Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικης Ισχυος

Εντοπισμός ρευματοκλοπών με μηχανική μάθηση

Διπλωματικη Εργασια

του

ΜΗΤΣΕΛΟΥ ΑΘΑΝΑΣΙΟΥ

Επιβλέπων: Χατζηαργυρίου Νικόλαος
Καθηγητής Ε.Μ.Π.

Εργαστηριο Συστηματων Ηλεκτρικης Ενεργειας
Αθήνα, Οκτώβριος 2017

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρικής Ισχύος
Εργαστήριο Συστημάτων Ηλεκτρικής Ενέργειας

Εντοπισμός ρευματοκλοπών με μηχανική μάθηση

Διπλωματικη Εργασια

του

ΜΗΤΣΕΛΟΣ ΑΘΑΝΑΣΙΟΣ

Επιβλέπων: Χατζηαργυρίου Νικόλαος
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 666 Οκτωβρίου 2017.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Χατζηαργυρίου Νικόλαος	Παπαθανασίου Σταύρος	Γεωργιλάκης Παύλος
Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.	Επ. Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2017

(Υπογραφή)

.....

ΜΗΤΣΕΛΟΥ ΑΘΑΝΑΣΙΟΥ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρικής Ισχύος
Εργαστήριο Συστημάτων Ηλεκτρικής Ενέργειας

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Βασίλειο Ασημακόπουλο για την ευκαιρία που μου έδωσε να εκπονήσω τη παρούσα διπλωματική και την υποστήριξή του σε όλη την πορεία της.

Επίσης, θα ήθελα να ευχαριστήσω τους καθηγητές κ. Ιωάννη Ψαρρά και κ. Δημήτριο Ασκούνη για την τιμή που μου έκαναν να συμμετάσχουν στην επιτροπή εξέτασης της διπλωματικής.

Ευχαριστώ ιδιαίτερα τον υποψήφιο διδάκτορα Ευάγγελο Σπηλιώτη για την καθοδήγηση, στήριξη και καθοριστική βοήθεια που μου παρείχε, όπως και τα υπόλοιπα μέλη της Μονάδας Προβλέψεων και Στρατηγικής.

Θερμές ευχαριστίες θα ήθελα να απευθύνω στον Δρ Χριστόφορο Αναγνωστόπουλο και την εταιρία Mentat Innovations για την καθοδήγησή τους στα πρώτα βήματα αυτής της εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου Γιώργο, Γρηγόρη, Κατερίνα και Μαρία.

Περίληψη

Αντικείμενο της διπλωματικής εργασίας είναι η ανάπτυξη μεθοδολογίας για τη βελτίωση της ακρίβειας στατιστικών μεθόδων πρόβλεψης σε χρονοσειρές που έχουν μικρό ιστορικό παρατηρήσεων μέσω τεχνικών συσταδοποίησης εποχιακών δεικτών από συναφείς χρονοσειρές.

Οι κλασικές μέθοδοι αποσύνθεσης απαιτούν ένα ελάχιστο πλήθος παρατηρήσεων για να μπορέσουν να εξάγουν το μοτίβο της εποχιακότητας μιας χρονοσειράς. Στη πράξη, όμως, συναντάμε συχνά χρονοσειρές που αποτελούνται από μικρό πλήθος τιμών, ενώ συγχρόνως περιγράφουν εποχιακά μεγέθη.

Παράλληλα, τα τελευταία χρόνια υπάρχει αφθονία στα δεδομένα που έχουμε στη διάθεσή μας. Η παρούσα εργασία βασίζεται στην υπόθεση ότι μπορούμε να χρησιμοποιήσουμε τη διαθέσιμη πληροφορία για να εξάγουμε αντιπροσωπευτικούς δείκτες εποχιακότητας που μπορούμε να χρησιμοποιήσουμε για να αναλύσουμε και να προεκτείνουμε χρονοσειρές που χαρακτηρίζονται από μικρό ιστορικό.

Για να το κάνουμε αυτό πρέπει αρχικά να συγκεντρώσουμε ένα πλήθος χρονοσειρών που περιγράφει παρόμοια φυσικά μεγέθη. Έπειτα, χρησιμοποιώντας τεχνικές συσταδοποίησης στους δείκτες εποχιακότητας αυτών που έχουν επαρκή δεδομένα για να εφαρμόσουμε τις κλασικές μεθόδους αποσύνθεσης, δημιουργούμε συστάδες παρόμοιας εποχιακής συμπεριφοράς. Ελέγχουμε, κατόπιν, αν οι μικρές χρονοσειρές μπορούν να υπαχθούν σε αυτές τις συστάδες και αν ναι, τις προβλέπουμε με δεδομένο ότι οι δείκτες εποχιακότητας τους είναι οι ίδιοι με τους μέσους δείκτες των συστάδων.

Για να ελέγξουμε την υπόθεση, εφαρμόσαμε την μεθοδολογία που περιγράφηκε σε ένα σύνολο χρονοσειρών ζήτησης φυσικού αερίου και λάβαμε θετικά αποτελέσματα. Συγκεκριμένα συγκρίναμε τη προτεινόμενη προσέγγιση με τη κλασική, που προεκτείνει τις μικρές χρονοσειρές βάσει των αρχικών τους δεδομένων και παρατηρήσαμε σημαντική βελτίωση της ακρίβειας.

Λέξεις Κλειδιά

Χρονοσειρές, Τεχνικές Προβλέψεων, Εποχιακότητα, Συσταδοποίηση, Μικρό ιστορικό, Φυσικό Αέριο.

Abstract

The purpose of this diploma thesis is to develop a methodology for improving the accuracy of statistical forecasting methods on timeseries with short history through the use of clustering techniques on the seasonal indices of other similar timeseries.

Classical decomposition methods require a minimum number of observations to be able to detect the seasonality pattern of a timeseries. In practice, however, we often encounter timeseries lacking enough data, while at the same time describing seasonal values.

Meanwhile, in recent years, there is an abundance of accessible data. This thesis draws upon the hypothesis that we can utilise the available infomation to extract representative seasonality indices that we can use in order to analyse and extend timeseries that are characterised by short history.

In order to achieve this, we initially have to gather a large number of timeseries describing similar values. Afterwards, we create clusters of similar seasonal behaviour by using clustering techniques on the seasonality indices of series with sufficient data. Then, we check if the shorter timeseries qualify to be a part of these clusters and if so, we predict their future values as they were characterised by the seasonal behaviour of the mean indices of the cluster members.

To test our hypothesis, we applied the described methodology to a set of natural gas demand timeseries and received positive results. In particular, we compared the proposed approach to the classical one, which forecasts short timeseries based on their original data, and we have measured a significant overall improvement in accuracy.

Keywords

Timeseries, Forecasting Techniques, Seasonality, Clustering, Short history, Natural Gas.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Είναι ευρέως διαδεδομένο πως η καθημερινότητα πολλών ανθρώπων συνδέεται άρρηκτα με τη χρήση ηλεκτρικών συσκευών, αλλά και με την ανάγκη ύπαρξης βιομηχανικών εγκαταστάσεων για την εκπλήρωση των καταναλωτικών τους επιθυμιών. Αυτό δημιουργεί μια αυξανόμενη ζήτηση στον τομέα της παραγωγής, της μεταφοράς και διανομής ηλεκτρικής ενέργεια, που με τη σειρά του οδηγεί στον συνεχή εκσυγχρονισμό των εγκαταστάσεων. Παράλληλα, διανύοντας την εποχή της Ψηφιακής Επανάστασης παρατηρείται η μετάβαση από τις αναλογικές τεχνολογίες στις ψηφιακές, γεγονός που δεν θα μπορούσε να αφήσει ανεπηρέαστο τον τομέα της ηλεκτρικής ενέργειας. Η μετάβαση αυτή στον τομέα που μελετάται σε αυτή τη διπλωματική εργασία σηματοδοτείται από την χρήση έξυπνων μετρητών, οι οποίοι έχουν τη δυνατότητα να παρέχουν σε πραγματικό χρόνο μεγάλο όγκο δεδομένων για τα επίπεδα της κατανάλωσης κάθε πελάτη.

Ανοίγεται, λοιπόν ένας νέος ορίζοντας εποπτείας και αναλυτικής μελέτης των χρονοσειρών που παράγονται από κάθε καταναλωτή. Η ταυτόχρονη και συνεχής αύξηση των ρευματοκλοπών στις περισσότερες περιοχές του κόσμου καθιστά επιτακτική ανάγκη την εύρεση μεθόδων εντοπισμού τους . Σύμφωνα με τα επίσημα στοιχεία του Διαχειριστή Δικτύου (ΔΕΔΔΗΕ), το 2016 εντοπίσηκαν 10.616 κρούσματα ρευματοκλοπών, μέγεθος που είναι ψηλότερο όλων των εποχών, έναντι 400 το 2006 [31]. Άμεσο επακόλουθο της επίλυσης αυτού προβλήματος είναι η ομαλή λειτουργία των παροχέων ενέργειας και η βελτίωση της ποιότητας των υπηρεσιών που παρέχουν οι ίδιες. Στη συνέχεια θα αναπτυχθεί το βαθύτερο αίτιο της παρούσας διατριβής και μια επισκόπηση του περιεχομένου της [25].

1.1 Κίνητρο και υπόβαθρο διπλωματικής

Το πρόβλημα της παράνομης αφαίρεσης ηλεκτρικής ενέργειας ενδιαφέρει τους διαχειριστές δικτύων. Οι χρήστες συχνά παραβιάζουν τους νόμους προσπαθώντας να αλλοιώσουν τα συστήματα μέτρησης. Σε κάποιες χώρες μόνο κάποιο κομμάτι της παραγωγής χρεώνεται, παραδείγματος χάριν στην Ινδία το 55% της παραγωγής ηλεκτρικής ενέργειας χρεώνεται (και μόνο ένα μέρος της πληρωμής καταλήγει στον πάροχο). Παρόλα αυτά, η παράνομη χρήση ενέργειας λαμβάνει χώρα και σε Ευρωπαϊκές χώρες. Μια από τις κινητήριες δυνάμεις για

το λανσάρισμα των αυτοματοποιημένων υποδομών ανάγνωσης μετρητών (Automated Meter Reading) για τον ιταλικό πάροχο ενέργειας (ENEL) ήταν η προσπάθεια ελαχιστοποίησης των μη τεχνικών απωλειών στο δίκτυα διανομής τους. Η μείωση των ρευματοκλοπών βοήθησε στην αιτιολόγηση μεγάλων επενδύσεων σε AMR και επί του παρόντος η Ιταλία πρωταγωνιστεί στην διείσδυση AMR [23],[5].
Μερικοί μπορεί να υποστηρίζουν ότι οι εταιρίες παραγωγής και διανομής, οι οποίες έχουν σημαντικό έργο παρέχουν κακή εξυπηρέτηση, υπερχρεώνουν, κερδίζουν ανεξαρτήτως αρκετά χρήματα και ως εκ τούτου, ένα ποσοστό κλοπής δεν θα καταστρέψει την εταιρία ή θα επηρεάσει δραστικά τις λειτουργίες και την κερδοφορία της. Άλλοι παρατηρώντας την ίδια κατάσταση θα υποστήριζαν ότι η κλοπή είναι έγκλημα και δεν θα έπρεπε να επιτρέπεται. Η Διεθνής Εταιρία Προστασίας Εσόδων των Πάροχων (International Utilities Revenue Protection Association) έχει καθιερωθεί για να προάγει τον εντοπισμό και την πρόληψη της κλοπής ρεύματος κυρίως για την οικονομική ασφάλεια των εταιριών παροχής ενέργειας.

Οι συνέπειες της κλοπής είναι εξαιρετικά σημαντικές και μπορούν να επηρεάσουν άμεσα τη βιωσιμότητα των υπηρεσιών που παρέχονται. Οι συνδιασμένες απώλειες (συμπεριλαμβάνοντας και τους απλήρωτους λογαριασμούς) σε μερικά συστήματα έχουν σοβαρές επιπτώσεις που έχουν ως αποτέλεσμα οι εγκαταστάσεις να λειτουργούν σε καθεστώς μεγάλων απωλειών και αναγκάζονται να αυξάνουν συνεχώς τα ηλεκτρικά φορτία. Απομονωμένες σε μια κουλτούρα αναποτελεσματικότητας και διαφθοράς, οι εταιρίες έχουν μεγάλη δυσκολία να παρέχουν αξιόπιστες υπηρεσίες. Ακόμη και σε αποτελεσματικά συστήματα ισχύος, όπως η Tenaga της Μαλαισίας, η κλοπή ρεύματος ανέρχεται στα $132 εκατομμύρια ετησίως [15]. Αντίστοιχα στην Ελλάδα η συνολική εγχεόμενη ενέργεια στο Δίκτυα Διανομής ανήλθε το 2016 σε 47.655.372 MWh, το σύνολο των ρευματοκλοπών εκτιμάται σε 1.525.292 MWh. Στην πραγματικότητα όμως το μέγεθος των ρευματοκλοπών είναι αρκετά μεγαλύτερο, επιβαρύνει δε κατά κύριο λόγο τη Δημόσια Επιχείρηση Ηλεκτρισμού (ΔΕΗ). Ωστόσο παίρνοντας ως δεδομένη την ποσότητα, που αναγνωρίζει η Ρυθμιστική Αρχή Ενέργειας (ΡΑΕ), τα έσοδα που διαφεύγουν κάθε χρόνο λόγω των ρευματοκλοπών με βάση τις μοναδιαίες τιμές του 2016 έχουν ως εξής [30]:

Εταιρίες	εκατ. €
ΔΕΗ	120-125
Υπηρεσίες Κοινής Ωφέλειας (ΥΚΩ)	21
ΕΤΜΕΑΡ	32
ΑΔΜΗΕ 4	7,3
ΔΕΔΔΗΕ 5	26,5
Σύνολο	206,8 έως 211,8

Πίνακας 1.1: Διαφεύγοντα έσοδα Ελληνικών εταιριών λόγω ρευματοκλοπών

1.1.1 Ορίζοντας τις ρευματοκλοπές

Σύμφωνα με το εγχειρίδιο ρευματοκλοπών της ΡΑΕ ρευματοκλοπή ορίζεται εν γένει η αυθαίρετη και με δόλο επέμβαση σε εξοπλισμό ή εγκαταστάσεις του Δικτύου, με σκοπό την

κατανάλωση ηλεκτρικής ενέργειας χωρίς αυτή να καταγράφεται, ή χωρίς να αντιστοιχίζεται με Εκπρόσωπο Φορτίου, και να μην τιμολογείται [32]. Υπάρχουν τέσσερα επικρατούντα είδη ¨κλοπής' σε όλα τα συστήματα ενέργειας. Η έκταση της κλοπής εξαρτάται από πλήθος παραγόντων από πολιτιστικές μέχρι τον τρόπο που διαχειρίζεται η ενέργεια.

Επέμβαση στο μετρητή

Επέμβαση στο μετρητή ορίζεται όταν ο καταναλωτής σκοπίμως προσπαθεί να εξαπατήσει τον πάροχο. Μια συνήθης πρακτική είναι να παραβιάζει το μετρητή ώστε να καταγράφει χαμηλότερα ποσά ενέργειας από τα πραγματικά. Αυτό εν γένει είναι μια επικίνδυνη διαδικασία για ένα ερασιτέχνη, και σε πολλές περιπτώσεις έχουν καταγραφεί ηλεκτροπληξίες. Στην Ελλάδα πρόκειται για τη συνηθέστερη περίπτωση ρευματοκλοπής [32].

Απευθείας Σύνδεση

Η κλοπή ενέργειας επιτευχθεί τραβώντας μια γραμμή από την από το δίκτυο διανομής μέχρι το επιθυμητό σημείο παρακάμπτοντας το μετρητή. Ένας καθιερωμένος τρόπος κλοπή ενέργειας στην Ελλάδα είναι η απευθείας σύνδεση με αγκίστρωση στους αγωγούς του εναέριου δικτύου, απουσία μετρητικής διάταξης ή παροχής ή νομίμως υφιστάμενου κτίσματος [32].

Ακανόνιστες χρεώσεις

Οι ακανόνιστες χρεώσεις μπορούν να συμβούν από πολλές πηγές. Κάποιο οργανισμοί παροχής ενέργειας μπορεί να μην είναι αρκετά αποτελεσματικοί στη μέτρηση της ενέργειας που έχει καταναλωθεί και ακούσια μπορεί να δώσουν υψηλότερη ή χαμηλότερη μέτρηση από την ακριβή. Αυτές οι ακανόνιστες χρεώσεις μπορεί να ισοζυγιστούν με την πάροδο του χρόνου. Παρόλα αυτά, είναι πολύ εύκολο σε μερικά συστήματα να κανονιστούν πολύ χαμηλότεροι λογαριασμοί από τους ρεαλιστικούς. Εργαζόμενοι μπορεί να δωροδοκηθούν για να καταγράψουν το μετρητή με μικρότερο νούμερο από αυτό που ενδεικνύεται. Ο καταναλωτής πληρώνει μικρότερο λογαριασμό και ο εργαζόμενος που καταγράφει τις μετρήσεις αποκτά ανεπίσημο μισθό.

Απλήρωτοι λογαριασμοί

Κάποια άτομα και κάποιοι οργανισμοί δεν πληρώνουν αυτά που οφείλουν για ηλεκτρική ενέργεια. Οικιακοί ή επιχειρηματικοί καταναλωτές μπορεί να έχουν φύγει από την πόλη ή την εγκατάσταση λόγω χρεωκοπίας. Στη Νότιο Αμερική, υπάρχει«καθεστώς μη πληρωμής» [14]. Στην Αρμενία, «τα επίπεδα μη πληρωμής είναι της τάξης του 80-90% για τον οικιακό τομέα. Οι απώλειες μετασχηματισμού και διανομής είναι άνω του 40%»[26].

Σε όλες τις χώρες, καθώς η τιμή της ηλεκτρικής ενέργειας αυξάνεται, κάποιοι άνθρωποι αδυνατούν να πληρώσουν τους λογαριασμούς τους με συνέπεια. Αυτό τους ενθαρρύνει να βρουν τρόπους να μειώσουν τους λογαριασμούς, όπως να πειράζουν τους μετρητές.

1.2 Δομή Διπλωματικής

Στον παρόν τόμο γίνεται μια διεξοδική αναζήτηση μεθόδων ανίχνευσης απάτης με μια πληθώρα διαφορετικών αλγορίθμων από την σκοπιά της μηχανικής μάθησης. Δεδομένου του εύρους των δυνατοτήτων της μηχανική μάθησης γίνεται προσπάθεια για αντιμετώπιση του προβλήματος από διαφορετικές οπτικές γωνίες, προσπαθώντας να επιτευχθεί η βέλτιστη αντιστάθμιση μεταξύ ευστοχίας και πρακτικότητας. Η εξισορρόπηση αυτών των παραγόντων είναι κύριο μέλημα κάθε μηχανικού [16]. Ειδικότερα, συνοψίζοντας κάθε κεφάλαιο εξάγεται η παρακάτω δομή:

Κεφάλαιο 1

Γνωστοποιείται η κινητήριος δύναμη αυτής της διπλωματικής, κάνοντας ένα σαφή ορισμό του προβλήματος προς αντιμετώπιση.

Κεφάλαιο 2

Γίνεται μια εισαγωγή στα εργαλεία που χρησιμοποιούνται για την λήψη των αρχικών χρονοσειρών, την επεξεργασία τους και ταξινόμηση των καταναλωτών, αλλά και για τις συνιστώσες που λαμβάνονται υπόψιν για τα τελικά αποτελέσματα.

Κεφάλαιο 3

Αναπτύσσεται η μορφή και φύση των δεδομένων, αλλά και η μεθοδολογία προεπεξεργασίας τους. Παράλληλα, διευκρινίζεται ο τρόπος προσομοίωσης και μοντελοποίησης της ρευματοκλοπής.

Κεφάλαιο 4

Δημιουργείται ένας άξονας αναφοράς για τα αποτελέσματα με τη χρήση αλγορίθμων επιβλεπόμενης μάθησης που φημίζονται για την μεγάλη ευστοχία τους, αλλά και την δυσκολία εφαρμογής τους σε πραγματικά προβλήματα.

Κεφάλαιο 5

Εξετάζονται λεπτομερώς τα συστατικά των αλγορίθμων μη-επιβλεπόμενης μάθησης, ενώ παράλληλα διεξάγεται δοκιμές για την εξερεύνηση των διαφορετικών μεθόδων επίλυσης του θέματος.

Κεφάλαιο 6

Επεξηγούνται οι δυσκολίες που αντιμετωπίστηκαν από το διαφορετικά πρίσματα. Αναλυτικότερα γνωστοποιούνται τα τεχνικά εμπόδια που αντιμετωπίστηκαν, αλλά και τα εμπόδια που θα αντιμετωπίσου οι καταναλωτές, προσπαθώντας να οριστεί ένα μονοπάτι αποφυγής τους και αρμονικής συνύπαρξης των δύο πλευρών.

Κεφάλαιο 7

Γίνεται σφαιρική εποπτεία των αποτελεσμάτων με γνώμονες τη φύση κάθε αλγορίθμου και την ευστοχία στην ταξινόμηση των καταναλωτών.

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

Η ηλεκτρική ενέργεια είναι ζωτικής σημασίας για την καθημερινότητά μας αλλά και ο ακρογωνιαίος λίθος της βιομηχανίας. Για αυτό το λόγο έννοια των μελλοντικών δικτύων (έξυπνα δίκτυα) στοχεύει στην αύξηση της αξιοπιστίας, της ποιότητας και της ασφάλειας της μελλοντικής παροχής ενέργειας. Για να συμβεί αυτό, απαιτούνται περαιτέρω πληροφορίες για την λειτουργία και την κατάσταση των δικτύων διανομής. Μια από τις σημαντικότερες προκλήσεις στα μελλοντικά δίκτυα διανομής είναι η αυξανόμενη διείσδυση κατανεμημένης παραγωγής (Distributed Generation) που συνδέεται στα κτίρια των καταναλωτών και η μετάβαση από την έννοια της παραδοσιακής παραγωγής ενέργειας με κυρίαρχους μεγάλους σταθμούς παραγωγής ενέργειας και ροές ενέργειας μονής κατεύθυνσης σε πιο περίπλοκες τροφοδοσίες ισχύος. Οι πληροφορίες λειτουργίας θα είναι καίριας σημασίας για τη λειτουργικότητα των μελλοντικών δικτύων διανομής και για τους διαχειριστές του δικτύου (Distribution Network Operators). Μια από της πηγές πληροφορίας θα είναι η υποδομή έξυπνων μετρητών. Εκτός των άλλων, οι έξυπνοι μετρητές πρέπει να διευρύνουν τους γνωστικούς ορίζοντες των καταναλωτών για την ηλεκτρική ενέργεια. Η έννοια αυτή θα παράξει ακόμη περισσότερη πληροφορία στου διαχειριστές δικτύου. Αυτό παρέχει τη δυνατότητα στο διαχειριστή του δικτύου να αναλύσει ροές ενέργειας και να εντοπίσει πιθανή κλοπή ρεύματος [18].

2.1 Έξυπνοι μετρητές

Η ηλεκτρική διανομή είναι ένας τομέας, που η εξέλιξη είναι σταδιακή, τουλάχιστον όσο αφορά τα περιουσιακά στοιχεία του δικτύου. Παρόλα αυτά, υπάρχει ένας κλάδος, στον οποίο η πρόοδος τα τελευταία χρόνια είναι ταχύτατη, με ταχύτητα τυπική για τον κλάδα των τηλεπικοινωνιών. Απομακρυσμένες μετρήσεις, αναγνώσεις και παρακολουθήσεις της κατανάλωσης αναφέρονται ως προηγμένη υποδομή μέτρησης (Advanced Metering Infrastructure). Η δραστική μείωση στις τιμές των μετρητών και στον εξοπλισμό τηλεπικοινωνιών κάνει την απόκτησή τους οικονομικά βιώσιμη, ξεκινώντας με μεγάλους καταναλωτές σταδιακά εφαρμόζοντάς τους και στους μέσους και μικρούς. Η αποτελεσματικότητα των εργαλείων στην αναγνώριση και αποθάρρυνση της κλοπής και άλλων τρόπων παράκαμψης μετρητών είναι τεράστια, όπως φαίνεται

να συμβαίνει σε αναπτυσσόμενες χώρες (συμπεριλαμβανομένου της Δομινικής Δημοκρατίας, της Χοντούρας και της Βραζιλίας).

Η ευρεία εφαρμογή AMI μπορεί να συμβάλει σημαντικά στην συνεχή ανάπτυξη και την αποτελεσματική λειτουργία. Οι AMI παρέχουν ισχυρά εργαλεία για να μειώσουν τις συνολικές απώλειες και να αυξήσουν τα ποσοστά συλλογής.

2.1.1 Θετικά αντίκτυπα εφαρμογής AMI

Η εφαρμογή των AMI θα έχει τα ακόλουθα θετικά αντίκτυπα:

1. Αίσθηση παρακολούθησης στους χρήστες. Οι καταναλωτές αντιλαμβάνονται πως η παροχή μπορεί να παρακολουθεί την κατανάλωση. Αυτό επιτρέπει στην εταιρία γρήγορη ανίχνευση οποιασδήποτε ανωμαλία στην κατανάλωση, λόγω αλλοίωσης του μετρητή ή παράκαμψής του και της δίνει τη δυνατότητα να κάνει διορθωτικές κινήσεις. Το αποτέλεσμα είναι η πειθάρχηση των καταναλωτών.

2. Ενίσχυση της εταιρικής διακυβέρνησης της εταιρίας και της καταπολέμησης της διαφθοράς. Τα παραδείγματα κλοπής μεγάλων καταναλωτών συνήθως συμπεριλαμβάνουν συνεννόηση μεταξύ αυτών και των ελεγκτών των μετρητών. Η διαφθορά είναι επίσης πιθανό να παρατηρηθεί και στις ενέργειες που συσχετίζονται με την αποσύνδεση του μετρητή, λόγω απλήρωτων λογαριασμών. Η είσοδος των μετρητών κάνει τις πληροφορίες των μετρητών διαθέσιμες στους καταναλωτές και τους διαχειριστές, επιβάλλοντας διαφάνεια.

3. Υλοποίηση προπληρωμένων καταναλώσεων. Η προ-πλήρωση των λογαριασμών είναι γενικώς κάτι πολύ καλό για τους καταναλωτές μικρού εισοδήματος. Οι AMI δίνουν τη δυνατότητα αντιγραφής του επιχειρηματικού μοντέλου των εταιριών κινητής τηλεφωνίας και στην τομέα της ενέργειας.

4. Ελαχιστοποίηση απωλειών σε μη διαχειρίσιμες περιοχές. Οι AMI έχει καθοριστικό ρόλο στην προσέγγιση της διανομής μέσης τάσης (Medium-Voltage Distribution), που χρησιμοποιείται για την κατασκευή και λειτουργία ηλεκτρικών δικτύων, για την παροχή ενέργειας σε περιοχές που η πρόσβαση της εταιρίας είναι περιορισμένη για λόγους ασφαλείας. Στα Μ΄Δ δίκτυα κάθε κάθε σύνδεση καταναλωτή ξεκινάει απευθείας από το μετασχηματιστή μέσης σε χαμηλή τάση, με το δίκτυο χαμηλής τάσης να εκλείπει.

5. Διαχείριση από την πλευρά της ζήτησης για μεγιστοποίηση αποτελεσματικότητας στην παροχή και κατανάλωση ενέργειας. Οι μόνιμοι AMI μέσα σε έξυπνο δίκτυο επιτρέπουν την βελτιστοποίηση της κατανάλωσης ενέργειας ενημερώνοντας τους χρήστες σε πραγματικό χρόνο για τις τιμές, την αρχή και το τέλος των περιόδων αιχμής της κατανάλωσης, το άθροισμα της κατανάλωσης, συναγερμούς κτλ [1].

2.2 Μηχανική μάθηση

Υπάρχουν διαφορετικοί τρόποι που ένας αλγόριθμος μπορεί να μοντελοποιήσει ένα πρόβλημα βασισμένος στην αλληλεπίδραση με την εμπειρία ή το περιβάλλον ή οτιδήποτε μπορεί να καλεστεί δεδομένα εισόδου. Είναι δημοφιλές στα βιβλία μηχανικής μάθησης και τεχνητής νοημοσύνης να εξεταστεί ο τρόπος εκμάθησης που ένας αλγόριθμος μπορεί να υιοθετήσει. Υπάρχουν μόνο μερικοί βασικοί τρόποι εκμάθησης ή μοντέλα εκμάθησης που ένας αλγόριθμος μπορεί χρησιμοποιήσει και θα αναφερθεί κάθε ένας με λίγα παραδείγματα από αλγορίθμους και τύπους προβλημάτων που ταιριάζει σε καθέναν. Αυτή η ταξινομία ή ο τρόπος οργάνωσης των αλγορίθμων είναι χρήσιμος, καθώς αναγκάζει το χρήστη να σκεφτεί το ρόλο των δεδομένων εισόδου και το μοντέλο επεξεργασίας και να επιλέξει τον κατάλληλο αλγόριθμο για το πρόβλημα, με στόχο τα βέλτιστα αποτελέσματα. Παρακάτω αναλύονται οι τρεις διαφορετικές κατηγορίες αλγορίθμων μηχανικής μάθησης με βάση τον τρόπο εκμάθησης.

2.2.1 Επιβλεπόμενη μάθηση

Τα δεδομένα εισόδου καλούνται δεδομένα εκπαίδευσης και είναι γνωστά τα δυαδικά χαρακτηριστικά ή τα αποτελέσματα όπως για παράδειγμα αν είναι ένα δεδομένο ανεπιθύμητο ή όχι ή η τιμή σε μια ορισμένη χρονική περίοδο. Ένα μοντέλο χτίζεται στη φάση της εκπαίδευσης κατά την οποία απαιτείται να κάνει προβλέψεις και να τις διορθώσει όταν είναι λάθος. Η διαδικασία της εκπαίδευσης συνεχίζει μέχρι το μοντέλο να επιτύχει το επίπεδο ευστοχίας στα δεδομένα εκπαίδευσης. Κάποια τέτοια προβλήματα είναι τα προβλήματα ταξινόμησης και παλινδρόμησης. Κάποιοι από τους δημοφιλείς αλγορίθμους είναι η λογιστική παλινδρόμησης και τα νευρωνικά δίκτυα.

2.2.2 Μη-επιβλεπόμενη μάθηση

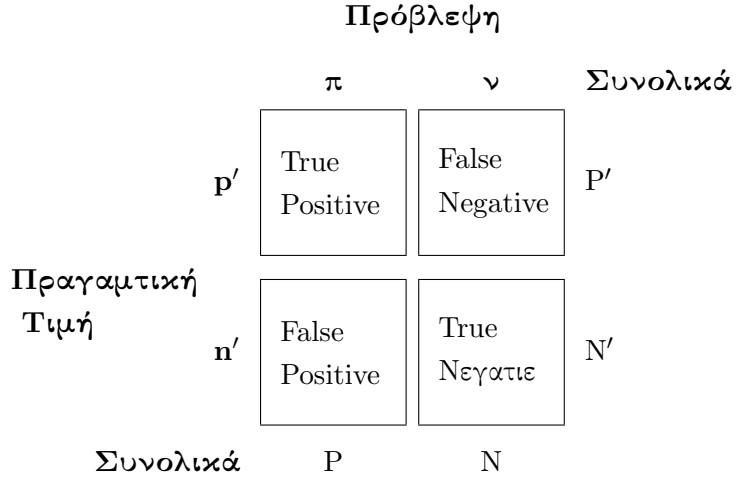
Τα δεδομένα εισόδου δεν έχουν δυαδικά χαρακτηριστικά και δεν είναι γνωστά τα αποτελέσματα. Ένα μοντέλο προετοιμάζεται μέσα από την εξαγωγή μιας παρούσας δομής στα δεδομένα εισόδου. Αυτό μπορεί να συμβεί εξάγοντας γενικούς κανόνες. Αυτό συνήθως συμβαίνει μέσο κάποιας μαθηματικής διαδικασίας που μειώνει συστηματικά την εφεδρεία, ή με οργάνωση των δεδομένων βάση ομοιότητας. Τέτοιου είδους προβλήματα είναι η συσταδοποίηση, η μείωση διάστασης και η εκπαίδευση μέσω κανόνων συσχέτισης. Τέτοιο αλγόριθμοι είναι το K-Means και το Principal Component Analysis (PCA).

2.2.3 Ημι-επιβλεπόμενη μάθηση

Τα δεδομένα εισόδου είμαι μια μίξη γνωστών και άγνωστων δυαδικών χαρακτηριστικών. Υπάρχει μια επιθυμητή πρόβλεψη τους προβλήματος, αλλά το μοντέλο πρέπει να μάθει τη δομή για να οργανώσει τα δεδομένα, αλλά και να κάνει τις τελικές προβλέψεις. Τέτοια προβλήματα είναι η ταξινόμηση και η παλινδρόμηση. Οι αλγόριθμοι που χρησιμοποιούνται είναι επέκταση άλλων ευέλικτων μεθόδων που κάνουν υποθέσεις για το μοντέλο χωρίς τα δυαδικά χαρακτηριστικά [3].

2.3 Μετρικές μηχανικής μάθησης

Για να γίνει αξιολόγηση της ταξινόμησης χρειάζεται να ληφθούν υπόψη κάποια κριτήρια και μετρικές. Ο ρυθμός ευστοχίας ή η μέση τιμή του λάθους αδυνατούν να μας περιγράψουν σαφώς τον ταξινομητή, οπότε εισάγεται η έννοια του confusion matrix. Σύμφωνα με τον πίνακα μετράμε τις εξής τιμές:



Σχήμα 2.1: Confusion Matrix

TP=πλήθος των σωστών προβλέψεων στο θετικό αποτέλεσμα
TN=πλήθος των σωστών προβλέψεων στο αρνητικό αποτέλεσμα
FN=πλήθος των λανθασμένων προβλέψεων στο θετικό αποτέλεσμα (αρνητική πρόβλεψη)
FP=πλήθος των λανθασμένων προβλέψεων στο αρνητικό αποτέλεσμα (θετική πρόβλεψη)

Με τις παραπάνω τιμές γίνεται να δομήσουμε τα κριτήρια ευστοχίας του συστήματος. Οι τέσσερις βασικοί άξονες της μέτρησης είναι το ποσοστό αναγνώρισης DR (Detection Rate), το ποσοστό λάθος συναγερμού FPR(False Positive Rate), το ποσοστό της ευστοχίας (Accuracy) και το F1 score που είναι ένας συνδυασμός μετρικών για να φανεί μια γενικότερη εικόνα της ακρίβειας του συστήματος.

$$DR = \frac{TP}{TP+FN},\ FPR = \frac{FP}{FP+TN},\ Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$
$$Precision = \frac{TP}{TP+FP},\ Recall = DR = \frac{TP}{TP+FN},\ F1 = 2\frac{precision \cdot recall}{precision+recall}$$

Ακόμη θα χρησιμοποιηθεί το ποσοστό αναγνώρισης του Bayes και η αντίστοιχή του άρνηση για να μας δώσουν μια πιθανοτική σκοπιά για την αναγνώριση απάτης και την αναγνώριση φυσιολογικής κατανάλωσης. Η $P(I)$ είναι η πιθανότητα να υπάρχει απάτη στα δεδομένα και αυτό σε πραγματικές συνθήκες δεν είναι εύκολο να υπολογιστεί με ακρίβεια. Το ενδεχόμενο A αντιστοιχεί στο συναγερμό που ενεργοποιείται στην αναγνώριση απάτης. Μπορεί στα συγκεκριμένα δεδομένα να οριστεί ως η πιθανότητα μια τυπική μέρα να βρεθεί απάτη στις μετρήσεις. Αυτό που έχει σημασία είναι και οι δύο πιθανότητες:

- $P(I|A)$−ότι ένας συναγερμός πραγματικά ενδεικνύει απάτη
- $P(\neg I|\neg A)$−ότι η απουσία του συναγερμού ενδεικνύει μη ικανοποιητικά δείγματα απάτης

να παραμείνουν όσο το δυνατόν μεγαλύτερες [2].
Μπορούμε να αντιστοιχίσουμε τα βασικά κριτήρια με τις πιθανότητες στο ποσοστό αναγνώρισης του Bayes.

$$P(A|I) = DR,\ P(A|\neg I) = FPR,\ P(\neg A|I) = 1 - P(A|I),\ P(\neg A|\neg I) = 1 - P(A|\neg I)$$
$$P(I|A) = \frac{P(I)P(A|I)}{P(I)P(A|I)+P(\neg I)\cdot P(A|\neg I)},\ P(\neg I|\neg A) = \frac{P(\neg I)\cdot P(\neg A|\neg I)}{P(\neg I)\cdot P(\neg A|\neg I)+P(I)\cdot P(\neg A|I)}$$
$$BDR = \frac{P(I)DR}{P(I)\cdot DR+P(\neg I)\cdot FPR}$$

Κεφάλαιο 3

Περιγραφή και οργάνωση δεδομένων

Απαραίτητη φάση της διαδικασίας εξόρυξης δεδομένων είναι η συλλογή και η προετοιμασία των δεδομένων. Η φάση κατανόησης των δεδομένων περιλαμβάνει τη συλλογή και εξερεύνησή τους. Ρίχνοντας μια πιο προσεκτική ματιά στα δεδομένα, καθίσταται εφικτός ο καθορισμός του πόσο καλά μπορούμε να αντιμετωπίσουμε το πρόβλημα. Η προσεκτική προετοιμασία δεδομένων μπορεί να βελτιώσει δραστικά τις πληροφορίες που μπορούν να εξαχθούν από την εξόρυξη δεδομένων[17].

3.1 Περιγραφή δεδομένων

Τα δεδομένα υπό εξερεύνηση αποτελούνται από καταναλώσεις έξυπνων μετρητών για σχεδόν 5.000 οικιακά νοικοκυριά και 600 επιχειρήσεις. Πιο συγκεκριμένα προέρχονται από την Commision for Energy Regulation (CER), η οποία αποτελεί την ανεξάρτητη αρχή για ενέργεια και νερό της Ιρλανδίας [7]. Οι ενδιαφερόμενοι πελάτες παρείχαν εθελοντικά τα δεδομένα των καταναλώσεων και ερωτηματολόγια για τις καταναλωτικές τους συνήθειες και τις υποδομές τους πράγμα που δίνει τη δυνατότητα να αναλυθούν διεξοδικά τα δεδομένα. Τα αντιπροσωπευτικά αυτά δείγματα συλλέχθηκαν ανώνυμα σε χρονικό παράθυρο σχεδόν 2 ετών, από το (2009-2011) και με συχνότητα λήψης 30 λεπτά για αυτό το διάστημα. Οι πληροφορίες των έξυπνων μετρητών είναι αποθηκευμένες σε έξι διαφορετικά αρχεία κειμένου (.txt), που καθένα έχει 24 εκατομμύρια καταχωρήσεις που αντιστοιχούν σε διάφορες μετρήσεις ενέργειας. Ο Πίνακας 3.1 αντιπροσωπεύει ένα μικρό δείγμα των αρχείων κειμένου, το οποίο αποτελείται από 3 στήλες. Η πρώτη στήλη αναπαριστά το ID του έξυπνου μετρητή που είναι ξεχωριστό για κάθε νοικοκυριό. Η δεύτερη στήλη δείχνει την ημερομηνία και την ώρα που σχετίζεται με τη συγκεκριμένη μέτρηση, ενώ η τρίτη στήλη αποτελεί την αντίστοιχη μέτρηση ενέργειας που καταναλώθηκε σε κιλοβατώρες (kWh)[28].

ID Μετρητή	Κωδικοποιημένη ημερομηνία/ώρα	Κατανάλωση ενέργειας kWh
1392	19503	0.140
1392	19504	0.138
...	...	...
1187	22028	1.367
1187	22029	1.425
1392	19940	0.234

Πίνακας 3.1: Στιγμιότυπα αρχείου δεδομένων

3.1.1 Επισκόπηση χρονοσειρών

Έχοντας διευκρινίσει, λοιπόν την προέλευση και τη δομή των δεδομένων αξίζει να γίνει μια αναλυτική επισκόπηση τους. Επειδή, καθίσταται αδιανόητη η μελέτη 4.500 ετήσιων καταναλώσεων, επιλέγονται ομάδες που να αντιπροσωπεύουν τον πληθυσμό. Για να μπορέσει να γίνει αυτό δημιουργήθηκαν 6 συστάδες (ομάδες) που να εκφράζουν είτε τη μορφή της καμπύλης είτε το ύψος της ημερήσιας κατανάλωσης. Με αυτό τον τρόπο ομαδοποιούνται τα δεδομένα και διευκολύνεται η διαδικασία παρατήρησης των χαρακτηριστικών 6 διαφορετικών ομάδων βάση 2 διαφορετικών κριτηρίων. Επιλέχθηκαν 6 συστάδες, καθώς έτσι επιτυγχάνεται ομοιομορφία στο πλήθος των μελών. Άμεσο αποτέλεσμα είναι οι συστάδες να αντιπροσωπεύουν κάποιο μετρήσιμο πλήθος μέλων. Στον Πίνακα 3.2 φαίνονται τα αποτελέσματα με τα μέλη κάθε συστάδας:

Συστάδα	Μέλη
1	1083
2	351
3	544
4	1078
5	420
6	1024

(α′) Συσταδοποίηση βάση των μορφών των χρονοσειρών

Συστάδα	Μέλη	Μέση κατανάλωση(kWh)
1	1680	26.75
2	163	77.35
3	721	42.67
4	49	330.51
5	1795	13.36
6	92	157.39

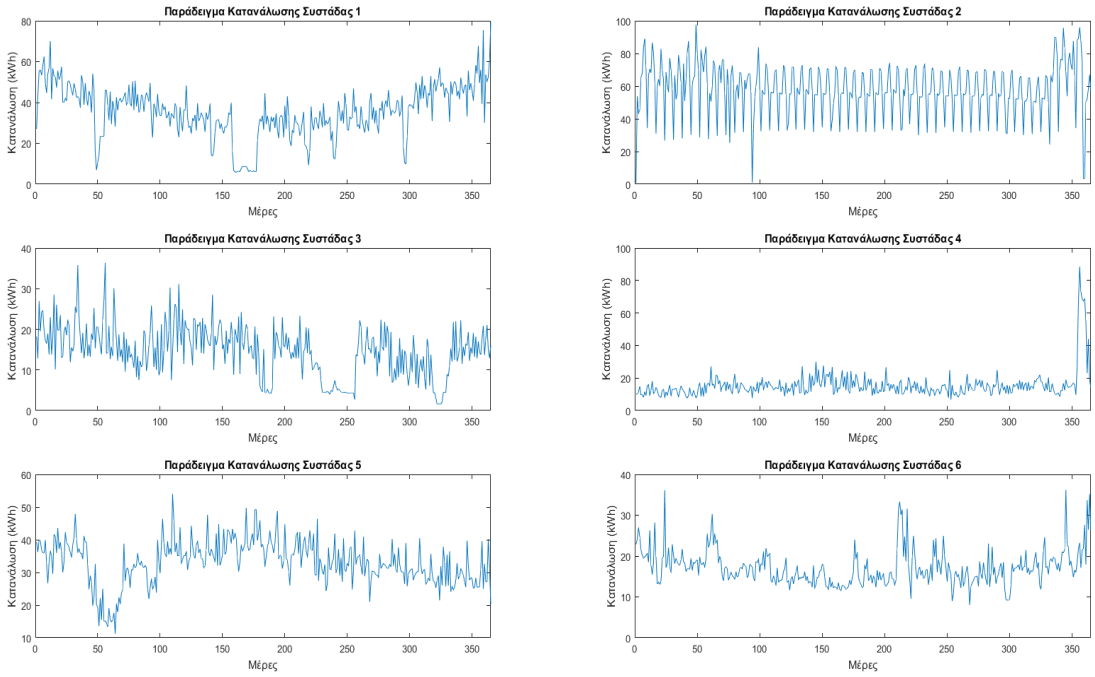
(β′) Συσταδοποίηση βάση του ύψους της κατανάλωση

Πίνακας 3.2: Ομαδοποιήσεις με 2 κριτήρια

Παρατηρείται, λοιπόν πως στον Πίνακα 3.2α′ η συσταδοποίηση βάση των μορφών των χρονοσειρών έχει 3 πολυμελείς συστάδες που συνοψίζουν τους 3.185 από τους 4.500 που επιλέχθηκαν για τη δοκιμή δημιουργώντας σχετικά ομοιόμορφες συστάδες. Παράλληλα, στον Πίνακα 3.2β′ η συσταδοποίηση βάση του ύψους της κατανάλωσης έχει 2 πολυμελείς συστάδες που συνοψίζουν τους 3.475 από τους 4.500 που επιλέχθηκαν και πρόκειται για απλούς οικιακούς πελάτες κρίνοντας από την μέση ημερήσια κατανάλωση κάθε συστάδας. Δεν μπορεί να παραληφθεί σε αυτό το σημείο το γεγονός πως υπάρχουν 2 ολιγομελείς ομάδες που απαριθμούν

αθροιστικά 141 μέλη και έχουν πολλαπλάσιες ημερήσιες καταναλώσεις από τους υπόλοιπους.

Για περαιτέρω εξερεύνηση των κριτηρίων ομαδοποίησης και των συστάδων δημιουργήθηκαν 2 σχήματα που αποτελούνται από παραδείγματα μελών κάθε συστάδας. Αναλυτικότερα στο Σχήμα 3.2 και στο Σχήμα 3.1 φαίνονται τυχαία επιλεγμένες καταναλώσεις για κάθε συστάδα. Έτσι δίνεται η δυνατότητα να αναλύσουμε τη μορφή 6 διαφορετικών ομάδων, αλλά και να παρατηρούμε τον διαχωρισμό των καταναλωτών και τις χρονοσειρές του με γνώμονα την ημερήσιά του κατανάλωση σε διάρκεια ενός έτους. Όπως φαίνεται παραπάνω υπάρχουν κάποιες



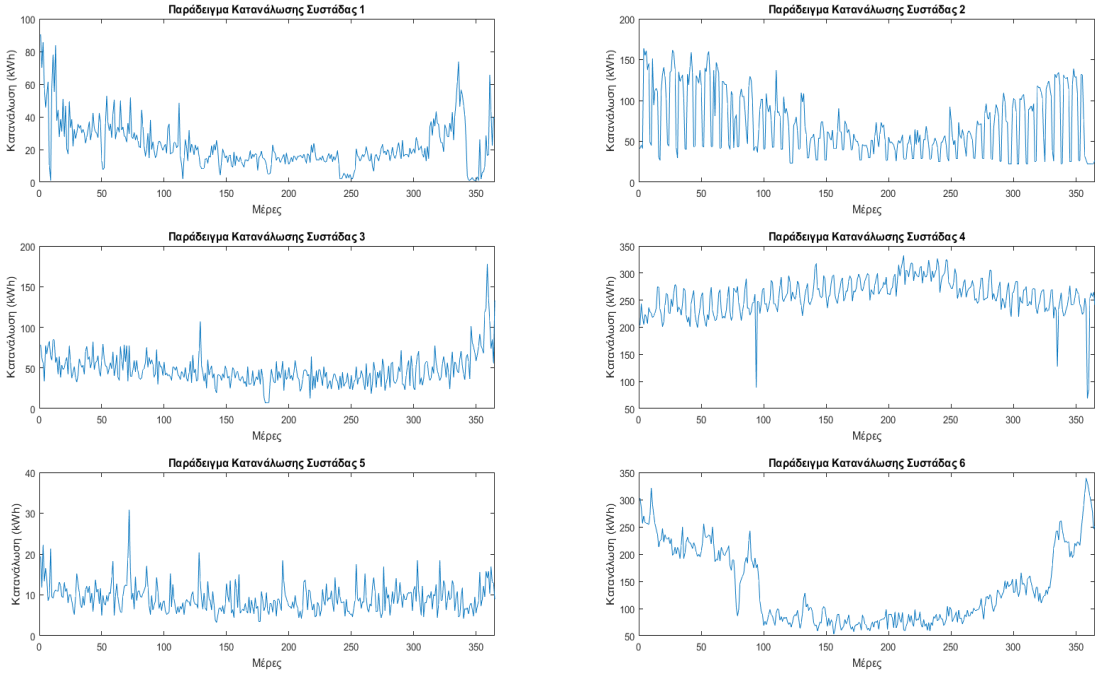
Σχήμα 3.1: Παραγείματα χρονοσειρών συσταδοποίησης βάση της μορφής των χρονοσειρών

αξιοσημείωτες ομοιότητες και διαφορές μεταξύ των μορφών των καμπυλών.

- Η συστάδα 1 φαίνεται πως στο ενδιάμεσο του έτους έχει μείωση της κατανάλωσης, ενώ κοντά στο χειμώνα οπού ξεκινά και τελειώνει η χρονοσειρά αύξησή της.
- Η συστάδα 2 έχει πολύ έντονες και συνεχείς διακυμάνσεις αλλά κρατά σχεδόν σταθερό μέσο όρο ανά της ημέρες, καθώς η διακύμναση είναι έντονη αλλά γύρω από μια νοητή γραμμή με ελάχιστη κλίση. Παράλληλα, είναι εμφανές πως στους χειμερινούς μήνες έχουμε αισθητή αύξηση της κατανάλωσης.
- Η συστάδα 3 εμφανίζει μια σχετικά ακανόνιστη, αλλά φθίνουσα εν γένει πορεία. Ειδικότερα υπάρχουν 2 σημαντικές βυθίσεις μία το Καλοκαίρι και μία το Φθινόπωρο.
- Η συστάδα 4 θυμίζει σημαντικά λευκό θόρυβο, καθώς δεν παρατηρείται έντονη απόκλιση από την μέση τιμή της καμπύλης, ενώ παράλληλα υπάρχει έντονος βαθμός τυχαιότητας στις διακυμάνσεις με την κατανάλωση να αυξάνεται μόνο τον τελευταίο μήνα του έτους.

- Η συστάδα 5 σημειώνει μια ύφεση στην κατανάλωση στο τέλος του χειμώνα που επιστρέφει στα κανονικά της επίπεδα μέσα στην άνοιξη. Κατά τα άλλα δεν φαίνεται να έχει κάποια άλλη έντονη κλίση.

- Η συστάδα 6 έχει εμφανώς αρχικά φθίνουσα τάση, ενώ μετά το καλοκαίρι ξεκινά βίαια και μετά ομαλότερα να αυξάνεται η ημερήσια κατανάλωση.



Σχήμα 3.2: Παραγείματα χρονοσειρών συσταδοποίησης βάση του ύψους της κατανάλωση

Στο παραπάνω Σχήμα εμφανίζονται τα παραδείγματα των συστάδων που δημιουργήθηκαν βάση του ύψους των ημερήσιων καταναλώσεων με τις εξής επισημάνσεις:

- Η συστάδα 1 που αποτελεί τη 2η μεγαλύτερη συστάδα έχει τιμές που κυμαίνονται γενικώς γύρω στις 27 kWh με 2 κύριες αλλαγές στη μονοτονία.

- Η συστάδα 2 εμπεριέχει καταναλωτές μικρομεσσαίων επιχειρήσεων με έντονες διακυμάνσεις και σχετικά μεγάλες καταναλώσεις.

- Η συστάδα 3 δεν έχει κάποιο ιδιαίτερο χαρακτηριστικό, καθώς εμφανίζει εξαιρετικές ομοιότητες με τη συστάδα 1 με μόνη διαφορά την μικρότερη κλίση στις μονοτονίες.

- Η συστάδα 4 εμφανίζει πολύ ξεχωριστή συμπεριφορά όντας καμπύλη μιας επιχείρησης με μεγάλες ενεργειακές απαιτήσεις που εμφανίζει τη μέγιστή της ζήτηση μετά τους καλοκαιρινούς μήνες.

- Η συστάδα 5 περιλαμβάνει ένα μεγάλο μέρος των οικιακών καταναλωτών που έχουν προσγειωμένες τιμές ημερήσιας κατανάλωσης, αλλά και μικρές διακυμάνσεις στη μονοτονία και στις μετρήσεις τους.

- Η συστάδα 6 περιγράφει καταναλωτές επιχειρήσεων με έντονη διακύμανση της κατανάλωσης ξεκινώντας με έντονη φθίνουσα πορεία και ακολουθώντας με ομαλή αύξουσα πορεία μετά το καλοκαίρι.

Ιστογράμματα Συχνοτήτων

Για την δημιουργία του ιστογράμματος απαιτείται ένα διάνυσμα δεδομένων που επιλέχθηκε να είναι ο μέσος όρος και η τυπική απόκλιση των ημερήσιων καταναλώσεων και η ετήσια κατανάλωση πελατών. Ο σκοπός ενός ιστογράμματος είναι να αναπαριστά γραφικά την κατανομή των δεδομένων με εξάρτηση από μια μεταβλητή. Το ιστόγραμμα χρησιμοποιείται ευρέως για να δώσει απάντηση στα παρακάτω ερωτήματα[6]:

1. Τι είδους κατανομή ακολουθεί ο πληθυσμός;

2. Που τοποθετούνται τα δεδομένα στον οριζόντια άξονα;

3. Πόσο αραιά είναι;

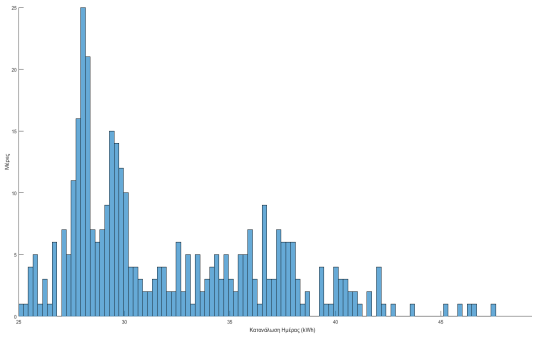
4. Υπάρχει εμφανής συμμετρία ή κυρτότητα;

5. Υπάρχουν ανωμαλίες στα δεδομένα;

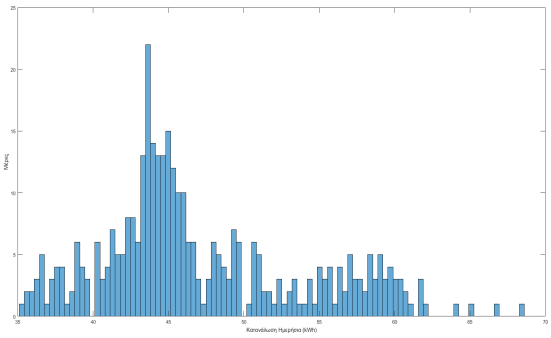
Εδώ ομαδοποιήθηκαν τα δεδομένα ως προς την ημερήσια κατανάλωση όλων των πελατών και την μέση ημερήσια κατανάλωση σε ένα έτος μετρούμενη σε kWh. Με αυτό τον τρόπο παρατηρούμε πως ομαδοποιούνται οι μέρες βάση της ημερήσιας κατανάλωσης και οι καταναλωτές βάση της ετήσιας συμπεριφοράς κατανάλωσης. Έτσι μπορούμε να παρατηρήσουμε ποσοτικά πόσες kWh καταναλώνονται σε μία μέρα, αλλά και πόσες kWh καταναλώνει κάθε πελάτης σε μία μέρα. Παράλληλα, είναι ιδιαίτερα χρήσιμη η παρατήρηση της απόκλισης των δεδομένων μεταξύ τους και του βαθμού συνέπειας τους παρακολουθώντας τα ιστογράμματα τυπικής απόκλισης.

Από τα σχήματα 3.3α΄ και 3.3β΄ φαίνεται πως και τα δύο ιστογράμματα έχουν θετική λοξότητα σε σχέση με το μέσο όρο του δείγματος. Παρόλα αυτά, υποθέτεται ότι η κατανομή του δείγματος προέρχεται από κανονική κατανομή πληθυσμού. Αντίστοιχα, τα ιστογράμματα των σχημάτων 3.3γ΄ και 3.3δ΄ δείχνουν επίσης θετική λοξότητα, αλλά με σημαντικά υψηλότερη κορυφή στο διάγραμμα, καθώς πρόκειται για πλήθος καταναλωτών.

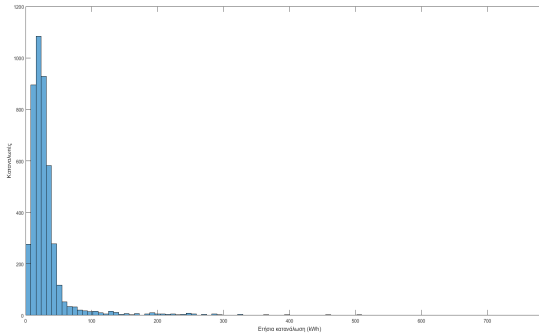
Σε αυτό το σημείο έχει νόημα να προσεγγιστούν οι ερωτήσεις που τέθηκαν παραπάνω. Γίνεται, λοιπόν σαφές πως τα δύο πρώτα ιστογράμματα έχουν μεγάλο εύρος και 2 κορυφές, ενώ τα επόμενα έχουν μικρό εύρος και μια μόνο κυριαρχούσα κορυφή. Παράλληλα, δεν επιβαρύνονται τα δεδομένα με ανωμαλίες ή ακραίες ομάδες με ιδιαίτερες καταναλωτικές συμπεριφορές. Παρόλα αυτά, τα τελευταία 2 σχήματα προδίδουν το γεγονός ύπαρξης καταναλωτών με μεγάλες ενεργειακές ανάγκες, αλλά λόγω του μικρού τους πλήθους δεν απαιτείται περαιτέρω εξερεύνηση προς τη συγκεκριμένη κατεύθυνση.



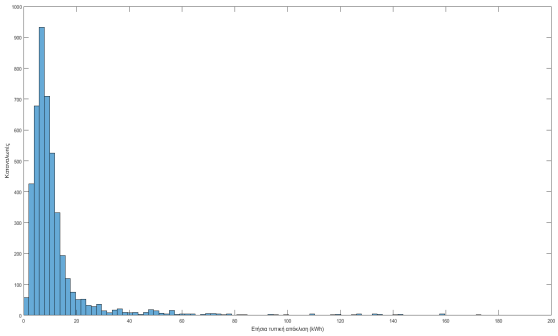
(α′) Μέσος όρος ημερήσιας κατανάλωσης



(β′) Τυπική απόκλιση ημερήσιας κατανάλωσης

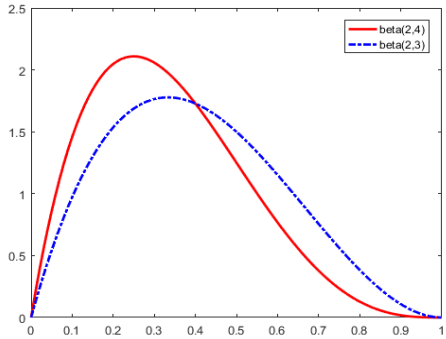


(γ′) Μέσος όρος ετήσιας κατανάλωσης

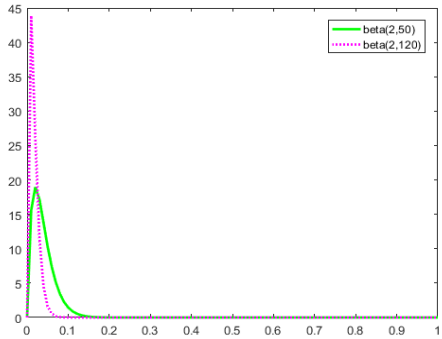


(δ′) Τυπική απόκλιση ετήσιας κατανάλωσης

Σχήμα 3.3: Ιστογράμματα για καταναλώσεις



(α′) Προσέγγιση Βήτα κατανομής στα σχήματα 3.3α′ και 3.3β′



(β′) Προσέγγιση Βήτα κατανομής στο σχήματα 3.3γ′ και 3.3δ′

Σχήμα 3.4: Εφαρμογή κατανομής Βήτα

Μέτρο	Σχήμα 3.3α΄	Σχήμα 3.3β΄	Σχήμα 3.3γ΄	Σχήμα 3.3δ΄
Μέσος Όρος	31.99	42.61	31.99	12.4111
Διάμεσος	29.82	40.24	23.85	8.34
Επικρατούσα Τιμή	24.71	30.44	23.50	9.95

Πίνακας 3.3: Ποσοτικά μέτρα περιγραφής ιστογραμμάτων

Γενικότερα, το είδος ασυμμετρίας που ακολουθούν τα ιστογράμματα παρουσιάζουν εξόγκωση προς τα αριστερά και έχουν μεγάλη ουρά προς τα δεξιά ($skewness > 0$). Για την προσέγγιση των κατανομών των ιστογραμμάτων χρησιμοποιήθηκε η κατανομή Βήτα, καθώς η συνάρτηση πυκνότητάς της είναι πολύ ευέλικτη στην αναπαράσταση μεγεθών και πιθανοτήτων [11]. Υπάρχουν δύο παράμετροι που θα εργαστούν ταυτοχρόνως για να καθορίσουν αν η κατανομή έχει επικρατούσα τιμή στο διάστημά της και αν αυτή είναι συμμετρική. Η κανονική Βήτα κατανομή παρέχει την πυκνότητα πιθανότητας της τιμής x στο διάστημα(0,1):

$$Beta(\alpha, \beta) : prob(x \| \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

όπου B είναι η βήτα συνάρτηση

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

Για την προσέγγιση των σχημάτων 3.3α΄ και 3.3β΄ χρησιμοποιήθηκαν οι κατανομές $Beta(2, 4)$ και $Beta(2, 3)$, ενώ τα σχήματα 3.3γ΄ και 3.3δ΄ αντιστοιχίζονται με τις κατανομές $Beta(2, 50)$ και $Beta(2, 120)$ δημιουργώντας σε κάθε παράδειγμα μια επικρατούσα τιμή. Παρακάτω μπορεί να φανεί η αναπαράστασή τους στο Σχήμα 3.4.

Ένα ακόμη απαραίτητο στάδιο στη μελέτη ιστογραμμάτων θετικής λοξότητας είναι η ποσοτικοποίηση μετρικών που να συνοψίζουν τα δεδομένα. Για αυτό το στάδιο επιλέχθηκαν ο μέσος όρος, ο διάμεσος και η επικρατούσα τιμή. Τα αποτελέσματα για κάθε ιστόγραμμα μπορούν να φανούν στον Πίνακα 3.3.

3.1.2 Μοντελοποίηση εποχιακών δεικτών

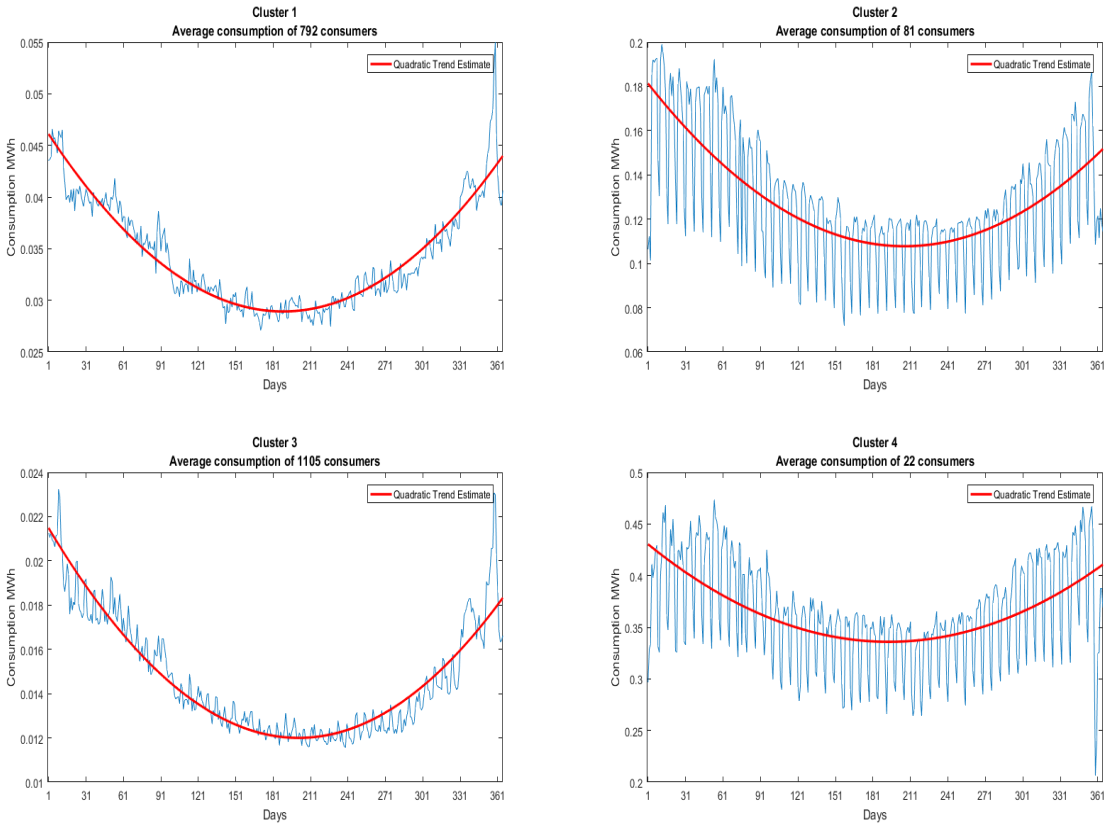
Για βαθύτερη κατανόηση των χρονοσειρών γίνεται εκτίμηση της εποχιακής και μη εποχιακής καταναλωτικής τάσης με τη χρήση παραμετρικών μοντέλων. Με αυτό τον τρόπο θα καταστεί δυνατή η παρατήρηση της επαναληψιμότητας και των μορφών των καταναλώσεων. Για να γίνει αυτό χρησιμοποιείται αρχικά ο αλγόριθμος K-Means για την ομαδοποίηση των καταναλωτών σε τέσσερις συστάδες βάση του ετήσιου μέσου όρου καθενός. Στη συνέχεια δημιουργείται ένα προφίλ κατανάλωσης για κάθε συστάδα βρίσκοντας το μέσο ημερήσιο όρο κατανάλωσης. Χρειάστηκαν 2000 καταναλωτές για αυτή την ανάλυση με περισσότερους 1800 να ομαδοποιούνται σε δύο ομάδες υποδεικνύοντας προφίλ οικιακών καταναλωτών.

Ανάλυση Παλινδρόμησης

Σκοπός, λοιπόν αυτού του μέρους είναι να γίνει στατιστική μελέτη του πολυωνυμικού μοντέλου στα δεδομένα μας και να δούμε αν οι χρονοσειρές κάθε συστάδας μπορούν να

περιγραφούν με πολυώνυμο δευτέρου βαθμού. [12]

$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2$$



Σχήμα 3.5: Εφαρμογή πολυωνύμου δευτέρου βαθμού

Όπως φαίνεται στο Σχήμα 3.5 οι συστάδες μπορούν να χαρακτηριστούν από μια παραβολική καμπύλη με θετικό συντελεστή μεγιστοβάθμιου όρου.

- Η συστάδα 1 αποτελείται από 792 καταναλωτές και έχει η παραβολική καμπύλη τάσης λαμβάνει ελάχιστη τιμή την 189η μέρα του έτους.
- Η συστάδα 2 αποτελείται από 81 καταναλωτές και έχει η παραβολική καμπύλη τάσης λαμβάνει ελάχιστη τιμή την 206η μέρα του έτους.
- Η συστάδα 3 αποτελείται από 81 καταναλωτές και έχει η παραβολική καμπύλη τάσης λαμβάνει ελάχιστη τιμή την 201η μέρα του έτους.
- Η συστάδα 4 αποτελείται από 81 καταναλωτές και έχει η παραβολική καμπύλη τάσης λαμβάνει ελάχιστη τιμή την 194η μέρα του έτους.

Εύκολα, λοιπόν, βγάνει το συμπέρασμα πως οι οικιακοί καταναλωτές έχουν την τάση να έχουν πιο ομοιόμορφα κατανεμημένα την παραβολική καμπύλη, ενώ οι επιχειρήσεις έχουν

μεγαλύτερο βαθμό τυχαιότητας και λιγότερο συμμετρική καμπύλη ως προς το ελάχιστο σημείο της.

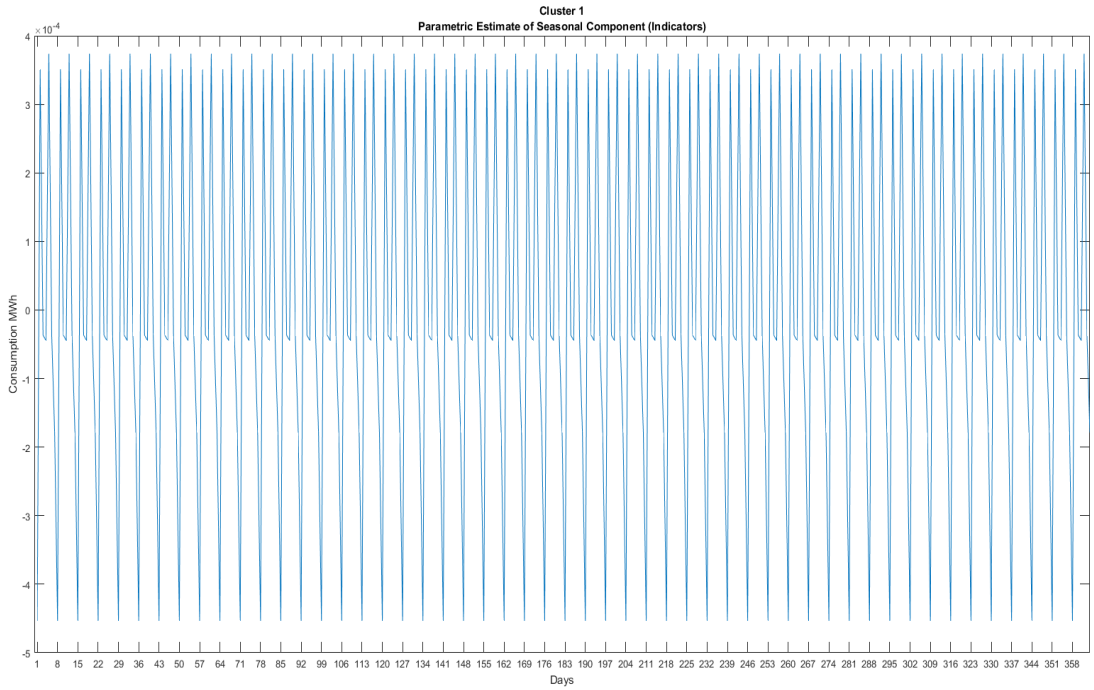
Εκτίμηση εποχιακών δεικτών

Αρχικά για την εκτίμηση των εποχιακών δεικτών απαιτείται η αφαίρεση του πολυώνυμου δευτέρου βαθμού από τις χρονοσειρές των ομάδων.[8] Δεδομένης της μικρής διάρκειας των καταναλώσεων (1 έτος) καθίσταται αδύνατη η εξαγωγή εποχιακών δεικτών ανά μήνα έτους ή ανά εποχή έτους. Για αυτό το λόγο οι εποχιακοί δείκτες μεταφέρθηκαν ανά ημέρα της εβδομάδας ή ανά ημέρα του μήνα. Για την πρώτη περίπτωση οι δείκτες αναφέρονται στις ημέρες κάθε εβδομάδας, ενώ για την δεύτερη αναφέρονται στις ημέρες κάθε μήνα δημιουργώντας 7 ή 30 δείκτες αντίστοιχα. Για την εβδομαδιαία εποχιακότητα έχω τις παρακάτω καμπύλες για κάθε ομάδα.

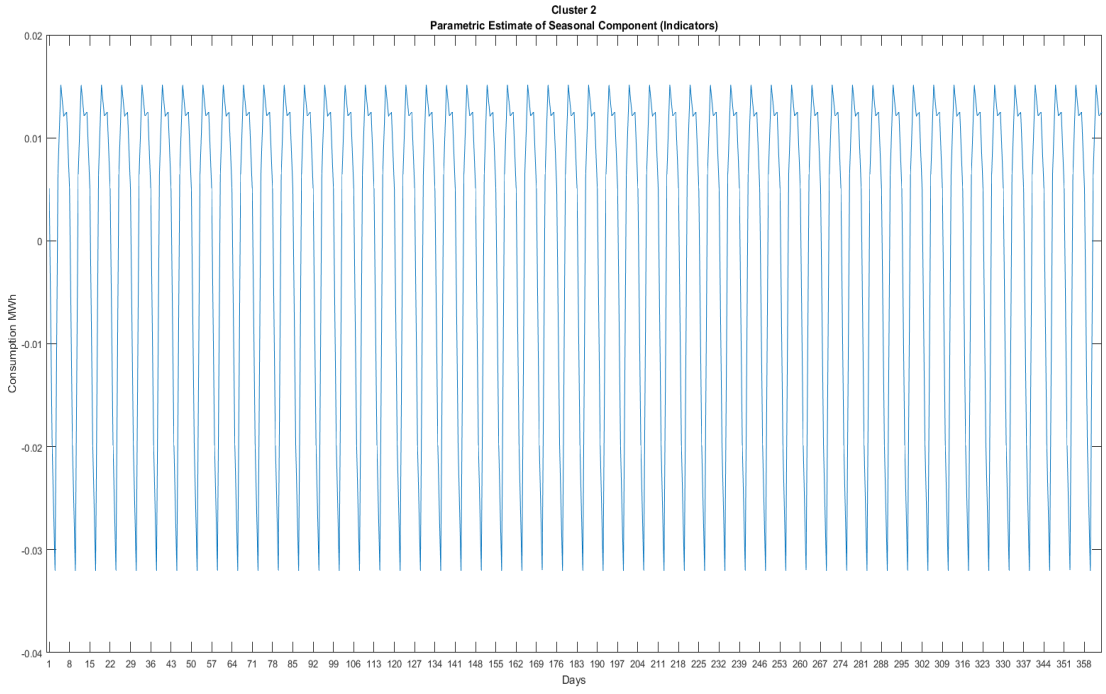
Εκτίμηση με διαστήματα ημέρας ανά εβδομάδα

Από την εβδομαδιαία εποχιακότητα λοιπόν εύκολα κάποιος αντιλαμβάνεται πως ανάλογα με τον τύπο των καταναλωτών οι μέρες που έχουμε μέγιστη και ελάχιστη κατανάλωση διαφέρουν ριζικά. Η πρώτη μέρα του έτους για το έτος που μελετάμε είναι Πέμπτη. Ειδικότερα:

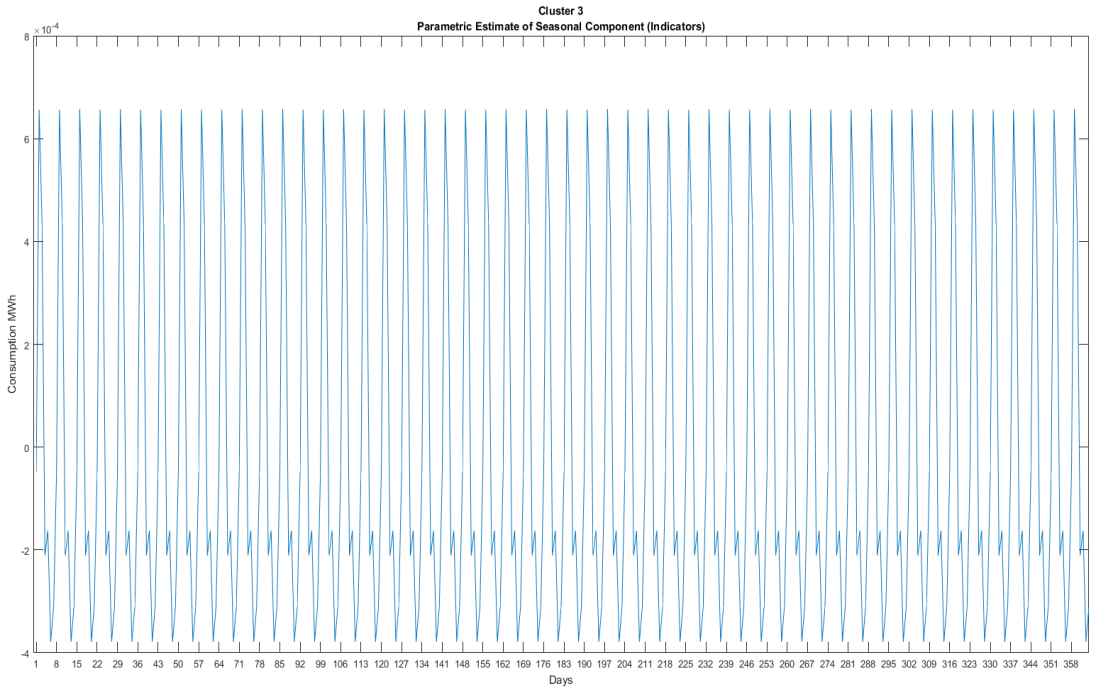
- Για τους καταναλωτές συστάδας 1 (οικιακοί καταναλωτές) έχουμε ελάχιστες καταναλώσεις τις Πέμπτες.
- Για τους καταναλωτές συστάδας 2 (επιχειρήσεις) έχουμε ελάχιστες καταναλώσεις τα Σάββατα.
- Για τους καταναλωτές συστάδας 3 (οικιακοί καταναλωτές) έχουμε ελάχιστες καταναλώσεις τις Τρίτες.
- Για τους καταναλωτές συστάδας 4 (επιχειρήσεις) έχουμε ελάχιστες καταναλώσεις τα Σάββατα.



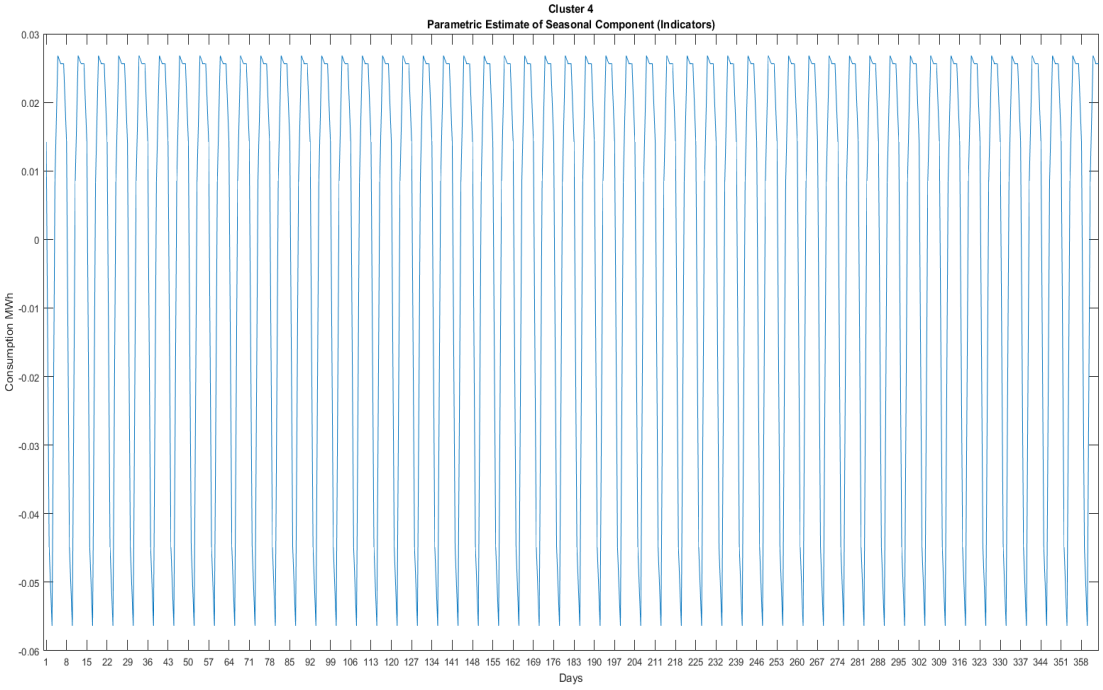
Σχήμα 3.6: Εβδομαδιαία εποχιακότητα ομάδας 1



Σχήμα 3.7: Εβδομαδιαία εποχιακότητα ομάδας 2



Σχήμα 3.8: Εβδομαδιαία εποχιακότητα ομάδας 3



Σχήμα 3.9: Εβδομαδιαία εποχιακότητα ομάδας 4

Εκτίμηση σε διαστήματα ημέρας ανά μήνα

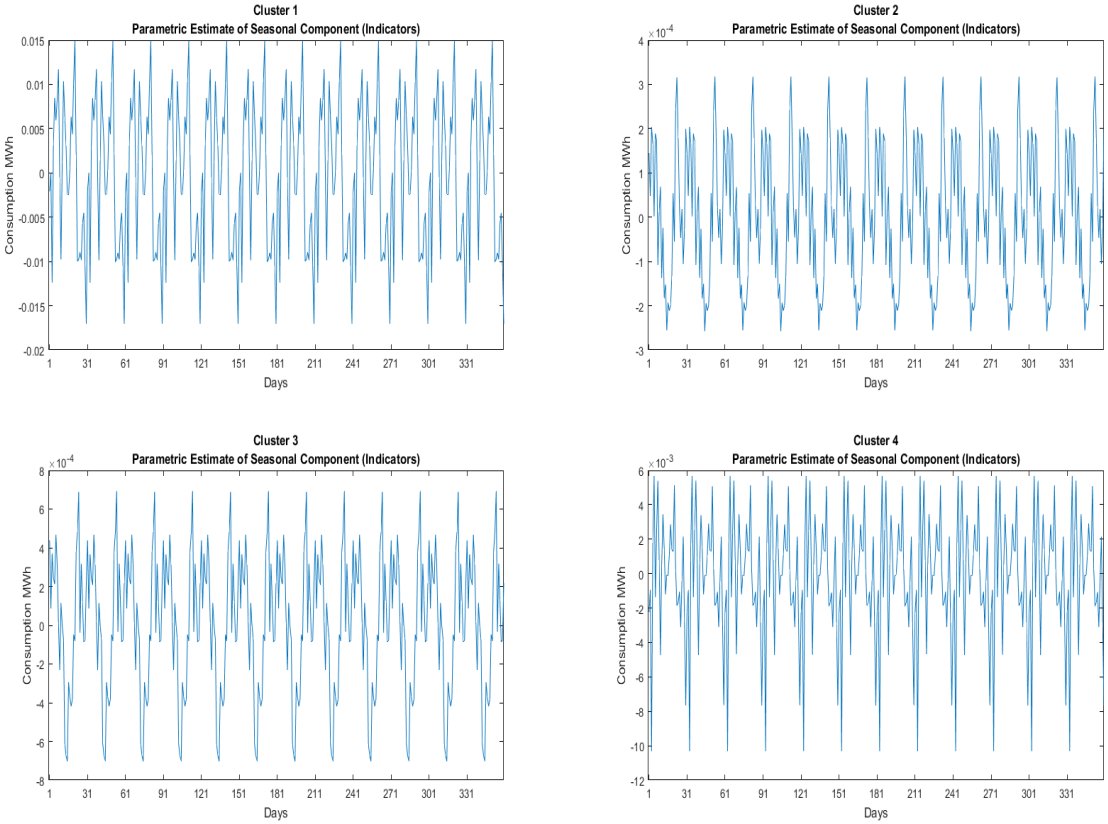
Το διάστημα ενός μήνα αφήνει μεγαλύτερα περιθώριο εποπτείας της χρονοσειράς, ενώ ταυτόχρονα δημιουργεί αποτελέσματα με μεγαλύτερη συνοχή. Από την άλλη πλευρά οι 12 μήνες του έτους δεν μπορούν να εξάγουν πολύ ασφαλή δεδομένα αν συγκριθούν με τις 52 εβδομάδες.

Από την μηνιαία εποχιακότητα γίνεται εύκολα αντιληπτό πως ανάλογα με τον τύπο των καταναλωτών οι μέρες που έχουμε μέγιστη και ελάχιστη κατανάλωση διαφέρουν ριζικά. Ειδικότερα:

- Για τους καταναλωτές συστάδας 1 (επιχειρήσεις) έχουμε ελάχιστες καταναλώσεις στις 30 του μηνός.
- Για τους καταναλωτές συστάδας 2 (οικιακοί καταναλωτές) έχουμε ελάχιστες καταναλώσεις στις 15 του μηνός.
- Για τους καταναλωτές συστάδας 3 (οικιακοί καταναλωτές) έχουμε ελάχιστες καταναλώσεις στις 15 του μηνός.
- Για τους καταναλωτές συστάδας 4 (επιχειρήσεις) έχουμε ελάχιστες καταναλώσεις στις 3 του μηνός.

Αφαίρεση εποχιακών δεικτών

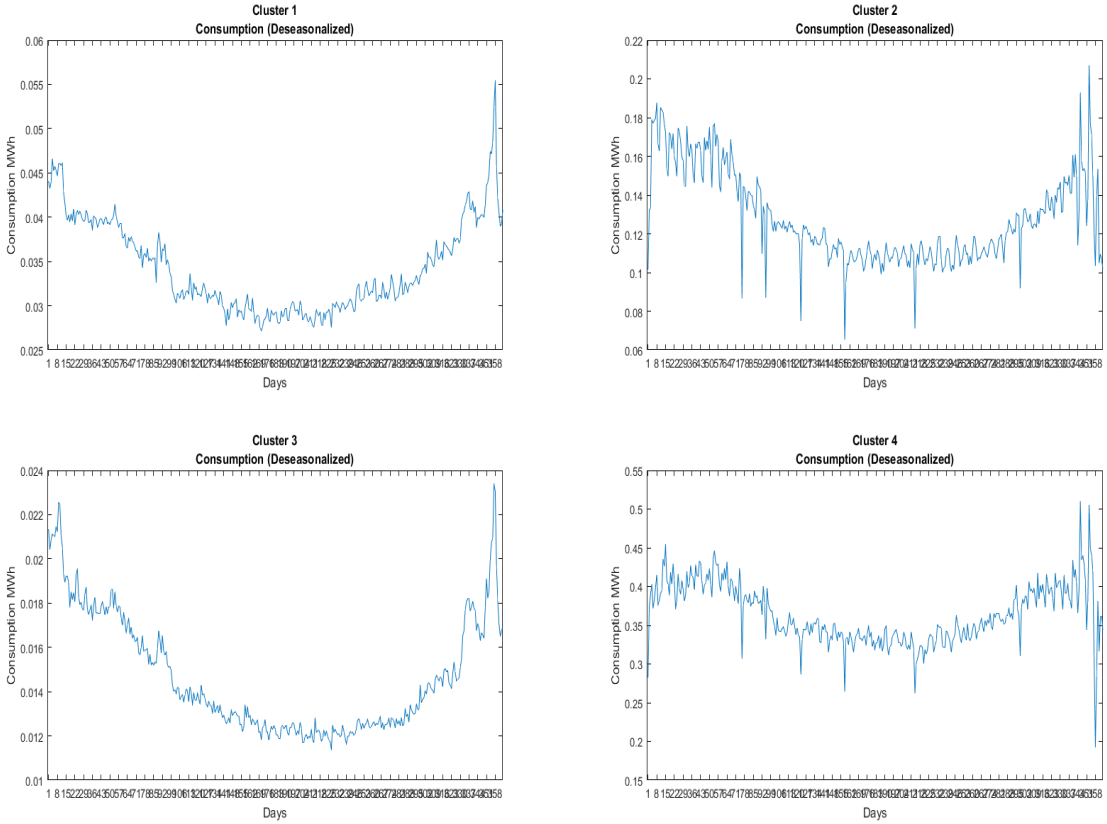
Σε αυτό το σημείο είναι σημαντικό να παρατηρηθεί η κατανάλωση χωρίς τους εποχιακούς δείκτες. Με αυτό τον τρόπο καθίσταται ευκολότερη η θεώρηση της μορφής των κυματομορφών και η σύγκρισή τους με τις αρχικές καταναλώσεις του πρώτου μέρους. Αφαιρώντας τα εποχιακά χαρακτηριστικά οι καμπύλες πλησιάζουν περισσότερο στην παραβολική συνάρτηση. Έτσι η καταναλωτική τους τάση χωρίς τους εποχιακούς δείκτες γίνεται πιο έντονη και ευδιάκριτη.



Σχήμα 3.10: Μηνιαία εποχιακότητα

Εκτίμηση ακανόνιστης συνιστώσας

Τέλος έχει ενδιαφέρουν να δούμε το βαθμό της τυχαιότητας που έχουμε στις καταναλώσεις των συστάδων που δημιουργήθηκαν. Αυτό επιτυγχάνεται αφαιρώντας την εποχιακή χρονοσειρά και την καταναλωτική τάση της αρχικής χρονοσειράς. Με αυτό τον τρόπο γίνεται σαφές ότι παρόλο την εποχιακότητα και την τάση οι χρονοσειρές έχουν αισθητό τυχαίο παράγοντα. Η αφαίρεση δημιουργεί αλλαγές στο επίπεδο της χεονοσειράς, σταθεροποιώντας έτσι το μέσο όρο της. Γίνεται αντιληπτό πως έχουν μη προβλέψιμα πρότυπα τουλάχιστον με δεδομένα διάρκειας ενός έτους. Τέτοιου τύπου δεδομένα λέγενται στατικές χρονοσειρές.[9]

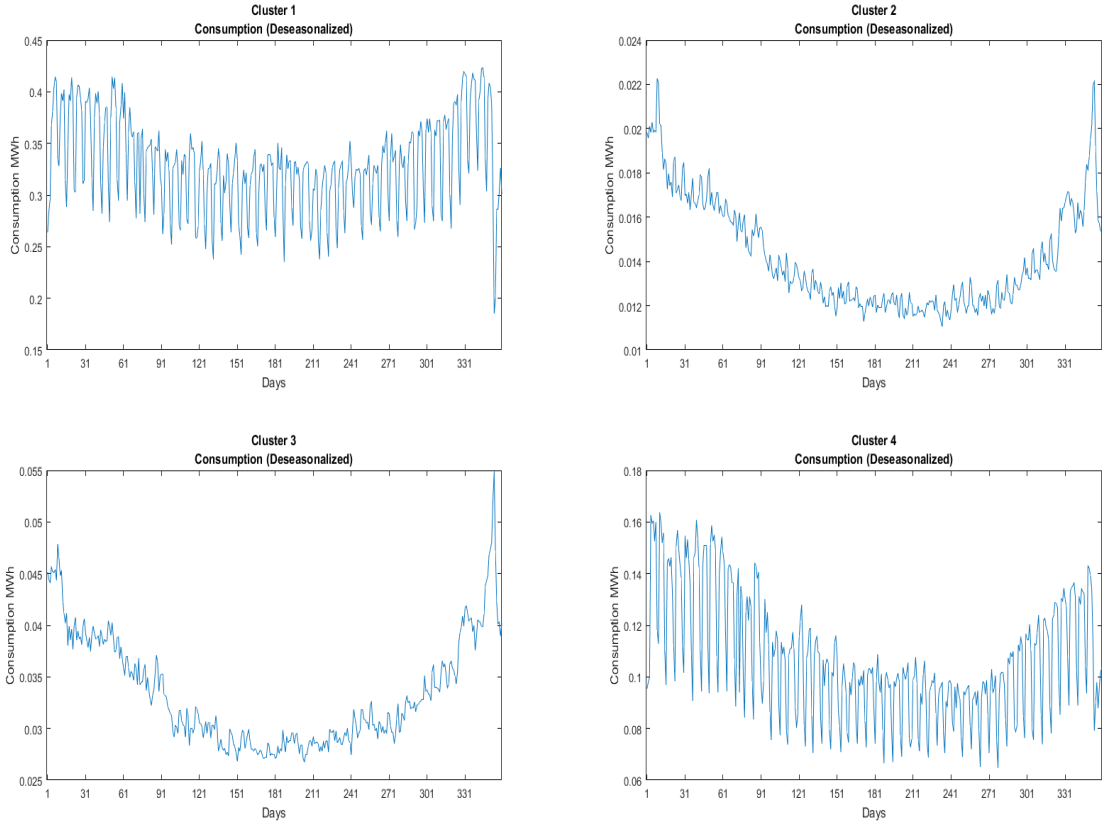


Σχήμα 3.11: Κατανάλωση χωρίς εποχιακούς δείκτες ανά εβδομάδα

Εξερεύνηση ημερών με χαμηλές καταναλώσεις

Για να αντληθούν περαιτέρω χαρακτηριστικά των χρονοσειρών χρειάστηκε η υλοποίηση αλγορίθμου με διπλή συσταδοποίηση. Σύμφωνα με τον αλγόριθμο πρώτα συσταδοποιούνται οι καταναλωτές με βάση την ημερήσια κατανάλωση, εν συνεχεία για κάθε συστάδα δημιουργείται νέα ομαδοποίηση με βάση την ομοιότητα κάθε ημερήσιας κατανάλωσης. Με αυτό τον τρόπο μπορεί να παρατηρηθεί ποιες μέρες όμοιων καταναλωτών έχουν παρόμοιες καταναλώσεις. Καθίσταται έτσι εφικτό, να φιλτράρουμε από τα δεδομένα μας μέρες με χαμηλή κατανάλωση που γνωρίζουμε πως θα δυσκόλευαν το πρόβλημα της ταξινόμησης σε αληθή και αλλοιωμένα δεδομένα.

Τα αποτελέσματα του αλγορίθμου έδειξαν πως μόνο τα Σάββατα μιας συστάδας εμφανίζουν έντονη ομοιότητα οικιακών καταναλώσεων. Οι Κυριακές κατά κύριο λόγο συσταδοποιούνται με την υπόλοιπη εβδομάδα δημιουργώντας την εβδομαδιαία τάση, γεγονός που δείχνει πως για τους περισσότερους καταναλωτές η Κυριακή είναι εργάσιμη ημέρα. Παράλληλα, παρατηρείται πως ανά περιόδους οι καταναλώσεις δημιουργούν νέες συστάδες αφήνοντας μόνο τα Σάββατα να σπάνε την συνεχόμενη συσταδοποίηση. Στον Πίνακα 3.4 φαίνεται πως ακόμη και στα Σάββατα δεν έχουμε απολύτως γεμάτες συστάδες.



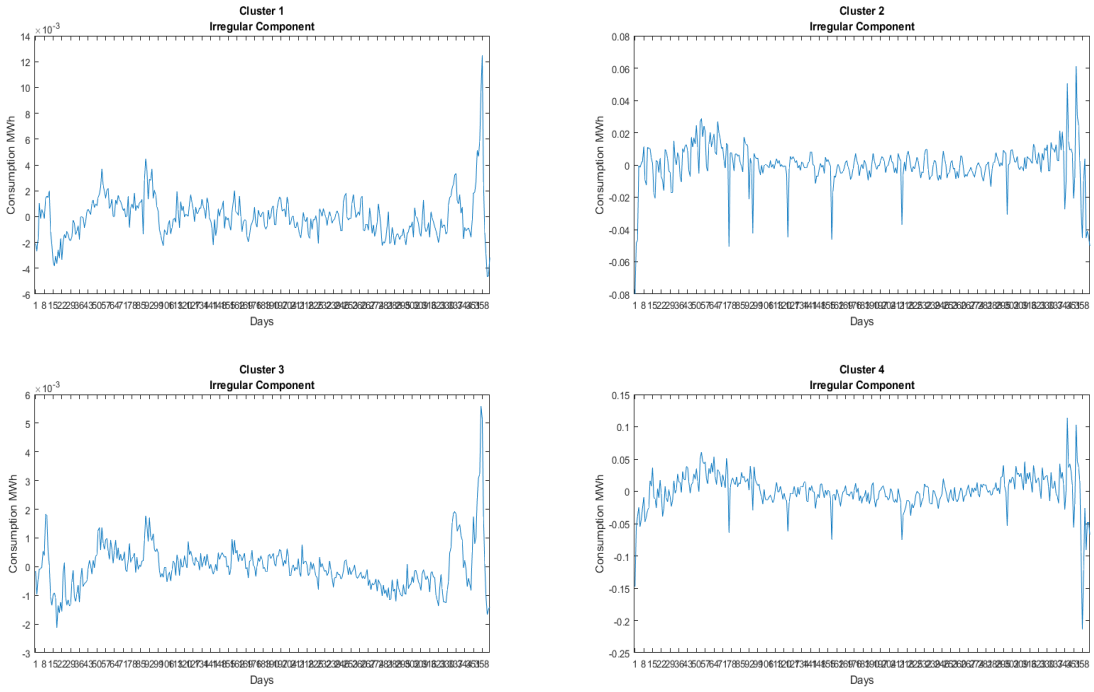
Σχήμα 3.12: Κατανάλωση χωρίς εποχιακούς δείκτες ανά μήνα

Συστάδες Καταναλωτών				
Συστάδες Σαββάτου	Συστάδα 1	Συστάδα 2	Συστάδα 3	Συστάδα 4
Συστάδα 1	0	24	30	19
Συστάδα 2	9	11	0	15
Συστάδα 3	0	9	0	0
Συστάδα 4	42	0	0	0
Συστάδα 5	0	2	0	0
Συστάδα 6	0	4	0	7
Συστάδα 7	0	1	21	10

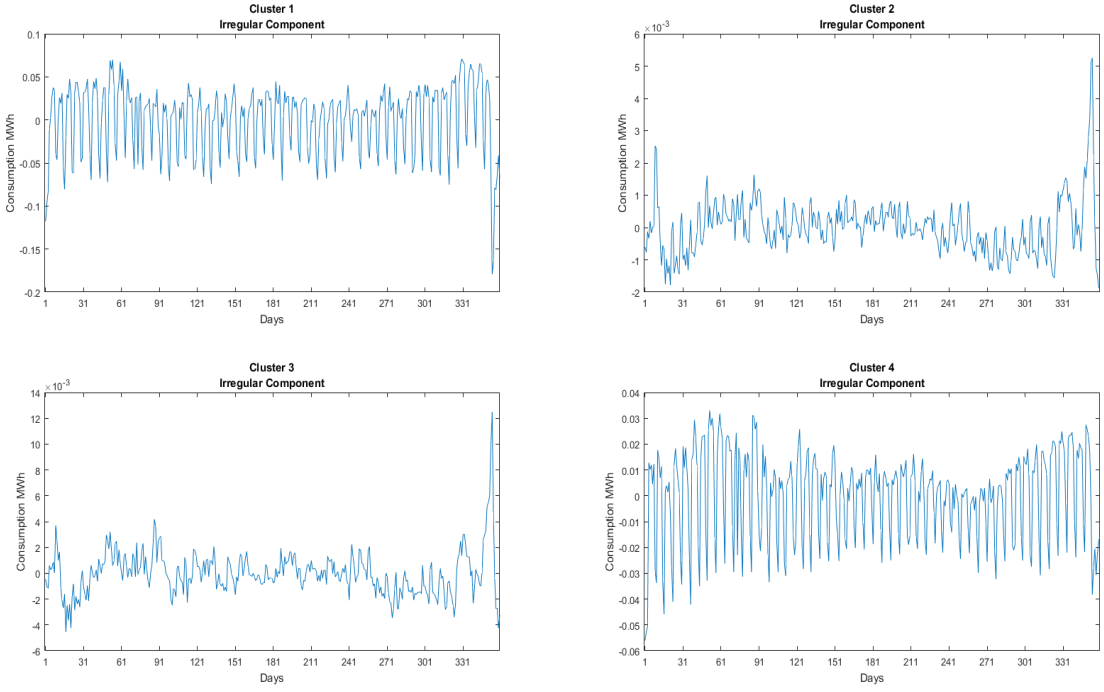
Πίνακας 3.4: Έλεγχος συσταδοποίησης Σαββάτου

Παρατηρήσεις

Τα εμφανή χαρακτηριστικά εποχιακότητας και η εφαρμογή πολυωνύμου δευτέρου βαθμού στις χρονοσειρές θέτει καλό υποψήφιο τα μοντέλα πρόβλεψης χρονοσειρών. Με ένα τέτοιο σύστημα θα δημιουργείται μια πρόβλεψη κατανάλωσης από έμπιστους καταναλωτές για κάποιο χρονικό διάστημα. Εν συνεχεία θα αλλοιώνονται τα χαρακτηριστικά κάποιου μέρους των



Σχήμα 3.13: Εκτίμηση ακανόνιστης συνιστώσας με εβδομαδιαία εποχιακότητα



Σχήμα 3.14: Εκτίμηση ακανόνιστης συνιστώσας με μηνιαία εποχιακότητα

καταναλωτών και θα ελέγχεται αν ο αλγόριθμος μπορεί να διαχωρίσει τις αλλοιωμένες τιμές από αυτές που προέβλεψε.

3.2 Προεπεξεργασία και καθάρισμα δεδομένων

Πριν τις αρχικές δοκιμές των ταξινομητών απαιτείται η επιλογή της τελικής μορφής των δεδομένων που θα χρησιμοποιηθούν στο υπόλοιπο σύστημα. Για να μπορέσουν τα δεδομένα να είναι κατανοητά και ξεκάθαρα χρειάζεται να οργανωθούν αν ID μετρητή που είναι ξεχωριστός για κάθε πελάτη και εν συνεχεία να οργανωθούν τα δεδομένα σε συνεχείς χρονικές περιόδους. Λαμβάνοντας υπόψη ότι τα δεδομένα είχαν χρονικό παράθυρο λιγότερο από 2 έτη, επιλέχθηκε πως κάθε καταναλωτής θα πρέπει να έχει ένα γεμάτο έτος μετρήσεων για να μπει σε οποιαδήποτε δοκιμή.

Έτσι λοιπόν όποιος καταναλωτής έχει πλήρη δεδομένα για όλα τα ημίωρα του έτους από την πρώτη Ιανουαρίου μέχρι και το Δεκέμβριο του 2010 περνάει στο τελικό σύνολο δεδομένων. Σε αυτό το στάδιο κάθε καταναλωτής περιγράφεται από ένα διάνυσμα με 17.520 μετρήσεις. Δυστυχώς, ακόμη και για τα σημερινά δεδομένα ένας πίνακας αποτελούμενος από τόσες μετρήσεις για κάθε καταναλωτή γίνεται δύσκολος στη διαχείριση και χρονοβόρος στην επεξεργασία. Για να δοθεί λύση στο πρόβλημα αυτό δημιουργήθηκαν δύο είδη πινάκων.

Το πρώτο είδος πίνακα περιέχει πολύ πληροφορία ώστε να γίνονται λεπτομερείς επεξεργασίες των δεδομένων, αλλά είναι δύσχρηστος στις δοκιμές, καθώς απαιτεί μεγάλη υπολογιστική δύναμη για να συμπεριληφθεί σε περίπλοκες πράξεις πινάκων. Ειδικότερα, κάθε καταναλωτής περιγράφεται από ένα πίνακα που περιέχει τις μετρήσεις του ανά ημέρα σε ημίωρα ή ανά μήνα σε ώρες ή ανά εβδομάδα σε ώρες κοκ. Το δεύτερο είδος πίνακα είναι λιγότερο περιεκτικό, αλλά έχει τη δυνατότητα να χειρίζεται πολύ πιο εύκολα και γρήγορα από τους αλγορίθμους που χρησιμοποιήθηκαν. Πιο συγκεκριμένα, κάθε καταναλωτής έχει ένα διάνυσμα που περιέχει τις τιμές κατανάλωσης ενός έτους σε ώρες, ημίωρα, μέρες, εβδομάδες ή και μήνες. Για να δοθεί ένα πρακτικό παράδειγμα των δύο ειδών πινάκων ένας περιγραφικός πίνακας για 2.000 καταναλωτές με ανάλυση σε ώρες ανά μέρα έχει 730.000 γραμμές και 24 στήλες, ενώ ο αντίστοιχος πίνακας για υπολογισμούς έχει 2.000 γραμμές και 24 στήλες. Ουσιαστικά, ο περιγραφικός πίνακας είναι 365 φορές μεγαλύτερος και κρίνεται ακατάλληλος για περίπλοκες πράξεις πινάκων.

Οι καρποί της προεπεξεργασίας και του καθαρίσματος των δεδομένων είναι ένα διάνυσμα με τα ID των μετρητών, ένας πίνακας με ετήσια διανύσματα κατανάλωσης και ένας πίνακας 3 διαστάσεων που περιγράφει αναλυτικά τη καταναλωτική συμπεριφορά των πελατών. Το διάνυσμα με τα ID των μετρητών χρησιμοποιείται για την αντιστοίχιση των πελατών με τις ετήσιες καταναλώσεις τους. Ο πίνακας διανυσμάτων κατανάλωσης χρησιμοποιείται για πολύπλοκες και επίπονες πράξεις, ενώ ο αναλυτικός πίνακας για λεπτομερή μελέτη και μικρή επεξεργασία.

3.3 Προσομοίωση απάτης

Δεδομένου ότι οι μετρήσεις που συλλέχθηκαν ήταν από αξιόπιστους καταναλωτές θα πρέπει να μοντελοποιηθεί η συμπεριφορά με μη τεχνικές απώλειες. Σε αυτό τον τόμο προσεγγίζεται η περίπτωση της φυσικής επίθεσης, οπού παράνομοι καταναλωτές αλλοιώνουν το σύστημα μέτρησης για να αναφέρει μικρότερα ποσά. Αυτό μπορεί να συμβεί με χρήση μαγνήτη που παρεμβαίνει στο μετρητή. Παράλληλα, επιθέσεις στο σύστημα μέτρησης μπορούν να επιτευχθούν και με ηλεκτρονικά μέσα (Cyber attacks), αλλοιώνοντας τις τιμές, συνοψίζοντας στους τρόπους επίθεσης στα δεδομένα. Σε κάθε περίπτωση ο καταναλωτής εισάγεται μια μέρα στην ρευματοκλοπή και ανάλογα με το σύστημα του αλλοιώνονται όλα ή μερικά από τα δεδομένα του με σταθερό ή μεταβλητό ρυθμό. Όπως γίνεται αντιληπτό μπορεί να εισαχθεί μεγάλος βαθμός τυχαιότητας στην ρευματοκλοπή. Στην περίπτωση της φυσικής επίθεσης, είναι ευκολότερος ο προσδιορισμός της απάτης και σχετικά σταθερός ο βαθμός αλλοίωσης των δεδομένων, ενώ στις επιθέσεις με ηλεκτρονικά μέσα μπορεί να εισαχθούν πολλοί εξωτερικοί και άγνωστοι παράγοντες, που μπορεί να έχουν στόχο τη απόκρυψη και ελαχιστοποίηση της κλοπής έτσι ώστε να μην γίνεται εύκολα αντιληπτό.

3.3.1 Τύποι απάτης

Έτσι, μοντελοποιήθηκαν 4 συμπεριφορές που καθεμία εισάγει έναν διαφορετικό παράγοντα [21]. Για όλους τους τύπους απάτης θεωρείται ότι μια μέρα ο καταναλωτής εισάγεται στην ρευματοκλοπή και από εκείνη τη μέρα χρησιμοποιεί με διαφορετικούς ρυθμούς το σύστημα αλλοίωσης. Παράλληλα, για την ένταση της κλοπής χρησιμοποιούνται διαφορετικές κατανομές για να επιλέγεται από αυτές η ένταση της επίθεσης. Η κατανομή Βήτα με παραμέτρους 6 και 3 (Σχήμα 3.16) θεωρήθηκε η πιο ρεαλιστική, καθώς έχει κορυφή στο 0.7 και σχετικά μεγάλο εύρος τιμών, εισάγοντας βαθμό τυχαιότητας, αλλά με κατεύθυνση τις έντονες επιθέσεις.

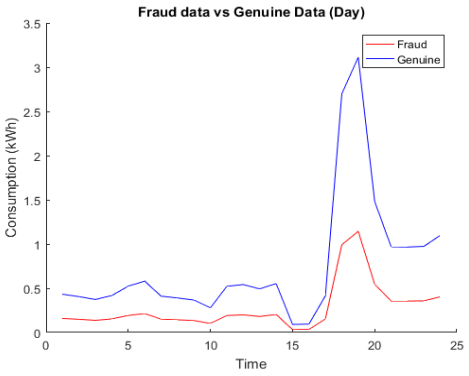
1. *Απώλειες Τύπου 1* Μοντελοποιεί τον καταναλωτή που θα χρησιμοποιεί αδιάκοπα και μόνιμα το σύστημα αλλοίωσης μετρήσεων με τον ίδια ένταση.

2. *Απώλειες Τύπου 2* Μοντελοποιεί τον καταναλωτή που θα χρησιμοποιεί τυχαίες μέρες και για τυχαία διάρκεια μέσα στη μέρα σύστημα που αλλοιώνει τη μέτρηση με διαφορετική ένταση ανά ημέρα.

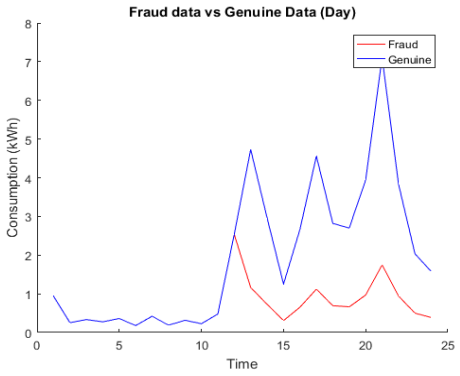
3. *Απώλειες Τύπου 3* Μοντελοποιεί τον καταναλωτή που θα χρησιμοποιεί τυχαίες μέρες και για τυχαία διάρκεια μέσα στη μέρα σύστημα που αλλοιώνει τη μέτρηση με διαφορετική ένταση ανά ώρα για κάθε διάρκεια.

4. *Απώλειες Τύπου 4* Μοντελοποιεί τον καταναλωτή που εκμεταλλεύεται την κυμαινόμενη χρέωση και αλλοιώνει τις τιμές του κατά τέτοιο τρόπο ώστε η μεγάλη κατανάλωση να μεταφέρεται τις ώρες μειωμένης χρέωσης.

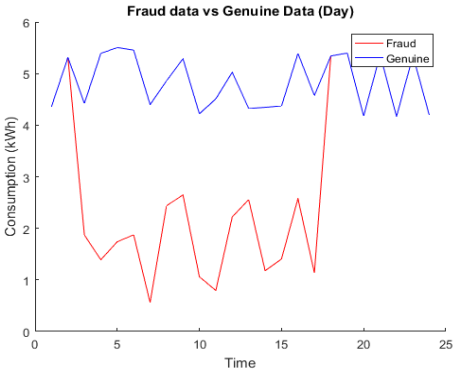
5. *Απώλειες Μικτών Τύπων* Μοντελοποιεί το 70% με απώλειες τύπου 1, το 20% με απώλειες τύπου 2 και το 10% με απώλειες τύπου 1. Η παρακάτω λογική βασίζεται στο



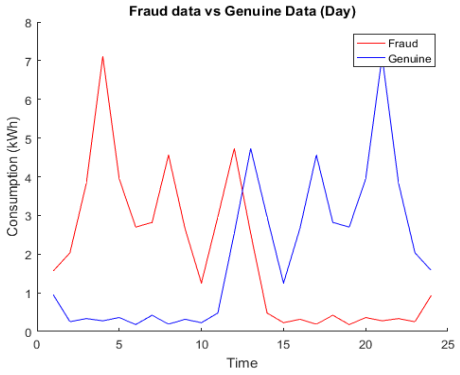
(α′) Απώλειες Τύπου 1



(β′) Απώλειες Τύπου 2



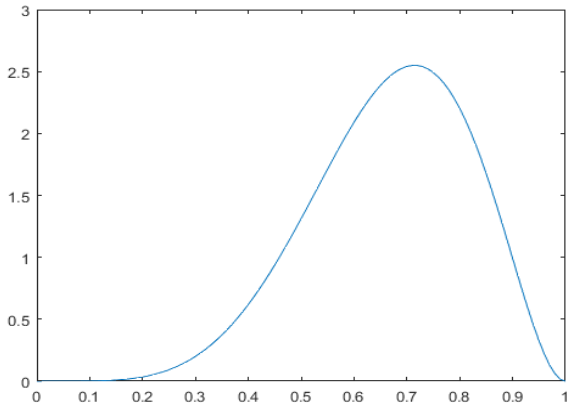
(γ′) Απώλειες Τύπου 3



(δ′) Απώλεια Τύπου 4

Σχήμα 3.15: Παραδείγματα απωλειών σε μια ημέρα

γεγονός πως ο ευκολότερος τύπος απώλειας συναντάται πολύ συχνότερα από τον πιο περίπλοκο.



Σχήμα 3.16: Πιθανοφάνεια κατανονής Βήτα(6,3)

Κεφάλαιο 4

Αλγόριθμοι επιβλεπόμενης μάθησης

Στο παρόν κεφάλαιο γίνεται μια εξερεύνηση στους αλγορίθμους επιβλεπόμενης μάθησης. Αυτό επιτεύχθηκε με τη χρήση γραμμικών και μη-γραμμικών ταξινομητών διερευνώντας διαφορετικά δεδομένα εισόδου για κάθε περίπτωση. Η βιβλιοθήκη που χρησιμοποιήθηκε για τη γραμμική ταξινόμηση ονομάζεται LIBLINEAR και χαρακτηρίζεται με εξαιρετικές επιδόσεις σε προβλήματα με μεγάλα σετ δεδομένων. Αντίστοιχα για τη μη-γραμμική ταξινόμηση χρησιμοποιήθηκε η βιβλιοθήκη LIBSVM, η οποία αναγάγει τα δεδομένα εισόδου σε μεγαλύτερο χώρο διαστάσεων.

4.1 Θεωρία γραμμικής ταξινόμησης

Η βιβλιοθήκη LIBLINEAR υποστηρίζει δύο δημοφιλείς δυαδικά γραμμικούς ταξινομητές: τη λογιστική παλινδρόμηση (Logistic Regression) και τη γραμμική μηχανή υποστήριξης διανυσμάτων (linear SVM). Δεδομένου ενός σετ εκπαίδευσης $(\mathbf{x}_i, y_i)$, $i = 1, ..., l$, όπου $\mathbf{x}_i \in \mathbb{R}^n$ είναι ένα χαρακτηριστικό διάνυσμα και $y_i = \pm 1$ είναι οι ετικέτες, ένας γραμμικός ταξινομητής βρίσκει ένα διάνυσμα βαρών $\mathbf{w} \in \mathbb{R}^n$ επιλύοντας το ακόλουθο πρόβλημα:

$$min_{\mathbf{w}} f(\mathbf{w}) \equiv \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l} \xi(y_i \mathbf{w}^T x_i)$$

όπου $\mathbf{w}^T\mathbf{w}/2$ είναι ο όρος ομαλοποίησης, $\xi(y_i\mathbf{w}^T x_i)$ είναι η συνάρτηση κόστους (loss function) και $C > 0$ είναι η παράμετρος ομαλοποίησης. Θεωρούμε τις συναρτήσεις κόστους στη λογιστική παλινδρόμηση (LR), στο L1-SVM, στο L2-SVM:

$$\xi_{LR}(y\mathbf{w}^T\mathbf{x}) = log(1 + exp(-y\mathbf{w}^T\mathbf{x}))$$
$$\xi_{L1}(y\mathbf{w}^T\mathbf{x}) = (max(0, 1 - y\mathbf{w}^T\mathbf{x}))$$
$$\xi_{L2}(y\mathbf{w}^T\mathbf{x}) = (max(0, 1 - y\mathbf{w}^T\mathbf{x}))^2$$

Σε μερικές περιπτώσεις, η συνάρτηση διακρίσεως του ταξινομητή περιλαμβάνει και ένα παράγοντα βάρους, b. Η LIBLINEAR χειρίζεται αυτό τον παράγοντα αυξάνοντας το διάνυσμα $\mathbf{w}$ και κάθε παράδειγμα $\mathbf{x}_i$ με μία επιπλέον διάσταση: $\mathbf{w}^T \leftarrow [\mathbf{w}^T, b]$, $\mathbf{x}_i^T \leftarrow [\mathbf{x}_i^T, B]$, όπου B είναι

μια σταθερά που ορίζεται από το χρήστη. Η προσέγγιση για το L1-SVM και το L2-SVM είναι μέσω της μεθόδου coordinate descent. Για το LR και το L2-SVM, η LIBLINEAR υλοποιεί μια μέθοδο περιοχής εμπιστοσύνης Newton. Στη φάση των δοκιμών, εκτιμάται ένα μέλος των δεδομένων $\mathbf{x}$ σαν θετικό εάν $\mathbf{w}^T\mathbf{x} > 0$, και αρνητικό σε αντίθετη περίπτωση[22] [4].

Η μηχανή διανυσμάτων υποστήριξης εντάσσεται στο γενικότερο πλαίσιο της βελτιστοποίησης κυρτών συναρτήσεων και έχει νόημα η προσέγγισή της για όλους τους γραμμικούς ταξινομητές. Σε αδρές γραμμές, η διαδικασία εξελίσσεται σε τέσσερα κύρια βήματα:

- Το πρόβλημα της εύρεσης του βελτίστου υπερεπιπέδου ξεκινά με μια δήλωση του προβλήματος στον πρωτεύοντα χώρο βαρών, ως ένα πρόβλημα βελτιστοποίησης με περιορισμούς.
- Κατασκευάζεται η συνάρτηση Lagrange του προβλήματος.
- Διατυπώνονται οι συνθήκες για τη βελτιστοποίηση της μηχανής.
- Στήνεται το σκηνικό για την επίλυση του προβλήματος βελτιστοποίησης στο δυικό χώρο των πολλαπλασιαστών Lagrange.

Όπως προαναφέρθηκε, το πρωτεύον πρόβλημα ασχολείται με μια κυρτή συνάρτηση κόστους και γραμμικούς περιορισμούς. Δοθέντος ενός τέτοιου προβλήματος βελτιστοποίησης με περιορισμούς, είναι δυνατό να κατασκευάσουμε ένα άλλο πρόβλημα, το αποκαλούμενο δυικό του πρωτεύοντος. Αυτό το δεύτερο πρόβλημα έχει την ίδια βέλτιστη τιμή με το πρωτεύον πρόβλημα, αλλά με τους πολλαπλασιαστές Lagrange να παρέχουν τη βέλτιστη λύση[29].

4.2 Εξερεύνηση γραμμικών ταξινομητών

Αρχικά έγινε μια εξερεύνηση των μεθόδων που παρέχει η LIBLINEAR για την επίλυση του δυαδικού προβλήματος. Λαμβάνοντας υπόψη 2.000 καταναλώσεις πελατών με ωριαίες μετρήσεις, επιλέχθηκε 10% ποσοστό ρευματοκλοπών για την προσομοίωση. Η βιβλιοθήκη που χρησιμοποιήθηκε περιλαμβάνει 7 διαφορετικούς συνδυασμούς ταξινομητών και συναρτήσεων κόστους για να μπορούν όσο το δυνατόν περισσότερα προβλήματα. Παρόλα αυτά οι μέθοδοι L1 είναι παλαιότερες εκδόσεις των L2 και αναμένεται να έχουν χειρότερα αποτελέσματα στις δοκιμές. Για την σφαιρική αντιμετώπιση του προβλήματος χρησιμοποιήθηκαν όλοι οι ταξινομητές που παρέχονται από τη βιβλιοθήκη σε κάθε τύπο απάτης. Παρακάτω παραθέτονται οι συνδυασμοί ταξινομητών και συναρτήσεων κόστους που δοκιμάστηκαν και τα αποτελέσματα σε κάθε τύπο απάτης.

1. L2 ομαλοποιημένη λογιστική παλινδρόμηση (πρωτεύον)
2. L2 ομαλοποιημένος ταξινομητής με L2 συνάρτηση κόστους διανυσμάτων υποστήριξης (δυικό)
3. L2 ομαλοποιημένος ταξινομητής με L2 συνάρτηση κόστους διανυσμάτων υποστήριξης (πρωτεύον)

4. L2 ομαλοποιημένη ταξινομητής με L1 συνάρτηση κόστους διανυσμάτων υποστήριξης (δυικό)

5. Ταξινόμηση διανυσμάτων υποστήριξης από Crammer και Singer

6. L1 ομαλοποιημένος ταξινομητής με L2 συνάρτηση κόστους διανυσμάτων υποστήριξης

7. L1 ομαλοποιημένη λογιστική παλινδρόμηση

8. L2 ομαλοποιημένη λογιστική παλινδρόμηση (δυικό)

Παρατηρώντας τους πίνακες αποτελεσμάτων εύκολα αποδεικνύεται η αρχική υπόθεση πως οι ταξινομητές και συναρτήσεις κόστους L2 έχουν καλύτερη συμπεριφορά ως προς την αντιμετώπιση του προβλήματος αναγνώρισης χρονοσειρών. Πιο συγκεκριμένα για την τελική επιλογή του συνδυασμού μεθόδων επιλέχθηκαν δύο μετρικές για να καθορίσουν την επιλογή του καλύτερου πακέτου. Λήφθηκε υπόψη η μεταβολή της ευστοχίας (αςςυραςψ) και παράχθηκε μέσος όρος για όλους τους τύπους. Παράλληλα, υπολογίστηκε μέσος όρος των δοκιμών με γνώμονα το καλύτερο F1 score, καθώς είναι μια αρκετά ζυγισμένη μετρική για τα προβλήματα ταξινόμησης. Βάση λοιπόν του Πίνακα 4.1 την καλύτερη επίδοση έχει το πρωτεύον πρόβλημα που αποτελείται από L2 ομαλοποιημένο ταξινομητή με L2 συνάρτηση κόστους διανυσμάτων υποστήριξης, καθώς όπως μπορεί και να φανεί στον Πίνακα Α΄.6 του Παραρτήματος η μηχανή διανυσμάτων υποστήριξης Crammer και Singer έχει καλύτερη επίδοση στους τύπους 2, 3 και στον μικτό. Αλλά, στην παρούσα φάση θα ασχοληθούμε με την απάτη τύπου 1.

Συνδυασμός	1	2	3	4	5	6	7	8
F1 score	23.92	31.99	30.19	28.67	32.66	29.28	20.43	24.04
Accuracy	91.36	90.41	90.46	90.56	90.15	90.37	91.43	91.35
Μέσος όρος	57.64	61.2	60.33	59.61	61.4	59,83	55.93	57.7

Πίνακας 4.1: Μέσος όρος Accuracy των δοκιμών

Συνδυασμός	DR	FPR	Accuracy	F1 score	BDR
1	77.44	1.56	96.37	80.78	0.85
2	79.70	1.81	96.37	81.23	0.83
3	78.95	2.22	95.93	79.25	0.80
4	78.95	2.05	96.07	79.85	0.81
5	78.20	1.81	96.22	80.31	0.83
6	77.44	2.14	95.85	78.63	0.80
7	75.94	1.81	96.00	78.91	0.82
8	79.70	1.81	96.37	81.23	0.83

Πίνακας 4.2: Αποτελέσματα δοκιμής τύπου 1 χωρίς κανονικοποίηση

4.3 Εξερεύνηση διαφορετικών τρόπων κανονικοποίησης

Το σκέλος της κανονικοποίησης των δεδομένων είναι ζωτικής σημασίας για κάθε σύστημα μηχανικής μάθησης. Η κανονικοποίηση των δεδομένων υλοποιείται, μειώνοντας το εύρος των τιμών σε οποιαδήποτε σχετικά μικρό εύρος. Συνηθέστερη πετυχημένη πρακτική είναι η αναγωγή των τιμών σε εύρος [0,1] ή [-1,1] με στόχο την βελτίωση της επίδοσης και της ταχύτητας του αλγορίθμου. Αυτό επιτυγχάνεται σε μεγάλο βαθμό στην συγκεκριμένη περίπτωση από την κανονικοποίηση στο εύρος [0,1], βελτιώνοντας σε μικρό βαθμό τις μετρικές και μειώνοντας σχεδόν 10 φορές τον χρόνο εκτέλεσης της εκπαίδευσης. Στον Πίνακα 4.3 παραθέτονται τα αποτελέσματα των βέλτιστων ταξινομητών σε κάθε είδος κανονικοποίησης.

Συνδυασμός	Κανονικοποίηση	DR	FPR	Accuracy	F1 score	BDR	senconds εκπ.
4	[0,1]	80.87	1.54	96.96	81.94	0.85	6.492741
4	[-1,1]	91.67	21.23	80.15	49.62	0.32	551.264250
2	-	79.70	1.81	96.37	81.23	0.83	58.246916

Πίνακας 4.3: Αποτελέσματα κανονικοποιήσεων

4.4 Εξερεύνηση χρονικής υποδιαίρεσης χρονοσειρών

Ολοκληρώνοντας την εξερεύνηση των ταξινομητών απαιτείται να γίνει έλεγχος στις χρονικές υποδιαιρέσεις των χρονοσειρών. Για αυτό το σκοπό έγινε δοκιμή του πιο εύστοχου ταξινομητή σε 2.000 καταναλωτές με ποσοστό ρευματοκλοπών 10% και μόνο απάτες τύπου 1. Στη δοκιμή οι χρονοσειρές διαιρέθηκαν σε ημερήσιες, ωριαίες και ημίωρες μετρήσεις λαμβάνοντας υπόψη όχι μόνο τις μετρικές ευστοχίας, αλλά και τον χρόνο εκτέλεσης της εκπαίδευσης κάθε ταξινομητή. Στον Πίνακα 4.4 εμφανίζεται όπως αναμενόταν πως όσο αυξάνεται η συχνότητα των μετρήσεων τόσο πιο εύστοχος γίνεται ο ταξινομητής. Παρόλα αυτά ο χρόνος εκτέλεσης της εκπαίδευσης φαίνεται να επηρεάζεται έντονα από διαφορετικές χρονικές υποδιαιρέσεις με την ταξινόμηση με συχνότητα λήξης ανά ημέρα να είναι σημαντικά γρηγορότερη από τις υπόλοιπες, αλλά παρουσιάζει σχετική δυσκολία στην αναγνώριση της απάτης.

Συχνότητα	DR	FPR	Accuracy	F1 score	BDR	χρόνος εκπαίδευσης (s)
μέρες	81.62	2.55	95.85	79.86	0.78	0.069182
ώρες	82.88	2.16	96.22	82.59	0.81	4.152410
ημίωρα	81.08	1.66	96.44	83.33	0.84	12.169304

Πίνακας 4.4: Αποτελέσματα δοκιμής χρονικής υποδιαίρεσης

4.5 Θεωρία Μηχανών Διανυσμάτων Υποστήριξης

Για την ταξινόμηση με μηχανές διανυσμάτων υποστήριξης επιλέχθηκε η βιβλιοθήκη LIBSVM, η οποία προέρχεται τους ίδιους δημιουργούς της LIBLINEAR. Σκοπός του SVM είναι η παραγωγή μοντέλων (βαση των δεδομένων εκπαίδευσης), τα οποία προβλέπουν τα χαρακτηριστικά των δεδομένων δοκιμής βάση μόνο των πληροφοριών που αντλούνται από τις τιμές των δεδομένων.

Ξεκινώντας από τα δεδομένα εκπαίδευσης έχουμε ζευγάρια παραδειγμάτων-δυαδικών χαρακτηριστικών $(\mathbf{x}_i, y_i)$,$i = 1, ..., l$ όπου $\mathbf{x}_i \in \mathbb{R}^n$ και $y \in \{1, -1\}^l$, οι μηχανές διανυσμάτων υποστήριξης (SVM) απαιτούν την λύση του παρακάτω προβλήματος βελτιστοποίησης:

$$\begin{gathered} min_{\mathbf{w},\mathbf{b},\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{δεδομένου } y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ \xi_i \geq 0 \end{gathered}$$

Εδώ τα διανύσματα εκπαίδευσης $\mathbf{x}_i$, ανάγονται σε μεγαλύτερο (ίσως άπειρο) χώρο διαστάσεων από τη συνάρτηση ϕ. Τα SVM βρίσκουν ένα γραμμικά διαχωρίσιμο υπερεπίπεδο με μέγιστο περιθώριο σε αυτό χώρο ανώτερων διαστάσεων. $C > 0$ είναι ο παράγοντας που θέτει ποινή στον παράγοντα λάθους (error term). Επιπροσθέτως, η σχέση $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ ονομάζεται συνάρτηση πυρήνα. Παρόλο που νέοι πυρήνες προτείνονται από ερευνητές, έχουν θεσπιστεί οι ακόλουθοι:

- Γραμμικός: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$.
- Πολυωνυμικός: $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma\mathbf{x}_i^T\mathbf{x}_j + r)^d$, $\gamma = \frac{1}{2\sigma^2} > 0$.
- RBF: $K(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, $\gamma > 0$.
- Σιγμοειδής: $K(\mathbf{x}_i, \mathbf{x}_j) = tanh(\gamma\mathbf{x}_i^T\mathbf{x}_j + r)$.

Εδώ τα γ, r και d είναι παράμετροι των πυρήνων [20].

Χρησιμοποιώντας τη μέθοδο των πολλαπλασιαστών Lagrange μπορεί να διατυπωθεί το δυικό πρόβλημα για τα μη-διαχωρίσιμα πρότυπα. Δοθέντος του δείγματος εκπαίδευσης $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, βρίσκονται οι πολλαπλασιαστές Lagrange $\{\alpha\}_{i=1}^N$ που μεγιστοποιούν την αντικειμενική συνάρτηση

$$Q(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$

υπό τους περιορισμούς

$$\begin{gathered} \sum_{i=1}^{N}\alpha_i d_i = 0 \\ 0 \leq \alpha_i \leq C \text{ για } i = 1, 2, ..., N \end{gathered}$$

όπου C είναι μια καθοριζόμενη από το χρήστη θετική παράμετρος [29].

4.5.1 Επιλογή πυρήνα RBF

Γενικώς, ο πυρήνας βάση δικτύου ακτινικής συνάρτησης βάσης (RBF) είναι μια λογική πρώτη επιλογή. Αυτός ο πυρήνας ανάγει μη-γραμμικά τα δείγματα σε υψηλότερο χώρο διαστάσεων που μπορεί να διαχειριστεί την περίπτωση που η σχέση μεταξύ της τάξης και της τιμής είναι μη-γραμμική. Επιπροσθέτως ο γραμμικός πυρήνας είναι μια ειδική περίπτωση του RBF, καθώς ο γραμμικός πυρήνας με την παράμετρο ποινής $\check{C}$ έχει την ίδια επίδοση με τον RBF με δύο παραμέτρους (C,γ). Επιπρόσθετα, ο πυρήνας με σιγμοειδή συνάρτηση συμπεριφέρεται όπως με RBF για συγκεκριμένες παραμέτρους.

Ο δεύτερος λόγος είναι ο αριθμός των υπερπαραμέτρων, οι οποίες επηρεάζουν την πολυπλοκότητα της επιλογής μοντέλου. Ο πολυωνυμικός πυρήνας έχει περισσότερες υπερπαραμέτρους από τον RBF πυρήνα.

Τέλος, ο πυρήνας RBF έχει λιγότερες αριθμητικές δυσκολίες. Το χαρακτηριστικό κλειδί είναι πως το $0 < K_{ij} \leq 1$ είναι σταθερά του πολυωνυμικού πυρήνα του οποίου οι τιμές μπορούν να φτάνουν το άπειρο ($\gamma \mathbf{x}_i^T \mathbf{x}_j + r > 1$) ή το μηδέν ($\gamma \mathbf{x}_i^T \mathbf{x}_j + r < 1$) ενώ ο βαθμός είναι ήδη μεγάλος. Επίσης, πρέπει να σημειωθεί πως π σιγμοειδής πυρήνας δεν είναι εφικτός με κάποιες παραμέτρους.

Υπάρχουν κάποιες περιπτώσεις όπου ο πυρήνας RBF δεν είναι κατάλληλος. Πιο συγκεκριμένα, όταν ο αριθμός των χαρακτηριστικών είναι πολύ μεγάλος, κάποιος θα μπορούσε να χρησιμοποιήσει το γραμμικό πυρήνα [20].

4.6 Δοκιμή ταξινόμησης με Μηχανές Διανυσμάτων Υποστήριξης

Η προτεινόμενη διαδικασία που ακολουθείται από τους δημιουργούς του LIBSVM είναι η εξής:

- Μετατροπή των δεδομένων σε μορφή αναγνωρίσιμη μορφή με το πακέτο SVM
- Κανονικοποίηση δεδομένων
- Εξέταση του RBF πυρήνα
- Χρήση cross-validation για την εύρεση των βέλτιστων παραμέτρων C και γ
- Χρήση των βέλτιστων παραμέτρων C και γ για την εκπαίδευση των δεδομένων εκπαίδευσης
- Δοκιμή

Έχοντας τη διαδικασία αυτή υπόψη δοκιμάστηκαν επιτυχώς δύο διαφορετικά σενάρια ταξινόμησης. Στο πρώτο σενάριο ταξινομήθηκαν οι χρονοσειρές κάθε καταναλωτή βάση της ετήσιας κατανάλωσης τους και αναγνωρίζοντας κάθε τύπο κλοπής. Στο δεύτερο σενάριο χρησιμοποιήθηκε ο πυρήνας RBF και έγινε μια προσέγγιση στην αναγνώριση των ημερήσιων μη

Τύπος	DR	FPR	Accuracy	F1 score	BDR	χρόνος εκτέλεσης
1	81.43	1.24	96.96	84.76	0.88	10.188667
2	22.63	7.25	85.63	24.22	0.26	39.489221
3	23.78	10.36	82.67	22.52	0.20	39.648516
Μικτός	27.13	7.37	86.37	27.56	0.29	36.836504

Πίνακας 4.5: Αποτελέσματα Γραμμικού SVM σε όλους τους τύπους απάτης

τεχνικών απωλειών ταξινομώντας σε πρώτη φάση τις ημέρες όλων των καταναλωτών και σε δεύτερη φάση κάθε καταναλωτή [20].

4.6.1 Δοκιμή χρονοσειρών χωρίς πυρήνα

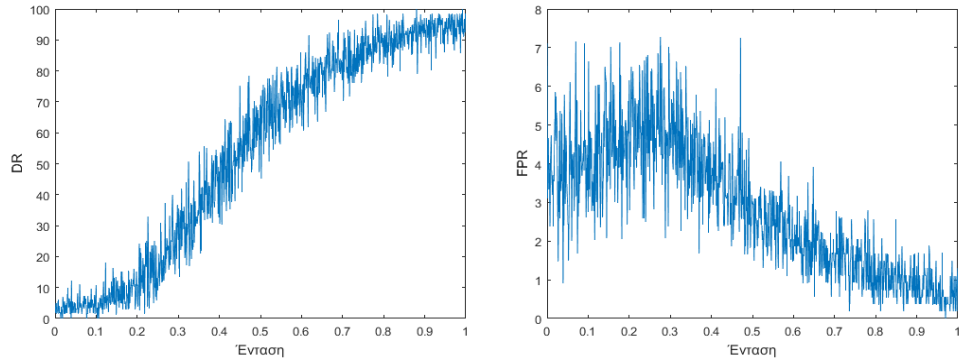
Δεδομένης της ευστοχίας των γραμμικών ταξινομητών θεωρήθηκε αναγκαία η δοκιμή του γραμμικού πυρήνα SVM. Παρόλα αυτά, η διαίσθηση δεν ήταν η μόνη κινητήριος δύναμη για την υλοποίηση αυτής της δοκιμής. Γενικότερα, αν ο αριθμός των μετρήσεων είναι μεγάλος δεν απαιτείται να αναχθούν τα δεδομένα σε χώρο ανώτερων διαστάσεων. Πρακτικά αυτό σημαίνει πως η μη-γραμμική αναγωγή δεν βελτιώνει την επίδοση του συστήματος. Ενώ, είναι γενικώς αποδεκτό ότι ο πυρήνας RBF είναι τουλάχιστον καλύτερος από το γραμμικό, αυτή η δήλωση είναι αληθής μόνο αφού έχουν επιλεχθεί οι παράμετροι (C,γ). Ένας γενικός κανόνας χρήσης του γραμμικού πυρήνα είναι η χρήση του όταν ο αριθμός των παραδειγμάτων (καταναλωτών) είναι μικρότερος ή σχετικός με τον αριθμό των χαρακτηριστικών (ωριαίες μετρήσεις έτους).

Αποτελέσματα δοκιμής

Η δοκιμή έγινε σε 4.500 καταναλωτές ελέγχοντας αρχικά την επίδοση του συστήματος σε κάθε τύπο απάτης με ποσοστό ρευματοκλοπής 10%. Στον Πίνακα 4.5 φαίνονται τα αποτελέσματα της δοκιμής. Γίνεται, λοιπόν σαφές πως ο ταξινομητής μπορεί να αναγνωρίσει με αξιοπιστία μόνο τις απάτες τύπου 1, όπως και οι αντίστοιχοι ταξινομητές της LIBSVM. Παρόλα αυτά ακόμα και στα χαμηλότερα αποτελέσματα έχουμε ικανοποιητικό Accuracy που δείχνει ότι ο ταξινομητής λειτουργεί όπως αναμενόταν.

Για να την βαθύτερη κατανόηση της λειτουργίας του ταξινομητή, απαιτείται η παρατήρηση της σχέσης των μετρικών με την ένταση κλοπής. Η ένταση κλοπής μαθηματικοποιείται σαν ένας παράγοντας που μπορεί να ποσοτικοποιήσει πόσο απέχουν τα αλλοιωμένα δεδομένα από τις πραγματικές μετρήσεις. Ουσιαστικά είναι ο συντελεστής υποδιαίρεσης των πραγματικών μετρήσεων.

Τα Γραφίματα 4.1 δείχνουν πως ο ταξινομητής ξεκινά να βελτιώνεται αφότου η ένταση αυξηθεί πάνω από 30%, καθώς το DR αυξάνεται σχεδόν γραμμικά με την ένταση και το FPR μειώνεται σταθερά μετά από αυτό το σημείο. Εκεί που ο ταξινομητής έχει την βέλτιστη απόδοση είναι στο εύρος [70%-90%] αφού η κλίση της καμπύλης σε αυτά τα σημεία είναι σημαντικά μικρότερη γεγονός που υποδεικνύει σύγκλιση.



(α′) DR συναρτήσει της έντασης της κλοπής

(β′) FPR συναρτήσει της έντασης της κλοπής

Σχήμα 4.1: Επίπτωση της έντασης στα αποτελέσματα

4.6.2 Δοκιμή ημερήσιων χαρακτηριστικών με πυρήνα RBF

Σε αυτή τη φάση, δημιουργήθηκε η ανάγκη για εξαγωγή χαρακτηριστικών, ώστε να μειωθούν οι διαστάσεις των πινάκων και να επιταχυνθεί η διαδικασία. Παράλληλα, παρέχει ένα επίπεδο αποπροσωποποίησης δημιουργώντας ένα αποτύπωμα της καταναλωτικής συνήθειας.[13]. Μετρώντας τα αθροίσματα, τα ελάχιστα, τα μέγιστα και τους μέσους όρους των καθημερινών καταναλώσεων δημιουργείται ένας βασικός κορμός χαρακτηριστικών για κάθε καταναλωτή που μπορεί εύκολα να επεκταθεί και σε άλλα γραμμικά και μη εξαρτώμενα χαρακτηριστικά.

- *Μέγιστο και ώρα μεγίστου*
- *Ελάχιστο και ώρα ελαχίστου*
- *Άθροισμα κατανάλωσης ανά ημέρα*
- *Μέσος όρος, διακύμανση και τυπική απόκλιση ανά ημέρα*
- *Παράγοντας φορτίου, ελάχιστο προς μέση τιμή, ελάχιστο προς μέγιστο*
- *Επίδραση βραδινής κατανάλωσης*
- *Λοξότητα και Κύρτωση*

Η πρώτη δοκιμή του SVM έγινε με επιλογή 300 τυχαίων καταναλωτών μιας περιοχής με σκοπό να εκπαιδευτεί το σύστημα ώστε να μπορεί να αναγνωρίζει ημέρες απάτης μέσα στο έτος. Η εκπαίδευση του ταξινομητή γινόταν με τα ημερήσια χαρακτηριστικά για κάθε καταναλωτή μαζί με τον confusion matrix. Τα δεδομένα διαχωρίζονται σε 2 κομμάτια, το κομμάτι της εκπαίδευσης που περιέχει ένα μεγάλο ποσοστό δεδομένων κάθε καταναλωτή και το κομμάτι της δοκιμής που περιέχει ένα ποσοστό της τάξης του 0.30 από τους αντίστοιχους καταναλωτές.

Ο ταξινομητής λοιπόν, εκπαιδεύεται με ημερήσια χαρακτηριστικά κάθε καταναλωτή, αλλά θα πρέπει να αποφανθεί στο τέλος αν ο καταναλωτής έχει νοθεύσει τις μετρήσεις του ή όχι. Η λύση δόθηκε εισάγοντας ένα όριο ημερών που αν ο ταξινιμητής το προσπερνούσε τότε ο

καταναλωτής θεωρείται πως έχει αλλοιώσει τα δεδομένα του. Για να βρούμε την βέλτιστη τιμή αυτού του ορίου χρησιμοποιήθηκαν ROC καμπύλες για να παρατηρηθεί η μεταβολή του DR και FPR, ενώ αλλάζει το όριο ημερών.

Αποτελέσματα δοκιμής

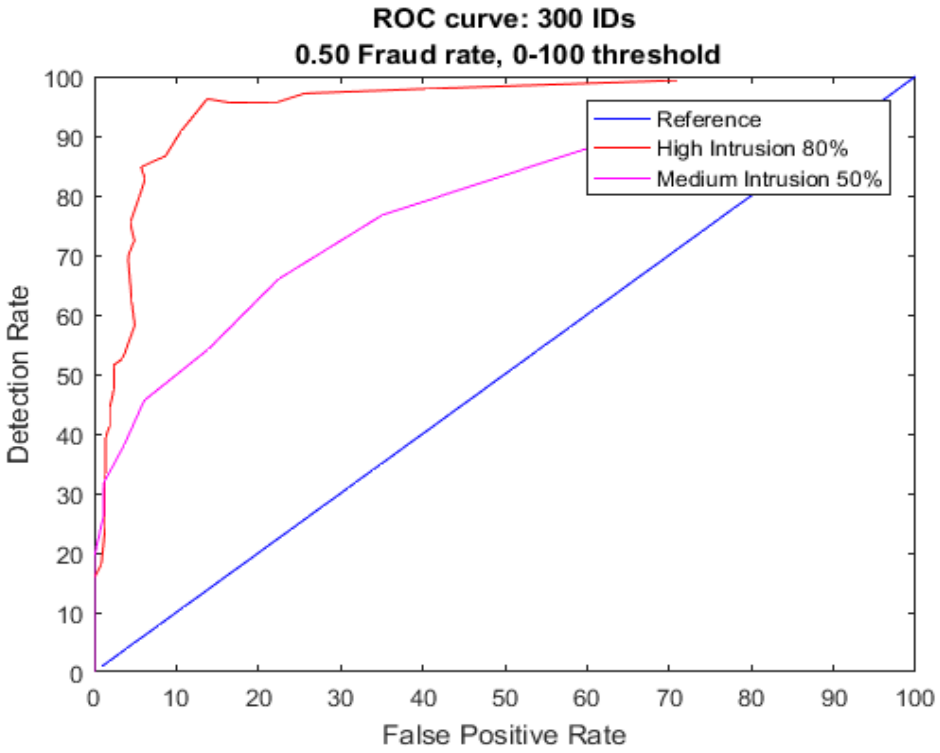
Ελέγχοντας τα αποτελέσματα του Πίνακα 4.3 παρατηρείται πως επιλέγοντας όριο στις 10 ημέρες επιτυγχάνεται ακρίβεια της τάξης τους 0.95 στην εύρεση της απάτης, αλλά με σχετικά υψηλό ποσοστό λάθος συναγερμού της τάξης του 0.15 για τις έντονες απάτες. Αν χρειαστεί να ελαχιστοποιηθεί το FPR θα πρέπει να επιλεχθεί μια μεγαλύτερη οριακή τιμή όπως το 14, που έχει ικανοποιητικό ποσοστό και στο DR που είναι της τάξης του 0.85 και του FPR που είναι της τάξης του 0.08. Οι απάτες που έγιναν με μικρότερη ένταση δεν γίνονται αντιληπτες από τον ταξινομητή που επιστρέφει καμπύλη με παρόμοια κλίση με της ευθείας αναφοράς.

Αντίστοιχα στον Πίνακα 4.5 φαίνεται πως η μείωση του FR επηρέασε το σύστημα, και ειδικότερα μείωσε το όριο στις 10 μέρες με DR=0.85 και FPR=0.09. Ουσιαστικά φαίνεται πως το σύστημα χρειάζεται και άλλους καταναλωτές ώστε να αποτυπωθούν και οι καμπύλες για χαμηλότερες εντάσεις διείσδυσης στα δεδομένα.

4.7 Σχόλια

Συνοψίζοντας, καθίσταται σαφές πως μπορεί να χρησιμοποιηθεί επιτυχώς επιβλεπόμενη μάθηση για τον εντοπισμό μη-τεχνικών απωλειών. Οι γραμμικοί ταξινομητές μπορούν να αναγνωρίσουν αξιόπιστα και γρήγορα τον πρώτο τύπο απάτης, ενώ έχουν δυσκολία εντοπισμού στους υπόλοιπους τύπους. Παρόλα, αυτά χρησιμοποιώντας τον πυρήνα RBF, γίνεται εφικτή η αναγνώριση μη τεχνικών απωλειών αρχικά κάθε ημέρας και εν συνεχεία κάθε καταναλωτή. Γενικότερα, όμως και οι δύο ομάδες ταξινομητών έχουν καλές επιδόσεις στον εντοπισμό ρευματοκλοπών με έντονη ένταση κλοπής, ενώ όσο μειώνεται οι ταξινομητές δείχνουν μεγαλύτερη δυσκολία να διαχωρίσουν αλλοιωμένα από κανονικά δεδομένα.

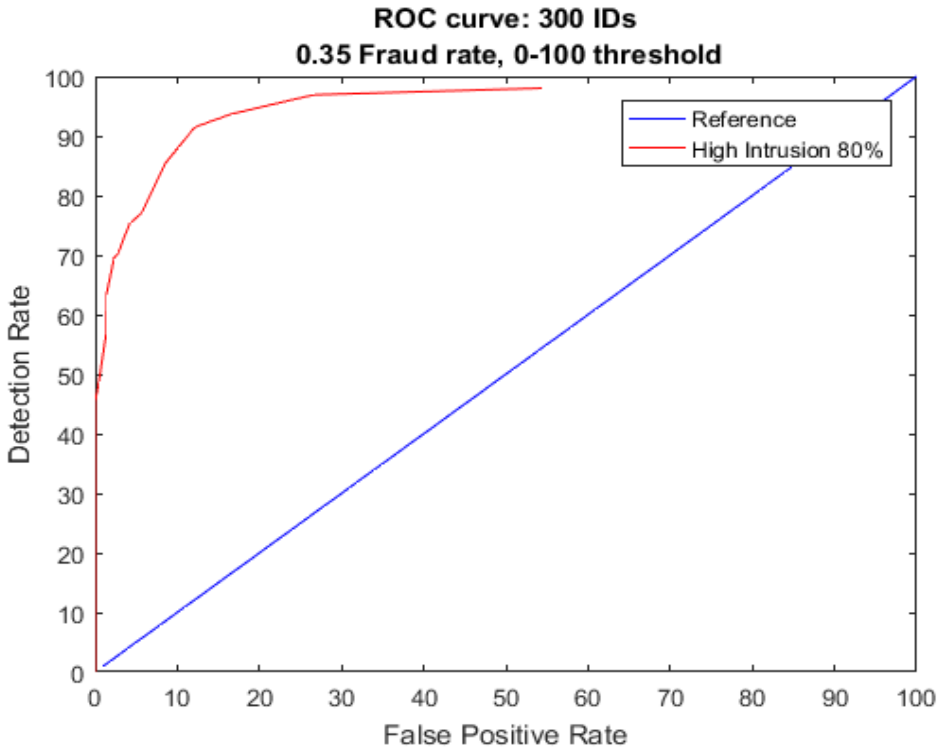
Παράλληλα, γίνεται εμφανής η ανάγκη για τη σωστή επιλογή της δομής των δεδομένων εισόδου, καθώς κάθε ταξινομητής απαιτεί διαφορετική μεταχείριση. Οι γραμμικοί ταξινομητές απαιτούν πολλά χαρακτηριστικά (μετρήσεις), ενώ οι μη γραμμικοί μπορούν να λειτουργήσουν με πολύ λιγότερα. Αντίστοιχα, η κανονικοποίηση προσφέρει άμεση βελτίωση στα αποτελέσματα και επιταχύνει τη διαδικασία εκπαίδευσης σε μεγάλο βαθμό.



Σχήμα 4.2: Καμπύλη ROC για FR=0.50

300 IDs, 0.5 rate, 0-100 threshold				
Όριο (Μέρες)	DR (0.8)	FPR (0.8)	DR (0.5)	FPR (0.5)
2	97,917	40,385	76,712	35,0649
4	97,143	25,625	65,972	22,436
6	95,683	22,360	54,225	13,924
8	95,588	16,463	45,588	6,098
10	96,241	13,772	37,879	3,571
12	90,698	10,526	31,783	1,17
14	86,614	8,671	26,190	1,149
16	84,8	5,714	19,355	0
18	82,787	6,18	15,702	0
20	79,832	5,525	11,667	0

Σχήμα 4.3: Πίνακας επιλογής ορίου FR=0.5



Σχήμα 4.4: Καμπύλη ROC για FR=0.35

300 IDs, 0.35 rate, 0-100 threshold		
Όριο (Μέρες)	DR (0.8)	FPR (0.8)
2	95,192	30,612
4	91,176	23,232
6	89,691	17,734
8	85,567	12,808
10	85,106	8,738
12	84,444	5,238
14	79,545	2,830
16	75	2,830
18	68,235	2,791
20	63,529	2,791

Σχήμα 4.5: Πίνακας επιλογής ορίου FR=0.35

Κεφάλαιο 5

Αλγόριθμοι μη-επιβλεπόμενης μάθησης

Ολοκληρώνοντας τον κύκλο των δοκιμών για τους αλγορίθμους επιβλεπόμενης μάθησης, δημιουργήθηκε η ανάγκη για περαιτέρω έρευνα σε διαφορετικούς αλγορίθμους. Οι αλγόριθμοι επιβλεπόμενης μάθησης έχουν ένα βασικό μειονέκτημα, όταν προσεγγίζεται ένα πραγματικό πρόβλημα. Αυτό είναι η δυσκολίας εφαρμογής του αλγορίθμου, λόγω της έλλειψης των τάξεων των δεδομένων που απαιτεί ένα τέτοιο σύστημα για να εκπαιδευτεί. Η δυσκολία αυτή παρακάμπτεται χρησιμοποιώντας μη-επιβλεπόμενους ή ημι-επιβλεπόμενους αλγορίθμους που απαιτούν λίγα ή και κανένα ταξινομημένο παράδειγμα. Σε αυτό το κεφάλαιο θα προσεγγιστεί το πρόβλημα της ταξινόμησης καταναλωτών με νέα συστήματα που θα μπορούν να έχουν άμεσα χρήση στη λύση του πραγματικού προβλήματος, κάνοντας μια ανασκόπηση στις νέες δυσκολίες που προέκυψαν.

5.1 Εξαγωγή Χαρακτηριστικών

Στο παρόν μέρος θα γίνει παρουσίαση και ανάλυση των χαρακτηριστικών που χρησιμοποιήθηκαν στο μερικώς επιβλεπόμενο σύστημα, αλλά και στο μη επιβλεπόμενο σύστημα. Κάθε παράδειγμα μπορεί να περιγραφεί από ένα συνδιασμό τιμών που αναφέρονται επίσης ως μεταβλητές, χαρακτηριστικά, πεδία ή διαστάσεις. Οι τιμές αυτές μπορούν να είναι διαφορετικού τύπου όπως συνεχείς, δυαδικές ή κατηγορίες. Κάθε παράδειγμα μπορεί να αποτελείται μόνο από μια τιμή (μονοπαραγοντικό) ή και από περισσότερες (πολυπαραγοντική). Στην περίπτωση των πολυπαραγοντικών παραδειγμάτων, όλες οι τιμές μπορεί να είναι ίδιου τύπου ή μπορεί να είναι ένας συνδυασμός διαφορετικών τύπων [27].
Παράλληλα, κάθε παράδειγμα μπορεί να οριστεί βάση ακόμη δύο δομών ως προς τον ορισμό του προβλήματος [27].

1. *Τιμές Συσχετισμού* Τέτοιου είδους τιμές χρησιμοποιούνται για να περιγράψουν ένα γενικό πλαίσιο που χαρακτηρίζει ένα παράδειγμα. Στις χρονοσειρές, ο χρόνος είναι μια τιμή που παρέχει μια σχετικότητα, η οποία καθορίζει τη θέση ενός παραδείγματος σε

μια ολόκληρη ακολουθία. Μία τιμή γενικού πλαισίου είναι η μηνιαία κατανάλωση ενός κατοίκου.

2. *Συμπεριφορικές Τιμές* Είναι οι τιμές που δεν προδίδουν ένα γενικό πλαίσιο για κάποιο παράδειγμα ή κάποια σχετικότητα. Ένα τέτοιο παράδειγμα θα μπορούσε να είναι η ετήσια παραγωγή ενέργειας σε όλο τον κόσμο.

5.1.1 Φύση Χαρακτηριστικών

Το μερικώς επιβλεπόμενο και μη επιβλεπόμενο σύστημα απαιτούν εισόδους που να δίνουν τη δυνατότητα να διαχωρίζονται σε δύο κλάσεις οι καταναλωτές. Για να γίνει αυτό απαιτείται η χρήση χαρακτηριστικών που να αντιπροσωπεύουν την κλάση, αλλά και χαρακτηριστικά που προσδίνουν γενικότητα στο κάθε παράδειγμα. Με αυτό τον τρόπο παρέχεται ένα περιθώριο στον αλγόριθμο, έτσι ώστε να μπορεί εύκολα να προσαρμόζεται σε καινούργια και ξεχωριστά παραδείγματα. Ένας απλοϊκός τρόπος να διαχωρίσουμε τα χαρακτηριστικά είναι σε χαρακτηριστικά γενίκευσης και σε χαρακτηριστικά διαχωρισμού κλάσεων. Όλα τα παρακάτω χαρακτηριστικά αποτελούν τιμές συσχέτισης.

Χαρακτηριστικά Γενίκευσης

Τα πλεονέκτημα των χαρακτηριστικών γενίκευσης είναι ότι βοηθούν στην κατάταξη του καταναλωτή σε σχέση με τους υπόλοιπους, ώστε να εξαχθούν πληροφορίες, όπως ο τύπος καταναλωτή (οικιακού ή βιομηχανικού) και το προφίλ κατανάλωσής του. Τέτοια χαρακτηριστικά πρέπει να περιορίζονται σε αριθμό όμως, καθώς ενδέχεται να δυσκολεύσουν τον διαχωρισμό με βάση το κριτήριο που θέτουμε παρέχοντας μεγάλο παράγοντα γενίκευσης. Τέτοιου είδους χαρακτηριστικά είναι τα παρακάτω:

1. *Ετήσια μέση τιμή ημίωρου* Βρίσκεται ο μέσος όρος ημίωρου κάθε μέρας και για όλες τις μέρες του έτους βρίσκεται ο ετήσιος μέσος όρος.

2. *Ετήσια τυπική απόκλιση ημίωρου* Βρίσκεται η τυπική απόκλιση κάθε μέρας και για όλες τις μέρες του έτους βρίσκεται ο ετήσιος μέσος όρος της τυπικής απόκλισης.

3. *Διαφορά Ετήσιου Ελάχιστου τάσης με όμοιους* Βάση αυτού του χαρακηριστικού ορίζεται για όμοιους καταναλωτές το ελάχιστο της τάσης κατανάλωσής τους και στην συνέχεια βρίσκεται η απόλυτη διαφορά σε ημέρες.

4. *Διαφορά μέσης τιμής με ομοίους* Με αυτό το χαρακτηριστικό βρίσκεται η διαφορά του ετήσιου μέσου όρου κάθε καταναλωτή με την ομάδα καταναλωτών που ανήκει.

5. *Διαφορά τυπικής απόκλισης με ομοίους* Με αυτό το χαρακτηριστικό βρίσκεται η διαφορά της ετήσιας τυπικής απόκλισης κάθε καταναλωτή με την ομάδα καταναλωτών που ανήκει.

Χαρακτηριστικά Διαχωρισμού

Τα χαρακτηριστικά διαχωρισμού επικεντρώνονται στην όξυνση των διαφορών μεταξύ των καταναλωτών διαφορετικών κλάσεων. Λειτουργούν, λοιπόν σαν οδηγοί για τον αλγόριθμο ώστε να κάνουν πιο εμφανείς τις διαφορές των κλάσεων. Το πλεονέκτημα τους είναι ο παράγοντας εξειδίκευσης που παρέχουν στον αλγόριθμο διευκολύνοντας τον να αναγνωρίζει με διαφορετικούς τρόπους κάθε κλάση. Το μειονέκτημα είναι πως λόγο της εξειδικευμένης τους φύσης μπορεί να μην εφαρμόζονται απόλυτα από όλους τους καταναλωτές ή στην χειρότερη περίπτωση να περιγράφουν μια σπάνια συμπεριφορά που δεν ενδιαφερόμαστε να διαχωρίσουμε.

1. *Κινούμενος μέσος όρος μηνιαίου μέσου όρου* Πρόκειται για υπό συνθήκη χαρακτηριστικό που αν παρατηρήσει κάποια σημαντική πτώση των καταναλώσεων τότε ψάχνει για την μέγιστη και την καταγράφει. Ορίζοντας ως min τον μήνα του ελαχίστου και c την κατανάλωση του αντίστοιχου i μήνα θα έχω την εξής φόρμουλα για αυτό το χαρακτηριστικό.

$$\bar{c_p} - \bar{c_a} = \frac{1}{k-1} \sum_{i=1}^{k} c_{m-i} - \frac{1}{w} \sum_{i=0}^{w} c_{m+i}$$

2. *Κινούμενος μέσος όρος μηνιαίας τυπικής απόκλισης* Πρόκειται για υπό συνθήκη χαρακτηριστικό που αν παρατηρήσει κάποια σημαντική πτώση της τυπικής απόκλισης τότε ψάχνει για την μέγιστη και την καταγράφει. Ορίζοντας ως min τον μήνα του ελαχίστου και std την τυπική απόκλιση της κατανάλωσης τον αντίστοιχο i μήνα θα έχω την εξής φόρμουλα για αυτό το χαρακτηριστικό.

$$\bar{std_p} - \bar{std_a} = \frac{1}{k-1} \sum_{i=1}^{k} std_{m-i} - \frac{1}{w} \sum_{i=0}^{w} std_{m+i}$$

3. *Συμμετρική διαφορά καταναλώσεων* Πρόκειται για υπό συνθήκη χαρακτηριστικό που παρατηρεί μια γενική συμπεριφορά όμοιων καταναλωτών ως προς τη χρονική στιγμή της ελάχιστης κατανάλωσης και ψάχνει για κάποια σημαντική πτώση της κατανάλωσης ανάμεσα σε 2 συμμετρικές χρονικές στιγμές με άξονα συμμετρίας την εκάστοτε χρονική στιγμή ελαχίστου. Ορίζοντας ως min την ημέρα του ελαχίστου και c την κατανάλωση της αντίστοιχης i ημέρα θα έχω τις εξής φόρμουλες εισάγοντας σε αυτό το σημείο και την ευκλείδεια απόσταση.

$$\bar{c_p} - \bar{c_a} = \frac{1}{n} \sum_{i=1}^{n+1} c_{min-i} - \frac{1}{n} \sum_{i=0}^{n} c_{min+i}$$
$$||c_p|| - ||c_a|| = \sqrt{\sum_{i=1}^{n+1} (c_{min-i})^2} - \sqrt{\sum_{i=0}^{n} (c_{min+i})^2}$$

4. *Συμμετρική διαφορά τυπικής απόκλισης* Πρόκειται για υπό συνθήκη χαρακτηριστικό που παρατηρεί μια γενική συμπεριφορά όμοιων καταναλωτών ως προς τη χρονική στιγμή της ελάχιστης κατανάλωσης και ψάχνει για κάποια σημαντική πτώση της τυπικής απόκλισης ανάμεσα σε 2 συμμετρικές χρονικές στιγμές με άξονα συμμετρίας την εκάστοτε χρονική στιγμή ελαχίστου. Ορίζοντας ως min την ημέρα του ελαχίστου και std την την τυπική απόκλιση της κατανάλωσης την αντίστοιχη i ημέρα θα έχω τις εξής φόρμουλα για αυτό το χαρακτηριστικό.

$$\bar{std}_p - \bar{std}_a = \frac{1}{n}\sum_{i=1}^{n+1} std_{min-i} - \frac{1}{n}\sum_{i=0}^{n} std_{min+i}$$
$$||std_p|| - ||std_a|| = \sqrt{\sum_{i=1}^{n+1}(std_{min-i})^2} - \sqrt{\sum_{i=0}^{n}(std_{min+i})^2}$$

5. *Τμηματική διαφορά κατανάλωσης με όμοιους καταναλωτές* Πρόκειται για υπό συνθήκη χαρακτηριστικό που παρατηρεί μια γενική συμπεριφορά όμοιων καταναλωτών ως προς τη χρονική στιγμή της ελάχιστης κατανάλωσης και ψάχνει για κάποια σημαντική πτώση της κατανάλωσης ανάμεσα στον καταναλωτή και τους όμοιούς του μετά την χρονική στιγμή της ελάχιστης κατανάλωσης. Πιο φορμαλιστικά θεωρώντας τους όρους c_{cl} την τυπική κατανάλωση μιας ομάδας και c_{co} την κατανάλωση ενός καταναλωτή έχουμε την παρακάτω διαφορά μέσων όρων και νορμών των καταναλώσεων.

$$\bar{c_{cl}} - \bar{c_{co}} = \frac{1}{n}\sum_{i=1}^{n+1} c_{cl,min-i} - \frac{1}{n}\sum_{i=0}^{n} c_{co,min+i}$$
$$||c_{cl}|| - ||c_{co}|| = \sqrt{\sum_{i=1}^{n+1}(c_{cl,min-i})^2} - \sqrt{\sum_{i=0}^{n}(c_{co,min+i})^2}$$

6. *Τμηματική διαφορά τυπικής απόκλισης με όμοιους καταναλωτές* Πρόκειται για υπό συνθήκη χαρακτηριστικό που παρατηρεί μια γενική συμπεριφορά όμοιων καταναλωτών ως προς τη χρονική στιγμή της ελάχιστης κατανάλωσης και ψάχνει για κάποια σημαντική πτώση της τυπικής απόκλισης ανάμεσα στον καταναλωτή και τους όμοιούς του μετά την χρονική στιγμή της ελάχιστης κατανάλωσης. Πιο φορμαλιστικά θεωρώντας τους όρους std_{cl} την τυπική κατανάλωση μιας ομάδας και std_{co} την κατανάλωση ενός καταναλωτή έχουμε την παρακάτω διαφορά μέσων όρων και νορμών των τυπικών αποκλίσεων των καταναλώσεων.

$$\bar{std}_{cl} - \bar{std}_{co} = \frac{1}{n}\sum_{i=0}^{n} std_{cl,min+i} - \frac{1}{n}\sum_{i=0}^{n} std_{co,min+i}$$
$$||std_{cl}|| - ||std_{co}|| = \sqrt{\sum_{i=0}^{n}(std_{cl,min+i})^2} - \sqrt{\sum_{i=0}^{n}(std_{co,min+i})^2}$$

7. *Χρονική Διαφορά Ελαχίστου* Πρόκειται για υπό συνθήκη χαρακτηριστικό που εξερευνά το ελάχιστο χρονικό σημείο της τάσης της καμπύλης κάθε καταναλωτή. Με βάση την ομάδα που ανήκει κάθε καταναλωτής υπολογίζεται η απόλυτη τιμή της χρονικής διαφοράς μεταξύ του ελαχίστου κάθε καταναλωτή με την ομάδα που ανήκει. Χρησιμοποιώντας ένα όριο για τη διαφορά αυτή γίνεται αντιληπτή οποιαδήποτε έντονη διακύμανση του καταναλωτή με την ομάδα του και καταγράφεται σαν χαρακτηριστικό διαχωρισμού από την αναμενόμενη συμπεριφορά κατανάλωσης.

$$|t_{cl,min} - t_{co,min}|$$

5.1.2 Δοκιμή Χαρακτηριστικών με σταθερή απάτη

Αφού οριστούν τα χαρακτηριστικά που εκτιμάται ότι μπορούν να βοηθήσουν στον διαχωρισμό των κλάσεων, έπεται φυσικά η δοκιμή τους με έναν αφελή τρόπο, έτσι ώστε να επιβεβαιωθεί ότι μπορούν να λειτουργήσουν όπως αναμένεται. Παράλληλα, η δοκιμή αυτή παρέχει μεγάλο όγκο πληροφορίας, αφού καθιστά εμφανή τα σημεία και τις προϋποθέσεις που τα χαρακτηριστικά έχουν μεγάλη ακρίβεια, αλλά και εκεί που υστερούν.

Ο κώδικας της δοκιμής θεωρεί δεδομένη και σταθερή την ένταση κλοπής και την ημέρα που κάθε καταναλωτής ξεκινά να αλλοιώνει τις τιμές του. Ειδικότερα, το ποσοστό των καταναλωτών που αλλοιώνει τις μετρήσεις του είναι 50 τοις εκατό, η ένταση της κλοπής είναι της τάξης του 80 τοις εκατό και η μέρα κλοπής ορίζεται η 182η, δηλαδή μετά από 6 μήνες κανονικής κατανάλωσης ο χρήστης εισάγει σύστημα αλλοίωσης της μέτρησής του. Δοκιμάζοντας ξεχωριστά τα χαρακτηριστικά διαχωρισμού ελέγχουμε το όριο κάθε χαρακτηριστικού έτσι ώστε να δώσει μεγαλύτερη ακρίβεια στις επιθέσεις δεδομένων υπό τις παραπάνω συνθήκες. Αν ο παρατηρηθούν τέτοια χαρακτηριστικά ο καταναλωτής θεωρείται θετικός στην κλοπή. Αντίθετα αν ο καταναλωτής δεν έχει την αναμενόμενη συμπεριφορά το χαρακτηριστικό δεν καταγράφει κάποια τιμή και ο καταναλωτής θεωρείται αρνητικός στην κλοπή. Αναλυτικότερα για κάθε χαρακτηριστικό διαχωρισμού λήφθηκαν τα παρακάτω αποτελέσματα:

1. *Κινούμενος μέσος όρος μηνιαίου μέσου όρου*
   Στην πρώτη δοκιμή δόθηκε έμφαση στη γενικότερη συμπεριφορά του χαρακτηριστικού ως προς το όριο που τίθεται κάθε φορά. Έτσι παρατηρείται εύκολα πως για μεγάλο όριο ο διαχωρισμός έχει χαμηλή ακρίβεια με εξαιρετικά μικρό ποσοστό αποτυχίας. Αντίθετα, αν το όριο χαμηλώσει αισθητά ο χάνεται η έννοια του διαχωρισμού και ο αλγόριθμος θεωρεί θετικούς σε κλοπές σχεδόν όλους τους καταναλωτές.

Όριο	DR	FPR	BDR	Accuracy	F1
0,8	44,8	1,4	0,97	71,7	61,29
0,7	98,7	1,9	0,98	98,4	98,4
0,6	99,3	3,6	0,97	97,85	97,88
0,5	99,8	7,5	0,93	96,15	96,15
0	99,9	91,5	0,52	54,2	68,57

Πίνακας 5.1: Δοκιμή 1ου χαρακτηριστικού

2. *Κινούμενος μέσος όρος μηνιαίας τυπικής απόκλισης*
   Αντίστοιχα και εδώ για παρόμοιες τιμές του ορίου με το προηγούμενο χαρακτηριστικό ο διαχωρισμός είναι εξαιρετικά εύστοχος και δεν αφήνει περιθώρια για αμφισβήτηση.

Όριο	DR	FPR	BDR	Accuracy	F1
0,7	98,2	2,3	0,98	98,3	98,31
0,6	99,8	4,1	0,96	97,85	97,89
0,5	99,5	8,2	0,92	95,65	95,81

Πίνακας 5.2: Δοκιμή 2ου χαρακτηριστικού

3. *Συμμετρική διαφορά καταναλώσεων*
   Το συγκεκριμένο χαρακτηριστικό δεν δίνει αξιόπιστα αποτελέσματα, καθώς η συμμετρία που προκύπτει από τον χρησιμοποιούμενο τύπο απάτης κάνει το συγκεκριμένο χαρακτηριστικό να αποτυγχάνει σε αυτή τη δοκιμή. Παρόλα αυτά, το χαμηλό ποσοστό αποτυχίας αφήνει δεύτερες σκέψεις, καθώς δεν επιβαρύνει αισθητά τα αποτελέσματα, αλλά βοηθά στη γενίκευση του τύπου κλοπής.

Όριο	DR	FPR	BDR	Accuracy	F1
0,1	26,3	5,7	0,82	60,3	39,85

Πίνακας 5.3: Δοκιμή 3ου χαρακτηριστικού

Η δοκιμή συνεχίστηκε και με τις νόρμες των καταναλώσεων, παρατηρώντας ελάχιστη βελτίωση στο DR, ενώ αισθητά καλύτερα αποτελέσματα παρατηρούνται στο FPR που μειώθηκε ακόμη περισσότερο.

Όριο	DR	FPR	BDR	Accuracy	F1
0,3	29	2,1	0,93	63,65	44,71

Πίνακας 5.4: Δοκιμή 3ου χαρακτηριστικού με νόρμες

4. *Συμμετρική διαφορά τυπικής απόκλισης*
   Αντίστοιχα συμπεράσματα ισχύουν και στη συμμετρική διαφορά τυπικής απόκλισης που οριακά ξεπερνά το 10 τοις εκατό στο FPR. Η γενίκευση που προσφέρει παρόλα αυτά το συγκεκριμένο χαρακτηριστικό είναι χρήσιμη, καθώς εν τέλει όλα τα χαρακτηριστικά θα ενώσουν τα δυνατά τους σημεία για να διαχωρίσουν απάτες με μεγαλύτερο τυχαίο παράγοντα.

Όριο	DR	FPR	BDR	Accuracy	F1
0,1	38,9	10,2	0,79	64,35	52,18

Πίνακας 5.5: Δοκιμή 4ου χαρακτηριστικού

Σε αυτό το σημείο η μείωση του FPR είναι αρκετά σημαντικό ζήτημα που τελικώς επιτεύχθει με τις νόρμες που μπόρεσαν να μειώσουν το FPR χωρίς να επηρεάσουν αρνητικά το DR.

Όριο	DR	FPR	BDR	Accuracy	F1
0,1	40,2	8,9	0,82	65,65	53,92

Πίνακας 5.6: Δοκιμή 4ου χαρακτηριστικού με νόρμες

5. *Τμηματική διαφορά κατανάλωσης με όμοιους καταναλωτές*
Σε αυτό το χαρακτηριστικό παρατηρείται σχετική αστοχία σε σχέση με τα πρώτα χαρακτηριστικά υποδεικνύοντας ανάγκη για καλύτερη ρύθμιση του χαρακτηριστικού.

Όριο	DR	FPR	BDR	Accuracy	F1
0,1	98,4	11,4	0,9	93,5	93,8
0,2	93,9	7	0,93	93,45	93,48
0,3	88,3	4,9	0,95	91,7	91,41

Πίνακας 5.7: Δοκιμή 5ου χαρακτηριστικού

Δεδομένης της διαφοράς των καταναλώσεων με την γενικευμένη κατανάλωση μιας ομάδας δημιουργείται η ανάγκη για κανονικοποίηση σε κάθε διάνυσμα κατανάλωσης. Με αυτό τον τρόπο επιτυγχάνονται πολύ καλύτερα αποτελέσματα.

Όριο	DR	FPR	BDR	Accuracy	F1
0,3	98,9	5,4	0,95	96,75	96,82

Πίνακας 5.8: Δοκιμή 5ου χαρακτηριστικού με κανονικοποίηση

Όριο	DR	FPR	BDR	Accuracy	F1
0,1	98,7	7	0,93	95,85	95,96
0,2	97,6	4,4	0,96	96,6	96,63

Πίνακας 5.9: Δοκιμή 5ου χαρακτηριστικού με κανονικοποίηση και νόρμες

6. *Τμηματική διαφορά τυπικής απόκλισης με όμοιους καταναλωτές*
Αντίστοιχη μεθοδολογία εφαρμόστηκε και σε αυτό το χαρακτηριστικό. Τα αποτελέσμα-

τα ήταν ικανοποιητικά, αλλά όχι αρκετά. Έτσι χρησιμοποιήθηκε κανονικοποίηση για μπορέσει να μειωθεί το FPR, ενώ αυξάνεται το DR.

Όριο	DR	FPR	BDR	Accuracy	F1
0,1	99,4	16,1	0,86	91,65	92,25
0,2	95,7	8,3	0,92	93,7	93,82
0,3	89,8	6,2	0,94	91,8	91,63

Πίνακας 5.10: Δοκιμή 6ου χαρακτηριστικού

Όριο	DR	FPR	BDR	Accuracy	F1
0,4	97	4,9	0,95	96,05	96,09
0,3	96,3	5,3	0,95	95,5	95,54

Πίνακας 5.11: Δοκιμή 6ου χαρακτηριστικού με κανονικοποίηση

Όριο	DR	FPR	BDR	Accuracy	F1
0,2	98,5	4,5	0,96	97	97,04
0,1	99,2	5,3	0,95	96,95	97,02

Πίνακας 5.12: Δοκιμή 6ου χαρακτηριστικού με κανονικοποίηση και νόρμες

7. *Χρονική Διαφορά Ελαχίστου*

Δοκιμάζοντας το μοναδικό χαρακτηριστικό που σχετίζεται με χρόνο και όχι με κατανάλωση καθίσταται σαφές πως δεν δίνει περισσότερη διακριτική ικανότητα στις κλάσεις. Αντίθετα, παρέχει μεγάλη γενικότητα στον αλγόριθμο δίνοντας μια ακόμη πληροφορία για τη συμπεριφορά κατανάλωσης, αλλά λόγω του αισθητά μεγάλου FPR αποτυγχάνει να διαχωρίσει.

Όριο	DR	FPR	BDR	Accuracy	F1
0,1	85,4	94,7	0,47	45,35	60,98
0,2	74,9	81,7	0,48	46,6	58,38
0,3	18,9	56,4	0,25	31,25	21,56
0,4	13,7	34	0,29	39,85	18,55

Πίνακας 5.13: Δοκιμή 7ου χαρακτηριστικού με κανονικοποίηση

5.1.3 Δοκιμή Χαρακτηριστικών με μεταβλητή απάτη

Τέλος έγινε μια ακόμη τελική δοκιμή στα χαρακτηριστικά, αυτή τη φορά με μεγαλύτερο τυχαίο παράγοντα. Η ένταση της κλοπής καθορίζεται από μια βήτα κατανομή με παραμέτρους 6 και 3, ενώ η ημέρα που ξεκινά η κλοπή επιλέγεται από μια κανονική κατανομή με παραμέτρους 182.5 και 56.1538. Σε κάθε καταναλωτή που έχει επιλεχθεί για κλοπή επιβάλλονται διαφορετικές τιμές των παραπάνω κατανομών κρατώντας όμως το γενικότερο πλαίσιο του τύπου της κλοπής που είδαμε προηγουμένως. Αυτό που μας ενδιαφέρει να δούμε σε αυτό το σημείο είναι χαμηλό FPR καθώς αναμένεται να χαμηλώσει σημαντικά η ακρίβεια, λόγο της απλότητας του κριτηρίου μας.

Χαρακτ.	Όριο	DR	FPR	BDR	Accuracy	F1
1	0,7	42,8	2,1	0,95	70,35	59,08
2	0,7	46,5	1,8	0,96	72,35	62,71
3	0,1	58	9,9	0,85	74,05	69,09
4	0,1	59,6	9,4	0,86	75,1	70,53
5	0,3	66,4	8,2	0,89	79,1	76,06
6	0,4	58,4	5,6	0,91	76,4	71,22
7	0,3	48,8	39,8	0,55	54,5	51,75

Πίνακας 5.14: Δοκιμή χαρακτηριστικών με τυχαίο παράγοντα

Δοκιμή χαρακτηριστικών με τυχαίο παράγοντα και νόρμες						
Χαρακτ.	Όριο	DR	FPR	BDR	Accuracy	F1
3	0,3	47,9	2,9	0,95	72,65	63,65
4	0,1	64,7	10,5	0,86	77,1	73,86
5	0,1	77,5	8,8	0,9	84,35	83,2
6	0,1	75,7	7,9	0,91	83,9	82,46

Πίνακας 5.15: Δοκιμή χαρακτηριστικών με τυχαίο παράγοντα και νόρμες

Καθίσταται λοιπόν σαφές πως τα χαρακτηριστικά σε γενικές γραμμές έχουν χαμηλότερη ακρίβεια, αλλά κρατούν χαμηλό FPR κάτι που μας ενδιαφέρει περισσότερο σε αυτό το σημείο. Παράλληλα, τα χαρακτηριστικά 3 και 4 που είχαν απογοητευτικά αποτελέσματα στην προηγούμενη δοκιμή, σε αυτήν δείχνουν να βελτιώνονται αισθητά σε σχέση με την επίδοση των υπόλοιπων. Αυτό μας πληροφορεί ότι η αρχική υπόθεση μας για το λόγο αστοχία τους

ήταν αληθής. Παράλληλα, το χαρακτηριστικό 7 που είχε επίσης εξαιρετικά αποθαρρυντικά αποτελέσματα στην προηγούμενη δοκιμή εξισορροπείται η σχέση μεταξύ του DR και FPR, παρόλο που ακόμη φαίνεται το χαρακτηριστικό με τα χειρότερα αποτελέσματα.

5.2 Αλγόριθμοι συσταδοποίησης

Η συσταδοποίηση είναι από τους δημοφιλέστερους τύπου μη-επιβλεπόμενης εκμάθησης. Σε αυτό τον τύπο εκμάθησης, ο στόχος δεν είναι η μεγιστοποίηση μιας συνάρτησης, αλλά είναι απλώς η εύρεση των ομοιοτήτων των δεδομένων. Η υπόθεση είναι συνήθως πως οι συστάδες που ανακαλύπτονται θα ταιριάξουν σχετικά καλά με τη διαισθητική ταξινόμηση [10]. Ειδικότερα ένα σύνολο παρατηρήσεων (σημείων δεδομένων) διαμερίζεται σε φυσικές ομαδοποιήσεις, ή συστάδες (clusters), προτύπως με τέτοιο τρόπο ώστε το μέτρο ομοιότητας μεταξύ οποιουδήποτε ζεύγους παρατηρήσεων αντιστοιχίζεται σε κάθε συστάδα να ελαχιστοποιεί μια καθορισμένη συνάρτηση κόστους.

5.2.1 K-Means

Δοθέντος ενός συνόλου N παρατηρήσεων, ζητείται ο κωδικοποιητής C που αντιστοιχίζει αυτές τις παρατηρήσεις στις K συστάδες με τέτοιο τρόπο ώστε, μέσα σε μια συστάδα, ο μέσος όρος του μέτρου ανομοιότητας των αντιστοιχισμένων παρατηρήσεων ως προς το μέσο της συστάδα ελαχιστοποιείται μέσω της συνάρτησης κόστους.

$$J(C) = \sum_{j=1}^{K} \sum_{C(i)=j} \|\mathbf{x}_i - \bar{\mu_j}\|^2$$

Με μαθηματικούς όρους, ο αλγόριθμος (K-Means) εξελίσσεται σε δύο βήματα:

1. Για ένα δεδομένο κωδικοποιητή C, η συνολική διακύμανση συστάδας ελαχιστοποιείται ως προς το σύνολο μέσων συστάδας $\{\bar{\mu_j}\}_{j=i}^{K}$, δηλαδή εκτελούμε την ακόλουθη ελαχιστοποίηση:

$$min_{\{\bar{\mu_j}\}_{j=i}^{K}} \sum_{j=i}^{K} \sum_{C(i)=j} \|\mathbf{x}_i - \bar{\mu_j}\|^2 \text{ για δεδομένο } C$$

2. Αφού υπολογιστούν οι βελτιστοποιημένοι μέσοι συστάδας $\{\bar{\mu_j}\}_{j=i}^{K}$, στη συνέχεια βελτιστοποιούμε τον κωδικοποιητή ως εξής:

$$C(i) = argmin_{1 \leq j \leq K} \|\mathbf{x}_i - \bar{\mu_j}\|^2$$

Ξεκινώντας από κάποια αρχική επιλογή κωδικοποιητή ", ο αλγόριθμος εναλλάσσεται μεταξύ αυτών των δυο βημάτων μέχρι να μην υπάρξει περαιτέρω αλλαγή στις αντιστοιχίσεις των συστάδων[29].

5.2.2 Fuzzy C-Means

Ο αλγόριθμος ασαφών C μέσων (FCM) είναι πολύ κοντά στη λογική του K-Means, αλλά εισάγει μια πιο πιθανοκρατική προσέγγιση. Επιλύσει το πρόβλημα της ελαχιστοποίησης των αποστάσεων μέσα σε μια συστάδα και της μεγιστοποίησης των αποστάσεων μεταξύ των συστάδων με βάση το παρακάτω κριτήριο βελτιστοποίησης[;]:

$$J_m = \sum_{k=1}^{N} \sum_{i=1}^{n} (u_{ik})^m \|\mathbf{x}_k - v_i\|^2$$

Έστω ένα σύνολο διανυσμάτων δεδομένων $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_3\}$ με $x_k \in \mathbb{R}^p$ $(1 \leq k \leq n)$, τα οποία ομαδοποιούνται σε ασαφείς συστάδες.

- Επιλογή αριθμού c των ασαφών συστάδων, της παραμέτρου m, των αρχικών τιμών για τα διανύσματα $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_c$ και της παραμέτρου c.

$$\text{όπου } \mathbf{v}_i = \frac{\sum_{i=1}^{n} (u_{ik})^m x_{ik}}{\sum_{i=1}^{n} (u_{ik}}$$

- Χρήση της παρακάτω εξίσωσης για τον υπολογισμό των συναρτήσεων συμμετοχής u_{ik}

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{\|\mathbf{x}_k - \mathbf{v}_i\|}{\|mathbf{x}_k - \mathbf{v}_j\|})}, \ (1 \leq k \leq n), \ (1 \leq i \leq c)$$

- Βάση της εξίσωσης του βήματος 1 γίνεται προσδιορισμός των νέων τιμών για τα κέντρα των ασαφών υποομάδων $\mathbf{v}_1^{new}, \mathbf{v}_2^{new}, ..., \mathbf{v}_c^{new}$

- Αν $max_i\{\|\mathbf{v}_i - \mathbf{v}_i^{new}\|^2\} < \epsilon$ τότε ο αλγόριθμος σταματάει, αλλιώς θέτει $\mathbf{v}_i = \mathbf{v}_i^{new}$ και η ροή του πηγαίνει στο βήμα 2.

5.2.3 SOM

Εμπνευσμένο από τα νευρωνικά δίκτυα στον εγκέφαλο, οι αυτο-οργανωμένοι χάρτες (SOM) χρησιμοποιούν ένα μηχανισμό ανταγωνισμού και συνεργασίας για να πετύχουν μη-επιβλεπόμενη εκμάθηση. Στην κλασσική περίπτωση του SOM, ένα μέρος από κόμβους οργανώνεται σε γεωμετρικό σχήμα, συνήθως δισδιάστατο πλέγμα. Κάθε κόμβος είναι σχετίζεται με ένα διάνυσμα βάρους με τις ίδιες διαστάσεις όπως η είσοδος. Ο σκοπός του SOM είναι να βρει μια χαρτογράφηση από τον υψηλό χώρο διαστάσεων της εισόδου σε δισδιάσταστη αναπαράσταση των κόμβων. Ένας τρόπος να χρησιμοποιηθεί για συσταδοποίηση είναι να αναμένεται τα αντικείμενα στο χώρο εισόδου να αναπαριστούν από τον ίδιο κόμβο όπως σχηματίστηκαν στη συστάδα. Στη διάρκεια της εκπαίδευσης, κάθε αντικείμενο στην είσοδο αναπαριστάται στο χάρτη και αναγνωρίζεται ο κόμβος που ταιριάζει βέλτιστα. Τυπικά, όταν η είσοδος και τα διανύσματα βαρών κανονικοποιηθούν, για δείγμα εισόδου $x(t)$ ο νικητής δείκτης c ορίζεται κάτω από τη συνθήκη:

$$\text{για όλα } i, \ \|x(t) - m_c(t)\| \leq \|x(t) - m_i(t)\|$$

όπου t είναι το χρονικό βήμα στην εκπαίδευση, m_i είναι το διάνυσμα βάρους του i κόμβου. Μετά από αυτό, το διάνυσμα βάρους γύρω από τον βέλτιστο κόμβο $c = c(x)$ ανανεώνεται ως εξής:

$$m_i(t+1) = m_i(t) + \alpha h_{c(x),i}(x(t) - m_i(t))$$

όπου α είναι ο ρυθμός εκμάθησης και $h_{c(x),i}$ είναι η «γειτονική συνάρτηση», μια φθίνουσα συνάρτηση της συνάρτησης μεταξή i και c κόμβων στο δίκτυο του χάρτη [;].

5.3 Συστατικά αλγορίθμου μη-επιβλεπόμενης μάθησης

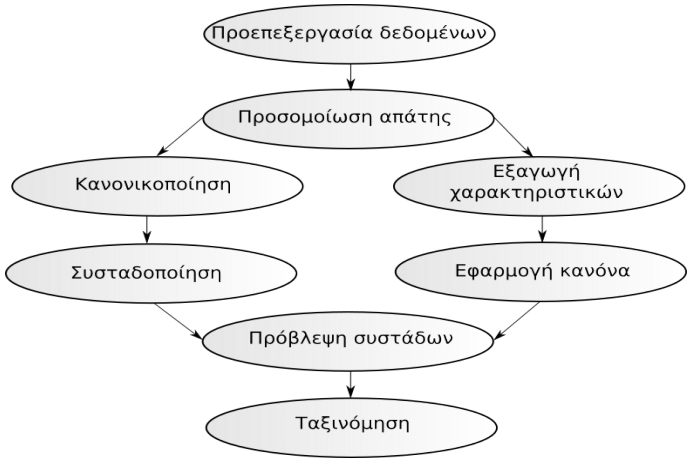
Οι δοκιμές στην επιβλεπόμενη μάθηση έδειξαν πως η ταξινόμηση των παρούσων χρονοσειρών δεν είναι εύκολη διαδικασία. Για αυτό το λόγο χρησιμοποιήθηκε συστοιχία μη-επιβλεπόμενων αλγορίθμων για την ταξινόμηση των καταναλωτών. Ειδικότερα, εισήχθηκε ένα σύστημα με ταξινόμηση βάση κανόνων, το οποίο συσταδοποιεί τα δεδομένα και εξάγει χαρακτηριστικά των χρονοσειρών και αποτελέσματα λαμβάνοντας υπόψη τα παραπάνω. Η βέλτιστη δομή του συστήματος επιτεύχθηκε, όπως φαίνεται στο Σχήμα 5.1.

Παρέχοντας περαιτέρω πληροφορίες για τα μέλη που απαρτίζουν το συστήμα προς έρευνα, έχουμε:

- *Προεπεξεργασία δεδομένων: Επιλέγονται και οργανώνονται τα δεδομένα σε συγκεκριμένους πίνακες και διανύσματα.*
- Προσομοίωση απάτης: Αλλοιώνονται οι μετρήσεις κάποιων καταναλωτών και ενημερώνονται οι προϋπάρχοντες πίνακες και διανύσματα.
- *Κανονικοποίηση: Κανονικοποιούνται οι ετήσιες χρονοσειρές κάθε καταναλωτή σε εύρος τιμών [-1,1].*
- Συσταδοποίηση: Συσταδοποιούνται οι καταναλωτές με βάση τις κανονικοποιημένες τιμές σε δύο συστάδες. Η μια συστάδα ομαλή και η άλλη η ανώμαλη.
- *Εξαγωγή χαρακτηριστικών: Βάση των χρονοσειρών δημιουργούνται ετήσια χαρακτηριστικά για κάθε καταναλωτή, προσπαθώντας να ανιχνευθεί ύποπτη συμπεριφορά.*
- Εφαρμογή κανόνα: Λαμβάνοντας υπόψη το πλήθος των χαρακτηριστικών απενοχοποιούνται κάποιοι καταναλωτές που βρίσκονται στην ανώμαλη συστάδα.
- *Πρόβλεψη συστάδων: Θέτονται δυαδικά χαρακτηριστικά στις συστάδες με σεβασμό στον κανόνα.*
- Ταξινόμηση: Ταξινομούνται οι καταναλωτές και παράγονται τα τελικά αποτελέσματα και μετρικές.

5.3.1 Μεθοδολογία εξαγωγής αποτελεσμάτων

Η εξαγωγή αποτελεσμάτων παίζει μεγάλο ρόλο στην τελική απόδοση του αλγορίθμου, οπότε χρειάζεται ιδιαίτερη προσοχή η τοποθέτηση δυαδικών χαρακτηριστικών. Η γενικότερη μεθοδολογία βασίζεται σε δύο σημαντικούς άξονες, καθώς η τομή των δύο είναι αυτή που



Σχήμα 5.1: Δομή μη-επιβλεπόμενου ταξινομητή

εξάγει τα βέλτιστα αποτελέσματα. Αυτό γίνεται ξεκάθαρο παρατηρώντας τον πίνακα αποτελεσμάτων 5.17.

Ο πρώτος άξονας αποτελείται από την κανονικοποίηση και την συσταδοποίηση. Κατά την διαδικασία της κανονικοποιήση ο πίνακας με τις χρονοσειρές αναστρέφεται, κανονικοποιείται και αναστρέφεται για δεύτερη φορά για να αποκτήσει την ίδια μορφή με την αρχική, αλλά με εύρος τιμών [-1,1]. Με αυτό τον τρόπο δίνεται έμφαση στη μορφή και όχι στα μεγέθη των χρονοσειρών. Έτσι, εκμεταλλεύεται το γεγονός ότι οι χρονοσειρές είναι αρκετά ομοιόμορφες ως προς το σχήμα. Σε επόμενη φάση εκτελείται συσταδοποίηση στα κανονικοποιημένα μεγέθη και διαχωρίζονται οι καταναλωτές σε μια μεγάλη συστάδα με αναμενόμενες μορφές και σε μια μικρή συστάδα με ακανόνιστες συμπεριφορές.

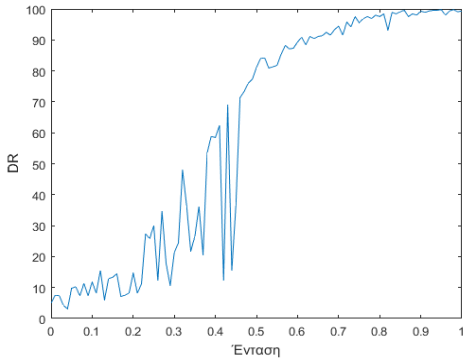
Ο δεύτερος άξονας αποτελείται από την εξαγωγή χαρακτηριστικών των χρονοσειρών και την εφαρμοφή του κανόνα. Η εξαγωγή χαρακτηριστικών δίνει τη δυνατότητα μέσω των χαρακτηριστικών διαχωρισμού να δημιουργηθούν ομάδες καταναλωτών που έχουν ύποπτες και αναμενόμενες μετρήσεις. Αν έχουμε έλλειψη χαρακτηριστικών, δηλαδή 0, πρακτικά σημαίνει πως ο καταναλωτής έχει αναμενόμενη συμπεριφορά. Στην αντίθετη περίπτωση ο καταναλωτής έχει αποκλίνουσα συμπεριφορά και θεωρείται ύποπτος. Εκεί έρχεται ο κανόνας που ορίζει πως αν ο καταναλωτής έχει λιγότερες από τρεις μετρήσεις στα χαρακτηριστικά διαχωρισμού ορίζεται αναμενόμενη συμπεριφορά.

Κανόνας	DR	FPR	Accuracy	F1 score	BDR
Συσταδ.	98.67	34.2	69.09	38.96	0.24
Χαρακτ.	87.78	6.72	92.73	70.73	0.59
Συνδ.	89.33	5.93	93.6	73.63	0.63

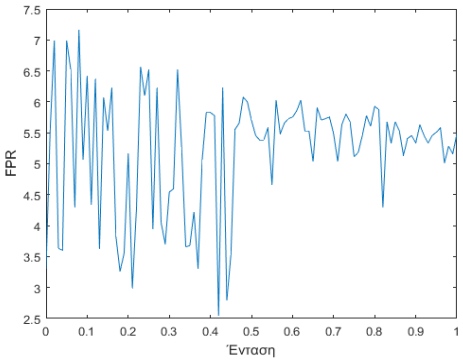
Πίνακας 5.16: Δοκιμή στους κανόνες

5.4 Δοκιμή αλγορίθμου μη επιβλεπόμενης μάθησης

Για να επιβεβαιωθεί η ορθή και βέλτιστη λειτουργία του συστήματος απαιτείται δοκιμή των παραμέτρων που το απαρτίζουν. Για να συμβεί αυτό επιλέχθηκαν 4.500 καταναλωτές και αλλοιώθηκαν τα δεδομένα μόνο του 10%. Ο τύπος απάτης που χρησιμοποιήθηκε για την αλλοίωση των δεδομένων είναι ο πρώτος, καθώς φάνηκε πως είναι ιδιαίτερα πολύπλοκο ακόμη και για επιβλεπόμενο σύστημα να παράξει αξιόπιστα αποτελέσματα.



(α′) DR συναρτήσει της έντασης της κλοπής



(β′) FPR συναρτήσει της έντασης της κλοπής

Σχήμα 5.2: Επίπτωση της έντασης στα αποτελέσματα

μεγάλα πλήγματα στην επίδοση του συστήματος που προδίδουν πως υπό κάποιες συνθήκες το σύστημα δυσκολεύεται να ορίσει την κλοπή, χωρίς όμως να ενοχοποιεί αδίκως.

Παράλληλα, αξίζει να σημειωθεί εδώ πως οι αλγόριθμοι συσταδοποίησης μπορούν να αλλάξουν σε κάποιο βαθμό τα χαρακτηριστικά του συστήματος και την επίδοσή του. Σαν αποτέλεσμα δημιουργήθηκε δοκιμή για τους διαφορετικούς αλγορίθμους που χρησιμοποιήθηκαν στην εξαγωγή των χαρακτηριστικών, καταλήγοντας σε αποτελέσματα για κάθε περίπτωση.

Αλγ.	DR	FPR	Accuracy	F1 score	BDR
K-Means	86.44	5.43	93.76	73.47	0.64
SOM	89.11	5.23	94.2	75.45	0.65
Fuzzy	85.78	4.99	94.09	74.37	0.66

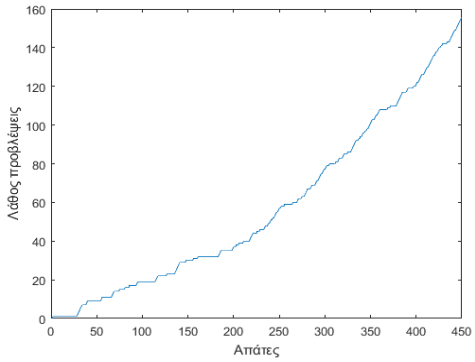
Πίνακας 5.17: Δοκιμή στους κανόνες

Γίνεται, λοιπόν αντιληπτό πως οι αλγόριθμοι συσταδοποίησης στην εξαγωγή δεδομένων παίζουν σχετικά μικρό ρόλο, αφού τα αποτελέσματα έχουν πολύ μικρές αποκλίσεις μεταξύ τους. Αυτό ήταν κάτι αναμενόμενο βέβαια καθώς μόνο δύο από τα οκτώ χαρακτηριστικά έχουν άμεση συσχέτιση με τη συσταδοποίηση.

5.4.2 Εξερεύνηση δυνατοτήτων FCM

Ο αλγόριθμος ασαφών κ μέσων μέσα από τον παράγοντα ασάφιας δίνει τη δυνατότητα να εξερευνηθούν οι συστάδες και με διαφορετικούς τρόπους. Ειδικότερα, ο παράγοντας αυτός καθορίζει την επικάλυψη των συστάδων. Παράλληλα, για να μπορέσει να διευκρινιστεί τελικά που ανήκει κάθε παράδειγμα παρέχεται μια τιμή για κάθε συστάδα με τη μεγαλύτερη από αυτή να υποδηλώνει μεγάλο βαθμό ομοιότητας του παραδείγματος με τη συστάδα.

Με αυτό το σκεπτικό δημιουργήθηκε μια δοκιμή κατά την οποία χωρίς εξαγωγή χαρακτηριστικών ταξινομούνται οι καταναλωτές. Ειδικότερα, γνωρίζοντας το ποσοστό των απατών τίθεται ένα όριο στο πλήθος που επιθυμεί κάποιος να ελέγξει. Ο αλγόριθμος βάση αυτού του πλήθους επιλέγει το δείγμα των καταναλωτών που φαίνεται πιο σίγουρο ότι ανήκει στη συστάδα με ακανόνιστες μετρήσεις. Στην πράξη αν από 450 κλοπές, τεθεί ένα όριο στην εύρεση μόνο των 100, ο αλγόριθμός έχει τη δυνατότητα να αναγνωρίσει σωστά 81, ενώ λάθος 19 όπως φαίνεται και στο παρακάτω σχήμα.

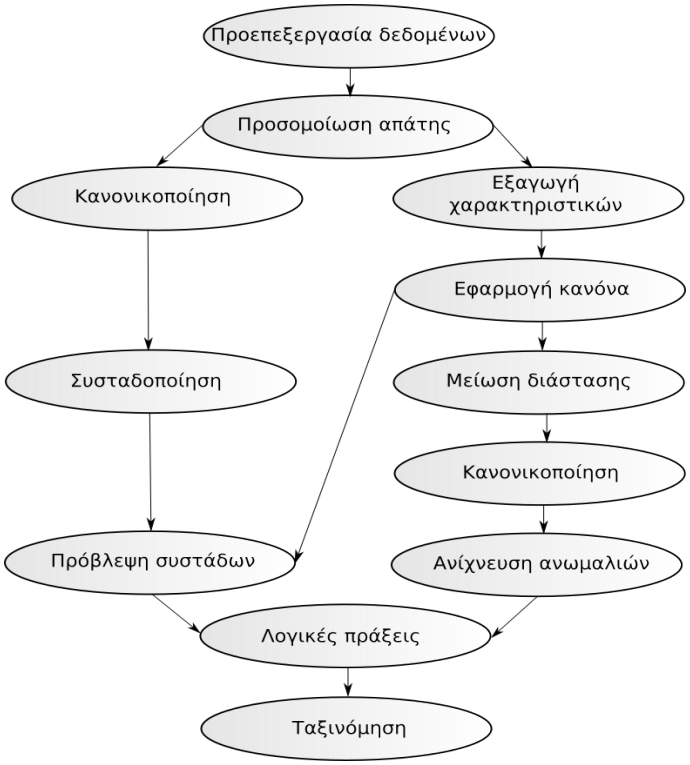


Σχήμα 5.3: Καμπύλη λάθος προβλέψεων με FCM

ταξινομούνται οι καταναλωτές. Παράλληλα, η προσθήκη νέων αλγορίθμων δίνει τη δυνατότητα εποπτείας των χαρακτηριστικών, αλλά και του μοντέλου που δημιουργήθηκε σε δισδιάστατο χώρο. Οπτικοποιούνται λοιπόν οι πληροφορίες και η εσωτερική λειτουργία του αλγορίθμου, ενώ παράλληλα δίνεται παρέχεται η δυνατότητα εκπαίδευσης προτύπων.

Αναλυτικότερα η δομή του αλγορίθμου αναπαρίσταται στο Σχήμα 5.4 ενώ αξίζει να γίνει μια εισαγωγή στα κομμάτια που απαρτίζουν το σύστημα:

- *Προεπεξεργασία δεδομένων: Επιλέγονται και οργανώνονται τα δεδομένα σε συγκεκριμένους πίνακες και διανύσματα.*
- Προσομοίωση απάτης: Αλλοιώνονται οι μετρήσεις κάποιων καταναλωτών και ενημερώνονται οι προϋπάρχοντες πίνακες και διανύσματα.
- *Κανονικοποίηση: Κανονικοποιούνται οι ετήσιες χρονοσειρές και τα χαρακτηριστικά κάθε καταναλωτή σε εύρος τιμών [-1,1] και [0,1] αντίστοιχα.*
- Συσταδοποίηση: Συσταδοποιούνται οι καταναλωτές με βάση τις κανονικοποιημένες τιμές σε δύο συστάδες. Η μια συστάδα ομαλή και η άλλη η ανώμαλη.
- *Εξαγωγή χαρακτηριστικών: Βάση των χρονοσειρών δημιουργούνται ετήσια χαρακτηριστικά για κάθε καταναλωτή, προσπαθώντας να ανιχνευθεί ύποπτη συμπεριφορά.*
- Εφαρμογή κανόνα: Λαμβάνοντας υπόψη το πλήθος των χαρακτηριστικών απενοχοποιούνται κάποιοι καταναλωτές που βρίσκονται στην ανώμαλη συστάδα.
- *Πρόβλεψη συστάδων: Θέτονται δυαδικά χαρακτηριστικά στις συστάδες με σεβασμό στον κανόνα.*
- Μείωση διάστασης: Ο πολυδιάστατος χώρος των χαρακτηριστικών μειώνεται σε δισδιάστατο.
- *Ανίχνευση ανωμαλιών: Εκπαιδεύεται το μοντέλο πρόβλεψης βάση των χαρακτηριστικών και βελτιστοποιούνται τα όρια ταξινόμησης.*



Σχήμα 5.4: Δομή ημί-επιβλεπόμενου ταξινομητή

- Λογικές πράξεις: Εκτελούνται λογικές πράξεις μεταξύ των δυαδικών χαρακτηριστικών που προέρχονται από την πρόβλεψη συστάδων και την ανίχνευση ανωμαλιών.
- *Ταξινόμηση: Ταξινομούνται οι καταναλωτές και παράγονται τα τελικά αποτελέσματα και μετρικές.*

5.5.1 Εφαρμογή αλγορίθμου μείωσης διάστασης

Το PCA είναι ένας μη-επιβλεπόμενος αλγόριθμος γραμμικής μείωσης διάστασης που στοχεύει στην εύρεση μιας βάσης ή ενός συστήματος συντεταγμένων με περισσότερο νόημα για τα δεδομένα και λειτουργεί βάση του πίνακα συνδιακύμανσης για την εύρεση ισχυρών χαρακτηριστικών.

Χρησιμοποιείται όταν χρειάζεται να αντιμετωπιστούν οι δυσκολίες των διαστάσεων σε δεδομένα με γραμμικές σχέσεις, καθώς το μεγάλο νούμερο διαστάσεων (χαρακτηριστικών) μπορεί να δημιουργήσει θόρυβο. Το φαινόμενο αυτό επιδεινώνεται όταν τα χαρακτηριστικά έχουν διαφορετικές κλίμακες.

Αυτό επιτυγχάνεται μειώνοντας διάσταση δηλαδή χαρακτηριστικά. Αλλά πότε πρέπει να μειώσουμε ή να αλλάξουμε διάσταση;

- *Καλύτερη εποπτεία και μικρότερη πολυπλοκότητα* Όταν απαιτείται μια πιο ρεαλιστική εποπτεία των διαστάσεων και υπάρχουν πολλά χαρακτηριστικά σε ένα σετ δεδομένων και ειδικότερα όταν υπάρχει διαισθητική γνώση πως δεν απαιτούνται πολλά χαρακτηριστικά.

- *Καλύτερη οπτικοποίηση* Όταν είναι αδύνατο να έχουμε καλή οπτικοποίηση λόγω του πλήθους των διαστάσεων χρησιμοποιείται PCA για να μειωθεί σε μια σκιά με δύο ή τρεις διαστάσεις.

- *Μείωση μεγέθους* Όταν υπάρχει μεγάλος όγκος δεδομένων και σκοπεύεται να χρησιμοποιηθούν χρονοβόροι αλγόριθμοι στα δεδομένα χρειάζεται να ελαχιστοποιηθούν οι πλεονασμοί.

- *Διαφορετική οπτική* Όταν υποβόσκει ανάγκη να αυξηθεί η γνώση πάνω στα δεδομένα. Το PCA μπορεί να δώσει τους καλύτερους γραμμικά ανεξάρτητους και διαφορετικούς συνδιασμούς χαρακτηριστικών, ώστε να περιγραφούν διαφορετικά τα δεδομένα.

Η πρακτική υλοποίηση του PCA είναι εύκολη και συνοψίζεται σε τρία βήματα [;]:

1. Οργάνωση των δεδομένων σε πίνακα $m \times n$, όπου m είναι ο αριθμός των μετρήσεων (χαρακτηριστικών) και n ο αριθμός των δοκιμών.

2. Αφαίρεση του μέσου όρου από κάθε μέτρηση ή από κάθε σειρά.

3. Υπολογισμός SVD των ιδιοδιανυσμάτων της συνδιακύμανσης.

Η συνδιακύμανση μεταξύ δύο χαρακτηριστικών υπολογίζεται ως εξή:

$$\sigma_{jk} = \frac{1}{n-1} \sum_{i=1}^{N} (x_{ij} - \bar{x_j})(x_{ik} - \bar{x_k})$$

Η παραπάνω μπορεί να γενικευθεί σε υπολογισμό του πίνακα συνδιακύμανσης με την ακόλουθη εξίσωση πινάκων:

$$\Sigma = \frac{1}{n-1}((X - \bar{x})^T (X - \bar{x}))$$
$$\text{όπου } \bar{x} \text{ είναι το διάνυσμα του μέσου όρου } \bar{x} = \sum_{k=1}^{n} x_i$$

Υπάρχουν τρεις προσεγγίσεις οι οποίες αποδίδουν τα ίδια ιδιοδιανύσματα και ζευγάρια ιδιοτιμών:

- Ιδιοπαραγοντοποίηση του πίνακα συνδιακύμανσης μετά από κανονικοποίηση δεδομένων.

- Ιδιοπαραγοντοποίηση του πίνακα συσχέτισης.

- Ιδιοπαραγοντοποίηση του πίνακα συσχέτισης μετά από κανονικοποίηση δεδομένων.

Στην παρούσα εργασία παρόλα αυτά χρησιμοποιείται παραγοντοποίηση ιδιόμορφων ιδιοδιανυσμάτων (SVD) για τη βελτίωση τη υπολογιστική επίδοση [;].

5.5.2 Εφαρμογή αλγορίθμου ανίχνευσης ανωμαλιών

Ο αλγόριθμος που χρησιμοποιήθηκε για την ανίχνευση ανωμαλιών είναι βασισμένος στο Γκαουσιανό μοντέλο. Τέτοιες τεχνικές υποθέτουν πως τα δεδομένα δημιουργούνται από μια Γκαουσιανή κατανομή. Οι παράμετροι υπολογίζονται με εκτιμητές μέγιστης πιθανοφάνειας

(MLE). Η απόσταση ενός παραδείγματος από το εκτιμώμενο μέσο είναι το αποτέλεσμα του ποσοστού ανωμαλίας. Ορίζεται ένα όριο στα ποσοστά αυτά για να οριστούν οι ανωμαλίες [27].

Ερμηνεύοντας αυτή την τεχνική πιο φορμαλιστικά θεωρούνται χαρακτηριστικά x_i που υποδεικνύουν ανώμαλα παραδείγματα. Για m παραδείγματα εκπαίδευσης και n χαρακτηριστικά ορίζονται τα δεδομένα εξόδου $\{x^{(1)}, ..., x^{(m)}\}$ που δημιουργούν τη μέση τιμή και διακύμανση κάθε χαρακτηριστικού $\mu_1, ..., \mu_n$,$\sigma_1^2, ..., \sigma_n^2$.

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)}$$
$$\sigma_j^2 = \frac{1}{m} \sum_{i=1}^{m} (x_j^{(i)} - \mu_j)^2$$

Δεδομένου ενός νέου παραδείγματος x, υπολογίζεται $p(x)$:

$$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} exp(-\frac{(x_j^{(i)} - \mu_j)^2}{2\sigma_j^2})$$

Η ανωμαλία λοιπόν ορίζεται αν $p(x) < \epsilon$ Αντίστοιχα το ϵ είναι προϊόν της διαδικασία βελτιστοποίησης του αλγορίθμου.

5.5.3 Μεθοδολογία εξαγωγής αποτελεσμάτων

Η μεθοδολογία που χρησιμοποιήθηκε σε αυτό το σύστημα κάποια κοινά στοιχεία με τη μεθοδολογία του αλγορίθμου μη-επιβλεπόμενης μάθησης. Συνεπώς, τα αποτελέσματα προέρχονται από δύο βασικές συνιστώσες με την πρώτη να είναι η κανονικοποίηση των καταναλωτών ανά έτος και εν συνεχεία η συσταδοποίησή τους σε δύο συστάδες. Η μία συστάδα έχει αναμενόμενες καταναλωτικές συνήθειες και η άλλη αποτελείται από ασυνήθιστες. Η συστάδα με τις ασυνήθιστες συνήθειες βελτιστοποιείται με την παρατήρηση των χαρακτηριστικών διαχωρισμού και έτσι δημιουργείται η πρώτη πρόβλεψη του αλγορίθμου.

Παράλληλα, επεκτείνεται η δεύτερη συνιστώσα, για να αποκτηθεί και δεύτερη πρόβλεψη μέσω των χαρακτηριστικών. Ειδικότερα, τα χαρακτηριστικά περνούν από αλγόριθμο μείωσης διάστασης για να γίνει εφικτή η εποπτεία των χαρακτηριστικών σε δισδιάστατο περιβάλλον. Εν συνεχεία χρησιμοποιείται ο αλγόριθμος ανίχνευσης ανωμαλιών για την ολοκλήρωση της δεύτερης πρόβλεψης με δύο διαφορετικές μεθόδους:

- Η πρώτη μέθοδος εξάγει το μέσο όρο και την διακύμανση από τα μικτά δεδομένα εκπαίδευσης που χρησιμοποιούνται για την εύρεση της πυκνότητας της πολυμεταβλητής κανονικής κατανομής. Τα δεδομένα δοκιμής και τα δυαδικά χαρακτηριστικά τους χρησιμοποιούνται για τη βελτιστοποίηση του ορίου ταξινόμησης για να χρησιμοποιηθεί από τα δεδομένα εκπαίδευσης.

- Η δεύτερη μέθοδος εκμεταλλεύεται τη γνώση που παράχθηκε από την πρώτη συνιστώσα εκπαιδεύοντας το μοντέλο μόνο με αρνητικά παραδείγματα που χρησιμοποιούνται για την εύρεση της πυκνότητας της πολυμεταβλητής κανονικής κατανομής στο ένα κομμάτι των δεδομένων δοκιμής. Το άλλο κομμάτι των δεδομένων δοκιμής και τα δυαδικά χαρακτηριστικά τους χρησιμοποιούνται για τη βελτιστοποίηση του ορίου ταξινόμησης που εφαρμόζεται στο πρώτο κομμάτι δεδομένων δοκιμής.

Και οι δύο μέθοδοι εξάγουν δυαδικές προβλέψεις για τη δεύτερη συνιστώσα, ολοκληρώνοντας με αυτό τον τρόπο τις προβλέψεις του ταξινομητή. Δεδομένου ότι ο ταξινομητής πρέπει να έχει μια και μόνο εκτίμηση τα δύο δυαδικά χαρακτηριστικά εκτελούν μεταξύ τους απλές δυαδικές πράξεις που καταλήγοντας στην τελική πρόβλεψη του αλγορίθμου.

5.6 Δοκιμή αλγορίθμου ημι-επιβλεπόμενης μάθησης

5.6.1 Αποτελέσματα δοκιμής αλγορίθμου

5.7 Σχόλια

Κεφάλαιο 6

Δυσκολίες και μελλονική κατεύθυνση

6.1 Τεχνικά εμπόδια

6.1.1 Έλλειψη μακροχρόνιων δεδομένων

6.1.2 Δυσκολία γενίκευσης σε άλλες καταναλωτικές συνήθειες

6.1.3 Δυσκολία επιλογής μετρικών

6.1.4 Εύρεση αξιόπιστων δυαδικών χαρακτηρισμών

6.1.5 Ανατροφοδότηση ελέγχων

6.2 Ασφάλεια Καταναλωτών

6.2.1 Ασφάλεια Μετρητών

6.2.2 Απειλή ιδιωτικότητας

Κεφάλαιο 7

Συμπεράσματα

7.1 Σύγκριση αποτελεσμάτων

7.2 Συμπερασματικές σημειώσεις

Βιβλιογραφία

[1] P. Antmann. Reducing technical and non-technical losses in the power sector. In *Transmission and Distribution Conference and Exposition*, pages 24–26. World Bank, 2009.

[2] S. Axelsson. The base-rate fallacy and its implications for the difficulty of intrusion detection. In *CCS '99 Proceedings of the 6th ACM conference on Computer and communications security*, pages 1–7. Computer and Communications Security, 1999.

[3] Jason Brownlee. A tour of machine learning algorithms, 2013. Accessed: 5 August 2017.

[4] Y. Zhu C.-Y. Hsia and Chih-Jen Lin. A study on trust region update rules in newton methods for large-scale linear classification. Technical report, JMLR, 2017.

[5] ERGEG. *Smart Metering with a Focus on Electricity Regulation*, 2007. E07-RMF-04-03.

[6] James J. Filliben and Alan Heckert. Nist/sematech ehandbook of statistical methods. Accessed: 25 August 2017.

[7] Commission for Energy Regulation. General information. Accessed: 24 August 2017.

[8] Gregory C. Reinsel George E. P. Box, Gwilym M. Jenkins and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, 2016.

[9] Rob J. Hyndman and George Athanasopoulos. Forecasting: principles and practice, 2012. Accessed: 5 August 2017.

[10] Paul Johnson and Matt Beverlin. Machine learning, part ii: Supervised and unsupervised learning. Accessed: 1 September 2017.

[11] Paul Johnson and Matt Beverlin. Beta distribution, 2013.

[12] Mathworks. Parametric trend estimation, 2017. Accessed: 4 August 2017.

[13] G. Messinis and A. Dimeas. Utilizing smart meter data for electricity fraud. In *First South East European Region CIGRE Conference*, pages 2–4. CIGRE, 2014.

[14] Mkhwanazi. Electricity as a birthright and the problem of non-payment. In *Third Annual South Africa Revenue Protection Conference*, 1999.

[15] Kiambang Nik. Tenaga out to short-circuit electricity thefts. 1 1999.

[16] Micheal J. North and Charles M. Macal. *Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation.* Oxford University Press, New York, 2007. ch. 2, Other angles on Nondeterminism.

[17] Oracle. Data mining concepts. Accessed: 24 August 2017.

[18] J. F. G. Cobben P. Kadurek, J. Blom and W.L.Kling. Theft detection and smart metering practices and expectations in the netherlands. *Innovative Smart Grid Technologies Conference Europe*, pages 1–2, 2010. IEEE.

[19] J. F. G. Cobben P. Kadurek, J. Blom and W.L.Kling. Theft detection and smart metering practices and expectations in the netherlands. In *Innovative Smart Grid Technologies Conference Europe*, page 1. IEEE, 2010.

[20] Nasim Arianpoo Paria Jokar and Victor C. M. Leung. A practical guide to support vector classification. 7:1–3 12–16, 2003. University of Freiburg.

[21] Nasim Arianpoo Paria Jokar and Victor C. M. Leung. Electricity theft detection in ami using customers' consumption patterns. *Innovative Smart Grid Technologies Conference Europe*, 7:216–226, 2016. IEEE.

[22] Cho-Jui Hsieh Xiang-Rui Wang Rong-En Fan, Kai-Wei Chang and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[23] A. Naveen S. De, R. Anand and S. Moinuddin. E-metering solution for checking energy thefts and streamlining revenue collection in india. In *Transmission and Distribution Conference and Exposition*, pages 654–658. IEEE, 2003.

[24] Jon Shlens. *A Tutorial on Principal Component Analysis.* PhD thesis, Princeton University, 1993.

[25] Thomas B. Smith. Electricity theft: a comparative analysis. *Energy Policy*, 32(18):2067–2076, 2004.

[26] TACIS. *Improving Residential Electricity Services*, 1998. Tacis Technical Dissemination Project.

[27] V. Kumar V. Chandola, A. Banerjee. Anomaly detection: A survey. Technical report, ACM Computing Surveys, 2009.

[28] A. Siraj V. Ford and W. Eberle. Smart grid energy fraud detection using articial neuralcomputational intelligence applications in smart grid (ciasg)transmission and distribution conference and exposition. pages 2–4. IEEE, 2014.

[29] Σιμον Ηαψκιν. *Νευρωνικά Δίκτυα και Μηχανική Μάθηση*. Παπασωτηρίου, 2010.

[30] ΔΕΗ. *Το Κόστος των Ρευματοκλοπών*, 5 2017. Δελτίο τύπου 552017.

[31] Θοδωρής Παναγούλης. «Εγχειρίδιο» από τη ΡΑΕ για την αντιμετώπιση των όλο και περισσότερων ρευματοκλοπών. Αςςεσσεδ: 6 Αυγυστ 2017.

[32] ΡΑΕ. *Εγχειρίδιο Ρευματοκλοπών σε εφαρμογή της παραγράφου 23 του άρθρου 95 του Κώδικα Διαχείρισης Δικτύου Διαχείρισης Διανομής Ηλεκτρικής Ενέργειας*, 5 2017. Εφημερίδα της κυβερνήσεως της Ελληνικής Δημοκρατίας.

Παράρτημα Α΄

Αναλυτικά αποτελέσματα γραμμικών ταξινομητών

Συνδυασμός	DR	FPR	Accuracy	F1 score	BDR
1	94.66	35.93	67.04	35.79	0.23
2	93.89	34.62	68.15	36.39	0.23
3	92.37	39.70	63.41	32.88	0.21
4	91.67	21.23	80.15	49.62	0.32
5	93.13	34.21	68.44	36.42	0.23
6	91.60	35.11	67.48	35.35	0.22
7	93.89	34.29	68.44	36.61	0.23
8	94.66	35.93	67.04	35.79	0.23

Πίνακας Α΄.1: Αποτελέσματα δοκιμής τύπου 1 κανονικοποίηση [-1,1]

Συνδυασμός	DR	FPR	Accuracy	F1 score	BDR
1	75.65	1.38	96.67	79.45	0.86
2	80.00	1.46	96.96	81.78	0.86
3	80.00	1.46	96.96	81.78	0.86
4	80.87	1.54	96.96	81.94	0.85
5	81.74	1.94	96.67	80.69	0.82
6	77.39	1.54	96.67	79.82	0.85
7	65.22	1.62	95.56	71.43	0.82
8	75.65	1.46	96.59	79.09	0.85

Πίνακας Α΄.2: Αποτελέσματα δοκιμής τύπου 1 κανονικοποίηση [0,1]

Συνδυασμός	DR	FPR	Accuracy	F1 score	BDR
1	4.65	0.82	90.15	8.28	0.39
2	12.40	3.77	88.22	16.75	0.27
3	10.08	3.19	88.52	14.36	0.26
4	9.30	2.87	88.74	13.64	0.26
5	13.18	4.01	88.07	17.44	0.27
6	8.53	3.44	88.15	12.09	0.22
7	0.78	0.41	90.15	1.48	0.17
8	4.65	0.82	90.15	8.28	0.39

Πίνακας Α΄.3: Αποτελέσματα δοκιμής τύπου 2 με κανονικοποίηση [0,1]

Συνδυασμός	DR	FPR	Accuracy	F1 score	BDR
1	2.05	0.83	88.67	3.77	0.21
2	9.59	2.82	87.70	14.43	0.27
3	8.22	2.66	87.70	12.63	0.25
4	8.22	2.33	88.00	12.90	0.28
5	10.27	3.16	87.48	15.08	0.26
6	8.90	2.41	88.00	13.83	0.29
7	0.68	0.50	88.81	1.31	0.13
8	2.05	0.83	88.67	3.77	0.21

Πίνακας Α΄.4: Αποτελέσματα δοκιμής τύπου 3 με κανονικοποίηση [0,1]

Συνδυασμός	DR	FPR	Accuracy	F1 score	BDR
1	1.55	0.74	89.93	2.86	0.19
2	10.85	3.03	88.74	15.56	0.28
3	10.08	3.03	88.67	14.53	0.27
4	5.43	2.70	88.52	8.28	0.18
5	13.18	3.69	88.37	17.80	0.28
6	8.53	2.87	88.67	12.57	0.25
7	0.00	0.25	90.22	NαN	0.00
8	1.55	0.66	90.00	2.88	0.21

Πίνακας Α΄.5: Αποτελέσματα δοκιμής μικτών τύπων με κανονικοποίηση [0,1]

μικρός	3	2	1
89.9300	88.6700	90.1500	96.6700
88.7400	87.7000	88.2200	96.9600
88.6700	87.7000	88.5200	96.9600
88.5200	88.0000	88.7400	96.9600
88.3700	87.4800	88.0700	96.6700
88.6700	88.0000	88.1500	96.6700
90.2200	88.8100	90.1500	96.5600
90.0000	88.6700	90.1500	96.5900

Πίνακας Α΄.6: πίνακαςAccuracy

1	2	3	μικτός
80.7800	8.2800	3.7700	2.8600
81.2300	16.7500	14.4300	15.5600
79.2500	14.3600	12.6300	14.5300
79.8500	13.6400	12.9000	8.2800
80.3100	17.4400	15.0800	17.8000
78.6300	12.0900	13.8300	12.5700
78.9100	1.4800	1.3100	0
81.2300	8.2800	3.7700	2.8800

Πίνακας Α΄.7: πίνακαςF1 score

Γλωσσάριο

Ελληνικός όρος	Αγγλικός όρος
στιβαρότητα	robustness
κινητοί μέσοι όροι	moving averages
επαναδειγματοληψία	resampling
δειγματοληψία προς τα πάνω	upsampling
δειγματοληψία προς τα κάτω	downsampling
βάση σύγκρισης	benchmark
εκθετική εξομάλυνση	exponential smoothing
γραμμές Θ	theta lines
μηχανική μάθηση	machine learning
ανάλυση συστάδων	cluster analysis
συστάδα	cluster
συσταδοποίηση	clustering
υπερπροσαρμογή	overfitting
περιηγητής	browser