ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΨΗΦΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ

ΚΑΤΕΥΘΥΝΣΗ: ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ & ΑΝΑΛΥΤΙΚΗ

ΜΑΘΗΜΑ: Μεγάλα Δεδομένα και Αναλυτική ΙΙ: Τεχνικές και Εργαλεία

ΔΙΔΑΣΚΟΝΤΕΣ: Χ. Δουλκερίδης - Α. Βλάχου

ΦΟΙΤΗΤΗΣ :

Αϊντίνι Μπαϊράμ ΜΕ1701

Περιεχόμενα

1. Εισαγωγή

Σε αυτή την έρευνα παρουσιάζεται ένα σύστημα για την παρακολούθηση της θαλάσσιας δραστηριότητας βάσει των θέσεων των πλοίων που πλέουν στη θάλασσα. Γενικά η ανίχνευση τέτοιων δεδομένων γίνεται και επικεντρώνεται στις σημαντικές αλλαγές πορείας του κάθε σκάφους στο πέρασμα του χρόνου και έτσι μπορούμε να κρατήσουμε ιστορικά δεδομένα για την πρόσφατη κίνηση των πλοίων. Επιπλέον, χάρη στην αναγνώριση σύνθετων γεγονότων, αυτό το σύστημα μπορεί επίσης να προσφέρει άμεση ειδοποίηση σε περίπτωση θαλάσσιας κατάστασης έκτακτης ανάγκης, όπως ο κίνδυνος των συγκρούσεων, ύποπτες κινήσεις σε προστατευόμενες ζώνες, ή ανταλλαγή παράνομων φορτίων στην ανοιχτή θάλασσα. Τα δεδομένα τα οποία θα ασχοληθούμε είναι χωροχρονικά όπως αναφέραμε αλλά και γεωγραφικά. Η ερεύνα μας θα εστιάσει στην παρακολούθηση αλιευτικών πλοίων που βρίσκονται εντος απαγορευμένων αλιευτικών χώρων. Θα γίνει μια ανάλυση χρονολογικής σειράς, η οποία θα μελετά την συχνότητα των πλοίων που είναι εντος αυτής στο βάθος του χρόνου με το μοντέλο AR(autoregressive). Τέλος θα εκτελέσουμε τον αλγόριθμο ομαδοποίησης K-means για τα αλιεύτηκα πλοία εντος των επιτρεπομένων περιοχών αλίευσης. Για τον λόγο ότι ο όγκος των δεδομένων που έχουμε είναι τεράστιος και δύσκολα επεξεργάσιμος χρησιμοποιήσαμε εργαλεία οπού προσφέρονται για την ανάλυση μεγάλης κλίμακας δεδομένων όπως είναι το Hdfs για την παράλληλη αποθήκευση των δεδομένων σε cluster υπολογιστών καθώς και το Spark για την γρήγορη επεξεργασίας μεγάλης κλίμακας παράλληλων δεδομένων.

2. Το Πρόβλημα Της Αλίευσης Σε Προστατευόμενες Περιοχές

Η αλιεία αποτελεί βασικό κομμάτι της οικονομίας και της παράδοσης κάθε χωράς η όποια περικλείετε από θάλασσα. Μεγάλος εχθρός είναι η υπεραλίευση καθώς και η αλίευση εντος απαγορευμένων περιοχών η οποία μπορεί να επηρεάσει τόσο τους πληθυσμούς των αλιευμάτων όσο και το θαλάσσιο οικοσύστημα. Για την προστασία της θαλάσσιας αλιείας τα αρμόδια Υπουργεία (Υπουργείο Ναυτιλίας) μέσω κάποιου συστήματος ελέγχου αλιείας είναι υπεύθυνο για το συντονισμό εφαρμογής των μέτρων που λαμβάνονται για τον έλεγχο της αλιείας και την παρακολούθηση της δραστηριότητας των Λιμενικών Αρχών της , στη δίωξη της παράνομης αλιείας. Ως Αλιευτική παράβαση χαρακτηρίζεται γενικά οποιαδήποτε δραστηριότητα που αντίκειται στην κείμενη νομοθεσία περί αλιείας, είτε σε θαλάσσιο χώρο, εντός των χωρικών υδάτων, είτε εντός των εσωτερικών υδάτων μιας χώρας, (π.χ. ποταμούς, λίμνες). Οι αλιευτικές παραβάσεις διακρίνονται σε αρκετά είδη. Οι παραβάσεις με τις όποιες ασχοληθήκαμε είναι παραβάσεις σε "απαγορευμένες περιοχές". Πρόκειται για ειδικές παραβάσεις αλιείας σε χώρους που αποκλείονται για αλιευτικές δραστηριότητες π.χ. εντός λιμένων, ναυστάθμων, παράλιων ή βιομηχανικών εγκαταστάσεων, ναυτικών οχυρών, διαύλων και ακρωτηρίων έντονης κυκλοφορίας πλωτών μέσων κ.λπ. Στην συγκεκριμένη έρευνα έχουμε δεδομένα τα όποια αφορούν προστατευόμενες περιοχές αλίευσης και πιο συγκεκριμένα έχουμε δυο περιοχές οι όποιες όπως φαίνονται και παρακάτω ότι είναι κοντά στην ακτή εικόνα 4.1. Αυτό όπου θέλουμε να βρούμε στο βασικό κομμάτι της εργασίας είναι πλοία εντος των περιοχών αυτών τα όποια βρίσκονται σε κατάσταση αλίευσης και το οποίο θα το εξηγήσουμε στο Κεφάλαιο 7 αργότερα.

3. Περιγραφή Δεδομένων

Τα δεδομένα τα όποια έχουμε χρησιμοποιήσει παρέχονται από την ιστοσελίδα του zenodo (Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance) και χρονικά αναφέρονται για το εξάμηνο 2015-10-01 έως 2016-03-31. Η συγκεκριμένη ιστοσελίδα είχε πολλά data sets τα όποια αφορούν δεδομένα κινούμενων πλοίων, περιβάλλοντος (μετεωρολογικά), γεωγραφικά δεδομένα καθώς και άλλα. Από όλα αυτά τα σύνολα δεδομένων στην παρούσα εργασία θα χρησιμοποιήσουμε τα εξής :

- [P1] AIS Data (nari_dynamic .csv, nari_static.csv)
- [P1] AIS Status Codes and Types
- [C5] Fishing Constraints
- [C4] Fishing Areas (European commission)

Στους παρακάτω πινάκες και εικόνες θα δούμε το σχήμα όπου έχει κάθε ένα από αυτά τα data sets καθώς και περιγραφικά μέτρα

1. Για το τα δεδομένα [P1] AIS Data - nari_dynamic έχουμε

- Sourcemmsi : Είναι το μοναδικό id κάθε πλοίου
- Navigationalstatus : Ακέραιος αριθμός που αναφέρετε στην κατάσταση του πλοίου
- Speedoverground : Ταχύτητα του πλοίου σε κόμβους
- Lon : Είναι το γεωγραφικό μήκος
- Lat : Είναι το γεωγραφικό πλάτος
- t : χρόνος σε epoch/unix timestamp

Οι καταστήσεις του Navigationalstatus είναι :

Πινάκας 3.1

Κίνηση με την χρήση κινητήρα	**0**
Πλοίο με άγκυρα ανοιχτή	1
Όχι υπό εντολή	2
περιορισμένη ελιγμών	3
constrained by her draught	4
Αγκυροβολημένο	5
Στην ξηρά	6
Σε κατάσταση αλίευσης	7
Υπό ιστιοπλοΐα	8
Αποκλειστικά για τη μελλοντική τροποποίηση της κατάστασης πλοήγησης για τα πλοία που μεταφέρουν τη DG	9
Προορίζονται για μελλοντική τροποποίηση της κατάστασης πλοήγησης για πλοία που μεταφέρουν επικίνδυνα εμπορεύματα DG	10
Αποκλειστικά για μελλοντική χρήση	11, 12, 13
AIS-SART (active)	14
Δεν έχει οριστεί	15

Στον παρακάτω πινάκα βλέπουμε μερικά περιγραφικά μετρά των μεταβλητών που περιέχει το συγκεκριμένο Data set. Παρατηρούμε πως οι συνολικές γραμμές που περιχούν τα δεδομένα αυτά είναι 19.035.360 .

Πίνακας 3.2 : Περιγραφικά μέτρα δεδομένων (nari_dynamic)

summary	sourcemmsi	navigationalstatus	speedoverground	lon	lat	t
count	19035630	18084115	19035630	19035630	19035630	19035630
mean	2.551771247438914E8	6.035147586708003	8.98630595889961	-4.568561120024541	48.285813586254896	1.4516568505196023E9
stddev	1.1575024337971106E8	6.825135013024206	21.95265285179425	0.2784973922289842	0.18590166957475818	4570324.795492887
min	923166	0.0	0.0	-9.713331	45.001045	1443650401
max	999999999	15.0	102.3	-0.015736667	50.887634000000006	1459461599

2. Για το τα δεδομένα [P1] AIS Data - nari_static έχουμε

- Sourcemmsi : Το μοναδικό id κάθε πλοίου
- shipname : Το όνομα του πλοίου
- shiptype : Ο τύπος του πλοίου
- t : χρόνος σε epoch/unix timestamp

Παρατηρούμε από τον παρακάτω πίνακα πως οι γραμμές που περιχέονται στο Data set αυτό είναι 1.078.617. Το συγκεκριμένο Data set μας φάνηκε αρκετά χρήσιμο διότι εμπεριέχει με βάσει το mmsi του κάθε πλοίου το όνομα καθώς και τον τύπο του.

Πίνακας 3.3 : Περιγραφικά μέτρα δεδομένων (nari_static)

summary	sourcemmsi	shipname	shiptype
count	1078617	1078558	1064192
mean	2.562654786015351E8	null	50.959855928253546
stddev	9.952139465312959E7	null	22.3658081929189
min	1178		0
max	999999999	ZUIDVLIET	99

3. Για το τα δεδομένα [P1] AIS Status Codes and Types έχουμε

- Shiptype(min,max) : Ο αριθμός που αντιστοιχεί στον τύπο του πλοίου.
- Ais_type_summary : Η ονομασία του τύπο των πλοίων.

Στον παρακάτω πινάκα θα δούμε τους διαφόρους τύπους πλοίων με βάση τον κωδικό που τους έχει οριστεί. Παρατηρούμε πως συνολικά οι τύποι όλων τον πλοίων που έχουμε εμείς είναι 17 καθώς και μια η όποια είναι η other και περιχεί όλες τις υπόλοιπες.

Πινάκας 3.4

shiptype_min	shiptype_max	ais_type_summary
10	19	Unspecified
20	28	Wing in Grnd
29	29	Search and Rescue
30	30	Fishing
31	32	Tug
33	35	Special Craft
36	36	Sailing Vessel
37	37	Pleasure Craft
38	39	Unspecified
40	49	High-Speed Craft
50	50	Special Craft
51	51	Search and Rescue
52	52	Tug
53	59	Special Craft
60	69	Passenger
70	79	Cargo
80	89	Tanker
90	99	Other

Πίνακας 3.5 : Περιγραφικά μέτρα δεδομένων [P1] AIS Status Codes and Types

summary	shiptype	shiptype_max	ais_type_summary
count	38	38	38
mean	54.44736842105263	55.81578947368421	null
stddev	20.904800132106022	20.831181491525125	null
min	10	19	Cargo
max	90	99	Wing in Grnd

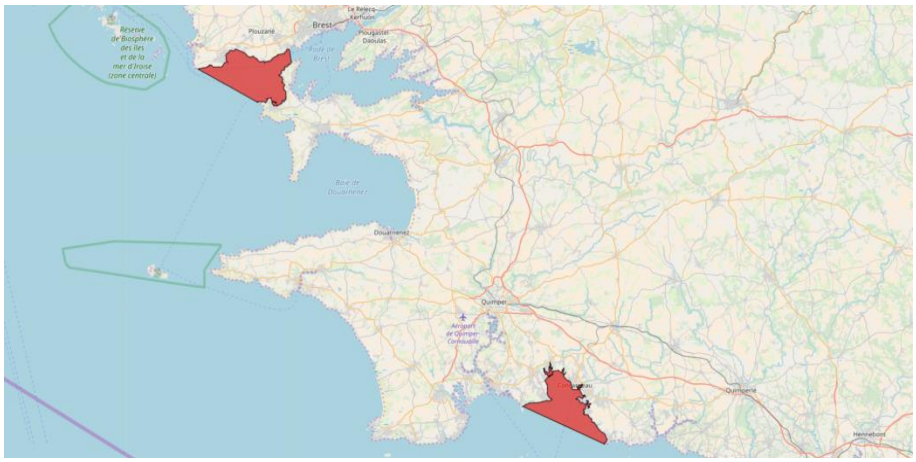
4. Για το τα δεδομένα [C5] Fishing Constraints έχουμε

- Geometry : Περιέχει τα πολύγωνα ‘’points’’ των 2 περιοχών

- Geoid : Περιέχει την αρίθμηση των περιοχών δηλαδή πρώτη περιοχή=0 και η δεύτερη =1

Στην Εικόνα 3.1 βλέπουμε της δυο απαγορευμένες περιοχές που είναι με κόκκινο χρώμα αποτυπωμένες στον χάρτη .

Εικόνα 3.1 : Τα δυο πολύγωνα που περιέχει το [C5] Fishing Constraints



5. Για το τα δεδομένα [C4] Fishing Areas (European commission) έχουμε

- Name : Το όνομα του κάθε πολυγώνου (Σε όλα είναι BREST)
- maxLat, maxLong, minLat, minLong : Είναι τα bounds του ενιαίου πολυγώνου
- geometry : Περιέχει τα πολύγωνα ''points'' των 2 περιοχών
- geoid : Περιέχει την αρίθμηση των περιοχών

Εικόνα 3.2 : Η επιτρεπόμενη περιοχή αλίευσης

4. Περιγραφή Εργαλείων Που Χρησιμοποιηθήκαν

Ο Μεγάλος όγκος δεδομένων προκάλεσε τεράστια αλλαγή στα εργαλεία όπου χρησιμοποιούνται για την αποθήκευση και την ανάλυση δεδομένων. Μια γλωσσά προγραμματισμού εγκατεστημένη σε έναν υπολογιστή δεν είναι ικανή να τα φέρει εις πέρας σε ένα Big Data πρόβλημα. Για την επίλυση αυτού το προβλήματος έχει δημιουργηθεί το Spark για την επεξεργασία των δεδομένων , καθώς και το HDFS του Hadoop για την αποθήκευση . Στην συγκεκριμένη εργασία χρησιμοποιήσαμε και τα δυο αυτά εργαλεία για την υλοποίηση της ερευνά που έγινε ως προς την παράβαση αλιευτικών πλοίων εντος απαγορευμένων περιέχων. Παρακάτω γίνεται μια μικρή επισκόπηση για τα δυο αυτά εργαλεία.

5.1 Εργαλείο Επεξεργασίας δεδομένων Spark

Το Apache Spark δημιουργήθηκε με την γλώσσα προγραμματισμού Scala άλλα έχει την δυνατότητα να χρησιμοποιηθεί και με τις γλώσσες (Python, R, Java). Είναι μία γρήγορη και γενικής χρήσεως μηχανή επεξεργασίας μεγάλης κλίμακας παράλληλων δεδομένων, η οποία επεκτείνει το δημοφιλές μοντέλο MapReduce. Μπορούμε να πούμε ότι ένα από τα πιο σημαντικά χαρακτηριστικά του Spark είναι ότι λειτουργεί στην μνήμη [1]. Το γεγονός αυτό, έχει ως αποτέλεσμα να είναι πιο αποτελεσματικό σε υπολογισμούς, λόγω της αποφυγής του δίσκου ανάγνωσης / εγγραφής συμφόρησης. Το έργο Spark περιέχει πολλές συνιστώσες: Spark SQL, Spark Streaming, MlLib και GraphX. Αυτές οι συνιστώσες έχουν σχεδιαστεί για να εργαστούν από κοινού, όποτε το θελήσει κανείς. Έτσι, μπορούν να συνδυαστούν σε ένα έργο λογισμικού, όπου το Spark αποτελεί τον πυρήνα τους. Τότε το SPARK είναι υπεύθυνο για τον προγραμματισμό, τη διανομή, και τις εφαρμογές παρακολούθησης ενός συμπλέγματος (cluster). Ένα από τα πιο σημαντικά εργαλεία στην Spark είναι η βιβλιοθήκη μηχανικής μάθησης . Αυτή η βιβλιοθήκη είναι μέρος του πυρήνα Spark, ως εκ τούτου μπορεί να χρησιμοποιηθεί από οποιαδήποτε άλλη από τις άλλες βιβλιοθήκες. Ο σχεδιασμός και η φιλοσοφία Mllib είναι απλή δηλαδή επιτρέπει να επικαλούνται διάφοροι αλγόριθμοι σε κατανεμημένα σύνολα δεδομένων, που αντιπροσωπεύουν όλα τα δεδομένα σε RDDs. Υλοποιήσαμε έναν αλγόριθμο (K-means) με την χρήση της βιβλιοθήκης Mllib όπως θα δούμε και στην υποενότητα 7.5.

5.2 Hadoop Distributed File System (HDFS)

Το σύστημα κατανομής αρχείων Hadoop (HDFS) είναι ένα κατανεμημένο σύστημα αρχείων που έχει σχεδιαστεί για να τρέχει σε όλα τα υπάρχοντα συστήματα Hardware [2]. Έχει πολλές ομοιότητες με τα υπάρχοντα κατανεμημένα συστήματα αρχείων. Ωστόσο, οι διαφορές από άλλα κατανεμημένα συστήματα αρχείων είναι σημαντικές. Το HDFS είναι εξαιρετικά ανεκτικό σε σφάλματα και έχει σχεδιαστεί για ανάπτυξη σε υλικό χαμηλού κόστους. Το HDFS παρέχει υψηλή πρόσβαση σε δεδομένα εφαρμογών και είναι κατάλληλο για εφαρμογές που έχουν μεγάλα σύνολα δεδομένων. Επίσης χαλαρώνει μερικές απαιτήσεις POSIX για να επιτρέψει τη ροή δεδομένων σε δεδομένα αρχείων. Το

HDFS δημιουργήθηκε αρχικά ως υποδομή για το πρόγραμμα μηχανών αναζήτησης του Apache Nutch και είναι μέρος του Apache Hadoop Core.

5.3 Γλωσσά προγραμματισμού Python - Pyspark

Ο προγραμματισμός των σύγχρονων συστημάτων μηχανικής μάθησης και cluster computing γίνεται συνήθως με Python. Ακόμη και όταν η Python δεν είναι η κύρια γλώσσα προγραμματισμού ενός τέτοιου συστήματος, υπάρχει ένα 'Python binding'. Χαρακτηριστικό παράδειγμα είναι το Spark που, αν και το πρωτεύον API του είναι σε Scala, κυρίως χρησιμοποιείται το Python API (PySpark). Η Python είναι επίσης η γλώσσα που χρησιμοποιούν πλέον οι περισσότεροι data scientists λόγω των εκτεταμένων βιβλιοθηκών έτοιμων αλγορίθμων μηχανικής μάθησης και επεξεργασίας δεδομένων διαφορετικών μορφών και μεγεθών.

5.4 FileZilla Client

Το πρόγραμμα FileZilla Client δίνει την δυνατότητα για την μεταφορά αρχείων από έναν υπολογιστή Α σε έναν άλλο μακρινό υπολογιστή Β. Το συγκεκριμένο εργαλείο χρησιμοποιήθηκε για την μεταφορά όλων μας των δεδομένων που είχαμε κατεβάσει από το zenodo στο vm Master όπου είχαμε φτιάξει στον okeano

5. Περιγραφή της αρχιτεκτονικής

6.1 Αρχιτεκτονική του Spark

Το Spark χρησιμοποιεί αρχιτεκτονική master/worker. Υπάρχει ένας driver που μιλάει με έναν συντονιστή που ονομάζεται master και διαχειρίζεται τους workers στους οποίους τρέχουν οι executors. Ένας master είναι ένα Spark instance που συνδέεται με έναν Cluster Manager για πόρους και αποκτά κόμβους της συστάδας για να εκτελέσει executors. Από την άλλη πλευρά, οι workers (γνωστοί και ως slaves) αποτελούν Spark instances στα οποία οι executors ζουν για να εκτελούν εργασίες. Αυτοί είναι οι υπολογιστικοί κόμβοι του Spark που επίσης επικοινωνούν μεταξύ τους χρησιμοποιώντας τα Block Manager instances που διαθέτουν [3]. Ο driver και οι executors τώρα, τρέχουν στις δικές τους Python διαδικασίες. Μπορούν να τρέξουν όλοι στην ίδια (οριζόντια συστάδα) ή σε ξεχωριστές μηχανές (κάθετη συστάδα) ή σε μικτή διαμόρφωση μηχανής. Όταν δημιουργείται ένα SparkContext, κάθε worker εκκινεί έναν executor. Οι executors συνδέονται πίσω στο πρόγραμμα του driver. Τώρα ο driver μπορεί να τους στείλει εντολές, όπως για παράδειγμα flatMap, map και reduceByKey. Όταν ο driver κλείσει, οι executors κλείνουν επίσης. Υπάρχουν τρείς διαφορετικοί τύποι cluster manager στο Apache Spark:

1. Spark-Standalone – Spark workers are registered with spark master

2. Yarn – Spark workers are registered with YARN Cluster manager.

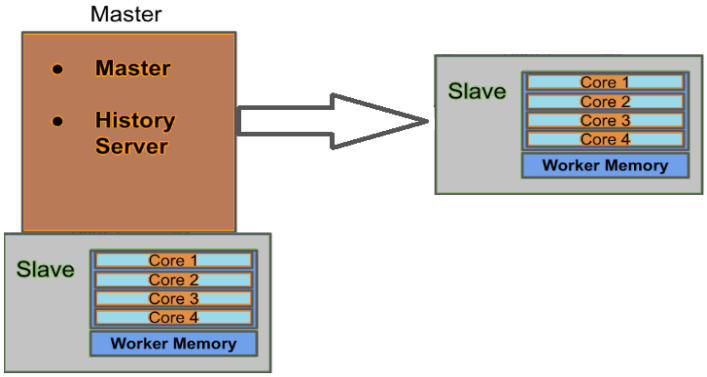
3. Mesos – Spark workers are registered with Mesos.

Στην συγκεκριμένη εργασία θα χρησιμοποιήσαμε ως cluster manager τον δεύτερο, δηλαδή το Yarn-SPARK. Με σκοπό την υλοποίηση της εργασίας μας δημιουργήσαμε δυο VMs στον okeano με τα έξης χαρακτηριστικά :

- Ubuntu Lts
- Cpu : x8
- Ram : 8
- Private network
- IPv6
- Storage Hdd : 40 Gb

Έπειτα κατεβάσαμε την 2.3.1 έκδοση του Spark και κάναμε τις ανάλογες ρυθμίσεις με βάση το Guide που μας δόθηκε. Τέλος ορίσαμε τον master και τους workers (slaves) όπως φαίνεται στην παρακάτω εικόνα .

Εικόνα 5.1 : Η αρχιτεκτονική του Spark cluster



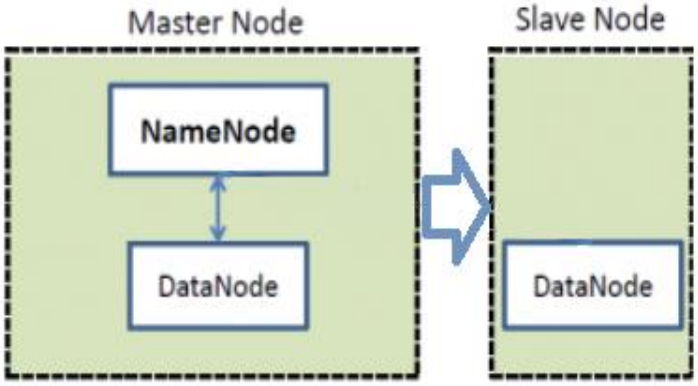
6.2 Αρχιτεκτονική του Hadoop

Το HDFS είναι ένα σύστημα σχεδιασμένο για την αποθήκευση πολύ μεγάλου όγκου δεδομένων, με υποστήριξη για πρότυπα προσπέλασης σε δεδομένα συνεχούς ροής (streaming data access patterns), σε cluster απλών υπολογιστών. Υπάρχουν σήμερα συστάδες Hadoop σε λειτουργία, που αποθηκεύουν Petabytes δεδομένων. Το HDFS είναι χτισμένο γύρω από την ιδέα ότι το πιο αποτελεσματική πρότυπο επεξεργασίας δεδομένων είναι αυτό της «εγγραφής-μια φορά, ανάγνωσης-πολλές φορές» (write-once, read-multiple times). Τυπικά, ένα σύνολο δεδομένων παράγεται ή αντιγράφεται από προϋπάρχουσα πηγή, και στη συνέχεια πραγματοποιούνται διάφορες αναλύσεις σε αυτό το σύνολο δεδομένων. Κάθε ανάλυση περιλαμβάνει ένα μεγάλο μέρος, αν όχι όλο, από το σύνολο των δεδομένων. Κατά συνέπεια, ο χρόνος για την ανάγνωση ολόκληρου του συνόλου δεδομένων είναι πιο σημαντικός από τον χρόνο ανάγνωσης της πρώτης εγγραφής. Στην δική μας περίπτωση έχουμε 2 VM μηχανήματα σε κάθε cluster τα οποία βρίσκονται στον cloud του okeanos . Είναι δυο Ubuntu LTS τα οποία βρίσκονται σε private network με NAT, σύμφωνα με το Guide για το ένα cluster κα για το δεύτερο cluster απλά με public IPs.

Στον master τρέχει ο NameNode , ο Secondary NameNode, ο NodeManager , ο ResourceManager και έχουμε βάλει και εκεί να τρέχει και ένας DataNode. Ενώ στον slave τρέχει μόνο ο NodeManager και ο Datanode

Στο πλαίσιο της εργασία μας ακλουθήσαμε το Guide για τις σωστές και λειτουργικές ρυθμίσεις ώστε να έχουμε το σύστημα του hdfs για την αποθήκευση των δεδομένων μας. Παρακάτω βλέπουμε μια εικόνα η οποία περιγράφει την αρχιτεκτονική του hdfs layer .

Εικόνα 5.2 : Αρχιτεκτονική Hdfs



6. Περιγραφή Του Κώδικα

Στο παρόν κεφάλαιο θα εξηγήσουμε με βήματα του κώδικα τον όποιο έχουμε υλοποιήσει για τον εντοπισμό των αλιευτικών πλοίων και μη, την ανάλυση χρονολογικής σειράς καθώς και την ομαδοποίηση που έγινε με τον αλγόριθμο k-means στα χωρικά δεδομένα.

Βήματα σύνδεσης με τα VM

Βήμα 1 : Συνδεόμαστε με το Vm Master μέσω cmd με την παρακάτω εντολή

- ssh -L 8000:localhost:8888 user@83.212.96.103

Βήμα 2 : Αφού έχουμε κάνει εγκατάσταση την IPython εκτελούμε την παρακάτω εντολή στον vm Master .

- ipython notebook --no-browser --port=7000

Βήμα 3 : Έπειτα ανοίξαμε ένα cmd οπού έδινε δικαίωμα στον υπολογιστή που χρησιμοποιούσαμε να έχει πρόσβαση στην πόρτα 6001 του VM Master

- ssh -N -f -L localhost:6001:localhost:7000 user@83.212.96.103

Ανοίξαμε ένα browser στο localhost:6001 ώστε να αρχίσουμε να γραφούμε των κώδικα μας.

- Τοποθέτηση αρχείων στο Hdfs

Βήμα 4 : Αφού μεταφέραμε τα αρχεία μας από το πρόγραμμα FileZillia στο Vm Master έπειτα συνεχίσαμε με την τοποθέτηση των δεδομένων μας στο Hdfs του Hadoop με την εντολή

- Hdfs dfs –put filename hdfs_path/user/user/Inputs

- Μέρος 1

Βήμα 5 :Κάναμε εισαγωγή των βιβλιοθηκών οι όποιες ήταν απαραίτητες για την εκτέλεση του κώδικα.

Βήμα 6 : Σύνδεση με το Spark server των οποίων δημιουργήσαμε 'spark://Master:7077' με spark.executor.memory : 3gb

Βήμα 7 : Διαβάζουμε τα 4 Data sets τα οποία αναφέραμε προηγουμένως από το hdfs

Βήμα 8 :Γίνεται μια περιγραφή ανάλυσης των δεδομένων αυτών με τις εντολές show και describe καθώς και ένα ραβδόγραμμα το όποιο αναφέρεται στην ταχύτητα των πλοίων εντος του συγκεκριμένου data set το όποιο θα το δούμε στο παρακάτω κεφάλαιο.

Βήμα 9 : Γίνετε μια τελική επιλογή για το ποιες θα είναι οι στήλες που θα χρησιμοποιήσουμε από κάθε data set ξεχωριστά ή θα μας φάνουνε χρήσιμες για την ανάλυση μας.

Βήμα 10 : Έπειτα καταχωρούμε σε μια μεταβλητή όλα τα διαφορετικά mmsi των πλοίων που έχουμε ώστε να κάνουμε ένα join πινάκα αυτού με τον πινάκα ο όποιος έχει τον τύπο του κάθε πλοίου .

Βήμα 11 : Από το Join ξέρουμε το κάθε mmsi που έχουμε το τι τύπου πλοίο είναι .Όλα αυτά έγιναν για να μπορέσουμε να εκτελέσουμε μια διαδικασία MapReduce ώστε να μετρήσουμε το πλήθος των διαφορετικών τύπων πλοίου που έχουμε .

- Μέρος 2

Στο data set το όποιο περιέχει τις περιοχές των δυο πολύγωνων θα δουλέψουμε με την ίδια μέθοδο .

Βήμα 12 : Εκχωρούμε σε μια μεταβλητή τα δεδομένα πλοίων τα οποία είναι εντός ενός νέου βέλτιστου παραλληλόγραμμου (Minimum bounding box) το όποιο δημιουργούμε εμείς με lat και lon να έχει τα bounds του παλιού πολυγώνου μας, δηλαδή το max και min των lat και lon. Αυτό έγινε για να έχουμε πιο γρήγορα αποτελέσματα για το συγκεκριμένο query μας .Το DataFrame το όποιο έχουμε πλέον με τα δεδομένα ais δεν είναι το πραγματικό διότι περιέχει παραπάνω δεδομένα για τον λόγο ότι πιάνει μια μικρή έκταση από περιοχή η οποία δεν είναι απαγορευμένη όπως θα δούμε και αργότερα στο έβδομο κεφάλαιο .Επίσης στην υποενότητα 8.3 θα κάνουμε μια επικύρωση αποτελεσμάτων .

Βήμα 13 : Παρουσιάζουμε τα δεδομένα Πίνακας 7.2 – Πίνακας 7.6.

Βήμα 14 : Υπολογίζουμε το πλήθος των πλοίων που είναι εντος της περιοχής

Βήμα 15 : Κάνουμε ένα join τον πινάκα των mmsi με το data set το όποιο περιέχει τον τύπο του κάθε πλοίου.

Βήμα 16 : Αφού ξέρουμε τους τύπους των πλοίων εντός της περιοχής , τρέχουμε την διαδικασία MapReduce ώστε να βρούμε το πλήθος των διαφορετικών πλοίων εντός της περιοχής και φτιαχτούμε τα ραβδογράμματα με την βοήθεια της βιβλιοθήκης ggplot στην R

Βήμα 17 :Για να βρούμε τα πλοία τα οποία είναι αλιευτικά εντος της περιοχής αυτής ορίσαμε δυο πιθανούς τρόπους

Ο $1^{ος}$ είναι : Συλλέγουμε τα πλοία τα οποία έχουν την στήλη navigationalstatus =7 (‘κατάσταση αλίευσης’)

Ο $2^{ος}$ είναι : Συλλέγουμε τα πλοία τα οποία έχουν την στήλη navigationalstatus =7 και η ταχύτητα τους να είναι μικρότερη από 2 κόμβους .

Βήμα 18 :Υπολογίζουμε το πλήθος των πλοίων τα οποία εκπληρούν τις παραπάνω προϋποθέσεις

Βήμα 19 :Παρουσιάζουμε τα mmsi και τα ονόματα των πλοίων αυτών

Βήμα 20 :Δημιουργούμε μια συνάρτηση οπού κάνουμε απεικόνιση μερικών πλοίων στα οποία βλέπουμε την διαδρομή που ακλουθήσανε με μισή ώρα νωρίτερα και αργότερα από την στιγμή που τα εντοπίσαμε καθώς και τα χάσαμε αντίστοιχα από την περιοχή

Βήμα 21 :Τέλος δημιουργούμε μια συνάρτηση οπού κάνουμε απεικόνιση των πλοίων οπού βρεθήκαν εντος των απαγορευμένων περιοχών καθ' όλη την διάρκεια οπού έχουν κινηθεί σε αυτούς τους έξι μήνες οπού αναφέρονται τα δεδομένα μας .Με περιορισμό να μας δείχνει μόνο τα στίγματα οπού η ταχύτητα των πλοίων να είναι μικρότερη από 2 κόμβους , ώστε να δούμε τα μέρη οπού πάνε τα αλιευτικά πλοία για αλίευση.

- Μέρος 3

Βήμα 22 : Αφού έχουμε πλέον υπολογίσει τα πλοία εντος των περιοχών αυτών και έχουμε μειώσει κατά πολύ τα δεδομένα τα οποία θα μας είναι χρήσιμα , εκτελούμε την εντολή within μόνο στην master

(local) ώστε να κάνουμε μια επικύρωση των δεδομένων μας για τα πραγματικά στίγματα των πλοίων που ήταν εντος της περιοχής .

- Ανάλυση χρονολογικής σειράς

Βήμα 23 : Υπολογίζουμε με βάση την στήλη t την οποία έχουμε μετατρέψει σε timestamp date το πλήθος των πλοίων που έχουν εισέρθει εντος των περιοχών αυτών ξεχωριστά για κάθε μια, και έπειτα τα αναπαριστούμε σαν χρονοσειρα.

Βήμα 24 : Κάνουμε τους προαπαιτούμενους έλεγχους όπως είναι ο έλεγχος της στασιμότητας , ώστε να μπορούμε να χρησιμοποιήσουμε μερικά μοντέλα χρονολογικών σειρών.

Βήμα 25 : Επειδή για μια από τις δυο περιοχές αυτή η χρονολογική σειρά η όποια πήραμε δεν ήταν στάσιμη , χρειάστηκε να κάνουμε μερικά τεχνάσματα όπως π.χ ο μετασχηματισμός καθώς και άλλα, ώστε να την μετατρέψουμε σε στάσιμη χρονολογική σειρά για να εκτελέσουμε τα κατάλληλα μοντέλα.

Βήμα 26 : Τέλος κάνουμε μια πρόβλεψη με το μοντέλο AR για τον τελευταίο μήνα τον οποίο ήδη έχουμε τα δεδομένα ώστε να δούμε την προβλεπτική ικανότητα του μοντέλου.

- K-means

Βήμα 27 : Δημιουργούμε μια νέα μεταβλητή η όποια θα περιέχει τα δεδομένα ais των πλοίων που έχουν την στήλη navigationalstatus =7 και την ταχύτητα τους μικρότερη από 1 κόμβο.

Βήμα 28 : Απεικονίζουμε τα δεδομένα αυτά σε συνδυασμό με τις επιτρεπόμενες περιοχές αλιείας ώστε να δούμε της συμπεριφορά τους εντος των περιοχών αυτών καθώς και τα μέρη τα όποια συνηθίζουν να επισκέπτονται.

Βήμα 29 : καθώς πήραμε τα αποτελέσματα από το παραπάνω βήμα παρατηρήσαμε πως μπορούμε να κάνουμε μια ομαδοποίηση των δεδομένων αυτόν για τον λόγο ότι υπήρχε μεγαλύτερη πυκνότητα επισκεψιμότητας σε μόνο 3 έκτασης από όλη την περιοχή αυτή.

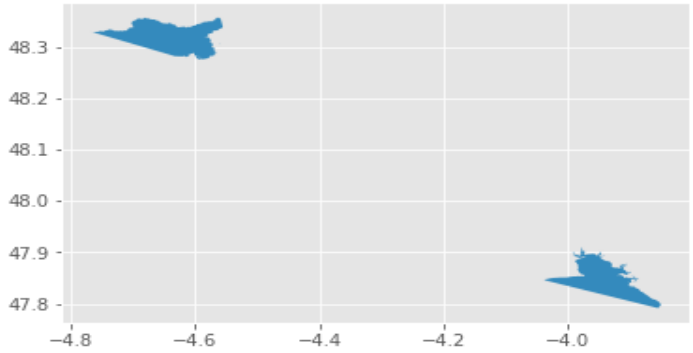
Βήμα 30 :Τέλος εισήγαμε την βιβλιοθήκη Mllib και εκτελέσαμε τον αλγόριθμο k-means με τους κατάλαλους ελέγχους.

7. Αποτελέσματα και Επικύρωση Αποτελεσμάτων

Στο κεφάλαιο αυτό θα παρουσιάσουμε τα αποτελέσματα από την ανάλυση της μελέτης μας πάνω στο θέμα του εντοπισμού αλιευτικών πλοίων εντος απαγορευμένων περιοχών καθώς και την ανάλυση της χρονολογικής σειράς η οποία βασίζεται στο πλήθος των πλοίων που είναι εντος των περιοχών αυτών σε καθημερινή βάση.

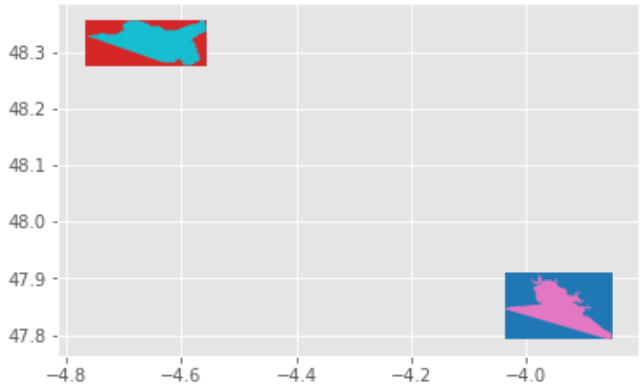
Στην Εικόνα 7.1 βλέπουμε τις δυο περιοχές οι όποιες είναι και στην Εικόνα 3.1 στο τέταρτο κεφάλαιο αλλά με την βιβλιοθήκη matplotlib.

Εικόνα 7.1



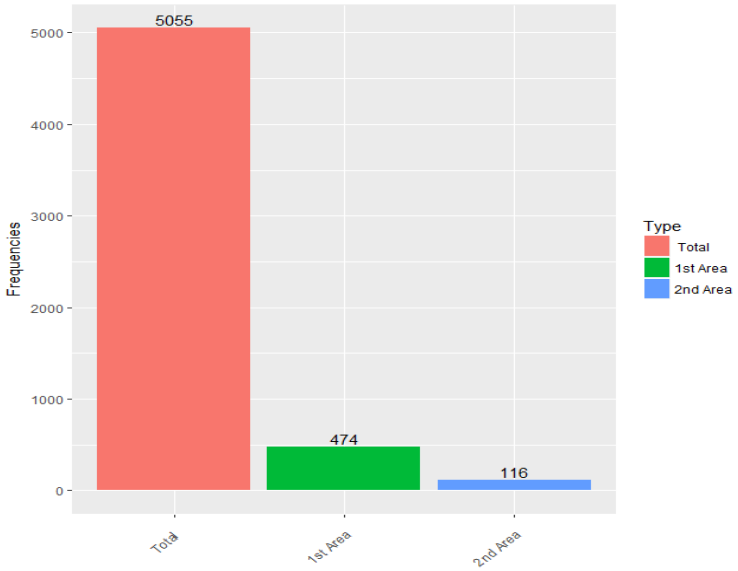
Όπως αναφέραμε και στο έβδομο κεφάλαιο , η προσέγγιση μας στο πρόβλημα του εντοπισμου αλιευτικών πλοίων εντος των περιοχών εγινε δημιουργόντας δυο νεα πολυγωνα τα οποια είναι με την προσεγγιση minimum bounding box όπως βλέπουμε στην παρακάτω Εικονα 7.2 με σκούρο κόκκινο και μπλε χρώμα στα οποία έχουμε πλεονεκτήματα αλλά και μειονεκτήματα. Τα πλεονεκτήματα είναι ότι το query που κάναμε είναι πιο γρήγορα και πιο σαφές διότι ρωτάμε εάν τα στήγματα των πλοίων είναι μεγαλύτερα ή μικρότερα από τέσσσερα σημεία , όπου τα σημεία αυτα είναι τα max και min των lat και lon των πολυγώνων.Το μειονέκτημα που έχουμε είναι ότι έχουμε και μερικά στήγματα πλοίων τα οποία δεν βρίσκονται εντός των πραγματικών πολυγώνων και ενδεχομένως πλοία τα οποία δεν είναι εντός των περιοχών αυτων αλλα εντοπίστηκαν λόγω των καινούριων πολυγώνων

Εικονα 7.2 : Παρακάτω βλέπουμε τα παραλληλόγραμα του minimum bounding box



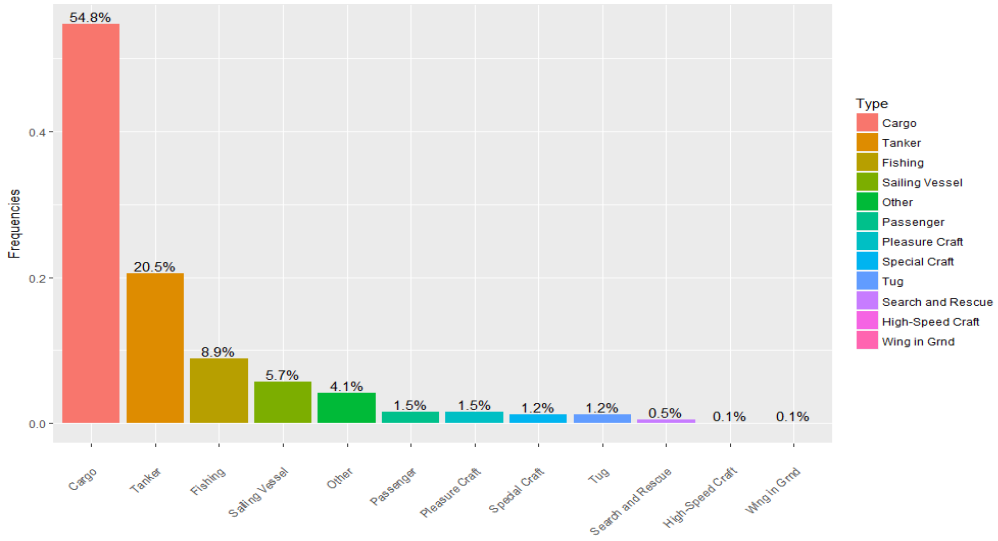
Στο Διάγραμμα 7.1 βλέπουμε με κόκκινο το πλήθος των συνολικών πλοίων που έχουμε στο data set μας, με πράσινο τα πλοία που είναι εντος της πρώτης απαγορευμένης περιοχής .Τέλος με μπλε είναι το πλήθος των πλοίων που βρεθήκαν εντος της δεύτερης περιοχής.

Διάγραμμα 7.1



Στο Διάγραμμα 7.2 βλέπουμε τα ποσοστά των διάφορων τύπου πλοίων που έχουμε στα δεδομένα μας, οπού συνολικά τα πλοία είναι 5055 όπως φαίνεται και παραπάνω .Παρατηρούμε πως το μεγαλύτερο ποσοστό αυτών είναι τύπου Cargo και αμέσως μετά είναι Tanker .Τρίτο έρχεται σε κατάταξη τα πλοία με τα οποία θα ασχοληθούμε δηλαδή αλιεύτηκα πλοία (Fishing).

Διάγραμμα 7.2



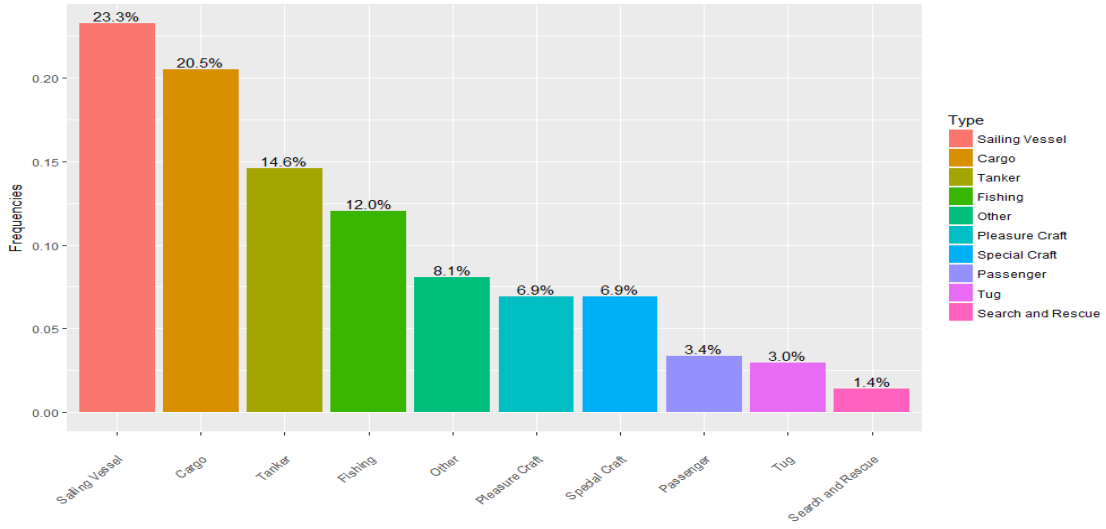
7.1 Πρώτη Περιοχή

Πίνακας 7.1 : Περιγραφικά μετρά των δεδομένων

summary	sourcemmsi	navigationalstatus	speedoverground	lon	lat	t
count	1120332	1012222	1120332	1120332	1120332	11203 32
mean	2.5322901677436063E8	4.530944792743094	19.277535230621396	-4.639560669920763	48.31314135079534	1.451673392044601 E9
stddev	9.995868429258257E7	6.576427849450526	28.63178526498597	0.05853057682286594	0.021470672859597348	4717799.5428278 56
min	205204000	0.0	0.0	-4.76704	48.275040000000004	14436511 81
max	999999999	15.0	102.3	-4.55494	48.356182000000004	14594615 67

Από ότι παρατηρούμε από τον πίνακα 7.1 το σύνολο των γραμμών είναι 1.120.332 και το πλήθος των πλοίων ανέρχεται στα 474 όπως είδαμε και προηγουμένως .Στο Διάγραμμα 7.3 βλέπουμε τα ποσοστά από τους διαφόρους τύπους πλοίων που έχουμε εντος της πρώτης περιοχής

Διάγραμμα 7.3 : Τα ποσοστά των διαφορετικών τύπων πλοίου στην πρώτη περιοχή (Βήμα 15)



Στον πινάκα 7.2 βλέπουμε μερικά από τα πλοία τα οποία βρεθήκαν εντος της περιοχής 1 .Τα αποτελέσματα αυτού του πινάκα πάρθηκαν από το Join που κάναμε στα δεδομένα που βρεθήκαν εντος της περιοχής από τα data set nari_dynamic καθώς και τον τύπο των πλοίων που ήταν στο data set nari_static .

Πινάκας 7.2

sourcemmsi	shipname	ais_type_summary	shiptype
227312180	F/V MARIE-LOU III	Fishing	30
244780246	BRUNEL	Sailing Vessel	36
227566870	JOANNA	Sailing Vessel	36
235034658	ELLIPTIC	Sailing Vessel	36
245257000	EEMS CARRIER	Cargo	70
249993000	BESIKTAS PERA	Tanker	80
205204000	NATO WARSHIP A960	Special Craft	35
227306100	LA RECOUVRANCE	Sailing Vessel	36
276700000	NATO WARSHIP M313	Special Craft	35
311164000	ATLANTICA HAV	Cargo	70
245547000	ABIS DOVER	Cargo	70
227337570	MAGIC KILI	Sailing Vessel	36
228267900	F/V FELIR	Fishing	30
227686540	EMERAUDE	Fishing	30
235003780	TAMARIND	Sailing Vessel	36
269669000	SCL ANITA	Cargo	70
227062780	BOURRE PIF	Special Craft	35
246058000	SIF R	Special Craft	33
265756690	S/Y TARANTELLA	Sailing Vessel	36
311000369	SKANDI AFRICA	Cargo	70

Στο πινάκα 7.3 βλέπουμε τα πλοία τα οποία έχουνε την στήλη Navigationalstatus ίσο με 7 δηλαδή κατάσταση αλίευσης ,πράγμα που σημαίνει πως ψαρεύανε εκείνη την στιγμή ή περνάγανε από την περιοχή αυτή ενώ είχαν πάει για αλίευση ή θα πήγαιναν .Η συγκεκριμένη περιοχή από ότι έχουμε δει στην Εικόνα 3.1 είναι πέρασμα , αρά πολύ πιθανό είναι απλώς να περνάγανε από αυτό το πολύγωνο.

Πινάκας 7.3

sourcemmsi	shipname	shiptype
227867000	F/V KREIZ AR MOR	30
227391000	F/V EFFERA	30
226084000	F/V KADEGE	30
227392000	F/V ALYA	90
228827000	F/V.OCEANE	30
227347000	F/V IROISE	30
226216000	FV DAMIEN FLORENT	30
227577000	F/V JUDINE	30
227091000	COTES DE LA MANCHE	90

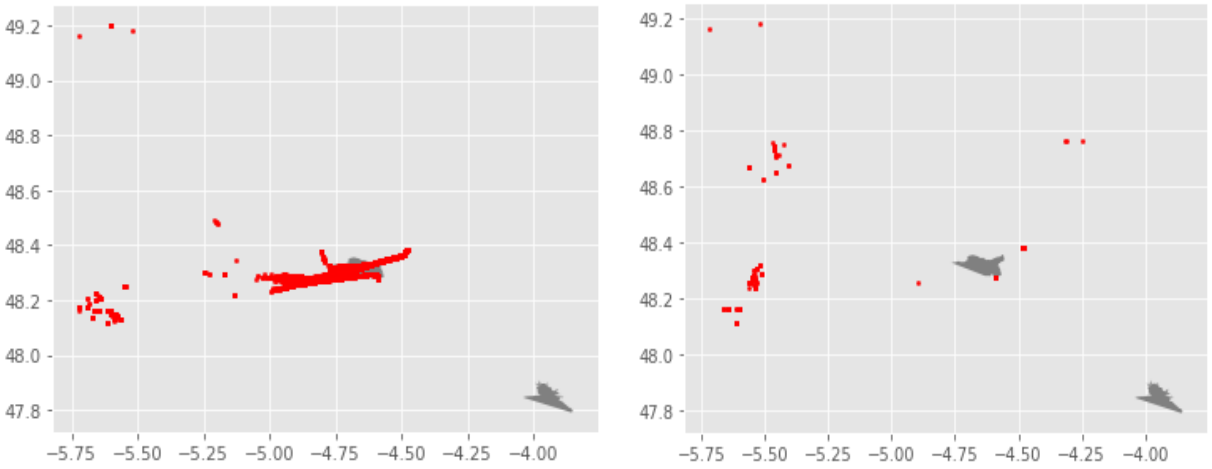
Στον πινάκα 7.4 βλέπουμε ένα και μοναδικό πλοίο το όποιο βρέθηκε σε κατάσταση αλίευσης καθώς και η ταχύτατα του ήταν κάτω από 2 κόμβους ,πράγμα που το καθίστα πιο πιθανό να αλιεύει στην περιοχή αυτή .

Πινάκας 7.4

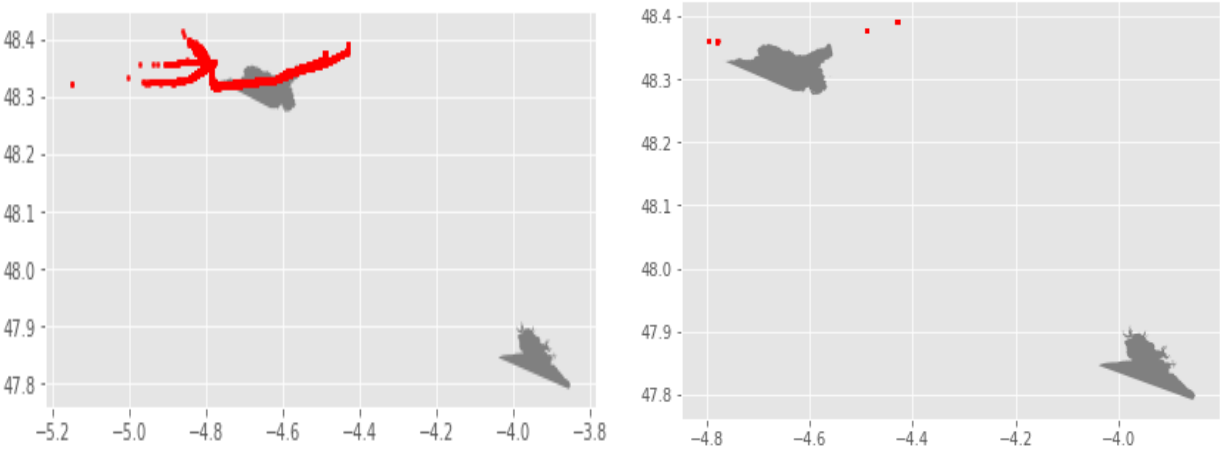
sourcemmsi	shipname	shiptype
228827000	F/V.OCEANE	30

Στις τρείς παρακάτω αριστερές εικόνες βλέπουμε την συμπεριφορά τριών πλοίων στο χρόνο μισή ώρα πριν και μετά την στιγμή που τα εντοπίσαμε στο πολύγωνο μας. Από ότι βλέπουμε κανένα δεν δίνει να έχει σκοπό για αλίευση παρά μόνο να περάσει από την περιοχή αυτή .Ένα πρόβλημα είναι πως στον αλγόριθμο μας είναι ότι δεν αντιμετωπίζουμε τα δεδομένα μας ως χρονοσειρα π.χ έναν ένα πλοίο έχει μπει δυο ή παραπάνω φόρες, ως ελάχιστο χρόνο θα πάρει την πρώτη φορά που μπήκε και ως μέγιστο την τελευταία φορά που μπήκε. Για τον λόγο αυτό στις δεξιές εικόνες θα δούμε τις συνήθειες αυτών των πλοίων όταν αλιεύουν. Έχουμε ορίσει εξαρχής πως ένα πλοίο για να αλιεύει θα πρέπει να έχει το Navigationalstatus ίσο με 7 και ταχύτατα κάτω από 2 κόμβους.

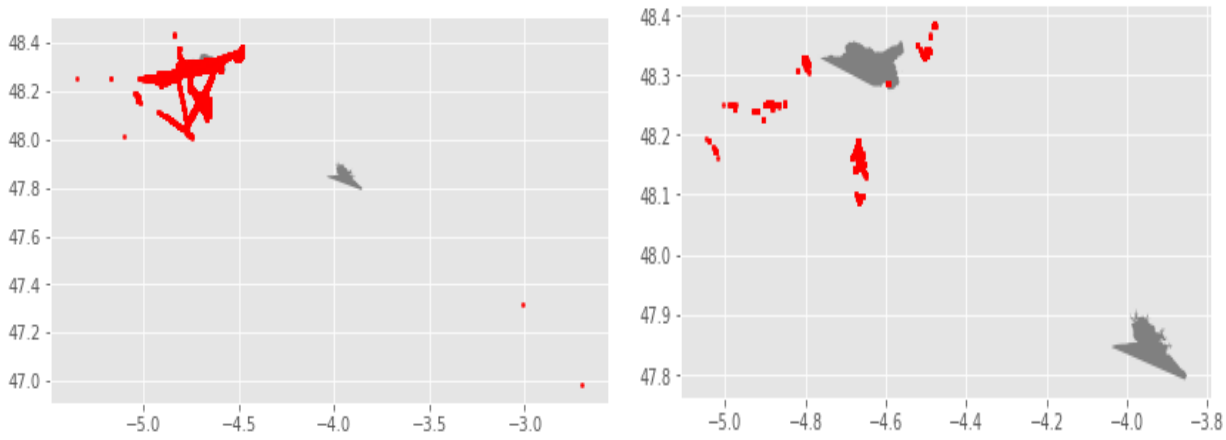
Εικόνα 7.3: MMSI - 228827000



Εικόνα 7.4: MMSI – 226084000



Εικόνα 7.5 : MMSI - 227091000



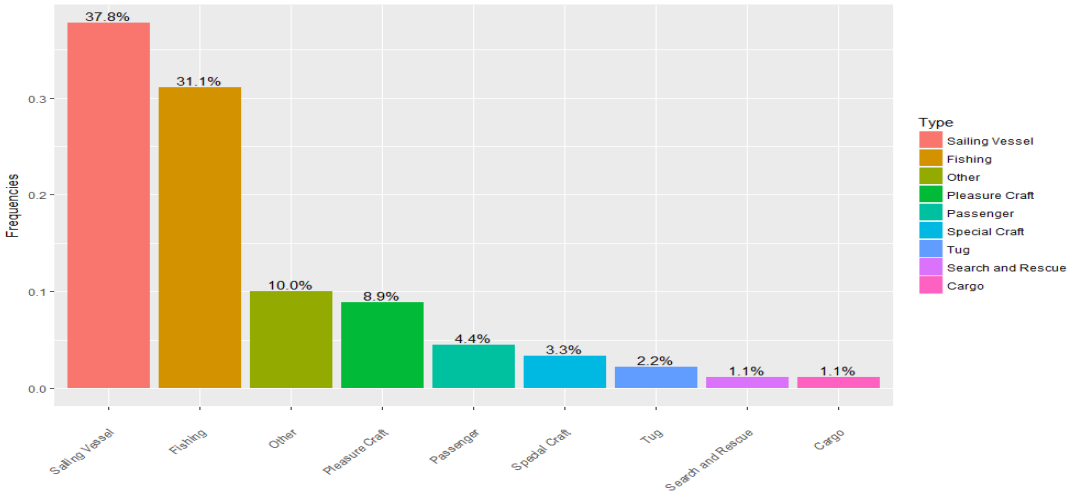
7.2 Δεύτερη Περιοχή

Από ότι παρατηρούμε από τον πίνακα 7.5 το σύνολο των γραμμών είναι 5.300 δηλαδή αρκετά μικρότερο σε σύγκριση με την πρώτη περιοχή καθώς και το πλήθος των πλοίων είναι 116 όπως είδαμε και προηγουμένως .Στο Διάγραμμα 7.4 βλέπουμε τα ποσοστά από τους διαφόρους τύπους πλοίων που έχουμε εντος της δεύτερης περιοχής .

Πίνακας 7.5 : Παρακάτω βλέπουμε περιγραφικά μετρά των δεδομένων

summary	sourcemmsi	navigationalstatus	speedoverground	lon	lat	t
count	5300	4663	5300	5300	5300	5300
mean	2.2915099013698113E8	9.788548144971049	6.550415094339629	-3.9481803897547207	47.83532737830194	1.453776994846415E9
stddev	5860512.264012891	6.907452010684565	3.974003774361546	0.02852466354202299	0.024807102754329673	4583915.890535743
min	211232180	0.0	0.0	-4.03829	47.791393	1443653640
max	319064400	15.0	25.2	-3.8545084000000003	47.900771999999996	1459461224

Διάγραμμα 7.4 : Τα ποσοστά των διαφορετικών τύπων πλοίου στην δεύτερη περιοχή



Στον πινάκα 7.6 βλέπουμε μερικά από τα πλοία τα οποία βρεθήκαν εντος της δεύτερης περιοχής. Ακολουθήσαμε ακριβώς την ίδια μεθοδολογία όπως και στην πρώτη περιοχή .

Πινάκας 7.6

sourcemmsi	shipname	ais_type_summary	shiptype
228240800	F/V MAILYS CHARLIE	Fishing	30
227306100	LA RECOUVRANCE	Sailing Vessel	36
227798280	JASLAND 2	Pleasure Craft	37
211232180	IZAR	Sailing Vessel	36
235081443	OC23 SOLO SAILOR	Sailing Vessel	36
235081443	ARTEMIS23 SOLOBAKER	Sailing Vessel	36
227962000	ALIXEL 2	Fishing	30
228387000	F/V LIBERTY-S	Fishing	30
227567680	PERZEL	Fishing	30
228218000	FRUGY	Fishing	30
228762000	FRENCH WARSHIP	Special Craft	35
227570000	F/V IBARDIN	Fishing	30
227191950	INITIATIVES SOLO	Sailing Vessel	36
228066900	GEVRED	Fishing	30
246774000	EN AVANT 10	Tug	52
228065900	BANQUE POPULAIRE 8	Sailing Vessel	36
227383920	EXCALIBUR	Sailing Vessel	36
227548120	CYRANO	Sailing Vessel	36
227688950	PECAB	Sailing Vessel	36
227701720	FOXY F EWE	Sailing Vessel	36

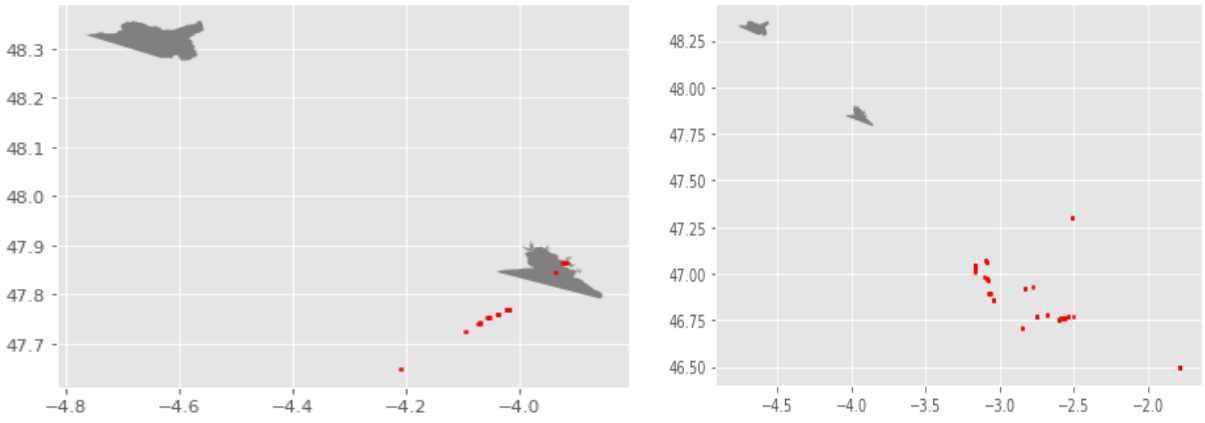
Στο πινάκα 7.7 βλέπουμε τα πλοία τα οποία είχανε την στήλη Navigationalstatus ίσο με 7 δηλαδή κατάσταση αλίευσης ,πράγμα που σημαίνει πως ψαρεύανε εκείνη την στιγμή ή περνάγανε από την περιοχή αυτή ενώ είχαν πάει για αλίευση ή θα πήγαιναν .

Πινάκας 7.7

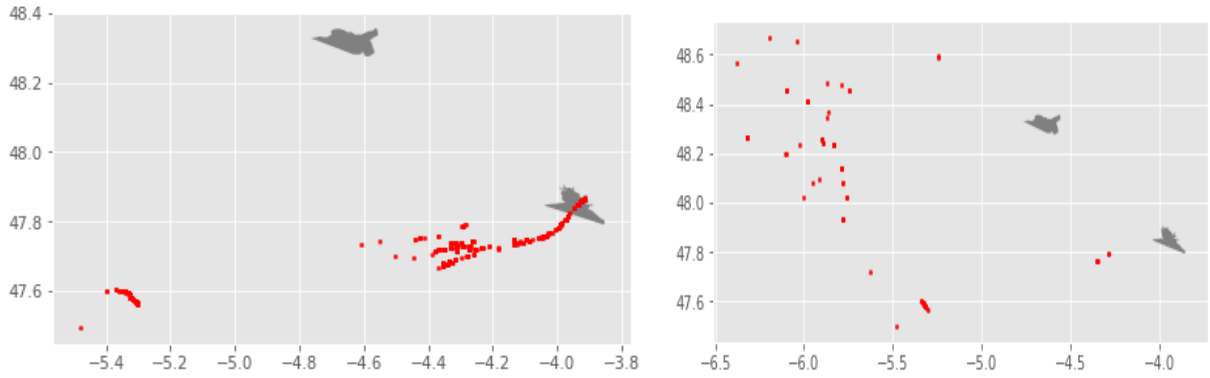
sourcemmsi	shipname	shiptype
227474000	F/V KERFLOUS	30
228042600	F/V MABON III	30
228126000	F/V L OCARINA	30
228910000	F/V L OKEANOS	90
228355800	F/V LUANTSA	30

Στις δυο παρακάτω αριστερά εικόνες βλέπουμε την συμπεριφορά δυο πλοίων στο χρόνο μισή ώρα πριν και μετά την στιγμή που τα εντοπίσαμε στο πολύγωνο μας. Από ότι βλέπουμε κανένα δεν δείχνει να έχει σκοπό να αλιεύσει διότι κατευθύνονται προς τον λιμένα όπως μπορούμε να παρατηρήσουμε στη εικόνα 3.1. Στις δεξιές εικόνες θα δούμε τις συνήθειες αυτών των πλοίων όταν αλιεύουν. Παρατηρούμε πως οι χώροι οι όποιοι αλιεύουν τα συγκεκριμένα πλοία είναι πολύ μακριά από το πολύγωνο μας.

Εικόνα 7.6 : MMSI – 228042600



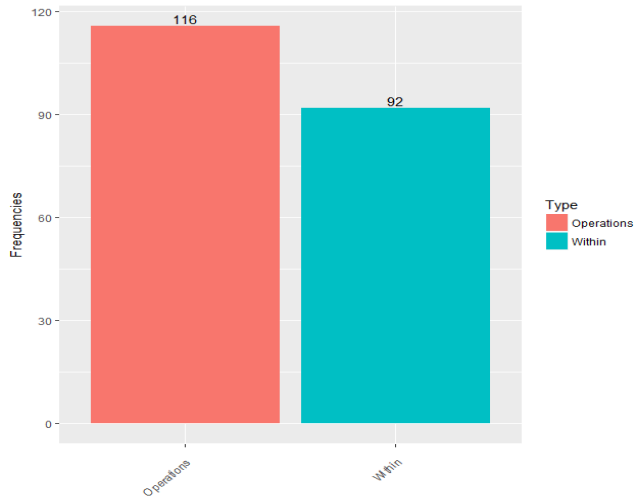
Εικόνα 7.7 MMSI - 227474000



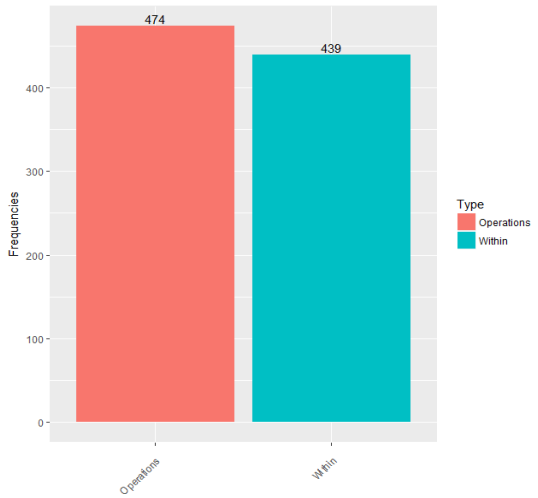
7.3 Επικύρωση αποτελεσμάτων

Καθώς τα πλέγματα που δημιουργήσαμε γύρω από τις απαγορευμένες περιοχές αλίευσης δεν είναι οι πραγματικές, στην υποενότητα αυτή θα κάνουμε μια επικύρωση των αποτελεσμάτων μας με στόχο να βρούμε ποσά πλοία βρίσκονται πραγματικά εντος των περιοχών αυτών. Η μεθοδολογία που ακολουθήσαμε είναι : Αφού ήδη είχαμε βρει τα δεδομένα των πλοίων που ήταν εντος των περιοχών αυτών ,οπού οι πραγματικές περιοχές είναι ένα υποσύνολο των καινούριων περιοχών μας, στην συνεχεία μετατρέψαμε τα δεδομένα αυτά σε Pandas DataFrame και εκτελέσαμε την συνάρτηση Within η οποία μας δίνει True εάν το points στην συγκεκριμένη γραμμή είναι εντος και False εάν όχι. Αφού πλέον έχουμε τις πραγματικές γραμμές που βρίσκονται εντος, τότε γνωρίζουμε το πραγματικό πλήθος των πλοίων που βρίσκονται μέσα μέσω της συνάρτησης unique. Στα διαγράμματα 7.5 και 7.6 βλέπουμε με μπλε χρώμα το πλήθος του πραγματικού πλήθους των πλοίων εντος της κάθε μια περιοχής και με κόκκινο το πλήθος το οποία ήταν εντος του πλέγματος που δημιουργήσαμε .

Διάγραμμα 7.5: Πρώτη Περιοχή



Διάγραμμα 7.6 Δεύτερη Περιοχή



7.4 Ανάλυσης χρονολογικής σειράς

Με αφορμή που ασχοληθήκαμε με τις συγκεκριμένη περιοχές, ενδιαφέρον θα ήταν να δούμε στο βάθος του χρόνου το πλήθος των πλοίων που εισέρχονται εντος αυτών. Όπως υποδηλώνει το όνομα, η χρονολογική σειρά είναι μια συλλογή σημείων δεδομένων που συλλέγονται σε σταθερά χρονικά διαστήματα. Αυτά αναλύονται για να καθορίσουν τη μακροπρόθεσμη τάση έτσι ώστε να προβλέπουν το μέλλον ή να εκτελούν κάποια άλλη μορφή ανάλυσης. Αλλά αυτό που κάνει μια χρονολογική σειρά διαφορετική από ένα κανονικό πρόβλημα παλινδρόμησης είναι τα έξης δυο:

1. Είναι εξαρτημένο από το χρόνο. Έτσι, η βασική παραδοχή ενός μοντέλου γραμμικής παλινδρόμησης που οι παρατηρήσεις είναι ανεξάρτητες δεν ισχύει στην περίπτωση αυτή.
2. Μαζί με μια αυξανόμενη ή μειούμενη τάση, οι περισσότερες χρονολογικές σειρές έχουν κάποια μορφή τάσεων ή εποχικότητας, δηλ. συγκεκριμένες αλλαγές τιμών σε ένα συγκεκριμένο χρονικό πλαίσιο.

Λόγω των ιδιοτήτων ενός χρονολογικού μοντέλου , υπάρχουν διάφορα βήματα που εμπλέκονται στην ανάλυση του. Αυτό θα εξετάσουμε παρακάτω. Στην συγκεκριμένη ερευνά χρήσιμο θα ήταν να περιγράψουμε και να αναφέρουμε τα μοντέλα τα όποια θα δούμε παρακάτω.

Αυτοπαλινδρομο Υπόδειγμα

Αυτοπαλινδρομο υπόδειγμα p τάξης συμβολίζετε με AR(p) και εκφράζετε από την σχέση :

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t$$

Όπου α_ι με $i \in [0, p]$ είναι οι σταθεροί παράμετροι και ε_t ο οποίος μετρά τυχαία σφάλματα. Στο αυτοπαλινδονουμενο υπόδειγμα , η εξαρτώμενη μεταβλητή Y_t παλινδρομείται στις προηγούμενες τιμές της . Η τάξη του αυτοπαλινδρονουμενου υποδείγματος συμβολίζεται με p και προσδιορίζει το μήκος της υστέρησης .

Υπόδειγμα Κινητού Μέσου

Ένα υπόδειγμα κινητού μέσου q τάξης συμβολίζεται με MA(q) και εκφράζετε από την σχέση :

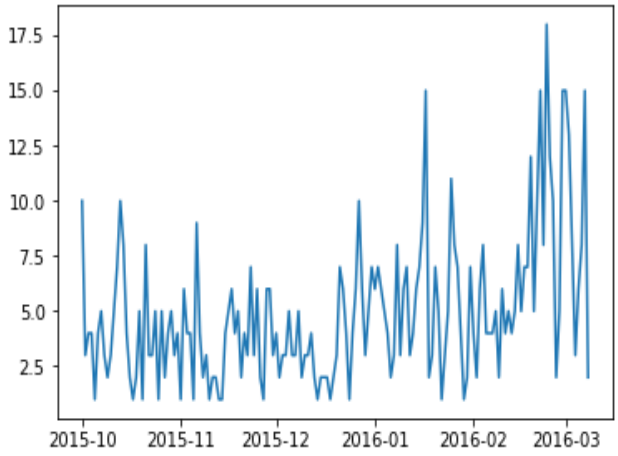
$$Y_\tau = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

Όπου θ_i με $i \in [0, q]$ οι σταθεροί παράμετροι και ε_t τα τυχαία σφάλματα .Στην διαδικασία κινητού μέσου η χρονολογική σειρά Y_t θεωρείται ότι δημιουργείται ως ένας σταθμικός μέσος τυχαίων σφαλμάτων των q προηγουμένων περιόδων .

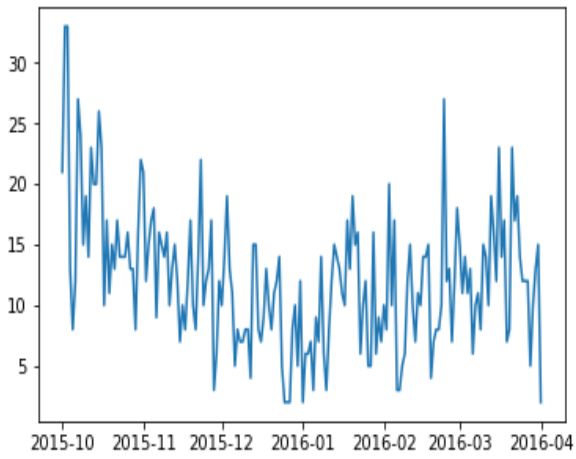
Ένα αυτοπαλινδρομο ολοκληρωμένο υπόδειγμα κινητού μέσου ARIMA(p,d,q) προκύπτει από τον συνδυασμό των αυτοπαλινδρομον διαδικασιών AR(p) και των διαδικασιών κινητού μέσου MA(q), όπου d είναι ο αριθμός των διαφόρων που απαιτούνται προκείμενου να μετατραπεί η σειρά σε στάσιμη

Αφού πλέον έχουμε υπολογίσει το πλήθος των πλοίων που είναι εντος και των δυο περιοχών και αναφέραμε τα μοντέλα τα όποια θα χρησιμοποιηθούν, θα δούμε παρακάτω τα δυο διαγράμματα σε μορφή χρονολογικών σειρών .

Διάγραμμα 7.7 :Δεδομένα δεύτερης περιοχής



Διάγραμμα 7.8: Δεδομένα πρώτης περιοχής

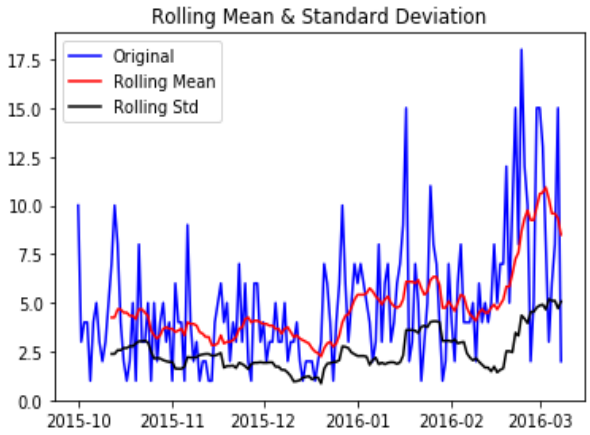


Μια χρονολογική σειρά είναι στάσιμη αν οι στατιστικές τις ιδιότητες όπως η μέση και η διακύμανση παραμένουν σταθερές με την πάροδο του χρόνου. Η στασιμότητα σε μοντέλα χρονολογικών σειρών είναι σημαντική διότι τα περισσότερα από τα μοντέλα αυτά λειτουργούν με την παραδοχή ότι είναι στάσιμες. Διαισθητικά, μπορούμε να υποθέσουμε ότι εάν μια χρονολογική σειρά έχει μια συγκεκριμένη συμπεριφορά με την πάροδο του χρόνου, υπάρχει πολύ μεγάλη πιθανότητα ότι θα ακολουθήσει το ίδιο και στο μέλλον. Παρακάτω θα εκτελέσουμε μερικούς ελέγχους στασιμότητας οι όποιοι είναι οι έξης :

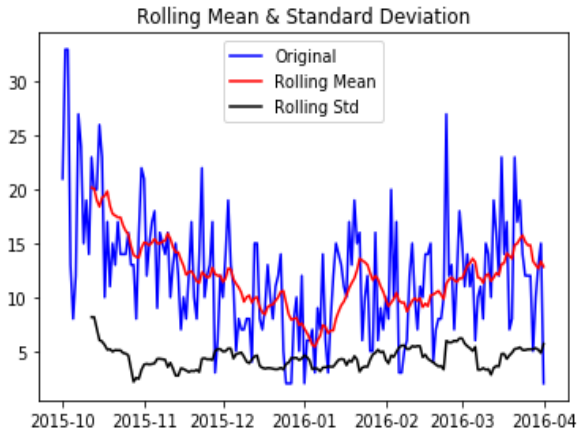
Απεικόνιση Στατιστικών Στοιχείων : Σχεδιάζουμε τον κινητό μέσο όρο και τη διακύμανση για να δούμε αν διαφέρει ανάλογα με το χρόνο. Με τη μετακίνηση μέσου όρου και την διακύμανσης όπου βασικά εννοούμε ότι σε κάθε στιγμή t, θα πάρουμε το μέσο όρο και την διακύμανση των προηγούμενων τιμών. Αλλά και πάλι αυτό είναι περισσότερο μια οπτική τεχνική.

Δοκιμή Dickey-Fuller: Αυτή είναι μία από τις στατιστικές δοκιμές για τον έλεγχο της στασιμότητας. Εδώ η μηδενική υπόθεση είναι ότι η χρονολογική σειρά είναι μη στατική. Τα αποτελέσματα των δοκιμών περιλαμβάνουν την τιμή της στατιστικής συνάρτησης, το P-value καθώς και ορισμένα διαστήματα εμπιστοσύνης. Εάν η «P-value» είναι μικρότερη από την «επίπεδο σημαντικότητας » το οποίο ορίζουμε εμείς , μπορούμε να απορρίψουμε τη μηδενική υπόθεση και να πούμε ότι η σειρά είναι στάσιμη.

Διάγραμμα 7.9 : Δεύτερη περιοχή



Διάγραμμα 7.10 : Πρώτη περιοχή



Results of Dickey-Fuller Test:	
Test Statistic	-2.111879
p-value	0.239765
#Lags Used	6.000000
Number of Observations Used	153.000000
Critical Value (5%)	-2.880623
Critical Value (1%)	-3.473830
Critical Value (10%)	-2.576945

Results of Dickey-Fuller Test:	
Test Statistic	-2.897765
p-value	0.045609
#Lags Used	6.000000
Number of Observations Used	153.000000
Critical Value (5%)	-2.880623
Critical Value (1%)	-3.473830
Critical Value (10%)	-2.576945

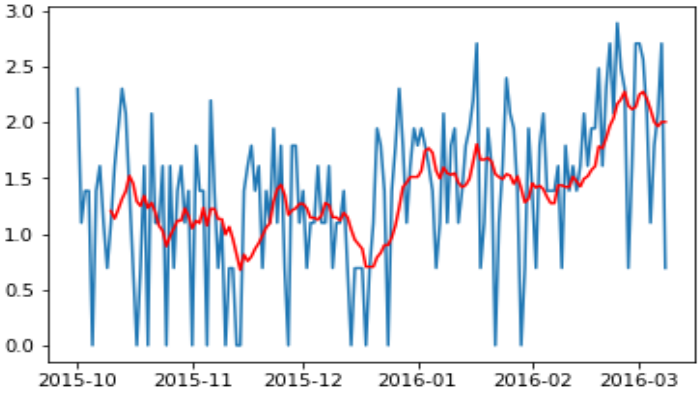
Στην δεύτερη περιοχή αν και η διακύμανση της τυπικής απόκλισης είναι μικρή, ο μέσος όρος αυξάνεται με την πάροδο του χρόνου και για αυτό το λόγο η χρονολογική σειρά μας δεν είναι στάσιμη. Επίσης η τιμή του p-value είναι μεγαλύτερη από το επίπεδο σημαντικότητας 5% που έχουμε ορίσει .Στην πρώτη περιοχή ισχύει το κριτήριο της στασιμότητας διότι το p-value είναι μικρότερο από το επίπεδο σημαντικότητας . Στη συνέχεια, θα συζητήσουμε τις τεχνικές που μπορούν να χρησιμοποιηθούν για να γίνουν τα δεδομένα της πρώτης περιοχής μια στάσιμη χρονολογική σειρά , ώστε να ισχύουν τα αποτελέσματα από της εκτίμησης του μοντέλου.

Ένα από τα πρώτα κόλπα για τη μείωση της τάσης είναι ο μετασχηματισμός. Για παράδειγμα, στην περίπτωση αυτή μπορούμε να διαπιστώσουμε σαφώς ότι υπάρχει μια σημαντική ανοδική τάση στα δεδομένα της πρώτης περιοχής. Έτσι μπορούμε να εφαρμόσουμε μετασχηματισμό που τιμωρεί τις υψηλότερες τιμές περισσότερο από τις μικρότερες τιμές με την λογαριθμική κλίμακα.

Διάγραμμα 7.11 : Δεδομένα δεύτερες περιοχής με λογαριθμική κλίμακα

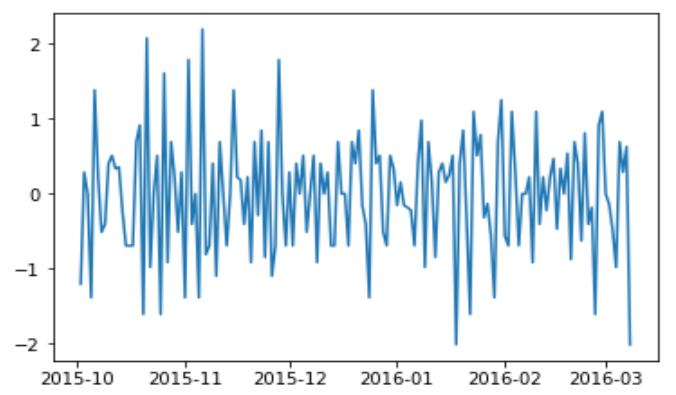
Σε αυτή την απλούστερη περίπτωση, παρατηρούμε πως και στην λογαριθμική κλίμακα έχουμε μια ανοδική τάση στα δεδομένα. Αλλά δεν είναι πολύ αντιληπτή η παρουσία του θορύβου. Έτσι, μπορούμε να χρησιμοποιήσουμε ορισμένες τεχνικές για να εκτιμήσουμε την τάση και στη συνέχεια να την αφαιρέσουμε από τη σειρά. Μια μέθοδος που θα ακολουθήσουμε είναι να πάρουμε τον κινητό μέσο όρο για τις 10 συνολικά μέρες πίσω και μπροστά από την κάθε τιμή.

Διάγραμμα 7.12 : Κινητός μέσος Όρος με παράμετρο 10 ημερών σε λογαριθμική κλίμακα



Η κόκκινη γραμμή δείχνει τον κινητό μέσο ορό. Παραιτούμε πως έχουμε μια εξομάλιση στην σειρά μας. Λαμβάνουμε υπόψη ότι από τη στιγμή που παίρνουμε το μέσο όρο των τελευταίων 10 τιμών, ο κινητός μέσος δεν ορίζεται για τις πρώτες 9 τιμές όπως φαίνεται και στο παραπάνω διάγραμμα. Κάνοντας όμως τον έλεγχο στασιμότητας με p-value > 5% βλέπουμε πως ούτε τώρα δεν είναι στάσιμη η χρονολογική σειρά. Μια άλλη τεχνική για να γίνει η χρονολογική σειρά στάσιμη είναι το να πάρουμε τις πρώτες διαφορές της σειράς , δηλαδή ορίζουμε ως πρώτες διαφορές $y_t' = y_t - y_{t-1}$

Διάγραμμα 7.13 : Δεδομένα με τις πρώτες διαφορές σε λογαριθμική κλίμακα

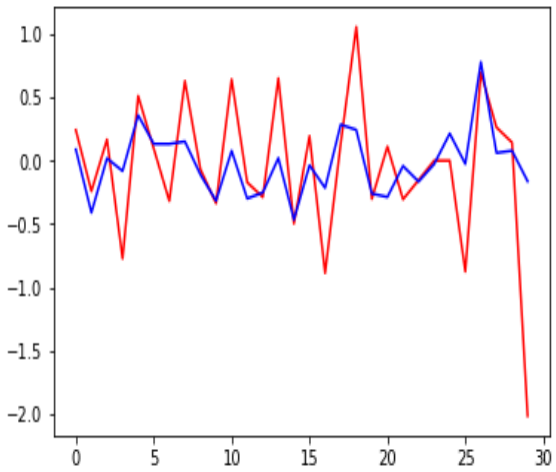


Τα δεδομένα που περιέχει το παραπάνω διάγραμμα είναι τα δεδομένα τα όποια θα έχουμε για το μοντέλο πρόβλεψης που θα κάνουμε παρακάτω διότι η παραπάνω χρονολογική σειρά είναι στάσιμη με p-valeu < 5%. Τα δεδομένα που θα χρησιμοποιήσουμε στα μοντέλα πρόβλεψης για την δεύτερη περιοχή δεν χρειάζονται μετασχηματισμό διότι ισχύουν οι κανόνες στασιμότητας , επομένως θα έχουμε τα πραγματικά δεδομένα .

Μοντέλο πρόβλεψης

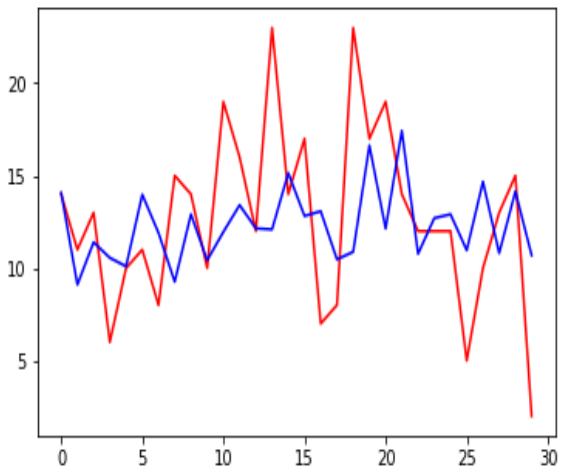
Ως μοντέλο για την πρόβλεψη των τιμών αυτόν χρησιμοποιήσαμε το AR(Autoregression) .Ορίσαμε να γίνει η πρόβλεψη για τις τελευταίες 30 ημέρες των δεόμενων μας . Στα παρακάτω διαγράμματα θα δούμε με μπλε γραμμή την πρόβλεψη του μοντέλου και με κόκκινη τα πραγματικά δεδομένα.

Διάγραμμα 7.14 : Πρόβλεψης δεύτερης περιοχής



MSE: 0.247

Διάγραμμα 7.15 : Πρόβλεψη πρώτης περιοχής

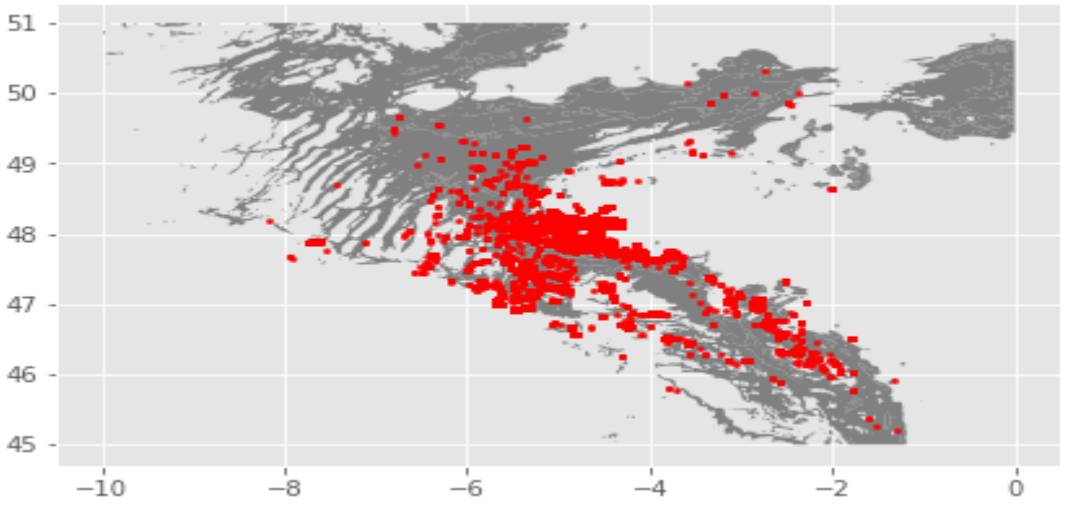


MSE: 22.323

7.5 K-Means

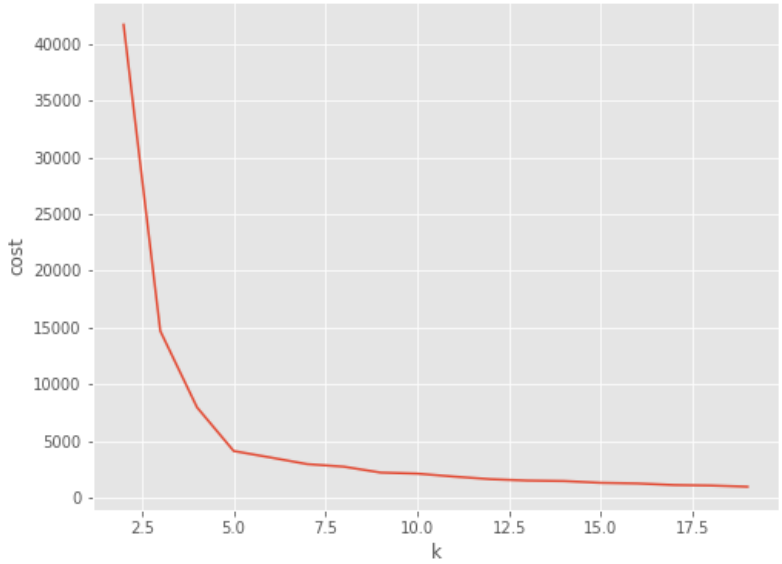
Με την αφορμή ότι δεν βρήκαμε κανένα αλιευτικό πλοίο να είναι εντος της περιοχής και να αλιεύει, όπου σαν ορισμό στο ότι ένα πλοίο αλιεύει που δώσαμε προηγουμένως στην μεθοδολογία (6^{o} κεφάλαιο, βήμα 17) .Αξιοσημείωτο θα ήταν να δούμε ποια είναι τα μέρη τα όποια πάνε τα αλιεύτηκα πλοία και εάν τηρούν τον κανονισμό δηλαδή εάν πάνε στα μέρη τα όποια επιτρέπετε η αλίευση.

Εικόνα 7.8



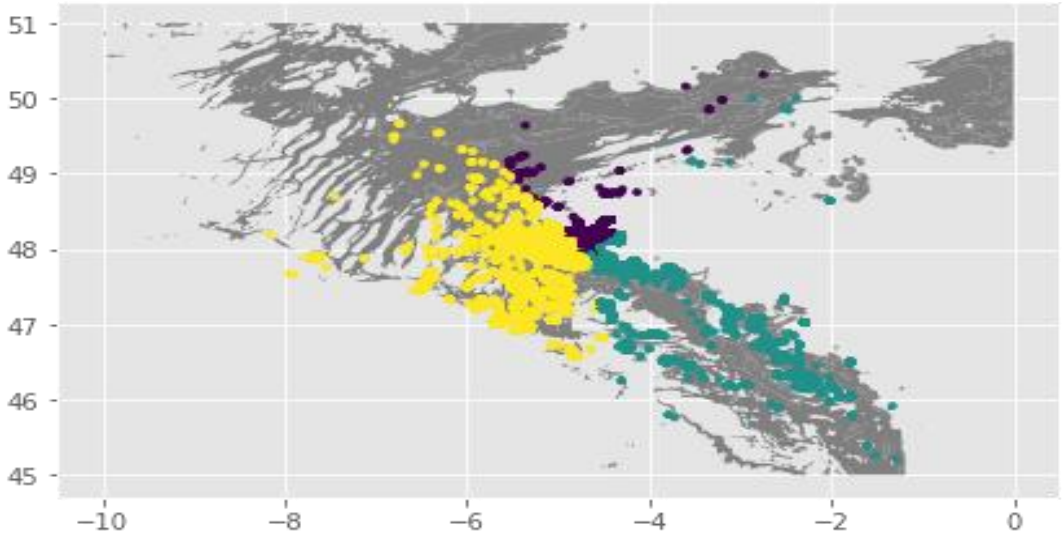
Στην εικόνα 7.8 παρατηρούμε πως τα αλιευτικά πλοία είναι εντός των επιτρεπόμενων περιοχών ώστε να αλιεύσουν. Βλέπουμε πως δεν καλύπτονται όλες οι περιοχές και επίσης ότι διαλέγουν κατά κύριο λόγο συγκεκριμένα μέρη ώστε να αλιεύσουν και ότι είναι 3 τα μέρη αυτά (ή και περισσότερα). Με αφορμή αυτό μπούμε να κάνουμε μια ομαδοποίηση αυτών των δεδομένων με τον αλγόριθμο K-means δίνοντας ως είσοδο μόνο το γεωγραφικό μήκος και πλάτος (lon,lat). χρησιμοποιήσαμε τη βιβλιοθήκη που προσφέρει το spark την MLLib για να δούμε αν διαχωρίζει τις περιοχές αυτές. Πρώτα πρέπει να βρούμε ποσά cluster θα φταίξουμε ώστε να έχουμε ένα ικανοποιητικό σφάλμα. Στην παρακάτω εικόνα βλέπουμε από 1 έως 18 cluster τα σφάλματα τα όποια θα έχουμε. Παρατηρούμε πως στα 3 cluster θα είναι ιδανικά με βάση έτσι όπως αντιλαμβανόμαστε την Εικόνα 7.8 . Σε συνδυασμό με τα συμπέρασμα που βγάζουμε από την Εικόνα 7.9 βλέπουμε πως στα 3 cluster έχουμε αρκετά καλό σφάλμα όμως από την άλλη , με την μέθοδο του αγκώνα (δηλαδή εάν δούμε το παρακάτω διάγραμμα σαν ένα χέρι, θέλουμε το σημείο εκείνο που βλέπουμε τον αγκώνα του χεριού). Αυτό το σημείο είναι στα 5 cluster, Για τον λόγο αυτό θα κάνουμε 2 ομαδοποιήσεις των ίδιων δεδομένων ώστε να συμπεράνουμε με βάσει των διαχωρισμό που θα κάνει.

Εικόνα 7.9 : Σφάλματα με βάσει τον αριθμό των clusters (1 έως 18)



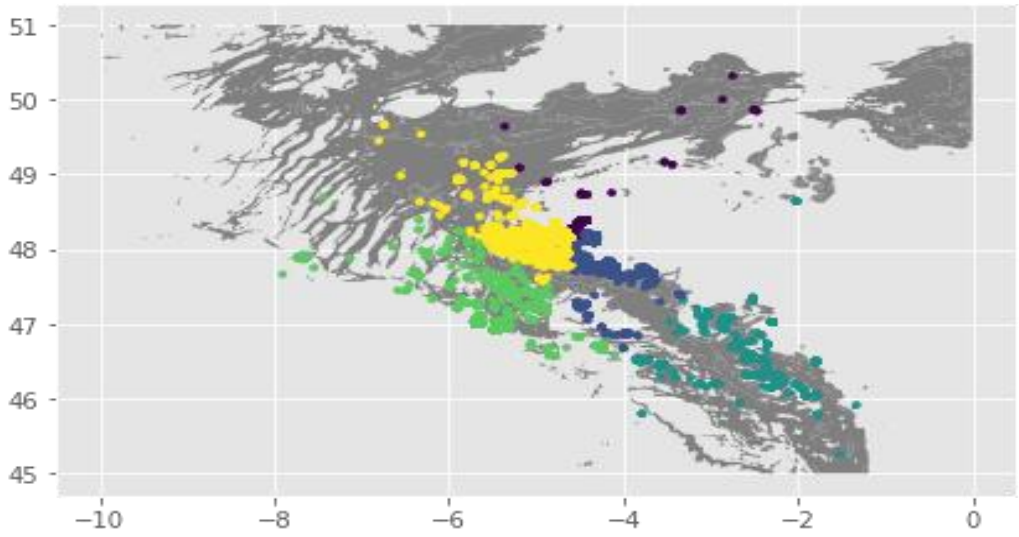
Στις παρακάτω εικόνες βλέπουμε τον διαχωρισμό που έκανε ο αλγόριθμος k-means στον χώρο τον οποίον βρίσκονται τα δεδομένα.

Εικόνα 7.10 : Ομαδοποίηση των δεδομένων σε 3 cluster



Within 3 Cluster Sum of Squared Errors = 45359.5653539

Εικόνα 7.11 : Ομαδοποίηση των δεδομένων σε 5 cluster



Within 5 Cluster Sum of Squared Errors = 4113.27894032

Παρατηρούμε πως στα 5 clusters το SSE είναι αρκετά μικρότερο σε σύγκριση με τα 3 clusters τα όποια φτιάξαμε. Αυτό είναι κάτι λογικό γιατί όσο ανεβαίνει ο αριθμός των cluster μειώνετε και το SSE διότι όλα τα σημεία θα είναι αρκετά κοντά στα κέντρα του κάθε cluster. Από ότι φαίνετε καλύτερος διαχωρισμός γίνετε στα 5 clusters διότι φαίνονται πιο ξεκάθαρα οι περιοχές τις όποιες επισκέπτονται περισσότερο.

8. Συμπεράσματα

Η ανάλυση χωροχρονικων δεδομένων είναι κάτι αρκετά δύσκολο αλλά ταυτόχρονο αρκετά ελκυστικό. Οι τεχνολογίες με τις όποιες ασχοληθήκαμε όπως για παράδειγμα το Spark για την επεξεργασία παραλλήλων δεδομένων ήταν αρκετά ικανοποιητικές και ταυτόχρονα αποτελεσματικές για τον λόγο ότι ένας υπολογιστής από μονός του θα χρειαζόταν αρκετό χρόνο παραπάνω ώστε να υλοποιήσει όλες αυτές τις πράξεις τις όποιες κάναμε στην συγκεκριμένη ερεύνα διότι αναφερόμασταν σε μεγάλη κλίμακα δεδομένων. Επιπροσθέτως το Hdfs του Hadoop το όποιο χρησιμοποιήσαμε για την παράλληλη αποθήκευση των δεδομένων, μας εξασφάλιζε την ασφάλεια των δεδομένων μας διότι είχαμε βάλει replication 1 , όποτε και ένα από τα δυο VMs που δημιουργήσαμε στον Okeano να χάλαγε, θα είχαμε πρόσβαση στα δεδομένα αυτά από τον άλλον. Ως συμπέρασμα στην ερεύνα που κάναμε είναι πως δεν εντοπίσαμε κάποιο πλοίο να είναι εντος των περιοχών αυτών και να αλιεύει ή δεν έδειχνε να αλιεύει διότι υπάρχει και η περίπτωση να έχει κλείσει το μηχάνημα το όποιο στέλνει σήμα για την τοποθεσία του καθώς και για τα αλλά δεδομένα τα οποία είχαμε και αρά δεν θα μπορούσαμε να το εντοπίσουμε. Μεγάλο ενδιαφέρον είχε και η ανάλυση της χρονολογικής σειράς με την οποία ασχοληθήκαμε διότι οι εκτιμήσεις του μοντέλου ήταν αρκετά ικανοποιητικές και για τις δυο περιοχές αυτές. Η ερεύνα χρονολογικών σειρών θα είναι χρήσιμη για τα δεδομένα τέτοιου τύπου, διότι θα γνωρίζουμε με μια μικρή απόκλιση τα πλήθος των πλοίων που εισέρχονται καθημερινώς (χειμώνα, καλοκαίρι) εντος κάποιων περιοχών(λιμένα, περιοχές αλίευσης κ.α) διότι κατά κύριο λόγο τα δρομολόγια τα όποια έχουν να κάνουν τα πλοία είναι σταθερά, όποτε εάν δούμε καμία ξαφνική πτώση (ή άνοδο) θα είναι θέμα ως προς μελέτη. Τέλος για τον λόγο ότι δεν βρήκαμε στο συγκριμένο data set (με τους δικούς μας ορισμούς ως προς την κατάσταση ενός πλοίου να αλιεύει) αλιεύτηκα πλοία να είναι εντος των δυο περιχέων αυτών και να αλιεύουν, εξετάσαμε την νόμιμη εκδοχή δηλαδή την εκδοχή ότι τα πλοία αυτά πάνε να αλιεύσουν στις επιτρεπόμενες περιοχές αλιείας. Αφού βεβαιωθήκαμε ότι αυτό ισχύει με τις ανάλογες εικόνες τις όποιες φτιάξαμε με βάση τα δεδομένα, παρατηρήσαμε πως δεν καλύπτονται όλοι οι επιτρεπόμενοι χώροι αλιείας καθώς και ότι 3 με 5 μέρη από όλοι αυτή την έκταση επισκέπτονται. Προσπαθήσαμε λοιπόν να ομαδοποιήσουμε αυτά τα δεδομένα με τον αλγόριθμο K-means ώστε να δούμε άμα διαχωρίζει τις περιοχές αυτές. Ως συμπέρασμα είχαμε ότι τις διαχώρισε αυτές τις περιοχές έχοντας μόνο ως είσοδο το γεωγραφικό μήκος και πλάτος. Αν και δεν είναι ο κατάλληλος αλγόριθμος για τέτοιου τύπου δεδομένων διότι ένα μεγάλο μειονέκτημα που έχει είναι πως τα cluster τα όποια φτιάχνει είναι κυκλικά (κάτι που δεν το θέλουμε σε όλες τις περιπτώσεις). Παρόλα αυτά ο διαχωρισμός των δεδομένων που έγινε ήταν ικανοποιητικός.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Εξόρυξη γνώσης από δεδομένα μεταφορικής αλυσίδας – Πειραματισμός με το εργαλείο Spark MLlib ΑΔΑΜΟΠΟΥΛΟΣ ΓΕΩΡΓΙΟΣ, ΜΠΣΠ-14001 Εισηγητής: ΓΙΑΝΝΗΣ ΘΕΟΔΩΡΙΔΗΣ

[2] Κατανεμημένη Λειτουργία Apache Spark με Hadoop Distributed File System (HDFS) Κλεφτάκης Σπυρίδων ,Τσότζολας Γεώργιος

[3] FPGA-Acceleration of Machine Learning in Cloud Computing, a case study using Logistic Regression Ηλίας Ν. Κορομηλάς

[4] ΑΝΑΛΥΣΗ ΣΥΝΘΕΤΩΝ ΓΕΓΟΝΟΤΩΝ ΤΟΥ ΑΥΤΟΜΑΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΑΝΑΓΝΩΡΙΣΗΣ (AIS) Μπουκουβάλας Προκόπιος