
Πανεπιστήμιο Θεσσαλίας
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

# Βραχυπρόθεσμη πρόβλεψη ηλεκτρικού φορτίου με χρήση τεχνικών μηχανικής μάθησης

Διπλωματική Εργασία

**ΑΠΟΣΤΟΛΟΥ Γ. ΤΑΜΒΑΚΗ**

**Επιβλέπων**
Μιχαήλ Βασιλακόπουλος
Αναπληρωτής Καθηγητής

Βόλος, Ιούνιος 2019

Πανεπιστήμιο Θεσσαλίας
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

# Βραχυπρόθεσμη πρόβλεψη ηλεκτρικού φορτίου με χρήση τεχνικών μηχανικής μάθησης

Διπλωματική Εργασία

**ΑΠΟΣΤΟΛΟΥ Γ. ΤΑΜΒΑΚΗ**

Επιτροπή επίβλεψης

| Επιβλέπων | Συνεπιβλέπων | Συνεπιβλέπουσα |
|---|---|---|
| Μιχαήλ Βασιλακόπουλος | Δημήτριος Μπαργιώτας | Ασπασία Δασκαλοπούλου |
| Αναπληρωτής Καθηγητής | Αναπληρωτής Καθηγητής | Επίκουρη Καθηγήτρια |

Βόλος, Ιούνιος 2019

Πανεπιστήμιο Θεσσαλίας
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών



Ο/Η συγγραφέας αυτής της εργασίας βεβαιώνει ότι κάθε βοήθεια την οποία είχε για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης βεβαιώνει ότι έχει αναφέρει τις όποιες πηγές από τις οποίες έκανε χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται επακριβώς, είτε παραφρασμένες.


University of Thessaly
Faculty of Engineering
Department of Electrical & Computer Engineering

# Short-term electrical load forecasting using machine learning techniques

Diploma Thesis

**APOSTOLOS G. TAMVAKIS**

**Supervisor**
Michael Vassilakopoulos
Associate Professor

Volos, June 2019

# Περίληψη

Η πρόβλεψη φορτίου στον τομέα της ενέργειας αποτελεί αναπόσπαστο κομμάτι του ηλεκτρικού συστήματος καθώς είναι κριτήριο για την ομαλή και βιώσιμη λειτουργία του. Η απελευθέρωση της αγοράς ηλεκτρικής ενέργειας, η είσοδος των ΑΠΕ στην παραγωγή και η ψηφιοποίηση των εποπτικών μέσων έχουν επιφέρει περισσότερη πολυπλοκότητα στο σύστημα που μεταφράζεται σε περισσότερες μεταβλητές.Τα μοντέλα μηχανικής μάθησης έχουν τη δυνατότητα να επεξεργάζονται μεγάλο πλήθος παραμέτρων και για αυτό τα καθιστά ελκυστικά στους ερευνητές. Στα πλαίσια αυτής της διπλωματικής εργασίας θα μελετηθούν οι πιο γνωστές μέθοδοι μηχανικής μάθησης για βραχυπρόθεσμη πρόβλεψη κατανάλωσης ενέργειας πάνω σε πραγματικά δεδομένα έξυπνων μετρητών. Σκοπός της εργασίας είναι η αξιολόγηση της απόδοσης των αλγορίθμων αλλά και η επιβεβαίωση των παρατηρήσεων άλλων ερευνών. που συνιστούν συνδυαστικές μεθοδολογίες που ενισχύουν την πρόβλεψη. Στο πειραματικό κομμάτι αξιολογείται η διαδικασία επιλογής των βέλτιστων παραμέτρων εισόδου των αλγορίθμων, η μέθοδος της ομαδοποίησης ενεργειακών προφίλ και η συνάθροιση των επιμέρους προβλέψεων αντί ενός συγκεντρωτικού δείγματος. Τα αποτελέσματα ήταν ενθαρρυντικά και έτσι δίνονται κίνητρα για περαιτέρω έρευνα και επέκταση αυτών των μεθόδων.

## Λέξεις Κλειδιά

Βραχυπρόθεσμη Πρόβλεψη,Μηχανική Μάθηση, Έξυπνοι Μετρητές.

# Abstract

Load forecasting in the field of power and energy is a crucial part to the power grid as it is a major factor in its smooth and sustainable operation. Energy liberalisation, RES integration to the production,the digitization and the expansion of digital monitoring equipment, all add extra complexity to the system which translates into more variables. Machine learning models posses the ability to process large amount of parameters and this is the reason they seem appealing to researchers. Over the course of this dissertation, the most common machine learning methods in response to short term load forecasting problems, will be studied on real smart meter data. The purpose of this project is to assess the performance of the algorithms and also confirm methodologies that are suggested in literature. The proposed methodologies include input parameter selection and smart meter clustering based aggregated forecast. In the experimental section of this projects both of this cases are evaluated. The results were motivating which lead to gather a greater interest into expanding further these methods

## Keywords

Short Term Load Forecasting, Machine Learning, Smart meters.

*Στην οικογένεια και τους φίλους μου.*

# Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ.Βασιλακόπουλο Μιχαήλ για την επίβλεψη αυτής της διπλωματικής εργασίας και για το ενδιαφέρον που έδειξε κατά την εκπόνηση αυτής.

Επίσης ευχαριστώ όλους τους καθηγητές της σχολής που κατά την διάρκεια των σπουδών μου προσέφεραν τόσο τις απαραίτητες επιστημονικές γνώσεις όσο και τις χρήσιμες συμβουλές για την μετέπειτα επαγγελματική μου καριέρα.

Ιδιαίτερη αναφορά θα ήθελα να κάνω στον καθηγητή κ.Ηλία Χούστη καθώς μέσα από τα μαθήματά του γνώρισα και αγάπησα το κομμάτι της Επιστήμης των Δεδομένων.

# Περιεχόμενα

# Κατάλογος σχημάτων

# Κατάλογος πινάκων

# Κεφάλαιο 1

# Εισαγωγή

## 1.1 Γενικά

Η ηλεκτρική ενέργεια αποτελεί για τον σύγχρονο κόσμο κάτι παραπάνω από απλό αγαθό. Είναι μια ανάγκη για την κοινωνία και ένας κύριος μοχλός για την ανάπτυξή της. Η σημερινή πραγματικότητα καθιστά πιο απαραίτητη από ποτέ τη μεταστροφή των συμβατικών συστημάτων ηλεκτρικής ενέργειας (ΣΗΕ) προς μία άλλη κατεύθυνση συνυφασμένη με τις εξελίξεις στον χώρο της αγοράς και τις οικολογικές συνέπειες των αυξανόμενων ενεργειακών απαιτήσεων στον πλανήτη.

Παραδοσιακά, οι απαιτήσεις του ηλεκτρικού δικτύου υπολογίζονταν βάσει προηγούμενων χρονικά τιμών κατανάλωσης ενέργειας. Η ολοένα αυξανόμενη ενεργειακή ζήτηση επέφερε κινδύνους διακοπής ρεύματος και αστάθειας στο δίκτυο με αποτέλεσμα να δημιουργηθούν μονάδες που θα καλύπτουν το φορτίο αιχμής, ωστόσο με υψηλό λειτουργικό κόστος. Το οικονομικό ζήτημα επιδεινώνεται περισσότερο με την μη αποδοτική εκμετάλλευση τη διαθέσιμης ισχύος [14]. Παράλληλα, παρατηρείται αύξηση των ενεργειακών ρύπων, καθώς στην πλειοψηφία τους οι μονάδες παραγωγής χρησιμοποιούν ορυκτά καύσιμα.

Για να διασφαλιστεί η αποδοτική επίδοση ενός ΣΗΕ, μία πολύ βασική και ζωτικής σημασίας προϋπόθεση για όλα τα μέρη του συστήματος είναι ο σωστός σχεδιασμός λειτουργίας και η συντήρηση του δικτύου. Χρήσιμο εργαλείο αποτελούν οι μέθοδοι πρόβλεψης που μπορούν να αναδείξουν μελλοντικά προβλήματα και να ορίσουν μελλοντικές ενεργειακές πολιτικές.

Η ανάγκη για αποδοτική και όσο το δυνατόν ακριβέστερη πρόβλεψη μπορεί να συνοψιστεί ως εξής:

- **Σχεδιασμός και Επέκταση Συστήματος Διανομής:** Για να αντιμετωπιστεί η αυξανόμενη ενεργειακή ζήτηση, ο σχεδιαστής του συστήματος θα πρέπει να μπορεί να εξυπηρετήσει τις απαιτήσεις στο αντίστοιχο χρονικό διάστημα με τον πιο οικονομικό και αξιόπιστο τρόπο. Αυτό βέβαια δεν είναι απλό, λόγω της μεγάλης έκτασης του δικτύου διανομής αλλά και επειδή σε αυτό το τμήμα συμβαίνουν οι μεγαλύτερες απώλειες σε όλο το σύστημα. Ακόμη υπάρχει ο κίνδυνος διακοπής λόγω βλαβών. Επιπρόσθετα, θα πρέπει να εντοπίζονται τα γεωγραφικά σημεία όπου συγκεντρώνονται μεγάλα μεγέθη φορτίων ώστε η κατασκευή υποσταθμών να βρίσκεται στο βέλτιστο σημείο.Το σκοπό αυτό εξυπηρετούν οι χωρικές προβλέ-

ψεις φορτίου . Το ζήτημα περιπλέκεται ακόμη περισσότερο με την είσοδο των Ανανεώσιμων Πηγών Ενέργειας (ΑΠΕ) καθώς επιφέρουν μεγαλύτερη αστάθεια και αβεβαιότητα στο σύστημα.

- **Λειτουργία και Διαχείριση Παραγωγής:** Η ακριβής πρόβλεψη φορτίου βοηθά τις εταιρείες ηλεκτρικής ενέργειας να λάβουν αποφάσεις ένταξης μονάδων, μείωση της ισχύος στρεφόμενης εφεδρείας, και τον ακριβή προγραμματισμό συντήρησης του εξοπλισμού. Αποτελεί βασικό ρόλο στην εξοικονόμηση παραγωγής ενέργειας και ταυτόχρονα συνιστά χρήσιμο εργαλείο για την αξιοπιστία του συστήματος. Οι διαχειριστές του συστήματος χρησιμοποιούν τα αποτελέσματα της πρόβλεψης φορτίου για να αποφασίσουν αν το δίκτυο είναι δυνητικά ευπαθές σε μεταβολές. Σε περίπτωση που ισχύσει αυτή η κατάσταση, διορθωτικές ενέργειες όπως η αποκοπή φορτίου, οι συμβάσεις αγοράς ενέργειας και οι συνδέσεις μονάδων αιχμής φορτίου χρειάζεται να είναι σε ετοιμότητα.

- **Οικονομικός Προγραμματισμός:** Ακριβείς προβλέψεις εξυπηρετούν την λήψη αποφάσεων ως προς τους οικονομικούς προϋπολογισμούς, την επέκταση και αναβάθμιση του δικτύου και την διαχείριση του ανθρώπινου δυναμικού. Ακόμη, τα ποσοστά ακριβείας έχουν άμεση σχέση με το υψηλό κόστος λειτουργίας της μονάδας. Χαμηλές εκτιμήσεις οδηγούν σε ανεπαρκή εφεδρεία ισχύος με αποτέλεσμα την συνεισφορά ακριβών μονάδων φορτίου αιχμής. Αντίθετα υψηλές εκτιμήσεις επιφέρουν περίσσεια εφεδρείας ξανά σχετιζόμενη με αυξημένα λειτουργικά κόστη. Χαρακτηριστικό είναι το παράδειγμα του Βρετανικού ηλεκτρικού δικτύου όπου 1% αύξηση στο σφάλμα πρόβλεψης επιφέρει αύξηση των λειτουργικών εξόδων κατά 10 εκατομμύρια λίρες τον χρόνο [12].

## 1.2 Ορισμός του προβλήματος

Σήμερα, η πρόβλεψη εντάσσεται στα πρότυπα των Ευφυών Δικτύων (Smart Grids) και την νέα αγορά ηλεκτρικής ενέργειας η οποία απαιτεί ακριβέστερα αποτελέσματα σε διαφορετικές κλίμακες ξεκινώντας από το επίπεδο ενός έξυπνου μετρητή ως και ολόκληρο το σύστημα.

Ωστόσο, η πρόβλεψη φορτίου δεν είναι μόνο σημαντική για την διαχείριση της παραγωγής, μεταφοράς και διανομής ενέργειας αλλά και για τους τελικούς χρήστες ώστε να είναι σε θέση να βελτιστοποιούν την διαχείριση της κατανάλωσης των κτιρίων.

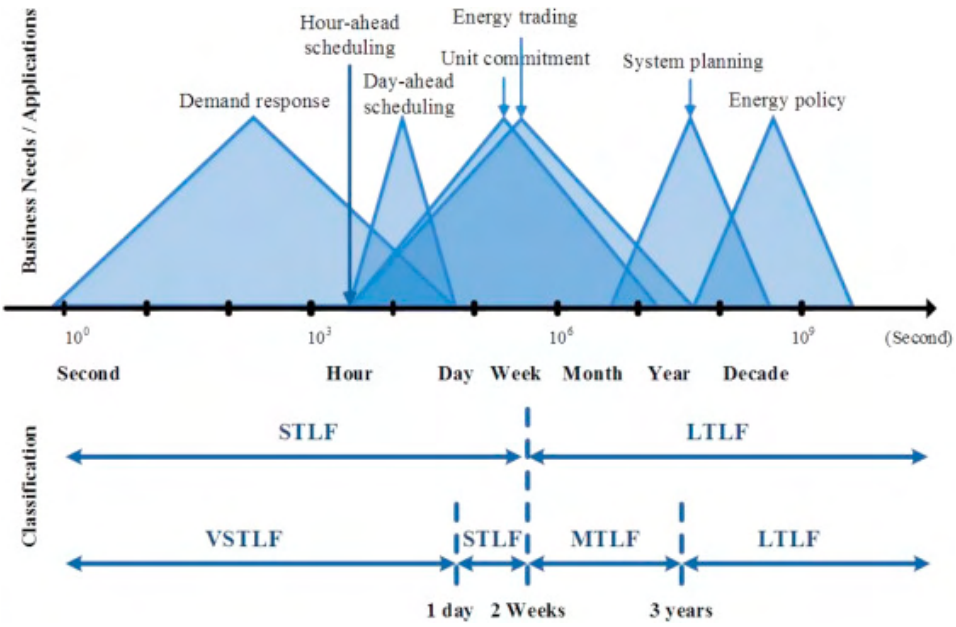
Η πρόβλεψη φορτίου υλοποιείται σε διαφορετική κλίμακα και σε διαφορετικό χρονικό ορίζοντα. Η κλίμακα πρόβλεψης ορίζεται ως το μέγεθος της μονάδας για την οποία επιτελείται η πρόβλεψη. Μπορεί να διαιρεθεί στα εξής επίπεδα:

- Πρόβλεψη σε επίπεδο μετρητή σπιτιού [28].
- Πρόβλεψη σε επίπεδο εμπορικού/βιομηχανικού κτιρίου [30].
- Συγκεντρωτικά σε επίπεδο περιοχής, περιφέρειας, χώρας [2].

Από πλευράς χρονικού ορίζοντα αυτός καθορίζεται σε 4 κατηγορίες [20] :

- Πολύ βραχυπρόθεσμη (VSTLF-Very Short Term Load Forecasting): Χρονικός ορίζοντας από μερικά λεπτά ως και μια ώρα μπροστά.
- Βραχυπρόθεσμη (STLF-Short Term Load Forecasting) : Περιλαμβάνει προβλέψεις μιας ημέρας έως και μίας εβδομάδας.
- Μεσοπρόθεσμη(MTLF-Mid Term Load Forecasting): Χρονικό εύρος μεταξύ 2 εβδομάδων και ενός έτους.
- Μακροπρόθεσμη(LTLF-Long Term Load Forecasting): Μεταξύ ενός χρόνου έως και 50 χρόνια.

Οι εφαρμογές των προβλέψεων για διαφορετικές χρήσεις απεικονίζονται στο παρακάτω διάγραμμα:




Σχήμα 1.1: Εφαρμογές πρόβλεψης σε ένα ηλεκτρικό σύστημα [15]

.

Η κατηγοριοποίηση ανάμεσα σε αυτές τις μεθόδους εξαρτάται επίσης με την συχνότητα δειγματοληψίας των μετρήσεων και την διαθεσιμότητα άλλων μεταβλητών που επηρεάζουν το φορτίο όπως είναι οι καιρικές συνθήκες και ο τύπος της ημέρας αναφορικά με τον αν είναι καθημερινή ή μη ή αν αποτελεί κάποια ξεχωριστή ημέρα ( γιορτές και αργίες). Οικονομικοί και δημογραφικοί δείκτες επίσης αποτελούν παράμετροι για διαφορετικού τύπου προβλέψεις.

Τα καιρικά δεδομένα που εξετάζονται αφορούν μετρήσεις θερμοκρασίας, υγρασίας, ταχύτητας ανέμου, βροχόπτωσης κ.α. . Έχει παρατηρηθεί ότι από όλες τις καιρικές παραμέτρους, αυτές που επηρεάζουν περισσότερο την κατανάλωση ενέργειας είναι κυρίως οι θερμοκρασία και η υγρασία. Η εξέλιξη της τεχνολογίας έδωσε πρόσβαση σε ακριβέστερες μετρήσεις που ενισχύουν τα καιρικά προγνωστικά με άμεση συνέπεια πιο σίγουρες προγνώσεις και για το ηλεκτρικό φορτίο [26].

Εκτός από τις καιρικές μεταβλητές, η ανθρώπινη δραστηριότητα μπορεί να θεωρηθεί ισχυρός παράγοντας επιρροής ενεργειακής κατανάλωσης. Αυτός ο παράγων μπορεί να διακριτοποιηθεί σε διάφορες χρονικές και ημερολογιακές παραμέτρους. Για παράδειγμα η ζήτηση το σαββατοκύριακο μπορεί να διαφέρει από τις καθημερινές και σε πολλές περιπτώσεις να παρουσιάζει παρόμοια συμπεριφορά με τις αργίες. Επίσης, οι κοινωνικοί και βιομηχανικοί δείκτες ανάπτυξης μπορούν να επηρεάσουν την συμπεριφορά κατανάλωσης ενέργειας [9].

## 1.3 Σκοπός της εργασίας

Το αντικείμενο της διπλωματικής εργασίας είναι η εφαρμογή και η αξιολόγηση των τεχνικών μηχανικής μάθησης για την ανάλυση ενεργειακών δεδομένων και την πρόβλεψη κατανάλωσης ενέργειας. Τα δεδομένα τα οποία εξετάζονται περιλαμβάνουν μετρήσεις οικιακών έξυπνων μετρητών σε μία περιοχή του Λονδίνου. Τα κριτήρια που λήφθηκαν υπόψιν για την κάλυψη αυτού του θέματος είναι οι πρακτικές εφαρμογές του, όσο και η μεγάλη έμφαση που έχει δοθεί από τους ερευνητές σε αυτό το είδος πρόβλεψης στον τομέα της ενέργειας τα τελευταία χρόνια .

Η ανάπτυξη του πειραματικού μέρους στοχεύει στην συγκριτική αξιολόγηση των πιο διαδεδομένων αλγορίθμων και κατά πόσο αυτοί προσεγγίζουν τις πραγματικές τιμές των δεδομένων. Επίσης, εξετάζονται οι παράμετροι που συσχετίζονται με την κατανάλωση φορτίου και πως επηρεάζουν την ακρίβεια της πρόβλεψης. Συμπληρωματικά, έχοντας πρόσβαση στα ενεργειακά δεδομένα κάθε μετρητή, επιχειρείται η ομαδοποίηση ενεργειακών προφίλ που ανταποκρίνεται στην φιλοσοφία του ευφυούς δικτύου και της δυναμικής τιμολόγησης. Είναι λοιπόν, εφικτή η υλοποίηση προσαρμοσμένων τιμολογίων ενέργειας που μπορούν να συνεισφέρουν στην μείωση της κατανάλωσης σε ώρες αιχμής όπως και της συνολικής κατανάλωσης του συστήματος. Με το τρόπο αυτό ελαττώνεται και το κόστος για τον καταναλωτή.

## 1.4 Οργάνωση του τόμου

Αρχικά παρατίθεται η περίληψη της πτυχιακής εργασίας στα ελληνικά και στα αγγλικά, όπου παρουσιάζεται συνοπτικά το κύριο θέμα της. Έπειτα ακολουθούν τα περιεχόμενα και τέλος το κύριο μέρος της εργασίας που αναπτύχθηκε σε 6 κεφάλαια.

**Κεφάλαιο 1:** Πρόκειται για το παρόν κεφάλαιο το οποίο εισάγει την έννοια της πρόβλεψης ενέργειας, όπως επίσης τα κίνητρα και τους στόχους.

**Κεφάλαιο 2:** Παρουσιάζονται οι συγγενικές εργασίες και η λογική που συσχετίζεται με αυτή την εργασία. Κατά την ολοκλήρωση αυτού του κεφαλαίου ο αναγνώστης θα μπορέσει να καταλάβει τόσο την λογική με την οποία αναπτύσσεται η βραχυπρόθεσμη πρόβλεψη όσο και τα κριτήρια επιλογής των μεθόδων πρόβλεψης.

**Κεφάλαιο 3**:Γίνεται η παρουσίαση και η μαθηματική ερμηνεία των μεθόδων που στην συνέχεια θα εφαρμοστούν.

**Κεφάλαιο 4:** Ξεκινά η προεργασία για το πειραματικό κομμάτι της διπλωματικής. Πιο συγκεκριμένα, γίνεται εξερεύνηση και οπτικοποίηση των δεδομένων και στην συνέχεια η ανάπτυξη της μεθοδολογίας.

**Κεφάλαιο 5:** Πρόκειται για το πειραματικό κομμάτι της εργασίας όπου δοκιμάζονται πάνω σε πραγματικά δεδομένα οι τεχνικές πρόβλεψης που αναφέρθηκαν σε προηγούμενα κεφάλαια.

**Κεφάλαιο 6:** Είναι το τελευταίο κεφάλαιο της εργασίας στο οποίο γίνεται η ανασκόπηση του κύριου θέματος αυτής και οι προοπτικές για περαιτέρω έρευνα πάνω στο αντικείμενο.

# Κεφάλαιο 2

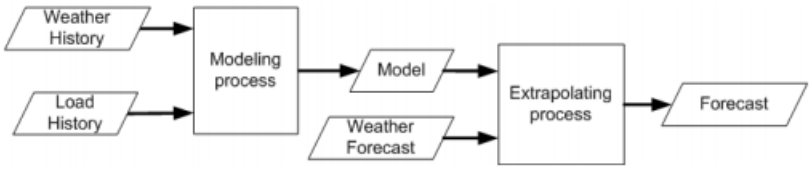
# Βιβλιογραφική αναφορά

## 2.1 Εισαγωγή

Η εικόνα 2.1 παρουσιάζει ένα τυπικό μηχανισμό βραχυπρόθεσμης πρόβλεψης που βασίζεται σε καιρικά δεδομένα. Ιστορικές μετρήσεις φορτίου και καιρού αποτελούν εισόδους για το μοντέλο και στην συνέχεια το παραγόμενο μοντέλο συνδυάζεται με δεδομένα πρόγνωσης καιρού για την εξαγωγή του τελικού αποτελέσματος. Στο στάδιο της μοντελοποίησης είναι κρίσιμης σημασίας η διάκριση της συστηματικής απόκλισης, ο θόρυβος των δεδομένων και η αντιμετώπισή τους καθώς επηρεάζουν σημαντικά την ακρίβεια του αποτελέσματος. Ως συνέπεια, μία πληθώρα πρωτοποριακών ερευνών και πρακτικών έχει επικεντρωθεί στο κομμάτι της μοντελοποίησης STLF. Η γενική προσέγγιση επίλυσης μπορεί να χωριστεί σε 2 μέρη:

1. Συλλογή και επεξεργασία δεδομένων.

2. Εφαρμογή και αξιολόγηση αλγορίθμων πάνω στα επεξεργασμένα δεδομένα.




Σχήμα 2.1: Μία τυπική STLF διαδικασία.

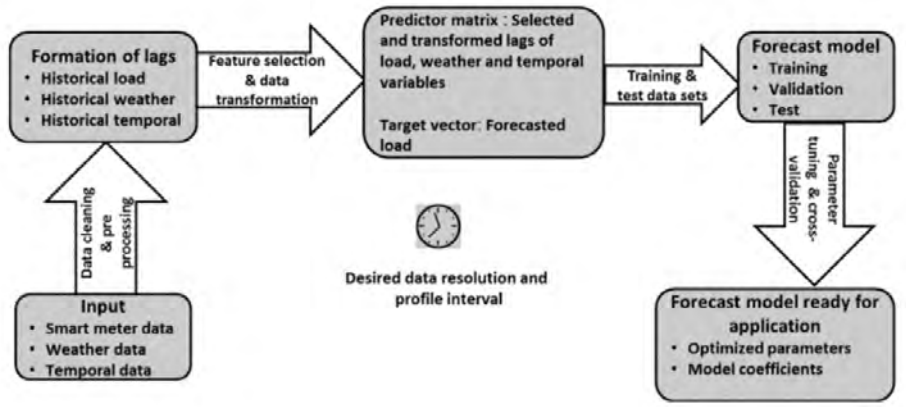
Η συλλογή δεδομένων έχει ως σκοπό την εύρεση των καταλληλότερων μεταβλητών για κάθε πρόβλημα ξεχωριστά και προσπαθεί εξάγει συμπεράσματα και να ερμηνεύσει την επιρροή αυτών στην κατανάλωση ενέργειας. Η διαδικασία μετέπειτα ενσωματώνεται στην εφαρμογή τεχνικών πρόβλεψης. Στην γενικότερη περίπτωση, το ηλεκτρικό φορτίο είναι άμεσα συνυφασμένο με την ανθρώπινη συμπεριφορά. Οι επιδράσεις της φύσης αντανακλώνται στις καιρικές μεταβλητές ενώ η ανθρώπινη επιρροή στις ημερολογιακές μεταβλητές, π.χ. ώρες εργασίας.

Επίσης με την απελευθέρωση τη αγοράς ενέργειας, ολοένα και περισσότερες εταιρείες εισέρχονται σε αυτήν, διότι η ενέργεια αντιμετωπίζεται ως εμπόρευμα. Σε μερικές περιπτώσεις, οι καταναλωτές που συμμετέχουν σε προγράμματα δυναμικής τιμολόγησης περιορίζουν την κατανάλωσή τους τις ώρες με υψηλή τιμολόγηση όταν αυτό είναι δυνατό, επιλέγοντας άλλες ώρες. Η πληροφορία της εξέλιξης της τιμής σε ένα τέτοιο περιβάλλον επηρεάζει τα προφίλ κατανάλωσης. Μοντέλα πρόβλεψης ευαίσθητα στις μεταβολές της τιμής έχουν προταθεί σε χώρες που υποστηρίζουν την δυναμική τιμολόγηση και επιδεικνύουν μεγάλη ακρίβεια σε σχέση με τις ήδη υπάρχουσες τεχνικές [1].

Στις παρακάτω ενότητες περιγράφονται η μέθοδος συλλογής και προ επεξεργασίας των δεδομένων , τα μοντέλα εκμάθησης που χρησιμοποιούνται και η βιβλιογραφική αναφορά σε σχετικές έρευνες.

## 2.2 Συλλογή και προ επεξεργασία δεδομένων

Αρχικές προσπάθειες για προβλέψεις κατανάλωσης ενέργειας συμπεριελάμβαναν λεπτομερή σετ δεδομένων με δημογραφικά στοιχεία. Ωστόσο, μετά την ευρεία εγκατάσταση έξυπνων μετρητών τα τελευταία χρόνια, οι έρευνες επικεντρώνονται σε τρόπους που εκμεταλλεύονται τα εκτενή δεδομένα που προσφέρονται από τους μετρητές σε συνδυασμό με τις ημερολογιακές μεταβλητές και τις ακριβείς μετρήσεις σε καιρικά δεδομένα. Οι αλγόριθμοι που υλοποιούν αυτή την προσέγγιση αναφέρονται στην βιβλιογραφία ως Smart Meter Based Model (SMBM) [31].




Σχήμα 2.2: Μοντέλο SMBM [21]

Τα δεδομένα που χρησιμοποιούνται για την πρόβλεψη μπορούν να οργανωθούν σε 3 τύπους:

- Πραγματικά δεδομένα συλλέγονται μέσω λογαριασμών εταιρειών ρεύματος , έξυπνων μετρητών, συστημάτων ενεργειακής διαχείρισης κτιρίων (BMS), προγνώσεις μετεωρολογικών σταθμών.

- Δεδομένα προσομοίωσης τα οποία παρέχονται από ειδικά λογισμικά μοντελοποίησης ηλεκτρικών δικτύων (Gridlab-D [13], EnergyPlus [11]).

- Δημόσια δεδομένα τα οποία έχουν δοθεί για διαγωνισμούς και για ερευνητικούς σκοπούς στα πλαίσια εύρεσης της καλύτερης επίδοσης διαφορετικών μεθόδων. Γνωστά datasets είναι το Irish Smart Meter Trial [16] και το πρόγραμμα Low Carbon London [22].

Το μέγεθος των πραγματικών δεδομένων κατανάλωσης τα οποία τροφοδοτούν τα μοντέλα μπορεί να διαφέρει ως προς την διάρκεια που γίνεται η δειγματοληψία και μπορεί να κρατά από δύο εβδομάδες ως και 4-5 χρόνια. Για να βελτιωθεί η ακρίβεια της πρόβλεψης έχουν προταθεί διάφορες τεχνικές. Οι δύο πιο δημοφιλής είναι :

- Ομαδοποίηση, μία πολύ γνωστή διαδικασία βελτίωσης των προβλέψεων. Έχει παρατηρηθεί πως η πρόβλεψη για ένα σπίτι οδηγεί σε μεγάλα σφάλματα, οπότε η ομαδοποίησή τους σε ομογενή γκρουπ ομαλοποιεί το δείγμα και βελτιώνει το σφάλμα. Στην πράξη ερευνά προηγούμενες χρονικά τιμές για να αναγνωρίσει περιόδους που η κατανάλωση ενέργειας εμφανίζει παρόμοια χαρακτηριστικά. Δηλαδή, εφαρμόζεται μια τεχνική clustering πριν το στάδιο της πρόβλεψης.

- Η δεύτερη μέθοδος βασίζεται στην βελτίωση της πρόγνωσης μέσω δεδομένων καιρού καθώς είναι αυτή η παράμετρος που επηρεάζει περισσότερο την συμπεριφορά κατανάλωσης ενέργειας.

Οι παραπάνω τεχνικές δρουν συνδυαστικά και αποτελούν στάδιο προ-επεξεργασίας στο γενικότερο μοντέλο πρόβλεψης. Βασική προϋπόθεση για αυτό το στάδιο είναι τα δεδομένα που παρέχονται να ‘‘καθαρίζονται’’ από τυχόν ακραίες ή μηδενικές τιμές που οφείλονται σε σφάλμα του μετρητή.

Έχοντας ένα μεγάλο διάνυσμα από παραμέτρους *n* που χαρακτηρίζουν ένα στοιχείο σε ένα δείγμα μεγέθους *m* είναι κατανοητό πως ο όγκος της πληροφορίας μπορεί να είναι αρκετά μεγάλος και η διάρκεια επεξεργασίας ολόκληρου αυτού του δείγματος να απαιτεί πολύ χρόνο. Σε αυτή την περίπτωση, προτείνεται η εύρεση των πιο χρήσιμων εισόδων για το μοντέλο. Αυτή η διαδικασία ονομάζεται επιλογή παραμέτρων (feature selection). Επιπρόσθετα, στην ενίσχυση των αλγορίθμων πρόβλεψης βοηθούν οι τεχνικές μετασχηματισμού δεδομένων (data transformation techniques). Αυτές οι τεχνικές τροποποιούν την αρχική μορφή των δεδομένων για να την αναπαραστήσουν σε μία πιο χρήσιμη. Γνωστά παραδείγματα αποτελούν η ανάλυση PCA και ο μετασχηματισμός Wavelet [31].

Μετά την επεξεργασία τα επιλεγμένα χαρακτηριστικά αναπαρίστανται με την μορφή ενός πίνακα (predicator matrix) με m αριθμό χρονικών στιγμιότυπων ως σειρές και n αριθμό χαρακτηριστικών ως στήλες. Αυτός αποτελεί την είσοδο της συνάρτησης που έχει ως έξοδο ένα διάνυσμα (target vector). Συνηθισμένη τακτική για την εκτέλεση αλγορίθμων πρόβλεψης είναι η πρότερη κανονικοποίηση (normalization) των δεδομένων για την μείωση του χρόνου εκτέλεσης .

Η προ-επεξεργασία είναι μια συνηθισμένη ρουτίνα στα πλαίσια της εξόρυξης δεδομένων και μπορεί να περιλαμβάνει τα εξής στάδια:

- Καθαρισμός δεδομένων (Data cleaning): Είναι η διαδικασία εντοπισμού και διόρθωσης (συμπλήρωση, τροποποίηση, αντικατάσταση, απαλοιφή) δεδομένων που είναι ημιτελή , ανα-

κριβή, ελλιπή ή άσχετα. Διαφορετικές μέθοδοι με συνηθέστερη την γραμμική παρεμβολή για την συμπλήρωση ελλιπών στοιχείων αλλά και άλλες αντιμετωπίζουν αυτά τα ζητήματα.

- Ενοποίηση δεδομένων (Data Integration): Είναι η διαδικασία που ενσωματώνει πολλαπλά δεδομένα από διαφορετικές πηγές.

- Μετασχηματισμός δεδομένων (Data Transformation): Περιλαμβάνει την κανονικοποίηση (normalization), την εξομάλυνση (smoothing) και την συγκέντρωση/διάσπαση (aggregation/disaggregation) των δεδομένων. Αυτές οι μέθοδοι τροποποιούν τα δεδομένα και τα αναπαριστούν σε μια πιο βολική μορφή. Για παράδειγμα η χρήση μετασχηματισμού Fourier μετατρέπει το διακριτής μορφής ηλεκτρικό φορτίο σε ημιτονοειδή μορφή διαφορετικής συχνότητας και πλάτους [8].

## 2.3 Αλγόριθμοι πρόβλεψης κατανάλωσης ενέργειας

Οι προτεινόμενοι τρόποι για την επίλυση της πρόβλεψης κατανάλωσης ενέργειας δεν καταλήγουν σε κάποιο σαφές συμπέρασμα ως προς το ποια αποδίδει καλύτερα καθώς το πρόβλημα είναι πολυσύνθετο και εξαρτάται από τον λόγο για τον οποίο αυτή επιτελείται. Στα άρθρα [10], [29] επιχειρείται η ανασκόπηση των επικρατέστερων μεθόδων πρόβλεψης.

Η διάκριση που γίνεται είναι ανάμεσα σε:

- *Στατιστικές τεχνικές:* Αυτές οι τεχνικές μπορεί να είναι ντετερμινιστικά μοντέλα τα οποία εκφράζουν την σχέση μεταξύ του φορτίου και των διάφορων παραμέτρων εισόδου. Επίσης μπορεί να είναι στοχαστικά μοντέλα που το φορτίο μπορεί να αναπαρίσταται με τη μορφή στοχαστικής διεργασίας. Μερικά από τα μοντέλα ανήκουν σε αυτή την κατηγορία είναι : πολυπαραμετρική παλινδρόμηση (multivariate regression), και η εκθετική εξομάλυνση (exponential smoothing).

- *Ανάλυση χρονοσειρών:* Τα μοντέλα αυτά υποθέτουν ότι κατά την απουσία διαταραχών σε σημαντικούς παράγοντες ενός επαναλαμβανόμενου γεγονότος, τα μελλοντικά δεδομένα αυτού του γεγονότος σχετίζονται με τα παρελθοντικά και μπορούν να αναπαρασταθούν μέσα από μοντέλα που αξιοποιούν τις χρονικά προηγούμενες παρατηρήσεις.Στον κλάδο των προβλέψεων είναι η συνηθέστερη προσέγγιση. Γνωστότερα μοντέλα αυτού του τύπου αποτελούν οι τεχνικές αυτοπαλινδρομικού κινητού μέσου όρου (ARIMA) και οι διάφορες προεκτάσεις του.

- *Μέθοδοι τεχνητής νοημοσύνης και μηχανικής μάθησης:* Τεχνητά νευρωνικά δίκτυα (ANN), μοντέλα ασαφούς λογικής(fuzzy logic), μηχανές διανυσμάτων υποστήριξης(SVM) και δένδρα απόφασης (Decision Trees), χρησιμοποιούνται για την επίλυση προβλημάτων πρόβλεψης με βάση τις απαιτήσεις του συστήματος. Έχει παρατηρηθεί πως ενώ θεωρητικά στην πλειοψηφία τους έχουν μελετηθεί στις δεκαετίες του ‘80 και ’90, έχουν βρει πρακτικές και ευρύτερες εφαρμογές τα τελευταία χρόνια. Αυτό οφείλεται στην βελτίωση της υπολογιστικής επεξεργασίας των υπολογιστών και στην ανάπτυξη βιβλιοθηκών που παρέχουν αυτούς τους αλγορίθμους.

## 2.4 Συγγενικές εργασίες

Σε αυτήν την ενότητα παρουσιάζονται δημοσιευμένες έρευνες που εφαρμόζουν τις τεχνικές που έχουν αναφερθεί στα πλαίσια της πρόβλεψης κατανάλωσης ενέργειας.

### 2.4.1 Πολλαπλή γραμμική παλινδρόμηση (Multiple linear regression)

Η ανάλυση παλινδρόμησης είναι η στατιστική διαδικασία για τον υπολογισμό συσχετίσεων μεταξύ μεταβλητών. Τέτοια μοντέλα χρησιμοποιούνται και στην περίπτωση της βραχυπρόθεσμης πρόβλεψης. Το φορτίο ή μία μετασχηματισμένη μορφή του φορτίου αντιμετωπίζεται ως η εξαρτημένη μεταβλητή, ενώ μεταβλητές όπως οι καιρικές και οι ημερολογιακές αντιμετωπίζονται ως ανεξάρτητες. Η μέθοδος αυτή, απαιτεί την αναπαράσταση μέσω μιας μορφής γραμμικής εξίσωσης μεταξύ των μεταβλητών.

Όταν λαμβάνονται υπόψιν μοντέλα γραμμικής παλινδρόμησης για προβλέψεις φορτίου, μία κοινή παρανόηση είναι ότι δεν είναι κατάλληλα να μοντελοποιήσουν την μη γραμμική σχέση μεταξύ του φορτίου και των καιρικών μεταβλητών. Η έννοια της 'γραμμικής' αναφέρεται στις γραμμικές εξισώσεις για κάθε ανεξάρτητη μεταβλητή που συνεισφέρει στην προσέγγιση της λύσης και όχι στον συσχετισμό μεταξύ των ανεξάρτητων μεταβλητών.

Οι Papalexopoulos και Herstenbeng [24] πρότειναν μία παλινδρομική προσέγγιση στο πρόβλημα της βραχυπρόθεσμης πρόβλεψης. Η μέθοδος δοκιμάστηκε πάνω σε δεδομένα της ενεργειακής εταιρείας PG&E για πρόβλεψη αιχμών φορτίου και ωριαίας κατανάλωσης ενέργειας μέσα στο επόμενο 24ωρο. Αυτή η μελέτη έχει αποτελέσει και την βάση για όλες τις έρευνες που βασίζονται στην πολλαπλή γραμμική παλινδρόμηση.

Μεταγενέστερες μελέτες [7] ενσωματώνουν και άλλες παραμέτρους όπως ο καιρός, ειδικές μέρες και οικονομετρικούς δείκτες.

### 2.4.2 Εκθετική εξομάλυνση (Exponential Smoothing)

Η εκθετική εξομάλυνση αναθέτει βάρη σε παρελθοντικές τιμές του φορτίου οι οποίες μειώνονται εκθετικά με την πάροδο του χρόνου. Έτσι δίνεται μεγαλύτερη βαρύτητα στις πιο κοντινές ώρες και λιγότερο σε αυτές που είναι αρκετά παλαιότερες. Ακόμη, βασίζεται μόνο στις τιμές του φορτίου, που σημαίνει λιγότερες απαιτήσεις αποθήκευσης δεδομένων σε σχέση με τις μεθόδους όπως η πολλαπλή παλινδρόμηση και τα νευρωνικά δίκτυα.

Οι έρευνα [19] και μελετά τους αλγορίθμους εκθετικής εξομάλυνσης και τις προεκτάσεις τους π.χ. διπλή και τριπλή εκθετική εξομάλυνση. Σε σύγκριση με άλλες μεθόδους οι οποίες και αυτές δεν συμπεριλαμβάνουν λοιπά δεδομένα, παρατηρήθηκε καλύτερη απόδοση. Ωστόσο στα πλαίσια του διαγωνισμού GEFCom2012 δεν σημείωσαν μεγάλη επιτυχία καθώς τα ενεργειακά δεδομένα ήταν σημαντικά επηρεασμένα από τις κλιματολογικές συνθήκες τις οποίες αυτό το μοντέλο ήταν αδύνατο να εντοπίσει.

### 2.4.3 Αυτοπαλινδρομικά μοντέλα κινητού μέσου όρου και ενσωματωμένου μέσου όρου (ARMA and ARIMA)

Αυτές οι μέθοδοι προτάθηκαν από τους Box και Jenkins και είναι από τις ευρύτερα διαδεδομένες στο τομέα της ανάλυσης χρονοσειρών. Το μοντέλο ARMA περιλαμβάνει 2 μέρη: το αυτοπαλινδρομικό (AR) και τον κινητό μέσο όρο (MA). Η τρέχουσα τιμή της χρονοσειράς αποτυπώνεται σε έναν όρο γραμμικής παλινδρόμησης της τρέχουσας τιμής ως προς τις προηγούμενες (AR) και ως ένα όρο γραμμικής παλινδρόμησης της τρέχουσας τιμής ως προς τον λευκό θόρυβο/σφάλμα (MA). Οι χρονοσειρές μπορούν να αντιμετωπιστούν και ως στοχαστικές διεργασίες και χαρακτηρίζονται από ιδιότητες όπως είναι : η τάση, η αυτοσυσχέτιση, η περιοδικότητα και η στασιμότητα. Για το ARMA μοντέλο βασική προϋπόθεση είναι η στασιμότητα της σειράς. Έτσι προτάθηκε η μέθοδος ARIMA που μετατρέπει μη στάσιμες χρονοσειρές σε στάσιμες. Η πρόβλεψη ωριαίας ζήτησης ενέργειας είναι μη στάσιμη και για αυτό προτιμάται το μοντέλο ARIMA και οι παραλλαγές του.

Ο Weron (2006) [25] παρουσιάζει μια ανασκόπηση πάνω σε πραγματικά δεδομένα σε σχέση τις παραπάνω τεχνικές.

### 2.4.4 Μηχανές διανυσμάτων υποστήριξης (SVM)

Οι μηχανές διανυσμάτων υποστήριξης παρουσιάστηκαν για πρώτη φορά από τον Vladimir Vapnik για την αντιμετώπιση προβλημάτων κατηγοριοποίησης (classification problems) όπως αναγνώριση συμβολοσειρών, αναγνώριση προσώπου κ.α. Ο ίδιος αργότερα δημοσίευσε και μία επέκταση αυτού του αλγορίθμου σε προβλήματα παλινδρόμησης [27]. Κατά συνέπεια βρήκε χρήση και στα ζητήματα πρόβλεψης φορτίου επιλύοντας μάλιστα προβλήματα υπερπροσαρμογής (soft margin SVM) που προέκυπταν από άλλους αλγορίθμους [6].

Στον διαγωνισμό του EUNITE network (2004) με θέμα την πρόβλεψη αιχμών φορτίου για τις επόμενες 31 μέρες, δηλαδή για πρόβλεψη μεσοπρόθεσμου διαστήματος (MTLF), οι νικητές βασίστηκαν στην μέθοδο SVM [5]. Πιο συγκεκριμένα, το μοντέλο βασιζόταν σε δεδομένα αποκλειστικά χειμερινής περιόδου και μάλιστα δεν χρησιμοποίησε καθόλου κλιματολογικά δεδομένα. Τα συμπεράσματα ήταν πως τελικά ίσως δεν είναι απαραίτητο να χρησιμοποιούνται καιρικές μεταβλητές για προβλήματα MTLF. Αν και ο διαγωνισμός ήταν επικεντρωμένος στο MTLF, οδήγησε στην υιοθέτηση SVM προσεγγίσεων για προβλήματα βραχυπρόθεσμης πρόβλεψης (STLF).

### 2.4.5 Διαβαθμιζόμενη Ενίσχυση (Gradient Boosting)

Η διαβαθμιζόμενη ώθηση είναι ένας αλγόριθμος μηχανικής μάθησης για επίλυση προβλημάτων παλινδρόμησης. Παράγει ένα ισχυρό μοντέλο πρόβλεψης (strong classifier) υπό την μορφή ενός πλήθους ασθενών μοντέλων πρόβλεψης (weak classifiers). Πολλοί αλγόριθμοι ώθησης είναι στην πραγματικότητα αλγόριθμοι καθόδου (descent algorithms), οι οποίοι βελτιστοποιούν μια κυρτή συνάρτηση κόστους [23].

Στον διαγωνισμό GEFCom2014 οι Taieb και Hyndman [3] προτείναν μία εξαιρετικά αποδοτική τεχνική gradient boosting η οποία αντιλαμβανόταν με επιτυχία μεταβολές στο φορτίο εξαιτίας εξωγενών παραμέτρων.

### 2.4.6 Νευρωνικά δίκτυα (Neural Networks)

Τα νευρωνικά δίκτυα είναι μία πολλά υποσχόμενη κατηγορία μηχανικής μάθησης καθώς δεν βασίζονται στην ανθρώπινη εμπειρία για την εκμάθηση πάνω στα δεδομένα, αλλά επιχειρούν να μάθουν τα ίδια την συναρτησιακή σχέση μεταξύ των εισόδων και εξόδων του συστήματος. Αυτή η προσέγγιση δεν βασίζεται σε μία ρητή αποδοχή μίας συναρτησιακής σχέσης μεταξύ παρελθοντικών τιμών φορτίου ή καιρικών παραμέτρων και της πρόγνωσης του φορτίου. Αντ' αυτού, τα νευρωνικά μαθαίνουν μόνα τους την συναρτησιακή σχέση εισόδων και εξόδου κατά την διαδικασία της εκπαίδευσης. Αρχικά το νευρωνικό δίκτυο εκπαιδεύεται με τα ιστορικά δεδομένα εισόδου και εξόδου, ενώ στην συνέχεια δίνει προβλέψεις πάνω σε δοσμένες εισόδους. [17]

Από τις αρχές του 1990 βρίσκουν μεγάλη απήχηση στον τομέα της βραχυπρόθεσμης πρόβλεψης εξαιτίας της μεγάλης προσαρμοστικότητας στα δεδομένα που τροφοδοτούνται .Το πρώτο νευρωνικό δίκτυο που εφαρμόστηκε για STLF σκοπούς προτάθηκε από τον Park κ.α. [18] και ήταν ένα δίκτυο τριών επιπέδων (layers) για πρόβλεψη ωριαίας κατανάλωσης, φορτίου ώρας αιχμής και συνολικής κατανάλωσης ενέργειας για την επόμενη μέρα. Συγκρίνοντάς το με τις υπόλοιπες προτεινόμενες τεχνικές το σφάλμα πρόβλεψης ήταν 2% χαμηλότερο βασιζόμενο σε στοιχεία μετρήσεων φορτίου και θερμοκρασίας τριών μηνών.

Έκτοτε, πολλοί τύποι νευρωνικών δικτύων έχουν χρησιμοποιηθεί για την πρόβλεψη φορτίου, όπως τα δίκτυα πρόσθιας τροφοδότησης (feed-forward neural networks), τα δίκτυα συναρτήσεων ακτινικής βάσης (radial basis function) και τα αναδραστικά νευρωνικά δίκτυα (recurrent neural networks).

# Κεφάλαιο 3

# Θεωρητικό υπόβαθρο

## 3.1 Εισαγωγή

Σε αυτό το κεφάλαιο παρουσιάζεται το θεωρητικό υπόβαθρο των τεχνικών πρόβλεψης που εφαρμόζονται για αυτή την εργασία. Οι επόμενες ενότητες περιγράφουν τους περισσότερους από τους αλγορίθμους μηχανικής που αναφέρθηκαν συνοπτικά στο προηγούμενο κεφάλαιο.

Θέτοντας το πρόβλημα της πρόβλεψης φορτίου ως ένα πρόβλημα παλινδρόμησης πρέπει πρώτα να ορίσουμε κάποιες βασικές έννοιες.

### 3.1.1 Δεδομένα εκπαίδευσης (Training data)

Ως $\mathfrak{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\} \subset \mathbb{R}^N \times \mathbb{R}$ ορίζουμε ένα σύνολο από δεδομένα εκπαίδευσης, όπου $x_i = \{x_{i1}, x_{i2}, ..., x_{iN}\}$ υποδεικνύει ένα διάνυσμα εισόδου N διαστάσεων όπου κάθε στοιχείο του διανύσματος αποτελεί και μία παράμετρο (feature) και $y_i$ η αντίστοιχη έξοδος (target variable). Για παράδειγμα, το $x_i$ θα μπορούσε να περιέχει παρελθοντικές τιμές του φορτίου και καιρικά δεδομένα ενώ η $y_i$ είναι η αντίστοιχη τιμή του φορτίου για την χρονική στιγμή που μελετάμε.

### 3.1.2 Σχέση πολυπλοκότητας μοντέλου- σφάλματος (Bias-Variance Tradeoff)

Όταν ασχολούμαστε με προβλήματα πρόβλεψης πρέπει πρώτα να κατανοήσουμε τα σφάλματα της πρόβλεψης. Οι δύο μεταβλητές που προσδιορίζουν το σφάλμα είναι:

- **Απόκλιση (Bias)**
- **Διασπορά (Variance)**

Υπάρχει ένα αντιστάθμισμα προσπαθώντας το μοντέλο να ελαχιστοποιήσει αυτές τις δύο μεταβλητές. Κατανοώντας την σημασία τους, εξυπηρετούν στην ανάπτυξη μοντέλων μεγαλύτερης ακρίβειας και την αποφυγή περιπτώσεων υπερπροσαρμογής υπό προσαρμογής. Πιο συγκεκριμένα για τις δύο συνθήκες:

**Απόκλιση (Bias)**

Η απόκλιση είναι η διαφορά μεταξύ της πρόβλεψη μοντέλου και της πραγματικής τιμής. Μοντέλα με υψηλή απόκλιση δεν επικεντρώνονται στα δεδομένα εκπαίδευσης και υπέρ απλοποιούν το μοντέλο. Επίσης, οδηγεί σε μεγάλα σφάλματα κατά την εκπαίδευση και την αξιολόγηση.

**Διασπορά**

Η διασπορά είναι η διακύμανση των προβλέψεων του μοντέλου ως προς μία τιμή. Μοντέλα με μεγάλη διασπορά δίνουν μεγάλη σημασία στα δεδομένα εκπαίδευσης και δεν έχουν καλές δυνατότητες σε γενίκευσης σε άγνωστα δεδομένα.

Μαθηματικά, μπορούμε να φτάσουμε στο μοντέλο πολυπλοκότητας-σφάλματος ξεκινώντας από μία απλή σχέση:

$$Y = f(x) + \epsilon$$

όπου $\epsilon$ ένας όρος σφάλματος με κανονική κατανομή και μέση τιμή 0.

Ορίζουμε το τετραγωνισμένο σφάλμα μεταξύ της πραγματικής τιμής $Y$ και της πρόβλεψης μέσα από την $\hat{f(x)}$ ως

$$Err(x) = E[(Y - (x))^2]$$

Το $Err(x)$ μπορεί να επεκταθεί περαιτέρω ως:

$$Err(x) = \big(E|\hat{f(x)}| - f(x)\big)^2 + E\big[(\hat{f(x)} - E|\hat{f(x)}|)^2\big] + \sigma_e$$

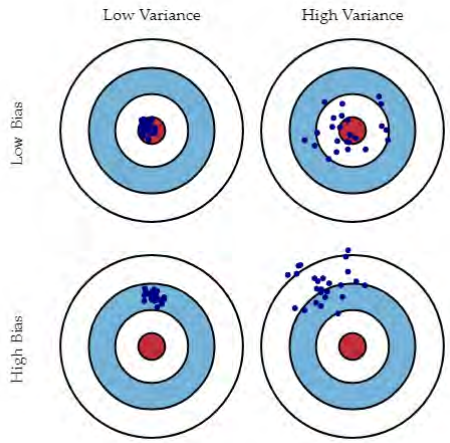
$$Err(x) = Bias^2 + Variance + IrreducibleError$$

όπου $Err(x)$ είναι το άθροισμα των τετραγωνισμένων αποκλίσεων, της διασποράς και ενός σφάλματος που δεν μπορεί να ελαττωθεί περισσότερο και είναι ένας δείκτης του θορύβου στα δεδομένα μας. Συνεπώς, όσο καλό να είναι το μοντέλο μας θα έχουμε πάντα ένα σφάλμα το οποίο δε θα μπορεί να απαλειφθεί.

Η εικόνα 3.1 απεικονίζει διαισθητικά την σχέση πολυπλοκότητας μοντέλου- σφάλματος. Στο κέντρο απεικονίζει ένα μοντέλο το οποίο προβλέπει τέλεια τις σωστές τιμές. Όσο φεύγουμε από τον στόχο, οι προβλέψεις γίνονται ολοένα και χειρότερες. Επαναλαμβάνοντας την διαδικασία προσπαθούμε να πετύχουμε περισσότερες φορές τον στόχο.

Στην επιτηρούμενη μάθηση ο όρος **underfitting** δηλώνει την αδυναμία του μοντέλου να προσαρμοστεί στα δεδομένα εκπαίδευσης. Αυτά τα μοντέλα συνήθως έχουν υψηλό bias και χαμηλο variance. Συμβαίνει στις περιπτώσεις όπου δεν έχουμε αρκετα δεδομένα για εκπαίδευση ή όταν χρησιμοποιούμε ένα γραμμικό μοντέλο για μη γραμμικά προβλήματα.

Ο όρος **overfitting** δηλώνει και τον θόρυβο εκτός από τα δεδομένα. Αυτό συμβαίνει όταν έχουμε ένα αρκετά αλλοιωμένο σετ δεδομένων. Αυτά το μοντέλα είναι περισσότερο σύνθετα από την περίπτωση της παλινδρόμησης, όπως είναι τα Δέντρα απόφασης τα οποία έχουν την τάση να υπερποσαρμόζουν τα δεδομένα.




Σχήμα 3.1: Γραφική αναπαράσταση bias-variance.
[4]

### 3.1.3 Μετρικές αξιολόγησης πρόβλεψης

Η επίδοση της πρόβλεψης φορτίου αξιολογείται από τις μετρικές απόδοσης. Στην ουσία, η μετρική σφάλματος προσδιορίζει το σφάλμα μεταξύ του προβλεπόμενου φορτίου και της πραγματικής παρατήρησης. Διαφορετικές μετρικές αξιολόγησης έχουν χρησιμοποιηθεί για την μέτρηση επίδοσης διαφορετικών τεχνικών πρόβλεψης, στον πίνακα 3.1 παρουσιάζονται οι πιο διαδεδομένοι.

Πίνακας 3.1: Μετρικές ακρίβειας πρόβλεψης.

| Μετρικές | Τύπος |
|---|---|
| Μέσο τετραγωνικό σφάλμα (MSE) | $\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}$ |
| Ρίζα του μέσου τετραγωνικού σφάλματος(RMSE) | $\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$ |
| Μέσο απόλυτο σφάλμα (MAE) | $\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$ |
| Μέσο ποσοστιαίο απόλυτο σφάλμα (MAPE) | $\frac{100}{n}\sum_{i=1}^{n}\lvert\frac{y_i - \hat{y}_i}{y_i}\rvert$ |
| Συμμετρικό ποσοστιαίο απόλυτο σφάλμα (SMAPE) | $\frac{100}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{(\lvert y_i\rvert + \lvert\hat{y}_i\rvert)/2}$ |

Κατά κόρον προτιμούνται οι μετρικές των MAPE και SMAPE καθώς μπορούν να εφαρμοστούν για δεδομένα διαφορετικής κλίμακας σε αντίθεση με τα MSE και RMSE που προϋποθέτουν δεδομένα της ίδιας κλίμακας. Ένα μειονέκτημα που πιθανώς να είχε να αντιμετωπίσει το MAPE είναι οι κοντά στο μηδέν ή οι μηδενικές τιμές των παρατηρήσεων μπορεί να προκύψουν στο δείγμα με αποτέλεσμα την εκτίναξη των ποσοστών των σφαλμάτων. Ωστόσο σε συγκεντρωτικό επίπεδο δεν υφίσταται κάτι τέτοιο. Ο δείκτης SMAPE συνίσταται σε περιπτώσεις όπου η διαφορά μεταξύ

πραγματικής και προβλεπόμενης τιμής είναι μεγάλη. Έτσιι μπορεί να αποφύγει ενδείξεις μεγάλων σφαλμάτων.

## 3.2 Γραμμική Παλινδρόμηση (Linear Regression)

Ως μοντέλο σύγκρισης ως προς την απόδοση των αλγορίθμων, θεωρούμε το πιο απλό μοντέλο προβλέψεων, αυτό της γραμμικής παλινδρόμησης.

Το μοντέλο της γραμμικής παλινδρόμησης είναι ένα ντετερμινιστικό μοντέλο. Δηλαδή οι εξισώσεις αυτού του τύπου επιτρέπουν τον καθορισμό της τιμής της εξαρτημένης μεταβλητής από την τιμή της ανεξάρτητης μεταβλητής όπως και κάποιου σφάλματος μέτρησης. Ωστόσο, στην πράξη είναι ελάχιστες οι περιπτώσεις όπου αυτές οι δύο μεταβλητές έχουν τέλεια γραμμική σχέση καθώς υπάρχουν και άλλοι παράγοντες που μπορεί να επηρεάζουν την εξαρτημένη μεταβλητή και που μπορεί να μην είναι πολλές φορές μετρήσιμοι. Οπότε στο μοντέλο πρέπει να εισαχθεί και το στοιχείο της τυχαιότητας που το μετατρέπει σε ένα μοντέλο πιθανότητας ή αλλιώς στοχαστικό. Για την πρόβλεψη χρονοσειρών υποθέτουμε, παρόλα αυτά ότι η σχέση μεταξύ αυτών των δύο μεταβλητών είναι γραμμική. Οι μαθηματικές σχέσεις που περιγράφουν αυτό το μοντέλο πρόβλεψης είναι οι ακόλουθες:

- **Απλή γραμμική παλινδρόμηση:**

$$\begin{aligned} \hat{y_i} &= a + b * X_i \\ a &= \overline{y} - b * \overline{(X)} \\ b &= \frac{\sum_{i=1}^{n}[(X_i - \overline{X}) * (y_i - \overline{y})]}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \end{aligned} \tag{3.1}$$

  $X_i$: Ανεξάρτητη μεταβλητή

  $y_i$: Εξαρτημένη μεταβλητή

  $\hat{y_i}$: Εκτιμώμενη τιμή της εξαρτημένης μεταβλητής

  a,b: Συντελεστές που προσδιορίζουν την ευθεία (regression coefficients)

  $\overline{X}$: Μέση τιμή του X

  $\overline{y}$: Μέση τιμή του y

  n: Πλήθος παρατηρήσεων

  Τα α,β ονομάζονται εκτιμήσεις ελαχίστων τετραγώνων των συντελεστών παλινδρόμησης.

- **Πολλαπλή παλινδρόμηση (multiple regression):**

$$\hat{y_i} = b_0 + b_1 * X_{1i} + b_2 * X_{2i} + ... + b_k * X_{ki} + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{3.2}$$

Στην πολλαπλή γραμμική παλινδρόμηση, το $y$ μπορεί να επηρεάζεται από περισσότερες ανεξάρτητες μεταβλητές . Έτσι η τιμή του φορτίου μπορεί να αναπαρασταθεί με τη μορφή της εξ.(3.2),

όπου y θεωρούμε το φορτίο, $X_k$ οι παράμετροι που το επηρεάζουν, $b_k$ οι συντελεστές της παλινδρόμησης ως προς το $X_i$ και $\epsilon_i$ ένας όρος σφάλματος με μηδενική μέση τιμή και σταθερή διασπορά.

Η παραπάνω εξίσωση μπορεί να αναπαρασταθεί και στην απλούστερη μορφή

$$Y = X * b + \epsilon$$

όπου

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,k} \\ 1 & X_{2,1} & \cdots & X_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,k} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Εφαρμόζοντας την μέθοδο των ελαχίστων τετραγώνων με όρους γραμμικής άλγεβρας έχουμε:

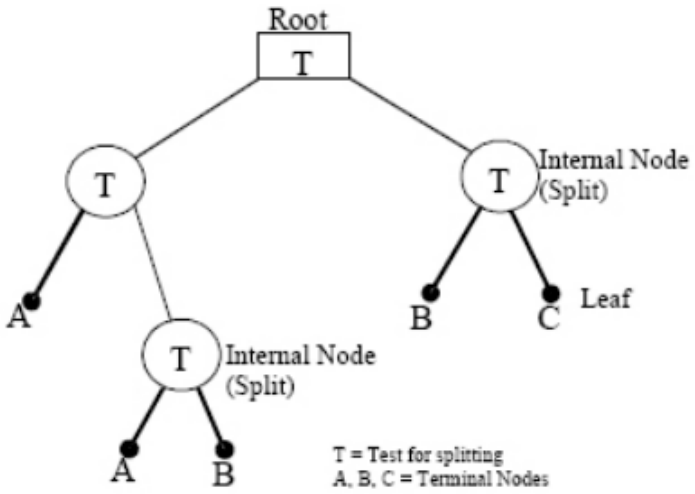
$$b = (X^T X)^{-1}(X^T Y)$$

Αφού υπολογιστούν οι παράμετροι, αυτό το μοντέλο μπορεί να χρησιμοποιηθεί για πρόβλεψη.

## 3.3 Δέντρα αποφάσεων (Decision trees)

### 3.3.1 Ανάλυση CART

Ένα δέντρο απόφασης είναι ένα μη-παραμετρικό προγνωστικό μοντέλο το οποίο μπορεί να έχει τον ρόλο ενός ταξινομητή (classifier) ή μιας μηχανής παλινδρόμησης (regressor). Η δημοφιλέστερη ανάλυση με βάση αυτό το μοντέλο λέγεται CART (Classification and Regression Tree analysis) (Breiman et.al.). Η CART ανάλυση στοχεύει στην πρόβλεψη μίας ποσοτικής εξαρτημένης μεταβλητής για προβλήματα παλινδρόμησης ή μίας κατηγορικής εξαρτημένης μεταβλητής για προβλήματα κατηγοριοποίησης, από ένα σύνολο ανεξαρτήτων μεταβλητών με τη μορφή δενδρικής ιεραρχίας. Επί της ουσίας, ένα δέντρο ταξινόμησης χωρίζει τα δεδομένα σε υποσύνολα με βάσει την ομοιογένεια, αποκόπτοντας έτσι τον 'θόρυβο' κάνοντάς το πιο 'καθαρό'. Έτσι προκύπτει και το κριτήριο 'καθαρότητας'. Στην περίπτωση που η εξαρτημένη μεταβλητή δεν έχει κλάσεις (κατηγορική), ένα μοντέλο παλινδρόμησης εφαρμόζεται στο σετ των ανεξάρτητων μεταβλητών, απομονώνοντας εκείνες τις μεταβλητές υπό την μορφή κόμβων οι οποίες κατά το τρέξιμο του αλγορίθμου μειώνουν το σφάλμα πρόβλεψης. Η ανάπτυξη του δέντρου είναι δυαδικής μορφής όπου το σύνολο των δεδομένων υποδιαιρείται αναδρομικά δημιουργώντας κόμβους/περιοχές βασιζόμενο σε μία συνθήκη 'αν-τότε'. Για παράδειγμα, αν ισχύει για μια μεταβλητή η συνθήκη $T$: $X_i \leq 0.5$, όπου το δεξί μέρος της εξίσωσης είναι η ανεξάρτητη μεταβλητή και στο αριστερό η εξαρτημένη, τότε οποιαδήποτε παρατήρηση υπακούει σε αυτή την συνθήκη κατατάσσεται αριστερά του σημείου διχοτόμησης ενώ για αντίθετη περίπτωση τοποθετείται δεξιά. Το τελευταίο επίπεδο για κάθε κλάδο καταλήγει σε μία συγκεκριμένη μεταβλητή που συνήθως αναφέρεται και ως φύλλο. Αυτή η μεταβλητή αντιπροσωπεύει τον υπολογισμό της εξαρτημένης μεταβλητής $y_i$ μέσα σε αυτή την περιοχή. Ένα παράδειγμα δέντρου παλινδρόμησης φαίνεται στο διάγραμμα 3.2.




Σχήμα 3.2: Δομή CART

**Μαθηματική αναπαράσταση CART**

Έχοντας ένα διάνυσμα $x_i \in \mathbb{R}^n$, i=1,...,l και μία εξαρτημένη μεταβλητή $y \in \mathbb{R}^l$, ένα δέντρο απόφασης αναδρομικά διαιρεί το σύνολο δεδομένων οπότε τα δείγματα με τις ίδιες ετικέτες να ομαδοποιούνται μαζί.

Ορίζουμε τα δεδομένα στον κόμβο $m$ ως $Q$. Για κάθε υποψήφια τμήση $\theta = (j, t_m)$ αποτελούμενη από μία παράμετρο (feature) $j$ και ένα όριο $t_m$, διαχωρίζουμε τα δεδομένα σε 2 υποσύνολα $Q_{left}(\theta)$ και $Q_{right}(\theta)$

$$\begin{aligned} Q_{left}(\theta) &= (x,y)|x_j \leq t_m \\ Q_{right}(\theta) &= Q \setminus Q_{left}(\theta) \end{aligned} \tag{3.3}$$

Οι κόμβοι διαχωρίζονται με βάση τον δείκτη νοθείας (impurity index), δηλαδή πόσο ομοιογενώς είναι χωρισμένες οι κλάσεις σε ένα δέντρο ταξινόμησης ή σε ένα δέντρο παλινδρόμησης βάσει ενός δοσμένου κόμβου, πόσο καλά/άσχημα τα δεδομένα του προσεγγίζουν το μοντέλο. Σε ένα δέντρο παλινδρόμησης για παράδειγμα, ο δείκτης νοθείας υπολογίζεται με το άθροισμα τετραγώνων υπολοίπων μέσα σε αυτό τον κόμβο. Για το δέντρο ταξινόμησης υπάρχουν πολλοί διάφοροι δείκτες λανθασμένης κατηγοριοποίησης όπως είναι ο δείκτης Gini και η εντροπία.

Αν θέλουμε,δηλαδή, να υπολογίσουμε τον δείκτη νοθείας στο κόμβο $m$ μέσω μιας συνάρτησης $H()$, εξαρτάται το πρόβλημα το οποίο πάμε να λύσουμε (ταξινόμηση ή παλινδρόμηση).

$$G(Q,\theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta) + \frac{n_{right}}{N_m} H(Q_{left}(\theta) \tag{3.4}$$

Επιλέγουμε τις παραμέτρους που ελαχιστοποιούν το σφάλμα.

$$\theta^* = argmin_\theta G(Q,\theta)$$

Εκτελείται αναδρομικά ο αλγόριθμος για τα υποσύνολα $Q_{left}(\theta)^*$ και $Q_{right}(\theta)^*$ μέχρι να φτάσει το μέγιστο επιτρεπόμενο βάθος, $N_m < min_{samples}$ ή $N_m = 1$ .

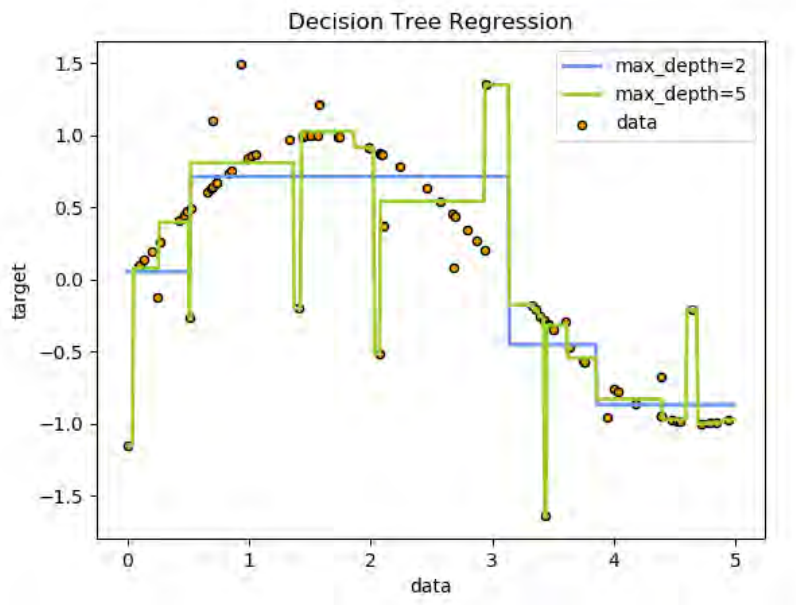
Για την περίπτωση της παλινδρόμησης όπου η τιμή της εξαρτημένης είναι συνεχής, για το κόμβο $m$, αντιπροσωπεύοντας μια περιοχή $R_m$ με $N_m$ παρατηρήσεις, τα κριτήρια τα οποία ελαχιστοποιούν το σφάλμα πρόβλεψης είναι το μέσο τετραγωνικό σφάλμα, το οποίο ελαχιστοποιεί το σφάλμα L2, υπολογίζοντας τις μέσες τιμές στους τερματικούς κόμβους.

Μέσο Τετραγωνικό Σφάλμα (MSE):

$$\begin{aligned} \overline{y_m} &= \frac{1}{N_m} \sum_{i \in N_m} y_i \\ H(X_m) &= \frac{1}{N_m} \sum_{i \in N_m} (y_i - \overline{y}_m)^2 \end{aligned} \tag{3.5}$$

όπου $X_m$ είναι τα δεδομένα προς εκπαίδευση στο κόμβο $m$.

Στο παρακάτω σχήμα 3.3 απεικονίζεται η προσσέγγιση ενός ημιτόνου με την μέθοδο των δέντρων παλινδρόμησης.



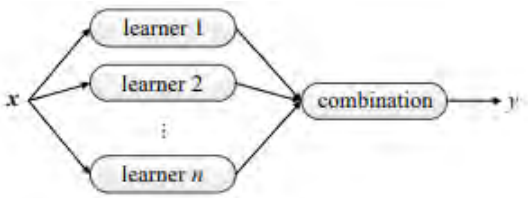

Σχήμα 3.3: Δέντρο παλινδρόμησης

Ο παραπάνω αλγόριθμος χωρίζει σε περιοχές το σύνολο των δεδομένων και μεγαλώνει έως το σημείο που δεν χρήζει περαιτέρω διαίρεσης. Όσο το δέντρο μεγαλώνει, τα δεδομένα εκπαίδευσης χωρίζονται σε ολοένα και μικρότερα δείγματα. Διαισθητικά, το δέντρο θα συνεχίζει να διαιρείται σε κόμβους έως ότου απομείνει ένα δείγμα σε κάθε κόμβο όπου το σφάλμα είναι μηδενικό. Αυτό είναι πιθανό να οδηγήσει σε προβλήματα υπερπροσαρμογής (overfitting). Δηλαδή ενώ αποδίδει εξαιρετικά στα δεδομένα εκπαίδευσης, όταν τροφοδοτείται με άγνωστα δεδομένα αποτυγχάνει την προσέγγισή τους.

Για την αντιμετώπιση της υπερπροσαρμογής υπάρχουν δύο κύριες τακτικές για τα δέντρα απόφασης. Αρχικά το δέντρο μπορεί να 'κλαδευτεί' εκ των προτέρων. Αυτό σημαίνει πως το δέντρο μεγαλώνει έως ένα προκαθορισμένο βάθος ή όταν κάθε κόμβος-φύλο περιέχει ένα συγκεκριμένο πλήθος δεδομένων. Στην δεύτερη περίπτωση, το δέντρο μπορεί να 'κλαδευτεί' εκ των υστέρων. Συνπεώς, τα φύλα περιέχουν μόνο ένα δείγμα και με τη χρήση ενός συνόλου δεδομένων προς αξιολόγηση (validation set), αφαιρούνται οι κόμβοι από κάτω προς τα πάνω εάν η ακρίβεια πρόβλεψης με βάση τα δεδομένα του σετ αξιολόγησης είναι ίδια ή καλύτερη με αυτήν του ακλάδευτου δέντρου. Ωστόσο, αυτές οι διαδικασίες απαιτούν μεγάλα ποσά χρόνου και επεξεργαστικής ισχύος για να εφαρμοστούν. Για αυτό το λόγο έχουν προταθεί διάφορες τεχνικές για την βελτίωση της επίδοσής τους.

### 3.3.2 Συνδυαστικές μέθοδοι (Ensemble methods)

Οι συνδυαστικές μέθοδοι εκμάθησης εκπαιδεύουν πολλαπλά μοντέλα εκμάθησης (learners/classifiers) για να επιλύσουν το ίδιο πρόβλημα. Σε αντίθεση με την συνηθισμένη προσέγγιση η οποία προσπαθεί να κατασκευάσει ένα μοντέλο εκμάθησης από τα δεδομένα εκπαίδευσης, τα σύνολα εκμάθησης προσπαθούν να κατασκευάσουν ένα σετ από μοντέλα τα οποία συνδυάζονται για την παραγωγή της τελικής πρόβλεψης/ταξινόμησης.




Σχήμα 3.4: Δομή Συνόλου εκμάθησης.

Η εικόνα 3.4 δείχνει μία τυπική δομή ενός συνόλου εκμάθησης το οποίο αποτελείται ένα σύνολο βασικών μοντέλων εκμάθησης (base learners). Αυτά τα μοντέλα παράγονται από τα δεδομένα εκπαίδευσης μέσα από κάποιο βασικό αλγόριθμο εκμάθησης όπως είναι για παράδειγμα τα δέντρα απόφασης. Τα σύνολα εκμάθησης έχουν βρει απήχηση, κυρίως γιατί μπορούν να ενισχύουν ασθενή μοντέλα (weak learners), που στην χειρότερη περίπτωση αποδίδουν καλύτερα από μία τυχαία πρόβλεψη, σε ισχυρά μοντέλα (strong learners) τα οποία μπορούν να κάνουν ακριβείς προβλέψεις.

#### Bagging (Bootstrap Aggregating)

Η μέθοδος του Bagging (Bootstrap Aggregating) χρησιμοποιείται για την βελτίωση της επίδοσης των δέντρων απόφασης. Η ιδέα είναι να οριστεί ένα πλήθος υποσυνόλου των δεδομένων προς εκπαίδευση με τη μέθοδο της δειγματοληψίας με επανάθεση (Bootstraping) και το κάθε υποσύνολο εκπαιδεύεται με ένα ασθενή αλγόριθμο (weak learner) και στο τέλος αξιολογούνται για την παραγωγή ενός τελικού ισχυρού μοντέλου (strong) μέσω ψηφοφορίας για προβλήματα ταξινόμησης ή υπολογισμού μέσου όρου για προβλήματα παλινδρόμησης.

Αλγοριθμικά η τεχνική Bagging μπορεί να οριστεί ως εξής:

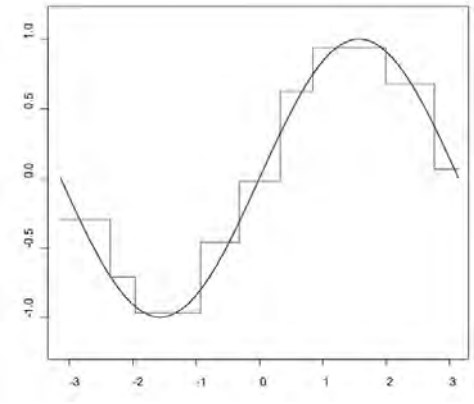
Έχοντας ένα σετ εκπαίδευσης $X = x_1, \ldots, x_n$ $, Y = y_1, \ldots, y_n$ η μέθοδος bagging επαναλαμβανόμενα (B φορές) επιλέγει ένα τυχαίο δείγμα μέσω επανάθεσης από το σετ εκπαίδευσης και εφαρμόζει έναν αλγόριθμο δέντρου απόφασης ώστε να προσεγγίσει το δείγμα, δηλαδή:

Για $b = 1, \ldots, B$:

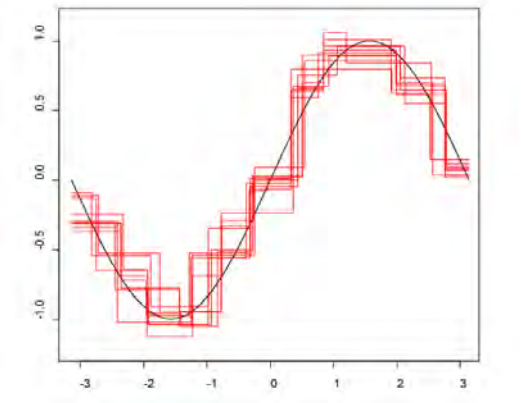
1. Εφαρμογή bootstrapping για $n$ αριθμό στιγμιοτύπων από το $X$, $Y$ και παραγωγή $X_b$, $Y_b$ δειγμάτων.
2. Εκπαίδευση συνάρτησης (ταξινόμησης/παλινδρόμησης) $f_b$ στο $X_b$, $Y_b$.
3. Υπολογίζουμε τον μέσο όρο των προβλέψεων (παλινδρόμηση).

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x)$$

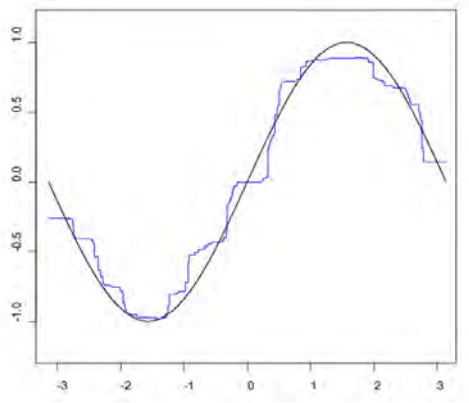
Στις εικόνες 3.5,3.6,3.7 αναπαρίσταται γραφικά η bagging μέθοδος.




Σχήμα 3.5: 1 δέντρο για προσέγγιση ημίτονου




Σχήμα 3.6: 10 δέντρα




Σχήμα 3.7: 100 δέντρα

**Random Forests**

Η μέθοδος Random Forest είναι μία βελτίωση της μεθόδου bagging για δέντρα απόφασης. Ένα πρόβλημα με τα δέντρα απόφασης όπως είναι τα CART, είναι πως επιλέγουν τη μεταβλητή διαχωρισμού, χρησιμοποιώντας έναν άπληστο αλγόριθμο ο οποίος ελαχιστοποιεί το σφάλμα. Ακόμη και οι Bagging τεχνικές παρουσιάζουν μεγάλες ομοιότητες στην δομή των δέντρων και παρουσιάζουν μεγάλο συσχετισμό στις προβλέψεις τους. Συνδυάζοντας τις προβλέψεις από πολλαπλά μοντέλα, στην λογική των συνόλων εκμάθησης, σημειώνεται καλύτερη επίδοση αν οι προβλέψεις από τα υποσύνολα είναι στην καλύτερη περίπτωση μη συσχετισμένα ή παρουσιάζουν μία ελαφριά συσχέτιση.

Η μέθοδος Random Forest αλλάζει το αλγόριθμο έτσι ώστε τα υπο-δέντρα μαθαίνουν με τέτοιο τρόπο που οι τελικές προβλέψεις τους σε σχέση με τα υπόλοιπα υπο-δέντρα έχουν μικρό συσχετισμό.

Στα CART, όταν επιλέγεται ο κόμβος στον οποίο γίνεται η τμήση, ο αλγόριθμος εκμάθησης έχει την δυνατότητα να ελέγξει όλες τις μεταβλητές και τις τιμές τους ώστε να επιλέξει το πιο βέλτιστο σημείο τμήσης. Ο αλγόριθμος Random Forest αλλάζει αυτή την διαδικασία με τέτοιο τρόπο που ο αλγόριθμος περιορίζεται σε ένα τυχαίο δείγμα από παραμέτρους στο οποίο κοιτάει.

Ο αριθμός των παραμέτρων τις οποίες μπορεί να ψάξει σε κάθε σημείο διαχωρισμού (m) πρέπει να οριστεί ως παράμετρος κατά την κλήση του αλγορίθμου. Μπορούν να δοκιμαστούν διάφορες παράμετροι και να τον ρυθμίσουμε έτσι μέσω τις μεθόδου cross-validation.

- Για ταξινόμηση μία καλή προεπιλογή είναι $m$=sqrt($p$)
- Για παλινδρόμηση μία καλή προεπιλογή είναι $m=p/3$

Όπου $m$ είναι οι τυχαία επιλεγμένες παράμετροι στο σημείο διαχωρισμού και $p$ ο αριθμός των μεταβλητών εισόδου.

**ExtraTrees**

Εφαρμόζοντας ένα ακόμα βήμα τυχαιότητας στη μέθοδο Random Forest προκύπτουν τα extremely randomized trees ή ExtraTrees. Παρουσιάζουν μία ομοιότητα με τα Random Forests ως

προς το ότι και οι δύο μέθοδοι συνδυάζουν σύνολα ξεχωριστών δέντρων, ωστόσο έχουν δύο βασικές διαφορές: κατά πρώτον, κάθε δέντρο εκπαιδεύεται με τα δεδομένα ολόκληρου του δείγματος εκπαίδευσης (σε αντίθεση με το bootstrap δείγμα), κατά δεύτερον, η από πάνω προς τα κάτω διχοτόμηση του δέντρου είναι τυχαία. Αντί δηλαδή να υπολογίζεται ένα τοπικά βέλτιστο σημείο τμήσης για κάθε παράμετρο που υπάρχει στο δείγμα, ένα τυχαίο σημείο τμήσης επιλέγεται. Από όλες τις τυχαία παραγόμενες τομές , επιλέγεται αυτή που φέρει το υψηλότερο σκορ. Δεν υπάρχει μεγάλη διαφορά ως προς τις επιδόσεις πρόβλεψης σε σχέση με την Random Forest, ωστόσο έχει παρατηρηθεί ότι ο χρόνος υπολογισμού είναι μικρότερος.

### Gradient Boosting

Η gradient boosting είναι μία τεχνική μηχανικής μάθησης για σκοπούς ταξινόμησης και παλινδρόμησης, η οποία παράγει ένα μοντέλο πρόβλεψης υπό την μορφή ενός συνδυασμού προβλέψεων από ασθενή μοντέλα (συνηθίζονται τα δέντρα απόφασης) με στόχο την παραγωγή ενός ισχυρού μοντέλου. Ωστόσο, το μοντέλο χτίζεται διαδοχικά και όχι παράλληλα όπως στην περίπτωση της bagging μεθόδου και γενικεύεται για διάφορα προβλήματα επιτρέποντας την βελτιστοποίηση μια αυθαίρετης συνάρτησης κόστους.

Για να εξηγήσουμε καλύτερα τη μέθοδο του gradient boosting θα την αναπτύξουμε στα πλαίσια τη παλινδρόμησης ελαχίστων τετραγώνων (least-square regression). Σε αυτή την περίπτωση, σκοπός είναι να μάθουμε στο μοντέλο $F$ να προβλέπει τις τιμές στη μορφή $\hat{y} = F(x)$ ελαχιστοποιώντας το μέσο τετραγωνικό σφάλμα $\frac{1}{n}\sum_i (\hat{y}_i - y_i)^2$, όπου $i$ ο δείκτης ενός σετ εκπαίδευσης $\mathfrak{D}$ μεγέθους $n$.

Σε κάθε στάδιο $m$, $1 \leq m \leq M$, του gradient boosting, μπορεί να θεωρηθεί ότι υπάρχει ένα ατελές μοντέλο $F_m$. Ο gradient boosting αλγόριθμος βελτιώνει το $F_m$ κατασκευάζοντας ένα νέο μοντέλο το οποίο προσθέτει μία εκτιμήτρια συνάρτηση (estimator) $h$ για δημιουργήσει ένα καλύτερο μοντέλο: $F_{m+1}(x) = F_m(x) + h(x)$. Για να βρεθεί η $h$ , αλγόριθμος gradient boosting ξεκινά την παρατήρηση με την συνθήκη:

$$F_{m+1}(x) = F_m(x) + h(x) = y$$

ή αντίστοιχα,

$$h(x) = y - F_m(x)$$

'Ετσι, ο αλγόριθμος gradient boosting θα προσεγγίσει την $h$ με βάση το υπόλοιπο $y - F(x)$. Στη γενικότερη περίπτωση των τεχνικών boosting, κάθε $F_{m+1}$ προσπαθεί να διορθώσει το σφάλμα του αμέσως προηγούμενου βήματος $F_m$.

Ο γενικός αλγόριθμος της gradient boosting είναι ο ακόλουθος:

**Algorithm 1** Gradient Boosting

**Είσοδος:** ένα training set $\{(x_i, y_i)\}_{i=1}^n$, μία συνάρτηση κόστους $L(y, F(x))$, ένας αριθμός επαναλήψεων $M$

1. Αρχικοποίησε το μοντέλο με μία σταθερή τιμή : $F_0(x) = argmin_\gamma \sum_{i=1}^n L(y_i, \gamma)$
2. For $m = 1$ to $M$:
   (α΄) Υπολόγισε τα υπόλοιπα:
   $$r_{im} = -\alpha \Big[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\Big]_{F(x)=F_{m-1}(x)}$$
   for $i = 1, \ldots, n$
   όπου το α είναι ο ρυθμός μάθησης(learning rate)
   (β΄) Χρησιμοποίησε έναν ασθενή εκτιμητή $h_m(x)$ (δέντρο παλινδρόμησης) με στόχο την εξαρτημένη μεταβλητή $r_im$, δηλαδή εκπαίδευσε το μοντέλο χρησιμοποιώντας το training σετ $\{x_i, r_{im}\}_{i=1}^n$.
   (γ΄) Υπολόγισε το $\gamma_m$
   $$\gamma_m = argmin_\gamma \sum i = 1^n L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)$$
   .
   (δ΄) Ανανέωσε το μοντέλο
   $$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$
3. **Έξοδος:** $\hat{F}(x) = F_M(x)$

## 3.4 Μηχανές διανυσματικής υποστήριξης (Support Vector Machines)

### 3.4.1 Η βασική ιδέα

Η ιδέα της $\epsilon$-SV παλινδρόμησης είναι να βρει μία συνάρτηση $f(x)$ η οποία έχει το πολύ μία $\epsilon$ απόκλιση από τις πραγματικές τιμές $y_i$ από τα δεδομένα εκπαίδευσης και παράλληλα να είναι όσο πιο 'επίπεδη' γίνεται. Με άλλα λόγια, δεν μας ενδιαφέρουν τα σφάλματα όσο αυτά είναι μικρότερα του $\epsilon$, αλλά δεν είναι επιτρεπτή μεγαλύτερη απόκλιση από αυτή.

Στην απλούστερη περίπτωση περιγράφουμε τη περίπτωση μιας γραμμικής συνάρτησης μορφής:

$$f(x) = \langle w, x \rangle + b, \text{με w} \in \mathbb{X}, b \in \mathbb{R} \tag{3.6}$$

Ο όρος 'επίπεδη' ερμηνεύεται ως η αναζήτηση μικρών βαρών $w$. Ένας τρόπος να επιτευχθεί αυτό είναι να ελαχιστοποιήσουμε τη νόρμα $||w||^2 = \langle w, w \rangle$. Αυτό το πρόβλημα μπορεί να εκφραστεί και ως ένα πρόβλημα κυρτής βελτιστοποίησης (convex optimization) ως:

$$\begin{aligned} &\text{ελαχιστοποίησε:} \quad \frac{1}{2}||w||^2 \\ &\text{τέτοιο ώστε:} \begin{cases} y_i - \langle w, x_i \rangle - b \leq \epsilon \\ \langle w, x_i \rangle + b - y_i \leq \epsilon \end{cases} \end{aligned} \tag{3.7}$$

Στη 3.7 υποθέσαμε ότι υπάρχει μια $f$ η οποία υπολογίζει όλα τα ζεύγη $(x_i, y_i)$ με $\epsilon$ ακρίβεια. Ωστόσο, δεν ισχύει πάντα αυτό ή μπορεί να θέλουμε να επιτρέψουμε στο σύστημά μας ένα βαθμό σφάλματος ορίζοντας κάποια 'χαλαρά' περιθώρια. Αντίστοιχα με τη μέθοδο *Soft Margin (Vapnik et.al.)*, για περιπτώσεις παλινδρόμησης εισάγουμε τις 'χαλαρές' μεταβλητές (slack variables) $\xi_i, \xi_i^*$ για να ικανοποιούνται οι περιορισμοί που τίθενται στα προβλήματα βελτιστοποίησης καταλήγοντας στην παρακάτω μορφή:

$$\begin{aligned} &\text{ελαχιστοποίησε:} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\xi_i, \xi_i^*) \\ &\text{τέτοιο ώστε:} \begin{cases} y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi \\ \langle w, x_i \rangle + b - y_i \leq \epsilon \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \tag{3.8}$$

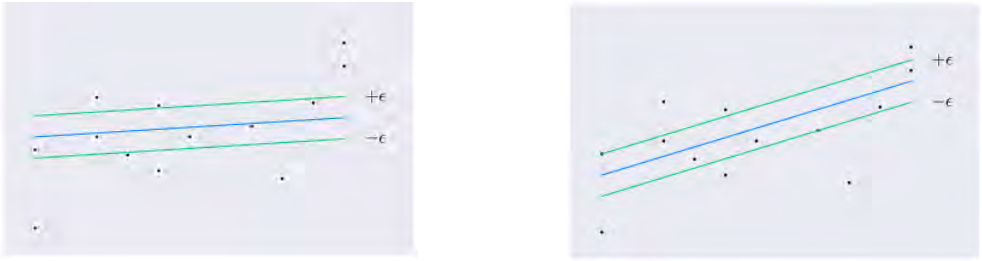
Η σταθερά κανονικοποίησης C>0 καθορίζει το ισοζύγιο μεταξύ του πόσο επίπεδη είναι η $f$ και το επιτρεπτό άνω όριο διακύμανσης σφάλματος μεγαλύτερο από το $\epsilon$. Ορίζουμε ,για αυτό το λόγο, μια συνάρτηση κόστους ($\epsilon-$insensitive loss function) $|\xi|_\epsilon$ τέτοια ώστε:

$$|\xi|_\epsilon = \begin{cases} 0 \quad \text{,if} |\xi| \leq \epsilon \\ |\xi| - \epsilon \quad \text{,σε κάθε άλλη περίπτωση.} \end{cases} \tag{3.9}$$

Ο ρόλος της $C$ φαίνεται γραφικά στο σχήμα 3.8. Μόνο τα σημεία εκτός της σκιαγραφημένης περιοχής συνεισφέρουν στο κόστος που σχετίζεται με την παράμετρο $C\sum_{i=1}^{l}(\xi_i, \xi_i^*)$ στην εξίσωση 3.8. Αν το C είναι μεγάλο τότε η αντίστοιχη βέλτιστη $f(x)$ θα ελαχιστοποιήσει τον αριθμό των σημείων εκτός της σκιαγραφημένης περιοχής ενώ για χαμηλό C κάνει την $f(x)$ πιο επίπεδη αλλά με μεγαλύτερο σφάλμα πρόβλεψης. Για γραμμικά προβλήματα η παραπάνω προσέγγιση συμπεριφέρεται ικανοποιητικά αλλά για μη γραμμικά συνίσταται η μέθοδος της διπλής αναπαράστασης (dual formulation).

### 3.4.2 Dual formulation

Η βασική ιδέα είναι να κατασκευαστούν μία συνάρτηση Lagrange από την πρωτεύουσα αντικειμενική συνάρτηση ( primal objective function) και οι ανάλογοι περιορισμοί, εισάγοντας ένα διπλό σύνολο από μεταβλητές. Η Langrangian συνάρτηση για την 3.8 είναι ως εξής:




(α΄) C=0.1 (β΄) C=1000

Σχήμα 3.8: Επίλυση Soft margin για γραμμικό SVM

$$\begin{aligned} L = \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\xi_i, \xi_i^*) - C\sum_{i=1}^{l}(\eta_i\xi_i, \eta_i^*\xi_i^*) \\ -\sum_{i=1}^{l}\alpha_i(\epsilon + \xi_i - y_i + \langle w, x_i\rangle + b) \\ -\sum_{i=1}^{l}\alpha_i^*(\epsilon + \xi_i + y_i - \langle w, x_i\rangle - b) \end{aligned} \tag{3.10}$$

όπου $\eta_i, \eta_i^*, \alpha_i, \alpha_i^* \geq 0$ είναι πολλαπλασιαστές Lagrange. Έχει αποδειχθεί ότι η λύση για την εξίσωση 3.8 δίνεται από το κρίσιμο σημείο (saddle point) της 3.10 , όπου στόχο έχει να ελαχιστοποιηθούν οι $w, b, \xi_i, \xi_i^*$ και παράλληλα να μεγιστοποιηθούν οι $\eta_i, \eta_i^*, \alpha_i, \alpha_i^*$. Το σημείο όπου υπολογίζεται το ελάχιστο βρίσκεται μέσα από τις μερικές παραγώγους της $L$:

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{l}(\alpha_i - \alpha_i) = 0 \tag{3.11}$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{l}(\alpha_i - \alpha_i)x_i = 0 \tag{3.12}$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \eta_i = 0 \tag{3.13}$$

$$\frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* = 0 \tag{3.14}$$

Αντικαθιστώντας τις εξισώσεις 3.11,3.12,3.13,3.14 στην 3.10 προκύπτει το πρόβλημα διπλής βελτιστοποίησης (dual optimization problem):

$$\begin{aligned} &\text{μεγιστοποίησε:} \begin{cases} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle \\ -\epsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \end{cases} \\ &\text{τέτοιο ώστε:} \quad \begin{cases} \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases} \end{aligned} \tag{3.15}$$

Στην 3.15 οι δίπλες μεταβλητές $\eta_i, \eta_i^*$ απαλείφονται μέσα από τις συνθήκες 3.13,3.14. Συνεπώς η 3.12 γράφεται ως:

$$w = \sum_{i=1}^{*}(\alpha_i - \alpha_i^*)x_i, \quad \text{άρα η } f(x) = \sum_{i=1}^{*}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b \tag{3.16}$$

Η ανάπτυξη της $f(x)$ σε αυτή την μορφή λέγεται ανάπτυγμα διανυσμάτων υποστήριξης (Support Vector expansion). Τα βάρη $w$, δηλαδή, εκφράζονται από ένα γραμμικό συνδυασμό των διανυσμάτων εισόδου $x_i$. Συνεπώς δεν χρειάζεται να υπολογίζουμε ρητά τα $w$ καθώς μπορούμε να εκφράσουμε τον αλγόριθμο με όρους εσωτερικού γινομένου μεταξύ των δεδομένων εισόδου. Αυτή η παρατήρηση είναι χρήσιμη γιατί ,όπως θα αναφερθεί και στην συνέχεια , αντιμετωπίζει και μη γραμμικά προβλήματα μέσω των πυρήνων (kernels).

### 3.4.3 Πυρήνες (Kernels)

Η βασική ιδέα πίσω από την μέθοδο των πυρήνων είναι ότι τα δεδομένα μας, τα οποία δεν είναι διαχωρίσιμα μεταξύ τους στον τρέχον χώρο χαρακτηριστικών (feature space) $n$ διαστάσεων, μπορεί να είναι όμως σε έναν χώρο ανώτερης διάστασης.

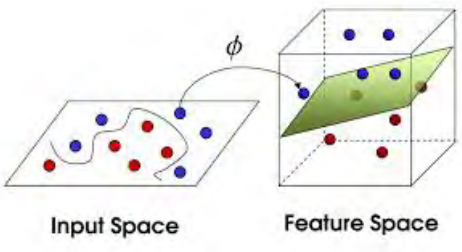
Το παράδειγμα του Vapnik(1995) μας δίνει την απλούστερη εκδοχή αυτού του προβλήματος :

Για να επιλύσουμε το πρόβλημα της προσέγγισης μη γραμμικών δεδομένων οι παράμετροι εισόδου $x_i$ μπορούν να προ επεξεργαστούν μέσω μιας συνάρτησης χαρτογράφησης $\Phi : \mathbb{R}^N \to \mathbb{F}$ στον χώρο χαρακτηριστικών (feature space) όπου εφαρμόζεται ο SV αλγόριθμος που περιγράφεται στη 3.15. Για μία μετάβαση από τον $\mathbb{R}^2$ στον $\mathbb{R}^3$ έχουμε:

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3 \tag{3.17}$$

$$\Phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \tag{3.18}$$

Όπως φαίνεται στο σχήμα 3.9 οι συντεταγμένες των σημείων $x_i$ διαχωρίζονται από το υπερεπίπεδο.




Σχήμα 3.9

Αυτή η προσέγγιση αποδίδει ικανοποιητικά,αλλά για μεγαλύτερες διαστάσεις το κόστος υπολογισμού αυξάνεται δραματικά.

Τη λύση σε αυτό το πρόβλημα δίνει το τρικ του πυρήνα (kernel trick) (Boser et al.). Αντί δημιουργίας μη γραμμικών μετασχηματισμών των παραμέτρων εισόδου $x_i$ και στη συνέχεια των

υπολογισμό των εσωτερικών γινομένων τους στον χώρο χαρακτηριστικών, δύο πρότυπα εισόδου $x_i$ μπορούν να συγκριθούν στις αρχικές διαστάσεις του χώρου εισόδου , μέσω μίας προκαθορισμένης συνάρτησης, πριν γίνει μη γραμμικός μετασχηματισμός των $x_i$. Με βάση το προηγούμενο παράδειγμα το εσωτερικό γινόμενο των προ-επεξεργασμένων παραμέτρων μπορεί να γραφεί ως:

$$\langle (x_1^2, \sqrt{2}x_1x_2, x_2^2), ({x'_1}^2, \sqrt{2}x'_1x'_2, {x'_2}^2) \rangle = \langle x, x' \rangle^2$$

Συνεπώς αρκεί να ξέρουμε την $K(x,x') = \langle \Phi(x_i), \Phi(x'_i)$ παρά ρητά την $\Phi$, το οποίο μας επιτρέπει να τροποποιήσουμε την 3.16 ως:

$$w = \sum_{i=1}^{*} (\alpha_i - \alpha_i^*)\Phi(x_i), \quad \text{άρα η } f(x) = \sum_{i=1}^{*} (\alpha_i - \alpha_i^*)K(x_i, x) + b \tag{3.19}$$

Η διαφορά με την γραμμική περίπτωση είναι ότι το $w$ δεν δίνεται άμεσα. Στις μη γραμμικές περιπτώσεις το πρόβλημα βελτιστοποίησης εμπίπτει στην εύρεση της πιο 'επίπεδης' συνάρτησης όχι στον χώρο εισόδου αλλά στον χώρο χαρακτηριστικών.

Πίνακας 3.2: Τύποι πυρήνων SVR και αντίστοιχες παράμετροι.

| Πυρήνας | $K(x,x')$ | Παράμετροι |
|---|---|---|
| Πολυωνυμικές | $(\langle x, x' \rangle + c)^p$ | $p \in \mathbb{N}, c \geq 0$ |
| Sigmoid | $\tanh(\gamma \langle x, x' \rangle + r)$ | $\gamma > 0, r < 0$ |
| RBF | $e^{-\gamma \|\|x - x'\|\|^2}$ | $\gamma > 0$ |

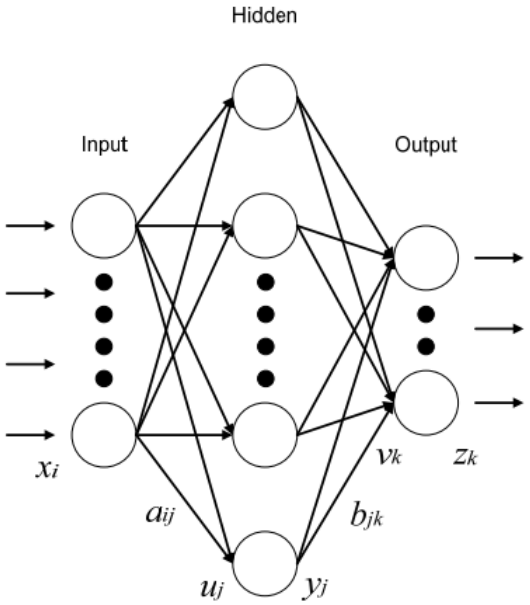
## 3.5 Τεχνητά Νευρωνικά Δίκτυα (ANN)

Τα ANN μιμούνται τον τρόπο εκμάθησης του ανθρώπινου εγκεφάλου μέσα από την σχέση μεταξύ συγκεκριμένων εισόδων και εξόδων με βάση την εμπειρία. Το σχήμα 3.10 δείχνει ένα τυπικό ANN τριών επιπέδων ή στρωμάτων (layers) εμπρόσθιας τροφοδότησης για σκοπούς STLF. Πιο συγκεκριμένα, περιλαμβάνει ένα επίπεδο εισόδου (input layer), ένα κρυμμένο επίπεδο (hidden layer) και ένα επίπεδο εξόδου (output layer). Αυτά είναι διασυνδεδεμένα μεταξύ τους μέσω βαρών (weights), η τιμή των οποίων είναι μεταβλητή. Οι υπολογιστικές μονάδες σε κάθε επίπεδο λέγονται νευρώνες (neurons). Οι κρυμμένοι νευρώνες και οι νευρώνες εξόδου υπολογίζουν τις εξόδους τους μέσω μίας συνάρτησης ενεργοποίησης όπου στα περισσότερα μοντέλα είναι συνάρτηση υπερβολικής εφαπτομένης (tanh) ή μία Γραμμική Συνάρτηση Ράμπας (ReLU).

$$\begin{aligned} tanh : g(x) &= \frac{1}{2 + e^{-2x}} - 1 \\ ReLU : g(x) &= max\{0, x\} \end{aligned} \tag{3.20}$$

Επίσης προστίθεται ένας όρος bias (κατώφλι ή μεροληψία) σε όλους του νευρώνες εκτός του επιπέδου εισόδου.

όπου:




Σχήμα 3.10: ANN εμπρόσθιας τροφοδότησης 3 επιπέδων.

- $x_i = 1, \ldots, I$, είσοδοι για κάθε νευρώνα $i$ του επιπέδου εισόδου.
- Βάρη από τον νευρώνα εισόδου $i$ προς τους κρυμμένους νευρώνες $j$: $a_{i,j}$. Ορίζουμε το $a_{0,j}$ ως το bias του κρυμμένου νευρώνα $j$, με $j = 1, \ldots, J$.
- Είσοδοι στο νευρώνα $j$:$u_j$, με:

$$u_j = a_{0j} + \sum_{i=1}^{I} a_{ij} x_i \tag{3.21}$$

- Έξοδοι του κρυμμένου νευρώνα $j$: $y_j$ ,με:

$$y_j = g(u_j) \tag{3.22}$$

- Βάρη από τον κρυμμένο νευρώνα $j$ προς την έξοδο του νευρώνα $k$: $b_{jk}$. Ως $b_{0k}$ ορίζουμε το bias του νευρώνα εξόδου $j$
- Είσοδοι στο νευρώνα εξόδου $k$: $v_k, k = 1, \ldots, K$, με:

$$v_k = b_{0k} + \sum_{k=1}^{K} a_{kj} y_j \tag{3.23}$$

- Έξοδοι του νευρώνα εξόδου $k$:$z_k$, με:

$$z_k = g(v_k) \tag{3.24}$$

Το ANN πρέπει να ρυθμιστεί έτσι ώστε οι παράμετροι εισόδου να παράγουν το επιθυμητό αποτέλεσμα. Αυτό επιτυγχάνεται με την μεταβολή βαρών μεταξύ των νευρώνων. Ένας τρόπος

καθορισμού των βαρών είναι η άμεση αποτίμησή μέσα από κάποιας γνώσης εκ των προτέρων. Ένας άλλος τρόπος, και ο ποιο συνηθισμένος είναι η εκπαίδευση του δικτύου σε κάποιο σετ εκπαίδευσης που το επιτρέπει να μαθαίνει πρότυπα και να αλλάζει τα βάρη σύμφωνα με κάποιο κανόνα εκμάθησης. Για τους σκοπούς της STLF ανάλυσης, το δίκτυο τροφοδοτείται με τις σχετικές παραμέτρους(ιστορικά δεδομένα φορτίου, καιρός, ημερολογιακοί δείκτες) και παράγει ένα αποτέλεσμα βάσει της μεταβλητής στόχου που του έχουμε ορίσει (π.χ.φορτίο). Η εκμάθηση του μοντέλου επιτυγχάνεται με τον υπολογισμό των συντελεστών/βαρών ($w = \{a_{ij}, bjk\}$), οι οποίο παρέχουν την καλύτερη δυνατή προσέγγιση μεταξύ της εξόδου του ANN ($z$)και της πραγματικής τιμής ($t$). Η οπισθοδιάδοση σφάλματος (backward error propagation) είναι μία από τις απλούστερες και πιο γενικές μεθόδους για την εκπαίδευση πολυεπίπεδων νευρωνικών δικτύων. Κατά την διαδικασία αυτή, για προβλήματα παλινδρόμησης, το μέσο τετραγωνικό σφάλμα ελαχιστοποιείται ως εξής:

$$\text{Min} \quad E = \frac{1}{2KN} \sum_{n=1}^{N} \sum_{k=1}^{K} (z_{kn} - t_{kn})^2 \tag{3.25}$$

με:

- N: ο αριθμός των παρατηρήσεων του σετ δεδομένων.
- K: ο αριθμός των εξόδων του δικτύου.
- $t_{km}$: είναι ο k-στη πραγματική τιμή για το n-οστο στοιχείο.
- $z_{km}$: είναι ο k-στη τιμή πρόβλεψης για το n-οστο στοιχείο.

## 3.6 K-means Clustering

Ο όρος clustering αναφέρεται στη μέθοδο ομαδοποίησης των παρατηρήσεων του dataset σε υπο-ομάδες, όπου οι παρατηρήσεις σε αυτό το group είναι περισσότερο όμοιες μεταξύ τους σε σχέση με άλλα cluster. Σε αντίθεση με τις μεθόδους SVR και Random Forest, οποίες αποτελούν μεθόδους επιτηρούμενης μάθησης (supervised learning) , η μέθοδος clustering είναι μη επιτηρούμενη (unsupervised). Στην επιτηρούμενη μάθηση προσπαθούμε να εξάγουμε σχέσεις μεταξύ των παραμέτρων και των εξαρτημένων μεταβλητών εκπαιδεύοντας τον αλγόριθμο, όπου ο στόχος είναι γνωστός. Στην μη επιτηρούμενη, επίσης προσπαθούμε να ανακαλύψουμε μια δόμη ή μια σχέση, ωστόσο δεν υπάρχει κάποια σαφής απάντηση ως προς το πώς πρέπει να υλοποιηθεί ή ποια είναι η σωστή απάντηση.

Ανάμεσα στις πιο γνωστές μεθόδους clustering είναι η μέθοδος K-means. Σκοπός είναι να διαχωρίσουμε το dataset σε ένα προκαθορισμένο αριθμό clusters . Αφού επιλεχθεί το $K$, κάθε παρατήρηση ανατίθεται σε ακριβώς ένα από τα $K$ clusters. Η καλύτερη μέθοδος clustering είναι αυτή που ελαχιστοποιεί μέσα σε κάθε cluster την ομοιότητα, υπολογισμένη, συνήθως μέσα από την τετραγωνική Ευκλείδια απόσταση.

Έχοντας ένα σύνολο παρατηρήσεων $\mathbb{X} = x_1, \ldots, x_n \subset \Re^{\mathfrak{N}}$ και ένα σύνολο από clusters $K$:$\mathbb{C} = C_1, \ldots, C_K$, για κάθε cluster $j$ η ανομοιότητα μέσα σε αυτό δίνεται από τη σχέση 3.26:

$$\sum_{x_i \in C_j} ||x_i - \mu_j||^2 \tag{3.26}$$

όπου:

- $\mu_j = \frac{1}{N} \sum_{x_i \in C_j} x_i$ : η μέση τιμή των παρατηρήσεων στο $C_j$ και $N_j$ ο αριθμός των παρατηρήσεων μέσα στο $C_j$.

Ο αλγόριθμος 4.4 δίνει την βέλτιστη τοπική λύση και παρουσιάζεται παρακάτω.Με δείκτη $m$ συμβολίζεται ο τρέχων αριθμός των επαναλήψεων του αλγορίθμου.

**Algorithm 2** Αλγόριθμος K-means

1. Δεδομένου ενός προκαθορισμένου αριθμού clusters , τυχαία αναθέτουμε ένα αρχικό σύνολο από μέσες τιμές $\mu_1^0, \ldots, \mu_K^0$.

2. Αναθέτουμε σε κάθε παρατήρηση $x_i$ στο cluster του οποίου η μέση τιμή είναι παρόμοια με την παρατήρηση. Αυτό αποτυπώνεται στην εύρεση της κοντινότερης μέσης τιμής του cluster μέσα από τον υπολογισμό της τετραγωνισμένης Ευκλείδειας απόστασης:

$$cluster(x_i) = \arg\min_j ||x_i - \mu_j^m||^2$$

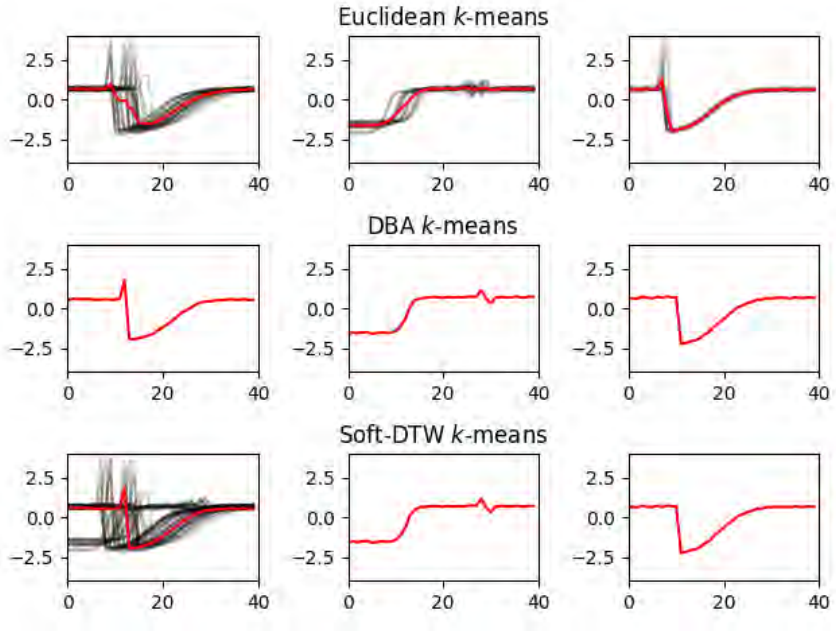
3. Για κάθε cluster $K$ υπολογίζουμε το νέο κεντροειδές, το οποίο είναι η νέα μέση τιμή του cluster:

$$\mu_j^{(m+1)} = \frac{1}{N_j^{(m)}} \sum_{x_i \in C_j^m}$$

4. Επέστρεψε στο 2 και επανέλαβε μέχρι οι αναθέσεις των παρατηρήσεων στα cluster σταματήσουν.

Ο αλγόριθμος πάντα συγκλίνει, ωστόσο εξαιτίας της αυθαίρετης αρχικοποίησης των κεντροειδών των clusters δεν εγγυάται η ομοιογένεια μεταξύ των παρατηρήσεων σε κάθε cluster. Για αυτό ο αλγόριθμος εκτελείται αρκετές φορές ώστε να προκύψει το επιθυμητό αποτέλεσμα.

Στην εικόνα 3.11 χρησιμοποιείται η μέθοδος k-means για clustering χρονοσειρών. Παρουσιάζονται 3 παραλλαγές αυτού του αλγορίθμου: Ευκλείδεια K-means,DBA-k-means,Soft-DTW k-means.




Σχήμα 3.11: K-means clustering για χρονοσειρές.

# Κεφάλαιο 4

# Εξερεύνηση και προ-επεξεργασία των δεδομένων

## 4.1 Εισαγωγή

Για το πειραματικό κομμάτι της εργασίας τα δεδομένα που χρησιμοποιήθηκαν παρέχονται από το project Low Carbon London του UK Power Networks. Οι μετρήσεις περιλαμβάνουν δείγματα από 5.567 έξυπνους μετρητές κατοικιών, το χρονικό διάστημα Νοέμβριος 2011 έως και Φεβρουάριος 2014.

Η δειγματοληψία των ενδείξεων έγινε ανά μισή ώρα. Επίσης παρέχονται κοινωνικοί δείκτες ως προς τους ενοίκους σύμφωνα με το πρότυπο CACI Acorn group (2010). Το δείγμα των κατοίκων είναι αντιπροσωπευτικό του πληθυσμού της ευρύτερης περιοχής του Λονδίνου.

Το dataset περιέχει την τιμής της κατανάλωσης ενέργειας σε kWh (ανά μισή ώρα), ένα μοναδικό αναγνωριστικό για κάθε οικία, ώρα και ημερομηνία, και ένα αναγνωριστικό κατηγορίας CACI Acorn group. Τα δεδομένα διατίθεται σε μορφή CSV μεγέθους περίπου 10 GB. Για τις απαιτήσεις αυτής της εργασίας η δειγματοληψία αναπροσαρμόστηκε ανά ώρα για την μείωση του όγκου των δεδομένων και του χρόνου υπολογισμού των αλγορίθμων.

Μέσα στο dataset οι πελάτες διακρίνονται σε δύο γκρουπ. Το πρώτο περιλαμβάνει περίπου 1100 καταναλωτές οι οποίοι υποβλήθηκαν στο πρόγραμμα Δυναμικής Τιμολόγησης (dToU). Το δεύτερο, περιλαμβάνει τους υπόλοιπους καταναλωτές σταθερής τιμολόγησης. Το κόστος της κιλοβατώρας δινόταν μια μέρα νωρίτερα μέσω του δικτύου των έξυπνων μετρητών ή μέσω μηνύματος στο κινητό. Οι καταναλωτές λάμβαναν 3 σήματα τιμολόγησης: Υψηλή(67,20 pence/kWh), Κανονική (11,76 p/kWh) και Χαμηλή (3,99 p/kWh). Οι υπόλοιποι καταναλωτές είχαν σταθερό τιμολόγιο 14,228 p/kWh.

Στις παρακάτω ενότητες γίνεται εξερεύνηση και οπτικοποίηση του dataset για την εξαγωγή χρήσιμων πληροφοριών που θα βοηθήσουν σε πρώτο επίπεδο στην ανάδειξη χρήσιμων συσχετίσεων μεταξύ φορτίου και εξωτερικών παραγόντων

Στη συνέχεια θα γίνει προ επεξεργασία των δεδομένων στο πνεύμα της μεθοδολογίας που αναφέρθηκε σε προηγούμενη ενότητα, ώστε τα δεδομένα να είναι στην κατάλληλη μορφή για την παραγωγή μοντέλων πρόβλεψης. Τέλος, τα μοντέλα πρόβλεψης θα αναπτυχθούν σε 3 επίπεδα:

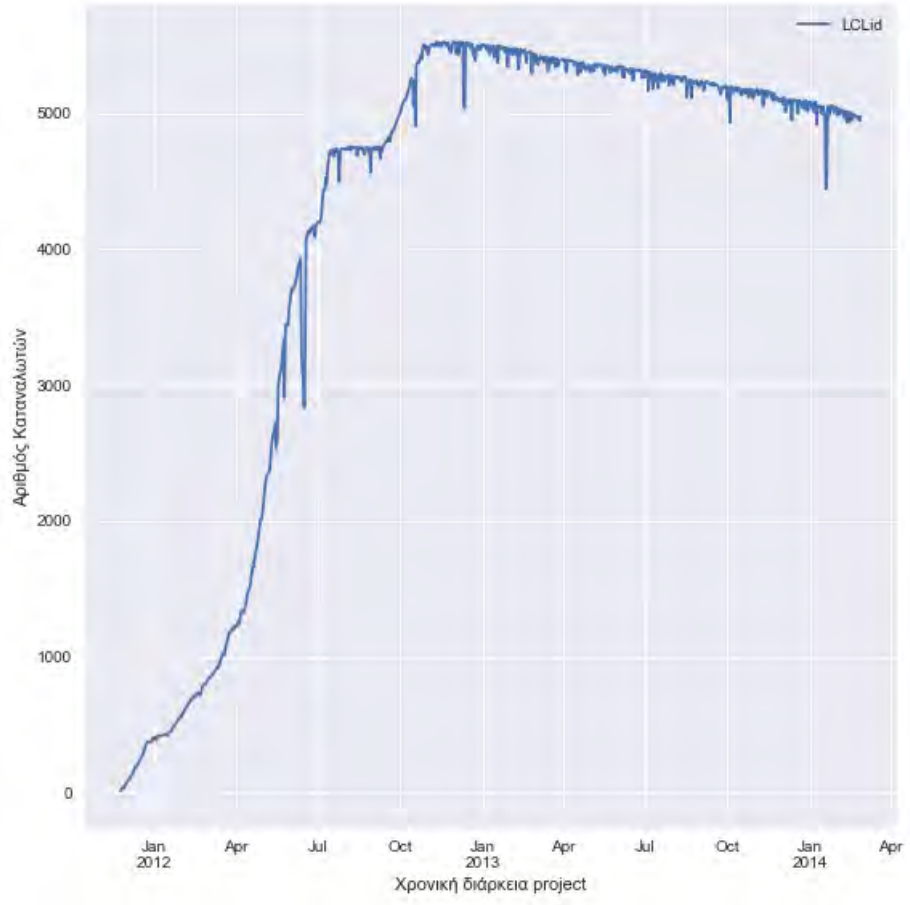
- Συγκεντρωτικό.
- Ανά γκρουπ/cluster
- Ατομικό

Σκοπός είναι να αναδειχθούν και να επιβεβαιωθούν επιδόσεις των μοντέλων πρόβλεψης σύμφωνα με την βιβλιογραφία. Σε αυτό το κεφάλαιο θα γίνει και η ομαδοποίηση των έξυπνων μετρητών σε clusters προφίλ ενέργειας.

## 4.2 Εξερεύνηση δεδομένων

### 4.2.1 Επιλογή πλήθους καταναλωτών

Αρχικά θα πρέπει να εξακριβωθεί πότε είναι η καταλληλότερη περίοδος να για συγκεντρώσουμε όσο το δυνατόν περισσότερα στοιχεία, καθώς ο αριθμός των καταναλωτών που συμμετέχουν στο πρόγραμμα είναι μεταβλητός στο διάστημα που διήρκεσε. Στην εικόνα 4.1 αποτυπώνεται αυτή η μεταβολή του πλήθους ως προς τον χρόνο.




Σχήμα 4.1: Χρήσιμες μέρες

### 4.2.2 Γραφική αναπαράσταση συνολικής καμπύλης φορτίου

Οι παρακάτω γραφικές παραστάσεις δείχνουν την εξέλιξη της κατανάλωσης ενέργειας στο χρονικό διάστημα που μελετάται, αθροίζοντας τις μετρήσεις όλων των μετρητών ανά χρονική στιγμή.

Στην 4.2 απεικονίζεται η ημερήσια κατανάλωση του δικτύου. Παρατηρείται μεγαλύτερη κατανάλωση ενέργειας την χειμερινή περίοδο ενώ το καλοκαίρι είναι η χαμηλότερη.


Σχήμα 4.2: Συνολική ημερήσια κατανάλωση του δικτύου.

Η 4.3 δίνει τη μέση ωριαία κατανάλωση για κάθε μήνα. Για όλους τους μήνες παρατηρείται το ίδιο πρότυπο συμπεριφοράς ωστόσο με διαφορετικά πλάτη.

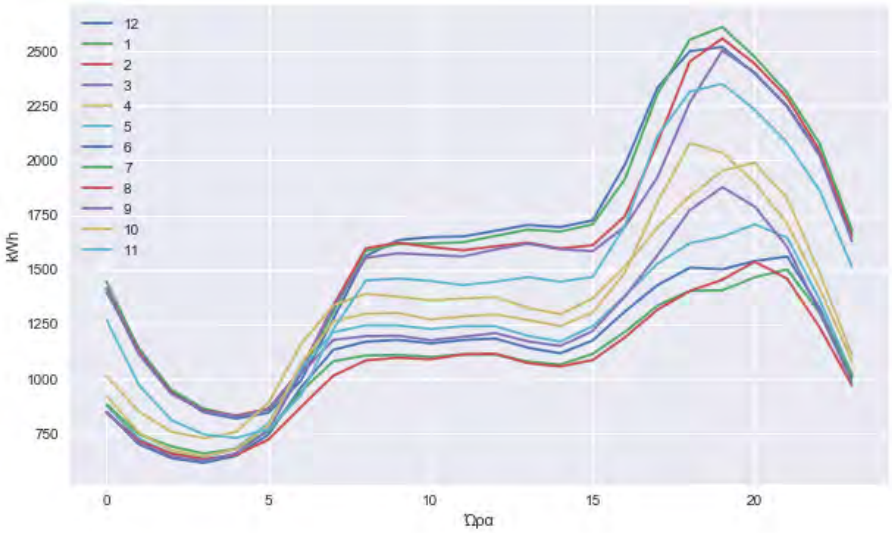
Μία κοινή ρουτίνα κατά την δημιουργία μοντέλων πρόβλεψης είναι προσέγγιση της παρόμοιας ημέρας. Προκύπτει έτσι ένας βασικός διαχωρισμός στο dataset σε καθημερινές ημέρες και Σαββατοκύριακο. Όπως φαίνεται και στις παρακάτω εικόνες 4.7 και4.5 οι καμπύλες φορτίου κατά τις καθημερινές ημέρες παρουσιάζουν παρόμοια συμπεριφορά ενώ τα Σ/Κ η κατανάλωση είναι μικρότερη με διαφορετική εξέλιξη.

### 4.2.3 Μετεωρολογικές παράμετροι

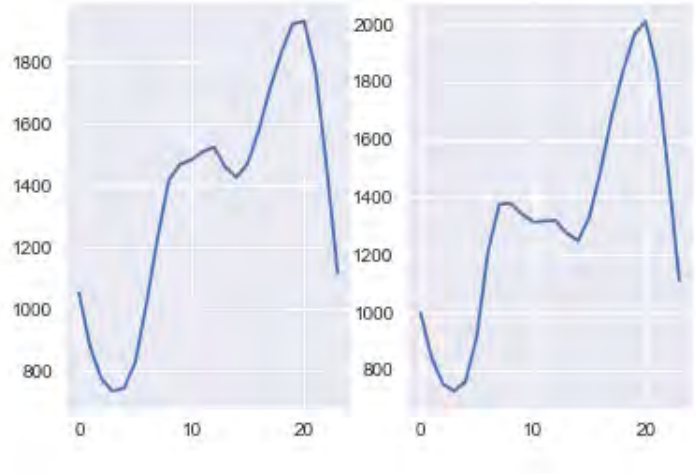
Οι μετεωρολογικές παράμετροι δεν παρέχονται από το επίσημο dataset του Low Carbon London. Ωστόσο, μέσω του Dark Sky API υπάρχει η δυνατότητα συλλογής πληροφοριών από μετεωρολογικούς σταθμούς στην ευρύτερη περιοχή του Λονδίνου και για το χρονικό διάστημα που μελετάμε.

Οι πιο σημαντικές παράμετροι που περιέχονται είναι:

- Θερμοκρασία




Σχήμα 4.3: Μέση ωριαία κατανάλωση ανα μήνα

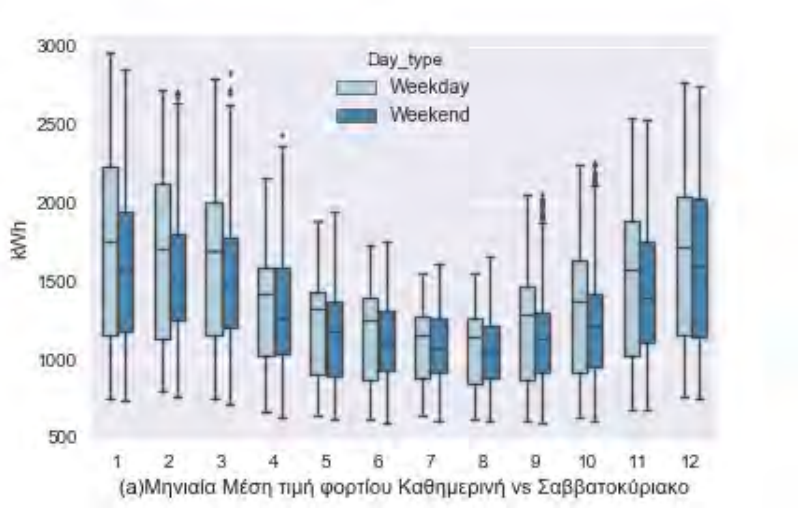



Σχήμα 4.4: Καθημερινή vs Σ/Κ

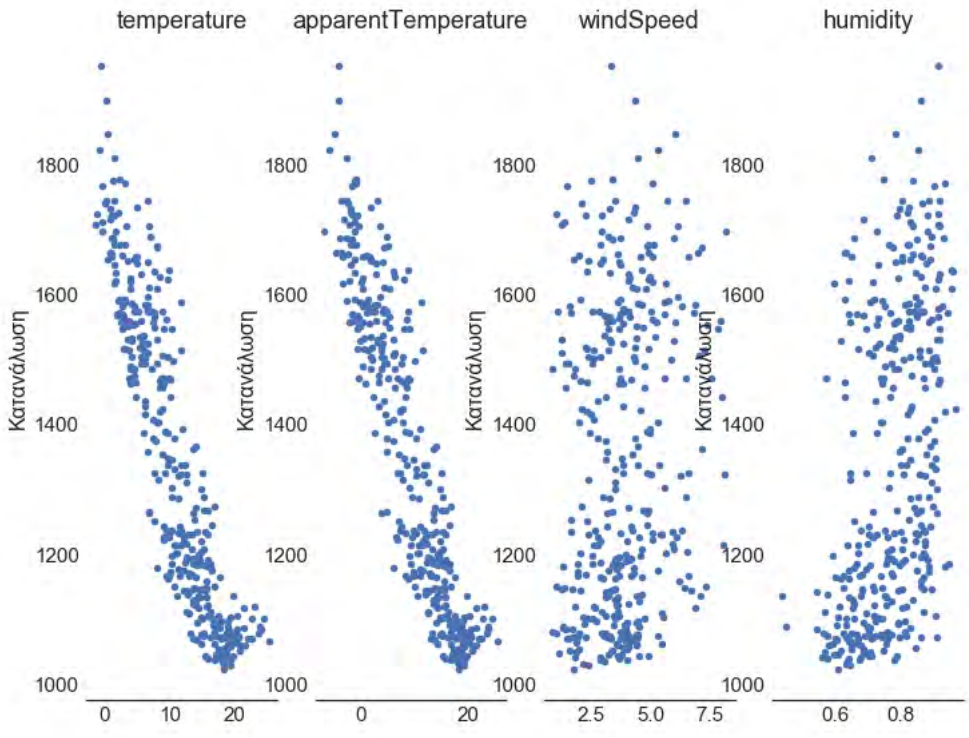
- Υγρασία
- Ταχύτητα ανέμου

Στη συνέχεια ελέγχουμε για συσχετισμό μεταξύ φορτίου και καιρικών παραμέτρων όπως φαίνεται και στο σχήμα 4.6 .

Παρατηρούμε πως υπάρχει γραμμική εξάρτηση μεταξύ φορτίου και θερμοκρασίας ενώ για τις υπόλοιπες μεταβλητές δεν είναι εμφανές κάτι τέτοιο. Οι μετρήσεις συσχέτισης δεν είναι σταθερές αλλά εξαρτώνται από την εποχή που γίνεται η μελέτη.



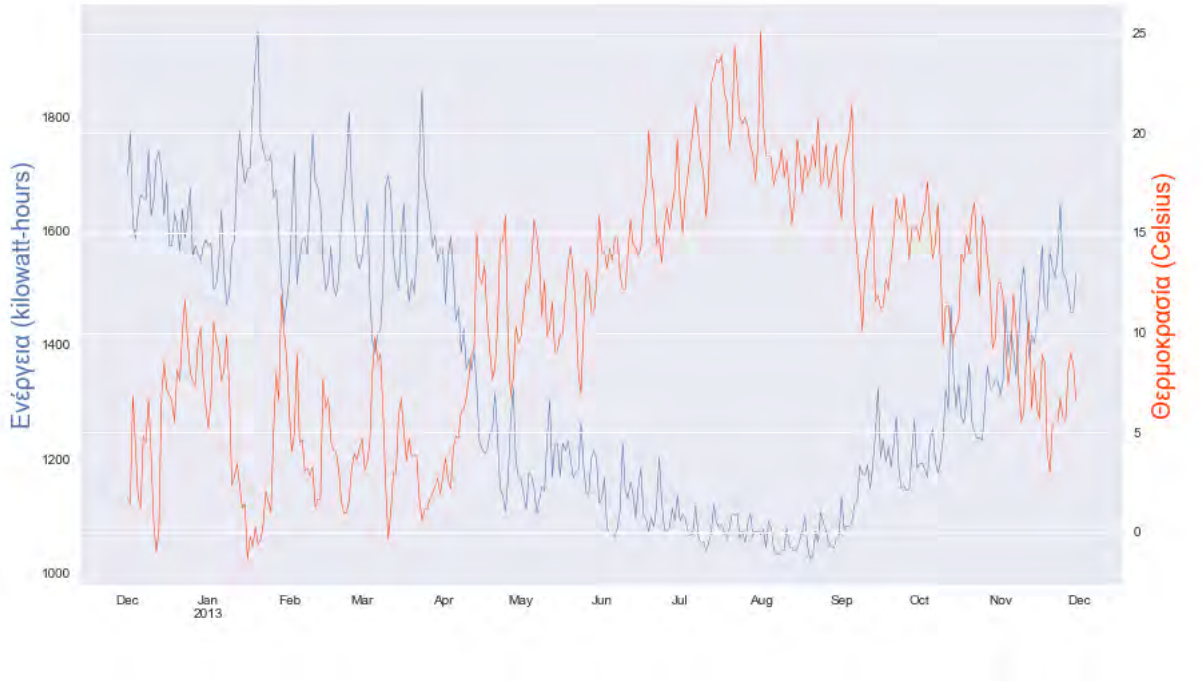

Σχήμα 4.5




Σχήμα 4.6: Scatter plot συσχετισμού καιρικών παραμέτρων με το φορτίο.

Στην εικόνα 4.7 φαίνεται η εξέλιξη της τιμής του φορτίου σε σχέση με την θερμοκρασία. Παρατηρούμε ότι κατά τις ψυχρές περιόδους είναι μεγαλύτερη η κατανάλωση. Αυτό υποδηλώνει πως στο Λονδίνο οι κάτοικοι χρησιμοποιούν ηλεκτρική ενέργεια για σκοπούς θέρμανσης. Παράλληλα τους καλοκαιρινούς μήνες επειδή δεν εκδηλώνονται ιδιαίτερα υψηλές θερμοκρασίες σε συνδυασμό με την πιθανότητα απουσίας των καταναλωτών από τις οικίες τους για παραθερισμό,

η κατανάλωση περιορίζεται αισθητά.




Σχήμα 4.7: Συγκριτικό διάγραμμα διακύμανσης των τιμών του φορτίου και της θερμοκρασίας.

## 4.3 Προεπεξεργασία δεδομένων

### 4.3.1 Χρονολογικές μεταβλητές

Η πρόβλεψη κατανάλωσης ενέργειας σε όλα τα επίπεδα είναι εξαρτημένη από τον χρόνο. Κάθε μέτρηση έχει μία χρονοσφραγίδα από την οποία εξάγουμε πληροφορίες για την ώρα και την ημερομηνία της μέτρησης που θα αποτελέσουν και αυτές υποψήφιες παράμετροι για την εκπαίδευση του μοντέλου.

Οι χρονολογικές μεταβλητές είναι εκ φύσεως κυκλικές και η τυπική αναπαράσταση του χρόνου δεν μπορεί να αποτυπώσει την πραγματική απόσταση μεταξύ δύο χρονικών στιγμών. Για παράδειγμα, χρησιμοποιώντας τις ώρες σε κλίμακα από 1-24 σημαίνει πως η ώρα 22 αναπαρίσταται ως κοντινότερη στην ώρα 10 (12 ώρες διαφορά) παρά στην ώρα 0 (22 ώρες διαφορά). Ωστόσο, η συμπεριφορά στην ώρα 0 είναι πιο πιθανό να είναι παρόμοια με την ώρα 22 καθώς η πραγματική διαφορά είναι δύο ώρες.

Για να αποτυπώσουμε σωστά τις κυκλικά συνεχείς μεταβλητές, πρέπει να μετατρέψουμε αυτές τις παραμέτρους σε παραστάσεις ημίτονου και συνημιτόνου. Για παράδειγμα, η αναπαράσταση της ώρας σε αυτή τη μορφή γίνεται με την εξίσωση:

$$Hour(h) = \left\{ \sin(\frac{2 * \pi * h}{24}), \cos(\frac{2 * \pi * h}{24} \right\}$$

### 4.3.2 Παράμετροι εισόδου

Ορίζουμε ένα γενικό σετ παραμέτρων εισόδου στο μοντέλο μας, οι οποίες θεωρούμε ότι σχετίζονται με το φορτίο και βοηθούν στην ακριβή πρόβλεψή του. Το σετ αυτό συνηθίζεται να είναι τεράστιο με εκατοντάδες παραμέτρους που περιλαμβάνουν ιστορικές τιμές του φορτίου και των καιρικών δεδομένων έως και 2 εβδομάδες πρίν. Αυτό βέβαια απαιτεί μεγάλη υπολογιστική ισχύ ώστε να καταλήξουμε στις πιο σημαντικές. Για το σκοπό αυτής της εργασίας οι παράμετροι που θα εισαχθούν θα είναι εμπειρικά επιλεγμένες βάσει της σχετικής βιβλιογραφίας και άλλων εργασιών σε αυτό το πρόβλημα.

Για να προβλέψουμε το φορτίο την ώρα $t$ για πρόβλεψη μίας ώρας μπροστά (1 hour ahead forecast), συμπεριλαμβάνουμε ιστορικές τιμές του φορτίου από την χρονική στιγμή $t-1$ έως $t-168$ με κάποια βήματα σε αυτό το διάστημα. Επίσης λαμβάνουμε υπόψη τις καιρικές μεταβλητές για την ώρα $t$ που μελετάμε αν θεωρήσουμε ότι εκ των προτέρων έχουμε τις καιρικές προγνώσεις. Συνοπτικά στον πίνακα 4.1 παρουσιάζονται οι υποψήφιες παράμετροι για τα μοντέλα.

| Είσοδοι | Περιγραφή |
|---|---|
| 1-12 | Load(t-h), h=1-12, step=1 |
| 13-19 | Load(t-h), h=24-168, step=24 |
| 20 | Temperature(t) |
| 21 | Windspeed(t) |
| 22 | Humidity(t) |
| 23 | DewPoint(t) |
| 24 | IsWeekday |
| 25-30 | $\sin(\frac{2*\pi*(\cdot)}{24})$ |

Πίνακας 4.1: Παράμετροι εισόδου για 1-h ahead forecast.

Η παράμετρος IsWeekday είναι μία δυαδική παράμετρος που περιγράφει αν η μέρα είναι καθημερινή ή Σαββατοκύριακο/Εορτή. Αν θέλουμε να μελετήσουμε την περίπτωση της πρόβλεψης 24 ωρών μπροστά, οι παρατηρήσεις μας θα ξεκινήσουν από τη στιγμή $t-24$ και πιο πίσω καθώς οι τιμές $t-1$ έως $t-23$ δεν είναι διαθέσιμες. Τροποποιούμε έτσι αντίστοιχα τις παραμέτρους μας. Οι μεταβλητές 25-30 περιγράφουν τις κυκλικές ημερολογιακές μεταβλητές hour,day,month.

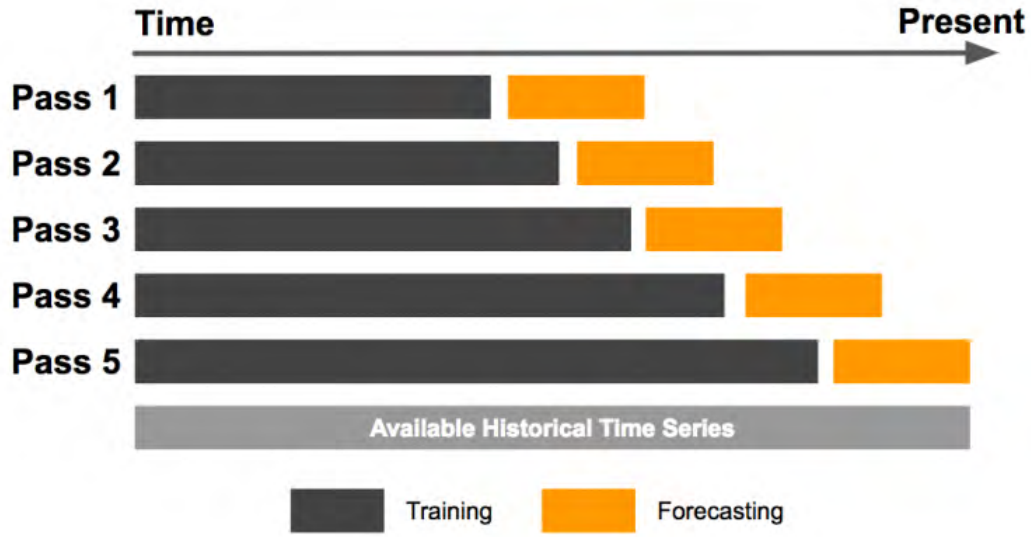
### 4.3.3 Παραγωγή δεδομένων εκπαίδευσης και αξιολόγησης

Σκοπός της εργασίας είναι να γίνει η πρόβλεψη φορτίου για 1 και 24 ώρες μπροστά. Μπορούμε να διαχωρίσουμε το training set από το test set με βάση τον ορίζοντα πρόβλεψης.

αρχή του test set =μήκος dataset - ημέρες προς προβλεψη * ημερήσιες μετρήσεις

Τα δεδομένα,δηλαδή, που προηγούνται της αρχής του test set μπορούν να χρησιμοποιηθούν ως δεδομένα εκπαίδευσης.

Η τεχνική αυτή ανταποκρίνεται στην λογική της μεθόδου **expanding window forecast** όπως φαίνεται και στο σχήμα 4.8.




Σχήμα 4.8: Διαδικασία πρόβλεψης Expanding window

### 4.3.4 Κανονικοποίηση παραμέτρων (Feature normalization)

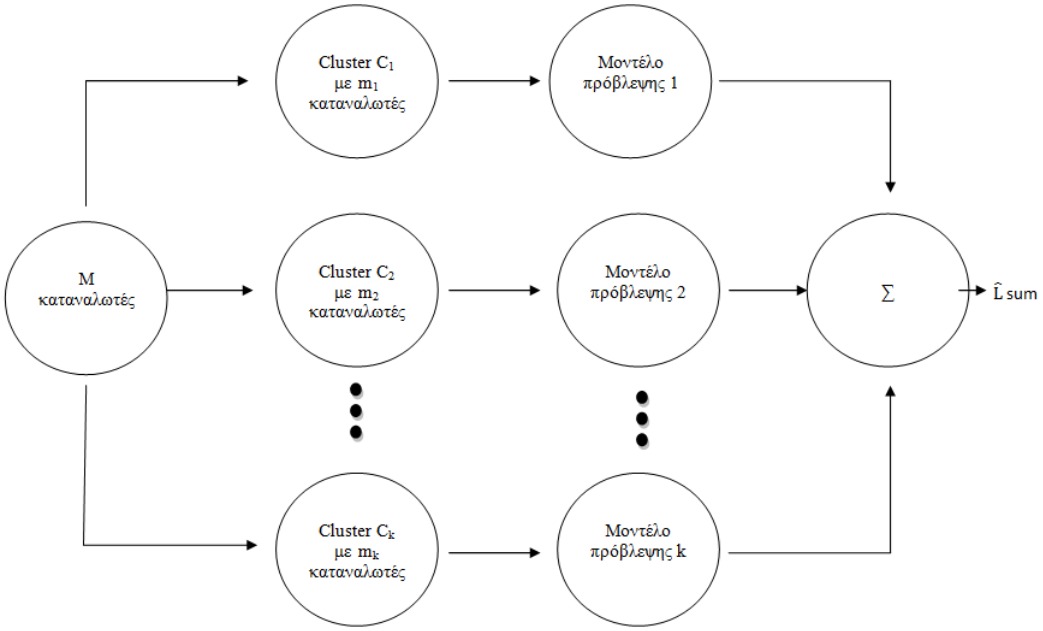
Η κανονικοποίηση παραμέτρων είναι η διαδικασία αλλαγής του εύρους τιμής μιας παραμέτρου. Αυτό χρειάζεται γιατί κάθε παράμετρος έχει διαφορετικές μονάδες μέτρησης και διαφορετικά εύρη. Για αλγόριθμους που βασίζονται σε υπολογισμούς αποστάσεων για να κάνουν προβλέψεις (SVM, KNN) οι διαφορετικές παράμετροι έχουν μεγάλη διαφορά στα εύρη τιμών και εμποδίζουν την μάθηση του μοντέλου.

## 4.4 Ομαδοποίηση (Clustering) προφίλ ενέργειας

Σκοπός της ομαδοποίησης καταναλωτών με βάση το πρότυπο κατανάλωσης ενέργειας που παρουσιάζουν, είναι να εξετάσει μία διαφορετική προσέγγιση στο πρόβλημα της πρόβλεψης κατανάλωσης ενέργειας σε συγκεντρωτικό επίπεδο. Η κλασσική μέθοδος αφορά την συνάθροιση όλων των μετρήσεων σε ατομικό επίπεδο και στη συνέχεια την εφαρμογή των τεχνικών STLF σε επίπεδο μικροδικτύου. Επίσης μία ακραία περίπτωση είναι να πάρουμε την πρόβλεψη για κάθε οικία ξεχωριστά και μετά να τις αθροίσουμε για να πάρουμε την συνολική πρόβλεψη.

Η πιο ενδιαφέρουσα προσέγγιση αφορά την ομαδοποίηση σπιτιών με παρόμοια χαρακτηριστικά και στη συνέχεια την ανάπτυξη ενός μοντέλου πρόβλεψης για κάθε cluster ξεχωριστά (Cluster based Aggregating Forecasting). Το πλήθος K των clusters μπορεί να είναι από K=1 (aggregated) έως και το πλήθος K=n των μετρητών (individual).

Οι παράμετροι που προκύπτουν για κάθε μετρητή προκύπτουν από την εξαγωγή ενός μέσου όρου ωριαίας κατανάλωσης για ένα 24ωρο. Δηλαδή κάθε μετρητής περιγράφεται από 24 παραμέτρους. Η ομαδοποίηση θα γίνει με τη μέθοδο K-means Clustering και θα πειραματιστούμε ως προς τον αριθμό των clusters βάσει μίας πρότερης διαισθητικής διαλογής ανάμεσα σε καταναλωτές δυναμικής και κανονικής τιμολόγησης και στη συνέχεια μέσω οπτικής αναπαράστασης θα




Σχήμα 4.9: Ροή εργασίας για κάθε cluster και παραγωγή συγκεντρωτικής πρόβλεψης.

συμπεράνουμε των αριθμό τον cluster τα οποία ως επί το πλείστων θα είναι ομογενή.

**Algorithm 3** Cluster Based Aggregated Forecasting

1. Διαχώρισε το training set σε μετρητές Σταθερής Τιμολόγησης και Δυναμικής Τιμολόγησης. ένα μέσο ωριαίο ημερήσιο προφίλ ενέργειας για κάθε μετρητή. Κάθε μετρητής δηλαδή περιγράφεται από 24 παραμέτρους.

2. Κανονικοποίησε τα δεδομένα να βρίσκονται στην ίδια κλίμακα για την καλύτερη ομογένεια των clusters

3. Επιλέγω ένα διάστημα πλήθους clusters $K$, ώστε να συγκρίνω πως επηρεάζει το πλήθος των Cluster την συγκεντρωτική πρόβλεψη. Επιλέγω το K με βάση την μέθοδο Elbow.

   Για κάθε $K$:

   (α΄) Εκτέλεσε για κάθε Cluster $C_i$ , $i$ =αρχική τιμή,. . . , $K$ τον επιλεγμένο αλγόριθμο πρόβλεψης

   (β΄) Άθροισε τις επιμέρους προβλέψεις και σύγκρινε τα αποτελέσματα ως προς τις πραγματικές τιμές.

   (γ΄) Επανέλαβε για νέο K

## 4.5 Εξέταση διαφορετικών σεναρίων πρόβλεψης φορτίου

Κατά την πειραματική αξιολόγηση αυτής της εργασίας θα εξεταστούν διαφορετικά σενάρια πρόβλεψης με σκοπό να αξιολογηθούν οι τεχνικές πρόβλεψης ως προς την συμπεριφορά της κα-

τανάλωσης ενέργειας. Για κάθε περίπτωση εφαρμόζονται οι τεχνικές προ επεξεργασίας των δεδομένων για την ανακάλυψη συσχετισμών και την αξιολόγηση χρήσιμων παραμέτρων. Τα πειράματα που θα διεξαχθούν είναι:

- **Πείραμα 1:** Αξιολόγηση σε συγκεντρωτικό επίπεδο των τεχνικών μηχανικής μάθησης ως προς την ακρίβεια της πρόβλεψης.
- **Πείραμα 2:** Εφαρμογή συνδυαστικής μεθόδου ομαδοποίησης προφίλ κατανάλωσης ενέργειας, συνάθροισης επιμέρους προβλέψεων και σύγκριση με το συγκεντρωτικό μοντέλο πρόβλεψης.
- **Πείραμα 3:** Εφαρμογή τεχνικών πρόβλεψης σε επίπεδο μετρητή.

Οι τεχνικές μηχανικής μάθησης που θα εφαρμοστούν είναι:

1. Linear Regression
2. Random Forest
3. ExtraTrees
4. Support Vector Regression
5. GradientBoosting
6. Neural Networks

Οι αλγόριθμοι αναπτύχθηκαν σε γλώσσα προγραμματισμού **Python** καθώς προσφέρεται για αποδοτική διαχείριση και επεξεργασία δεδομένων μέσω της βιβλιοθήκης **Pandas**. Το πιο σημαντικό κριτήριο ωστόσο είναι η εξαιρετικά πλούσια βιβλιοθήκη **Scikit-Learn** που παρέχει για αυτή την εργασία τις απαραίτητες συναρτήσεις.

# Κεφάλαιο 5

# Πειραματική αξιολόγηση

Σε αυτό το κεφάλαιο της εργασίας θα αναπτύξουμε τα σενάρια πρόβλεψης που αναφέρθηκαν και θα παρατεθούν οι μετρήσεις και τα αποτελέσματα για κάθε ένα από αυτά.

## 5.1 Πείραμα 1: Συγκεντρωτική πρόβλεψη

Έχοντας καθαρίσει τα δεδομένα ενός χρόνου από τυχόν ελλιπείς ή ακραίες μετρήσεις καταλήξαμε σε ένα σύνολο 4082 μετρητών των οποίων οι μετρήσεις αθροίστηκαν ώστε να γίνει η πρόβλεψη σε ένα υψηλότερο επίπεδο, όπου το πρότυπο της κατανάλωσης ενέργειας είναι πιο ομαλό και περιμένουμε να εξάγουμε αποτελέσματα κοντά στις πραγματικές τιμές.

Πέρα από την αξιολόγηση των επιδόσεων των τεχνικών μάθησης, θα εξεταστεί και η επιρροή των καιρικών παραμέτρων ως προς το παραγόμενο αποτέλεσμα και αν τελικά έχει σημασία να συμπεριληφθούν στο μοντέλο.

Οι προβλέψεις θα γίνουν για διαφορετικό πλήθος ημερών εκπαίδευσης το οποίο θα μεγαλώνει σε βάθος χρόνου. Οι προβλέψεις θα είναι για μία ώρα μπροστά σε βάθος 24ώρου (1h ahead next day forecast). Αρχικά χρησιμοποιούμε τις default τιμές των υπέρ-παραμέτρων για κάθε αλγόριθμο, που προσφέρονται από την βιβλιοθήκη Scikit-Learn. Έπειτα εξετάζουμε αν μπορεί να βελτιωθεί η επίδοσή τους ρυθμίζοντας κατάλληλα αυτές τις υπερ-παραμέτρους. Μπορούμε να θεωρήσουμε ως benchmark μέθοδο την linear regression καθώς αποτελεί το πιο απλό μοντέλο πρόβλεψης.

### 5.1.1 1h ahead πρόβλεψη, με default παραμέτρους

| | Χρόνος Εκτέλεσης | Training Scores | Testing Scores | MAPE |
|---|---|---|---|---|
| **LinearRegression** | 0.007 | 0.984 | 0.982 | 3.954 |
| **RandomForest** | 0.328 | 0.997 | 0.984 | 3.586 |
| **GradientBoosting** | 0.517 | 0.993 | 0.986 | 3.289 |
| **MLP** | 1.967 | 0.695 | 0.688 | 19.593 |
| **ExtraTrees** | 0.170 | 1.000 | 0.988 | 2.980 |

Πίνακας 5.1: 1h ahead προβλεψη για training data 90 ημερών

| | Χρόνος Εκτέλεσης | Training Scores | Testing Scores | MAPE |
|---|---|---|---|---|
| **LinearRegression** | 0.008 | 0.984 | 0.974 | 3.512 |
| **RandomForest** | 0.726 | 0.997 | 0.984 | 2.686 |
| **GradientBoosting** | 0.984 | 0.990 | 0.984 | 2.647 |
| **MLP** | 3.985 | 0.926 | 0.922 | 6.350 |
| **ExtraTrees** | 0.339 | 1.000 | 0.992 | 1.869 |

Πίνακας 5.2: 1h ahead προβλεψη για training data 180 ημερών

| | Χρόνος Εκτέλεσης | Training Scores | Testing Scores | MAPE |
|---|---|---|---|---|
| **LinearRegression** | 0.014 | 0.986 | 0.987 | 2.371 |
| **RandomForest** | 1.162 | 0.998 | 0.972 | 2.982 |
| **GradientBoosting** | 1.457 | 0.991 | 0.970 | 3.076 |
| **MLP** | 6.732 | 0.950 | 0.929 | 5.927 |
| **ExtraTrees** | 0.522 | 1.000 | 0.970 | 3.340 |

Πίνακας 5.3: 1h ahead προβλεψη για training data 260 ημερών

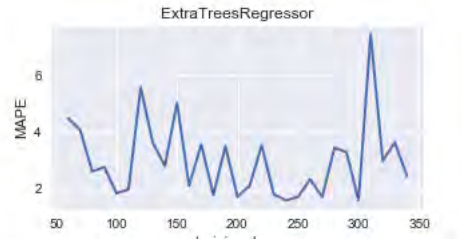
| | Χρόνος Εκτέλεσης | Training Scores | Testing Scores | MAPE |
|---|---|---|---|---|
| **LinearRegression** | 0.016 | 0.985 | 0.988 | 3.199 |
| **RandomForest** | 1.660 | 0.998 | 0.995 | 1.8192 |
| **GradientBoosting** | 2.286 | 0.989 | 0.992 | 2.164 |
| **MLP** | 9.167 | 0.961 | 0.976 | 4.316 |
| **ExtraTrees** | 0.786 | 1.000 | 0.994 | 1.762 |

Πίνακας 5.4: 1h ahead προβλεψη για training data 350 ημερών

Από τα παραγόμενα δεδομένα δεν είναι ξεκάθαρο ποιος αλγόριθμος υπερτερεί. Μπορούμε να ξεχωρίσουμε ότι αλγόριθμος του νευρωνικού δικτύου MLP υστερεί σε σχέση με τους υπόλοιπους, ωστόσο όσο αυξάνεται το δείγμα εκπαίδευσης το ποσοστό σφάλματος μειώνεται. Επίσης πολύ καλά συμπεριφέρεται και το απλούστερο μοντέλο της γραμμικής παλινδρόμησης που σε κάποιες περιπτώσεις αποδίδει και καλύτερα από τις θεωρητικά πιο προχωρημένες μεθόδους. Επίσης, πρέπει να αναφερθεί πως αυτές οι μέθοδοι δεν έχουν παραμετροποιηθεί κατάλληλα οπότε δεν έχουμε τη μέγιστη απόδοσή τους. Στις παρακάτω γραφικές παραστάσεις απεικονίζεται η εξέλιξη του ποσοστού σφάλματος MAPE ανάλογα με την αύξηση του δείγματος εκπαίδευσης.

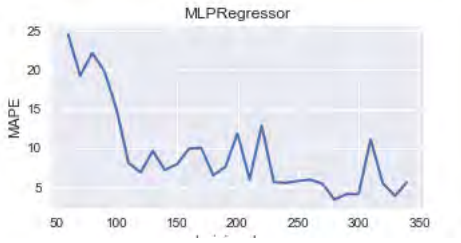
Το μέσο σφάλμα για κάθε μέθοδο ως προς τα διαφορετικά μεγέθη των training set συνοψίζεται στον πίνακα 5.5:

Βλέπουμε, λοιπόν, ότι κατά μέσο όρο η μέθοδος που προσεγγίζει πιο καλά τις πραγματικές τιμές είναι η ExtraTrees.

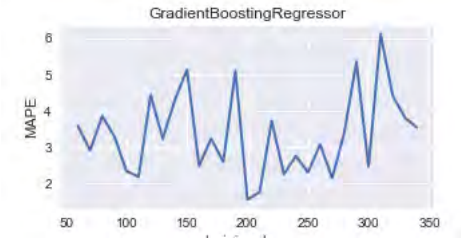



(α΄) ExtraTreesRegressor

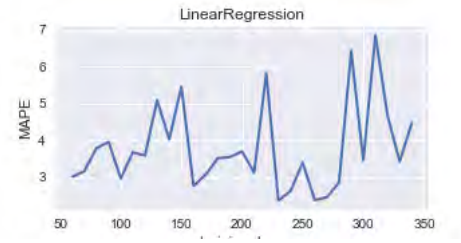

(β΄) RandomForestRegressor




(α΄) MLPRegressor




(β΄) GradientBoostingRegressor




(α΄) LinearRegression

| | MAPE |
|---|---|
| ExtraTrees | 2.98679384267228 |
| GradientBoosting | 3.3637438961074055 |
| LinearRegression | 3.7772806711174236 |
| MLP | 9.574905553340011 |
| RandomForest | 3.180971811305586 |

Πίνακας 5.5: Μέσος όρος σφαλμάτων στο σύνολο.

### 5.1.2 Ακρίβεια πρόβλεψης αλγορίθμων αποκλειστικά με ιστορικές τιμές του φορτίου

Κατά τον ίδιο τρόπο θα αξιολογήσουμε και το μοντέλο προβλέψεων χωρίς τις καιρικές παραμέτρους ώστε να εξετάσουμε αν και πόσο επηρεάζουν το αποτέλεσμα.

Στην γενικότερη περίπτωση εφαρμόζουμε τους αλγορίθμους για διαφορετικά training sets ώστε να εξάγουμε πιο ασφαλή συμπεράσματα ως προς την απόδοση της μεθόδου πρόβλεψης.

Στον πίνακα 5.6 παρατηρούμε πως για τις μεθόδους εκτός των RandomForest όλες οι υπόλοιπες βελτιώνουν τις προβλέψεις συμπεριλαμβάνοντας τις καιρικές παραμέτρους.

| | MAPE μαζί πλήρες σετ παραμέτρων | MAPE μόνο ιστορικές τιμές φορτίου |
|---|---|---|
| ExtraTrees | 2.98679384267228 | 3.1496399732774494 |
| GradientBoosting | 3.3637438961074055 | 3.4763789450330944 |
| LinearRegression | 3.7772806711174236 | 3.9496759089632016 |
| MLP | 9.574905553340011 | 11.191192580316232 |
| RandomForest | 3.180971811305586 | 3.1280147875467055 |

Πίνακας 5.6: Σύγκριση μεθόδων (1)

### 5.1.3 1h ahead με ρυθμισμένες υπερπαραμέτρους.

Για να βελτιώσουμε την επίδοση αλγορίθμων θα πρέπει να βρούμε τις καταλληλότερες υπερπαραμέτρους. Αυτό επιτυγχάνεται μέσα από την μέθοδο **GridCV** της Scikit-Learn.

1. Για την μέθοδο **SVR** εξετάζουμε τις παραμέτρους:

   Οι υποψήφιες παράμετροι είναι:

   - Kernel=[rbf, linear]
   - $\gamma$=[1e-4, 1e-3, 0.01, 0.1, 0.2, 0.5, 0.6, 0.9]
   - $C$=[1, 10, 100, 1000]

   Οι παράμετροι που δοκιμάζουμε είναι κυρίως αυθαίρετες και έχουν επιλεχθεί βάσει παρόμοιων προβλημάτων.

   Οι καλύτερες παράμετροι που εξάγονται από την GridCV είναι:

   - Kernel=rbf
   - $\gamma$=0.2
   - $C$=100

2. Για την μέθοδο **GradientBoost** εξετάζουμε τις παραμέτρους:

   - Number of estimators:[100,1000]
   - learning rate =[0.1, 0.05, 0.02, 0.01]
   - max depth:[4,6]
   - min samples leaf:[3,5,9,17]

   Οι καλύτερες παράμετροι που εξάγονται από την GridCV είναι:

   - Number of estimators:1000
   - learning rate =0.05
   - max depth:4
   - min samples leaf:17

3. Για την μέθοδο **RandomForest** εξετάζουμε τις παραμέτρους:

   - Number of estimators": [10, 50, 75, 100, 150],
   - max features: [auto, sqrt, log2],
   - max depth: [50, 100,150, 200, 250]

   Οι καλύτερες παράμετροι που εξάγονται από την GridCV είναι:

   - Number of estimators: 75,
   - max features:auto,
   - max depth: 250

4. Για την μέθοδο **ExtraTrees** εξετάζουμε τις παραμέτρους:

   - Number of estimators: [10, 50, 75, 100, 150],
   - max features: [auto, sqrt, log2],
   - max depth: [50, 100,150, 200, 250]

   Οι καλύτερες παράμετροι που εξάγονται από την GridCV είναι:

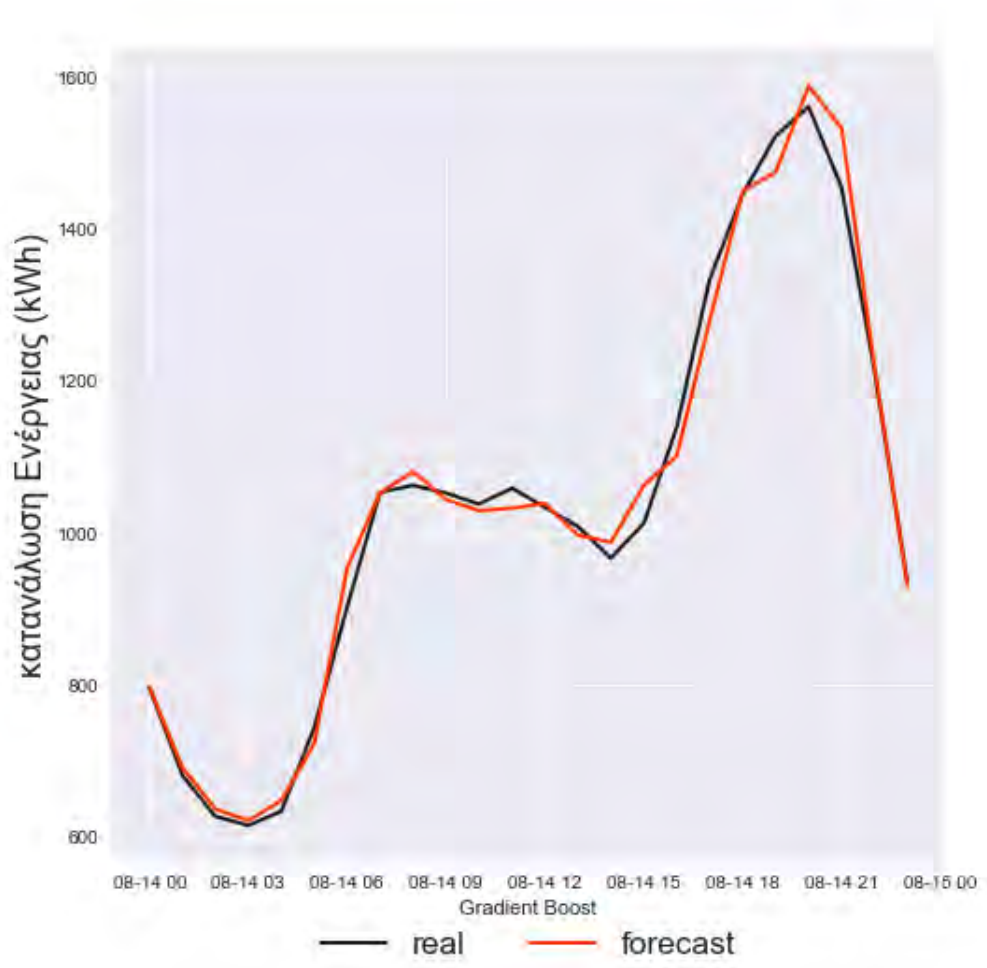
   - Number of estimators: 100,
   - max features:auto,
   - max depth: 250

Για αυτές τις παραμέτρους επαναλαμβάνουμε την διαδικασία για να εξετάσουμε αν θα υπάρξει βελτίωση στην ακρίβεια πρόβλεψης.

| | MAPE (default παράμετροι) | MAPE (βέλτιστες παράμετροι) |
|---|---|---|
| ExtraTrees | 2.98679384267228 | 2.6617931034482756 |
| GradientBoosting | 3.3637438961074055 | 2.5468620689655173 |
| SVR | | 2.9098620689655177 |
| RandomForest | 3.180971811305586 | 2.9394482758620692 |

Πίνακας 5.7: Σύγκριση μεθόδων (2)

Παρατηρούμε ότι απόδοση όλων των αλγορίθμων έχει αυξηθεί. Ιδιαίτερα, η μέθοδος GradientBoost έχει βελτιστοποιηθεί κατά μία περίπου ποσοστιαία μονάδα. Η μέθοδος SVR δεν συμπεριλήφθηκε στα προηγούμενα παραδείγματα γιατί χρειαζόταν εξαρχής ρύθμιση των default παραμέτρων της.




Σχήμα 5.4: Πρόβλεψη 1h ahead για τις επόμενες 24 ώρες με δείγμα εκπαίδευσης 250 μέρες.

## 5.2 Πείραμα 2

Η δεύτερη περίπτωση που θα εξετάσουμε είναι η αξιοποίηση ατομικών μετρητών μέσα από την ομαδοποίησή τους σε clusters παρόμοιων προφίλ ενέργειας. Αντί λοιπόν, να αθροίζουμε όλους τους μετρητές για να παράγουμε ένα συγκεντρωτικό μοντέλο, υπολογίζουμε την κατανάλωση ενέργειας για κάθε cluster και έπειτα αθροίζονται οι επιμέρους προβλέψεις για την παραγωγή του συγκεντρωτικού αποτελέσματος. Στη συνέχεια, μπορούμε να συγκρίνουμε τις δύο αυτές τεχνικές για να αποφανθούμε για το ποια είναι καλύτερη.

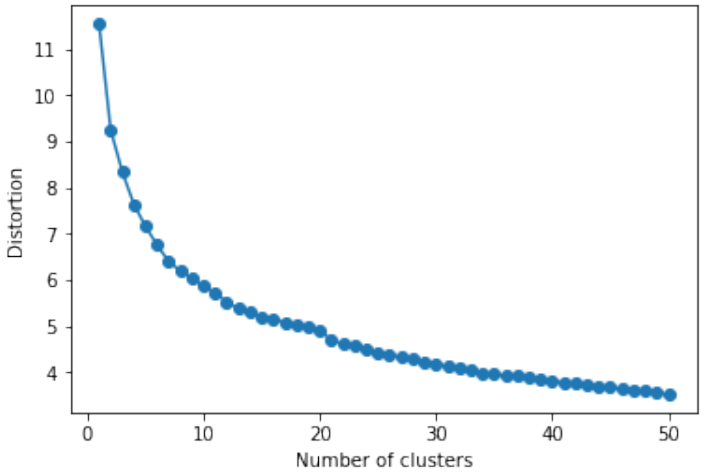
Για όλες τις μεθόδους ισχύει το ίδιο training set 60 ημερών.

### 5.2.1 Clustering Based Aggregated Forecast με την SVR μέθοδο

Ξεκινώντας με την μέθοδο SVR έχουμε:

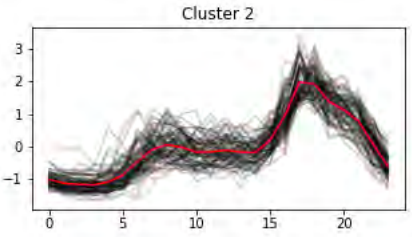
- Καταναλωτές δυναμικής τιμολόγησης πλήθους 796.

Στη εικόνα φαίνεται η μέθοδος του αγκώνα για την επιλογή βέλτιστου αριθμού των clusters. Με αυτό το κριτήριο θα τρέξουμε επαναληπτικά για K=1 μέχρι το K που ξεχωρίζει σε κάθε γράφημα.

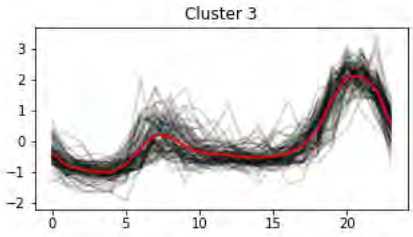



Σχήμα 5.5: Μέθοδος αγκώνα για clusters καταναλωτών δυναμικής τιμολόγησης.

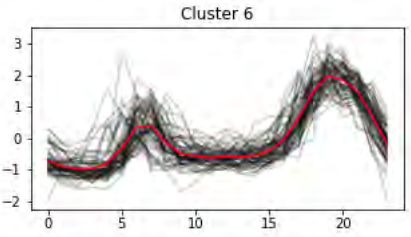
Επιλέγουμε για K=15 και προχωράμε την ομαδοποίηση και την πρόβλεψη των επιμέρους αθροιστικών φορτίων. Τα παραγόμενα clusters απεικονίζονται στις παρακάτω εικόνες. Σημειώνεται πως οι τιμές των μετρητών είναι κανονικοποιημένες για να γίνει καλύτερα αντιληπτή η συμπεριφορά της κατανάλωσης.



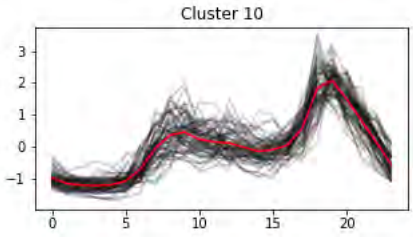

(α΄) cluster2




(β΄) cluster3

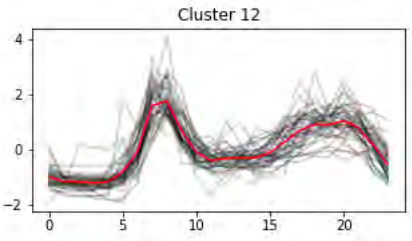



(α΄) cluster6

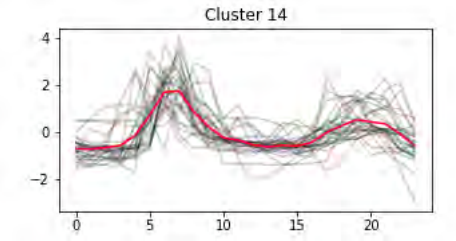



(β΄) cluster10

Τα παραγόμενα σφάλματα που αντιστοιχούν στα διαφορετικά πλήθη K=[1...15] cluster φαίνονται στο γράφημα 5.9. Θεωρούμε ότι για K=1 περιλαμβάνεται ολόκληρο το δείγμα και η πρόβλεψη για αυτή την περίπτωση ανταποκρίνεται στην λογική του πρώτου πειράμα-

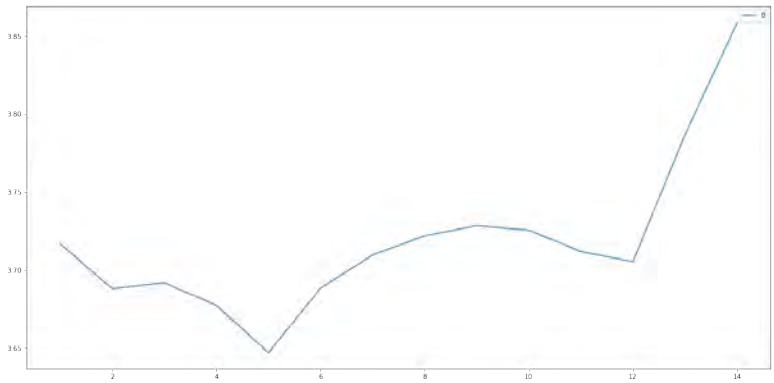



(α΄) cluster12




(β΄) cluster14

τος. Παρατηρούμε ότι η μέθοδος clustering δρα ευεργετικά στην ακρίβεια της πρόβλεψης σε συγκεντρωτικό επίπεδο. Για K=5, παρουσιάζει το μικρότερο σφάλμα, ενώ στη συνέχεια ακολουθεί μία ανοδική πορεία που ξεπερνά το σφάλμα για K=1.



Σχήμα 5.9: SVR MAPE για K clusters καταναλωτών δυναμικής τιμολόγησης

| | K= | MAPE |
|---|---|---|
| | 1 | 3.688056 |
| min | 5 | 3.647123 |
| max | 15 | 3.858682 |

Πίνακας 5.8: MAPE τιμές για διάφορα K

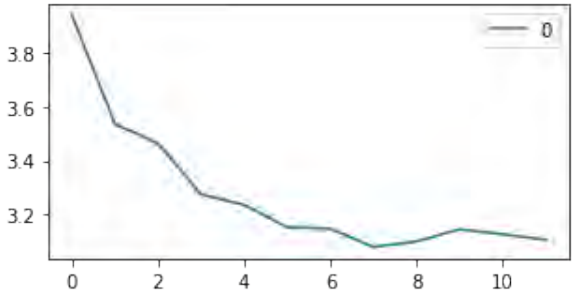
- **Καταναλωτές κανονικής τιμολόγησης, πλήθους 3025.**

  Κατά τον ίδιο τρόπο δουλεύουμε και σε αυτή την περίπτωση. Δοκιμάζουμε για K=[1...,12].

  Και σε αυτή την περίπτωση η μέθοδος clustering συνεισφέρει στην ακρίβεια της πρόβλεψης. Μάλιστα μειώνει τον δείκτη MAPE σχεδόν μία μονάδα.

| | K= | MAPE |
|---|---|---|
| | 1 | 3.9386 |
| min | 8 | 3.0800 |
| max | 15 | 3.9386 |

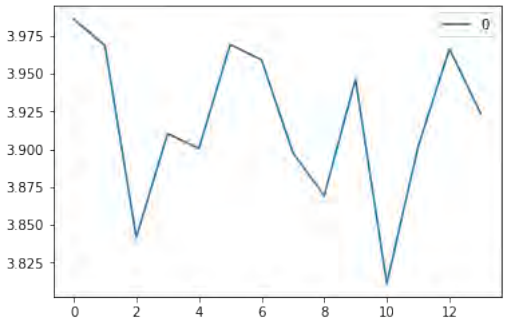
Πίνακας 5.9: SVR MAPE τιμές για διάφορα K



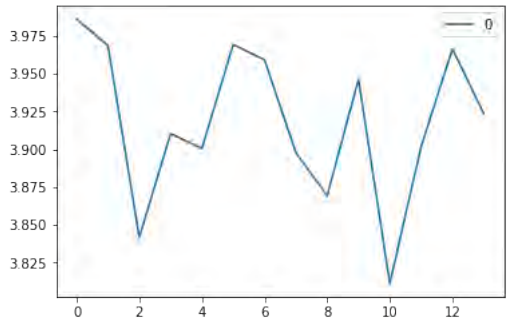

Σχήμα 5.10: SVR MAPE K clusters καταναλωτών σταθερής τιμολόγησης.

### 5.2.2 Cluster Based Aggregated Forecast με τις ExtraTrees,GradientBoosting και Linear Regression

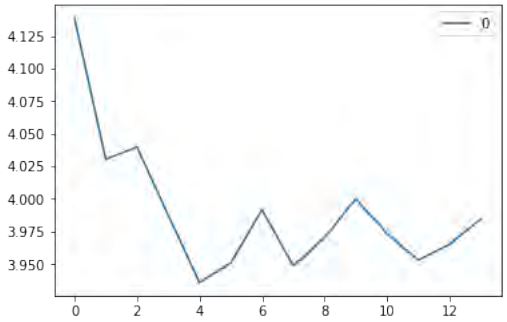
Για K=[1,...15] και για το σύνολο των καταναλωτών δυναμικής τιμολόγησης έχουμε:




(α΄) ExtraTress MAPE K clusters καταναλωτών δυναμικής τιμολόγησης




(β΄) GradienBoosting MAPE K clusters καταναλωτών δυναμικής τιμολόγησης




Σχήμα 5.12: Linear Regression K clusters καταναλωτών δυναμικής τιμολόγησης

Παρατηρούμε ότι η τεχνική Cluster based Aggregate Forecast load μπορεί να μειώσει το σφάλμα πρόβλεψης για όλους τους αλγορίθμους πρόβλεψης.

Πίνακας 5.10

| | K= | MAPE |
|---|---|---|
| | 1 | 3.9861 |
| min | 10 | 3.81105 |
| max | 3 | 3.96928 |

Πίνακας 5.11: ExtraTreesMAPE τιμές για διάφορα K

Πίνακας 5.12

| | K= | MAPE |
|---|---|---|
| | 1 | 4.0824 |
| min | 7 | 3.6717 |
| max | 3 | 4.0824 |

Πίνακας 5.13: GBM MAPE τιμές για διάφορα K

| | K= | MAPE |
|---|---|---|
| | 1 | 4.13865 |
| min | 8 | 3.96514 |
| max | 3 | 4.13865 |

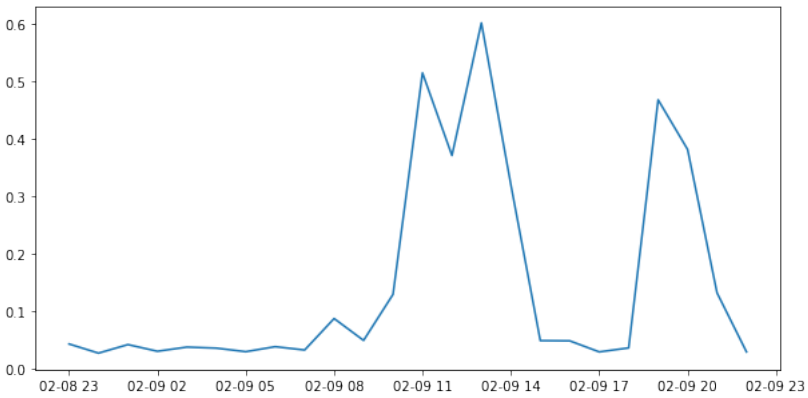
Πίνακας 5.14: Linear Regression MAPE τιμές για διάφορα K

## 5.3 Πρόβλεψη σε επίπεδο μετρητή (Meter Level Forecast)

Η πρόβλεψη σε επίπεδο σπιτιού αναμένουμε να έχει μεγαλύτερο σφάλμα πρόβλεψης από ότι στο συγκεντρωτικό επίπεδο εξαιτίας της υψηλής μεταβλητότητας του οικιακού φορτίου.

Επιλέγουμε τυχαία ένα κωδικό σπιτιού υποψήφιο για πρόβλεψη με δείγμα εκπαίδευσης 100 ημερών.

Η καμπύλη που θέλουμε να προσεγγίσουμε σύμφωνα με τις μεθόδους πρόβλεψης που έχουμε εφαρμόσει μέχρι τώρα είναι :



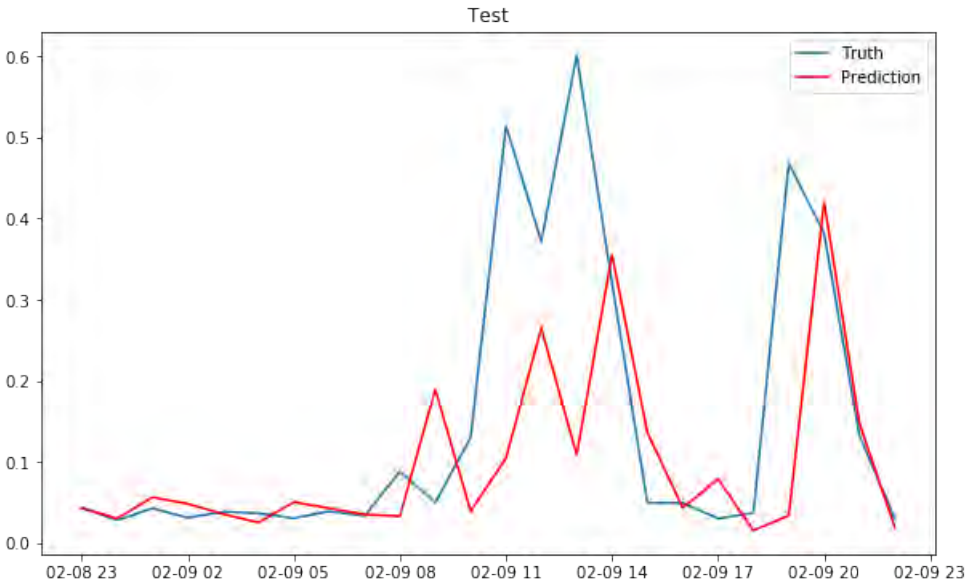

Σχήμα 5.13: Καμπύλη Φορτίου μετρητη με ID:MAC000349

Εφαρμόζοντας τις τεχνικές πρόβλεψης έχουμε :

Βλέπουμε ότι από τα ποσοστά ότι κανένα μοντέλο δεν μπορεί να προσεγγίσει τα δεδομένα. Παρατηρούμε μάλιστα και το φαινόμενο του overfitting για τις μεθόδους των Δέντρων αποφάσεων που ενώ σημειώνουν πολύ καλές επιδόσεις για το training set αποτυγχάνουν στο test set. Η μέθοδος που αποδίδει καλύτερα είναι η GradientBoost.

| Μέθοδος | Training Time | Training Score | Test Score | MAPE |
|---|---|---|---|---|
| SVR | 0.7448849678039551 | 63.16208256657367 | 21.292396008405888 | 102.42549155183349 |
| GradientBoosting | 3.8219516277313232 | 94.83592614336614 | 14.035813519746243 | 58.07811891474597 |
| RandomForest | 2.7499141693115234 | 91.8252117013704 | 28.145695083290313 | 64.78072184779198 |
| ExtraTrees | 1.3124535083770752 | 99.99999418955092 | 26.490613070189937 | 84.34523359145572 |
| LinearRegression | 0.0 | 41.94679692003761 | 36.939949307347256 | 84.34523359145572 |

Πίνακας 5.15: Πρόβλεψη φορτίου για μετρητή με ID:MAC000349




Σχήμα 5.14: Προσέγγιση καμπύλης φορτίου με την GradientBoost

# Κεφάλαιο 6

# Επίλογος

## 6.1 Τελικά συμπεράσματα

Σκοπός αυτής της εργασίας ήταν να εξετάσει τις πιο διαδεδομένες τεχνικές βραχυπρόθεσμης πρόβλεψης στον τομέα της ηλεκτρικής ενέργειας. Η αρχή της καταλληλότερης μεθόδου δεν μπορούμε να πούμε με σιγουριά ότι ισχύει, καθώς για διαφορετικά σετ εκπαίδευσης ο αλγόριθμος που σημείωνε την καλύτερη επίδοση δεν ήταν πάντα ο ίδιος.

Σαφή ήταν τα συμπεράσματα που εξήχθησαν σε σχέση με τα κριτήρια επιλογής των παραμέτρων. Για όλες σχεδόν τις μεθόδους παρουσιάστηκε βελτίωση στα ποσοστά σφάλματος. Ακόμη η προσέγγιση του Cluster Based Aggregated Forecast που αναφέρεται στην βιβλιογραφία πως ενισχύει τη ακρίβεια του μοντέλου διασταυρώθηκε και επιβεβαιώθηκε ότι ισχύει. Αυτό που κατάφερε λοιπόν, η ανάλυση της βραχυπρόθεσμης πρόβλεψης είναι να εξάγει κάποιες βασικές αρχές που ισχύουν σε κάθε περίπτωση και όχι για το ποιος είναι ο καλύτερος αλγόριθμος.

## 6.2 Μελλοντικές προτάσεις

Μία ερώτηση που πρέπει να κάνει κανείς είναι κατά πόσο ο ανθρώπινος παράγοντας επηρεάζει την κατανάλωση ενέργειας και πως αυτό εξάγεται μέσα από τα δεδομένα. Στην παρούσα εργασία η μέθοδος clustering που υλοποιήσαμε βασίζεται στη μέση ημερήσια κατανάλωση ανά ώρα. Τα αποτελέσματα ήταν ενθαρρυντικά, ωστόσο θα μπορούσαμε να εξετάσουμε και τους κοινωνικούς δείκτες των ομάδων που συμμετείχαν στο πρόγραμμα Low Carbon. Υπάρχει το ενδεχόμενο, το cluster να είναι περισσότερο ομογενές αν συμπεριλάβουμε καταναλωτές με τα ίδια κοινωνικά χαρακτηριστικά με συνέπεια να εξάγονται καλύτερες προβλέψεις για κάθε cluster. Τέλος, η ανάλυση που έγινε στα πλαίσια ερευνητικού ενδιαφέροντος θα μπορούσε να βρει και εμπορικές εφαρμογές στον χώρο της ενέργειας. Είναι δυνατό, να υλοποιηθεί μια πλατφόρμα παρακολούθησης ενεργειακών δεδομένων σε πραγματικό χρόνο που θα παράγει αναλυτικά δεδομένα ως προς την πρόβλεψη του φορτίου με άμεση συνέπεια την ρύθμιση της πολιτικής της δυναμικής τιμολόγησης.

# Βιβλιογραφία


[1] Z. Enwang A. Khotanzad και H. Elragal. A neuro-fuzzy approach to shortterm load forecasting in a price-sensitive environment. *IEEE Transactions on Power Systems*, 17:1273–1282, 2002.

[2] K. Amasyali και N.M.f El-Gohary. A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, 81:1192–1205, 2018.

[3] Souhaib Ben Taieb και Rob Hyndman. A Gradient Boosting Approach to the Kaggle Load Forecasting Competition. *International Journal of Forecasting*, 30, 2013.

[4] Understanding the Bias-Variance Tradeoff. `http://scott.fortmann-roe.com/docs/BiasVariance.html`.

[5] Bo-Juen Chen, Ming-Wei Chang και Chih-Jen lin. Load forecasting using support vector Machines: a study on EUNITE competition 2001. *IEEE Transactions on Power Systems*, 19(4):1821–1830, 2004.

[6] Liefeng Bo, Ling Wang και Licheng Jiao. Training Hard-Margin Support Vector Machines Using Greedy Stagewise Algorithm. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 19:1446–55, 2008.

[7] N. Charlton και C. Singleton. A refined parametric model for short term load forecasting. *International Journal of Forecasting*, 30:364–368, 2014.

[8] F.M. Bianchi Dang-Ha, T.-H. και R. Olsson. Local short term electricity load forecasting: Automatic approaches. in Neural Networks. *IJCNN*, 2017.

[9] Yong Ding, Martin Neumann, Per Silva και Michael Beigl. A framework for short-term activity-aware load forecasting, 2013. doi: 10.1145/2516911.2516919.

[10] E. Elham, A. Kareem και A. Hesham. A Review of Smart Meter Load Forecasting Techniques: Scale and Horizon. 2018.

[11] `https://energyplus.net/`.

[12] F. D. Galiana G. Gross. Short-term load forecasting. *Proceedings of the IEEE*, 75:1558 – 1571, 1987.

[13] `https://www.gridlabd.org/`.

[14] Strbac G. Demand side management: Benefits and challenges. *Energy Policy*, 36:4419–4426, 2008. doi:10.1016/j.enpol.2008.09.030.

[15] T. Hong και S. Fhan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32:914–938, 2016.

[16] `http://www.ucd.ie/issda/data/commissionforenergyregulationcer/`.

[17] V. Petridis J. Kiartzis, A. Bakirtzis. Short-term load forecasting using neural networks. *Electric Power Systems Research*, 33:1–6, 1995.

[18] Y.M. Park J.H. Park και K.Y. Lee. Composite modeling for adaptive short-term load forecasting. *1EEE Trans. Power Syst.*, 6:450–457, 1991.

[19] P. Ji, D. Xiong, P. Wang και J. Chen. A Study on Exponential Smoothing Model for Load Forecasting. σελίδες 1–4, 2012.

[20] J.L. Rueda Khuntia, S.R. και M.A.van der Meijden. Forecasting the load of electrical power systems in mid-and long-term horizons: a review. *IET Generation, Transmission Distribution*, 10:3971–3977, 2016.

[21] C. Kuster, Y. Rezgui και M. ourshed. Electrical load forecasting models: A critical systematic review. *Sustainable Cities and Society*, 35:257–270, 2017.

[22] `https://innovation.ukpowernetworks.co.uk/`.

[23] Alexey Natekin και Alois Knoll. Gradient Boosting Machines, A Tutorial. *Frontiers in neurorobotics*, 7:21, 2013.

[24] A. D. Papalexopoulos και T. C. Hesterberg. A regression-based approach to short-term system load forecasting. *IEEE Transactions on Power Systems*, 5:1535–1547, 1990.

[25] R. Weron. Rostamizadeh και A. Talwalkar. *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*. 2006.

[26] M. Wahab R.S. Elias, L. Fang. Electricity load forecasting based on weather variables and seasonalities: A neural network approach. *Proceedings of ICSSSM11*, 2011.

[27] S. Golowich V. Vapnik και A. Smola. Support vector method for function approximation, regression estimation, and signal processing. *Neural Information Processing Systems*, 9, 1997.

[28] D. J. Hill F. Luo W. Kong, Z. Y. Dong και Y. Xu. Short-Term Residential Load Forecasting Based on Resident Behaviour Learning. *IEEE Transactions on Power System*, 33:1087–1088, 2018.

[29] Yi Wang, Qixin Chen, Tao Hong και Chongqing Kang. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. *IEEE Transactions on Smart Grid*, PP, 2018.

[30] Ma. W, Fang. S, Liu. G και Zhou. R. Modeling of district load forecasting for distributed energy system. *Applied Energy*, 204:181–205, 2017.

[31] B. Yildiz. Recent advances in the analysis of residential electricity consumption and applications of smart meter data. *Applied Energy*, 208:402–427, 2017.