ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΑΝΑΚΑΛΥΨΗ ΕΠΑΝΑΛΗΠΤΙΚΩΝ ΠΡΟΤΥΠΩΝ ΣΕ ΔΕΔΟΜΕΝΑ ΚΙΝΗΣΗΣ ΨΑΡΑΔΙΚΩΝ

Ιωάννης Κοντογιαννίδης

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Σεπτέμβριος 2021

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Αναπλ. Καθηγητής Πελέκης Νικόλαος (Επιβλέπων)
- Καθηγητής Κούτρας Μάρκος
- Αναπλ. Καθηγητής Κοφίδης Ελευθέριος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

PERIODIC PATTERN DISCOVERY IN FISHERY VESSELS

By

John Kontogiannidis

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
September 2021

Στην οικογένεια μου

Ευχαριστίες

Περίληψη

Στην σύγχρονη εποχή όπου υπάρχει ένας τεράστιος όγκος πληροφοριών η αξιοποίηση αυτών αποτελεί ένα σημαντικό κομμάτι μελέτης. Ένας τρόπος για να αξιοποιήσουμε τις πληροφορίες αυτές και να καταλήξουμε σε χρήσιμα συμπεράσματα είναι με το να εντοπίσουμε κάποια μοτίβα μέσα από τα δεδομένα μας. Τι είναι μοτίβα? Με μία γρήγορη ερμηνεία του όρου αυτού θα μπορούσαμε να πούμε πως είναι ένα επαναλαμβανόμενο δομοστοιχείο. Ο εντοπισμός μοτίβων μπορεί να χρησιμοποιηθεί σε μία τεράστια γκάμα διαφορετικών τύπου δεδομένων, για παράδειγμα θα μπορούσαμε να εντοπίσουμε κάποια μοτίβα σε συμπεριφορές ζώων, σε συμπεριφορές καταναλωτών, στην μετεωρολογία και σε πολλά άλλα. Η συγκεκριμένη διπλωματική αναφέρετε στον εντοπισμό μοτίβων σε δεδομένα κίνησης ψαράδικων. Σκοπός της παρούσας εργασίας είναι μέσα από τα δεδομένα να μπορέσουν να αντληθούν κάποιες περιοδικές συμπεριφορές που παρουσιάζουν τα ψαράδικα και αργότερα να διαπιστωθεί εάν μπορούμε να βγάλουμε κάποια χρήσιμα συμπεράσματα από αυτά. Τα δεδομένα που χρησιμοποιήθηκαν είναι τύπου AIS και VMS τα οποία αναφέρονται σε δύο διαφορετικούς τύπους παρακολούθησης και καταγραφής των δεδομένων των σκαφών.

Στη συγκεκριμένη διπλωματική εργασία αναφέρονται δύο διαφορετικού τύποι δεδομένα που θα μπορούσαν να χρησιμοποιηθούν για τον εντοπισμό μοτίβων κινούμενων αντικειμένων. Ο πρώτος τύπος δεδομένων αναφέρεται σε σειριακές ακολουθίες, δηλαδή δισδιάστατα δεδομένα όπου έχουμε πληροφορία για το γεωγραφικό πλάτος και μήκος της εγγραφής. Ο δεύτερος τύπος δεδομένων που αναφέρθηκε είναι χωροχρονικά δεδομένα, δηλαδή τρισδιάστατα δεδομένα όπου στο γεωγραφικό πλάτος και μήκος προστίθεται και ο χρόνος της εγγραφής. Μαζί με τους τύπους των δεδομένων έγινε και η περιγραφή κάποιων αλγορίθμων που μπορούν να χρησιμοποιηθούν για τον εντοπισμό των μοτίβων ανάλογα με τον τύπο των δεδομένων.

Τέλος έγινε μία ανάλυση βασισμένη στον αλγόριθμο Periodica για τον εντοπισμό των περιοδικών συμπεριφορών στα δεδομένα κίνησης των ψαράδικων. Αρχικά καθαρίστηκαν τα δεδομένα μας, εντοπίστηκαν ποια από αυτά μπορούν να χρησιμοποιηθούν για ανάλυση μία φορά για τα AIS δεδομένα και μία για τα VMS. Στη συνέχεια έγινε μία μετατροπή των δεδομένων έτσι ώστε οι εγγραφές κάθε αλιευτικού σκάφους κατάλληλου για ανάλυση να αποτελείται από δεδομένα με σταθερό χρονικό βήμα. Τέλος βγήκαν κάποια από κοινού συμπεράσματα.

Για τον καθαρισμό των δεδομένων χρησιμοποιήθηκε η γλώσσα προγραμματισμού R και για την μετατροπή των δεδομένων η γλώσσα matlab. Επίσης ο αλγόριθμος Periodica ήταν σε γλώσσα matlab.

Abstract

In our modern era where massive amounts of information are available, many studies have been made to utilize these data. One way to make use of the available data and come to useful conclusions is by locating some patterns in them. To give a quick explanation on this term we could say that a pattern is a recurring module. Patterns can be found in a wide range of different types of data, for example we can find patterns in animal's behavior, in consumer's behavior, in meteorology and in many other fields. This thesis is about finding patterns in fishing vessels. The purpose of this study is at first to find periodic patterns in the fishing vessels and then ascertain whether we could or not make some use from those patterns. The data which was used for this study are AIS and VMS type, which refer to two different types of recording the vessel's location.

In this thesis we can find two different types of data. The first type of data are serial sequences, namely it is two-dimensional data where the information of each recording we have is about its latitude and longitude. The second type of data is space-time sequences, namely it is three-dimensional data where except the latitude and longitude we have as extra info the time of each recording. Along with serial and the space-time sequences it is described some algorithms which can be used to find periodic patterns, according to the type of the data.

In the final chapter an analysis has been made based on the Periodica algorithm to find some periodic patterns in the fishing vessels. The first step was the data cleaning, where only the data which could be used for analysis has been kept. This step was made one time for the AIS data and one time for VMS. Then the data were converted, so each fishing vessel would be consisted of data with a stable time frame. In the end some conclusions collectively for AIS and VMS data has been made.

For the data cleaning was used the R programming language. For the data conversion and Periodica algorithm was used the matlab programming language.

Περιεχόμενα

Κ Ε Φ Α Λ Α Ι Ο 1

1.1 Τι είναι τα μοτίβα

[1] Ο όρος πρότυπο ή αλλιώς όπως είναι ευρύτερα γνωστός με τον ξενικό του όρο μοτίβο βρίσκει αναφορές στις τέχνες όπως στη μουσική και στη ζωγραφική, στις επιστήμες όπως στα μαθηματικά, στη μετεωρολογία, στην αστρολογία ακόμη και στη καθημερινή μας ζωή. Άμα θέλουμε να δώσουμε μία πολλή απλή ερμηνεία του όρου αυτού θα λέγαμε πως αναφέρεται σε τμήματα που επαναλαμβάνονται συχνά μέχρι να συνθέσουν μία οντότητα. Για παράδειγμα πολύ εύκολα ακούγοντας μουσική μπορούμε να διαπιστώσουμε ότι σε ένα τραγούδι υπάρχουν ακολουθίες από νότες που επαναλαμβάνονται μέσα σε αυτό. Επίσης πολύ εύκολα μπορεί να γίνει αντιληπτό ένα από τα μοτίβα που ακολουθούνται στη μετεωρολογία κάθε χρόνο που είναι οι 4 εποχές. Η μελέτη και η ανακάλυψη των μοτίβων είναι πολύ χρήσιμη, διότι με αυτό το τρόπο μπορούμε να βγάλουμε πάρα πολλά και χρήσιμα συμπεράσματα για τον τρόπο συμπεριφοράς του αντικειμένου που μελετάμε. Ένα ακόμη όφελος που θα έχουμε εάν ανακαλύψουμε ένα μοτίβο, είναι πως μας δίνει την ευχέρεια να κάνουμε κάποιες υποθέσεις για το αντικείμενο μελέτης μας. Πράγμα πολύ σημαντικό όσων αφορά τις επιστήμες διότι ένα μεγάλο μέρος των επιστημών βασίζεται στην σύνταξη της υπόθεσης.

1.2 Ανακάλυψη μοτίβων σε δεδομένα κίνησης ψαράδικων

Στη συγκεκριμένη διπλωματική έγινε μία προσπάθεια να ανακαλυφθούν τα μοτίβα στις κινήσεις από αλιευτικά σκάφη. Δηλαδή αναλύοντας τα δεδομένα από διάφορα αλιευτικά σκάφη αρχικά είδαμε τις διαδρομές που ακολούθησαν τα σκάφη αυτά σε μία περίοδο 4 μηνών και αργότερα προσπαθήσαμε να ανακαλύψουμε μέσα σε αυτές τις διαδρομές εάν υπάρχουν κάποιες περιοδικές συμπεριφορές των σκαφών που επαναλαμβανόντουσαν ανά κάποια χρονικά διαστήματα, πόσες είναι αυτές και τη διάρκεια των χρονικών διαστημάτων για κάθε μία από αυτές. Αφού ανακαλυφθούν οι περιοδικές συμπεριφορές για κάθε αλιευτικό σκάφος ξεχωριστά θα αναλυθούν από κοινού όλα μαζί με κάποια περιγραφικά στατιστικά.

1.3 Τύπος δεδομένων που χρησιμοποιήθηκε

Τα δεδομένα που χρησιμοποιήθηκαν για την ανάλυση είναι δύο διαφορετικά datasets τύπου VMS και AIS, διάρκειας 4 μηνών από 1 Ιουλίου έως και 30 Σεπτεμβρίου του 2018.

[2] Σύστημα παρακολούθησης σκαφών (VMS)

Πρόκειται για ένα δορυφορικό σύστημα παρακολούθησης των αλιευτικών σκαφών, το οποίο, σε τακτικά διαστήματα, παρέχει στις αλιευτικές αρχές στοιχεία για τη θέση, την πορεία και την ταχύτητα των σκαφών. Το σύστημα είναι υποχρεωτικό για σκάφη της ΕΕ άνω των 15 μέτρων (από την 1η Ιανουαρίου 2012, για σκάφη άνω των 12 μέτρων). Τα σκάφη τρίτων χωρών του ιδίου μεγέθους οφείλουν, όταν βρίσκονται σε κοινοτικά ύδατα, να είναι εξοπλισμένα με λειτουργούσα συσκευή δορυφορικού εντοπισμού.

[3] Σύστημα αυτόματης αναγνώρισης (AIS)

Πρόκειται για αυτόνομο και συνεχές σύστημα εντοπισμού και παρακολούθησης των σκαφών το οποίο χρησιμοποιείται για λόγους ασφάλειας στη θάλασσα, καθώς δίνει στα σκάφη τη δυνατότητα να επικοινωνούν ηλεκτρονικά με παρακείμενα πλοία αλλά και με τις αρχές στην ξηρά και να ανταλλάσουν πληροφορίες για την ταυτότητα, τη θέση, την πορεία και την ταχύτητα του σκάφους.

Τα αλιευτικά σκάφη της ΕΕ πρέπει σταδιακά να εξοπλιστούν με πομπούς AIS, σύμφωνα με το ακόλουθο χρονοδιάγραμμα:

- από τις 31 Μαΐου 2012: όλα τα πλοία άνω των 24 μέτρων
- από τις 31 Μαΐου 2013: όλα τα πλοία άνω των 18 μέτρων
- στις 31 Μαΐου 2014: όλα τα πλοία άνω των 15 μέτρων

Τα κράτη μέλη μπορούν να χρησιμοποιούν στοιχεία του συστήματος AIS για λόγους ελέγχου και παρακολούθησης.

1.4 Εισαγωγή του προβλήματος

Στη συγκεκριμένη διπλωματική έγινε μία προσπάθεια ανάκτησης περιοδικών μοτίβων σε δεδομένα κίνησης ψαράδικων. Το πρώτο πρόβλημα που καλούμαστε να αντιμετωπίσουμε είναι η συλλογή των δεδομένων μιας και ο τύπος των συγκεκριμένων δεδομένων είναι κάπως ιδιαίτερος και δεν είναι εύκολος ο εντοπισμός τους στο διαδίκτυο. Τα δεδομένα που χρησιμοποιήθηκαν για τις ανάγκες τις διπλωματικής αποκτήθηκαν από τον ΕΛΚΕΘΕ. Στη συνέχεια ένα ακόμα πρόβλημα που καλούμαστε να αντιμετωπίσουμε είναι η διαχείριση των δεδομένων μιας και μιλάμε για δεδομένα πολύ μεγάλου όγκου (Πληροφορίες από πλοία για κάθε ημέρα). Τα δεδομένα επίσης θα πρέπει να μελετηθούν για τυχόν χρήσιμες πληροφορίες που μπορεί να λείπουν από κάποιες εγγραφές π.χ. να λείπει ο χρόνος καταγραφής σε μία εγγραφή, κάποιες λάθος εγγραφές π.χ. ο υπάλληλος που καταγράφει σε κάποια βάση τις εγγραφές να κάνει κάποιο λάθος και σε μία εγγραφή να έχει καταγραφεί γεωγραφικό μήκος και πλάτος που αντιστοιχεί σε στεριά ενώ μιλάμε για πλοία κ.α. Αφού τελειώσουν αυτές οι διαδικασίες ένα ακόμη πρόβλημα που θα κληθούμε να αντιμετωπίσουμε είναι η επιλογή του κατάλληλου μοντέλου για την σωστή εξόρυξη περιοδικών μοτίβων στα δεδομένα μας ή και

τυχόν τροποποιήσεις των μοντέλων που ίσως θα χρησιμοποιηθούν για να προσαρμοστεί το μοντέλο στα δεδομένα μας. Εφόσον τα δεδομένα που χρησιμοποιούνται είναι ιδιαίτερα, εκτός από γεωγραφικό μήκος και πλάτος έχουμε και τη μεταβλητή χρόνο οπότε μιλάμε για δεδομένα τριών διαστάσεων και επίσης οι εγγραφές δεν είναι ανά κάποια σταθερά χρονικά περιθώρια που σημαίνει πως δεν πρόκειται για κλασσικές χρονοσειρές.

1.5 Σύντομη εισαγωγή επόμενων Κεφαλαίων

Η δομή της διπλωματικής εργασίας αποτελείτε από 3 Κεφάλαιο. Στο 1ο Κεφάλαιο είδαμε κάποιες γενικές πληροφορίες για το τι είναι μοτίβο, τον τύπο των δεδομένων που χρησιμοποιήθηκε και μία εισαγωγή του προβλήματος. Στο 2ο Κεφάλαιο αρχικά γίνεται εισαγωγή σε δύο διαφορετικούς τύπους δεδομένων που μπορούν να χρησιμοποιηθούν για την εξόρυξη περιοδικών μοτίβων. Αρχικά αναφέρονται δεδομένα τα οποία ακολουθούν μία σειριακή διάταξη, δίνονται κάποια παραδείγματα για την καλύτερη κατανόηση των δεδομένων και πως μπορεί η αξιοποίηση αυτών να ωφελήσει τον εκάστοτε ενδιαφερόμενο. Στη συνέχεια της εισαγωγής αναφέρονται και τα χωροχρονικά δεδομένα δίνονται και σε αυτά κάποια παραδείγματα για την ευκολότερη κατανόηση τους και πως μπορούν να αξιοποιηθούν. Αναφέρονται κάποια προβλήματα που συναντώνται στα κινούμενα δεδομένα. Έπειτα στη συνέχεια του Κεφαλαίου περιγράφονται κάποιοι αλγόριθμοι η οποίοι έχουν χρησιμοποιηθεί για τον εντοπισμό μοτίβων μαζί με κάποιες χρήσιμες έννοιες για την βαθύτερη κατανόηση αυτών. Αρχικά περιγράφεται ο αλγόριθμος MPFPS ο οποίος εφαρμόζεται σε σειριακές ακολουθίες. Αργότερα αναλύονται οι αλγόριθμοι STPMine1, STPMine2 και STPMine2 – V2, ο αλγόριθμος Periodica και τέλος ο αλγόριθμος PRED οι οποίοι εφαρμόζονται σε χωροχρονικά δεδομένα. Στο 3ο Κεφάλαιο γίνεται η ανάλυση των δεδομένων έτσι ώστε να εντοπιστούν οι περίοδοι στα ψαράδικα. Αφού καθαριστούν τα δεδομένα και γίνει μία μετατροπή αυτών έτσι ώστε όλα τα ψαράδικα να αποτελούνται από εγγραφές με σταθερά χρονικά βήμα, εφαρμόστηκε ο αλγόριθμος Periodica για την εξόρυξη των περιόδων.

ΚΕΦΑΛΑΙΟ 2

2.1 Εισαγωγή

Η εξόρυξη συχνών μοτίβων αναφέρεται στον εντοπισμό των μοτίβων τα οποία εμφανίζονται συχνά σε μία βάση. Καταλαμβάνουν έναν πολύ σημαντικό ρόλο στο data mining και έχουν πάρα πολλές εφαρμογές. Ένας τύπος δεδομένων στον οποίο μπορούν να εφαρμοστούν οι τεχνικές αυτές είναι δεδομένα τα οποία ακολουθούν μία σειριακή διάταξη, όπως για παράδειγμα τα δεδομένα που αφορούν την συμπεριφορά ενός καταναλωτή όπου συνήθως ακολουθούν μία τέτοια διάταξη. Η ανακάλυψη μοτίβων σε αυτού του είδους τα δεδομένα μπορεί να μας δώσει πληροφορίες που μπορούν να μας φανούν πολύ χρήσιμες, για παράδειγμα στη συμπεριφορά του καταναλωτή μπορεί να μας δώσει πληροφορίες για το πως να σχεδιαστούν πιο αποτελεσματικές καμπάνιες ή τι στρατηγικές πωλήσεων θα ήταν προτιμότερο να ακολουθήσει μία διαφημιστική εταιρία.

Η ανακάλυψη μοτίβων σε δεδομένα τα οποία έχουν σειριακή διάταξη, έχει γενικευτεί ως το πρόβλημα της εξόρυξης σειριακών μοτίβων. Απαρτίζεται από την ανακάλυψη υποακολουθιών σε ένα σύνολο από ακολουθίες. Αν και η εξόρυξη σειριακών μοτίβων έχει αμέτρητες εφαρμογές και αρκετοί αλγόριθμοι έχουν σχεδιαστεί, ένας σημαντικός περιορισμός είναι πως δεν είναι κατάλληλοι στο να ανακαλύπτουν περιοδικά μοτίβα. Όμως τα περιοδικά μοτίβα εμφανίζονται σε πάρα πολλές περιπτώσεις. Για παράδειγμα, περιοδικά μοτίβα εμφανίζονται σε καταναλωτές οι οποίοι ψωνίζουν προϊόντα κάθε μέρα, βδομάδα η μήνα. Για την ανακάλυψη περιοδικών μοτίβων έχουν προταθεί πολύ αλγόριθμοι, όμως οι περισσότεροι από αυτούς έχουν σχεδιαστεί στο να βρίσκουν μοτίβα σε μία μόνο ακολουθία. Ίδια περιοδικά μοτίβα όμως συχνά εμφανίζονται σε πολλούς καταναλωτές και η ανακάλυψη τους για ένα σύνολο καταναλωτών θα μπορούσε να βοηθήσει με ποικίλους τρόπους τόσο τους καταναλωτές για να ενημερώνονται μόνο για τα προϊόντα που τους ενδιαφέρουν όσο και τις εταιρίες για να κάνουν πιο στοχευμένες διαφημίσεις. Ένας γνωστός αλγόριθμος που θα μπορούσε να λύσει αυτό το πρόβλημα είναι ο PHUSPM, όπου έχει τη δυνατότητα να μας εξορύξει περιοδικά μοτίβα σε μία βάση από ακολουθίες. Όμως αυτός ο αλγόριθμος έχει ένα σημαντικό μειονέκτημα, χρησιμοποιεί τις ίδιες μετρήσεις περιοδικότητας όπως οι αλγόριθμοι που χρησιμοποιούνται για μία ακολουθία. Το οποίο έχει σαν αποτέλεσμα, να μπορεί να ανακαλύψει μοτίβα που εμφανίζονται συχνά στο σύνολο των ακολουθιών, όμως δεν μπορεί να μας δώσει κάποια πληροφορία για το εάν αυτά τα μοτίβα είναι περιοδικά σε κάθε ακολουθία ξεχωριστά. Για παράδειγμα ο συγκεκριμένος αλγόριθμος εάν εφαρμοζόταν σε ένα super market, θα μπορούσε να μας δώσει τη πληροφορία πως στο κατάστημα πωλούνται

περιοδικά ψωμί και γάλα (εμφανίζεται συχνά στη βάση). Όμως δεν θα μας έδινε την πληροφορία πως πολλοί καταναλωτές αγοράζουν περιοδικά ψωμί και γάλα. (το ψωμί και το γάλα εμφανίζει περιοδικότητα σε πολλές ακολουθίες όπου η κάθε ακολουθία αναφέρεται σε έναν καταναλωτή). Για την αντιμετώπιση του συγκεκριμένου προβλήματος οι Philippe Fournier-Viger, Zhitian Li, Jerry Chun-Wei Lin, Rage Uday Kiran, και Hamido Fujita πρότειναν έναν αλγόριθμο με όνομα MPFPS [5] όπου θα δούμε παρακάτω περισσότερες λεπτομέρειες.

Εκτός από τα δεδομένα που ακολουθούν μία σειριακή διάταξη, πολύ μεγάλη άνθηση βρίσκουμε και στην ανακάλυψη περιοδικών μοτίβων στα χωροχρονικά δεδομένα. Η ραγδαία αύξηση της τεχνολογίας και των τηλεπικοινωνιών (π.χ. GPS, κινητά τηλέφωνα κτλ.) είναι ξεκάθαρο πως διευκόλυνε τις καθημερινότητες πολλών ανθρώπων με ποικίλους τρόπους. Εκτός από τις διευκολύνσεις που προσέφεραν στις καθημερινότητες μας, υπάρχει πια και η δυνατότητα να συλλέγετε τεράστιος όγκος χωροχρονικών δεδομένων. Για παράδειγμα μπορούν να συλλέγονται σε μία βάση οι κινήσεις διάφορων ζώων μέσω των δορυφόρων, οι κινήσεις ανθρώπων μέσω των κινητών τηλεφώνων τους, των αυτοκινήτων μέσω ενός GPS κ.α. .

Εφόσον τέτοιου είδους δεδομένα είναι πια ευρέος διαθέσιμα, η εξόρυξη και η κατανόηση τέτοιων δεδομένων αποκτά ολοένα και περισσότερη προσοχή. Η πιο κοινή δραστηριότητα που παρατηρείται σε αυτά τα δεδομένα είναι περιοδική συμπεριφορά. Μία περιοδική συμπεριφορά μπορεί να χαρακτηριστεί σαν επαναλαμβανόμενες δραστηριότητες σε συγκεκριμένες περιοχές ανά τακτικά χρονικά διαστήματα. π.χ. (λεωφορεία, πλοία, αεροπλάνα τρένα κ.λπ.). Επίσης περιοδικά μοτίβα συναντάμε και σε κινήσεις διάφορων ζώων π.χ. (ο χρυσός αετός αρχίζει να μεταναστεύει για τη νότια Αμερική τέλη του Οκτωβρίου και γυρίζει πίσω στην Αλάσκα γύρο στα μέσα Μαρτίου) όπως και σε ανθρώπους όπου ακολουθούν κάποια συγκεκριμένα μοτίβα π.χ. (τις εργάσιμες μέρες ξυπνάνε περίπου την ίδια ώρα, τρώνε, βουρτσίζουν τα δόντια τους και αλλάζουν ρούχα με την ίδια σειρά και ακολουθούν περίπου την ίδια διαδρομή για την δουλειά τους). Μία τέτοια συμπεριφορά ονομάζεται περιοδική συμπεριφορά.

Το να γνωρίζουμε τέτοιες συμπεριφορές έχει πολλά πλεονεκτήματα, όπως για παράδειγμα αυτές οι περιοδικές συμπεριφορές που συναντάμε μας προσφέρουν εξηγήσεις για διάφορες συμπεριφορές που μπορεί να συναντάμε. Επίσης μπορούν να χρησιμοποιηθούν αντί πραγματικών δεδομένων για να κερδίσουμε χώρο ή ακόμη και για να κάνουμε κάποιες μελλοντικές προβλέψεις. Τέλος εάν ένα αντικείμενο παρατηρήσουμε ότι δεν ακολουθεί κάποια συμπεριφορά που θα έπρεπε σύμφωνα με το μοτίβο που το είχαμε συνηθίσει, είναι ένα σημάδι ότι ίσως κάτι έχει αλλάξει και μπορεί να χρειαστεί μία περαιτέρω διερεύνηση.

Παρόλα τα πλεονεκτήματα δεν πρέπει να ξεχνάμε ότι η διαδικασία της ανακάλυψης περιοδικών συμπεριφορών από τα δεδομένα που έχουμε από το ιστορικό των κινήσεων των αντικειμένων είναι πολύ απαιτητική και συνήθως δεν είναι εμφανή. Η εξόρυξη περιοδικών συμπεριφορών από κινούμενα αντικείμενα τα οποία συνήθως έχουν ένα πολύ μεγάλο εύρος δεδομένων δεν είναι μία απλή διαδικασία. Η διαχείριση και η ανάλυση κινούμενων

δεδομένων είναι απαιτητική, διότι στην αρχή της συλλογής των δεδομένων θα έχουμε να αντιμετωπίσουμε τελείως αχανή δεδομένα τα οποία θα χρειαστούν πολλές τεχνικές (πχ. Καθαρισμός δεδομένων από θόρυβο ή outliers) για να καταλήξουμε σε ένα τελικό data set όπου θα μπορέσουμε να πάρουμε σημαντικές πληροφορίες από αυτό ή να το χρησιμοποιήσουμε για να χτίσουμε ένα μοντέλο.

Εντούτοις, ο καθαρισμός των δεδομένων δεν είναι και η μοναδική δυσκολία που συναντάμε. Υπάρχουν πολλές περίοδοι και περιοδικές συμπεριφορές που παρεμβάλλουν η μία με την άλλη. Η εξόρυξη περιοδικών φαινομένων μπορεί να γεφυρώσει τις διαφορές μεταξύ των προ επεξεργασμένων δεδομένων και της σημασιολογικής κατανόησης των δεδομένων, όμως υπάρχουν κάποια προβλήματα που συναντάμε συχνά.

i) Ένα σημαντικό πρόβλημα που συναντάμε είναι πως οι περίοδοι μας είναι τις περισσότερες φορές άγνωστοι.

ii) Εξίσου σημαντικό πρόβλημα που συναντάμε είναι πως ακόμη και όταν οι περίοδοι μας είναι γνωστοί, οι περιοδικές συμπεριφορές πρέπει να εξορυχθούν από τα δεδομένα μας, διότι υπάρχει η περίπτωση να έχουμε αρκετές περιοδικές συμπεριφορές σε ίδιες περιόδους.

iii) Επίσης, στην πράξη πολλές φορές τα μοτίβα αλλάζουν, για παράδειγμα αν ένας εργαζόμενος μετατεθεί, το μοτίβο της διαδρομής του προς την δουλειά του θα αλλάξει. Για αυτό το λόγο η εξόρυξη συχνών μοτίβων δεν είναι αρκετή, θα χρειαστεί να εκτιμήσουμε και το διάστημα που θα είναι σε ισχύ.

Σε πραγματικά δεδομένα, τα μοτίβα εμφανίζονται σε συγκεκριμένα περιοδικά διαστήματα και ενδέχεται να διαστρεβλωθούν ή να μετατοπιστούν στο χρόνο. Για παράδειγμα εάν ένας εργαζόμενος αργήσει να ξυπνήσει η διαδρομή για τη δουλειά του θα μετατοπιστεί π.χ.(δέκα λεπτά) ή άμα μία μέρα έχει πολύ κίνηση ενώ ακολουθεί το ίδιο μοτίβο θα καταλήξει να αργήσει στη δουλειά του επομένως το μοτίβο θα έχει διαστρεβλωθεί.

Μία άλλη κατηγορία χωροχρονικών δεδομένων που αποκτά ολοένα και περισσότερο ενδιαφέρον για το τρόπο που πρέπει να διαχειριστούν, είναι τα δεδομένα από χρήστες μέσων κοινωνικής δικτύωσης. Στις μέρες μας είναι πλέον πολύ συνηθισμένη η χρήση της ένδειξης τοποθεσίας από τους χρήστες των μέσων κοινωνικής δικτύωσης, όπως για παράδειγμα τα tweets, check-ins και οι φωτογραφίες στο Instagram. Εφ’ όσον τέτοιου είδους πληροφορίες είναι εύκολα διαθέσιμες, η ανακάλυψη μοτίβων του τρόπου κίνησης των χρηστών έχει γίνει πιο εύκολη υπόθεση. Τα περιοδικά μοτίβα κινητικότητας, μπορούμε να πούμε ότι είναι οι επαναλαμβανόμενες δραστηριότητες σε συγκεκριμένες περιοχές σε συγκεκριμένα χρονικά διαστήματα. Για παράδειγμα ένας άνθρωπος που επιλέγει να τρώει το πρωινό του κάθε πρωί σε μία συγκεκριμένη περιοχή με αρκετά μαγαζιά. Ενώ ο ίδιος άνθρωπος θα μπορούσε επίσης κάθε Παρασκευή να επιλέγει να τρώει σε κάποιο εστιατόριο το οποίο θα βρίσκεται σε ένα εύρος κοντά σε ένα συγκεκριμένο super market. Σε αυτή τη περίπτωση παρατηρούμε τις δύο διαφορετικές περιοχές όπου ο συγκεκριμένος άνθρωπος παρατηρείται να παρουσιάζει

περιοδικές συμπεριφορές, με περίοδο τη μία μέρα και τη μία βδομάδα αντίστοιχα. Η ανακάλυψη περιοδικών συμπεριφορών μας βοηθάει στην ευρύτερη κατανόηση του τρόπου που κινούνται οι χρήστες, όπως επίσης στην ενσωμάτωση διάφορων εφαρμογών που για παράδειγμα θα μπορούν να εντοπίζουν τυχόν ανωμαλίες των κινήσεων, προβλέψεις για το που θα βρίσκονται οι χρήστες ή ακόμη και διάφορες προτάσεις για τους χρήστες. Μπορούμε να ξαναφέρουμε το προηγούμενο παράδειγμα και να συμπληρώσουμε πως για παράδειγμα θα μπορούσαμε να εντοπίσουμε μία ασυνήθιστη συμπεριφορά του ατόμου αν αντί να βρίσκεται το πρωί στην περιοχή που καταναλώνει το πρωινό του να εντοπιζόταν το βράδυ. Επίσης εάν επισκεπτόταν το super market τη Παρασκευή όπως έχει συνηθίσει, θα μπορούσε να του προταθεί κάποιο διπλανό εστιατόριο. Η διαδικασία όμως για να ανακαλύψουμε αυτά τα περιοδικά μοτίβα από τα μέσα κοινωνική δικτύωσης, όπου περιέχονται γεωγραφικές συντεταγμένες και χρονικές πληροφορίες είναι πολύ απαιτητική για 3 βασικού λόγους.

- ✓ Έχουμε πληροφορίες μόνο από το GPS, όμως ούτε οι περιοχές που εμφανίζονται τα μοτίβα ούτε οι περίοδοι μας είναι γνωστές.
- ✓ Διαφορετικοί χρήστες έχουν και διαφορετικό αριθμό και διαφορετικές περιοχές που εμφανίζουν κάποια μοτίβα. Για παράδειγμα ένας φοιτητής μπορεί να έχει μία περιοχή τη σχολή του και ένας επιχειρηματίας να έχει τρείς το σπίτι του, το γραφείο και τη περιοχή που κάνει τα ψώνια του. Επίσης πολύ συχνά παρατηρείται να μην ακολουθούν τα μοτίβα τους και να βρίσκονται σε διαφορετικές περιοχές και ο εντοπισμός μη σχετικών περιοχών και ενός σημαντικού αριθμού περιοχών για να μοντελοποιηθούν τα δεδομένα μας είναι πολύ σημαντική διαδικασία.
- ✓ Λόγο της φύσης της λειτουργείας των μέσων κοινωνικής δικτύωσης το δείγμα μας πολλές φορές είναι μικρό και έχουμε πληροφορίες από πολύ μικρά χρονικά διαστήματα.(οι χρήστες δεν ακολουθούν αυστηρά περιόδους και δεν ποστάρουν όλες τους τις δραστηριότητες)

2.2 Προβλήματα στην κατανόηση κινούμενων δεδομένων και σημείο αναφοράς.

[4] Η αναζήτηση των μοτίβων στα χωροχρονικά δεδομένα είναι μία δύσκολη διαδικασία λόγου του ότι συχνά σε τέτοιου είδους δεδομένα παρατηρείται μεγάλη ετερογένεια και μεγάλη διακύμανση στα δεδομένα που σημαίνει ότι αν δεν καταφέρουμε να τα περιορίσουμε τα αποτελέσματα μας δεν θα είναι και τόσο αξιόπιστα. Επίσης τέτοιου είδους δεδομένα χρειάζονται μία προ-επεξεργασία όπως η κατάτμηση (το φαινόμενο κατά το οποίο τα πετρώματα χωρίζονται σε κομμάτια κανονικά ή ακανόνιστα εξαιτίας φυσικών επιδράσεων) , η χωρική κανονικοποίηση και η αναπαράσταση των δεδομένων. Κάποια ακόμη προβλήματα για την εύρεση των μοτίβων στα χωροχρονικά δεδομένα είναι η εξερεύνηση της υψηλής συσχέτισης μεταξύ των γειτονικών αντικειμένων, η υψηλή διαστατικότητα και η πολυπλοκότητα των συσχετίσεων. Όπου θα χρειαστούμε μία αποδοτική διαχείριση της τοπολογικής πληροφορίας / απόστασης και την αναπαράσταση της χωρικής γνώσης ώστε να διευκολυνθεί η αναζήτηση.

Όσον αφορά το σημείο αναφοράς, αρχικά μία ερμηνεία που θα μπορούσαμε να δώσουμε είναι πως αναφέρετε σε μία πυκνή περιοχή δηλαδή μία περιοχή όπου παρατηρείτε ότι το αντικείμενο που μελετάμε την επισκέπτεται πολύ συχνά. Για παράδειγμα σε έναν μέσο εργαζόμενο το 24ωρο παρατηρούμε δύο πολύ πυκνές περιοχές που επισκέπτεται το γραφείο του και το διαμέρισμα του. Θέλοντας να εντοπίσουμε περιοδικές συμπεριφορές σε κινούμενα δεδομένα, συναντάμε δύο προβλήματα. Το πρώτο πρόβλημα είναι να βρούμε το σημείο αναφοράς που προαναφέραμε και δεύτερον αφού εντοπίσουμε το σημείο αναφοράς, θα χρειαστεί να μετατρέψουμε τις περιόδους που βρήκαμε βάση του σημείου αναφοράς σε δυαδικές ακολουθίες. Θεωρούμε ότι είναι πολύ σημαντικό να εξηγηθεί γιατί το σημείο αναφοράς είναι αναγκαίο για τον εντοπισμό της περιόδου δίνοντας ένα παράδειγμα. Έστω ότι ένα ζώο κάθε μέρα ένα 8ωρο παραμένει στη φωλιά του και την υπόλοιπη μέρα αναζητεί τη τροφή του σε τυχαία μέρη κάθε φορά. Οπότε έχουμε δύο σημεία το ζώο βρίσκεται στη φωλιά του όπου το συμβολίζουμε με 1 ή το ζώο αναζητάει τροφή και βρίσκεται εκτός φωλιάς και το συμβολίζουμε με 0. Όπως είναι κατανοητό μπορούμε να αντιληφθούμε πολύ εύκολα τη τακτικότητα στη συγκεκριμένη δυαδική ακολουθία. Η έννοια του σημείου αναφοράς έχει αρκετά θετικά χαρακτηριστικά, αρχικά φιλτράρετε ο θόρυβος από τα χωρικά δεδομένα που είναι και πολύ συχνός σε αυτού του είδους τα δεδομένα. Ένα ακόμη πλεονέκτημα είναι πως μετατρέπουμε το πρόβλημα από δυσδιάστατο χωρικό σε μονοδιάστατο χωρικό. Επίσης, υπάρχει μεγάλη περίπτωση ενώ στη πραγματικότητα να έχουμε δύο περιόδους, π.χ. ανά μέρα και ανά βδομάδα, να μην μπορούν να ανιχνευτούν και οι δύο λόγο του ότι θα εντοπίζεται μόνο η περίοδος ανά ημέρα διότι θα εμφανίζεται πιο συχνά. Ενώ εάν χρησιμοποιηθούν δύο διαφορετικά σημεία αναφοράς θα μπορέσουμε να εντοπίσουμε και τις δύο περιόδους ξεχωριστά. Τέλος βασιζόμενοι στην υπόθεση ότι κάθε περιοδική συμπεριφορά σχετίζεται με κάποιο σημείο αναφοράς, όλες οι περίοδοι μπορούν να εντοπιστούν βασιζόμενοι σε κάποιο σημείο αναφοράς.

2.3 Αλγόριθμος εντοπισμού μοτίβων σειριακών ακολουθιών

2.3.1 Χρήσιμες έννοιες

Έστω ότι έχουμε ένα σύνολο από *I* αντικείμενα σε μία βάση όπως στο παρακάτω παράδειγμα.

Sequence_id	Sequence
1	*{(a, b, e), (a, b, e), (a, d), (a, e), (a, b, c)}*
2	*{(c), (a, b, c, e), (c, d), (a, b, c, e), (a, b, d)}*
3	*{(b, c), (a, b), (a, c, d), (a, c), (a, b)}*
4	*{(a, b, d, e), (a, b, e), (a, b, c), (a, b, d, e), (a, b)}*

Στο παραπάνω παράδειγμα μπορούμε εύκολα να καταλάβουμε πως το *I* θα ισούται με *{(a,b,c,d,e)}.*

Με X θα ονομάζουμε ένα σετ αντικειμένων υποσύνολο του I, $X \subseteq I$. Κάθε σετ αντικειμένων περιέχει k αντικείμενα. Π.χ. στη πρώτη ακολουθία το πρώτο σετ αντικειμένων (a, b, e) περιέχει $k = 3$ αντικείμενα.

Μία ακολουθία s θα ονομάζεται μία λίστα με σετ από αντικείμενα $s = \{T_1, T_2, ..., T_m\}$, όπου $T_j \subseteq I$ $(1 \leq j \leq m)$, j είναι το αναγνωριστικό της συναλλαγής του σετ αντικειμένων T_j, και T_j μία συναλλαγή.

Η βάση D θα είναι το σύνολο από όλες τις ακολουθίες, $D = \{s_1, s_2, ..., s_n\}$. Η ακολουθία s_i της βάσης D είναι η i-στη ακολουθία του D, και i ταυτότητα της ακολουθίας.

Μία ακολουθία $s_a = \{A_1, A_2, ..., A_k\}$ θα θεωρείτε υποσύνολο της ακολουθίας $sb = \{B_1, B_2, ..., B_l\}$ εάν υπάρχει ακέραιος $1 \leq i_1 \leq i_2 \leq ... \leq i_k \leq m$ τέτοιο ώστε $A_1 \subseteq B_{i1}$, $A_2 \subseteq B_{i2}$, ..., $A_k \subseteq B_{il}$ (συμβολίζεται ως $s_a \sqsubseteq s_b$) . Για παράδειγμα στο παραπάνω πίνακα που περιέχει τέσσερις ακολουθίες, παρατηρούμε ότι στη πρώτη ακολουθία, στο το πρώτο σετ αντικειμένων περιέχει τρία αντικείμενα το (a, b, e). Είναι δηλαδή ένα σετ 3 αντικειμένων, όπου η ακολουθία $\{(a, b), (a)\}$ είναι υποσύνολο της ακολουθίας s_1.

<u>Ορισμός 1</u> (η περίοδος ενός σετ αντικειμένων στη ακολουθία).

Θεωρούμε μία ακολουθία s_i της βάσης D και ένα σετ αντικειμένων X.

Έστω $TR(X, s_i) = \langle T_{g1}, T_{g.} \sqsubseteq ... T_{gk} \rangle$ s_i είναι το ταξινομημένο σετ συναλλαγών στην οποία το σετ αντικειμένων X εμφανίζεται στην ακολουθία s_i. Δύο συναλλαγές T_x και T_y στην s_i, θα λέμε ότι είναι συνεχόμενες εάν δεν υπάρχει συναλλαγή $T_z \in s_i$ τέτοιο ώστε $x < z < y$ και $X \subseteq T_z$. Η περίοδος δυο συνεχόμενων συναλλαγών T_x και T_y θα ισούται με $per(T_x, T_y) = y - x$.

Οι περίοδοι σε μία ακολουθία s_i είναι $pr(X, s_i) = \{per_1, per_2, ..., per_{k+1}\}$ όπου $per_1 = g_1 - g_0$, $per_2 = g_2 - g_1$, ..., $per_{k+1} = g_{k+1} - g_k$, όπου $g_1, g_2, ... g_k$ είναι οι αριθμοί των συναλλαγών των σετ αντικειμένων X που εμφανίζονται και g_0 και g_{k+1} ορίζονται ως $g_0 = 0$ και $g_{k+1} = n$.

Π.χ. το σετ αντικειμένων $\{a, b\}$ εμφανίζεται στην συναλλαγή T_1, T_2 και T_5 της ακολουθίας s_1. Συνεπώς $TR(\{a, b\}, s_1) = \{T_1, T_2, T_5\}$ και οι περίοδοι του μοτίβου $\{a, b\}$ είναι $pr(\{a, b\}, s_1) = \{1, 1, 3, 0\}$. Δηλαδή έχουμε $g1 - g0 = 1 - 0 = 1$, $g2 - g1 = 2 - 1 = 1$, $g5 - g2 = 5 - 2 = 3$ και $g5 - g5 = 5 - 5 = 0$

<u>Ορισμός 2</u> (μέγιστη περιοδικότητα)

Η μέγιστη περιοδικότητα ενός σετ αντικειμένων X σε μία ακολουθία ορίζεται ως $maxPr(X, S) = argmax(pr(X, s))$.

Ένα μοτίβο θεωρείται περιοδικό εάν η μέγιστη περίοδος είναι μικρότερη από ένα όριο που ορίζει ο χρήστης $maxPr$ και όχι μικρότερο από ένα κατώφλι που ορίζει και πάλι ο χρήστης $minSup$. Σαν support ενός σετ αντικειμένων X στην ακολουθία s θα εννοείται ο αριθμός των συναλλαγών που περιέχει το X σε μία ακολουθία s, $sup(X, s) = |TR(X, s)|$

Για παράδειγμα έστω ότι ο χρήστης ορίζει σαν *maxPr = 3* και *minSup = 3*. Το σετ αντικειμένων *{a, b}* είναι περιοδικό στην ακολουθία s_1 εφόσον οι περίοδοι σε αυτήν την ακολουθία είναι $pr(\{a, b\}, s_1) = \{1, 1, 3, 0\}$ και η μέγιστη περίοδος είναι $maxPr(\{a, b\}, s_1) = max\{1, 1, 3, 0\} = 3 \leq maxPr$ *και* $sup(\{a, b\}, s_1) = 3 \geq minSup$.

Ένα πρόβλημα του *maxPr* είναι πως εάν ο χρήστης δώσει μία χαμηλή τιμή τα μοτίβα θα απορρίπτονται εάν έχουν μόλις λίγες περιόδους μεγαλύτερες του *maxPr*, ενώ εάν στο *maxPr* τεθεί μία υψηλή τιμή μοτίβα που με περιόδους που διαφέρουν αρκετά μεταξύ τους θα θεωρούνται περιοδικά. Για την αντιμετώπιση του συγκεκριμένου προβλήματος οι συγγραφείς από [5] πρότειναν τον παρακάτω ορισμό.

Ορισμός 3 (τυπική απόκλιση περιόδων)

Η τυπική απόκλιση των περιόδων ενός σετ αντικειμένων *X* σε μία ακολουθία s ορίζεται ως *stanDev (X, s).*

Για παράδειγμα, το σετ αντικειμένων *{a, b}* έχει τέσσερις περιόδους στην ακολουθία s_1, οι οποίες είναι *pr(X) = {1, 1, 3, 0}*. Η μέση περίοδος του *{a, b}* είναι $avgPr(\{a, b\}, s_1) = (1 + 1 + 3 + 0) / 4 = 1.25$.

Η τυπική απόκλιση θα ισούται

$stanDev(\{a, b\}, s_1) = \sqrt{[(1 - 1.25)^2 + (1 - 1.25)^2 + (3 - 1.25)^2 + (0 - 1.25)^2]/4} = 1.09$

Ορισμός 4 (περιοδικά μοτίβα σε μία ακολουθία)

Έστω ότι έχουμε τρία όρια που δίνονται από τον χρήστη *maxPer, minSup και maxStd*. Ένα σετ αντικειμένων *X* είναι περιοδικό σε μία ακολουθία *s* εάν $maxPr(X, s) \leq max\ Pr$, $sup(X, s) \geq minSup$ και $stanDev(X, s) \leq maxStd$.

Για παράδειγμα εάν έχουμε το σετ αντικειμένων *{a, b}* θα θεωρείται περιοδικό σε μία ακολουθία s_1 για *maxStd =2.0*, εφόσον το $stanDev(\{a, b\}, s_1) = 1.09$. Ο παραπάνω ορισμός προτείνετε για τον εντοπισμό μοτίβων σε μία ακολουθία. Για τον εντοπισμό περιοδικών μοτίβων που είναι κοινά σε πολλαπλές ακολουθίες χρησιμοποιείται η αναλογία περιοδικής ακολουθίας.

Ορισμός 5 (αναλογία περιοδικής ακολουθίας)

Ο αριθμός των ακολουθιών όπου ένα σετ αντικειμένων *X* είναι περιοδικό σε μία βάση *D* από ακολουθίες ορίζεται ως *numSeq(X).*

Η αναλογία περιοδικής ακολουθίας του *X* στην *D* ορίζεται ως $ra\ (X) = numSeq(X)/\ |D|$, όπου $|D|$ ο αριθμός των ακολουθιών στην *D*.

Για παράδειγμα, ο αριθμός των ακολουθιών όπου *{a, b}* είναι περιοδικός είναι $numSeq(\{a, b\}) = 3$ (είναι περιοδικός στις s_1, s_2, s_4) . Ο αριθμός των ακολουθιών $|D| = 4$. Συνεπώς, η αναλογία περιοδικής ακολουθίας του *{a, b}* είναι $ra(\{a, b\}) = ¾ = 0.75$.

Ορισμός 5 (*boundRa*)

Για την βελτιστοποίηση του αλγορίθμου συστήνετε ένα ακόμη μέτρο που μειώνει τον χώρο στις αναζητήσεις το οποίο είναι ένα άνω όριο του μέτρου ra. Δεδομένης των εισακτέων τιμών που ορίζονται από τον χρήστη *maxPer* και *minSup*, ένα σετ αντικειμένων *X* είναι υποψήφιο για να ελεγχθεί εάν εμφανίζει κάποια περιοδικότητα εάν $maxPr(X, s) \leq maxPr$ και $sup(X, S) \geq minSup$. Ο αριθμός των ακολουθιών όπου ένα σετ αντικειμένων *X* είναι υποψήφιο για το εάν εμφανίζει κάποια περιοδικότητα στη βάση *D* ορίζεται ως *numCand(X).* Τα *boundRa* των *X* στην βάση *D* ορίζονται ως $boundRa(X) = numCand(X)/|D|$.

2.3.2 Αλγόριθμος MPFPS

Η μέθοδος που πρότειναν οι συγγραφείς από [5] αποτελείτε από 3 Αλγόριθμους :

Αλγόριθμος 1

Αρχικά σκανάρει τη βάση μία φορά για να υπολογίσει τα $sup(\{i\}, s)$, $pr(\{i\}, s)$, $maxPr(\{i\}, s)$ και $stanDev(\{i\}, s)$ για κάθε προϊόν *i* και ακολουθία *s*. Στη συνέχεια ελέγχει εάν κάθε προϊόν *i* εμφανίζει κάποια περιοδικότητα σε κάθε ακολουθία της βάσης (γραμμή 2 έως 4). Ένα προϊόν *i* θεωρείται ότι παρουσιάζει κάποια περιοδικότητα στην ακολουθία *s εάν* $sup(\{i\}, s) \geq minSup$, $maxPr(\{i\}, s) \leq maxPr$ και $stanDev(\{i\}, s) < maxStd$. Μετά ο αλγόριθμος υπολογίζει την αναλογία περιοδικής ακολουθίας του αντικειμένου *i* διαιρώντας τον αριθμό των ακολουθιών όπου το *i* είναι περιοδικό με τον αριθμό των ακολουθιών που βρίσκονται στη βάση. Εάν αυτή η τιμή δεν είναι μικρότερη από το *minRa*, το *i* θεωρείται ότι εμφανίζει συχνά περιοδικά μοτίβα κοινά σε πολλαπλές ακολουθίες (γραμμή 4). Επίσης, το *boundRa* του $\{i\}$ υπολογίζεται με σκοπό να μειώσει τις αναζητήσεις που θα χρειαστεί να γίνουν (γραμμή 5 και 6). Έπειτα η λίστα με τα μοναδικά προϊόντα τα οποία εμφανίζουν συχνά περιοδικά μοτίβα τα οποία είναι κοινά σε πολλαπλές ακολουθίες αποθηκεύονται σαν δεσμευμένα *PFPS Single* (Periodic Frequent Patterns common to multiple Sequences) (γραμμή 8). Στη συνέχεια, η πρώτη σε βάθος αναζήτηση *PFPS* ξεκινάει καλώντας την επαναλαμβανόμενη αναζήτηση με τις δεσμευμένες *PFPS Single, minSup, maxP r, maxStd,* minRa and *D*.

Παρακάτω βλέπουμε σε μορφή ψευδογλώσσας τον Αλγόριθμο 1

Algorithm 1. The MPFPS algorithm

input : D: a database with multiple sequences, $maxStd, minRa, maxPr, minSup$: the thresholds.
output: the set of periodic frequent patterns (PFPS).

1 Scan each sequence $s \in D$ to calculate $sup(\{i\}, s)$, $pr(\{i\}, s)$, $maxpr(\{i\}, s)$ and $stanDev(\{i\}, s)$ for each item $i \in I$;
2 **foreach** *item* $i \in I$ **do**
3 $\quad numSeq(\{i\}) \leftarrow |\{s|maxpr(\{i\}, s) \leq maxPr \wedge stanDev(\{i\}, s) \leq maxStd \wedge sup(\{i\}, s) \geq minSup \wedge s \in D\}|$;
4 $\quad ra(\{i\}) \leftarrow numSeq(\{i\})/|D|$;
5 $\quad numCand(\{i\}) \leftarrow |\{s|maxpr(\{i\}, s) \leq maxPr \wedge sup(\{i\}, s) \geq minSup \wedge s \in D\}|$;
6 $\quad boundRa(\{i\}) \leftarrow numCand(\{i\})/|D|$;
7 **end**
8 $boundPFPSSingle \leftarrow$ {PFPS-list of item $i|i \in I \wedge boundRa(\{i\}) \geq minRa\}$;
9 Sort $boundPFPSSingle$ by the order $\succ$ of ascending support values;
10 **Search** ($boundPFPSSingle$, $minSup$, $maxPr$, $maxStd$, $minRa$, D);

Αλγόριθμος 2

Δέχεται σαν εισακτέες τιμές τις τιμές που ορίζει ο χρήστης στα *minSup, maxP r, maxStd, minRa*, μία λίστα με *PFPS* προεκτάσεις ενός σετ αντικειμένων *P* και τη βάση.

Αρχικά θα δώσουμε τον ορισμό της προέκτασης ενός σετ αντικειμένων *P*, όπου θα ορίζεται ως P_z και θα ισούται με ένα σετ αντικειμένων *P* το οποίο αποκτήθηκε όταν προστέθηκε ένα αντικείμενο *z*. Όταν ξεκινάει η διαδικασία τα *P* είναι κενά και οι προεκτάσεις των *P* περιέχουν μόλις ένα αντικείμενο. Η διαδικασία αναζήτησης εφαρμόζει μία λούπα για κάθε προέκταση P_x του *P*. Ο αλγόριθμος αρχικά υπολογίζει τα $numCand(P_x)$ και $boundRa(P_x)$ χρησιμοποιώντας την *PFPS* λίστα από P_x, η οποία ορίζεται ως LP_x (γραμμή 2-3) . Στη συνέχεια ελέγχει εάν $boundRa(P_x) \geq minRa$, στη περίπτωση που ισχύει η συνθήκη υπολογίζεται το $ra(P_x)$ ως $numSeq(Px)/|D|$. Αφού υπολογιστεί το $ra(P_x)$ ελέγχετε εάν είναι μεγαλύτερο ή ίσο του *minRa*, στην περίπτωση που είναι ο αλγόριθμος κρατάει το συγκεκριμένο P_x (γραμμή 5-7). Τέλος, εφαρμόζεται μία λούπα για να συνδυάσει τα P_x με κάθε προέκταση P_y τέτοιο ώστε $y \succ x$, για να παραχθεί μία προέκταση P_{xy}, όπου θα περιέχει $|P_x| + 1$ αντικείμενα (γραμμή 8 με 11). Η *PFPS* λίστα όλων των P_{xy} προεκτάσεων θα ορίζεται ως LP_{xy}.

Παρακάτω βλέπουμε σε μορφή ψευδογλώσσας τον Αλγόριθμο 2

Algorithm 2. The *Search* procedure

```
   input  : ExtensionsOfP: a set of PFPS-lists of extensions of an itemset P,
            minSup, maxPr, maxStd, minRa: the thresholds, D: the database.
   output: the set of periodic frequent patterns that extend P.
 1 foreach PFPS-list LPx ∈ ExtensionsOfP and Px = LPx.i-set do
 2 |   numCand(Px) ← |{s|maxpr(Px, s) ≤ maxPr ∧ sup(Px, s) ≥ minSup ∧ s ∈ D}|;
 3 |   boundRa(Px) ← numCand(Px)/|D|;
 4 |   if boundRa(Px) ≥ minRa then
 5 |   |   numSeq(Px) ← |{s|maxpr(Px, s) ≤ maxPr ∧ stanDev(Px, s) ≤
   |   |   maxStd ∧ sup(Px, s) ≥ minSup ∧ s ∈ LPx.sid-list}|;
 6 |   |   ra(Px) ← numSeq(Px)/|D|;
 7 |   |   if ra(Px) ≥ minRa then output Px ;
 8 |   |   foreach PFPS-list LPy ∈ ExtensionsOfP and Py = LPy.i-set such that y ≻ x
   |   |   do
 9 |   |   |   LPxy ← Intersect (LPx, LPy);
10 |   |   |   ExtensionsOfPx ← ExtensionsOfPx ∪{LPxy};
11 |   |   end
12 |   end
13 |   Search (ExtensionsOfPx, minSup, maxPr, maxStd, minRa, D);
14 end
```

Αλγόριθμος 3

Αρχικά σαν εισακτέες τιμές δέχεται τις λίστες *PFPS* από τα σετ αντικειμένων P_x και P_y, τα οποία ορίζονται ως LP_x και LP_y και εξάγει τη λίστα *PFPS* του σετ αντικειμένων P_{xy}. Ο αλγόριθμος ξεκινά αρχικοποιώντας μία κενή λίστα LP_{xy} για το P_{xy} (γραμμή 1) . Στη συνέχεια, ο αλγόριθμος εκτελεί μία λούπα για να υπολογίσει κάθε ακολουθία στην οποία εμφανίζονται και το P_x και το P_y. Έστω s_i το αναγνωριστικό που θα χρησιμοποιείται για αυτές τις ακολουθίες. Το συγκεκριμένο αναγνωριστικό θα χρησιμοποιείται για να ανακτηθούν οι λίστες με τις θέσεις των συναλλαγών που περιέχουν P_x και P_y στη συγκεκριμένη ακολουθία και θα ορίζονται ως $tidListS_iP_x$ και $tidListS_iP_y$. (γραμμές 3-4). Αυτές οι δύο λίστες στη συνέχεια διασταυρώνονται για να ανακτηθεί η λίστα με τις θέσεις των συναλλαγών που περιέχουν P_{xy} στη συγκεκριμένη ακολουθία όπου θα ονομάζεται $tidListS_iP_{xy}$ (γραμμή 5). Εάν αυτή η λίστα δεν είναι κενή, το s_i προστίθεται στη *PFPS* λίστα των P_{xy} όπως και η λίστα με τις συναλλαγές $tidListS_iP_{xy}$. Η διαδικασία επιστρέφει την *PFPS* λίστα του P_{xy}.

Παρακάτω βλέπουμε σε μορφή ψευδογλώσσας τον Αλγόριθμο 3

Algorithm 3. The Intersect procedure

input : LPx and LPy: the PFPS-lists of two extensions Px and Py of an itemset
output: the PFPS-list $LPxy$ of itemset Pxy

```
1  LPxy.i-set ← Px ∪ {y}; LPxy.tidlist-list ← ∅; LPxy.sid-list ← ∅;
2  foreach sequence identifier si ∈ LPx.sid-list such that si ∈ LPy.sid-list do
3      tidListSiPx ← the tid list of si in LPx.tidlist-list;
4      tidListSiPy ← the tid list of si in LPy.tidlist-list;
5      tidListSiPxy ← tidListSiPx ∩ tidListSiPy;
6      if tidListSiPxy ≠ ∅ then
7          LPxy.sid-list.append(si); LPxy.tidlist-list.append(tidListSiPxy);
8      end
9  end
10 return LPxy;
```

2.3.4 Συμπεράσματα

Οι συγγραφείς από [5] πρότειναν έναν καινοτόμο αλγόριθμο για την εξόρυξη περιοδικών μοτίβων τα οποία είναι κοινά σε πολλαπλές ακολουθίες. Ο συγκεκριμένος αλγόριθμος ονομάζεται MPFPS ο οποίος βασίζεται σε μία λίστα *PFPS* (Periodic Frequent Patterns common to multiple Sequences). Πειράματα τα οποία διεξήγαγε η παραπάνω ομάδα σε διαφορετικές βάσεις δεδομένων έδειξε πως ο συγκεκριμένος αλγόριθμος είναι αποτελεσματικός και μπορεί να φιλτράρει με μεγάλη ακρίβεια τα μη περιοδικά μοτίβα.

2.4 Αλγόριθμοι εντοπισμού μοτίβων χωροχρονικών δεδομένων

2.4.1 STPMine1, STPMine2 και STPMine2 – V2 [6]

2.4.1.1 Χρήσιμες έννοιες

Οι συγγραφείς από [6] πρότειναν δύο μεθόδους για την εύρεση των μοτίβων, ο πρώτος αλγόριθμος SPMine1 (Spatio Temporal periodic Pattern Min(e)ing) χρησιμοποιεί μία bottom-up τεχνική η οποία μοιάζει με την συσσωρευτική Ιεραρχική μέθοδο κατά συστάδες δηλαδή ξεκινάει από ένα στοιχείο μόνο του και ανεβαίνει προς τα πάνω με τη μορφή δενδροδιαγράμματος σχηματίζοντας και σε κάθε πάνω βήμα από ένα μοτίβο μεγαλύτερου επιπέδου από το προηγούμενο. Ο δεύτερος αλγόριθμος STPMine2 ο οποίος είναι και πιο γρήγορος αναφέρεται σε μία top-down τεχνική η οποία μοιάζει με την διαιρετική μέθοδο κατά συστάδες, δηλαδή θα ξεκινήσει με το μεγαλύτερο επίπεδο και σε κάθε κάτω βήμα θα

διαιρείτε σε υποομάδες. Παρακάτω θα δούμε πως δουλεύουν οι δύο αλγόριθμοι λίγο πιο αναλυτικά.

Αρχικά θα αναφέρουμε κάποιες μεταβλητές που θα μας βοηθήσουν στην κατανόηση των αλγορίθμων:

T: Καλούμε μία σταθερά η οποία δίνεται από το χρήστη και ονομάζεται περίοδος(π.χ. μέρα, βδομάδα, χρόνος)

P: Ένα μοτίβο *P* είναι μία *T*-μήκους ακολουθία της μορφής $r_0 r_1 \dots r_{T1}$, όπου r_i είναι μία συχνή περιοχή ή ο ειδικός χαρακτήρας * όπου υποδηλώνει ότι το αντικείμενο μπορεί να βρίσκεται οπουδήποτε.

Για παράδειγμα, ένα μοτίβο *AB* C*** , σημαίνει πως στην αρχή του κύκλου το αντικείμενο βρίσκεται στην περιοχή *A*, στην επόμενη χρονική σήμανση βρίσκεται στη περιοχή *B*, αργότερα θα μπορούσε να βρίσκεται οπουδήποτε και στη συνέχεια θα βρίσκεται στη περιοχή *C*, μετά θα κινούνταν ξανά ακανόνιστα έως ότου ξανάρχιζε ο κύκλος του με τη περιοχή *A*.

S : Θα ονομάζουμε την ακολουθία μήκους n του δείγματος των περιοχών που επισκέφτηκε ένα αντικείμενο για κάθε χρονική σήμανση και θα είναι της μορφής $\{(l_0,t_0),(l_1,t_1), \dots, (l_{n-1},t_{n-1})\}$, όπου l_i η περιοχή που βρισκόταν το αντικείμενο τη χρονική στιγμή t_i

min sup : Μία παράμετρος που την ορίζει ο χρήστης $(0 < min\ sup \leq 1)$ και θα πρέπει να ισχύει το εξής. Το μοτίβο θα πρέπει να ακολουθείται από το αντικείμενο το λιγότερο α περιοδικά διαστήματα όπου $\alpha = min\ sup \times \left\lfloor \frac{n}{T} \right\rfloor$, στο εξής το ποσό $\left\lfloor \frac{n}{T} \right\rfloor$ θα το καλούμε *m*.

L_1: Αρχικά αφού χωρίσουμε την ακολουθία *S* σε *T* χωρικές βάσεις, δηλαδή περιοχές $\{l_i, l_{i+T}, \dots, l_{i+(m-1)\cdot T}\} = R_i$, για κάθε $0 \leq i \leq T$. Θα εφαρμόσουμε έναν αλγόριθμο της μεθόδου συστάδων βασισμένη στη πυκνότητα όπως π.χ. ο αλγόριθμος DBSCAN. Συστάδες με λιγότερο από α σημεία απορρίπτονται, οι υπόλοιπες συστάδες θα είναι τα 1-pattern που θα χρησιμοποιηθούν σαν input στους αλγόριθμους.

2.4.1.2. STPMine1 :

Ο αλγόριθμος STPMine1 για να εξάγει τα μοτίβα μοντελοποιεί το χρόνο στα χωροχρονικά δεδομένα μας. Βρίσκει τα πιο συχνά χωρικά 1-patterns (L_1) για προκαθορισμένο χρόνο από τα δεδομένα μας και για την ανακάλυψη μοτίβων μεγαλύτερου μήκους, χρησιμοποιεί τον αλγόριθμο Apriori-TID όπου σαν εισακτέα τιμή χρησιμοποιεί τα 1-patterns που αναφέραμε παραπάνω.

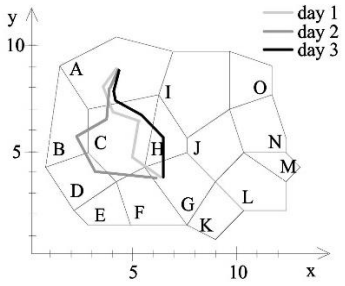
Παρακάτω βλέπουμε τον αλγόριθμο σε μία ψευδογλώσσα.

Algorithm STPMine1($\mathcal{L}_1, T, min\_sup$);
1). k:=2;
2). **while** ($\mathcal{L}_{k-1} \neq \emptyset \wedge k < T$)
3). $\mathcal{L}_k$:=$\emptyset$;
4). **for each** pair of patterns $(P_1, P_2) \in \mathcal{L}_{k-1}$
5). such that P_1 and P_2 agree on the first $k-2$
6). and have different $(k-1)$-th non-* position
7). P_{cand}:=**candidate_gen**(P_1, P_2);
8). **if** ($P_{cand} \neq null$) **then**
9). P_{cand}:=$P_1 \bowtie_{P_1.sid=P_2.sid} P_2$; //segment-id join
10). **if** ($|P_{cand}| \geq min\_sup \cdot m$) **then**
11). **validate_pattern**($P_{cand}, \mathcal{L}_k, min\_sup$);
12). k:=$k+1$;
13). **return** $\mathcal{P}$:=$\bigcup \mathcal{L}_k, \forall 1 \leq k < T$;

Βλέπουμε ότι σαν input ο αλγόριθμος έχει τα *(L1, T, min sup)* όπου έχει εξηγηθεί παραπάνω τι σημαίνει η κάθε μεταβλητή.

Τα ζευγάρια *<P_1, P_2> (k-1)-pattern στο L_k-1*, για να θεωρηθούν υποψήφιο *k-pattern* θα πρέπει οι πρώτες *k - 2* μη * (μη πυκνές περιοχές) περιοχές να είναι στην ίδια περιοχή και θέση και η *(k -1)* μη * θέση τους να είναι διαφορετική. Εάν ισχύουν οι παραπάνω προϋποθέσεις δημιουργείτε ένα υποψήφιο *k-pattern (P_{cand}).* Σε κάθε P_{cand} , ενώνουμε τα τμηματικά-id των P_1, P_2 και εάν ο αριθμός των τμημάτων που συμφωνεί και με τα δύο μοτίβα είναι το λιγότερο *α (min sup* × $\left\lfloor \frac{n}{T} \right\rfloor$*)*, εκτελούμε μία αξιολόγηση του μοτίβου για να ελέγξουμε εάν οι περιοχές του P_{cand} είναι ακόμη συστάδες. Αφού τα μοτίβα μήκους *k* έχουν ανακαλυφθεί, συνεχίζουμε για τα μοτίβα του επόμενου επιπέδου έως ότου δεν υπάρχουν άλλα στο συγκεκριμένο επίπεδο ή δεν υπάρχουν άλλα επίπεδα.

Παρακάτω θα δώσουμε ένα παράδειγμα για να εξηγήσουμε τη διαφορά της περιοχής με τη θέση.



(b)

events sequence:
A A C C C G | A A C B D G | A A A C H G

some partial periodic patterns:
support(AA***G) = 3
support(AAC**G) = 2
support(AA*C*G) = 2

(c)

Στη παραπάνω εικόνα παρατηρούμε πως το αντικείμενο μας και τις 3 μέρες βρίσκεται στην ίδια περιοχή και στην ίδια θέση τη στιγμή 1 όπου για παράδειγμα θα μπορούσε να είναι το σπίτι του. Τη στιγμή 2 επίσης και στις 3 μέρες βρίσκεται στην ίδια περιοχή και σε πολύ κοντινές θέσεις όπου για παράδειγμα θα μπορούσε να είναι κάποιες κοντινές διαδρομές που οδηγούν στο ίδιο μέρος. Τη στιγμή 3 παρατηρούμε ότι ενώ βρίσκεται σε κοντινές θέσεις τη τρίτη μέρα βρίσκεται σε διαφορετική περιοχή την *A* σε σύγκριση με τις άλλες δύο που βρίσκονται στη περιοχή *C*. Ενώ τις υπόλοιπες στιγμές παρατηρούμε να διαφοροποιούνται αρκετά οι θέσεις και οι περιοχές του αντικειμένου έως και τι στιγμή 5 όπου παρατηρούμε ότι βρίσκεται στην ίδια περιοχή και την ίδια θέση *G*. Που θα μπορούσε να είναι 3 διαφορετικά μονοπάτια που ακολούθησε το αντικείμενο μας για να καταλήξει στο ίδιο σημείο.

2.4.1.3. STPMine2 :

Ο αλγόριθμος STPMine2 επίσης έχει την ικανότητα να εξάγει μοτίβα για κινούμενα αντικείμενα με χωρικά και χρονικά δεδομένα. Μπορεί να εντοπίσει τα κινούμενα μοτίβα ανιχνεύοντας τη βάση μας μόλις δύο φορές, σε αντίθεση με τον STPMine1. Το οποίο σημαίνει ότι ο αλγόριθμος βελτιώνει την απόδοση σε χρόνο για δεδομένα με μοτίβα μεγάλου μήκους σε σύγκριση με τον αλγόριθμο STPMine1. Αν και ο αλγόριθμος STPMine2 μειώνει τον χρόνο σκαναρίσματος της βάσης, όσο αυξάνονται τα αντικείμενα που μελετάμε και συνεπώς αυξάνονται τα χωροχρονικά δεδομένα μας τόσο αυξάνονται και τα πιθανά μοτίβα κατεβαίνοντας το δενδροδιάγραμμα με γεωμετρική αύξηση. Συνεπώς, μπορεί να προκαλέσει προβλήματα στη μνήμη μιας και το σύστημα θα υπερφορτωθεί.

Παρακάτω βλέπουμε τον αλγόριθμο σε μία ψευδογλώσσα.

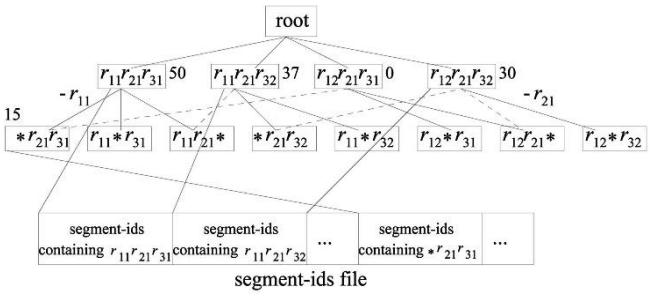
Algorithm STPMine2($\mathcal{L}_1, T, min\_sup$);
1). build max-subpattern tree $\mathcal{T}$ and pattern-file F;
2). sort F on $P'.id$ and connect it to the nodes of $\mathcal{T}$;
3). **for** k:=T down to 2
4). **for each** pattern P' at level k of $\mathcal{T}$
5). $|P'|$:=$P'.counter$ +$\sum_{P'' \supset P', length(P'')=k+1} |P''|$;
6). **if** ($|P'| \geq min\_sup \cdot m$) **then**
7). P_{cand}:=$\bigcup_{P'' \supseteq P'} P''.sids$;
8). **validate_pattern**($P_{cand}, \mathcal{L}, min\_sup$);
9). **if** ($\mathcal{P}$ has changed) **then**
10). remove from P' those $sids$ in new patterns of $\mathcal{P}$;
11). **if** (unassigned $sids$ less than $min\_sup \cdot m$) **then**
12). **return** $\mathcal{P}$;
13). **return** $\mathcal{P}$;

Αρχικά θα δώσουμε έναν ορισμό του τι ακριβώς είναι το superpattern, έστω ότι έχουμε δύο διαφορετικά patterns $P = r_0r_1r_2...r_{T-1}[b,e]$ και $P' = r'_0r'_1r'_2...r'_{T-1}\ [b',e']$. P είναι ένα superpattern του P' εάν:

1. $r_i = r'_i$ ή $r'_i = *$ για $0 \leq i < T$ και
2. $[b,e]$ είναι ένα υπερσύνολο του $[b',e']$

Πρώτα το δενδροδιάγραμμα και το αρχείο με τα τμηματικά-ids δημιουργούνται και συνδέονται. Αργότερα για κάθε επίπεδο ξεκινώντας από το max-pattern, βρίσκουμε το στήριγμα ενός ψευδό μοτίβου $|P'|$ επιπέδου k έχοντας πρόσβαση μόνο στο στήριγμα του superpattern του επιπέδου $k + 1$. Εάν $|P'| \geq \alpha$, αξιολογούμε το μοτίβο με την ίδια διαδικασία όπως και στον STPMine1, εφόσον ανακαλυφθούν κάποια μοτίβα, αφαιρούμε από το P' όλα τα τμήματα-ids που συμμορφώνονται με το μοτίβο. Έτσι, ο αριθμός των τμηματικών-ids μειώνεται όσο κατεβαίνουμε και επίπεδο στο δέντρο, έως ότου δεν είναι πια δυνατόν να ανακαλυφθούν περισσότερα μοτίβα ή δεν υπάρχουν άλλα επίπεδα.

Παρακάτω βλέπουμε ένα παράδειγμα ενός δέντρου.



2.4.1.4. STPMine2 – V2:

Επίσης υπάρχει και μία δεύτερη έκδοση του STPMine2 αλγόριθμου, όπου αποτελεί μία πιο απλοποιημένη έκδοση αυτού. Σε αυτή την έκδοση του αλγορίθμου αφού εντοπιστούν οι πυκνές περιοχές δεν γίνεται κάποια επαναξιολόγηση των συστάδων που σημαίνει ότι

κάποιο στοιχείο δεν θα έχει εκ νέου ανάθεση. Με αποτέλεσμα αυτός ο αλγόριθμος να είναι πολύ πιο γρήγορος σε σύγκριση με τον STPMine2 αλλά επίσης να είναι και πιο ανακριβής.

Παρακάτω βλέπουμε έναν ψευδό-αλγόριθμο που περιγράφει πως εξάγουμε συχνά μοτίβα από το P_{max}. (Ο αλγόριθμος που υιοθετήθηκε για να δημιουργηθεί ο STPMine2-V2)

Algorithm SPMine(P_{max}, T, min_sup);
1). **for** l:= ℓ_{max} down to 2 //ℓ_{max} is the length of P_{max}
2). **for** every subpattern of P_{max} with length l and with no frequent superpattern;
3). $|P|$:= 0; //$|P|$ is the support of P (Definition 2)
4). **for** each non-* position i of P with cluster r_i
5). p_i:=first point in r_i;
6). **if** ($\{p_1, p_2, \ldots, p_l\}$ is not a valid instance of P) **then**
7). **if** ($p_i = p_j$, for some $i < j$) **then**
8). p_j:= next point in r_j;
9). **else** j:=$\{i : p_i$ has the smallest timestamp$\}$;
10). p_j:= next point in r_j;
11). **else** //valid pattern instance
12). $|P|$:=$|P| + 1$;
13). **for** each non-* position i of P with cluster r_i
14). p_i:= next point in r_i;
15). **if** (more points in all r_i) **then goto** line 6;
16). **if** ($|P| \geq min\_sup \cdot m$) **then** report P;

Αρχικά εξετάζονται τα subpatterns του P_{max} ανά επίπεδο. Για κάθε υποψήφιο subpattern το οποίο σχηματίζεται από ένα σύνολο συστάδων, μία συστάδα για κάθε μη * περιοχή *i*, αρχικοποιούμε ένα δείκτη p_i στο πρώτο σημείο για κάθε συστάδα r_i. Αργότερα κάνουμε ένα merge-join ελέγχοντας το περιεχόμενο των συστάδων ενώ ταυτόχρονα επιχειρούμε να εντοπίσουμε τυχόν περιπτώσεις shifted/ distorted μοτίβων βασιζόμενοι σε κάποιους δείκτες για κάθε p_i. Δεδομένου της θέσης του τρέχον p_i, εάν το σύνολο των περιοχών είναι περιπτώσεις shifted/ distorted μοτίβων, αυξάνουμε όλα τα p_i, εφόσον δεν επιθυμούμε να μετρηθούν περισσότερες περιπτώσεις που έχουν κοινές περιοχές με το τρέχον p_i. Διαφορετικά, εάν υπάρχει ένα ζευγάρι από p_i με παρόμοιες θέσεις (οι οποίες έχουν συσταδοποιηθεί σε διαφορετικές αντισταθμίσεις), αυξάνουμε τους δείκτες στη συστάδα οι οποίοι αντιστοιχούν στο μεγαλύτερο συμψηφισμό. Εάν δεν υπάρχει ένα τέτοιο ζευγάρι με παρόμοια p_i, αυξάνουμε το p_i με τη μικρότερη χρονοσήμανση. Τέλος, αναφέρουμε το μοτίβα εάν εμφανίζεται συχνά.

2.4.1.6. Συμπεράσματα

Αναλύθηκαν τα προβλήματα στην εξόρυξη δεδομένων περιοδικών μοτίβων και προτάθηκαν κάποιοι αλγόριθμοι για την επίλυση των προβλημάτων αυτών. Η μέθοδος προτείνει χωρικό clustering για να ανακαλυφθούν τα 1-patterns και προσαρμόζει μία bottom-up και μία top-down τεχνική για μοτίβα μεγαλύτερου μήκους. Επιπλέον προτάθηκε μία μέθοδος όπου επαναπροσδιόρισε τα συχνά 1-patterns για να προσαρμόζεται στις περιπτώσεις που έχουμε shifted ή distorted μοτίβα.

2.4.2 Periodica [7]

2.4.2.1. Χρήσιμες έννοιες

Αρχικά θα εξηγηθούν κάποιες μεταβλητές και κάποιες έννοιες όπως αναφέρονται και από τους συγγραφείς του paper [7] και θα είναι χρήσιμες για την κατανόηση του αλγορίθμου.

i Ορίζουμε ως $D = \{(x_1,y_1,time_1),(x_2,y_2,time_2),\ldots\}$ τα αρχικά μας δεδομένα για ένα κινούμενο αντικείμενο.

ii Θα χρησιμοποιήσουμε μία γραμμική παρεμβολή στα αρχικά μας χωρικά δεδομένα με σταθερά ένα χρονικό όριο π.χ. μία ώρα ή μία μέρα. Σαν $LOC = loc_1,loc_2 \cdots loc_n$ θα ορίσουμε την παρεμβαλόμενη ακολουθία όπου κάθε loc_i εκφράζεται ως ένα χωρικό σημείο $(loci.x, loci.y)$.

iii Ως σημείο αναφοράς ορίζουμε μία πυκνή περιοχή, δηλαδή μία περιοχή που εντοπίζεται συχνά στα δεδομένα μας να επισκέπτεται από το κινούμενο αντικείμενο. Το σύνολο όλων των σημείων αναφοράς εκφράζεται ως $O = \{o_1,o_2,\ldots,o_d\}$, όπου d είναι αριθμός των σημείων αναφοράς. Για να εντοπιστούν οι συχνές περιοχές, σαν πρώτο βήμα εφόσον ο υπολογισμός για κάθε περιοχή σε συνεχή δεδομένα είναι μη πρακτικός θα μειωθούν τα χωρικά δεδομένα σε $w \times h$ κελιά (καθορίζονται αναλόγως και την εκάστοτε ανάλυση). Αργότερα προσαρμόστηκε μία μέθοδος Kernel, η οποία έχει αρχικά σχεδιαστεί για να εντοπίζει το εύρος των σπιτιών κάποιων ζώων. Εάν για παράδειγμα ένα ζώο εντοπίζεται να έχει συχνά κάποιες δραστηριότητες σε ένα μέρος, αυτό το μέρος θα έχει και μεγαλύτερη πιθανότητα να είναι το σπίτι του. Οπότε για κάθε κελί c η πυκνότητα εκτιμάται χρησιμοποιώντας τη σ.π.π kernel

$$f(c) = \frac{1}{n\gamma^2} + \sum_{i=1}^{n}\left(\frac{1}{2\pi}\, exp\,(-\frac{|c\ -\ loc_i|^2}{2\gamma^2}\right)$$

Όπου $|c - loc_i|$ η απόσταση μεταξύ του κελιού c και της περιοχής loc_i και γ μία παράμετρος εξομάλυνσης

$$\gamma = \frac{1}{2}\left(\sigma_x^2 + \sigma_y^2\right)^{\frac{1}{2}} n^{-\frac{1}{6}}$$

Με σ_x^2 , σ_y^2 οι τυπικές αποκλίσεις όλης της ακολουθίας LOC για τις x, y συντεταγμένες αντίστοιχα. Αφού βρούμε τις τιμές πυκνότητας ένα σημείο αναφοράς μπορεί να εντοπιστεί από μία ισοϋψής καμπύλη στο χάρτη, η οποία ενώνει τα κελία με ίσες τιμές πυκνότητας, χρησιμοποιώντας μία τιμή πυκνότητας ως κατώφλι. Το κατώφλι ορίζεται ως οι $p\%$ μεγαλύτερες τιμές πυκνότητας μεταξύ όλων των κελιών.

iv Ως περίοδο T ορίζεται ένα τακτικό χρονικό διάστημα ανά το οποίο εντοπίζονται κάποιες κινήσεις. Για τον εντοπισμό των περιόδων μέσα στις ακολουθίες, οι πιο γνωστές μέθοδοι είναι η μέθοδος μετατροπής του Fourier και οι αυτοσυσχετίσεις. Από τη μία μεριά η μέθοδος μετατροπής του Fourier έχει ένα μειονέκτημα στο να πάρει μία απόφαση όταν έχουμε χαμηλή συχνότητα στις περιοχές, ως εκ τούτου μας παρέχει κακές εκτιμήσεις στις μεγάλες περιόδους. Επίσης τείνει να μας δίνει πολλές λανθασμένα θετικές περιόδους στο περιοδιόγραμμα. Από την άλλη οι αυτοσυσχετίσεις μας προσφέρουν ακριβής εκτιμήσεις για μικρές και μεγάλες περιόδους, όμως είναι πιο δύσκολο να ορίσουμε το κατώφλι σημαντικότητας για τις σημαντικές περιόδους. Συνεπώς προτείνετε μία μέθοδος που συνδυάζει τις δύο παραπάνω μεθόδους.

Δεδομένου του συνόλου από σημεία αναφοράς, η ακολουθία από κινήσεις που είχαμε μπορεί να μετατραπεί σε μία ακολουθία από δυαδικές μεταβλητές $B = b_1 b_2 \ldots b_n$ όπου το κάθε b_i θα ισούται με 1 όταν το αντικείμενο είναι στο σημείο αναφοράς και 0 όταν δεν είναι. Χρησιμοποιώντας τη διακριτική κατανομή Fourier η ακολουθία $B = b_1 b_2 \ldots b_n$ μετατρέπεται σε μια ακολουθία n περίπλοκων αριθμών $X_1, X_2, \ldots, X_n$.

Όπου:

$$X\kappa = \sum_{j=0}^{n-1} b_j e^{-i\,2\pi \frac{k-1}{n}} \quad \text{για } i = \sqrt{-1} \quad \text{και } \kappa = 1,2,\ldots,n$$

Αφού βρούμε τους συντελεστές Fourier, για να εντοπίσουμε τις περιοδικότητες αρχικά πρέπει να εξετάσουμε την ισχύ τις κάθε φασματικής πυκνότητας. Για να τις εντοπίσουμε υπάρχουν δύο γνωστές μέθοδοι το περιοδιόγραμμα και οι κυκλικές αυτοσυσχετίσεις. Το περιοδιόγραμμα ορίζεται ως η τετραγωνική απόσταση του κάθε συντελεστή Fourier $F_k = ||X||^2$, όπου F_k η ισχύς της συχνότητας k. Για να αποφασίσουμε ποιες συχνότητες είναι σημαντικές, θα χρειαστεί να ορίσουμε ένα κατώφλι βάση του οποίου θα αναγνωρίζουμε τις υψηλές συχνότητες. Για να ορίσουμε το κατώφλι, αρχικά παίρνουμε μία τυχαία μεταλλαγμένη ακολουθία της B, έστω B' η οποία δεν θα πρέπει να παρουσιάζει κάποια περιοδικότητα λόγο της τυχαιότητας και καταγράφουμε την μέγιστη ισχύ της. Για να αποκτήσουμε 99% επίπεδο σημαντικότητας, συνεχίζουμε αυτή τη διαδικασία για άλλες 100 τυχαίες μεταλλαγμένες ακολουθίες και καταγράφουμε για κάθε μία από αυτές τη μέγιστη ισχύ της. Η $99_{\text{τη}}$ μεγαλύτερη τιμή από τις 100 που συλλέξαμε θα είναι το κατώφλι.

Η δεύτερη μέθοδος για να εκτιμήσουμε τις επικρατέστερες περιόδους σε μία χρονοσειρά είναι η μέθοδος με τις αυτοσυσχετίσεις, όπου εξετάζεται πόσο παρόμοιες είναι οι τιμές σε μία ακολουθία σε σύγκριση με τις προηγούμενες τιμές τις για διαφορετικά χρονικά περιθώρια.

$$ACF(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} x(\tau) \cdot x\,(n + \tau)$$

Παρόλο που μπορεί να εντοπίσει με μεγαλύτερη ακρίβεια ακόμη και μεγάλες περιόδους, δεν μπορεί να χρησιμοποιηθεί σε αυτοματοποιημένες διαδικασίες διότι πρώτον πρέπει ο χρήστης να ορίσει το επίπεδο σημαντικότητας, δεύτερον ακόμη και όταν ο χρήστης ορίζει το επίπεδο σημαντικότητας, πολλαπλάσια των βασικών περιόδων εμφανίζονται επίσης ως κορυφές που σημαίνει ότι θα πάρουμε λανθασμένα κάποιες περιόδους που πρέπει να τις διαγράψουμε σε μία μετανάλυση. Τρίτον γεγονότα με μικρό εύρος αλλά υψηλή συχνότητα μπορεί να φαίνονται μη σημαντικά.

Η συγκεκριμένη μέθοδος συνδυάζει τις παραπάνω δύο μεθόδους. Αρχικά χρησιμοποιεί το περιοδιόγραμμα για να εξάγει κάποιες περιόδους σαν ενδείξεις ή αλλιώς υποψήφιες περιόδους. Εάν αυτές οι υποψήφιες περίοδοι βρίσκονται σε μία κορυφή στο τότε θεωρείται σαν έγκυρη περίοδος διαφορετικά απορρίπτεται.

Και οι δύο αυτές μέθοδοι υπολογίζονται μετατρέποντας την ακολουθία σε διακριτή κατανομή Fourier, συνεπώς μπορεί να χρησιμοποιηθεί και η παραλλαγή FFT (Fast Transform Fourier) της DFT που σημαίνει ότι ο χρόνος για εκτέλεσης του αλγορίθμου θα μειωθεί από N_2 se $N\ log(N)$ δηλαδή όσο αυξάνεται το N από εκθετική αύξηση θα έχουμε σχεδόν γραμμική.

Παρακάτω μπορούμε να δούμε τα πλεονεκτήματα της συνδυαστικής μεθόδου σε σύγκριση με την κάθε μία ξεχωριστά

Method	Easy to threshold	Accurate short periods	Accurate large periods	Complexity
Periodogram	yes	yes	no	O(NlogN)
Autocorrelation	no	yes	yes	O(NlogN)
Combination	yes	yes	yes	O(NlogN)

2.4.2.3. Εξόρυξη Περιοδικών Συμπεριφορών

Αρχικά ανακτούμε όλα τα σημεία αναφοράς με περίοδο Τ. Συνδυάζοντας τα σημεία αναφοράς που έχουν την ίδια περίοδο θα καταφέρουμε να πάρουμε πιο κατατοπιστικές περιοδικές συμπεριφορές. Για παράδειγμα ένας εργαζόμενος θα μπορούσε να έχει σαν περιοδική συμπεριφορά την εξής 9.00-18.00 στο γραφείο και 20.00-8.00 στο διαμέρισμα του. Η εξής περιοδική αναφορά όπως γίνεται εύκολα αντιληπτό αναφέρεται σε περίοδο μίας μέρας και έχει δύο σημεία αναφοράς το γραφείο και το διαμέρισμα.

Έστω ότι το $O_T = \{o_1, o_2, \ldots, o_d\}$, θα αναφέρεται στα σημεία αναφοράς με περίοδο T και o_0 θα αναφέρεται οποιοδήποτε άλλο σημείο έξω από τα σημεία αναφοράς. Δεδομένου του $LOC = loc_1, loc_2 \ldots loc_n$, θα παράγουμε την αντίστοιχη ακολουθία των κινήσεων $S = s_1, s_2 \ldots s_n$, όπου $s_i = j$ εάν το σημείο loc_i ανήκει στο o_j. Αργότερα η ακολουθία S τμηματοποιείται ξανά σε $m = \left\lfloor \frac{n}{T} \right\rfloor$ τμήματα. Το I^j θα συμβολίζει στη σειρά το j τμήμα και το t_k $(1 \leq k \leq T)$ τη k-στη σχετική χρονοσήμανση σε μία περίοδο. Δηλαδή $I_k^j = i$ σημαίνει πως το αντικείμενο βρίσκεται στο σημείο αναφοράς o_i στη χρονικό σημείο t_k στο τμήμα στη σειρά j. Για να δώσουμε ένα παράδειγμα να γίνει πιο εύκολα αντιληπτό τι σημαίνει, έστω T = 24 ώρες, ένα τμήμα αντιστοιχεί σε μία μέρα και το t_9 συμβολίζει την ώρα 9.00. Οπότε το $I_9^5 = 2$ σημαίνει πως το αντικείμενο βρίσκεται στο σημείο αναφοράς o_2 στις 9.00 την 5η μέρα. Για την μοντελοποιήση της πιθανότητας αυτών των γεγονότων χρησιμοποιήθηκε η κατηγορική κατανομή.

Ορισμός κατηγορικής κατανομής:

Έστω $T = \{t_1, t_2,\ldots,t_T\}$ένα σύνολο από χρονικές σημάνσεις, x_k η τυχαία μεταβλητή της κατηγορικής κατανομής που υποδεικνύει το σημείο αναφοράς που βρίσκεται το αντικείμενο τη στιγμή t_k. Ο $P = [p_1,\ldots,p_T]$ είναι ένας πίνακας κατηγορικής κατανομής και κάθε κολόνα $p_k = [p(x_k = 0), p(x_k = 1), \ldots, p(x_k = d)]^T$ είναι ένα διάνυσμα το οποίο ικανοποιεί την εξίσωση

$$\sum_{i=0}^{d} p(x_k = i) = 1$$

Υποθέτουμε ότι τα I^1, I^2, ..., I^l ακολουθούν την ίδια περιοδική συμπεριφορά. Η πιθανότητα ότι το τμήμα $I = U_{j=1}^{l} Ij$ παράγεται από κάποιον πίνακα κατανομής P

$$P(I|P) = \prod_{IJ \in \chi} \prod_{k=1}^{T} p\,(\chi_k = I_k^j)$$

Σύμφωνα με τον Εκτιμητή Μέγιστης Πιθανοφάνειας, το καλύτερο δυνατό μοντέλο μπορεί να βρεθεί από την λύση της εξίσωσης:

$$\max_{P} \{ L\,(P|I) = log\ P\ (I|P) = \sum_{I^j \in \chi} \sum_{k=1}^{T} p\ (x_{k = I_k^j})\}$$

Όπου η λύση του είναι:

$$P\,(x_k = i) = \frac{\sum_{Ij \in \chi} 1_{I_k^j = i}}{|\chi|}$$

Έστω I ένα σύνολο από περιοχές. Η περιοδική συμπεριφορά όλων των περιοχών στο I, θα δηλώνεται ως $\boldsymbol{H(I)}$ και θα εκφράζει ένα ζευγάρι από $\{T,\boldsymbol{P}\}$, όπου T η περίοδος και $\boldsymbol{P}$ ο πίνακας της κατανομής των πιθανοτήτων.

Για την εξόρυξη περιοδικών συμπεριφορών αφού βρούμε τις περιόδους για κάθε σημείο αναφοράς, θα ξεκινήσουμε τη διαδικασία για την εξόρυξη των περιοδικών συμπεριφορών. Θα θεωρήσουμε τα σημεία αναφοράς με τις ίδιες περιόδους μαζί με σκοπό να πάρουμε πιο συνοπτικές και ενημερωτικές περιοδικές συμπεριφορές. Όμως, αφού μία συμπεριφορά μπορεί να υπάρχει μόνο σε μερικές κινήσεις, θα μπορούσαμε να έχουμε αρκετές περιοδικές συμπεριφορές στην ίδια περίοδο. Για παράδειγμα σε έναν μαθητή μπορούμε να δούμε δύο διαφορετικές περιοδικές συμπεριφορές εάν χρησιμοποιήσουμε για μία περίοδο την μέρα. Η μία περιοδική συμπεριφορά θα είναι για τις μέρες που ο μαθητής έχει σχολείο ενώ θα παρατηρήσουμε μία διαφορετική περιοδική συμπεριφορά τις καλοκαιρινές μέρες. Αυτό σημαίνει ότι δεν μπορούμε να γνωρίζουμε πόσες περιοδικές συμπεριφορές υπάρχουν στις κινήσεις και ποιες μέρες ανήκουν σε ποια περιοδική

συμπεριφορά. Για τον λόγο αυτό χρησιμοποιούμε τη μέθοδο συστάδων, διότι οι μέρες που έχουν ίδια περιοδική συμπεριφορά θα πρέπει να έχουν και παρόμοια μοτίβο χρονικής θέσης.

Θα εφαρμόσουμε τη μέθοδο συστάδων στα ζεύγη *{T,P}* και εφόσον το *T* είναι σταθερό, η απόσταση θα εφαρμοστεί στους πίνακες κατανομών πιθανοτήτων. Μικρή απόσταση μεταξύ δύο *{T,P}* είναι ένα δείγμα πως οι περιοχές του κάθε ζεύγους *{T,P}* είναι πολύ πιθανόν να έχουν παραχθεί από την ίδια περιοδική συμπεριφορά. Ως μονάδα μέτρησης της απόστασης οι συγγραφείς πρότειναν την Kullback-Leibler απόκλιση.

$$KL\,(\boldsymbol{P}\,||\,\boldsymbol{Q}) = \sum_{k=i}^{T} \sum_{i=0}^{d} p\,(x_k = i)\, log \frac{p\,(x_k = i)}{q\,(x_k = i)}$$

Επειδή ο $KL(\boldsymbol{P}\,||\boldsymbol{Q})$ απειρίζεται όταν το $p(xk = i)$ ή το $q(xk = i)$ ισούται με 0, προσθέτουμε στο $p(xk = i)$ και το $q(xk = i)$ μία μεταβλητή u η οποία κατανέμεται ομοιόμορφα σε όλα τα σημεία αναφοράς

$$\mathrm{p}(\mathrm{x_k} = \mathrm{i}) = (1 - \lambda)\mathrm{p}(\mathrm{x_k} = \mathrm{i}) + \lambda \mathrm{u}\,,$$

Όπου λ μία παράμετρος εξομάλυνσης $0 < \lambda < 1$.

Επειδή ο αριθμός των ομάδων δεν μας είναι γνωστός, η μέθοδος συστάδων που προτάθηκε είναι μία ιεραρχική αλγοριθμική μέθοδος όπου ομαδοποιεί τις περιοχές ενώ ταυτόχρονα μπορεί να καθοριστεί και ο ιδανικός αριθμός περιοδικών συμπεριφορών.

2.4.2.4. Ψευδογλώσσα Periodica

Για την εύρεση των περιοδικών δεδομένων προτείνεται ο αλγόριθμος periodica. Στον συγκεκριμένο αλγόριθμο όλη η διαδικασία βασίζεται σε δύο στάδια, όπου σε κάθε στάδιο αναπτύσσεται και από μία υπό διαδικασία. Στο πρώτο στάδιο αναζητούνται όλα τα σημεία αναφοράς με το τρόπο που έχει προαναφερθεί (γραμμή 2) και για κάθε σημείο αναφοράς εντοπίζονται οι αντίστοιχοι περίοδοι με τη συνδυαστική μέθοδο που έχει προταθεί στο paper (On Periodicity Detection and Structural Periodic Similarity) (γραμμή 3-5). Αμέσως μετά για κάθε περίοδο T, δεδομένου του σημείου αναφοράς γίνεται εξόρυξη των αντίστοιχων περιοδικών συμπεριφορών(γραμμή 7-10).

Παρακάτω βλέπουμε τον αλγόριθμο σε μία ψευδογλώσσα.

Algorithm1 Periodica

INPUT: A movement sequence $LOC = loc_1 loc_2 \cdots loc_n$.

OUTPUT: A set of periodic behaviors. ALGORITHM:

1: /* Stage 1: Detect periods (/

2: Find reference spots $O = \{o_1, o_2, \cdots, o_d\}$;

3: **for** each $o_i \in O$ **do**

4: Detect periods in o_i and store the periods in P_i;

5: $P_{set} \leftarrow P_{set} \cup P_i$;

6: /* Stage 2: Mine periodic behaviors /

7: **for** each $T \in P_{set}$ **do**

8: $O_T = \{o_i | T \in P_i\}$;

9: Construct the symbolized sequence S using O_T; 10: Mine periodic behaviors in S.

2.4.2.7. Συμπεράσματα

Απευθύνθηκε ένα σημαντικό πρόβλημα που είναι η εξόρυξη περιοδικών συμπεριφορών και οι δυσκολίες αυτού. Προτάθηκε ένα αλγόριθμος δύο σταδίων. Στο πρώτο στάδιο, εντοπίζονται οι περίοδοι μέσω των σημείων αναφοράς χρησιμοποιώντας τη μέθοδο Fourier και τις αυτοσυσχετίσεις. Στο δεύτερο στάδιο ανακεφαλαιώνονται οι περιοδικές συμπεριφορές χρησιμοποιώντας την ιεραρχική μέθοδο συστάδων. Εμπειρικά φάνηκε ότι η μέθοδος μπορεί να αντιμετωπίσει θόρυβο και περίπλοκες περιπτώσεις. Ενώ η προσέγγιση διορθώνει κάποια σημεία αναφοράς χρησιμοποιώντας μόνο χωρικές πληροφορίες, θα ήταν ενδιαφέρον να αναζητεί δυναμικά τα σημεία αναφοράς συμπεριλαμβανομένου και χρονικές πληροφορίες. Επίσης ένα ακόμη σημαντικό ζήτημα που δεν επιλύθηκε είναι να μπορέσουν να βρεθούν περιοδικά δεδομένα σε δεδομένα με αραιά και ασταθή δείγμα.

2.4.3 PRED [8]

2.4.3.1. Χρήσιμες έννοιες

Έστω Du μία συλλογή από εγγραφές τη χρήστη u, και κάθε εγγραφή di Є Du είναι ένα διάνυσμα {li, ti}, όπου li και ti είναι οι γεωγραφικές συντεταγμένες και η ώρα που έγινε μία δημοσίευση αντίστοιχα. Ένας χρήστης u θα λέμε ότι εμφανίζει περιοδική συμπεριφορά σε μία περιοχή r με περίοδο T, όταν είναι πολύ πιθανόν να επισκεφθεί την περιοχή r κάθε T ώρες. Με τον όρο περιοχές εννοούμε ένα σύνολο από γεωγραφικές συστάδες μέσα στις οποίες παρατηρούνται οι περισσότερες εγγραφές στο D_u. Οπότε δεδομένου του D_u ο στόχος είναι να βρεθεί ένα σύνολο από περιοχές R_u στις οποίες ο χρήστης παρουσιάζει περιοδική συμπεριφορές και η περίοδος T της κάθε περιοχής r Є R_u.

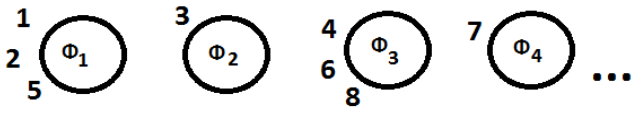
Αρχικά υποθέτουμε ότι κάθε περιοχή έχει μόνο μία περίοδο π.χ μέσα σε μία ημέρα μπορεί κάποιος να βρίσκεται στο σπίτι του να αποχωρήσει και να ξαναγυρίσει πάνω από μία φορές. Σε αυτή τη περίπτωση η περίοδος για τη τοποθεσία σπίτι θα είναι το ελάχιστο κοινό πολλαπλάσιο των επιμέρους περιόδων. Γενικώς οι συγγραφείς από [8] προτείνεται να χρησιμοποιηθούν από κοινού οι γεωγραφικές και χρονικές πληροφορίες που μας διατίθενται. Ο λόγος είναι πως άμα δοκιμάσουμε να ανακαλύψουμε τις περιοχές βασιζόμενοι μόνο σε γεωγραφικά δεδομένα, θα είναι δύσκολο να εκτιμήσουμε τη περίοδο εάν δύο περιοχές είναι σε κοντινή απόσταση αλλά έχουν διαφορετικές περιόδους. Ενώ άμα λάβουμε υπόψιν μας μόνο τα χρονικά δεδομένα είναι δύσκολο να ανιχνεύσουμε όλες τις πιθανές περιόδους, διότι οι περίοδοι παρεμβάλουν η μία με την άλλη αναμειγμένοι ταυτόχρονα και με θόρυβο.

Ο εντοπισμός του κατάλληλου αριθμού των περιοχών είναι ένα κρίσιμο πρόβλημα που καλούμαστε να αντιμετωπίσουμε. Τα περισσότερα μοντέλα υποθέτουν ότι όλοι οι χρήστες έχουν κοινό αριθμό περιοχών ο οποίος μας είναι γνωστός. Στα πραγματικά δεδομένα όμως κάτι τέτοιο δεν ισχύει. Διαφορετικοί χρήστες έχουν και διαφορετικό αριθμό περιοχών. Για παράδειγμα ένας φοιτητής θα μπορούσε να έχει μία μόνο περιοχή τη σχολή του, ενώ ένας εργαζόμενος ο οποίος κάνει μία δουλειάς γραφείου θα μπορούσε να έχει παραπάνω περιοχές όπου εργάζεται, δειπνεί, κάνει τις αγορές του κ.α. Η προτεινόμενη μέθοδος από τους προαναφερθέντες συγγραφείς για τον εντοπισμό του κατάλληλου αριθμού περιοχών χρησιμοποιεί τη μέθοδο Dirichlet Process . Μία γνωστή μεταφορική έννοια της μεθόδου αυτής είναι η CRP (Chinese Restaurant Process) , η οποία είναι μία στοχαστική διαδικασία όπου υποθέτουμε ότι πελάτες διαλέγουν θέσεις σε ένα εστιατόριο με απεριόριστο αριθμό τραπεζιών. Ο πρώτος πελάτης διαλέγει τη θέση του τελείως τυχαία, ο επόμενος πελάτης μπορεί να κάτσει στο ίδιο τραπέζι με τον πρώτο πελάτη η να επιλέξει ένα άλλο τραπέζι τυχαία με πιθανότητα ανάλογης μίας παραμέτρου α που την επιλέγουμε εμείς. Η διαδικασία αυτοί συνεχίζετε έως ότου όλοι οι πελάτες καθίσουν σε κάποιο τραπέζι. Για την ευκολότερη κατανόηση του συγκεκριμένου παραδείγματος μπορούμε να δώσουμε και ένα παράδειγμα. Έστω ότι ο πρώτος πελάτης έχει επιλέξει να καθίσει στο τραπέζι ένα ο δεύτερος πελάτης θα μπορεί να καθίσει στο ίδιο τραπέζι με τον πρώτο ή να καθίσει σε ένα άδειο τραπέζι με πιθανότητες αντίστοιχα: [9]

- Τραπέζι 1: $\frac{1}{(1+\alpha)}$

- Άδειο τραπέζι: $\frac{\alpha}{(1+\alpha)}$

Τώρα έστω ότι οι 8 πελάτες έχουν καθίσει στα τραπέζια με τον εξής τρόπο:



Οι πιθανότητες για το που θα καθίσει ο 9 πελάτης είναι οι εξής:

- Τραπέζι 1: $\frac{3}{(8+\alpha)}$

- Τραπέζι 2: $\frac{1}{(8+\alpha)}$

- Τραπέζι 3: $\frac{3}{(8+\alpha)}$

- Τραπέζι 4: $\frac{1}{(8+\alpha)}$

- Άδειο τραπέζι: $\frac{\alpha}{(8+\alpha)}$

Οπότε γενικεύοντας τους τύπους έχουμε
Έστω N_j ο αριθμός των ανθρώπων στο τραπέζι j. Για τον πελάτη (n+1) έχουμε: [10]

- Τραπέζι j: $\frac{N_j}{(N+\alpha)}$

- Άδειο τραπέζι: $\frac{\alpha}{(N+\alpha)}$

Στη συγκεκριμένη παρομοίωση αντί για πελάτες στην πραγματικότητα έχουμε και από μία εγγραφή . Θα εφαρμοστεί η μέθοδο CRP στις εγγραφές και με αυτό το τρόπο θα γίνει και μία εκτίμηση του αριθμού των περιοχών.

Η περιοχή r θα μοντελοποιείται από μία διμερής Gaussian κατανομή (γεωγραφικό μήκος και πλάτος), με παραμέτρους τον μέσο μ_r και τον πίνακα συσχετίσεων Σ_r. Η πιθανότητα η περιοχή r να παράγει μία τοποθεσία l είναι:

$$P(I \mid r) = N(I/\mu_r, \Sigma_r)$$

Οι περισσότερες μέθοδοι υποθέτουν ότι οι περίοδοι που επισκέπτονται οι χρήστες ένα μέρος είναι γνωστές π.χ. κάθε ένα 24ωρο και χρησιμοποιούν τη Gaussian κατανομή για να μοντελοποιήσουν το χρόνο επίσκεψης κάθε περιοδικού τμήματος. Όμως χρησιμοποιώντας αυτή τη μέθοδο δεν μπορούμε να εκτιμήσουμε τις πραγματικές περιόδους και συνεπώς θα κάνουμε κακές εκτιμήσεις. Η μέθοδος που χρησιμοποιήθηκε προτείνει αντί να μοντελοποιείται ο ακριβής χρόνος επίσκεψης, να μοντελοποιείται το κενό μεταξύ των συνεχόμενων εγγραφών. Ο λόγος είναι διότι εάν υπάρχει ένα μοτίβο περιοδικών επισκέψεων το κενό μεταξύ κάθε ζεύγους συνεχόμενων εγγραφών πρέπει να είναι προσεγγιστικά πολλαπλάσιο της πραγματικής περιόδου. Για παράδειγμα εάν ένας χρήστης επισκέπτεται κάθε μέρα ένα κατάστημα καφέ και καταγράφει τις επισκέψεις του, το κενό μεταξύ των επισκέψεων του θα είναι περίπου 24 ώρες όπως καταλαβαίνουμε. Όμως εάν κάποιες εγγραφές λείπουν τα κενά μεταξύ των επισκέψεων του θα είναι και 48 ώρες ή και 72 ώρες περίπου αναλόγως ποιες μέρες θα παραλείπονται οι εγγραφές. Παρ' όλα αυτά μπορούμε εύκολα να διαπιστώσουμε πως το ελάχιστο κοινό πολλαπλάσιο θα παραμένουν οι 24 ώρες.

Παρακάτω μπορούμε να δούμε και σε εικόνα το συγκεκριμένο παράδειγμα

ID	time	exact time	gap	rmdr.	ct.
d_1	D1 8:30 AM	8.50	–	–	–
d_2	D2 8:15 AM	32.25	23.75	23.75	1
d_3	D3 8:36 AM	56.60	24.35	24.35	1
d_4	D5 8:30 AM	104.50	47.90	23.90	2
d_5	D8 8:42 AM	176.70	72.20	24.20	3

Πιο συγκεκριμένα, ορίστηκαν τα κατάλοιπα $ei_{,j}$ του χρόνου διάρκειας του κενού $t_g(t_i, t_j)$ μεταξύ των εγγραφών d_i και d_j διαιρούμενα από τη περίοδο T ως εξής:

$$e_{i,j} = \begin{cases} mod\ (t_g\ (t_i, t_j), T\ ) \\ mod\ (t_g(t_i, t_j), T) + T \end{cases} \quad mod(t_g(t_i, t_j), T) > T/2$$

Αφού ορίστηκαν τα κατάλοιπα, υποθέτουμε ότι ένας χρήστης u επισκέπτεται μία περιοχή r τις χρονικές στιγμές $t_1, t_2, \ldots, t_h$ με περίοδο T, η πιθανότητα ο χρήστης u να επισκεφθεί τη περιοχή r τη στιγμή t_i είναι:

$$P(t_i \mid r\ ) = N(e_{h,i} \mid v_r, \sigma_r^2)$$

Όπου v_r και σ^2_r είναι ο μέσος και η διακύμανση της κατανομής Gauss.

Βέβαια το v_r δεν μπορεί να χρησιμοποιηθεί σαν περίοδος για τη περιοχή r, διότι εκτιμήθηκε βασισμένο στα χρονικά κατάλοιπα, όμως τα χρονικά κατάλοιπα υπολογίστηκαν βασισμένα στο v_r το οποίο υπολογίστηκε βάση του προηγούμενου δείγματος. Για το λόγο αυτό ορίζεται μία μεταβλητή που ονομάζεται period count (ct) όπου το $c_{i,j}$ του χρονικού κενού $t_g\ (t_i, t_j)$ μεταξύ των εγγραφών d_i και d_j ισούται με:

$$c_{i,j} = \frac{t_g(t_i, t_j) - e_{i,j} + v}{v}$$

Για να καταλάβουμε καλύτερα πως λειτουργεί η συγκεκριμένη μεταβλητή αν ανατρέξουμε πάλι στο παράδειγμα με τον χρήστη που επισκεπτόταν ένα συγκεκριμένο κατάστημα καφέ και υποθέσουμε ότι το v_r` ισούται με 8, θα έχουμε την αρίθμηση 3 δύο φορές (αντιστοιχούν στο χρονικά κενά 23.75 και 24.35) και από μία φορά την αρίθμηση 6 και 9 (47.90 και 72.20) Το ιδανικό v_r θα πρέπει να έχει τη μικρότερη διακύμανση στα κατάλοιπα και η αντίστοιχη αρίθμηση του θα πρέπει να εμφανίζεται περισσότερες φορές. Οπότε επιλέγουμε το v_r με το μικρότερο σκορ:

$$\frac{\sum_{e \in e_r} \frac{(e - v_r)^2}{|e_r| - 1}}{\#c}$$

Όπου $\#c$ ο αριθμός των φορών που η αρίθμηση εμφανίζεται. Στο συγκεκριμένο παράδειγμα τα σκορ για το v_r = 24, 48 και 96 είναι 0.0391, 563.278 και 579.278 αντίστοιχα που σημαίνει ότι το 24 είναι η βέλτιστη τιμή για το v_r.

2.4.3.2. Περιγραφή αλγόριθμου PRED

Η ανακάλυψη περιοχών με περιοδική συμπεριφορά μπορεί να μας ωφελήσει στο να μπορέσουμε να κάνουμε εκτιμήσεις για περιοχές που ενδεχόμενος να επισκεφθεί ένας χρήστης στο μέλλον. Δηλαδή δεδομένου ενός χρήστη u, με καταγραφές D_u και χρόνο t, μπορούμε να υπολογίσουμε ποια περιοχή θα είναι η επόμενη που θα επισκεφθεί ο χρήστης, δίνοντας και από μία πιθανότητα για κάθε περιοχή. Το μοντέλο που χρησιμοποιήθηκε δεν επηρεάζεται εύκολα από θόρυβο, διότι εντοπίζει τις περιοχές που δεν βρίσκονται στις περιοχές που επισκέπτεται συχνά ο χρήστης και τις αποκλείει ως θόρυβο. Για τον εντοπισμό των outliers στα χρονικά δεδομένα τα πράγματα είναι λίγο πιο περίπλοκα αφού το μοντέλο στην αρχή δεν μπορεί να εντοπίσει τα outliers και συμπεριλαμβάνει αυτές τις εγγραφές στην εκτίμηση της περιόδου. Όμως μετά από πολλές εγγραφές η εκτιμώμενη περίοδος αρχίζει να προσεγγίζει την πραγματική περίοδο και το μοντέλο μας αρχίζει να <<καταλαβαίνει>> ποιες εγγραφές πρέπει να αποκλειστούν. π.χ. Ο χρήστης που επισκέπτεται το κατάστημα καφέ κάθε μέρα μία συγκεκριμένη ώρα, έτυχε να το επισκεφθεί κάποια μέρα δύο φορές. Άμα αυτή η επίσκεψη έγινε στην αρχή του μοντέλου το κενό μεταξύ των δύο εγγραφών θα καταμετρηθεί και θα επηρεάσει την εκτιμώμενη περίοδο που ισούται με τη μέση τιμή των διαφορών των καταγεγραμμένων επισκέψεων αφού θα την μειώσει π.χ. από 24 ώρες σε 19. Μετά όμως από

πολλές εγγραφές η εκτιμώμενη περίοδος θα αρχίζει να προσεγγίζει το 24 και το μοντέλο θα αρχίζει να αντιλαμβάνεται ποιες τιμές αντιστοιχούν σε τυχαίες εγγραφές και δεν πρέπει να καταγραφούν.

Συνοπτικά ο αλγόριθμος PRED

- Αρχικά θα πάρει μία εγγραφή του χρήστη η οποία ανήκει στο σύνολο των εγγραφών του
- Θα ελέγξει εάν η εγγραφή αυτή ανήκει σε μία από τις περιοχές που δημιουργήθηκαν βάση του CRP
- Εάν η εγγραφή αυτή ανήκει στις περιοχές τότε θα μοντελοποιήσουμε τα κενά μεταξύ των συνεχόμενων εγγραφών

2.4.3.4. Συμπεράσματα

Μελετήθηκε το καινοτόμο πρόβλημα της εξαγωγής περιοδικών μοτίβων σε μη ολοκληρωμένα και με πολύ θόρυβο δεδομένα από χρήστες μέσων κοινωνικής δικτύωσης και προτάθηκε μία Μπεζυανή μη παραμετρική μέθοδος όπου από κοινού μοντελοποιεί γεωγραφικά και χρονικά δεδομένα. Παρατηρήθηκε ότι η προτεινόμενη μέθοδος σε σύγκριση με παλιότερες μεθόδους είναι ανθεκτική στον θόρυβο μοντελοποιώντας το κενό μεταξύ των συνεχόμενων εγγραφών. Συγκρίθηκε με άλλες μεθόδους και παρατηρήθηκε ότι ήταν αρκετά πιο αποτελεσματική σε όλους τους τομείς σύγκρισης.

ΚΕΦΑΛΑΙΟ 3

3.1. Ανάλυση AIS δεδομένα

Αρχικά αναλύσαμε τα δεδομένα του συστήματος αυτόματης αναγνώρισης (automatic identification system, AIS) που παρέχονται από το ΕΛ.ΚΕ.Θ.Ε και αφορούν μηχανότρατες για την περίοδο Ιούνιος – Σεπτέμβριος 2018. Όπως παρουσιάζεται στην παρακάτω εικόνα τα δεδομένα AIS περιέχουν τον χρόνο (t) στον οποίο έγινε η καταγραφή του σήματος, την ταυτότητα (Maritime Mobile Service Identity, vessel_id) του σκάφους, το γεωγραφικό μήκος (lat) και πλάτος (lon), την επωνυμία του σκάφους (heading), την ταχύτητά του (speed) και την πορεία του (course).

Πίνακας 1 Η διαμόρφωση των AIS δεδομένων στα ημερήσια .csv αρχεία όπως αυτά βρίσκονται αποθηκευμένα στο link που αναφέρεται μέσα στο κυρίως κείμενο.

A	B	C	D	E	F	G
t	vessel_id	lat	lon	heading	speed	course
1527811201000	237444000	21.72718	38.24495		0	53.5
1527811203000	237801000	24.06415	35.49128		0	0
1527811204000	237334000	25.88471	40.84518		0	
1527811205000	237665000	23.61278	37.95722		0	281.5
1527811206000	237698000	23.57441	37.96113		0	95.6
1527811206000	237688000	21.72779	38.24532		0	0
1527811209000	237431000	24.0641	35.49135		0	0
1527811212000	237334000	25.8847	40.84518		0	
1527811212000	237801000	24.06415	35.49127		0	0
1527811212000	237444000	21.72718	38.24494		0	0
1527811215000	237688000	21.72779	38.24532		0	0

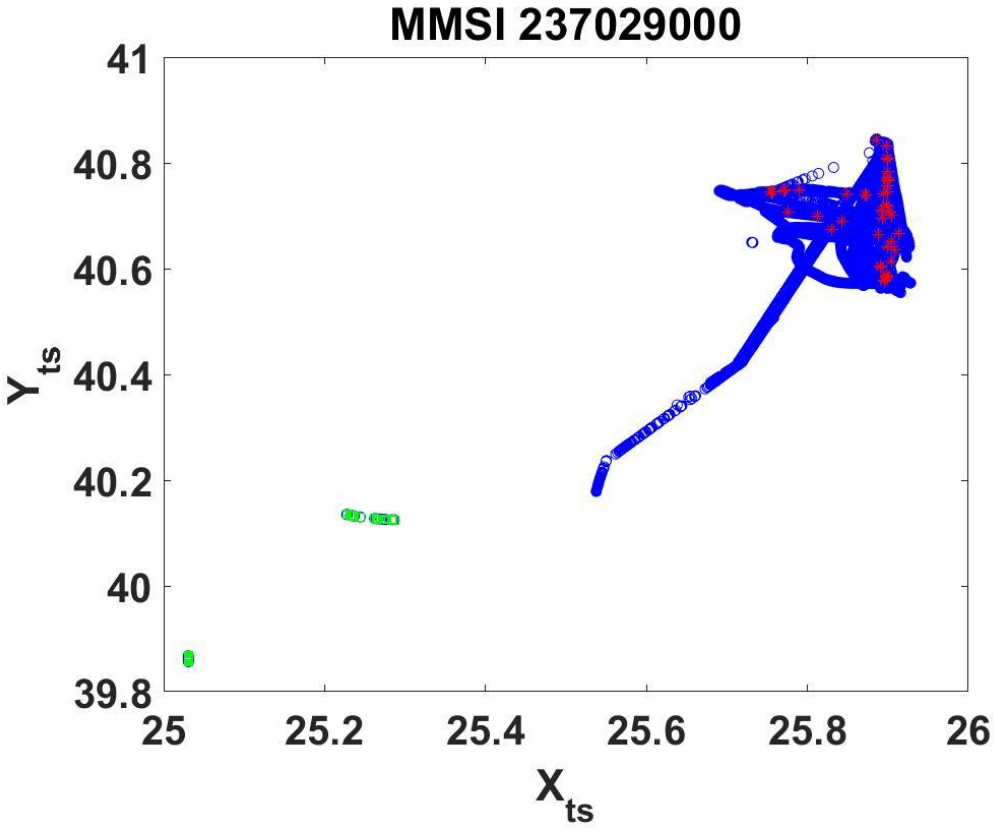
Η ανάλυση των δεδομένων για την εξόρυξη περιοδικών συμπεριφορών κάθε σκάφους έγινε με βάση την μέθοδο και τον υπολογιστικό κώδικα Periodica. Πριν όμως από την ανάλυση των δεδομένων έγινε επεξεργασία τους για την κατάλληλη διαμόρφωσή τους και την ανίχνευση λανθασμένων καταχωρήσεων μέσω κώδικα διατυπωμένο στην προγραμματιστική γλώσσα R καθώς και η παρεμβολή των δεδομένων για την δημιουργία χρονοσειρών με σταθερό χρονικό βήμα όπως απαιτείται από τον κώδικα Periodica.

Αρχικά έγινε η ενοποίηση των δεδομένων κάθε ημέρας σε ένα πίνακα. Εν συνεχεία με βάση ένα αρχείο που περιέχει πληροφορίες για το είδος κάθε σκάφους έγινε ταυτοποίηση του είδους κάθε σκάφους και διατηρήθηκαν τα δεδομένα μόνο για αλιευτικά σκάφη με τον

κωδικό “30”. Στο επόμενο βήμα απαλείφθηκαν τα δεδομένα με μηδενικές ή μη αριθμητικές καταχωρήσεις για το γεωγραφικό πλάτος και μήκος του σκάφους και τέλος έγινε διάκριση και καταχώρηση των δεδομένων για κάθε σκάφος σε ξεχωριστά αρχεία τύπου “.mat” ώστε να μπορούν να εισαχθούν στο προγραμματιστικό περιβάλλον του πακέτου Matlab.

Εν συνεχεία τα αρχεία αυτά εισήχθησαν στο περιβάλλον Matlab όπου έγινε παρεμβολή των χρονικών σημείων ώστε να δημιουργηθούν χρονοσειρές με σταθερό χρονικό βήμα.

Αρχικά στην παρεμβολή των χρονικών σημείων ανιχνεύθηκαν εσφαλμένες καταχωρήσεις εφαρμόζοντας δύο φίλτρα. Το πρώτο φίλτρο βασίζεται στον υπολογισμό της ταχύτητας του σκάφους. Κατόπιν σύγκρισης τη ταχύτητας του σκάφους με την μέση ταχύτητα του σκάφους όπως αυτή υπολογίζεται από τα δεδομένα και συγκεκριμένα αν η ταχύτητα του σκάφους είναι μεγαλύτερη από την μέση ταχύτητα επαυξημένη κατά 10 φορές την τυπική απόκλιση της ταχύτητας τότε τα χρονικά σημεία απορρίπτονται ως εσφαλμένα. Το δεύτερο φίλτρο βασίζεται στην σχετική θέση των σημείων. Συγκεκριμένα αν η θέση κάποιου σημείου είναι απέχει από την μέση θέση των σημείων επαυξημένη κατά 10 φορές την τυπική απόκλιση των θέσεων των σημείων της αρχικής χρονοσειράς, τότε αυτό απορρίπτεται ως εσφαλμένο.



Εικόνα 1 Με μπλε χρώμα παρουσιάζονται τα σημεία (N_o = 57418) της αρχικής χρονοσειράς για το σκάφος με κωδικό MMSI 237029000. Με κόκκινο χρώμα παρουσιάζονται τα σημεία (N_v = 91) που απορρίφθηκαν με βάση το κριτήριο

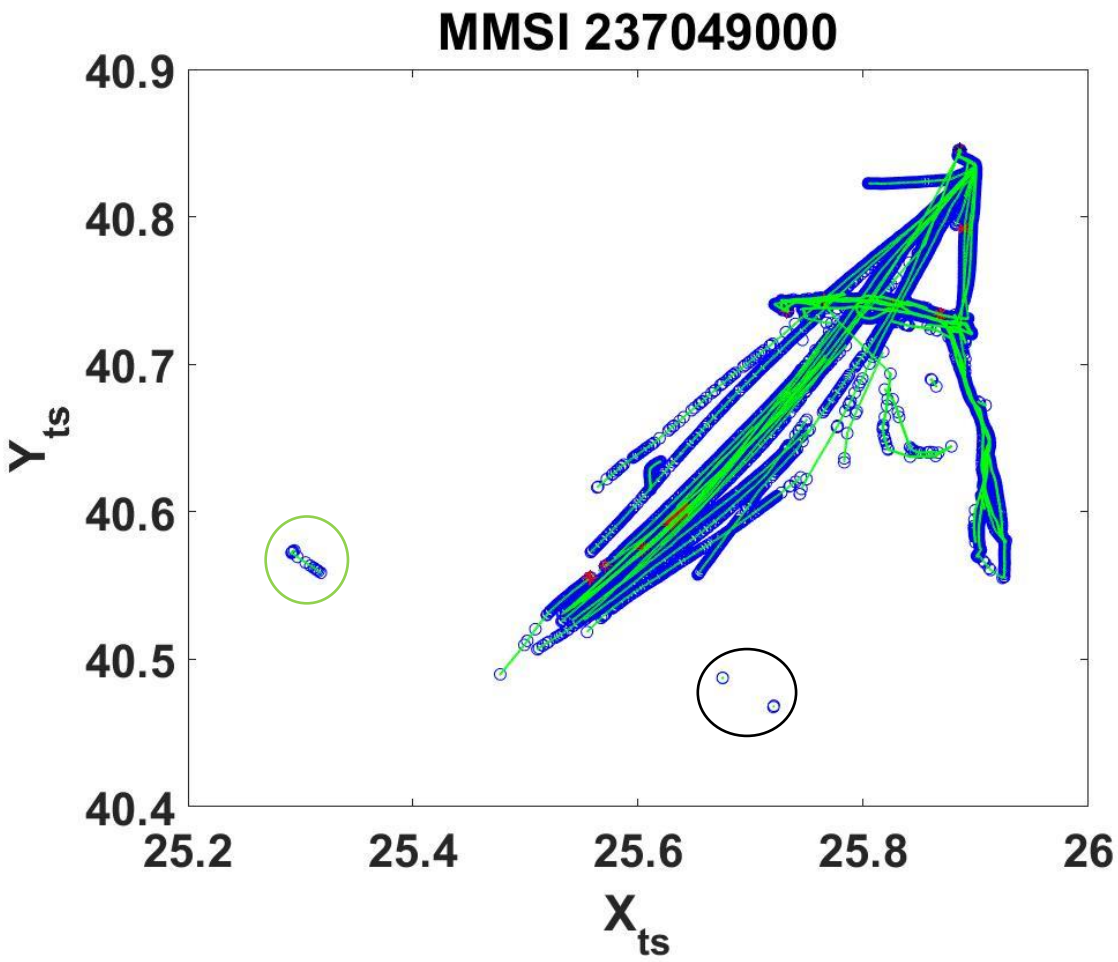
της ταχύτητας ενώ με πράσινο χρώμα τα σημεία (N_p = 73) που απορρίφθηκαν με βάση το κριτήριο της θέσης. Σημειώνουμε ότι Y_{ts} είναι το γεωγραφικό πλάτος ενώ X_{ts} είναι το γεωγραφικό μήκος.

Αφού καθαριστούν τα δεδομένα το επόμενο βήμα είναι η γραμμική παρεμβολή τους βάσει των σχέσεων

$$X_{c,m} = \frac{X_{s,n+1} - X_{s,n}}{T_{s,n+1} - T_{s,n}} \cdot \left(T_{c,m} - T_{s,n}\right) + X_{s,n}$$

$$Y_{c,m} = \frac{Y_{s,n+1} - Y_{s,n}}{T_{s,n+1} - T_{s,n}} \cdot \left(T_{c,m} - T_{s,n}\right) + Y_{s,n}$$

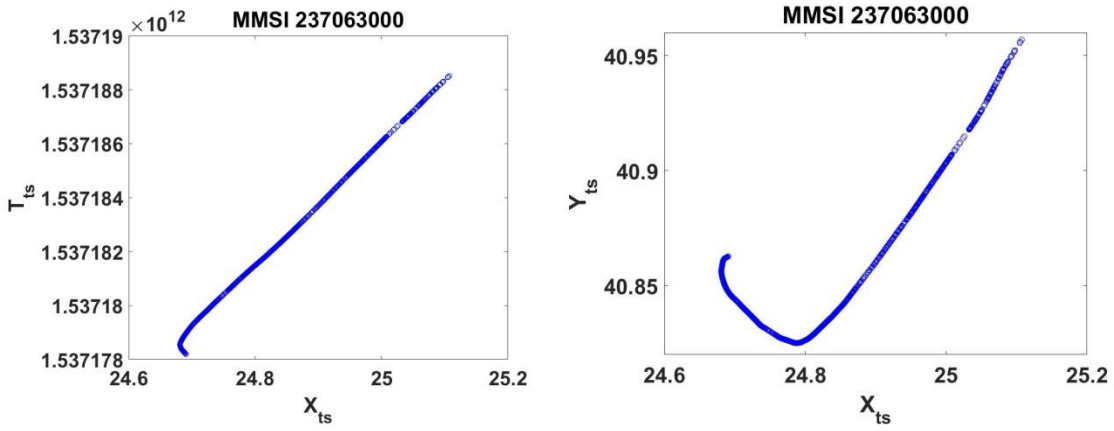
Όπου $(X_{c,m}, Y_{c,m})$ είναι το m-οστό σημείο της χρονοσειράς σταθερού χρονικού βήματος για την χρονική στιγμή $T_{c,m}$, και $(X_{s,n}, Y_{s,n}), (X_{s,n+1}, Y_{s,n+1})$ είναι τα σημεία που αντιστοιχούν στις διαδοχικές χρονικές στιγμές $T_{s,n}$ και $T_{s,n+1}$ της αρχικής χρονοσειράς οι οποίες ορίζουν το χρονικό διάστημα στο οποίο εμπίπτει η χρονική στιγμή $T_{c,m}$, δηλαδή $T_{s,n} < T_{c,m} < T_{s,n+1}$. Το χρονικό βήμα ΔT της χρονοσειράς ορίζεται ως $\Delta T = (T_{s,max} - T_{s,min})/N$ όπου $(T_{s,max} - T_{s,min})$ η χρονική διάρκεια καταγραφής δεδομένων του συγκεκριμένου σκάφους και N ο αριθμός των σημείων της νέας χρονοσειράς. Γενικά επιλέξαμε N=400000 για να διατηρήσουμε ένα μικρό χρονικό βήμα ακόμα και για μεγάλες χρονικές διάρκειες καταγραφής δεδομένων, ενώ όπου χρειάστηκε αυξήσαμε την παράμετρο N για να μην αλλοιωθεί η πυκνότητα των σημείων της αρχικής χρονοσειράς. Ο αριθμός των σημείων της νέας χρονοσειράς είναι πολύ μικρότερος από την παράμετρο N και αυτό γιατί δεν γίνεται παρεμβολή μεταξύ σημείων που απέχουν σημαντική χρονική διάρκεια μεταξύ τους. Ως σημαντική χρονική απόσταση μεταξύ δύο διαδοχικών χρονικών σημείων θεωρείται η $T_{s,n+1} - T_{s,n} > 100\Delta T$, δηλαδή δεν παρεμβάλουμε ποτέ περισσότερα από 100 σημεία μεταξύ δύο διαδοχικών σημείων της αρχικής χρονοσειράς. Έτσι αποφεύγουμε την αλλοίωση της καταγεγραμμένης πορείας του σκάφους. Γενικά προσπαθήσαμε ο αριθμός των σημείων της αρχικής και της νέας χρονοσειράς να έχουν μικρή σχετική διαφορά. Στην εικόνα 1 παρουσιάζουμε χαρακτηριστικό παράδειγμα της παρεμβολής των σημείων της αρχικής χρονοσειράς για το σκάφος MMSI 237029000.



Εικόνα 2 Η πορεία του σκάφους με κωδικό MMSI 237049000. Με μπλε χρώμα είναι τα σημεία της αρχικής χρονοσειράς ενώ με πράσινο τα σημεία της χρονοσειράς ύστερα από την γραμμική παρεμβολή. Παρατηρούμε ότι ακολουθώντας το κριτήριο που έχουμε θέσει τα σημεία που βρίσκονται στον μαύρο και πράσινο κύκλο παρόλο που δεν έχουν απορριφθεί με βάση το φίλτρο θέσης, δεν έχουν παρεμβληθεί με τα υπόλοιπα. Η επιλογή αυτή φαίνεται ρεαλιστική καθώς τα σημεία αυτά απέχουν σημαντικά στο χώρο και συνεπώς πιθανή γραμμική παρεμβολή μεταξύ αυτών των σημείων και των υπολοίπων θα αλλοίωνε της καταγεγραμμένη πορεία του σκάφους. Ο αριθμός των σημείων της αρχικής χρονοσειράς είναι N_i = 12539, ενώ της νέας χρονοσειράς είναι N_f = 13962. Με κόκκινο χρώμα είναι τα σημεία που απορρίφθηκαν με βάση το φίλτρο της ταχύτητας.

Αφού καθαρίσουμε τα δεδομένα μας προχωράμε με την ανάλυση τους με βάση την μέθοδο και τον κώδικα Periodica. Δυστυχώς όπως φαίνεται και από τα ενδεικτικά παραδείγματα που θα παρουσιάσουμε παρακάτω η συμπεριφορά για την πλειονότητα των σκαφών δεν μπορεί να αναλυθεί με βάση την μέθοδο Periodica.

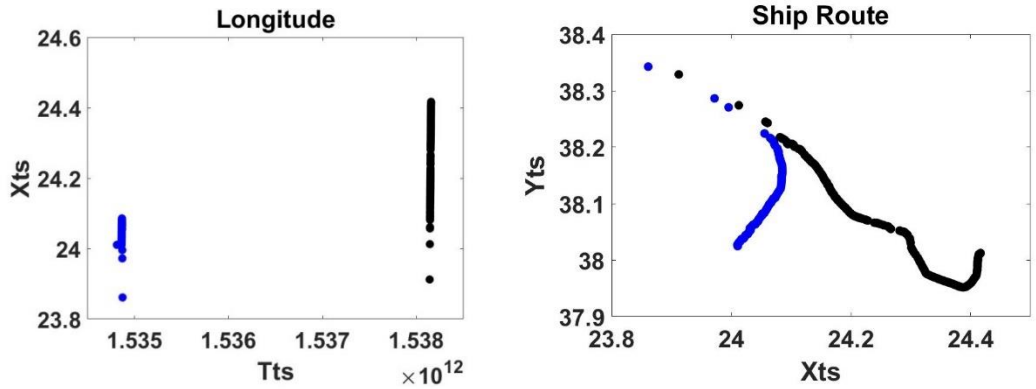
Στην εικόνα 3 παρουσιάζουμε την κίνηση του σκάφους με κωδικό MMSI 237063000. Παρατηρούμε ότι τα δεδομένα αντιστοιχούν σε μία μοναδική κίνηση του σκάφους η οποία είναι συνεχής προς μια κατεύθυνση. Επομένως δεν παρουσιάζεται καμία περιοδικότητα και συνεπώς δεν μπορεί να αναλυθεί με την μέθοδο Periodica.



Εικόνα 3 αριστερά) Το γεωγραφικό μήκος X_{ts} της πορείας του σκάφους με κωδικό MMSI 237063000 συναρτήσει του χρόνου T_{ts} . Παρατηρούμε ότι το γεωγραφικό μήκος του σκάφους αυξάνεται μονότονα (εκτός από κάποια σημεία στην αρχή της κίνησης) με το χρόνο και επομένως η κίνηση δεν παρουσιάζει καμία περιοδικότητα. Δεξιά) Η πορεία του σκάφους.

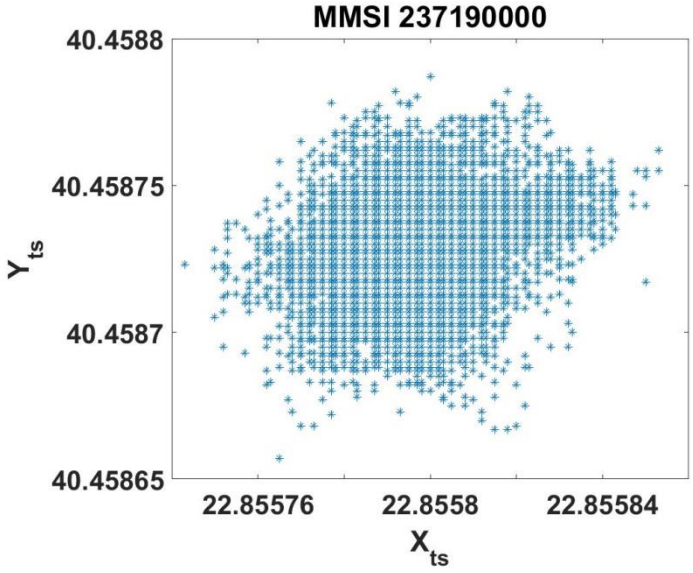
Στην Εικόνα 4 παρουσιάζουμε την αρχική χρονοσειρά για το σκάφος MMSI 237027000. Παρατηρούμε ότι σύμφωνα με τα δεδομένα η κίνηση του συγκεκριμένου σκάφους αποτελείται από δύο ξεχωριστές πορείες που εκτελέστηκαν σε δύο διαφορετικές στιγμές. Από τα σχήματα είναι προφανές ότι δεν μπορούμε να ορίσουμε κανένα σημείο αναφοράς (reference spot) που είναι και το αρχικό βήμα της μεθόδου Periodica. Επίσης είναι προφανές ότι στα συγκεκριμένα δεδομένα δεν εμφανίζεται καμία περιοδική συμπεριφορά αφού οι θέσεις του σκάφους δεν επαναλαμβάνονται. Παρατηρούμε δηλαδή ότι και στην περίπτωση που τα δεδομένα αφορούν περισσότερες από μία κινήσεις δεν σημαίνει απαραίτητα ότι υπάρχει κάποια περιοδικότητα στην κίνηση του σκάφους επομένως η ανάλυση με την μέθοδο

Periodica είναι ακατάλληλη.



Εικόνα 4 αριστερά) Το γεωγραφικό μήκος του σκάφους MMSI 237027000 συναρτήσει της χρονικής μεταβλητής. Παρατηρούμε την πραγματοποίηση δύο ξεχωριστών κινήσεων που χωρίζονται από μεγάλη χρονική διάρκεια. δεξιά) Η πορεία του σκάφους όπως προκύπτει από τα αρχεία AIS. Οι δύο κινήσεις που πραγματοποιεί το σκάφος διακρίνονται με μπλε και μαύρο χρώμα αντίστοιχα.

Στην Εικόνα 4 παρουσιάζουμε μια άλλη περίπτωση, το σκάφος με κωδικό MMSI 237190000, το οποίο όπως φαίνεται από τα δεδομένα είναι στάσιμο. Αυτό συνεπάγεται καθώς το γεωγραφικό πλάτος του σκάφους μεταβάλλεται μόνο κατά 0.00015 μοίρες που αντιστοιχούν σε απόσταση λιγότερο από 15μ, αν εύλογα θεωρήσουμε ότι μια μοίρα γεωγραφικού πλάτους αντιστοιχεί περίπου σε 111χιλιομετρα, καθώς το μήκος μια μοίρας γεωγραφικού πλάτους δίνεται από τον τύπο $\frac{\pi}{180^0} \cdot R_{\gamma\eta\varsigma} \cong 111km$ όπου $R_{\gamma\eta\varsigma} = 6371km$ η ακτίνα της γης. Αντίστοιχα το γεωγραφικό μήκος του σκάφους μεταβάλλεται κατά περίπου 0.0001 μοίρες που αντιστοιχεί σε περίπου 9μ καθώς μία μοίρα γεωγραφικού μήκους αντιστοιχεί σε 85km, σύμφωνα με το τύπο $\frac{\pi}{180^0} \cdot R_{\gamma\eta\varsigma} \cdot cos\varphi$, όπου φ=40^o το γεωγραφικό πλάτος στην συγκεκριμένη περίπτωση. Επομένως βλέπουμε ότι η κίνηση του σκάφους περιορίζεται σε ένα χώρο 9μ x 15μ που κατά πάσα πιθανότητα οφείλεται στον κυματισμό της θάλασσας και στην ακρίβεια καταγραφής των δεδομένων.



Εικόνα 5 Τα δεδομένα για το σκάφος με κωδικό MMSI 237190000. Παρατηρούμε ότι το σκάφος παραμένει ουσιαστικά στο ίδιο σημείο οπότε δεν έχει κανένα νόημα η ανάλυση των δεδομένων αυτών με τον κώδικα Periodica.

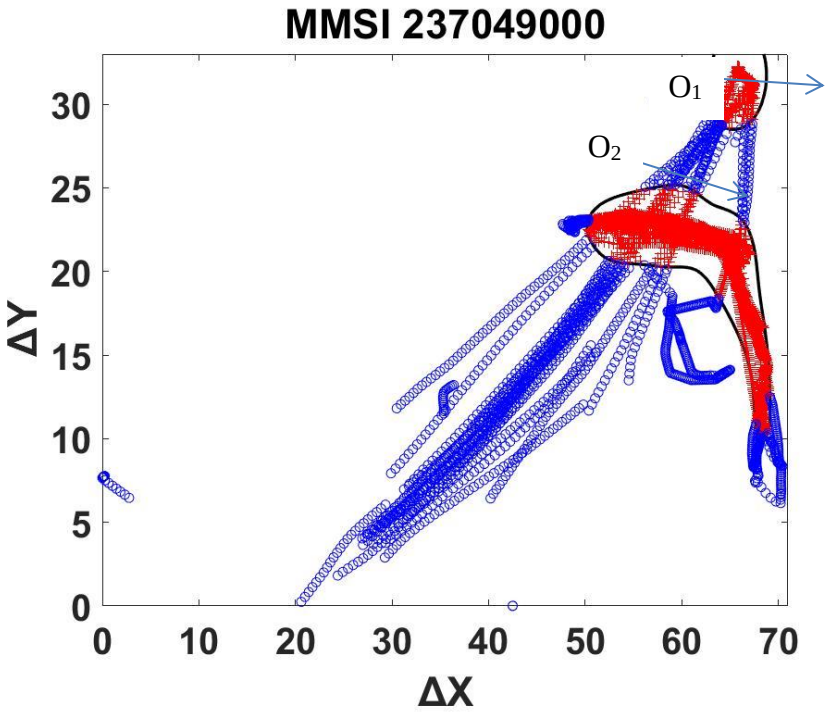
Επίσης δεν εξετάστηκαν τα δεδομένα για σκάφη στα οποία η αρχική χρονοσειρά αποτελούνταν από λιγότερα από 100 σημεία θεωρώντας ότι είναι αδύνατη η εύρεση οποιασδήποτε περιοδικής συμπεριφοράς σε τόσο μικρές χρονοσειρές.

Παρακάτω παρουσιάζουμε ένα ενδεικτικό παράδειγμα για το οποίο η μέθοδος Periodica δίνει κάποια ικανοποιητικά αποτελέσματα και εν συνεχεία παρουσιάζουμε συγκεντρωτικά τα αποτελέσματα που προέκυψαν για όλα τα σκάφη για τα οποία ο κώδικας Periodica έχει κάποιο νόημα.

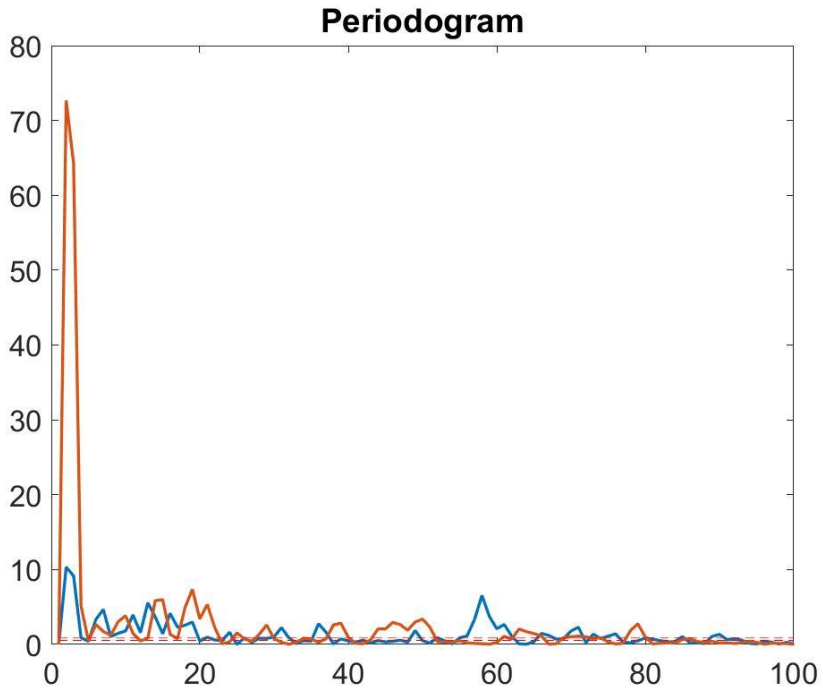
Το σκάφος που εξετάζουμε στην παρούσα παράγραφο έχει τον κωδικό MMSI 237049000 για το οποίο παρουσιάσαμε αποτελέσματα στην αρχή του κεφαλαίου. Εικόνα 6 παρουσιάζονται οι 2 περιοχές αναφοράς (refrence spots) που προέκυψαν θεωρώντας ως όριο

την top-10% πυκνότητα σημείων όπως προέκυψε για κάθε κελί του πλέγματος. Σημειώνουμε ότι στο συγκεκριμένο γράφημα αντί για τις γεωγραφικές συντεταγμένες παρουσιάζονται τα σημεία της χρονοσειράς με βάση την απόσταση τους ΔΧ και ΔΥ από την μικρότερη τιμή του γεωγραφικού πλάτους και μήκους αντίστοιχα εκφρασμένες σε km. Πιστεύουμε ότι έτσι έχουμε μια καλύτερη εικόνα της κίνησης του σκάφους. Το πλέγμα στο οποίο υπολογίσαμε την πυκνότητα σημείων της χρονοσειράς έχει ανάλυση 0.5km X 0.5km.

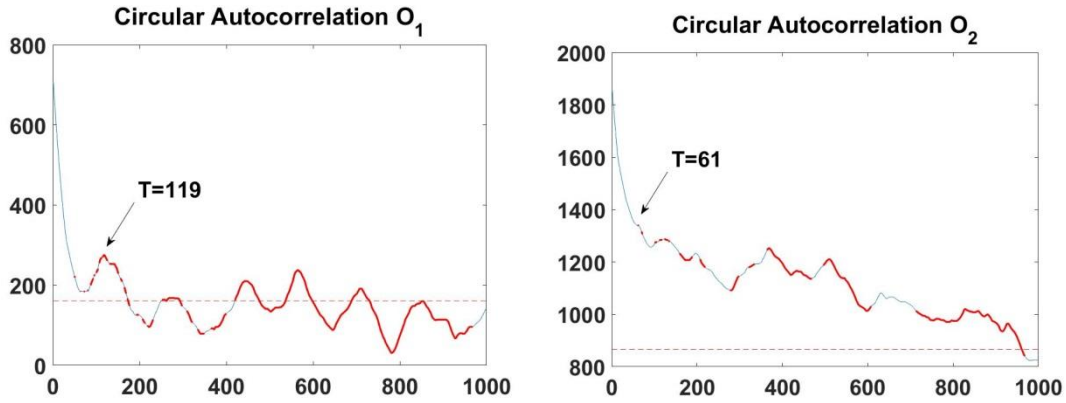
Εφόσον προκύπτουν δύο περιοχές αναφοράς αναζητούμε τουλάχιστον δυο περιόδους, μια για κάθε περιοχή αναφοράς. Από το περιοδιόγραμμα που παρουσιάζεται στην Εικόνα 7 παρατηρούμε ότι πολλές συχνότητες εμφανίζονται με ένταση πάνω από το όριο που τίθεται με επίπεδο εμπιστοσύνης 99%.



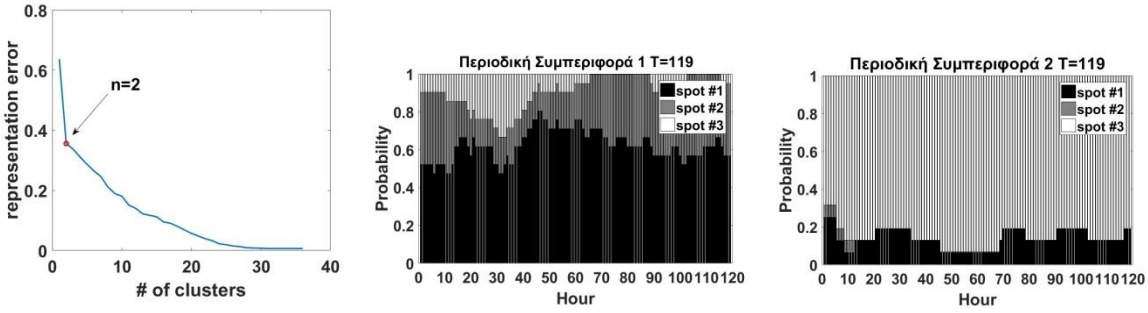
Εικόνα 6 Οι 2 περιοχές αναφοράς (reference spots) O_1 και O_2 όπως ορίζονται από τις μαύρες γραμμές. Με κόκκινο παρουσιάζονται τα σημεία της χρονοσειράς, η οποία προέκυψε ύστερα από την παρεμβολή, τα οποία βρίσκονται μέσα στις περιοχές αναφοράς.



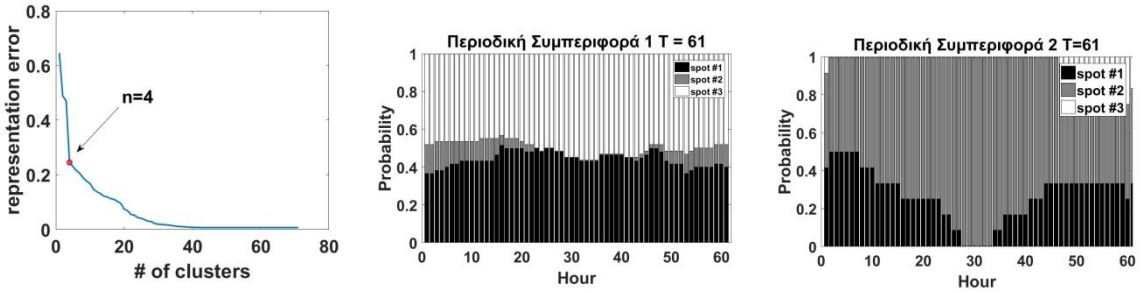
Εικόνα 7 Το περιοδιόγραμμα που προκύπτει για τις δύο περιοχές αναφοράς που παρουσιάζονται στην Εικόνα 6. Παρατηρούμε ότι πολλές συχνότητες είναι πάνω από την τιμή του ορίου το οποίο είναι διαφορετικό για τις δύο περιοχές.



Εικόνα 8 Τα διαγράμματα κυκλικής αυτοσισχέτισης για τις δύο περιοχές αναφοράς. Οι περίοδοι που προκύπτουν είναι $T_{O_1} = 119$ και $T_{O_2} = 61$ για τις περιοχές O_1 και O_2 αντίστοιχα.



Εικόνα 9 Σύμφωνα με την ανάλυση των δεδομένων με τον κώδικα Periodica οι πραγματικές περιοδικές συμπεριφορές για την περίοδο T = 119 που σχετίζονται με την πρώτη περιοχή αναφοράς είναι 2. Παρουσιάζονται δύο από τις πραγματικές συμπεριφορές.



Εικόνα 10 Οι πραγματικές περιοδικές συμπεριφορές για την περίοδο T = 61 που σχετίζονται με την δεύτερη περιοχή αναφοράς είναι 4. Παρουσιάζονται δύο από τις πραγματικές συμπεριφορές.

Κωδικός MMSI σκάφους	Περιοχές Αναφοράς	Περίοδοι (πραγματικές περιοδικές συμπεριφορές)
237049000	2	T_1=119 (2) , T_2=61 (4)
237058000	2	T_1=191 (2) , T_2=164 (2)
237119000	1	T_1=242 (2)
237168000	1	T_1=50 (3)
237287000	1	T_1=122 (2)
237302000	1	T_1=92 (2)
237343000	3	T_1=175 (2), T_2=169 (2), T_3=88 (2)
237562000	4	T_1=179 (2), T_2=841 (2), T_3=58 (2), T_4=88 (2)
237567000	1	T_1=184 (4), T_2=149 (5),
237589000	3	T_1=220 (2), T_2=183 (2), T_3=108 (2)
237623000	4	T_1=363 (2), T_2=248 (2), T_3=188 (3), T_4=143 (2)
237638000	1	T_1=85 (3)
237687000	2	T_1=239 (2), T_2=138 (2)
237698000	3	T_1=775 (2), T_2=185 (2), T_3=153 (3)
237802000	3	T_1=428 (3), T_2=277 (2), T_3=188 (2)
237805000	2	T_1=985 (3), T_2=157 (2)
237920000	1	T_1=95 (2)
237930000	2	T_1=117 (2), T_2=785 (2)
239013000	3	T_1=358 (3), T_2=221 (3), T_3=143 (3)

239383000	1	T_1=195 (2)
240102000	1	T_1=90 (2)
240112000	2	T_1=91 (2), T_2=91 (2)
240131000	1	T_1=693 (4)
240238000	2	T_1=298 (4), T_2=49 (2)
240250000	1	T_1=126 (2)
240259000	1	T_1=227 (6)
240427000	1	T_1=121 (2)
240430000	2	T_1=351 (5), T_2=196 (2)
241045000	3	T_1=337 (4), T_2=94 (3), T_3=87 (3)
241065000	5	T_1=149 (2), T_2=151 (2), T_3=189 (2), T_3=211 (2), T_3=266 (2)

Παρατηρούμε ότι τα μισά σχεδόν σκάφη (43%) σχετίζονται με μια περιοχή αναφοράς η οποία κατά πάσα πιθανότητα είναι το λιμάνι που αγκυροβολούν. Συνεπώς, σύμφωνα με τα δεδομένα, τα σκάφη αυτά δεν σχετίζονται με κάποια συγκεκριμένη περιοχή ψαρέματος. Αντίθετα το 27% και 20% των σκαφών παρουσιάζει δύο και τρεις περιοχές αναφοράς αντίστοιχα. Ενδεχομένως οι περιοχές αυτές να είναι είτε δύο ή τρία λιμάνια όπου αγκυροβολούν, είτε ένα λιμάνι και μια ή δύο συγκεκριμένες περιοχές ψαρέματος. Τέλος, μόνο 3 σκάφη παρουσιάζουν πάνω από τρεις περιοχές αναφοράς γεγονός το οποίο μπορεί να οφείλεται στο μεγάλο αριθμό δεδομένων για τα σκάφη αυτά, MMSI 237562000 **75089** (4) MMSI 237623000 **106722** (4) και MMSI 241065000 **203799** (5), και συνεπώς με παραπάνω από μια περιοχές ψαρέματος ανάλογα με την περίοδο συλλογής των δεδομένων.

Αριθμός Περιοχών Αναφοράς	1	2	3	4	5
Αριθμός Σκαφών	13	8	6	2	1
Ποσοστό Σκαφών	43%	27%	20%	7%	3%

Στον παρακάτω πίνακα παρατηρούμε επίσης μεγάλο εύρος και τυπική απόκλιση στις περιόδους των σκαφών. Αυτό καταδεικνύει την διαφορετική συμπεριφορά μεταξύ των κινήσεων του κάθε σκάφους. Πάντως σημειώνουμε ότι η πλειονότητά (69%) των περιόδων κίνησης εμφανίζουν 2 περιοδικές συμπεριφορές.

	Ελάχιστη	Μέγιστη	Μέση Τιμή	Τυπική Απόκλιση
Περίοδοι	49	985	230	199
Περιοδικές Συμπεριφορές	2	6	2.5	0.9

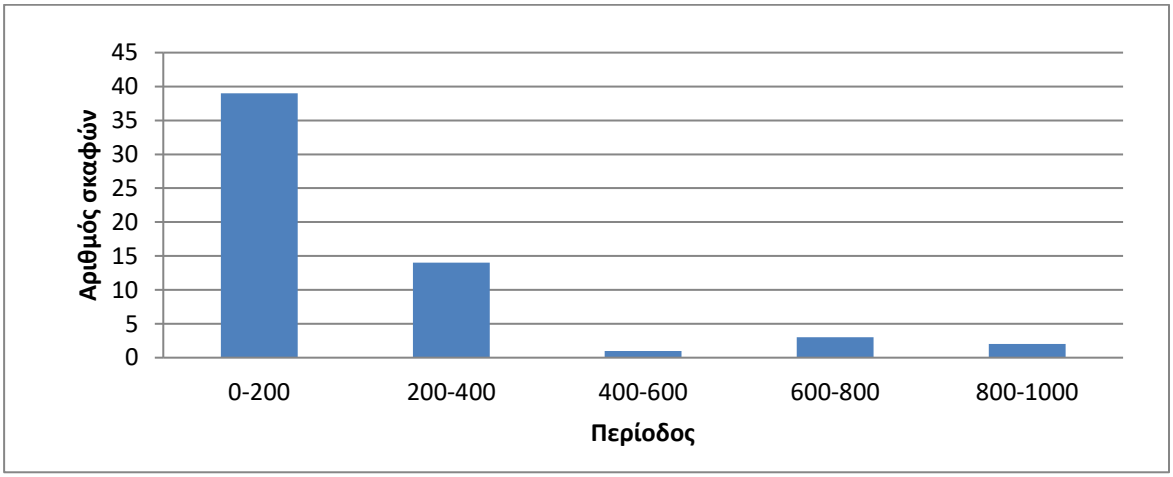
Πίνακας 2 Στατιστικά στοιχεία για τις περιόδους και τις περιοδικές συμπεριφορές των σκαφών.

Περίοδοι	0-200	200-400	400-600	600-800	800-1000
Αριθμός Σκαφών	39	14	1	3	2
% Σκαφών	66%	24%	2%	5%	3%

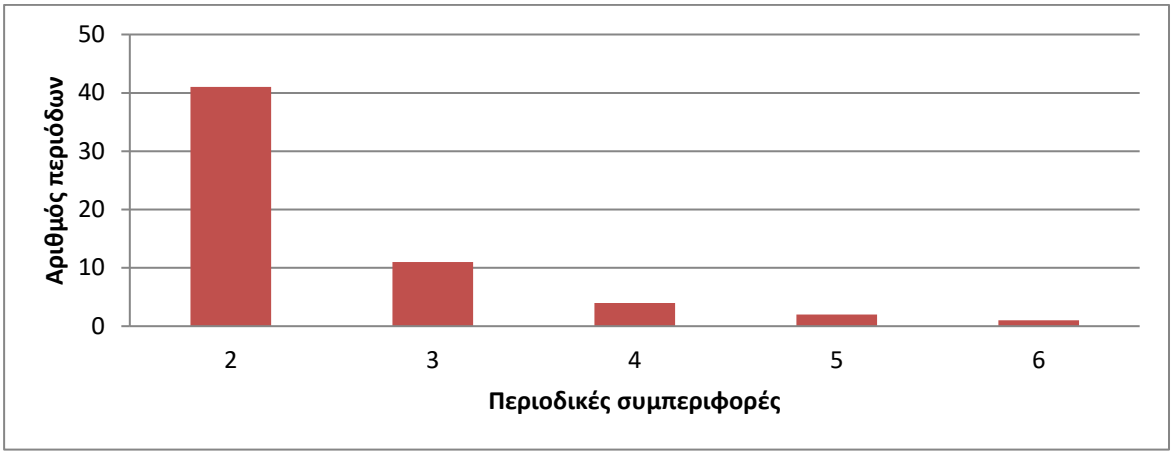
Πίνακας 3 Κατανομή σκαφών σε σχέση με την περίοδο που παρατηρείται στην κίνηση τους.

Περιοδικές Συμπεριφορές	2	3	4	5	6
Αριθμός Σκαφών	41	11	4	2	1
% Σκαφών	69%	19%	7%	3%	2%

Πίνακας 4 Κατανομή των περιόδων κίνησης των σκαφών σε σχέση με τον αριθμό περιοδικών συμπεριφορών που εμφανίζουν.



Εικόνα 11 Κατανομή σκαφών με βάση την περίοδο κίνησης



Εικόνα 12 Κατανομή περιόδων με βάση τον αριθμό περιοδικών συμπεριφορών

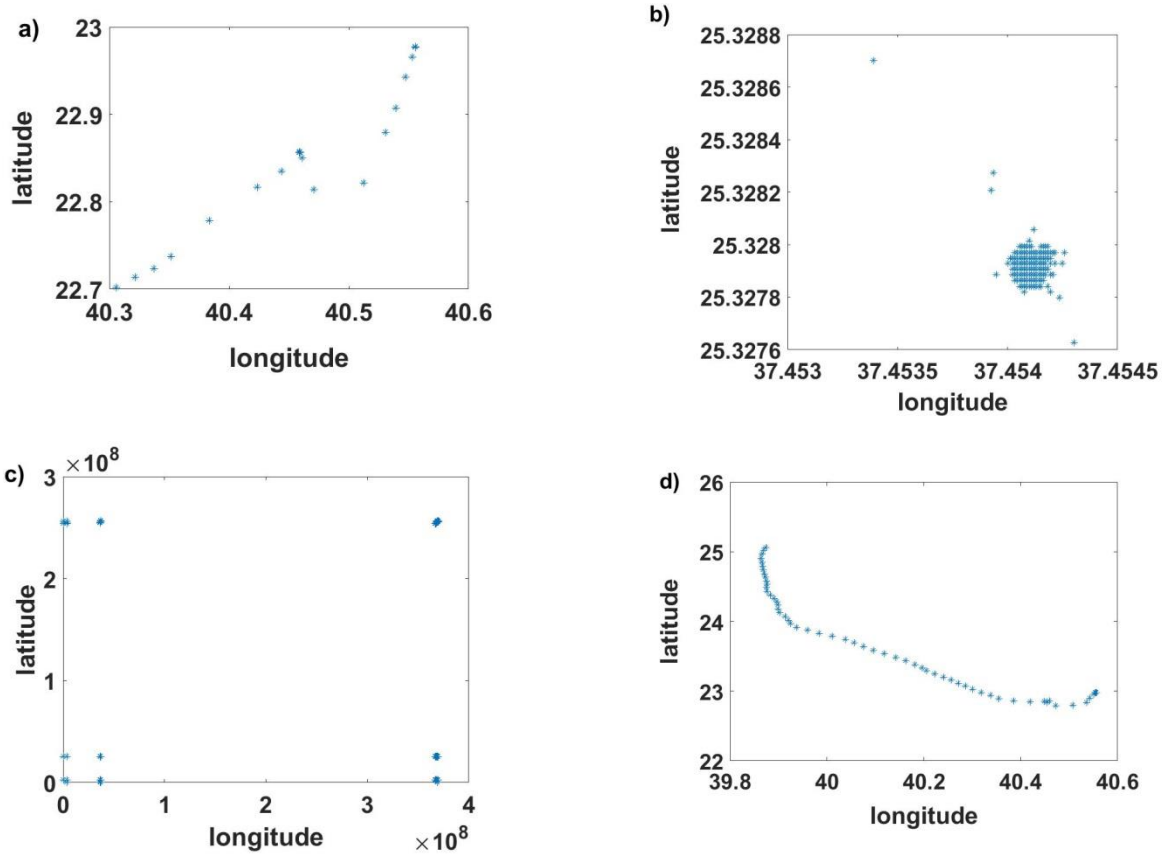
3.2. Ανάλυση VMS δεδομένα

Στη συνέχεια αναλύουμε τα δεδομένα vms. Τα δεδομένα αυτά αναφέρονται στην περίοδο Ιούνιος – Σεπτέμβριος 2018.

	A	B	C	D	E	F	G	H	I
1	objectid	vesselid	lat	lon	speed	heading	posdate	recdate	satellitename
2	20886736	159ca3d053099e68ff1733600c392d2ed34a7f137879ad9f709ebfd433a46084	39.9453	23.528	0	216	26/6/2018 6:02	26/6/2018 6:03	IRIDIUM-1
3	20886737	960b25288cc0b843bda12079283a8741d36a8aac849e68bc8aaa5c205e3f44cc	40.1966	23.3243	0	81	26/6/2018 14:34	26/6/2018 14:35	IRIDIUM-1
4	20886738	6b74d5626d6df14df61a40ec19d071a97c527199374380c4d3753d8ff6cc4243	37.4157	23.1272	0	190	26/6/2018 14:35	26/6/2018 14:35	IRIDIUM-1
5	20886739	f81a290bf70b4437dba8d8341b8d18b68c7202b4e9a67132c07d9abfe93b2098	37.4165	23.1277	0	279	26/6/2018 14:36	26/6/2018 14:36	IRIDIUM-1
6	20886740	c120378bb2d75ffe360101a56495daf4a0400de6045aa55d47b3ec612dd93076	38.6647	20.7842	0	213	26/6/2018 14:36	26/6/2018 14:37	IRIDIUM-1
7	20886741	08e0bf725ab73ab86fcf5ca6f44c974296db41af76bedb641719ed0e9fc76117	38.2455	21.7282	0	173	26/6/2018 17:41	26/6/2018 17:41	IRIDIUM-1
8	20886742	acbf4026434b2e3ee94165be700ada837b117491e1c29f219d761bbe1ac16c06	40.849	24.3138	0	29	26/6/2018 17:41	26/6/2018 17:41	IRIDIUM-1
9	20886743	61d6125726bc2d9f26a71e3f4345129b19a3df619e76b20a965322c8203a6bb2	39.008	26.1703	0.1	354	26/6/2018 17:41	26/6/2018 17:41	IRIDIUM-1
10	20886744	fbe86193e2a8e2fd926d8e57310573de584f20adb5d5eb3271926cf58a6c0ddb	36.7366	26.9722	0	56	26/6/2018 17:41	26/6/2018 17:42	IRIDIUM-1
11	20886745	900b983cc259ba72cfd9f34ecc670fd03b6450cf6ed6915cf80dfa7878113df7	37.2795	23.0982	0.2	69	26/6/2018 17:42	26/6/2018 17:42	IRIDIUM-1
12	20886746	926bf14d1bfaedad5ba4d82b368d836faf379f0ed10654d413807d301aa0a15b	40.2603	22.5948	0	21	26/6/2018 17:42	26/6/2018 17:42	IRIDIUM-1
13	20886747	de25373918258d8ab4b3479ccc53e2f559bff268afc7eeb3f0a49e4ffd280aaf	36.9895	24.6742	0	47	26/6/2018 17:39	26/6/2018 17:42	IRIDIUM-1
14	20886748	20a196edf10d948e6b5951ddad42ffbd75ee60e5ffd93f9e5a36b610dd140c92	37.1424	24.5162	0	271	26/6/2018 20:13	26/6/2018 20:13	IRIDIUM-1
15	20886749	1ff223e59212c359ebc626e35dc06ee3d78dcb9aaa7f2d475b82f9d4275833d3	36.9512	26.9883	0	165	26/6/2018 20:14	26/6/2018 20:15	IRIDIUM-1

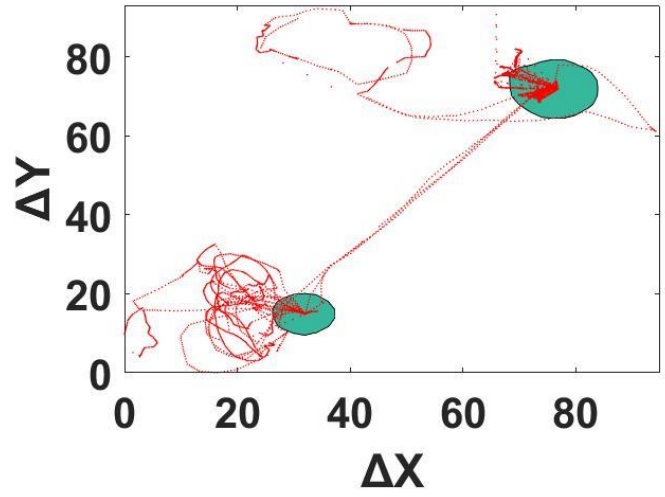
Εικόνα 13 Η δομή του αρχείου vms_dynamic_gr_jun_sep_2018.csv με τα vms δεδομένα.

Όπως και με τα ais αρχεία, έτσι και το συγκεκριμένο το επεξεργαζόμαστε μέσω κώδικα R για την ανίχνευση σημείων που δεν περιέχουν κατάλληλες τιμές για τα στοιχεία “lat” και “lon” για τις συντεταγμένες δηλαδή. Μετέπειτα αναγνωρίζουμε τα σκάφη για τα οποία περιέχονται δεδομένα μέσω της στήλης “vesselid” και καταγράφουμε τα δεδομένα “vesselid”, “lat”, “lot” και “recdate” για κάθε σκάφος σε ξεχωριστό αρχείο μορφής .mat ώστε να μπορούν να εισαχθούν σε περιβάλλον matlab. Με την μέθοδο αυτή αναγνωρίσθηκαν 621 διαφορετικά σκάφη. Δυστυχώς όμως, όπως παρουσιάζουμε παρακάτω για πολλά από αυτά τα σκάφη τα δεδομένα είναι εσφαλμένα, είτε αναφέρονται σε στάσιμα σκάφη, είτε είναι ακατάλληλα για επεξεργασία μέσω του κώδικα Periodica.

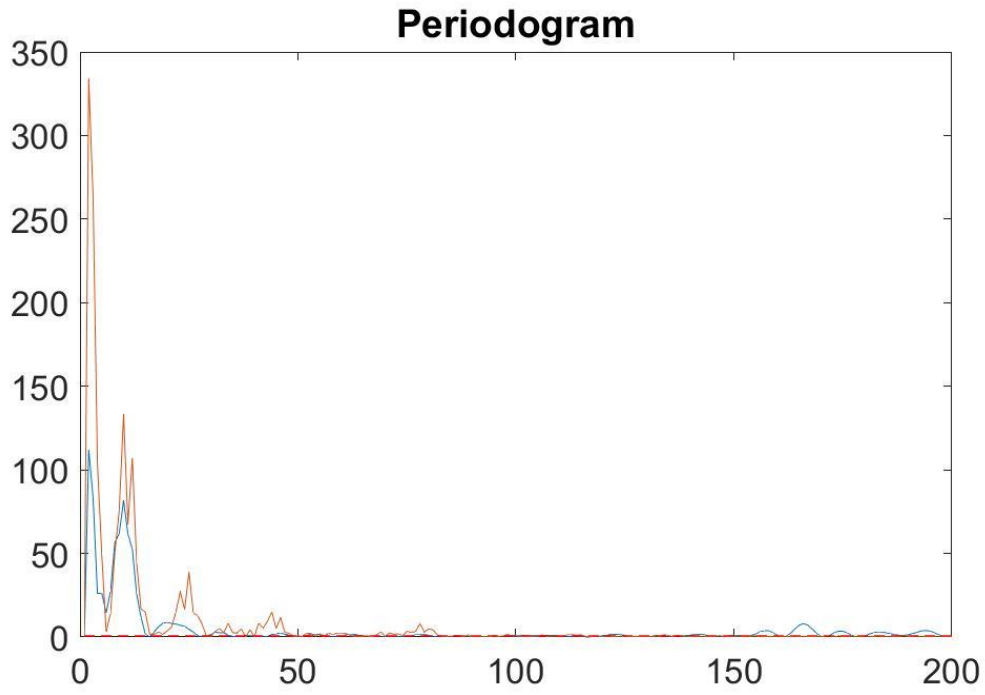


Εικόνα 14 Οι 4 περιπτώσεις δεδομένων ακατάλληλων για περαιτέρω ανάλυση με την μέθοδο Periodica. a) Μικρός αριθμός σημείων, β) στάσιμο σκάφος, c) ακατάλληλα δεδομένα και d) μη περιοδική κίνηση σκάφους.

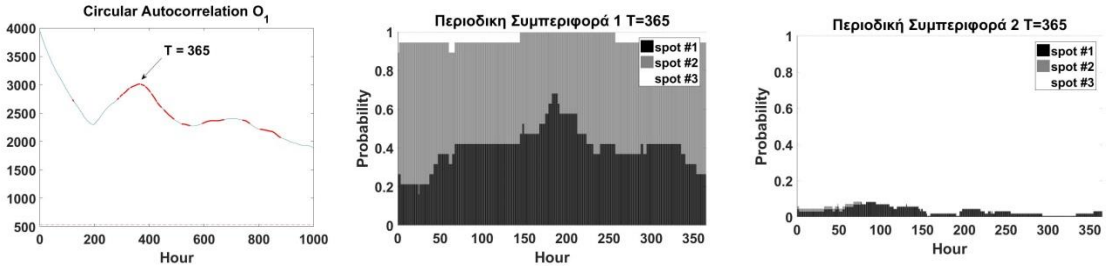
Παρακάτω παρουσιάζουμε αναλυτικά τα αποτελέσματα για μια από τις 67 περιπτώσεις για τις οποίες τα δεδομένα είναι κατάλληλη για επεξεργασία με την μέθοδο Periodica.



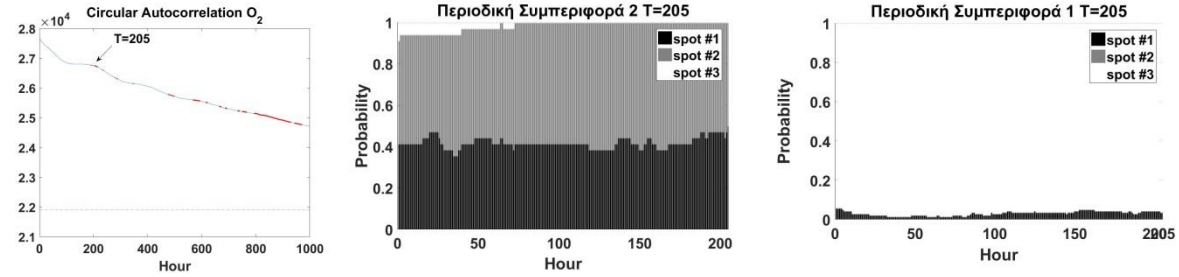
Εικόνα 15 Οι περιοχές αναφοράς. ΔX και ΔY είναι οι αποστάσεις σε km από το σημείο αναφοράς (0,0) που αντιστοιχεί στο ελάχιστο γεωγραφικό μήκος και πλάτος της πορείας του σκάφους.



Εικόνα 16 Το περιοδιόγραμμα της πορείας του σκάφους.



Εικόνα 17 Το διάγραμμα κυκλικής αυτοσυσχέτισης για την περίοδο T=365 ώρες και οι δύο περιοδικές συμπεριφορές της συγκεκριμένης περιόδου.



Εικόνα 18 Το διάγραμμα κυκλικής αυτοσυσχέτισης για την περίοδο T=205 ώρες και οι δύο περιοδικές συμπεριφορές της συγκεκριμένης περιόδου.

Παρατηρούμε ότι και για τις δύο περιόδους η δεύτερη περιοδική συμπεριφορά εμπλέκει μόνο μια περιοχή αναφοράς.

Εν συνεχεία παρουσιάζουμε ένα συγκεντρωτικό πίνακα με τα σημεία αναφοράς, τις περιόδους και τις περιοδικές συμπεριφορές για τα 67 σκάφη που αναλύθηκαν με την μέθοδο Periodica.

A/A	Περιοχές Αναφοράς	Περίοδοι (πραγματικές περιοδικές συμπεριφορές)
1 (9)	1	T_1=421 (17)
2 (10)	3	T_1=476 (2), T_2=421 (2), T_3=287 (2)
3 (12)	3	T_1=380 (2), T_2=248 (2), T_3=106 (2)
4 (18)	3	T_1=421 (17), T_2=123 (3), T_3=70 (2)
5 (21)	1	T_1=191 (17)
6 (23)	1	T_1=421 (2)
7 (24)	1	T_1=101 (2)
8 (27)	4	T_1=421 (17), T_2=281 (3), T_3=170 (2), T_4=148 (2)
9 (28)	3	T_1=415 (2), T_2=410 (2), T_3=166 (2)
10 (48)	3	T_1=434 (5), T_2=141 (7), T_3=129 (5)
11 (51)	1	T_1=105 (2)
12 (55)	2	T_1=419 (4), T_2=276 (5)
13 (65)	3	T_1=394 (3), T_2=411 (3), T_3=299 (3)
14 (66)	2	T_1=166 (2)
15 (67)	1	T_1=323 (2)
16 (68)	1	T_1=421 (2)
17 (86)	2	$T_{1,2}$=421 (2)
18 (87)	2	T_1=145 (2), T_1=207 (2)
19 (93)	1	T_1=498 (5)
20 (95)	1	T_1=71 (2)
21 (96)	1	T_1=175 (2)
22 (107)	3	T_1=598 (2), T_2=172 (2), T_3=35 (2)
23 (111)	2	T_1^a=421 (4), T_1^b=194 (10), T_2=164 (9)
24 (114)	1	T_1=410 (11)
25 (116)	3	T_1=873 (2), T_2=416 (2), T_3=408 (2)
26 (119)	3	T_1=580 (2), T_2=417 (2), T_3=392 (2)
27 (125)	3	T_1=732 (2), T_2=469 (2), T_3=157 (2)
28 (127)	1	T_1=162 (2)
29 (134)	2	$T_{1,2}$=411 (3)
30 (140)	2	T_1=365 (2), T_2=205 (2)
31 (142)	4	T_1=411 (4), T_2=410 (5) , T_3=398 (6) , T_4=305 (2)
32 (148)	1	T_1=71 (4)
33 (149)	1	T_1=205 (2)
34 (151)	1	T_1=114 (2)
35 (156)	2	T_1=172 (2), T_1=633 (2)
36 (161)	4	T_1=384 (3), T_2=206 (3) , T_3=155 (2) , T_4=101 (2)
37 (163)	3	T_1=833 (8) , T_2=138 (8), T_3=69 (6)

38 (166)	1	T_1=167 (2)
39 (170)	2	T_1=135 (2), T_2=164 (2)
40 (175)	3	T_1=405 (2), T_2=274 (2) , T_3=103 (2)
41 (180)	1	T_1=148 (2)
42 (189)	3	T_1=293 (3), T_2=205 (2) , T_3=204 (2)
43 (213)	1	T_1=310 (2)
44 (216)	1	T_1=820 (2)
45 (217)	1	T_1=399 (8)
46 (219)	2	T_1=429 (2), T_2=272 (2)
47 (226)	2	T_1=541 (2), T_2=213 (2)
48 (227)	2	T_1=421 (2), T_2=401 (2)
49 (228)	1	T_1=393 (2)
50 (240)	3	T_1=422 (3), T_2=358 (3) , T_3=200 (2)
51 (244)	2	T_1=580 (2), T_2=445 (2)
52 (273)	2	T_1=518 (2), T_2=98 (2)
53 (291)	2	T_1=541 (2), T_2=636 (2)
54 (295)	2	T_1=411 (2), T_2=404 (2)
55 (308)	1	T_1=421 (19)
56 (313)	4	T_1=231 (5), T_2=210 (4) , T_3=126 (3), T_4=502 (3)
57 (314)	6	T_1=693 (3), T_2=615 (4) , T_3=421 (2), T_4=294 (2), T_5=196 (3), T_6=183 (3)
58 (350)	3	T_1=189 (7), T_2=148 (2) , T_3=145 (2)
59 (406)	3	T_1=222 (3), T_2=198 (4) , T_3=111 (5)
60 (408)	1	T_1=432 (4)
61 (414)	2	T_1=407 (2), T_2=879 (2)
62 (500)	2	T_1=916 (2), T_2=338 (2)
63 (555)	3	T_1=771 (3), T_2=550 (3), T_3=236 (3)
64 (563)	3	T_1=709 (3) , T_2=273 (2), T_3=111 (2)
65 (574)	3	T_1=377 (8), T_2=346 (7) , T_3=185 (4)
66 (585)	2	T_1=239 (2), T_2=191 (2)
67 (595)	4	T_1=669 (11), T_2=240 (3) , T_3=211 (2), T_4=69 (2)

Πίνακας 5 Οι περιοχές αναφοράς, οι περίοδοι και οι περιοδικές συμπεριφορές για τα 67 σκάφη που αναλύθηκαν με την μέθοδο Periodica.

Βάσει των αποτελεσμάτων που παρουσιάζονται στον πίνακα 5 προκύπτουν τα παρακάτω στατιστικά στοιχεία για τις περιοχές αναφοράς, τις περιόδους και τις περιοδικές συμπεριφορές των 67 σκαφών.

Εικόνα 19 Κατανομή περιοχών αναφοράς.

Αριθμός Περιοχών Αναφοράς	1	2	3	4	5	6
Αριθμός Σκαφών	26	15	20	5	0	1
Ποσοστό Σκαφών	38.8%	22.4%	29.9%	7.5%	0.0%	1.5%

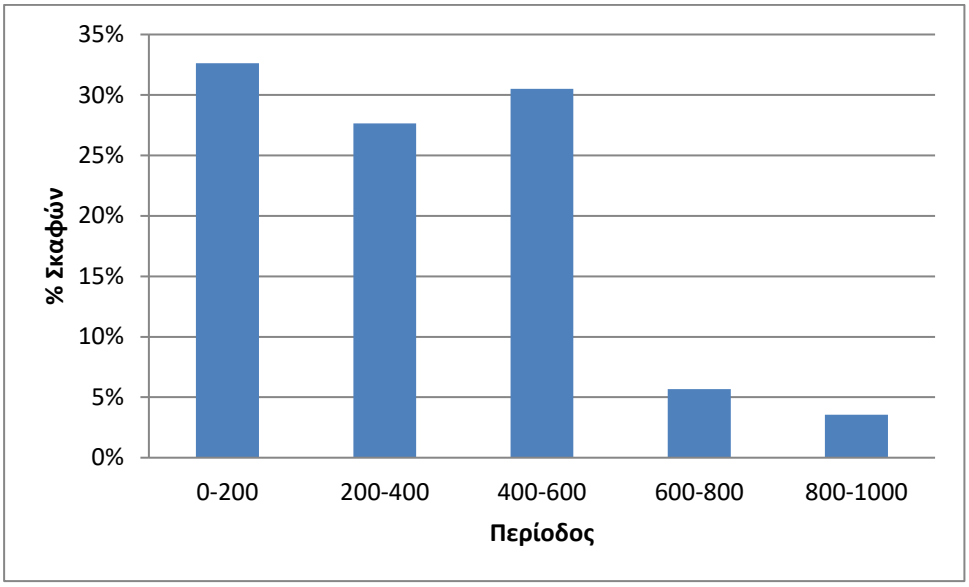
Πίνακας 6 Κατανομή περιοχών αναφοράς για τα 67 σκάφη που αναλύθηκαν με την μέθοδο Periodica.

	Ελάχιστη	Μέγιστη	Μέση Τιμή	Τυπική Απόκλιση
Περίοδοι	35	916	330	192
Περιοδικές Συμπεριφορές	2	19	3.5	3.2

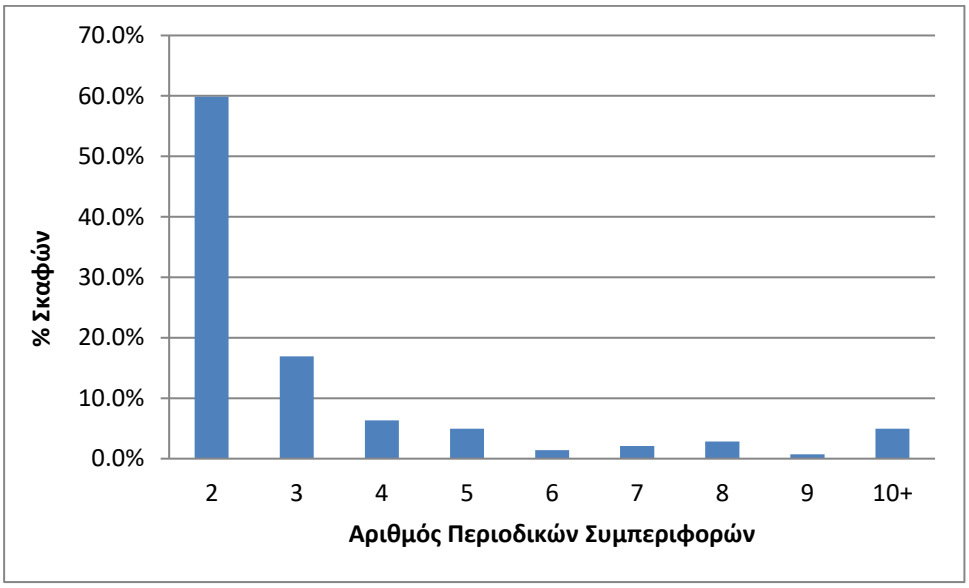
Πίνακας 7 Στατιστικά στοιχεία για τις περιόδους και τις περιοδικές συμπεριφορές για τα 67 σκάφη με δεδομένα vms.

Περίοδοι	0-200	200-400	400-600	600-800	800-1000
Αριθμός Περιόδων	46	39	43	8	5
% Περιόδων	33%	28%	30%	6%	4%

Πίνακας 8 Κατανομή περιόδων για τα 67 σκάφη με δεδομένα vms.



Εικόνα 20 Κατανομή του αριθμού των περιόδων που παρατηρήθηκαν για κάθε σκάφος στα δεδομένα vms.



Εικόνα 21 Κατανομή των περιόδων που παρατηρήθηκαν στα δεδομένα vms.

3.3. Συγκεντρωτικά AIS & VMS δεδομένα

Παρακάτω παρουσιάζουμε συγκεντρωτικά στοιχεία για τα 96 σκάφη που μελετήθηκαν συνολικά, όπως προέκυψαν από δεδομένα vms και AIS, με την μέθοδο Periodica.

Αριθμός Περιοχών Αναφοράς	1	2	3	4	5	6
Αριθμός Σκαφών	36	25	24	8	2	1
Ποσοστό Σκαφών	37.5%	26.0%	25.0%	8.3%	2.1%	1.0%

Πίνακας 9 Η κατανομή του αριθμού περιοχών αναφοράς που παρατηρήθηκαν.

Παρατηρούμε ότι η πλειονότητα των σκαφών σχετίζεται με 1-3 περιοχές αναφοράς που κατά πάσα πιθανότητα αντιστοιχούν σε 1 ή 2 λιμάνια και 1 ή 2 περιοχές ψαρέματος .

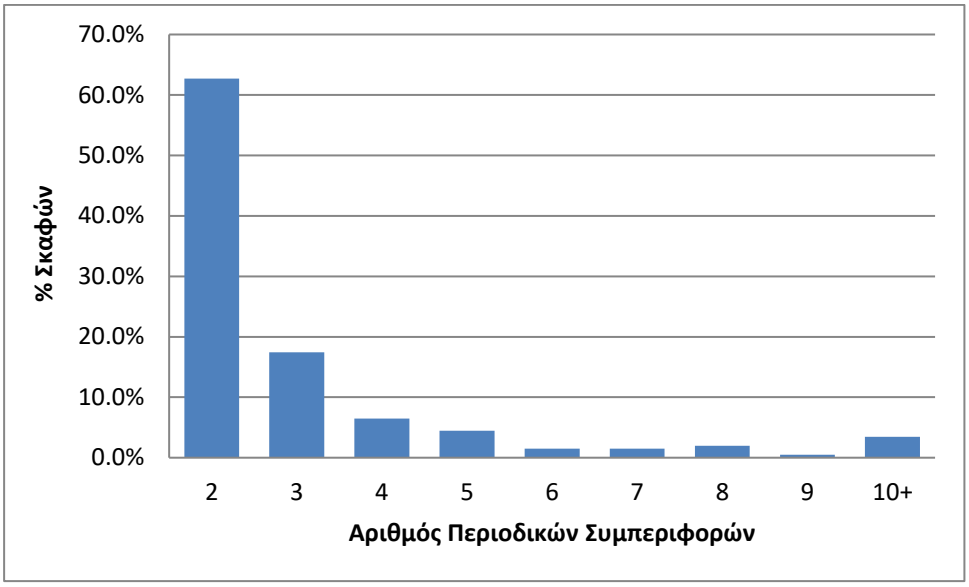
	Ελάχιστη	Μέγιστη	Μέση Τιμή	Τυπική Απόκλιση
Περίοδοι	35	985	301	199
Περιοδικές Συμπεριφορές	2	19	3.2	2.8

Πίνακας 10 Στατιστικά στοιχεία για τις περιόδους και τις περιοδικές συμπεριφορές που παρατηρήθηκαν στα 96 σκάφη που μελετήθηκαν.

Επίσης παρατηρούμε ότι η συντριπτική πλειονότητα των σκαφών παρουσιάζει 2 ή 3 περιοδικές συμπεριφορές για κάθε περίοδο.

Περιοδικές Συμπεριφορές	2	3	4	5	6	7	8	9	10+
Αριθμός Σκαφών	126	35	13	9	3	3	4	1	7
% Σκαφών	62.7%	17.4%	6.5%	4.5%	1.5%	1.5%	2.0%	0.5%	3.5%

Πίνακας 11 Κατανομή του αριθμού των περιοδικών συμπεριφορών για τις περιόδους που παρατηρήθηκαν στο σύνολο των σκαφών.

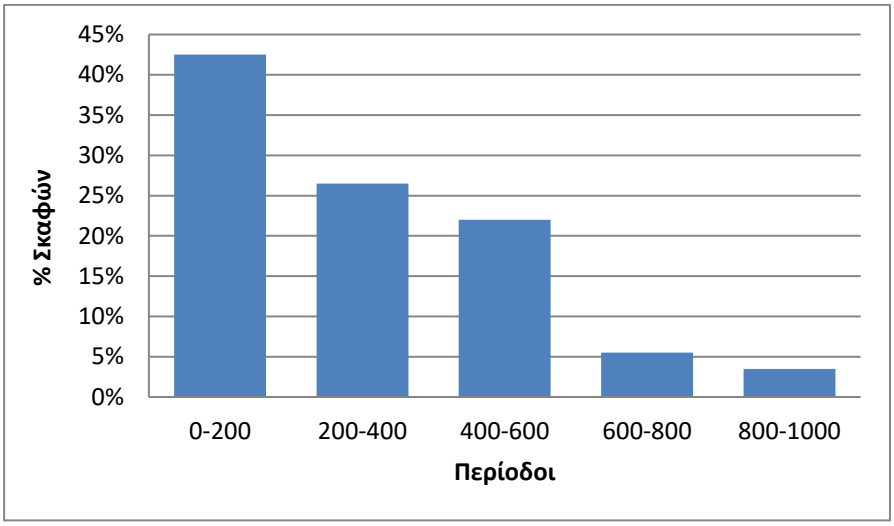


Εικόνα 22 Κατανομή αριθμού περιοδικών συμπεριφορών για τα 96 σκάφη που μελετήθηκαν.

Περίοδοι	0-200	200-400	400-600	600-800	800-1000
Αριθμός Περιόδων	85	53	44	11	7
% Περιόδων	42.5%	26.5%	22%	5.5%	3.5%

Πίνακας 12 Κατανομή των περιόδων που παρατηρήθηκαν στο σύνολο των σκαφών.

Τέλος παρατηρούμε ότι οι περισσότερες περίοδοι που παρατηρούνται (43%) στην κίνηση των σκαφών είναι προσεγγιστικά της τάξεως της βδομάδας, αλλά παρατηρούνται και μεγαλύτερες περίοδοι των δύο ή τριών εβδομάδων. Πιο σπάνιες είναι περιοδικές κινήσεις των αλιευτικών σκαφών σε μηνιαία ή διμηνιαία βάση.



Εικόνα 23 Κατανομή περιόδων για τα 96 σκάφη που μελετήθηκαν.

3.4. Συμπεράσματα

Αναλύσαμε τα δεδομένα τροχιάς αλιευτικών σκαφών τα οποία περισυλλέγησαν με τις μεθόδους Automatic identification system (AIS) και Vessel monitoring system (VMS). Συγκεκριμένα 29 σκάφη σχετίζονται με δεδομένα AIS και 67 με δεδομένα VMS. Τα δεδομένα αναλύθηκαν με την μέθοδο Periodica.

Σημειώνουμε αρχικά ότι η μέθοδος Periodica δεν φαίνεται να είναι η καταλληλότερη για την ανάλυση των δεδομένων AIS και αυτό ισχύει γιατί οι συγκεκριμένες χρονοσειρές παρουσιάζουν μεγάλα κενά τα οποία δεν μπορούν να καλυφθούν αξιόπιστα με κάποια μέθοδο παρεμβολής, ενώ ο κώδικας Periodica είναι κατασκευασμένος να λειτουργεί με συνεχείς χρονοσειρές με σταθερό χρονικό βήμα. Τα δεδομένα vms θεωρούνται καταλληλότερα για ανάλυση με την μέθοδο Periodica γιατί αντιστοιχούσαν σε σχετικά συνεχείς χρονοσειρές.

Επιπλέον, παρατηρήθηκε ότι 37.5% των σκαφών σχετίζονται με ένα σημείο αναφοράς, το οποίο εύλογα υποθέτουμε ότι είναι ένα λιμάνι από το οποίο τα σκάφη εκτελούν τις κυκλικές διαδρομές τους. Τα σκάφη αυτά δεν σχετίζονται με κάποια συγκεκριμένη ζώνη ψαρέματος. Η μέση περιοδικότητα που παρατηρήθηκε στα σκάφη αυτά είναι 290 ώρες ή περίπου 11 ημέρες, ενώ το 60% των σκαφών αυτών εμφανίζει δύο περιοδικές συμπεριφορές.

Για το 26% των σκαφών η κίνηση τους σχετίζεται με δύο περιοχές αναφοράς, από τις οποίες η μία θα είναι σίγουρα ένα λιμάνι ενώ η δεύτερη μπορεί να είναι μια συγκεκριμένη περιοχή ψαρέματος ή κάποιο άλλο λιμάνι. Η μέση περίοδος της κίνηση των σκαφών αυτών είναι 378 ώρες ή περίπου 15 μέρες ενώ το 80% των σκαφών αυτών εμφανίζει 2 περιοδικές συμπεριφορές.

Το 25% των σκαφών σχετίζεται με τρεις περιοχές αναφοράς ενώ το υπόλοιπο 11% των σκαφών σχετίζεται με περισσότερες από τρεις περιοχές αναφοράς. Για τα σκάφη με 3 περιοχές αναφοράς η μέση περίοδος είναι περίπου 470 ώρες ή 19 ημέρες ενώ μόνο το 46% από αυτά εμφανίζει 2 περιοδικές συμπεριφορές με τα υπόλοιπα να εμφανίζουν περισσότερες.

Τέλος σημειώνουμε ότι το 43% των περιόδων που παρατηρήθηκαν ήταν της τάξεως της μιας εβδομάδας ενώ το 27% των 2 εβδομάδων και το 22% των 3 εβδομάδων, ενώ 62.7% των περιόδων που παρατηρήθηκαν εμφανίζει 2 περιοδικές συμπεριφορές και το 17.4% τρεις με τις υπόλοιπες περιόδους να εμφανίζουν περισσότερες περιοδικές συμπεριφορές.

Σαν τελικό συμπέρασμα για τον αλγόριθμο Periodica στα συγκεκριμένα δεδομένα μπορούμε να πούμε ότι κατάφερε να μας βγάλει κάποια αποτελέσματα μετά την μετατροπή των δεδομένων παρά τις δυσκολίες που αντιμετωπίστηκαν. Ίσως ένας αλγόριθμος ο οποίος θα εφαρμοζόταν στα πραγματικά δεδομένα και να μπορούσε να διαχειριστεί τις περιπτώσεις όπου ο αλγόριθμος Periodica αδυνατούσε με σκοπό να αφαιρέσουμε ένα μεγάλο ποσοστό των δεδομένων να λειτουργούσε καλύτερα στην περίπτωση μας.

Σε μελλοντικές μελέτες θα μπορούσε να αναλυθεί η απόσταση μεταξύ των περιοχών αναφοράς που παρατηρούνται για κάθε σκάφος, προφανώς αυτό ισχύει για τα σκάφη που παρουσιάζουν περισσότερες της μιας περιοχής αναφοράς. Επίσης, θα μπορούσε να αναλυθεί πως σχετίζεται η απόσταση αυτή με τις αντίστοιχες περιόδους που παρατηρούνται. Επιπλέον, θα ήταν ενδιαφέρον να διαπιστώσουμε αν υπάρχουν κοινές περιοχές αναφοράς μεταξύ των σκαφών που αναλύθηκαν και εν γένει πως κατανέμονται και σχετίζονται μεταξύ τους οι περιοχές αναφοράς συνολικά των σκαφών που μελετήθηκαν. Τέλος, με ανάλυση των περιοδικών συμπεριφορών θα μπορούσε να γίνει αξιολόγηση των περιοχών αναφοράς υπολογίζοντας την χρονική διάρκεια για την οποία κάθε σκάφος βρίσκεται μέσα σε αυτές.

Π1

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική Βιβλιογραφία

Ν. Πελέκης , Πειραιάς 2019. Data Mining, Πανεπιστημιακές Σημειώσεις

Πηγή της παρακάτω ιστοσελίδας Van de Walle, John A. (2005):
https://tsetsosstavros.blogspot.com/2011/08/blog-post_27.html [1]

aganet.gr:
https://aganet.gr/what-is-ais/ [2]

Ευρωπαϊκό Ελεγκτικό συνέδριο:
https://op.europa.eu/webpub/eca/special-reports/fisheries-08-2017/el/ [3]

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ ΤΜΗΜΑ ΜΗΧ/ΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ:
Spatial And Temporal Data Mining Βασίλειος Μεγαλοοικονόμου [4]

Ξενόγλωσση Βιβλιογραφία

Fourier Viger P., Jerry Lin C.W. Kiran Rage U. (2018) Discovering Periodic Patterns Common to Multiple Sequences: 20th International Conference
DOI: 10.1007/978-3-319-98539-8_18 [5]

Cao, H., Mamoulis, N., & Cheung, D. W. (2007). Discovery of periodic patterns in spatiotemporal sequences. IEEE Transactions on Knowledge and Data Engineering, 19(4), 453–467. https://doi.org/10.1109/TKDE.2007.1002 [6]

Li, Z., Ding, B., Han, J., Kays, R., & Nye, P. (2010). Mining periodic behaviors for moving objects. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1099–1108. https://doi.org/10.1145/1835804.1835942 [7]

Yuan, Q., Zhang, W., Zhang, C., Geng, X., Cong, G., & Han, J. (2017). PRED: Periodic region detection for mobility modeling of social media users. WSDM 2017 - Proceedings of the 10th ACM International Conference on Web Search and Data Mining, 263–272. https://doi.org/10.1145/3018661.3018680 [8]

Stephanie Glen :

https://www.statisticshowto.com/chinese-restaurant-process/ [9]

Panupong (Ice) Pasupat :

https://ppasupat.github.io//a9online/bayesian-nonparametrics/023-chinese-view.html [10]