ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Ομαδοποίηση οικιακών φορτίων ηλεκτρικής ενέργειας για την διαχείριση ζήτησης με αλγορίθμους μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αφροδίτη Φραγκιαδάκη

Επιβλέπων: Ευάγγελος Μαρινάκης , Επίκουρος Καθηγητής Ε.Μ.Π
Υπεύθυνος: Ελισσαίος Β. Σαρμάς

Αθήνα, Φεβρουάριος 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Ομαδοποίηση οικιακών φορτίων ηλεκτρικής ενέργειας για την διαχείριση ζήτησης με αλγορίθμους μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αφροδίτη Φραγκιαδάκη

Επιβλέπων: Ευάγγελος Μαρινάκης , Επίκουρος Καθηγητής Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 29 Φεβρουαρίου 2024.

.....	.....	.....
Μαρινάκης Ε.	Ασκούνης Δ.	Δούκας Χ.
Επικ. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2024

.....

Αφροδίτη Φραγκιαδάκη

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

Περίληψη

Στη σημερινή εποχή της περιβαλλοντικής ευαισθητοποίησης και των βιώσιμων πρακτικών, η ανάπτυξη αποτελεσματικών στρατηγικών αντιμετώπισης της ζήτησης προϋποθέτει τη λεπτομερή κατανόηση των προτύπων κατανάλωσης ενέργειας των νοικοκυριών. Τέτοιες στρατηγικές βασίζονται στη δυναμική προσαρμογή της κατανάλωσης ενέργειας ανάλογα με τις συνθήκες προσφοράς, γεγονός που καθιστά την κατανόηση της συμπεριφοράς των χρηστών ζωτικής σημασίας. Η παρούσα διπλωματική εργασία εξετάζει διεξοδικά τις ενεργειακές συμπεριφορές των νοικοκυριών χρησιμοποιώντας ένα εκτεταμένο σύνολο δεδομένων από το London Data Store. Το σύνολο δεδομένων παρουσιάζει μετρήσεις κατανάλωσης ενέργειας από 5.567 νοικοκυριά του Λονδίνου που συμμετείχαν στο έργο Low Carbon London του UK Power Networks.

Ξεκινώντας με το ακατέργαστο σύνολο δεδομένων, εφαρμόστηκε μια στρατηγική προσέγγιση δειγματοληψίας για να καταστούν τα δεδομένα πιο διαχειρίσιμα και ευνοϊκά για ανάλυση. Μέσω σχολαστικής προεπεξεργασίας - που περιλαμβάνει εργασίες όπως ο χειρισμός των ελλιπών τιμών, ο μετασχηματισμός και η κανονικοποίηση - το σύνολο δεδομένων προετοιμάστηκε για περαιτέρω διερεύνηση. Η καινοτομία της έρευνας έγκειται στη φάση της μηχανικής των χαρακτηριστικών, όπου προέκυψαν χρήσιμα χαρακτηριστικά, για να αποτυπωθούν οι λεπτές διακυμάνσεις στην ενεργειακή κατανάλωση των νοικοκυριών. Αυτά τα χαρακτηριστικά άνοιξαν το δρόμο για την εφαρμογή μιας σειράς αλγορίθμων ομαδοποίησης μηχανικής μάθησης, συμπεριλαμβανομένων των αλγορίθμων K-Means++, Fuzzy C-means, Ιεραρχικής ομαδοποίησης, Αυτοοργανωτικών χαρτών (SOMs), BIRCH, Μοντέλων μίξης Gaussian (GMMs) και Spectral. Επιπλέον, διερευνήθηκε η συσταδοποίηση Ensembe.

Για να προσδιορίσουμε τον βέλτιστο αριθμό ομάδων και να συγκρίνουμε τις επιδόσεις των αλγορίθμων ομαδοποίησης, χρησιμοποιήσαμε μια σειρά από μετρικές αξιολόγησης. Αυτές περιλάμβαναν το Silhouette Score, το Davies-Bouldin Score, το Calinski-Harabasz Score και το Dunn Index. Χρησιμοποιώντας αυτές τις μετρικές, αξιολογήσαμε την αποτελεσματικότητα κάθε αλγορίθμου και επιλέξαμε τον καταλληλότερο για τη μελέτη περίπτωσης σχηματισμού έξι ομάδων. Μετά την ομαδοποίηση, δημιουργήθηκαν οπτικοποιήσεις που παρέχουν πληροφορίες για τα μοναδικά μοτίβα κατανάλωσης ενέργειας στις διάφορες ομάδες, αναδεικνύοντας μια σαφή διαφοροποίηση μεταξύ των κανονικοποιημένων και μη κανονικοποιημένων προφίλ φορτίου των ομάδων. Χρησιμοποιήθηκαν τεχνικές επεξηγήσιμης τεχνητής νοημοσύνης (ΤΝ) για την περαιτέρω διερεύνηση των πιο σημαντικών στοιχείων κάθε συστάδας.

Τα αποτελέσματα αυτής της έρευνας θέτουν τα θεμέλια για τη βελτίωση των τακτικών απόκρισης-ζήτησης με την προσαρμογή τους ώστε να ανταποκρίνονται καλύτερα στις συμπεριφορές των καταναλωτών και την αναγνώριση της αποτελεσματικότητας της μεθόδου Ensemble Clustering και της συμβολής της επεξηγήσιμης ΤΝ.

Λέξεις-κλειδιά: Στρατηγικές απόκρισης-ζήτησης, έξυπνοι μετρητές, μηχανική μάθηση, αλγόριθμοι ομαδοποίησης, K-Means++ , δεδομένα προεπεξεργασία, εξαγωγή χαρακτηριστικών, μετρικές αξιολόγησης συστάδων, Οπτικοποίηση δεδομένων , Ensemble ομαδοποίηση, Επεξηγίσιμη ΤΝ.

Abstract

In today's age of environmental awareness and sustainable practices, the development of effective Demand-Response (DR) strategies requires an intricate understanding of household energy consumption patterns. Such strategies hinge on dynamically adjusting energy consumption in response to supply conditions, making insights into user behavior crucial. This study extensively examines household energy behaviors using an extensive data set sourced from the London Data Store. The dataset presents energy consumption readings from 5,567 London households that actively participated in the UK Power Networks' Low Carbon London project from November 2011 to February 2014.

Beginning with the raw dataset, a strategic sampling approach was implemented to make the data more manageable and conducive to analysis. Through meticulous preprocessing — encompassing tasks like handling missing values, transformation, and scaling — the dataset was readied for further exploration. The research's novelty lies in its phase of feature engineering, where informative features, not previously included in the literature, were derived to capture the subtle variations in households' energy consumption. These engineered features paved the way for the application of a diverse suite of machine learning clustering algorithms, including K-Means++, Fuzzy C-means, Hierarchical clustering (Ward method), Self-Organizing Maps (SOMs), BIRCH, Gaussian Mixture Models (GMMs) and Spectral Clustering. Additionally, Ensembe Clustering was investigated, although it has received limited scholarly attention.

To determine the optimal number of clusters and rigorously compare the performance of multiple clustering algorithms, we employed a suite of cluster evaluation metrics. These included the Silhouette Score, Davies-Bouldin Score, Calinski-Harabasz Score, and the Dunn Index. By utilizing these metrics, we successfully assessed the efficacy of every algorithm and chose the most suitable one for the given case study of six cluster formation. Upon clustering, visualizations were generated to provide insights into the unique energy consumption patterns across the different clusters, highlighting a clear differentiation between normalized and non-normalized cluster load profiles. Explainable AI (XAI) techniques were utilized to further investigate the most significant elements of each cluster as determined by the clustering algorithm.

The results of this research establish the foundation for improving Demand-Response tactics by customizing them to better suit consumer behaviors and recognizing the effectiveness of Ensemble Clustering and the contribution of Explainable AI. To gain a more thorough knowledge of these consuming patterns in the future, it would be beneficial to conduct a more extensive investigation into their underlying factors, by incorporating sociodemographic data. Additionally, it would be worthwhile to further explore the approaches of XAI.

Keywords: Household Energy Consumption, Demand-Response Strategies, Smart Meter, Machine Learning, Clustering Algorithms, K-Means++ , Data Preprocessing, Feature Engineering, Cluster Evaluation Metrics, Data Visualization , Ensemble Clustering , XAI.

Ευχαριστίες

Καθώς αυτό το ακαδημαϊκό ταξίδι φτάνει στο τέλος του, βρίσκομαι μέσα σε ένα μείγμα συναισθημάτων. Το τέλος αυτής της διαδρομής δεν είναι απλώς μια κατάληξη, αλλά και η έναρξη νέων διαδρομών.

Εκφράζω τη βαθύτατη ευγνωμοσύνη μου στον επιβλέποντα μου, τον Ελισσαίο Σάρμα, του οποίου η καθοδήγηση και η στήριξη υπήρξε ανεκτίμητη. Στον καθηγητή μου, Βαγγέλη Μαρινάκη, ευχαριστώ για την ευκαιρία που μου δόθηκε να συνεργαστώ μαζί σας στο εργαστήριο, μια πραγματικά εποικοδομητική εμπειρία.

Στους φίλους και την οικογένειά μου, που υπήρξαν το ακλόνητο σύστημα υποστήριξής μου καθ' όλη τη διάρκεια αυτού του ακαδημαϊκού ταξιδιού. Η πίστη σας στις φιλοδοξίες μου, η κατανόησή σας κατά τη διάρκεια των ατελείωτων ωρών εργασίας και το γέλιο και η παρηγοριά που μου προσφέρατε ήταν το στήριγμα μου. Αυτό το επίτευγμα δεν είναι μόνο δικό μου, αλλά μια απόδειξη της διαρκούς αγάπης και της υποστήριξης που μου δώσατε απλόχερα. Για κάθε στιγμή αμφιβολίας που μετατρέψατε σε ελπίδα και για κάθε πρόκληση που με βοηθήσατε να ξεπεράσω, είμαι αιώνια ευγνώμων.

Αφροδίτη Φραγκιαδάκη,

Αθήνα, Φεβρουάριος 2024

Contents

List of Figures

List of Tables

Ευρεία Περίληψη

Κεφάλαιο 1: Εισαγωγή

Η κατανάλωση ενέργειας εξακολουθεί να αποτελεί μείζον ζήτημα στον σημερινό κόσμο της ραγδαίας προόδου, συνυφαίνοντας τις απαιτήσεις βιωσιμότητας με την πολυπλοκότητα της σύγχρονης ζωής. Αναπόφευκτα, καθώς η αστικοποίηση και η αύξηση του πληθυσμού συνεχίζουν να αυξάνονται, το ίδιο συμβαίνει και με την ανάγκη για ενέργεια. Αυτή η αυξημένη ζήτηση δημιουργεί τεράστια πίεση στις ενεργειακές υποδομές που έχουμε, καθιστώντας επιτακτική τη δημιουργία πολιτικών που ενθαρρύνουν την ενεργειακή απόδοση.

Η στρατηγική ζήτησης-απόκρισης είναι ένας τέτοιος μηχανισμός που είναι απαραίτητος για τα ευφυή ενεργειακά συστήματα και τα έξυπνα δίκτυα. Η ζήτηση-απόκριση είναι μια στρατηγική διαχείρισης της ενέργειας που έχει σχεδιαστεί για να μειώσει το χάσμα μεταξύ μεταξύ της προσφοράς ενέργειας και της ζήτησης των καταναλωτών. Στόχος της είναι να παρακινήσει τους καταναλωτές, τόσο ιδιώτες και βιομηχανίες, να αλλάξουν τις τυπικές καταναλωτικές συνήθειές τους ως απάντηση σε συγκεκριμένα σήματα, συνήθως διακυμάνσεις των τιμών ή κίνητρα. Με την ενσωμάτωση ευφυούς τεχνολογίας όπως οι έξυπνοι μετρητές και οι συσκευές IoT, η ανταπόκριση στη ζήτηση περιλαμβάνει κάτι περισσότερο από απλά μείωση ή μετατόπιση της χρήσης ενέργειας σε περιόδους υψηλής ζήτησης. Περιλαμβάνει επίσης τη χρήση δεδομένων σε πραγματικό χρόνο για τη λήψη δυναμικών και τεκμηριωμένων επιλογών.

Ωστόσο, η σύνθετη φύση της ζήτησης-απόκρισης εκτείνεται πέρα από την τεχνολογία. Η ζήτηση-απόκριση περιστρέφεται γύρω από την ανθρώπινη συμπεριφορά, καθιστώντας την πολύπλοκη και πολύπλευρη πρόκληση. Η κατανόηση του τρόπου με τον οποίο οι καταναλωτές, υπό τις ιδιαίτερες περιστάσεις τους, αντιδρούν στα σήματα απόκρισης ζήτησης είναι εξαιρετικά σημαντική. Η συσχέτιση μεταξύ της ενεργειακής απόδοσης και της ανταπόκρισης στη ζήτηση έγκειται στην κατανόηση των βαθύτερων λόγων που κρύβονται πίσω από τις πρότυπα κατανάλωσης ενέργειας, προκειμένου να σχεδιαστούν στρατηγικές για πιο συνετή χρήση.

Η επιδίωξη της ενεργειακής απόδοσης δεν αποτελεί μόνο απαίτηση για τη διατήρηση της περιβαλλοντικής βιωσιμότητας, αλλά και μια πρακτική στρατηγική για τη διασφάλιση της οικονομικής σταθερότητας. Η βελτιστοποίηση της κατανάλωσης ενέργειας ελαχιστοποιεί τις περιττές ενεργειακές δαπάνες, μειώνει τα ποσοστά των πόρων εξάντλησης των πόρων και μετριάζει τις περιβαλλοντικές επιπτώσεις. Σήμερα, η ενεργειακή απόδοση δεν είναι απλώς ένας μοντέρνος όρος- είναι απαίτηση, καθήκον και, κυρίως, μια έξυπνη προσέγγιση για την επίτευξη βιώσιμης ανάπτυξης.

Ένα βασικό συστατικό αυτής της στρατηγικής βασίζεται στην ικανότητα ταξινόμησης και διαχωρισμού των καταναλωτές ενέργειας. Η κατανόηση των μοναδικών καταναλωτικών προτύπων, συμπεριφορών και προτιμήσεών τους δεν είναι ένα απλό ή καθολικό έργο. Η έννοια της ομαδοποίησης προκύπτει ως ένα ισχυρό εργαλείο για την ανάλυση των ομοιογενών ομάδων σε μια ποικιλόμορφη αγορά ενέργειας. Μέσω της διαδικασίας ταξινόμησης των χρηστών ενέργειας σύμφωνα με τις ομοιότητες στα καταναλωτικά τους πρότυπα, καθίσ-

ταται εφικτή η προσαρμογή στρατηγικών ζήτησης-απόκρισης που ευθυγραμμίζονται με τις ξεχωριστέςτάσεις συμπεριφοράς κάθε ομάδας. Οι προσαρμοσμένες στρατηγικές έχουν την ικανότητα να δημιουργούναυξημένα επίπεδα δέσμευσης στα προγράμματα ανταπόκρισης στη ζήτηση, βελτιώνοντας έτσι τη συνολική αποτελεσματικότητα του συστήματος.

Η ομαδοποίηση όχι μόνο διευκολύνει την αποτελεσματική ανάπτυξη στρατηγικής αλλά και ενισχύει την αποτελεσματική επικοινωνία. Η μετάδοση των σημάτων απόκρισης στη ζήτηση ή η επιλογή των κατάλληλων κινήτρων σε μία σαφώς καθορισμένη ομάδα αυξάνει την πιθανότητα να ληφθεί μια ευνοϊκή αντίδραση από τους καταναλωτές.

Τελικά, καθώς πλοηγούμαστε στην πορεία της τεχνολογικής προόδου, η σημασία της ανθρωποκεντρικών μεθοδολογιών στη διαχείριση της ενέργειας γίνεται όλο και πιο εμφανής. Αυτή η μελέτη εξετάζει διεξοδικά τη συγχώνευση της τεχνολογίας και της ανθρώπινης συμπεριφοράς, χρησιμοποιώντας την ομαδοποίηση για την ενίσχυση του παραδείγματος ζήτησης-απόκρισης.

Στόχος και αντικείμενο της διπλωματικής

Όπως αναφέρθηκε προηγουμένως, η απόκτηση γνώσεων σχετικά με τα καταναλωτικά πρότυπα είναι ζωτικής σημασίας για την αποτελεσματικότερη διαχείριση της ενέργειας και τη μακροπρόθεσμη βιωσιμότητα στο δυναμικό ενεργειακό μας περιβάλλον. Η παρούσα διπλωματική διερευνά τον τομέα της διαχείρισης της ενέργειας, με ιδιαίτερη έμφαση στη βελτίωση των στρατηγικών ζήτησης-απόκρισης μέσω της ανάλυσης δεδομένων.

Οι πρωταρχικοί στόχοι αυτής της έρευνας είναι η διεξοδική εξέταση των δεδομένων κατανάλωσης ενέργειας των νοικοκυριών, προκειμένου να αποκαλυφθούν διακριτά πρότυπα και συμπεριφορές. Θα χρησιμοποιήσουμε μια σειρά από εξελιγμένους αλγορίθμους ομαδοποίησης για να διαχωρίσουμε τους καταναλωτές σε διακριτές ομάδες, επιτρέποντας την ανάπτυξη προσαρμοσμένων στρατηγικών απόκρισης στη ζήτηση. Μετά την εφαρμογή αυτών των τεχνικών, θα αξιολογήσουμε την απόδοσή τους και θα πραγματοποιήσουμε μια ολοκληρωμένη ανάλυση των αποτελεσμάτων για να διασφαλίσουμε ότι πληρούν τις απαιτήσεις των πρακτικών εφαρμογών.

Η παρούσα διπλωματική επιδιώκει να επιτύχει αυτούς τους στόχους προκειμένου να δημιουργήσει μια σύνδεση μεταξύ της ακαδημαϊκής έρευνας και των πρακτικών εφαρμογών στον τομέα της διαχείρισης της ενέργειας. Εν τέλει, στοχεύει να συμβάλει πολύτιμα προς ένα μέλλον που θα είναι πιο αποδοτικό ενεργειακά.

Συνεισφορά και αξία της διπλωματικής

Η κατανόηση των προτύπων ενεργειακής κατανάλωσης των νοικοκυριών είναι ζωτικής σημασίας στον πολύπλοκο τομέα της διαχείρισης της ενέργειας και των βιώσιμων πρακτικών. Η παρούσα έρευνα προωθεί, ωφελώντας τους παραπάνω τομείς:

Ενεργειακές στρατηγικές και στρατηγικές ζήτσης-απόρικρης

Η παρούσα μελέτη παρέχει πολύτιμες πληροφορίες για τη βελτίωση των στρατηγικών διαχείρισης της ενέργειας και της απόκρισης στη ζήτηση, όχι μόνο ακαδημαϊκές λεπτομέρειες. Η κατανόηση και η κατηγοριοποίηση της συμπεριφοράς των καταναλωτών βοηθά τους ενδιαφερόμενους ενεργειακούς παράγοντες να δημιουργήσουν πιο αποτελεσματικές στρατηγικές, εξασφαλίζοντας μια ισορροπημένη προσφορά και ζήτηση ενέργειας.

Αξιολόγηση Αλγορίθμων
Στην παρούσα διπλωματική συγκρίνονται και εφαρμόζονται διάφοροι αλγόριθμοι ομαδοποίησης σε δεδομένα κατανάλωσης ενέργειας. Μέσω δοκιμών, παρέχουμε μια ολοκληρωμένη αξιολόγηση της ικανότητας των αλγορίθμων να εντοπίζουν πρότυπα κατανάλωσης ενέργειας. Η μέθοδος αυτή αναδεικνύει τα πλεονεκτήματα και τα μειονεκτήματα κάθε αλγορίθμου και αποκαλύπτει ποιοι είναι οι καλύτεροι για τα ενεργειακά δεδομένα που έχουμε στην περίπτωση μας.

Ανάλυση κατανάλωσης ενέργειας
Η παρούσα μελέτη αναλύει την κατανάλωση ενέργειας των συστάδων πέραν της συλλογής δεδομένων. Η ταξινόμηση των νοικοκυριών με βάση τη χρήση ενέργειας μας δίνει πολύτιμες πληροφορίες για τη συμπεριφορά της κατανάλωσης. Αυτή η συστηματική προσέγγιση μετατρέπει τα ακατέργαστα δεδομένα σε χρήσιμη γνώση και ωφελεί τους ενδιαφερόμενους φορείς διαχείρισης ενέργειας.

Ευφυΐα δεδομένων στα ενεργειακά δεδομένα
Αυτή η διατριβή χρησιμοποιεί μια νέα μεθοδολογία ανάλυσης δεδομένων για τον εντοπισμό καινοτόμων χαρακτηριστικών σε μη επεξεργασμένα δεδομένα. Η μελέτη αυτή επικεντρώνεται στην προσεκτική προεπεξεργασία δεδομένων, τον μετασχηματισμό και την προηγμένη μηχανική χαρακτηριστικών. Η μελέτη αυτή βελτιώνει την ανάλυση δεδομένων κατανάλωσης ενέργειας με τον εντοπισμό και την εξαγωγή ξεχωριστών χαρακτηριστικών.

Κοινωνικός και περιβαλλοντικός αντίκτυπος
Η παρούσα έρευνα επηρεάζει την κοινωνία. Η έμφαση στην ενεργειακή απόδοση και τη βιωσιμότητα υποστηρίζει τις παγκόσμιες προσπάθειες για την κλιματική αλλαγή και την περιβαλλοντική διαχείριση.

Η έρευνα αυτή έχει σημαντικό αντίκτυπο πέραν των ακαδημαϊκών κύκλων. Στοχεύει στην καθοδήγηση των ενδιαφερόμενων φορέων της βιομηχανίας προς ένα βιώσιμο μέλλον, σύμφωνα με την αυξανόμενη σημασία της ενεργειακής απόδοσης και της βιωσιμότητας.

Δομή της διπλωματικής

Η διπλωματική αποτελείται από έξι κεφάλαια και ένα παράρτημα. Ακολουθεί μια σύντομη περιγραφή του περιεχομένου τους:

Κεφάλαιο 1
Το πρώτο κεφάλαιο της διπλωματικής παρέχει μια σύντομη εισαγωγή στο πρόβλημα, συμπεριλαμβανομένων των κύριων χαρακτηριστικών του και του ιστορικού πλαισίου. Καθορίζεται ο στόχος της διπλωματικής και η συμβολή στην επιστημονική κοινωνία. Τέλος, περιγράφεται η δομή της διπλωματικής.

Κεφάλαιο 2
Στο δεύτερο κεφάλαιο της διπλωματικής αναλύεται η απόκριση ζήτησης στα οικιακά ενεργειακά συστήματα, συμπεριλαμβανομένων των πλεονεκτημάτων, των προκλήσεων και της σημασίας της ομαδοποίησης των καταναλωτών. Η εισαγωγή στην απόκριση-ζήτηση ακολουθείται από μια λεπτομερή συζήτηση των πλεονεκτημάτων και των προκλήσεων της σε οικιακά περιβάλλοντα. Τονίζεται η σημασία της ομαδοποίησης των καταναλωτών για τη βελτίωση των προγραμμάτων ζήτησης-απόκρισης και εξετάζεται πώς η μηχανική μάθηση μπορεί να βελτιώσει τις στρατηγικές ομαδοποίησης.

Κεφάλαιο 3
Στο Κεφάλαιο 3, γίνεται ανασκόπηση της βιβλιογραφίας σε τέσσερις φάσεις. Στην πρώτη φάση, Προ-ομαδοποίησης, εξετάζουμε τη βιβλιογραφία για την προετοιμασία δεδομένων και την εξαγωγή χαρακτηριστικών. Στη δεύτερη φάση, Μεθοδολογίες ομαδοποίησης, εξετάζουμε διάφορους αλγορίθμους που χρησιμοποιούνται στη βιβλιογραφία. Ακολουθεί η τρίτη φάση, Μετρικές αξιολόγησης επιδόσεων, η οποία καλύπτει τις μετρικές αξιολόγησης αλγορίθμων. Η τελευταία φάση, Μετά-ομαδοποίησης, εξετάζει τη βιβλιογραφία σχετικά με την ερμηνεία των αποτελεσμάτων της ομαδοποίησης.

Κεφάλαιο 4
Η μεθοδολογία της διπλωματικής παρουσιάζεται στο κεφάλαιο 4. Το κεφάλαιο ξεκινά με μια επισκόπηση του συνόλου δεδομένων, ακολουθούμενη από τις διαδικασίες προεπεξεργασίας και εξαγωγής χαρακτηριστικών. Στη συνέχεια, παρουσιάζονται οι αλγόριθμοι ομαδοποίησης που θα χρησιμοποιηθούν. Το κεφάλαιο περιγράφει επίσης τις μετρικές αξιολόγησης για την αξιολόγηση της απόδοσης των αλγορίθμων.

Κεφάλαιο 5
Στο κεφάλαιο 5 της διπλωματικής, παρουσιάζουμε τα αποτελέσματα από διάφορους αλγορίθμους ομαδοποίησης. Το κεφάλαιο αυτό παρέχει μια διεξοδική ανάλυση αυτών των αλγορίθμων, αξιολογώντας την απόδοσή τους σε διάφορες δοκιμές. Μετά από ενδελεχή αξιολόγηση, επιλέγεται ο αλγόριθμος με την καλύτερη απόδοση για περαιτέρω ανάλυση. Χρησιμοποιώντας τον αλγόριθμο που επιλέχθηκε, εντοπίζονται και εξάγονται μοτίβα κατανάλωσης, παρέχοντας μια περιγραφή κάθε ομάδας καταναλωτών.

Κεφάλαιο 6
Το 6ο κεφάλαιο της διπλωματικής ολοκληρώνει το έργο και εξετάζει τις μελλοντικές προοπτικές.

Παράρτημα Α
Στο Παράρτημα Α παρουσιάζονται τα αποτελέσματα του Κεφαλαίου 5 που δεν συμπεριλήφθηκαν στο κυρίως κείμενο λόγω περιορισμένης έκτασης. Αυτό το παράρτημα συγκεντρώνει πρόσθετα δεδομένα και αναλύσεις για να παρέχει μια πιο ολοκληρωμένη κατανόηση των αλγορίθμων ομαδοποίησης που συζητήθηκαν στο Κεφάλαιο 5.

Κεφάλαιο 2: Το πρόβλημα της Απόκρισης-Ζήτησης

Στο κεφάλαιο 2 της διπλωματικής, αναλύεται εκτενώς ο μηχανισμός απόκρισης-ζήτησης στα σύγχρονα συστήματα διαχείρισης ενέργειας. Αρχικά, περιγράφεται λεπτομερώς η εξέλιξη του ενεργειακού δικτύου και η αυξανόμενη πολυπλοκότητα στη διαχείριση της προσφοράς και της ζήτησης ηλεκτρικής ενέργειας, ιδίως με την ενσωμάτωση των ανανεώσιμων πηγών ενέργειας. Στη συνέχεια, το κεφάλαιο διερευνά τους διάφορους τύπους προγραμμάτων ζήτησης-απόκρισης, κάνοντας διάκριση μεταξύ μοντέλων που βασίζονται σε κίνητρα και σε τιμές, και περιγράφει λεπτομερώς τη στρατηγική σημασία τους στη διαμόρφωση της συμπεριφοράς κατανάλωσης ηλεκτρικής ενέργειας.

Το κεφάλαιο εμβαθύνει περαιτέρω στα οφέλη της στατηγικής απόκρισης-ζήτησης , τονίζοντας τον κρίσιμο ρόλο της στην ενσωμάτωση των ανανεώσιμων πηγών ενέργειας και τη διασφάλιση της σταθερότητας του δικτύου. Υπογραμμίζει τα πλεονεκτήματα που προσφέρει η στρατηγική απόκρισης-ζήτησης τόσο στους καταναλωτές όσο και στην αγορά ενέργειας,

όπως η μείωση της ζήτησης αιχμής, η μείωση των δαπανών ηλεκτρικής ενέργειας και η προώθηση της αποτελεσματικότητας της αγοράς.

Ωστόσο, το κεφάλαιο καταδεικνύει και τις προκλήσεις στην εφαρμογή της οικιακής απόκρισης-ζήτησης , εστιάζοντας σε τεχνολογικά εμπόδια, όπως η ενσωμάτωση έξυπνων μετρητών και συσκευών IoT, και σε προκλήσεις συμπεριφοράς που σχετίζονται με τη δέσμευση των καταναλωτών και την ανταπόκριση στα σήματα απόκρισης-ζήτησης. Υπογραμμίζει τη σημασία της ομαδοποίησης των καταναλωτών για τον σχηματισμό αποδοτικών στρατηγικών απόκρισης-ζήτησης , υποστηρίζοντας τη χρήση της μηχανικής μάθησης, ιδίως των αλγορίθμων μάθησης χωρίς επίβλεψη, όπως η συσταδοποίηση, για την αποτελεσματική κατηγοριοποίηση των καταναλωτών με βάση τα πρότυπα χρήσης ενέργειας.

Εν κατακλείδι, το κεφάλαιο υπογραμμίζει τη σημασία του συνδυασμού της απόκρισης-ζήτησης με την ομαδοποίηση των καταναλωτών με βάση τη μηχανική μάθηση για τη δημιουργία ενός πιο προσαρμοστικού, εξορθολογισμένου και εστιασμένου στον καταναλωτή ενεργειακού μέλλοντος. Θέτει τις βάσεις για την περαιτέρω διερεύνηση των διαφόρων αλγορίθμων ομαδοποίησης και της εφαρμογής τους σε επόμενα κεφάλαια, με στόχο την παροχή μιας ολοκληρωμένης εικόνας των πλεονεκτημάτων, των περιορισμών και των μοναδικών προοπτικών που προσφέρουν στον μετασχηματισμό του ενεργειακού τομέα.

Κεφάλαιο 3: Επισκόπηση συναφών μεθοδολογιών

Το κεφάλαιο 3 της διπλωματικής εργασίας προσφέρει μια διεξοδική βιβλιογραφική ανασκόπηση των τεχνικών ομαδοποίησης με μηχανική μάθηση για την ομαδοποίηση των καταναλωτών ενέργειας, καθώς και μια ανάλυση των εργαλείων που χρησιμοποιήθηκαν για τη συγκέντρωση μιας ολοκληρωμένης βιβλιοθήκης. Η εν λόγω ανασκόπηση είναι μεθοδικά δομημένη σε διακριτές φάσεις, καθεμία από τις οποίες είναι αφιερωμένη στη διερεύνηση των υφιστάμενων μεθοδολογιών σε βάθος. Το κεφάλαιο παρέχει μια κριτική και συστηματική ανάλυση της τρέχουσας κατάστασης των τεχνικών ομαδοποίησης με μηχανική μάθηση, αναδεικνύοντας την εφαρμογή τους στο πλαίσιο της ανάλυσης και τμηματοποίησης της συμπεριφοράς των καταναλωτών ενέργειας.

Φάση Ι:Προετοιμασία δεδομένων πριν την ομαδοποίηση

Η προ-ομαδοποίηση είναι κρίσιμη για την προετοιμασία των δεδομένων, την απόκτηση γνώσεων σχετικά με τα αρχικά μοτίβα και τη δημιουργία μιας βάσης για αποτελεσματική ομαδοποίηση. Στην υπάρχουσα βιβλιογραφία χρησιμοποιήθηκαν πολλαπλές μεθοδολογίες για την αντιμετώπιση αυτού του συγκεκριμένου σταδίου. Η προεπεξεργασία δεδομένων περιλαμβάνει την κανονικοποίηση των δεδομένων και τον χειρισμό των ελλιπών τιμών, ενώ η εξαγωγή χαρακτηριστικών αντιμετωπίζει την "κατάρα της διάστασης" επιλέγοντας τα κατάλληλα χαρακτηριστικά για την ομαδοποίηση.

Οι μέθοδοι μείωσης του μεγέθους των δεδομένων περιλαμβάνουν την ανάλυση κύριων συνιστωσών (PCA), τον χάρτη Sammon και την ανάλυση κυρτών συνιστωσών (CCA). Ο αλγόριθμος ομαδοποίησης Hopfield-K-Means εισάγει δείκτες ισχύος για τη μείωση των δεδομένων, εστιάζοντας σε βασικά χρονικά διαστήματα για την αποτελεσματική σύνοψη των πληροφοριών της καμπύλης φορτίου. Η τεχνική SAX χρησιμοποιείται επίσης για τη

μετατροπή των καμπυλών φορτίου σε συμβολικές συμβολοσειρές, απλοποιώντας την πολυπλοκότητα των δεδομένων.

Φάση ΙΙ: Μεθοδολογίες ομαδοποίησης

Στη βιβλιογραφία παρουσιάζονται διάφορες μεθοδολογίες ομαδοποίησης για την κατηγοριοποίηση των καταναλωτών ηλεκτρικής ενέργειας, καθεμία με τα δικά της πλεονεκτήματα. Τα αντιπροσωπευτικά πρότυπα φορτίου (RLP) είναι ζωτικής σημασίας για την κατανόηση των συμπεριφορών κατανάλωσης ηλεκτρικής ενέργειας και είναι κανονικοποιημένα προφίλ φορτίου μεμονωμένων πελατών. Οι παραδοσιακοί αλγόριθμοι ομαδοποίησης χωρίς επίβλεψη χρησιμοποιούνται ευρέως και θεωρούνται η θεμελιώδης προσέγγιση για την κατηγοριοποίηση των καταναλωτών ηλεκτρικής ενέργειας. Η εργασία των Chicco et al. με τίτλο "Comparisons Among Clustering Techniques for Electricity Customer Classification" αξιολογεί διάφορες μεθοδολογίες ομαδοποίησης, συμπεριλαμβανομένων των μεθόδων K-means, ιεραρχικής ομαδοποίησης, ασαφούς K-means και της τροποποιημένης προσέγγισης follow-the-leader.

Οι μεθοδολογίες πολλαπλών σταδίων έχουν πρωταγωνιστήσει στη βιβλιογραφία σχετικά με την κατηγοριοποίηση των καταναλωτών ηλεκτρικής ενέργειας, όπως "A Hybrid Machine Learning Model for Electricity Consumer Categorization Using Smart Meter Data" και "A Clustering Approach to Domestic Electricity Load Profile". Η τεχνική ταχείας αναζήτησης και εύρεσης κορυφών πυκνότητας (Fast Search and Find of Density Peaks - CFSFDP) είναι κομβικής σημασίας για τη σκιαγράφηση των συμπεριφορών κατανάλωσης ηλεκτρικής ενέργειας, η οποία εκτιμάται για τη χαμηλή χρονική πολυπλοκότητά της και την ανθεκτικότητά της στο θόρυβο. Η μελέτη ενσωματώνει μια προσέγγιση διαίρει και βασίλευε, εφαρμόζοντας προσαρμοστικό k-means σε τοπικές τοποθεσίες για την απόκτηση αντιπροσωπευτικών προφίλ πελατών, ακολουθούμενη από μια τροποποιημένη μέθοδο CFSFDP σε παγκόσμιες τοποθεσίες για αποτελεσματική επεξεργασία δεδομένων.

Φάση ΙΙΙ: Μετρικές αξιολόγησης της απόδοσης

Οι δείκτες εγκυρότητας ομαδοποίησης (CVIs) χρησιμοποιούνται για τη μέτρηση της συμπαγούς μορφής των μοτίβων φόρτωσης εντός μιας συστάδας και του διαχωρισμού τους από άλλες συστάδες.

Ο τροποποιημένος δείκτης Dunn (MDI) επικεντρώνεται στον λόγο της μικρότερης απόστασης μεταξύ των συστάδων προς τη μεγαλύτερη απόσταση εντός των συστάδων. Ο δείκτης διασποράς (SI) προκύπτει από την αναλογία της διασποράς που αντιπροσωπεύει η συσταδοποίηση. Ο δείκτης Davies-Bouldin (DB) αντιπροσωπεύει το μέσο μέτρο ομοιότητας κάθε συστάδας με την πιο παρόμοια συστάδα της. Το 2011, η Chicco συμπλήρωσε τους δείκτες CVI με δείκτες όπως ο δείκτης εντός συστάδας (IAI), το κριτήριο αναλογίας διακύμανσης (VRC), ο δείκτης μεταξύ συστάδων (IEI), ο μέσος δείκτης επάρκειας (MIA), ο δείκτης πίνακα ομοιότητας (SMI) και ο λόγος του αθροίσματος τετραγώνων εντός συστάδας προς τη διακύμανση μεταξύ συστάδων (WCBCR). Ο δείκτης Calinski-Harabasz (CH) είναι μια ευρέως χρησιμοποιούμενη μετρική για την αξιολόγηση της ποιότητας των αλγορίθμων συσταδοποίησης. Η βαθμολογία σιλουέτας (SIL) μετρά πόσο παρόμοιο είναι ένα αντικείμενο με τη δική του συστάδα (συνοχή) σε σύγκριση με άλλες συστάδες (διαχωρισμός). Ο δείκτης Entropy of Eigenvalues (EoE) προτείνεται για την εύρεση του βέλτιστου αριθμού συστάδων,

αλλά καταγράφει κυρίως τις γραμμικές σχέσεις μεταξύ των χρονοσειρών των συστάδων, περιορίζοντας ενδεχομένως την αποτελεσματικότητά του σε σενάρια με ισχυρές μη γραμμικές συσχετίσεις.

Φάση IV: Μεθοδολογίες μετά την ομαδοποίηση

Το τελικό στάδιο των μεθοδολογιών ομαδοποιήσης περιλαμβάνει τη σύγκριση και αξιολόγηση αλγορίθμων, τη δημιουργία κατηγοριών καταναλωτών με βάση τα χαρακτηριστικά των συστάδων και τον προσδιορισμό των τελικών προφίλ φορτίου για κάθε κατηγορία καταναλωτών, μετατρέποντας την ανάλυση συσταδοποίησης σε πρακτικές στρατηγικές.

Η μελέτη με τίτλο "Comparisons among Clustering Techniques for Electricity Customer Classification" εξετάζει διάφορους αλγορίθμους ομαδοποίησης για την ταξινόμηση πελατών ηλεκτρικής ενέργειας. Η τροποποιημένη follow-the-leader και η ιεραρχική ομαδοποίηση βρέθηκαν να είναι οι πιο αποτελεσματικές μέθοδοι, με πρακτικό όριο τις 15 έως 20 κλάσεις πελατών. Η μελέτη υπογράμμισε επίσης τη σημασία ενός μικρότερου αριθμού συστάδων για το σχεδιασμό προσαρμοσμένων προγραμμάτων για κάθε κατηγορία πελατών.

Η διαδικασία επιλογής του αλγορίθμου συσταδοποίησης περιλαμβάνει τον προσδιορισμό των κλάσεων προφίλ (PCs) με την ανάλυση των καμπυλών φορτίου κάθε συστάδας. Αυτό βοηθά στη διάκριση μοναδικών μοτίβων κατανάλωσης ενέργειας και στην τμηματοποίηση των πελατών σε διακριτές κλάσεις προφίλ. Η δημιουργία των PCs περιλαμβάνει τη μέση τιμή για αντιπροσωπευτικά προφίλ, την κανονικοποίηση των τιμών, τη συγχώνευση μικρών συστάδων με παρόμοια μοτίβα, την εξέταση της μεταβλητότητας και τη σύνδεση με τα χαρακτηριστικά των νοικοκυριών.

Συμπεράσματα

Η βιβλιογραφική ανασκόπηση που πραγματοποιήθηκε στο πλαίσιο της παρούσας διπλωματικής αποκάλυψε αρκετά σημαντικά κενά στην στην υπάρχουσα βιβλιογραφία σχετικά με τις τεχνικές ομαδοποίησης για τους καταναλωτές ενέργειας:

- **Μηχανική χαρακτηριστικών:** Υπάρχει αξιοσημείωτη έλλειψη λεπτομερούς διερεύνησης σχετικά με τα χαρακτηριστικά στο πλαίσιο της ομαδοποίησης για τη χρήση ενέργειας από τους καταναλωτές. Οι περισσότερες μελέτες βασίζονται κυρίως σε καμπύλες φορτίου, με περιορισμένη διερεύνηση ενούς ευρέως φάσματος χαρακτηριστικών. Η ανάδειξη νέων και καινοτόμων χαρακτηριστικών θα μπορούσε να ενισχύσει σημαντικά την αποτελεσματικότητα της ομαδοποίησης.
- **Αλγόριθμοι ομαδικής συσταδοποίησης:** Η βιβλιογραφία δείχνει μια σπανιότητα στην εφαρμογή αλγορίθμων ομαδικής ομαδοποίησης (ensemble clustering algorithms). Αυτοί οι αλγόριθμοι, γνωστοί για την ευρωστία τους και την ακρίβειά τους, θα μπορούσαν ενδεχομένως να προσφέρουν πιο διαφοροποιημένες γνώσεις στην τμηματοποίηση των καταναλωτών.
- **Εξηγήσιμη τεχνητή νοημοσύνη στη δημιουργία προφίλ**: Υπάρχει κενό στη χρήση της επεξηγήσιμης ΤΝ για τη δημιουργία προφίλ καταναλωτών. Η επεξηγήσιμη ΤΝ θα μπορούσε να παρέχει διαφάνεια και κατανοητότητα στη διαδικασία ομαδοποίησης, βοηθώντας στην υιοθέτηση αυτών των τεχνικών από μη ειδικούς ενδιαφερόμενους.

Κεφάλαιο 4: Μεθοδολογία

Εισαγωγή

Οπώς προειπώθηκε, η διπλωματική εργασία χρησιμοποιεί αλγορίθμους μηχανικής μάθησης για την ομαδοποίηση της κατανάλωσης ενέργειας. Η μελέτη χρησιμοποιεί δεδομένα από το London Data Store, 5.567 νοικοκυριών του Μεγάλου Λονδίνου. Το σύνολο δεδομένων, το οποίο καλύπτει την περίοδο από τον Νοέμβριο του 2011 έως τον Φεβρουάριο του 2014, περιλαμβάνει 167 εκατομμύρια σειρές ημίωρων μετρήσεων. Το σύνολο δεδομένων περιλαμβάνει βασικές μεταβλητές όπως LCLid, stdorToU (Standard ή Time of Use), DateTime και KWH/hh (Kilowatt-Hours per Half-Hour).Για να καταστεί το σύνολο δεδομένων διαχειρίσιμο για ανάλυση με τους διαθέσιμους υπολογιστικούς πόρους, χρησιμοποιήθηκε μια μέθοδος στρατηγικής δειγματοληψίας. Το σύνολο δεδομένων μειώθηκε ώστε να περιλαμβάνει μόνο τις δύο πρώτες εβδομάδες του Δεκεμβρίου, του Οκτωβρίου, του Απριλίου και του Ιουλίου του 2013. Η επιλογή αυτή βασίστηκε στην εποχιακή αντιπροσώπευση, στις περιόδους αιχμής και τις περιόδους εκτός αιχμής, στην ποικιλομορφία των δεδομένων και στην υπολογιστική σκοπιμότητα. Τα βήματας της μεθοδολογίας που ακολουθήθηκε παρουσιάζονται στο διάγραμμα 1 αναλυτικά.

Εξαγωγή χαρακτηριστικών

Στην ενότητα "Μηχανική χαρακτηριστικών" του κεφαλαίου 4, η διπλωματική εργασία υπογραμμίζει τον κρίσιμο ρόλο της εξαγωγής χαρακτηριστικών στην ανάπτυξη μοντέλων μηχανικής μάθησης για την ανάλυση της κατανάλωσης ενέργειας. Η διαδικασία περιλαμβάνει την κανονικοποίηση των δεδομένων και τη δημιουργία ειδικών χαρακτηριστικών που σχετίζονται με την κατανάλωση ενέργειας. Μερικά απο τα χαρακτηριστικά που υπολογίστηκαν είναι τα παρακάτω:

- Μέση κατανάλωση
- Μέση κατανάλωση αιχμής
- Συχνότητα αιχμής σε κάθε χρονοθυρίδα
- Μέση πρωινή/απογευματινή/βραδινή/απογευματινή κατανάλωση
- Μέση εποχιακή κατανάλωση
- Μέση κατανάλωση σε ώρες αιχμής και εκτός αιχμής
- Μέση χρήση τις καθημερινές και τα Σαββατοκύριακα
- Χαρακτηριστικά αυτοσυσχέτισης
- Εποχιακή τάση χρήσης
- Αύξηση της κατανάλωσης
- Μεταβλητότητα στις ώρες αιχμής

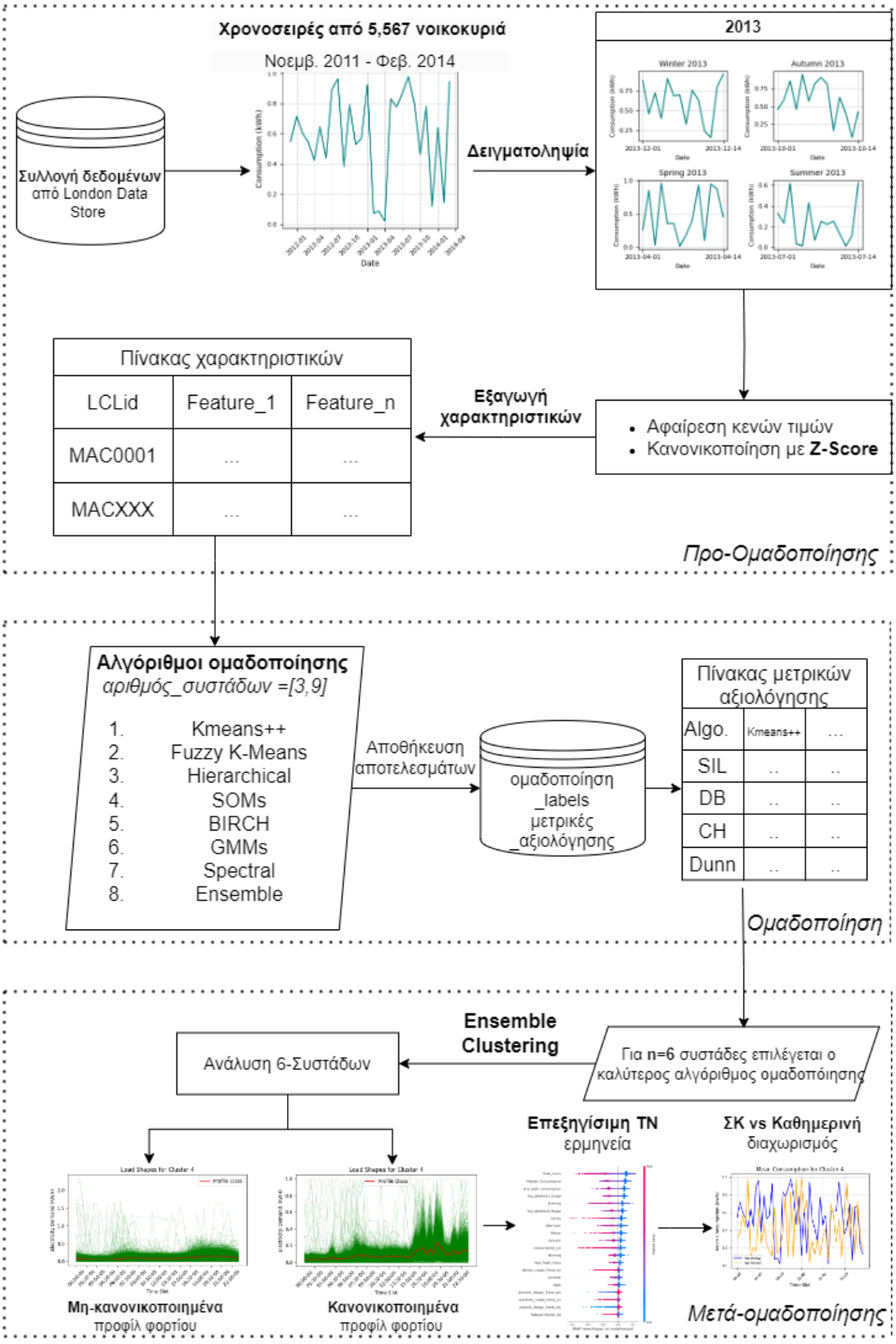


Figure 1: Βήματα μεθοδολογίας

Αλγόριθμοι μηχανικής μάθησης

Στην ενότητα "Αλγόριθμοι μηχανικής μάθησης" του κεφαλαίου 4, η διπλωματική εργασία περιγράφει μια σειρά αλγορίθμων που χρησιμοποιούνται για την ανάλυση ομαδοποίησης στην τμηματοποίηση των καταναλωτών ενέργειας. Οι αλγόριθμοι που εξετάζονται περιλαμβάνουν τους K-Means++, Fuzzy K-Means, Hierarchical Clustering, Self-Organizing Maps (SOMs), BIRCH, Gaussian Mixture Models (GMMs), Spectral Clustering και Ensemble Clustering.

Κάθε αλγόριθμος αναλύεται λεπτομερώς όσον αφορά την εφαρμογή του και τα μοναδικά χαρακτηριστικά του, παρέχοντας μια ολοκληρωμένη κατανόηση των εφαρμογών και της αποτελεσματικότητάς τους στο πλαίσιο των δεδομένων κατανάλωσης ενέργειας. Η ενότητα αυτή όχι μόνο διερευνά τις τεχνικές πτυχές αυτών των αλγορίθμων, αλλά συζητά επίσης και την πρακτική τους αλγοριθμική εφαρμογή στο περιβάλλον Python και την χρήση των κατάλληλων βιβλιοθηκών για την υλοποίηση τους.

Μετρικές αξιολόγησης της απόδοσης

Στο μέρος "Μετρικές αξιολόγησης" του Κεφαλαίου 4, η διπλωματική εργασία επικεντρώνεται στις μετρικές που χρησιμοποιούνται για την αξιολόγηση της απόδοσης των αλγορίθμων ομαδοποίησης που εφαρμόζονται στην τμηματοποίηση των καταναλωτών ενέργειας. Το τμήμα αυτό τονίζει τη σημασία της επιλογής των κατάλληλων μετρικών αξιολόγησης για την αξιολόγηση της αποτελεσματικότητας κάθε αλγορίθμου. Εξετάζονται διάφορες μετρικές όπως το Silhouette Score, το Davies-Bouldin Score, το Calinski-Harabasz Score και η Dunn Index, οι οποίες είναι κρίσιμες για τη μέτρηση της επιτυχίας της συσταδοποίησης όσον αφορά τον διαχωρισμό και τη συνοχή των συστάδων. Αυτή η ολοκληρωμένη ανάλυση των μετρικών αξιολόγησης είναι κομβικής σημασίας για τη διασφάλιση της αξιοπιστίας και της ακρίβειας των αποτελεσμάτων ομαδοποίησης που λαμβάνονται από τον αλγόριθμο μηχανικής μάθησης.

Κεφάλαιο 5: Αποτελέσματα

Το κεφάλαιο 5 περιγράφει την προσέγγιση για τη δοκιμή και τη σύγκριση των επιδόσεων διαφόρων αλγορίθμων ομαδοποίησης που εφαρμόζονται στα δεδομένα που έχουμε. Δίνει έμφαση στην ευθυγράμμιση αυτής της ανάλυσης με τις αρχές της επεξηγήσιμης ΤΝ, τονίζοντας τη σημασία όχι μόνο της αξιολόγησης της αποτελεσματικότητας των αλγορίθμων ομαδοποίησης αλλά και της διασφάλισης ότι τα αποτελέσματα είναι ερμηνεύσιμα και κατανοητά. Η προσέγγιση αυτή υπογραμμίζει τη σημασία της διαφάνειας και της προσβασιμότητας στις εφαρμογές μηχανικής μάθησης, ιδίως στο πλαίσιο της ανάλυσης της κατανάλωσης ενέργειας.

Αξιολόγηση αλγορίθμων

Στην ενότητα "Αξιολόγηση αλγορίθμων" του κεφαλαίου 5, η διπλωματική εργασία αξιολογεί τους διάφορους αλγορίθμους ομαδοποίησης χρησιμοποιώντας τους δείκτες εγκυρότητας συστάδων (CVI) που παρουσιάστηκαν στο κεφάλαιο 4. Η ανάλυση γίνεται

σε ένα εύρος συστάδων [3,9] και αποσκοπεί στην αξιολόγηση της αποτελεσματικότητας και της καταλληλότητας κάθε αλγορίθμου για την τμηματοποίηση των καταναλωτών ενέργειας. Αυτή η αυστηρή διαδικασία αξιολόγησης είναι ζωτικής σημασίας για τον προσδιορισμό του πιο αποτελεσματικού αλγορίθμου για πρακτική εφαρμογή στην ανάλυση της κατανάλωσης ενέργειας.Τα διαγραμματικά αποτελέσματα που προκύπτουν φαίνονται στα διαγράμματα 2,3,4 και 5.

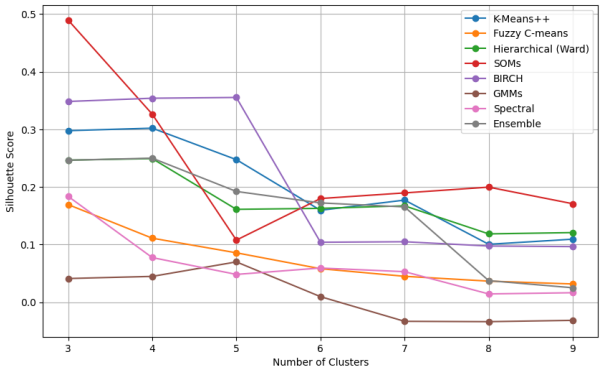


Figure 2: Silouette Score μετρική για διαφορετικούς αλγορίθμους ομαδοποίησης σε ένα εύρος αριθμών συστάδων [3,9].

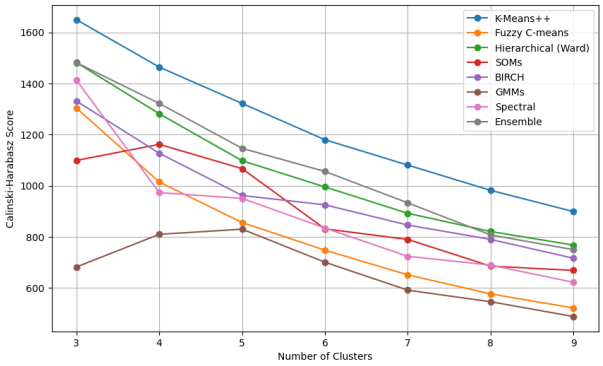


Figure 3: Calinski-Harabasz μετρική για διαφορετικούς αλγορίθμους ομαδοποίησης σε ένα εύρος αριθμών συστάδων [3,9].

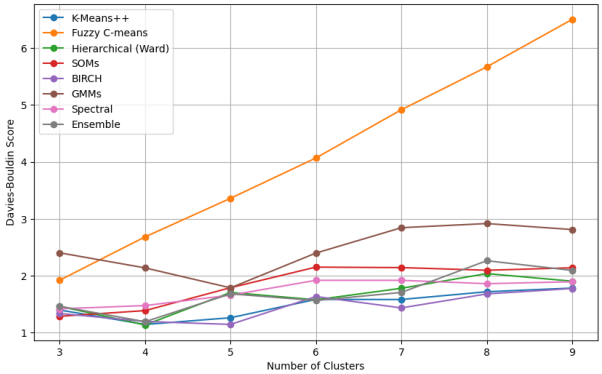


Figure 4: Davies-Bouldin Calinski-Harabasz μετρική για διαφορετικούς αλγορίθμους ομαδοποίησης σε ένα εύρος αριθμών συστάδων [3,9].

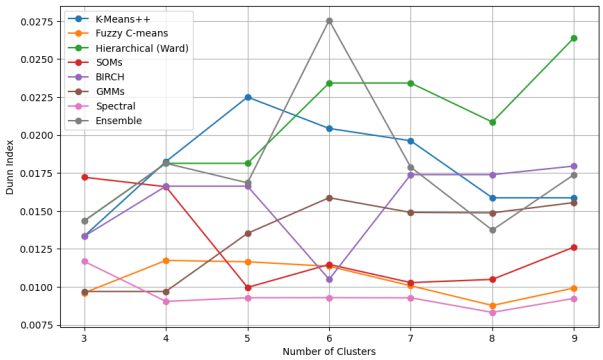


Figure 5: Dunn Index Calinski-Harabasz μετρική για διαφορετικούς αλγορίθμους ομαδοποίησης σε ένα εύρος αριθμών συστάδων [3,9].

Ανάλυση ομάδων με Ensemble Clustering

Στην ενότητα "Ανάλυση ομάδων με Ensemble Clustering " του κεφαλαίου 5, διεξάγεται διεξοδική ανάλυση των συστάδων που σχηματίζονται από τον αλγόριθμο Ensemble. Αυτή η εις βάθος εξέταση περιλαμβάνει τη διερεύνηση των μεγεθών και των χαρακτηριστικών των συστάδων, με συγκεκριμένη απόφαση να χρησιμοποιηθούν **έξι συστάδες**. Χρησιμοποιούνται προηγμένες τεχνικές οπτικοποίησης για να ενισχυθεί η κατανόηση αυτών των συστάδων, απεικονίζοντας τα διαφορετικά πρότυπα και τις συμπεριφορές χρήσης ενέργειας εντός κάθε ομάδας. Αυτή η ολοκληρωμένη ανάλυση είναι ζωτικής σημασίας για την ακριβή ερμηνεία των αποτελεσμάτων της ομαδοποίησης και την κατανόηση των διαφορετικών τμημάτων καταναλωτών στην αγορά ενέργειας.

Τα ενδεικτικά προφίλ (representative load profiles) που προκύπτουν απο τις μη-κανονικοποιημένες τιμές για κάθε ομάδα φαίνονται στο διάγραμμα 6.

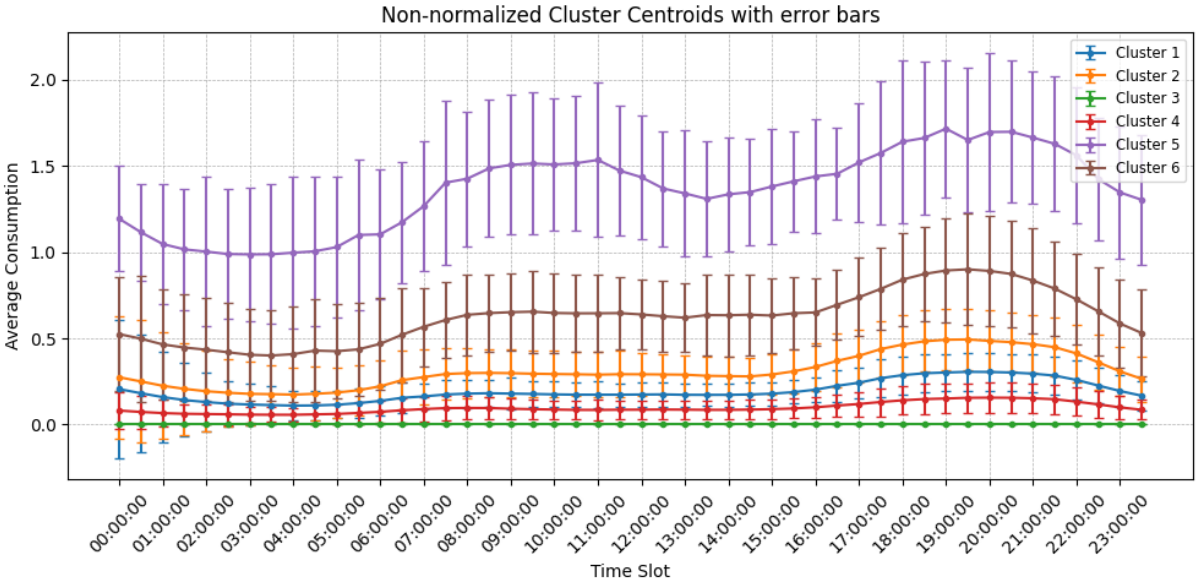


Figure 6: Ενδεικτικά προφίλ φορτίου ομάδας

Παρατηρούμε ότι έχει γίνει ενας σαφής διαχωρισμός των ομάδων ανάλογα με την κλί-

μακα καταναλάωσης τους (πολυ υψηλή - υψηλή - μέτρια - χαμηλή - μηδενική κατανάλωση). Για να διακρίνουμε καθημερινά διαφοροποιημένα μοτίβα κατανάλωσης, κανονικοποιούμε τις καταναλώσεις κάθε ομάδες και λαμβάνουμε τα κανονικοποιημένα ενδεικτικά προφίλ , τα οποία φαίνονται στο διάγραμμα 7.

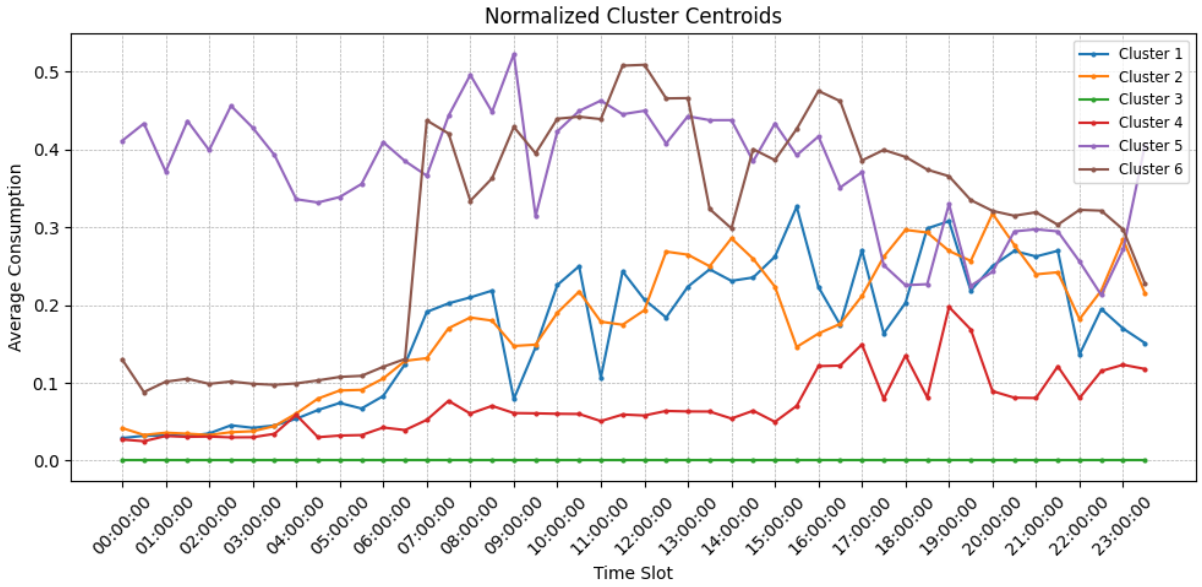


Figure 7: Κανονικοποιημένα Ενδεικτικά προφίλ φορτίου ομάδας

Επιπλέον, τα κανονικοποιημένα προφίλ κατανάλωσης όλων των νοικοκυριών μαζί με την μέση γραμμής κατανάλωσης κάθε ομάδας απεικονίζονται στο 8.

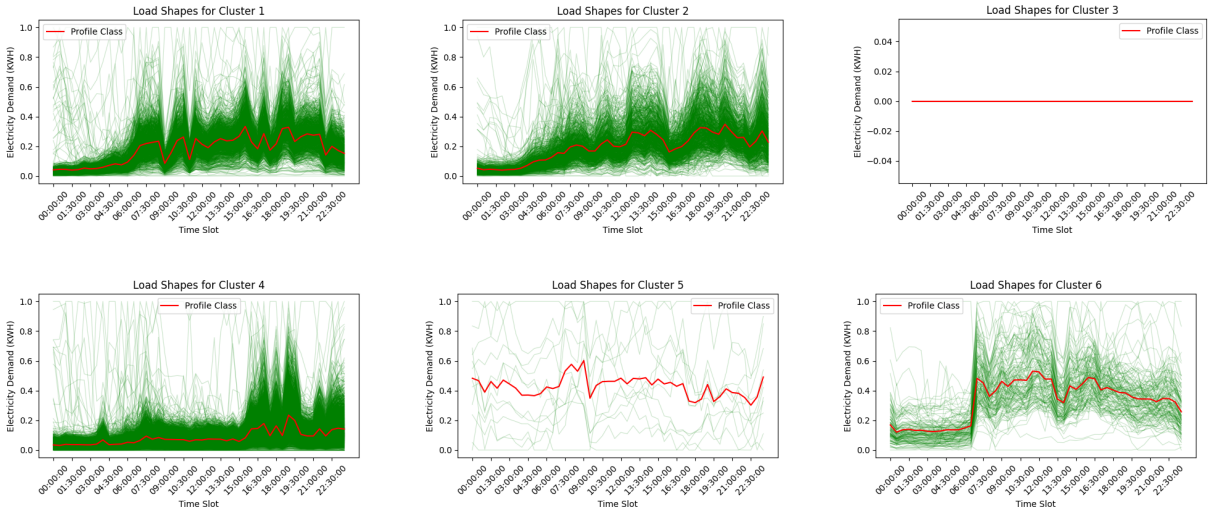


Figure 8: Προφίλ φορτίου όλων των νοικοκυριών κάθε ομάδας

Από τα κανονικοποιημένα προφίλ παρατηρούμε διαφορές ανάμεσα στις καθημερινές συνήθειες της κάθε ομάδας , πέραν του ύψους κατανάλωσης όπως π.χ η ομάδα 4 τις πρωινές ώρες έχει πολύ χαμηλή κατανάλωση (ενδεχομένως δεν βρίσκεται κάποιος στο σπίτι) και το απόγευμα υπάρχει μια σχετικά υψηλότερη κατανάλωση (ενδεχομένως επιστρέφουν σπίτι). Αντίθετα, στην ομάδα 2 παρατηρούμε κατανάλωση καθ'όλη την διάρκεια της ημέρας, γεγονός το οποίο μπορεί να υποδηλώνει ότι βρίσκονται σπίτι όλη την ημέρα.

Στην συνέχεια, οι μέθοδοι της εξηγήσιμης τεχνητής νοημοσύνης (XAI) μας βοήθησαν να κατανοήσουμε γιατί οι συστάδες ομαδοποιήθηκαν ξεχωριστά. Χρησιμοποιήσαμε τη βιβλιοθήκη SHAP (SHapley Additive exPlanations), που χρησιμοποιεί τη θεωρία παιγνίων για να εξηγήσει την έξοδο των μοντέλων μηχανικής μάθησης. Η βιβλιοθήκη SHAP μας βοήθησε να κατανοήσουμε τη λογική του αλγορίθμου Ensemble, δείχνοντας πώς κάθε χαρακτηριστικό επηρεάζει την ομαδοποίηση.

Τέλος, έγινε ένας φασής διαχωρισμός ανάμεσα στα ημερήσια καθημερινά προφίλ και τα προφίλ του σαββατοκύριακου.

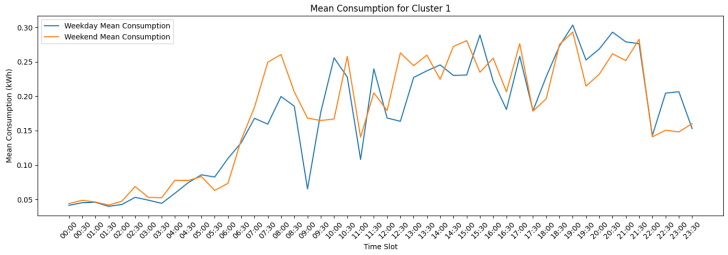


Figure 9: RLPs Σαββατοκύριακου και Εβδομαδιαίας ημέρας την ομάδα 1

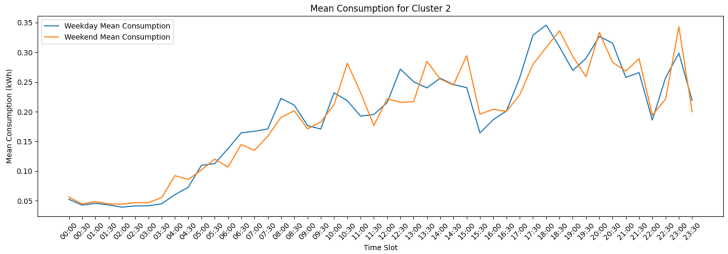


Figure 10: RLPs Σαββατοκύριακου και Εβδομαδιαίας ημέρας την ομάδα 2

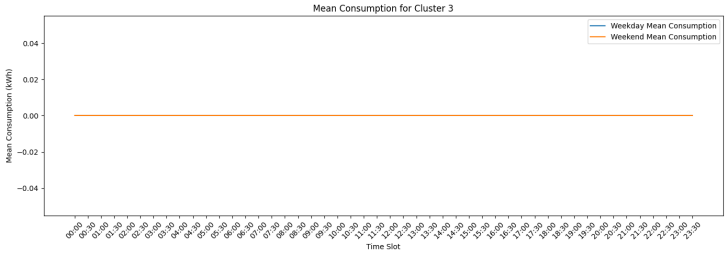


Figure 11: RLPs Σαββατοκύριακου και Εβδομαδιαίας ημέρας την ομάδα 3

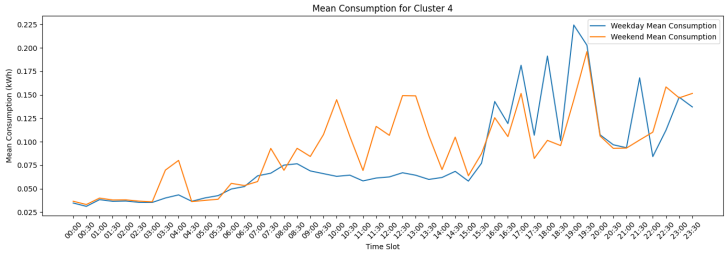


Figure 12: RLPs Σαββατοκύριακου και Εβδομαδιαίας ημέρας την ομάδα 4

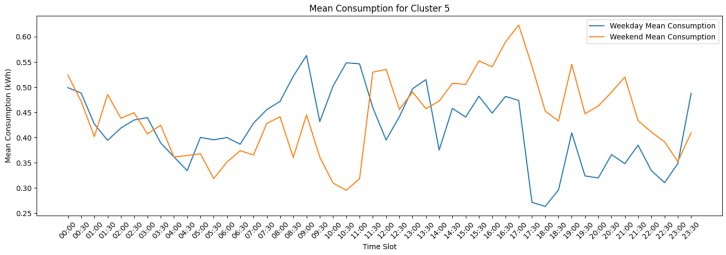


Figure 13: RLPs Σαββατοκύριακου και Εβδομαδιαίας ημέρας την ομάδα 5

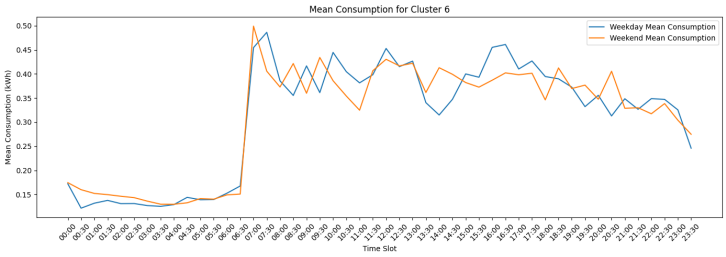


Figure 14: RLPs Σαββατοκύριακου και Εβδομαδιαίας ημέρας την ομάδα 6

Κεφάλαιο 6: Συμπεράσματα και μελλοντικές προοπτικές

Η ενδελεχής αξιολόγηση των αλγορίθμων ομαδοποίησης σε δεδομένα κατανάλωσης ενέργειας αποκάλυψε την ασυνέπεια στην απόδοση των αλγορίθμων με βάση τον επιλεγμένο αριθμό των ομάδων. Τα αποτελέσματα της μελέτης μας δείχνουν ότι ορισμένοι αλγόριθμοι, όπως ο K-means++ και η ομαδοποίηση Ensemble, επιδεικνύουν σταθερά ισχυρές επιδόσεις. Ωστόσο, η αποτελεσματικότητα άλλων αλγορίθμων ποικίλλει όταν λαμβάνονται υπόψη διαφορετικοί αριθμοί συστάδων. Αξιοσημείωτη είναι η αποτελεσματικότητα του Ensemble Clustering, καθώς επιτυγχάνει σταθερά από mooderate έως υψηλές κατατάξεις. Ωστόσο, δεν ξεπερνά πάντα τις επιδόσεις των κορυφαίων μεμονωμένων αλγορίθμων, όπως ο K-means++.

Η περαιτέρω έρευνά μας έχει κατηγοριοποιήσει επιτυχώς τους καταναλωτές σε διακριτές κατηγορίες προφίλ φορτίου, κυρίως με βάση τα συνολικά επίπεδα ενεργειακής κατανάλωσης - χαμηλά, μεσαία και υψηλά. Εντοπίσαμε ομάδες ακραίων τιμών που παρουσιάζουν μη φυσιολογικά πρότυπα κατανάλωσης.

Σε αντίθεση με τις αρχικές μας παρατηρήσεις, οι οποίες έδειχναν ότι οι συστάδες διαχωρίζονταν κυρίως με βάση τα επίπεδα κατανάλωσης, η διαδικασία κανονικοποίησης αποκάλυψε πιο σύνθετα μοτίβα κατανάλωσης εντός κάθε συστάδας. Τονίζεται η σημασία της κανονικοποίησης στις μελέτες κατανάλωσης ενέργειας, καθώς επιτρέπει την ακριβέστερη σύγκριση των προτύπων χρήσης μεταξύ διαφόρων ομάδων καταναλωτών.

Επιπλέον, σε περιπτώσεις όπου οι συστάδες επέδειξαν στενά συσχετιζόμενες συμπεριφορές, η XAI προσέφερε διευκρινίσεις σχετικά με τα διαφοροποιημένα χαρακτηριστικά που διαφοροποιούν τη μία συστάδα από την άλλη. Για παράδειγμα, αν και ορισμένες ομάδες εμφάνισαν ομοιότητες όσον αφορά την ποσότητα της κατανάλωσης, ο XAI αποκάλυψε διακριτά χαρακτηριστικά, όπως ο χρόνος της υψηλότερης χρήσης ή η κανονικότητα της κατανάλωσης, που τις διέκριναν μεταξύ τους.

Συνοψίζοντας, η παρούσα μελέτη υπογραμμίζει τη σημασία της επιλογής κατάλληλων

αλγορίθμων ομαδοποίησης που είναι ειδικά σχεδιασμένοι για τα μοναδικά χαρακτηριστικά των δεδομένων κατανάλωσης ενέργειας. Η σημασία του Ensemble Clustering στην ανάλυσή μας υπογραμμίζει την αποτελεσματικότητά τους στον εντοπισμό διακριτών τμημάτων καταναλωτών. Επιπλέον, η χρήση του XAI αποδείχθηκε ότι αποτελεί θεμελιώδες στοιχείο για τη βελτίωση της κατανοητότητας των αποτελεσμάτων της ομαδοποίησης, παρέχοντας βαθιά κατανόηση της ενεργειακής συμπεριφοράς των καταναλωτών. Η παρούσα μελέτη προωθεί την προσεκτική εφαρμογή της κανονικοποίησης και της επιλογής χαρακτηριστικών, υποστηριζόμενη από το XAI, για τη βελτίωση της ακρίβειας των στρατηγικών διαχείρισης ενέργειας και δέσμευσης πελατών. Αυτό υπερβαίνει την απλή κατηγοριοποίηση των καταναλωτών με βάση τα επίπεδα κατανάλωσής τους και, αντίθετα, στοχεύει στην αποκάλυψη των περίπλοκων μοτίβων που χαρακτηρίζουν κάθε τμήμα καταναλωτή.

Προοπτικές

Η παρούσα εργασία δημιουργεί ενδιαφέρουσες μελλοντικές προοπτικές έρευνας, οι κυριότερες εκ των οποίων παρουσιάζονται παρακάτω:

Συσχέτιση με κοινωνικοδημογραφικά χαρακτηριστικά: Η μελλοντική έρευνα θα πρέπει να εξετάσει τη συσχέτιση μεταξύ των προτύπων κατανάλωσης ενέργειας και των κοινωνικοδημογραφικών χαρακτηριστικών των νοικοκυριών. Η κατανόηση του τρόπου με τον οποίο παράγοντες όπως το μέγεθος του νοικοκυριού, το επίπεδο εισοδήματος, η γεωγραφική θέση και οι επιλογές του τρόπου ζωής επηρεάζουν τη χρήση ενέργειας θα μπορούσε να προσφέρει βαθύτερη κατανόηση της συμπεριφοράς των καταναλωτών. Η προσέγγιση αυτή μπορεί να βοηθήσει στην ανάπτυξη πιο εξατομικευμένων στρατηγικών εξοικονόμησης ενέργειας και στην προσαρμογή της επικοινωνίας σε διαφορετικά δημογραφικά τμήματα.

Εκτεταμένη εφαρμογή της επεξηγήσιμης τεχνητής νοημοσύνης: Μια πολλά υποσχόμενη κατεύθυνση είναι η περαιτέρω αξιοποίηση της Εξηγήσιμης ΤΝ. Τα εργαλεία της Εξηγήσιμης ΤΝ, όπως η βιβλιοθήκη SHAP (SHapley Additive exPlanations), έχουν ήδη αποδειχθεί πολύτιμα στην τρέχουσα ανάλυσή μας. Προχωρώντας προς τα εμπρός, μια βαθύτερη ενσωμάτωση των μεθοδολογιών Εξηγήσιμης ΤΝ θα μπορούσε να προσφέρει μεγαλύτερη κατανόηση των παραγόντων που επηρεάζουν τα αποτελέσματα της ομαδοποίησης. Αυτή η αυξημένη διαφάνεια στα μοντέλα ΤΝ δεν θα βοηθήσει μόνο στην κατανόηση των πολύπλοκων συμπεριφορών των καταναλωτών, αλλά και στην επικύρωση και βελτίωση των ίδιων των μοντέλων. Η δυνατότητα της Εξηγήσιμης ΤΝ να αποκαλύπτει περίπλοκες σχέσεις μέσα στα δεδομένα μπορεί να οδηγήσει σε πιο στοχευμένες και αποτελεσματικές στρατηγικές διαχείρισης της ενέργειας, προσαρμοσμένες στις συγκεκριμένες ανάγκες και τα πρότυπα των διαφόρων τμημάτων καταναλωτών.

Μείωση διαστάσεων και βελτιστοποίηση μεγέθους δεδομένων: Μια σημαντική πτυχή της διαχείρισης τεράστιων συνόλων δεδομένων είναι η αποτελεσματική μείωση της διαστατικότητας των δεδομένων χωρίς να διακυβεύεται η ακεραιότητα των υποκείμενων προτύπων. Τεχνικές όπως η ανάλυση κύριων συνιστωσών (PCA), η t-διανεμημένη στοχαστική ενσωμάτωση γειτόνων (t-SNE) και οι αυτοκωδικοποιητές έχουν δείξει υποσχέσεις σε άλλους τομείς. Η αξιολόγηση της αποτελεσματικότητας αυτών των μεθόδων στο πλαίσιο των δεδομένων κατανάλωσης ενέργειας θα μπορούσε να οδηγήσει σε πιο εκλεπτυσμένη ομαδοποίηση χωρίς το βάρος της επεξεργασίας και ανάλυσης μεγάλου όγκου δεδομένων.

Chapter 1

Introduction

1.1 Introduction

Energy consumption is still a major concern in today's world of rapid advancement, entwining sustainability requirements with the complexities of modern living. Unavoidably, as urbanization and population growth continue to rise, so does the need for energy. This elevated demand places tremendous strain on the energy infrastructure we currently have, making it imperative to create policies that encourage energy efficiency. The Demand-Response (DR) strategy is one such mechanism that is essential to intelligent energy systems and smart grids.

Demand-Response is an energy management strategy designed to narrow the gap between energy supply and consumer demand. Its objective is to motivate consumers, both individuals and industries, to alter their typical consumption habits in response to specific indicators, typically fluctuations in prices or incentives. By incorporating intelligent technologies like smart meters and IoT devices, DR involves more than just reducing or shifting energy usage during periods of high demand. It also involves utilizing real-time data to make dynamic and well-informed choices.

Nevertheless, the complex nature of DR extends beyond technology. DR revolves around human behavior, making it a complex and multifaceted challenge. Understanding how consumers, in their particular circumstances, react to DR signals is of utmost importance. The correlation between energy efficiency and DR lies in comprehending the underlying reasons behind people's energy consumption patterns in order to devise strategies for more prudent utilization.

The pursuit of energy efficiency is not only a requirement for maintaining environmental sustainability, but also a practical strategy for ensuring economic stability. Optimizing energy consumption minimizes superfluous energy expenses, diminishes rates of resource exhaustion, and alleviates environmental repercussions. Today, energy efficiency is not just a trendy term; it is a requirement, a duty, and, most importantly, a smart approach for achieving sustainable development.

An essential component of this strategy relies on the ability to classify and divide energy consumers. Gaining insight into their unique consumption patterns, behaviors, and preferences is not a simple or universal task. The concept of clustering arises as a powerful tool for analyzing the homogeneous groups within a diverse energy market. Through the process of classifying energy users according to similarities in their consumption patterns, it becomes feasible to customize DR strategies that align with the distinct behavioral tendencies of each group. Customized strategies have the capacity to generate increased levels of engagement in DR programs, thereby improving the overall effectiveness of the

system.

Clustering not only facilitates efficient strategy development but also enhances effective communication. Familiarizing yourself with your clusters is akin to understanding your target demographic. Transmitting DR signals or incentives to a clearly defined group enhances the probability of receiving a favorable reaction from consumers.

Ultimately, as we navigate the course of technological progress, the significance of human-centered methodologies in energy management becomes increasingly evident. This study thoroughly examines the merging of technology and human behavior, using clustering to strengthen the DR paradigm.

1.2 Thesis target and objectives

As previously mentioned, acquiring knowledge about consumption patterns is vital for efficient energy administration and long-term sustainability in our dynamic energy environment. This thesis explores the field of energy management, with a specific emphasis on improving DR strategies through the use of data analytics.

The primary goals of this research are to thoroughly examine household energy consumption data in order to reveal distinct patterns and behaviors. We will utilize a range of sophisticated clustering algorithms to divide consumers into distinct groups, allowing for the development of customized DR strategies. Upon implementing these techniques, we will assess their performance and conduct a comprehensive analysis of the outcomes to ensure they meet the requirements of practical applications and foster long-term viability.

This thesis seeks to accomplish these objectives in order to establish a connection between academic research and practical applications in the field of energy management. Ultimately, it aims to make a valuable contribution towards a future that is more energy-efficient.

1.3 Thesis contribution and value

Within the highly complex field of energy management and sustainable practices, comprehending the nuances of household energy consumption patterns is of utmost importance. This research makes significant progress in this field, providing diverse benefits as illustrated in the diagram 1.1.

Energy Management and DR Strategies
The results of this study are not just insignificant details in academia; they function as precise plans for the development and improvement of energy management and demand response strategies. Through comprehending and categorizing consumer behaviors, energy stakeholders can develop more knowledgeable and efficient strategies, guaranteeing a harmonious equilibrium between energy supply and demand.

Algorithmic Evaluation
This thesis entails a comprehensive comparative analysis of different clustering algorithms

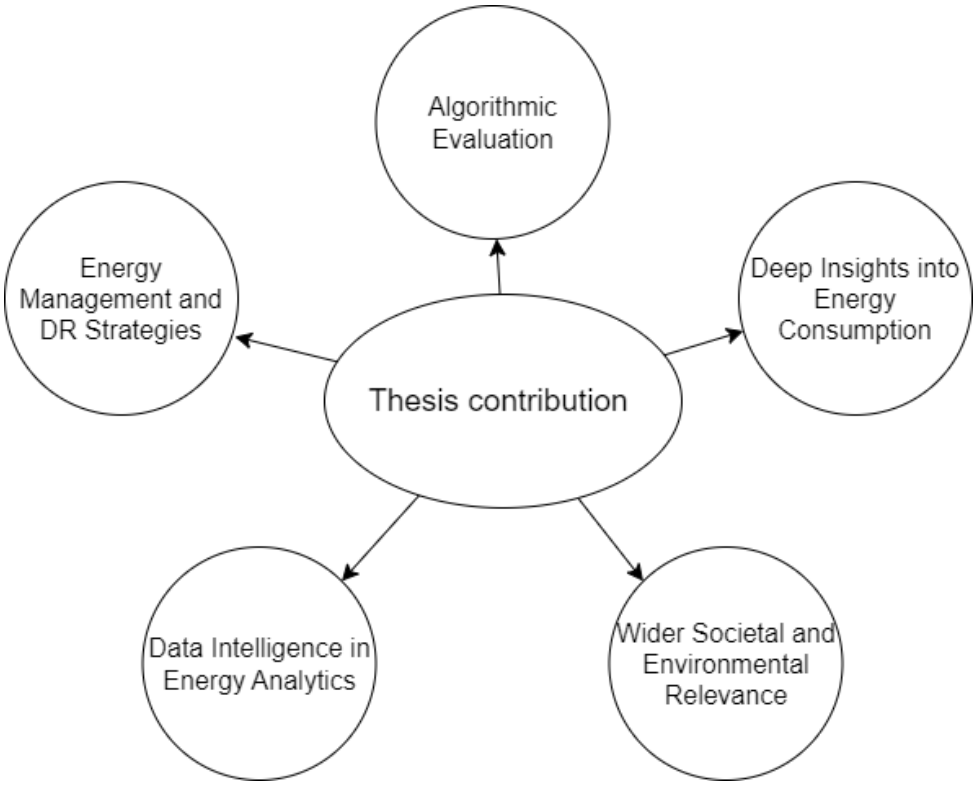


Figure 1.1: Thesis contribution

as they are applied to energy consumption data. By running these algorithms and meticulously comparing their performance, we offer a detailed evaluation of their effectiveness in deciphering energy consumption patterns. This approach not only emphasizes the advantages and drawbacks of each algorithm but also offers essential insights into which algorithms are most appropriate for particular types of energy data.

Deep Insights into Energy Consumption
This study goes beyond simple data collection by examining the analysis of energy consumption in the clusters. By classifying households according to their energy usage patterns, we are able to obtain valuable insights into various consumption behaviors. This systematic approach not only converts raw data into practical knowledge but also becomes an invaluable asset for stakeholders in energy management.

Data Intelligence in Energy Analytics
This thesis employs a unique methodology for data analytics, with a particular emphasis on extracting innovative characteristics from unprocessed data. The main focus of this study revolves around the rigorous procedure of data preprocessing, transformation, and sophisticated feature engineering. The research enhances the level of understanding that can be derived from energy consumption data by effectively identifying and extracting distinctive features.

Wider Societal and Environmental Relevance
This research carries substantial societal ramifications. By emphasizing the significance of energy efficiency and sustainability, it reinforces global endeavors to tackle climate change and advocate for environmental stewardship.

This research has a broader impact that goes beyond the boundaries of academic

circles. The thesis provides valuable guidance for industry stakeholders, ensuring a more sustainable future, as energy sustainability and efficiency become increasingly crucial in today's world.

1.4 Thesis structure

The thesis consists of 6 chapters and 1 appendix. Here is a brief description of their contents.

Chapter 1
In the 1st chapter of the thesis, a brief introduction to the problem was made, defining the main attributes and the historical background. The thesis target and objective were defined, as well as its contribution to the scientific society. Finally, there was a description of the thesis structure.

Chapter 2
In the 2nd chapter of the thesis, a comprehensive analysis of DR in residential energy systems is presented, covering its overview, benefits, challenges, and the essential role of consumer segmentation. The chapter begins with an introduction to DR, followed by a detailed discussion of its advantages and the specific challenges encountered in residential contexts. It emphasizes the necessity of consumer segmentation in enhancing the efficacy of DR programs and explores the integration of Machine Learning to refine these segmentation strategies. The chapter concludes by summarizing these key aspects.

Chapter 3
In Chapter 3, we present a literature review segmented into four phases. The first phase, Pre-Clustering, examines existing literature on data preparation and feature extraction. The second phase, Clustering Methodologies, reviews the various clustering algorithms used in literature. This is followed by the third phase, Performance Evaluation Metrics, which discusses the metrics used in evaluating the algorithms.The final phase, Post-Clustering, explores literature on the interpretation of clustering results.

Chapter 4
In the 4th chapter of the thesis, we present the proposed methodology.The chapter commences with a synopsis of the data set, followed by the procedures included in preprocessing and feature extraction. Next, the clustering algorithms that will be utilized are presented. The chapter also presents the evaluation metrics that will be utilized to assess the performance of the algorithms.

Chapter 5
In the 5th chapter of the thesis, we present the he results derived from the application of various clustering algorithms.This chapter presents a comprehensive analysis of these algorithms, assessing their performance across a range of tests. After conducting a thorough evaluation, the algorithm that demonstrates superior performance is chosen for a detailed analysis. The selected algorithm is subsequently employed to identify and extract indicative load patterns, offering a comprehensive description of each consumer cluster.

Chapter 6
Finally, in the 6th chapter of the thesis, a conclusion of the whole project is made and the future prospects are discussed.

Appendix A

In Appendix A, we provide a comprehensive collection of results related to Chapter 5 that were not included in the main text due to length considerations. This appendix serves as an extensive repository of additional data and analyses, offering a deeper and more detailed insight into the outcomes of the clustering algorithms discussed in Chapter 5.

Chapter 2

Problem Setting

2.1 Introduction

The Demand-Response (DR) mechanism is a key component of contemporary energy systems in the ever-changing field of energy management. Demand-Response is the deliberate adjustment of energy consumption patterns by end-users in reaction to signals from energy suppliers or grid operators[1]. Traditionally, the energy grid operated primarily in a unidirectional manner, where electricity was generated by power plants and transmitted to consumers. The integration of intermittent renewable energy sources, such as solar and wind, along with the growing variability of consumer demands, has resulted in a heightened complexity in the management of electricity supply and demand[2].

Given the increasing energy demands and the pressing need to incorporate sustainable energy practices, the effectiveness of the DR system becomes crucial. It offers a flexible solution that enables the adaptation of energy requirements to match the available supply. This promotes energy efficiency and helps to decrease expenses related to high-demand periods[3].

Nevertheless, the practical application of DR has its share of its difficulties. A major challenge is comprehending and forecasting consumer behavior, which exhibits significant variations among different demographic groups[4]. The presence of variability diminishes the effectiveness of a one-size-fits-all approach, thereby emphasizing the necessity for a more customized strategy.

In this chapter, we will explore in detail the intricacies of the DR challenge, emphasizing the need for consumer segmentation, the potential role of machine learning in addressing these challenges, and laying the foundation for the discussions and analyses that follow in this thesis.

2.2 Overview of DR

DR has steadily grown in prominence, transitioning from a niche concept to a central pillar in the energy sector. It operates on the basic principle of aligning energy demand with supply, but there's more depth to its functionality and application than meets the eye.

Definition and Fundamental Principles

DR serves as both a proactive and reactive approach in energy management, aiming to harmonize consumption with generation. What sets DR apart is its capacity to influence consumers - be it households, industries, or businesses - to modify their energy use in response to real-time electricity prices or specific external cues[1]. Typically, these cues come from grid operators or utilities and indicate various scenarios, like a supply shortage, abundant renewable energy output, or grid congestion.

Leveraging advanced technologies, an effective DR system incorporates real-time metering, predictive analytics, and automated controls[2]. This integration permits almost immediate feedback and adjustments, fortifying a flexible and adaptive energy ecosystem. Additionally, there are diverse DR programs: some offer financial incentives to consumers who adjust their energy use, while others have fluctuating energy prices based on time or demand levels.

In essence, DR symbolizes a mutual endeavor between energy providers and consumers to create an energy framework that's more sustainable and adaptable. As we move forward, we will dissect the various types of DR strategies, each catering to unique scenarios and requirements.

Types of DR Programs

The diverse classification of DR programs underlines their strategic importance in shaping electricity consumption behaviors. These programs are primarily segregated into two major categories: incentive-based DR programs and price-based DR programs. Figure 2.1 depicts the classification of DR programs.

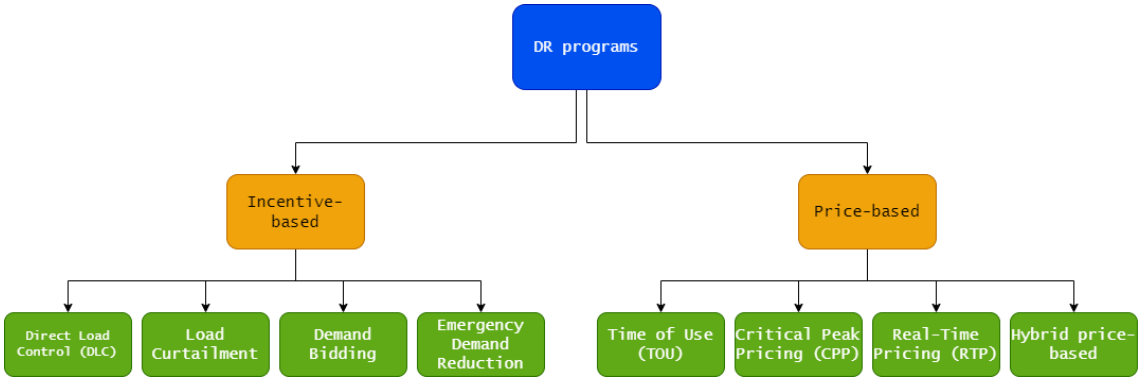


Figure 2.1: Classification of DR programs

Incentive-based DR Programs

Incentive-based programs reward consumers with incentives for modifying their electricity usage in accordance with the needs of the supply-side. By providing tangible rewards, these programs create a direct link between consumer behavior and the operational needs of the energy grid. For example, during times of peak demand or specific events, utilities may need to ease the load on the grid. With incentive-based programs,

consumers can volunteer or sign up to reduce their energy consumption during these critical periods[1]. In return, they receive financial incentives or other benefits, making it a win-win situation: the grid experiences reduced strain, and consumers benefit financially. Different types of incentive-based DR programs are discussed and detailed below according to [5],[2],[1].

- **Direct Load Control (DLC) Programs**: Here, certain consumers or appliances are pre-enrolled, granting the utility permission to shut them down or cycle them as per requirements, primarily during peak demand or specific events. Participating consumers receive incentives for their cooperation.
- **Load Curtailment Programs**: In these setups, enrolled consumers receive incentives for reducing their electricity usage based on the utility's needs. Notably, if these consumers don't adhere to the program, they may face significant penalties.
- **Demand Bidding Programs**: This kind of program is typically available for large-scale consumers with consumptions surpassing 1 MW. In situations of high demand or contingencies, these consumers can propose to reduce a part of their consumption for a specific bid price.
- **Emergency Demand Reduction Programs**: This program activates during severe system contingencies. Participating consumers are generously incentivized to decrease their electricity usage, bolstering the overall reliability of the power system. Penalties are not imposed for noncompliance.

Price-based DR Programs

In price-based DR programs, the fluctuations in electricity tariffs serve as an implicit signal to consumers about the current state of the grid. When electricity is abundant and the grid is under low stress, tariffs may drop, encouraging consumption. On the contrary, during peak times or when supply is limited, tariffs rise, nudging consumers to either curtail their use or shift it to a more favorable time. This dynamic electricity pricing not only reflects the real-time costs associated with generating and distributing electricity but also encourages more strategic energy consumption patterns among users.The primary forms of price-based DR programs are mentioned below.

- **Time of Use (TOU) Pricing**: Electricity pricing in this scheme hinges on specific intervals within the day. Typically, the day gets divided into peak, mid-peak, and off-peak intervals. Consumption during peak times incurs higher charges, nudging consumers to shift their usage to off-peak hours[5].
- **Critical Peak Pricing (CPP)**: Similar to TOU, CPP increases the electricity rate dramatically during times when the power system's reliability is at stake. These high rates are usually in effect for a few hours annually, aiding in maintaining the power system's stability[6].
- **Real-Time Pricing (RTP)**: This dynamic pricing structure alters the electricity tariffs, often on an hourly basis, mirroring the ongoing wholesale electricity market prices. Consumers typically receive notifications about these prices a day or an hour in advance. As real-time pricing grows in popularity, especially with the rise of smart

homes, some cases require a price prediction module, particularly when prices aren't declared on a day-ahead basis[5].

In addition to the conventional TOU structures, advancements in DR strategies have led to the inception of hybrid price-based programs, specifically tailored for residential micro-grids. These hybrid programs integrate both fixed and dynamic pricing mechanisms. They are designed to synergize the benefits of stable pricing structures like TOU with the flexibility of real-time pricing (RTP). This fusion allows for more efficient electricity consumption management within micro-grids as stated in [7],

2.3 Benefits of DR

Within the transition towards more intelligent and resilient energy systems, DR arises as a crucial innovation, improving grid efficiency and empowering consumers. DR enables a more agile and environmentally-friendly electricity grid by adjusting energy usage in accordance with grid conditions. The following paragraphs highlight the diverse advantages of DR:

Integration of Renewable Resources and Grid Stability

DR plays a crucial role in tackling the difficulties related to incorporating renewable energy resources (RES) into the power grid, specifically because of their intermittent and fluctuating characteristics. As the European Union (EU) begins its ambitious pursuit of a sustainable and resilient energy future, as outlined in the European Green Deal[8], DR emerges as a crucial facilitator. The EU's dedication to achieving climate neutrality by 2050 requires a swift escalation in the adoption and integration of RES, which will fundamentally reshape its energy system.Dynamic DR facilitates the real-time modification of electricity consumption, thereby mitigating the variability of renewable energy generation and guaranteeing the stability and dependability of the power grid. This adaptation is vital during periods of elevated renewable energy production, such as when there is abundant sunlight or strong winds, as well as during periods of limited renewable energy generation. DR can optimize the utilization of the current grid infrastructure, minimizing the requirement for expensive upgrades and improving the grid's ability to withstand fluctuations[2]. Furthermore, through its support for a less centralized energy system, DR promotes the use of small-scale renewable generators like solar panels on homes or wind turbines in local communities.

Benefits for Consumers

Both consumers and businesses participating in DR programs directly reap financial advantages. To achieve substantial reductions in their electricity bills, individuals can optimize their energy consumption by aligning it with periods of lower energy prices, such as off-peak periods or times when renewable energy production is at its peak[2]. These modifications not only lead to financial savings but also encourage a more environmentally friendly and accountable energy usage pattern. Moreover, participants acquire enhanced visibility and authority over their energy consumption, enabling them to make

more knowledgeable choices and cultivating a feeling of empowerment and engagement in the wider energy domain. The improved regulation and the possibility of reducing expenses contribute to higher levels of customer contentment and can also bolster the public's perception of utility providers.

Benefits for the Energy Market

DR programs not only facilitate the integration of renewable resources and provide direct benefits to consumers, but they also play a pivotal role in enhancing the overall efficiency and competitiveness of the energy market. DR aids in mitigating price fluctuations and lowering electricity expenses throughout the market by promoting a dynamic equilibrium between supply and demand[9]. The stabilization of this market is especially crucial in areas with a significant prevalence of renewable energy sources.

Moreover, DR programs encourage innovation in the energy industry, promoting the advancement of intelligent power grids, sophisticated metering infrastructure, and other technologies that improve grid control and effectiveness [1]. These technological advancements contribute to the development of energy markets that are more adaptable and responsive, capable of accommodating the increasing proportion of renewable energy in the energy mix.

DR not only improves market efficiency, but also fosters transparency and equity in the energy market. By facilitating the involvement of consumers in the decision-making process, DR ensures that all participants in the market, as referenced by [1], have the opportunity to influence market outcomes. The process of democratizing the energy market not only results in fairer pricing and improved access to energy, but also promotes responsible energy usage and encourages investments in clean energy technologies.

2.4 Challenges of Residential DR

The concept of residential DR signifies a fundamental change in the manner in which electricity is utilized and controlled within households. It involves modifying consumer electricity consumption in accordance with supply conditions, such as price signals or power grid requirements. Although residential DR offers notable advantages, such as improved energy efficiency, decreased peak demand, and reduced electricity expenses, its execution encounters various obstacles.

Technological Challenges

The main technological obstacles in residential DR programs revolve around the integration and efficient utilization of sophisticated technologies. This encompasses the implementation of intelligent meters and Internet of Things (IoT) devices, which are essential for real-time monitoring and regulation of energy consumption. Nevertheless, the extensive implementation and maintenance of these devices in various residential areas present notable difficulties.

An additional crucial factor is the administration and examination of the substantial quantities of data produced by these programs. Processing this data and generating actionable insights necessitates the use of advanced data analytics and artificial intelligence algorithms. Moreover, the success of DR programs heavily depends on reliable and secure communication networks to transmit data and DR signals between utilities and consumers. The complexity is increased by the need to maintain consistent communication in different geographical and infrastructural environments, as well as ensuring compatibility between various devices and systems.

Moreover, it is imperative to tackle the issues related to cybersecurity and privacy. Ensuring the security of consumer data and energy usage patterns against cyberthreats, as well as addressing privacy concerns associated with the extensive monitoring of household energy usage, are crucial for upholding consumer confidence and program reliability.

Behavioral Challenges

On the behavioral side, the heterogeneity of residential consumers in terms of their energy needs and their responsiveness to DR signals presents a complex challenge. Louf and Barthelemy (2016) emphasize that the spatial arrangement of resources and socioeconomic factors plays a crucial role in determining the adoption and efficacy of DR strategies[10]. Changing consumer behavior to accommodate DR (DR) strategies necessitates not just knowledge and instruction, but also rewards that are in line with their individual and financial motivations.

Consumers' inclination to engage in DR (DR) programs can be impacted by multiple factors, such as their comprehension of the program's advantages, apprehensions regarding comfort and convenience, and confidence in the utility providers. To overcome these behavioral barriers, it is necessary to develop DR (DR) programs that are both financially appealing and user-friendly.

Furthermore, the adoption and effectiveness of DR (DR) strategies are significantly influenced by socioeconomic factors and the spatial distribution of resources. This encompasses the presence of technology, the financial condition of consumers, and their ability to obtain information regarding energy management. Customizing DR (DR) programs to cater to the distinct requirements and situations of various consumer segments is crucial for achieving widespread acceptance and efficacy.

Forecasting and Scenario Planning

Comprehending the critical importance of forecasting in DR (DR) programs is essential. DR programs, which seek to modify energy consumption during periods of high demand, heavily depend on precise forecasts of energy usage. Precision is essential for optimizing the distribution of energy and guaranteeing reliability in the energy grid. Forecasting allows energy providers to predict increases in demand and modify supply accordingly, making it an essential tool in the management and success of DR (DR) initiatives.

Predicting residential energy usage with precision poses considerable difficulties. Zhang and Zhang (2019) underscore the challenges, emphasizing the capriciousness of household energy consumption patterns[11]. These patterns are influenced by various

factors, such as individual behaviors, weather conditions, and the implementation of energy-efficient technologies.

A major challenge in this forecasting project arises from the intricate and diverse characteristics of residential energy usage. Residential energy consumption is subject to various personal habits and preferences, which introduces unpredictability, unlike commercial or industrial energy usage. The diversity of daily routines, appliance usage, and home-based activities across households results in substantial variations in energy consumption patterns. The inherent unpredictability presents a significant obstacle in developing precise and dependable energy predictions for the successful implementation of DR programs.

2.5 Necessity of Consumer Segmentation in DR

Consumer segmentation in DR (DR) involves categorizing a large consumer population into smaller, more similar groups based on specific factors like energy consumption patterns, demographic traits, lifestyle preferences, and receptiveness to DR signals. This approach enables a more customized and effective implementation of DR strategies. Through comprehending the distinct requirements and actions of various segments, utilities and energy providers can devise DR (DR) programs that are more enticing and efficient for each demographic.For example, a group of consumers that uses a lot of energy during busy times of the day may be more open to pricing models that change based on demand. On the other hand, another group that is environmentally conscious may be more likely to participate in programs that focus on the environmental advantages. Segmentation allows for the creation of focused communications and incentives, enhancing the consumer-centricity and effectiveness of DR initiatives.

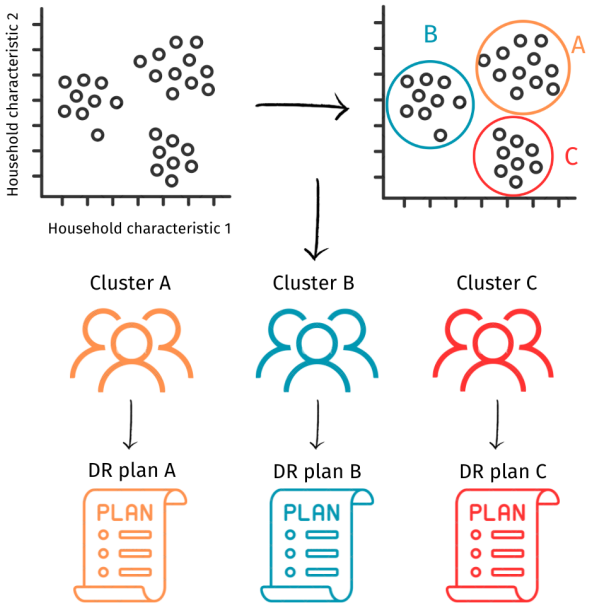


Figure 2.2: Consumer Segmentation in DR

A key example of this in practice is a Smart Grid pilot project[12], which provided

a clear demonstration of the importance of consumer segmentation in DR. The project utilized advanced metering infrastructure to gather data on consumer energy usage, enabling effective segmentation. The outcomes showed notable differences in how various consumer segments responded to DR strategies, emphasizing that consumer-specific approaches significantly improve participation rates and energy savings.

This approach brings multiple benefits:

1. **Increased Participation and Engagement:** By understanding and addressing the unique motivations of different segments, utilities can design DR programs that are more appealing to each group. This personalized approach often results in higher participation rates.
2. **Efficiency in Energy Management:** Segmentation allows for more precise targeting of energy-saving efforts. For example, segments with high energy use during peak times can be targeted with specific incentives to reduce consumption, leading to more effective load balancing.
3. **Enhanced Customer Satisfaction:** When consumers feel that their specific needs and preferences are being considered, their satisfaction with utility providers increases. This positive relationship fosters trust and long-term engagement with DR programs.
4. **Cost-Effectiveness:** By focusing resources and strategies on specific segments, utilities can achieve more with less, reducing the overall cost of implementing DR programs.

2.6 Machine Learning in Consumer Segmentation

ML is a transformative branch of artificial intelligence that empowers systems to learn and improve from experience without explicit programming [13]. It stands as a pivotal tool in contemporary data analysis, facilitating the automation of decision-making processes and the extraction of meaningful insights from voluminous datasets.

ML is broadly categorized into three principal types:

1. **Supervised Learning:** This approach involves training models on a labeled dataset, where the desired output is known. The model learns to map input data to the corresponding output, and it finds extensive use in applications like spam detection and image recognition [14].
2. **Unsupervised Learning:** Here, models are trained on unlabeled data, aiming to discover hidden patterns or intrinsic structures within the data. This category is crucial in exploratory data analysis, clustering, and dimensionality reduction [15].
3. **Reinforcement Learning:** This type focuses on how agents should act in an environment to maximize a cumulative reward. It has applications in diverse domains, including robotics and gaming [16].

Our primary focus will be on unsupervised learning, given its relevance to our objectives.

2.6.1 Unsupervised Learning

In unsupervised learning, the algorithm receives a set of inputs $\{x_1, x_2, ..., x_n\}$ without any corresponding output labels. The objective is to uncover the underlying structure of the data, grouping similar data together, or representing the data in a more informative manner[17]. Unsupervised learning encompasses a variety of methods, each serving specific purposes:

- **Clustering** - This process mathematically involves partitioning the dataset into distinct groups, or clusters, ensuring that data points within the same cluster exhibit greater similarity to each other than to those in other clusters. Similarity is often quantified using measures like Euclidean distance [18].
- **Dimensionality Reduction** (e.g., PCA, t-SNE) - for reducing the number of variables in the dataset while retaining its essential features [19].
- **Association Rule Learning** (e.g., Apriori, Eclat) - useful in market basket analysis to find common itemsets [20].

Unsupervised learning Clustering Algorithms

Clustering is one of the most popular unsupervised machine learning approaches. There are several types of unsupervised learning algorithms that are used for clustering, which include exclusive, overlapping, hierarchical, and probabilistic.

- **Exclusive clustering:** Data is grouped in a way where a single data point can only exist in one cluster. This is also referred to as "hard" clustering. A common example of exclusive clustering is the K-means clustering algorithm, which partitions data points into a user-defined number K of clusters.
- **Overlapping clustering:** Data is grouped in a way where a single data point can exist in two or more clusters with different degrees of membership. This is also referred to as "soft" clustering. The difference of exclusive and overlapping is illustrated in 2.3.

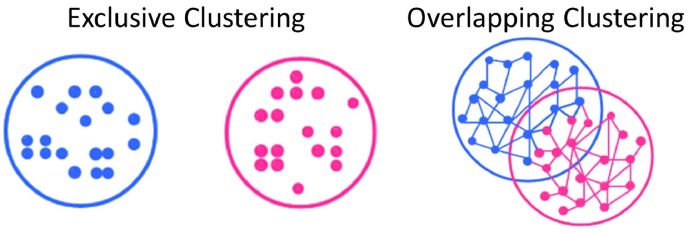


Figure 2.3: Graphical representation of overlapping and exclusive clustering [21].

- **Hierarchical clustering:** Data is divided into distinct clusters based on similarities, which are then repeatedly merged and organized based on their hierarchical relationships. There are two main types of hierarchical clustering as seen in 2.4: agglomerative and divisive clustering. This method is also referred to as HAC (hierarchical cluster analysis).

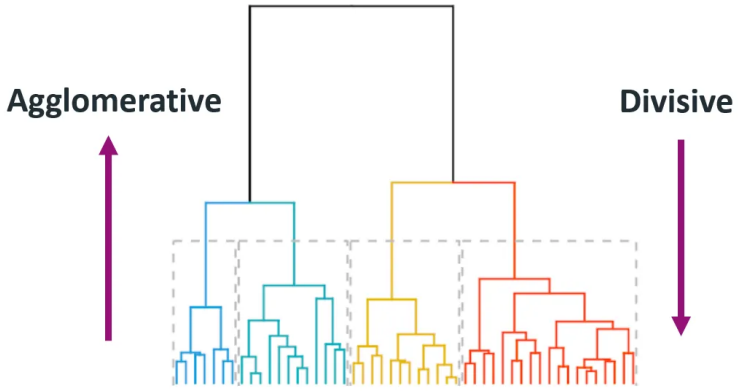


Figure 2.4: Dendrogram displaying the two main hierarchical clustering techniques

- **Probabilistic clustering:** Data is grouped into clusters based on the probability of each data point belonging to each cluster. This approach differs from the other methods, which group data points based on their similarities to others in a cluster. The Gaussian Mixture Model (GMM) is the one of the most commonly used probabilistic clustering methods and a visual representation of how GMMs work can be seen in]2.5.

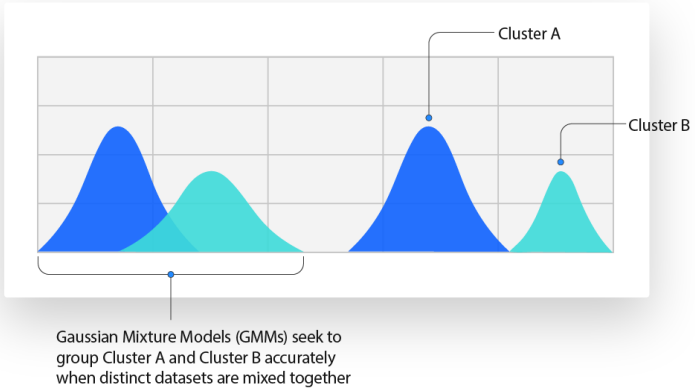


Figure 2.5: Gaussian Mixture model illustration example [22].

According to [23], a clear distinction is made between modern and traditional clustering algorithms. Below, Tables 2.1 and 2.2 summarize the categorization and typical algorithms associated with each.

Table 2.1: Traditional Clustering Algorithms

Category	**Typical Algorithm**
Based on partition	K-means, K-medoids, PAM, CLARA, CLARANS
Based on hierarchy	BIRCH, CURE, ROCK, Chameleon
Based on fuzzy theory	FCM, FCS, MM
Based on distribution	DBCLASD, GMM
Based on density	DBSCAN, OPTICS, Mean-shift
Based on graph theory	CLICK, MST
Based on grid	STING, CLIQUE
Based on fractal theory	FC
Based on model	COBWEB, GMM, SOM, ART

Table 2.2: Modern Clustering Algorithms

Category	**Typical Algorithm**
Based on kernel	Kernel K-means, Kernel SOM, Kernel FCM, SVC, MMC, MKC
Based on ensemble	CSPA, HGPA, MCLA, VM, HCE, LAC, WPCK, sCSPA , sMCLA, sHBGPA
Based on swarm intelligence	ACO_based(LF), PSO_based, SFLA_based, ABC_based
Based on quantum theory	QC, DQC
Based on spectral graph theory	SM, NJW
Based on affinity propagation	AP
Based on density and distance	DD
For spatial data	DBSCAN, STING, Wavecluster, CLARANS
For data stream	STREAM, CluStream, HPStream, DenStream
For large-scale data	K-means, BIRCH, CLARA, CURE, DBSCAN, DENCLUE, Wavecluster, FC

Applications in Consumer Segmentation

In the context of energy consumer segmentation, particularly for DR programs, unsupervised learning algorithms like K-means play a crucial role. These algorithms are adept at segmenting energy consumers based on their usage patterns, peak demand times, and consumption behaviors. By analyzing factors such as time-of-use data, seasonal consumption variations, and load profiles, unsupervised learning helps utilities identify distinct consumer groups with similar energy demands and responsiveness to DR initiatives. This segmentation is vital for optimizing DR strategies, as it allows energy providers to tailor their approaches to specific segments, ensuring more effective management of energy demand and supply. It also aids in enhancing the efficiency of energy distribution systems and in promoting more sustainable energy consumption practices.

2.7 Conclusion

Throughout this chapter, we have examined how DR plays a crucial role in the modernization of energy systems. It provides significant advantages for grid stability, consumers, and the overall energy market. DR strategies are uniquely positioned to manage the variability introduced by the integration of renewable resources, which is beneficial for sustainability. DR programs offer consumers the chance to save money and enable them to actively participate in managing their energy usage. DR plays a significant role in the energy market by improving efficiency, facilitating the integration of renewable energy sources, and maintaining price stability.

Nevertheless, the implementation of residential DR encounters certain difficulties. Overcoming technological and behavioral challenges, actively involving the community, addressing perception issues, and dealing with the intricacies of predicting and planning for different scenarios are all major obstacles that require careful and precise navigation. The importance of consumer segmentation becomes evident in this context. Energy providers can customize DR strategies to effectively address the varied requirements of their customer base by analyzing and classifying consumers according to their energy load patterns and socio-demographic characteristics.

ML, particularly unsupervised learning methods such as clustering, play a leading role in this segmentation endeavor. Through the utilization of ML, utilities can analyze intricate datasets to discover patterns and clusters that provide insights for more sophisticated and efficient DR strategies. This not only improves the accuracy of DR programs but also guarantees their ability to adjust and stay pertinent as consumer behaviors and the energy landscape continue to change.

Ultimately, the combination of DR and ML-driven consumer segmentation holds great potential for creating a more adaptive, streamlined, and consumer-focused energy future. As we finish this chapter, we are about to embark on a more extensive investigation into the capabilities of different clustering algorithms. In the upcoming chapters, we will analyze these algorithms by implementing them on actual datasets to determine their advantages, constraints, and the distinct perspectives they can offer. Additionally, we will investigate the extensive range of existing academic studies. This upcoming analysis is not just an academic exercise; it is a crucial step towards realizing the potential of machine learning (ML) to completely transform the energy sector. It will enable the development of a more intelligent and adaptable grid that can effectively address the challenges of the 21st century.

Chapter 3

Related Work

3.1 Introduction

This chapter provides a comprehensive literature review focused on machine learning clustering techniques employed for the segmentation of consumer energy usage, which is a fundamental aspect of our thesis. We analyze different clustering techniques that are essential for comprehending the segments of energy consumers, and discuss the difficulties encountered in this ever-changing domain. The aim of our project is to establish connections between various methodologies, assess their efficacy, and investigate novel research opportunities.

The chapter has been structured to systematically explore the complicated landscape of clustering algorithms in energy consumption. In order to offer a more comprehensive analysis, we partitioned our literature review into distinct phases that align with the entirety of the clustering problem. This framework enables a concentrated investigation and direct contrast of various clustering methodologies and their utilization in partitioning energy consumers.

3.2 Approach to Literature Review

In compiling the literature for this thesis, a systematic and methodical approach was employed to ensure a comprehensive understanding of clustering techniques in electricity consumer segmentation. The process was guided by the following steps:

1. **Search Methodology**: The literature was identified through targeted searches in academic databases and journals. Keywords such as "clustering techniques," "electricity consumer segmentation," "energy consumption segmentation," and names of specific algorithms like "K-Means," "DBSCAN" were used. This approach ensured a diverse yet focused collection of relevant scholarly work.

2. **Criteria for Inclusion:** The papers included in this review were selected based on their direct relevance to clustering techniques in energy consumption segmentation. Special attention was given to studies that demonstrated significant citation counts, indicating their impact and recognition in the field.

3. **Analysis Approach:** The selected papers will be analyzed with the aim of comparing methodologies, results, and their applicability to the problem of electricity consumer segmentation. This analysis will also explore the implications of these

studies on the current research, identifying gaps and opportunities for further investigation.

In addition to the above methodologies, we employed InfraNodus [24], a graph visualization tool, to create a network representation of our literature library. This graphical depiction allowed for an intuitive understanding of the relationships and clusters within the selected papers, facilitating an efficient filtering and analysis process. The use of InfraNodus underscores our commitment to leveraging advanced tools for enhancing the effectiveness and depth of our literature review. The resulting network provided for our research can be found in Figure 3.1.

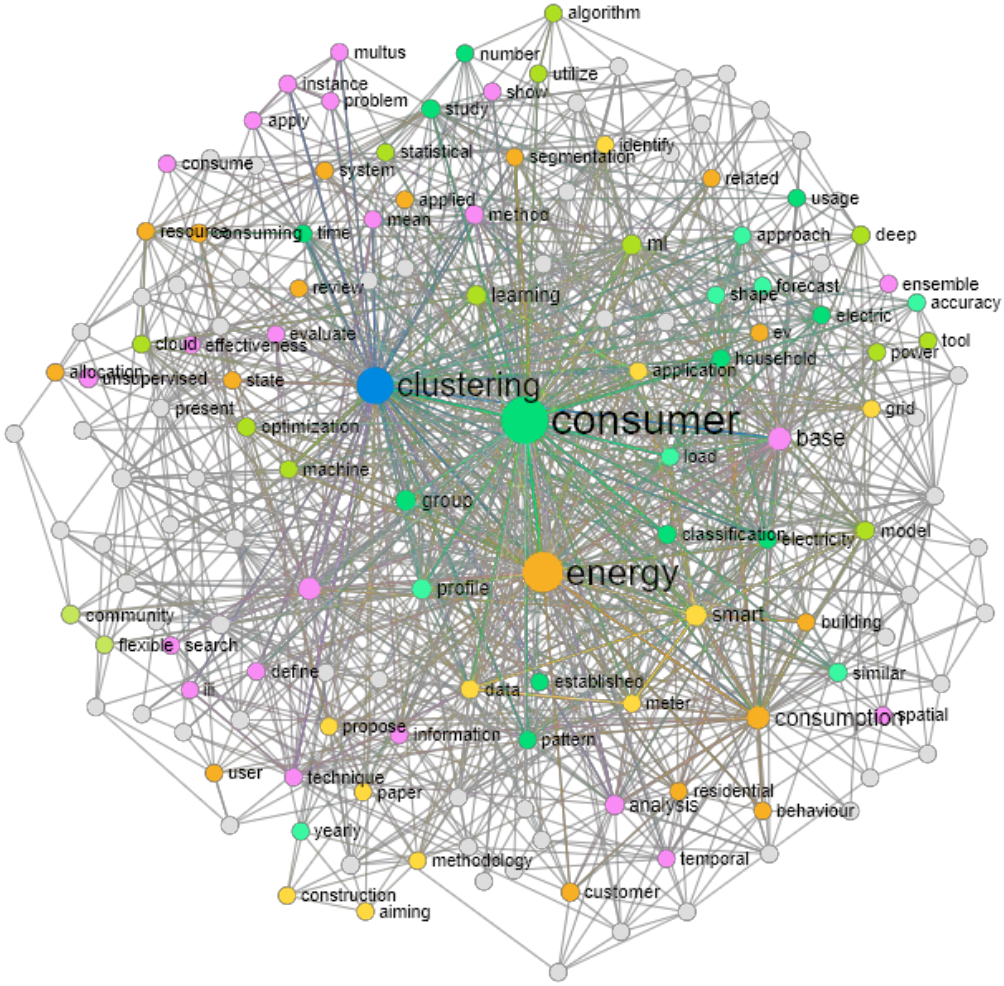


Figure 3.1: Network visualization of the literature library using InfraNodus

This structured approach to literature review guarantees a thorough exploration of the field, providing a solid foundation for the research presented in this thesis.

3.3 Phase I: Pre-Clustering

Prior to the main clustering process, the pre-clustering phase plays a vital role in the classification of electricity consumers. This phase is crucial for data preparation,

gaining insights into initial patterns, and establishing a foundation for efficient clustering. Multiple methodologies were employed in the existing literature to address this particular stage, which we will now present.

Data pre-processing

In the context of data normalization and handling missing values, Yilmaz et al. in [25] address these aspects effectively. They adopt a strategy where datasets with more than five consecutive days of missing data are not filled, preserving the dataset's integrity. Outliers are also carefully identified and removed, such as households with exceptionally high or low electricity consumption. In [26] the authors specifically ignore daily usage data with very small sums in the dictionary generation process. This decision is based on the rationale that such small usage patterns tend to be irregular and can distort the representative shapes in the dictionary after normalization. The study also sets a practical threshold for considering load patterns: those with total energy consumption lower than 3 kWh (which is less than the 6% quantile) are excluded from analysis.

Clustering techniques can be applied to customer load profiles using various types of data, as well-roundly explained in [27].The methods include:

- **Raw consumption data**, which are the actual time series data of customer electricity usage.
- **User-defined features** based on the characteristics of the load shapes and the specific application at hand.
- **Features extracted** from load shapes using techniques such as frequency domain analysis, which unveil intrinsic patterns like periodicity.
- **A reduced data set** obtained from the original data through dimensionality reduction techniques like principal component analysis (PCA).

Feature extraction

An important challenge faced when implementing machine learning is the phenomenon known as the **"curse of dimensionality"**. Several algorithms that exhibit good performance in low dimensions become unmanageable when applied to high-dimensional input[28]. Put simply, as the number of features increases, clustering becomes significantly more challenging.

The paper "Overview and Performance Assessment of the Clustering Methods for Electrical Load Pattern Grouping" [29] underscores the importance of selecting appropriate features for clustering in the field of feature engineering. While Räsänen and Kolehmainen in [30] elaborates on this, detailing that customer load profiles can be efficiently represented using a concise set of seven features, namely *mean, standard deviation, skewness, kurtosis, chaos, energy, and periodicity.*

In addressing the challenges, the work by Haben et al. in [31] offers a sophisticated approach to mitigate the curse of dimensionality. The study strategically selects attributes

that capture the essence of customer energy behavior, focusing on four distinct time periods: the overnight period from 10:30 P.M. to 6:30 A.M., the breakfast period from 6:30 A.M. to 9:00 A.M., the daytime from 9:00 A.M. to 3:30 P.M., and the evening from 3:30 P.M. to 10:30 P.M. For each customer, the relative average power across these periods is calculated, along with a mean relative standard deviation to gauge variability. Seasonal and day-type variations are encapsulated by a seasonal score, reflecting the absolute differences between summer and winter mean powers, and a weekend versus weekday difference score, both normalized and summed across all time periods.

Moreover, [32] introduces additional advanced time series features that are critical for capturing the dynamic nature of electrical load patterns. These encompass *cross-correlation measures* that reflect the degree of similarity between the time series at varying lags, *autocorrelation-based distances* that emphasize the self-similarity of a load profile, *periodogram-based distances* that assess the distribution of power across frequencies, and *Fourier coefficients-based distances* that isolate the essential frequency components of the time series. These measures enable a deeper understanding of the temporal dynamics within the electrical load data, which is essential for developing more refined and insightful clustering models.

Data size reduction methods

For data size reduction, Chicco et al. [33] in "Comparisons Among Clustering Techniques for Electricity Customer Classification" discuss the use of **Principal Component Analysis (PCA)** , **Sammon map**, and **Curvilinear component analysis (CCA)**, emphasizing the necessity of reducing dataset size for manageable and accurate clustering. Additionally, the "Hopfield–K-Means clustering algorithm" paper [34] introduces the concept of power indexes for data reduction, focusing on key time intervals to summarize load curve information effectively. Furthermore, the "Clustering of Electricity Consumption Behavior Dynamics Toward Big Data Applications" [35] outlines the utilization of the SAX technique for transforming load curves into symbolic strings, thereby simplifying data complexity.The SAX (Symbolic Aggregate Approximation) technique is also used in [36].

3.4 Phase II: Clustering Methodologies

In the following section, we will present various clustering approaches found in literature, acknowledging that there is no consensus on a singular 'best' method. Instead, we observe a range of popular methodologies, each with its own merits in the context of clustering electricity consumers.

Before discussing the various methodologies, it is important to define a commonly used term in the literature: *Representative Load Patterns (RLPs).* RLPs are crucial for understanding electricity consumption behaviors and are defined as the normalized load profiles of individual electricity customers. Mathematically, an RLP for a customer is expressed as:

$$RLP(t) = \frac{Load(t)}{P_{\max}} \tag{3.1}$$

where $Load(t)$ represents the power consumed at time t, and $P_{\max}$ is the maximum power consumption recorded for that customer. This normalization process, by scaling the load pattern relative to $P_{\max}$, allows for a standardized comparison of consumption patterns across different customers, highlighting their relative consumption intensity and timing, independent of their absolute levels of power usage [33],[29],[36].

Traditional clustering algorithms

Traditional unsupervised clustering algorithms are widely used and considered the fundamental approach in classifying electricity consumers. Their prominence stems from their demonstrated efficacy in segmenting data on electricity consumption, making them an essential component of the literature.

"Comparisons Among Clustering Techniques for Electricity Customer Classification" by Chicco et al.[33] is a seminal work that evaluates several clustering methodologies, including K-means, hierarchical clustering, fuzzy K-means, and the modified follow-the-leader approach. Focused on classifying electricity customers based on Representative Load Patterns (RLPs), these techniques are analyzed within a Euclidean distance framework, essential for measuring distances between RLPs and forming meaningful clusters.

In a similar vein, "Overview and performance assessment of the clustering methods for electrical load pattern grouping" [29] delves into the specifics of K-means and Fuzzy K-means algorithms, highlighting their need for a predefined number of clusters. The K-means algorithm is noted for its use of a fixed number of iterations to refine clusters, starting with centroids randomly selected from RLPs. Conversely, the Fuzzy K-means algorithm introduces a fuzziness degree to determine each RLP's cluster membership, with lower degrees often leading to more effective clustering.

Contrasting these methods, hierarchical clustering variants and the Follow the Leader (FDL) method offer a more adaptable clustering approach. Hierarchical clustering, employing an agglomerative procedure, merges RLPs based on similarity measures, while the FDL method eschews the need for predefined cluster numbers and relies on a user-defined distance threshold to determine cluster centroids.

Clustering Algorithm	**References**
K-means	[33], [29], [37], [38], [25] , [39]
Fuzzy K-means	[33], [29], [38]
Weighted Fuzzy Average K-means	[39]
Hierarchical Clustering	[33] , [38] , [39]
Follow the Leader (FDL)	[29], [40], [38] , [39]
Self Organising Maps (SOMs)	[37] , [40], [38] , [39]
Gaussian Mixture Models (GMMs)	[36]

Table 3.1: Summary of Clustering Algorithms and Corresponding References

Multi-stage methodologies

In the literature on electricity consumer categorization, two-stage methodologies have been prominently featured. One such approach, detailed in "A Hybrid Machine Learning

Model for Electricity Consumer Categorization Using Smart Meter Data," [41] utilizes unsupervised clustering algorithms in its first stage to extract typical electricity consumption behaviors. This phase is followed by a classification stage, where fuzzy consumer categorization and supervised classification algorithms are employed for in-depth consumer analysis.

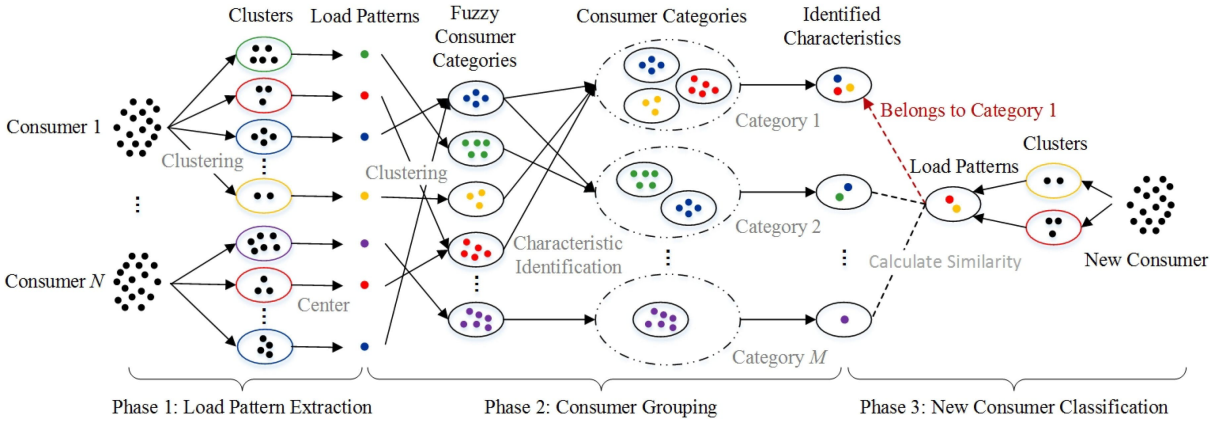


Figure 3.2: Diagram of the two-stage methodology in [41]

Another notable two-stage methodology is presented in "A clustering approach to domestic electricity load profile." [37] The initial phase involves clustering techniques like k-means, k-medoid, and Self-Organizing Maps (SOM).The second stage focuses on characterizing and classifying electricity load profile classes (PCs), utilizing a multinomial logistic regression to classify consumers based on their electricity usage patterns.Additionally, "Hopfield–K-Means Clustering Algorithm" [34] introduces a method that combines the Hopfield neural network with K-Means, enhancing customer segmentation accuracy through a two-stage process.

Distributed clustering

In "Clustering of Electricity Consumption Behavior Dynamics Toward Big Data Applications,"[35] the Fast Search and Find of Density Peaks (CFSFDP) technique is pivotal in profiling electricity consumption behaviors, valued for its low time complexity and robustness to noise. To tackle the complexities of large, distributed datasets, the study integrates a divide-and-conquer approach, applying adaptive k-means, also used in [26], at local sites to obtain representative customer profiles, followed by a modified CFSFDP method at global sites for efficient data processing. This innovative methodology, combining local adaptive clustering with global density-based clustering, addresses the challenges of big data in electricity consumption analysis, optimizing both the computational efficiency and the accuracy of the clustering process.

3.5 Phase III: Performance Evaluation Metrics

As stated in [33] a crucial aspect of the research is the Clustering Validity Assessment. The study conducts repeated executions of various clustering algorithms, varying the

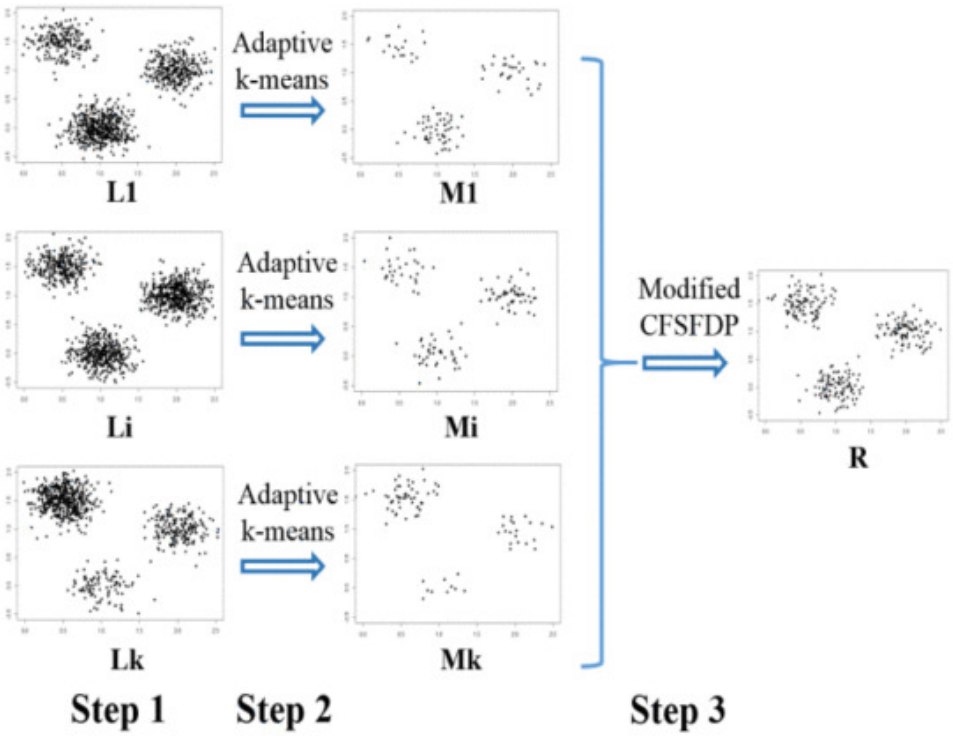


Figure 3.3: Distributed framework proposed in [35]

number of customer classes from 5 to 100. This range was chosen to provide a broad basis for method comparison, though it's noted that having as many as 100 classes is practically too high for real-world applications like tariff association.

Chicco et al.(2004) refer to the metrics used for the assessment of each clustering algorithm as **Adequacy Measures** while in 2006 [33] they refer to them as **Clustering validity indicators**. They propose the following CVIs:

- **Clustering dispersion indicator (CDI)**, which measures the compactness of load patterns within a cluster and their separation from other clusters, also used in [40].
- **Modified Dunn index (MDI)**, adapted to use Euclidean distances, focuses on the ratio of the smallest inter-cluster distance to the largest intra-cluster distance.
- **Scatter index (SI)**, derived from the proportion of scatter accounted for by clustering.
- **Davies–Bouldin index (DB)**, which represents the average similarity measure of each cluster with its most similar cluster ([37],[34],[38],[36]).

In 2011, Chicco [29] complemented the CVIs with the following indicators:

- **Intra-cluster index (IAI)**: This metric evaluates the homogeneity within each cluster. A lower IAI value indicates that the elements within a cluster are more similar to each other, suggesting a better clustering quality.
- **Variance Ratio Criterion (VRC)**: measures the ratio of the sum of variances within clusters to the variance between clusters. High values of VRC indicate dis-

tinct, well-separated clusters, as it implies that the variation within clusters is small compared to the variation between clusters.

- **Inter-cluster index (IEI)**: In contrast to IAI, the IEI assesses the distinctness or separation between different clusters. A higher IEI value suggests that clusters are more distinct from each other, which is a desirable trait in clustering.
- **Mean Index Adequacy (MIA)**: measures how well each object lies within its cluster. It is an average measure that assesses the adequacy of the clustering process for each data point individually, thus providing a more detailed view of the clustering performance ([40],[34],[38],[36])
- **Similarity Matrix Indicator (SMI)**:uses a similarity matrix to evaluate the clustering structure. It helps in visualizing and understanding the relationships between different clusters and the degree of separation or overlap among them.
- **Ratio of within cluster sum of squares to between cluster variation (WCBCR)**: compares the sum of squares within each cluster to the variation between different clusters. A higher value indicates that the data points within each cluster are closely packed together, while being well-separated from other clusters.

The **Calinski-Harabasz index (CH)**, also known as the Calinski index, is a widely-used metric for assessing the quality of clustering algorithms and was used in [34]. Introduced by T. Calinski and J. Harabasz in 1974, it evaluates the clustering by calculating the ratio of the sum of between-clusters dispersion to within-cluster dispersion for different numbers of clusters (See Table 3.2 where TSS = total sum of squares, WSS = the within-cluster sum of squares, N = number of data points in the dataset, and k = number of clusters).

The **silhouette score (SIL)** is also a popular metric used to assess the quality of clusters in a clustering algorithm.It measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation) as was used in [25] as well as in [36]. (See Table 3.2 where a = average intra-cluster distance, b = average shortest distance to another cluster)

Optimal Number of clusters

All the above metrics, including the silhouette score, Davies-Bouldin index, Calinski-Harabasz index, and others, can be utilized for determining the optimal number of clusters in a dataset. Whether these metrics are minimized or maximized, they provide valuable insights into the clustering structure, helping to identify the most effective number of clusters for a given dataset.

Addressing the challenge of determining the optimal number of clusters, the study "A Novel Clustering Index to Find Optimal Clusters Size With Application to Segmentation of Energy Consumers"[42] proposes an Entropy of Eigenvalues (EoE) index. This index assesses the distinguishability between clusters, aiming to identify the optimal cluster size. However, it's important to note that the EoE index primarily captures linear relationships between time series of clusters, potentially limiting its effectiveness in scenarios with strong nonlinear correlations.

Metric	Equation	Rule
CDI	$\frac{1}{d(C)^N}\left[\frac{1}{K}\sum_{k=1}^{K} d^2\left(X^{(k)}\right)\right]$	min
MDI	$\max_{1\leq q\leq K}\left\{\hat{d}\left(X^{(q)}\right)\right\} \times \left(\min_{i\neq j}\left\{d\left(c^{(i)},c^{(j)}\right)\right\}\right)^{-1}$	max
SI	$\left(\sum_{m=1}^{M} d^2\left(x^{(m)},p\right)\right)\left(\sum_{k=1}^{K} d^2\left(c^{(k)},p\right)\right)^{-1}$	min
DB	$\frac{1}{K}\sum_{k=1}^{K}\max_{i\neq j}\left\{\frac{d\left(X^{(i)}\right)+d\left(X^{(j)}\right)}{d\left(c^{(i)},c^{(j)}\right)}\right\}$	min
MIA	$\sqrt{\frac{1}{K}\sum_{k=1}^{K} d^2(c(k),L(k))}$	min
IAI	$\sum_{k=1}^{K}\sum_{x(i)\in L(k)} d^2(c(k),x(i))$	min
VRC	$\frac{1}{M}\left(1+\frac{K}{M-K}W^{-1}\right)$	max
IEI	$\sum_{k=1}^{K} n(k)d^2(c(k),p)$	min
SMI	$\max_{i\neq j}\left\{1-\ln\left(d(c(i),c(j))\right)\right\}$	min
WCBCR	$\frac{\sum_{k=1}^{K}\sum_{x(i)\in L(k)} d^2(c(k),x(i))}{\sum_{i\neq j} d^2(c(i),c(j))^{-1}}$	max
CH	$\frac{TSS-WSS}{WSS}\times\frac{(N-k)}{(k-1)}$	max
SIL	$\frac{b-a}{\max\{a,b\}}$	max

Table 3.2: Summary of Clustering Validity Metrics in Literature

3.6 Phase IV: Post-clustering methodologies

The final stage of clustering methodologies involves the critical tasks of comparing and evaluating the clustering algorithms, and subsequently utilizing the acquired knowledge to create customer classes based on cluster attributes. This stage also entails identifying the ultimate load profiles for each customer category, thereby converting clustering analysis into practical strategies.

Comparison of the algorithms

The study titled "Comparisons among Clustering Techniques for Electricity Customer Classification" presents a detailed comparison of various clustering algorithms, examining a range of cluster numbers from 10 to 30 based on Cluster Validity Indexes (CVIs). The analysis revealed that the modified follow-the-leader and hierarchical clustering using the average distance linkage criterion outperformed other methods in effectiveness. Furthermore, the study advises that a practical limit of 15 to 20 customer classes is optimal for meeting supplier needs, despite theoretical criteria sometimes suggesting a higher number of clusters [33].

Another significant contribution in this field is found in [36], which focuses on hierarchical clustering, particularly employing ward linkage, and identifies it as the most effective method for analyzing daily load patterns of customers. Contrarily, the Self-

Organizing Map (SOM) algorithm was deemed the least effective in this context. The study also evaluated the Fuzzy C-Means (FCM) clustering method, noting its potential yet highlighting its sensitivity to minor adjustments in the fuzziness degree. The CVI values pointed towards an optimal cluster range of 8-10.

In [39] the results indicated that the choice of the appropriate clustering algorithm is somewhat dependent on the clustering's objective. For more distinct clusters, Modified Follow the Leader is the best, while for compact clusters, Weighted Fuzzy Average K-Means is recommended. Overall, Weighted Fuzzy Average K-Means demonstrated better performance across both objectives.

Upon summarizing the comparison of clustering algorithms in these studies, it is clear that although a wide range of clusters is initially taken into account for testing the algorithms, practical considerations of usefulness require a more targeted approach. It is recommended to have a smaller number of clusters so that utilities can easily design and implement effective programs that are customized for each customer class.

Profile Classes

After selecting the right clustering algorithm and determining the cluster number, the focus shifts to identifying **Profile Classes (PCs)** by analyzing each cluster's load curves. This analysis helps to discern unique energy consumption patterns, segmenting customers into distinct profile classes. The creation of these PCs is complex and involves multiple important factors, as mentioned in the literature:

Averaging for Representative Profiles: The most common method found in literature involves averaging the daily electricity demand of all households within a cluster for each time period. This process results in a representative daily load profile for the cluster, effectively smoothing out individual variations and highlighting the prevailing pattern of energy use [43]. However, this method can sometimes oversimplify the diversity within the cluster, masking important nuances in consumption behavior.

Normalization of Values: Prior to averaging, normalizing the values is recommended. This step adjusts for variations in overall consumption levels among different households, allowing for a more equitable comparison of patterns based on the shape of the load curve, rather than the absolute magnitude of consumption[44].The importance of normalization can be seen in the above diagram:

Combining Small Clusters with Similar Patterns: In instances where smaller clusters exhibit only minor differences in magnitudes or timing of electricity use, it may be beneficial to merge them [37]. This strategy not only simplifies the profiling process but also ensures that essential consumption patterns are preserved, thereby generating more robust PCs.

Consideration of Variability: It's crucial to acknowledge the inherent variability in electricity demand profiles within each cluster. While averaging leads to a more uniform profile, the underlying demand may exhibit significant fluctuations [25]. Recognizing and understanding this variability is key to accurately representing the energy usage behaviors of each cluster.

Linking to Household Characteristics: An integral part of creating PCs is linking

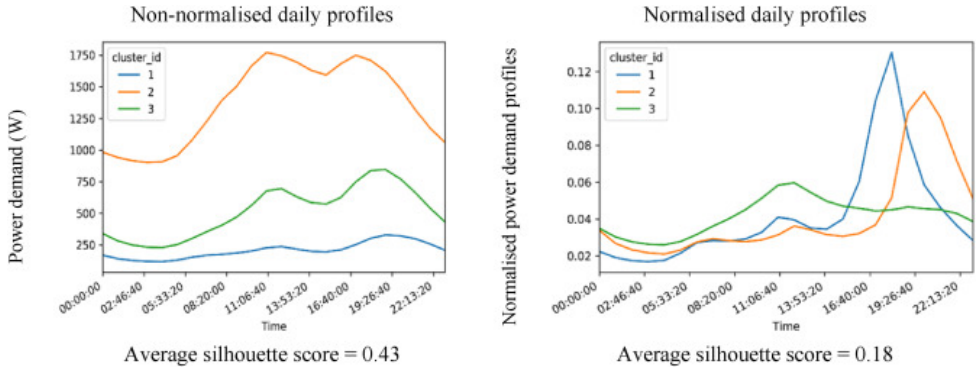


Figure 3.4: The centroids of the cluster found for non-normalised and normalised daily electricity profiles[25]

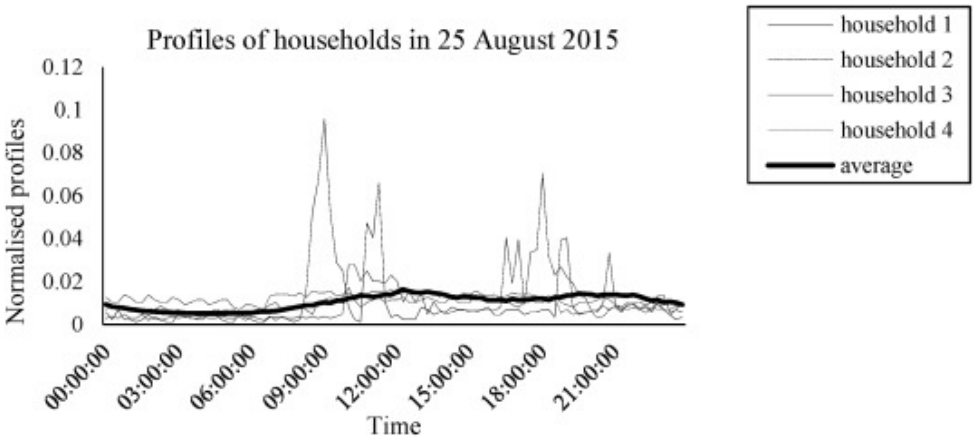


Figure 3.5: Daily electricity demand profiles for four households chosen at random and the average value of 656 households on 25th August 2015 illustrating variation among the households [25]

them to specific dwelling and household characteristics as was illustrated in [37]. This connection not only provides valuable context but also facilitates a deeper understanding of the factors influencing consumption behaviors. Such insights are invaluable for designing targeted energy efficiency programs and informing policy decisions.

3.7 Conclusion

The literature review conducted in this thesis has revealed several significant gaps in the existing literature regarding clustering techniques for energy consumers:

- **Feature Engineering:** There is a noticeable lack of detailed exploration on feature engineering in the context of clustering for consumer energy usage. Most studies primarily rely on load curves, with a limited variety exploring a diverse range of features. Highlighting new and innovative features could significantly enhance clustering effectiveness.
- **Ensemble Clustering Algorithms:** The literature shows a scarcity in the application of ensemble clustering algorithms. These algorithms, known for their robustness

and accuracy, could potentially offer more nuanced insights in consumer segmentation.

- **Explainable AI in Profile Creation:** There's a gap in the utilization of explainable AI for creating consumer profiles. Explainable AI could provide transparency and understandability in the clustering process, aiding in the adoption of these techniques by non-expert stakeholders.

Chapter 4

Methodology

4.1 Introduction

The research project's methodology provides the essential framework that guides the analytical structure, overseeing the investigation from hypothesis development to the ultimate conclusion. This chapter offers an intricate elucidation of the methodology employed in this thesis. This text discusses the methods and analytical approaches employed to group energy consumers by utilizing machine learning algorithms.

The methodology in this chapter is carefully organized in a systematic and sequential manner, guaranteeing clarity and facilitating comprehension. Commencing with a comprehensive outline, we present a lucid and succinct introduction to the methodology to facilitate better understanding. Subsequently, we thoroughly examine each stage of our methodology, providing extensive elaboration. This encompasses a comprehensive depiction of each algorithm employed, along with a detailed examination of every metric implicated.

Furthermore, the methodology is presented with meticulousness and clarity, making it conducive to effortless replication and application in future research. We have meticulously documented every aspect of the process, guaranteeing that anyone desiring to replicate or expand upon our work can do so effortlessly and with utmost precision. The level of detail in our research not only emphasizes its rigor but also enhances its value as a resource that can be reused and adapted in the field.

4.2 Methodology Overview

This research systematically explores London's households - energy consumers clustering to find patterns and structures that can inform energy management strategies and DR programs. The sequential steps and analytical methods that make up the research process are described at the diagram 4.1.

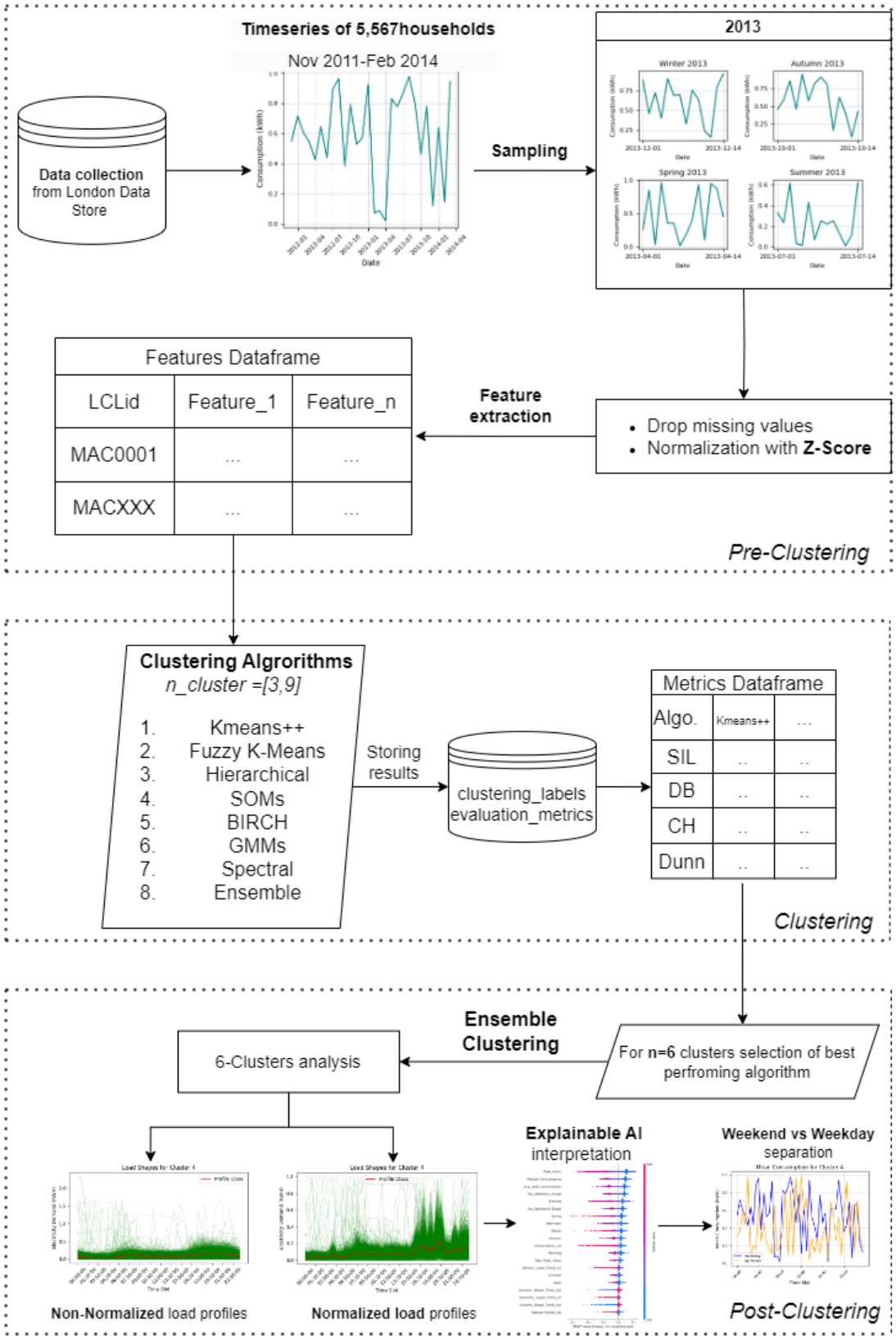


Figure 4.1: Detailed Overview of methodology

The quantitative research uses machine learning algorithms to analyze a large energy consumption dataset. The methodology begins with dataset acquisition, followed by exploratory data analysis (EDA), feature engineering, and unsupervised learning for clustering.

Feature engineering requires a foundational understanding of the data's characteristics from the EDA. In this crucial stage, domain knowledge is used to construct informative features that may indicate energy consumption patterns. To reveal data structure for clustering, this process involves selecting relevant variables, creating new composite features, and transforming variables.

After feature engineering, unsupervised machine learning algorithms cluster the data. Algorithms are chosen based on their compatibility with the dataset and their multidimensional cluster detection methods. The algorithms used include Kmeans++, Fuzzy K-means, and Hierarchical clustering.

Internal validity indices measure cluster structure fitness and evaluate clustering results. The Dunn index and other metrics assess cluster compactness and separation. These measures compare algorithm performance across cluster numbers and algorithms.

The methodology determines the best clustering solution using evaluation criteria. This solution is then analyzed to interpret the clusters in the context of energy consumer behavior, advancing energy consumption analysis.

Our research methodology relies on a solid technical framework to efficiently handle and analyze the large dataset. Our primary work environment is **Google Colab**, a cloud-based platform with powerful computing and ease of use. The infrastructure supports large-scale data processing and analysis.

Alongside Google Colab, **Python** was employed as the main programming language. Known for its versatility and wide range of libraries, Python is particularly well-suited for data analysis and machine learning tasks. Its intuitive syntax and rich ecosystem of tools enable efficient handling of complex data operations, which are fundamental to our study.

In conclusion, the methodology overview prepares the data for systematic and precise analysis. Each step will be detailed in the following sections to explain the research process.

4.3 Data description and EDA

This study used data from the London Data Store, a Greater London Authority(GLA) initiative to make data public. The dataset "Smart-meter Energy Use Data in London Households" contains half-hourly energy consumption readings from smart meters in London households. This detailed dataset of energy usage patterns is essential to our demand response management analysis.

Academic and research use of the dataset is free. To access the dataset, visit the London Data Store website at London Datastore to download the data in various formats. The open dataset shows the city's commitment to public service transparency and innovation.

Registration was optional, making data collection easy. Following data protection laws like the GDPR, the dataset is anonymized to protect privacy. The anonymization process ensures ethical research by preventing personal data from being linked to individuals. Despite anonymity, researchers should follow ethical guidelines when handling data.

4.3.1 Dataset Overview

The dataset meticulously records the energy consumption of 5,567 Greater London households, representing the city's diverse population. The record covers **November 2011 to February 2014**, allowing for longitudinal analysis. Half-hourly energy consumption readings result in a large dataset of 167 million rows. The dataset includes several key variables:

- **LCLid**: Serves as the anonymized unique identifier for each household, essential for discrete data analysis without compromising privacy.
- **stdorToU (Standard or Time of Use):** This categorizes the tariff the customer is on.
- **DateTime:** This timestamp records the date and time when the energy consumption was logged.
- **KWH/hh (Kilowatt-Hours per Half-Hour)**: Represents the energy consumed in kilowatt-hours during each half-hour interval.

Figure 4.2 below presents a snapshot of the dataset in Excel, highlighting key variables and their recorded values.

LCLid	stdorToU	DateTime	KWH/hh (per half hour)
MAC000002	Std	12/1/2012 0:00	0.215
MAC000002	Std	12/1/2012 0:30	0.217
MAC000002	Std	12/1/2012 1:00	0.237
MAC000002	Std	12/1/2012 1:30	0.204
MAC000002	Std	12/1/2012 2:00	0.243
MAC000002	Std	12/1/2012 2:30	0.199
MAC000002	Std	12/1/2012 3:00	0.237
MAC000002	Std	12/1/2012 3:30	0.125

Figure 4.2: Snapshot of the energy consumption dataset

The dataset identifies two primary tariff structures:

1. **Standard Tariff (Std):** A consistent rate applied across the board, regardless of the time of day, offering a baseline for energy costs.
2. **Dynamic Time of Use (dToU) Tariff:** A variable pricing model experienced by a subset of around 1,100 households, with costs fluctuating based on time-of-day demand, aiming to incentivize consumption during off-peak hours. The dToU structure had three pricing levels—High, Low, and Normal—communicated to customers a day in advance.

Those on the dToU tariff were informed of the rates via their Smart Meter In-Home Display or by text message, aligning energy consumption with periods of high renewable generation or to alleviate grid stress during peak demand.

4.3.2 Exploratory Data Analysis

Given the extensive size of the dataset, encompassing approximately 167 million rows of half-hourly energy consumption data, a pragmatic approach was necessary to make the dataset manageable for analysis with the available computational resources. The constraints primarily arose from limitations in RAM and GPU power, which are critical for processing and analyzing large datasets efficiently.

To address this, a strategic sampling method was employed. The dataset was reduced to include only the first two weeks of December, October, April, and July from the year 2013. This selection was made based on several considerations:

1. **Seasonal Representation:** These months were chosen to represent different seasons—winter, autumn, spring, and summer—ensuring that the sample captured varying energy consumption patterns influenced by seasonal changes.
2. **Peak and Off-Peak Periods:** December and July typically represent peak energy usage months due to heating and cooling needs, respectively, whereas April and October are generally considered off-peak months. Including both peak and off-peak periods allows for a more comprehensive analysis of energy consumption behaviors.
3. **Data Diversity**: Sampling from different months ensures a diverse range of data, contributing to a more robust and generalizable analysis. This diversity is crucial for developing machine learning models that are effective across various conditions.
4. **Computational Feasibility:** Limiting the dataset to specific weeks reduces the computational load, making it feasible to process and analyze the data with the available resources without compromising the integrity and representativeness of the analysis.

Pandas was extensively used for data analysis and manipulation in Python. The powerful Pandas data analysis toolkit makes structured data manipulation easy. Its numerical table and time series manipulation data structures and operations make it ideal for our comprehensive EDA. To gain preliminary insights into the sampled dataset's consumption patterns, mean, median, standard deviation, minimum, and maximum values were calculated.

For in-depth analysis, the sampled data was critically assessed for integrity and accuracy. Key steps in this quality assessment process:

- **Missing Values:** The dataset was thoroughly scanned for missing values. Only three missing data points were found, a negligible percentage of the dataset. These missing values had little effect on the dataset's representativeness, and imputation could introduce bias, so they were dropped. This method preserves data authenticity while minimizing analysis impact.

- **Outliers Detection:** This study avoided outlier detection. This choice was based on the dataset and research focus. Since the dataset represents real-world energy consumption patterns, it was important to preserve its natural variability and extremes to gain insights into unusual but potentially significant consumption behaviors.

The 2012 dataset was added to the data quality assessment. This inclusion was strategic to enable a year-on-year energy consumption comparison to identify trends.The study focused on households from both 2012 and 2013 datasets for consistency and accurate comparisons. This method ensures that energy consumption trends are due to changes over time rather than household variations.

4.4 Feature Engineering

Feature engineering is crucial to machine learning model development. Changing features to make raw data analyzeable is the key. This step can improve a model by highlighting key data patterns or removing noise and irrelevant details. We want to create features that reveal data structures to help machine learning algorithms learn and predict.We chose feature engineering over using load curves, which have dominated literature. To explore new features that may be indicative, we went off the beaten path. This method may yield valuable insights and a deeper understanding of the data, improving model performance and predictive accuracy.

Normalization

Our study began feature engineering with normalization, which was crucial due to the dataset's diverse energy consumption values. Normalization scales data between 0 and 1, ensuring that all features contribute equally to model performance. When dealing with features of different scales, this process prevents larger features from dominating model learning.

We normalized using **standard scaling**, also known as **Z-score normalization**. Each feature's mean is subtracted and divided by its standard deviation. Each feature will have a zero mean and one standard deviation. Standard scaling standardises independent variables, which is useful when dataset features have different units or scales. A uniform scale ensures that each feature contributes proportionally to the final prediction, improving the model's ability to learn from data.

$$z = \frac{(x - \mu)}{\sigma} \tag{4.1}$$

In this formula:

- z represents the standardized value after scaling.
- x is the original value of the feature.
- μ is the mean of the feature.

- σ is the standard deviation of the feature.

After normalization, we developed and selected energy consumption-related features.

Median Consumption

The *Median Consumption* is a key feature of our machine learning model. A household's median energy consumption over a period is represented by this feature. The median is more robust than the mean and less sensitive to outliers. It is important in energy consumption analysis, where usage spikes can skew the mean. The median is the middle number in a sorted list. It is the average of the two middle numbers if the list has even observations.

$$\text{Median}(X) = \begin{cases} x_{\frac{n+1}{2}}, & \text{if } n \text{ is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{if } n \text{ is even} \end{cases} \tag{4.2}$$

- $X = \{x_1, x_2, \ldots, x_n\}$ represents a sorted set of observations.
- n is the number of observations in X.
- x_i represents the i-th value in X.

Average Peak Consumption

Another key feature engineered for our analysis is the *Average Peak Consumption*. This feature represents the average energy usage during peak consumption periods for each household. Peak periods are critical as they often correspond to the highest demand on the energy grid, and understanding these periods is crucial for efficient energy management and load balancing. The process for calculating this feature involves two primary steps:

1. **Identification of Peak Periods:** Peak periods were identified using the 'find_peaks' function from the 'scipy.signal' library. This method involves a simple comparison of neighboring values to determine the peaks in energy consumption.
2. **Calculation of Average Peak Consumption:** After identifying peak periods, the average consumption during these times is calculated. This involves averaging the energy consumption values (x_i) at each peak, where N is the number of peaks. The formula is given by:

$$\text{Average Peak Consumption} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{4.3}$$

Here, $\sum_{i=1}^{N} x_i$ sums the energy consumption during peak periods, and dividing by N yields the average.

Peak Frequency in Each Time Slot

The fourth feature we focused on is *Peak Frequency in Each Time Slot*. This feature quantifies the frequency of peak energy consumption within specific time slots throughout the day for each household. The process for developing this feature involved several steps:

1. **Time Slot Identification:** Each record in the dataset was associated with a specific time slot based on the time of day of the energy consumption. This step involved extracting the time component from each timestamp.
2. **Peak Counting per Time Slot:** For each household, the number of peak energy consumption occurrences within each time slot was counted.
3. **Total Count Calculation:** The total number of records for each household over the analysis period was calculated.
4. **Frequency Calculation:** The frequency of peak consumption for each time slot was computed by dividing the peak count by the total number of records, yielding a frequency value for each time slot.
5. **Matrix Formation:** The frequencies were arranged in a matrix format, with each row corresponding to a household and each column to a time slot. This matrix offers a comprehensive view of peak consumption patterns throughout the day for each household.
6. **Data Cleaning:** Frequencies were set to zero for time slots with no recorded peaks to maintain consistency.

The *Peak Frequency in Each Time Slot* feature provides valuable insights into the temporal distribution of high-energy consumption events, aiding energy suppliers and policymakers in understanding and managing energy demand dynamics throughout the day.

Consumption Standard Deviation

The *Consumption Standard Deviation* is a key metric in our energy consumption analysis, measuring the variability in a household's energy use over time.A high standard deviation indicates irregular energy use, while a low value suggests consistency. The standard deviation is mathematically expressed as:

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N-1}} \tag{4.4}$$

In this equation, x_i are the individual consumption values, μ is the mean consumption, and N is the number of observations. The formula calculates the square root of the average squared deviations from the mean, using $N-1$ in the denominator. The Pandas 'std' function uses the N-1 denominator.

Average Consumption by Time of Day

In our analysis, we further segmented the energy consumption data into different parts of the day to understand the consumption patterns during specific timeframes. This segmentation resulted in four key features:

- **Average Morning Consumption**
- **Average Afternoon Consumption**
- **Average Evening Consumption**
- **Average Night Consumption**

Each of these features is calculated by averaging the energy consumption readings within the respective time slots for each household.

$$\text{Average Time-of-Day Consumption} = \frac{\sum_{t \in T} x_t}{|T|} \tag{4.5}$$

In equation 4.5, x_t represents the energy consumption in time slot t, T is the set of time slots corresponding to morning, afternoon, evening, or night, and $|T|$ is the number of slots in T. This calculation is repeated for each of the time segments to derive the respective average consumption values.

Average Seasonal Consumption

Understanding how energy consumption varies with the seasons is crucial for our analysis. To this end, we divided the data into four distinct seasonal categories, resulting in the creation of the following features:

- **Average Winter Consumption**
- **Average Autumn Consumption**
- **Average Spring Consumption**
- **Average Summer Consumption**

These seasonal features are computed by averaging the energy consumption readings for each household across the respective seasons.

Data Filtering: It is important to note that during the feature creation process, we encountered households that lacked data for one or more seasons. To maintain the integrity and consistency of our analysis, these households were excluded from the dataset. This decision was made to ensure that our seasonal consumption analysis is based on complete and representative data for each household across all seasons.

$$\text{Average Seasonal Consumption} = \frac{\sum_{d \in S} x_d}{|S|} \tag{4.6}$$

In equation 4.6, x_d represents the energy consumption on day d, S denotes the set of days in the considered season, and $|S|$ is the total number of days in S. This formula is applied separately for each season to obtain the average consumption values.

Average Peak-Hours and Off-Peak-Hours Usage

In our energy consumption analysis, distinguishing between peak-hours and off-peak-hours usage is essential to understand the dynamics of energy demand. Based on the study referenced in [45], we define peak hours as the period from **16:00 to 20:00**. Accordingly, we have developed two features:

- **Average Peak-Hours Usage**
- **Average Off-Peak-Hours Usage**

The calculation of these features involves aggregating the energy consumption data within the respective time frames and computing the average usage for each household.

$$\text{Average Peak-Hours Usage} = \frac{\sum_{t \in T_{peak}} x_t}{|T_{peak}|} \tag{4.7}$$

$$\text{Average Off-Peak-Hours Usage} = \frac{\sum_{t \in T_{off-peak}} x_t}{|T_{off-peak}|} \tag{4.8}$$

In these equations, x_t denotes the energy consumption at time t, T_{peak} represents the set of peak hours, $T_{off-peak}$ represents the off-peak hours, and $|T_{peak}|$ and $|T_{off-peak}|$ are the number of time slots in peak and off-peak periods, respectively.

Average Weekday and Weekend Usage

Understanding the difference in energy consumption patterns between weekdays and weekends is vital for a comprehensive analysis of household energy usage. To this end, we have developed two distinct features:

- **Average Weekday Usage**
- **Average Weekend Usage**

The calculation of these features involves segmenting the energy consumption data into weekdays and weekends and then computing the average consumption for each segment for every household.

$$\text{Average Weekday Usage} = \frac{\sum_{d \in D_{weekday}} x_d}{|D_{weekday}|} \tag{4.9}$$

$$\text{Average Weekend Usage} = \frac{\sum_{d \in D_{weekend}} x_d}{|D_{weekend}|} \tag{4.10}$$

In these equations, x_d represents the energy consumption on day d, $D_{weekday}$ is the set of all weekdays in the analysis period, $D_{weekend}$ is the set of all weekend days, and $|D_{weekday}|$ and $|D_{weekend}|$ are the total number of days in each set, respectively. These formulas provide the average consumption values that are essential for understanding the variations in household energy consumption across different days of the week.

Seasonal Usage Trend

To understand how energy consumption patterns change throughout the year, we conducted a seasonal analysis. We calculated usage trends individually for each of the four seasons - Spring, Summer, Autumn, and Winter.The process involved:

1. Aggregating daily energy consumption data for each household.
2. Creating a 'DayOfYear' column to represent each day numerically within its respective season.
3. Using the *LinearRegression* model from *scikit-learn* to fit a linear model to the daily usage data for each season.
4. Calculating the slope of the regression line for each season and household, indicating consumption trends. A positive slope suggests an increase, while a negative slope indicates a decrease.

The *Seasonal Usage Trend* is defined as the slope of the linear regression model fitted to each season's data. It quantifies the direction and magnitude of consumption trends within specific seasons.

Autocorrelation Features

Autocorrelation is a statistical measure that helps in understanding the degree of similarity between a time series and a lagged version of itself over successive time intervals. In the context of energy consumption, it is crucial for identifying patterns and predictability in usage, which are key factors in the effectiveness of DR strategies. We focused on two specific autocorrelation features:

- **Autocorrelation 7 days**: This feature calculates the autocorrelation of energy consumption with a lag of seven days, helping to understand weekly consumption patterns.
- **Autocorrelation 1 day**: Similarly, this feature measures the autocorrelation with a lag of one day.

Households with high autocorrelation, especially over these time lags, are ideal candidates for DR programs, as their energy usage patterns are more predictable and can be managed more effectively. The autocorrelation is calculated using the formula:

$$\text{Autocorrelation} = \frac{\sum_{t=1}^{N-lag}(x_t - \mu)(x_{t+lag} - \mu)}{\sum_{t=1}^{N}(x_t - \mu)^2} \tag{4.11}$$

In this equation, x_t is the energy consumption at time t, μ is the mean energy consumption, N is the total number of observations, and lag is the lag period (1 day or 7 days). This calculation provides a numerical value indicating the strength and direction of the relationship between the energy consumption and its lagged values.

Consumption Growth

As part of our energy consumption study, we assessed the change in consumption from 2012 to 2013, captured by the "Consumption Growth" feature. This metric quantifies the year-over-year variation in average energy consumption for each household. The process involved:

1. Calculating the average energy consumption for each household in both 2012 and 2013.
2. Determining the consumption growth by subtracting the 2012 average from the 2013 average for each household, representing the change in energy usage over the two years.
3. Adding the consumption growth values to our features dataset as 'Consumption_Growth.'
4. Handling missing values by filling gaps with the median consumption growth to ensure data consistency.

$$\text{Consumption Growth} = \text{Average Consumption}_{2013} - \text{Average Consumption}_{2012} \tag{4.12}$$

Equation 4.12 succinctly represents the calculation of Consumption Growth, quantifying the year-over-year change in energy usage for each household.

The *Consumption Growth* feature is essential for understanding how energy usage patterns evolve over time, benefiting energy providers and policymakers in demand planning and forecasting.

Variability in Peak-Usage Times

The *Variability in Peak-Usage Times* quantifies the consistency of peak energy consumption times for each household. High variability implies irregular peak times, while low variability suggests a more consistent pattern. The calculation involved:

1. **Identification of Peak Times (in seconds):**

   Let P_{ij} represent the peak time in seconds for household i at peak j.

2. **Calculation of Standard Deviation of Peak Times (Time Variability):**

   For each household i, calculate the standard deviation (σ_i) of peak times (P_{ij}) using the formula:

$$\sigma_i = \sqrt{\frac{1}{N_i - 1} \sum_{j=1}^{N_i} (P_{ij} - \mu_i)^2}$$

   Where N_i is the number of peak times for household i and μ_i is the mean peak time for household i.

3. **Normalizing Standard Deviations:**

   Normalize the standard deviations (σ_i) across households by calculating the mean (μ_σ) and standard deviation (σ_σ) of all household standard deviations:

$$\mu_\sigma = \frac{1}{M} \sum_{i=1}^{M} \sigma_i$$

$$\sigma_\sigma = \sqrt{\frac{1}{M - 1} \sum_{i=1}^{M} (\sigma_i - \mu_\sigma)^2}$$

   Where: - M is the total number of households.

4. **Handling Missing Data:**

   Fill gaps in the calculated standard deviations (σ_i) with zeros for households with insufficient peak data.

The "Variability in Peak-Usage Times" feature, represented as "Time Variability (normalized)," is defined as:

$$\text{Time Variability (normalized)} = \frac{\sigma_i - \mu_\sigma}{\sigma_\sigma}$$

This equation quantifies the normalized variability in peak energy consumption times for each household, allowing for consistency and comparison across households.

4.5 Machine Learning Algorithms

This section of the thesis describes the machine learning algorithms utilized in our analysis. The choice of algorithms was guided by the nature of our dataset, the specific characteristics of our features, and the goals of the study. We employed a range of algorithms, each offering unique strengths and suited for different aspects of energy consumption analysis.

Feature Transformation for Algorithm Application

To optimize our features for machine learning algorithms, we focused on transforming the seasonal usage trend features that represent energy consumption patterns in each season (Autumn, Spring, Summer, Winter). In this transformation:

- For each seasonal trend feature, we introduced two new columns: one for the sine and another for the cosine of the trend value. These new columns represent the cyclical components of the original trends.
- We retained the new sine and cosine columns while dropping the original linear trend features.

This approach allows machine learning algorithms to better understand and utilize the cyclical patterns in energy consumption data, leading to improved model performance and interpretability.

4.5.1 K-Means++

In our analysis, the K-Means++ clustering algorithm was implemented to segment households based on energy consumption patterns. K-Means++ is an extension of the standard K-Means algorithm, with an improved initialization method that enhances the quality of the resulting clusters.

Algorithm Description

K-Means++ clustering algorithm seeks to partition the dataset into a predetermined number of clusters, K, by minimizing the within-cluster variance. The algorithm involves the following steps:

Sklearn Implementation

The implementation of K-Means++ in our study was carried out using the `KMeans` class from the `scikit-learn` library in Python. We explored a range of values for K to find the optimal number of clusters for our dataset.This range is chosen to be between 3 and 10 clusters, allowing for a comprehensive analysis of the dataset. The `KMeans` function in sklearn is invoked as follows:

```
from sklearn.cluster import KMeans

kmeans_plusplus = KMeans(n_clusters=K, init='k-means++',
                         random_state=0, n_init=20)
kmeans_plusplus.fit(data)
```

Algorithm 1 K-Means++ Clustering

```
procedure KMeans++(Data, K)
    Initialize an empty set of centroids, C
    Select the first centroid c_1 randomly from the data points
    C ← C ∪ {c_1}
    while |C| < K do
        Select the next centroid c_i using a weighted probability distribution
        C ← C ∪ {c_i}
    end while
    repeat
        for each point x in Data do
            Find the nearest centroid c_j ∈ C
            Assign x to cluster j
        end for
        for each cluster j in C do
            Calculate the new centroid c_j as the mean of all points in cluster j
        end for
    until centroids do not change or maximum iterations are reached
    return clusters and centroids
end procedure
```

In this implementation: - `n_clusters` represents the number of clusters, K. - `init='k-means++'` specifies the use of the K-Means++ initialization method. - `random_state` ensures reproducibility of results. - `n_init` denotes the number of times the algorithm will run with different centroid seeds.

$$\text{Objective Function} = \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{4.13}$$

Equation 4.13 represents the objective function of K-Means++, where C_i is the i^{th} cluster, x is a data point in C_i, and μ_i is the centroid of C_i. The goal is to minimize this objective function, which quantifies the within-cluster variance.

4.5.2 Fuzzy K-Means

Fuzzy K-Means, also known as Fuzzy C-Means (FCM), is an advanced clustering technique used in our study to analyze energy consumption patterns. This algorithm extends the idea of traditional K-Means clustering by allowing data points to belong to multiple clusters with varying degrees of membership.

Algorithm Description

In Fuzzy K-Means clustering, each data point is assigned a membership level for each cluster, ranging from 0 (no membership) to 1 (full membership). This approach contrasts with the hard clustering of K-Means, where each point is assigned to only one cluster.

Algorithm 2 Fuzzy K-Means Clustering

procedure FuzzyKMeans(Data, K, m)
    Initialize membership matrix U with random values between 0 and 1
    Normalize each column of U to ensure that the sum of memberships for each point is 1
    repeat
        for each cluster k from 1 to K **do**
            Calculate the centroid c_k as a weighted mean of all points
        end for
        for each point x_i and each cluster k **do**
            Calculate the membership u_{ik} using the distance to centroid c_k
            Update u_{ik} using the fuzziness parameter m
        end for
        Normalize each column of U to maintain the constraint on memberships
    until the change in U is below a threshold or a maximum number of iterations is reached
    return membership matrix U and centroids C
end procedure

Implementation Using Scikit-Fuzzy

The implementation of Fuzzy K-Means in our study was conducted using the `scikit-fuzzy` library in Python [46]. The algorithm was applied across the same range of cluster numbers as before [3,10], and for each configuration, cluster centers and membership degrees were calculated.

```
import skfuzzy as fuzz

# Initialize a dictionary to store membership matrices
#for each cluster countmemberships_fuzzycmeans = {}

 #Apply the Fuzzy C-Means algorithm
    cntr, u, u0, d, jm, p, fpc = fuzz.cluster.cmeans(
       data.T, n_clusters, 2, error=0.005, maxiter=1000, init=None
    )
# Determine the cluster assignments based
#on the highest membership value
    cluster_assignments_fuzzy = np.argmax(u, axis=0)
```

In the context of this implementation: - The ‘fuzz.cluster.cmeans‘ function is the central component, where ‘data_df.T‘ is the transposed data frame input. - The parameter ‘n_clusters‘ represents the number of clusters K. - The fuzziness coefficient is set to 2, controlling the degree of cluster fuzziness. - The ‘error‘ parameter and ‘maxiter‘ define the stopping criteria of the algorithm. - The ‘init‘ parameter is set to ‘None‘, allowing the algorithm to select initial cluster centers automatically.

4.5.3 Hierarchical Clustering

Agglomerative Hierarchical Clustering is a technique used in our study to uncover the hierarchical structure within the energy consumption dataset. This method is particularly

useful for its ability to provide a detailed view of the data's clustering at various levels of granularity.

Mathematical Basis of the Algorithm

The Agglomerative Hierarchical Clustering algorithm operates on the principle of iteratively merging clusters based on a certain measure of dissimilarity or distance between them. The process starts with each data point as a single cluster and then successively merges clusters until all points are merged into a single cluster.

Distance Metrics and Linkage Criteria The key to this algorithm lies in the choice of distance metric and linkage criterion:

- **Distance Metric**: This defines the distance between data points. Common metrics include Euclidean distance, Manhattan distance, etc.
- **Linkage Criterion**: This determines the distance between clusters. The Ward method, used in our study, minimizes the total within-cluster variance at each merging step.

Mathematically, the Ward linkage criterion minimizes the sum of squared differences within all clusters. It is a variance-minimizing approach and is given by:

$$\Delta(\text{SS}_W) = \sum_{i=1}^{N} ||x_i - \mu_{C \cup C'}||^2 - \sum_{i=1}^{N} ||x_i - \mu_C||^2 - \sum_{i=1}^{N} ||x_i - \mu_{C'}||^2 \tag{4.14}$$

where $\Delta(\text{SS}_W)$ is the increase in the total within-cluster sum of squares as a result of merging cluster C and C', x_i are the data points, and μ denotes the mean of the points in a cluster.

Sklearn Implementation of Hierarchical Clustering

The implementation of Hierarchical Clustering in our study was conducted using the `AgglomerativeClustering` class from the `scikit-learn` library in Python. To explore the data's inherent structure, we applied the algorithm across a range of cluster numbers, varying from 3 to 11. This range was chosen to comprehensively analyze the dataset and understand the natural groupings within it.

In our implementation, we specifically used the Ward linkage method within the `AgglomerativeClustering` class, as it minimizes the total within-cluster variance, thereby creating more homogenous clusters. The `AgglomerativeClustering` function in sklearn is invoked as follows:

```
from sklearn.cluster import AgglomerativeClustering
```

```
# Applying Agglomerative Clustering with the Ward linkage method
hierarchical_ward = AgglomerativeClustering(n_clusters=K,
linkage='ward')hierarchical_ward.fit(data)
```

In this implementation: - `n_clusters` represents the number of clusters, K. - `linkage='ward'` specifies the use of the Ward linkage method. - The `fit` method applies the algorithm to the data, executing the agglomerative clustering process.

This approach to Hierarchical Clustering using the Ward method allowed us to systematically reveal and analyze the layered structure of household energy consumption behaviors, enhancing our understanding of various consumer groups and patterns.

4.5.4 Self-Organizing Maps (SOMs)

Self-Organizing Maps (SOMs) represent another significant machine learning algorithm employed in our analysis. SOMs are a type of unsupervised learning algorithm used for clustering and visualizing high-dimensional data in a lower-dimensional space, typically two dimensions.

Algorithm Overview

SOMs operate by mapping the input data onto a grid of neurons, where each neuron has a position in the grid and a weight vector of the same dimensionality as the input data. The training process involves adjusting these weights to preserve the topological properties of the input space, leading to a form of dimensionality reduction. The key steps in the SOM algorithm include:

Algorithm 3 Self-Organizing Maps (SOMs)

```
procedure SelfOrganizingMap(Data, grid_size, iterations)
    Initialize a grid of neurons with random weights
    for each iteration do
        for each input vector v in Data do
            Find the Best Matching Unit (BMU) on the grid
            Determine the neighborhood of the BMU
            for each neighbor n of BMU do
                Adjust the weights of n to be more like v
            end for
        end for
        Decrease the neighborhood radius
        Decrease the learning rate
    end for
    return trained grid
end procedure
```

Implementation Using MiniSom

In our study, SOMs were implemented using the 'MiniSom' library in Python [47]. The algorithm was applied to cluster energy consumption data, exploring various config-

urations to identify meaningful groupings:

```
from minisom import MiniSom

# Define function to train SOM
def train_som(som_shape, data):
    som = MiniSom(som_shape[0], som_shape[1], data.shape[1],
    sigma=0.9, learning_rate=1)
    som.train_random(data.values, 1000)
    winner_coordinates = np.array([som.winner(x) for x
    in data.values]).T
    som_cluster_index = np.ravel_multi_index(winner_coordinates,
    som_shape)
    return som_cluster_index

# Example of training a SOM
som_shape = (number_of_clusters, 1)  # Example shape
som_cluster_index = train_som(som_shape, data)
```

In this implementation, we trained SOMs of different shapes and configurations, experimenting with various numbers of clusters to best capture the patterns in our dataset. The 'train_som' function defines the training process, including grid size (determined by 'som_shape'), learning parameters, and the training algorithm.

4.5.5 BIRCH Clustering

Our analysis utilizes the BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm, which is especially adept at clustering large datasets. BIRCH incrementally processes data, efficiently constructing a hierarchical tree structure.

Algorithm Overview

BIRCH is ideal for large datasets due to its incremental and memory-efficient approach. The algorithm's process involves several key steps:

Implementation Using Scikit-Learn

For implementing BIRCH, we used the `Birch` class from the `scikit-learn` library. We configured the BIRCH parameters to align with our dataset's characteristics, focusing on energy consumption patterns:

```
from sklearn.cluster import Birch

# Configuration of the BIRCH algorithm
birch_model = Birch(n_clusters=number_of_clusters,
```

Algorithm 4 BIRCH Clustering

```
procedure BIRCH(Data, threshold, branching_factor)
    Initialize a Clustering Feature (CF) Tree with given threshold and branching factor
    for each data point d in Data do
        Insert d into the CF Tree
        if any CF Tree node exceeds the threshold then
            Split the node according to the branching factor
        end if
        Update the tree structure to reflect the new data
    end for
    Optionally perform global clustering on the leaf nodes
    Handle outliers and noise in the dataset
    return clusters and updated CF Tree
end procedure
```

```
threshold=0.5, branching_factor=50)
birch_model.fit(data)
```

In this configuration: - `n_clusters` specifies the number of clusters. - `threshold` determines the radius of the sub-cluster obtained in the leaf node of the CF Tree. - `branching_factor` defines the number of sub-clusters a node in the tree can have.

4.5.6 Gaussian Mixture Models (GMMs)

Gaussian Mixture Models (GMMs) are employed in our analysis to cluster household energy consumption data. GMMs are based on the assumption that all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

Mathematical Foundation of GMMs

The GMM represents the data using a probabilistic model where each component corresponds to a Gaussian distribution. Mathematically, the probability of observing a data point x is given by:

$$P(x) = \sum_{i=1}^{K} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i) \tag{4.15}$$

In this equation: - K is the number of Gaussian distributions (clusters). - π_i is the weight of the i^{th} Gaussian in the mixture. - $\mathcal{N}(x|\mu_i, \Sigma_i)$ is the i^{th} Gaussian distribution with mean μ_i and covariance matrix Σ_i. - $P(x)$ is the probability density of x under the model.

Expectation-Maximization in GMMs

The parameters of GMMs (π_i, μ_i, Σ_i) are typically estimated using the Expectation-Maximization (EM) algorithm, which iteratively applies the following steps:

- **Expectation (E) Step**: Calculate the probability of each data point belonging to each cluster.
- **Maximization (M) Step**: Update the parameters (π_i, μ_i, Σ_i) to maximize the likelihood of the data given these probabilities.

Algorithm 5 Clustering with Gaussian Mixture Models

```
procedure ClusterGMM(Data, K)
    Initialize the parameters π_i, μ_i, Σ_i for each Gaussian i in the mixture
    repeat
        Expectation Step:
        for each data point x in Data do
            Calculate the probability of x belonging to each Gaussian
            Assign x to the cluster with the highest probability
        end for
        Maximization Step:
        for each cluster do
            Recalculate the mean μ and covariance Σ based on assigned data points
            Update the weight π of the cluster
        end for
    until cluster assignments do not change or a maximum number of iterations is reached
    return cluster assignments and updated parameters π_i, μ_i, Σ_i
end procedure
```

Implementation Using Scikit-Learn

We implemented GMMs in our study using the `GaussianMixture` class from the `scikit-learn` library. The algorithm was configured to explore a range of clusters, determining the best fit for our dataset:

```
from sklearn.mixture import GaussianMixture

    gmm = GaussianMixture(n_components=n_clusters, random_state=0)
    gmm.fit(data)
    cluster_assignments_gmm = gmm.predict(data)

    labels_gmm[n_clusters] = cluster_assignments_gmm
```

This process involved estimating the GMM parameters for different numbers of clusters (specified by `n_components`) and assigning data points to clusters based on the fitted model.

The GMM's probabilistic approach and flexibility in modeling complex data distributions made it a valuable tool for identifying and understanding the patterns in household energy consumption.

4.5.7 Spectral Clustering

Spectral Clustering, used in our analysis, is particularly adept at identifying complex structures in data that are not necessarily linearly separable. This method involves transforming the data into a space where clusters are more apparent.

Algorithm Overview

The foundation of Spectral Clustering lies in graph theory and linear algebra. Its process can be described as follows:

Algorithm 6 Spectral Clustering

Input: Data set $S = \{x_i\}_{i=1}^{n} \in \mathbb{R}^p$, number of clusters k, affinity parameter σ
Output: Cluster labels for the data set

function SpectralClustering(S, k, σ)
    Form the affinity matrix $A \in \mathbb{R}^{n \times n}$ defined by:
    $A_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$ for $i \neq j$, and $A_{ij} = 0$ otherwise
    Construct the diagonal matrix D with $D_{ii} = \sum_j A_{ij}$
    Compute the normalized Laplacian $L = D^{-1/2} A D^{-1/2}$
    Perform eigenvalue decomposition on L to obtain the k smallest eigenvalues and their corresponding eigenvectors
    Form matrix X by stacking the eigenvectors associated with the k smallest eigenvalues
    Normalize the rows of X to have unit length to obtain matrix Y
    Apply K-Means clustering on the rows of Y to identify k clusters
    Assign each original point x_i to a cluster based on the K-Means result
end function

Implementation Using Scikit-Learn

The practical implementation of Spectral Clustering was done using the `SpectralClustering` class from Python's `scikit-learn` library. Our approach included varying the number of clusters to best fit the energy consumption data:

```
from sklearn.cluster import SpectralClustering

    spectral = SpectralClustering(n_clusters=n_clusters,
                                  affinity='nearest_neighbors',
                                  n_neighbors=15,
                                  assign_labels="kmeans",
                                  random_state=42)
    cluster_assignments_spectral = spectral.fit_predict(data)
```

```
labels_spectral[n_clusters] = cluster_assignments_spectral
```

In this setup, the `affinity` parameter determines how the affinity matrix is computed, and `n_neighbors` is used in the nearest neighbor approach for affinity calculation. The `assign_labels` parameter, set to 'kmeans', indicates the method used for clustering in the reduced eigenvector space.

4.5.8 Ensemble Clustering

Ensemble Clustering received limited attention in the literature.We have chosen to investigate this approach and conduct a comparative analysis with the other unsupervised algorithms used in this study. Jain (2010) [48] classifies the techniques for creating multiple fundamental partitions for ensemble clustering into three distinct approaches:

(i) **Using Different Clustering Algorithms:** This approach involves applying various clustering algorithms to the same dataset. Each algorithm, due to its unique mechanism and principles, can provide different clustering results, contributing to the ensemble's diversity.

(ii) **Applying the Same Clustering Algorithm with Different Initializations:** In this method, the same clustering algorithm is run multiple times on the dataset, but with different initial conditions or parameters each time. This variability in initialization can lead to different clustering outcomes, which can be combined in the ensemble process.

(iii) **Running Clustering Algorithms on Different Feature Spaces:** This technique employs the same or different clustering algorithms on varied feature spaces of the dataset. By transforming or selecting different subsets of features, the algorithm can capture different aspects of the data, leading to diverse clustering solutions.

We followed Jain's first suggestion and used different clustering algorithms to advance our research. Our feature matrix was clustered using KMeans++, Hierarchical, and Birch. We then created a similarity matrix from these algorithms. We used this matrix to perform hierarchical clustering again and get an ensemble clustering result. Our methodology is described more explicitly below:

4.6 Evaluation Metrics

Evaluation metrics for clustering are crucial to understanding algorithm effectiveness. Internal and external validation methods exist. Internal validation metrics evaluate clustering structures without external data. However, external validation metrics require ground truth labels to assess clustering accuracy.

We use internal validation metrics because our study lacks ground truth data, as in unsupervised learning. These metrics are essential because they reveal the clustering

Algorithm 7 Ensemble Clustering Approach

Input: Data points, D
Output: Cluster labels, $cluster\_labels$
Initialize $n \leftarrow$ number of data points in D
Initialize co-association matrix $co\_association\_matrix$ as a zero matrix of size $n \times n$
for each clustering algorithm ($kmeans++$, $hierarchical$, $birch$) **do**
    Generate $labels$
    for $i \leftarrow 1$ to n **do**
        for $j \leftarrow 1$ to n **do**
            if $labels[i] == labels[j]$ **then**
                $co\_association\_matrix[i][j] \leftarrow co\_association\_matrix[i][j] + 1$
            end if
        end for
    end for
end for
Normalize $co\_association\_matrix$ by dividing each element by 3
Compute $distance\_matrix \leftarrow 1 - co\_association\_matrix$
Convert $distance\_matrix$ to condensed form $condensed\_distance\_matrix$
Perform hierarchical clustering on $condensed\_distance\_matrix$ using 'average' linkage method
Decide on the number of clusters, $num\_clusters \leftarrow 6$
Obtain final cluster labels $cluster\_labels$ using $fcluster$
return $cluster\_labels$

process intrinsically. They analyze the dataset and cluster compactness and separation to determine clustering formation goodness.

Internal validation metrics help compare clustering algorithms. They help us rank algorithms based on their ability to find meaningful data structures. By using multiple metrics, we can assess each algorithm's clustering quality strengths and weaknesses.

The following internal validation metrics were used to compare clustering algorithm performance:

- Silhouette Score (SI)
- Davies-Bouldin Score (DB)
- Calinski-Harabasz Score (CH)
- Dunn Index (DI)

Each of these metrics captures a different aspect of clustering quality, such as cluster cohesion, separation, and overall structure. The subsequent subsections will elaborate on the mathematical formulation of each metric and the rationale behind its use in our evaluation framework.

4.6.1 Silhouette Score

The Silhouette Score is a widely used internal metric for measuring the quality of a clustering. For each data point, the score is a comparative measure of how similar the

point is to its own cluster (cohesion) versus other clusters (separation). The score for a single data point i is calculated as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{4.16}$$

where:

- $a(i)$ is the mean intra-cluster distance, or the average distance between the i^{th} data point and all other points in its own cluster.
- $b(i)$ is the mean nearest-cluster distance, or the smallest mean distance from the i^{th} data point to points in a different cluster, minimized over all clusters.

The silhouette score for the entire dataset is the mean of all individual silhouette scores $s(i)$, and it ranges from -1 to 1. A high silhouette score close to 1 indicates that the data point is very similar to other points in its cluster and dissimilar to points of other clusters, suggesting that the clusters are well apart from each other and clearly defined. Conversely, a silhouette score close to -1 implies that the point is closer to points of neighboring clusters than to points in its own cluster, indicating poor clustering. Therefore, when comparing clustering algorithms, higher average silhouette scores across the dataset typically signify superior clustering performance.

4.6.2 Davies-Bouldin Score

The Davies-Bouldin Score (DB) is an internal evaluation metric that captures the average "similarity" between each cluster and its most closely adjacent cluster. Here, "similarity" is a function that compares the within-cluster cohesion against the between-cluster separation. The Davies-Bouldin Score is mathematically expressed as:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \left\{ \frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right\} \tag{4.17}$$

where:

- k represents the number of clusters.
- σ_i denotes the average distance of all points in cluster i to their centroid c_i, which measures the extent of dispersion within the cluster.
- $d(c_i, c_j)$ is the distance between centroids c_i and c_j, signifying the separation between clusters.

The DB score quantifies the ratio of the sum of within-cluster scatter to the separation between clusters. A lower Davies-Bouldin Score indicates a better clustering scheme where clusters are densely packed and well separated from each other. Thus, when assessing clustering results, one seeks to minimize the DB score as an objective criterion for quality.

4.6.3 Calinski-Harabasz Score

The Calinski-Harabasz Score, also known as the Variance Ratio Criterion, evaluates cluster validity based on the ratio of between-cluster variance to within-cluster variance. A higher ratio corresponds to clusters with better definition, that is, clusters that are more compact and well-separated from each other. The score is formally defined by the following equation:

$$CH = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \times \frac{N-k}{k-1} \tag{4.18}$$

In this equation:

- $\mathrm{Tr}(B_k)$ is the trace of the between-group dispersion matrix and measures the dispersion between the clusters. It is computed as the sum of squared distances from each cluster centroid to the overall centroid of the data, scaled by the size of the respective clusters.
- $\mathrm{Tr}(W_k)$ is the trace of the within-cluster dispersion matrix, which quantifies the dispersion of data points within each cluster. It is the sum of squared distances from each data point to its respective cluster centroid.
- N is the total number of data points.
- k is the number of clusters.

A high Calinski-Harabasz Score indicates that the clusters are dense and well-separated, which generally represents a better clustering structure. Consequently, when comparing different clustering results, models with higher Calinski-Harabasz Scores are preferred as they suggest more clearly defined clusters.

4.6.4 Dunn Index

The Dunn Index is an internal validation metric that assesses the quality of clustering by simultaneously considering the compactness within clusters and the separation between clusters. It is especially valuable for identifying sets of clusters that are both cohesive and distinct from each other. The Dunn Index for a set of clusters is mathematically defined as:

$$Dunn = \min_{1\leq i\leq k} \left\{ \min_{1\leq j\leq k, i\neq j} \left\{ \frac{d(i,j)}{\max_{1\leq l\leq k}\{diam(l)\}} \right\} \right\} \tag{4.19}$$

In this formula:

- $d(i,j)$ represents the inter-cluster distance, which is the distance between clusters i and j. This distance can be measured in various ways, such as the distance between cluster centroids or the closest points of the clusters.

- $diam(l)$ denotes the intra-cluster diameter, or the largest distance between any two points within cluster l. This measure reflects the spread or size of the cluster.

The Dunn Index is the ratio of the smallest inter-cluster distance to the largest intra-cluster diameter across all clusters. The goal is to maximize this ratio; thus, a higher Dunn Index is indicative of a clustering structure where the clusters are compact (small diameters) and well-separated (large inter-cluster distances). It is a direct measure of the overall clustering validity, with higher values suggesting a more robust partitioning of the data into distinct groups.

Dunn Index Calculation

The Dunn Index was calculated using a custom Python function, due to the absence of its implementation in `scikit-learn`. The code listing below demonstrates the computation of the Dunn Index:

```
from scipy.spatial.distance import cdist
from itertools import combinations
import numpy as np

def dunn(data, labels):
    unique_labels = np.unique(labels)
    # Compute inter-cluster distances
    min_intercluster_distance = float('inf')
    for cluster_i, cluster_j in combinations(unique_labels, 2):
        data_i = data[labels == cluster_i].to_numpy()
        data_j = data[labels == cluster_j].to_numpy()
        distances = cdist(data_i, data_j, 'euclidean')
        min_distance = distances.min()
        if min_distance < min_intercluster_distance:
            min_intercluster_distance = min_distance

    # Compute intra-cluster diameters
    max_intracluster_diameter = 0
    for cluster in unique_labels:
        data_cluster = data[labels == cluster].to_numpy()
        distances = cdist(data_cluster, data_cluster, 'euclidean')
        max_distance = distances.max()
        if max_distance > max_intracluster_diameter:
            max_intracluster_diameter = max_distance

    return min_intercluster_distance / max_intracluster_diameter
```

Listing 4.1: Python code for computing the Dunn Index

Chapter 5

Results

5.1 Introduction

We test several clustering algorithms on energy consumption data in this chapter. We used cluster numbers from 3 to 10 to rigorously test and compare each algorithm's ability to reveal energy usage patterns.

Our evaluation relies on carefully selected metrics to quantify these algorithms' energy data clustering performance. This method finds the most efficient algorithm and illuminates clustering configurations.

How well our analysis fits explainable AI principles is crucial. In the rapidly evolving field of artificial intelligence, understanding and trusting AI model decision-making is crucial. Our study follows this philosophy by presenting clustering results and explaining their rationale. Energy stakeholders who make decisions using AI-driven insights need this transparency.

Businesses, especially energy providers, need a manageable number of customer profiles to implement demand-response policies. Here, granularity and manageability must be balanced. A 6-cluster solution analyzed using the Ensemble Clustering algorithm is a viable and practical option. This focused approach makes clusters statistically sound, meaningful, and actionable for business.

We go beyond technical analysis of the 6-cluster model using ensemble clustering. It aims to explain and apply energy clustering AI. Each cluster is dissected to reveal its unique traits and behavior. These insights help energy providers improve demand-response policies, energy distribution, and customer engagement and satisfaction.

In conclusion, this chapter connects advanced clustering methods to energy management applications. It emphasizes the technical sophistication of algorithms and the importance of explainable AI in turning data-driven insights into energy management strategies.

5.2 Performance of the Algorithms

In this section, we present the performance evaluation of the implemented clustering algorithms through the use of Cluster Validity Indices (CVIs). The following subsections display and discuss the graphical representations of these indices for each clustering algorithm.

We will first examine the **silhouette score** of the algorithms. As was previously mentioned, clustering is improved as the silhouette score increases.The silhouette scores for clusters 3 through 9 for each of our testing algorithms are presented in the 5.1 diagram:

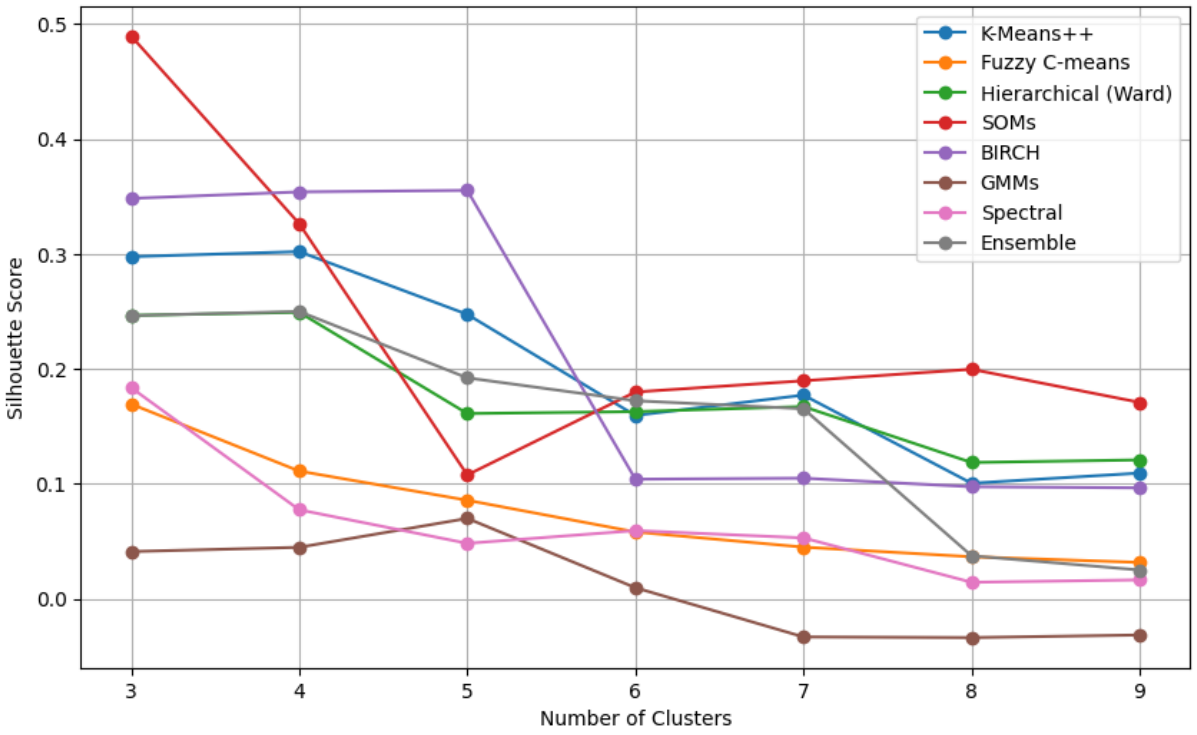


Figure 5.1: Comparison of Silhouette Scores across Clustering Algorithms for a Range of 3 to 9 Clusters

From the graph, key insights include:

- **Low Performers:** 'GMMs', 'Spectral', and 'Fuzzy C-Means' exhibit the lowest silhouette scores.
- **Stable Algorithms:** 'KMeans++' ,'Ensemble' and 'Hierarchical' maintain stable performance across varying cluster numbers, suggesting robustness, while 'SOMs' and 'Birch' are more unstable.
- **Ensemble Method:** The 'Ensemble' demonstrates steady results, underscoring its utility as a reliable clustering approach. However, for a small number of clusters it does not outperform the best-performing individual algorithms.

We should add here that it is a logical outcome for silhouette scores to generally decrease as the number of clusters increases due to the inherent increase in within-cluster dispersion and decrease in between-cluster separation.

Next CVI we will examine is the **Davies-Bouldin** and as we mentioned before a lower DB score indicates better clustering results. Figure 5.2 depicts the score of each algorithm for the range [3,9] of number of clusters again:

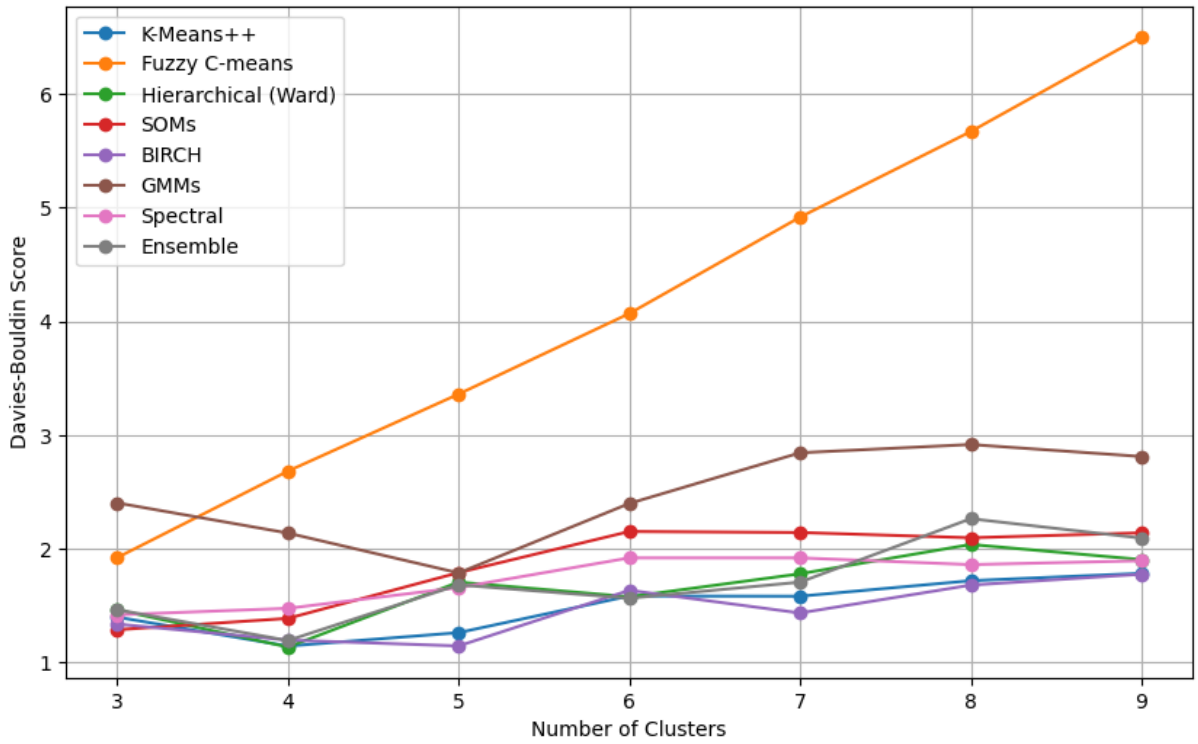


Figure 5.2: DB for different clustering algorithms across a range of cluster numbers.

Observations from the graph are as follows:

- **Low Performers:** 'Fuzzy C-Means', 'GMMs', and 'SOMs'. It is noteworthy that Fuzzy C-Means's DB score appears to increase linearly as the number of clusters increases.
- **Good Performers:** Birch now demonstrates good scores.The low Davies-Bouldin Score for BIRCH indicates that, despite the lack of cohesion within clusters (as reflected by the low Silhouette Score), the algorithm effectively separates the clusters from each other. In essence, BIRCH might be generating clusters that are not tight or cohesive (hence the low Silhouette Score), but are well-separated from each other (resulting in a low Davies-Bouldin Score).
- **Ensemble Method:** Overall steady behavior and similar to most of the other good-performing algorithms.

Moving forward we will examine the **Calinski-Harabasz** CVI. Below 5.3 are the clustering results:

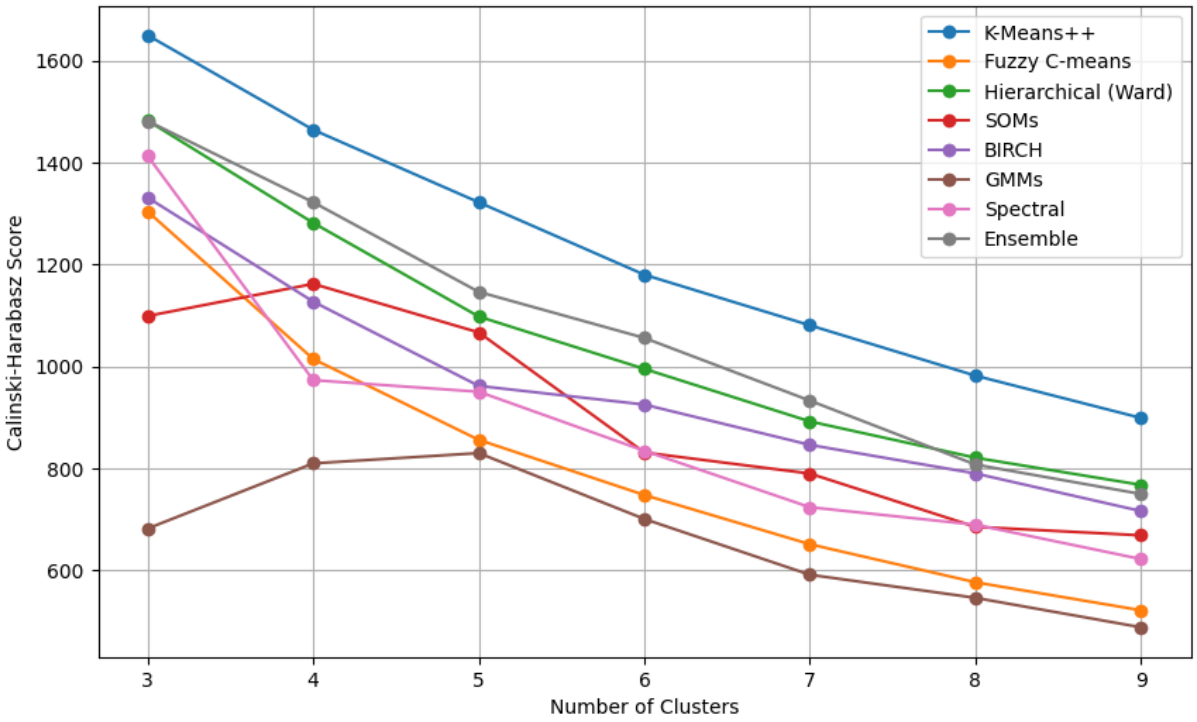


Figure 5.3: Calinski-Harabasz Score for different clustering algorithms across a range of cluster numbers.

The graph reveals several noteworthy observations:

- **Top Performer:** 'KMeans++' manages to maintain the highest scores across the range of clusters.
- **Low Performer:** 'GMMs' exhibit the lowest performance again indicating that they were not very suitable for our dataset.

Lastly, we examine the Dunn Index scores in 5.4 and we conclude the following:

- **Fluctuating Performance:** No single algorithm consistently maintains a high Dunn Index across the range of cluster numbers. Instead, each algorithm exhibits significant variability in performance as the number of clusters changes. Especially, Ensemble Clustering has a spike in 6 clusters.

The overall instability in Dunn Index scores across the algorithms and number of clusters underscores the complex nature of clustering. It suggests that there is not a one-size-fits-all approach to clustering and that algorithm performance can significantly depend on the chosen number of clusters. Furthermore, it highlights the importance of selecting the appropriate number of clusters for each algorithm and dataset to optimize clustering outcomes. The graph serves as a reminder of the need to combine multiple validity indices to comprehensively assess and understand the behavior of different clustering algorithms.

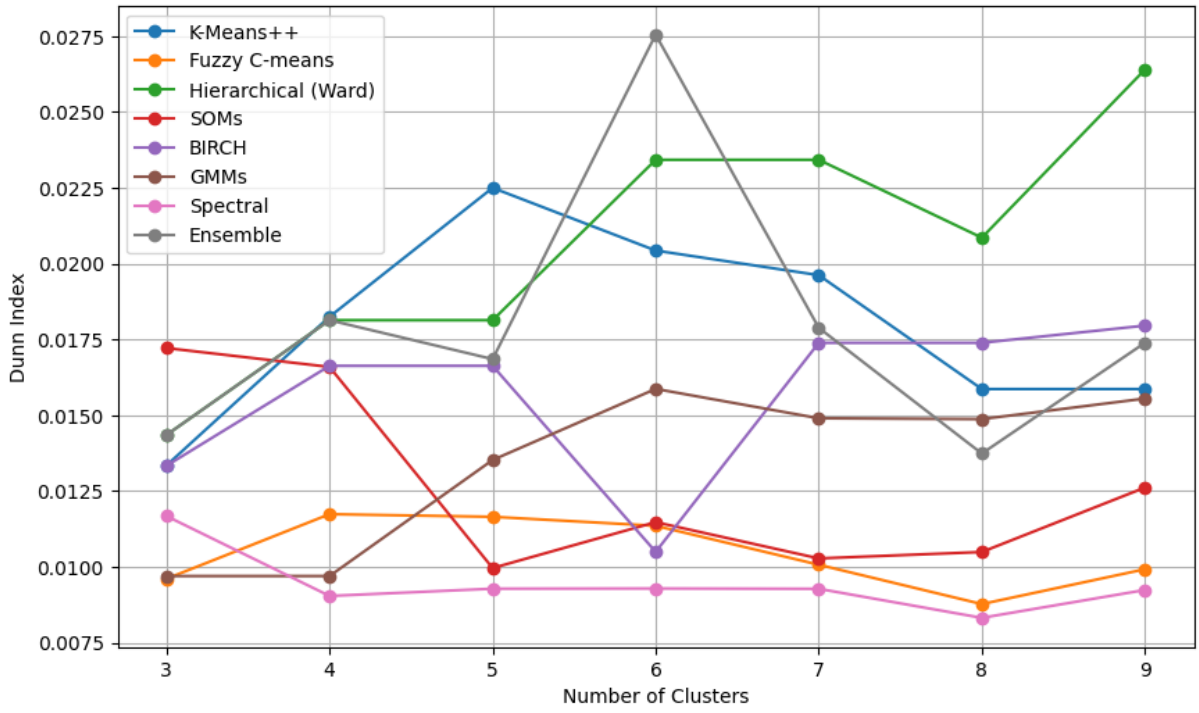


Figure 5.4: Dunn Index for different clustering algorithms across a range of cluster numbers.

The above graphs demonstrate a fundamental cluster analysis principle: **different algorithms are optimized for different clustering scenarios and perform better with specific cluster counts.** The optimal number of clusters for practical applications like utility DR is often determined by business considerations rather than unsupervised learning metrics. Clusters may match operational segments, customer categorizations, or strategic initiatives for utilities. After determining the desired number of clusters based on business needs, utility companies can use performance graphs like the one provided to choose the clustering algorithm with the best results. This ensures that the algorithm selected delivers actionable insights that match business goals and data structure.

Within the scope of this analysis, We chose a **six-cluster solution** as a prime example for this analysis. This cluster number will be used to define Profile Classes (PCs) that best represent the dataset's structure in DR strategies.

More specifically the CVIs results for 6 clusters are the following:

Table 5.1: Clustering Algorithm Performance Metrics for 6 Clusters

	Sil Score	DB Score	CH Score	Dunn Index
K-Means++	0.1594	1.5856	1,180.0197	0.0204
Fuzzy C-means	0.0580	4.0677	748.0819	0.0114
Hierarchical (Ward)	0.1627	1.5791	995.3937	0.0234
SOMs	0.1798	2.1524	831.4561	0.0115
BIRCH	0.1039	1.6358	925.4142	0.0105
GMMs	0.0095	2.3987	701.0690	0.0159
Spectral	0.0592	1.9206	834.7180	0.0093
Ensemble	0.1722	1.5660	1,056.0420	0.0275

The exact results-table for other number of clusters can be found in the Appendix in table A.1.

According to the table 5.1, the *best performorming algorithm* for 6 clusters is **Ensemble Clustering** which has the top score in 2 out of 4 CVIs and overall very close performance with the top-perfromer in the other 2 CVIs. Given its consistent and superior performance across multiple metrics, **Ensemble Clustering** is selected as the algorithm of choice for our cluster analysis.

5.3 Cluster Analysis Using Ensemble Clustering

In this section, we present a detailed analysis of the clusters obtained through the Ensemble algorithm. We begin by examining the sizes of the clusters, which provide initial insight into the distribution of data points among the identified groups. Subsequently, we utilize visualization techniques such as 3D plots, Principal Component Analysis (PCA), and t-Distributed Stochastic Neighbor Embedding (t-SNE) to facilitate a more intuitive understanding of the clustering results.

Cluster Sizes

The distribution of data points among the clusters is a fundamental characteristic that can shed light on the underlying structure of the data. The sizes of the clusters obtained from the Ensemble algorithm are shown in 5.5:

Figure 5.5: Cluster Sizes

The cluster size bar chart shows a large data point distribution variance.Cluster 4

stands out with a notably high count of 3107 consumers, suggesting it represents the most common energy usage behavior among the dataset. In stark contrast, Clusters 3 and 5 contain markedly fewer consumers, with just 5 and 12 respectively, indicating these clusters represent niche or atypical consumption patterns. Clusters 1, 2, and 6 exhibit more moderate sizes, with Cluster 2 being the second most populous cluster, containing 1123 consumers. The substantial difference in cluster sizes suggests a diverse range of energy usage behaviors, with a dominant pattern captured by Cluster 4 and more unique or irregular profiles characterized by Clusters 3 and 5.

Visualizing Clustering Results

We will display clustering results using multiple visualization methods in the next section. **Principal Component Analysis (PCA)** is a highly regarded statistical technique for reducing dimensionality. A large dataset of potentially correlated variables is transformed into a structured set of linearly uncorrelated principal components by PCA [49]. This method reduces data dimensionality, allowing a two-dimensional or three-dimensional visualization that captures most of its variance. Our clustering results are presented in a two-dimensional PCA visualization in figure 5.6

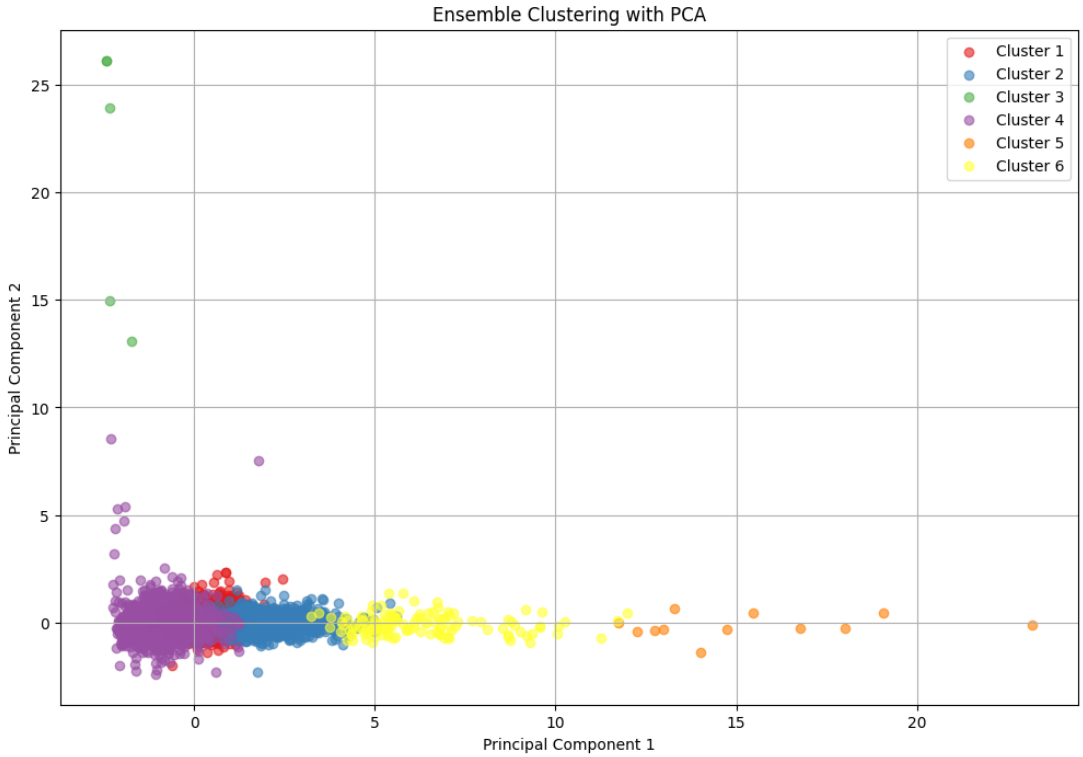


Figure 5.6: Clusters visualisation with PCA

We also created a **3D plot** to visualize our clustering analysis and explore data relationships. We chose 'Consumption_Growth' for the z-axis, 'avg_peak_consumption' for the y-axis, and 'median_consumption' for the x-axis. These characteristics were chosen to describe consumption patterns. Other analyses may prioritize different features based on research goals. The 3D visualization shows how data clusters around these features in 5.7.

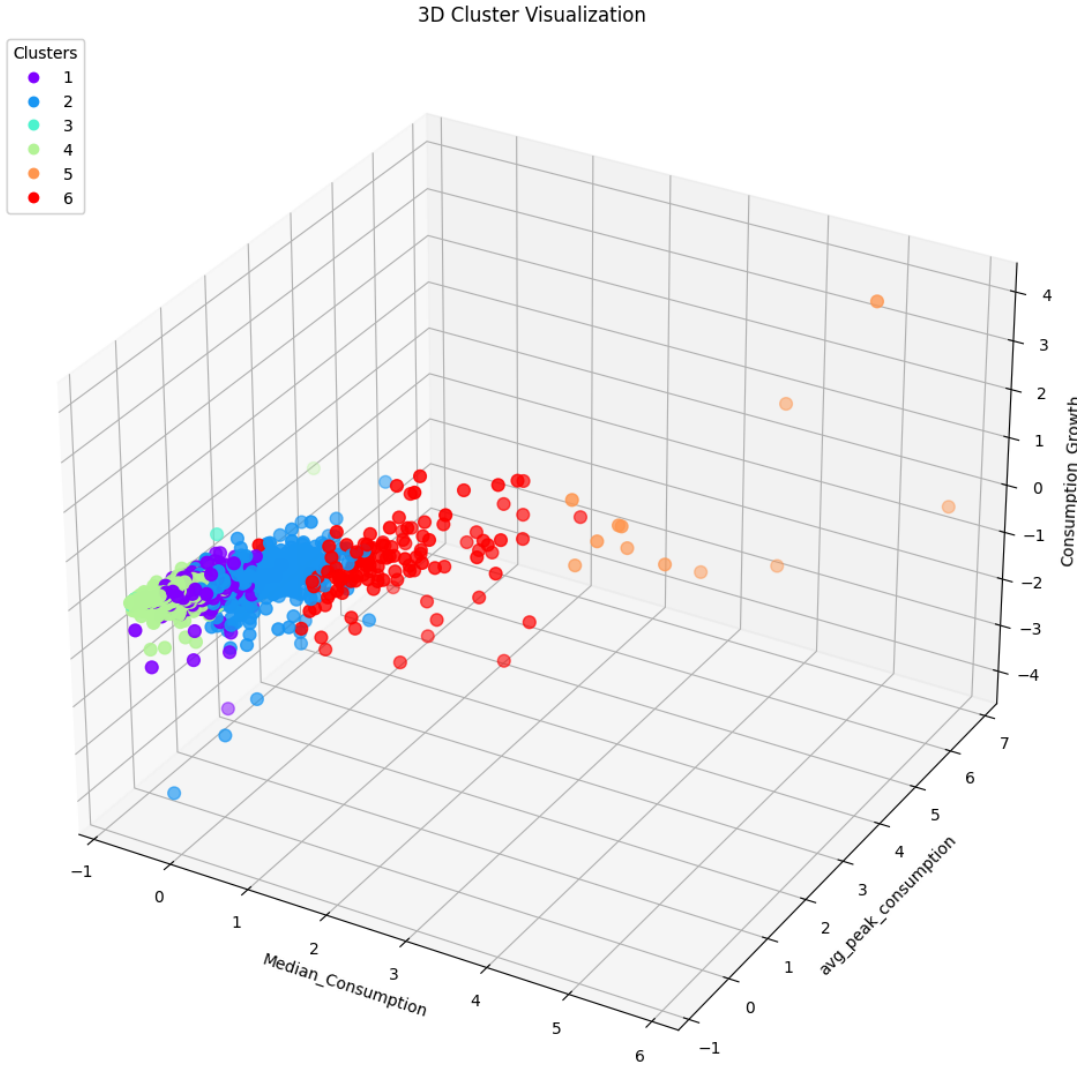


Figure 5.7: 3D Clusters visualisation

We concluded our visualization with **t-Distributed Stochastic Neighbor Embedding (t-SNE)** for its effectiveness in high-dimensional data visualization. t-SNE allows us to observe data structure in a reduced two-dimensional space, revealing clusters and patterns that are not immediately apparent in higher dimensions. Employing the following parameters: perplexity set to 30, iteration count at 300, and a fixed random state for reproducibility, we performed t-SNE on the scaled dataset and can be seen in figure 5.8.

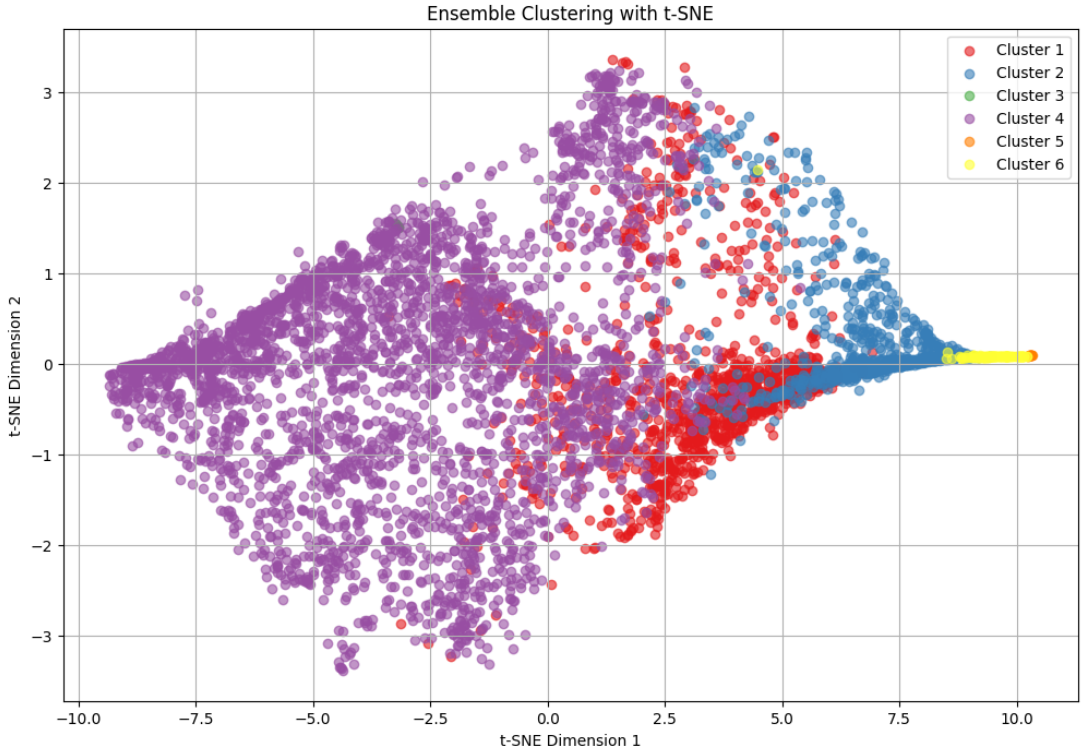


Figure 5.8: Clusters visualisation with t-SNE

5.3.1 Load shape analysis

As the next step in our analysis, we will present the load shapes characteristic of each cluster. These load shapes encapsulate the representative load profiles (RLPs) for all households within a given cluster. By aggregating the individual load curves of the households and computing the median RLP for each cluster, we obtain a cluster-specific load shape. It provides a visual summary of the typical energy usage behavior within each cluster, which can be a valuable asset for designing targeted DR strategies and tailoring energy services to meet specific customer needs. The figures 5.9,5.10,5.11,5.12,5.13,5.14 are with the non-normalized values.

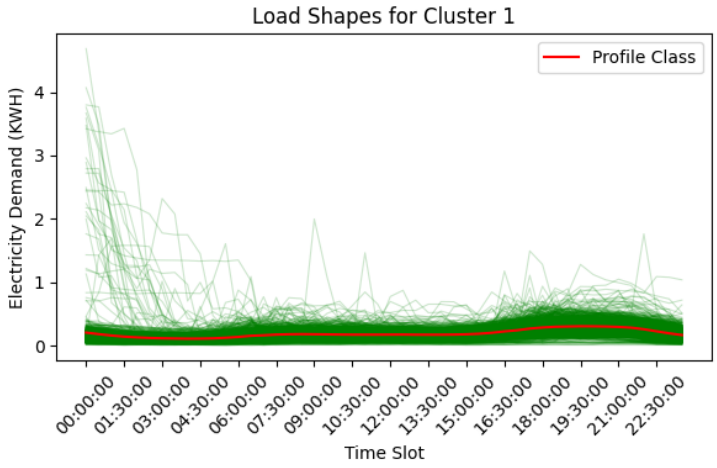


Figure 5.9: Individual and Mean Residential Load Profiles for Cluster 1

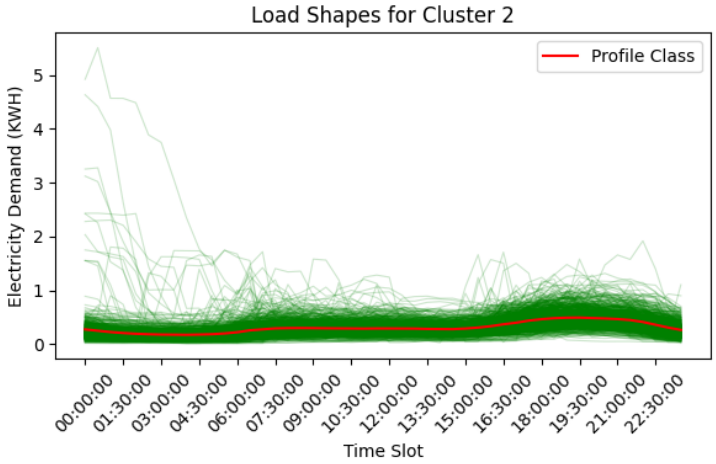


Figure 5.10: Individual and Mean Residential Load Profiles for Cluster 2

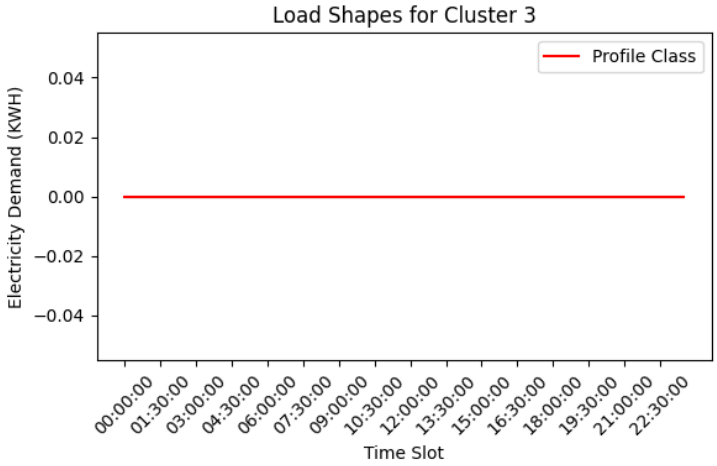


Figure 5.11: Individual and Mean Residential Load Profiles for Cluster 3

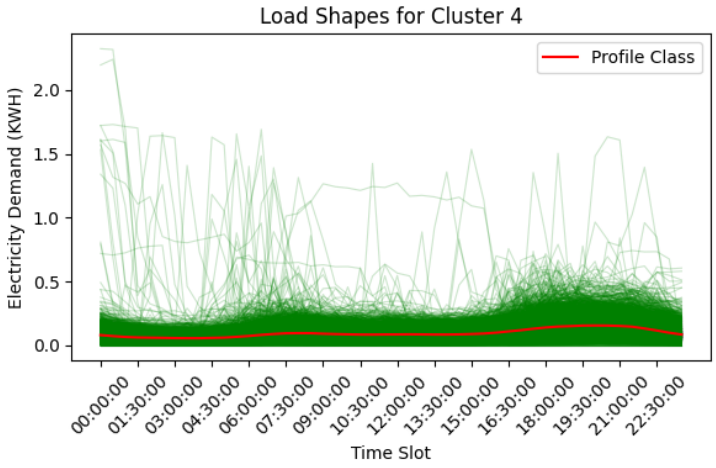


Figure 5.12: Individual and Mean Residential Load Profiles for Cluster 4

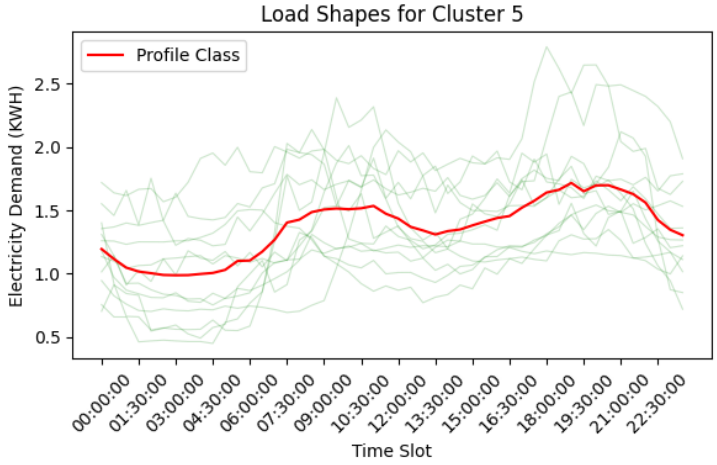


Figure 5.13: Individual and Mean Residential Load Profiles for Cluster 5

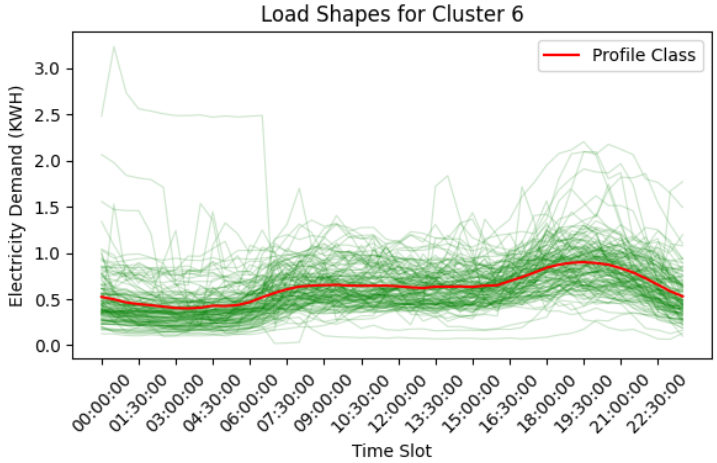


Figure 5.14: Individual and Mean Residential Load Profiles for Cluster 6

We also plot the representative load profile for all clusters in one diagram 5.15 so that the scales of consumption would be comparable among clusters:

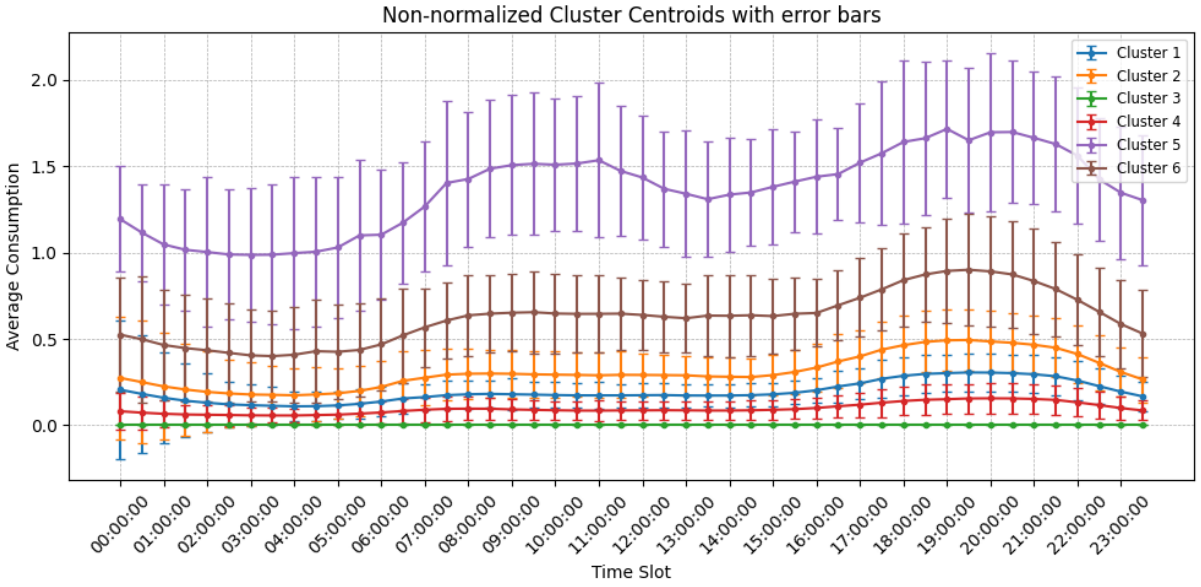


Figure 5.15: Cluster Representatives Load Profiles

It is evident that the clustering algorithms have partitioned the consumers into distinct levels of consumption:

- **Clusters 1** (Blue Line) and **4** (Red Line): These clusters show the lowest consumption levels among all, indicating either energy-efficient households or small-scale consumers.
- **Cluster 2** (Orange Line): Presents a moderate consumption level.
- **Cluster 3** (Green Line): Shows zero consumption, indicated by the flat line at the bottom of the chart. This could represent unoccupied properties(vacant homes or unused office spaces).
- **Cluster 5** (Purple Line): This cluster stands out with the highest energy consumption throughout the day, which could represent large commercial spaces or industrial operations with high energy demands.
- **Cluster 6** (Brown Line): Displays consistently high consumption with less variability than Cluster 5. This pattern might be associated with high-occupancy residential buildings or businesses with steady energy needs.

The above plots do not exhibit any discernible patterns. As stated by [25], normalization is crucial in order to uncover more distinct household patterns.Thus, we will now present the updated, refined normalized load shapes for each cluster in 5.16,5.17,5.18,5.19,5.20,5.21.

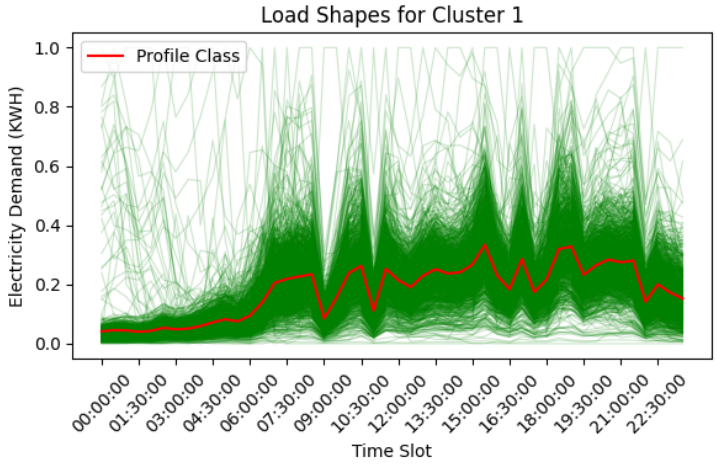


Figure 5.16: Cluster 1 Load Shapes with normalized values

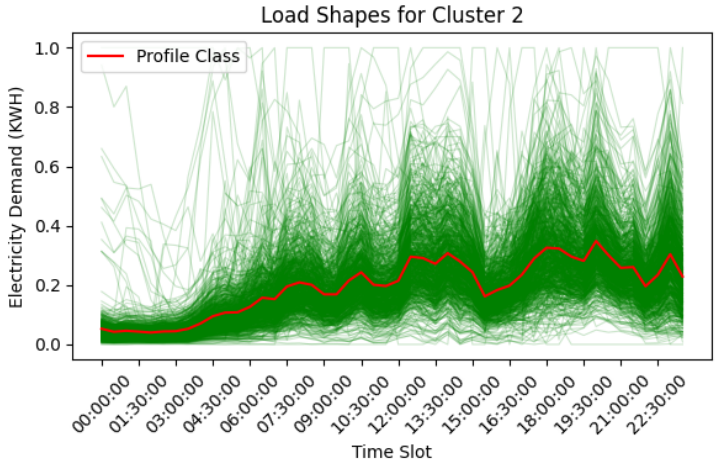


Figure 5.17: Cluster 2 Load Shapes with normalized values

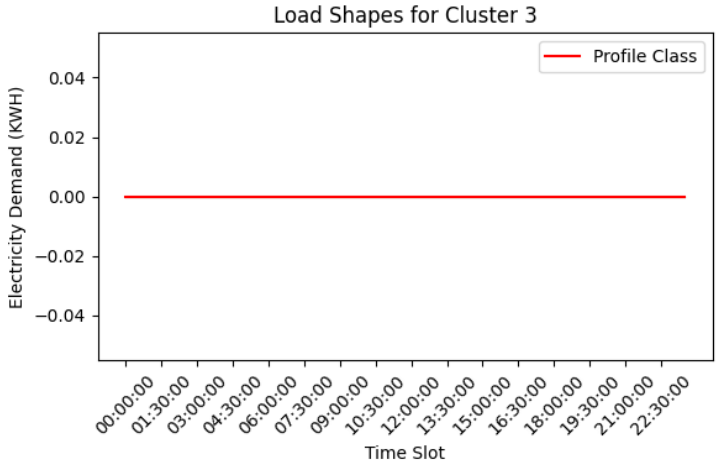


Figure 5.18: Cluster 3 Load Shapes with normalized values

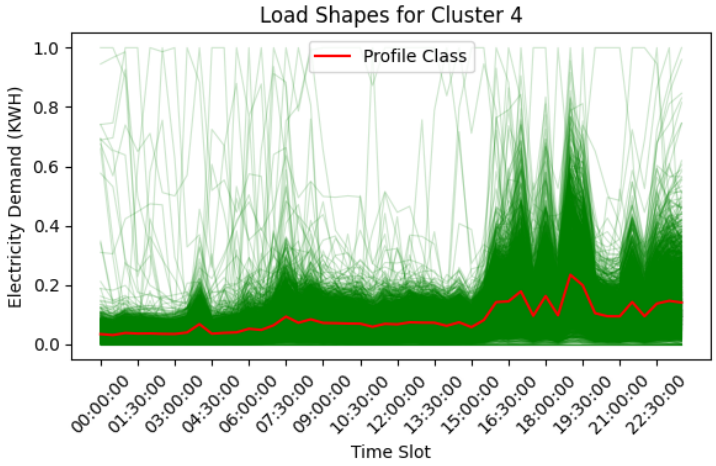


Figure 5.19: Cluster 4 Load Shapes with normalized values

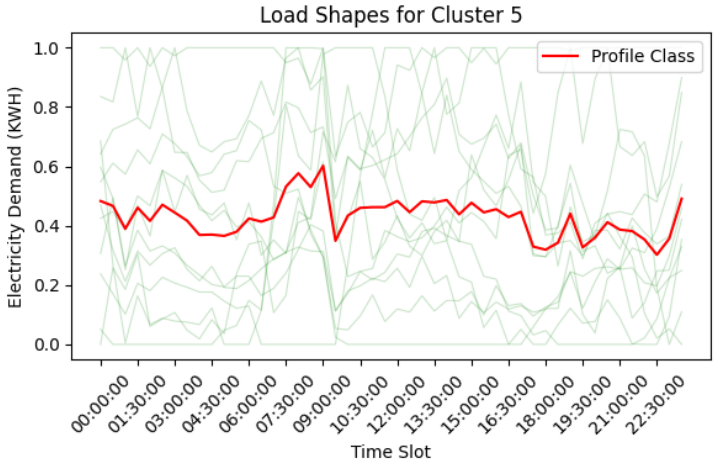


Figure 5.20: Cluster 5 Load Shapes with normalized values

Figure 5.21: Cluster 6 Load Shapes with normalized values

And the centroids of the normalized values are depicted in 5.22.

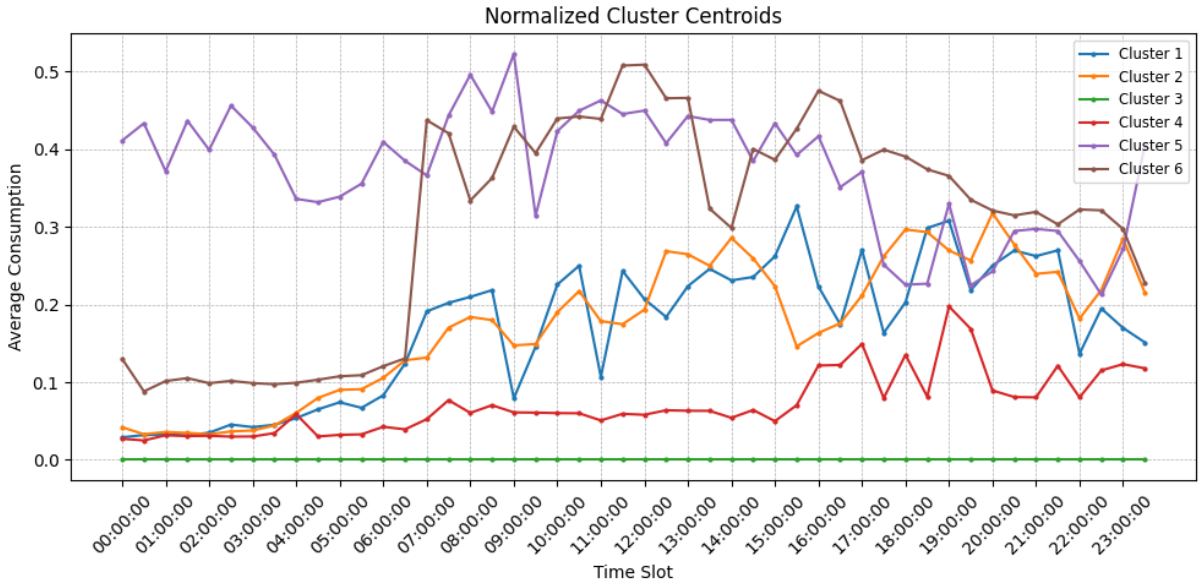


Figure 5.22: Cluster Representatives Normalized Load Profiles

- **Cluster 1** (Blue Line): Displays consistent energy use throughout the day, suggesting occupants are likely at home, using energy fairly regularly.

- **Cluster 2** (Orange Line): Dual peak pattern, with morning and evening spikes, typical of residential routines.

- **Cluster 3** (Green Line): Flat baseline, zero consumption, potentially empty properties.

- **Cluster 4** (Red Line): Flat usage with evening peak, possibly indicating households that go to work in the morning and return home in evening , which is the most common energy consumption pattern.

- **Cluster 5** (Purple Line): Presents a very high level of consumption as we seen before, and a decline in the after-work hours suggesting these could be commercial spaces or offices that also are running some equipment during the night.

- **Cluster 6** (Brown Line): This pattern shows a sharp rise in the morning, a plateau during working hours, and a decline after work, which, along with high consumption levels, may represent commercial spaces or offices.

5.3.2 Explainable AI

Explainable AI (XAI) methods helped us understand why Clusters were clustered separately. We used the SHAP (SHapley Additive exPlanations) library, a cutting-edge XAI tool that uses game theory to explain machine learning model output. The SHAP library helps us understand the Ensemble algorithm's rationale by showing how each feature affects clustering. SHAP value analysis of our clustering model reveals the features that most influenced clusters' separation. This method improves clustering interpretability and helps us identify distinguishing traits between similar clusters.

Figure 5.23: SHAP Values for **Cluster 1**

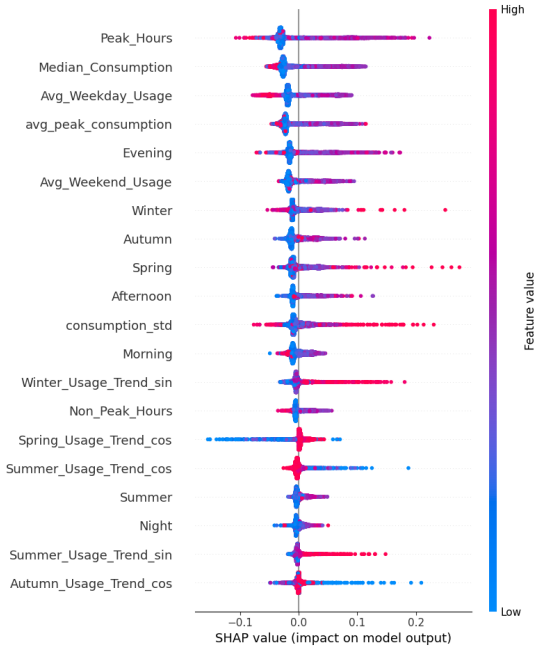
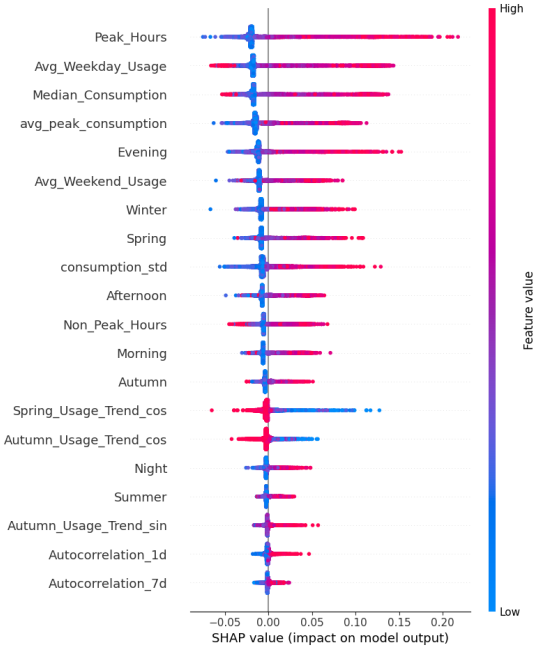


Figure 5.24: SHAP Values for **Cluster 2**



For Cluster 1 and Cluster 2, while they present a similar scale of consumption and appear to exhibit related behavior patterns at first glance, the SHAP values allow us to discern more nuanced differences between them.

Cluster 1's energy consumption is characterized by steady, consistent use, with significant features such as Peak Hours and Median Consumption indicating a regular and predictable pattern that doesn't change much from day to day, as evidenced by the absence of autocorrelation features.

Cluster 2, although similar in overall consumption levels, shows evidence of being influenced by recent consumption patterns (albeit slightly), as indicated by the presence (but low impact) of autocorrelation features in the SHAP plot. Additionally, the Average Weekday Usage has a more uniform and pronounced effect across instances, suggesting a more routine-based consumption with noticeable morning and evening peaks typical of residential lifestyles.

Figure 5.25: SHAP Values for **Cluster 3**

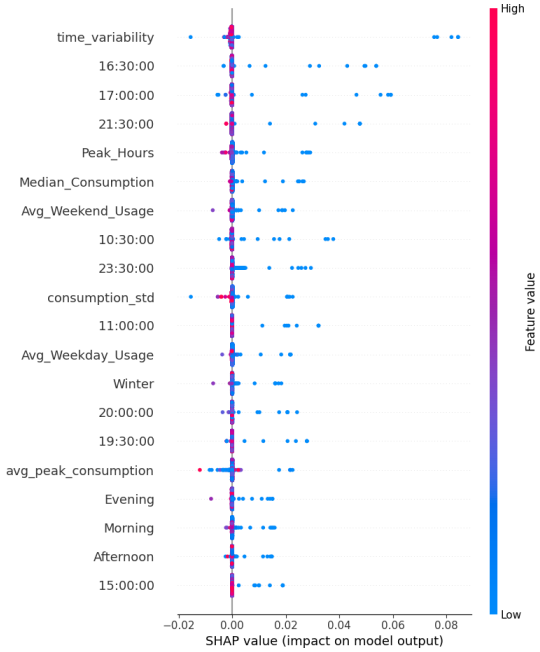


Figure 5.26: SHAP Values for **Cluster 4**

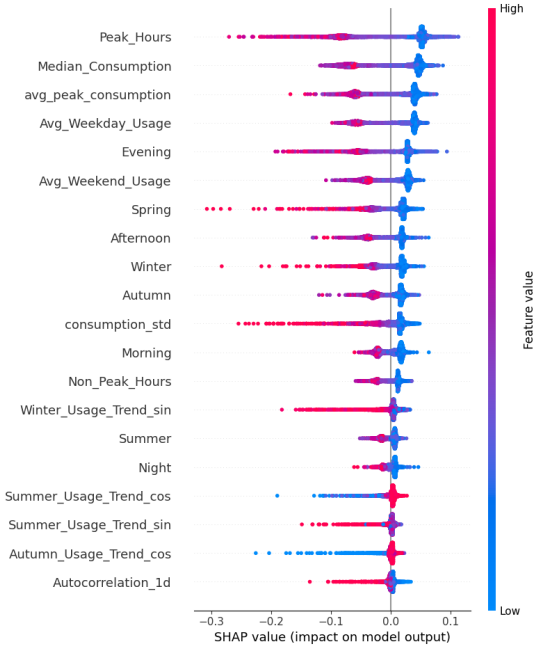


Figure 5.27: SHAP Values for **Cluster 5**

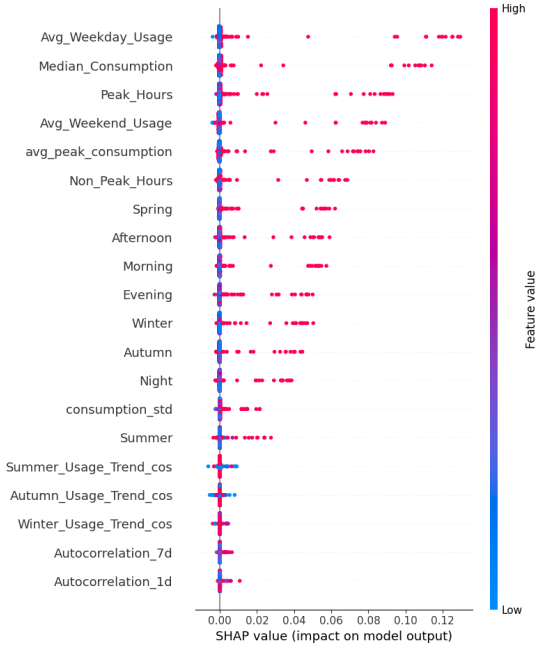
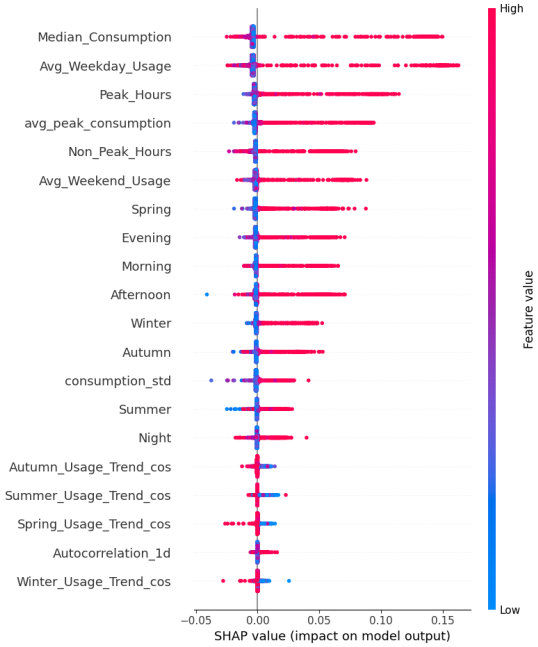


Figure 5.28: SHAP Values for **Cluster 6**



The high consumption levels in both clusters align with the expected energy usage profiles of commercial spaces or offices. However, the nuanced differences in feature impacts between Cluster 5 and Cluster 6, as indicated by their respective SHAP plots, offer a more refined understanding of their consumption behaviors. For instance, Cluster 5 might represent businesses that continue to consume energy into the evening, while Cluster 6 may include operations with a clear delineation between working and non-working hours. These insights could guide energy management strategies, such as targeted energy efficiency measures for after-work hours in Cluster 5 and DR programs during peak hours for Cluster 6.

5.3.3 Weekend - Weekday load profiles

Lastly, we must distinguish weekday and weekend energy consumption patterns. Weekend and weekday lifestyles and occupancy patterns can significantly affect energy usage. Thus, we will show weekday and weekend load profiles for each cluster. The profile for Cluster 1 is depicted in Figure 5.29, Cluster 2 in Figure 5.30, Cluster 3 in Figure 5.31, Cluster 4 in Figure 5.32, Cluster 5 in Figure 5.33, and Cluster 6 in Figure 5.34.

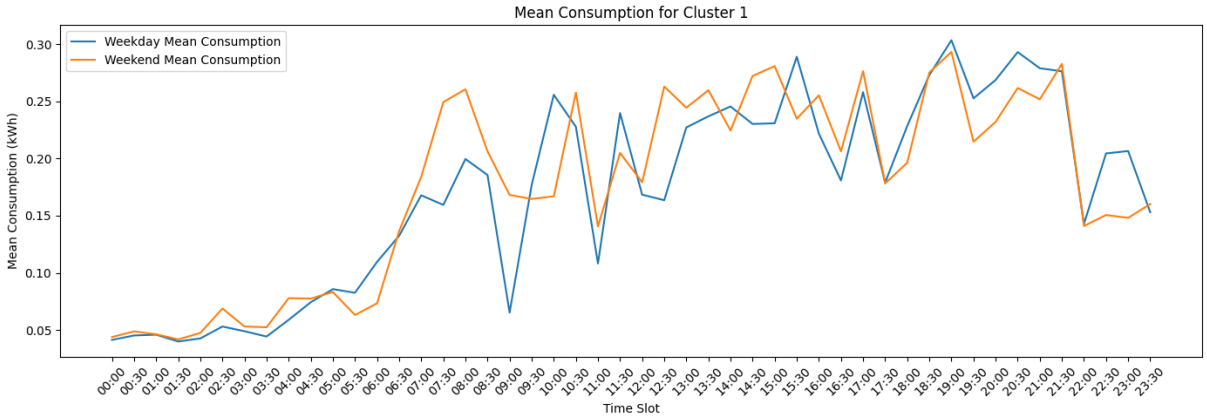


Figure 5.29: Weekend and weekday profile for cluster 1

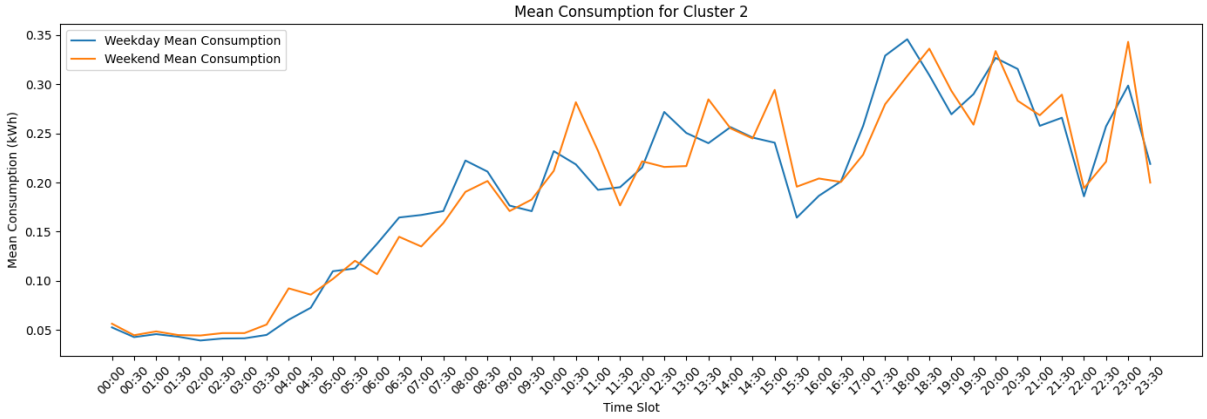


Figure 5.30: Weekend and weekday profile for cluster 2

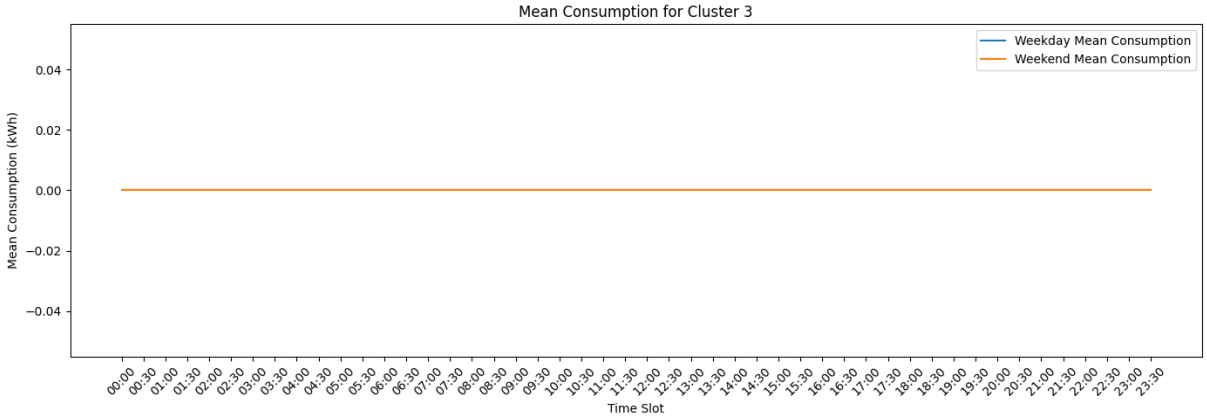


Figure 5.31: Weekend and weekday profile for cluster 3

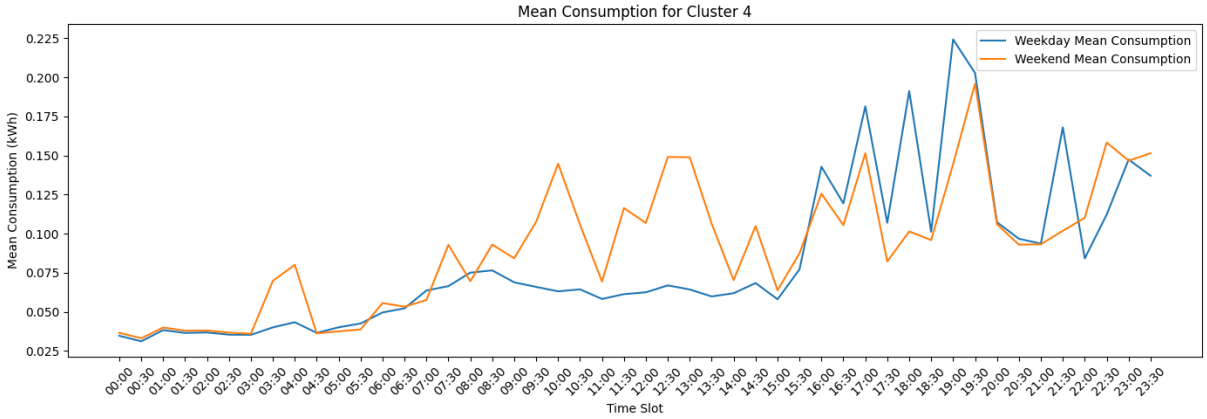


Figure 5.32: Weekend and weekday profile for cluster 4

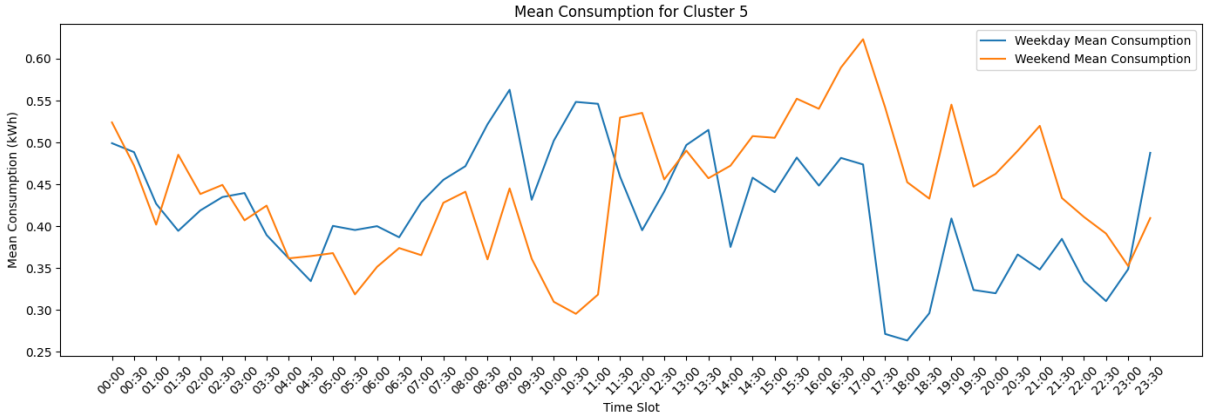


Figure 5.33: Weekend and weekday profile for cluster 5

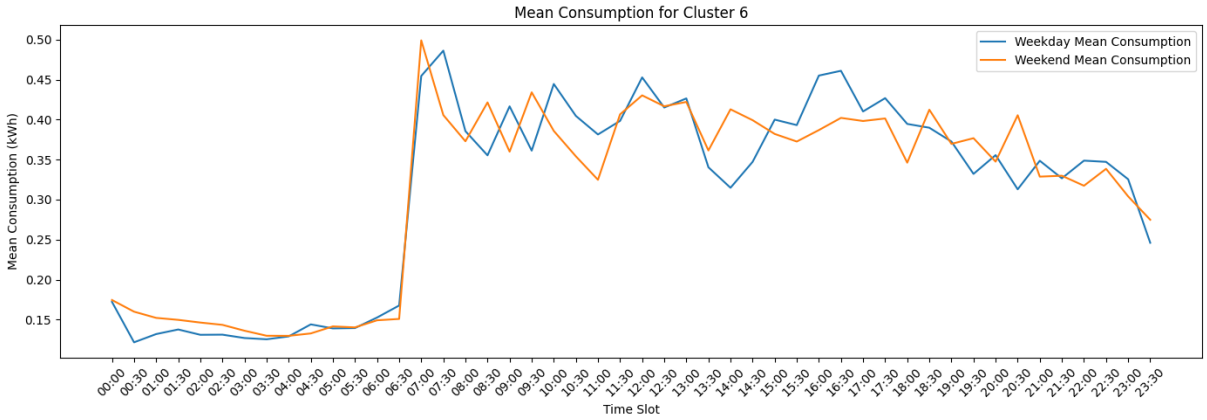


Figure 5.34: Weekend and weekday profile for cluster 6

Chapter 6

Conclusion - Future Prospects

6.1 Conclusion

Our thorough assessment of clustering algorithms on consumer energy usage data has revealed the inconsistency in algorithm performance based on the selected number of clusters. The results of our study suggest that certain algorithms, such as K-means++ and Ensemble clustering, consistently demonstrate strong performance. However, the effectiveness of other algorithms varies when considering different cluster counts. The effectiveness of ensemble clustering is noteworthy, as it consistently achieves mooderate-to-high rankings. However, it does not always surpass the performance of top-performing individual algorithms such as K-means++.

Our further investigation has successfully categorized consumers into distinct Load Profile Classes, primarily based on their overall energy consumption levels—low, medium, and high. We have identified outlier clusters that exhibit abnormal consumption patterns.

Contrary to our initial observations, which indicated that clusters were mainly separated based on consumption levels, the normalization process uncovered more complex consumption patterns within each cluster. The importance of normalization in energy consumption studies is emphasized, as it enables a more precise comparison of usage patterns among various consumer groups.

Moreover, in instances where clusters demonstrated closely correlated behaviors, XAI offered elucidation on the nuanced characteristics that differentiate one cluster from another. For instance, although certain clusters showed similarities in terms of the amount of consumption, XAI uncovered distinct characteristics, such as the timing of the highest usage or the regularity of consumption, that distinguished them from each other.

To summarize, this study emphasizes the importance of choosing suitable clustering algorithms that are specifically designed for the unique features of energy consumption data. The significance of Ensemble clustering in our analysis underscores their effectiveness in identifying distinct consumer segments. Furthermore, the utilization of XAI has demonstrated itself to be a fundamental element in improving the comprehensibility of clustering results, providing deep understanding into consumer energy behaviors. This study promotes the careful implementation of normalization and feature selection, supported by XAI, to enhance the accuracy of energy management and customer engagement strategies. This goes beyond simply categorizing consumers based on their consumption levels, and instead aims to uncover the intricate patterns that characterize each consumer segment.

6.2 Future Work

As we consider future developments in the field of energy consumers segmentation, some key areas emerge as promising avenues for research and methodology enhancement:

Correlation with Socio-Demographic Features: Future research should also consider the correlation between energy consumption patterns and socio-demographic characteristics of households. Understanding how factors like household size, income level, geographical location, and lifestyle choices impact energy usage could provide deeper insights into consumer behavior. This approach could aid in developing more personalized energy-saving strategies and tailoring communication to different demographic segments.

Expanded Application of Explainable AI (XAI): Another promising direction is the further utilization of Explainable AI. XAI tools, such as the SHAP (SHapley Additive exPlanations) library, have already proven valuable in our current analysis. Going forward, a deeper integration of XAI methodologies could provide greater insights into the factors influencing clustering outcomes. This increased transparency in AI models will not only aid in understanding complex consumer behaviors but also in validating and improving the models themselves. XAI's potential to reveal intricate relationships within data can lead to more targeted and effective energy management strategies, tailored to the specific needs and patterns of different consumer segments.

Dimensionality Reduction and Data Size Optimization: An important aspect of managing vast datasets is the efficient reduction of data dimensionality without compromising the integrity of the underlying patterns. Techniques such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and autoencoders have shown promise in other domains. Evaluating the effectiveness of these methods in the context of energy consumption data could lead to more refined clustering without the burden of processing and analyzing large volumes of data.

Appendix A

The purpose of Appendix A in this document is to provide a detailed expansion on the main analysis, specifically focusing on the intricacies of clustering algorithms used for analyzing consumer energy usage data. The analysis involves a comprehensive assessment of the performance metrics for different clustering algorithms across a spectrum of cluster solutions, ranging from 3 to 9 clusters. This provides a thorough evaluation of the effectiveness of each algorithm. Furthermore, this appendix offers comprehensive visual representations for the six-cluster solution, shedding light on the unique consumption patterns and characteristics of each identified consumer cluster. This additional section is intended for readers who desire a comprehensive quantitative comprehension and aims to strengthen the reliability and clarity of the research methodology and findings presented in the main part of the study.

A.1 Detailed Clustering Algorithms Performance

This section presents a comprehensive tabular summary of the performance metrics for a range of clustering algorithms applied to our consumer energy usage dataset. The table A.1 spans cluster solutions from 3 to 9 clusters, providing a broad perspective on how each algorithm scales with increasing cluster numbers.

The performance metrics encapsulated in the table include the Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Score, and the Dunn Index. These metrics collectively offer insights into the compactness, separation, and overall suitability of the clustering solutions generated by each algorithm.

The detailed performance table thus serves as a crucial tool for selecting the most appropriate clustering algorithm for specific energy consumer segmentation tasks. It also underpins the necessity for rigorous evaluation of clustering outcomes to ensure meaningful and actionable segmentation.

Clusters	Algorithm	Sil Score	DB Score	CH Score	Dunn Index
3	**K-Means++**	0.2975	1.3997	1649.5400	0.0133
	Fuzzy C-means	0.1693	1.9228	1303.8637	0.0096
	Hierarchical (Ward)	0.2464	1.4638	1482.0378	0.0143
	SOMs	0.4895	1.2893	1099.1569	0.0172
	BIRCH	0.3482	1.3389	1331.2656	0.0133
	GMMs	0.0410	2.4035	682.3985	0.0097
	Spectral	0.1836	1.4205	1412.7718	0.0117
	Ensemble	0.2463	1.4652	1481.4723	0.0143
4	**K-Means++**	0.3019	1.1458	1464.2405	0.0182
	Fuzzy C-means	0.1110	2.6827	1014.7781	0.0117
	Hierarchical (Ward)	0.2491	1.1371	1281.6889	0.0181
	SOMs	0.3262	1.3890	1162.1289	0.0166
	BIRCH	0.3538	1.1984	1127.2532	0.0166
	GMMs	0.0447	2.1385	810.1209	0.0097
	Spectral	0.0773	1.4772	973.3099	0.0090
	Ensemble	0.2498	1.1942	1322.1354	0.0181
5	**K-Means++**	0.2473	1.2637	1322.0442	0.0225
	Fuzzy C-means	0.0856	3.3597	856.2704	0.0116
	Hierarchical (Ward)	0.1610	1.7075	1098.0232	0.0181
	SOMs	0.1076	1.7909	1066.8957	0.0099
	BIRCH	0.3552	1.1469	962.0519	0.0166
	GMMs	0.0698	1.7884	830.2905	0.0135
	Spectral	0.0481	1.6599	950.6170	0.0093
	Ensemble	0.1921	1.6833	1146.4227	0.0168
6	**K-Means++**	0.1594	1.5865	1180.0833	0.0204
	Fuzzy C-means	0.0581	4.0676	748.1149	0.0113
	Hierarchical (Ward)	0.1627	1.5790	995.4433	0.0234
	SOMs	0.1798	2.1523	831.4943	0.0115
	BIRCH	0.1039	1.6357	925.4578	0.0105
	GMMs	0.0096	2.3986	701.0983	0.0158
	Spectral	0.0592	1.9206	834.7551	0.0093
	Ensemble	0.1722	1.5660	1056.0957	0.0275
7	**K-Means++**	0.1771	1.5837	1080.8693	0.0196
	Fuzzy C-means	0.0448	4.9141	651.5604	0.0101
	Hierarchical (Ward)	0.1671	1.7806	892.3846	0.0234
	SOMs	0.1896	2.1436	790.2138	0.0103
	BIRCH	0.1048	1.4362	846.3774	0.0174
	GMMs	-0.0333	2.8439	591.7261	0.0149
	Spectral	0.0529	1.9206	724.1489	0.0093
	Ensemble	0.1652	1.7075	933.1757	0.0179
8	**K-Means++**	0.1004	1.7201	982.1360	0.0158
	Fuzzy C-means	0.0365	5.6663	576.9787	0.0087
	Hierarchical (Ward)	0.1185	2.0386	821.2460	0.0208
	SOMs	0.1996	2.0972	685.6861	0.0105
	BIRCH	0.0974	1.6816	790.3500	0.0174
	GMMs	-0.0339	2.9167	546.4884	0.0148
	Spectral	0.0143	1.8613	689.7913	0.0083
	Ensemble	0.0373	2.2647	807.7934	0.0137
9	**K-Means++**	0.1093	1.7845	899.1727	0.0158
	Fuzzy C-means	0.0315	6.5000	522.0603	0.0099
	Hierarchical (Ward)	0.1207	1.9054	767.9334	0.0264
	SOMs	0.1708	2.1405	669.0390	0.0126
	BIRCH	0.0964	1.7770	716.6024	0.0179
	GMMs	-0.0316	2.8111	488.6858	0.0155
	Spectral	0.0163	1.8951	622.4718	0.0092
	Ensemble	0.0250	2.0941	750.0807	0.0174

Table A.1: Clustering Algorithm Performance Metrics for 3 to 9 Clusters

A.2 Extended Clustering Results for Six-Cluster Solution

Within this subsection, we explore a seasonal analysis for each of the six clusters that have been identified in our study. This analysis aims to compare the mean energy consumption patterns of each cluster during various seasons, including Spring, Summer, Autumn, and Winter. Through the analysis of these profiles in an non-normalized format, our objective is to reveal the influence of seasonal variations on the energy consumption patterns of various consumer segments.

For each cluster, the seasonal analysis will provide:

- **Spring Profile:** Understanding how the onset of warmer weather affects energy consumption, particularly in clusters that may indicate residential or commercial spaces.
- **Summer Profile:** Insights into the peak energy usage during the hottest part of the year, which can be critical for managing energy demands and optimizing grid performance.
- **Autumn Profile:** Observations on the transitional energy usage patterns as temperatures begin to drop, potentially affecting heating and lighting needs.
- **Winter Profile:** Analysis of energy consumption during the coldest season, which is often associated with increased heating demands.

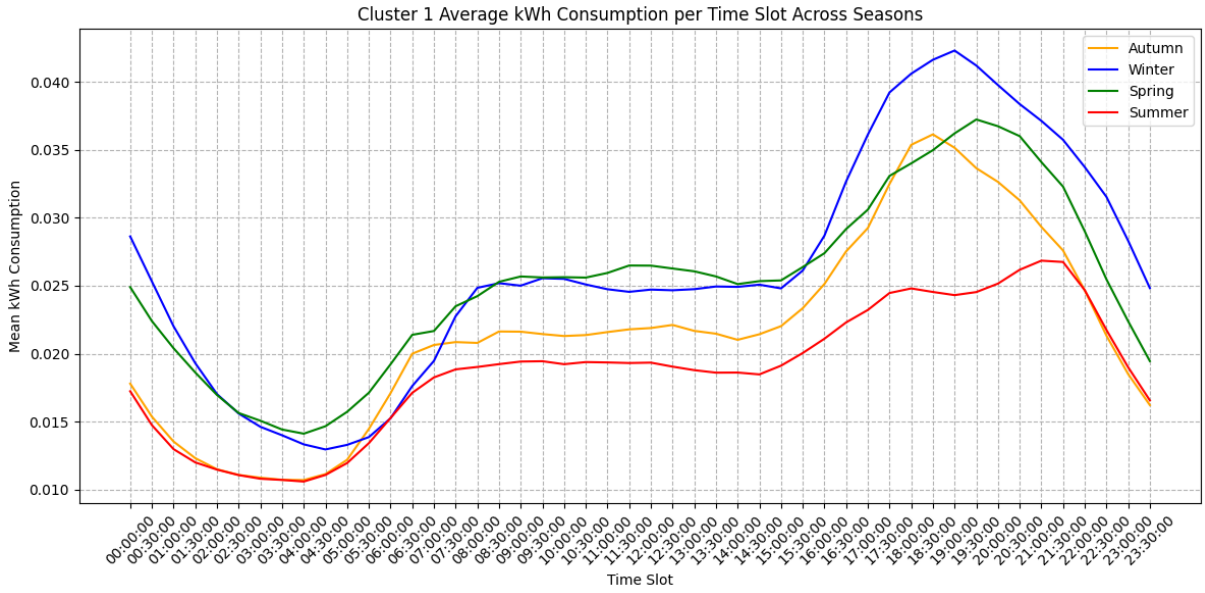


Figure A.1: Seasonal Consumption Profile for Cluster 1

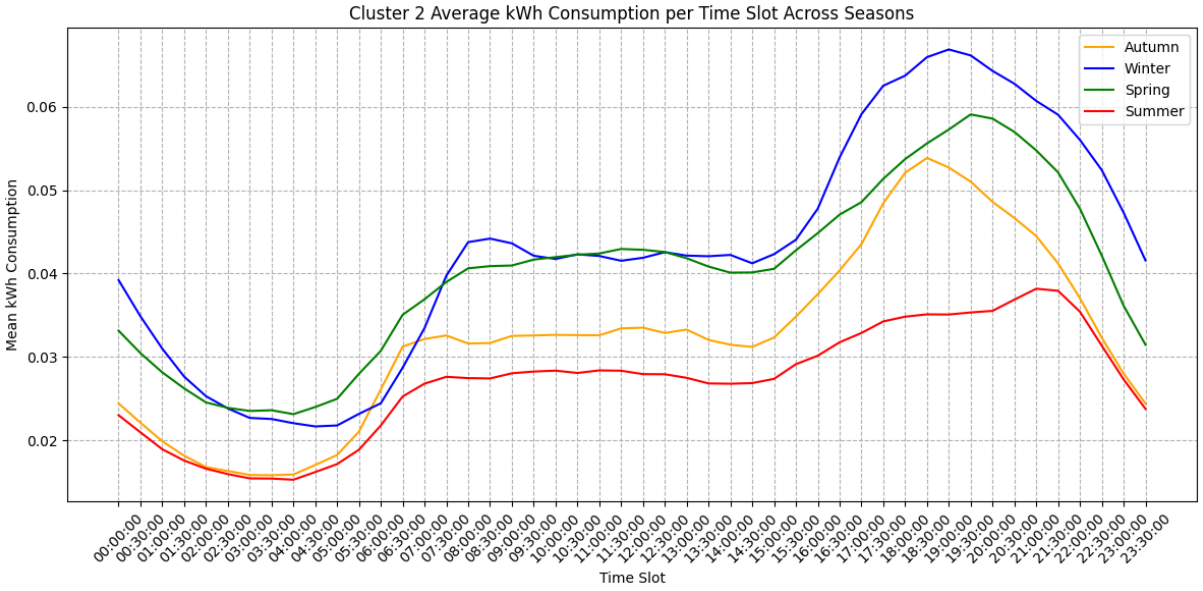


Figure A.2: Seasonal Consumption Profile for Cluster 2

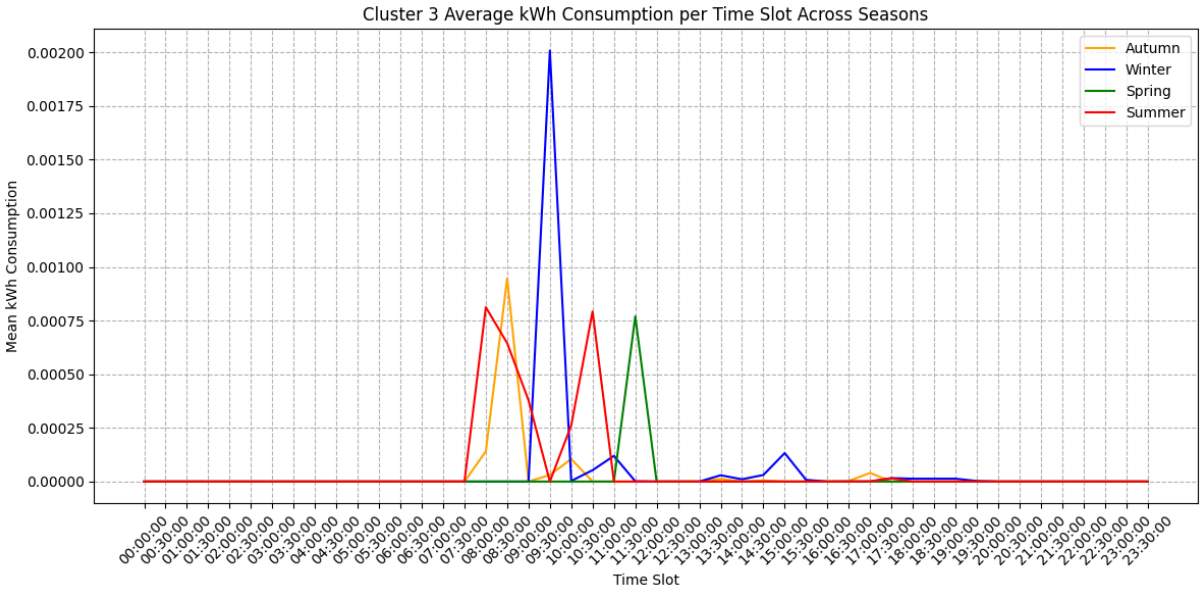


Figure A.3: Seasonal Consumption Profile for Cluster 3

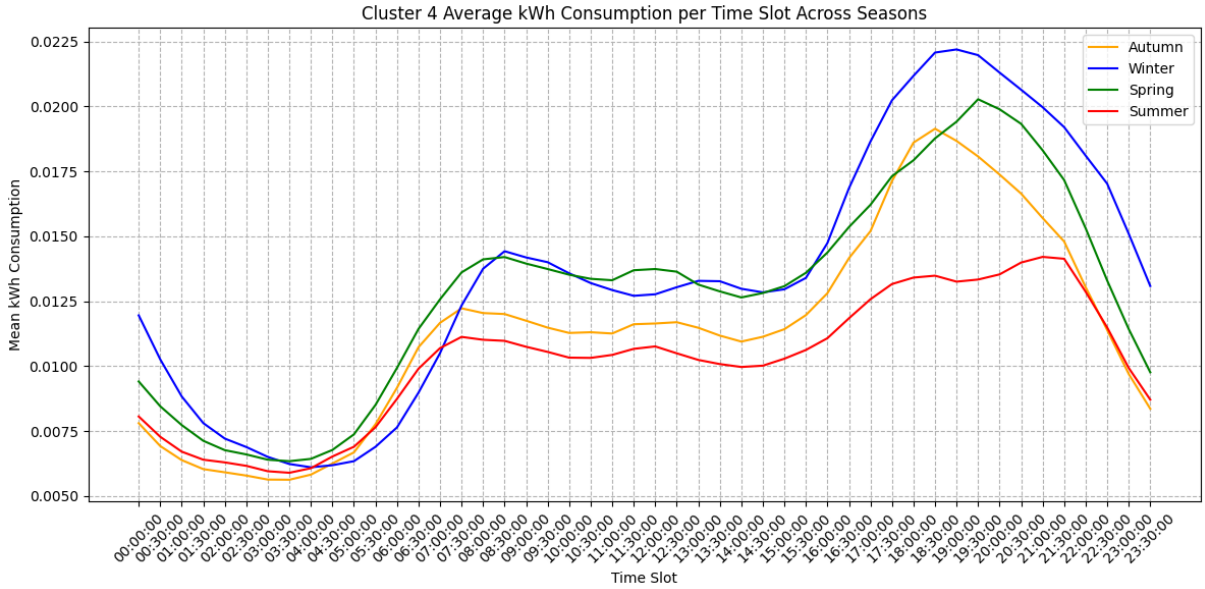


Figure A.4: Seasonal Consumption Profile for Cluster 4

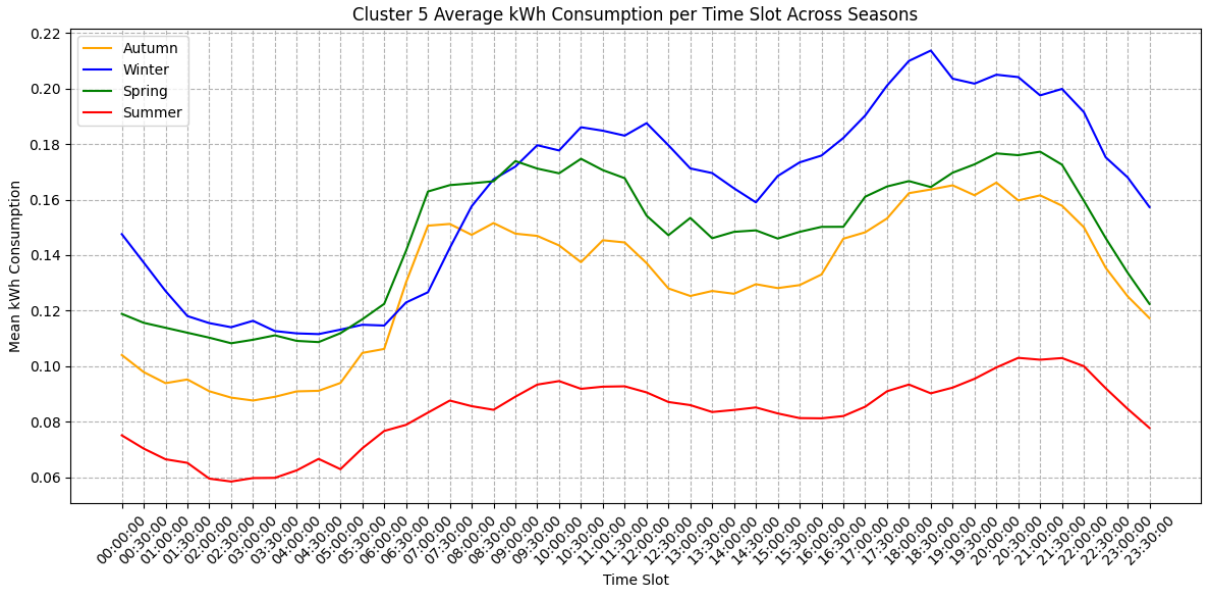


Figure A.5: Seasonal Consumption Profile for Cluster 5

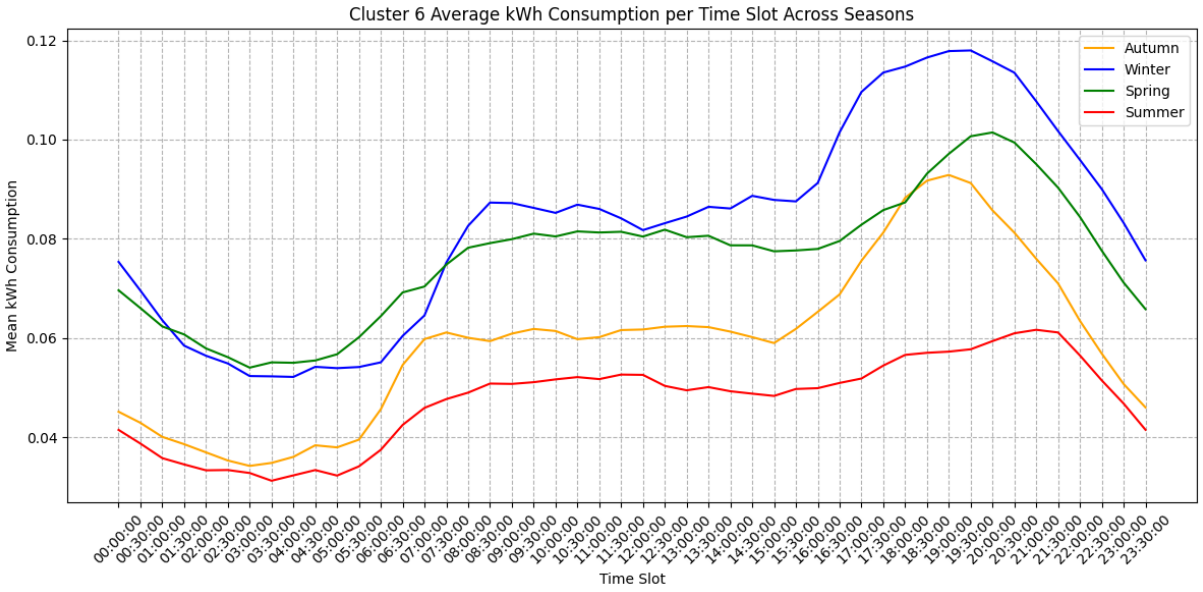


Figure A.6: Seasonal Consumption Profile for Cluster 6

A.3 Refined Load Shape Profiles Post-Outlier Removal

In this section, we present the load shape profiles for each of the six clusters after the removal of outlier values. Outliers in energy consumption data can often skew the analysis, leading to misrepresentations of the typical usage patterns within each cluster. By eliminating these extreme values, we aim to achieve a more accurate and realistic depiction of the average energy consumption behaviors in each consumer segment.

The refined load shape profiles are expected to provide:

- A clearer understanding of the typical daily energy usage patterns in each cluster, devoid of extreme variations caused by outliers.
- Enhanced insights into the operational characteristics and lifestyle habits of consumers within each cluster, facilitating more targeted energy management strategies.

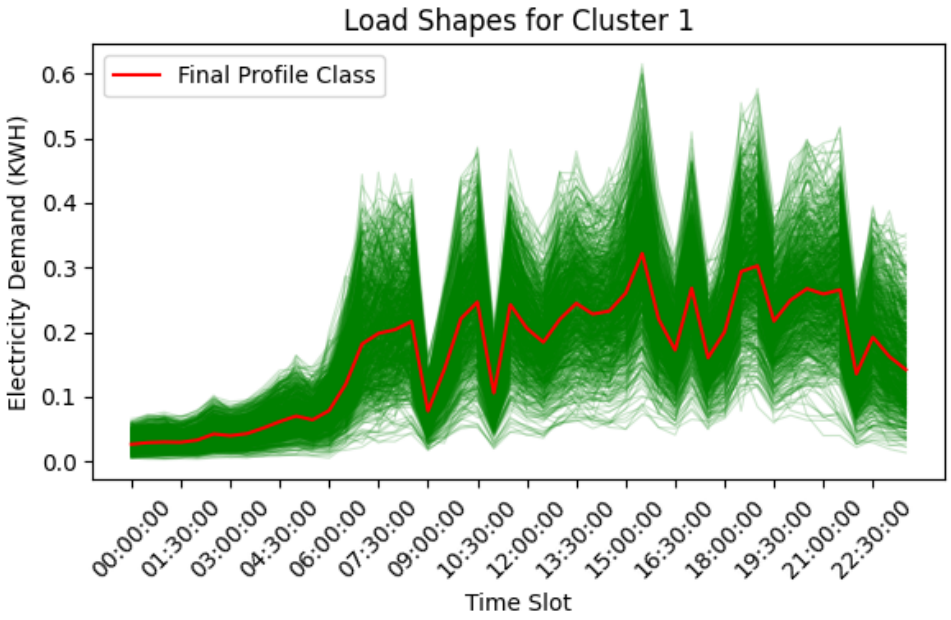


Figure A.7: Normalized Load Shapes Post-Outlier Removal for Cluster 1

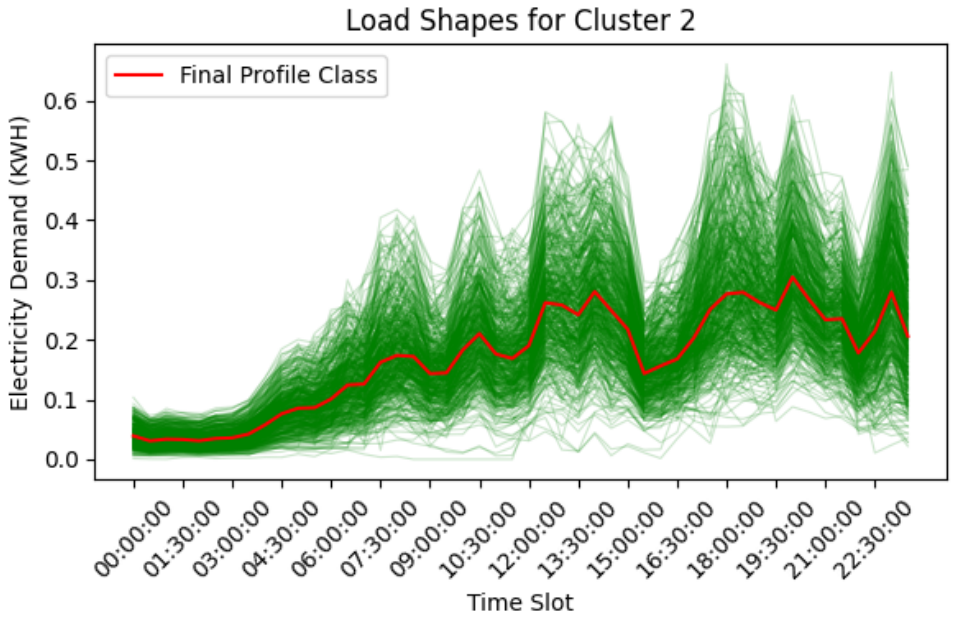


Figure A.8: Normalized Load Shapes Post-Outlier Removal for Cluster 2

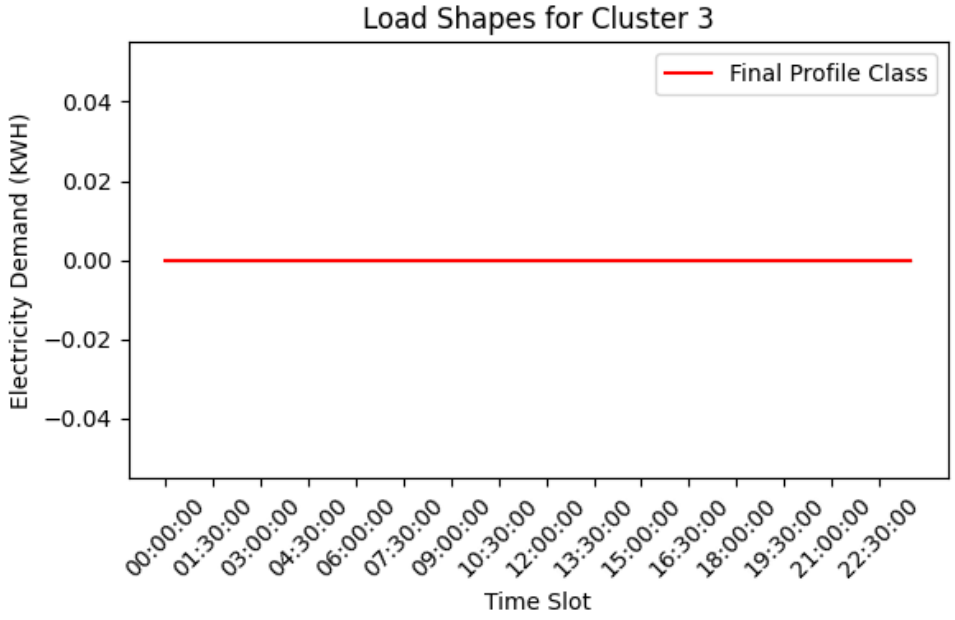


Figure A.9: Normalized Load Shapes Post-Outlier Removal for Cluster 3

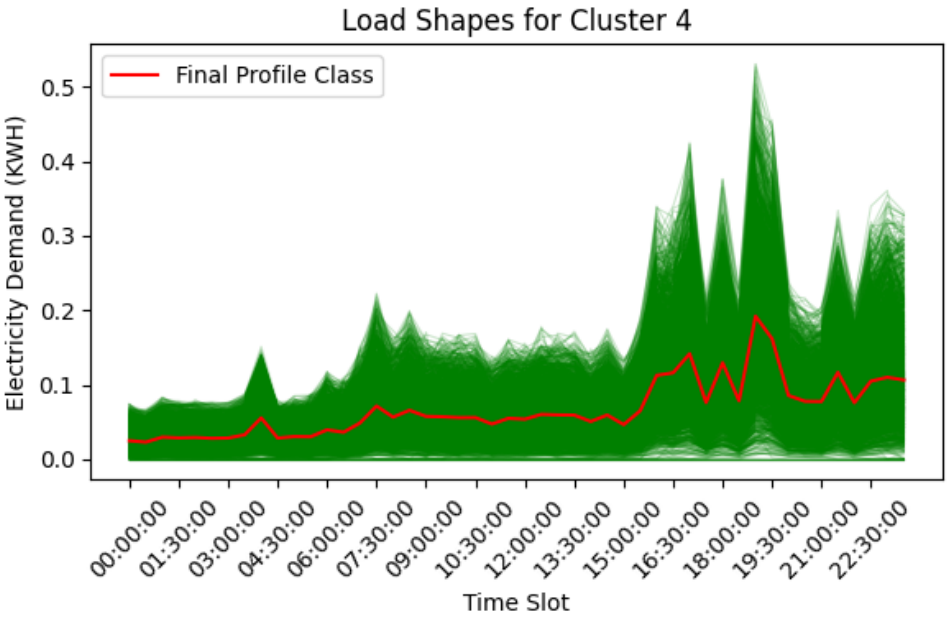


Figure A.10: Normalized Load Shapes Post-Outlier Removal for Cluster 4

Figure A.11: Normalized Load Shapes Post-Outlier Removal for Cluster 5

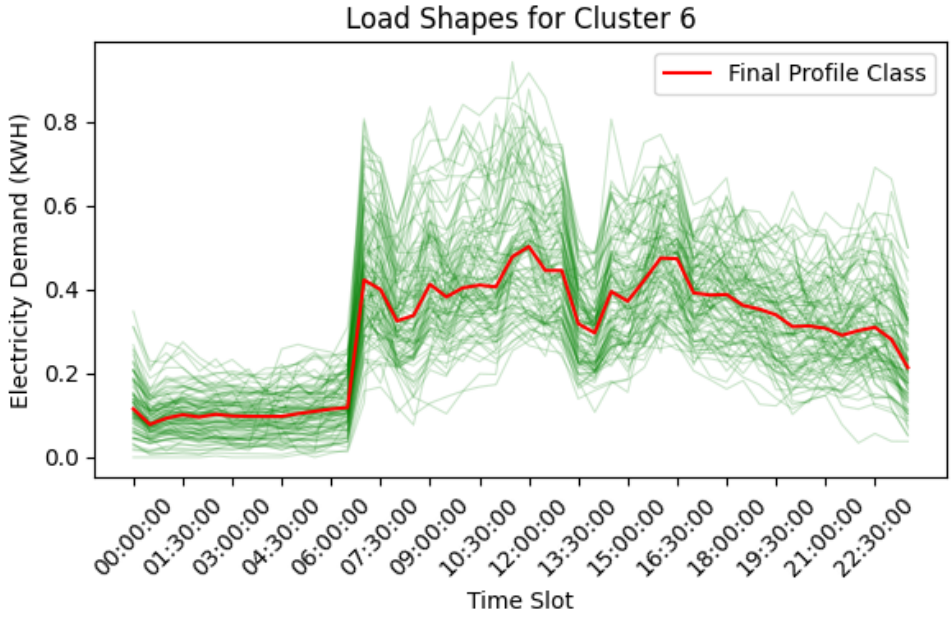


Figure A.12: Normalized Load Shapes Post-Outlier Removal for Cluster 6

A.4 Code Availability

All the code developed and utilized for the analyses presented in this document, including the clustering algorithms, data preprocessing, normalization, outlier removal, and generation of visualizations, is available for review and use. The code repository, which offers a comprehensive collection of the scripts and notebooks, can be accessed on GitHub.

Interested readers, researchers, and practitioners can find the code at the following GitHub repository: github.com/afragiadaki/Clustering-EnergyConsumers. This repository includes detailed documentation and instructions for running the code, allowing for replication of the results or further exploration and adaptation of the methodologies for other datasets or purposes.

Bibliography

[1] Pierluigi Siano. "Demand response and smart grids—A survey". In: *Renewable and Sustainable Energy Reviews* 30 (2014), pp. 461–478. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2013.10.022. URL: https://www.sciencedirect.com/science/article/pii/S1364032113007211.

[2] Nikolaos G. Paterakis, Ozan Erdinç, and João P.S. Catalão. "An overview of Demand Response: Key-elements and international experience". In: *Renewable and Sustainable Energy Reviews* 69 (2017), pp. 871–891. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2016.11.167. URL: https://www.sciencedirect.com/science/article/pii/S1364032116308966.

[3] Amir Safdarian, Mahmud Fotuhi-Firuzabad, and Matti Lehtonen. "Benefits of Demand Response on Operation of Distribution Networks: A Case Study". In: *IEEE Systems Journal* 10.1 (2016), pp. 189–197. DOI: 10.1109/JSYST.2013.2297792.

[4] Milad Afzalan and Farrokh Jazizadeh. "Residential loads flexibility potential for demand response using energy consumption patterns and user segments". In: *Applied Energy* 254 (2019), p. 113693. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2019.113693. URL: https://www.sciencedirect.com/science/article/pii/S0306261919313807.

[5] A. Rezaee Jordehi. "Optimisation of demand response in electric power systems, a review". In: *Renewable and Sustainable Energy Reviews* 103 (2019), pp. 308–319. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2018.12.054. URL: https://www.sciencedirect.com/science/article/pii/S1364032118308566.

[6] Sima Davarzani et al. "Residential Demand Response Strategies and Applications in Active Distribution Network Management". In: *Renewable and Sustainable Energy Reviews* 138 (2021), p. 110567. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2020.110567. URL: https://www.sciencedirect.com/science/article/pii/S1364032120308510.

[7] Houman Jamshidi Monfared et al. "A hybrid price-based demand response program for the residential micro-grid". In: *Energy* 185 (2019), pp. 274–285. ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2019.07.045. URL: https://www.sciencedirect.com/science/article/pii/S0360544219313726.

[8] European Commission et al. *The European green deal*. 2019.

[9] Peter Bradley, Matthew Leach, and Jacopo Torriti. "A review of the costs and benefits of demand response for electricity in the UK". In: *Energy Policy* 52 (2013). Special Section: Transition Pathways to a Low Carbon Economy, pp. 312–327. ISSN: 0301-4215. DOI: https://doi.org/10.1016/j.enpol.2012.09.039. URL: https://www.sciencedirect.com/science/article/pii/S0301421512008142.

[10] Rémi Louf and Marc Barthelemy. "Patterns of residential segregation". In: *PloS one* 11.6 (2016), e0157476.

[11] Ling Zhang and Baosen Zhang. "Scenario forecasting of residential load profiles". In: *IEEE Journal on Selected Areas in Communications* 38.1 (2019), pp. 84–95.

[12] Eunjung Lee, Jinho Kim, and Dongsik Jang. "Load Profile Segmentation for Effective Residential Demand Response Program: Method and Evidence from Korean Pilot Study". In: *Energies* 13.6 (2020). ISSN: 1996-1073. DOI: 10.3390/en13061348. URL: https://www.mdpi.com/1996-1073/13/6/1348.

[13] Tom M Mitchell. *Machine learning*. 1997.

[14] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[15] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[16] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[17] Zoubin Ghahramani. "Unsupervised learning". In: *Summer school on machine learning*. Springer, 2003, pp. 72–112.

[18] R. Xu and D. C. Wunsch. "Survey of clustering algorithms". In: *IEEE Transactions on Neural Networks* 16 (3 2005), pp. 645–678. DOI: 10.1109/tnn.2005.845141.

[19] L. Rokach. "A survey of clustering algorithms". In: *Data Mining and Knowledge Discovery Handbook* (2009), pp. 269–298. DOI: 10.1007/978-0-387-09823-4_14.

[20] M. Noroozi and P. Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles". In: *Computer Vision – ECCV 2016* (2016), pp. 69–84. DOI: 10.1007/978-3-319-46466-4_5.

[21] Sina Khanmohammadi, Naiier Adibeig, and Samaneh Shanehbandy. "An improved overlapping k-means clustering method for medical applications". In: *Expert Systems with Applications* 67 (2017), pp. 12–18.

[22] IBM. *Unsupervised Learning*. https://www.ibm.com/topics/unsupervised-learning.

[23] Dongkuan Xu and Yingjie Tian. "A comprehensive survey of clustering algorithms". In: *Annals of Data Science* 2 (2015), pp. 165–193.

[24] Dmitry Paranyushkin. "InfraNodus: Generating Insight Using Text Network Analysis". In: *The World Wide Web Conference*. WWW '19. San Francisco, CA, USA: Association for Computing Machinery, 2019, pp. 3584–3589. ISBN: 9781450366748. DOI: 10.1145/3308558.3314123. URL: https://doi.org/10.1145/3308558.3314123.

[25] Selin Yilmaz, Jonathan Chambers, and Martin Kumar Patel. "Comparison of clustering approaches for domestic electricity load profile characterisation-Implications for demand side management". In: *Energy* 180 (2019), pp. 665–677.

[26] Jungsuk Kwac, June Flora, and Ram Rajagopal. "Household energy consumption segmentation using hourly data". In: *IEEE Transactions on Smart Grid* 5.1 (2014), pp. 420–430.

[27] Amin Rajabi et al. "A review on clustering of residential electricity customers and its applications". In: *2017 20th International Conference on Electrical Machines and Systems (ICEMS)*. IEEE. 2017, pp. 1–6.

[28] Richard Bellman and Robert Kalaba. "On adaptive control processes". In: *IRE Transactions on Automatic Control* 4.2 (1959), pp. 1–9.

[29] Gianfranco Chicco. "Overview and performance assessment of the clustering methods for electrical load pattern grouping". In: *Energy* 42.1 (2012), pp. 68–80.

[30] Teemu Räsänen and Mikko Kolehmainen. "Feature-based clustering for electricity use time series data". In: *Adaptive and Natural Computing Algorithms: 9th International Conference, ICANNGA 2009, Kuopio, Finland, April 23-25, 2009, Revised Selected Papers 9*. Springer. 2009, pp. 401–412.

[31] Stephen Haben, Colin Singleton, and Peter Grindrod. "Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data". In: *IEEE Transactions on Smart Grid* 7.1 (2016), pp. 136–144. DOI: 10.1109/TSG.2015.2409786.

[32] Krzysztof Gajowniczek and Tomasz Ząbkowski. "Simulation study on clustering approaches for short-term electricity forecasting". In: *Complexity* 2018 (2018).

[33] G. Chicco, R. Napoli, and F. Piglione. "Comparisons among clustering techniques for electricity customer classification". In: *IEEE Transactions on Power Systems* 21.2 (2006), pp. 933–940. DOI: 10.1109/TPWRS.2006.873122.

[34] José J. López et al. "Hopfield–K-Means clustering algorithm: A proposal for the segmentation of electricity customers". In: *Electric Power Systems Research* 81.2 (2011), pp. 716–724. ISSN: 0378-7796. DOI: https://doi.org/10.1016/j.epsr.2010.10.036. URL: https://www.sciencedirect.com/science/article/pii/S0378779610002713.

[35] Yi Wang et al. "Clustering of electricity consumption behavior dynamics toward big data applications". In: *IEEE transactions on smart grid* 7.5 (2016), pp. 2437–2447.

[36] Amin Rajabi et al. "A comparative study of clustering techniques for electrical load pattern segmentation". In: *Renewable and Sustainable Energy Reviews* 120 (2020), p. 109628.

[37] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. "A clustering approach to domestic electricity load profile characterisation using smart metering data". In: *Applied energy* 141 (2015), pp. 190–199.

[38] Gianfranco Chicco, Roberto Napoli, and Federico Piglione. "Application of clustering algorithms and self organising maps to classify electricity customers". In: *2003 IEEE Bologna Power Tech Conference Proceedings,* vol. 1. IEEE. 2003, 7–pp.

[39] SM Bidoki et al. "Evaluating different clustering techniques for electricity customer classification". In: *IEEE PES T&D 2010*. IEEE. 2010, pp. 1–5.

[40] Gianfranco Chicco et al. "Load pattern-based classification of electricity customers". In: *IEEE Transactions on Power Systems* 19.2 (2004), pp. 1232–1239.

[41] Zigui Jiang, Rongheng Lin, and Fangchun Yang. "A Hybrid Machine Learning Model for Electricity Consumer Categorization Using Smart Meter Data". In: *Energies* 11.9 (2018). ISSN: 1996-1073. DOI: 10.3390/en11092235. URL: https://www.mdpi.com/1996-1073/11/9/2235.

[42] Nameer Al Khafaf, Mahdi Jalili, and Peter Sokolowski. "A novel clustering index to find optimal clusters size with application to segmentation of energy consumers". In: *IEEE transactions on industrial informatics* 17.1 (2020), pp. 346–355.

[43] V. Figueiredo et al. "An electric energy consumer characterization framework based on data mining techniques". In: *IEEE Transactions on Power Systems* 20.2 (2005), pp. 596–602. DOI: 10.1109/TPWRS.2005.846234.

[44] Reem Al-Otaibi et al. "Feature Construction and Calibration for Clustering Daily Load Curves from Smart-Meter Data". In: *IEEE Transactions on Industrial Informatics* 12.2 (2016), pp. 645–654. DOI: 10.1109/TII.2016.2528819.

[45] Gareth Powells et al. "Peak electricity demand and the flexibility of everyday life". In: *Geoforum* 55 (2014), pp. 43–52. ISSN: 0016-7185. DOI: https://doi.org/10.1016/j.geoforum.2014.04.014. URL: https://www.sciencedirect.com/science/article/pii/S0016718514000931.

[46] Josh Warner et al. *JDWarner/scikit-fuzzy: Scikit-Fuzzy version 0.4.2*. Version v0.4.2. Nov. 2019. DOI: 10.5281/zenodo.3541386. URL: https://doi.org/10.5281/zenodo.3541386.

[47] Giuseppe Vettigli. *MiniSom: minimalistic and NumPy-based implementation of the Self Organizing Map*. 2018. URL: https://github.com/JustGlowing/minisom/.

[48] Anil K Jain. "Data clustering: 50 years beyond K-means". In: *Pattern recognition letters* 31.8 (2010), pp. 651–666.

[49] Caterina Labrín and Francisco Urdinez. "Principal component analysis". In: *R for Political Data Science*. Chapman and Hall/CRC, 2020, pp. 375–393.