνονται στους άξονες ενός πίνακας διαστάσεων i, j όπου αντιστοιχούν στα μήκη της κάθε ακολουθίας. Η διαφορά της απόστασης κάθε στοιχείου των ακολουθιών υπολογίζεται από τον τύπο:

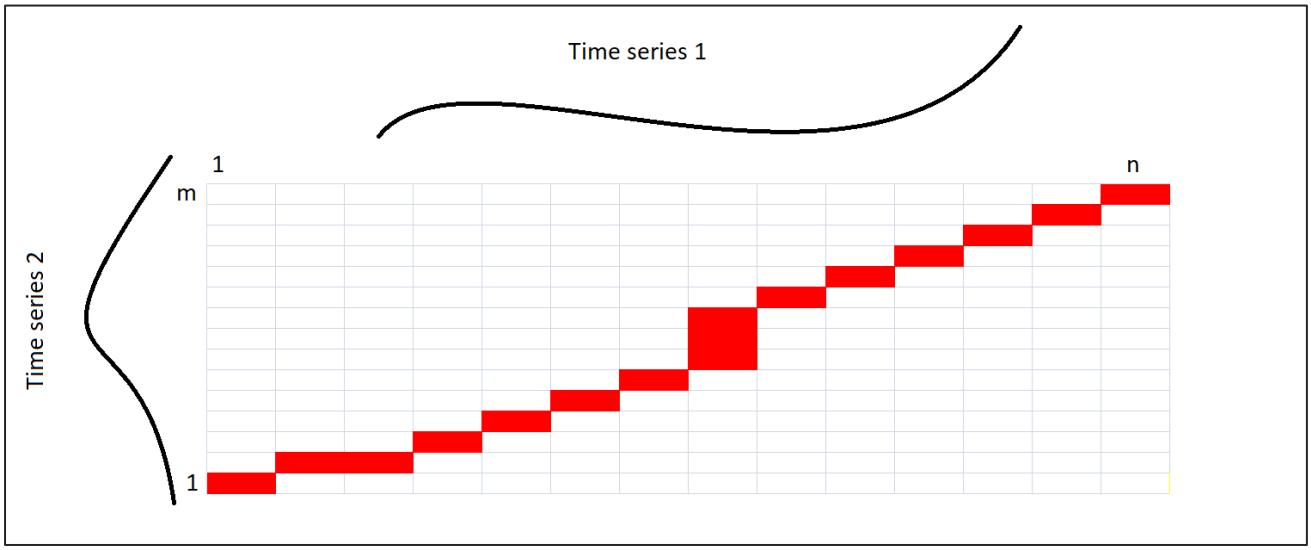
$$d(xi,yi) = (xi - yi)^2$$

Στη συνέχεια υπολογίζεται ένα wrapping path όπου ορίζεται ως:

$$W = w_1, w_2, w_3, \ldots, w_K \qquad \text{με} \qquad max(n,m) \leq K \leq m + n - 1$$

Τέλος, η συνολική απόσταση υπολογίζεται από τον τύπο:

$$D_{DTW}(x,y) = min\sqrt{\sum_{k=1}^{K} wk}$$




*Εικόνα 9. Υπολογισμός DTW Απόστασης*

Το μονοπάτι που σχηματίζεται με τις ελάχιστες αποστάσεις ονομάζεται μονοπάτι συγχρονισμού (alignment path) (Εικόνα 9). Για να μην αυξηθούν εκθετικά οι διαφορετικοί πιθανοί συνδυασμοί υπολογισμού του alignment path υπάρχουν 4 περιορισμοί (Niennattrakul & Ratanamahatana, 2007; Kyaagba, 2019):

- **Boundary Conditions:** Το alignment path πάντα θα ξεκινάει από την κάτω αριστερά γωνία και θα καταλήγει στην πάνω δεξιά γωνία. Με αυτόν τον τρόπο διασφαλίζουμε ότι δεν θα η αντιστοίχιση δεν θα περιοριστεί σε κάποια συγκεκριμένη ακολουθία.
- **Warping Window:** Το alignment path δεν θα αποκλίνει σημαντικά από τη διαγώνιο του πίνακα.
- **Monotonicity:** Το alignment path δεν μπορεί να κινηθεί αντίστροφα στην ροή του χρόνου.
- **Continuity:** Το alignment path δεν παρουσιάζει χρονικές ασυνέχειες

## 2. Longest Common Subsequence (LCSS)

Το μέτρο ομοιότητα LCSS βασίζεται πάνω στο Longest Common Subsequence πρόβλημα. Σύμφωνα με τη βιβλιογραφία (Aghabozorgi et al., 2011), έστω δύο δοθείσες ακολουθίες $x_i$, $y_i$ οπού περιέχουν αλφαριθμητικά δεδομένα. Εξετάζουμε να βρούμε τη μεγαλύτερη κοινή ακολουθία χαρακτήρων που περιέχεται και στις δύο ακολουθίες $x_i$, $y_i$. Προκειμένου να γίνει περισσότερο αντιληπτό ορίζουμε το $T_1$=ACDTGH και το $T_2$= SFACDTGS. Στο συγκεκριμένο παράδειγμα η LCS είναι η CDTG και μεγέθους ίσο με 4.

Το παραπάνω μέτρο ομοιότητας μπορεί να εφαρμοστεί και μεταξύ δύο χρονοσειρών με μία μικρή παραλλαγή. Ειδικότερα στις χρονοσειρές η ομοιότητα των ακολουθιών καθορίζεται λαμβάνοντας υπόψη και ένα όριο ομοιότητας θ. Το κατώφλι αυτό χρησιμοποιείται για να διακρίνει εάν δύο πραγματικοί αριθμοί ταυτίζονται ή όχι. Συγκεκριμένα εξετάζοντας δύο

σημεία εάν ταυτίζονται ή όχι υπολογίζεται η Ευκλείδεια απόστασή τους. Αν η απόσταση είναι μικρότερη της τιμή θ τα σημεία ταυτίζονται διαφορετικά δεν ταυτίζονται (Guo et al., 2016; Bergroth, 2000; Columbia University, n.d.). Έστω 2 χρονοσειρές $X_i=\{x_1,x_2,x_3,\ldots,x_n\}$ και $Y_i=\{y_1,y_2,y_3,\ldots,y_n\}$. Η LCSSD απόσταση μεταξύ των χρονοσειρών μπορεί να υπολογιστεί από τον παρακάτω τύπο:

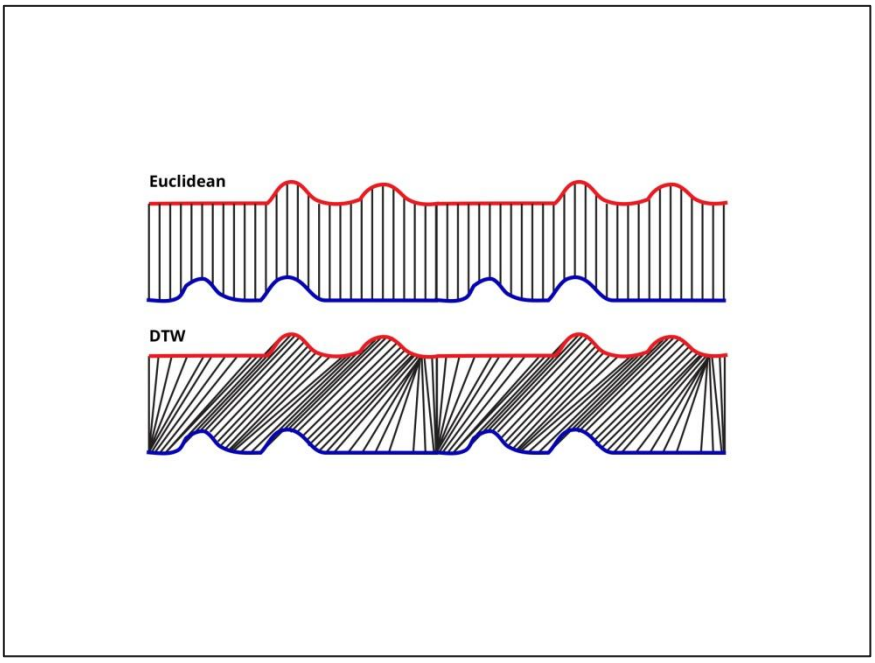
$$D_{lcss}(i,j) = \begin{cases} 0 & if\ i = 0\ or\ j = 0 \\ c(i-1, j-1) + 1 & if\ i,j > 0\ and\ xi = yi \\ \max\big(c(i-1,j), c(i,j-1)\big) & if\ \ i,j > 0\ and\ xi \neq yi \end{cases}$$

Συνοψίζοντας θα συγκρίνουμε τα παραπάνω μέτρα. Σε αυτό το σημείο θα προσθέσουμε μία ακόμα διάσταση για την αξιολόγηση των μέτρων απόστασης η οποία ονομάζεται χρονική πολυπλοκότητα. Ως χρονική πολυπλοκότητα ορίζουμε τη χρονική διάρκεια που απαιτείται για να ολοκληρώσει τους υπολογισμούς του ένας αλγόριθμος και συμβολίζεται με το γράμμα O(n) για n σημεία (Studytonight, n.d). Η χρονική πολυπλοκότητα υπολογίζεται μετρώντας τα βήματα που πραγματοποιούνται από έναν αλγόριθμο. Στον πίνακα 2 παρουσιάζεται η χρονική πολυπλοκότητα των παραπάνω μέτρων απόστασης.

*Πίνακας 2. Μέτρα απόστασης-χρονική πολυπλοκότητα*

| Μέτρα απόστασης | Χρονική πολυπλοκότητα |
|---|---|
| Ευκλείδεια | $O(nN^2)$ |
| Manhattan | $O(nN^2)$ |
| DTW | $O(nmN^2)$ |
| LCSS | $O(nmN^2)$ |

Γενικά στη συσταδοποίηση χρησιμοποιούνται ως επί το πλείστον τα lock step measures μέτρα απόστασης όπως η Ευκλείδεια, η Jaccard καθώς και η Pearson. Ενώ τα παραπάνω μέτρα βρίσκουν εφαρμογή σε πληθώρα εφαρμογών και μάλιστα αρκετά αποτελεσματικά δεν είναι εξίσου αποτελεσματικά όταν καλούνται να χειριστούν δεδομένα χρονοσειρών. Το πρόβλημα έγκειται στο γεγονός ότι δεν λαμβάνουν υπόψη την ύπαρξη χρονικών υστερήσεων, τα διαφορετικά μήκη χρονοσειρών καθώς και τις ακραίες τιμές (Wang et al., 2013). Αντίθετα τα elastic measures όπως το DTW και το LCSS χρησιμοποιούν τη χρονική διαστρέβλωση ώστε να αντιμετωπίσουν το πρόβλημα των ασυγχρόνιστων χρονοσειρών καθώς και των ακραίων τιμών όπως απεικονίζεται στην εικόνα 10. Το βασικό μειονέκτημα των elastic step measures βρίσκεται στο μήκος των χρονοσειρών. Παρατηρούμε στον πίνακα 2 ότι για μεγάλο μήκος χρονοσειρών αυξάνεται σημαντικά η πολυπλοκότητα και ο χρόνος υπολογισμού των αποστάσεων (Dunham, 2006; Oregi et al., 2019).




*Εικόνα 10. Ευκλείδεια Απόσταση vs Απόσταση Δυναμικής Χρονικής Διαστρέβλωσης*

### 4.4.2 Επιλογή Αλγορίθμου Συσταδοποίησης

Σε αυτήν την ενότητα θα παρουσιαστούν οι μέθοδοι συσταδοποίησης που υλοποιήθηκαν στην πειραματική διαδικασία. Η συσταδοποίηση όπως αναφέρθηκε αποτελεί ένα από τους πιο διαδεδομένους τρόπους εξόρυξης δεδομένων και γνώσης. Τελικός στόχος είναι η ομαδοποίηση των δεδομένων με βάση την ομοιότητα τους σε ομάδες.
Η συσταδοποίηση μπορεί να διακριθεί σε 5 διαφορετικές κατηγορίες (Dunham, 2006; Θεοδωρίδης & Πελέκης, χ.η.):

1. Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)
2. Συσταδοποίηση Διαμέρισης (Partitioning Clustering)
3. Συσταδοποίηση βασισμένη στην πυκνότητα (Density-based Clustering)
4. Συσταδοποίηση βασισμένη σε πλέγμα (Grid-based Clustering)
5. Συσταδοποίηση βασισμένη σε μοντέλα (Model-based Clustering)

Ειδικότερα σε συσταδοποίηση χρονοσειρών μπορούν να εφαρμοστούν οι περισσότερες από τις παραπάνω κατηγορίες, αλλά η επιλογή του κατάλληλου μέτρου ομοιότητας είναι επιλογή υψίστης σημασίας. Για την ανάλυση καταλήξαμε στην χρήση αλγορίθμων της πρώτης κατηγορίας οι οποίες και αναλύονται παρακάτω.

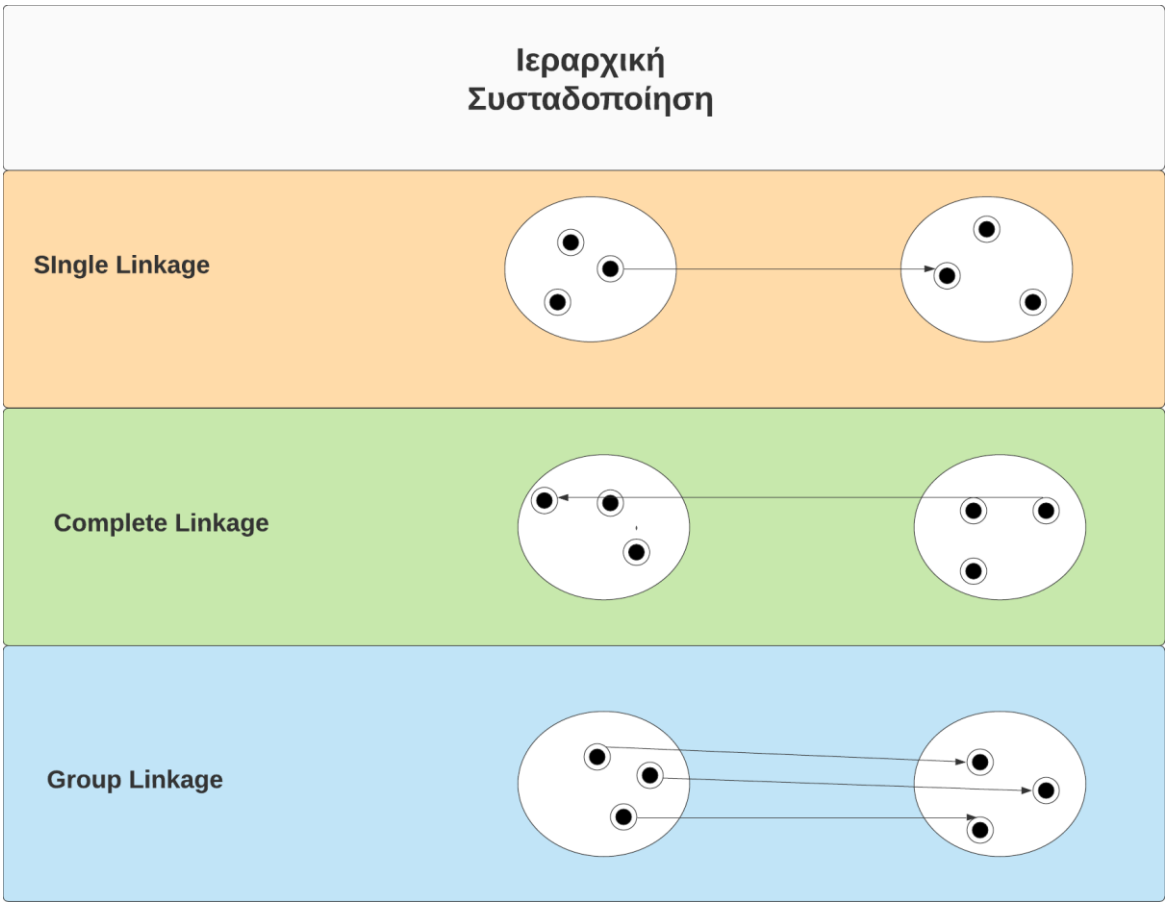
**Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)**

Στους αλγόριθμους ιεραρχικής συσταδοποίησης όπως δηλώνει και το όνομα τους δημιουργείται μια ιεραρχία εμφωλευμένων συστάδων. Οι συστάδες σχηματίζονται σταδιακά είτε με συνένωση μικρότερων ομάδων (συσσωρευτική μέθοδος) είτε με διαίρεση ομάδων σε μικρότερες (διαιρετική μέθοδος). Οι αλγόριθμοι μπορούν να αναπαρασταθούν με δενδρογράμματα όπου παρουσιάζεται η διάταξη των συστάδων που δημιουργήθηκαν κατά τη συσταδοποίηση. Βασικό πλεονέκτημα της ιεραρχικής συσταδοποίησης είναι η απουσία ορισμού αριθμών των συστάδων μιας και απλά μπορεί να επιτευχθεί ο επιθυμητός αριθμός κόβοντας το δενδρόγραμμα στο κατάλληλο επίπεδο. Επίσης είναι καλό να αποφεύγεται η χρήση τους σε μεγάλα σύνολα δεδομένων (Repositorykallipos, n.d.). Τέλος, διακρίνονται σε δύο κατηγορίες:

1. **Συσσωρευτικοί αλγόριθμοι (agglomerative algorithms)**

Οι συγκεκριμένη κατηγορία αλγορίθμων ξεκινάει με το σύνολο n των παρατηρήσεων να ανήκει σε n ομάδες-συστάδες. Σε κάθε επανάληψη συγχωνεύονται οι δύο πιο κοντινές σύμφωνα με τα κριτήρια ομοιότητας μέχρις ότου όλες οι παρατηρήσεις να ανήκουν σε μία μοναδική συστάδα. Επίσης, είναι απαραίτητο να ορίσουμε τους τρόπους προσδιορισμού της απόστασης μεταξύ των συστάδων (Εικόνα 11). Οι κυριότεροι είναι οι εξής:

- Ελάχιστης Απόστασης ή απλού συνδέσμου (Single linkage): Η ομοιότητα μεταξύ δύο συστάδων υπολογίζεται μεταξύ των δύο πιο κοντινών σημείων των συστάδων, δηλαδή εκείνα με την ελάχιστη απόσταση, όπως φαίνεται στην εικόνα στο πρώτο σχήμα.
- Μέγιστης Απόστασης ή πλήρους συνδέσμου (Complete linkage): Η ομοιότητα μεταξύ δύο συστάδων υπολογίζεται μεταξύ των δύο πιο μακρινών σημείων των συστάδων, δηλαδή εκείνα με την μεγαλύτερη απόσταση, όπως φαίνεται στην εικόνα στο δεύτερο σχήμα.
- Απόσταση μέσων όρων συστάδων (Group linkage): Η ομοιότητα μεταξύ δύο συστάδων υπολογίζεται ως η μέση τιμή των αποστάσεων των σημείων μεταξύ των συστάδων, όπως φαίνεται στην εικόνα στο τρίτο σχήμα.
- Απόσταση κεντρικών σημείων: Η ομοιότητα μεταξύ δύο συστάδων υπολογίζεται μεταξύ των δύο κέντρων των συστάδων.




*Εικόνα 11. Ιεραρχική συσταδοποίηση*

## 2. Διαιρετικοί αλγόριθμοι (Divisive algorithm)

Σε αυτήν την κατηγορία αλγορίθμων αρχικά όλα οι παρατηρήσεις ανήκουν στην ίδια συστάδα. Σε κάθε επανάληψη μία συστάδα διασπάται σε δύο νέες μέχρις ότου καταλήξουμε σε n πλήθος συστάδων όσες και το πλήθος n των παρατηρήσεων.

### Συσταδοποίηση Διαμέρισης (Partitioning Clustering)

Η συσταδοποίηση διαμέρισης αποτελεί μία πολύ κοινή μέθοδο συσταδοποίησης όπου το πλήθος n δεδομένων διαμοιράζονται σε k διαφορετικές συστάδες. Σε αντίθεση με τους ιεραρχικούς αλγόριθμος συσταδοποίησης, στη συσταδοποίηση διαμέρισης είναι απαραίτητη προϋπόθεση ο καθορισμός των k συστάδων από το χρήστη. Σκοπός της συσταδοποίησης διαμέρισης είναι η ελαχιστοποίηση των αποστάσεων των δεδομένων μέσα στις ομάδες και η μεγιστοποίηση των αποστάσεων μεταξύ των ομάδων. Παράλληλα ο σχηματισμός των συστάδων επιτυγχάνεται χρησιμοποιώντας ένα κριτήριο διαχωρισμού. Ακολουθεί η παρουσίαση των αλγορίθμων που χρησιμοποιήθηκαν στην πειραματική διαδικασία.

**K-means**

Ο K-Means αποτελεί το πιο γνωστό αλγόριθμο συσταδοποίησης παρατηρήσεων σε προκαθορισμένο από το χρήστη αριθμό συστάδων. Αρχικά, καθορίζεται ο αριθμός των συστάδων και στη συνέχεια ακολουθεί η ανάθεση των αρχικών κεντρικών τους σημείων. Ενδεικτικοί τρόποι την ανάθεση τους είναι είτε με τυχαία επιλογή k παρατηρήσεων και ορισμό τους ως κεντρικά σημεία των συστάδων (forgy method) είτε διαμοιράζονται τυχαία τις παρατηρήσεις στις k συστάδες και θέτοντας τις μέσες τιμές τις κάθε ομάδας ως το κεντρικό σημείο (Random Partitioning Method) (MacQueen, 1967; Θεοδωρίδης & Πελέκης, χ.η.).

Ο παραπάνω αλγόριθμος λειτουργεί επαναληπτικά δηλαδή οι παρατηρήσεις ανατίθενται στις συστάδες, υπολογίζονται τα νέα κεντρικά σημεία, υπολογίζονται οι νέες αποστάσεις των σημείων από τις κεντρικές τιμές και ανακατανέμονται οι παρατηρήσεις. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να ικανοποιηθούν τα κριτήρια τερματισμού. Ως κριτήριο τερματισμού μπορεί να οριστεί είτε ένας πεπερασμένος αριθμός επαναλήψεων είτε όταν δεν παρατηρείται νέα ανακατανομή των παρατηρήσεων μεταξύ των ομάδων και ελαχιστοποιείται το κριτήριο του μέσου τετραγωνικού σφάλματος.

**k-medoids**

Ο αλγόριθμος k-medoids όπως και ο k-means είναι διαιρετικός αλγόριθμος δηλαδή διαμοιράζει τα δεδομένα σε ομάδες. Η διαφορά έγκειται στο γεγονός ότι επιλέγει ως αντιπροσωπευτικό σημείο της συστάδας το στοιχείο που απέχει τη μικρότερη απόσταση από τα υπόλοιπα. Το αντιπροσωπευτικό σημείο της συστάδας ονομάζεται medoid (Θεοδωρίδης & Πελέκης, χ.η.). Τα medoid είναι τα πιο κεντρικά σημεία μιας συστάδας ή καλύτερα ορισμένα τα σημεία στα οποία η ανομοιότητα μεταξύ των σημείων είναι η ελάχιστη (Kaufman & Rousseeuw, 2009).

**k-shape**

Ο αλγόριθμος k-shape αναπτύχθηκε από τους Paparrizos & Gravano (2015) και αποτελεί έναν αλγόριθμο συσταδοποίησης διαμέρισης πολύ όμοιο με τον K-means. Ο υπολογισμός του μέτρου απόστασης βασίζεται στην cross-correlation συσχέτιση μεταξύ των χρονοσειρών.

Συνοψίζοντας, θα συγκρίνουμε τις παραπάνω μεθόδους ιεραρχικής συσταδοποίησης και συσταδοποίησης διαμέρισης. Αρχικά, το βασικό χαρακτηριστικό των αλγορίθμων ιεραρχικής συσταδοποίησης είναι η απουσία ορισμού του αριθμού συστάδων πρίν την υλοποίηση τους. Κυρίως με τη χρήση του δενδρογράμματος έχουμε μία καλύτερη και

πληρέστερη εικόνα για τις πιθανές επιλογές του κατάλληλου αριθμού συστάδων. Αντίθετα οι αλγόριθμοι διαμέρισης βάση του παρακάτω πίνακα φαίνεται να υπερτερούν έναντι των ιεραρχικών. Ειδικότερα καθώς αυξάνεται το πλήθος των παρατηρήσεων, η πολυπλοκότητα στους ιεραρχικούς αυξάνει εκθετικά όπως φαίνεται στον πίνακα 3.

*Πίνακας 3. Αλγόριθμοι συσταδοποίησης-χρονική πολυπλοκότητα*

| Αλγόριθμος Συσταδοποίησης | Χρονική πολυπλοκότητα |
|---|---|
| Ιεραρχική Συσταδοποίηση | $O(N^2 \log N)$ |
| K-Means | O(INkn) |
| K-Medoids | $O(Ik(N-k)^2)$ |
| K-shape | $O(\max(nkm\log(m), nm^2, km^3))$ |

### 4.4.3 Επιλογή αριθμού k συστάδων

Όπως είδαμε στην προηγούμενη ενότητα, αρκετές μέθοδοι συσταδοποίησης λαμβάνουν ως όρισμα από τον χρήστη τον αριθμό των συστάδων που θα δημιουργήσουν στην πορεία. Γενικά το συγκεκριμένο χαρακτηριστικό αποτελεί γνώρισμα των μη ιεραρχικών αλγορίθμων. Ο αριθμός των συστάδων αποτελεί μία άγνωστη παράμετρο η οποία μπορεί να καθοριστεί με πολλές μεθόδους (Datasciencecentral, 2019). Για τις ανάγκες της παρούσας έρευνας χρησιμοποιήθηκαν δύο προσεγγίσεις, όπου η πρώτη αφορούσε τον υπολογισμό του δείκτη Silhouette και η δεύτερη τον δείκτη Calinski & Harabasz index.

- **Silhouette index:** Ο δείκτη Silhouette υπολογίζει τη μέση απόσταση ενός σημείου από τα υπόλοιπα στοιχεία της συστάδας που ανήκει και ταυτόχρονα εν συγκρίσει με τα στοιχεία των υπολοίπων συστάδων. Ο τύπος υπολογισμού της παραπάνω απόστασης είναι ο παρακάτω (Βρυώνης & Τσούτσας, 2011):

$$S = \frac{bi - ai}{\max(ai, bi)}$$

Όπου Ci, Cj ορίζονται ως οι συστάδες με $i \neq j$ και ai, bi ορίζονται ως η μέση απόσταση του σημείου i από τα στοιχεία της συστάδας Ci και Cj αντίστοιχα. Η κατανομή των τιμών του δείκτη ανήκουν στο διάστημα [-1,1].

  - -1: Οι αρνητικές τιμές δηλώνουν ότι το στοιχείο i απέχει περισσότερο από τα στοιχεία της συστάδας του συνεπώς υπάρχουν ενδείξεις ότι τα στοιχεία έχουν ανατεθεί λάθος.
  - 0: Η απόσταση μεταξύ των συστάδων δεν είναι σημαντική.

- 1: Οι θετικές τιμές δηλώνουν ότι το στοιχείο i απέχει κατά μέσο όρο λιγότερο από τα στοιχεία εντός της συστάδας του, άρα έχουν ανατεθεί σωστά.

- **Calinski-Harabasz:** Αποτελεί ένα μέτρο το οποίο στην ουσία είναι μια αναλογία για αυτό και είναι γνωστό ως το κριτήριο αναλογίας διακύμανσης. Βασίζεται στην διασπορά εντός της συστάδας καθώς και στη διασποράς μεταξύ των συστάδων (Wei, 2020).

  ΟΙ υψηλές τιμές του συγκεκριμένου δείκτη αποτελούν ένδειξη καλού διαχωρισμού μεταξύ των συστάδων. Τέλος, ορίζεται από τον παρακάτω τύπο (Ντουντούμι, 2020).

$$CH = \frac{\frac{SSB}{k-1}}{\frac{SSE}{k}}$$

# Κεφάλαιο 5

## 5.1 Πειραματική μελέτη

Στην προηγούμενη ενότητα αναφέρθηκαν οι διαφορετικοί τύποι συσταδοποίησης, οι βασικές τους κατηγορίες, ορισμένα μέτρα ομοιότητας και απόστασης όπως επίσης και τεχνικές για την εύρεση του βέλτιστου αριθμού συστάδων καθώς και ορισμένους αλγόριθμους συσταδοποίησης. Σε αυτό το κεφάλαιο θα παρουσιάσουμε τα αποτελέσματα που προέκυψαν από την πειραματική διαδικασία. Αρχικά στην ενότητα 5.2 περιγράφεται το σύνολο των δεδομένων που χρησιμοποιήθηκε και γίνεται αναφορά στις εξεταζόμενες μεταβλητές. Έπειτα ακολουθεί η προ-επεξεργασία του συνόλου των δεδομένων. Στην ενότητα 5.3 γίνεται αναφορά στην επιλογή των μέτρων απόστασης όπως επίσης και στον καθορισμό του αριθμού των συστάδων. Στη συνέχεια γίνεται εφαρμογή των αλγορίθμων συσταδοποίησης. Τέλος, στην ενότητα 5.4 παρουσιάζονται τα αποτελέσματα της πειραματικής διαδικασίας.

Η εισαγωγή των δεδομένων, η προ-επεξεργασία τους καθώς και η ανάπτυξη των αλγορίθμων μηχανικής μάθησης έγινε με τη χρήση της γλώσσας προγραμματισμού Python έκδοσης 3.7.9. Η ανάπτυξη και η εκτέλεση τους πραγματοποιήθηκε σε τοπικό υπολογιστή με τα ακόλουθα χαρακτηριστικά πίνακας 4:

*Πίνακας 4. Χαρακτηριστικά υπολογιστή*

| CPU | RAM | ROM |
|---|---|---|
| I7 9th Gen 4c/8th | 16gb | 512gb ssd |

## 5.2 Περιγραφή συνόλου δεδομένων

Όπως αναφέρθηκε στην ενότητα 4.3 τα δεδομένα περιέχουν μετρήσεις από το CAN BUS σύστημα ενός οχήματος. Προήλθαν κατά την διαδικασία της οδήγησης ενός οχήματος ιδίων χαρακτηριστικών από 3 διαφορετικούς οδηγούς με 2 διαφορετικές διαδρομές ανά οδηγό. Τα δεδομένα ήταν σε μορφή hdf αρχείου και πραγματοποιήθηκε η εισαγωγή τους σε δομές δεδομένων της μορφής data frame. Ειδικότερα αποφασίστηκε η διατήρηση των μεταβλητών της επιτάχυνσης και της ταχύτητας στον άξονα X όπου διαχωρίστηκαν ανά οδηγό και ανά διαδρομή. Το τελικό subset, αποτελούμενο από 5 διαφορετικά data-frames αποτυπώνεται στον πίνακα 5.

*Πίνακας 5. Διαστάσεις πινάκων δεδομένων*

| Driver | Trip | Dimensions of DF | Trip duration hours:minutes:seconds |
|---|---|---|---|
| **A** | 1 | (158659,2) | 02:12:12 |
| **A** | 2 | (54835,2) | 00:45:41 |
| **B** | 1 | (41817,2) | 00:34:50 |
| **B** | 2 | (136154,2) | 01:53:27 |
| **C** | 1 | (36216,2) | 00:30:10 |
| **C** | 2 | (74032,2) | 01:01:41 |

Το πρώτο βήμα πριν προχωρήσουμε στην ανάλυση ήταν να διασφαλιστεί ότι οι χρονοσειρές έχουν κοινό αρχικό σημείο αναφοράς. Εξετάζοντας τα σύνολα των δεδομένων την χρονική στιγμή t = 0 διακρίνεται ότι όλες οι χρονοσειρές έχουν κοινό σημείο αναφοράς t=0, U=0. Όπως αναφέρθηκε και στον κεφάλαιο 4, απαραίτητη προϋπόθεση για τη χρήση Lock step measures είναι η χρήση χρονοσειρών ίσου μήκους αποτυπώνοντας τη χρονική διάρκεια των δεδομένων σε επίπεδο οδηγού και διαδρομή. Διαπιστώθηκε ότι η μικρότερη διάρκεια καταγραφής είναι 30 λεπτά και η μεγαλύτερη είναι 2 ώρες και 12 λεπτά, με σταθερό ρυθμό καταγραφής στα 0.05 seconds. Δεδομένου ότι μεταβλητή του χρόνου είναι κοινή σε όλες τις χρονοσειρές με ταυτόσημες τιμές 1:1., η στήλη αυτή χρησιμοποιήθηκε ως κλειδί για την συγχώνευση των πινάκων. Για τον καλύτερο χειρισμό των δεδομένων δημιουργήθηκε ένας νέος πίνακας διαστάσεων (36216 , 6) εφαρμόζοντας inner join στους 6 διακριτούς πίνακες με primary key τη στήλη New Timestamp. Παράλληλα, ορίστηκε ως Index τύπου datetime και εφαρμόστηκε resampling στον πίνακα στα 10 sec με τη μέθοδο της μέσης τιμής. Ο τελικός πίνακας όπως επίσης και η γραμμογράφηση των μεταβλητών αποτυπώνονται στους παρακάτω πίνακες (Εικόνα 12, Πίνακας 6):

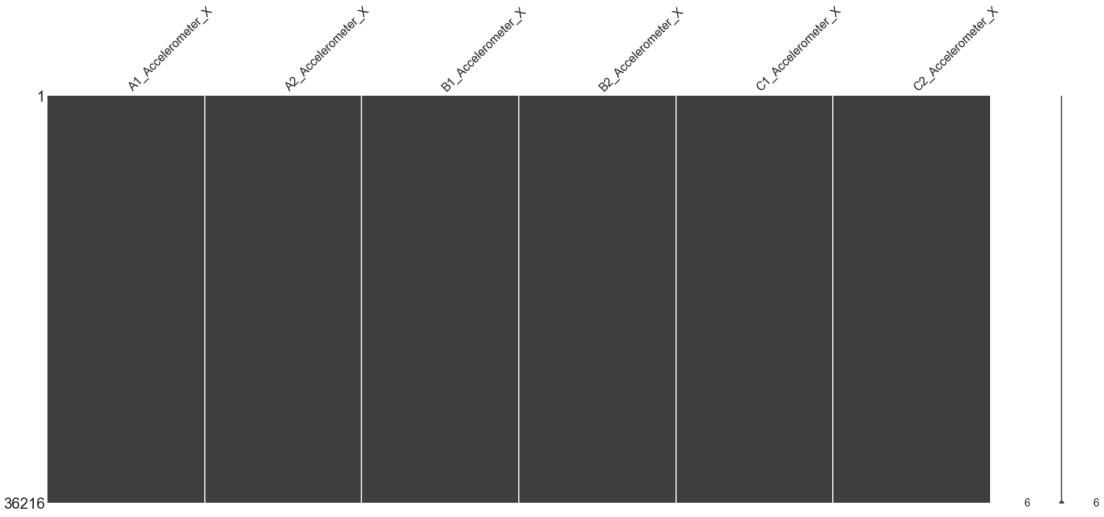
| New Timestamp | A1_Accelerometer_X | A2_Accelerometer_X | B1_Accelerometer_X | B2_Accelerometer_X | C1_Accelerometer_X | C2_Accelerometer_X |
|---|---|---|---|---|---|---|
| **2019-04-01 00:00:00** | 0.117096 | -0.266846 | -0.171145 | 0.118700 | -0.146133 | -0.453960 |
| **2019-04-01 00:00:10** | 0.562144 | 0.611638 | -0.092775 | 1.109128 | -0.148238 | -0.456119 |
| **2019-04-01 00:00:20** | -0.066634 | -0.283170 | 0.924883 | 0.378887 | -0.146906 | -0.454541 |
| **2019-04-01 00:00:30** | -0.339243 | 0.273816 | -0.885970 | 0.051724 | 0.112342 | -0.453110 |
| **2019-04-01 00:00:40** | 0.320866 | 0.677633 | 0.521110 | 0.109337 | -0.265295 | -0.519580 |

*Εικόνα 12. Πίνακας τιμών συνόλου δεδομένων (Προ επεξεργασίας)*

*Πίνακας 6. Γραμμογράφηση μεταβλητών*

| Όνομα μεταβλητής | Περιγραφή |
|---|---|
| **A1_Accelerometer_X** | Οδηγός Α διαδρομή 1 |
| **A2_Accelerometer_X** | Οδηγός Α διαδρομή 2 |
| **B1_Accelerometer_X** | Οδηγός Β διαδρομή 1 |
| **B2_Accelerometer_X** | Οδηγός Β διαδρομή 2 |
| **C1_Accelerometer_X** | Οδηγός C διαδρομή 1 |
| **C2_Accelerometer_X** | Οδηγός C διαδρομή 2 |

Το επόμενο βήμα ήταν ο έλεγχος για ακραίες τιμές. Τα μέτρα ομοιότητας και απόστασης είναι ευαίσθητα στην ύπαρξη θορύβου, Null values και ακραίων τιμών συνεπώς κρίνεται αναγκαία η διερεύνηση τους. Η βιβλιοθήκη missingno προσφέρει ένα αποδοτικό και γρήγορο τρόπο γραφικής απεικόνισης των θέσεων των missing values. Όπως φαίνεται και στο παρακάτω ραβδόγραμμα (εικόνα 13), κάθε στήλη αναπαριστά τις στήλες του data frame. Αριστερά κάτω αποτυπώνεται το πλήθος των γραμμών του data frame και σε κάθε ράβδο εμφανίζονται με λευκές γραμμές οι κενές τιμές. Με τη χρήση της παραπάνω βιβλιοθήκης εξάγουμε την παρακάτω εικόνα (Εικόνα 13) για το σύνολο των δεδομένων και παρατηρείται ότι δεν περιέχει κενές τιμές.

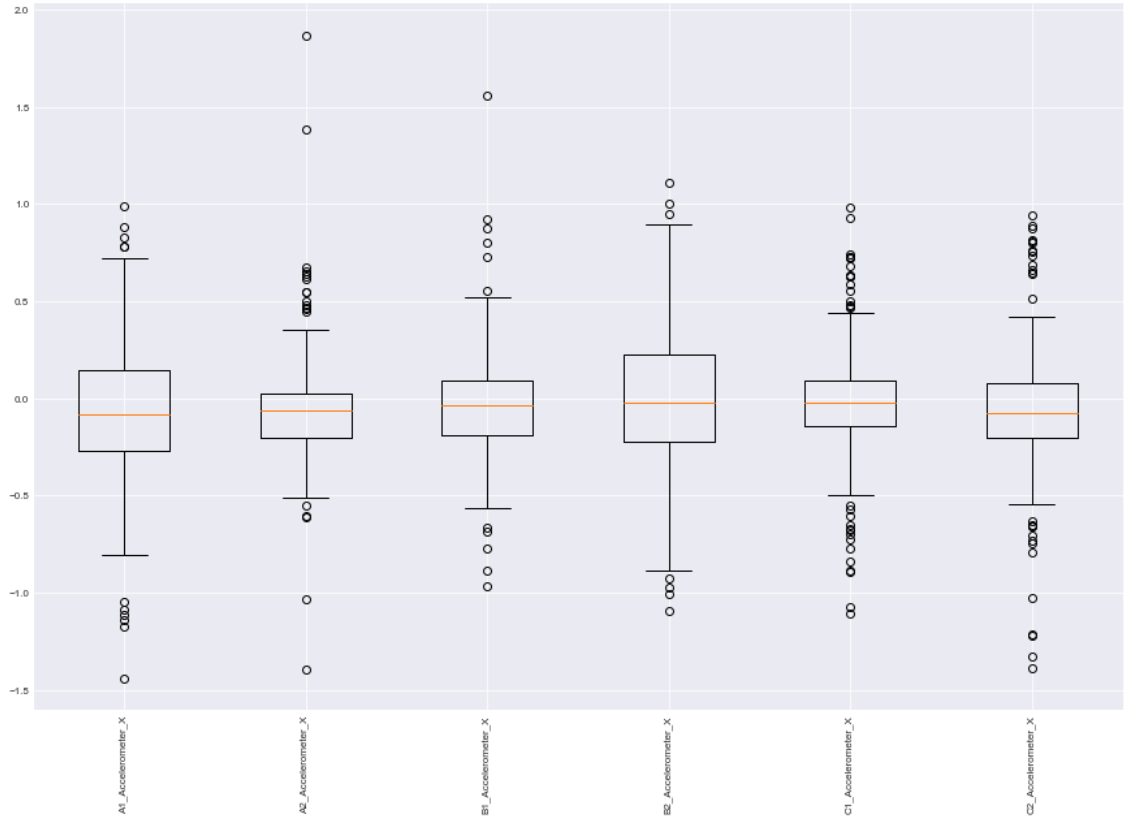



*Εικόνα 13. Missing values*

Παράλληλα με τη χρήση της βιβλιοθήκης pandas και ειδικότερα της συνάρτησης describe απεικονίζονται στην παρακάτω εικόνα ορισμένα βασικά μέτρα θέσης και διασποράς (Εικόνα 14).

| | A1_Accelerometer_X | A2_Accelerometer_X | B1_Accelerometer_X | B2_Accelerometer_X | C1_Accelerometer_X | C2_Accelerometer_X |
|---|---|---|---|---|---|---|
| **count** | 182.000000 | 182.000000 | 182.000000 | 182.000000 | 182.000000 | 182.000000 |
| **mean** | -0.074766 | -0.046081 | -0.039657 | 0.011169 | -0.013810 | -0.061909 |
| **std** | 0.378749 | 0.318253 | 0.308424 | 0.400286 | 0.338395 | 0.388338 |
| **min** | -1.439256 | -1.395435 | -0.968873 | -1.093977 | -1.103209 | -1.384425 |
| **25%** | -0.268630 | -0.199916 | -0.191938 | -0.219991 | -0.144205 | -0.200729 |
| **50%** | -0.083703 | -0.064399 | -0.037287 | -0.022605 | -0.019878 | -0.075089 |
| **75%** | 0.146552 | 0.022820 | 0.094873 | 0.227543 | 0.094450 | 0.078962 |
| **max** | 0.986929 | 1.869074 | 1.559312 | 1.109128 | 0.981166 | 0.943482 |

*Εικόνα 14. Πινακας τιμών συνόλου δεδομένων (Μετά την επεξεργασία)*




*Εικόνα 15. Boxplots*

Παρατηρώντας τα παραπάνω γραφήματα (Εικόνα 15) διαπιστώθηκε η ύπαρξη ακραίων τιμών με μεγάλη διακύμανση. Προκειμένου να εξαλειφθεί ο θόρυβος, να μειωθούν τα σημεία με μεγάλες διακυμάνσεις και να ομαλοποιηθούν οι χρονοσειρές επιλέχθηκε η τεχνική του κινητού μέσου. Ο κινητός μέσος προκύπτει ως ο μέσος των k προηγούμενων παρατηρήσεων. Εφαρμόστηκε πειραματικά για διάφορες τιμές του K ο κινητός μέσος και στη συνέχεια ακολούθησε διαγραμματική απεικόνιση των χρονοσειρών. Επιλέχτηκε το K=5 ώστε

να μην αφαιρεθεί εντελώς η μεταβλητότητα αλλά ταυτόχρονα να μην εξομαλυνθεί τελείως η τάση. Στόχος είναι να αποτυπώνεται η κάθε προσωποποιημένη συμπεριφορά στα δεδομένα.

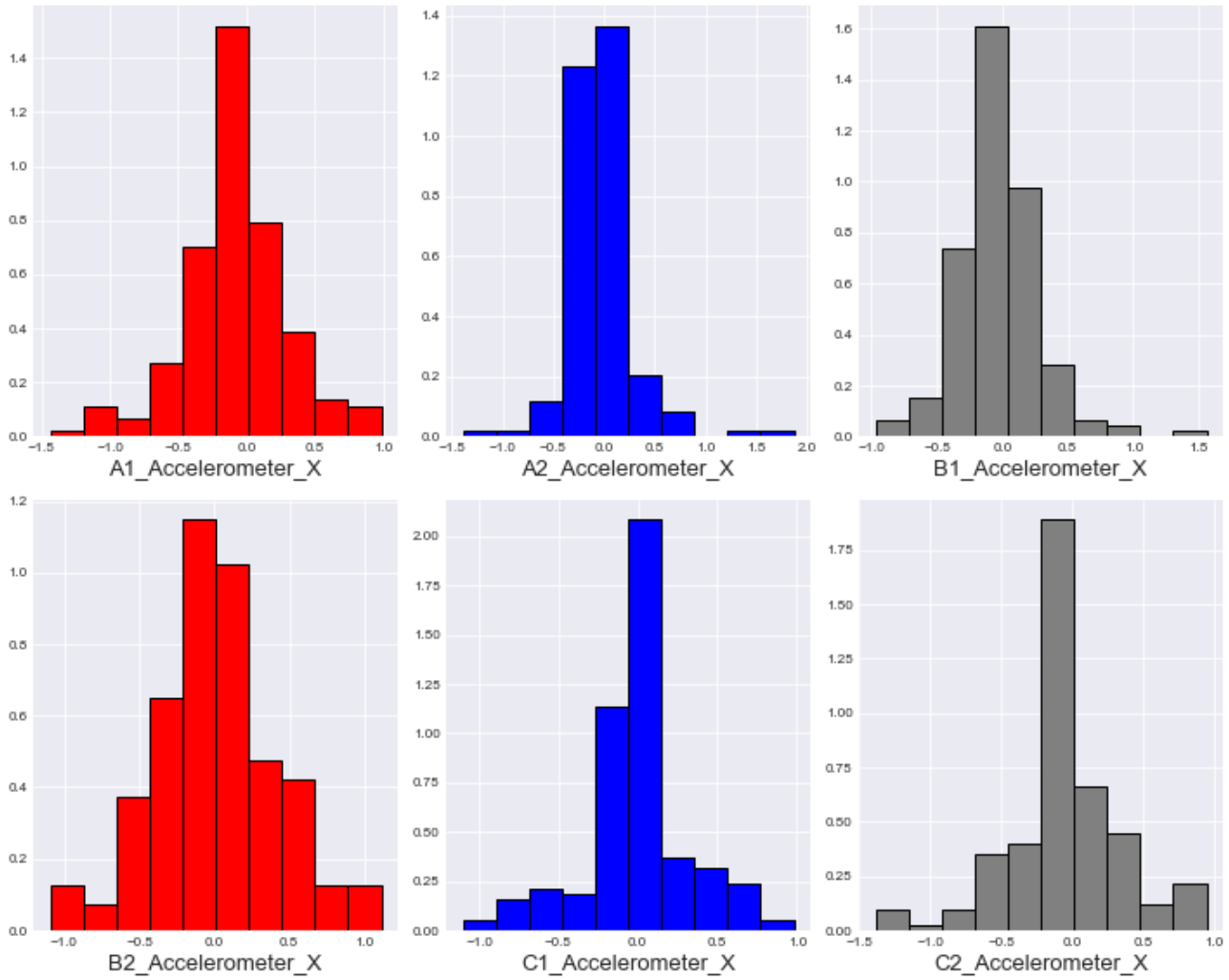
Το επόμενο βήμα στην ανάλυση των χρονοσειρών είναι η απεικόνιση των μετρήσεων σε ένα διάγραμμα χρονοσειράς. Μέσω της γραφικής απεικόνισης των δεδομένων τα χαρακτηριστικά της χρονοσειράς θα εμφανιστούν ως μοτίβα. Ως μέθοδος απεικόνισης επιλέχθηκαν τα γραφήματα πολυγωνικών γραμμών (Εικόνα 16). Η πολλαπλή πολυγωνική γραμμή αποτελεί την πιο ευρέως διαδεδομένη γραφική απεικόνιση καθώς προσφέρει τη δυνατότητα αναπαράστασης της πορείας που έχουν καταγράψει οι εξεταζόμενες μεταβλητές συναρτήσει του χρόνου. Χρησιμοποιώντας τη βιβλιοθήκη της matplotlib διακρίνεται γραφικά η ύπαρξη οποιασδήποτε τάσης (ανοδική ή καθοδική) καθώς και η περιοδική διακύμανση όπως εποχικότητα ή κυκλικότητα. Όπως φαίνεται στην εικόνα (εικόνα 16) παρουσιάζεται γραφικά η εξέλιξη των μεταβλητών.


*Εικόνα 16. Γραφήματα πολυγωνικών γραμμών*

Η γραφική απεικόνιση των διαγραμμάτων κατανομής των μεταβλητών συντελούν στον εντοπισμό της κατάλληλης κατανομής που μπορεί να περιγράψει τα δεδομένα. Η λήψη ενός δείγματος από έναν πληθυσμό θεωρείται επιτυχής όταν προκύπτουν όσο το δυνατόν

ακριβέστερα αποτελέσματα. Δηλαδή, να είναι γενικεύσιμα και να βρίσκονται πιο κοντά στα αντίστοιχα αποτελέσματα που δημιουργούνται από το σύνολο του πληθυσμού.

Συνεπώς σε αυτό το βήμα θα εξετασθεί η υπόθεση την κανονικής κατανομή των μεταβλητών. Ο έλεγχος μπορεί να πραγματοποιηθεί με δύο μεθόδους, είτε διαγραμματικά είτε μέσω Normality tests. Με τη χρήση της βιβλιοθήκης matplotlib παράγεται μια γραφική παράσταση της κατανομής των δεδομένων ώστε μέσα από τα διαγράμματα να εξεταστεί εάν οι τιμές των μεταβλητών τείνουν να συγκεντρώνονται γύρω από τη μέση τιμή. Εάν δηλαδή η γραφική παράσταση έχει σχήμα 'καμπάνας' τότε η μεταβλητή ακολουθεί την κανονική κατανομή. Παρακάτω απεικονίζονται αναλυτικά οι γραφικές παραστάσεις των μεταβλητών (Εικόνα 17):




*Εικόνα 17. Γραφικές παραστάσεις των μεταβλητών*

Μέσω της διαγραμματικής απεικόνισης παρατηρούνται ενδείξεις αλλά χωρίς να καταλήγουν σε αξιόπιστο συμπέρασμα. Για αυτό και ακολουθεί το Jarque-Bera Normality Test όπου εξετάζεται κατά πόσο το δείγμα παρουσιάζει την ασυμμετρία και την κύρτωση ενός δείγματος που θα ακολουθούσε την Γκαουσιανή (κανονική) κατανομή.

Ορίζεται η υπόθεση:

H0: Το δείγμα ακολουθεί την κανονική κατανομή

H1: Το δείγμα δεν ακολουθεί την κανονική κατανομή

*Πίνακας 7. Τεστ κανονικότητας(Jarque-Bera Normality Test) p-values*

| Variables | P-values | Reject Ho | Fail to reject Ho |
|---|---|---|---|
| A1_Accelerometer_X | p=0.000 | ✓ | |
| A2_Accelerometer_X | p=0.000 | ✓ | |
| B1_Accelerometer_X | p=0.000 | ✓ | |
| B2_Accelerometer_X | p= 0.644 | | ✓ |
| C1_Accelerometer_X | p=0.000 | ✓ | |
| C2_Accelerometer_X | p=0.000 | ✓ | |

Σε επίπεδό σημαντικότητας 5% παρατηρείται ότι τα p-values σε όλες τις μεταβλητές είναι μικρότερα του 0.05 με ακρίβεια 3 δεκαδικών ψηφίων εκτός της B2_Accelerometer_X. Οι μεταβλητές A1_Accelerometer_X, A2_Accelerometer_X, B1_Accelerometer_X, C1_Accelerometer_X, C2_Accelerometer_X σε επίπεδο σημαντικότητας 0.05 παρουσιάζονται ισχυρές ενδείξεις απόρριψης της μηδενικής υπόθεσης, συνεπώς ακολουθούν την κανονική κατανομή. Αντίθετα η B2_Accelerometer_X δεν έχει επαρκή στοιχεία για απόρριψη της Ho υπόθεσης επομένως ακολουθεί την κανονική κατανομή (πίνακας 7).

Τέλος, στα πλαίσια της έρευνας δεν εφαρμόστηκαν τεχνικές ανάλυσης χρονοσειρών για την εξάλειψη τυχόν τάσης και περιοδικότητας. Εφαρμόζοντας τεχνικές shaped-based συσταδοποίησης πρέπει να συμπεριληφθούν αυτά τα χαρακτηριστικά των χρονοσειρών στην μεταξύ τους ομαδοποίηση. Επιπρόσθετα, ενώ το εύρος των τιμών των εξεταζόμενων μεταβλητών επιτάχυνσης δεν διαφέρει σημαντικά εντούτοις πραγματοποιήθηκε κανονικοποίηση στις τιμές των δεδομένων χρησιμοποιώντας τη βιβλιοθήκη sklearn.

## 5.3 Εκτίμηση n αριθμού συστάδων

Σε αυτήν την υποενότητα θα εφαρμοστούν δύο μέθοδοι για την επιλογή του κατάλληλου αριθμού των συστάδων. Οι μέθοδοι που θα εφαρμοστούν είναι ο υπολογισμός των δεικτών Silhouette και Calinski-Harabasz οι οποίοι αναφέρονται στο κεφάλαιο 4. Ο υπολογισμός των δύο δεικτών έγινε εφαρμόζοντας τεχνικές συσταδοποίησης σε συνδυασμό με τρία διαφορετικά μέτρα απόστασης. Ο υπολογισμός των παραπάνω δεικτών υλοποιήθηκε μέσω της βιβλιοθήκη sklearn και ειδικότερα του πακέτου metrics. Τέλος, ο πίνακας 8 περιέχει τα αποτελέσματα της εκτίμησης τους αριθμού των συστάδων με τις παραπάνω μεθόδους.

*Πίνακας 8. Εύρεση αριθμού συστάδων*

| Clustering algorithm | Index | Μετρική Απόστασης | N Clusters for best score |
|---|---|---|---|
| TS K-Means | Sillouette | Ευκλείδεια | 2 |
| | | DTW | 2 |
| | | Soft-DTW | 3 |
| | Calinski-Harabasz | Ευκλείδεια | 2 |
| | | | |
| K-Medoid | Sillouette | Ευκλείδεια | 2 |
| | | DTW | 2 |
| | | Pearson correlation | 2 |
| | Calinski-Harabasz | Ευκλείδεια | 2 |
| | | | |

Στο πίνακα 7 αναγράφονται όλοι οι συνδυασμούς που υλοποιήθηκαν για τον υπολογισμό των δεικτών. Το δείγμα αποτελείται από τρείς οδηγούς και δύο διαδρομές ανά οδηγό συνεπώς ορίστηκε ως εύρος k συστάδων από 2 μέχρι 5. Οι αλγόριθμοι συσταδοποίησης που επιλέχθηκαν ήταν ο K-Means και ο K-Medoid κάνοντας χρήση της βιβλιοθήκης sklearn. Ειδικότερα ο τρόπος υπολογισμού του δείκτη Calinski-Harabasz περιορίζεται μόνο για την Ευκλείδεια απόσταση. Καταλήγοντας, παρατηρείται ότι η επικρατούσα τιμή των αποτελεσμάτων είναι το k=2. Ταυτόχρονα πραγματοποιήθηκε και γραφική απεικόνιση των k συστάδων για επιβεβαίωση των αποτελεσμάτων αλλά λόγο του μικρού εξεταζόμενου δείγματος δεν κατέληξε σε κάποιο αποτέλεσμα.

## 5.4 Αποτελέσματα αλγορίθμων συσταδοποίησης

Στην τελευταία υποενότητα του κεφαλαίου 5 θα παρουσιαστούν οι αλγόριθμοι συσταδοποίησης και τα μέτρα θέσης που υλοποιήθηκαν και θα ακολουθήσει η απεικόνιση των

αποτελεσμάτων. Ο αριθμός των συστάδων που επιλέχθηκε βάση των ελέγχων της ενότητας 5.3 ισούται με δύο. Τα βήματα της διαδικασίας συνοψίζονται στα εξής:

1. Επιλογή μέτρου ομοιότητας
2. Επιλογή του αριθμού των συστάδων
3. Συσταδοποίηση των δεδομένων
4. Υπολογισμός της μέσης τιμής των επιταχύνσεων των χρονοσειρών ανά συστάδα
5. Χαρακτηρισμός των συστάδων με κριτήριο τη μέση τιμή τους
6. Απεικόνιση των χρονοσειρών ανά συστάδα

Τα βήματα 1, 2, 3 αναλύθηκαν σε προηγούμενες υποενότητες των κεφαλαίων 4, 5. Οι συνδυασμοί των αλγορίθμων και των μέτρων απόστασης αποτυπώνονται στον παρακάτω πίνακα 9 . Αρχικά πραγματοποιείται η διαμέριση των 6 χρονοσειρών σε συστάδες. Κατόπιν υπολογίζεται η μέση τιμή των επιταχύνσεων ανά συστάδα. Στη συστάδα με τη μικρότερη μέση τιμή τα στοιχεία της χαρακτηρίζονται ως ήπια-φυσιολογική οδήγηση και στη συστάδα με τη μεγαλύτερη μέση τιμή τα στοιχεία της χαρακτηρίζονται ως επιθετική οδήγηση. Ο χαρακτηρισμός της συστάδας αποτυπώνεται στον τίτλο του γραφήματος και τα στοιχεία της κάθε συστάδας αποτυπώνονται εντός του γραφήματος σύμφωνα με την παραπάνω γραμμογράφηση (Πίνακας 6).

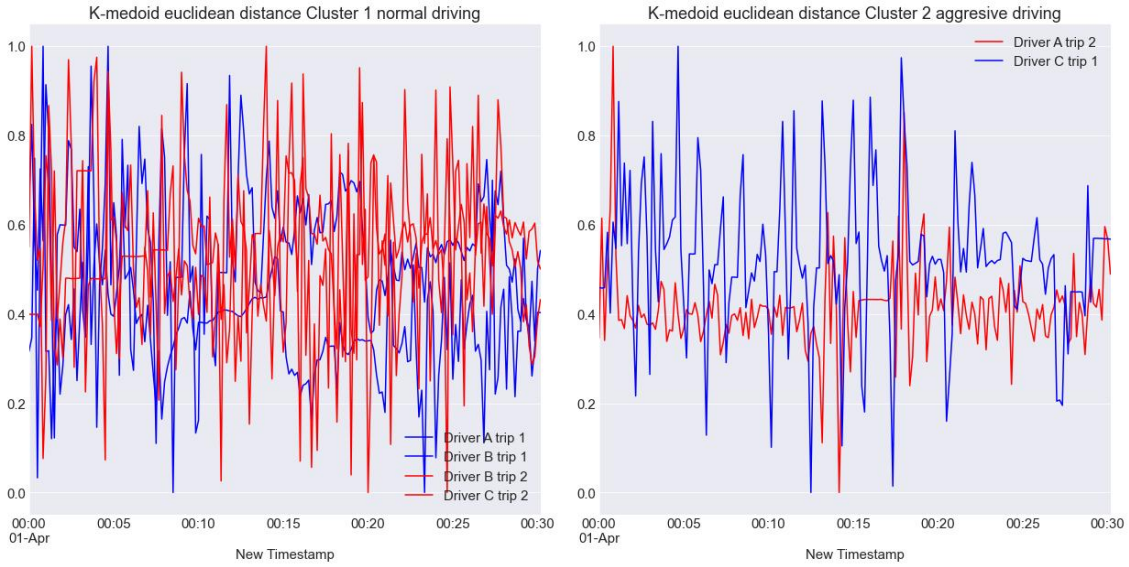
*Πίνακας 9. Μέτρα απόστασης*

| Μέτρα Απόστασης | TSK-means | K-shape | K-Medoids |
|---|---|---|---|
| **Ευκλείδεια** | ✔ | | ✔ |
| **Manhattan** | | | ✔ |
| **DTW** | ✔ | | ✔ |
| **Custom** | | ✔ | |

**K-Medoids**: Επιλέχθηκε η βιβλιοθήκη sklearn για την υλοποίηση της συσταδοποίησης και η βιβλιοθήκη tslearn.metrics για τον υπολογισμό των αντίστοιχων μέτρων απόστασης. Υλοποιήθηκε για k=2 clusters, με πλήθος επαναλήψεων ίσο με 10, με μέτρα απόστασης την

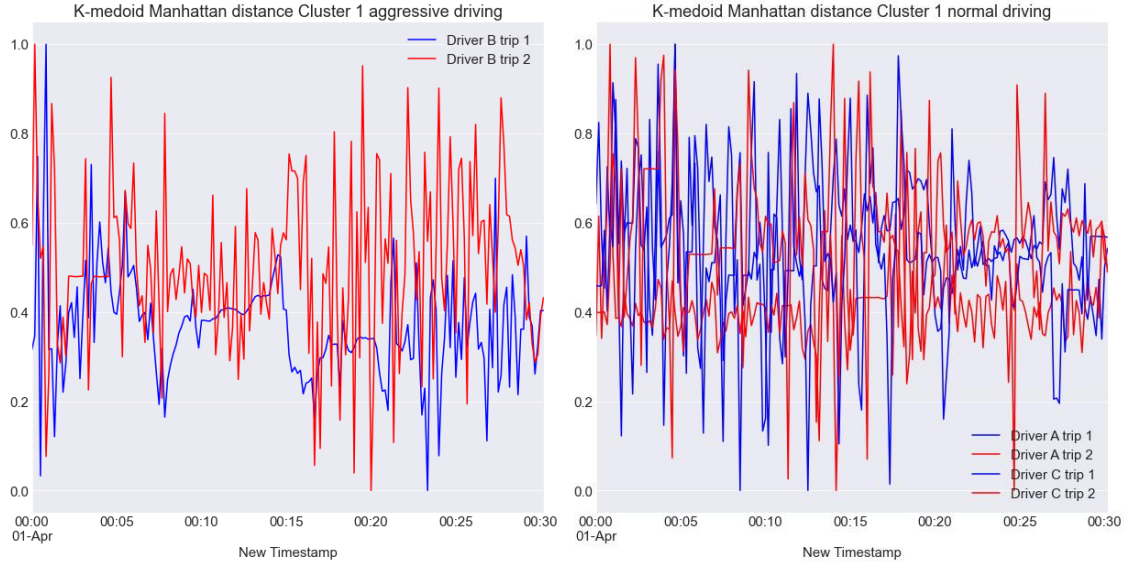
Ευκλείδεια, την DTW και τη Manhattan καθώς και τυχαία επιλογή αρχικών θέσεων medoids. Τα αποτελέσματα αναπαρίστανται γραφικά στις παρακάτω εικόνες (Εικόνα 18-20).

- K-Medoids συσταδοποίηση με χρήση της Ευκλείδειας απόσταση



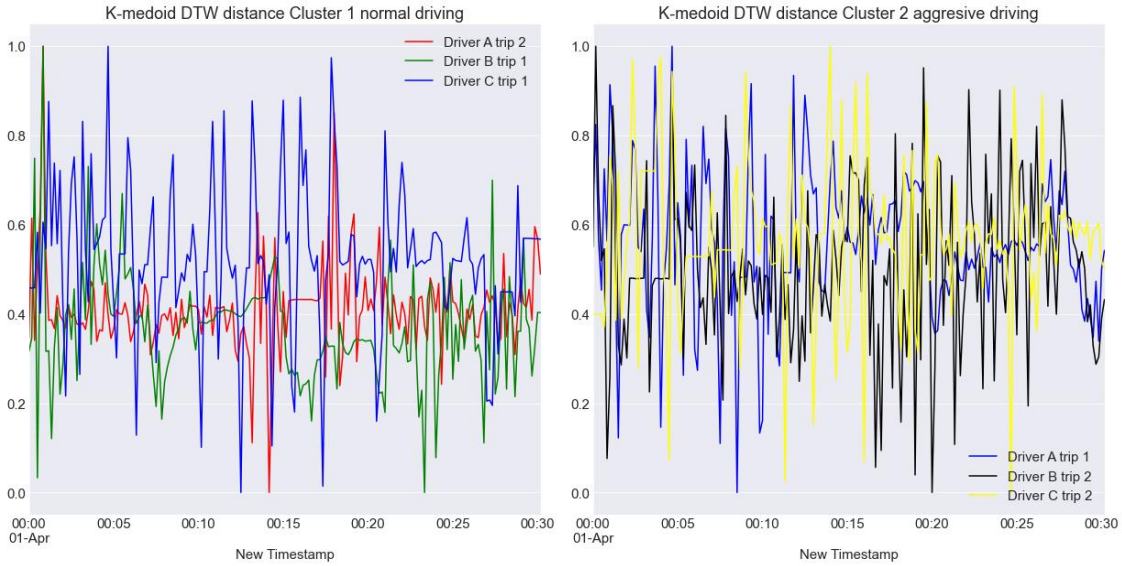

*Εικόνα 18. K-Medoids clustering (Euclidean)*

- K-Medoids συσταδοποίηση με χρήση της Manhattan απόσταση




*Εικόνα 19. K-medoids clustering (Manhattan)*

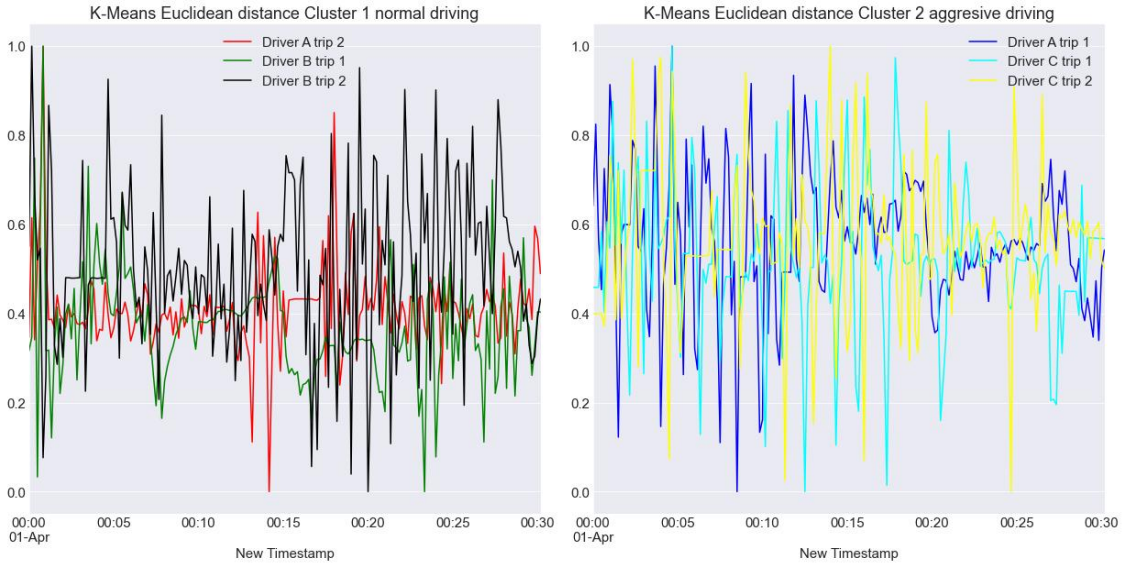
- K-Medoids συσταδοποίηση με τη χρήση της DTW απόστασης




*Εικόνα 20. K-medoids clustering (DTW)*

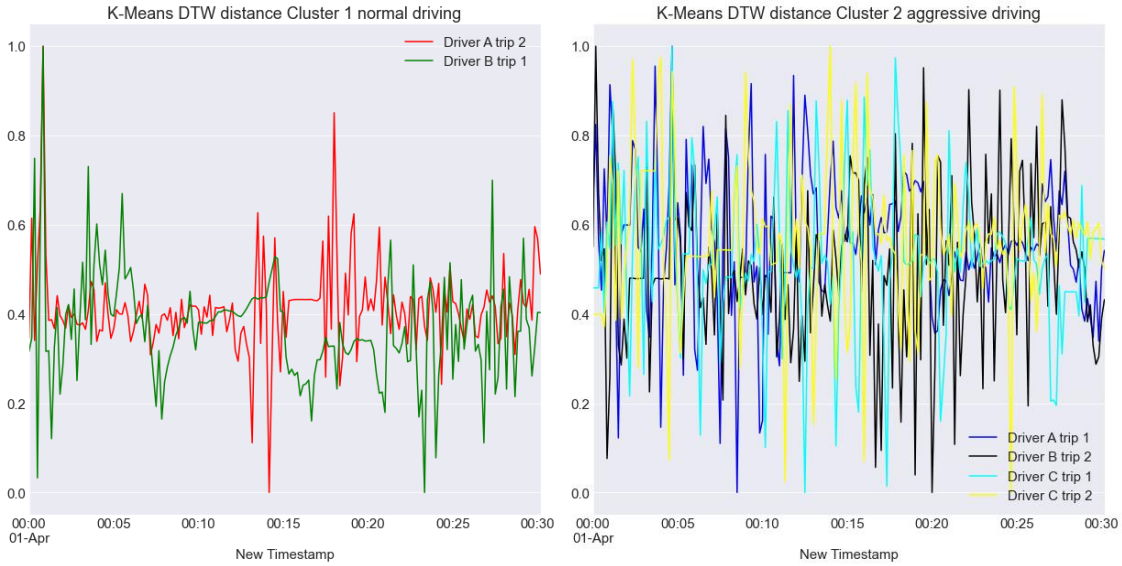
**K-Means**: Επιλέχθηκε η βιβλιοθήκη sklearn για την υλοποίηση της συσταδοποίησης και η βιβλιοθήκη tslearn.metrics για τον υπολογισμό των αντίστοιχων μέτρων απόστασης. Υλοποιήθηκε για k=2 clusters, με πλήθος επαναλήψεων ίσο με 10, για την ευκλείδεια και την DTW απόσταση καθώς και τυχαία επιλογή αρχικών θέσεων centroids. Τα αποτελέσματα αναπαρίστανται γραφικά στις παρακάτω εικόνες (Εικόνα 21-22).

- K-Means συσταδοποίηση με τη χρήση της Ευκλείδειας απόστασης



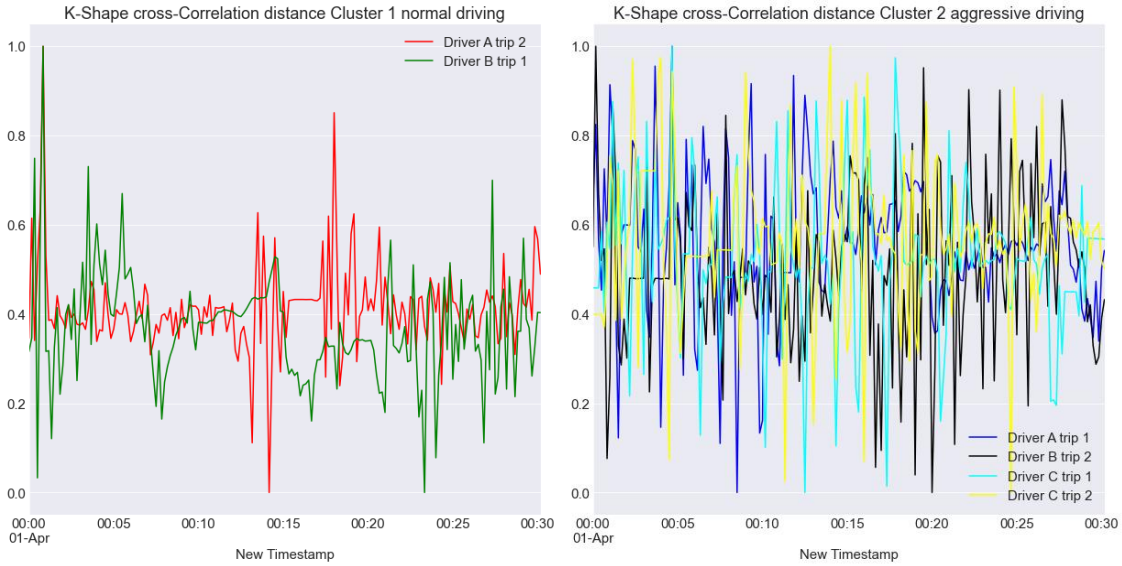

*Εικόνα 21. K-means clustering (Euclidean)*

- K-means συσταδοποίηση με τη χρήση της DTW απόστασης




*Εικόνα 22. K-means clustering (DTW)*

**K-Shape**: Για την υλοποίηση του αλγόριθμου K-shape χρησιμοποιήθηκε η βιβλιοθήκη sklearn. Υλοποιήθηκε για k=2 clusters χρησιμοποιώντας την built in απόσταση ομοιότητας normalized cross correlation. Τα αποτελέσματα αναπαρίστανται γραφικά στις παρακάτω εικόνες (Εικόνα 23).




*Εικόνα 23. K-shape clustering (normalized cross-Correlation)*

# Κεφάλαιο 6

## 6.1 Συμπεράσματα

Στη συγκεκριμένη εργασία εξετάστηκε το θέμα της αναγνώρισης της οδηγικής συμπεριφοράς μέσω της συσταδοποίησης δεδομένων διαφορετικών οδηγών σε διαφορετικές διαδρομές. Για να καταστεί εφικτή η συγκεκριμένη έρευνα, αρχικά αναφέρθηκαν και αναλύθηκαν οι κυριότερα προσεγγίσεις στην συσταδοποίηση χρονολογικών δεδομένων (shape-based, model-based, feaute-based). Επίσης παρουσιάστηκαν ορισμένα βασικά μέτρα ομοιότητας (shape-based, edit-based, feauter based, structure-based) και εφαρμόστηκαν οι αλγόριθμοι συσταδοποίησης k-means, k-medoids και k-shape. Τέλος, υπολογίστηκαν και συγκρίθηκαν οι μέσες τιμές των συστάδων για μια αρχική εκτίμηση της οδηγικής συμπεριφοράς των οδηγών και των διαδρομών που ακολούθησαν.

Ειδικότερα για τα παραπάνω αποτελέσματα παρατηρείται ότι στο 50% των περιπτώσεων και συγκεκριμένα με τη χρήση της elastic step measure απόστασης DTW και του cross-correlation μέτρου ομοιότητας οι οδηγοί Α, Β στις διαδρομές 2, 1 αντίστοιχα, κατηγοριοποιήθηκαν ως οδηγοί-διαδρομές με φυσιολογική οδήγηση. Επιπλέον κατά την γραφική απεικόνιση των χρονοσειρών ανά cluster παρατηρείται ότι οι χρονοσειρές παρουσιάζουν γραφική ομοιότητα.

Αντίστοιχα στις περιπτώσεις χρήσης των lock step measure Ευκλείδειας και Manhattan δεν φαίνεται να παρατηρείται το ίδιο μοτίβο ενώ τα αποτελέσματα παρουσιάζουν ασάφεια. Συνεπώς επιβεβαιώνεται η υπεροχή στη χρήση των elastic step measures στη συσταδοποίηση ασυγχρόνιστων χρονοσειρών, γεγονός το οποίο αναφερόταν στις βιβλιογραφικές παραπομπές.

Καταλήγοντας, η συσταδοποίηση χρονοσειρών μπορεί να αποτελέσει εξαιρετικά πολύπλοκο ζήτημα ειδικά όσο αυξάνονται οι διαστάσεις των δεδομένων. Στη παρούσα εργασία το πλήθος και το μήκος των χρονοσειρών δεν αποτέλεσε πρόβλημα. Σε μελλοντική όμως εργασία κρίνεται απαραίτητη η επιλογή μεγαλύτερου συνόλου δεδομένων για εξαγωγή ακριβέστερων αποτελεσμάτων. Τέλος, μία ενδιαφέρουσα προσθήκη θα ήταν η διερεύνηση διαφορετικών προσεγγίσεων μείωσης των διαστάσεων των χρονοσειρών και εφαρμογή των αντίστοιχων μεθόδων συσταδοποίησης.

# Βιβλιογραφία

## Ελληνόγλωσση

Βρυώνης, Ε., & Τσούτσας, Δ. (2011). Γραφο-θεωρητικές μέθοδοι συσταδοποίησης και ο αλγόριθμος Jarvis-Patrick σε βάσεις δεδομένων Oracle. Ανακτήθηκε 2 Φεβρουαρίου, 2021, από: http://apothesis.teicm.gr/xmlui/bitstream/handle/123456789/830/vrionis.pdf?sequence=1&isAllowed=y)

Θεοδωρίδης Γ. & Πελέκης Ν., (χ.η.). Εξόρυξη Γνώσης από Δεδομένα (Data Mining). Συσταδοποίηση (clustering). Πανεπιστήμιο Πειραιώς Τμήμα Πληροφορικής. Ανακτήθηκε 8 Ιανουαρίου, 2021, από: http://infolab.cs.unipi.gr/pre-eclass/courses/dwdm/ slides/5-clustering.pdf

Καλαμβόκη, Γ. (2017). Μέθοδοι πρόβλεψης χρονοσειρών: χρονοσειρές στην Ελληνική Οικονομία (Doctoral dissertation).

Κάρλος, Σ. (2020). Ανάπτυξη μερικώς επιβλεπόμενων αλγορίθμων μηχανικής μάθησης (Doctoral dissertation, Πανεπιστήμιο Πατρών. Σχολή Θετικών Επιστημών. Τμήμα Μαθηματικών).

Κουγιουμτζής, Δ. (2011). Χρονοσειρες. Ανακτήθηκε 2 Φεβρουαρίου, 2021, από: https://users. auth.gr/dkugiu/Teach/DataAnalysis/Lecture7NoPause.pdf?fbclid=IwAR0B5-g6Zb6snE9k1fUrDR_xbas8aKos9N2UBOpQ6ba8-KsZnLPQgy_NPiY

Μαργιά, Γ. Π. (2009). Ανάλυση και πρόβλεψη χρονοσειρών (No. GRI-2009-2689). Aristotle University of Thessaloniki.

Μπεγκόμ, Τ. (2013). Ανάλυση των χρηματιστηριακών δεδομένων με χρήση των αλγορίθμων εξόρυξης (Doctoral dissertation).

Ντουντούμι, Κ. (2020). Ομαδοποίηση ιατρικών προφίλ από δημιουργημένη βάση δεδομένων με τεχνικές μηχανικής μάθησης. Ανακτήθηκε 22 Ιανουαρίου, 2021, από: https://dspace.lib.ntua.gr/xmlui/bitstream/handle/123456789/52500/thesis-final-kl-dudumi.pdf ?sequence=1

## Ξενόγλωσση

AA1car. (2019). How to Diagnose a Controller Area Network (CAN). Ανακτήθηκε 9 Οκτωβρίου, 2020, από: https://www.aa1car.com/library/can_systems.htm

Aghabozorgi, S. R., Wah, T. Y., Amini, A., & Saybani, M. R. (2011). A new approach to present prototypes in clustering of time series.

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–a decade review. Information Systems, 53, 16-38.

Analytics-vidhya. (2020). Time Series Forecasting of COVID19 data with FBProphet. Ανακτήθηκε 4 Ιανουαρίου, 2021, από: https://medium.com/analytics-vidhya/time-series-forecasting-of-covid19-data-with-fbprophet-3a73dba59106

Aljaafreh, A., Alshabatat, N., & Al-Din, M. S. N. (2012, July). Driving style recognition using fuzzy logic. In 2012 IEEE International Conference on Vehicular Electronics and Safety (ICVES 2012) (pp. 460-463). IEEE.

Ayodele, T. O. (2010). Types of machine learning algorithms. New advances in machine learning, 3, 19-48.

Bergroth, L., Hakonen, H., & Raita, T. (2000, September). A survey of longest common subsequence algorithms. In Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000 (pp. 39-48). IEEE.

Biliri, E., Kokkinakos, P., Michailitsi, A., Papaspyros, D., Tsapelas, J., Mouzakitis, S., ... & Kirstein, F. (2017, June). Big data analytics in public safety and personal security: challenges and potential. In 2017 International Conference on Engineering, Technology and Innovation (ICE/ITMC) (pp. 1382-1386). IEEE.

Borgatti, S. (n.d). Distance and Correlation. Boston College. Ανακτήθηκε 29 Δεκεμβρίου, 2020, από: https://cmci.colorado.edu/classes/INFO-1301/files/borgatti.htm

Bozdal, M., Samie, M., Aslam, S., & Jennions, I. (2020). Evaluation of can bus security challenges. Sensors, 20(8), 2364.

Brownlee, J . (2020). 4 Distance Measures for Machine Learning. Ανακτήθηκε 9 Ιανουαρίου, 2021, από: https://machinelearningmastery.com/distance-measures-for-machine-learning/#: ~:text = Manhattan %20Distance%20(Taxicab%20or%20City%20Block%20Distance),a%20chessboard%20or%20city% 20blocks.

Csselectronics. (n.d.). CAN Bus Explained - A Simple Intro (2021). Ανακτήθηκε 20 Δεκεμβρίου, 2020, από: https://www.csselectronics.com/screen/page/simple-intro-to-can-bus/language/en?fbclid=IwAR 0B5- g6Zb 6snE9k1fUrDR_xbas8aKos9N2UBOpQ6ba8-KsZnLPQgy_NPiY

Chandrasiri, N. P., Nawa, K., & Ishii, A. (2016). Driving skill classification in curve driving scenes using machine learning. Journal of Modern Transportation, 24(3), 196-206.

Civilprotection. (2020). 3 η Έκδοση Γενικού Σχεδίου Αντιμετώπισης Τεχνολογικών Ατυχημάτων Μεγάλης Έκτασης. Ανακτήθηκε 19 Μαίου, 2020, από: https://www.civilprotection.gr/sites/default/ gscp_uploads/shedio_irakleitos.pdf

Columbia University. (n.d.). Longest Common Subsequence. Ανακτήθηκε 16 Ιανουαρίου, 2021, από: http://www.columbia.edu/~cs2035/courses/csor4231.F11/lcs.pdf

Copperhilltech. (n.d.). A Brief Introduction to Controller Area Network. Ανακτήθηκε 2 Δεκεμβρίου, 2020, από: https://copperhilltech.com/a-brief-introduction-to-controller-area-network/?fbclid=IwAR 0F3Nof37LVs6Bnq3bbBgRztH-bn4ulrP_3JvsnYL2d26HJXQye9H_xWA4

Datasciencecentral. (2019). How to Automatically Determine the Number of Clusters in your Data - and more. Ανακτήθηκε 12 Δεκεμβρίου, 2020, από: https://www.datasciencecentral.com/profiles/blogs /how-to-automatically-determine-the-number-of-clusters-in-your-dat)

Dunham, M. H. (2006). Data mining: Introductory and advanced topics. Pearson Education India.

Esling, P., & Agon, C. (2012). Time-series data mining. ACM Computing Surveys (CSUR), 45(1), 1-34.

European Commission (2017). Safer roads for all, The EU Good Practice Guide. Ανακτήθηκε 17 Μαίου, 2020, από: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/safer_ roads4all.pdf

European Commission. (2019a). Road safety. Ανακτήθηκε 30 Μαίου, 2020, από: https://ec.europa.eu/ commission/news/road-safety-2019-apr-04_en

European Commission. (2019b). Advanced Big Data Value Chain for Public Safety and Personal Security Ανακτήθηκε 27 Μαίου, 2020, από: https://cordis.europa.eu/project/id/732189

European Transport Safety Council. (2019). Road safety: Data show improvements in 2018 but further concrete and swift actions are needed. Ανακτήθηκε 23 Απριλίου, 2020, από: https://ec.europa.eu/ commission/presscorner/detail/en/IP_19_1951

European Parliament. (2019). Road fatality statistics in the EU (infographic). Ανακτήθηκε 10 Απριλίου, 2020, από: https://www.europarl.europa.eu/news/en/headlines/society/20190410STO36615/road-fatality-statistics-in-the-eu-infographic

Ferreira, L. N., Ferreira, N. C., Gava, M. L. L., Zhao, L., & Macau, E. E. (2019). The influence of time series distance functions on climate networks. Ανακτήθηκε 8 Δεκεμβρίου, 2020, από: https://www.researchgate.net/figure/Time-series-comparison-using-lock-step-and-elastic-measures-The-total-distance-is_fig1_330970291

Figueirêdo, I., Guarieiro, L. L. N., & Nascimento, E. G. S. (2020). Multivariate Real Time Series Data Using Six Unsupervised Machine Learning Algorithms. In Anomaly Detection-Recent Advances, Issues and Challenges. IntechOpen.

Guo, G., Huang, K., & Yang, C. (2016). Time Series Classification Based On The Longest Common Subsequence Similarity And Ensemble Learning. Ανακτήθηκε 20 Δεκεμβρίου, 2020, από: http:// par.cse.nsysu.edu.tw/~cbyang/person/publish/c16time_series.pdf

Jarašūniene, A., & Jakubauskas, G. (2007). Improvement of road safety using passive and active intelligent vehicle safety systems. Transport, 22(4), 284-289.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.

Keogh, E., & Lin, J. (2005). Clustering of time-series subsequences is meaningless: implications for previous and future research. Knowledge and information systems, 8(2), 154-177.

Kyaagba, S. (2019). An Intro to Graph Databases. Ανακτήθηκε 16 Ιανουαρίου, 2021, από: https://medium.com/@shachiakyaagba_41915/dynamic-time-warping-with-time-series-1f5c05fb8950

Liao, T. W. (2005). Clustering of time series data—a survey. Pattern recognition, 38(11), 1857-1874.

Madhukar, B. (2020). Tutorial On Missingno – Python Tool To Visualize Missing Values. Ανακτήθηκε 14 Ιανουαρίου, 2021, από: https://analyticsindiamag.com/tutorial-on-missingno-python-tool-to-visualize-missing-values/

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

Marous, J. (2012). Big Data: Big Opportunity In Banking… Or Big B.S.?. THE FINANCIAL BRAND. Ανακτήθηκε 10 Νοεμβρίου 2020, από: https://thefinancialbrand.com/26363/big-data-analytics-retail-banking-jm/

National Highway Traffic Safety Administration – NHTSA. (2016). Traffic safety facts research note. Technical Report DOT HS 812 318, U.S. Department of Transportation. Year:2018

Niennattrakul, V., & Ratanamahatana, C. A. (2007, April). On clustering multimedia time series data using k-means and dynamic time warping. In 2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07) (pp. 733-738). IEEE.

Nzfaruqui. (2019). Types of Machine Learning. Ανακτήθηκε 4 Ιανουαρίου, 2021, από: https://www.nzfaruqui.com/types-of-machine-learning/

Oesincorp. (n.d.). Three-axis Yaw, Pitch and Roll Stage. Ανακτήθηκε 16 Ιουνίου, 2020, από: http://www.oesincorp.com/motorized-multi-axis-stages/YPR60-60-60.htm

Oregi, I., Pérez, A., Del Ser, J., & Lozano, J. A. (2019). On-line Elastic Similarity Measures for time series. Pattern Recognition, 88, 506-517.

Özkoç, E. E. (2020). Clustering of time-series data. In Data Mining-Methods, Applications and Systems. IntechOpen.

Paparrizos, J., & Gravano, L. (2015, May). k-shape: Efficient and accurate clustering of time series. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (pp. 1855-1870).

Repositorykallipos. (n.d.). Κεφάλαιο 5: Ανάλυση Συστάδων. Ανακτήθηκε 16 Ιανουαρίου 2021, από: https://repository.kallipos.gr/bitstream/11419/2130/1/06_chapter05.pdf?fbclid=IwAR3ELZSo5HdA2-nAoCuKdRTbqjJYc7C8eP99ylkm7dnuz6PhtHF0XU55iPA

Rhdjapan. (n.d.). MIDORI SEIBI CENTER DIGITAL YAW RATE SENSOR UNIT - BNR34 V-SPEC. Ανακτήθηκε 10 Ιανουαρίου 2021, από: https://www.rhdjapan.com/midori-seibi-center-digital-yaw-rate-sensor-unit-bnr34-v-spec.html

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

Slideshare. (2019). Lesson 2 stationary_time_series. Ανακτήθηκε 6 Ιανουαρίου 2021, από: https:// www.slideshare.net/ankit_ppt/lesson-2-stationarytimeseries

Statisticshowto. (2013). Time Series Transformation. Ανακτήθηκε 26 Οκτωβρίου 2020, από: https://www.statisticshowto.com/timeplot/

Stefanatos, D. (2015). Γυροσκόπια και Εφαρμογές Αυτών στην Πλοήγηση. Space Sciences, Hellenic Military Academy. Ανακτήθηκε 10 Οκτωβρίου 2020, από: http://users.otenet.gr/~elpida_1/ Dionisis _Stefanatos/Gyroskopia_Stefanatos_Shmeiwseis.pdf

Studytonight. (n.d.). Time Complexity of Algorithms. Ανακτήθηκε 17 Ιανουαρίου 2021, από: https://www.studytonight.com/data-structures/time-complexity-of-algorithms.

Turing, I. B. A. (1950). Computing machinery and intelligence-AM Turing. Mind, 59(236), 433. Ανακτήθηκε 1 Δεκεμβρίου, 2020, από: https://academic.oup.com/mind/article/LIX/236/433/986238

Unipi-eclass. (n.d). Χρονοσειρές - Μέθοδοι Εξομάλυνσης. Ανακτήθηκε 17 Δεκεμβρίου, 2020, από: http://stat.unipi.gr/eclass/modules/document/file.php/EFA158/%CF%87%CF%81%CE%BF%CE%BD%CE%BF%CF%83%CE%B5%CE%B9%CF%81%CE%B5%CF%82.pdf

University of Minnesota. (2018). Definition of Vehicle Heading and Steeing Angle. Ανακτήθηκε 12 Δεκεμβρίου, 2020, από: http://street.umn.edu/VehControl/javahelp/HTML/Definition_of _Vehicle_ Heading_ and_Steeing_Angle.htm

Van Ly, M., Martin, S., & Trivedi, M. M. (2013, June). Driver classification and driving style recognition using inertial sensors. In 2013 IEEE Intelligent Vehicles Symposium (IV) (pp. 1040-1045). IEEE.

Van Overeem, S., Alen, L., Brouwer, Y., Van Dam, A. D., VAN, G. M., DEKKEN, G. H. G., ... & HARRY, W. (2017). Wind assessment for micro wind turbines in an urban environment. Ανακτήθηκε 9 Οκτωβρίου, 2020, από: https://www.researchgate.net/figure/Coordinate-system-with-velocity-vector-U-xyz-and-relevant-angles_fig3_320628520

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery, 26(2), 275-309.

Wei, H. (2020). How to measure clustering performances when there are no ground truth?. Ανακτήθηκε 19 Δεκεμβρίου, 2020, από: https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c

Yadav, P., Jung, S., & Singh, D. (2019, April), Machine learning based real-time vehicle data analysis for safe driving modeling. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (pp. 1355-1358).

Zardosht, M., Beauchemin, S. S., & Bauer, M. A. (2018), Identifying driver behavior in preturning maneuvers using in-vehicle CANbus signals. Journal of Advanced Transportation, 2018.

Zenodo. (2019). Automotive CAN bus data: An Example Dataset from the AEGIS Big Data Project. Open access data set.