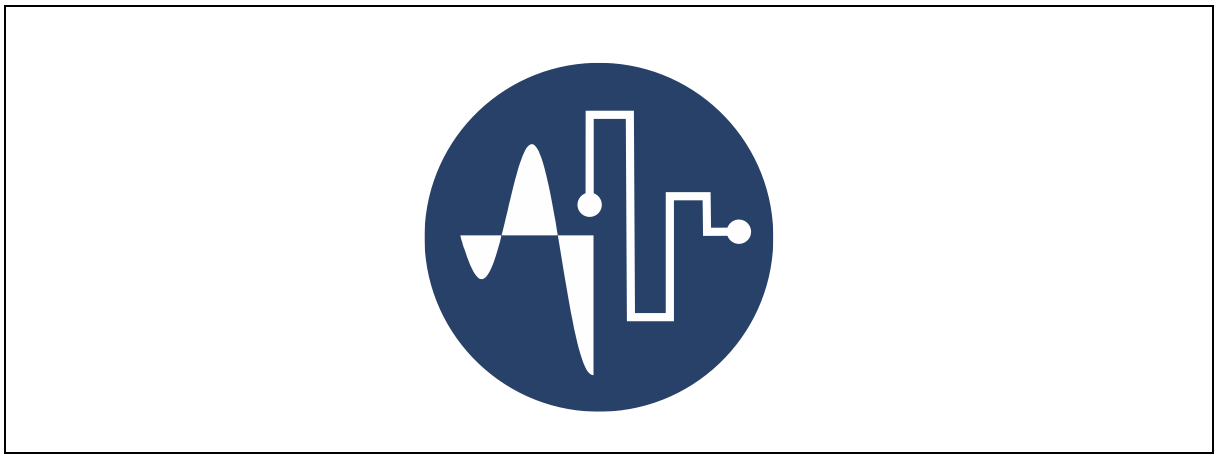
ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Μελέτη αλγορίθμων Matrix Profile και πειραματική αξιολόγησή τους στην αναζήτηση μοτίβων σε χρονοσειρές με την χρήση της γλώσσας προγραμματισμού R.»



Της φοιτήτριας
Βαβατζιάνη Αικατερίνη
Αρ. Μητρώου: 164640

Επιβλέπων
Καραμητόπουλος Λεωνίδας
Βαθμίδα Εργαστηριακό Διδακτικό Προσωπικό

Ημερομηνία 11/09/2025

Τίτλος Δ.Ε. Μελέτη αλγορίθμων Matrix Profile και πειραματική αξιολόγησή τους στην αναζήτηση μοτίβων σε χρονοσειρές με την χρήση της γλώσσας προγραμματισμού R.
Κωδικός Δ.Ε. 23317
Ονοματεπώνυμο φοιτήτριας Βαβατζιάνη Αικατερίνη
Ονοματεπώνυμο εισηγητή Καραμητόπουλος Λεωνίδας
Ημερομηνία ανάληψης Δ.Ε 05/11/2023
Ημερομηνία περάτωσης Δ.Ε. 11/09/2025

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

«Θα ήθελα να αφιερώσω την παρούσα πτυχιακή εργασία στην οικογενειά μου»

Πρόλογος

Η παρούσα πτυχιακή αναλύει την μεθοδολογία του Matrix Profile και την χρήση του για την αναγνώριση μοτίβων και ανωμαλιών σε χρονοσειρές χρησιμοποιώντας την γλώσσα προγραμματισμού R. Στόχος είναι να παρουσιάσει τα πλεονεκτήματα και μειονεκτήματα της χρήσης του αλγορίθμου Matrix Profile και να συμβάλει στην κατανόησή της χρησιμότητάς του στην ανάλυση δεδομένων.
Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες προς τον επιβλέποντα της πτυχιακής μου εργασίας, κ. Λεωνίδα Καραμητόπουλο για την καθοδήγηση και την πολύτιμη βοήθεια που προσέφερε σε κάθε στάδιο εκπόνησης της παρούσας πτυχιακής εργασίας.

Περίληψη

Στο παρόν έργο, παρουσιάζεται και αναλύεται η δομή δεδομένων Matrix Profile, όπου καινοτομεί στην ανάλυση χρονοσειρών και συγκεκριμένα στην αναζήτηση μοτίβων αλλά και την ανίχνευση ανωμαλιών, σε πρωτογενή δεδομένα διαφορετικών μεγεθών αλλά και ποικίλων επιστημονικών πεδίων. Αρχικά, γίνεται μια εισαγωγική αποτύπωση του θεωρητικού πλαισίου της Ανάλυσης των Χρονοσειρών αλλά και της Εξόρυξης Δεδομένων. Κατόπιν, παρουσιάζεται η βασική μεθοδολογία του MP, τα κύρια τεχνικά χαρακτηριστικά του και έπειτα αναλύονται οι συνήθεις χρησιμοποιημένοι επιμέρους αλγόριθμοι του. Παρουσιάζονται οι σχετικοί κώδικες των αλγορίθμων αυτών στα βασικά λειτουργικά προγράμματα(R,C++,MATLAB, PYTHON), διενεργείται μια σύγκριση τους και παρουσιάζονται τα θετικά και αρνητικά χαρακτηριστικά τους. Επίσης, αποτυπώνονται τα πλεονεκτήματα και μειονεκτήματα του MP όσον αφορά τα βασικά ζητούμενα του(μοτίβα, ανωμαλίες), γίνεται αναφορά στη συμβατότητα του αναλόγως των αναγκών της κάθε έρευνας και της φύσης των δεδομένων, ενώ εν συνεχεία σημειώνονται κάποιες ενδεικτικές αναφορές χρήσης του MP σε διαφόρους επιστημονικούς κλάδους. Ακόμη, παρουσιάζονται οι δομές δεδομένων πέραν του MP στον ανωτέρω τομέα ενδιαφέροντος. Τονίζεται τόσο το που υπερτερούν σε σχέση με τον MP, όσο και το που υστερούν, αντίστοιχα. Στο πειραματικό μέρος, εξετάζονται 3 διαφορετικά set δεδομένων στον τομέα της Βιοϊατρικής, όπου μέσω του αλγορίθμου STOMP στο περιβάλλον του R STUDIO, εκτελείται ο σχετικός κώδικας και παρουσιάζονται τα βασικά αποτελέσματα σχετικά με τα μοτίβα και της ανωμαλίες που ανιχνεύονται αντιστοίχως. Τα αποτελέσματα, έρχονται σε σχετική συμφωνία με την υπάρχουσα βιβλιογραφία σχετικά με τα συγκεκριμένα δεδομένα.

«Matrix Profile Algorithms in Finding Motifs: Experimental Evaluation Using the Language R»

«Katerina Vavatziani»

Abstract

In the present document, Matrix Profile is presented and analyzed, as it plays a leading role in Time Series analysis, particularly in motifs discovery and anomalies detection, on raw data of various sizes and diverse scientific fields.
Initially, an introduction of the theoretical framework of Time Series Analysis and Data Mining is provided. Subsequently, the core methodology of the Matrix Profile (MP), its main technical characteristics, and the most commonly used associated algorithms are presented and analyzed.
The relevant algorithmic codes are shown in the programming tool of R-studio, followed by an analysis, highlighting its strengths and weaknesses.
The advantages and disadvantages of MP with regard to its key objectives (motifs, anomalies) are discussed, along with its compatibility depending on the needs of each research case and the nature of the data. Moreover, certain bibliography, related to the use of MP in various scientific topics is presented. Furthermore, alternative data structures used in the same area of interest are examined, with emphasis on where they outperform or fall short in comparison to MP.
In the experimental section, three different datasets from the field of biomedicine are analyzed. Using the STOMP algorithm within the R STUDIO environment, the corresponding code is executed, and key results are presented, related to detected motifs and anomalies. The results are found to be in general agreement with the existing literature related to these specific datasets.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την οικογενειά μου για την συμπαράσταση και κατανόηση που έδειξαν σε όλη τη διάρκεια της συγγραφής.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Συντομογραφίες

MP	Matrix Profile
FFP	Fast Fourier Tranform
MASS	Mueen's Algorithm for Similarity Search
DM	Data Mining
ACAMP	Anytime Computation of Matrix Profile
SCRIMP	Scalable and Accurate Matrix Profile
STOMP	Scalable Time Series Ordered Matrix Profile
STAMP	Scalable Time Series Anytime Matrix Profile
MPX	Matrix Profile Exact

ΕΙΣΑΓΩΓΗ

Η ανάλυση χρονοσειρών αποτελεί βασικό τομέα ενδιαφέροντος σε διάφορες επιστημονικές και βιομηχανικές εφαρμογές, όπως η βιοπληροφορική, τα χρηματοοικονομικά, η ανάλυση αισθητήρων κ.α .[1] [2] [3] [4] Κύρια ζητήματα διερεύνησης σε αυτόν τον τομέα περιλαμβάνουν την αναζήτηση συγκεκριμένων επαναλαμβανόμενων μοτίβων (motifs) και τη ανίχνευση ανωμαλιών (anomalies) μέσα από μεγάλα σύνολα δεδομένων. [5] [6] [7] Ένα κρίσιμο πεδίο έρευνας αποτελεί η ανάπτυξη μεθόδων που μπορούν να αντιμετωπίσουν αυτές τις προκλήσεις, διατηρώντας την επεκτασιμότητα και την αποδοτικότητα τους. [8] Ο αλγόριθμος Matrix Profile εισήχθη ως μια γενική μέθοδος για την ανάλυση χρονοσειρών, επιτρέποντας την εύρεση επαναλαμβανόμενων μοτίβων και ανωμαλιών με μεγάλη ακρίβεια και χαμηλό υπολογιστικό κόστος. [9] Η βασική του αρχή στηρίζεται στον υπολογισμό της απόστασης μεταξύ όλων των δυνατών υποακολουθιών μιας χρονοσειράς, αποτυπώνοντας σχέσεις που σε άλλη περίπτωση θα ήταν δύσκολο να εμφανιστούν. [10] Αυτή η προσέγγιση διενεργεί μια συνολική επισκόπηση της εξεταζόμενης χρονοσειράς και προβαίνει σε διάγνωση απαιτητικών αποτελεσμάτων.

Στο παρόν, διερευνάται η χρήση του Matrix Profile για την ανάλυση μοτίβων και ανωμαλιών κάτω από το πρίσμα μιας συγκεκριμένης πειραματικής έρευνας σε συμβατά σύνολα δεδομένων. Ειδικότερα, εξετάζεται η εφαρμογή του σε set δεδομένων που αφορούν πληροφορίες στον τομέα της Βιοατρικής.

Μέσα από τη χρήση του Matrix Profile οι στόχοι είναι οι εξής:

- Να εντοπιστούν χαρακτηριστικά επαναλαμβανόμενα μοτίβα που υποδεικνύουν κανονική λειτουργία ή συμπεριφορά. (motifs)
- Να ανιχνευτούν ασυνήθιστες υποακολουθίες που μπορεί να σχετίζονται με σημαντικά συμβάντα και δυσλειτουργίες στην κατάσταση του συστήματος (anomalies).

Η εργασία δομείται και οργανώνεται ως εξής:

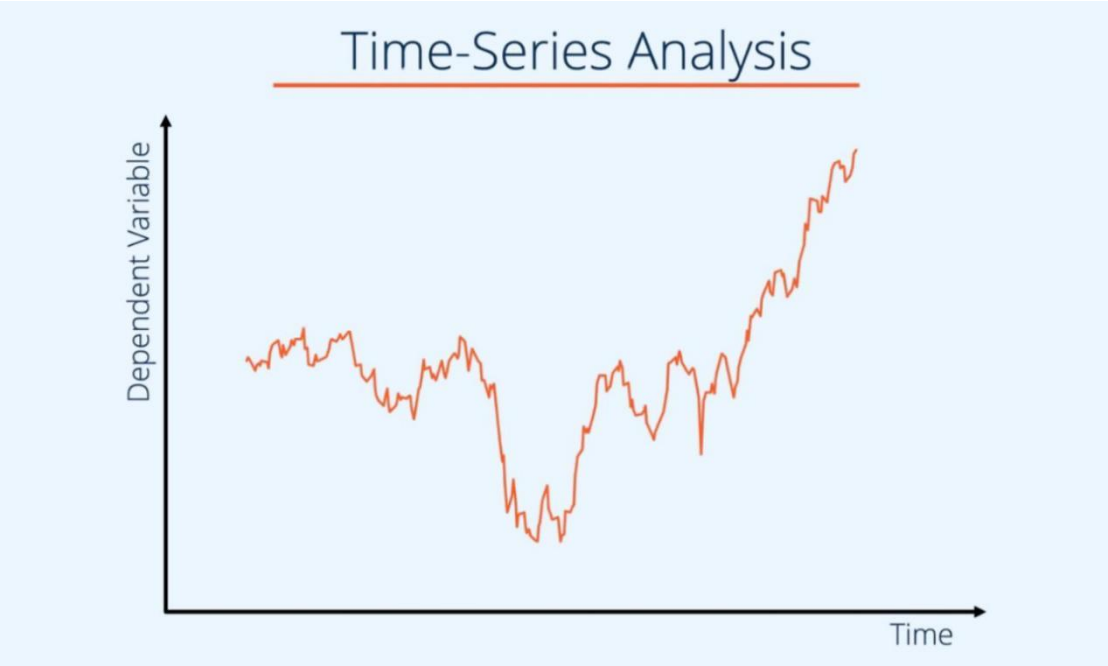
- Στο κεφάλαιο 1 παρουσιάζεται το βασικό θεωρητικό πλαίσιο πάνω στο οποίο στηρίζεται η ανάλυση χρονοσειρών και εν συνεχεία η εξόρυξη δεδομένων στπ πεδίο αυτό.
- Στο κεφάλαιο 2 περιγράφεται αναλυτικά η λειτουργία του MP, καθώς τα πλεονεκτήματα και μειονεκτήματα του και οι εφαρμογές του στα διάφορα επιστημονικά πεδία
- Στο κεφάλαιο 3, αναλύονται οι έννοιες των μοτίβων και των ανωμαλιών, καθώς επίσης και η σύνδεση τους με τη δομή δεδομένων Matrix Profile
- Στο κεφάλαιο 4, επιοσημαίονται επιμέρους αλγόριθμοι που χρησιμοποιεί ο Mp, η σύγκριση μεταξύ τους, καθώς και ο συνδεδεμένος κώδικας στο περιβάλλον της R
- Στο κεφάλαιο 5, αποτυπώνονται ,παρουσιάζονται και αναλύονται τα αποτελέσματα που προκύπτουν από τη σχετική στατιστική ανάλυση και χρήση του MP στα εξεταζόμενα δεδομένα, σε 3 διαφορετικά datasets.

Κεφάλαιο 1ο : Εισαγωγικές έννοιες και ορισμοί

Στο κεφάλαιο αυτό θα παρουσιαστούν οι βασικές έννοιες στην ανάλυση των χρονοσειρών, στην εξόρυξη δεδομένων και η σύνδεση τους με τον Αλγόριθμο Matrix Profile.

1.1 Ανάλυση χρονοσειρών

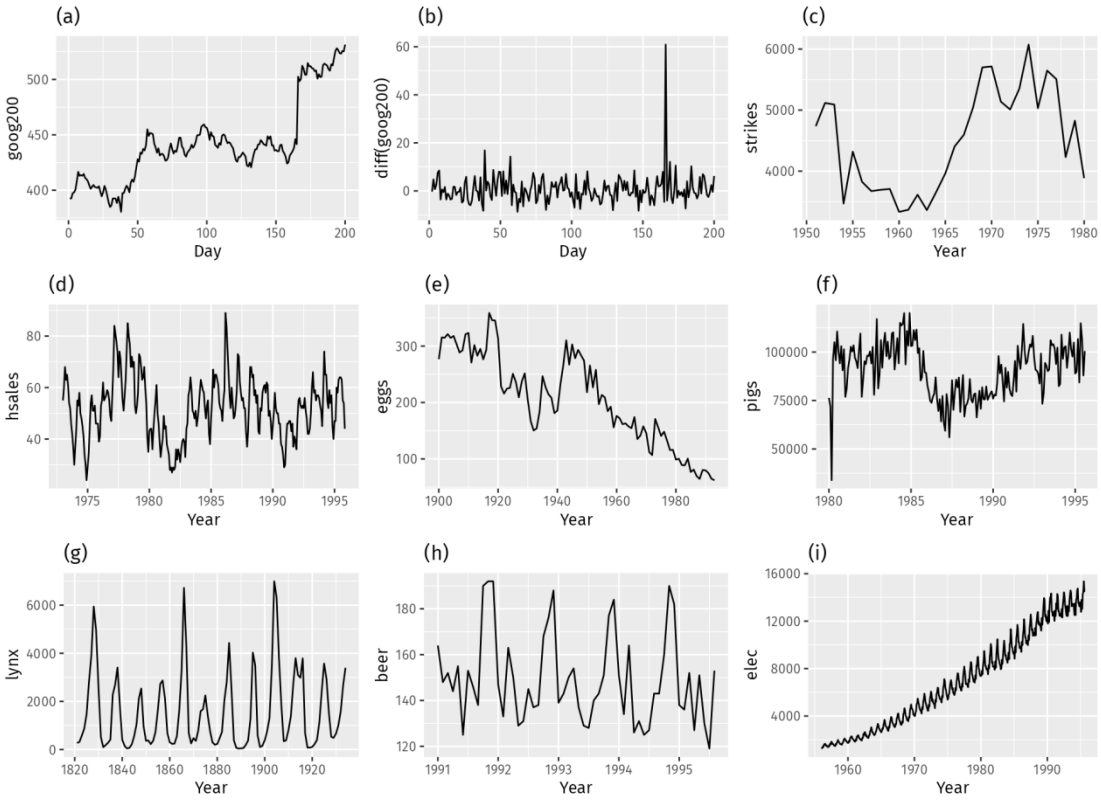
Η ανάλυση χρονοσειρών (time series analysis) είναι ο τομέας αυτός της Στατιστικής και της Επιστήμης των δεδομένων που εστιάζει στη μελέτη και επεξεργασία πρωτότυπων δεδομένων, τα οποία αφορούν καταγραφές σε διαδοχικές χρονικές στιγμές. Χρονοσειρά ορίζεται μια ακολουθία καταγραφών σε τακτά χρονικά διαστήματα, για παράδειγμα η καταγραφή των τιμών μιας μετοχής ανά ημέρα, οι θερμοκρασίες μιας περιοχής ανά ώρα, ο αριθμός επισκεπτών σε μια ιστοσελίδα κάθε λεπτό, των επισκεπτών ανά ώρα σε ένα αεροδρόμιο κ.α.
Για την καλύτερη κατανόηση της έννοιας της χρονοσειράς, ακολουθεί μια τυπική γραφική αποτύπωση της, στην Εικόνα 1.1.1 παρακάτω.



ΕΙΚΟΝΑ 1.1.1 : ΘΕΩΡΗΤΙΚΟ ΔΙΑΓΡΑΜΜΑ ΔΕΔΟΜΕΝΩΝ ΣΕ ΧΡΟΝΟΣΕΙΡΑ
Πηγή : https://bigblue.academy/gr/time-series-analysis

Στην παραπάνω Εικόνα , γίνεται σαφές, ότι μια Χρονοσειρά αποτυπώνει τις μετρήσεις μιας υπό εξέτασης μεταβλητής με το πέρας του χρόνου.
Η ανάλυση των Χρονοσειρών διέπεται από κάποιες βασικές έννοιες και ορισμούς.
Ορισμένες από τις πιο θεμελιώδεις έννοιες στην Ανάλυση Χρονοσειρών περιλαμβάνουν την Τάση, την εποχικότητα, τη Στασιμότητα, την Αυτοσυσχέτιση και τον θόρυβο.
Αναλυτικότερα, Τάση (Trend) μιας Χρονοσειράς ονομάζεται η μακροχρόνια εξέλιξη των τιμών μιας χρονοσειράς και αποτυπώνει την συστηματική και σταθερή αύξηση ή μείωση των παρατηρήσεων με

την χρονική πρόοδο που δεν συσχετίζεται από εποχιακές διακυμάνσεις.[11] Εποχικότητα (Seasonality) ορίζεται η καταγραφή περιοδικών μοτίβων που εμφανίζονται σε σταθερά χρονικά διαστήματα που αφορούν για παράδειγμα ένα τρίμηνο ή μήνα και έχει εφαρμογές στους οικονομικούς κύκλους, στο κλίμα κ.α. [12] Στασιμότητα (Stationarity) εμφανίζει μια χρονοσειρά, όταν όταν οι στατιστικές της μετρήσεις όπως :μέσος, διακύμανση κλπ. παραμένουν σταθερές με το πέρας του χρόνου. Σύμφωνα μ' αυτό το φαινόμενο, οι διαφορές μεταξύ των τιμών παραμένουν σταθερές στο χρόνο. [13] [14]
Οι παραπάνω έννοιες γίνονται πιο εύληπτα κατανοητές με τη βοήθεια της παρακάτω εικόνας:

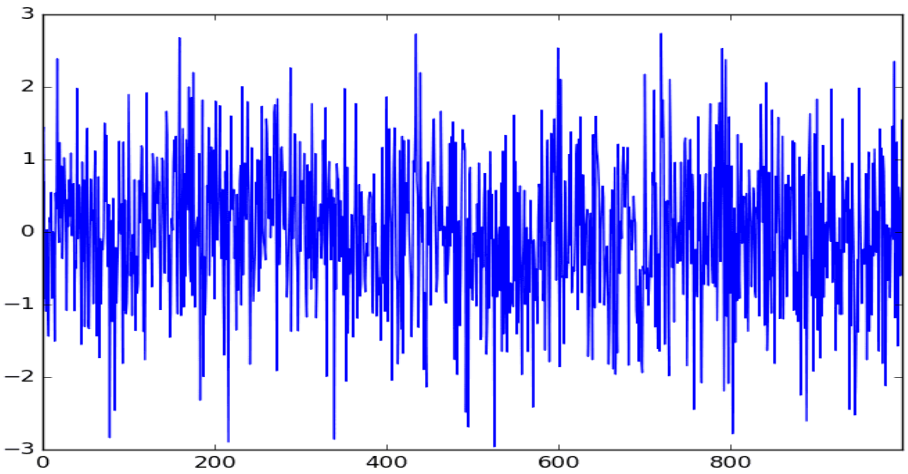


ΕΙΚΟΝΑ 1.1.2 : ΣΤΑΣΙΜΟΤΗΤΑ ΚΑΙ ΕΠΟΧΙΚΟΤΗΤΑ ΣΤΙΣ ΧΡΟΝΟΣΕΙΡΕΣ

Πηγή : https://otexts.com/fpp2/stationarity.html

Τα παραπάνω σετ χρονοσειρών αφορούν : Την τιμή της μετοχής της Google για 200 συνεχόμενες μέρες (a), Την ημερήσια διαφορά της ανωτέρω τιμής (b), Τον ετήσιο αριθμό των Απεργιών στις ΗΠΑ (c) , Τις μηνιαίες πωλήσεις σπιτιών νεοσύστατων οικογενειών στις ΗΠΑ (d), Την ετήσια τιμή της ντουζίνας αυγών στις ΗΠΑ (e), την Μηνιαία σφαγή γουρουνιών στη Ν.Ζηλανδία (f) , Τον ετήσιο αριθμό θηραμάτων Λύγκας στον ποταμό Μακ Κενζι στη Βικτόρια της Αυστραλίας (g) , τη Μηνιαία παραγωγή μπύρας στην Αυστραλία (h) και Τη μηνιαία παραγωγή ηλεκτρικής ενέργειας στην Αυστραλία.
Με μια βαθύτερη ανάλυση των ανωτέρω γραφημάτων, συμπεραίνεται πως τα γραφήματα d,h,i αποτυπώνουν Εποχικότητα, τα αντίστοιχα (c), (e), (f) και (i) αποτυπώνουν Τάση, ενώ το γράφημα g περιέχει μη περιοδικές αλλαγές που επηρεάζονται από πολλούς αστάθμητους παράγοντες ,επομένως μακροπρόθεσμα είναι μη προβλέψιμες και αυτό συνιστά μια Στάσιμη χρονοσειρά.
Αυτοσυσχέτιση (Autocorrelation) ορίζεται το φαινόμενο κατά το οποίο οι τιμές μιας χρονοσειράς επηρεάζονται και αλληλοσυσχετίζονται από προηγούμενες τιμές της, δηλαδή οι μελλοντικές τιμές επηρεάζονται από τις παρελθούσες με συγκεκριμένη χρονική υστέρηση. [15] [11] Θόρυβος (Noise):

ορίζεται το φαινόμενο των τυχαίων διακυμάνσεων που δεν ακολουθούν κάποιο προβλέψιμο και συγκεκριμένο μοτίβο και αποτυπώνει τα ‘’κατάλοιπα’’ αν αφαιρεθούν οι τάσεις και οι τυχόν εποχικότητες σε μια Χρονοσειρά, όπως αποτυπώνεται στην παρακάτω εικόνα. [16]



ΕΙΚΟΝΑ 1.1.3: ΘΟΡΥΒΟΣ ΣΤΙΣ ΧΡΟΝΟΣΕΙΡΕΣ
Πηγή : https://machinelearningmastery.com/

Αναλόγως του επιστημονικού στόχου της, μια διαδικασία Ανάλυσης Χρονοσειρών ενδέχεται να είναι περιγραφική, προγνωστική ή ανιχνευτική . Πιο συγκεκριμένα, η περιγραφική ανάλυση έχει ως σκοπό την κατανόηση των βασικών χαρακτηριστικών μιας χρονοσειράς, δηλαδή της τάσης, των θορύβων, της εποχικότητας, μέσω γραφημάτων και υπολογισμών βασικών Στατιστικών μετρήσεων χωρίς να προβαίνει σε προβλέψεις. Η προγνωστική ανάλυση έχει ως σκοπό την πρόβλεψη μελλοντικών τιμών μιας χρονοσειράς με βάση τις προ υπάρχουσες τιμές, χρησιμοποιώντας εργαλεία όπως η Μηχανική μάθηση για να εκτιμήσει τις μελλοντικές τάσεις. Τέλος, η ανιχνευτική ανάλυση έχει ως στόχο την ανακάλυψη ασυνήθιστων μοτίβων, αλλαγών ή ανωμαλιών μέσα στη χρονοσειρά. Αναφέρεται και ως ανίχνευση ανωμαλιών (anomaly detection). [11] [17] [18]

Ιστορικά, η μελέτη των χρονοσειρών αποτέλεσε μέρος της Κλασικής Στατιστικής. Στις αρχές του 20ού αιώνα, οι πρώτες αξιόπιστες εφαρμογές αφορούσαν τους τομείς των Οικονομικών και της Μετεωρολογίας. Οι George Udny Yule και Gilbert Walker ήταν από τους πρώτους επιστήμονες που ανέλυσαν χρονοσειρές μέσω μοντέλων αυτοπαλινδρόμησης (AR) , δηλαδή μοντέλων στα οποία η τρέχουσα τιμή μιας μεταβλητής εξαρτάται γραμμικά από τις προηγούμενες τιμές της ίδιας μεταβλητής και από έναν στοχαστικό όρο (θόρυβο ή κατάλοιπο). [19]
Τη δεκαετία του 1970, εξέλιξη των ανωτέρω μοντέλων αποτυπώθηκε με τη θεωρία των ARIMA από τους Box και Jenkins , τα οποία κυριάρχησαν για δεκαετίες στη μοντελοποίηση χρονοσειρών. και αφορούν συνδυαστικά μοντέλα αυτοσυσχέτισης, ολοκλήρωσης και ανάλυσης Κινητού Μέσου. [12]
Η τεχνολογική πρόοδος της υπολογιστικής ισχύος και η εξέλιξη της Επιστήμης δεδομένων, συνέβαλε στο γεγονός ότι στην Ανάλυση χρονοσειρών άρχισαν να χρησιμοποιούνται πιο σύνθετες μεθοδοι, όπως τα νευρωνικά δίκτυα, οι αλγόριθμοι clustering και οι τεχνικές εξόρυξης μοτίβων (pattern discovery). Βαθύτερα, τα Νευρωνικά δίκτυα, αντλούν ως μέθοδο την έμπνευση τους από τη λειτουργία του ανθρωπίνου εγκεφάλου, με τα δεδομένα που περιέχονται σ αυτά να εκπαιδεύονται από τα ίδια τα δεδομένα και να παράξουν προβλέψεις. [20] Ακόμη, ομαδοποίηση (clustering) ονομάζεται η τεχνική

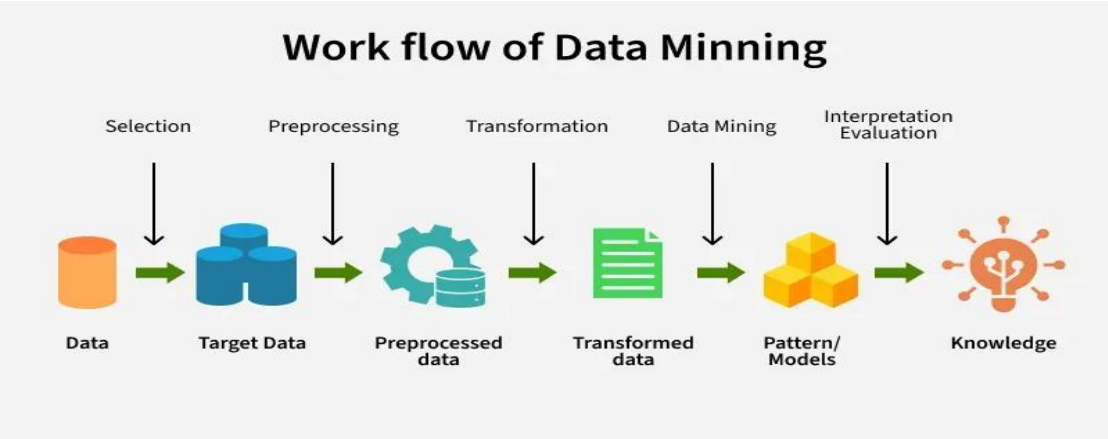
μη εποπτευόμενης μάθησης (unsupervised learning) που έχει ως σκοπό τη διάκριση και την ομαδοποίηση των παρατηρήσεων σε ομάδες (clusters) με βάση κάποια ομοιότητα ή εγγύτητα. [21]
Επίσης, Η εξόρυξη μοτίβων (pattern mining) είναι κλάδος της εξόρυξης δεδομένων που αποσκοπεί στον εντοπισμό επαναλαμβανόμενων, σημαντικών ή μη προφανών μοτίβων σε μεγάλα σύνολα δεδομένων. [22] Οι επιστημονικές εφαρμογές της ανάλυση χρονοσειρών ποικίλουν. Για παράδειγμα χρησιμοποιείται στον τομέα της Χρηματοοικονομικής (π.χ. πρόβλεψη τιμών μετοχών) , της Βιοϊατρικής, (π.χ. παρακολούθηση βιολογικών σημάτων), των Τηλεπικοινωνιών και cuber security (π.χ. εντοπισμός ανωμαλιών) αλλά και της Μετεωρολογίας (π.χ. πρόβλεψη καιρού)
Σε αντίθεση με προηγούμενες παραδοσιακές μεθόδους, το Matrix Profile δεν απαιτεί παραμετροποίηση, είναι αυτόνομο, εφαρμόσιμο σε μεγάλες χρονοσειρές και χρησιμοποιείται ως βάση για πολλές εφαρμογές. [9] Η συνεισφορά του Matrix Profile στην ανάλυση χρονοσειρών είναι καινοτόμα, καθώς επιτρέπει με ενιαίο τρόπο να εκτελούνται εργασίες που παλαιότερα απαιτούσαν διαφορετικά εργαλεία. Λόγω της απλότητας στην εφαρμογή του, έχει ενσωματωθεί για παράδειγμα σε βιβλιοθήκες της Python, όπως το stumpy, και χρησιμοποιείται σε μεγάλο έυρος σε εργασίες που απαιτούν ερμηνεία αποτελεσμάτων από δεδομένα σε χρονοσειρές.

Συμπερασματικά, το Matrix Profile προσφέρει μια ενιαία και υπολογιστικά αποδοτική λύση για την Ανάλυση χρονοσειρών, και αναδεικνύεται σε ένα από τα πιο σημαντικά εργαλεία των τελευταίων ετών.

1.2 Εξόρυξη δεδομένων σε χρονοσειρές

Data Mining (Εξόρυξη Δεδομένων) ονομάζεται η επιστημονική μέθοδος εξαγωγής πληροφοριών ή μοτίβων από μεγάλα σύνολα δεδομένων, με τη χρήση αλγορίθμων και αφορά τους τομείς της Στατιστικής, της Μηχανικής μάθησης και της Επιστήμης δεδομένων.
Η κεντρική ιδέα αποτυπώνει την υπόθεση ότι στους μεγάλους όγκους δεδομένων (big data) υπάρχουν κρυμμένα πρότυπα τα οποία μπορούν να αξιοποιηθούν για πρόβλεψη, κατανόηση , λήψη αποφάσεων, κ.ά. [22]

Η ροή της επιστημονικής διαδικασίας πίσω από το Data Mining αποτυπώνεται στην παρακάτω εικόνα:

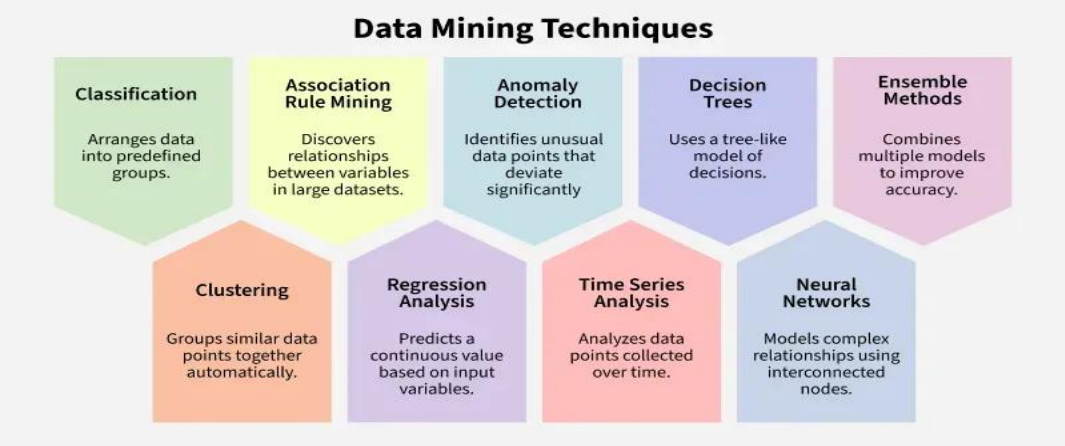


ΕΙΚΟΝΑ 1.2.1: ΡΟΗ ΕΡΓΑΣΙΩΝ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Πηγή: https://botpenguin.com/glossary/data-mining
Ειδικότερα, στην παραπάνω εικόνα, αναφέρεται η χρονική ροή των εργασιών πίσω από την Εξόρυξη δεδομένων, ήτοι: Τα πρωτότυπα δεδομένα και η συλλογή τους, η βασική προεργασία, ο

μετασχηματισμός τους, η εξαγωγή των μοτίβων και των προτύπων και τα συμπεράσματα που μπορούν να εξαχθούν.

Οι βασικές τεχνικές του DM είναι : Η Κατηγοριοποίηση (clasification) , η Ομαδοποίηση(clustering), Ανακάλυψη κανόνων συσχέτισης(Association Rule Learning), η Ανίχνευση ανωμαλιών(Anomaly Detection) και η Ανίχνευση διατεταγμένων προτύπων (Sequential Pattern Mining). [23] Αναλυτικότερα, Κατηγοριοποίηση ονομάζουμε μια εποπτευόμενη μαθησιακή διαδικασία (supervised learning), στην οποία ένα σύστημα δεδομένων 'εκπαιδεύεται' από υπάρχοντα κατηγοριοποιημένα δεδομένα (labelled data) με σκοπό να προβλέψει τις κατηγορίες (class) των νέων δεδομένων. Δηλαδή, σκοπός της ανωτέρω τεχνικής είναι να αποκαλύπτει την κατηγορία που ανήκει μια νέα παρατήρηση. Χρήσιμες εφαρμογές της τεχνικής αυτής χρησιμοποιούνται στην Κυβερνοασφάλεια, την Ιατρική και την Τραπεζική. [22] [23] . Ακόμη, Ομαδοποίηση καλείται μια μη εποπτευόμενη τεχνική (unsupervised learning) που στοχεύει στην ομαδοποίηση των παρατηρήσεων χωρίς προ υπάρχουσες κατηγορίες , με βάση την ομοιότητα μεταξύ τους δηλαδή κύριος σκοπός της τεχνικής είναι να αποκαλυφθούν οι 'φυσικές' ομάδες (clusters) που αποτυπώνουν παρόμοια χαρακτηριστικά.
Οι παραπάνω τεχνικές με τα επιμέρους χαρακτηριστικά τους, βρίσκονται στην παρακάτω εικόνα.



ΕΙΚΟΝΑ 1.2.2 : Οι βασικές τεχνικές του Data Mining
Πηγή: https://botpenguin.com/glossary/data-mining

Εφαρμογές της ανωτέρω τεχνικής μπορούν να βρεθούν για παράδειγμα στην Ομαδοποίηση πελατών αναλόγως του καταναλωτικού τους προφίλ, στην Ανάλυση κοινωνικών δικτύων (π.χ., κοινότητες) κ.α [24] . Ανακάλυψη κανόνων συσχέτισης, ονομάζεται η τεχνική εξόρυξης μιας συσχέτισης τύπου : 'Αν Α τότε Β', μεταξύ των στοιχείων στο σύνολο δεδομένων. Για παράδειγμα, η αγορά 'συμπληρωματικών' αγαθών, όπως Γάλα και Δημητριακά, ή οι προτάσεις της επόμενης θέασης σε μια πλατφόρμα όπως το Youtube μετά από μια παρακολούθηση ενός βίντεο. [25]. Επιπρόσθετα, Ανίχνευση ανωμαλιών ονομάζεται η τεχνική ανακάλυψης ασυνήθιστων μοτίβων και ''εκτός κανόνα'' παρατηρήσεων (δείτε αναλυτικότερα ενότητα 2.2) [18] [9] Τέλος, Ανίχνευση διατεταγμένων προτύπων, ονομάζεται η τεχνική αυτή, που αποτυπώνει τις χρονικά ακολουθούμενες παρατηρήσεις σε μια χρονοσειρά, όπως για παράδειγμα το μια ποια σειρά και αν αυτή σχετίζεται ένας καταναλωτής αγοράζει προϊόντα ή ένας χρήστης επισκέπτεται κάποιες ιστοσελίδες.

Κεφάλαιο 2ο: Matrix profile: Έννοιες, ορισμοί, λειτουργία και εφαρμογές του

Το Matrix Profile είναι ένα ολοκληρωμένο εργαλείο στο πεδίο της ανάλυσης των χρονοσειρών, μια σχετικά πρόσφατη τεχνική στην ανάλυση τους, που παρουσιάστηκε από ερευνητές του Πανεπιστημίου της Καλιφόρνια. [9] Οι λειτουργίες του MP επεκτείνονται σε πολλά πεδία.
Αρχικά, αφορά έναν αποτελεσματικό αλγόριθμο που επιτρέπει την αναζήτηση μοτίβων (motifs), την ανίχνευση ανωμαλιών (discords) και την εύρεση όμοιων υποακολουθιών μέσα σε μια χρονοσειρά με εξαιρετικά αποδοτικό τρόπο. Επιπρόσθετα, οι δυνατότητες του MP αφορούν την ομαδοποίηση κάποιων τμημάτων μιας χρονοσειράς, την κατάτμηση της, τον εντοπισμό ‘΄σημείων καμπής/αλλαγής’, καθώς και την συμπίεση μιας χρονοσειράς. [26] Παρακάτω, αναλύονται οι ανωτέρω τεχνικές της δομής δεδομένων Matrix Profile.

2.1 Η λειτουργία του Matrix Profile και οι βασικές έννοιες

Η βασική λειτουργία του MP συνοψίζεται στα παρακάτω βήματα:

- Αντλείται μια Χρονοσειρά T, συνολικού αριθμού παρατηρήσεων n.
- Δημιουργείται το ‘Παράθυρο’ που διαχωρίζει τα υποσύνολα , μεγέθους m
- Δημιουργούνται τα υποτμήματα μήκους m : $T_1,T_2,T_3,.....T_{n-m+1}$
- Υπολογίζεται η Ευκλείδεια απόσταση κάθε υπό τμήματος από τα άλλα
- Δημιουργείται ένας πίνακας διαστάσεων (n-m+1) * (n-m+1)
- Για κάθε υπό τμήμα T_i , υπολογίζεται η ελάχιστη απόσταση με άλλο υπό τμήμα και όχι με τον εαυτό του
- Η παραπάνω ελάχιστη απόσταση αποθηκεύεται στο διάνυσμα του MP
- Και η αντίστοιχη θέση (index) αποθηκεύεται αντίστοιχα στον Πίνακα MP

Η παραπάνω μεθοδολογία γίνεται πιο εύληπτη από το παρακάτω υποθετικό αριθμητικό παράδειγμα:

‘Εστω μια χρονοσειρά ‘T’ με τα παρακάτω στοιχεία:
T= [4,7,6,5,8,6], που δύναται για παράδειγμα, να αποτυπωνει τους βαθμούς Κελσίου μέσα σε μια συγκεκριμένη εβδομάδα του χειμώνα στο κέντρο της Θεσσαλονίκης.
Το πρώτο βήμα στη μεθοδολογία είναι ο ορισμός του μεγέθους του παραθύρου. Έστω λοιπόν m = 3, δηλαδή 3 μέρες.
Εν συνεχεία, ο αριθμός των υποσειρών που απορρεύν από την ανωτέρω χρονοσειρά είναι:
n−m+1=7−3+1=5, και πιο συγκεκριμένα :
S1=[5,6,7]
S2=[6,7,6]
S3=[7,6,5]
S4=[6,5,4]
S5=[5,4,6]

Το επόμενο βήμα για να δημιουργηθεί ο πίνακας MP είναι ο υπολογισμός των αντίστοιχων ευκλιδείων αποστάσεων. Μετά τους απαραίτητους υπολογισμούς, ενδεικτικά, κάποιες από τις αποστάσεις είναι οι εξής:

dist(S1,S2) = $\sqrt{3}$

dist(S1,S3) = $\sqrt{8}$

dist(S1,S4) = $\sqrt{11}$

dist(S1,S5) = $\sqrt{5}$

Το τελευταίο βήμα για τον πίνακα MP είναι ο υπολογισμός του ελάχιστου της κάθε παραπάνω απόστασης και της τοποθέτησης του στο διάνυσμα, ήτοι:

MP=[$3^{1/2}$,$3^{1/2}$,$3^{1/2}$,$3^{1/2}$,$5^{1/2}$]

Το παραπάνω αριθμητικό παράδειγμα, παρουσιάζεται συνοπτικά στον παρακάτω πίνακα: 1.2.1

Πίνακας 1.2.1 : Συνοπτικά στοιχεία αριθμητικού παραδείγματος MP

Υποακολουθία	Στοιχεία	Απόσταση	Τιμή MP	Κοντινότερος Γείτονας
S_1	[4, 7, 6]	d(S_1,S_2) = 3.3	Min = 1.4	S_4
S_2	[7, 6, 5]	d(S_1,S_3) = 3.4	Min = 3	S_4
S_3	[6, 5, 8]	d(S_1,S_4) = 1.4	Min = 3.3	S_2
S_4	[5, 8, 6]	d(S_2,S_3) = 3.3	Min = 1.4	S_1
		d(S_2,S_4) = 3		
		d(S_3,S_4) = 3.7		

Υποθετικά, το μέγεθος παραθύρου στο συγκεκριμένο παράδειγμα, μπορεί να οδηγήσει σε συμπεράσματα σε σχέση με μοτίβα που σχετίζονται με μετεωρολογικά ‘κύματα’ ανά τριήμερο.

Μέσω συγκεκριμένης γλώσσας προγραμματισμού, των σχετικών βιβλιοθηκών κτλπ, ο MP μπορεί να χρησιμεύσει - όπως αναφέρθηκε και ανωτέρω - σε ποικίλες τεχνικές. Αρχικά, μέσω το σχετικού πακέτου ‘Julia’ που μπορεί να εγκατασταθεί σε windows, linux κ.α, και συγκεκριμένα της συνάρτησης ‘snippets(T, K, length)’, αφού ορίζονται οι βασικές παράμετροι (μήκος χρονοσειράς, αριθμός υποτμημάτων) , επιστρέφονται ως αποτελέσματα οι ‘K’ πιο αντιπροσωπευτικές χρονοσειρές. [27] Επίσης, για τη λειτουργία της κατάτμησης (αλλά και των changing points), χρησιμοποιείται συνηθέστερα η μέθοδος ‘FLUSS/FLOSS/ , σύμφωνα με την οποία , υπολογίζεται ένας πίνακας ‘’τμηματοποιημένος’’ (Segmentation Profile), που υποδεικύνονται online και με μεγάλη ταχύτητα, τα σημεία αλλαγής. Η παραπάνω μέθοδος είναι εγκατεστημένη και στην Python, και στην R , αλλά και στη βιβλιοθήκη Julia του MP [28] . Ακόμη, μέσω της ανωτέρω εντολής ‘’‘snippets’’ , ο MP μπορεί να

χρησιμεύσει στην συμπίεση/περίληψη μιας Χρονοσειράς, καθώς μπορεί να παρουσιάσει τα σημαντικότερα και πιο αντιπροσωπευτικά κομμάτια της.

2.2 Οι εφαρμογές του MATRIX PROFILE

Η μεθοδολογία του MP, χρησιμοποιείται κατά κόρον σε ποικίλα επιστημονικά πεδία, αναλόγως των επιμέρους στόχων που έχουν τεθεί στο κάθε project. Στην έρευνα του 2021 των Σι Γινκ και άλλων, ο επιστημονικός σκοπός της χρήσης του MP ήταν η παρακολούθηση των προτύπων Kpis σε συστήματα πληροφορικής. [36] Ο MP χρησιμοποιείται on line και συγκρίνει την τρέχουσα υπό - σειρά με τις προϋπάρχουσες για ανίχνευση μοτίβων και ανωμαλιών. Η μέθοδος αυτή λειτουργεί ως πρότυπο (generic) με σημαντική αποδοτικότητα. Σε άλλη έρευνα του 2023, στον τομέα της Ιατρικής, ο σκοπός χρήσης του MP ήταν η εμφάνιση μοτίβων και ανωμαλιών στους ηλικιωμένους μέσω του περπατήματος. [29] Μάλιστα, η εξόρυξη των δεδομένων διενεργείται από ενσωματωμένες συσκευές στα υποκείμενα της έρευνας, με αποτελέσματα που αποτυπώνουν ακρίβεια, αλλά σε σχετικά χαμηλή ταχύτητα υπολογισμού. Το ίδιο έτος, σε έρευνα των Νικιφόροφ και Αλαμανιώτη, [30] στον τομέα της παραγωγής, το υπό εξέταση ζήτημα ήταν η κατανάλωση ενέργειας στις Πανεπιστημιακές μονάδες. Το ενδιαφέρον στοιχείο της εν λόγω εργασίας, είναι πως χρησιμοποιεί ως πρότυπο κάποιες υποσειρές με ασυνήθιστη συμπεριφορά από τους Καταναλωτές. Τα αποτελέσματα χαρακτηρίζονται από ακρίβεια, αξιοπιστία και ισχυρή αντοχή στον θόρυβο. Το 2024, οι Τσαο και Λιν, στο πεδίο της Πληροφορικής και συγκεκριμένα στον κλάδο Internet of Thinks(IoT), [31] διενήργησαν τον αλγόριθμο MP σε πολλές διαστάσεις (multidimensional), έχοντας ως σκοπό την αναγνώριση ομάδων μοτίβων. Η συγκεκριμένη μεθοδολογία έχει ορθή εφαρμογή για ανάλυση δικτύων αισθητήρων (sensor networks) , καθώς εντοπίζει μοτίβα που εμφανίζονται σε πολλές διαστάσεις(αισθητήρες). Το ίδιο έτος (2024) [8] σε έρευνα των Γιε κ.α , διενεργήθηκε η προηγούμενη μεθοδολογία του 'Πολυδιάστατου' MP, με τη χρήση ενός ειδικού διανύσματος που δημιουργείται , γίνεται η σύγκριση με 119 προϋπάρχοντα σετ δεδομένων. Τα σενάρια που εκτελούνται είναι σε εκπαιδευόμενα και μη δεδομένα και ο MP κρίνεται ως σταθερή και αποτελεσματική λύση σε κάθε περίπτωση. Το 2022, ο Τανμάι κ.α [32] εξειδικεύουν την έρευνα τους στο να συγχωνεύσουν δεδομένα μικρού μήκους Χρονοσειρών σε μια ενιαία μεγαλύτερη. Η συγκεκριμένη μεθοδολογία κρίνεται αποτελεσματική σε μοτίβα που ανιχνεύονται από πολλές, διαφορετικές, μικρότερες πηγές και ομαδοποιούνται με τη μέθοδο της 'γειτονικότητας' (Kn-MP). Στον κλάδο της Βιοατρικής, σε έρευνα του 2021 [33] διενεργείται η κατηγοριοποίηση των Χρονοσειρών στα εξειδικευμένα Ιατρικά σήματα(ECG). Σύμφωνα με τη μεθοδολογία έρευνας της ανωτέρω εργασίας, αντλούνται κάποιες 'διακριτές' υποσειρές, εν συνεχεία δημιουργείτε ένα 'δένδρο απόφασης' (decision tree), που παρέχει έναν εύληπτο και αποτελεσματικό τρόπο ταξινόμησης των δεδομένων μέσω του MP. Στον κλάδο των Χρηματοοικονομικών, σε εργασία του 2021, διενεργήθηκε ανάλυση χρηματιστηριακών δεδομένων που αφορούν τις σχετικές 'κινήσεις' της αγοράς. [34]
Πιο συγκεκριμένα, το σετ δεδομένων ελέγχθηκε για επαναλαμβανόμενες ''συμπεριφορές'' σε μετοχές και δείκτες των Χρηματιστηρίων. Σύμφωνα με τα δημοσιευμένα αποτελέσματα, η ανωτέρω μεθοδολογία έχει εφαρμογή στην ερμηνεία Χρηματιστηριακών κρίσεων και εν γένει τάσεων των αγορών και έχει τη δυνατότητα επεκτασιμότητας σε πολυδιάστατες Χρονοσειρές. Το 2018, στον τομέα της Γεωλογίας και της Σεισμολογίας, οι Keogh κ.α [35] χρησιμοποίησαν τον MP πάνω σε σεισμικά

σήματα. Συγκεκριμένα, διενεργήθηκε ανίχνευση μοτίβων σε σεισμούς και άλλα σχετικά φαινομένα, μέσω της τμηματοποίησης (segmentation) και κατηγοριοποίησης (clustering) των δεδομένων. Ακόμη, στον κλαδο της Κυβερνοασφάλειας, σε έρευνα του 2022 [36], διενεργήθηκε έλεγχος αναζήτησης ανωμαλιών που πιθανώς να σχετίζονται με Κυβερνοεπιθέσεις. Ειδικότερα, εφαρμόστηκε η μέθοδος εντοπισμού αλλαγών στις σχετικές αποστάσεις των επιμέρους σειρών, κατά τη διάρκεια μιας ‘επίθεσης’, δηλαδή μια ‘real time’’ έρευνα για τον εντοπισμό ανωμαλιών στις εξαταζόμενες Χρονοσειρές.

Κεφάλαιο 3ο: Motifs και Discords (Ανωμαλίες)

Στο κεφάλαιο αυτό, θα αναλυθούν οι έννοιες των μοτίβων και των ανωμαλιών, καθώς επίσης και θα αποτυπωθεί ο τρόπος κατά τον οποίο ο Matrix Profile προβαίνει στην ανίχνευση και την αναζήτησή τους.

3.1 Motifs and Discords: Ορισμοί και βασικές έννοιες

Μοτίβο καλείται το φαινόμενο κατά το οποίο κάποιο ή περισσότερα υπό-τμήματα επαναλαμβάνονται με παρόμοιο τρόπο. Αρχικά, σαν ιδέα αποτυπώθηκε από τους Chiu, Keogh [9] και αργότερα εξελίχθηκε από τον Mueen κ.α [37] Τα μοτίβα δηλαδή, αποτυπώνουν δομικά στοιχεία μιας επαναλαμβανόμενης στατιστικής συμπεριφοράς.

Η Μαθηματική έκφραση του Μοτίβου, για 2 υπό-τμήματα Ti,Tj για $i \neq j$, είναι:

$$\text{Motif} = \arg\min_{i,j} \text{Euc. Dis}\,(T_i, T_j) \quad (3.1)$$

δηλαδή: 'Βρες τα i,j που έχουν την ελάχιστη ευκλείδεια απόσταση'.

Ακόμη, ανωμαλία (anomaly) καλείται το φαινόμενο κατά το οποίο ένα υπο - τμήμα δεν μοιάζει με κανένα άλλο — δηλαδή είναι το πιο μοναδικό/διαφορετικό κομμάτι της χρονοσειράς. Συγκεκριμένα η έννοια 'discord' ορίστηκε από Keogh, Lin κ.ά. [38] ως η υπο-ακολουθία με τη μεγαλύτερη απόσταση από τον πλησιέστερο γείτονά της (nearest neighbor), δηλαδή το πιο ''ασυνήθιστο'' κομμάτι μιας ακολουθίας.

Αντίστοιχά με τη σχέση 3.1 , η μαθηματική σχέση των ανωμαλίας είναι:

$$\text{Discord} = \arg\max_{i,j} \text{Euc. Dis}\,(T_i, T_j) \quad (3.2)$$

Οι ανωμαλίες, αναλόγως της φύσης και της μορφής τους, ταξινομούνται σε: (i) σημειακές (point), (ii) συλλογικές/τμηματικές (collective) (iii) συμφραστικές (contextual) . Η έρευνα των Chandola κ.ά. παραμένει σημείο αναφοράς για το ζήτημα [18] .

Αναλυτικότερα, σημειακές ανωμαλίες , καλούνται μεμονωμένες τιμές/παρατηρήσεις που διαφέρουν σημαντικά από το υπόλοιπο dataset. Παράδειγμα, αν ένα σύνολο δεδομένων αποτυπώνει τις τραπεζικές κινήσεις ενός πελάτη, μια μοναδική ανάληψη 10κ ευρώ σε ένα λοιπό σύνολο αναλήψεων απο κινήσεις 200 ευρώ, τότε αυτό αποτελεί σημειακή ανωμαλια. Συλλογικές ανωμαλίες, αποτελεί μια ομάδα τιμών που εμφανίζει διαφορετική συμπεριφορά από το λοιπό σύνολο και όχι μια μεμονωμένη τιμή. Συμφραστική ανωμαλία , καλείται μια ακραία διαφορετική τιμή μιας παρατήρησης, ενταγμένη όμως σ' ένα πλαίσιο. [39] Για παράδειγμα, αν παρατηρηθεί μια θερμοκρασία 35 βαθμών σε χειμερινούς μήνες.

Τα κύρια κριτήρια που οδηγούν στην άντληση των βασικών ζητουμένων (μοτίβα, ανωμαλίες κ.α) είναι το μέγεθος των παρατηρήσεων n αλλά κυρίως το μέγεθος του παραθύρου m. Ειδικά για το δεύτερο, η σχετική βιβλιογραφία σημειώνει την επιστημονική σημασία του και κατά πόσο επηρεάζει τα αποτελέσματα. [40] [41]

Στα τεχνικά χαρακτηριστικά, χρήσιμες έννοιες για τον υπολογισμό των μοτίβων είναι οι: Z-κανονικοποίηση, η Ευκλίδεια απόσταση και ο Γρήγορος μετασχηματισμός Fourier. (FFT)

'Z-κανονικοποίηση' ονομάζουμε τον μετασχηματισμό αυτό των δεδομένων που έχει ως σκοπό την εξάλειψη της διαφοράς της μέσης τιμής και της διακύμανσης μεταξύ των υπό- τμημάτων. Η σχέση που ορίζει τον ανωτέρω μετασχηματισμό, για ένα υπό-τμήμα

με μέση τιμή 'μ' και τυπική απόκλιση 'σ' είναι η εξής:

$$\text{Si}: \quad Z = \frac{S_i - \mu}{\sigma}, \quad \forall i \quad , \ (2.1)$$

Μέσα από την κανονικοποίηση αυτή, όλα τα υπο τμήματα έχουν μ=1 και σ=0.

Η ευκλείδεια απόσταση που υπολογίστηκε στο παρόν, στην ενότητα 2.1 και στο σχετικό αριθμητικό παράδειγμα, αποτελεί ένα σημαντικό υπολογιστικό βήμα για την αναζήτηση μοτίβων και ανωμαλιών.

Επιπρόσθετα, ο Γρήγορος μετασχηματισμός Fourier (FFT)- ο οποίος μειώνει δραστικά την ταχύτητα των λειτουργιών και αυξάνει η αποδοτικότητα του Αλγορίθμου- διασπά την Χρονοσειρά σε συχνότητες (ημίτονα και συνημίτονα) με αποτέλεσμα να επιταχύνεται ο υπολογισμός συσχέτισης των δεδομένων. [42] Η ταχύτητα της σύγκρισης με τον FFT μειώνεται από : O(n)^2 (τετραγωνική) σε :Onlogn (λογαριθμική).

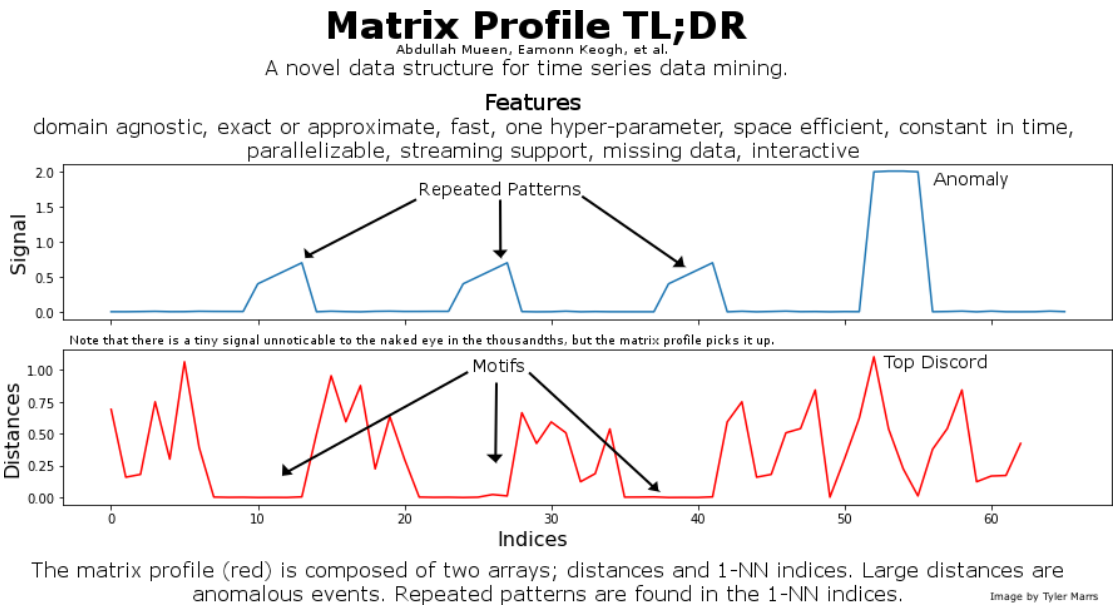
3.2 Χρήση του Matrix Profile για τον Εντοπισμό Motifs and Discords

Όπως προαναφέρθηκε και παραπάνω, οι βασικοί στόχοι που τίθενται μέσω του MP, είναι η αναζήτηση Μοτίβων (motifs) και ανωμαλιών (discords) σε μια εξεταζόμενη χρονοσειρά.

Το 2016, η δομή δεδομένων Matrix Profile παρουσιάστηκε ώς ένα ενοποιημένο και ολιστικό πλαίσιο αναζήτησης και ανίχνευσης μοτίβων και ανωμαλιών. Η διαδικασία αυτή, έχει εξελιχθεί χρονικά και όπως θα παρουσιαστεί παρακάτω στο Κεφ 4, μέσω των εργαλείων, των επιμέρους αλγορίθμων και των λειτουργικών, έχει επιτευχθεί υψηλή ακρίβεια αλλά και ταχύτητα. [] Ακόμη, μέσω του MP πραγματοποιείται και μια online ανάλυση των δεδομένων ,καθώς επίσης προχωράει και στην τμηματοποίηση μιας χρονοσειράς για τους ανωτέρω σκοπούς.
Στα τεχνικά κριτήρια, το μήκος της χρονοσειράς. το μέγεθος της του παραθύρου, η ζώνη αποκλεισμού συντελούν στην ανίχνευση των μοτίβων και των ανωμαλιών.

Η δομή δεδομένων Matrix Profile, έρχεται να επεκτείνει και να εξελίξει σε ακρίβεια και σε ταχύτητα την εύρεσή μοτίβων και ανωμαλιών , σε σχέση με τις προηγούμενες μεθόδους, όπως : οι πιθανοκρατικές/προβολικές μέθοδοι [43] ή η ακριβής απαρίθμηση. [28]

Γραφικά , τα μοτίβα και οι ανωμαλίες, αποκαλύπτονται με τις κορυφές και τα βάθη που εμφανίζονται στα σχετικά γραφήματα του MP, σύμφωνα με την παρακάτω εικόνα:

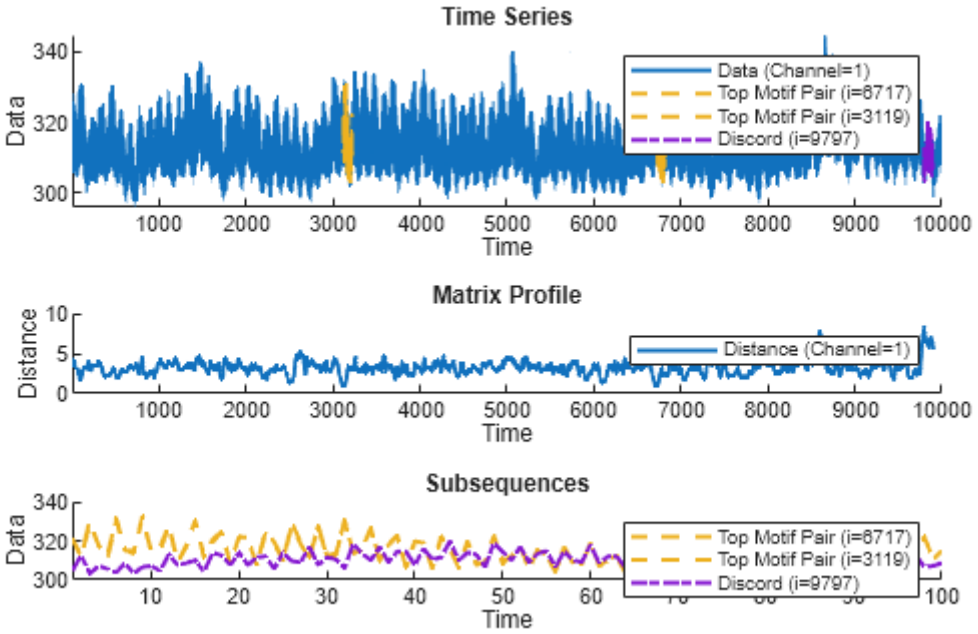


ΕΙΚΟΝΑ 3.2.1 : Μοτίβα, επαναλαμβανόμενα πρότυπα και ανωμαλίες
Πηγή : https://medium.com/data-science/introduction-to-matrix-profiles-5568f3375d90

Σύμφωνα με το ανωτέρω γράφημα:

Στον κάθετο άξονα μετριέται η απόσταση μεταξύ των υποσειρών.
Δύναται να ανακαλυφθούν τα μοτίβα ως τα τοπικά ελάχιστα των ανωτέρω γραφημάτων στην εικόνα 3.2.1, δηλαδή οι μικρότερες αποστάσεις των υποτμημάτων.
Επίσης, σε αντιστροφή, τα τοπικά μέγιστα αποτυπώνουν τις ανωμαλίες, καθώς καταγράφουν τις μεγαλύτερες δυνατόν αποστάσεις μεταξύ των υποσειρών. Τέλος, στην Εικόνα 3.2.1 γίνεται εύληπτη η παρατήρηση των επαναλαμβανόμενων προτύπων , δηλαδή μια συμπεριφορά των παρατηρήσεων συγκεκριμένη ως προς τις κορυφές και τα μεγέθη τους.

Στην κάτωθι εικόνα 3.2.2 , αποτυπώνεται το πως η δομή δεδομένων Matrix Profile συνεισφέρει στην αναζήτηση μοτίβων και ανωμαλιών :



Εικόνα 3.2.2 : Γραφική αποτύπωση του MP για μοτίβα και ανωμαλίες
Πηγή : https://www.mathworks.com/help/predmaint/ref/matrixprofile.html

Όπως μπορεί να παρατηρηθεί από την ανωτέρω εικόνα, ο αρχικός θόρυβος της Χρονοσειράς, αντιμετωπίζεται με την καταγραφή των αποστάσεων στον κάθετο άξονα του δεύτερου γραφήματος. Μ' αυτό τον τρόπο και αφού έχει οριστεί το μέγεθος παραθύρου, στο τρίτο γράφημα μπορούν να αποτυπωθούν τα μοτίβα και οι ανωμαλίες των καθορισμένων υποτμημάτων.

Συμπερασματικά, ο MP αποτελεί μια Δομή δεδομένων που μετά απο καθορισμένα βήματα αποθηκεύει την ελάχιστη απόσταση από οποιαδήποτε άλλη υποακολουθία (Nearest Neighbor Distance). για κάθε υποακολουθία μιας χρονοσειράς, σύμφωνα με το μέγεθος παραθύρου. Αποτελεί ένα ενιαίο πλαίσιο αναζήτησης μοτίβων και ανωμαλιών, χρησιμοποιώντας προγραμματιστικά εργαλεία όπως ο STAPM και ο STOMP , παρουσιάζει με απλό, γρήγορο και ακριβή τρόπο τα ζητούμενα. Οι εφαρμογές του MP για την αναζήτηση μοτίβων και ανωμαλιών ποικίλουν στα διάφορα επιστημονικά πεδία, όπως η βιοατρική, τα χρηματοοικονομικά και οι αυτοματισμοί.

Κεφάλαιο 4ο: Matrix profile: Αλγόριθμοι και Κώδικας στην γλώσσα R

4.1 Οι αλγόριθμοι που τρέχουν τη δομή δεδομένων Matrix Profile

Για τις ανάγκες της διενέργειας της δομής δεδομένων Matrix Profile , χρησιμοποιούνται διάφοροι αλγόριθμοι, αναλόγως του σκοπού, των δυνατοτήτων , της φύσης των πρωτότυπων δεδομένων κ.α
Οι πιο συνηθισμένοι αλγόριθμοι είναι οι :

- STAMP(Scalable Time series Anytime Matrix Profile)
- STOMP (Scalable Time series Ordered Matrix Profile)
- SCRIMP++ (Scalable and Accurate Matrix Profile)
- ACAMP (Anytime Computation of Matrix Profile (εξέλιξη του ο SCRIMP)
- MPX

Πιο συγκεκριμένα, ο STAMP (Scalable Time series Anytime Matrix Profile) είναι ο πρώτος αλγόριθμος που χρησιμοποιήθηκε για τον υπολογισμό του MP, [9] και αποτελεί τη βάση στην οποία αναπτύχθηκαν οι υπόλοιποι αλγόριθμοι (STOMP, SCRIMP++, κ.α). Κύριο χαρακτηριστικό του, ότι μπορεί να διακοπεί ανά πάσα στιγμή και να δώσει ένα αποτέλεσμα που βελτιώνεται όσο συνεχίζεται ο υπολογισμός του. Υπολογίζει το Matrix Profile μιας χρονοσειράς, δηλαδή την ελάχιστη απόσταση (Ευκλίδεια) κάθε υπό-σειράς με όλες τις υπόλοιπες. Βασίζεται στον υπολογισμό της συσχέτισης μέσω FFT (Fast Fourier Transform) με τη χρήση της μεθόδου MASS (Mueen's Algorithm for Similarity Search), δηλαδή ενός αλγορίθμου που χρησιμοποείται για να υπολογίσει τις επιμέρους αποστάσεις της κάθε υπό -σειράς από μια συγκεκριμένη υπό σειρά αναφοράς [37] [44]
Επίσης, τα λοιπά βασικά χαρακτηριστικά του, είναι η αποδοτική του κλιμάκωση (scalability) στις Χρονοσειρές μεγάλου μήκους, η σχετική απλότητα του και χρήση του από άλλους αλγορίθμους, αλλά στα βασικά αρνητικά του η έλλειψη πολύ μεγάλης ταχύτητας σε συνδυασμό με την απαιτούμενη ακρίβεια στους υπολογισμούς του.

Η βασική του λειτουργία συνοψίζεται στα επόμενα βήματα:

- Επιλογή μιας τυχαίας υπό- σειράς από τη χρονοσειρά.
- Υπολογισμός της απόστασή της με όλες τις άλλες υπό-σειρές χρησιμοποιώντας την MASS
- Ενημέρωση του Matrix Profile με τις ελάχιστες αποστάσεις που προέκυψαν.
- Επανάληψη της ανωτέρω διαδικασίας με νέα τυχαία υποσειρά.

Όσο συνεχίζεται η διαδικασία, το Matrix Profile συγκλίνει στις πραγματικές ελάχιστες αποστάσεις.

Οι κύριες εφαρμογές του STAMP είναι : η ανακάλυψη μοτίβων, η ανίχνευση ανωμαλιών, η τμηματοποίηση και η σύγκριση μεταξύ διαφορετικών χρονοσειρών (similarity joins).
Τα λειτουργικά που έχουν ενσωματωμένο τον STAMP στις βιβλιοθήκες τους είναι τα : PYTHON, R και MATLAB. Ακολουθούν στις παρακάτω εικόνες, ο βασικός κώδικας στο περιβάλλον του R - STUDIO :

```
ΛΕΙΤΟΥΡΓΙΚΟ : R
ΠΑΚΕΤΟ : tsmp

library(tsmp)  #κλήση βιβλιοθήκης
set.seed(42)   #αρχικοποίηση τυχαιότητας
data <- cumsum(sample(c(-1,1), 5000, replace=TRUE))  #
διάβασμα βάσης δεδομένων


mp_obj <- tsmp(data, window_size=100, mode="stamp",
               s_size = 0.1 * (length(data) - 100 + 1),
               exclusion_zone = 50, n_workers=2)
# Τρέξε STAMP με 10% δειγμα για 'anytime' προσέγγιση


mp_motif <- find_motif(mp_obj, n_motifs = 3, radius =
sd(mp_obj$mp))
mp_disc  <- find_discord(mp_obj, n_discords = 2)
# Βρες motifs και discords
```

ΕΙΚΟΝΑ 4.1.1 - ΒΑΣΙΚΟΣ ΚΩΔΙΚΑΣ STAMP ΣΤΗΝ R

Το πακέτο που χρησιμοποιεί η R είναι το tsmp [45] ενώ η δομή του κώδικα παρουσιάζει μια απλή σειρά από εντολές με σχετικό μικρό μέγεθος. [46]

Ο Αλγόριθμος STOMP είναι ο δεύτερος αλγόριθμος για την υπολογισμό του Matrix Profile, βελτιωμένος σημαντικά συγκριτικά με τον STAMP σε ταχύτητα.
Ο STOMP υπολογίζει το Matrix Profile μιας χρονοσειράς με διαδοχική (ordered) αξιολόγηση των distance profiles (πίνακας αποστάσεων) αντί για τυχαία σειρά όπως στον STAMP. [47]
Χρησιμοποιεί τη δομή εξάρτησης, δηλαδή την αλγεβρική σχέση ανάμεσα στους διαδοχικούς distance profiles, επιτρέποντας αναδρομικό υπολογισμό και μείωση της πολυπλοκότητας, καθώς δεν διενεργεί τον μετασχηματισμό FFT κάθε φορά. Πιο συγκεκριμένα, ο STOMP ‘ανακυκλώνει’ τους υπολογισμούς του προηγούμενου distance profile για να βγάλει το επόμενο, πιο γρήγορα και πιο αποτελεσματικά. [44] [45]

Η βασική του λειτουργία συνοψίζεται στα επόμενα βήματα:

- Υπολογίζει ‘dot products’ μεταξύ της πρώτης υπό-σειράς και του σήματος μέσω FFT

- Για κάθε επόμενο παράθυρο, ενημερώνει τους dot products αντί να τους υπολογίζει εκ νέου ως εξής: $Q_{i,j} = Q_{i-1,j-1} - T_{i-1,j-1} + T_{i+m-1,j+m-1}$ (2.4) , T : Χρονοσειρά, m: Μέγεθος παραθύρου, Q : Υπό -σειρά, I,j : Θέσεις παραθύρων
- Με τη νέα τιμή, υπολογίζει το distance profile, ενημερώνει το MP και MPi, εφαρμόζοντας ζώνη αποκλεισμού για αποφυγή επικαλύψεων
- Επανάληψη μέχρι να καλυφθούν όλα τα παράθυρα

Ακολουθούν (όπως και προηγουμένως με τον STAMP), οι αντίστοιχες εντολές στην R για τον αλγόριθμο STOMP :

```
ΛΕΙΤΟΥΡΓΙΚΟ : R
ΒΙΒΛΙΟΘΗΚΗ : tsmp

library(tsmp) # κλήση πακέτου

# Δημιουργία χρονοσειράς
set.seed(1)
ts = sin(seq(0, 20*pi, length.out = 5000)) + 0.5 * rnorm(5000)
m = 100

# Εκτέλεση STOMP
mp_obj = tsmp(ts, window_size = m, mode = "stomp", n_workers = 2)

# Εύρεση 3 motifs & 3 discords
mp_obj = find_motif(mp_obj, n_motifs = 3, n_neighbors = 5, radius = 3)
mp_obj = find_discord(mp_obj, k = 3, neighbor_count = 5, radius = 3)


# Εμφάνιση αποτελεσμάτων
print(mp_obj$motif_idx)
print(mp_obj$discord_idx)
plot(mp_obj)
```

ΕΙΚΟΝΑ 4.1.2 - ΒΑΣΙΚΟΣ ΚΩΔΙΚΑΣ STOMP ΣΤΗΝ R

Παραπάνω, αποτυπώνεται ένα αριθμητικό παράδειγμα, δημιουργίας τυχαίας Χρονοσειράς δεδομένων, στην R μέσω του πακέτου tsmp , με τα σχετικά βασικά ζητούμενα του Matrix Profile.

Ο αλγόριθμος SCRIMP (Scalable and Accurate Matrix Profile) είναι ένας αποδοτικός αλγόριθμος για τον υπολογισμό του Matrix Profile, με σκοπό την ανίχνευση μοτίβων και ανωμαλιών μεταξύ άλλων. [47]

Τα βασικά βήματα υλοποίησης του, είναι τα εξής:

- Υπολογίζει τις αποστάσεις (Euclidean) ανάμεσα σε όλες τις υπό σειρές ενός μήκους n της χρονοσειράς.
- Χρησιμοποιεί FFT (Fast Fourier Transform) και τεχνικές πρόβλεψης συσχετίσεων για ταχύτερο υπολογισμό.
- Υλοποιεί μια προοδευτική στο χρόνο , 'anytime' προσέγγιση ήτοι το αποτέλεσμα γίνεται πιο ακριβές όσο περνά ο χρόνος.

Τα κύρια πλεονεκτήματα του SCRIMP είναι η σημαντική αποδοτικότητα για μεγάλου μεγέθους χρονοσειρές (scalability), έχει προοδευτική ακρίβεια, δηλαδή βελτιώνεται με το χρόνο, μπορεί να τερματίσει τη λειτουργία του ανά πάσα στιγμή (anytime), καθώς επίσης ότι έχει εφαρμογή σε ποικίλα λειτουργικά συστήματα. Ο Αλγόριθμος αυτός έχει εφαρμογές και χρησιμοποιείται σε Βιοϊατρικά δεδομένα, Οικονομικά, γεωδυναμικά κλπ . Κύριο χαρακτηριστικό στον υπολογισμό των αποστάσεων, είναι το γεγονός ότι υπολογίζεται διαγώνια (σταδιακά), με δημιουργία υπό πινάκων, επιτρέποντας να σταματήσει οποιαδήποτε στιγμή.(Incremental)

Ακολουθεί στην παρακάτω εικόνα ο βασικός κώδικας του Αλγορίθμου SCRIMP στο λειτουργικό της R :

```
ΛΕΙΤΟΥΡΓΙΚΟ R
ΠΑΚΕΤΟ : tsmp

library(tsmp) # κλήση βιβλιοθήκης
set.seed(123) #αρχικοποίηση
ts <- rnorm(5000)
m <- 50
mp <- compute(ts, window_size = m, mode = "scrimp")
# επιστροφή αποτελέσματος
# Μοτίβα
motifs <- find_motif(mp, k=3)
# Ανωμαλίες
discords <- find_discord(mp, k=3)
```

ΕΙΚΟΝΑ 4.1.3 - ΒΑΣΙΚΟΣ ΚΩΔΙΚΑΣ ΤΟΥ SCRIMP ΣΤΗΝ R

Όπως και σε κάθε αλγόριθμο, το κύριο χαρακτηριστικού της R είναι η απλότητα στον κώδικα της. Επιστρέφει μέσω της εντολής *compute* τον πίνακα και τον αντίστοιχοι δείκτη.

Εν συνεχεία , μια εξέλιξη του ανωτέρω αλγορίθμου SCRIMP είναι ο αλγόριθμος ACAMP(Anytime Comprehensive Approximate Matrix Profile), εστιάζοντας στην ακρίβεια και την ταχύτητα, διατηρώντας τη δυνατότητα να σταματήσει οποτεδήποτε. (anytime) [48]

Τα βασικά βήματα του ACAMP είναι τα εξής:

- Υπολογίζει το Matrix Profile (το ελάχιστο απόστασης για κάθε υπό - σειρά από την πιο κοντινή μη επικαλυπτόμενη).
- Βασίζεται σε τεχνικές sliding dot product με FFT για να επιταχύνει τους υπολογισμούς.
- Ονομάζεται “Anytime” επειδή μπορεί να επιστρέψει προσεγγιστικό αποτέλεσμα σταδιακά, βελτιώνοντας με τον χρόνο (παρόμοια με SCRIMP αλλά πιο πλήρης από STAMP/STOMP).
- Παρέχει ολική κάλυψη των ζευγών υπό - σειρών — "comprehensive".

Αναλυτικότερα, το sliding dot product είναι η πράξη όπου υπολογίζεται το εσωτερικό γινόμενο (dot product) μεταξύ: μιας υπό σειράς Q μήκους m και κάθε υποσειράς του ίδιου μήκους με μέγεθος το δείγμα της Χρονοσειράς T. Η ανωτέρω διαδικασία είναι εν γένει αργή, αλλά επισπεύδεται χρονικά με τη βοήθεια του μετασχηματισμού FFT. Ακόμη με τον όρο ‘comprehensive’, εννοείται η δυνατότητα του αλγορίθμου να καλύπτει όλη τη Χρονοσειρά. Ο αλγόριθμος ACAMP είναι ιδανικός για μεγάλου μήκους datasets. [49] Στα βασικά μειονεκτήματα του ACAMP είναι το γεγονός ότι έχει μεγαλύτερη πολυπλοκότητα σε σχέση με τους STAMP και STOMP, ενώ δεν έχει τον ίδιο βαθμό συμβατότητας με όλα τα λειτουργικά (πχ δεν περιλαμβάνεται άμεσα στην PYTHON)
Όπως τονίστηκε και παραπάνω, για την πλήρη εφαρμογή του ACAMP, απαιτούνται επιπλέον χειροκίνητοι υπολογισμοί σε κάθε λειτουργικό.

Τέλος, ο Αλγόριθμος MPX (Matrix Profile eXact), είναι ένας από τους πρώτους και απλούς αλγορίθμους για τον υπολογισμό του Matrix Profile. Αναπτύχθηκε στο πλαίσιο της εργασίας των Yeh κ.α [9] όταν παρουσιάστηκε για πρώτη φορά η έννοια του Matrix Profile. Είναι και αυτός ένας αλγόριθμος που βασίζεται σε sliding dot product με τον μετασχηματισμό FFT.

Τα κύρια βήματα που ακολουθεί είναι τα εξής:

- Για κάθε υπό σειρά Q μήκους m στη χρονοσειρά T:

Υπολογίζει το distance profile με όλες τις άλλες υποσειρές του T μέσω FFT.

- Χρησιμοποιεί z-normalization για σταθερότητα και αντοχή σε κλίμακα.
- Εφαρμόζει exclusion zone (αποκλείει τις γειτονικές θέσεις για αποφυγή trivial matches).
- Ενημερώνει το Matrix Profile (MP) και το Profile Index (PI), αν βρεθεί μικρότερη απόσταση.

Στα βασικά του πλεονεκτήματα του MPX είναι η απλότητα στην εφαρμογή του, καθώς επίσης οτι είναι ιδανικός σε ‘’off line’’ δεδομένα. Τα αρνητικά του MPX είναι ότι σε αντίθεση με τους STOMP,SCRIMP δεν είναι incremental και ότι δεν είναι κατάλληλος για on line ανάλυση.

Κατά συνέπεια, ο αλγόριθμος MPX είναι ο πρώτος χρονικά που εμφανίστηκε και μάλιστα ως ‘exact’ αλγόριθμος στο πεδίο του Matrix Profile. Εν συνεχεία, ο STAMP είναι μια ‘’προσαρμογή’’ και εξέλιξη του MPX , καθώς υπολογίζει τα ‘distance profiles ‘ με τυχαία σειρά. Κύριο χαρακτηριστικό του τα πρόωρα αποτελέσματα. Επιπρόσθετα, ο STOMP είναι μια νέα, γρήγορη και ακριβής εκδοχή, που χρησιμοποιεί επαναληπτικά υπολογισμούς από τα σχετικά εσωτερικά γινόμενα (dot product) για κάθε νέα υπό σειρά. Από την άλλη ο αλγόριθμος SCRIMP, είναι μια υβριδική έκδοση, που διενεργεί τις εντολές τους αρχικά με βάση την τυχαιότητα και εν συνεχεία με ακρίβεια. Τέλος, ο
ACAMP έχει το κύριο χαρακτηριστικό την "χειροκίνητη" υλοποίηση, ενώ υπολογίζει πλήρως κάθε distance profile από την αρχή, χωρίς βελτιστοποιήσεις. Παρακάτω παρουσιάζονται 2 πίνακες, που συνοψίζουν τα κύρια χαρακτηριστικά κάθε ενός από τους ανωτέρω αλγορίθμους, καθώς επίσης και τη συμβατότητα τους με τα πιο δημοφιλή λειτουργικά.

ΠΙΝΑΚΑΣ 4.1.1- ΒΑΣΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ ΓΙΑ ΤΟΝ MATRIX PROFILE

ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ	MPX	STAMP	STOMP	SCRIMP++	ACAMP
Ακρίβεια	ΥΨΗΛΗ	ΠΡΟΣΕΓΓΙΣΤΙΚΑ	ΥΨΗΛΗ	ΥΨΗΛΗ	REFINEMENT
Τυχαιότητα στους υπολογισμούς	ΟΧΙ	ΝΑΙ	ΟΧΙ	ΝΑΙ	ΟΧΙ
Incremental	ΟΧΙ	ΟΧΙ	ΝΑΙ	ΝΑΙ	ΟΧΙ
Χρηση FFT	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ
Χρήση Εσωτερικού Γινομένου	ΝΑΙ	ΝΑΙ	REUSE	REUSE	ΝΑΙ
Ταχύτητα	ΜΕΤΡΙΑ	ΜΕΤΡΙΑ ΠΡΟΣ ΓΡΗΓΟΡΗ	ΠΟΛΥ ΓΡΗΓΟΡΗ	ΠΟΛΥ ΓΡΗΓΟΡΗ	ΑΡΓΟΣ
OFF/ON LINE	OFFLINE	OFFLINE	ON LINE	ON LINE	OFFLINE
Πολυπλοκότητα	O(n^2logn)	O(n^2logn)	O(n^2)	O(n^2)	O(n^2logn)

Επεξηγηματικά στον ανωτέρω πίνακα, ως ‘REFINEMENT’ ονομάζεται η διαδικασία προοδευτικής βελτίωσης της ακρίβειας ενός αποτελέσματος, ξεκινώντας από μια αρχική κατά προσέγγιση λύση, ενώ

ως ‘REUSE’ καλείται η τεχνική στην οποία, επαναχρησιμοποιούνται ενδιάμεσα αποτελέσματα από προηγούμενους υπολογισμούς για να αποφευχθεί την επανάληψη των ίδιων υπολογισμών.
Επίσης, σχετικά με την πολυπλοκότητα, ως ‘O(n^2logn)’ ορίζουμε τον βαθμό πολυπλοκότητας αυτό, σύμφωνα με τον οποίο, ο χρόνος εκτέλεσης αυξάνεται αναλογικά με το τετράφωνο της αρχικής πληροφορίας που εισήχθη, και ως ‘O(n^2)’ καλείται η πολυπλοκότητα που αυξάνει λογαριθμικά σε σχέση με την πληροφορία εισόδου.

ΠΙΝΑΚΑΣ 4.1.2 -ΣΥΜΒΑΤΟΤΗΤΑ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ ΜΕ ΤΑ ΛΕΙΤΟΥΡΓΙΚΑ ΣΥΣΤΗΜΑΤΑ

Αλγόριθμος	PYTHON	R	MATLAB	C++	ΒΙΒΛΙΟΘΗΚΕΣ
MPX	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	Stumpy, matrixprofile,TSclust
STAMP	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	Stumpy, matrixprofile
STOMP	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	Stumpy, matrixprofile
SCRIMP++	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	Stumpy, matrixprofile
ACAMP	ΧΕΙΡΟΚΙΝΗΤΗ ΥΛΟΠΟΙΗΣΗ	ΧΕΙΡΟΚΙΝΗΤΗ ΥΛΟΠΟΙΗΣΗ	ΧΕΙΡΟΚΙΝΗΤΗ ΥΛΟΠΟΙΗΣΗ	ΧΕΙΡΟΚΙΝΗΤΗ ΥΛΟΠΟΙΗΣΗ	

Συνοψίζοντας, ο αλγόριθμος MPX είναι κατάλληλος για εκπαιδευτικούς λόγους, ακριβής αλλά πιο αργός συγκριτικά. Ο STAMP, εργάζεται προσεγγιστικά, ενώ επιστρέφει γρήγορα ‘’πρόωρα’’ αποτελέσματα. Ο STOMP, είναι ταχύτατος ως αλγόριθμος, ιδανικός για ‘’real time’’ και μεγάλου μήκους Χρονοσειρές. Ο SCRIMP++ λειτουργεί ‘’υβριδικά’’ με αρχική προσέγγιση στο αποτέλεσμα και μετά με τη διαδικασία του refinement. Τέλος, ο ACAMP έχει μια κύρια χειροκίνητη υλοποίηση, αλλά μπορεί να συνεισφέρει στην κατανόηση της κεντρικής ιδέας του MATRIX PROFILE.

4.2 Σύγκριση του Matrix Profile με ανταγωνιστικούς Αλγορίθμους

Στον τομέα της Ανάλυσης Χρονοσειρών για την ανίχνευση μοτίβων και ανωμαλιών, ο Αλγόριθμος Matrix Profile εμφανίζει σημαντικά πλεονεκτήματα αλλά και κάποια επιμέρους μειονεκτήματα. Αρχικά, αποτυπώνει μια σημαντική ακρίβεια στον εντοπισμό μοτίβων και ανωμαλιών, μέσω της μεθοδολογίας της ευκλίδειας απόστασης αλλά και του σχετικού εσωτερικού γινομένου που περιέχεται στους εξειδικευμένους αλγόριθμους που χρησιμοποιεί. Επίσης, στα πλεονεκτήματα του MP, είναι το γεγονός ότι πέραν του ορισμού του μεγέθους παραθύρου, δεν απαιτείται καμία άλλη παραμετροποίηση για την υλοποίηση του. Επιπρόσθετα, στα θετικά του MP, αποτελεί το γεγονός της ‘’καθολικότητας’’ του στον έρευνα που διενεργεί, καθώς αναλύει όλες τις πιθανές υποσειρές της συνολικής Χρονοσειράς. Εν συνεχεία, κυρίως μέσω των αλγορίθμων STOMP και SCRIMP , εμφανίζει μια υψηλή αποδοτικότητα και ταχύτητα στα προς εξέταση ζητούμενα.

Από την άλλη, στα πιο εξειδικευμένα προβλήματα, ο MP εμφανίζει κάποιες αδυναμίες και μειονεκτήματα. Πιο συγκεκριμένα, τα αποτελέσματα έχουν άμεση εξάρτηση από την παράμετρο του μεγέθους παραθύρου, γεγονός που το καθιστά ‘’ευαίσθητο’ σ αυτήν την επιλογή. Ακόμη, όσον αφορά τις πολύπλοκες δομές δεδομένων (π.χ ανωμαλίες που έχουν συσχέτιση από το περιεχόμενο των δεδομένων), δεν είναι ο ιδανικότερος αλγόριθμος υλοποίησης. Επίσης, παρόλο που έχει εφαρμογή όπως

περιγράφηκε στο παρόν κείμενο, σε πολυδιάστατα δεδομένα, μειονεκτεί σε σχέση με τους 'ανταγωνιστές' όπως θα περιγραφεί και παρακάτω. Εν συνεχεία, σε συστήματα ανάλυσης που απαιτούν αποκλειστική on line υλοποίηση, πιθανά να εμφανίσει καθυστερήσεις , αναλόγως του ποιος επιμέρους αλγόριθμος θα χρησιμοποιηθεί. Τέλος, ο MP δύναται να εμφανίσει αδυναμίες και ευαισθησία αναλόγως της κλιμάκωσης(scalability) που μπορεί να απαιτηθεί στην ανάλυση του δείγματος, καθώς επίσης , σημαντικό προαπαιτούμενο της ορθής υλοποίησης του MP είναι πως πρέπει να χρησιμοποιήσει όλη τη Χρονοσειρά για ασφαλή αποτελέσματα.

Στον τομέα ανίχνευσης μοτίβων και ανωμαλιών, ποικίλες δομές δεδομένων και σχετικές μεθοδολογίες έχουν αναπτυχθεί πέραν του MP. Οι αλγόριθμοι που ξεχωρίζουν για την αποτελεσματικότητα και την εξειδίκευσή τους, είναι οι παρακάτω:

- DTW (Dynamic Type Warping)
- SAX (Synbolic Aggregate ApproXimation)
- Isolation Forest
- One Class SVM
- FB PROPHET

Συγκεκριμένα, ο Αλγόριθμος DTW (Dynamic Type Warping) έχει ως κεντρική ιδέα το ότι μετρά την απόσταση μεταξύ δύο χρονικών σειρών επιτρέποντας μη γραμμική ευθυγράμμιση (time warping), ώστε να προσαρμόζει αποτελέσματα που εμφανίζονται σε διαφορετικούς ρυθμούς με έντονη διακύμανση. [50] Αναλυτικότερα, ο DTW είναι ιδανικός για μοτίβα που αποτυπώνουν μεταβλητό ρυθμό, όπως στους τομείς 'αναγνώριση ομιλίας' ή χειρονομιών, και δύναται να χρησιμοποιηθεί για ανίχνευση ανωμαλιών συγκρίνοντας τμήματα με γνωστά μοτίβα. [51] Ακόμη, ο SAX, έχει ως κεντρική ιδέα το ότι μετατρέπει μια συνεχόμενη χρονική σειρά σε συμβολοσειρά (string) μέσω της μεθόδου της κβαντοποίησης, επιτρέποντας τη δημιουργία αλγορίθμων σύγκρισης, κατηγοριοποίησης για την ανίχνευση μοτίβων και ανωμαλιών. [52] Συγκεκριμένα, διενεργείται μια απλοποίηση των πολύπλοκων δεδομένων μειώνοντας τις διαστάσεις και εμφανίζει υψηλή αποδοτικότητα και συμβατότητα σε δεδομένα που έχουν δημογραφική ευαισθησία. Χρησιμοποιείται σε συνδυασμό με τον Matrix profile. [43] Εν συνεχεία, όσον αφορά τον Αλγόριθμο Isolation Forest, έχει ως κεντρική ιδέα υλοποίησης την εξής Μεθοδολογία: Απομονώνει σημεία ανωμαλίας στη Χρονοσειρά μέσω τυχαίων διαχωρισμών σε δέντρα απόφασης, με τη λογική να απομονώνονται πιο εύκολα οι ανωμαλίες από τα φυσιολογικά σημεία, μέσω του μικρότερου μήκους διαδρομής. Αναλυτικότερα, η ανωτέρω μέθοδος χαρακτηρίζεται από υψηλή ταχύτητα αλλά και συμβατότητα στην κλιμάκωση των δεδομένων. Ακόμη, δεν απαιτεί μεγάλη παραμετροποίηση και έχει αποτελεσματική λειτουργία σε υψηλές διαστάσεις. [26]
Σχετικά με τον One Class SVM, η ιδέα υλοποίησης αφορά την εκπαίδευση ενός σχετικού μοντέλου πάνω στα ''ομαλοποιημένα'' δεδομένα , δημιουργώντας μια 'σφαίρα' πάνω στην οποία, συγκρίνονται όλα τα λοιπά δεδομένα για ανίχνευση ανωμαλιών. Η ανωτέρω μέθοδος, είναι κατάλληλη για δεδομένα

που εμφανίζουν περιορισμένες έως καθόλου ανωμαλίες. [18] [53] Τέλος, όσον αφορά τον FB Prophet, χρησιμοποιεί ένα μοντέλο συσχέτισης της τάσης, της εποχικότητας και του τυχαίου σφάλματος, που εμφανίζει ευκολία στην υλοποίηση, αλλά και δυνατότητα αυτόματης επιλογής 'υπερπαραμέτρων', δηλαδή παραμέτρων που ορίζονται εξωτερικά. Η μεθοδολογία είναι βασισμένη σε μια ''αθροιστική'' παλινδρόμησή, με άθροιση των επιμέρους συνιστωσών που αναφέρθηκαν παραπάνω , ήτοι: Τάση, Εποχικότητα, τυχαιότητα. Η κύρια συμβατότητα των ανωτέρω μεθόδου είναι σε χρονοσειρές που εμφανίζουν σαφείς εποχικότητες και τάσεις. [54]

Στον παρακάτω πίνακα, συνοψίζονται τα συγκριτικά στοιχεία μεταξύ του Matrix Profile και των ανωτέρω αλγορίθμων, ως προς τα συνήθη επιστημονικά προβλήματα που αντιμετωπίζουν.

ΠΙΝΑΚΑΣ 4.2.1 -ΣΥΓΚΡΙΣΗ MATRIX PROFILE ΚΑΙ ΛΟΙΠΩΝ ΑΛΓΟΡΙΘΜΩΝ

Κριτήριο	Matrix Profile	DTW	SAX	Isolation Forest	One Class SVM	FB PROPHET
Ακρίβεια για μοτίβα	Υψηλή	Υψηλή	Μέτρια	Καλή	Καλή	Χαμηλή
Ανίχνευση Ανωμαλιών	Πολύ Καλή	Καλή	Μέτρια	Πολύ Καλή	Πολύ Καλή	Μέτρια
Ερμηνευσιμότητα	Πολύ Καλή	Μέτρια	Υψηλή	Χαμηλή	Χαμηλή	Καλή
Πολυδιαστατικότητα	Χαμηλή	Χαμηλή	Υψηλή	Υψηλή	Υψηλή	Υψηλή
Real Time Υλοποίηση	Μέτρια	Χαμηλή	Υψηλή	Υψηλή	Υψηλή	Υψηλή
Ταχύτητα στην υλοποίηση	Πολύ γρήγορη	Αργή	Πολύ γρήγορη	Αργή	Αργή	Υψηλή

Κεφάλαιο 5ο : Εφαρμογή του Matrix profile

Για τις ανάγκες του παρόντος , έχει διενεργηθεί η μεθοδολογία του Matrix Profile στο πρόγραμμα R STUDIO , μέσω του Αλγορίθμου STOMP, σε 3 διαφορετικά set δεδομένων, για την εξεύρεση μοτίβων και ανωμαλιών.

5.1 Μοτίβα και ανωμαλίες στο σύνολο δεδομένων: Muscle Activation

Το πρωτογενές σύνολο δεδομένων ονομάζεται 'Muscle Activation' και αντλήθηκε από τη δεδομένων του Github [55] και χρησιμοποιήθηκε πρωτογενώς σε έρευνα του 2017 [56]. Αφορά ένα δείγμα 29000 παρατηρήσεων που ''ταλαντώνονται'' γύρω από το 0 με θετικές και αρνητικές τιμές. Πιο συγκεκριμένα, αποτυπώνει τις 'ενεργοποιήσεις των μυών' και τα ηλεκτρικά σήματα που παράγονται απ τους μυες των υπό εξέταση αθλητών, κάτω από συνθήκες πατινάζ ταχύτητας. Το συνολικό χρονικό διάστημα είναι διάρκειας 30 δευτερολέπτων, ήτοι συχνότητα δειγματοληψίας = 100 Hz, δηλαδή 100 δείγματα ανά δευτερόλεπτο.

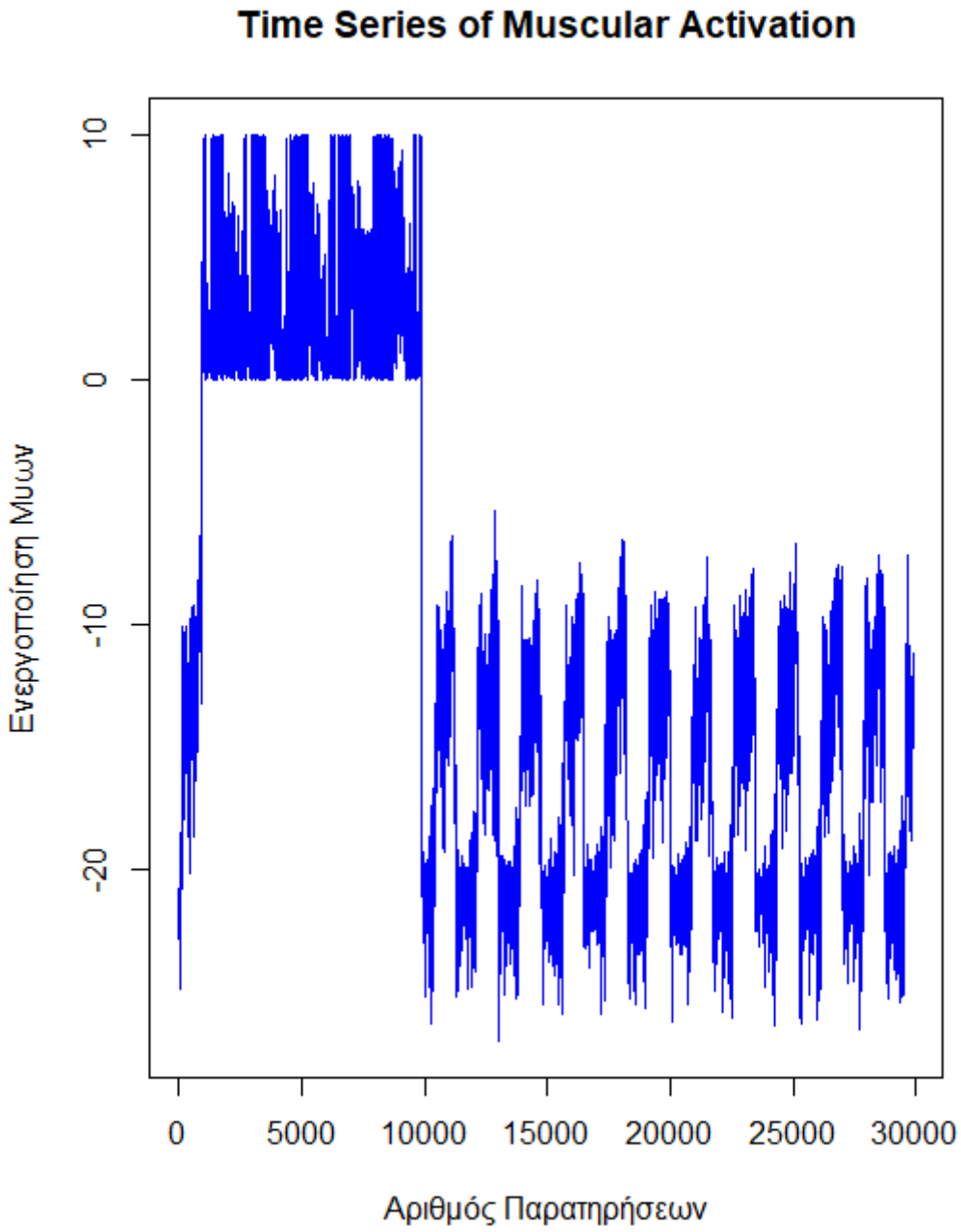
Στον Πίνακα 5.1.1 καταγράφονται οι βασικές στατιστικές μετρήσεις της μεταβλητής που εξετάζεται. Στο περιβάλλον της R έχουν εκτελεστεί οι εντολές : *str(), mean(), max(), min()* κ.α στη μεταβλητή που ορίστηκε με τη σχετική ανάθεση και ανάλογη ονομασία στο περιβάλλον του R-STUDIO.

ΠΙΝΑΚΑΣ 5.1.1 : Βασικά Στατιστικά στοιχεία της υπό εξέτασης μεταβλητής | Muscle Activation

Μέση τιμή	Διάμεσος	Τυπική Απόκλιση	Εύρος	Αριθμός παρατηρήσεων	
-10.85805	-12.93387	9.881005	-26.98534 9.99999	29900	
Τεταρτημόρια	0%	25%	50%	75%	100%
	-26.98534	-20.34564	-12.93387	0.62598	9.99999

Από τον ανωτέρω πίνακα , μπορούν να εξαχθούν κάποια πρώτα βασικά συμπεράσματα σχετικά με την υπό εξέταση μεταβλητή του παρόντος. Κατ αρχάς, οι 29000 παρατηρήσεις αφορούν κατά προσέγγιση την μέτρηση των ενεργοποιήσεων σε 30 δευτερόλεπτα (100 ανά λεπτό). Γίνεται σαφές, ότι τα δεδομένα είναι ασύμμετρα προς τα αρνητικά, καθώς κάτω απ το 50% των παρατηρήσεων είναι υπό του μηδενός. Ακόμη, έχει σημαντική στατιστική τυπική απόκλιση ίση με 9.88 , κάτι που επιβεβαιώνεται και από το εύρος (-26 έως 9.99).

Επιπρόσθετα , για μια πρώτη αποτύπωση και έρευνα των πρωτογενών δεδομένων, παρουσιάζεται στο Διάγραμμα 5.1.1, το διάγραμμα ροής του διανύσματος μέσω της εντολής: *Timeplot(),* με τον άξονα Y και X να αποτυπώνουν τον παράγοντα χρόνο και το διάνυσμα προς εξέταση, αντίστοιχα.



ΔΙΑΓΡΑΜΜΑ 5.1.1: Διάγραμμα ροής | Muscle Activation

Όπως γίνεται αντιληπτό από το ανωτέρω γράφημα, οι ενεργοποιήσεις των μυών ταλαντώνονται γύρω από το 0 και συσσωρεύονται κατά πλειοψηφία στις αρνητικές τιμές.

Ακόμη, παρουσιάζεται έστω και προκαταρκτικά κάποιες αιχμές και κορυφές στα πρωτογενή δεδομένα, στοιχείο που πρέπει να επιβεβαιωθεί μέσω της ανάλυσης του Matrix Profile.

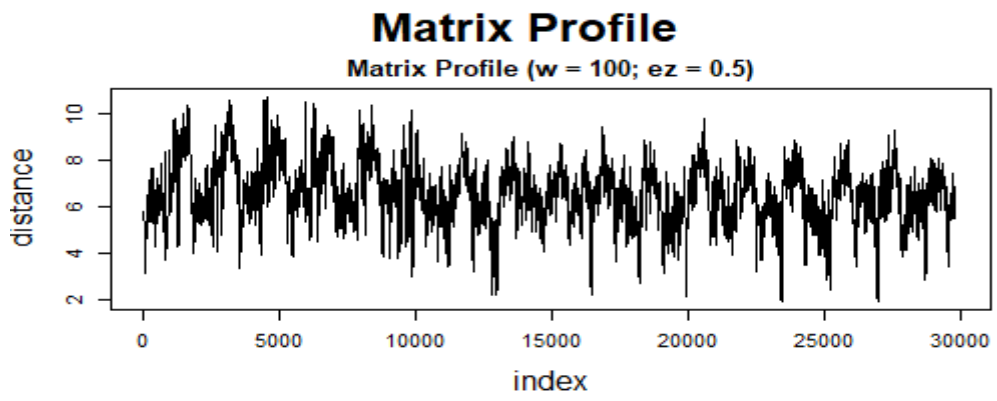
Περαιτέρω, μέσω της βιβλιοθήκης *tsmp* και των εντολών που την ακολουθούν, γίνεται μια πρώτη παρουσιάση των μοτίβων που ανιχνεύτηκαν με μέγεθος παραθύρου ίσο με 100. Ο ορισμός του μεγέθους του παραθύρου είναι μια κρίσιμη παράμετρος και επηρεάζει στις περισσότερες των περιπτώσεων τα αποτελέσματα που αποτυπώνονται.
Μέσω της εντολής print(mp) παρουσιάζεται ο Πίνακας 5.1.2 με τα ευρισκόμενα μοτίβα στο dataset. Αυτόματα, στις ανωτέρω εντολές έχει ενσωματωθεί η ‘’Z κανονικοποίηση’’ -οχι του διανύσματος - αλλά των ορισμένων παραθύρων.

ΠΙΝΑΚΑΣ 5.1.2: Παρουσίαση μοτίβων | Muscle Activation

Αριθμός Παρατηρήσεων	Μέγεθος παραθύρου	Ζώνη αποκλεισμού	Χρόνος περάτωσης	Αριθμός μοτίβων που ανιχνεύτηκαν
29801	100	50	2.79 λεπτά	3
Διανύσματα μοτίβων				
[23414, 26948]	[58, 19079]	[10790, 17334]		

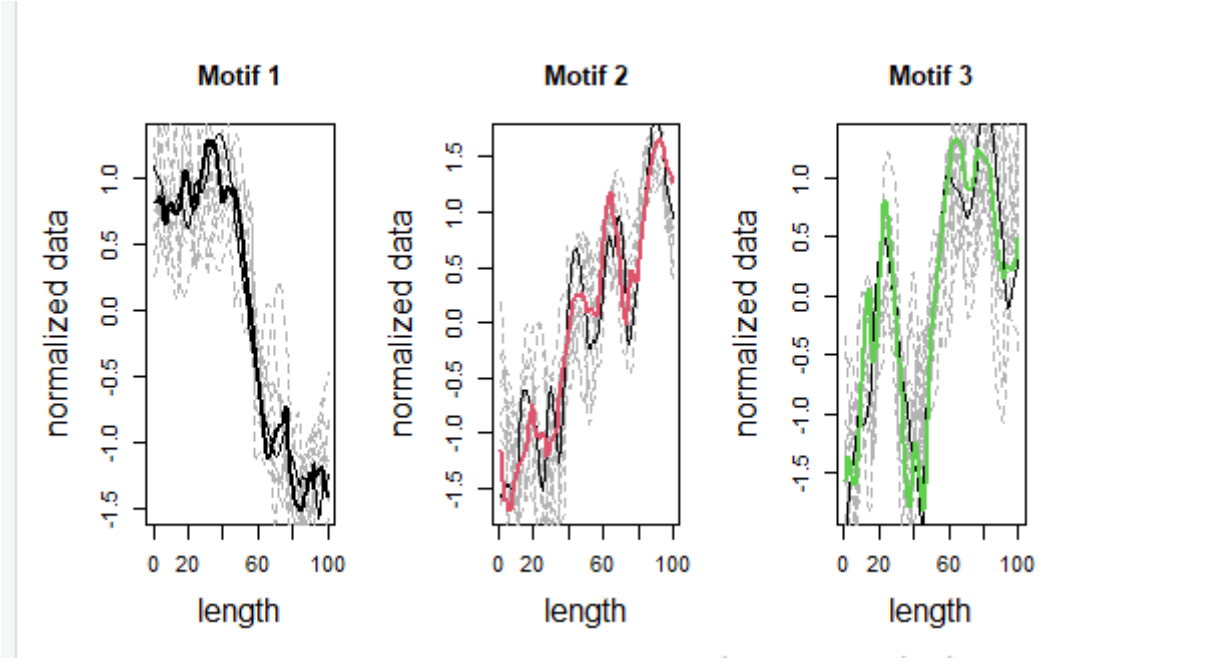
Από τον παραπάνω πίνακα, εξάγεται το συμπέρασμά ότι με τη διενέργεια του αλγορίθμου STOMP στα πρωτότυπα δεδομένα, ανιχνεύτηκαν 3 μοτίβα. Συγκεκριμένα, από τα ανωτέρω διανύσματα, φαίνεται πως το υποσύνολο που ξεκινάει απ την παρατήρηση 23414 έχει παρόμοια ‘’συμπεριφορά’’ με το αντίστοιχο υποσύνολο που ξεκινά από την παρατήρηση 26948, με οριζόμενο μέγεθος παραθύρου ίσο με 100. Αντιστοίχως ισχύουν τα ανάλογα για τα διανύσματα : [58, 19079] & [10790, 17334] του Πίνακα 5.1.2.

Ακόμη, μέσω της εντολής *plot(mp)* η R επιστρέφει το ‘’κεντρικό διάγραμμα’’ των μοτίβων , που στον άξονα X έχει τις παρατηρήσεις και στον άξονα Y τις αποστάσεις που συγκρίνονται ανά υποενότητα και σύμφωνα με το μέγεθος παραθύρου.



ΔΙΑΓΡΑΜΜΑ 5.1.2 : Γραφική απεικόνιση του Matrix Profile | Muscle Activation

Αναλυτικότερα, στον άξονα του Y φαίνονται οι αποστάσεις μεταξύ των υποενοτήτων και πιο συγκεκριμένα χαμηλές τιμές του άξονα Y καταδεικνύουν μοτίβα, ενώ υψηλές πιθανές ανωμαλίες. Τα ανωτέρω σε συνδυασμό με τα στοιχεία του Πίνακα 5.1.2, επαναβεβαιώνεται, καθώς τα διανύσματα των ευρισκόμενων Μοτίβων εμφανίζουν ‘’τοπικά ελάχιστα’’ στο Διάγραμμα 5.1.2



ΔΙΑΓΡΑΜΜΑ 5.1.3 : ΜΟΤΙΒΑ ΠΟΥ ΑΝΙΧΝΕΥΤΗΚΑΝ | Muscle Activation

Μέσω της εντολής: Print(motifs) η R έχει τη δυνατότητα να αναπαραστήσει γραφικά ξεχωριστά το κάθε κύριο μοτίβο, ακριβώς όπως φαίνεται και στο ανωτέρω Διάγραμμα 5.1.2.
Στον άξονα X είναι οι παρατηρήσεις της κάθε υποενότητας που έχουν μέγιστο μέγεθος 100 , όσο το μέγεθος παραθύρου. Στον άξονα των Y αποτυπώνονται ‘’Z κανονικοποιημένες’’ οι τιμές των σημείων του μοτίβου. Θετικές τιμές κοντά στο 1 καταδεικνύουν σημεία πάνω από το μ.ο και τιμές κοντά στο -1 το αντίστροφο.

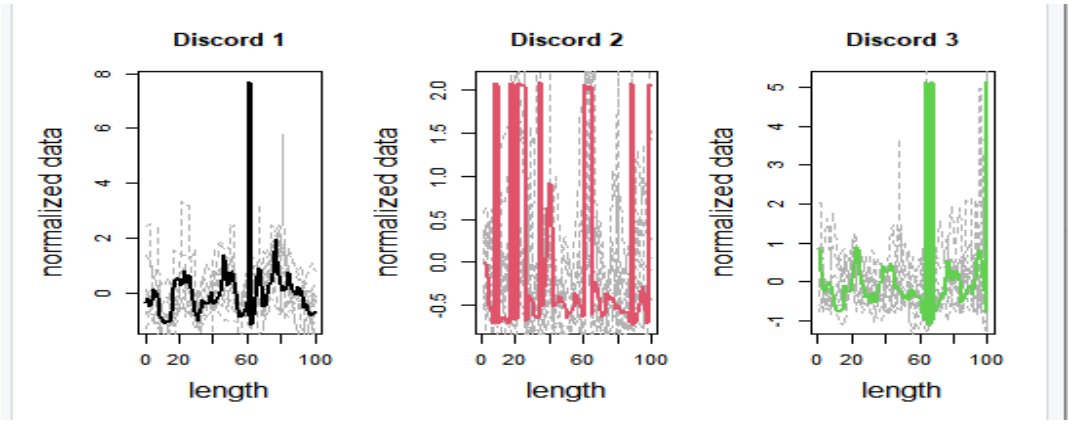
Εν συνεχεία, με την ίδια λογική και τις εντολές print, plot εμφανίζονται σε πίνακα αλλά και σε αντίστοιχα διαγράμματα οι ευρισκόμενες ανωμαλίες που αποθηκεύονται στη μεταβλητή discords και αποτυπώνονται αμέσως παρακάτω στον Πίνακα 5.1.3 και στα διαγράμματα 5.1.4 & 5.1.5.

Πίνακας 5.1.3 : Παρουσίαση ανωμαλιών | Muscle Activation

Αριθμός Παρατηρήσεων	Μέγεθος παραθύρου	Ζώνη αποκλεισμού	Χρόνος περάτωσης	Αριθμός ανωμαλιών που ανιχνεύτηκαν
29801	100	50	2.79 λεπτά	3
Διανύσματα ανωμαλιών				
[4539]	[4428]	[3157]		

Σύμφωνα με τον ανωτέρω πίνακα , βρέθηκαν 3 υποενότητες που αποτελούν ανωμαλίες στο υπό εξέταση δείγμα και παρουσιάζουν κορυφές και τοπικά μέγιστα στο Διάγραμμα 5.1.2 και στα αντίστοιχα σημεία που αναφέρονται στον ανωτέρω πίνακα.

Παρακάτω παρατίθενται και ξεχωριστά σε 3 διαφορετικές παραστάσεις τα 3 κύρια discords του δείγματος.



ΔΙΑΓΡΑΜΜΑ 5.1.4 : ΑΝΩΜΑΛΙΕΣ ΠΟΥ ΑΝΙΧΝΕΥΤΗΚΑΝ | Muscle Activation

Η Βασική διαφορά μεταξύ των διαγραμμάτων 5.1.3 και 5.1.4 είναι πως το 5.1.4 επειδή παρουσιάζει τα κύρια discords του δείγματος, εμφανίζει outliers στον άξονα των Y, δηλαδή ακραία υψηλές τιμές σε διακύμανση στην κάθε υποενότητα.

Οι ανωτέρω ενέργειες στο περιβάλλον του R -STUDIO για τις ανάγκες του παρόντος, συνοψίζονται ως εξής:

- Υλοποίηση του STOMP μέσω του πακέτου tsmp
- Φόρτωση αυτού του πακέτου με την εντολή library(tsmp).
- Ορισμός της παράμετρου 'window_size', η οποία καθορίζει το μήκος των υποσειρών
- Εκτέλεση τον STOMP μέσω του: tsmp(ts = data1, window_size = window_size)
- Χρησιμοποίηση του FFT, για να υπολογίσει αποστάσεις ανάμεσα σε υποσειρές με υψηλή υπολογιστική απόδοση.
- Το αποτέλεσμα αυτής της εντολής είναι ένα αντικείμενο που περιέχει το matrix profile, δηλαδή τις μικρότερες αποστάσεις κάθε υποσειράς με άλλες, καθώς και τους αντίστοιχους δείκτες των πιο παρόμοιων περιοχών.
- Παρουσίαση των βασικων στοιχείων του αποτελέσματος, όπως τις τιμές του matrix profile και τις θέσεις των αντίστοιχων υποσειρών, μέσω του print(mp_result).
- Οπτικοποίηση του matrix profile μέσω της εντολής plot(mp_result)
- Οι περιοχές με χαμηλές τιμές αντιπροσωπεύουν περιοχές της χρονοσειράς που είναι παρόμοιες με άλλες(υποψήφια μοτίβα), ενώ οι υψηλές τιμές δείχνουν σημεία που είναι μοναδικά και δεν μοιάζουν με άλλες περιοχές.
- Εντοπισμός των σημειών της χρονοσειράς που επαναλαμβάνονται με τον μικρότερο βαθμό απόκλισης, μέσω του motifs(mp_result), για την εύρεση των πιο ισχυρών μοτίβων.
- Απεικόνιση επιτρέπει την αυτών των μοτίβων επάνω στη χρονοσειρά, παρέχοντας ένα οπτικό αποτέλεσμα μέσω της εντολής plot(motif_result)
- Χρήση της συνάρτησης discords(mp_result) για την ανίχνευση ανωμαλιών ή ακραίων τιμών, η οποία βρίσκει τις υποσειρές με τις μεγαλύτερες αποστάσεις από οποιαδήποτε άλλη.
- Γραφική απεικόνιση των discords με την εντολή plot(discord_result)

5.2 Μοτίβα και ανωμαλίες στο σύνολο δεδομένων: Terminate Dna

Στην ενότητα αυτή, θα εκτελεστεί ο MP στο περιβάλλον του R- STUDIO, μέσω του αλγορίθμου STOMP, στο set δεδομένων που ονομάζεται 'Terminate Dna'. Το ανωτέρω dataset είναι αποθηκευμένο στη σχετική βάση δεδομένων της google για τον SKIMP [53] , ενώ χρησιμοποιήθηκε πρωτογενώς στην έρευνα του 2010 [10]. Αφορά μια χρονοσειρά απο Μιτοχονδριακά στοιχεία του Dna. Τα πρωτότυπα δεδομένα είναι σε μορφή *.mat*, δηλαδή συμβατά με την Matlab. Για τον λόγο αυτό, απαραίτητη είναι η εγκατάσταση και η κλήση της βιβλιοθήκης *R.MATLAB,* καθώς και για το 'διάβασμα' του αρχείου μέσω της εντολής *readMat.*

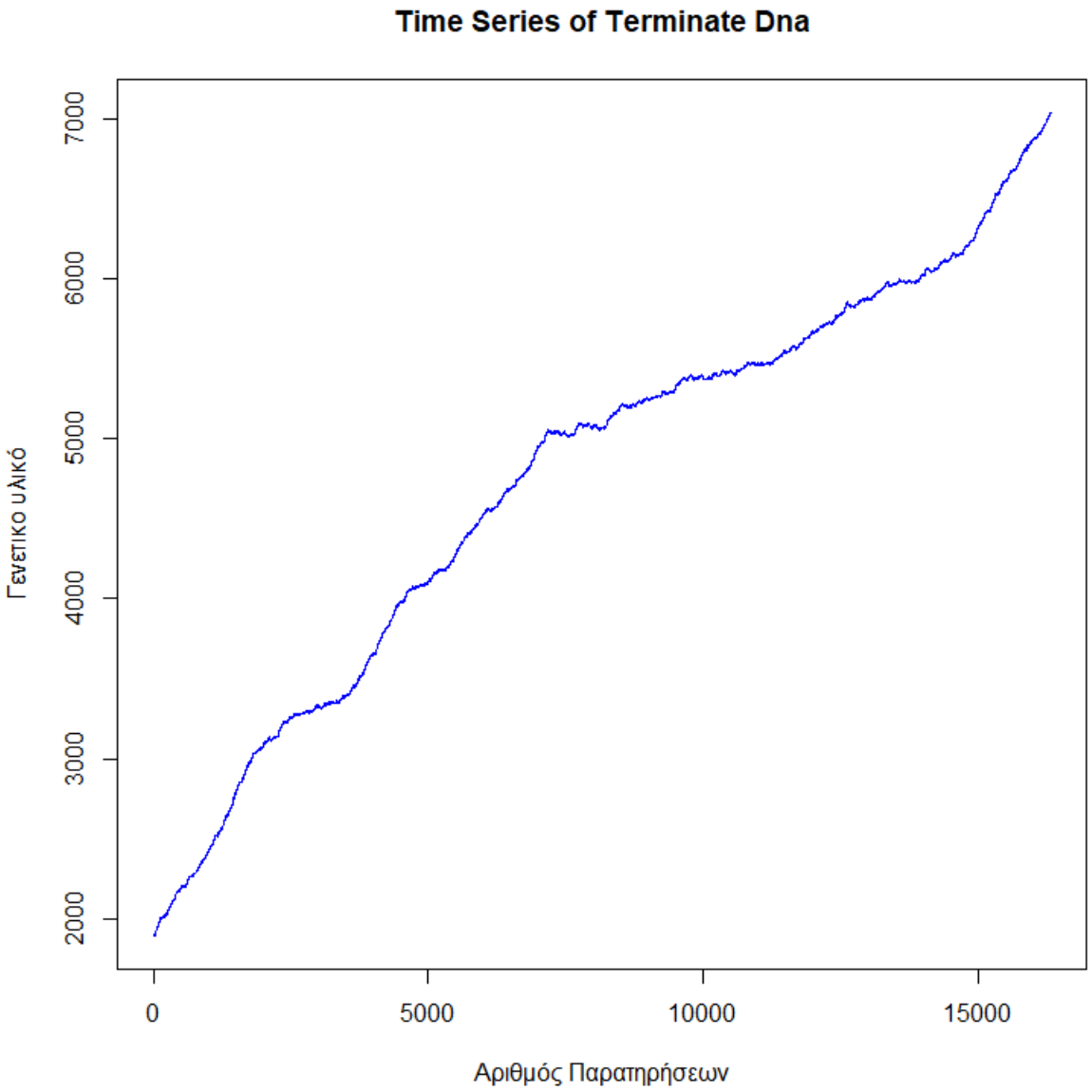
Μία ακόμη διαφοροποίηση σε σχέση με το σετ δεδομένων της ενότητας 5.1, είναι πως τα πρωτότυπα δεδομένα, δεν είναι αποθηκευμένα σε αριθμητικό διάνυσμα , αλλά σ ένα ιδιότυπο ‘πίνακα’, επομένως για να τρέξει ο STOMP, απαραίτητη είναι η χρήση της εντολής *as.vector.*
Κατά τα λοιπά, ο κώδικας που εκτελείται δεν έχει καμία διαφοροποίηση σε σχέση με την ενότητα 3.1 και τα βασικά αποτελέσματα συνοψίζονται ως εξής:

Στον Πίνακα 5.2.1 καταγράφονται οι βασικές στατιστικές μετρήσεις της μεταβλητής που εξετάζεται. Έχουν εκτελεστεί οι εντολές : *str(), mean(), max(), min()* κ.α στη μεταβλητή που ορίστηκε με τη σχετική ανάθεση και ανάλογη ονομασία στο περιβάλλον του R-STUDIO.

ΠΙΝΑΚΑΣ 5.2.1 : Βασικά Στατιστικά στοιχεία της υπό εξέτασης μεταβλητής | Terminate Dna

Μέση τιμή	Διάμεσος	Τυπική Απόκλιση	Εύρος	Αριθμός παρατηρήσεων	
4764.97	5091	1280.272	1893 7041	16326	
Τεταρτημόρια	0%	25%	50%	75%	100%
	1893	3715	5091	5718	7041

Από τον ανωτέρω πίνακα, μπορεί να εξαχθεί το πρώτο συμπέρασμα, ότι εφόσον η διάμεσος είναι μεγαλύτερη της μέσης τιμής, τα δεδομένα έχουν μια ‘’αριστερή ουρά’ , ασυμμετρία δηλαδή προς τις μικρότερες τιμές. Εν συνεχεία, αποτυπώνεται μια σχετικά υψηλή διασπορά γύρω από τη μέση τιμή, κάτι που ‘ενισχύεται’ από το μεγάλο εύρος των παρατηρήσεων. Τέλος, η ασυμμετρία επιβεβαιώνεται και από το εύρημα που δείχνει πως η απόσταση από το πρώτο τεταρτημόριο εως το διάμεσο, είναι μεγαλύτερη, από την αντίστοιχη του τρίτου τεταρτημορίου απ’ τη διάμεσο.



ΔΙΑΓΡΑΜΜΑ 5.2.1: Διάγραμμα ροής | Terminate Dna

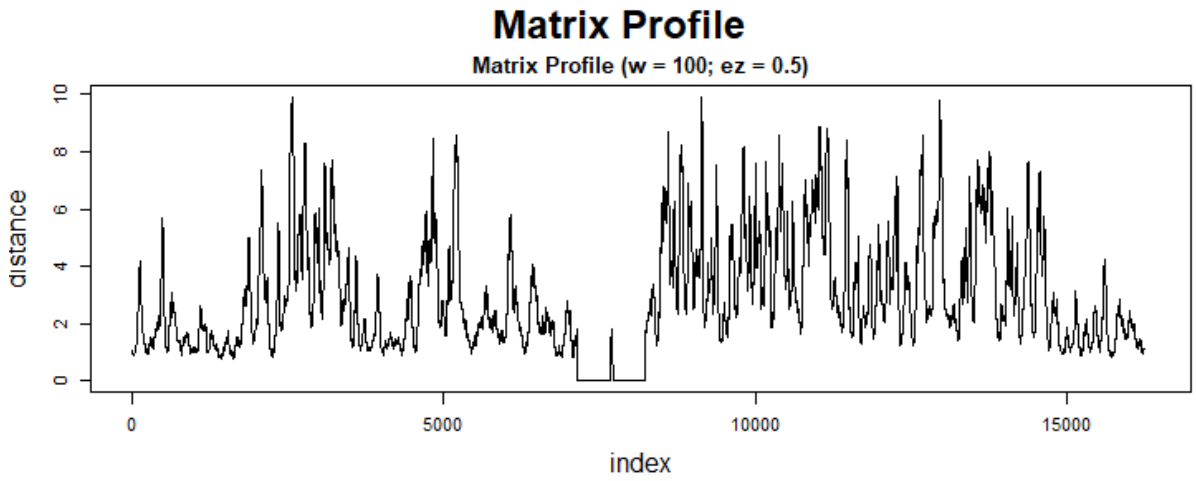
Από το ανωτέρω γράφημα, γίνεται αντιληπτό, πως οι περισσότερες παρατηρήσεις έχουν σωρευθεί στις μικρότερες τιμές, ενώ αποτυπώνεται και μια ανερχόμενη τάση των τιμών.

ΠΙΝΑΚΑΣ 5.2.2: Παρουσίαση μοτίβων | Terminate Dna

Αριθμός Παρατηρήσεων	Μέγεθος παραθύρου	Ζώνη αποκλεισμού	Χρόνος περάτωσης	Αριθμός μοτίβων που ανιχνεύτηκαν
16772	100	50	56''	3
Διανύσματα μοτίβων				
[7149, 7714]	[7200, 7765]	[7251, 7816]		

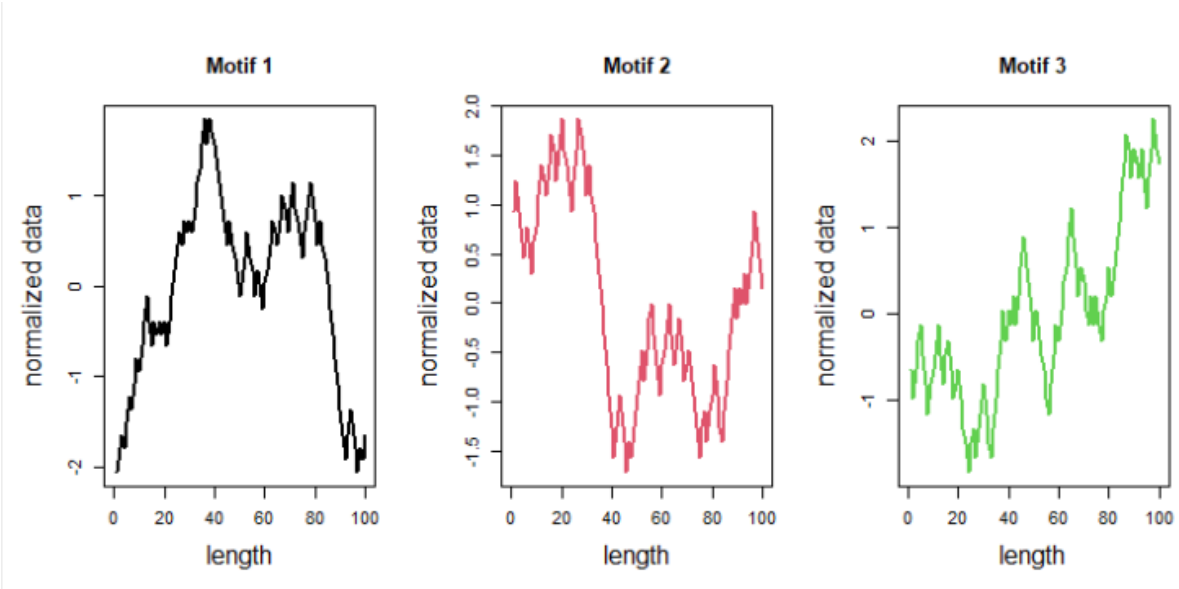
Από τον ανωτέρω πίνακα, εμφανίζονται (ομοίως με την ενότητα 3.1), 3 επαναλαμβανόμενα μοτίβα , σε συγκεκριμένες θέσεις (διανύσματα), ήτοι : [7149, 7714] , [7200, 7765] , [7251, 7816], κάτι που θα επιχειρηθεί να

επιβεβαιωθεί και από το επόμενο γράφημα.



ΔΙΑΓΡΑΜΜΑ 5.2.2 : Γραφική απεικόνιση του Matrix Profile | Terminate Dna

Από το ανωτέρω γράφημα, επαναβεβαιώνονται τα τοπικά ελάχιστα στις περιοχές που εμφανίζονται στον ανωτέρω πίνακα , δηλαδή στις παρατηρήσεις κοντά στο 7000. Αυτό, δείχνει την ύπαρξη επαναλαμβανόμενων μοτίβων στα σημεία αυτά, για το υπό εξέταση σετ δεδομένων.



ΔΙΑΓΡΑΜΜΑ 5.2.3 : ΜΟΤΙΒΑ ΠΟΥ ΑΝΙΧΝΕΥΤΗΚΑΝ | Terminate Dna

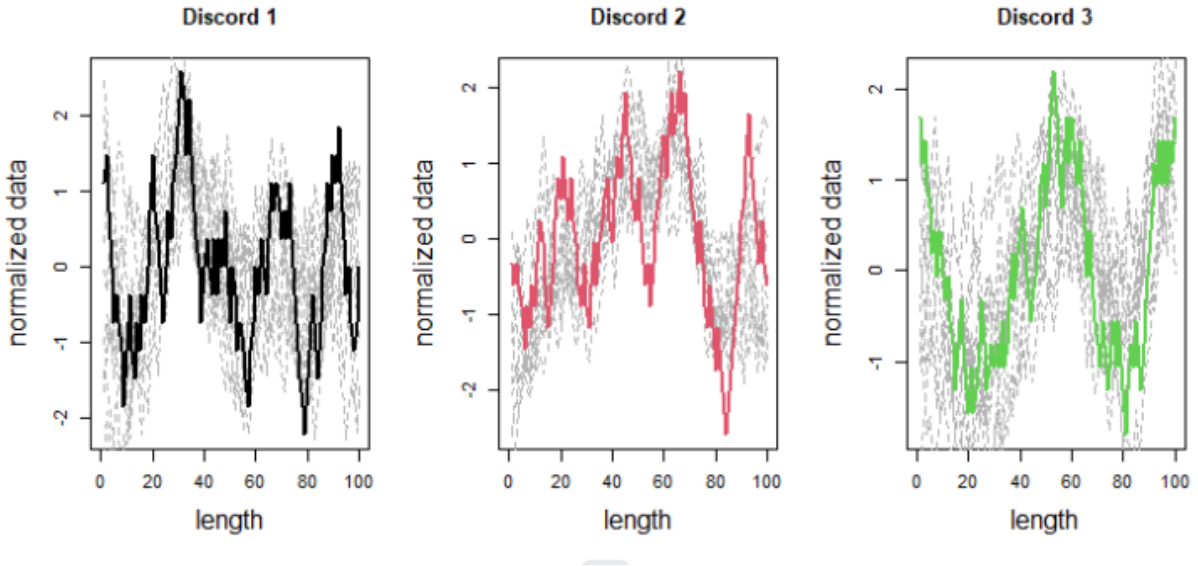
Σε συνέχεια της ενότητας 5.1, με την ίδια εκτέλεση στο R-STUDIO, παρουσιάζονται τα κύρια μοτίβα σε ένα κοινό γράφημα. Γίνεται αντιληπτό, το φαινόμενο των έντονων διακυμάνσεων γύρω απο το 0 του κάθετου άξονα.

Πίνακας 5.2.3 : Παρουσίαση ανωμαλιών | Terminate Dna

Αριθμός Παρατηρήσεων	Μέγεθος παραθύρου	Ζώνη αποκλεισμού	Χρόνος περάτωσης	Αριθμός ανωμαλιών που ανιχνεύτηκαν
16772	100	50	56''	3
Διανύσματα ανωμαλιών				
[2572]	[11141]	[8609]		

Μέσω του ανωτέρω πίνακα, αποτυπώνονται συνοπτικά τα κύρια διανύσματα που αφορούν ανωμαλίες στο υπό εξέταση σετ δεδομένων. Τα εν λόγω διανύσματα, επιβεβαιώνονται και από το διάγραμμα 5.2.2, ως τοπικά μέγιστα στο ''κεντρικό'' γράφημα του MP για τα δεδομένα που εξετάζονται.

Τέλος, στο παρακάτω γράφημα, παρουσιάζονται συνοπτικά σ ένα κοινό διάγραμμα οι κύριες ανωμαλίες του σετ δεδομένων που αναλύεται.



ΔΙΑΓΡΑΜΜΑ 5.2.4 : ΑΝΩΜΑΛΙΕΣ ΠΟΥ ΑΝΙΧΝΕΥΤΗΚΑΝ |Terminate Dna

5.3 Μοτίβα και ανωμαλίες στο σύνολο δεδομένων: Eog Multiple Scale

Στην ίδια βάση δεδομένων με το σετ της ενότητας 5.2, καθώς επίσης και στην ίδια δημοσίευση του 2010 [10], είναι αποθηκευμένο το dataset με ονομασία : Electro-oculography Dataset, όπου αφορά τα ηλεκτρικά σήματα της αριστερής όρασης των υπό εξέταση υποκειμένων. Οι καταγραφές εμφανίζουν συχνότητα 50 Hz. Ο σκοπός χρήσης είναι η αναζήτηση μοτίβων, μέσω των καταγραφών απόκλισης του σήματος της όρασης λόγω του εύρους μέτρησης.
Αντίστοιχα, με τις προηγούμενες ενότητες, για τις ανάγκες του παρόντος, θα διενεργηθεί ο MP μέσω του STOMP στο R -STUDIO.
Εκτελούνται -σε αντιστοιχία με την ενότητα 5.2- οι κατάλληλοι μετασχηματισμοί των πρωτότυπων δεδομένων, καθώς επίσης και οι συμβατές εντολές λόγω του γεγονότος πως τα δεδομένα είναι της μορφής *.mat*

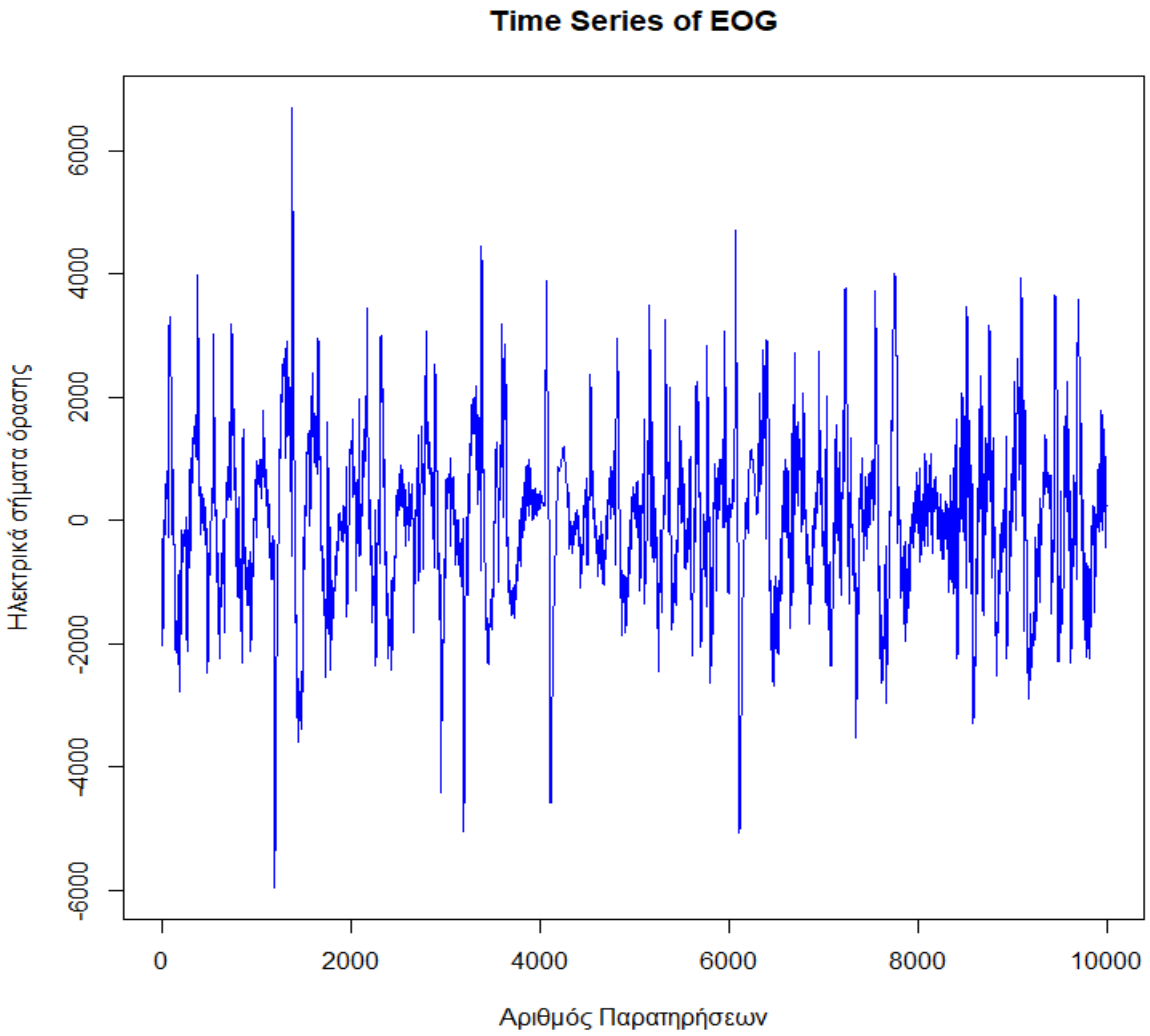
Προτού, παρουσιαστούν τα βασικά αποτελέσματα του MP στο ανωτέρω σετ, θα συνοψιστούν τα κύρια στατιστικά μέτρα της υπό εξέτασης μεταβλητής, στον ακόλουθο πίνακα.

ΠΙΝΑΚΑΣ 5.3.1 : Βασικά Στατιστικά στοιχεία της υπό εξέτασης μεταβλητής | Eog Multiple Scale

Μέση τιμή	Διάμεσος	Τυπική Απόκλιση	Εύρος	Αριθμός παρατηρήσεων	
23.35	66.6	1306	-5962 6697	10000	
Τεταρτημόρια	0%	25%	50%	75%	100%
	-5692	-762	66.6	722	6697

Από τα ευρήματα του ανωτέρω πίνακα, το στατιστικά πιο χρήσιμο σε σχέση με τα υπό εξέταση ζητήματα του παρόντος, είναι η έντονη αρνητική ασυμμετρία και οι πολλές αρνητικές τιμές, φαινόμενο που πιθανά να οδηγήσει σε ανίχνευση ανωμαλιών, κάτι που θα πρέπει να επιβεβαιωθεί από τα παρακάτω αποτελέσματα.

Στο παρακάτω αρχικό διάγραμμα της υπό εξέτασης μεταβλητής, αποτυπώνεται ή έντονη διακύμανση της, οι πολλές αρνητικές τιμές των παρατηρήσεων της. Τα παραπάνω αποτελούν μια πρώτη ένδειξη για την ανίχνευση μοτίβων και ανωμαλιών. Εν συνεχεία, στον πίνακα 5.3.1,



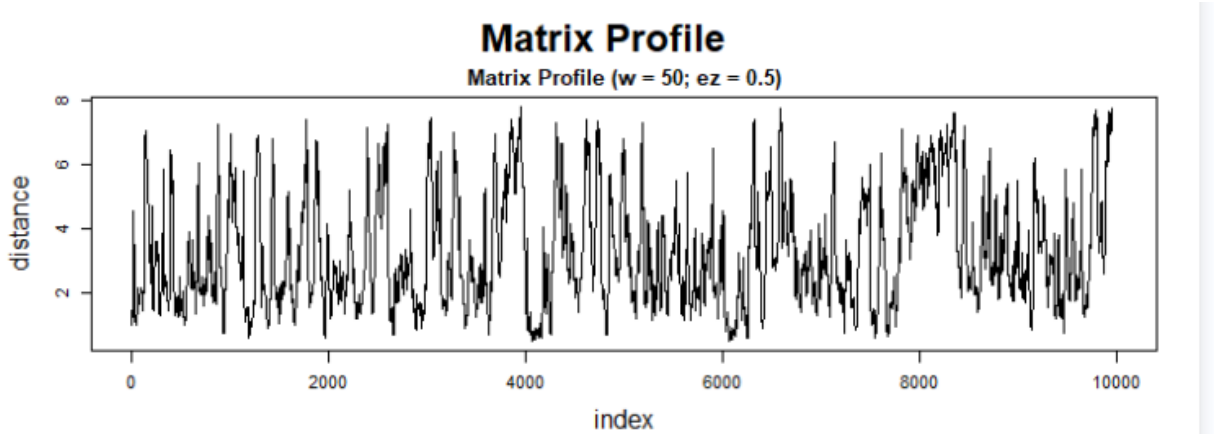
ΔΙΑΓΡΑΜΜΑ 5.3.1: Διάγραμμα ροής | EOG

ΠΙΝΑΚΑΣ 5.3.2: Παρουσίαση μοτίβων | EOG

Αριθμός Παρατηρήσεων	Μέγεθος παραθύρου	Ζώνη αποκλεισμού	Χρόνος περάτωσης	Αριθμός μοτίβων που ανιχνεύτηκαν
100	50	25	32''	3
Διανύσματα μοτίβων				
[4061, 6064]	[4091, 6091]	[6256, 7548]		

Με ορισμένο μέγεθος παραθύρου ίσο με 50, λόγω του γεγονότος ότι η συχνότητα μέτρησης είναι 50 hz (1 παρατήρηση ανά λεπτό), τα μοτίβα που επιστρέφει η R μέσω του STOMP είναι 3 στον αριθμό. Τα σχετικά διανύσματα που αφορούν τα ανωτέρω μοτίβα, παρουσιάζονται στον σχετικό πίνακα 3.3.1

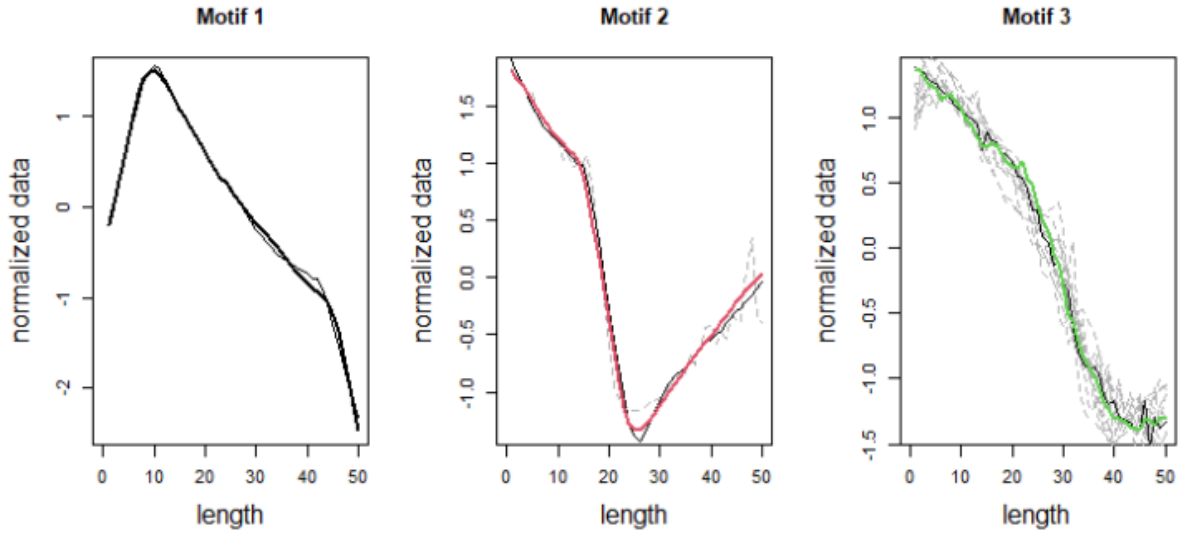
και μένει να επιβεβαιωθούν από το παρακάτω γράφημα που παρουσιάζει το κεντρικό διάγραμμα του MP για το σχετικό σετ δεδομένων.



ΔΙΑΓΡΑΜΜΑ 5.3.2 : Γραφική απεικόνιση του Matrix Profile | EOG

Πράγματι, στα διανύσματα που αναφέρονται τα επαναλαμβανόμενα μοτίβα, παρουσιάζεται τοπικό ελάχιστο στο διάγραμμα 5.3.2 ,δηλαδή ισχυρή ένδειξη μοτίβων για το υπό εξέταση δείγμα.

Σε αντιστοιχία, με τις προηγούμενες ενότητες, παρουσιάζεται παρακάτω το κοινό γράφημα που περιλαμβάνει τα κύρια μοτίβα στο σετ ‘EOG’.



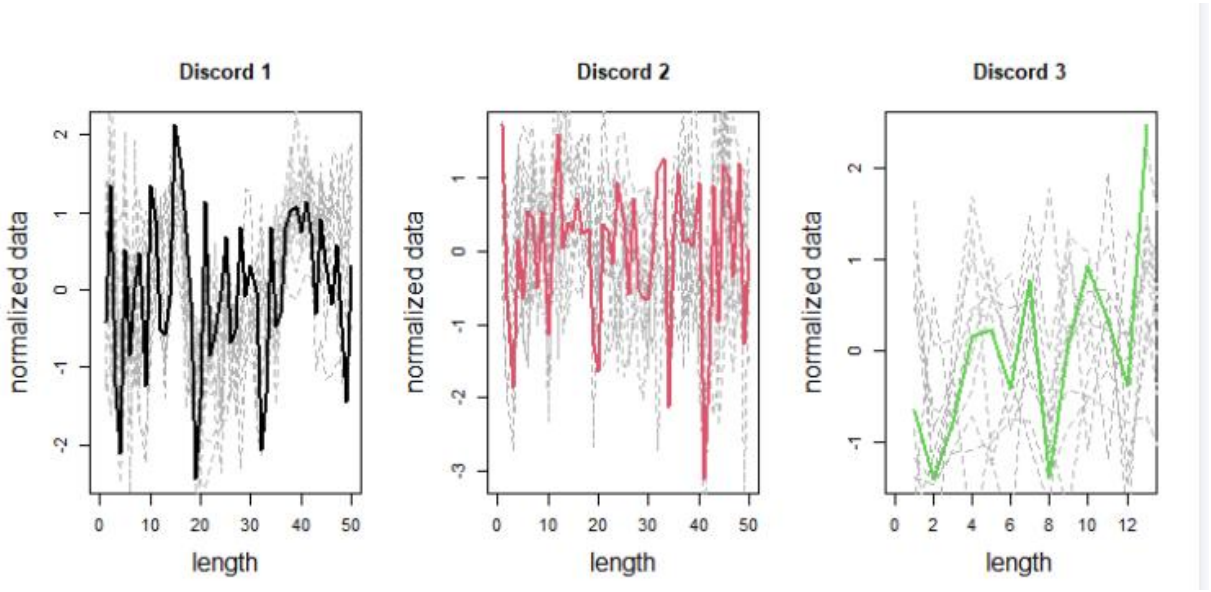
ΔΙΑΓΡΑΜΜΑ 5.3.3 : ΜΟΤΙΒΑ ΠΟΥ ΑΝΙΧΝΕΥΤΗΚΑΝ | EOG

Στον κάτωθι πίνακα, παρουσιάζονται οι ανωμαλίες που ανιχνεύτηκαν μέσω του αλγορίθμου STOMP και των σχετικών εντολών που τον διέπουν.

Πίνακας 5.3.3 : Παρουσίαση ανωμαλιών | EOG

Αριθμός Παρατηρήσεων	Μέγεθος παραθύρου	Ζώνη αποκλεισμού	Χρόνος περάτωσης	Αριθμός ανωμαλιών που ανιχνεύτηκαν
10000	50	25	32΄΄	3
Διανύσματα ανωμαλιών				
[3946]	[6591]	[9939]		

Τέλος, παρουσιάζονται στο κάτωθι γράφημα, οι κύριες ανωμαλίες του δείγματος (3 στον αριθμό), που επιβεβαιώνονται και από το κεντρικό γράφημα 5.3.2.



ΔΙΑΓΡΑΜΜΑ 5.3.4 : ΑΝΩΜΑΛΙΕΣ ΠΟΥ ΑΝΙΧΝΕΥΤΗΚΑΝ |EOG

Στο ανωτέρω γράφημα, παρουσιάζεται εύληπτα το φαινόμενο των έντονων ‘’outliers’’ στο σετ δεδομένων που αναλύεται, δηλαδή ακραία διασκορπισμένες τιμές των σχετικών παρατηρήσεων.

Κεφάλαιο 6 : Συμπεράσματα

Από την παρούσα έρευνα, μπορούν να εξαχθούν ορισμένα βασικά συμπεράσματα που αφορούν τη μεθοδολογία του Matrix Profile (MP) και την εφαρμογή του στην Ανάλυση Χρονοσειρών. Αρχικά, η συγκεκριμένη μεθοδολογία αναδεικνύεται ως μια καινοτόμα διαδικασία, η οποία προσφέρει ακριβή, γρήγορα και αποδοτικά αποτελέσματα στον τομέα της ανίχνευσης μοτίβων και ανωμαλιών. Η συμβολή της έγκειται στο γεγονός ότι επιτρέπει τη μελέτη πολύπλοκων δεδομένων με τρόπο που συνδυάζει αποτελεσματικότητα και υπολογιστική οικονομία.
Μία από τις σημαντικότερες παραμέτρους που καθορίζουν την ποιότητα και την αξιοπιστία των αποτελεσμάτων του MP είναι ο ορισμός του μεγέθους του παραθύρου. Ο σωστός προσδιορισμός του επηρεάζει άμεσα την ικανότητα της μεθοδολογίας να εντοπίζει μοτίβα και ανωμαλίες με ακρίβεια, ενώ τυχόν αστοχίες στην επιλογή του μπορούν να οδηγήσουν σε παραπλανητικά συμπεράσματα.

Επιπλέον, οι αλγόριθμοι που χρησιμοποιεί το MP εμφανίζουν διαφορετικού βαθμού καταλληλότητα, η οποία εξαρτάται τόσο από τη φύση και τον όγκο των δεδομένων όσο και από τους ερευνητικούς στόχους κάθε μελέτης. Κάτι τέτοιο σημαίνει ότι η επιλογή του κατάλληλου αλγορίθμου πρέπει να γίνεται με βάση τα ιδιαίτερα χαρακτηριστικά του εκάστοτε προβλήματος, προκειμένου να αξιοποιηθεί στο μέγιστο η αξία της μεθοδολογίας.
Αξίζει επίσης να αναφερθεί ότι οι ανταγωνιστικοί αλγόριθμοι σε σχέση με το MP παρουσιάζουν συγκριτικά πλεονεκτήματα σε ορισμένα πεδία έρευνας. Αν και ο MP υπερέχει σε ταχύτητα και ακρίβεια, δεν μπορεί να παραλειφθεί η περίπτωση όπου ένας εναλλακτικός αλγόριθμος αποδεικνύεται πιο αποτελεσματικός υπό συγκεκριμένες συνθήκες και ερευνητικές απαιτήσεις.
Στο πειραματικό μέρος της παρούσας εργασίας χρησιμοποιήθηκε ο αλγόριθμος STOMP, ο οποίος επιβεβαίωσε την αποδοτικότητα και την ακρίβειά του. Η εφαρμογή του έδειξε ότι ανταποκρίνεται με συνέπεια στα ζητούμενα, ενώ τα αποτελέσματά του συμφωνούν σε μεγάλο βαθμό με τα πορίσματα της σχετικής βιβλιογραφίας, γεγονός που ενισχύει την αξιοπιστία του.

Τέλος, υπογραμμίζεται η σημασία της συνδυαστικής χρήσης διαφορετικών αλγορίθμων στον τομέα της Ανάλυσης Χρονοσειρών. Η αξιοποίηση πολλαπλών προσεγγίσεων σε συνεργασία μπορεί να αποδειχθεί αποτελεσματικότερη από την αποκλειστική χρήση μιας μεμονωμένης δομής δεδομένων, προσφέροντας έτσι μια πιο σφαιρική και ολοκληρωμένη κατανόηση των μοτίβων και των ανωμαλιών που αναδύονται στα δεδομένα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] S. S. W. Fatima και A. Rahimi, 'A Review of Time-Series Forecasting Algorithms for Industrial Manufacturing Systems', *Machines*, τ. 12, τχ. 6, σ. 380, Ιουνίου 2024, doi: 10.3390/machines12060380.

[2] A. Casolaro, V. Capone, G. Iannuzzo, και F. Camastra, 'Deep Learning for Time Series Forecasting: Advances and Open Problems', *Information*, τ. 14, τχ. 11, σ. 598, Νοεμβρίου 2023, doi: 10.3390/info14110598.

[3] Y. K. Dwivedi *κ.ά.*, 'Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy', *International Journal of Information Management*, τ. 57, σ. 101994, Απριλίου 2021, doi: 10.1016/j.ijinfomgt.2019.08.002.

[4] Z. Liu, Z. Zhu, J. Gao, και C. Xu, 'Forecast Methods for Time Series Data: A Survey', *IEEE Access*, τ. 9, σσ. 91896–91912, 2021, doi: 10.1109/ACCESS.2021.3091162.

[5] Z. Z. Darban, G. I. Webb, S. Pan, C. C. Aggarwal, και M. Salehi, 'Deep Learning for Time Series Anomaly Detection: A Survey', 2022, doi: 10.48550/ARXIV.2211.05244.

[6] A. Iqbal, R. Amin, F. S. Alsubaei, και A. Alzahrani, 'Anomaly detection in multivariate time series data using deep ensemble models', *PLoS ONE*, τ. 19, τχ. 6, σ. e0303890, Ιουνίου 2024, doi: 10.1371/journal.pone.0303890.

[7] Q. Liu, P. Boniol, T. Palpanas, και J. Paparrizos, 'Time-Series Anomaly Detection: Overview and New Trends', *Proc. VLDB Endow.*, τ. 17, τχ. 12, σσ. 4229–4232, Αυγούστου 2024, doi: 10.14778/3685800.3685842.

[8] Z. Deng, J. Kang, και X. Wang, 'Multidimensional Time Series Analysis for Anomaly Pattern Detection and Interpretation', στο *2024 IEEE 4th International Conference on Power, Electronics and Computer Applications (ICPECA)*, Shenyang, China: IEEE, Ιανουαρίου 2024, σσ. 1371–1375. doi: 10.1109/ICPECA60615.2024.10471100.

[9] C.-C. M. Yeh *κ.ά.*, 'Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets', στο *2016 IEEE 16th International Conference on Data Mining (ICDM)*, Barcelona, Spain: IEEE, Δεκεμβρίου 2016, σσ. 1317–1322. doi: 10.1109/ICDM.2016.0179.

[10] F. Madrid, S. Imani, R. Mercer, Z. Zimmerman, N. Shakibay, και E. Keogh, 'Matrix Profile XX: Finding and Visualizing Time Series Motifs of All Lengths using the Matrix Profile', στο *2019 IEEE International Conference on Big Knowledge (ICBK)*, Beijing, China: IEEE, Νοεμβρίου 2019, σσ. 175–182. doi: 10.1109/ICBK.2019.00031.

[11] A. R.Hyndman, 'Forecasting : Principles and Practice,2nd edition', 2018.

[12] J. G.Box, *Time Series Analysis, Forecasting and control*. Wiley and Sons, 2016.

[13] James D, Hamilton, 'Time Series Analysis', 1994.

[14] R. Tsay, *Analysis of Finacial Time Series*, 3rd έκδ. Wiley and Sons, 2010.

[15] D. Brockwell, *Introduction to Time Series and forecasting*, Latest. Springer International Publishing : Imprint: Springer, 2016.

[16] Chatfield,Chris, *The Analysis of Time Series, an introduction*. Taylor and Francis Group, 2003.

[17] S. Shumway, *Time Series Analysis and Its Applications,With R Examples*. Springer International Publishing : Imprint: Springer, 2025.

[18] kumar Chandola, 'Anomaly detection: A survey'.

[19] 'VII. On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers', *Phil. Trans. R. Soc. Lond. A*, τ. 226, τχ. 636–646, σσ. 267–298, Ιανουαρίου 1927, doi: 10.1098/rsta.1927.0007.

[20] B. Goodfellow,I, *Deep Learning*. M.I.T Press, 2016.

[21] F. Hastie, *The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition*. Springer International Publishing : Imprint: Springer, 2009.

[22] K. Han,J, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2012.

[23] kumar Tan, *Introduction to Data Mining (Second Edition)*. 2023.
[24] A. K. Jain, 'Data clustering: 50 years beyond K-means', *Pattern Recognition Letters*, τ. 31, τχ. 8, σσ. 651–666, Ιουνίου 2010, doi: 10.1016/j.patrec.2009.09.011.
[25] R. Agrawal, T. Imieliński, και A. Swami, 'Mining association rules between sets of items in large databases', *SIGMOD Rec.*, τ. 22, τχ. 2, σσ. 207–216, Ιουνίου 1993, doi: 10.1145/170036.170072.
[26] F. T. Liu, K. M. Ting, και Z.-H. Zhou, 'Isolation-Based Anomaly Detection', *ACM Trans. Knowl. Discov. Data*, τ. 6, τχ. 1, σσ. 1–39, Μαρτίου 2012, doi: 10.1145/2133360.2133363.
[27] 'Github.com/baggepinnen/MatrixProfile', 2025.
[28] Z. Mueen, 'Time Series Data Mining Using the Matrix Profile: A Unifying View of Motif Discovery, Anomaly Detection, Segmentation, Classification, Clustering and Similarity Joins'. 2017.
[29] B. Gerazov *κ.ά.*, 'Matrix Profile based Anomaly Detection in Streaming Gait Data for Fall Prevention', στο *2023 30th International Conference on Systems, Signals and Image Processing (IWSSIP)*, Ohrid, North Macedonia: IEEE, Ιουνίου 2023, σσ. 1–5. doi: 10.1109/IWSSIP58668.2023.10180243.
[30] M. Czosnyka, B. Wnukowska, και K. Karbowa, 'Electrical energy consumption and the energy market in Poland during the COVID-19 pandemic', στο *2020 Progress in Applied Electrical Engineering (PAEE)*, Koscielisko, Poland: IEEE, Ιουνίου 2020, σσ. 1–5. doi: 10.1109/PAEE50669.2020.9158771.
[31] D. Cao και Z. Lin, 'Multidimensional time series motif group discovery based on matrix profile', *Knowledge-Based Systems*, τ. 304, σ. 112509, Νοεμβρίου 2024, doi: 10.1016/j.knosys.2024.112509.
[32] T. Mondal, R. Akbarinia, και F. Masseglia, 'kNN matrix profile for knowledge discovery from time series', *Data Min Knowl Disc*, Φεβρουαρίου 2023, doi: 10.1007/s10618-022-00883-8.
[33] R. Guidotti και M. D'Onofrio, 'Matrix Profile-Based Interpretable Time Series Classifier', *Front. Artif. Intell.*, τ. 4, σ. 699448, Οκτωβρίου 2021, doi: 10.3389/frai.2021.699448.
[34] E. Cartwright, M. Crane, και H. J. Ruskin, 'Financial Time Series: Market Analysis Techniques Based on Matrix Profiles †', στο *The 7th International Conference on Time Series and Forecasting*, MDPI, Ιουλίου 2021, σ. 45. doi: 10.3390/engproc2021005045.
[35] Chin-Chia Michael Yeh, 'Towards a Near Universal Time Series Data Mining Tool: Introducing the Matrix Profile', 2018.
[36] Sontowski, 'Detecting Cyber Attacks using the Matrix Profile', 2022.
[37] A. Mueen, 'The Fastest Similarity Search Algorithm for Time Series Subsequences under Euclidean Distance', 2015.
[38] E. Keogh, J. Lin, S.-H. Lee, και H. V. Herle, 'Finding the most unusual time series subsequence: algorithms and applications', *Knowl Inf Syst*, τ. 11, τχ. 1, σσ. 1–27, Δεκεμβρίου 2006, doi: 10.1007/s10115-006-0034-6.
[39] R. Foorthuis, 'On the Nature and Types of Anomalies: A Review of Deviations in Data', 2020, doi: 10.48550/ARXIV.2007.15634.
[40] A. Koran, 'Unveiling the Flaws: A Critical Analysis of Initialization Effect on Time Series Anomaly Detection', 2023.
[41] A. Hien, N. Beldiceanu, C.-G. Quimper, και M.-I. Restrepo, 'Automata Based Multivariate Time Series Analysis for Anomaly Detection over Sliding Time Windows', στο *ITISE 2023*, MDPI, Ιουλίου 2023, σ. 65. doi: 10.3390/engproc2023039065.
[42] T. Cooley, 'An algorithm for the machine calculation of complex Fourier series.', 1965.
[43] F. Guigou, P. Collet, και P. Parrend, 'Anomaly detection and motif discovery in symbolic representations of time series', 2017, doi: 10.13140/RG.2.2.20158.69447.
[44] Y. Zhu, 'Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins', 2016.
[45] F. Bischoff, 'tsmp: Time Series with Matrix Profile'. σ. 0.4.15, 17 Αύγουστος 2018. doi: 10.32614/CRAN.package.tsmp.
[46] R. Akbarinia και B. Cloez, 'Efficient Matrix Profile Computation Using Different Distance Functions', 2019, *arXiv*. doi: 10.48550/ARXIV.1901.05708.

[47] Zhu, ‘Matrix Profile XI: SCRIMP++: Time Series Motif Discovery at Interactive Speeds’, 2018.
[48] S. Jin και R. Zafarani, ‘Representing Networks with 3D Shapes’, στο *2018 IEEE International Conference on Data Mining (ICDM)*, Singapore: IEEE, Νοεμβρίου 2018, σσ. 177–186. doi: 10.1109/ICDM.2018.00033.
[49] E. Keogh, S. Chu, D. Hart, και M. Pazzani, ‘An online algorithm for segmenting time series’, στο *Proceedings 2001 IEEE International Conference on Data Mining*, San Jose, CA, USA: IEEE Comput. Soc, 2001, σσ. 289–296. doi: 10.1109/ICDM.2001.989531.
[50] E. Cartwright, M. Crane, και H. J. Ruskin, ‘Side-Length-Independent Motif (SLIM): Motif Discovery and Volatility Analysis in Time Series—SAX, MDL and the Matrix Profile’, *Forecasting*, τ. 4, τχ. 1, σσ. 219–237, Φεβρουαρίου 2022, doi: 10.3390/forecast4010013.
[51] S. Alaee, K. Kamgar, και E. Keogh, ‘Matrix Profile XXII: Exact Discovery of Time Series Motifs under DTW’, 2020, *arXiv*. doi: 10.48550/ARXIV.2009.07907.
[52] T. P. Q. Nguyen, T. N. Tran, H. T. N. H. Giang, και T. T. Nguyen, ‘Timeseries Anomaly Detection Using SAX and Matrix Profiles Based Longest Common Subsequence’, στο *Computational Science – ICCS 2023*, τ. 14074, J. Mikyška, C. De Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra, και P. M. A. Sloot, Επιμ., στο Lecture Notes in Computer Science, vol. 14074. , Cham: Springer Nature Switzerland, 2023, σσ. 221–229. doi: 10.1007/978-3-031-36021-3_21.
[53] D. Hawkins, *Identification of Outliers*. στο Monographs on Statistics and Applied Probability Ser. Dordrecht: Springer Netherlands, 1980.
[54] S. J. Taylor και B. Letham, ‘Forecasting at Scale’, *The American Statistician*, τ. 72, τχ. 1, σσ. 37–45, Ιανουαρίου 2018, doi: 10.1080/00031305.2017.1380080.
[55] Github, ‘motifletshttps://github.com/patrickzib/motiflets’. 2022.
[56] F. Mörchen και A. Ultsch, ‘Efficient mining of understandable patterns from multivariate interval time series’, *Data Min Knowl Disc*, τ. 15, τχ. 2, σσ. 181–215, Αυγούστου 2007, doi: 10.1007/s10618-007-0070-1.

ΠΑΡΑΡΤΗΜΑ Α : ΚΩΔΙΚΑΣ ΚΑΙ ΠΡΩΤΟΤΥΠΑ ΔΕΔΟΜΕΝΑ | MUSCLE ACTIVATION

Α.1 οι πρώτες 100 παρατηρήσεις των πρωτότυπων δεδομένων

No	T1
1	-2,25E+07
2	-2,29E+07
3	-2,25E+07
4	-2,17E+07
5	-2,11E+07
6	-2,05E+07
7	-2,06E+07
8	-2,09E+07
9	-2,03E+07
10	-1,97E+07
11	-1,93E+07
12	-1,91E+07
13	-1,89E+07
14	-1,87E+07
15	-1,89E+07
16	-1,90E+07
17	-1,89E+07
18	-1,87E+07
19	-1,85E+07
20	-1,85E+07
21	-1,87E+07
22	-1,88E+07
23	-1,87E+07
24	-1,88E+07
25	-1,89E+07
26	-1,97E+07
27	-2,06E+07
28	-2,11E+07
29	-2,18E+07
30	-2,21E+07
31	-2,24E+07
32	-2,24E+07
33	-2,29E+07
34	-2,44E+07
35	-2,49E+07
36	-2,31E+07
37	-2,19E+07
38	-2,17E+07
39	-2,27E+07
40	-2,35E+07
41	-2,32E+07
42	-2,24E+07
43	-2,18E+07
44	-2,14E+07
45	-2,08E+07

46	-2,05E+07
47	-2,05E+07
48	-2,02E+07
49	-1,99E+07
50	-1,97E+07
51	-1,99E+07
52	-1,98E+07
53	-1,99E+07
54	-1,99E+07
55	-1,96E+07
56	-1,95E+07
57	-2,00E+07
58	-2,06E+07
59	-2,16E+07
60	-2,23E+07
61	-2,22E+07
62	-2,22E+07
63	-2,26E+07
64	-2,26E+07
65	-2,20E+07
66	-2,16E+07
67	-2,13E+07
68	-2,12E+07
69	-2,10E+07
70	-2,09E+07
71	-2,09E+07
72	-2,06E+07
73	-2,03E+07
74	-1,97E+07
75	-1,93E+07
76	-1,90E+07
77	-1,91E+07
78	-1,97E+07
79	-2,01E+07
80	-2,01E+07
81	-2,01E+07
82	-1,99E+07
83	-2,01E+07
84	-2,06E+07
85	-2,08E+07
86	-2,06E+07
87	-2,03E+07
88	-2,02E+07
89	-2,00E+07
90	-1,98E+07
91	-1,94E+07
92	-1,88E+07
93	-1,82E+07
94	-1,78E+07
95	-1,74E+07
96	-1,70E+07
97	-1,65E+07
98	-1,60E+07
99	-1,57E+07
100	-1,55E+07

Πηγή : Github, ‘motifletshttps://github.com/patrickzib/motiflets’. 2022.

Α.2 Ο Κώδικας στο R -STUDIO για τον αλγόριθμο STOMP

```
Library(tsmp)
newdata <- readcsv("C:/Users/User/Downloads/muscle_activation.csv")
names(newdata)
data<-newdata$testdata
str(data)
head(data)
length(data)
str(data)
mean(data)
max(data)
min(data)
median(data)
sd(data)
range(data)
quantile(data)
window_size = 100
mp_result <- tsmp(ts = data, window_size = window_size)
print(mp_result)
plot(mp_result, main = "Matrix Profile")
motif_result <- motifs(mp_result)
print(motif_result)
plot(motif_result)
discord_result <- discords(mp_result)
print(discord_result)
plot(discord_result)
--------------------
```

ΠΑΡΑΡΤΗΜΑ Β : ΚΩΔΙΚΑΣ ΚΑΙ ΠΡΩΤΟΤΥΠΑ ΔΕΔΟΜΕΝΑ | TERMINATE DNA

Β.1 Τα πρώτα 100 στοιχεία του dataset

15	1895
16	1897
17	1895
18	1894
19	1896
20	1898
21	1897
22	1899
23	1901
24	1903
25	1905
26	1903
27	1905
28	1907
29	1909
30	1908
31	1907
32	1909

33	1911
34	1913
35	1911
36	1910
37	1912
38	1913
39	1915
40	1914
41	1916
42	1918
43	1920
44	1922
45	1921
46	1923
47	1925
48	1927
49	1928
50	1927
51	1929
52	1931
53	1933
54	1932
55	1934
56	1933
57	1932
58	1931
59	1933
60	1935
61	1937

62	1939
63	1941
64	1943
65	1942
66	1944
67	1946
68	1945
69	1947
70	1949
71	1951
72	1953
73	1951
74	1950
75	1952
76	1954
77	1953
78	1952
79	1951
80	1953
81	1955
82	1956
83	1958
84	1960
85	1962
86	1964
87	1966
88	1964
89	1966
90	1968

91	1969
92	1971
93	1973
94	1975
95	1977
96	1975
97	1977
98	1979
99	1981
100	1979

Πηγή : https://sites.google.com/view/pan-matrix-profile/datasets?authuser=0

B.2 Ο Κώδικας στο R -STUDIO για τον αλγόριθμο STOMP

```
library(R.matlab)
data <- readMat("C:/Users/User/Downloads/termite_DNA_circular_shift.mat")
names(data)
str(data$t2)
newbie<-data$t2
head(newbie)
View(data$t2)
t2 <- as.vector(data$t2)
length(t2)
str(t2)
mean(t2)
max(t2)
min(t2)
median(t2)
sd(t2)
range(t2)
quantile(t2)

window_size <- 100
plot(data$t2, type = "l", col = "blue",
main = "Time Series of Terminate Dna",
xlab = "Αριθμός Παρατηρήσεων" , ylab ="Γενετικο υλικό" )
library(tsmp)
mp_result <- tsmp(ts = data$t2, window_size = window_size)
print(mp_result)
plot(mp_result, main = "Matrix Profile")
motif_result <- motifs(mp_result)
print(motif_result)
plot(motif_result)
discord_result <- discords(mp_result)
print(discord_result)
plot(discord_result)
```

ΠΑΡΑΡΤΗΜΑ Γ : ΚΩΔΙΚΑΣ | EOG

```
library(R.matlab)
newdata <-
readMat("C:/Users/User/Downloads/eog_multiple_scale_example.mat")
names(newdata)
t3<-newdata$testdata
str(t3)
head(t3)
t3 <- as.vector(newdata$testdata)
length(t3)
str(t3)
mean(t3)
max(t3)
min(t3)
median(t3)
sd(t3)
range(t3)
quantile(t3)
mp_result <- tsmp(ts = t3, window_size = window_size)
print(mp_result)
plot(mp_result, main = "Matrix Profile")
motif_result <- motifs(mp_result)
print(motif_result)
plot(motif_result)
discord_result <- discords(mp_result)
print(discord_result)
plot(discord_result)
```