ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΗΣ ΙΣΧΥΟΣ

Μηχανική Μάθηση σε Ενεργειακά Δεδομένα με σκοπό την Εξαγωγή Γνώσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Απόστολος Δ. Μαυρόπουλος

Επιβλέπων : Νικόλαος Δ. Χατζηαργυρίου
Καθηγητής Ε.Μ.Π

Αθήνα, Ιούλιος 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΗΣ ΙΣΧΥΟΣ

Μηχανική Μάθηση σε Ενεργειακά Δεδομένα με σκοπό την Εξαγωγή Γνώσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Απόστολος Δ. Μαυρόπουλος

Επιβλέπων : Νικόλαος Δ. Χατζηαργυρίου
Καθηγητής Ε.Μ.Π

.....	.....	.....
Νικόλαος Χατζηαργυρίου	Σταύρος Παπαθανασίου	Παύλος Γεωργιλάκης
Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π

Αθήνα, Ιούλιος 2021

.....

Απόστολος Δ. Μαυρόπουλος

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

ΕΥΧΑΡΙΣΤΙΕΣ

Νιώθω την ανάγκη να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Νικόλαο Χατζηαργυρίου για τη δυνατότητα που μου έδωσε να ασχοληθώ με τη παρούσα διπλωματική εργασία.

Επίσης, θέλω να ευχαριστήσω τον κ. Άρη Δημέα και τον κ. Κυριάκο Ανδρεσάκη που με την υπομονετική τους καθοδήγηση με βοήθησαν να την ολοκληρώσω.

Η εργασία αυτή είναι αφιερωμένη στην οικογένεια μου και σε όλους τους ανθρώπους που με στήριξαν όλα αυτά τα χρόνια.

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία εντάσσεται στο ερευνητικό πεδίο της ανάλυσης ενεργειακών δεδομένων και πιο συγκεκριμένα της εφαρμογής τεχνικών μηχανικής μάθησης και εξόρυξης δεδομένων σε χρονοσειρές φορτίου ηλεκτρικής ενέργειας. Σε αυτό το πλαίσιο παρουσιάζονται και εφαρμόζονται τεχνικές και εργαλεία για την ανάλυση χρονοσειρών ενεργειακών δεδομένων με σκοπό την εξαγωγή γνώσης.

Στο πρώτο κεφάλαιο γίνεται αναφορά στον τύπο των δεδομένων ανάλυσης καθώς και στον Ευρωπαϊκό σύνδεσμο ENTSO-E, διαχειριστή της πλατφόρμας διαφάνειας (TP) μέσω της οποίας διατίθενται δημόσια τα εν λόγω δεδομένα. Επίσης, παρουσιάζονται συνοπτικά τα ανοιχτής πηγής εργαλεία – βιβλιοθήκες και υπολογιστικό περιβάλλον ανάπτυξης που αξιοποιήθηκαν για την υλοποίηση της ανάλυσης και των εφαρμογών.

Στο δεύτερο κεφάλαιο γίνεται αναφορά στη μηχανική μάθηση. Παρουσιάζονται οι βασικές κατηγορίες μάθησης, μέθοδοι αξιολόγησης και βελτιστοποίησης μοντέλων, θεμελιώδεις αλγόριθμοι καθώς και προβλήματα που αναδύονται. Εν συνεχεία, στο τρίτο κεφάλαιο παρουσιάζονται μέθοδοι και τεχνικές που αξιοποιούνται στην εξόρυξη δεδομένων και στη μηχανική χαρακτηριστικών.

Στο τέταρτο κεφάλαιο παρουσιάζονται αναλυτικά οι εφαρμογές μη επιβλεπόμενης μάθησης που υλοποιήσαμε. Ιδιαίτερη έμφαση δίνεται στην εξαγωγή προτοτύπων ως χαρακτηριστικές καμπύλες – προφίλ φορτίου. Τέλος, στο πέμπτο κεφάλαιο παρουσιάζονται αναλυτικά οι εφαρμογές επιβλεπόμενης μάθησης και μηχανικής χαρακτηριστικών καθώς και η συγκριτική αξιολόγηση των μοντέλων ταξινόμησης που αναπτύχθηκαν.

Λέξεις – Κλειδιά

Ενεργειακά Δεδομένα, Χρονοσειρές Φορτίου, Μηχανική Μάθηση, Μηχανική Χαρακτηριστικών, Εξόρυξη Δεδομένων, Εξαγωγή Γνώσης, Προφίλ Φορτίου, Χαρακτηριστικές Καμπύλες Φορτίου, Συσταδοποίηση, Ομαδοποίηση, Ταξινόμηση, Κατηγοριοποίηση.

ABSTRACT

This diploma thesis is part of the research field of energy data analysis and more specifically of the application of machine learning and data mining techniques in load time series. In this context, tools and techniques for energy time series data analysis are being presented and applied for the purpose of knowledge extraction.

The first chapter presents the set of data for analysis as well as the European Network of Transmission System Operators for Electricity ENTSO-E, the manager of the Transparency Platform (TP) through which the above data are publicly available. In addition, the open source libraries, tools and Integrated Development Environment (IDE) that were utilized for the implementation of the analysis and applications are briefly presented.

The second chapter presents main definitions and categories of machine learning, model evaluation and optimization methods, fundamental algorithms as well as emerging drawbacks. Consequently, in the third chapter, methods and techniques utilized in data mining and feature engineering are exhibited.

The fourth chapter thoroughly presents the unsupervised machine learning applications that we developed. We stress out the importance of load time series clustering regarding the extraction of prototype based load profiles also known as typical load curves. Finally, the fifth chapter presents the supervised machine learning models and feature engineering applications that we developed as well as a benchmarking of the above classification models in detail.

Keywords

Energy Data, Load Time Series, Machine Learning, Feature Engineering, Data Mining, Knowledge Extraction, Typical Load Curves, Load Profiles, Clustering, Classification.

Περιεχόμενα

Ευρετήριο Σχημάτων

Ευρετήριο Πινάκων

Κεφάλαιο 1: Εισαγωγή

Η ραγδαία ανάπτυξη της τεχνολογίας έχει προκαλέσει αντίστοιχη αύξηση στον όγκο των αποθηκευμένων δεδομένων. Η αξιοποίηση της μεγάλης αυτής ποσότητας δεδομένων με σκοπό την ανακάλυψη γνώσης παραμένει μια σπουδαία πρόκληση, ιδιαίτερα στον τομέα των παραγωγικών διαδικασιών. Συγκεκριμένα, στη βιομηχανία της ενέργειας δημιουργείται η ανάγκη για ανάλυση και ερμηνεία δεδομένων που αντικατοπτρίζουν την δυναμική εξέλιξη των συστημάτων παραγωγής και κατανάλωσης. Κατά συνέπεια, τα ενεργειακά δεδομένα στο μεγαλύτερο ποσοστό τους παρουσιάζονται ως χρονοσειρές. Στα πλαίσια της μηχανικής μάθησης και εξόρυξης ενεργειακών δεδομένων συναντάμε κυρίως τεχνικές ομαδοποίησης και ανίχνευσης προτύπων, κατηγοριοποίησης, πρόβλεψης και ανίχνευσης ανωμαλιών [54]. Στη παρούσα διπλωματική εργασία επίκεντρο αποτελεί η υλοποίηση τεχνικών εφαρμογών ομαδοποίησης και κατηγοριοποίησης χρονοσειρών κατανάλωσης ενέργειας ηλεκτρικού φορτίου.Τα δεδομένα της ανάλυσης αφορούν τις ενεργειακές καταναλώσεις ευρωπαϊκών χωρών και είναι διαθέσιμα μέσω της πλατφόρμας διαφάνειας του ευρωπαϊκού διαχειριστή μεταφοράς ηλεκτρικής ενέργειας ENTSO-E [20].

1.1 Χρονοσειρές Ενεργειακών Δεδομένων

Οι χρονοσειρές αποτελούν σύνολα διαδοχικών παρατηρήσεων της τιμής ενός μετρούμενου μεγέθους που μεταβάλλεται χρονικά. Συνεπώς, αφορούν σύνολα δεδομένων με φυσική χρονική διάταξη. Η επιστήμη των υπολογιστών και της πληροφορικής αντιμετωπίζει τις χρονοσειρές ως σύνολα δεδομένων διακριτού χρόνου και πεπερασμένης ακρίβειας καθώς υπάρχει πεπερασμένο πλήθος τιμών που δύνανται να αναπαρασταθούν στη μνήμη λόγω της αριθμητικής κινητής υποδιαστολής και των φαινομένων της υπερχείλισης και ανεπάρκειας. Επίσης υπάρχει διαχωρισμός μεταξύ μονομεταβλητών (univariate) και πολυμεταβλητών (multivariate) χρονοσειρών. Στη περίπτωση των μονομεταβλητών εξετάζεται μία μόνο εξαρτημένη ως προς τον χρόνο μεταβλητή ενώ στις πολυμεταβλητές εξετάζονται δύο ή περισσότερες. Η παρούσα διπλωματική εργασία πραγματεύεται την ανάλυση μονομεταβλητών χρονοσειρών. Μαθηματικά αναπαρίστανται ως εξής :

$$T = (t_0, \dots, t_n), \qquad t_i \in \mathbb{R}, \ \ n \in \mathbb{N} \tag{1.1}$$

Στη περίπτωση των χρονοσειρών ενεργειακών δεδομένων οι παρατηρήσεις t_i αφορούν συνήθως μετρήσεις κατανάλωσης και παραγωγής ηλεκτρικής ενέργειας, θέρμανσης και φυσικού αερίου [54]. Η συχνότητα δειγματοληψίας των μετρήσεων είναι ομοιόμορφη και καθορίζει τη διάσταση των δεδομένων. Βασικά χαρακτηριστικά ενεργειακών χρονοσειρών είναι η ύπαρξη ή μη ύπαρξη στασιμότητας (stationarity), γραμμικότητας (linearity) κανονικότητας (regularity), κυκλικότητας (cyclicity), εποχικότητας (seasonality), τάσης (trend) και κλασματικότητας (fractality) [2, 14].

Εκτός από τη πληροφορία που είναι χρήσιμη, οι χρονοσειρές ενδέχεται να περιέχουν παρεμβολές που αποτελούν θόρυβο, ακραίες και ελλιπείς παρατηρήσεις καθώς και παρατηρήσεις που χρήζουν ημερολογιακής προσαρμογής. Συνεπώς, πριν προβεί κανείς σε τεχνικές εφαρμογές απαιτείται η επισκόπηση των χρονοσειρών μέσω γραφικών παραστάσεων και ο υπολογισμός βασικών στατιστικών δεικτών όπως η μέση τιμή, η τυπική απόκλιση και η διακύμανση. Η καλή αντίληψη και αντιπροσωπευτική εικόνα των δεδομένων καθιστούν δυνατή τη σωστή επιλογή μοντέλων ανάλυσης και εφαρμογών.

1.2 Πλατφόρμα Διαφάνειας ENTSO-E

Το ευρωπαϊκό δίκτυο διαχειριστών συστημάτων μεταφοράς ηλεκτρικής ενέργειας ENTSO-E (European Network Transmission System Operator – Electricity) ιδρύθηκε τον Δεκέμβριο του 2008 στο πλαίσιο του Τρίτου Πακέτου Ενέργειας που αποσκοπεί στην απελευθέρωση της ευρωπαϊκής αγοράς ηλεκτρικής ενέργειας και φυσικού αερίου [55]. Αντιπροσωπεύει 42 διαχειριστές συστημάτων μεταφοράς από 35 ευρωπαϊκές χώρες οι οποίοι είναι υπεύθυνοι για τη μαζική μεταφορά ηλεκτρικής ενέργειας στα δίκτυα υψηλής τάσης. Οι κύριες αρμοδιότητες του ENTSO-E είναι ο σχεδιασμός ενός ευρωπαϊκού δικτύου ηλεκτρικής ενέργειας και η ανάπτυξη κοινών εργαλείων για τη διαχείριση των δικτύων σε ευρωπαϊκή κλίμακα.

Στα πλαίσια της διαφάνειας και ακεραιότητας της χονδρικής αγοράς ηλεκτρικής ενέργειας σύμφωνα με τον κανονισμό 543/2013 [17] τέθηκε σε λειτουργία στις 5 Ιανουαρίου του 2015 η Πλατφόρμα Διαφάνειας (Transparency Platform). Ο ENTSO-E έχει δημοσιεύσει εγχειρίδιο διαδικασιών που περιέχει τις απαραίτητες πληροφορίες για την υποβολή και την καταφόρτωση δεδομένων μέσω της πλατφόρμας [62]. Τα δεδομένα και οι πληροφορίες που δημοσιεύονται και διατίθενται μέσω της πλατφόρμας διαφάνειας αφορούν κυριώς τη παραγωγή, τη μεταφορά, τις διακοπές, τα φορτία και τη διαχείριση συμφόρησης στα ευρωπαϊκά δίκτυα ηλεκτρικής ενέργειας [18].

Πρόσφατες μελέτες ερευνητών έχουν επιχειρήσει να αξιολογήσουν την ίδια την πλατφόρμα διαφάνειας καθώς και την ποιότητα και αξιοπιστία των δεδομένων που δημοσιεύονται σε αυτή. Οι εν λόγω μελέτες αναδεικνύουν προβλήματα και προκλήσεις που απαιτείται να αντιμετωπιστούν ώστε να αποτελέσει ένα πραγματικά χρήσιμο εργαλείο για τη βιομηχανία της ενέργειας και τους ερευνητές [20, 21]. Από προσωπική εμπειρία, το βασικό πρόβλημα που αναδύθηκε κατά τη χρήση της πλατφόρμας για την καταφόρτωση δεδομένων "Συνολικού Πραγματικού Φορτίου" (Actual Total Load) είναι το γεγονός ότι στα αρχεία πολλές φορές παρουσιάζονται ελλιπείς τιμές. Το γεγονός αυτό δεν επηρέασε σημαντικά την εκπόνηση της παρούσας διπλωματικής εργασίας καθώς το ποσοστό των ελλιπών τιμών από το σύνολο των δεδομένων ήταν λιγότερο από 1.2%.

1.3 Εργαλεία Βιβλιοθήκες και Υπολογιστικό Περιβάλλον

Η υλοποίηση των εφαρμογών μηχανικής μάθησης και εξόρυξης δεδομένων πραγματοποιήθηκε στο ολοκληρωμένο περιβάλλον ανάπτυξης PyCharm 2020.2.4 σε γλώσσα προγραμματισμού Python 3.7.0. Η Python υποστηρίζει έναν σημαντικό αριθμό βιβλιοθηκών ανοιχτής πηγής (open source) που αποτέλεσαν τον σκελετό για την ανάπτυξη των εν λόγω εφαρμογών και μοντέλων σε χρονοσειρές φορτίου ηλεκτρικής ενέργειας.

Στη συνέχεια παρουσιάζονται συνοπτικά οι βιβλιοθήκες που αξιοποιήθηκαν.

- **pandas :** *Διαχείριση δομών δεδομένων και αρχείων.*
- **scikit-learn :** *Αλγόριθμοι και μοντέλα μηχανικής μάθησης.*
- **tslearn :** *Αλγόριθμοι μηχανικής μάθησης που υποστηρίζουν χρονοσειρές.*
- **tsfresh :** *Αλγόριθμοι για την εξαγωγή και επιλογή χαρακτηριστικών από χρονοσειρές.*
- **numpy :** *Επιστημονικοί υπολογισμοί και διαχείριση πινάκων.*
- **matplotlib :** *Δημιουργία γραφημάτων και οπτικοποίηση.*
- **yellowbrick :** *Διαγνωστικά εργαλεία, αξιολόγηση μοντέλων και οπτικοποίηση.*

Πίνακας 1.1 : Βιβλιοθήκες της Python και πεδία εφαρμογών.

Algorithms & Tasks	pandas	sklearn	tslearn	tsfresh	numpy	matplotlib	yellowbrick
Clustering		✕	✕				
Classification		✕	✕				
Preprocessing	✕	✕	✕		✕		
Model Evaluation		✕					✕
Feature Extraction		✕		✕			
Feature Selection		✕		✕			✕
Visualization						✕	✕

Η σπουδαιότητα της διαθεσιμότητας των παραπάνω βιβλιοθηκών είναι αναμφισβήτητη. Προσφέρουν και υποστηρίζουν πολύπλοκες μαθηματικές διεργασίες μέσω απλών εντολών, συναρτήσεων και κλάσεων οι οποίες έχουν ελεγχθεί συστηματικά για την ακεραιότητα τους από έναν σημαντικό αριθμό προγραμματιστών και επαγγελματιών στον χώρο της πληροφορικής και της μαθηματικής επιστήμης. Επιπρόσθετα, διατίθενται και κυκλοφορούν οδηγοί που τεκμηριώνουν όλες τις λεπτομέρειες και τα τεχνικά χαρακτηριστικά που τις διέπουν [56 - 61]. Ακόμα, το γεγονός ότι είναι ανοιχτής πηγής καθιστά δυνατή τη προσαρμογή και τροποποίηση του πηγαίου κώδικα εφόσον το απαιτούν οι συνθήκες και το είδος της εφαρμογής προς υλοποίηση. Τέτοιες πρακτικές όμως συνιστάται να αποφεύγονται όταν δεν υπάρχει δυνατότητα τεκμηρίωσης της ορθής λειτουργίας των τροποποιημένων συναρτήσεων και κλάσεων. Τέλος, η διαλειτουργικότητα των βιβλιοθηκών υποστηρίζει την ομαλή ροή εγασιών και προσφέρει στον χρήστη τη δυνατότητα ανάπτυξης επαρκώς δομημένων και αποδοτικών εφαρμογών.

Κεφάλαιο 2 : Θεωρητικό Υπόβαθρο Μηχανικής Μάθησης

Στο παρόν κεφάλαιο πραγματοποιείται μια εισαγωγή στο θεωρητικό υπόβαθρο της Μηχανικής Μάθησης (Machine Learning). Στα κεφάλαια 2.1 και 2.2 παρουσιάζεται ο ορισμός και προσδιορίζονται οι δύο θεμελιώδεις κατηγορίες Μηχανικής Μάθησης αντίστοιχα. Στο κεφάλαιο 2.3 και 2.4 παρουσιάζονται τεχνικές, μέθοδοι αξιολόγησης καθώς και αλγόριθμοι και μοντέλα μηχανικής μάθησης που εφαρμόσαμε σε δεδομένα ενεργειακών χρονοσειρών ηλεκτρικού φορτίου. Στο κεφάλαιο 2.5 προσδιορίζεται η έννοια των υπερπαραμέτρων και η σημασία τους στη διαδικασία της βελτιστοποίησης μοντέλων. Τέλος, στο κεφάλαιο 2.6 προσδιορίζεται το πρόβλημα της υπερμοντελοποίησης και υποδεικνύονται μέθοδοι αντιμετώπισης. Βιβλία στα οποία βασιστήκαμε και προτείνονται για μια εισαγωγή στη Μηχανική Μάθηση είναι πρώτον το "The Hundred – Page Machine Learning Book" του Andriy Burkov [49], δεύτερον το "Data Mining : Practical Machine Learning Tools and Techniques" των Witten, Frank, Hall και Pall [51] και τρίτον το "The Elements of Statistical Learning : Data Mining Inference and Prediction" των Hastie, Tibshirani και Friedman [52].

2.1 Ορισμός

Η Μηχανική Μάθηση (Machine Learning) αποτελεί υποπεδίο της Επιστήμης Υπολογιστών και μπορεί να προσδιοριστεί ως η διαδικασία της επίλυσης πρακτικών προβλημάτων μέσω της αλγοριθμικής ανάπτυξης στατιστικών μοντέλων που αναλύουν και επεξεργάζονται δεδομένα. Ένας επίσημος ορισμός δόθηκε το 1997 από τον Tom M. Mitchell όπου διατυπώνει την έννοια της Μηχανικής Μάθησης ως εξής : « Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από εμπειρία E ως προς μια κλάση εργασιών T και ένα μέτρο επίδοσης P, αν η επίδοση του σε εργασίες της κλάσης T, όπως αποτιμάται από το μέτρο P, βελτιώνεται με την εμπειρία E ». Σύμφωνα με τον παραπάνω ορισμό είναι δυνατό να θεωρήσουμε τη Μηχανική Μάθηση ως τη διαδικασία σχεδιασμού και υλοποίησης αλγορίθμων και συστημάτων που εκπαιδεύονται σε ένα σύνολο δεδομένων εισόδου και βελτιστοποιούνται αυτόματα χρησιμοποιώντας στατιστική ανάλυση με απώτερο στόχο την αποδοτικότερη λήψη αποφάσεων και αναγνώριση προτύπων.

2.2 Κατηγορίες Μηχανικής Μάθησης

Οι δύο βασικές κατηγορίες Μηχανικής Μάθησης είναι η **Επιβλεπόμενη (Supervised)** και η **Μη Επιβλεπόμενη (Unsupervised)**. Ο διαχωρισμός έγκειται στον τρόπο που πραγματοποιείται η διαδικασία μάθησης στο υπό εκπαίδευση σύστημα. Στη κατηγορία της Επιβλεπόμενης Μάθησης τα διαθέσιμα δεδομένα εισόδου συνοδεύονται από επισημασμένα δεδομένα επιθυμητής εξόδου (labels). Η επισήμανση (labeling) των δεδομένων εισόδου πραγματοποιείται συνήθως από ανθρώπους και αποτελεί σε γενικές γραμμές μια αρκετά χρονοβόρα και δαπανηρή διαδικασία. Αντίθετα, στη κατηγορία της Μη Επιβλεπόμενης Μάθησης δεν ορίζονται επισημασμένα δεδομένα και επιχειρείται η ανακάλυψη δομών στα δεδομένα εισόδου χωρίς την καθοδήγηση από παραδείγματα επιθυμητών αποτελεσμάτων.

Σημειώνεται επιπλέον η ύπαρξη τριών ακόμη κατηγοριών Μηχανικής Μάθησης, της **Ημι-Επιβλεπόμενης Μάθησης (Semi-Supervised Learning)** της **Ενισχυτικής Μάθησης (Reinforcement Learning)** και της **Μετα-Εκπαίδευσης (Meta Learning)**.

Οι παραπάνω τρεις κατηγορίες Μηχανικής Μάθησης αναφέρονται για λόγους πληρότητας. Η παρούσα διπλωματική εργασία επικεντρώνεται μόνο σε εφαρμογές Επιβλεπόμενης και Μη Επιβλεπόμενης Μάθησης.

2.2.1 Επιβλεπόμενη Μάθηση

Ορίζουμε το σύνολο δεδομένων εισόδου ως X και το σύνολο επισημασμένων δεδομένων εξόδου ως Y. Συγκεκριμένα, το σύνολο Y περιέχει τις ετικέτες (labels) των παρατηρήσεων του συνόλου εισόδου. Σκοπός της Επιβλεπόμενης Μάθησης είναι η εκπαίδευση ενός μοντέλου που θα αντιστοιχεί τις παρατηρήσεις του X στις αντίστοιχες εξόδους του Y.

$$X \to Y \quad \text{ή} \quad Y = f(X) \tag{2.1}$$

Η συνάρτηση f καλείται συνάρτηση απεικόνισης (mapping function). Η σχέση (2.1) αποτελεί μια ιδανική προσέγγιση του προβλήματος δεδομένου ότι κάθε υπαρκτό σύστημα παρουσιάζει κάποιο σφάλμα ε. Συνεπώς, το πρόβλημα έγκειται στην εύρεση κατάλληλης συνάρτησης απεικόνισης που παράγει το ελάχιστο δυνατό σφάλμα ε ($Y = f(X) + \varepsilon$). Η φύση των συνόλων εξόδου καθορίζει δύο διαφορετικές κατηγορίες προβλημάτων. Στη περίπτωση που το Y περιέχει διακριτές τιμές αναφερόμαστε σε προβλήματα **Ταξινόμησης (Classification)**, ενώ στη περίπτωση που το Y περιέχει συνεχείς τιμές αναφερόμαστε σε προβλήματα **Παλινδρόμησης (Regression)**. Τα μοντέλα Ταξινόμησης ταξινομούν τις παρατηρήσεις του συνόλου εισόδου σε κατηγορίες και τα μοντέλα Παλινδρόμησης αποδίδουν συνεχείς τιμές στις παρατηρήσεις του συνόλου εισόδου.

Στη βιβλιογραφία τα προβλήματα Ταξινόμησης και Παλινδρόμησης αναφέρονται και ως προβλήματα Κατηγοριοποίησης και Πρόβλεψης αντίστοιχα. Στη παρούσα διπλωματική εργασία, όσον αφορά την Επιβλεπόμενη Μάθηση θα επικεντρωθούμε σε εφαρμογές Ταξινόμησης χρονοσειρών.

2.2.2 Μη Επιβλεπόμενη Μάθηση

Στην Μη Επιβλεπόμενη Μάθηση έχουμε στη διάθεση μας μόνο το σύνολο δεδομένων εισόδου X καθώς οι ετικέτες των δεδομένων εισόδου είτε δεν υπάρχουν είτε απλά τις αγνοούμε. Σε τέτοιου είδους προβλήματα επιχειρούμε να ανακαλύψουμε δομές και ιεραρχίες που υποκρύπτονται στο σύνολο των παρατηρήσεων της εισόδου. Οι βασικές κατηγορίες προβλημάτων που εντοπίζονται στην Μη Επιβλεπόμενη Μάθηση είναι τα προβλήματα **Ομαδοποίησης (Clustering)** και τα προβλήματα **Ανίχνευσης Ανωμαλιών (Anomaly Detection)** [27]. Το πρόβλημα της Ομαδοποίησης έγκειται στον σχηματισμό κατάλληλων ομάδών από τα δεδομένα εισόδου με βάση κάποιο προκαθορισμένο μέτρο ομοιότητας. Το πρόβλημα της Ανίχνευσης Ανωμαλιών έγκειται στην ανίχνευση προτύπων με αποκλίνοντα χαρακτηριστικά από το σύνολο των δεδομένων εισόδου.

Στη βιβλιογραφία τα προβλήματα Ομαδοποίησης και Ανίχνευσης Ανωμαλιών αναφέρονται και ως προβλήματα Συσταδοποίησης και Ανίχνευσης Ακραίων Τιμών (Outlier Detection) αντίστοιχα. Στη παρούσα διπλωματική εργασία, όσον αφορά την Μη Επιβλεπόμενη Μάθηση θα επικεντρωθούμε σε εφαρμογές Ομαδοποίησης χρονοσειρών.

Στη συνέχεια παρουσιάζεται ο ορισμός του προβλήματος Ομαδοποίησης χρονοσειρών [1].

Ορισμός 2.2.

$$\textit{Δεδομένου ενός συνόλου χρονοσειρών } X = \{T_1, \dots, T_i, \dots, T_n\}$$

$$\textit{και ενός μέτρου ομοιότητας } D(T_i, T_j), \textit{ανακάλυψε το σύνολο ομάδων}$$

$$C = \{c_1, \dots, c_i, \dots, c_m\}, \textit{όπου } m \leq n \textit{ και } c_i = \{X_i \mid X_i \subseteq X\}$$

$$\textit{που μεγιστοποιεί την απόσταση μεταξύ των ομάδων και ελαχιστοποιεί}$$

$$\textit{.τη διακύμανση εντός των ομάδων.}$$

$$\textit{Δηλαδή } \forall\, i_1, i_2, j \textit{ τέτοια ώστε } T_{i_1}, T_{i_2} \in c_i \textit{ και } T_j \in c_j, \textit{ισχύει } D(T_{i_1}, T_{i_2}) \ll D(T_{i_1}, T_j)$$

2.3 Μετρικές και Μέθοδοι Αξιολόγησης

Οι μετρικές αξιολόγησης παρέχουν ποσοτικές πληροφορίες για την εκτίμηση της επίδοσης και της ευρωστίας των μοντέλων μηχανικής μάθησης και αποτελούν πολύτιμα εργαλεία για την διαδικασία της επιλογής της κατάλληλης μεθόδου εκμάθησης και της βελτίωσης των μοντέλων. Η διαδικασία της ανάπτυξης ενός μοντέλου στηρίζεται σε μια εποικοδομητική αρχή ανατροφοδότησης όπου επαναληπτικά προσαρμόζουμε και ρυθμίζουμε το μοντέλο με βάση τα αποτελέσματα των εν λόγω μετρικών. Για κάθε πεδίο εφαρμογής μηχανικής μάθησης υπάρχουν κατάλληλες μετρικές αξιολόγησης και η επιλογή τους γίνεται σύμφωνα με τον στόχο της εφαρμογής. Στα υποκεφάλαια 2.3.1 και 2.3.2 παρουσιάζονται βασικές μετρικές αξιολόγησης για εφαρμογές Ταξινόμησης και Ομαδοποίησης αντίστοιχα. Οι εν λόγω μετρικές αξιοποιήθηκαν για την αξιολόγηση των εφαρμογών μηχανικής μάθησης που υλοποιήθηκαν στα πλαίσια της διπλωματικής.

2.3.1 Μετρικές Αξιολόγησης Ταξινόμησης

Για την αξιολόγηση των μοντέλων Ταξινόμησης απαιτείται η κατανόηση βασικών μέτρων που παρουσιάζονται στη συνέχεια. Οι εν λόγω μετρικές και μέθοδοι αφορούν μοντέλα δυαδικής ταξινόμησης (binary classifiers). Συνεπώς, ορίζονται δύο κλάσεις, η κλάση Positive (ή αλλιώς True) και η κλάση Negative (η αλλίως False). Παράδειγμα δυαδικής ταξινόμησης αποτελεί μια εφαρμογή φίλτρου ηλεκτρονικής αλληλογραφίας η οποία ταξινομεί τα μηνύματα είτε στη κατηγορία επιθυμητών είτε στη κατηγορία μη επιθυμητών μηνυμάτων. Η κλάση Negative μοντελοποιεί τη κατηγορία επιθυμητά και η κλάση Positive μοντελοποιεί τη κατηγορία μη επιθυμητά. Η εφαρμογή αυτή ορίζει ένα "Spam Filter". Σημειώνεται ότι οι εν λόγω μετρικές αξιολόγησης μοντέλων ταξινόμησης μπορούν να γενικευτούν και εφαρμόζονται σε μοντέλα ταξινόμησης πολλών κλάσεων (multiclass classifiers) μέσω στρατηγικών υπολογισμού του μέσου όρου των κλάσεων.

Παρουσίαση βασικών μέτρων αξιολόγησης :

- **True Positives (TP) :** Ο αριθμός των παρατηρήσεων που ταξινομήθηκαν ορθά στη κλάση Positive.
- **True Negatives (TN) :** Ο αριθμός των παρατηρήσεων που ταξινομήθηκαν ορθά στη κλάση Negative.
- **False Positives (FP) :** Ο αριθμός των παρατηρήσεων που ταξινομήθηκαν εσφαλμένα στη κλάση Positive.
- **False Negatives (FN) :** Ο αριθμός των παρατηρήσεων που ταξινομήθηκαν εσφαλμένα στη κλάση Negative.

Με βάση τα παραπάνω υλοποιείται ο **Πίνακας Σύγχυσης (Confusion Matrix)**, ένα πολύτιμο εργαλείο αξιολόγησης που προσφέρει μια βοηθητική αναπαράσταση της επίδοσης του μοντέλου στην ταξινόμηση των παρατηρήσεων σε διάφορες κατηγορίες – κλάσεις. Με τη βοήθεια του μπορούμε να παρατηρήσουμε σε ποιες κλάσεις παρουσιάζει αδυναμία το μοντέλο. Ο κάθετος άξονας του πίνακα αναφέρεται στις πραγματικές κλάσεις των παρατηρήσεων και ο οριζόντιος άξονας αναφέρεται στις κλάσεις που προβλέφθηκαν από το μοντέλο.

Πίνακας 2.1 : Μορφή Πίνακα Σύγχυσης δυαδικής ταξινόμησης.

Actual Class		*Predicted Positive*	*Predicted Negative*
	Actual Positive	**TP**	**FN**
	Actual Negative	**FP**	**TN**

Predicted Class

Σε προβλήματα ταξινόμησης πολλών κλάσεων (multiclass classification) ο Πίνακας Σύγχυσης έχει τη παρακάτω μορφή.

Πίνακας 2.2 : Μορφή Πίνακα Σύγχυσης ταξινόμησης πολλών κλάσεων.

Actual Class		Predicted Class 1	Predicted Class 2	……	Predicted Class n-1	Predicted Class n
	Actual Class 1	TP_1	$E_{1,2}$	……	$E_{1,n-1}$	$E_{1,n}$
	Actual Class 2	$E_{2,1}$	TP_2	……	$E_{2,n-1}$	$E_{2,n}$
	…….	……...	……...	……	……	……...
	Actual Class n-1	$E_{n-1,1}$	$E_{n-1,2}$	……	TP_{n-1}	$E_{n-1,n}$
	Actual Class n	$E_{n,1}$	$E_{n,2}$	……...	$E_{n,n-1}$	TP_n

Predicted Class

Με βάση τον Πίνακα 2.2 τα μέτρα *TP, TN, FP, FN* για κάθε κλάση υπολογίζονται ως εξής:

- $\boldsymbol{TP_i}$ **:** Το στοιχείο (i, i) της διαγωνίου.
- $\boldsymbol{TN_i}$ **:** Το άθροισμα όλων των στοιχείων εκτός αυτών που ανήκουν στη στήλη i και στη σειρά i.
- $\boldsymbol{FP_i}$ **:** Το άθροισμα όλων των στοιχείων της i στήλης εκτός του στοιχείου (i, i) της διαγωνίου.
- $\boldsymbol{FN_i}$ **:** Το άθροισμα όλων των στοιχείων της i σειράς εκτός του στοιχείου (i, i) της διαγωνίου.
- Ο συνολικός αριθμός των παρατηρήσεων που ανήκουν στην κλάση i είναι το άθροισμα $TP_i + FN_i$.

Με βάση τα προαναφερθέντα μέτρα αξιολόγησης ορίζονται τα εξής μέτρα αξιολόγησης :

- **Accuracy** $\triangleq \frac{TP+TN}{TP+TN+FP+FN}$
- **Precision** $\triangleq \frac{TP}{TP+FP}$
- **Recall** $\triangleq \frac{TP}{TP+FN}$
- **Specificity** $\triangleq \frac{TN}{TN+FP}$
- **False Positive Rate** $\triangleq \frac{FP}{FP+TN}$

Σημειώνεται ότι στη βιβλιογραφία το μέτρο Specificity αναφέρεται και ως True Negative Rate. Επίσης το Precision αναφέρεται και ως Type I Error. Ακόμα, το Recall αναφέρεται είτε ως Type II Error είτε ως Sensitivity είτε ως True Positive Rate. Επίσης, ως σφάλμα κατηγοριοποίησης (classification error) ορίζεται η ποσότητα $1 - Accuracy$.

Επιπρόσθετα ορίζονται τα παρακάτω μέτρα αξιολόγησης :

- **F1 Score** $\triangleq 2 * \frac{Precision*Recall}{Precision+Recall}$
- **F beta Score** $\triangleq (1 + \beta^2) * \frac{Precision*Recall}{(\beta^2*Precision)+Recall}$
- **Balanced Accuracy** $\triangleq \frac{Specificity+Recall}{2}$
- **Jaccard Similarity Index** $\triangleq \frac{|Y_{true} \cap Y_{pred}|}{|Y_{true} \cup Y_{pred}|}$

Όλα τα μέτρα αξιολόγησης που παρουσιάστηκαν μετά τον Πίνακα Σύγχυσης παίρνουν τιμές στο διάστημα [0,1]. Για όλα τα εν λόγω μέτρα εκτός του False Positive Rate η καλύτερη τιμή που αντιστοιχεί σε ιδανικό μοντέλο ταξινόμησης είναι η μονάδα ενώ για το False Positive Rate η καλύτερη τιμή είναι το μηδέν. Η παράμετρος β στο F beta Score καθορίζει το βάρος του μέτρου Recall στο αποτέλεσμα και παίρνει τιμές στο διάστημα $[0, +\infty)$. Για τιμές του $\beta < 1$ μεγαλύτερη επιρροή έχει το Precision ενώ για τιμές του $\beta > 1$ επικρατεί η επιρροή του Recall. Στη βιβλιογραφία η μετρική F-Score αναφέρεται και ως F-Measure.

Είναι ιδιαίτερα σημαντικό να κατανοήσει κανείς σε βάθος τη σημασία αυτών των μέτρων αξιολόγησης ώστε να αποφεύγονται τυχόν παρερμηνείες και λανθασμένες εντυπώσεις όταν εξετάζεται ένα μοντέλο ως προς την επίδοση του.

Το μέτρο αξιολόγησης Accuracy αναφέρεται στο ποσοστό των παρατηρήσεων που έχουν κατηγοριοποιηθεί σωστά από τον συνολικό αριθμό των παρατηρήσεων του συνόλου εισόδου. Σημειώνεται ότι σε εφαρμογές όπου το σύνολο δεδομένων εισόδου παρουσιάζει μεγάλη ανισορροπία μεταξύ των κλάσεων, το εν λόγω μέτρο κρίνεται ακατάλληλο. Για παράδειγμα, σε περίπτωση που τα δεδομένα απαρτίζονται από 100 στιγμιότυπα μιας κλάσης και 10 στιγμιότυπα άλλης κλάσης, ένα μοντέλο που τυφλά κατηγοριοποιεί όλα τα δεδομένα στη πρώτη κλάση θα παρουσιάζει 90% Accuracy score. Το Precision αναφέρεται στον λόγο των παρατηρήσεων που έχουν κατηγοριοποιηθεί σωστά σε μια κατηγορία προς το άθροισμα τους με τις παρατηρήσεις που εσφαλμένα κατηγοριοποιήθηκαν στην ίδια κατηγορία. Το Recall αναφέρεται στον λόγο των παρατηρήσεων που κατηγοριοποιήθηκαν σωστά σε μια κατηγορία προς τον συνολικό αριθμό των παρατηρήσεων που ανήκουν σε αυτή τη κατηγορία. Το Specificity αναφέρεται στην ικανότητα του μοντέλου να κατηγοριοποιεί σωστά παρατηρήσεις μιας συγκεκριμένης κλάσης. Το False Positive Rate αποτελεί συμπληρωματική έννοια του Specificity, δηλαδή την ανικανότητα του μοντέλου να κατηγοριοποιεί σωστά παρατηρήσεις μιας κλάσης. Τα μέτρα F1 Score, F beta Score και Balanced Accuracy αποτελούν συνδιασμούς μετρικών και είναι κατάλληλα για την αξιολόγηση μοντέλων όποτε τα δεδομένα εισόδου παρουσιάζουν ανισορροπία μεταξύ των κλάσεων.

Όπως αναφέραμε στην αρχή, η γενίκευση των μετρικών αξιολόγησης στη περίπτωση μοντέλων ταξινόμησης πολλών κλάσεων επιτυγχάνεται με την αξιοποίηση στρατηγικών υπολογισμού μέσων όρων. Υπάρχουν τρεις στρατηγικές που μπορεί να ακολουθήσει κανείς, την **micro**, την **macro** και την **weighted** Averaging.

1. **micro Averaging** **:** Χρησιμοποιούνται απευθείας στη σχέση υπολογισμού της μετρικής όλα τα μέτρα (TP, TN, FP, FN) που ενδεχομένως αξιοποιούνται, όλων των κλάσεων του συστήματος

2. **macro Averaging** **:** Μέση τιμή των μετρικών όλων των κλάσεων του συστήματος.

3. **weighted Averaging :** Σταθμισμένη μέση τιμή των μετρικών όλων των κλάσεων του συστήματος.

Στη συνέχεια παρουσιάζουμε παραδείγματα υπολογισμού του μέσου Recall για την αξιολόγηση μοντέλων ταξινόμησης πολλών κλάσεων με βάση τις παραπάνω στρατηγικές. Με αντίστοιχο τρόπο υπολογίζονται οι μέσες τιμές των υπολοίπων μετρικών που ήδη έχουμε αναλύσει.

Έστω ότι υπάρχουν n κλάσεις – κατηγορίες και το σύνολο δεδομένων εισόδου αποτελείται από k στιγμιότυπα με k_i τα στιγμιότυπα της κλάσης i. Ο υπολογισμός του μέσου Recall μοντέλων ταξινόμησης πολλών κλάσεων για κάθε στρατηγική έχει ως εξής :

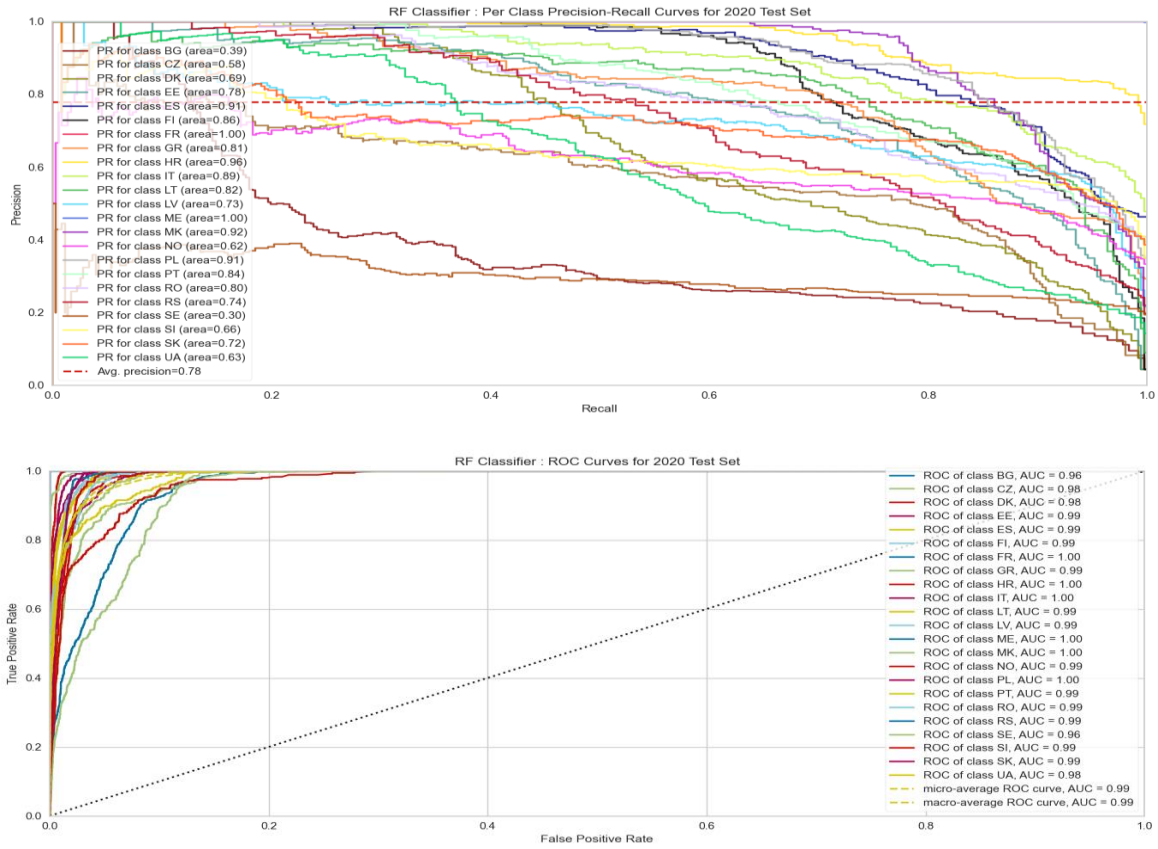
- micro Recall $= \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i}$
- macro Recall $= \frac{\sum_{i=1}^{n} Recall_i}{n}$
- weighted Recall $= \sum_{i=1}^{n} \frac{k_i}{k} * Recall_i$

Η φύση και οι ιδιαιτερότητες της εφαρμογής που καλείται να υλοποιήσει το εκάστοτε μοντέλο ταξινόμησης καθορίζουν την σωστή επιλογή στρατηγικής. Επομένως, αξίζει να αναλύσουμε επαρκώς τον μηχανισμό κάθε στρατηγικής.

Όσον αφορά τη στρατηγική weighted αποδίδεται μεγαλύτερη βαρύτητα στις πλειοψηφικές κλάσεις. Σε περίπτωση που τα δεδομένα εισόδου παρουσιάζουν μεγάλη ανισορροπία κλάσεων και ταυτόχρονα μας ενδιαφέρουν ισοδύναμα τα στιγμιότυπα κάθε κλάσης, η στρατηγική αυτή προφανώς δεν αποτελεί τη βέλτιστη επιλογή. Η στρατηγική macro υπολογίζει τη μετρική ανεξάρτητα για κάθε κλάση και στη συνέχεια υπολογίζει τον μέσο όρο αυτών. Συνεπώς, αντιμετωπίζει ισοδύναμα όλες τις κλάσεις. Αποτελεί καλή επιλογή όταν μας ενδιαφέρει η συνολική επίδοση του μοντέλου και η ανάλυση δεν απαιτεί τον διαχωρισμό των κλάσεων σε ιεραρχίες με βάση τη σπουδαιότητα τους. Αντίθετα, η στρατηγική micro συγκεντρώνει τις συνεισφορές όλων των κλάσεων για τον υπολογισμό της μέσης μετρικής. Συνεπώς, σε εφαρμογές μοντέλων ταξινόμησης πολλών κλάσεων η στρατηγική micro αποτελεί καλύτερη λύση όταν το σύνολο των δεδομένων εισόδου παρουσιάζει σημαντική ανισορροπία μεταξύ των κλάσεων και ταυτόχρονα μας ενδιαφέρουν ιδιαίτερα τα στιγμιότυπα της κλάσης μειονότητας.

Δύο ακόμα σημαντικά εργαλεία για την αξιολόγηση μοντέλων ταξινόμησης είναι οι καμπύλες **Precision-Recall (PR Curves)** και οι καμπύλες **Receiver Operating Characteristics (ROC Curves)**. Αποτελούν γραφικές μεθόδους αξιολόγησης μοντέλων δυαδικής ταξινόμησης και η γενίκευση τους για την αξιολόγηση ταξινομητών πολλών κλάσεων επιτυγχάνεται είτε μέσω των στρατηγικών που έχουμε ήδη αναφέρει, είτε με την απεικόνηση των καμπυλών κάθε κλάσης στο ίδιο γράφημα (class reference scheme). Σε ένα διάγραμμα ROC ο άξονας των x αναφέρεται στο μέτρο False Positive Rate ενώ ο άξονας των y στο μέτρο Recall. Αντίστοιχα, σε ένα διάγραμμα PR ο άξονας των x αναφέρεται στο μέτρο Recall ενώ ο άξονας των y στο μέτρο Precision.

Η καμπύλη ROC δείχνει τη σχέση μεταβολής των μέτρων Recall και False Positive Rate ενός μοντέλου σε διάφορες ρυθμίσεις κατωφλίου. Οι ρυθμίσεις κατωφλίου αναφέρονται στις πιθανές οριακές τιμές διαχωρισμού των κλάσεων που καθορίζονται από το εκάστοτε μοντέλο. Αντίστοιχα, η καμπύλη PR δείχνει τη σχέση μεταβολής των μέτρων Recall και Precision.



Σχήμα 2.1 : Παράδειγμα καμπυλών PR (άνω διάγραμμα) και ROC (κάτω διάγραμμα) μοντέλου κατηγοριοποίησης πολλών κλάσεων.

Το σημείο (1,1) σε μια καμπύλη PR αντιστοιχεί σε ένα ιδανικό μοντέλο ενώ το σημείο (0,0) σε ένα προβληματικό μοντέλο με μηδενική ικανότητα ταξινόμησης. Αντίστοιχα, το σημείο (0,1) σε μια καμπύλη ROC αντιστοιχεί σε ένα ιδανικό μοντέλο ενώ το σημείο (1,0) σε ένα προβληματικό μοντέλο.

Μια μέθοδος αξιολόγησης μοντέλων ταξινόμησης είναι ο υπολογισμός του εμβαδού της δισδιάστατης περιοχής κάτω από την καμπύλη PR και ROC αντίστοιχα. Το εμβαδό κάτω από τη καμπύλη (Area Under the Curve - AUC) αντικατοπτρίζει τη συνολική επίδοση του μοντέλου [22]. Ιδανικά μοντέλα παρουσιάζουν μοναδιαίο εμβαδό ενώ προβληματικά μοντέλα παρουσιάζουν $AUC \leq 0.5$.

Σημειώνεται ότι ένα μοντέλο που βελτιστοποιεί το εμβαδό κάτω από τη καμπύλη ROC δεν βελτιστοποιεί απαραίτητα και το εμβαδό κάτω από τη καμπύλη PR [23]. Επίσης, αξίζει να αναφερθεί το γεγονός πως οι καμπύλες ROC παρουσιάζουν συνήθως υπερβολικά αισιόδοξη άποψη για την επίδοση των μοντέλων όταν το σύνολο δεδομένων εισόδου διέπεται από μεγάλο βαθμό ασυμμετρίας (skewness) [23]. Σε εφαρμογές με δεδομένα που διέπονται από ασύμμετρες κατανομές η καμπύλη PR προσφέρει μια πιο αντιπροσωπευτική εικόνα της επίδοσης των μοντέλων.

2.3.2 Μετρικές Αξιολόγησης Ομαδοποίησης

Η αξιολόγηση των μοντέλων Ομαδοποίησης καθώς και η αξιολόγηση των εξαγόμενων ομάδων είναι μια δύσκολη διαδικασία και αποτελεί ανοιχτό πρόβλημα [35, 41]. Δεδομένου ότι στην Μη Επιβλεπόμενη Μάθηση τα δεδομένα εισόδου είναι κατά κανόνα μη επισημασμένα, ο ορισμός των ομάδων είναι γενικά υποκειμενικός και εξαρτάται κυρίως από το πεδίο εφαρμογής και τον χρήστη. Στη προσπάθεια αξιολόγησης της ποιότητας των αποτελεσμάτων των εφαρμογών ομαδοποίησης, η επιστημονική κοινότητα έχει επινοήσει μετρικές που στη βιβλιογραφία αναφέρονται ως **Δείκτες Επικύρωσης Ομαδοποίησης (Clustering Validation Indices –CVI)**. Εκτενής συγκριτική μελέτη από τους ερευνητές Arbelaitz, Gurrutxaga, Muguerza, Pérez και Perona υποστηρίζει ότι δεν υπάρχει μετρική που υπερέχει έναντι των υπολοίπων σε όλα τα πλαίσια εφαρμογών [36]. Παρ' όλα αυτά, η εν λόγω έρευνα υπέδειξε ένα σύνολο τεσσάρων μετρικών που ξεχώρισαν για την συνέπεια και την αποδοτικότητά τους. Οι τέσσερεις αυτές μετρικές αξιοποιήθηκαν στις εφαρμογές ομαδοποίησης χρονοσειρών ηλεκτρικού φορτίου που υλοποιήθηκαν στα πλαίσια της εν λόγω διπλωματικής.

Για λόγους πληρότητας αναφέρεται ότι οι μετρικές CVI χωρίζονται σε τρεις κατηγορίες : τις **εσωτερικές (internal)**, τις **εξωτερικές (external)** και τις **σχετικές (relative)** μετρικές επικύρωσης [35]. Οι εσωτερικές μετρικές βασίζονται μόνο στο σύνολο δεδομένων εισόδου, οι εξωτερικές βασίζονται στην εκ των προτέρων γνώση του σχηματισμού των ομάδων (ετικέτες), ενώ οι σχετικές μετρικές αξιοποιούνται για την σύγκριση διαφορετικών ομαδοποιήσεων που πραγματοποιεί ένα μοντέλο όταν μεταβάλλονται οι παράμετροι του. Στην εν λόγω διπλωματική εργασία θα ασχοληθούμε μόνο με εσωτερικές και σχετικές μετρικές επικύρωσης καθώς οι εξωτερικές μετρικές εφαρμόζονται κυρίως σε ομαδοποιήσεις συνθετικών δεδομένων και αξιοποιούνται για τη συγκριτική ανάλυση αλγορίθμων ομαδοποίησης [27].

Στη συνέχεια παρουσιάζονται βασικές έννοιες, απαραίτητες για την κατανόηση των δεικτών επικύρωσης ομαδοποίησης.

I. <u>Διαχωρισμός Ομάδων (Cluster Separation)</u> :

Αναφέρεται στον βαθμό διαχωρισμού των ομάδων, δηλαδή στον βαθμό που δύνανται να ξεχωρίζουν οι ομάδες μεταξύ τους. Ο διαχωρισμός των ομάδων υπολογίζεται με το μέτρο BCSS (Between Cluster Sum of Squares).

$$BCSS = \sum_{i=1}^{N_C} |C_i| * d\left(\bar{x}_{c_i}, \bar{x}\right)^2 \quad (2.3)$$

Όπου :

N_c : # *ομάδων*, C_i : *η i ομάδα*, $\bar{x}_{c_i}$: *το κέντρο της i ομάδας και* $\bar{x}$: *η μέση τιμή όλων των ομάδων*.

Η μεγιστοποίηση του διαχωρισμού των ομάδων επιτυγχάνεται με τη μεγιστοποίηση του BCSS.

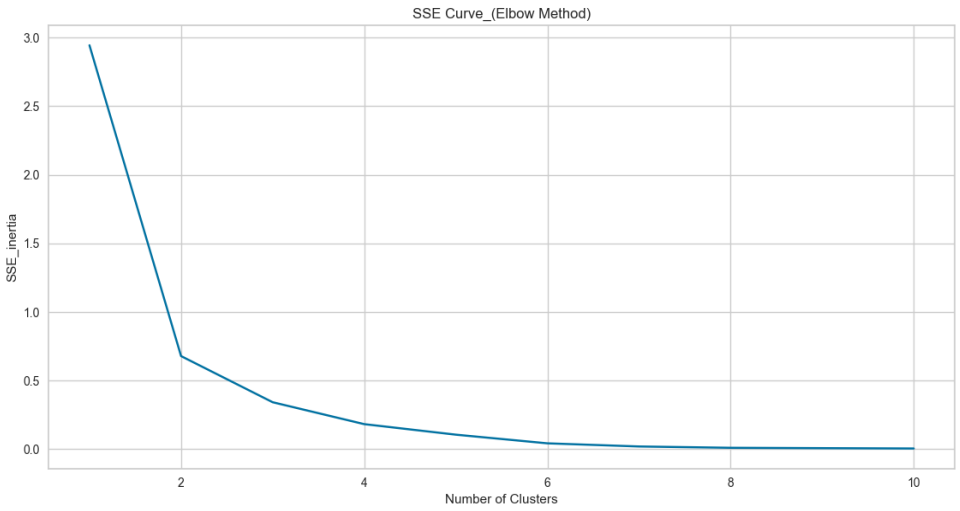
II. Συνοχή Ομάδων (Cluster Cohesion – Compactness) :

Αναφέρεται στον βαθμό συνοχής των ομάδων, δηλαδή στο βαθμό συμπακτότητας των παρατηρήσεων που απαρτίζουν τις ομάδες. Η συνοχή των ομάδων υπολογίζεται με το μέτρο WCSS (Within Cluster Sum of Squares).

$$WCSS = \sum_{i=1}^{N_c} \sum_{x \, \epsilon \, C_i} d\left(\bar{x}_{c_i}, x\right)^2 \tag{2.4}$$

Η μεγιστοποίηση της συνοχής των ομάδων επιτυγχάνεται με την ελαχιστοποίηση του WCSS.

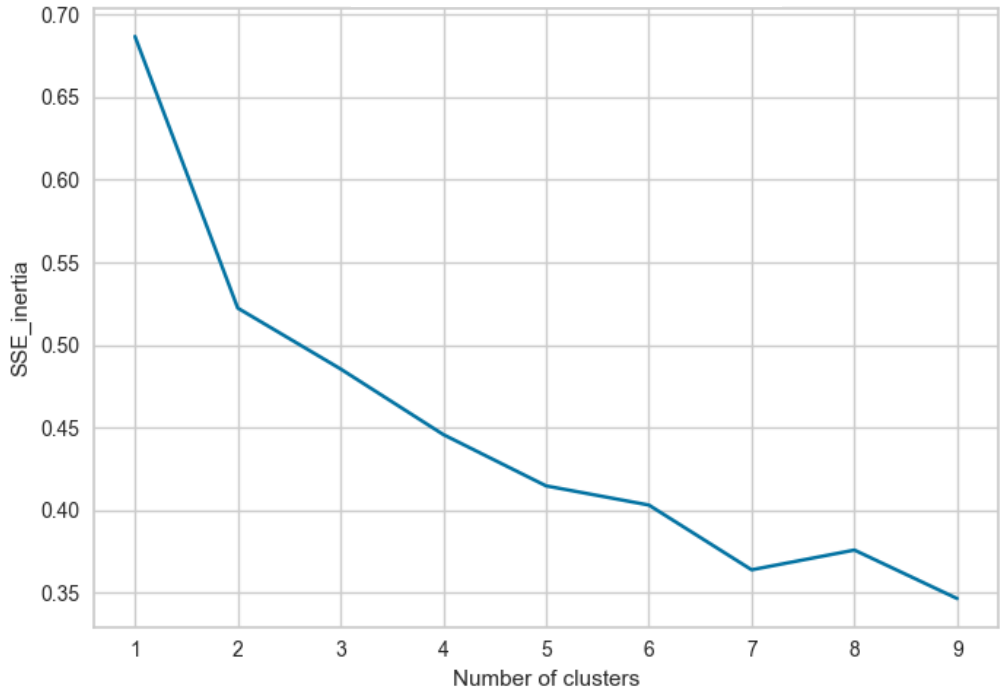
Το μέτρο WCSS αναφέρεται στη βιβλιογραφία ως SSE (Sum of Squares Error) και ως Inertia. Μια μέθοδος για τον υπολογισμό της "βέλτιστης" ομαδοποίησης είναι ο σχεδιασμός του αποτελέσματος του SSE συναρτήσει όλων των δυνατών ομαδοποιήσεων ($N_c = 1, \dots, n$) που υλοποιεί ένας αλγόριθμος ομαδοποίησης. Η μέθοδος αυτή αναφέρεται ως "Elbow Plot Method" και το γράφημα που προκύπτει έχει τυπικά την παρακάτω μορφή :



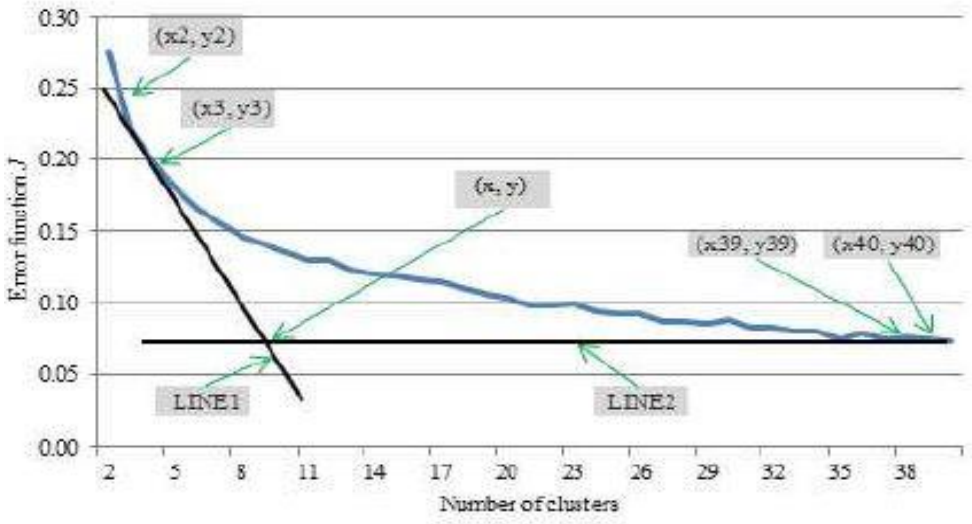
Σχήμα 2.2 : Παράδειγμα καμπύλης SSE (Elbow Plot Method) για τον προσδιορισμό των ομάδων.

Με βάση το σχήμα 2.2 παρατηρούμε ότι όσο αυξάνεται ο αριθμός των ομάδων τόσο μειώνεται το SSE. Η μείωση του SSE μεταφράζεται σε αύξηση της συνοχής των ομάδων όπως ήδη έχουμε αναφέρει. Το εν λόγω φαινόμενο είναι απόλυτα λογικό αν σκεφτεί κανείς ότι στην ακραία περίπτωση που ο αριθμός των ομάδων ισούται με τον αριθμό των παρατηρήσεων (κάθε παρατήρηση αποτελεί μια ομάδα), το SSE θα ισούται με το μηδέν καθώς κάθε παρατήρηση θα αποτελεί το κέντρο της ομάδας στην οποία ανήκει. Συνεπώς, η απόσταση κάθε παρατήρησης από το κέντρο της ομάδας της θα είναι μηδενική. Με βάση αυτή τη λογική, ο προσδιορισμός της "βέλτιστης" ομαδοποίησης υποδεικνύεται από το γόνατο της καμπύλης όπου στο παράδειγμα του σχήματος 2.2 παρουσιάζεται στις δύο ομάδες. Στο σημείο αυτό παρατηρείται δραματική αλλαγή στην κλίση της καμπύλης και αποτελεί ένα κατώφλι πέραν του οποίου η ομαδοποίηση, σύμφωνα με τη συγκεκριμένη μετρική κρίνεται παθογενική.

Στη περίπτωση που το σύνολο δεδομένων προς ομαδοποίηση διέπεται από περίπλοκες δομές, η καμπύλη SSE ενδέχεται είτε να παρουσιάζει πολλαπλά γόνατα είτε να μην παρουσιάζει δραματική αλλαγή στη κλίση της. Όσον αφορά τις καμπύλες SSE με πολλαπλά γόνατα, η "βέλτιστη" επιλογή γονάτου απαιτεί τη κρίση του χρήστη και εξαρτάται από τη φύση του προβλήματος και τον στόχο της εφαρμογής.



Σχήμα 2.3 : Παράδειγμα καμπύλης SSE με πολλαπλά γόνατα.



Σχήμα 2.4 : Παράδειγμα καμπύλης SSE που δεν παρουσιάζει γόνατο [9].

Στη περίπτωση που η καμπύλη SSE δεν παρουσιάζει γόνατο, δηλαδή δεν παρουσιάζεται ακραία μεταβολή στη κλίσης της, ο υπολογισμός του "γονάτου" επιτυγχάνεται μέσω του υπολογισμού του σημείου διασταύρωσης των ασυμπτώτων [9] όπως υποδεικνύεται στο σχήμα 2.4.

Η μετρική SSE και κατά συνέπεια η μέθοδος Elbow Plot αξιοποιεί μόνο τη πληροφορία της συνοχής των ομάδων για την υπόδειξη της "βέλτιστης" ομαδοποίησης. Στη συνέχεια, παρουσιάζουμε τρεις υβριδικές μετρικές αξιολόγησης που συνδιάζουν τη πληροφορία συνοχής και διαχωρισμού των ομάδων. Σημειώνεται επίσης ότι αποτελεί καλή πρακτική η σχεδίαση των αποτελεσμάτων των εν λόγω μετρικών συναρτήσει του αριθμού των ομάδων, όπως αντίστοιχα υποδείξαμε με την μέθοδο Elbow Plot της μετρικής SSE.

i. Calinski – Harabasz Index (CH CVI)

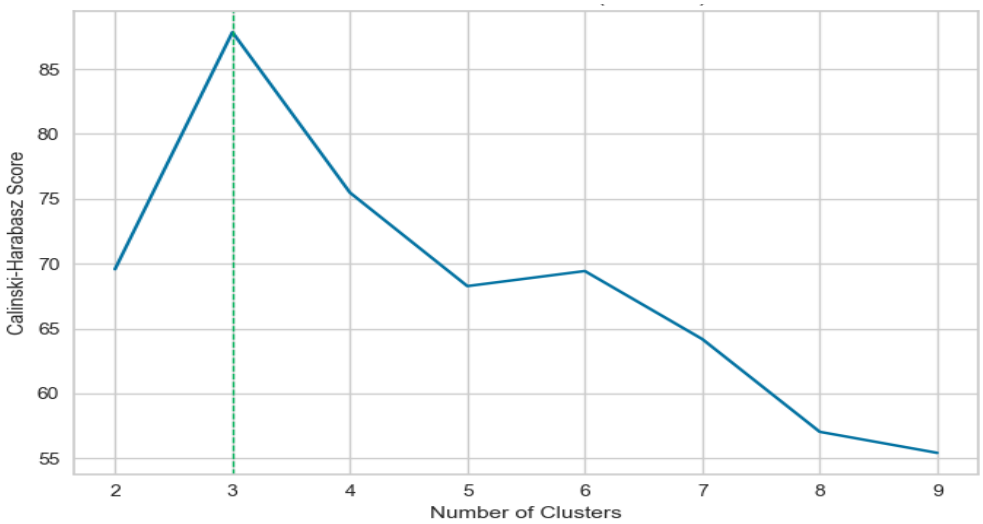
Ο δείκτης επικύρωσης ομαδοποίησης **CH** προτάθηκε από τους Calinski και Harabasz το 1974 [37] και υπολογίζεται ως εξής :

$$CH(k) = \frac{N-k}{k-1} * \frac{BCSS(k)}{WCSS(k)} \tag{2.5}$$

Όπου :

k : # *ομάδων με* $k > 1$, N: # *παρατηρήσεων και* $BCSS, WCSS$ *από τις σχέσεις* 2.3 *και* 2.4 *αντίστοιχα*.

Για τον υπολογισμό της "βέλτιστης" ομαδοποίησης ψάχνουμε εκείνο το k, δηλαδή τον αριθμό των ομάδων, που μεγιστοποιεί το αποτέλεσμα του δείκτη CH. Εναλλακτικά, μέσω του γραφήματος των αποτελεσμάτων του CH συναρτήσει του k ψάχνουμε τη λύση που δίνει μια κορυφή ή ένα απότομο γόνατο στη καμπύλη. Αντιθέτως, εάν το γράφημα είναι ομαλό (οριζόντιο ή αύξον ή κατηφορικό), τότε δεν υπάρχει λόγος να προτιμήσουμε μια λύση έναντι των υπολοίπων.



Σχήμα 2.5 : Παράδειγμα καμπύλης CH για την επιλογή ομάδων.

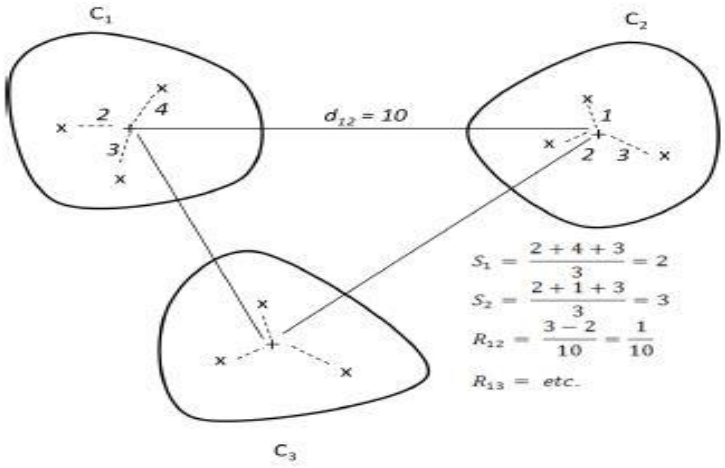
ii. Davies – Bouldin Index (DB CVI)

Ο δείκτης επικύρωσης ομαδοποίησης **DB** προτάθηκε από τους Davies και Bouldin το 1979 [38] και υπολογίζεται ως εξής :

$$DB(k) = \frac{1}{k} * \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \qquad (2.6) \qquad \text{και} \qquad R_{ij} = \frac{s_i + s_j}{d_{ij}} \qquad (2.7)$$

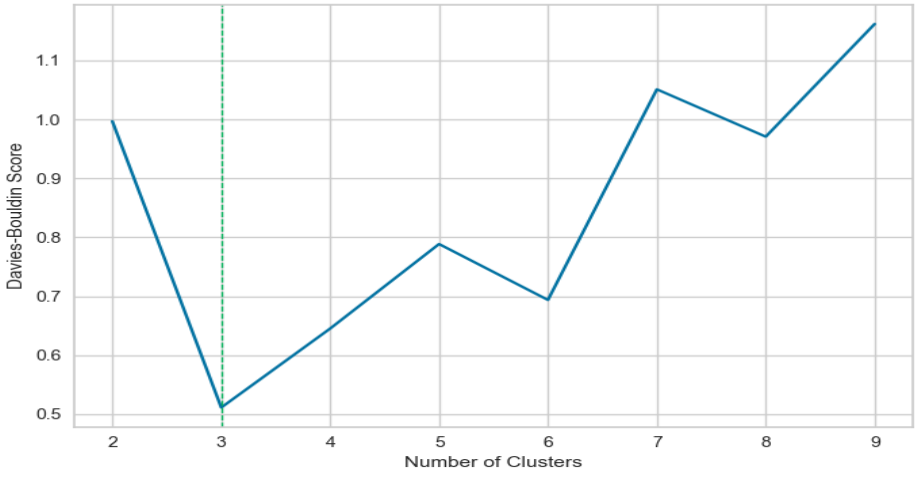
Όπου :

- s_i : η μέση απόσταση μεταξύ κάθε παρατήρησης της ομάδας i και του κέντρου της ομάδας i *(αναφέρεται και ως διάμετρος της ομάδας).*
- d_{ij} : η απόσταση μεταξύ των κέντρων των ομάδων i και j.



Σχήμα 2.6 : Παράδειγμα υπολογισμού του δείκτη επικύρωσης DB [39].

Για τον υπολογισμό της "βέλτιστης" ομαδοποίησης ψάχνουμε εκείνο το k, δηλαδή τον αριθμό των ομάδων, που ελαχιστοποιεί το αποτέλεσμα του δείκτη DB. Εναλλακτικά, μέσω του γραφήματος των αποτελεσμάτων του DB συναρτήσει του k ψάχνουμε τη λύση που δίνει το κατώτατο σημείο στη καμπύλη.



Σχήμα 2.7 : Παράδειγμα καμπύλης DB για την επιλογή ομάδων.

iii. Silhouette Score Index (Silhouette CVI)

Ο δείκτης επικύρωσης ομαδοποίησης **Silhouette** ή αλλιώς **Silhouette Width Coefficient (SWC)** προτάθηκε από τον Rousseeuw το 1987 [42] και με τη συμβολή του Kaufman [65] υπολογίζεται ως εξής :

$$MeanSilhouette(k) = SWC(k) = \frac{\sum_{i=1}^{k} \frac{\sum_{j=1}^{|C_i|} s(x_j)}{|C_i|}}{k} \tag{2.8}$$

Όπου :

$$s(x_j) = \frac{b(x_j) - a(x_j)}{\max\{a(x_j), b(x_j)\}} \quad \text{ή} \quad s(x_j) = \begin{cases} 1 - \frac{a(x_j)}{b(x_j)} & \text{εάν } a(x_j) < b(x_j) \\ 0 & \text{εάν } a(x_j) = b(x_j) \\ \frac{b(x_j)}{a(x_j)} - 1 & \text{εάν } a(x_j) > b(x_j) \end{cases} \tag{2.9}$$

$$a(x_j) = \frac{1}{|C_j| - 1} * \sum_{x_i \in C_j,\ x_j \neq x_i} d(x_i, x_j) \tag{2.10}$$

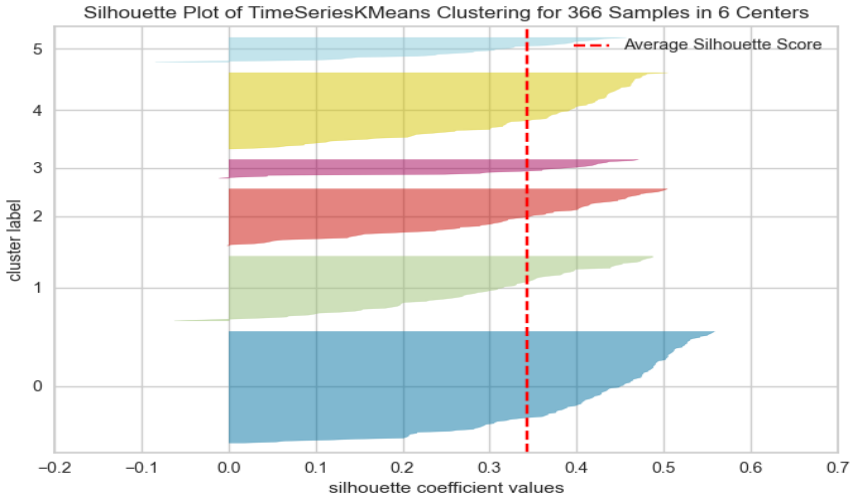
$$b(x_j) = \min_{C_k \neq C_j} \frac{1}{|C_k|} * \sum_{x_i \in C_k} d(x_i, x_j) \tag{2.11}$$

Το $a(x_j)$ αναφέρεται στη συνοχή (cohesion) της παρατήρησης x_j με τις παρατηρήσεις της ομάδας στην οποία ανήκει. Το $b(x_j)$ αναφέρεται στον βαθμό διαχωρισμού (separation) της παρατήρησης x_j από τις παρατηρήσεις που ανήκουν στην ομάδα η οποία απέχει την ελάχιστη απόσταση από την ομάδα της x_j.

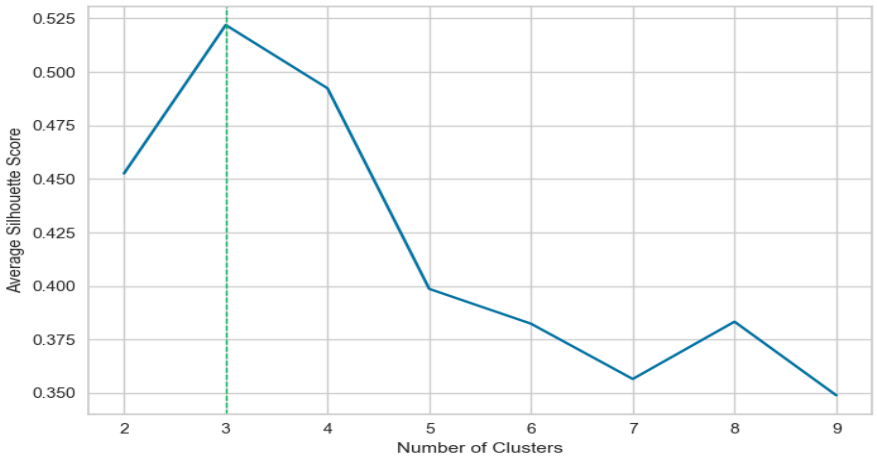
Ο δείκτης $s(x_j)$ παίρνει τιμές στο διάστημα [-1,1] για κάθε παρατήρηση x_j του συνόλου δεδομένων X όπως είναι προφανές από τη σχέση (2.9). Το ίδιο ισχύει και για τον μέσο δείκτη Silhouette όλων των παρατηρήσεων του συνόλου X που δίνεται από τη σχέση (2.8). Αρνητικές τιμές του SWC υποδεικνύουν προβληματικές ομαδοποιήσεις δεδομένων. Θετικές τιμές κοντά στο μηδέν υποδεικνύουν την ύπαρξη αλληλοεπικαλυπτόμενων ομάδων, ενώ τιμές κοντά στη μονάδα υποδεικνύουν "βέλτιστες" ομαδοποιήσεις.

Συνεπώς, για τον υπολογισμό της "βέλτιστης" ομαδοποίησης ψάχνουμε εκείνο το k, δηλαδή τον αριθμό των ομάδων, που μεγιστοποιεί το αποτέλεσμα του δείκτη SWC. Εναλλακτικά, μέσω του γραφήματος των αποτελεσμάτων του SWC συναρτήσει του k ψάχνουμε τη λύση που δίνει τη κορυφή της καμπύλης.

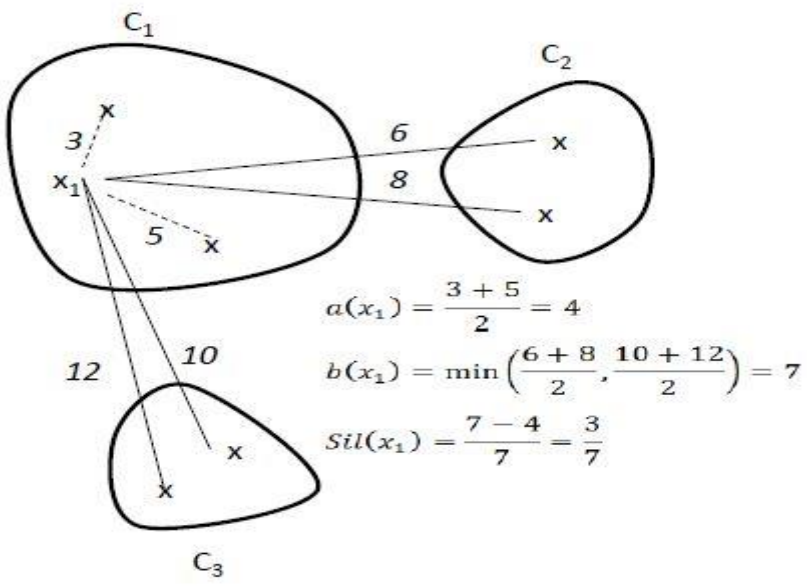
Σημειώνεται ότι ο δείκτης επικύρωσης ομαδοποίησης Silhouette Width Coefficient αναφέρεται στη βιβλιογραφία και ως Silhouette Width Criterion. Επίσης, υπάρχει η δυνατότητα βοηθητικής γραφικής αναπαράστασης της εκάστοτε ομαδοποίησης, που συνδιάζει το αποτέλεσμα κάθε παρατήρησης αλλά και το μέσο συνολικό αποτέλεσμα όλων των παρατηρήσεων με έναν ιδιαίτερα έξυπνο τρόπο διευκολύνοντας την ερμηνεία των αποτελεσμάτων από τον χρήστη.



Σχήμα 2.8 : Παράδειγμα γραφικής αναπαράστασης των δεικτών $s(x_j)$ και SWC**.**



Σχήμα 2.9 : Παράδειγμα καμπύλης SWC για την επιλογή ομάδων.



Σχήμα 2.10 : Παράδειγμα υπολογισμού του δείκτη επικύρωσης Silhouette [39].

Στη γραφική αναπαράσταση του σχήματος 2.8 κάθε "τριγωνοειδές" σχήμα αποτελεί μια ομάδα παρατηρήσεων που έχει δημιουργηθεί και η βάση τους αντικατοπτρίζει τον αριθμό των παρατηρήσεων που ανήκουν στην αντίστοιχη ομάδα. Για παράδειγμα, στο σχήμα 2.8, η ομάδα με τις λιγότερες παρατηρήσεις αντιστοιχεί στο cluster label 3. Τα αποτελέσματα των δεικτών $s(x_j)$ των παρατηρήσεων x_j αποτυπώνονται ως οριζόντιες γραμμές που συνθέτουν τα εν λόγω "τριγωνοειδή" και το μήκος κάθε γραμμής ισούται με την απόλυτη τιμή του score της αντίστοιχης παρατήρησης. Οι οριζόντιες αυτές γραμμές έχουν κατεύθυνση προς τα θετικά του άξονα των x για θετικά score παρατηρήσεων, ενώ για αρνητικά score έχουν κατεύθυνση προς τα αρνητικά του άξονα των x. Συνεπώς, στη περίπτωση που όλες οι παρατηρήσεις x_j μιας ομάδας έχουν το ίδιο score $s(x_j)$, τότε η ομάδα θα αποτυπώνεται στο γράφημα ως ένα τέλειο παραλληλόγραμμο.

2.3.3 Μέτρα Ομοιότητας και Ανομοιότητας

Τα μέτρα ομοιότητας και ανομοιότητας αποτελούν συναρτήσεις που δέχονται ως όρισμα ένα ζεύγος χρονοσειρών ή διανυσμάτων και επιστρέφουν ως αποτέλεσμα έναν θετικό αριθμό που αντιπροσωπεύει τον βαθμό ομοιότητας και ανομοιότητας αντίστοιχα μεταξύ των δύο ορισμάτων. Στη συνέχεια παρουσιάζουμε τον ορισμό των εν λόγω μέτρων [26].

1. **<u>Μέτρο Ομοιότητας</u>**

   Έστω X ένα σύνολο δεδομένων. Μια συνάρτηση $S : X \times X \rightarrow \mathbb{R}$ καλείται μέτρο ομοιότητας όταν $\forall\, x, y \in X$ ικανοποιεί τις ακόλουθες ιδιότητες :

   - Μη αρνητικότητα : $S(x, y) \geq 0$
   - Συμμετρία : $S(x, y) = S(y, x)$
   - Εάν $x \neq y$ τότε : $S(x, x) = S(y, y) > S(x, y)$

   Το πεδίο τιμών μιας συνάρτησης S ανήκει στο διάστημα [0,1]. Όταν τα δύο ορίσματα είναι όμοια τότε επιστρέφει τιμές κοντά στη μονάδα, ενώ όταν είναι ανόμοια επιστρέφει τιμές κοντά στο μηδέν.

2. **<u>Μέτρο Ανομοιότητας</u>**

   Έστω X ένα σύνολο δεδομένων. Μια συνάρτηση $D : X \times X \rightarrow \mathbb{R}$ καλείται μέτρο ανομοιότητας όταν $\forall\, x, y \in X$ ικανοποιεί τις ακόλουθες ιδιότητες :

   - Μη αρνητικότητα : $D(x, y) \geq 0$
   - Συμμετρία : $D(x, y) = D(y, x)$
   - Ανακλαστικότητα : $D(x, x) = 0$

   Το πεδίο τιμών μιας συνάρτησης D ανήκει στο διάστημα [0,∞). Όταν δύο ορίσματα είναι όμοια τότε επιστρέφει τιμές κοντά στο μηδέν.

Κάθε μέτρο ομοιότητας μετασχηματίζεται σε μέτρο ανομοιότητας και αντίστροφα μέσω των παρακάτω σχέσεων :

- $D(x,y) = \big(S(x,x) - S(x,y)\big)$
- $D(x,y) = \frac{S(x,x) - S(x,y)}{S(x,y)}$
- $D(x,y) = \big(S(x,x) - S(x,y)\big)^{\frac{1}{2}}$

Συνήθως ο βαθμός ομοιότητας δύο αντικειμένων μεταφράζεται στην μεταξύ τους απόσταση. Όσον αφορά το μέτρο απόστασης, είναι σημαντικό να κατανοήσουμε εάν μπορεί να θεωρηθεί ως μετρική (metric).

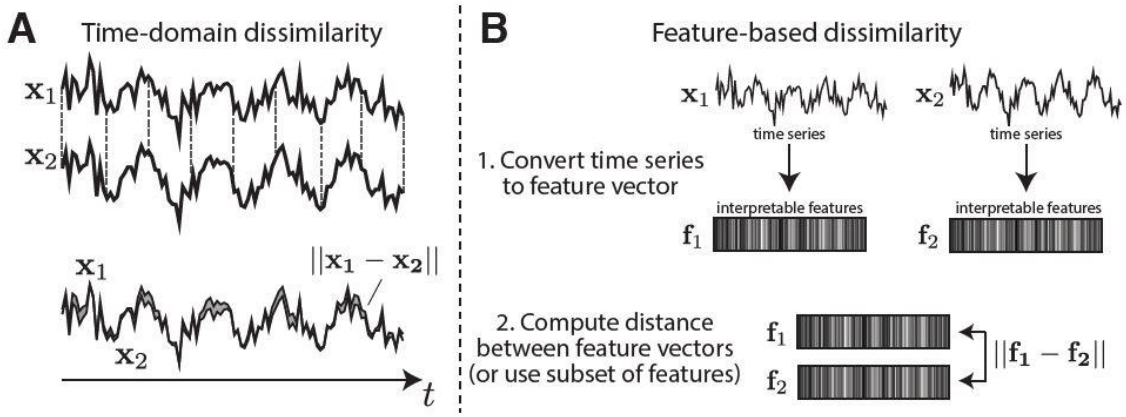
Έστω X ένα σύνολο δεδομένων. Μια συνάρτηση $\mathcal{D} : X \times X \to \mathbb{R}$ καλείται μετρική απόστασης όταν $\forall\, x, y, z \,\epsilon\, X$ ικανοποιεί τις ακόλουθες ιδιότητες :

- Μη αρνητικότητα : $\mathcal{D}(x,y) \geq 0$
- Συμμετρία : $\mathcal{D}(x,y) = \mathcal{D}(y,x)$
- Ανακλαστικότητα : $\mathcal{D}(x,y) = 0$ $\textit{εάν και μόνο εάν}$ $x = y$
- Τριγωνική Ανισότητα : $\mathcal{D}(x,y) \leq \mathcal{D}(x,z) + \mathcal{D}(z,y)$

Εάν οποιαδήποτε από τις παραπάνω ιδιότητες παραβιάζεται, τότε το μέτρο απόστασης δεν μπορεί να θεωρηθεί ως μετρική. Σημειώνεται ότι σε γενικές γραμμές είναι επιθυμητό να ικανοποιούνται οι τέσσερεις αυτές ιδιότητες από ένα μέτρο (αν)ομοιότητας. Παρ' όλα αυτά, ένα μέτρο που δεν αποτελεί μετρική δύναται να είναι εξίσου αποτελεσματικό [26].

Όσον αφορά τη σύγκριση χρονοσειρών, η ευκλείδεια απόσταση καθώς και διάφορες άλλες Lp νόρμες αποτελούν τα πιο ευρέως χρησιμοποιούμενα μέτρα απόστασης [69]. Ακόμα ένα μέτρο που συναντά κανείς στη βιβλιογραφία είναι η απόσταση Canberra [70]. Τα μέτρα αυτά ικανοποιούν τις ιδιότητες που πρέπει να ακολουθεί μια μετρική απόστασης αλλά η χρήση τους περιορίζεται μόνο σε χρονοσειρές ίδιου μήκους. Τα μέτρα που η χρήση τους περιορίζεται μόνο σε χρονοσειρές ίδιου μήκους αναφέρονται στη βιβλιογραφία ως "lock-step measures". Ένα μέτρο απόστασης που αντιμετωπίζει το πρόβλημα χρονοσειρών διαφορετικών διαστάσεων είναι το μέτρο ανομοιότητας DTW (Dynamic Time Warping) [45-48]. Το μέτρο DTW όμως δεν αποτελεί μετρική και η τετραγωνική χρονική πολυπλοκότητα του το καθιστά ακατάλληλο για εφαρμογές όπου το σύνολο δεδομένων είναι ιδιαίτερα μεγάλο. Τα μέτρα που δύνανται να δεχτούν ως ορίσματα χρονοσειρές διαφορετικού μήκους αναφέρονται στη βιβλιογραφία ως "elastic measures". Ένα επίσης δημοφιλές μέτρο ομοιότητας χρονοσειρών που όμως δεν αποτελεί μετρική είναι ο συντελεστής συσχέτισης Pearson [74]. Ωστόσο, σύμφωνα με έρευνες, η ευκλείδεια απόσταση σε εφαρμογές χρονοσειρών ίδιας διάστασης αποδίδει καλύτερα σε σύγκριση με άλλα μέτρα [31, 68].

Μια μέθοδος για την αντιμετώπιση των προβλημάτων που παρουσιάζονται στη μέτρηση του βαθμού ομοιότητας των χρονοσειρών στο πεδίο του χρόνου είναι ο υπολογισμός της ομοιότητας των χαρακτηριστικών (features) που μπορούν να εξαχθούν από τις χρονοσειρές. Κάθε χρονοσειρά αναπαρίσταται ως ένα διάνυσμα χαρακτηριστικών τα οποία εισάγονται ως ορίσματα στις συναρτήσεις (αν)ομοιότητας (feature based (dis)similarity).



Σχήμα 2.11 : Μέθοδοι υπολογισμού ανομοιότητας χρονοσειρών [10].

2.4 Μοντέλα και Αλγόριθμοι Μηχανικής Μάθησης

Στην ενότητα αυτή πραγματοποιείται η περιγραφή των αλγορίθμων και μοντέλων Μηχανικής Μάθησης που χρησιμοποιήθηκαν στη παρούσα διπλωματική εργασία. Για την υλοποίηση των εφαρμογών Μη Επιβλεπόμενης Μάθησης χρησιμοποιήθηκε ο αλγόριθμος K-Means που αποτελεί τον πιο διαδεδομένο αλγόριθμο ομαδοποίησης. Για την υλοποίηση των εφαρμογών Επιβλεπόμενης Μάθησης χρησιμοποιήθηκε ο αλγόριθμος K-Κοντινότερων Γειτόνων (K-Nearest Neighbors) ή αλλιώς K-NN, ο αλγόριθμος Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines) ή αλλιώς SVM, ο αλγόριθμος CART Δένδρων Απόφασης (Decision Trees) καθώς και ο αλγόριθμος Τυχαίου Δάσους (Random Forest).

2.4.1 Αλγόριθμος K-Means

Σύμφωνα με τη βιβλιογραφία ο αλγόριθμος K-Means είναι ο πιο ευρέως διαδεδομένος αλγόριθμος ομαδοποίησης λόγω της απλότητας της λειτουργίας του και της ταχύτητας του. Ανήκει στην κατηγορία των διαμεριστικών αλγορίθμων ομαδοποίησης (partition based clustering algorithms) και προτάθηκε το 1956 από τον Steinhaus, ενώ το όνομα που του αποδόθηκε χρησιμοποιήθηκε για πρώτη φορά από τον MacQueen το 1967 [84].

Η διαδικασία ομαδοποίησης ενός συνόλου δεδομένων με βάση τον K-Means απαιτεί τον καθορισμό του αριθμού των ομάδων που θα προκύψουν και το γεγονός αυτό αποτελεί ένα από τα βασικά μειονεκτήματα του εν λόγω αλγορίθμου. Επίσης, ένα ακόμα σημαντικό μειονέκτημα είναι το γεγονός πως διαφορετικές αρχικές συνθήκες παράγουν διαφορετικές ομαδοποιήσεις. Η αντιμετώπιση των παραπάνω προβλημάτων απαιτεί πολλαπλά τρεξίματα του αλγορίθμου με διαφορετικές αρχικές συνθήκες για κάθε δυνατό αριθμό ομάδων που ορίζεται από τον χρήστη καθώς και τη τελική σύγκριση των αποτελεσμάτων μέσω μετρικών επικύρωσης ομαδοποίησης.

Η κύρια ιδέα του αλγορίθμου είναι να προσδιοριστούν με βάση τις αρχικές συνθήκες *k* κεντροειδή (centroids), ένα για κάθε ομάδα. Στη συνέχεια, επιλέγεται κάθε παρατήρηση του συνόλου δεδομένων και συσχετίζεται με το κοντινότερο σε αυτή κεντροειδές. Η συσχέτιση γίνεται βάση κάποιου μέτρου (αν)ομοιότητας που επιλέγεται από τον χρήστη. Όταν αυτό γίνει για όλες τις παρατηρήσεις του συνόλου δεδομένων και κάθε παρατήρηση ανήκει πλέον σε μια ομάδα, ο αλγόριθμος έχει υλοποιήσει μια "πρόχειρη" πρώτη ομαδοποίηση. Στη

συνέχεια υπολογίζονται εκ νέου οι θέσεις των κεντροειδών των ομάδων μέσω του υπολογισμού του κέντρου βάρους των παρατηρήσεων που ανήκουν σε αυτές. Έπειτα, υλοποιείται ξανά η συσχέτιση όλων των παρατηρήσεων με τα νέα κεντροειδή που προέκυψαν και κάθε παρατήρηση τοποθετείται στην ομάδα με το κοντινότερο σε αυτήν κεντροειδές. Η διαδικασία αυτή επαναλαμβάνεται συνεχώς και ο αλγόριθμος τερματίζει όταν δεν σημειωθούν περαιτέρω αντιμεταθέσεις παρατηρήσεων. Το αποτέλεσμα που προκύπτει είναι η ομαδοποίηση των παρατηρήσεων του συνόλου δεδομένων σε k ομάδες.

Ο αλγόριθμος στοχεύει να ελαχιστοποιήσει τη παρακάτω αντικειμενική συνάρτηση η οποία ορίζεται ως εξής :

Δεδομένου ενός συνόλου παρατηρήσεων $X = (x_1, \dots, x_n)$, όπου κάθε παρατήρηση είναι ένα διάνυσμα διάστασης m και ενός συνόλου $S = (S_1, \dots, S_k)$ από k σύνολα–ομάδες όπου $C = (c_1, \dots, c_k)$ το σύνολο των κεντροειδών των ομάδων, η αντικειμενική συνάρτηση ορίζεται ως εξής :

$$J = \sum_{i=1}^{k} \sum_{x \in S_i} d(x, c_i) \qquad (2.12)$$

Ο τελεστής d εκφράζει ένα μέτρο απόστασης που χρησιμοποιείται για τη μέτρηση της απόστασης κάθε παρατήρησης x που ανήκει στην ομάδα i από το κεντροειδές της ομάδας i και συνήθως αναφέρεται στην ευκλείδεια απόσταση. Σημειώνεται ότι η επιλογή του μέτρου d επηρεάζει σημαντικά τα αποτελέσματα του αλγορίθμου [30, 35]. Επίσης, αξίζει να σημειωθεί ότι η αντικειμενική συνάρτηση J είναι ουσιαστικά το μέτρο WCSS που ορίσαμε μέσω της σχέσης (2.4) στην ενότητα 2.3.2.

Στη συνέχεια παρουσιάζουμε τον ψευδοκώδικα του αλγορίθμου K-Means [34].

Αλγόριθμος K-Means (Ψευδοκώδικας)
1: **Θέσε** την τιμή του $\boldsymbol{k}$
2: **Αρχικοποίησε** $\boldsymbol{k}$ κεντροειδή $\boldsymbol{\mu_1, \mu_2, \dots, \mu_k}$
3: **Επανάλαβε Όσο** το κριτήριο τερματισμού δεν έχει επιτευχθεί :
4: **Για** $\boldsymbol{i = 1}$ **μέχρι** $\boldsymbol{n}$ **Επανάλαβε :**
5: $label_i \leftarrow arg\min_l d(x_i, \mu_l)$
6: **Τέλος Επανάληψης**
7: **Για** $\boldsymbol{j = 1}$ **μέχρι** $\boldsymbol{k}$ **Επανάλαβε :**
8: $\boldsymbol{\mu_j} \leftarrow \frac{\sum_{i=1}^{n}\{label_i = j\}x_i}{\sum_{i=1}^{n}\{label_i = j\}}$
9: **Τέλος Επανάληψης**
10: **Έλεγχος** κριτηρίου τερματισμού
11: **Επίστρεψε** $label_1, \dots, label_n$ και $\mu_1, \dots, \mu_n$

Η χρονική πολυπλοκότητα του αλγορίθμου είναι γραμμική, δηλαδή $O(n)$.

Ο αλγόριθμος K-Means αποδεικνύεται ότι τερματίζει πάντα [84], όμως τα αποτελέσματα δεν είναι πάντα βέλτιστα καθώς επηρεάζεται σημαντικά από την αρχικοποίηση των κεντροειδών. Αυτό σημαίνει ότι ο αλγόριθμος θα συγκλίνει σε τοπικό ελάχιστο εάν δεν επιτευχθεί κάποια αποδοτική αρχικοποίηση των κεντροειδών. Για την αντιμετώπιση αυτού

του προβλήματος έχουν προταθεί παραλλαγές του αλγορίθμου K-Means όπου συνδιάζεται με ευρετικές συναρτήσεις (heuristics) για την αποδοτική αρχικοποίηση των κεντροειδών. Μια τέτοια παραλλαγή είναι ο αλγόριθμος K-Means++ που προτάθηκε από τους Arthur και Vassilvitskii [32], τον οποίο χρησιμοποιήσαμε για την υλοποίηση των εφαρμογών στα πλαίσια της διπλωματικής.

2.4.2 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

Οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ) ή αλλιώς Support Vector Machines (SVM) προτάθηκαν από τους Vapnik και Chervonenkis το 1992 και αποτελούν πανίσχυρα μοντέλα Επιβλεπόμενης Μάθησης. Χρησιμοποιούνται σε εφαρμογές γραμμικής και μη γραμμικής Ταξινόμησης καθώς και σε εφαρμογές Παλινδρόμησης. Η βασική ιδέα της λειτουργίας τους έχει ως εξής [85] :

Έστω ότι έχουμε ένα σύνολο δεδομένων X με δεδομένα $x \epsilon \mathcal{R}^d$ τα οποία χαρακτηρίζονται από δύο διαφορετικές κλάσεις και έστω $y \epsilon \{-1,1\}$ οι ετικέτες των κλασεων. Οι ΜΔΥ μεταφέρουν τα δεδομένα στο "χώρο χαρακτηριστικών" (feature space) στον οποίο αυξάνονται οι διαστάσεις τους με αποτέλεσμα να είναι γραμμικά διαχωριζόμενα μέσω του υπολογισμού ενός βέλτιστου "υπερεπίπεδου" (optimal hyperplane) διαχωρισμού.

Η μεταφορά των δεδομένων στο χώρο χαρακτηριστικών επιτυγχάνεται μέσω συναρτήσεων πυρήνων (kernels) οι οποίες δίνουν το εσωτερικό γινόμενο στο χώρο χαρακτηριστικών εκτελώντας υπολογισμούς στο χώρο των δεδομένων.

Ένα "υπερεπίπεδο" δίνεται από μια συνάρτηση της μορφής :

$$f(x) = w^T x + b \tag{2.13}$$

Όπου $w, b \ \epsilon \mathcal{R}^d$ οι παράμετροι του μοντέλου.

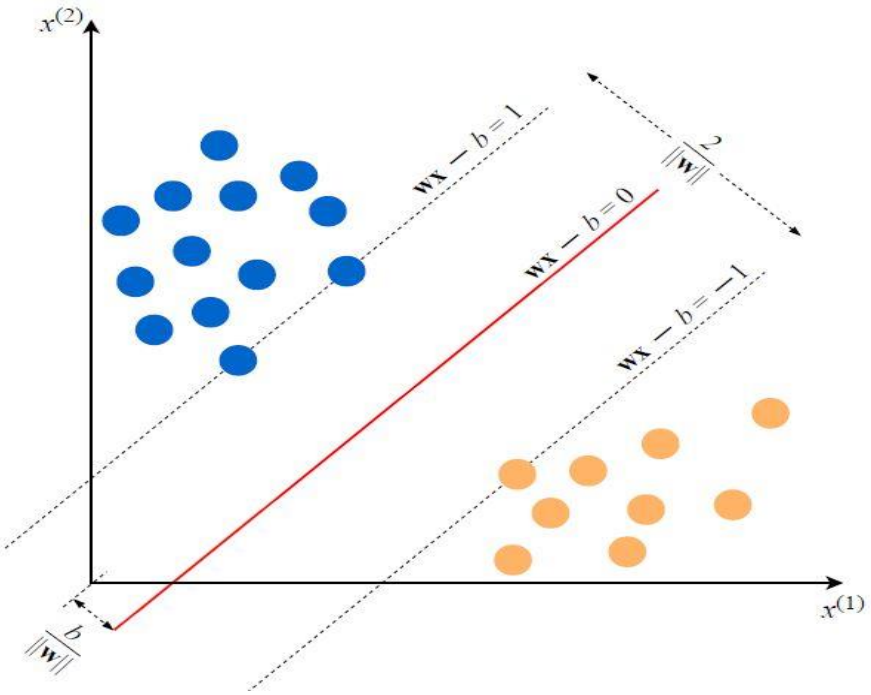
Υπάρχουν άπειρα υπερεπίπεδα που έχουν την δυνατότητα να ταξινομήσουν τα δεδομένα εκπαίδευσης x με επιτυχία. Το βέλτιστο υπερεπίπεδο είναι αυτό που μεγιστοποιεί το περιθώριο (margin) ανάμεσα στις δύο κλάσεις. Ένα μεγάλο περιθώριο θα συνεισφέρει σε μια καλύτερη γενίκευση του μοντέλου. Ο όρος γενίκευση αναφέρεται στην ικανότητα ενός μοντέλου να κατηγοριοποιεί σωστά στο μέλλον νέες παρατηρήσεις οι οποίες δεν ανήκουν στο σύνολο εκπαίδευσης.

Η απόσταση μιας παρατήρησης x από το υπερεπίπεδο δίνεται από τη σχέση :

$$z(x) = \frac{|f(x)|}{||w||} \tag{2.14}$$

Όπου $||w|| = \sqrt{\sum_{j=1}^{d} (w^{(j)})^2}$ η ευκλείδεια νόρμα του w.

Συνεπώς, όπως είναι φανερό από τη σχέση (2.14) μέσω της ελαχιστοποίησης της ευκλείδειας νόρμας του w μεγιστοποιούμε το περιθώριο ανάμεσα στις δύο κλάσεις και υπολογίζουμε το βέλτιστο υπερεπίπεδο.



Σχήμα 2.12 : Παράδειγμα μοντέλου Ταξινόμησης SVM δισδιάστατων δεδομένων [49].

Με βάση το παράδειγμα του σχήματος 2.11 όπου οι δύο κλάσεις είναι γραμμικά διαχωρίσιμες, η επιλογή κατάλληλων παραμέτρων w, b δίνει ως αποτέλεσμα [49]:

$$f(x) \geq 1 \quad \forall \ x \quad με \quad y = +1$$

$$f(x) \leq -1 \quad \forall \ x \quad με \quad y = -1$$

Με μέγιστο μήκος περιθωρίου $\frac{2}{||w||}$

Σύμφωνα με τα παραπάνω, το πρόβλημα ταξινόμησης ανάγεται στον υπολογισμό της ελάχιστης ευκλείδειας νόρμας του w υπό τον περιορισμό :

$$y_i(w^T x_i + b) \geq 1 \ \forall \ i = 1, \dots, n$$

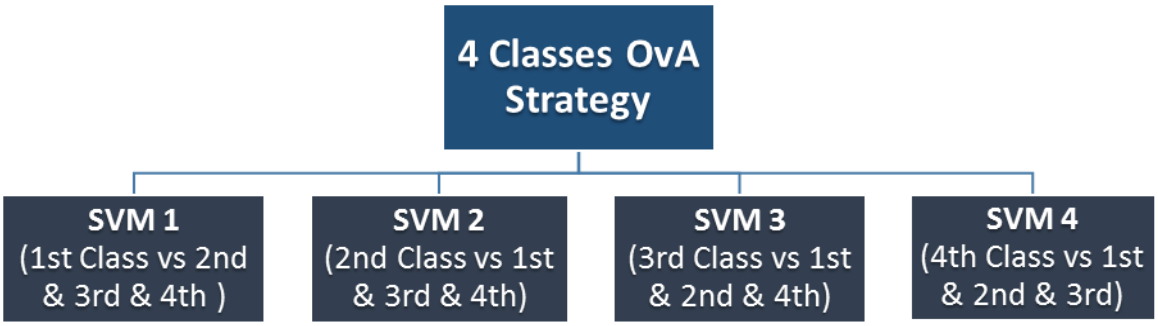
και λύνεται με τη μέθοδο Πολλαπλασιαστών Lagrange [52].

Στο παράδειγμα του σχήματος 2.11 το μοντέλο που αναπτύχθηκε είναι γραμμικό καθώς το σύνορο διαχωρισμού (decision boundary) το οποίο παρουσιάζεται με κόκκινο χρώμα είναι μία ευθεία γραμμή. Τα μοντέλα SVM μπορούν να ενσωματώσουν συναρτήσεις πυρήνων (kernels) που δίνουν τη δυνατότητα δημιουργίας μη γραμμικών συνόρων διαχωρισμού. Η μορφή των συνόρων διαχωρισμού καθορίζει την ακρίβεια (accuracy) των μοντέλων και αποτελεί τον βασικό παράγοντα διαφοροποίησης τους.

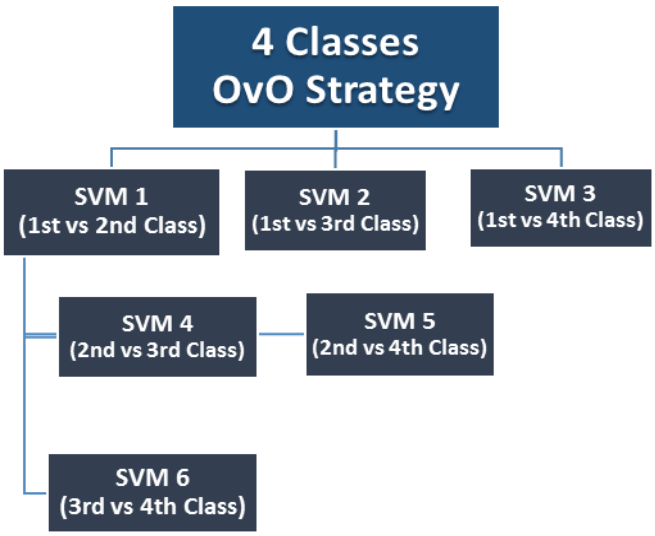
Στο σημείο αυτό αξίζει να σημειωθεί ότι οι αλγόριθμοι SVM αποτελούν μοντέλα δυαδικής ταξινόμησης αυστηρά. Η αξιοποίηση τους σε εφαρμογές ταξινόμησης πολλών κλάσεων επιτυγχάνεται μέσω δύο στρατηγικών, τη στρατηγική OvA (One vs All) ή αλλιώς OvR (One vs Rest) και τη στρατηγική OvO (One vs One).

Στη στρατηγική OvA εκπαιδεύονται τόσα μοντέλα δυαδικής ταξινόμησης όσες είναι και οι διαφορετικές κλάσεις των δεδομένων εισόδου και κάθε μοντέλο μαθαίνει να αναγνωρίζει μία κλάση έναντι των υπολοίπων. Συνεπώς, με βάση τη στρατηγική OvA κάθε μοντέλο SVM εκπαιδεύεται σε όλο το σύνολο των δεδομένων εισόδου εκπαίδευσης και υπολογίζει ένα σύνορο διαχωρισμού μεταξύ μίας συγκεκριμένης κλάσης και όλων των υπολοίπων κλάσεων.

Αντίθετα, στη στρατηγική OvO εκπαιδεύονται τόσα μοντέλα δυαδικής ταξινόμησης όσα είναι και τα δυνατά ζεύγη των κλάσεων και κάθε μοντέλο μαθαίνει να αναγνωρίζει μία συγκεκριμένη κλάση έναντι μίας άλλης συγκεκριμένης κλάσης. Συνεπώς, με βάση τη στρατηγική OvO κάθε μοντέλο SVM εκπαιδεύεται μόνο στα δεδομένα που ανήκουν σε δύο διαφορετικές κλάσεις από το σύνολο των δεδομένων εκπαίδευσης και υπολογίζει ένα σύνορο διαχωρισμού μεταξύ των δύο συγκεκριμένων κλάσεων. Για παράδειγμα, όταν το σύνολο δεδομένων εκπαίδευσης αποτελείται από δεδομένα δέκα διαφορετικών κλάσεων τότε εκπαιδεύονται $\binom{10}{2} = \frac{10!}{2!(10-2)!} = 45$ διαφορετικά μοντέλα SVM.



Σχήμα 2.13 : Στρατηγική OvA κατηγοριοποίησης τεσσάρων κλάσεων με μοντέλα SVM.



Σχήμα 2.14 : Στρατηγική OvO κατηγοριοποίησης τεσσάρων κλάσεων με μοντέλα SVM.

2.4.3 Κ - Κοντινότεροι Γείτονες (K-Nearest Neighbors)

Ο αλγόριθμος των Κ – Κοντινότερων Γειτόνων (K-NN) είναι από τους πιο διαδεδομένους αλγορίθμους Μηχανικής Μάθησης για εφαρμογές μη γραμμικής Ταξινόμησης. Προτάθηκε από τους Fix και Hodges το 1951 [86] και αποτελεί μη παραμετρική μέθοδο καθώς δεν απαιτείται η μάθηση κάποιου συνόλου παραμέτρων των παρατηρήσεων για την ταξινόμηση τους. Συνεπώς, ο εν λόγω αλγόριθμος δεν μοντελοποιεί τα δεδομένα και δεν κάνει υποθέσεις σχετικά με την κατανομή τους. Αντιθέτως, αποθηκεύει όλο το σύνολο εκπαίδευσης στη μνήμη [52]. Για αυτόν τον λόγο η μέθοδος Ταξινόμησης των Κ – Κοντινότερων γειτόνων αναφέρεται στη βιβλιογραφία και ως Ταξινόμηση Βασιζόμενη σε Μνήμη (Memory Based Classification). Ο μηχανισμός λειτουργίας του είναι ιδιαίτερα απλός καθώς η υπό ταξινόμηση παρατήρηση κατατάσσεται στη κλάση στην οποία ανήκουν οι k κοντινότερες σε αυτή παρατηρήσεις εκπαίδευσης σύμφωνα με κάποια μετρική απόστασης. Τυχόν ισοπαλίες αντιμετωπίζονται με τυχαίο τρόπο ενώ σε κάθε άλλη περίπτωση χρησιμοποιείται η ψήφος πλειοψηφίας [52].

Η παράμετρος k και η μετρική απόστασης καθορίζονται από τον χρήστη πριν το τρέξιμο του αλγορίθμου. Συνήθως το k επιλέγεται ως περιττός αριθμός ώστε να αποφεύγονται τυχόν ισοπαλίες. Οι μετρικές απόστασης που χρησιμοποιούνται είναι κυρίως η ευκλείδεια απόσταση (L_2 νόρμα) καθώς και διάφορες άλλες L_P νόρμες. Επίσης, δημοφιλείς μετρικές απόστασης που χρησιμοποιούνται συχνά είναι η ομοιότητα συνιμιτόνου (cosine similarity), η απόσταση Chebyshev, η απόσταση Hamming και η απόσταση Mahalanobis [49].

Στη συνέχεια παρουσιάζουμε τον ψευδοκώδικα του αλγορίθμου K-NN.

Αλγόριθμος Κ – Nearest Neighbors (Ψευδοκώδικας)
1: **Φόρτωσε** τα δεδομένα εκπαίδευσης και ελέγχου στη μνήμη
2: **Θέσε** την τιμή του k
3: **Καθόρισε** τη μετρική απόστασης
4: **Για κάθε** δεδομένο στο σύνολο ελέγχου :
5: **Υπολόγισε** τις αποστάσεις από όλα τα δεδομένα εκπαίδευσης
6: **Αποθήκευσε** τις αποστάσεις σε μια λίστα
7: **Εκτέλεσε Αύξουσα Ταξινόμηση** στη λίστα
8: **Επίλεξε** τα k δεδομένα εκπαίδευσης που αντιστοιχούν στα πρώτα k στοιχεία της ταξινομημένης λίστας
9: **Ανάθεσε** το δεδομένο ελέγχου στη κλάση που ανήκει η πλειοψηφία των k επιλεγμένων δεδομένων εκπαίδευσης
10: **Τέλος Επανάληψης**
11: **Επίστρεψε** τις ετικέτες των κλάσεων των δεδομένων ελέγχου

Ο αλγόριθμος των Κ – Κοντινότερων Γειτόνων έχει πολυπλοκότητα $O(k * n * d)$, όπου n το πλήθος των δεδομένων εκπαίδευσης και d η διάσταση των δεδομένων.

2.4.4 Δένδρα Απόφασης (Decision Trees)

Τα Δένδρα Απόφασης είναι άκυκλοι γράφοι και αποτελούν μοντέλα Επιβλεπόμενης Μάθησης. Χρησιμοποιούνται σε εφαρμογές Ταξινόμησης όπου αναφέρονται ως Δένδρα Ταξινόμησης καθώς και σε εφαρμογές Πρόβλεψης όπου αναφέρονται ως Δένδρα Παλινδρόμησης.

Τα βασικά δομικά στοιχεία ενός Δένδρου Απόφασης είναι :

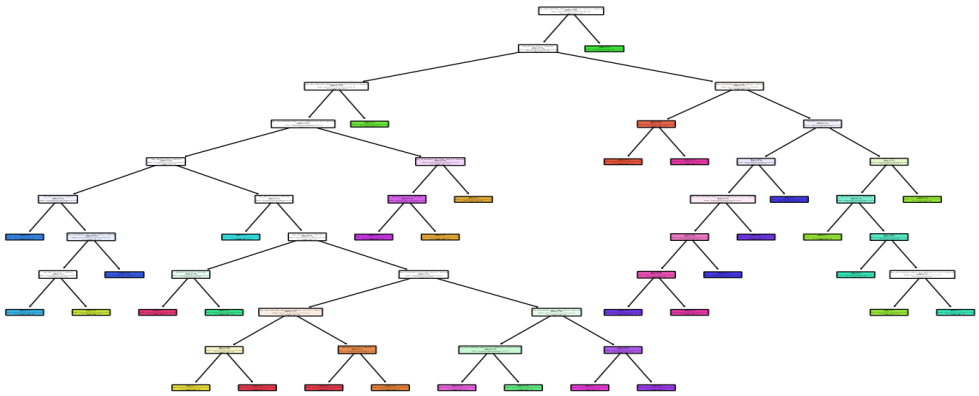
- **Η ρίζα (root node) του δένδρου :**
  Αντιπροσωπεύει το σύνολο των δεδομένων εισόδου υπό κατηγοριοποίηση και βρίσκεται στη κορυφή του δένδρου.

- **Οι κλάδοι (branches) :**
  Αποτελούν συνδέσμους μεταξύ των κόμβων και έχουν κατεύθυνση αυστηρά από τα υψηλότερα επίπεδα προς τα χαμηλότερα επίπεδα του δένδρου.

- **Οι κόμβοι απόφασης (decision nodes) :**
  Αποτέλουν ενδιάμεσους κόμβους του δένδρου και αντιπροσωπεύουν υποσύνολα των δεδομένων εισόδου. Στους κόμβους απόφασης εισέρχονται και εξέρχονται κλάδοι.

- **Τα φύλλα (leafs) :**
  Αποτελούν τερματικούς κόμβους του δένδρου και αντιπροσωπεύουν συγκεκριμένες κλάσεις.

Στη ρίζα και σε κάθε κόμβο του δένδρου εκτός των φύλλων αντιστοιχεί ένα συγκεκριμένο χαρακτηριστικό – γνώρισμα των δεδομένων εισόδου. Το **κριτήριο διάσπασης (splitting criterion)** αξιολογεί και ταξινομεί ιεραρχικά τα χαρακτηριστικά ως προς την ικανότητα τους να διαχωρίζουν τις παρατηρήσεις με βάση τις ετικέτες τους. Κάθε παρατήρηση του συνόλου εισόδου ξεκινάει από την ρίζα και καταλήγει σε κάποιο φύλλο μέσω διαδοχικών συγκρίσεων των χαρακτηριστικών του. Στη ρίζα του δένδρου επιλέγεται εκείνο το γνώρισμα – χαρακτηριστικό που δύναται να διαχωρίσει βέλτιστα τις παρατηρήσεις. Στη συνέχεια, με βάση την ιεραρχία των χαρακτηριστικών προσδιορίζονται διαδοχικά οι εσωτερικοί κόμβοι εώς ότου κατηγοριοποιηθούν σωστά όλες οι παρατηρήσεις με αποτέλεσμα τη δημιουργία των φύλλων.



Σχήμα 2.15 : Παράδειγμα μοντέλου Δένδρου Απόφασης.

Δεδομένου ενός συνόλου παρατηρήσεων εισόδου και των χαρακτηριστικών τους, δύναται να δημιουργηθούν πολλά διαφορετικά Δένδρα Απόφασης. Η διαφοροποίηση τους έγκειται στην αντιστοίχηση συγκεκριμένων χαρακτηριστικών σε συγκεκριμένους κόμβους και καθορίζεται από το κριτήριο διάσπασης καθώς και από τον αλγόριθμο κατασκευής του δένδρου. Στη συνέχεια αναφέρουμε βασικά κριτήρια διάσπασης καθώς και αλγορίθμους που χρησιμοποιούνται για την ανάπτυξη των μοντέλων Δένδρων Απόφασης [24, 53].

Βασικά Κριτήρια Διάσπασης (Splitting Criteria) :

- Entropy
- Information Gain
- Reduction in Variance
- Chi – Square *(ποιοτικές μεταβλητές αυστηρά)*
- Gain Ratio
- Gini Impurity Index

Σε έναν κόμβο t το μέτρο **Gini Impurity Index** ορίζεται ως εξής :

$$i(t) = \sum_{i,j} C(i|j)\, p(i|t)\, p(j|t) \tag{2.15}$$

Όπου :

- $C(i|j)$: το κόστος εσφαλμένης κατηγοριοποίησης μιας παρατήρησης στη κλάση i ενώ ανήκει στην κλάση j. Δηλαδή $C(i|j) = 1$ εάν $i \neq j$ και $C(i|j) = 0$ εάν $i = j$

- $p(i|t)$: η πιθανότητα μιας παρατήρησης που ανήκει στην κλάση i να καταλήξει στον κόμβο t

- $p(j|t)$: η πιθανότητα μιας παρατήρησης που ανήκει στην κλάση j να καταλήξει στον κόμβο t

Το **κριτήριο διάσπασης Gini** στοχεύει στην ελαχιστοποίηση της παρακάτω συνάρτησης :

$$\Delta_i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{2.16}$$

Όπου :

- $\Delta_i(s,t)$: η μεταβολή του Gini Impurity Index στον κόμβο t μέσω της διάσπασης s

- p_L : η πιθανότητα μιας παρατήρησης να καταλήξει στον κόμβο t_L, δηλαδή στο αριστερό παιδί του κόμβου t

- p_R : η πιθανότητα μιας παρατήρησης να καταλήξει στον κόμβο t_R, δηλαδή στο δεξί παιδί του κόμβου t

Οι πιο δημοφιλείς αλγόριθμοι κατασκευής Δένδρων Απόφασης είναι :

- IDE3
- C4.5 & C5
- CHAID
- CART

Στο πλαίσιο ανάπτυξης εφαρμογών κατηγοριοποίησης χρονοσειρών ενεργειακών δεδομένων χρησιμοποιήσαμε τον αλγόριθμο CART (Classification And Regression Trees) ο οποίος εκτελεί πάντα δυαδικούς διαχωρισμούς, δηλαδή παράγει δυαδικά δένδρα τα οποία είναι συνήθως ισοζυγισμένα. Επίσης, ο εν λόγω αλγόριθμος χρησιμοποιεί το Gini Impurity Index ως κριτήριο διάσπασης.

Στη συνέχεια παρουσιάζουμε τον ψευδοκώδικα του αλγορίθμου CART [5].

Αλγόριθμος CART (Ψευδοκώδικας)
1: **Για κάθε χαρακτηριστικό που δεν έχει αντιστοιχηθεί σε κόμβο βρες το καλύτερο κατώφλι διαχωρισμού που μεγιστοποιεί το κριτήριο διάσπασης**. (Υπάρχουν K-1 πιθανά κατώφλια διαχωρισμού για κάθε χαρακτηριστικό με K διαφορετικές τιμές)
2: **Βρες τη καλύτερη διάσπαση του κόμβου** : Ανάμεσα στους καλύτερους διαχωρισμούς που επιλέχθηκαν στο βήμα 1 βρες τον διαχωρισμό που μεγιστοποιεί το κριτήριο διάσπασης
3: **Σχημάτισε το αριστερό και δεξί παιδί του κόμβου** με βάση τη καλύτερη διάσπαση κόμβου που επιλέχθηκε στο βήμα 2
4: **Επανάλαβε** τα βήματα 1 εως 3 **εώς ότου** επιτευχθεί το κριτήριο τερματισμού

Στο σημείο αυτό αξίζει να αναφέρουμε πως όταν επιτευχθεί το κριτήριο τερματισμού του αλγορίθμου, δηλαδή όταν έχουν κατηγοριοποιηθεί όλα τα δεδομένα εκπαίδευσης επιτυχώς, το Δένδρο Απόφασης που έχει σχηματιστεί είναι συνήθως ιδιαίτερα μεγάλο καθώς έχει υπερμοντελοποιηθεί πάνω στα δεδομένα εκπαίδευσης. Για τη βελτίωση της γενίκευσης του μοντέλου σε μελλοντικά δεδομένα ελέγχου αξιοποιούνται τεχνικές "κλαδέματος" (pruning techniques). Το κλάδεμα ενός δένδρου περιλαμβάνει την αφαίρεση περιττών συγκρίσεων μέσω της διαγραφής ορισμένων κόμβων και κλάδων οι οποίοι υλοποιούν διασπάσεις μικρής στατιστικής σημασίας. Η πιο δημοφιλής τεχνική κλαδέματος είναι η τεχνική Cost-Complexity Pruning όπου το κλάδεμα παραμετροποιείται μέσω της παραμέτρου ccp_alpha [66].

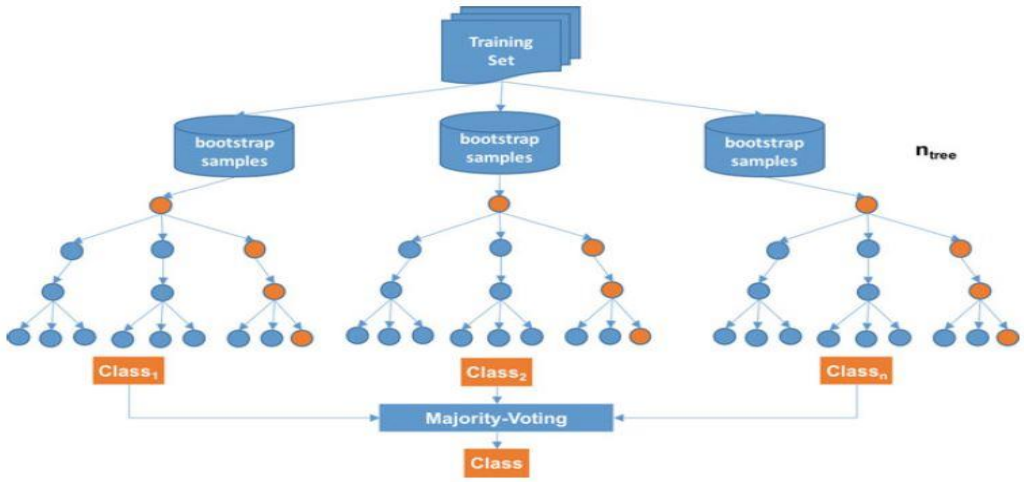
Τα κύρια πλεονεκτήματα των μοντέλων Δένδρων Απόφασης σε εφαρμογές Επιβλεπόμενης Μάθησης είναι [24] :

- Είναι ιδιαίτερα αποτελεσματικά και εύκολα στη χρήση
- Δεν επηρεάζονται σημαντικά από τον θόρυβο των δεδομένων
- Παρουσιάζουν ανοχή σε ελλιπείς τιμές (missing values)
- Τα μοντέλα μεταφράζονται σε ένα σύνολο κανόνων if-then, γεγονός που καθιστά βατή την ερμηνεία τους
- Δεν απαιτείται προεπεξεργασία και μετασχηματισμός (scaling) των δεδομένων

2.4.5 Τυχαίο Δάσος (Random Forest)

Ο αλγόριθμος αυτός αναπτύχθηκε από τους Breiman και Cutler το 2001 [5] και αποτελεί μια τροποποίηση του αλγορίθμου Bootstrap aggregating ή αλλιώς Bagging. Η βασική ιδέα του αλγορίθμου Bagging έγκειται στην δημιουργία πολλών αντιγράφων ενός συνόλου δεδομένων μέσω της ομοιόμορφης δειγματοληψίας τους με επανατοποθέτηση όπου κάθε αντίγραφο χρησιμοποιείται για την εκπαίδευση ενός διαφορετικού μοντέλου. Τα διαφορετικά αυτά μοντέλα είναι του ίδιου τύπου και συνδιάζονται σε ένα μοντέλο συνάθροισης (ensemble model) το οποίο παράγει ως έξοδο την μέση τιμή των αποτελεσμάτων (παλινδρόμηση) ή την ψήφο πλειοψηφίας (κατηγοριοποίηση) των μοντέλων που το απαρτίζουν.

Συνεπώς, ένα μοντέλο ταξινόμησης Τυχαίου Δάσους αποτελεί μια συνάθροιση πολλαπλών Δένδρων Απόφασης όπου κάθε δένδρο μπορεί να θεωρηθεί ως ένας μεμονωμένος ταξινομητής και η έξοδος του μοντέλου συνάθροισης ψηφίζεται από όλα τα δένδρα.



Σχήμα 2.16 : Δομή μοντέλου Ταξινόμησης Τυχαίου Δάσους [67].

Έστω ότι έχουμε ένα σύνολο δεδομένων X με δεδομένα $x_i \epsilon \mathcal{R}^d$ όπου $i = (1, \dots, n)$. Συνεπώς, κάθε παρατήρηση x_i αποτελεί ένα διάνυσμα d χαρακτηριστικών μεταβλητών. Η διαδικασία κατασκευής ενός ταξινομητή Τυχαίου Δάσους που αποτελείται από Τ Δένδρα Απόφασης έχει ως εξής [67] :

1. Παράγουμε Τ νέα σύνολα εκπαίδευσης μεγέθους n μέσω ομοιόμορφης δειγματοληψίας με αντικατάσταση *(bootstrapping)* από το σύνολο X. Σημειώνεται ότι ενδέχεται να μην χρησιμοποιηθούν όλα τα δεδομένα του X λόγω του μηχανισμού αντικατάστασης.

2. Κάθε Δένδρο Απόφασης εκπαιδεύεται σε ένα συγκεκριμένο νέο σύνολο εκπαίδευσης και επιλέγονται τυχαία από κάθε παρατήρηση του αντίστοιχου συνόλου m χαρακτηριστικά, όπου $m \ll d$. Ο καλύτερος διαχωρισμός σε αυτά υπολογίζεται για τον διαχωρισμό του κόμβου με βάση το επιλεγμένο κριτήριο διάσπασης.

3. Κάθε δένδρο συνεχίζει να αναπτύσσεται πλήρως μέχρις ότου να κατηγοριοποιηθούν ορθά όλες οι παρατηρήσεις του συνόλου εκπαίδευσης.

Σύμφωνα με τους δημιουργούς του αλγορίθμου το ποσοστό σφάλματος του Τυχαίου Δάσους εξαρτάται από δύο παράγοντες [5] :

1. Τη **συσχέτιση (correlation)** των δένδρων του δάσους. Το ποσοστό σφάλματος του δάσους είναι ανάλογο της συσχέτισης.

2. Τη **δύναμη (strength)** κάθε δένδρου του δάσους. Το ποσοστό σφάλματος είναι αντιστρόφως ανάλογο της δύναμης.

Για την μείωση της συσχέτισης μεταξύ των δένδρων απαιτείται η μείωση του αριθμού των επιλεγμένων χαρακτηριστικών, δηλαδή η μείωση του m. Ωστόσο, η μείωση του m εκτός από τη συσχέτιση μειώνει και τη δύναμη κάθε δένδρου. Συνεπώς, πρέπει να βρεθεί κατάλληλη ισορροπία μεταξύ συσχέτισης και δύναμης ώστε να επιτευχθεί η βέλτιστη επίδοση του μοντέλου.

Στη συνέχεια παρουσιάζουμε τον αλγόριθμο του Τυχαίου Δάσους σε ψευδοκώδικα [52].

Αλγόριθμος Random Forest (Ψευδοκώδικας)
1: **Φόρτωσε** το σύνολο δεδομένων εκπαίδευσης $\boldsymbol{X}$
2: **Θέσε** τον αριθμό $\boldsymbol{T}$ που αντιστοιχεί στο πλήθος των Δένδρων Απόφασης
3: **Κατασκεύασε** $\boldsymbol{T}$ νέα σύνολα εκπαίδευσης από το $\boldsymbol{X}$ με τη μέθοδο Bootsrap
4: **Για κάθε** $t = 1$ εώς T :
5: **Καταχώρισε** το σύνολο εκπαίδευσης $\boldsymbol{X\_new_i}$ στο Δένδρο Απόφασης $\boldsymbol{T_i}$
6: **Για κάθε** κόμβο που πρόκειται να προστεθεί στο Δένδρο Απόφασης $\boldsymbol{T_i}$:
7: **Επίλεξε** m χαρακτηριστικά από το σύνολο των d χαρακτηριστικών του διανύσματος
8: **Υπολόγισε** τον δείκτη Gini Impurity για τα m αυτά χαρακτηριστικά
9: **Διαχώρισε** τα δεδομένα στο κόμβο με βάση τη μείωση του Gini Impurity
10: **Τέλος Επανάληψης**
11: **Τέλος Επανάληψης**

Για την κατηγοριοποίηση μιας νέας παρατήρησης x :

Έστω $\hat{C}_i(x)$ η πρόβλεψη του i-οστού Δένδρου Απόφασης του Δάσους για τη κλάση της παρατήρησης x.

Τότε η πρόβλεψη του μοντέλου Τυχαίου Δάσους για την κλάση της παρατήρησης x είναι :

$$\hat{C}_{RF}(x) = majority\ vote\ \{\hat{C}_i(x)\}_1^T$$

Σημειώνεται πως η διαφορά του αλγορίθμου Random Forest από τον αλγόριθμο Bagging έγκειται στο ότι κατά την κατασκευή κάθε δένδρου T_i από το νέο σύνολο εκπαίδευσης X_new_i, όταν πρόκειται να προστεθεί ένας νέος κόμβος, δεν χρησιμοποιούνται όλα τα διαθέσιμα χαρακτηριστικά των δεδομένων (d συνολικά), αλλά μόνο m από αυτά. Κατά την εκτέλεση του αλγορίθμου ο αριθμός m των χαρακτηριστικών είναι σταθερός.

Στο σημείο αυτό αξίζει να αναφέρουμε τη δυνατότητα αυτοαξιολόγησης που παρουσιάζει ο αλγόριθμος Random Forest μέσω της μεθόδου Out of Bag Estimate ή αλλιώς Out of Bag Error. Κατά τη δημιουργία των T νέων συνόλων εκπαίδευσης μέσω της μεθόδου Bootstraping, τα δεδομένα του αρχικού συνόλου εκπαίδευσης X που δεν επιλέχθηκαν από κανένα νέο σύνολο εκπαίδευσης καλούνται Out of Bag Samples και αποτελούν συνήθως το 33% των συνολικών δεδομένων του συνόλου X [5]. Για την αξιολόγηση της ακρίβειας του μοντέλου, εφόσον έχει ολοκληρωθεί η φάση εκπαίδευσης, ταξινομούνται τα Out of Bag δεδομένα και η ακρίβεια (accuracy) της ταξινόμησης τους δίνει το σφάλμα γενίκευσης του μοντέλου.

Συνοψίζοντας, τα πλεονεκτήματα του αλγορίθμου Random Forest είναι :

- Μπορεί να χειριστεί αποδοτικά δεδομένα μεγάλων διαστάσεων
- Δίνει μια εκτίμηση για το ποια χαρακτηριστικά είναι τα πιο σημαντικά στη κατηγοριοποίηση
- Δεν χρειάζεται διαφορετικό σύνολο δεδομένων για την επικύρωση του καθώς η εκτίμηση του λάθους γενίκευσης γίνεται από τον ίδιο τον αλγόριθμο με βάση το OOB score
- Μπορεί να χειριστεί αποδοτικά δεδομένα με ελλιπείς τιμές
- Δεν παρουσιάζει φαινόμενα υπερμοντελοποίησης όταν ο αριθμός των δένδρων είναι μεγάλος
- Δεν απαιτείται προεπεξεργασία και μετασχηματισμός (scaling) των δεδομένων
- Παρουσιάζει ευρωστία στις ακραίες παρατηρήσεις.

Τα μειονεκτήματα του αλγορίθμου Random Forest είναι :

- Ο αλγόριθμος είναι ιδιαίτερα αργός όταν ο αριθμός των δένδρων είναι πολύ μεγάλος
- Τα μοντέλα που δημιουργούνται δεν δύνανται να ερμηνευτούν εύκολα

2.4.6 Ταξινομητές Ψηφοφορίας (Voting Classifiers)

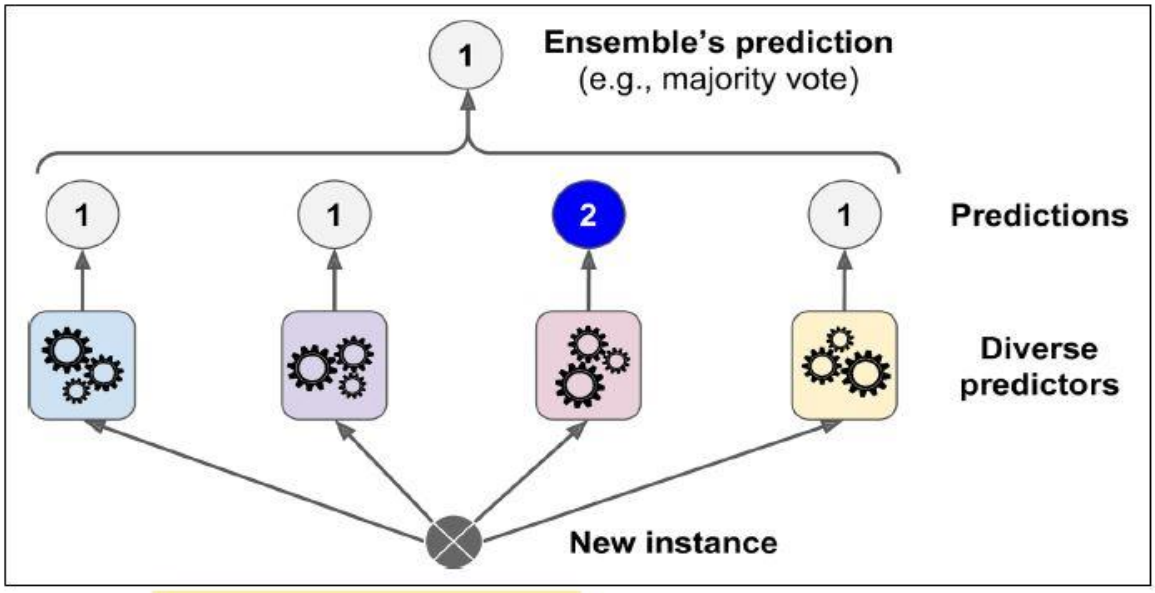
Ένα μοντέλο συνάθροισης (ensemble model) δύναται να αποτελείται από πολλά διαφορετικά μοντέλα ίδιου τύπου των οποίων η έξοδος είναι είτε πιθανότητες, όπως στη περίπτωση ενός Δένδρου Απόφασης, είτε κλάσεις, όπως στη περίπτωση μιας Μηχανής Διανυσμάτων Υποστήριξης. Οι Ταξινομητές Ψηφοφορίας αποτελούν μοντέλα συνάθροισης που δύνανται να συνδιάζουν διαφορετικά μοντέλα διαφορετικού τύπου και παράγουν ως έξοδο την ψήφο πλειοψηφίας τους. Σε γενικές γραμμές ένας Ταξινομητής Ψηφοφορίας επιτυγχάνει μεγαλύτερη ακρίβεια από τον καλύτερο ταξινομητή που έχει ενσωματώσει στο σύνολο του [50]. Ακόμα και αν οι ταξινομητές είναι λίγο καλύτεροι από έναν τυχαίο ταξινομητή, δηλαδή παρουσιάζουν ακρίβεια λίγο μεγαλύτερη του 50%, τότε ένας Ταξινομητής Ψηφοφορίας που τους συναθροίζει δύναται να παρουσιάσει αρκετά μεγάλη ακρίβεια υπό τον όρο ότι υπάρχει επαρκής αριθμός διαφορετικών και ασυσχέτιστων (uncorrelated) ταξινομητών στο σύνολο συνάθροισης. Το φαινόμενο αυτό παρουσιάζεται εξαιτίας του νόμου των μεγάλων αριθμών (law of large numbers).

Οι Ταξινομητές Ψηφοφορίας διακρίνονται σε δύο κατηγορίες :

1. **Hard Voting :** Όπου οι ταξινομητές του συνόλου συνάθροισης παράγουν ως έξοδο κλάσεις και ο ταξινομητής ψηφοφορίας παράγει ως έξοδο την ψήφο πλειοψηφίας τους.

2. **Soft Voting :** Όπου οι ταξινομητές του συνόλου συνάθροισης παράγουν ως έξοδο πιθανότητες, δηλαδή την εκτίμηση της πιθανότητας των παρατηρήσεων να ανήκουν σε μια κλάση. Ο ταξινομητής ψηφοφορίας σε αυτή τη περίπτωση παράγει ως έξοδο την κλάση με τη μεγαλύτερη πιθανότητα η οποία εκτιμάται ως ο μέσος όρος των εκτιμήσεων των ταξινομητών που συναθροίζονται.

Οι Ταξινομητές Ψηφοφορίας της δεύτερης κατηγορίας συνήθως παρουσιάζουν καλύτερη επίδοση καθώς προσδίδουν περισσότερη βαρύτητα στις εκτιμήσεις μεγαλύτερων πιθανοτήτων [50].

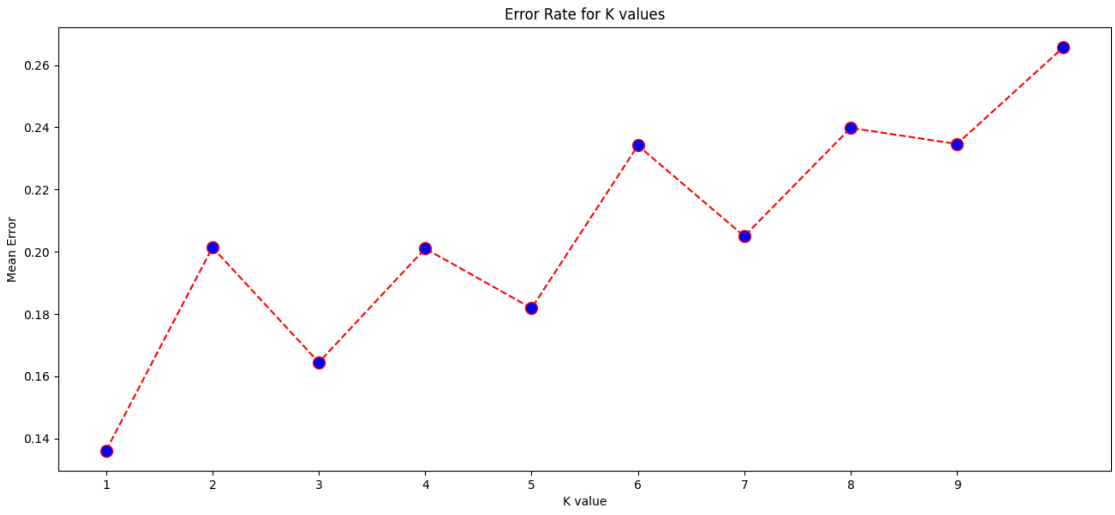


Σχήμα 2.17 : Δομή μοντέλου Ταξινομητή Ψηφοφορίας [50].

2.5 Υπερπαράμετροι και Βελτιστοποίηση

Από τις προηγούμενες ενότητες γίνεται αντιληπτό ότι οι αλγόριθμοι μηχανικής μάθησης διαθέτουν ρυθμίσεις οι οποίες ορίζονται από τον χρήστη με στόχο τον έλεγχο της συμπεριφοράς τους και τη βελτίωση της επίδοσης τους. Οι εν λόγω ρυθμίσεις κατονομάζονται ως υπερπαράμετροι (hyperparameters) και διαχωρίζονται από τις βασικές παραμέτρους οι οποίες ρυθμίζονται από το ίδιο το μοντέλο με βάση τα δεδομένα εκπαίδευσης [49]. Για παράδειγμα, ο αριθμός *k* στον αλγόριθμο K-Means αποτελεί μία υπερπαράμετρο καθώς ορίζεται από τον χρήστη και δεν διαμορφώνεται από τον αλγόριθμο κατά την λειτουργία του. Αντιθέτως, οι συντεταγμένες των κεντροειδών των ομάδων στον K-Means αποτελούν παραμέτρους του μοντέλου καθώς διαμορφώνονται επαναληπτικά από τα ίδια τα δεδομένα κατά την εκτέλεση του αλγορίθμου. Προς αποφυγή κάθε παρανόησης, σημειώνεται ότι ο χρήστης δεν ορίζει άμεσα την αρχική θέση των κεντροειδών, αλλά έμμεσα μέσω της υπερπαραμέτρου "τυχαία κατάσταση" (random state) που μοντελοποιεί ουσιαστικά μια γεννήτρια "τυχαίων" αριθμών.

Τα ορίσματα των υπερπαραμέτρων είναι συνήθως αριθμητικές τιμές ενώ υπάρχουν και υπερπαράμετροι μοντέλων που δέχονται ως ορίσματα συναρτήσεις, όπως για παράδειγμα το κριτήριο διάσπασης στον αλγόριθμο κατασκευής ενός Δένδρου Απόφασης. Η διαδικασία της εύρεσης κατάλληλων ορισμάτων των υπερπαραμέτρων ενός μοντέλου αποτελεί πρόβλημα βελτιστοποίησης καθώς στοχεύει όπως αναφέραμε προηγουμένως στην βελτίωση της επίδοσης του μοντέλου. Η κύρια μεθοδολογία υπολογισμού των βέλτιστων υπερπαραμέτρων ενός μοντέλου έγκειται στην εφαρμογή πολλαπλών εκτελέσεων του αλγορίθμου υπό μια συλλογή δυνατών τιμών ή συναρτήσεων που δύνανται να αποτελούν ορίσματα των εν λόγω υπερπαραμέτρων. Κάθε διαφορετική εκτέλεση αξιολογείται συγκριτικά μέσω κατάλληλων μετρικών και επιλέγονται τα ορίσματα αυτά που προκάλεσαν την αύξηση της επίδοσης. Σημειώνεται ότι κάθε υπερπαράμετρος εξετάζεται ανεξάρτητα των υπολοίπων για την αποφυγή τυχόν συγκρούσεων. Επίσης, είναι ιδιαίτερα σημαντικό η παραπάνω διαδικασία βελτιστοποίησης να εκτελείται είτε στα δεδομένα εκπαίδευσης είτε στα δεδομένα επικύρωσης και ποτέ στα δεδομένα ελέγχου διότι με αυτή τη πρακτική παρουσιάζεται μία υπεραισιόδοξη εικόνα της επίδοσης του μοντέλου που διέπεται από προκαταλήψεις [51].



Σχήμα 2.18 : Διάγραμμα μέσου σφάλματος μοντέλου K-NN για τον υπολογισμό της βέλτιστης υπερπαραμέτρου ***k***.

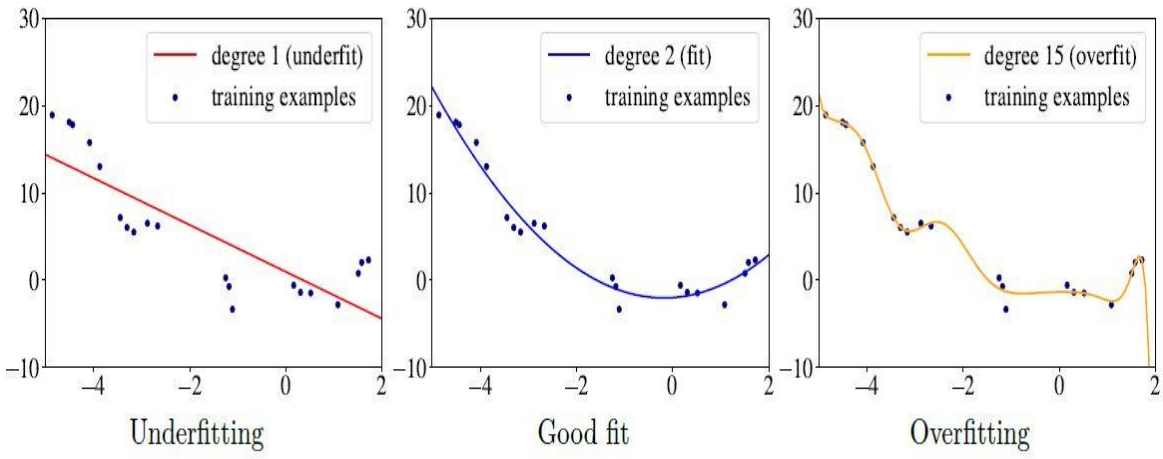
2.6 Υπερμοντελοποίηση και Υπομοντελοποίηση

Η αξιολόγηση ενός μοντέλου μηχανικής μάθησης είναι πρακτικά μια διαδικασία υπολογισμού του σφάλματος του μοντέλου και συμβαίνει σε δύο στάδια. Όταν το σφάλμα υπολογίζεται στο σύνολο των δεδομένων εκπαίδευσης αναφέρεται ως σφάλμα εκπαίδευσης (training error) ενώ όταν υπολογίζεται στο σύνολο των δεδομένων ελέγχου αναφέρεται ως σφάλμα γενίκευσης (generalization error). Το φαινόμενο της **υπερμοντελοποίησης (overfitting)** ή αλλιώς **υπερεκπαίδευσης (overtraining)** προκύπτει όταν ένα μοντέλο παρουσιάζει μικρό ή αμελητέο σφάλμα εκπαίδευσης και μεγάλο σφάλμα γενίκευσης. Στη βιβλιογραφία το πρόβλημα αυτό αναφέρεται και ως μεγάλη **διακύμανση (high variance)** του μοντέλου καθώς τότε παρουσιάζει μεγάλη ευαισθησία στο θόρυβο προσδίδοντας βαρύτητα σε μικρές διακυμάνσεις των δεδομένων εκπαίδευσης [49]. Δύο βασικοί παράγοντες που οδηγούν στην υπερεκπαίδευση ενός μοντέλου είναι :

- Η μεγάλη πολυπλοκότητα του μοντέλου ως προς τα δεδομένα
- Η μεγάλη διάσταση και ταυτόχρονα ο μικρός αριθμός των δεδομένων εκπαίδευσης

Για την αντιμετώπιση της υπερεκπαίδευσης απαιτείται συνήθως η απλοποίηση του μοντέλου (για παράδειγμα το κλάδεμα ενός Δένδρου Απόφασης), η μείωση των διαστάσεων και η εξάλειψη του θορύβου των δεδομένων, καθώς και η συλλογή μεγαλύτερου συνόλου δεδομένων εκπαίδευσης.

Στη περίπτωση που ένα μοντέλο μηχανικής μάθησης παρουσιάζει μεγάλο σφάλμα εκπαίδευσης τότε το πρόβλημα αυτό αναφέρεται ως φαινόμενο **υπομοντελοποίησης (underfitting)**. Στη βιβλιογραφία αναφέρεται και ως **υψηλή μεροληψία (high bias)** του μοντέλου καθώς τότε βασίζεται σε λανθασμένες υποθέσεις μεταξύ των χαρακτηριστικών των δεδομένων και των αντίστοιχων ετικετών τους [49]. Ένας βασικός παράγοντας που οδηγεί στο φαινόμενο αυτό είναι η απλοϊκότητα των μοντέλων ως προς τα δεδομένα (για παράδειγμα ένα γραμμικό μοντέλο). Για την αντιμετώπιση της υπομοντελοποίησης απαιτείται συνήθως η επιλογή μη γραμμικών μοντέλων με περισσότερες παραμέτρους καθώς και η εξαγωγή κατάλληλων χαρακτηριστικών από τα δεδομένα εκπαίδευσης που αποτελεί πεδίο έρευνας της μηχανικής χαρακτηριστικών (feature engineering).



Σχήμα 2.19 : Παράδειγμα υπερμοντελοποίησης και υπομοντελοποίησης [49].

Κεφάλαιο 3 : Εξόρυξη Δεδομένων και Εξαγωγή Γνώσης

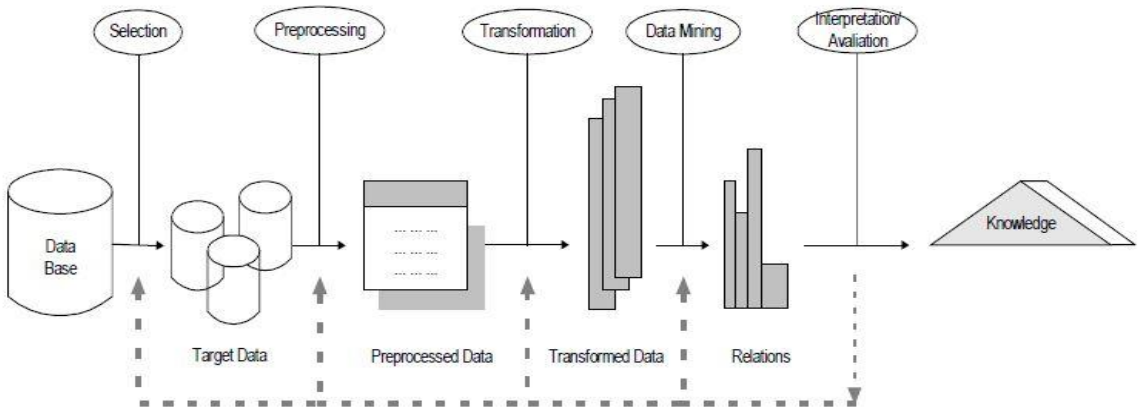
Στο κεφάλαιο αυτό πραγματοποιείται μια εισαγωγή στην έννοια της Εξόρυξης Δεδομένων (Data Mining) και Εξαγωγής Γνώσης (Knowledge Extraction). Στα κεφάλαια 3.1 και 3.2 παρουσιάζεται ο ορισμός και οι βασικές τεχνικές της Εξόρυξης Δεδομένων αντίστοιχα. Στο κεφάλαιο 3.3 αναλύουμε θεμελιώδεις τεχνικές που αξιοποιούνται κατά την προετοιμασία των υπό ανάλυση δεδομένων. Τέλος, στο κεφάλαιο 3.4 προσδιορίζεται ο κλάδος της επιστήμης της Μηχανικής Χαρακτηριστικών και παρουσιάζονται τεχνικές, μέθοδοι καθώς και αλγόριθμοι που εφαρμόσαμε για την εξαγωγή και επιλογή χαρακτηριστικών σε χρονοσειρές φορτίου ηλεκτρικής ενέργειας. Βιβλία στα οποία βασιστήκαμε είναι πρώτον το "Data Mining : Practical Machine Learning Tools and Techniques" των Witten, Frank, Hall και Pall [51], δεύτερον το "The Elements of Statistical Learning : Data Mining, Inference and Prediction" των Hastie, Tibshirani και Friedman [52] και τρίτον το "Data Mining : A Knowledge Discovery Approach" των Cios, Pedrycz, Swiniarski και Kurgan [24].

3.1 Ορισμός

Η Εξόρυξη Δεδομένων (Data Mining) ενσωματώνει μεθοδολογίες από τη Μηχανική Μάθηση, τη Στατιστική και τις Βάσεις Δεδομένων για την ανακάλυψη γνώσης υπό τη μορφή συσχετίσεων, προτύπων και τάσεων από μεγάλους όγκους δεδομένων. Αποτελεί στάδιο της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Databases – KDD) αν και σε επίπεδο ορολογίας αυτοί οι δύο όροι χρησιμοποιούνται ως συνώνυμα.

Η Εξαγωγή Γνώσης μέσω της Εξόρυξης Δεδομένων είναι μία επαναληπτική διαδικασία και αποτελείται από τα ακόλουθα βήματα [24] :

- Κατανόηση του προβλήματος και των δεδομένων (Data Understanding)
- Επιλογή κατάλληλων δεδομένων (Data Selection)
- Ενσωμάτωση δεδομένων (Data Integration)
- Προεπεξεργασία δεδομένων (Data Preprocessing)
- Μετασχηματισμός Δεδομένων (Data Transformation)
- Μοντελοποίηση και Εξόρυξη Δεδομένων (Modeling and Data Mining)
- Αξιολόγηση (Evaluation)
- Αναπαράσταση γνώσης (Knowledge Representation)



Σχήμα 3.1 : Διαδικασία Εξαγωγής Γνώσης [29].

3.2 Τεχνικές Εξόρυξης Δεδομένων

Οι βασικές τεχνικές και αλγόριθμοι που αξιοποιούνται στο στάδιο της Εξόρυξης Δεδομένων χωρίζονται σε δύο κατηγορίες, τις περιγραφικές μεθόδους και τις μεθόδους πρόβλεψης.

1. **Περιγραφικές Μέθοδοι (Descriptive Methods)**

   Στη κατηγορία αυτή ανήκουν μέθοδοι και τεχνικές αναζήτησης προτύπων που στοχεύουν στο σχηματισμό μίας διορατικής περιγραφής των δεδομένων και των ιδιοτήτων τους. Οι βασικές τεχνικές της εν λόγω κατηγορίας είναι οι εξής:

   - Ομαδοποίηση ή αλλιώς Συσταδοποίηση (Clustering)

     Στο δεύτερο κεφάλαιο έχει πραγματοποιηθεί αναλυτική περιγραφή του προβλήματος Ομαδοποίησης.

   - Κανόνες Συσχέτισης (Association Rules)

     Στόχος αυτών των μεθόδων είναι η ανακάλυψη κανόνων "if-then" που εκφράζουν την εξάρτηση μεταξύ των μεταβλητών του συνόλου δεδομένων

   - Σύνοψη (Summarization)

     Αναφέρεται σε τεχνικές που στοχεύουν σε μία σύντομη παρουσίαση των βασικών χαρακτηριστικών των δεδομένων. Μερικά παραδείγματα είναι η δημιουργία στατιστικών πινάκων και τα διαγράμματα Q-Q Plots (quantile – quantile plots) που αξιοποιούνται για την ανίχνευση ασσυμετρίας (skewness) στην κατανομή των δεδομένων.

2. **Μέθοδοι Πρόβλεψης (Prediction Methods)**

   Πρόκειται για μεθόδους που αποσκοπούν στην πρόβλεψη μελλοντικών και άγνωστων τιμών. Οι βασικές τεχνικές που ανήκουν σε αυτή τη κατηγορία είναι :

   - Παλινδρόμηση (Regression)
   - Ταξινόμηση ή αλλιώς Κατηγοριοποίηση (Classification)
   - Ανίχνευση Ανωμαλιών (Anomaly Detection)

   Η περιγραφή των προβλημάτων Παλινδρόμησης, Ταξινόμησης και Ανίχνευσης Ανωμαλιών έχει πραγματοποιηθεί στο δεύτερο κεφάλαιο.

3.3 Προετοιμασία των Δεδομένων

Για την επίτευξη του στόχου της εκάστοτε εφαρμογής Εξόρυξης Δεδομένων απαιτείται αρχικά η δημιουργία μίας βάσης κατάλληλων δεδομένων η οποία θα αποτελέσει το σύνολο δεδομένων εισόδου σε μοντέλα Μηχανικής Μάθησης. Η επίδοση όμως κάθε μοντέλου επηρεάζεται δραματικά από τη ποιότητα και τη μορφή των υπό ανάλυση δεδομένων. Συνεπώς, ένα από τα βασικότερα στάδια της Εξαγωγής Γνώσης είναι η προετοιμασία των δεδομένων και αποτελεί το πιο χρονοβόρο στάδιο της όλης διαδικασίας. Το στάδιο αυτό αναφέρεται στα βήματα επιλογής, ενσωμάτωσης, προεπεξεργασίας και μετασχηματισμού των δεδομένων και αξίζει να αναφέρουμε το γεγονός ότι καταλαμβάνει συνήθως το 50% με 90% του συνολικού χρόνου που απαιτείται για την ολοκλήρωση της Εξαγωγής Γνώσης [25,72]. Στις υποενότητες που ακολουθούν παρουσιάζουμε βασικές μεθόδους και εργαλεία που στοχεύουν στη κατάλληλη προετοιμασία των δεδομένων.

3.3.1 Καθαρισμός Δεδομένων (Data Cleansing)

Ο καθαρισμός δεδομένων αποτελεί τη διαδικασία εντοπισμού, διόρθωσης ή και αφαίρεσης ελλιπών και λανθασμένων τιμών που ενδέχεται να περιέχονται στο σύνολο μιας βάσης δεδομένων. Οι ασυνέπειες αυτές συνήθως προκαλούνται από σφάλματα κατά τη καταχώρηση, μεταφορά και αποθήκευση των δεδομένων. Σημειώνεται ότι μια βάση δεδομένων αποτελείται ουσιαστικά από μια συλλογή πινάκων με ετικέτες – δείκτες στους οποίους είναι καταχωρημένα τα δεδομένα. Κάθε αντικείμενο – οντότητα αποθηκεύεται σε μια γραμμή του πίνακα και οι στήλες αποτελούν τα χαρακτηριστικά (για στατικά δεδομένα) ή τις χρονικές παρατηρήσεις στη περίπτωση που τα δεδομένα είναι χρονοσειρές.

Για την αντιμετώπιση των **ελλιπών τιμών (missing values)** υπάρχουν διάφορες στρατηγικές που μπορεί να ακολουθήσει κανείς και η επιλογή τους καθορίζεται από τις απαιτήσεις της εφαρμογής καθώς και από τα υπό ανάλυση δεδομένα. Οι βασικότερες από αυτές για δεδομένα χρονοσειρών είναι :

- Διαγραφή οντοτήτων

  Σύμφωνα με αυτή τη στρατηγική απομακρύνεται ολοκληρωτικά από τον πίνακα κάθε γραμμή (χρονοσειρά) που περιέχει ελλιπείς τιμές. Η μέθοδος αυτή ενδείκνυται μόνο σε περιπτώσεις όπου οι απούσες τιμές επηρεάζουν έναν μικρό αριθμό από το σύνολο των οντοτήτων (χρονοσειρών).

- Γραμμική Παρεμβολή

  Η μέθοδος αυτή παρεμβάλλει γραμμικά τις παρατηρήσεις που βρίσκονται αμέσως πριν και μετά από μια απούσα παρατήρηση. Αν θεωρήσουμε τη παρατήρηση $f(x_0)$ τη στιγμή x_0 και τη παρατήρηση $f(x_1)$ τη στιγμή x_1 τότε η ενδιάμεση άγνωστη παρατήρηση $f(x)$ υπολογίζεται ως εξής :

$$f(x) = f(x_0) + \frac{f(x_1)-f(x_0)}{x_1-x_0}\ (x - x_0) \qquad (3.1)$$

- Αντικατάσταση με τη μέση τιμή ή τη διάμεσο της οντότητας

  Η μέθοδος αυτή αντικαθιστά τις απούσες παρατηρήσεις της χρονοσειράς με τη διάμεσο ή την μέση τιμή των χρονικών παρατηρήσεων που είναι καταχωρημένες στην αντίστοιχη γραμμή του πίνακα (χρονοσειρά)

- Πρόβλεψη

  Η μέθοδος αυτή αξιοποιεί αλγορίθμους Επιβλεπόμενης Μάθησης για την πρόβλεψη των ελλιπών τιμών. Ουσιαστικά απαιτείται να λύσουμε ένα πρόβλημα Παλινδρόμησης.

- Αντικατάσταση με τύπο δεδομένων "NaN"

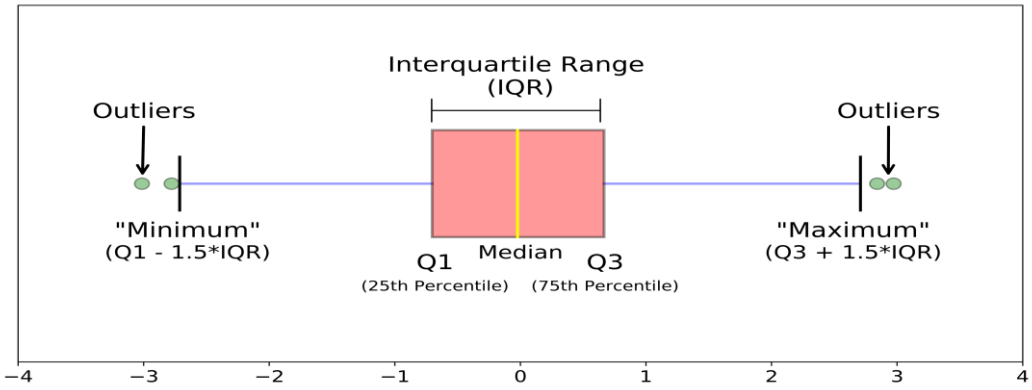
  Ο ειδικός τύπος δεδομένων "NaN" (Not a Number) χρησιμοποιείται για την αναπαράσταση οποιασδήποτε τιμής που δεν είναι καθορισμένη ή δεν δύναται να εμφανιστεί. Η τεχνική αυτή είναι συνήθως ακατάλληλη καθώς οι περισσότεροι αλγόριθμοι μηχανικής μάθησης δεν δύνανται να επεξεργαστούν "NaN" τύπους δεδομένων.

Στη περίπτωση που στο σύνολο δεδομένων υπάρχουν λανθασμένες καταχωρήσεις τιμών, τότε στα δεδομένα υπάρχει **θόρυβος**. Ο θόρυβος αποτελεί συχνό φαινόμενο και απαιτεί ιδιαίτερη μεταχείρηση καθώς επηρεάζει σημαντικά τους αλγορίθμους Μηχανικής Μάθησης. Συγκεκριμένα, ένα μοντέλο καθίσταται λιγότερο αξιόπιστο όταν τα δεδομένα περιέχουν λανθασμένες τιμές που δεν αντιστοιχούν στη πραγματικότητα αλλά ανήκουν στο επιτρεπτό σύνολο τιμών. Επίσης, ενδέχεται το σύνολο δεδομένων να περιέχει τιμές που είναι είτε διαφορετικές είτε ανώμαλες σε σχέση με τις υπόλοιπες τιμές των οντοτήτων του συνόλου δεδομένων και αναφερόμαστε σε αυτές ως **ακραίες τιμές (outliers)** [76].

Για τον εντοπισμό του θορύβου και των ακραίων τιμών σε ένα σύνολο δεδομένων χρησιμοποιούμε κυριώς τις παρακάτω τεχνικές :

- Διαγράμματα BoxPlots

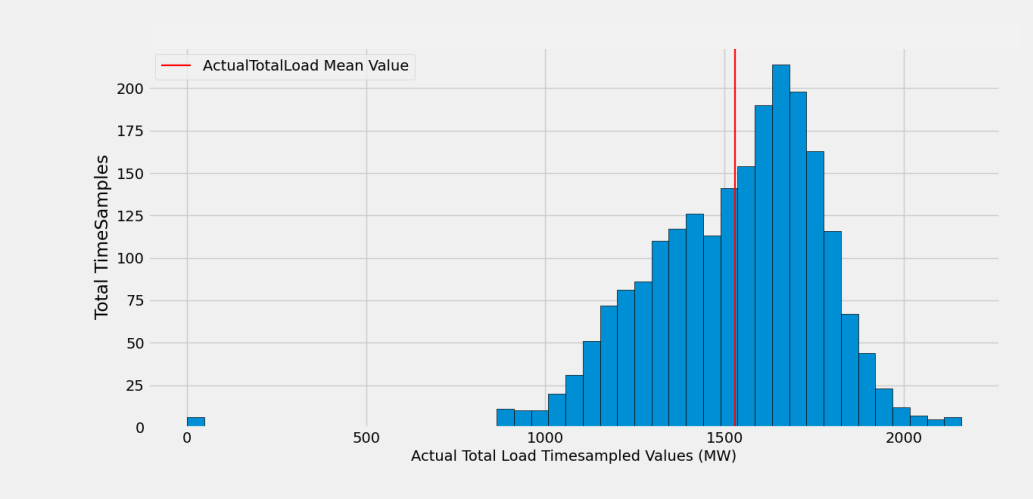
  Τα διαγράμματα BoxPlots ή αλλιώς Whisker Plots αποτελούν διαγνωστικά εργαλεία οπτικοποίησης των ακραίων τιμών [75]. Η παρακάτω εικόνα παρουσιάζει τη μορφή και τα χαρακτηριστικά ενός τυπικού διαγράμματος BoxPlot.



Σχήμα 3.2 : Παράδειγμα διαγράμματος BoxPlot.[*towardsdatascience.com*].

- Ιστογράμματα

  Τα ιστογράμματα παρέχουν μια οπτική ερμηνεία των δεδομένων υποδεικνύοντας τον αριθμό των παρατηρήσεων που βρίσκονται εντός ενός εύρους τιμών. Αυτά τα εύρη τιμών ονομάζονται κλάσεις ή κάδοι (bins). Όταν ο αριθμός των κάδων είναι επαρκώς μεγάλος, τότε μπορούμε να εκτιμήσουμε προσεγγιστικά τον τύπο κατανομής των δεδομένων. Επίσης, η ύπαρξη ακραίων τιμών μπορεί να εντοπιστεί μέσω ενός ιστογράμματος καθώς τότε θα παρουσιάζονται κάδοι που απέχουν σημαντική απόσταση από τους υπόλοιπους του διαγράμματος. Στη συνέχεια παρουσιάζουμε ένα παράδειγμα ιστογράμματος όπου παρατηρείται το εν λόγω φαινόμενο.



Σχήμα 3.3 : Ιστόγραμμα που υποδεικνύει την ύπαρξη ακραίων τιμών.

- Ανάλυση Ομάδων (Clustering)

  Εφαρμόζοντας αλγορίθμους ομαδοποίησης στο σύνολο δεδομένων επιτυγχάνεται η ομαδοποίηση παρόμοιων αντικειμένων, Έτσι, έχουμε ως αποτέλεσμα τα ανόμοια αντικείμενα να εντάσσονται σε διαφορετικές ομάδες και να τα εντοπίζουμε με ευκολία.

- Στατιστικός εντοπισμός

  Συγκρίνοντας τις αποκλίσεις των εξεταζόμενων τιμών με την μέση τιμή των οντοτήτων στις οποίες ανήκουν, μπορούμε στη περίπτωση που είναι εξαιρετικά μεγάλες να θεωρήσουμε τις εν λόγω τιμές ως θόρυβο ή ως ακραίες τιμές.

Μετά τον εντοπισμό του θορύβου και των ακραίων τιμών στο σύνολο των δεδομένων μπορούμε να τις αντικαταστήσουμε με τον τύπο δεδομένων "NaN" που αντιπροσωπεύει ελλιπείς τιμές και μετέπειτα να τις διαχειριστούμε σύμφωνα με τις τεχνικές αντιμετώπισης ελλιπών τιμών που έχουμε αναφέρει.

3.3.2 Κανονικοποίηση Δεδομένων (Data Scaling – Normalization)

Συνήθως, σε σύνολα δεδομένων μεγάλου όγκου παρουσιάζονται οντότητες που περιέχουν τιμές διαφορετικής κλίμακας και τάξης μεγέθους. Οι περισσότεροι αλγόριθμοι Μηχανικής Μάθησης παρουσιάζουν ευαισθησία σε τέτοιου είδους αποκλίσεις τιμών με αποτέλεσμα να μειώνεται δραστικά η επίδοση τους. Για παράδειγμα, στον αλγόριθμο K-Means όπου υπολογίζονται αποστάσεις μεταξύ των δεδομένων και των κεντροειδών των ομάδων, προκύπτουν προβληματικές ομαδοποιήσεις όταν η κλίμακα των τιμών διαφέρει πολύ καθώς οι μικρές τιμές επηρεάζουν ελάχιστα το μέτρο απόστασης ενώ οι μεγάλες το επηρεάζουν σημαντικά. Για την αντιμετώπιση του εν λόγω προβλήματος απαιτείται ο μετασχηματισμός των δεδομένων σε ένα κατάλληλο εύρος τιμών. Η διαδικασία αυτή ονομάζεται κανονικοποίηση. Οι μέθοδοι κανονικοποίησης διακρίνονται σε δύο διαφορετικές κατηγορίες, τις γραμμικές και τις μη γραμμικές μεθόδους.

Έστω ένα διάνυσμα $X = (x_1, \dots, x_i, \dots, x_n)$ και $X_{new} = (x_{1_{new}}, \dots, x_{i_{new}}, \dots, x_{n_{new}})$ το κανονικοποιημένο διάνυσμα που προκύπτει από τη κανονικοποίηση του X. Στη συνέχεια παρουσιάζουμε διάφορες τεχνικές κανονικοποίησης μιας τυχαίας μεταβλητής x_i του διανύσματος X.

1. **Γραμμικές Μέθοδοι Κανονικοποίησης**

   Όταν τα δεδομένα κανονικοποιούνται με γραμμικές μεθόδους, τότε οι μεταξύ τους σχετικές αποστάσεις παραμένουν ίδιες και δεν επηρεάζεται η μορφή της καμπύλης της κατανομής που ακολουθούν.

   - Z-Score (Standardization)

     Η μέθοδος αυτή ενδείκνυται όταν τα δεδομένα ακολουθούν κανονικές κατανομές και μετασχηματίζει κάθε διάνυσμα ώστε να έχει μέση τιμή μηδέν και τυπική απόκλιση μονάδα. Όταν υπάρχουν πολλές ακραίες τιμές στα δεδομένα η μέθοδος αυτή κρίνεται ακτάλληλη.

$$z_{i_{new}} = x_{i_{new}} = \frac{x_i - Mean(X)}{Std(X)} \qquad (3.2)$$

   - Min – Max Scaling

     Η μέθοδος αυτή κανονικοποιεί όλες τις μεταβλητές εντός του διαστήματος [*min, max*]. Η συνηθέστερη περίπτωση είναι η επιλογή του min ως μηδέν και του max ως μονάδα. Προσφέρει ικανοποιητικά αποτελέσματα ακόμα και όταν τα δεδομένα δεν ακολουθούν κανονικές κατανομές αλλά κρίνεται ακατάλληλη όταν παρουσιάζονται πολλές ακραίες τιμές.

$$x_{i_{new}} = \frac{x_i - Min(X)}{Max(X) - Min(X)} \ (max - min) + min \qquad (3.3)$$

- Max Absolute Scaling

  Η μέθοδος αυτή κανονικοποιεί όλες τις μεταβλητές εντός του διαστήματος [-1,1] και είναι ακατάλληλη όταν παρουσιάζονται ακραίες τιμές.

$$x_{i_{new}} = \frac{x_i}{Abs(Max(X))} \tag{3.4}$$

- Robust Scaler

  Η μέθοδος αυτή είναι κατάλληλη όταν τα δεδομένα ακολουθούν ασύμμετρες κατανομές και όταν παρουσιάζονται πολλές ακραίες τιμές. Τα ποσοστιαία σημεία (quantiles) Q_i αποτελούν σημεία κοπής που διαιρούν το εύρος μιας κατανομής πιθανότητας σε συνεχή ισοπίθανα διαστήματα. Η διαφορά του τρίτου ποσοστιαίου σημείου από το πρώτο ονομάζεται Interquantile Range (IQR).

$$x_{i_{new}} = \frac{x_i - Q_2(X)}{Q_3(X) - Q_1(X)} = \frac{x_i - Median(X)}{IQR(X)} \tag{3.5}$$

- Vector Normalization

  Η μέθοδος αυτή κανονικοποιεί τα δεδομένα στο διάστημα [0,1] όταν όλα τα δεδομένα είναι θετικά. Στη περίπτωση που υπάρχουν αρνητικά δεδομένα τότε τα κανονικοποιεί στο διάστημα [-1,1]. Είναι κατάλληλη όταν τα δεδομένα πρόκειται να αποτελέσουν το σύνολο εισόδου σε αλγορίθμους Ομαδοποίησης ενώ κρίνεται ακτάλληλη για εφαρμογές Ταξινόμησης. Επίσης, δεν ενδείκνυται όταν υπάρχουν ακραίες τιμές στα δεδομένα.

$$x_{i_{new}} = \frac{x_i}{\left|\left|X\right|\right|_p} \tag{3.6}$$

  Όπου $\left|\left|X\right|\right|_p$ η L_p νόρμα του διανύσματος X.
  Συνήθως επιλέγεται $p = 2$ ή $p = 1$.

- Mean Normalization

$$x_{i_{new}} = \frac{x_i}{Mean(X)} \tag{3.7}$$

- Sum Normalization

$$x_{i_{new}} = \frac{x_i}{Sum(X)} \tag{3.8}$$

Οι μέθοδοι Mean Normalization και Sum Normalization επηρεάζονται σε μεγάλο βαθμό αρνητικά από την ύπαρξη ακραίων τιμών. Συνεπώς, συνιστάται να αποφεύγονται.

2. Μη Γραμμικές Μέθοδοι Κανονικοποίησης

Όταν τα δεδομένα κανονικοποιούνται με μη γραμμικές μεθόδους τότε οι μεταξύ τους σχετικές αποστάσεις και η μορφή της καμπύλης της κατανομής που ακολουθούν μεταβάλλονται.

- Logarithmic Normalization

  Η μέθοδος αυτή χρησιμοποιείται όταν η κλίμακα των δεδομένων διαφέρει σε πολύ μεγάλο βαθμό. Συνήθως κρίνεται ακατάλληλη όταν τα δεδομένα δεν είναι αυστηρά θετικά. Σε γεννικές γραμμές επιλέγεται είτε ο δεκαδικός είτε ο φυσικός λογάριθμος.

$$x_{i_{new}} = \log_n x_i \tag{3.9}$$

- Power Transformations (Yoe - Johnson, Box - Cox)

  Οι μετασχηματισμοί ισχύος (Power Transformations) στοχεύουν στον περιορισμό της διακύμανσης και την ελαχιστοποίηση της ασυμμετρίας των δεδομένων.

  i. Yeo – Johnson Transform :

$$x_{i_{new}} \begin{cases} \frac{(x_i+1)^{\lambda}-1}{\lambda}, & \lambda \neq 0 \ \kappa\alpha\iota \ x_i \geq 0 \\ \ln(x_i+1), & \lambda = 0 \ \kappa\alpha\iota \ x_i \geq 0 \\ -\frac{(-x_i+1)^{2-\lambda}-1}{2-\lambda}, & \lambda \neq 2 \ \kappa\alpha\iota \ x_i < 0 \\ -\ln(-x_i+1), & \lambda = 2 \ \kappa\alpha\iota \ x_i < 0 \end{cases} \tag{3.10}$$

  ii. Box – Cox Transformation :

$$x_{i_{new}} = \begin{cases} \frac{x_i^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln(x_i), & \lambda = 0 \end{cases} \tag{3.11}$$

  Ο μετασχηματισμός Box – Cox εφαρμόζεται αυστηρά μόνο σε θετικά δεδομένα. Και στις δύο μεθόδους ο μετασχηματισμός παραμετροποιείται από το λ, το οποίο καθορίζεται μέσω της εκτίμησης της μέγιστης πιθανοφάνειας [77, 78].

3.3.3 Μείωση Διαστάσεων (Dimensionality Reduction)

Όπως ήδη έχουμε αναφέρει στην ενότητα 3.3.1 σε μια βάση δεδομένων κάθε αντικείμενο – οντότητα αποθηκεύεται σε μια γραμμή ενός πίνακα και οι στήλες αποτελούν τα χαρακτηριστικά (features) όταν τα δεδομένα είναι στατικά ή τις χρονικές παρατηρήσεις στη περίπτωση που τα δεδομένα είναι χρονοσειρές. Σημειώνεται ότι τα χαρακτηριστικά αναφέρονται στη βιβλιογραφία και ως διαστάσεις (dimensions) ή γνωρίσματα (attributes) των οντοτήτων. Η ύπαρξη πολλών διαστάσεων επιβαρύνει τη διαδικασία Εξαγωγής Γνώσης καθώς αυξάνει το υπολογιστικό κόστος και την πολυπλοκότητα του προβλήματος. Το πρόβλημα της ύπαρξης δεδομένων πολλών διαστάσεων αναφέρεται επίσης ως «κατάρα των διαστάσεων» (curse of dimensionality) [73]. Ειδικά σε εφαρμογές ανάλυσης χρονοσειρών το πρόβλημα των διαστάσεων αποτελεί σοβαρό εμπόδιο με αποτέλεσμα την αναζήτηση τεχνικών και μεθόδων που μεταφέρουν την ανάλυση από το πεδίο του χρόνου σε ένα πεδίο κατάλληλων επιλεγμένων χαρακτηριστικών που εξάγονται από τις χρονοσειρές με σκοπό την μείωση των διαστάσεων τους. Ο κλάδος που πραγματεύεται την εξαγωγή και επιλογή κατάλληλων χαρακτηριστικών από οντότητες δεδομένων ονομάζεται μηχανική χαρακτηριστικών (feature engineering) και θα αναλυθεί ενδελεχώς στην επόμενη ενότητα.

Σημειώνεται ότι η μείωση των διαστάσεων (dimensionality reduction) και η επιλογή χαρακτηριστικών (feature selection) δεν είναι ταυτόσημες έννοιες δεδομένου ότι η μείωση των διαστάσεων μπορεί να πραγματοποιηθεί είτε με την επιλογή χαρακτηριστικών είτε με την προβολή των αρχικών δεδομένων σε ένα νέο χώρο μικρότερης διάστασης. Οι τεχνικές μείωσης διαστάσεων αφαιρούν περιττά και εξαιρετικά συσχετισμένα χαρακτηριστικά, μειώνουν τον θόρυβο και συμβάλλουν γενικά στην ευκολότερη ερμηνεία των μοντέλων. Μια ευρέως χρησιμοποιούμενη τεχνική μείωσης διαστάσεων είναι η Ανάλυση Κύριων Συνιστωσών (Principal Components Analysis – PCA).

- **<u>Ανάλυση Κύριων Συνιστωσών (PCA)</u>** [24]

  Έστω ένα αρχικό σύνολο δεδομένων με k διανύσματα n διαστάσεων . Με την μέθοδο PCA υπολογίζουμε ένα σύστημα m ορθογώνιων διανυσμάτων ($m < n$) τα οποία αποτελούν ένα νέο σύστημα αξόνων για τη προβολή των αρχικών δεδομένων σε ένα διαφορετικό χώρο μικρότερων διαστάσεων. Συγκεκριμένα, αρχικά κάθε ένα από τα συνολικά n χαρακτηριστικά διανύσματα μετασχηματίζεται γραμμικά ώστε να έχει μηδενική μέση τιμή. Έπειτα, υπολογίζεται ο πίνακας συνδιασποράς (covariance matrix) των νέων μετασχηματισμένων διανυσμάτων, ο οποίος στη συνέχεια διαγωνοποιείται για τον υπολογισμό των ιδιοδιανυσμάτων (eigenvectors) του. Τα εν λόγω υπολογισμένα ιδιοδιανύσματα ονομάζονται κύριες συνιστώσες (principal components). Συνεπώς, μέσω των κύριων συνιστώσών συντίθεται ένα νέο σύστημα αξόνων των δεδομένων. Σε επόμενο βήμα υπολογίζονται οι ιδιοτιμές των κύριων συνιστωσών οι οποίες ταξινομούνται σε φθίνουσα σειρά με βάση το μέγεθος της ιδιοτιμής τους. Οι πρώτες κύριες συνιστώσες στην εν λόγω σειρά κατάταξης σημαντικότητας ενσωματώνουν το μεγαλύτερο μέρος της πληροφορίας που αφορά τη διασπορά των υπό ανάλυση δεδομένων. Τέλος, απαλείφονται οι λιγότερο σημαντικές κύριες συνιστώσες των οποίων η απουσία δεν επιφέρει σημαντική απώλεια πληροφορίας. Με την κατάλληλη επιλογή των l πρώτων σημαντικότερων ιδιοδιανυσμάτων έχουμε στη διάθεση μας μια ικανοποιητική προσέγγιση των αρχικών δεδομένων. Η μέθοδος PCA αξιοποιείται επίσης και για την οπτικοποίηση πολυδιάστατων δεδομένων καθώς με την επιλογή των τριών ή

δύο πρώτων κύριων συνιστωσών καθίσταται εφικτό να προβληθούν τα δεδομένα σε έναν τρισδιάστατο ή δισδιάστατο χώρο αντίστοιχα.

Αξίζει να αναφέρουμε ότι η οπτικοποίηση πολυδιάστατων δεδομένων σε δισδιάστατους και τρισδιάστατους χώρους μπορεί επίσης να επιτευχθεί μέσω πολλών διαφορετικών αλγορίθμων και τεχνικών που έχουν αναπτυχθεί τα τελευταία τριάντα χρόνια. Οι πιο δημοφιλείς από αυτές είναι :

- t-Distributed Stohastic Neighbor Embedding (t-SNE)
- Linear Discriminant Analysis (LDA)
- Uniform Manifold Approximation and Projection (UMAP)
- Radial Visualization (RadViz)
- Parallel Coordinates

3.4 Μηχανική Χαρακτηριστικών (Feature Engineering)

Η Μηχανική Χαρακτηριστικών (Feature Engineering) είναι η διαδικασία μετατροπής δεδομένων σε κατάλληλα χαρακτηριστικά που αντιπροσωπεύουν προβλήματα και στοχεύει στη βελτίωση της επίδοσης των μοντέλων μηχανικής μάθησης. Συγκεκριμένα, πραγματεύεται την ανάλυση, την επιλογή, την αξιολόγηση και τον μετασχηματισμό χαρακτηριστικών καθώς και την εξαγωγή χαρακτηριστικών (feature extraction) από πηγαία δεδομένα. Συνεπώς, η μηχανική χαρακτηριστικών εφαρμόζεται σε οποιοδήποτε στάδιο της εξαγωγής γνώσης από βάσεις δεδομένων αξιοποιώντας τεχνικές και αλγορίθμους που επενεργούν είτε σε ακατέργαστα είτε σε προεπεξεργασμένα δεδομένα.

Καθώς η μηχανική χαρακτηριστικών στοχεύει στη βελτίωση της επίδοσης των μοντέλων μηχανικής μάθησης, είναι ιδιαίτερα σημαντικό να διατυπωθεί η διαδικασία αξιολόγησης της εφαρμογής της στα πλαίσια της επίλυσης προβλημάτων βελτιστοποίησης. Τα κύρια βήματα της εν λόγω διαδικασίας είναι [72] :

1. Εκπαιδεύουμε και αξιολογούμε τα μοντέλα μηχανικής μάθησης πριν την εφαρμογή οποιασδήποτε διαδικασίας μηχανικής χαρακτηριστικών. Τα αποτελέσματα των αξιοποιημένων μετρικών αξιολόγησης συνθέτουν μια εικόνα της βασικής επίδοσης (baseline performance).

2. Εφαρμόζουμε στο σύνολο δεδομένων συνδιασμούς από διαδικασίες που εμπίπτουν στον κλάδο της μηχανικής χαρακτηριστικών.

3. Για κάθε διαδικασία μηχανικής χαρακτηριστικών που υλοποιήσαμε αξιολογούμε εκ νέου τα μοντέλα και συγκρίνουμε τα αποτελέσματα με την αρχική βασική επίδοση.

4. Αν η επίδοση των μοντελων βελτιωθεί θεωρούμε ότι η συγκεκριμένη διαδικασία μηχανικής χαρακτηριστικών είναι χρήσιμη και την ενσωματώνουμε στο σκελετό της εφαρμογής.

3.4.1 Εξαγωγή Χαρακτηριστικών (Feature Extraction)

Ο όρος εξαγωγή χαρακτηριστικών αναφέρεται σε αλγορίθμους μετασχηματισμού του αρχικού συνόλου δεδομένων σε ένα νέο σύνολο συνήθως μικρότερης διάστασης. Στη βιβλιογραφία αναφέρεται επίσης είτε ως παραγωγή χαρακτηριστικών (feature generation) είτε ως κατασκευή χαρακτηριστικών (feature construction). Η εξαγωγή χαρακτηριστικών πραγματοποιείται μετά από την ενσωμάτωση και τον καθαρισμό των συλλεχθέντων δεδομένων και έχει ως στόχο την παραγωγή χαρακτηριστικών τα οποία περιέχουν πληροφορία κατάλληλη για χρήση σε εφαρμογές πρόβλεψης και ταξινόμησης. Η πληροφορία αυτή θα πρέπει να υποστηρίζει τον ευκολότερο διαχωρισμό των διαφορετικών κλάσεων του προβλήματος και την αποδοτικότερη αναγνώριση των διαφόρων προτύπων [24].

Οι αλγόριθμοι και οι συναρτήσεις που εξάγουν χαρακτηριστικά από ένα σύνολο δεδομένων αναφέρονται στη βιβλιογραφία ως υπολογιστές χαρακτηριστικών (feature calculators). Στη περίπτωση που τα δεδομένα από τα οποία εξάγονται τα χαρακτηριστικά είναι χρονοσειρές, τότε οι αλγόριθμοι εξαγωγής χαρακτηριστικών διακρίνονται στις εξής τρεις βασικές κατηγοριές :

1. **Εξαγωγής χαρακτηριστικών στο πεδίο του χρόνου**

   Παραδείγματα αποτελούν μέθοδοι αποσύνθεσης χρονοσειρών σε στατιστικούς δείκτες όπως τάσεις (trends), ποσοστιαία σημεία (quantiles) και συντελεστές αυτοσυσχέτισης (autocorrelation).

2. **Εξαγωγής χαρακτηριστικών στο πεδίο της συχνότητας**

   Για παράδειγμα, ο υπολογισμός συντελεστών του μετασχηματισμού Fourier (FT), του γήγορου μετασχηματισμού Fourier (FFT) και του διακριτού μετασχηματισμού Fourier (DFT).

3. **Εξαγωγής χαρακτηριστικών στο πεδίο χρόνου – συχνότητας**

   Για παράδειγμα, ο υπολογισμός συντελεστών του μετασχηματισμού Fourier βραχέως χρόνου (STFT), του συνεχή μετασχηματισμού κυματιδίων (CWT) και του διακριτού μετασχηματισμού κυματιδίων (DWT).

Σε εφαρμογές και προβλήματα στα οποία δεν έχουμε επαρκή εμπειρία, εφόσον το επιτρέπουν οι διαθέσιμες υπολογιστικές δυνατότητες, συνιστάται η εξαγωγή όλων των δυνατών χαρακτηριστικών που επιτρέπει η μορφή των αρχικών δεδομένων και στη συνέχεια η επιλογή των πλέον κατάλληλων μέσω τεχνικών αξιολόγησης και υπολογισμού της στατιστικής τους σημασίας υπό το πλαίσιο της εκάστοτε εφαρμογής.

3.4.2 Επιλογή Χαρακτηριστικών (Feature Selection)

Η επιλογή κατάλληλων χαρακτηριστικών αποτελεί ένα NP-δύσκολο (NP-hard) πρόβλημα καθώς σε ένα σύνολο δεδομένων με n χαρακτηριστικά υπάρχουν 2^n δυνατά υποσύνολα που πρέπει να εξεταστούν ως προς τη σημαντικότητα τους σύμφωνα με το υπό ανάλυση πρόβλημα. Στόχος των μεθόδων αυτών είναι η επιλογή κατάλληλων χαρακτηριστικών τα οποία οδηγούν σε μεγάλες αποστάσεις των διαφορετικών κλάσεων μεταξύ τους και σε μικρές διακυμάνσεις εντός των προτύπων της ίδιας κλάσης [24]. Οι μέθοδοι επιλογής χαρακτηριστικών χωρίζονται σε τρεις ευρείες κατηγορίες, τις στατιστικές μεθόδους (statistical based), τις μεθόδους βασιμένες σε μοντέλα (model based) και τις ενσωματωμένες μεθόδους (embedded methods) [15]. Οι στατιστικές μέθοδοι βασίζονται στην εφαρμογή στατιστικών ελέγχων σημαντικότητας (statistical significance tests) και στη βιβλιογραφία αναφέρονται επίσης ως μέθοδοι φίλτρου (filter methods). Οι μέθοδοι μοντέλων βασίζονται στην εκπαίδευση ενός δευτερεύοντος μοντέλου μηχανικής μάθησης όπου αξιοποιοείται η προγνωστική του ισχύς για την επιλογή των κατάλληλων χαρακτηριστικών. Στη βιβλιογραφία οι μέθοδοι που βασίζονται σε μοντέλα αναφέρονται επίσης ως μέθοδοι περιτυλίγματος (wrapper methods). Οι ενσωματωμένες μέθοδοι συνδιάζουν τις ιδιότητες των μεθόδων φίλτρου και περιτυλίγματος καθώς αξιοποιούνται μοντέλα που διαθέτουν μεθόδους αξιολόγησης χαρακτηριστικών, όπως για παράδειγμα τα Δένδρα Απόφασης.

3.4.2.1 Μέθοδοι Φίλτρου (Filter Methods)

Οι μέθοδοι φίλτρου επιλογής χαρακτηριστικών χρησιμοποιούνται λόγω της απλότητας τους και της ανεξαρτησίας τους από την μεροληψία (bias) που εισάγεται από τα μοντέλα μηχανικής μάθησης. Συνεπώς, μπορούμε να εφαρμόσουμε μια μέθοδο φίλτρου για να επιλέξουμε ένα "βέλτιστο" υποσύνολο χαρακτηριστικών – γνωρισμάτων που μπορεί να αποτελέσει την είσοδο σε διαφορετικά μοντέλα ταξινόμησης. Στη κατηγορία αυτή ανήκουν αλγόριθμοι που βασίζονται στην αξιοποίηση στατιστικών μέτρων για τον εντοπισμό του βαθμού εξάρτησης μεταξύ χαρακτηριστικών και κλάσεων. Ουσιαστικά, αξιολογούν τη σχετικότητα των χαρακτηριστικών ως προς τις κλάσεις του προβλήματος ερευνώντας μόνο τις ιδιότητες των υπό εξέταση χαρακτηριστικών.

Το κύριο μειονέκτημα των μεθόδων αυτών είναι ότι αγνοούν την αλληλεπίδραση που ενδεχομένως υπάρχει μεταξύ των χαρακτηριστικών που επιλέγονται και του μοντέλου ταξινόμησης που πρόκειται να χρησιμοποιηθεί, καθώς κατά την διαδικασία της επιλογής των χαρακτηριστικών δεν υπάρχει οποιαδήποτε αλληλεπίδραση με το μοντέλο.

Οι μέθοδοι φίλτρου διακρίνονται σε δύο βασικές κατηγορίες, τις μονομεταβλητές (univariate) και τις πολυμεταβλητές (multivariate) μεθόδους [16]. Οι μονομεταβλητές μέθοδοι αξιολογούν μεμονωμένα κάθε χαρακτηριστικό με αποτέλεσμα να επιλέγονται χαρακτηριστικά τα οποία ενδέχεται να είναι ισχυρά συσχετισμένα μεταξύ τους και έτσι ο συνδιασμός τους δεν προσφέρει πολύ περισσότερη πληροφορία για τις κλάσεις από αυτή που θα μπορούσε να προσφέρει κάθε χαρακτηριστικό από μόνο του. Αντίθετα, οι πολυμεταβλητές μέθοδοι αξιολογούν τα χαρακτηριστικά λαμβάνοντας υπόψη τη σχετική τους ανεξαρτησία, προσπαθώντας έτσι να αποφύγουν την επιλογή περιττών χαρακτηριστικών.

Στη συνέχεια παρουσιάζονται βασικές μέθοδοι φίλτρου επιλογής χαρακτηριστικών, κατάλληλες για εφαρμογές Ταξινόμησης (Classification).

- **Pearson Correlation Matrix**

  Ο συντελεστής γραμμικής συσχέτισης Pearson μπορεί να χρησιμοποιηθεί για την απαλοιφή χαρακτηριστικών τα οποία είναι ισχυρά συσχετισμένα μεταξύ τους. Σύμφωνα με τη μέθοδο αυτή, υπολογίζουμε τη συσχέτιση μεταξύ κάθε δυνατού ζεύγους χαρακτηριστικών μέσω του πίνακα συσχέτισης (correlation matrix) και επιλέγουμε εκείνα τα χαρακτηριστικά τα οποία παρουσιάζουν ασθενή συσχέτιση μεταξύ τους. Σημειώνεται ότι η μέθοδος αυτή συνήθως δεν παράγει βέλτιστα αποτελέσματα καθώς δεν συμπεριλαμβάνει την πληροφορία των κλάσεων του προβλήματος. Το γεγονός αυτό προκύπτει από τη φύση της εξόδου των προβλημάτων ταξινόμησης καθώς δεν έχει νόημα να υπολογίσουμε τη γραμμική συσχέτιση μεταξύ ποσοτικών μεταβλητών και ποιοτικών μεταβλητών οι οποίες δεν υπονοούν διάταξη. Παρ' όλα αυτά είναι μια απλή και υπολογιστικά γρήγορη μέθοδος που σε γενικές γραμμές βελτιώνει την επίδοση των μοντέλων.

  Αν θεωρήσουμε ως X, Y δύο διαφορετικά διανύσματα – χαρακτηριστικά, τότε ο συντελεστής συσχέτισης Pearson των διανυσμάτων υπολογίζεται ως εξής :

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{var(X)\, var(Y)}} \tag{3.12}$$

  Όπου $\rho_{XY} \epsilon [-1,1]$.

  Αν $\rho_{XY} = \pm 1$ τότε υπάρχει **τέλεια γραμμική συσχέτιση** μεταξύ των X και Y.
  Αν $-0.25 \leq \rho_{XY} \leq 0.25$ τότε **δεν υπάρχει γραμμική συσχέτιση**. Αυτό όμως, δεν αποκλείει την ύπαρξη άλλου είδους συσχέτισης μεταξύ των δύο διανυσμάτων. Ο βαθμός γραμμικής συσχέτισης καθορίζεται από την απόλυτη τιμή του ρ και όχι από το πρόσημο του.

- **Hypothesis Tests**

  Έστω Y το διάνυσμα στόχος του προβλήματος και $X = (X_1, \dots, X_i, \dots, X_n)$ το σύνολο των χαρακτηριστικών. Όπως ήδη έχουμε αναφέρει τα δεδομένα αποθηκεύονται σε πίνακες που περιέχουν m οντότητες (γραμμές) και κάθε οντότητα περιγράφεται από n χαρακτηριστικά (στήλες). Συνεπώς, κάθε χαρακτηριστικό μπορεί να θεωρηθεί ως ένα διάνυσμα. Για κάθε διάνυσμα - χαρακτηριστικό X_i του συνόλου X αναπτύσσουμε μια στατιστική δοκιμή ελέγχοντας τις εξής υποθέσεις :

$$H_0^i = \{X_i \, είναι \, ασήμαντο \, στη \, πρόβλεψη \, του \, Y\}$$

$$H_1^i = \{X_i \, είναι \, σημαντικό \, στη \, πρόβλεψη \, του \, Y\}$$

  Ο έλεγχος υπόθεσης H_0 αναφέρεται στη βιβλιογραφία ως Null Hypothesis Testing και είναι ένας τύπος εικασίας που χρησιμοποιείται στη στατιστική ο οποίος προτείνει ότι δεν υπάρχει διαφορά μεταξύ ορισμένων χαρακτηριστικών ενός πληθυσμού ή μιας διαδικασίας δημιουργίας δεδομένων.

Το αποτέλεσμα κάθε ελέγχου υπόθεσης H_0^i αναφέρεται ως p-value (p_i) και ποσοτικοποιεί την πιθανότητα το χαρακτηριστικό X_i να είναι ασήμαντο ως προς τον διαχωρισμό των κλάσεων του προβλήματος. Τιμές του $p_i < 0.05$ υποδεικνύουν ότι το χαρακτηριστικό X_i είναι σημαντικό για τη πρόβλεψη του στόχου.

Ανάλογα με τη φύση των διανυσμάτων – χαρακτηριστικών X_i και του στόχου Y υπάρχουν διαφορετικές στατιστικές τεχνικές ελέγχου υποθέσεων. Οι πιο δημοφιλείς τεχνικές είναι οι εξής [13] :

- **Fisher Test** (για δυαδικά X_i και Y)
- **Kolmogorov – Sminrov Test** (για δυαδικά X_i και συνεχή ή δυαδικό Y)
- **Kendal Rank Test** (όταν τα X_i και Y δεν είναι δυαδικά)

Κατά τον έλεγχο υποθέσεων ενδέχεται ένα χαρακτηριστικό X_i να επιλεχθεί λανθασμένα όταν απορρίπτεται ο έλεγχος H_0^i ενώ στη πραγματικότητα ισχύει. Όταν πραγματοποιούνται πολλαπλές συγκρίσεις υποθέσεων και χαρακτηριστικών ταυτόχρονα, τέτοια λάθη τείνουν να συσσωρεύονται [13]. Το αναμενόμενο ποσοστό λανθασμένων απορρίψεων του H_0 μεταξύ όλων των απορρίψεων ονομάζεται Ψευδές Ποσοστό Ανακάλυψης (False Discovery Rate – FDR). Για την αντιμετώπιση του εν λόγω προβλήματος αξιοποιούνται τεχνικές που προσαρμόζουνε τα αποτελέσματα των υποθέσεων H_0^i, αυξάνοντας τις τιμές των p_i με αποτέλεσμα να μειώνεται το αναμενόμενο ποσοστό των λανθασμένων απορρίψεων της υπόθεσης H_0.

- **Analysis of Variance (ANOVA)**

Η ανάλυση διασποράς (ANOVA) αποτελεί μια ευρέως διαδεδομένη μέθοδο ελέγχου στατιστικής σημασίας μεταβλητών. Συνήθως χρησιμοποιείται ως ένα πρώτο στάδιο επιλογής χαρακτηριστικών όταν το πλήθος αυτών είναι ιδιαίτερα μεγάλο.

Έστω $X = (X_1, \dots, X_i, \dots, X_n)$ το σύνολο των χαρακτηριστικών και μ_i η μέση τιμή του διανύσματος – χαρακτηριστικού X_i. Ο έλεγχος που πραγματοποιείται κατά την ανάλυση διασποράς είναι :

$$H_0 = \{\mu_1 = \cdots = \mu_i = \cdots = \mu_n\}$$

$$H_1 = \{Τουλάχιστον\ μια\ εκ\ των\ μέσων\ τιμων\ είναι\ διαφορετική\}$$

Κάθε X_i θεωρείται ως μια ομάδα (group). Για την απόρριψη ή την αποδοχή της υπόθεσης H_0 υπολογίζεται ο στατιστικός δείκτης F που ορίζεται ως εξής :

$$F = \frac{SSB}{SSW} = \frac{Sum\ of\ Squares\ Between\ Group}{Sum\ of\ Squares\ Within\ Group} \tag{3.13}$$

Όπου :

$$SSB = \sum_{i=1}^{n} \frac{m_i(\mu_i - \mu)^2}{n-1} \quad \text{και} \quad SSW = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{(x_{ij} - \mu_i)^2}{N-n}$$

n : ο αριθμός των χαρακτηριστικών
m_i : η διάσταση του X_i
μ : η συνολική μέση τιμή όλων των X_i
N : ο συνολικός αριθμός όλων των παρατηρήσεων όλων των X_i
x_{ij} : η παρατήρηση j του X_i

Ο στατιστικός δείκτης F ακολουθεί την κατανομή F με βαθμούς ελευθερίας $d_1 = n - 1$ και $d_2 = N - n$ υπό την υπόθεση H_0. Μεγάλες τιμές του στατιστικού δείκτη F υπολογίζονται όταν η μεταβλητότητα μεταξύ των ομάδων είναι μεγάλη σε σχέση με τη μεταβλητότητα εντός των ομάδων (δηλαδή των χαρακτηριστικών). Κάτι τέτοιο δύναται να ισχύει μόνο αν οι μέσες τιμές των ομάδων είναι διαφορετικές μεταξύ τους και στη περίπτωση αυτή απορρίπτουμε την υπόθεση H_0.

Στα πλαίσια της επιλογής χαρακτηριστικών θεωρούμε ως εξαρτημένη μεταβλητή το διάνυσμα – στόχο Y και ως ανεξάρτητες μεταβλητές τα διανύσματα – χαρακτηριστικά X_i. Για την αξιολόγηση της σημαντικότητας κάθε χαρακτηριστικού ως προς τον στόχο χρησιμοποιούμε την ανάλυση διασποράς κατά έναν παράγοντα (OneWay ANOVA) όπου σε αυτή τη μέθοδο κάθε στατιστικός έλεγχος πραγματοποιείται μόνο ανάμεσα σε ένα χαρακτηριστικό X_i και το στόχο Y. Για κάθε X_i υπολογίζουμε το στατιστικό δείκτη F_i (ο οποίος ακολουθεί τη κατανομή t^2) και στη συνέχεια κατατάσσουμε τα χαρακτηριστικά ως προς τη σημαντικότητα τους με βάση τα αποτελέσματα των δεικτών. Συνεπώς, σύμφωνα με τη σειρά κατάταξης των χαρακτηριστικών μπορούμε να επιλέξουμε τα k πρώτα από αυτά.

- **Mutual Information (MI)**

Στη θεωρία πιθανοτήτων και στη θεωρία πληροφοριών (information theory) η Αμοιβαία Πληροφορία (Mutual Information – MI) δύο τυχαίων μεταβλητών είναι ένα μέτρο της αμοιβαίας εξάρτησης μεταξύ των δύο μεταβλητών. Πιο συγκεκριμένα, ποσοτικοποιεί το μέγεθος της πληροφορίας (σε μονάδες όπως Shannon Bits (Sh)) που λαμβάνεται για μια τυχαία μεταβλητή παρατηρώντας μια άλλη τυχαία μεταβλητή. Η έννοια της αμοιβαίας πληροφορίας συνδέεται στενά με αυτήν της εντροπίας (entropy), που ποσοτικοποιεί την αναμενόμενη πληροφορία που διατηρείται σε μια τυχαία μεταβλητή. Στη βιβλιογραφία ο όρος Αμοιβαία Πληροφορία αναφέρεται και ως Κέρδος Πληροφορίας (Information Gain).

Έστω Y το διάνυσμα στόχος του προβλήματος και $X = (X_1, \dots, X_i, \dots, X_n)$ το σύνολο των χαρακτηριστικών. Οι τυχαίες μεταβλητές Y και X_i θεωρούνται διακριτές. Η εντροπία της τυχαίας μεταβλητής Y σύμφωνα με τον τύπο του Shannon υπολογίζεται ως εξής [15]:

$$H(Y) = -\sum_{y \in Y} p(y) \log\big(p(y)\big) \tag{3.14}$$

Η σχέση (3.14) αντιπροσωπεύει την αβεβαιότητα ως προς το περιεχόμενο της πληροφορίας στην έξοδο Y [81]. Αν υποθέσουμε ότι παρατηρούμε μια μεταβλητή X_i, τότε η δεσμευμένη εντροπία (conditional entropy) υπολογίζεται ως εξής :

$$H(Y|X_i) = -\sum_{x \in X_i} \sum_{y \in Y} p(x,y)\, log\big(p(y|x)\big) \tag{3.15}$$

Μέσω της σχέσης (3.15) υπονοείται ότι παρατηρώντας μια μεταβλητή X_i, μειώνεται η αβεβαιότητα στην έξοδο Y. Η μείωση της αβεβαιότητας υπολογίζεται ως εξής :

$$I(Y, X_i) = H(Y) - H(Y|X_i) \tag{3.16}$$

Η σχέση (3.16) δίνει την Αμοιβαία Πληροφορία (MI) μεταξύ των μεταβλητών Y και X_i. Στη περίπτωση που οι μεταβλητές είναι ανεξάρτητες τότε το μέτρο MI δίνει ως αποτέλεσμα το μηδέν ενώ σε αντίθετη περίπτωση δίνει ως αποτέλεσμα έναν θετικό αριθμό. Όσο μεγαλύτερο είναι το αποτέλεσμα τόσο μεγαλύτερη η εξάρτηση που υπάρχει μεταξύ των δύο μεταβλητών. Σημειώνεται ότι η εν λόγω εξάρτηση δεν είναι απαραίτητα γραμμική. Επίσης, οι παραπάνω σχέσεις εφαρμόζονται σε συνεχείς μεταβλητές αντικαθιστώντας τα αθροίσματα με ολοκληρώματα.

Σύμφωνα λοιπόν με τη μέθοδο της Αμοιβαίας Πληροφορίας, για κάθε X_i υπολογίζουμε τον στατιστικό δείκτη MI_i από τη σχέση (3.16) και κατατάσσουμε τα χαρακτηριστικά ως προς τη σημαντικότητα τους με βάση τα αποτελέσματα των δεικτών. Στη συνέχεια, σύμφωνα με την εν λόγω σειρά κατάταξης των χαρακτηριστικών μπορούμε να επιλέξουμε τα k πρώτα από αυτά.

3.4.2.2 Μέθοδοι Περιτυλίγματος (Wrapper Methods)

Οι μέθοδοι περιτυλίγματος ενσωματώνουν μοντέλα ταξινόμησης για την αναζήτηση κατάλληλων υποσυνόλων χαρακτηριστικών και χρησιμοποιούν την ακρίβεια ταξινόμησης των μοντέλων ως κριτήριο αξιολόγησης. Οι εν λόγω μέθοδοι δύνανται να αξιοποιήσουν οποιοδήποτε μοντέλο και αλγόριθμο ταξινόμησης καθώς δεν εξαρτώνται από τον τρόπο λειτουργίας των μοντέλων. Η επίδοση των μοντέλων είναι το κριτήριο αξιολόγησης των υπό εξέταση χαρακτηριστικών τα οποία αξιολογούνται συνήθως σε ομάδες και όχι ανεξάρτητα.

Ο τρόπος λειτουργίας των μεθόδων περιτυλίγματος συνοψίζεται ως εξής :

i. Τα διαθέσιμα δεδομένα (παραδείγματα) χωρίζονται σε δύο σύνολα, το σύνολο εκπαίδευσης (training set) και το σύνολο επικύρωσης (validation set).

ii. Επιλέγεται ένα υποψήφιο υποσύνολο χαρακτηριστικών προς αξιολόγηση και διαγράφονται από τα αντίγραφα των συνόλων εκπαίδευσης και επικύρωσης όσα χαρακτηριστικά δεν ανήκουν στο επιλεγμένο υποσύνολο.

iii. Το μοντέλο ταξινόμησης εκπαιδεύεται με βάση το τροποποιημένο σύνολο εκπαίδευσης και βάση αυτής της εκπαίδευσης ταξινομεί τα δεδομένα του τροποποιημένου συνόλου επικύρωσης.

iv. Το μοντέλο αξιολογείται ως προς την ταξινόμηση που υλοποίησε με κατάλληλες μετρικές (συνήθως χρησιμοποιείται το μέτρο accuracy). Το αποτέλεσμα της αξιολόγησης αντιστοιχεί στην καταλληλότητα του επιλεγμένου υποσυνόλου χαρακτηριστικών.

v. Η διαδικασία επαναλαμβάνεται από το δεύτερο στάδιο εώς ότου επιτευχθεί το κριτήριο τερματισμού και τότε επιστρέφεται το καταλληλότερο υποσύνολο χαρακτηριστικών.

Το κύριο μειονέκτημα αυτών των μεθόδων είναι το αυξημένο υπολογιστικό κόστος καθώς η αξιολόγηση των χαρακτηριστικών με βάση την επίδοση του ταξινομητή απαιτεί την εκπαίδευση και την αξιολόγηση του μοντέλου για κάθε δυνατό υποσύνολο χαρακτηριστικών [73]. Για την αντιμετώπιση του εν λόγω προβλήματος, καθώς ο αριθμός των δυνατών υποσυνόλων αυξάνεται εκθετικά με το πλήθος των διαθέσιμων χαρακτηριστικών, χρησιμοποιούνται ευρετικοί (heuristics) αλγόριθμοι αναζήτησης.

Επίσης, σε περίπτωση που το αρχικό σύνολο εκπαίδευσης είναι μικρό, δεν υπάρχει δυνατότητα να σχηματιστούν επαρκώς μεγάλα σύνολα εκπαίδευσης και επικύρωσης. Το ενδεχόμενο αυτό αποτελεί σημαντικό πρόβλημα καθώς στη περίπτωση που το σύνολο εκπαίδευσης είναι μικρό το μοντέλο υπερμοντελοποιείται, ενώ όταν το σύνολο επικύρωσης είναι μικρό δεν μπορεί να εκτιμηθεί με αξιοπιστία η επίδοση του μοντέλου [73]. Για την αντιμετώπιση αυτού του προβλήματος αξιοποιούνται τεχνικές επαναληπτικής δειγματοληψίας όπως για παράδειγμα οι μέθοδοι Bootstraping, k-fold cross-validation [79] και stratified k-fold cross-validation σε περίπτωση που υπάρχει επίσης ανισορροπία κλάσεων [80]. Ωστόσο, η χρήση οποιασδήποτε τεχνικής επαναληπτικής δειγματοληψίας αυξάνει ακόμα περισσότερο το υπολογιστικό κόστος των μεθόδων περιτυλίγματος.

Αξίζει να σημειωθεί το γεγονός ότι τα χαρακτηριστικά που επιλέγονται ως κατάλληλα από τις μεθόδους περιτυλίγματος συνήθως βελτιώνουν την επίδοση μόνο των αντίστοιχων μοντέλων που ενσωματώθηκαν σε αυτές. Το γεγονός αυτό παρατηρείται εξαιτίας της μεροληψίας (bias) των μοντέλων καθώς κάθε μοντέλο προσδιορίζεται από διαφορετικούς παραμέτρους και απεικονίζει με διαφορετικό τρόπο τα δεδομένα εισόδου ως έξοδο. Συνεπώς, όσον αφορά τις εν λόγω μεθόδους, το καλύτερο υποσύνολο χαρακτηριστικών για ένα μοντέλο ταξινόμησης δεν αποτελεί απαραίτητα το καλύτερο υποσύνολο και για μοντέλα διαφορετικού τύπου.

Σύμφωνα με όσα έχουμε αναλύσει, οι μέθοδοι περιτυλίγματος ενδεχομένως εντοπίζουν αλληλεπιδράσεις μεταξύ των χαρακτηριστικών καθώς τα διαθέσιμα χαρακτηριστικά αξιολογούνται ως μέλη συνόλων. Η ανακάλυψη όμως των χαρακτηριστικών που αλληλεπιδρούν εξαρτάται από το αν θα τύχει να βρεθούν στο ίδιο υποσύνολο αξιολόγησης, γεγονός που καθορίζεται από την ευρετική μέθοδο αναζήτησης που χρησιμοποιείται. Στη συνέχεια θα αναλύσουμε δύο βασικούς ευρετικούς αλγορίθμους αναζήτησης υποσυνόλων χαρακτηριστικών που αξιοποιούνται από τις μεθόδους περιτυλίγματος [82].

1. **<u>Βηματική Πρόσθια Επιλογή (Stepwise Forward Selection)</u>**

Έστω S το σύνολο που περιέχει όλα τα διαθέσιμα χαρακτηριστικά με $|S| = n$. Ο αλγόριθμος αυτός αρχικά ορίζει το σύνολο των επιλεγμένων χαρακτηριστικών να είναι το κενό σύνολο $\Phi_{\emptyset}$ και στη συνέχεια (πρώτο βήμα) αναζητά το πρώτο χαρακτηριστικό έστω X_a που μεγιστοποιεί το κριτήριο επίδοσης, δηλαδή την ακρίβεια ταξινόμησης. Το χαρακτηριστικό X_a προστίθεται στο σύνολο $\Phi_{\emptyset}$ και προκύπτει ένα νέο σύνολο Φ_1. Η διαδικασία συνεχίζεται για τον εντοπισμό του δεύτερου χαρακτηριστικού έστω X_b από το σύνολο $\{S - \Phi_1\}$ το οποίο προστίθεται στο σύνολο Φ_1 και προκύπτει ένα νέο σύνολο Φ_2. Στο επόμενο βήμα εντοπίζεται το τρίτο χαρακτηριστικό έστω X_c από το σύνολο $\{S - \Phi_2\}$ και ούτω καθεξής. Σε κάποιο τυχαίο βήμα k έχει διαμορφωθεί ένα σύνολο $\Phi_k \subseteq S$ που περιέχει ως μέλη k χαρακτηριστικά. Συνεπώς, στο βήμα $k+1$ το σύνολο χαρακτηριστικών Φ_{k+1} διαμορφώνεται ως εξής :

$$\Phi_{k+1} = \Phi_k \cup \{X_i \epsilon \{S - \Phi_k\} \,|X_i = argmax\ C_A(\Phi_k \cup X_i)\}$$

Όπου C_A η ακρίβεια ταξινόμησης (Classification Accuracy).

Ο αλγόριθμος συνεχίζει την εκτέλεση βημάτων εώς ότου επιτευχθεί το κριτήριο τερματισμού. Ως κριτήριο τερματισμού ορίζεται είτε η κατάσταση όπου σε κάποιο επαναληπτικό βήμα δεν παρουσιαστεί αύξηση της ακρίβειας ταξινόμησης πάνω από ένα ορισμένο κατώφλι, είτε όταν όλα τα διαθέσιμα χαρακτηριστικά του συνόλου S έχουν επιλεχθεί. Είναι σημαντικό να οριστεί ένα κατάλληλο κατώφλι καθώς ένα ιδιαίτερα μεγάλο (αυστηρό) κατώφλι μπορεί να οδηγήσει τον αλγόριθμο σε πρόωρο σταμάτημα.

2. **<u>Βηματική Οπίσθια Εξάλειψη (Stepwise Backward Elimination)</u>**

Ο αλγόριθμος αυτός αρχικά ορίζει το σύνολο των χαρακτηριστικών Φ_n να είναι ίσο με το σύνολο S όλων των διαθέσιμων χαρακτηριστικών, όπου $|S| = n$. Σε κάθε βήμα αφαιρεί εκείνο το χαρακτηριστικό X_i του οποίου η απαλοιφή μεγιστοποιεί το κριτήριο επίδοσης, δηλαδή την ακρίβεια ταξινόμησης. Στο πρώτο βήμα απαλείφεται το χαρακτηριστικό έστω X_a από το σύνολο Φ_n και προκύπτει το σύνολο Φ_{n-1}, στο δεύτερο βήμα απαλείφεται το χαρακτηριστικό έστω X_b από το σύνολο Φ_{n-1} και προκύπτει το σύνολο Φ_{n-2} και ούτω καθεξής. Σε κάποιο τυχαίο βήμα k έχει διαμορφωθεί ένα σύνολο $\Phi_{n-k} \subseteq S$ που περιέχει ως μέλη $n-k$ χαρακτηριστικά. Συνεπώς, στο βήμα $k+1$ το σύνολο χαρακτηριστικών $\Phi_{n-(k+1)}$ διαμορφώνεται ως εξής :

$$\Phi_{n-(k+1)} = \Phi_{n-k} - \{X_i \epsilon \Phi_{n-k} \mid X_i = argmax\ C_A(\Phi_{n-k} - X_i)\}$$

όπου C_A η ακρίβεια ταξινόμησης (Classification Accuracy).

Ο αλγόριθμος συνεχίζει την εκτέλεση βημάτων εώς ότου επιτευχθεί το κριτήριο τερματισμού. Ως κριτήριο τερματισμού ορίζεται είτε η κατάσταση όπου σε κάποιο επαναληπτικό βήμα δεν παρουσιαστεί αύξηση της ακρίβειας ταξινόμησης πάνω από

ένα ορισμένο κατώφλι, είτε όταν έχουν εκτελεστεί $n-1$ βήματα όπου θα έχουμε καταλήξει στο σύνολο Φ_1 με $|\Phi_1| = 1$, δηλαδή όταν έχει επιλεχθεί τελικά ένα μοναδικό κατάλληλο χαρακτηριστικό. Όπως και στον προηγούμενο αλγόριθμο, είναι σημαντικό να οριστεί ένα κατάλληλο κατώφλι, καθώς ένα ιδιαίτερα μεγάλο (αυστηρό) κατώφλι μπορεί να οδηγήσει τον αλγόριθμο σε πρόωρο σταμάτημα.

Ο αλγόριθμος αυτός αναφέρεται στη βιβλιογραφία και ως **Αναδρομική Εξάλειψη Χαρακτηριστικών (Recursvie Feature Elimination – RFE)**.

3.4.2.3 Ενσωματωμένες Μέθοδοι (Embedded Methods)

Σε αντίθεση με τις μεθόδους περιτυλίγματος που χρησιμοποιούν την ακρίβεια ταξινόμησης των μοντέλων για την αξιολόγηση των χαρακτηριστικών, οι ενσωματωμένες μέθοδοι επιλέγουν χαρακτηριστικά με βάση τις μεθόδους αξιολόγησης χαρακτηριστικών που ενσωματώνουν ορισμένα μοντέλα μηχανικής μάθησης. Για παράδειγμα, τα Δένδρα Απόφασης επιλέγουν χαρακτηριστικά με βάση το κριτήριο διάσπασης που ορίζεται σύμφωνα με αντικειμενικές στατιστικές συναρτήσεις όπως αυτή της σχέσης (2.16). Συνεπώς, οι ενσωματωμένες μέθοδοι αποτελούν έναν συνδιασμό των μεθόδων φίλτρου και περιτυλίγματος.

Ο τρόπος λειτουργίας των ενσωματωμένων μεθόδων συνοψίζεται ως εξής :

i. Εκπαιδεύεται κάποιο μοντέλο ταξινόμησης το οποίο ενσωματώνει στατιστικές μεθόδους που υπολογίζουν τη σημαντικότητα των χαρακτηριστικών (feature importance).

ii. Κατά τη διαδικασία εκπαίδευσης υπολογίζεται η σημαντικότητα όλων των χαρακτηριστικών σύμφωνα με το βαθμό που δύνανται να βελτιστοποιήσουν τα αποτελέσματα της στατιστικής αντικειμενικής συνάρτησης του μοντέλου.

iii. Απαλείφονται τα ασήμαντα χαρακτηριστικά.

Υπάρχουν δύο βασικές κατηγορίες ενσωματωμένων μεθόδων επιλογής χαρακτηριστικών :

1. **Μέθοδοι Ομαλοποίησης (Regularization Methods)**
   (Lasso Regression, Ridge Regression, Elastic Nets)

2. **Μέθοδοι Βασισμένοι σε Δένδρα Απόφασης (Tree Based Methods)**
   (Decision Trees, Random Forests)

Οι ενσωματωμένες μέθοδοι πλεονεκτούν σε σχέση με τις μεθόδους περιτυλίγματος καθώς δεν χαρακτηρίζονται από ιδιαίτερα μεγάλο υπολογιστικό κόστος και δεν απαιτείται ο σχηματισμός συνόλων επικύρωσης δεδομένου ότι τα χαρακτηριστικά επιλέγονται κατά τη διαδικασία της εκπαίδευσης των μοντέλων. Επίσης, οι εν λόγω μέθοδοι παρουσιάζουν μεγαλύτερη ακρίβεια από τις μεθόδους φίλτρου υπό την προϋπόθεση ότι τα επιλεγμένα χαρακτηριστικά θα αποτελέσουν την είσοδο σε τύπους μοντέλων που αξιοποιήθηκαν κατά τη διαδικασία επιλογής τους. Σε γενικές γραμμές οι ενσωματωμένες μέθοδοι βελτιώνουν τη γενίκευση των μοντέλων και δρουν αμυντικά ως προς το φαινόμενο της υπερεκπαίδευσης.

Κεφάλαιο 4 : Εφαρμογές Μη Επιβλεπόμενης Μάθησης

Στο κεφάλαιο αυτό παρουσιάζουμε αναλυτικά τη διαδικασία και τα αποτελέσματα των εφαρμογών Ομαδοποίησης και Εξόρυξης Δεδομένων που υλοποιήσαμε σε ενεργειακά δεδομένα και συγκεκριμένα σε χρονοσειρές φορτίου ηλεκτρικής ενέργειας. Στόχος των εν λόγω εφαρμογών είναι η εξαγωγή χαρακτηριστικών ημερήσιων καμπυλών φορτίου που αντιπροσωπεύουν την ενεργειακή συμπεριφορά ευρωπαϊκών χωρών. Οι χαρακτηριστικές αυτές καμπύλες αναφέρονται στη βιβλιογραφία ως "Προφίλ Φορτίου" (Load Profiles) καθώς και ως "Τυπικές Χρονολογικές Καμπύλες Φορτίου" (Typical Chronological Load Curves) [4, 9, 29, 31, 33]. Η γνώση και η αναλυτική περιγραφή της ενεργειακής συμπεριφοράς των καταναλωτών αποτελεί χρήσιμη πληροφορία στα πλαίσια της αποδοτικής διαχείρισης και του σχεδιασμού των συστημάτων ηλεκτρικής ενέργειας και μπορεί να αξιοποιηθεί σε εφαρμογές πρόβλεψης και κατηγοριοποίησης.

4.1 Περιγραφή Δεδομένων Ανάλυσης

Τα δεδομένα ανάλυσης αφορούν το "Συνολικό Πραγματικό Φορτίο" (Actual Total Load) ευρωπαϊκών χωρών και είναι διαθέσιμα μέσω της πλατφόρμας διαφάνειας (TP) του ευρωπαϊκού διαχειριστή συστημάτων μεταφοράς ηλεκτρικής ενέργειας ENTSO-E. Ο ορισμός του Συνολικού Πραγματικού Φορτίου (Σ.Π.Φ) κάθε χώρας, είτε κάθε ζώνης προσφοράς (bidding zone), είτε κάθε περιοχής ελέγχου (control area) έχει ως εξής [18] :

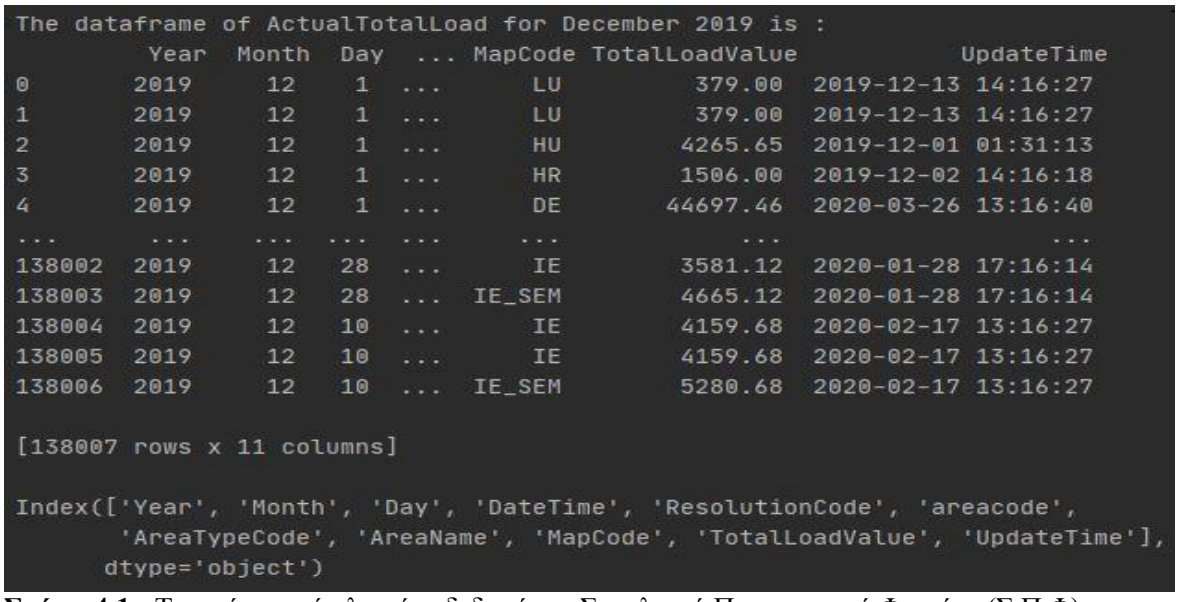
$$\boldsymbol{Actual\ Total\ Load} = \boldsymbol{P_{G_{TSO}}} + \boldsymbol{P_{G_{DSO}}} - \boldsymbol{Balance} - \boldsymbol{P_S} \qquad (4.1)$$

Όπου :

- $\boldsymbol{P_{G_{TSO}}}$ **:** Η ισχύς που παράγεται από σταθμούς στα δίκτυα μεταφοράς
- $\boldsymbol{P_{G_{DSO}}}$ **:** Η ισχύς που παράγεται από σταθμούς στα δίκτυα διανομής
- $\boldsymbol{Balance}$ **:** Εξαγωγές μείον Εισαγωγές ισχύος στις διασυνδέσεις γειτονικών ζωνών προσφοράς
- $\boldsymbol{P_S}$ **:** Η ισχύς που απορροφάται στα συστήματα αποθήκευσης ενέργειας

Οι μονάδες μέτρησης των παραπάνω μεγεθών είναι σε MegaWatt (MW) και η δήλωση του μέσου Σ.Π.Φ κάθε οντότητας (χώρας είτε ζώνης προσφοράς είτε περιοχής ελέγχου) γίνεται συνήθως ανά ώρα ή ανά τριάντα λεπτά, ανάλογα με το σύστημα που ακολουθείται.

Στη πλατφόρμα διαφάνειας τα δεδομένα για όλες τις οντότητες αποθηκεύονται σε μηνιαία αρχεία τιμών διαχωρισμένων με "tab" ("\t"), δηλαδή σε αρχεία TSV, υπό το πρότυπο κωδικοποίησης "UTF-16" και ανανεώνονται σε ημερήσια βάση. Οδηγίες για την καταφόρτωση των εν λόγω δεδομένων μπορούν να βρεθούν στην ιστοσελίδα της πλατφόρμας διαφάνειας. Στη συνέχεια παρουσιάζουμε τη τυπική μορφή ενός αρχείου TSV από τη πλατφόρμα διαφάνειας, με δεδομένα που αφορούν το Σ.Π.Φ. Κάθε τέτοιο αρχείο αναπαρίσταται ως ένα πλαίσιο δεδομένων (dataframe) με ετικέτες (labels) και πεδία τιμών.



```
The dataframe of ActualTotalLoad for December 2019 is :
         Year  Month  Day  ... MapCode TotalLoadValue          UpdateTime
0        2019     12    1  ...      LU         379.00  2019-12-13 14:16:27
1        2019     12    1  ...      LU         379.00  2019-12-13 14:16:27
2        2019     12    1  ...      HU        4265.65  2019-12-01 01:31:13
3        2019     12    1  ...      HR        1506.00  2019-12-02 14:16:18
4        2019     12    1  ...      DE       44697.46  2020-03-26 13:16:40
...       ...    ...  ...  ...     ...            ...                  ...
138002   2019     12   28  ...      IE        3581.12  2020-01-28 17:16:14
138003   2019     12   28  ...  IE_SEM        4665.12  2020-01-28 17:16:14
138004   2019     12   10  ...      IE        4159.68  2020-02-17 13:16:27
138005   2019     12   10  ...      IE        4159.68  2020-02-17 13:16:27
138006   2019     12   10  ...  IE_SEM        5280.68  2020-02-17 13:16:27

[138007 rows x 11 columns]

Index(['Year', 'Month', 'Day', 'DateTime', 'ResolutionCode', 'areacode',
       'AreaTypeCode', 'AreaName', 'MapCode', 'TotalLoadValue', 'UpdateTime'],
      dtype='object')
```

Σχήμα 4.1 : Τυπική μορφή πλαισίου δεδομένων Συνολικού Πραγματικού Φορτίου (Σ.Π.Φ).

Με βάση το σχήμα 4.1 παρατηρούμε ότι το πλαίσιο αποτελείται από έντεκα στήλες που είναι οι ετικέτες - δείκτες των τύπων των δεδομένων. Κάθε αντικείμενο, δηλαδή κάθε δήλωση Σ.Π.Φ αποθηκεύεται σε μια σειρά του πίνακα και προσδιορίζεται επαρκώς μέσω αυτών των δεικτών. Συνεπώς, κάθε αντικείμενο αποτελείται από έντεκα πεδία τιμών τα οποία αναλύονται στη συνέχεια.

- **Year, Month, Day :** Δηλώνουν τη χρονολογική ταυτότητα του αντικειμένου.
- **DateTime :** Δηλώνει τη χρονική ταυτότητα του αντικειμένου. Οι υποβολές Σ.Π.Φ πραγματοποιούνται υπό σταθερή συχνότητα η οποία είναι προκαθορισμένη για κάθε οντότητα.
- **ResolutionCode :** Δηλώνει τη συχνότητα υποβολής του Σ.Π.Φ της αντίστοιχης οντότητας σύμφωνα με το πρότυπο κωδικοποίησης ISO 8601. Για παράδειγμα, ο κωδικός "PT60M" αναφέρεται σε ωριαία συχνότητα υποβολής.
- **areacode :** Αποτελεί σύστημα κωδικοποίησης για την αναγνώριση των οντοτήτων σύμφωνα με το πρότυπο Energy Identification Code (EIC) το οποίο έχει αναπτυχθεί από τον ENTSO-E.
- **AreaTypeCode :** Δηλώνει τον τύπο της οντότητας. Υπάρχουν τέσσερεις διαφορετικοί τύποι οντοτήτων (CTY, BZN, CTA, MBA).
- **AreaName :** Αναφέρεται στο όνομα των οντοτήτων.
- **MapCode :** Κωδικοποίηση για την αναγνώριση της γεωγραφικής περιοχής στην οποία ανήκουν οι οντότητες.
- **TotalLoadValue :** Η τιμή του μέσου Σ.Π.Φ.
- **UpdateTime :** Η χρονική στιγμή διόρθωσης της τιμής του μέσου Σ.Π.Φ.

Οι πλέον χρήσιμες ετικέτες (στήλες) για την ανάπτυξη των εφαρμογών της διπλωματικής εργασίας είναι οι εξής :

i. DateTime
ii. ResolutionCode
iii. AreaTypeCode
iv. TotalLoadValue
v. MapCode

Όπως είναι φανερό από το σχήμα 4.1, οι τιμές Σ.Π.Φ που βρίσκονται στα πεδία της στήλης "TotalLoadValue" δεν είναι διατεταγμένες χρονικά. Συνεπώς, απαιτείται η φυσική χρονική τους διάταξη για τη κατασκευή των χρονοσειρών φορτίου οι οποίες θα αποτελέσουν τα δεδομένα ανάλυσης. Η στήλη "DateTime" περιέχει την απαραίτητη πληροφορία που θα μας βοηθήσει να υλοποιήσουμε τη σωστή χρονική διάταξη των τιμών Σ.Π.Φ και κατά συνέπεια να συνθέσουμε τις αντίστοιχες χρονοσειρές φορτίου. Η μορφή ενός οποιουδήποτε πεδίου "DateTime" είναι :

$$\{Χρόνος - Μήνας - Μερα \quad Ώρα : Λεπτό : Δευτερόλεπτο\}$$

Η χρονική διάταξη των τιμών Σ.Π.Φ απαιτείται να πραγματοποιηθεί ξεχωριστά για κάθε χώρα ώστε να προκύψουν οι αντίστοιχες χρονοσειρές. Μέσω της στήλης "AreaTypeCode" μπορούμε να επιλέξουμε τιμές Σ.Π.Φ που αφορούν οντότητες οι οποίες χαρακτηρίζονται ως χώρες, με τον κωδικό "CTY". Επίσης, μέσω της στήλης "MapCode" μπορούμε να ομαδοποιήσουμε τις τιμές Σ.Π.Φ κάθε χώρας η οποία προσδιορίζεται μέσω του αντίστοιχου της κωδικού. Για παράδειγμα η Ελλάδα έχει ως κωδικό MapCode το "GR" ενώ η Ιταλία το "IT". Η στήλη "ResolutionCode" περιέχει την απαραίτητη πληροφορία ώστε να επιλέξουμε οντότητες οι οποίες χαρακτηρίζονται από την ίδια συχνότητα υποβολής των μέσων τιμών Σ.Π.Φ. Η συχνότητα υποβολής ή αλλιώς συχνότητα δήλωσης καθορίζει την διάσταση των χρονοσειρών δεδομένου ενός χρονικού παραθύρου (time window). Για παράδειγμα, για ένα χρονικό παράθυρο ενός έτους, μια χώρα με ωριαία συχνότητα υποβολής παράγει μια χρονοσειρά φορτίου διάστασης $d_1 = 8760$, ενώ αντίστοιχα μια άλλη χώρα με συχνότητα υποβολής τριάντα λεπτών παράγει μια χρονοσειρά φορτίου διπλάσιας διάστασης. Κάθε χρονοσειρά μεγαλύτερης διάστασης δύναται να συμπιεστεί σε μια χρονοσειρά μικρότερης διάστασης μέσω του υπολογισμού των μέσων όρων διαδοχικών παρατηρήσεων. Ωστόσο, για λόγους διευκόλυνσης, στη παρούσα διπλωματική εργασία επιλέξαμε να επεξεργαστούμε μόνο χρονοσειρές φορτίου χωρών που διέπονται από ωριαία συχνότητα υποβολής.

Η πλατφόρμα διαφάνειας παρέχει δεδομένα από το Δεκέμβρη του 2014 και έπειτα. Συνεπώς, είχαμε στις διάθεση μας δεδομένα από το 2015 μέχρι και τις αρχές του 2021 κατά τη περίοδο εκπόνησης της παρούσας διπλωματικής. Αξίζει να αναφερθεί ότι κάποιες ευρωπαϊκές χώρες όπως για παράδειγμα η Βοσνία – Ερζεγοβίνη και η Ουκρανία, ξεκίνησαν την υποβολή τιμών του μέσου Σ.Π.Φ τους σχετικά πρόσφατα. Σύγκεκριμένα, η Βοσνία Ερζεγοβίνη παρουσιάζεται για πρώτη φορά στη πλατφόρμα διαφάνειας το 2017 ενώ η Ουκρανία παρουσιάζεται για πρώτη φορά το 2018.

Σύμφωνα με τις παραπάνω παραδοχές, καταλήγουμε να έχουμε στη διάθεση μας δεδομένα Σ.Π.Φ από είκοσι πέντε (25) ευρωπαϊκές χώρες, εκ των οποίων οι είκοσι τρεις (23) είναι διαθέσιμες στο χρονικό παράθυρο "2015" εώς "2020". Στη συνέχεια παρουσιάζουμε τις είκοσι πέντε αυτές χώρες μαζί με τον αντίστοιχο τους "MapCode" κωδικό. Ο αστερίσκος "*" στο όνομα της χώρας υπονοεί την μερική διαθεσιμότητα των αντίστοιχων δεδομένων όσον αφορά το συνολικό χρονικό παράθυρο.

Πίνακας 4.1 : Ευρωπαϊκές χώρες που εμπίπτουν στο σύνολο δεδομένων ανάλυσης.

Χώρα	**MapCode**
***Βοσνία - Ερζεγοβίνη*	BA
Βουλγαρία	BG
Ελβετία	CH
Τσεχία	CZ
Δανία	DK
Εσθονία	EE
Ισπανία	ES
Φινλανδία	FI
Γαλλία	FR
Ελλάδα	GR
Κροατία	HR
Ιταλία	IT
Λιθουανία	LT
Λετονία	LV
Μαυροβούνιο	ME
Βόρεια Μακεδονία	MK
Νορβηγία	NO
Πολωνία	PL
Πορτογαλία	PT
Ρουμανία	RO
Σερβία	RS
Σουηδία	SE
Σλοβενία	SI
Σλοβακία	SK
**Ουκρανία*	UA

4.2 Προετοιμασία Δεδομένων για Εφαρμογές Ομαδοποίησης

Στο κεφάλαιο αυτό παρέχουμε την αναλυτική περιγραφή της διαδικασίας που ακολουθήσαμε για τη προετοιμασία των ανεπεξέργαστων δεδομένων. Στόχος της εν λόγω διαδικασίας είναι η σύνθεση των χρονοσειρών φορτίου ηλεκτρικής ενέργειας των ευρωπαϊκών χωρών που εμπίπτουν στο σύνολο δεδομένων ανάλυσης. Η προετοιμασία των δεδομένων πραγματοποιήθηκε σε δύο στάδια. Στο πρώτο στάδιο εφαρμόσαμε τεχνικές για την σύνθεση ετήσιων χρονοσειρών φορτίου ηλεκτρικής ενέργειας των ευρωπαϊκών χωρών και δημιουργήσαμε μια βάση δεδομένων η οποία αποτελείται από έξι πλαίσια δεδομένων τα οποία αφορούν τα έτη 2015 εώς και το 2020. Στο δεύτερο στάδιο εφαρμόσαμε τεχνικές για την διάσπαση των ετήσιων χρονοσειρών φορτίου σε ημερήσιες χρονοσειρές φορτίου και δημιουργήσαμε μια δεύτερη βάση δεδομένων που αποτελείται επίσης από έξι πλαίσια δεδομένων τα οποία αντιστοιχούν στα προαναφερθέντα έτη ανάλυσης και περιέχουν τις εν λόγω ημερήσιες χρονοσειρές.

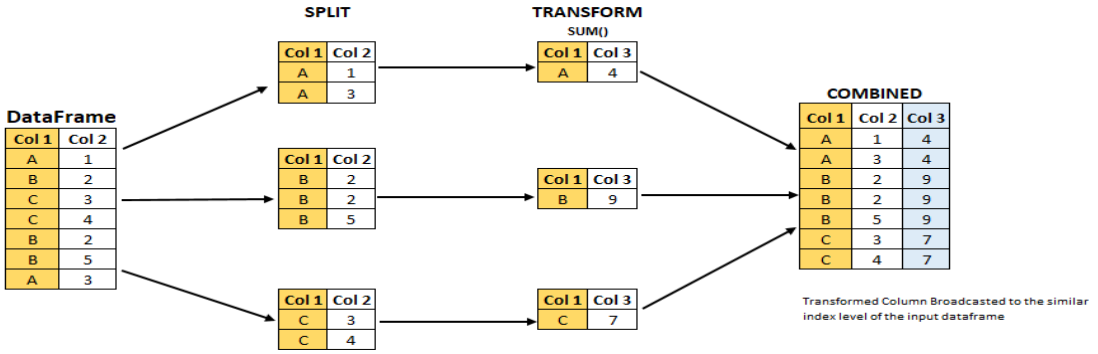
Σχήμα 4.2 : Διάγραμμα ροής προετοιμασίας δεδομένων.

Η αναλυτική περιγραφή της διαδικασίας θα πραγματοποιηθεί μόνο για το σενάριο του 2019, καθώς η περιγραφή για όλα τα χρονικά σενάρια κρίνεται καταχρηστική, δεδομένου ότι η διαδικασία που ακολουθείται είναι ανάλογη για όλα τα έτη. Σημειώνεται ότι κάθε ετήσια χρονοσειρά αναφέρεται στο χρονικό διάστημα που ορίζεται από τη 1η Μαρτιου του εν λόγω έτους εως το τέλος Φεβρουαρίου του επόμενου έτους. Για παράδειγμα όταν αναφερόμαστε στο έτος 2018, στη πραγματικότητα αναφερόμαστε στο διάστημα "01/03/2018" εώς "28/02/2019". Η θεώρηση αυτή κρίνεται βολική σε περίπτωση που θέλουμε να αναλύσουμε το πρόβλημα εποχικά, δηλαδή συναρτήσει των συνθηκών φόρτισης. Πιο συγκεκριμένα :

- **Άνοιξη** *(Μάρτιος, Απρίλιος Μάιος)*
- **Καλοκαίρι** *(Ιούνιος, Ιούλιος, Αύγουστος)*
- **Φθινόπωρο** *(Σεπτέμβριος, Οκτώβριος, Νοέμβριος)*
- **Χειμώνας** *(Δεκέμβριος, Ιανουάριος, Φεβρουάριος)*

Σημειώνεται επίσης, ότι η παρπάνω παραδοχή δεν ισχύει για τις χρονοσειρές του 2015, καθώς αυτές ορίζονται στο χρονικό διάστημα "01/01/2015" εώς "31/12/2015". Συνεπώς, από τις βάσεις δεδομένων που δημιουργήσαμε απουσιάζουν τα δεδομένα μεταξύ "01/01/2016" και "29/02/2016".

Για την υλοποίηση του πρώτου και δεύτερου σταδίου προεπεξεργασίας εφαρμόσαμε τη μέθοδο "Split - Apply - Combine". Το στάδιο "Split" αναφέρεται στη δημιουργία ομάδων δεδομένων, όπου στη δική μας περίπτωση κάθε ομάδα θα αντιστοιχεί σε μία συγκεκριμένη χώρα. Το στάδιο "Apply" αναφέρεται σε συναρτήσεις που επενεργούν στις ομάδες δεδομένων και το στάδιο "Combine" αναφέρεται στην συνάθροιση των τροποποιημένων ομάδων δεδομένων σε ένα ενιαίο πλαίσιο δεδομένων.



Σχήμα 4.3 : Μέθοδος προεπεξεργασίας δεδομένων "Split-Apply-Combine" [*medium.com*].

.

Στη συνέχεια παρουσιάζουμε αναλυτικά τη μεθοδολογία προετοιμασίας των δεδομένων για τη περίοδο "01/03/2019" εώς "29/02/2020" στην οποία αναφερόμαστε ως έτος "2019".

1. **<u>1ο Στάδιο Προεπεξεργασίας</u> :**

   Στο στάδιο αυτό εισάγουμε την βιβλιοθήκη "pandas" με τη παρακάτω εντολή :

   ```
   import pandas as pd
   ```

   Στη συνέχεια διαβάζουμε τα δώδεκα αρχεία tsv που αντιστοιχούν στους μήνες που ορίζουν το χρονικό παράθυρο "01/03/2019" εώς "29/02/2020". Σύμφωνα με τον ENTSO-E κάθε μηνιαίο αρχειο Σ.Π.Φ ακολουθεί το παρακάτω πρότυπο όσον αφορά την ονομασία του :

   $$Έτος\_Μήνας\_ActualTotalLoad.csv$$

   <u>Μέσω της βιβλιοθήκης "pandas" διαβάζουμε ένα μηνιαίο αρχείο και το αποθηκεύουμε σε ένα πλαίσιο δεδομένων (dataframe) με τη παρακάτω εντολή</u> :

   ```
   df1 = pd.read_csv("2019_3_ActualTotalLoad.csv", sep="\t", encoding="utf_16")
   ```

   Αντίστοιχα διαβάζουμε τα υπόλοιπα έντεκα μηνιαία αρχεία.

   <u>Για παράδειγμα, το τελευταίο μηνιαίο αρχείο</u> :

   ```
   df12 = pd.read_csv("2020_2_ActualTotalLoad.csv", sep="\t", encoding="utf_16")
   ```

Στη συνέχεια "φιλτράρουμε" κάθε πλαίσιο δεδομένων df ώστε να εξάγουμε σε ένα πλαίσιο df_f μόνο τις σειρές του πίνακα df που αντιστοιχούν σε οντότητες "CTY" (δηλαδή σε χώρες) με ωριαία συχνότητα δειγματοληψίας ("PT60M"). Έπειτα, εκτελούμε διαγνωστικό έλεγχο για να βεβαιωθούμε ότι η διαδικασία "φιλτραρίσματος" εκτελέστηκε με επιτυχία.

Σε επόμενο στάδιο, για κάθε πλαίσιο δεδομένων df_f εξαλείφουμε τις πλεονάζουσες στήλες και κρατάμε μόνο τις εξής :

$$\{\text{"DateTime"}, \text{"}TotalLoadValue\text{"}, \text{"}MapCode\text{"}\}$$

σε ένα νέο πλαίσιο δεδομένων df_r.

Έπειτα βρίσκουμε το σύνολο των κοινών κωδικών "MapCode" σε όλα τα πλαίσια δεδομένων df_r και τους αποθηκεύουμε σε μια λίστα "common_mapcodes". Το βήμα αυτό εκτελείται καθώς έχει παρατηρηθεί ότι υπάρχει περίπτωση μια χώρα να απουσιάζει τελείως από κάποιο μηνιαίο αρχείο Σ.Π.Φ.

Στο σημείο αυτό ξεκινάμε την εφαρμογή τεχνικών "Split-Apply-Combine". Αρχικά, για κάθε πλαίσιο δεδομένων df_r εκτελούμε τη μέθοδο "Split" όπου ομαδοποιούμε τα δεδομένα κάθε χώρας με βάση τον αντίστοιχο αναγνωριστικό της κωδικό "MapCode" μέσω της εντολής $groupby$. Συνεπώς, δημιουργούμε πλαίσια δεδομένων g τα οποία έχουν δεικτοποιηθεί (indexed) με τους κωδικούς των χωρών που εμπίπτουν στο σύνολο ανάλυσης.

Στη συνέχεια, μέσω της εντολής get_group, εξάγουμε τα δεδομένα κάθε χώρας από τα "indexed" πλαίσια g μέσω του αντίστοιχου δείκτη και τα αποθηκεύουμε σε ένα νέο πλαίσιο δεδομένων το οποίο ακολουθεί το εξής πρότυπο ονομασίας :

$$\{\ \mathit{ΚωδικόςΧώρας\_fΑριθμόςτουΜήνα}\ \}$$

Στο σημείο αυτό έχουμε στη διάθεση μας όλα τα διαθέσιμα δεδομένα Σ.Π.Φ ανά χώρα και ανά μήνα. Όμως τα δεδομένα αυτά δεν είναι απαραίτητα διατεταγμένα ορθά ως προς τη χρόνική στιγμή δήλωσης των τιμών Σ.Π.Φ. Συνεπώς, απαιτείται να εκτελέσουμε τη φυσική χρονική διάταξη των τιμών Σ.Π.Φ σύμφωνα με τα πεδία τιμών της στήλης "DateTime" για κάθε μηνιαίο αρχείο κάθε χώρας. Στη συνέχεια, ενδέχεται να υπάρχουν ελλιπείς τιμές Σ.Π.Φ (απουσία χρονικών δηλώσεων Σ.Π.Φ) σε κάθε μηνιαίο αρχείο κάθε χώρας. Για την αντιμετώπιση του προβλήματος των ελλιπών τιμών δημιουργήσαμε τις συναρτήσεις $Month\_Populate()$ (μια για κάθε μηνιαίο αρχείο) που ελέγχουν τα πεδία της στήλης "DateTime" και εντοπίζουν τις χρονικές στιγμές (datetime instances) οι οποίες απουσιάζουν. Οι συναρτήσεις αυτές, αφού ταξινομήσουν χρονικά ορθά το αντίστοιχο μηνιαίο πλαίσιο ως προς τις τιμές των πεδίων της στήλης "DateTime", έπειτα εντοπίζουν όλες τις απούσες χρονικές στιγμές, δημιουργούν μια νέα στήλη "DateTime" που περιέχει όλες τις χρονικές στιγμές ορθά ταξινομημένες και στη συνέχεια συμπληρώνουν τις αντίστοιχες απούσες τιμές Σ.Π.Φ με τον τύπο δεδομένων "NaN".

Συνεπώς, οι συναρτήσεις $Month\_Populate()$ εκτελούν τους παρακάτω ελέγχους :

i. Αν η χρονική παρατήρηση υπάρχει, τότε συμπληρώνουν τη νέα στήλη "TotalLoadValue" με την αντίστοιχη τιμή Σ.Π.Φ που βρίσκεται στην αρχική στήλη "TotalLoadValue".

ii. Αν η χρονική παρατήρηση απουσιάζει, τότε συμπληρώνουν το πεδίο που αντιστοιχεί σε αυτή τη χρονική παρατήρηση της νέας στήλης "TotalLoadValue" με την ειδική τιμή "NaN".

Οι συναρτήσεις της οικογένειας συναρτήσεων $Month\_Populate()$ διαφοροποιούνται μόνο ως προς τα παρακάτω :

- Ορίσματα του "datetime object" που κατασκευάζεται στο σώμα τους με την εντολή $pd.date\_range$.
- Τα όρια των εντολών επανάληψης που εκτελούνται στο σώμα τους.

Για παράδειγμα, όσον αφορά τον μήνα Σεπτέμβριο, κατασκευάζουμε ένα "datetime object" ως εξής :

```
dts = pd.date_range(start='2019/09/01', periods=720, freq="H")
```

και οι εντολές επανάληψης θα τρέχουν από το βήμα 0 εώς και το βήμα 719.

Συνεπώς, στο στάδιο "Apply", με βάση τα παραπάνω, καλούμε τις εν λόγω συναρτήσεις που επενεργουν στα μηνιαία αρχεία κάθε χώρας και τα επιστρέφουν στην επιθυμητή μορφή.

Στο σημείο αυτό, έχουμε στη διάθεση μας όλες τις μηνιαίες χρονοσειρές κάθε χώρας. Κάθε πλαίσιο δεδομένων (dataframe) που περιέχει οποιαδήποτε μηνιαία χρονοσειρά οποιασδήποτε χώρας αναγνωρίζεται σύμφωνα με το πρότυπο ονομασίας :

$$\{MapCodeΑριθμόςΜήνα\_nomiss\}$$

Βρισκόμαστε τώρα στο στάδιο "Combine" όπου απαιτείται να συνδιάσουμε τις εκάστοτε μηνιαίες χρονοσειρές σε μια ενιαία ετήσια χρονοσειρά. Η διαδικασία αυτή θα πραγματοποιηθεί για κάθε χώρα που ανήκει στο σύνολο ανάλυσης.

Για την επίτευξη αυτού του στόχου δημιουργήσαμε τη συνάρτηση $merge\_12months\_load\_values()$ που δέχεται ως ορίσματα τα δώδεκα αντίστοιχα μηνιαία αρχεία της χώρας και συγχωνεύει τις αντίστοιχες μηνιαίες χρονοσειρές σε μια ενιαία ετήσια χρονοσειρά η οποία επιστρέφεται ως λίστα.

Στη συνέχεια παραθέτουμε τον ψευδοκώδικα της παραπάνω συνάρτησης .

Ψευδοκώδικας συνάρτησης $merge\_12months\_load\_values()$:
1. Κατασκεύασε μια κενή λίστα whole_period[ ]
2. Για κάθε μηνιαίο αρχείο *(με τη χρονική σειρά που τα διέπει)*:
3. Φόρτωσε διαδοχικά τις τιμές Σ.Π.Φ με τη σειρά που ορίζει η στήλη "DateTime"στη λίστα whole_period
4. Έλγχος διάστασης λίστας whole_period
5. Επίστρεψε τη λίστα whole_period

Εφόσον έχουμε καλέσει την παραπάνω συνάρτηση για όλες τις χώρες του συνόλου ανάλυσης, δημιουργούμε ένα "datetime object" το οποίο θα αποτελέσει τη στήλη "DateTime" στο τελικό πλαίσιο δεδομένων df_final . Επίσης, εισάγουμε ως στήλες με κατάλληλη ονομασία τις ετήσιες χρονοσειρές Σ.Π.Φ κάθε χώρας.

Η τυπική μορφή του πλαισίου δεδομένων (dataframe) df_final παρουσιάζεται στη συνέχεια.

DateTime	BA_TotalLoadValue	...	UA_TotalLoadValue
		...	
2019-03-01 00:00:00	1192.09	...	16738.0
2019-03-01 01:00:00	1136.57	...	16938.0
2019-03-01 02:00:00	1109.53	...	17854.0
2019-03-01 03:00:00	1138.52	...	19369.0
2019-03-01 04:00:00	1234.47	...	20181.0
...	...	...	...
2020-02-29 19:00:00	1655.67	...	17998.0
2020-02-29 20:00:00	1580.91	...	17137.0
2020-02-29 21:00:00	1535.09	...	16354.0
2020-02-29 22:00:00	1414.46	...	15543.0
2020-02-29 23:00:00	1273.31	...	15010.0

[8784 rows x 25 columns]

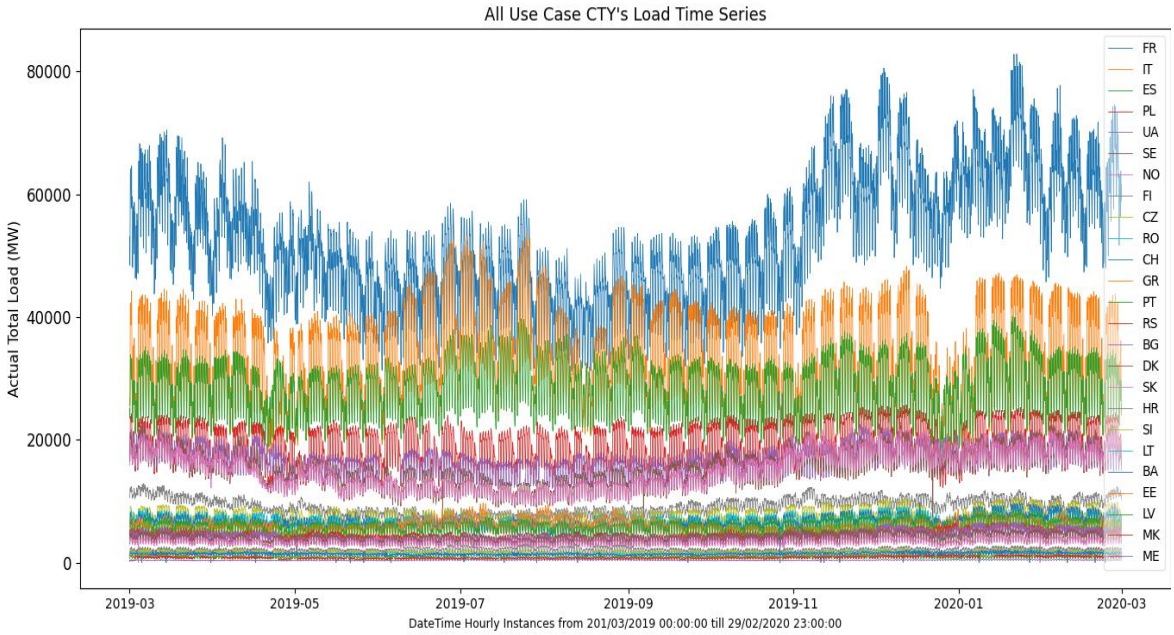
Σχήμα 4.4 : Μορφή τελικού πλαισίου δεδομένων (dataframe) του 1ου σταδίου προεπεξεργασίας.

Τέλος, αποθηκεύουμε το τελικό πλαίσιο δεδομένων με τη παρακάτω εντολή :

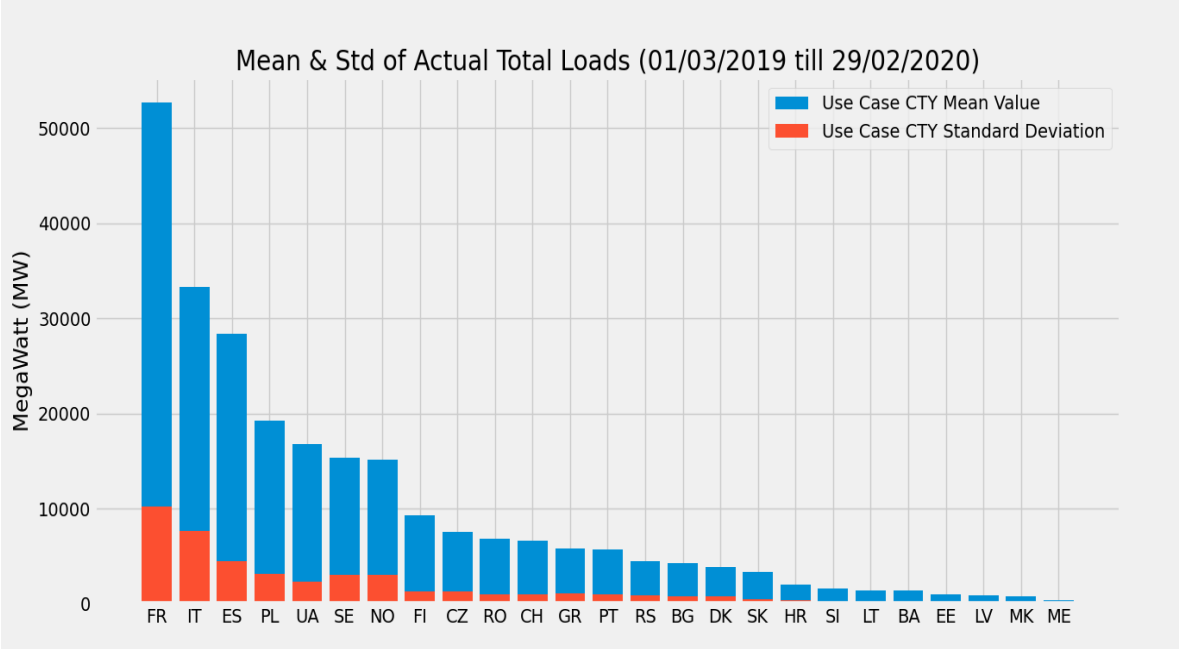
```
final_frame.to_csv("25CTY_ActualTotalLoad_2019Year.csv", sep="\t",
                                encoding="utf_16", index=False)
```

Προτού προχωρήσουμε στο δεύτερο στάδιο προεπεξεργασίας, αποτελεί καλή πρακτική η επισκόπηση των χρονοσειρών που συνθέσαμε μέσω διαφόρων γραφικών παραστάσεων καθώς και ο υπολογισμός βασικών στατιστικών τους μεγεθών. Η εν λόγω διαδικασία θα αναδείξει τυχόν προβλήματα όπως η ύπαρξη ακραίων τιμών και θορύβου. Επίσης, θα αποτελέσει έναν έλεγχο για την εγκυρότητα των αποτελεσμάτων του πρώτου σταδίου προεπεξεργασίας.

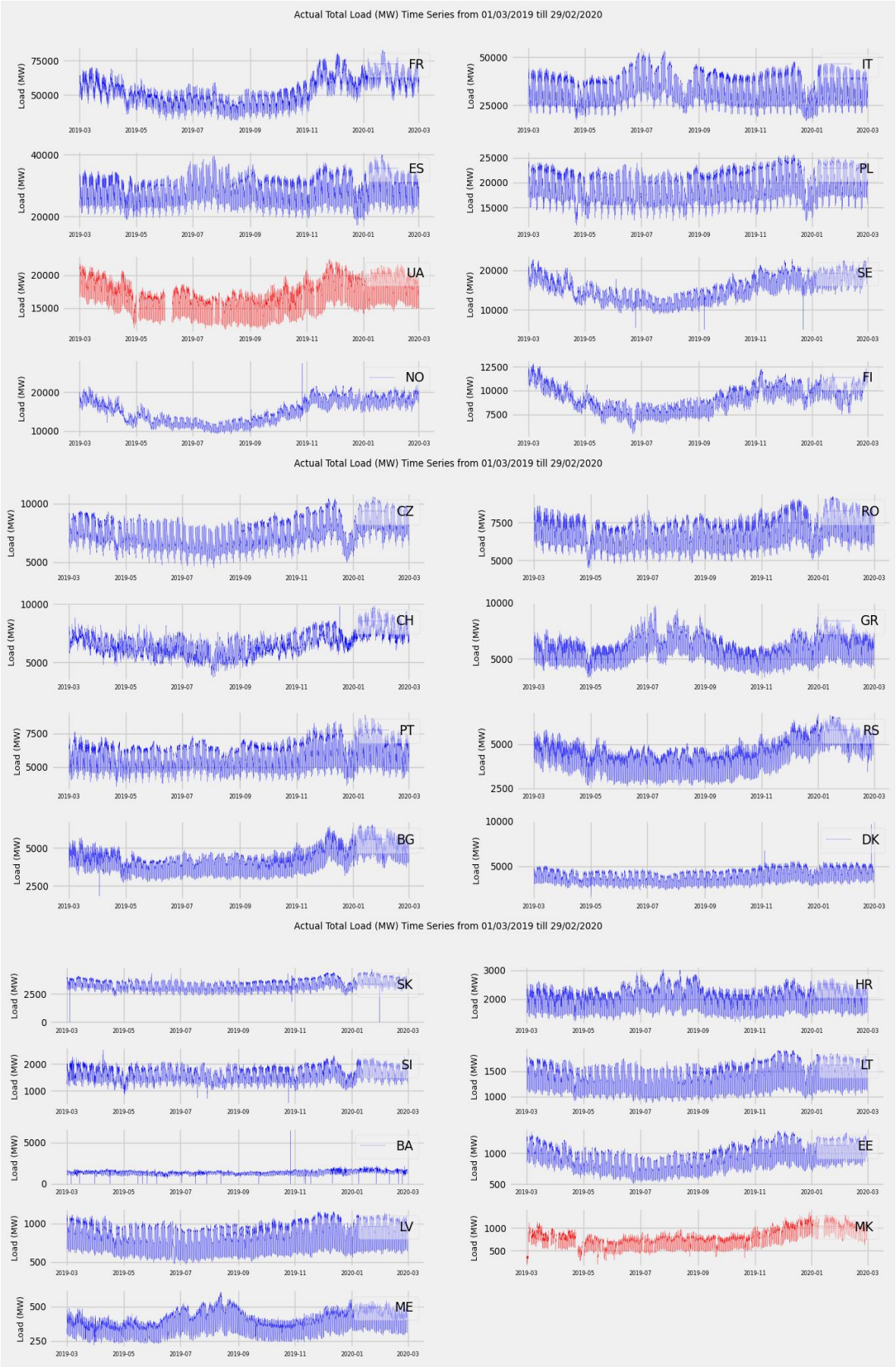
Στη συνέχεια, παρουσιάζουμε τα εν λόγω γραφήματα για τις ετήσιες χρονοσειρές του χρονικού παραθύρου ["01/03/2019", "29/02/2020"]. Στο χρονικό αυτό παράθυρο ανάλυσης αναφερόμαστε ως έτος "2019". Τα αντίστοιχα γραφήματα των δεδομένων των υπολοίπων χρονικών παραθύρων ανάλυσης παρουσιάζονται στο Παράρτημα Α.



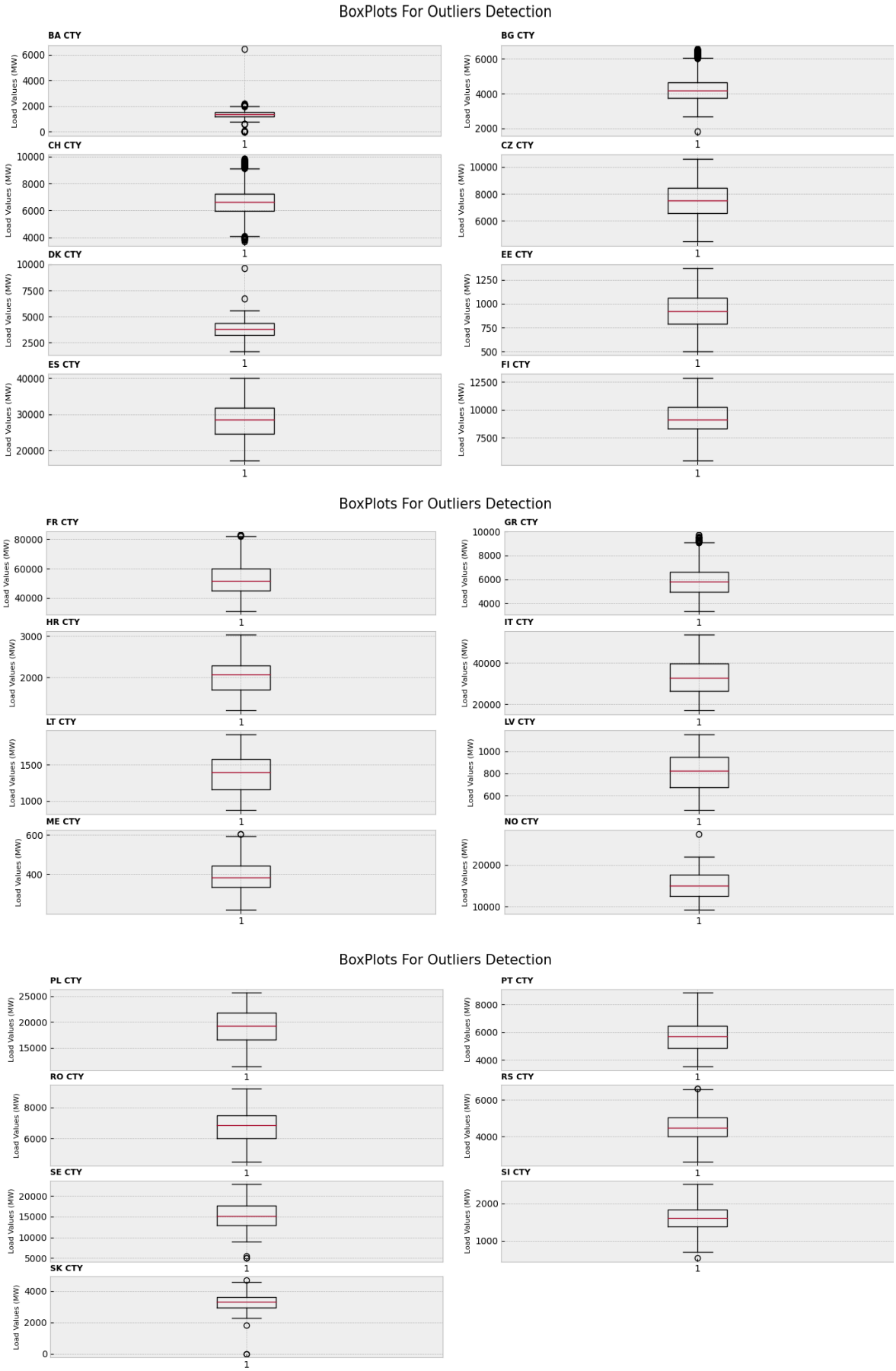
Σχήμα 4.5 : Γραφική παράσταση των χρονοσειρών του έτους "2019".



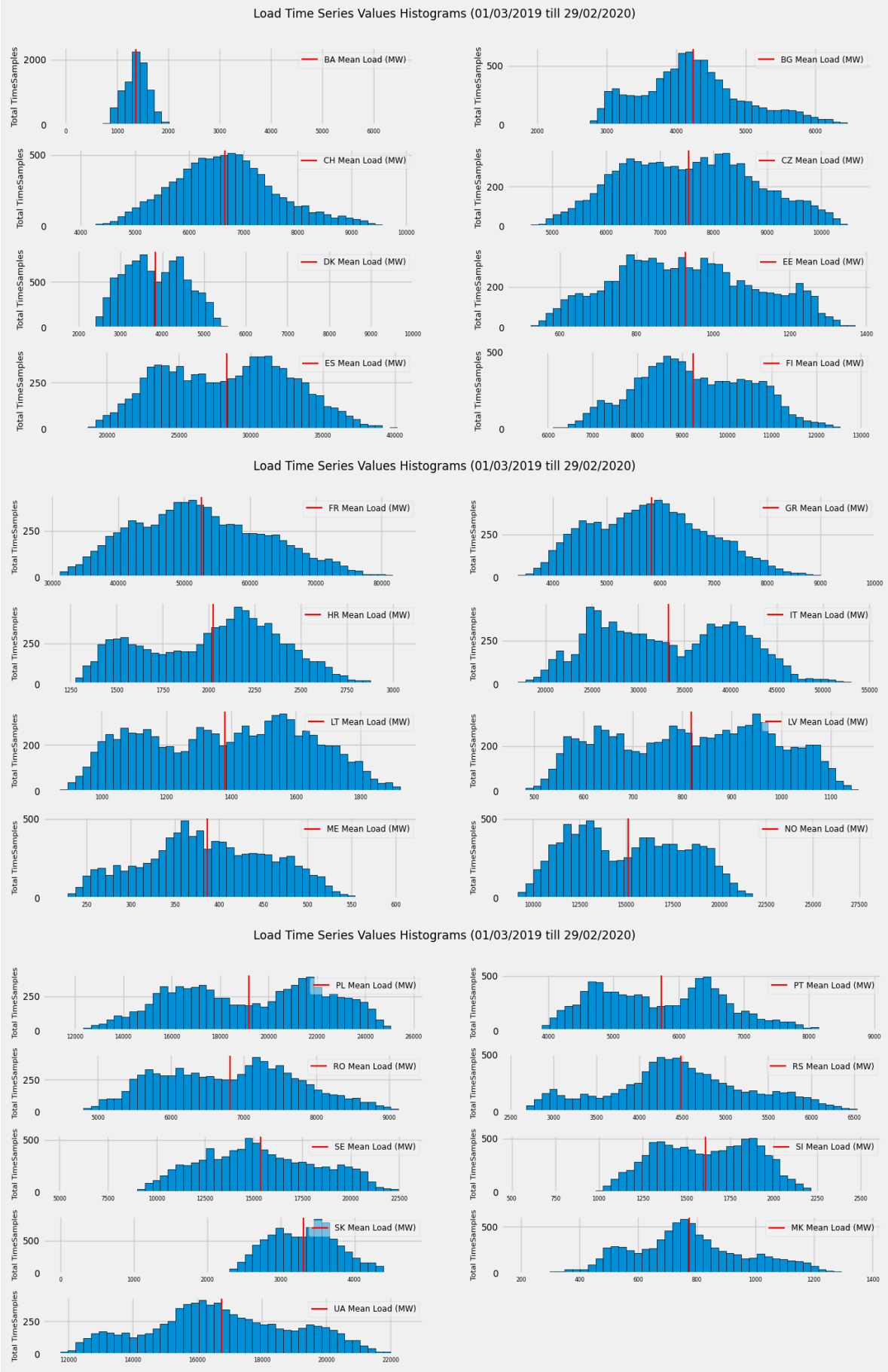
Σχήμα 4.6 : Ραβδόγραμμα μέσης τιμής και τυπικής απόκλισης των χρονοσειρών του "2019".



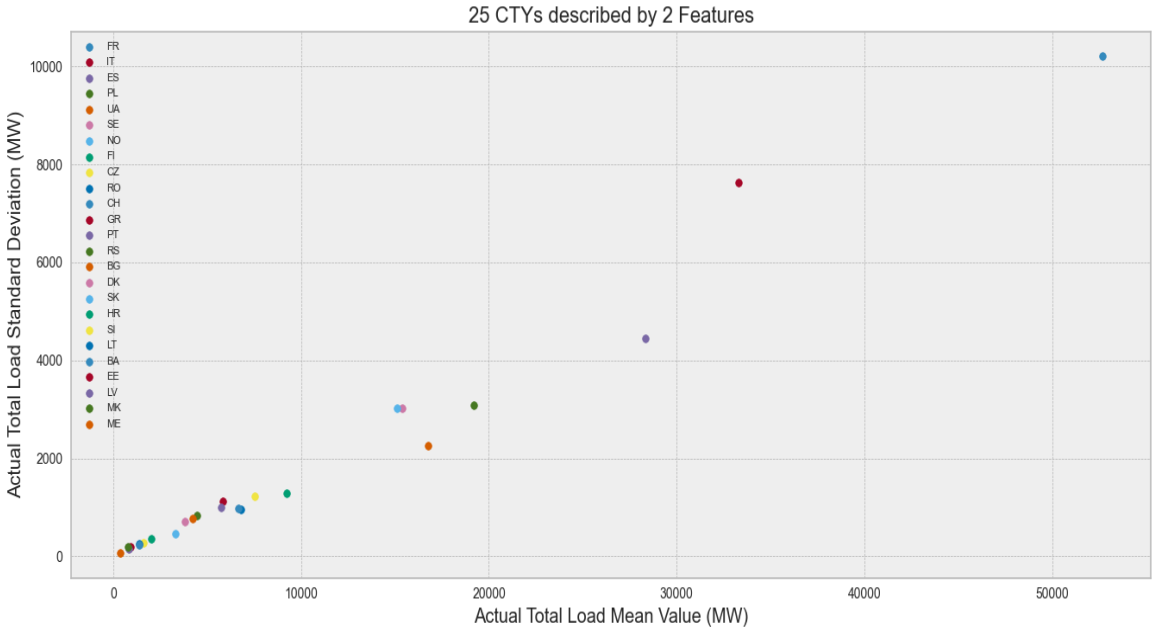
Σχήμα 4.7 : Γραφικές παραστάσεις των ετήσιων χρονοσειρών του "2019".



Σχήμα 4.8 : Διαγράμματα BoxPlots των ετήσιων χρονοσειρών του "2019".



Σχήμα 4.9 : Ιστογράμματα χρονοσειρών του έτους "2019".



Σχήμα 4.10 : Δισδιάστατη οπτικοποίηση χρονοσειρών του έτους "2019" ως προς τη μέση τιμή και την τυπική τους απόκλιση.

<u>Με βάση τα σχήματα 4.5 εώς 4.10 συμπεραίνουμε τα εξής</u> :

- Οι χρονοσειρές UA και MK παρουσιάζουν ελλιπείς τιμές – παρατηρήσεις.
- Οι χρονοσειρές BA, BG, DK, CH, ME, MK, NO, SE, SK, SI παρουσιάζουν ακραίες τιμές (outliers).
- Η χρονοσειρά της Βοσνίας-Ερζεγοβίνης (BA) παρουσιάζει πολλές διακοπές φορτίου. Πιθανώς η χώρα αυτή δεν έχει σχεδιάσει επαρκώς το Σ.Η.Ε της.
- Από τα ιστογράμματα παρατηρείται για τις περισσότερες χρονοσειρές ότι οι παρατηρήσεις αυτών προέρχονται από δύο διαφορετικούς πληθυσμούς, καθώς παρουσιάζουν προσεγγιστικά διτροπικές κατανομές (bimodal distributions). Το γεγονός αυτό είναι λογικό και δεν πρέπει να μας ανησυχήσει, καθώς οι καμπύλες φορτίου επηρεάζονται και διαφοροποιούνται σημαντικά κατά τους καλοκαιρινούς και χειμερινούς μήνες.
- Οι χρονοσειρές FR, IT και ES παρουσιάζουν ιδιαίτερα μεγάλες τιμές Σ.Π.Φ σε σχέση με τις υπόλοιπες χώρες του συνόλου ανάλυσης. Το γεγονός αυτό είναι απόλυτα λογικό καθώς οι χώρες αυτές είναι ιδιαίτερα μεγάλης έκτασης και επίσης διαθέτουν βαριά βιομηχανία.
- Σε γεννικές γραμμές φαίνεται ότι το πρώτο στάδιο προεπεξεργασίας εκτελέστηκε με επιτυχία.

Στο σημείο αυτό είμαστε σε θέση να προβούμε στο δεύτερο στάδιο προεπεξεργασίας έχοντας υπόψη την ύπαρξη ακραίων και ελλιπών τιμών σε ορισμένες χρονοσειρές του συνόλου ανάλυσης.

2. 2ο Στάδιο Προεπεξεργασίας

Στο στάδιο αυτό εισάγουμε τις βιβλιοθήκες "pandas" και "numpy" με τις παρακάτω εντολές :

```
import pandas as pd
import numpy as np
```

Αρχικά, διαβάζουμε το αρχείο tsv που περιέχει τις ετήσιες χρονοσειρές φορτίου των ευρωπαϊκών χωρών του συνόλου ανάλυσης και το αποθηκεύουμε σε ένα πλαίσιο δεδομένων.

Στη συνέχεια, εξάγουμε τις ετικέτες των στηλών του πλαισίου δεδομένων και τις αποθηκεύουμε σε μία λίστα. Οι ετικέτες αυτές αποτελούν το αναγνωριστικό των χρονοσειρών που εμπεριέχονται στο πλαίσιο δεδομένων.

Σε επόμενο βήμα εξάγουμε τα δύο πρώτα σύμβολα από κάθε συμβολοσειρά που αποτελεί ουσιαστικά την ετικέτα – αναγνωριστικό της αντίστοιχης στήλης και τα αποθηκεύουμε σε μία άλλη λίστα. Με αυτό το τρόπο έχουμε όλους τους "MapCode" κωδικούς των χωρών που εμπίπτουν στο σύνολο ανάλυσης. Οι ετικέτες των στηλών όπως φαίνεται και από το σχήμα 4.4 είναι της μορφής $MapCode\_TotalLoadValue$.

Έπειτα, με βάση τα σχήματα 4.8 και 4.9 αντικαθιστούμε τις ακραίες παρατηρήσεις των χρονοσειρών με την ειδική τιμή "NaN". Για τον σκοπό αυτό υλοποιήσαμε τη συνάρτηση $Outliers\_Removal()$ η οποία αναλύεται στη συνέχεια.

Η συνάρτηση $Outliers\_Removal()$ δέχεται πέντε ορίσματα :

- df : Το πλαίσιο δεδομένων που περιέχει τις ετήσιες χρονοσειρές.
- headercol : Το αναγνωριστικό της στήλης του πλαισίου δεδομένων df.
- threshold : Το κατώφλι με βάση το οποίο οι τιμές αντικαθίστανται με τον ειδικό τύπο δεδομένων "NaN".
- bool_upper : Εάν "TRUE" τότε οι τιμές που βρίσκονται πάνω από το κατώφλι αντικαθίστανται με "NaN".
- bool_lower : Εάν "TRUE" τότε οι τιμές που βρίσκονται κάτω από το κατώφλι αντικαθίστανται με "NaN".

Συνεπώς, καλούμε τη συνάρτηση $Outliers\_Removal()$ για κάθε χρονοσειρά χώρας που περιέχει ακραίες τιμές. Ο εντοπισμός των ακραίων τιμών έχει πραγματοποιηθεί μέσω της εποπτείας των σχημάτων 4.8 και 4.9.

Στο σημείο αυτό, μπορούμε να προβούμε στη διάσπαση των ετήσιων χρονοσειρών φορτίου σε ημερήσιες χρονοσειρες φορτίου (daily load curves – DLC).

Η κατασκευή των ημερήσιων χρονοσειρών πραγματοποιείται μέσω της συνάρτησης $DLC\_extraction()$ που δέχεται ως όρισμα το τροποποιημένο ετήσιο πλαίσιο χρονοσειρών και το αναγνωριστικό "MapCode"κωδικό της αντίστοιχης χώρας.

Ψευδοκώδικας συνάρτησης $DLC\_extraction()$
1. Δημιούργησε ονόματα, για κάθε μέρα i του έτους, ως συμβολοσειρές "Day_i" και φόρτωσέ αυτές με τη σωστή χρονική σειρά σε μία λίστα days_list.
2. Δημιούργησε μία βοηθητική κενή λίστα period_list[ ].
3. Για i = 1 εως και τον αριθμό της διάστασης της ετήσιας χρονοσειράς, με βήμα 24 :
4. Φόρτωσε σε μία βοηθητική λίστα aux_list, τα 24 πρώτα δεδομένα Σ.Π.Φ της ετήσιας χρονοσειράς ξεκινόντας από την χρονική της θέση i
5. Φόρτωσε τη λίστα aux_list στη λίστα period_list.
6. Κατασκεύασε ένα πλαίσιο δεδομένων με στήλες τις λίστες που είναι αποθηκευμένες στην λίστα period_list και δεικτοποίησε τις εν λόγω στήλες με τις κατάλληλες συμβολοσειρές της λίστας days_list.
7. Δεικτοποίησε το πλαίσιο δεδομένων με τον "MapCode" κωδικό της χώρας.
8. Επίστρεψε το πλαίσιο δεδομένων.

Στο σημείο αυτό, εφόσον έχουμε καλέσει τη συνάρτηση $DLC\_extraction()$ για κάθε χώρα, έχουμε στη διαθεσή μας όλες τις ημερήσιες καμπύλες φορτίου των χωρών που εμπίπτουν στο σύνολο ανάλυσης.

Τελευταίο βήμα είναι η ενσωμάτωση όλων αυτών των πλαισίων δεδομένων που περιέχουν τις ημερήσιες καμπύλες φορτίου των αντίστοιχων χωρών σε ένα τελικό ενιαίο πλαίσιο δεδομένων. Το τελικό αυτό πλαίσιο θα δεικτοποιηθεί με βάση τους "MapCode" κωδικούς των χωρών για να μπορούμε να εξάγουμε ξεχωριστά τις ημερήσιες καμπύλες φορτίου κάθε χώρας.

Το εν λόγω τελικό πλαίσιο δημιουργείται εκτελώντας τη παρακάτω εντολή :

```
df_final = pd.concat([df_ba, df_bg, df_ch, df_cz, df_dk, df_ee, df_es, df_fi, df_fr,
                      df_gr, df_hr, df_it ,df_lt, df_lv, df_me, df_mk, df_no, df_pl,
                      df_pt , df_ro, df_rs ,df_se ,df_si ,df_sk, df_ua],
                      keys = cty_mapcode_list)
```

Η τυπική μορφή του πλαισίου δεδομένων df_final που περιέχει τις ημερήσιες καμπύλες Σ.Π.Φ όλων των χωρών που εμπίπτουν στο σύνολο ανάλυσης παρουσιάζετε στη συνέχεια.

		Day_1	Day_2	Day_3	...	Day_365	Day_366	MapCode
BA	0	1192.09	1062.73	1193.93	...	1244.17	1302.74	BA
	1	1136.57	1019.89	1129.71	...	1205.68	1238.21	BA
	2	1109.53	984.13	1104.98	...	1174.45	1217.91	BA
	3	1138.52	981.84	1102.80	...	1216.94	1239.84	BA
	4	1234.47	1065.97	1125.80	...	1335.62	1290.33	BA
...		...	...	...	...	...	...	...
UA	19	19153.00	19187.00	18888.00	...	18410.00	17998.00	UA
	20	18464.00	18430.00	18158.00	...	17441.00	17137.00	UA
	21	17599.00	17732.00	17236.00	...	16708.00	16354.00	UA
	22	16948.00	17113.00	16800.00	...	15734.00	15543.00	UA
	23	16774.00	16748.00	16726.00	...	15354.00	15010.00	UA

[600 rows x 367 columns]

Σχήμα 4.11 : Τυπική μορφή πλαισίου δεδομένων που περιέχει όλες τις ημερήσιες καμπύλες φορτίου όλων των χωρών που εμπίπτουν στο σύνολο ανάλυσης του έτους.

Μετά το πέρας του δεύτερου σταδίου προεπεξεργασίας, έχουμε δημιουργήσει τις εξής βάσεις δεδομένων :

1. $\boldsymbol{DB_1 = \{X_i,\ i = 2015, \dots, 2020\}}$

   Όπου :

   - $X_i = \{x_i^{(m_i)}\ ,\ m_{2015} = m_{2016} = 1, \dots, 23\ \wedge\ m_{2017} = m_{2020} = 1, \dots, 24 \wedge\ m_{2018} = m_{2019} = 1, \dots ,25\}$

   - $x_i^{(m_i)} = \left[x_1^{m_i}, \dots , x_{d_i}^{m_i}\right]^T, d_{2015} = d_{2016} = d_{2017} = d_{2018} = d_{2020} = 8760 \wedge\quad d_{2019} = 8784$

2. $\boldsymbol{DB_2 = \{Z_i,\ i = 2015, \dots, 2020\}}$

   Όπου :

   - $Z_i = \{Z_i^{(m_i)},\ m_{2015} = m_{2016} = 1, \dots ,23\ \wedge\ m_{2017} = m_{2020} = 1, \dots ,24 \wedge\ m_{2018} = m_{2019} = 1, \dots ,25\ \}$

   - $Z_i^{(m_i)} = \{z_i^{(j^i)^{m_i}}, j^{2015} = j^{2016} = j^{2017} = j^{2018} = j^{2020} = 1, \dots, 365 \wedge\ j^{2019} = 1, \dots ,366\}$

   - $z_i^{(j^i)^{m_i}} = \left[z_{1_{m_i}}^{j^i}, \dots , z_{24_{m_i}}^{j^i}\right]^T$

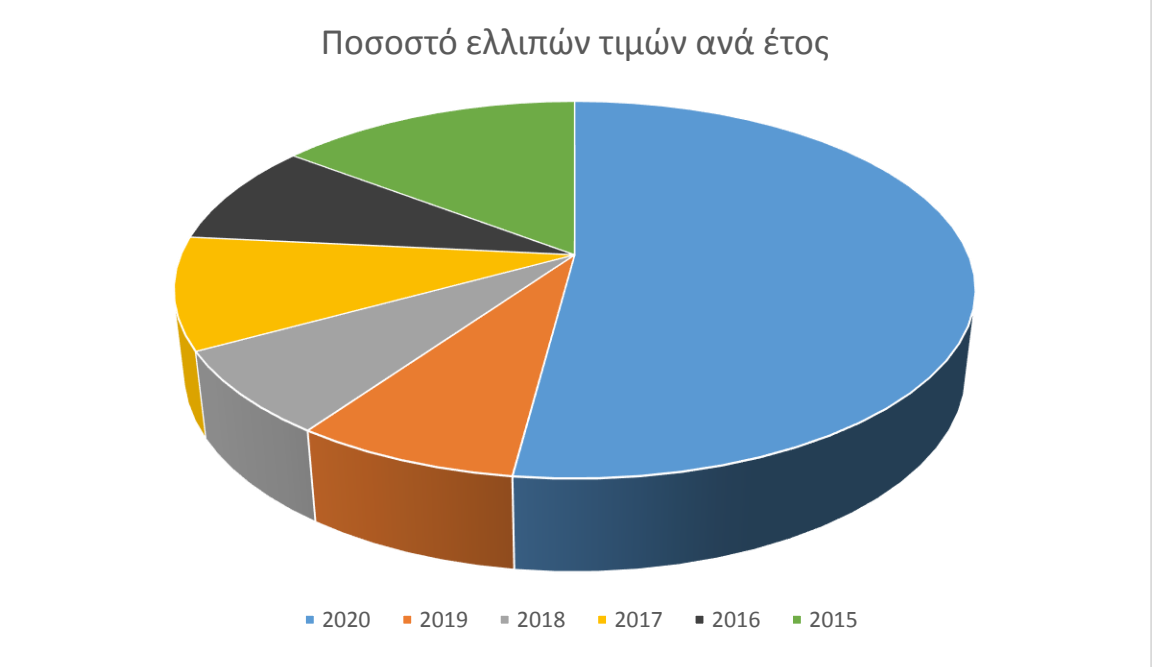
Στο σημείο αυτό αναφέρουμε για μία ακόμα φορά πως έχουμε θεωρήσει κάθε έτος, εκτός του 2015, να ξεκινάει από τη 1η Μαρτίου μέχρι το τέλος Φεβρουαρίου του επόμενου έτους. Για παράδειγμα, όταν ανφερόμαστε στο έτος 2019 τότε εννοούμε το χρονικό παράθυρο "01/03/2019" εώς "29/02/2020". Για αυτό το λόγο, όσον αφορά το έτος 2019, παρουσιάζεται στη 2η βάση δεδομένων ότι σε κάθε χώρα αντιστοιχούν 366 ημερήσιες χρονοσειρές φορτίου (αντί για 365) και αντίστοιχα στο έτος 2020 ότι αντιστοιχούν 365.

Τα ετήσια αρχεία του 2015 και του 2016 περιέχουν πληροφορία από είκοσι τρεις χώρες καθώς η Βοσνία –Ερζεγοβίνη και η Ουκρανία απουσιάζουνε από τη βάση δεδομένων όσον αφορά τα έτη αυτά. Επίσης, σημειώνεται ότι η Ουκρανία απουσιάζει και από τα ετήσια δεδομένα του 2017 και η Βοσνία –Ερζεγοβίνη από τα δεδομένα του 2020.

Στη συνέχεια παρουσιάζουμε σε πίνακα το ποσοστό των ελλιπών τιμών (missing values) σε κάθε ετήσιο πλαίσιο δεδομένων X_i ως προς τον συνολικό αριθμό των τιμών που περιέχονται. Σημειώνεται ότι οι ελλιπείς τιμές έχουν εισαχθεί σε σωστή χρονική σειρά διάταξης ως ειδικές τιμές "NaN" μέσω των συναρτήσεων $Month\_Populate()$.

Πίνακας 4.2 : Ελλιπείς τιμές στη βάση δεδομένων $\boldsymbol{DB_1}$ των ετήσιων χρονοσειρών Σ.Π.Φ.

Ετήσιες Χρονοσειρές	Χώρες	Συνολικός αριθμός τιμών Σ.Π.Φ	Ελλιπείς τιμές Σ.Π.Φ	Ποσοστό Ελλιπών τιμών
2020	*24*	*210240*	*7190*	*3,42 %*
2019	*25*	*219600*	*1062*	*0,48 %*
2018	*25*	*219000*	*1001*	*0,46 %*
2017	*24*	*210240*	*1324*	*0,63 %*
2016	*23*	*201480*	*1212*	*0,60 %*
2015	*23*	*201480*	*2024*	*1%*



Σχήμα 4.12 : Διάγραμμα πίτας συνολικών ελλιπών τιμών ανά έτος.

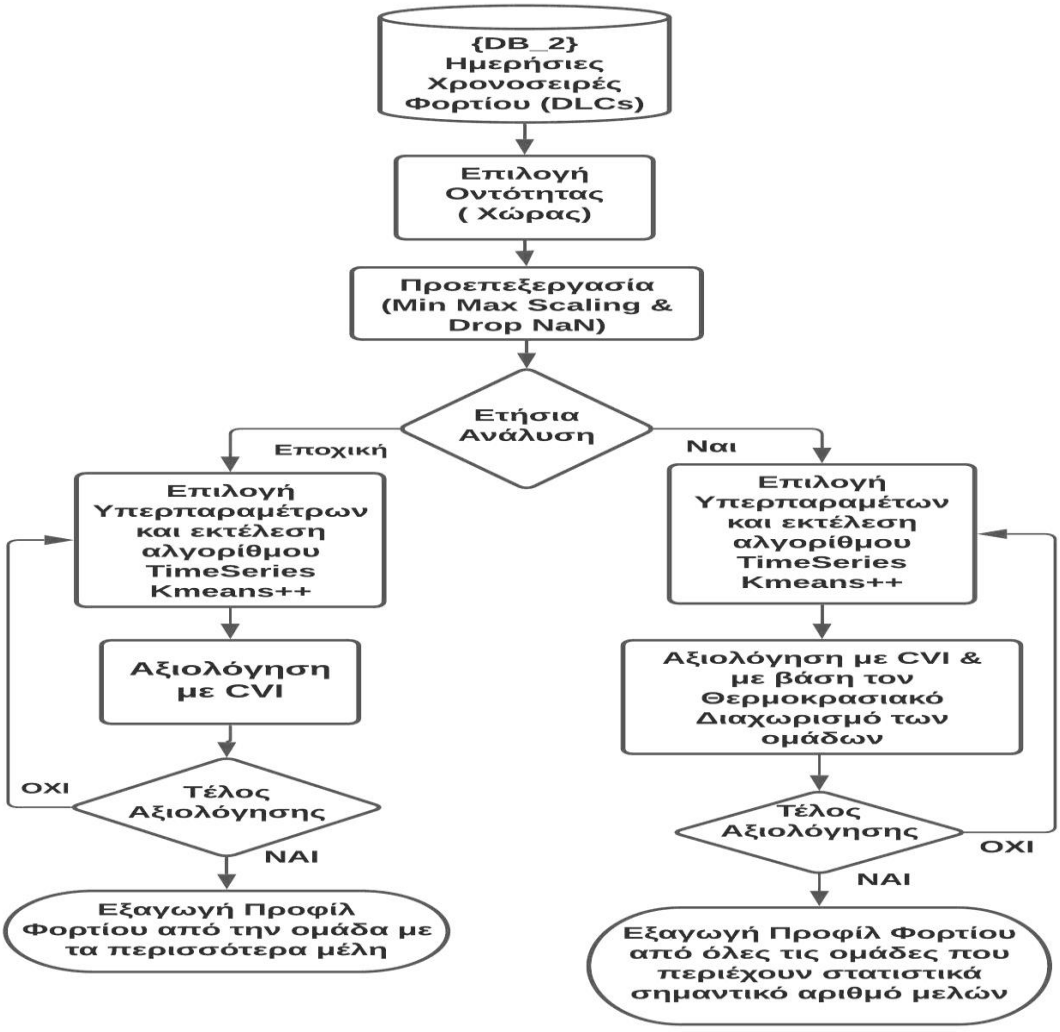
4.3 Μεθοδολογία Εφαρμογών Ομαδοποίησης

Για την υλοποίηση των εφαρμογών ομαδοποίησης χρησιμοποιήθηκε ο αλγόριθμος TimeSeriesKMeans της βιβλιοθήκης tslearn σε συνδιασμό με τον ευρετικό αλγόριθμο αρχικοποίησης κεντροειδών kmeans++. Πλήρης περιγραφή των αλγορίθμων μπορεί να βρεθεί εδώ [32, 60]. Για την αξιολόγηση των αποτελεσμάτων αξιοποιήθηκαν οι μετρικές επικύρωσης ομαδοποίησης (clustering validation indices – CVI) που παρουσιάστηκαν αναλυτικά στο κεφάλαιο 2.3.2.

Στόχος των εν λόγω εφαρμογών είναι η ομαδοποίηση χρονοσειρών ηλεκτρικού φορτίου των ευρωπαϊκών χωρών του συνόλου ανάλυσης και η εξαγωγή των τυπικών χαρακτηριστικών τους καμπυλών. Οι χαρακτηριστικές καμπύλες φορτίου αναφέρονται στη βιβλιογραφία ως "Προφίλ" Φορτίου (Load Profiles). Η ομαδοποίηση αυτών των χρονοσειρών απαιτείται να υλοποιηθεί με βάση το σχήμα τη καμπύλης τους, καθώς αυτό αντικατοπτρίζει την ενεργειακή συμπεριφορά των μηχανισμών παραγωγής τους. Όπώς είναι προφανές, με βάση τα σχήματα 4.6 και 4.8, οι καμπύλες φορτίου του συνόλου ανάλυσης παρουσιάζουν μεγάλες αποκλίσεις ως προς τη τάξη μεγέθους των αντίστοιχων τιμών τους. Συνεπώς, απαιτείται να πραγματοποιηθεί κατάλληλος μετασχηματισμός των δεδομένων που θα διατηρήσει τη πληροφορία του σχήματος των καμπυλών και ταυτόχρονα θα περιορίσει τις σχετικές τους αριθμητικές αποκλίσεις. Οι μη γραμμικές μέθοδοι κανονικοποίησης επηρεάζουν το σχήμα των καμπυλών και κατά συνέπεια κρίνονται ακατάλληλες σύμφωνα με τις παραδοχές του εν λόγω προβλήματος. Συνεπώς, οι πλέον κατάλληλες μέθοδοι κανονικοποίησης ανήκουν στη κατηγορία των γραμμικών μεθόδων. Επίσης, για να έχουν φυσική σημασία οι κανονικοποιημένες τιμές Σ.Π.Φ θα πρέπει να ανήκουν σε ένα θετικό εύρος τιμών. Σύμφωνα με τα παραπάνω, και δεδομένου ότι οι προεπεξεργασμένες χρονοσειρές ανάλυσης δεν παρουσιάζουν σημαντικό αριθμό ακραιών παρατηρήσεων, επιλέξαμε τη μέθοδο Min Max Scaling με παραμέτρους min = 0 και max = 1. Η περιγραφή της μεθόδου κανονικοποίησης Min Max Scaling έχει πραγματοποιηθεί στο κεφάλαιο 3.3.2. Σημειώνεται ότι η εν λόγω μέθοδος κανονικοποίησης έχει αξιοποιηθεί στο παρελθόν σε εφαρμογές ομαδοποίησης καμπυλών φορτίου και έχει διαπιστωθεί η καταλληλότητα της [33].

Όσον αφορά το μέτρο ανομοιότητας με βάση το οποίο θα ομαδοποιεί τις καμπύλες φορτίου ο αλγόριθμος ομαδοποίησης, επιλέξαμε την ευκλείδεια απόσταση, η οποία αποτελεί μετρική και παρουσιάζει ικανοποιητικά αποτελέσματα σε εφαρμογές ομαδοποίησης καμπυλών ηλεκτρικού φορτίου [31, 33]. Για την εξαγωγή των "Πορφίλ" Φορτίου κάθε χώρας υπολογίσαμε τη μέση τιμή των αντίστοιχων ομαδοποιημένων ημερήσιων καμπυλών φορτίου (daily load curves – DLC). Η εν λόγω εφαρμογή υλοποιήθηκε σε δύο διαφορετικά πλαίσια. Στο πρώτο πλαίσιο πραγματοποιήθηκε βάση των συνθηκών φόρτισης, δηλαδή εποχικά (άνοιξη, καλοκαίρι, φθινόπωρο χειμώνας). Επίσης, σε δεύτερο πλαίσιο η εξαγωγή των "Προφίλ" Φορτίου πραγματοποιήθηκε στο σύνολο του έτους.

Στη συνέχεια, απεικονίζεται το διάγραμμα ροής της μεθοδολογίας που ακολουθήσαμε για την εξαγωγή των "Προφίλ" Φορτίου.



Σχήμα 4.13 : Διάγραμμα ροής μεθοδολογίας εξαγωγής των "Προφίλ" Φορτίου.

Στο διάγραμμα ροής του σχήματος 4.13 παρουσιάζεται η διαφορετική στρατηγική εξαγωγής των "Προφίλ" Φορτίου όσον αφορά την ετήσια και την εποχική ανάλυση. Κατά την εποχική ανάλυση επιλέγουμε να εξάγεται το "Προφίλ" Φορτίου από την ομάδα η οποία περιέχει τα περισσότερα μέλη, δηλαδή τις περισσότερες ημερήσιες καμπύλες Σ.Π.Φ. Η εν λόγω μέθοδος αναφέρεται στη βιβλιογραφία ως "Most Populated Approach". Κατά την ετήσια ανάλυση, παρατηρήσαμε ότι οι μετρικές επικύρωσης ομαδοποίησης υποδείκνυαν ομαδοποιήσεις ημερήσιων καμπυλών Σ.Π.Φ οι οποίες ήταν σε γενικές γραμμές ικανοποιητικές κατά την οπτικοποίηση των ομάδων, όμως δεν παρουσιαζόταν επαρκής θερμοκρασιακός διαχωρισμός μεταξύ των ομάδων. Πιο συγκεκριμένα, κατά την ανά χώρα ετήσια ανάλυση, παρατηρήσαμε σε αρκετές περιπτώσεις την ανάθεση ημερήσιων καμπυλών φορτίου της καλοκαιρινής και της χειμερινής περιόδου σε ίδιες ομάδες. Δεδομένου ότι η θερμοκρασία επηρεάζει σημαντικά τα συστήματα παραγωγής ηλεκτρικής ενέργειας, επιλέξαμε να αξιολογήσουμε τις ομαδοποιήσεις της ετήσιας ανάλυσης όχι μόνο βάση των μετρικών επικύρωσης αλλά και σύμφωνα με την ποιότητα θερμοκρασιακού διαχωρισμού των υπό αξιολόγηση ομάδων. Αυξάνοντας τον αριθμό των ομάδων, δηλαδή το k, πέρα από τις υποδείξεις των μετρικών επικύρωσης, με απώτερο σκοπό να επιτευχθεί ο κατάλληλος θερμοκρασιακός διαχωρισμός των ομάδων, παρατηρήσαμε ότι οι μετρικές επικύρωσης υπολείπονται ως ένα βαθμό σε επιθυμητή ακρίβεια, τουλάχιστον όσον αφορά το συγκεκριμένο πρόβλημα.

Στη συνέχεια παρουσιάζουμε τη μεθοδολογία των εφαρμογών ομαδοποίησης ημερήσιων καμπυλών Σ.Π.Φ καθώς και της εξαγωγής των "Προφίλ" Φορτίου. Το εν λόγω παράδειγμα αφορά τα ετήσια δεδομένα του χρονικού παραθύρου "01/01/2019" εώς "29/02/2020" στο οποίο αναφερόμαστε ως έτος "2019". Αντίστοιχα πράττουμε για τα υπόλοιπα ετήσια αρχεία της βάσης δεδομένων DB_2.

Εισάγουμε τις απαραίτητες βιβλιοθήκες με τις παρακάτω εντολές :

```
import pandas as pd
from tslearn.clustering import TimeSeriesKMeans
from tslearn.clustering import silhouette_score
from matplotlib import pyplot as plt
from yellowbrick.cluster import SilhouetteVisualizer
from sklearn.metrics import calinski_harabasz_score
from sklearn.metrics import davies_bouldin_score
from statistics import mean
```

Διαβάζουμε το ετήσιο αρχείο που περιέχει τις ημερήσιες καμπύλες Σ.Π.Φ των ευρωπαϊκών χωρών.

```
df = pd.read_csv("366DLC_25CTY_2019Year.csv", sep="\t", encoding="utf_16")
```

Στο αρχείο αυτό, όπως είναι γνωστό από το 2^{ο} στάδιο προεπεξεργασίας, ενδέχεται να υπάρχουν ελλιπείς τιμές οι οποίες υποδεικνύονται από την ειδική τιμή "NaN". Συγκεκριμένα, με βάση τον πίνακα 4.2 το εν λόγω αρχείο περιέχει 1062 ελλιπείς τιμές οι οποίες αποτελούν το 0,48% των συνολικών δεδομένων του αρχείου. Ο αλγόριθμος ομαδοποίησης K-Means όμως δεν δύναται να εκτελεστεί σε δεδομένα που περιέχουν ελλιπείς τιμές, δηλαδή πεδία που περιέχουν τιμές "NaN". Συνεπώς, για την εκτέλεση του αλγορίθμου απαιτείται να απαλείψουμε διανύσματα εισόδου (ημερήσιες καμπύλες Σ.Π.Φ) τα οποία περιέχουν ελλιπείς τιμές .

Σε επόμενο βήμα απαιτείται η κανονικοποίηση των δεδομένων για τους λόγους που ήδη έχουμε αναφέρει με τη μέθοδο Min Max Scaling με παραμέτρους min = 0 και max = 1.

Στο σημείο αυτό, τα δεδομένα είναι κατάλληλα προετοιμασμένα για να αποτελέσουν την είσοδο στον αλγόριθμο ομαδοποίησης. Για τις εφαρμογές ομαδοποίησης, σύμφωνα με το διάγραμμα ροής του σχήματος 4.13, απαιτείται η αξιολόγηση των ομαδοποιήσεων που εκτελεί ο αλγόριθμος για κάθε διαφορετική ρύθμιση των υπερπαραμέτρων του.

Συνεπώς, εκτελούμε πολλαπλές φορές τον αλγόριθμο ομαδοποίησης με διαφορετικούς δυνατούς συνδιασμούς των υπερπαραμέτρων του και υπολογίζουμε τις βέλτιστες υπερπαραμέτρους με βάση τα αποτελέσματα των μετρικών επικύρωσης (CVI). Η διαδικασία αυτή πραγματοποιείται ξεχωριστά για κάθε χώρα.

Εφόσον έχουμε υπολογίσει τις βέλτιστες υπερπαραμέτρους, τις διοχετεύουμε ως ορίσματα στον αλγόριθμο ομαδοποίησης και τον "τρέχουμε" για να μας επιστρέψει τις υποψήφιες τελικές ομάδες.

Στη συνέχεια, εφόσον ακολουθούμε την ετήσια ανάλυση, αξιολογούμε την ομαδοποίηση που προέκυψε με βάση τον θερμοκρασιακό διαχωρισμό των ομάδων. Για το λόγο αυτό, υλοποιήσαμε τη συνάρτηση $Count\_Days\_Months()$ η οποία δέχεται ως όρισμα το πλαίσιο δεδομένων που περιέχει τις ημερήσιες καμπύλες Σ.Π.Φ της αντίστοιχης ομάδας και τυπώνει στην οθόνη πόσες ημερήσιες καμπύλες αντιστοιχούν σε κάθε μήνα του έτους. Στο σημείο αυτό, προς αποφυγή παρεξηγήσεων, αναφέρουμε για μία ακόμα φορά ότι το ετήσιο χρονικό παράθυρο αντιστοιχεί στο διάστημα που ορίζεται από τη $1^{η}$ Μαρτίου του έτους μέχρι το τέλος Φεβρουαρίου του επόμενου έτους. Συνεπώς, η "Μέρα_1" ανήκει στον μήνα Μάρτιο και η "Μέρα_365" στο μήνα Φεβρουάριο. Με βάση τα αποτελέσματα της εν λόγω συνάρτησης μπρούμε να αποφανθούμε ως προς την ποιότητα του θερμοκρασιακού διαχωρισμού των ομάδων. Για παράδειγμα, εφόσον μια ομάδα περιέχει σημαντικό αριθμό καλοκαιρινών ημερήσιων καμπυλών Σ.Π.Φ καθώς επίσης και σημαντικό αριθμό χειμερινών καμπυλών, τότε η ομαδοποίηση κρίνεται ακατάλληλη και συμβουλευόμαστε τα γραφήματα των μετρικών αξιολόγησης για το επόμενο καλύτερο k που υποδεικνύουν.

Συνεπώς, ομαδοποιούμε ξανά τα δεδομένα με τον αλγόριθμο TimeSeriesKMeans θέτοντας στη συνάρτηση του τα νέα υποψήφια ορίσματα. Η διαδικασία εκτελείται επαναληπτικά μέχρι να παρατηρήσουμε έναν επαρκή θερμοκρασιακό διαχωρισμό μεταξύ των ομάδων. Σύμφωνα με τα προαναφερθέντα, ενδέχεται να καταφύγουμε σε λύσεις οι οποίες κρίνονται ως μη "βέλτιστες" σύμφωνα με τα αποτελέσματα των μετρικών επικύρωσης ομαδοποίησης.

Στο σημείο αυτό έχουμε αποφασίσει και επιλέξει τις υπερπαραμέτρους που ορίζουν την πλέον κατάλληλη ομαδοποίηση των δεδομένων. Συνεπώς, εκτελούμε τον αλγόριθμο TimeSeriesKMeans της βιβλιοθήκης tslearn με τα αντίστοιχα ορίσματα και μας επιστρέφει τις τελικές ομάδες .

Όπως υποδεικνύει το διάγραμμα ροής του σχήματος 4.13, στη περίπτωση της ετήσιας ανάλυσης, η εξαγωγή των "Προφίλ" Φορτίου κάθε χώρας θα πραγματοποιηθεί μόνο στις αντίστοιχες ομάδες οι οποίες περιέχουν στατιστικά σημαντικό αριθμό μελών. Επίσης , όσον αφορά την εποχική ανάλυση, η εξαγωγή των "Προφίλ" Φορτίου κάθε χώρας θα πραγματοποιηθεί στην αντίστοιχη ομάδα που περιέχει τα περισσότερα μέλη.

Για την εξαγωγή των "Προφίλ" Φορτίου από τις πλέον κατάλληλες ομάδες υλοποιήσαμε τη συνάρτηση $Load\_Profile\_Extraction()$ η οποία δέχεται τέσσερα ορίσματα :

i. Τον "MapCode" κωδικό της χώρας.

ii. Το λεξικό που περιέχει ως κλειδιά τα ονόματα των διανυσμάτων εισόδου (δηλαδή των ημερήσιων καμπυλών Σ.Π.Φ.) και ως τιμές τις αντίστοιχες μέγιστες τιμές των αρχικών μη κανονικοποιημένων διανυσμάτων.

iii. Το λεξικό που περιέχει ως κλειδιά τα ονόματα των διανυσμάτων εισόδου (ημερήσιες καμπύλες Σ.Π.Φ) και ως τιμές τις αντίστοιχες ελάχιστες τιμές των αρχικών μη κανονικοποιημένων διανυσμάτων.

iv. Το πλαίσιο δεδομένων που περιέχει τις ημερήσιες καμπύλες Σ.Π.Φ της αντίστοιχης ομάδας.

Η συνάρτηση $Load\_Profile\_Extraction()$ αρχικά αποκανονικοποιεί τα δεδομένα της ομάδας εκτελώντας "MinMax_Unscaling", στη συνέχεια υπολογίζει τη μέση τιμή των αποκανονικοποιημένων μελών της ομάδας και επιστρέφει το "Προφίλ Φορτίου" ως λίστα. Επίσης, η εν λόγω συνάρτηση παρέχει το γράφημα του υπολογισμένου "Προφίλ" Φορτίου και επιστρέφει τις αποκανονικοποιημένες ημερήσιες καπύλες Σ.Π.Φ ως ένα πλαίσιο δεδομένων.

Έπειτα, ελέγχουμε τα αποτελέσματα μέσω γραφημάτων που υπερθέτουν όλες τις αποκανονικοποιημένες ημερήσιες καμπύλες Σ.Π.Φ της αντίστοιχης ομάδας καθώς και το αντίστοιχο "Προφίλ" Φορτίου με έντονη γραμμή για την διευκόλυνση της οπτικοποίησης.

Σε επόμενο βήμα, εφόσον τα αποτελέσματα είναι ικανοποιητικά, αποθηκεύουμε τα δεδομένα των "Προφίλ" Φορτίου σε ένα λεξικό με κλειδιά το όνομα που χαρακτηρίζει το αντίστοιχο "Προφίλ" Φορτίου, και τιμές την λίστα που περιέχει τις τιμές του αντίστοιχου "Προφίλ" Φορτίου.

Τέλος, εφόσον έχουμε υλοποιήσει όλη την παραπάνω διαδικασία για κάθε χώρα που εμπίπτει στο σύνολο ανάλυσης και έχουμε αποθηκεύσει στο λεξικό $diction\_lp\{\}$ όλα τα αντίστοιχα "Προφίλ" Φορτίου, δημιουργούμε ένα πλαίσιο δεδομένων που ενσωματώνει όλα τα υπολογισμένα "Προφίλ" Φορτίου ως στήλες με κατάλληλους δείκτες οι οποίοι είναι τα κλειδιά του λεξικού.

Σημειώνεται ότι στη περίπτωση που εκτελούμε εποχική ανάλυση, δεν χρησιμοποιούμε τη συνάρτηση $Count\_Days\_Months()$ καθώς η αξιολόγηση των ομαδοποιήσεων τελείται μόνο βάση των αποτελεσμάτων των μετρικών επικύρωσης ομαδοποίησης (CVI). Επίσης, όπως έχουμε αναφέρει προηγουμένως, η εξαγωγή των "Προφίλ" Φορτίου στη περίπτωση της εποχικής ανάλυσης κάθε χώρας πραγματοποιείται μόνο στις αντίστοιχες ομάδες με τα περισσότερα μέλη (Most Populated Approach – MPA).

Στο σημείο αυτό αναφέρουμε ότι κατά την εποχική ανάλυση κάθε χώρας απαιτείται αρχικά να επιλέξουμε τις αντίστοιχες ημερήσιες καμπύλες Σ.Π.Φ που ανιστοιχούν στην υπό ανάλυση εποχή.

```
df_mapcode_season = df_mapcode.loc[:, "Day_StartSeason" :
                                                    "Day_EndSeason"].copy()
```

Για παράδειγμα, υπό το χρονικό παράθυρο ανάλυσης "01/03/2019" εώς "29/02/2020" ,για την εποχή άνοιξη, όσον αφορά την περίπτωση της Ελλάδας εκτελούμε τη παρακάτω εντολή.

```
df_gr_spring = df_gr.loc[:, "Day_1" : "Day_92"]
```

και στη συνέχεια εργαζόμαστε κατά τα γνωστά.

4.4 Αποτελέσματα και Χαρακτηριστικές Καμπύλες Φορτίου

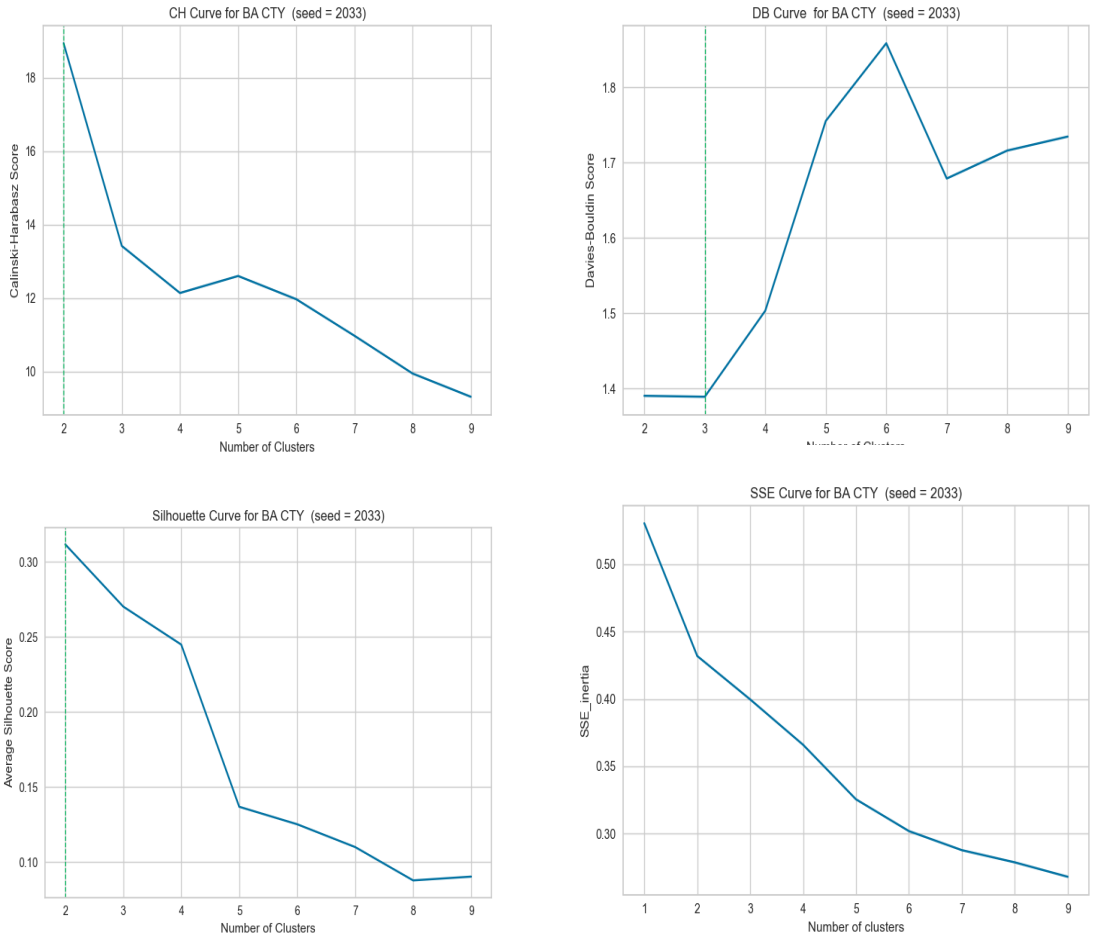
Στο κεφάλαιο αυτό παρουσιάζουμε τα αποτελέσματα των εφαρμογών ομαδοποίησης της εποχικής καθώς και της ετήσιας ανάλυσης που ορίζονται στο χρονικό παράθυρο "01/03/2019" εώς "29/02/2020" (στο οποίο αναφερόμαστε ως έτος "2019"). Τα εν λόγω αποτελέσματα θα παρουσιαστούν μέσω γραφημάτων καθώς και μέσω πινάκων με σκοπό την διευκόλυνση της οπτικοποίησης και της ερμηνείας τους.

4.4.1 Αποτελέσματα Εποχικής Ανάλυσης

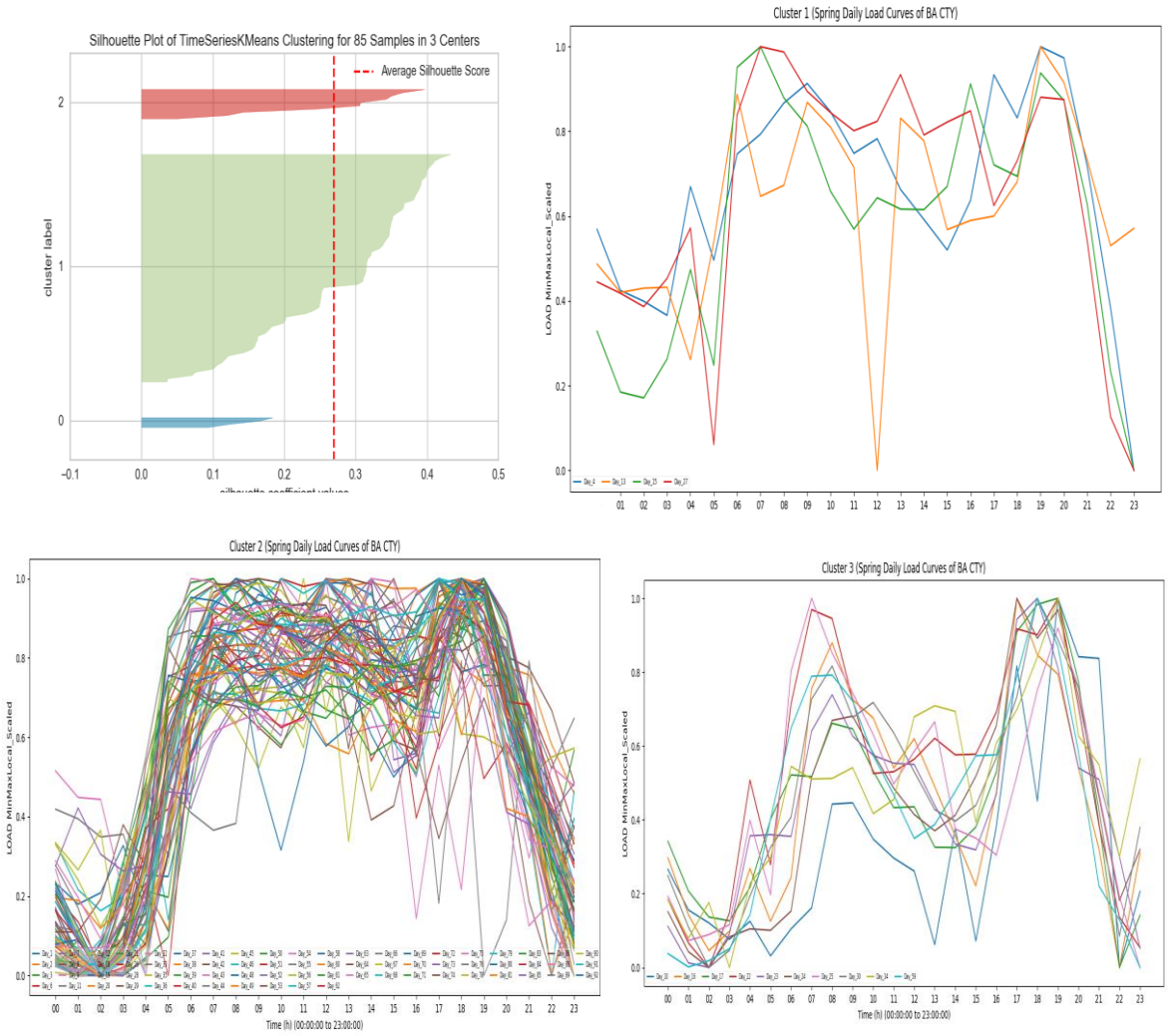
Στη συνέχεια παρουσιάζουμε κατά εποχή τα αποτελέσματα των εφαρμογών ομαδοποίησης εποχικής ανάλυσης (με βάση τις συνθήκες φόρτισης) κάθε ευρωπαϊκής χώρας που εμπίπτει στο σύνολο ανάλυσης. Επίσης παρουσιάζουμε τα αντίστοιχα "Προφίλ" Φορτίου.

- **Άνοιξη ("01/03/2019" εώς "31/05/2019")**

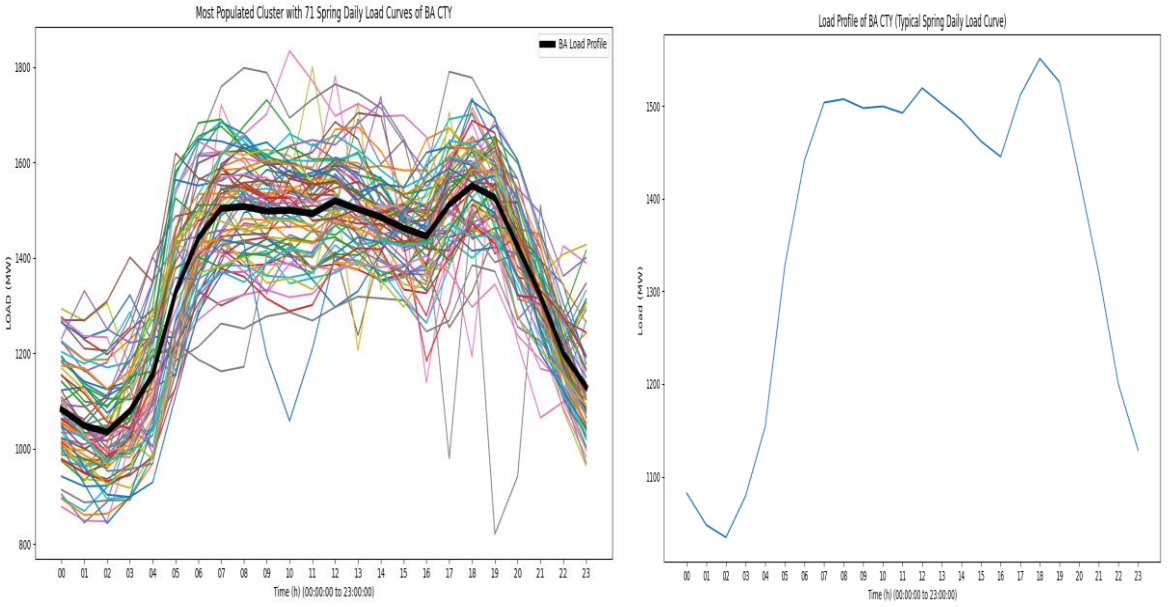
1) **Βοσνία – Ερζεγοβίνη (BA)**



Σχήμα 4.14 : (BA) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

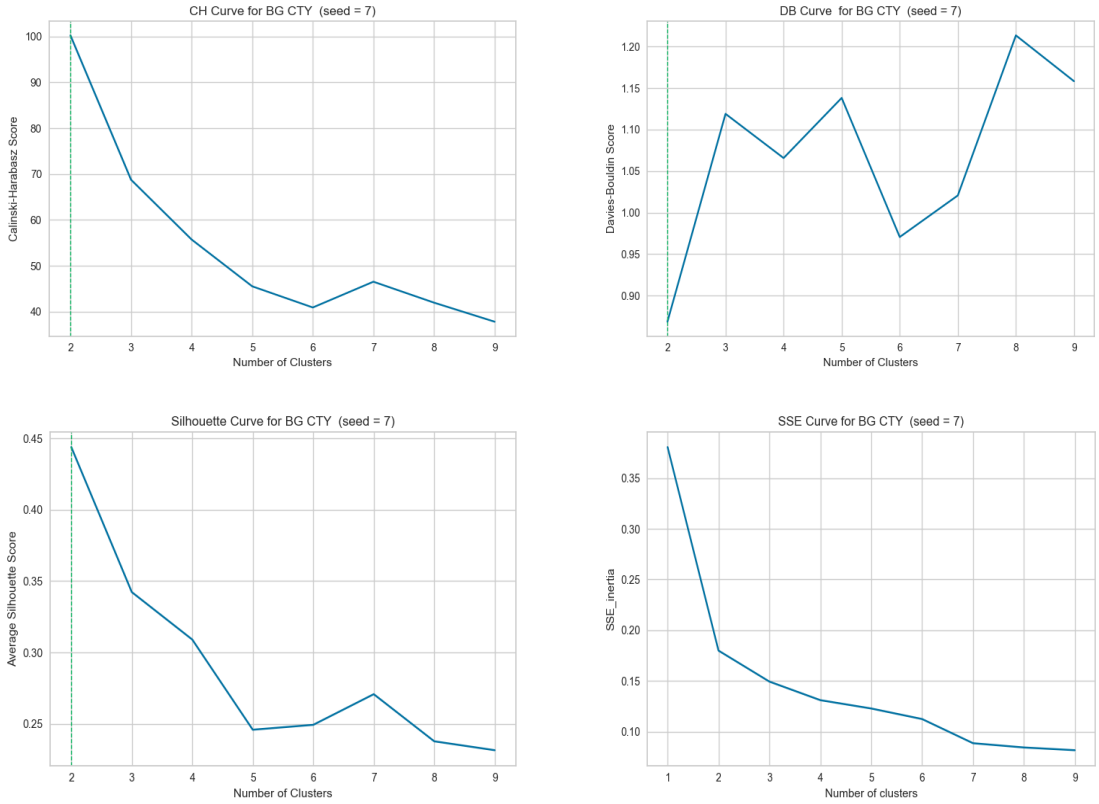


Σχήμα 4.15 : (BA) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

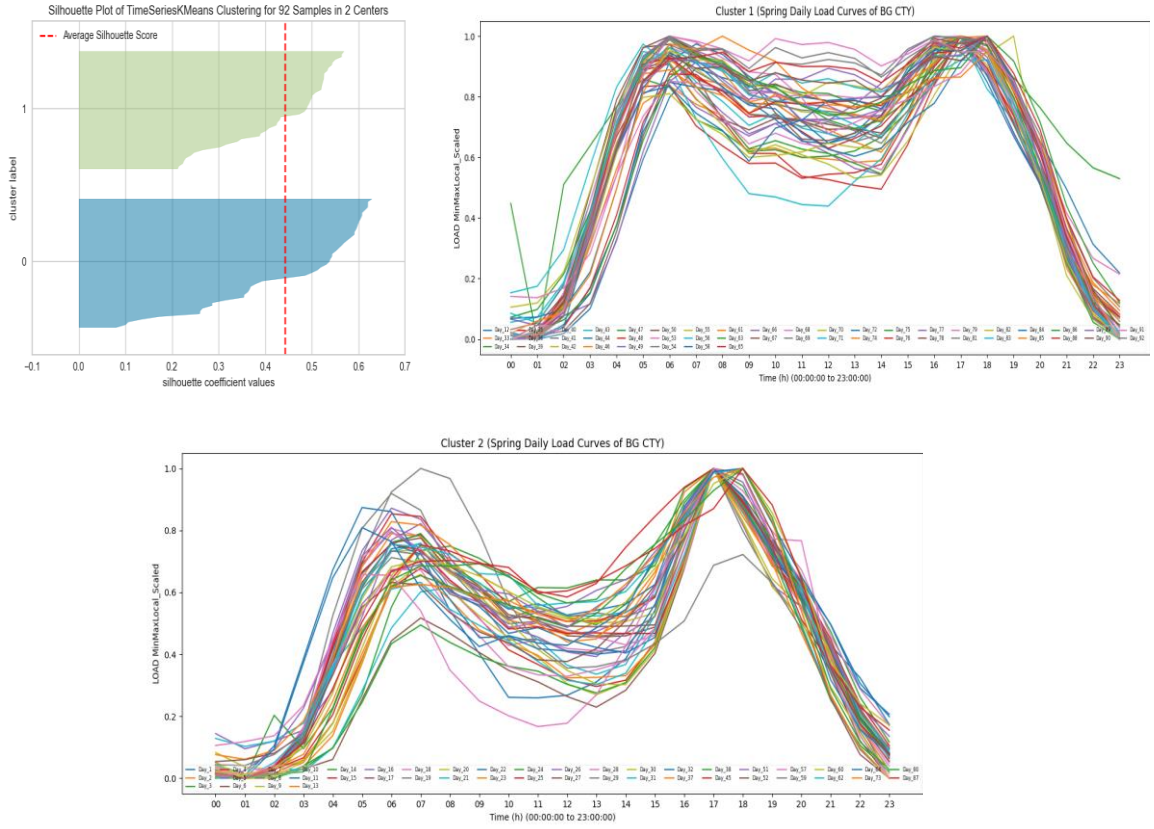


Σχήμα 4.16 : Προφίλ Φορτίου Άνοιξης (BA).

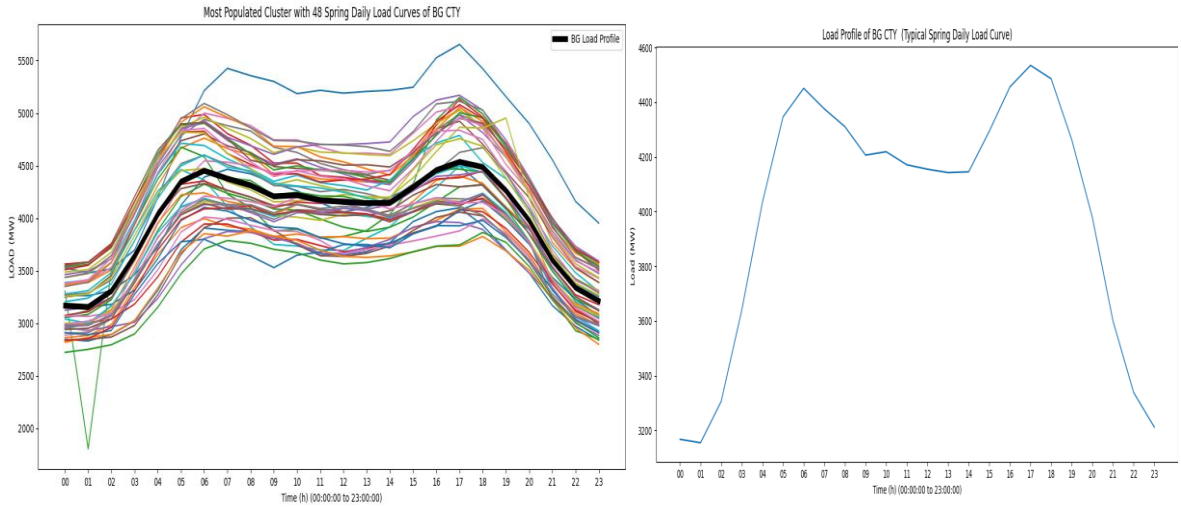
2) Βουλγαρία (BG)



Σχήμα 4.17 : (BG) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

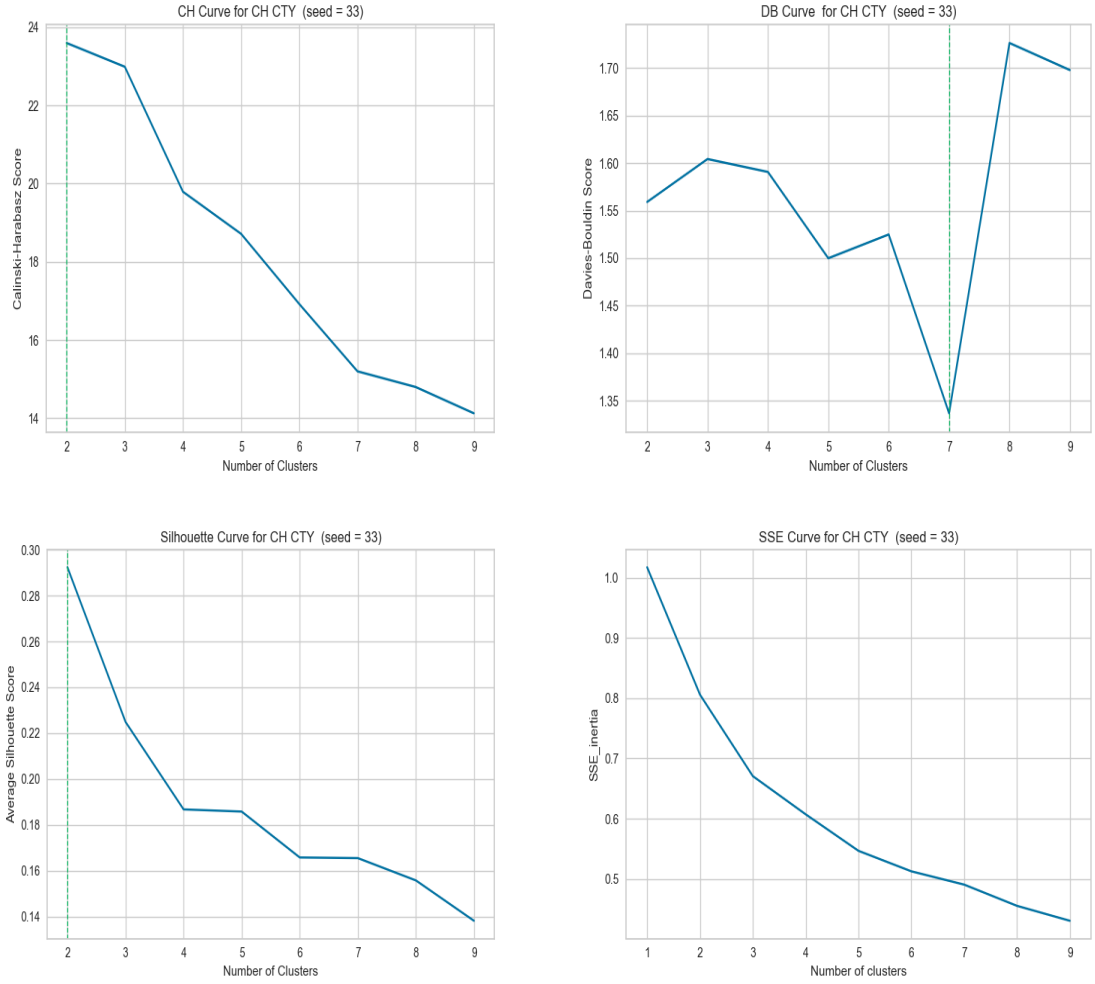


Σχήμα 4.18 : (BG) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

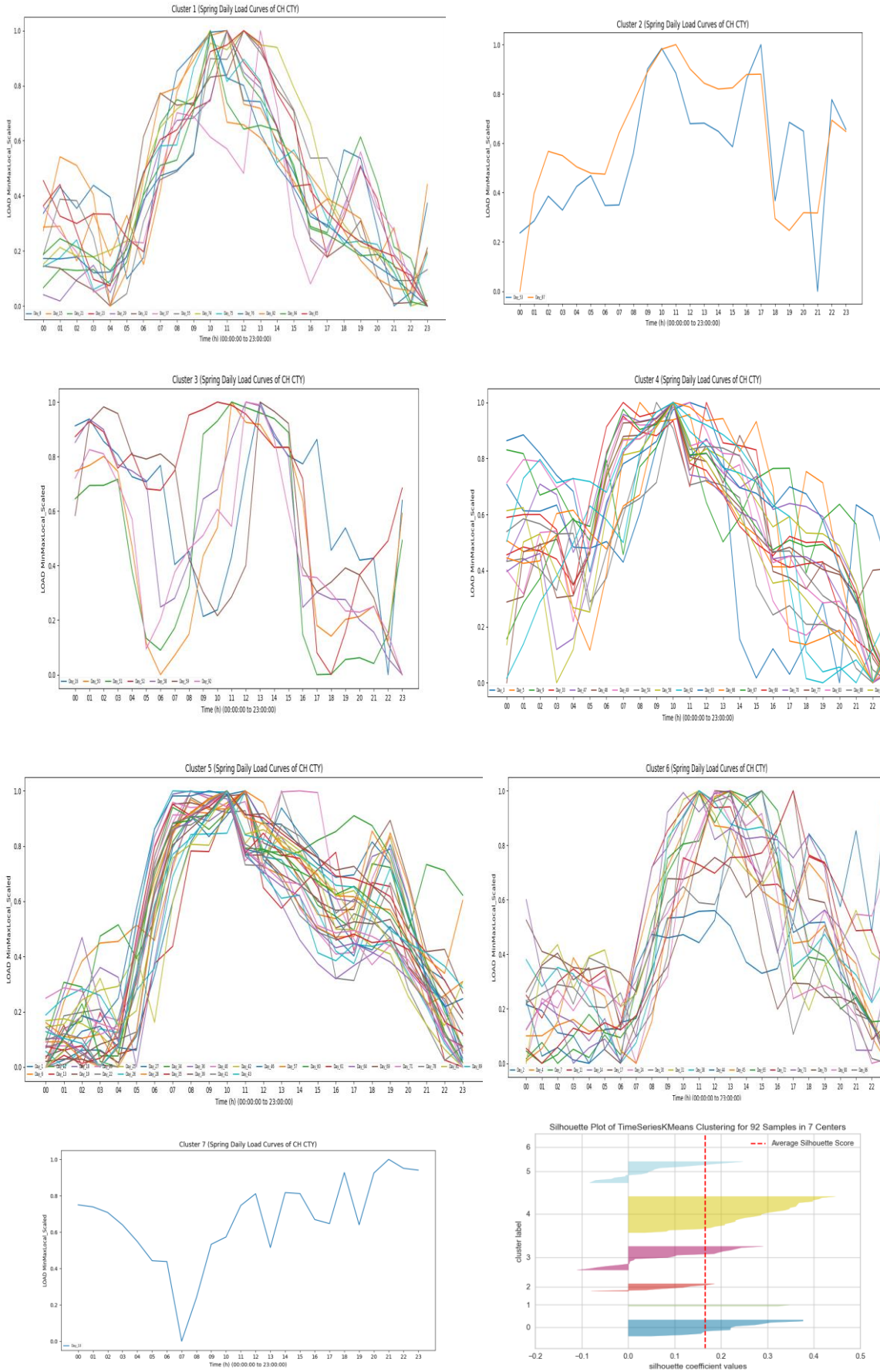


Σχήμα 4.19 : Προφίλ Φορτίου Άνοιξης (BG).

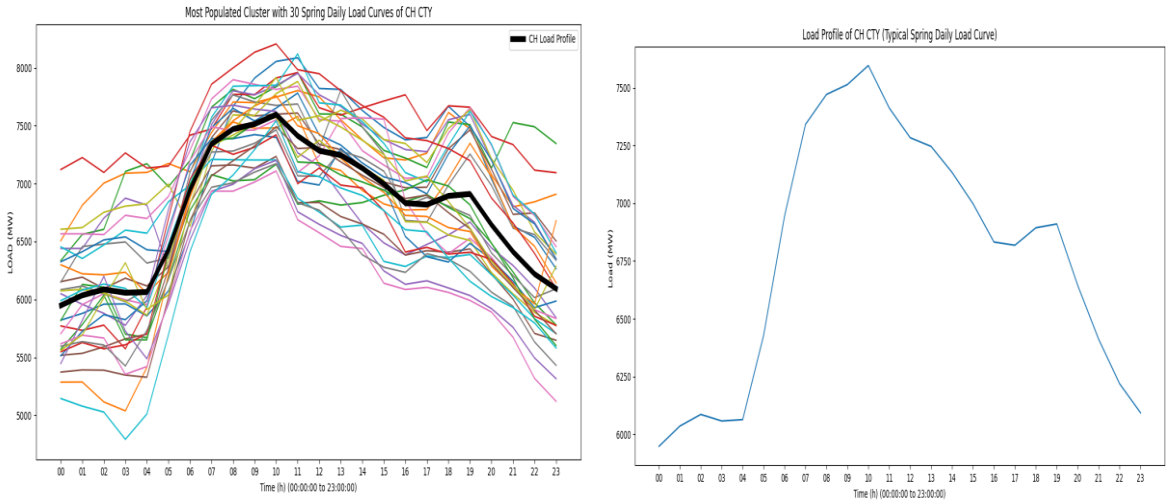
3) Ελβετία (CH)



Σχήμα 4.20 : (CH) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

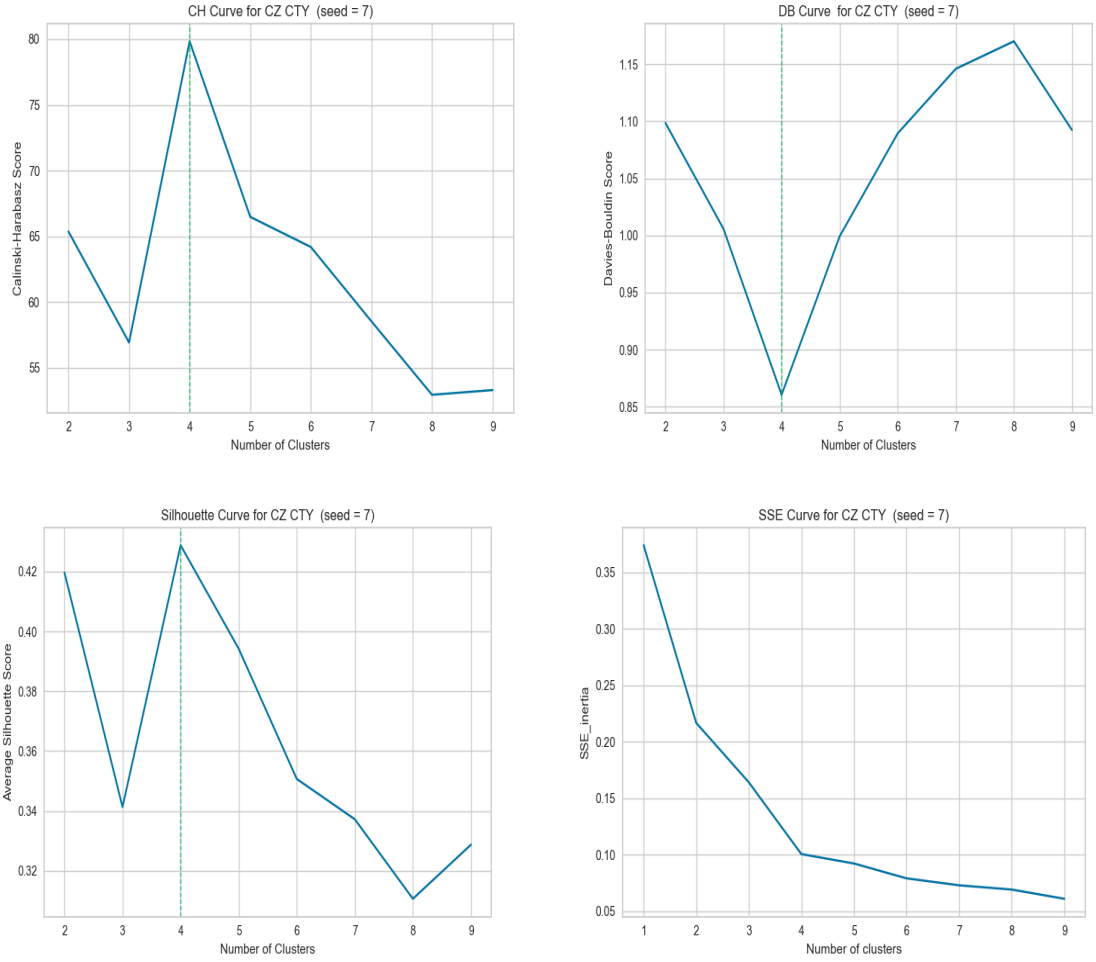


Σχήμα 4.21 : (CH) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

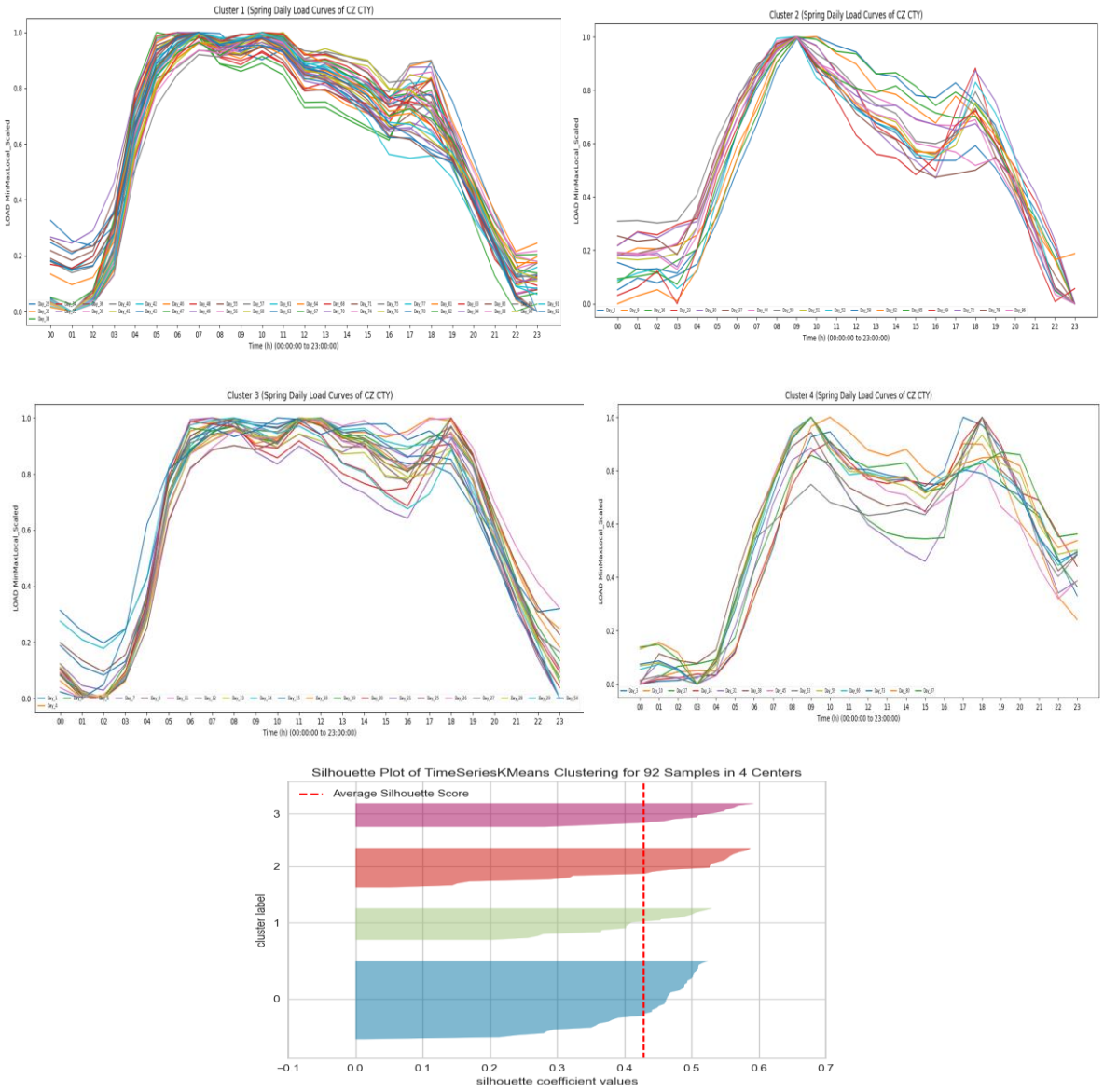


Σχήμα 4.22 : Προφίλ Φορτίου Άνοιξης (CH).

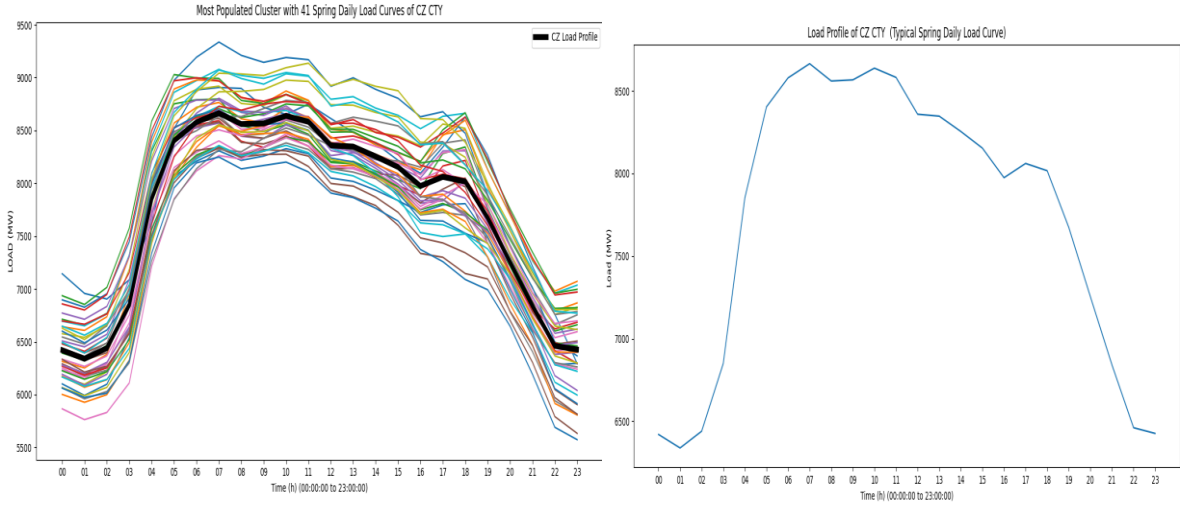
4) <u>Τσεχία (CZ)</u>



Σχήμα 4.23 : (CZ) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

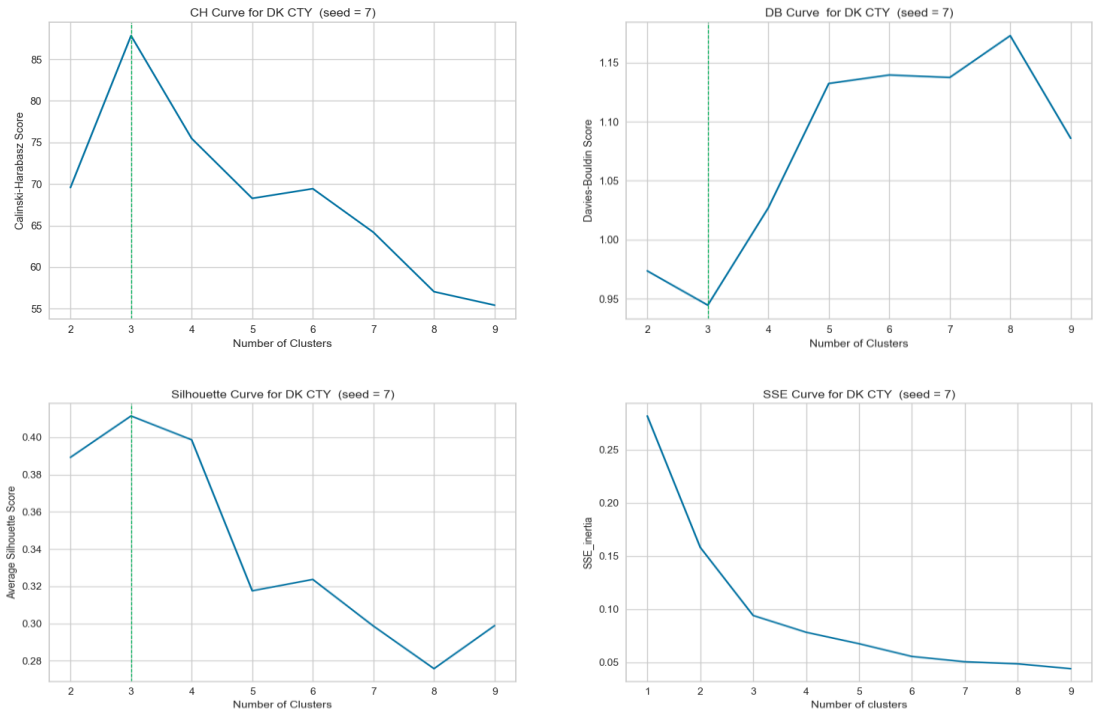


Σχήμα 4.24 : (CZ) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

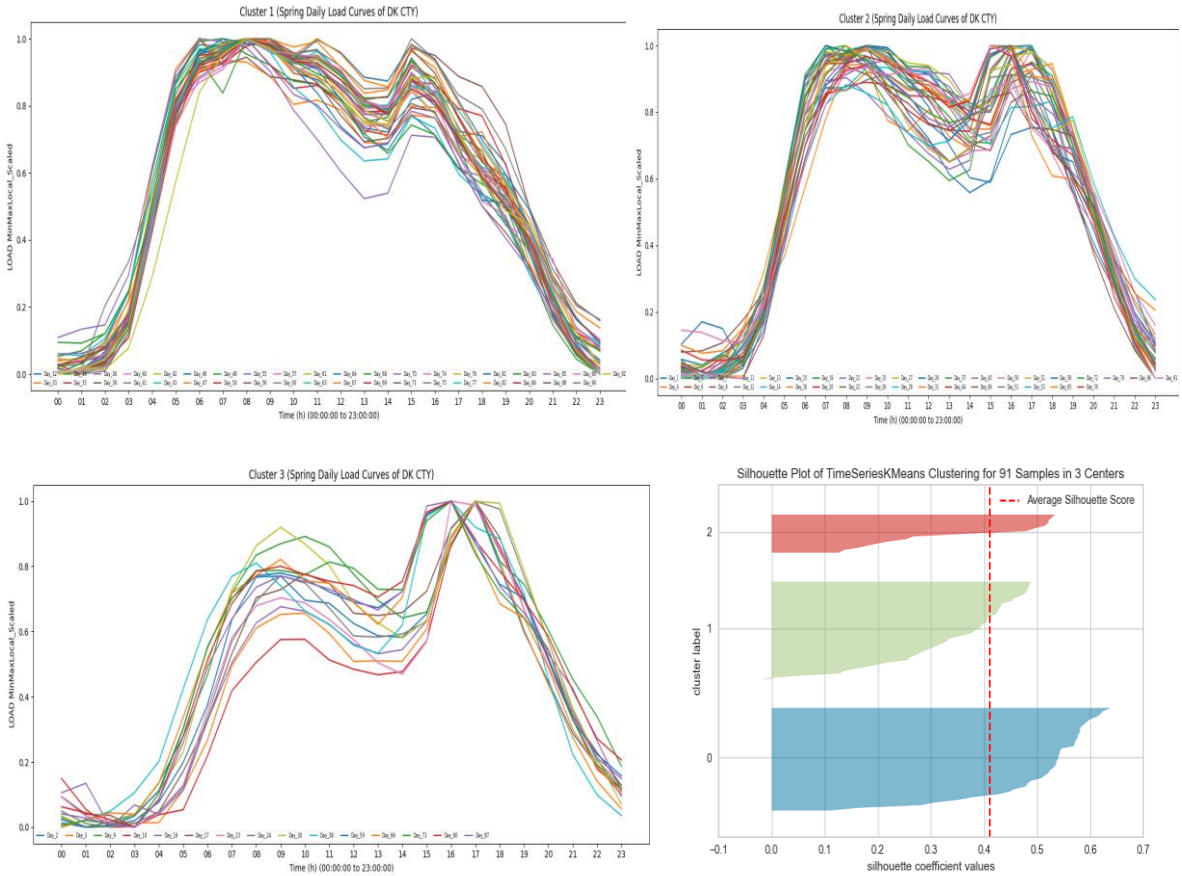


Σχήμα 4.25 : Προφίλ Φορτίου Άνοιξης (CZ).

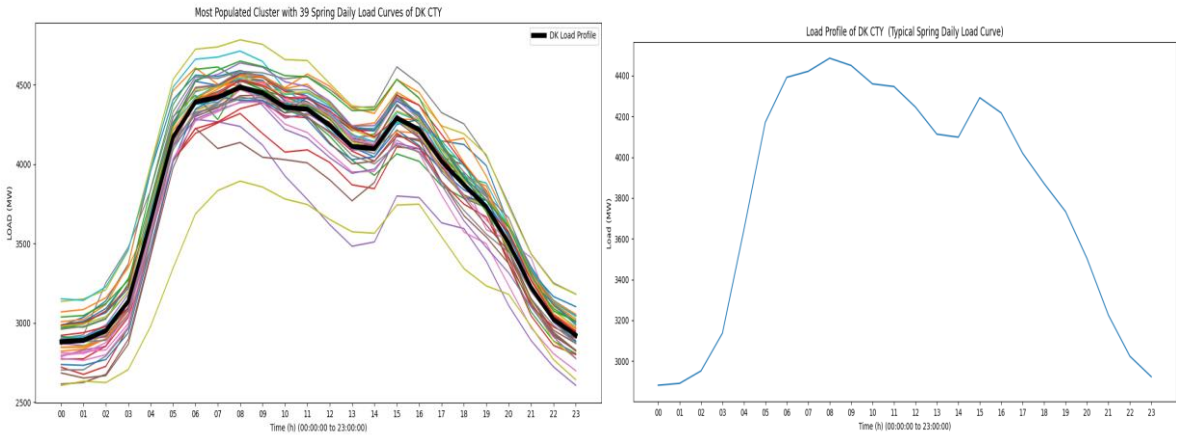
5) Δανία (DK)



Σχήμα 4.26 : (DK) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

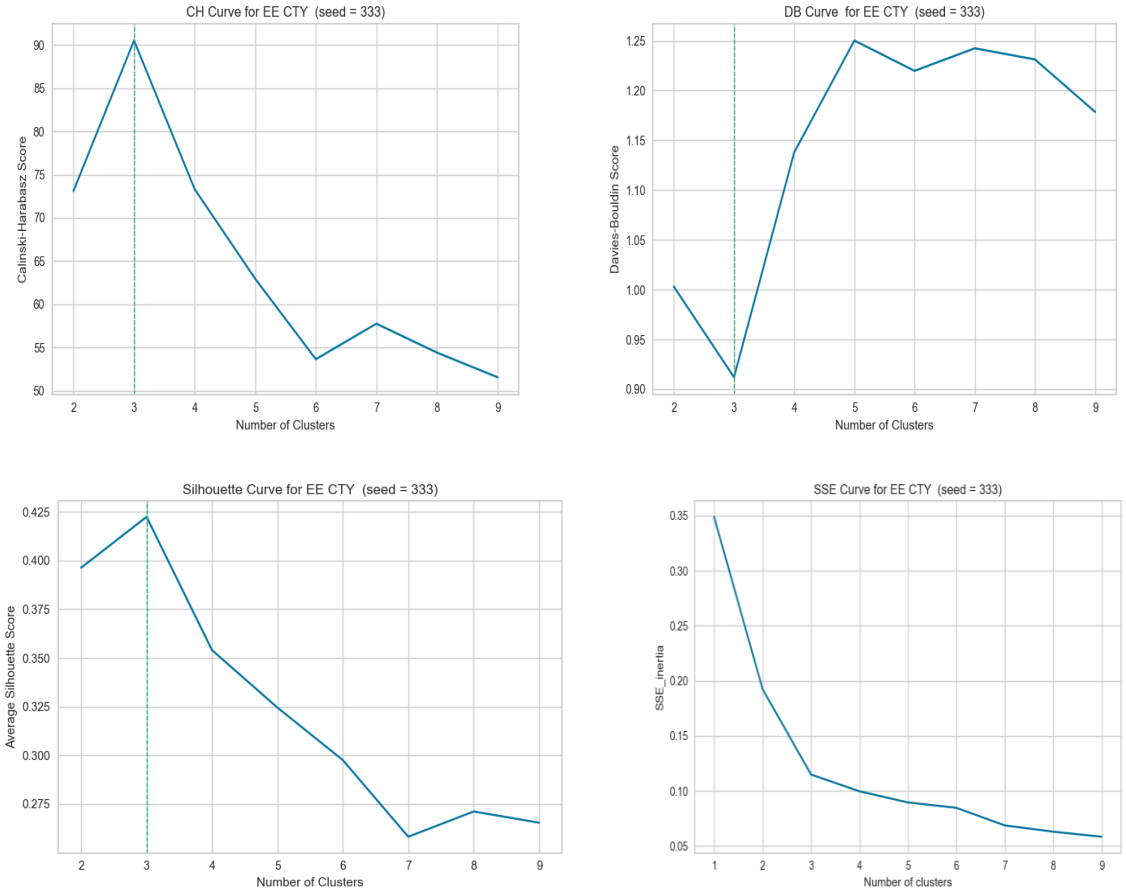


Σχήμα 4.27 : (DK) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

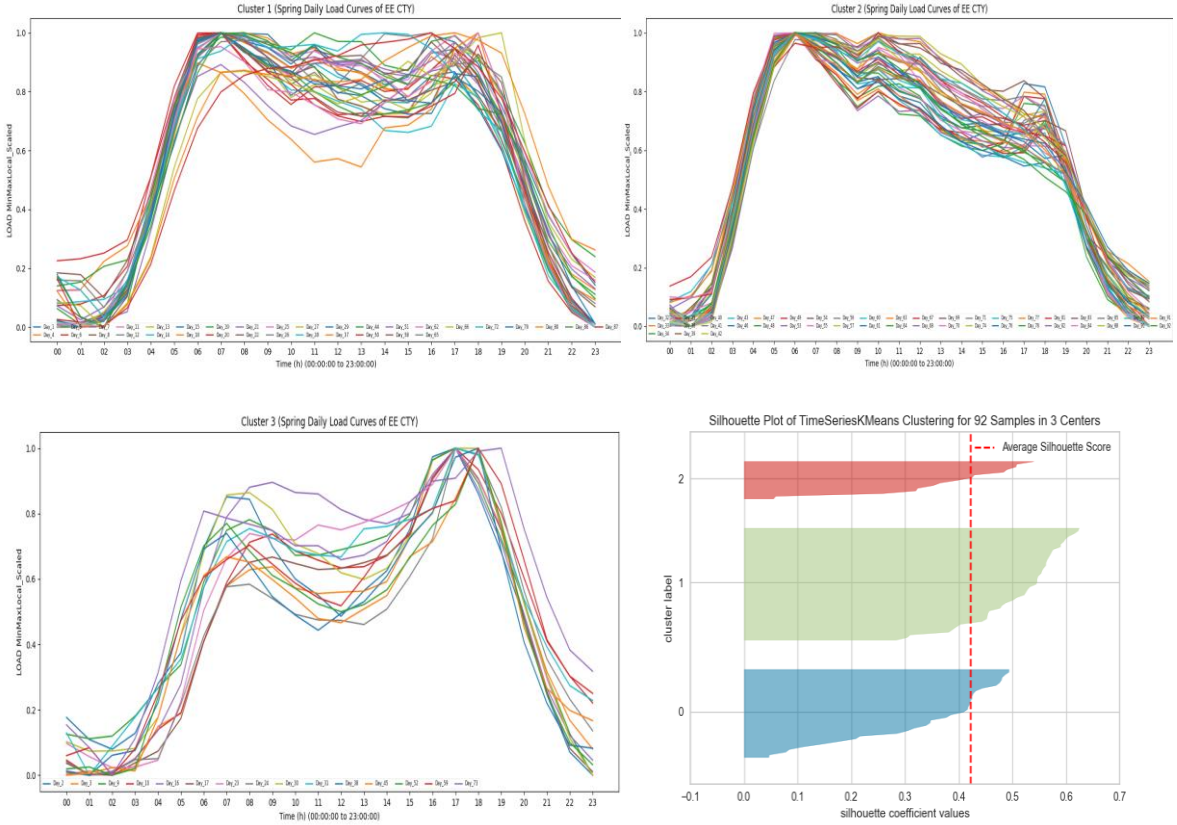


Σχήμα 4.28 : Προφίλ Φορτίου Άνοιξης (DK).

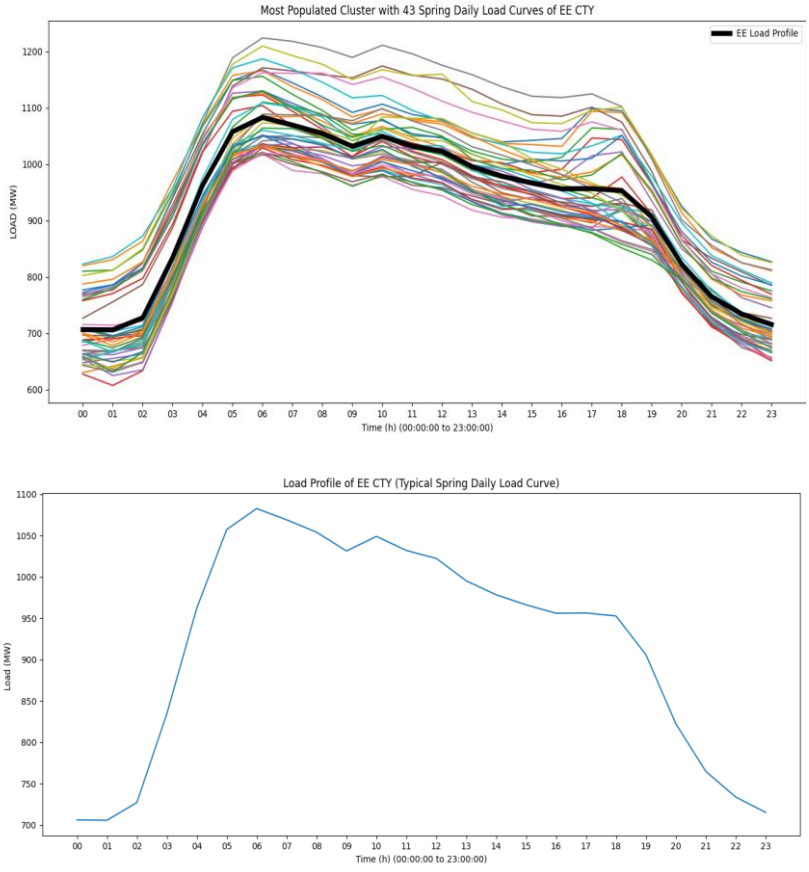
6) Εσθονία (EE)



Σχήμα 4.29 : (EE) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

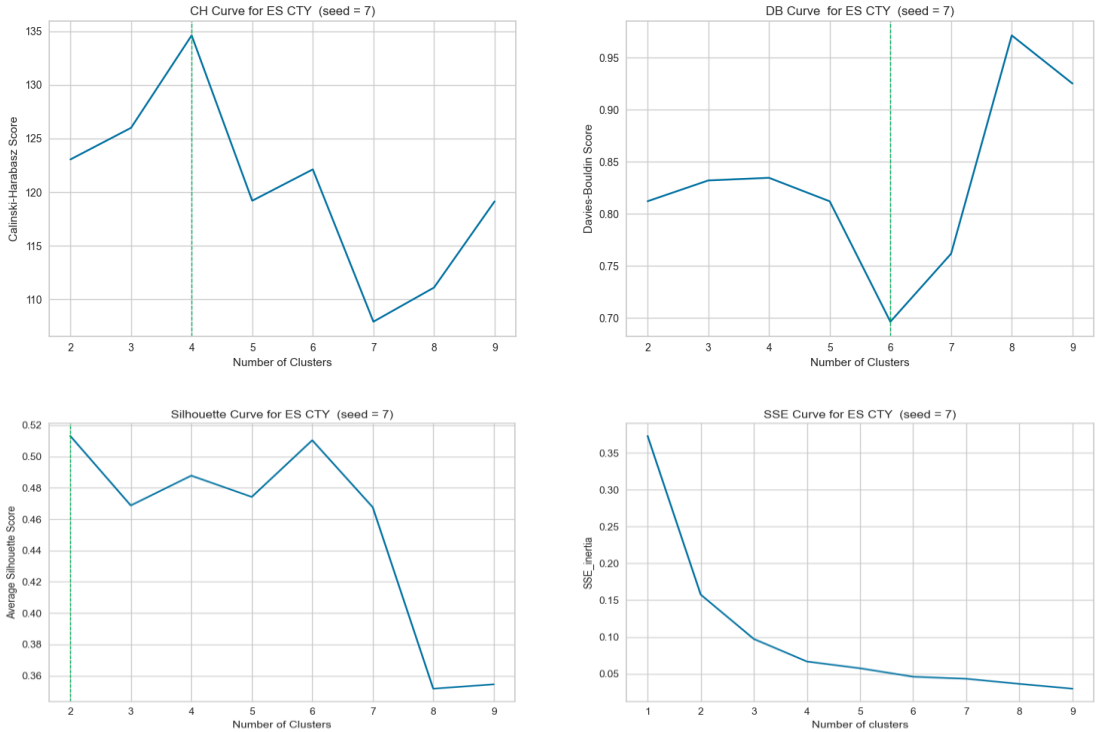


Σχήμα 4.30 : (ΕΕ) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

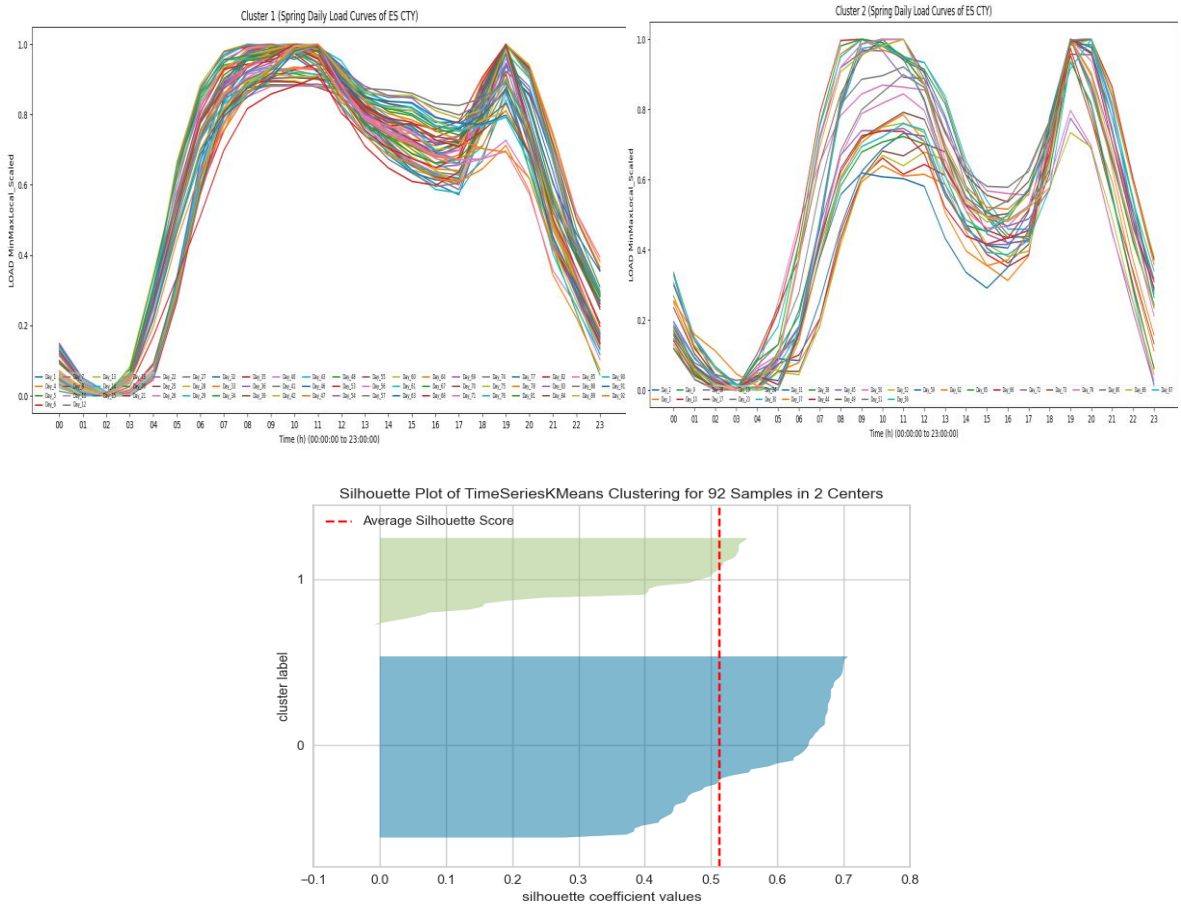


Σχήμα 4.31: Προφίλ Φορτίου Άνοιξης (ΕΕ).

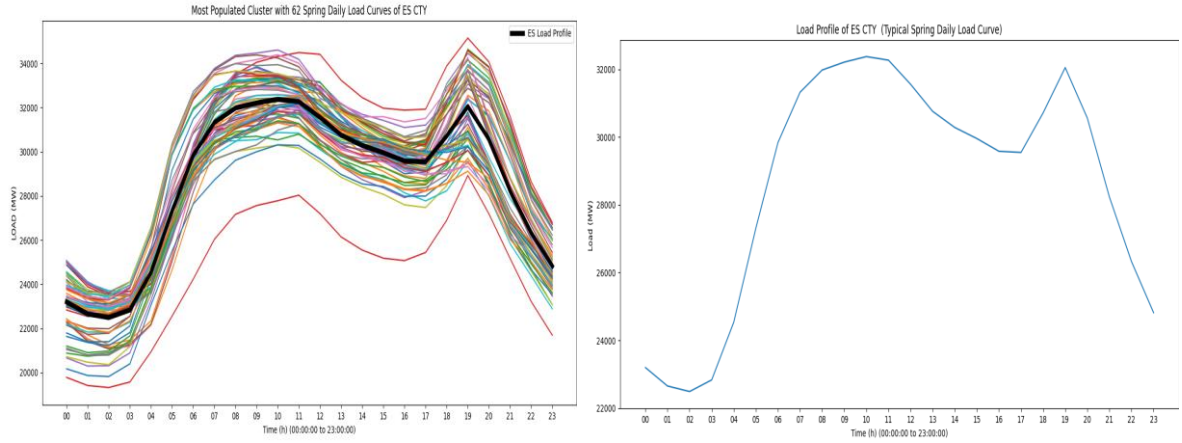
7) Ισπανία (ES)



Σχήμα 4.32 : (ES) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

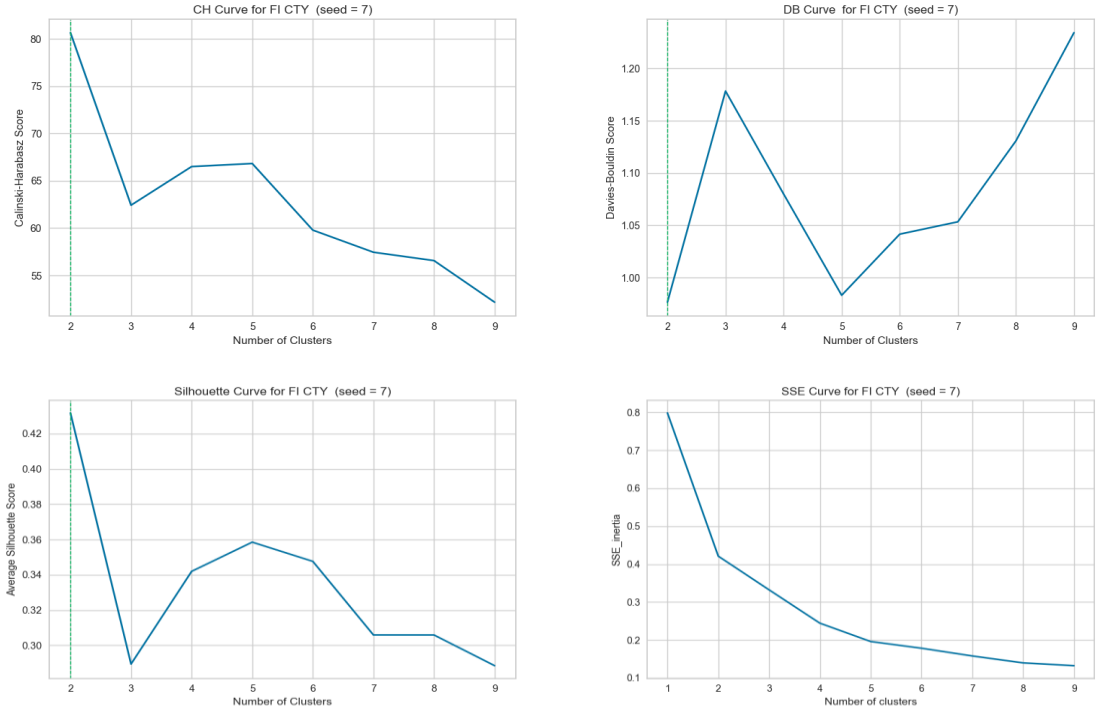


Σχήμα 4.33 : (ES) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

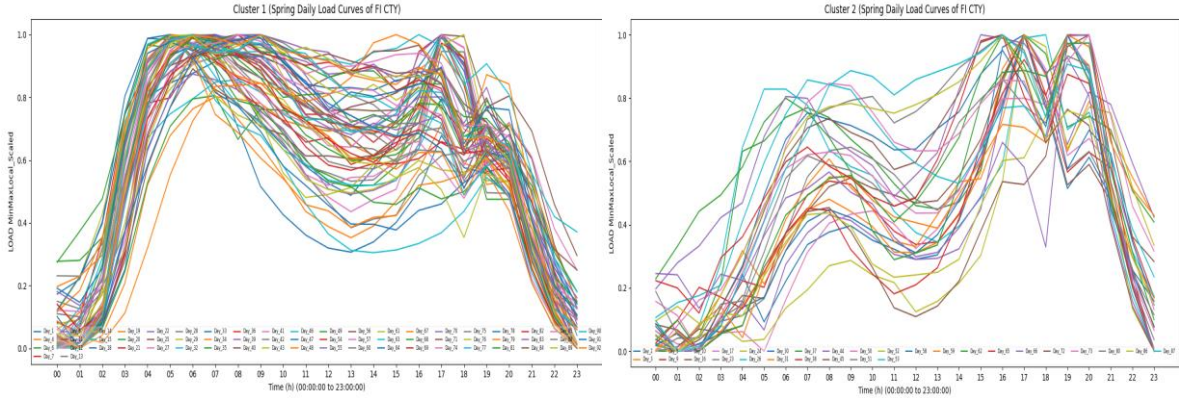


Σχήμα 4.34 : Προφίλ Φορτίου Άνοιξης (ES).

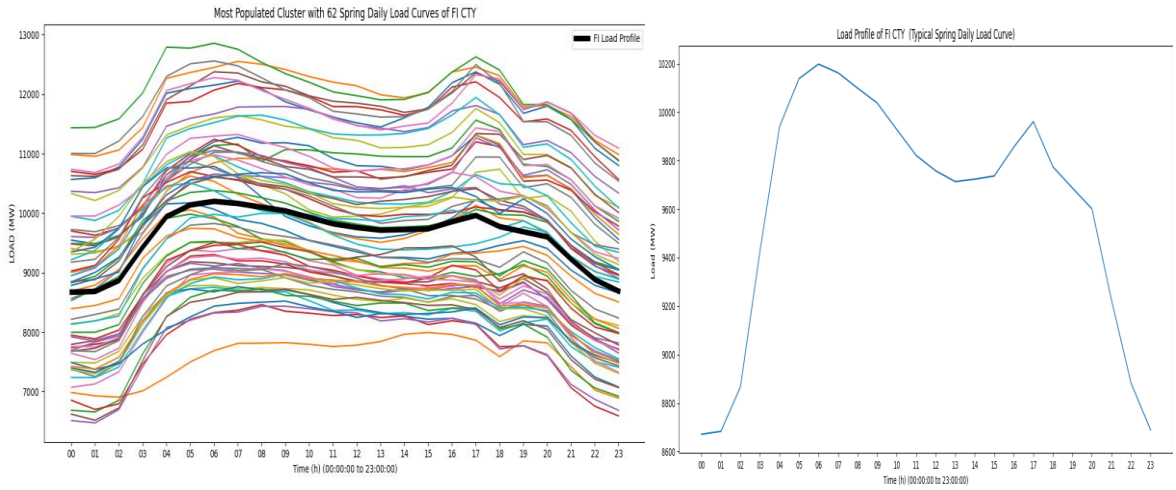
8) <u>Φινλανδία (FI)</u>



Σχήμα 4.35 : (FI) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

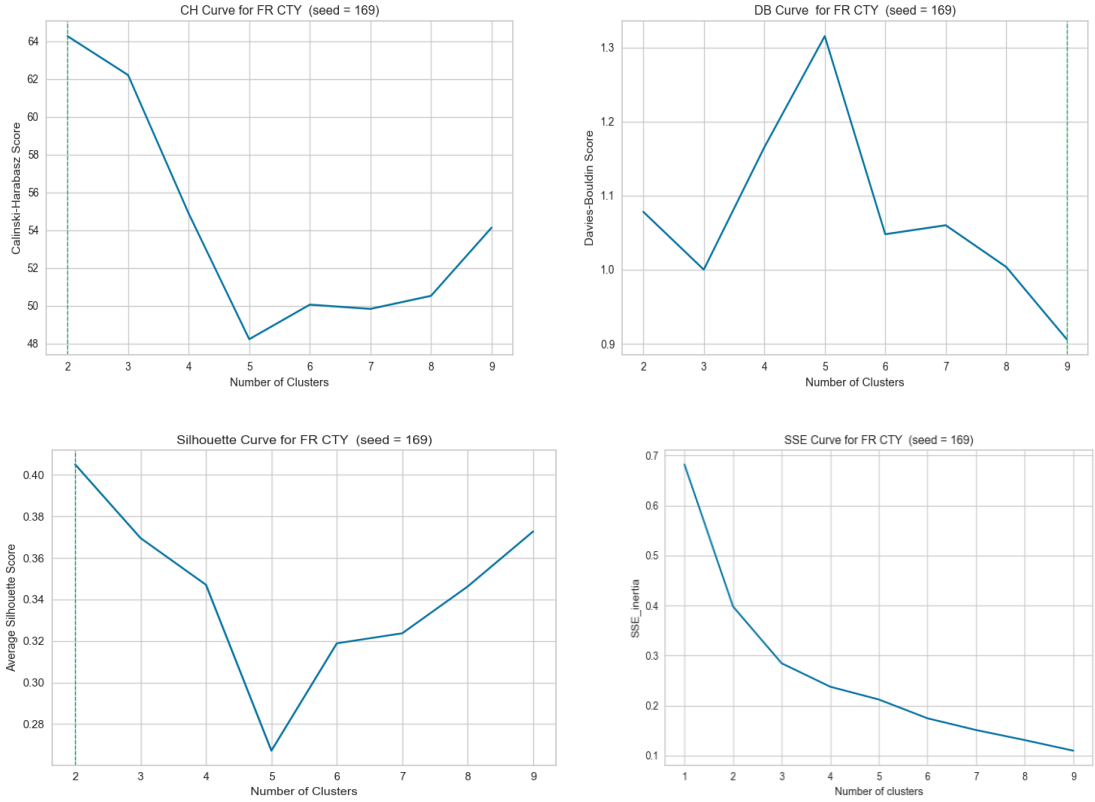


Σχήμα 4.36 : (FI) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

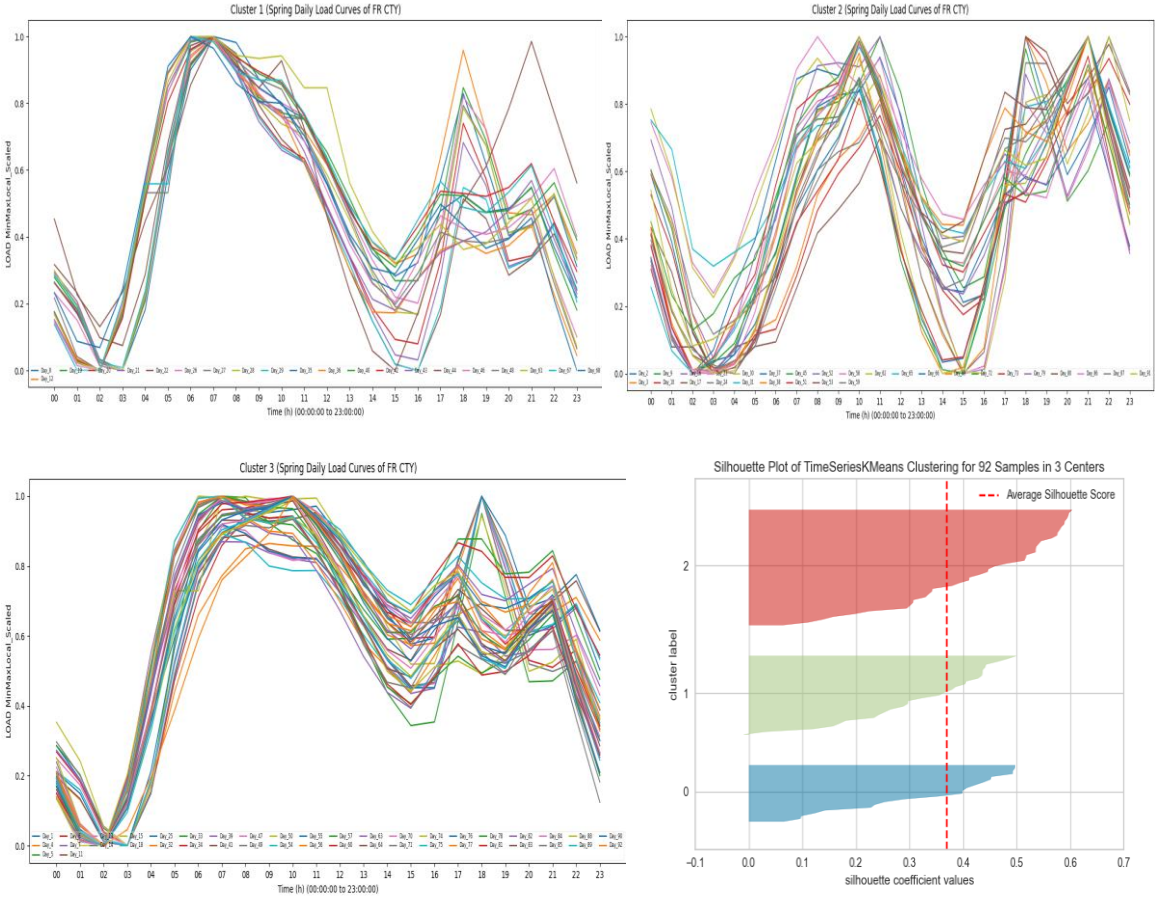


Σχήμα 4.37 : Προφίλ Φορτίου Άνοιξης (FI).

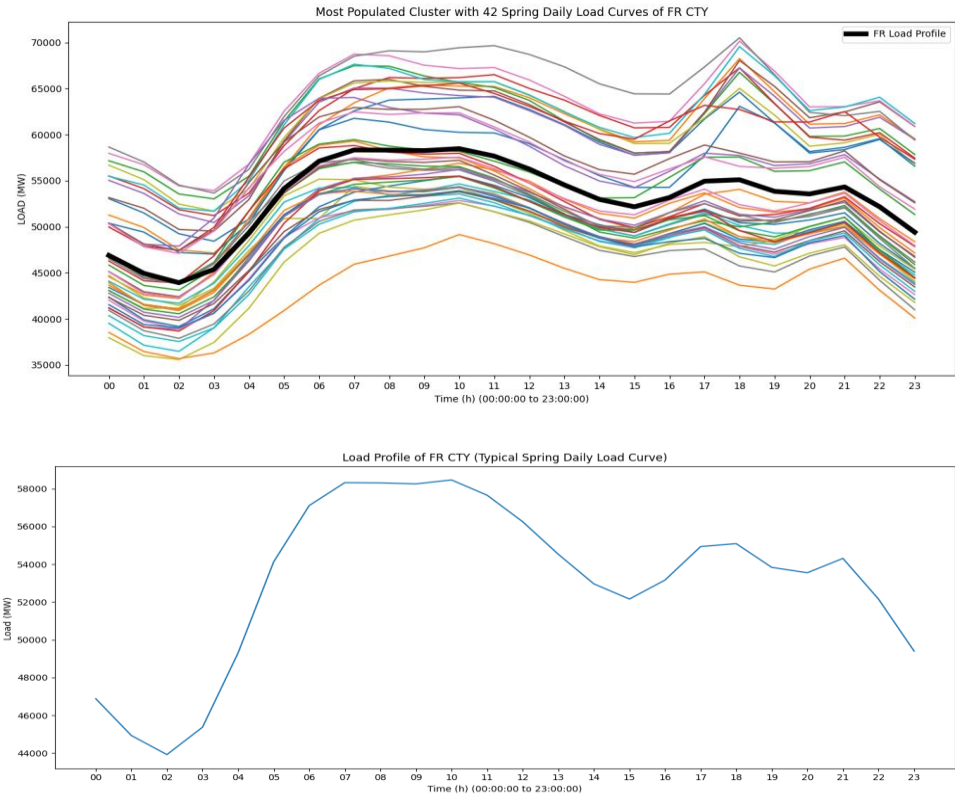
9) Γαλλία (FR)



Σχήμα 4.38 : (FR) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

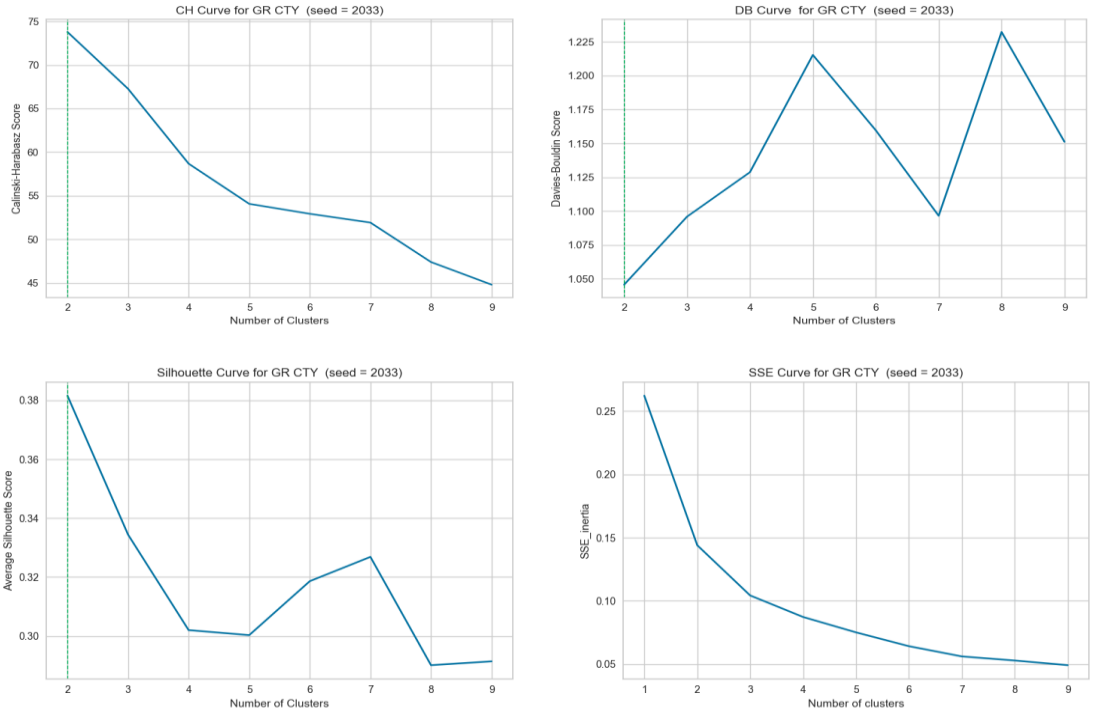


Σχήμα 4.39 : (FR) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

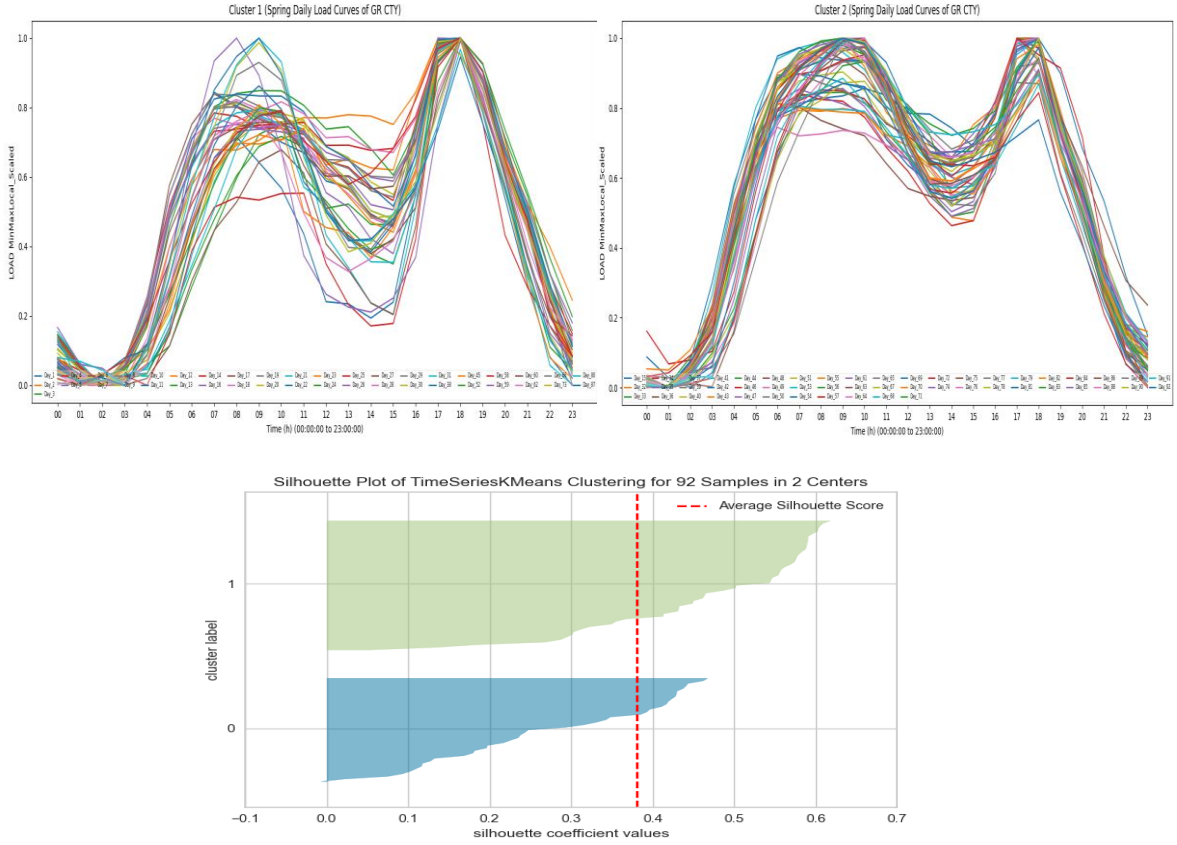


Σχήμα 4.40 : Προφίλ Φορτίου Άνοιξης (FR).

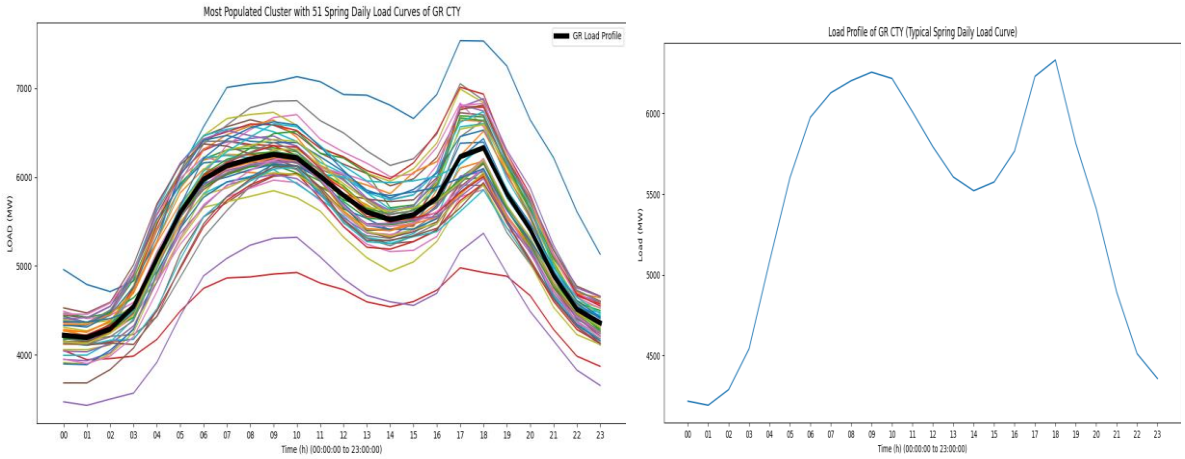
10) Ελλάδα (GR)



Σχήμα 4.41 : (GR) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

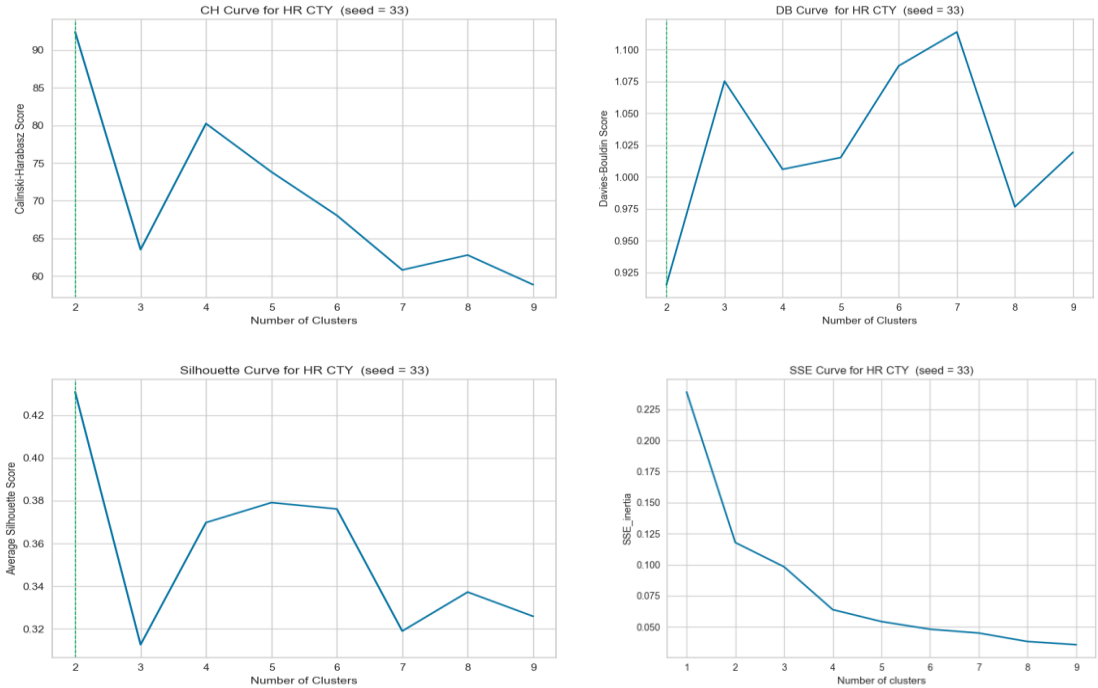


Σχήμα 4.42 : (GR) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

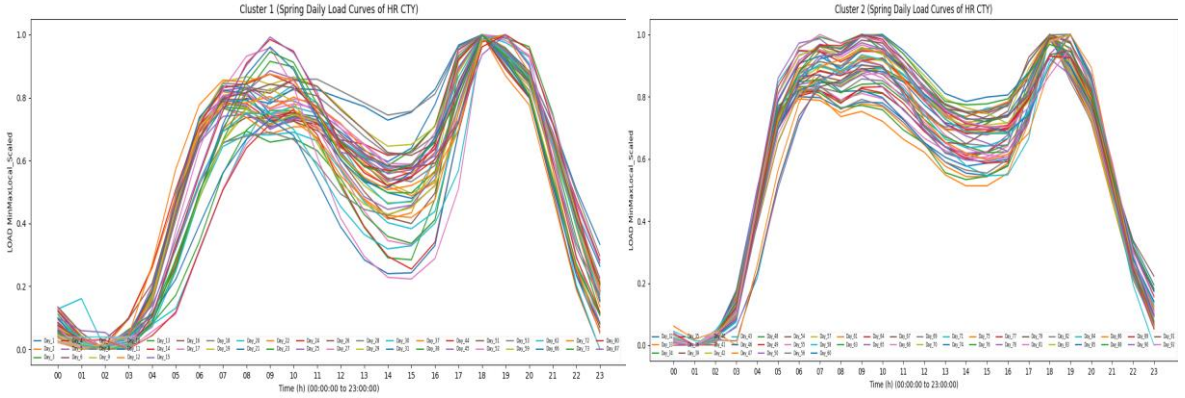


Σχήμα 4.43 : Προφίλ Φορτίου Άνοιξης (GR).

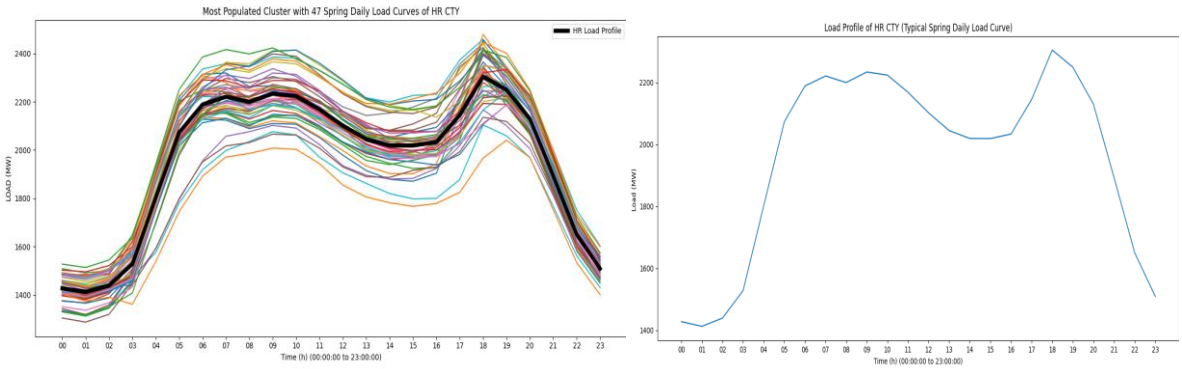
11) <u>Κροατία (HR)</u>



Σχήμα 4.44 : (HR) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

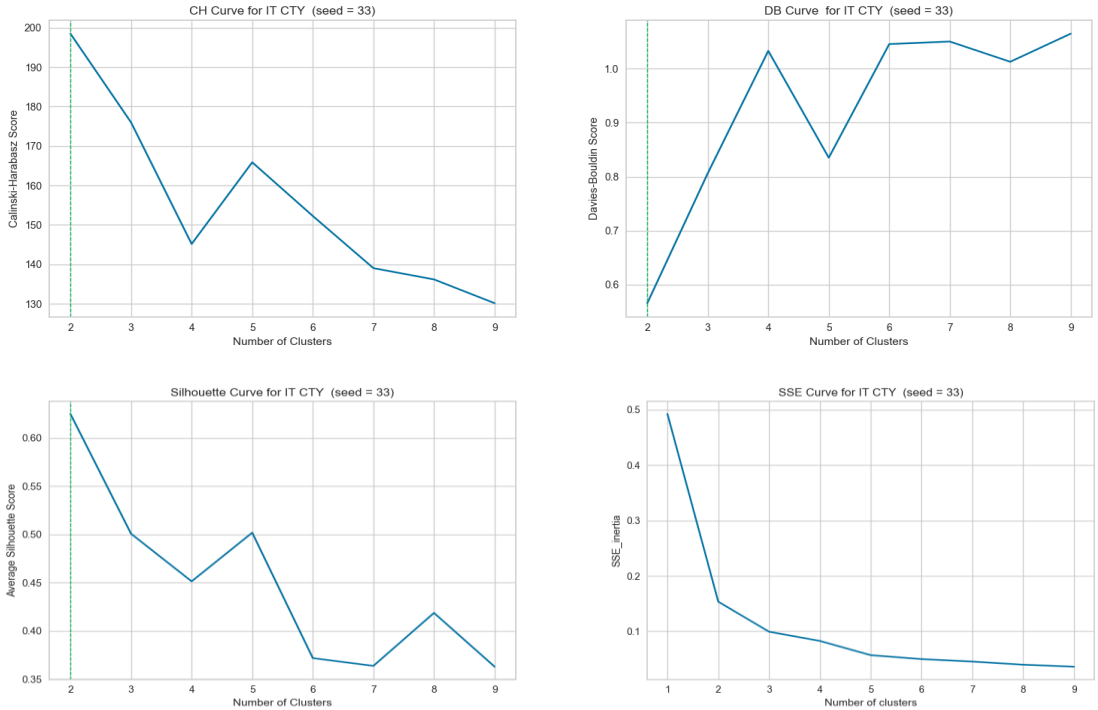


Σχήμα 4.45 : (HR) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

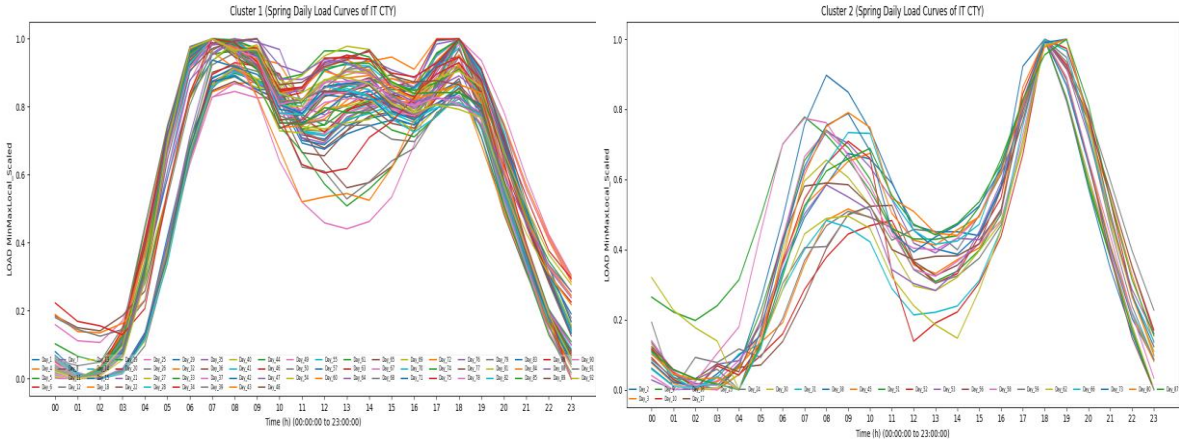


Σχήμα 4.46 : Προφίλ Φορτίου Άνοιξης (HR).

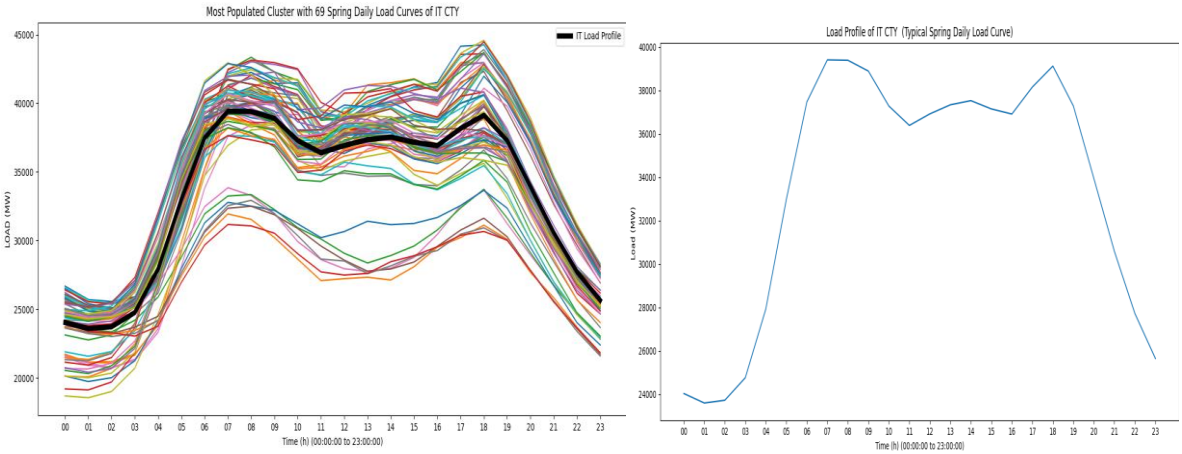
12) Ιταλία (IT)



Σχήμα 4.47 : (IT) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

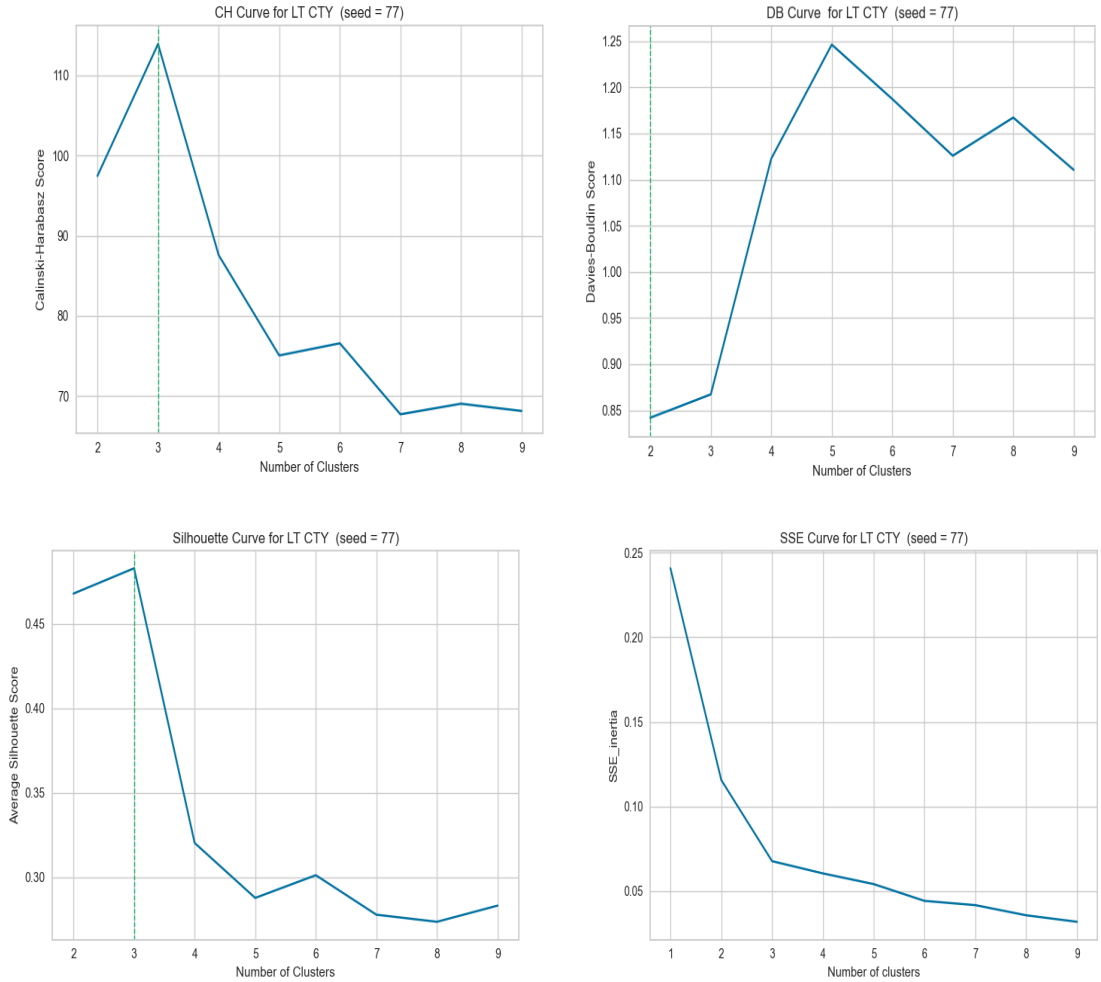


Σχήμα 4.48 : (IT) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

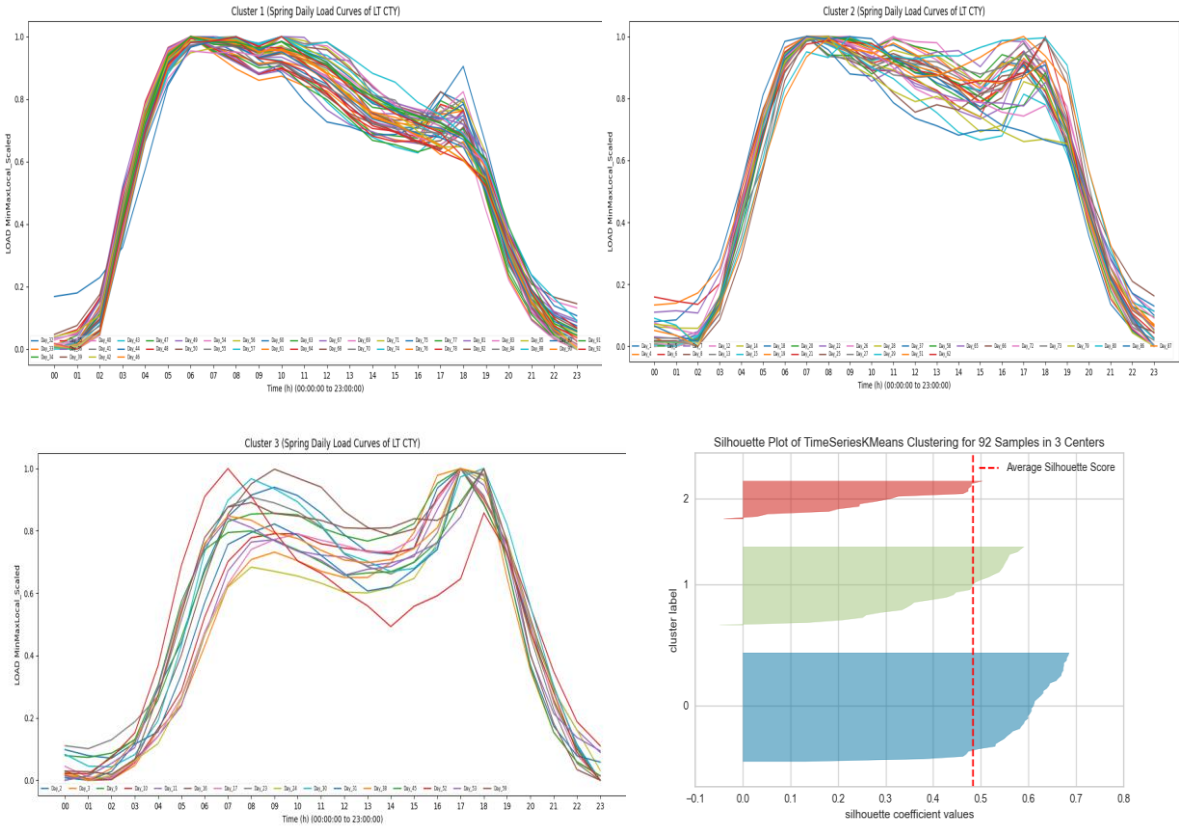


Σχήμα 4.49 : Προφίλ Φορτίου Άνοιξης (IT).

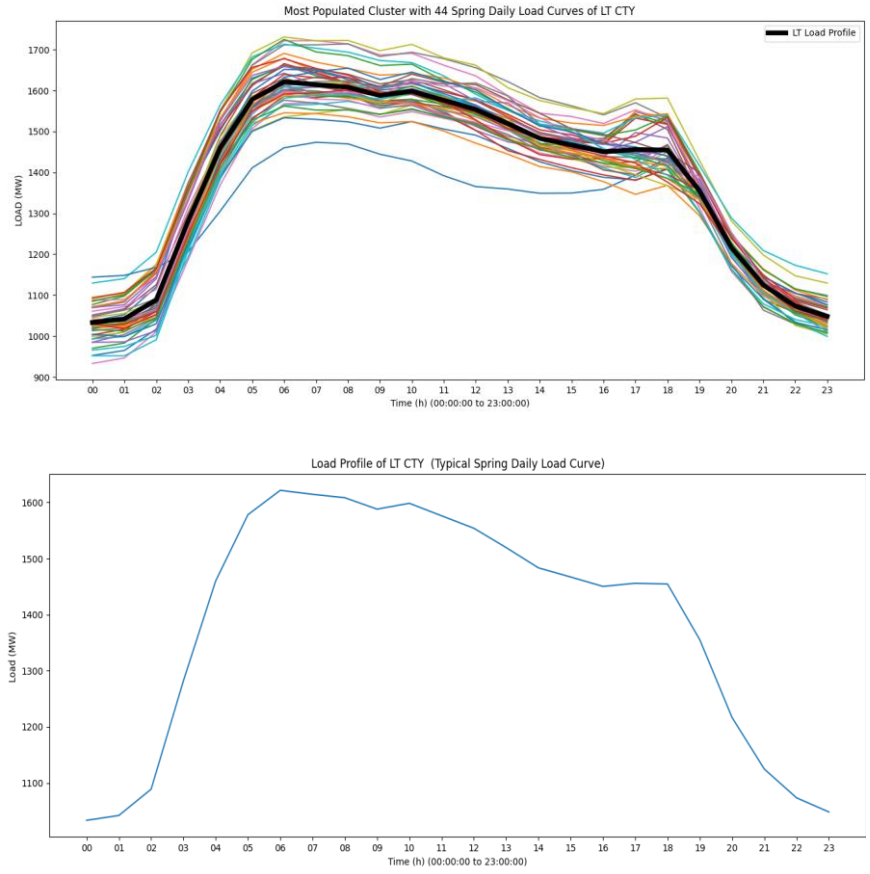
13) <u>Λιθουανία (LT)</u>



Σχήμα 4.50 : (LT) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

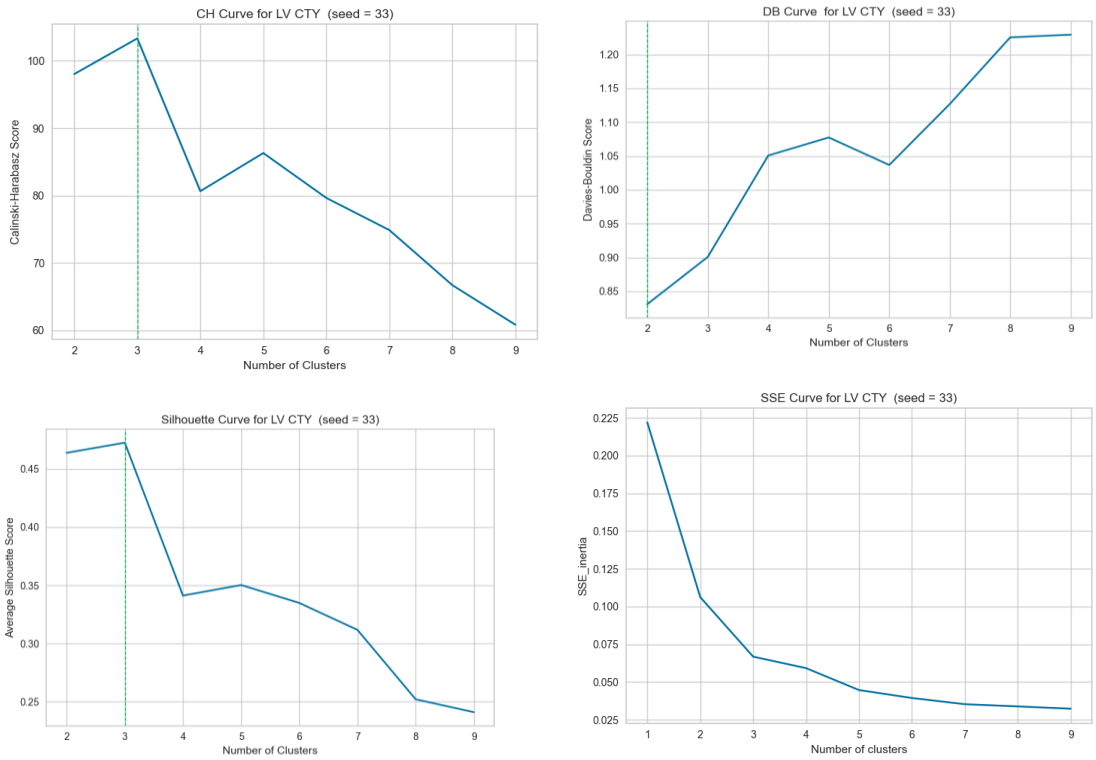


Σχήμα 4.51 : (LT) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

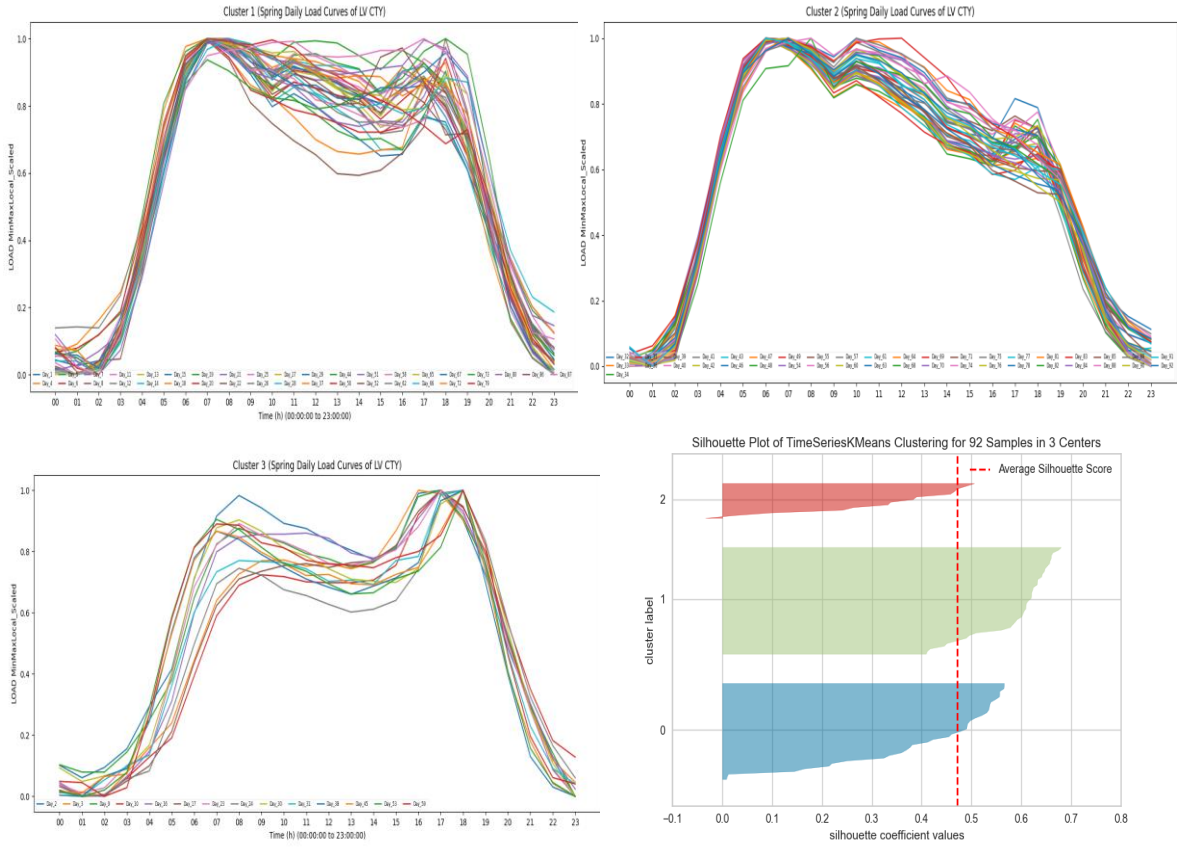


Σχήμα 4.52 : Προφίλ Φορτίου Άνοιξης (LT).

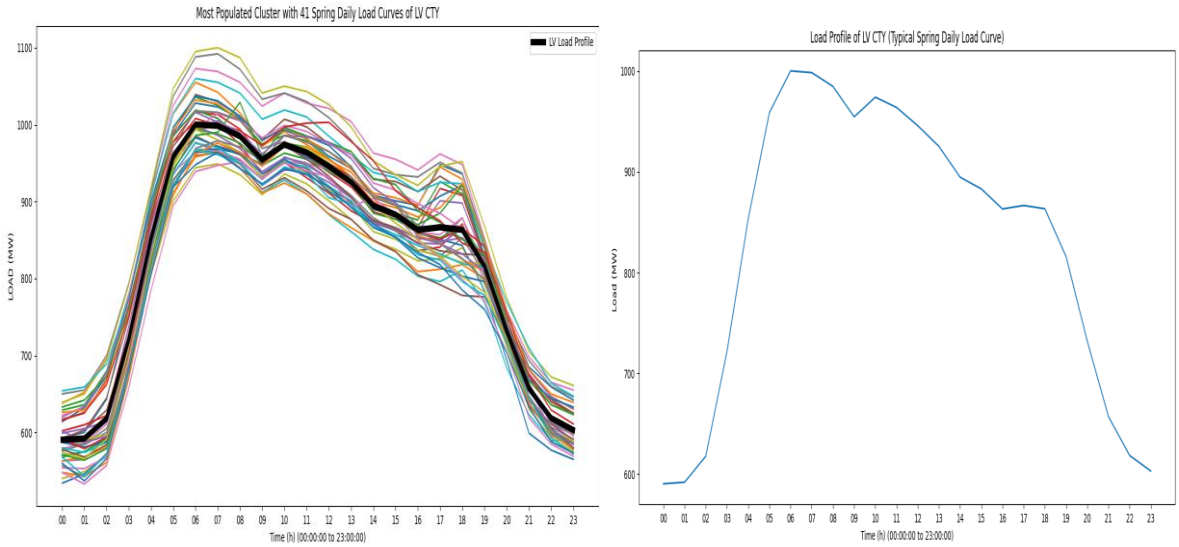
14) Λετονία (LV)



Σχήμα 4.53 : (LV) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

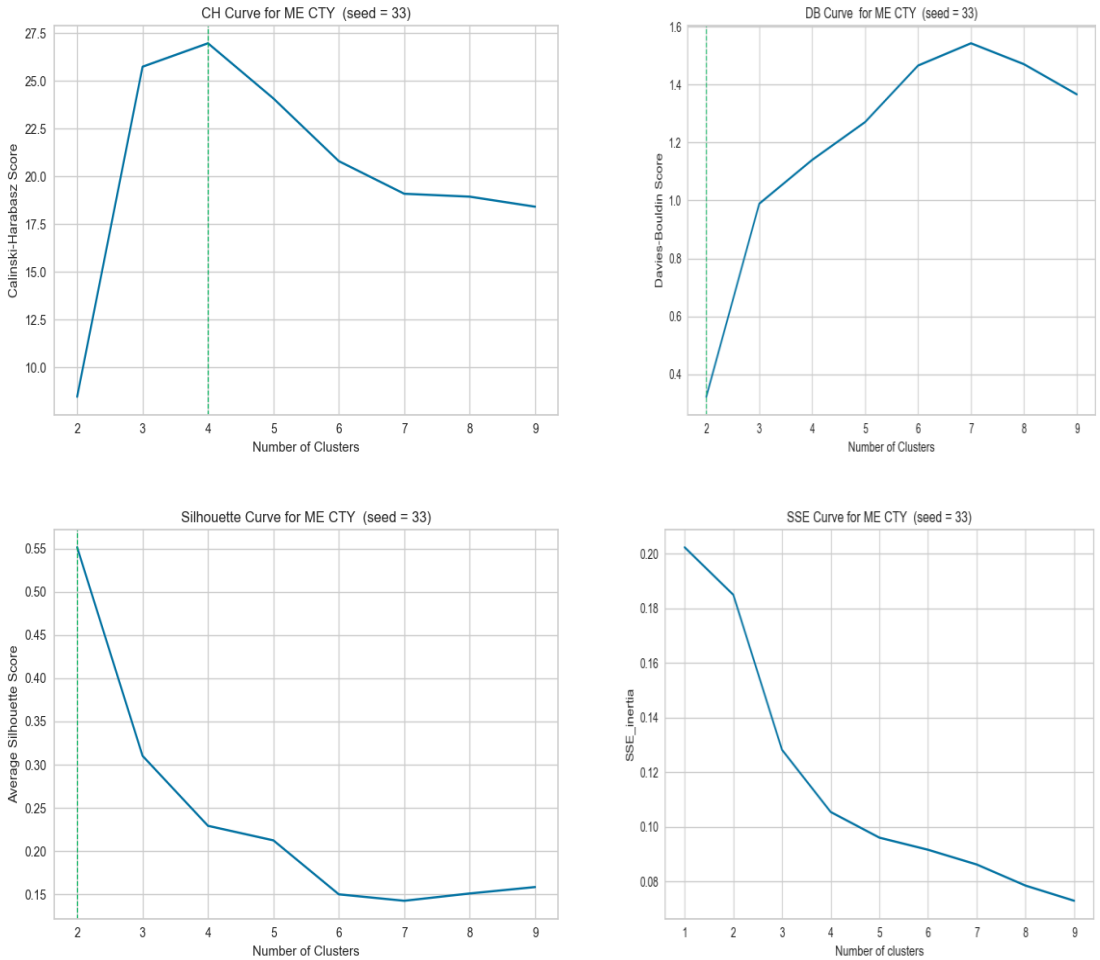


Σχήμα 4.54 : (LV) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

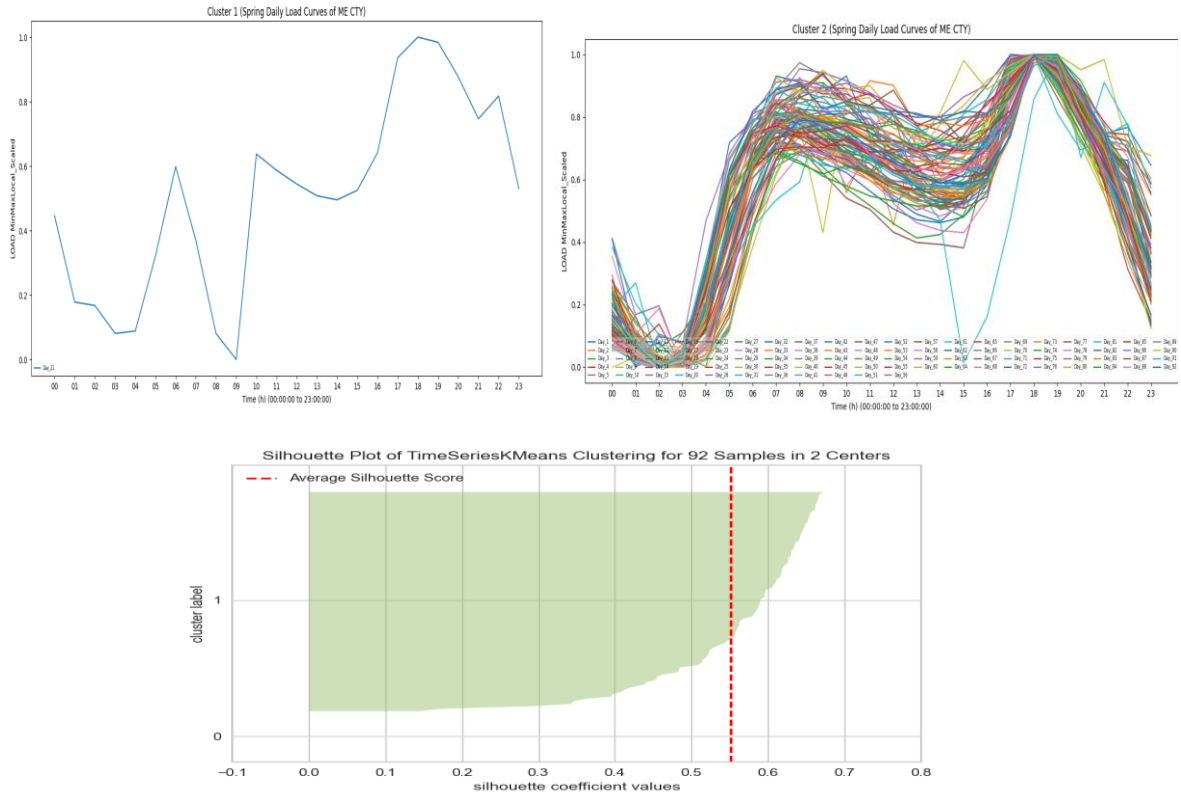


Σχήμα 4.55 : Προφίλ Φορτίου Άνοιξης (LV).

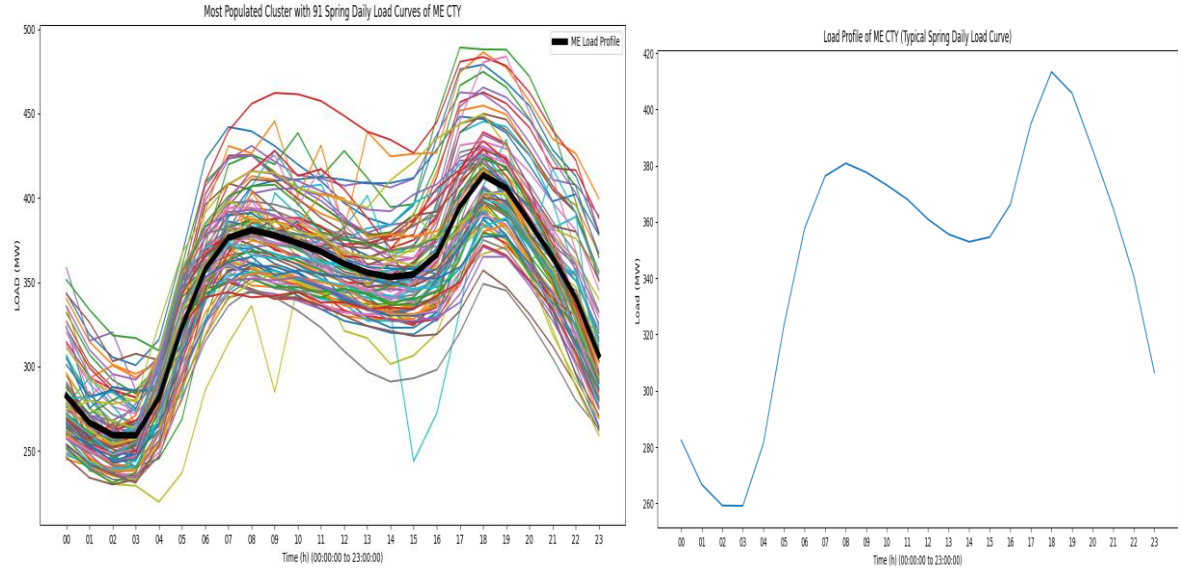
15) <u>Μαυροβούνιο (ME)</u>



Σχήμα 4.56 : (ME) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

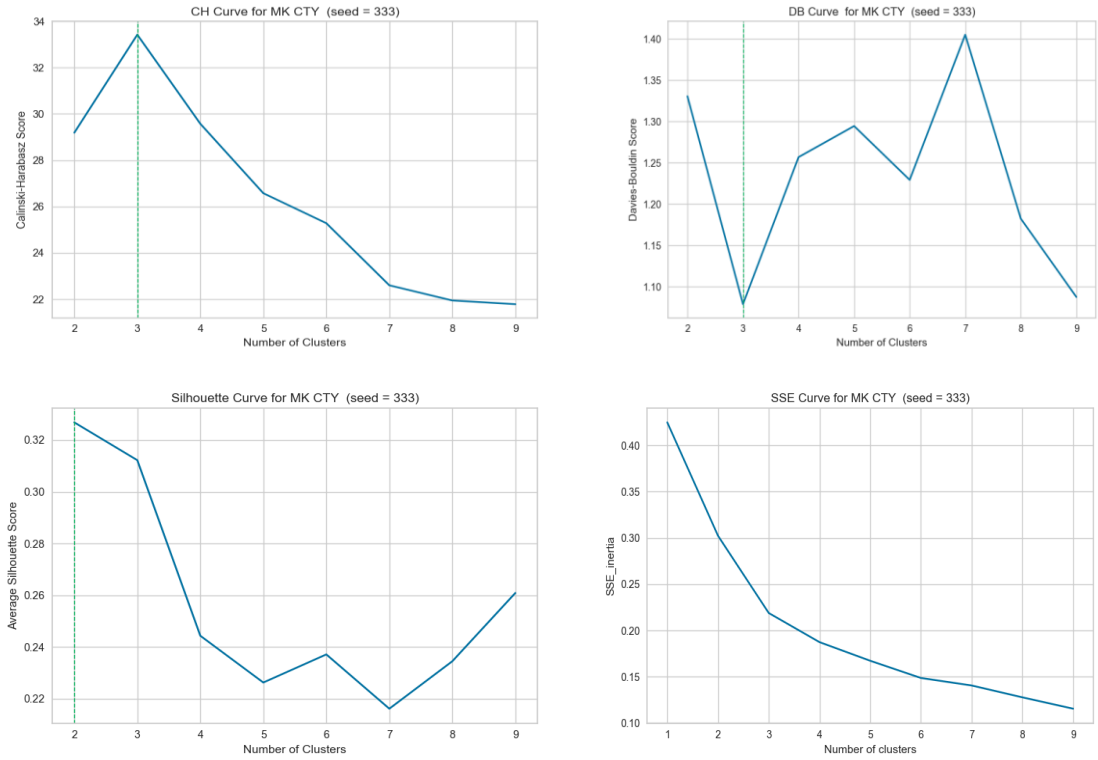


Σχήμα 4.57 : (ΜΕ) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

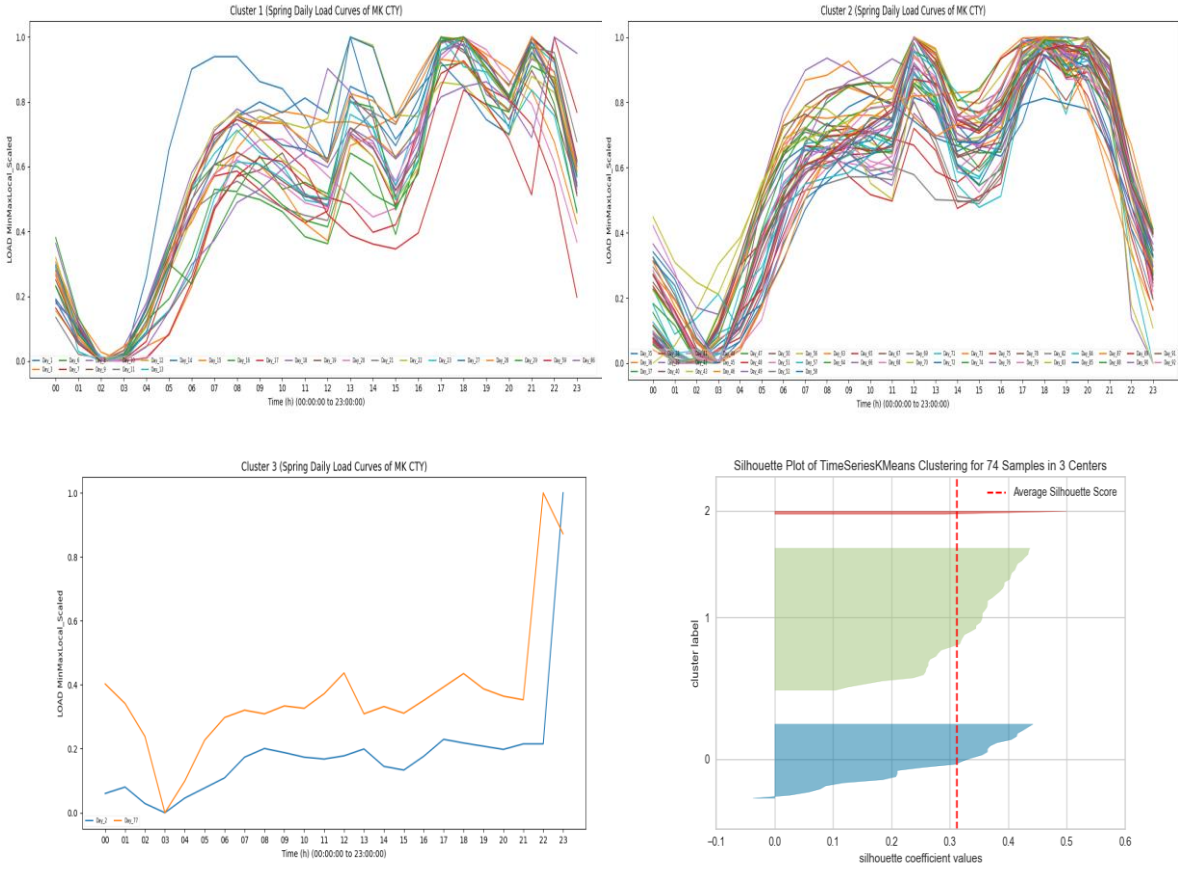


Σχήμα 4.58 : Προφίλ Φορτίου Άνοιξης (ΜΕ).

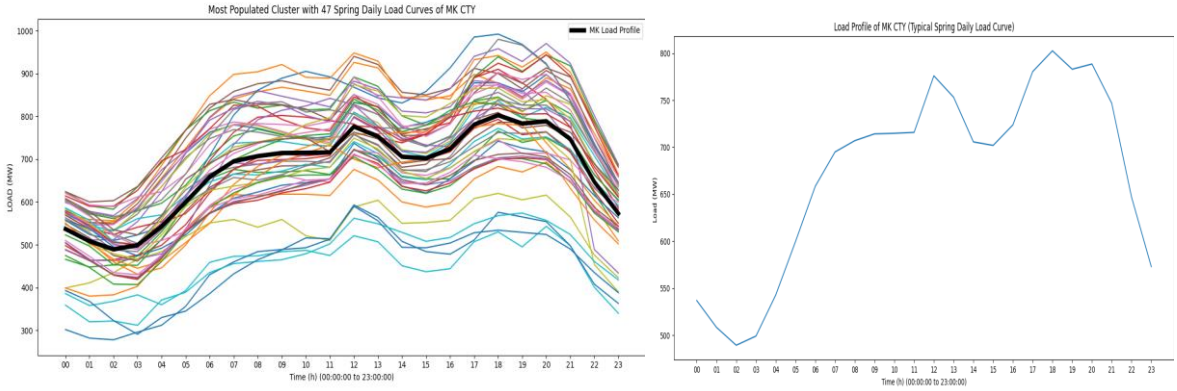
16) Βόρεια Μακεδονία (MK)



Σχήμα 4.59 : (MK) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

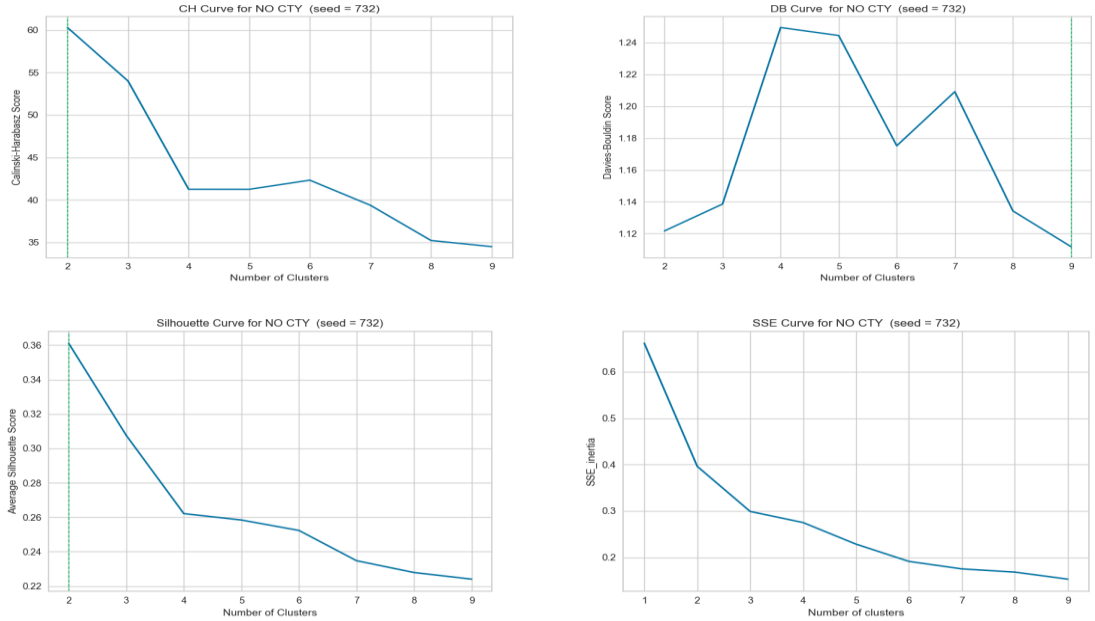


Σχήμα 4.60 : (MK) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

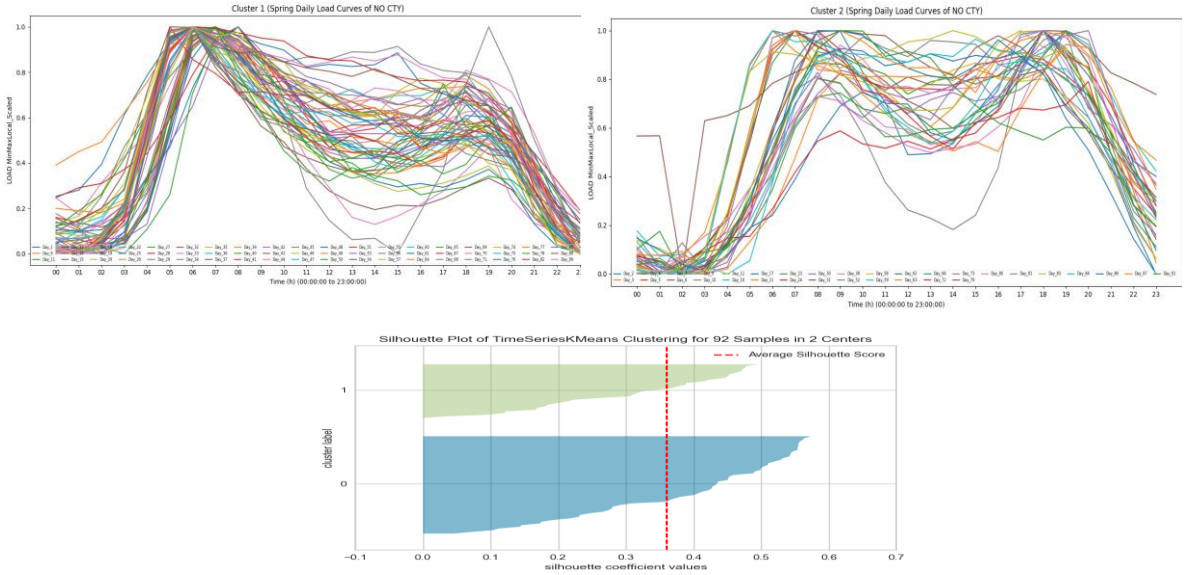


Σχήμα 4.61 : Προφίλ Φορτίου Άνοιξης (MK).

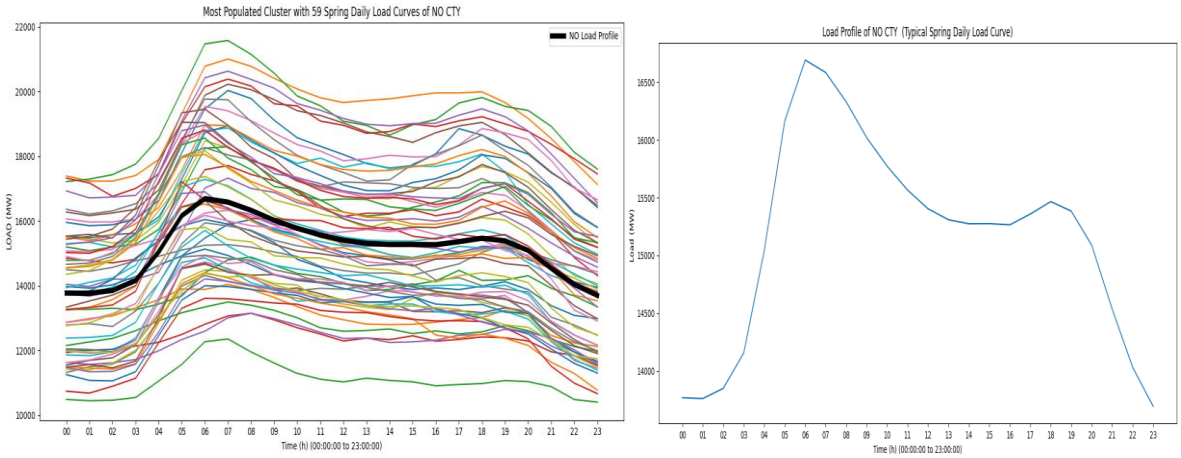
17) Νορβηγία (NO)



Σχήμα 4.62 : (NO) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

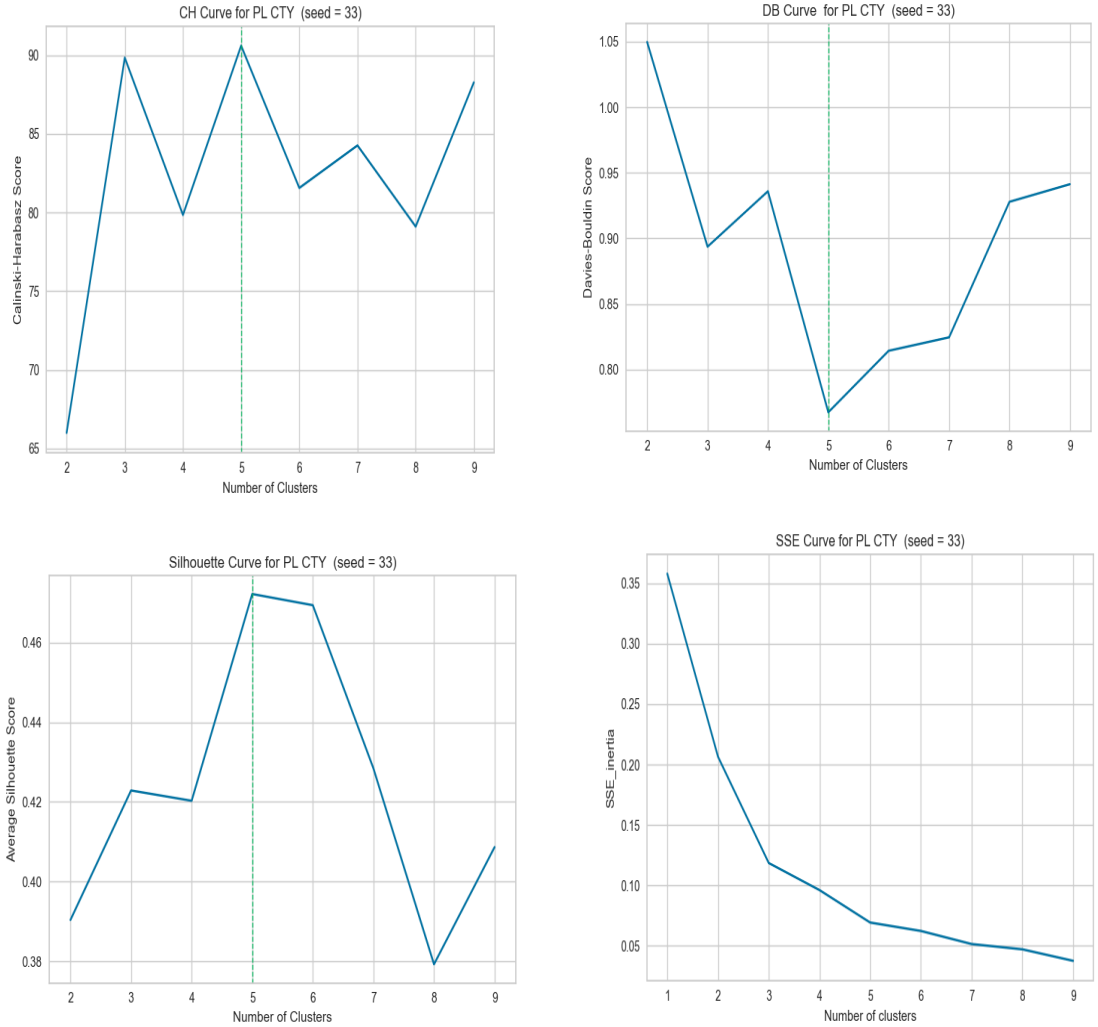


Σχήμα 4.63 : (NO) Ομαδοποίηση καμπυλών Σ.Π.Φ Άνοιξης.

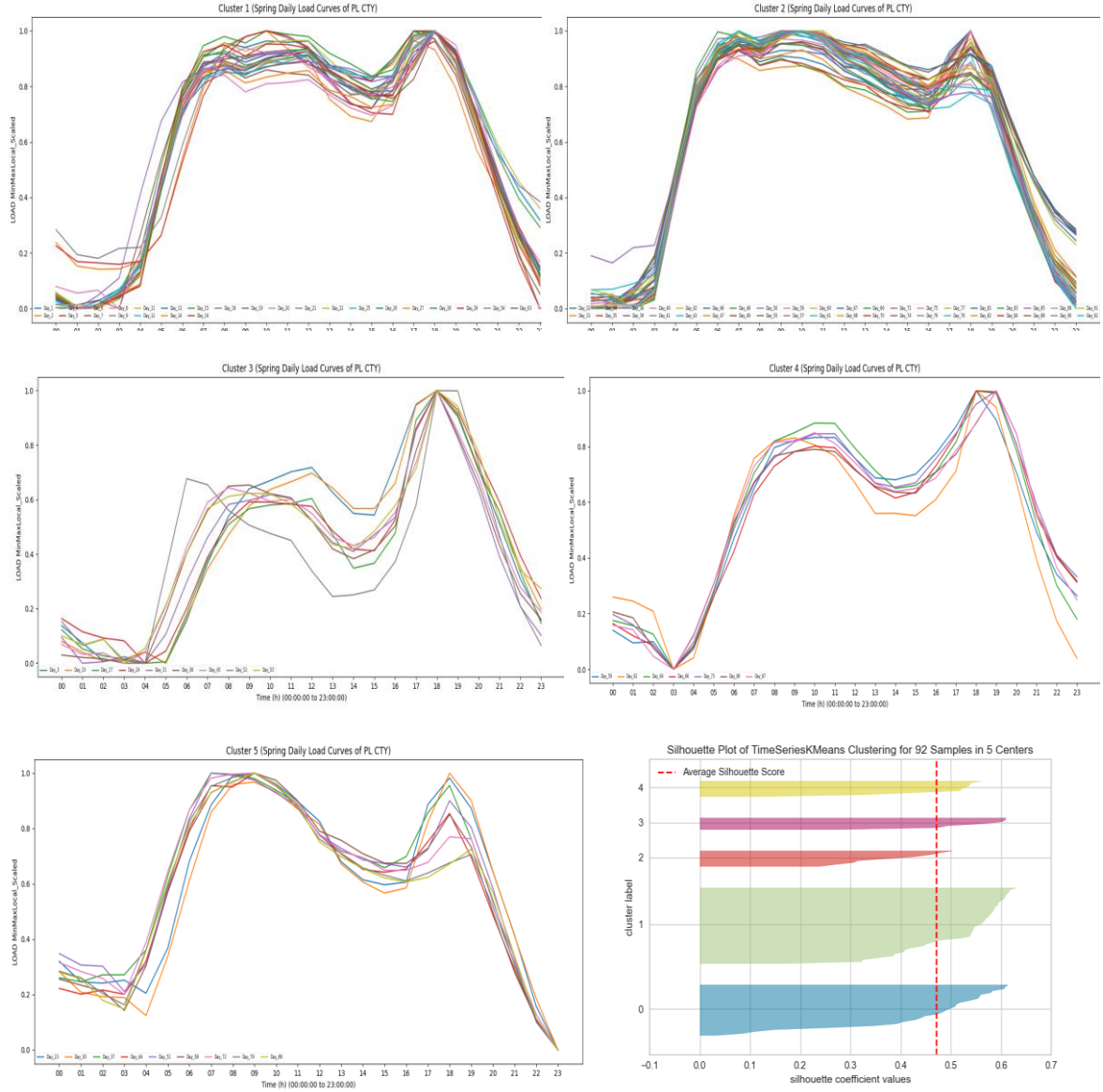


Σχήμα 4.64 : Προφίλ Φορτίου Άνοιξης (NO).

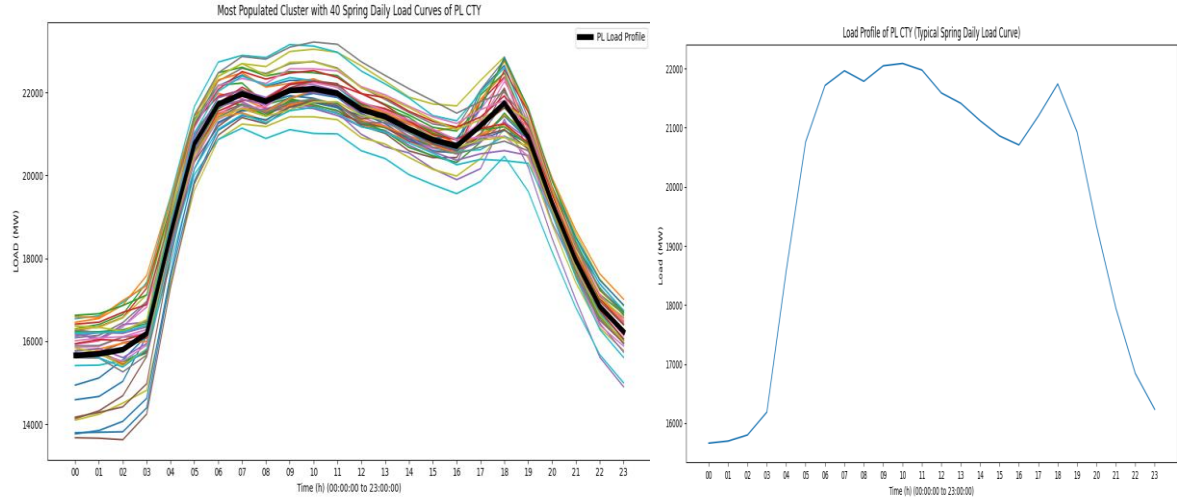
18) <u>Πολωνία (PL)</u>



Σχήμα 4.65 : (PL) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

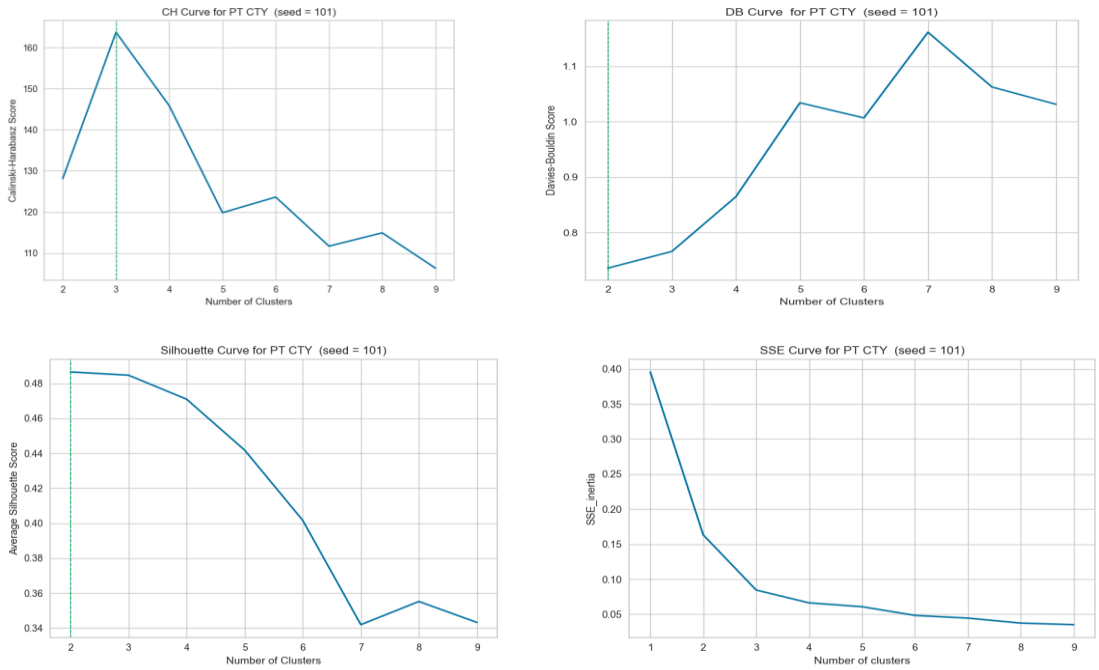


Σχήμα 4.66 : (PL) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

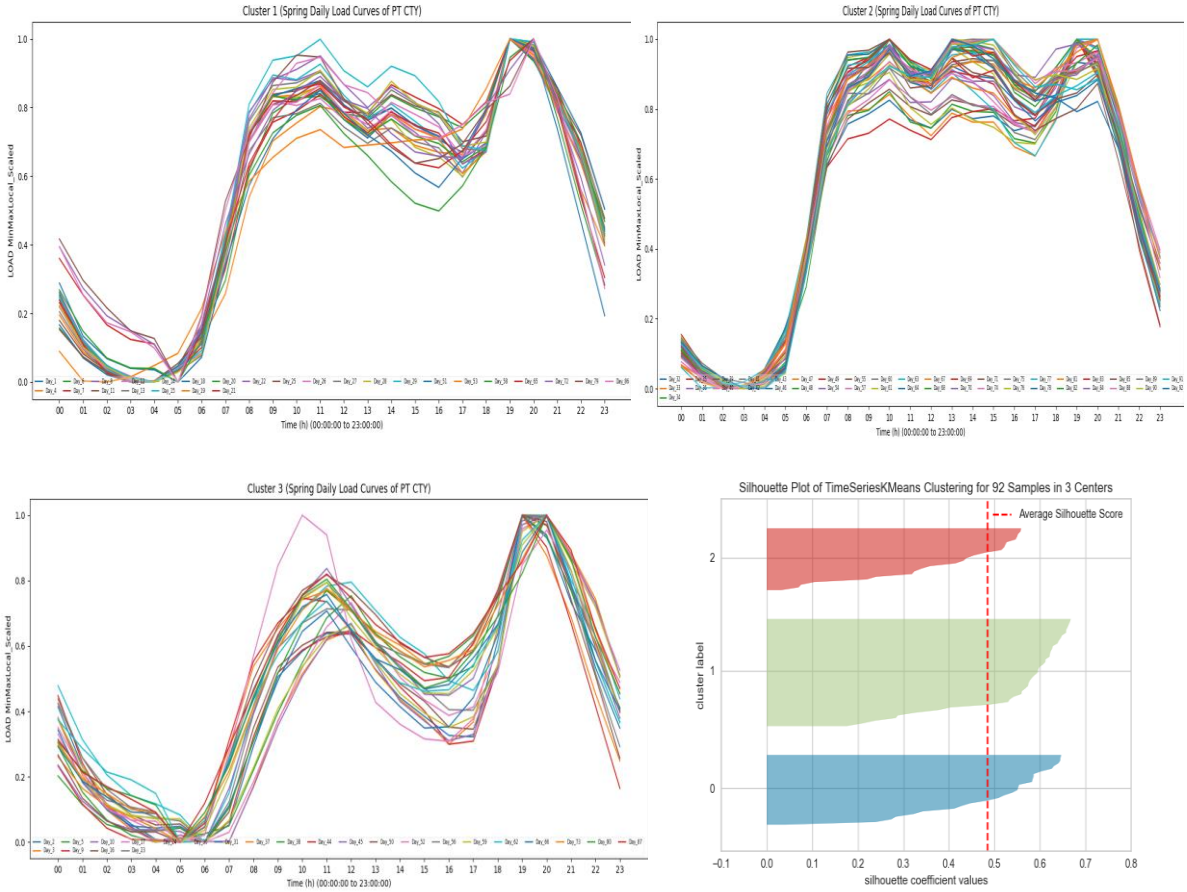


Σχήμα 4.67 : Προφίλ Φορτίου Άνοιξης (PL).

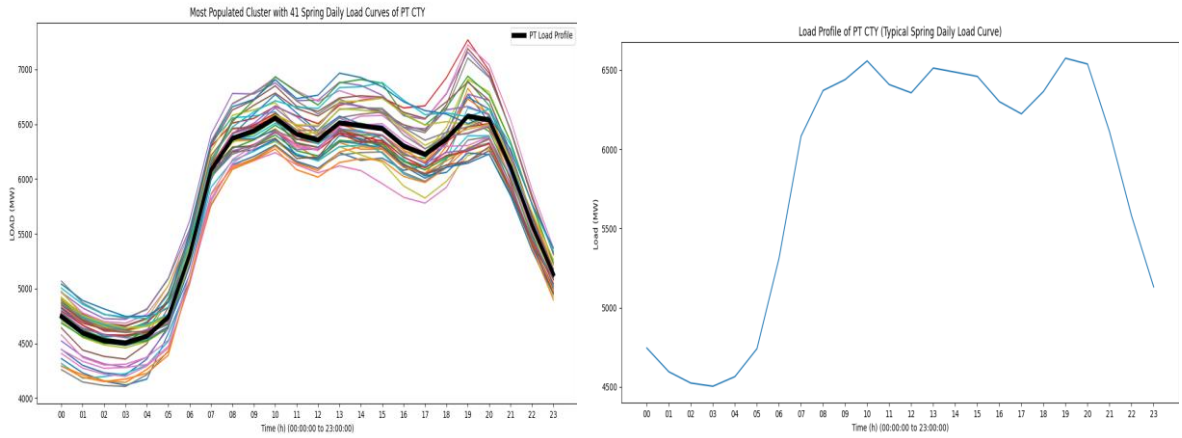
19) Πορτογαλία (PT)



Σχήμα 4.68 : (PT) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

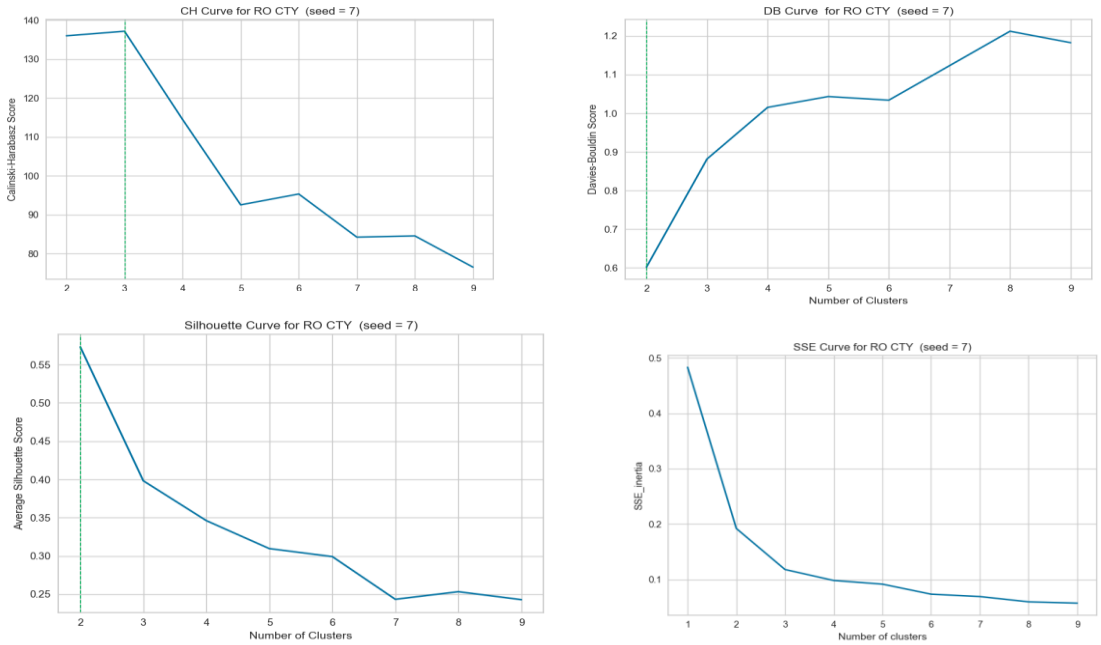


Σχήμα 4.69 : (PT) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

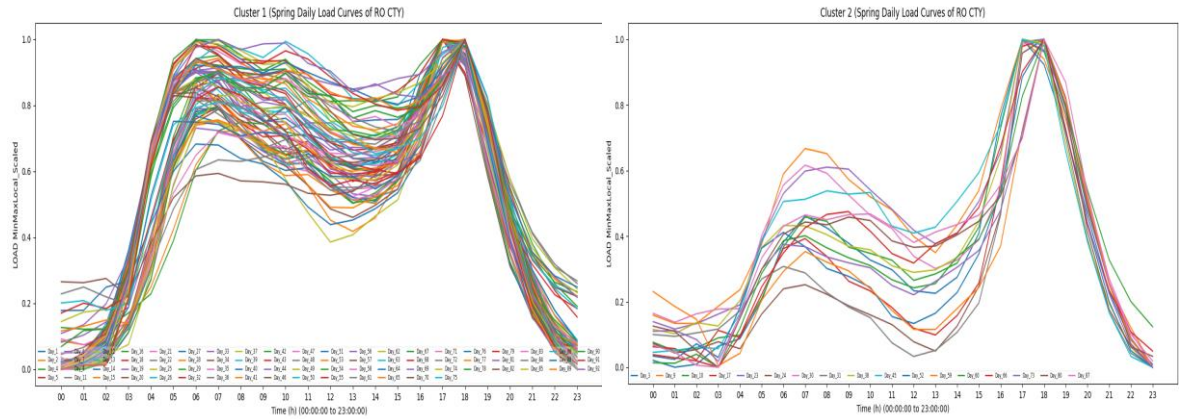


Σχήμα 4.70 : Προφίλ Φορτίου Άνοιξης (PT).

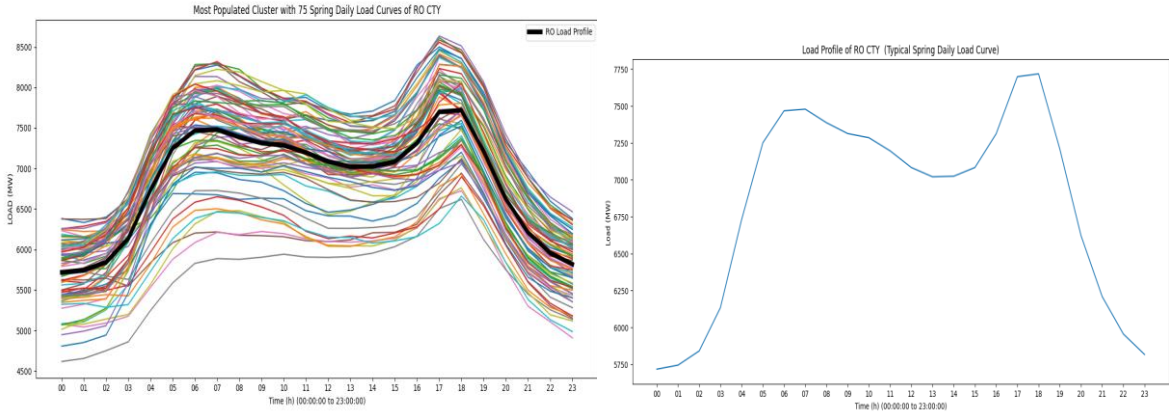
20) <u>Ρουμανία (RO)</u>



Σχήμα 4.71 : (RO) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

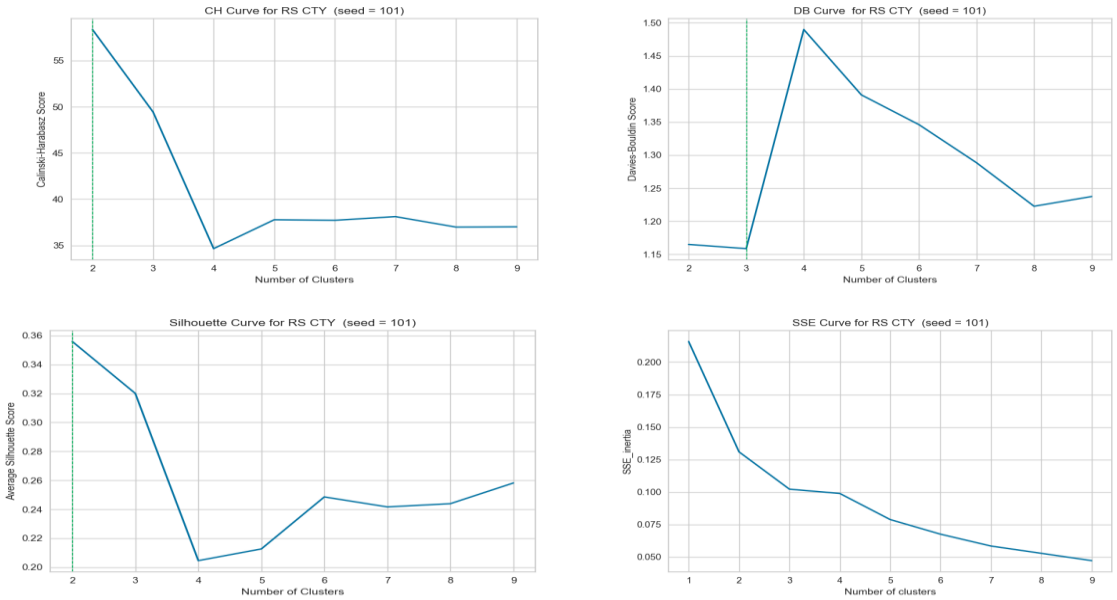


Σχήμα 4.72 : (RO) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

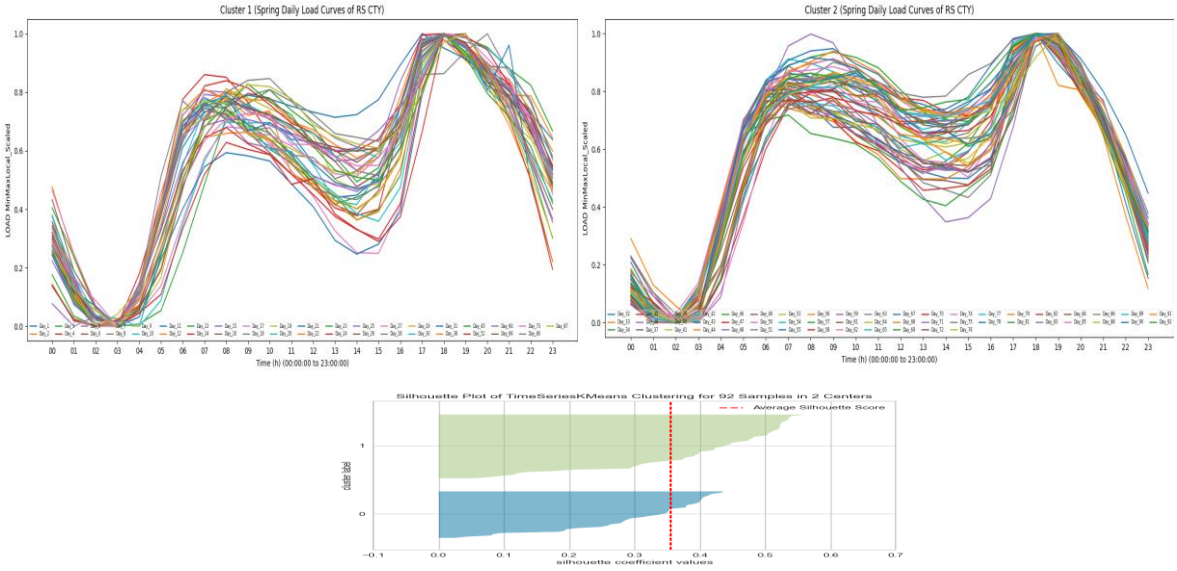


Σχήμα 4.73 : Προφίλ Φορτίου Άνοιξης (RO).

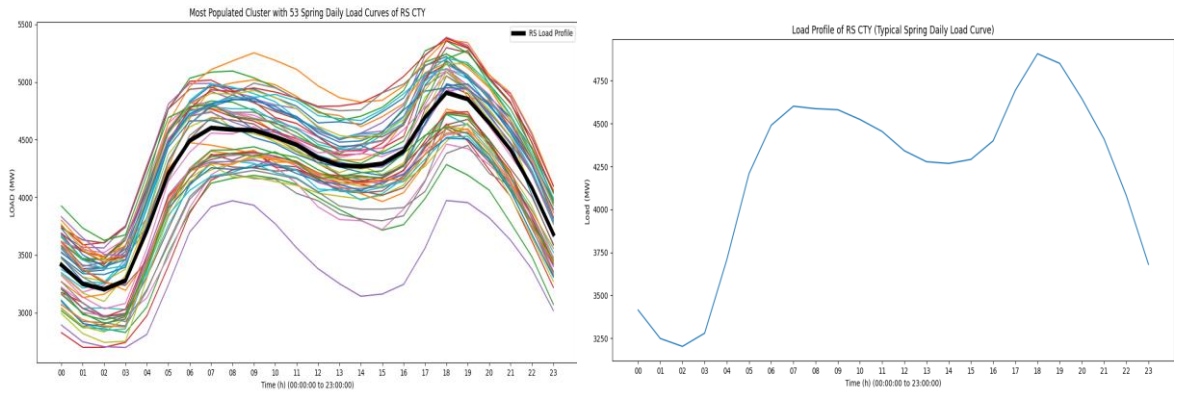
21) <u>Σερβία (RS)</u>



Σχήμα 4.74 : (RS) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

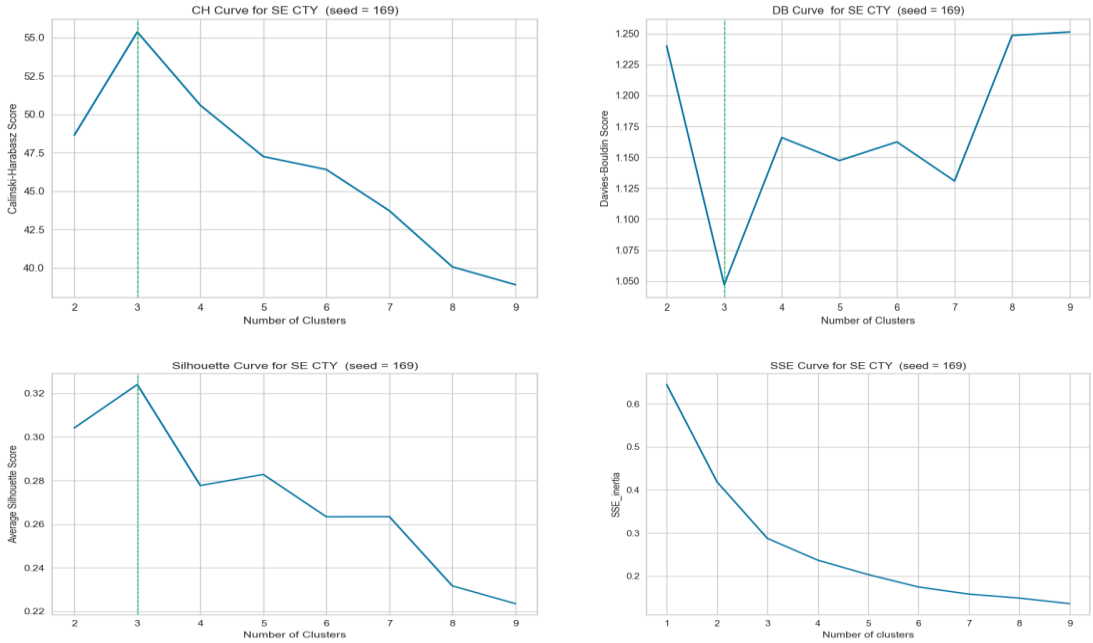


Σχήμα 4.75 : (RS) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

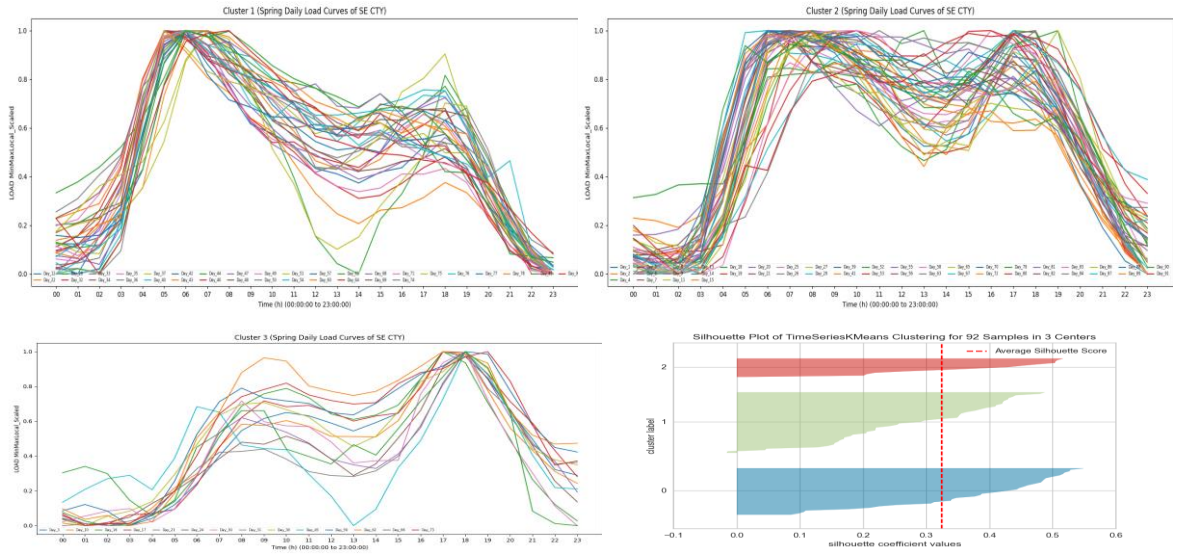


Σχήμα 4.76 : Προφίλ Φορτίου Άνοιξης (RS).

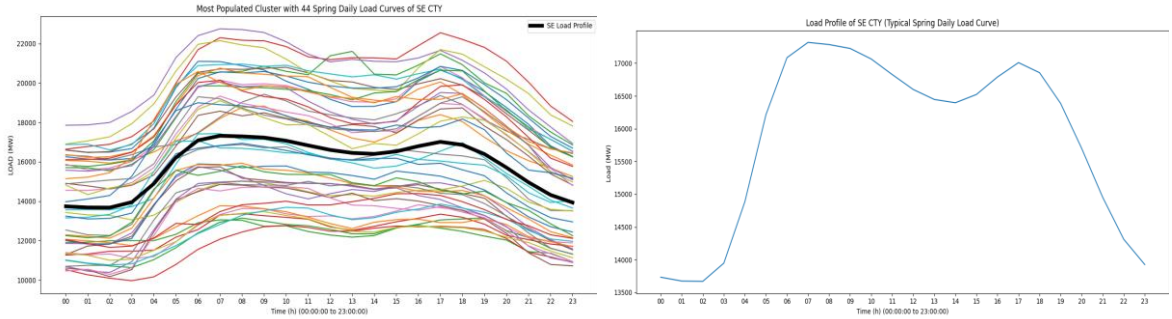
22) <u>Σουηδία (SE)</u>



Σχήμα 4.77 : (SE) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

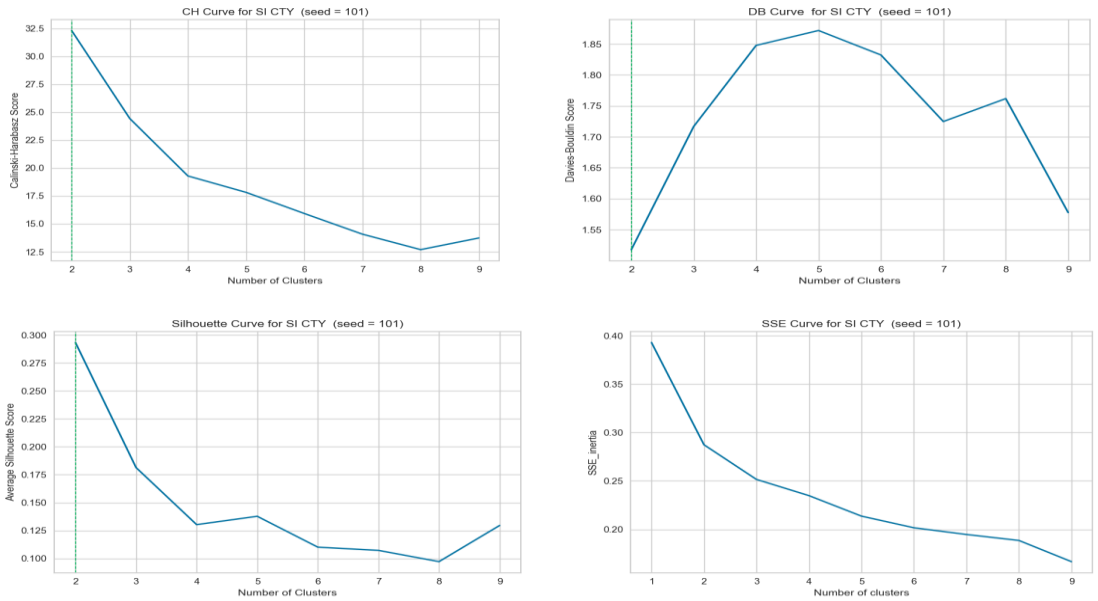


Σχήμα 4.78 : (SE) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

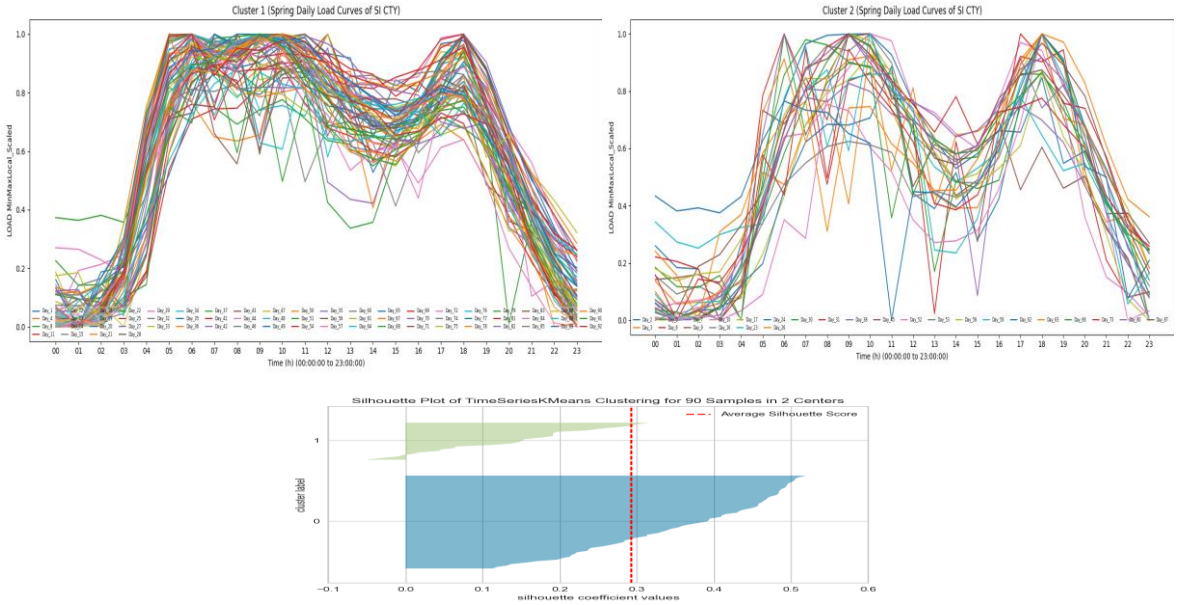


Σχήμα 4.79 : Προφίλ Φορτίου Άνοιξης (SE).

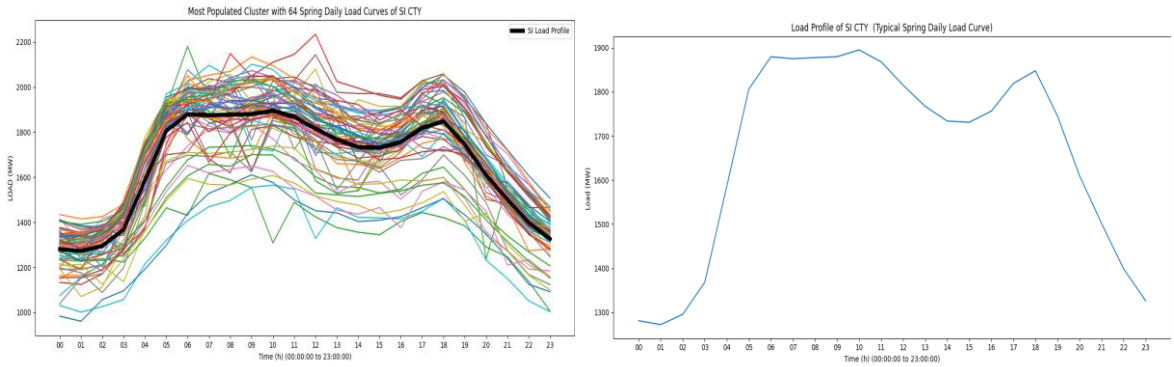
23) <u>Σλοβενία (SI)</u>



Σχήμα 4.80 : (SI) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

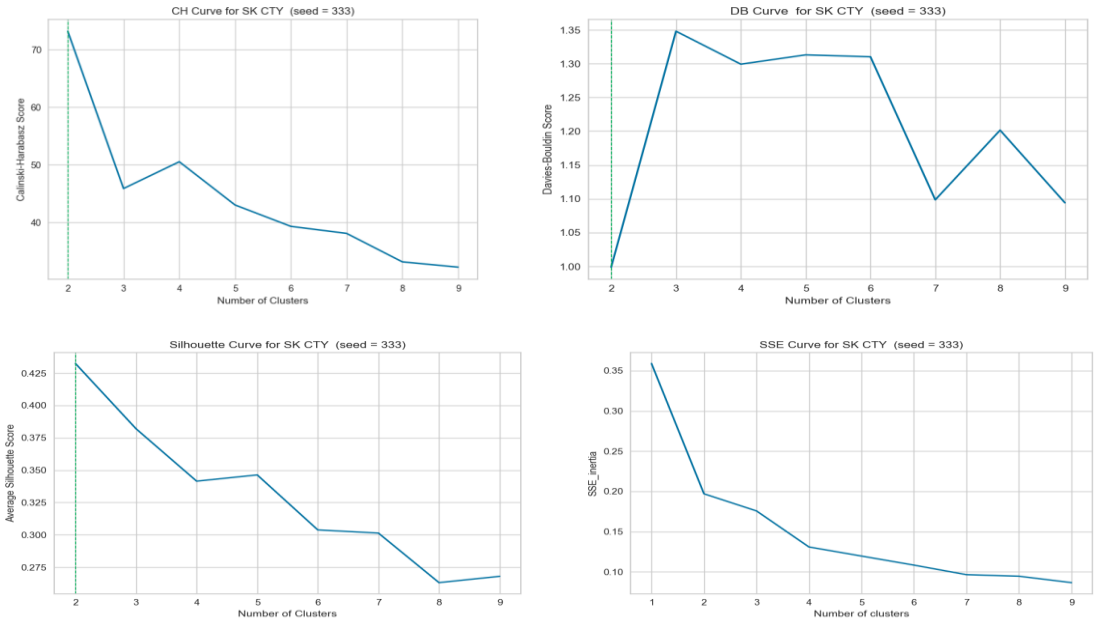


Σχήμα 4.81 : (SI) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

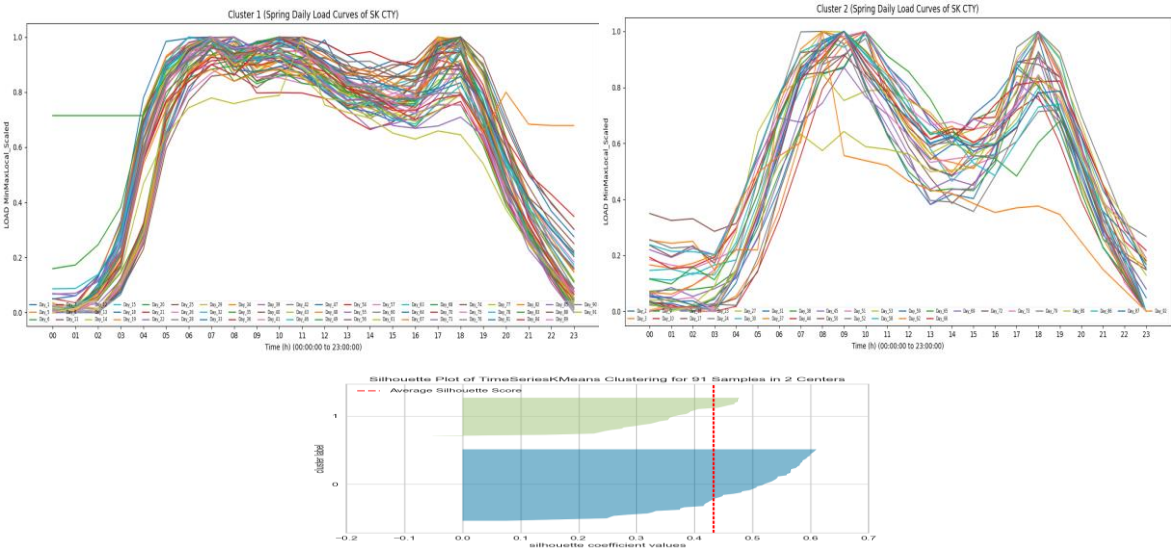


Σχήμα 4.82 : Προφίλ Φορτίου Άνοιξης (SI).

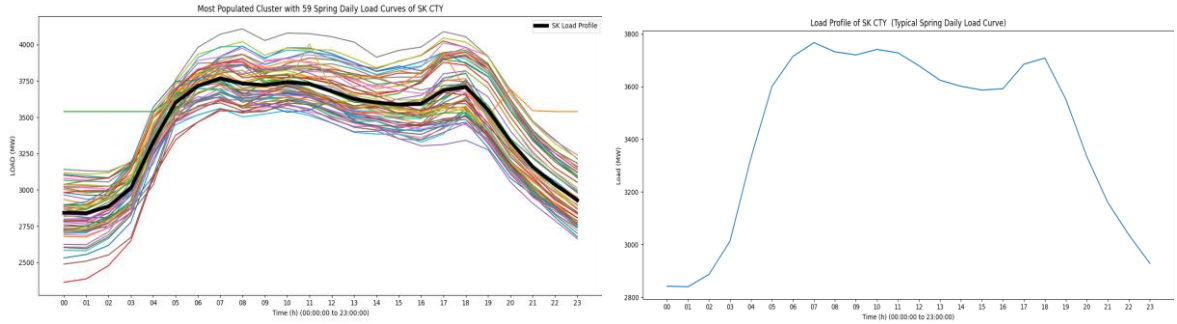
24) <u>Σλοβακία (SK)</u>



Σχήμα 4.83 : (SK) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.

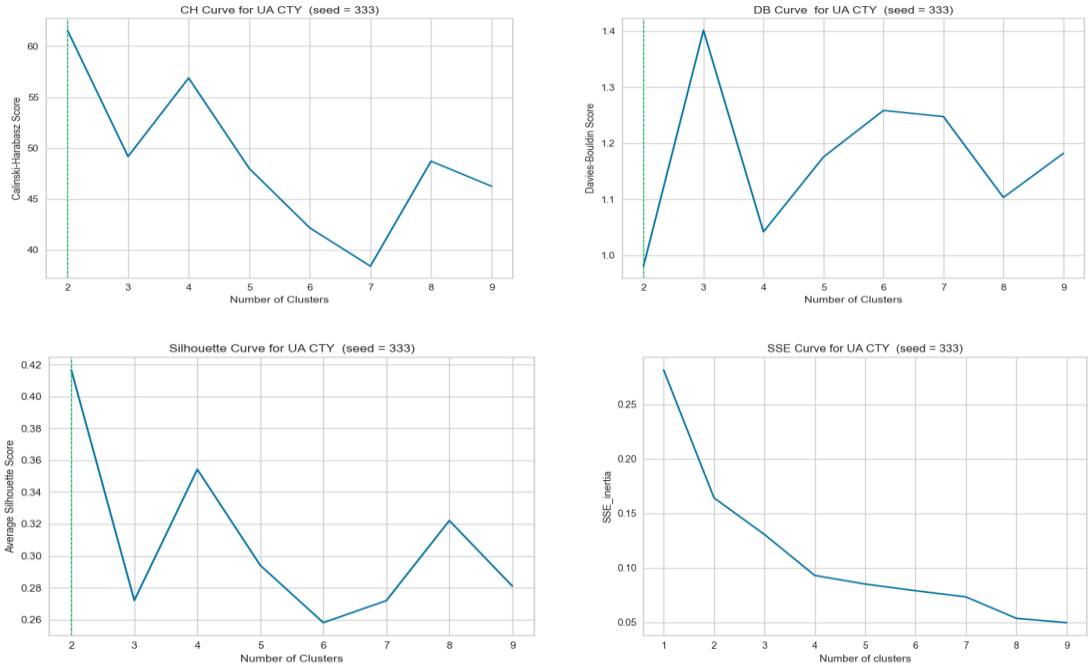


Σχήμα 4.84 : (SK) Ομάδες ημερήσιων καμπυλών Σ.Π.Φ Άνοιξης.

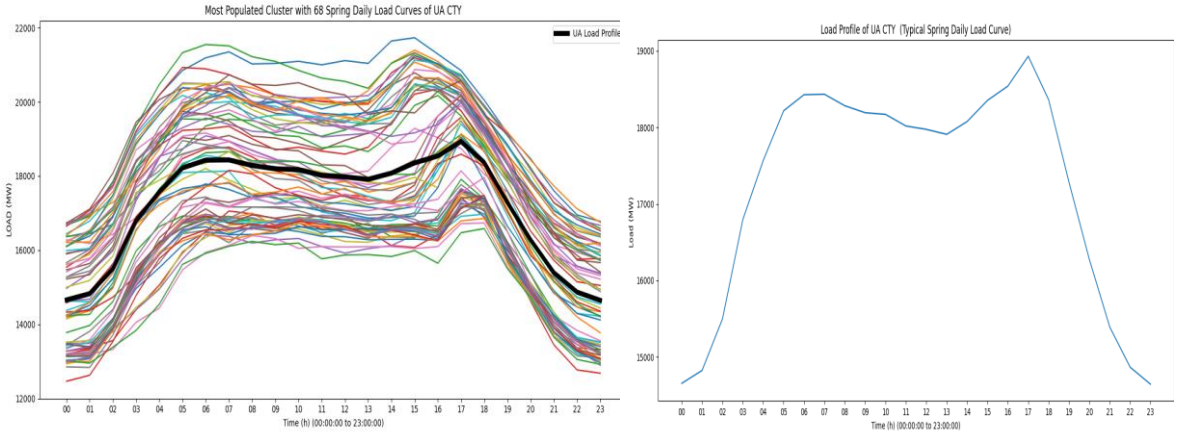


Σχήμα 4.85 : Προφίλ Φορτίου Άνοιξης (SK).

25) <u>Ουκρανία (UA)</u>



Σχήμα 4.86 : (UA) Αποτελέσματα Άνοιξης μετρικών επικύρωσης ομαδοποίησης.



Σχήμα 4.87 : Προφίλ Φορτίου Άνοιξης (UA).

Όσον αφορά τις υπόλοιπες εποχές, παρουσιαζουμε τα "Προφίλ" Φορτίου των ευρωπαϊκών χωρών του συνόλου ανάλυσης στο Παράρτημα Β.

Στη συνέχεια παρουσιάζουμε συνοπτικά τα αποτελέσματα των μετρικών επικύρωσης ομαδοποίησης για κάθε εποχή ανάλυσης.

Πίνακας 4.3 : Αποτελέσματα Μετρικών Επικύρωσης κατά την εποχική ανάλυση της Άνοιξης.

CTY	Spring				*Same optimal "k"*	Final "k"
	CH	DB	Silhouette	SSE		
BA	18	1.4	0.3	0.4	Y	3
BG	100	0.9	0.45	0.2	Y	2
CH	24	1.35	0.3	0.6	N	7
CZ	80	0.9	0.42	0.1	Y	4
DK	85	0.95	0.4	0.1	Y	3
EE	90	0.9	0.42	0.1	Y	3
ES	135	0.7	0.5	0.15	N	2
FI	80	1	0.42	0.4	Y	2
FR	64	0.9	0.4	0.3	N	3
GR	75	1	0.38	0.1	Y	2
HR	90	0.9	0.42	0.125	Y	2
IT	200	0.6	0.6	0.1	Y	2
LT	110	0.85	0.48	0.06	Y	3
LV	100	0.85	0.47	0.073	N	3
ME	27.5	0.4	0.55	0.15	N	2
MK	33	1.1	0.32	0.21	N	3
NO	60	1.1	0.36	0.35	N	2
PL	90	0.73	0.47	0.1	N	5
PT	160	0.8	0.48	0.1	Y	3
RO	135	0.6	0.55	0.2	Y	2
RS	57	1.16	0.36	0.125	Y	2
SE	55	1.05	0.32	0.3	Y	3
SI	32.5	1.5	0.29	0.27	Y	2
SK	73	1	0.42	0.15	Y	2
UA	60	1	0.42	0.15	Y	2

Πίνακας 4.4 : Αποτελέσματα Μετρικών Επικύρωσης κατά την Θερινή εποχική ανάλυση.

CTY	Summer				*Same optimal "k"*	Final "k"
	CH	DB	Silhouette	SSE		
BA	22	1.5	0.38	0.5	N	2
BG	63	1	0.42	0.08	N	3
CH	35	1.33	0.29	0.4	N	3
CZ	220	0.5	0.61	0.05	Y	3
DK	118	0.8	0.5	0.08	Y	2
EE	115	0.83	0.52	0.079	Y	2
ES	112	0.6	0.5	0.06	N	3
FI	160	0.75	0.58	0.1	Y	2
FR	155	0.7	0.57	0.08	Y	2
GR	46	1.1	0.32	0.065	Y	3
HR	60	1	0.44	0.06	Y	2
IT	180	0.5	0.64	0.1	Y	2
LT	105	0.75	0.52	0.05	Y	2
LV	112	0.45	0.55	0.05	Y	4
ME	29	0.75	0.5	0.07	N	2
MK	27	0.75	0.37	0.08	Y	3
NO	80	1	0.44	0.165	Y	2
PL	130	0.6	0.61	0.06	Y	3
PT	158	0.68	0.57	0.08	Y	2
RO	130	0.68	0.56	0.08	Y	2
RS	33	1.25	0.37	0.05	N	2
SE	88	0.9	0.45	0.14	Y	2
SI	44	1	0.42	0.170	N	2
SK	105	0.84	0.52	0.08	Y	2
UA	65	0.9	0.4	0.06	Y	3

Πίνακας 4.5 : Αποτελέσματα Μετρικών Επικύρωσης κατά τη Φθινοπωρινή εποχική ανάλυση.

CTY	Autumn				*Same optimal "k"*	Final "k"
	CH	DB	Silhouette	SSE		
BA	17	1.3	0.2	0.2	N	7
BG	53	1,05	0.36	0.075	Y	4
CH	32	1.36	0.35	0.55	N	3
CZ	97	0.81	0.49	0.1	Y	4
DK	112	0.83	0.45	0.08	Y	3
EE	110	0.75	0.46	0.08	Y	3
ES	115	0.7	0.53	0.06	Y	4
FI	115	0.73	0.52	0.2	Y	2
FR	120	0.8	0.5	0.15	Y	2
GR	83	0.9	0.43	0.1	Y	2
HR	77	0.85	0.46	0.05	Y	4
IT	160	0.64	0.55	0.06	Y	4
LT	146	0.72	0.52	0.055	Y	3
LV	145	0.73	0.525	0.05	Y	3
ME	75	0.9	0.4	0.08	N	3
MK	63	1	0.42	0.15	Y	2
NO	57	1.07	0.35	0.15	N	4
PL	110	0.7	0.5	0.07	Y	4
PT	135	0.65	0.55	0.08	N	3
RO	80	0.85	0.5	0.2	Y	2
RS	95	0.86	0.48	0.1	Y	2
SE	80	0.85	0.45	0.15	Y	2
SI	57	1	0.37	0.13	Y	3
SK	78	0.88	0.46	0.08	N	3
UA	50	1	0.37	1	N	3

Πίνακας 4.6 : Αποτελέσματα Μετρικών Επικύρωσης κατά τη Χειμερινή εποχική ανάλυση.

CTY	Winter				*Same optimal "k"*	Final "k"
	CH	DB	Silhouette	SSE		
BA	26	1.7	0.27	0.5	Y	2
BG	60	1.1	0.38	0.12	N	3
CH	70	1,1	0.4	0.6	Y	2
CZ	102	0.8	0.52	0.15	Y	3
DK	190	0.6	0.6	0.075	Y	2
EE	178	0.7	0.6	0.1	N	2
ES	180	0.64	0.6	0.07	Y	3
FI	110	0.8	0.5	0.2	Y	2
FR	85	0.9	0.46	0.2	Y	2
GR	65	1.1	0.4	0.1	Y	3
HR	88	0.9	0.47	0.08	Y	2
IT	200	0.6	0.6	0.08	Y	3
LT	180	0.6	0.6	0.07	Y	2
LV	240	0.6	0.63	0.05	Y	2
ME	34	1.3	0.25	0.08	Y	3
MK	27	1	0.32	0.12	N	3
NO	67	1	0.42	0.24	N	3
PL	155	0.53	0.63	0.1	Y	3
PT	225	0.58	0.64	0.08	Y	2
RO	150	0.7	0.55	0.11	Y	2
RS	47	1.2	0.4	0.15	Y	2
SE	90	0.9	0.44	0.2	Y	2
SI	58	0.85	0.4	0.15	N	3
SK	95	0.9	0.48	0.13	Y	2
UA	67.5	0.55	0.47	0.1	N	4

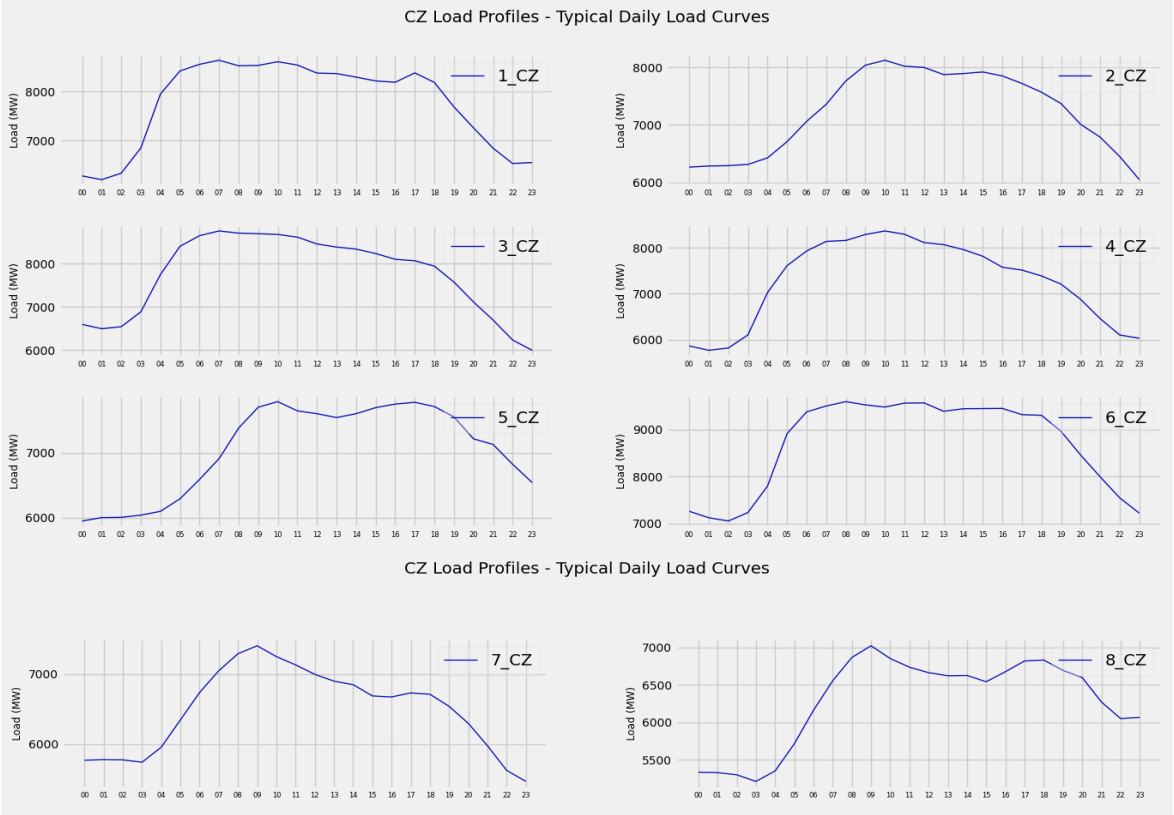
4.4.2 Αποτελέσματα Ετήσιας Ανάλυσης

Στην ενότητα αυτή παρουσιάζουμε τα "Προφίλ" Φορτίου που υπολογίσαμε κατά την ετήσια ανάλυση η οποία προσδιορίζεται στο χρονικό παράθυρο "01/03/2019" εώς "29/02/2020". Σημειώνεται ότι, όσον αφορά την ετήσια ανάλυση, δεν λήφθηκε υπόψη το υποσύνολο των δεδομένων της Βοσνίας – Ερζεγοβίνης και της Ελβετίας, καθώς κατά την εποχική ανάλυση οι εν λόγω χώρες παρουσίασαν μη ικανοποιητικά αποτελέσματα. Συνεπώς, τα "Προφίλ" Φορτίου που παρατίθενται στη συνέχεια αφορούν τις υπόλοιπες είκοσι τρεις (23) χώρες που εμπίπτουν στο σύνολο ανάλυσης.

1) <u>Βουλγαρία (BG)</u>

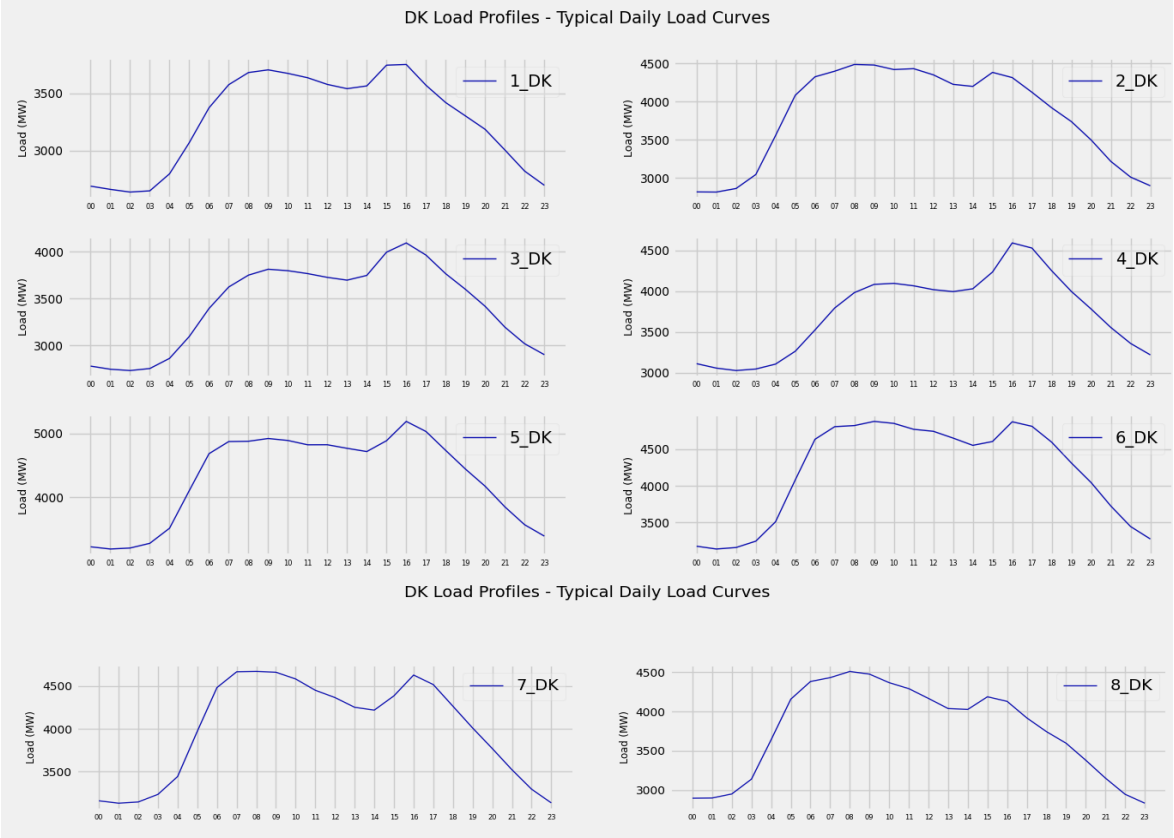
Σχήμα 4.88 : Προφίλ Φορτίου Βουλγαρίας (ετήσια ανάλυση).

2) Τσεχία (CZ)



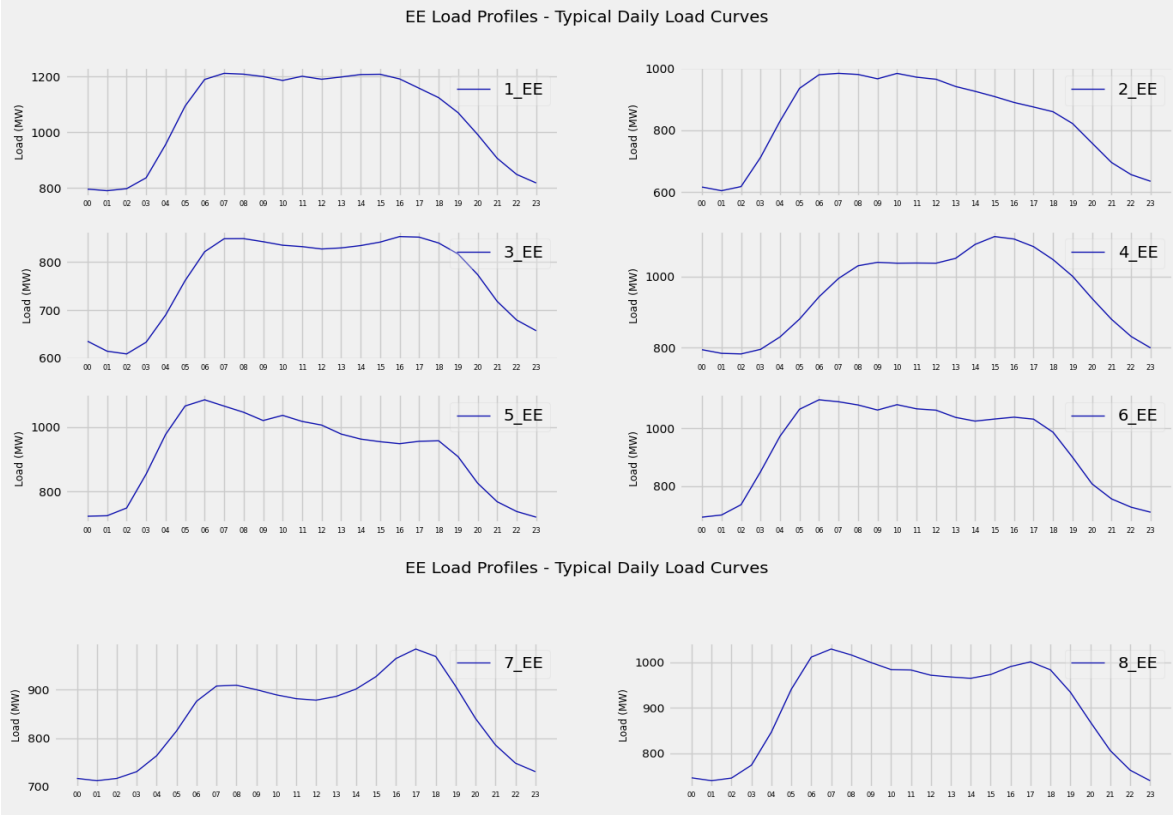
Σχήμα 4.89: Προφίλ Φορτίου Τσεχίας (ετήσια ανάλυση).

3) Δανία (DK)



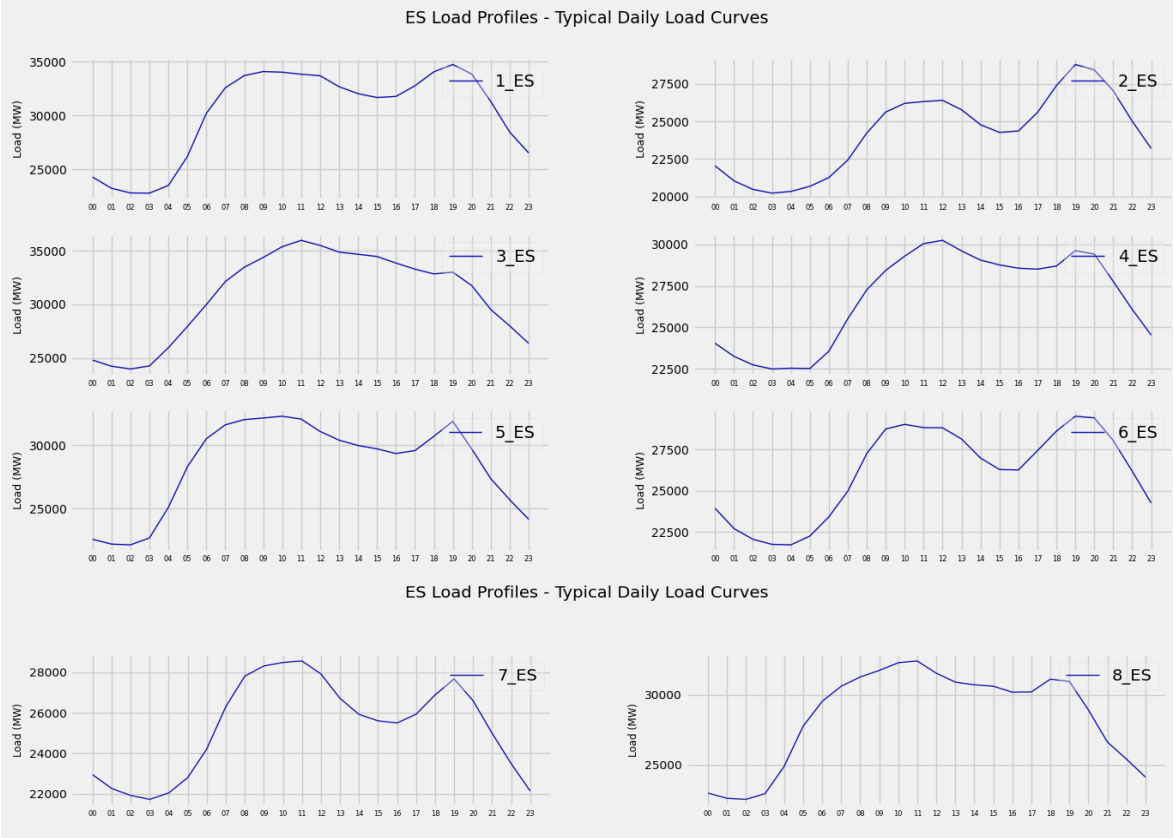
Σχήμα 4.90 : Προφίλ Φορτίου Δανίας (ετήσια ανάλυση).

4) Εσθονία (EE)



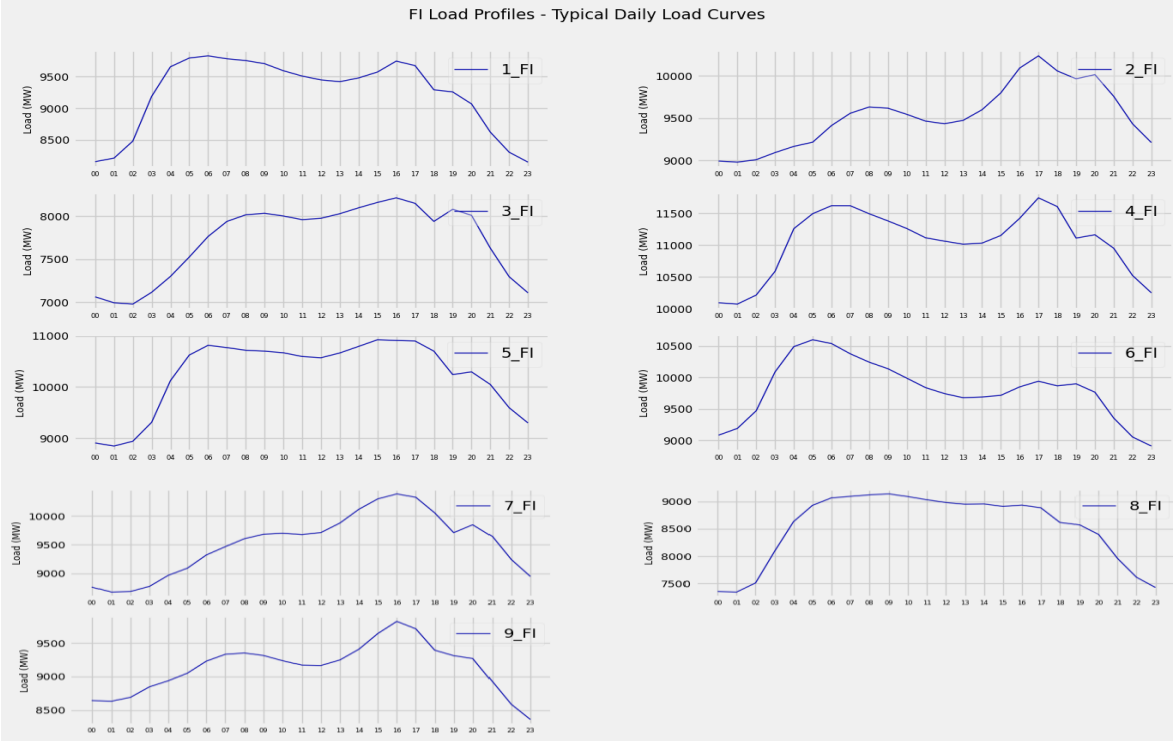
Σχήμα 4.91 : Προφίλ Φορτίου Εσθονίας (ετήσια ανάλυση).

5) Ισπανία (ES)



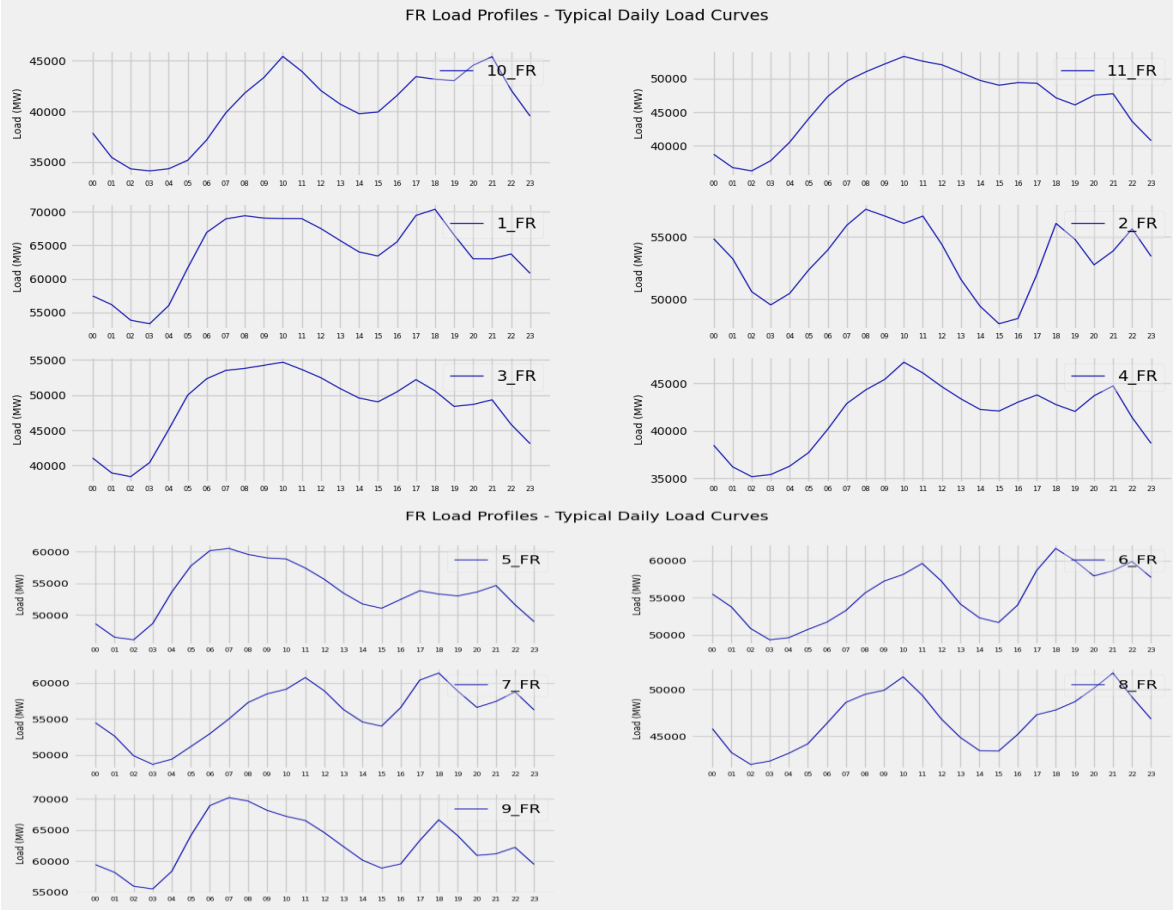
Σχήμα 4.92 : Προφίλ Φορτίου Ισπανίας (ετήσια ανάλυση).

6) Φινλανδία (FI)



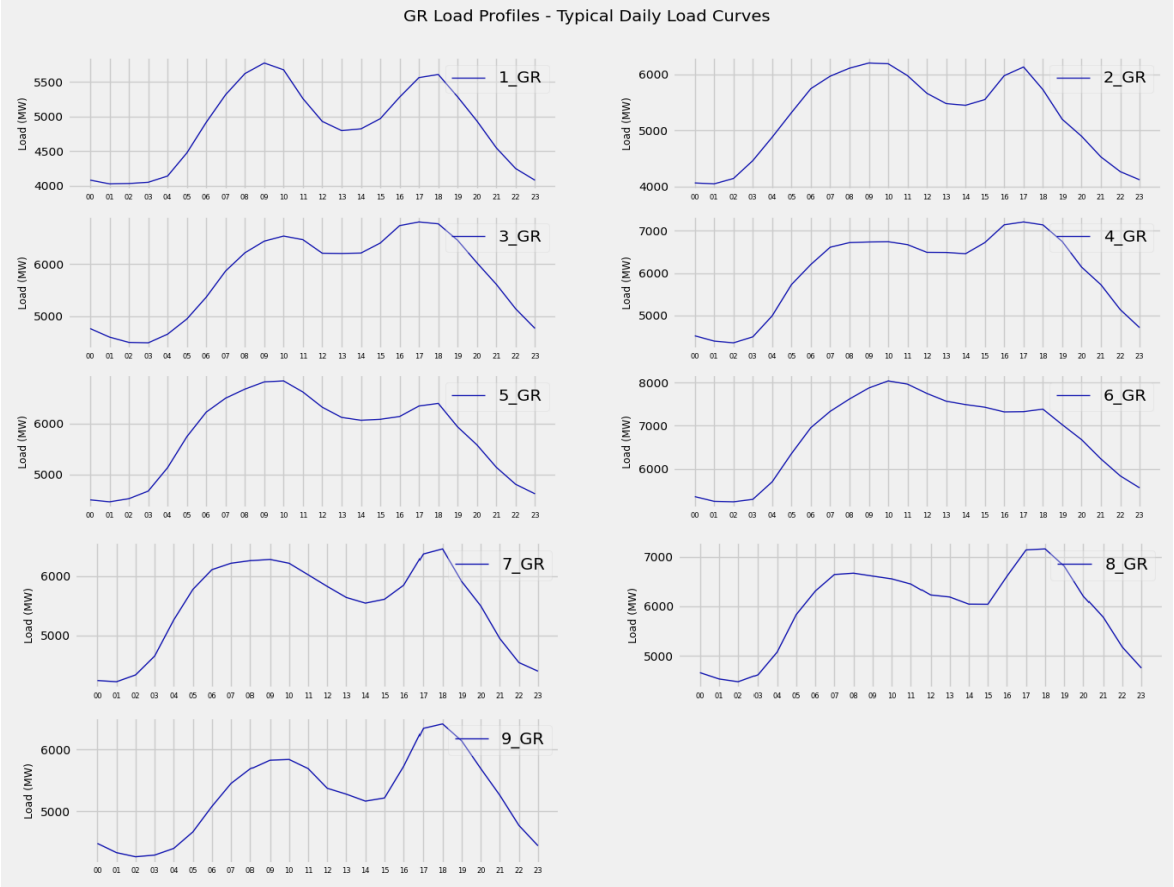
Σχήμα 4.93 : Προφίλ Φορτίου Φινλανδίας (ετήσια ανάλυση).

7) Γαλλία (FR)



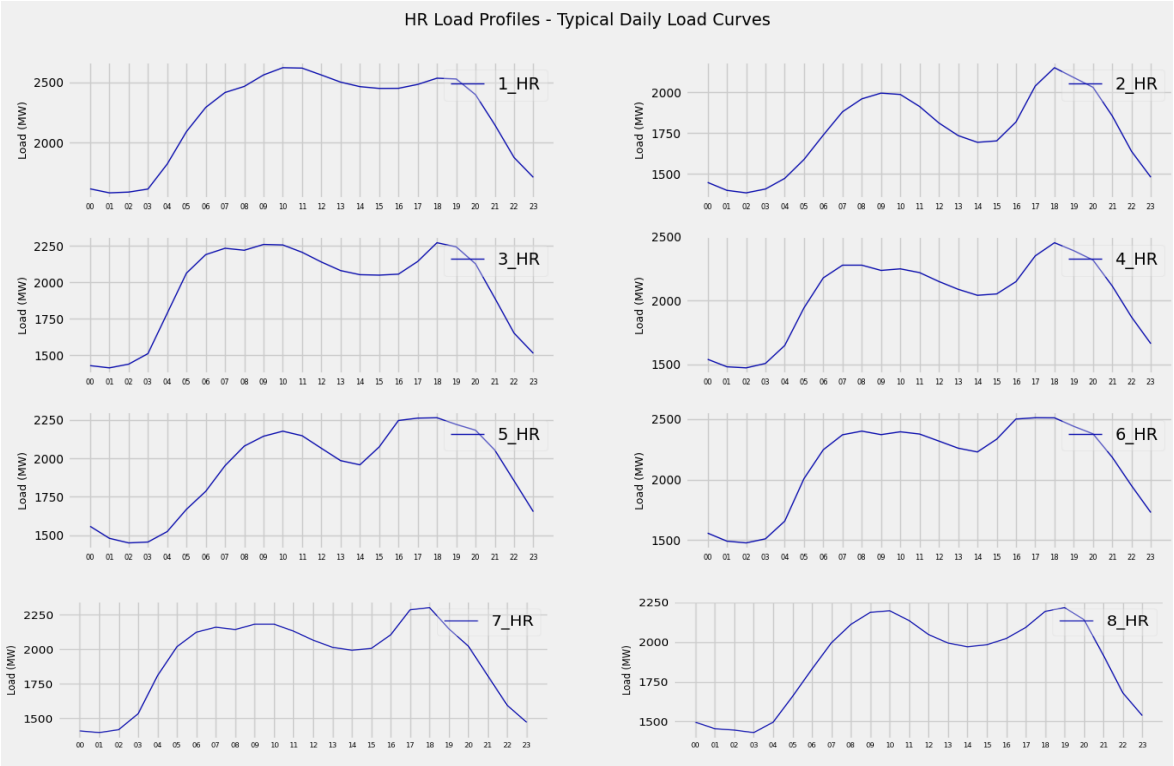
Σχήμα 4.94 : Προφίλ Φορτίου Γαλλίας (ετήσια ανάλυση).

8) Ελλάδα (GR)



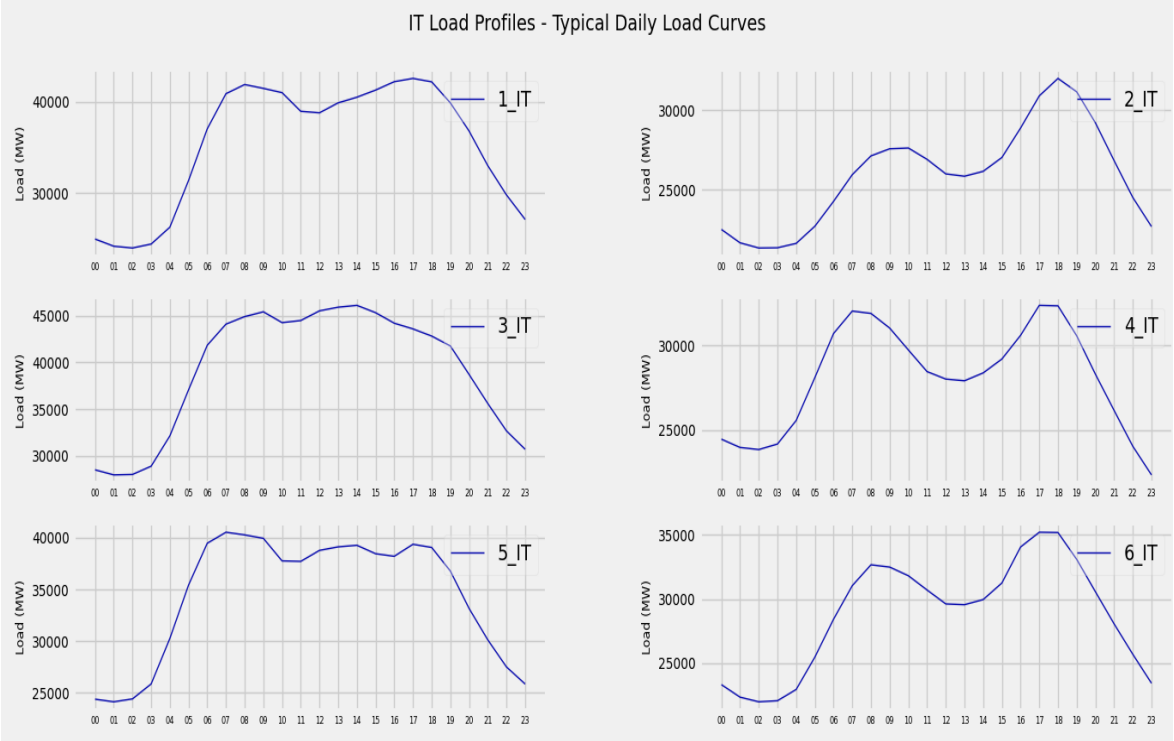
Σχήμα 4.95 : Προφίλ Φορτίου Ελλάδας (ετήσια ανάλυση).

9) Κροατία (HR)



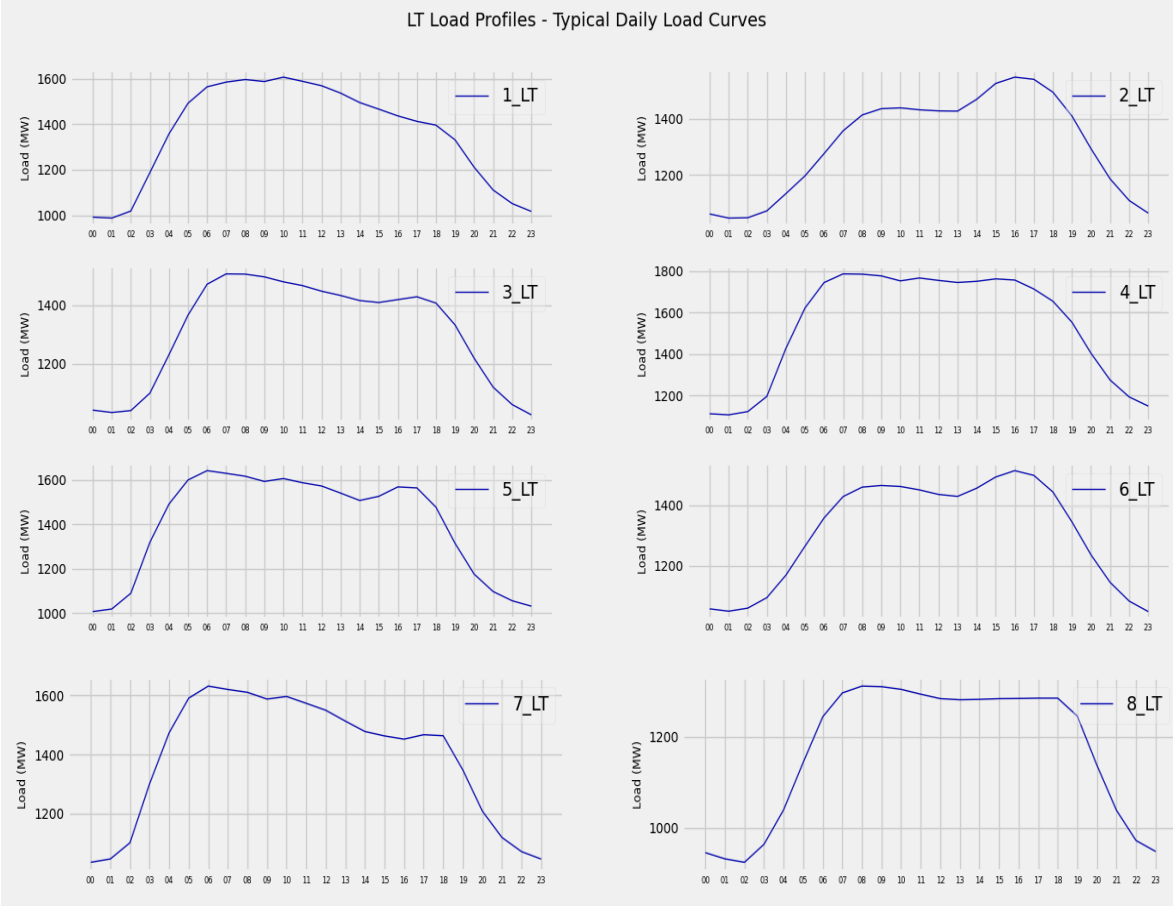
Σχήμα 4.96 : Προφίλ Φορτίου Κροατίας (ετήσια ανάλυση).

10) Ιταλία (IT)



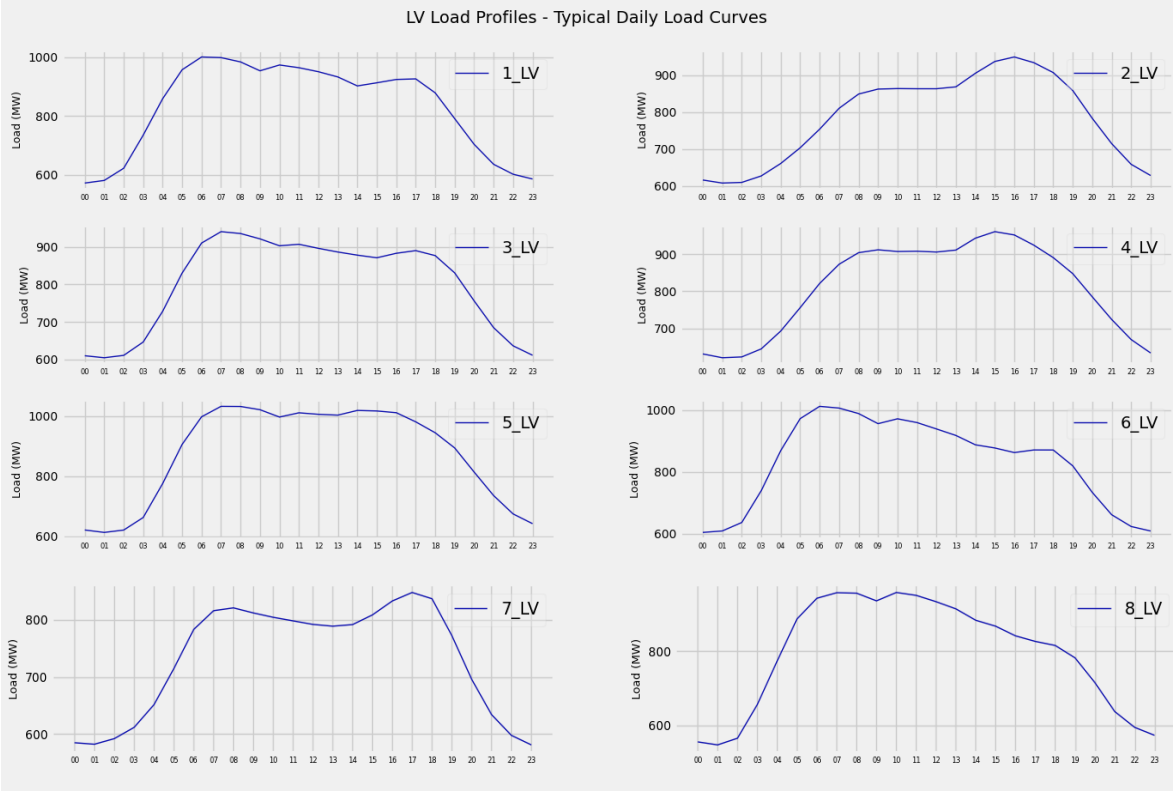
Σχήμα 4.97 : Προφίλ Φορτίου Ιταλίας (ετήσια ανάλυση).

11) Λιθουανία (LT)



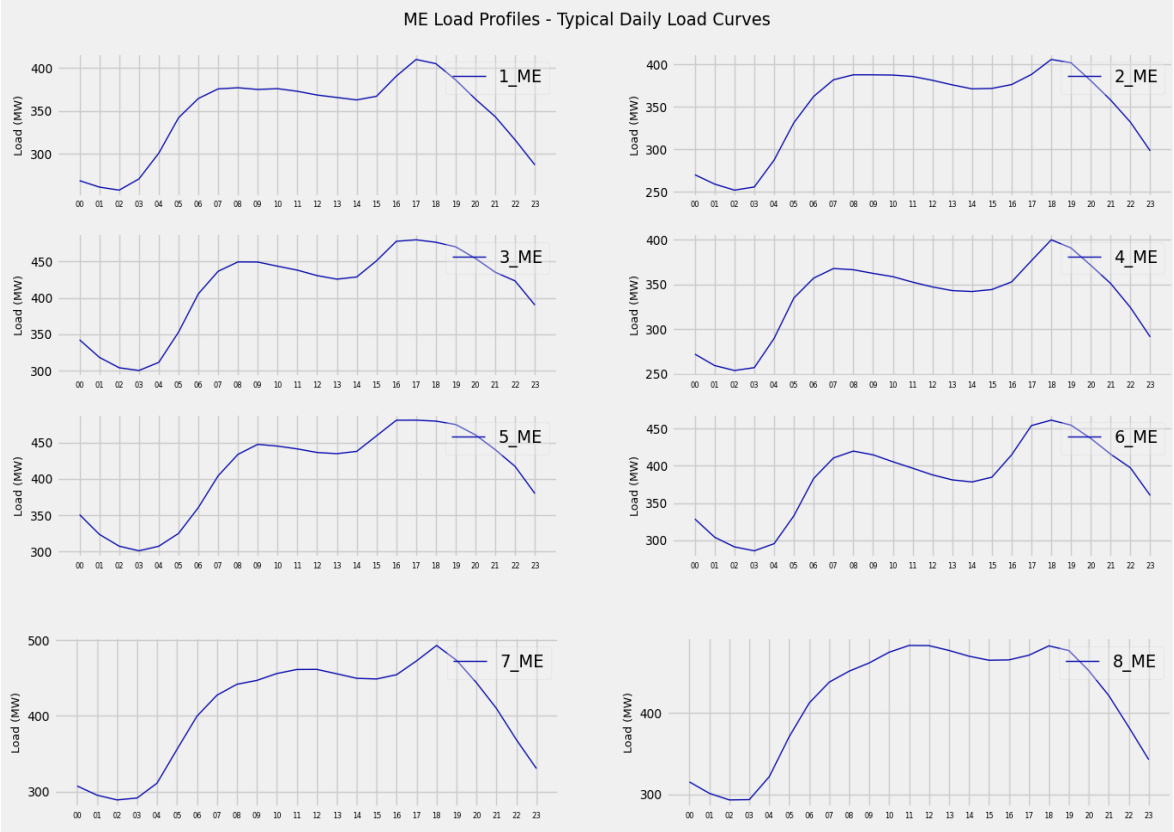
Σχήμα 4.98 : Προφίλ Φορτίου Λιθουανίας (ετήσια ανάλυση).

12) Λετονία (LV)



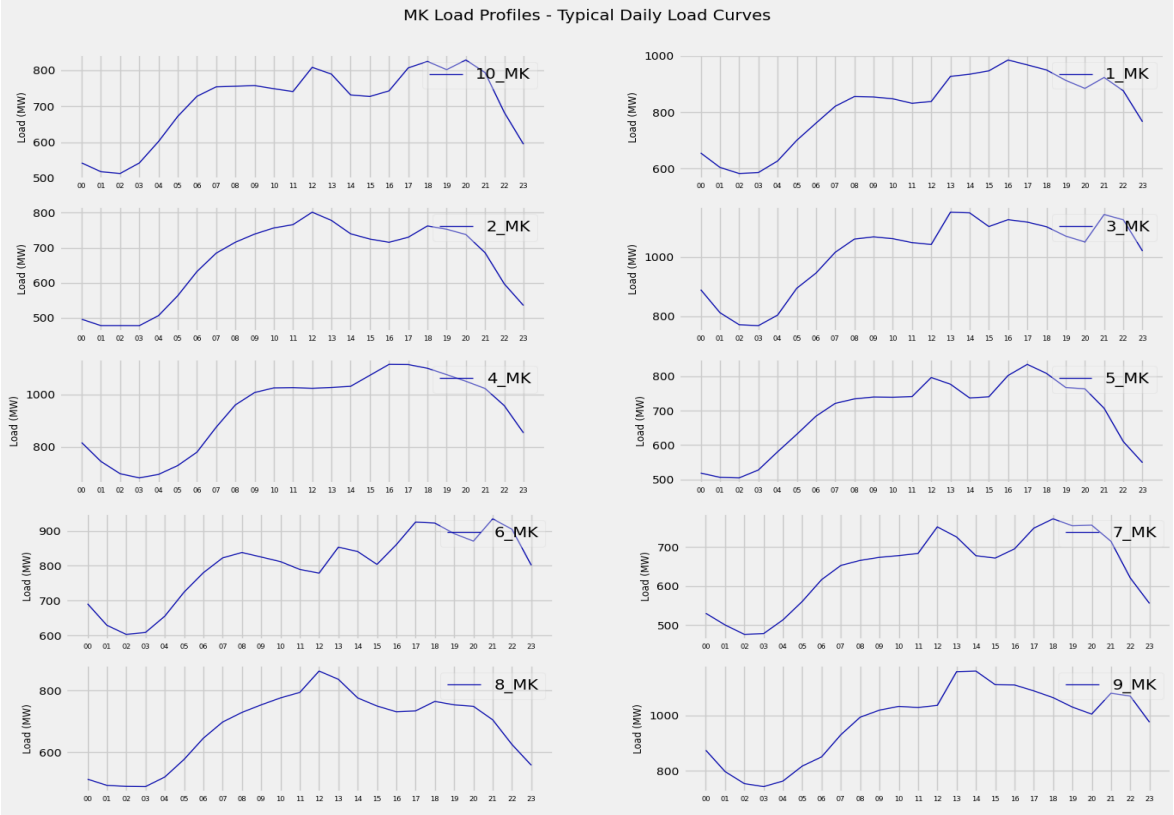
Σχήμα 4.99 : Προφίλ Φορτίου Λετονίας (ετήσια ανάλυση).

13) Μαυροβούνιο (ΜΕ)



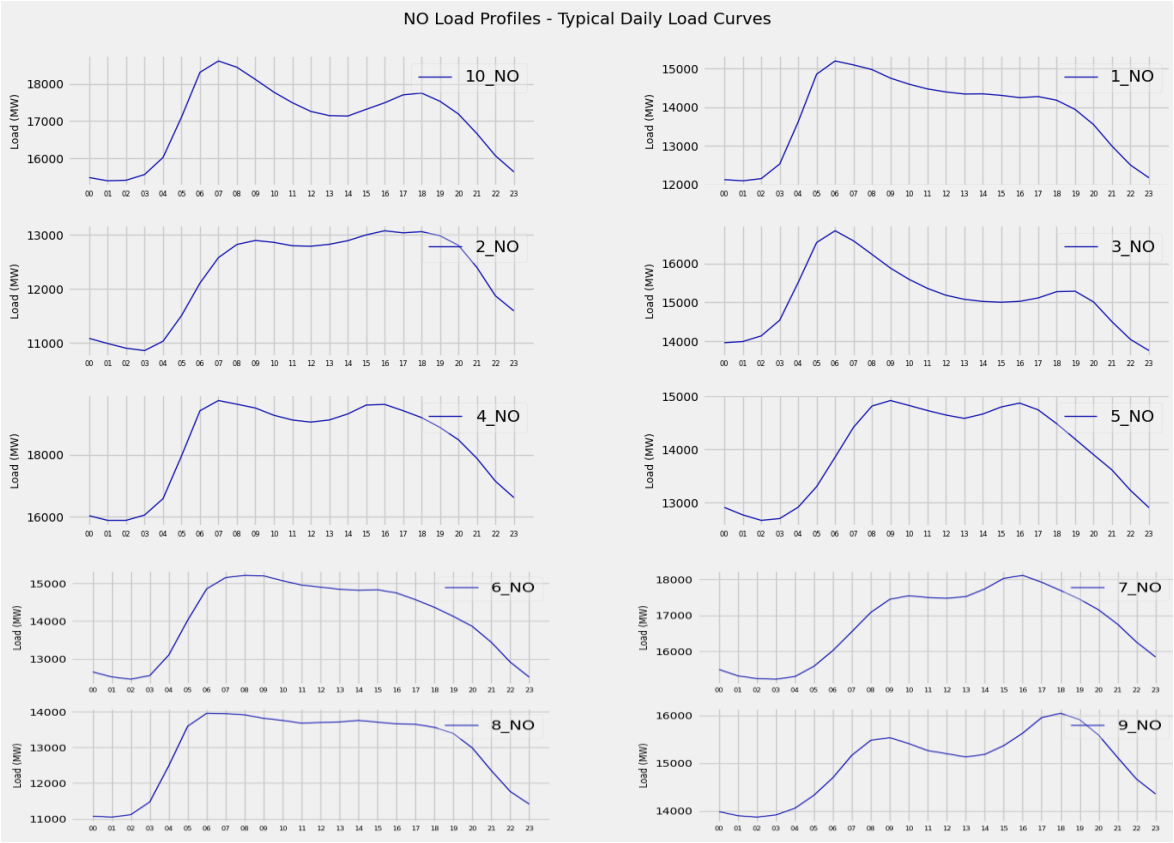
Σχήμα 4.100 : Προφίλ Φορτίου Μαυροβουνίου (ετήσια ανάλυση).

14) Βόρεια Μακεδονία (MK)



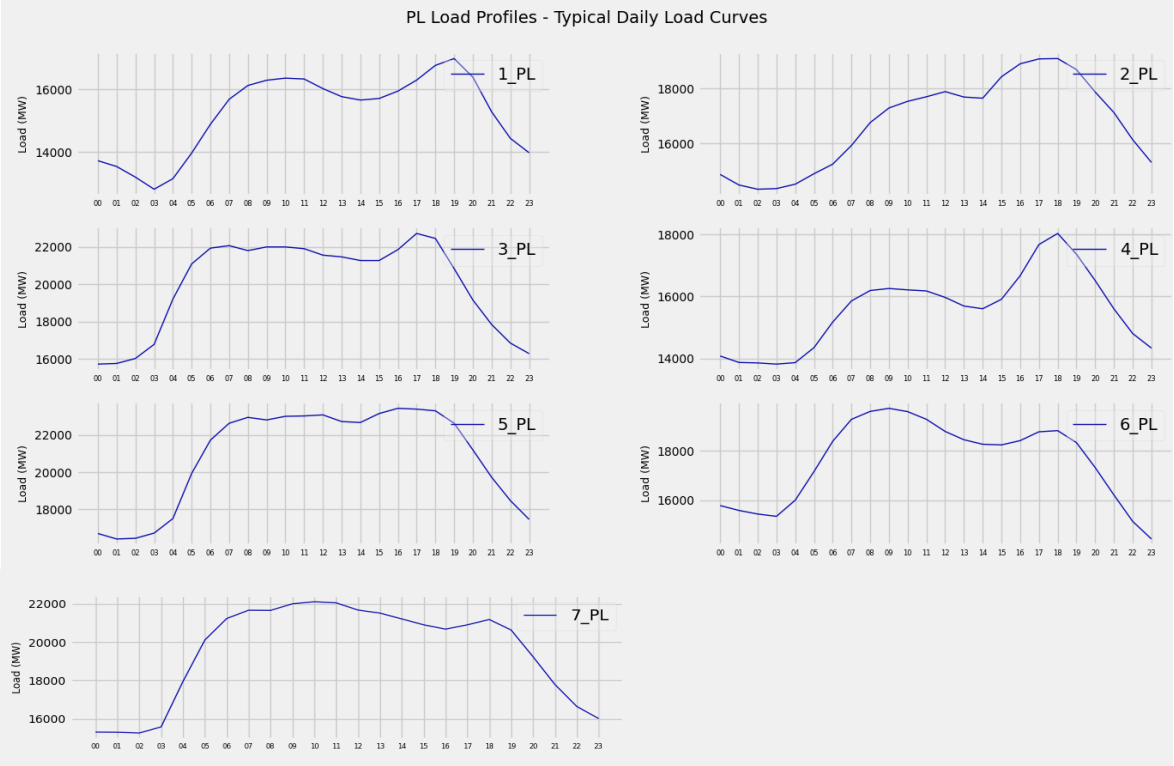
Σχήμα 4.101 : Προφίλ Φορτίου Βόρειας Μακεδονίας (ετήσια ανάλυση).

15) Νορβηγία (NO)



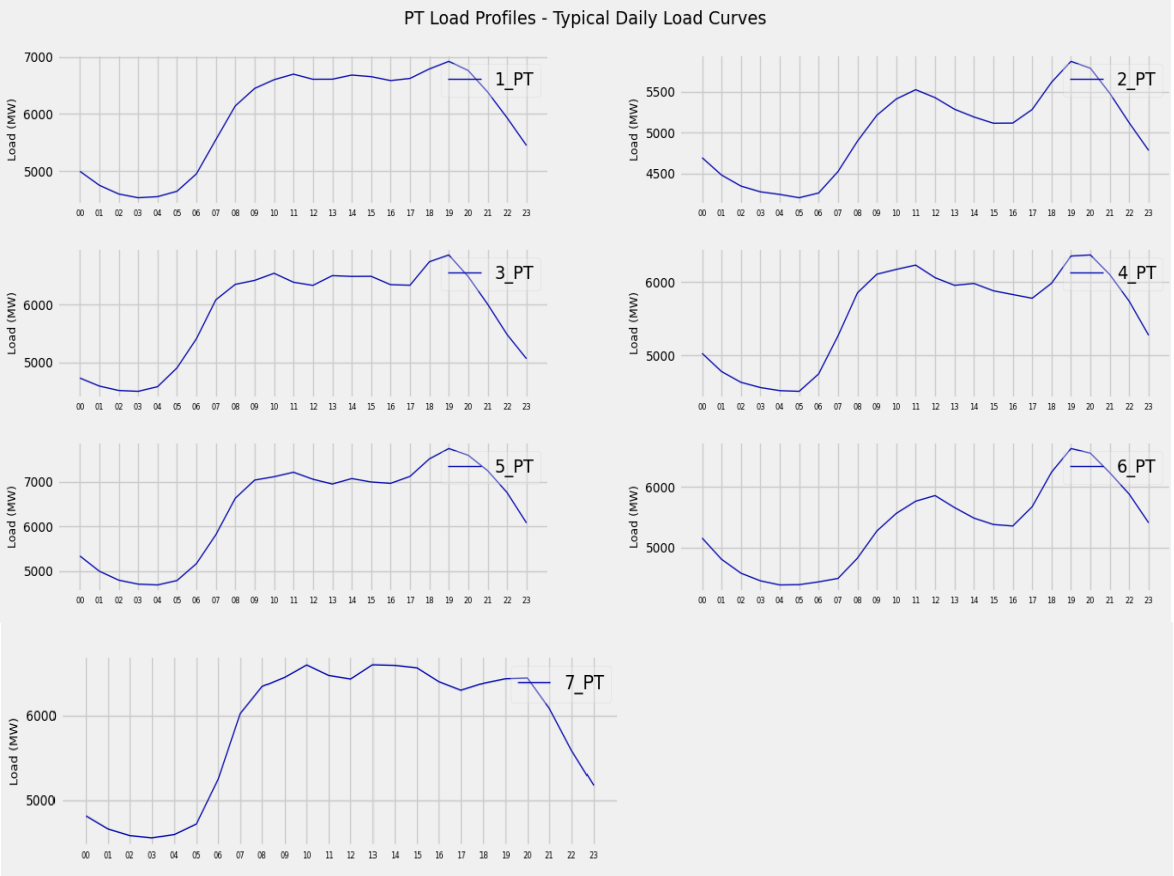
Σχήμα 4.102 : Προφίλ Φορτίου Νορβηγίας (ετήσια ανάλυση).

16) Πολωνία (PL)



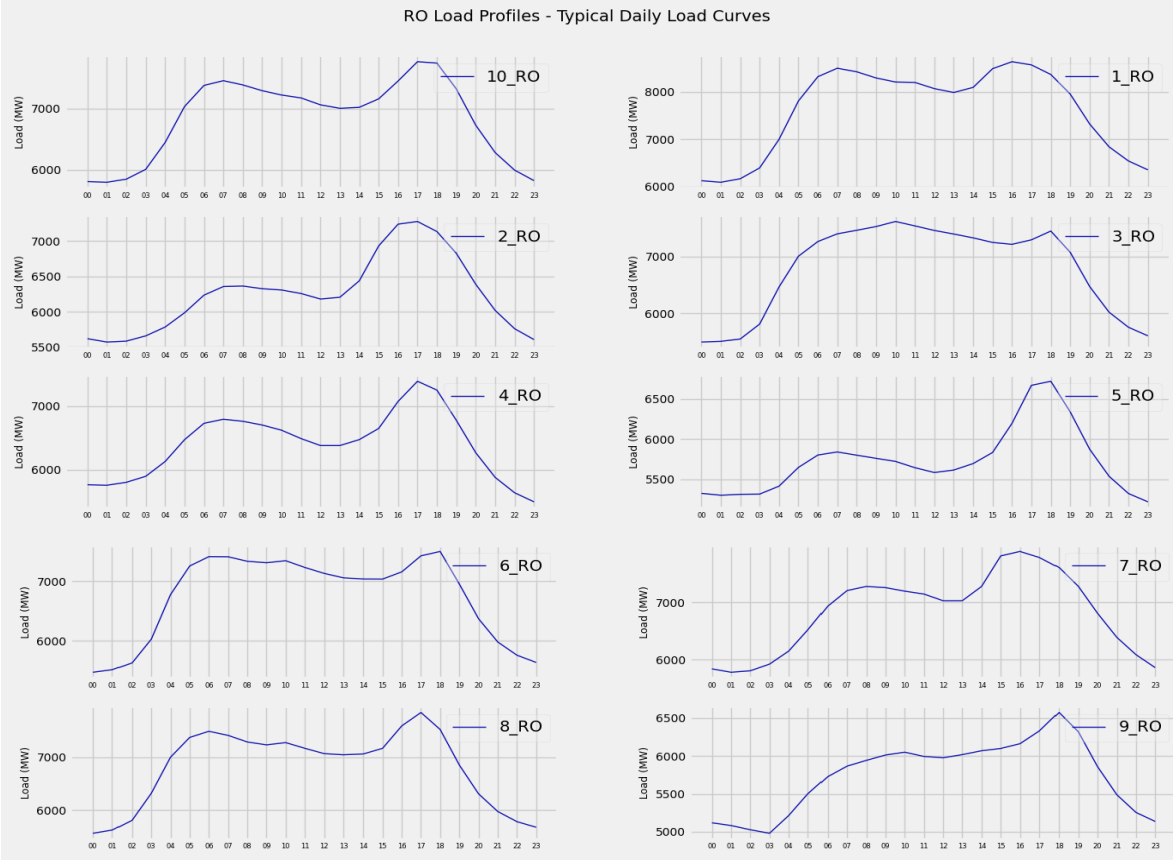
Σχήμα 4.103 : Προφίλ Φορτίου Πολωνίας (ετήσια ανάλυση).

17) Πορτογαλία (PT)



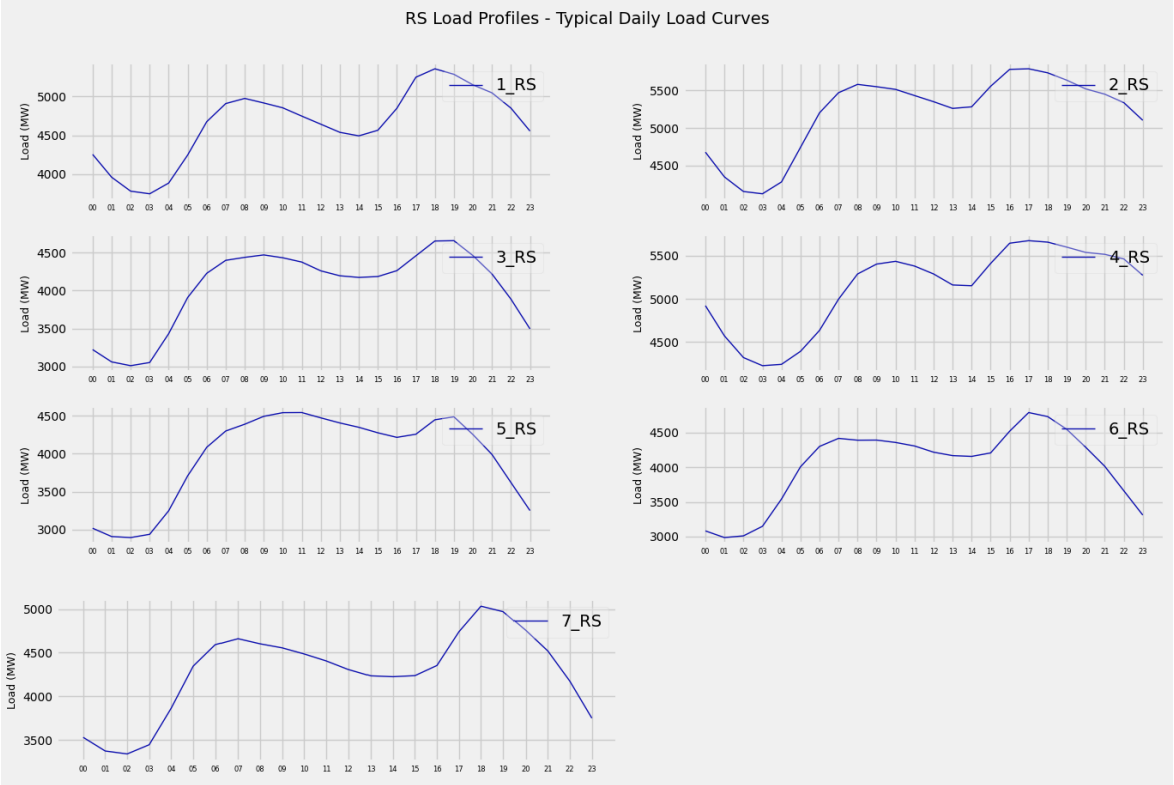
Σχήμα 4.104 : Προφίλ Φορτίου Πορτογαλίας (ετήσια ανάλυση).

18) Ρουμανία (RO)



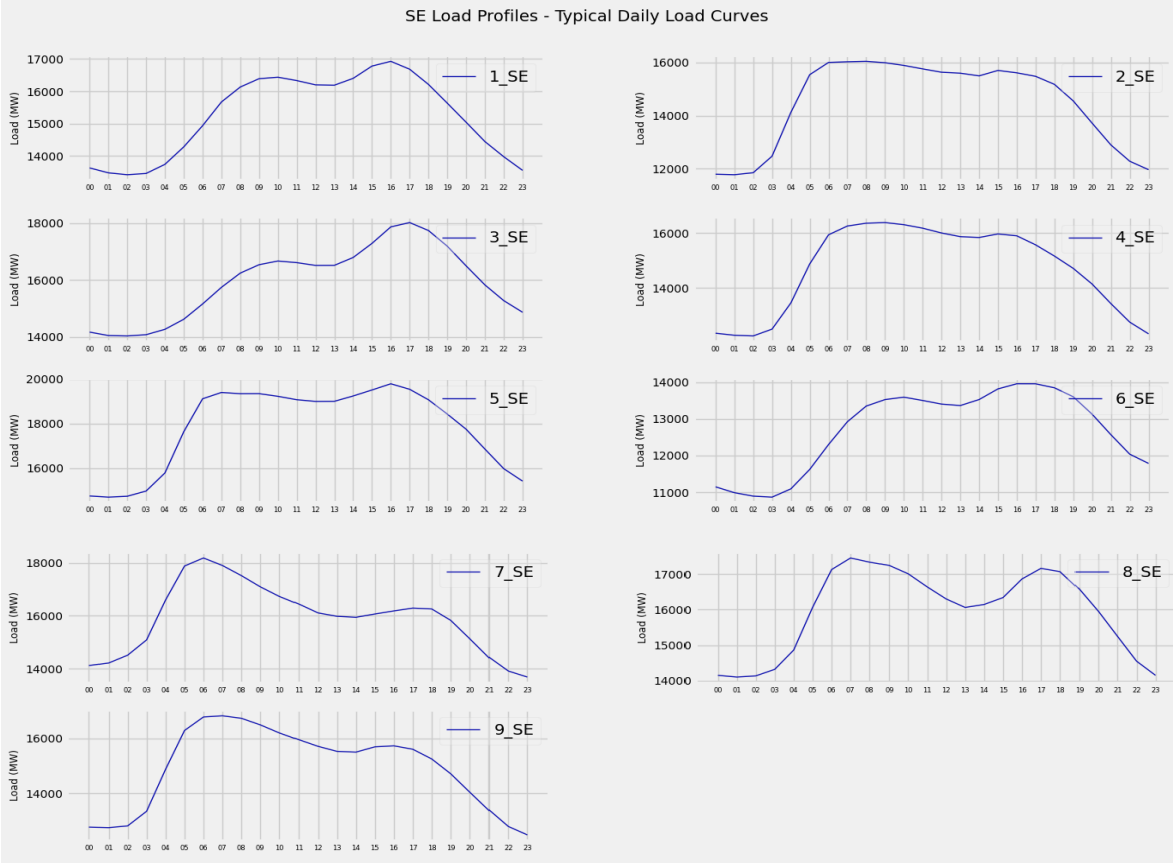
Σχήμα 4.105 : Προφίλ Φορτίου Ρουμανίας (RO) (ετήσια ανάλυση).

19) Σερβία (RS)



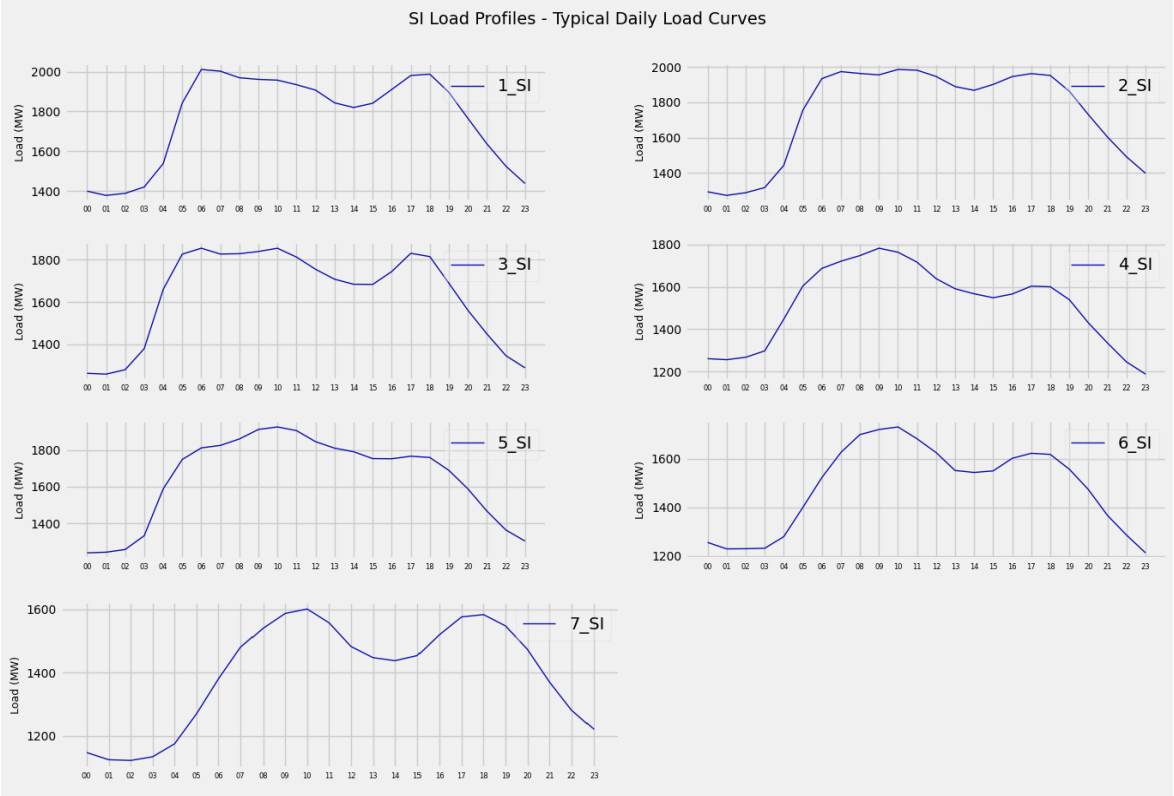
Σχήμα 4.106 : Προφίλ Φορτίου Σερβίας (ετήσια ανάλυση).

20) Σουηδία (SE)



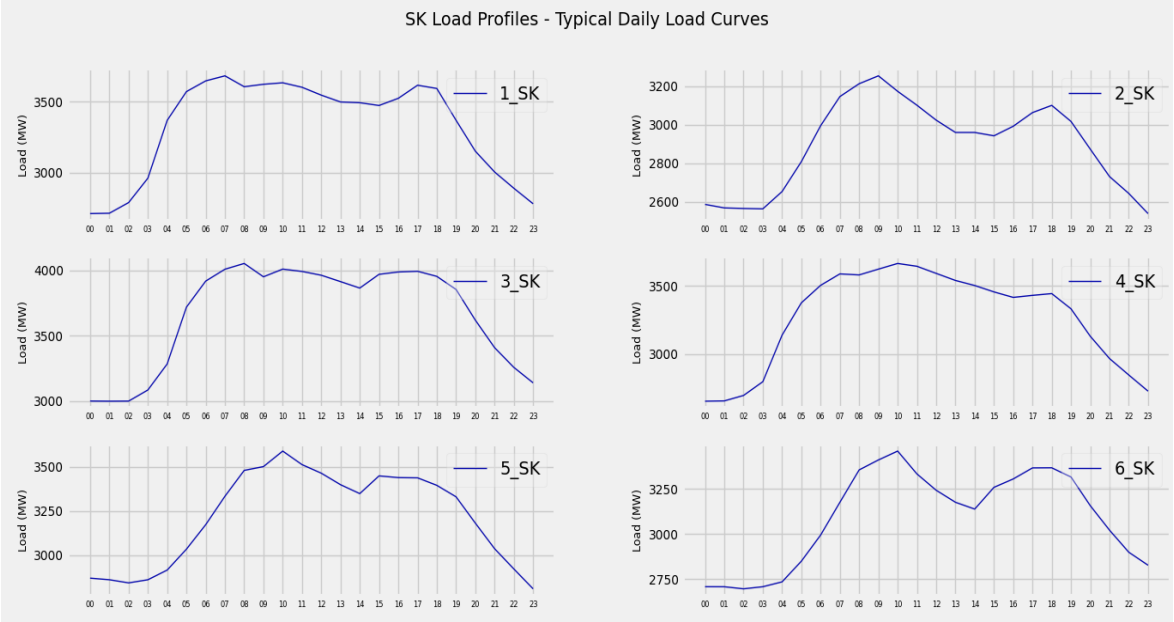
Σχήμα 4.107 : Προφίλ Φορτίου Σουηδίας (ετήσια ανάλυση).

21) Σλοβενία (SI)



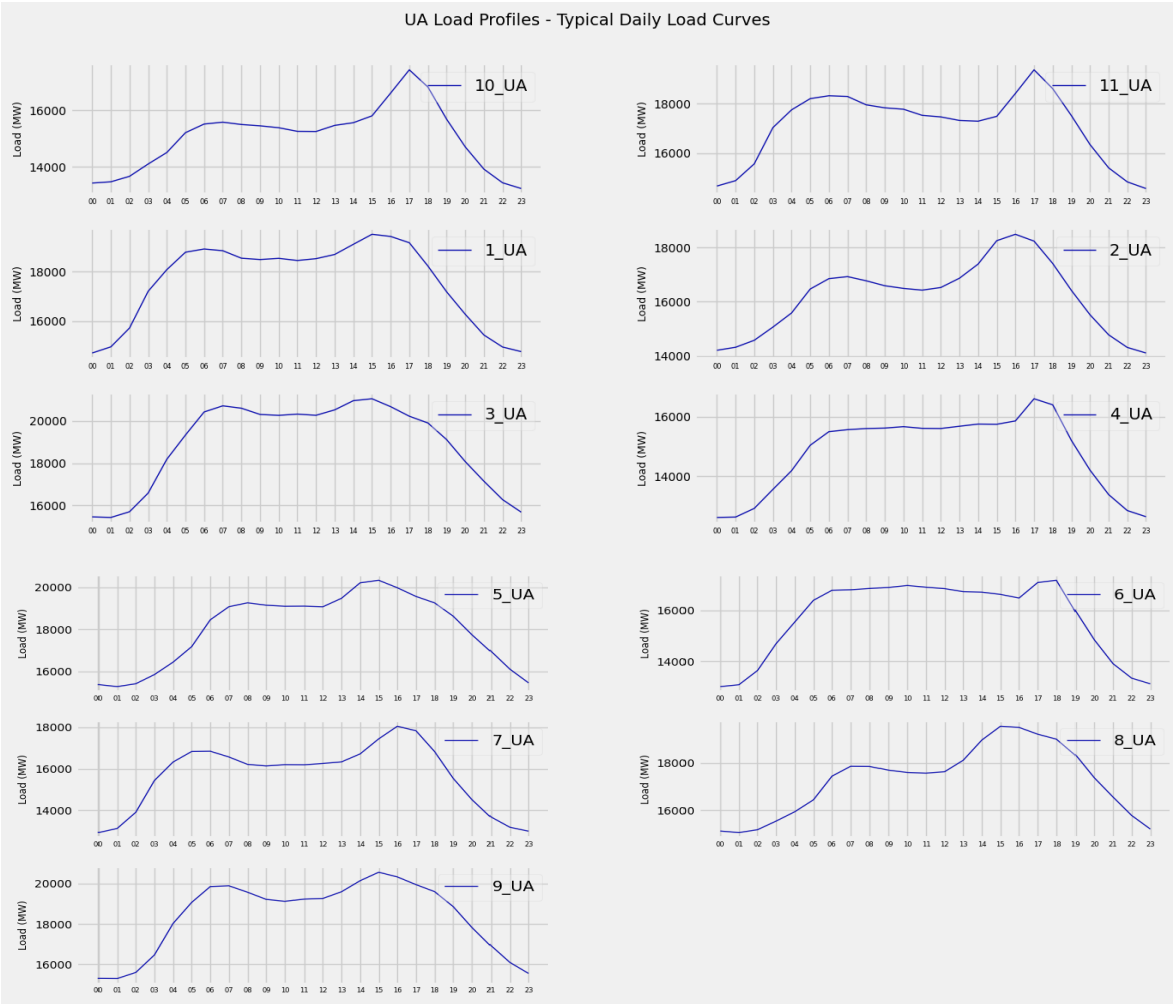
Σχήμα 4.108 : Προφίλ Φορτίου Σλοβενίας (ετήσια ανάλυση).

22) Σλοβακία (SK)



Σχήμα 4.109 : Προφίλ Φορτίου Σλοβακίας (ετήσια ανάλυση).

23) Ουκρανία (UA)



Σχήμα 4.110 : Προφίλ Φορτίου Ουκρανίας (ετήσια ανάλυση).

Πίνακας 4.7 : Αποτελέσματα Μετρικών Επικύρωσης κατά την ετήσια ανάλυση.

CTY	01/03/2019 - 29/02/2020				*Same optimal "k"*	Final "k"
	CH	DB	Silhouette	SSE		
BG	240	1.15	0.33	0.22	Y	9
CZ	238	1	0.41	0.22	N	8
DK	327	0.85	0.45	0.1	N	9
EE	280	1	0.35	0.12	Y	8
ES	318	0.9	0.46	0.2	Y	8
FI	280	1	0.44	0.35	Y	9
FR	210	1.1	0.38	0.2	N	11
GR	150	1.15	0.3	0.1	N	9
HR	155	1.13	0.31	0.1	N	8
IT	520	0.72	0.54	0.15	Y	6
LT	320	0.8	0.39	0.08	Y	8
LV	290	0.9	0.41	0.1	Y	10
ME	145	1.25	0.28	0.14	Y	8
MK	160	1.27	0.34	0.25	Y	12
NO	190	1.1	0.33	0.28	Y	10
PL	290	0.9	0.5	0.15	Y	7
PT	430	0.8	0.48	0.1	Y	7
RO	320	0.95	0.44	0.24	Y	10
RS	210	1.1	0.34	0.14	N	7
SE	182	1.15	0.34	0.27	Y	11
SI	120	1.4	0.22	0.2	Y	10
SK	240	1.15	0.39	0.18	Y	7
UA	150	1.05	0.32	0.16	N	12

4.5 Συμπεράσματα

Αρχικά, σύμφωνα με τα αποτελέσματα της εποχικής ανάλυσης, παρατηρούμε ότι η ποιότητα των ομαδοποιήσεων των ημερήσιων καμπυλών Σ.Π.Φ της Βοσνίας – Ερζεγοβίνης (BA) και της Ελβετίας (CH) ήταν ιδιαίτερα χαμηλή σε σχέση με τις ομαδοποιήσεις των υπολοίπων χωρών του συνόλου ανάλυσης. Συγκεκριμένα, οι δύο αυτές χώρες, σε κάθε εποχή ανάλυσης, παρουσιάζουν συστηματικά τα χειρότερα αποτελέσματα όσον αφορά τις μετρικές επικύρωσης ομαδοποίησης (CVI). Το παραπάνω γεγονός επιβεβαιώνεται παρατηρώντας τα αντίστοιχα γραφήματα των εν λόγω χωρών, τα οποία υποδεικνύουν τη σχετικά απρόβλεπτη ενεργειακή συμπεριφορά τους που αποτυπώνεται στο σχήμα των αντίστοιχων ημερήσιων καμπυλών συνολικού πραγματικού φορτίου (Σ.Π.Φ). Για αυτό το λόγο, αποφασίσαμε να μην συμπεριλάβουμε τις δύο αυτές χώρες κατά την ετήσια ανάλυση, τα αποτελέσματα της οποίας αξιοποιήθηκαν σε εφαρμογές ταξινόμησης.

Ανά εποχή, οι χώρες των οποίων οι ομάδες ημερήσιων καμπυλών Σ.Π.Φ αξιολογήθηκαν ως οι πιο ποιοτικές σύμφωνα με τις μετρικές επικύρωσης ομαδοποίησης είναι οι εξής :

- Άνοιξη : {DK, EE, ES, HR, IT, PL, PT, RO}
- Καλοκαίρι : {CZ, FI, FR, IT, PL, PT, RO}
- Φθινόπωρο : {ES, FR, IT, LT, LV, PL, PT}
- Χειμώνας : {DK, ES, IT, LT, LV, PL, PT}

Αντίστοιχα, κατά την ετήσια ανάλυση τα καλύτερα αποτελέσματα προέκυψαν από τις εξής χώρες : {DK, ES, IT, LT, LV, PL, PT}.

Σύμφωνα με τους πίνακες 4.3 εώς 4.6, για κάθε χώρα κατά τη θερινή και χειμερινή περίοδο ανάλυσης, προέκυψε σε γενικές γραμμές μικρότερο πλήθος ομάδων ημερήσιων καμπυλών Σ.Π.Φ σε αντίθεση με τη φθινοπωρινή και την ανοιξιάτικη περίοδο. Το γεγονός αυτό πιθανώς αιτιολογείται λόγω της σχετικά αυξημένης θερμοκρασιακής μεταβλητότητας που παρουσιάζεται εν γένει τους ανοιξιάτικους και φθινοπωρινούς μήνες, με αποτέλεσμα να επηρεάζεται ανάλογα η παραγωγή και η κατανάλωση ενέργειας.

Επίσης, παρατηρώντας τα Προφίλ Φορτίου της Λιθουανίας (LT), της Λετονίας (LV) και της Εσθονίας (EE) συμπεραίνουμε ότι η ενεργειακή συμπεριφορά τους είναι σε μεγάλο βαθμό παραπλήσια. Δεδομένου ότι οι εν λόγω χώρες είναι ιδιαίτερα μικρής έκτασης και ταυτόχρονα συνορεύουν, η ομοιότητα της ενεργειακής τους συμπεριφοράς ενδεχομένως αιτιολογείται λόγω της ύπαρξης πανομοιότυπων περιβαλλοντικών συνθηκών στην αντίστοιχη γεωγραφική περιοχή καθώς και ότι οι χώρες αυτές μοιράζονται κοινά πολιτιστικά στοιχεία και ιστορία. Αντίστοιχα παρατηρώντας τα Προφίλ Φορτίου της Σερβίας (RS) και του Μαυροβουνίου (ME) συμπεραίνεται ο μεγάλος βαθμός ομοιότητας της ενεργειακής τους συμπεριφοράς.

Η χώρα με τη πιο ιδιαίτερη ενεργειακή συμπεριφορά είναι η Βόρεια Μακεδονία (MK), καθώς οι περισσότερες ημερήσιες καμπύλες Σ.Π.Φ της εν λόγω χώρας παρουσιάζουν αιχμή φορτίου κατά τις μεσημεριανές ώρες. Αντίθετα, όλες οι υπόλοιπες χώρες του συνόλου ανάλυσης εκτός της Ιταλίας (IT), της Βοσνίας-Ερζεγοβίνης (BA) και της Πορτογαλίας (PT) παρουσιάζουν πτώση στο φορτίο τους κατά τις μεσημεριανές ώρες. Η Ιταλία, η Πορτογαλία και η Βοσνία – Ερζεγοβίνη διέπονται από κάποιες θερινές ημερήσιες καμπύλες Σ.Π.Φ οι οποίες παρουσιάζουν επίσης αιχμή φορτίου κατά τις μεσημεριανές ώρες.

Όσον αφορά την Ιταλία, το 3ο "Προφίλ" του σχήματος 4.97 είναι το αντίστοιχο "Προφίλ" Φορτίου που μοντελοποιεί τις θερινές ημερήσιες καμπύλες Σ.Π.Φ οι οποίες παρουσιάζουν αιχμή κατά τις μεσημεριανές ώρες. Η αιχμή φορτίου της Ιταλίας κατά τις θερινές μεσημεριανές ώρες πιθανώς παρατηρείται εξαιτίας του ιδιαίτερα υψηλού τουρισμού που έχει το καλοκαίρι καθώς και λόγω των υψηλών θερμοκρασιών που παρατηρούνται τη περίοδο αυτή.

Όσον αφορά τη Πορτογαλία και τη Βοσνία – Ερζεγοβίνη, η ύπαρξη αιχμής φορτίου κατά τις θερινές μεσημεριανές ώρες πιθανώς προκύπτει λόγω διαφορετικών παραγόντων οι οποίοι μας διαφεύγουν. Οι εν λόγω χώρες διέπονται από μεσογειακό κλίμα με ιδιαίτερα θερμά καλοκαίρια, παρ' όλα αυτά, το εν λόγω φαινόμενο δεν παρατηρείται στις υπόλοιπες χώρες του συνόλου ανάλυσης που διέπονται επίσης από μεσογειακό κλίμα.

Αξίζει να σημειωθεί ότι οι περισσότερες χώρες του συνόλου ανάλυσης που βρίσκονται στη νότια και κεντρική Ευρώπη παρουσιάζουν σε γενικές γραμμές αιχμή φορτίου κατά τις πρωινές και απογευματινές ώρες, ενώ οι περισσότερες χώρες που βρίσκονται στη βόρεια Ευρώπη σε γενικές γραμμές παρουσιάζουν πιο συχνά αιχμή φορτίου κατά τις πρωινές ώρες. Ειδικές περιπτώσεις αποτελούν η Γαλλία και η Βόρεια Μακεδονία, οι οποίες παρουσιάζουν σε ορισμένες ημερήσιες καμπύλες Σ.Π.Φ αιχμή και κατά τις βραδινές ώρες. Επίσης η Ουκρανία και το Μαυροβούνιο που παρουσιάζουν ως επί το πλείστον αιχμή φορτίου κατά τις απογευματινές ώρες, η Πολωνία η οποία παρουσιάζει συνήθως αιχμή φορτίου κατά τις απογευματινές ώρες καθώς και η Δανία και η Φιλανδία οι οποίες παρουσιάζουν αιχμή φορτίου κατά τις πρωινές αλλά και κατά τις απογευματινές ώρες.

Σημαντικό επίσης είναι το γεγονός ότι η Ελλάδα είναι η μόνη χώρα του συνόλου ανάλυσης η οποία παρουσιάζει μέγιστο ηλεκτρικό φορτίο κατά τους θερινούς μήνες, ενώ σε γενικές γραμμές όλες οι υπόλοιπες ευρωπαϊκές χώρες του συνόλου ανάλυσης παρουσιάζουν μέγιστο φορτίο κατά τους χειμερινούς μήνες. Ειδικές περιπτώσεις αποτελούν η Ισπανία και η Ιταλία οι οποίες παρουσιάζουν μέγιστο φορτίο κατά τους θερινούς καθώς και κατά τους χειμερινούς μήνες. Επίσης, ειδική περίπτωση αποτελεί η Γαλλία η οποία παρουσιάζει μέγιστο φορτίο κατά τους χειμερινούς καθώς και κατά τους φθινοπωρινούς μήνες. Αξίζει επίσης να σημειωθεί ότι όλες οι χώρες της βόρειας Ευρώπης παρουσιάζουν ελάχιστο ηλεκτρικό φορτίο κατά τους θερινούς μήνες.

Όσον αφορά την αποδοτικότητα των δεικτών επικύρωσης ομαδοποίησης, σημειώνεται ότι εν γένει οι δείκτες Silhouette και Davies – Bouldin υποδείκνυαν με μεγαλύτερη συνέπεια ποιοτικές ομαδοποιήσεις από ότι ο δείκτης Calinski – Harabasz. Συγκεκριμένα, παρατηρήσαμε σε περιπτώσεις ανάλυσης όπως αυτές του Μαυροβουνίου, της Σερβίας της Νορβηγίας και της Ουκρανίας, ότι ο δείκτης Calinski – Harabasz παρουσίαζε απαισιόδοξα αποτελέσματα σε αντίθεση με τους δείκτες Silhouette και Davies – Bouldin. Στις εν λόγω περιπτώσεις, μέσω της εποπτείας των γραφημάτων των αντίστοιχων ομάδων, οδηγηθήκαμε στο συμπέρασμα ότι ο δείκτης Calinski – Harabasz αξιολογούσε ελαττωματικά τη ποιότητα των ομάδων που είχαν προκύψει. Επίσης, ο δείκτης SSE παρουσίασε τη μεγαλύτερη ευρωστία σε σχέση με τους υπόλοιπους δείκτες καθώς σε καμία περίπτωση ανάλυσης δεν υπερεκτίμησε αλλά ούτε υποτίμησε την ποιότητα των ομαδοποιήσεων. Ένα παράδειγμα αποτελεί η περίπτωση της θερινής ανάλυσης της Ελλάδας, όπου ο δείκτης SSE παρουσίασε αισιόδοξα αποτελέσματα σε αντίθεση με όλους τους υπόλοιπους δείκτες. Μέσω της εποπτείας των θερινών ομάδων ημερήσιων καμπυλών Σ.Π.Φ της Ελλάδας, παρατηρήσαμε ότι τα αποτελέσματα όλων των δεικτών επικύρωσης εκτός του SSE ήταν ιδιαίτερα αυστηρά.

Σε περιπτώσεις ανάλυσης όπως της Ελλάδας, όπου επεξεργαζόμαστε ημερήσιες καμπύλες Σ.Π.Φ με αρκετά παρόμοιο σχήμα οι οποίες όμως διαφοροποιούνται ως προς τη χρονική στιγμή που παρουσιάζουν το μέγιστο φορτίο τους (κατά μία ώρα), οι εν λόγω δείκτες εκτός του SSE αξιολογούν αυστηρά τις ομαδοποιήσεις των αντίστοιχων δεδομένων καθώς θεωρούν ότι ο διαχωρισμός των ομάδων είναι ανεπαρκής. Όμως, υπό το πρίσμα της ενεργειακής συμπεριφοράς, η μετατόπιση του φορτίου αιχμής κατά μια ώρα είναι επαρκώς σημαντική ώστε να θεωρήσουμε δύο διαφορετικές ομάδες ακόμα και αν το σχήμα των ημερήσιων καμπυλών Σ.Π.Φ είναι πρακτικά όμοιο.

Επίσης, σημειώνεται ότι σε γενικές γραμμές οι δείκτες επικύρωσης συμφωνούσαν ως προς το βέλτιστο αριθμό των ομάδων, δηλαδή το βέλτιστο "k". Συγκεκριμένα, δεδομένου ότι υλοποιήθηκαν εκατόν είκοσι τρεις (123) εφαρμογές ομαδοποίησης, οι δείκτες επικύρωσης υπέδειξαν τον ίδιο αριθμό "k" στο 71.5% των εφαρμογών.

Τέλος, όσον αφορά την ετήσια ανάλυση, όπως ήδη έχουμε αναφέρει κατά τη περιγραφή της μεθοδολογίας των εφαμοργών ομαδοποίησης, οι μετρικές επικύρωσης υποδείκνυαν "βέλτιστες" ομαδοποιήσεις οι οποίες ήταν σε γενικές γραμμές ικανοποιητικές κατά την οπτικοποίηση των ομάδων, όμως συνήθως οι ομάδες αυτές δεν διέπονταν από επαρκή θερμοκρασιακό διαχωρισμό μεταξύ τους. Αυξάνοντας τον αριθμό των ομάδων μέχρι να επιτευχθεί ένας επαρκής θερμοκρασιακός διαχωρισμός, οδηγηθήκαμε στο συμπέρασμα ότι οι μετρικές επικύρωσης υπολείπονται ως ένα βαθμό σε επιθυμητή ακρίβεια, καθώς με την αύξηση των ομάδων αναδύθηκαν σημαντικά πρότυπα τα οποία δεν εντοπίστηκαν από τις εν λόγω μετρικές επικύρωσης.

Κεφάλαιο 5 : Εφαρμογές Επιβλεπόμενης Μάθησης

Στο κεφάλαιο αυτό παρουσιάζουμε αναλυτικά τη διαδικασία και τα αποτελέσματα των εφαρμογών Επιβλεπόμενης Μάθησης που υλοποιήσαμε σε ενεργειακά δεδομένα. Συγκεκριμένα, οι εφαρμογές αφορούν την ταξινόμηση ημερήσιων χρονοσειρών ηλεκτρικού φορτίου σε είκοσι τρεις (23) κλάσεις, οι οποίες αντιστοιχούν στις είκοσι τρεις ευρωπαϊκές χώρες που επεξεργαστήκαμε κατά την ετήσια ανάλυση του τετάρτου κεφαλαίου. Στόχος των εν λόγω εφαρμογών είναι η αξιοποίηση των "Προφίλ" Φορτίου για την εκπαίδευση των μοντέλων ταξινόμησης που παρουσιάσαμε στο δεύτερο κεφάλαιο, και η αξιολόγηση των εν λόγω μοντέλων ως προς την ικανότητα τους να προβλέπουν τις χώρες στις οποίες ανήκουν οι ημερήσιες καμπύλες φορτίου του συνόλου ελέγχου. Συνεπώς, το πρόβλημα που καλούμαστε να λύσουμε είναι ένα πρόβλημα ταξινόμησης πολλών κλάσεων (multiclass classification). Η επίδοση των μοντέλων θα αποτελέσει ένα μέτρο της ποιότητας των αποτελεσμάτων των εφαμογών ομαδοποίησης και εξαγωγής "Προφίλ" Φορτίου, καθώς επίσης και μία γενική εικόνα για το πόσο αποδοτικά δύνανται να διαχωριστούν τα δεδομένα.

Πιο αναλυτικά, το σύνολο δεδομένων εκπαίδευσης (training set) αποτελείται από τις εκατόν ενενήντα δύο (192) χαρακτηριστικές ημερήσιες καμπύλες φορτίου ("Προφίλ" Φορτίου) των είκοσι τριών (23) ευρωπαϊκών χωρών που υπολογίσαμε κατά την ετήσια ανάλυση του χρονικού παραθύρου $T.W_1 = [\text{"01/03/2019"}, \text{"29/02/2020"}]$. Το σύνολο επικύρωσης (validation set) αποτελείται από όλες τις ημερήσιες καμπύλες φορτίου του χρονικού παραθύρου $T.W_1$ των ευρωπαϊκών χωρών – κλάσεων του προβλήματος. Σημειώνεται ότι το σύνολο επικύρωσης αξιοποιήθηκε για τη βελτιστοποίηση των μοντέλων.

Για την αξιολόγηση των μοντέλων αξιοποιήθηκαν τα εξής πέντε σύνολα ελέγχου :

i. $\boldsymbol{TestSet_1}$ **:** Ημερήσιες καμπύλες φορτίου του χρονικού παραθύρου "01/01/2015" εώς "31/12/2015".

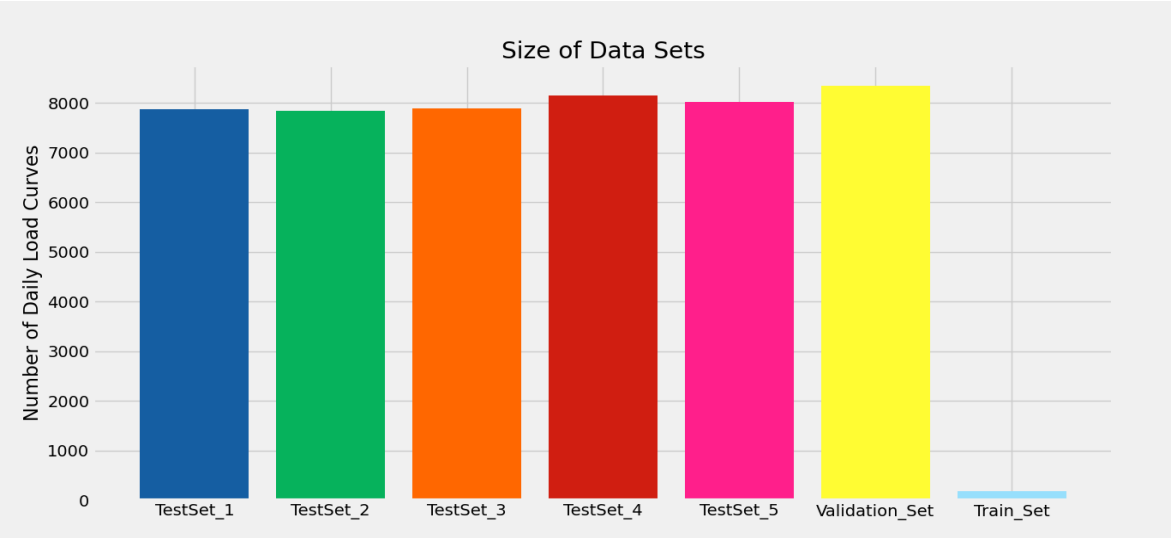
ii. $\boldsymbol{TestSet_2}$ **:** Ημερήσιες καμπύλες φορτίου του χρονικού παραθύρου "01/03/2016" εώς "28/02/2017".

iii. $\boldsymbol{TestSet_3}$ **:** Ημερήσιες καμπύλες φορτίου του χρονικού παραθύρου "01/03/2017" εώς "28/02/2018".

iv. $\boldsymbol{TestSet_4}$ **:** Ημερήσιες καμπύλες φορτίου του χρονικού παραθύρου "01/03/2018" εώς "28/02/2019".

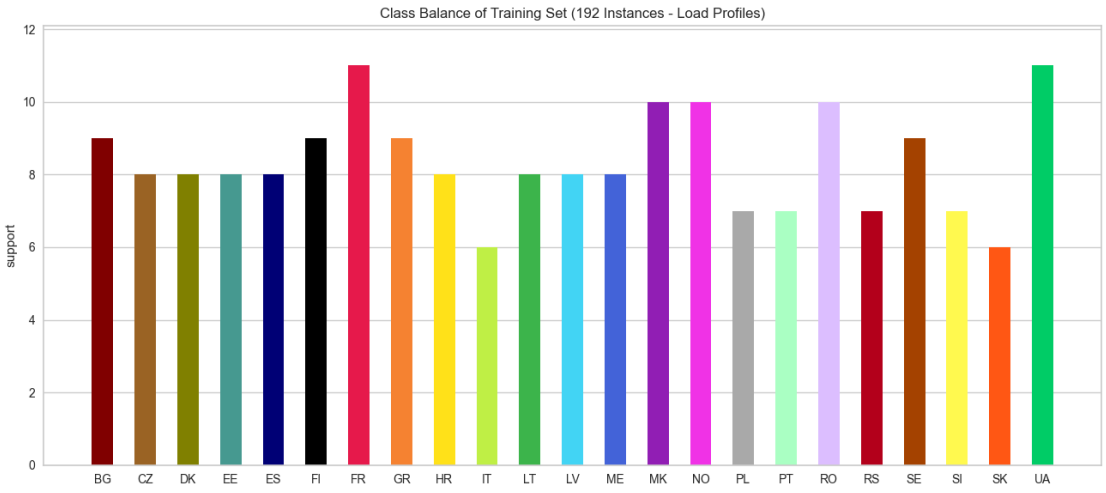
v. $\boldsymbol{TestSet_5}$ **:** Ημερήσιες καμπύλες φορτίου του χρονικού παραθύρου "01/03/2020" εώς "28/02/2021".

Ο συνολικός αριθμός των αντίστοιχων δεδομένων (χρονοσειρών) των συνόλων εκπαίδευσης, επικύρωσης και ελέγχου υποδεικνύεται στο παρακάτω γράφημα.



Σχήμα 5.1 : Μέγεθος συνόλων δεδομένων (Ελέγχου, Επικύρωσης και Εκπαίδευσης).

Στο επόμενο σχήμα υποδεικνύεται το πλήθος των δεδομένων εκπαίδευσης ("Προφίλ" Φορτίου) ανά κατηγορία - κλάση.



Σχήμα 5.2 : Κατανομή των δεδομένων εκπαίδευσης ως προς τις κλάσεις του προβλήματος.

Στο σημείο αυτό, όπως υποδεικνύεται από το σχήμα 5.1 αναφέρουμε ότι το σύνολο δεδομένων εκπαίδευσης είναι ιδιαίτερα μικρό σε σχέση με τα αντίστοιχα σύνολα ελέγχου και επικύρωσης. Το γεγονός αυτό, όπως έχουμε ήδη αναφέρει σε προηγούμενα κεφάλαια, ενδέχεται να οδηγήσει στην υπερμοντελοποίηση (overfitting) των μοντέλων ταξινόμησης. Στο κεφάλαιο 2.6 έχουμε αναφέρει τις βασικές αιτίες καθώς και μεθοδολογίες αντιμετώπισης του εν λόγω προβλήματος. Συνεπώς, το να καταφέρουμε να κατασκευάσουμε μοντέλα ταξινόμησης (τα οποία έχουν εκπαιδευτεί με τα "Προφίλ" Φορτίου) τα οποία παρουσιάζουν υψηλές επιδόσεις όσον αφορά τα δεδομένα ελέγχου, αποτελεί μια πρόκληση μέσω της οποίας θα επιχειρήσουμε να προσδιορίσουμε τη καταλληλότητα των αποτελεσμάτων των εφαρμογών ομαδοποίησης και εξαγωγής "Προφίλ" Φορτίου.

5.1 Προετοιμασία Δεδομένων για Εφαρμογές Ταξινόμησης

Στο κεφάλαιο αυτό παρέχουμε την αναλυτική περιγραφή της διαδικασίας που ακολουθήσαμε για τη κατάλληλη προετοιμασία των δεδομένων ώστε να αξιοποιηθούν σε εφαρμογές ταξινόμησης. Στόχος της εν λόγω διαδικασίας είναι η επισήμανση (labeling) των δεδομένων εισόδου, δηλαδή η δημιουργία των διανυσμάτων - στόχων Y_i που περιέχουν τις επιθυμητές εξόδους – κλάσεις των δεδομένων. Η εν λόγω διαδικασία πραγματοποιήθηκε για κάθε σύνολο δεδομένων (εκπαίδευσης, επικύρωσης και ελέγχου) που αναλύσαμε στην εισαγωγή του εν λόγω κεφαλαίου. Σημειώνεται ότι τα συνολα επικύρωσης και ελέγχου αποτελούν ουσιαστικά τη βάση δεδομενων ημερήσιων χρονοσειρών φορτίου DB_2 που δημιουργήσαμε κατά το δεύτερο στάδιο προεπεξεργασίας του τετάρτου κεφαλαίου. Όπως είναι ήδη γνωστό, στα αρχεία της βάσης δεδομένων DB_2 περιέχονται ορισμένα διανύσματα εισόδου τα οποία περιέχουν ελλιπείς τιμές που υποδεικνύονται από τον ειδικό τύπο δεδομένων "NaN". Συνεπώς, σε αρχικό στάδιο απαιτείται η απαλοιφή των εν λόγω διανυσμάτων. Σε δεύτερο στάδιο, καθώς τα αρχικά πλαίσια δεδομένων (dataframes) είναι δεικτοποιημένα (indexed) με τους "MapCode" κωδικούς, δημιουργούμε νέα πλαίσια δεδομένων που περιέχουν ως στήλες όλα τα διανύσματα εισόδου. Ονοματίζουμε κάθε στήλη με βάση το αρχικό αναγνωριστικό (Day_n) του αντίστοιχου διανύσματος καθώς και τον αντίστοιχο του "MapCode" κωδικό ως εξής :

$$Day\_n\_"MapCode"$$

Η τυπική μορφή των αρχικών δεικτοποιημένων πλαισίων δεδομένων υποδεικνύεται στο σχήμα 4.11 .

Στη συνέχεια απαιτείται η αναστροφή (transpose) των νέων πλαισίων δεδομένων ώστε να έχουμε ως σειρές τα διανύσματα εισόδου (input vectors – samples) και ως στήλες τις χρονικά διατεταγμένες τιμές Σ.Π.Φ (timesampled measurments). Κατά την αναστροφή των νέων πλαισίων, τα αναγνωριστικά των στηλών μετατρέπονται σε δείκτες των αντίστοιχων σειρών. Στο σημείο αυτό, τα ανεστραμμένα πλαίσια δεδομένων df_i_final είναι σε κατάλληλη μορφή για να αποτελέσουν σύνολα εισόδου σε μοντέλα ταξινόμησης.

Σε τελικό στάδιο απαιτείται η κατάλληλη επισήμανση των διανυσμάτων εισόδου που περιέχονται στα πλαίσια δεδομένων df_i_final. Για την δημιουργία των στόχων Y_i υλοποιήσαμε κατάλληλη συνάρτηση που εξάγει τα δύο τελευταία σύμβολα (που αντιστοιχούν σε "MapCode" κωδικούς) των δεικτών κάθε σειράς και τα αποθηκεύει σε μια λίστα. Στη συνέχεια, η εν λόγω συνάρτηση επιστρέφει τη λίστα ως ένα πλαίσιο δεδομένων.

	0	1	2	...	21	22	23
Day_1_BG	5259.0	5240.0	5271.0	...	5836.0	5192.0	4930.0
Day_2_BG	4824.0	4824.0	4838.0	...	5270.0	5097.0	4775.0
Day_3_BG	4585.0	4533.0	4523.0	...	4889.0	4543.0	4296.0
Day_4_BG	4166.0	4101.0	4090.0	...	5195.0	4591.0	4291.0
Day_5_BG	4170.0	4112.0	4146.0	...	5294.0	4594.0	4234.0
...	...	...	...	...	...	...	...
Day_361_UA	17531.0	17709.0	17928.0	...	17796.0	17230.0	16938.0
Day_362_UA	16953.0	17353.0	18396.0	...	18017.0	17542.0	17377.0
Day_363_UA	17387.0	17620.0	18541.0	...	18121.0	17589.0	17377.0
Day_364_UA	17423.0	17652.0	18350.0	...	17787.0	17231.0	17093.0
Day_365_UA	17052.0	17359.0	18252.0	...	17368.0	16810.0	16618.0

[8152 rows x 24 columns]

X_dataset

	id	target
0	Day_1_BG	BG
1	Day_2_BG	BG
2	Day_3_BG	BG
3	Day_4_BG	BG
4	Day_5_BG	BG
...	...	...
8147	Day_361_UA	UA
8148	Day_362_UA	UA
8149	Day_363_UA	UA
8150	Day_364_UA	UA
8151	Day_365_UA	UA

[8152 rows x 2 columns]

Σχήμα 5.3 : Τυπική μορφή πλαισίων δεδομένων X και Y για κατηγοριοποίηση.

5.2 Εφαρμογές Ταξινόμησης στο Πεδίο του Χρόνου (Instance Based Classification Applications)

Για την υλοποίηση των εν λόγω εφαρμογών Ταξινόμησης ή αλλιώς Κατηγοριοποίησης αξιοποιήθηκε η βιβλιοθήκη scikit-learn η οποία παρέχει πληθώρα αλγορίθμων ταξινόμησης και μετρικών αξιολόγησης. Η περιγραφή των αλγορίθμων και των μετρικών αξιολόγησης που αξιοποιήθηκαν έχει υλοποιηθεί στο δεύτερο κεφάλαιο της παρούσας διπλωματικής εργασίας.

Κάθε μοντέλο ταξινόμησης που κατασκευάσαμε εκπαιδεύτηκε, βελτιστοποιήθηκε και αξιολογήθηκε στα ίδια σύνολα εκπαίδευσης, επικύρωσης και ελέγχου αντίστοιχα. Η περιγραφή των εν λόγω συνόλων έχει πραγματοποιηθεί στην αρχή του πέμπτου κεφαλαίου. Επίσης, η περιγραφή της μεθοδολογίας βελτιστοποίησης των υπερπαραμέτρων των μοντέλων έχει πραγματοποιηθεί στη πέμπτη ενότητα του δεύτερου κεφαλαίου.

Στη συνέχεια παρουσιάζουμε τα τελικά μοντέλα ταξινόμησης καθώς και τα αντίστοιχα αποτελέσματα των μετρικών αξιολόγησης.

1. **<u>Μοντέλο Ταξινόμησης Δένδρου Απόφασης</u>**

   Τα μοντέλα ταξινόμησης Δένδρων Αποφάσεων (Decision Trees) δεν απαιτούν τη κανονικοποίηση των δεδομένων εισόδου. Συνεπώς, τα δεδομένα εισόδου του εν λόγω μοντέλου δεν κανονικοποιήθηκαν.

   Το τελικό βελτιστοποιημένο μοντέλο ταξινόμησης Δένδρου Απόφασης είναι το εξής

```
clf_dt = DecisionTreeClassifier(criterion="entropy", random_state=101, max_depth=10, min_samples_split=2,
                                min_samples_leaf=1, max_features=8, max_leaf_nodes=None)
```

   Σημειώνεται ότι το συγκεκριμένο μοντέλο είναι ένα πλήρες ανεπτυγμένο δένδρο. Στο εν λόγω τελικό μοντέλο δεν εφαρμόσαμε τεχνικές κλαδέματος δεδομένου ότι παρατηρήσαμε πτώση στην επίδοση του ως προς τα δεδομένα επικύρωσης κατά το κλάδεμα του.

Σχήμα 5.4 : Διάγραμμα Cost Complexity Pruning του Instance Based ταξινομητή Δένδρου Απόφασης.

Στη συνέχεια παρουσιάζουμε τον πίνακα με τα αποτελέσματα των μετρικών αξιολόγησης. Σημειώνεται ότι τα εν λόγω μέτρα έχουν υπολογισθεί με βάση τη στρατηγική "macro averaging" και έχουν στρογγυλοποιηθεί στο τρίτο δεκαδικό ψηφίο. Το ίδιο ισχύει και για όλα τα υπόλοιπα μοντέλα ταξινόμησης.

Πίνακας 5.1 : Αποτελέσματα μετρικών αξιολόγησης Instance Based Decision Tree Classifier.

	Instance Based Decision Tree Classifier					
Evaluation Metrics	$TestSet_1$	$TestSet_2$	$TestSet_3$	$TestSet_4$	$TestSet_5$	*Validation_Set*
Precision	0.659	0.626	0.653	0.637	0.61	0.692
Recall	0.64	0.615	0.646	0.636	0.603	0.695
Accuracy	0.668	0.64	0.673	0.638	0.598	0.695
Balanced Accuracy	0.669	0.643	0.675	0.636	0.6	0.695
F1_Score	0.641	0.61	0.643	0.626	0.599	0.691
Jaccard Score	0.5	0.47	0.5	0.469	0.427	0.53

Σύμφωνα με τα αποτελέσματα των μετρικών αξιολόγησης συμπεραίνουμε ότι το μοντέλο έχει υπερμοντελοποιηθεί στα δεδομένα εκπαίδευσης καθώς η επίδοση του στα δεδομένα ελέγχου είναι σχετικά χαμηλή.

2. <u>Μοντέλο Ταξινόμησης Κ – Κοντινότερων Γειτόνων</u>

Ο αλγόριθμος Κ – Κοντινότερων Γειτόνων (K-NN) επηρεάζεται αρνητικά όταν υπάρχουν μεγάλες αποκλίσεις ως προς τη τάξη μεγέθους των τιμών των δεδομένων. Συνεπώς, για τη κατασκευή του εν λόγω μοντέλου ταξινόμησης απαιτείται η κανονικοποίηση των δεδεομένων εισόδου. Δοκιμάσαμε δύο διαφορετικές μεθόδους κανονικοποίησης, τη μέθοδο Min Max Scaling και τη μέθοδο Standard Scaling εκ των οποίων επιλέξαμε τη δεύτερη καθώς προσέφερε λίγο καλύτερα αποτελέσματα.

Το τελικό βελτιστοποιημένο μοντέλο ταξινόμησης K-NN είναι το εξής :

```
print("=========================      K-NN CLASSIFIER Hyperparameter Optimized   ======================", end="\n\n")
clf_knn = neighbors.KNeighborsClassifier(n_neighbors=1, weights="distance", algorithm="auto", leaf_size=30, p=2)
```

Πίνακας 5.2 : Αποτελέσματα μετρικών αξιολόγησης Instance Based K-NN Classifier.

	Instance Based K-NN Classifier					
Evaluation Metrics	$TestSet_1$	$TestSet_2$	$TestSet_3$	$TestSet_4$	$TestSet_5$	*Validation_Set*
Precision	0.579	0.616	0.635	0.656	0.621	0.727
Recall	0.566	0.598	0.623	0.653	0.615	0.723
Accuracy	0.593	0.628	0.652	0.658	0.612	0.722
Balanced Accuracy	0.591	0.625	0.651	0.653	0.615	0.723
F1_Score	0.563	0.597	0.624	0.647	0.613	0.723
Jaccard Score	0.422	0.458	0.483	0.49	0.441	0.565

Με βάση τα αποτελέσματα του πίνακα 5.2 συμπεραίνουμε ότι ο Time Instance Based ταξινομητής K-NN έχει επίσης υπερμοντελοποιηθεί καθώς η επίδοση του στα δεδομένα ελέγχου είναι σχετικά χαμηλή. Επίσης, παρουσιάζει μεγαλύτερο σφάλμα γενίκευσης από τον ταξινομητή Δένδρου Απόφασης.

3. Μοντέλο Ταξινόμησης Μηχανών Διανυσμάτων Υποστήριξης

Το μοντέλο αυτό επίσης αποδίδει καλύτερα όταν τα δεδομένα εισόδου είναι κανονικοποιημένα. Για την κανονικοποίηση των δεδομένων αξιοποίησαμε τη μέθοδο Standard Scaling καθώς οδήγησε το μοντέλο σε μεγαλύτερες επιδόσεις σε σχέση με τις υπόλοιπες μεθόδους. Επίσης, σημειώνεται ότι αξιοποιήθηκε η στρατηγική OvR (ή αλλιώς OvA) για τη κατασκευή του μοντέλου, καθώς οι Μηχανές Διανυσμάτων Υποστήριξης αποτελούν μοντέλα δυαδικής ταξινόμησης και το πρόβλημα ανήκει στη κατηγορία της ταξινόμησης πολλών κλάσεων.

Το τελικό βελτιστοποιημένο μοντέλο ταξινόμησης OvR - SVM είναι το εξής :

```
print("========================      OvA SVM CLASSIFIER Hyperparameter Optimized   ======================", end="\n\n")

clf_svm = OneVsRestClassifier(SVC(C=650, kernel="rbf", gamma=2.2, probability=False, shrinking=False,
                                  break_ties=False, class_weight=None))
```

Στη συνέχεια παρουσιάζουμε τα αποτελέσματα των μετρικών αξιολόγησης.

Πίνακας 5.3 : Αποτελέσματα μετρικών αξιολόγησης Instance Based OvR-SVM Classifier.

	Instance Based OvR - SVM Classifier					
Evaluation Metrics	$TestSet_1$	$TestSet_2$	$TestSet_3$	$TestSet_4$	$TestSet_5$	*Validation_Set*
Precision	0.587	0.624	0.65	0.68	0.655	0.737
Recall	0.582	0.612	0.636	0.677	0.651	0.733
Accuracy	0.611	0.641	0.666	0.681	0.648	0.732
Balanced Accuracy	0.608	0.639	0.664	0.677	0.651	0.733
F1_Score	0.573	0.608	0.635	0.67	0.645	0.731
Jaccard Score	0.44	0.472	0.499	0.516	0.479	0.577

Σύμφωνα με τα αποτελέσματα του πίνακα 5.3 καταλήγουμε στο συμπέρασμα ότι και το Time Instance Based OvR – SVM μοντέλο ταξινόμησης έχει υπερμοντελοποιηθεί. Παρ' όλα αυτά, παρουσιάζει σε γενικές γραμμές λίγο καλύτερη επίδοση από τα δύο προηγούμενα μοντέλα.

4. Μοντέλο Ταξινόμησης Τυχαίου Δάσους

Όπως ήδη έχουμε αναφέρει στο δεύτερο κεφάλαιο, για τη κατασκευή του εν λόγω μοντέλου ταξινόμησης δεν απαιτείται η κανονικοποίηση των δεδομένων.

Το τελικό βελτιστοποιημένο μοντέλο ταξινόμησης Τυχαίου Δάσους είναι το εξής :

```
print("=========================      RF CLASSIFIER Hyperparameter Optimized   ======================", end="\n\n")

clf_rf = RandomForestClassifier(n_estimators=350, criterion= "gini", max_depth=12, min_samples_split=2,
                                min_samples_leaf=1, max_features=3, max_leaf_nodes=60, max_samples=192,
                                min_impurity_decrease=0.0, bootstrap=True, class_weight="balanced",
                                 min_weight_fraction_leaf=0.0, ccp_alpha=0.0, random_state=33)
```

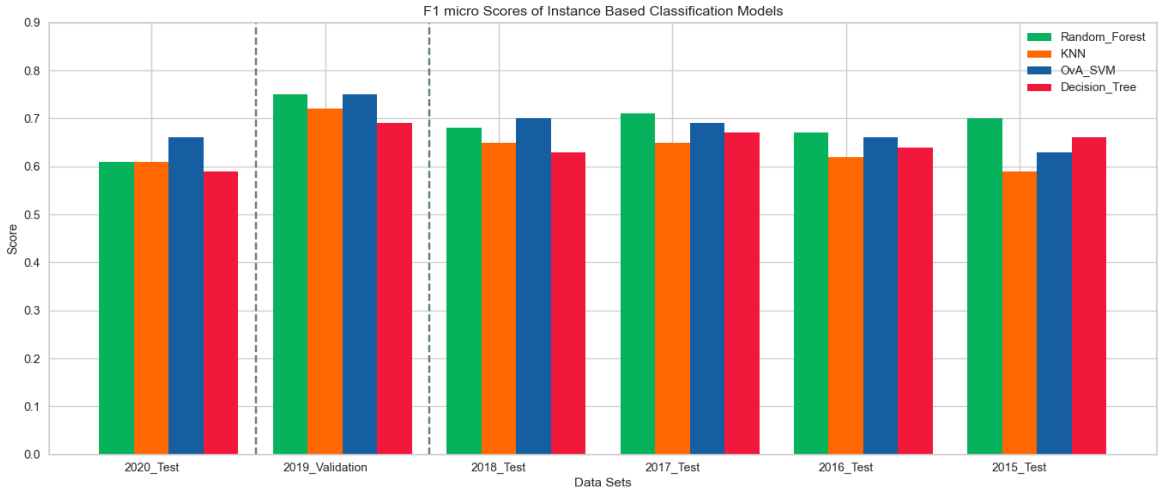
Στη συνέχεια παρουσιάζουμε τα αποτελέσματα των μετρικών αξιολόγησης.

Πίνακας 5.4 : Αποτελέσματα μετρικών αξιολόγησης Instance Based RF Classifier.

	Instance Based Random Forest Classifier					
Evaluation Metrics	$TestSet_1$	$TestSet_2$	$TestSet_3$	$TestSet_4$	$TestSet_5$	*Validation_Set*
Precision	0.699	0.66	0.692	0.675	0.623	0.745
Recall	0.676	0.646	0.686	0.682	0.624	0.755
Accuracy	0.706	0.673	0.715	0.684	0.619	0.753
Balanced Accuracy	0.706	0.675	0.717	0.682	0.623	0.754
F1_Score	0.678	0.643	0.681	0.671	0.614	0.744
Jaccard Score	0.545	0.507	0.556	0.52	0.448	0.60

Σύμφωνα με τα αποτελέσματα του πίνακα 5.4 συμπεραίνουμε ότι το μοντέλο αυτό παρουσιάζει σε γενικές γραμμές μικρότερο σφάλμα γενίκευσης έναντι των υπολοίπων μοντέλων ταξινόμησης που κατασκευάσαμε. Παρ'όλα αυτά η επίδοση του ως προς τα δεδομένα ελέγχου είναι σχετικά μέτρια.

Στη συνέχεια παρουσιάζουμε σε γραφήμα τη συγκριτική αξιολόγηση των μοντέλων.



Σχήμα 5.5 : Γράφημα συγκριτικής αξιολόγησης Instance Based μοντέλων ταξινόμησης.

Σύμφωνα με το σχήμα 5.5, όσον αφορά το μοντέλο K-NN παρατηρούμε ότι όσο τα δεδομένα ελέγχου απομακρύνονται χρονικά από τα δεδομένα εκπαίδευσης, τόσο αυξάνεται το σφάλμα γενίκευσης. Το γεγονός αυτό είναι λογικό καθώς το εν λόγω μοντέλο δεν παραμετροποιεί τα δεδομένα και συνεπώς βασίζεται περισσότερο στο σχήμα των καμπυλών που λογικά διαφοροποιείται με τα χρόνια. Επίσης το μοντέλο Τυχαίου Δάσους υπερτερεί σε γενικές γραμμές έναντι των υπολοίπων. Αξίζει επίσης να αναφέρουμε την σχετική ευρωστία του μοντέλου ταξινόμησης OvR – SVM ως προς τα δεδομένα ελέγχου του "2020" που ορίζονται στο χρονικό παράθυρο "01/03/2020" εώς "28/02/2021". Δεδομένου ότι κατά την εν λόγω περίοδο η πανδημία του "Covid" επηρέασε σημαντικά τη παγκόσμια ζήτηση και κατανάλωση ηλεκτρικής ενέργειας, ήταν αναμενόμενο το αυξημένο σφάλμα γενίκευσης των μοντέλων όσον αφορά το συγκεκριμένο σύνολο ελέγχου.

Όλα τα παραπάνω αποτελέσματα και προβλήματα που αναδύθηκαν ήταν σε γενικές γραμμές αναμενόμενα, εφόσον λάβουμε υπόψη το ιδιαίτερα μικρό συνόλο εκπαίδευσης και τη σχετικά μεγάλη διάσταση των δεδομένων. Συνεπώς, για την αντιμετώπιση των εν λόγω προβλημάτων απαιτείται να μειώσουμε τις διαστάσεις των δεδομένων και να αυξήσουμε το σύνολο δεδομένων εκπαίδευσης. Καθώς με τις εφαρμογές ταξινόμησης στοχεύουμε στην αξιολόγηση της ποιότητας των αποτελεσμάτων των εφαμοργών ομαδοποίησης και εξαγωγής "Προφίλ" Φορτίου του τετάρτου κεφαλαίου, το μέγεθος του συνόλου δεδομένων εκπαίδευσης θα παραμείνει ως έχει. Κατά συνέπεια, θα περιοριστούμε στη λύση της μείωσης των διαστάσεων των δεδομένων. Μια δημοφιλής μέθοδος αντιμετώπισης του προβλήματος των μεγάλων διαστάσεων που παρουσιάζεται σε δεδομένα χρονοσειρών είναι η εξαγωγή και η επιλογή κατάλληλων χαρακτηριστικών αξιοποιώντας τεχνικές και αλγορίθμους που εμπίπτουν στον κλάδο της μηχανικής χαρακτηριστικών.

5.3 Εφαρμογές Ταξινόμησης στο Πεδίο Χαρακτηριστικών (Feature Based Classification Applications)

Για την υλοποίηση των εφαρμογών αξιοποίηθηκαν οι παρακάτω βιβλιοθήκες :

- **tslearn** : Εξαγωγή και επιλογή χαρακτηριστικών από χρονοσειρές.
- **scikit-learn** : Μέτρα αξιολόγησης, αλγόριθμοι ταξινόμησης και αλγόριθμοι επιλογής χαρακτηριστικών.
- **yellowbrick** : Μετρικές και εργαλεία αξιολόγησης.

Στόχος των εν λόγω εφαρμογών είναι η εξαγωγή κατάλληλων χαρακτηριστικών που δύνανται να διαχωρίσουν τα δεδομένα ως προς τις κλάσεις τους, καθώς και η αξιολόγηση της ποιότητας τους μέσω της επίδοσης των μοντέλων ταξινόμησης. Η επίδοση των μοντέλων θα αποτελέσει επίσης μια ένδειξη για το πόσο αποδοτικά δύναται να μοντελοποιηθεί η ενεργειακή συμπεριφορά των ευρωπαϊκών χωρών του συνόλου ανάλυσης, δεδομένου ότι η εκπαίδευση των μοντέλων θα πραγματοποιηθεί με βάση τα "Προφίλ" Φορτίου των εν λόγω χωρών.

5.3.1 Εξαγωγή και Επιλογή Χαρακτηριστικών για Εφαρμογές Ταξινόμησης

Για την εξαγωγή και επιλογή χαρακτηριστικών αξιοποιήθηκαν τα δεδομένα του συνόλου εκπαίδευσης, δηλαδή τα "Προφίλ" Φορτίου της ετήσιας ανάλυσης. Αρχικά εξήχθησαν όλα τα δυνατά χαρακτηριστικά από τα δεδομένα εκπαίδευσης και στη συνέχεια εφαρμόστηκαν στατιστικοί ελέγχοι και αλγόριθμοι επιλογής χαρακτηριστικών. Σημειώνεται ότι για την εξαγωγή των χαρακτηριστικών μέσω της βιβλιοθήκης tsfresh απαιτείται πρωτίστως η κατάλληλη προετοιμασία των δεδομένων που θα αποτελέσουν το σύνολο εισόδου στις συναρτήσεις της εν λόγω βιβλιοθήκης. Αναλυτική περιγραφή της πρότυπης μορφοποίησης των πλαισίων δεδομένων την οποία απαιτεί η βιβλιοθήκη tsfresh μπορεί να βρεθεί εδώ [60].

Αξίζει να αναφερθεί ότι η εξαγωγή των χαρακτηριστικών πραγματοποιήθηκε κάτω από δύο διαφορετικές συνθήκες. Σε πρώτο πλαίσιο εξήχθησαν όλα τα χαρακτηριστικά από τα μη κανονικοποημένα δεδομένα και σε δεύτερο πλαίσιο εξήχθησαν όλα τα χαρακτηριστικά από τα κανονικοποιημένα δεδομένα. Κατά το στάδιο της αξιολόγησης οδηγηθήκαμε στο συμπέρασμα ότι η κανονικοποίηση των χρονοσειρών προκαλεί τη μείωση της στατιστικής σημασίας των εξαχθέντων χαρακτηριστικών. Το γεγονός αυτό είναι λογικό καθώς με τη κανονικοποίηση των αρχικών δεδομένων (χρονοσειρές) χάνεται ένα σημαντικό μέρος της πληροφορίας που περιέχεται στα αρχικά διανύσματα εισόδου και κατά συνέπεια περιορίζεται η διακύμανση των τελικών χαρακτηριστικών διανυσμάτων. Επίσης η επίδοση των μοντέλων ταξινόμησης στο πεδίο των χαρακτηριστικών τα οποία εξήχθησαν από τα κανονικοποιημένα δεδομένα ήταν πολύ χειρότερη από την επίδοση των μοντέλων ταξινόμησης στο πεδίο του χρόνου. Συνεπώς, στη συνέχεια παρουσιάζουμε μόνο τα αποτελέσματα της εξαγωγής χαρακτηριστικών από τα μη κανονικοποιημένα δεδομένα καθώς επίσης και τα αποτελέσματα της αξιολόγησης αυτών των χαρακτηριστικών.

- **<u>Διαδικασία Εξαγωγής και Επιλογής Χαρακτηριστικών</u>**

  Σε πρώτο στάδιο χρησιμοποιήσαμε τη συνάρτηση $extract\_features()$ της βιβλιοθήκης tslearn η οποία δέχτηκε ως είσοδο το κατάλληλο μορφοποιημένο πλαίσιο με τις χρονοσειρές του συνόλου εκπαίδευσης ("Προφίλ" Φορτίου) και ως αποτέλεσμα επιστράφηκαν 787 χαρακτηριστικά διανύσματα.

```
Feature Extraction: 100%|██████████| 10/10 [00:21<00:00,  2.20s/it]
====================        FEATURE EXTRACTION (TSFRESH)      ============================
The 787 extracted features dataframe is :
      MW__variance_larger_than_standard_deviation  ...  MW__matrix_profile__feature_"75"__threshold_0.98
10_FR                                          1.0  ...                                               NaN
10_MK                                          1.0  ...                                               NaN
10_NO                                          1.0  ...                                               NaN
10_RO                                          1.0  ...                                               NaN
10_UA                                          1.0  ...                                               NaN
...                                            ...  ...                                               ...
9_MK                                           1.0  ...                                               NaN
9_NO                                           1.0  ...                                          2.993138
9_RO                                           1.0  ...                                               NaN
9_SE                                           1.0  ...                                          3.831748
9_UA                                           1.0  ...                                               NaN

[192 rows x 787 columns]
```

Σχήμα 5.6 : Πλαίσιο δεδομένων που επιστράφηκε από τη συνάρτηση extract_features().

Στο σημείο αυτό σημειώνεται ότι κάποια από αυτά τα 787 χαρακτηριστικά διανύσματα περιέχουν πεδία συμπληρωμένα με τιμές "NaN". Μετά από την απαλοιφή των εν λόγω διανυσμάτων καταλήγουμε να έχουμε στη διάθεση μας 414 χαρακτηριστικά διανύσματα. Καθώς ο αριθμός των χαρακτηριστικών είναι ιδιαίτερα

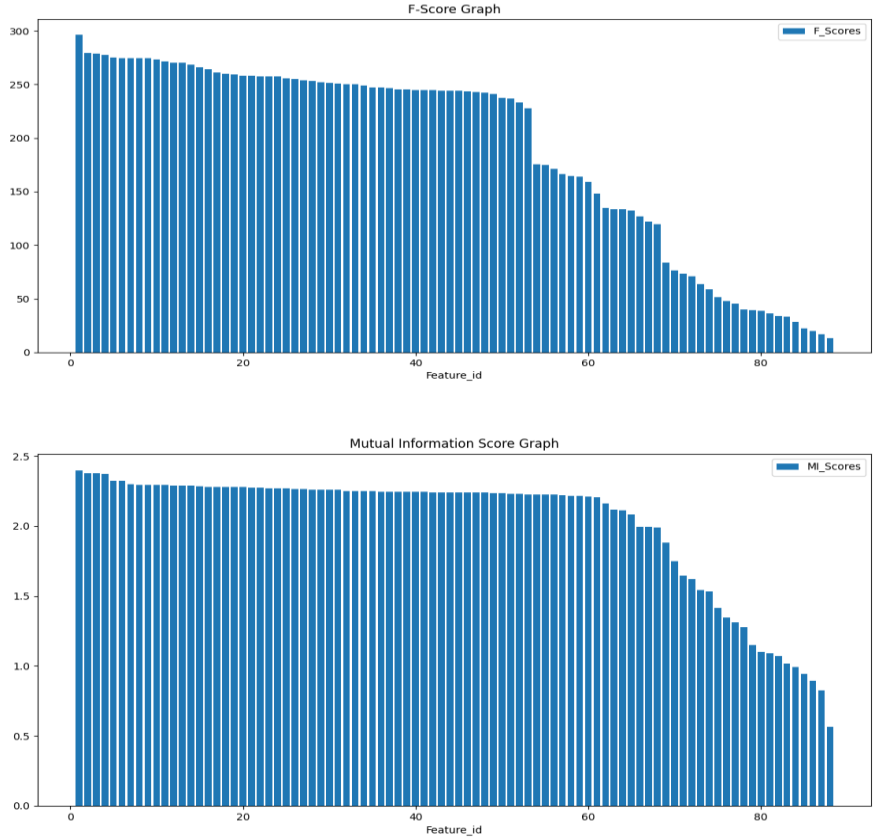
μεγάλος χρησιμοποιήσαμε τη συνάρτηση $select\_features()$ της βιβλιοθήκης tslearn η οποία αποτελεί μέθοδο φίλτρου επιλογής χαρακτηριστικών. Η εν λόγω συνάρτηση δέχεται τα εξής ορίσματα :

- Το πλαίσιο δεδομένων με τα υποψήφια χαρακτηριστικά διανύσματα.
- Το διάνυσμα στόχος Y των δεδομένων.
- Ρύθμιση κατωφλίου για το εκτιμώμενο Ποσοστό Ψευδών Ανακαλύψεων (FDR).
- Τη μέθοδο ελέγχου υποθέσεων.
- Το είδος της εφαρμογής (Ταξινόμηση, Παλινδρόμηση).

Επιλέξαμε ως μέθοδο ελέγχου υποθέσεων τον έλεγχο Kendal Rank Test και θέσαμε το κατώφλι FDR ως $1e-06$ ώστε να απαλειφθούν όσα το δυνατόν περισσότερα χαρακτηριστικά. Επίσης ορίσαμε τον τύπο της εφαρμογής ως ταξινόμηση. Μετά την εκτέλεση της παραπάνω συνάρτησης με τα αντίστοιχα κατάλληλα ορίσματα καταλήγουμε να έχουμε στη διάθεση μας 88 χαρακτηριστικά διανύσματα.

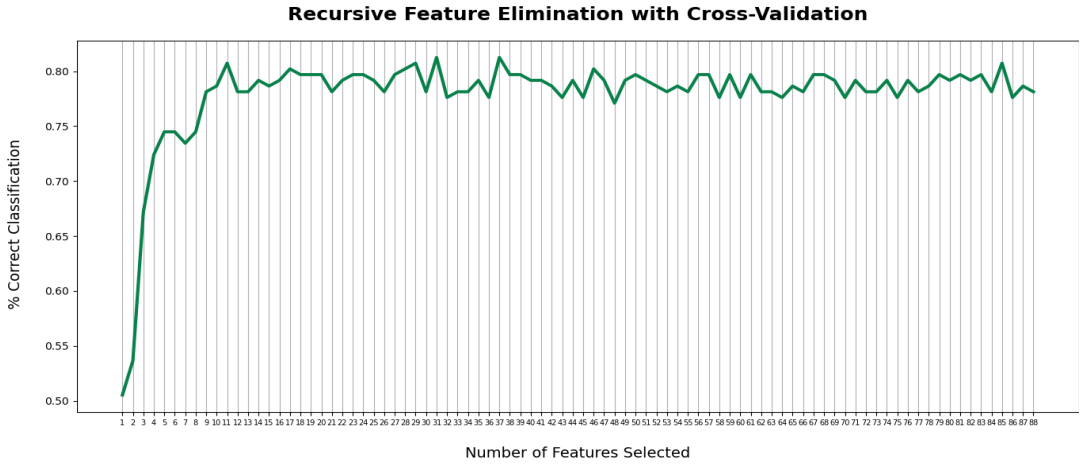
Στη συνέχεια υπολογίσαμε τον πίνακα συσχέτισης Pearson για να επιλέξουμε όλα τα χαρακτηριστικά διανύσματα τα οποία δεν συσχετίζονται γραμμικά. Προς απογοήτευση μας, ο πίνακας συσχέτισης υπέδειξε ότι υπάρχει ισχυρή γραμμική συσχέτιση ανάμεσα σε όλα τα χαρακτηριστικά διανύσματα (88 συνολικά).

Στη συνέχεια υπολογίσαμε τους δείκτες Anova F και Αμοιβαίας Πληροφορίας (MI) των χαρακτηριστικών διανυσμάτων ως προς το στόχο, ώστε να έχουμε μια γενική εικόνα της στατιστικής τους σημασίας.



Σχήμα 5.7 : Αποτελέσματα δεικτών Anova F και MI των υποψήφιων χαρακτηριστικών.

Σε επόμενο βήμα χρησιμοποιήσαμε τη μέθοδο της Βηματικής Οπίσθιας Εξάλειψης (ή αλλιώς Recursive Feature Elimination - RFE) η οποία ανήκει στις μεθόδους περιτυλίγματος και αξιοποιήσαμε την επίδοση ενός ταξινομητή Τυχαίου Δάσους (RF) για την αξιολόγηση των χαρακτηριστικών. Καθώς το σύνολο εκπαίδευσης είναι ιδιαίτερα μικρό (192 "Προφίλ" Φορτίου) και παρουσιάζει ανισορροπία μεταξύ των κλάσεων, χρησιμοποιήσαμε τη τεχνική επαναληπτικής δειγματοληψίας Stratified K-Folds Cross Validation. Τα αποτελέσματα της μεθόδου παρατίθενται στη συνέχεια.



Σχήμα 5.8 : Αποτελέσματα Μεθόδου Random Forest - RFE επιλογής χαρακτηριστικών.

Αξίζει να σημειωθεί ότι το χρονικό κόστος της παραπάνω μεθόδου ήταν ιδιαίτερα υψηλό. Το σχήμα 5.8 υποδεικνύει την ύπαρξη ενός υποσυνόλου χαρακτηριστικών που αποτελείται από δέκα συγκεκριμένα χαρακτηριστικά τα οποία δύνανται να διαχωρίσουν "βέλτιστα" τις κλάσεις του προβλήματος όταν επεξεργάζονται από μοντέλα ταξινόμησης Τυχαίου Δάσους. Συνεπώς, επιλέξαμε αρχικά τα δέκα αυτά χαρακτηριστικά που υπέδειξε η μέθοδος Random Forest - RFE. Σημειώνεται επίσης ότι με βάση το σχήμα 5.8, η περαιτέρω αύξηση του συνόλου των επιλεγμένων χαρακτηριστικών δεν προκαλεί σημαντική μείωση της επίδοσης του ταξινομητή RF.

Καθώς επιδιώκουμε να εκπαιδεύσουμε και να αξιολογήσουμε μοντέλα διαφόρων τύπων, επιλέξαμε επίσης τα πρώτα τέσσερα χαρακτηριστικά τα οποία δεν εμπεριέχονται στο σύνολο που υπέδειξε η μέθοδος RF-RFE αλλά υποδεικνύονται ως "βέλτιστα" με βάση τον στατιστικό έλεγχο Mutual Information (MI). Καθώς το σύνολο υποψήφιων χαρακτηριστικών διέπεται από σημαντικό βαθμό ασσυμετρίας αποφύγαμε την επιλογή χαρακτηριστικών μέσω της μεθόδου Anova. Επίσης, υπολογίσαμε τρία χαρακτηριστικά τα οποία αναφέρονται στη βιβλιογραφία ως "Load Shape Features" ή αλλιώς "Load Shape Indices" και προτείνονται για εφαρμογές ταξινόμησης χρονοσειρών ηλεκτρικού φορτίου [4, 11].

Τα χαρακτηριστικά "Load Shape Features" ορίζονται ως εξής [11] :

i. $d_1 = Load\ Factor = \frac{P_{avg,day}}{P_{max,day}}$

ii. $d_2 = Night\ Impact = \frac{1}{3}\frac{P_{avg,night}}{P_{avg,day}}$ $(8\ hours\ night\ 23{:}00 - 07{:}00)$

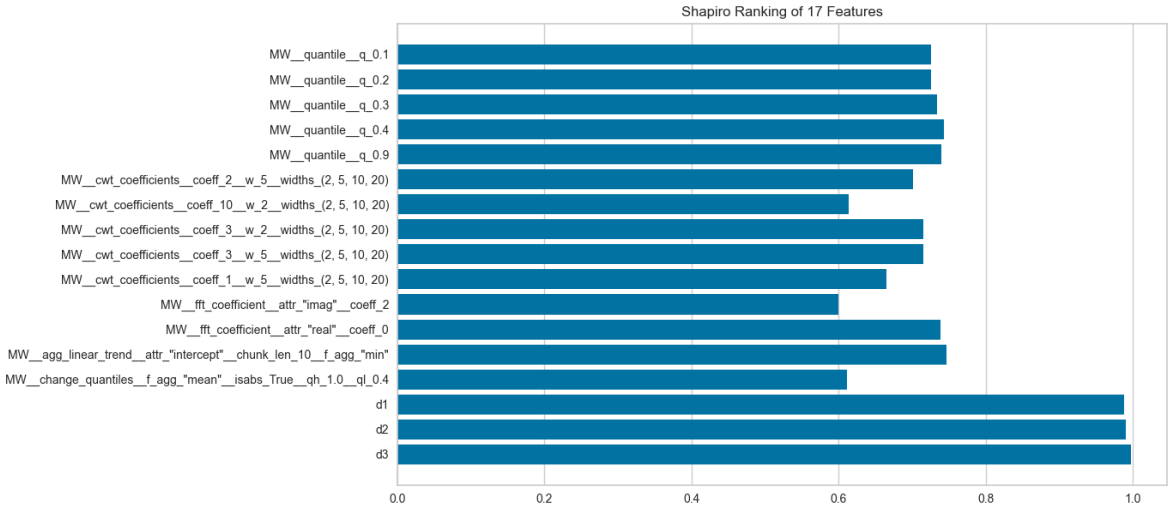
iii. $d_3 = Lunch\ Impact = \frac{1}{8}\frac{P_{avg,lunch}}{P_{avg,day}}$ $(3\ hours\ lunch\ 12{:}00 - 15{:}00)$

5.3.2 Τελικά Χαρακτηριστικά

Τα τελικά επιλεγμένα χαρατκηριστικά είναι τα εξής :

1. quantile__q_0.1
2. quantile__q_0.2
3. quantile__q_0.3
4. quantile__q_0.4
5. quantile__q_0.9
6. cwt_coefficients__coeff_2__w_5__widths_(2, 5, 10, 20)
7. cwt_coefficients__coeff_10__w_2__widths_(2, 5, 10, 20)
8. cwt_coefficients__coeff_3__w_2__widths_(2, 5, 10, 20)
9. cwt_coefficients__coeff_3__w_5__widths_(2, 5, 10, 20)
10. cwt_coefficients__coeff_1__w_5__witdths_(2, 5, 10, 20)
11. fft_coefficient__attr_"imag"__coeff_2
12. fft_coefficient__attr_"real"__coeff_0
13. agg_linear_trend_attr_"intercept"_chunk_len_10_f_agg_"min"
14. change_quantiles__f_agg_"mean"__isabs_True__gh_1.0__gl_0.4
15. d1 (Load Factor)
16. d2 (Night Imapct)
17. d3 (Lunch Impact)

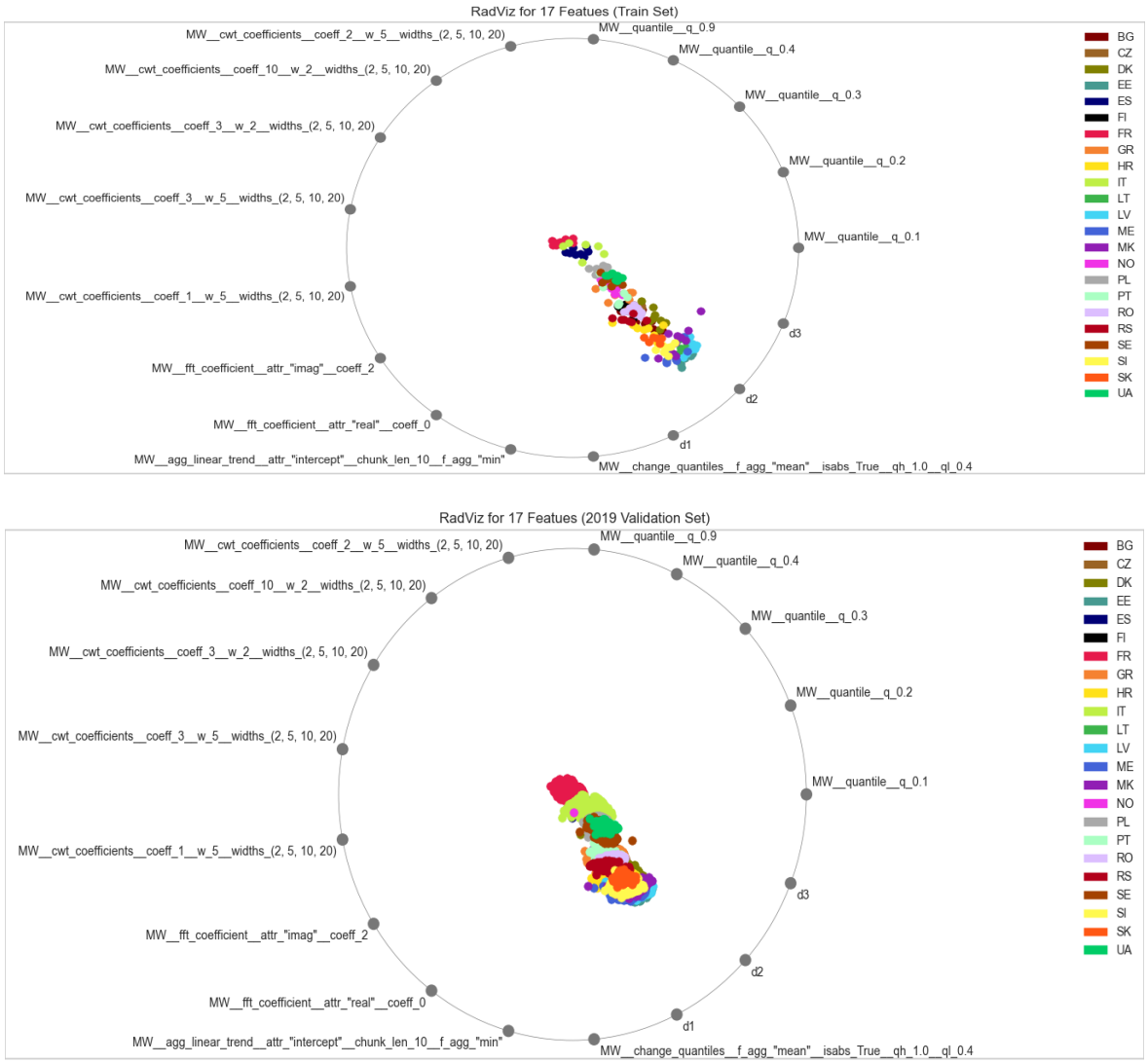
Αναλυτική περιγραφή των χαρακτηριστικών 1 εως 14 μπορεί να βρεθεί εδώ [60]. Αξίζει να σημειωθεί ότι τα πρώτα δέκα τέσσερα χαρακτηριστικά, τα οποία εξήχθησαν μέσω της βιβλιοθήκης tsfresh, ακολουθούν ασύμμετρες κατανομές.



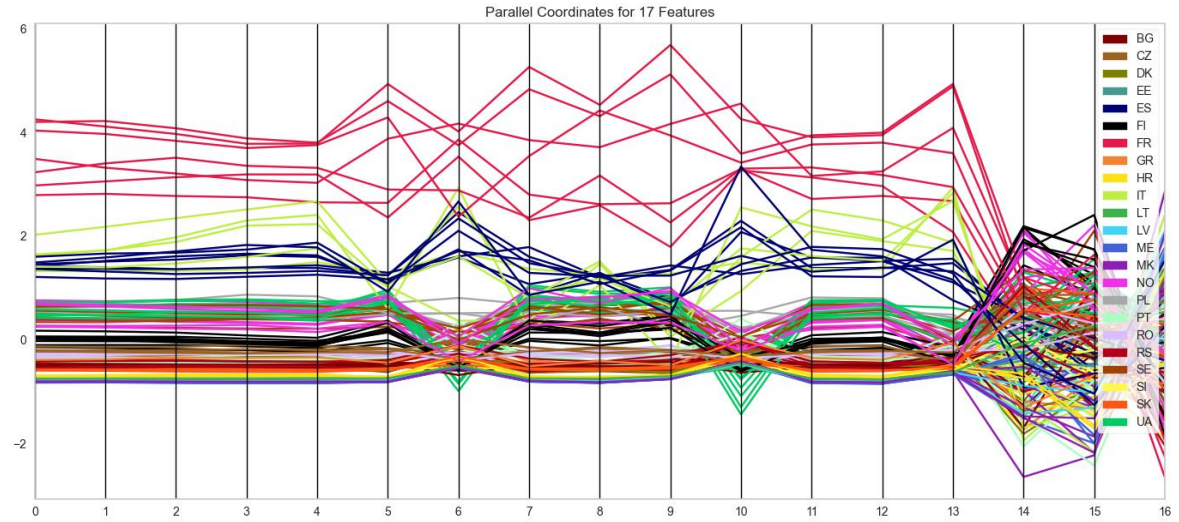
Σχήμα 5.9 : Αποτελέσματα ελέγχου κανονικότητας Shapiro Ranking τελικών χαρακτηριστικών.

Επίσης, μετά από μια πρόχειρη αξιολόγηση των μοντέλων ταξινόμησης οδηγηθήκαμε στο συμπέρασμα ότι όλα τα μοντέλα εκτός του K-NN παρουσιάζουν μεγαλύτερες επιδόσεις όταν εκπαιδεύονται σε δεδομένα που μοντελοποιούνται με όλα τα παραπάνω χαρακτηριστικά. Το μοντέλο K-NN όμως παρουσίασε μεγαλύτερες επιδόσεις όταν εκπαιδεύτηκε σε δεδομένα που μοντελοποιούνται με τα πρώτα δέκα τέσσερα χαρακτηριστικά της βιβλιοθήλης tsfresh. Συνεπώς, όσον αφορά την εκπαίδευση και την αξιολόγηση του ταξινομητή K-NN αξιοποιήσαμε μόνο τα χαρακτηριστικά 1 εως 14.

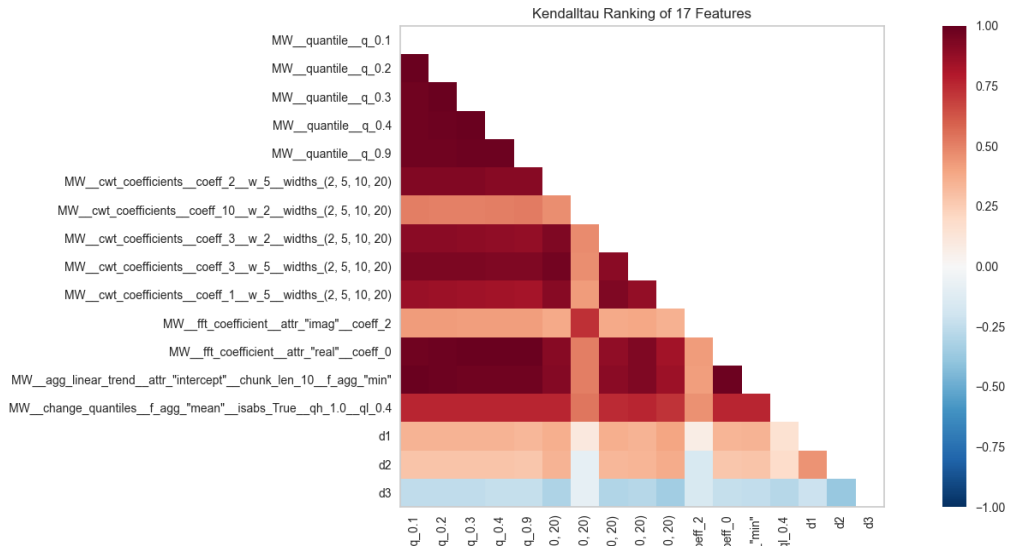
Στη συνέχεια παρουσιάζονται γραφήματα οπτικοποίησης των τελικών χαρακτηριστικών.



Σχήμα 5.10 : Γραφήματα μεθόδου οπτικοποίησης RadViz των τελικών χαρακτηριστικών.



Σχήμα 5.11 : Γραφημα μεθόδου οπτικοποίησης Παράλληλων Συντεταγμένων των τελικών χαρακτηριστικών.



Σχήμα 5.12 : Πίνακας συσχέτισης Kendall των τελικών χαρακτηριστικών διανυσμάτων.

Με βάση το σχήμα 5.10 παρατηρούμε την ευεργετική συμβολή των χαρακτηριστικών "Load Shape Features" καθώς υποδεικνύεται ότι οι διαφορετικές κλάσεις διαχωρίζονται σε αρκετά μεγάλο βαθμό με βάση τους εν λόγω δείκτες. Στο σημείο αυτό αξίζει να αναφέρουμε ότι εκπαιδεύσαμε μοντέλα ταξινόμησης μόνο με αυτά τα τρία χαρακτηριστικά (d1, d2, d3) και τα αποτελέσματα της αξιολόγησης τους ήταν μη ικανοποιητικά. Παρ' όλα αυτά, ο συνδιασμός των "Load Shape Features" με τα χαρακτηριστικά τα οποία εξήχθησαν μέσω της βιβλιοθήκης tsfresh είχε ως αποτέλεσμα τη σημαντική βελτίωση της γενίκευσης των μοντέλων. Δεδομένου ότι τα αποτελέσματα της αξιολόγησης των μοντέλων που εκπαιδεύτηκαν με βάση όλα τα τελικά χαρακτηριστικά (εκτός του K-NN που εκπαιδεύτηκε με τα πρώτα δέκα τέσσερα) ήταν ιδιαίτερα ικανοποιητικά, οδηγηθήκαμε στο συμπέρασμα ότι οι εφαρμογές μηχανικής χαρακτηριστικών ήταν επιτυχείς. Συνεπώς, τροποποιήσαμε όλα τα σύνολα δεδομένων (εκπαίδευσης, επικύρωσης και ελέγχου) εξάγοντας τα τελικά επιλεγμένα χαρακτηριστικά από τα αντίστοιχα διανύσματα εισόδου τους (ημερήσιες καμπύλες Σ.Π.Φ) και δημιουργήσαμε νέα αντίστοιχα σύνολα δεδομένων στο πεδίο αυτών των τελικών χαρακτηριστικών. Τα εν λόγω σύνολα αξιοποιήθηκαν για την εκπαίδευση, βελτιστοποίηση και αξιολόγηση των μοντέλων ταξινομήσης που παρουσιάζονται στην επόμενη ενότητα.

5.3.3 Αποτελέσματα Μοντέλων Ταξινόμησης στο Πεδίο Χαρακτηριστικών

Σημειώνεται ότι όλα τα μοντέλα που κατασκευάσαμε, εκτός του ταξινομητή K-NN, επεξεργαστήκαν σύνολα δεδομένων τα οποία έχουν ενσωματώσει όλα τα τελικά επιλεγμένα χαρακτηριστικά. Ο ταξινομήτής K-NN όμως επεξεργάστηκε δεδομένα που αναπαρίστανται μόνο από τα πρώτα δέκα τέσσερα χαρακτηριστικά που παρουσιάσαμε στη προηγούμενη ενότητα καθώς η εισαγωγή των "Load Shape Features" στα σύνολα δεδομένων επηρέαζε αρνητικά την επίδοση του.

Στη συνέχεια παρουσιάζουμε τα τελικά Feature Based μοντέλα ταξινόμησης καθώς και τα αντίστοιχα αποτελέσματα των μετρικών αξιολόγησης.

1. **<u>Μοντέλο ταξινόμησης Δένδρου Απόφασης (DT) στο Πεδίο Χαρακτηριστικών</u>**

Σημειώνεται ότι τα δεδομένα εισόδου του εν λόγω μοντέλου δεν κανονικοποιήθηκαν.

Το τελικό βελτιστοποιημένο μοντέλο Feature Based DT Classifier είναι το εξής :

```
print("==========  Feature Based DECISION TREE CLASSIFIER Hyperparameter Optimized ==========", end="\n\n")

clf_dt = DecisionTreeClassifier(criterion="gini", random_state=101, max_depth=10, min_samples_split=2,
                                min_samples_leaf=5, max_features=8, max_leaf_nodes=None)
```

Στη συνέχεια παρουσιάζουμε τον πίνακα με τα αποτελέσματα των μετρικών αξιολόγησης. Σημειώνεται ότι τα εν λόγω μέτρα έχουν υπολογισθεί με βάση τη στρατηγική "macro averaging". Το ίδιο ισχύει και για όλα τα υπόλοιπα μοντέλα ταξινόμησης.

Πίνακας 5.5 : Αποτελέσματα μετρικών αξιολόγησης Feature Based Decision Tree Classifier.

	Feature Based Decision Tree Classifier					
Evaluation Metrics	$TestSet_1$	$TestSet_2$	$TestSet_3$	$TestSet_4$	$TestSet_5$	*Validation_Set*
Precision	0.65	0.617	0.652	0.647	0.635	0.706
Recall	0.628	0.602	0.639	0.641	0.627	0.698
Accuracy	0.656	0.629	0.667	0.648	0.624	0.697
Balanced Accuracy	0.657	0.629	0.668	0.641	0.627	0.698
F1_Score	0.622	0.59	0.631	0.631	0.608	0.689
Jaccard Score	0.488	0.459	0.5	0.479	0.454	0.535

Σύμφωνα με τα αποτελέσματα του πίνακα 5.5 συμπεραίνουμε ότι το εν λόγω μοντέλο έχει υπερμοντελοποιηθεί. Συγκεκριμένα, προς απογοήτευση μας, παρουσιάζει λίγο μεγαλύτερο σφάλμα γενίκευσης σε σχέση με το ταξινομητή Δένδρου Απόφασης στο πεδίο του χρόνου (Instance Based DT Classifier).

2. **<u>Μοντέλο ταξινόμησης K-NN στο Πεδίο Χαρακτηριστικών</u>**

Τα δεδομένα εισόδου κανονικοποιήθηκαν σύμφωνα με τη μέθοδο Standard Scaling.

Το τελικό βελτιστοποιημένο μοντέλο Feature Based K-NN Classifier είναι το εξής :

```
print("=========================      K-NN CLASSIFIER Hyperparameter Optimized   ======================", end="\n\n")
clf = neighbors.KNeighborsClassifier(n_neighbors=1, weights="distance", algorithm="auto", leaf_size=30, p=3)
```

Στη συνέχεια παρουσιάζουμε το πίνακα με τα αποτελέσματα των μετρικών αξιολόγησης.

Πίνακας 5.6 : Αποτελέσματα μετρικών αξιολόγησης Feature Based K-NN Classifier.

	Feature Based K-NN Classifier					
Evaluation Metrics	$TestSet_1$	$TestSet_2$	$TestSet_3$	$TestSet_4$	$TestSet_5$	*Validation_Set*
Precision	0.618	0.592	0.642	0.793	0.789	0.87
Recall	0.614	0.59	0.639	0.789	0.787	0.869
Accuracy	0.615	0.59	0.638	0.787	0.786	0.867
Balanced Accuracy	0.61	0.583	0.633	0.785	0.781	0.866
F1_Score	0.614	0.589	0.638	0.787	0.785	0.867
Jaccard Score	0.443	0.418	0.468	0.649	0.646	0.765

Σύμφωνα με τα αποτελέσματα του πίνακα 5.6 παρατηρούμε ότι σε γενικές γραμμές το Feature Based K-NN μοντέλο ταξινόμησης υπερτερεί έναντι του Instance Based K-NN μοντέλου.

3. <u>Μοντέλο ταξινόμησης OvR - SVM στο Πεδίο Χαρακτηριστικών</u>

Τα δεδομένα εισόδου κανονικοποιήθηκαν σύμφωνα με τη μέθοδο Standard Scaling.

Το τελικό βελτιστοποιημένο μοντέλο ταξινόμησης OvR – SVM είναι το εξής :

```
print("=========================      SVM CLASSIFIER MODEL    ======================", end="\n\n")
clf_svm = OneVsRestClassifier(SVC(C=420, kernel="rbf", gamma=0.15, probability=False, shrinking=False,
                                  break_ties=False, class_weight=None))
```

Στη συνέχεια παρουσιάζουμε το πίνακα με τα αποτελέσματα των μετρικών αξιολόγησης.

Πίνακας 5.7 : Αποτελέσματα μετρικών αξιολόγησης Feature Based OvR - SVM Classifier.

	Feature Based OvR - SVM Classifier					
Evaluation Metrics	$TestSet_1$	$TestSet_2$	$TestSet_3$	$TestSet_4$	$TestSet_5$	*Validation_Set*
Precision	0.697	0.69	0.72	0.767	0.769	0.812
Recall	0.645	0.635	0.679	0.765	0.766	0.81
Accuracy	0.674	0.664	0.71	0.767	0.767	0.81
Balanced Accuracy	0.673	0.664	0.71	0.765	0.766	0.809
F1_Score	0.633	0.621	0.67	0.762	0.762	0.805
Jaccard Score	0.508	0.497	0.55	0.622	0.622	0.68

Σύμφωνα με τα αποτελέσματα του πίνακα 5.7 συμπεραίνουμε ότι το Feature Based OvR SVM μοντέλο ταξινόμησης υπερτερεί έναντι του Instance Based OvR - SVM μοντέλου. Επίσης, σε γενικές γραμμές παρουσιάζει μικρότερο σφάλμα γενίκευσης από ότι το Feature Based K-NN μοντέλο ταξινόμησης.

4. Μοντέλο ταξινόμησης Τυχαίου Δάσους (RF) στο Πεδίο Χαρακτηριστικών

Το τελικό βελτιστοποιημένο μοντέλο ταξινόμησης Τυχαίου Δάσους είναι το εξής :

```
clf_rf = RandomForestClassifier(n_estimators=400, criterion="gini", random_state=54, bootstrap=True,
                                max_features=1, min_samples_leaf=1, max_samples=145, max_depth=12,
                                min_samples_split=2, max_leaf_nodes=None, min_impurity_decrease=0.0001,
                                oob_score=True)
```

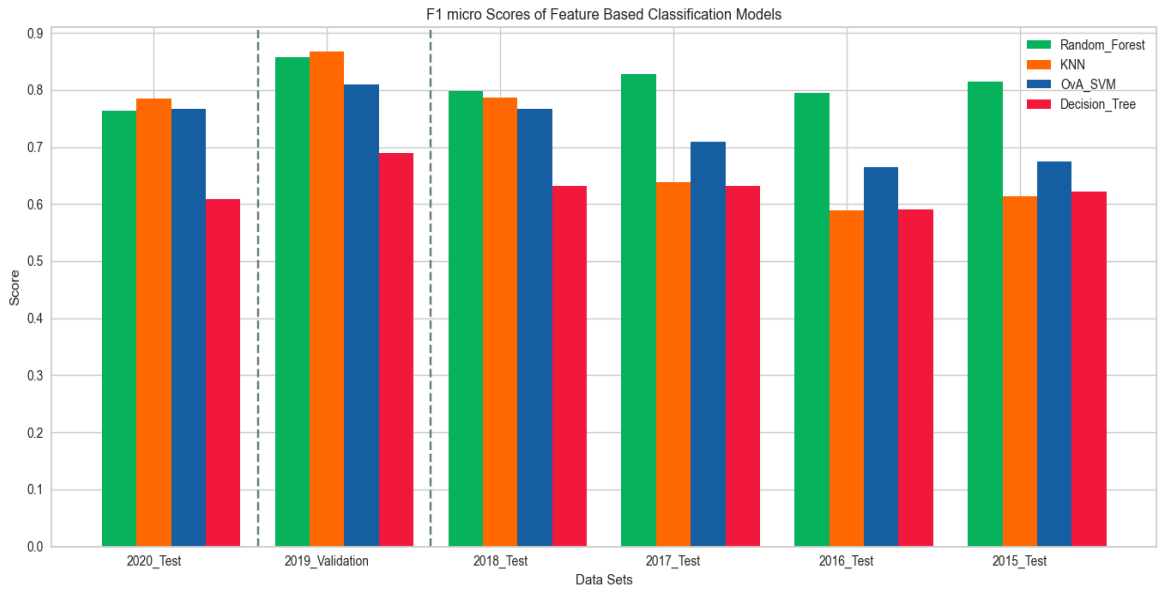
Στη συνέχεια παρουσιάζουμε το πίνακα με τα αποτελέσματα των μετρικών αξιολόγησης.

Πίνακας 5.8 : Αποτελέσματα μετρικών αξιολόγησης Feature Based RF Classifier.

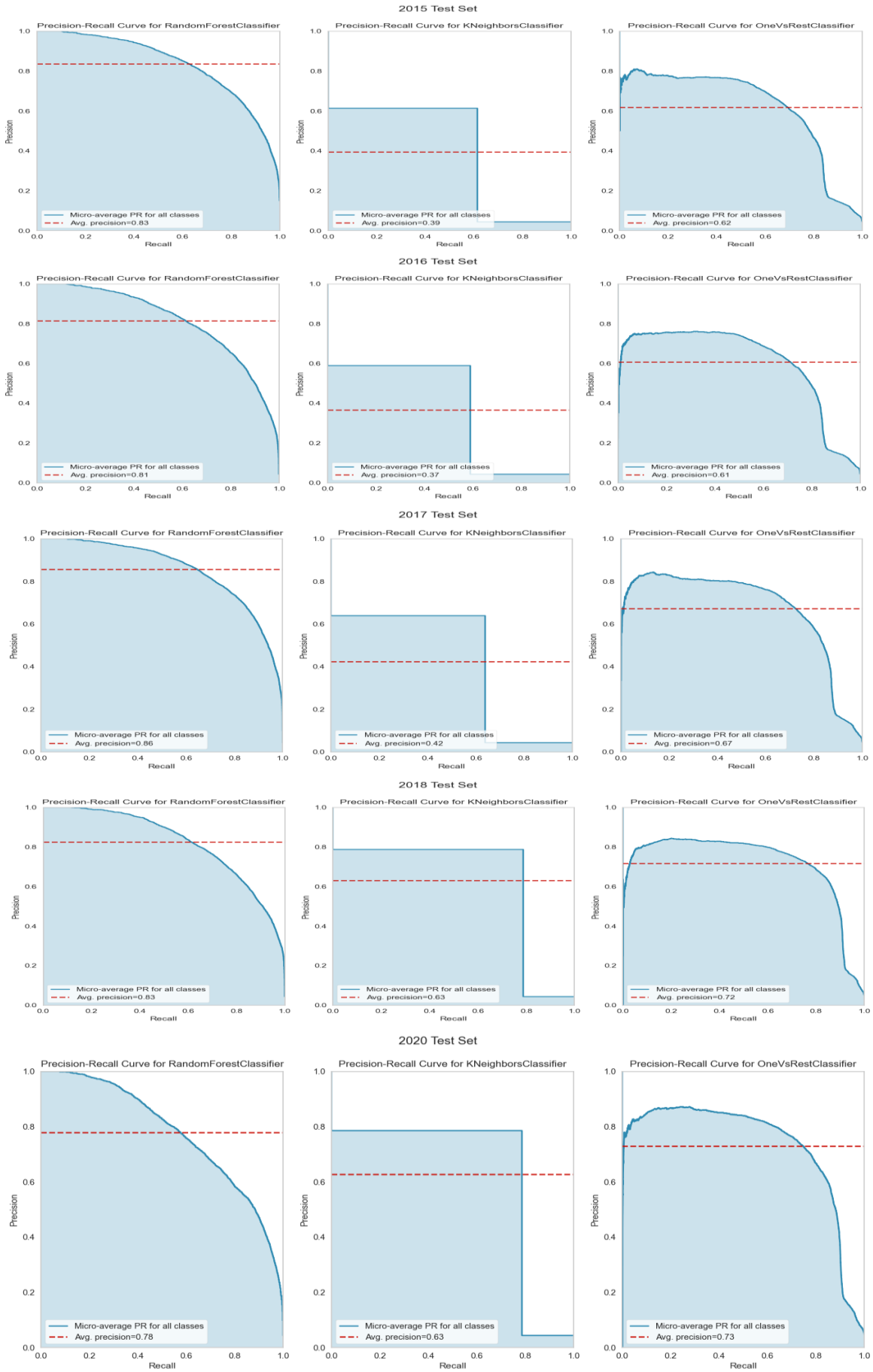
	Feature Based Random Forest Classifier					
Evaluation Metrics	$TestSet_1$	$TestSet_2$	$TestSet_3$	$TestSet_4$	$TestSet_5$	*Validation_Set*
Precision	0.818	0.785	0.824	0.793	0.764	0.859
Recall	0.805	0.78	0.813	0.784	0.76	0.855
Accuracy	0.814	0.795	0.828	0.798	0.764	0.858
Balanced Accuracy	0.813	0.796	0.829	0.793	0.767	0.858
F1_Score	0.81	0.78	0.82	0.79	0.762	0.858
Jaccard Score	0.687	0.66	0.70	0.66	0.618	0.751

Σύμφωνα με τα αποτελέσματα του πίνακα 5.8 επιβεβαιώνεται η υπεροχή του Feature Based Random Forest μοντέλου ταξινόμησης έναντι των υπολοίπων μοντέλων. Συγκεκριμένα, παρουσιάζει το μικρότερο σφάλμα γενίκευσης από όλα τα μοντέλα που έχουμε παρουσιάσει μέχρι στιγμής.

Στη συνέχεια παρουσιάζουμε σε γραφήματα τη συγκριτική αξιολόγηση των Feature Based μοντέλων ταξινόμησης.



Σχήμα 5.13 : Γράφημα συγκριτικής αξιολόγησης Feature Based μοντέλων ταξινόμησης.



Σχήμα 5.14 : Διαγράμματα Precision - Recall των Feature Based μοντέλων ταξινόμησης.

Κάθε μοντέλο παρουσιάζει διαφορετικά σφάλματα καθώς απεικονίζει με διαφορετικό τρόπο τα δεδομένα εισόδου ως έξοδο. Συνεπώς, μπορούμε πιθανώς να βελτιώσουμε ακόμα περισσότερο το σφάλμα γενίκευσης της πρόβλεψης συναθροίζοντας όλα τα παραπάνω Feature Based μοντέλα για τη κατασκευή ενός Ταξινομητή Ψηφοφορίας.

5. Ταξινομητής Ψηφοφορίας στο Πεδίο Χαρακτηριστικών (Feature Based Voting Classifier)

Μετά από δοκιμές καταλήξαμε στο συμπέρασμα ότι η συνάθροιση των Feature Based μοντέλων ταξινόμησης βελτιώνει όντως το σφάλμα γενίκευσης της πρόβλεψης. Συγκεκριμένα, η συνάθροιση των μοντέλων RF, OvR-SVM και K-NN παρουσίασε τα καλύτερα αποτελέσμα. Η ενσωμάτωση του ταξινομητή Δένδρου Απόφασης στο σύνολο συνάθροισης επηρέασε αρνητικά την επίδοση του Ταξινομητή Ψηφοφορίας και κατά συνέπεια δεν συμπεριλήφθηκε στο τελικό βελτιστοποιημένο μοντέλο συνάθροισης.

Σημειώνεται πως όλα τα μοντέλα του τελικού συνόλου συνάθροισης εκπαιδεύτηκαν, επικυρώθηκαν και ελέγχθηκαν με βάση τα ίδια εν γένει αντίστοιχα σύνολα δεδομένων. Όμως όσον αφορά τα μοντέλα K-NN και OvR-SVM τα εν λόγω σύνολα δεδομένων κανονικοποιήθηκαν με τη μέθοδο Standard Scaling. Επίσης μόνο για το μοντέλο K-NN απαλείψαμε τα χαρακτηριστικά d1, d2 και d3, δηλαδή τις στήλες που αντιστοιχούν σε αυτά τα διανύσματα χαρακτηριστικών. Υπενθυμίζουμε ότι όλες οι ημερήσιες καμπύλες Σ.Π.Φ είναι αποθηκευμένες ως σειρές στα αντίστοιχα πλαίσια δεδομένων και οι στήλες αυτών των πλαισίων αναφέρονται στα χαρακτηριστικά (features) τους. Για να διαχωρίσουμε και να αυτοματοποιήσουμε αποδοτικά τη προαναφερθείσα διαδικασία προεπεξεργασίας των δεδομένων εισόδου και να αποφύγουμε τη διαρροή δεδομένων (data leakage), αξιοποιήσαμε "αγωγούς" μηχανικής μάθησης (machine learning pipelines). Κάθε μοντέλο ταξινόμησης κατασκευάζεται σύμφωνα με τον δικό του "αγωγό". Αρχικά αποθηκεύσαμε σε λίστες με κατάλληλα ονόματα τα αναγνωριστικά (labels) των στηλών – χαρακτηριστικών που επεξεργάζεται κάθε μοντέλο και στη συνέχεια δημιουργήσαμε μια κλάση με όνομα $Column\_Extractor()$ που ορίζεται ως εξής :

```
class ColumnExtractor(TransformerMixin, BaseEstimator):
    def __init__(self, cols):
        self.cols = cols

    def transform(self, X):
        X_prime = X.loc[:, self.cols]
        return X_prime

    def fit(self, X, y=None):
        return self
```

Σε επόμενο βήμα δημιουργήσαμε έναν "αγωγό" (pipeline) με κατάλληλα ορίσματα για κάθε μοντέλο. Τα εν λόγω ορίσματα καθορίζουν το μοντέλο ταξινόμησης, την επιλογή των χαρακτηριστικών του αντίστοιχου μοντέλου (μεσω της παραπάνω κλάσης) και την κανονικοποίηση των δεδομένων. Στη συνέχεια ενσωματώνονται τα βελτιστοποιημένα Feature Based μοντέλα ταξινόμησης στο σύνολο συνάθροισης με βάση τον αντίστοιχο τους "αγωγό".

Το τελικό Feature Based μονέλο του Ταξινομητή Ψηφοφορίας είναι το εξής :

```
voting_clf = VotingClassifier(estimators = [('rf', pipe1), ('knn', pipe2), ('svm', pipe3)],
                              voting='soft')
```

Οι "αγωγοί" μηχανικής μάθησης ορίστηκαν ως εξής :

```
pipe1 = Pipeline([('rf', clf_rf)])
pipe1.fit(X_train, Y_train['target'])

pipe2 = Pipeline([ ('x_prime', ColumnExtractor(cols=knn_features)),
                   ('scale',StandardScaler()), ('knn', clf_knn)])
pipe2.fit(X_train, Y_train['target'])

pipe3 = Pipeline([('scale', StandardScaler()), ('svc', clf_svm)])
pipe3.fit(X_train, Y_train['target'])
```

Στη συνέχεια παρουσιάζουμε το πίνακα με τα αποτελέσματα των μετρικών αξιολόγησης.

Πίνακας 5.9 : Αποτελέσματα μετρικών αξιολόγησης Feature Based Voting Classifier.

	Feature Based Voting Classifier					
Evaluation Metrics	$TestSet_1$	$TestSet_2$	$TestSet_3$	$TestSet_4$	$TestSet_5$	*Validation_Set*
Precision	0.843	0.847	0.86	0.868	0.853	0.914
Recall	0.833	0.839	0.852	0.862	0.851	0.913
Accuracy	0.87	0.876	0.89	0.867	0.849	0.912
Balanced Accuracy	0.871	0.877	0.891	0.862	0.851	0.913
F1_Score	0.832	0.838	0.852	0.86	0.849	0.911
Jaccard Score	0.77	0.78	0.8	0.766	0.738	0.843

Με βάση τα αποτελέσματα του πίνακα 5.9, καταλήγουμε στο συμπέρασμα ότι ο Feature Based Ταξινομητής Ψηφοφορίας παρουσιάζει το μικρότερο σφάλμα γενίκευσης σε σχέση με όλα τα υπόλοιπα μοντέλα που κατασκευάσαμε.

Στο Παράρτημα Γ παρουσιάζουμε όλα τα γραφήματα αξιολόγησης του εν λόγω Μοντέλου ταξινόμησης.

5.4 Συμπεράσματα

Οι εφαρμογές του εν λόγω κεφαλαίου που υλοποιήσαμε, κατέδειξαν την αξία της συμβολής των "Προφίλ" Φορτίου και της μηχανικής χαρακτηριστικών στη κατηγοριοποίηση χρονοσειρών ηλεκτρικού φορτίου. Η μεγάλη διάσταση των υπό ανάλυση δεδομένων στο πεδίο του χρόνου επιδρά αρνητικά στην επίδοση των αντίστοιχων μοντέλων ταξινόμησης καθώς προκαλεί την υπερμοντελοποίηση τους. Επίσης, ο πολύ μικρός αριθμός δεδομένων εκπαίδευσης αποτελεί εξίσου σημαντικό παράγοντα υπερμοντελοποίησης καi περιορίζει τη γενίκευση των μοντέλων. Με τεχνικές που πηγάζουν από τη μηχανική χαρακτηριστικών καταφέραμε να ξεπεράσουμε τα παραπάνω προβλήματα που αναδύονται κατά τη μοντελοποίηση χρονοσειρών ηλεκτρικού φορτίου. Συγκεκριμένα, καταφέραμε να αναπτύξουμε ένα επαρκές μοντέλο ταξινόμησης πολλών κλάσεων (multiclass classifier), των ημερήσιων καμπυλών φορτίου είκοσι τριών (23) ευρωπαϊκών χωρών. Το σύνολο εκπαίδευσης αποτελείται από εκατόν ενενήντα δύο (192) πρότυπες ημερήσιες χρονοσειρές συνολικού πραγματικού φορτίου (Σ.Π.Φ) οι οποίες αντιστοιχούν στις είκοσι τρεις (23) ευρωπαϊκές χώρες, και είναι το αποτέλεσμα των εφαρμογών ομαδοποίησης και εξαγωγής "Προφίλ" Φορτίου που αναλύσαμε στο τέταρτο κεφάλαιο. Δεδομένου της υψηλής επίδοσης του τελικού μοντέλου ταξινόμησης (Feature Based Voting Classifier) σε πέντε διαφορετικά σύνολα ελέγχου τα οποία περιέχουν πάνω από εβδομήντα πέντε χιλιάδες (75000) παρατηρήσεις το καθένα, συμπεραίνουμε την εγκυρότητα της πληροφορίας που ενσωματώνουν τα "Προφίλ" Φορτίου. Συνεπώς, μέσω του υπολογισμού κατάλληλων χαρακτηριστικών καμπυλών φορτίου ή αλλιώς "Προφίλ" Φορτίου, δύναται να αποθορυβοποιήσουμε το σύνολο των δεδομένων ανάλυσης και να να μοντελοποιήσουμε ικανοποιητικά την ενεργειακή συμπεριφορά των καταναλωτών (ευρωπαϊκές χώρες).

Συγκεκριμένα, το τελικό μοντέλο ταξινόμησης παρουσίασε τις καλύτερες επιδόσεις όσον αφορά τα δεδομένα που αντιστοιχούν στις κατηγορίες ES, FI, FR, HR, IT, ME, MK και RS. Οι αμέσως καλύτερες επιδόσεις παρουσιάστηκαν στις κατηγορίες GR, PL και PT και ικανοποιητικές επιδόσεις παρουσιάστηκαν στις κατηγορίες BG, CZ, DK, LT, LV, RO, SI και SK. Οι χειρότερες επιδόσεις παρουσιάστηκαν στις κατηγορίες EE NO, SE και UA. Μέσω των πινάκων σύγχησης που παρατίθενται στο Παράρτημα Γ, είναι προφανές ότι το τελικό μοντέλο ταξινόμησης διέπεται από ιδιαίτερη σύγχυση ανάμεσα στα δεδομένα των κλάσεων SE και NO , SE και UA καθώς και SE και PL. Επίσης πολλά δεδομένα της κατηγορίας EE κατηγοριοποιούνται εσφαλμένα ως LT και LV. Όσον αφορά τις χώρες EE, LT και LV, μέσω απλής εποπτείας του σχήματος των αντίστοιχων "Προφίλ" Φορτίου που παρουσιάσαμε στην ενότητα 4.2 του κεφαλαίου 4, συνειδητοποιούμε ότι οι χώρες αυτές έχουν πρακτικά όμοια ενεργειακή συμπεριφορά. Παρ' όλα αυτά, το τελικό μοντέλο ταξινόμησης κατηγοριοποιούσε εσφαλμένα μόνο τα δεδομένα EE ως LT και LV και όχι αντιστρόφως. Επίσης, όσον αφορά τις χώρες SE, NO, PL και UA, μέσω της εποπτείας των αντίστοιχων "Προφίλ" Φορτίου υποδεικνύεται ο μεγάλος βαθμός ομοιότητας της ενεργειακής τους συμπεριφοράς, δεδομένου ότι το σχήμα των αντίστοιχων καμπυλών αλλά και η τάξη μεγέθους των τιμών Σ.Π.Φ δεν διαφέρουν σημαντικά. Συνεπώς, τα αποτελέσματα της αξιολόγησης ως προς κάθε κλάση του τελικού Feature Based μοντέλου ταξινόμησης προσέφεραν μια γενική εικόνα της ομοιότητας και ως ένα βαθμό της ανομοιότητας της ενεργειακής συμπεριφοράς μεταξύ των κλάσεων, δηλαδή των ευρωπαϊκών χωρών του συνόλου ανάλυσης. Για παράδειγμα, όσον αφορά την ανομοιότητα ενεργειακής συμπεριφοράς, το τελικό μοντέλο ταξινόμησης σε κάθε σύνολο ελέγχου κατηγοριοποίησε ορθά όλες τις ημερήσιες καμπύλες Σ.Π.Φ της Γαλλίας. Συγκρίνοντας το σχήμα καθώς και τις τιμές Σ.Π.Φ των "Προφίλ" Φορτίου της Γαλλίας με των υπολοίπων ευρωπαϊκών χωρών, επιβεβαιώνεται ότι η ενεργειακή συμπεριφορά της Γαλλίας δεν ομοιάζει με άλλης χώρας

του συνόλου ανάλυσης. Επίσης αξίζει να σημειωθεί πως το τελικό μοντέλο ταξινόμησης δεν παρουσίασε σφάλματα μεταξύ των κλάσεων ΜΕ και RS παρόλο που τα προφίλ φορτίου των εν λόγω χωρών παρουσιάζουν μεγάλη ομοιότητα ως προς το σχήμα τους. Το γεγονός αυτό πιθανώς αιτιολογείται λόγω της μεγάλης διαφοράς που παρουσιάζουν αυτές οι δύο χώρες ως προς τη τάξη μεγέθους των τιμών τους.

Με βάση τα παραπάνω, και έχοντας υπόψη ότι η επεξεργασία - κατηγοριοποίηση των δεδομένων πραγματοποιήθηκε στο πεδίο των χαρακτηριστικών που παρουσιάζονται στην ενότητα 3.2 αυτού του κεφαλαίου, συμπεραίνεται ότι τα εν λόγω χαρακτηριστικά δύνανται να ενσωματώνουν αποδοτικά τη πληροφορία του σχήματος καθώς και της τάξης μεγέθους των τιμών των ημερήσιων καμπυλών φορτίου των χωρών του συνόλου ανάλυσης.

Όσον αφορά την ευρωστία του τελικού μοντέλου, παρατηρήσαμε όπως ήταν αναμενόμενο την επιδείνωση της επίδοσης του όσο τα δεδομένα ελέγχου απομακρύνονται χρονικά από τα δεδομένα εκπαίδευσης *("Προφίλ" Φορτίου),* τα οποία ορίζονται στο χρονικό παράθυρο ["01/03/2019", "29/02/2020"]. Παρ' όλα αυτά, το μέγιστο ποσοστό συνολικού σφάλματος υπολογίσθηκε ως 15% και παρουσιάστηκε στο σύνολο ελέγχου $TestSet_5$ το οποίο ορίζεται στο χρονικό παράθυρο ["01/03/2020", "29/02/2021] . Επίσης το ελάχιστο ποσοστό συνολικού σφάλματος υπολογίσθηκε ως 11% και παρουσιάστηκε στο σύνολο ελέγχου $TestSet_3$ το οποίο ορίζεται στο χρονικό παράθυρο ["01/03/2017", "28/02/2018"]. Δεδομένου ότι κατά τη περίοδο της πανδημίας του COVID-19 η παραγωγή και κατανάλωση ηλεκτρικής ενέργειας ενδέχεται να επηρεάστηκε σε μεγάλο βαθμό, η επαρκής επίδοση του μοντέλου στα αντίστοιχα δεδομένα ελέγχου είναι είτε ένδειξη αυξημένης ευρωστίας, είτε υποδεικνύει ότι η βιομηχανία της ενέργειας επηρεάστηκε αμυδρά κατά την εν λόγω περίοδο. Επίσης, ένδειξη ικανοποιητικής ευρωστίας είναι το γεγονός ότι το τελικό μοντέλο ταξινόμησης παρουσίασε εν γένει σε κάθε σύνολο ελέγχου πανομοιότυπα ανά κλάση σφάλματα.

Στο σημείο αυτό, αξίζει να σημειωθεί το γεγονός ότι τα τελικά δέκα πρώτα χαρακτηριστικά (ποσοστιαία σημεία και συστελεστές μετασχηματισμου κυματιδίων) επιλέχθηκαν βάση της επίδοσης του μοντέλου Τυχαίου Δάσους μέσω της μεθόδου περιτυλύγματος RF – RFE. Τα επόμενα τέσσερα χαρακτηριστικά *(συντελεστές FFT, και μεταβολή ποσοστιαίων σημειών)* από τα υπόλοιπα επτά τελικά χαρακτηριστικά, επιλέχθηκαν σύμφωνα με τη γενική μέθοδο φίλτρου "Mutual Information" (MI) ενώ τα τελευταία τρία (d1, d2, d3) επιλέχθηκαν βάση των υποδείξεων στη βιβλιογραφία [4, 11]. Συνεπώς, υπάρχει πιθανότητα περαιτέρω βελτίωσης της γενίκευσης της εφαρμογής, εφόσον υπολογίσουμε και επιλέξουμε για κάθε τύπο μοντέλου που συναθροίζεται στον Ταξινομητή Ψηφοφορίας σύνολα κατάλληλων χαρακτηριστικών για εφαρμογές ταξινόμησης, μέσω των μεθόδων επιλογής περιτυλίγματος KNN – RFE, SVM – RFE και RF – RFE.

Τέλος, όσον αφορά τα μέτρα αξιολόγησης, παρατηρώντας τις καμπύλες Receiver Operating Characteristics (ROC Curves) επιβεβαιώθηκε το γεγονός ότι τα γραφήματα ROC παρουσιάζουν υπερβολικά αισιόδοξη άποψη για την επίδοση των μοντέλων όταν τα σύνολα δεδομένων εισόδου διέπονται από μεγάλο βαθμό ασυμμετρίας (skewness).

Κεφάλαιο 6 : Συμπεράσματα – Μελλοντικές Επεκτάσεις

Στο κεφάλαιο αυτό πραγματοποιείται μια σύνοψη της παρούσας διπλωματικής εργασίας καθώς επίσης και μία αναφορά σε πιθανές μελλοντικές επεκτάσεις. Αναλυτικότερα, τα αποτελέσματα και η διαδικασία των εφαρμογών παρουσιάζονται στα αντίστοιχα κεφάλαια καθώς και στα Παραρτήματα Α, Β και Γ.

6.1 Σύνοψη και Συμπεράσματα

Στη παρούσα διπλωματική εργασία εφαρμόσαμε τεχνικές και αλγορίθμους μηχανικής μάθησης και εξόρυξης δεδομένων σε ενεργειακά δεδομένα και συγκεκριμένα σε χρονοσειρές ηλεκτρικού φορτίου είκοσι πέντε (25) ευρωπαϊκών χωρών. Στα πρώτα τρία κεφάλαια παρουσιάσαμε αναλυτικά όλο το θεωρητικό υπόβαθρο των εν λόγω αλγορίθμων και τεχνικών. Στα επόμενα δύο κεφάλαια παρουσιάσαμε επιμελώς όλα τα στάδια των εφαρμογών προσδίδωντας ιδιαίτερη έμφαση στη προετοιμασία των δεδομένων και στην αξιολόγηση των αποτελεσμάτων.

Όσον αφορά τις εφαρμογές, σε πρώτο στάδιο αξιοποιήσαμε τον διαμεριστικό αλγόριθμο ομαδοποίησης TimeSeriesK-Means++ [60, 32] της βιβλιοθήκης tslearn για την ομαδοποίηση των ημερήσιων καμπυλών φορτίου με βάση το σχήμα τους, και στη συνέχεια υπολογίσαμε για όλες τις ευρωπαϊκές χώρες του συνόλου ανάλυσης πρότυπες ημερήσιες καμπύλες φορτίου, γνωστές ως "Προφίλ" Φορτίου. Μέσω αυτών των προτύπων επιδιώκουμε σε επόμενο στάδιο να μοντελοποιήσουμε την ενεργειακή συμπεριφορά των ευρωπαϊκών χωρών του συνόλου ανάλυσης. Οι εφαρμογές ομαδοποίησης υλοποιήθηκαν σε δύο διαφορετικά πλαίσια ανάλυσης. Στο πρώτο πλαίσιο, πραγματοποιήσαμε την ανάλυση βάση των συνθηκών φόρτισης, δηλαδή εποχικά, κατά την οποία το σύνολο των δεδομένων διασπάστηκε σε υποσύνολα ανάλυσης σύμφωνα με τα χρονικά παράθυρα που ορίζονται από τις εποχές του έτους. Σε δεύτερο πλαίσιο, πραγματοποιήσαμε την ετήσια ανάλυση κατά την οποία επεξεργαστήκαμε όλες τις ημερήσιες καμπύλες φορτίου στο σύνολο του έτους. Οι μετρικές επικύρωσης ομαδοποίησης (CVI), που αξιοποιήθηκαν για την αξιολόγηση των ομαδοποιήσεων στο πλαίσιο της ετήσιας ανάλυσης, υπέδειξαν σε γενικές γραμμές το σχηματισμό ομάδων οι οποίες δεν διαχωρίζονταν επαρκώς θερμοκρασιακά. Συνεπώς, αυξήσαμε τον αριθμό των ομάδων πέραν των υποδείξεων των CVI μέχρι να σχηματιστούν ομάδες οι οποίες διέπονται από επαρκή θερμοκρασιακό διαχωρισμό. Συγκρίνοντας τα αποτελέσματα των ομαδοποιήσεων που προέκυψαν με βάση μόνο τις μετρικές CVI και τα αποτελέσματα των ομαδοποιήσεων που προέκυψαν με βάση το συνδιασμό των CVI και την ποιότητα του θερμοκρασιακού διαχωρισμού, οδηγηθήκαμε στο συμπέρασμα ότι οι εν λόγω μετρικές CVI υπολείπονται ως ένα βαθμό σε επιθυμητή ακρίβεια. Μετά το πέρας των εφαρμογών ομαδοποίησης, μέσω της εποπτείας των γραφημάτων των τελικών ομάδων και των αποτελεσμάτων των CVI, παρατηρήσαμε τη προβληματική και απρόβλεπτη ενεργειακή συμπεριφορά της Βοσνίας – Ερζεγοβίνης (BA) καθώς και της Ελβετίας (CH) που αντικατοπτρίζεται στο σχήμα των αντίστοιχων ημερήσιων καμπυλών φορτίου. Κατά συνέπεια, αποφασίσαμε να μην συμπεριλάβουμε τις παραπάνω χώρες στις εφαρμογές ταξινόμησης καθώς τα αντίστοιχα τους δεδομένα παρουσιάζουν αρκετό "θόρυβο".

Για την υλοποίηση των εφαρμογών ταξινόμησης σχηματίσαμε κατάλληλα μορφοποιημένα σύνολα δεδομένων τα οποία αξιοποιήθηκαν για την εκπαίδευση, βελτιστοποίηση και αξιολόγηση των μοντέλων ταξινόμησης που κατασκευάσαμε. Συγκεκριμένα, για την εκπαίδευση των μοντέλων ταξινόμησης αξιοποιήθηκαν τα εκατόν ενενήντα δύο (192)

"Προφίλ" Φορτίου της ετήσιας ανάλυσης τα οποία ορίζονται στο χρονικό παράθυρο ["01/03/2019", "29/02/2020"] . Για τη βελτιστοποίηση των μοντελων αξιοποιήθηκαν όλες οι ημερήσιες καμπύλες φορτίου του παραπάνω χρονικού παραθύρου. Επίσης, για τον ελέγχο και την αξιολόγηση των μοντέλων αξιοποιήθηκαν όλες οι ημερήσιες καμπύλες φορτίου που αντιστοιχούν στα χρονικά παράθυρα ["01/01/2015", "31/12/2015"], ["01/03/2016", "28/02/2019"] και ["01/03/2020", "28/02/2021"]. Προς αποφυγή παρεξηγήσεων, αναφερόμαστε σε όλες τις διαθέσιμες ημερήσιες καμπύλες φορτίου που προέκυψαν μετά το στάδιο της προετοιμασίας των δεδομένων, καθώς ένα ιδιαίτερα μικρό ποσοστό των αρχικών δεδομένων απορρίφθηκε λόγω των ακραίων τιμών και των ελλιπών παρατηρήσεων που ανιχνεύθηκαν.

Οι εφαρμογές ταξινόμησης ημερήσιων καμπυλών φορτίου πραγματοποιήθηκαν σε δύο διαφορετικά πλαίσια. Στο πρώτο πλαίσιο επεξεργαστήκαμε τα δεδομένα στο πεδίο του χρόνου (Time Instance Based Classification), δηλαδή ως χρονοσειρές, και σε δεύτερο πλαίσιο επεξργαστήκαμε τα δεδομένα στο πεδίο των χρακτηριστικών (Feature Based Classification). Οι κλασεις, δηλαδή οι κατηγορίες του προβλήματος, είναι οι είκοσι τρεις ευρωπαϊκές χώρες του συνόλου της ετήσιας ανάλυσης. Μέσω των εν λόγω εφαρμογών επιχειρήσαμε τη μοντελοποίηση της ενεργειακής συμπεριφοράς των ευρωπαϊκών χωρών του συνόλου ανάλυσης καθώς και την αξιολόγηση των αποτελεσμάτων των εφαρμογών ομαδοποίησης. Για την εξαγωγή των χαρακτηριστικών (features) από τις ημερήσιες χρονοσειρές φορτίου αξιοποιήθηκαν αλγόριθμοι της βιβλιοθήκης tsfresh [59] ενώ για την επιλογή των τελικά κατάλληλων χαρακτηριστικών αξιοποιήθηκαν αλγόριθμοι και μέθοδοι επιλογής και αξιολόγησης χαρακτηριστικών από τις βιβλιοθήκες tsfresh και scikit-learn [58]. Σημειώνεται επίσης ότι ενσωματώσαμε στο σύνολο των τελικών χαρακτηριστικών τρία χαρακτηριστικά τα οποία αναφέρονται στη βιβλιογραφία [4, 11] ως "Load Shape Features". Και στα δύο πλαίσια εφαρμογών ταξινόμησης εκπαιδεύτηκαν βελτιστοποιήθηκαν και αξιολογήθηκαν όλα τα μοντέλα ταξινόμησης που παρουσιάζονται στο δεύτερο κεφάλαιο της παρούσας διπλωματικής. Κατά την αξιολόγηση των εν λόγω μοντέλων ταξινόμησης, οδηγηθήκαμε στο συμπέρασμα ότι η επεξεργασία των δεδομένων στο πεδίο κατάλληλων χαρακτηριστικών προσφέρει σημαντικά πλεονεκτήματα. Συγκεκριμένα, μειώνονται οι διαστάσεις των υπό ανάλυση δεδομένων και αντιμετωπίζεται το πρόβλημα της υπερμοντελοποίησης των μοντέλων ταξινόμησης. Συνεπώς, δεδομένου της ικανοποιητικής επίδοσης του τελικού Feature Based ταξινομητή (Voting Classifier) συμπεραίνεται ότι μέσω κατάλληλων χαρακτηριστικών δύναται να διαχειριστούμε επαρκώς τη πληροφορία της τάξης μεγέθους των τιμών και του σχήματος των ημερήσιων καμπυλών φορτίου του συνόλου ανάλυσης, καθώς και να αντιμετωπίσουμε προβλήματα που αναδύονται στο πεδίο του χρόνου. Επιπρόσθετα, μέσω των πινάκων σύγχυσης (Confusion Matrices) του τελικού μοντέλου που παρουσιάζονται στο Παράρτημα Γ, καταφέραμε να σχηματίσουμε μια γενική εικόνα της ομοιότητας και ως ένα βαθμό της ανομοιότητας της ενεργειακής συμπεριφοράς των χωρών του συνόλου ανάλυσης. Στους εν λόγω πίνακες σύγχησης, αναδύθηκαν σημαντικά συστηματικά λάθη του τελικού μοντέλου ταξινόμησης μεταξύ δεδομένων συγκεκριμένων κατηγοριών. Μέσω της εποπτείας των "Προφίλ" Φορτίου που αντιστοιχούν στις εν λόγω κατηγορίες, επιβεβαιώθηκε η προφανής ομοιότητα της ενεργειακής τους συμπεριφοράς. Επίσης, δεδομένου ότι το σύνολο εκπαίδευσης, το οποίο αποτελείται από τα "Προφίλ" Φορτίου της ετήσιας ανάλυσης, είναι δραματικά μικρό σε σχέση με τα σύνολα ελέγχου, καταλήγουμε στο συμπέρασμα ότι μέσω της εξαγωγής κατάλληλων πρότυπων καμπυλών φορτίου δύναται η αποθορυβοποίηση των δεδομένων ανάλυσης καθώς και η υποστήριξη της μοντελοποίησης της ενεργειακής συμπεριφοράς των καταναλωτών.

6.2 Μελλοντικές Επεκτάσεις

Στο τελευταίο αυτό κεφάλαιο παραθέτουμε ως επίλογο τις πιθανές ενδιαφέρουσες μελλοντικές επεκτάσεις που φέρει η μελέτη της παρούσας διπλωματικής εργασίας.

- Η αξιοποίηση διαφορετικών μέτρων ομοιότητας και αλγορίθμων συσταδοποίησης καθώς και τεχνητών νευρωνικών δικτύων για την ομαδοποίηση χρονοσειρών ηλεκτρικού φορτίου.

- Υπάρχει πληθώρα μετρικών επικύρωσης ομαδοποίησης (CVI) που δεν αξιοποιήθηκαν στις εφαρμογές μη επιβλεπόμενης μάθησης που υλοποιήσαμε. Συνεπώς, η ενσωμάτωση περισσότερων μετρικών επικύρωσης στο στάδιο της αξιολόγησης των αποτελεσμάτων ενδέχεται να προσφέρει μεγαλύτερη αξιοπιστία στις εφαρμογές καθώς και καλύτερη αντίληψη της ποιότητας των ομαδοποιήσεων.

- Η μελέτη της τεχνικής Dynamic Barycenter Averaging (DBA) [44] για την εξαγωγή πρότυπων – χαρακτηριστικών καμπυλών φορτίου ("Προφίλ" Φορτίου).

- Η αξιοποίηση των "Προφίλ" Φορτίου σε εφαρμογές Πρόβλεψης (Παλινδρόμησης) ηλεκτρικού φορτίου.

- Η επέκταση του συνόλου ανάλυσης των εφαρμογών ώστε να συμπεριληφθούν και τα δεδομένα συνολικού πραγματικού φορτίου των υπολοίπων δέκα ευρωπαϊκών χωρών οι οποίες αντιπροσωπεύονται από τον ENTSO-E.

- Η βελτίωση της επίδοσης του Feature Based Voting Classifier μέσω της στοχευμένης εξαγωγής και επιλογής χαρακτηριστικών για κάθε μοντέλο ταξινόμησης του συνόλου συνάθροισης.

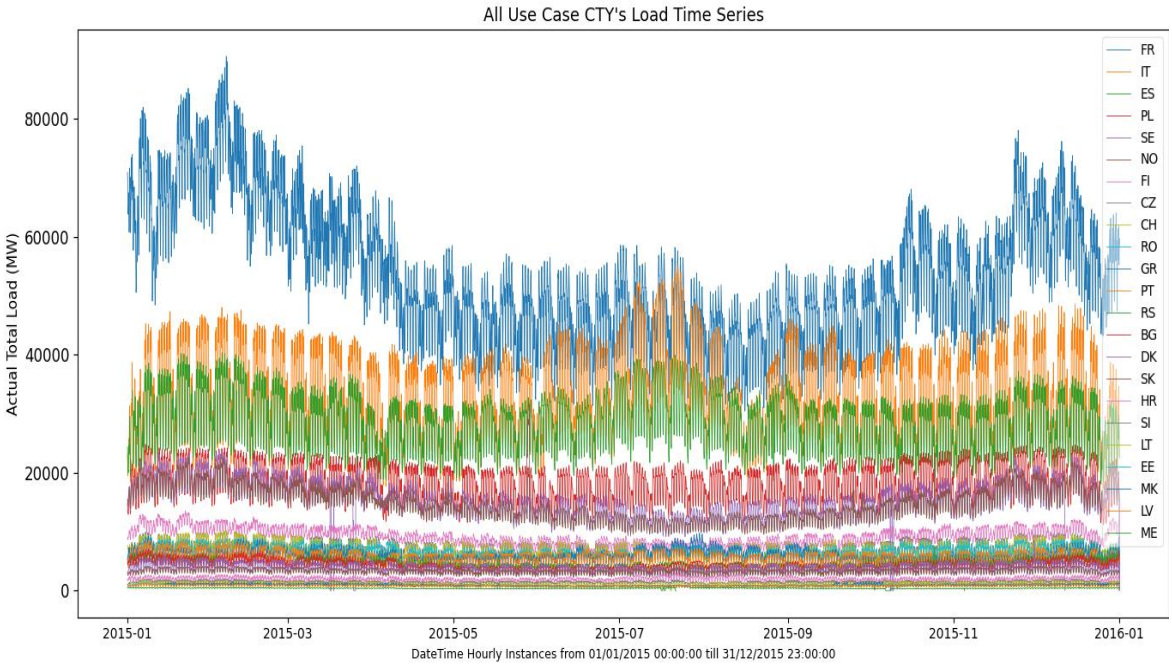
- Η αξιοποίηση αναδρομικών νευρωνικών δικτύων (Recursive Neural Networks - RNN) για τη κατηγοριοποίηση χρονοσειρών ηλεκτρικού φορτίου.

- Η επέκταση της εφαρμογής κατηγοριοποίησης πολλών αμοιβαία αποκλειόμενων κλάσεων (multiclass classification) της εν λόγω διπλωματικής σε μια εφαρμογή κατηγοριοποίησης πολλών κλάσεων με πολλές εξόδους (multiclass multilabel classification). Για παράδειγμα, μπορούμε να επισημάνουμε (label) τα δεδομένα όχι μόνο ως προς τη γεωγραφική περιοχή (χώρα) αλλά και ως προς τις συνθήκες φόρτισης (εποχή) που αντιπροσωπεύουν.

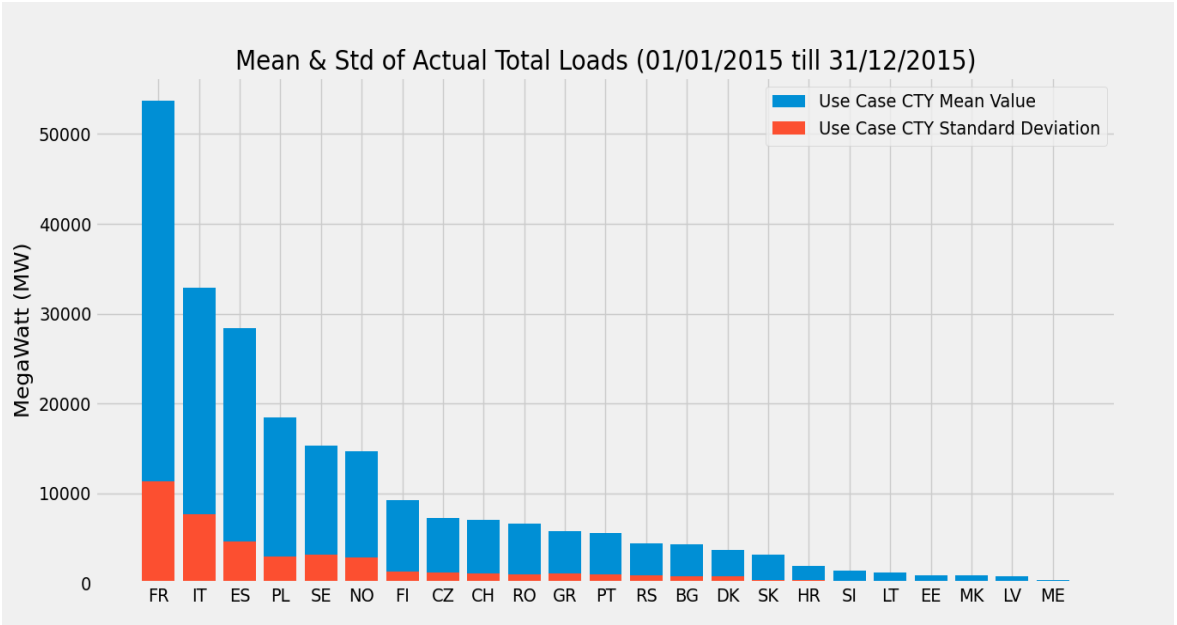
Παράρτημα Α : Σύνοψη Δεδομένων Πρώτου Σταδίου Προεπεξεργασίας

Στο παράρτημα αυτό παρουσιάζουμε συνοπτικά τα δεδομένα (ετήσιες χρονοσειρές φορτίου ευρωπαϊκών χωρών) μέσω κατάλληλων γραφημάτων. Τα αντίστοιχα γραφήματα των δεδομένων που ορίζονται από το χρονικό παράθυρο "01/03/2019" εώς "29/02/2020" παρουσιάζονται στη δεύτερη ενότητα του τετάρτου κεφαλαίου.

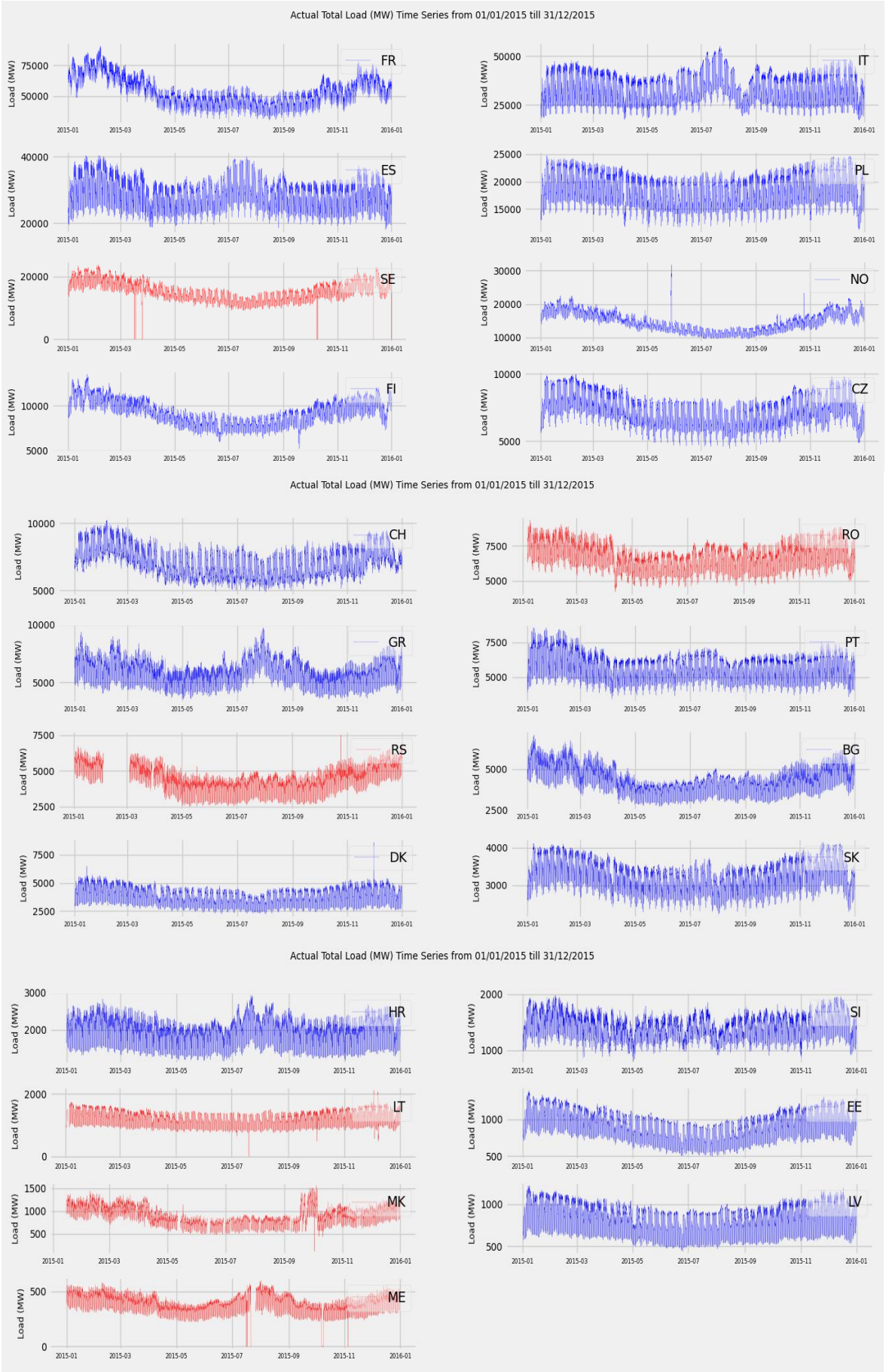
- **"01/01/2015" εώς "31/12/2015"**



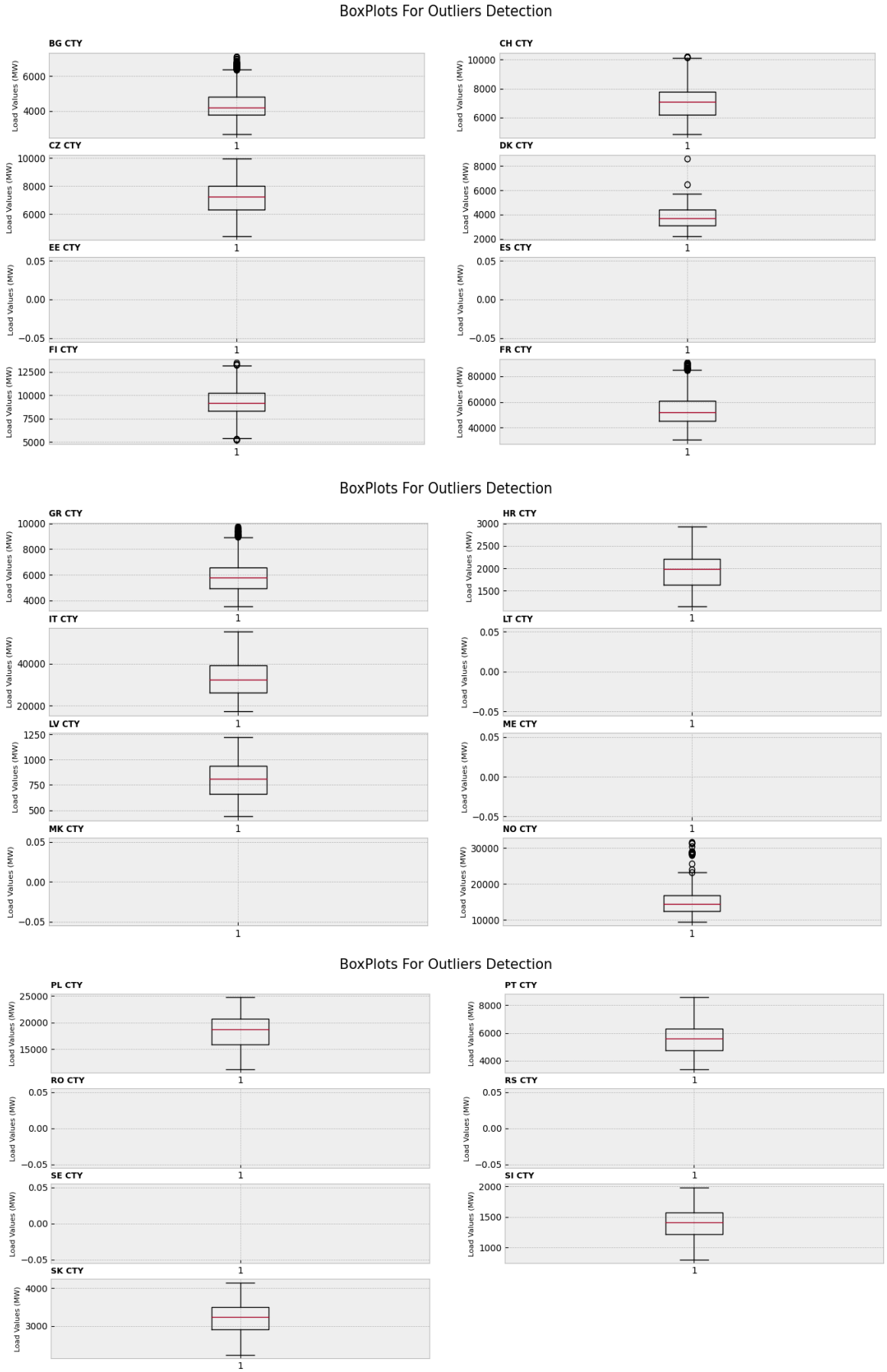
Σχήμα Π.Α.1 : Γραφική παράσταση των χρονοσειρών του έτους "2015".



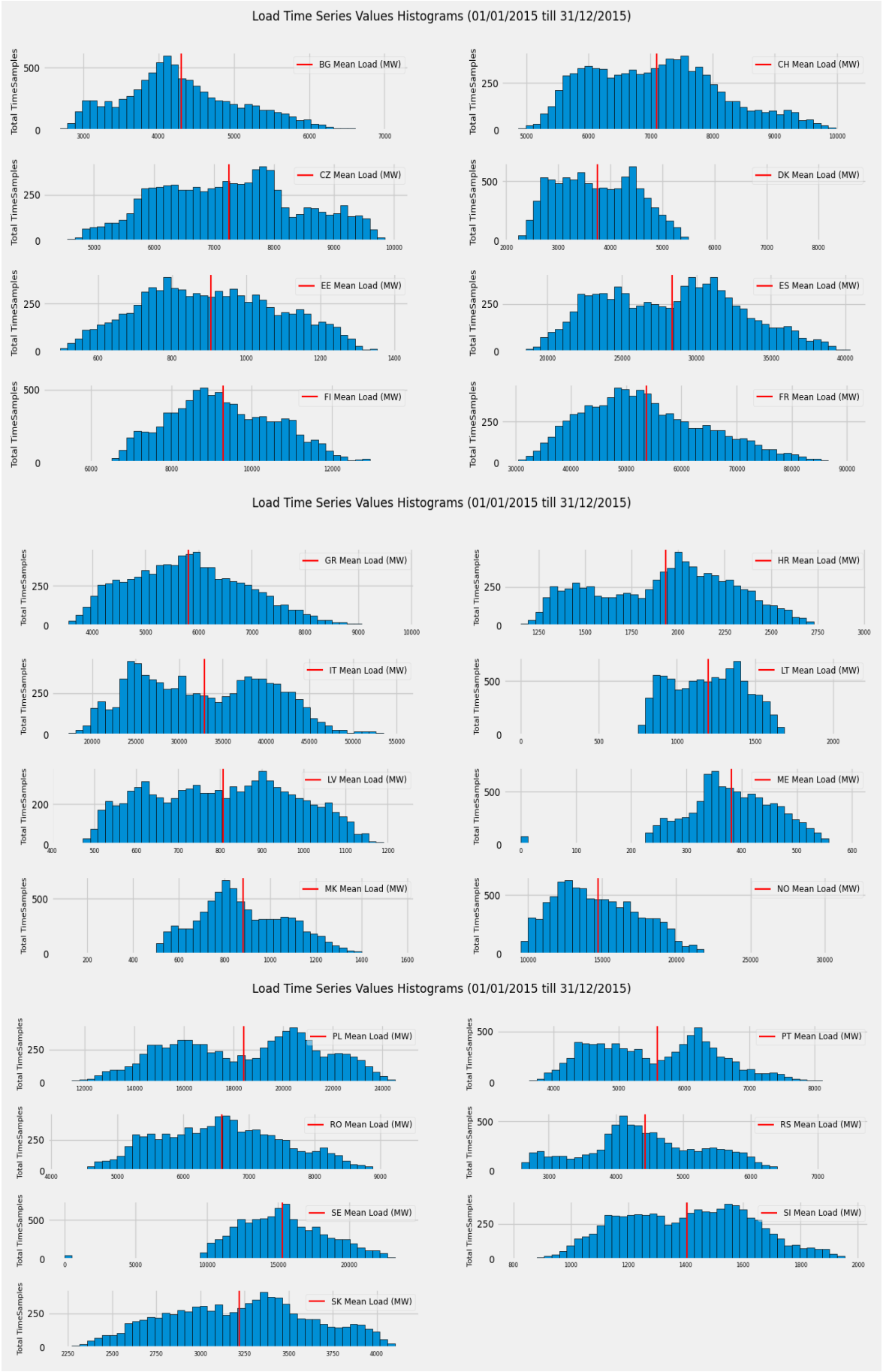
Σχήμα Π.Α.2 : Ραβδόγραμμα μέσης τιμής και τυπικής απόκλισης των χρονοσειρών του "2015".



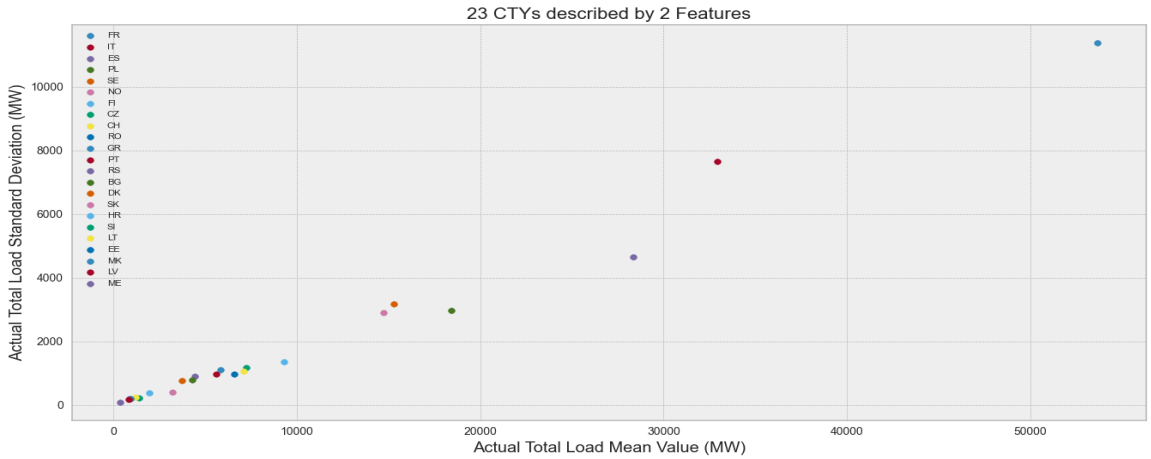
Σχήμα Π.Α.3 : Γραφικές παραστάσεις των ετήσιων χρονοσειρών του "2015".



Σχήμα Π.Α.4 : Διαγράμματα BoxPlots των ετήσιων χρονοσειρών του "2015".

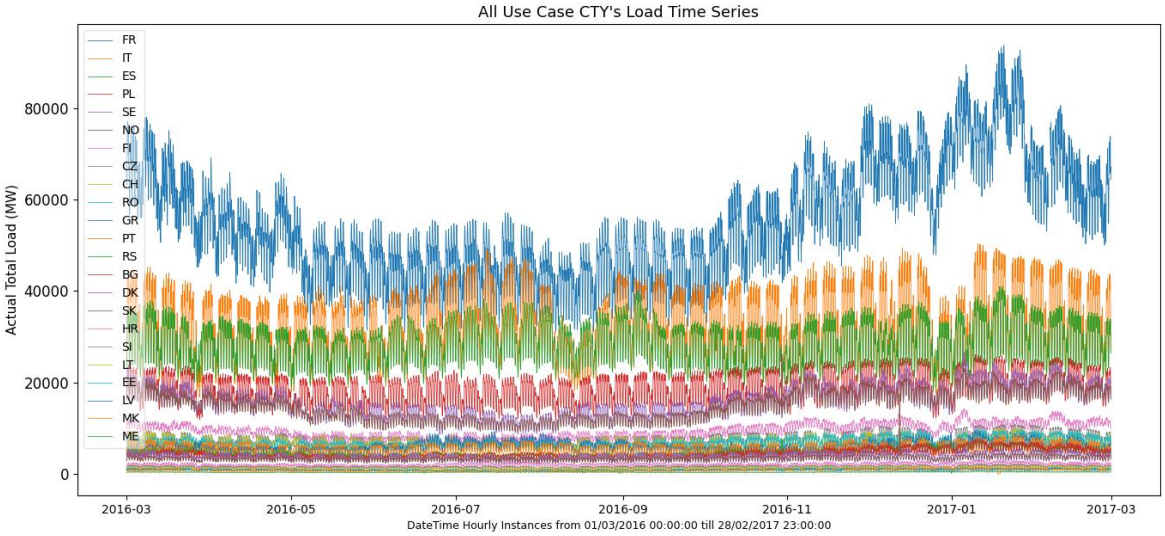


Σχήμα Π.Α.5 : Ιστογράμματα των ετήσιων χρονοσειρών του "2015".

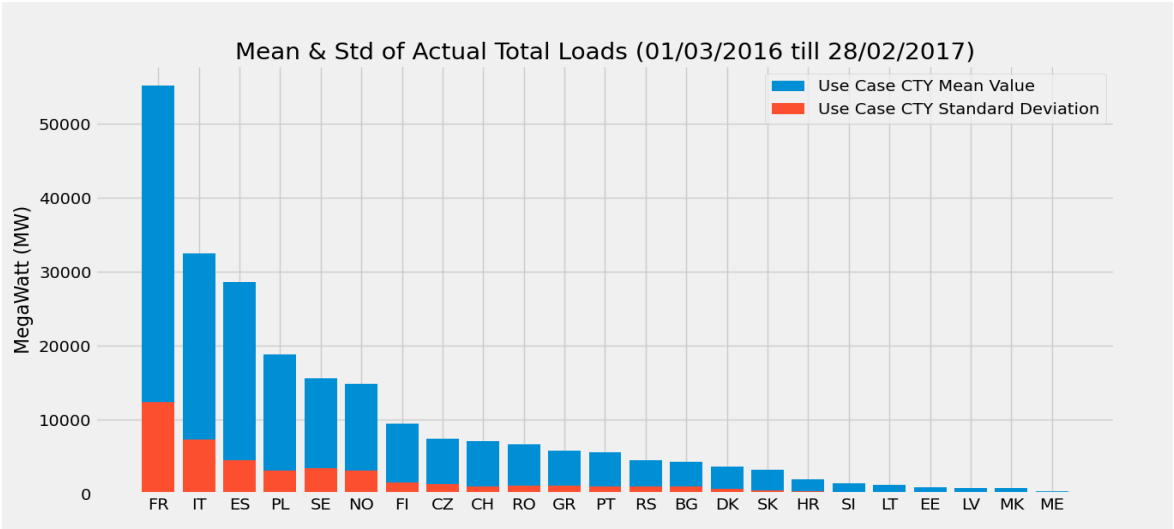


Σχήμα Π.Α.6 : Δισδιάστατη οπτικοποίηση χρονοσειρών του "2015" ως προς τη μέση τιμή και τη τυπική τους απόκλιση.

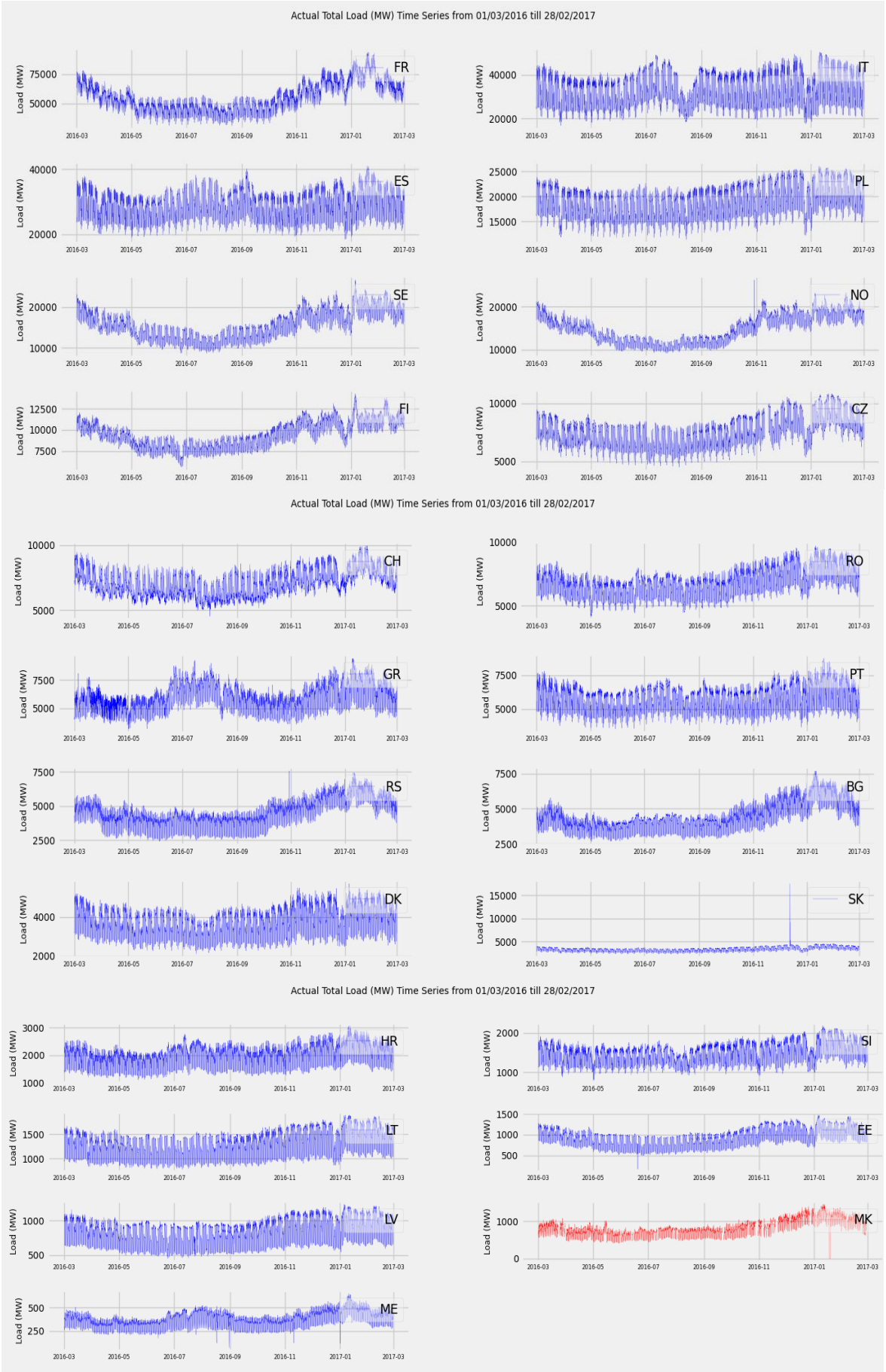
- **"01/03/2016" εώς "28/02/2017"**



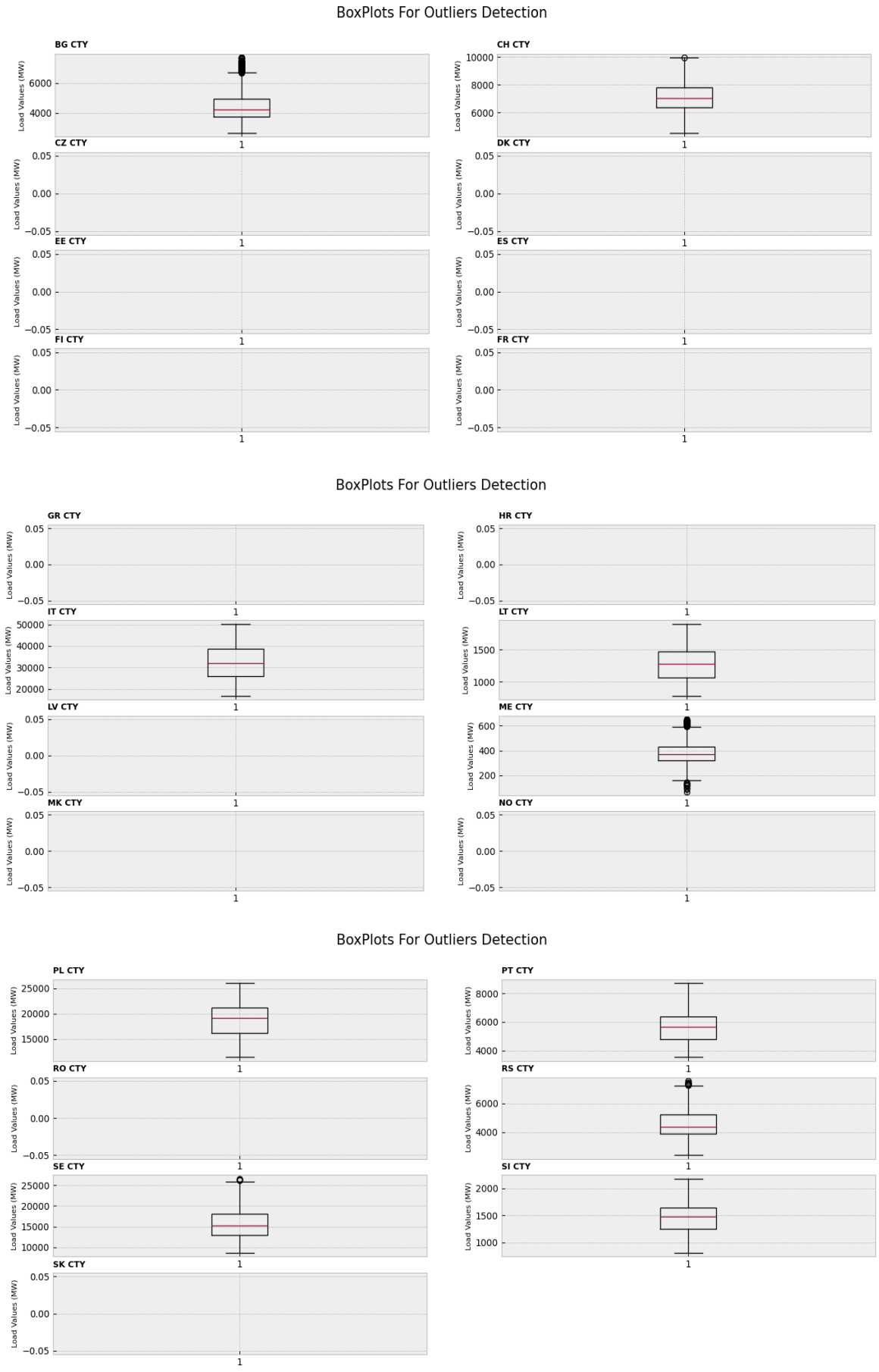
Σχήμα Π.Α.7 : Γραφική παράσταση των χρονοσειών του "2016".



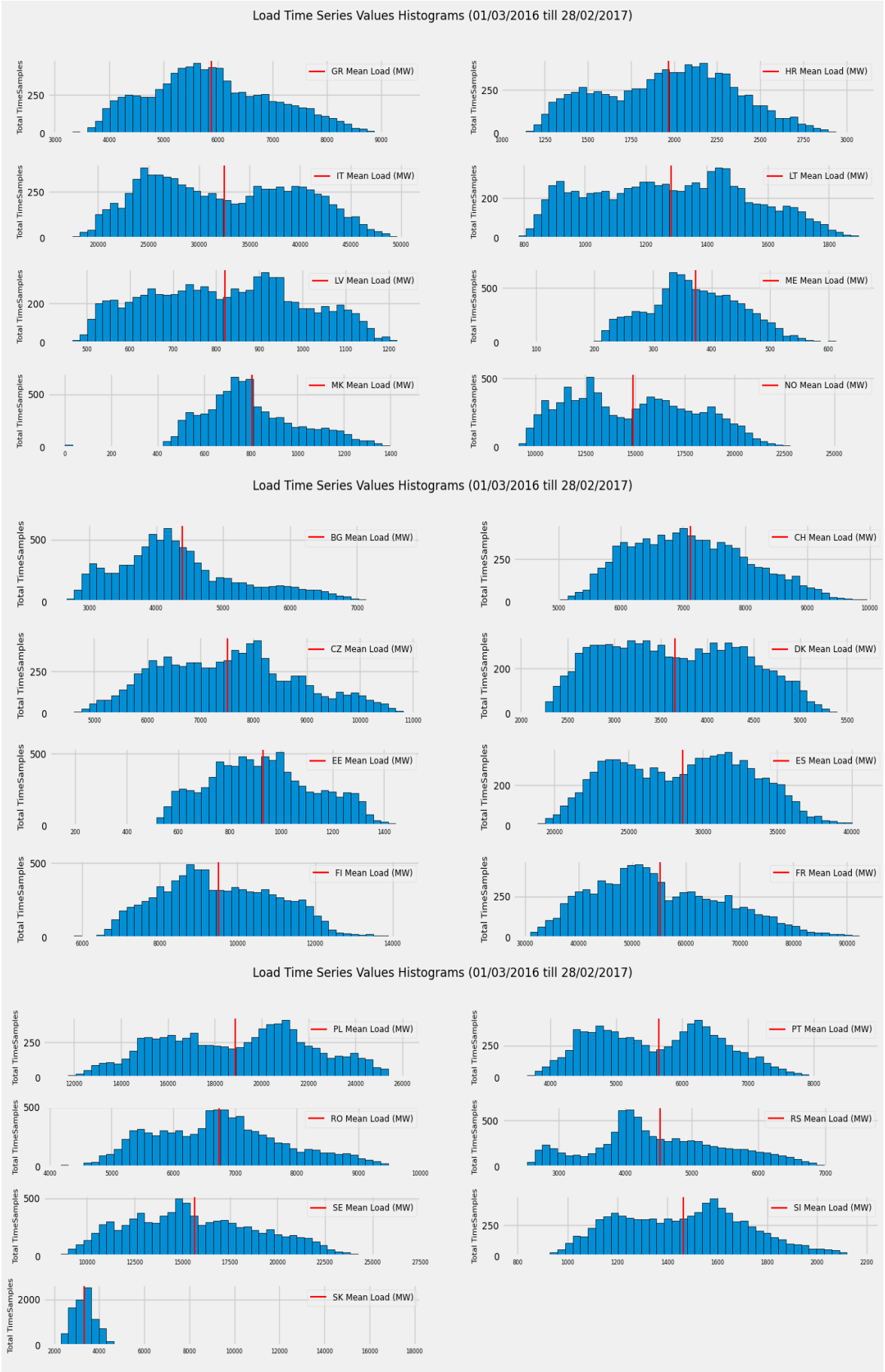
Σχήμα Π.Α.8 : Ραβδόγραμμα μέσης τιμής και τυπικής απόκλισης των χρονοσειρών του "2016".



Σχήμα Π.Α.9 : Γραφικές παραστάσεις των ετήσιων χρονοσειρών του "2016".

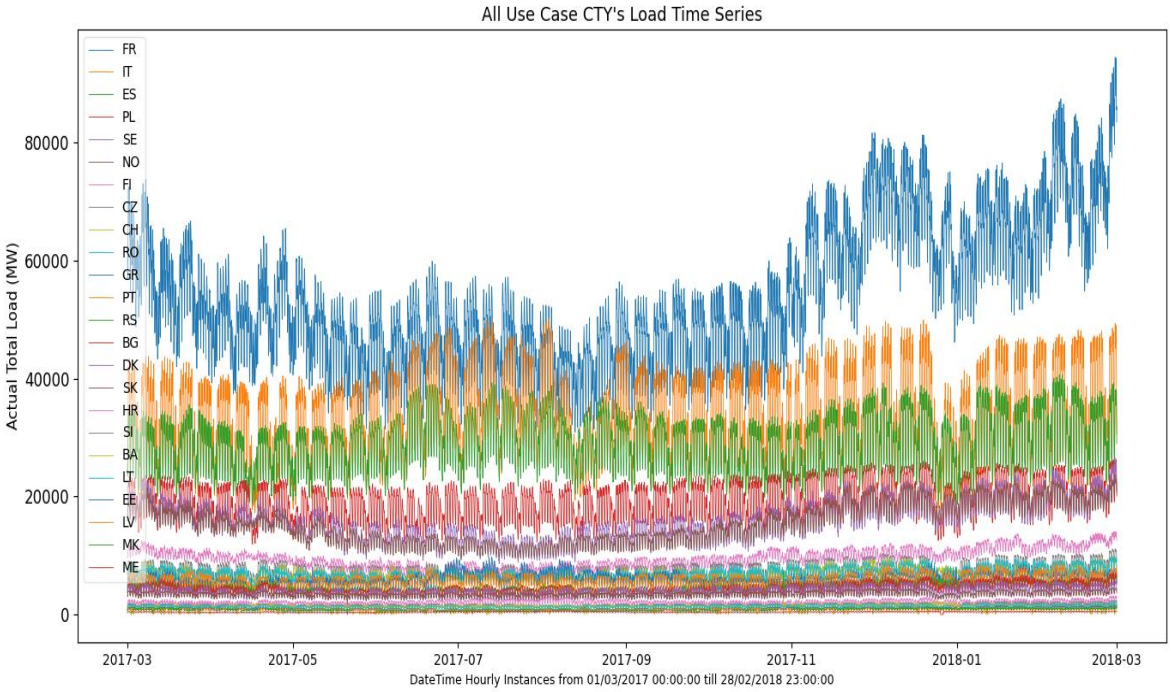


Σχήμα Π.Α.10 : Διαγράμματα BoxPlots των ετήσιων χρονοσειρών του "2016".

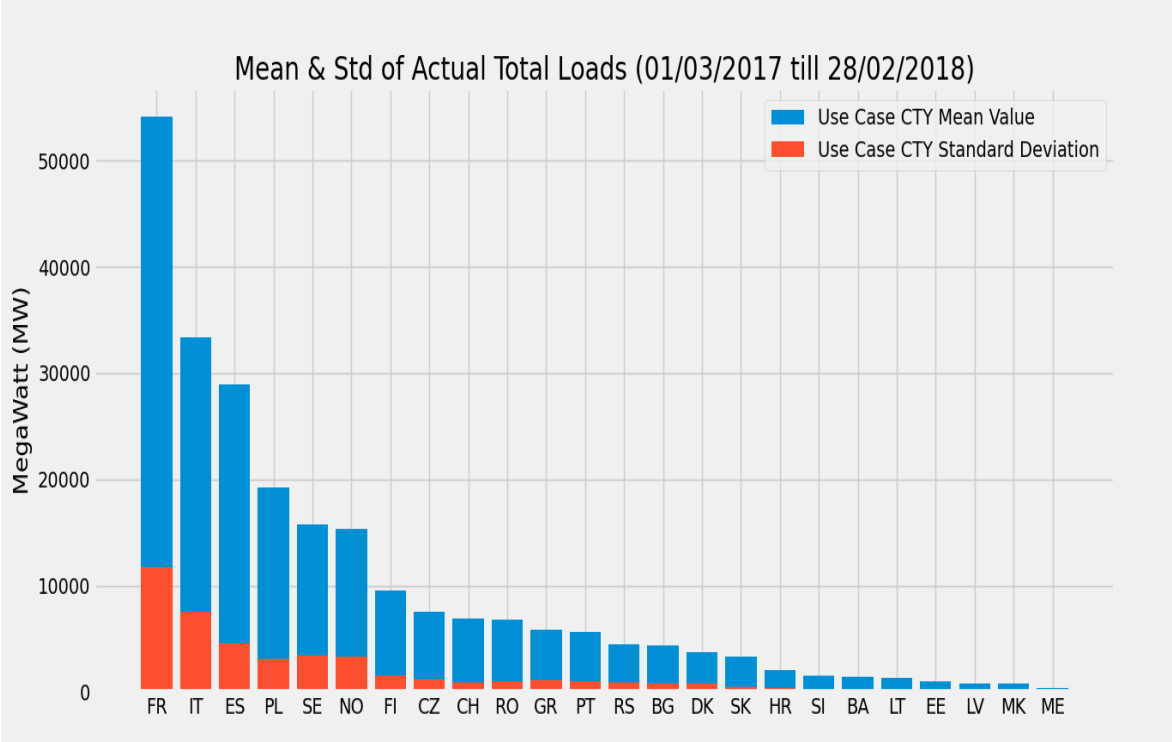


Σχήμα Π.Α.11 : Ιστογράμματα ετήσιων χρονοσειρών του "2016".

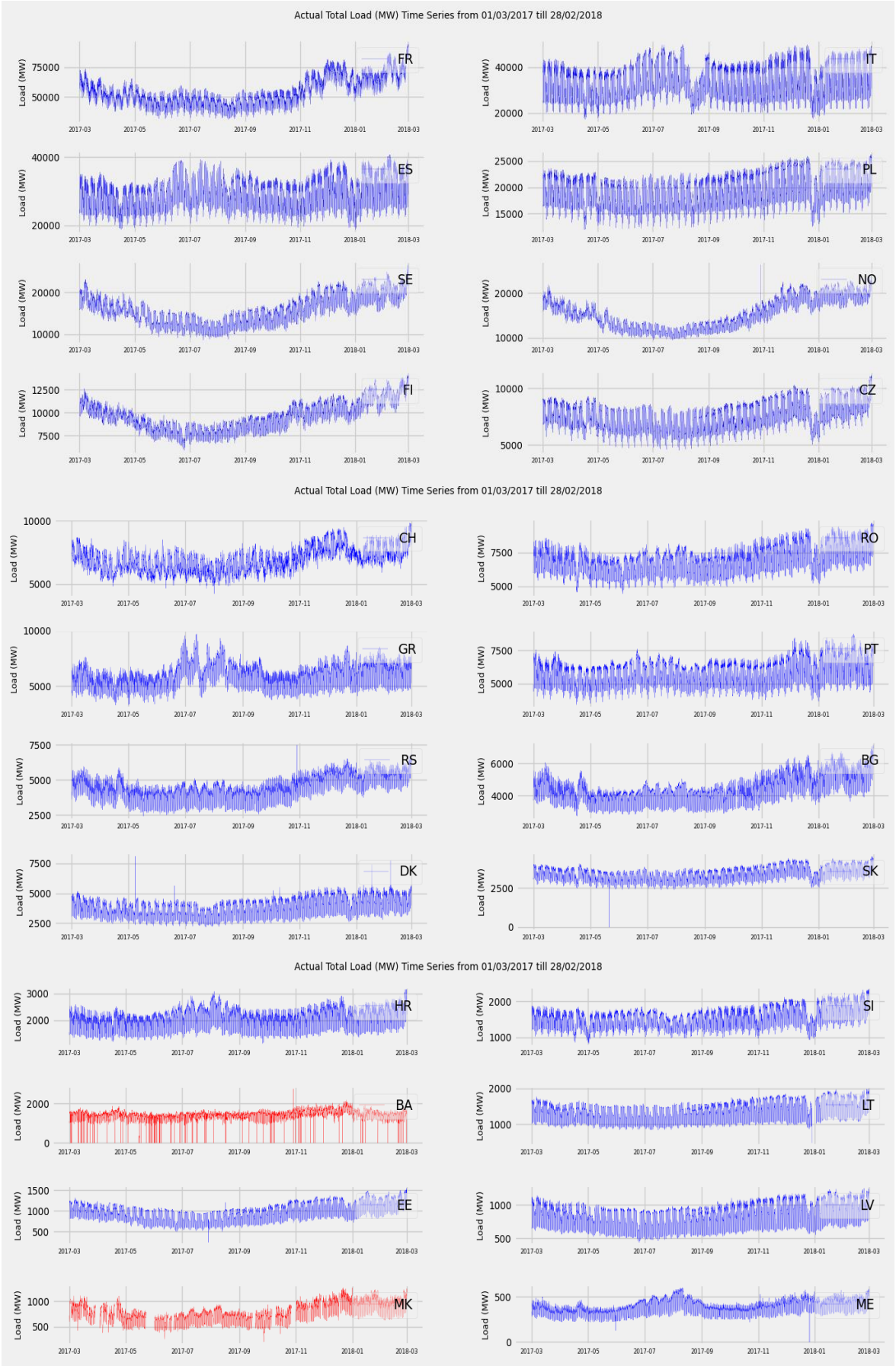
- **"01/03/2017" εώς "28/02/2018"**



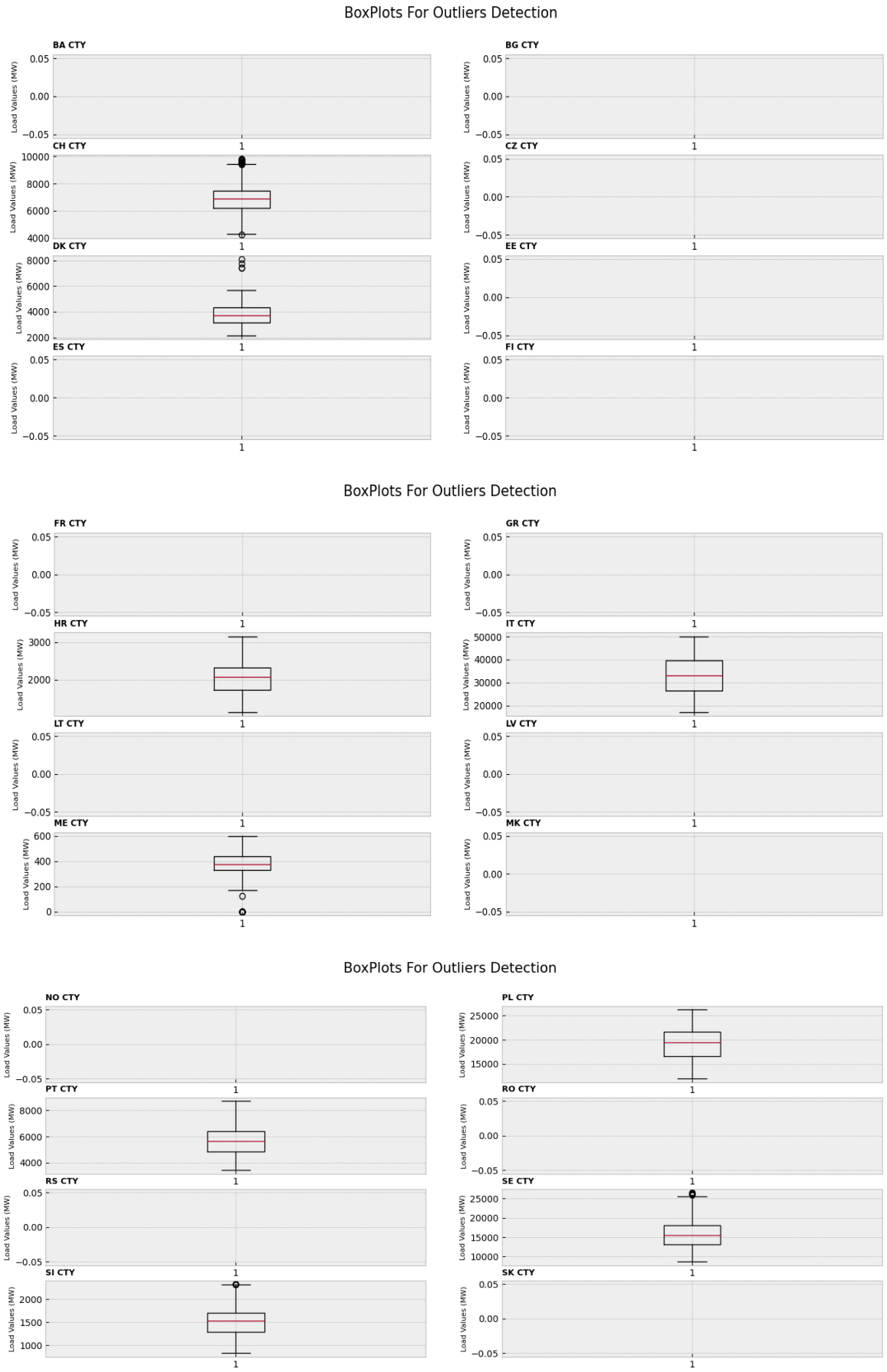
Σχήμα Π.Α.12 : Γραφική παράσταση των χρονοσειρών του "2017".



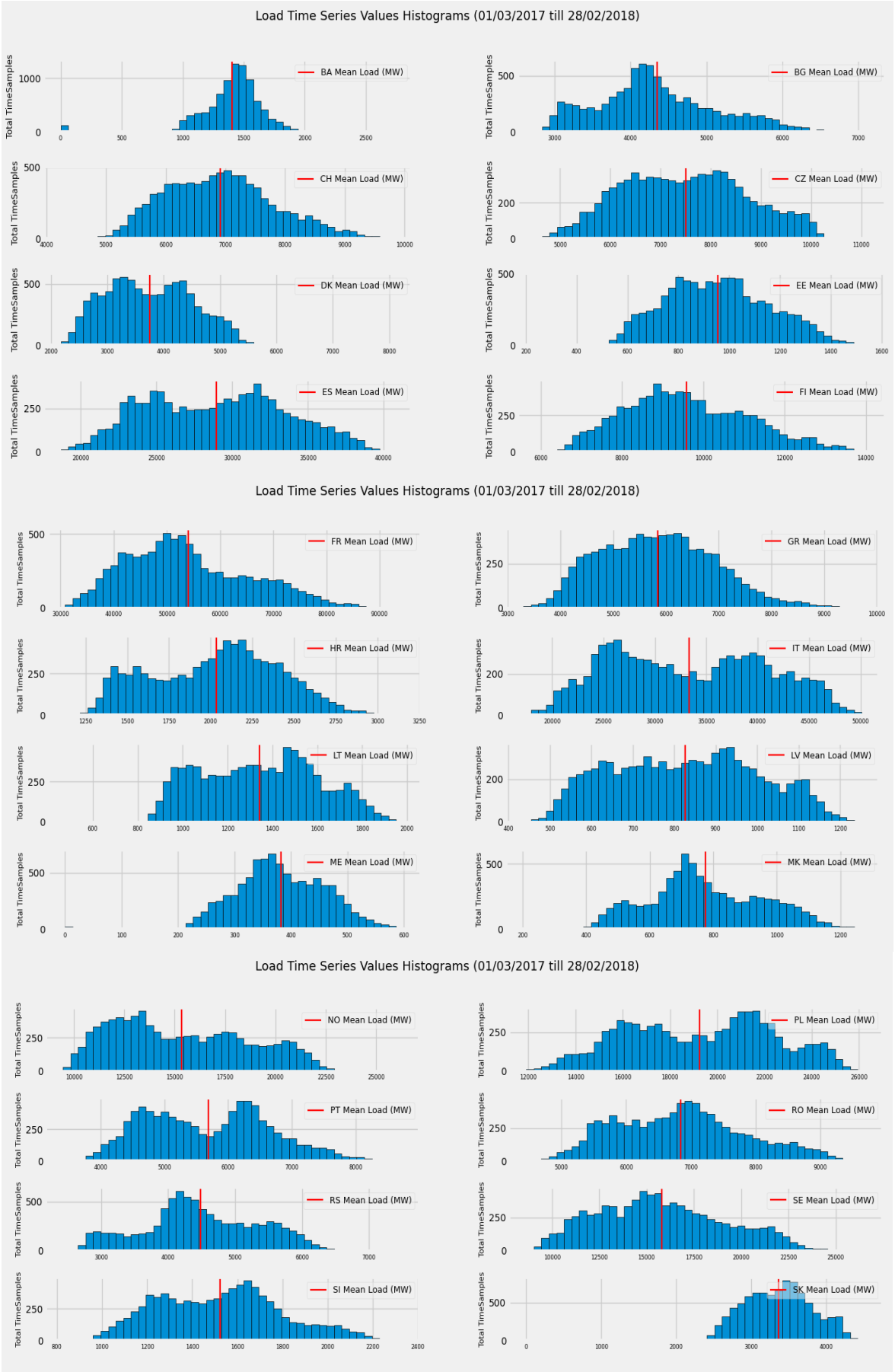
Σχήμα Π.Α.13 : Ραβδόγραμμα μέσης τιμής και τυπικής απόκλισης των χρονοσειρών του "2017".



Σχήμα Π.Α.14 : Γραφικές παραστάσεις των ετήσιων χρονοσειρών του "2017".

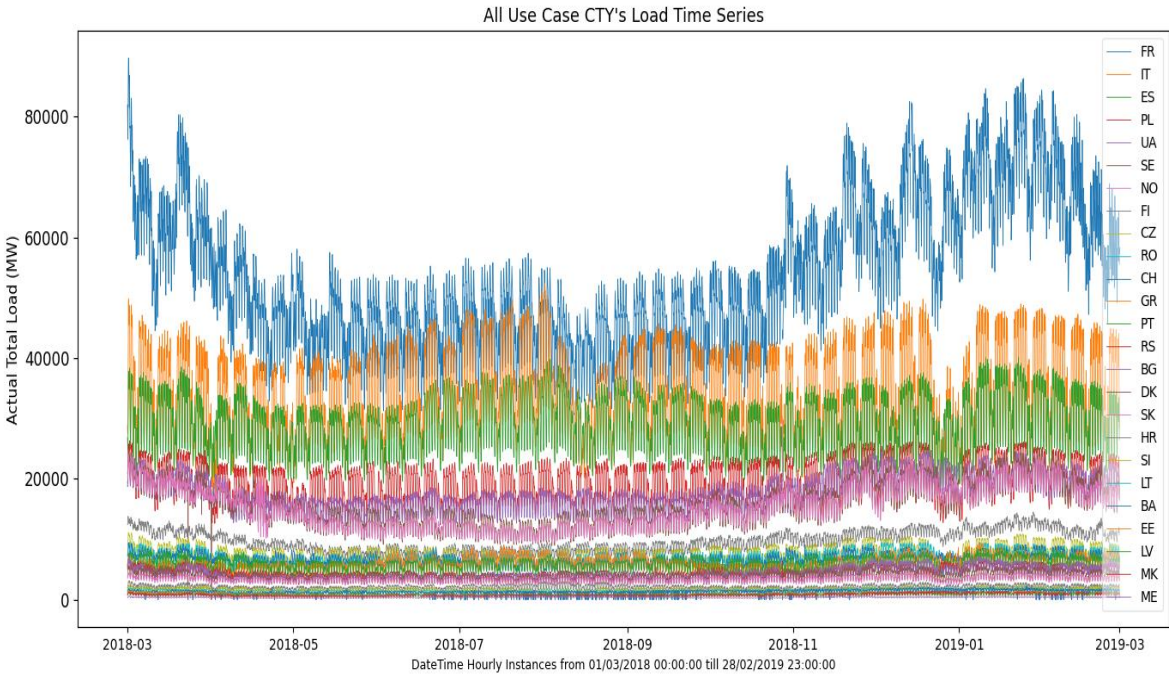


Σχήμα Π.Α.15 : Διαγράμματα BoxPlots των ετήσιων χρονοσειρών του "2017".

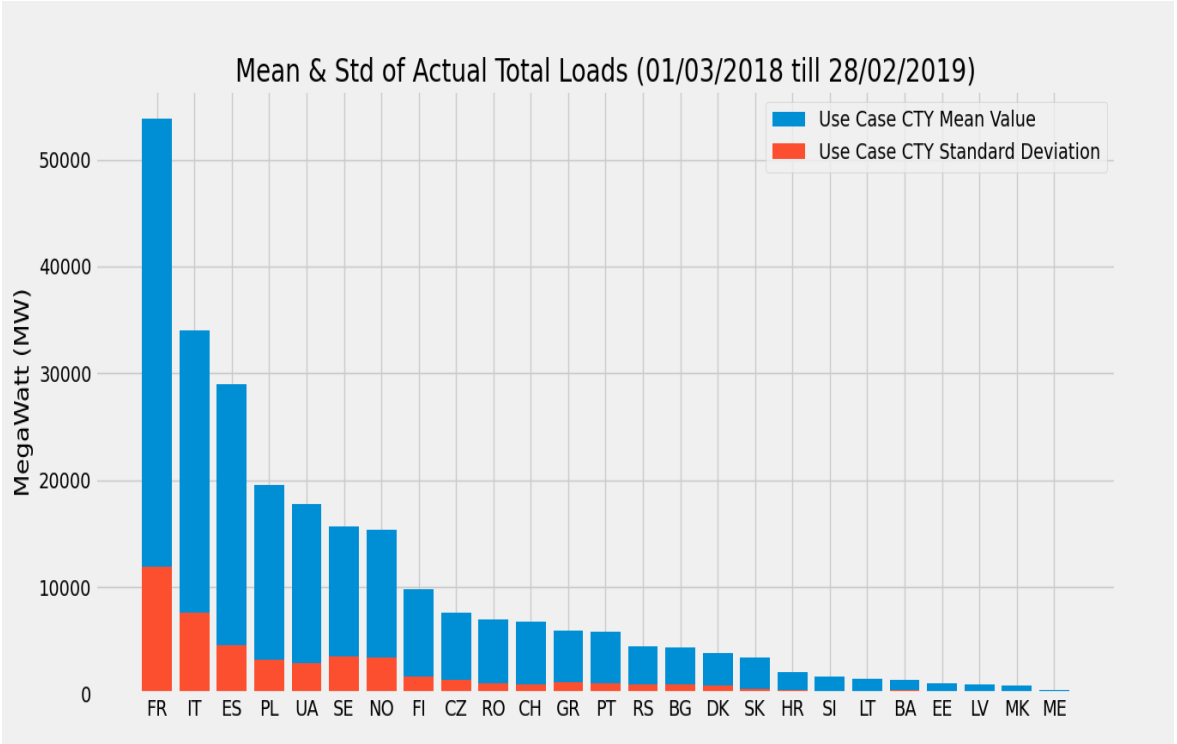


Σχήμα Π.Α.16 : Ιστογράμματα ετήσιων χρονοσειρών του "2017".

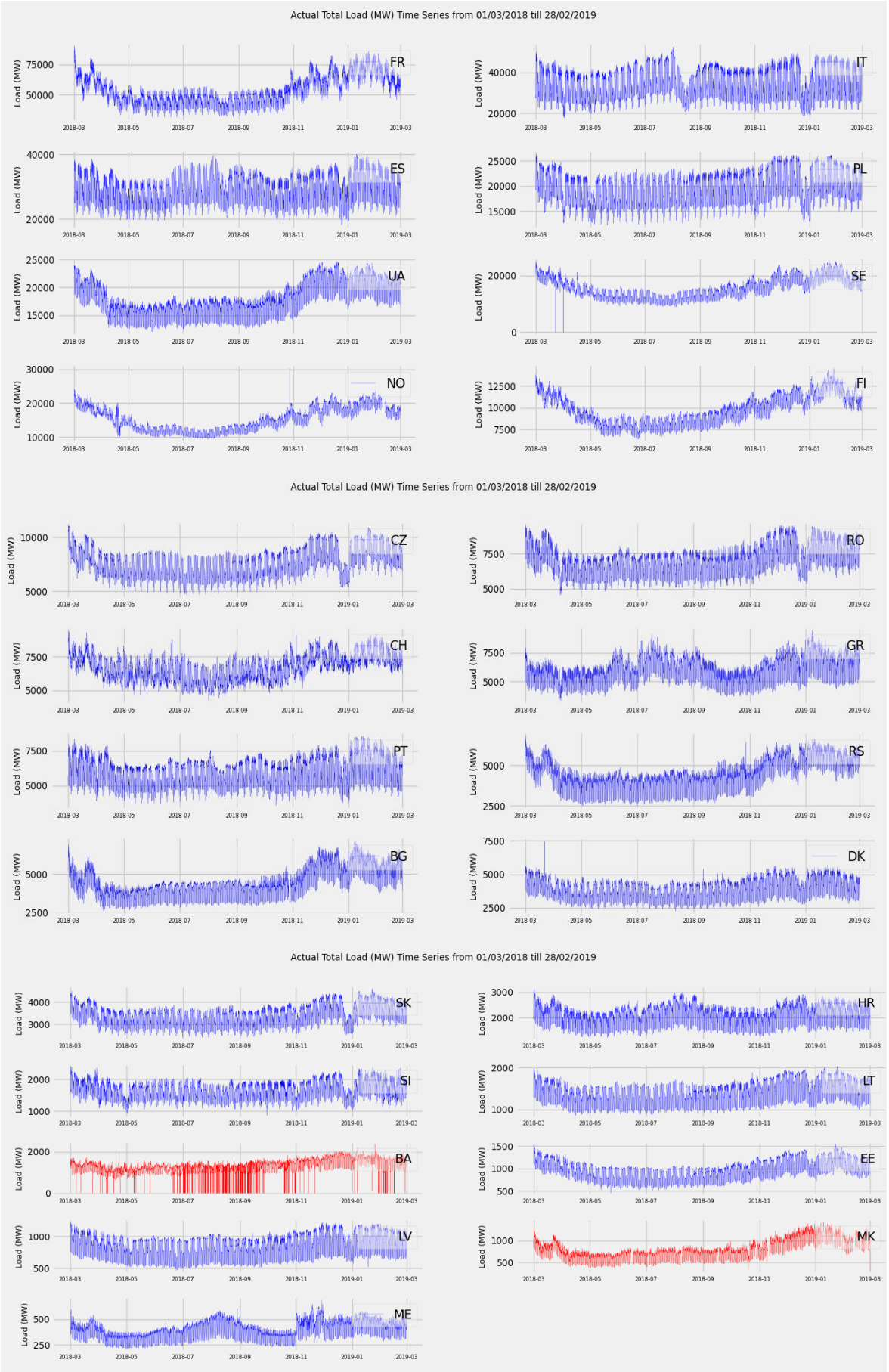
- **"01/03/2018" εώς "28/02/2019"**



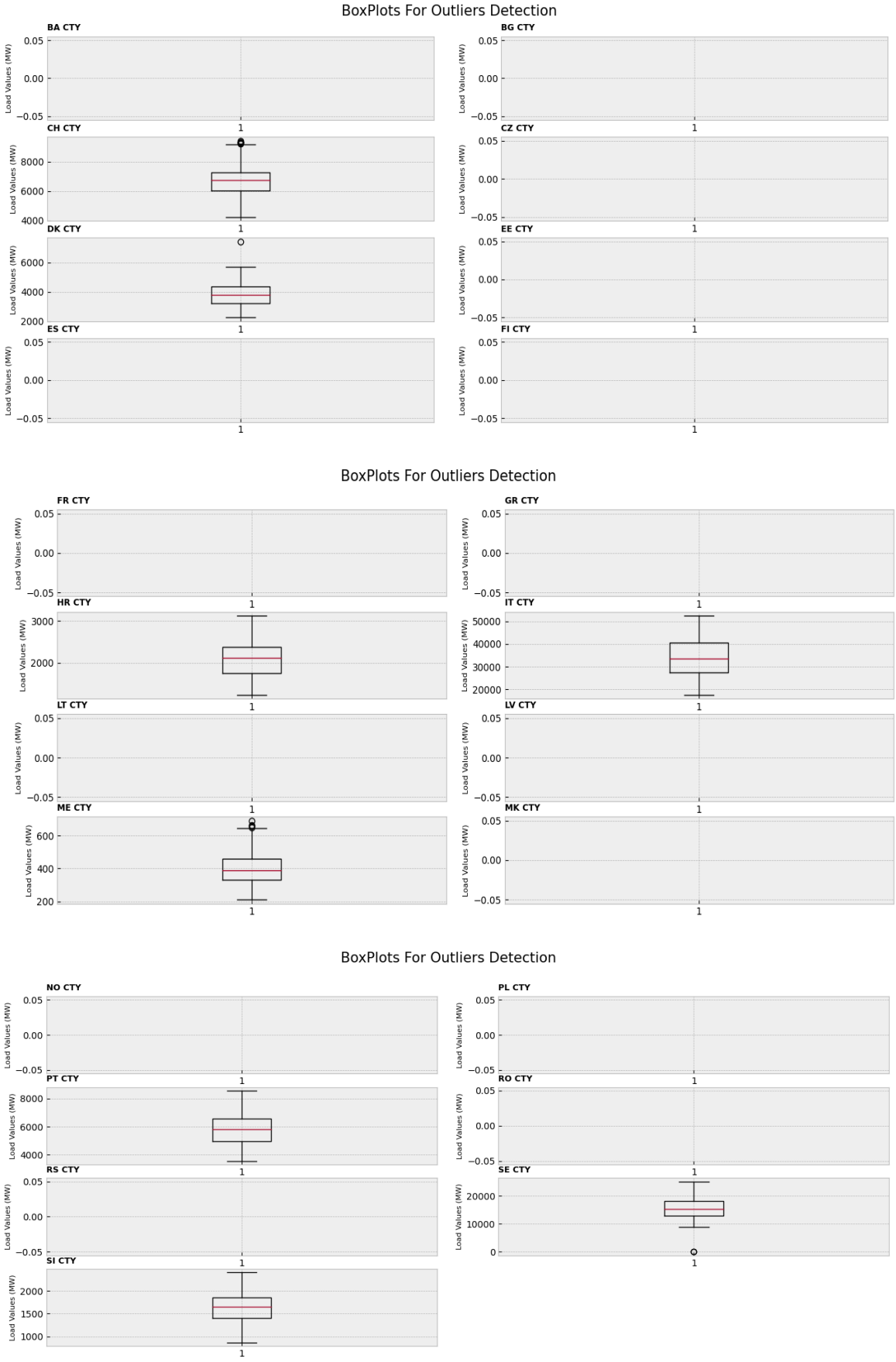
Σχήμα Π.Α.17 : Γραφική παράσταση των χρονοσειρών του έτους "2018".



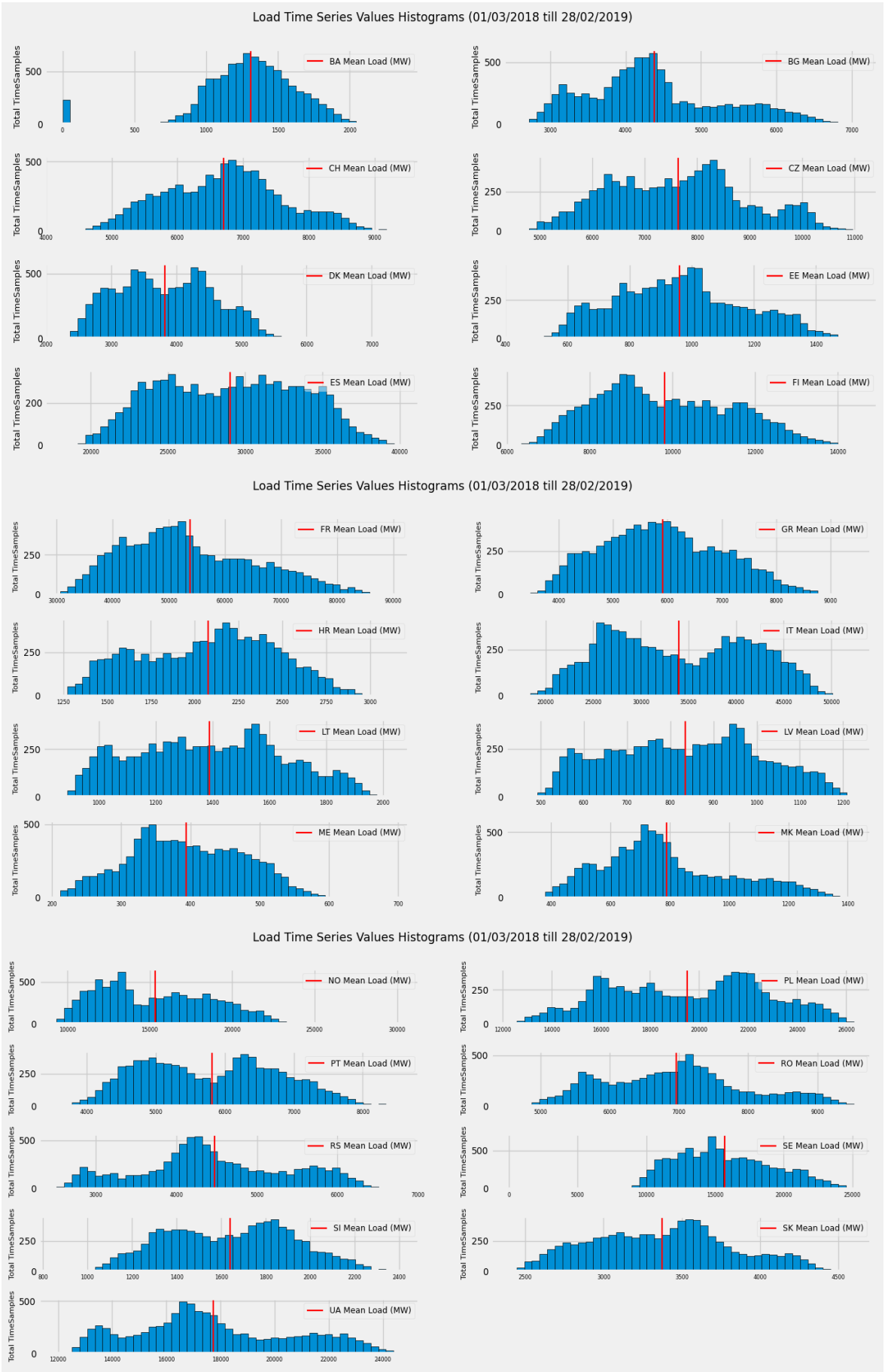
Σχήμα Π.Α.18 : Ραβδόγραμμα μέσης τιμής και τυπικής απόκλισης χρονοσειρών του έτους "2018".



Σχήμα Π.Α.19 : Γραφικές παραστάσεις ετήσιων χρονοσειρών του "2018".

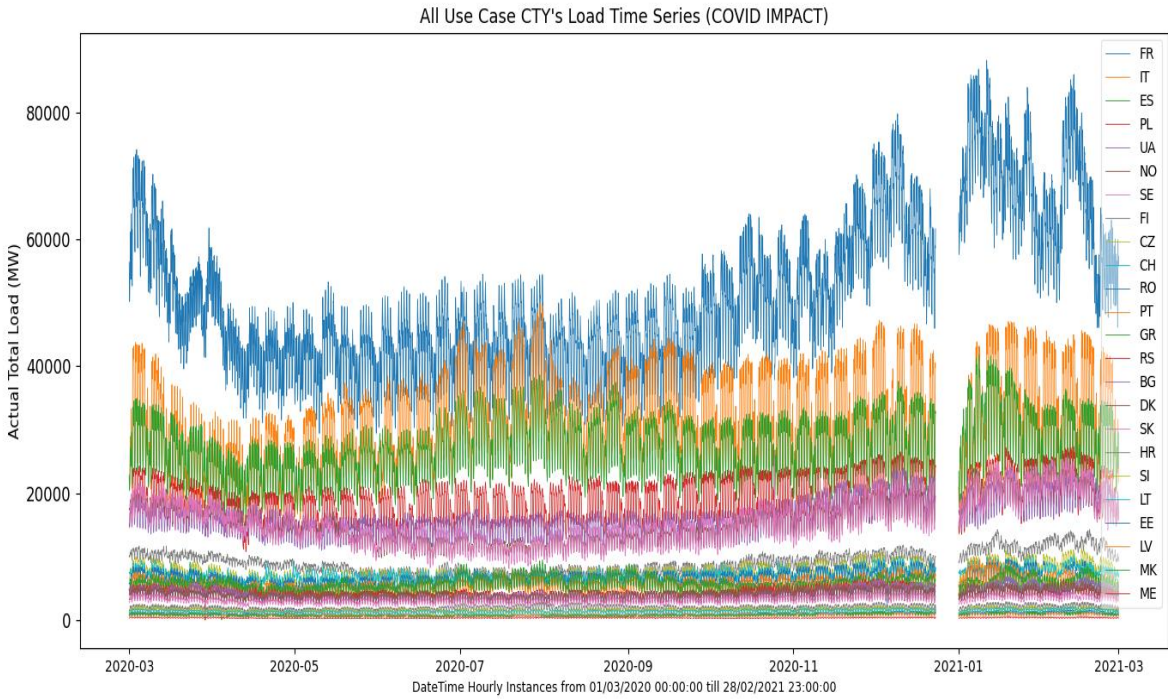


Σχήμα Π.Α.20 : Διαγράμματα BoxPlots των ετήσιων χρονοσειρών του "2018".

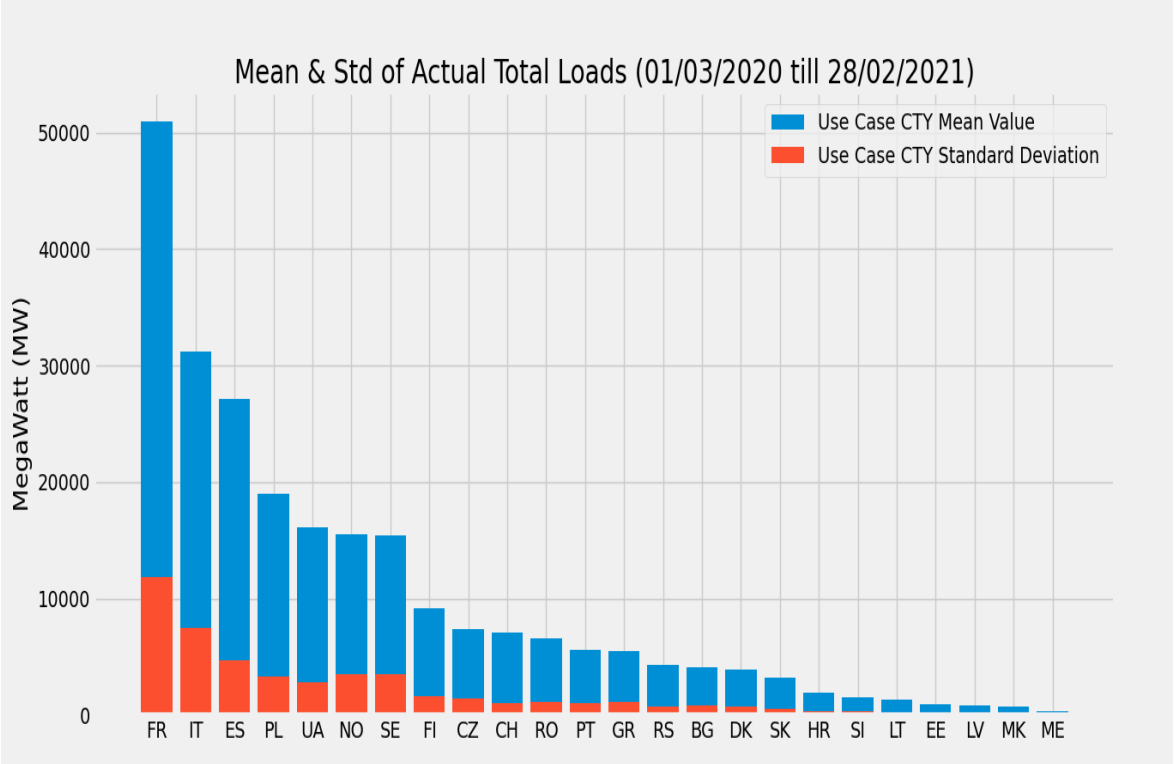


Σχήμα Π.Α.21 : Ιστογράμματα ετήσιων χρονοσειρών του "2018".

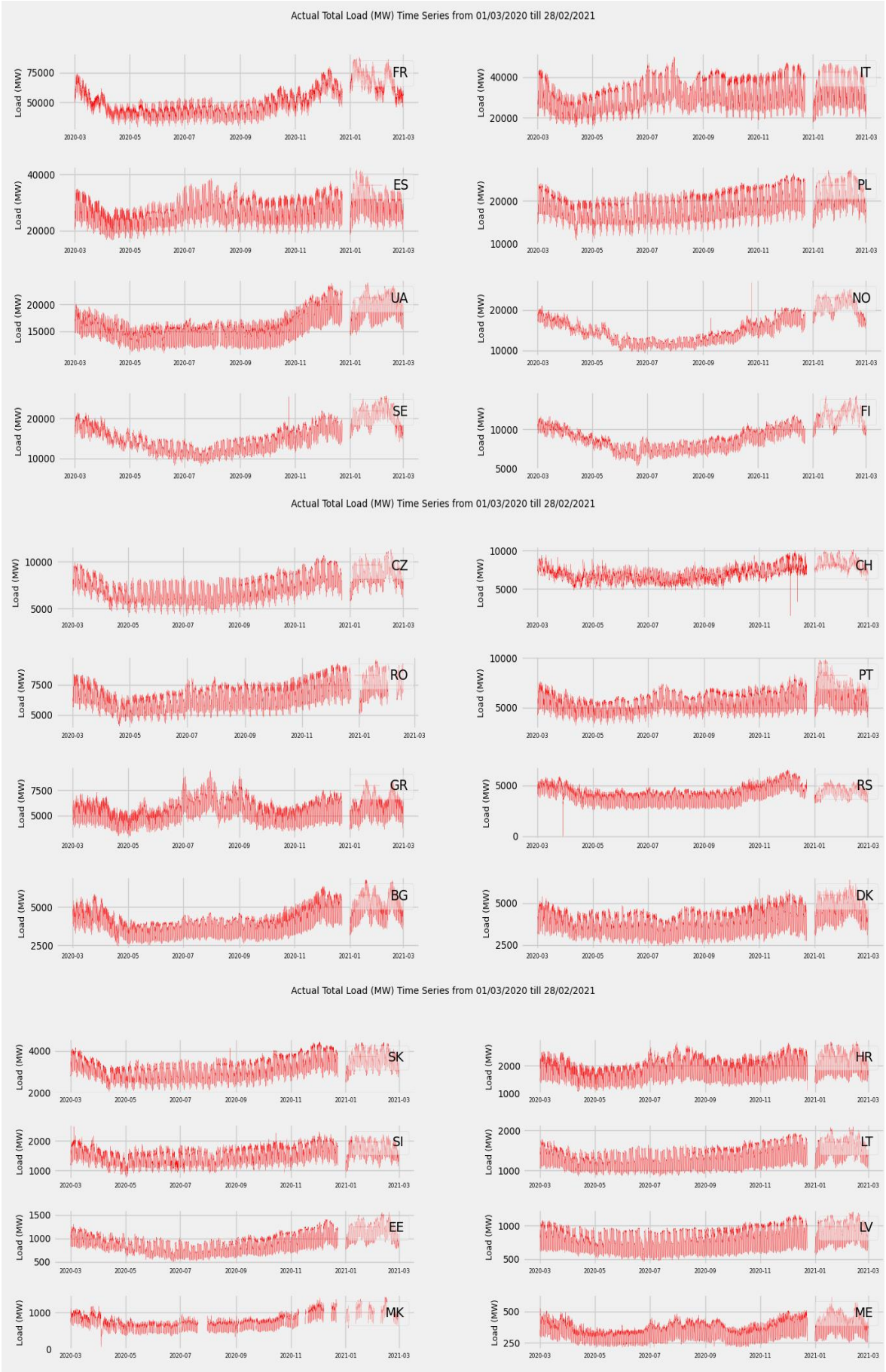
- **"01/03/2020" εώς "28/02/2021"**



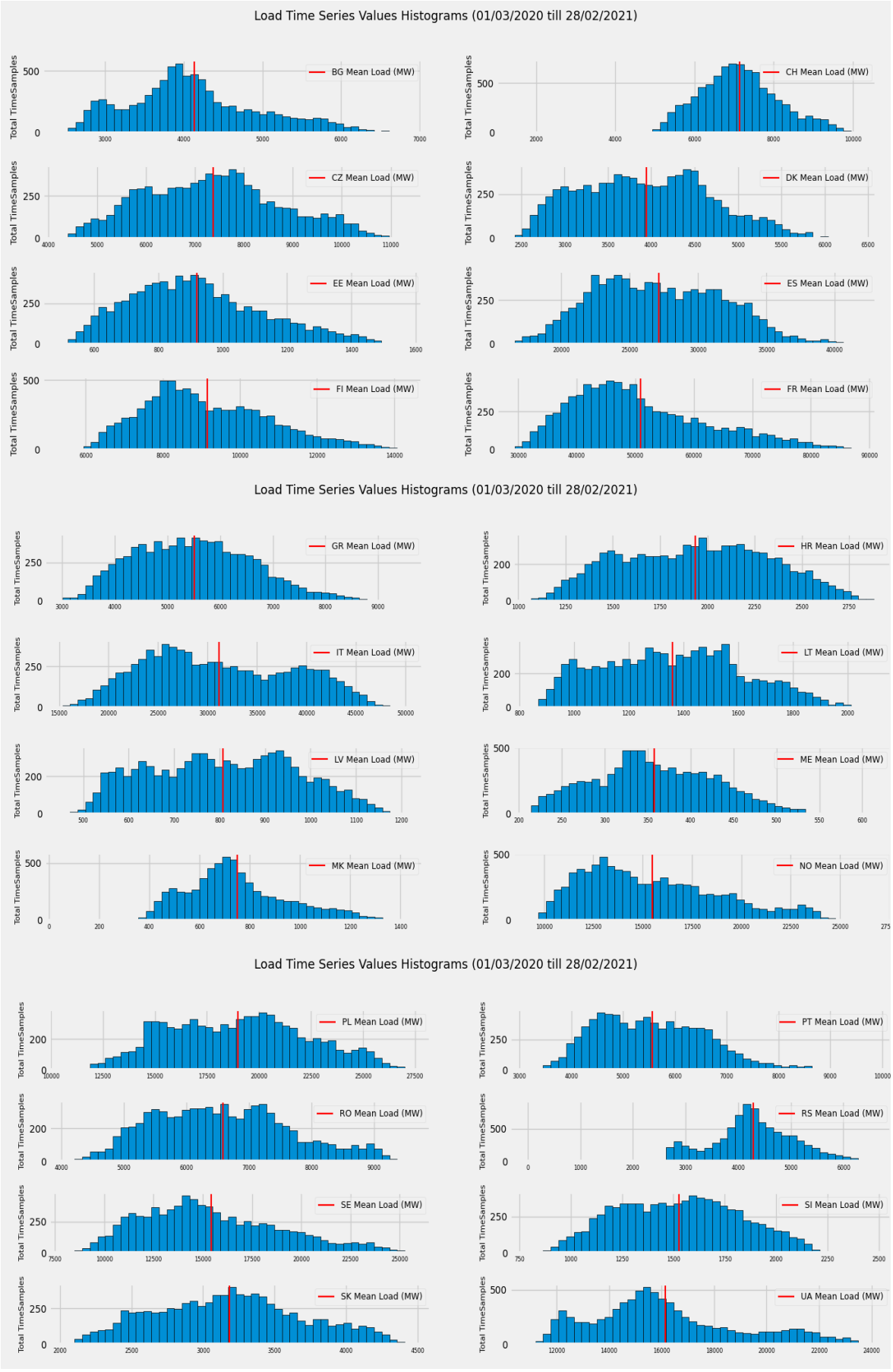
Σχήμα Π.Α.22 : Γραφική παράσταση των ετήσιων χρονοσειρών του "2020".



Σχήμα Π.Α.23 : Ραβδόγραμμα μέσης τιμής και τυπικής απόκλισης των χρονοσειρών του "2020".



Σχήμα Π.Α.24 : Γραφικές παραστάσεις ετήσιων χρονοσειρών του "2020".



Σχήμα Π.Α.25 : Ιστογράμματα ετήσιων χρονοσειρών του "2020".

Παράρτημα Β: Προφίλ Φορτίου Εποχικής Ανάλυσης

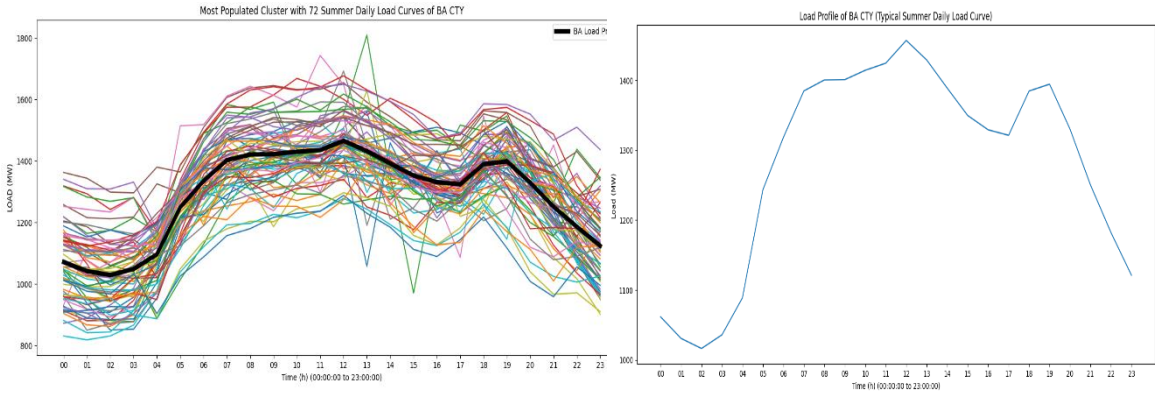
Σε αυτό το παράρτημα παρουσιάζονται τα τελικά "Προφίλ" Φορτίου των ευρωπαϊκών χωρών του συνόλου ανάλυσης τα οποία υπολογίσαμε κατά την εποχική ανάλυση των παρακάτω εποχών :

- **Καλοκαίρι :** ["01/06/2019" εώς "31/08/2019"]
- **Φινόπωρο :** ["01/09/2019" εώς "30/11/2019"]
- **Χειμώνας :** ["01/12/2019" εώς "29/02/2020"]

Τα "Προφίλ"Φορτίου της Άνοιξης έχουν παρουσιαστεί στην ενότητα 4.1 του κεφαλαίου 4.

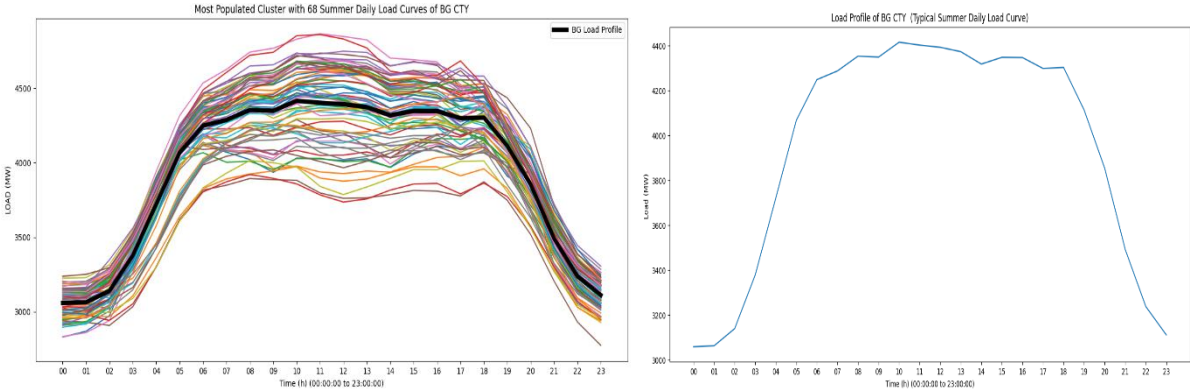
- **Καλοκαίρι ("01/06/2019" εώς "31/08/2019")**

1) **Βοσνία –Ερζεγοβίνη (BA)**



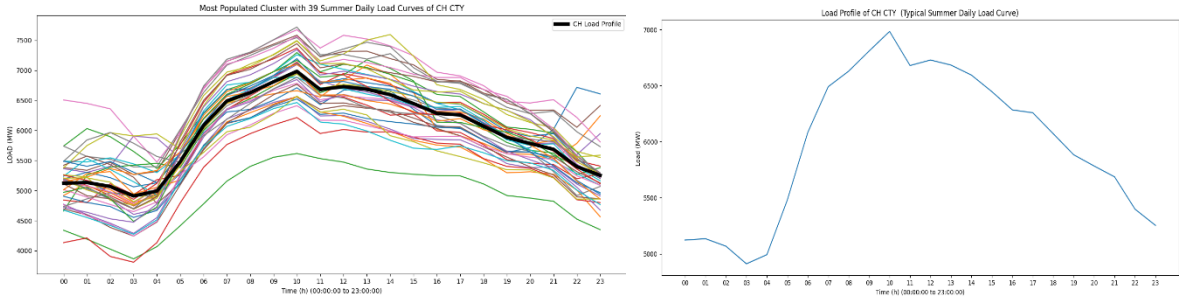
Σχήμα Π.Β.1 : Θερινό Προφίλ Φορτίου (BA).

2) **Βουλγαρία (BG)**



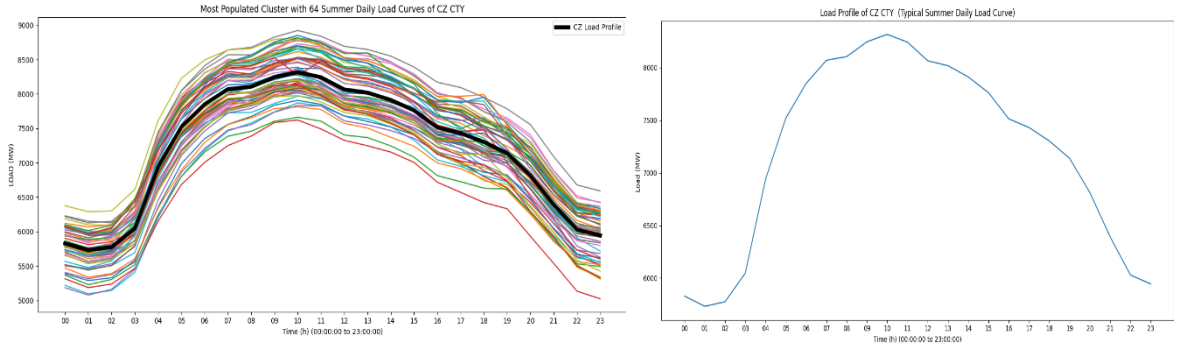
Σχήμα Π.Β.2 : Θερινό Προφίλ Φορτίου (BG).

3) Ελβετία (CH)



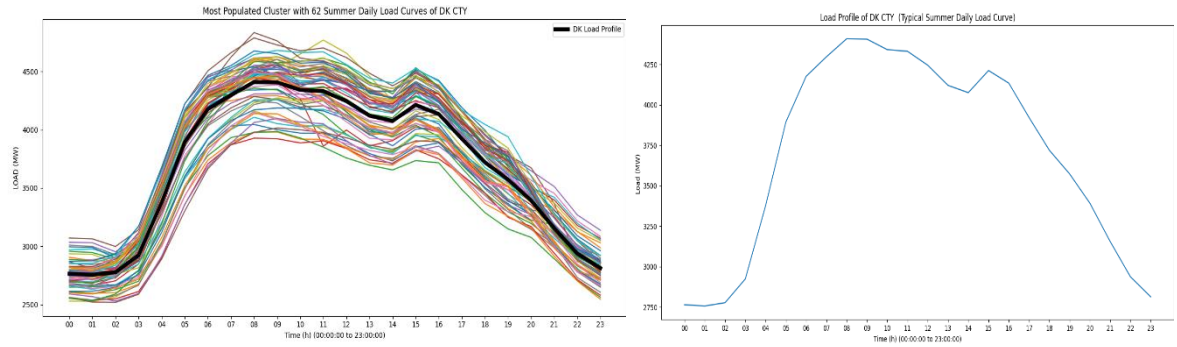
Σχήμα Π.Β.3 : Θερινό Προφίλ Φορτίου (CH).

4) Τσεχία (CZ)



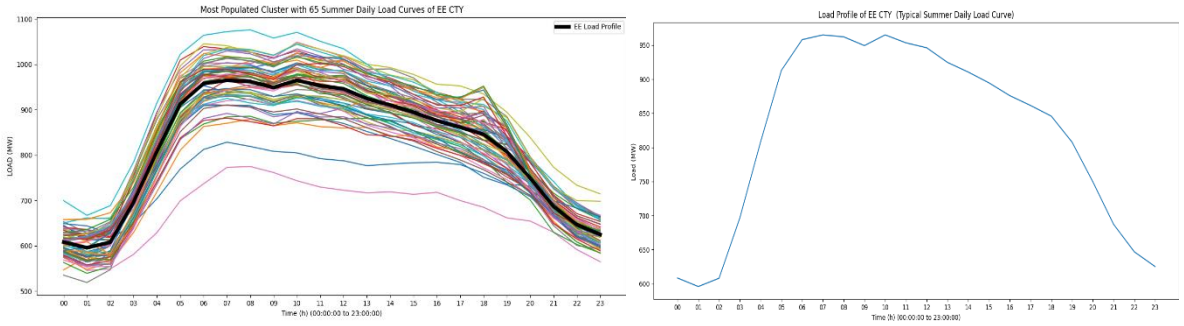
Σχήμα Π.Β.4 : Θερινό Προφίλ Φορτίου (CZ).

5) Δανία (DK)



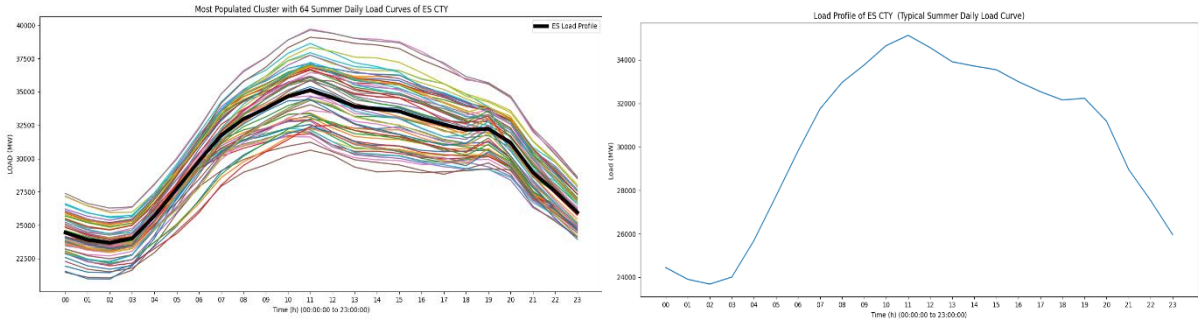
Σχήμα Π.Α.5 : Θερινό Προφίλ Φορτίου (DK).

6) Εσθονία (EE)



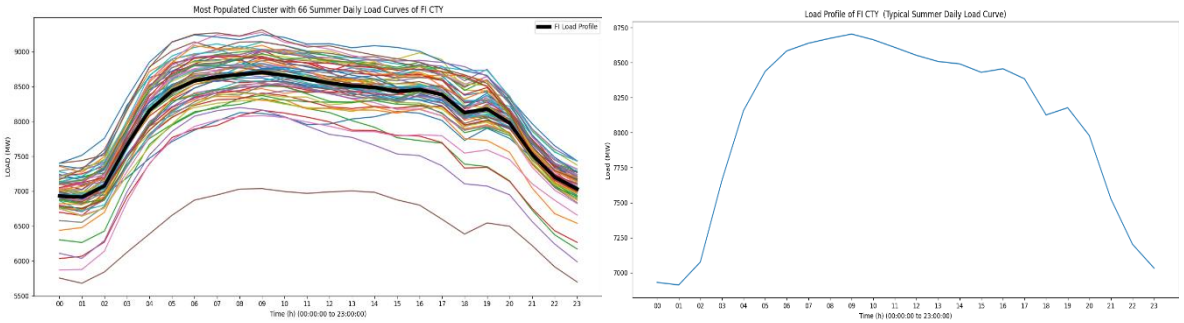
Σχήμα Π.Β.6 : Θερινό Προφίλ Φορτίου (EE).

7) Ισπανία (ES)



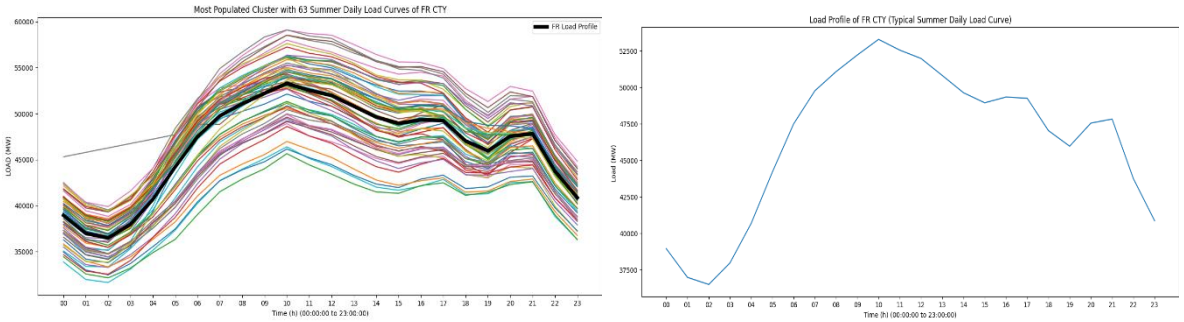
Σχήμα Π.Β.7 : Θερινό Προφίλ Φορτίου (ES).

8) Φινλανδία (FI)



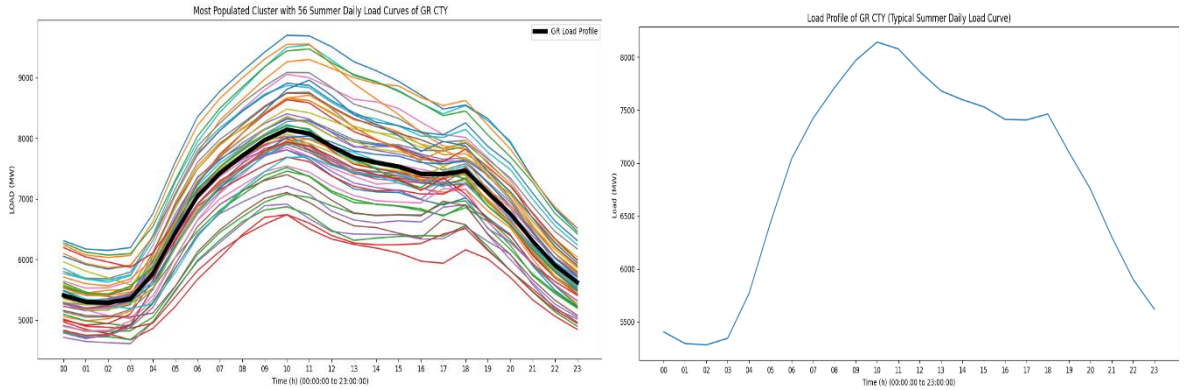
Σχήμα Π.Β.8 : Θερινό Προφίλ Φορτίου (FI).

9) Γαλλία (FR)



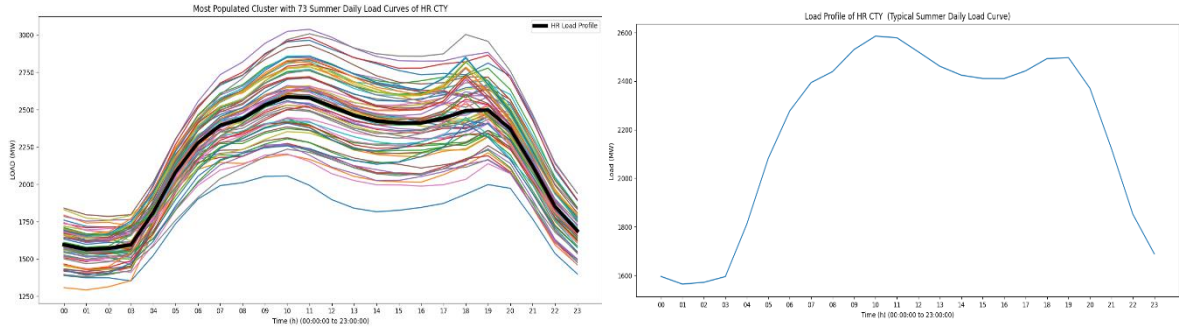
Σχήμα Π.Β.9 : Θερινό Προφίλ Φορτίου (FR).

10) Ελλάδα (GR)



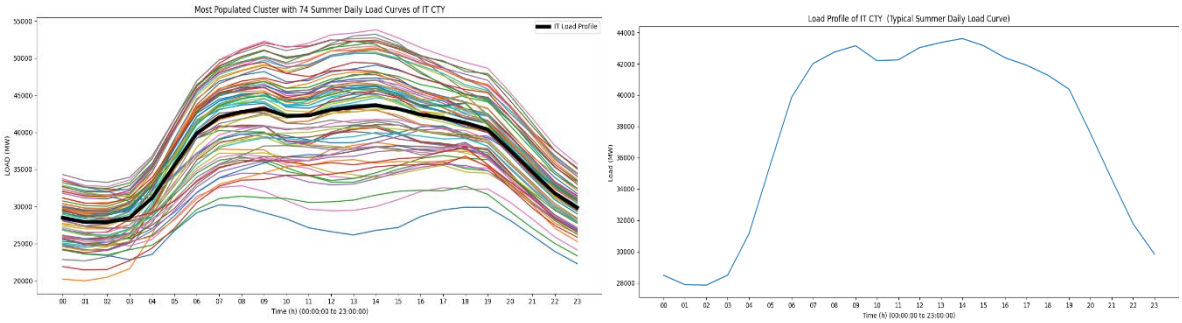
Σχήμα Π.Β.10 : Θερινό Προφίλ Φορτίου (GR).

11) Κροατία (HR)



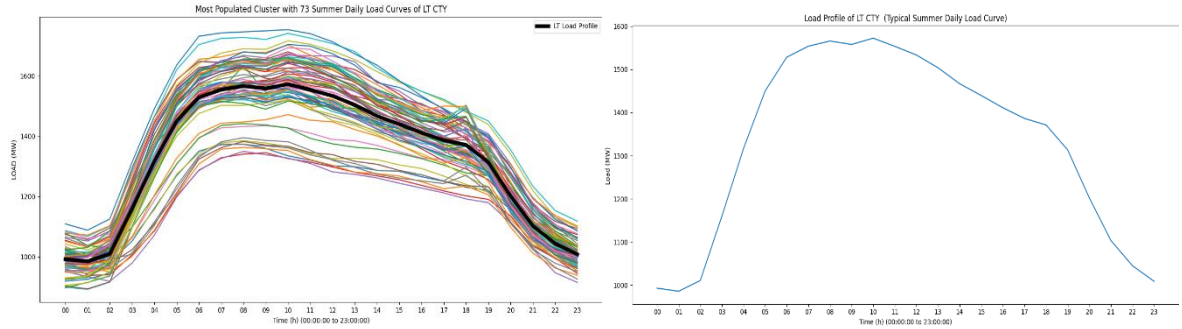
Σχήμα Π.Β.11 : Θερινό Προφίλ Φορτίου (HR).

12) Ιταλία (IT)



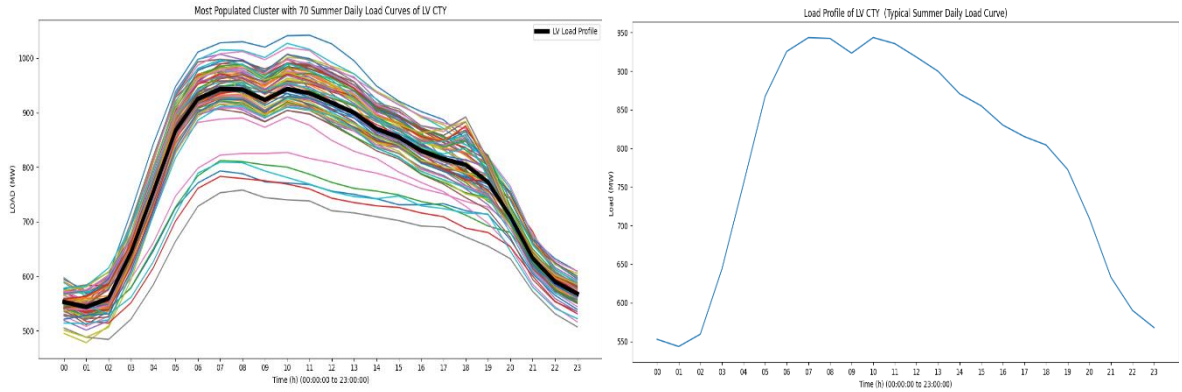
Σχήμα Π.Β.12 : Θερινό Προφίλ Φορτίου (IT).

13) Λιθουανία (LT)



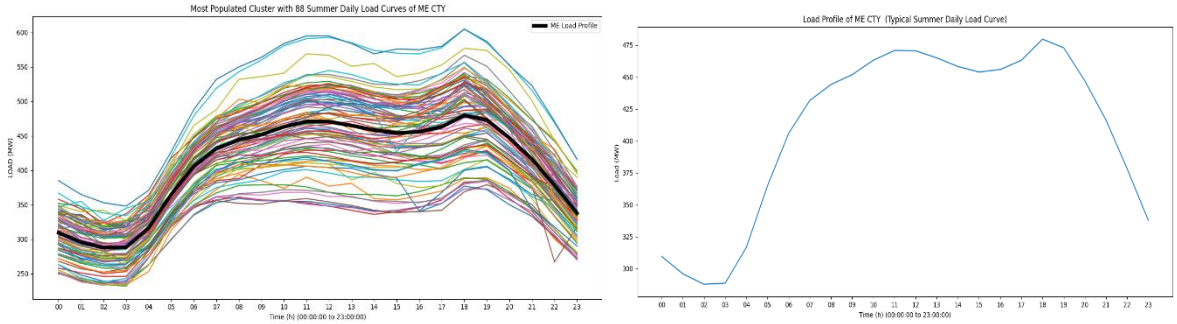
Σχήμα Π.Β.13 : Θερινό Προφίλ Φορτίου (LT).

14) Λετονία (LV)



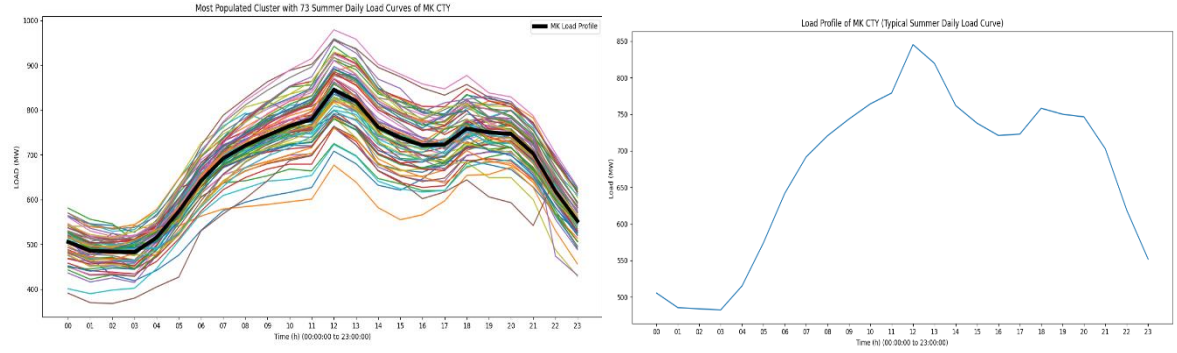
Σχήμα Π.Β.14 : Θερινό Προφίλ Φορτίου (LV).

15) Μαυροβούνιο (ΜΕ)



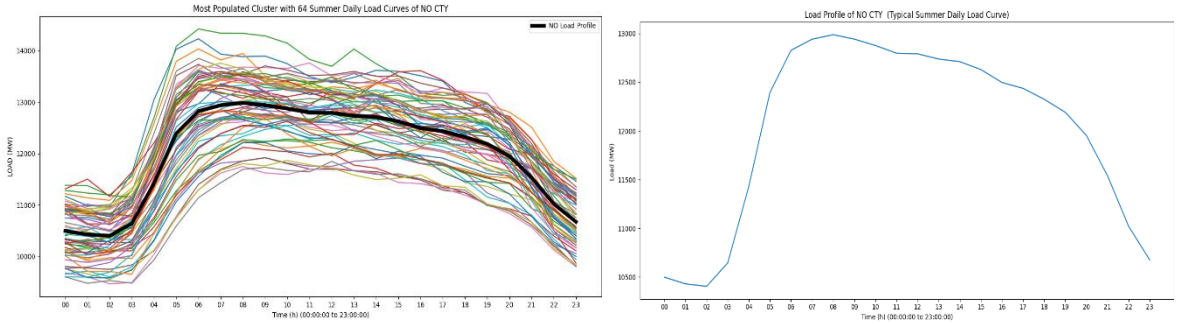
Σχήμα Π.Β.15 : Θερινό Προφίλ Φορτίου (ΜΕ).

16) Βόρεια Μακεδονία (ΜΚ)



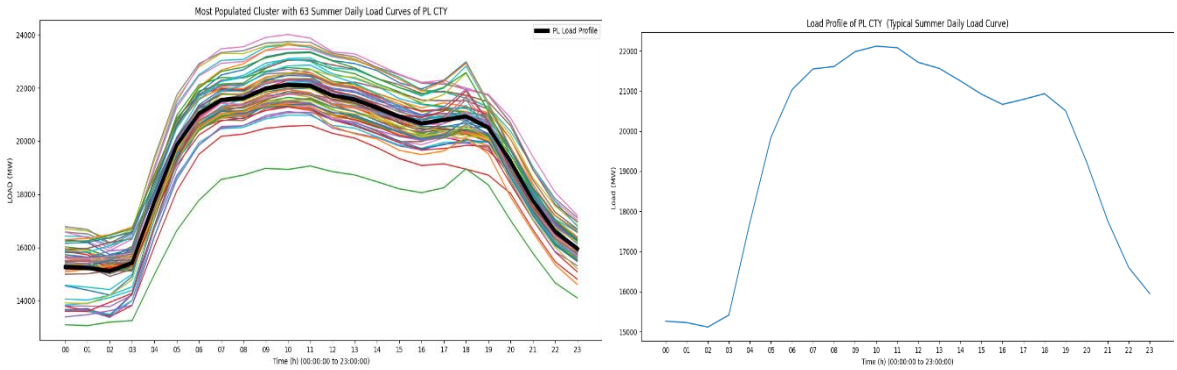
Σχήμα Π.Β.16 : Θερινό Προφίλ Φορτίου (ΜΚ).

17) Νορβηγία (NO)



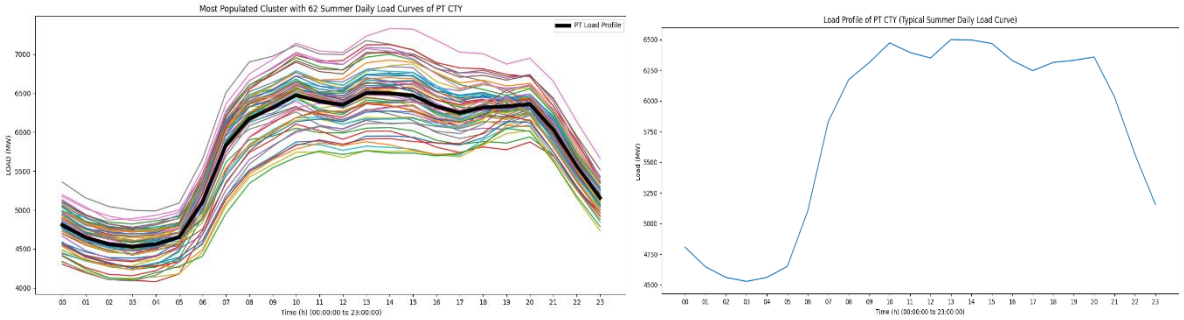
Σχήμα Π.Β.17 : Θερινό Προφίλ Φορτίου (NO).

18) Πολωνία (PL)



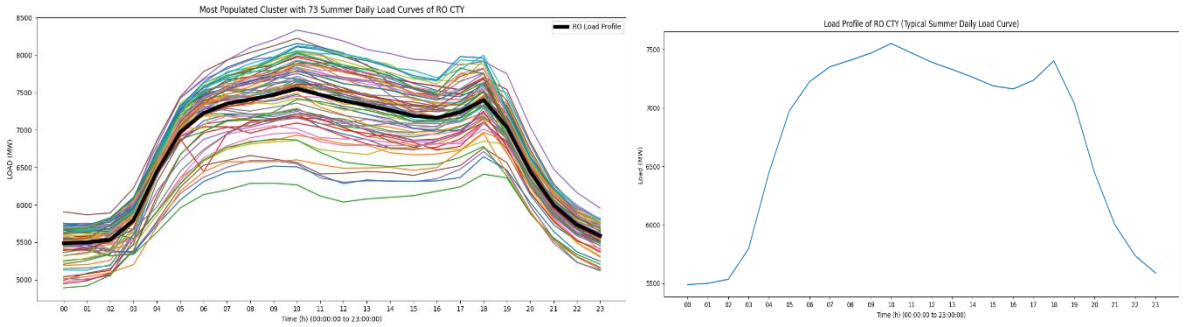
Σχήμα Π.Β.18 : Θερινό Προφίλ Φορτίου (PL).

19) Πορτογαλία (PT)



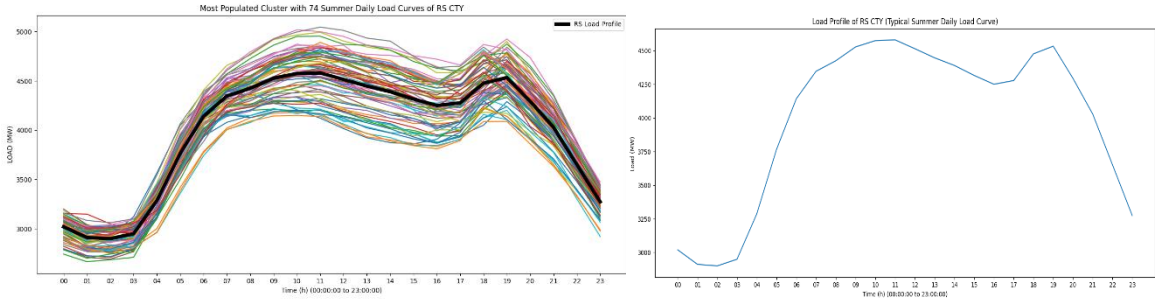
Σχήμα Π.Β.19 : Θερινό Προφίλ Φορτίου (PT).

20) Ρουμανία (RO)



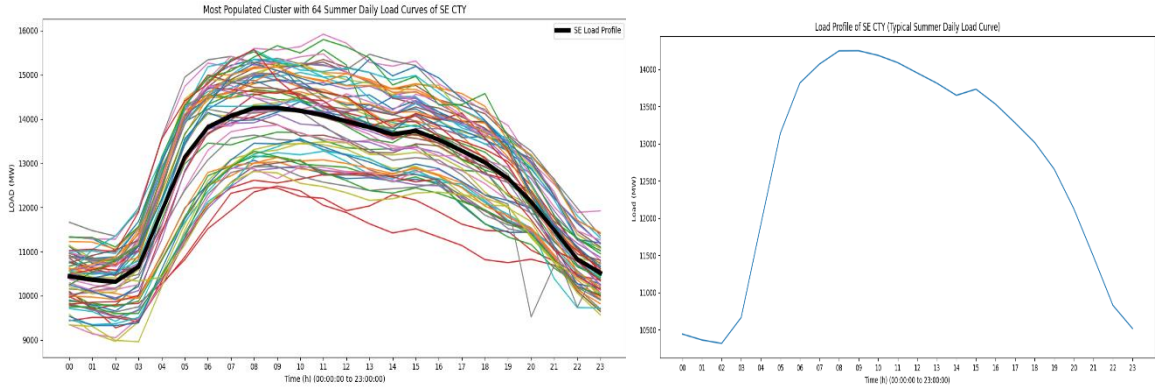
Σχήμα Π.Β.20 : Θερινό Προφίλ Φορτίου (RO).

21) Σερβία (RS)



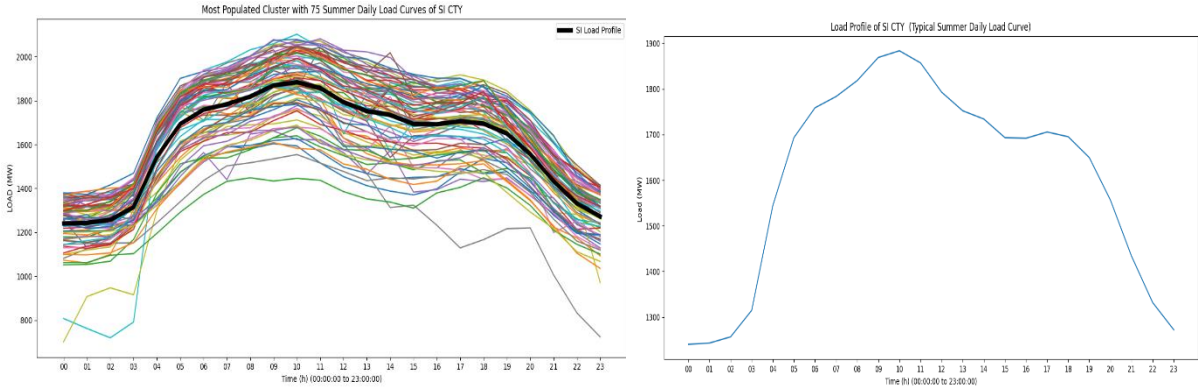
Σχήμα Π.Β.21 : Θερινό Προφίλ Φορτίου (RS).

22) Σουηδία (SE)



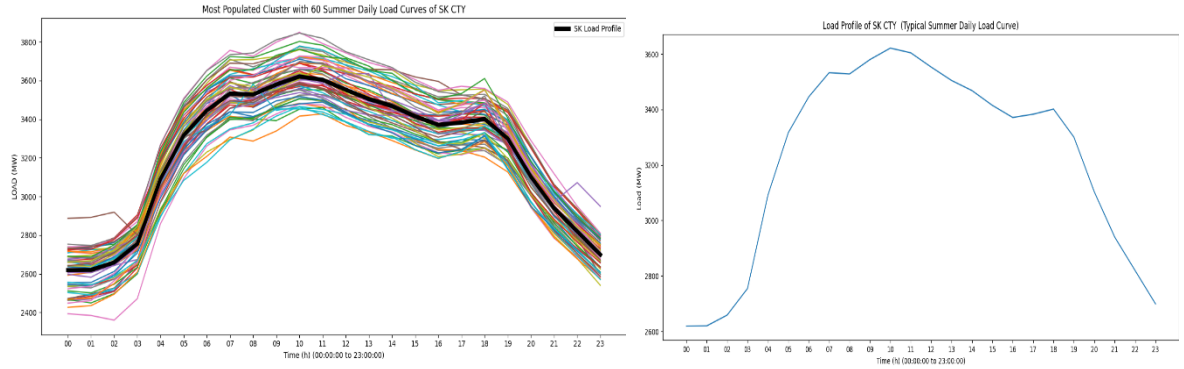
Σχήμα Π.Β.22 : Θερινό Προφίλ Φορτίου (SE).

23) Σλοβενία (SI)



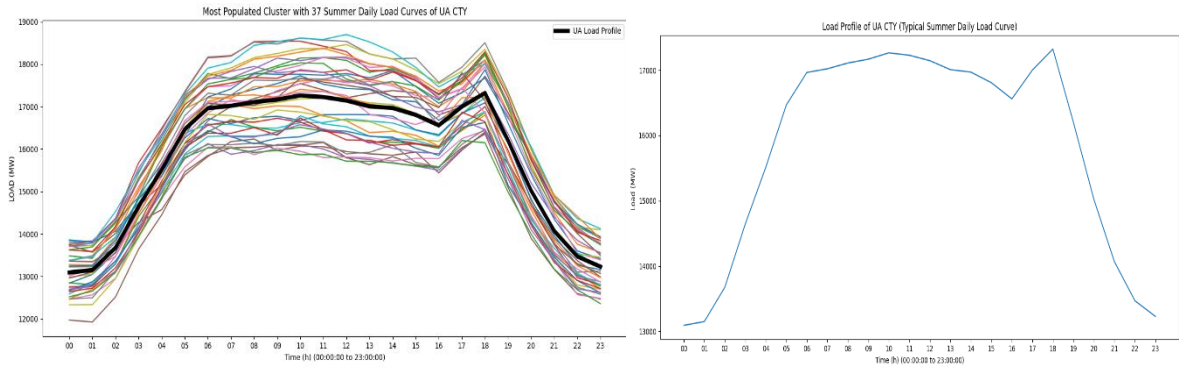
Σχήμα Π.Β.23 : Θερινό Προφίλ Φορτίου (SI).

24) Σλοβακία (SK)



Σχήμα Π.Β.24 : Θερινό Προφίλ Φορτίου (SK).

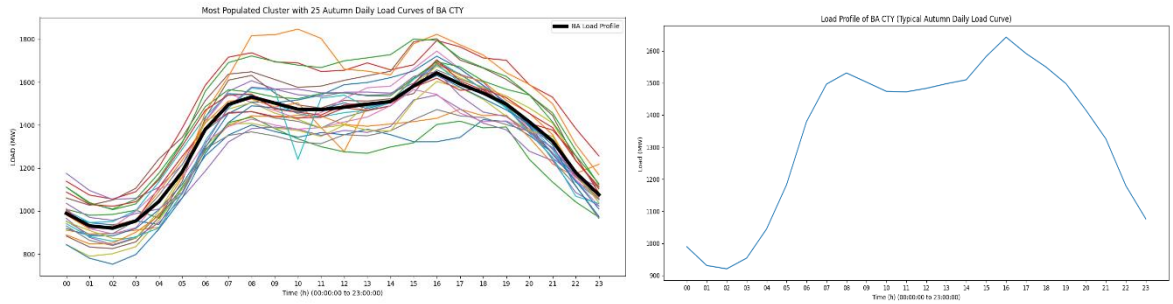
25) Ουκρανία (UA)



Σχήμα Π.Β.25 : Θερινό Προφίλ Φορτίου (UA).

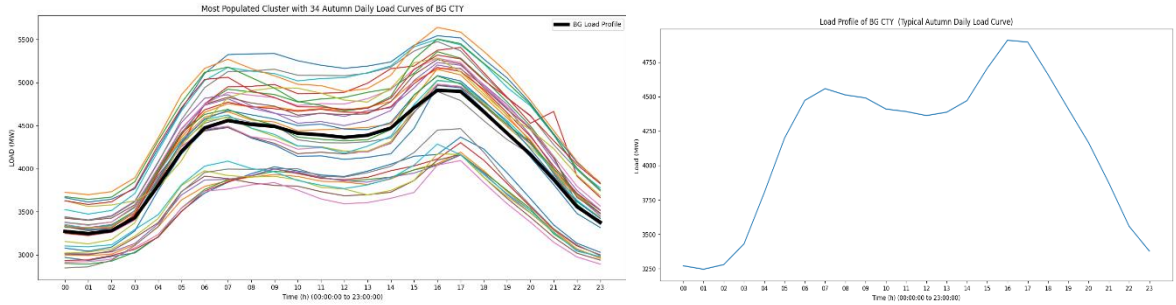
- **Φθινόπωρο ("01/09/2019" εώς "30/11/2019")**

1) Βοσνία – Ερζεγοβίνη (BA)



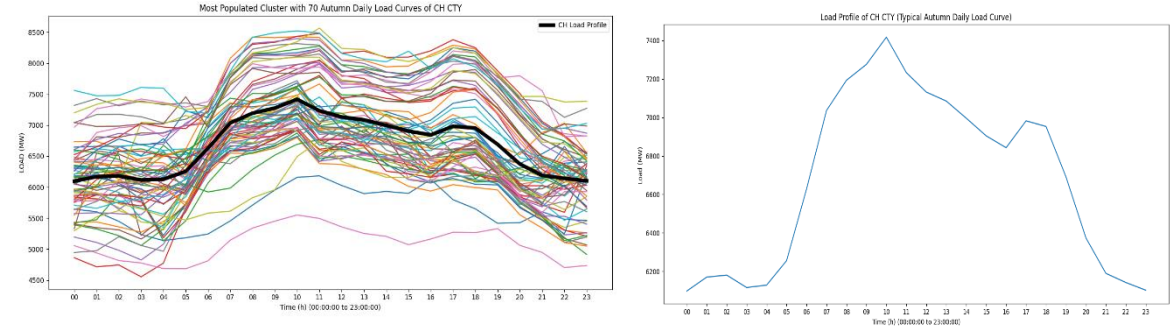
Σχήμα Π.Β.26 : Φθινοπωρινό Προφίλ Φορτίου (BA).

2) Βουλγαρία (BG)



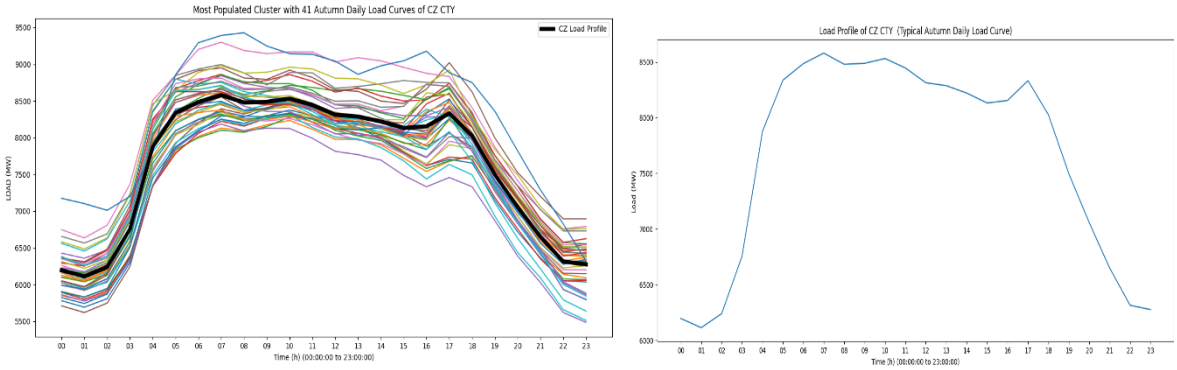
Σχήμα Π.Β.27 : Φθινοπωρινό Προφίλ Φορτίου (BG).

3) Ελβετία (CH)



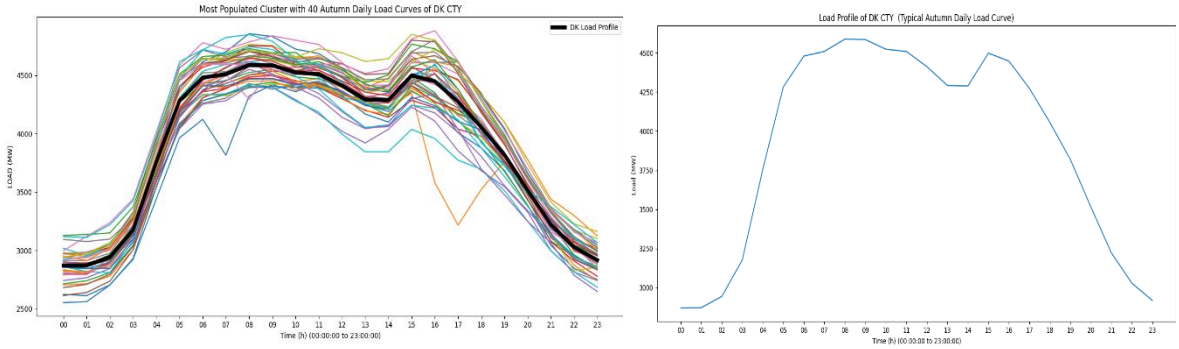
Σχήμα Π.Β.28 : Φθινοπωρινό Προφίλ Φορτίου (CH).

4) Τσεχία (CZ)



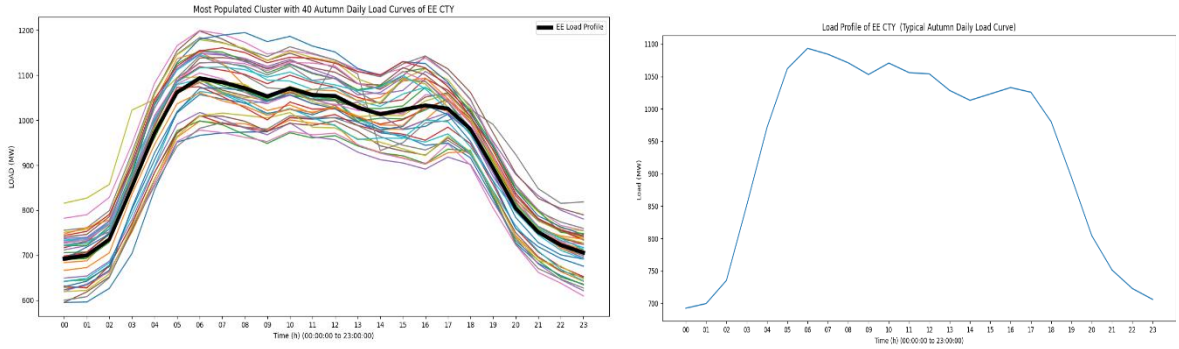
Σχήμα Π.Β.29 : Φθινοπωρινό Προφίλ Φορτίου (CZ).

5) Δανία (DK)



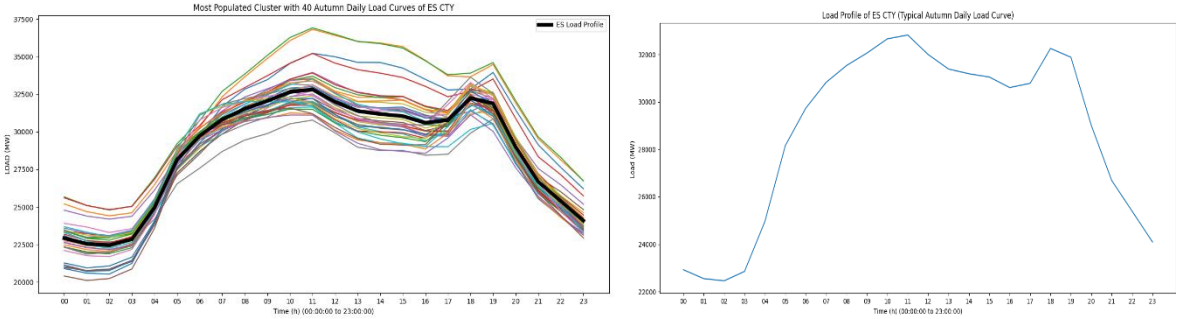
Σχήμα Π.Β.30 : Φθινοπωρινό Προφίλ Φορτίου (DK).

6) Εσθονία (EE)



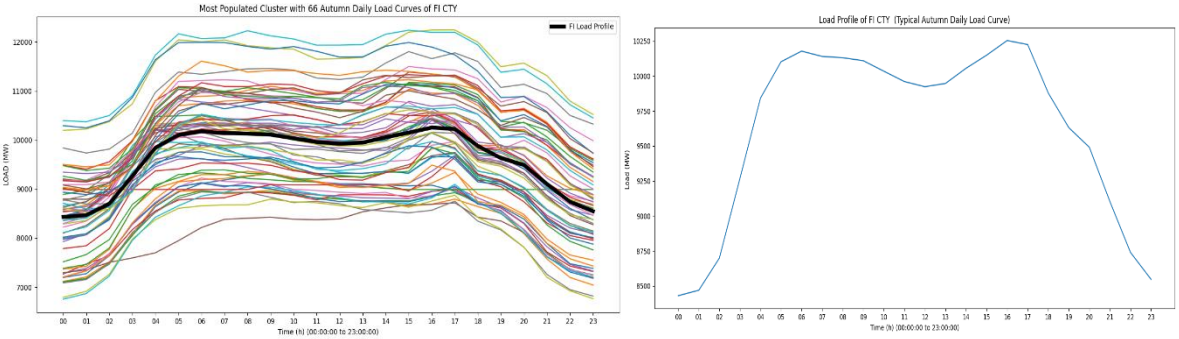
Σχήμα Π.Β.31 : Φθινοπωρινό Προφίλ Φορτίου (EE).

7) Ισπανία (ES)



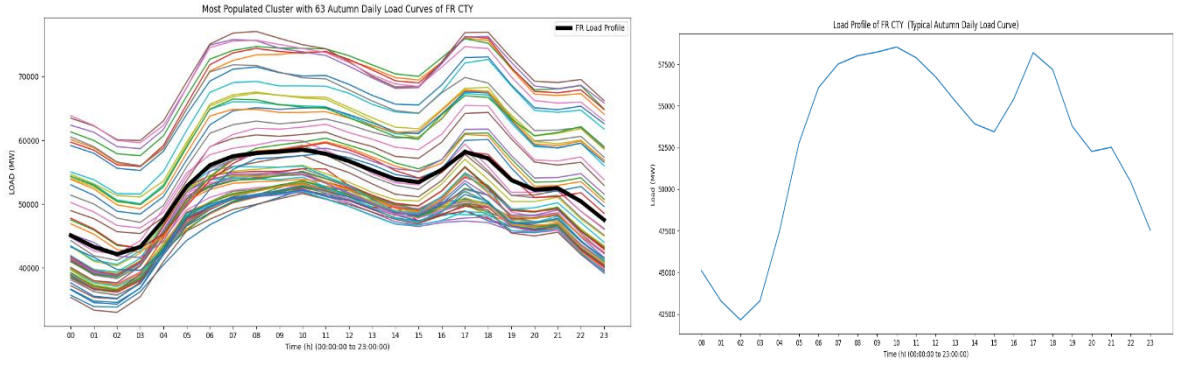
Σχήμα Π.Β.32 : Φθινοπωρινό Προφίλ Φορτίου (ES).

8) Φινλανδία (FI)



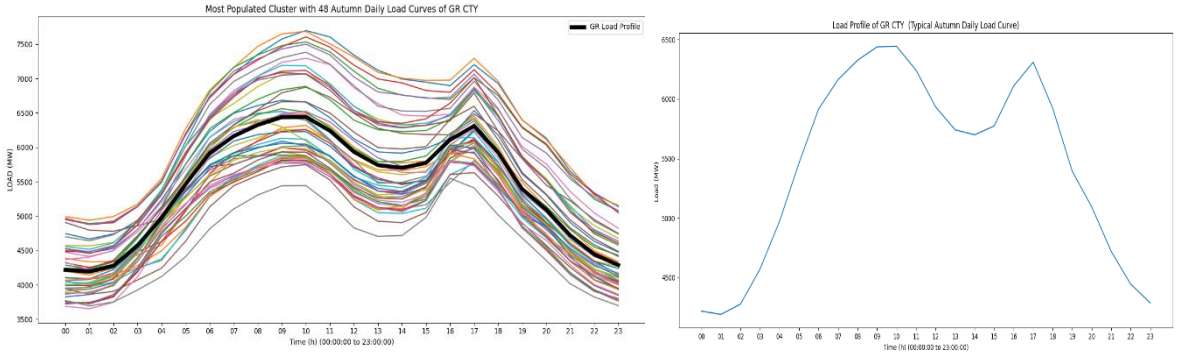
Σχήμα Π.Β.33 : Φθινοπωρινό Προφίλ Φορτίου (FI).

9) Γαλλία (FR)



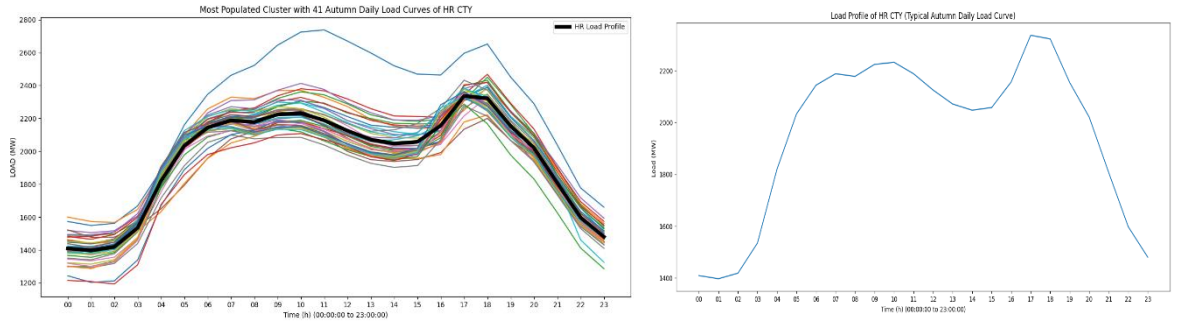
Σχήμα Π.Β.34 : Φθινοπωρινό Προφίλ Φορτίου (FR).

10) Ελλάδα (GR)



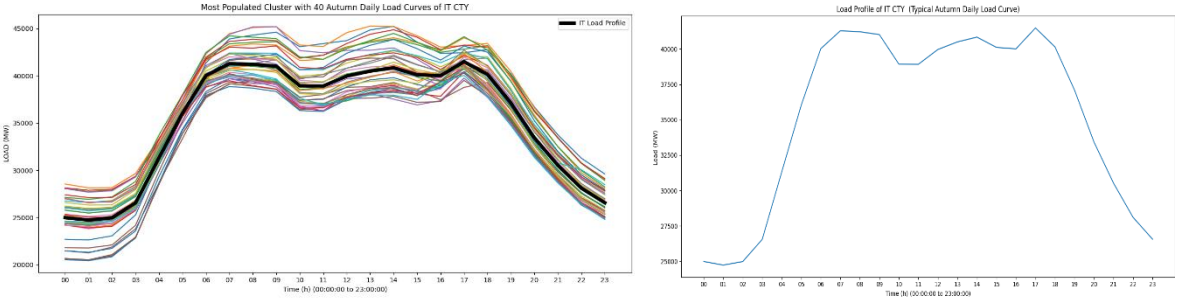
Σχήμα Π.Β.35 : Φθινοπωρινό Προφίλ Φορτίου (GR).

11) Κροατία (HR)



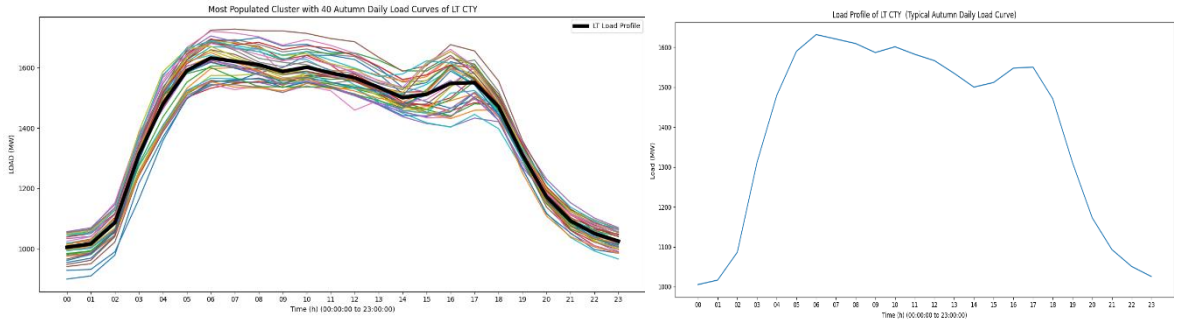
Σχήμα Π.Β.36 : Φθινοπωρινό Προφίλ Φορτίου (HR).

12) Ιταλία (IT)



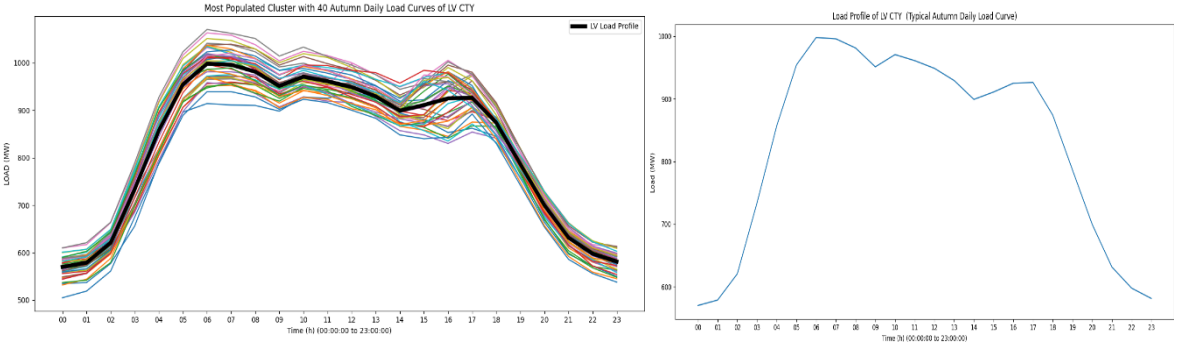
Σχήμα Π.Β.37 : Φθινοπωρινό Προφίλ Φορτίου (IT).

13) Λιθουανία (LT)



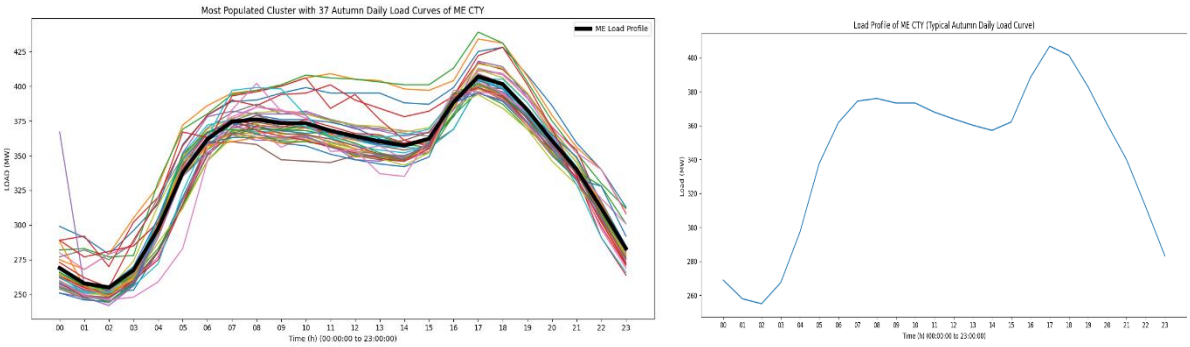
Σχήμα Π.Β.38 : Φθινοπωρινό Προφίλ Φορτίου (LT).

14) Λετονία (LV)



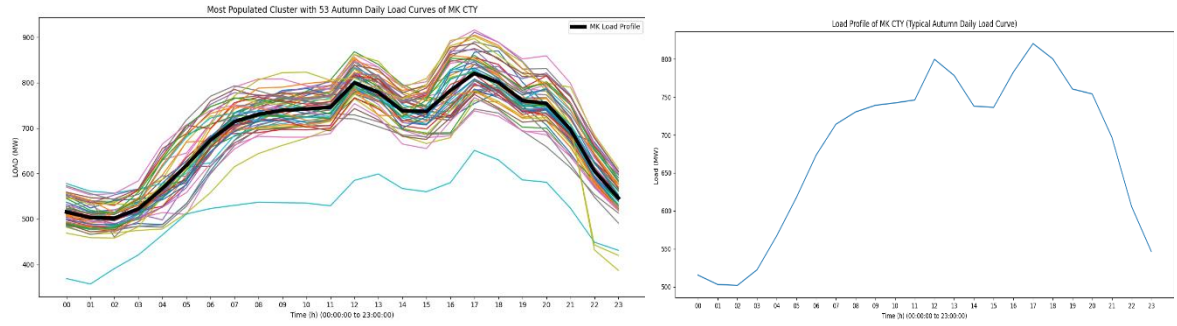
Σχήμα Π.Β.39 : Φθινοπωρινό Προφίλ Φορτίου (LV).

15) Μαυροβούνιο (ME)



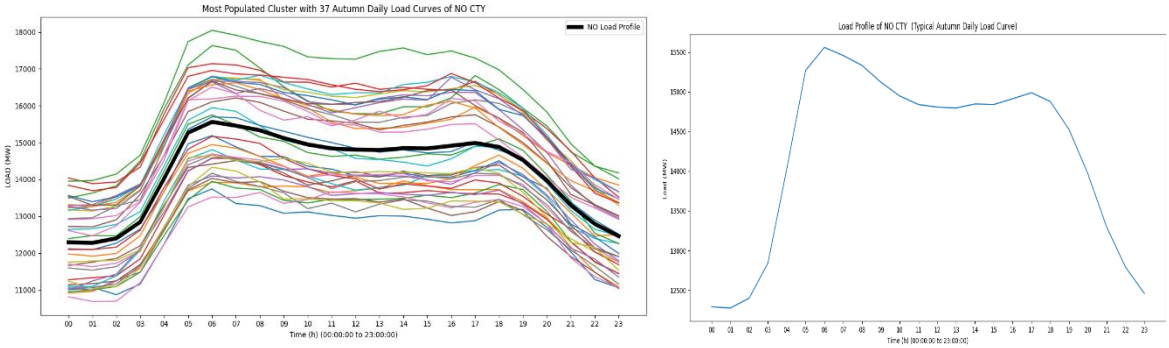
Σχήμα Π.Β.40 : Φθινοπωρινό Προφίλ Φορτίου (ME).

16) Βόρεια Μακεδονία (MK)



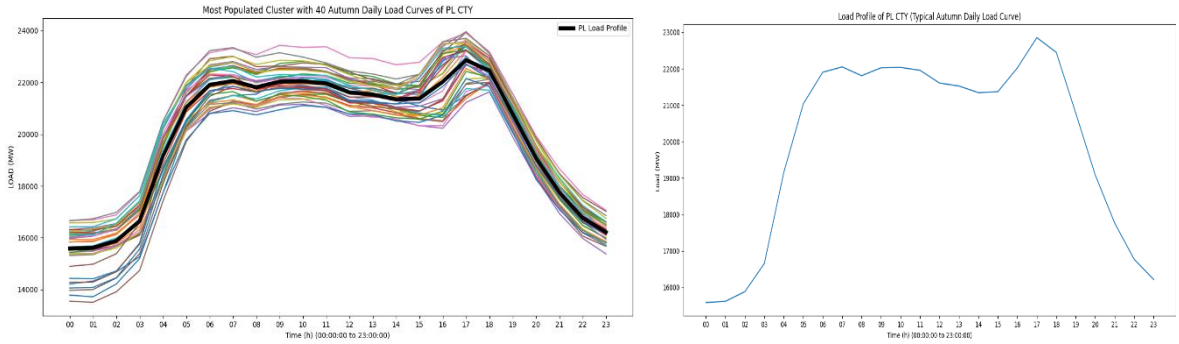
Σχήμα Π.Β.41 : Φθινοπωρινό Προφίλ Φορτίου (MK).

17) Νορβηγία (NO)



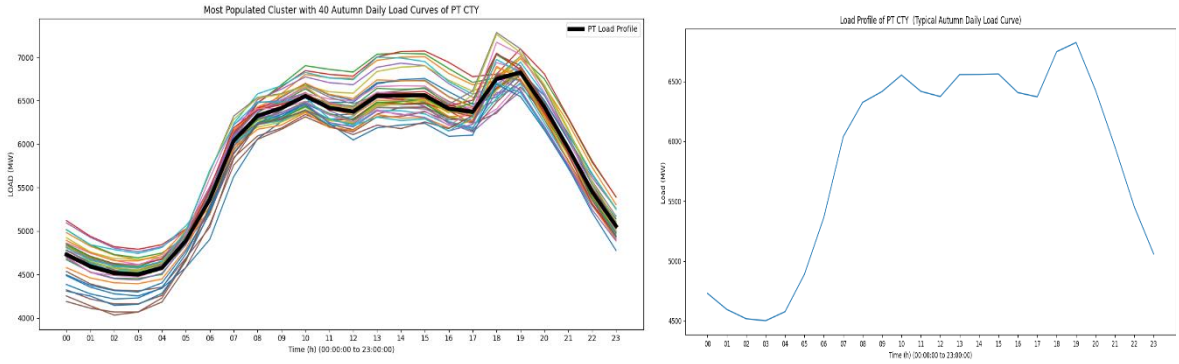
Σχήμα Π.Β.42 : Φθινοπωρινό Προφίλ Φορτίου (NO).

18) Πολωνία (PL)



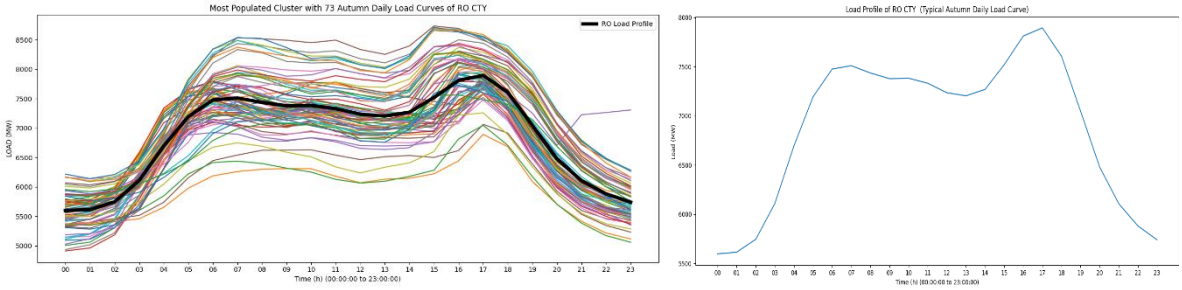
Σχήμα Π.Β.43 : Φθινοπωρινό Προφίλ Φορτίου (PL).

19) Πορτογαλία (PT)



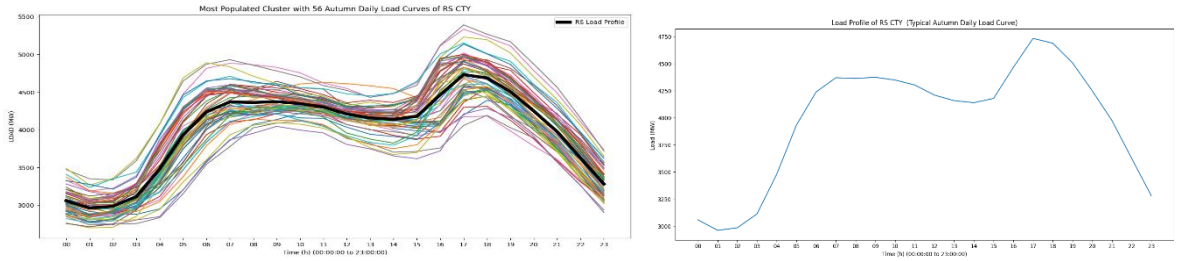
Σχήμα Π.Β.44 : Φθινοπωρινό Προφίλ Φορτίου (PT).

20) Ρουμανία (RO)



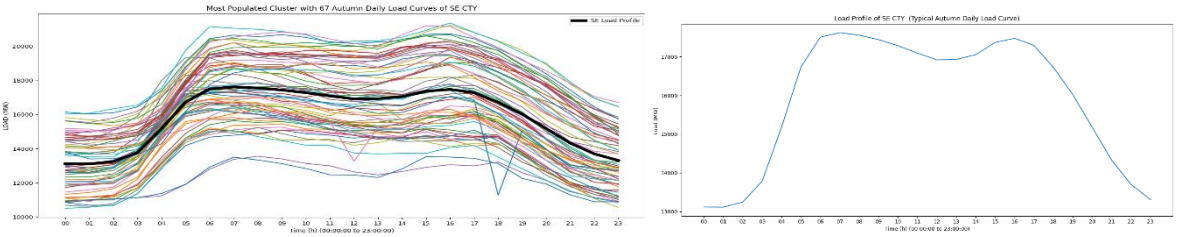
Σχήμα Π.Β.45 : Φθινοπωρινό Προφίλ Φορτίου (RO).

21) Σερβία (RS)



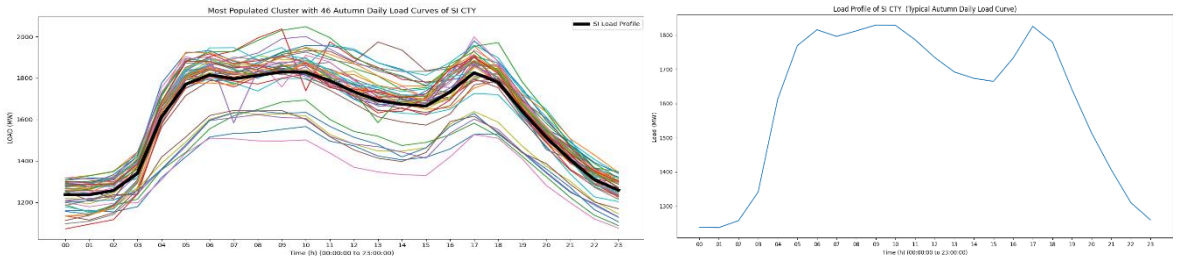
Σχήμα Π.Β.46 : Φθινοπωρινό Προφίλ Φορτίου (RS).

22) Σουηδία (SE)



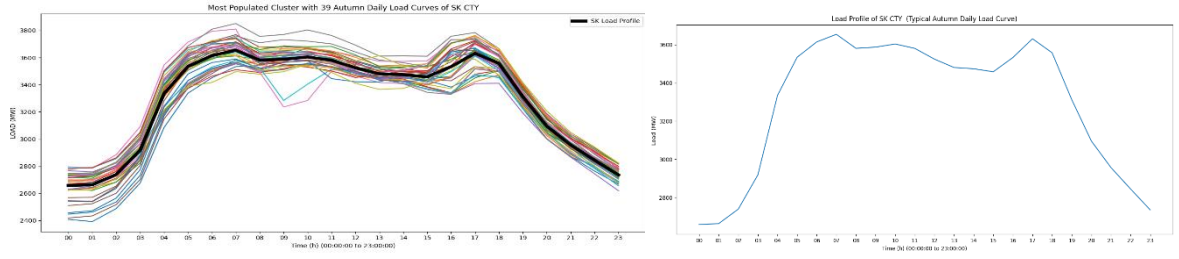
Σχήμα Π.Β.47 : Φθινοπωρινό Προφίλ Φορτίου (SE).

23) Σλοβενία (SI)



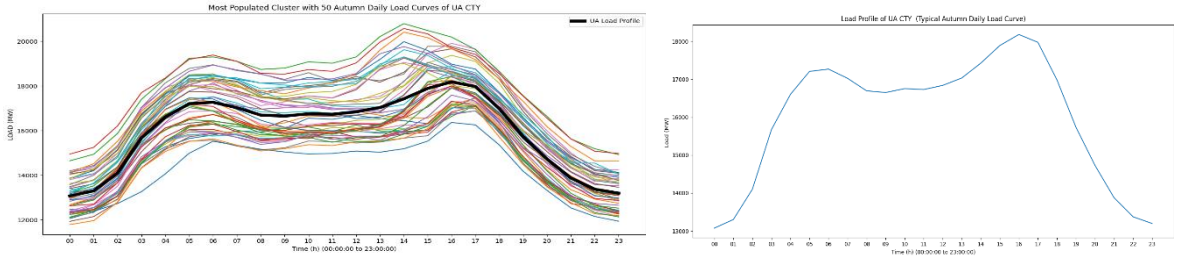
Σχήμα Π.Β.48 : Φθινοπωρινό Προφίλ Φορτίου (SI).

24) Σλοβακία (SK)



Σχήμα Π.Β.49 : Φθινοπωρινό Προφίλ Φορτίου (SK).

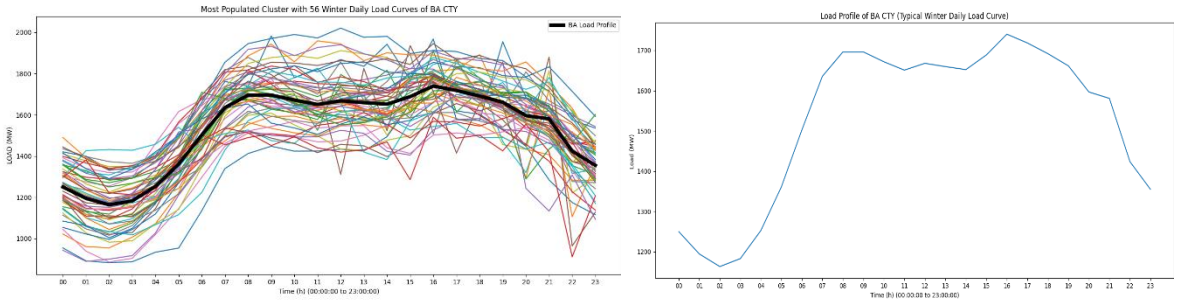
25) Ουκρανία (UA)



Σχήμα Π.Β.50 : Φθινοπωρινό Προφίλ Φορτίου (UA).

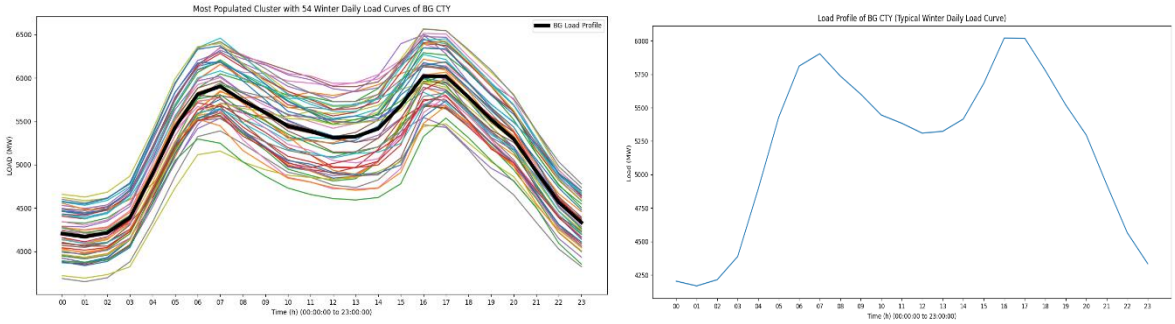
- **Χειμώνας ("01/12/2019" εώς "29/02/2020")**

1) Βοσνία – Ερζεγοβίνη (BA)



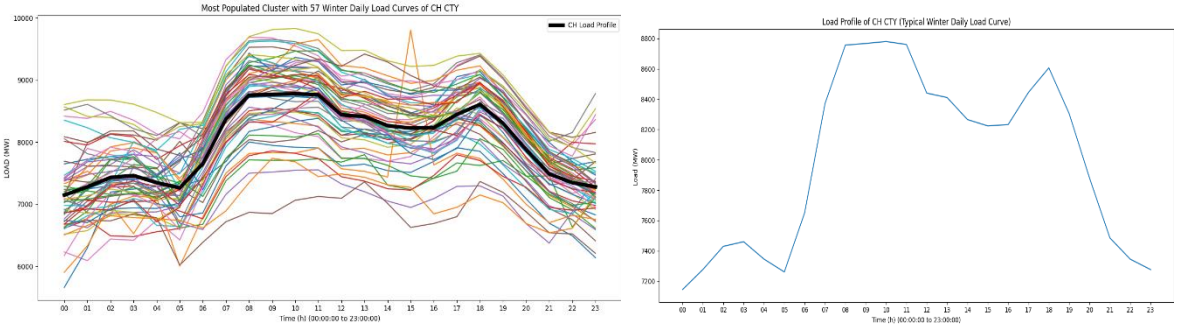
Σχήμα Π.Β.51 : Χειμερινό Προφίλ Φορτίου (BA).

2) Βουλγαρία (BG)



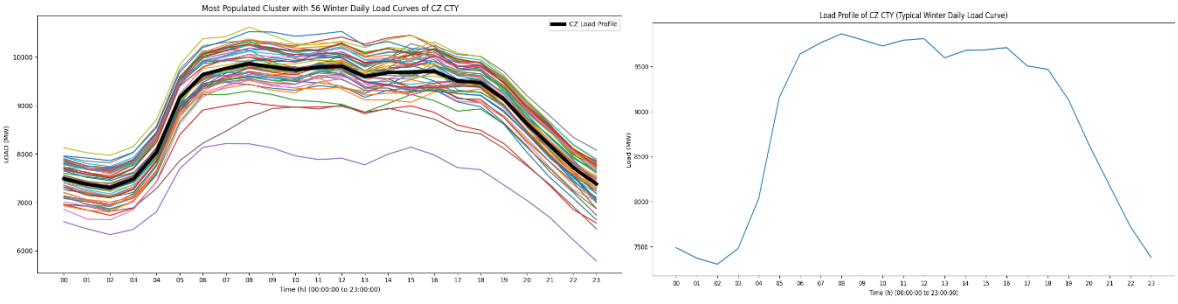
Σχήμα Π.Β.52 : Χειμερινό Προφίλ Φορτίου (BG).

3) Ελβετία (CH)



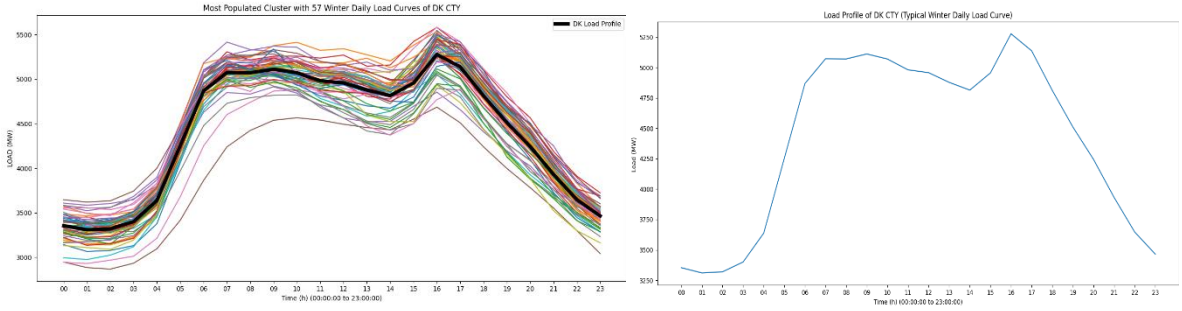
Σχήμα Π.Β.53 : Χειμερινό Προφίλ Φορτίου (CH).

4) Τσεχία (CZ)



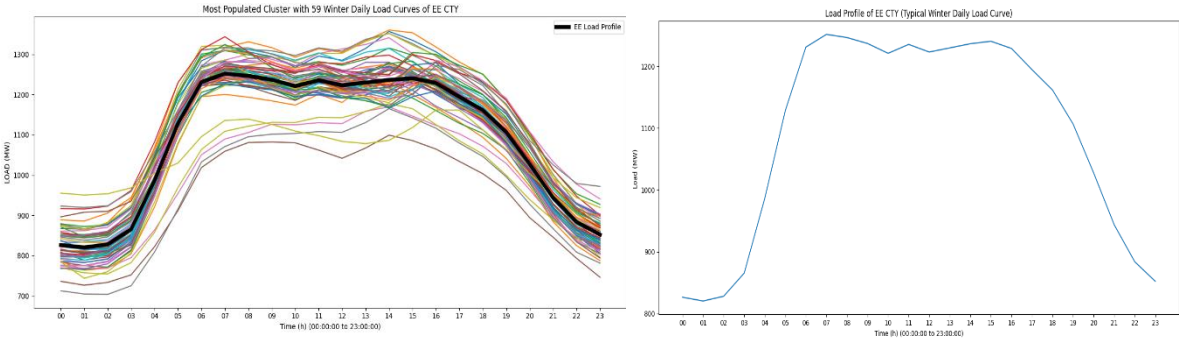
Σχήμα Π.Β.54 : Χειμερινό Προφίλ Φορτίου (CZ).

5) Δανία (DK)



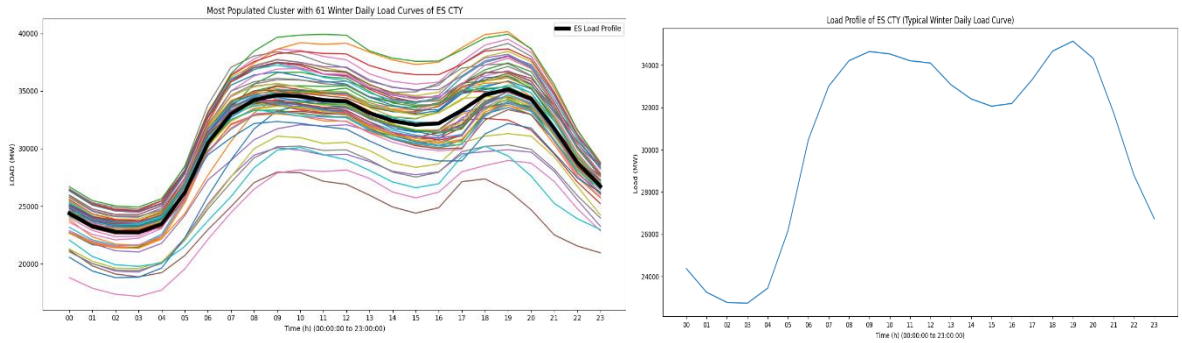
Σχήμα Π.Β.55 : Χειμερινό Προφίλ Φορτίου (DK).

6) Εσθονία (EE)



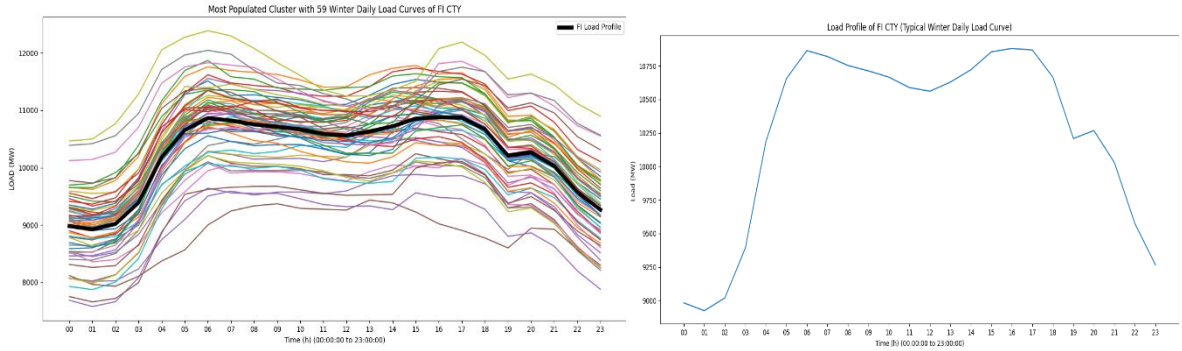
Σχήμα Π.Β.56 : Χειμερινό Προφίλ Φορτίου (EE).

7) Ισπανία (ES)



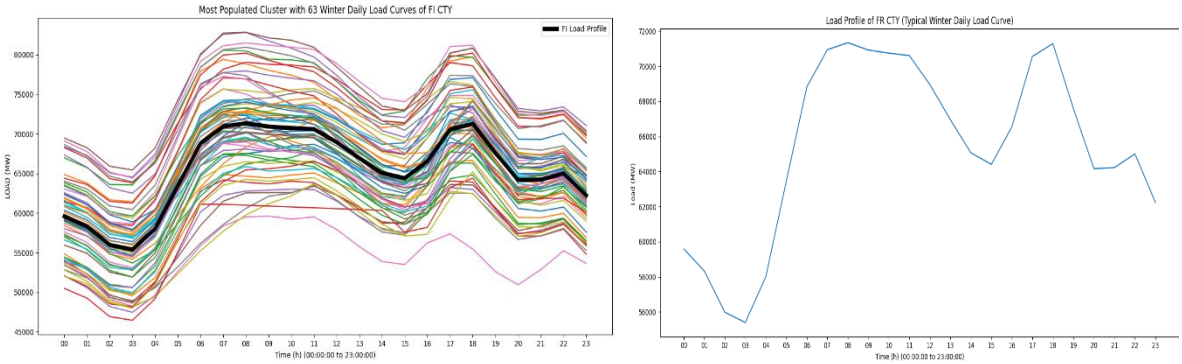
Σχήμα Π.Β.57 : Χειμερινό Προφίλ Φορτίου (ES).

8) Φινλανδία (FI)



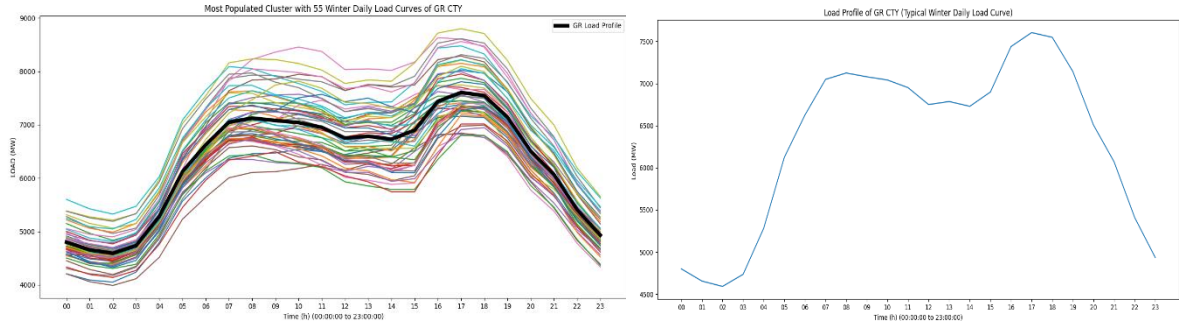
Σχήμα Π.Β.58 : Χειμερινό Προφίλ Φορτίου (FI).

9) Γαλλία (FR)



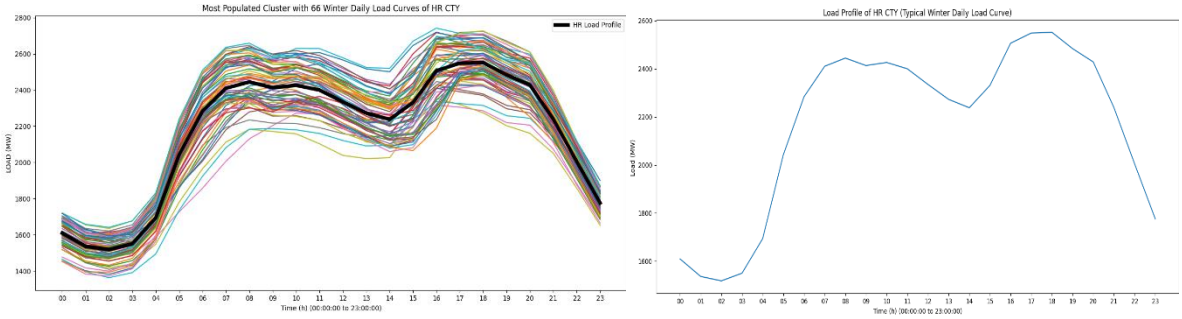
Σχήμα Π.Β.59 : Χειμερινό Προφίλ Φορτίου (FR).

10) Ελλάδα (GR)



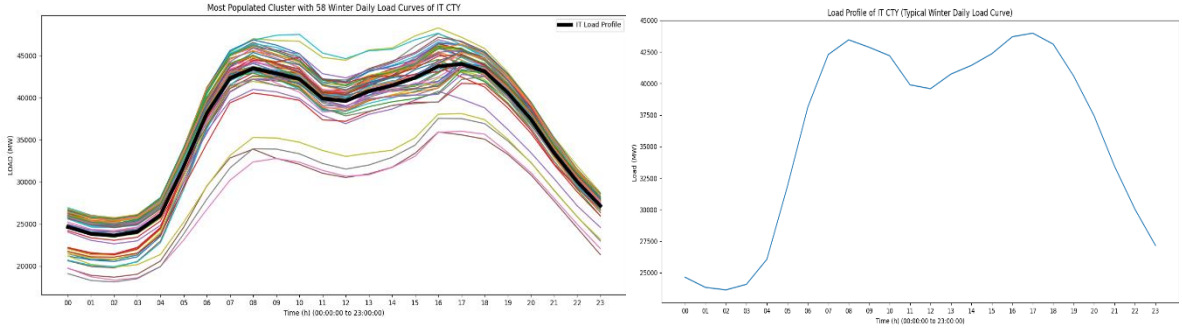
Σχήμα Π.Β.60 : Χειμερινό Προφίλ Φορτίου (GR).

11) Κροατία (HR)



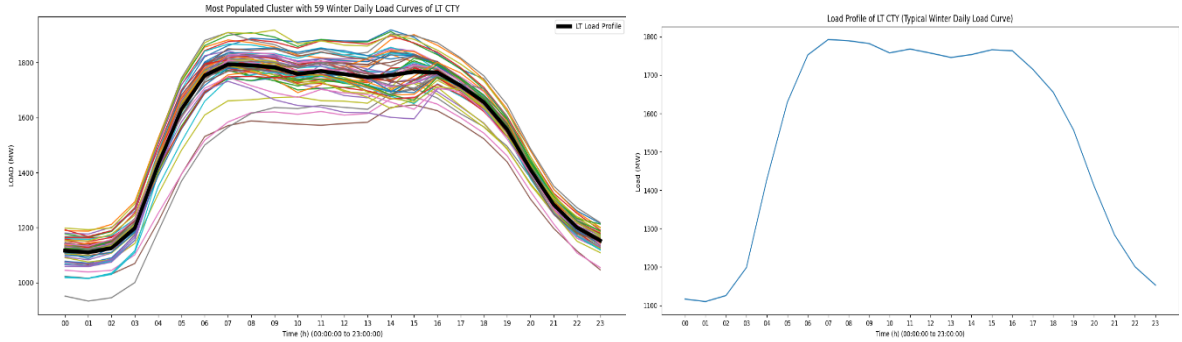
Σχήμα Π.Β.61 : Χειμερινό Προφίλ Φορτίου (HR).

12) Ιταλία (IT)



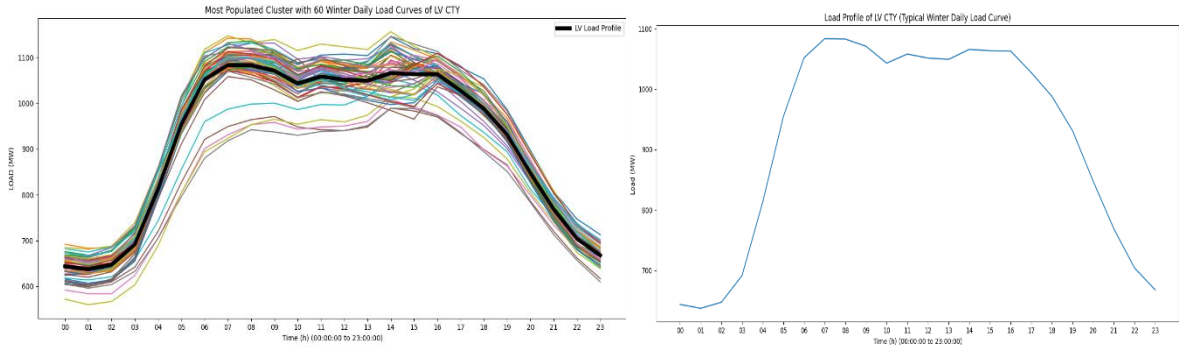
Σχήμα Π.Β.62 : Χειμερινό Προφίλ Φορτίου (IT).

13) Λιθουανία (LT)



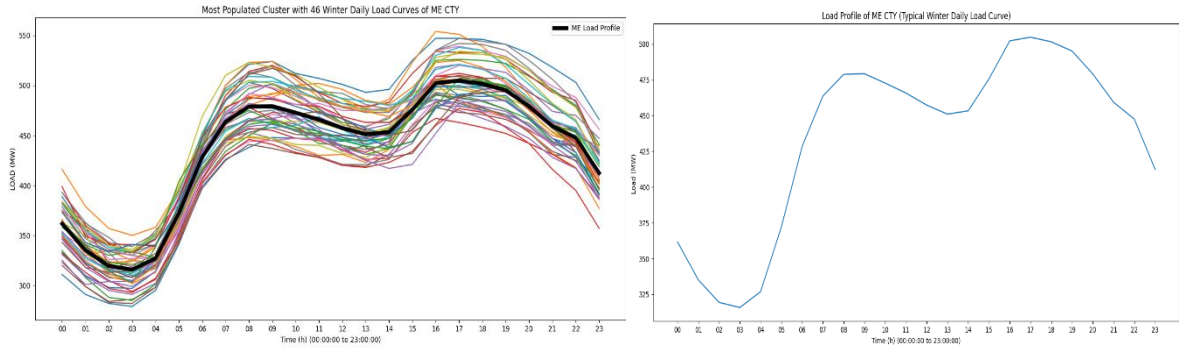
Σχήμα Π.Β.63 : Χειμερινό Προφίλ Φορτίου (LT).

14) Λετονία (LV)



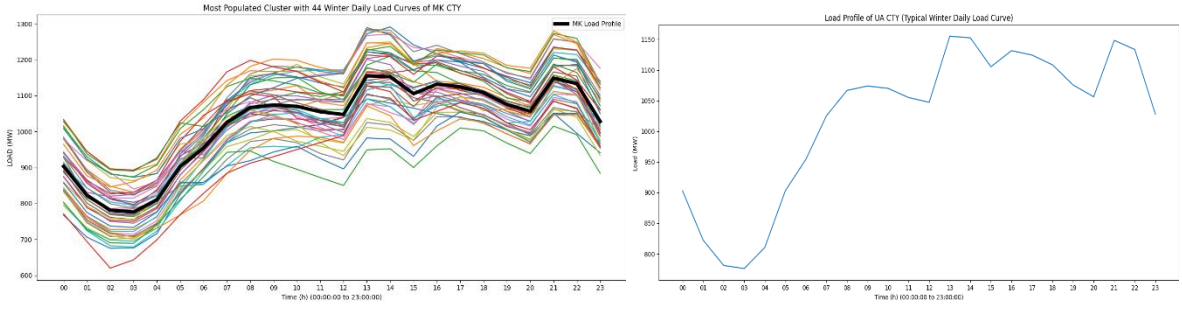
Σχήμα Π.Β.64 : Χειμερινό Προφίλ Φορτίου (LV).

15) Μαυροβούνιο (ME)



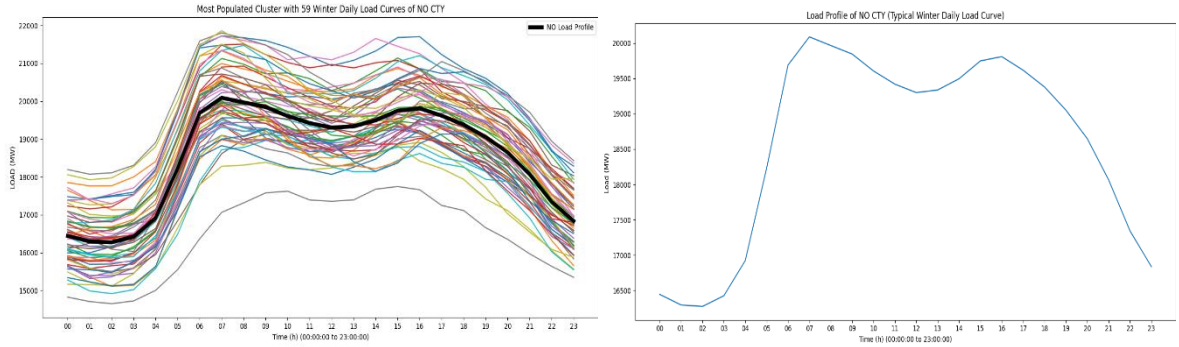
Σχήμα Π.Β.65 : Χειμερινό Προφίλ Φορτίου (ME).

16) Βόρεια Μακεδονία (MK)



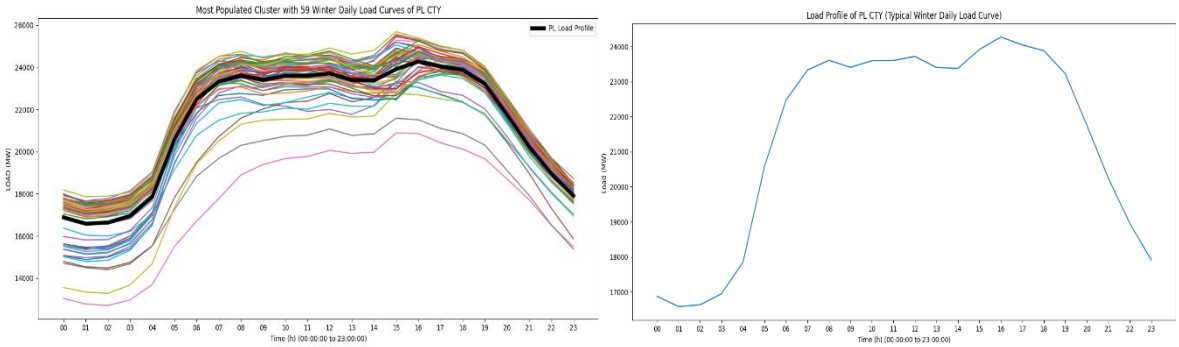
Σχήμα Π.Β.66 : Χειμερινό Προφίλ Φορτίου (MK).

17) Νορβηγία (NO)



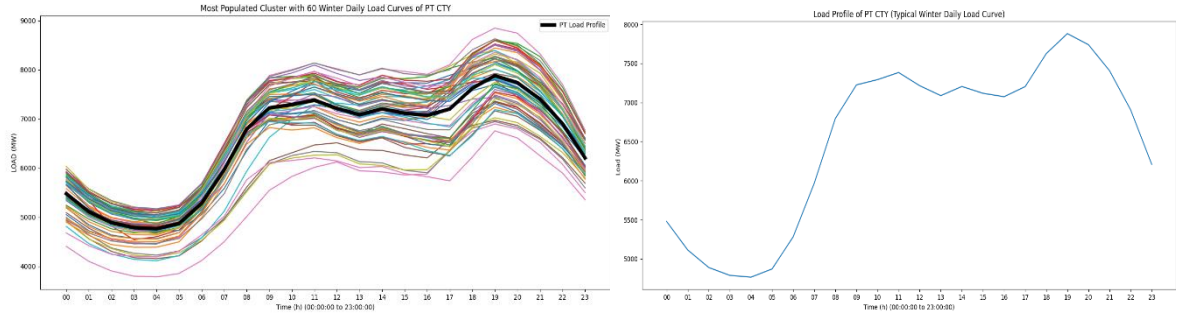
Σχήμα Π.Β.67 : Χειμερινό Προφίλ Φορτίου (NO).

18) Πολωνία (PL)



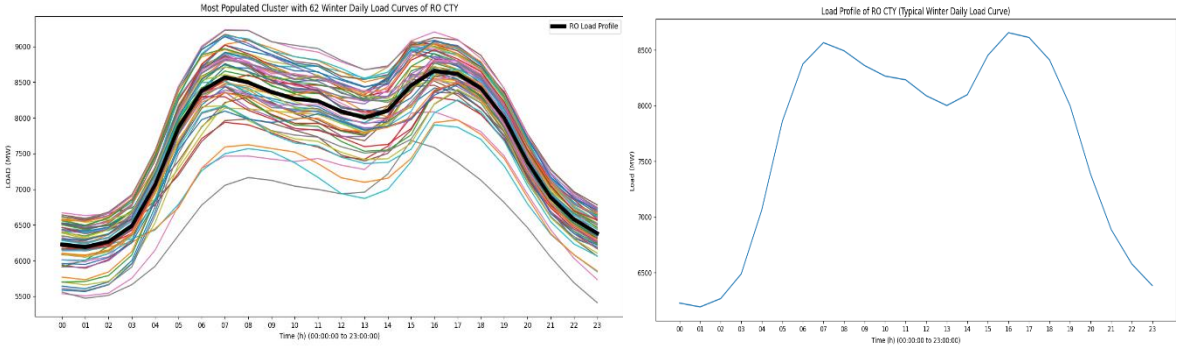
Σχήμα Π.Β.68 : Χειμερινό Προφίλ Φορτίου (PL).

19) Πορτογαλία (PT)



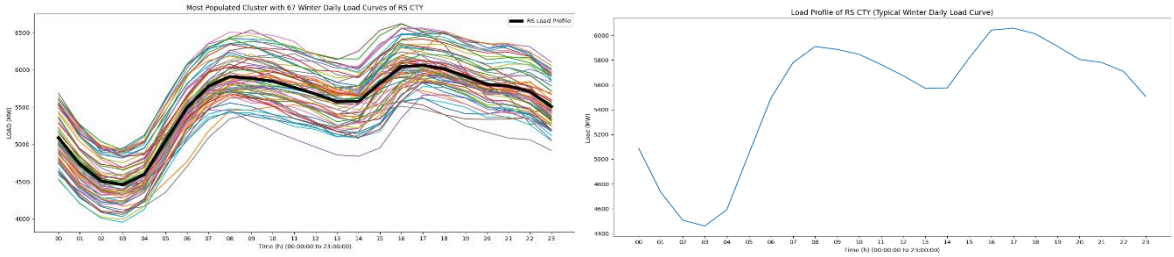
Σχήμα Π.Β.69 : Χειμερινό Προφίλ Φορτίου (PT).

20) Ρουμανία (RO)



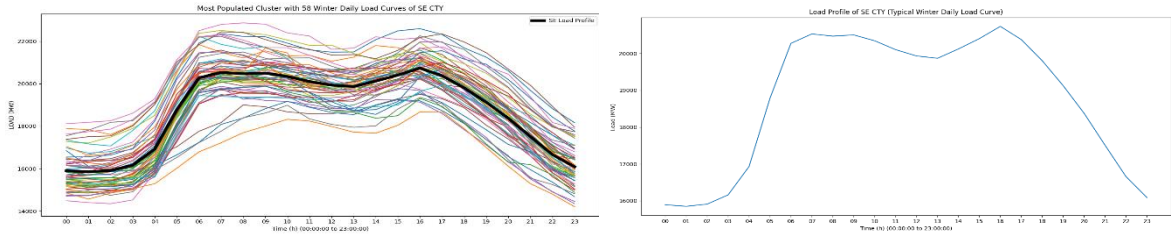
Σχήμα Π.Β.70 : Χειμερινό Προφίλ Φορτίου (RO).

21) Σερβία (RS)



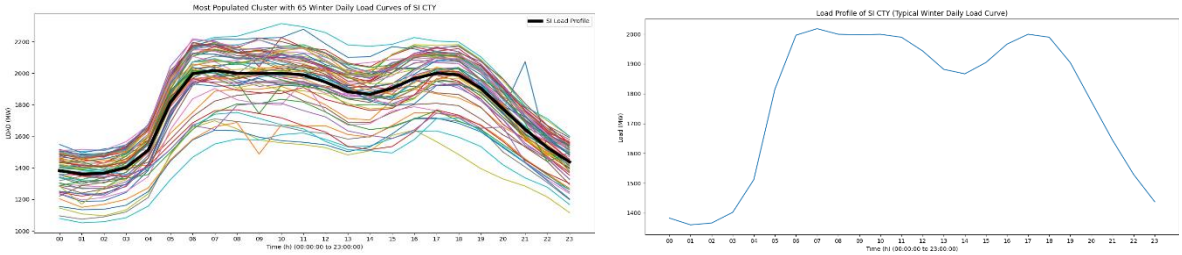
Σχήμα Π.Β.71 : Χειμερινό Προφίλ Φορτίου (RS).

22) Σουηδία (SE)



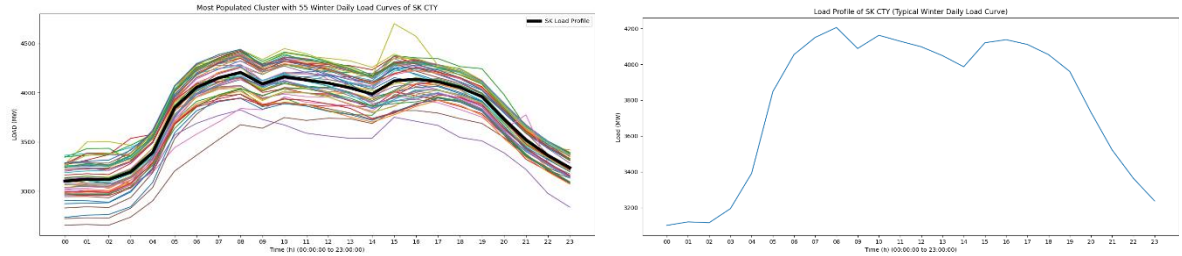
Σχήμα Π.Β.72 : Χειμερινό Προφίλ Φορτίου (SE).

23) Σλοβενία (SI)



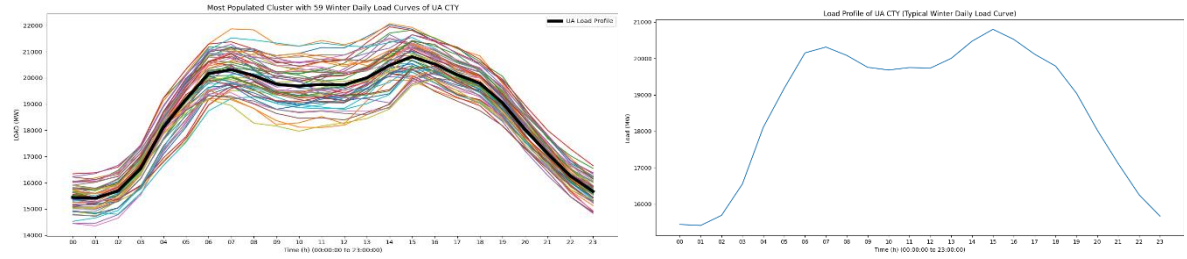
Σχήμα Π.Β.73 : Χειμερινό Προφίλ Φορτίου (SI).

24) Σλοβακία (SK)



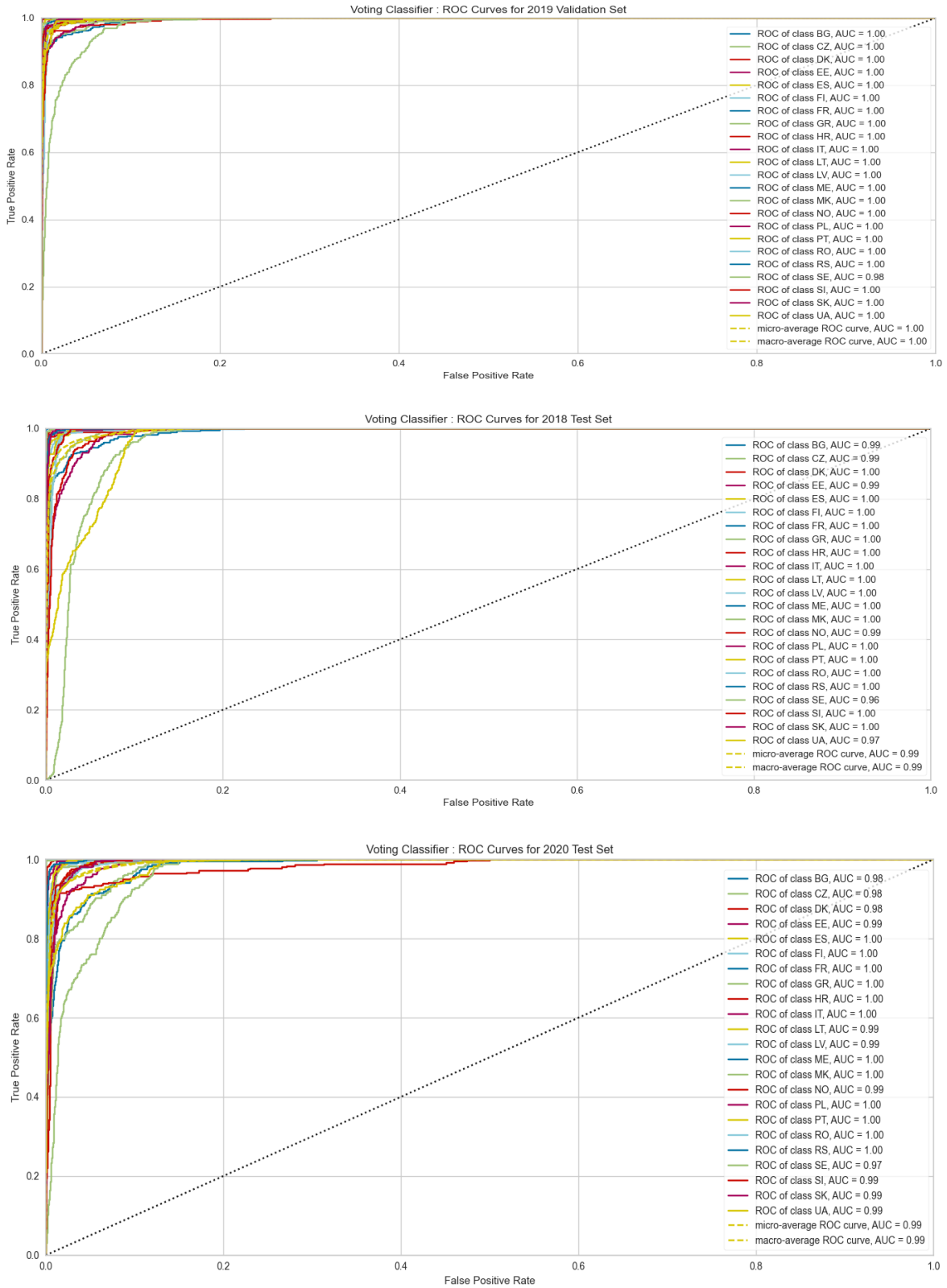
Σχήμα Π.Β.74 : Χειμερινό Προφίλ Φορτίου (SK).

25) Ουκρανία (UA)

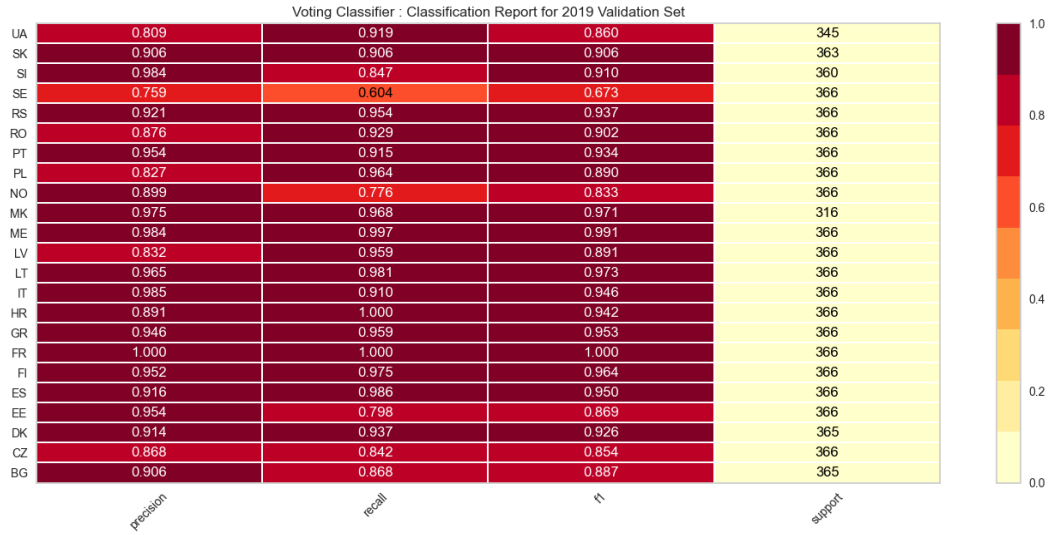


Σχήμα Π.Β.75 : Χειμερινό Προφίλ Φορτίου (UA).

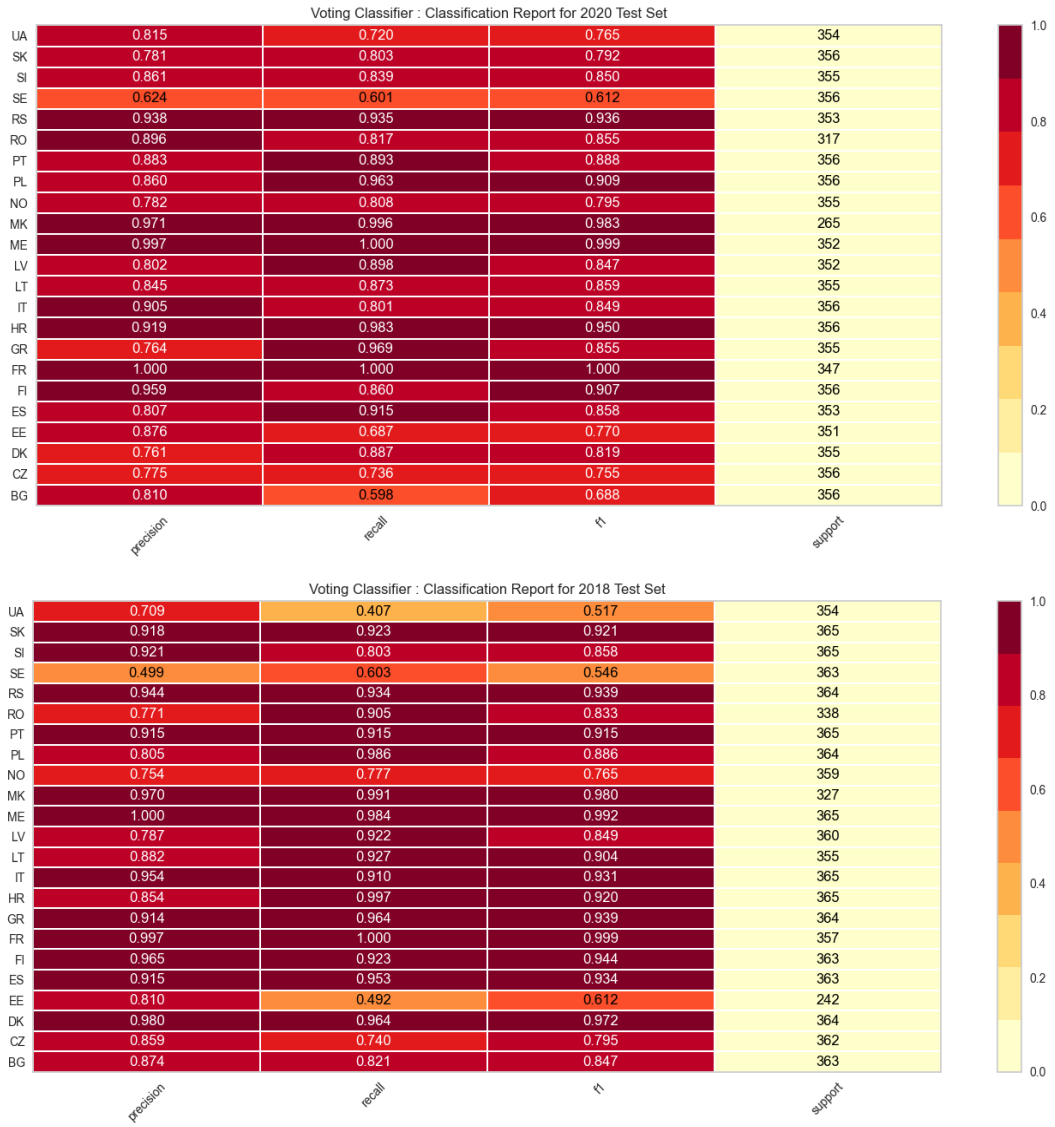
Παράρτημα Γ : Αποτελέσματα Αξιολόγησης του Ταξινομητή Ψηφοφορίας στο Πεδίο Χαρακτηριστικών



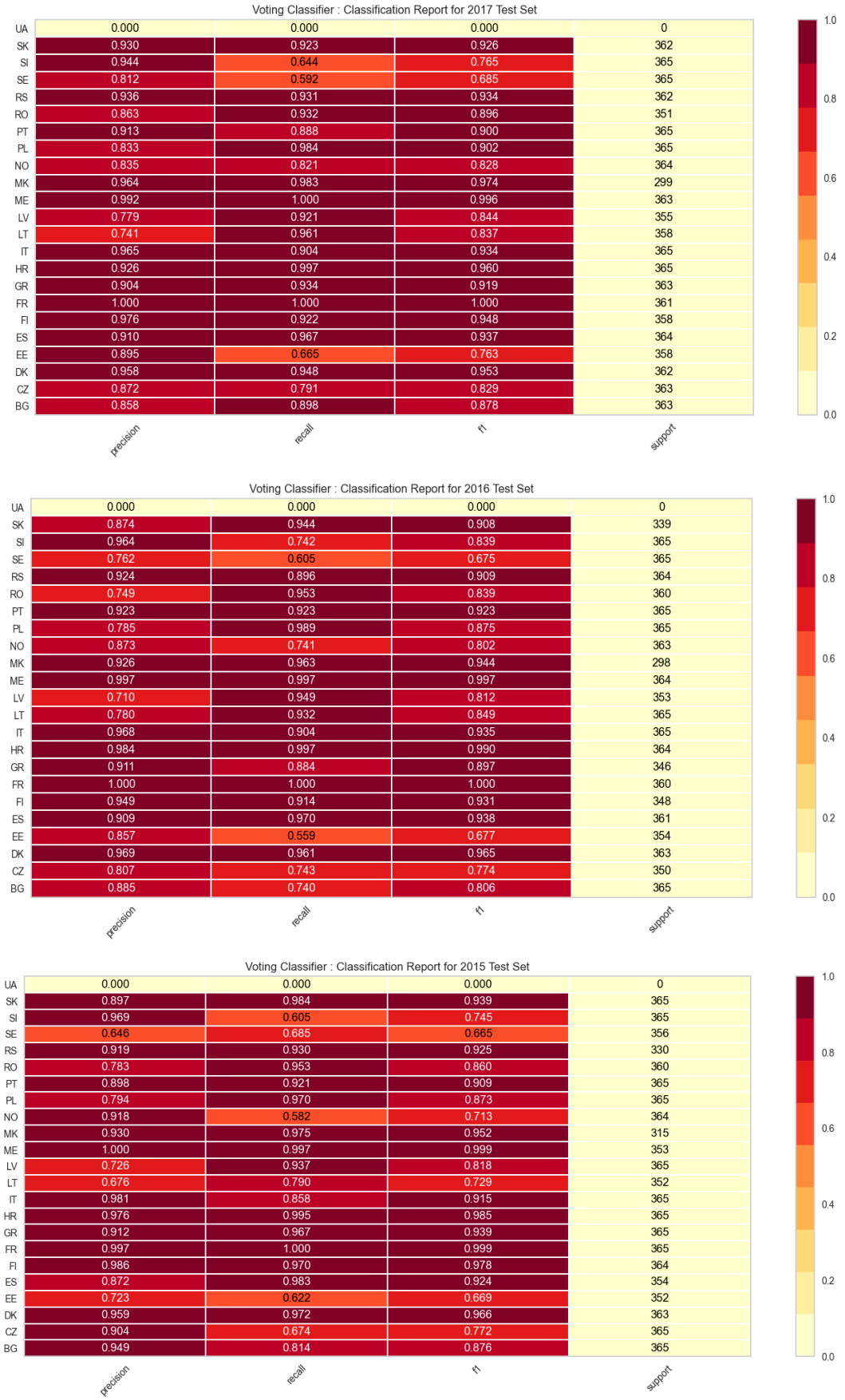
Σχήμα Π.Γ.1 : Διαγράμματα ROC του Feature Based Voting Classifier.



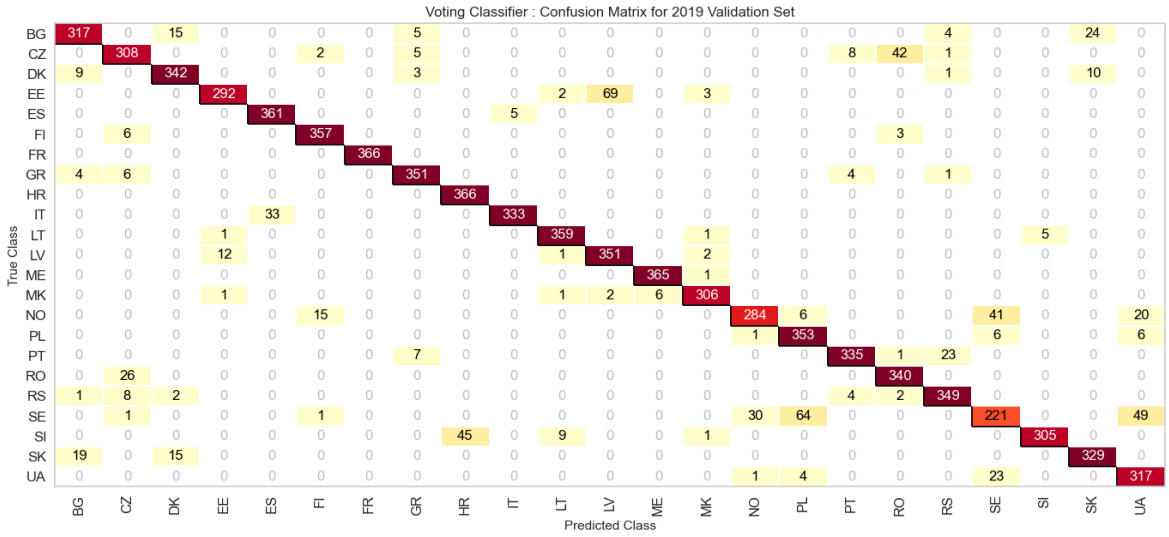
Σχήμα Π.Γ.2 : Classification Report του Feature Based Voting Classifier για το σύνολο επικύρωσης.



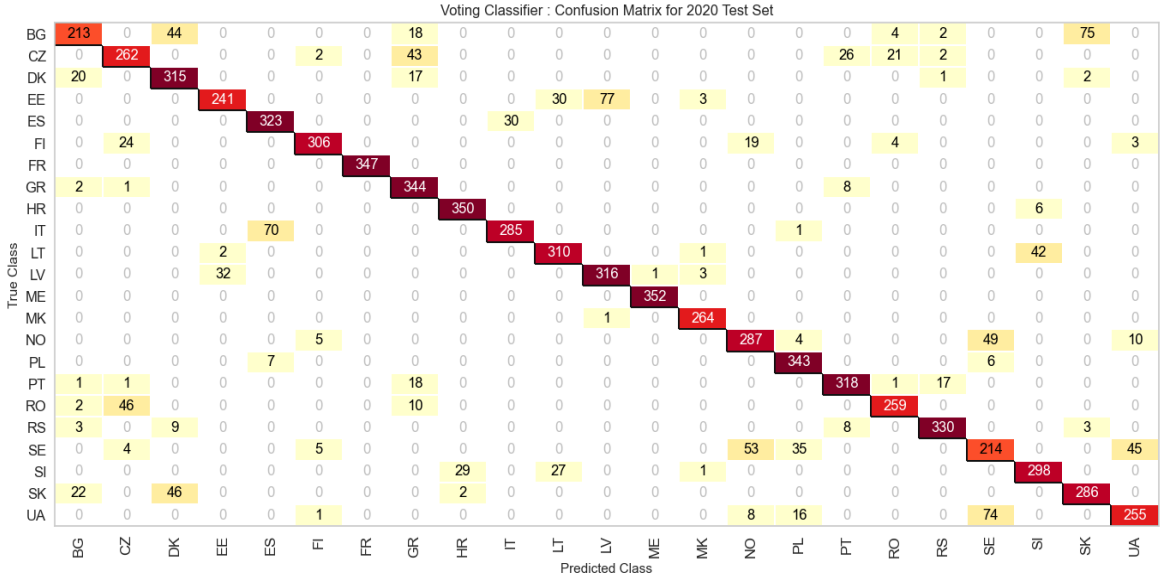
Σχήμα Π.Γ.3 : Classification Report του Feature Based Voting Classifier για τα σύνολα ελέγχου "2020" (TestSet_5) και "2018" (TestSet_4).

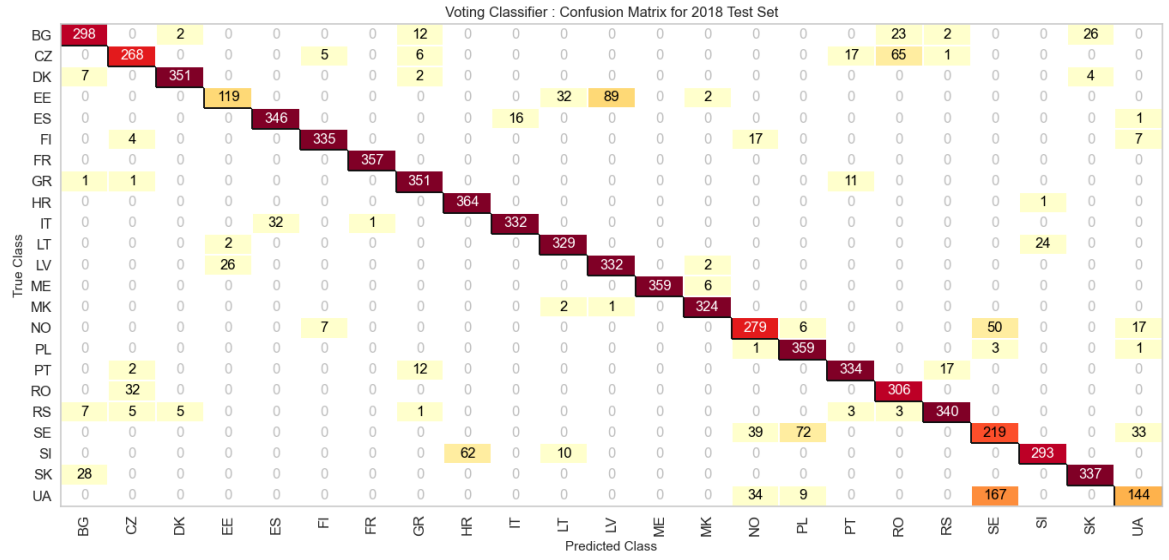


Σχήμα Π.Γ.4 : Classification Report του Feature Based Voting Classifier για τα σύνολα ελέγχου "2017" (TestSet_3) , "2016" (TestSet_2) και "2015" (TestSet_1).

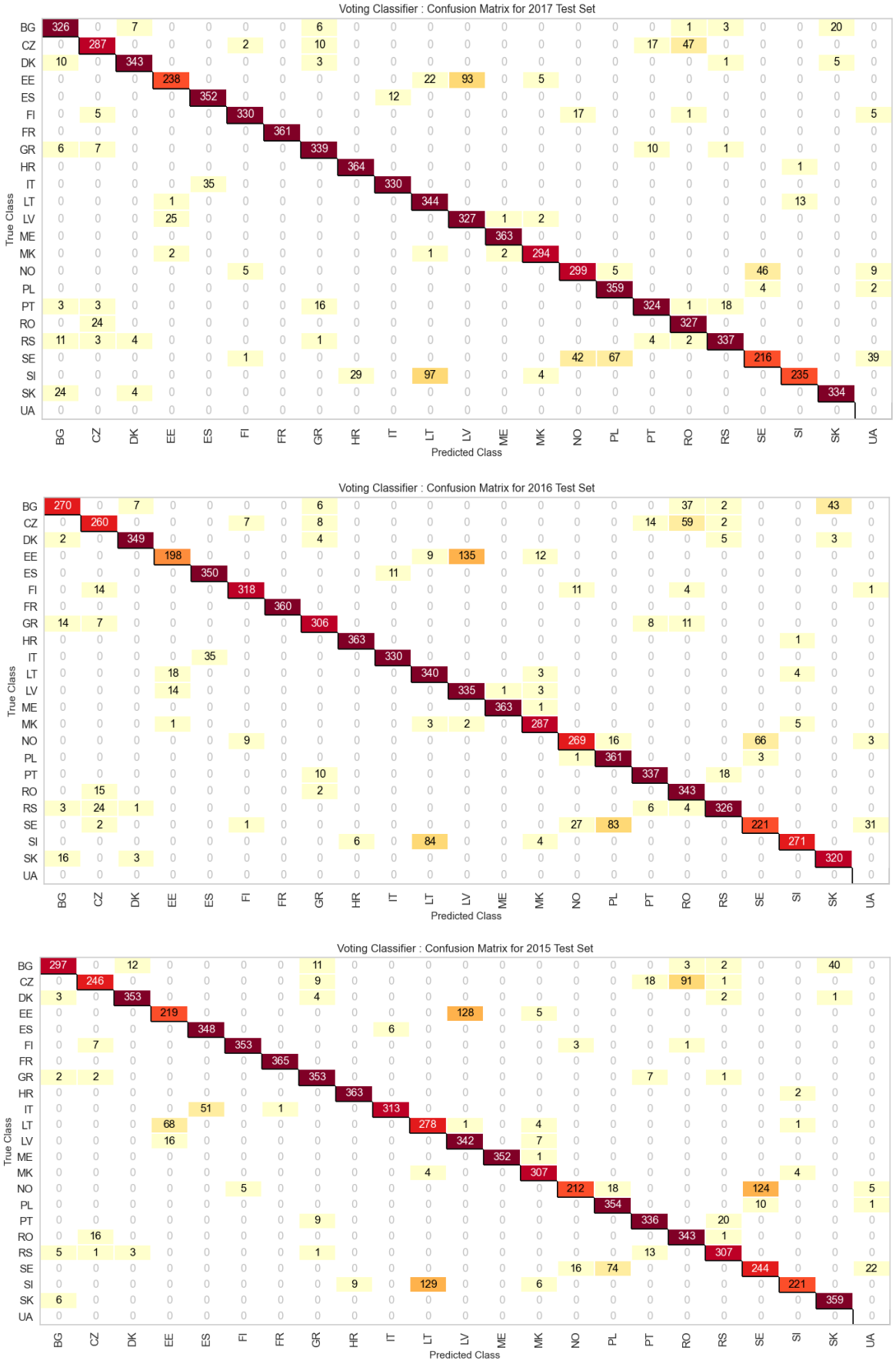


Σχήμα Π.Β.5 : Πίνακας Σύγχησης του Feature Based Voting Classifier για το σύνολο επικύρωσης.

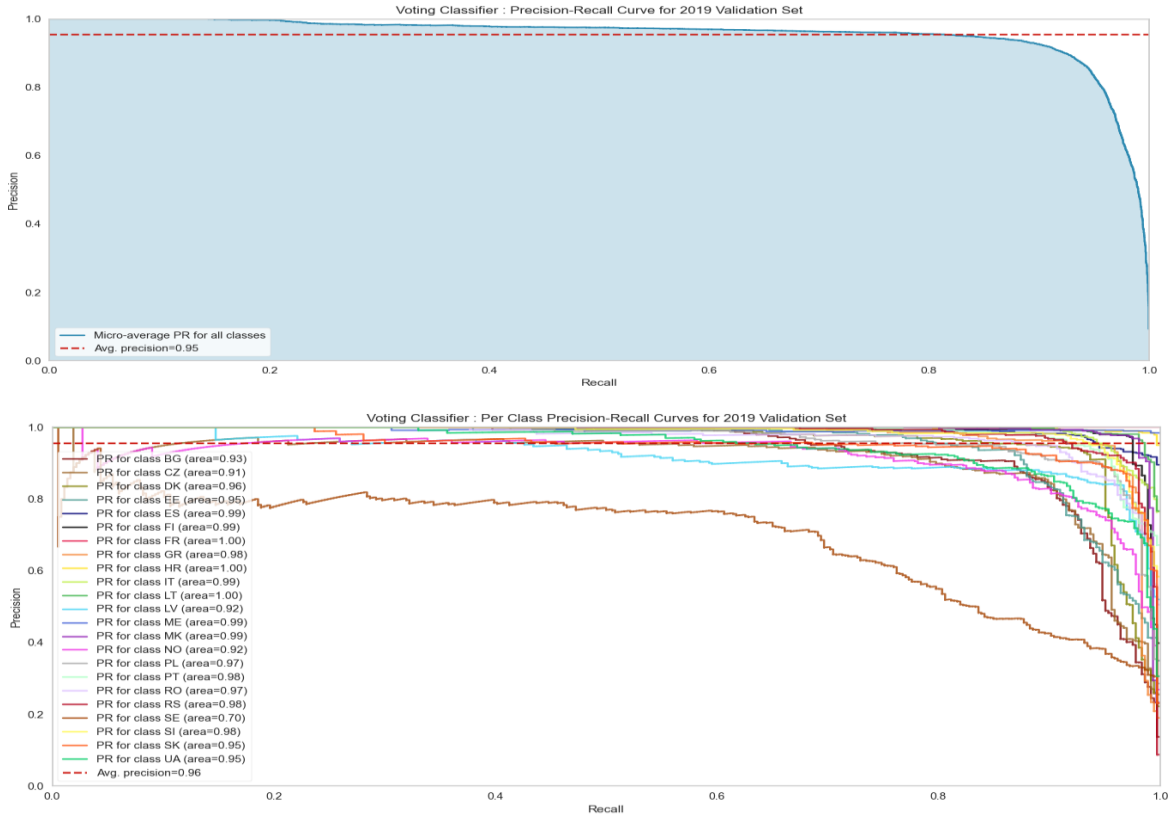




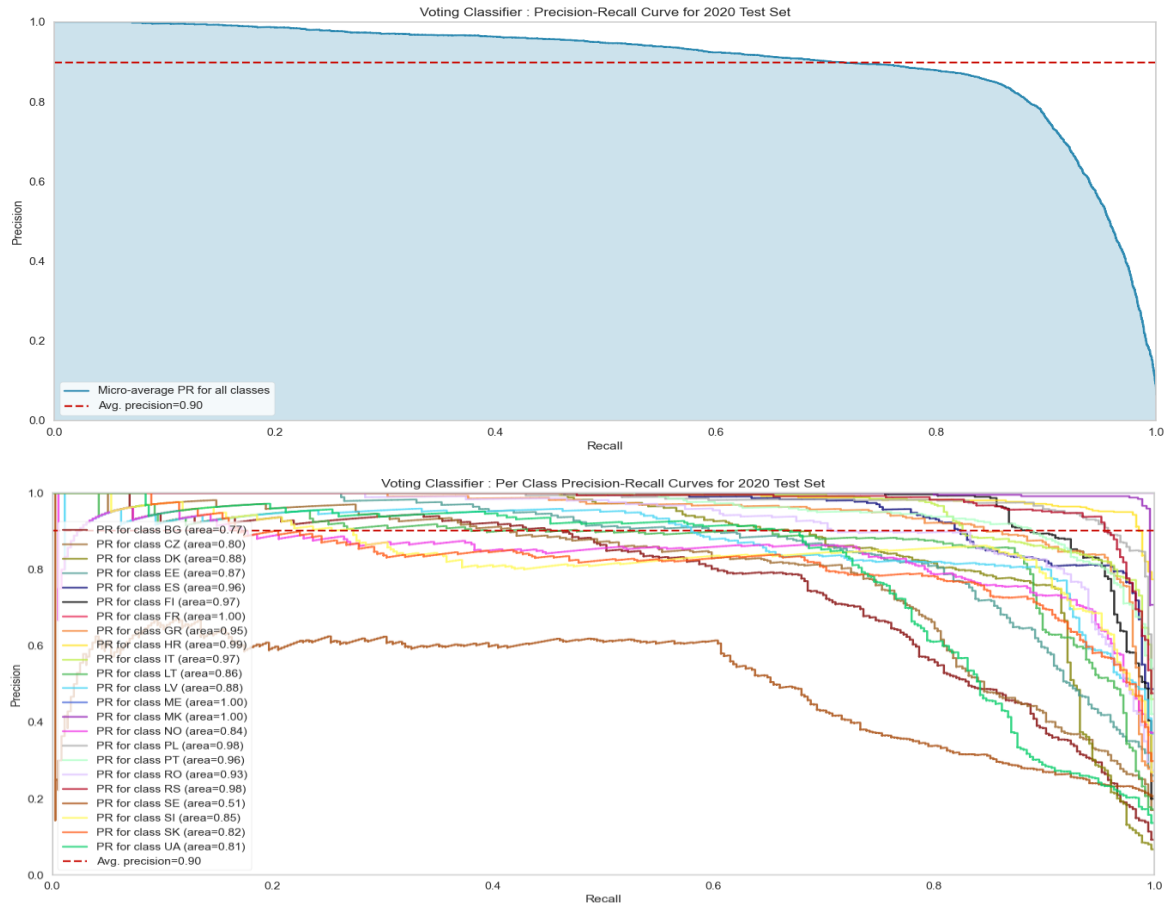
Σχήμα Π.Γ.6 : Πίνακες Σύγχησης του Feature Based Voting Classifier για τα σύνολα ελέγχου "2020" (TestSet_5) και "2018" (TestSet_4).



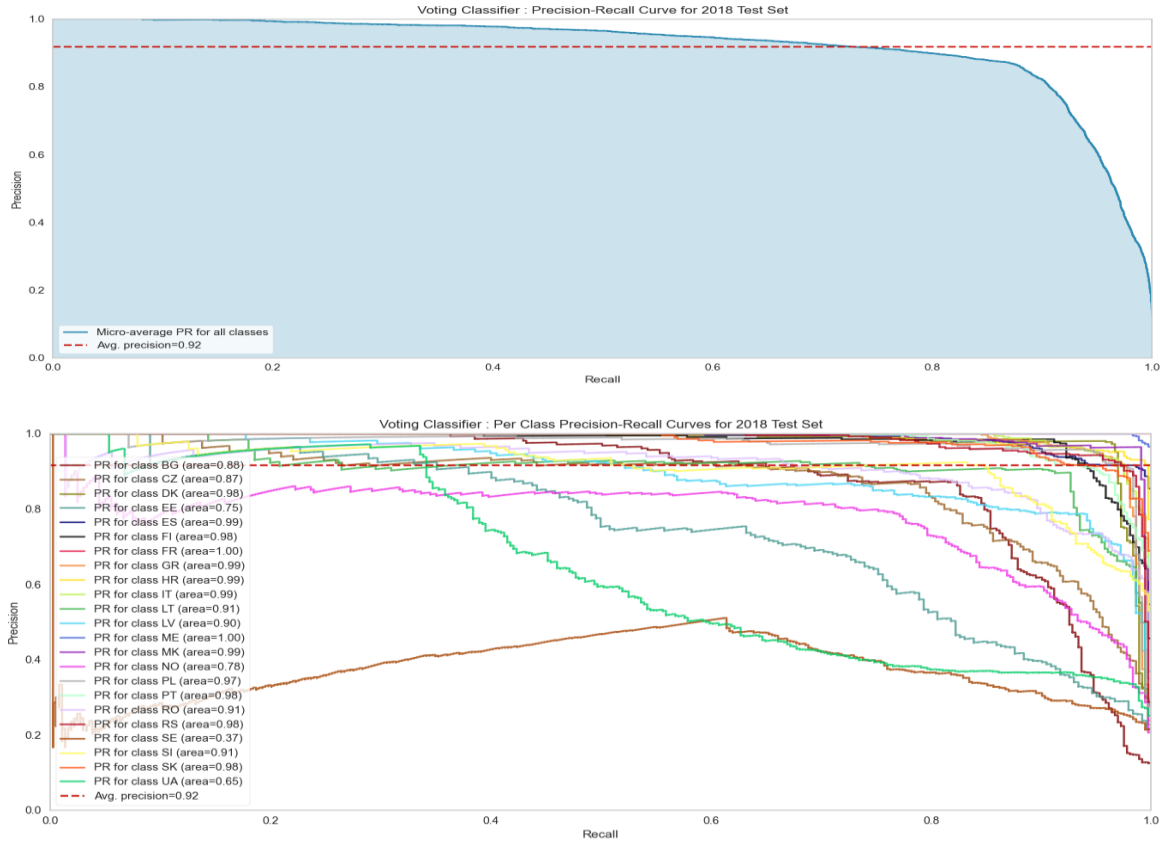
Σχήμα Π.Γ.7 : Πίνακες Σύγχησης του Feature Based Voting Classifier για τα σύνολα ελέγχου "2017" (TestSet_3), "2016" (TestSet_2) και "2015" (TestSet_1).



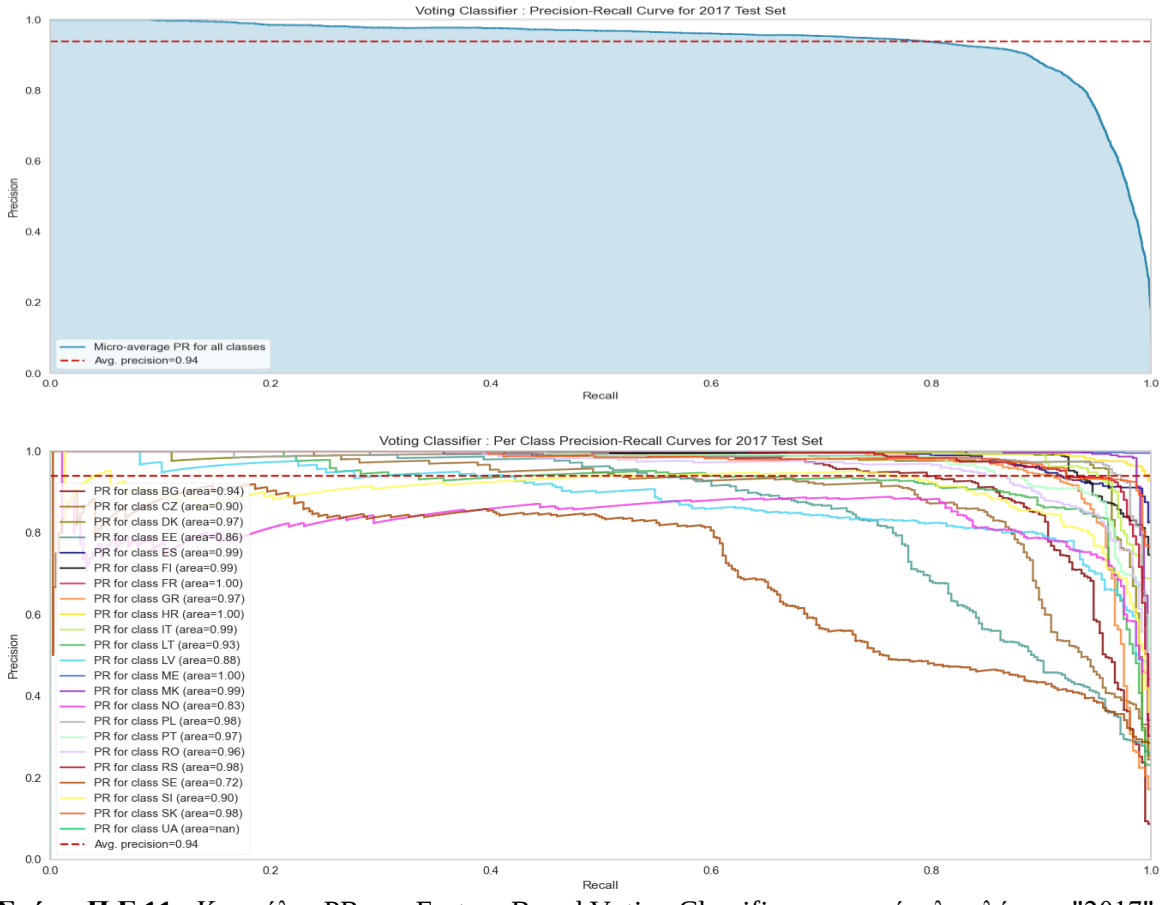
Σχήμα Π.Γ.8 : Καμπύλες PR του Feature Based Voting Classifier για το σύνολο επικύρωσης.



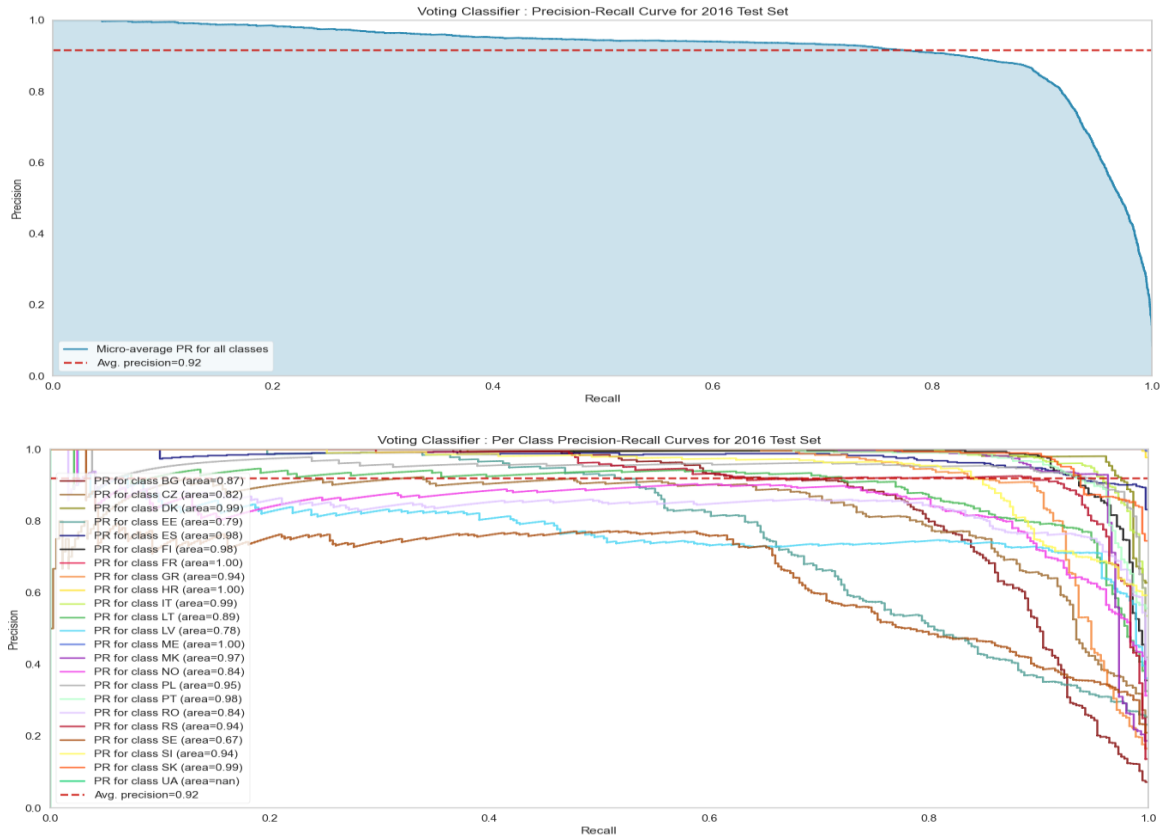
Σχήμα Π.Γ.9 : Καμπύλες PR του Feature Based Voting Classifier για το σύνολο ελέγχου "2020".



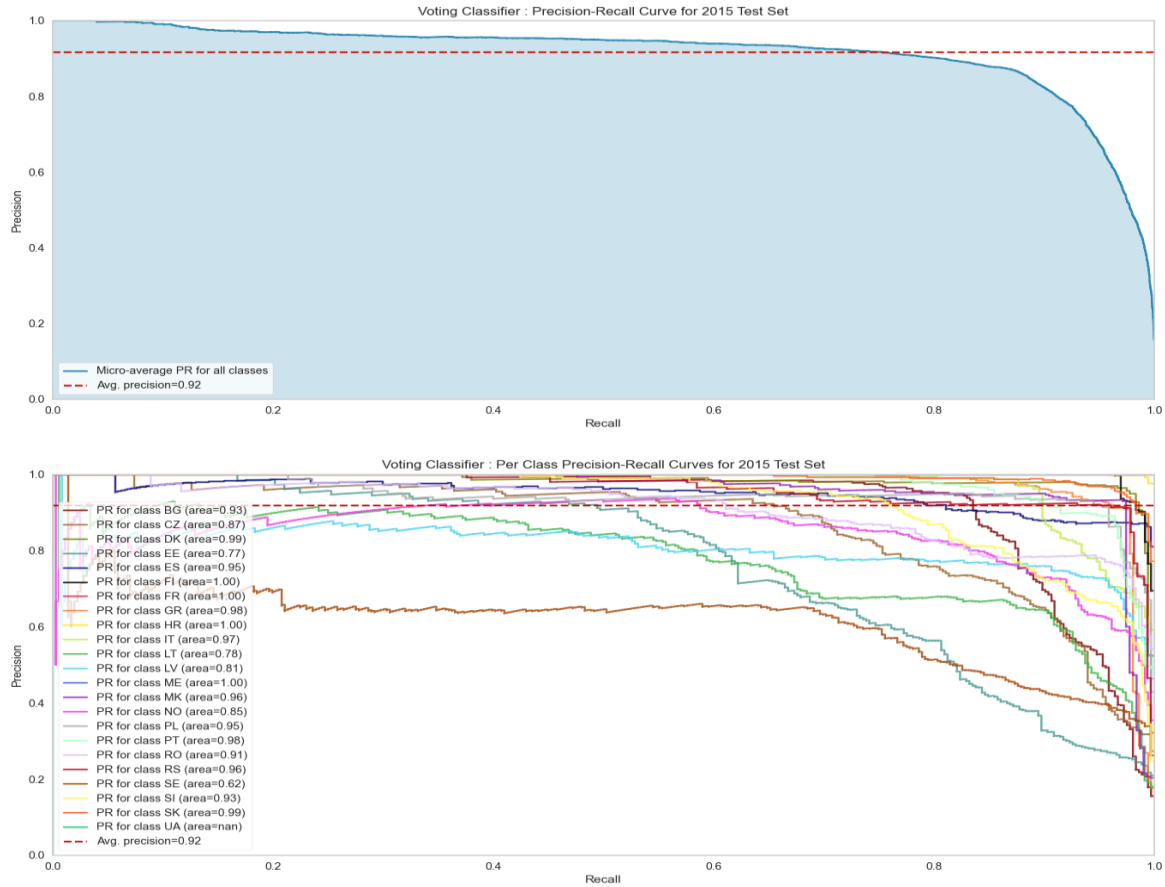
Σχήμα Π.Γ.10 : Καμπύλες PR του Feature Based Voting Classifier για το σύνολο ελέγχου "2018".



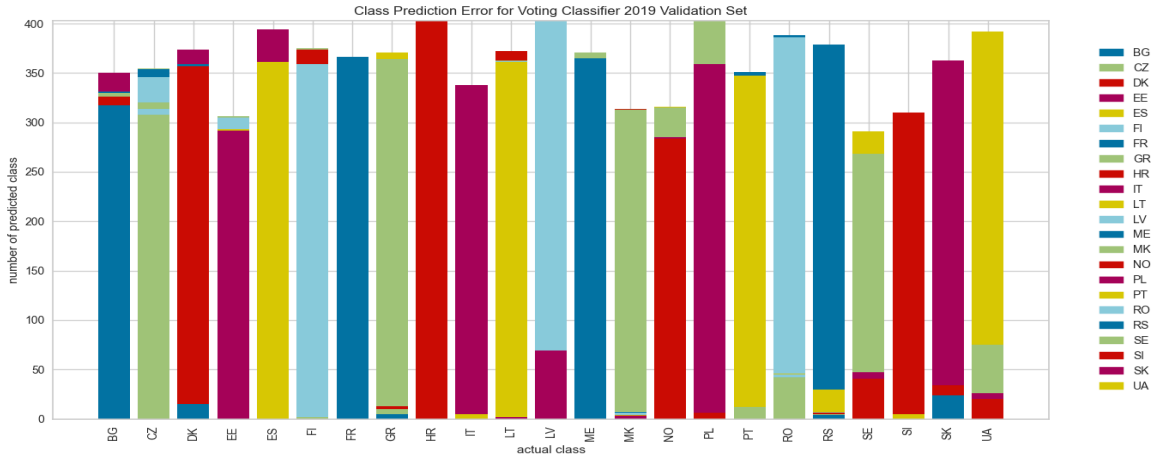
Σχήμα Π.Γ.11 : Καμπύλες PR του Feature Based Voting Classifier για το σύνολο ελέγχου "2017".



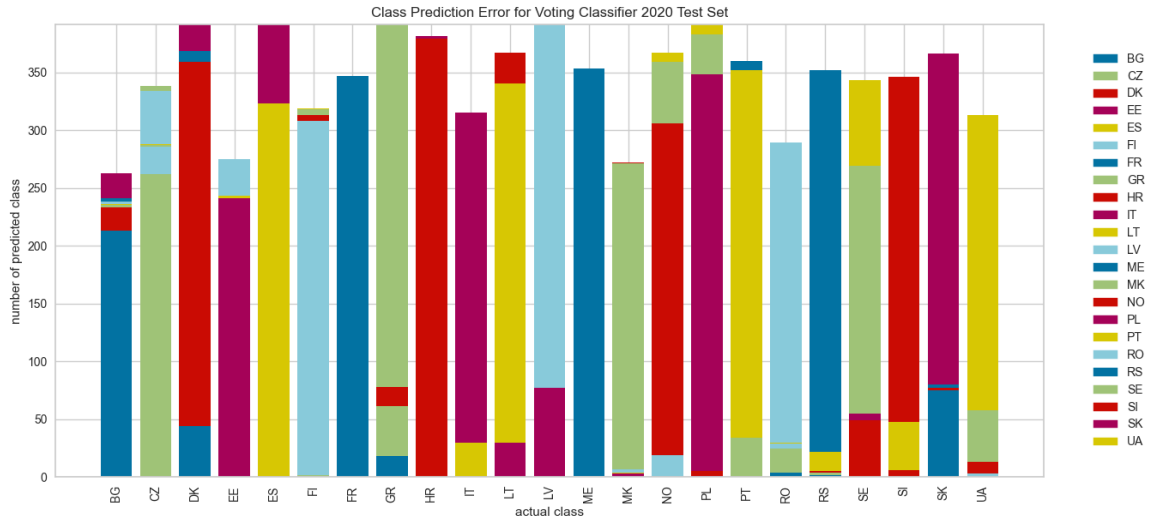
Σχήμα Π.Γ.12 : Καμπύλες PR του Feature Based Voting Classifier για το σύνολο ελέγχου "2016".



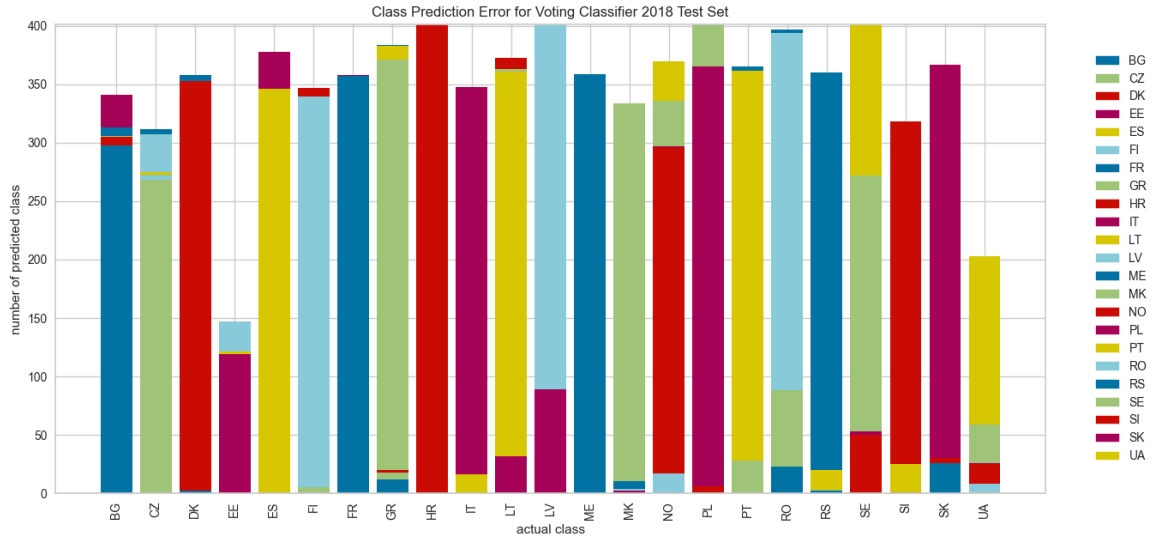
Σχήμα Π.Γ.13 : Καμπύλες PR του Feature Based Voting Classifier για το σύνολο ελέγχου "2015".



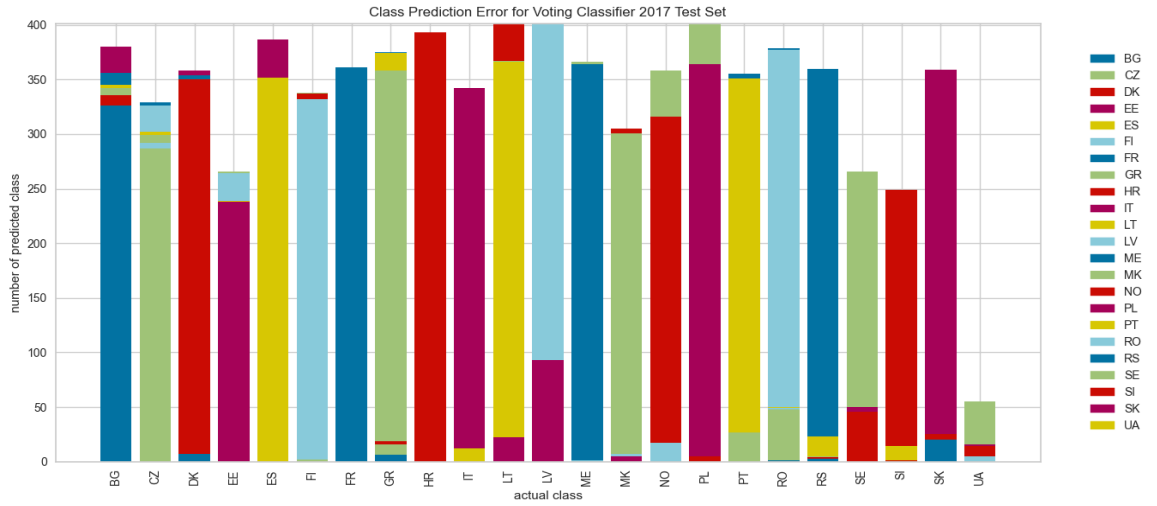
Σχήμα Π.Γ.14 : Class Prediction Error του Feature Based Voting Classifier για το σύνολο Επικύρωσης.



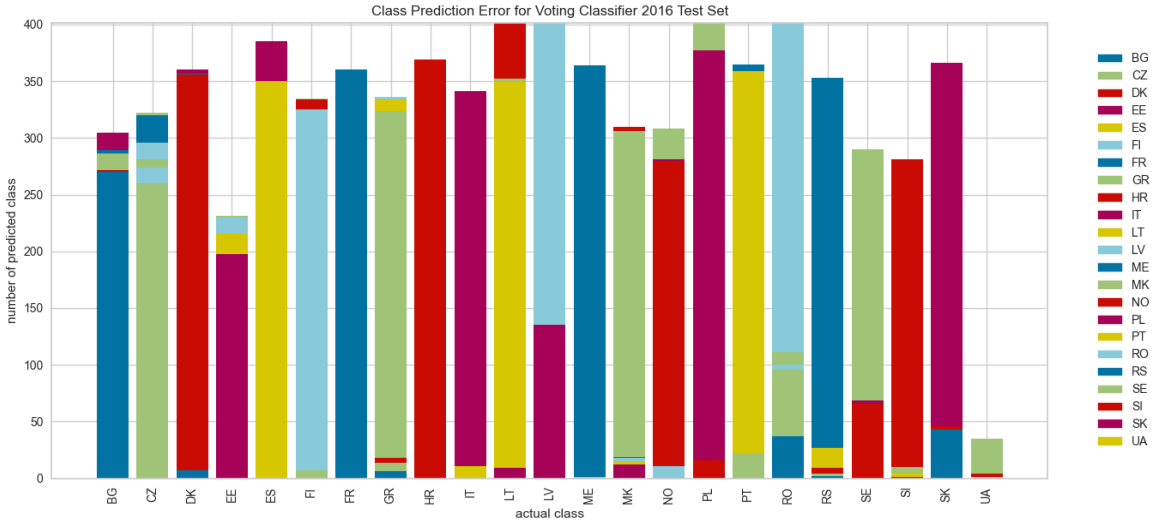
Σχήμα Π.Γ.15 : Class Prediction Error του Feature Based Voting Classifier για το σύνολο ελέγχου "2020" (TestSet_5).



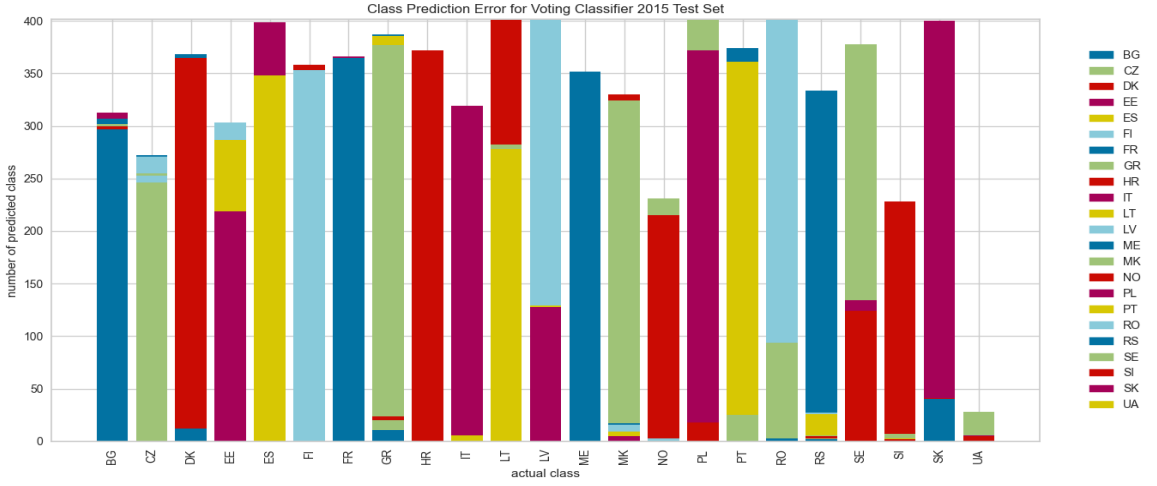
Σχήμα Π.Γ.16 : Class Prediction Error του Feature Based Voting Classifier για το σύνολο ελέγχου "2018" (TestSet_4).



Σχήμα Π.Γ.17 : Class Prediction Error του Feature Based Voting Classifier για το σύνολο ελέγχου "2017" (TestSet_3).



Σχήμα Π.Γ.18 : Class Prediction Error του Feature Based Voting Classifier για το σύνολο ελέγχου "2016" (TestSet_2).



Σχήμα Π.Γ.19 : Class Prediction Error του Feature Based Voting Classifier για το σύνολο ελέγχου "2015" (TestSet_1).

Βιβλιογραφία

[1] Philippe E., Carlos A., Time-series data mining. *ACM Computing Surveys, Association for Computing Machinery*, 2012, 45 (1), pp.12. 10.1145/2379776.2379788.

[2] Tang L., Wang C., Wang S., Energy time series data analysis based on a novel integrated data characteristic testing approach, *Procedia Computer Science*, (2013), pp. 759-756

[3] Zor K., Çelik Ö., Timur O., Yıldırım, B., Yıldırım H. B., Teke A., Simple Approaches to Missing Data for Energy Forecasting Applications, *16th International Conference on Clean Energy(ICCE-2018)*

[4] Piao M., Lee H.G., Park J.H., Ryu K.H. (2008) Application of Classification Methods for Forecasting Mid-Term Power Load Patterns. In: Huang DS., Wunsch D.C., Levine D.S., Jo KH. (eds) *Advanced Intelligent Computing Theories and Applications. With Aspects of Contemporary Intelligent Computing Techniques*. ICIC 2008. Communications in Computer and Information Science, v. 15. Springer, Berlin, Heidelberg.

[5] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[6] Fulcher B., Jones N., Highly comparative feature-based time-series classification, *IEEE Trans. Knowl. Data Eng. 26, 3026* (2014)

[7] Serrà, J., Arcos J., An empirical evaluation of similarity measures for time series classification, *Knowledge-Based Systems*, (2014), pp. 305-314, v. 67

[8] Ogasawara E., Martinez L., Oliveira D., Zimbrão G., Pappa G., Mattoso M., Adaptive Normalization: A Novel Data Normalization Approach for Non-Stationary Time Series, *Evolutionary Computation (CEC), 2010 IEEE Congress*

[9] I. P. Panapakidis, M. C. Alexiadis and G. K. Papagiannis, "Electricity customer characterization based on different representative load curves," *2012 9th International Conference on the European Energy Market*, 2012, pp. 1-8.

[10] Fulcher B., (2017). Feature-based time-series analysis

[11] V. Figueiredo, F. Rodrigues, Z. Vale and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," in *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 596-602, May 2005

[12] Kraskov AStögbauer HGrassberger P, Estimating mutual information, *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, (2004), 16, v. 69(6)

[13] Christ, Maximilian & Kempa-Liehr, Andreas & Feindt, Michael. (2016). Distributed and parallel time series feature extraction for industrial big data applications.

[14] Wang X., Smith,K., Hyndman R., (2006). Characteristic-Based Clustering for Time Series Data. Data Min. Knowl. Discov., v. 13. pp. 335-364.

[15] Chandrashekar G., Sahin F., A survey on feature selection methods, *Computers & Electrical Engineering*, v. 40, Issue 1, 2014, pp. 16-28.

[16] Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. In Data Classification: Algorithms and Applications (pp. 37-64). CRC Press.

[17] Commission Regulation (EU) No 543/2013 of 14 June 2013 on submission and publication of data in electricity markets and amending Annex I to Regulation (EC) No 714/2009 of the European Parliament and of the Council Text with EEA relevance

[18] ENTSO-E Detailed Data Descriptions, https://eepublicdownloads.entsoe.eu/clean-documents/pre2015/resources/Transparency/MoP_Ref_02_-_Detailed_Data_Descriptions_v1r2.pdf

[19] Load and consumption data: Specificities of member countries, https://eepublicdownloads.entsoe.eu/clean-documents/pre2015/publications/ce/Load_and_Consumption_Data.pdf

[20] A review of the ENTSO-E Transparency Platform, https://ec.europa.eu/energy/sites/ener/files/documents/review_of_the_entso_e_plattform.pdf

[21] Hirth, Lion & Mühlenpfordt, Jonathan & Bulkeley, Marisa. (2018). The ENTSO-E Transparency Platform – A review of Europe's most ambitious electricity data platform. *Applied Energy*. 225. 1054-1067. 10.1016/j.apenergy.2018.04.048.

[22] Hossin, Mohammad & M.N, Sulaiman. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process. 5. 01-11. 10.5121/ijdkp.2015.5201.

[23] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (ICML '06). Association for Computing Machinery, New York, NY, USA, 233–240.

[24] Cios, Krzysztof & Pedrycz, Witold & Swiniarski, Roman & Kurgan, Lukasz. (2007). Data Mining: A Knowledge Discovery Approach. 10.1007/978-0-387-36795-8.

[25] Introduction to Data Mining and Knowledge Discovery, Two Crows Corporation 3, Two Crows Corporation, 1999, 1892095025, 9781892095022, pp. 36

[26] Saeid Soheily-Khah. Generalized k-means-based clustering for temporal data under time warp. Other[cs.OH]. Université Grenoble Alpes, 2016. English. NNT: 2016GREAM064. tel-01680370v2

[27] Anil K. Jain and Richard C. Dubes. 1988. Algorithms for clustering data. Prentice-Hall, Inc., USA.

[28] Javed, Ali & Lee, Byung & Rizzo, Donna. (2020). A benchmark study on time series clustering. Machine Learning with Applications. 1. 10.1016/j.mlwa.2020.100001.

[29] Silva, Vera & Duarte, F. & Rodrigues, Fátima & Vale, Zita & Ramos, Cathleen & Ramos, Sérgio & Borges Gouveia, Joaquim. (2003). Electric Energy Customer Characterization by Clustering.

[30] T. Warren Liao, Clustering of time series data—a survey, *Pattern Recognition*, Volume 38, Issue 11, 2005, Pages 1857-1874, ISSN 0031-3203,

[31] Iglesias F, Kastner W. Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns. Energies. 2013; 6(2):579-597.

[32] David Arthur and Sergei Vassilvitskii. 2007. K-means++: the advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (SODA '07). Society for Industrial and Applied Mathematics, USA, 1027–1035.

[33] G. J. Tsekouras, N. D. Hatziargyriou and E. N. Dialynas, "Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers," in IEEE Transactions on Power Systems, vol. 22, no. 3, pp. 1120-1128, Aug. 2007

[34] Roelofsen P., (2018), Time series clustering, (Master's Thesis), Vrije Universiteit, Amsterdam, Retrieved from https://beta.vu.nl/nl/Images/stageverslag-roelofsen_tcm235-882304.pdf

[35] Aghabozorgi, Sr & Shirkhorshidi, Ali Seyed & Wah, Teh. (2015). Time-series clustering - A decade review. *Information Systems*. 53. 10.1016/j.is.2015.04.007.

[36] Arbelaitz, Olatz & Gurrutxaga, Ibai & Muguerza, Javier & Pérez, Jesús & Perona, Iñigo. (2013). An extensive comparative study of cluster validity indices. Pattern Recognition. 46. 243-256. 10.1016/j.patcog.2012.07.021.

[37] T. Caliński & J Harabasz (1974) A dendrite method for cluster analysis, *Communications in Statistics*, 3:1, 1-27, DOI: 10.1080/03610927408827101

[38] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224-227, April 1979, doi: 10.1109/TPAMI.1979.4766909.

[39] Chaimontree, Santhana & Atkinson, Katie & Coenen, Frans. (2010). Best Clustering Configuration Metrics: Towards Multiagent Based Clustering. 48-59. 10.1007/978-3-642-17316-5_5.

[40] Hämäläinen, Joonas & Jauhiainen, Susanne & Kärkkäinen, Tommi. (2017). Comparison of Internal Clustering Validation Indices for Prototype-Based Clustering. Algorithms. 10. 105. 10.3390/a10030105.

[41] Nasser, Alissar. (2019). Investigating K-means and Kernel K-means Algorithms with Internal Validity Indices for Cluster Identification. Journal of Advances in Mathematics and Computer Science. 30. 1-12. 10.9734/JAMCS/2019/45837.

[42] Rousseeuw, Peter. (1987). Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math. 20, 53-65. Journal of Computational and Applied Mathematics. 20. 53-65. 10.1016/0377-0427(87)90125-7.

[43] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 12, pp. 1650-1654, Dec. 2002, doi: 10.1109/TPAMI.2002.1114856.

[44] Schultz, David & Jain, Brijnesh. (2018). Nonsmooth Analysis and Subgradient Methods for Averaging in Dynamic Time Warping Spaces. Pattern Recognition. 74. 340-358. 10.1016/j.patcog.2017.08.012.

[45] Petitjean, François & Ketterlin, Alain & Gancarski, Pierre. (2011). A global averaging method for dynamic time warping, with applications to clustering. Pattern Recognition. 44. 678-. 10.1016/j.patcog.2010.09.013.

[46] Ratanamahatana, Chotirat & Keogh, E.. (2004). Everything you know about dynamic time warping is wrong.

[47] Petitjean, François & Forestier, Germain & Webb, Geoffrey & Nicholson, Ann & Chen, Yanping & Keogh, Eamonn. (2016). Faster and more accurate classification of time series by exploiting a novel dynamic time warping averaging algorithm. Knowledge and Information Systems. 47. 10.1007/s10115-015-0878-8.

[48] Kurbalija, Vladimir & Radovanovic, Milos & Geler, Zoltan & Ivanovic, Mirjana. (2011). The Influence of Global Constraints on DTW and LCS Similarity Measures for Time-Series Databases. 10.1007/978-3-642-23163-6_10.

[49] Andriy Burkov, The Hundred-Page Machine Learning Book, illustrated, Andriy Burkov, 2019, 1999579518, 9781999579517, pp. 160

[50] Aurélien Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, (2019), O'Reilly Media Inc., 9781492032649

[51] Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J., Data Mining: Practical Machine Learning Tools and Techniques , 4th Edition, (2016), 978-0128042915

[52] Hastie T., Tibshirani R., Friedman J., The Elements of Statistical Learning, 2nd Edition, (2009), Springer Series in Statistics, 978-0387848570

[53] Alharan, Abbas & Alsagheer, Radhwan & Al-Haboobi, Ali. (2017). Popular Decision Tree Algorithms of Data Mining Techniques: A Review. International Journal of Computer Science and Mobile Computing. 6. 133-142.

[54] Yang Zhao, Chaobo Zhang, Yiwen Zhang, Zihao Wang, Junyang Li, A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis, *Energy and Built Environment*, Volume 1, Issue 2, 2020, Pages 149-164, ISSN 2666-1233,

[55] 2009/73/EK retrieved from https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:211:0094:0136:EL:PDF

[56] matplotlib Documentation, v. 3.1.1, retrieved from https://matplotlib.org/3.1.1/Matplotlib.pdf

[57] pandas Documentation, v.1.1.3, retrieved from https://pandas.pydata.org/pandas-docs/version/1.1.3/pandas.pdf

[58] scikit-learn user guide, v. 0.23.2, retrieved from https://scikit-learn.org/0.23//_downloads/scikit-learn-docs.pdf

[59] tsfresh Documentation, v. 0.11.1, retrieved from https://tsfresh.readthedocs.io/_/downloads/en/v0.11.1/pdf/

[60] tslearn Documentation, v 0.5.0.4, retrieved https://readthedocs.org/projects/tslearn/downloads/pdf/latest/

[61] Yellowbrick Documentation, v1.3.post1, retrieved from https://buildmedia.readthedocs.org/media/pdf/yellowbrick/stable/yellowbrick.pdf

[62] Manual of Procedures for the ENTSO-E Central Information Transparency Platform, Version 2.1, 12 December 2016, retrieved from https://eepublicdownloads.entsoe.eu/clean-documents/mc-documents/transparency-platform/MOP/00_ENTSO-E%20Manual%20of%20Procedures_V2R1.pdf

[63] Tavenard, Romain & Faouzi, Johann & Vandewiele, Gilles & Divo, Felix & Androz, Guillaume & Holtz, Chester & Payne, Marie & Yurchak, Roman & Rußwurm, Marc & Kolar, Kushal & Woods, Eli. (2020). Tslearn, A Machine Learning Toolkit for Time Series Data.

[64] Mitchell, T. (1997). Machine Learning, McGraw Hill, Machine Learning, McGraw Hill, p.2

[65] Leonard Kaufman; Peter J. Rousseeuw (1990). Finding groups in data : An introduction to cluster analysis. Hoboken, NJ: Wiley-Interscience. p. 87. doi:10.1002/9780470316801

[66] Bertsimas, D., Dunn, J. Optimal classification trees. Mach Learn 106, 1039–1082 (2017). https://doi.org/10.1007/s10994-017-5633-9

[67] Wang S., Ren W., Zhang Y., Liang F. (2019) Random Forest Classifier for Distributed Multi-plant Order Allocation. In: Huang G., Chien CF., Dou R. (eds) Proceeding of the 24th International Conference on Industrial Engineering and Engineering Management 2018. Springer, Singapore. https://doi.org/10.1007/978-981-13-3402-3_14

[68] Keogh, E., Lin, J., and Truppel,W. 2003. Clustering of time series subsequences is meaningless: Implications for past and future research. In Proc. of the 3rd IEEE International Conference on Data Mining, Melbourne, FL, USA, pp. 115–122.

[69] Keogh, E., Kasetty, S. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. *Data Mining and Knowledge Discovery* **7,** 349–371 (2003). https://doi.org/10.1023/A:1024988512476

[70] D. J. Weller-Fahy, B. J. Borghetti and A. A. Sodemann, "A Survey of Distance and Similarity Measures Used Within Network Intrusion Anomaly Detection," in *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 70-91, Firstquarter 2015, doi: 10.1109/COMST.2014.2336610.

[71] Zheng A., Casari A., *Feature Engineering for Machine Learning*, 2018, O'Reilly Media, Inc., ISBN: 9781491953242

[72] Ozdemir S., Susarla D., *Feature Engineering Made Easy*, 2018, Packt Publishing, ISBN: 9781787287600

[73] Duboue, P. (2020). *The Art of Feature Engineering: Essentials for Machine Learning*. Cambridge: Cambridge University Press. doi:10.1017/9781108671682

[74] Benesty J., Chen J., Huang Y., Cohen I. (2009) Pearson Correlation Coefficient. In: Noise Reduction in Speech Processing. Springer Topics in Signal Processing, vol 2. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00296-0_5

[75] Krzywinski, M., Altman, N. Visualizing samples with box plots. *Nat Methods* **11,** 119–120 (2014). https://doi.org/10.1038/nmeth.2813

[76] Salgado, Cátia & Azevedo, Carlos & Manuel Proença, Hugo & Vieira, Susana. (2016). Noise Versus Outliers. 10.1007/978-3-319-43742-2_14.

[77] Weisberg, Sanford. (2001). Yeo-Johnson Power Transformations.

[78] Atkinson, Anthony C. (2020) The box-cox transformation: review and extensions. Statistical Science. ISSN 0883-4237

[79] Berrar, Daniel. (2018). Cross-Validation. 10.1016/B978-0-12-809633-8.20349-X.

[80] Xinchuan Zeng & Tony R. Martinez (2000) Distributionbalanced stratified cross-validation for accuracy estimation, Journal of Experimental & Theoretical Artificial Intelligence, 12:1, 1-12, DOI: 10.1080/095281300146272

[81] Zhang Z., Hancock E.R. (2011) Mutual Information Criteria for Feature Selection. In: Pelillo M., Hancock E.R. (eds) Similarity-Based Pattern Recognition. SIMBAD 2011. Lecture Notes in Computer Science, vol 7005. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-24471-1_17

[82] Max Khun, Kjell Johnson, Feature Engineering And Selection : A Practical Approach for Predictive Models, 2020, pp. 314, ISBN : 9781032090856

[83] Staffell, I., & Pfenninger, S. (2018). The increasing impact of weather on electricity supply and demand. Energy, 145, 65–78. doi:10.1016/j.energy.2017.12.051

[84] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations

[85] Cortes, C., & Vapnik, V. (2004). Support-Vector Networks. *Machine Learning, 20*, 273-297.

[86] Silverman, B. W., and M. C. Jones. "E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951)." *International Statistical Review / Revue Internationale De Statistique* 57, no. 3 (1989): 233-38. Accessed June 21, 2021. doi:10.2307/1403796.