



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Τεχνολογιών Πληροφορικής και Επικοινωνιών

Τμήμα Πληροφορικής

Πτυχιακή Εργασία

Τίτλος πτυχιακής εργασίας	Εύρεση στατιστικά σημαντικών χωρικών συστάδων σε δεδομένα πλοίων Detecting statistically significant spatial clusters in vessel data
Ονοματεπώνυμο φοιτητή	Μίμογλου Γεώργιος
Πατρώνυμο	Κωνσταντίνος
Αριθμός μητρώου	Π17073
Επιβλέπων καθηγητής	Θεοδωρίδης Ιωάννης, Καθηγητής

Σεπτέμβριος 2021

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Ιωάννη Θεοδορίδη για την πολύτιμη βοήθεια και καθοδήγηση, τόσο καθ' όλη τη διάρκεια των σπουδών μου στο Τμήμα Πληροφορικής, όσο και κατά την εκπόνηση της πτυχιακής μου εργασίας. Επιπλέον, θέλω να ευχαριστήσω τον υποψήφιο διδάκτορα κ. Ιωάννη Κοντούλη για τη στήριξη και τη βοήθεια στη εκπόνηση της παρούσας εργασίας. Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για όσα μου έχουν προσφέρει, την τεράστια υπομονή που έχουν για εμένα και που αποτελούν πυλώνα της ζωής μου. Σας ευχαριστώ για όλα.

Περίληψη

Η επιτήρηση θαλάσσιων περιοχών, τόσο σε ευρωπαϊκό όσο και σε παγκόσμιο επίπεδο, αποτελεί βασικό άξονα ανάπτυξης της σύγχρονης ναυτιλίας. Αυτό οφείλεται στην πολύπλοκη φύση του πεδίου όσον αφορά την προβλεψιμότητα, την ασφάλεια, τη διαχείριση παράνομων ενεργειών, καθώς και τα περιβαλλοντικά ζητήματα που εγείρονται. Σε αυτή την κατεύθυνση, η θαλάσσια επιτήρηση συμβάλει τα μέγιστα στην ανακάλυψη μοτιβών, η αποκωδικοποίηση των οποίων συνιστά πολύτιμο εργαλείο για πληθώρα δράσεων και ενεργειών, όπως η αποφυγή συγκρούσεων, η διαχείριση της κυκλοφορίας και η μείωση της θαλάσσιας ρύπανσης, μεταξύ άλλων. Ανάμεσα στα διάφορα εποπτικά μέσα που χρησιμοποιούνται ευρέως τα τελευταία χρόνια, το Automatic Identification System (AIS) έχει καθιερωθεί σαν μία από τις πλέον αξιόπιστες τεχνολογίες επιτήρησης, που απαντάται στους περισσότερους τύπους πλοίων. Τα εν λόγω δεδομένα συνίστανται σε κινηματικά και στατικά μηνύματα, παρέχοντας εξαιρετικά χρήσιμες πληροφορίες, αναφορικά με την κίνηση και την ταυτότητα των πλοίων. Στο πλαίσιο της παρούσης εργασίας, προτείνονται μέθοδοι και αλγόριθμοι για την επίτευξη των στόχων επεξεργασίας και ανάλυσης θαλάσσιων δεδομένων, όπως οι προαναφερθείσες περιπτώσεις. Ειδικότερα, παρουσιάζονται δύο αλγόριθμοι χωροχρονικής συσταδοποίησης, που συμβάλουν στον εντοπισμό αγκυροβολίων και τα χαρακτηρίζουν ως μακροπρόθεσμα ή βραχυπρόθεσμα βάσει της διάρκειάς τους (Spatio-Temporal Clustering – STC) και ως στατιστικά σημαντικά ή μη (hot/cold spots), λαμβάνοντας υπόψη το πλήθος των πλοίων για συγκεκριμένο χρονικό διάστημα. Τέλος, διεξήχθησαν πειράματα με τη χρήση πραγματικών δεδομένων AIS που ελήφθησαν από την κεραία του Πανεπιστημίου Πειραιώς κατά το έτος 2018 και αφορούν χωρικά στην ευρύτερη περιοχή του Σαρωνικού Κόλπου. Επιπλέον, χρησιμοποιήθηκαν προσομοιωμένα δεδομένα για τη διεξαγωγή πειραμάτων κλιμάκωσης.

Λέξεις κλειδιά: Θαλάσσια Επιτήρηση, Συσταδοποίηση, Στατιστικά Σημαντικές Περιοχές, Αγκυροβόλια, AIS, Χωροχρονικές Διαστάσεις

Abstract

Surveillance of marine areas is major pillar of the modern maritime evolution, in both European and International level. This is due to the complex nature of the domain which concerns the predictability, the safety, the management of illegal activities, as well as the environmental issues that are raised. In this regard, maritime surveillance contributes the most to the discovery of patterns, the decoding of which is a valuable tool for a variety of actions and actions, such as collision avoidance, traffic management and the reduction of marine pollution, among others. Among the various surveillance tools widely used in recent years, the Automatic Identification System (AIS) has established itself as one of the most reliable surveillance technologies found on most ship types. This data consists of kinematic and static messages, providing extremely useful information regarding the movement and identity of ships. In the context of this work, methods and algorithms are proposed to achieve the objectives of processing and analyzing marine data, such as the aforementioned cases. In particular, two spatio-clustering algorithms are presented, which help to locate anchorages and characterize them as long-term or short-term based on their duration (Spatio-Temporal Clustering – STC) and as statistically significant or not (hot/cold spots), taking into account the number of ships for a certain period of time. Finally, experiments were performed using real AIS data received from the antenna of the University of Piraeus in the year 2018 and spatially concern the wider area of the Saronic Gulf. In addition, simulated data were used to conduct escalation experiments.

Keywords: Maritime Surveillance, Clustering, Statistically Significant Regions, Anchorages, AIS, Spatio-temporal Dimensions

Αφιερωμένο στην υπέροχη οικογένειά
μου και στη λαμπρή σταδιοδρομία της
Φελίσιας και της Μελίνας.

Περιεχόμενα

Ευχαριστίες	i
Περίληψη	ii
Abstract	ii
Περιεχόμενα.....	iv
Κατάλογος εικόνων.....	v
Κατάλογος πινάκων	vii
1 Εισαγωγή	1
2 Προκαταρκτικές έννοιες και ορισμός του προβλήματος	1
2.1 Χωρική αυτοσυσχέτιση και στατιστική σημαντικότητα	1
2.2 Χωρικές αναλύσεις.....	2
2.3 Ορισμός του προβλήματος.....	4
3 Σχετικές εργασίες	6
3.1 Μια αρχική προσέγγιση: THS για περιοχές αγκυροβολίων	7
4 Spatio-Temporal Clustering.....	8
4.1 Στατιστική σημασία χωρίς στατιστική	8
4.2 Μοντελοποίηση.....	9
4.3 Μετα-ανάλυση αποτελεσμάτων	13
4.4 Πλεονεκτήματα και μειονεκτήματα του αλγορίθμου.....	14
5 Spatio-Temporal Hot Spot.....	14
5.1 Αυτοσυσχέτιση στο χωροχρονικό πεδίο	15
5.2 Μοντελοποίηση.....	16
5.3 Σύγκριση Spatio-Temporal Hot Spot και Trajectory Hot Spot	19
6 Υλοποίηση και παραδείγματα	19
6.1 STC και STHS σε δεδομένα AIS	20
6.2 STC και STHS σε προσομοιωμένα δεδομένα	26
7 Συμπεράσματα και μελλοντική μελέτη	33
Αναφορές	34

Κατάλογος εικόνων

Fig. 1. Απεικόνιση της χωρικής αυτοσυσχέτισης. https://pro.arcgis.com/	2
Fig. 2. Απεικόνιση της συσταδοποίησης High/Low. https://pro.arcgis.com/	3
Fig. 3. Απεικόνιση υπολογισμού z-scores και p-values με το Gi*. https://pro.arcgis.com/	3
Fig. 4. Απεικόνιση υπολογισμού z-scores και p-values με το I, για τον εντοπισμό συστάδων και ακραίων τιμών. https://pro.arcgis.com/	4
Fig. 5. Μορφή αγκυροβολίου σε πλέγμα με βάση το πλήθος των πλοίων.....	5
Fig. 6. Η εξέλιξη ενός αγκυροβολίου με την πάροδο του χρόνου.	6
Fig. 7. Η τροχιά ενός πλοίου και το "κουβάρι" που δημιουργεί.	7
Fig. 8. Συστάδα από "κουβάρια" τροχιών.....	8
Fig. 9. Αφηρημένη απεικόνιση ενός βραχυχρόνιου αγκυροβολίου.	9
Fig. 10. Αφηρημένη απεικόνιση ενός μακροπρόθεσμου αγκυροβολίου.....	9
Fig. 11. Επιλογή σημείων εισόδου, εξόδου και μέσου από μια τροχιά.....	10
Fig. 12. Χρονική διαμέριση και τοποθέτηση των χρόνων εισόδου και εξόδου	10
Fig. 13. Συσταδοποίηση 8 μέσων σημείων m_i , σε δυο διαδοχικές χρονικές φέτες t_j και $t_j + 1$ και εξαγωγή των centroids u_{jk}	11
Fig. 14. Συσταδοποίηση των κεντροειδών των συστάδων-αγκυροβολίων, υποδηλώνει την ύπαρξη του αγκυροβολίου κατά τη διάρκεια του χρονικού παραθύρου.	12
Fig. 15. Γραφική απεικόνιση της εξασθένησης για διάφορες τιμές του z και $m = 1$	16
Fig. 16. Απεικόνιση χωροχρονικής εξασθένησης.....	16
Fig. 17. Χαρακτηρισμός συστάδων για διάφορα επίπεδα εμπιστοσύνης. https://pro.arcgis.com/	18
Fig. 18. Διάγραμμα διασποράς σημείων και ιστόγραμμα διάρκειας παραμονής... ..	21
Fig. 19. STC AIS test 1.....	22
Fig. 20. STC AIS test 2.....	22
Fig. 21. STHS AIS test 1.....	23
Fig. 22. STHS AIS test 2.....	23
Fig. 23. Ιστόγραμμα χαρακτηριστικών τιμών στα V_j	24
Fig. 24. STC AIS test 3.....	25
Fig. 25. STHS AIS test 3.....	25
Fig. 26. STC AIS test 4.....	26
Fig. 27. STHS AIS test 4.....	26
Fig. 28. Dummy data test 1.....	27
Fig. 29. STC dummy test 1.....	28
Fig. 30. STHS dummy test 1.....	28
Fig. 31. Dummy data test 2.....	29
Fig. 32. STC dummy test 2.....	29
Fig. 33. STHS dummy test 2.....	30
Fig. 34. Dummy data test 3.....	30
Fig. 35. STC dummy test 3.....	31
Fig. 36. STHS dummy test 3.....	31

Fig. 37. Dummy data test 4.....	32
Fig. 38. STC dummy test 4.	32
Fig. 39. STHS dummy test 4.....	33

Κατάλογος πινάκων

Table 1. Διαφορές στις ερμηνείες μεταξύ High/Low Clustering και Spatial Autocorrelation.....	3
Table 2. Ερμηνεία των αποτελεσμάτων της τεχνικής Hot Spot Analysis.....	3
Table 3. Χαρακτηρισμός συστάδων.	17
Table 4. Παράμετροι δοκιμών	21
Table 5. Παράμετροι δοκιμών προσομοιωμένων δεδομένων	27

1 Εισαγωγή

Αντικείμενο της παρούσας εργασίας είναι η επιτήρηση θαλάσσιων περιοχών, που τα τελευταία χρόνια διαδραματίζει πρωταγωνιστικό ρόλο στον τομέα της ναυτιλίας, αναφορικά με ζητήματα που σχετίζονται με την ασφάλεια και την εποπτεία των θαλάσσιων δραστηριοτήτων. Οι δραστηριότητες αυτές απαντώνται σε ενέργειες σχετικά με την ασφαλή πλοήγηση των πλοίων, τη διαχείριση της κυκλοφορίας, τον περιορισμό της θαλάσσιας ρύπανσης (π.χ. μέσω του ελέγχου εκπομπής ρύπων), την αποτροπή παραβατικών συμπεριφορών (π.χ. παράνομη αλιεία), κ.ά. Οι σύγχρονες τεχνολογίες επιτήρησης, σε συνδυασμό με τη δυνατότητα αδιάλειπτης επικοινωνίας και ανταλλαγής πληροφοριών σχεδόν σε κάθε μέρος του πλανήτη, συμβάλουν στη δημιουργία ενός εξαιρετικά μεγάλου όγκου δεδομένων, τα οποία αποτελούν την πρώτη ύλη για την εξυπηρέτηση πληθώρας σεναρίων επεξεργασίας και αναλυτικής δεδομένων κίνησης.

Συγκεκριμένα, η εργασία εστιάζει στην ανάλυση της κίνησης των πλοίων και τον εντοπισμό αγκυροβολίων, περιοχές για τις οποίες παρέχονται χρήσιμες πληροφορίες, δυνητικά αξιοποιήσιμες για τους ειδικούς του χώρου (domain experts). Στην κατεύθυνση αυτή, παρουσιάζονται δύο αλγόριθμοι, που αποσκοπούν τόσο στον εντοπισμό αγκυροβολίων, όσο και στον χρονικό (Spatio-Temporal Clustering) και στατιστικό (Spatio-Temporal Hot Spot) χαρακτηρισμό αυτών. Έναυσμα για τη δημιουργία του πρώτου αλγόριθμου αποτέλεσε η αδυναμία ένταξης της στατιστικής σημαντικότητας στη λύση του προβλήματος, ζήτημα το οποίο επιλύθηκε με τη δημιουργία του δεύτερου αλγόριθμου, παρέχοντας έτσι μια άρτια και ολοκληρωμένη προσέγγιση στο πρόβλημα της εύρεσης στατιστικά σημαντικών χωρικών συστάδων. Οι αλγόριθμοι αυτοί προτείνονται ως αυτοτελείς μέθοδοι με διαφορετική χρήση η κάθε μία.

2 Προκαταρκτικές έννοιες και ορισμός του προβλήματος

2.1 Χωρική αυτοσυσχέτιση και στατιστική σημαντικότητα

Το πρόβλημα το οποίο προσπαθεί να επιλύσει αυτή η εργασία τοποθετείται στον χώρο της θαλάσσιας επιτήρησης, αφού στόχος είναι η ανακάλυψη περιοχών στις οποίες παρατηρείται συγκέντρωση πλοίων και παρουσιάζουν στατιστική σημαντικότητα. Τι είναι όμως η στατιστική σημαντικότητα στα χωρικά δεδομένα και πώς υπολογίζεται; Οι πιο γνωστές χωρικές στατιστικές, όπως Moran's I [1] και Getis-Ord G_i^* [2] ορίζουν ως null hypothesis (H_0) την πλήρη χωρική τυχαιότητα, δηλαδή ένα χωρικό φαινόμενο πραγματοποιείται με εντελώς τυχαίο τρόπο. Όταν η τιμή της στατιστικής καταρρίπτει την H_0 , λέμε πως η περιοχή αυτή είναι στατιστικά σημαντική. Με άλλα λόγια, οι στατιστικές αυτές εκφράζουν τη χωρική αυτοσυσχέτιση μιας περιοχής ή αλλιώς το πόσο πολύ, λίγο ή καθόλου ένα φαινόμενο της περιοχής επηρεάζεται θετικά ή αρνητικά από τα γειτονικά φαινόμενα.

Για παράδειγμα, αν υποθέσουμε πως η στατιστική G_i^* μετράει την αυτοσυσχέτιση του πληθυσμού των δήμων της Αττικής, θα παρατηρήσουμε πως οι δήμοι κοντά στο κέντρο της Αθήνας λαμβάνουν υψηλές τιμές, αφού έχουν υψηλό πληθυσμό και γειτνιάζουν με περιοχές υψηλού πληθυσμού. Συνεπώς, ενδέχεται να είναι και

στατιστικά σημαντικές. Η τελευταία πρόταση χαρακτηρίζεται ως ενδεχόμενο, αφού εξαρτάται από το διάστημα εμπιστοσύνης, το οποίο θα θέσει ο αναλυτής. Αντιθέτως, οι δήμοι στα βόρεια προάστεια θα εμφανίζουν μικρότερη τιμή.

2.2 Χωρικές αναλύσεις

Σε αυτό το σημείο πρέπει να αναφερθεί πως κάθε στατιστική χρησιμοποιείται για διαφορετικές αναλύσεις, έχοντας όμως τη βασική ιδέα κοινή, δηλαδή τη χωρική αυτοσυσχέτιση. Ορισμένες από αυτές είναι: ο υπολογισμός ολικής χωρικής αυτοσυσχέτισης [3], η συσταδοποίηση High/Low [4], η ανάλυση Hot Spot [5], η ανάλυση Cluster και Outlier [6]. Οι στατιστικές που αναφέρονται παρακάτω χρησιμοποιούνται από τα γνωστά εργαλεία του Environmental Systems Research Institute 1 (ESRI), ενώ υπάρχουν περιπτώσεις που θα μπορούσαν να χρησιμοποιηθούν διαφορετικές προσεγγίσεις.

Spatial Autocorrelation. Η πρώτη ανάλυση χρησιμοποιεί το Moran's I (ή το Geary's C [7], που ποσοτικοποιεί τη χωρική αυτοσυσχέτιση βάσει της τοποθεσίας και της τιμής ενός χαρακτηριστικού, δηλαδή αν τα παρατηρούμενα μοτίβα τιμών είναι συσταδοποιημένα, διασκορπισμένα ή τυχαία κατανεμημένα.

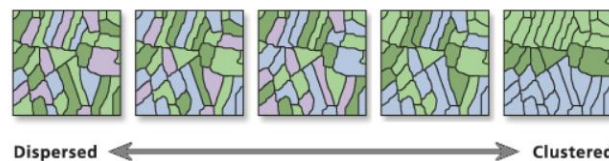


Fig. 1. Απεικόνιση της χωρικής αυτοσυσχέτισης. <https://pro.arcgis.com/>

High/Low Clustering. Η συσταδοποίηση High/Low χρησιμοποιεί τον Getis-Ord General G, ο οποίος αποτελεί μια γενικότερη στατιστική του G_i^* και μετράει τον βαθμό συσταδοποίησης είτε υψηλών είτε χαμηλών τιμών ενός χαρακτηριστικού των περιοχών (βλ. Fig. 2). Το μειονέκτημα αυτής της ανάλυσης είναι πως αν η περιοχή μελέτης περιέχει συστάδες υψηλών και χαμηλών τιμών, τότε ο συνολικός δείκτης G δεν μπορεί να αποδώσει ορθό αποτέλεσμα. Η διαφορά της προηγούμενης ανάλυσης και της ανάλυσης High/Low βρίσκεται στην ερμηνεία των z-scores. Το Table 1 παρουσιάζει τις διαφορές αυτές.

¹ <https://www.esri.com>

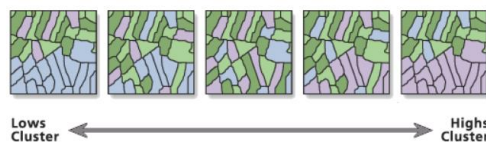


Fig. 2. Απεικόνιση της συσταδοποίησης High/Low. <https://pro.arcgis.com/>

Table 1. Διαφορές στις ερμηνείες μεταξύ High/Low Clustering και Spatial Autocorrelation

Αποτέλεσμα	High/Low Clustering	Spatial Autocorrelation
Το p-value δεν είναι στατιστικά σημαντικό	Δεν απορρίπτεται η null hypothesis	
Το p-value είναι στατιστικά σημαντικό και το z-score είναι <i>θετικό</i>	Απορρίπτεται η null hypothesis, οι υψηλές τιμές είναι χωρικά συγκεντρωμένες	Απορρίπτεται η null hypothesis, οι υψηλές τιμές είναι χωρικά συγκεντρωμένες
Το p-value είναι στατιστικά σημαντικό και το z-score είναι <i>αρνητικό</i>	Απορρίπτεται η null hypothesis, οι χαμηλές τιμές είναι χωρικά συγκεντρωμένες	Απορρίπτεται η null hypothesis, οι υψηλές και χαμηλές τιμές είναι διασκορπισμένες

Hot Spot Analysis. Οι προηγούμενες δύο αναλύσεις έχουν ως αποτέλεσμα μια συμπερασματική στατιστική, η οποία περιγράφει τα φαινόμενα σε όλη την κλίμακα. Η Hot Spot Analysis εντοπίζει υπο-περιοχές hot spot με υψηλές τιμές, οι οποίες γειτνιάζουν με υπο-περιοχές με υψηλές τιμές, μετρώντας τον βαθμό χωρικής συσταδοποίησης με τη βοήθεια της στατιστικής Getis-Ord G_i^* για κάθε μια υπο-περιοχή μέσα στην ευρύτερη περιοχή μελέτης. Αντίστοιχα, εντοπίζει τις cold spot περιοχές στις περιπτώσεις χαμηλών τιμών. Αυτό βοηθάει στην αντιμετώπιση του μειονεκτήματος της τελευταίας ανάλυσης.



Fig. 3. Απεικόνιση υπολογισμού z-scores και p-values με το G_i^* . <https://pro.arcgis.com/>

Table 2. Ερμηνεία των αποτελεσμάτων της τεχνικής Hot Spot Analysis

Αποτέλεσμα	High/Low Clustering
Το p-value δεν είναι στατιστικά σημαντικό	-
Το p-value υπο-περιοχής είναι στατιστικά σημαντικό και το z-score είναι <i>θετικό</i>	Απορρίπτεται η null hypothesis, οι υψηλές τιμές είναι χωρικά συγκεντρωμένες (hot spot)

Το p-value υπο-περιοχής είναι στατιστικά σημαντικό και το z-score είναι αρνητικό

Απορρίπτεται η null hypothesis, οι χαμηλές τιμές είναι χωρικά συγκεντρωμένες (cold spot)

Cluster and Outlier Analysis. Αυτή η ανάλυση αποτελεί επέκταση της προηγούμενης και εντοπίζει outliers, πέρα από hot & cold spots. Πιο συγκεκριμένα, με τη βοήθεια της στατιστικής Local Moran's I, η ανάλυση εντοπίζει τις περιοχές με υψηλή τιμή, που γειτνιάζουν με περιοχές υψηλής τιμής (High-High), τις περιοχές με χαμηλή τιμή, που γειτνιάζουν με περιοχές χαμηλής τιμής (Low-Low) και τις περιοχές με χαμηλή (αντ. υψηλή), που γειτνιάζουν με περιοχές με υψηλή (αντ. χαμηλή) τιμή (Low-High και High-Low, αντίστοιχα). Οι δύο πρώτες περιπτώσεις αποτελούν τα hot και cold spots, ενώ οι δύο τελευταίες τα outliers.

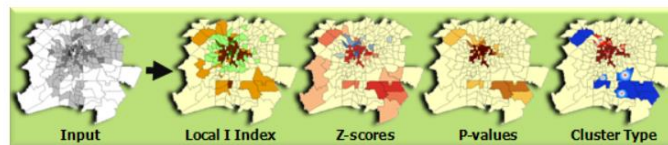


Fig. 4. Απεικόνιση υπολογισμού z-scores και p-values με το I, για τον εντοπισμό συστάδων και ακραίων τιμών. <https://pro.arcgis.com/>

2.3 Ορισμός του προβλήματος

Έχοντας ορίσει τη χωρική αυτοσυσχέτιση και τη στατιστική σημαντικότητα στα χωρικά δεδομένα, πρέπει να ορισθεί η στατιστική σημαντικότητα αναφορικά με το ζητούμενο της εργασίας αυτής. Θεωρήθηκε σκόπιμο η εργασία να επικεντρωθεί στον εντοπισμό αγκυροβολίων. Ορίζουμε το αγκυροβόλιο ως:

Αγκυροβόλιο: η περιοχή όπου παρατηρείται στατική συγκέντρωση πλοίων για κάποιο χρονικό διάστημα.

Πώς μπορούμε να εισάγουμε την έννοια της στατιστικής σημαντικότητας σε αυτό το πεδίο; Αν θεωρήσουμε την περιοχή μελέτης ως ένα πλέγμα, όπου κάθε κελί έχει ως χαρακτηριστικό το πλήθος των πλοίων όπου βρίσκονται μέσα σε αυτό το κελί, αναμένουμε πως η τιμή χωρικής αυτοσυσχέτισης, δεδομένου ότι υπάρχει ένα αγκυροβόλιο σε όλη την περιοχή, θα είναι θετική. Με άλλα λόγια, τα κελιά με μεγάλο πλήθος πλοίων είναι συγκεντρωμένα.

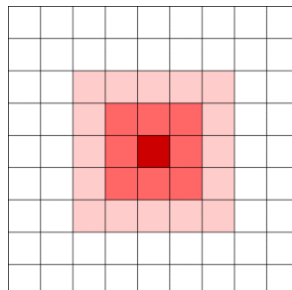


Fig. 5. Μορφή αγκυροβολίου σε πλέγμα με βάση το πλήθος των πλοίων.

Είναι σίγουρο πως μια ολική χωρική στατιστική εξαρτάται σημαντικά από την κλίμακα. Στις περισσότερες περιπτώσεις αναλύσεων χωρικής αυτοσυσχέτισης οι περιοχές συνορεύουν μεταξύ τους, για παράδειγμα οι δήμοι μιας περιφέρειας ή τα κελιά στην περίπτωση ύπαρξης πλέγματος (βλ. THS παρακάτω), πράγμα το οποίο δημιουργεί μια σχετική πυκνότητα των δεδομένων, αναδεικνύοντας περισσότερο τα μοτίβα συσχέτισης που υπάρχουν. Στην περίπτωση της θαλάσσιας επιτήρησης, όπου η κλίμακα είναι τεράστια και τα δεδομένα είναι αραιά στη χωροχρονική διάσταση, ο εντοπισμός των μοτίβων αποτελεί μια δύσκολη και πολλές φορές παραπλανητική διαδικασία. Έτσι, οι τοπικές χωρικές στατιστικές είναι μονόδρομος για τον εντοπισμό στατιστικά σημαντικών περιοχών.

Το πρόβλημα της εργασίας δημιουργεί ένα ακόμα ζήτημα στον καθορισμό της στατιστικής σημαντικότητας. Είναι λογικό πως η ύπαρξη ενός αγκυροβολίου σε μια περιοχή θα έχει αρνητική επίδραση στις γειτονικές περιοχές. Με άλλα λόγια, δεν αναμένουμε να παρατηρήσουμε δυο διαφορετικά γειτονικά αγκυροβόλια. Οπότε, μια ανάλυση hot spot δεν θα έδινε κάποιο ιδιαίτερο αποτέλεσμα. Αντιθέτως, η ανάλυση Cluster and Outlier θα χαρακτήριζε όλα τα αγκυροβόλια ως outliers, αφού στη μεγάλη επιφάνεια της θάλασσας δεν περιλαμβάνει μεγάλο πλήθος στατικών πλοίων. Να σημειωθεί ξανά πως όλα τα παραπάνω εξαρτώνται και από την υπό μελέτη κλίμακα.

Είναι εμφανές πως τα αγκυροβόλια ενδέχεται να μην παρουσιάσουν κάποια ολική στατιστική σημαντικότητα σε μεγάλη κλίμακα μελέτης, ή θα χαρακτηρίζονται στατιστικά σημαντικά σε περιπτώσεις μελέτης υπο-περιοχών. Όμως, όλα τα παραπάνω αναφέρθηκαν κυρίως στις χωρικές διαστάσεις. Τι συμβαίνει όταν εισάγουμε την έννοια του χρόνου; Μήπως έχει νόημα η στατιστική σημαντικότητα στις τρεις διαστάσεις; Παρατηρούμε πως ένα αγκυροβόλιο δεν εξαρτάται από τον χρόνο, δηλαδή δεν υπάρχει κάποιο χρονικό κατώφλι το οποίο χαρακτηρίζει μια περιοχή ως αγκυροβόλιο. Όμως, μπορούμε να κατανοήσουμε πως υπάρχουν διαφορετικά αγκυροβόλια βάσει του χρόνου. Για παράδειγμα, μια μαρίνα στο λιμάνι της Σύρου αποτελεί μακροπρόθεσμο αγκυροβόλιο, ενώ μια περιοχή στα ανοιχτά της Μυκόνου, την καλοκαιρινή περίοδο, την οποία προσεγγίζουν πλοία αναψυχής, αποτελεί βραχυπρόθεσμο αγκυροβόλιο. Έτσι, έχουμε τους εξής ορισμούς:

Μακροπρόθεσμο αγκυροβόλιο: η περιοχή όπου παρατηρείται στατική συγκέντρωση πλοίων για μεγάλο χρονικό διάστημα.

Βραχυπρόθεσμο αγκυροβόλιο: η περιοχή όπου παρατηρείται στατική συγκέντρωση πλοίων για μικρό χρονικό διάστημα.

Έτσι, μακροπρόθεσμο αγκυροβόλια μπορεί να είναι επίσημα λιμάνια και μαρίνες, ενώ βραχυπρόθεσμο αγκυροβόλια μπορεί να είναι περιοχές αρόδου και περιοχές ψαρέματος, μεταξύ άλλων. Πρέπει να σημειωθεί πως το χρονικό κατώφλι, το οποίο διαχωρίζει τις δυο κατηγορίες, εξαρτάται από το πεδίο του προβλήματος και το είδος της ανάλυσης.

Πλέον, έχει νόημα να εξεταστεί η *χωροχρονική* αυτοσυσχέτιση και η στατιστική σημασία των περιοχών, δεδομένου της νέας *null hypothesis*, δηλαδή την *ολική χωροχρονική τυχαιότητα*, ερευνώντας τη χωρική εξέλιξη ενός αγκυροβολίου στη διάσταση του χρόνου.

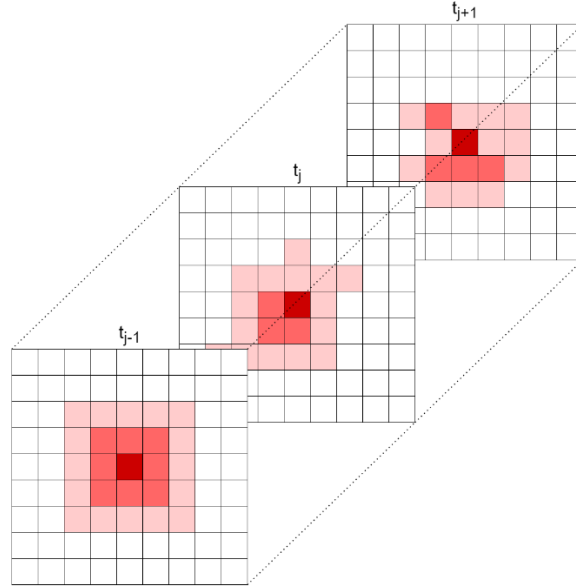


Fig. 6. Η εξέλιξη ενός αγκυροβολίου με την πάροδο του χρόνου.

3 Σχετικές εργασίες

Η εύρεση στατιστικά σημαντικών περιοχών ξεκινάει με την εύρεση συστάδων, ενώ στη συνέχεια προστίθεται και η στατιστική σημασία. Στην περίπτωση της συσταδοποίησης, οι λύσεις OPTICS [8] και DBSCAN [9] αποτελούν τους διασημότερους αλγόριθμους μεγάλων βάσεων δεδομένων, με βάση την πυκνότητα, ενώ οι T-OPTICS [10] και TraClus [11] αποτελούν παραλλαγές των παραπάνω, οι οποίες εισάγουν την έννοια του χρόνου για τη συσταδοποίηση τροχιών. Οι αλγόριθμοι αυτοί αποτελούν το εννοιολογικό δομικό λίθο στο πρόβλημα αυτής της εργασίας, αφού τα αγκυροβόλια χαρακτηρίζονται σημαντικά από την πυκνότητα των πλοίων για κάποια χρονική περίοδο.

Η στατιστική σημαντικότητα χωρικών δεδομένων βασίζεται κυρίως σε στατιστικούς δείκτες, όπως οι Moran's I και G_i^* statistic. Ο πρώτος χρησιμοποιείται ως μέτρο αυτοσυσχέτισης χωρικών δεδομένων σε παγκόσμια κλίμακα. Ο G_i^* αναδεικνύει στατιστικά σημαντικά περιοχές σε σχέση με τα γειτονικά φαινόμενα. Στο [12] παρουσιάζεται μια μέθοδο εύρεσης στατιστικά σημαντικών τροχιών, Trajectory Hot Spot (THS), χρησιμοποιώντας έναν κατάλληλα προσαρμοσμένο δείκτη G_i^* . Πιο συγκεκριμένα, προτείνουν τη χρήση μιας χωροχρονικής διαμέρισης και την ανάθεση μιας τιμής G_i^* σε κάθε κελί βάσει της γειτονίας τους, χρησιμοποιώντας τον χρόνο παραμονής στο κελί ως χαρακτηριστική τιμή x_i . Έστερα, γίνεται εξαγωγή των στατιστικά σημαντικών περιοχών χρησιμοποιώντας τις τιμές αυτές. Το πρόβλημα ανάγεται στην επιλογή των παραμέτρων της διαμέρισης και του πλήθους των κελιών που θα θεωρηθούν ως γειτονιά. Όμως, η εφαρμογή αυτής της μεθόδου σε δεδομένα

πλοίων οξύνει το πρόβλημα, λόγω της χωρικής κλίμακας και της πιθανής αραιής πυκνότητας των σημείων.

Μια ακόμα σημαντική εργασία στον εντοπισμό αγκυροβολίων είναι η [13] στην οποία προτείνεται η μέθοδος Moving and Stationary Evolving Clusters (MaSEC). Πιο συγκεκριμένα, η μέθοδος αποτελείται από δυο εκδοχές του αλγορίθμου Evolving Clusters [14], εκ των οποίων η μια επικεντρώνεται στην εύρεση αγκυροβολίων.

3.1 Μια αρχική προσέγγιση: THS για περιοχές αγκυροβολίων

Όπως αναφέρθηκε παραπάνω, ο THS χρησιμοποιείται για την εύρεση στατιστικά σημαντικών τροχιών. Το πρώτο ερώτημα που τέθηκε ήταν αν θα μπορούσε να λύσει ο THS το πρόβλημα της παρούσας εργασίας. Πώς θα μπορούσε ένα αγκυροβόλιο να φαίνεται στο σύνολο τροχιών; Μια τροχιά παρουσιάζει ορισμένα χαρακτηριστικά όταν το πλοίο παραμένει στάσιμο, όπως για παράδειγμα χαμηλή ταχύτητα, απότομη αλλαγή πορείας και, γενικότερα, ένα «κουβάρι» στιγμάτων. Η παρακάτω εικόνα δείχνει το εν λόγω «κουβάρι».

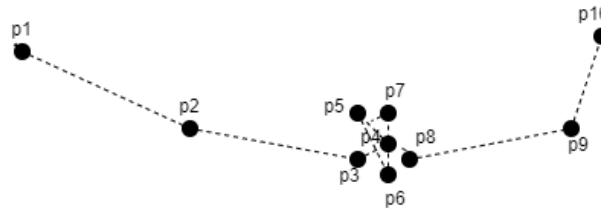


Fig. 7. Η τροχιά ενός πλοίου και το "κουβάρι" που δημιουργεί.

Πλέον, μπορούμε να θεωρήσουμε ένα αγκυροβόλιο ως μια συστάδα από «κουβάρια». Χωρίς βλάβη της γενικότητας, σε μια χρονική φέτα της τρισδιάστατης διαμέρισης, οι εν λόγω συστάδες δημιουργούν γειτονιές κελιών, με υψηλές τιμές x_i – όπως ορίζονται από τη μέθοδο THS – σε σχέση με τα υπόλοιπα κελιά (βλ. Fig. 8).

Έτσι, αν αφαιρέσουμε τα περιττά μέρη των τροχιών, δηλαδή αυτά που βρίσκονται έξω από το «κουβάρι», και εφαρμόσουμε την μέθοδο THS, θα λάβουμε τις στατιστικά σημαντικές περιοχές. Όμως, υπάρχει η περίπτωση στην οποία μια μόνο τροχιά παραμένει για τεράστιο χρονικό διάστημα σε ένα κελί, δημιουργώντας μεγάλη τιμή x_i , με αποτέλεσμα η μέθοδος να θεωρήσει το κελί αυτό ως στατιστικά σημαντική περιοχή υπό τις κατάλληλες γειτονικές συνθήκες. Αυτό ίσως προκαλέσει προβλήματα στην ανάλυση, ειδικά όταν ο αναλυτής αναζητά αγκυροβόλια με περισσότερα από ένα πλοία. Επιπλέον, ένα ακόμα πρόβλημα του THS αποτελεί ο κατάλληλος ορισμός των παραμέτρων για τη χωροχρονική διαμέριση. Κελιά μικρής επιφάνειας αυξάνουν την ακρίβεια των αποτελεσμάτων, όμως αυξάνουν το υπολογιστικό κόστος. Στην αντίθετη περίπτωση, κελιά μεγάλης επιφάνειας μειώνουν την ακρίβεια, τόσο τη χωροχρονική όσο και τη στατιστική, μειώνοντας παράλληλα και το υπολογιστικό κόστος.

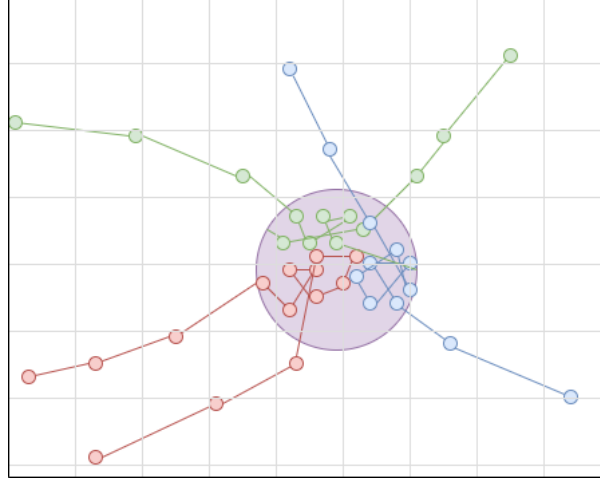


Fig. 8. Συστάδα από "κουβάρια" τροχιών.

Γενικά, η πολυπλοκότητα του THS είναι $O(n^2)$, έτσι οι ερευνητές πρότειναν μια προσεγγιστική λύση του αρχικού αλγορίθμου, τον aTHS, μειώνοντας την πολυπλοκότητα. Πιο συγκεκριμένα, ο aTHS δέχεται μια επιπλέον παράμετρο h η οποία ορίζει το σύνολο των γειτονικών κελιών που θα ληφθούν υπόψιν στο υπολογισμό της στατιστικής G_i^* . Στην περίπτωση του πεδίου της εργασίας αυτής, δηλαδή της θαλάσσια επιτήρησης, η επιλογή της παραμέτρου h είναι δύσκολη λόγω της φύσης των χωρικά αραιών δεδομένων και της κλίμακας του προβλήματος.

4 Spatio-Temporal Clustering

Στην εργασία αυτή υλοποιήθηκε μια γενικευμένη μέθοδος εξαγωγής συστάδων στο χωροχρονικό πεδίο, δίνοντας έμφαση στη διάσταση του χρόνου, δίχως τη χρήση κάποιου πλέγματος, Spatio-Temporal Clustering (STC). Ο αναλυτής έχει τη δυνατότητα να ερμηνεύσει αυτές τις συστάδες ανάλογα με το πρόβλημα και τη μετα-ανάλυση που θα διεξάγει. Πριν τη παρουσίαση της μεθόδου, έχει νόημα να διατυπωθεί η μοντελοποίηση του προβλήματος.

4.1 Στατιστική σημασία χωρίς στατιστική

Η αρχική ιδέα του STC προήλθε από τις τεχνικές δυσκολίες ένταξης της στατιστικής σημαντικότητας και, ακολούθως, της ερμηνείας της στις συστάδες των πλοίων για τους λόγους που αναφέρθηκαν προηγουμένως. Η προτεινόμενη μέθοδος εξάγει τις ζητούμενες συστάδες και αποδίδει μια ψευδο-στατιστική σημαντικότητα, χρησιμοποιώντας την έννοια της χωρικής αυτοσυσχέτισης, χωρίς, πρακτικά, να χρησιμοποιεί στατιστική! Δεδομένης της ολικής χωροχρονικής τυχαιότητας, οι συστάδες που δημιουργούνται στο χωροχρόνο αναμένεται να είναι στατιστικά σημαντικές. Πιο συγκεκριμένα, η παρούσα ιδέα βασίζεται στη λογική του THS και της

χωρικής αυτοσυσχέτισης, όπου επιπλέον χρησιμοποιούνται τεχνικές συσταδοποίησης και διαμέρισης του χρόνου.

Στην περίπτωση του THS, ένα βραχυπρόθεσμο αγκυροβόλιο θα φαινόταν όπως στο Figure 9. Να σημειωθεί πως οι παρακάτω εικόνες μπορούν να θεωρηθούν ως διάφορες καταστάσεις του παραδείγματος του Figure 6, όμως σε χαμηλή ευκρίνεια για απλοποίηση του προβλήματος. Η κόκκινη περιοχή υποδηλώνει υψηλή τιμή x_i , ενώ η λευκή την απουσία της - μηδέν. Σε αυτό το παράδειγμα, ο THS δεν είναι ικανός να αναγνωρίσει την περιοχή αυτή ως στατιστικά σημαντική, λόγω της φύσης του Getis-Ord Gi*. Αξίζει να σημειωθεί πως το κόκκινο κελί αποτελεί outlier (βλ. Fig. 9). Αντιθέτως, ένας λιμένας ή μια μαρίνα αναμένεται να έχει την μορφή όπως φαίνεται στο Figure 10. Στο τελευταίο παράδειγμα, παρατηρείται μια συστάδα στη χρονική διάσταση και αναμένεται να αναγνωρισθεί από τον THS και να χαρακτηρισθεί ως στατιστικά σημαντική.

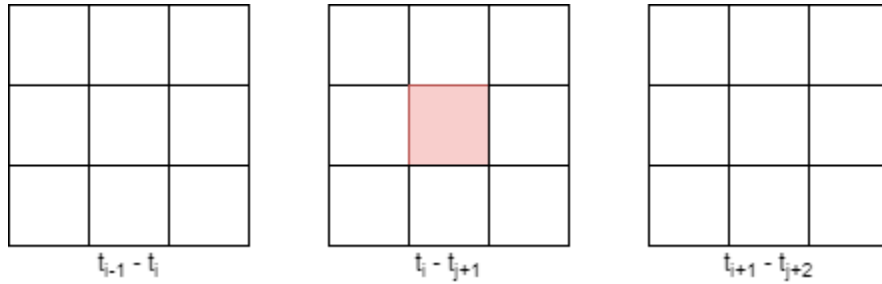


Fig. 9. Αφηρημένη απεικόνιση ενός βραχυπρόθεσμου αγκυροβολίου.

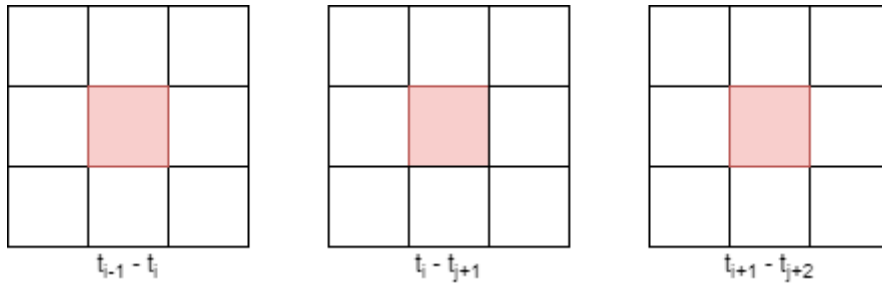


Fig. 10. Αφηρημένη απεικόνιση ενός μακροπρόθεσμου αγκυροβολίου.

4.2 Μοντελοποίηση

Το πρόβλημα της εργασίας, λοιπόν, επικεντρώνεται στον εντοπισμό των βραχυπρόθεσμων αγκυροβολίων (outliers) και των μακροπρόθεσμων αγκυροβολίων (pseudo-hot spots/pseudo-statistically significant). Για αυτόν το λόγο σχεδιάστηκε η παρακάτω μοντελοποίηση.

Οι τροχιές των πλοίων εμφανίζουν ορισμένα χαρακτηριστικά όταν βρίσκονται σε πιθανές περιοχές αγκυροβολίων, όπως για παράδειγμα μειωμένη ταχύτητα, γρήγορη μεταβολή πορείας, εξαφάνιση του σήματος για κάποια χρονικό διάστημα, κλπ. Έχει νόημα να αγνοήσουμε το μεγαλύτερο μέρος της τροχιάς και να κρατήσουμε τα σημεία ενδιαφέροντος, δηλαδή εκείνα τα σημεία της τροχιάς όπου το πλοίο φαίνεται να μπαίνει και να βγαίνει στο αγκυροβόλιο, ή διαφορετικά το «κουβάρι» που περιεγράφηκε προηγουμένως.

Πιο συγκεκριμένα, υποθέτουμε ότι $\tau_i = (p_{it_1}, \dots, p_{it_j}, \dots, p_{it_{a_i}})$ μια τροχιά του πλοίου i , όπου $p_{it_j} = (c_{it_j}, a_{it_j})$ οι πληροφορίες του πλοίου τη χρονική στιγμή t_j , c_{it_j} οι συντεταγμένες και a_{it_j} ένα σύνολο χαρακτηριστικών του πλοίου. Θέτουμε p_{it_s} και p_{it_e} τα σημεία στάσης, δηλαδή τα σημεία εισόδου και εξόδου του πλοίου από το πιθανό αγκυροβόλιο και m_i το μέσο των σημείων αυτών. Διαφορετικά, ορίζουμε $f(\tau_i)$ ως την συνάρτηση εξαγωγής των σημείων $(p_{it_s}, p_{it_e}, m_i)$, βάσει των χαρακτηριστικών a_i .

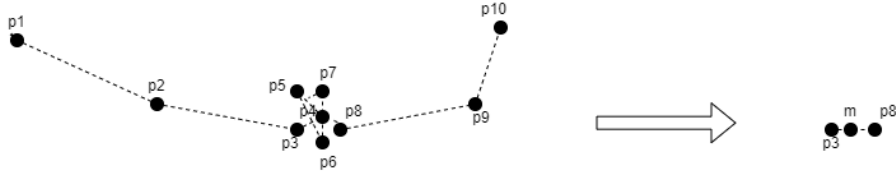


Fig. 11. Επιλογή σημείων εισόδου, εξόδου και μέσου από μια τροχιά.

Με την παραπάνω διαδικασία μειώνεται σε μεγάλο ποσοστό το πλήθος των δεδομένων, όμως διατηρείται η χρήσιμη πληροφορία. Αξίζει να σημειωθεί πως η διαδικασία αυτή αποτελεί σημαντικό βήμα στην προπαρασκευή των δεδομένων και την ερμηνεία των αποτελεσμάτων των αλγορίθμων που θα παρουσιαστούν. Για παράδειγμα, μια διαφορετική συνάρτηση εξαγωγής θα μπορούσε να έχει ως έξοδο το (p_1, p_{10}, m') , επηρεάζοντας την ποιότητα των αποτελεσμάτων. Παρόλα αυτά, η f εξαρτάται από το πεδίο του προβλήματος.

Έστω μια χρονική διαμέριση $T = \{t_1, \dots, t_j, \dots, t_v\}$ με $t_{j+1} - t_j \neq 0$ και $G = \{g_1, \dots, g_i, \dots, g_v\}$, όπου $g_i = (m_i, t_{is}, t_{ie})$. Επιπλέον, θέτουμε:

$$G_j = \{g_i \in G: t_{is} \leq t_{j+1} \text{ and } t_{ie} \geq t_j\}.$$

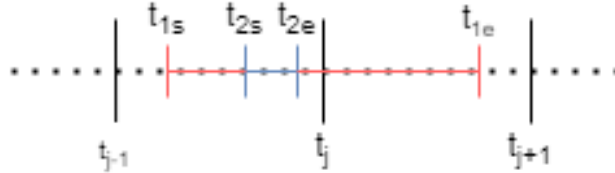


Fig. 12. Χρονική διαμέριση και τοποθέτηση των χρόνων εισόδου και εξόδου

Στη συνέχεια, εκτελείται ο αλγόριθμος συσταδοποίησης για κάθε G_j και εξάγονται οι χωρικές συστάδες ως $K_j = \{K_{j1}, \dots, K_{jk}, \dots, K_{ja_j}\}$ και τα αντίστοιχα κεντροειδή $U_j = \{u_{j1}, \dots, u_{jk}, \dots, u_{ja_j}\}$, ενώ αγνοούμε τον θόρυβο.

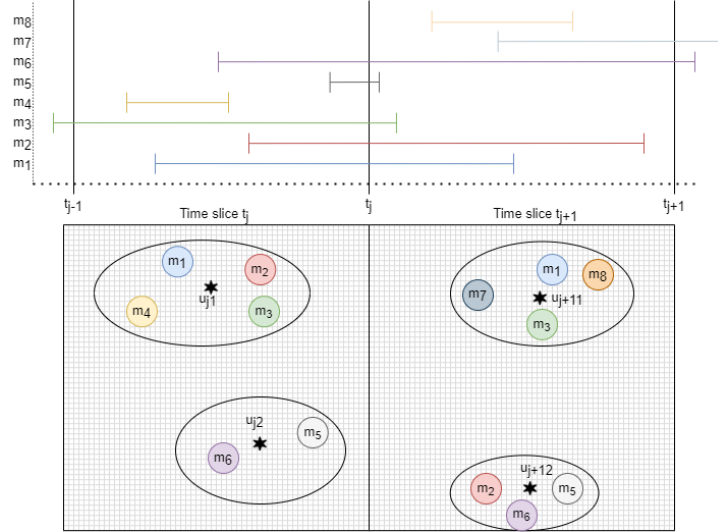


Fig. 13. Συσταδοποίηση 8 μέσων σημείων m_i , σε δυο διαδοχικές χρονικές φέτες t_j και t_{j+1} και εξαγωγή των centroids u_{jk}

Επιλέγεται $1 \leq w \leq v$ το μέγεθος του χρονικού παραθύρου (time window). Εκτελείται εκ νέου ο αλγόριθμος συσταδοποίησης για κάθε σύνολο

$$V_j = \bigcup_{z=j}^{j+w-1} U_z, \quad j = 1, \dots, v - w$$

και λαμβάνονται οι συστάδες ως σύνολο R_j και ο θόρυβος ως σύνολο N_j . Το κεντροειδές μοντελοποιεί την ύπαρξη της συστάδας μέσα στη διάσταση του χρόνου και όχι το χωρικό μέγεθός της ή το πλήθος των σημείων της (βλ. Fig. 14). Βάσει αυτού, προκύπτει ο χαρακτηρισμός «βραχυπρόθεσμο» ή «μακροπρόθεσμο».

Σε αυτό το σημείο, η μέθοδος μας δίνει τη δυνατότητα να ερμηνεύσουμε δύο σύνολα (σύνολο συστάδων και σύνολο θορύβου/outlier) με διαφορετικό τρόπο. Το σύνολο R_j αποτελείται από μακροπρόθεσμα αγκυροβόλια, ενώ το N_j αποτελείται από τα βραχυπρόθεσμα αγκυροβόλια, δηλαδή τα αγκυροβόλια που εμφανίστηκαν για σχετικά μικρό χρονικό διάστημα, π.χ. αγκυροβόλια σκαφών στη Μύκονο την περίοδο του καλοκαιριού. Εύλογα, θα μπορούσε να διατυπωθεί ο ισχυρισμός ότι δεν παρέχεται ο τρόπος υπολογισμού της στατιστικής σημαντικότητας. Όμως, έχοντας υπόψιν το παράδειγμα που παρουσιάστηκε στην αρχή της ενότητας με τον THS και τη σημασία της χωρικής αυτοσυσχέτισης, μπορούμε να πούμε πως η στατιστική σημαντικότητα

ενθυλακώνεται στη χρονική διαμέριση και τη δεύτερη συσταδοποίηση. Με άλλο λόγια, αν υπολογίσουμε το *Getis-Ord* G_i^* ανά w χρονικές φέτες και υπολογιστούν τα z-scores, αναμένεται να ληφθούν τα ίδια αποτελέσματα (σύνολο R_j) ως στατιστικά σημαντικά. Μένει μόνο η σχετική απόδειξη, μιας και η παραπάνω υπόθεση αποτελεί προϊόν συλλογιστικής σκέψης.

$$w=3$$

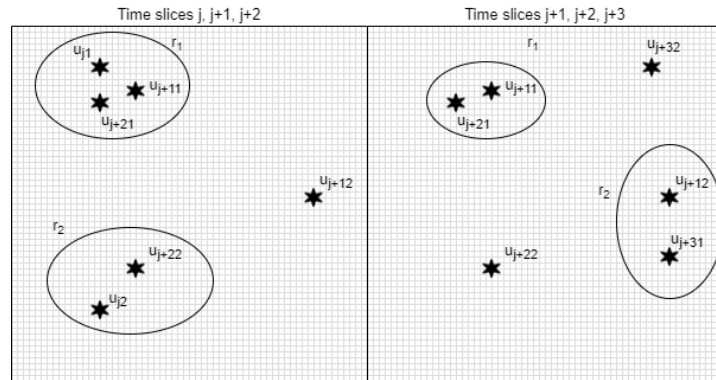


Fig. 14. Συσταδοποίηση των κεντροειδών των συστάδων-αγκυροβολίων, υποδηλώνει την ύπαρξη του αγκυροβολίου κατά τη διάρκεια του χρονικού παραθύρου.

Ενδεικτικά, ο αλγόριθμος είναι:

```

Input
-----
data: Set of trajectories
timeslice_size: Size of the time partition
w: Size of the time window
-----

[1] G = extract_trajectory_centroids(data)
[2] T = [G.min_timestamp + i*timeslice_size for i in range(0,
G.max_timestamp/timeslice_size)]
[3] U = []
[4] for j in range(0, T.length):
[5]     Gj = SELECT g FROM G
           WHERE g.s <= T[j+1]
           AND g.e >= T[j]
[6]     Kj = clustering_algorithm_without_noise(Gj)
[7]     U.append(calculate_centroids(Kj))
    end loop
[8] R = []
[9] N = []
[10] for j in range(0, T.length - w + 1):
[11]     Vj = get_union(U[i] for j in range(i, i + w))
[12]     Rj, Nj = clustering_algorithm_with_noise(Vj)
[13]     R.append(Rj)
[14]     N.append(Nj)

```

```

        end loop
[15] return R, N

```

Ο αλγόριθμος δέχεται ως είσοδο ένα σύνολο τροχιών (data), το μέγεθος της χρονικής διαμέρισης (timeslice_size) και το πλήθος των χρονικών διαμερίσεων που θα συγχωνευθούν κατά την δεύτερη συσταδοποίηση (w). Αρχικά, στο βήμα 1 πραγματοποιείται εξαγωγή των σημείων στάσης των τροχιών και ο υπολογισμός των μέσων σημείων m . Η συνάρτηση αυτή εξαρτάται από το πεδίο και αφήνεται στην κρίση του αναλυτή. Στα βήματα 2 και 3, δημιουργείται η χρονική διαμέριση και αρχικοποιείται λίστα των μέσων των συστάδων, αντίστοιχα. Στη συνέχεια, για κάθε χρονική φέτα, επιλέγονται τα g τα οποία ανήκουν στην αντίστοιχη χρονική περίοδο (βήμα 5), ως σύνολο G_j και εκτελείται αλγόριθμος συσταδοποίησης για το σύνολο αυτό, αγνοώντας τον «θόρυβο» (βήμα 6). Στο τελευταίο βήμα του βρόγχου, υπολογίζονται τα κεντροειδή των συστάδων που εντοπίστηκαν και προστίθενται στο σύνολο U .

Στο δεύτερο σκέλος του *STC*, αρχικοποιούνται οι λίστες των «μακροπρόθεσμων» και των «βραχυπρόθεσμων» αγκυροβολίων, R και N αντίστοιχα. Έπειτα, δημιουργείται επαναληπτικά η ένωση των κεντροειδών w διαδοχικών χρονικών διαμερίσεων V_j (βήμα 11) και εκτελείται αλγόριθμος συσταδοποίησης (βήμα 12). Αυτή τη φορά ο «θόρυβος» διατηρείται και προστίθεται στη λίστα N , όπως και οι συστάδες στην αντίστοιχη λίστα R .

4.3 Μετα-ανάλυση αποτελεσμάτων

Έχοντας λάβει τα δύο σύνολα από τη μέθοδο, ο αναλυτής έχει τη δυνατότητα να εκτελέσει, επιπλέον, άλλες αναλύσεις και να λάβει τα ζητούμενα αποτελέσματα. Για παράδειγμα, τα σύνολα R_j και N_j τις συστάδες και τον θόρυβο των g_i , όμως το ζητούμενο μπορεί να είναι ο σχηματισμός των περιοχών. Στην περίπτωση του πρώτου συνόλου, μπορεί να εκτελέσει κάποιον αλγόριθμο παραγωγής convex/concave hull και να λάβει μια σχετικά φτωχή προσέγγιση της περιοχής, λόγω του τρόπου υπολογισμού των centroids, ο οποίος μειώνει την πληροφορία. Μια άλλη, ακριβέστερη αλλά ακριβότερη λύση, αποτελεί η δημιουργία Heatmap του συνόλου V_j . Όμως, για να έχει νόημα η απεικόνιση, απαιτείται παράλληλη απεικόνιση των R_j και N_j .

Επιπλέον, μια μετα-ανάλυση είναι την ταξινόμηση των αρχικών σημείων m ως προς τα εντοπισμένα αγκυροβόλια. Η ταξινόμηση αυτή θα έχει ως αποτέλεσμα την ανάδειξη των περιοχών τους και μπορεί να επιτευχθεί με τη χρήση του αλγορίθμου K-Nearest Neighbors [15]. Πιο συγκεκριμένα, τα σημεία m ταξινομούνται με την βοήθεια K γειτόνων από το σύνολο $R_j \cup N_j$ και τελικά υπολογίζονται τα αντίστοιχα convex/concave hulls.

Τέλος, υπάρχει δυνατότητα συγχώνευσης των συστάδων των συνόλων R_j για $j = 1, \dots, v - w$, έτσι ώστε να παραχθεί ένα τελικό σύνολο συστάδων R το οποίο αποτυπώνει όλες τις συστάδες. Αξίζει να σημειωθεί πως οι μετα-αναλύσεις δεν περιορίζονται στις προαναφερθείσες, όμως δεν θα παρουσιαστούν και αναλυθούν περεταίρω στο παρόν κείμενο.

4.4 Πλεονεκτήματα και μειονεκτήματα του αλγορίθμου

Το μεγαλύτερο πλεονέκτημα του STC μπορεί να θεωρηθεί η δυνατότητα ανάλυσης των αραιών δεδομένων, ανεξαρτήτως κλίμακας, ένα πρόβλημα το οποίο συναντάται στις αναλύσεις χωρικής αυτοσυσχέτισης. Η προτεινόμενη μέθοδος είναι εύκολο να παραμετροποιηθεί ανάλογα με το υπό εξέταση πρόβλημα, τόσο στα δεδομένα εισόδου, όσο και στις διαδικασίες που πραγματοποιούνται, αντίστοιχα. Για παράδειγμα, η επιλογή του αλγορίθμου συσταδοποίησης είναι ελεύθερη, αλλά εξαρτάται και από τη φύση του προβλήματος. Το ίδιο ισχύει και για την επιλογή των αντίστοιχων παραμέτρων, οι οποίοι ερμηνεύονται με ευκολία και δεν αποτελούν προϊόν τυχαίας επιλογής. Το παραπάνω γίνεται κατανοητό στην ενότητα 9 όπου παρουσιάζονται παραδείγματα με τη χρήση του DBSCAN. Επιπλέον, το πρώτο στάδιο, δηλαδή η εξαγωγή των p_{it_s} και p_{it_e} μπορεί να διαφέρει ανάλογα με το πρόβλημα. Σίγουρα η αποφυγή χρήσης κάποιου χωροχρονικού πλέγματος, συνεπάγεται τη μείωση του κόστους μνήμης κατά την εκτέλεση, όμως μένει να μελετηθεί το κόστος της προτεινόμενης μεθόδου. Ενδεικτικά, το βήμα 1 του αλγορίθμου έχει κόστος $O(n)$ όπου n το πλήθος των σημείων. Το κόστος των βημάτων 4 - 15 εξαρτάται σε μεγάλο βαθμό από τη χρονική διαμέριση αλλά και από τα διαθέσιμα ευρετήρια και τους αλγορίθμους συσταδοποίησης. Τέλος, υπάρχει δυνατότητα υλοποίησης σε κατανεμημένα συστήματα λόγω της χρονικής διαμέρισης, μειώνοντας σημαντικά το κόστος υπολογισμού.

Όπως αναφέρθηκε, οι συστάδες χαρακτηρίζονται από «ψευδο-στατιστική» σημαντικότητα, οπότε δεν αποδεικνύεται από τα παραπάνω αν όντως και σε ποιο βαθμό ισχύει η στατιστική σημασία. Επιπλέον, τα αποτελέσματα εξαρτώνται σε μεγάλο βαθμό από τους αλγορίθμους συσταδοποίησης και τις παραμέτρους τους. Παράλληλα, όμως, καθορίζουν και την ερμηνεία των αποτελεσμάτων (βλ. Ενότητα 9).

Σε αυτό το σημείο θα μπορούσε να γίνει μια σύγκριση με τη μέθοδο MaSEC [13] και συγκεκριμένα με τη διαδικασία εντοπισμού αγκυροβολίων stationary ECs. Η εν λόγω διαδικασία θα μπορούσε να θεωρηθεί ως εναλλακτική μέθοδος για το πρώτο στάδιο του STC, δηλαδή της συσταδοποίησης των συνόλων G_j . Εν αντιθέσει με τα stationary ECs, που περιορίζονται στον εντοπισμό των αγκυροβολίων, η προτεινόμενη μέθοδος δύναται να χαρακτηρίσει επιπλέον τα αγκυροβόλια ως μακροπρόθεσμα ή βραχυπρόθεσμα, αντίστοιχα.

5 Spatio-Temporal Hot Spot

Χρησιμοποιώντας την ίδια φιλοσοφία με τον αλγόριθμο STC, έγινε προσπάθεια εισαγωγής της πραγματικής στατιστικής σημαντικότητας στη λύση. Ο αλγόριθμος Spatio-Temporal Hot Spot (STHS) εντοπίζει τις στατιστικές σημαντικές συστάδες (hot/cold spots) με τη βοήθεια της στατιστικής G_i σε κάθε χρονικό παράθυρο (time window). Πιο συγκεκριμένα, ο STHS αποτελεί μερική παραλλαγή του STC, αφού αντί για τη συσταδοποίηση των V_j , πραγματοποιείται ανάλυση στατιστικής σημαντικότητας στο σύνολο αυτό.

5.1 Αυτοσυσχέτιση στο χωροχρονικό πεδίο

Στις χωροχρονικές στατιστικές αναλύσεις κυρίαρχο ρόλο έχει η χωρική αυτοσυσχέτιση. Όπως αναφέρθηκε στην ενότητα 2, η αυτοσυσχέτιση των δεδομένων βασίζεται σε κάποιο χαρακτηριστικό τους, για παράδειγμα στο πλήθος των οδικών ατυχημάτων των περιφερειών μιας χώρας. Στην περίπτωση του προβλήματος αυτής της εργασίας, το χαρακτηριστικό ενδιαφέροντος είναι το πλήθος των πλοίων σε ένα αγκυροβόλιο. Έτσι, η χωρική αυτοσυσχέτιση μπορεί να τεθεί σε μορφή ερώτησης: Πόσο πολύ και με ποιο τρόπο (θετικά ή αρνητικά) το πλήθος των πλοίων σε ένα αγκυροβόλιο επηρεάζει, αντίστοιχα, το πλήθος των πλοίων στα γειτονικά αγκυροβόλια ως προς το χώρο (απόσταση) και το χρόνο (χρονικά διαστήματα); Για παράδειγμα, η ύπαρξη ενός «διάσημου» αγκυροβολίου πιθανότατα να μειώνει το πλήθος των πλοίων στα γειτονικά αγκυροβόλια με την πάροδο του χρόνου.

Σημαντικό σημείο πριν την ανάλυση αποτελεί ο ορισμός του τρόπου με τον οποίο «συνδέονται» τα χαρακτηριστικά των δεδομένων. Σύμφωνα με το Πρώτο νόμο της Γεωγραφίας του Tobler, «Τα πάντα σχετίζονται μεταξύ τους, αλλά τα πιο κοντινά σχετίζονται περισσότερο από ότι τα απομακρυσμένα». Θα μπορούσε κανείς να μοντελοποιήσει τον νόμο αυτό μέσω μιας συνάρτησης εξασθένησης ως προς την απόσταση. Όμως, στο πρόβλημα αυτό δεν υπάρχει μόνο η χωρική απόσταση, αλλά και η χρονική, αφού στόχος είναι η εύρεση των hot spots στο σύνολο $V_j = K_j \cup K_{j+1} \cup \dots \cup K_{j+w-1}$, όπου K_j οι συστάδες της χρονικής φέτας j . Ποια είναι όμως η απόσταση μεταξύ των συστάδων K_j και K_{j+1} ;

Σύμφωνα με τα παραπάνω, ορίζεται η χωροχρονική εξασθένηση ως:

$$w_{z,a,b} = \frac{m}{e^{(z+1) \cdot \varphi(K_{xa}, K_{yb})}}, \quad z = |x - y|$$

όπου $K_{xa}, K_{yb} \in V_j$ δυο συστάδες με $x, y \in \{j, j+1, \dots, j+w-1\}$, $\varphi: V_j \times V_j \rightarrow R^+$ μια μετρική μεταξύ των συστάδων, z η απόλυτη διαφορά των δεικτών των χρονικών διαμερίσεων των συστάδων και $m > 0$ η παράμετρος που εξαρτάται από την κλίμακα των δεδομένων μελέτης. Η παραπάνω εξασθένηση αποτελεί παραμετροποίηση της αρνητικής εκθετικής εξασθένησης. Στην εικόνα 15 φαίνεται η γραφική απεικόνιση της εξασθένησης για διάφορες τιμές του z , δηλαδή για διαφορετικές χρονικές αποστάσεις και $m = 1$. Διαισθητικά, αν μια συστάδα δεν μεταβάλλει την τοποθεσία της κατά τη διάρκεια του χρόνου ($x = 0$ στο Fig. 15), τότε θα έχει υψηλή επιρροή. Οι πιο απομακρυσμένες συστάδες, τόσο χωρικά όσο και χρονικά, επηρεάζουν όλο και λιγότερο την πρώτη. Στο Figure 16 παρουσιάζεται ένα σχετικό παράδειγμα. Οι συστάδες A, C, E και B, D, F αντιστοιχούν σε δύο διαφορετικά αγκυροβόλια για τις χρονικές φέτες $j, j+1$, και $j+2$. Έτσι, η συστάδα A επηρεάζεται από τις υπόλοιπες με βαθμό εξασθένησης w , όπως φαίνεται στις ακμές.

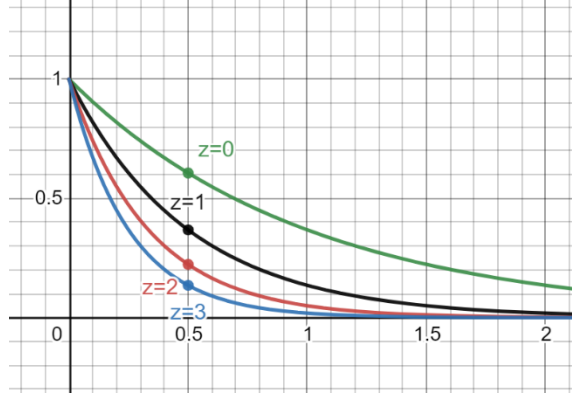


Fig. 15. Γραφική απεικόνιση της εξασθένησης για διάφορες τιμές του z και $m = 1$.

5.2 Μοντελοποίηση

Ακολουθώντας τα ίδια βήματα με τον STC, δημιουργούνται τα σύνολα V_j για $j = 1, \dots, v - w$. Στη συνέχεια, για κάθε συστάδα στο V_j υπολογίζεται το αντίστοιχο Getis-Ord G_i^* . Έχοντας ορίσει το χαρακτηριστικό ενδιαφέροντος αλλά και την εξασθένηση στην προηγούμενη ενότητα, είναι δυνατός ο ορισμός της στατιστικής Getis-Ord G_i^* για κάθε συστάδα $K_{\chi\alpha}$ του συνόλου V_j ως εξής:

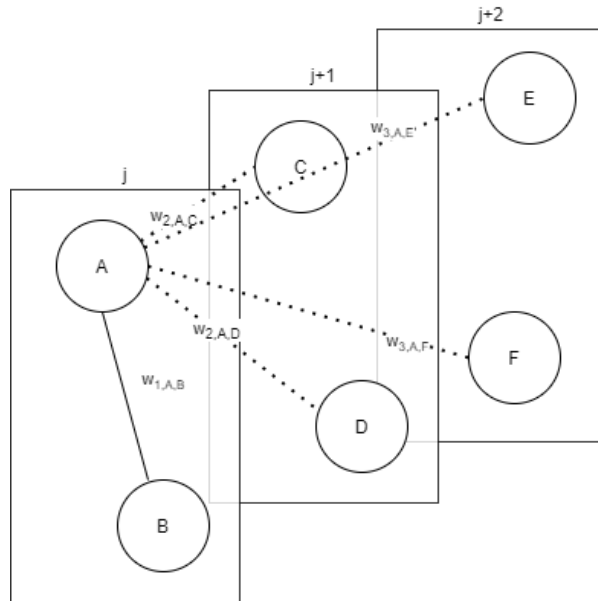


Fig. 16. Απεικόνιση χωροχρονικής εξασθένησης.

$$G_{xa}^* = \frac{\sum_{K_{yb} \in V_j} w_{z,a,b} \cdot \ddot{x}_a - \bar{X} \sum_{K_{yb} \in V_j} w_{z,a,b}}{S \sqrt{\frac{n \sum_{K_{yb} \in V_j} w_{z,a,b}^2 - \left(\sum_{K_{yb} \in V_j} w_{z,a,b}\right)^2}{n-1}}}$$

όπου

$$\begin{aligned} \ddot{x}_a &= |K_{xa}|, \quad n = |V_j|, \quad x, y \in \{j, \dots, j+w-1\} \\ w_{z,a,b} &= \frac{1}{m} e^{(z+1) \cdot \phi(K_{xa}, K_{yb})}, \quad z = |x - y| \\ \bar{X} &= \frac{\sum_{K_{xa} \in V_j} \ddot{x}_a}{n} \\ S &= \sqrt{\frac{\sum_{K_{xa} \in V_j} \ddot{x}_a^2}{n} - (\bar{X})^2} \end{aligned}$$

Ο χαρακτηρισμός των συστάδων ως hot/cold spot ή μη-στατιστικά σημαντικό από τα προηγούμενα προκύπτει από την τιμή της στατιστικής. Το G_i^* αποτελεί z-score, έτσι επιλέγοντας το επίπεδο εμπιστοσύνης α διακρίνονται οι περιπτώσεις που παρουσιάζονται στο Table 3. Πρέπει να σημειωθεί, πως επειδή το Getis-Ord αποτελεί z-score και δεν είναι γνωστό αν η τυχαία μεταβλητή ακολουθεί την κανονική μεταβλητή, το πλήθος των συστάδων στο σύνολο V_j θα πρέπει να είναι μεγαλύτερο από 30, ώστε τα αποτελέσματα να είναι έγκυρα. Για αυτό το λόγο, πρέπει να γίνει προσεκτική επιλογή των παραμέτρων του αλγορίθμου, όπως το time window w και το χρονικό βήμα της διαμέρισης.

Table 3. Χαρακτηρισμός συστάδων.

Συνθήκη	Χαρακτηρισμός
$G_i^* \geq \Phi\left(\frac{\alpha}{2}\right)$	Hot spot
$G_i^* \leq -\Phi\left(\frac{\alpha}{2}\right)$	Cold spot
$-\Phi\left(\frac{\alpha}{2}\right) < G_i^* < \Phi\left(\frac{\alpha}{2}\right)$	-

Ενδεικτικά, ο αλγόριθμος είναι:

```

Input
-----
data: Set of trajectories
timeslice_size: Size of the time partition
w: Size of the time window
a: significance level
-----

[1] G = extract_trajectory_centroids(data)
[2] T = [G.min_timestamp + i*timeslice_size for i in range(0,
G.max_timestamp/timeslice_size)]
[3] K = []
[4] for j in range(0, T.length):

```

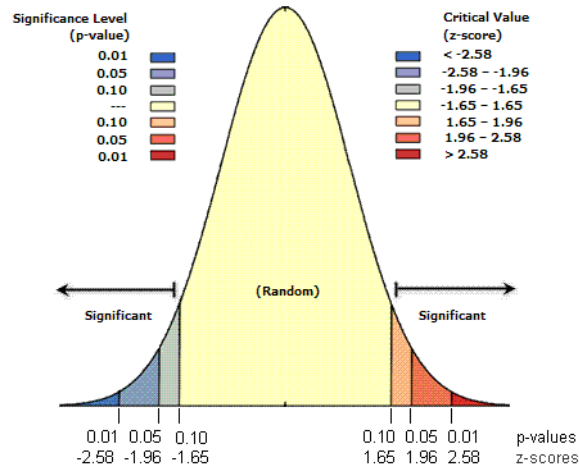


Fig. 17. Χαρακτηρισμός συστάδων για διάφορα επίπεδα εμπιστοσύνης. <https://pro.arcgis.com/>

```

[5]     Gj = SELECT g FROM G
          WHERE g.s <= T[j+1]
          AND g.e >= T[j]
[6]     Kj = clustering_algorithm_without_noise(Gj)
[7]     K.append(Kj)
    end loop
[8]     V = []
[9]     for j in range(0, T.length - w + 1):
[10]        Vj = get_union(K[i] for j in range(i, i + w))
[11]        n = Vj.length
[12]        x_bar = sum(k.x for k in Vj)
[13]        S = sqrt(sum(k.x^2 for k in Vj) / n - x_bar^2)
[14]        for ka in Vj:
[15]            swx, sw, sw2 = 0
[16]            for kb in Vj:
[17]                z = |ka.t - kb.t|
[18]                w = weight(ka, kb)
[19]                swx += w * kb.x
[20]                sw += w

[21]            sw2 += w^2
        end loop
[22]        ka.gi = (swx - x_bar * sw) / (S * sqrt((n * sw2 - sw^2) / (n - 1)))
[23]        if ka.gi <= -Φ(a/2):
[24]            ka.spot = 'Cold'
[25]        else if ka.gi >= Φ(a/2):
[26]            ka.spot = 'Hot'
        end if
    end loop
[27]     V.append(Vj)
    end loop
[28]     return V

```

Ο αλγόριθμος STHS δέχεται ως είσοδο τις ίδιες παραμέτρους με τον αλγόριθμο STC, συν της παραμέτρου a , δηλαδή το επίπεδο εμπιστοσύνης. Τα βήματα 1 – 10 είναι ίδια με τα αντίστοιχα βήματα του STC, δηλαδή η διαδικασία εντοπισμού αγκυροβολίων.

Το δεύτερο σκέλος του αλγορίθμου είναι υπεύθυνο για τον στατιστικό χαρακτηρισμό των αγκυροβολίων. Πιο συγκεκριμένα, στα βήματα 11 – 13 υπολογίζονται στατιστικές τιμές που αφορούν όλα τα αγκυροβόλια του V_j . Στα βήματα 14 – 21 υπολογίζονται επιπλέον στατιστικές τιμές που αφορούν ανά δύο τα αγκυροβόλια του συνόλου V_j , ενώ στο βήμα 22 υπολογίζεται το Getis-Ord G_i^* του αγκυροβολίου που εξετάζεται. Στα βήματα 23 – 26 γίνεται ο στατιστικός χαρακτηρισμός των αγκυροβολίων ως Hot spot, Cold spot ή τίποτα, βάσει της τιμής G_i^* και του επιπέδου εμπιστοσύνης. Τέλος, επιστρέφονται τα χαρακτηρισμένα αγκυροβόλια, για κάθε ένωση χρονικών διαμερίσεων.

5.3 Σύγκριση Spatio-Temporal Hot Spot και Trajectory Hot Spot

Η βασική διαφορά των δύο αλγορίθμων, του Spatio-Temporal Host Spot (STHS) και του Trajectory Hot Spot (THS), αντίστοιχα, εντοπίζεται στη χρήση τους. Στην πρώτη περίπτωση, το ενδιαφέρον εστιάζεται στον εντοπισμό και των χαρακτηρισμό των αγκυροβολίων, ενώ στην περίπτωση του STHS αφενός γίνεται εντοπισμός αγκυροβολίων, αφετέρου λαμβάνεται υπόψιν η στατιστική σημασία τους ως προς το πλήθος των πλοίων. Μια επιπλέον διαφορά είναι ο ορισμός του χαρακτηριστικού ενδιαφέροντος. Στην περίπτωση του αλγορίθμου STHS, το χαρακτηριστικό ενδιαφέροντος είναι το πλήθος των πλοίων της συστάδας, ενώ στον THS είναι η συνολική χρονική διάρκεια που παρέμεινε κάποιο πλοίο μέσα στο κελί. Επιπρόσθετα, η οντότητα της συστάδας διαφέρει, εννοιολογικά, σε σχέση με το κελί. Μια συστάδα μπορεί να μοιράζεται σε διαφορετικά κελιά μιας διαμέρισης, συναρτήσει πάντα των αντίστοιχων παραμέτρων και δεδομένων. Έτσι, δεν είναι δυνατό να γίνει συζήτηση περί συσχετισμού αγκυροβολίων αλλά περί συσχετισμού κελιών.

Από την πλευρά του κόστους, δεν υπάρχει ξεκάθαρη γραμμή διαχωρισμού ως προς τη βέλτιστη λύση. Η επιλογή των παραμέτρων και στις δύο περιπτώσεις καθορίζουν σημαντικά την πολυπλοκότητα των αλγορίθμων. Σίγουρα, στον STHS ο υπολογισμός του G_i^* για το V_j έχει πολυπλοκότητα $O(n^2)$, όπου n το πλήθος των συστάδων στο V_j και η διαδικασία επαναλαμβάνεται $n - w$ φορές. Στην περίπτωση του aTHS η επιλογή της παραμέτρου h , δηλαδή η μέγιστη απόσταση-πλήθος κελιών που θα συμπεριληφθούν στον υπολογισμό του G , θα πρέπει να είναι $3 \leq h_0 \leq h$, ώστε να ικανοποιηθεί η συνθήκη των 30 συνολικών χαρακτηριστικών στο σύνολο δεδομένων (3) και να καλύπτει ένα ικανοποιητικό πλήθος διαφορετικών συστάδων στα κελιά αυτά (h_0 , το πρόβλημα της κλίμακας).

6 Υλοποίηση και παραδείγματα

Οι παραπάνω αλγόριθμοι υλοποιήθηκαν σε Python 3.8, ενώ παράλληλα δημιουργήθηκαν οι απαραίτητες συναρτήσεις για τη δημιουργία τυχαίων δεδομένων και την οπτικοποίηση των αποτελεσμάτων. Ως αλγόριθμος συσταδοποίησης

επιλέχθηκε η υλοποίηση DBSCAN2 της scikit-learn βιβλιοθήκης. Πιο συγκεκριμένα, δημιουργήθηκε η βασική κλάση ST η οποία υλοποιεί τις κοινές λειτουργίες των STC και STHS αλγορίθμων, δηλαδή τη δημιουργία της χρονικής διαμέρισης και τη συσταδοποίηση των μέσων σημείων.

Επιπλέον, δημιουργήθηκαν οι κλάσεις STC και STHS (spatio_temporal.py) οι οποίες κληρονομούν την κλάση ST και εκτελούν τους αντίστοιχους αλγορίθμους. Για τις ανάγκες των παραπάνω κλάσεων και τη διαχείριση των δεδομένων, υλοποιήθηκαν οι κλάσεις μοντέλων δεδομένων Point, TPoint, ClusterPoint, Trajectory, TPartition και GetisCluster (models.py). Τέλος, υπάρχουν βοηθητικά αρχεία για την προ-επεξεργασία των δεδομένων AIS (preprocessing.py). Λόγω του μεγάλου όγκου των δεδομένων, προτείνεται η χρήση των δεδομένων ms.csv, τα οποία είναι το αποτέλεσμα της εν λόγω διαδικασίας. Ο κώδικας μπορεί να βρεθεί στο συνοδευτικό φάκελο ή στο Github3 με πλήρη τεκμηρίωση και σχόλια.

Να σημειωθεί πως οι παρακάτω εικόνες αποτελεσμάτων των STC και STHS ενδέχεται να μην προβάλλουν τις ίδιες χρονικές περιόδους, ώστε να αναδειχθούν οι πιο ενδιαφέρουσες από άποψη ερμηνείας και σύγκρισης καταστάσεις.

6.1 STC και STHS σε δεδομένα AIS

Οι αλγόριθμοι δοκιμάστηκαν χρησιμοποιώντας ένα σύνολο δεδομένων σημάτων AIS των πλοίων που βρέθηκαν στην ευρύτερη περιοχή του Σαρωνικού Κόλπου την περίοδο 01/01/2018 – 31/12/2018. Τα σήματα AIS συλλέχθηκαν από κεραία εγκατεστημένη στην οροφή του Πανεπιστήμιου του Πειραιά. Αφού αποθηκεύτηκαν σε βάση δεδομένων (PostgreSQL) και έγινε καθαρισμός του θορύβου, επιλέχθηκαν τα δεδομένα που έχουν ταχύτητα το πολύ έως 2 κόμβους (2 knots). Στη συνέχεια, έγινε η δημιουργία των τροχιών με κριτήριο η απόσταση δύο διαδοχικών σημείων να είναι μικρότερη των 10 χλμ. Δεν αποτέλεσε, σκοπίμως, επιλογή κριτηρίου ο χρόνος, αφού ένα πλοίο μπορεί να αγκυροβολήσει και για μεγάλο χρονικό διάστημα. Τέλος, υπολογίστηκαν απευθείας τα σημεία m_i ως το κεντροειδές των σημείων της τροχιάς και αποθηκεύτηκαν στο αρχείο ms.csv.

Έχοντας δημιουργήσει το αρχείο ms.csv, ο χρήστης μπορεί εκτελέσει διάφορα παραδείγματα εκτελώντας τα κελιά του Jupyter Notebook test_ais.ipynb. Η παρακάτω εικόνα δείχνει τις τοποθεσίες των κεντροειδών και το ιστόγραμμα της διάρκειας παραμονής (te-ts). Έτσι, εκτελέστηκαν οι δοκιμές που παρουσιάζονται στο Table 4.

Η πρώτη δοκιμή εκτελέστηκε με διάστημα 2 ημερών και χρονικό παράθυρο 15 διαστημάτων. Οπότε, η πρώτη συσταδοποίηση g_clustering εντοπίζει τα αγκυροβόλια (τουλάχιστον 4 πλοία) για τις 2 ημέρες του έτους. Η δεύτερη συσταδοποίηση v_clustering εντοπίζει τα μακροπρόθεσμα και βραχυπρόθεσμα αγκυροβόλια σε διάστημα $2*15=30$ ημερών. Ο χαρακτηρισμός «μακροπρόθεσμο» και «βραχυπρόθεσμο» αγκυροβόλιο βασίζεται, εν προκειμένω στην περίπτωση του DBSCAN, από το min_samples της συσταδοποίησης clustering. Άρα, σε αυτή τη δοκιμή αγκυροβόλια

² <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>

³ <https://github.com/geomimo/Diploma-Thesis>

που εμφανίστηκαν λιγότερο από 6 φορές σε μια περιοχή κατά τη διάρκεια 30 ημερών θεωρούνται βραχυπρόθεσμα.

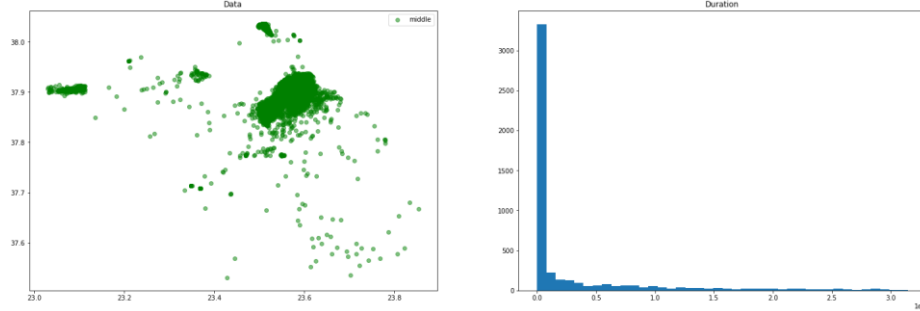


Fig. 18. Διάγραμμα διασποράς σημείων και ιστόγραμμα διάρκειας παραμονής.

Table 4. Παράμετροι δοκιμών

Test	Interval (days)	w	g_clustering_args (eps, min_samples)	v_clustering_args (eps, min_samples)	a
1	2	15	0.010, 4	0.010, 6	0.01
2	2	15	0.010, 2	0.010, 15	0.01
3	7	4	0.013, 4	0.020, 4	0.01
4	31	12	0.020, 10	0.030, 12	0.01

Η δεύτερη δοκιμή εκτελέστηκε με παρόμοιες παραμέτρους, μειώνοντας το κατώφλι χαρακτηρισμού αγκυροβολίου ($\text{min_samples/g_clustering_args}$) και αυξάνοντας παράλληλα το κατώφλι χαρακτηρισμού των βραχυπρόθεσμων αγκυροβολίων ($\text{min_samples/v_clustering_args}$). Στα Figures 19 & 20 φαίνονται οι διαφορές για την ίδια περίοδο. Στη δεύτερη δοκιμή, παρατηρούμε πως εντοπίστηκαν περισσότερα αγκυροβόλια (γενική ύπαρξη σημείου, κύκλου ή αστερίσκου) λόγω της μείωσης του κατωφλίου και χαρακτηρίστηκαν ως βραχυπρόθεσμα για την αντίστοιχη αύξηση της παραμέτρου για τη δεύτερη συσταδοποίηση (αστερίσκος). Σε αυτό το σημείο παρατηρείται ότι το $\text{min_samples/v_clustering_args}$ πρέπει να είναι μικρότερο του w, διαφορετικά όλα τα αγκυροβόλια θα χαρακτηριστούν ως θόρυβος, άρα «βραχυπρόθεσμα».

Στην περίπτωση του STHS, από τις δυο δοκιμές εξήχθησαν hot spots και cold spots (βλ. Fig. 21), με τη δεύτερη να εντοπίζει περισσότερα μικρά cold spots (βλ. Fig. 22) λόγω της μικρότερης τιμής της παραμέτρου $\text{min_samples/g_clustering_args}$.

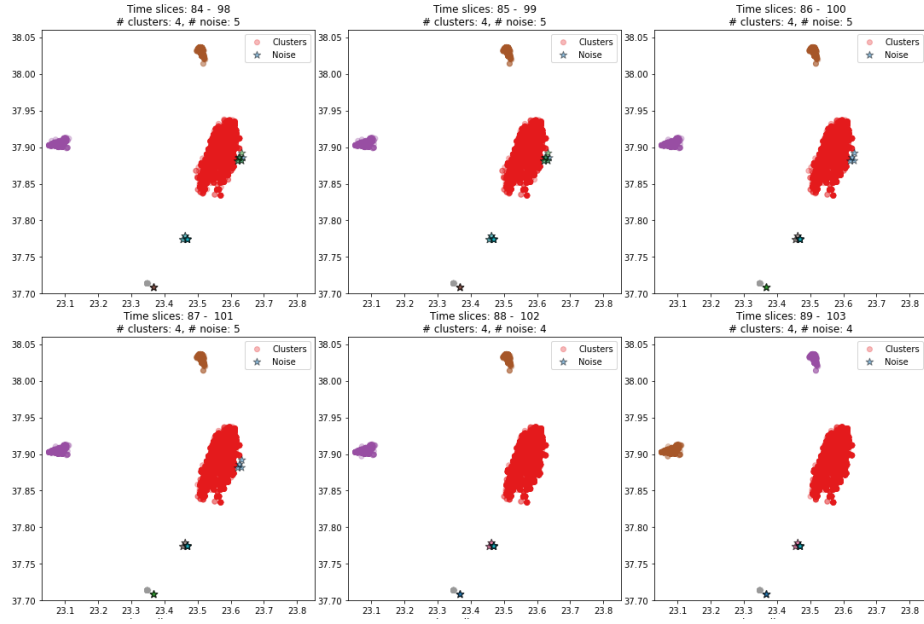


Fig. 19. STC AIS test 1.

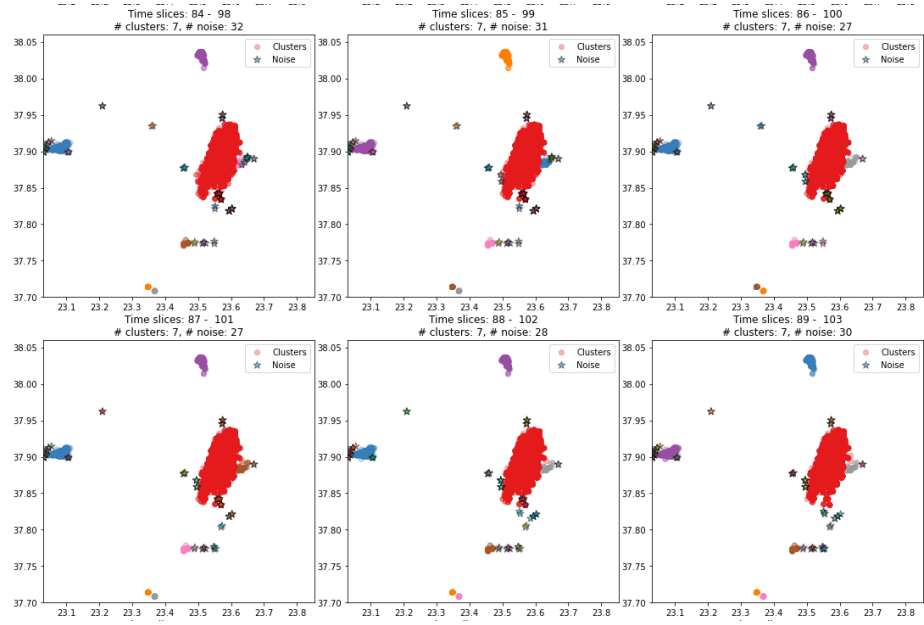


Fig. 20. STC AIS test 2.

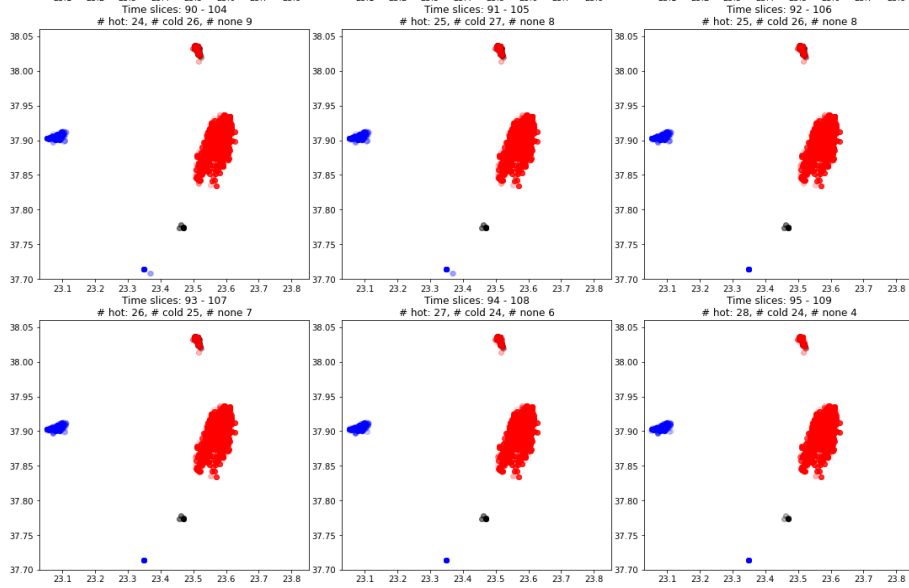


Fig. 21. STHS AIS test 1.

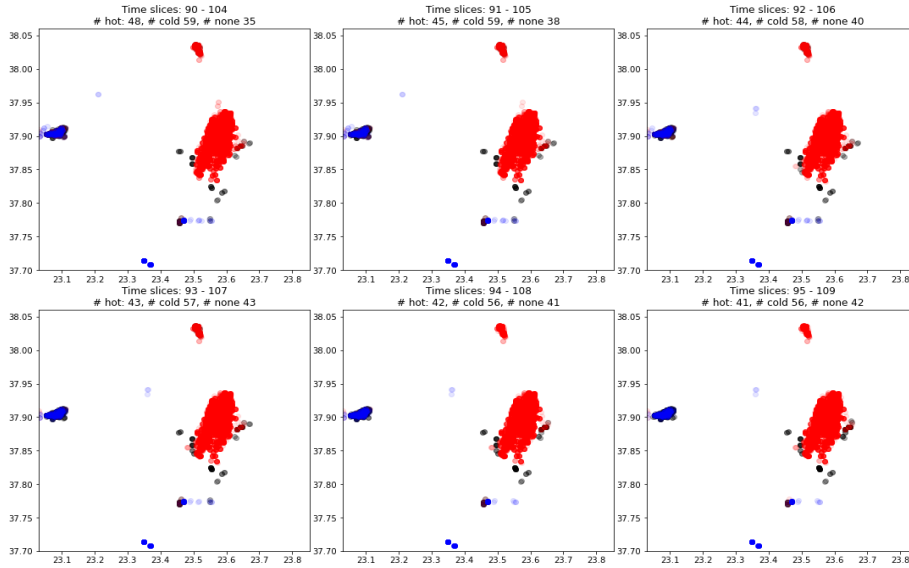


Fig. 22. STHS AIS test 2.

Παρατηρώντας τα αποτελέσματα προκύπτει πως όλα τα αγκυροβόλια έχουν χαρακτηριστεί ως στατιστικά σημαντικά (statistical significant) με ελάχιστες εξαιρέσεις. Αυτό οφείλεται στην κατανομή των χαρακτηριστικών τιμών χ_i των συστάδων στο κάθε V_j . Στο Figure 23 φαίνεται η εν λόγω κατανομή για 3 τυχαία V_j . Είναι κατανοητό πως οι

χαρακτηριστικές τιμές συσσωρεύονται σε δυο περιοχές με μεγάλη απόκλιση μεταξύ τους δημιουργώντας έτσι μια παραπλανητική μέση τιμή (πορτοκαλί γραμμή). Αυτό έχει ως αποτέλεσμα τη δημιουργία ακραίων G_i^* τιμών και τον χαρακτηρισμό σχεδόν όλων των συστάδων ως στατιστικά σημαντικές. Να σημειωθεί πως και σε όλα τα υπόλοιπα V_j παρατηρείται το ίδιο φαινόμενο.

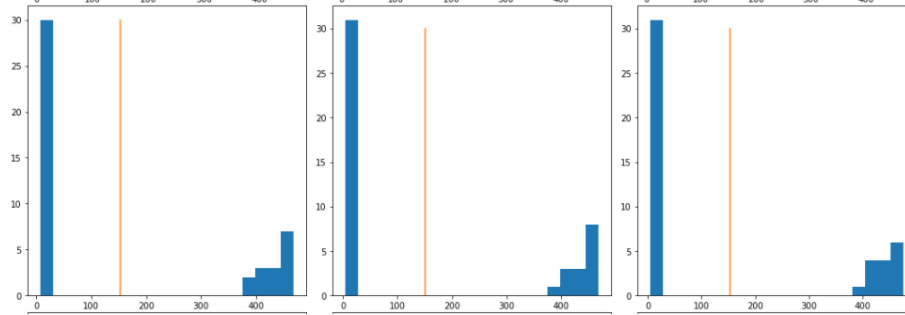


Fig. 23. Ιστόγραμμα χαρακτηριστικών τιμών στα V_j .

Στην τρίτη δοκιμή, το διάστημα γίνεται 7 ημέρες και το χρονικό παράθυρο 4 διαστήματα, αντίστοιχα, οπότε συνολικά θα εξεταστούν 28 ημέρες. Επιπλέον, υιοθετούνται πιο χαλαρές τιμές για τους χωρικούς περιορισμούς συσταδοποίησης (`min_samples`), όμως απαιτείται η ύπαρξη 4 συστάδων, δηλαδή αγκυροβόλιο στην ίδια περιοχή και στα 4 διαστήματα, έτσι ώστε να χαρακτηριστεί ως «μακροπρόθεσμο». Δεν παρατηρείται μεγάλη διαφορά στις συστάδες που εντοπίστηκαν στις προηγούμενες δοκιμές (βλ. Fig. 24), αφού αποτελούν ήδη γνωστά αγκυροβόλια στο οποία παραμένουν καθημερινά πλοία, π.χ. η κόκκινη συστάδα είναι το μακροπρόθεσμο αγκυροβόλιο των εμπορικών πλοίων νότια της νήσου Σαλαμίνας. Στην περίπτωση του STHS στην τρίτη δοκιμή (βλ. Fig. 25), εντοπίστηκαν στατιστικά σημαντικές συστάδες. Το αξιοσημείωτο σε αυτή τη δοκιμή είναι πως οι παράμετροι του `g_clustering` οδήγησαν στη δημιουργία λιγότερων συστάδων, συγκεκριμένα `#hot + #cold + #none < 30`, γεγονός το οποίο επηρεάζει σημαντικά την αξιοπιστία της στατιστικής G_i^* και τον αντίστοιχο έλεγχο υποθέσεων. Να σημειωθεί πως και σε αυτή την περίπτωση το πρόβλημα με την κατανομή των χαρακτηριστικών τιμών παραμένει.

Για την τελευταία δοκιμή στα δεδομένα AIS επιλέχθηκε διάστημα 31 ημερών και χρονικό παράθυρο 12 διαστημάτων, ώστε να εξετασθεί η φύση των αγκυροβολίων συνολικά σε ένα έτος. Όσον αφορά το `g_clustering`, τέθηκε `min_samples` ίσο με 10, δηλαδή ύπαρξη 10 τουλάχιστον πλοίων στο διάστημα 31 ημερών, ώστε να θεωρηθεί μια συστάδα ως αγκυροβόλιο και όχι θόρυβος. Το `v_clustering/min_samples` τέθηκε ίσο με 12, το οποίο θα οδηγήσει σε χαρακτηρισμό των συστάδων ως μακροπρόθεσμες, μόνο όταν τα αγκυροβόλια υπάρχουν κάθε μήνα του έτους. Ως εκ τούτου, έχουμε το μοναδικό αποτέλεσμα του Figure 26.

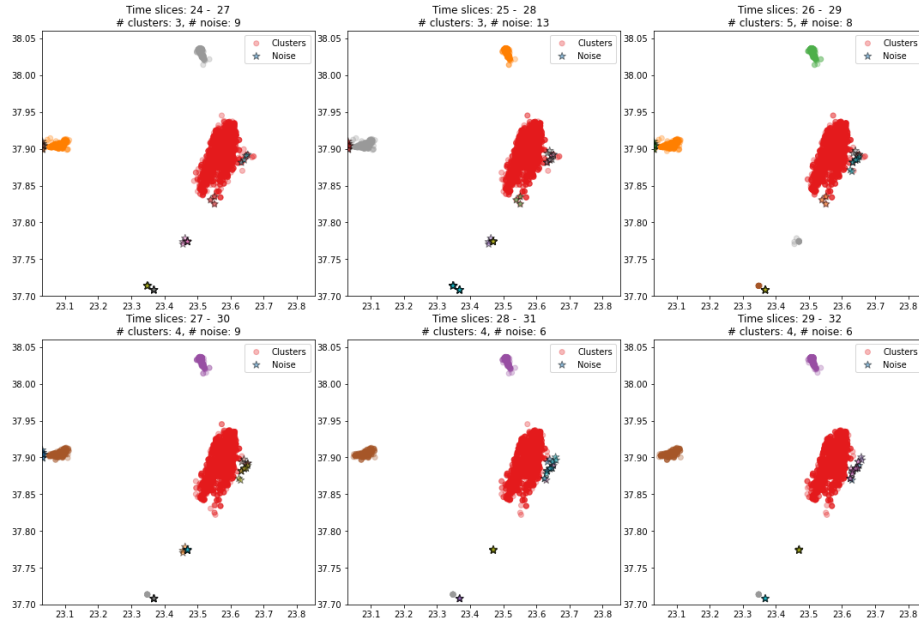


Fig. 24. STC AIS test 3.

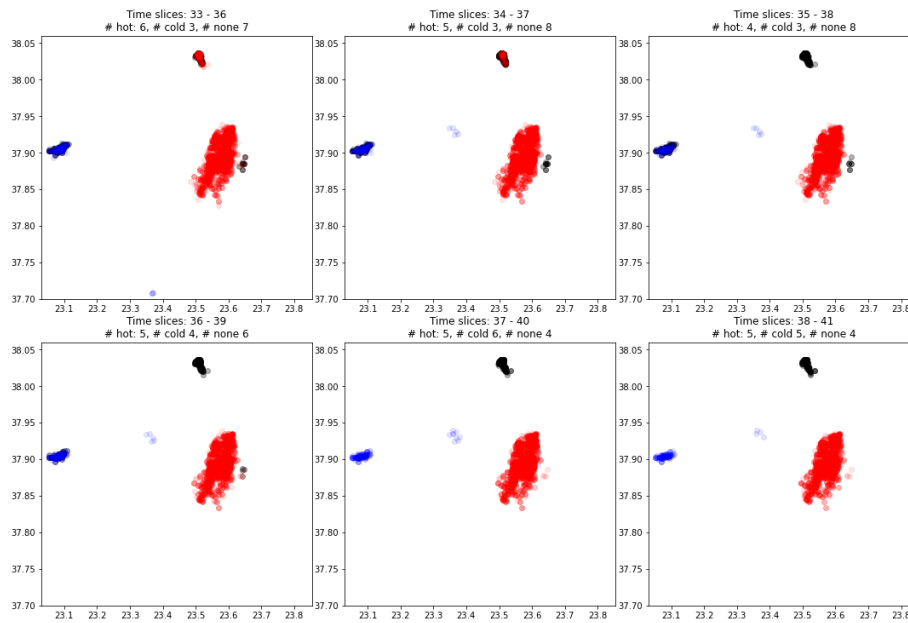


Fig. 25. STHS AIS test 3.

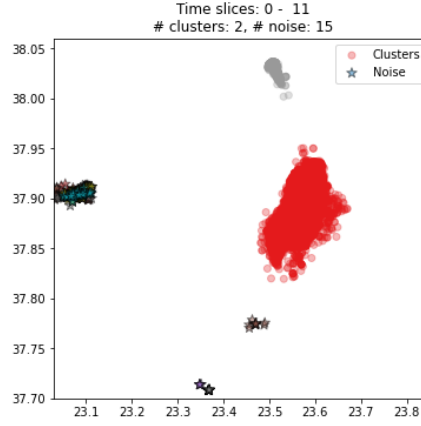


Fig. 26. STC AIS test 4.

Ο STHS εντόπισε ξανά στατιστικά σημαντικές συστάδες με αξιοπιστία (βλ. Fig. 27), λόγω του συνολικού πλήθους των συστάδων (>30), όμως και εδώ η κατανομή των χαρακτηριστικών τιμών ίσως αποτελεί πρόβλημα.

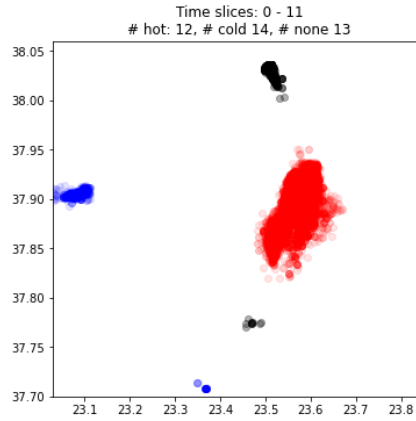


Fig. 27. STHS AIS test 4.

6.2 STC και STHS σε προσομοιωμένα δεδομένα

Για τη μελέτη των αλγορίθμων, εκτός των δεδομένων AIS, χρησιμοποιήθηκαν και προσομοιωμένα δεδομένα τα οποία παρήχθησαν από συναρτήσεις τυχαίων αριθμών. Πιο συγκεκριμένα, υλοποιήθηκε συνάρτηση η οποία παράγει ομοιόμορφα n συστάδες που περιέχουν συνολικά $N * (p - 1)$ τροχιές (p_{is}, p_{ie}), ενώ $N * p$ τροχιές αποτελούν θόρυβο. Επιπλέον, οι χρόνοι t_s και t_e είναι εξίσου τυχαίοι, με $t_s < t_e$. Πρέπει να σημειωθεί πως τα δεδομένα περιορίζονται στο σύνολο $[23, 24] \times [37, 38]$, ώστε να προσομοιωθεί η χωρική έκταση του Σαρωνικού. Στα παρακάτω παραδείγματα είναι $p = 0.40$. Οι δοκιμές που διεξήχθησαν είναι οι εξής:

Table 5. Παράμετροι δοκιμών προσομοιωμένων δεδομένων

Test	N	n	Interval (days)	w	g_clustering_args (eps, min_samples)	v_clustering_args (eps, min_samples)	a
1	5000	5	2	15	0.010, 4	0.010, 6	0.05
2	1000	3	2	15	0.010, 2	0.010, 15	0.05
3	10000	2	7	4	0.013, 4	0.020, 4	0.05
4	5000	5	31	12	0.020, 10	0.030, 12	0.05

Στην πρώτη δοκιμή, οι χωρικοί περιορισμοί είναι «χαλαροί», ενώ απαιτείται η ύπαρξη αγκυροβολίου στα 3/4 των συσσωρευμένων χρονικών διαμερίσεων, ώστε να χαρακτηριστεί μια συστάδα ως «μακροπρόθεσμη».

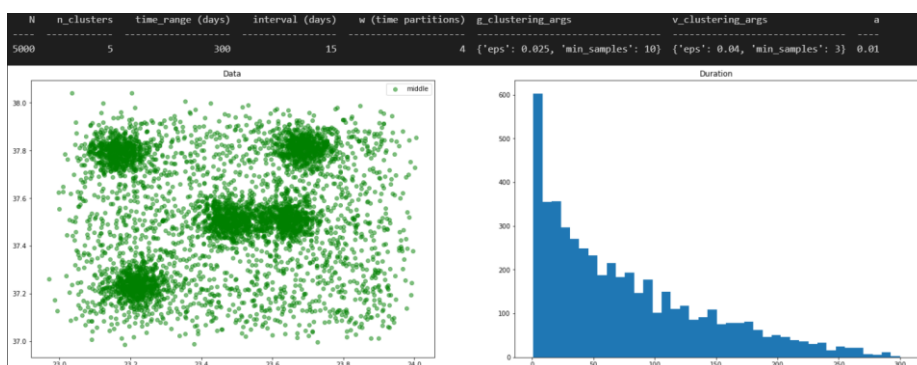


Fig. 28. Dummy data test 1.

Στα Figures 29 & 30 φαίνονται μερικά αποτελέσματα των STC και STHS αλγορίθμων, οι οποίοι εντόπισαν συμπαγείς συστάδες. Το πλήθος των χαρακτηρισμένων αγκυροβολίων είναι 4, ενώ δόθηκε ως παράμετρος $n = 5$ συστάδες. Αυτό συμβαίνει διότι 2 συστάδες παρήχθησαν σε κοντινή απόσταση, οπότε η πυκνότητα των δεδομένων οδήγησε στον εντοπισμό μιας συστάδας. Στην περίπτωση του STHS, εντοπίστηκαν hot spots, ενώ τα αποτελέσματα δεν είναι αξιόπιστα λόγω του μικρού πλήθους συστάδων (<30). Γενικά, παρατηρείται χαρακτηρισμός της μεσαίας «διπλής» συστάδας ως hot spot, αφού τα πλοία προέρχονται από την παραγωγή δυο διαφορετικών συστάδων, διαφοροποιώντας το πλήθος πλοίων της εν λόγω συστάδας από τις υπόλοιπες.

Η δεύτερη δοκιμή αποτελείται από αραιά δεδομένα (βλ. Fig. 31), και σε συνδυασμό με τις παραμέτρους συσταδοποίησης οδηγούν σε εντοπισμό μικρότερων αγκυροβολίων (βλ. Fig. 32). Η επιλογή των time range = 300, interval = 5 και w = 10 οδηγεί στη δημιουργία πολλών αραιών αγκυροβολίων σε κάθε V_j , αυξάνοντας την αξιοπιστία της στατιστικής Getis-Ord στην περίπτωση του STHS (βλ. Fig. 33).

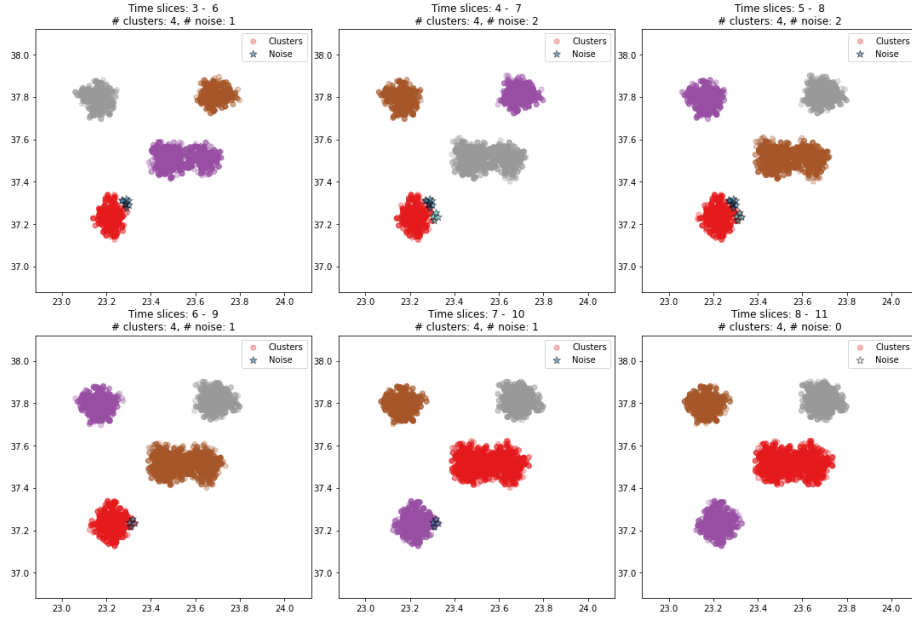


Fig. 29. STC dummy test 1.

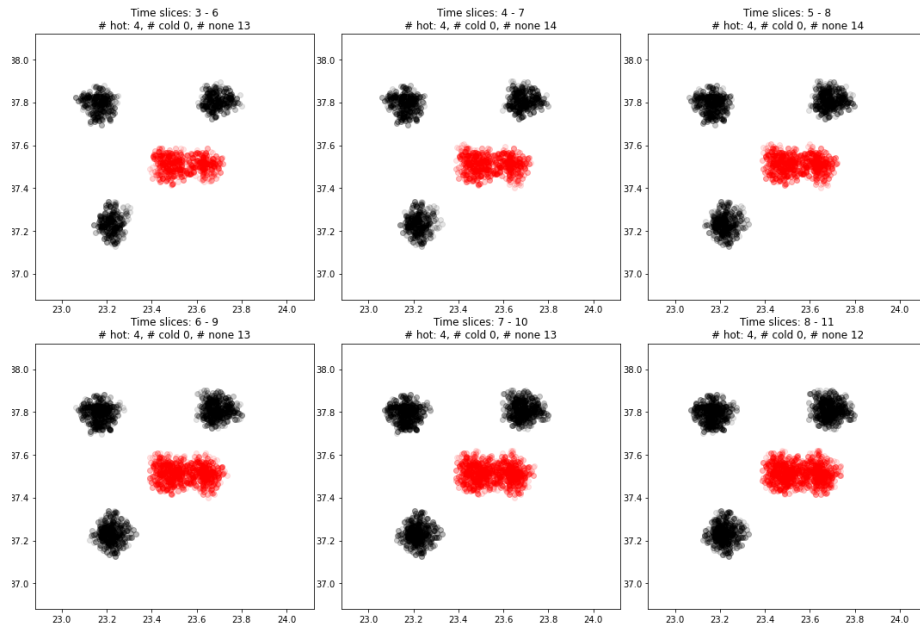


Fig. 30. STHS dummy test 1.

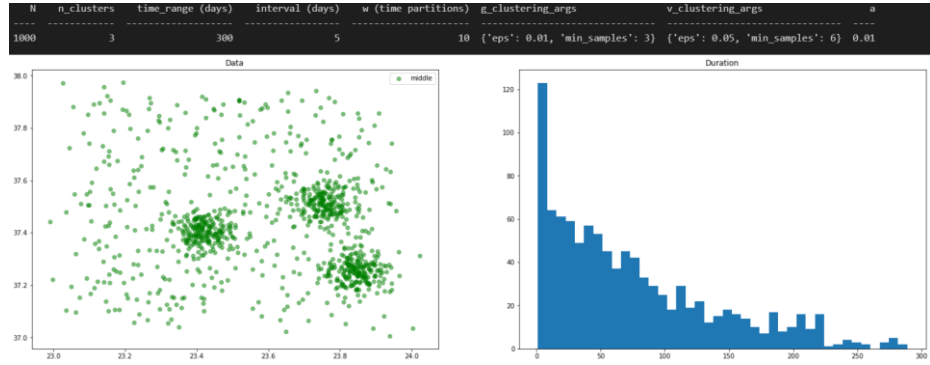


Fig. 31. Dummy data test 2.

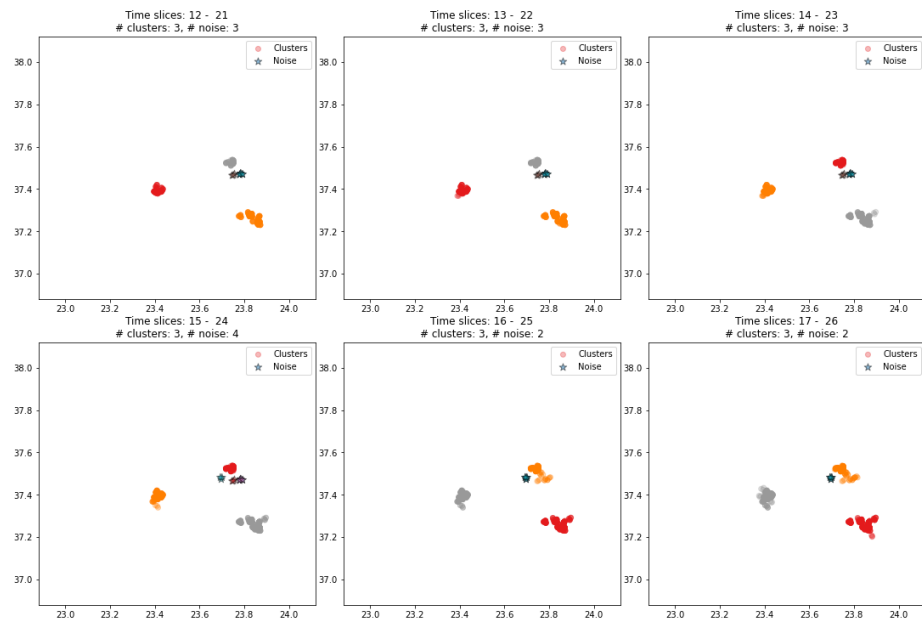


Fig. 32. STC dummy test 2.

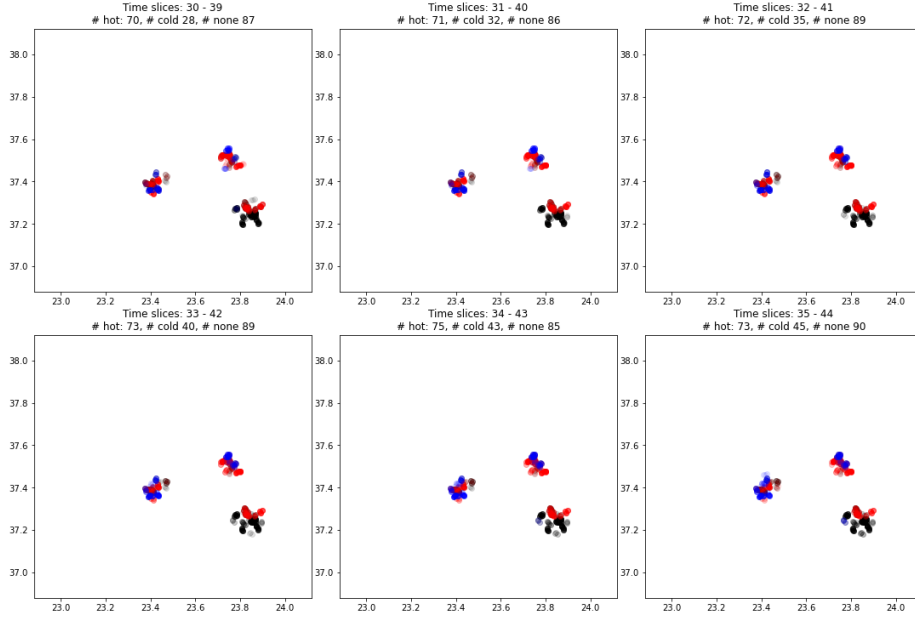


Fig. 33. STHS dummy test 2.

Η τρίτη δοκιμή περιλαμβάνει τεράστιο πλήθος δεδομένων και η συγκεκριμένη τυχαία κατανομή των δεδομένων θα μπορούσε να ερμηνευθεί ως η ύπαρξη δύο λιμανιών και η θαλάσσια δραστηριότητα ενδιάμεσα (βλ. Fig. 34).

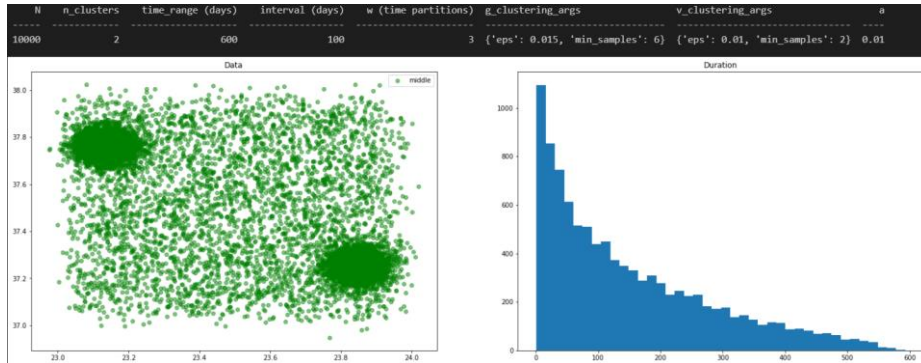


Fig. 34. Dummy data test 3.

Ο STC εντοπίζει τα «λιμάνια», καθώς και αρκετά «μακροπρόθεσμα» και «βραχυχρόνια» αγκυροβόλια ενδιάμεσα. Να σημειωθεί πως η απόδοση κόκκινου χρώματος και δύο στα «λιμάνια» συνέβη λόγω έλλειψης χρωμάτων στην χρωματική παλέτα (βλ. Fig. 35). Στο Figure 36 παρουσιάζεται άλλη μία έμπιστη εφαρμογή του STHS.

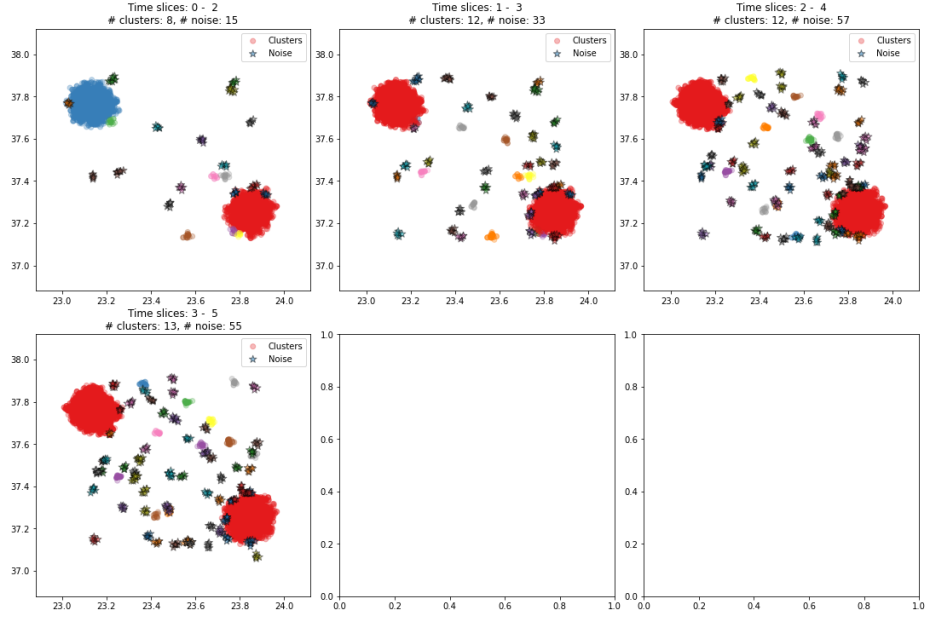


Fig. 35. STC dummy test 3.

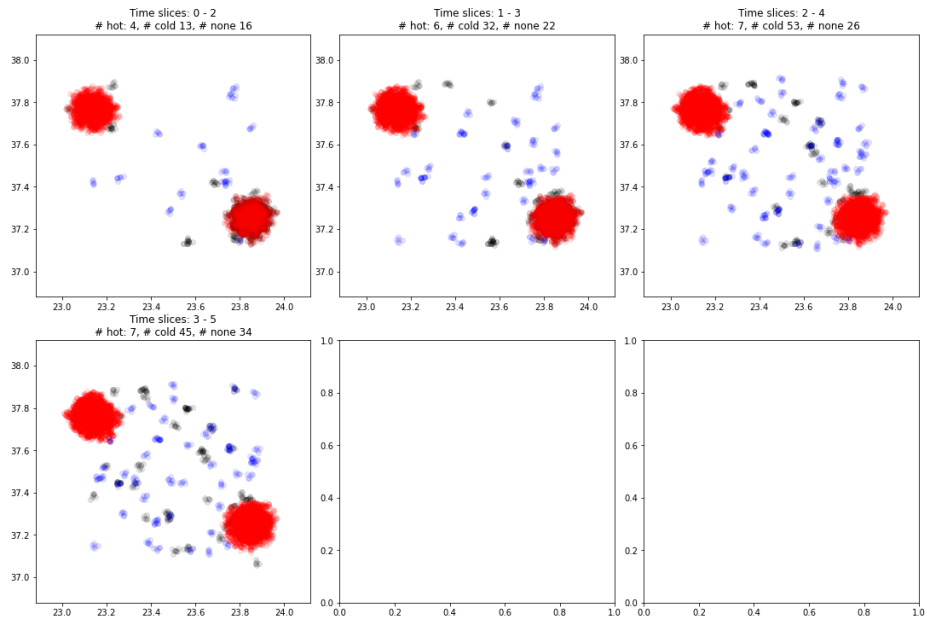


Fig. 36. STHS dummy test 3.

Στο τελευταίο παράδειγμα, μπορεί κάποιος να ερμηνεύσει τις παραμέτρους ως τον εντοπισμό αγκυροβολίων με διάρκεια μιας ημέρας (interval), στα οποία εντοπίστηκαν

τουλάχιστον 10 ($g_clustering_args/min_samples$) πλοία, και χαρακτηρίστηκαν «μακροπρόθεσμα» αν διατηρήθηκαν για τουλάχιστον 25 ημέρες (w) σε διάστημα 30 ημερών ($v_clustering_args/min_samples$).

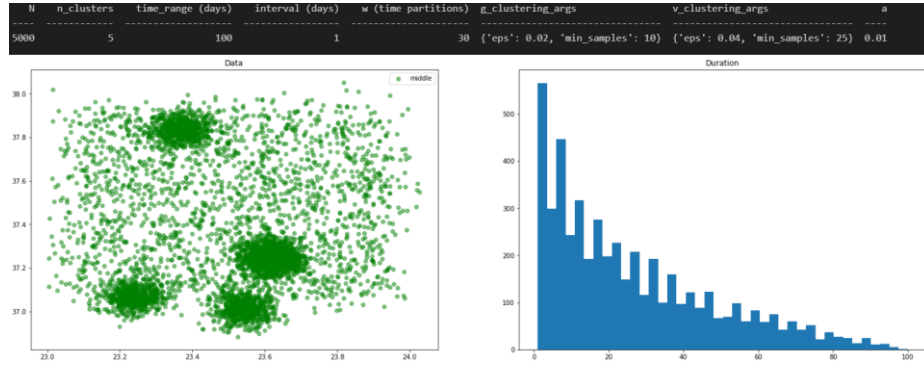


Fig. 37. Dummy data test 4.

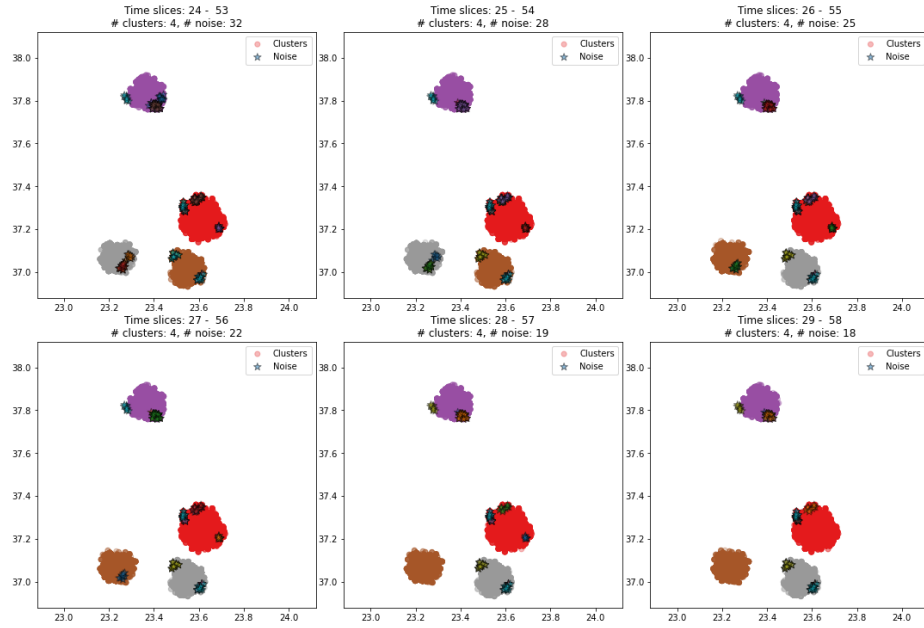


Fig. 38. STC dummy test 4.

Αξιοσημείωτο είναι πως ένα αγκυροβόλιο μπορεί να χαρακτηριστεί ως hot spot και cold spot σε διαφορετικές χρονικές διαμερίσεις, όπως συμβαίνει στις νότιες συστάδες.

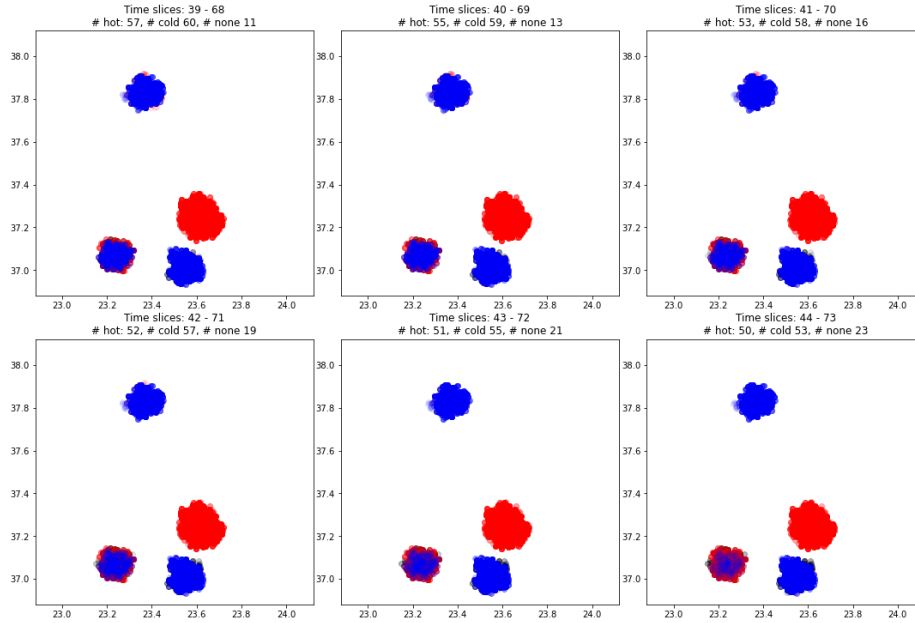


Fig. 39. STHS dummy test 4.

7 Συμπεράσματα και μελλοντική μελέτη

Στην παρούσα εργασία παρουσιάστηκαν δύο αλγόριθμοι για τον εντοπισμό και χαρακτηρισμό αγκυροβολίων ως προς τη χρονική διάσταση αλλά και τη στατιστική σημαντικότητα. Παράλληλα, ξεπερνούν το έμφυτο πρόβλημα της κλίμακας, το οποίο εμφανίζεται στο πεδίο της θαλάσσιας επιτήρησης. Τα σημαντικά πλεονεκτήματά τους είναι, μεταξύ άλλων, η γενικευμένη μορφή τους η οποία επιτρέπει τη χρήση τους και σε άλλα πεδία με κατάλληλες παραμετροποιήσεις, καθώς και η εύκολη ερμηνεία και κατανόηση των παραμέτρων τους.

Επιπλέον, ενδιαφέρον παρουσιάζει η δυνατότητα υλοποίησης των αλγορίθμων σε κατανεμημένες αρχιτεκτονικές συστημάτων λόγω της χρονικής διαμέρισης. Η υλοποίηση αυτή αποτελεί αντικείμενο για μελλοντική μελέτη, επιτρέποντας την επεξεργασία τεράστιου όγκου δεδομένων. Ακόμη, η οπτικοποίηση των αποτελεσμάτων παραμένει μια δύσκολη διεργασία λόγω της χρονικής διάστασης και του πλήθους των διαμερίσεων, ωστόσο αποτελεί ενδιαφέρον θέμα προς έρευνα και υλοποίηση. Μια μελλοντική εργασία θα μπορούσε να έχει θέμα τη μελέτη της «ψευδό-στατιστικής» σημαντικότητας του αλγορίθμου STC και το κατά πόσο αυτή προσεγγίζει την πραγματικότητα, είτε την αυστηρώς καθορισμένη στατιστική σημαντικότητα.

Τέλος, αξίζει να μελετηθεί η αξιοπιστία των αλγορίθμων σε διάφορες μη-κυκλικές μορφές συστάδων, αφού η επιλογή των κεντροειδών και η χρήση αλγορίθμων συσταδοποίησης με βάση την πυκνότητα ενδέχεται να παρουσιάζουν παραπλανητικά αποτελέσματα στις αντίστοιχες περιπτώσεις.

Αναφορές

- [1] P. Moran, «Notes on Continuous Stochastic Phenomena,» *Biometrika*, pp. 17-23, 1950.
- [2] A. Getis και K. Ord, «The Analysis of Spatial Association by Use of Distance Statistics,» *Geographical Analysis*, pp. 189 - 206, 1992.
- [3] ESRI, «How Spatial Autocorrelation (Global Moran's I) works,» [Ηλεκτρονικό]. Available: <https://pro.arcgis.com/en/pro-app/2.7/tool-reference/spatial-statistics/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm>.
- [4] ESRI, «How High/Low Clustering (Getis-Ord General G) works,» [Ηλεκτρονικό]. Available: <https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-how-high-low-clustering-getis-ord-general-g-spat.htm>.
- [5] ESRI, «How Hot Spot Analysis (Getis-Ord Gi*) works,» [Ηλεκτρονικό]. Available: <https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-how-hot-spot-analysis-getis-ord-gi-spatial-stati.htm>.
- [6] ESRI, «How Cluster and Outlier Analysis (Anselin Local Moran's I) works,» [Ηλεκτρονικό]. Available: <https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-how-cluster-and-outlier-analysis-anselin-local-m.htm>.
- [7] R. C. Geary, «The Contiguity Ratio and Statistical Mapping,» *The Incorporated Statistician*, pp. 115-146, 1954.
- [8] M. Ankerst, M. M. Breunig, H.-P. Kriegel και J. Sander, «OPTICS: Ordering Points to Identify the Clustering Structure,» *SIGMOD Rec.*, pp. 49-60, 1999.
- [9] M. Ester, H.-P. Kriegel, J. Sander και X. Xu, «A density-based algorithm for discovering clusters in large spatial databases with noise,» *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226-231, 1996.
- [10] M. Nanni και D. Pedreschi, «Time-Focused Clustering of Trajectories of Moving Objects,» *J. Intell. Inf. Syst.*, pp. 267-289, 2006.
- [11] J. Chen, «A new trajectory clustering algorithm based on TRACCLUS,» *Proceedings of 2012 2nd International Conference on Computer Science and Network Technology*, pp. 789-787, 2012.
- [12] P. Nikitopoulos, A.-I. Paraskevopoulos, C. Doulkeridis, N. Pelekis και Y. Theodoridis, «Hot Spot Analysis over Big Trajectory Data,» *2018 IEEE International Conference on Big Data (Big Data)*, pp. 761-770, 2018.
- [13] A. Tritsarolis, N. P. Yiannis Kontoulis και Y. Theodoridis, «MaSEC: Discovering Anchorages and Co-movement Patterns on Streaming Vessel Trajectories,» *SSTD*, 2021.

- [14] G.-S. Theodoropoulos, A. Tritsarolis και Y. Theodoridis, «EvolvingClusters: Online Discovery of Group Patterns in Enriched Maritime Data.,» *Multiple-Aspect Analysis of Semantic Trajectories*, 2020.
- [15] E. Fix και J. J. L. Hodges, «Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties,» *Technical Report 4* , *USAF School of Aviation Medicine, Randolph Field*, 1951.