ΕΡΓΑΣΤΗΡΙΟ ΘΕΡΜΙΚΩΝ ΣΤΡΟΒΙΛΟΜΗΧΑΝΩΝ
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΑΝΑΠΤΥΞΗ ΚΑΙ ΕΦΑΡΜΟΓΗ ΜΕΘΟΔΩΝ CLUSTERING ΣΤΗ ΔΙΑΓΝΩΣΤΙΚΗ ΑΕΡΙΟΣΤΡΟΒΙΛΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΦΛΟΥΡΗΣ ΝΙΚΟΛΑΟΣ

Επίβλεψη:
Επ. Καθηγητής Ν. Αρετάκης

ΑΘΗΝΑ
Δεκέμβριος, 2019

ΠΡΟΛΟΓΟΣ

Από τη θέση αυτή, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Νικόλαο Αρετάκη, αρχικά για την ανάθεση ενός τόσου ενδιαφέροντος θέματος, αλλά και για την καθοριστική συμβολή και βοήθεια του για την εκπόνηση της παρούσας διπλωματικής εργασίας. Επίσης, θα ήθελα να ευχαριστήσω τον κ. Αλέξιο Αλεξίου ο οποίος βοήθησε στην διαπίστευση της εγκυρότητας και την συμπλήρωση ατελειών της εργασίας μου. Οι συμβουλές, η καθοδήγηση και η υποστήριξη τους υπήρξαν πολύ σημαντικές καθ' όλη τη διάρκεια, και για αυτό τους ευχαριστώ πραγματικά.

ΠΕΡΙΛΗΨΗ

Σκοπός της παρούσας διπλωματικής εργασίας είναι η μελέτη των μεθόδων διάγνωσης που χρησιμοποιούνται ευρέως, η διαλογή των καταλληλότερων για αεριοστρόβιλους, η υλοποίηση τους σε προγραμματιστικό περιβάλλον και η σύγκριση μεταξύ τους .

Πρώτα, πραγματοποιήθηκε εκτενής βιβλιογραφική ανασκόπηση σχετικά με τις διαδεδομένες μεθόδους διάγνωσης που χρησιμοποιούνται τον 21ο αιώνα. Γίνεται αναφορά σε διάφορες δημοφιλείς μεθόδους που έχουν εφαρμογή σε πολλούς τομείς όπως η ανίχνευση: διαδικτυακής εισβολής, τραπεζικής ή ασφαλιστικής απάτης καθώς και βιομηχανικής ζημιάς σε μονάδες εργοστασίων. Ειδικότερα, δίνεται βάση στις μεθόδους που αφορούν ανίχνευση βλάβης σε αεριοστρόβιλους.

Στη συνέχεια, μετά από εφαρμογή κριτηρίων υλοποίησης του συνόλου των μεθόδων, διακρίνονται μερικές από αυτές για ανάπτυξη και σύγκριση. Η διάκριση των μεθόδων αυτών γίνεται μέσω διαδικασίας επικύρωσης όπου ελέγχεται η διαγνωστική ικανότητα κάθε μεθόδου. Αφού γίνει διαλογή των καταλληλότερων μεθόδων, αναλύεται η φιλοσοφία που κρύβεται πίσω από κάθε μέθοδο καθώς και τα μεγέθη που την χαρακτηρίζουν. Αναφέρονται οι παραδοχές λειτουργίας κάθε μεθόδου, το υπολογιστικό κόστος και η διαδικασία λειτουργίας της. Παράλληλα, γίνεται αναφορά στα είδη συναρτήσεων απόστασης που χρησιμοποιήθηκαν ή δε χρησιμοποιήθηκαν για τις μεθόδους που αναπτύχθηκαν καθώς και τα κριτήρια επιλογής αυτών.

Έπειτα, περιγράφεται το φαινόμενο υποβάθμισης των κινητήρων, ένα πολύ συχνό φαινόμενο στη λειτουργία αεριοστρόβιλων. Απαριθμούνται τα αίτια και οι μορφές πρόκλησης του φαινομένου, οι παραδοχές που υποτέθηκαν για την ανάλυση, καθώς και στόχοι που πρέπει να πετύχει η κάθε μέθοδος ώστε να θεωρείται αξιόπιστη.

Ακολουθούν οι τρόποι χρήσης των μεθόδων μέσω των δύο λειτουργιών που αναπτύχθηκαν για αυτές. Η πρώτη λειτουργία, η διαγνωστική, περιλαμβάνει συλλογή, αδιαστατοποίηση, και ομαδοποίηση των δεδομένων σε clusters ώστε να διακρίνεται η υγιής από την εσφαλμένη λειτουργία. Η δεύτερη λειτουργία, η λειτουργία ομαδοποίησης προφίλ θερμοκρασιών, περιλαμβάνει συλλογή πραγματικών δεδομένων για τρία μοντέλα στροβιλοκινητήρων, διαχείριση και επεξεργασία των μετρήσεων, εφαρμογή των μεθόδων και ομαδοποίηση των προφίλ αναφοράς (υπογραφών) σε μέσα προφίλ για λειτουργία σε μεγάλο εύρος ισχύ. Έτσι αναπτύσσονται δύο αλγόριθμοι για κάθε μέθοδο.

Τελικά, γίνεται ανακεφαλαίωση και σύγκριση των μεθόδων μέσω της απόδοσης στις μελέτες που έλαβαν χώρα. Προκύπτουν συμπεράσματα σχετικά με το πόσο κατάλληλες είναι οι μέθοδοι στο πεδίο των θερμικών στροβιλοκινητήρων, την αξιοπιστία αυτών και τη συμβολή τους στη λύση των προβλημάτων που τέθηκαν. Κλείνοντας, γίνεται αναφορά σε μελλοντικές προτάσεις επέκτασης της μελέτης αυτής, μέσω βελτιστοποίησης των ήδη υπάρχοντών μεθόδων ή μέσω υιοθέτησης νέων.

ABSTRACT

The purpose of this thesis is to study the widely used diagnostic methods, to select the most suitable for gas turbines, to implement them in a programming environment and to compare them.

First, there was an extensive bibliographic review of the widespread diagnostic methods used in the 21st century. Reference is made to various popular methods that are applied in many areas such as: Network invasion, banking or insurance fraud as well as industrial damage to factory units. In particular, analysis is focused on methods based on gas turbine failure detection.

Then, following the implementation criteria of all the methods, some of them are distinguished for development and comparison. These methods are distinguished by a validation process that checks the diagnostic capability of each method. After selecting the most appropriate methods, the philosophy behind each method and the sizes that characterize it are analysed. The operating assumptions of each method, its computational cost and its operating procedure are reported. At the same time, reference is made to the types of metric distances used or not used for the methods developed and their selection criteria.

Next, the phenomenon of engine degradation is described, a very common phenomenon in gas turbine operation. They list the causes and forms of the phenomenon, the assumptions assumed for the analysis, and the goals that each method must achieve in order to be considered reliable.

The ways to use the methods are presented through two functions. In addition to diagnosing engine failures, the methods are engineered in such a way that another objective is achieved. The first function involves collecting, dimensioning, and clustering data into Clusters to distinguish healthy from malfunctioning status. The second function, the temperature profile clustering function, involves collecting real data for three turbine engine models, managing and processing the measurements, applying the methods that are selected and grouping the reference profiles (signatures) into profile profiles for wide range operation. Thus for each method there are two algorithms that are developed.

Finally, methods are summarized and compared through performance in the studies that took place. Conclusions are drawn on the suitability of the methods in the field of thermal turbines, their reliability and their contribution to the solution of the problems raised. In conclusion, reference is made to future proposals for extension of this study, either by optimizing existing methods or by adopting new ones.

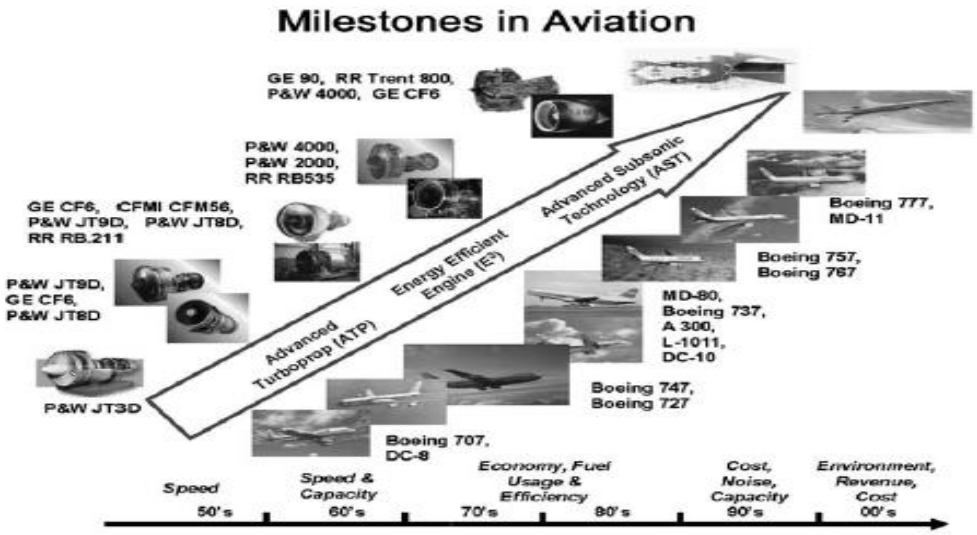
Περιεχόμενα

1 ΕΙΣΑΓΩΓΗ

Σε αυτό το κεφάλαιο γίνεται περιγραφή της βιβλιογραφικής έρευνας που πραγματοποιήθηκε. Αναλύεται η σημαντικότητα της επιστήμης της Διαγνωστικής για το παρόν και το μέλλον και εξετάζονται οι τρόποι εφαρμογής της. Δίνεται ιδιαίτερη προσοχή στην παρακολούθηση και τη συντήρηση των στροβιλοκινητήρων και στους τρόπους εφαρμογής της Διαγνωστικής σε αυτούς.

1.1 Εισαγωγικά στοιχεία

Η εύρεση ανωμαλιών είναι ένα αρκετά σημαντικό πρόβλημα που έχει ερευνηθεί σε διάφορους ερευνητικούς τομείς και πεδία εφαρμογών. Πολλές τεχνικές ανίχνευσης ανωμαλιών έχουν αναπτυχθεί ειδικά για συγκεκριμένη εφαρμογή, ενώ άλλες τεχνικές είναι πιο γενικές. Η επιστήμη της Διαγνωστικής ειδικότερα, είναι ένας κλάδος που αναπτύσσεται όλο και περισσότερο τα τελευταία χρόνια. Η ανάγκη της πρόληψης και της διάγνωσης είναι πλέον απαραίτητη ανεξαρτήτως αν πρόκειται για επίγειες ή εναέριες εφαρμογές. Η πρώτη αφορμή ανάπτυξης μίας τέτοιας επιστήμης καταγράφεται στα τέλη της δεκαετίας του 1940 όπου εμφανίστηκαν τα πρώτα πολιτικά αεροσκάφη. Η μαζική μεταφορά ατόμων από το ένα αεροδρόμιο στο άλλο είχε ως πρώτο στόχο την ασφάλεια των πολιτών αυτών. Ήταν αναγκαίο όλες οι παράμετροι υγείας των μηχανών των αεροσκαφών να είναι εντός επιτρεπτών ορίων προκειμένου να πραγματοποιηθεί πτήση. Αν και σε πρώιμο στάδιο, η Διαγνωστική σαν επιστήμη εμπλουτιζόταν με δεδομένα, στοιχεία και μοτίβα που κατοχυρώνουν την ασφάλεια. Μερικές δεκαετίες αργότερα, η ασφάλεια στα αεροσκάφη ήταν πλέον δεδομένη και ο τομέας της Διαγνωστικής στράφηκε στη βελτιστοποίηση της απόδοσης, την εξοικονόμηση καυσίμου και τη γενικότερα οικονομικότερη αποστολή αεροσκάφους. Πλέον χρησιμοποιούνται και επίγειες εφαρμογές στον τομέα της ηλεκτροπαραγωγής. Όπως είναι λογικό, τα δεδομένα αλλάζουν ή εμπλουτίζονται, νέες παράμετροι γίνονται κρίσιμες προς παρακολούθηση, νέες τεχνικές αναπτύσσονται για την ακριβέστερη διάγνωση. Η Διαγνωστική εξελίσσεται. Φτάνοντας στο σήμερα, πέρα από τους τομείς – πυλώνες που απασχολούσαν το μισό προηγούμενο αιώνα τη Διαγνωστική, προστίθενται και η εκπομπή ρύπων και τα ποσοστά παραγωγής ήχου. Είναι λοιπόν χρήσιμο να εξεταστούν οι κυριότερες μέθοδοι που χρησιμοποιούνται, να αναλυθούν μερικές από αυτές και μέσω της σύγκρισης τους να εξαχθούν αποτελέσματα σχετικά με τη χρησιμότητα, την ακρίβεια και τη διαγνωστική τους ικανότητα.

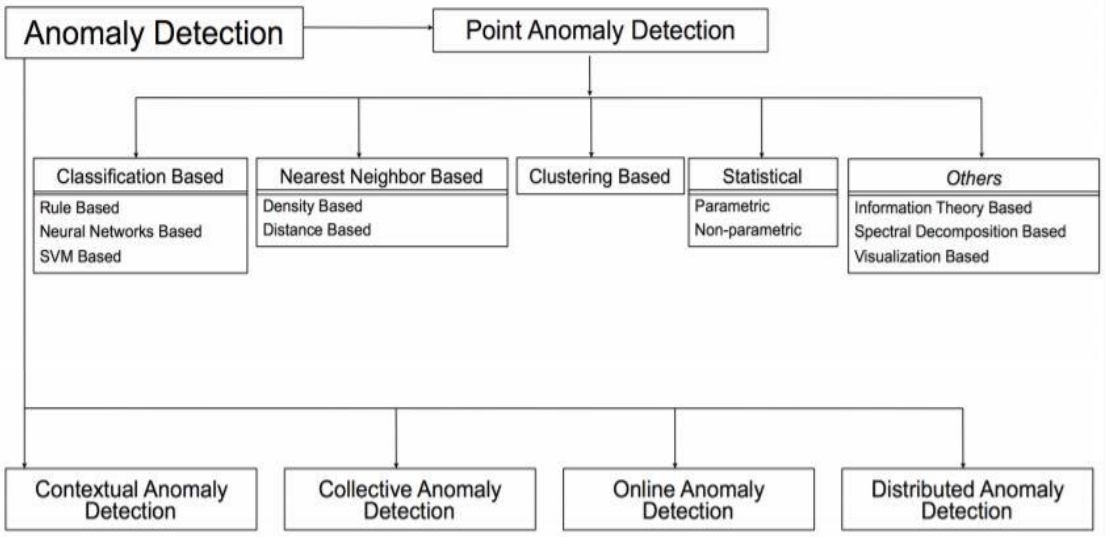


Σχήμα 1.1: Στόχοι αεροπορίας σε συνάρτηση με το χρόνο

Το σχήμα 1.1 βοηθά στην καλύτερη απεικόνιση όσων περιεγράφηκαν νωρίτερα για τους εναέριους αεριοστρόβιλους. Μέσω της έρευνας που ακολουθεί, γίνεται ανασκόπηση στις μεθόδους εύρεσης ανωμαλιών που εφαρμόζονται ευρέως σήμερα, επιλογή και σύγκριση των καταλληλότερων αυτών για τον τομέα της Διαγνωστικής αεριοστρόβιλων και τελικά πρόταση μεθοδολογίας βελτιστοποίησης αυτών.

1.2 Βιβλιογραφική ανασκόπηση

Οι μέθοδοι διάγνωσης ποικίλουν ανάλογα με το πεδίο εφαρμογής. Ο κύριος τρόπος διαχωρισμού τους είναι μέσω του είδους ανωμαλιών που ανιχνεύουν. Όπως φαίνεται στο σχήμα 1.2 η ανίχνευση ανωμαλιών χωρίζεται σε: Σημειακή, Συλλογική, Κατανεμημένη και Ανίχνευση βάσει γειτονικών δεδομένων.



Σχήμα 1.2: Είδη ανίχνευσης σημειακών ανωμαλιών σε σχέση με άλλα είδη ανίχνευσης

Στην παρούσα διπλωματική εργασία αναλύονται οι εφαρμογές των μεθόδων πρόγνωσης στους αεριοστρόβιλους οπότε δίνεται ιδιαίτερη βάση στη σημειακή και συλλογική μόνο ανίχνευση ανωμαλιών. Σε αυτούς τους τομείς ανίχνευσης ανωμαλιών οι δημοφιλέστερες άρα και πιο επιτυχημένες μέθοδοι είναι οι εξής 5:

1. K-means Clustering
2. DBSCAN (Μέθοδος ομαδοποίησης δεδομένων σε εφαρμογές με θόρυβο)
3. Mean-Shift Clustering (Μοντέλο μέσης μετατόπισης)
4. Agglomerative Hierarchical Clustering (Αθροιστική ιεραρχική μέθοδος)
5. Expectation-Maximization algorithm (Μοντέλα μίξης Gauss Ομαδοποίησης-Μεγιστοποίησης)

Από αυτές τις πέντε μεθόδους, οι τρείς κρίθηκαν ικανές προς ανάπτυξη. Η μέθοδος μέσης μετατόπισης, αν και είναι εξαιρετική στην ταχύτητα και την ακρίβεια αποτελεσμάτων για μία μεταβλητή, κατά την ανάλυση της φαίνεται να υστερεί σε σχέση με τις άλλες μεθόδους σε δεδομένα πολλών διαστάσεων (μεταβλητών). Παρόμοιο πρόβλημα αντιμετώπισε και η μέθοδος μοντέλων μίξης Gauss. Η μέθοδος αυτή χρησιμοποιεί διάφορες κατανομές Gauss που αντιστοιχίζονται στο πλήθος δεδομένων που της δόθηκαν για διάγνωση παράγοντας έτσι ένα μοντέλο πιθανοτήτων για το χαρακτηρισμό κάθε νέου σημείου ως υγιές ή βλαβερό. Ενώ το πρόβλημα των πολλαπλών διαστάσεων αντιμετωπίζεται από την ικανότητα προσαρμογής του μοντέλου Gauss μέσω επιπλέον κατανομών, για μεταβλητές (άρα και διαστάσεις) μεγαλύτερες των δέκα το μοντέλο Gauss χάνει την αξιοπιστία και την ακρίβεια του.

Οι μέθοδοι λοιπόν που αναλύονται παρακάτω είναι η K-means η DBSCAN και η AHC. Για τις μεθόδους αυτές γίνεται εκτενής περιγραφή, ανάπτυξη για δύο διαφορετικούς σκοπούς διάγνωσης και σύγκριση μεταξύ τους ώστε να διακριθεί η πιο αποτελεσματική.

Η διαδικασία αυτή εμπλουτίζεται με τη συμβολή διάφορων συναρτήσεων απόστασης που μπορούν να χρησιμοποιηθούν και θα αναλυθούν παρακάτω. Παράλληλα, αναλύεται το πρόβλημα της υποβάθμισης λειτουργίας επίγειων αλλά και εναέριων αεριοστρόβιλων και γίνεται προσπάθεια συνδυασμού διάγνωσης και υποβάθμισης ώστε να υπάρχει εικόνα από ένα πληρέστερο μοντέλο.

1.3 Δομή εργασίας

Η δομή της εργασίας έχει ως εξής:

Στο πρώτο κεφάλαιο, πραγματοποιήθηκε αναλυτική βιβλιογραφική ανασκόπηση σχετικά με δημοφιλείς και μη μεθόδους εύρεσης ανωμαλιών είτε στην διαγνωστική στροβιλοκινητήρων είτε σε άλλα επιστημονικά πεδία. Αναφέρονται οι πιθανές εφαρμογές, οι προϋποθέσεις λειτουργίας, ο υπολογιστικός χρόνος, τα θετικά και τα αρνητικά κάθε μίας από τις μεθόδους. Γίνεται σύγκριση μεταξύ των μεθόδων, με τελικό σκοπό τη διαλογή των καταλληλότερων από αυτές για ανάπτυξη και εφαρμογή σε πραγματικές μετρήσεις αεριοστρόβιλων.

Στο δεύτερο κεφάλαιο, αφού επιλέχθηκαν οι μέθοδοι που κρίνονται ιδανικές για διαγνωστική λειτουργία, ακολουθεί η ανάπτυξη της. Αρχικά αναφέρεται η ιστορία και η φιλοσοφία κάθε μεθόδου, ενώ στη συνέχεια περιγράφονται τα βήματα λειτουργίας, το υπολογιστικό κόστος, οι χαρακτηριστικές παράμετροι και πλεονεκτήματα ή μειονεκτήματα της μεθόδου έναντι των άλλων μεθόδων που επιλέχτηκαν. Αναφέρονται οι συναρτήσεις μέτρησης που μπορούν να χρησιμοποιηθούν και οι συναρτήσεις που τελικά επιλέχθηκαν και ακολουθεί μία διαδικασία επικύρωσης των μεθόδων που υλοποιείται μέσω δύο τεστ.

Στο τρίτο κεφάλαιο, αναλύεται η πρώτη λειτουργία των μεθόδων, η δημιουργία προφίλ αναφοράς θερμοκρασίας εξόδου καυσαερίων. Η διαδικασία αυτή βοηθά στην παραγωγή ενός σετ πρότυπων καταστάσεων για την υγιή λειτουργία της μηχανής σε όλο το εύρος λειτουργίας της. Αναπτύσσεται αλγόριθμος για κάθε μέθοδο ο οποίος συνοδεύεται από διάγραμμα ροής των αποφάσεων που θα πρέπει να πάρει ο αλγόριθμος για την κατάταξη των δεδομένων σε clusters με εκτενή εξήγηση των βημάτων του. Στη συνέχεια, εφαρμόζεται ο αλγόριθμος αυτός σε πραγματικά δεδομένα τριών αεριοστρόβιλων. Τα αποτελέσματα συγκρίνονται με συμβατικούς αλγόριθμο που χρησιμοποιείται σήμερα.

Στο τέταρτο κεφάλαιο, αναλύεται η δεύτερη λειτουργία των μεθόδων, η διάγνωση βλαβών αισθητήρων. Παράγονται υπογραφές αναφοράς που περιγράφουν την υγιή κατάσταση αλλά και διάφορους τύπους εσφαλμένης λειτουργίας. Αναπτύσσεται επιπλέον αλγόριθμος αφαίρεσης υποβάθμισης ώστε να εξασφαλισθεί ότι η σύγκριση των μετρήσεων με τις υπογραφές αναφοράς θα είναι άμεση και ίση. Ελέγχεται η διαγνωστική ικανότητα σε προσομοιωμένα δεδομένα απότομης ή σταδιακής μεταβολής και εφαρμόζεται η μέθοδος σε πραγματικά δεδομένα τριών αεριοστροβίλων.

Στο πέμπτο κεφάλαιο, ανακεφαλαιώνονται τα βήματα που ακολουθήθηκαν συνολικά στην εργασία. Προκύπτουν συμπεράσματα για τις μελέτες που εκπονήθηκαν επισημαίνοντας τα πλεονεκτήματα και τα μειονεκτήματα κάθε μεθόδου καθώς και συμπεράσματα για τις πραγματικές μετρήσεις που αναλύθηκαν. Ακολουθεί αναφορά σε μελλοντική δουλειά που δύναται να επεκτείνει την ανάλυση αυτή ή να δημιουργήσει νέα σημεία ενδιαφέροντος στη μελέτη αεριοστρόβιλων. Προτείνονται νέες μέθοδοι προς ανάπτυξη ή τρόποι βελτιστοποίησης των ήδη υλοποιημένων αλγορίθμων που θα συνεισφέρουν σημαντικά στη μελέτη της Διαγνωστικής.

2 Μέθοδοι Clustering

Στο κεφάλαιο αυτό αναλύονται οι μέθοδοι που αναπτύχθηκαν για τη διάγνωση βλαβών και τη δημιουργία προφίλ αναφοράς. Αναφέρεται η ιστορία, ο τρόπος ανάπτυξης, τα χαρακτηριστικά μεγέθη και το υπολογιστικό κόστος κάθε μίας μεθόδου. Γίνεται επίσης αναφορά στις συναρτήσεις απόστασης που χρησιμοποιήθηκαν αλλά και σε άλλες συναρτήσεις που υπάρχουν.

2.1 Γενικά

Οι μέθοδοι clustering χρησιμοποιούνται ήδη σε διάφορα επιστημονικά πεδία για ποικίλους σκοπούς. Στόχος αυτής της εργασίας είναι να διερευνηθεί η απόδοση τους στον τομέα της Διαγνωστικής και να συγκριθούν τα αποτελέσματα τους με τις μεθόδους που χρησιμοποιούνται στη Διαγνωστική ως τώρα. Παρακάτω παρουσιάζονται οι πιο διαδεδομένες μέθοδοι clustering. Οι μέθοδοι clustering που χρησιμοποιούνται στη διαγνωστική αλλά και σε άλλα επιστημονικά πεδία είναι η εξής:

1. Density Based Spatial Clustering of Applications with Noise (DBSCAN)
2. K-means Clustering
3. Agglomerative Hierarchical Clustering (AHC)
4. Μέθοδος προσέγγισης μεγιστοποίησης μέσω Γκαουσιανών μοντέλων μίξης (GMM)

Από αυτές αναπτύχθηκαν οι πρώτες τρεις καθώς η GMM έχει ελλιπή ικανότητα διαχείρισης δεδομένων πολλών διαστάσεων. Λαμβάνοντας υπόψη ότι οι μέθοδοι που επιλέχθηκαν θα πρέπει να επεξεργαστούν δεδομένα δεκάδων διαστάσεων λόγω των εξεταζόμενων μετρήσεων που τα περιγράφουν η μέθοδος αυτή απορρίφθηκε. Για τις υπόλοιπες μεθόδους γίνεται περιγραφή της αρχής λειτουργίας τους και των χαρακτηριστικών τους.

2.2 DBSCAN

2.2.1 Γενικά

Η μέθοδος DBSCAN (Density Based Spatial Clustering of Applications with Noise) ήταν η πρώτη μέθοδος που αναπτύχθηκε για αυτήν την εργασία. Είναι μία μέθοδος

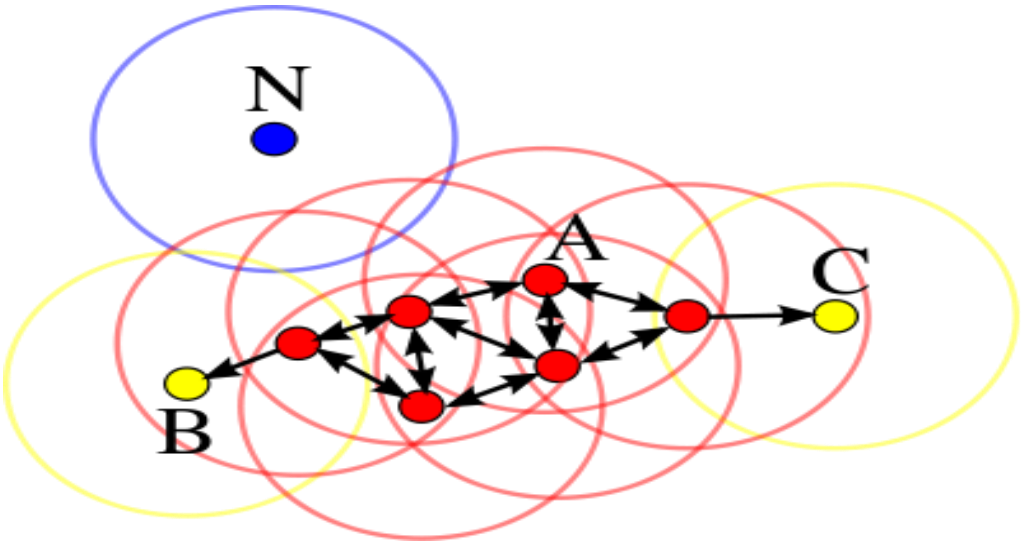
ομαδοποίησης δεδομένων που προτάθηκε από τους Martin Ester, Hans-Peter Kriegel, Jörg Sander και Xiaowei Xu το 1996. Πρόκειται για μία μέθοδο που βασίζεται στην πυκνότητα των δεδομένων, επικεντρώνεται δηλαδή σε τόπους με πολλά γειτονικά σημεία, ανεξαρτήτως θορύβου. Η μέθοδος αυτή, χρησιμοποιεί δύο χαρακτηριστικά μεγέθη για να κατατάξει κάθε γραμμή δεδομένων σε μία κατηγορία. Τα μεγέθη αυτά είναι το *έψιλον*, το οποίο περιγράφει τη μέγιστη απόσταση που μπορούν να έχουν δύο σημεία ώστε να θεωρούνται γείτονες, και το *MinPts*, το οποίο περιγράφει τον ελάχιστο αριθμό σημείων-γειτόνων (δηλαδή σημεία με απόσταση μεταξύ τους μικρότερη του *έψιλον*) που μπορούν να αποτελέσουν ένα cluster-γειτονιά. Πιο συγκεκριμένα, αν σκεφτούμε ότι έχουμε ένα σύνολο σημείων σε κάποιο χώρο με οποιοδήποτε αριθμό διαστάσεων, μπορούμε μέσω αυτών των δύο μεγεθών να κατατάξουμε όλα τα σημεία του χώρου σε: α.) Σημεία πυρήνα, β.) Άμεσα προσβάσιμα σημεία και γ.) Ακραίες τιμές ή αλλιώς Σημεία θορύβου.

- Ένα σημείο *A* είναι ένα *σημείο πυρήνα* εάν τουλάχιστον *MinPts* αριθμός σημείων βρίσκονται εντός της απόστασης *έψιλον* (συμπεριλαμβανομένου του *A*).
- Ένα σημείο *B* είναι *άμεσα προσβάσιμο* από το *A* εάν το σημείο *B* βρίσκεται μέσα στην απόσταση *έψιλον* από το σημείο πυρήνα *A* . Τα σημεία αυτά λέγονται μόνο *άμεσα προσβάσιμα* και είναι διαφορετικά από τα βασικά σημεία ακόμα και αν συμπεριλαμβάνονται στο cluster-γειτονιά.
- Ένα σημείο *B* είναι *προσβάσιμο* από το *p* αν υπάρχει διαδρομή A_1 έως A_n με $A_1 = A$ και $A_n = B$, όπου κάθε A_{i+1} είναι *άμεσα προσβάσιμο* από το A_i. Σημειώστε ότι αυτό σημαίνει ότι όλα τα σημεία της διαδρομής πρέπει να είναι πυρήνα, με την πιθανή εξαίρεση του *B*.
- Όλοι οι βαθμοί που δεν είναι *προσβάσιμοι* από οποιοδήποτε άλλο σημείο είναι *ακραίες τιμές ή αλλιώς σημεία θορύβου* .

Αν το *A* είναι ένα σημείο πυρήνα, τότε σχηματίζει ένα *σύμπλεγμα* μαζί με όλα τα σημεία (πυρήνα ή μη πυρήνα) που είναι *προσβάσιμα* από αυτό. Κάθε ομάδα περιλαμβάνει τουλάχιστον ένα σημείο πυρήνα. τα μη πυρήνα σημεία μπορεί να είναι μέρος ενός συμπλέγματος, αλλά σχηματίζουν το "άκρο" του, αφού δεν μπορούν να χρησιμοποιηθούν για να φτάσουν σε περισσότερα σημεία.

Στο παρακάτω διάγραμμα παρατηρείται μία ομαδοποίηση με τη μέθοδο DBSCAN. Το χαρακτηριστικό μέγεθος MinPts παίρνει την τιμή τέσσερα (MinPts=4), δηλαδή για να ονομαστεί μία γειτονιά-cluster χρειάζονται μόλις 4 σημεία που θεωρούνται γείτονες μεταξύ τους. Το χαρακτηριστικό μέγεθος *έψιλον* δε θα μας απασχολήσει σε αυτό το διάγραμμα, παρόλα αυτά εκφράζεται με το δακτύλιο που περικλείει κάθε σημείο οριοθετώντας έτσι τη μέγιστη δυνατή ακτίνα για να βρεθεί γείτονας με το εξεταζόμενο σημείο. Παρατηρούμε λοιπόν, ότι όλα τα κόκκινα σημεία γύρω από το εξεταζόμενο σημείο Α βρίσκονται σε απόσταση μικρότερη από το μέγεθος *έψιλον* ενώ παράλληλα πληρούν τις προδιαγραφές για να θεωρούνται cluster-γειτονιά, επειδή ο αριθμός τους είναι μεγαλύτερος του 4. Για αυτό λοιπόν, τα σημεία αυτά ονομάζονται *σημεία πυρήνα.* Με παρόμοια λογική, τα σημεία Β και C δεν είναι σημεία πυρήνα, αλλά είναι *προσβάσιμα σημεία* από το Α (και τα υπόλοιπα σημεία πυρήνα), και για αυτό ανήκουν και αυτά στο cluster-γειτονιά. Τέλος, το σημείο Ν είναι *σημείο θορύβου* γιατί δεν είναι προσβάσιμο από

σημείο πυρήνα ώστε να θεωρείται *προσβάσιμο σημείο* ούτε «συνδέει» κάποιο άλλο σημείο με τη γειτονιά ώστε να θεωρείται σημείο πυρήνα.



Σχήμα 2.1: Σημεία που θεωρούνται ή δε θεωρούνται γείτονες βάσει της DBSCAN

Η δυνατότητα προσέγγισης δεν είναι συμμετρική, αφού εξ ορισμού, κανένα σημείο δεν μπορεί να είναι προσβάσιμο από ένα μη κεντρικό σημείο, ανεξάρτητα από την απόσταση (έτσι μπορεί να είναι προσβάσιμο ένα σημείο μη πυρήνα, αλλά δεν μπορεί να επιτευχθεί τίποτε από αυτό). Επομένως, χρειάζεται μια περαιτέρω έννοια της *συνδεσιμότητας* για να καθορίσει τυπικά την έκταση των ομάδων που βρέθηκαν από το DBSCAN. Δύο σημεία *A* και *B* συνδέονται με πυκνότητα αν υπάρχει ένα σημείο *O* έτσι ώστε τόσο το *A* όσο και το *B* να είναι προσβάσιμα από *O*.

Ένα σύμπλεγμα ικανοποιεί δύο ιδιότητες:

1. Όλα τα σημεία μέσα στο σύμπλεγμα είναι αμοιβαία συνδεδεμένα μέσω της πυκνότητας.
2. Εάν ένα σημείο είναι προσεγγίσιμο μέσω πυκνότητας από οποιοδήποτε σημείο του συμπλέγματος, είναι και αυτό μέρος του συμπλέγματος.

2.2.2 Αλγόριθμος

Οι αλγόριθμοι που παράχθηκαν για αυτήν την μέθοδο έχουν παρόμοια βάση και αρχή λειτουργίας, παρόλα αυτά υπάρχουν σημαντικές διαφορές στα δεδομένα που ζητούν από το χρήστη καθώς και τα αποτελέσματα που εξάγουν. Η αρχή λειτουργίας του κάθε αλγόριθμου λοιπόν, ξεκινά εισάγοντας τα δεδομένα των χρονοσειρών που αποθηκεύονται σε υπολογιστικό φύλλο Excel. Τα δεδομένα αυτά έχουν τη μορφή Πίνακα όπου οι γραμμές αφορούν τις διαφορετικές χρονικές στιγμές ενώ οι στήλες αφορούν τις διαστάσεις του σημείου δηλαδή τις διάφορες μετρήσεις που υπάρχουν για τη δεδομένη χρονική στιγμή. Έπειτα, διαλέγεται ένα τυχαίο σημείο εκκίνησης. Ελέγχεται αν υπάρχουν άλλα σημεία κοντά στο εξεταζόμενο σημείο υπολογίζοντας τις αποστάσεις όλων των σημείων μεταξύ

τους σε ένα μητρώο *n* * *n* διαστάσεων, όπου *n* ο αριθμός των σημείων που εισήχθησαν. Η συνάρτηση που παράγει αυτό το μητρώο στο προγραμματιστικό περιβάλλον της MATLAB ονομάζεται pdist2. Η συνάρτηση αυτή μπορεί να βρίσκει την Ευκλείδεια απόσταση μεταξύ σημείων οσοδήποτε διαστάσεων. Έτσι, μπορεί κανείς να βρει τους γείτονες του εξεταζόμενου σημείου και να το κατατάξει σε σημείο με γείτονες ή σημείο θορύβου. Τα σημεία-γείτονες αργότερα αν πληρούν την προϋπόθεση των ελάχιστων σημείων (MinPts) μπορούν να καταχωρηθούν ως σημεία πυρήνα. Αντίστοιχα, αν συγκεντρωθούν αρκετά σημεία θορύβου σε έναν δεδομένο χώρο, μπορούν να δημιουργήσουν μία δική τους γειτονιά. Αν ένα σημείο Α, δεν πληροί της προϋποθέσεις (ακτίνας) για να είναι γείτονας με ένα σημείο Β, αλλά πληροί αυτές τις προϋποθέσεις όμως με ένα τρίτο σημείο Γ, το οποίο έχει καταχωρηθεί νωρίτερα ως γείτονας του σημείου Β, τότε το σημείο Β είναι γείτονας του σημείου Α, ανεξάρτητα με το αν είναι άμεσα συνδεδεμένα σα σημεία. Τέλος, με αντίστοιχη λογική, ένα σημείο που ανήκει σε μία γειτονιά Χ, αν αργότερα βρεθεί ότι ανήκει και σε μία άλλη γειτονιά Ψ, τότε οι δύο αυτές γειτονιές γίνονται μία. Δεν υπάρχει δηλαδή περίπτωση ένα σημείο να ανήκει σε δύο διαφορετικά cluster-γειτονιές ταυτόχρονα. Ο αλγόριθμος συνεχίζει να κατατάσσει σημεία, μέχρι να μην υπάρχουν πλέον σημεία χωρίς ταυτότητα.

Ο αλγόριθμος DBSCAN επισκέπτεται κάθε σημείο της βάσης δεδομένων, πιθανώς πολλαπλές φορές (π.χ. ως υποψήφιο σε διαφορετικά clusters). Για πρακτικούς λόγους ωστόσο, η πολυπλοκότητα του χρόνου εξαρτάται κατά κύριο λόγο από τον αριθμό των επικλήσεων της συνάρτησης RegionQuery, της συνάρτησης δηλαδή που βρίσκει τους άμεσους γείτονες ενός σημείου. Ο DBSCAN επικαλείται της συνάρτησης ακριβώς μία φορά για κάθε σημείο και εάν χρησιμοποιείται μια δομή ευρετηρίου που εκτελεί τη συνάρτηση αυτή σε O(log(n)) , λαμβάνεται μια συνολική μέση πολυπλοκότητα χρόνου εκτέλεσης O(n*log(n)) (εάν η παράμετρος επιλέγεται στο με τρόπο ουσιαστικό, δηλαδή ότι κατά μέσο όρο μόνο O(log(n)) τα σημεία επιστρέφονται). Χωρίς τη χρήση δομής επιταχυνόμενου δείκτη ή με εκφυλισμένα δεδομένα (π.χ. όλα τα σημεία σε απόσταση μικρότερη από έψιλον), η δυσκολότερη χρονική πολυπλοκότητα παραμένει $O(n^2)$. Το μητρώο των αποστάσεων μεγέθους $(n^2-n)/2$ (επειδή είναι συμμετρικό) μπορεί να υλοποιηθεί για να αποφευχθούν οι αναπροσαρμογές απόστασης, αλλά αυτό χρειάζεται μνήμη $O(n^2)$, ενώ μια μη βασιζόμενη σε μητρώο αποστάσεων υλοποίηση του DBSCAN χρειάζεται μόνο μνήμη O(n).

Ένας ψευδοκώδικας που περιγράφει επιγραμματικά τα βήματα του αλγορίθμου που περιεγράφηκαν παραπάνω φαίνεται στο επόμενο σχήμα:

```
DBSCAN(DB, distFunc, eps, minPts) {
    C = 0                                                   /* Cluster counter */
    for each point P in database DB {
        if label(P) ≠ undefined then continue               /* Previously processed in inner loop */
        Neighbors N = RangeQuery(DB, distFunc, P, eps)      /* Find neighbors */
        if |N| < minPts then {                              /* Density check */
            label(P) = Noise                                /* Label as Noise */
            continue
        }
        C = C + 1                                           /* next cluster label */
        label(P) = C                                        /* Label initial point */
        Seed set S = N \ {P}                                /* Neighbors to expand */
        for each point Q in S {                             /* Process every seed point */
            if label(Q) = Noise then label(Q) = C           /* Change Noise to border point */
            if label(Q) ≠ undefined then continue           /* Previously processed */
            label(Q) = C                                    /* Label neighbor */
            Neighbors N = RangeQuery(DB, distFunc, Q, eps) /* Find neighbors */
            if |N| ≥ minPts then {                          /* Density check */
                S = S U N                                   /* Add new neighbors to seed set */
            }
        }
    }
}
```

Σχήμα 2.2: Ψευδοκώδικας αλγορίθμου DBSCAN

2.2.3 Δομή και προετοιμασία δεδομένων

Τα δεδομένα πρέπει να είναι αριθμητικό πίνακα με:

- Σειρές που αντιπροσωπεύουν παρατηρήσεις (άτομα)
- Στήλες που αντιπροσωπεύουν μεταβλητές

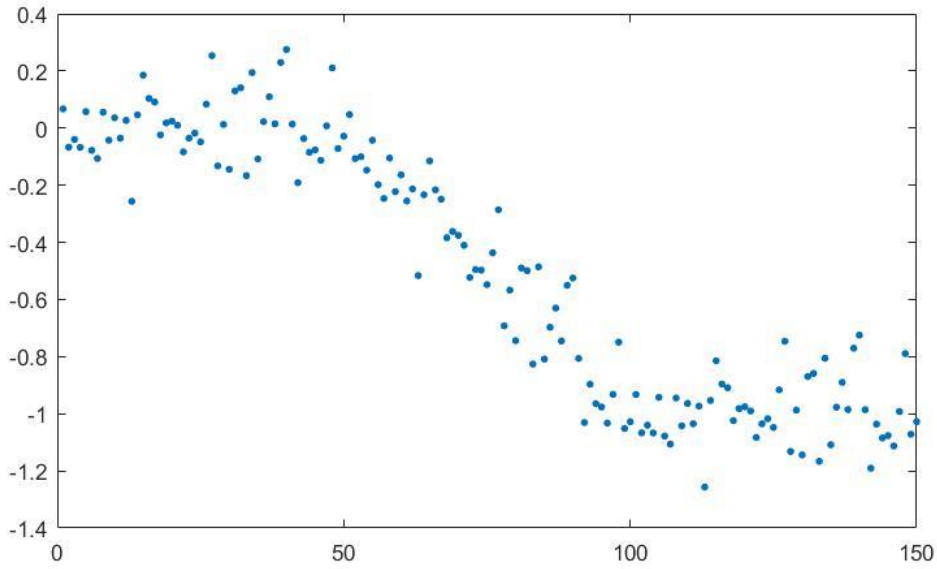
2.2.4 Κριτήρια επιλογής χαρακτηριστικών μεγεθών

Όπως είναι λογικό, κάθε ομάδα δεδομένων είναι διαφορετική από την προηγούμενη. Για αυτό το λόγο είναι σημαντικό να οριστούν κάποιες γενικές αρχές καθορισμού των χαρακτηριστικών μεγεθών της μεθόδου DBSCAN, δηλαδή του *έψιλον* και του *MinPts*. Για να επιλέξουμε καλές τιμές για τις παραμέτρους πρέπει να κατανοήσουμε πώς χρησιμοποιούνται και να έχουν τουλάχιστον μια βασική προηγούμενη γνώση σχετικά με το σύνολο δεδομένων που χρησιμοποιείται.

Έψιλον

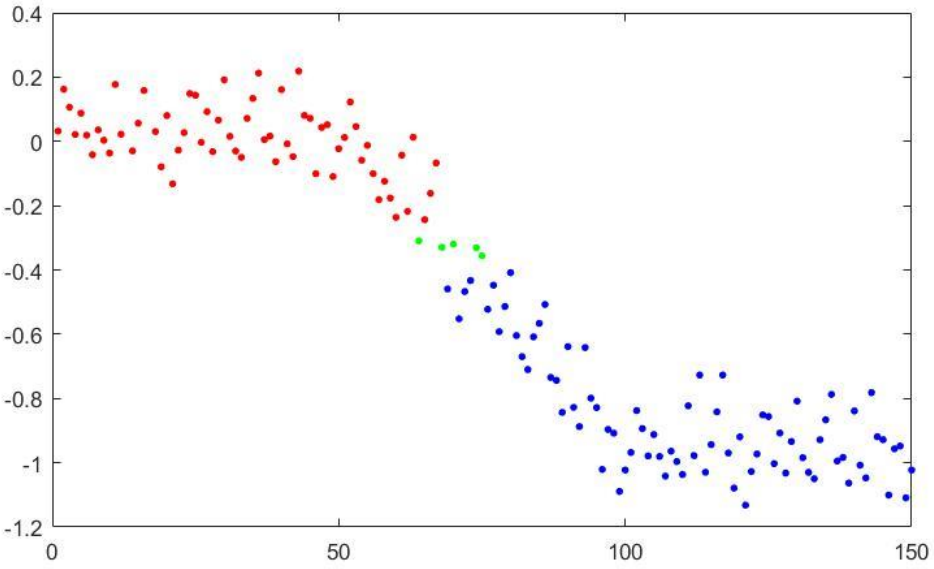
Το *έψιλον* όπως αναφέρθηκε παραπάνω αντιπροσωπεύει την μέγιστη δυνατή ακτίνα που μπορεί να βρίσκεται ένα ή παραπάνω σημεία για να θεωρηθεί γείτονας του εξεταζόμενου σημείου. Αν η επιλεγμένη τιμή *έψιλον* είναι πολύ μικρή, ένα μεγάλο μέρος των δεδομένων δεν θα ομαδοποιηθεί και θα καταταχθεί ως θόρυβος. Θα θεωρηθούν ως υπερβολικές τιμές επειδή δεν ικανοποιούν τον αριθμό (MinPts) των σημείων για να

δημιουργήσουν μια πυκνή γειτονιά. Από την άλλη πλευρά, εάν η επιλεγμένη τιμή για το *έψιλον* είναι πολύ υψηλή, τα clusters θα συγχωνευθούν και η πλειοψηφία των δεδομένων θα ανήκει στο ίδιο σύμπλεγμα. Το *έψιλον* θα πρέπει να επιλέγεται με βάση την απόσταση του συνόλου δεδομένων. Για να υπολογίσουμε αυτήν την απόσταση, μπορούμε να χρησιμοποιήσουμε ένα γράφημα απόστασης-K, αλλά γενικά είναι προτιμότερες οι μικρές τιμές *έψιλον*. Εισάγονται λοιπόν στο προγραμματιστικό περιβάλλον της MATLAB, δεδομένα που αφορούν μία χρονοσειρά και έχουν την εξής μορφή:

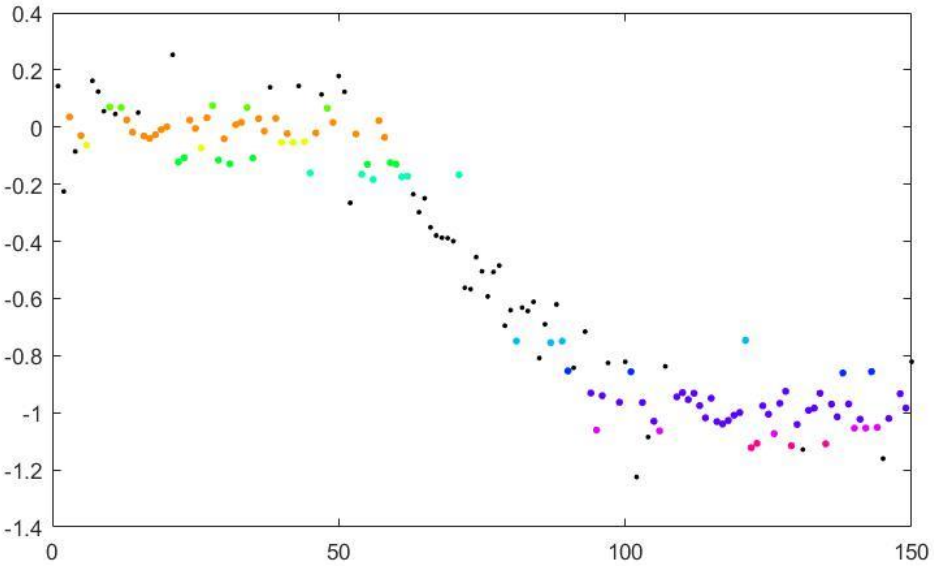


Σχήμα 2.3: Διάγραμμα μεταβολής χρονοσειράς που περιγράφει σταδιακή πτώση ενός συντελεστή

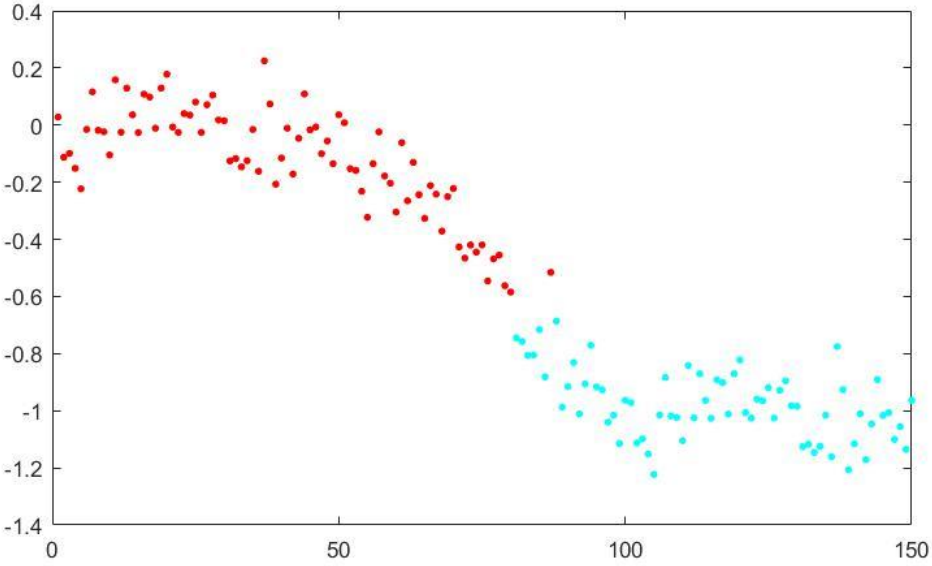
Παρατηρείται ότι το διάγραμμα αυτό περιγράφει 150 μετρήσεις μίας εξεταζόμενης μεταβλητής βάσει του οριζόντιου άξονα. Φαίνεται επίσης ότι οι τιμές της μεταβλητής αυτής είναι φυσιολογικές μέχρι περίπου τη μέτρηση 50. Μετά τη μέτρηση 50 ακολουθεί σταδιακή πτώση του εξεταζόμενου μεγέθους μέχρι το -1% όπου σταθεροποιείται. Προφανώς μία τέτοια μεταβολή συνιστά ότι υπάρχει βλάβη στα εξεταζόμενα δεδομένα. Σκοπός ανάλυσης των παραπάνω δεδομένων είναι η εύρεση της απόδοσης του αλγορίθμου DBSCAN ρυθμίζοντας διαφορετικά κάθε φορά τις μεταβλητές εισόδου. Τα παρακάτω διαγράμματα βοηθούν στην περαιτέρω κατανόηση της εξεταζόμενης μεθόδου καθώς και την κρισιμότητα επιλογής σωστών μεταβλητών εισόδου για αυτήν:



Σχήμα 2.4: Αποτελέσματα DBSCAN για έψιλον=0.05 και MinPts=5



Σχήμα 2.5: Αποτελέσματα DBSCAN για έψιλον=0.01 και MinPts=5



Σχήμα 2.6: Αποτελέσματα DBSCAN για έψιλον=0.1 και MinPts=5

Στα παραπάνω διαγράμματα παρατηρείται ότι η μεταβλητή εισόδου *MinPts* παραμένει σταθερή. Αυτό συμβαίνει επειδή τα εξεταζόμενα δεδομένα απαρτίζονται από μία μόνο χρονοσειρά οπότε είναι μονοδιάστατα. Αν υπήρχαν παραπάνω παρατηρούμενες μεταβλητές τότε θα είχε περισσότερο νόημα να αλλάξει αυτή η μεταβλητή. Το μείζον ενδιαφέρον όμως επικεντρώνεται στο *έψιλον*.

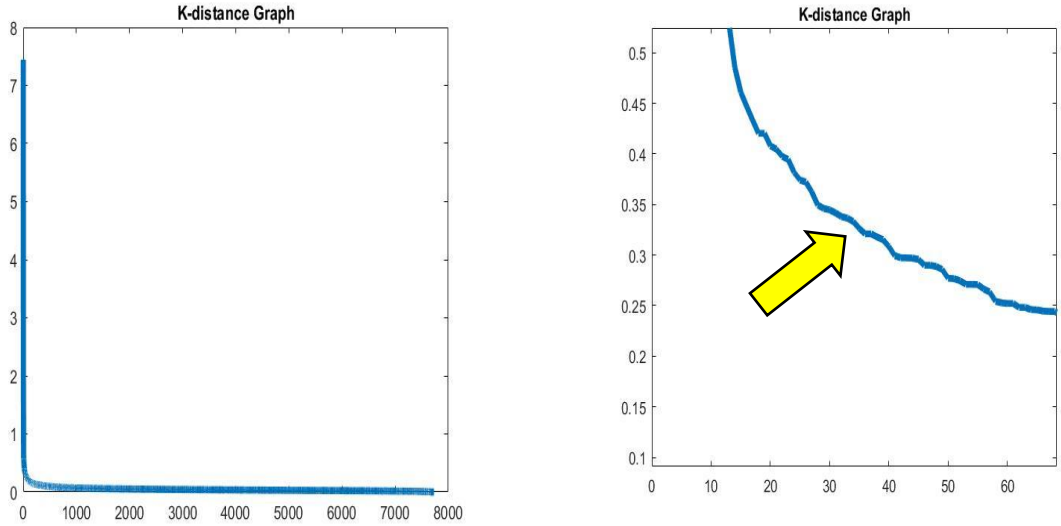
Με διαφορετικό χρώμα απεικονίζονται στα διαγράμματα τα διαφορετικά clusters. Είναι προφανές ότι για μείωση του μεγέθους *έψιλον* ο συνολικός αριθμός clusters αυξάνεται (όπως φαίνεται συγκρίνοντας τα Διαγράμματα 2.4 και 2.5) καθώς μειώνονται οι πιθανοί γείτονες κάθε εξεταζόμενου σημείου. Επίσης, μειώνεται και η πιθανότητα ομαδοποίησης – συγχώνευσης 2 clusters μεταξύ τους μέσω της αρχής που ορίζει ότι αν ένα σημείο ανήκει σε δύο clusters ταυτόχρονα, αυτά τα δύο clusters συγχωνεύονται. Επίσης στο Σχήμα 2.5 παρατηρούνται και κάποιες μαύρες μικρότερες τελείες. Αυτές αφορούν δεδομένα που ο αλγόριθμος δεν ομαδοποίησε σε κανένα cluster και τα ονόμασε σημεία θορύβου. Τα σημεία θορύβου μεταβάλλονται κατά κανόνα αντιστρόφως ανάλογα με το *έψιλον*. Το ιδανικό αποτέλεσμα του αλγορίθμου θα ήταν να χωρίσει τα δεδομένα σε 2 clusters όπως και έκανε στο Σχήμα 2.6 και στην Τρίτη επιλογή του συντελεστή *έψιλον*. Μπορεί λοιπόν η γενική αρχή να ορίζει ότι η επιλογή μικρού *έψιλον* βοηθά στη μείωση της πιθανότητας ομαδοποίησης διαφορετικών clusters μεταξύ τους, από την άλλη πλευρά όμως μεγάλο *έψιλον* μπορεί να οδηγήσει σε μείωση αποδοτικότητας όταν ο στόχος της μεθόδου είναι να διαχωριστεί η κατάσταση υγείας από την κατάσταση βλαβών αλλά και να διαχωριστούν οι πιθανές βλάβες μεταξύ τους. Η επιλογή του έψιλον λοιπόν εξαρτάται απαραίτητα από τους σκοπούς του χρήστη σχετικά με τα δεδομένα.

Τι γίνεται όμως όταν ο χρήστης έχει στα χέρια του ένα σετ δεδομένων άγνωστου περιεχομένου; Τότε χρησιμοποιείται ένα γράφημα απόστασης-Κ για την εύρεση της τιμής του *έψιλον*. Το γράφημα απόστασης-Κ είναι μία καμπύλη που στον οριζόντιο άξονα έχει

τους δείκτες όλων των σημείων του εξεταζόμενου σετ δεδομένων ενώ στον κατακόρυφο έχει την απόσταση του κάθε σημείου με τον K-κοντινότερο γείτονά του. Το K συνήθως έχει τιμή που ορίζεται από τη σχέση:

$$K = MinPts - 1 \quad (2.1)$$

Το γράφημα αυτό έχει μορφή όμοια με της συνάρτησης y=1/x με χαμηλότερο βαθμό καμπυλότητας λόγω των διακριτών τιμών K που αντιστοιχεί σε κάθε σημείο. Ένα τέτοιο διάγραμμα παρίσταται παρακάτω όπου βλέπουμε 7000 μετρήσεις του σετ δεδομένων «deltas» που θα αναλυθούν ξανά στη συνέχεια. Δίπλα σε αυτό βρίσκεται μία μεγέθυνση του ίδιου διαγράμματος για να διακρίνουμε την κορυφή (elbow) που θα ορίσει το *έψιλον*. Λαμβάνοντας το σημείο που δείχνει το κίτρινο βέλος στο παρακάτω σχήμα τραβάμε οριζόντια γραμμή μέχρι τον κατακόρυφο άξονα και πλέον μπορούμε να βρούμε τη μέση απόσταση σημείων μεταξύ τους ώστε να θεωρούνται γείτονες, ή με άλλα λόγια το *έψιλον*. Έτσι λοιπόν, γίνεται με ασφάλεια να ορισθεί το *έψιλον* για δεδομένα άγνωστης προέλευσης. Η μεθοδολογία αυτή λειτουργεί και σε δεδομένα που περιέχουν «θόρυβο» καθώς το επιθυμητό μέγεθος είναι η μέση απόσταση, και άρα ο θόρυβος απαλείφεται χωρίς να επηρεάσει το *έψιλον*. Στο σχήμα 2.7 φαίνεται η διαδικασία που μόλις περιεγράφηκε.



Σχήμα 2.7: Μεταβολή της μέγιστης απόστασης του K-γείτονα συναρτήσει του μεγέθους MinPts

MinPts

Για την επιλογή της παραμέτρου *MinPts* ένας εμπειρικός κανόνας είναι να προκύψει από τον αριθμό των διαστάσεων D (στήλες) στο σύνολο των δεδομένων. Η σχέση που μας δίνει τον παραπάνω κανόνα είναι:

$$MinPts \geq D + 1 \quad (2.2)$$

Η χαμηλή τιμή του *MinPts* = 1 δεν έχει νόημα, καθώς κάθε σημείο θα αποτελεί ένα σύμπλεγμα (cluster). Με *MinPts* ≤ 2, το αποτέλεσμα θα είναι το ίδιο με την ιεραρχική ομαδοποίηση την οποία θα δούμε παρακάτω. Συνεπώς, ο *MinPts* πρέπει να επιλεγούν

τουλάχιστον 3. Ωστόσο, οι μεγαλύτερες τιμές είναι συνήθως καλύτερες για σύνολα δεδομένων με θόρυβο και θα αποδώσουν πιο σημαντικές ομάδες. Κατά κανόνα, *MinPts* = 2 * *dim* μπορεί να χρησιμοποιηθεί αλλά μπορεί να χρειαστεί να επιλεχθούν μεγαλύτερες τιμές για πολύ μεγάλα δεδομένα, για θορυβώδη δεδομένα ή για δεδομένα που περιέχουν πολλά αντίγραφα.

2.2.5 Συνάρτηση απόστασης

Για όλες τις μεθόδους αυτής της εργασίας χρησιμοποιήθηκε η Ευκλείδεια Απόσταση και ο Συντελεστής Αλληλοσυσχέτισης. Για κάθε τρέξιμο του κώδικα, είτε με τη μία συνάρτηση απόστασης είτε με την άλλη, το *έψιλον* πρέπει να ρυθμίζεται ανάλογα. Όπως είδαμε παραπάνω η τιμή του *έψιλον* συνδέεται άρρηκτα με τις τιμές των αποστάσεων των σημείων, άρα με αλλαγή των αποστάσεων είναι λογικό να επακολουθήσει αλλαγή του *έψιλον*. Περισσότερες πληροφορίες για τις συναρτήσεις απόστασης υπάρχουν στο κεφάλαιο 2.5.

2.3 K-means

2.3.1 Γενικά

Η K-means όπως και η DBSCAN είναι μέθοδος ανάλυσης σημάτων χρονοσειρών και ομαδοποίησης σε cluster – γειτονιές. Η K-means είναι μία από τις πιο διάσημες μεθόδους clustering στον τομέα εξόρυξης δεδομένων (data mining). Η κύρια διαφορά με την DBSCAN είναι ότι η K-means χρειάζεται ως δεδομένο τον αριθμό των clusters που θα σχηματιστούν. Όπως και η DBSCAN, η K-means είναι μία μέθοδος unsupervised machine learning, δηλαδή δεν χρειάζεται προηγούμενα δεδομένα εκμάθησης για να καταχωρήσει σε ομάδες τα νέα δεδομένα. Ο Andrey Bu, ο οποίος έχει περισσότερα από 5 χρόνια εμπειρίας στο machine learning και διδάσκει σήμερα τις ικανότητές του, λέει ότι "ο στόχος της K-means είναι απλός: ομαδοποίηση παρόμοιων δεδομένων και συγκέντρωση των υποκειμενικών μοτίβων. Για να επιτευχθεί αυτός ο στόχος, η K-means αναζητά έναν σταθερό αριθμό (*K*) ομάδων σε ένα σύνολο δεδομένων". Ο αριθμός K ορίζεται ως ο αριθμός των κέντρων των clusters στα οποία ομαδοποιούνται όλα τα εξεταζόμενα σημεία. Αντίθετα με την DBSCAN, σε αυτήν την μέθοδο δεν υπάρχει πιθανότητα εκχώρησης δεδομένων ως θόρυβο, παρά μόνο σε cluster που φαίνεται οπτικά ότι είναι ομάδα ανωμαλιών. Η K-means είναι μία επαναληπτική μέθοδος, δηλαδή τα βήματα του αλγορίθμου της επαναλαμβάνονται μέσω μίας συνθήκης η οποία εξετάζεται μετά από κάθε επανάληψη ώστε να δούμε αν η συνθήκη αυτή συγκλίνει ώστε να τερματιστούν οι επαναλήψεις. Τα βήματα της επαναληπτικής διαδικασίας του αλγόριθμου θα αναλυθούν περαιτέρω παρακάτω. Όσων αφορά το κριτήριο σύγκλισης, χρησιμοποιήθηκαν δύο διαφορετικές παράμετροι. Η μία από αυτές τις παραμέτρους εξετάζει τον αριθμό επαναλήψεων, ώστε να περιορίζεται το υπολογιστικό κόστος, ενώ

η άλλη παράμετρος εξετάζει την απόσταση των κέντρων των clusters της $i-1$ επανάληψης με τα κέντρα της i επανάληψης. Για την πρώτη παράμετρο αρκεί να οριστεί ένας λογικός αριθμός επαναλήψεων ανά τρέξιμο του αλγόριθμου, τον οποίο όταν φτάνει το εν λόγω τρέξιμο, θα σταματά την λούπα που εκτελεί και θα παρουσιάζει τα αποτελέσματα που έχει τη δεδομένη στιγμή ανεξαρτήτως απόστασης των κέντρων. Για τη δεύτερη παράμετρο, ορίζεται μία μικρή απόσταση ως μέγεθος ανοχής (tolerance) την οποία συγκρίνουμε με το άθροισμα των ευκλείδειων αποστάσεων των κέντρων μεταξύ επαναλήψεων. Αντίστοιχα με την πρώτη παράμετρο, όταν το άθροισμα γίνει μικρότερο από την ανοχή που τέθηκε από τον χρήστη, τότε ο αλγόριθμος βγαίνει από τη λούπα που εκτελεί ανεξαρτήτως πάλι από τον αριθμό επαναλήψεων. Η ρύθμιση των δύο αυτών παραμέτρων χρειάζεται ιδιαίτερη προσοχή. Πολύ μικρός αριθμός επαναλήψεων ενδέχεται να παρουσιάσει αποτελέσματα ανακριβή ενώ πολύ μεγάλος αριθμός επαναλήψεων δύναται να καθυστερήσει χρονικά τη διάγνωση. Παράλληλα, η μεγάλη απόσταση ανοχής έχει παρόμοια προβλήματα με τον μικρό αριθμό επαναλήψεων λόγω του μεγάλου περιθωρίου τυχαίας διαλογής κέντρων κοντινά μεταξύ τους ικανοποιώντας έτσι το κριτήριο ανοχής.

2.3.2 Αλγόριθμος

Ο πιο συνηθισμένος αλγόριθμος χρησιμοποιεί μια επαναληπτική τεχνική βελτίωσης. Δεδομένου ότι ένα αρχικό σύνολο clusters k σημαίνει $m_1^{(1)}$, ..., $m_k^{(1)}$, ο αλγόριθμος επαναλαμβάνει τα εξής δύο βήματα:

- **Βήμα εκχώρησης** : Αντιστοίχιση κάθε παρατήρησης στη γειτονιά του οποίου ο μέσος όρος έχει την ελάχιστη τετραγωνική ευκλείδεια απόσταση , είναι διαισθητικά ο «πλησιέστερος» μέσος όρος. (Μαθηματικά, αυτό σημαίνει χωρισμό των παρατηρήσεων σύμφωνα με το διάγραμμα Voronoi που παράγεται από τις μέσες τιμές).

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k \right\} \tag{2.3}$$

Όπου η κάθε μία παρατήρηση x_p ανατίθεται σε ένα ακριβώς cluster $S^{(t)}$, ακόμη και αν μπορεί να ανατεθεί σε δύο ή περισσότερα από αυτά.

- **Βήμα ενημέρωσης** : Υπολογίζονται εκ νέου τα κέντρα των clusters των παρατηρήσεων. Αυτό γίνεται μέσω της σχέσης:

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j \tag{2.4}$$

Ο αλγόριθμος έχει συγκλίνει όταν οι αναθέσεις δεν αλλάζουν πλέον. Ο αλγόριθμος δεν εγγυάται την εύρεση του βέλτιστου. Ο αλγόριθμος αυτός συνήθως χρησιμοποιείται για την ανάθεση σημείων στο πλησιέστερο σύμπλεγμα με κριτήριο την απόσταση. Χρησιμοποιώντας μια διαφορετική συνάρτηση απόστασης εκτός από

(τετραγωνισμένη) Ευκλείδεια απόσταση μπορεί να σταματήσει ο αλγόριθμος να συγκλίνει. Για τους σκοπούς αυτής της εργασίας πέρα από την απόσταση αυτή χρησιμοποιήθηκε και ο Συντελεστής Αλληλοσυσχέτισης ως μέτρο απόστασης με παρόμοια αποτελέσματα.

Πολυπλοκότητα

Η εύρεση της βέλτιστης λύσης στο πρόβλημα της ομαδοποίησης *k* -means για παρατηρήσεις σε διαστάσεις *d* είναι:

- NP Ευκλείδειο διάστημα (των διαστάσεων *d*) ακόμη και για 2 clusters
- NP για έναν γενικό αριθμό ομάδων *k* ακόμα και στο επίπεδο,
- εάν τα *k* και *d* (η διάσταση) είναι σταθερά, το πρόβλημα μπορεί να λυθεί ακριβώς στο χρόνο $O(n^{dk+1})$, όπου *n* είναι ο αριθμός των οντοτήτων που θα συγκεντρωθούν.

Έτσι, γενικά χρησιμοποιούνται διάφοροι ευρετικοί αλγόριθμοι. Ο χρόνος εκτέλεσης του αλγόριθμου (και οι περισσότερες παραλλαγές) είναι $O(nkdi)$ όπου:

- *n* είναι ο αριθμός των *d*- διαστάσεων διανυσμάτων (που πρέπει να συγκεντρωθούν)
- *k* είναι ο αριθμός των clusters
- *i* είναι ο αριθμός των απαιτούμενων επαναλήψεων μέχρι τη σύγκλιση.

2.3.3 Δομή και προετοιμασία δεδομένων

Τα δεδομένα πρέπει να είναι αριθμητικό πίνακα με:

- Σειρές που αντιπροσωπεύουν παρατηρήσεις (άτομα)
- Στήλες που αντιπροσωπεύουν μεταβλητές

2.3.4 Κριτήρια επιλογής χαρακτηριστικών μεγεθών

Το κύριο χαρακτηριστικό μέγεθος που πρέπει να οριστεί είναι το K, δηλαδή ο αριθμός των clusters που θα σχηματιστούν. Το K καθορίζεται από το χρήστη του αλγορίθμου ανάλογα με τη λειτουργία που θέλει η K-means να επιτελεί. Για παράδειγμα, αν χρησιμοποιείται για διαγνωστικούς σκοπούς το K μπορεί να πάρει την τιμή δύο (K=2) ώστε να ξεχωρίσει σε ένα σετ μετρήσεων στροβιλοκινητήρα την υγιή λειτουργία από την ανώμαλη. Αν όμως χρησιμοποιείται για τη δημιουργία προφίλ αναφοράς για διάφορα εύρη Ισχύος του κινητήρα αυτού όπως θα δούμε παρακάτω, θα πρέπει να επιλεγεί ένα K μεγαλύτερο του δύο. Αυτό θα επιφέρει καλύτερη ομαδοποίηση των περιοχών Ισχύος χωρίς συγχώνευση τυχόν κοντινών αλλά διαφορετικών περιοχών. Μία τελευταία προς εξέταση περίπτωση είναι αυτή με K ίσο με τη μονάδα. Για K=1, όλα τα σημεία εν τέλει

ομαδοποιούνται σε ένα cluster. Αυτό που κάνει αυτήν την περίπτωση ιδιαίτερη είναι ο τρόπος ομαδοποίησης των σημείων σε αυτήν τη μία ομάδα. Ελέγχοντας τη διαδικασία αυτή, τα βήματα που ακολουθούνται είναι ίδια με αυτά της ιεραρχικής μεθόδου που θα αναλυθεί παρακάτω.

2.4 Agglomerative Hierarchical Clustering (AHC)

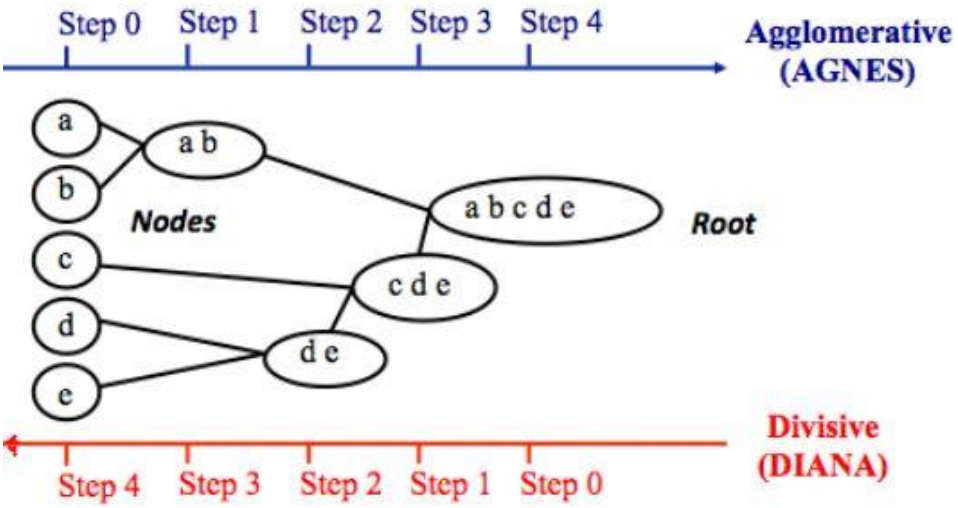
2.4.1 Γενικά

Η μέθοδος Αθροιστική Ιεραρχική μέθοδος ομαδοποίησης είναι η 3^{η} μέθοδος που θα εξεταστεί. Η αθροιστική ομαδοποίηση είναι ο πιο συνηθισμένος τρόπος ομαδοποίησης με την ιεραρχική μέθοδο λόγω της απλότητας και της ακρίβειας του. Ο αλγόριθμος ξεκινά θεωρώντας ότι κάθε ένα από τα δεδομένα που εισήχθησαν στο πρόγραμμα αποτελεί ένα cluster. Στη συνέχεια υπολογίζεται η απόσταση των σημείων μεταξύ τους σε ένα πίνακα D διαστάσεων $n * n$, όπου n ο αριθμός στηλών στον πίνακα δεδομένων, με άλλα λόγια ο αριθμός παρατηρήσεων στις εξεταζόμενες χρονοσειρές, ακριβώς όπως και στη DBSCAN. Ο πίνακας αυτός παίρνει ως τιμές την ευκλείδεια απόσταση των σημείων μεταξύ τους από προεπιλογή αλλά μπορεί να χρησιμοποιήσει και άλλου είδους αποστάσεις όπως αυτές που αναφέρθηκαν παραπάνω. Στη συνέχεια συγκρίνει τις αποστάσεις αυτές και συγχωνεύει τα clusters – σημεία που έχουν τη μικρότερη απόσταση μεταξύ τους. Κρατώντας ιστορικό των ενώσεων και των αποστάσεων συνεχίζει επαναληπτικά μέχρι να ομαδοποιηθούν όλα τα σημεία σε ένα cluster. Τότε παράγει ένα δενδρόγραμμα στο οποίο απεικονίζονται όλοι οι δεσμοί των σημείων μεταξύ τους καθώς και σε ποια στιγμή δημιουργήθηκαν. Τέλος, είναι στο χέρι του χρήστη να αποφασίσει πόσα cluster θεωρεί ότι χρειάζονται για να έχουμε ικανοποιητική ομαδοποίηση των δεδομένων (σαν την K-means) ή ποια είναι η μέγιστη απόσταση στην οποία ο αλγόριθμος θα σταματά να ομαδοποιεί προφανώς επειδή πάνω από αυτή τα σημεία δε θα έπρεπε να θεωρούνται γείτονες, και άρα να συγχωνευτούν (σαν την DBSCAN). Όποιο από τα δύο κριτήρια χρησιμοποιήσει ο χρήστης, θα εφαρμοστεί με τον ίδιο τρόπο. Ο αλγόριθμος αφού ομαδοποιήσει όλα τα σημεία σε ένα cluster, θα ψάξει να βρει σε ποιο σημείο της ομαδοποίησης ικανοποιείται το κριτήριο που δόθηκε από το χρήστη και θα εμφανίσει αποτελέσματα βάσει της συγκεκριμένης κατάστασης ομαδοποίησης. Είναι φανερό λοιπόν ότι αυτή η μέθοδος συνδυάζει χαρακτηριστικά και από τις δύο μεθόδους, δίνοντας έτσι στο χρήστη το πλεονέκτημα της ευελιξίας.

2.4.2 Αλγόριθμος

Η αθροιστική ομαδοποίηση λειτουργεί με τρόπο γνωστό και ως bottom up (από τη βάση προς τα πάνω). Δηλαδή, κάθε δεδομένο αρχικά θεωρείται ως cluster. Σε κάθε βήμα του αλγορίθμου, τα δύο clusters που είναι τα πιο παρόμοια συνδυάζονται σε ένα νέο μεγαλύτερο cluster (κόμβοι). Αυτή η διαδικασία επαναλαμβάνεται μέχρις ότου όλα τα σημεία είναι μέλη μόνο ενός μεγάλου συμπλέγματος (ρίζα). Έτσι λειτουργεί και

ένας από τους γνωστότερους αλγόριθμους τέτοιου είδους, ο AGNES. Στο παρακάτω Σχήμα 2.8 φαίνεται πως λειτουργεί ο AGNES αλλά και ένας αντίστροφος αλγόριθμος (top down) γνωστός με το όνομα DIANA.



Σχήμα 2.8: Τρόπος ομαδοποίησης σε ένα cluster με τη μέθοδο AHC

Το αντίστροφο της αθροιστικής ομαδοποίησης είναι η *διαιρετική ομαδοποίηση* , η οποία είναι επίσης γνωστή ως DIANA (*Divide Analysis*) και λειτουργεί κατά τρόπο top-down (από πάνω προς τα κάτω). Αρχίζει με τη ρίζα (root), στην οποία περιλαμβάνονται όλα τα αντικείμενα σε ένα μόνο σύμπλεγμα. Σε κάθε βήμα επανάληψης, το πιο ετερογενές σύμπλεγμα χωρίζεται σε δύο. Η διαδικασία επαναλαμβάνεται έως ότου όλα τα αντικείμενα βρίσκονται στη δική τους ομάδα. Και οι δύο αλγόριθμοι που περιεγράφηκαν ακολουθούν το μοτίβο που αναλύθηκε παραπάνω. Ο χρήστης οριοθετεί που θέλει να διακοπεί η ομαδοποίηση στην προκειμένη ή η διαίρεση στην περίπτωση του αλγόριθμου DIANA και προβάλλονται βάσει αυτής της ρύθμισης. Για την εφαρμογή αυτής της μεθόδου θα επικεντρωθούμε στην bottom – up εκδοχή της.

Βήματα για την αθροιστική ιεραρχική ομαδοποίηση

Ακολουθώντας τα παρακάτω βήματα εκτελείται η αθροιστική ιεραρχική ομαδοποίηση χρησιμοποιώντας το λογισμικό της MATLAB:

1. Προετοιμασία των δεδομένων
2. Υπολογίζεται η απόσταση ομοιότητας μεταξύ κάθε ζεύγους δεδομένων στο σύνολο δεδομένων.
3. Χρησιμοποιώντας τη λειτουργία συγχώνευσης για την ομαδοποίηση αντικειμένων σε ιεραρχική δομή δενδρογράμματος, με βάση τις πληροφορίες απόστασης που δημιουργούνται στο βήμα 2. Τα δεδομένα / clusters που βρίσκονται σε κοντινή απόσταση συνδέονται μεταξύ τους χρησιμοποιώντας τη λειτουργία σύνδεσης.

4. Προσδιορισμός «κοπής» στο ιεραρχικό δέντρο, και παρουσιάσει των ήδη διαμορφωμένων clusters.

Τα βήματα αυτά περιγράφονται παρακάτω.

Δομή και προετοιμασία δεδομένων

Τα δεδομένα πρέπει να είναι αριθμητικό πίνακα με:

- Σειρές που αντιπροσωπεύουν παρατηρήσεις (άτομα)
- Στήλες που αντιπροσωπεύουν μεταβλητές

Υπολογισμός απόστασης δεδομένων

Ο υπολογισμός απόστασης δεδομένων γίνεται με τρόπο ίδιο με αυτόν της μεθόδου DBSCAN. Το μητρώο D διαστάσεων $n * n$n * n περιέχει τις αποστάσεις όλων των διαφορετικών παρατηρήσεων (γραμμών) και είναι συμμετρικό. Όπως και στις προηγούμενες μεθόδους, χρησιμοποιείται ευκλείδεια απόσταση καθώς και συντελεστής αλληλοσυσχέτισης για τις αποστάσεις των δεδομένων.

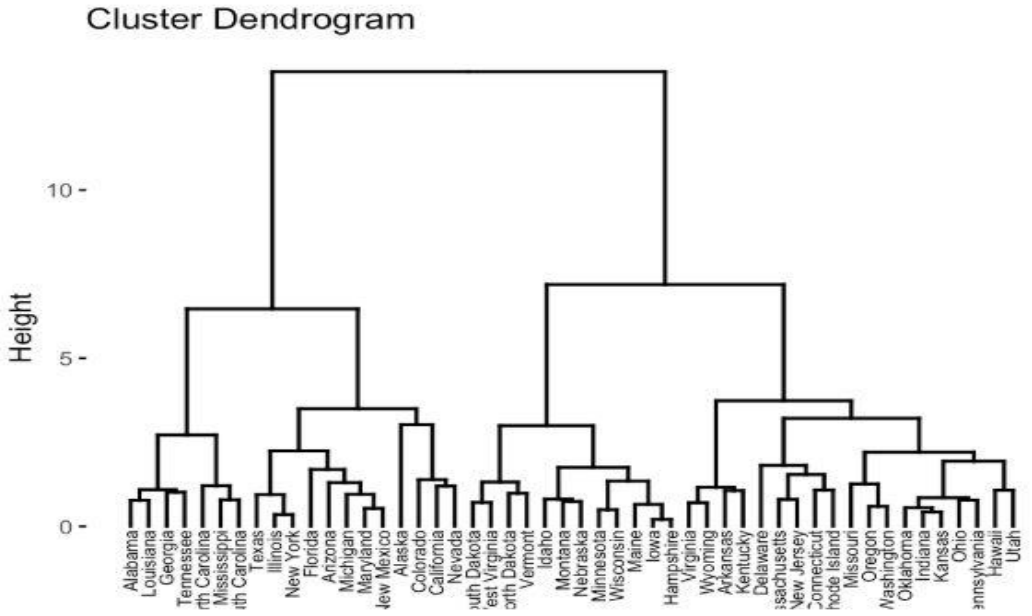
Λειτουργία συγχώνευσης – ομαδοποίησης

Όπως θα αναφερθεί και στο Κεφάλαιο Αποστάσεων, υπάρχουν διάφοροι τρόποι σύνδεσης δύο γειτονιών. Για την ανάλυση που επιτελείται σε αυτήν την εργασία θα χρειαστούν οι συνδέσεις ελάχιστης απόστασης και απόστασης κέντρων. Η πρώτη σύνδεση, ελάχιστης απόστασης, αφορά συνδέσεις clusters-σημείων (clusters που αποτελούνται από ένα σημείο) μεταξύ τους. Σε αυτήν την περίπτωση υπολογίζεται η απόσταση σημείων μεταξύ τους καθώς κάθε τέτοιο σημείο αποτελεί και το κέντρο του cluster. Η δεύτερη σύνδεση, απόστασης κέντρων, αφορά συνδέσεις clusters-σημείων με άλλα γειτονικά clusters (με αριθμό σημείων μεγαλύτερο της μονάδας), ή συνδέσεις μεταξύ γειτονικών clusters. Σε αυτήν την περίπτωση, η ζητούμενη απόσταση είναι αυτή του μεμονωμένου σημείου με κέντρο γειτονικού cluster ή των κέντρων γειτονικών clusters μεταξύ τους. Σε αυτό το σημείο αξίζει να σημειωθεί ότι το κέντρο ενός cluster υπολογίζεται όπως και στην K-means δηλαδή με την μέση τιμή για κάθε διάσταση (μέτρηση) των σημείων που ανήκουν στη γειτονιά αυτή.

Δενδρόγραμμα και διαδικασία κοπής

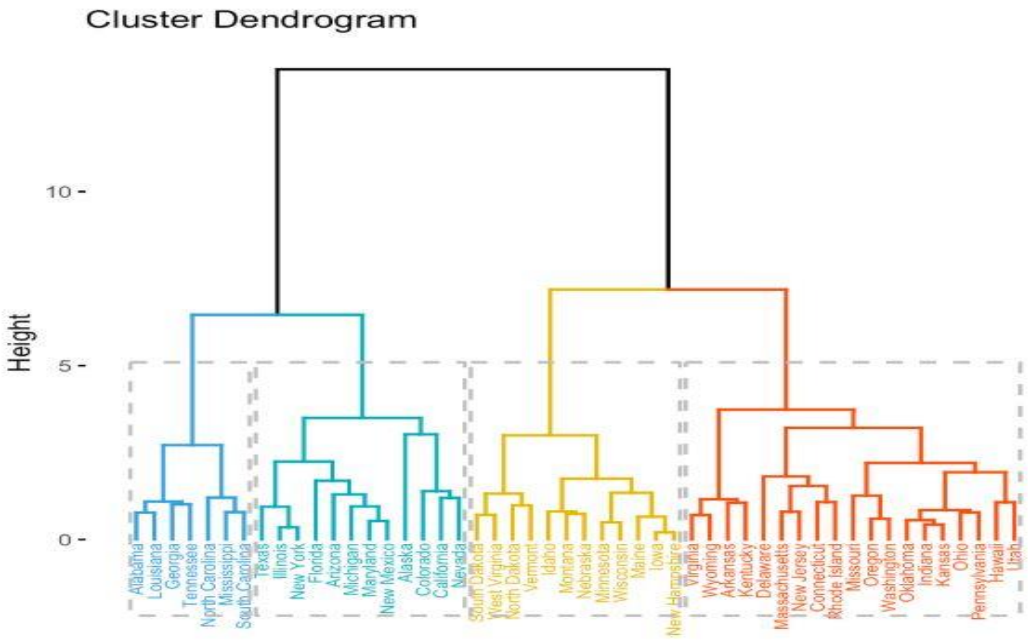
Ένα από τα προβλήματα με την ιεραρχική ομαδοποίηση είναι ότι δεν υπάρχει τρόπος αυτόματης επιλογής αριθμού clusters που θα σχηματιστούν ή πού πρέπει να κοπεί το δενδρόγραμμα για να σχηματιστούν clusters. Το πρόβλημα αυτό αντιμετωπίζεται με τον ίδιο τρόπο που αντιμετωπίζεται και στη μέθοδο K-means καθώς είναι στην ευχέρεια του χρήστη να κόψει το δενδρόγραμμα οπουδήποτε χρειαστεί ανάλογα με τη διαγνωστική λειτουργία που επιθυμεί.

Για παράδειγμα, στο δενδρόγραμμα που εμφανίζεται παρακάτω, κάθε «φύλλο» αντιστοιχεί σε ένα δεδομένο (γραμμή πίνακα). Καθώς ανεβαίνουμε το δέντρο, τα δεδομένα που είναι παρόμοια μεταξύ τους συνδυάζονται σε κλάδους, οι οποίοι είναι οι ίδιοι συγχωνευμένοι σε υψηλότερο ύψος.



Σχήμα 2.9: Δενδρόγραμμα σχηματισμού ενός Cluster από συνολικά 250 Clusters

Το ύψος της σύντηξης, που παρέχεται στον κάθετο άξονα, δείχνει την ομοιότητα / απόσταση μεταξύ δύο δεδομένων / clusters. Όσο υψηλότερο είναι το ύψος της σύντηξης, τόσο λιγότερο είναι τα αντικείμενα. Αυτό το ύψος είναι γνωστό ως η *συγγενική απόσταση* μεταξύ των δύο clusters και πήρε το όνομα του από την αρχική χρήση δενδρογραμμάτων, την αναπαράσταση γενεαλογικού δένδρου. Όπως ανεβαίνοντας σε ένα γενεαλογικό δένδρο είναι πιθανότερο δύο τυχαία διαλεγμένα άτομα να είναι άμεσοι συγγενείς, έτσι και στο δενδρόγραμμα, ανεβαίνοντας βρίσκουμε σημεία – παρατηρήσεις που είναι πιθανότερο να είναι γείτονες. Σημαντικό είναι να σημειωθεί ότι συμπεράσματα για τη συγγένεια δύο ή παραπάνω σημείων από ένα δενδρόγραμμα μπορούμε να βγάλουμε μόνο από το ύψος των διακλαδώσεων και όχι από το μήκος. Ο οριζόντιος άξονας δεν θεωρείται κριτήριο ομοιότητας.



Σχήμα 2.10: Δενδρόγραμμα σχηματισμού τριών Clusters από συνολικά 250 Clusters

Ομαδοποιώντας τα δεδομένα του παραπάνω παραδείγματος και κόβοντας το δενδρόγραμμα σε ύψος ίσο με 5 ή σε τελικό συνολικό αριθμό clusters ίσο με 4 παίρνουμε το παραπάνω σχήμα. Τα σημεία που ανήκουν στο cluster 1 έχουν κόκκινο χρώμα, στο cluster 2 έχουν κίτρινο, στο cluster 3 έχουν γαλάζιο ενώ στο 4 έχουν μπλε. Όπως φαίνεται η γραμμή που θα τραβηχτεί πρέπει να είναι οριζόντια, και κάθε φορά που συναντά κάθετα γραμμή του δενδρογράμματος σημαίνει ότι τα συνολικά clusters που θα σχηματιστούν θα αυξηθούν κατά ένα. Αντίστοιχα, όταν τα δεδομένα ομαδοποιούνται με κριτήριο το συνολικό αριθμό clusters, τότε ο αλγόριθμος αφού τελειώσει την ομαδοποίηση ψάχνει από το μικρότερο προς το μεγαλύτερο ύψος, ποιο είναι το ελάχιστο ύψος που ικανοποιεί αυτήν τη συνθήκη.

Επαλήθευση Δενδρογράμματος

Αφού συνδεθούν όλα τα δεδομένα σε ένα ιεραρχικό δένδρο – cluster, ίσως είναι χρήσιμο να επαληθευτούν τα «ύψη» των αποστάσεων που αναφέρθηκαν παραπάνω για να ελεγχθεί η ακρίβεια του αλγορίθμου. Ένας τρόπος για να μετρηθεί το πόσο καλά παράγεται το δένδρο – cluster είναι από τη *συνάρτηση hclust*(), η οποία υπολογίζει τη συσχέτιση μεταξύ των *γειτονικών* αποστάσεων και τα αρχικά δεδομένα απόστασης που παράγονται από το μητρώο D. Εάν η ομαδοποίηση είναι έγκυρη, η σύνδεση αντικειμένων στο δέντρο συμπλέγματος θα πρέπει να έχει ισχυρή συσχέτιση με τις αποστάσεις μεταξύ αντικειμένων στην αρχική μήτρα αποστάσεων. Όσο πιο κοντά η τιμή του συντελεστή συσχέτισης είναι στη μονάδα, τόσο πιο ακριβής η λύση συγκέντρωσης αντανακλά τα δεδομένα σας. Οι τιμές άνω του 0.95 θεωρούνται ότι είναι καλές. Η "μέση" μέθοδος σύνδεσης φαίνεται να παράγει υψηλές τιμές αυτού του στατιστικού στοιχείου. Αυτό μπορεί να είναι ένας λόγος που είναι τόσο δημοφιλής.

2.4.3 Δομή και προετοιμασία δεδομένων

Τα δεδομένα πρέπει να είναι αριθμητικό πίνακα με:

- Σειρές που αντιπροσωπεύουν παρατηρήσεις (άτομα)
- Στήλες που αντιπροσωπεύουν μεταβλητές

2.4.4 Κριτήρια επιλογής χαρακτηριστικών μεγεθών

Το κύριο χαρακτηριστικό μέγεθος που πρέπει να οριστεί είναι το K, δηλαδή ο αριθμός των clusters που θα σχηματιστούν. Το K καθορίζεται από το χρήστη του αλγορίθμου ανάλογα με τη λειτουργία που θέλει η K-means να επιτελεί. Για παράδειγμα, αν χρησιμοποιείται για διαγνωστικούς σκοπούς το K μπορεί να πάρει την τιμή δύο (K=2) ώστε να ξεχωρίσει σε ένα σετ μετρήσεων στροβιλοκινητήρα την υγιή λειτουργία από την ανώμαλη. Αν όμως χρησιμοποιείται για τη δημιουργία προφίλ αναφοράς για διάφορα εύρη Ισχύος του κινητήρα αυτού όπως θα δούμε παρακάτω, θα πρέπει να επιλεγεί ένα K μεγαλύτερο του δύο. Αυτό θα επιφέρει καλύτερη ομαδοποίηση των περιοχών Ισχύος χωρίς συγχώνευση τυχόν κοντινών αλλά διαφορετικών περιοχών. Μία τελευταία προς εξέταση περίπτωση είναι αυτή με K ίσο με τη μονάδα. Για K=1, όλα τα σημεία εν τέλει ομαδοποιούνται σε ένα cluster. Αυτό που κάνει αυτήν την περίπτωση ιδιαίτερη είναι ο τρόπος ομαδοποίησης των σημείων σε αυτήν τη μία ομάδα. Ελέγχοντας τη διαδικασία αυτή, τα βήματα που ακολουθούνται είναι ίδια με αυτά της ιεραρχικής μεθόδου που θα αναλυθεί παρακάτω.

2.5 Συναρτήσεις απόστασης

Για την ανάλυση των μεθόδων που χρησιμοποιήθηκαν είναι σημαντικό να γίνει αναφορά στις διάφορες συναρτήσεις απόστασης και τους τρόπους σύνδεσης των cluster που υπάρχουν. Από τις διάφορες συναρτήσεις απόστασης θα χρησιμοποιηθούν δύο και από τους τρόπους σύνδεσης θα χρησιμοποιηθεί ένας στην μετέπειτα ανάλυση για να συγκριθούν τα αποτελέσματα.

2.5.1 Ευκλείδεια απόσταση

Οι συναρτήσεις απόστασης που χρησιμοποιήθηκαν για κάθε μέθοδο ήταν δύο, η Ευκλείδεια απόσταση καθώς και ο Συντελεστής Αλληλοσυσχέτισης CCD (Cross Correlation Coefficient). Οι συναρτήσεις αυτές χρησιμοποιήθηκαν επειδή έχουν ευρύ πεδίο εφαρμογής, ειδικά στις Επιστήμες της Εξόρυξης Δεδομένων (Data Mining) και της Διαγνωστικής Στροβιλοκινητήρων. Παρακάτω παρατίθεται η αρχή λειτουργίας τους αλλά

και άλλες συναρτήσεις απόστασης που δεν χρησιμοποιηθήκαν. Η απόσταση μεταξύ δύο σημείων με Ευκλείδεια απόσταση δίνεται από τη σχέση:

$$Eud_k = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{(i)} - x_{(i)}\right)^2} \tag{2.5}$$

Όπου:

- N: Οι διαστάσεις των εξεταζόμενων σημείων. Προφανώς, τα δύο σημεία θα πρέπει να έχουν ίδιο αριθμό διαστάσεων.
- X_i και Y_i: Υπογραφές των εξεταζόμενων σημείων. Δηλαδή το σύνολο των τιμών του σημείου αυτού σε όλες τις διαστάσεις του.

2.5.2 Συντελεστής Αλληλοσυσχέτισης

Ένας άλλος τρόπος σύγκρισης των σημείων μεταξύ τους πέρα από την Ευκλείδεια απόσταση είναι ο Συντελεστής Αλληλοσυσχέτισης. Για αρχή, παρατίθεται λόγω ακαδημαϊκού ενδιαφέροντος, ενώ αργότερα θα επαναφερθεί για υπολογιστικούς σκοπούς. Η σχέση που μας δίνει το Συντελεστή Αλληλοσυσχέτισης CCD είναι:

$$Ccd_k = \frac{\left|\sum_{i=1}^{N}(p(i) - \bar{p}) * (p_{rk}(i) - \overline{p_{rk}})\right|}{\sqrt{\sum_{i=1}^{N}(p(i) - \bar{p})^2 * \sum_{i=1}^{N}(p_{rk}(i) - \overline{p_{rk}})^2}} \tag{2.6}$$

Όπου:

- N: Οι διαστάσεις των εξεταζόμενων σημείων. Προφανώς, τα δύο σημεία θα πρέπει να έχουν ίδιο αριθμό διαστάσεων.
- P: Υπογραφή του εξεταζόμενου σημείου. Δηλαδή το σύνολο των τιμών του σημείου αυτού σε όλες τις διαστάσεις του.
- P_{rk}: Υπογραφή αναφοράς για τη βλάβη k.
- $\bar{P}$: Μέση τιμή της υπογραφής P
- $\overline{P_{rk}}$: Μέση τιμή της υπογραφής για την k βλάβη.

Κάθε μία από τις υπογραφές πρέπει να έχει N τιμές, ενώ οι μέσες τιμές έχουν μία μόνο τιμή. Η συνάρτηση pdist2, που αναφέρθηκε παραπάνω, χρησιμοποιεί από προεπιλογή την Ευκλείδεια απόσταση αλλά υπάρχουν και άλλες επιλογές συναρτήσεων που μπορεί κανείς να χρησιμοποιήσει. Ο Συντελεστής Αλληλοσυσχέτισης δεν ήταν μία από τις επιλογές, για αυτόν το λόγο παράχθηκε επιπλέον συνάρτηση πηγαίου κώδικα με σκοπό τη σύγκριση αποτελεσμάτων με αυτά της Ευκλείδειας απόστασης.

2.5.3 Άλλες συναρτήσεις απόστασης

Οι πιθανές επιλογές συναρτήσεων απόστασης αποστάσεων μεταξύ των σημείων ονομαστικά ήταν:

- Squared Euclidean Distance: Η απόσταση που μελετήσαμε παραπάνω υψωμένη στο τετράγωνο.
- Standardized Euclidean Distance: Κάθε μία από τις διαστάσεις των εξεταζόμενων μεγεθών κανονικοποιείται διαιρώντας το με το αντίστοιχο στοιχείο της τυπικής απόκλισης.
- Mahalanobis Distance: Χρησιμοποιεί το μητρώο συνδιακύμανσης του Πίνακα X (εξεταζόμενα σημεία). Το μητρώο αυτό θα πρέπει να είναι συμμετρικό και θετικά ορισμένο.
- Minkowski Distance: Μέτρηση απόστασης σε ένα κανονικό διανυσματικό χώρο που μπορεί να θεωρηθεί ως γενίκευση της Ευκλείδειας απόστασης. Ο χρήστης ορίζει τον εκθέτη της συνάρτησης αρκεί να είναι θετικός ακέραιος αριθμός.
- Cityblock Distance: Αντιπροσωπεύει την απόσταση μεταξύ σημείων σε ένα αστικό δίκτυο. Εξετάζει τις απόλυτες διαφορές μεταξύ συντεταγμένων ενός ζεύγους αντικειμένων.
- Chebychev Distance: Μέγιστη διαφορά συντεταγμένων των εξεταζόμενων σημείων.
- Spearman Distance: Απόσταση ίση με τη μονάδα μείον τη συσχέτιση του δείγματος Spearman μεταξύ των παρατηρήσεων.
- Hamming Distance: Η απόσταση Hamming, μετρά τον ελάχιστο αριθμό αντικαταστάσεων που χρειάζονται ώστε να μετατραπεί η μία συμβολοσειρά στην άλλη, ή αλλιώς, τον αριθμό των λαθών που μετέτρεψαν την μία συμβολοσειρά στην άλλη.

2.5.4 Τρόποι σύνδεσης clusters

Πέρα από τις συναρτήσεις απόστασης είναι εξίσου σημαντικό να εξεταστούν τα είδη σύνδεσης των clusters. Αρκετά συχνά στις μεθόδους που θα παρουσιαστούν παρακάτω, είναι αναγκαία η συγχώνευση δύο clusters μεταξύ τους σε ένα μεγαλύτερο γενικότερο cluster. Οι λόγοι για τους οποίους γίνεται κάποια τέτοια συγχώνευση διαφέρουν από μέθοδο σε μέθοδο και θα αναλυθούν περισσότερο παρακάτω. Σημαντικό σε αυτό το σημείο είναι να εξεταστούν τα είδη των συγχωνεύσεων που μπορούν να πραγματοποιηθούν. Οι πιο συνηθισμένες μέθοδοι συγχώνευσης περιγράφονται παρακάτω.

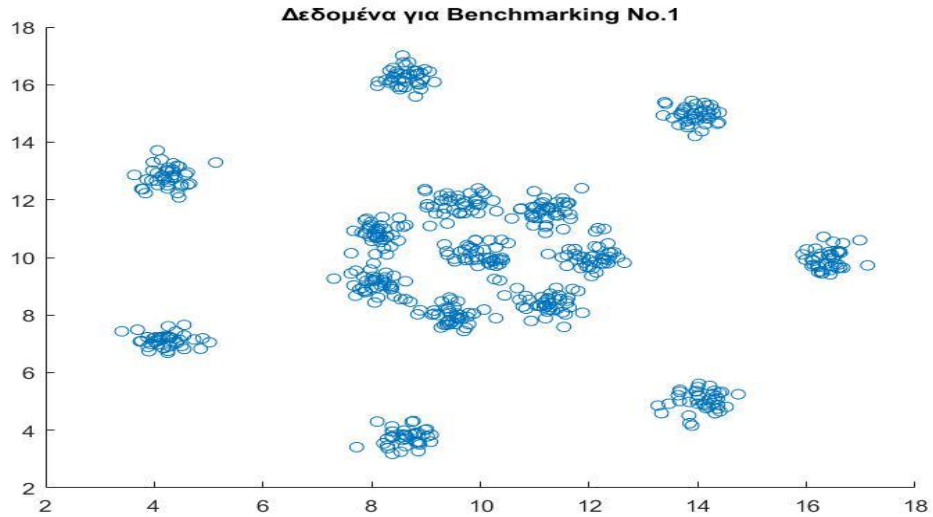
- Μέγιστη ή πλήρης σύνδεση: Η απόσταση μεταξύ δύο clusters ορίζεται ως η μέγιστη τιμή όλων των ζευγών αποστάσεων μεταξύ των στοιχείων του cluster 1 και των στοιχείων του cluster 2. Τείνει να παράγει πιο συμπαγή συμπλέγματα.

- Ελάχιστη ή απλή σύνδεση: Η απόσταση μεταξύ δύο clusters ορίζεται ως η ελάχιστη τιμή όλων των ζευγών αποστάσεων μεταξύ των στοιχείων του cluster 1 και των στοιχείων του cluster 2. Τείνει να παράγει μακρά, "χαλαρά" συμπλέγματα.
- Μέση σύνδεση: Η απόσταση μεταξύ δύο clusters ορίζεται ως η μέση απόσταση μεταξύ των στοιχείων του cluster 1 και των στοιχείων του cluster 2.
- Σύνδεση κέντρου : Η απόσταση μεταξύ των clusters ορίζεται ως η απόσταση μεταξύ του κέντρου για του cluster 1 (μέσος φορέας των μεταβλητών μήκους p) και του κέντρου για του cluster 2.
- Μέθοδος ελάχιστης διακύμανσης του Ward: Μειώνει τη συνολική διακύμανση μεταξύ clusters. Σε κάθε βήμα το ζεύγος clusters με την ελάχιστη απόσταση μεταξύ cluster συγχωνεύονται.

Από τους παραπάνω τρόπους σύνδεσης, αυτός που χρησιμοποιήθηκε για τη συγχώνευση των clusters ήταν η σύνδεση κέντρου. Η συγχώνευση θα χρησιμοποιηθεί κυρίως στην Αθροιστική Ιεραρχική Μέθοδο στην οποία, εξ ορισμού, συγχωνεύονται τα κέντρα των clusters. Το φαινόμενο αυτό θα αναλυθεί παραπάνω στη συνέχεια αφού πρώτα αναλυθούν οι μέθοδοι που χρησιμοποιήθηκαν.

2.6 Διαδικασία επικύρωσης μεθόδων

Σημαντικό βήμα πριν την ανάλυση των λειτουργιών που ακολουθούν είναι να επικυρωθούν οι μέθοδοι που χρησιμοποιήθηκαν. Η επικύρωση μεθόδων χρησιμοποιείται για να περιγράψει τη διαδικασία αξιολόγησης της πιστότητας των αποτελεσμάτων του αλγόριθμου ομαδοποίησης. Η διαδικασία αυτή είναι σημαντική για την σύγκριση των μεθόδων και την ανακάλυψη τυχόν αδυναμιών που μπορεί να χαρακτηρίζουν κάποια από αυτές. Για τη διαδικασία αυτή λοιπόν χρησιμοποιήθηκαν δύο τεστ δεδομένων. Τα τεστ αφορούν δισδιάστατα δεδομένα και καθένα από αυτά υποβάλλει τις μεθόδους σε διαφορετικές δυσκολίες. Το πρώτο τεστ φαίνεται παρακάτω:



Σχήμα 2.11: Τεστ δισδιάστατων δεδομένων για επικύρωση των μεθόδων No.1

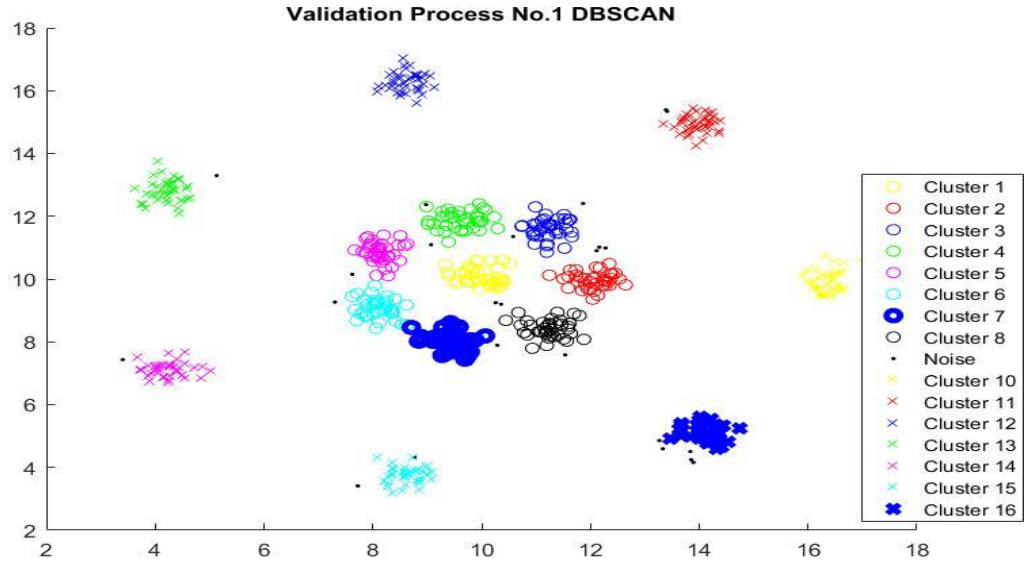
Στο παραπάνω σχήμα είναι διακριτό για έναν εξωτερικό παρατηρητή ότι τα Cluster που σχηματίζονται είναι 15, οι μέθοδοι που αναπτύχθηκαν όμως δεν το γνωρίζουν αυτό. Έτσι, εισάγοντας τα δεδομένα μπορεί εύκολα ο χρήστης να καταλάβει το ποσοστό επιτυχίας των μεθόδων. Πρώτα ελέγχθηκε η ακρίβεια της μεθόδου DBSCAN. Για μεγέθη εισόδου ίσα με:

$$epsilon = 0.4$$

Και

$$MinPts = 3$$

Παράγονται τα παρακάτω αποτελέσματα:



Σχήμα 2.12: Αποτελέσματα Validation Process No.1 DBSCAN

Όπως φαίνεται στο Σχήμα 2.12 η DBSCAN είναι ικανή να ομαδοποιήσει σωστά τα δεδομένα που της δόθηκαν. Το αποτέλεσμα αυτό βέβαια προκύπτει μετά από σωστή επιλογή των μεγεθών εισόδου. Το *epsilon* επιλέχθηκε ίσο με 1.5 λόγω του διαγράμματος αποστάσεων κ-γειτόνων που αναφέρθηκε παραπάνω, ενώ το μέγεθος *MinPts* λόγω των διαστάσεων των δεδομένων (δηλαδή 2) βάσει του θεωρητικού κανόνα:

$$MinPts = D + 1 \qquad (2.7)$$

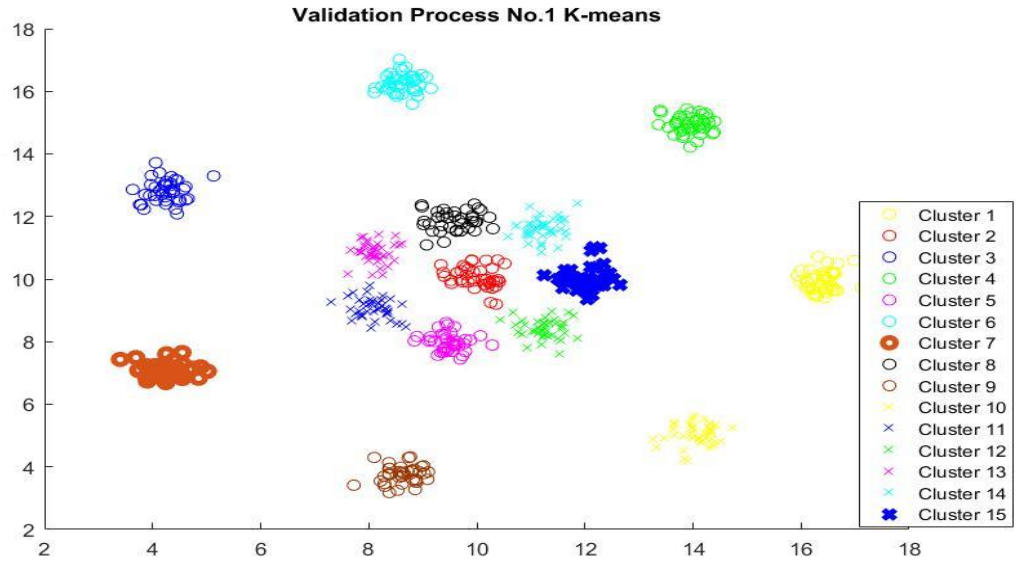
Πρακτικά πρόκειται για μία επαναληπτική διαδικασία όπου το *epsilon* αυξάνεται σταδιακά μέχρι να ικανοποιηθούν 2 περιορισμοί:

1. Να σχηματιστεί ο επιθυμητός αριθμός Clusters για δεδομένο *epsilon* μέσω της εντολής:
$$if\ \max(I) < 15.5$$
2. Να υπάρχουν όσο το δυνατόν ελάχιστα σημεία θορύβου στο δείγμα μέσω της εντολής:
$$if\ numel\big(find(I == 0)\big) < 3$$

Μέσω αυτής της διαδικασίας το μέγεθος εισόδου προκύπτει ίσο με:

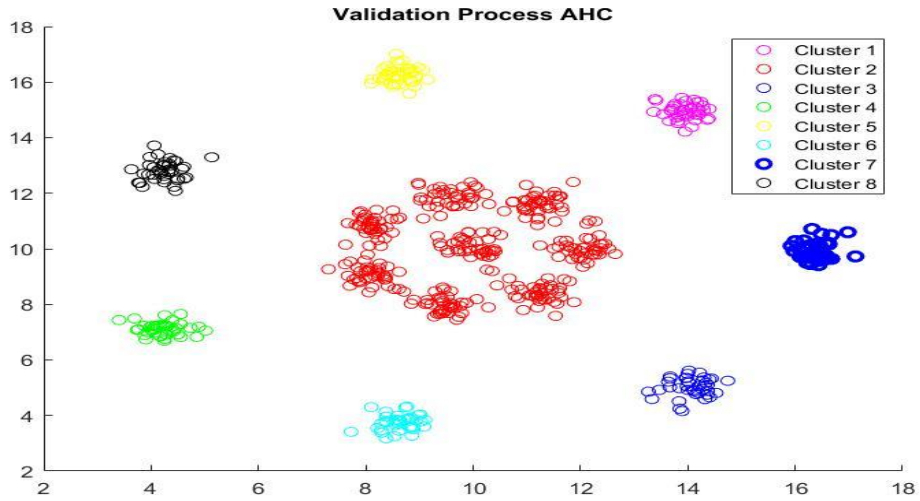
$$epsilon = 0.43$$

Δηλαδή πολύ κοντά σε αυτό που προκύπτει από το σχήμα κ-γειτόνων. Η διαδικασία έμμεσου υπολογισμού του *epsilon* θα φανεί χρήσιμη και παρακάτω καθώς για τη δημιουργία προφίλ θερμοκρασιακής αναφοράς είναι απαραίτητος ως είσοδος ο αριθμός επιθυμητών Clusters. Το *epsilon* αυτό είναι λίγο μεγαλύτερο επειδή έχει ως περιορισμό τον αριθμό σημείων θορύβου. Όπως γνωρίζουμε, η αύξηση του *epsilon* οδηγεί σε μείωση των σημείων θορύβου πέρα από μείωση των σχηματιζόμενων Clusters.
Τα αποτελέσματα της K-means για το ίδιο τεστ φαίνονται στο Σχήμα 2.13:



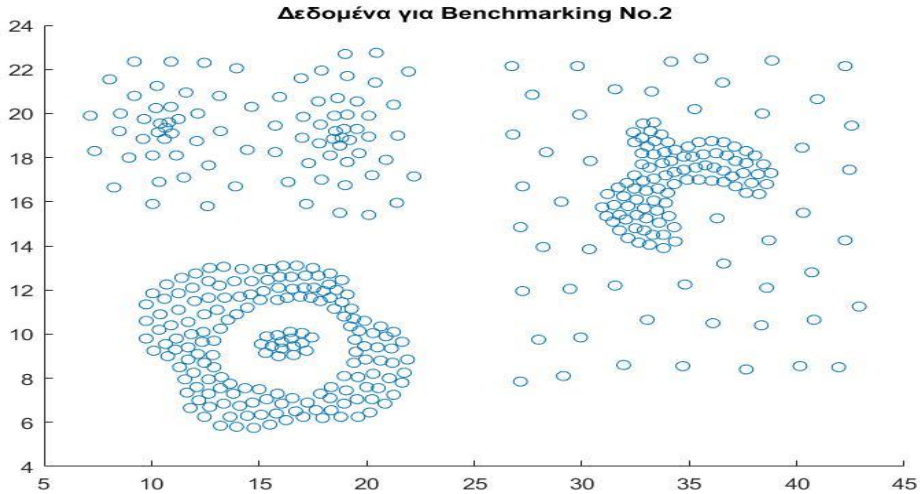
Σχήμα 2.13: Αποτελέσματα Validation Process No.1 K-means

Η αρχή λειτουργίας της K-means είναι διαφορετική από αυτήν της DBSCAN. Στην προκειμένη, ο επιθυμητός αριθμός Clusters επιλέγεται από το χρήστη. Τα αποτελέσματα που προέκυψαν είναι ακόμα καλύτερα από αυτά της DBSCAN καθώς η K-means δεν έχει ονοματίσει σημεία θορύβου. Η K-means είναι ιδανική για διάγνωση συλλογικών ανωμαλιών οπότε για δεδομένα μορφής όμοιας με την παραπάνω είναι λογικό να έχει επιτυχία 100%.
Τα αποτελέσματα της AHC για το ίδιο τεστ φαίνονται στο Σχήμα 2.14:



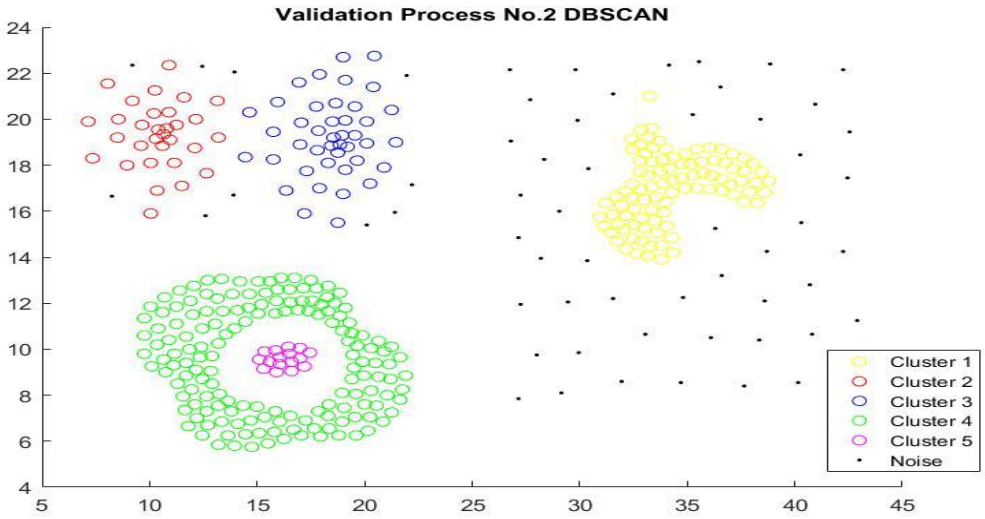
Σχήμα 2.14: Αποτελέσματα Validation Process No.1 AHC

Η μέθοδος AHC παρουσιάζει σχετικά αποτελέσματα. Όπως φαίνεται στο Σχήμα 2.14 έχει ομαδοποιήσει σωστά τα περιφερειακά Clusters και προφανώς λόγω της φύσης της λειτουργίας της δεν έχει παραγάγει σημεία θορύβου όπως η DBSCAN. Η αστοχία αυτής της μεθόδου περιγράφεται μέσω των 8 γειτονιών σημείων στο κέντρο του διαγράμματος. Εκεί, τα σημεία της κάθε γειτονιάς ομαδοποιούνται σε ένα μεγάλο Cluster αντί για 8 ξεχωριστά. Γνωρίζοντας τη θεωρητική βάση πίσω από τη μέθοδο αυτή, δε δύναται να κατοχυρωθεί ως λάθος της μεθόδου. Πιο συγκεκριμένα, η AHC ομαδοποιεί ιεραρχικά από τη μικρότερη προς τη μεγαλύτερη απόσταση σημείων-Clusters. Πολλά από τα σημεία του Cluster 2 λοιπόν μοιράζονται μικρότερες αποστάσεις μεταξύ τους, από ότι σημεία του Cluster 8για παράδειγμα. Έτσι ομαδοποιούνται προτού σχηματιστεί ο επιθυμητός αριθμός Clusters που σηματοδοτεί τη λήξη της διαδικασίας. Στη συνέχεια οι μέθοδοι αυτές ελέγχθηκαν σε ένα 2^{ο} τεστ που φαίνεται στο Σχήμα 2.15:



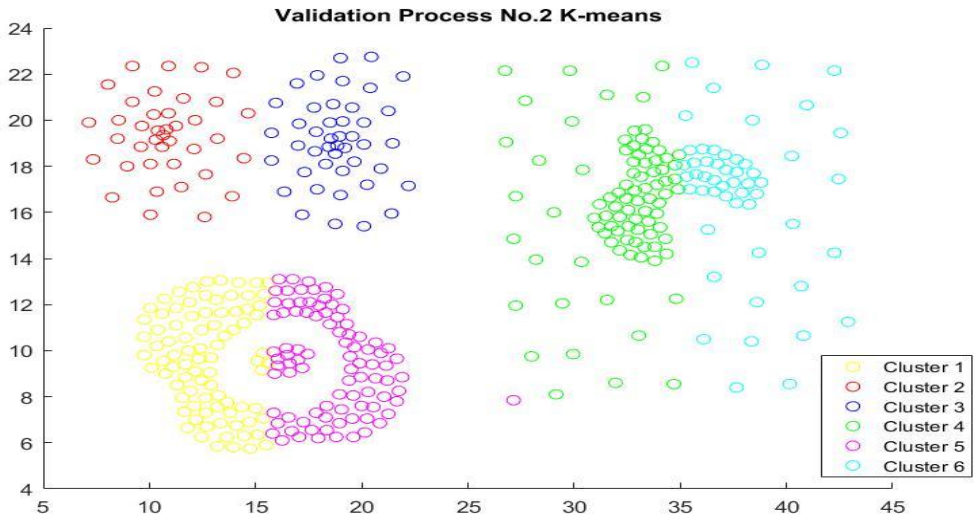
Σχήμα 2.15: Τεστ δισδιάστατων δεδομένων για επικύρωση των μεθόδων No.2

Το τεστ αυτό περιέχει 6 Clusters και ελέγχει την ακρίβεια των μεθόδων σε πυκνά και μη δεδομένα. Ενδιαφέρον προκαλούν τα αποτελέσματα όσων αφορά τα ομόκεντρα Clusters που φαίνονται στις συντεταγμένες: (10-20,6-12) περίπου καθώς και τα αποτελέσματα στα Clusters που σχηματίζονται στις συντεταγμένες: (30-42,8-22) όπου οι αλγόριθμοι καλούνται να αναγνωρίσουν τη διαφορά πυκνότητας δεδομένων ώστε να τα διαχωρίσουν. Τα αποτελέσματα για την DBSCAN φαίνονται παρακάτω:



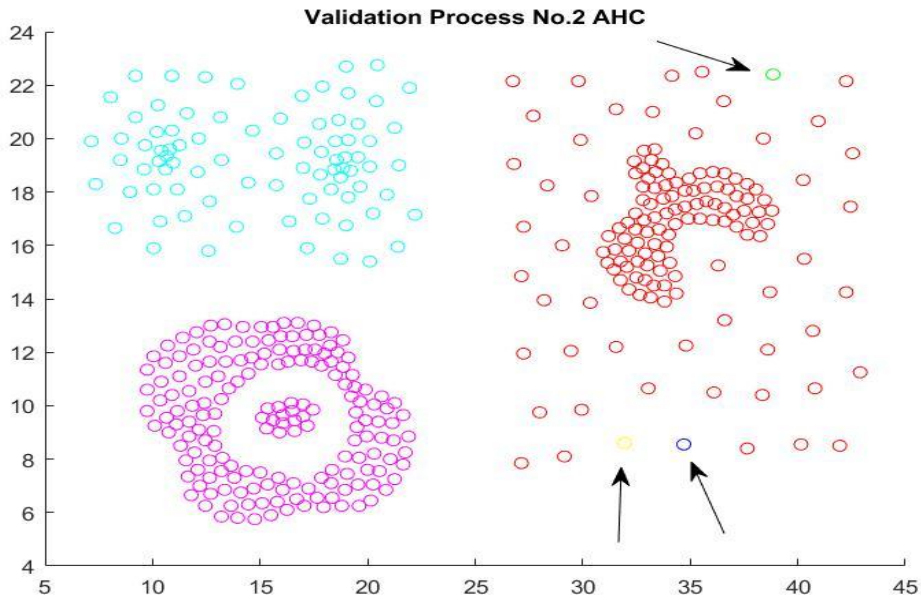
Σχήμα 2.16: Αποτελέσματα Validation Process No.1 DBSCAN

Οι στόχοι που αναφέρθηκαν παραπάνω φαίνεται να εκπληρώνονται στα αποτελέσματα της DBSCAN. Ειδικότερα, τα έκκεντρα Clusters ξεχωρίζονται μεταξύ τους, όπως και τα Cluster 2 και 3 είναι ξεκάθαρα διαχωρισμένα. Το σημαντικότερο όμως χαρακτηριστικό των αποτελεσμάτων της DBSCAN είναι ο διαχωρισμός των Cluster 1 και των σημείων θορύβου. Σε εκείνη την περιοχή ο αλγόριθμος όχι απλά διαχώρισε την περιοχή μεταξύ του πυκνού και αραιού Cluster αλλά εκφράζει και την αμφιβολία ύπαρξης Cluster στην περιοχή που υπάρχουν σημεία θορύβου. Για *epsilon* μικρότερο από το ιδανικό που επιλέχθηκε τα σημεία θορύβου δημιουργούν ένα Cluster όλα μαζί. Το νέο *epsilon* όμως θα ομαδοποιούσε τα Clusters 2 και 3 σε ένα Cluster.



Σχήμα 2.17: Αποτελέσματα Validation Process No.1 K-means

Στο Σχήμα 2.17 φαίνεται τα Clusters 2 και 3 να είναι καθαρά ορισμένα χωρίς σημεία θορύβου όπως παράγει η DBSCAN. Το δεύτερο τεστ εμφανίζει περισσότερο όμως τις αδυναμίες της k-means. Τα Clusters 1, 4, 5 και 6 είναι λάθος διαχωρισμένα. Αυτό οφείλεται στον τρόπο λειτουργίας της K-means. Ο υποχρεωτικός ορισμός των κέντρων των Cluster αποκαλύπτει την αδυναμία της μεθόδου να ξεχωρίσει Clusters όπως τα παραπάνω. Από τη μία πλευρά, τα Clusters 1 και 5 είναι ομόκεντροι κύκλοι άρα ο διαχωρισμός τους από τη μέθοδο είναι αδύνατος, οπότε λανθασμένα η μέθοδος χωρίζει τους κύκλους σε ημικύκλια ώστε να παρουσιάσει αποτέλεσμα. Από την άλλη πλευρά στα Clusters 4 και 6 δεν αναγνωρίζονται οι περιοχές διαφορετικής πυκνότητας.



Σχήμα 2.18: Αποτελέσματα Validation Process No.1 AHC

Στο συγκεκριμένο τεστ η μέθοδος AHC είναι αυτή με τα χειρότερα αποτελέσματα. Στο Σχήμα 2.18 φαίνεται η αδυναμία διαχωρισμού όλων των κοντινών Clusters μεταξύ τους. Είναι γνωστό ότι η μέθοδος αυτή ζητά από το χρήστη ως είσοδο τον επιθυμητό αριθμό Clusters, επομένως γεννιούνται απορίες σχετικά με την αποτυχία της μεθόδους στα κοντινά Clusters. Τα βέλη στο παραπάνω σχήμα βοηθούν στην κατανόηση του τρόπου λειτουργίας της μεθόδου. Πιο συγκεκριμένα, η μέθοδος ομαδοποιεί τα κοντινά μεταξύ τους σημεία, και τα σημεία με βελάκι είναι τα σχετικά μακρύτερα από τους γείτονες τους με αποτέλεσμα η μέθοδος να τα παρουσιάζει ως ξεχωριστά Clusters μεταξύ τους. Τα συνολικά αποτελέσματα των μεθόδων φαίνονται στο παρακάτω Σχήμα:

Μέθοδος	Πακέτο δεδομένων No.1	Πακέτο δεδομένων No.2
DBSCAN	92%	100%
K-means	100%	55%
AHC	84%	61%

Σχήμα 2.19: Ποσοστό επιτυχίας μεθόδων ανάλογα με το τεστ επικύρωσης

Από το Σχήμα 2.19 φαίνεται η DBSCAN να έχει τα καλύτερα αποτελέσματα. Για την k-means, όπως είπαμε και πριν το τεστ No.2 είναι που αποκαλύπτει τις αδυναμίες της και για αυτό έχει αρκετά χαμηλό σκορ εκεί. Η AHC είναι χειρότερη από τις δύο μεθόδους στο πρώτο τεστ αλλά καλύτερη από την K-means στο δεύτερο. Άρα φαίνεται καλύτερη να είναι η DBSCAN με δεύτερη την K-means και Τρίτη την AHC.

3 Δημιουργία προφίλ αναφοράς θερμοκρασίας εξόδου καυσαερίων

Στο παρόν κεφάλαιο γίνεται η χρήση των μεθόδων clustering για τη δημιουργία προφίλ αναφοράς θερμοκρασίας εξόδου καυσαερίων Αναλύεται η σημασία της ομαδοποίησης αυτής και ο τρόπος ομαδοποίησης που χρησιμοποιείται ως σήμερα. Προτείνεται νέος τρόπος βάσει των επιλεγμένων μεθόδων και συγκρίνονται τα αποτελέσματα σε δεδομένα τριών αεριοστροβίλων.

3.1 Γενικά

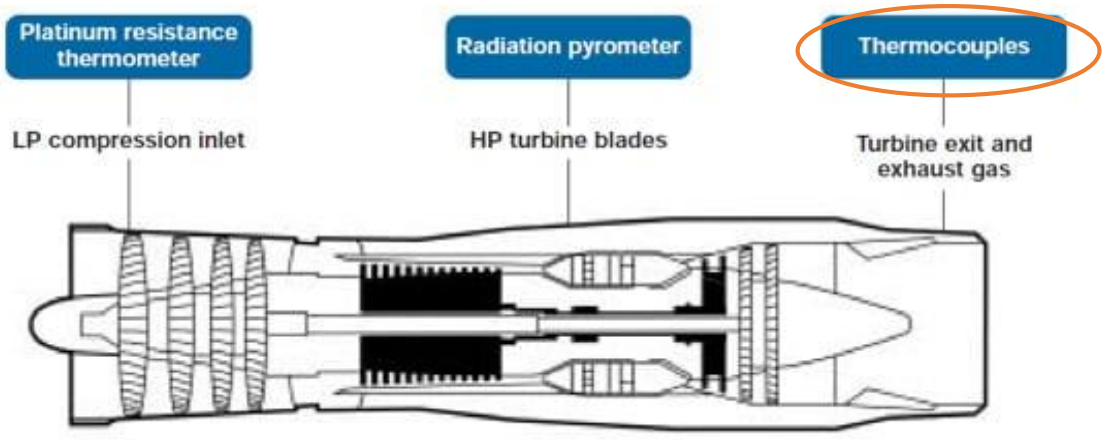
Στο κεφάλαιο αυτό αναλύεται η διαδικασία επεξεργασίας πραγματικών δεδομένων αεριοστρόβιλου και ο διαχωρισμός αυτών για τη δημιουργία προφίλ αναφοράς. Πιο συγκεκριμένα, λαμβάνονται μετρήσεις από κινητήρες που παρακολουθούνται συστηματικά από το Εργαστήριο Θερμικών Στροβιλοκινητήρων του Εθνικού Μετσόβιου Πολυτεχνείου, με όνομα GTA, GTB1 και GTB2. Οι μετρήσεις αυτές αφορούν τη λειτουργία των κινητήρων για τη χρονική περίοδο 2018 έως 2019, χρονική περίοδος κατά την οποία οι κινητήρες δεν υπέστη καμία βλάβη.

Η ανάλυση αυτή έχει σκοπό την αναγνώριση προτύπων. Οι τεχνικές αναγνώρισης προτύπων έχουν ευρεία εφαρμογή στις διαγνωστικές μεθόδους, δεδομένου ότι με τη χρήση τους μπορεί να αυτοματοποιηθεί η αναγνώριση βλαβών, μία διαδικασία που αλλιώς απαιτεί την παρέμβαση του ανθρώπινου παράγοντα. Πιο συγκεκριμένα, η διαδικασία που ακολουθείται παρακάτω ομαδοποιεί προφίλ θερμοκρασιών εξόδου από το στρόβιλο ανάλογα με την ομοιότητα τους. Γενικά, η μελέτη της θερμοκρασίας εξόδου από το στρόβιλο είναι μία σύνθετη διαδικασία καθώς επηρεάζεται από αρκετούς παράγοντες. Η διαδικασία αυτή εξαρτάται από:

- Το ποσοστό ανάμειξης αέρα-καυσίμου
- Τον αριθμό καυστήρων στο θάλαμο καύσης
- Το σχήμα του θαλάμου καύσης
- Το ενδεχόμενο ύπαρξης μετακαυστήρα
- Το ποσοστό μείγματος που καίγεται τέλεια στον θάλαμο καύσης

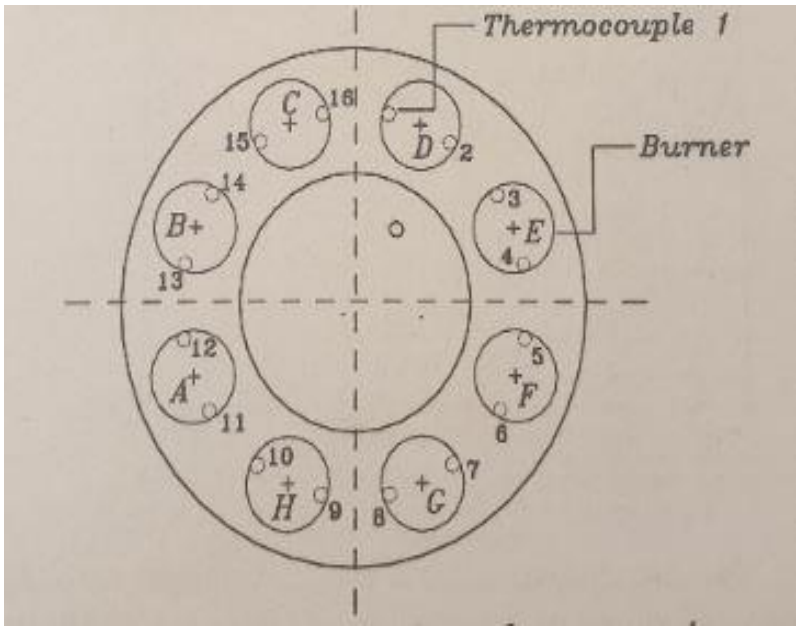
Είναι κατανοητό επομένως ότι η παρακολούθηση του δεδομένη και εξίσου σύνθετη με τη διαδικασία. Αναλυτικότερα, ο θάλαμος καύσης έχει συνήθως σχήμα δακτυλίου, σε τέτοια περίπτωση τοποθετούνται στην έξοδο του στροβίλου περιφερειακά διάφορα θερμοστοιχεία για τον έλεγχο του προφίλ θερμοκρασίας. Τα θερμοστοιχεία αυτά

ισαπέχουν, καταγράφουν το εύρος θερμοκρασιών που επικρατούν ανά πάσα στιγμή και είναι 24 στον αριθμό για τους δύο κινητήρες που εξετάζονται.



Σχήμα 3.1: Κάθετη τομή κινητήρα turbojet με τους διάφορους αισθητήρες θερμοκρασίας σε όλο το μήκος του

Στην παραπάνω τομή ενός κινητήρα turbojet φαίνονται οι διάφοροι αισθητήρες θερμότητας που τοποθετούνται για παρακολούθηση της θερμοκρασίας σε κάθε συνιστώσα. Τα δεδομένα για τους κινητήρες αφορούν τους αισθητήρες που φαίνονται στον κόκκινο κύκλο. Για τους κινητήρες που μελετώνται, παρακολουθούνται 16 ή 24 θερμοστοιχεία ανάλογα με το μοντέλο του κινητήρα, τα οποία είναι τοποθετημένα περιμετρικά όπως φαίνεται στο σχήμα 3.2.



Σχήμα 3.2: Κατανομή θερμοστοιχείων στην έξοδο του στροβίλου σε σύγκριση με τη θέση των θαλάμων καύσης

Στη συνέχεια, η διαδικασία ομαδοποίησης που αναφέρθηκε παραπάνω δεν περιλαμβάνει τις τιμές διορθωμένης ισχύος κάθε μέτρησης. Ελέγχεται αν η ομαδοποίηση προφίλ θερμοκρασιών οδηγεί σε ομαδοποίηση προφίλ διορθωμένης ισχύος. Έτσι δίνεται η δυνατότητα παραγωγής ενός σετ προφίλ θερμοκρασιών ανάλογα με τη διορθωμένη ισχύ που λειτουργούν οι κινητήρες.

3.2 Μέθοδος παραγωγής μέσων προφίλ χωρίς clustering

Η μέθοδος που χρησιμοποιείται μέχρι σήμερα για την παραγωγή προφίλ αναφοράς χωρίζει το σύνολο εύρους λειτουργίας του αεριοστρόβιλου σε πεπερασμένα διαστήματα. Αν για παράδειγμα μελετάται ένας κινητήρας που έχει εύρος λειτουργίας 100 με 200 MW διορθωμένης ισχύος και ο χρήστης επιθυμεί να χωρίσει το εύρος αυτό σε 10 διαστήματα, τότε θα δημιουργηθούν προφίλ αναφοράς των 10 MW. Κάθε σημείο από τα εξεταζόμενα δεδομένα θα κατατάσσεται σε ένα από τα διαστήματα βάσει της διορθωμένης ισχύος του. Για την παραγωγή του μέσου προφίλ κάθε διαστήματος υπολογίζεται η μέση τιμή για κάθε θερμοστοιχείο των δεδομένων που έχουν εκχωρηθεί στο διάστημα αυτό. Εύκολα συμπεραίνεται λοιπόν ότι η μέθοδος αυτή δεν εξετάζει τις τιμές των θερμοστοιχείων, παρά μόνο τη διορθωμένη ισχύ που περιγράφει τα δεδομένα. Το αρνητικό αυτής της μεθόδου είναι ότι το πλήθος των διαστημάτων ορίζεται από το χρήστη. Έτσι, αν τα διαστήματα αυτά είναι μικρά τότε ο διαθέσιμος αριθμός σημείων μπορεί να μην είναι ικανός για τη δημιουργία προφίλ αναφοράς. Αν τα διαστήματα είναι μεγάλα τότε τα προφίλ που παράγονται δεν αντιπροσωπεύουν καλά την περιοχή ισχύος που περιγράφουν. Για αυτό λοιπόν προτείνεται διαφορετική μέθοδος.

3.3 Μέθοδος παραγωγής μέσων προφίλ με clustering

Η μέθοδος που προτείνεται ακολουθεί διαφορετική προσέγγιση από τη μέθοδο που χρησιμοποιείται ως τώρα. Μέσω των τιμών των θερμοστοιχείων ομαδοποιούνται τα δεδομένα σε διάφορα clusters.

3.3.1 Επεξεργασία και εισαγωγή δεδομένων

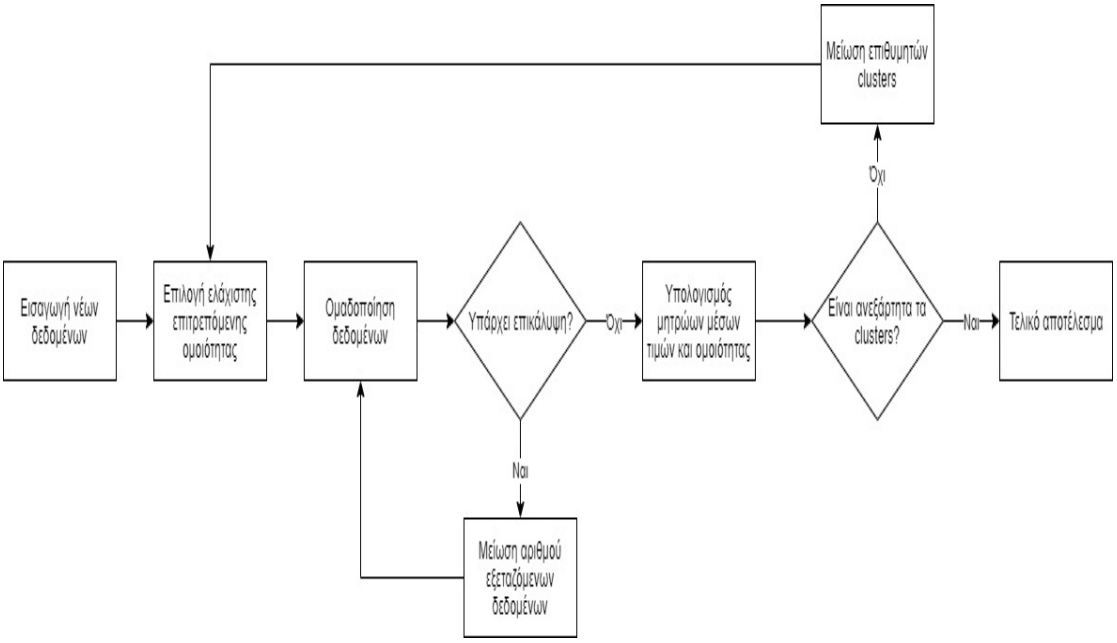
Η διαδικασία ξεκινά με την επεξεργασία δεδομένων σε υπολογιστικό φύλλο Excel. Αρχικά, οι μετρήσεις των θερμοκρασιών είναι σε βαθμούς Celsius οπότε μετατρέπονται σε βαθμούς Kelvin μέσω της σχέσης:

$$\theta\ [°K] = \theta\ [°C] + 273.15 \quad (3.1)$$

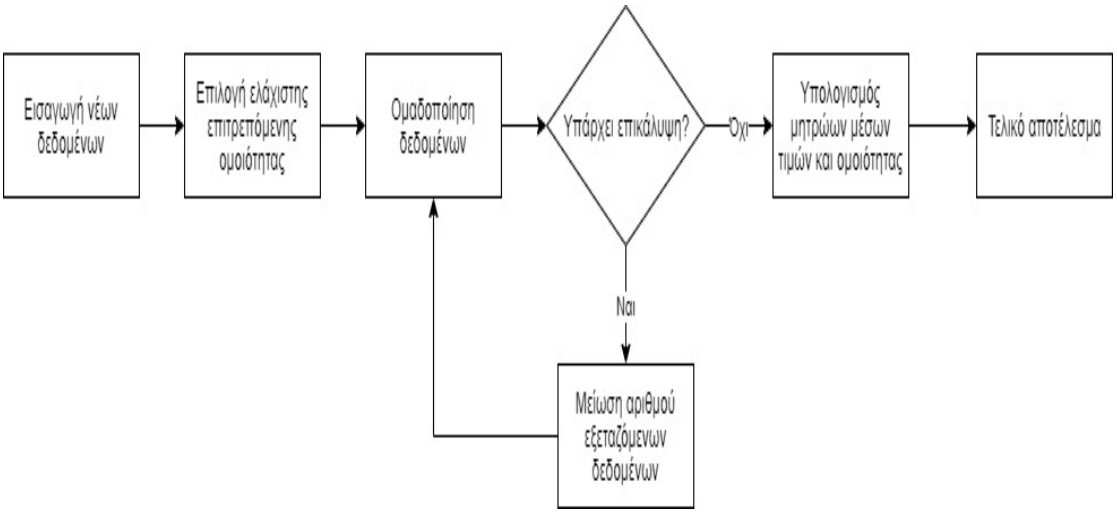
Έτσι παράγεται ένα νέο μητρώο ίδιων διαστάσεων με τις μετρήσεις που αφορούν τη θερμοκρασία να έχουν μετατραπεί σε βαθμούς Kelvin. Στη συνέχεια, αδιαστατοποιείται το μητρώο διαιρώντας κάθε μέτρηση της χρονικής στιγμής i με το μέσο όρο τιμών της συγκεκριμένης χρονικής στιγμής. Έτσι πλέον οι τιμές των θερμοστοιχείων έχουν τιμές γύρω από τη μονάδα με ποσοστό απόκλισης ανάλογο της απόκλισης της

πραγματικής τιμής από τη μέση τιμή κάθε χρονικής στιγμής. Ένας επιπλέον έλεγχος που μπορεί να γίνει για να διεκπεραιωθεί η ορθότητα της επεξεργασίας που μόλις προηγήθηκε είναι να υπολογιστεί ο μέσος όρος των τιμών όλων των θερμοστοιχείων σε μία δεδομένη χρονική στιγμή. Αν αυτός ο μέσος όρος έχει τιμή διαφορετική της μονάδας τότε αποδεικνύεται μαθηματικά ότι η διαδικασία επεξεργασίας δεν έγινε σωστά. Το μητρώο με τις πλέον επεξεργασμένες μετρήσεις εισάγεται στον αλγόριθμο ομαδοποίησης.

Τα βήματα που ακολουθεί ο κάθε αλγόριθμος με τις πλέον επεξεργασμένες μετρήσεις είναι κοινά μεταξύ των μεθόδων. Χρησιμοποιώντας όμως διαφορετική συνάρτηση απόστασης για την ομαδοποίηση αλλάζουν και τα βήματα που θα ακολουθήσει ο αλγόριθμος. Τα διαγράμματα ροής των αλγορίθμων φαίνονται στα παρακάτω σχήματα και εξηγηθούν επιγραμματικά οι διαφορές τους:



Σχήμα 3.3: Διάγραμμα ροής για ευκλείδεια απόσταση



Σχήμα 3.4: Διάγραμμα ροής για συντελεστή αλληλοσυσχέτισης

Παρατηρούμε διάφορες ομοιότητες και διαφορές στα παραπάνω διαγράμματα ροής. Οι ομοιότητες είναι οι εξής: Εισάγονται τα νέα δεδομένα, ορίζει ο χρήστης την ελάχιστη επιτρεπόμενη ομοιότητα στα δεδομένα που αναλύονται, γίνεται ομαδοποίηση βάσει των τιμών των θερμοστοιχείων έλεγχος για επικάλυψη των clusters και υπολογίζονται τα μητρώα μέσης τιμής και ομοιότητας. Μέχρι εδώ η διαδικασία είναι όμοια για τα δύο Διαγράμματα, η διαφορά τους βρίσκεται στην ανεξαρτησία των clusters. Ορίζοντας ως συνάρτηση απόστασης τον συντελεστή αλληλοσυσχέτισης δεν χρειάζεται έλεγχος για την ανεξαρτησία των Clusters καθώς αυτή εξασφαλίζεται από τον τρόπο σύγκρισης και ομαδοποίησης των δεδομένων. Δηλαδή, για ομαδοποίηση με CCD τα Clusters είναι σίγουρα ανεξάρτητα. Η διαδικασία όμως που ακολουθεί η κάθε μέθοδος ως προς την εφαρμογή αυτών των διαγραμμάτων ροής είναι διαφορετική.

DBSCAN

Η DBSCAN ομαδοποιεί τα δεδομένα βάσει των χαρακτηριστικών μεγεθών *epsilon* και *MinPts*. Γνωρίζοντας ότι η DBSCAN δε δίνει τη δυνατότητα στο χρήστη προεπιλογής συνολικού αριθμού Clusters που θέλει να εξάγει ως αποτέλεσμα, είναι αναγκαίο ο αριθμός των clusters να οριστεί μέσω των χαρακτηριστικών μεγεθών της μεθόδου. Προστέθηκε λοιπόν στον αλγόριθμο της DBSCAN η παρακάτω συνθήκη:

```
%% Gia to veltisto (elaxisto) e kathe cluster anaforas (ref)
for epsilon=0.0001:0.001:10
    [I, Noise, Neighbors]=DBSCAN(EGTref,epsilon,MinPts,w);
    if max(I)<1.5 && numel(find(I==0))<3
        break
    end
end
```

Σχήμα 3.5: Κομμάτι κώδικα εύρεσης του κατάλληλου epsilon

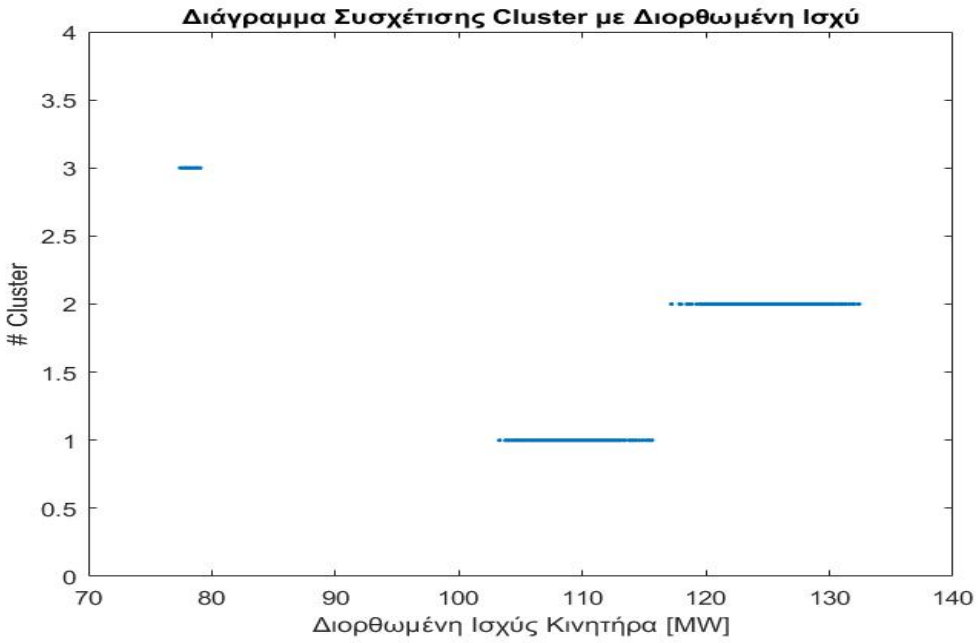
Η παραπάνω σειρά εντολών μέσω πολλών επαναλήψεων κάθε φορά για διαφορετικό *epsilon* πραγματοποιεί έλεγχο DBSCAN ώστε τα σημεία που ανήκουν στο cluster θορύβου να μην ξεπερνούν τα 3 ενώ παράλληλα, ο μέγιστος αριθμός clusters να παραμένει μικρότερος από Χ. Το Χ αντιπροσωπεύει το μέγιστο αριθμό clusters ανά επανάληψη και είναι ο «χειροκίνητος» τρόπος επιβολής συγκεκριμένου αριθμού clusters στη DBSCAN μέθοδο. Πρακτικά, η παραπάνω σειρά εντολών βρίσκει το ελάχιστο *epsilon* για το οποίο έχουμε Χ clusters. Έτσι παράγονται Χ τιμές *epsilon* που κάθε μία αντιπροσωπεύει την ιδανική τιμή του για τον σχηματισμό Χ clusters. Στη συνέχεια, χρησιμοποιώντας ένα-ένα τα *epsilon* που υπολογίσθηκαν παραπάνω, εισάγονται στον αλγόριθμο και γίνεται ομαδοποίηση των επεξεργασμένων δεδομένων.

K-means και AHC

Η K-means και η AHC ομαδοποιούν τα δεδομένα βάσει του χαρακτηριστικού μεγέθους K που περιγράφει τον αριθμό των κέντρων (άρα και clusters) που θα σχηματιστούν. Η διαδικασία λοιπόν που έγινε νωρίτερα με τη DBSCAN μέθοδο θεωρείται περιττή. Τα άλλα δύο μεγέθη που χαρακτηρίζουν την K-means, ο αριθμός επαναλήψεων και ο αριθμός ανοχής απόστασης για αυτή τη λειτουργία μένουν στις μεγαλύτερες δυνατές τιμές ώστε να μην επηρεάσουν το αποτέλεσμα. Πιο συγκεκριμένα, ο αριθμός επαναλήψεων μένει ίσος με $iter = 500000$ ενώ η ανοχή απόστασης ίση με $tol = 10^{-8}$. Έτσι, αναγκάζεται ο αλγόριθμος να παραγάγει αποτέλεσμα μόνο όταν συγκλίνουν οι συντεταγμένες των κέντρων των Clusters και όχι λόγω πεπερασμένου αριθμού επαναλήψεων της διαδικασίας. Για την AHC ο αλγόριθμος χρειάζεται μόνο ένα ακόμα δεδομένο για να τρέξει το μέγεθος D που ορίζει τη μέγιστη δυνατή απόσταση συγχώνευσης ο οποίος για τη συγκεκριμένη ανάλυση παίρνει μία μεγάλη τιμή μιας και δεν επηρεάζει το αποτέλεσμα.

3.3.2 Έλεγχος επικάλυψης

Μετά την ομαδοποίηση των δεδομένων παράγεται ένας Πίνακας-Στήλη μήκους ίσου με τον αριθμό των εξεταζόμενων δεδομένων που περιγράφει σε ποιο cluster κατατάχθηκε κάθε σημείο. Φτιάχνοντας ένα διάγραμμα των τιμών αυτών και της διορθωμένης ισχύος που περιγράφει κάθε μέτρηση προκύπτει ένα σχήμα σαν το παρακάτω:



Σχήμα 3.6: Διάγραμμα ελέγχου επικάλυψης

Στο διάγραμμα αυτό παρατηρούμε αν τα σημεία κάθε cluster περιγράφουν παρόμοιες περιοχές φορτίου καθώς και αν υπάρχει επικάλυψη μεταξύ περιοχών φορτίου και Clusters. Στο συγκεκριμένο διάγραμμα δεν υπάρχει επικάλυψη. Αν όμως υπήρχε επικάλυψη τότε θα σήμαινε ότι κάποια μεταβολή έχει επιβληθεί στη μηχανή και λόγω αυτής έχουν μεταβληθεί τα προφίλ αναφοράς. Για αυτό λοιπόν προτείνετε από τα διαγράμματα ροής μείωση των εξεταζόμενων δεδομένων για να βρεθεί σε ποιο σημείο έχει προκύψει αυτή η μεταβολή και να παραχθούν νέα προφίλ για την περιοχή μετά τη μεταβολή.

3.3.3 Μητρώα μέσης τιμής και ομοιότητας

Ακολούθως, αν δεν υπάρχει επικάλυψη των clusters πλέον, ο αλγόριθμος παράγει ένα μητρώο t * k που περιέχει τη μέση τιμή κάθε θερμοστοιχείου (από τα t θερμοστοιχεία) σε κάθε ένα από τα σχηματισμένα Clusters. Ο πίνακας αυτός έχει k στήλες επειδή είναι ο μέγιστος αριθμός clusters που χρησιμοποιείται. Για *epsilon* μεγαλύτερο από αυτό που αντιστοιχεί στα k clusters ο αλγόριθμος αφήνει τις απαραίτητες στήλες κενές. Ένα παράδειγμα του πίνακα που σχηματίζεται για *epsilon* 0.0029 στο δείγμα της μηχανής GTB1 είναι το εξής:

Results	# Cluster								
	1	2	3	4	5	6	7	8	9
1	0.9985148	0.994949	0.994949	0.998049	1.000375	0.99805	0.998287	0.997589	0.997065
2	1.0055005	1.003628	1.003628	1.007916	1.005155	1.005598	1.005441	1.006062	1.006151
3	0.9991113	0.998768	0.998768	1.003487	1.010531	0.998396	0.998525	0.998055	0.997917
4	0.9957176	0.99881	0.99881	0.994871	0.995767	0.995962	0.995928	0.995903	0.995782
5	1.0109386	0.999842	0.999842	1.012032	1.008686	1.011502	1.011188	1.012343	1.013153
6	1.0019283	1.001225	1.001225	1.006204	1.008143	1.001573	1.001671	1.001303	1.001422
7	1.0065679	1.00113	1.00113	0.999231	0.994553	1.006519	1.007105	1.004773	1.003511
8	1.0011156	1.000614	1.000614	1.001863	1.002683	1.001155	1.000928	1.001967	1.001446
9	0.9991587	1.002499	1.002499	1.006165	1.009646	0.998601	0.998546	0.998824	0.999337
10	0.9933629	1.003163	1.003163	1.001274	1.004353	0.99307	0.992938	0.99344	0.994332
11	0.9884302	1.006551	1.006551	0.987164	0.990921	0.988424	0.988592	0.987613	0.987181
12	1.0034442	1.005619	1.005619	0.996241	0.99534	1.004218	1.004185	1.004002	1.003262
13	0.9951586	1.001743	1.001743	0.994533	0.991867	0.995371	0.995312	0.995598	0.995969
14	0.9939899	1.002197	1.002197	0.992257	0.990856	0.994034	0.994296	0.993082	0.992399
15	1.0023696	0.995114	0.995114	1.001908	0.998206	1.002255	1.002464	1.001841	1.00206
16	1.0052974	0.997972	0.997972	1.005983	1.007787	1.005905	1.005385	1.007281	1.008505
17	1.015158	1.000563	1.000563	1.009967	1.006765	1.016068	1.015821	1.016535	1.01718
18	0.990437	0.995331	0.995331	0.997462	0.998874	0.989886	0.990011	0.989558	0.989667
19	0.994032	0.999544	0.999544	0.98616	0.984402	0.994311	0.994681	0.992762	0.991986
20	1.0064139	0.997557	0.997557	1.00522	1.002924	1.006071	1.006156	1.006558	1.007255
21	0.994808	0.99929	0.99929	0.986252	0.987063	0.995099	0.995249	0.994634	0.994426
22	0.9906013	0.99895	0.99895	0.994771	0.997865	0.990342	0.989954	0.991208	0.990735
23	1.0074988	0.9999	0.9999	1.006664	1.003209	1.007523	1.00743	1.008094	1.007787
24	1.0004449	0.995041	0.995041	1.004326	1.004028	1.000066	0.999908	1.000973	1.00147

Σχήμα 3.7: Ενδεικτικές τιμές πίνακα μέσων τιμών

Το παραπάνω Σχήμα δεν παρέχει τόσες πληροφορίες όσες ο επόμενος. Ο επόμενος Πίνακας είναι διαστάσεων k * k και περιέχει του Συντελεστή Ομοιότητας μεταξύ των clusters του παραπάνω Πίνακα. Πρακτικά υπολογίζεται η ομοιότητα των αντίστοιχων τιμών κάθε στήλης με όλες τις υπόλοιπες. Όπως είναι λογικό, ο Πίνακας αυτός είναι διαγώνιος συμμετρικός με μονάδες στη διαγώνιο του, καθώς για παράδειγμα η συσχέτιση της στήλης 4 με τη στήλη 5 είναι ίδια με αυτήν της στήλης 5 με τη στήλη 4 και η στήλη 3 με τη στήλη 3 έχει ομοιότητα ίση με τη μονάδα. Στη συνέχεια, για κάθε στήλη του νέου πίνακα, υπολογίζεται η μέγιστη τιμή (εξαιρέτως της μονάδας σε κάθε στήλη) και η γραμμή με τις μέγιστες τιμές εκφράζει τη συνολική ομοιότητα των Clusters του δείγματος. Το αποτέλεσμα μοιάζει κάπως έτσι:

Στήλες	1	2	3	4	5	6	7	8	9
1	1	-0.061	-0.061	0.78	0.547	0.998	0.998	0.994	0.986
2	-0.061	1	1	-0.157	-0.142	-0.04	-0.041	-0.054	-0.069
3	-0.061	1	1	-0.157	-0.142	-0.04	-0.041	-0.054	-0.069
4	0.78	-0.157	-0.157	1	0.918	0.757	0.746	0.792	0.815
5	0.547	-0.142	-0.142	0.918	1	0.518	0.503	0.565	0.597
6	0.998	-0.04	-0.04	0.757	0.518	1	0.999	0.995	0.987
7	0.998	-0.041	-0.041	0.746	0.503	0.999	1	0.991	0.981
8	0.994	-0.054	-0.054	0.792	0.565	0.995	0.991	1	0.997
9	0.986	-0.069	-0.069	0.815	0.597	0.987	0.981	0.997	1
MAX	0.998	1	-0.04	0.918	0.597	0.999	0.991	0.997	1

Σχήμα 3.8: Ενδεικτικές τιμές πίνακα στηλών ομοιότητας

Παρατηρείται ότι όντως οι τιμές της διαγώνιου είναι ίσες με τη μονάδα καθώς και ότι ο Πίνακας είναι συμμετρικός. Κάθε κελί του Πίνακα αναπαριστά την ομοιότητα των στηλών στις οποίες αντιστοιχεί, για παράδειγμα η στήλη 5 με τη στήλη 1 είναι όμοιες κατά 54.7%. Επίσης φαίνεται ότι όλες οι στήλες είναι συμπληρωμένες άρα πρόκειται για τρέξιμο του αλγόριθμου με *epsilon* ικανό να παραγάγει 9 Clusters.
Ορίζοντας ως κριτήριο ότι η ελάχιστη δυνατή ομοιότητα ώστε να θεωρούνται δύο Clusters αρκετά όμοια για ομαδοποίηση ίση με 95%, εκτελούνται τα τρεξίματα του αλγορίθμου με διαφορετικό *epsilon* κάθε φορά αντίστοιχο του μέγιστου αριθμού Clusters. Μετά από κάθε τρέξιμο, υπολογίζονται οι μέγιστες ομοιότητες και κρίνεται αν γίνεται περαιτέρω ομαδοποίηση. Στον συγκεκριμένο Πίνακα, υπάρχουν πολλές μέγιστες ομοιότητες πάνω από την τιμή 0.95 άρα το *epsilon* θα πρέπει να αυξηθεί και μάλιστα αρκετά. Όταν τελικά καταλήξει ο αλγόριθμος σε ένα μητρώο με ομοιότητες χαμηλότερες του 0.95 τότε σταματά και παράγει το διάγραμμα συσχέτισης στοιχείων που ανήκουν σε κάθε Cluster ανάλογα με τη Διορθωμένη Ισχύ τους. Το τελικό αυτό διάγραμμα περιέχει πληροφορίες για:

- Το μέγιστο βαθμό ομαδοποίησης που μπορεί να επιτευχθεί στο δείγμα χωρίς αλλαγή των θερμοκρασιακών προφίλ,
- Το εύρος λειτουργίας της μηχανής με το ίδιο θερμοκρασιακό προφίλ αλλά και

- Ποιες ομάδες σημείων συγχωνεύτηκαν σε ποιο cluster ώστε να γίνει διακριτή η ομαδοποίηση.

Έτσι, αν τα clusters δεν είναι ανεξάρτητα μεταξύ τους, τότε θα πρέπει να γίνει συγχώνευση η οποία επιτυγχάνεται μειώνοντας τον αριθμό των σχηματιζόμενων Clusters και κάνοντας ξανά ομαδοποίηση. Όταν λοιπόν δεν υπάρχουν επικαλύψεις και τα clusters βρεθούν ανεξάρτητα, έχουμε το τελικό μας αποτέλεσμα.

3.4 Εφαρμογή σε πραγματικά δεδομένα

3.4.1 Αεριοστρόβιλος GTA

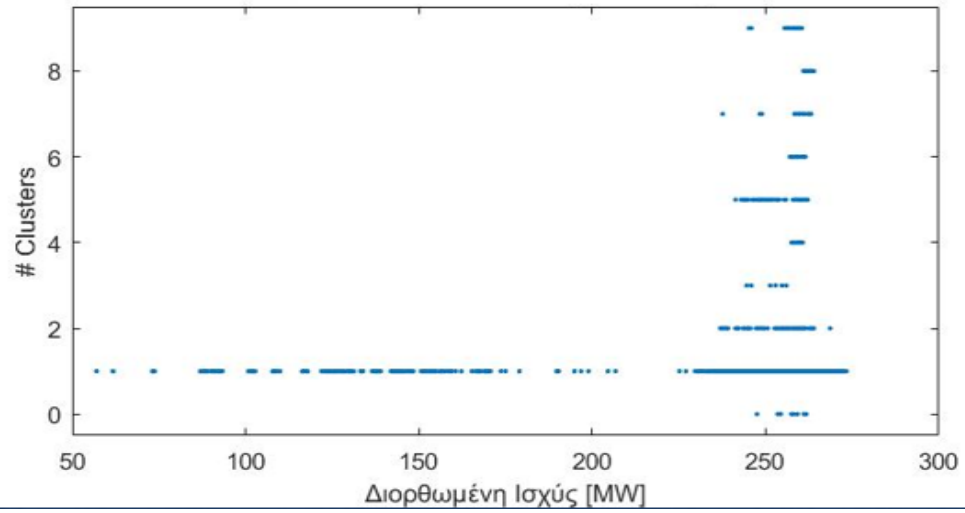
Σε αυτήν τη μελέτη εξετάζεται η λειτουργία ενός στροβιλοκινητήρα, στον οποίο καταγράφονται διάφορες μεταβλητές για τη χρονική περίοδο Φεβρουαρίου μέχρι Δεκεμβρίου του 2016. Για αυτήν την περίοδο παρέχονται διάφορα δεδομένα μεταξύ των οποίων, 16 τιμές θερμοστοιχείων και η διορθωμένη ισχύς του κινητήρα για αυτές τις τιμές. Ακολουθώντας την ίδια διαδικασία που περιεγράφηκε νωρίτερα, οι τιμές των θερμοστοιχείων μετατρέπονται σε βαθμούς Kelvin, αδιαστατοποιούνται βάσει της μέσης τιμής ανά χρονική στιγμή με τη βοήθεια υπολογιστικού φύλλου Excel. Στη συνέχεια, τα δεδομένα εισάγονται στο προγραμματιστικό περιβάλλον της MATLAB και εφαρμόζεται η μέθοδος ομαδοποίησης DBSCAN. Για την ανάλυση των δεδομένων χρησιμοποιήθηκαν τρείς τρόποι ομαδοποίησης. Οι δύο από αυτούς αφορούν διαφορετικές ρυθμίσεις του ίδιου αλγόριθμου ομαδοποίησης της DBSCAN ενώ ο τρίτος είναι ένας πιο απλός έλεγχος που εφαρμόζεται από το Εργαστήριο Θερμικών Στροβιλοκινητήρων ο οποίος θα αναλυθεί παρακάτω. Για τους δύο πρώτους τρόπους οι τιμές των μεταβλητών εισόδου ήταν οι εξής:

- $Epsilon = 0.05, MinPts = 17, w = 2$
- $Epsilon = 0.10, MinPts = 17, w = 2$

Οι τιμές αυτές δεν επιλέχθηκαν τυχαία καθώς κάθε μία από αυτές εκπληρώνει διαφορετικό σκοπό. Βάσει της θεωρίας εφαρμογής της DBSCAN το μέγεθος *MinPts* διαλέγεται συνήθως ίσο με τον αριθμό των διαστάσεων των δεδομένων επαυξημένο κατά ένα, άρα για 16 θερμοστοιχεία το μέγεθος αυτό παίρνει την τιμή 17. Επίσης, η μέθοδος αυτή αποσκοπεί στην ομαδοποίηση των δεδομένων και όχι στην διαγνωστική ταξινόμηση οπότε θεωρήθηκε καταλληλότερο να χρησιμοποιηθεί ως συνάρτηση απόστασης ο Συντελεστής αλληλοσυσχέτισης, για αυτό το w παίρνει την τιμή 2. Τέλος, εφόσον χρησιμοποιείται ο συντελεστής αλληλοσυσχέτισης ως συνάρτηση απόστασης, δίνεται η δυνατότητα ρύθμισης της ελάχιστης επιτρεπτής ομοιότητας μέσω του *epsilon*. Τα *epsilon* στις δύο παραπάνω περιπτώσεις παίρνουν τιμές 0.05 και 0.10 για ελάχιστη επιτρεπτή ομοιότητα ίση με 95% και 90% αντίστοιχα, που προκύπτει από τη σχέση:

$$Min(ccd) = 1 - epsilon$$

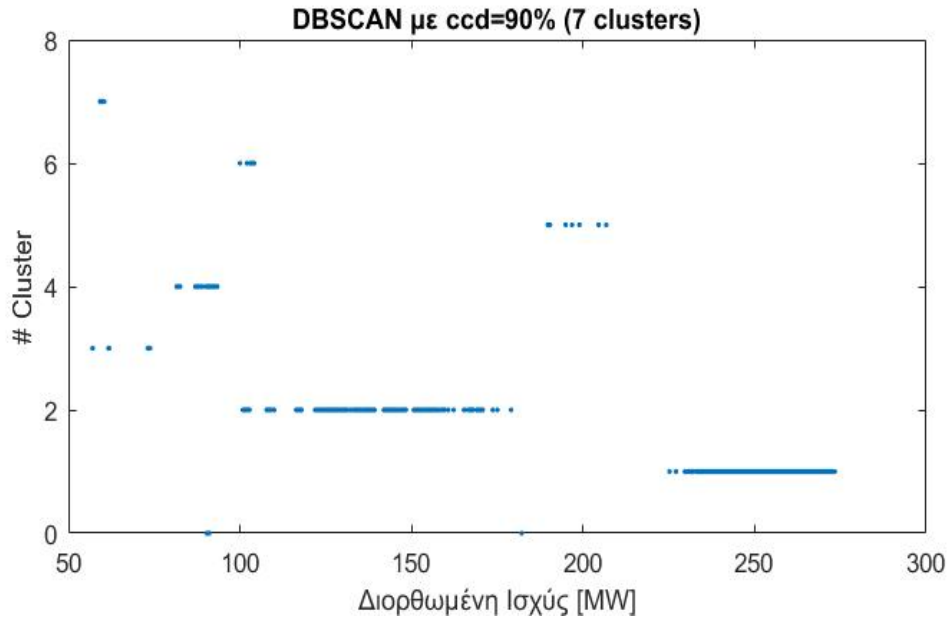
Ακολουθώντας το παραπάνω διάγραμμα ροής, γίνεται μία πρώτη ομαδοποίηση του συνόλου των δεδομένων για CCD=90%. Τα αποτελέσματα φαίνονται στο παρακάτω σχήμα:



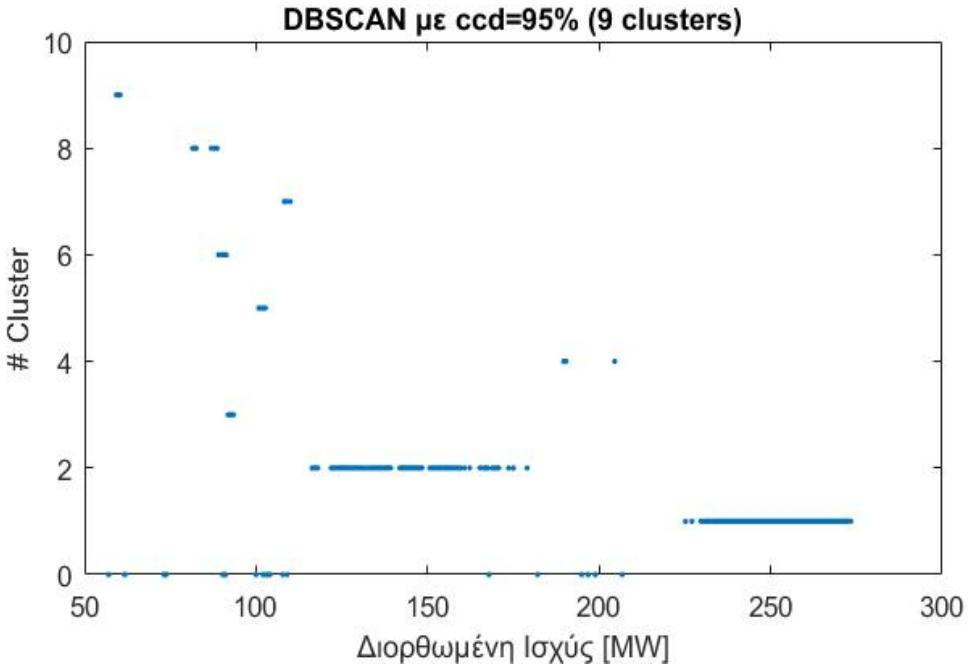
Σχήμα 3.9: Αποτέλεσμα clustering όλων των δεδομένων με CCD=90%

Παρατηρείται ότι υπάρχει έντονη επικάλυψη μεταξύ των 9 clusters που σχηματίστηκαν, ειδικά στην περιοχή των 250 MW. Βάσει λοιπόν της θεωρίας και του διαγράμματος ροής, έχει επιβληθεί κάποια μεταβολή στη μηχανή οπότε θα πρέπει να μειωθούν τα εξεταζόμενα δεδομένα για να βρεθεί σε ποιο σημείο συνέβη η μεταβολή αυτή. Για τους δύο παραπάνω λόγους λοιπόν, χρησιμοποιήθηκαν οι πρώτες 6000 (από τις συνολικά 22500) μετρήσεις για την επόμενη ομαδοποίηση. Βάσει αυτών των μετρήσεων, σχηματίζονται cluster ομοιότητας το καθένα με το δικό του εύρος διορθωμένης ισχύος. Διαγράμματα που περιγράφουν τη συσχέτιση cluster ομοιότητας με τη διορθωμένη ισχύ παράχθηκαν και για τους άλλους κινητήρες που αναλύθηκαν νωρίτερα. Αξίζει να θυμηθούμε ότι σε αυτά τα διαγράμματα ιδανικό είναι να υπάρχουν ξεκάθαρα ορισμένες περιοχές ισχύος για κάθε cluster χωρίς επικαλύψεις μεταξύ τους. Από το διάγραμμα αυτό μπορεί εύκολα να ορισθεί το εύρος ισχύος κάθε cluster.

Ακολούθως, από τα clusters που σχηματίστηκαν, είναι δυνατό να παραχθεί ο Πίνακας των μέσων προφίλ που προαναφέρθηκε, διαστάσεων $D * n$, όπου D οι διαστάσεις των δεδομένων και n ο αριθμός των clusters που σχηματίστηκαν από τον αλγόριθμο ομαδοποίησης. Έχοντας λοιπόν τα μέσα προφίλ και το εύρος ισχύος κάθε cluster για τα πρώτα 6000 σημεία, είναι δυνατή η ταξινόμηση του συνόλου των δεδομένων. Η ταξινόμηση γίνεται χρησιμοποιώντας την διορθωμένη ισχύ κάθε μέτρησης ως ενδεικτική τιμή. Αφού ταξινομηθούν όλα τα σημεία στο cluster αναφοράς που θεωρητικά ανήκουν, υπολογίζεται ο συντελεστής αλληλοσυσχέτισης μεταξύ κάθε σημείου και του μέσου προφίλ ανάλογα με το cluster που ταξινομήθηκε. Η διαδικασία αυτή μπορεί να δώσει μία συνολική εικόνα για την κατάσταση λειτουργίας του κινητήρα, σε βάθος πολλών μετρήσεων. Παρακάτω φαίνονται τα διαγράμματα συσχέτισης cluster ταξινόμησης και διορθωμένης ισχύος για τις δύο ρυθμίσεις του αλγόριθμου που αναφέρθηκαν παραπάνω:



Σχήμα 3.10: Συσχέτιση cluster ταξινόμησης και διορθωμένης ισχύος για CCD=90%



Σχήμα 3.11: Συσχέτιση cluster ταξινόμησης και διορθωμένης ισχύος για CCD=95%

Όπως φαίνεται παραπάνω ο ελάχιστος αποδεκτός βαθμός ομοιότητας επηρεάζει τον αριθμό των clusters. Όσο αυξάνεται ο CCD αυξάνεται και ο συνολικός αριθμός clusters. Το φαινόμενο αυτό είναι λογικό, καθώς ο αλγόριθμος κατατάσσει περισσότερα σημεία στην ίδια γειτονιά όταν το κριτήριο γειτονίας είναι χαλαρότερο. Έτσι, δε χρειάζονται 9 clusters για παράδειγμα όταν με 90% ομοιότητα τα δεδομένα «χωρούν» σε

7 clusters. Επίσης παρατηρείται ότι τα εύρη ισχύος στις δύο διαδικασίες καλύπτουν τις ίδιες τιμές στον οριζόντιο άξονα, φαινόμενο λογικό αφού η ομαδοποίηση έγινε με τα ίδια δεδομένα, τις πρώτες 6000 μετρήσεις. Πιο συγκεκριμένα, το δεύτερο διάγραμμα φαίνεται περισσότερο ως αναλυτική μορφή του πρώτου. Τα clusters στο δεύτερο διάγραμμα καλύπτουν σχεδόν τις ίδιες περιοχές με αυτά του πρώτου με εξαίρεση τα δύο επιπλέον clusters που σκοπός της δημιουργίας τους είναι η αύξηση της ακρίβειας όπως θα φανεί παρακάτω.

Συνεχίζοντας, εφόσον παράχθηκαν τα διαγράμματα συσχέτισης Ισχύος-Clusters, μπορούν πλέον να οριστούν τα εύρη λειτουργίας κάθε cluster για τη μετέπειτα ταξινόμηση. Οι παρακάτω πίνακες δίνουν τα εύρη για κάθε cluster για τις δύο ομαδοποιήσεις που πραγματοποιήθηκαν:

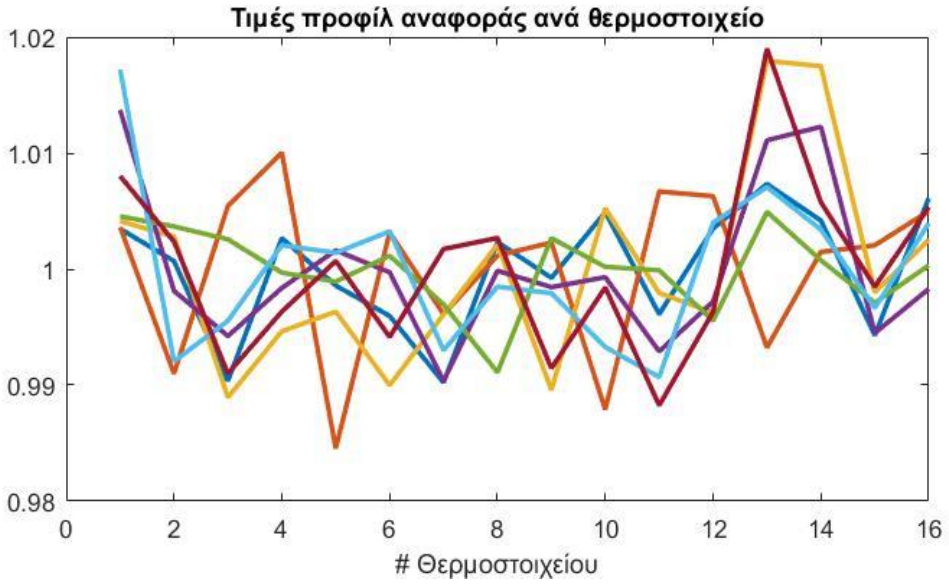
# Cluster	7	3	4	6	2	5	1
Min	35	60	80	95	105	185	220
Max	60	80	95	105	185	220	280

Σχήμα 3.12: Εύρος ισχύος ανάλογα με το cluster ταξινόμησης

# Cluster	9	8	6	3	5	7	2	4	1
Min	35	60	80	92	95	105	115	185	220
Max	60	80	92	95	105	115	185	220	280

Σχήμα 3.13: Εύρος ισχύος ανάλογα με το cluster ταξινόμησης

Παρατηρείται ότι οι τιμές κυμαίνονται σε παρόμοια όρια, με κύρια διαφορά την αυξημένη διακριτοποίηση στο δεύτερο πίνακα καθώς σχηματίζονται περισσότερα clusters. Επίσης να σημειωθεί ότι η ελάχιστη και η μέγιστη τιμή στους παραπάνω πίνακες, δεν είναι ίσες με αυτές των διαγραμμάτων συσχέτισης. Αυτό συνέβη επειδή στην ομαδοποίηση του συνόλου των δεδομένων (22500 σημείων) οι τιμές διορθωμένης ισχύος κυμαίνονται σε μεγαλύτερο εύρος από ότι αυτές των πρώτων 6000 και κάθε σημείο πρέπει αναγκαστικά να κατανεμηθεί σε ένα cluster. Στα Σχήματα 3.12 και 3.13 φαίνεται επίσης η ομοιότητα των ευρών λειτουργίας κάθε cluster σε όλο το μήκος του οριζόντιου άξονα με μερικές εξαιρέσεις. Για παράδειγμα, τα cluster 2 και 7 του Σχήματος 3.13 είναι πρακτικά ένας λεπτομερέστερος διαχωρισμός του cluster 2 του Σχήματος 3.12. Το ίδιο ισχύει για τα cluster 6 και 3 του Σχήματος 3.13 με το cluster 4 του Σχήματος 3.12. Επίσης, από τον αλγόριθμο ομαδοποίησης παράχθηκαν και τα μέσα προφίλ θερμοκρασιών ανά Cluster όπως προαναφέρθηκε. Η τάση κίνησης του φαίνεται στα παρακάτω διαγράμματα για τις δύο περιπτώσεις ομαδοποίησης:

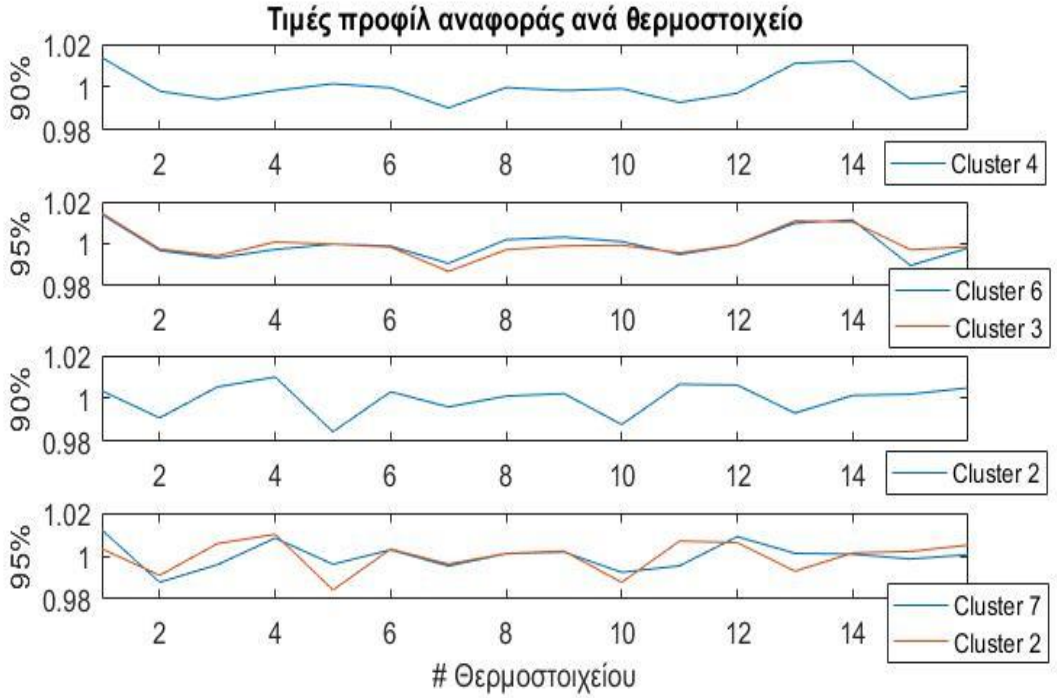


Σχήμα 3.14: Τιμές θερμοκρασιακών προφίλ για ομαδοποίηση με CCD=90%, 7 καμπύλες

Σχήμα 3.15: Τιμές θερμοκρασιακών προφίλ για ομαδοποίηση με CCD=95%, 9 καμπύλες

Εδώ φαίνεται η ανεξαρτησία και η διαφορετικότητα των προφίλ θερμοκρασιών που σχηματίζονται, καθώς οι καμπύλες δεν έχουν καθόλου όμοια τάση κίνησης. Επίσης παρατηρείται ότι οι τιμές κάθε καμπύλης είναι πολύ κοντά στη μονάδα, φαινόμενο που αποδεικνύει ότι πρόκειται για μέσα προφίλ. Τα σχήματα 3.14 και 3.15 έχουν όμως παρόμοια μορφή μεταξύ τους μίας και πρόκειται για ομαδοποιήσεις των ίδιων δεδομένων και σε σχεδόν ίδιες περιοχές φορτίου. Όπως φάνηκε και νωρίτερα, το σχήμα 3.12 αποτελεί γενίκευση του σχήματος 3.13 και αυτό φαίνεται και στα παραπάνω διαγράμματα. Πιο συγκεκριμένα, μελετώντας τα προφίλ που η διαδικασία του σχήματος 3.13 συγχώνευσε

προκύπτει ότι τα clusters 4 και 2 του σχήματος 3.12 αντιστοιχούν στα clusters 6, 3, 7 και 2. Στο παρακάτω σχήμα φαίνεται η μεταβολή των θερμοστοιχείων για τα δεδομένα που ανήκουν σε αυτά τα clusters.



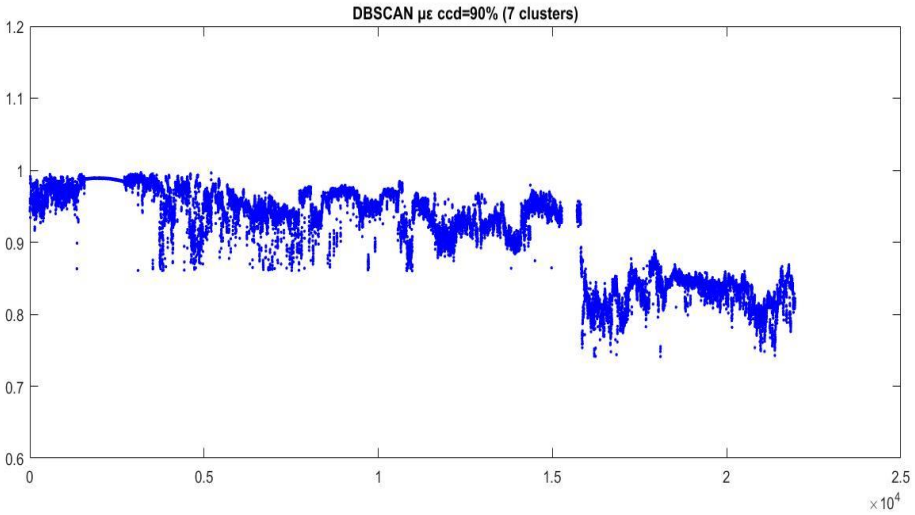
Σχήμα 3.16: Τιμές θερμοκρασιακών προφίλ για ομαδοποίηση με CCD=95% και CCD=90%

Όπως είναι λογικό οι τιμές των καμπυλών 90% μοιάζουν με τις αντίστοιχες των 95% καθώς περιγράφουν τα ίδια συνολικά δεδομένα. Επίσης παρατηρείται στα διαγράμματα 95% ότι οι καμπύλες έχουν μικρές διαφορές και συνολικά ίδια τάση κίνησης. Αυτό οφείλεται στην υψηλή ομοιότητα των clusters αυτών. Ένας επιπλέον τρόπος να αποδειχτεί ότι τα clusters αυτά έχουν υψηλή ομοιότητα μεταξύ τους είναι η σύγκριση των Πινάκων 3.12 και 3.13. Δηλαδή, για την ίδια περιοχή φορτίου τα δύο διαφορετικά τρεξίματα της DBSCAN αποφάσισαν να εκχωρήσουν τα σημεία σε ένα ή δύο clusters αντίστοιχα ανάλογα με την ελάχιστη αποδεκτή ομοιότητα των προφίλ. Μεταξύ των δύο τρεξιμάτων η ομοιότητα αυτή μειώνεται και εμφανίζονται λιγότερα clusters (λόγω μεγαλύτερης ομαδοποίησης) φαινόμενο που προδίδει ότι τα clusters που «εξαφανίστηκαν» από το τρέξιμο 95% στο τρέξιμο 90% είχαν ομοιότητα 90 με 95%. Αυτό αποδεικνύεται και στην πράξη καθώς τα εν λόγω clusters έχουν τις εξής ομοιότητες:

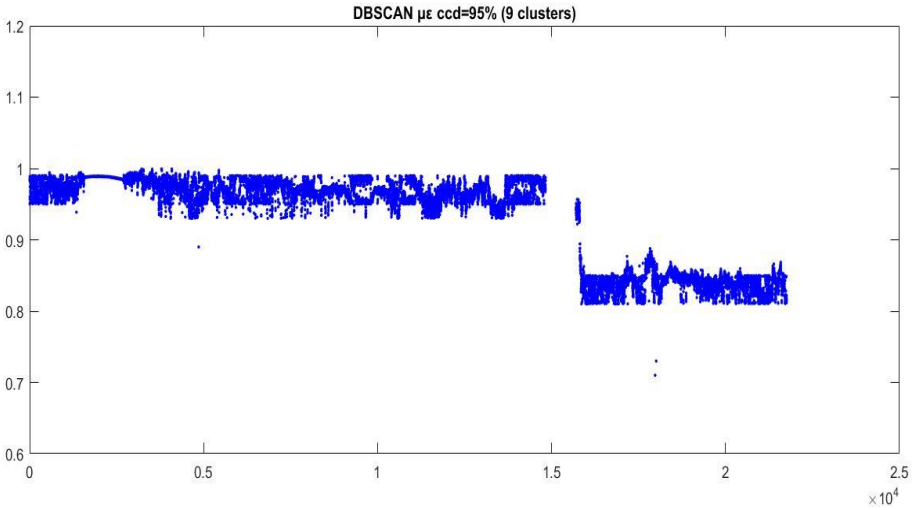
$$CCD_{6-3} = 92.13\%$$
$$CCD_{7-2} = 94.39\%$$

Στη συνέχεια, βάσει των προηγούμενων σχημάτων που αναλύθηκαν πραγματοποιείται ταξινόμηση και υπολογισμός του CCD για όλα τα σημεία του δείγματος. Δηλαδή υπολογίζεται για κάθε σημείο από τα 22000 ο συντελεστής ομοιότητας με το μέσο προφίλ που αντιστοιχεί στη διορθωμένη ισχύ που το περιγράφει. Παρακάτω

φαίνονται τα διαγράμματα που παράχθηκαν βάσει του συντελεστή αλληλοσυσχέτισης για όλα τα σημεία συναρτήσει των χρονικών μετρήσεων:



Σχήμα 3.17: Συντελεστής ομοιότητας CCD μετά από ταξινόμηση σε 7 clusters

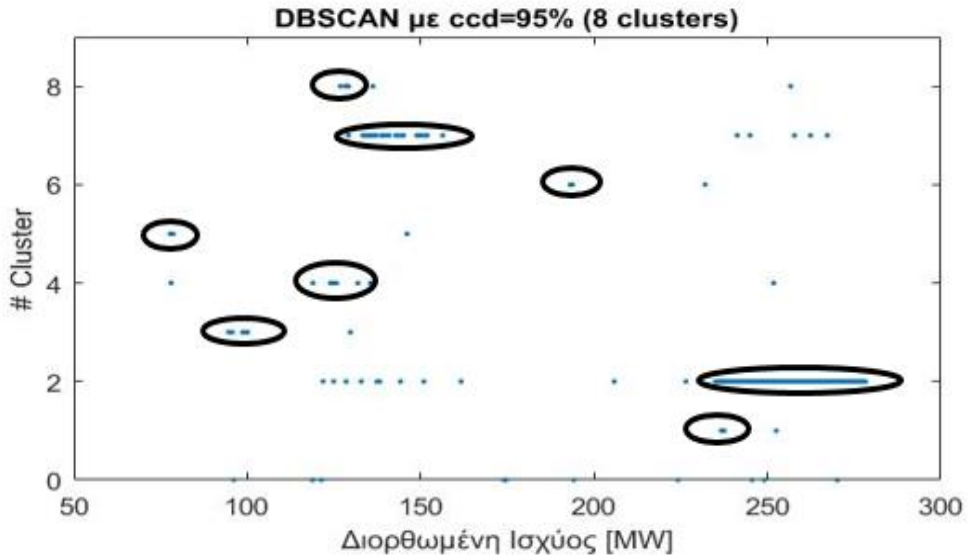


Σχήμα 3.18: Συντελεστής ομοιότητας CCD μετά από ταξινόμηση σε 9 clusters

Εδώ φαίνεται ότι μέσω της μεθόδου ταξινόμησης τα δεδομένα κατανέμονται σωστά καθώς ο συντελεστής ομοιότητας είναι πολύ κοντά στη μονάδα. Επίσης, φαίνεται ότι κατά μέσο όρο οι τιμές του σχήματος 3.18 προσεγγίζουν καλύτερα τη μονάδα από ότι αυτές του σχήματος 3.17. Αυτό συμβαίνει επειδή κατανέμοντας τα σημεία σε περισσότερα clusters επιτυγχάνεται μεγαλύτερο CCD καθώς οι μέσες τιμές εκπροσωπούν καλύτερα τα σημεία που κατανέμονται εκεί. Η τρίτη και κυριότερη παρατήρηση πάνω στα διαγράμματα αυτά αφορά στις τιμές μετά τη μέτρηση 16000 όπου υπάρχει μία σταθερή και σημαντική μείωση στο συντελεστή.

Μετά από διερεύνηση των δεδομένων προέκυψε ότι ο κινητήρας σε αυτό το διάστημα διέκοψε τη λειτουργία του και επανήλθε με διαφορετικές ρυθμίσεις, συγκεκριμένα συνέβη hot section inspection. Έτσι, όπως είναι λογικό, τα παλιά προφίλ δε

μπορούσαν να πετύχουν υψηλή ομοιότητα με τα νέα δεδομένα. Για το λόγο αυτό, πραγματοποιήθηκε νέα ομαδοποίηση στα νέα δεδομένα με την ίδια διαδικασία και τα αποτελέσματα φαίνονται παρακάτω:

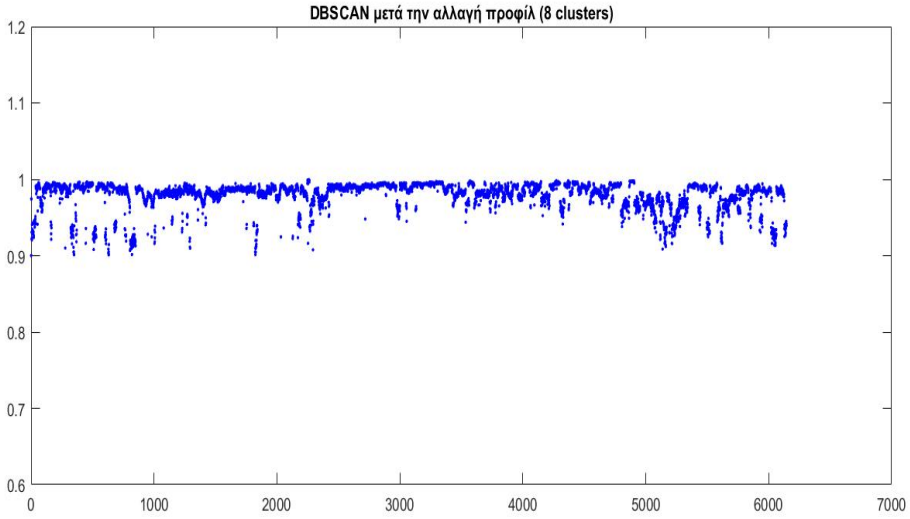


Σχήμα 3.19: Συσχέτιση cluster ταξινόμησης και διορθωμένης ισχύος για CCD=95% του 2ου τμήματος

Σχήμα 3.20: Τιμές θερμοκρασιακών προφίλ για ομαδοποίηση με CCD=95%, 7 καμπύλες του 2ου τμήματος

# Cluster	4	2	3	7	6	5	1
Min	75	95	110	125	135	185	220
Max	95	110	125	135	185	220	280

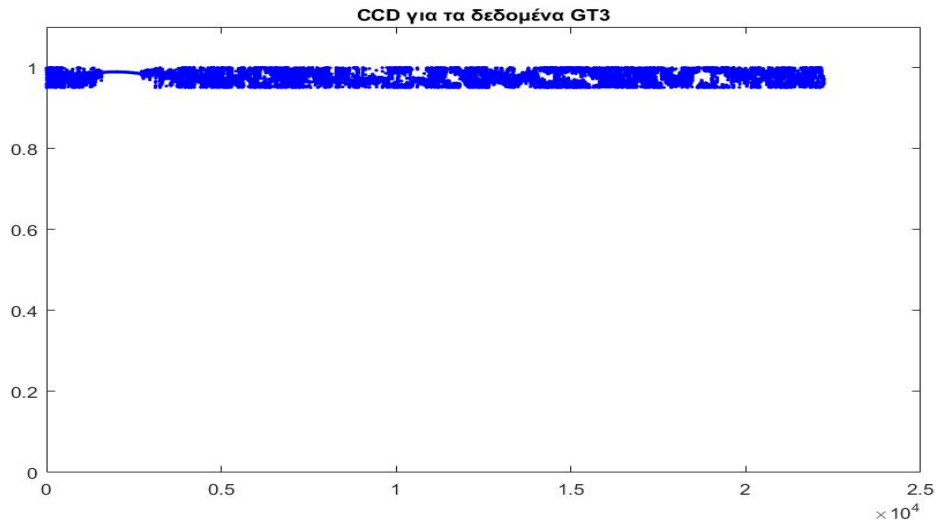
Σχήμα 3.21: Εύρος ισχύος ανάλογα με το cluster ταξινόμησης του 2ου τμήματος



Σχήμα 3.22: Συντελεστής ομοιότητας CCD μετά από ταξινόμηση σε 8 clusters

Σύμφωνα με την πρώτη ανάλυση η DBSCAN λειτουργεί ακριβέστερα με ελάχιστη αποδεκτή ομοιότητα ίση με 95% οπότε ο έλεγχος της δεύτερης περιοχής θα γίνει με αυτές τις ρυθμίσεις. Τα Σχήματα 3.19 και 3.20 μας δείχνουν τα αποτελέσματα του αλγορίθμου ομαδοποίησης ο οποίος λειτούργησε με *epsilon*=0.05 για ελάχιστη αποδεκτή ομοιότητα 95% και παρήγαγε 7 clusters. Στο Σχήμα 3.21 φαίνονται τα εύρη κάθε Cluster ταξινόμησης με τα όρια να καταλαμβάνουν όλες τις τιμές μεταξύ της ελάχιστης και μέγιστης των 75 και 280 MW αντίστοιχα. Τα εύρη των clusters που σχηματίζονται μοιάζουν με αυτά που σχηματίστηκαν για τα πρώτα 6000 σημεία με κυριότερη διαφορά τις ακραίες τιμές που εξαρτώνται αποκλειστικά από τα δεδομένα παρακολούθησης της δεύτερης περιόδου. Στο Σχήμα 3.22 φαίνεται ο συντελεστής ομοιότητας συναρτήσει του χρόνου με ικανοποιητικές τιμές καθώς τα σημεία κατανέμονται σωστά και δεν υπάρχουν περαιτέρω αλλαγές στον στροβιλοκινητήρα.

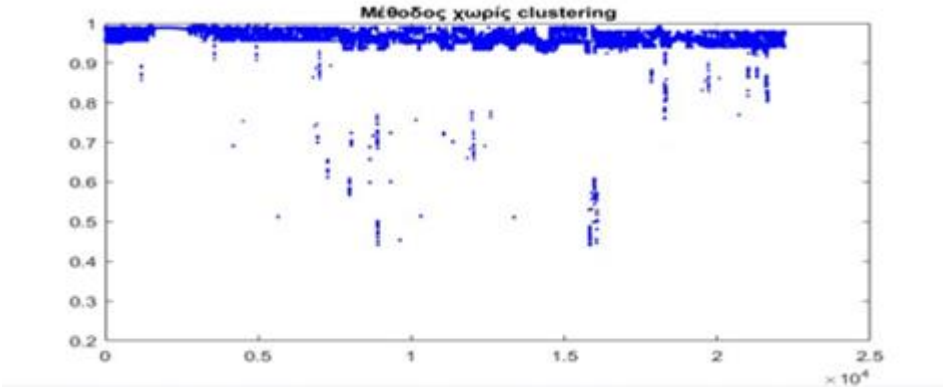
Ο σκοπός παραγωγής νέων μέσων προφίλ για τη δεύτερη περιοχή είναι η συνολικά σωστή παρακολούθηση του κινητήρα με αποφυγή τυχόν “false alarm” στον κινητήρα λόγω μίας απλής ρύθμισης του. Έχοντας λοιπόν προφίλ για όλες τις τιμές του κινητήρα, είναι δυνατόν πλέον να υπολογισθεί ο συντελεστής ομοιότητας στο σύνολο των μετρήσεων ενώ κάθε σημείο θα ταξινομείται σε ένα cluster αναφοράς από τα παραπάνω με κριτήριο της ισχύ που το χαρακτηρίζει. Το τελικό διάγραμμα φαίνεται παρακάτω:



Σχήμα 3.23: Συντελεστής ομοιότητας CCD στο σύνολο των μετρήσεων

Όπως φαίνεται τα αποτελέσματα είναι πολύ θετικά καθώς όλες οι μετρήσεις έχουν συντελεστή ομοιότητας πολύ κοντά στη μονάδα και καμία τιμή δεν πέφτει κάτω από 90% ομοιότητα. Έτσι, παράχθηκε ένα μοντέλο παρακολούθησης του στροβιλοκινητήρα για όλα τα εύρη λειτουργίας, βάσει της ομαδοποίησης των θερμοκρασιακών δεδομένων μέσω της DBSCAN ανεξαρτήτως αλλαγής του κινητήρα.

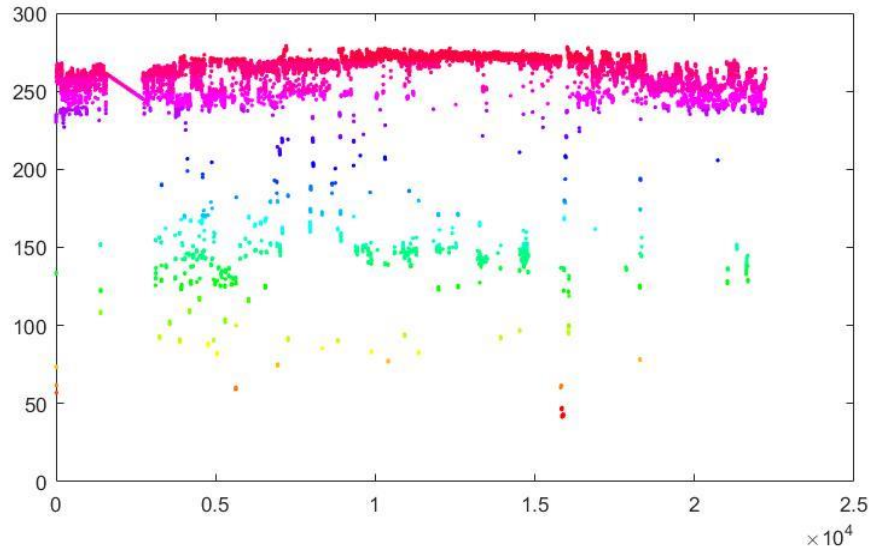
Παράλληλα με την παραπάνω μελέτη, αναλύθηκε και η διαδικασία που χρησιμοποιείται ήδη από το εργαστήριο για την ταξινόμηση των σημείων. Η διαδικασία αυτή δεν περιέχει αλγόριθμο clustering και κατανέμει όλα τα εισαγόμενα δεδομένα ανάλογα με τη διορθωμένη ισχύ που τα χαρακτηρίζει. Πιο συγκεκριμένα, η μέθοδος αυτή χωρίζει το συνολικό εύρος ισχύος σε μικρά ίσα τμήματα των 10 MW εκτός από τις πολύ μικρές τιμές που στην περίπτωση μας ομαδοποιούνται σε ένα τμήμα 0 έως 60 MW. Έτσι, κάθε νέο σημείο κατανέμεται ανάλογα με την διορθωμένη ισχύ που το χαρακτηρίζει. Μετά την κατανομή αυτή παράγονται τα μέσα προφίλ ως οι μέσες τιμές ανά θερμοστοιχείο για κάθε τμήμα. Για τα σημεία που έχουν κατανεμηθεί σε κάθε τμήμα υπολογίζεται ο αντίστοιχος συντελεστής ομοιότητας που φαίνεται παρακάτω:



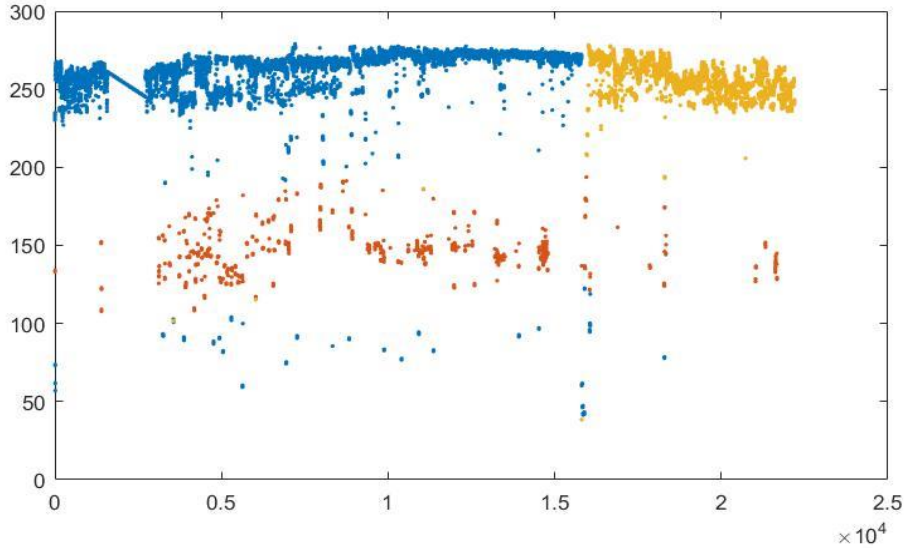
Σχήμα 3.24: Αποτελέσματα μεθόδου ταξινόμησης του εργαστηρίου

Η σημαντική διαφορά του σχήματος 3.24 από τα σχήματα 3.22 και 3.23 είναι στις ακραίες τιμές του διαγράμματος όπου παρατηρούνται διάφορες ακραίες αποκλίσεις για όλο το χρονικό διάστημα σποραδικά αλλά κυρίως στη μεταβλητή κατάσταση της μηχανής που προαναφέρθηκε. Μάλιστα οι τιμές εκεί έχουν τέτοια απόκλιση που κάποιος αλγόριθμος ομοιότητας θα σήμανε κίνδυνο βλάβης. Για αυτό είναι προτιμότερη η επιλογή κάποιας μεθόδου clustering για την παραγωγή μέσων προφίλ.

Η διαφορά των μεθόδων clustering και της μεθόδου του εργαστηρίου περιγράφονται και με ένα δεύτερο τρόπο πέρα από το συντελεστή ομοιότητας. Συγκρίνοντας τα διαγράμματα διορθωμένης ισχύος των διαφορετικών μεθόδων χρωματίζοντας τα δεδομένα ανάλογα με τον τρόπο ομαδοποίησης προκύπτουν σημαντικές παρατηρήσεις:



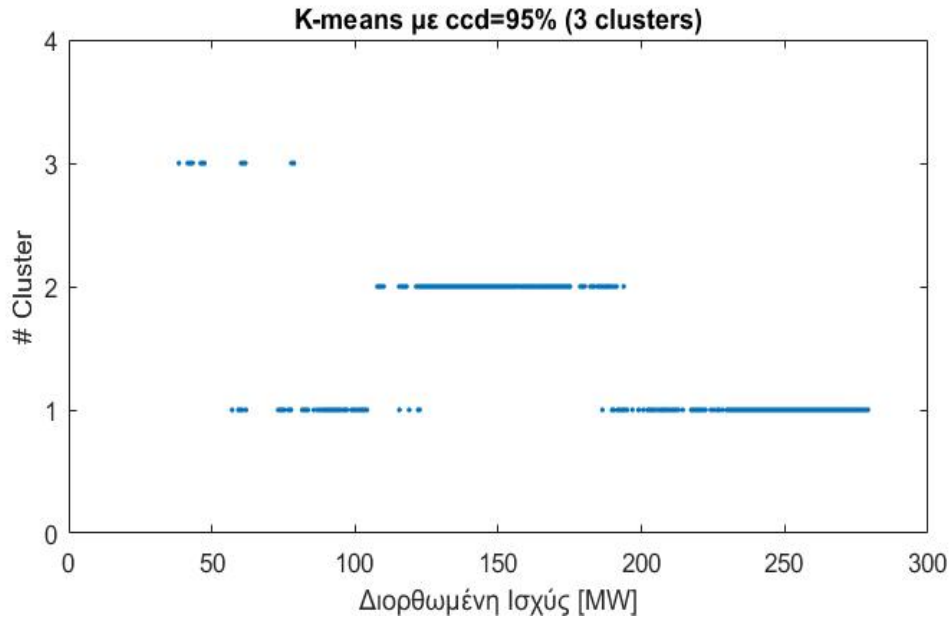
Σχήμα 3.25: Κατανομή δεδομένων χωρίς clustering



Σχήμα 3.26: Κατανομή δεδομένων με clustering

Στα δύο παραπάνω διαγράμματα φαίνεται η κατανομή της διορθωμένης ισχύος σε ομάδες. Το πάνω σχήμα κατανέμει τα σημεία ανάλογα με τη διορθωμένη ισχύ, δηλαδή βάσει της μεθόδου χωρίς clustering, ενώ το αμέσως επόμενο τα κατανέμει με DBSCAN και άρα clustering. Η βασικότερη διαφορά είναι ο αριθμός σχηματιζόμενων clusters καθώς χωρίς clustering σχηματίζονται 24 ενώ με clustering 3. Επίσης παρατηρείται ότι με clustering, μεγάλο κομμάτι της υγειούς κατάστασης περιγράφεται μέσω του μπλε cluster, τυχόν αποκλίσεις περιγράφονται με κόκκινο, ενώ οι μετρήσεις μετά τη μεταβολή των προφίλ λόγω Hot Section Inspection, περιγράφονται με κίτρινο. Η ομαδοποίηση αυτή βοηθά το χρήστη να διαχωρίσει τις τρεις βασικές καταστάσεις της μηχανής αρκετά ικανοποιητικά. Από την άλλη πλευρά η μέθοδος χωρίς clustering χωρίζει την υγιή λειτουργία σε πολλά κομμάτια θεωρώντας ότι για μεταβολή 5 MW και πάνω στη διορθωμένη ισχύ ότι αλλάζει η κατάσταση της μηχανής, φαινόμενο που οδηγεί σε αποκλίσεις του συντελεστή ομοιότητας όπως φάνηκε και νωρίτερα. Επίσης, για την περιοχή μετά τη μεταβολή των προφίλ, γνωρίζουμε ότι λόγω της μεταβολής δε γίνεται τα σημεία να ομαδοποιούνται σε ίδιο προφίλ. Η χρήση ροζ και μωβ χρώματος σε όλο το μήκος των μετρήσεων από τη μέθοδο χωρίς clustering, οδηγεί στο συμπέρασμα ότι δεν υπάρχει μεταβολή, απορρίπτοντας έτσι τη διαγνωστική ικανότητα σε αυτό το σετ δεδομένων.

Στη συνέχεια μελετάται η απόδοση των άλλων δύο μεθόδων clustering για την παραγωγή περιοχών αλλαγής φορτίου. Ξεκινώντας με την K-means όπως αναφέρθηκε και νωρίτερα η επιλογή αριθμού clusters γίνεται ευκολότερα από ότι στη DBSCAN καθώς εισάγεται από το χρήστη ως μεταβλητή εισόδου. Ξεκινώντας λοιπόν και πάλι από υψηλές τιμές αριθμού επιθυμητών clusters τα δεδομένα ομαδοποιούνται, παράγουν τα μέσα προφίλ και μέσω της ομοιότητας των τελευταίων αποφασίζεται αν θα μειωθεί ο αριθμός των clusters ή όχι. Επίσης με την K-means αλλά και με την ιεραρχική μέθοδο όπως θα δούμε αργότερα, δε χρειάζεται έλεγχος μερικών σημείων πριν την κατανομή του συνόλου των δεδομένων. Οι μέθοδοι αυτές, λόγω της απλότητας των υπολογισμών τους μπορούν να διαχειριστούν μεγαλύτερους όγκους δεδομένων από ότι η DBSCAN χωρίς να επηρεάζεται αισθητά το υπολογιστικό κόστος ή η απαιτούμενη μνήμη. Για να υπάρχει άμεση σύγκριση με τη DBSCAN χρησιμοποιήθηκε ο ίδιος αποδεκτός βαθμός ομοιότητας μεταξύ των μεθόδων. Παρακάτω φαίνονται τα αποτελέσματα της K-means με αυτές τις παραμέτρους:

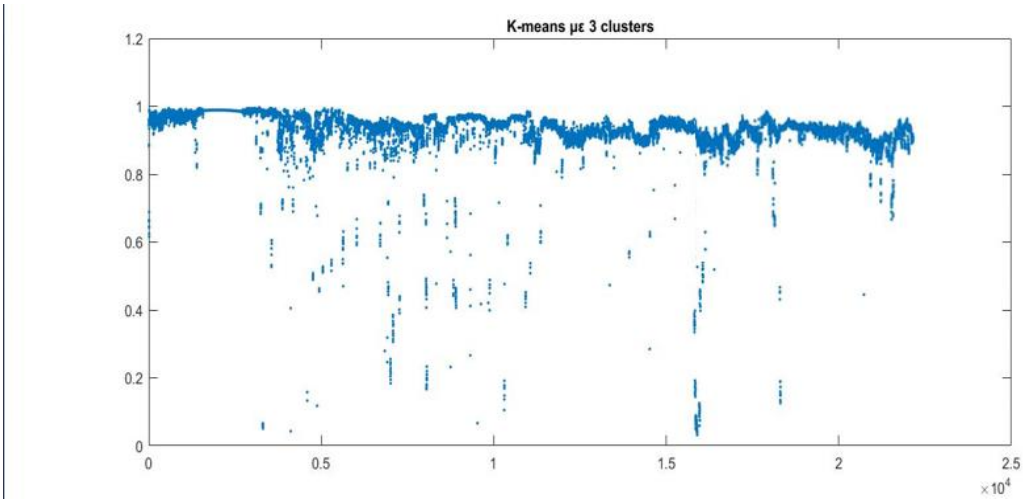


Σχήμα 3.27: Συσχέτιση cluster ταξινόμησης και διορθωμένης ισχύος για CCD=95%

Σχήμα 3.28: Τιμές θερμοκρασιακών προφίλ για ομαδοποίηση με CCD=95%, 3 καμπύλες

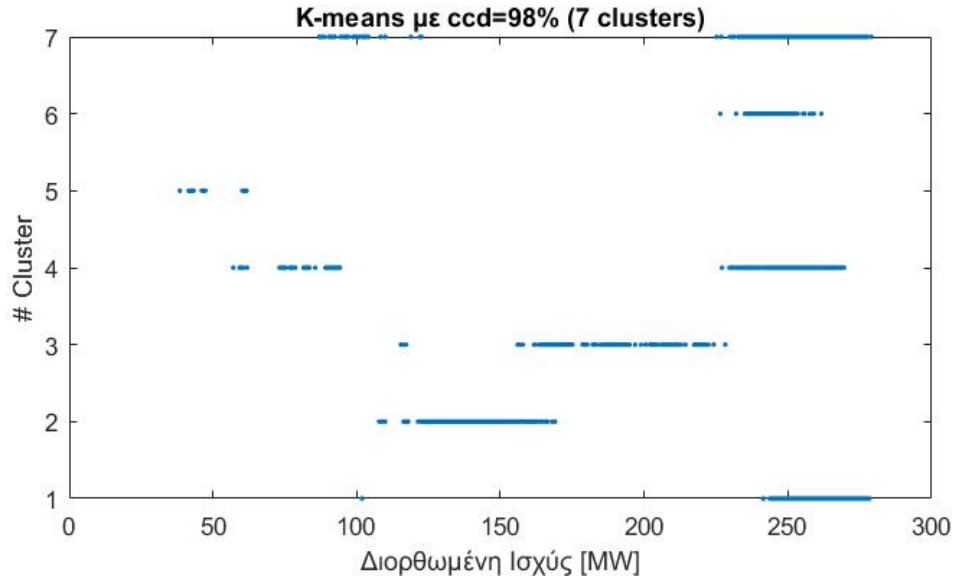
# Cluster	1	2	3	1
Min	38	110	150	205
Max	110	150	205	280

Σχήμα 3.29: Εύρος ισχύος ανάλογα με το cluster ταξινόμησης



Σχήμα 3.30: Συντελεστής ομοιότητας CCD στο σύνολο των μετρήσεων

Όπως φαίνεται η K-means με αποδεκτή ομοιότητα 95% παράγει συνολικά 3 clusters με κυριότερο αυτό με τον αριθμό 1. Παρατηρείται ότι οι μεσαίες τιμές ισχύος καλύπτονται από τα clusters 2 και 3, ενώ οι ακραίες τιμές ομαδοποιούνται στο cluster 1. Πηγαίνοντας ένα βήμα πίσω στην ομαδοποίηση, για 4 clusters οι περιοχές χαμηλών και υψηλών τιμών διαχωρίζονται σε 2 διαφορετικά clusters ενώ τα clusters 2 και 3 μένουν ίδια. Το φαινόμενο αυτό προδίδει ότι οι περιοχές ακραίων τιμών έχουν υψηλή ομοιότητα στο θερμοκρασιακό τους προφίλ λογικά λόγω απόκλισης από το σημείο λειτουργίας. Στο Σχήμα 3.28 φαίνονται τα μέσα θερμοκρασιακά προφίλ και μπορεί κανείς να διακρίνει την ομοιότητα στην τάση κίνησης των θερμοστοιχείων με αυτά της DBSCAN. Το φαινόμενο αυτό αποδεικνύει ότι οι μέθοδοι αυτοί ομαδοποιούν με ίδιο τρόπο τα δεδομένα που τους παρέχονται. Τέλος, στο Σχήμα 3.30 φαίνονται οι τιμές του συντελεστή ομοιότητας ανάλογα με το σημείο εξέτασης. Παρατηρείται η ίδια απότομη πτώση των μέσων τιμών περίπου στη μέτρηση 16000 λόγω της αλλαγής ρυθμίσεων που αναφέρθηκε και παραπάνω. Σημαντική είναι επίσης και η διερεύνηση της ομαδοποίησης με ομοιότητα 95% καθώς εκεί σχηματίζεται παρόμοιος αριθμός clusters με τη DBSCAN. Για μέγιστη αποδεκτή ομοιότητα 98% τα αποτελέσματα της K-means είναι τα εξής:

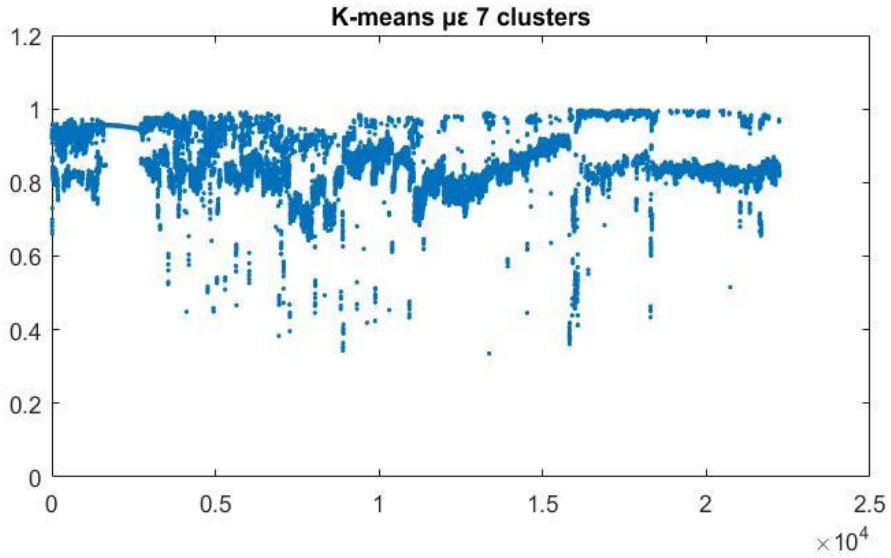


Σχήμα 3.31: Συσχέτιση cluster ταξινόμησης και διορθωμένης ισχύος για CCD=98%

Σχήμα 3.32: Τιμές θερμοκρασιακών προφίλ για ομαδοποίηση με CCD=98%, 7 καμπύλες

# Cluster	5	4	2	3	6	7	1
Min	35	55	105	160	225	250	260
Max	55	105	160	225	250	260	280

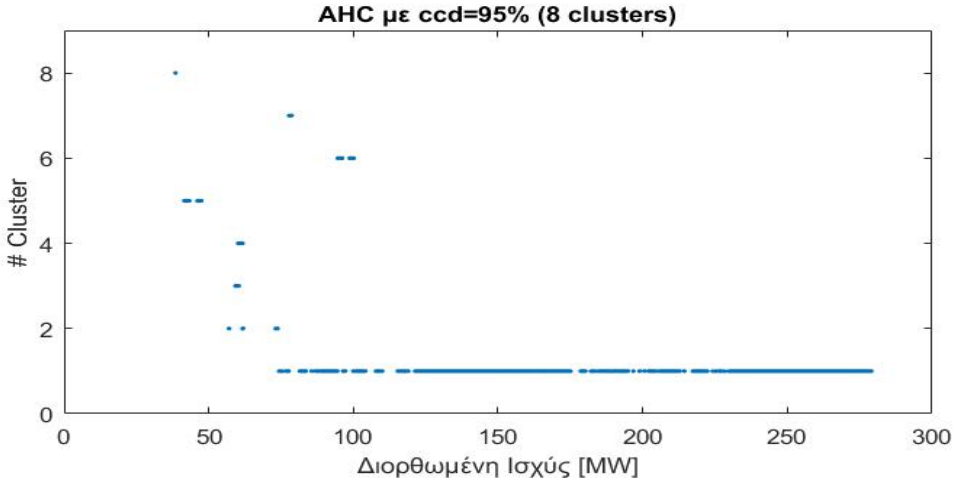
Σχήμα 3.33: Εύρος ισχύος ανάλογα με το cluster ταξινόμησης



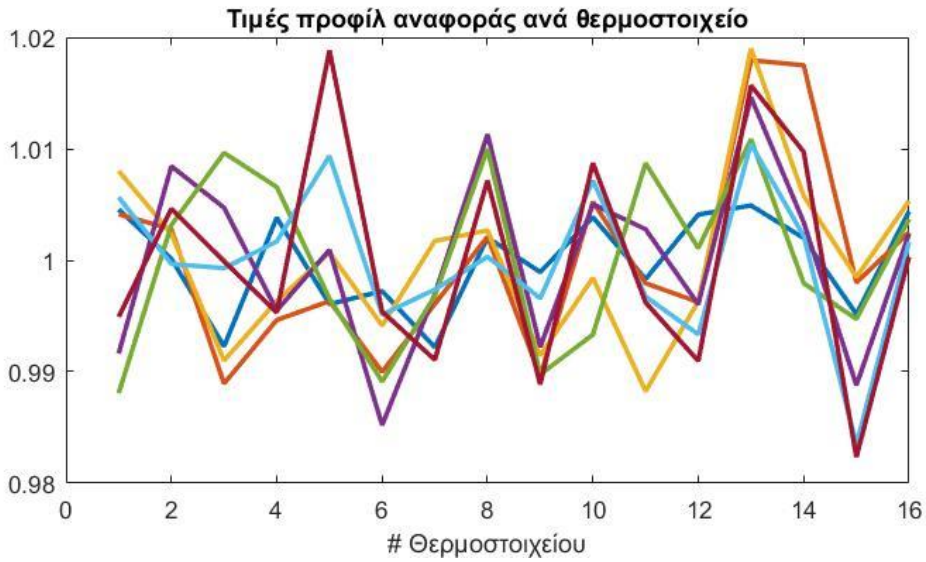
Σχήμα 3.34: Συντελεστής ομοιότητας CCD στο σύνολο των μετρήσεων

Όπως παρατηρήθηκε και νωρίτερα η αύξηση της αποδεκτής ομοιότητας αυξάνει και τον αριθμό σχηματιζόμενων clusters. Επίσης, στο Σχήμα 3.31 φαίνεται ότι κατά τον σχηματισμό παραπάνω clusters ορισμένα επικαλύπτονται και γίνεται δυσκολότερη η οριοθέτηση περιοχών για τον υπολογισμό των συντελεστών ομοιότητας. Αυτό προφανώς επηρεάζει τη μορφή του Σχήματος 3.34, που ενώ έχει σχετικά υψηλότερη μέση τιμή από το Σχήμα 3.30 των τριών clusters, έχει μεγαλύτερη διασπορά στις τιμές του. Η διασπορά αυτή οφείλεται στο γεγονός ότι η K-means (και η ιεραρχική μέθοδος όπως θα δούμε στη συνέχεια) δεν έχουν την ικανότητα διαχωρισμού των σημείων θορύβου από τα υπόλοιπα σημεία. Τέλος, οι περιοχές στο Σχήμα 3.33 φαίνεται να μοιάζουν με αυτές της DBSCAN καθώς σχηματίζεται ίδιος αριθμός clusters και με σχετικά όμοια όρια.

Ακολουθεί συνοπτικά η μελέτη της ιεραρχικής μεθόδου για τη διαδικασία αυτή καθώς η DBSCAN αποδεικνύεται ότι έχει τα ακριβέστερα αποτελέσματα. Παρακάτω φαίνονται τα αντίστοιχα διαγράμματα και πίνακες για την ιεραρχική μέθοδο:



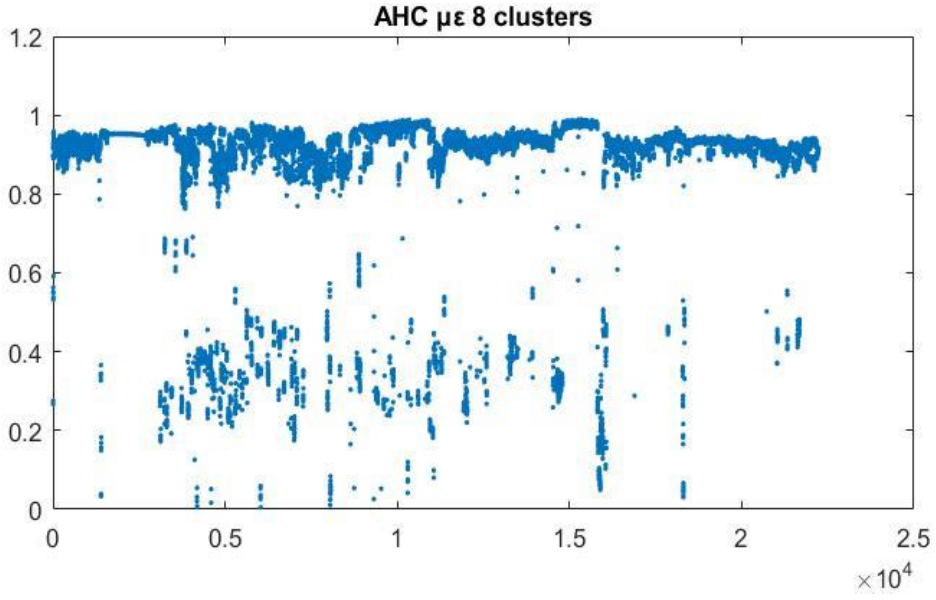
Σχήμα 3.35: Συσχέτιση cluster ταξινόμησης και διορθωμένης ισχύος για CCD=95%



Σχήμα 3.36: Τιμές θερμοκρασιακών προφίλ για ομαδοποίηση με CCD=95%, 7 καμπύλες

# Cluster	5	3	4	2	7	6	1
Min	35	55	60	65	75	90	100
Max	55	60	65	75	90	100	280

Σχήμα 3.37: Εύρος ισχύος ανάλογα με το cluster ταξινόμησης



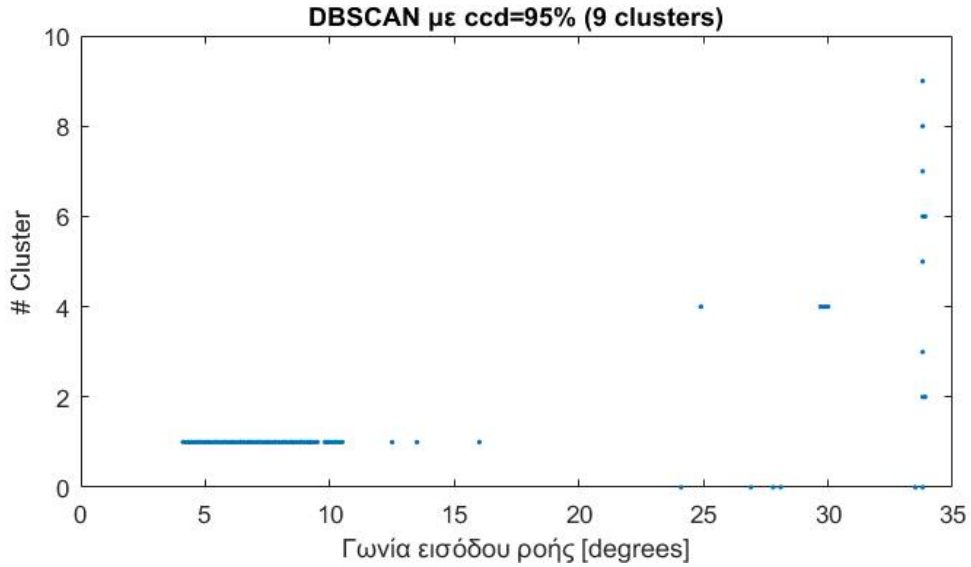
Σχήμα 3.38: Συντελεστής ομοιότητας CCD στο σύνολο των μετρήσεων

Όπως και με την K-means διατηρήθηκε το ίδιο ποσοστό αποδεκτής ομοιότητας ώστε τα αποτελέσματα να είναι συγκρίσιμα. Το αδιάστατο προφίλ στο Σχήμα 3.36 φαίνεται όμοιο με αυτή των άλλων 2 μεθόδων και ο αριθμός clusters που σχηματίστηκε

είναι σχεδόν ίδιος. Η κύρια διαφορά της ιεραρχικής μεθόδου είναι η κατανομή της ισχύος στα clusters καθώς προτιμάται η υψηλή διακριτοποίηση των χαμηλών τιμών ισχύος ενώ οι υψηλές ομαδοποιούνται σε ένα μοναδικό cluster. Τέλος, στο Σχήμα 3.38 φαίνεται η αδυναμία που μοιράζεται η ιεραρχική μέθοδος με την K-means. Η αδυναμία δηλαδή διαχωρισμού των σημείων θορύβου από το σύνολο των μετρήσεων. Έτσι, υπάρχουν πολλά σημεία με χαμηλό συντελεστή ομοιότητας που οφείλεται στην λανθασμένη κατανομή σημείων σε εύρος ισχύος που δεν ανήκουν.

Εξετάζοντας και τις τρεις μεθόδους στη διαδικασία παραγωγής περιοχών ίδιων μέσων θερμοκρασιακών προφίλ, παρατηρείται σαφές προβάδισμα της DBSCAN ως μέθοδος που βασίζεται σε διάφορα χαρακτηριστικά. Η DBSCAN μπορεί να διαχειρίζεται τα σημεία θορύβου ξεχωριστά από το σύνολο των υπόλοιπων μετρήσεων με αποτέλεσμα να γίνεται αποτελεσματικότερη ομαδοποίηση με ακριβέστερα όρια ισχύος κάθε cluster και καλύτερη προσέγγιση των μέσων τιμών ανά θερμοστοιχείο για τα μέσα προφίλ. Οι αδυναμίες τις μεθόδου περιορίζονται στο υπολογιστικό κόστος όπως προαναφέρθηκε και την απαιτούμενη μνήμη για τη διεργασία η οποία αναγκάζει το χρήστη να επεξεργαστεί ποσοστό των συνολικών μετρήσεων όταν η συχνότητα δειγματοληψίας είναι υψηλή. Επίσης, για ομαδοποίηση με συγκεκριμένο αριθμό clusters η DBSCAN δεν έχει τη δυνατότητα προεπιλογής, παρά μόνο εμμέσως από το μέγεθος εισόδου *epsilon*. Παρόλα αυτά, τα αποτελέσματα της μεθόδου όπως φαίνεται είναι τα θετικότερα, και σε σύγκριση με τις ήδη υπάρχουσες τεχνικές παραγωγής μέσων προφίλ, η DBSCAN υπερτερεί.

Ως επιπλέον ανάλυση τώρα που η βέλτιστη μέθοδος είναι γνωστή, μπορεί να χρησιμοποιηθεί η ομαδοποίηση δεδομένων βάσει της γωνίας εισόδου IGV του αέρα εισαγωγής αντί για τη διορθωμένη ισχύ ώστε να ελεγχθεί αν είναι αξιόπιστη μεταβλητή παρακολούθησης για το συγκεκριμένο φαινόμενο.



Σχήμα 3.39: Συσχέτιση cluster ταξινόμησης και γωνίας εισόδου ροής για CCD=95%

Είναι γνωστό ότι όσο μεγαλύτερη είναι η γωνία εισόδου σε μοίρες τόσο μικρότερη παροχή αναρροφά ο συμπιεστής του στροβιλοκινητήρα. Για μειωμένη παροχή λοιπόν,

παράγεται αντίστοιχα και μειωμένη ώση. Το φαινόμενο αυτό μπορεί να διασταυρωθεί και από το παραπάνω διάγραμμα καθώς συγκρίνοντας το με το Σχήμα 3.37 παρατηρείται παρόμοια κατανομή των clusters. Πιο συγκεκριμένα, τα clusters 1 και 4 περιέχουν την πλειοψηφία των υψηλών τιμών διορθωμένης ισχύος άρα σύμφωνα με την παραπάνω θεώρηση έχουν μικρότερη γωνία εισόδου από τα υπόλοιπα clusters πράγμα που διασταυρώνεται από το Σχήμα 3.39. Αντίστοιχα, τα υπόλοιπα clusters περιγράφουν τις περιοχές χαμηλότερων φορτίων, και για αυτό έχουν γωνία εισόδου κοντά στις 35 μοίρες. Όπως φαίνεται στο Σχήμα 3.39 η κατανομή των clusters έχει πολλές επικαλύψεις. Αυτό συμβαίνει επειδή οι μετρήσεις χαμηλών φορτίων χαρακτηρίζονται από παρόμοιες γωνίες εισόδου. Σε αντίθεση με τη διορθωμένη ισχύ του Σχήματος 3.35, το Σχήμα 3.39 δε μπορεί να προσφέρει χρήσιμες πληροφορίες για την αλλαγή των μέσων προφίλ ή τον τρόπο κατάταξης των clusters άρα η γωνία εισόδου δεν είναι ικανό κριτήριο διακριτοποίησης.

3.4.2 Αεριοστρόβιλος GTB1

Ακολουθεί η εύρεση προφίλ αναφοράς για τον κινητήρα GTB1. Ο κινητήρας αυτός είναι πανομοιότυπος με τον επόμενο κινητήρα, τον GTB2, μελετάται η λειτουργία του για το διάστημα Οκτώβρη του 2018 με Απρίλιο του 2019 και η ενδεικνυόμενη τιμή διορθωμένης ισχύος του κινητήρα είναι 135 MW. Το μητρώο που εισάγεται από το Excel αποτελείται από 9950 μετρήσεις χρόνου (ανά 5 λεπτά λειτουργίας του κινητήρα) για συνεχή λειτουργία 24 ώρες καθημερινά. Οι μεταβλητές που παρακολουθούνται σε αυτές τις χρονικές στιγμές είναι οι:

- Μέγιστη ισχύς κινητήρα
- Γωνία εισαγωγής αέρα από το περιβάλλον
- Θερμοκρασία αέρα στην είσοδο του συμπιεστή
- Ατμοσφαιρική πίεση
- Διαφορά πίεσης μεταξύ αέρα εισόδου στον συμπιεστή και ατμοσφαιρικής πίεσης
- 24 διαφορετικές θερμοκρασίες για το προφίλ της θερμοκρασίας εξόδου των καυσαερίων από το στρόβιλο
- Μέγιστη διορθωμένη ισχύς κινητήρα

Το μητρώο λοιπόν αποτελείται από 9950 γραμμές και 30 στήλες και εισάγεται μέσω των κατάλληλων εντολών στο Προγραμματιστικό Περιβάλλον της MATLAB. Εκεί, γίνεται επεξεργασία των μετρήσεων ώστε να έρθουν σε αξιοποιήσιμη μορφή.

Μαζί με τις μετρήσεις για τον κινητήρα, χρησιμοποιήθηκε και πίνακας ομαδοποίησης των δεδομένων σε 9 προφίλ ανάλογα με το εύρος Διορθωμένης Ισχύος, ο οποίος παρέχει χρήσιμες πληροφορίες για τη ρύθμιση της μηχανής σε επόμενη χρήση ανάλογα με την ισχύ που διαλέγεται να λειτουργήσει ο κινητήρας. Τα 9 αυτά προφίλ καλύπτουν όλο το εύρος λειτουργίας της Ισχύος (75 με 135 MW) με αλλαγή προφίλ ανά 5 MW. Η αλλαγή ανά 5 MW συμβαίνει λόγω ομοιότητας των συνιστωσών σε ένα σχετικά μικρό εύρος. Πρόκειται για μία ασφαλή λύση στο πρόβλημα της λειτουργίας σε μεγάλο εύρος συντελεστών που επιτρέπει μεν τη δυνατότητα ομαδοποίησης των συντελεστών

αυτών σε μεγαλύτερα σύνολα, δεν εγγυάται δε τη βέλτιστη λειτουργία βάσει των ρυθμίσεων που επιβάλλονται πριν από την αρχή λειτουργίας. Το έργο των τεχνικών διάγνωσης που αναπτύσσονται σε αυτήν τη διπλωματική εργασία είναι να δώσουν λύση στο παραπάνω πρόβλημα, κάνοντας ευρύτερη και παράλληλα αποτελεσματικότερη ομαδοποίηση των προφίλ λειτουργίας.

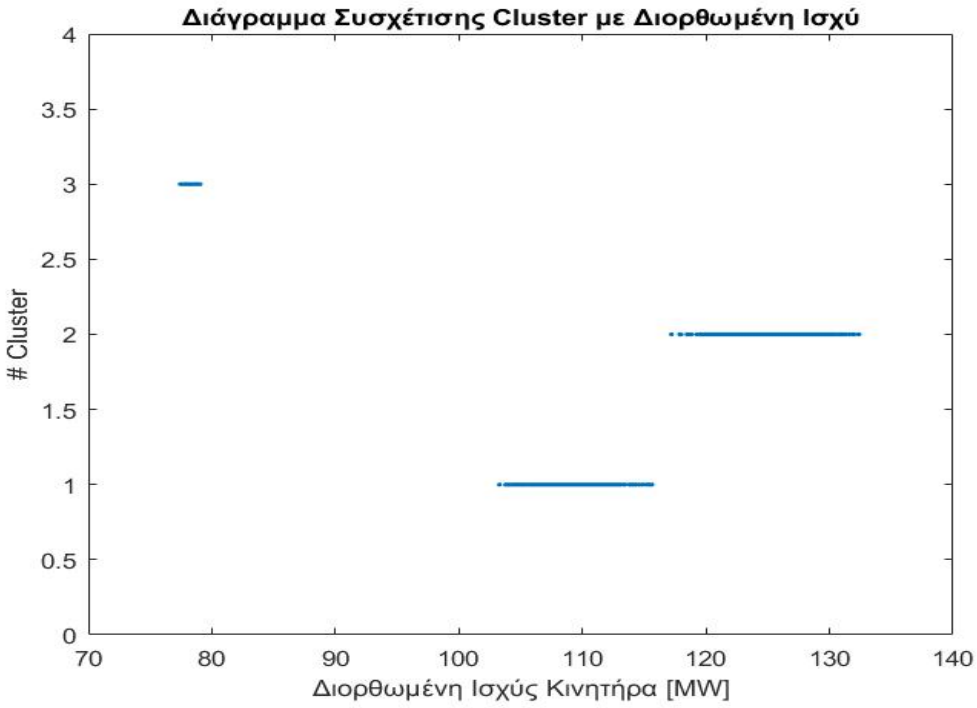
DBSCAN

Η DBSCAN ομαδοποιεί τα δεδομένα βάσει των χαρακτηριστικών μεγεθών *epsilon* και *MinPts*. Γνωρίζοντας ότι η DBSCAN δε δίνει τη δυνατότητα στο χρήστη προεπιλογής συνολικού αριθμού Clusters που θέλει να εξάγει ως αποτέλεσμα, είναι αναγκαίο ο αριθμός των clusters να οριστεί μέσω των χαρακτηριστικών μεγεθών της μεθόδου. Βάσει αυτών, μπορεί να παραχθεί ο παρακάτω πίνακας που περιέχει τις τιμές του χαρακτηριστικού συντελεστή *epsilon* συναρτήσει του μέγιστου αριθμού clusters:

# Clusters	9	8	7	6	5	4	3	2	1
Epsilon	0.0029	0.00295	0.0035	0.005	0.004	0.008	0.011	0.012	>0.012

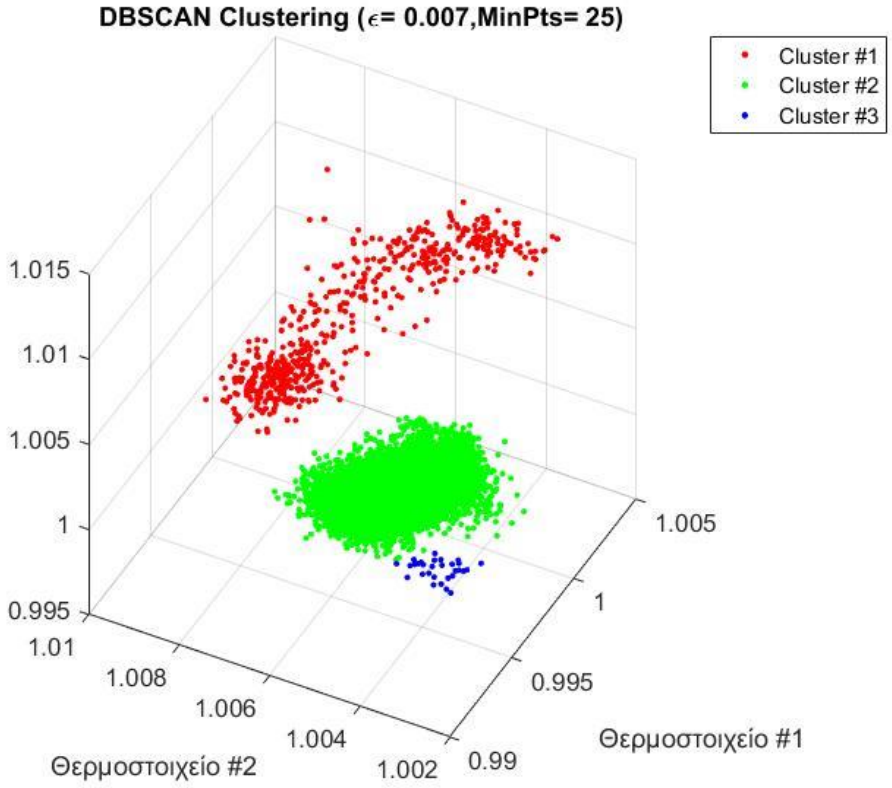
Σχήμα 3.40: Ελάχιστα epsilon ανάλογα με το πλήθος Clusters για την DBSCAN

Στο παραπάνω Σχήμα παρατηρείται ότι όσο μειώνεται ο αριθμός των Clusters τόσο αυξάνεται η τιμή της μεταβλητής. Το φαινόμενο αυτό είναι λογικό καθώς το *epsilon* εκφράζει την μέγιστη απόσταση που μπορούν να έχουν δύο σημεία για να θεωρούνται γείτονες. Έτσι λοιπόν, όταν διαλέγεται μεγαλύτερη ομαδοποίηση (μικρότερος αριθμός Clusters δηλαδή) υπάρχει η φυσική ανάγκη συγχώνευσης clusters που δεν αποτελούνται από πανομοιότυπα στοιχεία, συγχώνευσης δηλαδή λόγω *epsilon* αρκετά μεγάλου ώστε να θεωρηθούν σημεία, που υπό άλλες δε θα θεωρούνταν κοντινά, γείτονες. Ο ίδιος πίνακας έχει παραχθεί και για DBSCAN με CCD μέθοδο αλλά δε θα αναλυθεί σε αυτό το σημείο. Το διάγραμμα για τη μέθοδο DBSCAN με ευκλείδεια απόσταση στα δεδομένα της μηχανής GTB1:



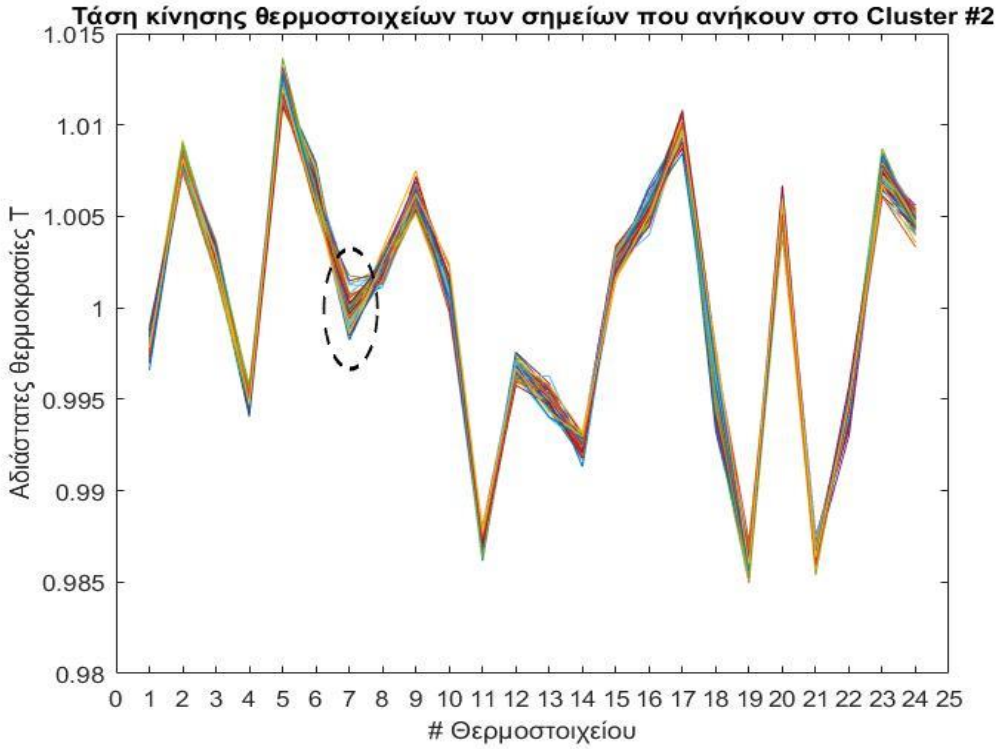
Σχήμα 3.41: Διάγραμμα συσχέτισης Cluster με διορθωμένη ισχύ για τη DBSCAN

Πρώτα από όλα παρατηρείται ότι τα εύρη των clusters μεταξύ τους δεν επικαλύπτονται. Φαινόμενο λογικό καθώς αν για παράδειγμα δύο ή παραπάνω σημεία είχαν από διαφορετικό cluster είχαν Ισχύ ίση με 130 MW τότε κάποια μεταβολή θα είχε συμβεί στη μηχανή κατά το εξεταζόμενο διάστημα. Επίσης λογικό είναι το γεγονός ότι όλες οι μετρήσεις κοντά στα 80 MW ανήκουν στο ίδιο cluster. Εφόσον δεν υπάρχουν δεδομένα για λειτουργία σε 85 με 100 MW, θα θεωρούταν απίθανο οι ρυθμίσεις της μηχανής για 80 και 120 MW για παράδειγμα να είναι ίδιες (δηλαδή να ανήκουν στο ίδιο cluster). Τέλος, με την ομαδοποίηση αυτή υπάρχει πλέον η δυνατότητα σύγκρισης των θερμοκρασιών λειτουργίας του κινητήρα με τις υπογραφές για τα διάφορα εύρη λειτουργίας με σκοπό τη διεξαγωγή διαγνωστικών αποτελεσμάτων. Παρακάτω μπορούμε να δούμε και σε τρισδιάστατο χώρο τα επεξεργασμένα δεδομένα. Το διάγραμμα που ακολουθεί έχει όλες τις μετρήσεις (9950) αλλά όχι όλων των θερμοστοιχείων καθώς δε μπορεί να υπάρξει μία τέτοια γραφική απεικόνιση.



Σχήμα 3.42: Τρισδιάστατη απεικόνιση κατανομής σημείων ανάλογα με το Cluster που ανήκουν μέσω DBSCAN

Μέσω της τρισδιάστατης απεικόνισης γίνεται ευκολότερος ο διαχωρισμός σε ομάδες, παρόλα αυτά τα δεδομένα δεν έρχονται πάντα σε τριάδες μετρήσεων. Το πλεονέκτημα των μεθόδων αυτών είναι ότι υπολογίζουν αποστάσεις ανεξαρτήτως διαστάσεων των παρατηρήσεων. Σε αυτό το διάγραμμα επίσης παρατηρείται και η τιμή του άλλου χαρακτηριστικού μεγέθους της μεθόδου DBSCAN, ο αριθμός *MinPts*. Κατά κανόνα, όπως αναφέρθηκε και στη θεωρία της μεθόδου, το μέγεθος αυτό υπολογίζεται ανάλογα με τις διαστάσεις του δείγματος. Στην προκειμένη περίπτωση, υπάρχουν 24 θερμοστοιχεία, άρα και 24 διαστάσεις οπότε το χαρακτηριστικό μέγεθος παίρνει την τιμή 25. Ένας επιπλέον έλεγχος που μπορεί να εφαρμοσθεί και θα φανεί αργότερα χρησιμότερος είναι η συσχέτιση των σημείων του ίδιου cluster μεταξύ τους. Ο έλεγχος αυτός γίνεται μέσω διαγράμματος των χρονικών στιγμών που νωρίτερα ομαδοποιήθηκαν σε ένα cluster συναρτήσει των τιμών που λαμβάνουν για κάθε ένα από τα θερμοστοιχεία. Από το διάγραμμα που παράγεται διακρίνεται αν τα στοιχεία του ίδιου cluster έχουν ίδια τάση κίνησης, με αλλά λόγια, είναι όμοια μεταξύ τους. Στην προκειμένη μελέτη φάνηκε νωρίτερα ότι δεν επικαλύπτονται τα εύρη ισχύος των clusters μεταξύ τους, οπότε το διάγραμμα των θερμοστοιχείων θα έχει ίδια τάση κίνησης. Ένα παράδειγμα τέτοιου διαγράμματος φαίνεται παρακάτω:



Σχήμα 3.43: Τάση κίνησης θερμοστοιχείων του Cluster #2 για DBSCAN

Το παραπάνω διάγραμμα παράχθηκε για $epsilon = 0.007$, $MinPts = 25$ και παρουσιάζει τις τιμές όλων των θερμοστοιχείων (24 στον αριθμό) για τα σημεία που κατατάχθηκαν στο Cluster #2. Η απόκλιση ανά θερμοστοιχείο υπολογίζεται εύκολα αν θεωρηθεί ότι το διάγραμμα αποτελείται από μία γραμμή και όχι πολλές κοντινές συναρτήσει των θερμοστοιχείων. Η μέγιστη απόκλιση λοιπόν προκύπτει ίση με το πάχος της γραμμής που θεωρήθηκε. Η μέγιστη απόκλιση βρίσκεται στο 7ο θερμοστοιχείο και έχει σημειωθεί με έναν μαύρο κύκλο. Υπολογίστηκε ότι η απόκλιση αυτή έχει πλάτος $b = 0.0031\%$ δηλαδή 2.2 βαθμούς Kelvin. Η απόκλιση αυτή δε θεωρείται σημαντική ακόμα και αν είναι η μέγιστη, άρα είναι διακριτό ότι οι τιμές ανά θερμοστοιχείο για όλα τα σημεία του Cluster δεν αποκλίνουν σημαντικά φαινόμενο λογικό σύμφωνα με όσα περιεγράφηκαν πριν το διάγραμμα.

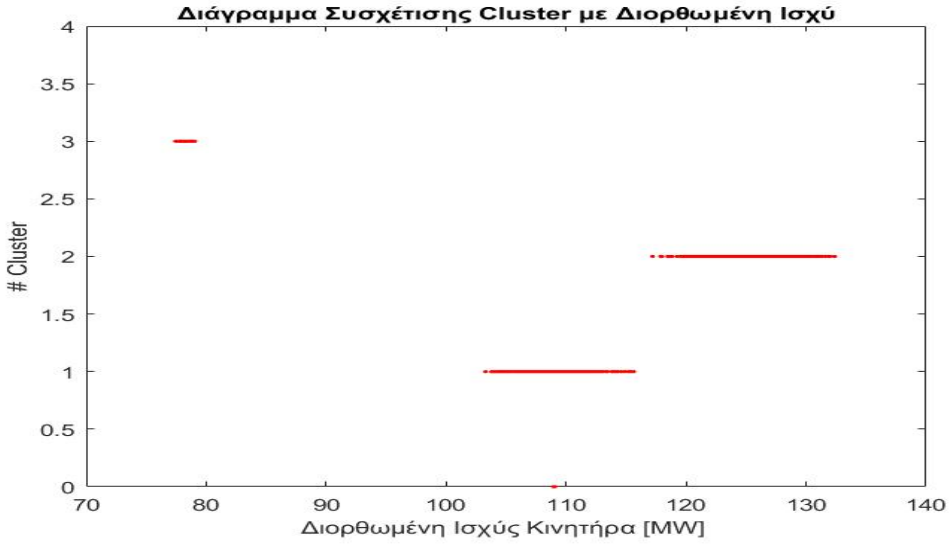
Στη συνέχεια, ακολουθείται παρόμοια διαδικασία με αυτήν της μεθόδου DBSCAN όπου μειώνεται ο αριθμός Clusters που επιδιώκεται να σχηματιστούν ανάλογα με τη μέγιστη ομοιότητα που παρατηρείται στις μέσες τιμές ανά θερμοστοιχείο ανά Cluster. Η διαδικασία ξεκινά με $K = 9$ ώστε να ελεγχθεί εάν η ομαδοποίηση που έγινε από το Εργαστήριο είναι η βέλτιστη. Αφού προκύψει ότι η μέγιστη ομοιότητα είναι μεγαλύτερη του 0.95 όπως αναφέρθηκε παραπάνω, η διαδικασία συνεχίζεται με $K = 8$, $K = 7$ κοκ. Εισάγοντας στον αλγόριθμο τις τιμές αυτές γίνεται ομαδοποίηση των επεξεργασμένων δεδομένων. Για κάθε γραμμή-χρονική παρατήρηση του δείγματος αντιστοιχεί ένας αριθμός από το 1 μέχρι το μέγιστο αριθμό clusters στον οποίο αντιστοιχεί το K που χρησιμοποιήθηκε. Ο αριθμός αυτός αντιστοιχεί κάθε μέτρηση με το cluster στο οποίο ανήκει. Επιπλέον, ο αλγόριθμος παράγει ένα μητρώο t * k που περιέχει τη μέση τιμή κάθε

θερμοστοιχείου (από τα k) σε κάθε ένα από τα σχηματισμένα Clusters. Ο πίνακας αυτός έχει k στήλες επειδή είναι ο μέγιστος αριθμός clusters που χρησιμοποιείται. Για *K* μεγαλύτερο από αυτό που αντιστοιχεί στα k clusters ο αλγόριθμος αφήνει τις απαραίτητες στήλες κενές. Ο επόμενος Πίνακας, όπως και με τη DBSCAN, είναι διαστάσεων k * k και περιέχει του Συντελεστές Αλληλοσυσχέτισης μεταξύ των στηλών του παραπάνω Πίνακα. Πρακτικά υπολογίζεται η ομοιότητα των αντίστοιχων τιμών κάθε στήλης με όλες τις υπόλοιπες. Στον παρακάτω πίνακα φαίνεται η επιτυχία ή η αποτυχία να ξεπεραστεί το όριο της μέγιστης ομοιότητας, ανάλογα με το K που αντιστοιχεί:

# Clusters	9	8	7	6	5	4	3	2	1
	V	V	V	V	V	V	X	X	X

Σχήμα 3.44: Πίνακας μέγιστης ομαδοποίησης με τη μέθοδο K-means

Ακολουθεί το τελικό διάγραμμα για τη μέθοδο K-means με ευκλείδεια απόσταση στα δεδομένα της μηχανής GTB1:

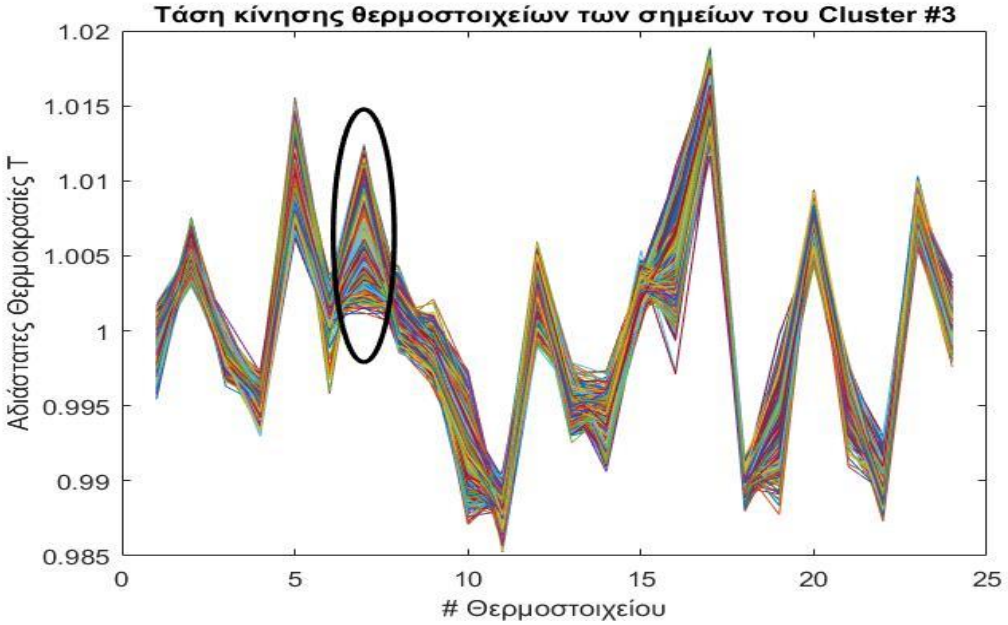


Σχήμα 3.45: Διάγραμμα συσχέτισης Cluster με διορθωμένη ισχύ για τη K-means

Πρώτα από όλα παρατηρείται ότι τα εύρη των clusters μεταξύ τους δεν επικαλύπτονται όπως και στη DBSCAN. Φαινόμενο λογικό καθώς αν για παράδειγμα δύο ή παραπάνω σημεία είχαν από διαφορετικό cluster είχαν Ισχύ ίση με 130 MW τότε οι ρυθμίσεις της μηχανής θα γίνονταν περίπλοκες. Επίσης λογικό είναι το γεγονός ότι όλες οι μετρήσεις κοντά στα 80 MW ανήκουν στο ίδιο cluster. Εφόσον δεν υπάρχουν δεδομένα για λειτουργία σε 85 με 100 MW, θα θεωρούταν απίθανο οι ρυθμίσεις της μηχανής για 80 και 120 MW για παράδειγμα να είναι ίδιες (δηλαδή να ανήκουν στο ίδιο cluster). Τέλος, με την ομαδοποίηση αυτή υπάρχει πλέον η δυνατότητα σύγκρισης των θερμοκρασιών λειτουργίας του κινητήρα με τις υπογραφές για τα διάφορα εύρη λειτουργίας με σκοπό τη διεξαγωγή διαγνωστικών αποτελεσμάτων. Τα αποτελέσματα αυτά είναι σχεδόν πανομοιότυπα για τις 2 μεθόδους με μόνη διαφορά το χρόνο επεξεργασίας και εξαγωγής αποτελέσματος των αλγορίθμων. Η K-means λειτούργησε αισθητά πιο γρήγορα καθώς

παρείχε αποτελέσματα σε 18-19 δευτερόλεπτα σε αντίθεση με τα 40-45 δευτερόλεπτα της DBSCAN. Με παρόμοια λογική λειτούργησε και η μέθοδος K-means με τη συνάρτηση του Συντελεστή Αλληλοσυσχέτισης. Η διαφορά της τελευταίας με αυτήν που εφάρμοσε ευκλείδεια απόσταση ήταν απλά μία υστέρηση στα αποτελέσματα της τάξης των 10 με 12 δευτερολέπτων η οποία πιθανότατα οφείλεται στην παραγωγή μητρώου αποστάσεων με συνάρτηση διαφορετική από τις ήδη υπάρχουσες στο λογισμικό της MATLAB. Παρόλα αυτά, η υστέρηση αυτή είναι η βέλτιστη δυνατή καθώς για την παραγωγή της συνάρτησης του Συντελεστή Αλληλοσυσχέτισης εφαρμόσθηκαν και τεχνικές ελαχιστοποίησης των περιττών πράξεων καθώς ο πίνακας αποστάσεων που παράγεται κάθε φορά είναι διαγώνιος συμμετρικός.

Ο επιπλέον έλεγχος που εφαρμόσθηκε και στη μέθοδο DBSCAN, μπορεί να εφαρμοσθεί φυσικά και στην K-means. Ο έλεγχος αυτός γίνεται μέσω διαγράμματος των χρονικών στιγμών που νωρίτερα ομαδοποιήθηκαν σε ένα cluster συναρτήσει των τιμών που λαμβάνουν για κάθε ένα από τα θερμοστοιχεία. Από το διάγραμμα που παράγεται διακρίνεται αν τα στοιχεία του ίδιου cluster έχουν ίδια τάση κίνησης, με αλλά λόγια, είναι όμοια μεταξύ τους. Στην προκειμένη μελέτη φάνηκε νωρίτερα ότι δεν επικαλύπτονται τα εύρη ισχύος των clusters μεταξύ τους, οπότε το διάγραμμα των θερμοστοιχείων θα έχει ίδια τάση κίνησης. Το διάγραμμα αυτό φαίνεται παρακάτω:



Σχήμα 3.46: Τάση κίνησης θερμοστοιχείων του Cluster #3 για K-means

Το παραπάνω διάγραμμα παράχθηκε για $K = 3$ και παρουσιάζει τις τιμές όλων των θερμοστοιχείων (24 στον αριθμό) για τα σημεία που κατατάχθηκαν στο Cluster #3. Η απόκλιση ανά θερμοστοιχείο υπολογίζεται εύκολα αν θεωρηθεί ότι το διάγραμμα αποτελείται από μία γραμμή και όχι πολλές κοντινές συναρτήσει των θερμοστοιχείων. Η μέγιστη απόκλιση λοιπόν προκύπτει ίση με το πάχος της γραμμής που θεωρήθηκε. Η μέγιστη απόκλιση βρίσκεται και πάλι στο 7ο θερμοστοιχείο και έχει σημειωθεί με έναν μαύρο κύκλο. Υπολογίστηκε ότι η απόκλιση αυτή έχει πλάτος $b = 0.0127\%$ δηλαδή 7.9

βαθμούς Kelvin. Η απόκλιση αυτή θεωρείται σχετικά σημαντική σε σχέση με αυτήν που υπολογίστηκε με τη DBSCAN, διότι είναι περίπου 4 φορές μεγαλύτερη. Από ό,τι φαίνεται, το υπολογιστικό κόστος και η ακρίβεια αποτελεσμάτων είναι οι κύριοι πυλώνες διαφοροποίησης των δύο μεθόδων, καθώς η μία μέθοδος αστοχεί εκεί που επιτυγχάνει η άλλη.

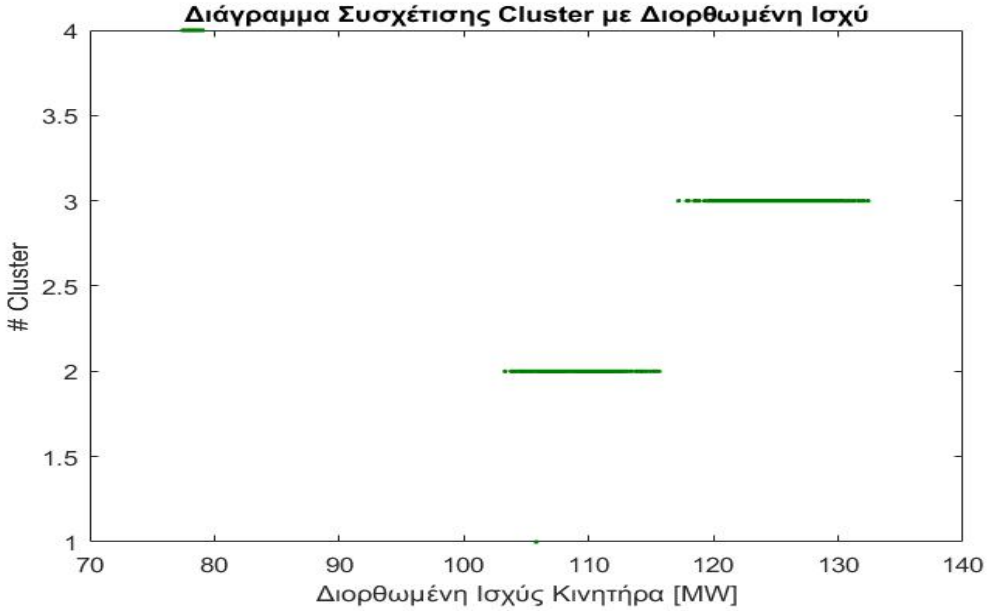
AHC

Όπως αναφέρθηκε και παραπάνω η μέθοδος AHC ομαδοποιεί παρόμοια με την K-means. Υπολογίζει τις αποστάσεις των επεξεργασμένων δεδομένων στο t-διάστατο χώρο και συνδέει τα δεδομένα που βρίσκονται στη μικρότερη απόσταση. Η χαρακτηριστική μεταβλητή που επηρεάζει το τελικό αποτέλεσμα της μεθόδου είναι η τελική τιμή Clusters που θα δώσει ο χρήστης ώστε να διακοπούν οι συγχωνεύσεις. Ρίχνοντας λοιπόν την τιμή της μεταβλητής από το k όπως και στις άλλες μεθόδους προς τα κάτω, υπολογίζονται κάθε φορά τα δύο μητρώα t * k και k * k που αντιπροσωπεύουν την ομοιότητα των μετρήσεων. Με κριτήριο ομοιότητας το 0.95, η μέθοδος καταλήγει ότι 4 Clusters είναι η βέλτιστη ομαδοποίηση. Μάλιστα, η μέγιστη ομοιότητα στα 4 Clusters είναι ίση με 90,9% φαινόμενο που καθιστά την τελευταία συγχώνευση αμφισβητήσιμη. Στον παρακάτω πίνακα φαίνεται η επιτυχία ή η αποτυχία να ξεπεραστεί το όριο της μέγιστης ομοιότητας, ανάλογα με το K που αντιστοιχεί:

# Clusters	9	8	7	6	5	4	3	2	1
	V	V	V	V	V	X	X	X	X

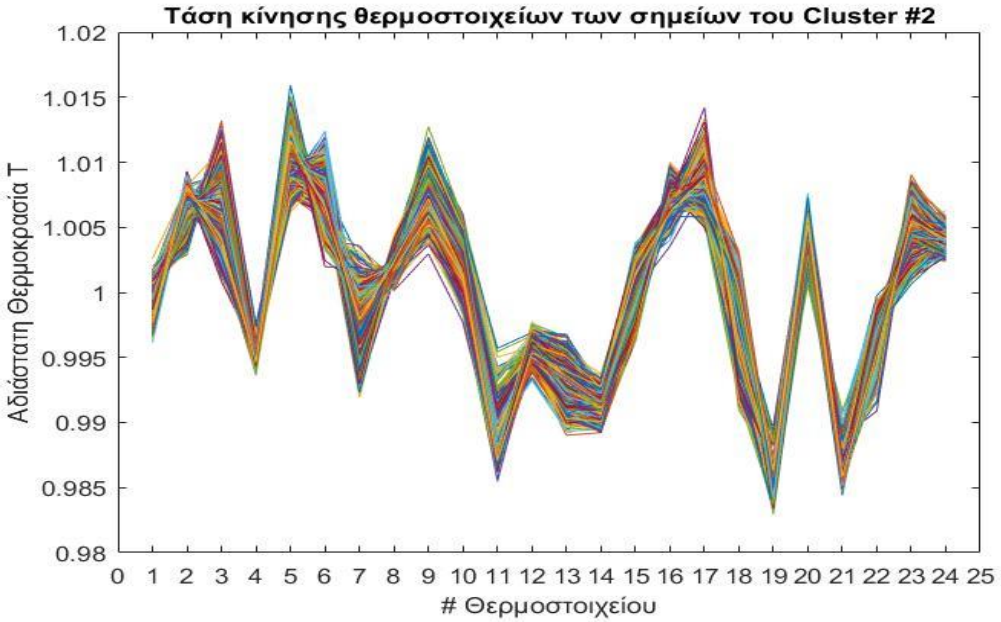
Σχήμα 3.47: Πίνακας μέγιστης ομαδοποίησης με τη μέθοδο AHC

Ακολουθεί το τελικό διάγραμμα συσχέτισης για τη μέθοδο AHC με ευκλείδεια απόσταση στα δεδομένα της μηχανής GTB1:



Σχήμα 3.48: Διάγραμμα συσχέτισης Cluster με διορθωμένη ισχύ για την AHC

Το διάγραμμα φαίνεται πανομοιότυπο με αυτό που παράχθηκε μέσω της K-means. Τα εύρη των Clusters δεν επικαλύπτονται και αυτό είναι αναμενόμενο να φανεί και στο επόμενο διάγραμμα. Στη συνέχεια ακολουθεί το διάγραμμα τάσης κίνησης των θερμοστοιχείων των σημείων που κατατάχθηκαν στο Cluster #2 μέσω της μεθόδου AHC:



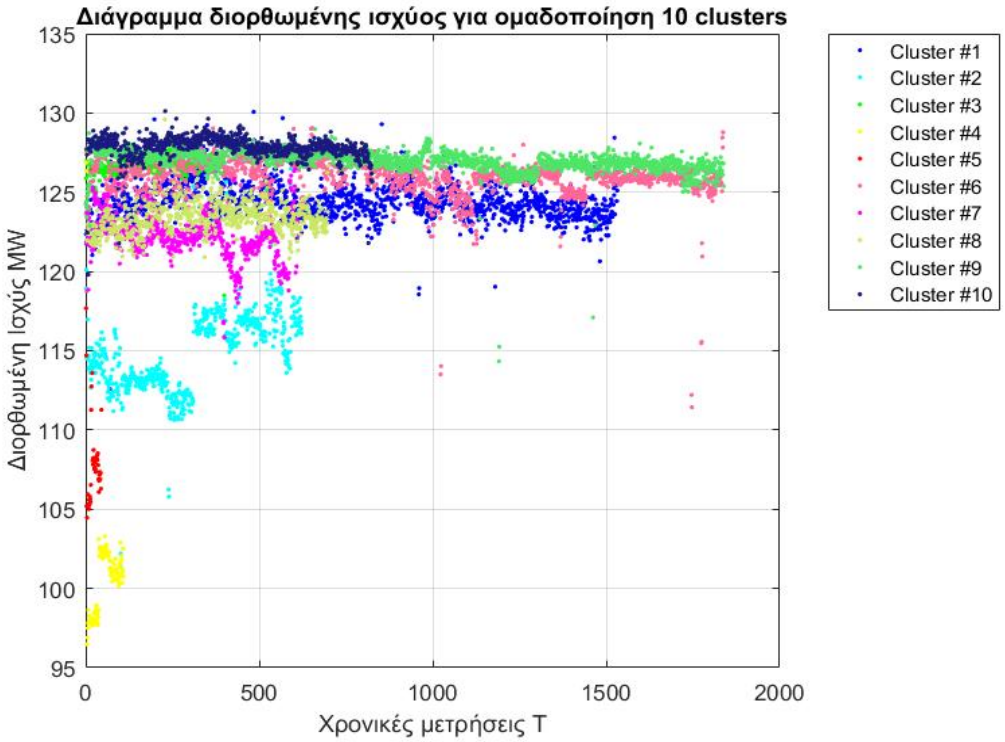
Σχήμα 3.49: Τάση κίνησης θερμοστοιχείων του Cluster #3 για AHC

Όπως και με τις προηγούμενες μεθόδους, η μέγιστη απόκλιση υπολογίζεται εκεί που η θεωρητική γραμμή έχει μέγιστο πάχος. Η μέγιστη αυτή απόκλιση φαίνεται να βρίσκεται στο 7ο θερμοστοιχείο και πάλι. Το φαινόμενο αυτό, θα μπορούσε να διερευνηθεί παραπάνω καθώς αυξάνεται το ενδεχόμενο βλάβης στο συγκεκριμένο θερμοστοιχείο, αλλά δεν αποτελεί κομμάτι της συγκεκριμένης ανάλυσης. Η μέγιστη απόκλιση είναι ίση με $b = 0.09$ δηλαδή απόκλιση της τάξης των 7.2 βαθμών Kelvin. Στο υπόλοιπο διάγραμμα η τάση κίνησης φαίνεται να είναι ομοιόμορφη ανά θερμοστοιχείο, άρα επιβεβαιώνεται η πιστότητα του αποτελέσματος της μεθόδου.

Ανάλυση ευαισθησίας

Για να γίνει αποτελεσματική ομαδοποίηση των θερμοκρασιακών προφίλ, θα πρέπει να οριστεί ένας αρχικός αριθμός επιθυμητών clusters. Η πρώτη ομαδοποίηση των μεθόδων ακολουθεί και ο αλγόριθμος αποφασίζει αν τα clusters που σχηματίστηκαν είναι ανεξάρτητα μεταξύ τους ή όχι. Αν ο αρχικός αριθμός clusters επιλεχθεί πολύ μικρότερος από τον τελικά επιθυμητό αριθμό ανεξάρτητων clusters, τότε ο αλγόριθμος αδυνατεί να αυξήσει των αριθμό των clusters που θα σχηματιστούν. Αυτό συμβαίνει επειδή από τα μητρώα ομοιότητας που αναφέρθηκαν παραπάνω εξαρτάται μόνο το αν θα μειωθούν τα clusters και όχι το αν θα αυξηθούν. Έτσι λοιπόν συμπεραίνεται ότι ο αρχικός αριθμός clusters είναι κρίσιμης σημασίας.

Για την ανάλυση ευαισθησίας που ακολουθεί εξετάστηκαν διάφοροι αριθμοί αρχικών clusters, μεγάλοι και μικροί. Ξεκινώντας, με μεγάλους αριθμούς clusters παρατηρήθηκε ότι ο τελικός αριθμός ανεξάρτητων clusters παραμένει ίδιος. Για τη διαδικασία αυτή, οι αλγόριθμοι κλήθηκαν να παραγάγουν αποτελέσματα για αρχικό αριθμό clusters ίσο με: 10, 12, 14, 16, 18 και 20. Τα αρχικά clusters που σχηματίζονταν, είχαν μορφή σαν αυτή του σχήματος 3.50:



Σχήμα 3.50: Διάγραμμα διορθωμένης ισχύος για τρέξιμο 10 επιθυμητών clusters

Όπως φαίνεται στην παραπάνω εικόνα, τα clusters δεν κατανέμονται σε συγκεκριμένα εύρη φαινόμενο που προδίδει την εξάρτηση clusters μεταξύ τους. Με τη διαδικασία που περιεγράφηκε παραπάνω τα clusters αυτά μειώνονται μέχρι τελικά να σχηματιστούν 4 ανεξάρτητα. Το ίδιο ισχύει και για επιλογή μεγαλύτερου αρχικού αριθμού clusters, απλά τα 10 clusters που σχηματίστηκαν τώρα θα χωριστούν σε ακόμα μικρότερες γειτονιές. Έτσι από την ανάλυση αυτή συμπεραίνεται ότι ως αρχικό αριθμό επιθυμητών clusters διαλέγεται ένας αριθμός σχετικά μεγάλος περίπου 15 με 20 clusters και οι αλγόριθμοι θα καταλήξουν οριστικά στα ανεξάρτητα clusters ανεξαρτήτως του αρχικού αριθμού.

Ακολούθως, προκύπτουν τα τελικά αποτελέσματα. Στον παρακάτω πίνακα φαίνονται οι μέθοδοι με τις συναρτήσεις απόστασης που χρησιμοποιήθηκαν καθώς και το πλήθος γειτονιών που αναγνωρίστηκαν με τα εύρη λειτουργίας τους και τον βαθμό ομοιότητας που τα χαρακτηρίζει:

Μέθοδος	Μετρική	Αριθμός Clusters	Μέγιστη Ομοιότητα	Υπολογιστικός Χρόνος (sec)
DBSCAN	Ευκλείδεια	3	63%	45
	CCD	3	63%	800
K-means	Ευκλείδεια	3	60%	18
	CCD	3	63%	30
AHC	Ευκλείδεια	4	90%	12
	CCD	4	91%	14

Σχήμα 3.51: Αποτελέσματα αριθμού Clusters, μέγιστης ομοιότητας και χρόνου για τις μεθόδους που αναπτύχθηκαν

Παρατηρώντας την τελευταία στήλη του πίνακα είναι πλέον ξεκάθαρη η απόκλιση της DBSCAN με CCD και ο λόγος που δε χρησιμοποιήθηκε. Παρά τη μεγάλη χρονική καθυστέρηση του υπολογισμού CCD στην DBSCAN, τα αποτελέσματα ήταν όμοια με αυτά της Ευκλείδειας. Αξιοσημείωτο επίσης είναι ότι η K-means κατέληξε σε χαμηλότερο αριθμό Clusters από ότι η DBSCAN. Η τελική ομαδοποίηση στην οποία προχώρησε η K-means και με τις δύο συναρτήσεις απόστασης ήταν η συγχώνευση των ομάδων 105-115 MW με αυτήν των 118-132 MW. Σε περίπτωση διαφωνίας των μεθόδων, η λύση δίνεται από το διάγραμμα τάσης κίνησης των θερμοστοιχείων. Όπως φάνηκε και παραπάνω, το 3ο Cluster της K-means με Ευκλείδεια απόσταση είχε αρκετά μεγάλη απόκλιση στο θερμοστοιχείο #7. Η AHC από την άλλη πλευρά φαίνεται να διατηρεί μία πιο συντηρητική γραμμή στην ομαδοποίηση καθώς κατέληξε σε 4 Clusters. Παρόλα αυτά, φαίνεται από τη στήλη μέγιστης ομοιότητας ότι και αυτή η μέθοδος ήταν αρκετά κοντά σε συγχώνευση με τελικό αριθμό Clusters ίσο με 3. Είναι βέβαιο δηλαδή ότι η επιπλέον ομαδοποίηση της K-means ήταν περιττή ενώ της AHC απαραίτητη και ότι η τελική ομαδοποίηση της μηχανής GTB1 θα έπρεπε να είναι:

- 1ο Cluster: 70 με 90 MW
- 2ο Cluster: 100 με 115 MW
- 3ο Cluster: 115 με 140 MW

Γνωρίζοντας ότι με τα δεδομένα αυτά η λειτουργία του αεριοστρόβιλου μπορεί να ομαδοποιηθεί στα τρία αυτά Clusters, είναι πλέον δυνατή η παραγωγή Μέσου Προφίλ Αναφοράς (υπογραφή) των θερμοστοιχείων για κάθε ένα από τα Clusters. Είναι δυνατή δηλαδή η λειτουργία του αεριοστρόβιλου σε διάφορα εύρη ισχύος με τις κατάλληλες υπογραφές να συνοδεύουν το μοντέλο και να συγκρίνονται με σκοπό τη διεξαγωγή διαγνωστικής πληροφορίας για την υγεία του κινητήρα.

3.4.3 Αεριοστρόβιλος GTB2

Για τον κινητήρα GTB2 τα δεδομένα που δόθηκαν καλύπτουν την περίοδο Απρίλη με Ιούλη του 2019 για λειτουργία σε 135 MW. Μαζί με τις μετρήσεις για τον κινητήρα, χρησιμοποιήθηκε και πίνακας ομαδοποίησης των δεδομένων σε 8 προφίλ ανάλογα με το εύρος Διορθωμένης Ισχύος, ο οποίος παρέχει χρήσιμες πληροφορίες για τη ρύθμιση της

μηχανής σε επόμενη χρήση ανάλογα με την ισχύ που διαλέγεται να λειτουργήσει ο κινητήρας. Τα 8 αυτά προφίλ καλύπτουν όλο το εύρος λειτουργίας της Ισχύος (95 με 135 MW) με αλλαγή προφίλ ανά 5 MW. Η αρχή λειτουργίας των μεθόδων δεν αλλάζει προφανώς από μηχανή σε μηχανή. Το μόνο που αλλάζει στη διαδικασία είναι ο πίνακας με το χαρακτηριστικό μέγεθος *epsilon* για την DBSCAN. Ο ανανεωμένος πίνακας βάσει των νέων επεξεργασμένων μετρήσεων φαίνεται παρακάτω:

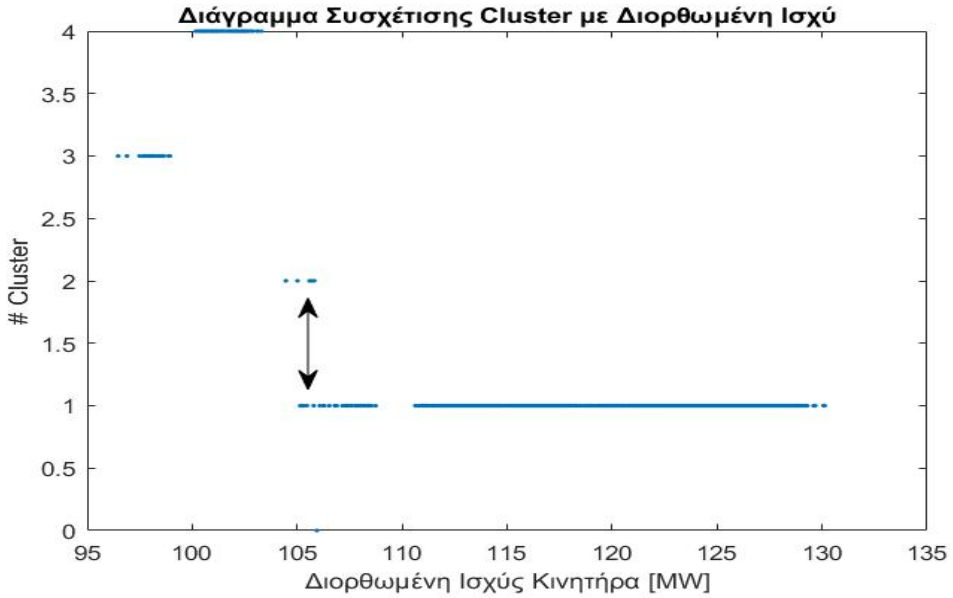
# Clusters	8	7	6	5	4	3	2	1
Epsilon	0.0028	0.0032	0.005	0.02	0.03	0.045	0.075	>0.08

Σχήμα 3.52: Ελάχιστο epsilon ανάλογα με τον επιθυμητό αριθμό Clusters

Όπως και προηγουμένως, παρατηρείται ότι οι τιμές του *epsilon* αυξάνονται όσο μειώνονται τα απαιτούμενα Clusters.
Συνεχίζοντας, για όλες τις μεθόδους όπως και προηγουμένως, παράγονται τα μητρώα:

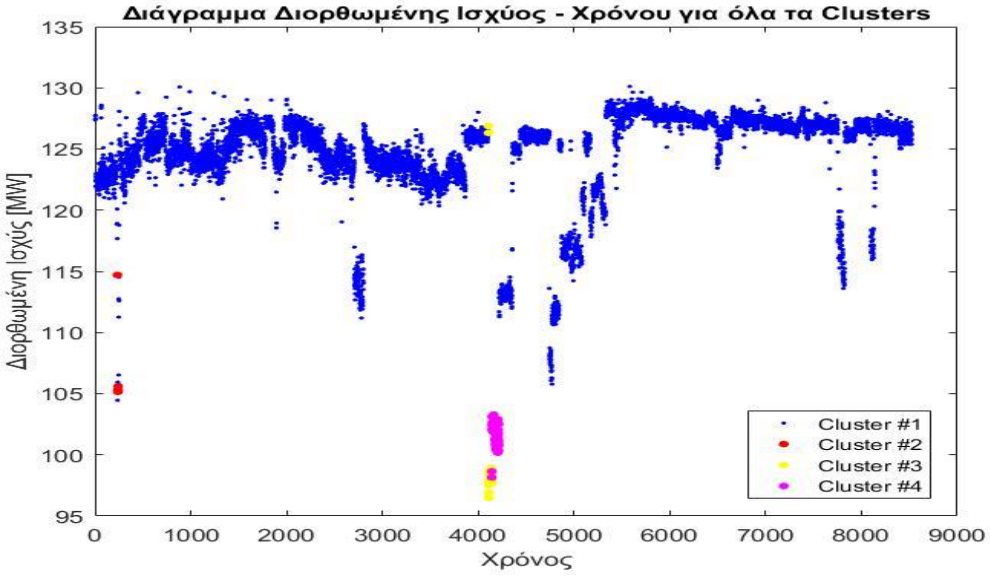
- 24 * 8, που περιέχει τις μέσες τιμές ανά θερμοστοιχείο ανά Cluster και
- 8 * 8, που περιέχει τους Συντελεστές Αλληλοσυσχέτισης των στηλών του παραπάνω μητρώου και είναι διαγώνιο συμμετρικό.

Ορίζοντας ως κριτήριο ότι η ελάχιστη δυνατή ομοιότητα ώστε να θεωρούνται δύο Clusters αρκετά όμοια για ομαδοποίηση ίση με 98%, εκτελούνται τα τρεξίματα του αλγορίθμου με διαφορετικό *epsilon* κάθε φορά αντίστοιχο του μέγιστου αριθμού Clusters. Για τις υπόλοιπες μεθόδους αρκεί να δοθεί ο αριθμός των Clusters. Η διαδικασία διεξαγωγής αποτελεσμάτων και παραγωγής διαγραμμάτων είναι παρόμοια με αυτήν που χρησιμοποιήθηκε για τις μετρήσεις του κινητήρα GTB1. Έτσι λοιπόν, συνολικά τα διαγράμματα που ακολουθούν είναι τα τελικά για τις μεθόδους που αναπτύχθηκαν με ευκλείδεια απόσταση στα δεδομένα της μηχανής GTB2:



Σχήμα 3.53: Ελάχιστο epsilon ανάλογα με τον επιθυμητό αριθμό Clusters

Οι μέθοδοι DBSCAN και K-means παράγουν πανομοιότυπο το παραπάνω διάγραμμα για επιθυμητό αριθμό Clusters (έμμεσα ή άμεσα) ίσο με 4. Η μέθοδος AHC δεν θα ελεγχθεί στη συγκεκριμένη διαδικασία καθώς τα αποτελέσματα της μετά από πολλές επαναλήψεις κρίθηκαν ασταθή και με πολλά σημεία θορύβου, δηλαδή σημεία με αριθμό Cluster ίσο με το 0 που αλλοιώνουν τις μετρήσεις. Για τις άλλες δύο μεθόδους λοιπόν, από το παραπάνω διάγραμμα, γίνεται διακριτό ότι οι μετρήσεις των 95 με 100 MW ανήκουν στο ίδιο Cluster, ως αναμενόμενο, και το ίδιο ισχύει για τις μετρήσεις των 100 με 105 MW. Το πρόβλημα που προκύπτει στο συγκεκριμένο διάγραμμα είναι η επικάλυψη στα εύρη του 1ου και του 2ου Cluster. Πιο συγκεκριμένα, φαίνεται ότι τιμές ισχύος κοντά στα 105 MW υπάρχουν παράλληλα και στο 1ο και στο 2ο Cluster. Για περαιτέρω διερεύνηση του φαινομένου αυτού θα χρειαστεί το διάγραμμα μεταβολής της Διορθωμένης Ισχύος συναρτήσει του χρόνου ανά Cluster που ακολουθεί:



Σχήμα 3.54: Διορθωμένη ισχύς συναρτήσει του χρόνου για όλα τα Clusters

Στο διάγραμμα αυτό φαίνονται όλες οι μετρήσεις διορθωμένης ισχύος συναρτήσει του χρόνου, με διαφορετικό χρώμα, ανάλογα με το Cluster στο οποίο κατατάχθηκαν. Φάνηκε παραπάνω ότι τα Cluster που επικαλύπτονται μεταξύ τους είναι τα Cluster 1 και 2. Οι μετρήσεις αυτές αντιστοιχούν σε χρόνο περίπου ίσο με 250 όπως φαίνεται παραπάνω. Για τις μετρήσεις κοντά σε αυτόν το χρόνο, παρατηρείται ότι η διορθωμένη ισχύς ακολουθεί μία απότομη μεταβολή από τα 120 με 125 MW στα 105 MW. Η μεταβολή αυτή μπορεί να οφείλεται σε διάφορους λόγους όπως:

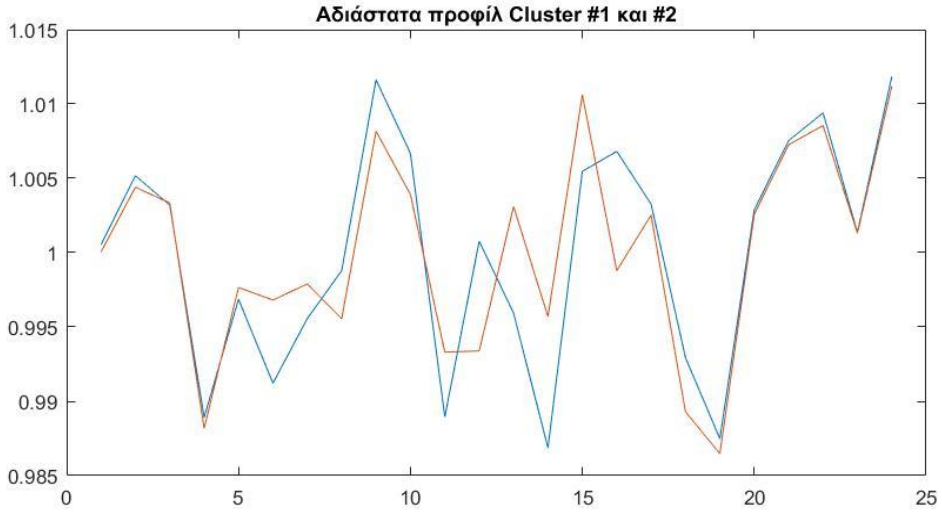
- Αλλαγή συνθηκών λειτουργίας
- Λάθος μετρητικού οργάνου
- Απότομη δυσλειτουργία και επαναφορά στην κανονική λειτουργία

Σε κάθε περίπτωση, τα σημεία του Cluster 2 φαίνεται ότι διαφέρουν σημαντικά από αυτά του Cluster 1. Πιο συγκεκριμένα, η διαφορά των δύο Cluster μπορεί να φανεί σε δύο ακόμα σημεία. Πρώτα στη μητρώο C που σχηματίζεται. Το μητρώο C υπολογίζει τους συντελεστές αλληλοσυσχέτισης των στηλών του μητρώου B. Το μητρώο B περιγράφει τις μέσες τιμές των θερμοστοιχείων ανά Cluster. Για το τρέξιμο των τεσσάρων Clusters τα

μητρώα B και C έχουν διαστάσεις 24 * 4 και 4 * 4, αντίστοιχα. Η τιμή που περιγράφει την ομοιότητα των Cluster 1 και 2 προκύπτει από το μητρώο C ίση με:

$$Ομοιότητα(1,2) = 0,4599$$

Το δεύτερο σημείο διαφοράς του Cluster 1 με το Cluster 2 προκύπτει αν ελέγχουν οι τιμές των 24^{ων} θερμοστοιχείων μία προς μία. Παρακάτω φαίνονται τα αδιάστατα προφίλ για τα Clusters #1 και #2.



Σχήμα 3.55: Αδιάστατα προφίλ Cluster #1 και Cluster #2

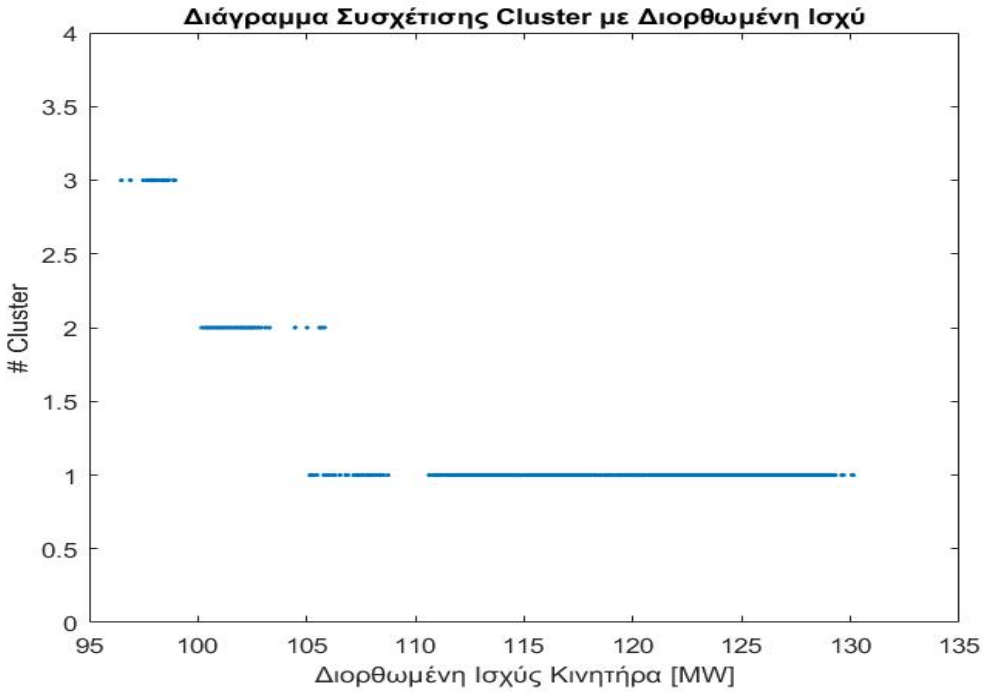
Στο παραπάνω Σχήμα φαίνονται οι τιμές του αδιάστατου προφίλ κάθε cluster. Παρατηρώντας τις τιμές της εικόνας μεταξύ σημείων του Cluster 1 και του Cluster 2 μπορεί να διαπιστωθεί η ανομοιότητα των δύο αυτών Cluster. Αναλυτικότερα, 3 από τα 24 θερμοστοιχεία έχουν σημαντικές αποκλίσεις (μεγαλύτερες του 2%) μεταξύ των δύο Clusters άρα η διαφοροποίηση τους θεωρείται απαραίτητη. Η αμφίβολη ανεξαρτησία του Cluster 2, προκαλεί ενδιαφέρον περαιτέρω διερεύνησης, για αυτό λοιπόν το λόγο εισάγεται θόρυβος στα δεδομένα μέσω της εντολής:

$$randbetween = \pm 5\%$$

Μέσω της ανάλυσης ευαισθησίας, μπορεί να ελεγχθεί η ανεξαρτησία του Cluster 2. Πράγματι, εφαρμόζοντας την αλλαγή αυτή στους αλγόριθμους όλων των μεθόδων, το αποτέλεσμα που προέκυψε από κοινού ήταν διαφορετικό. Και οι τρεις μέθοδοι κατέληξαν ότι ο βέλτιστος αριθμός ομαδοποίησης είναι:

$$3\ Clusters, μέγιστη\ ομοιότητα: 60 - 65\%$$

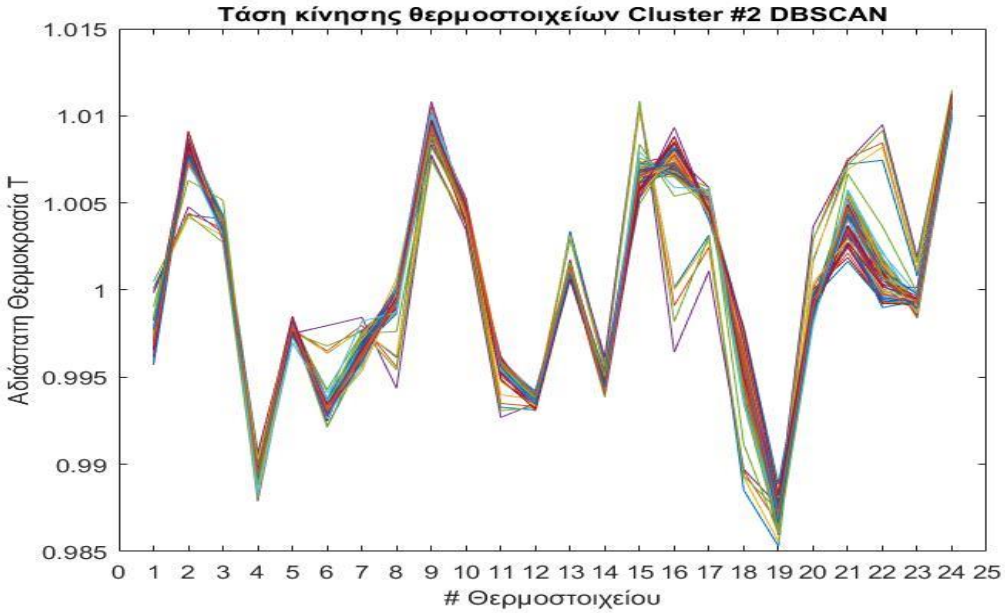
Το διάγραμμα συσχέτισης Διορθωμένης Ισχύος με Clusters στα οποία ανήκουν τα σημεία προέκυψε από κοινού και φαίνεται παρακάτω:



Σχήμα 3.56: Συσχέτιση Cluster με διορθωμένη ισχύ για μικρότερο epsilon

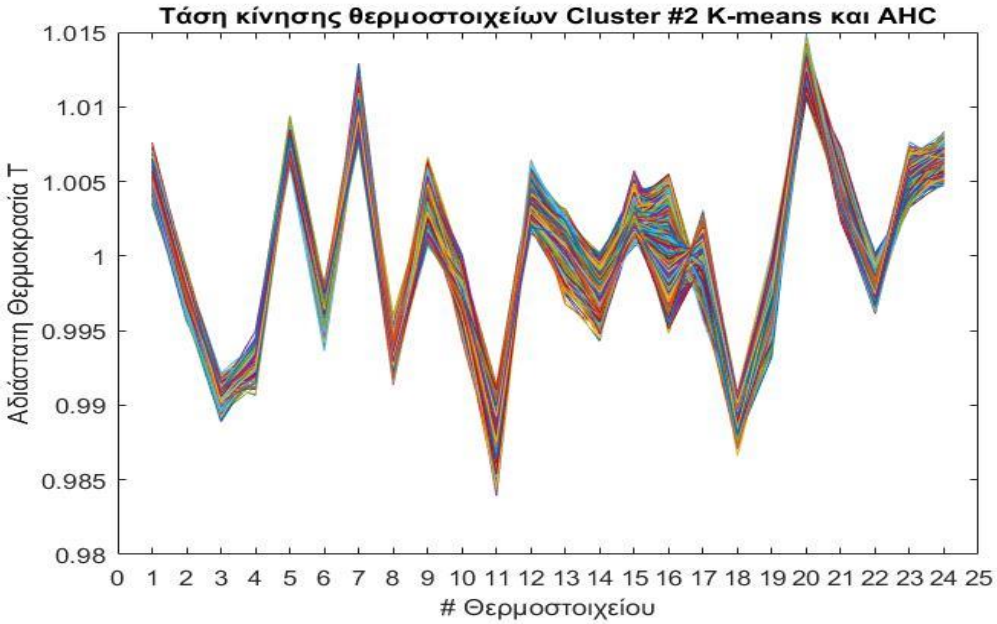
Όπως φαίνεται στο διάγραμμα οι τιμές που νωρίτερα ανήκαν σε ξεχωριστό Cluster 2, τώρα έχουν ομαδοποιηθεί με τις τιμές που νωρίτερα ήταν το Cluster 4. Σε αυτό το σημείο να επισημανθεί ότι το παραπάνω διάγραμμα είναι αποτέλεσμα Αριθμού Διασταυρωμένης Επικύρωσης (Cross Validation Number) ίσο με 5. Δηλαδή, επειδή η δημιουργία νέων τυχαίων μετρήσεων μπορεί να επιφέρει ασταθή λύση, παράγονται 5 αντίτυπα του αρχικού μοντέλου με διαφορετικές τυχαίες τιμές το καθένα των οποίων τα αποτελέσματα διασταυρώνονται και παράγουν έναν αριθμό Ομοιότητας που στην προκειμένη περίπτωση είναι: 99.86%. Άρα, το παραπάνω διάγραμμα αποτελεί ευσταθές αποτέλεσμα της μεταβολής που επιβλήθηκε.

Για να ελεγχθεί εάν η ομαδοποίηση των παλιών Clusters 2 και 4 στο νέο Cluster 2, είναι χρήσιμο να παρασταθούν τα διαγράμματα τάσης κίνησης των θερμοστοιχείων για όλες τις μεθόδους:



Σχήμα 3.57: Τάση κίνησης θερμοστοιχείων του Cluster #2 για DBSCAN

Η μέθοδος DBSCAN από ό,τι φαίνεται έχει σημαντικές αποκλίσεις στα θερμοστοιχεία όπως τα: 6, 14, 16 και 22.



Σχήμα 3.58: Τάση κίνησης θερμοστοιχείων του Cluster #2 για K-means και AHC

Οι μέθοδοι K-means και AHC παράγουν παρόμοιο διάγραμμα τάσης κίνησης οπότε μελετώνται μαζί. Οι αποκλίσεις εδώ δεν είναι τόσο έντονες όσο στο διάγραμμα της DBSCAN, αλλά και εδώ έχουμε μέσο «πάχος» μεγαλύτερο από αυτά για παράδειγμα των μετρήσεων του GTB1 αεριοστρόβιλου. Το συμπέρασμα λοιπόν που προκύπτει είναι ότι το Cluster 2 (των 3 τιμών) θα πρέπει να θεωρηθεί ανεξάρτητο μέχρι περαιτέρω διερεύνησης. Επίσης, μπορεί κατά κανόνα η DBSCAN για 24 θερμοστοιχεία να λαμβάνει

τιμή: $MinPts = 25$, και άρα ένα Cluster σαν το 2 να θεωρείται θόρυβος, ή μέρος άλλου Cluster, παρόλα αυτά, η διαφορά των τιμών των θερμοστοιχείων του Cluster 2 δείχνει ότι η μη ανεξαρτησία του από τα άλλα Cluster είναι σίγουρα αμφισβητήσιμη.

Τέλος, στον παρακάτω πίνακα φαίνονται οι μέθοδοι με τις συναρτήσεις απόστασης που χρησιμοποιήθηκαν καθώς και το πλήθος γειτονιών που αναγνωρίστηκαν με τα εύρη λειτουργίας τους και τον βαθμό ομοιότητας που τα χαρακτηρίζει:

Μέθοδος	Μετρική	Αριθμός Clusters	Μέγιστη Ομοιότητα	Υπολογιστικός Χρόνος (sec)
DBSCAN	Ευκλείδεια	4	80%	9
	CCD	4	81%	760
K-means	Ευκλείδεια	4	61%	21
	CCD	4	68%	28
AHC	Ευκλείδεια	4	80%	12
	CCD	4	84%	14

Σχήμα 3.59: Αποτελέσματα αριθμού Clusters, μέγιστης ομοιότητας και χρόνου για τις μεθόδους που αναπτύχθηκαν

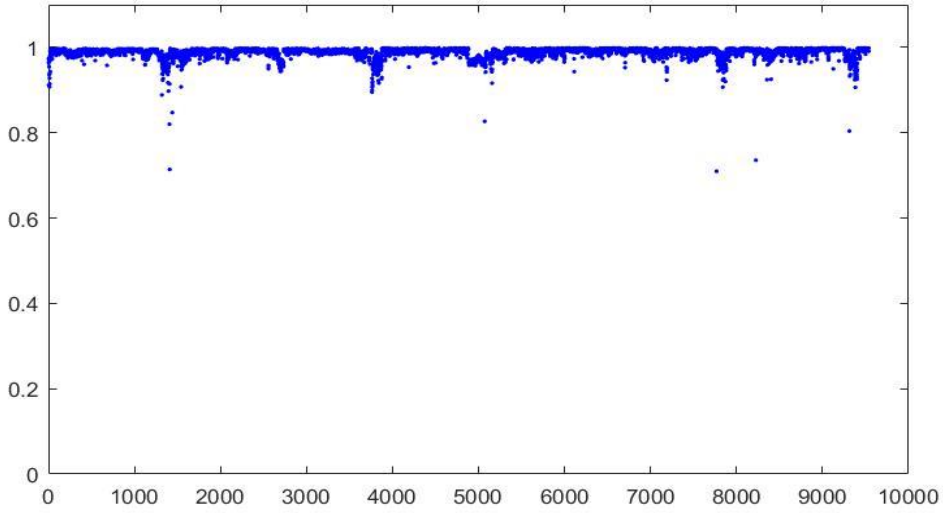
Παρακάτω φαίνονται το εύρος ισχύος των τελικών τεσσάρων Clusters:

- **1ο Cluster: 106 με 130 MW**
- **2ο Cluster: 105 με 106 MW**
- **3ο Cluster: 90 με 100 MW**
- **4ο Cluster: 100 με 105 MW**

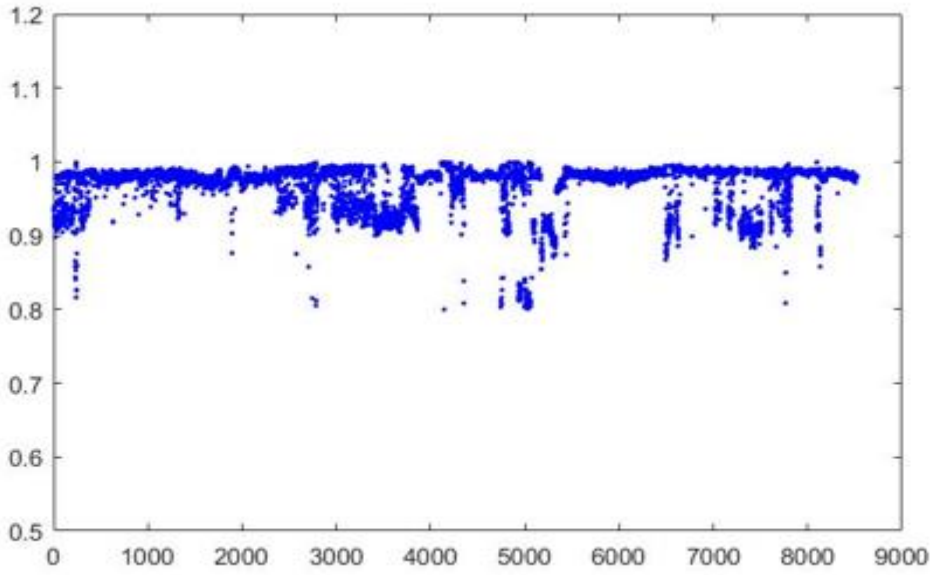
Γνωρίζοντας ότι με τα δεδομένα αυτά η λειτουργία του αεριοστρόβιλου μπορεί να ομαδοποιηθεί στα τέσσερα αυτά Clusters, είναι πλέον δυνατή η παραγωγή Μέσου Προφίλ Αναφοράς (υπογραφή) των θερμοστοιχείων για κάθε ένα από τα Clusters. Είναι δυνατή δηλαδή η λειτουργία του αεριοστρόβιλου σε διάφορα εύρη ισχύος με τις κατάλληλες υπογραφές να συνοδεύουν το μοντέλο και να συγκρίνονται με σκοπό τη διεξαγωγή διαγνωστικής πληροφορίας για την υγεία του κινητήρα. Επίσης, στα συγκεκριμένα δεδομένα μπορεί κανείς να συμπεράνει ότι η μέθοδος αυτή, πέρα από παραγωγή προφίλ αναφοράς μπορεί να λειτουργήσει και διαγνωστικά.

Υπολογισμός CCD για τους κινητήρες GTB1 και GTB2

Το κριτήριο τάσης κίνησης θερμοστοιχείων, είναι ένα αξιόπιστο εργαλείο για να ονοματιστεί ένα cluster ως ανεξάρτητο ή όχι, υπάρχουν όμως και άλλοι τρόποι. Εφόσον έγινε η κατανομή των δεδομένων σε cluster αναφοράς είναι δυνατό πλέον να ελεγχθεί η ορθότητα της κατανομής ενός σημείου στο εκάστοτε cluster μέσω του συντελεστή CCD. Για κάθε δεδομένο λοιπόν των δύο κινητήρων, υπολογίστηκε το CCD με το cluster που κατανεμήθηκε. Αυτή η διαδικασία γίνεται εφικτή παράγοντας το μέσο προφίλ κάθε cluster που σχηματίστηκε από τις τιμές των δεδομένων που κατανεμήθηκαν εκεί. Έτσι τα αποτελέσματα για τους κινητήρες GTB1 και GTB2 είναι τα εξής:

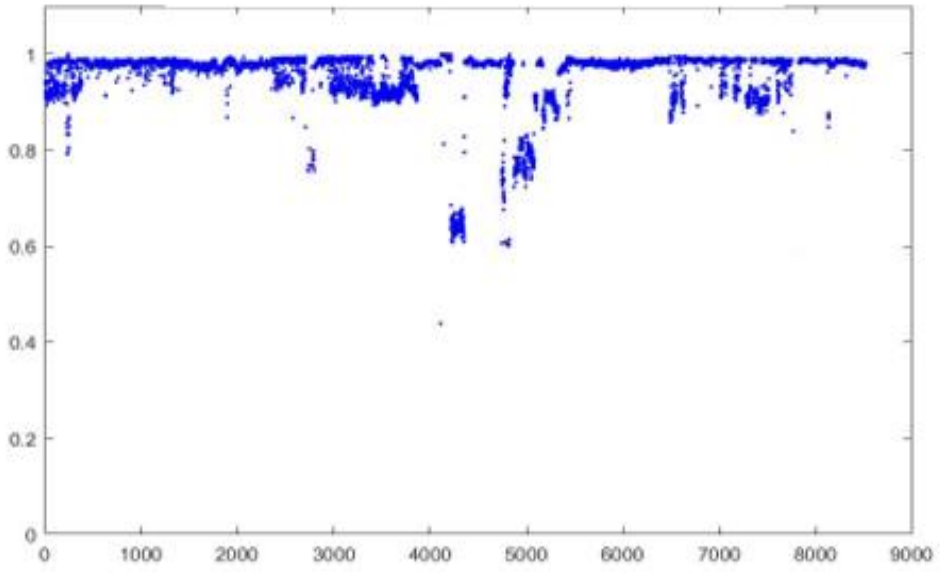


Σχήμα 3.60: CCD για τα δεδομένα GTB1



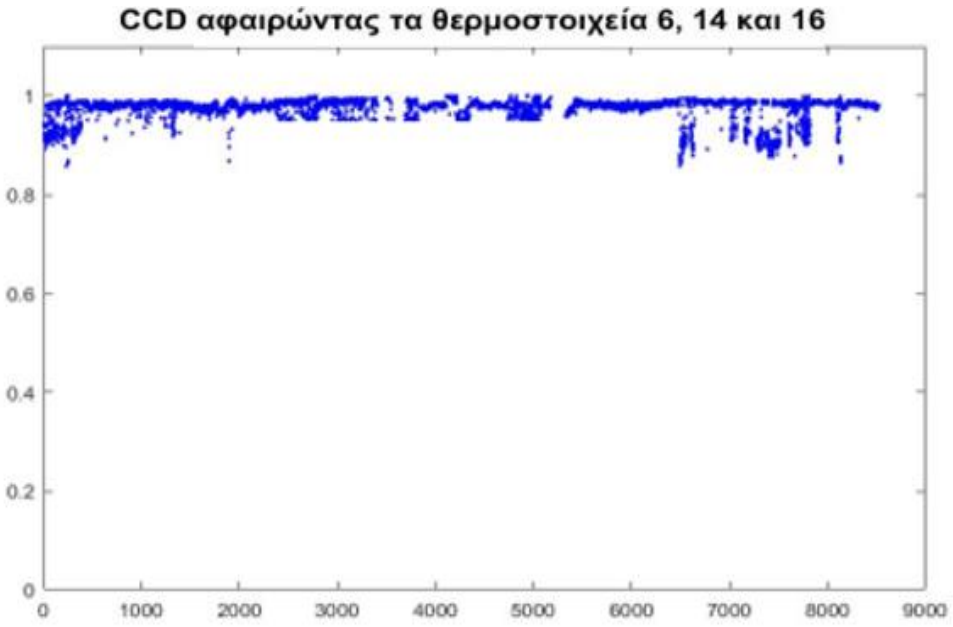
Σχήμα 3.61: CCD για τα δεδομένα GTB2

Όπως φαίνεται, τα αποτελέσματα για τη GTB1 μηχανή είναι πολύ ικανοποιητικά. Οι τιμές του Συντελεστή Αλληλοσυσχέτισης είναι κοντά στη μονάδα για όλα σχεδόν τα σημεία. Πιο συγκεκριμένα, οι τιμές που έχουν CCD μικρότερο του 95% είναι 54, δηλαδή 0.57%. Αυτό σημαίνει ότι ο αλγόριθμος είχε επιτυχία 99.43% για τα δεδομένα της GTB1. Από την άλλη πλευρά τα αποτελέσματα της GTB2 μηχανής είναι απλώς ικανοποιητικά. Πάλι η πλειοψηφία των σημείων βρίσκεται κοντά στη μονάδα, αλλά υπάρχουν αρκετές μετρήσεις που έχουν σχετικά χαμηλό CCD. Πολλές εκ των οποίων αντιστοιχούν σε θέσεις όπως για παράδειγμα του Cluster 2 για τη GTB2 μηχανή, που η ανεξαρτησία του αμφισβητείται. Όπως αναφέρθηκε και νωρίτερα, υπάρχουν θερμοστοιχεία (3 από τα 24) που δεν συμμορφώνονται πλήρως με τα υπόλοιπα του cluster που ανήκουν. Ας δούμε λοιπόν τι θα συμβεί αν αφαιρέσουμε αυτά τα σημεία από τον υπολογισμό CCD. Επαναλαμβάνεται λοιπόν η παραπάνω διαδικασία για όλα τα θερμοστοιχεία πέρα από το 16 πρώτα. Τα αποτελέσματα φαίνονται παρακάτω:



Σχήμα 3.62: CCD με αφαίρεση του 16 θερμοστοιχείου για τη GTB2

Παρατηρούνται καλύτερα αποτελέσματα. Ενδιαφέρον έχει και η εκδοχή όπου αφαιρούνται και τα υπόλοιπα θερμοστοιχεία δηλαδή τα: 6, 14 και 16. Τα αποτελέσματα είναι τα εξής:



Σχήμα 3.63: CCD με αφαίρεση όλων των εσφαλμένων θερμοστοιχείων για τη GTB2

Παρατηρείται ότι τα αποτελέσματα είναι εμφανώς καλύτερα. Υπάρχουν ακόμα μετρήσεις που αποκλίνουν της μονάδας αλλά επί το πλείστων υπάρχει ομοιομορφία. Δικαίως λοιπόν διαχωρίστηκαν τα συγκεκριμένα θερμοστοιχεία και αξίζει να μελετηθεί παραπάνω η κατάσταση λειτουργίας τους.

4 Διάγνωση βλαβών αισθητήρων

Σε αυτό το κεφάλαιο οι μέθοδοι χρησιμοποιούνται ως διαγνωστικά εργαλεία για τη διάγνωση βλαβών αισθητήρων στους κινητήρες. Γίνεται περιγραφή της μεθόδου που χρησιμοποιείται για διάγνωση, χρήση σε προσομοιωμένα δεδομένα για τον έλεγχο της διαγνωστικής ικανότητας και χρήση σε πραγματικά δεδομένα τριών αεριοστρόβιλων.

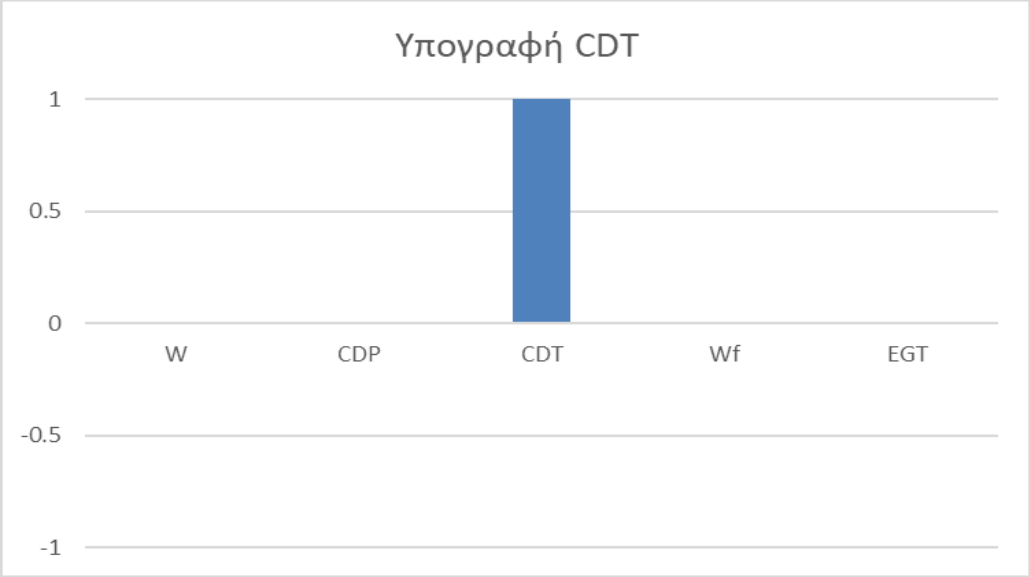
4.1 Γενικά

Η διάγνωση βλαβών αισθητήρων είναι ένα πολύ σημαντικό κομμάτι στην ανάλυση λειτουργίας ενός αεριοστρόβιλου. Στην παρούσα εργασία δόθηκε έμφαση στη θερμοδυναμική διαγνωστική οπότε οι μετρήσεις που μας ενδιαφέρουν είναι οι εξής:

- Παροχή αέρα (W)
- Πίεση εξόδου του Συμπιεστή (CDP)
- Θερμοκρασία εξόδου του Συμπιεστή (CDT)
- Παροχή καυσίμου (Wf)
- Θερμοκρασία εξόδου καυσαερίων (EGT)

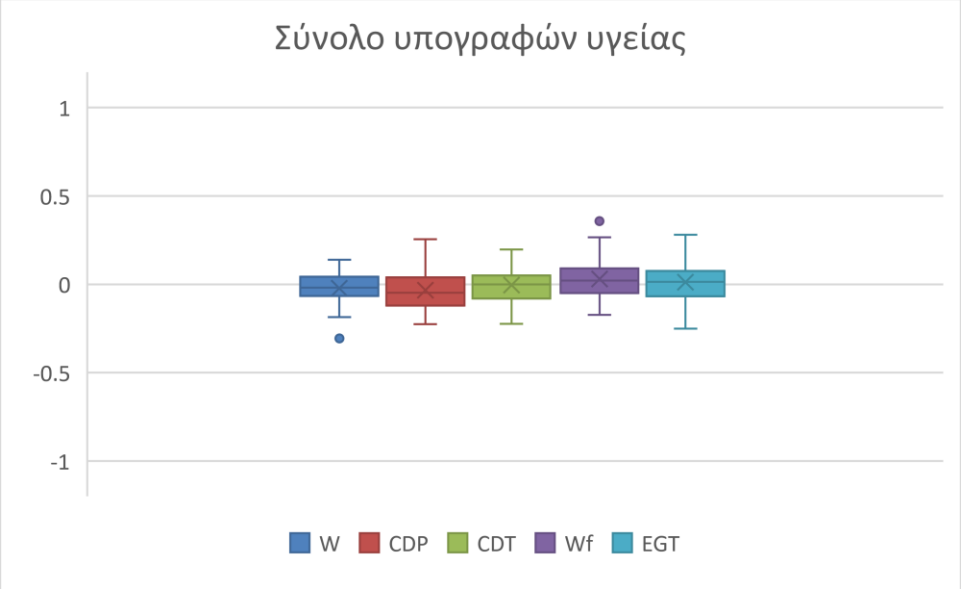
Αυτές είναι οι βασικές μετρήσεις που παρέχει το μοντέλο λειτουργίας ενός αεριοστρόβιλου για δεδομένες συνθήκες περιβάλλοντος και παραγόμενη ισχύ. Από τους πέντε αυτούς αισθητήρες θεωρήθηκε για την εργασία αυτή ότι μπορεί να υπάρχει βλάβη ταυτόχρονα σε δύο από αυτούς. Ο τρόπος με τον οποίο έγινε εισαγωγή αυτής της θεώρησης στον αλγόριθμο που αναλύεται στη συνέχεια είναι μέσω υπογραφών βλαβών. Οι υπογραφές αυτές είναι οι ποσοστιαίες διαφορές των μετρήσεων από τις αντίστοιχες που εκτιμά το θερμοδυναμικό μοντέλο για υγιή λειτουργία, γνωστές και ως, δέλτας. Μελετήθηκαν οι ποσοστιαίες διαφορές ώστε οι υπογραφές αυτές να είναι ανεξάρτητες της μηχανής που αναλύεται. Έτσι λοιπόν, ο χρήστης φορτώνοντας αυτές τις υπογραφές και τις μετρήσεις που επιθυμεί να αναλύσει στον αλγόριθμο που αναλύεται στην επόμενη υπο-ενότητα, μπορεί να έχει έγκυρο διαγνωστικό αποτέλεσμα.

Ο τρόπος παραγωγής αυτών των υπογραφών για τα cluster υγείας και βλάβης αποτελείται από δύο βήματα. Το βήμα παραγωγής μίας υπογραφής, δημιουργώντας μία αδιάσταση μεταβολή 1% των δέλτας στις μετρήσεις για κάθε βλάβη. Για παράδειγμα στο σχήμα 4.1 φαίνεται η υπογραφή της βλάβης CDT:

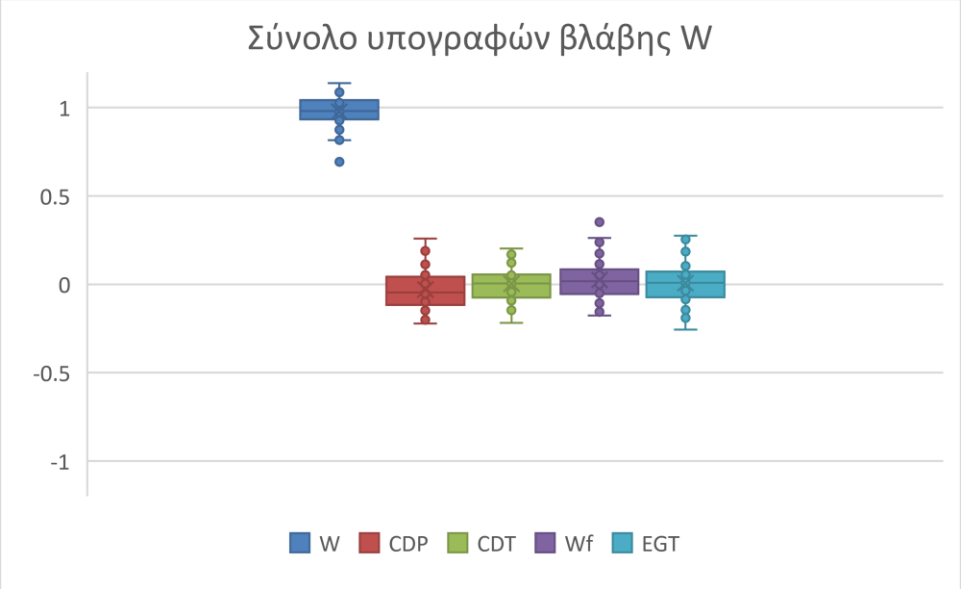


Σχήμα 4.1: Υπογραφή που περιγράφει τη βλάβη CDT

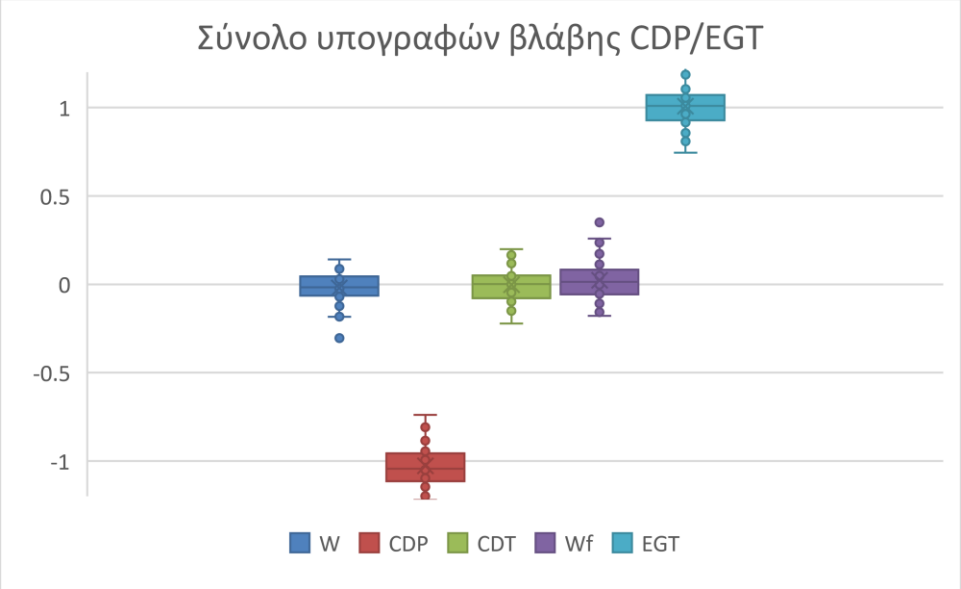
Όπως φαίνεται οι υπογραφές αυτές μοντελοποιούν τα cluster υγείας, τα cluster απλής βλάβης, τα cluster συνδυασμένης βλάβης και τα cluster συνδυασμένης βλάβης με ομόρροπη και αντίρροπη φορά βλάβης των συνιστωσών. Οι υπογραφές αυτές αφορούν το 50% των περιπτώσεων καθώς υπάρχουν και οι αντίθετες εκδοχές τους, οι οποίες για λακωνικότητα δεν δημιουργήθηκαν αλλά ελέγχονται παρόλα αυτά στον αλγόριθμο κατά τη σύγκριση με τις υπογραφές βλάβης ταιριάζοντας το ενδεχόμενο εξεταζόμενο σημείο με τα cluster βλάβης και με τα αντίθετα τους πολλαπλασιάζοντας το με -1.
Στο δεύτερο βήμα, στις υπογραφές αυτές προστίθεται θόρυβος $\sigma = \pm 0.15\%$ ώστε να σχηματιστούν 50 από αυτά για κάθε υπογραφή. Έτσι προστίθεται ο τυχαίος παράγοντας με σκοπό την καλύτερη αντιμετώπιση των νέων σημείων και των τυχαίων μοτίβων που περιέχουν. Παρακάτω φαίνονται τρία παραδείγματα υπογραφών για τα clusters υγείας, βλάβης W και βλάβης CDP/EGT:



Σχήμα 4.2: Σύνολο υπογραφών υγείας



Σχήμα 4.3: Σύνολο υπογραφών βλάβης W

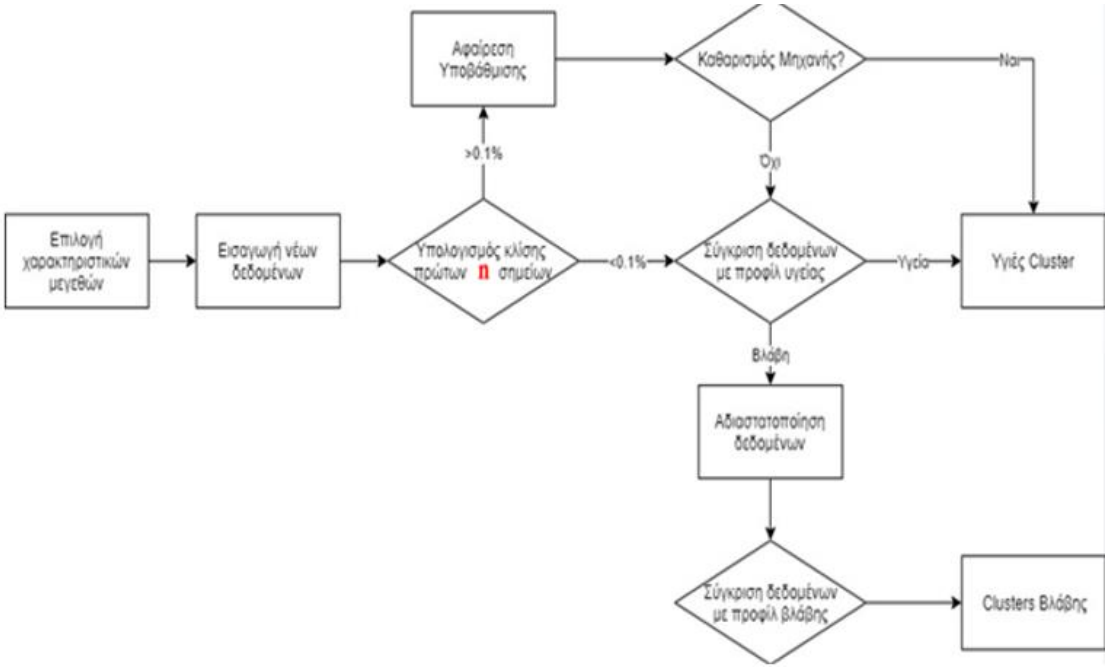


Σχήμα 4.4: Σύνολο υπογραφών συνδυασμένης αντίρροπης βλάβης CDP/EGT

Στα παραπάνω σχήματα φαίνεται το σύνολο των υπογραφών υγείας και βλάβης. Οι τιμές όλων των μεταβλητών να κινούνται γύρω από το 0 ενώ την τιμή του W για παράδειγμα στη δεύτερη υπογραφή κινείται γύρω από την μονάδα. Οι υπογραφές που παράχθηκαν συνολικά ήταν 26 και έχουν την εξής μορφή για κάθε κατάσταση: Υγεία, Βλάβη W, Βλάβη CDP, Βλάβη CDT, Βλάβη Wf, Βλάβη EGT, Βλάβη W/CDP, Βλάβη W/CDP αντίρροπη, Βλάβη W/CDT, και ούτω καθεξής. Στη συνέχεια εξηγείται πως χρησιμοποιεί ο αλγόριθμος τις υπογραφές αυτές και πως η κάθε μία από τις μεθόδους που επιλέχθηκαν συγκρίνει τα εξεταζόμενα δεδομένα με τις υπογραφές αυτές.

4.2 Αλγόριθμος

Κάθε μία από τις μεθόδους που αναπτύχθηκαν χαρακτηρίζονται από διαφορετικές εντολές και επαναλήψεις καθώς η αρχή λειτουργίας τους είναι διαφορετική. Παρόλα αυτά, οι μέθοδοι έχουν μερικές κοινές διαδικασίες διότι ο γενικός τρόπος λειτουργίας τους είναι ίδιος. Στο σχήμα 4.5 φαίνεται το διάγραμμα ροής που ακολούθησαν από κοινού οι αλγόριθμοι των μεθόδων που επιλέχθηκαν:



Σχήμα 4.5: Διάγραμμα ροής μεθόδου διάγνωσης βλαβών

Όπως φαίνεται η διαδικασία έχει διάφορα βήματα που αναλύονται στις επόμενες ενότητες. Η λειτουργία των αλγορίθμων ξεκινά διαβάζοντας την τιμή που εισάγει ο χρήστης στη μεταβλητή w:

- $αν\ w = 1, τότε\ θα\ χρησιμοποιηθεί\ Ευκλείδεια\ Απόσταση$
- $αν\ w = 2, τότε\ θα\ χρησιμοποιηθεί\ Συντελεστής\ Αλληλοσυσχέτισης$

Έπειτα, διαβάζεται το εξεταζόμενο μητρώο από τη MATLAB από αρχείο τύπου Excel μέσω της εντολής: $B = xlsread('deltas.xlsx')$; Το μητρώο αυτό συνήθως περιέχει μία ή παραπάνω στήλες μετρήσεων συναρτήσει του χρόνου, δηλαδή χρονοσειρές. Στη συνέχεια, αναπτύχθηκαν δύο τρόποι διάγνωσης. Ο πρώτος ήταν με εφαρμογή των μεθόδων στο σύνολο των μετρήσεων παράγοντας έτσι όλα τα ενδεχόμενα Clusters (με τα σημεία που τους ανήκουν) ταυτόχρονα. Αυτή η μέθοδος μπορεί να έχει εφαρμογή μόνο αν μηχανικοί της εγκατάστασης έχουν παραδώσει μετρήσεις προς εξέταση των κινητήρων μετά τη διακοπή λειτουργίας τους (Offline diagnosis). Προσφέρεται λοιπόν συνολική εικόνα της κατάστασης των κινητήρων εκ των υστέρων. Εφόσον ο σκοπός αυτής της εργασίας είναι η πρόγνωση, η μέθοδος αυτή διερευνήθηκε λόγω ακαδημαϊκού ενδιαφέροντος, αλλά δεν παρίσταται στα αποτελέσματα.

Ο δεύτερος τρόπος ήταν η εφαρμογή των μεθόδων για κάθε νέα μέτρηση που προκύπτει ξεχωριστά (Online diagnosis). Πρόκειται για την περίπτωση που ο αεριοστρόβιλος εκτελεί κάποια αποστολή ενώ παράλληλα ελέγχονται οι συντελεστές υγείας του. Με αυτή τη μέθοδο, τα αποτελέσματα είναι αμεσότερα, πιο ακριβή και βοηθούν στην αποτροπή κινδύνου εν ώρα λειτουργίας. Το αρνητικό αυτής της μεθόδου είναι ότι απαιτεί μεγάλο κομμάτι μνήμης για την αποθήκευση στοιχείων κάθε χρονικής στιγμής για την οποία βρίσκεται σε λειτουργία. Για το λόγο αυτό οι μέθοδοι που

χρησιμοποιήθηκαν, εφαρμόσθηκαν σε σετ n+1 σημείων ανά φορά. Η εφαρμογή αυτή γίνεται μέσω των παρακάτω εντολών:

```
X1=[Href; Xn(k,:)];
[I, Noise, Neighbors]=DBSCAN(X1,e01,MinPts,w);
if I(end)==1
    IX(k,1)={'Normal'};
    IA(k,1)=1;
else
    IX(k,1)={'Faulty'};
    IA(k,1)=2;
    i=i-1;
end
```

Σχήμα 4.6: Αλγόριθμος ταξινόμησης σημείων σε υγιή και επιβλαβή

Όπως φαίνεται στο Σχήμα 4.6 ο πίνακας X1 αποτελείται από τους πίνακες Href και την k-γραμμή του πίνακα Xn. Ο πίνακας Xn έχει το σύνολο των επεξεργασμένων μετρήσεων που εισάγονται για διάγνωση στους αλγόριθμους. Ο πίνακας αυτός αποτελείται από X γραμμές, ανάλογα με τις παρατηρήσεις που γίνονται, και πέντε στήλες που περιγράφουν τις τιμές των μετρήσεων που χρησιμοποιούμε για θερμοδυναμική διαγνωστική.

Θεωρώντας ότι μελετάμε σύνολο μετρήσεων πέντε στηλών λοιπόν, ο πίνακας Href έχει διαστάσεις 50 * 5 ενώ η γραμμή k του Xn, 1 * 5. Βάσει της εικόνας και πάλι, στον πίνακα X1 διαστάσεων 51 * 5 πλέον εφαρμόζεται η μέθοδος DBSCAN, η οποία παράγει το μητρώο-στήλη I που περιλαμβάνει τιμές 1 ή 2 για τις 51 γραμμές του πίνακα X1 ανάλογα με το αν το σημείο που περιγράφεται από κάθε γραμμή θεωρείται υγιές ή εσφαλμένο. Στην προκειμένη περίπτωση ενδιαφέρον έχει η τιμή μόνο της τελευταίας γραμμής του μητρώου I, για αυτό εμφανίζεται η εντολή: $if\ I(end) == 1$; Ενδιαφέρον για τη μελέτη αυτή έχει μόνο στην τελευταία γραμμή καθώς είναι γνωστό ότι όλες οι προηγούμενες τιμές του μητρώου είναι ίσες με τη μονάδα. Οι παραπάνω τιμές είναι μονάδα λόγω της δομής του πίνακα Href, η οποία αναλύεται περαιτέρω στη συνέχεια.

Στη συνέχεια, για κάθε μέθοδο η σύγκριση με τις υπογραφές αναφοράς είναι διαφορετική. Συγκεκριμένα για τη DBSCAN, ακολουθεί διαδικασία εύρεσης κατάλληλου αριθμού *epsilon* για κάθε ένα από τα μητρώα. Η επιλογή αυτή γίνεται θεωρώντας ότι όλα τα σημεία του αδιάστατου μητρώου είναι γειτονικά και θα πρέπει να ανήκουν όλα σε ένα Cluster.

```
%% Gia to veltisto (elaxisto) e kathe cluster anaforas (ref)
for epsilon=0.0001:0.001:10
    [I, Noise, Neighbors]=DBSCAN(EGTref,epsilon,MinPts,w);
    if max(I)<1.5 && numel(find(I==0))<3
        break
    end
end
```

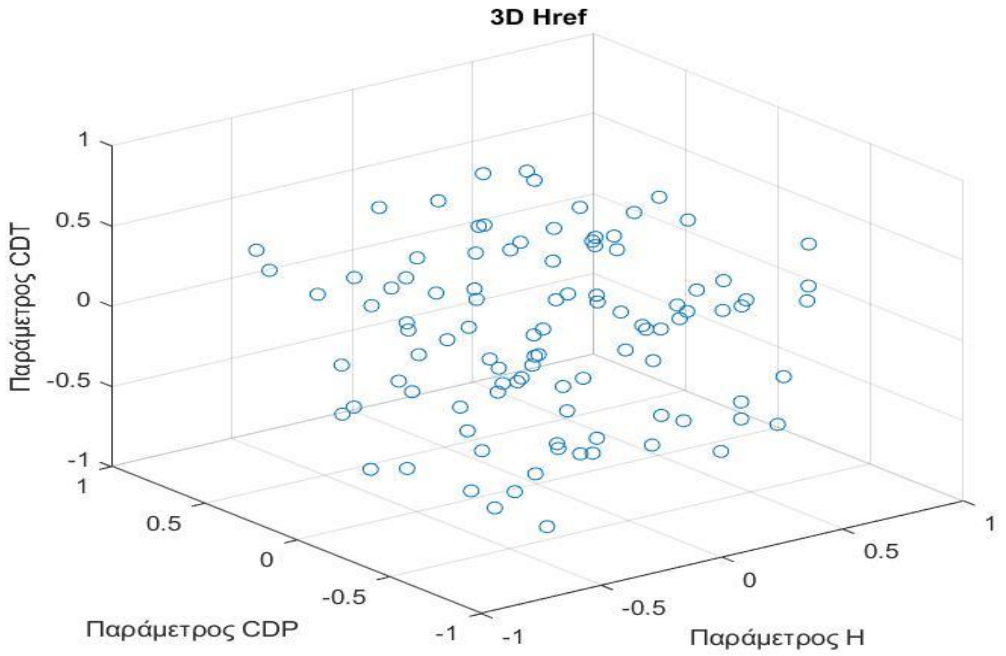
Σχήμα 4.7: Κομμάτι κώδικα εύρεσης ελάχιστου epsilon

Το παραπάνω σύνολο εντολών υπολογίζει το ελάχιστο δυνατό *epsilon* ώστε όλα τα σημεία του μητρώου (EGTref στην προκειμένη) να είναι γείτονες. Αυξάνοντας τις τιμές του *epsilon* και εφαρμόζοντας τη μέθοδο απαιτείται από το μητρώο εξόδου *I* να παρουσιάζει μόνο ένα Cluster και τα σημεία θορύβου να είναι μικρότερα από τρία. Η μέθοδος αυτή εφαρμόζεται για όλα τα μητρώα βλαβών. Έτσι, παράγεται μία σειρά από ελάχιστα *epsilon* που χρησιμοποιείται παρακάτω και έχει τις εξής τιμές:

Epsilon	1	2	3	4	5	6	7	8	9	10	11
Euclidean	0.246	0.144	0.129	0.089	0.156	0.107	0.118	0.176	0.281	0.148	0.189
CCD	0.072	0.043	0.052	0.024	0.021	0.019	0.011	0.015	0.032	0.15	0.19

Σχήμα 4.8: Πίνακας ελάχιστων epsilon ανάλογα με την υπογραφή αναφοράς

Όπως φαίνεται η διαδικασία αυτή επαναλήφθηκε για Ευκλείδεια αλλά και Συντελεστή Αλληλοσυσχέτισης συνάρτηση απόστασης. Στο παραπάνω σχήμα φαίνονται μόνο 11 από τις συνολικά 26 τιμές, μία για κάθε υπογραφή αναφοράς. Με αυτόν τον τρόπο μοντελοποίησης κάθε ένα από τα μητρώα αναφοράς έχει τη μορφή παρόμοια με αυτή της σφαίρας με ακτίνα περίπου ίση με 0.75 (λόγω της εντολής Randbetween). Τα πενήντα σημεία που σχηματίζουν αυτή τη σφαίρα φαίνονται στο παρακάτω τρισδιάστατο σχήμα:



Σχήμα 4.9: Τρισδιάστατο διάγραμμα των τιμών του μητρώου υγιούς υπογραφής

Βάσει του χώρου που καταλαμβάνουν αυτά τα σημεία αργότερα καθορίζεται αν τα νέα εξεταζόμενα σημεία χαρακτηρίζονται ως υγιή ή σημεία βλάβης.

Οι μέθοδοι K-means και AHC δεν χρειάζονται την παράμετρο *epsilon*. Για τις μεθόδους αυτές, γίνεται μία διαφορετική διαδικασία ορισμού των περιοχών βλαβών και των αποδεκτών αποστάσεων από αυτές. Αναφέρθηκε παραπάνω ότι ο αλγόριθμος της K-means έχει ως έξοδο:

- Το μητρώο χαρακτηρισμού I των σημείων ανάλογα με το cluster που ανήκουν
- Το μητρώο C που περιλαμβάνει τα κέντρα των clusters που σχηματίστηκαν

Σε αυτό το σημείο έχει ενδιαφέρον η ανάλυση του μητρώου C. Αφού έχουν σχηματιστεί και εισαχθεί τα μητρώα αναφοράς στον αλγόριθμο, δημιουργείται ένα μητρώο που περιέχει όλα τα μητρώα αναφοράς. Στη συνέχεια, εφαρμόζεται η μέθοδος K-means στο συνολικό αυτό μητρώο, παράγοντας έτσι ένα μητρώο C με τις συντεταγμένες των κέντρων των μητρώων αναφοράς. Η διαδικασία αυτή έχει ως στόχο την αδιαστατοποίηση των μητρώων αναφοράς.

```
%Apostaseis simeiwn anaforas apo to antistoixo kentro tous.
for i=1:length(Href)
    DHr(i,:)=Href(i,:)-Cref(1,:);
    DWr(i,:)=Wref(i,:)-Cref(2,:);
    DCDPr(i,:)=CDPref(i,:)-Cref(3,:);
    DCDTr(i,:)=CDTref(i,:)-Cref(4,:);
    DEGTr(i,:)=EGTref(i,:)-Cref(5,:);
end
%Aktines: Therwntas oti ta cluster mas einai sfaires
RH=max(max(abs(DHr)));
RW=max(max(abs(DWr)));
RCDP=max(max(abs(DCDPr)));
RCDT=max(max(abs(DCDTr)));
```

Σχήμα 4.10: Κομμάτι κώδικα που υπολογίζει τη θέση των νέων κέντρων αναφοράς για την K-means

Οι εντολές που περιγράφονται στην παραπάνω εικόνα αποτελούν το κύριο μέρος αυτής της αδιαστατοποίησης. Πρακτικά η έλλειψη που σχηματίζει κάθε ένα από τα μητρώα υγείας ή βλάβης τοποθετούνται στο σημείο (0,0,0,0) του τετραδιάστατου χώρου. Αφού γίνει αυτό, υπολογίζεται η ακτίνα της κάθε σφαίρας R_i.

4.2.1 Σύγκριση με υγιές cluster

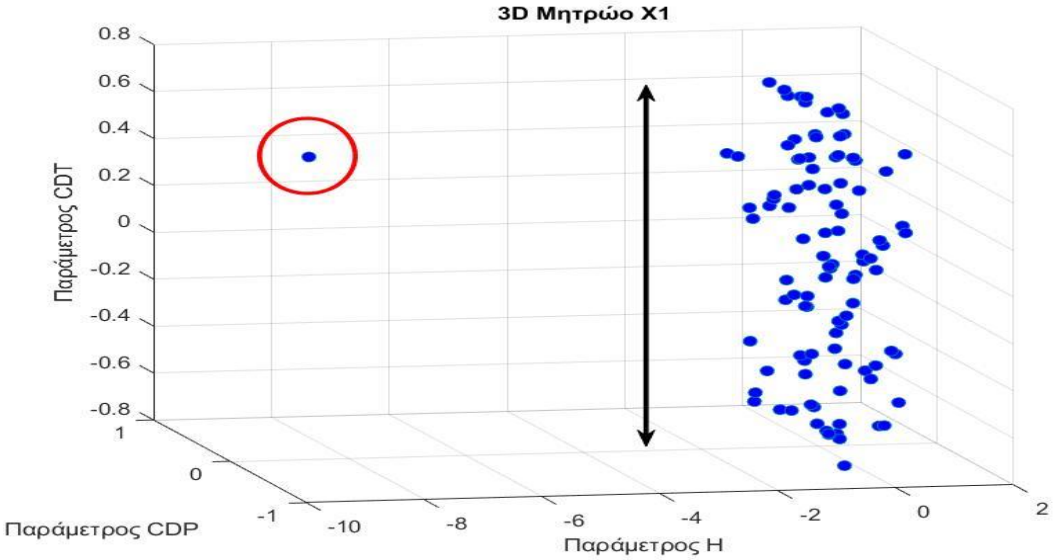
Η διάγνωση των εισαγόμενων σημείων γίνεται σε δύο στάδια διαχωρισμού. Το πρώτο στάδιο διαχωρίζει σημεία υγιούς από σημεία εσφαλμένης λειτουργίας. Το πρώτο στάδιο είναι κοινό σε όλες τις μεθόδους. Πρακτικά ελέγχεται αν το κάθε νέο εξεταζόμενο σημείο ανήκει στο cluster αναφοράς υγείας Href. Ο έλεγχος αυτός προϋποθέτει τη δημιουργία του μητρώου X1 που αναπτύχθηκε παραπάνω.

Στη μέθοδο DBSCAN παράγεται το μητρώο I όπου ελέγχεται μόνο η τελευταία τιμή. Αν αυτή η τιμή είναι ίση με 1 τότε το σημείο είναι υγιές, αν είναι ίση με 0 τότε η DBSCAN θεωρεί ότι το σημείο αυτό αποτελεί θόρυβο σε σχέση με το υπόλοιπο σύνολο σημείων, των σημείων δηλαδή του μητρώου Href. Είναι γνωστό ότι η DBSCAN για να

λειτουργήσει σωστά χρειάζεται και το κατάλληλο *epsilon*. Λαμβάνοντας την αντίστοιχη τιμή από τον Πίνακα με τα *epsilon*, εξάγεται το επιθυμητό αποτέλεσμα.

Η διαδικασία δε διαφέρει πολύ στη μέθοδο K-means. Εκεί, η σύγκριση υγιούς/εσφαλμένου γίνεται μέσω της μοντελοποιημένης έλλειψης και της αντίστοιχης ακτίνας. Το νέο εξεταζόμενο σημείο αδιαστατοποιείται και αυτό βάσει του κέντρου του υγιούς Cluster και ελέγχεται αν η απόσταση του σημείου από το κέντρο είναι μεγαλύτερη ή μικρότερη της ακτίνας του.

Τέλος, για τη μέθοδο AHC ακολουθείται παρόμοια λογική με αυτήν της DBSCAN. Η AHC βρίσκει τις αποστάσεις όλων των σημείων μεταξύ τους και αρχίζει να ομαδοποιεί τα κοντινότερα σημεία μεταξύ τους μέχρι να καταλήξει σε 2 Clusters. Εκεί, αν το Cluster 2 περιέχει μόνο ένα σημείο και ταυτόχρονα αυτό το σημείο είναι το εξεταζόμενο, τότε το εξεταζόμενο σημείο δεν ανήκει στο Cluster υγείας. Ο συγκεκριμένος έλεγχος επαναλαμβάνεται μέσω του μοντέλου διασταύρωσης τυχαίων σημείων (Cross-Validation Process) ώστε να εξασφαλισθεί ότι το εξεταζόμενο σημείο δεν ονοματίζεται υγιές ή με βλάβη λόγω τύχης.



Σχήμα 4.11: Τρισδιάστατη απεικόνιση Cluster αναφοράς και νέου εξεταζόμενου σημείου (κυκλωμένο)

Στο παραπάνω σχήμα φαίνονται οι τιμές του μητρώου X1 για ένα τυχαίο εξεταζόμενο σημείο (στον κόκκινο κύκλο). Από το σχήμα αυτό είναι διακριτό στο ανθρώπινο μάτι ότι το εξεταζόμενο σημείο αποτελεί θόρυβο ή βλάβη, όμως οι μέθοδοι που αναπτύχθηκαν παραπάνω έχουν τους δικούς τους τρόπους για να διευκρινίσουν κάτι τέτοιο. Το χαρακτηριστικό μέγεθος *epsilon* της DBSCAN φαίνεται παραπάνω ότι για το μητρώο Href να παίρνει την τιμή:

$$e_{01} = 1.51$$

Αυτή η τιμή αναπαριστά τη μέγιστη απόσταση μεταξύ δύο σημείων ώστε να θεωρούνται γείτονες. Κοιτώντας τον κατακόρυφο άξονα το μαύρο βέλος του διαγράμματος αναπαριστά ακριβώς αυτήν την τιμή. Επίσης φαίνεται ότι το εξεταζόμενο σημείο έχει ευκλείδεια απόσταση αρκετά μεγαλύτερη από αυτή που μετρά το βέλος, άρα λογικό είναι να θεωρείται το σημείο ως θόρυβο.

Αντίστοιχα, για την K-means η ακτίνα της μοντελοποιημένης έλλειψης, είναι μικρότερη της απόστασης του κέντρου της έλλειψης από το εξεταζόμενο σημείο. Με παρόμοια λογική, η μέθοδος AHC ομαδοποιεί τα κοντινά σημεία και ξεχωρίζει το κυκλωμένο-εξεταζόμενο σημείο ως θόρυβο λόγω της απόστασης. Η επεξήγηση του παραπάνω διαγράμματος με χρήση του Συντελεστή Αλληλοσυσχέτισης γίνεται πιο σύνθετη καθώς δεν δόκιμη οπτικά. Για αυτό προτιμήθηκε η ανάλυση μέσω Ευκλείδειας απόστασης.

4.2.2 Σύγκριση με clusters βλάβης

Αφού ξεχωριστούν τα υγιή από τα σημεία βλάβης, είναι σημαντικό να χαρακτηριστεί κάθε ένα από τα σημεία βλάβης με τη βλάβη που αντιστοιχεί. Η διαδικασία δεν διαφέρει καθόλου από αυτήν του διαχωρισμού υγείας. Αυτή τη φορά τα εξεταζόμενα σημεία συγκρίνονται με αυτά των μητρώων βλάβης. Πρώτα γίνεται σύγκριση με τα μητρώα απλής βλάβης και στη συνέχεια γίνεται επιπλέον έλεγχος για συνδυασμό βλαβών.

Η κύρια διαφορά με την παραπάνω διαδικασία είναι η αδιαστατοποίηση των εξεταζόμενων σημείων πριν τη σύγκριση με τις υπογραφές βλαβών. Αυτό συμβαίνει επειδή οι τιμές των εξεταζόμενων μετρήσεων μπορεί να απέχουν αρκετά από 2 ή περισσότερες υπογραφές βλάβης και οι συναρτήσεις απόστασης που χρησιμοποιούνται να μην καταλήγουν αξιόπιστα στους σωστούς χαρακτηρισμούς βλάβης. Είναι σημαντικό να εξεταζόμενα σημεία να είναι προσαρμοσμένα στην ίδια κλίμακα με τα σημεία αναφοράς.

Και πάλι η κάθε μέθοδος εφαρμόζει διαφορετικά τη σύγκριση αυτή. Η DBSCAN μέσω των *epsilon* και των ήδη σχηματισμένων Cluster βλαβών καταχωρεί κάθε σημείο ανάλογα με την απόσταση. Η K-means μεταφέρει κάθε εξεταζόμενο σημείο από το απόλυτο σύστημα στο σχετικό μέσω της αφαίρεσης συντεταγμένων και στη συνέχεια συγκρίνει την απόσταση από το κέντρο με την ακτίνα. Ενώ η AHC όπως και παραπάνω εξετάζει κάθε σημείο σε σχέση με κάθε Cluster βλάβης και παρέχει πληροφορίες για την ανεξαρτησία ή μη του εξεταζόμενου σημείου. Οι αλγόριθμοι και των τριών μεθόδων παρέχουν ως έξοδο το μητρώο *I* στο οποίο ελέγχεται η τελευταία τιμή για να καθοριστεί το Cluster στο οποίο ανήκει το σημείο.

4.3 Προβλήματα στη διαγνωστική διαδικασία

Στο κεφάλαιο αυτό αναλύονται τα προβλήματα που αντιμετώπισε η διαγνωστική διαδικασία με κυριότερο το φαινόμενο υποβάθμισης αεριοστροβίλων. Γίνεται αναφορά στη σημασία του για τη διεξαγωγή ορθών αποτελεσμάτων και επιλέγεται ο τρόπος υποβάθμισης που χρησιμοποιείται στη συνέχεια της ανάλυσης και μελετώνται οι τρόποι εφαρμογής αυτού.

4.3.1 Γενικά

Η λειτουργία ενός αεριοστρόβιλου είναι το αποτέλεσμα της συντονισμένης συνεργασίας πολλών διαφορετικών συνιστωσών. Οποιοδήποτε από αυτά τα μέρη μπορεί να παρουσιάσει φθορά κατά τη διάρκεια ζωής της συσκευασίας και συνεπώς μπορεί να επηρεάσει δυσμενώς τη λειτουργία του συστήματος. Η κατανόηση των μηχανισμών που προκαλούν υποβάθμιση καθώς και οι επιπτώσεις που μπορεί να έχει η υποβάθμιση ορισμένων συνιστωσών στο συνολικό σύστημα είναι ένα μείζον θέμα στη μελέτη αεριοστρόβιλων. Συγκεκριμένα, τα αεροδυναμικά εξαρτήματα, όπως ο συμπιεστής του κινητήρα, οι στρόβιλοι, οι διαχύτες ή τα ακροφύσια είναι αναγκαίο να λειτουργούν ακόμα και σε περιβάλλον που θα υποβαθμίσει σταθερά την απόδοσή τους.

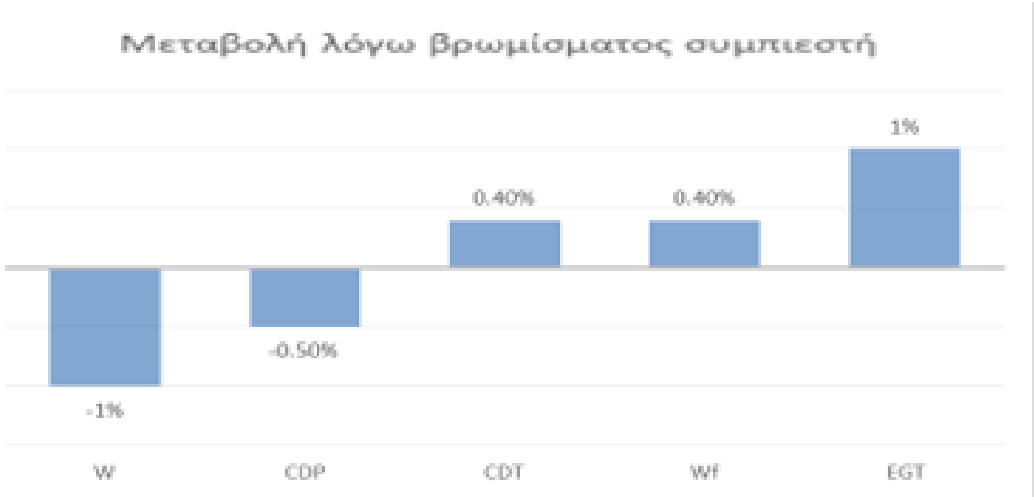
4.3.2 Υποβάθμιση λόγω βρομίσματος συμπιεστή

Γνωρίζοντας λοιπόν ότι τα αίτια πιθανής υποβάθμισης ενός αεριοστρόβιλου ποικίλουν και προκύπτουν συχνά κατά τη διάρκεια ζωής του, είναι σημαντικό να ληφθούν υπόψη στα μοντέλα διάγνωσης που χρησιμοποιήθηκαν. Έχοντας ως στόχο τη μοντελοποίηση προβλήματος που εμφανίζεται συχνότερα, χρησιμοποιήθηκε το πιο διαδεδομένο αίτιο υποβάθμισης ενός αεριοστρόβιλου, δηλαδή η συγκέντρωση σκόνης/ βρομίσματος του συμπιεστή.

Λαμβάνοντας υπόψη την υποβάθμιση αυτή, χρησιμοποιήθηκε το λογισμικό μοντελοποίησης λειτουργίας αεριοστρόβιλου του εργαστηρίου για την προσομοίωση της βλάβης σε βιομηχανικό αεριοστρόβιλο. Σε αυτή την ανάλυση μας ενδιαφέρει η βλάβη αισθητήρων στη θερμοδυναμική διαγνωστική, όποτε οι μετρήσεις που αναλύονται είναι οι εξής:

- Η παροχή αέρα (W)
- Η πίεση στην έξοδο του συμπιεστή (CDP)
- Η θερμοκρασία στην έξοδο του συμπιεστή (CDT)
- Η παροχή καυσίμου στο Θάλαμο Καύσης (Wf)
- Η θερμοκρασία εξόδου των καυσαερίων (EGT)

Στο Σχήμα 4.12 φαίνονται οι μεταβολές των μετρήσεων που επιβαρύνονται από μία υποβάθμιση τέτοιου είδους.



Σχήμα 4.12: Διάγραμμα μεταβολής συντελεστών παρακολούθησης λόγω σκόνης στο Συμπιεστή.

Για τους αλγόριθμους που αναπτύχθηκαν έγινε μοντελοποίηση υποβάθμισης γραμμικής μεταβολής καθώς είναι η συνηθέστερη υποβάθμιση. Θέτονται λοιπόν οι στόχοι που πρέπει να πετυχαίνει το μοντέλο υποβάθμισης του αεριοστρόβιλου. Οι στόχοι αυτοί αποφασίστηκαν μετά από μελέτη της διάρκειας ζωής του αεριοστρόβιλου. Αποφασίστηκαν βάσει της ασφαλούς λειτουργίας και των αναγκών σε συντήρηση που μπορεί να χρειάζεται η μηχανή.

Οι στόχοι αυτοί απαριθμούνται παρακάτω:

1. Αρχικά, ο αλγόριθμος πρέπει να αναγνωρίζει ότι τα δεδομένα που παρέχονται είναι αποτέλεσμα υποβάθμισης. Χωρίς αυτόν τον έλεγχο κανένας από τους παρακάτω στόχους δε μπορεί να επιτευχθεί. Για αυτό, ο αλγόριθμος αρχικά υπολογίζει την κλίση των 300 πρώτων σημείων με σκοπό να ελέγξει αν τα δεδομένα ακολουθούν κάποια κλίση. Διαλέγονται μόνο τα πρώτα 300 σημεία ώστε αν υπάρχει αλλαγή κλίσης υποβάθμισης (αναλύεται στον τρίτο στόχο). Αν όντως υπάρχει κλίση ακολουθούν οι παρακάτω στόχοι. Οι εντολές που υπολογίζουν την κλίση και αποφασίζουν αν θα ακολουθηθεί η παρακάτω διαδικασία είναι οι εξής:

$$y_i = X(1{:}300, i)\ ; \tag{4.1}$$

Όπου: X(1:300,i) είναι τα πρώτα 300 δεδομένα σε κάθε στήλη i

$$a_i = polyfit(T(1{:}300), y_i, 1)\ ; \tag{4.2}$$

Όπου: polyfit είναι η εντολή δημιουργίας προσεγγιστικού πολυωνύμου βάσει των συντεταγμένων T και Y που αντιπροσωπεύουν το χρόνο και τη μέτρηση i.

$$alpha = [a_i\ ; a_{i+1}\ ; a_{i+2}\ ; \dots]; \tag{4.3}$$

Για την παραγωγή ενός πίνακα – στήλης με τους συντελεστές των πολυωνύμων που παράχθηκαν για κάθε μετρητικό.

$$if\ \max\big(abs(alpha)\big) > 10^{-4} \tag{4.4}$$

Η συνθήκη για να υπάρξει αλλαγή κλίσης

2. Ο αλγόριθμος θα πρέπει να αναγνωρίζει ότι η μηχανή σε βάθος χρόνου υπόκειται σε υποβάθμιση λειτουργίας και αν οι μετρήσεις που δεχθεί ως είσοδο δεν «ταιριάζουν» με τις αρχικές μετρήσεις που ορίστηκαν ως Κατάσταση Υγείας να μην τις κατατάσσει σε ομάδα ή ομάδες ανωμαλιών. Το πρόβλημα αυτό περιορίζεται από τη βασική αρχή διαβάσματος των δεδομένων από κάθε αλγόριθμο. Πιο συγκεκριμένα, όπως ειπώθηκε νωρίτερα, ο αλγόριθμος διαβάζει ένα προς ένα τα νέα σημεία και τα κατατάσσει σε ομάδες. Όσο επιτελεί αυτήν τη λειτουργία, παράλληλα κρατά τις τιμές των τελευταίων 49 δεδομένων, ώστε μαζί με το νέο (50ο) δεδομένο να παράγεται ένα σύνολο δεδομένων στο οποίο υπολογίζεται η κλίση του. Περνώντας παραμετρικά ένα πολυώνυμο πρώτου βαθμού της μορφής: $y = ax + b$ από την 50άδα δεδομένων ο αλγόριθμος υπολογίζει τις τιμές των α και b. Στη συνέχεια, ο αλγόριθμος αφαιρεί τη μοντελοποιημένη υποβάθμιση από το εξεταζόμενο σημείο μέσω της σχέσης:

$$X_{new,i} = X_i - (a * T_i + b) \tag{4.5}$$

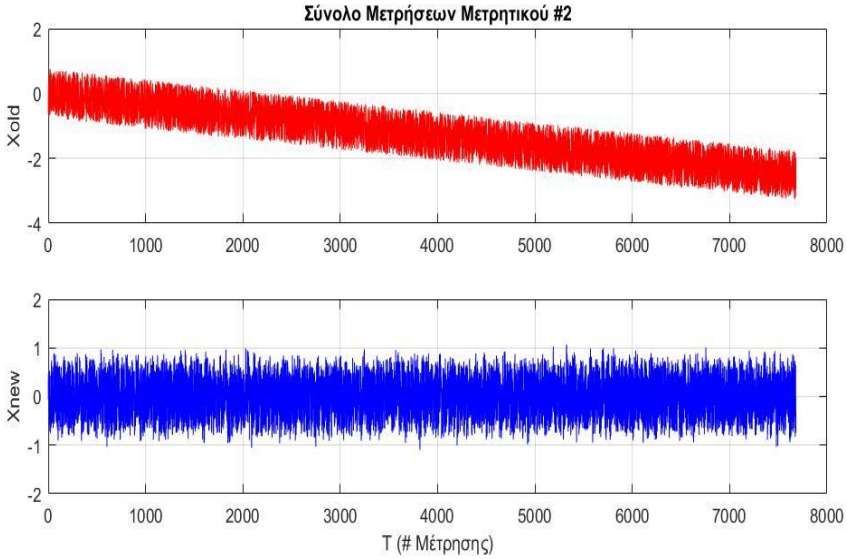
Όπου:

$X_{new,i}$: είναι η νέα θέση i του εξεταζόμενου δεδομένου

X_i: είναι η θέση i του εξεταζόμενου δεδομένου με υποβάθμιση

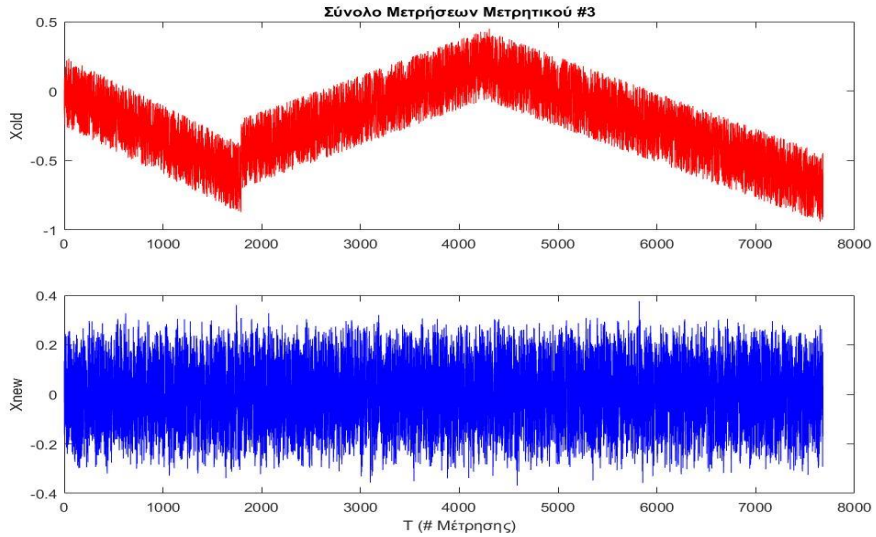
α και b: οι συντελεστές κλίσης και μετατόπισης που χαρακτηρίζει το πολυώνυμο πρώτου βαθμού.

Έτσι, κάθε νέο δεδομένο κατηγοριοποιείται βάσει της τιμής που θα είχε αν δεν υπήρχε υποβάθμιση της μηχανής, στηριζόμενο βέβαια στο κατά πόσο αξιόπιστα τα προηγούμενα 49 σημεία αντιπροσωπεύουν την κλίση και μετατόπιση του νέου (50ου)σημείου. Για δεδομένα χωρίς βλάβη που απλά δείχνουν ότι η μηχανή υπόκειται σε υποβάθμιση το σύνολο δεδομένων ενός μετρητικού έχει την παρακάτω μορφή:



Σχήμα 4.13: Διάγραμμα κλίσης λόγω υποβάθμισης της μηχανής vs. Διορθωμένο διάγραμμα συναρτήσει του χρόνου

3. Όπως αναφέρθηκε και στον πρώτο στόχο, ο αλγόριθμος πρέπει να είναι σε θέση να βρίσκει το Xnew ανεξαρτήτως αλλαγής κλίσης. Για αυτό χρησιμοποιούνται μόνο τα τελευταία 49 και όχι τα τελευταία 1000 για παράδειγμα. Έτσι, ακόμα και αν τα δεδομένα μετά από κάποια χρονική στιγμή αλλάξουν κλίση ή φορά κλίσης, ο αλγόριθμος να καταφέρει να τα αντιμετωπίσει ανεπηρέαστα κατατάσσοντας τα στην κατάσταση υγείας ή ως ανώμαλα σημεία. Στο παρακάτω σχήμα απεικονίζεται ένα τέτοιο παράδειγμα:



Σχήμα 4.14: Διάγραμμα πολλαπλής κλίσης λόγω υποβάθμισης της μηχανής vs. Διορθωμένο διάγραμμα συναρτήσει του χρόνου

Όπως φαίνεται στο Σχήμα 4.14 οι τιμές του μετρητικού #3 έχουν αλλαγή στην τάση κίνησης του. Παρόλα αυτά, μετά τη μετατροπή μέσω των εντολών στο δεύτερο στόχο οι νέες μετρήσεις έχουν κανονική μορφή με μέγιστη διακύμανση ±0.4%.

4. Ο αλγόριθμος πρέπει να ξεχωρίζει την απότομη μεταβολή μίας βλάβης από την απότομη μεταβολή καθαρίσματος της μηχανής. Πιο συγκεκριμένα, αν για παράδειγμα η μηχανή υπόκειται σε βρόμισμα του συμπιεστή της, η παροχή αέρα σταδιακά και σχεδόν γραμμικά θα μειώνεται. Σε αυτή τη φάση, όπως ειπώθηκε στο δεύτερο στόχο, ο αλγόριθμος δεν εντοπίζει ανώμαλη λειτουργία. Αν όμως μετά από μερικούς κύκλους λειτουργίας η μηχανή καθαριστεί και ληφθούν νέα δεδομένα, ο αλγόριθμος θα πρέπει να ξεχωρίσει ότι πρόκειται για καθάρισμα και να κατατάσσει τις νέες μετρήσεις ως υγιή λειτουργία. Αυτό επιτυγχάνεται με παράλληλη παρακολούθηση όλων των μετρητικών οργάνων (παρατηρήσεων) ταυτόχρονα. Πιο συγκεκριμένα η συνθήκη καθαρισμού δίνεται από τη σχέση:

$$if \sum_{j=1,2,3,..} abs\left(X_{new}(i)\right) < 1 \tag{4.6}$$

Όπου:

i: Η εξεταζόμενη χρονική στιγμή της χρονοσειράς

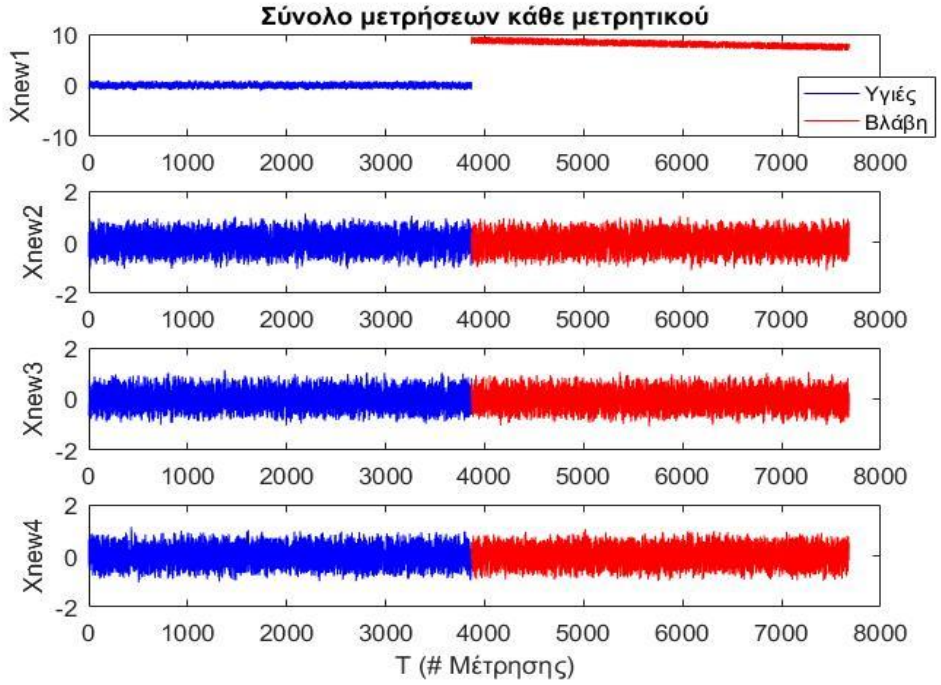
j: Ο αριθμός των μετρητικών οργάνων πάνω στη μηχανή

Ένας άλλος τρόπος που μπορεί να ελέγχεται η συνθήκη καθαρισμού είναι από τη σχέση:

$$if\ abs\big(X_{new}(i,j)\big) < 0.25, \qquad \forall\, j > 1 \tag{4.7}$$

Και οι δύο τρόποι ελέγχουν πόσο κοντά είναι οι μεταβολές των μετρήσεων στο 0. Η πρώτη το κάνει σα σύνολο μετρήσεων, ενώ η δεύτερη ελέγχει κάθε μετρητικό ξεχωριστά. Στην τελική μορφή του αλγορίθμου χρησιμοποιήθηκε η πρώτη έκφραση καθώς είχε συστηματικότερα αποτελέσματα, αλλά και επειδή προέβλεπε βλάψιμο αισθητήρα που ακόμα και μετά το καθάρισμα της μηχανής, οι τιμές που παρείχε ήταν εσφαλμένες.

5. Ο αλγόριθμος πρέπει να λειτουργεί σωστά όταν υπάρχει βλάβη στη μηχανή παράλληλα με υποβάθμιση. Ο τελευταίος αυτός στόχος δεν ήταν δύσκολο να επιτευχθεί αφού μέσω του δεύτερου στόχου προβλέπεται η λειτουργία με βλάβη και υποβάθμιση. Ο αλγόριθμος υπολογίζει την κλίση και μετατόπιση του πρωτοβάθμιου πολυωνύμου που παρεμβάλλει τις μετρήσεις και αφαιρεί την υποβάθμιση όπου θεωρεί κρίσιμο. Οι μετρήσεις με βλάβη όπως θα φανεί και παρακάτω δεν αποτελούν εξαίρεση στον παραπάνω κανόνα γιατί η απότομη μεταβολή μία βλάβης επηρεάζει 5 το πολύ μετρήσεις ώστε να ανιχνευτεί. Έτσι, ο αριθμός των 50 σημείων που χρησιμοποιείται για το προσεγγιστικό πολυώνυμο βοηθά στην εξομάλυνση της απότομης κλίσης που δημιουργούν τα λίγα σημεία που περιγράφουν τη βλάβη. Αυτό έχει ως αποτέλεσμα το προσεγγιστικό πολυώνυμο $1^{ου}$ βαθμού να περνά και πάλι σχεδόν τέλεια από τις μετρήσεις βοηθώντας έτσι τον αλγόριθμο να επαναφέρει τα σημεία με υποβάθμιση σε τέτοιο επίπεδο ώστε να μπορεί να κρίνει αν είναι ανώμαλα ή υγιή. Στο σχήμα 4.15 φαίνονται ειδικά διαμορφωμένες χρονοσειρές τεσσάρων μετρητικών στοιχείων οι οποίες απεικονίζουν υποβάθμιση της μηχανής με τους συντελεστές που παράχθηκαν από τη μοντελοποίηση λειτουργίας με βρόμισμα του συμπιεστή, ενώ παράλληλα, γύρω στις 4000 μετρήσεις παρατηρείται απότομη βλάβη.

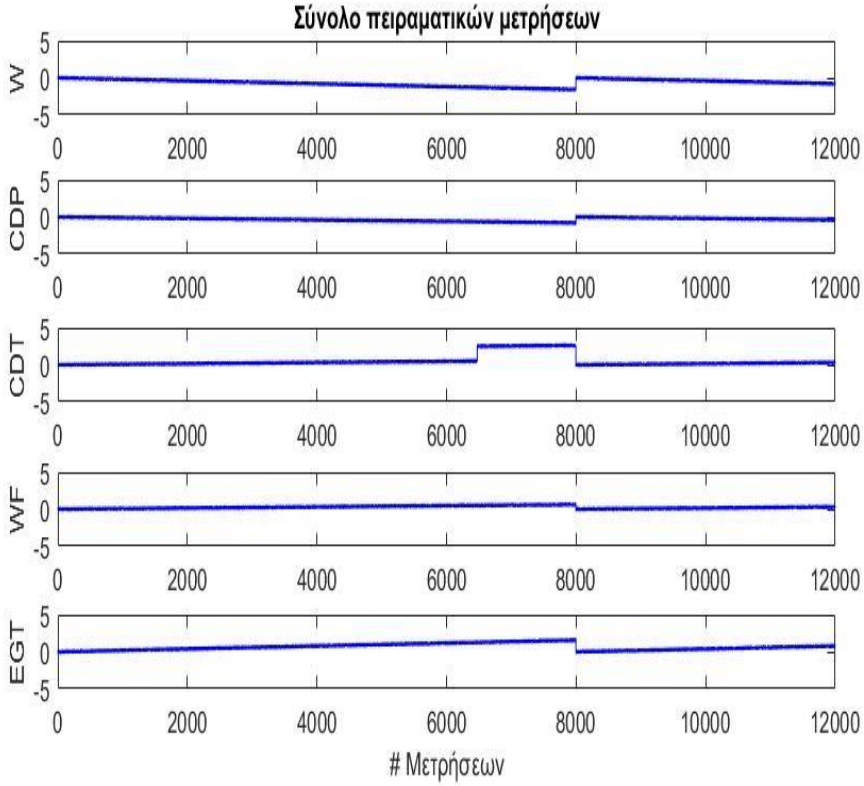


Σχήμα 4.15: Σύνολο διορθωμένων διαγραμμάτων λόγω υποβάθμισης της μηχανής αφού εφαρμοσθεί πρόγνωση

Όπως φαίνεται ο αλγόριθμος «αγνοεί» την υποβάθμιση στις μετρήσεις 2, 3 και 4 καθώς το νέο Χ για αυτά τα μετρητικά προκύπτει με απόκλιση της τάξεως του 1% μεταβολή, επιτυχημένα. Από την άλλη πλευρά, το μετρητικό #1, αυτό δηλαδή που ορίστηκε να έχει βλάβη, μετά τη μέτρηση 4000 φαίνεται να έχει μεταβολή περίπου 10%. Η βλάβη λοιπόν εντοπίστηκε από τον αλγόριθμο και οι μετρήσεις από 4000 και μετά προστέθηκαν στο μητρώο βλάβης.

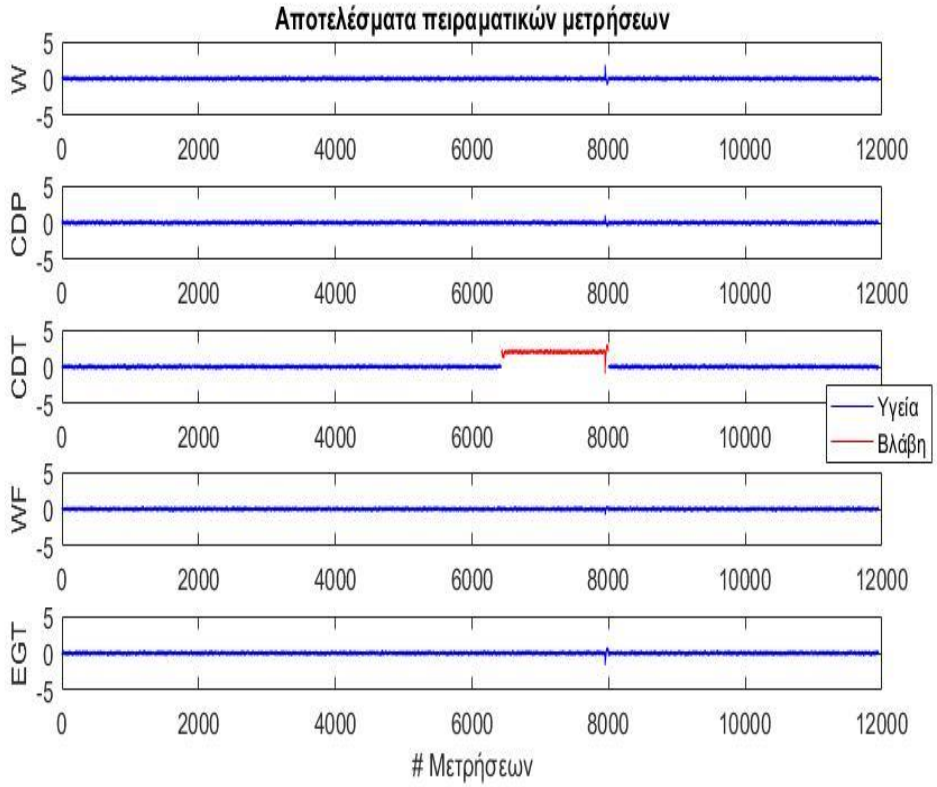
6. Τέλος, ο αλγόριθμος πρέπει να λειτουργεί ορθά όταν υπάρχει επιδιόρθωση εσφαλμένης λειτουργίας με παράλληλη υποβάθμιση του κινητήρα όπως για παράδειγμα, μέσω σκονίσματος στο συμπιεστή. Εφόσον μέσω του στόχου 5 ο αλγόριθμος είναι ικανός να διαγνώσει λειτουργία εσφαλμένη λειτουργία, είναι σημαντικό το αρμόδιο προσωπικό να ενημερωθεί και να δράσει εγκαίρως ώστε να αποκατασταθεί η κανονική λειτουργία της μηχανής. Έτσι λοιπόν, όταν ένας αεριοστρόβιλος διαγνωστεί με βλάβη, η λειτουργία του σταματά, όπως και οι μετρήσεις των διαφόρων μετρητικών στοιχείων πάνω στον αεριοστρόβιλο. Η μηχανή καθαρίζεται και επιδιορθώνεται για όσο χρονικό διάστημα χρειαστεί και μετά επιστρέφει στην κανονική λειτουργία. Τα μετρητικά συστήματα σε αυτήν την περίπτωση επιστρέφουν στην προκαθορισμένη τους δουλειά από εκεί που σταμάτησαν, δηλαδή την εσφαλμένη λειτουργία με αποτέλεσμα να υπάρχει απότομη μεταβολή στις χρονοσειρές που μελετώνται. Το πρόβλημα αυτό αντιμετωπίζεται εύκολα όπως παρουσιάστηκε και παραπάνω. Η πολυπλοκότητα του στόχου αυτού οφείλεται στην συνύπαρξη βλάβης με υποβάθμισης. Χρησιμοποιώντας το σύνολο των τεσσάρων χρονοσειρών που μελετήθηκε παραπάνω για τη μοντελοποίηση δεδομένων που

αντιπροσωπεύουν τη μεταβολή που περιεγράφηκε παραπάνω παράγεται το παρακάτω σχήμα:



Σχήμα 4.16: Σύνολο διαγραμμάτων κλίσης λόγω υποβάθμισης της μηχανής συναρτήσει του χρόνου

Όπως φαίνεται στις παραπάνω μετρήσεις το μετρητικό CDT καταγράφει ότι ίσως υπάρχει βλάβη, μέσω απότομης μεταβολής περίπου στη χρονική μέτρηση 6300. Παράλληλα, όλα τα υπόλοιπα μετρητικά μέχρι εκείνο το σημείο, κάνουν διακριτό ότι ο αεριοστρόβιλος που μελετάται υπόκειται σε υποβάθμιση μέσω της σταθερής κλίσης που επιτηδευμένα επιβλήθηκε όπως ακριβώς και στα παραπάνω διαγράμματα. Στη μέτρηση 8000 περίπου, η βλάβη που επιβλήθηκε στη συνιστώσα που ελέγχει το μετρητικό CDT διορθώνεται, μαζί με αυτήν, διορθώνεται και η απόκλιση λόγω υποβάθμισης της μηχανής σε όλες τις συνιστώσες. Έτσι, ο στόχος που περιεγράφηκε μοντελοποιήθηκε και μένει μόνο να παρουσιαστούν τα αντίστοιχα αποτελέσματα από τον αλγόριθμο, παρακάτω:



Σχήμα 4.17: Σύνολο διορθωμένων διαγραμμάτων κλίσης λόγω υποβάθμισης της μηχανής συναρτήσει του χρόνου

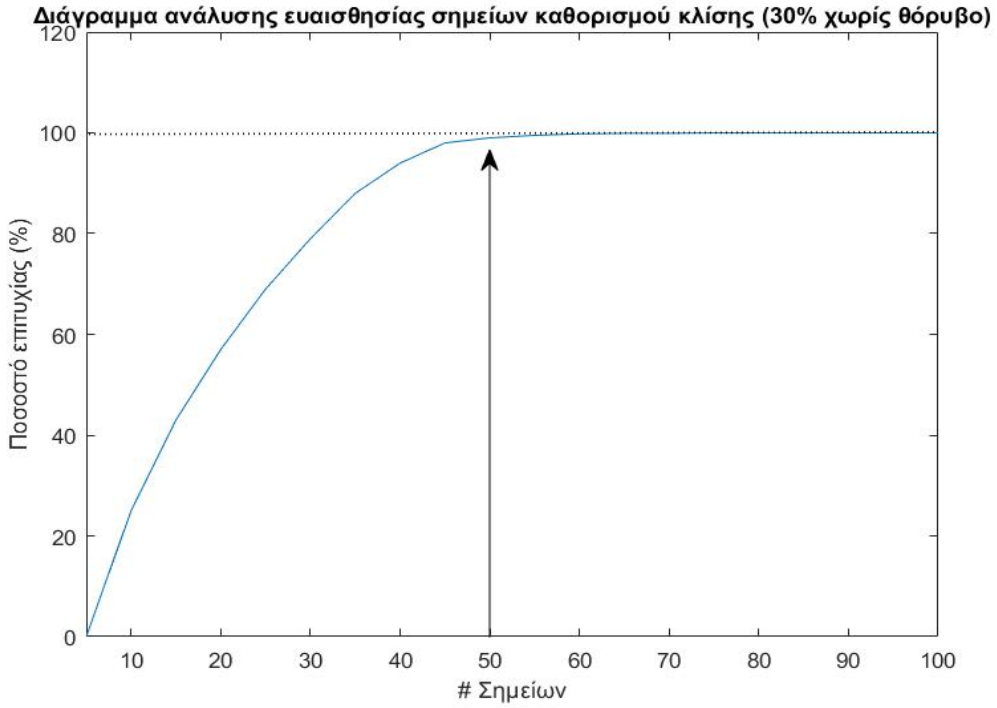
Στο παραπάνω σχήμα, φαίνεται η κατανομή των δεδομένων σε γειτονιές υγείας ή σφάλματος αφού πρώτα προηγηθεί επεξεργασία αυτών. Όπως και στο στόχο 2, η κλίση λόγω υποβάθμισης της μηχανής αφαιρείται από όλες τις συνιστώσες για την εξαγωγή πιο έγκυρων αποτελεσμάτων. Ορίζοντας βήμα υπολογισμού της κλίσης των δεδομένων, αρκετά μεγάλο ώστε να μην επηρεάζεται από την τυχαία κατανομή των μοντελοποιημένων μετρήσεων αλλά και αρκετά μικρό ώστε να ανιχνεύονται πιο έγκαιρα οι απότομες μεταβολές, ο αλγόριθμος καταφέρνει να διατηρεί τη μεταβολή των συνιστωσών που εξετάζονται σε απόκλιση $\pm 1\%$ από τη μέση τιμή. Παράλληλα, θα πρέπει να ικανοποιούνται οι στόχοι 1 και 5, οι στόχοι δηλαδή της διάγνωσης ανώμαλης λειτουργίας. Πράγματι, για τις μετρήσεις 6300 έως 8000 περίπου παρατηρείται ότι η μέτρηση 3 έχει ανώμαλη λειτουργία και ότι σωστά κατατάσσονται αυτά τα δεδομένα για όλες τις μετρήσεις (1 έως 5) ως εσφαλμένη λειτουργία. Ο αλγόριθμος είναι φτιαγμένος με τέτοιο τρόπο ώστε οι μετρήσεις της εσφαλμένης λειτουργίας να μην επηρεάζονται από την κλίση των υγειών σημείων δίνοντας ξεκάθαρη διαφοροποίηση μεταξύ των δύο γειτονικών συμπλεγμάτων. Τέλος, οι μετρήσεις μετά το $8000^{ο}$ σημείο στον οριζόντιο άξονα φαίνεται να κατατάσσονται στην υγιή λειτουργία, όπως πρέπει δηλαδή, καθώς η μηχανή μετά από αυτήν τη μέτρηση δέχεται επιδιόρθωση και της εσφαλμένης λειτουργίας και της βαθμιαίας υποβάθμισης.

Για την ανάλυση αυτή χρησιμοποιήθηκαν όλοι οι παραπάνω αλγόριθμοι επαυξημένοι όμως από συνθήκες υποβάθμισης και γενικής κλίσης των δεδομένων. Όπως φάνηκε και στο κεφάλαιο υποβάθμισης της μηχανής, είναι σημαντικό να ελεγχθούν οι παρακάτω στόχοι:

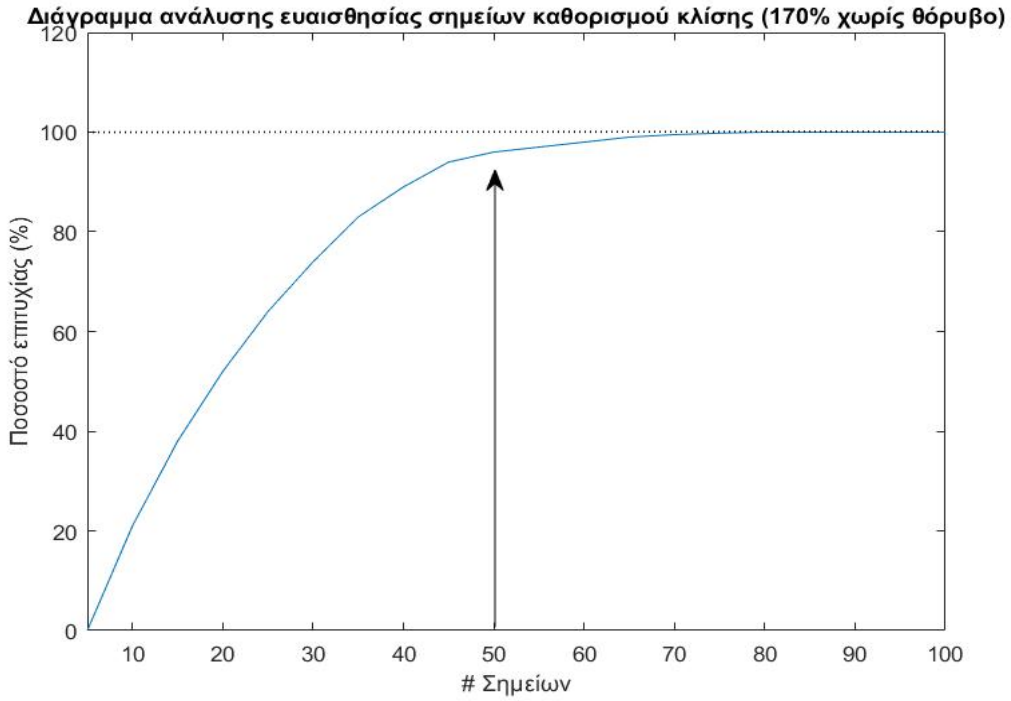
1. Αν υπάρχει υποβάθμιση
2. Τι είδους υποβάθμιση είναι
3. Αν υπάρχει υποβάθμιση παράλληλα με εσφαλμένη λειτουργία
4. Αν σε κάποια χρονική στιγμή διορθώθηκε η υποβάθμιση αυτή

Υποβάθμιση σε μία μηχανή μπορεί να υπάρχει για ποικίλους λόγους και να εκφραστεί στα εξεταζόμενα δεδομένα με διάφορους τρόπους. Για το κομμάτι της μελέτης που αφιερώθηκε στην υποβάθμιση αυτής της εργασίας υπήρξαν μερικές παραδοχές που λήφθηκαν. Πρώτα από όλα, θεωρήθηκε ότι η υποβάθμιση ακολουθεί γραμμική μεταβολή. Ο τρόπος εξέτασης των εισαγόμενων σημείων είναι αρκετά ενδελεχής ώστε να μπορεί να χαρακτηρίσει σωστά πλήθος δεδομένων ακόμα και αν αυτά υπόκεινται σε άλλου είδους υποβάθμιση. Στη συνέχεια, θεωρήθηκε ότι για να υπάρχουν πιο αξιόπιστοι συντελεστές πρωτοβάθμιου πολυωνύμου που περιγράφει τη γραμμική υποβάθμιση χρειάζεται δείγμα τουλάχιστον 50 σημείων πριν από το εξεταζόμενο σημείο. Έτσι προβλέπεται η θέση του εξεταζόμενου σημείου ακριβέστερα και γίνεται ορθότερη διάγνωση.

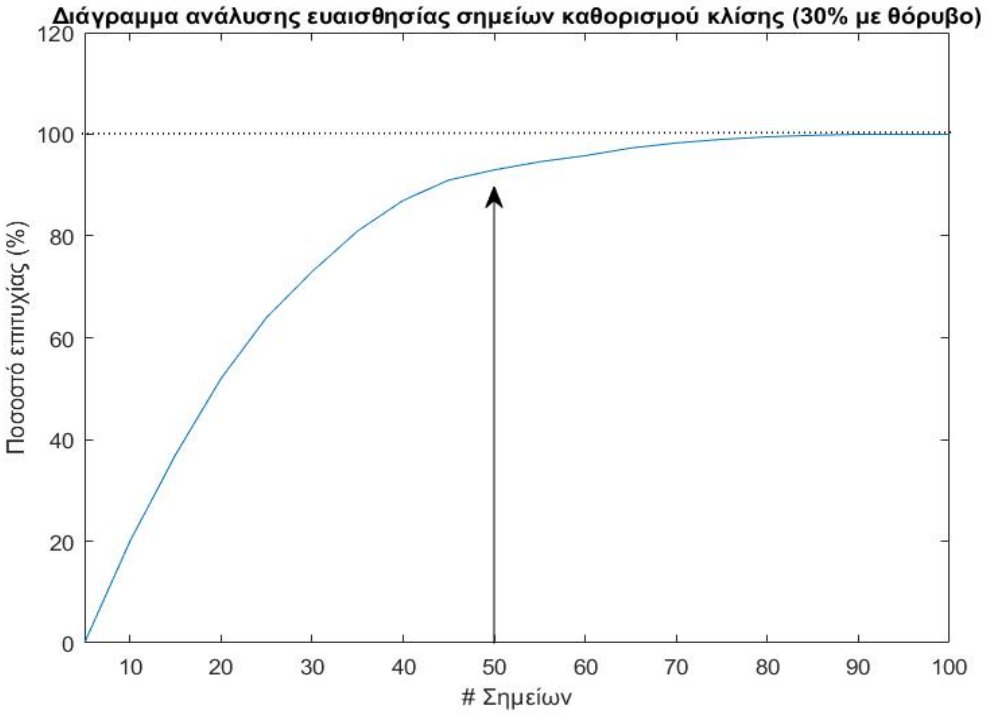
Γιατί όμως χρησιμοποιήθηκαν 50 σημεία και όχι 40 ή 80? Η απάντηση στην ερώτηση αυτή έρχεται από την ανάλυση ευαισθησίας που έγινε για την επιλογή αριθμού επαρκών σημείων για τον υπολογισμό της κλίσης. Συγκεκριμένα, για ένα δεδομένο σετ σημείων που ο χρήστης γνωρίζει ποιο πρέπει να είναι το διαγνωστικό αποτέλεσμα, τίθεται ο αριθμός σημείων κλίσης ίσος με 100. Με τόσα σημεία παρατηρείται ότι η διαγνωστική ικανότητα έχει 100% επιτυχία. Οπότε στη συνέχεια, μειώνοντας τον αριθμό υπολογισμού κλίσης και συγκρίνοντας μέσω του Συντελεστή ομοιότητας τα αποτελέσματα, μπορούμε να έχουμε μία καλύτερη εικόνα για τον ιδανικό αριθμό χρήσιμων σημείων. Η ανάλυση αυτή έγινε πάνω στα δεδομένα που χρησιμοποιήθηκαν στο δεύτερο μοντέλο της Ενότητας 5.1 που περιγράφει τη διαγνωστική ικανότητα των μεθόδων. Εκεί, η υποβάθμιση ελέγχεται σε δύο στάδια, το ήπιο και το απότομο 30% και 170% αντίστοιχα της υποκείμενης υποβάθμισης.



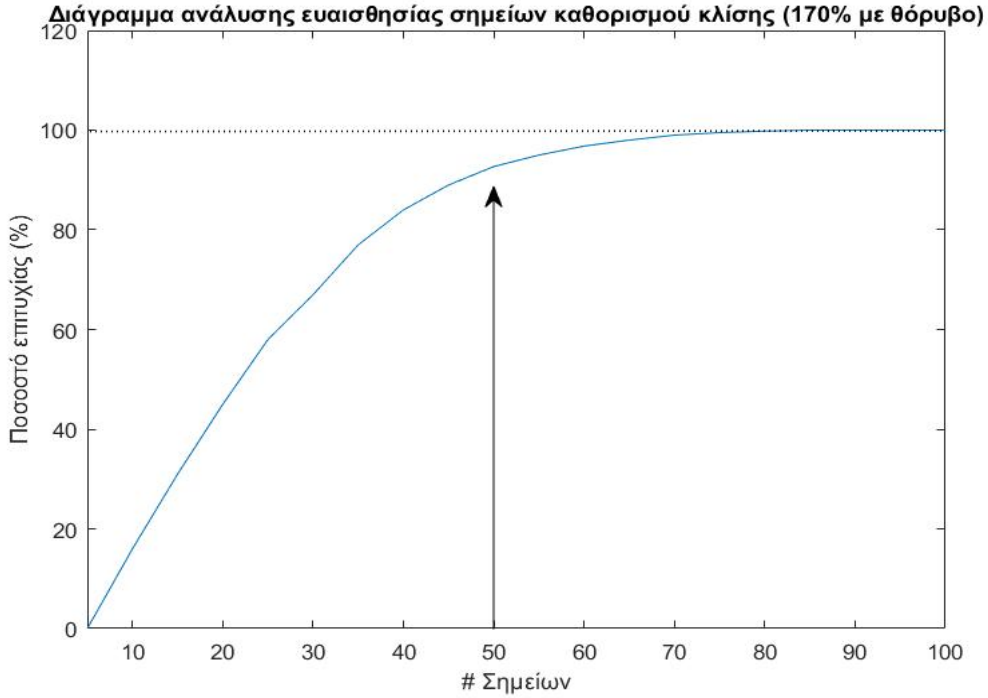
Σχήμα 4.18: Διάγραμμα ανάλυσης ευαισθησίας σημείων καθαρισμού ήπιας κλίσης χωρίς θόρυβο



Σχήμα 4.19: Διάγραμμα ανάλυσης ευαισθησίας σημείων καθαρισμού απότομης κλίσης χωρίς θόρυβο



Σχήμα 4.20: Διάγραμμα ανάλυσης ευαισθησίας σημείων καθαρισμού ήπιας κλίσης με θόρυβο



Σχήμα 4.21: Διάγραμμα ανάλυσης ευαισθησίας σημείων καθαρισμού απότομης κλίσης με θόρυβο

Ξεκινώντας από τα 5 σημεία πρόβλεψης και πηγαίνοντας μέχρι τα 100 σημεία παρατηρείται ότι το ποσοστό επιτυχίας αυξάνεται, φαινόμενο λογικό. Σε κάθε ένα από τα διαγράμματα υπάρχει ένα βέλος δείχνοντας ότι στα 50 σημεία έχει σημειωθεί αρκετή επάρκεια, έχοντας μεγάλο ποσοστό επιτυχίας, σχεδόν όσο τα 100 σημεία, χωρίς όμως να περνά πολύτιμος χρόνος δειγματοληψίας οδηγώντας σε αλλοίωση της διαγνωστικής ικανότητας. Τα ποσοστά επιτυχίας για κάθε μία από τις περιπτώσεις και για 50 σημεία πρόβλεψης κυμαίνονται από 91% έως 96% ορίζοντας αρκετά υψηλή επιτυχία. Επίσης αξιοσημείωτη είναι η κλίση των διαγραμμάτων ανάλογα με την περίπτωση καθώς για ήπια κλίση οι αλγόριθμοι πετυχαίνουν μεγαλύτερη ακρίβεια με λιγότερα σημεία σε σχέση με την απότομη κλίση. Αυτό οφείλεται στο γεγονός ότι η μικρότερη κλίση χρειάζεται μικρότερη προσπάθεια να εξομαλυνθεί, επειδή απέχει λιγότερο από τον άξονα αναφοράς. Τελευταία και μικρότερη διαφορά μεταξύ των διαγραμμάτων είναι η ομαλή μετάβαση από τα χαμηλά στα υψηλά επίπεδα επιτυχίας χωρίς «γόνατα» στην καμπύλη που υπάρχει στις μετρήσεις χωρίς θόρυβο αλλά απουσιάζει σε αυτές με θόρυβο. Είναι λογικό το φαινόμενο αυτό καθώς με τυχαίες μετρήσεις η πρόβλεψη της κλίσης γίνεται πιο ασταθής και το νέο εξεταζόμενο σημείο μπορεί να κατανεμηθεί άστοχα κάπου που δεν έπρεπε λόγω της μέσης κλίσης που παρέχει ο τυχαίος παράγοντας των προηγούμενων σημείων. Έτσι τελικά, συμπεραίνεται ότι τα επαρκή σημεία πρόβλεψης της κλίσης είναι 50 πρώτα για υπάρχει μεγάλη ακρίβεια και έπειτα για να μη χάνεται διαγνωστική πληροφορία.

Στη συνέχεια, λήφθηκε υπόψη η περίπτωση καθαρισμού της μηχανής μετά από κάποιο διάστημα λειτουργίας με υποβάθμιση. Λόγω της περίπτωσης αυτής αναπτύχθηκε επιπλέον έλεγχος των κλίσεων όλων των μεταβλητών μεγεθών σε κάθε επανάληψη ώστε να μη θεωρηθεί ο καθαρισμός της μηχανής ως ένα είδος βλάβης μέσω των απότομων μεταβολών των εξεταζόμενων μεγεθών.

Για το στόχο 1, πριν τη μεγάλη επανάληψη λήψης νέου εξεταζόμενου σημείου για να χαρακτηριστεί από κάποια βλάβη, προστίθεται ένας έλεγχος της κλίσης των πρώτων 50 δεδομένων κάθε μετρούμενου μεγέθους. Ο έλεγχος της κλίσης γίνεται μέσω της εντολής:

$$\alpha_\iota = polyfit(T(1{:}50)', XX(1{:}50, i), 1); \tag{4.8}$$

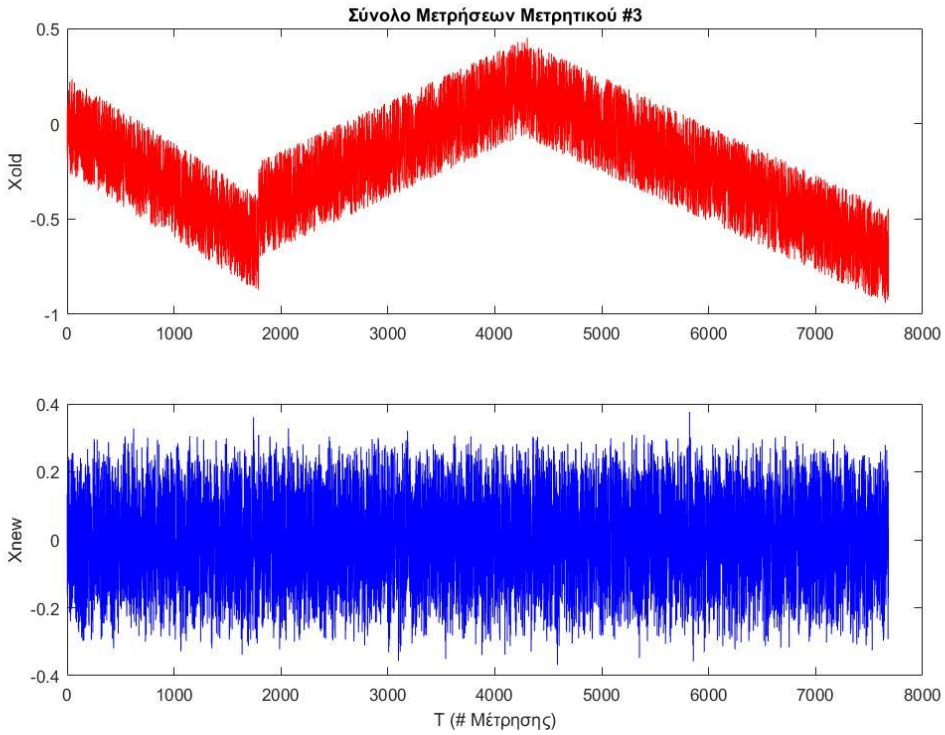
Και

```
if max(abs(alphas))>1e-4
    Anum=50; %arithmos simeiwn pou epireazoun tin klisi
    i=0;
    for k=1:length(X)-Anum
        i=i+1;
        for j=1:4
            y1=X(i:i+Anum-1,j); %Anum proigoumena simeia
            a=polyfit(T(i:i+Anum-1)',y1,1); %angle
            %afairesi degradation
            Xn(k,j)=X(k+Anum,j)-(T(i+Anum)*a(1)+a(2))';
        end
```

Σχήμα 4.22: Κομμάτι κώδικα αφαίρεσης υποβάθμισης από το αρχικό δείγμα δεδομένων

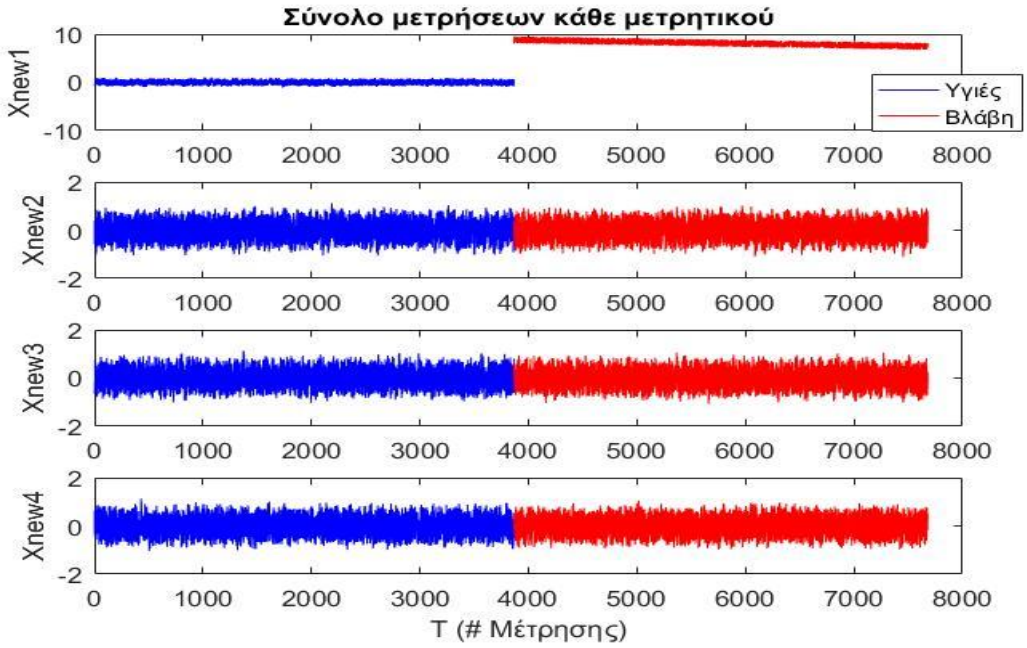
Έτσι σε περίπτωση υποβάθμισης, ελέγχεται ένα μέρος του δείγματος ώστε να αφαιρεθεί πριν τη διάγνωση. Η αφαίρεση γίνεται μέσω των εντολών που φαίνονται παρακάτω, όπου η τελική κλίση του δείγματος γίνεται μηδέν.

Ο δεύτερος στόχος αφορά την αλλαγή κλίσης της υποβάθμισης αλλά και τον τύπο υποβάθμισης. Ο τύπος υποβάθμισης δε θα αναλυθεί περαιτέρω λόγω των παραδοχών που ορίσθηκαν. Η αλλαγή κλίσης όμως, ικανοποιείται από την παραπάνω εικόνα. Αυτή η σειρά εντολών μπορεί ανά 50 σημεία να διακρίνει την αλλαγή κλίσης και να εφαρμόσει αφαίρεση υποβάθμισης καταλήγοντας και πάλι σε δεδομένα μηδενική κλίσης όπως φαίνεται και παρακάτω.



Σχήμα 4.23: Αφαίρεση σύνθετης υποβάθμισης μεταβλητής

Ο τρίτος στόχος χρησιμοποιεί παρόμοια λογική με τους δύο προηγούμενους. Το κρίσιμο σημείο εδώ είναι η απότομη κλίση που δημιουργείται λόγω της βλάβης. Η κλίση αυτή εξαλείφεται στα επόμενα σημεία καθώς ανά 50 σημεία γίνεται νέο έλεγχος. Στη χειρότερη των περιπτώσεων μπορεί να καθυστερήσει η διάγνωση για χρονικό διάστημα ίσο με 49 μετρήσεις (ανάλογα με την ταχύτητα δειγματοληψίας). Η υστέρηση αυτή δεν είναι σημαντική συνυπολογίζοντας τον όγκο των δεδομένων. Για τον καθορισμό της ακριβούς θέσης των εξεταζόμενων σημείων αφαιρώντας την υποβάθμιση είναι πολύ σημαντικός ο ρόλος της δεύτερης μεταβλητής που παράγει η εντολή polyfit Όπως αναφέρθηκε και στο κεφάλαιο υποβαθμισμένης λειτουργίας, ο συντελεστής b υπολογίζει την απόσταση της ευθείας (που προσεγγίζει τα δεδομένα προς εξέταση) από την αρχή των αξόνων. Μέσω του συντελεστή αυτού δίνεται η αίσθηση της κλίμακας της βλάβης στις χρονοσειρές. Παρακάτω φαίνεται η κλίμακα αυτή σε δεδομένα που ήδη έχει αφαιρεθεί η υποβάθμιση και έχουν διαγνωσθεί ως λειτουργία βλάβης.



Σχήμα 4.24: Σύνολο αποτελεσμάτων μετά από αφαίρεση υποβάθμισης και πρόγνωσης βλαβών

Τέλος, για τον τέταρτο στόχο προστέθηκε ένας ακόμα έλεγχος στους αλγόριθμους των μεθόδων. Η διαφορά αυτού του ελέγχου από τους υπόλοιπους είναι ότι εξετάζει όλες τις δεδομένες χρονοσειρές ταυτόχρονα. Ο έλεγχος αυτός φαίνεται στο παρακάτω σχήμα.

```
if sum(abs(X(k,:)))<1 %se periptwsi katharismou mixanis
    IX(k,1)={'Normal'};
    IA(k,1)=1;
```

Σχήμα 4.25: Κώδικας συνθήκης καθαρισμού της μηχανής

Πρακτικά εξετάζει αν το άθροισμα των απόλυτων τιμών των μετρήσεων σε μία χρονική στιγμή k είναι μικρότερο της απόκλισης 1% τότε τα δεδομένα ανήκουν στην υγιή λειτουργεία καθώς υπάρχει περίπτωση καθαρισμού της μηχανής. Ένα άλλο είδος ελέγχου που δοκιμάστηκε, εξέταζε κάθε μία από τις εξεταζόμενες μεταβλητές ξεχωριστά ελέγχοντας αν η τιμή τους ξεπερνά το 0.2% ανά χρονική στιγμή. Η εντολή είχε την παρακάτω μορφή:

$$if\ abs\big(X(k,1)\big) < 0.2\ \&\&\ abs\big(X(k,2)\big) < 0.2\ \&\&\ abs\big(X(k,3)\big) < 0.2\ \&\&\ abs\big(X(k,4)\big) < 0.2 \tag{4.9}$$

Η διαλογή μεταξύ των δύο ελέγχων έχει να κάνει με το είδος των δεδομένων και μόνο. Στη συγκεκριμένη μελέτη ο δεύτερος έλεγχος παρείχε ασταθή αποτελέσματα όποτε δε χρησιμοποιήθηκε.

4.3.3 Σύγκριση αποτελεσμάτων

Τα αποτελέσματα των μεθόδων στα διάφορα τεστ που υποβλήθηκαν είχαν κοινά αποτελέσματα. Στον παρακάτω Πίνακα συγκρίνονται οι μέθοδοι, οι συναρτήσεις απόστασης, το υπολογιστικό κόστος και οι διάφοροι στόχοι για την αφαίρεση υποβάθμισης που αναλύθηκαν παραπάνω:

Μέθοδος	Μετρική	Στόχος 1	Στόχος 2	Στόχος 3	Στόχος 4	Στόχος 5	Run time (sec)
DBSCAN	Ευκλείδεια	√	√	√	√	√	56
	CCD	√	√	√	√	√	148
K-means	Ευκλείδεια	√	√	√	√	√	22
	CCD	√	√	√	√	X	44
AHC	Ευκλείδεια	√	√	√	X	X	49
	CCD	√	√	√	X	X	55

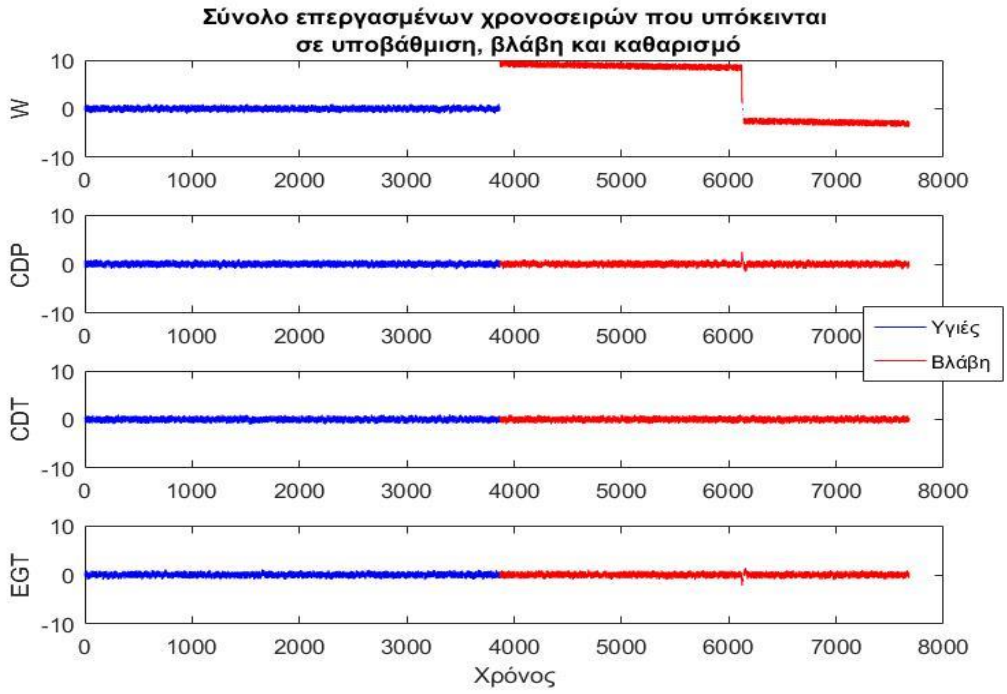
Σχήμα 4.26: Πίνακας επιτυχίας-αποτυχίας των μεθόδων ανάλογα με το στόχο

Ο Πίνακας αυτός βασίζεται σε αποτελέσματα επιτυχίας των μεθόδων στους εξής στόχους:

1. Δεδομένα με μία βλάβη
2. Δεδομένα με πολλαπλές βλάβες
3. Δεδομένα με υποβάθμιση
4. Δεδομένα με υποβάθμιση και καθαρισμό μηχανής
5. Δεδομένα με υποβάθμιση, βλάβη και καθαρισμό

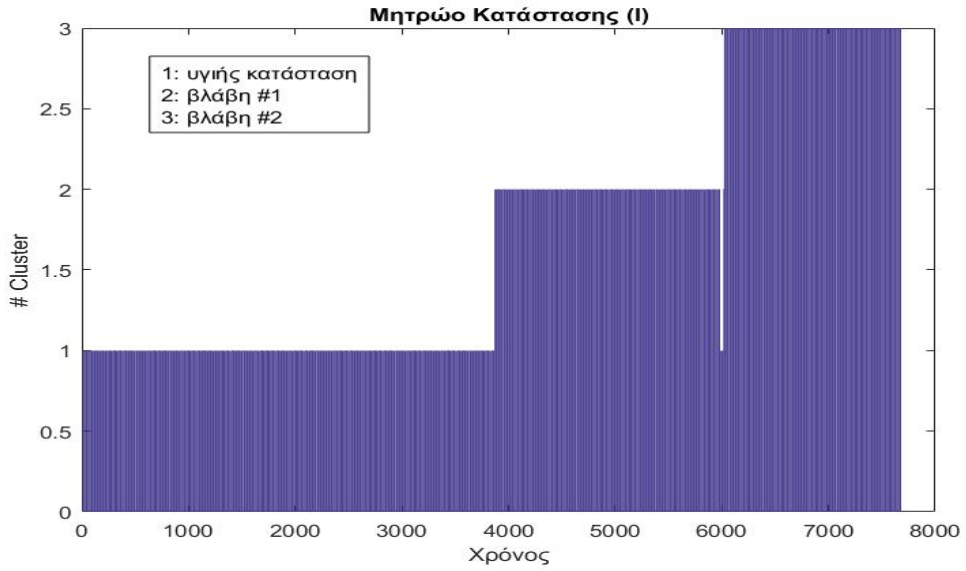
Παρατηρώντας τους στόχους που επιτεύχθηκαν φαίνεται ότι οι μέθοδοι διαφέρουν μεταξύ τους αλλά και ως προς τη συνάρτηση απόστασης που εφαρμόζεται. Οι πιο πετυχημένες φαίνεται να είναι οι DBSCAN και K-means με ευκλείδεια απόσταση καθώς πετυχαίνουν όλους τους στόχους σε σχετικά χαμηλό χρονικό διάστημα. Ο πιο σύνθετος στόχος ήταν ο στόχος 5 και φαίνεται ότι τουλάχιστον μία εκδοχή της κάθε μεθόδου έβγαλε σωστά αποτελέσματα.

Πέρα από τις επιτυχίες όμως είναι σημαντικό να αναφερθούν και οι αποτυχίες. Η πιο «αποτυχημένη» μέθοδος βάσει του παραπάνω Πίνακα ήταν η AHC. Δυστυχώς, η μέθοδος αυτή είχε διάφορες άστοχες προβλέψεις κυρίως σε δεδομένα που σταδιακά απέκλιναν από τη μέση τιμή και μετά επανέρχονταν. Η αδυναμία αυτή φαίνεται και από τους στόχους που δεν κατάφερε να πετύχει η μέθοδος. Και ο καθαρισμός μηχανής και η επιδιόρθωση βλάβης απαιτεί απότομο άλμα στη μέση τιμή της χρονοσειράς. Το σημείο αυτό ήταν που αδυνατούσε να προβλέψει η μέθοδος και αντί να κατανέμει τα σημεία μετά από αυτό στο υγιές Cluster, δημιουργούσε νέο Cluster βλάβης και εκχωρούσε κάθε νέο υγιές σημείο σε αυτό. Ενδεικτικά ένα διάγραμμα που περιγράφει τον 5^{ο} στόχο:



Σχήμα 4.27: Αποτυχία πρόγνωσης μετά από καθαρισμό της μηχανής για τη μέθοδο AHC

Στο παραπάνω Σχήμα 4.27 φαίνεται το αποτέλεσμα της AHC για δείγμα τεσσάρων υποβαθμισμένων μεταβλητών εκ των οποίων η μία υπόκειται και σε βλάβη (η παροχή W). Η επεξεργασία των μετρήσεων φαίνεται ότι έχει γίνει σωστά καθώς καμία χρονοσειρά δεν έχει κλίση. Έπειτα, η μέθοδος αναγνώρισε σωστά ότι περίπου στη μέτρηση 4000 υπάρχει βλάβη στη μεταβλητή W. Το πρόβλημα εμφανίζεται όταν γίνεται η επιδιόρθωση της μηχανής, κοντά στη μέτρηση 6000 όπου όλες οι μεταβλητές επανέρχονται στη μηδενική απόκλιση. Εκεί, για την παροχή W, φαίνεται ότι υπάρχουν μερικά υγιή σημεία αλλά η λειτουργία χαρακτηρίζεται εκ νέου σε βλαβερή. Διερευνώντας παραπάνω τα δεδομένα αυτά έχει ενδιαφέρον να μελετηθεί η κατανομή του μητρώου κατάσταση κάθε σημείου *I* το οποίο παρουσιάζεται παρακάτω:



Σχήμα 4.28: Κατανομή σημείων ανάλογα με το Cluster που ανήκουν για τη μέθοδο AHC

Παραπάνω παρατηρείται η κατανομή των χρονικών μετρήσεων ανάλογα με το Cluster στο οποίο ανήκουν. Τα σημεία από 1 μέχρι περίπου 6000 έχουν κατανεμηθεί σωστά στην υγιή και εσφαλμένη λειτουργία αντίστοιχα. Το πρόβλημα εμφανίζεται στις μετρήσεις μετά την 6000η όπου τα σημεία εκχωρούνται στο Cluster 3. Σύμφωνα με τον τρόπο λειτουργίας της AHC μεθόδου, η ομαδοποίηση των σημείων μετά τη μέτρηση 6000 μπορεί να είναι τόσο πυκνή που να θεωρεί η μέθοδος ότι δεν είναι όμοια αυτής της υγιούς λειτουργίας. Όπως και να έχει, αποτελεί ατόπημα της μεθόδου και μόνο με την κριτική ικανότητα κάποιου επιβλέποντα θα μπορούσε να αποφευχθεί αυτή η λάθος ομαδοποίηση. Για το λόγο αυτό η μέθοδος αυτή υστερεί από τις άλλες δύο.
Τέλος, αξίζει να σημειωθεί ότι όλες οι μέθοδοι είχαν κάποια υστέρηση στην CCD εκδοχή τους. Αυτό συνέβη επειδή η λειτουργία της εκδοχής αυτής βασίζεται στην παραγωγή μητρώων αποστάσεων ανάλογων του πλήθους σημείων που συγκρίνονται σε κάθε μία από τις επαναλήψεις με αποτέλεσμα να αυξάνεται το υπολογιστικό κόστος.

4.3.4 Cluster καθαρισμού

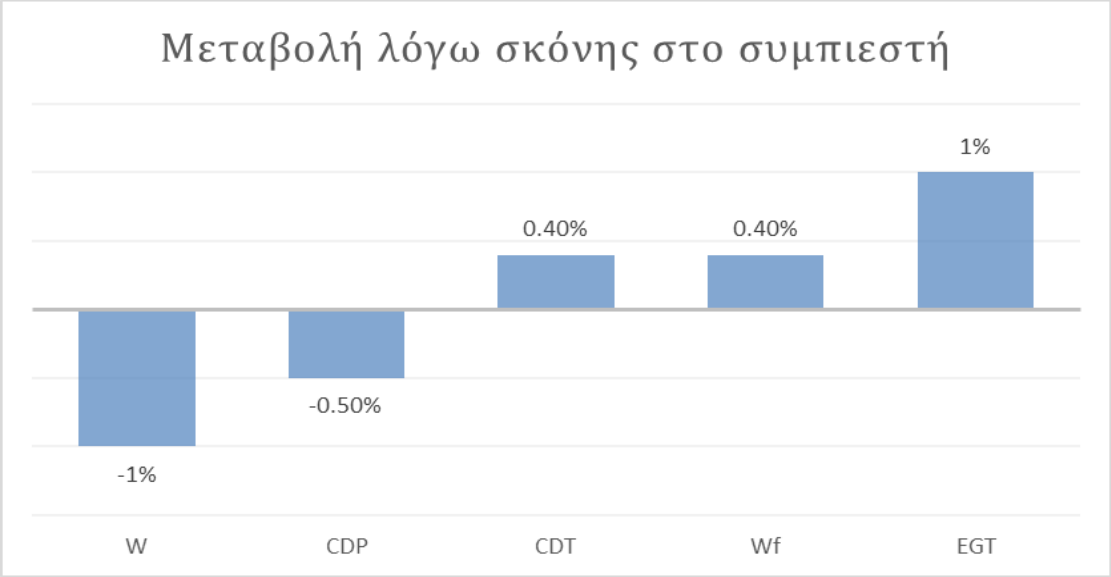
Σε συνδυασμό με τον αλγόριθμο που δημιουργήθηκε για τη διαχείριση της υποβάθμισης, παράχθηκε και μία ακόμα υπογραφή για να αντιμετωπίσει το πρόβλημα μεταβολής λόγω καθαρισμού της μηχανής. Όπως είναι λογικό, στο χρονικό διάστημα εξέτασης μίας μηχανής μπορεί να υπάρχει μία τάση υποβάθμισης των μετρήσεων όπως αναφέρθηκε και παραπάνω. Αυτή η υποβάθμιση αφαιρείται μέσω offline ή online καθαρισμού της μηχανής. Η μεταβολή στις μετρήσεις που προκύπτει λόγω του καθαρισμού, και ιδίως του offline, προβλέπεται ότι θα επηρεάσει τον αλγόριθμο στο να διαγνώσει βλάβη ενώ επικρατεί υγιής λειτουργία (false alarm). Για να περιοριστεί αυτό το φαινόμενο όσο το

δυνατόν περισσότερο, δημιουργήθηκε μία υπογραφή που περιγράφει τη μεταβολή λόγω καθαρισμού. Η υπογραφή αυτή, όπως είναι λογικό έχει αντίθετες τιμές από αυτές που προβλέπει το μοντέλο υποβάθμισης λόγω βρομίσματος του συμπιεστή και οι αναλυτικές τιμές του φαίνονται στο παρακάτω σχήμα:

Σχήμα 4.29: Υπογραφή λόγω καθαρισμού βρομίσματος του συμπιεστή

4.4 Εφαρμογή σε προσομοιωμένα δεδομένα

Πέρα από τις μετρήσεις deltas που διαγνώστηκαν και τροποιήθηκαν για να ελεγχθεί το ποσοστό επιτυχίας των στόχων που τέθηκαν παραπάνω, έγινε ανάλυση και σε νέα μοντελοποιημένα δεδομένα. Με τη βοήθεια των υπολογιστικών φύλλων Excel, μοντελοποιήθηκε η λειτουργία αεριοστρόβιλου που χρησιμοποιείται στο εργοστάσιο αλουμινίου της Αττικής για παραγωγή αλουμινίου και ατμού που παρέχει ενέργεια στο ηλεκτρικό δίκτυο. Παράχθηκε μοντέλο παρακολούθησης των τεσσάρων μεγεθών που είδαμε παραπάνω συν την παροχή καυσίμου Wf. Πέρα από την υγιή κατάσταση, παράχθηκε μοντέλο που περιγράφει την υποβαθμισμένη λειτουργία του κινητήρα για τη συνηθέστερη μορφή υποβάθμισης που αναφέρθηκε και παραπάνω, τη συγκέντρωση σκόνης στο συμπιεστή. Τα ποσοστά υποβάθμισης των συνιστωσών φαίνονται στο παρακάτω σχήμα:

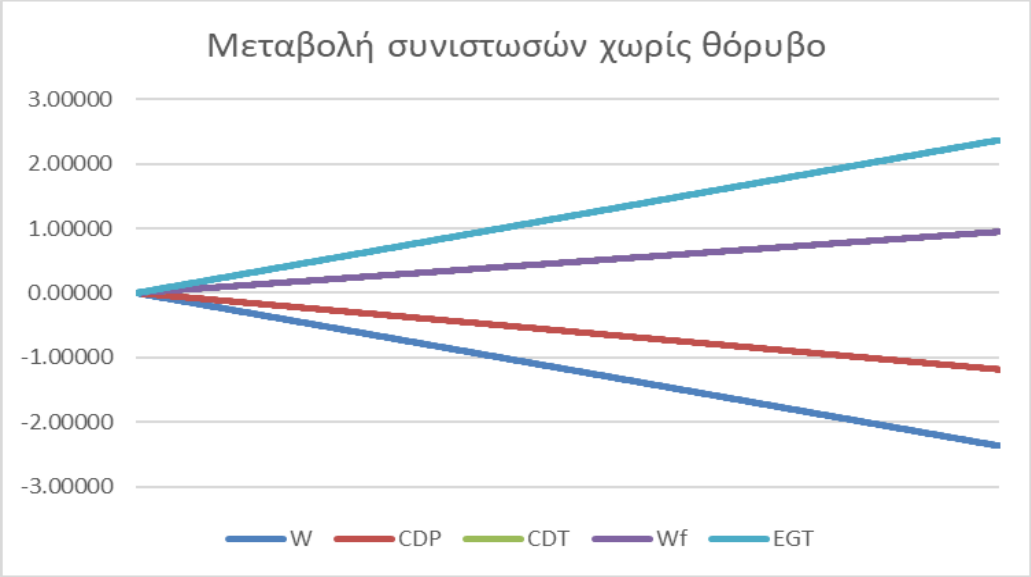


Σχήμα 4.30: Μεταβολή των 5 κύριων συνιστωσών λόγω σκόνης στο συμπιεστή

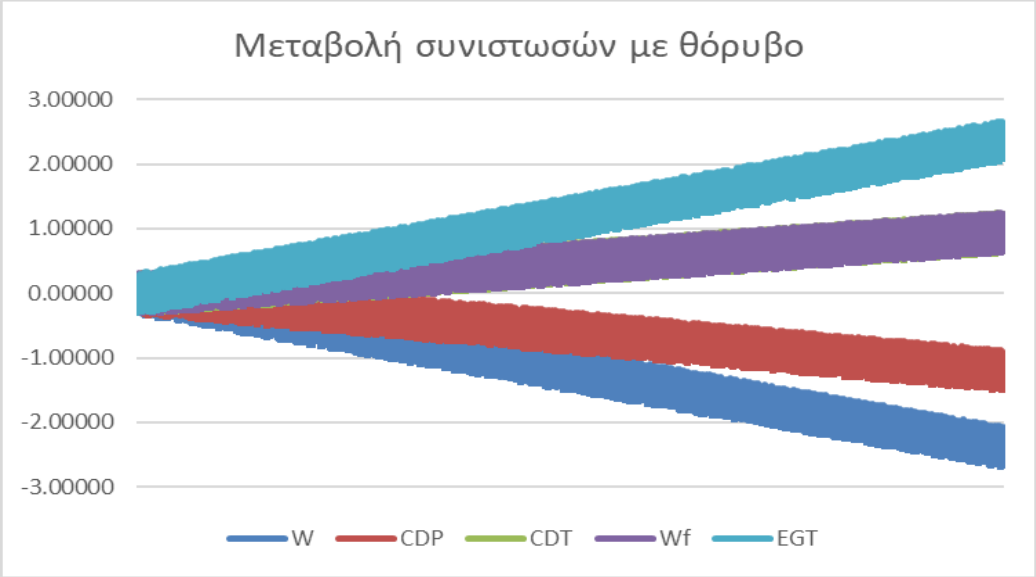
Τα ποσοστά αυτά αφορούν βρόμισμα του συμπιεστή και επιφέρουν υποβάθμιση τέτοιου μεγέθους σε διάρκεια ενός μήνα. Επίσης η βλάβη αυτή, επηρεάζει τόσο το βαθμό απόδοσης όσο και την παροχή του συμπιεστή καθώς για να προκύψουν αυτές οι τιμές, το μοντέλο υποβάθμισης χρησιμοποίησε μείωση 1% και στις 2 αυτές συνιστώσες. Το βρόμισμα του συμπιεστή όπως αναφέρθηκε και παραπάνω είναι πολύ συχνό φαινόμενο υποβάθμισης και εξαρτάται από το περιβάλλον λειτουργίας του κινητήρα. Πιο συγκεκριμένα, η διακύμανση της θερμοκρασίας μέσα σε διάστημα ενός χρόνου μπορεί να δώσει διαφορετικές εκτάσεις στην υποβάθμιση. Για παράδειγμα, το καλοκαίρι, η ατμόσφαιρα έχει περισσότερα σωματίδια που μπορούν να επηρεάσουν τη λειτουργία του κινητήρα από ότι το χειμώνα. Αυτό συμβαίνει επειδή το καλοκαίρι η βροχόπτωση, ειδικά στην Ελλάδα είναι μηδαμινή και έτσι τα σωματίδια αιωρούνται στην ατμόσφαιρα και δυνητικά καταλήγουν στον αγωγό εισόδου της εξεταζόμενης μηχανής. Άρα οι παραπάνω μεταβολές δεν παραμένουν σταθερές όλο το χρόνο.

Επίσης, σημαντικό είναι και το μέγεθος του δείγματος που θα μοντελοποιηθεί. Στην παραπάνω μελέτη χρησιμοποιήθηκαν 7700 σημεία που περιγράφουν την περίοδο περίπου ενός μήνα. Αυτή τη φορά θα μοντελοποιηθεί η περίοδος 1.5 μήνα λειτουργίας. Δεδομένου ότι υπάρχει συχνότητα δειγματοληψίας: $f = 0.2\ λεπτά^{-1}$ δηλαδή γίνεται μία μέτρηση ανά 5 λεπτά, και ότι ο κινητήρας λειτουργεί 24 ώρες, 30 μέρες το μήνα, προκύπτει ότι υπάρχουν περίπου 8600 μετρήσεις ανά μήνα. Μετά από κάθε μήνα θεωρείται ότι ο κινητήρας σταματά για λίγο τη λειτουργία του, καθαρίζεται και ξανά ξεκινά. Για την ανάλυση αυτή θα χρησιμοποιηθούν 13000 μετρήσεις για τη μοντελοποίηση 1.5 μήνα, οπότε θα αναλυθεί και το φαινόμενο καθαρισμού της μηχανής και πάλι.

Οι μετρήσεις και τα αποτελέσματα θα αφορούν δεδομένα με ή χωρίς θόρυβο και τα αποτελέσματα θα συγκριθούν μεταξύ τους. Παρακάτω φαίνονται οι μεταβολές των μεγεθών για υποβάθμιση 170% μεγαλύτερη με και χωρίς θόρυβο.



Σχήμα 4.31: Διάγραμμα μεταβολής συνιστωσών χωρίς θόρυβο



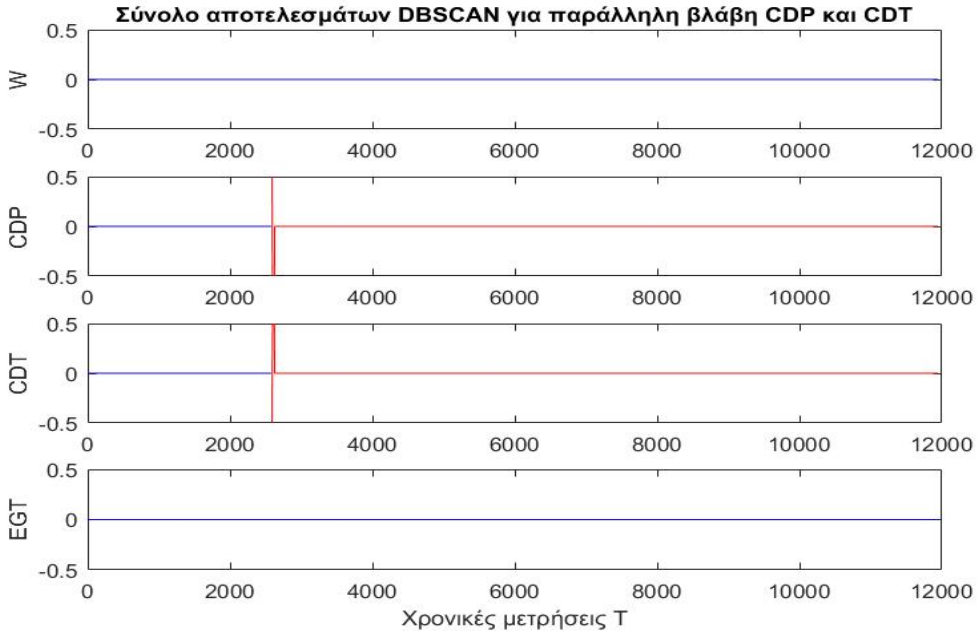
Σχήμα 4.32: Διάγραμμα μεταβολής συνιστωσών με θόρυβο

Ο θόρυβος στο παραπάνω σχήμα έχει διακύμανση ίση με $3\sigma = \pm 0.3\%$. Στη συνέχεια δοκιμάζονται οι στόχοι που αναφέρθηκαν και παραπάνω. Δηλαδή, κανονική λειτουργία, λειτουργία με υποβάθμιση, ήπια ή απότομη, λειτουργία με υποβάθμιση και καθαρισμό του κινητήρα, λειτουργία με υποβάθμιση και βλάβη και τέλος λειτουργία με υποβάθμιση, βλάβη και καθαρισμό, όλα αυτά με ή χωρίς θόρυβο.

Στην ανάλυση που ακολουθεί, η μεταβολή του μεγέθους Wf είναι ακριβώς ίδια με αυτή του μεγέθους CDT. Για λόγους απλότητας λοιπόν, στα παρακάτω αποτελέσματα θα παρίστανται μόνο 4 από τις 5 χρονοσειρές που παράγονται. Ξεκινώντας από την περίπτωση κανονική λειτουργίας, οι αλγόριθμοι από κοινού παρείχαν διάγνωση για δεδομένα με ή χωρίς θόρυβο χωρίς σφάλμα. Εισάγοντας την υποβάθμιση που

μοντελοποιήθηκε μέσω του συνόλου εντολών αφαίρεσης υποβάθμισης που αναφέρθηκε παραπάνω, όλοι οι αλγόριθμοι καταφέρνουν να κατατάξουν κάθε σημείο στην υγιή κατάσταση. Ακολουθούν οι περιπτώσεις μεταβολών κατά τη λειτουργία.

Πρώτα, εξετάζοντας την κατάσταση χωρίς θόρυβο, παρατηρείται ότι όλοι οι αλγόριθμοι αναγνωρίζουν απότομη βλάβη ή συνδυασμού βλάβης θετικού ή αρνητικού πρόσημου είτε υπάρχει είτε όχι το φαινόμενο της υποβάθμισης. Παρακάτω φαίνεται ένα παράδειγμα της μεταβολής αυτής όπου οι μετρητές CDP και CDT υπόκεινται σε βλάβη ενώ στο στροβιλοκινητήρα γενικά επικρατεί υποβάθμιση:



Σχήμα 4.33: Σύνολο αποτελεσμάτων μεθόδων για παράλληλη βλάβη

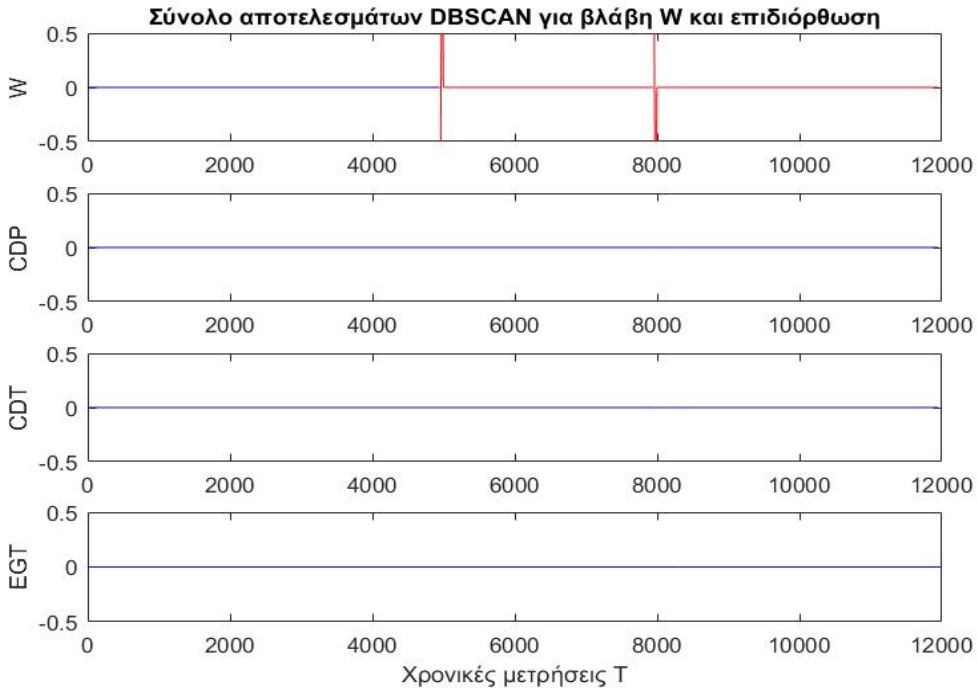
Όπως φαίνεται στο Σχήμα 4.33, οι βλάβες συμβαίνουν ταυτόχρονα στους αισθητήρες CDP και CDT. Οι αλγόριθμοι από κοινού αναγνωρίζουν ότι πρόκειται για βλάβη #8 δηλαδή βλάβη των συνιστωσών CDP και CDT. Σε αυτό το σημείο να σημειωθεί ότι η βλάβη που μοντελοποιήθηκε είναι διαφορετικής κλίμακας για τα 2 αυτά μεγέθη. Σχεδιάστηκε λοιπόν βλάβη +2% για το CDP και -1% για το CDT ώστε να διαπιστωθεί και η ακρίβεια των μεθόδων σε βλάβες διαφορετικού πρόσημου. Επίσης η απόλυτη τιμή των μεταβολών αυτών δεν είναι τυχαία καθώς ανήκουν στα αποδεκτά όρια βλάβης που αναγνωρίζουν οι αλγόριθμοι. Αναλυτικότερα, μετά από εκτενείς ελέγχους, μετρήθηκε ότι η μέγιστη δυνατή μεταβολή που μπορούν να ανιχνεύσουν οι μέθοδοι ως βλάβη είναι ±0.4%, νούμερο λογικό καθώς εξαρτάται από το *epsilon* που διαλέγει ο χρήστης και τον τρόπο μέτρησης απόστασης γειτόνων.

Ακολουθεί η περίπτωση βλάβης ή συνδυασμού βλαβών με επιδιόρθωση της μετά από ορισμένο χρονικό διάστημα. Η διαδικασία αυτή για λειτουργία χωρίς υποβάθμιση είναι αρκετά απλή αλλά γίνεται πιο σύνθετη σε περίπτωση λειτουργίας με υποβάθμιση. Στο σενάριο αυτό, μελετώνται δύο διαφορετικά ενδεχόμενα:

- Βλάβη ή συνδυασμός βλαβών και επιδιόρθωση σε τυχαία χρονική στιγμή
- Βλάβη ή συνδυασμός βλαβών και επιδιόρθωση κατά το μηνιαίο πλύσιμο του κινητήρα

Η επιδιόρθωση χωρίς υποβάθμιση του κινητήρα ανιχνεύεται ως μεταβολή σαν τη μεταβολή λόγω βλάβης οπότε υπάρχει επιτυχία από όλες τις μεθόδους. Η επιδιόρθωση με υποβάθμιση γίνεται πιο σύνθετη όμως για τα δύο παραπάνω ενδεχόμενα.

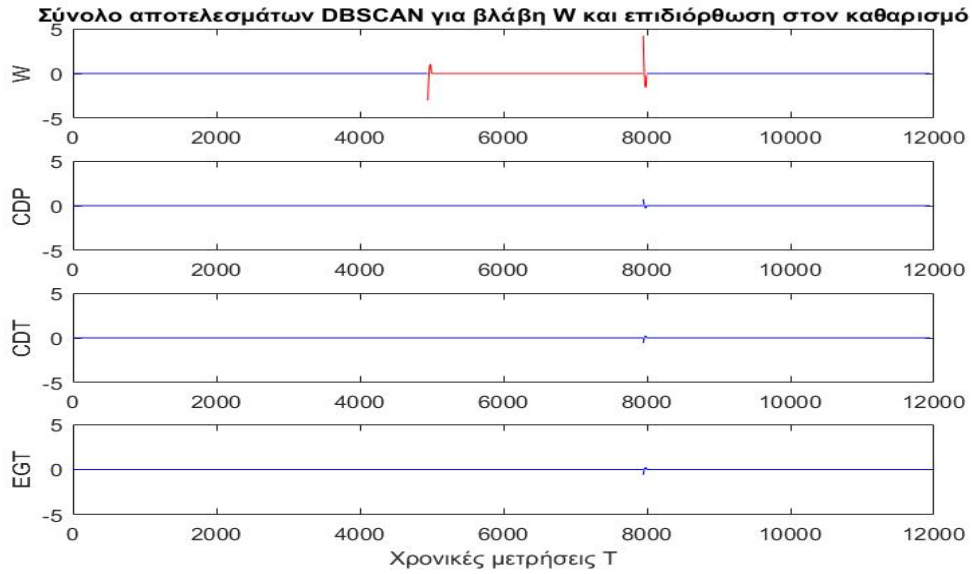
Στο πρώτο ενδεχόμενο παρατηρείται αδυναμία αναγνώρισης της επιστροφής στην υποβαθμισμένη λειτουργία χωρίς βλάβη πλέον και από τους τρεις αλγόριθμους. Έτσι, για μικρές επιδιορθώσεις στον κινητήρα που δεν επιφέρουν κατέβασμα και ολικό έλεγχος του κινητήρα, προτείνεται στο χρήστη η προσεκτική παρακολούθηση των συνιστωσών που αφορούν τις διορθώσεις αυτές. Τα αποτελέσματα φαίνονται παρακάτω:



Σχήμα 4.34: Σύνολο αποτελεσμάτων για βλάβη και επιδιόρθωση ενδεχόμενου 1

Όπως φαίνεται στο παραπάνω σχήμα, με κόκκινο υποδηλώνεται η εσφαλμένη λειτουργία, η οποία επεκτείνεται και μετά τη μέτρηση 8000 παρά την απότομη μεταβολή στη μέτρηση αυτή. Τα θετικά των αποτελεσμάτων αυτών είναι ότι ο αλγόριθμος και πάλι αναγνωρίζει τη βλάβη αυτή σωστά ενώ το αρνητικό είναι ότι δεν αναγνωρίζει πότε διορθώνεται.

Από την άλλη πλευρά, στο δεύτερο ενδεχόμενο προκύπτουν διαφορετικά αποτελέσματα. Στην περίπτωση βλάβης και επιδιόρθωσης κατά το κατέβασμα και καθαρισμό του κινητήρα, γίνεται διάγνωση της βλάβης κανονικά ενώ παράλληλα αναγνωρίζεται και η επιστροφή στην κανονική λειτουργία μετά τον καθαρισμό της μηχανής περίπου στη μέτρηση 8000 που σηματοδοτεί το πέρας ενός μήνα λειτουργίας. Τα αποτελέσματα φαίνονται παρακάτω:



Σχήμα 4.35: Σύνολο αποτελεσμάτων για βλάβη και επιδιόρθωση ενδεχόμενου 2

Τα αποτελέσματα στο ενδεχόμενο 2 είναι διαφορετικά από αυτά του ενδεχομένου 1. Εδώ ο αλγόριθμος αναγνωρίζει σωστά τη βλάβη που μοντελοποιήθηκε, και την επιδιόρθωση της. Επίσης ενδιαφέρον έχουν οι μικρές μεταβολές στη μέτρηση 8000 στα άλλα 3 μεγέθη που προφανώς αφορούν τον καθαρισμό. Οι αλγόριθμοι είναι έτσι κατασκευασμένοι ώστε να αναγνωρίζουν τον καθαρισμό μηχανής λόγω ενδεχόμενης υποβάθμισης σε όλες τις λογικές κλίσεις υποβάθμισης, απότομες και ήπιες.

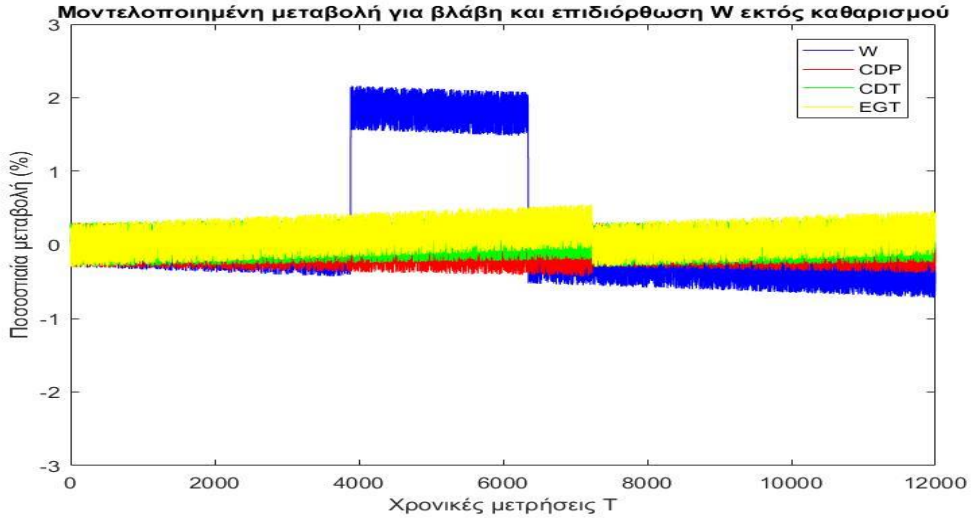
Προχωρώντας στη λειτουργία των μεθόδων σε δεδομένα με θόρυβο, οι στόχοι που επιτυγχάνονται δε διαφέρουν και πολύ. Η διαδικασία ελέγχου είναι ίδια, καθώς τα δεδομένα ελέγχονται σε ήπια και απότομη κλίση, σε υποβαθμισμένη ή βλαμμένη λειτουργία, με ή χωρίς καθαρισμό του κινητήρα σε όλους τους δυνατούς συνδυασμούς. Η μοναδική διαφορά είναι η προσθήκη θορύβου διακύμανσης $3\sigma = \pm 0.3\%$ όπως αναφέρθηκε και παραπάνω.

Ξεκινώντας, με απλή βλάβη ή συνδυασμό βλάβης χωρίς υποβάθμιση τα αποτελέσματα είναι ακριβή με όλες τις μεθόδους να αναγνωρίζουν την περιοχή σφάλματος εξ ολοκλήρου. Όπως και στα δεδομένα χωρίς θόρυβο πραγματοποιείται έλεγχος ελάχιστης δυνατής ανιχνεύσιμης μεταβολής. Υπενθυμίζεται ότι η ελάχιστη ανιχνεύσιμη μεταβολή για δεδομένα χωρίς θόρυβο ήταν ±0.4%. Στα δεδομένα με θόρυβο, δύο υγιή σημεία μπορούν να έχουν μέγιστη κατακόρυφη απόσταση 0.6% λόγω της διακύμανσης των τιμών που αναφέρθηκε παραπάνω. Όπως είναι λογικό, το *epsilon* που επιλέχθηκε για λειτουργία χωρίς θόρυβο δε μπορεί να κατατάξει όλα τα σημεία υγείας στο υγιές cluster λόγω ενδεχόμενων αποστάσεων σημείων μεγαλύτερων του *epsilon*. Για τα δεδομένα με θόρυβο λοιπόν επιλέγεται μεγαλύτερο *epsilon* που ορίζεται παρακάτω:

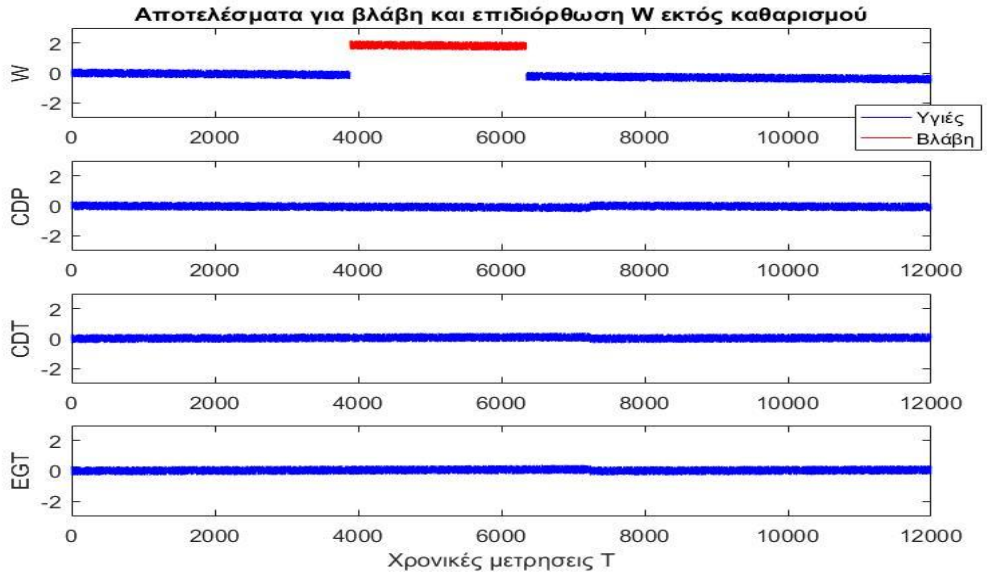
$$epsilon = 0.246$$

Έτσι, αυξάνεται και η τιμή της ελάχιστης δυνατής ανιχνεύσιμης μεταβολής που μετά επαναληπτικές δοκιμές παίρνει την τιμή ±0.9%. Αυτή η τιμή είναι λογική καθώς για σχεδόν τριπλασιασμό του *epsilon* ακολουθεί σχεδόν τριπλασιασμός των ανεκτών ορίων.

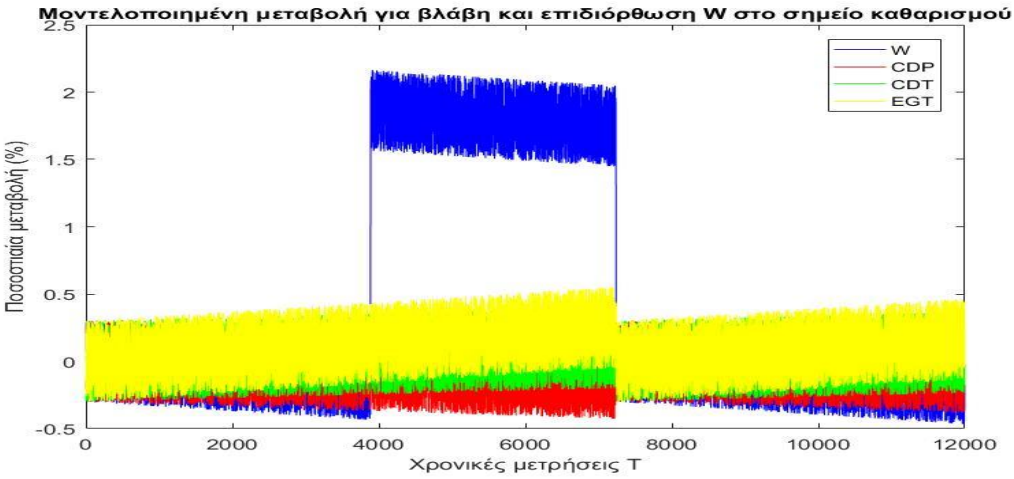
Στη συνέχεια ελέγχθηκε λειτουργία με βλάβη ή συνδυασμό βλαβών και διόρθωση μετά από αόριστο χρονικό διάστημα και οι αλγόριθμοι λειτουργούν με 100% επιτυχία. Τα αποτελέσματα δε διαφέρουν καθόλου από αυτά των δεδομένων χωρίς θόρυβο. Το ίδιο ισχύει και για λειτουργία με απλή υποβάθμιση των μετρούμενων συνιστωσών και καθαρισμό της μηχανής σε διάστημα ενός μήνα. Προχωρώντας στους σύνθετους στόχους, οι αλγόριθμοι έρχονται αντιμέτωποι με τα ενδεχόμενα 1 και 2 που περιεγράφηκαν παραπάνω. Τα αποτελέσματα για τα δύο αυτά ενδεχόμενα φαίνονται παρακάτω:



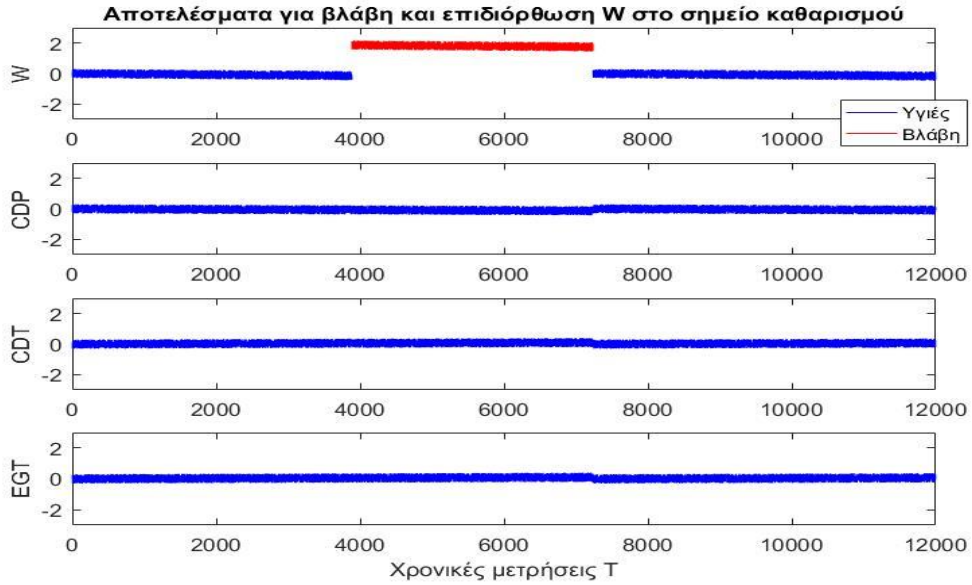
Σχήμα 4.36: Μεταβολή χρονοσειρών για βλάβη και επιδιόρθωση W εκτός καθαρισμού



Σχήμα 4.37: Αποτελέσματα K-means για βλάβη και επιδιόρθωση W εκτός καθαρισμού



Σχήμα 4.38: Μεταβολή χρονοσειρών για βλάβη και επιδιόρθωση W στο σημείο καθαρισμού



Σχήμα 4.39: Αποτελέσματα K-means για βλάβη και επιδιόρθωση W στο σημείο καθαρισμού

Τα τέσσερα παραπάνω σχήματα περιλαμβάνουν τις μεταβολές που μοντελοποιήθηκαν βάσει των ενδεχομένων που αναπτύχθηκαν παραπάνω καθώς και τα αποτελέσματα των μεθόδων. Όπως φαίνεται τα αποτελέσματα είναι θετικά και κοινά για τις τρεις μεθόδους και για τα δύο ενδεχόμενα. Και οι τρεις μέθοδοι ανιχνεύουν 100% των σημείων βλάβης όπως φαίνεται στα Σχήματα 4.36 και 4.38. Λόγω των συνθηκών καθαρισμού που έχουν εισαχθεί, οι αλγόριθμοι δεν ομοιάζουν τις μεταβολές λόγω καθαρισμού με τις μεταβολές λόγω βλάβης. Από τα Σχήματα 4.37 και 4.39 επίσης φαίνεται ότι ανιχνεύεται σωστά και το είδος της βλάβης, τα αποτελέσματα όμως είναι όμοια και για συνδυασμό βλάβης.

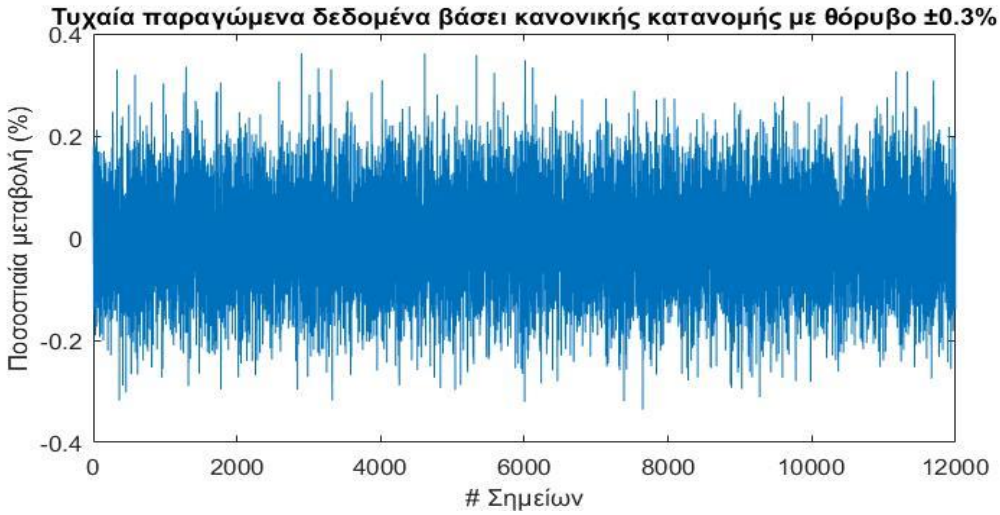
Επίσης σημαντικός τομέας αξιολόγησης της διαγνωστικής ικανότητας είναι η επιβολή σταδιακής μετατόπισης σε τυχαία δεδομένα. Για την αξιολόγηση αυτή χρησιμοποιείται μόνο η μέθοδος DBSCAN εφόσον παραπάνω φάνηκε ότι παρέχει τα ακριβέστερα αποτελέσματα σε step-change μεταβολή και επειδή είναι η μόνη μέθοδος που ειδικεύεται σε σημειακές ανωμαλίες και όχι σε υπολογισμό θέσης κέντρων clusters όπως οι άλλες δύο εξεταζόμενες μέθοδοι.

Πρακτικά εξετάστηκαν τα ενδεχόμενα που παρουσιάστηκαν και παραπάνω για step change μεταβολή μόνο για τη DBSCAN με κύρια διαφορά, τα επιπλέον στάδια μεταβολής μέχρι την τελική βλάβη. Δηλαδή, η παραπάνω ανάλυση αφορά την ικανότητα των μεθόδων να ανιχνεύσουν βλάβη που συμβαίνει σε ένα χρονικό βήμα, ενώ τώρα εξετάζεται η ικανότητα της DBSCAN να ανιχνεύσει βλάβη σε περισσότερα από ένα. Αυτό επιτυγχάνεται μέσω παραγωγής τυχαίων δεδομένων σε υπολογιστικό φύλλο Excel.

Εκεί, παράγονται τυχαία δεδομένα βάσει κανονικής κατανομής με:

- Μέση τιμή: $\mu = 0$ και,
- Τυπική απόκλιση $\sigma = 0.1$

Αυτό σημαίνει ότι τα δεδομένα που παράγονται έχουν αυξημένη πιθανότητα να έχουν τιμή κοντά στο μηδέν, ενώ περίπου 0.25% πιθανότητα να έχουν μία από τις ακραίες τιμές δηλαδή ±0.3%. Στο παρακάτω σχήμα γίνεται ευκρινέστερο αυτό που περιεγράφηκε μόλις:



Σχήμα 4.40: Παραγωγή τυχαίων δεδομένων βάσει κανονικής κατανομής για τη μοντελοποίση βλάβης

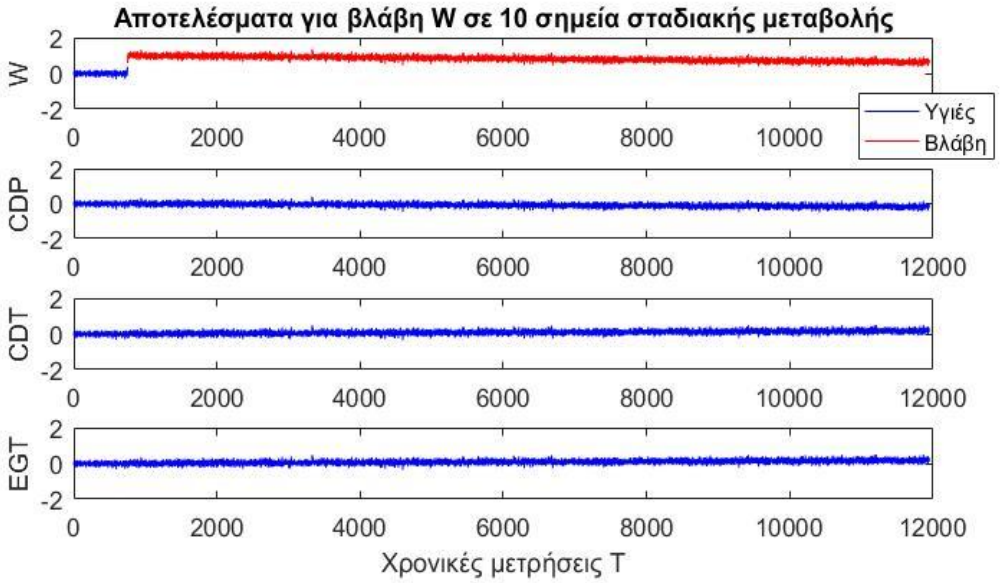
Βάσει αυτού του μοντέλου παράχθηκαν δεδομένα που μοντελοποιούν βλάβη ή πολλαπλές βλάβες, σε παραπάνω από ένα σημεία ώστε να ελεγχθεί η ικανότητα της μεθόδου να διαγνώσει σταδιακές μεταβολές. Εφόσον παραπάνω βρέθηκε ότι η μέθοδος είναι ικανή να διαγνώσει μεταβολή ±0.9% με ή χωρίς θόρυβο ανεξαρτήτου υποβάθμισης, για την ανάλυση που ακολουθεί, χρησιμοποιούνται μεταβολές παρόμοιας κλίμακας.

Χρησιμοποιώντας λοιπόν συνολική βλάβη της τάξης του ±1% ελέγχεται σε πόσα βήματα μέτρησης μπορεί να επιτευχθεί ή να αποτύχει η διάγνωση της βλάβης. Ξεκινώντας

από δύο, τρία, τέσσερα βήματα μέχρι να φτάσει η μεταβολή την απόλυτη τιμή του 1% βλάβης, ο αλγόριθμος εξετάζεται αρχικά στο αν καταφέρνει να ανιχνεύσει τη βλάβη και έπειτα, αν η πρώτη εξέταση είναι επιτυχής, να ορισθεί το σημείο ανίχνευσης μεταξύ της μέσης τιμής δεδομένων 0 και 1%. Μετά από αρχικό έλεγχο της μεθόδου με δεδομένα που αντικατοπτρίζουν τις περιπτώσεις που αναπτύχθηκαν και παραπάνω (με ή χωρίς θόρυβο, με ή χωρίς υποβάθμιση) βρέθηκε ότι τα αποτελέσματα είναι κοινά για εφαρμογή ή μη της υποβάθμισης οπότε ακολούθησε ανάλυση μόνο με την συνθετότερη των περιπτώσεων, δηλαδή με υποβάθμιση. Όπως και νωρίτερα παράγονται δεδομένα με δύο διαφορετικές κλίσης υποβάθμισης, ήπιας και απότομης που περιγράφουν τις περιόδους λειτουργίας ενός κινητήρα, καλοκαίρι και χειμώνα με 170% και 30% κλίση αντίστοιχα.
Τα αποτελέσματα που παράχθηκαν για τις περιπτώσεις χωρίς ή με θόρυβο ήταν τα εξής:

- Χωρίς θόρυβο, ανίχνευση βλάβης μέχρι και σε 15 σημεία σταδιακής μεταβολής
- Με θόρυβο 0.15%, ανίχνευση βλάβης μέχρι και σε 10 σημεία σταδιακής μεταβολής
- Με θόρυβο 0.30%, ανίχνευση βλάβης μέχρι και σε 6 σημεία σταδιακής μεταβολής

Τα αποτελέσματα που παρουσιάζονται παρακάτω αφορούν δεδομένα με θόρυβο κανονικής κατανομής όπως φάνηκε παραπάνω και όχι χωρίς θόρυβο καθώς η μέθοδος οφείλει να εξεταστεί σε δεδομένα που ανταποκρίνονται στην πραγματικότητα.



Σχήμα 4.41:Αποτελέσματα DBSCAN για βλάβη σε 10 σημεία σταδιακής μεταβολής

Όπως είναι λογικό, όσο μεγαλύτερος είναι ο αριθμός σημείων διακριτοποίησης της σταδιακής μεταβολής που ανιχνεύει η μέθοδος, τόσο καλύτερη διαγνωστική ικανότητα την χαρακτηρίζει. Έτσι, για δεδομένα χωρίς θόρυβο η μέθοδος αντιλαμβάνεται τη βλάβη μέχρι και σε 15 σημεία διαχωρισμού της συνολικής μεταβολής του 1%. Αυξάνοντας το θόρυβο στο 0.15% τα σημεία μέγιστης διακριτοποίησης μειώνονται σε 10, ενώ για θόρυβο 0.30% η ανίχνευση επιτυγχάνεται μέχρι και σε 6 σημεία. Γνωρίζοντας το

θεωρητικό υπόβαθρο της μεθόδου συμπεραίνεται ότι τα αποτελέσματα αυτά είναι λογικά και ότι ένα τέτοιο τεστ για τη μέθοδο βασίζεται στην ικανότητα διαχωρισμού κοντινών σημείων από υγιείς γείτονες, σε επιβλαβείς. Αργότερα παρατηρήθηκε ότι τα αποτελέσματα είναι κοινά για υποβάθμιση διαφορετικής κλίσης. Το φαινόμενο αυτό δικαιολογείται βάσει της εξαιρετικής ικανότητας της μεθόδου για αφαίρεση της υποβάθμισης ακόμα και σε δεδομένα με θόρυβο.

Ακολούθησε μία ανάλυση ευαισθησίας ώστε να ελεγχθεί η εγκυρότητα της επιλογής 50 σημείων ως κατευθυντήρια γραμμή για την αφαίρεση της υποβάθμισης. Εφόσον ο χρήστης πλέον γνωρίζει την ικανότητα διαχωρισμού των υγιών σημείων από οποιαδήποτε σημειακή ή σταδιακή μεταβολή, θεωρείται κατάλληλο να γνωρίζει και την ικανή ποσότητα σημείων που πρέπει να τροφοδοτεί στον αλγόριθμό του ώστε να αφαιρεί την υποβάθμιση αξιόπιστα. Τα αποτελέσματα της ανάλυσης ευαισθησίας είναι τα εξής για δεδομένα:

- Χωρίς θόρυβο, η αναγνώριση μέχρι και σε 15 σημεία γίνεται για αριθμό σημείων πρόβλεψης της κλίσης που δίνεται από τη σχέση: $20 \leq x \leq 50$. Για περισσότερα από 50 η βλάβη δεν αναγνωρίζεται.
- Με θόρυβο, η αναγνώριση μέχρι και σε 10 ή 6 σημεία αντίστοιχα γίνεται για αριθμό σημείων πρόβλεψης της κλίσης που δίνεται από τη σχέση: $x \geq 20$. Αρκούν δηλαδή 20 σημεία για τον καθορισμό της κλίσης και οποιοσδήποτε αριθμός μεγαλύτερος του 20 πετυχαίνει το ίδιο αποτέλεσμα.

Τέλος, νωρίτερα αναφέρθηκε ότι στην εξέταση που υποβλήθηκε η μέθοδος DBSCAN παρακολουθήθηκε και το ακριβές σημείο διαχωρισμού υγειούς και επιβλαβούς συνόλου σημείων. Το σημείο αυτό, όπως και με την στιγμιαία-απότομη μεταβολή, ήταν διαφορετικό για δεδομένα με ή χωρίς θόρυβο. Το φαινόμενο αυτό είναι λογικό καθώς ο θόρυβος που εισάγεται στα δεδομένα είναι τυχαίος και δε μπορεί να προβλεφθεί η γραμμή διαχωρισμού με την ίδια ακρίβεια των δεδομένων χωρίς θόρυβο. Συγκεκριμένα παρατηρήθηκε ότι ανάλογα με το είδος δεδομένων η βλάβη 1% στο δείγμα ανιχνευόταν σε:

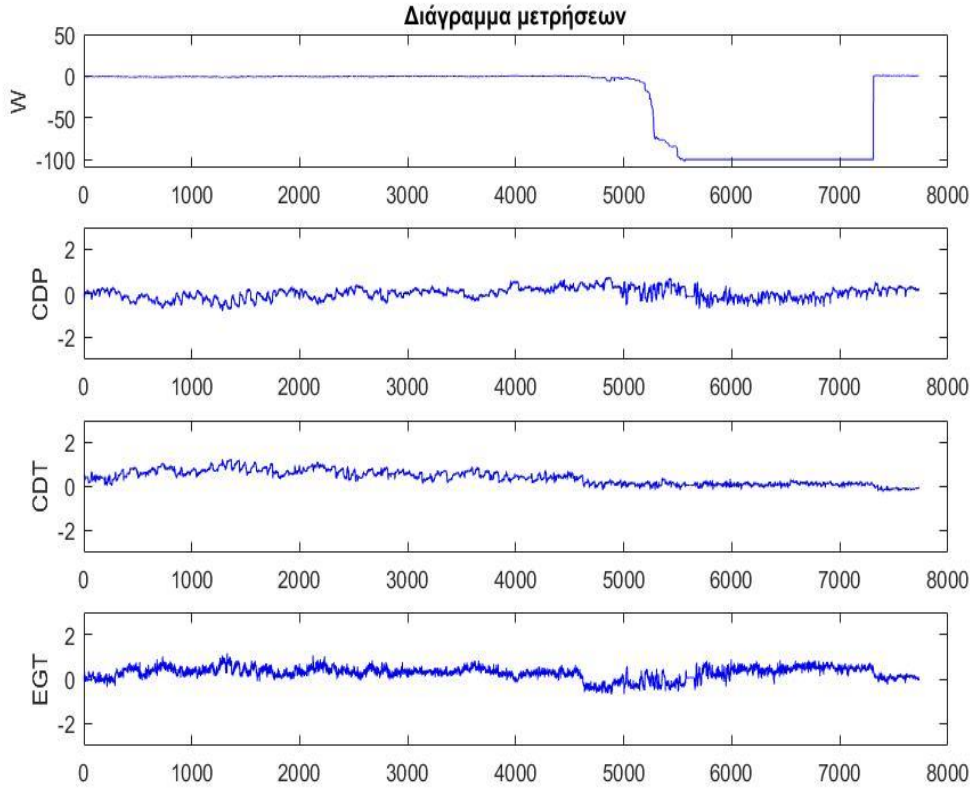
- 0.5-0.6% για δεδομένα χωρίς θόρυβο
- 0.7-0.8% για δεδομένα με θόρυβο

Το αποτέλεσμα αυτό είναι λογικό όπως αναφέρθηκε και νωρίτερα. Ένας επιπλέον τρόπος επαλήθευσης του αποτελέσματος αυτού είναι μέσω σύγκρισης με την ελάχιστη δυνατή ανιχνεύσιμη μεταβολή σημειακής βλάβης. Παρακολουθώντας τις τιμές των δύο αυτών μεγεθών με ή χωρίς θόρυβο αντίστοιχα, παρατηρείται σημαντική ομοιότητα ως προς την τάξη μεγέθους αλλά και ως προς τη διαφορά μεγεθών μεταξύ δεδομένων χωρίς ή με θόρυβο των δύο μεγεθών.

4.5 Εφαρμογή σε πραγματικά δεδομένα

4.5.1 Αεριοστρόβιλος GTA

Σε αυτό το κομμάτι εξετάζονται οι μέθοδοι που περιεγράφηκαν παραπάνω σε πραγματικά δεδομένα. Βάσει της αρχής λειτουργίας της διαγνωστικής έκφανσης των μεθόδων που αναλύθηκε νωρίτερα, εισάγονται δεδομένα στροβιλοκινητήρων που λειτουργούν σε πραγματικό χρόνο ώστε να διαπιστωθεί η εγκυρότητα των μεθόδων.
Το πρώτο σύνολο δεδομένων που εξετάζεται είναι οι μετρήσεις "deltas" που παρουσιάστηκαν αρχικά στο Κεφάλαιο 2 για την περαιτέρω κατανόηση των μεταβλητών εισόδου της DBSCAN: *epsilon* και *MinPts*. Αυτές οι μετρήσεις αφορούν τον κινητήρα GTA για τη λειτουργία του το 2017. Όπως αναφέρθηκε και στο παραπάνω κεφάλαιο, τα δεδομένα αποτελούνται από τέσσερις χρονοσειρές που περιγράφουν τα μεγέθη W, CDP, CDT και EGT για 7700 μετρήσεις. Παρακάτω φαίνεται το σύνολο των μετρήσεων για τις τέσσερις αυτές μεταβλητές:



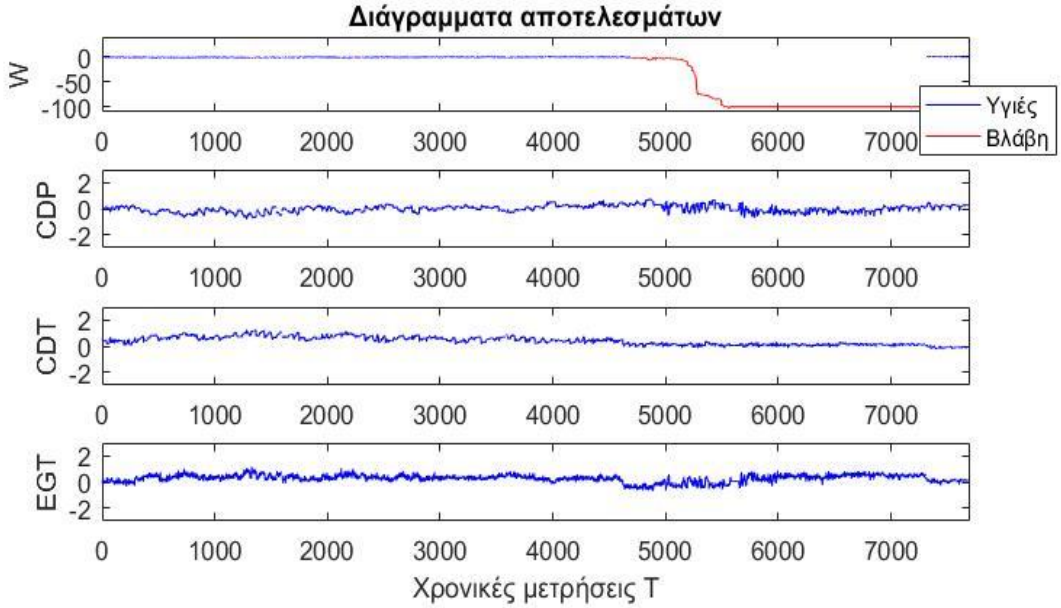
Σχήμα 4.42: Σύνολο μεταβολής τεσσάρων χρονοσειρών (μεταβλητών) συναρτήσει του χρόνου

Είναι φανερό από το Σχήμα 4.42 ότι η πρώτη χρονοσειρά δηλαδή αυτή που μετρά το μέγεθος W υπόκειται σε βλάβη που ξεκινά λίγο πριν τη μέτρηση 5500 και έχει πλέον απότομη κλίση στην παροχή αέρα (-100%) κατά τη μέτρηση 5500 και μετά. Η μηχανή κοντά στη μέτρηση 7500 επανέρχεται σε κανονικές συνθήκες μέσω επιδιόρθωσης οπότε και παρατηρείται μικρή η μεγάλη μεταβολή σε όλες τις μεταβλητές. Με αυτό το σύνολο δεδομένων θα απαντηθούν για τις μεθόδους που αναπτύχθηκαν τα εξής ερωτήματα:

- Αν θεωρεί ο αλγόριθμος ότι υπάρχει βλάβη στο δείγμα
- Σε ποιο σημείο θα αρχίσει να αντιλαμβάνεται ο αλγόριθμος ότι ξεκινά η βλάβη

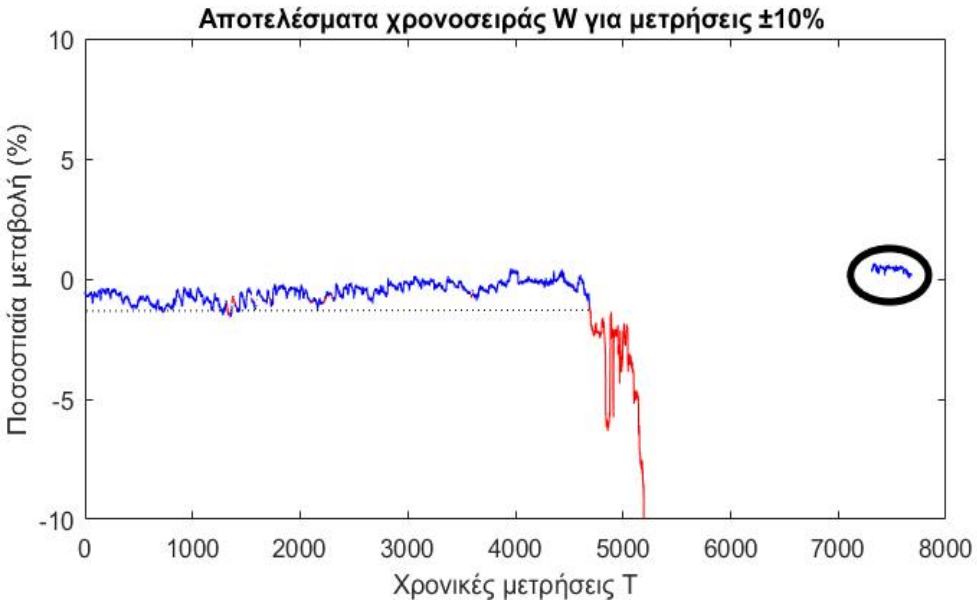
- Εφόσον αναγνωρίσει βλάβη, αν θα καταφέρει να την ξεχωρίσει από το σύνολο μετρήσεων μετά την επιδιόρθωση της μηχανής

Εισάγοντας λοιπόν τα δεδομένα στο περιβάλλον της MATLAB η κάθε μέθοδος διαβάζει τα δεδομένα και ακολουθεί τα βήματα που περιεγράφηκαν παραπάνω. Ελέγχεται δηλαδή η κλίση των δεδομένων, αφαιρείται αν ξεπερνά μία ορισμένη τιμή, συγκρίνεται κάθε σημείο με την υγιή υπογραφή αναφοράς, αν προκύψουν σημεία βλάβης, συγκρίνονται με τις υπογραφές βλάβης για να κατανεμηθούν και πλοτάρονται τα αποτελέσματα. Τα αποτελέσματα που προκύπτουν για τη DBSCAN φαίνονται παρακάτω:



Σχήμα 4.43: Αποτελέσματα μεταβολής των τεσσάρων χρονοσειρών (μεταβλητών) συναρτήσει του χρόνου

Όπως φαίνεται ο αλγόριθμος της μεθόδου αναγνωρίζει επιτυχία ότι υπάρχει βλάβη στην παροχή αέρα καθώς είναι το μόνο διάγραμμα με κόκκινες μετρήσεις. Μεγεθύνοντας λοιπόν σε αυτό το σχήμα έχουμε το εξής αποτέλεσμα:

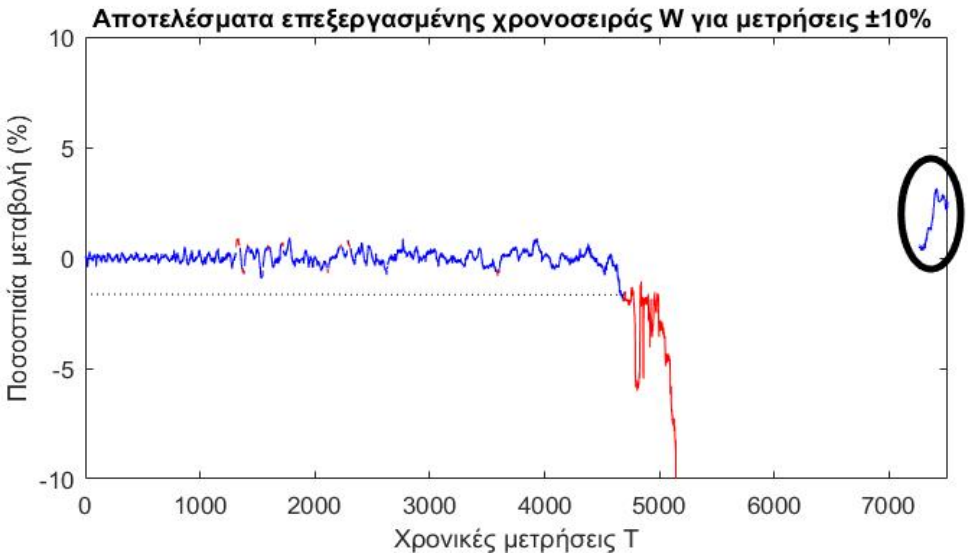


Σχήμα 4.44: Μεγεθυμένη έκδοση του πρώτου διαγράμματος

Φαίνεται καθαρά στο Σχήμα 4.44 ότι η μέθοδος απαντά σωστά στα ερωτήματα που τέθηκαν παραπάνω. Αρχικά αναγνωρίζει ότι υπάρχει βλάβη, που ήταν ο ευκολότερος στόχος. Μέσω της διακεκομμένης γραμμής μπορεί να ορισθεί το πρώτο σημείο που ονομάστηκε επιβλαβές ώστε να γνωρίζει ο χρήστης μετά από πόσες μετρήσεις η μέθοδος πραγματοποιεί διάγνωση της βλάβης. Συγκεκριμένα αυτό γίνεται στη μέτρηση 4635 που έχει τιμή 0.91% πολύ κοντά στις τιμές που παράχθηκαν με τα ιδανικά δεδομένα που αναλύθηκαν νωρίτερα για σταδιακή μεταβολή 1%. Επίσης, με μαύρο κύκλο περικλείονται τα σημεία μετά την επιδιόρθωση του κινητήρα που φαίνεται να έχουν κατανεμηθεί σωστά λόγω του μπλε χρώματός τους, δηλαδή στην υγιή περιοχή. Στη συνέχεια, παρατηρούνται συνολικά 4 σημεία κοντά στη μέτρηση 1500 που έχουν ονοματιστεί εσφαλμένα ότι ανήκουν στην επιβλαβή περιοχή. Είναι όμως λογικά να υπάρχουν μερικά μεμονωμένα δεδομένα σφάλματος, ειδικά όταν πρόκειται για 4 στα συνολικά 7730, δηλαδή 0.05% των μετρήσεων.

Μιλώντας για αριθμό σημείων, καλό είναι σε αυτό το σημείο να αναφερθεί η συχνότητα δειγματοληψίας των δεδομένων. Σε αυτά τα δεδομένα υπάρχει συχνότητα: 1 μέτρηση ανά 15 λεπτά, που οδηγεί στο συμπέρασμα ότι οι 4 αυτές ανεξάρτητες λανθασμένες ενδείξεις οδηγούν σε αναμονές 30 λεπτών μέχρι την επαναφορά στην κανονική λειτουργία. Κρίνοντας από αυτό, και έχοντας υπόψη ότι οι ενδείξεις αυτές είναι ανεξάρτητες, η DBSCAN έκανε συνολικά άριστη δουλειά με αυτά τα πραγματικά δεδομένα.

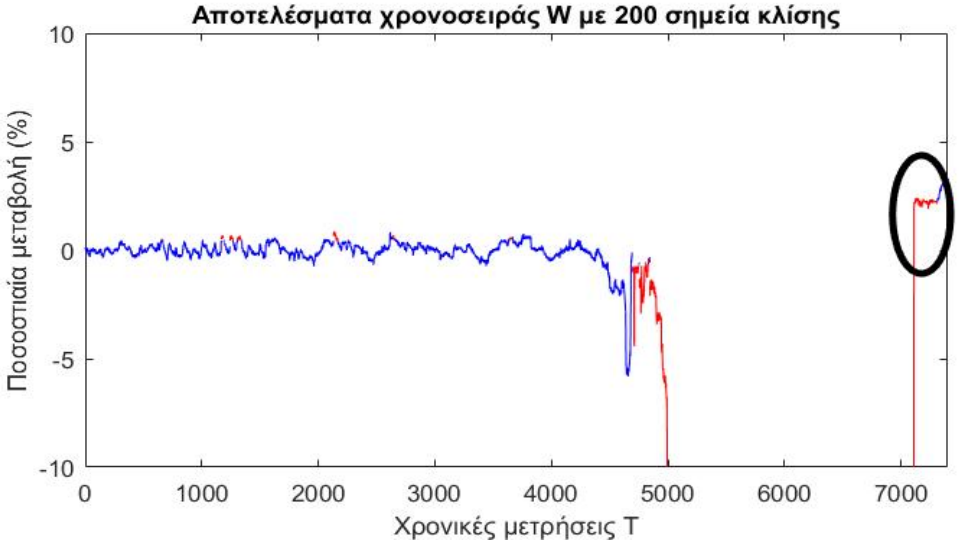
Τα αποτελέσματα του σχήματος 4.45 περιέχουν τη μεταβολή των πραγματικών δεδομένων, αξίζει όμως να παρουσιαστούν και να δικαιολογηθούν τα επεξεργασμένα δεδομένα που προκύπτουν μετά την αφαίρεση υποβάθμισης. Τα επεξεργασμένα δεδομένα φαίνονται παρακάτω:



Σχήμα 4.45: Αποτελέσματα επεξεργασμένης χρονοσειράς W

Φαίνεται ότι ο αλγόριθμος λειτουργεί κανονικά αφού δεν υπάρχει κάποια απότομη ανεξήγητη μεταβολή. Η περιοχή επαναφοράς λόγω επιδιόρθωσης ήταν ένα ενδιαφέρον σημείο διερεύνησης καθώς είναι λογικό να είναι ευάλωτη σε απότομες αυξομειώσεις σημείων. Οι τιμές όμως που περικλείονται από το μαύρο κύκλο φαίνεται να κινούνται σε αποδεκτά όρια. Ακόμα και οι τελευταίες τιμές που καταλήγουν σε τιμές μεγαλύτερες του 1% θετική μεταβολή κατατάσσονται στην υγιή περιοχή λόγω των πολλών διακριτών σημείων μεταξύ μηδενικής και τελικής μεταβολής.

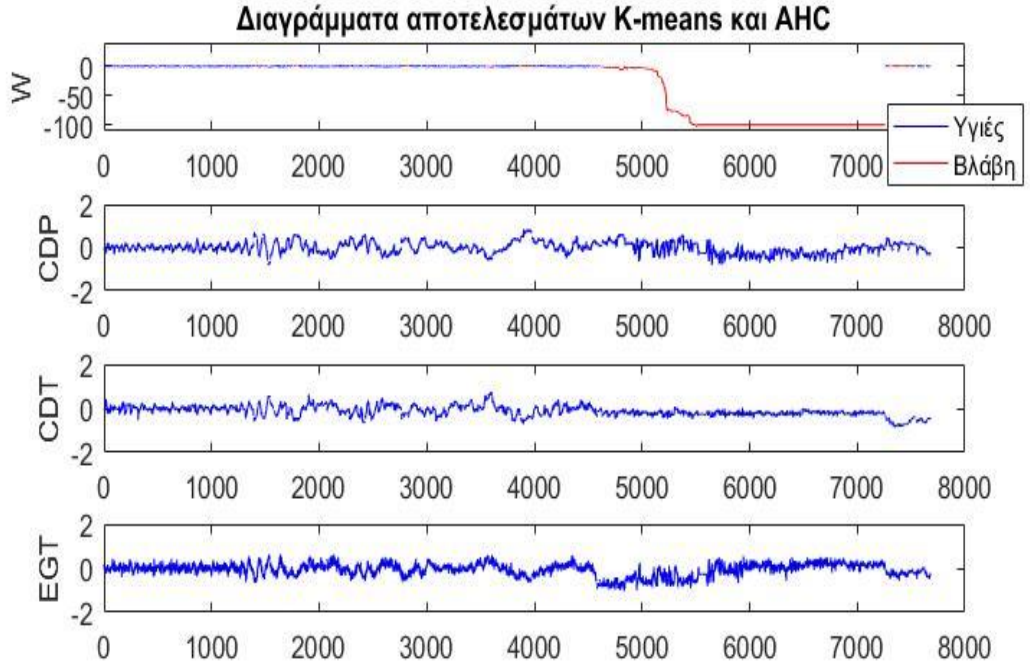
Τέλος, όσον αφορά τη DBSCAN πραγματοποιήθηκε μία ανάλυση ευαισθησίας και πάλι για τον αριθμό σημείων που επηρεάζουν την κλίση. Τα αποτελέσματα ήταν κοινά για αριθμό σημείων μικρότερο του 200, για 200 και πάνω τα αποτελέσματα ήταν αρνητικά καθώς δεν γινόταν διαχωρισμός της υγιούς κατάστασης μετά την επιδιόρθωση και της εσφαλμένης λειτουργίας.



Σχήμα 4.46: Αποτελέσματα επεξεργασμένης χρονοσειράς W για 200 σημεία κλίσης

Φαίνεται καθαρά ο μη διαχωρισμός των σημείων μετά τη μέτρηση 7000. Έτσι επαληθεύεται η προηγούμενη ανάλυση ευαισθησίας που μας ενημερώνει ότι ο ιδανικός αριθμός είναι περίπου στα 50 σημεία.

Αναφέρθηκε και νωρίτερα ότι ο τρόπος χτισίματος των αλγορίθμων είναι κοινός οπότε δυνητικές διαφορές στα αποτελέσματα μπορούν να προκύψουν μόνο λόγω της θεωρητικής βάσης των μεθόδων. Πιο συγκεκριμένα, τα ίδια δεδομένα (deltas) εισάχθηκαν και στις άλλες δύο μεθόδους που μελετώνται, την K-means και την AHC. Αυτές με τη σειρά τους, ακολούθησαν τα βήματα που αναπτύχθηκαν στο θεωρητικό κομμάτι των μεθόδων αλλά και στις τροποποιήσεις για τη διαγνωστική ικανότητα των μεθόδων. Τα αποτελέσματα λοιπόν ήταν ίδια όσον αφορά την έγκυρη ανίχνευση βλάβης, την σωστή ταξινόμηση και το πρώτο σημείο βλάβης. Η μόνη διαφορά όπως φαίνεται παρακάτω είναι στην κατάταξη των σημείων μετά την επιδιόρθωση της βλάβης.



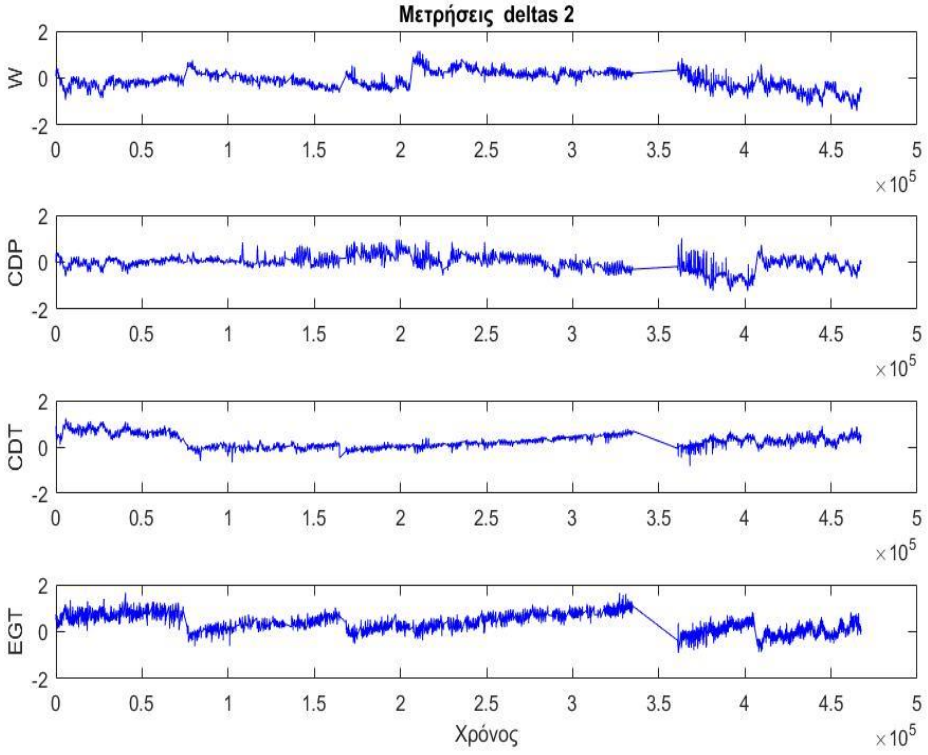
Σχήμα 4.47: Αποτελέσματα συνόλου χρονοσειρών για τις K-means και AHC

Όπως φαίνεται στο σχήμα της μεταβλητής W λοιπόν, υπάρχουν ορισμένα σημεία μετά την επιδιόρθωση που κατατάσσονται στο επιβλαβές cluster. Μετά από προσεκτική μελέτη σημείο προς σημείο, φάνηκε ότι αυτό οφείλεται στις αποστάσεις των σημείων αυτών μεταξύ τους μετά από την επεξεργασία λόγω αφαίρεσης υποβάθμισης. Όπως είναι γνωστό και οι δύο αυτές μέθοδοι υπολογίζουν αποστάσεις από κέντρα cluster, έτσι είναι λογικό, όπως και στα validation test, μερικά σημεία να καταταχθούν λάθος λόγω κοντινών τιμών αποστάσεων των δύο σχηματιζόμενων clusters. Τέλος, τα σημεία αυτά μετρήθηκαν και είναι 131 στον αριθμό, δίνοντας την αίσθηση ότι πρόκειται για μικρό ποσοστό εσφαλμένων ενδείξεων. Αν όμως αναλογιστεί κανείς ότι πρόκειται για συχνότητα δειγματοληψίας: 1 μέτρηση ανά 15 λεπτά, μπορεί εύκολα να βγάλει το συμπέρασμα ότι

131 μετρήσεις μεταφράζονται σε μεγάλο χρονικό διάστημα εσφαλμένων ενδείξεων. Η DBSCAN για άλλη μία φορά λοιπόν είναι ένα βήμα πιο μπροστά.

Νέα δεδομένα αεριοστρόβιλου GTA

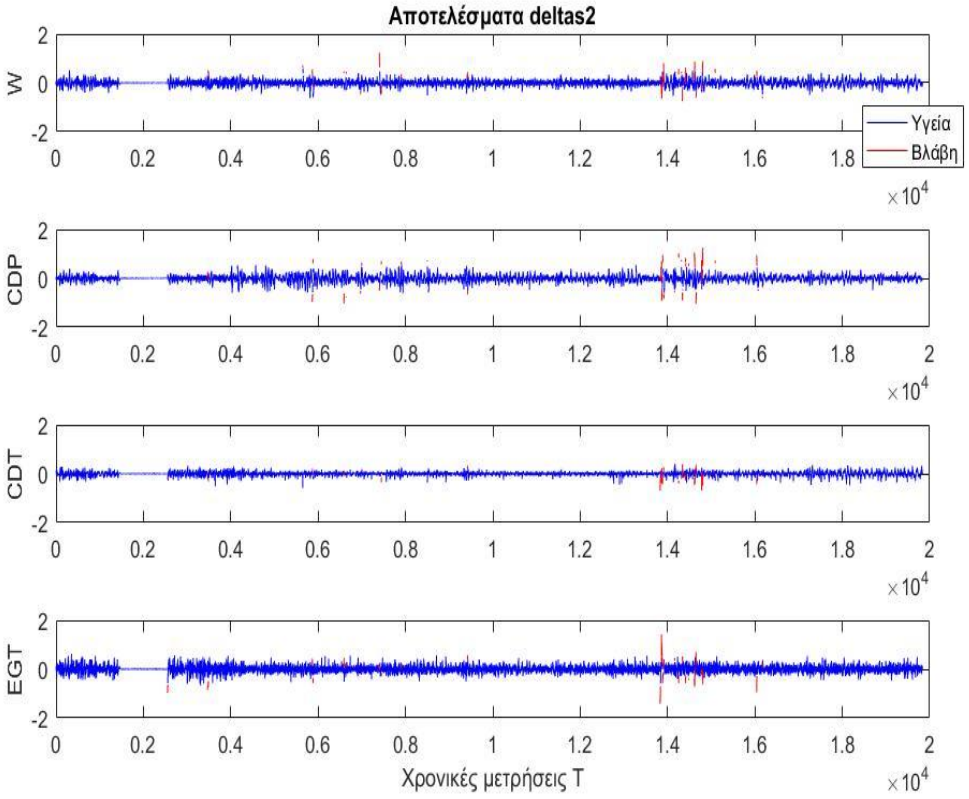
Χρησιμοποιώντας τον ίδιο αλγόριθμο και τις ίδιες υπογραφές αναφοράς μπορούν να εξεταστούν διάφοροι τύποι δεδομένων. Εισάγονται λοιπόν στον αλγόριθμο νέα δεδομένα προς διάγνωση. Αυτά, αφορούν τον ίδιο κινητήρα αλλά για τη χρονιά 2016. Τα δεδομένα αυτά αποτελούνται και αυτά από 4 χρονοσειρές που περιγράφουν τα ίδια μεγέθη με νωρίτερα, W, CDP CDT και EGT. Η παροχή καυσίμου W_f δεν παρείχε σταθερά αποτελέσματα όποτε δεν αναλύθηκε. Συνολικά παρατηρήθηκαν 20000 μετρήσεις οι οποίες φαίνονται στο παρακάτω διάγραμμα:



Σχήμα 4.48: Σύνολο χρονοσειρών GTA 2016

Με την εισαγωγή αυτών των δεδομένων στο περιβάλλον της MATLAB ακολουθούν οι ρυθμίσεις χρήστη. Για το κομμάτι σύγκρισης και εκχώρησης του cluster υγείας χρησιμοποιήθηκε ευκλείδεια απόσταση καθώς παρείχε καλύτερα αποτελέσματα από το συντελεστή αλληλοσυσχέτισης. Για το κομμάτι της σύγκρισης των σημείων που θεωρούνται βλαβερά με τα cluster βλάβης είναι καταλληλότερο να χρησιμοποιηθεί CCD απόσταση καθώς τα δεδομένα είναι αδιάστατα και η σύγκριση τάσης κίνησης πριμοδοτεί περισσότερο τα σημεία ακραίων μεταβολών οπότε γίνεται ευκολότερη η κατανομή τους σε cluster βλάβης.

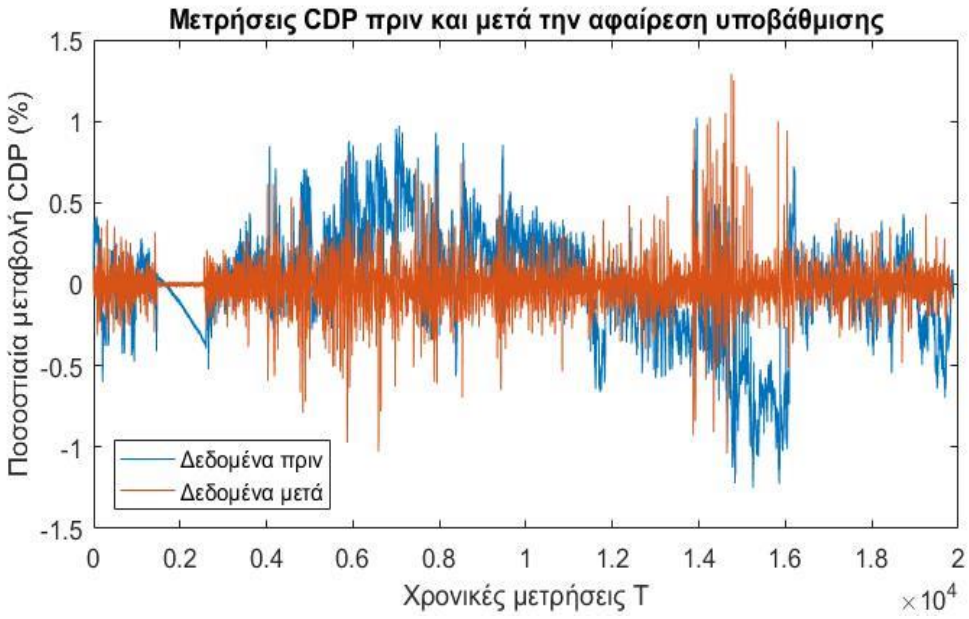
Στη συνέχεια, έγινε μία σημαντική αλλαγή στα μοντελοποιημένα cluster βλάβης καθώς πλέον έχουν θόρυβο κανονικής κατανομής ώστε να θεωρείται πλήρως τυχαία η κατανομή τους στο χώρο. Τα αποτελέσματα είναι κοινά αλλά θεωρείται καταλληλότερο να υπάρχει θόρυβος κανονικής κατανομής. Τα αποτελέσματα αυτών φαίνονται παρακάτω:



Σχήμα 4.49: Αποτελέσματα συνόλου χρονοσειρών για τη GTA

Όπως φαίνεται ο αλγόριθμος αναγνωρίζει κάποια σημεία βλάβης, αλλά είναι ελάχιστα σε σχέση με τα υγιή σημεία. Συγκεκριμένα από 20000 σημεία, μόνο 200 βγήκαν βλαβερά και μόνο 4 (διαδοχικά) σημεία ταίριαξαν με το cluster καθαρισμού που είδαμε παραπάνω. Η συνολική εικόνα μας δείχνει ότι πρόκειται για υγιή δεδομένα, υγιή λειτουργία του κινητήρα δηλαδή, που σε κάποιο χρονικό διάστημα, γύρω στις 14000 μετρήσεις σημειώθηκε καθαρισμός του κινητήρα.

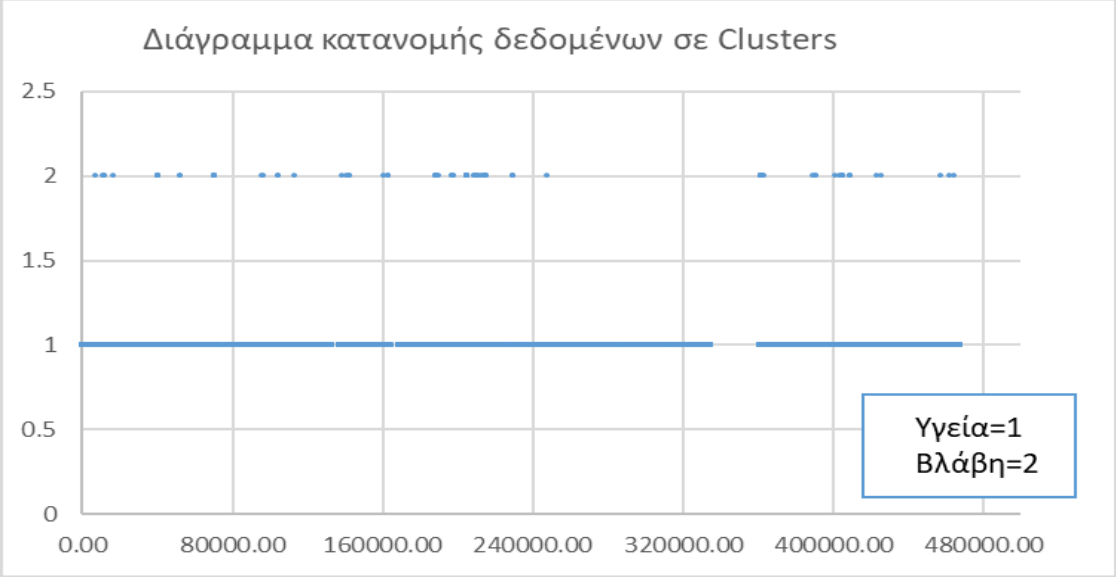
Μεγάλο ποσοστό της 99.5% επιτυχίας που σημείωσε ο αλγόριθμος οφείλεται στην αποτελεσματική αφαίρεση υποβάθμισης που πραγματοποιείται από τον αλγόριθμο. Η εύρεση της κλίσης και η αφαίρεση της από τον αλγόριθμο έχεις ως αποτέλεσμα το εξής διάγραμμα:



Σχήμα 4.50: Μετρήσεις CDP πριν και μετά την αφαίρεση υποβάθμισης

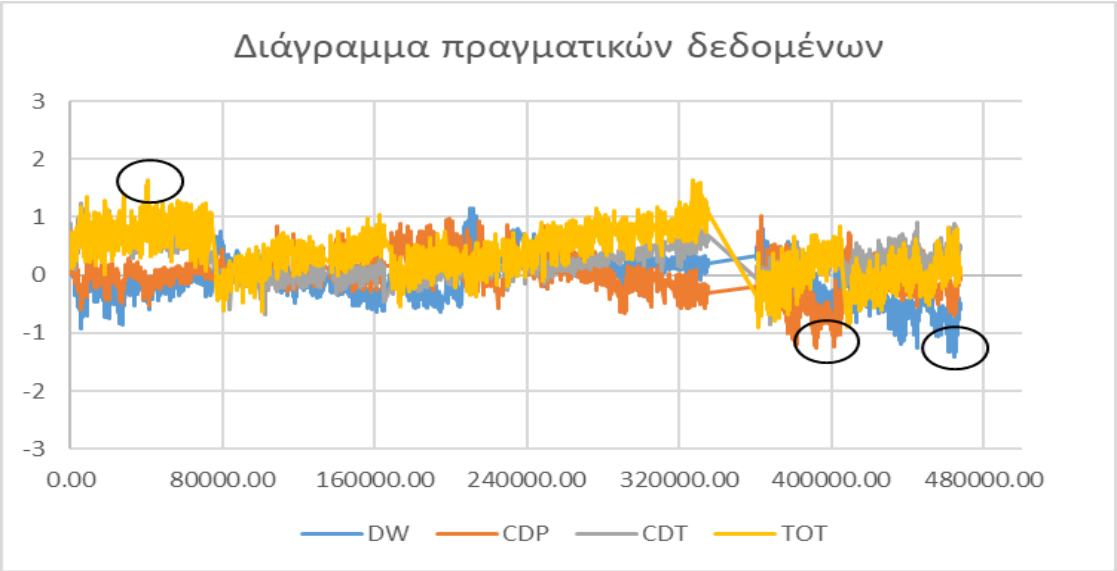
Η μπλε γραμμή περιγράφει τα δεδομένα CDP όπως τα δεχτήκαμε ενώ η πορτοκαλή τα νέα δεδομένα μετά την αφαίρεση της υποβάθμισης. Σε αυτό το σημείο γεννάται το ερώτημα, ποια θα ήταν η επίδοση του αλγορίθμου χωρίς αφαίρεση υποβάθμισης? Τα αποτελέσματα ήταν συντριπτικά χειρότερα καθώς 80% των μετρήσεων διαγνώστηκαν λανθασμένα. Έτσι συμπεραίνεται ότι η αφαίρεση υποβάθμισης είναι αξιοσημείωτη και σωστά αναπτύχθηκε για τη διαγνωστική μέθοδο.

Ακολουθεί το διάγραμμα κατανομής των δεδομένων σε 2 Clusters. Κάνοντας τον πρώτο έλεγχο των αλγορίθμων όπως αναφέρθηκε και παραπάνω, τα δεδομένα χωρίζονται απλά σε υγιή ή επιβλαβή. Άρα με ένδειξη 1 συμβολίζονται τα υγιή, ενώ με ένδειξη 2 τα επιβλαβή όπως φαίνεται παρακάτω:



Σχήμα 4.51: Διάγραμμα κατανομής δεδομένων σε Clusters

Μόνο 0.5% των δεδομένων ανήκουν σε διάφορες διαγνώσεις βλάβης. Παρόλα αυτά, μελετώντας τις τιμές των χρονοσειρών που μας δόθηκαν μπορεί κανείς να διακρίνει γιατί προέκυψαν ακόμα και αυτές.



Σχήμα 4.52: Διάγραμμα πραγματικών δεδομένων

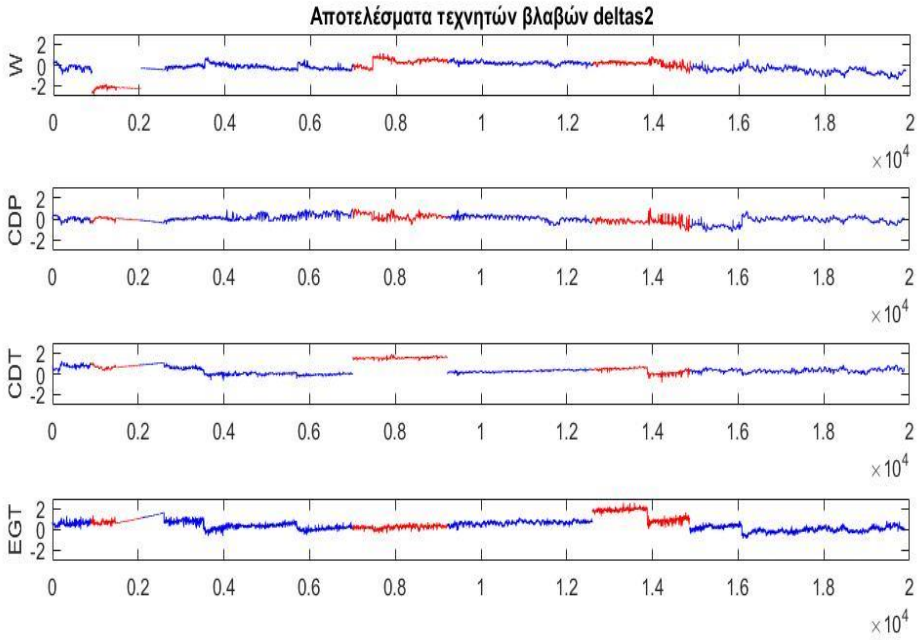
Στο παραπάνω σχήμα φαίνονται όλες οι τιμές των σημείων που αναλύθηκαν. Οι 3 κύκλοι δείχνουν τις μοναδικές περιοχές που δεν ονοματίζονται ως υγεία μετά την αύξηση του epsilon κατά 0,2. Στον πρώτο μαύρο κύκλο υπάρχει τιμή που εκχωρήθηκε ως βλάβη EGT για παράδειγμα. Τα σημεία στους τρείς κύκλους είναι 50-60 στα 20000 οπότε η συνολική επιτυχία του αλγόριθμου με την επαύξηση του epsilon ανέρχεται στο 99.7%.

Στη συνέχεια, ακολούθησε διερεύνηση για το βέλτιστο αριθμό σημείων προσδιορισμού της κλίσης υποβάθμισης. Όπως είπαμε νωρίτερα ο ιδανικότερος αριθμός για τα προηγούμενα δεδομένα ήταν 50, αυτή τη φορά τα αποτελέσματα δε διαφέρουν και πολύ. Η κύρια διαφορά είναι στο στόχο που προσπαθεί κανείς να πετύχει με τα δεδομένα του. Σε αυτά τα δεδομένα, που όλα σχεδόν προκύπτουν υγιή, σημαντικό είναι να παρατηρήσουμε το ποσοστό σφάλματος. Μειώνοντας λοιπόν τον αριθμό σημείων κλίσης το αποτέλεσμα παραμένει ίδιο, αλλά δεν προσδιορίζεται σωστά το σημείο καθαρισμού. Κατά τα άλλα τα αποτελέσματα είναι κοινά μέχρι τα σημεία κλίσης να γίνουν λιγότερα από 10. Τότε, η τυχαία κατανομή των σημείων επηρεάζει πολύ τον προσδιορισμό μετατόπισης των νέων σημείων παρέχοντας αρκετά ασταθή αποτελέσματα. Αυξάνοντας τον αριθμό σημείων κλίσης, αυξάνονται και τα σημεία βλάβης. Έτσι από 100 σημεία βλάβης που υπάρχουν με 50 σημεία κλίσης τα σημεία βλάβης σε 200 σημεία κλίσης γίνονται 500 με τον αριθμό αυτό να αυξάνεται για περισσότερα σημεία προσδιορισμού κλίσης. Έτσι συμπεραίνεται ότι ο βέλτιστος αριθμός είναι 50 και σε αυτό το σετ δεδομένων καθώς καθορίζεται σωστά το σημείο καθορισμού κρατώντας παράλληλα τα σημεία βλάβης στο ελάχιστο δυνατό.

Όσον αφορά το χρόνο, δεδομένου ότι τα δεδομένα που δόθηκαν έχουν συχνότητα δειγματοληψίας περίπου 1/15 λεπτά, συμπεραίνεται ότι για τη διάγνωση 1 βλαβερού

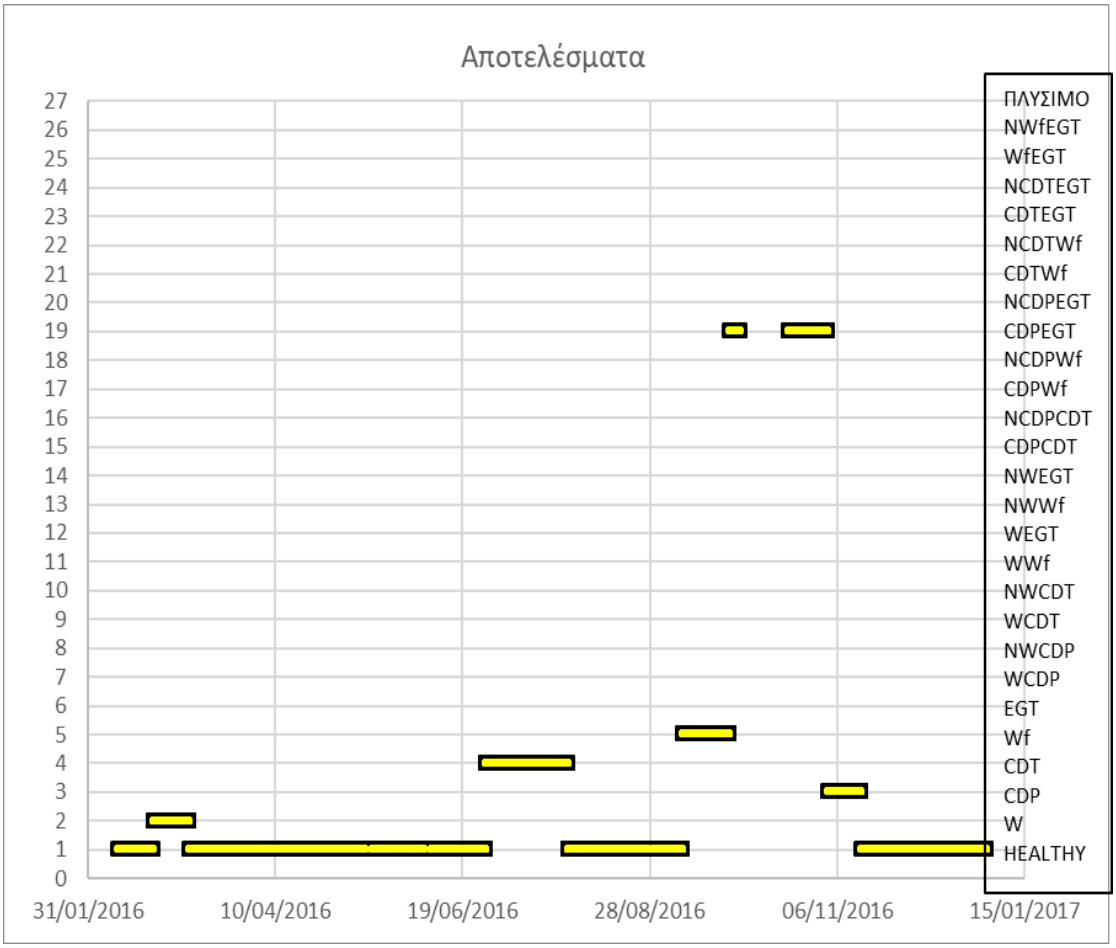
σημείου πρέπει να υπάρχουν περίπου 12 ώρες καταγραφής νωρίτερα ώστε να είμαστε βέβαιοι ότι δεν πρόκειται για απλή υποβάθμιση του κινητήρα.

Τέλος, ενδιαφέρον έχει να εφαρμοσθεί στα δεδομένα αυτά επίτηδες επιπλέον βλάβη. Όπως φάνηκε νωρίτερα, τα δεδομένα αυτά περιγράφουν στη συντριπτική πλειοψηφία τους υγιή λειτουργία του κινητήρα. Εισάγοντας τώρα σε ένα τυχαίο σημείο σε μία τυχαία μεταβλητή πολλαπλές τεχνητές βλάβες, ας δούμε τι αποτελέσματα παρέχει ο αλγόριθμος. Για τις τυχαίες βλάβες, στα δεδομένα εισάχθηκε αρχικά μεγάλη step βλάβη, και αν ο αλγόριθμος έκανε σωστή διάγνωση, μειωνόταν το step και επαναλαμβανόταν η διαδικασία. Το τελικό αποτέλεσμα προκύπτει για βλάβη 1% και προς τις 2 κατευθύνσεις, με τα αποτελέσματα του αλγορίθμου των μεθόδων να φέρνουν εις πέρας τη δοκιμασία αυτή όπως φαίνεται στο παρακάτω σχήμα:



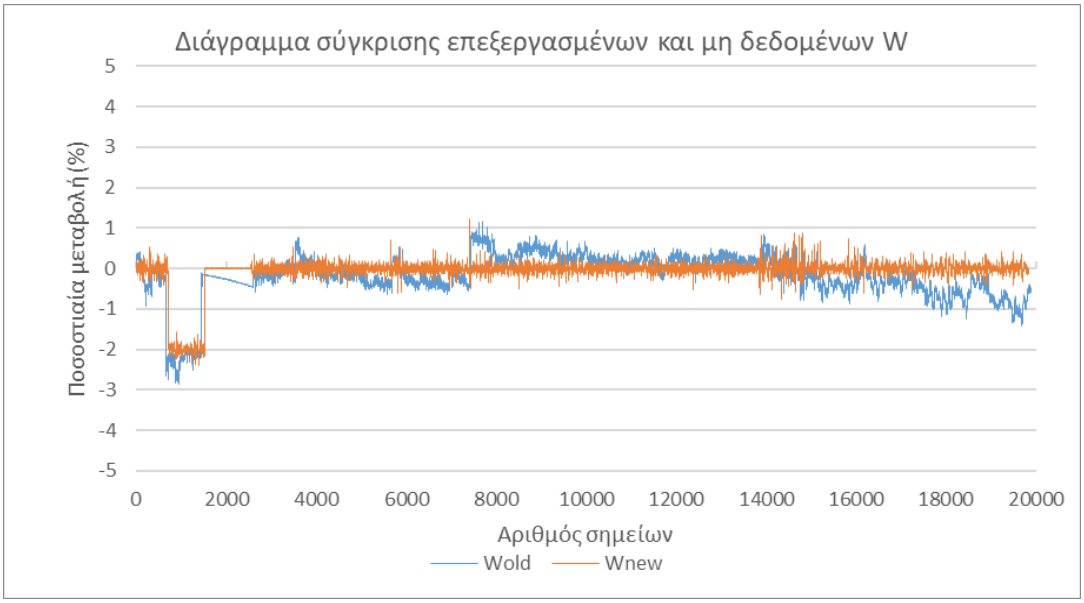
Σχήμα 4.53: Αποτελέσματα τεχνητών βλαβών GTA

Αρχικά από το παραπάνω Σχήμα είναι ξεκάθαρο, παρά την προσπάθεια αφαίρεσης υποβάθμισης, ότι οι βλάβες στο W, CDP, CDT και EGT ανιχνεύονται σωστά ανιχνεύει ο αλγόριθμος. Επίσης σημαντικό είναι ότι ο αλγόριθμος αντιμετωπίζει κι άλλες σχετικά μεγάλες μεταβολές στις υπόλοιπες μεταβλητές, παρόλα αυτά όμως τις ομαδοποιεί στο υγιές cluster. Αυτό συμβαίνει λόγω της μεταβλητής εισόδου *epsilon*. Αν δεν ήταν τόσο στενά ορισμένη η μεταβλητή αυτή, τυχαίες μεταβολές όπως αυτές που μόλις αναφέρθηκαν θα κατανέμονταν στα cluster βλάβης μπερδεύοντας το χρήστη και το σύστημα. Επίσης σημαντικό σε αυτή τη διαδικασία είναι η χρήστη του συντελεστή αλληλοσυσχέτισης όπως είπαμε και νωρίτερα, για την κατανομή βλαβών. Όπως βλέπουμε, κόκκινες περιοχές υπάρχουν μόνο στις περιοχές βλάβης και το είδος της βλάβης φαίνεται στο παρακάτω σχήμα:

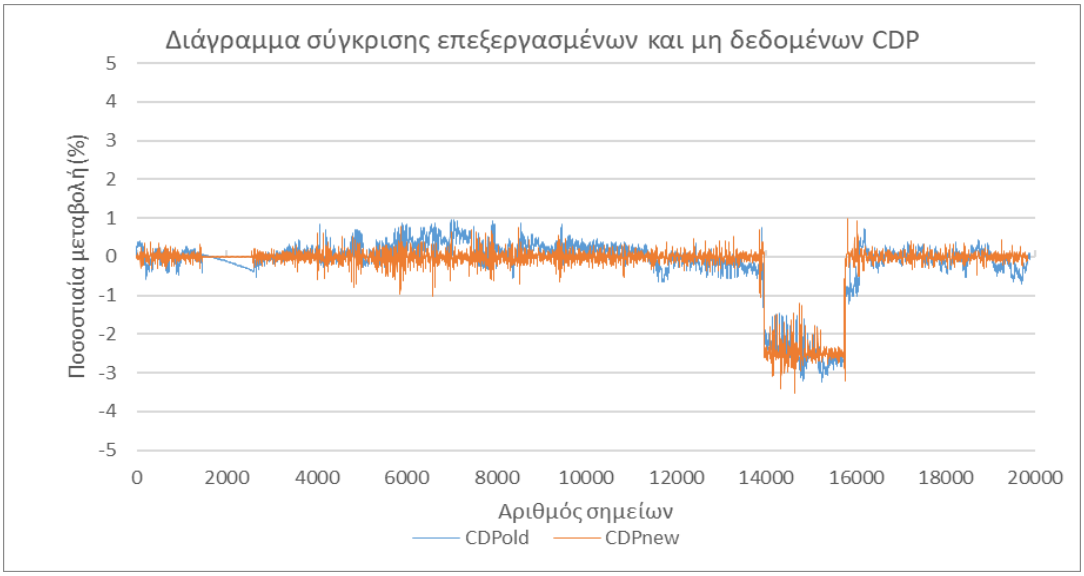


Σχήμα 4.54: Διάγραμμα κατανομής επεξεργασμένων δεδομένων σε Clusters

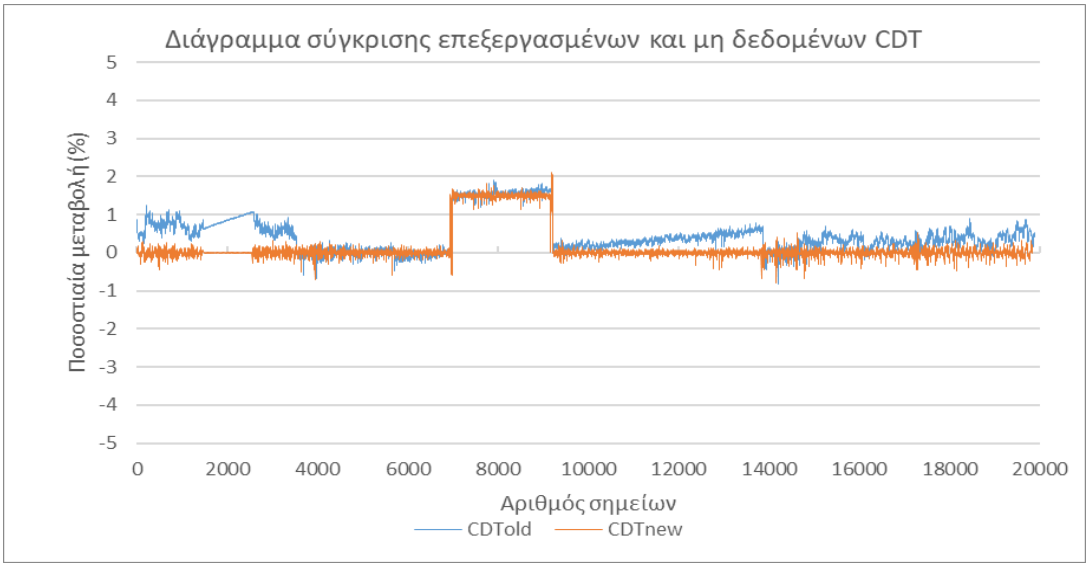
Όπου cluster #1 είναι το cluster υγείας και οι αριθμοί 2 μέχρι 11 περιγράφουν όλους τους συνδυασμούς βλάβης. Αναλυτικά όπως φαίνεται, υπάρχει βλάβη W, βλάβη CDT, βλάβη Wf, βλάβη CDP και EGT ταυτόχρονα, και βλάβη CDP όσο αυξάνεται ο αριθμός σημείων, βλάβες που συμβολίζονται με 2, 4, 5, 19 και 3 αντίστοιχα.
Ακολουθούν τα διαγράμματα διορθωμένων και μη μετρήσεων λόγω degradation για κάθε μεταβλητή που μελετήθηκε. Με πορτοκαλί χρώμα απεικονίζεται η τελικά επεξεργασμένη μορφή των δεδομένων μετά την αφαίρεση της υποβάθμισης. Όπως φαίνεται ο αλγόριθμος καλύπτει αρκετά καλά τα σημεία απλής υποβάθμισης αλλά και τα σημεία απότομης βλάβης. Το σημείο καθαρισμού προσδιορίζεται και πάλι σωστά ενώ ο αλγόριθμος αναγνωρίζει βλάβες στο 100%.



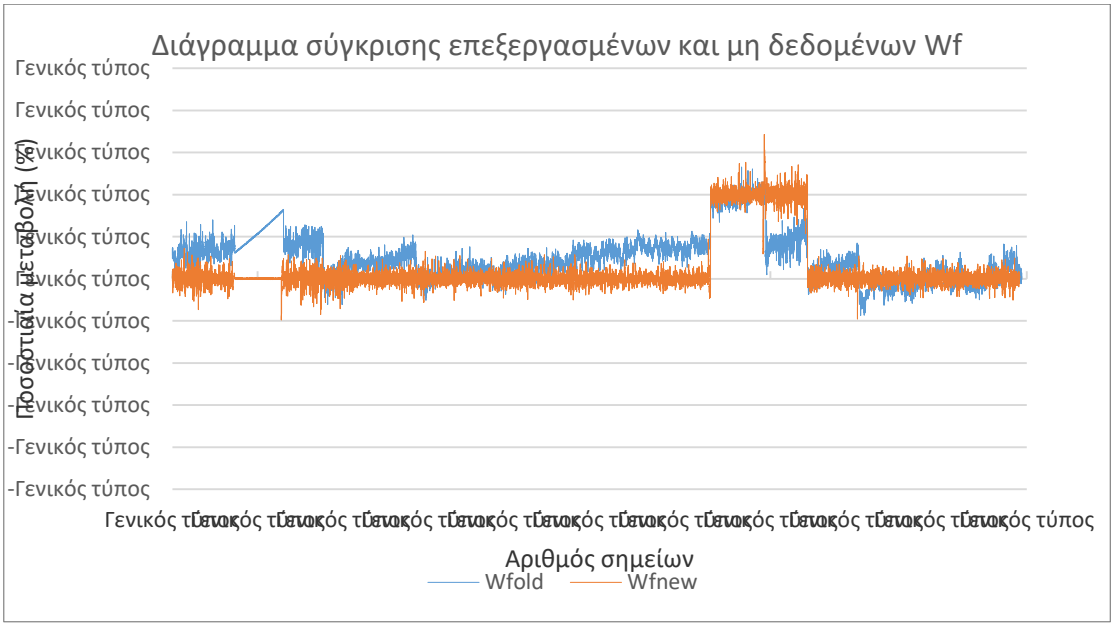
Σχήμα 4.55: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων W



Σχήμα 4.56: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων CDP



Σχήμα 4.57: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων CDT



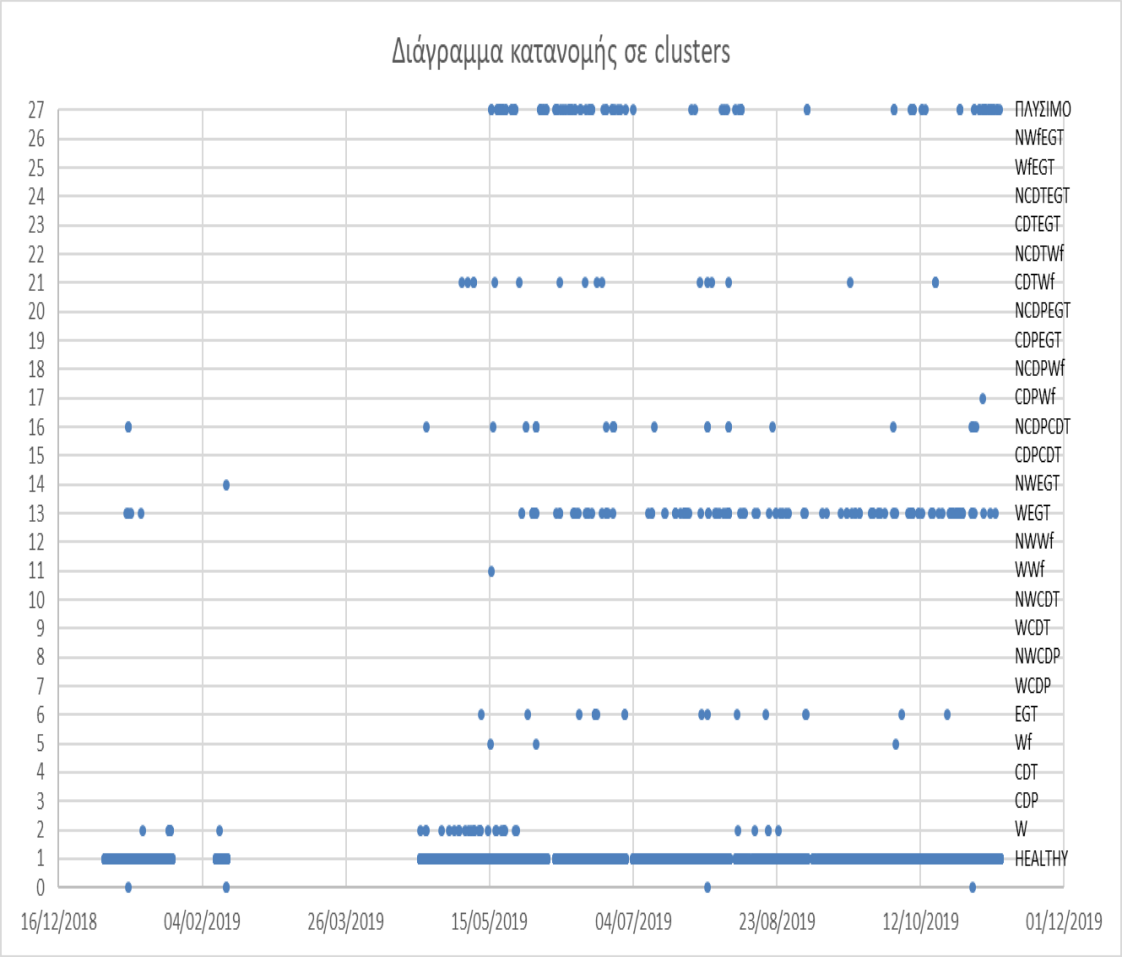
Σχήμα 4.58: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων Wf

4.5.2 Αεριοστρόβιλος GTB2

Στο τελευταίο κομμάτι της ανάλυσης πραγματικών δεδομένων μελετώνται οι κινητήρες GTB1 και GTB2 που χρησιμοποιούνται για την παραγωγή αλουμινίου και ηλεκτρικής ισχύος. Οι κινητήρες αυτοί παρακολουθούνται στη συγκεκριμένη ενότητα για τη χρονική περίοδο 2018-2019. Στο διάστημα αυτό οι κινητήρες δεν παρουσίασαν κάποιο σημαντικό πρόβλημα, παρόλα αυτά πραγματοποιήθηκαν αρκετοί καθαρισμοί σε Offline ή Online λειτουργία. Για αυτό λοιπόν αξίζει η μελέτη των δεδομένων που παρέχονται για την περαιτέρω διεκπεραίωση της ακρίβειας, πρώτα του φίλτρου υγείας-βλάβης, και σε

επόμενο στάδιο του φίλτρου εκχώρησης μίας μέτρησης με το Cluster καθαρισμού που ορίστηκε νωρίτερα.

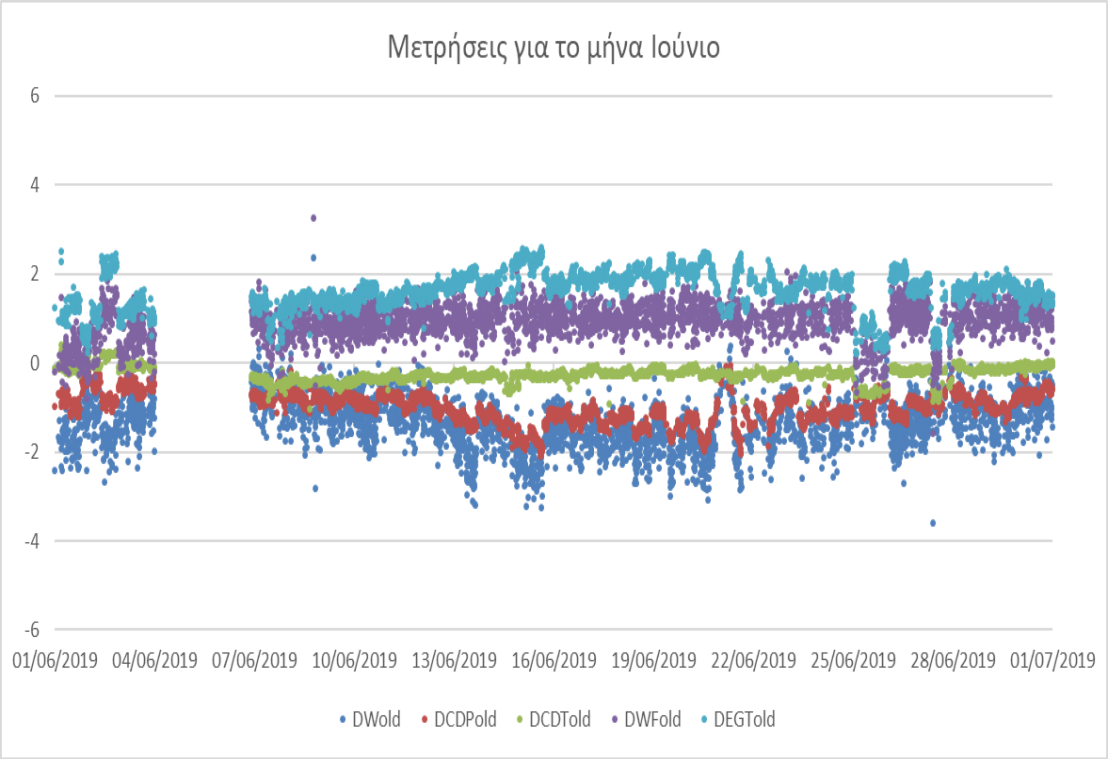
Πρώτα μελετήθηκαν οι μετρήσεις του κινητήρα GTB2 για κάθε φορτίο για όλο το έτος 2019 με συχνότητα δειγματοληψίας μία ανά 5 λεπτά. Μετά από τρέξιμο των γνωστών πλέον αλγορίθμων, το σύστημα παρέχει ως έξοδο το εξής σχήμα:



Σχήμα 4.59: Διάγραμμα κατανομής σε clusters

Στο σχήμα αυτό φαίνονται χρονολογικά οι μετρήσεις από Δεκέμβρη του 2018 έως Δεκέμβρη του 2019 όπως προαναφέρθηκε, ενώ στον κατακόρυφο άξονα φαίνεται ο αριθμός cluster εκχώρησης κάθε σημείου. Δεξιά φαίνεται η διάγνωση που προκύπτει για κάθε αριθμό cluster, όπως έγινε και στις προηγούμενες πραγματικές μετρήσεις. Με N-διάγνωση σηματοδοτούνται τα clusters συνδυασμένης βλάβης αντίρροπης κατεύθυνσης συντελεστών. Όπως φαίνεται λοιπόν η πλειοψηφία των μετρήσεων έχει εκχωρηθεί στο cluster υγείας ως αναμενόμενο, ενώ το δεύτερο δημοφιλέστερο cluster είναι το cluster καθαρισμού που περιλαμβάνει 110 από τα συνολικά 200 σημεία «βλάβης» που εκχώρησε ο αλγόριθμος, στα οποία θα εστιάσουμε αργότερα. Το τρίτο δημοφιλέστερο cluster είναι το cluster συνδυασμένης βλάβης W/EGT καθώς έχει όμοιες τιμές με αυτές του καθαρισμού και λόγω της συνάρτησης απόστασης που χρησιμοποιήθηκε ο αλγόριθμος εκχώρησε σημεία καθαρισμού ως σημεία βλάβης. Το ποσοστό συνολικής επιτυχίας των αλγορίθμων σε αυτή τη διαδικασία ήταν 100/20000 δηλαδή 99.5%. Κάνοντας μία

περαιτέρω ανάλυση στο μήνα Ιούνιο, μπορεί κανείς να παρατηρήσει τη διασπορά και τον λόγο εκχώρησης των διαφόρων δεδομένων στα clusters βλάβης.

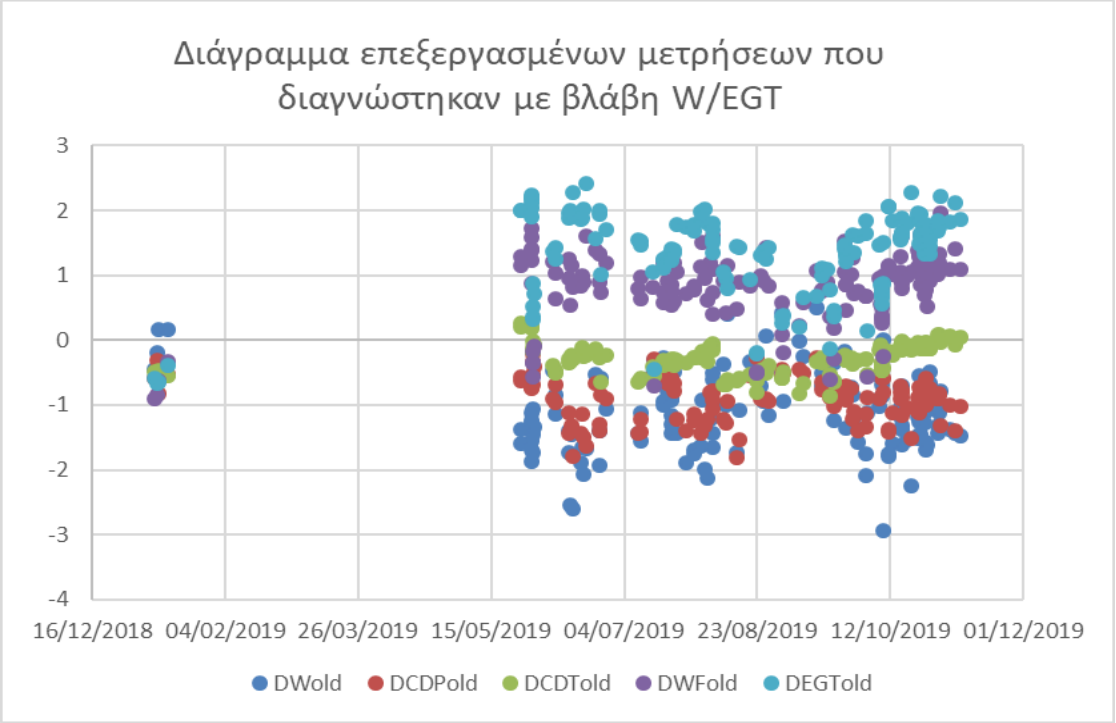


Σχήμα 4.60: Μετρήσεις για το μήνα Ιούνιο

Όπως φαίνεται και παραπάνω οι μετρήσεις του Ιουνίου παρέχουν κυρίως διαγνώσεις:

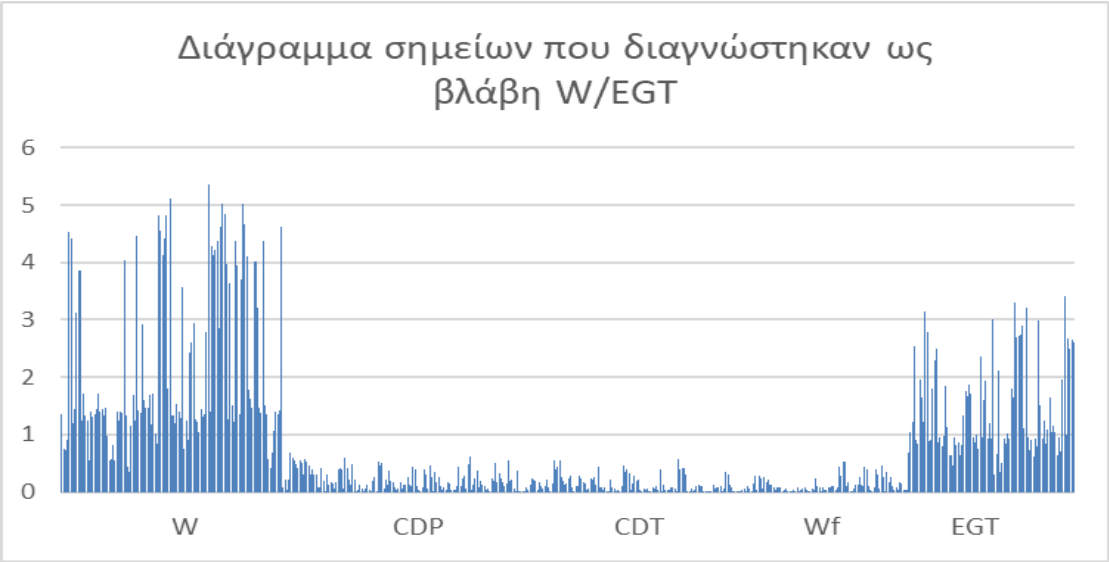
1. Υγείας
2. Πλυσίματος
3. Βλάβης W/EGT

Τα δύο παραπάνω διαγράμματα βοηθούν στην κατανόηση λειτουργίας του αλγορίθμου καθώς μπορεί κανείς να δει τα σημεία στα οποία αντιστοιχούν καθαρισμοί κατά τον αλγόριθμο τραβώντας κάθετη γραμμή μεταξύ των δύο διαγραμμάτων. Χαρακτηριστική είναι κυρίως η αυξομείωση των μεταβλητών W, EGT και CDP που αυξάνουν την ομοιότητα των εξεταζόμενων σημείων με τις υπογραφές καθαρισμού που έχουν παραχθεί. Ακόμη, είναι σημαντικό να μελετηθεί και το κύριο σφάλμα των αλγορίθμων, δηλαδή η εκχώρηση δεδομένων στο συνδυασμό βλάβης W/EGT που είδαμε νωρίτερα.



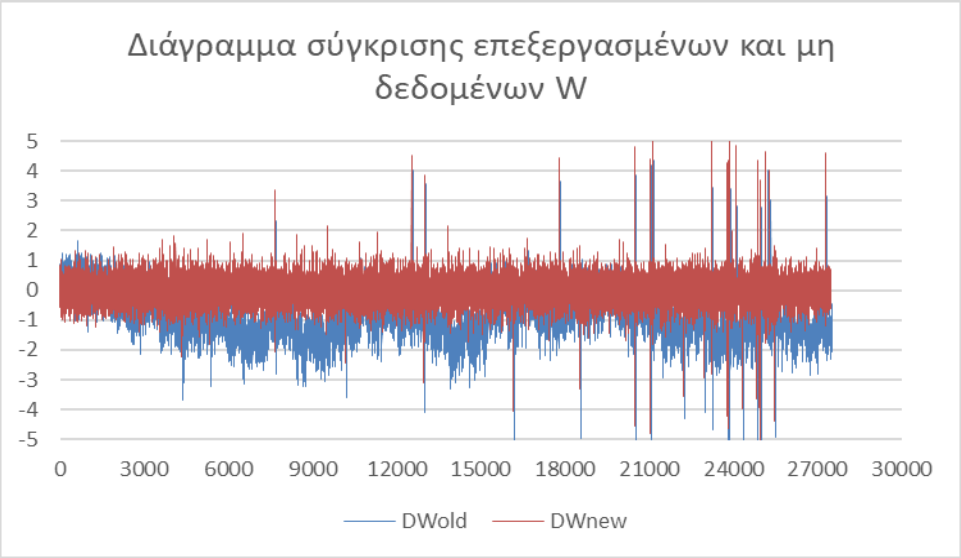
Σχήμα 4.61: Διάγραμμα επεξεργασμένων μετρήσεων που διαγνώστηκαν με βλάβη W/EGT

Τα παραπάνω διαγράμματα βοηθούν στην ανάλυση αυτή. Φαίνεται και στις πραγματικές αλλά και στις επεξεργασμένες μετρήσεις ότι οι ακραίες τιμές καταλαμβάνονται κυρίως από μετρήσεις W και EGT. Το φαινόμενο αυτό σε συνδυασμό με τις μικρές έως μηδενικές μετρήσεις των άλλων δύο μεταβλητών οδηγούν τους αλγόριθμους στη λανθασμένη εκχώρηση.

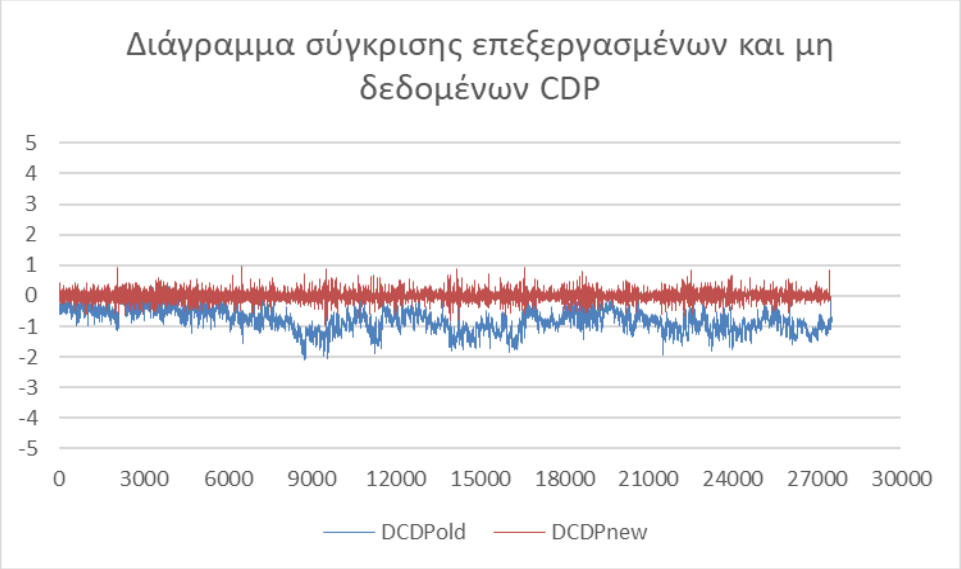


Σχήμα 4.62: Διάγραμμα επεξεργασμένων μετρήσεων που διαγνώστηκαν με βλάβη W/EGT

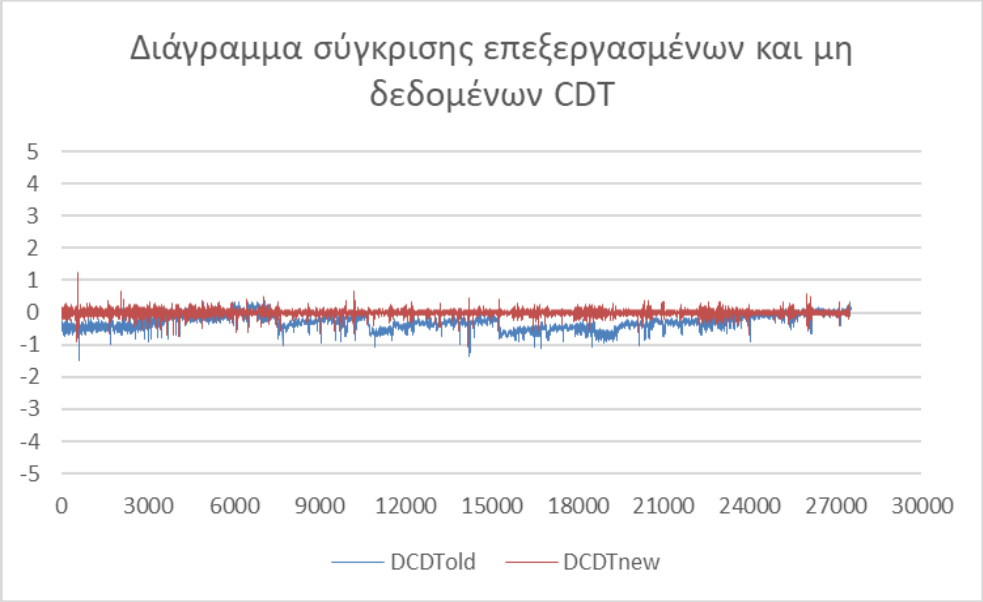
Με την παραπάνω γραφική αναπαράσταση γίνεται ακόμα πιο δικαιολογημένη η εκχώρηση σημείων στο cluster συνδυασμένης βλάβης αντί για αυτό του καθαρισμού. Τέλος, όπως και νωρίτερα, παρατίθενται τα διαγράμματα σύγκρισης επεξεργασμένων και μη μετρήσεων για κάθε μία από τις μεταβλητές.



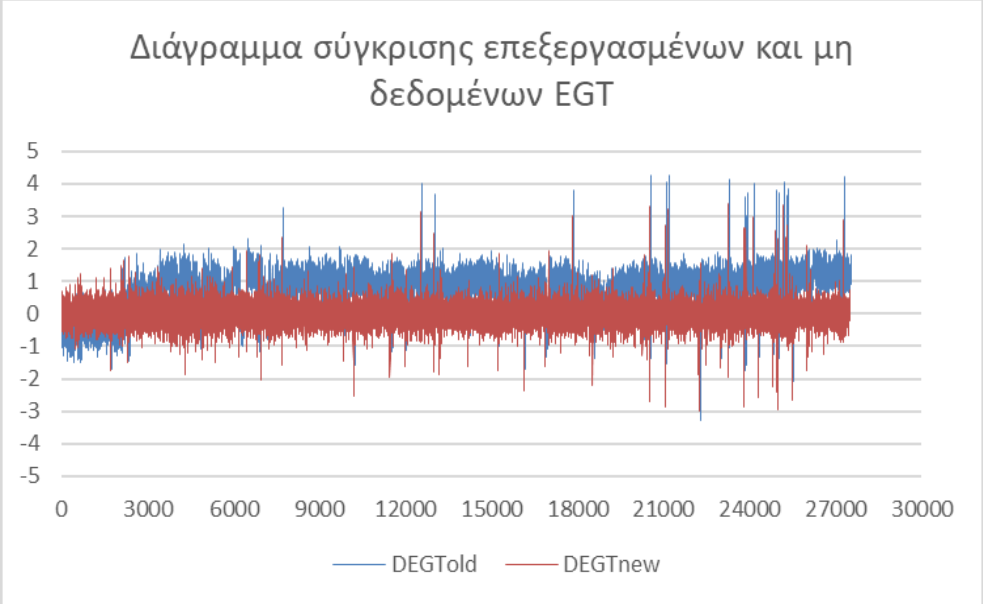
Σχήμα 4.63: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων W



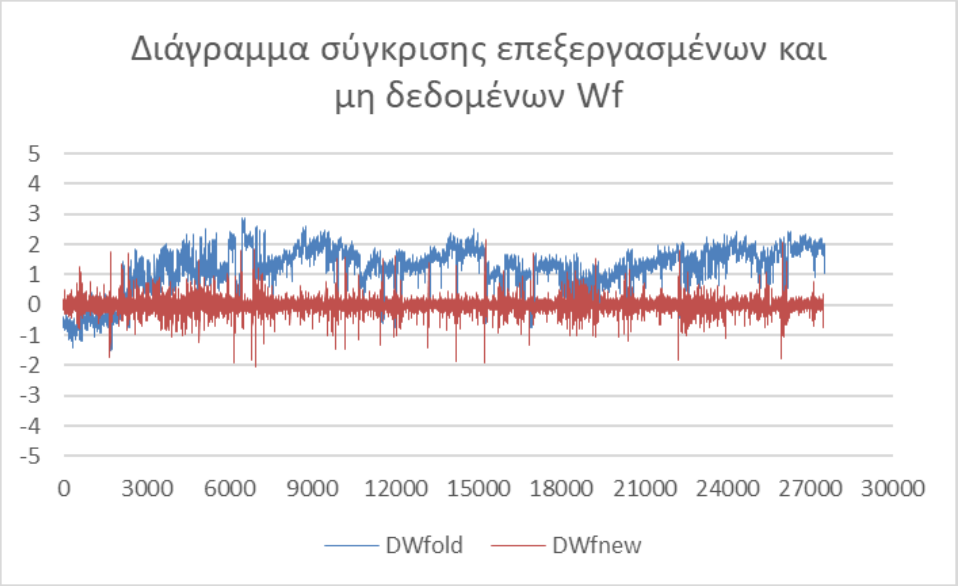
Σχήμα 4.64: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων CDP



Σχήμα 4.65: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων CDT



Σχήμα 4.66: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων EGT

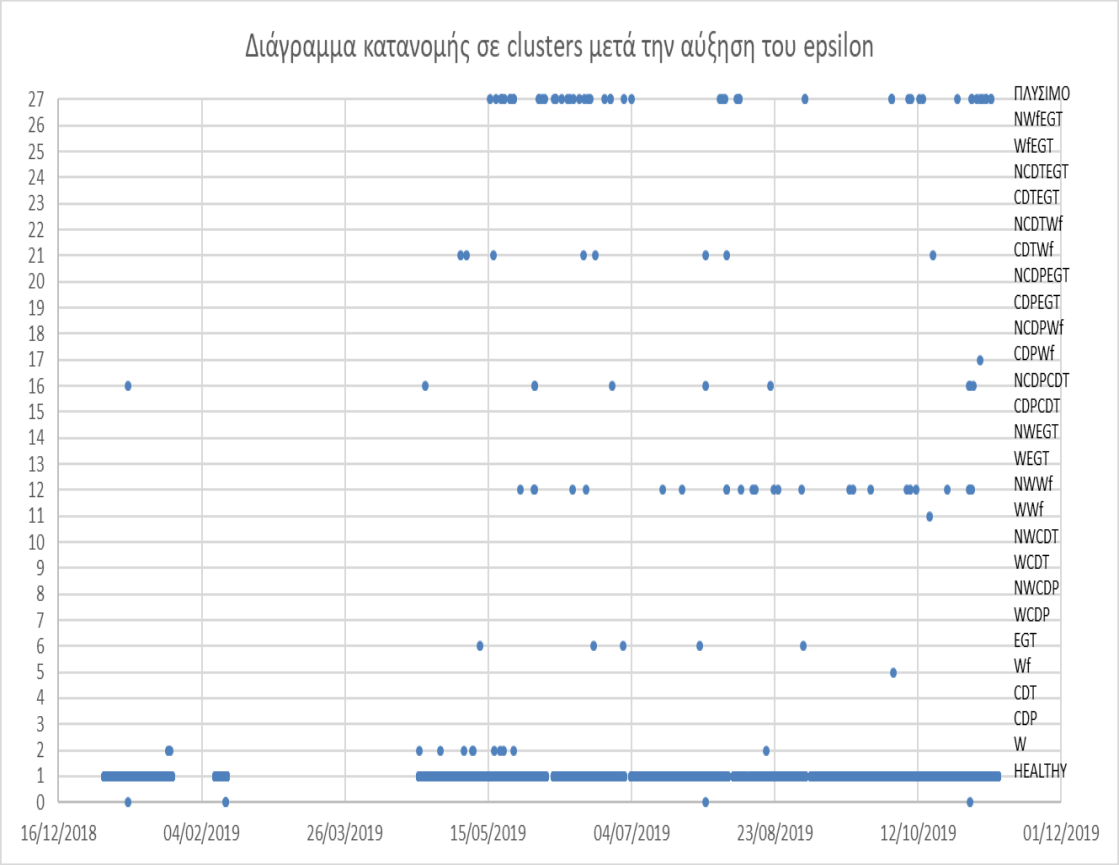


Σχήμα 4.67: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων Wf

Όπως φαίνεται, για άλλη μία φορά, το κομμάτι των αλγορίθμων που ρυθμίζει την υποβάθμιση λειτουργεί σωστά και επαναφέρει όχι όλα, αλλά όσα πρέπει, από τα δεδομένα που εισάχθηκαν. Οι απότομες μεταβολές παραμένουν απότομες δίνοντας έτσι σημαντική πληροφορία στον αλγόριθμο σύγκρισης υπογραφών για την κατάσταση κάθε μέτρησης.

Ακολουθεί μία ανάλυση παρόμοια με αυτή που μόλις μελετήθηκε αλλά για μεγαλύτερες τιμές μεταβλητών εισόδου. Μετά τα παραπάνω αποτελέσματα, είναι άξιο απορίας αν το διαγνωστικό αποτέλεσμα θα ήταν ακόμη καλύτερο αυξάνοντας τη μεταβλητή *epsilon*. Λόγος για τη μεταβλητή *MinPts* δε γίνεται καθώς παραπάνω αποδείχθηκε ότι δεν επηρεάζει το διαγνωστικό αποτέλεσμα.

Αυξάνοντας λοιπόν τη μεταβλητή *epsilon* ο αριθμός σημείων βλάβης μειώνεται, όπως είναι φυσικό. Το φαινόμενο αυτό όμως δε συνεπάγεται με την αύξηση αποδοτικότητας. Το παρακάτω διάγραμμα βοηθά στην περαιτέρω κατανόηση της παραπάνω δήλωσης:

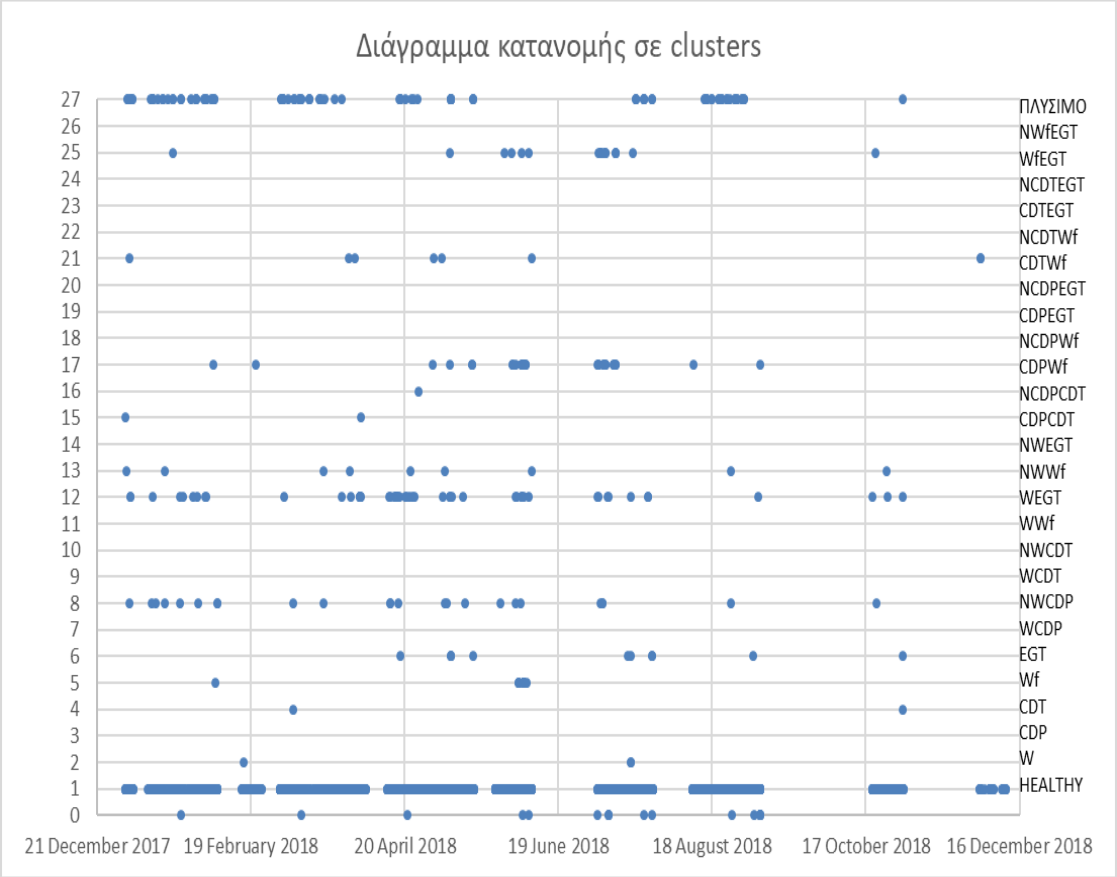


Σχήμα 4.68: Διάγραμμα κατανομής σε clusters μετά την αύξηση του epsilon

Τα αποτελέσματα όπως φαίνεται είναι θετικότερα όσον αφορά τα σημεία βλάβης. Τα σημεία υγείας αποτελούν το 99.8%, ποσοστό αυξημένο σε σχέση με την προηγούμενη ανάλυση. Παράλληλα όμως, έχει μειωθεί και το ποσοστό σημείων καθαρισμού. Η αύξηση του epsilon πρακτικά κανονικοποίησε κάποια σημεία που οριακά κατατάχθηκαν ως σημεία βλάβης στο πρώτο φίλτρο της προηγούμενης ανάλυσης. Το φαινόμενο αυτό είχε ως αποτέλεσμα να μειωθούν τα σημεία βλάβης μεν, που λανθασμένα κατατάχθηκαν ως βλάβη στην προηγούμενη ανάλυση καθώς αργότερα αποδείχθηκε ότι πρόκειται για σύνολο δεδομένων κανονικής λειτουργίας, ενώ παράλληλα να μειωθούν και τα πιθανά σημεία καθαρισμού διότι και ο καθαρισμός του κινητήρα όπως περιεγράφηκε νωρίτερα μοντελοποιήθηκε ως ένα είδος βλάβης-απότομης μεταβολής.

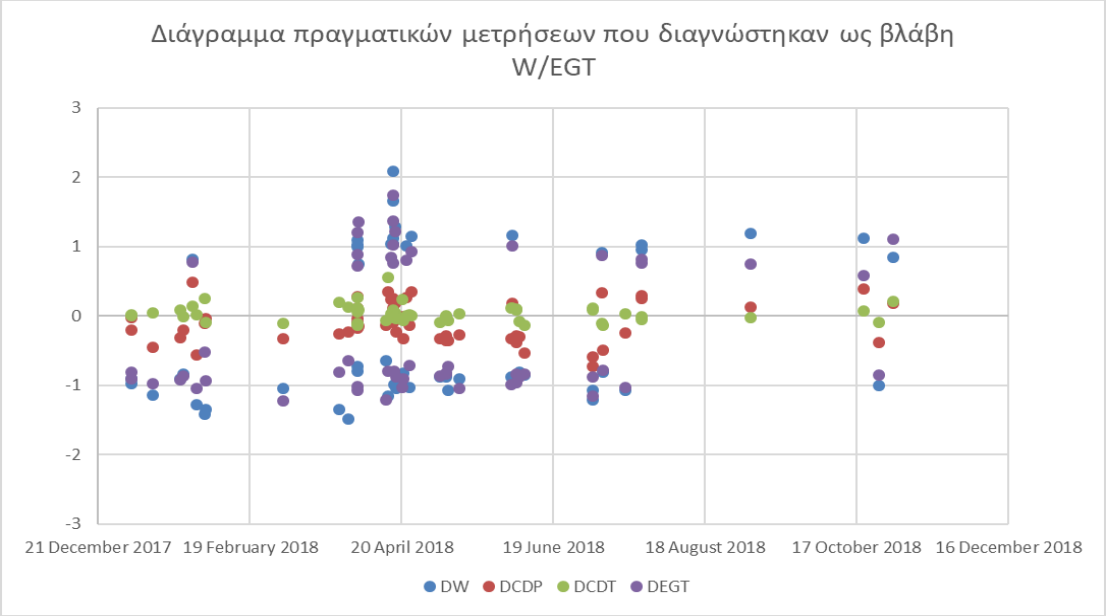
Νέα δεδομένα αεριοστρόβιλου GTB2

Στη συνέχεια γίνεται η ίδια ανάλυση για δεδομένα του κινητήρα GTB2 τη χρονική περίοδο του 2018 και αυτή τη φορά εισάγοντας μόνο τις μετρήσεις πλήρους φορτίου. Οι μετρήσεις αυτές αφορούν λειτουργία όπου τα IGVs είναι πλήρως ανοιχτά επιτυγχάνοντας έτσι μέγιστη ώση ή αλλιώς πλήρες φορτίο.

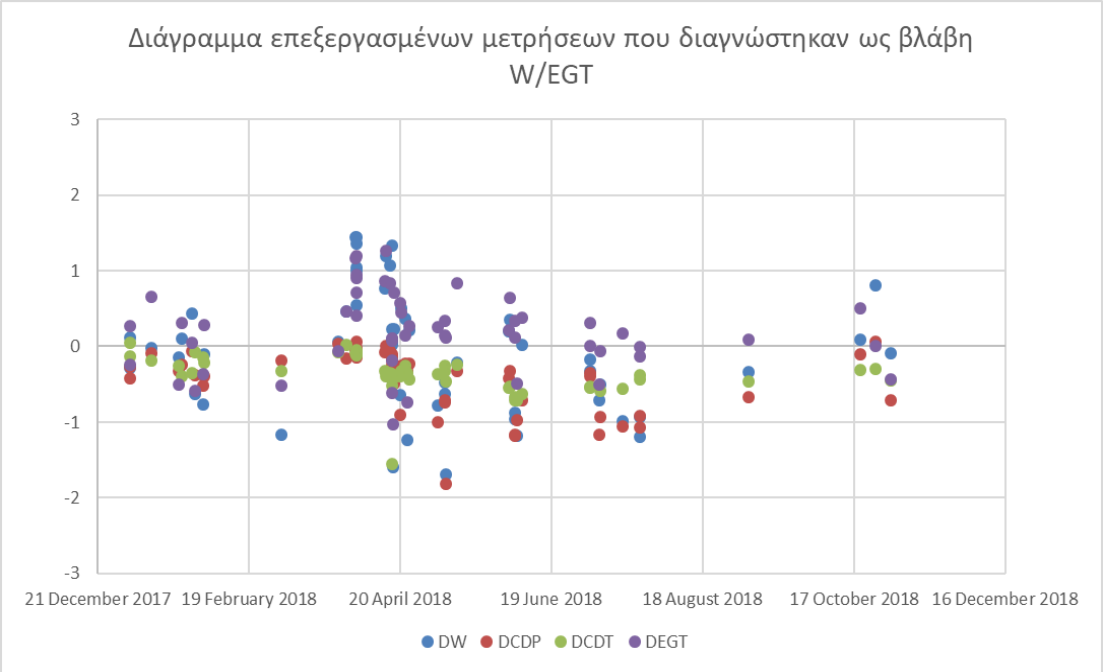


Σχήμα 4.69: Διάγραμμα κατανομής σε clusters

Το διάγραμμα κατανομής των συνολικών φορτίων είναι παρόμοιο με το προηγούμενο. Παρατηρούνται μερικά false alarm καθ' όλη τη διάρκεια του χρόνου, συγκεκριμένα γύρω στα 100, αλλά επίσης και αρκετά σημεία καθαρισμού καθώς είναι η δημοφιλέστερη «βλάβη». Τα αποτελέσματα δε διαφέρουν σχεδόν καθόλου με τα προηγούμενα, και αυτό φαίνεται και στα επόμενα διαγράμματα.

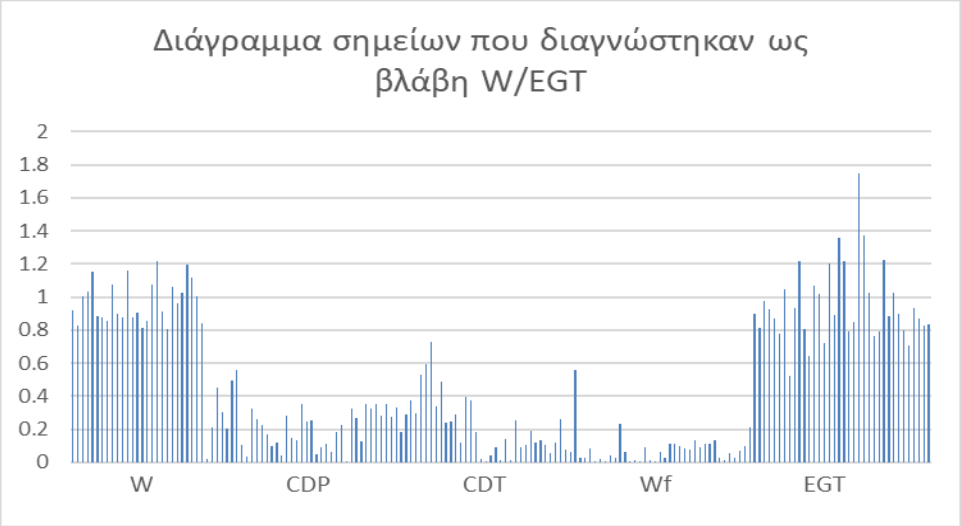


Σχήμα 4.70: Διάγραμμα πραγματικών μετρήσεων που διαγνώστηκαν ως βλάβη W/EGT



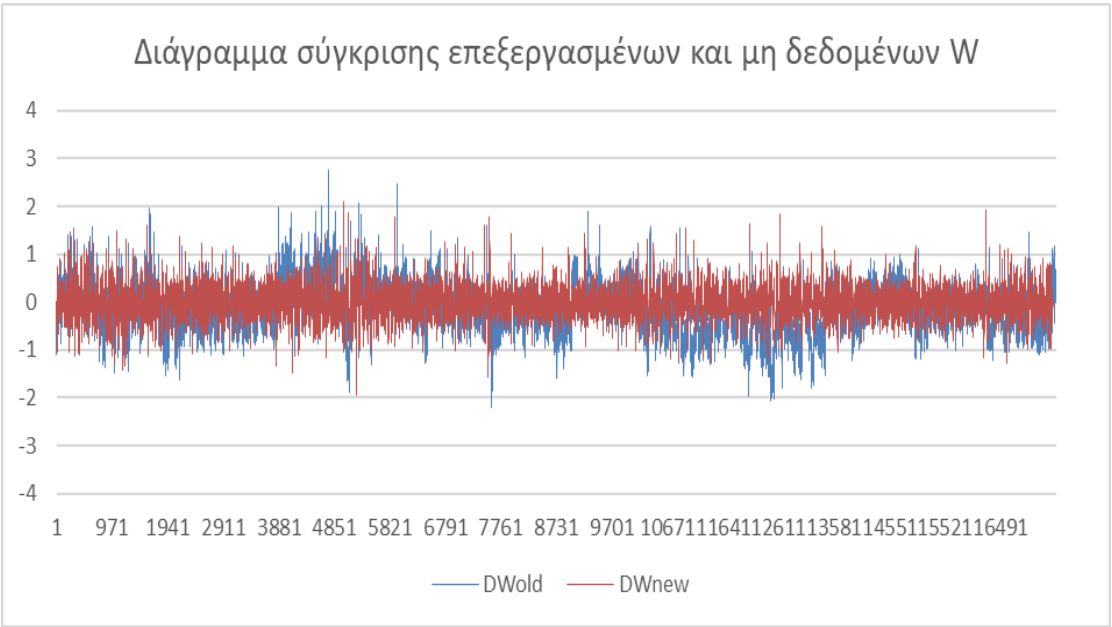
Σχήμα 4.71: Διάγραμμα επεξεργασμένων μετρήσεων που διαγνώστηκαν ως βλάβη W/EGT

Τα σημεία βλάβης W/EGT που αποτελούν τη δεύτερη σημαντικότερη βλάβη έχουν παρόμοια κατανομή μέσα στο χρόνο, με τις τιμές των προαναφερθέντων μεταβλητών να ξεχωρίζουν από αυτές των CDP και CDT.

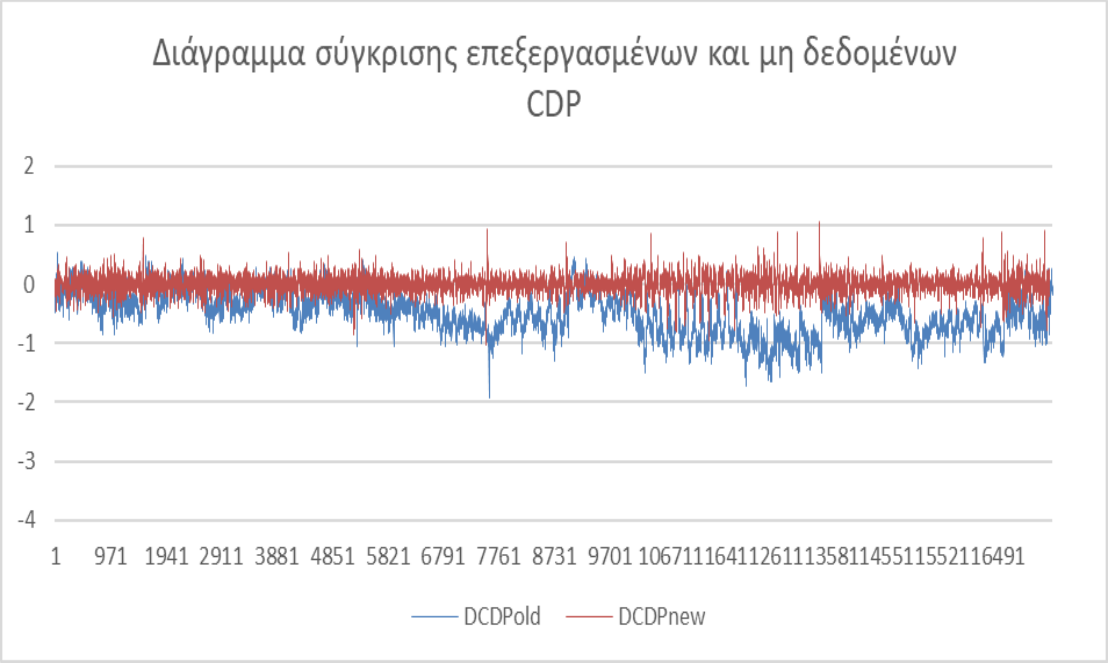


Σχήμα 4.72: Διάγραμμα σημείων που διαγνώστηκαν ως βλάβη W/EGT

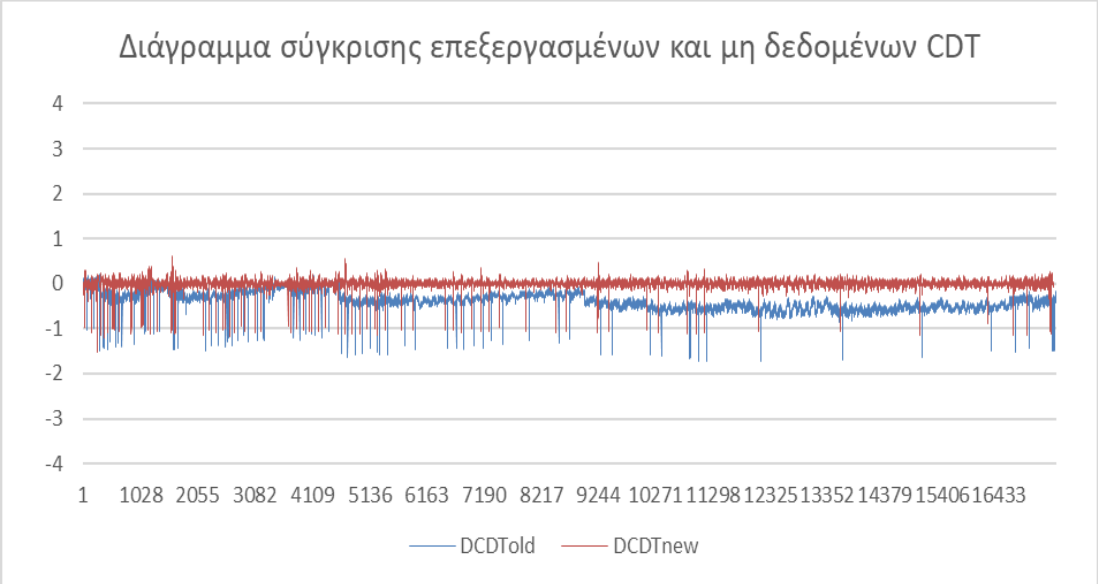
Οι παρατηρήσεις στο εξής διάγραμμα δε διαφέρουν με το προηγούμενο της GTB2 για το 2019. Τέλος, ακολουθούν τα διαγράμματα σύγκρισης επεξεργασμένων και μη δεδομένων για όλες τις συνιστώσες.



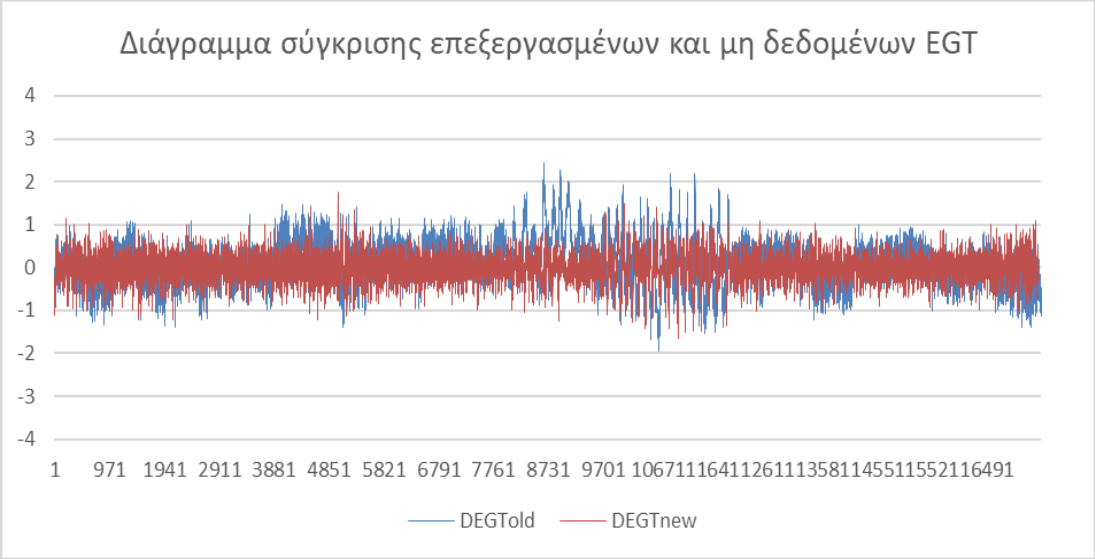
Σχήμα 4.73: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων W



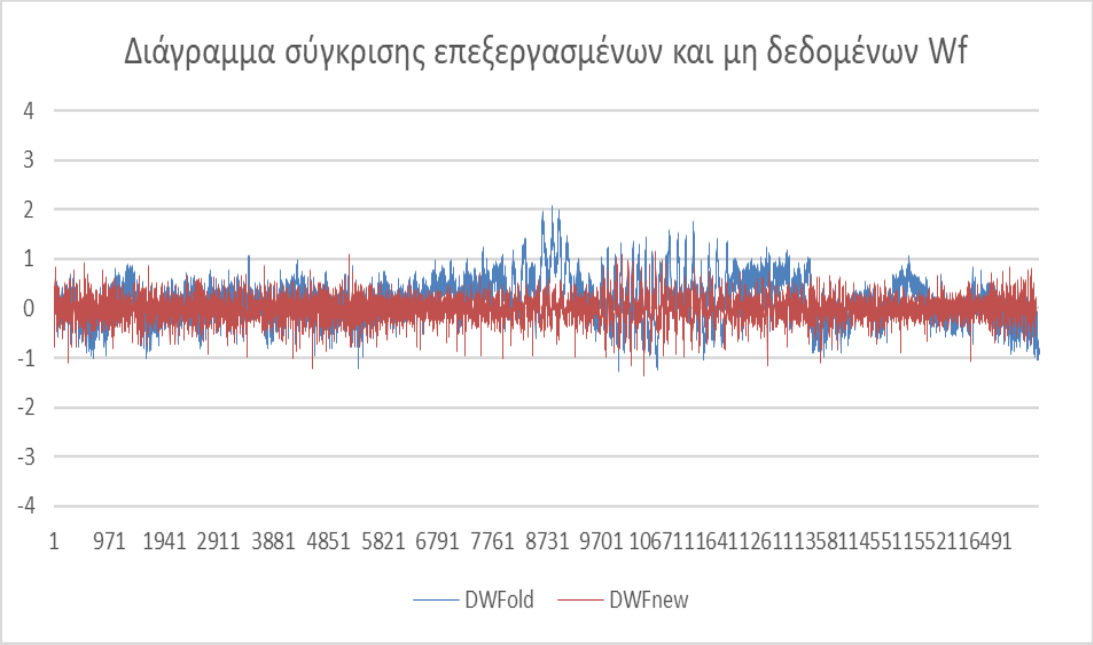
Σχήμα 4.74: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων CDP



Σχήμα 4.75: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων CDT



Σχήμα 4.76: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων EGT

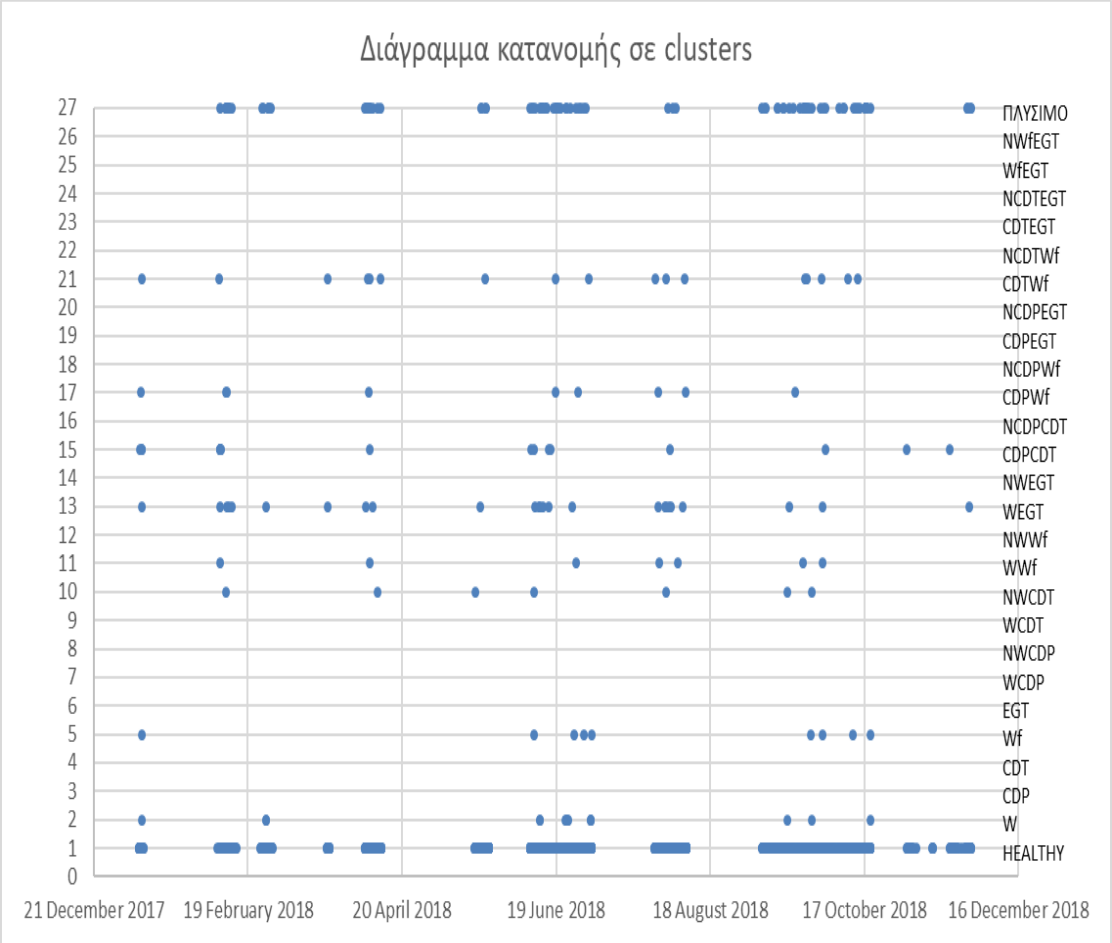


Σχήμα 4.77: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων Wf

Τα αποτελέσματα ως αναμενόμενο έχουν σωστή διόρθωση υποβάθμισης.

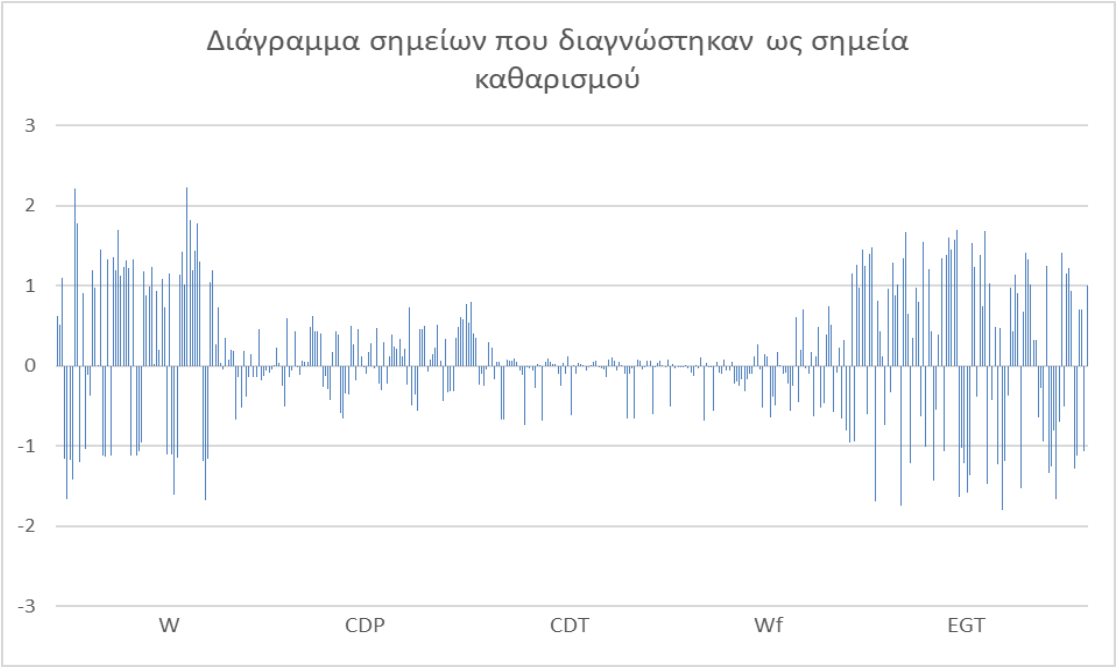
4.5.3 Αεριοστρόβιλος GTB1

Τέλος, γίνεται η ίδια ανάλυση για δεδομένα του κινητήρα GTB1 τη χρονική περίοδο του 2018 εισάγοντας πάλι μόνο τις μετρήσεις πλήρους φορτίου. Οι μετρήσεις αυτές αφορούν λειτουργία όπου τα IGVs είναι πλήρως ανοιχτά επιτυγχάνοντας έτσι μέγιστη ώση ή αλλιώς πλήρες φορτίο.



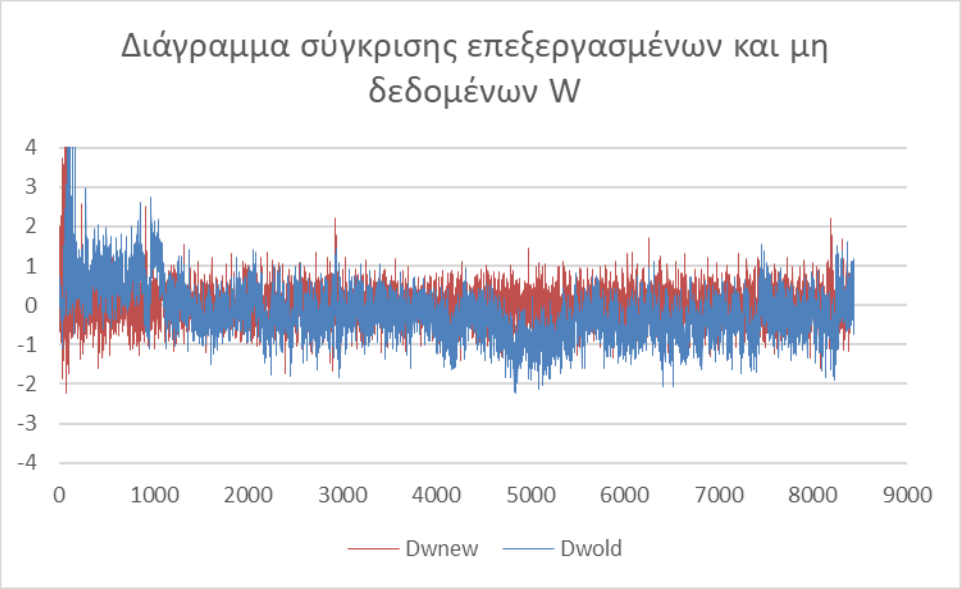
Σχήμα 4.78: Διάγραμμα κατανομής σε clusters

Παρόλες τις διακοπές λειτουργίας του κινητήρα, οι μετρήσεις δεν είχαν αρκετές ανωμαλίες και το αποτέλεσμα ήταν κοινό.

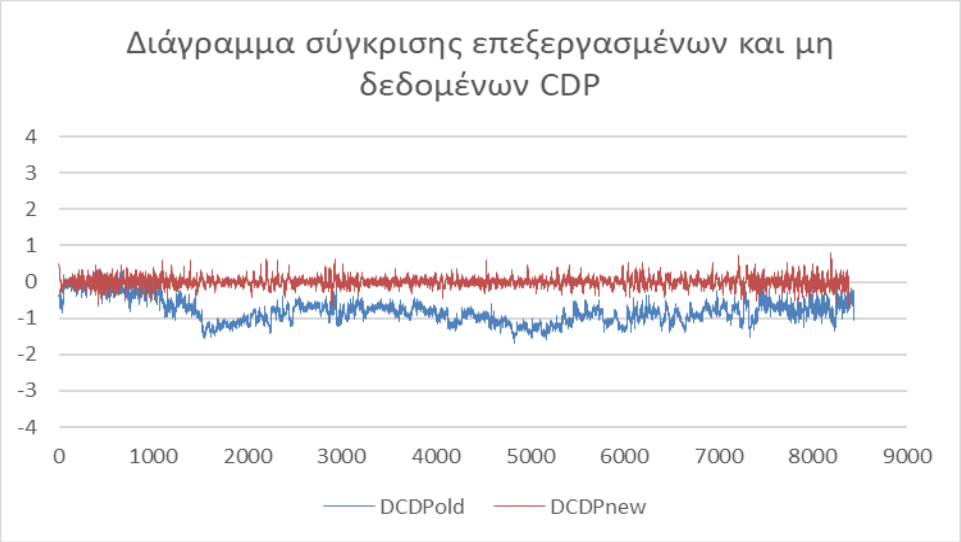


Σχήμα 4.79: Διάγραμμα σημείων διαγνώστηκαν ως σημεία καθαρισμού

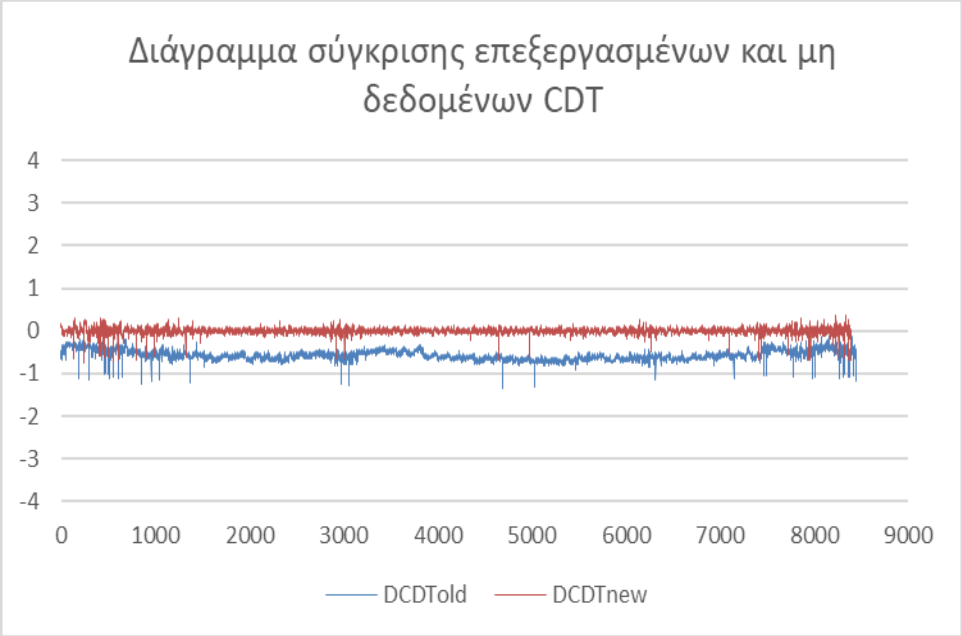
Στο συγκεκριμένο παράδειγμα διατηρήθηκαν και οι αρνητικές τιμές για να αποδειχθεί ότι ο αλγόριθμος ξεχωρίζει τις μετρήσεις βλαβών ανεξαρτήτως πρόσημου. Εξίσου σημαντική είναι η παρατήρηση των τιμών CDP και CDT όπως είδαμε και νωρίτερα, καθώς για να καταταχθούν αυτές οι μετρήσεις σε αυτή τη βλάβη σημαίνει ότι δεν βρήκαν αντιστοιχία με τα cluster καθαρισμού ή υγείας. Τέλος, ως συνήθως ακολουθούν τα διαγράμματα διόρθωσης της υποβάθμισης όλων των μεταβλητών που μελετώνται.



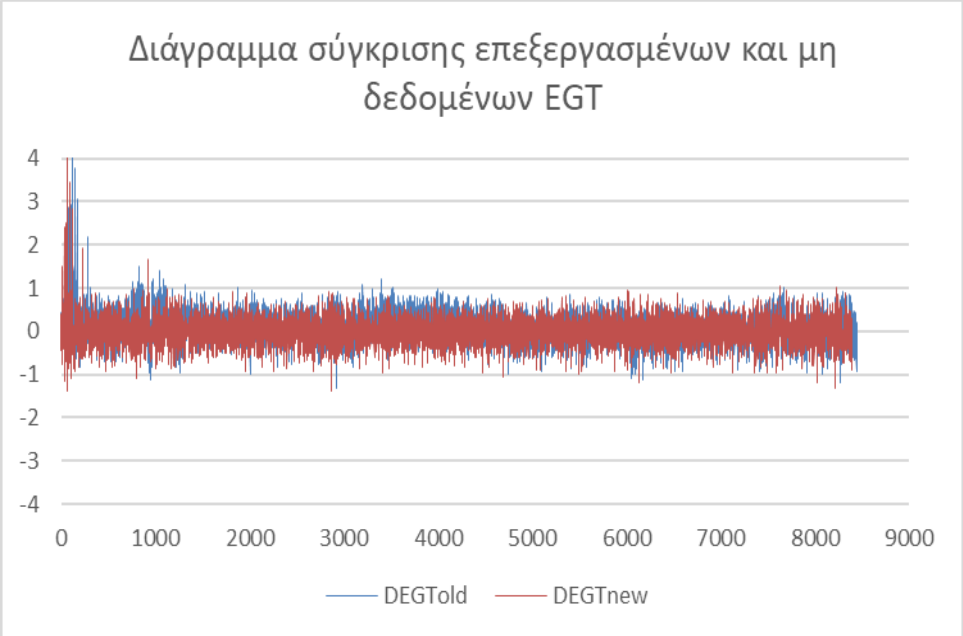
Σχήμα 4.80: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων W



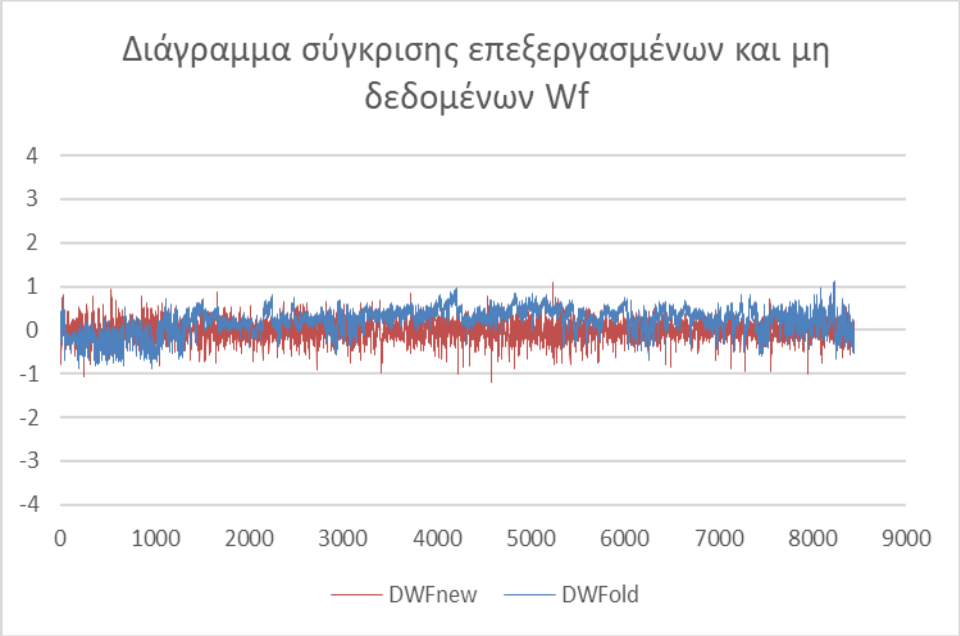
Σχήμα 4.81: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων CDP



Σχήμα 4.82: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων CDT



Σχήμα 4.83: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων EGT



Σχήμα 4.84: Διάγραμμα σύγκρισης επεξεργασμένων και μη δεδομένων Wf

Προφανώς τα αποτελέσματα ήταν κοινά με αυτά των προηγούμενων δύο αναλύσεων. Οι μέσες τιμές κάθε δείγματος είναι κοντά στο μηδέν και οι ακραίες μεταβολές στις μεταβλητές μας δεν μηδενίζονται ώστε οι αλγόριθμοι να μπορούν να κατατάξουν ανάλογα με τα χαρακτηριστικά κάθε μέτρησης, τα δεδομένα στις αντίστοιχες βλάβες που τους ταιριάζουν.

5 Ανακεφαλαίωση-Συμπεράσματα-Προτάσεις

Σε αυτό το κεφάλαιο γίνεται μία σύντομη ανακεφαλαίωση όσων αναλύθηκαν στις παραπάνω ενότητες, διεξάγονται συμπεράσματα για τα αποτελέσματα των μεθόδων στα διάφορα τεχνητά ή πραγματικά δεδομένα και απαριθμούνται προτάσεις για μελλοντική έρευνα πάνω στο αντικείμενο.

5.1 Ανακεφαλαίωση

Στα πλαίσια της παρούσας διπλωματικής εργασίας πραγματοποιήθηκε μελέτη διαφόρων μεθόδων Clustering, ~~έγινε~~ ανάπτυξη μερικών από ~~αυτών~~ αυτές, σύγκριση αποτελεσμάτων μεταξύ τους και εξαγωγή συμπερασμάτων βάσει της απόδοσης κάθε μεθόδου.

Αρχικά, πραγματοποιήθηκε εκτενής βιβλιογραφική ανασκόπηση σχετικά με διάφορες μεθόδους πρόγνωσης που χρησιμοποιούνται σήμερα στη Διαγνωστικής αλλά και σε άλλα επιστημονικά πεδία. Η ανάλυση επικεντρώθηκε αργότερα σε μεθόδους που αφορούν τη διάγνωση για επίγειους αεριοστρόβιλους.

Διαλέγοντας τις μεθόδους που θεωρούνται ιδανικότερες για τη μελέτη που ακολούθησε, ξεκινά η περιγραφή της ιστορίας και των βημάτων των αλγορίθμων τους. Εφόσον πρόκειται για επαναληπτικές μεθόδους, δόθηκε ~~βάση~~ έμφαση στις διαδικασίες που απαρτίζουν μία επανάληψη και στα κριτήρια σύγκλισης και ομοιότητας. Παράλληλα Αναφέρονται τα πλεονεκτήματα και τα μειονεκτήματα των μεθόδων βάσει της αρχής λειτουργίας τους και συγκρίνονται μεταξύ τους. Στη συνέχεια οι μέθοδοι υποβλήθηκαν σε δύο τεστ επικύρωσης της διαγνωστικής τους ικανότητας. Τα τεστ αυτά έλεγχαν μεταξύ άλλων την ικανότητα διαχωρισμού: ομόκεντρων clusters και clusters διαφορετικής πυκνότητας. Τα καλύτερα αποτελέσματα είχε η DBSCAN, ακολουθούσε η K-means και τέλος η AHC. Παράλληλα, αναλύονται τα είδη συναρτήσεων απόστασης που χρησιμοποιήθηκαν, η Ευκλείδεια απόσταση και ο Συντελεστής Αλληλοσυσχέτισης. Περιγράφονται τα διάφορα είδη καθώς και τα πεδία εφαρμογής τους στο σήμερα. Ακολούθησε επιγραμματική περιγραφή των βημάτων υπολογισμού τους και αιτιολόγηση διαλογής των δύο τελικά συναρτήσεων απόστασης που αναφέρθηκαν παραπάνω.

Οι μέθοδοι χρησιμοποιήθηκαν για δύο στόχους, τη δημιουργία προφίλ αναφοράς θερμοκρασίας εξόδου καυσαερίων και τη διάγνωση βλαβών. Πρώτα χρησιμοποιούνται για ομαδοποίηση προφίλ λειτουργίας βιομηχανικών αεριοστρόβιλων καθώς υπάρχει

δυνατότητα χρήσης των προφίλ ως πρότυπες τιμές λειτουργίας οι οποίες μπορούν έπειτα να συγκριθούν με μετρήσεις που παράγονται εν ώρα λειτουργίας και να προσφέρουν διαγνωστικές πληροφορίες για την υγεία του κινητήρα. Η διαδικασία αυτή εφαρμόστηκε σε πραγματικά δεδομένα τριών αεριοστροβίλων, τον GTA και τους GTB1 και GTB2 οι οποίοι είναι ίδιοι κινητήρες και λειτουργούν στην ίδια εγκατάσταση.

Για το δεύτερο στόχο, αναπτύσσεται η μέθοδος διάγνωσης βλαβών και ελέγχεται η διαγνωστική ικανότητα σε προσομοιωμένα δεδομένα απότομης και σταδιακής μεταβολής. Ακολούθως, οι μέθοδοι δοκιμάζονται σε πραγματικά δεδομένα των ίδιων τριών αεριοστροβίλων, παράγοντας χρήσιμα αποτελέσματα και συμπεράσματα.

5.2 Συμπεράσματα

Σύμφωνα με τα αποτελέσματα των δυο λειτουργιών που έλαβαν χώρα μπορούν πλέον να εξαχθούν τα παρακάτω συμπεράσματα:

- Οι μέθοδοι clustering αποτελούν χρήσιμο εργαλείο για την δημιουργία προφίλ αναφοράς. Ομαδοποιώντας τα δεδομένα βάσει των τιμών των θερμοστοιχείων, και όχι βάσει της διορθωμένης ισχύος, εξασφαλίζεται ότι τα σημεία με αυξημένη ομοιότητα μεταξύ τους θα καταλήξουν στο ίδιο cluster, ενώ παράλληλα τα συνολικά clusters που σχηματίζονται θα παραμένουν ανεξάρτητα μεταξύ τους.
- Τα αδιάστατα προφίλ και ο CCD παρέχουν χρήσιμες πληροφορίες για την ανεξαρτησία των clusters που σχηματίζονται. Βάσει αυτών των δύο εργαλείων ο χρήστης μπορεί να συμπεράνει αν τα clusters που σχηματίζει ο αλγόριθμος σαν αποτέλεσμα είναι πραγματικά ανεξάρτητα ή όχι.
- Από τα αποτελέσματα προκύπτει ότι τα προφίλ αναφοράς μεταβάλλονται ακόμα και σε μηχανές ίδιου τύπου όπως επίσης μεταβάλλονται όταν στην ίδια μηχανή γίνει κάποια επέμβαση όπως ένα hot section inspection.
- Η μέθοδος διάγνωσης που αναπτύχθηκε χρησιμοποιεί υπογραφές ανεξάρτητες από τη μηχανή. Ο χρήστης έχει τη δυνατότητα να συγκρίνει δεδομένα οποιασδήποτε μηχανής με τις υπογραφές αυτές παρέχοντας το ίδιο καλό διαγνωστικό αποτέλεσμα.
- Είναι απαραίτητη η αφαίρεση της υποβάθμισης για καλύτερη διαγνωστική ικανότητα. Σε πολλές περιπτώσεις φάνηκε ότι η μη χρήση της αφαίρεσης υποβάθμισης οδηγεί σε αποτελέσματα με χαμηλότερη επιτυχία.
- Για τα προσομοιωμένα δεδομένα, παρατηρήθηκε ότι ο αλγόριθμος αναγνώριζε βλάβη για step change μεταβολή μεγαλύτερη του 0.9% και για σταδιακή μεταβολή αν αυτή δεν ήταν πολύ ομαλή.
- Κατατάσσοντας τις μεθόδους που αναπτύχθηκαν βάσει της απόδοσης τους, η DBSCAN είχε τη μεγαλύτερη επιτυχία. Ακολούθησε η K-means ενώ η AHC, παρόλο που ξεπέρασε τις δυσκολίες των πολλαπλών βλαβών, δεν κατάφερε λόγω της δομής της να κατανέμει τα νέα σημεία υγείας στο υγιές Cluster μετά τον καθαρισμό της μηχανής.

5.3 Προτάσεις

Για τη βελτίωση και περαιτέρω ανάπτυξη της παρούσας ανάλυσης προτείνονται οι παρακάτω κατευθύνσεις:

- Ανάπτυξη νέων μεθόδων clustering όπως τη δημοφιλή μέθοδο K-Nearest Neighbour και μελέτη των αποτελεσμάτων τους σε παρόμοιες συνθήκες με αυτές των ήδη αναλυμένων μεθόδων. Άλλες γνωστές μέθοδοι Classification που χρησιμοποιούνται σε άλλους τομείς αλλά δεν είχαν ικανοποιητική εφαρμογή στη Διαγνωστική ήταν οι: 1.) Μέθοδος προσέγγισης-μεγιστοποίησης μέσω Γκαουσσιανών μοντέλων μίξης και 2.) Η μέθοδος μέσων μεταβολών.
- Για τη διάγνωση βλαβών, βελτιστοποίηση των ήδη υπαρχόντων μεθόδων. Αυτό μπορεί να συμβεί μέσω ανάλυσης λειτουργίας των μεθόδων βήμα-βήμα για ελαχιστοποίηση των πράξεων του αλγόριθμου ή μέσω εισαγωγής νέων κριτηρίων κατανομής των σημείων στα διάφορα Clusters για πρόγνωση ακόμα πιο κοντά στην πραγματική κατάσταση. Ένας άλλος τρόπος βελτίωσης είναι η χρήση περισσότερων πολυπλοκότερων υπογραφών βλάβης για περισσότερες μεταβλητές ή/και για περισσότερους συνδυασμούς αυτών.
- Για την ομαδοποίηση των προφίλ λειτουργίας προτείνεται ανάπτυξη αυτόματης μεθόδου μείωσης των απαιτούμενων Clusters ανάλογα με την ικανοποίηση του κριτηρίου ομοιότητας. Η αυτόματη αυτή μέθοδος είναι σημαντικό να διαχειρίζεται τα σημεία θορύβου και τα Clusters λίγων σημείων με ακρίβεια αντίστοιχη της κριτικής ικανότητας ενός Μηχανικού με σκοπό την απόφαση αναφορικά με το αν θα συγχωνευτούν με άλλα προφίλ αναφοράς ή όχι. Ως αποτέλεσμα ο έλεγχος βλαβών αισθητήρων θα επιτυγχάνεται μέσω της σύγκρισης των προφίλ αναφοράς με αυτό που παράγει κατά τη λειτουργία ο κινητήρας.

6 Βιβλιογραφία

Ξενόγλωσση Βιβλιογραφία

[1] Chandola, V., Banerjee, A., and Kumar, V. 2009. Anomaly detection: A survey. ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages. DOI = 10.1145/1541880.1541882.

[2] Li L., and Hansman R. J. 2013. Anomaly detection in airline routine operations using flight data recorder data. MIT, Report No. ICAT-2013-4, (June 2013), 147 pages.

[3] Andrenelli L., Paone N., Rossi G., Tomasini E. P., 1991. "Non-Itrusive Measurement of Blade Tip Vibration in Turbomachines", ASME Paper 91-GT-301.

[4] Wu, J.; Wu, C.; Cao, S.; Or, W.; Deng, C.; Shao, X. Degradation Data-Driven Time-To-Failure Prognostics Approach for Rolling Element Bearings in Electrical Machines. IEEE Trans. Ind. Electron. 2018, 66, 529–539.

[5] X.J. Luo, K.F. Fong, Y.J. Sun, M.K.H. Leung Development of clustering-based sensor fault detection and diagnosis strategy for chilled water system. Energy & Buildings 186 (2019) 17-36, 20 pages.

[6] F. Khan, O.F. Eker, A.Khan, W. Orfali, 2018. Adaptive degradation prognostic reasoning by particle filter with a neural network degradation model for turbofan jet engine. College of Engineering, Taibah University, Al-Medina Al-Munawara, Medina 42353, Saudi Arabia, 21 pages.

Ελληνική Βιβλιογραφία

[1] Μαθιουδάκης Κ., 2007. ΔΙΑΓΝΩΣΤΙΚΗ ΑΕΡΙΟΣΤΡΟΒΙΛΩΝ, Τομέας Ρευστών Εργαστηρίου Θερμικών Στροβιλομηχανών, Ε. Μ. Πολυτεχνείο Σχολή Μηχανολόγων Μηχανικών, Αθήνα 2007, 253 σελίδες.

[2] Μαθιουδάκης Κ., 2007. ΛΕΙΤΟΥΡΓΙΑ ΑΕΡΟΠΟΡΙΚΩΝ ΚΙΝΗΤΗΡΩΝ, Τομέας Ρευστών Εργαστηρίου Θερμικών Στροβιλομηχανών, Ε. Μ. Πολυτεχνείο Σχολή Μηχανολόγων Μηχανικών, Αθήνα 2007, 253 σελίδες.

[3] Κατσούλη Μ., 2018. ΤΕΧΝΙΚΕΣ ΠΡΟΒΛΕΨΗΣ ΜΕ ΕΦΑΡΜΟΓΗ ΣΤΗΝ ΔΙΑΓΝΩΣΤΙΚΗ ΒΙΟΜΗΧΑΝΙΚΩΝ ΑΕΡΙΟΣΤΡΟΒΙΛΩΝ. ΕΜΠ (Ιούλιος 2018), 92 σελίδες.