Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρικών Βιομηχανικών Διατάξεων και
Συστημάτων Αποφάσεων

Ανάλυση και Αξιοποίηση Δεδομένων Κίνησης Πλοίων για τη Δημιουργία Καινοτόμων Υπηρεσιών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΦΩΤΕΙΝΗ ΠΑΝΑΓΙΩΤΟΥ

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2023

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρικών Βιομηχανικών Διατάξεων και
Συστημάτων Αποφάσεων

Ανάλυση και Αξιοποίηση Δεδομένων Κίνησης Πλοίων για τη Δημιουργία Καινοτόμων Υπηρεσιών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΦΩΤΕΙΝΗ ΠΑΝΑΓΙΩΤΟΥ

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Οκτωβρίου 2023

.....	.....	.....
Δημήτριος Ασκούνης Καθηγητής Ε.Μ.Π.	Ιωάννης Ψαρράς Καθηγητής Ε.Μ.Π.	Χρυσόστομος Δούκας Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2023

Φωτεινή Παναγιώτου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η ναυτιλιακή βιομηχανία διαδραματίζει έναν ζωτικό ρόλο στο παγκόσμιο εμπόριο, καθιστώντας απαραίτητη την αξιοποίηση τεχνολογιών αιχμής όπως η Τεχνητή Νοημοσύνη, για τη βελτίωση της αποτελεσματικότητας και της διαδικασίας λήψης αποφάσεων. Η παρούσα εργασία εστιάζει σε μια εκτενή ανάλυση διαφορετικών συνόλων δεδομένων, την ενοποίηση τους σε γεωχωρικά ερωτήματα, και την προσομοίωση της πορείας των πλοίων. Η έρευνα ξεκινά με μια λεπτομερή εξέταση των δεδομένων, περιλαμβάνοντας ανάλυση, εναρμόνιση, αξιολόγηση και οπτικοποίηση των δεδομένων. Μέσω αυτής της διαδικασίας, εντοπίζονται γεωμετρίες με ιδιαίτερα χαρακτηριστικά, αποκαλύπτοντας σημαντικές ενδείξεις σχετικά με τη συμπεριφορά των πλοίων και τα πρότυπα κυκλοφορίας. Επιπλέον, πραγματοποιείται ένα πείραμα πρόβλεψης χρησιμοποιώντας ένα κλασσικό μονομεταβλητό υπόδειγμα χρονολογικής σειράς. Εφαρμόζονται διάφορες μέθοδοι εκτίμησης που χρησιμοποιούνται στην βιβλιογραφία των χρονολογικών σειρών και της μηχανικής μάθησης ενώ η συγκριτική αξιολόγηση τους γίνεται με βάση ορισμένα στατιστικά κριτήρια. Τα αποτελέσματα καταδεικνύουν ότι οι προτεινόμενες μέθοδοι καταφέρνουν να προβλέψουν με υψηλή ακρίβεια, παρέχοντας πολύτιμα εργαλεία για την προγνωστική ανάλυση και τον σχεδιασμό νέων υποδειγμάτων ανάλυσης σύνθετων δεδομένων.

Λέξεις κλειδιά

Ναυτιλία, AIS, Χρονολογικές Σειρές, Πρόβλεψη, Μηχανική Μάθηση, Γραμμική Παλινδρόμηση, Τυχαίο Δάσος, Κ-Κοντινότεροι Γείτονες

Abstract

The shipping industry plays a vital role in global trade, making the utilization of cutting-edge technologies like Artificial Intelligence essential for improving efficiency and decision-making processes. This research focuses on an in-depth analysis of different datasets, consolidating them into geospatial queries, and simulating ship trajectories. The research begins with a detailed examination of the data, including analysis, harmonization, evaluation, and visualization of the data. Through this process, geometries with distinctive characteristics are identified, revealing significant insights into ship behavior and traffic patterns. Additionally, a time series forecasting exercise is conducted using a classic univariate time series model. Various estimation methods commonly used in time series and machine learning literature are applied, and their comparative evaluation is based on specific statistical criteria. The results demonstrate that the proposed methods achieve high prediction accuracy, providing valuable tools for predictive analysis and the design of new modelling frameworks for more complex data.

Key words

Maritime, AIS, Time Series, Forecasting, Machine Learning, Linear Regression, Random Forest, K-Nearest Neighbors

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες προς τον επιβλέποντα μου, κ. Ασκούνη, ο οποίος με ενέπνευσε από τα πρώιμα χρόνια της φοιτητικής μου πορείας και μου έδωσε την ευκαιρία να συνεργαστώ με αξιόλογους ανθρώπους.

Επίσης, δεν μπορώ παρά να εκφράσω την ειλικρινή μου ευγνωμοσύνη στον Χρήστο Κοντζίνο για την εξαιρετική συνεργασία κατά την εκπόνηση της διπλωματικής μου εργασίας, αλλά και για την αδιάκοπη καθοδήγησή του σε κάθε μου εγχείρημα κατά τη διάρκεια της φοίτησής μου.

Τέλος θα ήθελα να ευχαριστήσω από καρδιάς την οικογένεια μου για την συνεχή τους υποστήριξη. Ιδιαίτερη αναφορά θα ήθελα να κάνω στους φίλους και συνάδελφους μου, Αθηνά, Βασιλική, Κώστα, Χριστίνα και Χριστίνα-Μαρία, καθώς έκαναν το ταξίδι αυτό μοναδικό.

Φωτεινή Παναγιώτου,

Αθήνα, 18η Οκτωβρίου 2023

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

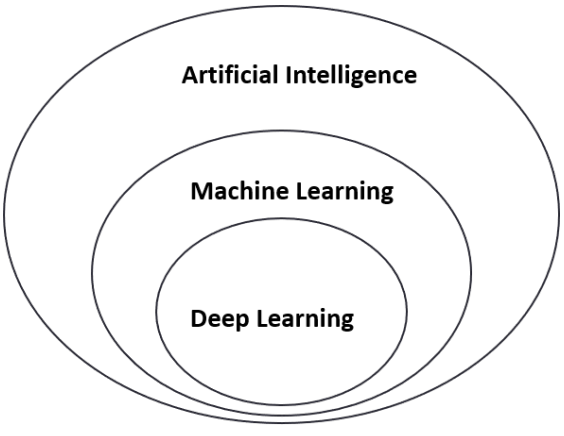
Κεφάλαιο: Εισαγωγή

1.1 Αντικείμενο και Σκοπός

Η αξιοποίηση της τεχνητής νοημοσύνης στο χώρο της ναυτιλίας προμηνύει μια νέα εποχή με σημαντικές αναμορφώσεις στην αγορά. Οι δυνατότητες που προσφέρει η ανάλυση δεδομένων και οι εφαρμογές της μηχανικής μάθησης έχουν γίνει ορατές, ενώ αναμένεται η ευρεία υιοθέτησή τους με στόχο τη βελτιστοποίηση των ναυτιλιακών υπηρεσιών και της απόδοσης τους.

Ενα από τα σημαντικότερα υποσύνολα της τεχνητής νοημοσύνης είναι η μηχανική μάθηση. Η Μηχανική μάθηση (Machine learning- ML) αποτελεί έναν εξελισσόμενο κλάδο των υπολογιστικών αλγορίθμων που σχεδιάζονται για να μιμούνται την ανθρώπινη νοημοσύνη, μαθαίνοντας από το περιβάλλον τους, και βελτιώνουν αυτόματα την απόδοσή τους καθώς αποκτούν εμπειρία. Αυτό το πεδίο βρίσκεται στο σημείο τομής της επιστήμης των υπολογιστών και της στατιστικής και είναι ένα θεμελιώδες στοιχείο της τεχνητής νοημοσύνης και της επιστήμης των δεδομένων. Οι τεχνικές που βασίζονται στη μηχανική μάθηση έχουν εφαρμοστεί με επιτυχία σε διάφορους τομείς, συμπεριλαμβανομένης της επεξεργασίας φυσικής γλώσσας, της υπολογιστικής όρασης, της αναγνώρισης φωνής, των αυτόνομων οχημάτων, της οικονομετρίας, της ιατρικής διάγνωσης και πολλών άλλων.

Η Μηχανική μάθηση έχει κρίσιμο ρόλο στην ανάλυση πληροφοριών και τη διατύπωση



Σχήμα 1.1: Σχέση μεταξύ των ειδών μάθησης

προβλέψεων από μεγάλα και σύνθετα σύνολα δεδομένων, που ονομάζονται ”big data” (μεγάλα δεδομένα). Χρησιμοποιεί μαθηματικούς αλγόριθμους για να κάνει προβλέψεις ή εκτιμήσεις σχετικά με μελλοντικά ή μη ορατά σημεία δεδομένων. Στην επιβλεπόμενη μάθηση (Supervised learning), μια συγκεκριμένη κατηγορία της μηχανικής μάθησης, αυτή η διαδικασία καθοδηγείται από ένα σαφές και δομημένο πλαίσιο. Στην επιβλεπόμενη μάθηση, ο αλγόριθμος τροφοδοτείται με ένα σύνολο δεδομένων που περιλαμβάνει τόσο τα δεδομένα εισόδου όσο και τις αντίστοιχες τιμές εξόδου (target values).Ο στόχος είναι ο αλγόριθμος να μάθει να αντιστοιχίζει τα δεδομένα εισόδου με την σωστή έξοδο, αποτυπώνοντας αποτελεσματικά τη σχέση μεταξύ των δύο. Κατά τη διάρκεια της φάσης εκπαίδευσης, ο αλγόριθμος αναλύει το σύνολο δεδομένων, εντοπίζοντας μοτίβα, συσχετίσεις και εξαρτήσεις ανάμεσα στα δεδομένα. Χρησιμοποιεί μαθηματικές τεχνικές και μεθόδους βελτιστοποίησης για να δημιουργήσει ένα μοντέλο που ενσωματώνει αυτή τη γνώση. Αυτό το μοντέλο μπορεί στη συνέχεια να χρησιμοποιηθεί για να κάνει αποδοτικές προβλέψεις ή ταξινομήσεις σε νέα δεδομένα.

Οι αλγόριθμοι της μηχανικής μάθησης περιλαμβάνουν δέντρα αποφάσεων, γραμμική παλινδρόμηση, νευρωνικά δίκτυα και πολλά άλλα. Η επιλογή του αλγορίθμου εξαρτάται από το συγκεκριμένο πρόβλημα και τα χαρακτηριστικά των δεδομένων. Σε αυτό το πλαίσιο, επιλέξαμε 7 αλγορίθμους μηχανικής μάθησης και 3 στατιστικά μοντέλα για την διεξαγωγή του πειράματος αυτής της εργασίας.

1.2 Μεθοδολογία

Αρχικά, συλλέξαμε σύνολα δεδομένων με σκοπό την εκπαίδευση του μοντέλου για πρόβλεψη χρονοσειρών για τις γεωγραφικές θέσεις πλοίων σε κατά μήκος ενός συγκεκριμένου χρονικού διαστήματος. Ως πρώτο βήμα, απαιτείται η μετατροπή του συνόλου δεδομένων της χρονοσειράς σε ένα πρόβλημα επίβλεψης. Στη συνέχεια, πραγματοποιείται ένα πείραμα πρόβλεψης χρησιμοποιώντας ένα κλασσικό μονομεταβλητό υπόδειγμα χρονολογικής σειράς. Εφαρμόζονται 10 μοντέλα, που χρησιμοποιούνται στην βιβλιογραφία των χρονολογικών σειρών και της μηχανικής μάθησης, που εκτιμούν τις μεταβλητές ενδιαφέροντος ενώ η συγκριτική αξιολόγηση τους γίνεται με βάση ορισμένα στατιστικά κριτήρια.

1.3 Οργάνωση Κειμένου

Η υπόλοιπη δομή της εργασίας οργανώνεται ως εξής: Το Κεφάλαιο 2 παρέχει συνοπτικές πληροφορίες σχετικά με το υπόβαθρο της τεχνητής νοημοσύνης στην ναυτιλία και το εννοιολογικό πλαίσιο που χρησιμοποιείται για την καθοδήγηση της ανάλυσης αυτής της εργασίας. Στο κεφάλαιο 3 περιγράφονται τα εργαλεία που χρησιμοποιήθηκαν για την διαχείριση και την ανάλυση των δεδομένων, καθώς και για την εκτέλεση των πειραμάτων. Το Κεφάλαιο 4 περιλαμβάνει την ανάλυση των συνόλων δεδομένων και η οπτικοποίησή τους. Επιπλέον, παρουσιάζεται η εμπειρική μεθοδολογία του πειράματος, το μοντέλο που χρησιμοποιείται, και οι μέθοδοι εκτίμησης. Στο Κεφάλαιο 5 αποτυπώνονται τα αποτελέσματα των κριτηρίων υπολογισμού σφαλμάτων με βάση τις μεθόδους εκτίμησης του μοντέλου, ενώ επίσης παρέχεται γραφική σύγκριση των μεθόδων εκτίμησης με βάση διάφορους χρονικούς ορίζοντες που

επιλέγονται. Το Κεφάλαιο 6 ολοκληρώνει την εργασία με συμπεράσματα και προτάσεις για μελλοντικές επεκτάσεις.

Κεφάλαιο: Βιβλιογραφική Ανασκόπηση

2.1 Τεχνητή Νοημοσύνη στην Ναυτιλία

Η ναυτιλιακή μεταφορά αποτελεί αναμφίβολα τον κυριότερο τρόπο μεταφοράς εμπορευμάτων παγκοσμίως. Είναι ένας βασικός πυλώνας του παγκόσμιου εμπορίου και της διεθνούς εφοδιαστικής αλυσίδας. Σε γενικές γραμμές, περισσότερο από το 90% των εμπορευμάτων που κυκλοφορούν παγκοσμίως μεταφέρονται μέσω των θαλάσσιων δρόμων [1].

Η εξαιρετική σημασία της ναυτιλιακής μεταφοράς προκύπτει από την ικανότητά της να μεταφέρει μεγάλες ποσότητες εμπορευμάτων με αποτελεσματικό τρόπο σε διεθνή κλίμακα. Η συνεχής ανάπτυξη του παγκόσμιου εμπορίου έχει ενισχύσει ακόμη περισσότερο τη θέση της ναυτιλιακής βιομηχανίας ως κινητήρια δύναμη της παγκόσμιας οικονομίας. Αν ανατρέξουμε σε ιστορικά δεδομένα, θα δούμε ότι η ναυτιλιακή μεταφορά έχει αναπτυχθεί και προσαρμοστεί στις ανάγκες του χρόνου. Αρχικά, ήταν κυρίως ένας τρόπος μεταφοράς ανθρώπων και εμπορευμάτων ανά τις θάλασσες. Κατά τη διάρκεια των αιώνων, εξελίχθηκε σε μια πολύπλοκη βιομηχανία με υψηλή τεχνολογία, αποτελώντας τον πρωταρχικό τρόπο μεταφοράς εμπορευμάτων παγκοσμίως.

Παρά την κρίσιμη σημασία της, η ναυτιλιακή βιομηχανία αντιμετωπίζει πολλές προκλήσεις, σε μια περίοδο σημαντικών αλλαγών. Ο ραγδαίος ρυθμός αύξησης της παγκόσμιας εμπορικής δραστηριότητας, σε συνδυασμό με τις αυξημένες απαιτήσεις για περιβαλλοντική αειφορία και ασφάλεια, απαιτεί από τη βιομηχανία αυτήν να αντιμετωπίσει προκλήσεις που προκύπτουν σε ποικίλους τομείς. Η ανάγκη για εκσυγχρονισμό στη ναυτιλία είναι πρωταρχική, καθώς η βιομηχανία πρέπει να προσαρμοστεί στις σύγχρονες απαιτήσεις και να υιοθετήσει νέες τεχνολογίες που θα διασφαλίζουν την μέγιστη αποδοτικότητα στις ναυτιλιακές δραστηριότητες.

Η εφαρμογή της τεχνητής νοημοσύνης(Artificial intelligence- AI) και των μεγάλων δεδομένων(Big Data) έχει αρχίσει να κατακτά εδάφη σε διάφορους τομείς της ναυτιλιακής βιομηχανίας και έχει προωθήσει τα όρια της παραγωγικής αποδοτικότητας των ναυτιλιακών εταιρειών. Οι μελέτες για τα μεγάλα δεδομένα και την τεχνητή νοημοσύνη στη ναυτιλιακή βιομηχανία έχουν αυξηθεί αισθητά από το 2012, γεγονός που αναδεικνύει την αυξανόμενη σημασία του θέματος [2].

Η έρευνα στην τεχνητή νοημοσύνη αρχικά στόχευε στο να αντιγράψει την ανθρώπινη διαδικασία λήψης αποφάσεων χρησιμοποιώντας μηχανές και έναν μεγάλο όγκο δεδομένων. Σήμερα, η τεχνητή νοημοσύνη είναι ικανή να επιτελέσει πράγματα που ήταν αδύνατο να

πραγματοποιηθούν πριν από μία δεκαετία. Για παράδειγμα, προηγμένα συστήματα AI συμβάλλουν στην ανάπτυξη αυτόνομων πλοίων, τα οποία μπορούν να λειτουργούν ανεξάρτητα, χωρίς ανθρώπινη αλληλεπίδραση, και το ποσοστό σφάλματος είναι χαμηλότερο από αυτό των πλοίων που λειτουργούν με ανθρώπους [3]. Έτσι, η τεχνητή νοημοσύνη μεταμορφώνει σταδιακά τον παραδοσιακό τρόπο λειτουργίας της ναυτιλιακής βιομηχανίας.

Η ναυτιλία παράγει τεράστιες ποσότητες δεδομένων ανά λεπτό, οι προοπτικές των οποίων ωστόσο παραμένουν ανεκμετάλλευτες λόγω της συμμετοχής εντυπωσιακού αριθμού ενδιαφερομένων και της πολυπλοκότητας του σύγχρονου σχεδιασμού και λειτουργίας των πλοίων. Για την διαχείριση αυτών των προκλήσεων, απαιτούνται νέες μέθοδοι τεχνητής νοημοσύνης, αλγόριθμοι ανάλυσης μεγάλων δεδομένων και μοντέλα βελτιστοποίησης, προκειμένου να διευκολυνθούν οι διαδικασίες λήψης αποφάσεων σε κάθε στάδιο της ναυτιλιακής βιομηχανίας. Οι πληροφορίες που συγκεντρώνονται από την τεχνολογία της τεχνητής νοημοσύνης προσφέρουν στις εταιρείες ναυτιλίας τη δυνατότητα απόκτησης ανταγωνιστικού πλεονεκτήματος, παρέχοντας πληροφορίες σχετικά με τον καιρό, τη ναυτιλιακή κίνηση, την κίνηση στους λιμένες κ.ά [4].

Σημαντικό ρόλο στις ναυτιλιακές της δραστηριότητες παίζουν η αποτελεσματική διαχείριση και ο συντονισμός της εφοδιαστικής αλυσίδας. Η γνώση περιοχών με διαθέσιμα πλοία για κράτηση σε συγκεκριμένο χρονικό διάστημα, βοηθά στην έγκαιρη εξασφάλιση φορτίων προς μεταφορά από αυτές τις περιοχές και στην επωφελή χρήση πλοίων με χαμηλό κόστος λόγω του έντονου ανταγωνισμού [5]. Παρά τα πλεονεκτήματα που μπορούν να προκύψουν από την γνώση του αριθμού των πλοίων σε μια συγκεκριμένη περιοχή, υπάρχει έλλειψη έρευνας που στοχεύει στην πρόβλεψη του εφοδιασμού πλοίων. Το κλειδί για την αντιμετώπιση αυτού του προβλήματος, βρίσκεται σε νέους αλγόριθμους, εργαλεία και πλατφόρμες στο φάσμα της τεχνητής νοημοσύνης, προσφέροντας στοιχεία σχετικά με τον αναμενόμενο αριθμό πλοίων σε μια περιοχή σε μια συγκεκριμένη χρονική στιγμή.

Καθώς οι τεχνολογικές εφαρμογές αυξάνονται με τους αλγόριθμους μηχανικής μάθησης, η πλειονότητα των πλοίων θα πρέπει να αφομοιώσει την τεχνητή νοημοσύνη, καθώς θα πρέπει να προσαρμοστούν σε νέα συστήματα. Αυτές οι υλοποιήσεις μηχανικής μάθησης υιοθετούνται όχι μόνο από εταιρείες ναυτιλίας, αλλά και από βιομηχανικούς και κυβερνητικούς φορείς. Ορισμένες από τις εφαρμογές που έχουν παρουσιάσει σοβαροί φορείς όπως οι IMO, BIMCO και Lloyd's Register, έχουν ήδη δείξει ότι η υιοθέτηση της τεχνητής νοημοσύνης στο μέλλον πιθανότατα θα γίνει υποχρεωτική για πολλές εταιρείες[6].

2.2 Εφαρμογές Τεχνητής Νοημοσύνης στη Ναυτιλία

Η τεχνητή νοημοσύνη αποτελεί ένα ισχυρό εργαλείο που μετασχηματίζει τον τομέα της ναυτιλίας, προσφέροντας προηγμένες λύσεις για τη βελτιστοποίηση της απόδοσης και την αύξηση της ασφάλειας στη θάλασσα. Στα επόμενα υποκεφάλαια, θα εξετάσουμε πώς η τεχνητή νοημοσύνη εφαρμόζεται σε διάφορους τομείς της ναυτιλίας, προσφέροντας καινοτόμες λύσεις που συμβάλλουν στην ανάπτυξη και τη βελτίωση του κλάδου.

2.2.1 Πρόληψη ναυτιλιακών ατυχημάτων

Η θαλάσσια μεταφορά αναδεικνύεται ως ζωτική συνιστώσα της παγκόσμιας οικονομίας, καθώς αποτελεί τον κύριο τρόπο μεταφοράς αγαθών παγκοσμίως. Εκτός από τη μείωση του κόστους μεταφοράς, οι θαλάσσιες μεταφορές συχνά επιτρέπουν την αποδοτική μεταφορά μεγάλων ποσοτήτων αγαθών, προωθώντας το διεθνές εμπόριο. Οι υδάτινες οδοί είναι απαραίτητες για τη διασφάλιση της συνεχούς ευημερίας και οικονομικής ανάπτυξης των χωρών παραθαλάσσιων περιοχών. Ωστόσο, οι δραστηριότητες που σχετίζονται με τις μεταφορές πλοίων ενέχουν υψηλούς κινδύνους για τις ανθρώπινες ζωές, την περιβαλλοντική ασφάλεια και την οικονομική βιωσιμότητα.[7]

Τα τελευταία χρόνια, παρατηρείται αύξηση στις θαλάσσιες μεταφορές λόγω της αυξανόμενης ζήτησης για εισαγωγές και εξαγωγές αγαθών παγκοσμίως. Αυτή η αύξηση τονίζει τη σημασία των θαλασσίων μεταφορών ως βασικό μέσο για την κίνηση αγαθών και την ανάπτυξη της παγκόσμιας οικονομίας. Ωστόσο, με την εξέλιξη αυτή, τα θαλάσσια ατυχήματα έχουν αρχίσει να αποτελούν έναν μη αμελητέο κίνδυνο για τα άτομα και τις κοινωνίες σε διάφορους τομείς, όπως οι ανθρώπινες και οικονομικές απώλειες, οι περιβαλλοντικές συνέπειες κ.λ.π. Η σύγκρουση πλοίων, ως ένα συχνά εμφανιζόμενο ατύχημα κατά την ναυτική κυκλοφορία (16% του συνόλου των θαλασσίων ατυχημάτων), αποτελεί έναν από τους κύριους παράγοντες [8].

Από το 2014 έως το 2021, καταγράφηκε συνολικά η απώλεια 563 ανθρώπινων ζωών σε 376 ναυτικά ατυχήματα. Η κύρια αιτία αυτών των απωλειών ήταν η σύγκρουση πλοίων. Επιπλέον, από το 2014 έως το 2021, σημειώθηκαν συνολικά 6.155 τραυματισμοί σε 5.394 ναυτικά ατυχήματα και περιστατικά. Το κύριο γεγονός που οδήγησε σε τραυματισμούς ήταν επίσης η σύγκρουση πλοίων [9].

Είναι γνωστό ότι η τρέχουσα κατάσταση στον τομέα της ναυτιλίας απαιτεί την ενσωμάτωση τεχνολογιών βασισμένων στην τεχνητή νοημοσύνη και στα νευρωνικά δίκτυα για τον σχεδιασμό και την ανάπτυξη μοντέλων που αφορούν την πρόβλεψη ατυχημάτων. [10]. Με την αξιοποίηση δεδομένων από αισθητήρες, ανιχνεύονται πρότυπα κινδύνου με σκοπό την προειδοποίηση για πιθανά προβλήματα. Επιπλέον, μέσω της ανάλυσης ιστορικών δεδομένων από προηγούμενα ατυχήματα, η τεχνητή νοημοσύνη μπορεί να προτείνει βελτιώσεις στις ναυτιλιακές διαδικασίες, με στόχο τη μείωση του κινδύνου.

Καταρχάς, με την εφαρμογή της τεχνητής νοημοσύνης, δηλαδή την χρήση μοντέλων και αλγορίθμων, μπορούμε να εξαλείψουμε τον ανθρώπινο παράγοντα, που αποτελεί τον κύριο λόγο ατυχημάτων στη θάλασσα. Όπως αναφέρουν οι Antão και Soares, ο αριθμός των συγκρούσεων που οφείλονται σε ανθρώπινο λάθος είναι υψηλός και κυμαίνεται σε ποσοστά

έως και 80%, ανάλογα με τον χρόνο και την περιοχή [11]. Επιπλέον, ο αλγόριθμος δεν αντιμετωπίζει στρες σε περίπτωση σύγκρουσης και δεν αισθάνεται άγχος. Αντίθετα, δημιουργεί μια διαδικασία ελιγμού, η οποία εκτελείται από τον αυτόματο πιλότο, αποτρέποντας έτσι τη σύγκρουση [12].

Πέραν των ατυχημάτων που οφείλονται σε ανθρώπινο παράγοντα, υπάρχουν ατυχήματα που συνδέονται με την αλλαγή του κέντρου βάρους του πλοίου. Αυτή η αλλαγή μπορεί να οδηγήσει σε αυξημένη κλίση ή ακόμα και ανατροπή. Οι Gaponenko και Malyshev προτείνουν μία πιθανή λύση για το πρόβλημα της μεταφοράς ταλαντευόμενων φορτίων σε πλοία σε συνθήκες ταραγμένων νερών. Στην έρευνα τους παρουσιάζουν την ενσωμάτωση παραμέτρων που σχετίζονται με αυτό το φαινόμενο και τη θεωρία μετατόπισης κραδασμών σε μοντέλα πρόβλεψης που βασίζονται στην τεχνητή νοημοσύνη και σε νευρωνικά δίκτυα [13].

Ωστόσο, με την αύξηση της ποσότητας, της κλίμακας και της ταχύτητας των πλοίων, τα θαλάσσια ατυχήματα εξακολουθούν να αποτελούν αυξανόμενο κίνδυνο, ιδίως μεταξύ πλοίων. Οι Chen και Huang παρουσιάζουν μια εκτενή ανασκόπηση των μεθόδων πιθανολογικής ανάλυσης κινδύνου για σύγκρουση πλοίου με πλοίο. Επισημαίνουν ότι οι τρέχουσες προσεγγίσεις συνήθως βασίζονται στην ανάλυση δεδομένων σε συγκεκριμένα χρονικά διαστήματα, πράγμα που μπορεί να οδηγήσει σε υπερβολικές ή υποτιμητικές εκτιμήσεις των αποτελεσμάτων. Ως εναλλακτική, προτείνουν τη χρήση στατιστικής ανάλυσης, ανάλυσης δέντρου σφαλμάτων (Fault tree analysis) και μοντέλων δικτύων Bayesian ως κύριες προσεγγίσεις [14].

Οι Wang και Yin εφάρμοσαν την τεχνική εξόρυξης κειμένου(text mining) και την εκμάθηση κανόνων συσχέτισης(association rule mining) σε 536 αναφορές ατυχημάτων μεταφοράς πλοίων μεταξύ των ετών 2000 και 2018, με σκοπό την αναγνώριση και εξέταση των μεταβλητών κινδύνου. Οι μεταβλητές κινδύνου μπορούν να κατηγοριοποιηθούν σε τέσσερις κατηγορίες, δηλαδή παράγοντες του πλοίου, περιβαλλοντικοί παράγοντες, ανθρώπινοι παράγοντες και παράγοντες του ατυχήματος. Συνολικά, το άρθρο προσφέρει νέες μεθοδολογικές συνεισφορές και εισηγήσεις για τη διαχείριση της ναυτικής ασφάλειας και την πρόληψη ατυχημάτων με τη χρήση αλγορίθμων text mining [15].

Ένα ακόμα παράδειγμα εφαρμογής της τεχνητής νοημοσύνης για την αποφυγή ατυχημάτων σύγκρουσης μεταξύ πλοίων παρουσιάζεται από τους Ma και Yong. Συγκεκριμένα, προτείνουν ένα μοντέλο Βαθιάς Ενισχυτικής Μάθησης (Deep Reinforcement Learning - DRL) για την αντιμετώπιση του προβλήματος συγκρούσεων αυτόνομων θαλάσσιων οχημάτων (Unmanned Surface Vehicles - USVs) σε πολύπλοκα σενάρια. Το μοντέλο DRL απέδειξε την αποτελεσματικότητά του σε πειραματικά σενάρια που περιλάμβαναν πέντε αυτόνομα θαλάσσια οχήματα [16].

2.2.2 Βελτιστοποίηση θαλάσσιων διαδρομών

Οι θαλάσσιες διαδρομές αποτελούν τη βάση για την ανάλυση των χαρακτηριστικών της ναυτιλίας και τη διασφάλιση της ασφάλειας στις υδάτινες οδούς. Ωστόσο, η ελευθερία κίνησης στη θάλασσα δυσκολεύει τη συλλογή δεδομένων σχετικά με τις θαλάσσιες διαδρομές. Για αυτό τον λόγο, προτείνεται μια μέθοδος εξαγωγής δεδομένων βασισμένη στο αυτόματο σύστημα αναγνώρισης ιστορικού των πλοίων (Automatic Identification System- AIS) [17]. Το πρόβλημα είναι ότι ο όγκος των δεδομένων που παράγει το AIS είναι τεράστιος και συχνά

τα δεδομένα εμφανίζουν ανωμαλίες. Ως αποτέλεσμα, απαιτούνται αυτοματοποιημένες διαδικασίες για να φιλτράρουν τα δεδομένα και να τα καταστήσουν χρήσιμα για αξιοποίηση. Τα δεδομένα που σχετίζονται με τη ροή της κυκλοφορίας των πλοίων μπορούν να κατηγοριοποιηθούν ως δεδομένα χρονοσειρών, και επομένως οι συμβατικές μέθοδοι Μηχανικής Μάθησης (Machine Learning- ML), Νευρωνικών Δικτύων (Neural Networks- NN) και Βαθιάς Μάθησης (Deep Learning- DL) αποτελούν τις πιο γενικές και αποτελεσματικές μεθόδους πρόβλεψης [18]. Οι Chen προτείνουν μια λύση για την αντιμετώπιση απουσιάζοντων δεδομένων και ασυνεχειών στα δεδομένα AIS με τον σχεδιασμό του μοντέλου GRU και θεσπίζουν ένα μοντέλο γραμμικής παλινδρόμησης για τις τροχιές των πλοίων με σκοπό τη βελτίωση της ακρίβειας προσαρμογής. [19]

Τα παραπάνω ενισχύονται από το τους Chondrodima και Mandalis. Στην έρευνα τους υποστηρίζεται πως οι μέθοδοι Μηχανικής Μάθησης (ML) μπορούν να αξιοποιήσουν τον μεγάλο όγκο πληροφοριών παρακολούθησης πλοίων για να διευκολύνουν την εμβάθυνση της ψηφιοποίησης στον τομέα της ναυτιλίας και να αντιμετωπίσουν το πρόβλημα πρόβλεψης διαδρομής πλοίων [20].

Προκειμένου να ανακουφιστεί αποτελεσματικά η κυκλοφοριακή συμφόρηση, να πραγματοποιηθεί δυναμική διαχείριση και αποτελεσματικός έλεγχος της ροής της ναυσιπλοΐας σε πολυσύχναστα ύδατα, ο Dong κατασκευάζει ένα ακριβές μοντέλο πρόβλεψης της ροής της ναυσιπλοΐας με χρήση της πρόβλεψης της χρονοσειράς LSTM [21].

Οι Li και συνεργάτες κάνουν μια σημαντική συνεισφορά στη βιβλιογραφία με τη δημοσίευσή τους. Το άρθρο τους παρουσιάζει μια επισκόπηση της πιο πρόσφατης βιβλιογραφίας σχετικά με την πρόβλεψη της πορείας των πλοίων, επισημαίνοντας δώδεκα προηγμένες μεθόδους μηχανικής μάθησης και βαθιάς μάθησης από το 2000 έως το 2022. Τα αποτελέσματα τους δείχνουν ότι οι παραδοσιακές μέθοδοι πρόβλεψης της πορείας των πλοίων, που βασίζονται στη μηχανική μάθηση, δεν είναι πλέον ικανές να ανταποκριθούν στην αυξανόμενη ζήτηση για ακρίβεια και πραγματικό χρόνο. Συνολικά, οι μέθοδοι πρόβλεψης της πορείας των πλοίων που βασίζονται στη βαθιά μάθηση έχουν κερδίσει αυξανόμενο ενδιαφέρον και έχουν παρουσιάσει ελπιδοφόρα αποτελέσματα [22].

2.2.3 Μείωση εκπομπών

Η ναυτιλία συνεισφέρει σημαντικά στην ατμοσφαιρική ρύπανση, κυρίως με την εκπομπή διοξειδίων του θείου, διοξειδίων του αζώτου, σωματιδίων και διοξειδίου του άνθρακα. Ωστόσο, αναμένεται ότι η συμβολή της ναυσιπλοΐας στις παγκόσμιες εκπομπές διοξειδίου του άνθρακα θα μειωθεί σημαντικά τα επόμενα χρόνια.[23]

Επιπλέον, οι περιβαλλοντικοί κανονισμοί του IMO (International Maritime Organization) θα επηρεάσουν τη ναυτιλιακή βιομηχανία, αφού απαιτούν από τις ναυτιλιακές εταιρείες να μειώσουν το περιεχόμενο θείου στα καύσιμα στο 0,5% από το 2020 και έπειτα. Η χρήση της τεχνητής νοημοσύνης αναμένεται να συμβάλει στην μείωση των εκπομπών διοξειδίου του άνθρακα που σχετίζονται με πλοία, μέσω της υιοθέτησης περιβαλλοντικά βιώσιμων λύσεων [4].

Για να μειωθεί η κατανάλωση καυσίμου και οι εκπομπές CO_2 από τα πλοία, είναι κρί-

σιμο να υπολογιστεί με ακρίβεια η επίδραση των ανέμων και των κυμάτων στην ταχύτητα των πλοίων και η αποδοτικότητα καυσίμου. Οι υπάρχουσες μέθοδοι εκτίμησης της απόδοσης των πλοίων, που βασίζονται σε πειράματα που έγιναν σε ελεγχόμενα περιβάλλοντα ή σε προσομοιώσεις βασισμένες στη φυσική, συνήθως οδηγούν σε μεγάλα σφάλματα σε σύγκριση με τις πραγματικές θαλάσσιες συνθήκες λόγω των πολύπλοκων αλληλεπιδράσεων των ανέμων, των κυμάτων και των θαλάσσιων ρευμάτων. Για να αντιμετωπίσουν αυτό το ζήτημα, ο Anan κ.ά. προτείνουν τη χρήση της τεχνολογίας ανάλυσης μεγάλων δεδομένων και της τεχνητής νοημοσύνης. Αυτές οι τεχνολογίες στοχεύουν στην βελτίωση του προσδιορισμού των καλύτερων διαδρομών με βάση τον καιρό, λαμβάνοντας υπόψη παράγοντες του πραγματικού κόσμου που επηρεάζουν την απόδοση των πλοίων, επιτρέποντας πιο ακριβή και αποτελεσματικό προγραμματισμό της διαδρομής και κατ' επέκταση την ελαχιστοποίηση της εκπομπής αερίων. [24].

2.2.4 Παρακολούθηση κατανάλωσης καυσίμων

Σε συνέχεια των παραπάνω, το ποσό των εκπομπών αερίων στην ατμόσφαιρα εξαρτάται απευθείας από την ποσότητα του καυσίμου που καταναλώνεται. Συνεπώς, έχουν εξεταστεί διάφοροι τρόποι και μέθοδοι για τη βελτιστοποίηση της απόδοσης στην κατανάλωση καυσίμου στα πλοία.

Η Μηχανική Μάθηση (ML) αποτελεί την πιο υποσχόμενη μέθοδο στην εκτίμηση και τη βελτιστοποίηση της κατανάλωσης καυσίμου. Η δημιουργία ενός μοντέλου εκτίμησης θα παρέχει πιο αποδοτικά αποτελέσματα στην παρακολούθηση και την ερμηνεία των εξωτερικών παραγόντων που σχετίζονται με την κατανάλωση καυσίμου, προσφέροντας παράλληλα στις ναυτιλιακές εταιρείες πιο οικονομικές λύσεις με την προσαρμογή νέων συστημάτων στο στόλο τους για τη διαχείριση της αποδοτικότητας. Κατά συνέπεια, οι εκπομπές μπορούν να διατηρηθούν υπό έλεγχο, και το κόστος του καυσίμου και τα συναφή έξοδα, σύμφωνα με το πλάνο διαδρομής του πλοίου, μπορούν να προγραμματιστούν εκ των προτέρων [25].

Για να μειώσουν την κατανάλωση καυσίμου στα πλοία, οι ναυτιλιακές εταιρείες επικεντρώνονται κυρίως στον προσδιορισμό της ποσότητας καυσίμου που χρησιμοποιείται κατά την πλοήγηση. Στη μελέτη τους, οι Uyanik και Arslanoglu χρησιμοποίησαν τεχνικές τεχνητής νοημοσύνης για να εκτιμήσουν την κατανάλωση καυσίμου από τα πλοία κατά τη διάρκεια των ταξιδιών [26]. Αυτό πραγματοποιήθηκε με τη χρήση της μεθόδου πολλαπλής γραμμικής παλινδρόμησης(Multiple Linear Regression). Ως αποτέλεσμα, η εκτίμηση της κατανάλωσης καυσίμου πραγματοποιήθηκε με επιτυχία, λαμβάνοντας υπόψη εσωτερικούς και εξωτερικούς παράγοντες.

Η κατανάλωση καυσίμων αποτελεί πάνω από το 25% του συνολικού λειτουργικού κόστους ενός πλοίου, και η ακριβής πρόβλεψη μπορεί να έχει σημαντική επίδραση στη βιωσιμότητα και την κερδοφορία της λειτουργίας του πλοίου. Ο Gkerekos κ.ά. παρουσιάζουν μια σύγκριση αλγορίθμων πολλαπλής παλινδρόμησης βασισμένων σε δεδομένα για την πρόβλεψη της κατανάλωσης καυσίμου της κύριας μηχανής του πλοίου. Για τον σκοπό αυτό, χρησιμοποιούνται διάφοροι αλγόριθμοι πολλαπλής παλινδρόμησης, συμπεριλαμβανομένων των Support Vector Machines, Extra Trees Regressors, Random Forest Regressors, Artificial Neural Networks και μεθόδων συνόλου (ensemble methods). Τα προκύπτοντα μοντέλα μπορούν να προβλέψουν με

ακρίβεια την κατανάλωση καυσίμου των πλοίων που πλέουν υπό διάφορες συνθήκες φόρτωσης, καιρικών συνθηκών, ταχύτητας και απόστασης πλοήγησης [27].

2.2.5 Εντοπισμός παράνομων δραστηριοτήτων

Μεταξύ των διαφορετικών μέσων μεταφοράς, η ναυσιπλοΐα αντιμετωπίζει τον υψηλότερο κίνδυνο στον τομέα της ασφάλειας. Αυτό οφείλεται στο γεγονός ότι αποτελεί τον κύριο τρόπο διευκόλυνσης του διεθνούς εμπορίου. Πέραν αυτού, μέσω των πλοίων δημιουργούνται περιθώρια για την παράνομη μεταφορά ανθρώπων, ναρκωτικών και όπλων, ενώ αποτελούν επίσης πιθανούς στόχους τρομοκρατικών επιθέσεων [28]. Για την προώθηση της ασφάλειας στη ναυτιλία είναι σημαντική η λήψη μέτρων και η συστηματική παρακολούθηση των υδάτινων οδών.

Ο Wan κ.ά [29] παρουσιάζουν τη δημιουργία ενός γράφου γνώσης (Knowledge graph) με σκοπό την ανάλυση παράνομων συμπεριφορών πλοίων και την ενίσχυση της δυνατότητας επιθεώρησης και λήψης αποφάσεων για την ασφάλειά τους. Με τη δημιουργία αυτού του γράφου γνώσης και την εφαρμογή γραφικών υπολογισμών και αναλύσεων, επιτυγχάνονται πιο αποτελεσματικές αναζητήσεις για πλοία, ανάλυση των σχέσεών τους και ανίχνευση κρυφών συνδέσεων. Το δίκτυο μπορεί να αναγνωρίσει γρήγορα παρανομίες στον τομέα της θαλάσσιας παρακολούθησης, όπως τον έλεγχο βασικών πλοίων, τα πλοία που πλέουν σε εσωτερικούς ποταμούς και ωκεανούς, τη λήξη πιστοποιητικών, την ασυνήθιστη αναφορά λιμένων, την ασυνήθιστη τροχιά και τον κίνδυνο απάτης.

Ο Schwehr και συνεργάτες [30] πραγματοποιούν εκτενή συζήτηση για την χρήση βελτιωμένων δυνατοτήτων για την σχεδόν συνεχή, σε πραγματικό χρόνο, παρακολούθηση της θέσης της ναυτιλιακής κυκλοφορίας, χρησιμοποιώντας το Αυτόματο Σύστημα Ταυτοποίησης των Πλοίων (AIS), το οποίο θα διευκολύνει την αναγνώριση των πλοίων που ευθύνονται για την παράνομη εκπομπή πετρελαιοειδών αποβλήτων.

Ο Abouheaf κ.ά [31] παρουσιάζουν έναν μηχανισμό λήψης αποφάσεων χρησιμοποιώντας την ενισχυτική μάθηση (Reinforcement Learning - RL) για την αντιμετώπιση των αρνητικών επιπτώσεων των περιστατικών παράνομης, μη καταγεγραμμένης και ανεξέλεγκτης αλιείας (Illegal, unreported and unregulated fishing- IUU) κατά μήκος των ακτών του Καναδά. Η ενσωμάτωση αλγορίθμων μηχανικής μάθησης ενισχύει την ασφάλεια στις θαλάσσιες περιοχές και συμβάλλει στην καταπολέμηση της παράνομης αλιείας, η οποία έχει σοβαρές περιβαλλοντικές και οικονομικές επιπτώσεις λόγω της εξάντλησης των φυσικών πόρων.

2.3 Σχετική Βιβλιογραφία

Στο κεφάλαιο αυτό, πραγματοποιήθηκε εκτενής έρευνα και αναζήτηση σε παρόμοια βιβλιογραφία και προηγούμενες επιστημονικές εργασίες που ασχολούνται με την πρόβλεψη της πορείας πλοίων. Στόχος της αναζήτησης ήταν η κατανόηση των τεχνικών τεχνητής νοημοσύνης που έχουν χρησιμοποιηθεί σε παρόμοιες μελέτες και πειράματα. Κατά τη διάρκεια αυτής της ερευνητικής διαδικασίας, πραγματοποιήθηκε ανάλυση των προσεγγίσεων που έχουν χρησιμοποιηθεί για την πρόβλεψη της πορείας πλοίων, καθώς και των διάφορων τεχνικών τεχνητής νοημοσύνης που έχουν εφαρμοστεί. Η ανασκόπηση της σχετικής βιβλιογραφίας αποτέλεσε τη βάση για την περαιτέρω ανάπτυξη της εργασίας μας και την επιλογή των κατάλληλων μεθοδολογιών πρόβλεψης που θα εφαρμοστούν στη συγκεκριμένη μελέτη. Οι εμπειρίες και τα αποτελέσματα προηγούμενων ερευνών αποτέλεσαν πολύτιμη συνεισφορά στην κατανόηση και την προετοιμασία του ερευνητικού μας έργου.

Στην μελέτη των Uyanik κ.ά., δημιουργήθηκαν διάφορα μοντέλα πρόβλεψης, όπως Multiple Linear Regression, Ridge Regression, LASSO Regression, Support Vector Regression, Boosting Algorithms, Tree-Based Algorithms. Η ακρίβεια των μοντέλων καθορίζεται από μια τεχνική ονόματι *K-fold cross-validation*. Για την αξιολόγηση της ορθότητας των μεθόδων εκτίμησης και την ανάλυση της συσχέτισης μεταξύ των μεταβλητών, χρησιμοποιούνται κριτήρια υπολογισμού σφαλμάτων, όπως η τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος, το μέσο απόλυτο σφάλμα και ο συντελεστής προσδιορισμού (Coefficient of determination) [25]. Η *K-fold cross-validation* δεν λαμβάνει υπόψη την διαδοχική σειρά δεδομένων σε χρονοσειρές. Σε αυτήν την διαδικασία, το σύνολο δεδομένων διαιρείται σε K περίπου ίσου μεγέθους υποσύνολα. Κάθε υποσύνολο μπορεί να περιέχει δεδομένα από διάφορες χρονικές περιόδους, το οποίο μπορεί να μην αντικατοπτρίζει πραγματικές καταστάσεις. Αντίθετα, στην παρούσα εργασία επιλέγεται η τεχνική *walk-forward validation* η οποία λαμβάνει ρητά υπόψη την χρονική σειρά των σημείων δεδομένων, καθιστώντας την κατάλληλη για πειράματα πρόβλεψης χρονοσειρών, όπου τα παλαιά δεδομένα χρησιμοποιούνται για την πρόβλεψη των μελλοντικών δεδομένων.

Το επόμενο άρθρο που εξετάσθηκε, στοχεύει στην συστηματική ανάλυση της απόδοσης των μεθόδων πρόβλεψης της τροχιάς των πλοίων και στην πραγματοποίηση πειραματικών δοκιμών για να αποκαλυφθεί η καταλληλότητά τους σε διάφορα σενάρια. Χρησιμοποιούνται πέντε μέθοδοι μηχανικής μάθησης και επτά μέθοδοι βαθιάς μάθησης για να υλοποιηθεί πρόβλεψη της τροχιάς και να συγκριθεί η απόδοσή τους στον πραγματικό κόσμο. Συλλέγονται τρία σύνολα δεδομένων AIS, δυο από αυτά αφορούν την θαλάσσια κυκλοφορία και ένα την περιοχή ενός λιμένα. Εξετάζονται έξι δείκτες αξιολόγησης και κρίνεται η αποτελεσματικότητα των δώδεκα μεθόδων πρόβλεψης τροχιάς, καθώς και η καταλληλότητα κάθε μεθόδου σε διάφορα σενάρια ναυτικής κυκλοφορίας [22]. Αυτή η μελέτη έχει ως βασικό στόχο την σύγκριση της απόδοσης της μηχανικής μάθησης και της βαθιάς μάθησης στην πρόβλεψη της τροχιάς των πλοίων. Αντίθετα, στη δική μας εργασία, ο κύριος στόχος μας είναι η δημιουργία ενός προτύπου αξιολόγησης (benchmark) για τις μεθόδους πρόβλεψης της τροχιάς των πλοίων στον τομέα της ναυτιλίας, εστιάζοντας αποκλειστικά στην παλινδρόμηση (regression). Η παλινδρόμηση αναφέρεται σε μια κατηγορία εποπτευόμενων προβλημάτων μηχανικής μάθησης με σκοπό την πρόβλεψη μιας συνεχούς μεταβλητής με βάση ένα ή περισ-

σότερα χαρακτηριστικά εισόδου.

Στην έρευνα των Zhang και Bin, προτείνεται μια λύση βασισμένη σε δεδομένα AIS για τη γενική πρόβλεψη του προορισμού των πλοίων στις υπηρεσίες παγκόσμιας ναυτιλίας. Δημιουργείται ένα μοντέλο βασισμένο σε τυχαίο δάσος (Random Forest) για την εκτίμηση της ομοιότητας μεταξύ της τροχιάς ταξιδιού του πλοίου και των ιστορικών τροχιών. Προβλέπεται ως προορισμός του πλοίου ο προορισμός της ιστορικής τροχιάς που έχει τη μεγαλύτερη ομοιότητα με την τρέχουσα τροχιά. Αυτή η μέθοδος διαφέρει από προηγούμενες μελέτες που χρησιμοποιούν ναυτιλιακές εγγραφές ως είσοδο για την πρόβλεψη του προορισμού [32]. Στο πλαίσιο αυτών των μελετών κινείται και το δικό μας πείραμα, καθώς χρησιμοποιούμε δεδομένα AIS ως είσοδο και αξιοποιούμε χρονοσειρές για να προβλέψουμε τον προορισμό. Στόχος μας δεν είναι μόνο η χρήση του Random Forest, αλλά και η εξερεύνηση της καταλληλότητας και της απόδοσης άλλων μοντέλων που δεν δημιουργούν πολλαπλά δέντρα αποφάσεων, αλλά χρησιμοποιούν μια μοναδική δομή δέντρου αποφάσεων. Συγκεκριμένα, ανάμεσα στα μοντέλα που εξετάζουμε περιλαμβάνονται τα εξής: DecisionTreeRegressor και ExtraTreesRegressor.

Ο Bodunov κ.α μέσω της μελέτης τους παρέχουν πρόβλεψη για (i) έναν προορισμό και (ii) την ώρα άφιξης των πλοίων με χρήση γεωχωρικών δεδομένων στο πλαίσιο της ναυτιλίας. Η προσέγγισή τους περιλαμβάνει τη χρήση συνόλων μάθησης βασισμένων στις μεθόδους Random Forest, Gradient Boosting Decision Trees (GBDT), XGBoost Trees και Extremely Randomized Trees (ERT) προκειμένου να παράξουν πρόβλεψη για έναν προορισμό [33]. Σε αυτή τη μελέτη, εξετάζεται ο τελικός προορισμός ενός πλοίου καθώς ταξιδεύει και γίνεται σύγκριση μεταξύ τριών διαφορετικών μεθόδων. Αυτό το πρόβλημα, παρότι μπορεί αρχικά να φαίνεται ότι είναι κατάλληλο για μοντέλα παλινδρόμησης, καθώς μπορούν να λάβουν υπόψη ενδιάμεσες καταστάσεις που προκύπτουν από τα εισερχόμενα γεωχωρικά δεδομένα, στην πραγματικότητα αντιμετωπίζεται ως ένα πρόβλημα ταξινόμησης. Εδώ οι συγγραφείς εστιάζουν στην πρόβλεψη της υψηλού επιπέδου συμπεριφοράς (high-level behavior) του πλοίου. Δεδομένου ότι τα ταξίδια των πλοίων σπάνια ακολουθούν τυχαίες τροχιές, η πρόβλεψη του προορισμού αντιμετωπίζεται ως πρόβλημα ταξινόμησης αντί να προβλεφθούν μελλοντικές συντεταγμένες για γεωγραφικό μήκος και πλάτος χρησιμοποιώντας παλινδρόμηση. Στο δικό μας πείραμα, αντίθετα, στόχος μας είναι να προβλέψουμε μια συνεχή αριθμητική τιμή που αφορά την επόμενη θέση του πλοίου, και γι' αυτό χρησιμοποιούμε μοντέλα παλινδρόμησης.

Κατόπιν της ανάλυσης της σχετικής βιβλιογραφίας, είναι εμφανές ότι η λύση για την επιλογή της καλύτερης μεθόδου πρόβλεψης της τροχιάς πλοίων προς έναν προορισμό, προσαρμοσμένη σε συγκεκριμένες ναυτιλιακές συνθήκες, παραμένει ανεξερεύνητη. Είναι αναγκαίο να αναπτυχθεί μια συστηματική ανάλυση και αξιολόγηση προηγμένων μεθόδων πρόβλεψης βασισμένων σε πραγματικά σύνολα δεδομένων, προκειμένου να δημιουργηθεί ένα αξιόπιστο σημείο αναφοράς για την επιλογή της βέλτιστης μεθόδου πρόβλεψης της τροχιάς πλοίων βασισμένης σε δεδομένα AIS. Αυτό το σημείο αναφοράς θα συμβάλει στη βελτίωση της ναυτιλιακής ασφάλειας και της αποδοτικότητας των λιμενικών και ναυτιλιακών δραστηριοτήτων, ενισχύοντας παράλληλα την αειφορία της ναυτιλιακής βιομηχανίας.

2.4 Εννοιολογικό πλαίσιο

Οι Hyndman και Athanasopoulos [34] καθορίζουν πέντε γενικά βήματα που πρέπει να ακολουθηθούν για τη διεξαγωγή προβλέψεων:

1. Ορισμός του προβλήματος
2. Συγκέντρωση πληροφοριών
3. Προκαταρκτική εξερευνητική ανάλυση
4. Επιλογή και εφαρμογή μοντέλων
5. Αξιολόγηση ενός προτύπου πρόβλεψης

Τα παραπάνω βήματα έχουν ακολουθηθεί και στην παρούσα εργασία, διασφαλίζοντας έτσι ότι η διαδικασία πρόβλεψης πραγματοποιήθηκε σύμφωνα με ένα δομημένο και αξιόπιστο πλαίσιο.

Ανεξάρτητα από τον τύπο των διαθέσιμων δεδομένων, οι πληροφορίες πρέπει να αναλυθούν εκτενώς προτού επιλεγεί η μέθοδος πρόβλεψης. Αυτό είναι απαραίτητο διότι αρκετά μοντέλα διαφέρουν στην καταλληλότητά τους. Για την παρούσα εργασία επιλέξαμε να χρησιμοποιήσουμε σύνολα δεδομένων AIS, λόγω της ευρείας χρήσης τους αλλά και της δυνατότητάς τους να παρέχουν λεπτομερείς χωροχρονικές πληροφορίες, οι οποίες είναι κατάλληλες για τον εντοπισμό συμπεριφορών πλοίων και μοτίβων κυκλοφορίας.

Τα δεδομένα AIS(Automatic Identification System) αποτελούν μια πηγή πληροφοριών που μπορεί να έχει υψηλή αξία για τον προγραμματισμό στη ναυτιλιακή βιομηχανία. Αυτά τα δεδομένα μεταδίδονται από πομποδέκτες AIS που είναι εγκατεστημένοι σε πλοία και μεταδίδουν αυτόματα πληροφορίες, όπως η θέση τους, η ταχύτητά τους και η ναυτική τους κατάσταση. Η ακρίβεια της θέσης των δεδομένων AIS είναι γενικά 0.0001° [35]. Ως αποτέλεσμα, τα δεδομένα AIS καταγράφουν λεπτομερώς την πορεία των πλοίων και έχουν εκτενή εφαρμογή στην έρευνα της ναυτιλιακής κυκλοφορίας.

Στόχος της εργασίας δεν είναι μόνο η εύρεση της βέλτιστης μεθόδου πρόβλεψης της τροχιάς πλοίων, αλλά και ο καθορισμός ενός δικτύου θαλάσσιων μεταφορών που ακολουθεί συγκεκριμένα μοτίβα. Στην ανάλυση των γεωχωρικών δεδομένων σκοπός είναι να εντοπίσουμε γεωμετρίες όπως στάσιμες περιοχές, κόμβοι λιμένων αλλά και συμπεριφορές, όπως η στάση και στάθμευση πλοίων.

Σε σύγκριση με τα ανοιχτά ύδατα, οι λιμένες αντιμετωπίζουν μεγαλύτερη πίεση λόγω χωρικών περιορισμών και της υψηλής πυκνότητας κίνησης των πλοίων. Η αυξημένη ζήτηση για την επίβλεψη των δραστηριοτήτων των πλοίων σε αυτούς τους περιορισμένους χώρους είναι επείγουσα και καθοριστική για την ασφάλεια και την αποτελεσματική λειτουργία των λιμένων [36]. Για να ανταποκριθούμε σε αυτές τις ανάγκες, μελετούμε δύο σύνολα δεδομένων που αφορούν τις θέσεις των λιμανιών, με στόχο την εξαγωγή συμπερασμάτων για την κίνηση σε αυτούς τους κόμβους.

Οι στάσιμες περιοχές είναι επίσης ιδιαίτερα σημαντικές και χρήζουν ανάλυσης, επειδή συχνά αντιστοιχούν στη θέση στρατηγικών κόμβων στο δίκτυο ναυτιλίας. Οι στάσιμες περιοχές που βρίσκονται μακριά από την ακτογραμμή και δεν ταιριάζουν με τη θέση των υπεράκτιων πλατφορμών, μπορεί να είναι είτε περιοχές αγκυροβολίας, είτε περιοχές αλιείας είτε περιοχές αναμονής [37]. Η ανάλυσή τους είναι σημαντική για την αποτελεσματική διαχείριση και ασφάλεια των θαλάσσιων περιοχών.

Τέλος, ο εντοπισμός και η ανάλυση των στάσεων ελλιμενισμού και αγκυροβόλησης των πλοίων μπορεί να χρησιμοποιηθεί για την κατανόηση της δραστηριότητάς τους. Αυτό μπορεί να συμβάλει στην ανάλυση της ροής κυκλοφορίας των πλοίων και στην παρακολούθηση των δραστηριοτήτων εισόδου και εξόδου από το λιμάνι.

Κεφάλαιο: Εργαλεία

3.1 Εισαγωγή

Για την αποτελεσματική διαχείριση και ανάλυση του μεγάλου όγκου δεδομένων που συλλέχθηκαν σε αυτήν την ερευνητική εργασία, απαιτήθηκε η χρήση μια σειράς από ισχυρά εργαλεία και τεχνολογίες. Η βάση δεδομένων PostgreSQL αποδείχθηκε αξιόπιστη και επεκτάσιμη, παρέχοντας τη δυνατότητα οργάνωσης και διαχείρισης των δεδομένων. Επιπλέον, η επέκταση PostGIS ενίσχυσε και επέκτεινε την επεξεργασία γεωγραφικών δεδομένων, προσφέροντας την δυνατότητα για χωρική αναπαράσταση και ανάλυση. Το pgAdmin μας επέτρεψε να διαχειριστούμε αποτελεσματικά τη βάση δεδομένων, ενώ το Jupyter Notebook και η πλατφόρμα Kaggle παρείχαν τη δυνατότητα προεπεξεργασίας των δεδομένων καθώς επίσης τη δυνατότητα εξέτασης και δοκιμής των προσεγγίσεών μας. Με τη συνδυασμένη χρήση αυτών των εργαλείων και τεχνολογιών, καταφέραμε να ανταποκριθούμε στους στόχους της έρευνάς μας και να ανακαλύψουμε σημαντικά ευρήματα από τα δεδομένα μας.

Σχήμα 3.1: Διάγραμμα Ροής Τεχνολογιών

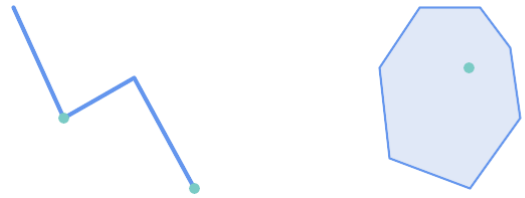
3.2 PostgreSQL

Για την διαχείριση του μεγάλου όγκου δεδομένων και την αρχική ανάλυση τους χρησιμοποιήθηκε η βάση δεδομένων PostgreSQL. Η PostgreSQL είναι ένα αξιόπιστο και επεκτάσιμο σύστημα διαχείρισης βάσεων δεδομένων (Database Management System - DBMS) ανοικτού κώδικα που ακολουθεί το μοντέλο της σχεσιακής βάσης δεδομένων και επεκτείνει την προγραμματιστική γλώσσα SQL μέσω της εκτέλεσης πολύπλοκων ερωτημάτων [38]. Η PostgreSQL επιτρέπει στα δεδομένα να οργανωθούν, με περιορισμούς μοναδικότητας, σε πίνακες που ομαδοποιούνται σε προκαθορισμένα σχήματα και συνδέονται μεταξύ τους μέσω ξένων κλειδιών.

Η PostgreSQL ξεχωρίζει μέσα σε άλλες σχεσιακές βάσεις για την εξαιρετική υποστήριξη γεωγραφικών δεδομένων μέσω της επέκτασης PostGIS. Το PostGIS είναι μια επέκταση ανοιχτού κώδικα για το σύστημα διαχείρισης βάσεων δεδομένων PostgreSQL που προσφέρει τη δυνατότητα επεξεργασίας γεωγραφικών και χωρικών δεδομένων. Η συνδυασμένη χρήση της PostgreSQL και της επέκτασης PostGIS αποτελεί ισχυρό μέσο για την διαχείριση γεωγραφικών δεδομένων και την χωρική τους αναπαράσταση.

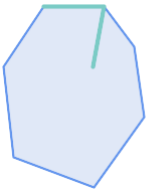
Το PostGIS προσθέτει νέους τύπους δεδομένων, όπως σημεία, γραμμές, πολύγωνα αλλά και ένα ευρύ φάσμα συναρτήσεων που επιτρέπουν την ανάλυση χωρικών δεδομένων. Με αυτές τις συναρτήσεις είναι δυνατή η επεξεργασία και η ανάλυση γεωγραφικών χαρακτηριστικών, επιτρέποντας διάφορες λειτουργίες, όπως υπολογισμός αποστάσεων, ανίχνευση γεωμετρικών επικαλύψεων, υπολογισμός χωρικών συναθροίσεων κ.ά. [39]. Για παράδειγμα, η συνάρτηση `ST_Distance` καλείται για τον υπολογισμό της απόστασης μεταξύ δύο σημείων. Η συνάρτηση `ST_Buffer` μπορεί να χρησιμοποιηθεί για την δημιουργία ενός πολυγώνου που αντιπροσωπεύει μια περιοχή γύρω από ένα σημείο, ενώ η συνάρτηση `ST_Union` εφαρμόζεται για την ένωση γεωγραφικών αντικειμένων, όπως πολύγωνα ή γραμμές, σε ένα μεγαλύτερο γεωγραφικό αντικείμενο.

Επιπλέον η συνάρτηση `ST_Contains` αξιοποιείται για την ανίχνευση σχέσεων συνύπαρξης μεταξύ διαφορετικών χωρικών χαρακτηριστικών. Προσδιορίζει εάν ένα συγκεκριμένο σημείο/αντικείμενο (π.χ. ένα σημείο, μια γραμμή ή ένα πολύγωνο) βρίσκεται μέσα σε μια συγκεκριμένη περιοχή (π.χ. ένα πολύγωνο) ή εάν μια γεωγραφική περιοχή περικλείει πλήρως μια άλλη. Η συνάρτηση επιστρέφει μια boolean τιμή, υποδεικνύοντας εάν το πρώτο αντικείμενο περιέχει το δεύτερο αντικείμενο. Οι συναρτήσεις αυτές μπορούν να εφαρμοστούν μεταξύ γεωμετριών διαφόρων τύπων και να συνδυαστούν με χρονικούς τελεστές, επιτρέποντας την δημιουργία σύνθετων χωροχρονικών ερωτημάτων.



(a) LINESTRING/ MULTIPOINT

(b) POLYGON/ POINT



(a) POLYGON/ LINESTRING

(b) POLYGON/ POLYGON

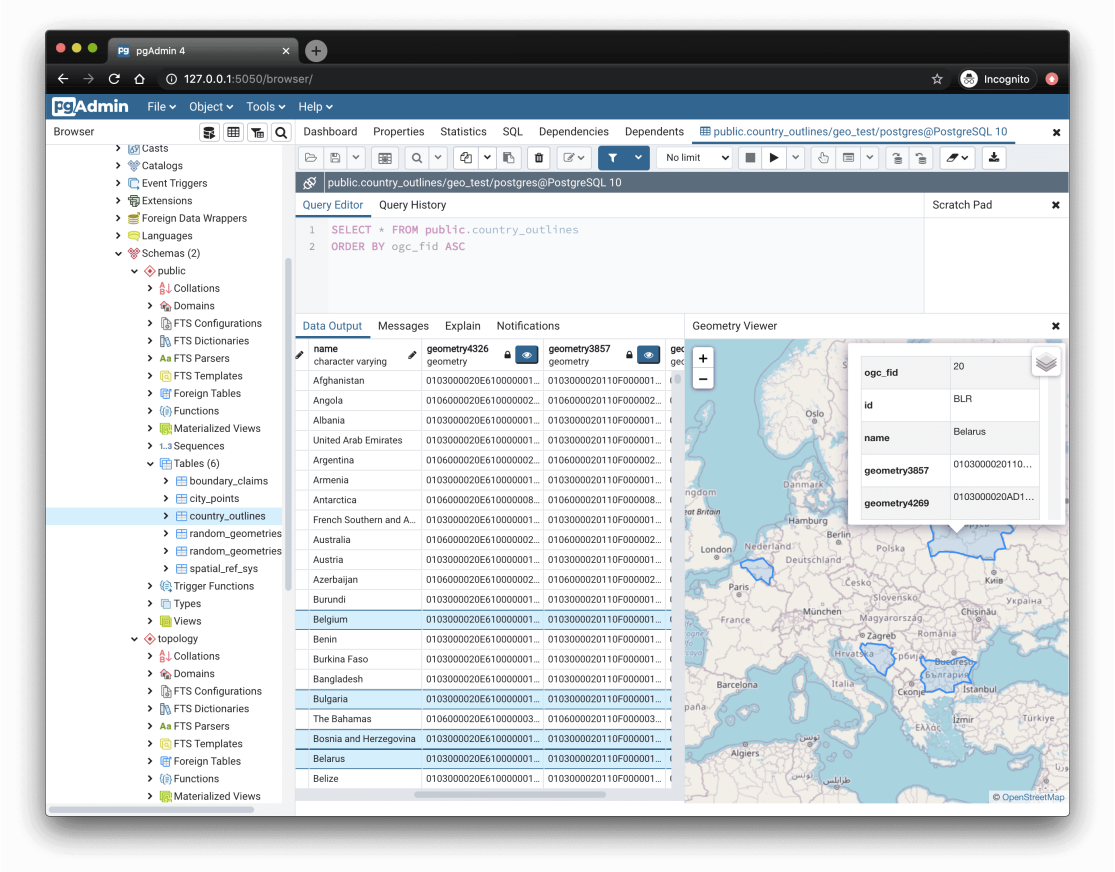
Σχήμα 3.3: Συνάρτηση ST_Contains

3.3 pgAdmin

Το pgAdmin είναι ένα δημοφιλές, ανοιχτού κώδικα εργαλείο διαχείρισης βάσεων δεδομένων PostgreSQL. Σκοπός του είναι να παρέχει μια χρήσιμη γραφική διεπαφή (GUI) που επιτρέπει στους χρήστες να δημιουργούν, να τροποποιούν και να διαχειρίζονται βάσεις δεδομένων PostgreSQL αντί να χρησιμοποιούν εντολές SQL απευθείας στο τερματικό [40].

Οι βασικές λειτουργίες που παρέχονται από το pgAdmin περιλαμβάνουν τη δημιουργία και διαγραφή πινάκων, την εισαγωγή και επεξεργασία δεδομένων, την εκτέλεση ερωτημάτων SQL, την παρακολούθηση της απόδοσης της βάσης δεδομένων και πολλές άλλες λειτουργίες που απαιτούνται για την αποτελεσματική διαχείριση μιας βάσης δεδομένων PostgreSQL.

Μέσω του pgAdmin είναι δυνατή η απεικόνιση γεωγραφικών δεδομένων, χρησιμοποιώντας εργαλεία και επεκτάσεις που υποστηρίζουν γεωγραφικές λειτουργίες, όπως η επέκταση PostGIS. Για την αξιοποίηση του PostGIS μέσω του pgAdmin, πρέπει πρώτα να εγκατασταθεί η επέκταση αυτή στη βάση δεδομένων και στη συνέχεια, να δημιουργηθούν γεωγραφικοί πίνακες. Με αυτόν τον τρόπο θα είναι δυνατή η προβολή, επεξεργασία και η εκτέλεση ερωτημάτων που σχετίζονται με τα γεωγραφικά δεδομένα των πινάκων.



Σχήμα 3.4: Περιβάλλον pgAdmin

3.4 Jupyter Notebook

Το Jupyter Notebook αποδείχθηκε ένα πολύτιμο μέσο στη μεταφορά των αρχείων δεδομένων στην PostgreSQL και στην αρχική επεξεργασία τους. Πρόκειται για μια δημοφιλή ανοιχτού κώδικα πλατφόρμα που χρησιμοποιείται για διαδραστική ανάλυση δεδομένων και προγραμματισμό. Ως εφαρμογή, το Jupyter Notebook επιτρέπει τη δημιουργία και την κοινοποίηση εγγράφων που περιλαμβάνουν κώδικα, σχόλια, σχήματα, εξισώσεις και γραφικές αναπαραστάσεις [41].

Χάρη στην ευελιξία του Jupyter Notebook, μπορέσαμε να δημιουργήσουμε και να εκτελέσουμε προσαρμοσμένο κώδικα που απλοποίησε την διαχείριση και τη μεταφορά των δεδομένων στη βάση δεδομένων PostgreSQL. Εξαιτίας του όγκου των δεδομένων, κρίθηκε αναγκαία η αρχική τους επεξεργασία μέσω του διαδραστικού αυτού περιβάλλοντος. Για την ανάλυση των δεδομένων χρησιμοποιήθηκαν συναρτήσεις από τη βιβλιοθήκη `pandas` της προγραμματιστικής γλώσσας Python. Επίσης, αξιοποιήθηκαν οι βιβλιοθήκες `psycopg2` και `sqlalchemy` για την τελική μετάπτωση και αποθήκευση των δεδομένων στην νεοσύστατη βάση.

Η εγκατάσταση του Jupyter Notebook προϋποθέτει την εγκατάσταση της Python. Έπειτα ανοίγοντας το cmd από το παράθυρο διαλόγου, πληκτρολογούμε τις ακόλουθες εντολές:

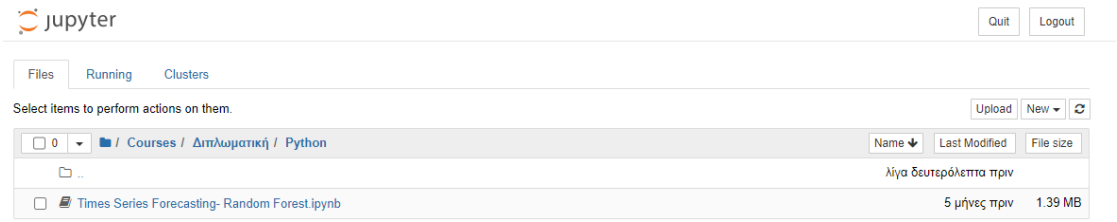
```
pip install jupyter
```

Listing 3.1: Εγκατάσταση Jupyter Notebook

```
jupyter notebook
```

Listing 3.2: Εκκίνηση Jupyter Notebook

Η παραπάνω εντολή θα ανοίξει μια νέα καρτέλα στο πρόγραμμα περιήγησης με τη διεπαφή Jupyter Notebook, όπως φαίνεται παρακάτω:



Σχήμα 3.5: Περιβάλλον Jupyter Notebook

Για το άνοιγμα της διεπαφής μπορεί να χρησιμοποιηθεί και η παρακάτω εντολή:

```
$ py -m notebook
```

Listing 3.3: Άνοιγμα Jupyter Notebook

Για την σύνδεση της Python με την εφαρμογή pgAdmin και την πρόσβαση στην βάση δεδομένων χρησιμοποιούνται οι βιβλιοθήκες *psycopg2* και *sqlalchemy* της Python. Η SQLAlchemy είναι μια υψηλού επιπέδου βιβλιοθήκη που επιτρέπει τη δημιουργία, ανάγνωση, ενημέρωση και διαγραφή εγγραφών στη βάση δεδομένων μέσω κώδικα Python. Αντιθέτως, η βιβλιοθήκη psycopg2 είναι χαμηλού επιπέδου και παρέχει μια άμεση διεπαφή για αλληλεπίδραση με βάσεις δεδομένων PostgreSQL, καθώς επιτρέπει την εκτέλεση ερωτημάτων SQL και την ανάκτηση των αποτελεσμάτων. Δημιουργώντας ένα νέο notebook, εισάγουμε τον παρακάτω κώδικα:

Σχήμα 3.6: Σύνδεση με τη βάση δεδομένων

Για τη δημιουργία, λοιπόν, μιας νέας σύνδεσης με τη βάση δεδομένων γίνεται κλήση της συνάρτησης *connect*. Στη συνέχεια πρέπει να δοθούν τα κατάλληλα διαπιστευτήρια, όπως το όνομα της βάσης, του χρήστη και ο κωδικός, ώστε η σύνδεση να κριθεί επιτυχής. Στη συνέχεια, μέσω του jupyter, μπορούν να γραφούν ερωτήματα(queries) με τον ίδιο τρόπο που αποτυπώνονται στο περιβάλλον pgAdmin:

Σχήμα 3.7: Queries μέσω Jupyter Notebook

Αναλυτικά για τον παραπάνω κώδικα:

- **cursor_obj=con.cursor()** : Αυτή η γραμμή δημιουργεί έναν δρομέα με το όνομα *cursor_obj* που σχετίζεται με τη σύνδεση της βάσης δεδομένων. Ο δρομέας χρησιμοποιείται για την εκτέλεση ερωτημάτων SQL και την ανάκτηση αποτελεσμάτων από τη βάση δεδομένων.
- **cursor_obj.execute** : Σε αυτό το σημείο του κώδικα εκτελείται ένα ερώτημα SQL χρησιμοποιώντας τη μέθοδο *execute*.
- **result=cursor_obj.fetchall()** : Σε αυτή τη γραμμή χρησιμοποιείται η μέθοδος *fetchall* η οποία ανακτά όλες τις σειρές που επιστρέφονται από το εκτελεσμένο ερώτημα.

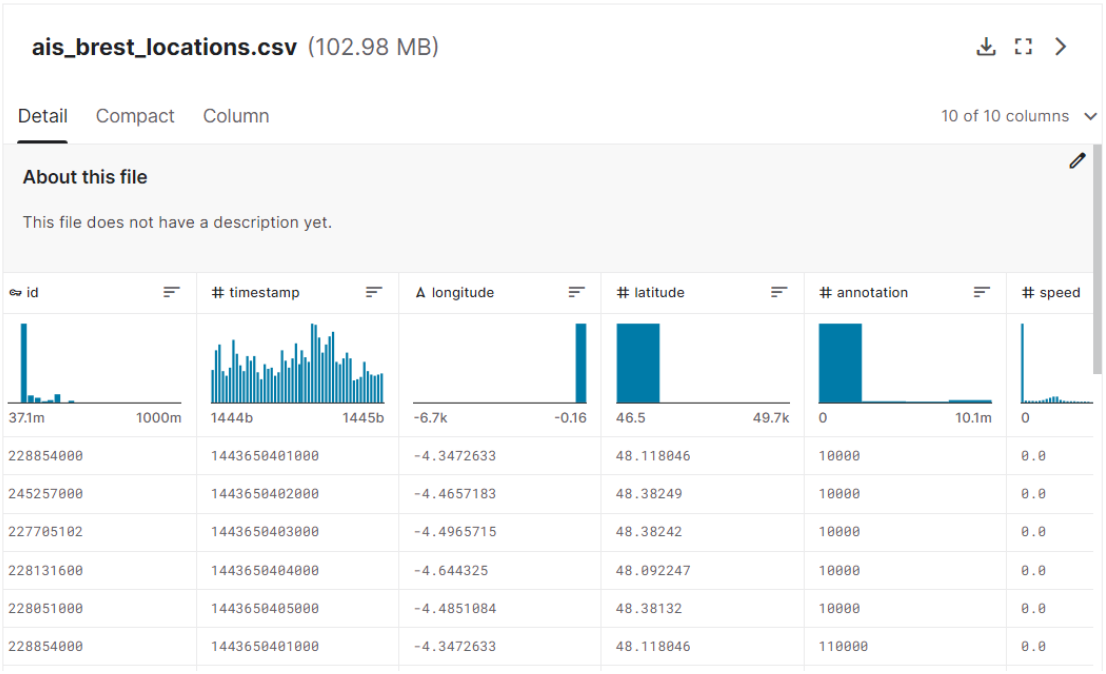
3.5 Kaggle

Η διαδικτυακή πλατφόρμα Kaggle αποδείχθηκε επίσης απαραίτητο εργαλείο στη διαδικασία διεξαγωγής των πειραμάτων για την παρούσα εργασία. Το Kaggle είναι μια διαδραστική πλατφόρμα που προσφέρει στους χρήστες τη δυνατότητα να γράφουν και να εκτελούν κώδικα σε διάφορες γλώσσες προγραμματισμού, όπως Python ή R, απευθείας μέσω του προγράμματος περιήγησής τους. Τα κυριότερα πλεονεκτήματα της χρήσης του Kaggle περιλαμβάνουν την εξοικονόμηση χρόνου και πόρων, την πρόσβαση σε ισχυρούς υπολογιστικούς πόρους, και τη δυνατότητα αποθήκευσης και κοινής χρήσης του κώδικα.

Ένα από τα βασικά προβλήματα που παρουσιάστηκαν κατά την εκτέλεση των πειραμάτων, ήταν η ανάγκη για μεγάλη υπολογιστική ισχύ η οποία υπερέβαινε τις δυνατότητες του

τοπικού υπολογιστή. Το κόλλημα αυτό επιλύθηκε μέσω των Kaggle Kernels, που χρησιμοποιούν τους διαθέσιμους πόρους CPU και GPU της πλατφόρμας και προσφέρουν παράλληλη επεξεργασία. Αυτό επέτρεψε την εκτέλεση έως και 12 σεναρίων κώδικα ταυτόχρονα, με χρονικό όριο 12 ωρών για κάθε εκτέλεση.

Μία από τις δυνατότητες του Kaggle είναι επίσης η μεταφόρτωση και κοινοποίηση συνόλων δεδομένων (datasets). Στην Ενότητα *Datasets* ο κάθε χρήστης μπορεί να ανεβάσει τα δεδομένα του και να προσθέσει περιγραφή, τίτλο, και τις ετικέτες που περιγράφουν το dataset. Για κάθε dataset δίνεται η δυνατότητα της γραφικής απεικόνισης του εύρους των τιμών που λαμβάνει κάθε στήλη, όπως φαίνεται παρακάτω:



Σχήμα 3.8: Περιβάλλον Kaggle - Dataset

Κεφάλαιο: Παρουσίαση Δεδομένων & Πειραμάτων

4.1 Περιγραφή Πειράματος

Σκοπός της εμπειρικής μελέτης της παρούσας εργασίας είναι η επεξεργασία, οπτικοποίηση, και ανάλυση δεδομένων πλοήγησης. Συγκεκριμένα, εξετάζονται ιστορικά δεδομένα θέσεων AIS πλοίων που προέρχονται από διάφορα ανοικτά σύνολα δεδομένων στον ναυτικό τομέα. Ο στόχος είναι να κατανοήσουμε τη συμπεριφορά των πλοίων και να διερευνήσουμε τη θαλάσσια κυκλοφορία που προκύπτει από αυτά τα δεδομένα.

Για τη διεξαγωγή αυτής της μελέτης, κρίθηκε απαραίτητη η συλλογή AIS δεδομένων από διάφορες πηγές. Επίσης, έγινε αναζήτηση ανοικτών συνόλων δεδομένων που μπορούν να χρησιμοποιηθούν για τον εμπλουτισμό των πληροφοριών που περιέχονται στα δεδομένα παρακολούθησης πλοίων, με σκοπό τον συνδυασμό τους σε σύνθετα ερωτήματα (queries).

Για την σωστή διαχείριση των δεδομένων στήθηκε μια βάση δεδομένων PostgreSQL που επέτρεψε την σύνδεση πινάκων και την εκτέλεση πολύπλοκων ερωτημάτων και αναλύσεων. Μέσα από αυτή τη διαδικασία επιδιώκεται η εύρεση νέων πληροφοριών και η ανακάλυψη συσχετίσεων που πιθανώς να μην ήταν εμφανείς, όπως δηλαδή ο εντοπισμός χωροχρονικών σχέσεων μεταξύ πλοίων και περιοχών ενδιαφέροντος, όταν π.χ. βρίσκονται σε κοντινή απόσταση μεταξύ τους. Με αυτό τον τρόπο επιτυγχάνεται η βαθύτερη κατανόηση των φαινομένων και, τελικά, η εξαγωγή συμπερασμάτων και ερμηνειών που αφορούν τις ναυτιλιακές δραστηριότητες πλοίων.

4.2 Εξερεύνηση Δεδομένων

Το Σύστημα Αυτόματης Ταυτοποίησης (Automatic Identification System- AIS) είναι ένα από τα ηλεκτρονικά συστήματα που επιτρέπουν στα πλοία να μεταδίδουν τη θέση τους και πληροφορίες σχετικά με την ταυτότητά τους μέσω ραδιοεπικοινωνίας [42]. Τα κινηματικά μηνύματα AIS που παράγονται παρέχουν πληροφορίες που αφορούν την ταυτότητα του πλοίου, την ταχύτητα, την κατεύθυνση, το χρονικό στίγμα, την κατηγορία πλοίου κοκ. Τα δεδομένα αυτά συλλέγονται από σταθμούς που βρίσκονται στην ακτή, από δορυφόρους αλλά και από παρακείμενα πλοία.

Τα δεδομένα AIS έχουν αποδειχθεί χρήσιμα για την παρακολούθηση των πλοίων και την εξαγωγή πολύτιμων πληροφοριών σχετικά με τη συμπεριφορά των πλοίων, τα λειτουργικά πρότυπα και τα στατιστικά στοιχεία απόδοσης [43]. Χρησιμοποιούνται για τον εντοπισμό και την καταγραφή των θαλάσσιων κινήσεων, την εξασφάλιση της ασφάλειας της ναυσιπλοΐας, τον προγραμματισμό διαδρομών, καθώς και για την πραγματοποίηση αναλύσεων που αφορούν τις μεταβολές και τις τάσεις στη ναυτική κυκλοφορία.

Το πρώτο σύνολο δεδομένων (Dataset 1) που αξιοποιήθηκε παρουσιάζεται σε μορφή CSV και είναι δημόσια διαθέσιμο [44]. Το Dataset 1 επικεντρώνεται στην καταγραφή 1.048.576 δυναμικών μηνυμάτων πλοίων καθ' όλη τη διάρκεια του έτους 2009, εντός της εκτεταμένης περιοχής της Βρετάνης. Τα δεδομένα αυτά αντικατοπτρίζουν τις ναυτιλιακές κινήσεις που εκδηλώθηκαν σε αυτήν την συγκεκριμένη περιοχή, ενσωματώνοντας παραμέτρους που είναι απαραίτητες για την παρακολούθηση της πορείας των πλοίων. Συγκεκριμένα, το Dataset 1 περιέχει τα ακόλουθα χαρακτηριστικά: Longitude(Γεωγραφικό μήκος), Latitude(Γεωγραφικού πλάτος), Heading (κατεύθυνση), Speed (ταχύτητα), Course over Ground (πορεία πάνω από το έδαφος), Rate of Turn (ρυθμός στροφής) και shipCode.

Τα δεδομένα έχουν οργανωθεί σε έναν πίνακα της νεοσύστατης βάσης, με όνομα *ship_data*, ο οποίος λαμβάνει την ακόλουθη μορφή:

MMSI	Time	Longitude	Latitude	Heading	Speed	COG	ROT
227635210	2009-02-05 09:22:53	-4.4161982	48.2842025	511.0	0.00	274.70	-128.00

Πίνακας 4.1: Dataset 1

Το δεύτερο σύνολο δεδομένων (Dataset 2) παρουσιάζεται επίσης σε μορφή CSV και είναι δημόσια διαθέσιμο [45]. Το Dataset 2 και περιλαμβάνει 19.035.630 εγγραφές AIS κινηματικών μηνυμάτων από πλοία που εκτελούσαν κινήσεις σε κοντινή εμβέλεια από το λιμάνι της Βρέστης στη Γαλλία, κατά την περίοδο μεταξύ Οκτωβρίου 2015 και Μαρτίου 2016. Από το σύνολο των διαθέσιμων πεδίων του Dataset 2 επιλέγονται τα παρακάτω βασικά χαρακτηριστικά για την εισαγωγή τους στην βάση δεδομένων, στον πίνακα *brest2015*:

ship_id	time	longitude	latitude	sog	heading	turn	cog
228854000	1443650401000	-4.3472633	48.118046	0.0	0.0	0.0	260.8

Πίνακας 4.2: Dataset 2

Ένα άλλο παρόμοιο, ανοικτό σύνολο δεδομένων (Dataset 3) [46] αφορά τις κινήσεις πλοίων στην ευρύτερη περιοχή του Ατλαντικού ωκεανού κατά τη διάρκεια της περιόδου από Οκτώβριο 2015 έως Μάρτιο 2016. Το Dataset 3 αποτελείται από 18.648.556 εγγραφές και διαμορφώνεται ως εξής στον πίνακα *eu2015* της βάσης:

Τα Dataset 2 και Dataset 3 επεκτείνονται σε κοινά σημεία, εστιάζοντας στο λιμάνι της Βρέστης, καλύπτοντας την ίδια χρονική περίοδο. Η δυνατότητα σύνδεσης και συνδυασμού αυτών των δύο dataset παρέχει τη δυνατότητα για βαθύτερη ανάλυση και ερμηνεία της ναυ-

ship_id	rot	sog	cog	heading	longitude	latitude	time
228186700	0.0	0.0	38.0	137	-4.5128317	48.370667	1446742140

Πίνακας 4.3: Dataset 3

τιλιακής δραστηριότητας και συμπεριφοράς των πλοίων σε αυτήν τη συγκεκριμένη περιοχή. Επιπλέον, επιτρέπει την ανίχνευση πιθανών μοτίβων και τάσεων που ενδέχεται να επηρεάζουν την ναυτιλιακή δραστηριότητα στην εν λόγω περιοχή κατά τη διάρκεια της συγκεκριμένης περιόδου.

Ένας επιπρόσθετος στόχος είναι ο εντοπισμός συνδέσμων μεταξύ ετερογενών συνόλων δεδομένων. Οι πληροφορίες που περιέχονται σε ένα ενιαίο σύνολο δεδομένων, όπως οι θέσεις των πλοίων, μπορούν να εμπλουτιστούν αν συσχετιστούν με διαφορετικά σύνολα δεδομένων που αναφέρονται π.χ. σε καιρικές συνθήκες ή θέσεις λιμανιών. Για την επίτευξη των παραπάνω, χρησιμοποιήθηκαν δύο επιπλέον datasets που αφορούν τις συντεταγμένες λιμανιών.

Το Dataset 4 είναι δημόσια διαθέσιμο [47], διατίθεται σε μορφή CSV και περιλαμβάνει αναλυτικές πληροφορίες για λιμάνια σε διάφορες περιοχές της Γαλλίας. Ο αντίστοιχος πίνακας ονομάζεται *`French_Ports`* στη σχεσιακή βάση λαμβάνει την ακόλουθη μορφή:

portNumber	portName	regNumber	regName	cntrCode	cntrName	latitude	longitude
37170	Ambes	36450	FRANCE WEST COAST	F	France	45.0100	0.3200

Πίνακας 4.4: Dataset 4

Τέλος, το Dataset 5 με την ονομασία "Ports of Brittany", διατίθεται σε μορφή shapefile [46]. Αυτό το σύνολο δεδομένων περιλαμβάνει τα ονόματα και τις συντεταγμένες των λιμανιών στην περιοχή της Βρετάνης. Η χρήση της μορφής shapefile επιτρέπει την αναπαράσταση γεωγραφικών δεδομένων, παρέχοντας πληροφορίες για την τοποθεσία και τη γεωμετρία των λιμανιών σε αυτήν την περιοχή. Το Dataset 5 εισαγάγεται στον πίνακα *`Brittany_Ports`* και επιλέγονται τα εξής χαρακτηριστικά:

gml_id	por_id	libelle_po	insee_comm	por_x	por_y
port.1	1	Le Vivier-sur-Mer	35361	297025	2408370

Πίνακας 4.5: Dataset 5

Μέσω των ανωτέρω datasets, προστίθενται στον συνδυασμό δεδομένων οι γεωγραφικές πληροφορίες για τις τοποθεσίες λιμανιών στη Γαλλία, προσφέροντας έτσι την δυνατότητα για περαιτέρω ανάλυση και κατανόηση των συνδέσεων ανάμεσα στις ναυτιλιακές κινήσεις και τη θέση των λιμανιών σε αυτήν την περιοχή.

Στον παρακάτω πίνακα παρουσιάζεται η χαρτογράφηση των οντοτήτων της βάσης μεταξύ των datasets 1, 2, 3 που αναφέρονται σε μηνύματα AIS:

Οντότητα	Πίνακες		
	ship_data	brest2015	eu2015
Κωδ. πλοίου	MMSI	ship_id	ship_id
Timestamp	Time	time	time
Γεωγ. Πλάτος	Latitude	latitude	latitude
Γεωγ. Μήκος	Longitude	longitude	longitude
Ταχύτητα	Speed	sog	sog
Κατεύθυνση	Heading	heading	heading
Πορεία	COG	cog	cog
Ρυθμός Στροφής	ROT	turn	rot

Πίνακας 4.6: Χαρτογράφηση πινάκων με δεδομένα AIS

Όπως φαίνεται παραπάνω, για την συμπλήρωση κάθε πίνακα επιλέγονται χαρακτηριστικά που είναι κοινά σε όλα τα σύνολα δεδομένων. Αυτή η πρακτική επιτρέπει την δημιουργία μιας ενιαίας δομής για τα δεδομένα και διευκολύνει τη σύγκριση και ανάλυση τους. Επιπλέον, δίνει τη δυνατότητα σύνδεσης των πινάκων με βάση αυτά τα κοινά χαρακτηριστικά, επιτρέποντας την ανίχνευση συσχετίσεων, όπως οι τομές ή οι ενώσεις των δεδομένων. Για παράδειγμα, χρησιμοποιώντας το κοινό πεδίο "Κωδ. πλοίου" σε όλους τους πίνακες, μπορεί να γίνει σύνδεση μεταξύ τους για κοινούς κωδικούς πλοίων.

Στον παρακάτω πίνακα παρουσιάζεται η χαρτογράφηση των οντοτήτων της βάσης μεταξύ των datasets 4, 5 που αναφέρονται σε λιμάνια της Γαλλίας:

Οντότητα	Πίνακες	
	French_Ports	Brittany_Ports
Κωδ. Λιμανιού	portNumber	gml_id
Όνομα Λιμανιού	portName	libelle_po
Αριθμός Περιοχής	regionNumber	insee_comm
Όνομα Περιοχής	regionName	-
Κωδ. Χώρας	countryCode	-
Όνομα Χώρας	countryName	-
Γεωγ. Πλάτος	latitude	por_x
Γεωγ. Μήκος	longitude	por_y

Πίνακας 4.7: Χαρτογράφηση πινάκων με δεδομένα Λιμένων

Για την αποτελεσματική διαχείριση των δεδομένων και την σύζευξή τους, τα παραπάνω datasets εισάγονται στη βάση δεδομένων PostgreSQL και σε διαφορετικούς πίνακες. Στο περιβάλλον pgAdmin4, πριν την εισαγωγή των δεδομένων, πρέπει να δημιουργηθεί μια ενιαία βάση και οι επιμέρους πίνακες στους οποίους θα εισαχθεί το εκάστοτε dataset. Με δεξί κλικ πάνω στο πεδίο "Databases" επιλέγεται "Create" → Database. Ορίζεται το όνομα "Ais" καθώς χρησιμοποιούνται AIS data. Από την ενότητα "Schemas" της νέας βάσης επιλέγεται το πεδίο "Tables" ώστε να δημιουργηθούν οι απαραίτητοι πίνακες και με δεξί κλικ→Create→ Table. Ενδεικτικά για τον πίνακα `Ship_data` ορίζονται το όνομα (Name) και οι στήλες (Columns). Στη συνέχεια καθορίζονται τα παρακάτω ονόματα στηλών και τα αντίστοιχα Data Types,

δηλαδή οι τύποι δεδομένων που ορίζουν το είδος τους:

ColumnName	DataType
MMSI_Number	integer NOT NULL
Time	timestamp without time zone NOT NULL
Longitude	numeric NOT NULL
Latitude	numeric NOT NULL
Heading	numeric
Speed	numeric
COG	numeric
ROT	numeric
ShipCode	numeric
Point	geometry

Στον πίνακα δημιουργούνται πρωτεύοντα κλειδιά (primary keys) εξασφαλίζοντας έτσι ότι οι τιμές σε μία ή περισσότερες στήλες είναι μοναδικές και μη κενές. Γι' αυτό το λόγο πριν την εισαγωγή του εκάστοτε dataset στη βάση γίνεται μια προεργασία με τη βοήθεια του Jupyter Notebook. Πιο συγκεκριμένα, με τη βοήθεια της συνάρτησης *drop_duplicates* της βιβλιοθήκης pandas, εξαλείφονται τα διπλότυπα που τυχόν υπάρχουν στο dataset και θα μπορούσαν να παραβιάσουν την ακεραιότητα των primary keys. Επιπλέον, λόγω του μεγάλου όγκου δεδομένων, επιλέγεται ένα δείγμα περίπου ενός εκατομμυρίου εγγραφών από κάθε σύνολο δεδομένων AIS για την εισαγωγή του στη βάση δεδομένων PostgreSQL. Αυτή η προσέγγιση θεωρείται απαραίτητη για τη διασφάλιση της αποδοτικότητας των εκτελούμενων σεναρίων και εργασιών.

Η εισαγωγή των δεδομένων γίνεται είτε μέσω του περιβάλλοντος pgAdmin4 είτε με τη βοήθεια του Jupyter Notebook. Ενδεικτικά ο αντίστοιχος κώδικας:

```
import pandas as pd
data=pd.read_csv(r'D:/Courses/Thesis/AIS/AIS Breast 2015/ais_brest_locations.csv')
print(data.shape[0])
data.drop_duplicates(subset=['id','speed','longitude','latitude','timestamp'], keep='last', inplace=True)
print(data.shape[0])
data.to_csv('ais_brest_locations_clean.csv', index=False)
```

Listing 4.1: Καθαρισμός δεδομένων

```
import psycopg2
import pandas as pd
conn = psycopg2.connect("host=localhost dbname=ais user=postgres password=****")
cur = conn.cursor()
cur.execute("""
    CREATE TABLE BREST2015(
```

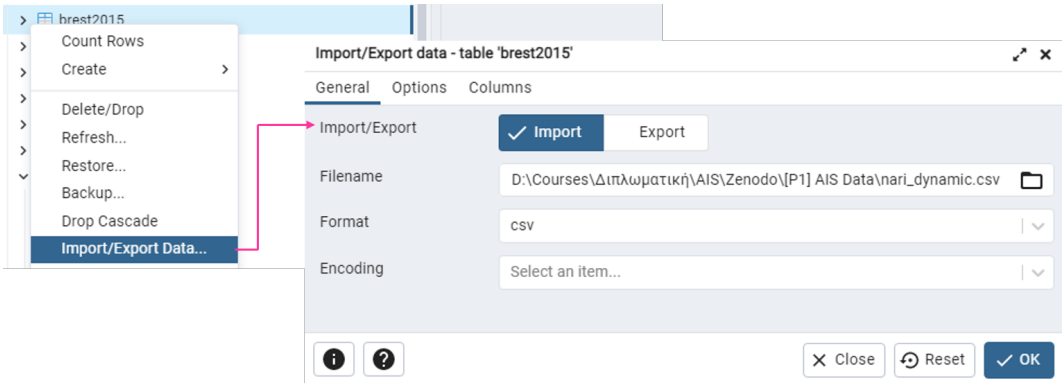
```
        ship_id bigint,timestamp bigint,longitude numeric,latitude numeric,
        annotation numeric,SOG numeric,Heading numeric,turn numeric,
        COG numeric )
 """)

with open('D:/Courses/Thesis/AIS/python_scripts/ais_brest_locations_clean.csv
    ', 'r') as f:
     next(f) # Skip the header row.
     cur.copy_from(f, 'brest2015', sep=',')
conn.commit()
```

Listing 4.2: Εισαγωγή Δεδομένων μέσω Jupyter Notebook

Το παραπάνω κομμάτι κώδικα Python πραγματοποιεί την ανάγνωση ενός αρχείου CSV και την εισαγωγή των δεδομένων του σε έναν πίνακα με το όνομα *brest2015* στην βάση δεδομένων *ais*. Το επιπλέον χαρακτηριστικό του κώδικα είναι η δυνατότητα δημιουργίας του πίνακα, με καθορισμένα πεδία (στήλες) και τους τύπους δεδομένων τους. Αυτό επιτρέπει την ευέλικτη διαχείριση των δεδομένων της βάσης δεδομένων, είτε μέσω της εφαρμογής pgAdmin4, είτε μέσω του περιβάλλοντος Jupyter Notebook.

Παρακάτω εμφανίζονται σε στιγμιότυπο τα βήματα για την εισαγωγή ενός αρχείου στο περιβάλλον pgAdmin4:



Σχήμα 4.1: Εισαγωγή Δεδομένων μέσω pgAdmin4

Στη συνέχεια ακολουθεί η ενσωμάτωση του PostGIS στη βάση δεδομένων PostgreSQL χρησιμοποιώντας το pgAdmin, ακολουθώντας τα εξής βήματα:

```
CREATE EXTENSION IF NOT EXISTS postgis;
```

Listing 4.3: Προσθήκη επέκταση Postgis

```
SELECT postgis_version();
```

Listing 4.4: Επαλήθευση εγκατάστασης Postgis

4.3 Εναρμόνιση Δεδομένων

Ο Εναρμονισμός Δεδομένων (Data Harmonization) είναι η διαδικασία ενοποίησης δεδομένων από διάφορες πηγές, συχνά με διαφορετικές μορφές, και η μετατροπή τους σε μια ενιαία μορφή. Η ετερογένεια και η αποκέντρωση των πηγών δεδομένων επηρεάζουν την οπτικοποίηση των δεδομένων και την πρόβλεψη, αλλοιώνοντας έτσι τα αναλυτικά αποτελέσματα [48]. Ο εναρμονισμός δεδομένων αποτρέπει αυτές τις ανομοιότητες και βοηθά στον εντοπισμό και τη διόρθωση σφαλμάτων, ακραίων ή κενών τιμών στο σύνολο δεδομένων, βελτιώνοντας έτσι τη συνολική ακρίβεια και ποιότητα των δεδομένων. Επίσης, εξασφαλίζεται η συνοχή των δεδομένων και η προσαρμογή στις κοινές μονάδες μέτρησης. Ο εναρμονισμός δεδομένων είναι αναγκαίος για να διασφαλιστεί ότι τα δεδομένα είναι ευανάγνωστα, συνεκτικά και χρήσιμα για ανάλυση και λήψη αποφάσεων.

Κατά τον προκαταρκτικό καθαρισμό των δεδομένων πριν την εισαγωγή τους στη βάση, απορρίφθηκαν εγγραφές που εμφάνιζαν διπλότυπα με βάση τα κλειδιά του εκάστοτε πίνακα. Έπειτα, πραγματοποιήθηκε έλεγχος για τυχόν κενές τιμές σε πεδία που αποτελούν πρωτεύοντα κλειδιά. Σε περίπτωση που αυτό επαληθευόταν, οι εγγραφές αυτές επίσης απορρίπτονταν. Αυτά τα βήματα αποτελούν σημαντικό μέρος του προκαταρκτικού καθαρισμού των δεδομένων. Έτσι διασφαλίζεται η ακεραιότητα και η ποιότητα των δεδομένων που αποθηκεύονται στη βάση.

Στη συνέχεια, παρατηρείται πως σε κάθε dataset η στήλη που αφορά τη χρονική στιγμή μετάδοσης του μηνύματος, δηλαδή το timestamp των σημάτων, παρουσιάζεται σε μορφή UTC epoch, με το χρόνο εκφρασμένο σε μιλιδευτερόλεπτα (millisecond). Αυτή η αποτύπωση δυσκολεύει την ανάλυση των δεδομένων, καθώς δεν είναι εύκολο να αναγνωριστεί η συγκεκριμένη ημερομηνία και ώρα κάθε εγγραφής. Για ευκολότερη επεξεργασία των χρονικών δεδομένων, πραγματοποιούμε τη μετατροπή του χρόνου στην εξής μορφή: 'YYYY-MM-DD HH:MI:SS'. Αυτό επιτυγχάνεται με τον παρακάτω τρόπο:

```
1 alter table public.brest2015 add column timestamp timestamp without time zone
2 update public.brest2015 set timestamp=to_timestamp(time/1000);
```

Listing 4.5: Timestamp Alteration

Στον καθορισμό των συντεταγμένων υπάρχουν συγκεκριμένα διαστήματα τιμών για το γεωγραφικό πλάτος και το γεωγραφικό μήκος που πρέπει να τηρούνται. Πιο αναλυτικά, το γεωγραφικό πλάτος καθορίζεται σε μοίρες και πρέπει να βρίσκεται εντός του εύρους από -90° έως 90°, ενώ το γεωγραφικό μήκος καθορίζεται επίσης σε μοίρες και πρέπει να βρίσκεται εντός του διαστήματος από -180° έως 180°. Για να αποφευχθούν κενές ή λανθασμένες τιμές, σε κάθε dataset επιβάλλονται αυτά τα όρια στο γεωγραφικό πλάτος και μήκος:

```
1 delete from public.brest2015 where latitude>=90    or latitude<=-90
2 delete from public.brest2015 where longitude<=-180 or longitude>180
```

Listing 4.6: Coordinates

Σε κάθε πίνακα έχουμε ορίσει τις στήλες *longitude* και *latitude* στις οποίες αναγράφονται οι συντεταγμένες κάθε πλοίου. Το γεωγραφικό πλάτος ενός σημείου μετριέται σε μοίρες και υποδεικνύει την απόσταση του σημείου από τον ισημερινό κύκλο ενώ το γεωγραφικό μήκος ενός σημείου επίσης μετριέται σε μοίρες και υποδεικνύει την απόσταση του σημείου από τον μεσημβρινό κύκλο. Αυτές οι πληροφορίες δεν είναι ευανάγνωστες από τον χρήστη, καθώς δεν μπορεί να υπολογίσει την ακρίβεια της τοποθεσίας ενός σημείου στην επιφάνεια της Γης.

Οι γεωγραφικές θέσεις μπορούν να προβληθούν σε ένα δισδιάστατο επίπεδο, διατηρώντας ιδιότητες όπως η κατεύθυνση, η περιοχή κλπ. και παραμορφώνοντας άλλες. Για την σωστή αναπαράσταση των συντεταγμένων πρέπει να εφαρμοστεί ένα Σύστημα Συντεταγμένων (CRS) το οποίο προσαρμόζεται σε μια συγκεκριμένη περιοχή. Τα σύνολα δεδομένων που χρησιμοποιούμε αφορούν την Ευρώπη και παρέχουν δεδομένα που εκφράζονται χρησιμοποιώντας τα CRS WGS84 (EPSG:4326) και ETRS89/LAEA Europe (EPSG:3035a). Χρησιμοποιώντας λοιπόν την συνάρτηση *ST_MakePoint* του PostGIS, ορίζουμε ένα σημείο (POINT) από δύο συντεταγμένες δεδομένου του συστήματος αναφοράς τους (CRS):

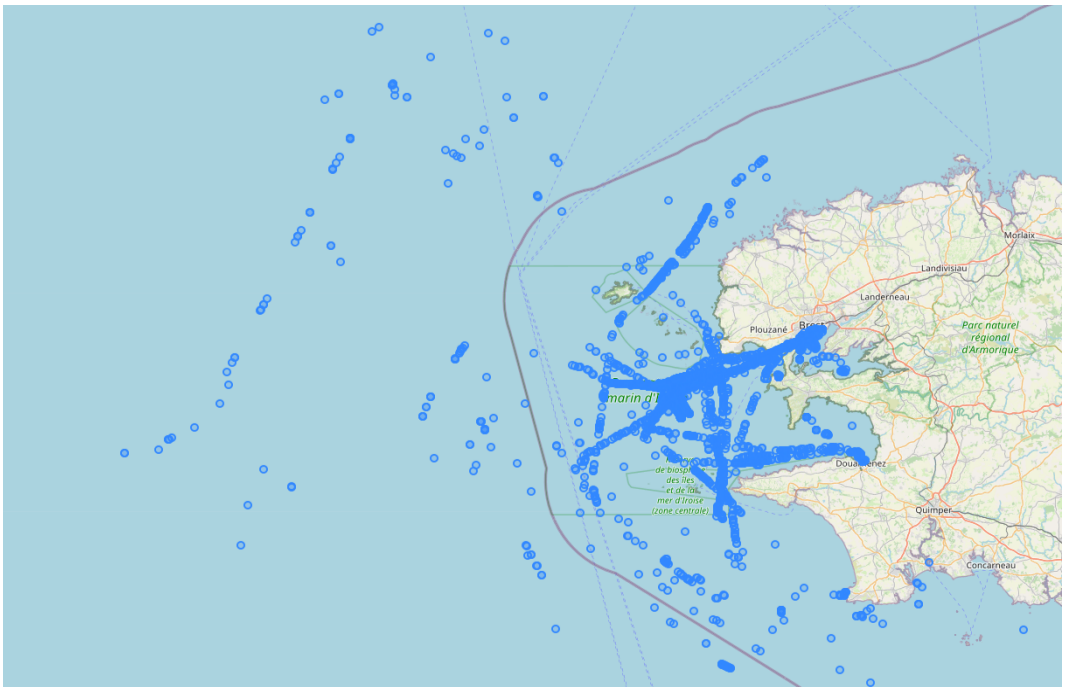
```
1 alter table public.eu2015 add column point geometry(Point)
2
3 update public.eu2015
4 set point=ST_SetSRID(ST_MakePoint(longitude,latitude),3035)
```

Listing 4.7: Points

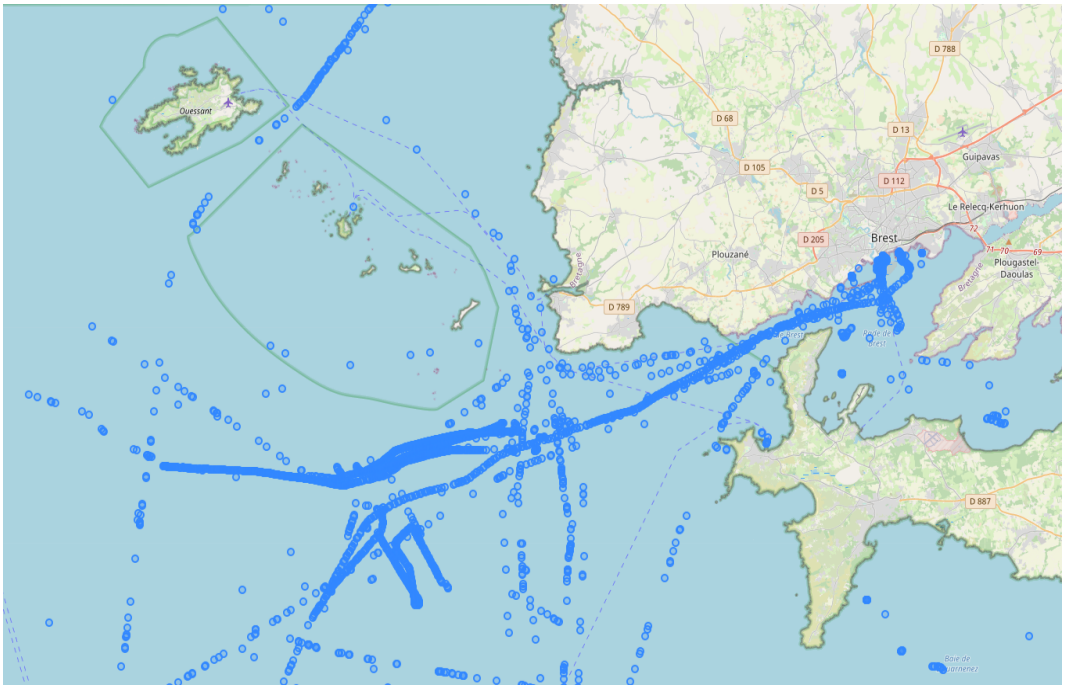
4.4 Οπτικοποίηση Δεδομένων

Η οπτικοποίηση δεδομένων είναι η αναπαράσταση των δεδομένων με γραφικά μέσα με σκοπό την ευκολότερη κατανόηση των τάσεων, των πληροφοριών και των συσχετίσεων που περιέχονται σε ένα σύνολο δεδομένων. Οι οπτικοποιήσεις βοηθούν στην απλοποίηση της πολυπλοκότητας των δεδομένων και μπορούν να αποκαλύψουν κρυμμένα πρότυπα και ακραία σημεία σε αυτά. Επιπλέον, μπορούν να ανιχνεύσουν προβλήματα εναρμονισμού των δεδομένων ή ανωμαλίες, προσφέροντας τη δυνατότητα καθαρισμού των δεδομένων και την βελτίωση της ακεραιότητάς τους.

Στο προηγούμενο κεφάλαιο πραγματοποιήθηκε η εναρμόνιση των δεδομένων του κάθε πίνακα με βάση το ίδιο σύστημα αναφοράς (CRS). Αυτή η διαδικασία επιτρέπει την απεικόνιση της θέσης κάθε πλοίου στο χάρτη κατά τη χρονική στιγμή της μετάδοσης του σήματος. Παρακάτω παρουσιάζεται ένα στιγμιότυπο του χάρτη, ο οποίος δημιουργείται με τη χρήση της επέκτασης PostGIS, με την αναπαράσταση δείγματος πλοίων από τον πίνακα *brest2015*:



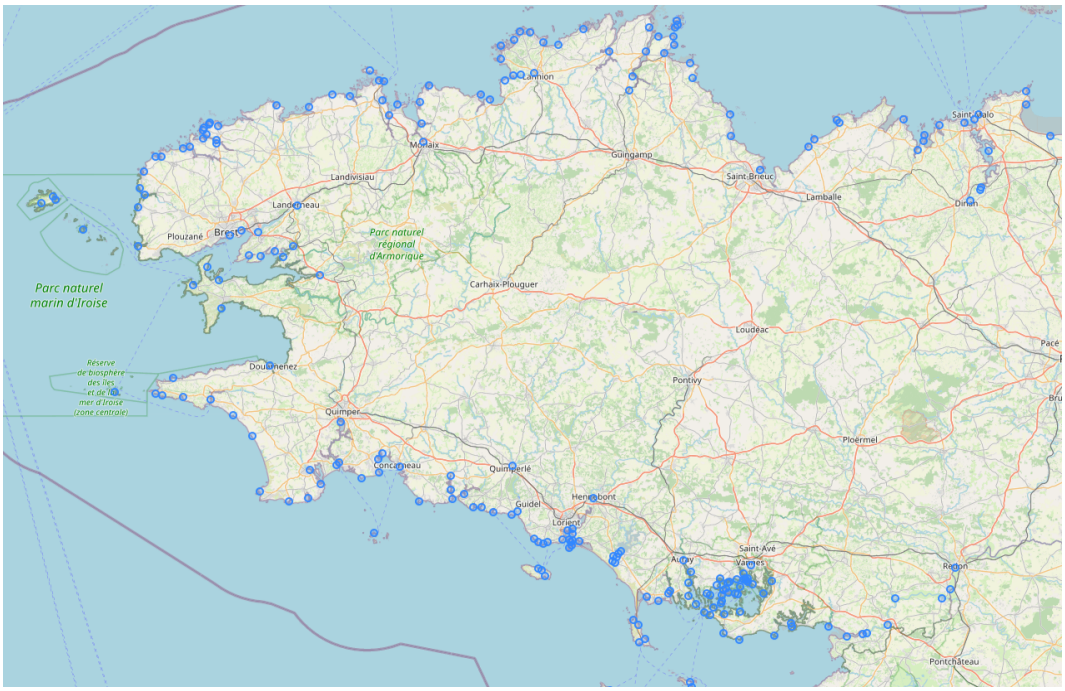
Σχήμα 4.2: Σημεία πλοίων στο χάρτη



Σχήμα 4.3: Σημεία πλοίων στο χάρτη

Εκτός από την απεικόνιση των γεωγραφικών θέσεων των πλοίων, η οπτικοποίηση των θέσεων των λιμανιών αποτελεί εξίσου σημαντικό στοιχείο στην ανάλυση των δεδομένων. Μέσω αυτής της διαδικασίας, επιτυγχάνεται όχι μόνο η κατανόηση της γεωγραφικής διάταξης των λιμανιών και των θαλάσσιων διαδρομών, αλλά δίνεται και η δυνατότητα σύνδεσης των πινάκων της βάσης μέσω γεωμετρικών στοιχείων.

Παρακάτω παρουσιάζεται μια εικόνα του χάρτη, η οποία απεικονίζει τις τοποθεσίες των λιμανιών στη Βρετάνη. Τα δεδομένα αυτά προέρχονται από τη νέα στήλη *points* του πίνακα *brest2015*:



Σχήμα 4.4: Λιμάνια στην περιοχή της Βρετάνης

Αυτή η οπτικοποίηση παρέχει μια εικόνα της γεωγραφικής θέσης των λιμανιών στη Βρετάνη και μπορεί να χρησιμοποιηθεί για την κατανόηση της διακύμανσης και της κατανομής αυτών των λιμανιών.

Οι δυνατότητες της επέκτασης PostGIS είναι πολλές και πολύπλευρες. Ένα ακόμα χρήσιμο παράδειγμα οπτικοποίησης είναι η αναπαράσταση της συνολικής πορείας του κάθε πλοίου. Χρησιμοποιώντας τη συνάρτηση ST_MakeLine, συνδέουμε τα σημεία από τα οποία το εκάστοτε πλοίο μεταδίδει σήματα, προκειμένου να απεικονίσουμε τη συνεχή διαδρομή του.

```
1 select ship_id, ST_Makeline( point order by timestamp) as line
2 from public.brest2015
3 group by ship_id
```

Listing 4.8: ST_Makeline

Σχήμα 4.5: Διαδρομή τυχαίου πλοίου

Στο παραπάνω στιγμιότυπο παρουσιάζεται η διαδρομή ενός τυχαίου πλοίου, η οποία αντιστοιχεί στο σύνολο των σημείων από τα οποία έχει μεταδώσει σήματα. Αυτή η οπτικοποίηση βοηθάει στην επισκόπηση της διαδρομής του πλοίου και των περιοχών που έχει επισκεφθεί ή διασχίσει κατά τη διάρκεια της μετάδοσης σημάτων.

Σε αυτό το σημείο, επιλέγονται όλα τα πλοία από τους πίνακες και υπολογίζεται η συνολική διαδρομή τους. Έπειτα, αυτές οι διαδρομές εισάγονται ανά πλοίο στον νέο πίνακα με το όνομα *Routes* στη βάση δεδομένων:

```
CREATE TABLE Routes (Ship_id bigint, Route geometry)

insert into Routes (
select ship_id, ST_Makeline( point order by timestamp) as Route
from public.brest2015
group by ship_id
union all
select ship_id, ST_Makeline( point order by timestamp) as Route
from public.eu2015
group by ship_id)
union all
select MMSI, ST_Makeline( point order by timestamp) as Route
from public.eu2015
group by ship_id)
```

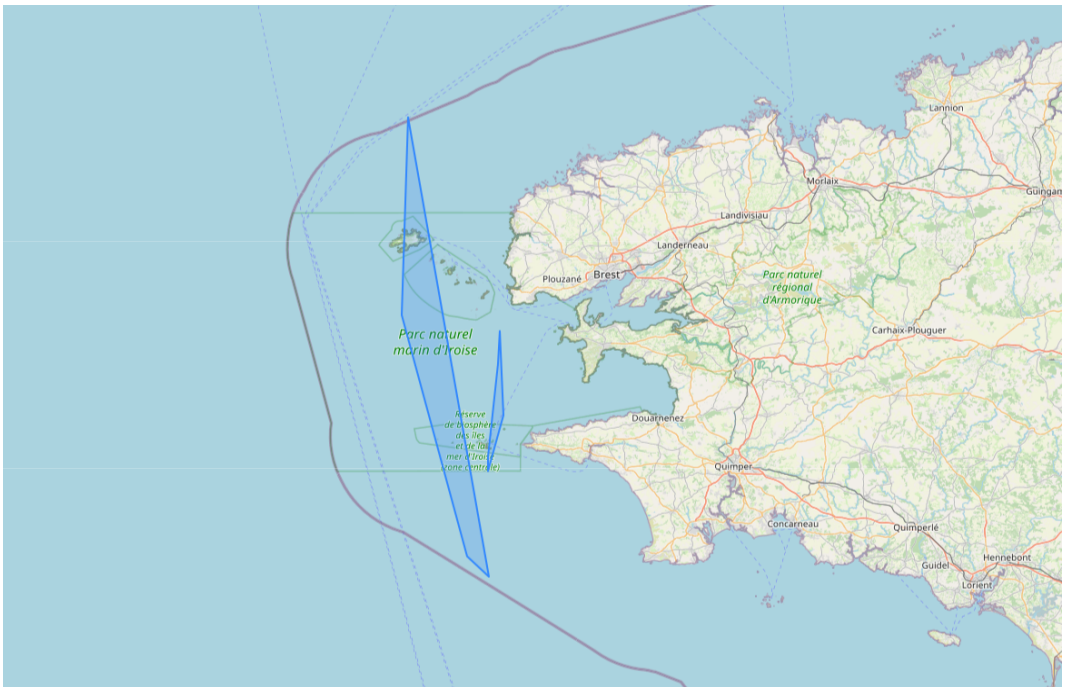
Listing 4.9: Δημιουργία πίνακα Route

Μέσω του νέου πίνακα γίνεται πολύ πιο άμεση και δυναμική η αναζήτηση της διαδρομής ενός πλοίου, ειδικά όταν χρειάζεται να εντοπίσουμε πλοία με συγκεκριμένα χαρακτηριστικά. Αφού εντοπίσουμε τα πλοία που έχουν, για παράδειγμα, διανύσει μεγαλύτερες αποστάσεις ή έχουν πλεύσει κοντά σε λιμένες ή σε πιο ανοιχτά νερά, μπορούμε να εξετάσουμε πιο λεπτομερώς τις πορείες τους και να ανακαλύψουμε σχέσεις και πρότυπα που μπορεί να μας παρέχουν σημαντική πληροφορία για τη ναυτιλιακή τους δραστηριότητα.

Με τη χρήση της συνάρτησης *ST_ConcaveHull*, μπορούμε να δημιουργήσουμε γεωμετρικά σχήματα όπως πολύγωνα, τα οποία αντιπροσωπεύουν περιοχές με κάποιο διακριτό χαρακτηριστικό. Για παράδειγμα, μπορούμε να χρησιμοποιήσουμε αυτήν τη συνάρτηση για να δημιουργήσουμε πολύγωνα που αντιπροσωπεύουν τις περιοχές όπου πλοία έχουν διασχίσει ή έχουν εκπέμψει σήματα. Αυτά τα πολύγωνα μπορούν να αποδειχθούν πολύ χρήσιμα για την οπτικοποίηση της συνολικής κίνησης των πλοίων σε μια περιοχή.

```
1 select ST_ConcaveHull(ST_Collect(point order by ship_id), 0.99, false) as R
2 from Rsting_Areas
```

Listing 4.10: ST_ConcaveHull



Σχήμα 4.6: Πολύγωνο Στάσιμων Περιοχών

Στο παραπάνω στιγμιότυπο, παρουσιάζονται δύο πολύγωνα που δημιουργούνται από τα δεδομένα των πινάκων *brest2015* και *eu2015*. Αυτά τα πολύγωνα αντιπροσωπεύουν περιοχές που μπορούν να χαρακτηριστούν ως "Περιοχές Ανάπαυσης," όπου πλοία έχουν μηδενική

ταχύτητα για περισσότερο από μία ώρα. Ο προσδιορισμός τέτοιων περιοχών είναι σημαντικός για τον ναυτιλιακό τομέα, καθώς αποτελούν περιοχές ανάπαυσης και αναμονής για τα πλοία. Τα χαρακτηριστικά της διαμονής των πλοίων σε αυτές τις περιοχές μπορεί να έχουν σημαντικές επιπτώσεις και να συμβάλλουν στην ασφάλεια και την ομαλή ροή της ναυτιλίας.

Ο συνδυασμός των πινάκων *brest2015* και *eu2015* έχει σημαντική αξία, καθώς τα δεδομένα εκτείνονται στην ίδια χρονική περίοδο. Η δημιουργία πολυγώνων με διαφορετικά χαρακτηριστικά μπορεί να αναδείξει τις τάσεις και τα φαινόμενα που λαμβάνουν χώρα σε συγκεκριμένες περιοχές και πως αυτά επηρεάζουν τις κινήσεις των πλοίων.

4.5 Queries

Για την πιο εκτενή κατανόηση των δεδομένων, αναπτύχθηκαν διάφορα ερωτήματα που επιστρέφουν σημαντικές πληροφορίες, συμπεριλαμβανομένων των εξής [49]:

- **Το πιο πολυσύχναστο λιμάνι**: Δίνει πληροφορίες σχετικά με το λιμάνι που χρησιμοποιείται συχνότερα, προσδιορίζοντας κρίσιμες θαλάσσιες διαδρομές.
- **Τη διάρκεια της μετάδοσης σημάτων του εκάστοτε πλοίου/μέρα**: Αυτό το ερώτημα βοηθά να κατανοήσουμε πόσο χρόνο αφιέρωνε κάθε πλοίο στη μετάδοση σημάτων καθημερινά.
- **Πλοία προσαραγμένα σε ανοιχτά νερά**: Επιτρέπει τον εντοπισμό πλοίων που βρίσκονται σε κατάσταση ακινησίας σε ανοιχτές θαλάσσιες περιοχές, προσφέροντας πληροφορίες για τη θαλάσσια κυκλοφορία.
- **Πλοία σε ανοιχτά νερά με τη μεταξύ τους απόσταση <200m**: Αυτό το ερώτημα μας επιτρέπει να παρακολουθήσουμε πλοία που ενδέχεται να βρίσκονται κοντά το ένα στον άλλο, προκειμένου να ανιχνεύσουμε ενδεχόμενες αλληλεπιδράσεις.
- **Τις στάσεις των πλοίων και την διάρκεια αυτών**: Επιτρέπει τον εντοπισμό των χρονικών στιγμών όπου τα πλοία έχουν σταματήσει την κίνησή τους και βρίσκονται σε στάση.

Ενδεικτικά μερικά από τα παραπάνω:

```
select a.EU_ship,a.eu_position,a.Brest_ship,a.dt_brest,a.brest_position,
from
  (
  select eu.ship_id as EU_ship,cast(eu.timestamp as date) as dt_eu,eu.point
    as eu_position,
  bre.ship_id as Brest_ship,cast(bre.timestamp as date) as dt_brest,bre.point
     as brest_position
  from public.eu2015 eu
  inner join public.brest2015 bre
  on ST_dWithin(bre.point,eu.point,0.002)
  inner join public.''Brittany_Ports'' bri
  on ST_Disjoint(bri.geometry,bre.point)
```

```
  where 1=1
  and eu.sog>0
  and bre.sog>0
  )a
  where a.dt_eu=a.dt_brest
```

Listing 4.11: Πλοία σε ανοιχτά νερά με απόσταση <200m

```
CREATE TABLE Movements AS
select a.ship_id,a.ts as Starting_time,a.tf Ending_time,a.sog1,a.sog2,a.p1 as
     point1,a.p2 as point2, st_distance(a.p1,a.p2) as distance,
extract(epoch from (a.tf-a.ts)) as duration, -- seconds remain
(st_distance(a.p1,a.p2)/extract(epoch from (a.tf-a.ts))) as speed_per_s --m/s
from (
    select ship_id,
    LEAD(ship_id) OVER (ORDER BY ship_id, timestamp) AS id2, --next ship
    timestamp as ts, --starting time
    LEAD(timestamp) OVER (ORDER BY ship_id, timestamp) AS tf, --final time
    sog as sog1, -- starting sog
    LEAD(sog) OVER (ORDER BY ship_id, timestamp) AS sog2, --final sog
    point as p1, --starting point
    LEAD(point) OVER (ORDER BY ship_id, timestamp) AS p2 --final point
    from public.brest2015
    )a
where a.ship_id=a.id2;

--Stop begins when velocity reduces
drop table if exists stop_start
select ship_id,Starting_time as Starting_tm
into temp table stop_start
from public.movements
where sog1>0.1 and sog2<=0.1

--Stop ends when velocity increases
drop table if exists stop_end
select ship_id,Ending_time  as Ending_tm
into temp table stop_end
from public.movements
where sog1<=0.1 and sog2>0.1

--filtered stops
CREATE TABLE Stops as (select ss.ship_id,ss.Starting_tm,a.Ending_tm,(
    Ending_tm-Starting_tm) as duration
from stop_start ss
inner join lateral (
  select Ending_tm
  from stop_end se
  where ss.ship_id=se.ship_id
  and Starting_tm<=Ending_tm
  order by Ending_tm limit 1
        )a on (true)
)t
```

Listing 4.12: Στάσεις των πλοίων

Στο παραπάνω query δημιουργείται ένας πίνακας με την ονομασία *Stops* για την αποθήκευση δεδομένων που περιγράφουν τις στάσεις των πλοίων. Ο πίνακας περιέχει πληροφορίες όπως ο μοναδικός αριθμός πλοίου, ο χρόνος έναρξης της στάσης και ο χρόνος λήξης της στάσης, καθώς και η διάρκεια της στάσης σε δευτερόλεπτα.

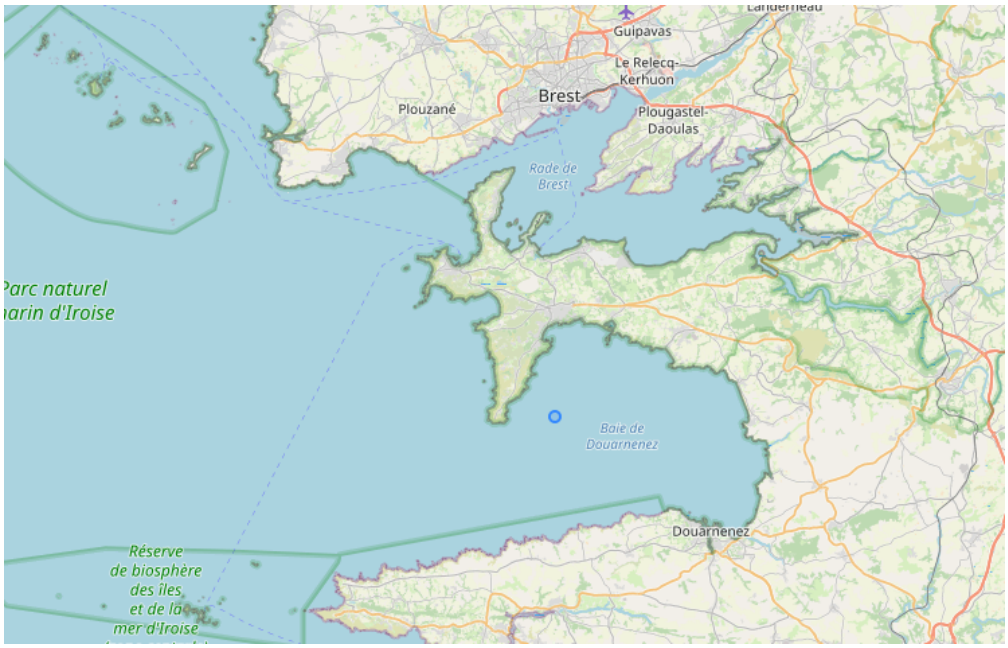
ship_id	starting_tm	ending_tm	duration
37100300	2015-10-07 03:48:24	2015-10-07 05:32:03	01:43:39
37100300	2015-10-07 05:32:24	2015-10-08 09:29:13	1 day 03:56:49
20520400	2015-10-02 10:40:43	2015-10-02 10:41:13	00:00:30
20520400	2015-10-02 10:43:43	2015-10-02 10:44:13	00:00:30
20520400	2015-10-02 15:08:43	2015-10-08 16:02:44	00:54:01
20520400	2015-10-03 11:33:37	2015-10-03 11:33:41	00:00:04
20520400	2015-10-03 11:33:53	2015-10-03 11:34:55	00:01:02
20520400	2015-10-03 11:35:01	2015-10-03 11:35:05	00:00:04
20520400	2015-10-03 11:35:33	2015-10-03 11:35:45	00:00:12
20520400	2015-10-03 11:35:55	2015-10-03 12:00:44	00:24:49
20520400	2015-10-03 12:09:45	2015-10-05 09:34:41	1 day 21:24:56

Πίνακας 4.8: Πίνακας *Stops*

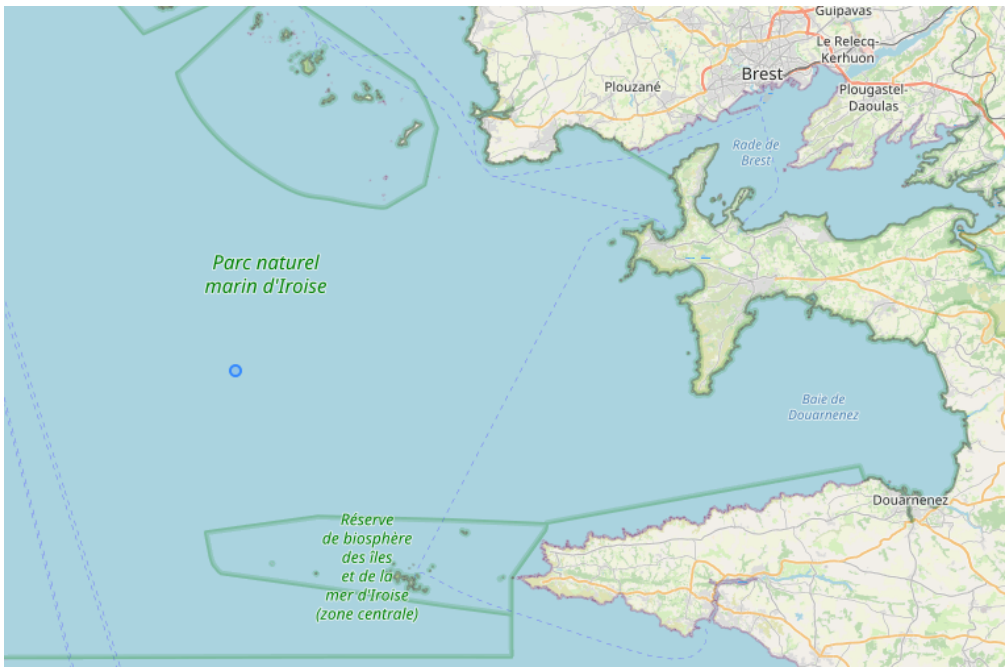
Στον πίνακα 4.8 περιλαμβάνονται δείγματα εγγραφών του πίνακα *Stops*, που περιγράφουν τις στάσεις των πλοίων. Αυτές οι πληροφορίες μας επιτρέπουν να αναλύσουμε και να κατανοήσουμε τη συμπεριφορά των πλοίων κατά τη διάρκεια των στάσεών τους. Αξίζει να μελετηθούν και να χαρτογραφηθούν τα πλοία που φαίνεται να είναι στάσιμα για περισσότερο από μία ημέρα και να εξετασθεί η τοποθεσία τους. Σε αυτό το πλαίσιο, μπορεί επίσης να ελεγχθεί εάν οι στάσεις αυτές λαμβάνουν χώρα κοντά σε λιμάνια ή σε ανοιχτές θαλάσσιες περιοχές και στην δεύτερη περίπτωση, να αναρωτηθούμε γιατί συμβαίνει αυτό. Αυτός ο πίνακας, λοιπόν, μπορεί να χρησιμοποιηθεί για περαιτέρω ανάλυση των ναυτιλιακών δραστηριοτήτων και των μοτίβων κίνησης των πλοίων.

Στα παρακάτω στιγμιότυπα εμφανίζονται τα σημεία στα οποία ξεκινούν και σταματούν οι στάσεις ενός τυχαίου πλοίου. Μελετώντας τις κινήσεις και τα γεωγραφικά σημεία όπου παρατηρούνται παρατεταμένες στάσεις, μπορούμε να αποφανθούμε για την φύση των στάσεων αυτών. Πιο αναλυτικά, μπορούμε να διακρίνουμε δύο βασικές κατηγορίες στάσεων: τις στάσεις ελλιμενισμού και τις στάσεις αγκυροβόλησης. Οι στάσεις αγκυροβόλησης συνήθως υποδεικνύουν την παραμονή του πλοίου στο εσωτερικό των λιμένων ή την αναμονή του προτού αναχωρήσει. Από την άλλη πλευρά, οι στάσεις ελλιμενισμού μπορεί να οφείλονται σε ποικίλους παράγοντες, όπως τον προσδιορισμό της σωστής κατεύθυνσης, την αποφυγή εμποδίων, την προσέγγιση του πλοίου σε περιοχές αλιείας, τις στάσιμες περιοχές κτλ.

Αυτή η ανάλυση μας επιτρέπει να κατανοήσουμε καλύτερα την κίνηση και τις δραστηριότητες του πλοίου καθώς και τους λόγους που το οδηγούν να πραγματοποιεί στάσεις κατά τη διάρκεια της διαδρομής του.



Σχήμα 4.7: Starting Location



Σχήμα 4.8: Stop Location

Σχήμα 4.9: Σημεία Στάσεων

4.6 Πρόβλεψη χρονοσειρών

4.6.1 Εισαγωγή

Σε αυτή την ενότητα, πραγματοποιείται πρόβλεψη χρονοσειρών για τις γεωγραφικές θέσεις πλοίων σε ένα χρονικό διάστημα. Αρχικά, απαιτείται η μετατροπή του συνόλου δεδομένων της χρονοσειράς σε ένα πρόβλημα επίβλεψης. Αυτό επιτυγχάνεται χρησιμοποιώντας τα προηγούμενα χρονικά βήματα των δεδομένων ως μεταβλητές εισόδου(Ορίζοντες) και το επόμενο χρονικό βήμα ως τη μεταβλητή εξόδου. Στο πείραμα επιλέγονται 4 διακριτοί ορίζοντες, δηλαδή 4 διαφορετικά μήκη των ακολουθιών εισόδου, προκειμένου να εξετασθεί ποιοι ενδέχεται να οδηγήσουν σε καλύτερα, πιο αποδοτικά μοντέλα. Επίσης, εφαρμόζεται μια εξειδικευμένη τεχνική για την αξιολόγηση του μοντέλου που ονομάζεται *walk-forward validation*. Η τεχνική αυτή περιλαμβάνει την εκπαίδευση του μοντέλου σε ένα σταθερό εύρος ιστορικών δεδομένων και, στην συνέχεια, την πραγματοποίηση διαδοχικών προβλέψεων ενός βήματος για κάθε επόμενη χρονική στιγμή. Το σύνολο δεδομένων διαιρείται αρχικά σε 2 υποσύνολα, το σύνολο εκπαίδευσης (train dataset) και το σύνολο ελέγχου (test dataset) επιλέγοντας ένα διακριτό σημείο διαχωρισμού. Ένας συνηθισμένος διαχωρισμός για μεγάλα datasets είναι 80% του συνόλου δεδομένων να προορίζεται για εκπαίδευση και 20% για έλεγχο. Σε αυτό το πλαίσιο, χρησιμοποιούνται 10 διαφορετικές στατιστικές μέθοδοι που εκτιμούν τις μεταβλητές ενδιαφέροντος. Στόχος μας είναι η συγκριτική αξιολόγηση των αποτελεσμάτων κάθε μεθόδου και η επιλογή της πιο κατάλληλης για την ακριβέστερη πρόβλεψη (forecast) τους. Επιπλέον, χρησιμοποιούνται 6 διαφορετικά κριτήρια υπολογισμού σφαλμάτων με βάση τα οποία γίνεται η τελική αξιολόγηση των αποτελεσμάτων.

4.6.2 Δεδομένα

Οι μεταβλητές ενδιαφέροντος είναι το γεωγραφικό πλάτος και γεωγραφικό μήκος ενός πλοίου. Το πείραμα πραγματοποιείται χρησιμοποιώντας ιστορικά δεδομένα θέσεων ενός πλοίου κατά την χρονική περίοδο μεταξύ Οκτωβρίου 2015 και Μαρτίου 2016. Τα δεδομένα εμφανίζουν καθημερινή συχνότητα χωρίς σταθερό μοτίβο, καθώς αφορούν μεταδόσεις σημάτων που γίνονται διάσπαρτα κατά τη διάρκεια αυτής της περιόδου.

4.6.3 Συντελεστής Συσχέτισης Pearson

Για την διεξαγωγή της εμπειρικής μελέτης, είναι χρήσιμη η εύρεση της γραμμικής σχέσης μεταξύ του γεωγραφικού πλάτους και μήκους. Ένα μέτρο της συσχέτισης μεταξύ των δύο μεταβλητών είναι μέσω του συντελεστή συσχέτισης Pearson, ο οποίος υπολογίζεται στον ακόλουθο πίνακα:

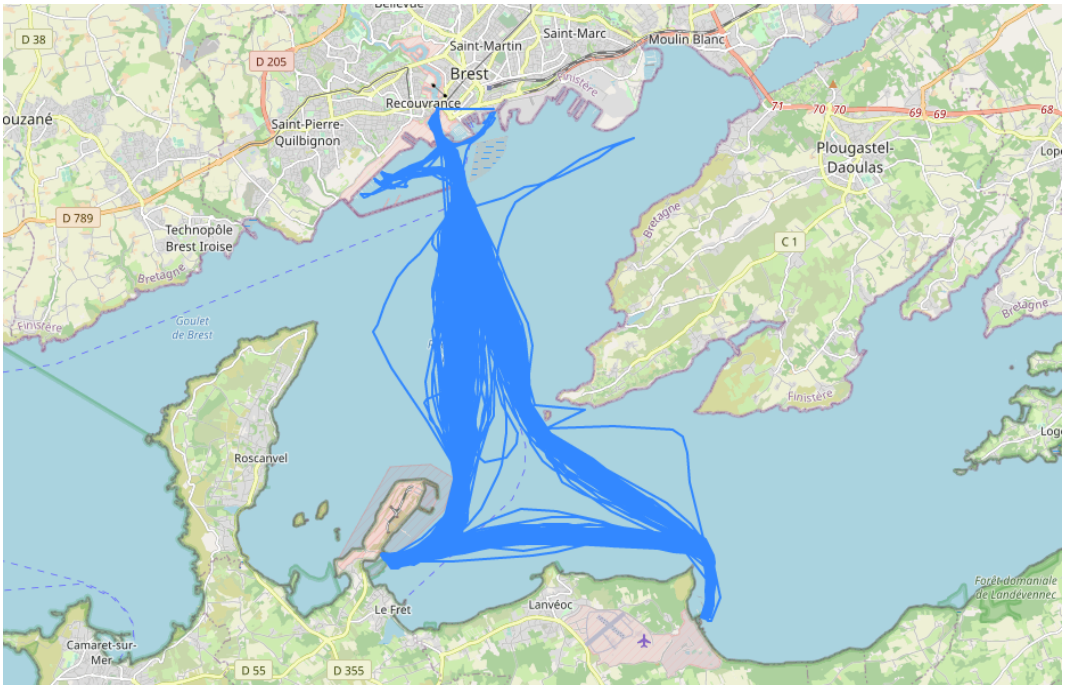
$$\begin{bmatrix} 1 & -0.3563 \\ -0.3563 & 1 \end{bmatrix}$$

Η τιμή **1** κατά μήκος της διαγώνιου υποδεικνύει την τέλεια συσχέτιση κάθε μεταβλητής με τον εαυτό της. Το μη διαγώνιο στοιχείο **-0,3563** υποδηλώνει μια μέτρια αρνητική συσχέτιση

μεταξύ γεωγραφικού μήκους και γεωγραφικού πλάτους, καθώς η απόλυτη τιμή του είναι μικρότερη από 0,5. Μια αρνητική συσχέτιση υποδηλώνει ότι καθώς η μία μεταβλητή αυξάνεται, η άλλη μεταβλητή τείνει να μειώνεται.

4.6.4 Μετασχηματισμός Δεδομένων

Για την πρόβλεψη της χρονοσειράς της διαδρομής ενός πλοίου, επιλέγεται ένα μοναδικό αναγνωριστικό (ID) πλοίου με 22.347 εγγραφές από το Dataset 2. Αυτή η επιλογή λαμβάνει υπόψη τους περιορισμούς των χρονικών περιθωρίων του Kaggle, καθώς και τη διαθεσιμότητα των υπολογιστικών πόρων, όπως η δέσμευση της μονάδας επεξεργασίας γραφικών (GPU). Το συγκεκριμένο πλοίο, όπως φαίνεται στο παρακάτω στιγμιότυπο, πραγματοποίησε σημαντικό αριθμό επαναλαμβανόμενων διαδρομών μεταξύ 3 λιμανιών.



Σχήμα 4.10: Διαδρομές πλοίου προς πρόβλεψη

Παρατηρώντας τα δεδομένα για το συγκεκριμένο αναγνωριστικό (ID), παρατηρούμε ότι υπάρχουν ανωμαλίες στις τιμές των γεωγραφικών συντεταγμένων, δηλαδή του γεωγραφικού πλάτους (latitude) και του γεωγραφικού μήκους (longitude). Αυτές οι ανωμαλίες μπορεί να επηρεάσουν αρνητικά τον υπολογισμό των προβλέψεων για αυτές τις μεταβλητές, καθώς οι προβλέψεις βασίζονται στις προηγούμενες τιμές.

Για να διασφάλιση την ακεραιότητα και την αξιοπιστία των δεδομένων, δεν λαμβάνονται υπόψη οι τιμές που παρουσιάζουν σημαντικές αποκλίσεις ή ανωμαλίες. Αυτό γίνεται για

να αποφευχθεί η απορρόφηση αυτών των ανωμαλιών στον υπολογισμό των προβλέψεων, καθώς οι τέτοιες αποκλίσεις μπορεί να οδηγήσουν σε εσφαλμένα αποτελέσματα. Στο παρακάτω στιγμιότυπο φαίνεται η αποκοπή των ακραίων τιμών από το σύνολο των παρατηρήσεων του επιλεγμένου πλοίου:

```
import numpy as np

ship_data = df[df['id'] ==227635210] #22347

ship_data = ship_data[ship_data['latitude'] <= 50]
ship_data = ship_data[ship_data['longitude'] >= -5]
# Select only the desired columns and sort by timestamp
desired_columns = ['timestamp','longitude', 'latitude']
ship_data = ship_data[desired_columns].sort_values(by='timestamp')

print(ship_data)
```

Listing 4.13: Drop out of range values

4.7 Εμπειρική Μεθοδολογία

4.7.1 Βασικό Μοντέλο

Χρησιμοποιείται ένα βασικό πλαίσιο AR(p)

$$y_t = f(y_{t-1}, \ldots, y_{t-p}) + e_t$$

[50] όπου:

- y_t είναι η τρέχουσα τιμή της χρονοσειράς.
- f αντιπροσωπεύει τη σχέση που μεταξύ της τρέχουσας τιμής της χρονοσειράς και των προηγούμενων υστερήσεων
- y_{t-p} αναφέρεται στην τιμή της χρονοσειράς σε προηγούμενη χρονική στιγμή $t-p$, όπου p είναι η υστέρηση ή η χρονική καθυστέρηση.
- e_t Είναι ο όρος σφάλματος στη χρονική στιγμή t.

Σε αυτό το πλαίσιο, εκτελούνται τέσσερις παλινδρομήσεις με (p1, p2, p3, p4) = (1, 2, 5, 10) υστερήσεις, αντιπροσωπεύοντας τους Ορίζοντες 1, 2, 5 και 10 αντίστοιχα.

4.7.2 Μέθοδοι

Παρακάτω, περιγράφονται σύντομα τα διάφορα μοντέλα που εφαρμόστηκαν, χρησιμοποιώντας την παραπάνω διάταξη AR(p). Ο στόχος αυτής της άσκησης είναι να προβλέψουμε τις εξαρτημένες μεταβλητές την περίοδο t+1 [50].

1. **Extreme Gradient Boosting (XGBoost)**
   Το XGBoost είναι μια μέθοδος εκμάθησης συνόλου που συνδυάζει πολλά αδύναμα μοντέλα (δέντρα) για να δημιουργήσει ένα ισχυρό μοντέλο πρόβλέψης.

$$minL_{\text{XGBoost}} = \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} f_k(y_{i-1}, \ldots, y_{i-p}) \right)^2 + \sum_{k=1}^{K} \Omega(f_k)$$

2. **Random Forest**
   Η μέθοδος Random Forest κατασκευάζει μια συλλογή από δέντρα αποφάσεων και συγκεντρώνει τις προβλέψεις τους για να πραγματοποιήσει ακριβείς προβλέψεις.

$$minL_{\text{RandomForest}} = \sum_{i=1}^{n} \left(y_i - \text{RF}(y_{i-1}, \ldots, y_{i-p}) \right)^2$$

3. **Decision Tree**
   Τα Δέντρα Αποφάσεων χρησιμοποιούν την ιεραρχική διαίρεση των μεταβλητών πρόβλεψης για να πραγματοποιήσουν προβλέψεις βασισμένες στα χαρακτηριστικά εισόδου.

$$minL_{\text{DecisionTree}} = \sum_{i=1}^{n} \left(y_i - \text{DT}(y_{i-1}, \ldots, y_{i-p})\right)^2$$

4. **Extra Trees**
   Η μέθοδος Extra Trees είναι παρόμοια με το Random Forest αλλά χρησιμοποιεί τυχαία όρια για τον διαχωρισμό των χαρακτηριστικών, δημιουργώντας έτσι διάφορα και ανεξάρτητα δέντρα.

$$minL_{\text{ExtraTrees}} = \sum_{i=1}^{n} \left(y_i - \text{ET}(y_{i-1}, \ldots, y_{i-p})\right)^2$$

5. **K-nearest neighbors (KNN)**
   Η μέθοδος K-Nearest Neighbors πραγματοποιεί προβλέψεις βάσει του μέσου όρου των K πλησιέστερων σημείων δεδομένων, κάνοντάς τον ευαίσθητο στα τοπικά πρότυπα.

$$minL_{\text{KNN}} = \sum_{i=1}^{n} \left(y_i - \text{KNN}(y_{i-1}, \ldots, y_{i-p})\right)^2$$

6. **Ridge Regression**
   Η μέθοδος Ridge Regression προσθέτει κανονικοποίηση στην Μέθοδο των Ελαχίστων Τετραγώνων (OLS) για να αποτρέψει το overfitting, επιβάλλοντας ποινές στις μεγάλες τιμές των συντελεστών.

$$minL_{\text{Ridge}} = \sum_{t=p+1}^{T} (y_t - \beta_0 - \beta_1 y_{t-1} - \ldots - \beta_p y_{t-p})^2 + \lambda(\beta_0^2 + \beta_1^2 + \ldots + \beta_p^2)$$

7. **Huber Loss**
   Η μέθοδος Huber Loss ισορροπεί το τετραγωνικό και απόλυτο σφάλμα, κάνοντάς το λιγότερο ευαίσθητο στις ακραίες τιμές σε σύγκριση με το τετραγωνικό σφάλμα.

$$minL_{\text{Huber}} = \sum_{i=p+1}^{n} \begin{cases} \frac{1}{2}\left(y_i - \beta_0 - \beta_1 y_{i-1} - \ldots - \beta_p y_{i-p}\right)^2, \\ \quad \text{if } |y_i - \beta_0 - \beta_1 y_{i-1} - \ldots - \beta_p y_{i-p}| \leq \delta, \\ \delta\left(|y_i - \beta_0 - \beta_1 y_{i-1} - \ldots - \beta_p y_{i-p}| - \frac{\delta}{2}\right), \\ \quad \text{otherwise,} \end{cases}$$

8. **Linear Regression (OLS)**
   Η Μέθοδος των Ελαχίστων Τετραγώνων (OLS) εκτιμά τους συντελεστές (coefficients) με τον τρόπο που ελαχιστοποιεί το άθροισμα των τετραγώνων των διαφορών μεταξύ των παρατηρούμενων και προβλεπόμενων τιμών.

$$minL_{\text{OLS}} = \sum_{t=p+1}^{T} (y_t - \beta_0 - \beta_1 y_{t-1} - \ldots - \beta_p y_{t-p})^2$$

9. **Bayesian Ridge Regression**
   Η μέθοδος Bayesian Ridge Regression χρησιμοποιεί Μπεϋζιανή μοντελοποίηση για να εκτιμήσει τους συντελεστές, "ενημερώνοντας" τις προβλέψεις με βάση έναν πιθανοτικό αλγόριθμο (priors and posteriors).

$$minL_{\text{Bayesian}} = \sum_{t=p+1}^{T} (y_t - \beta_0 - \beta_1 y_{t-1} - \ldots - \beta_p y_{t-p})^2 + \lambda(\beta_0^2 + \beta_1^2 + \ldots + \beta_p^2)$$

10. **Stochastic Gradient Descent Regressor (SGDR)**
    Η μέθοδος Stochastic Gradient Descent ενημερώνει τους συντελεστές επαναληπτικά, χρησιμοποιώντας πληροφορίες για την ελαχιστοποίηση της συνάρτησης L.

$$minL_{\text{SGD}} = \sum_{i=p+1}^{n} (y_i - \beta_0 - \beta_1 y_{i-1} - \ldots - \beta_p y_{i-p})^2$$

Κεφάλαιο: Αποτελέσματα Πειραμάτων

5.1 Πίνακες Αποτελεσμάτων και Ανάλυση

Μέθοδος Εκτίμησης	Horizon 1	Horizon 2	Horizon 5	Horizon 10
XGBRegressor	0.99977	0.99976	0.99978	0.99982
RandomForestRegressor	0.99978	0.99982	off limits	off limits
DecisionTreeRegressor	0.99952	0.99963	0.99960	0.99952
ExtraTreesRegressor	0.99986	0.99991	0.99990	0.99989
KNeighborsRegressor	0.99930	0.99921	0.99896	0.99880
Ridge Regressor	0.99671	0.99878	0.99847	0.99666
Huber Regressor	0.99965	0.99965	0.99917	0.99843
Linear Regressor	0.99981	0.99994	0.99996	0.99996
Bayesian Ridge	0.99981	0.99994	0.99996	0.99996
SGDR Regressor	0.99755	0.99902	0.99843	0.99512

Πίνακας 5.1: R^2 - Αποτελέσματα για Γεωγραφικό μήκος

Μέθοδος Εκτίμησης	Horizon 1	Horizon 2	Horizon 5	Horizon 10
XGBRegressor	0.99993	0.99993	0.99995	0.99995
RandomForestRegressor	0.99994	0.99996	off limits	off limits
DecisionTreeRegressor	0.99989	0.99991	0.99992	0.99991
ExtraTreesRegressor	0.99995	0.99997	0.99998	0.99998
KNeighborsRegressor	0.99986	0.99982	0.99975	0.99972
Ridge Regressor	0.99976	0.99984	0.99960	0.99917
Huber Regressor	0.99995	0.99989	0.99959	0.99865
Linear Regressor	0.99995	0.99999	0.99999	1
Bayesian Ridge	0.99995	0.99999	0.99999	1
SGDR Regressor	0.99755	0.99915	0.99855	0.99542

Πίνακας 5.2: R^2 - Αποτελέσματα για Γεωγραφικό πλάτος

Στον πίνακα 5.1, όπου παρουσιάζονται τα αποτελέσματα του κριτηρίου R^2 για την πρό-

βλεψη του γεωγραφικού μήκους με διάφορες τιμές υστέρησης (lags), παρατηρούνται διακριτές επιδόσεις στην εφαρμογή κάθε μεθόδου. Το R^2 χρησιμοποιείται για την αξιολόγηση της καλής προσαρμογής ενός μοντέλου στα παρατηρούμενα δεδομένα και κυμαίνεται από 0 έως 1. Ο XGBRegressor εμφανίζει συνεχώς υψηλή απόδοση σε όλα τα επίπεδα υστέρησης. Αντίστοιχα, ο RandomForestRegressor φαίνεται να έχει εξαιρετική απόδοση στα πρώτα lags, αλλά ξεπερνά το χρονικό κατώφλι των 12 ωρών που έχει οριστεί για την εκτέλεση των πειραμάτων, κάτι που τον καθιστά χρονοβόρο και υπολογιστικά απαιτητικό. Οι DecisionTreeRegressor και ExtraTreesRegressor εκδηλώνουν ανταγωνιστική απόδοση, με τον ExtraTreesRegressor να υπερτερεί σε όλους τους ορίζοντες. Ο KNeighborsRegressor παρουσιάζει καλή επίδοση για μικρά lags, αλλά με την αύξηση τους, αυτή φαίνεται να φθίνει. Η ακρίβεια των Ridge Regressor και Huber Regressor μεταβάλλεται συνεχώς και εξασθενεί για Horizon 5 και 10. Οι Linear Regressor και Bayesian Ridge παρουσιάζουν σταθερή ισχυρή απόδοση, προσεγγίζοντας στην τιμή 1 στα Horizon 5 και Horizon 10, κάτι που υποδηλώνει πως μπορούν να καταγράφουν αποτελεσματικά γραμμικές σχέσεις μεταξύ των χαρακτηριστικών εισόδου και της προβλεπόμενη τιμής. Αντίθετα, ο SGDR Regressor αντιμετωπίζει δυσκολίες, ειδικά σε μεγαλύτερους ορίζοντες, όπως φαίνεται από την απόδοσή του στο Horizon 10.

Στον πίνακα 5.2, κατά τον ίδιο τρόπο, παρουσιάζονται τα αποτελέσματα του κριτηρίου R^2 για την πρόβλεψη του γεωγραφικού πλάτους και παρατηρείται παρόμοια συμπεριφορά των μοντέλων. Συνολικά οι Linear Regressor και Bayesian Ridge εμφανίζουν συνεχώς υψηλές τιμές για το R^2 σε όλους τους ορίζοντες, τόσο για το γεωγραφικό μήκος, όσο και το γεωγραφικό πλάτος. Αυτά τα μοντέλα είναι γραμμικά, και η ικανότητά τους να προσεγγίζουν την τιμή 1 για το R^2 σε διάφορους ορίζοντες, επαληθεύεται και από την επιλογή του παλίνδρομου μοντέλου για την πρόβλεψη των μεταβλητών.

Μέθοδος Εκτίμησης	**Horizon 1**	**Horizon 2**	**Horizon 5**	**Horizon 10**
XGBRegressor	0.0001322561	0.0001292156	0.0001184734	0.0001148898
RandomForestRegressor	0.0000961401	0.0000828332	off limits	off limits
DecisionTreeRegressor	0.0001292689	0.0001124104	0.0001191302	0.0001277038
ExtraTreesRegressor	0.0000854424	0.0000673921	0.0000732706	0.0000787775
KNeighborsRegressor	0.0001987762	0.0002141674	0.0002321913	0.0002607813
Ridge Regressor	0.0004857308	0.0003028875	0.0003459182	0.0004900866
Huber Regressor	0.0001401224	0.0001436887	0.0001977401	0.0002590106
Linear Regressor	0.0001070076	0.0040709229	0.0040459768	0.0040310704
Bayesian Ridge	0.0001064097	0.0040671449	0.0040683144	0.0040543363
SGDR Regressor	0.0004802457	0.0003147883	0.0042063541	0.0052591019

Πίνακας 5.3: Mean Absolute Error - Αποτελέσματα για Γεωγραφικό μήκος

Μέθοδος Εκτίμησης	Horizon 1	Horizon 2	Horizon 5	Horizon 10
XGBRegressor	0.0001692766	0.0001654985	0.0001443725	0.0001387394
RandomForestRegressor	0.0001235798	0.0000960256	off limits	off limits
DecisionTreeRegressor	0.0001541648	0.0001295241	0.0001222135	0.0001285755
ExtraTreesRegressor	0.0001125084	0.0000773946	0.0000708306	0.0000735306
KNeighborsRegressor	0.0001929467	0.0002281969	0.0002384376	0.0002250602
Ridge Regressor	0.0004406627	0.0003384998	0.0004512206	0.0006196807
Huber Regressor	0.0001446534	0.0002142541	0.0004138395	0.0007472835
Linear Regressor	0.0001428289	0.0028213229	0.0029594811	0.0029552838
Bayesian Ridge	0.0001428690	0.0028198094	0.0029565173	0.0029515710
SGDR Regressor	0.0004601337	0.0003027556	0.0040657012	0.0050709543

Πίνακας 5.4: Mean Absolute Error - Αποτελέσματα για Γεωγραφικό πλάτος

Στον πίνακα 5.3, αξιολογείται η απόδοση των μοντέλων μέσω του Μέσου Απόλυτου Σφάλματος (Mean Absolute Error- MAE) στην εκτίμηση του γεωγραφικού μήκους στους ίδιους ορίζοντες. Το MAE μετρά τη μέση απόλυτη διαφορά μεταξύ των προβλεπόμενων και των πραγματικών τιμών χωρίς να λαμβάνει υπόψη τη σχετική κλίμακα των σφαλμάτων.

Από τα δεδομένα του πίνακα, παρατηρείται πως ο XGBRegressor επιτυγχάνει συνεχώς τις χαμηλότερες τιμές MAE. Ο RandomForestRegressor έχει επίσης καλή απόδοση για τους σύντομους ορίζοντες, αλλά υπερέβη το χρονικό όριο για τα πειράματα που αφορούν τον Ορίζοντα 5 και 10. Οι DecisionTreeRegressor και ExtraTreesRegressor εμφανίζουν συγκρίσιμη απόδοση, με τον τελευταίο όμως να υπερτερεί και σε αυτό το κριτήριο. 0 KNeighborsRegressor παρουσιάζει υψηλότερες τιμές MAE σε σχέση με τα προηγούμενα μοντέλα, ιδίως για τους μεγαλύτερους ορίζοντες, υποδηλώνοντας δυσκολίες στην αποτύπωση των τάσεων του γεωγραφικού μήκους. Ο Ridge Regressor και ο Huber Regressor εμφανίζουν τις μεγαλύτερες τιμές MAE και ιδιαίτερα μεταβαλλόμενα επίπεδα ακρίβειας ανάμεσα στους ορίζοντες. Οι Linear Regressor και Bayesian Ridge διαθέτουν την καλύτερη απόδοση, σχεδόν ταυτόσημη σε όλες τις χρονικές διαστάσεις. Ωστόσο, ο SGDR Regressor αντιμετωπίζει προκλήσεις σε όλους τους ορίζοντες, με σημαντική αύξηση του MAE για τον Ορίζοντα 10, τάση που παρατηρήθηκε και στον υπολογισμό του R^2.

Στον πίνακα 5.4 παρουσιάζονται επίσης τα αποτελέσματα για τις τιμές του Μέσου Απόλυτου Σφάλματος κατά την πρόβλεψη του γεωγραφικού πλάτους. Σε συμφωνία με τα παραπάνω, το εκάστοτε μοντέλο παρουσιάζει παρόμοια συμπεριφορά. Η σταθερή απόδοση των μοντέλων τόσο για γεωγραφικό πλάτος όσο και για γεωγραφικό μήκος μπορεί να αποδοθεί στο ότι τα δεδομένα εμφανίζουν ισχυρά γραμμικά πρότυπα.

Στους πίνακες 5.5, 5.6 αποτυπώνονται τα αποτελέσματα της μέτρησης του Μέσου τετραγωνικού σφάλματος (Mean Squared Error- MSE). Το MSE μετρά τον μέσο όρο των τετραγωνικών διαφορών μεταξύ των προβλεπόμενων και των πραγματικών τιμών και είναι χρήσιμο για τον εντοπισμό ακραίων τιμών με μεγάλα σφάλματα. Είναι προφανές πως τα μοντέλα παρουσιάζουν παρόμοια απόδοση όσον αφορά και το κριτήριο MSE. Η συνεχής τους απόδοση και μπορεί να υποδηλώνει την προσαρμοστικότητά τους σε διάφορες πτυχές των δεδομένων.

Μέθοδος Εκτίμησης	Horizon 1	Horizon 2	Horizon 5	Horizon 10
XGBRegressor	0.0000000723	0.0000000752	0.0000000701	0.0000000574
RandomForestRegressor	0.0000000686	0.0000000560	off limits	off limits
DecisionTreeRegressor	0.0000001490	0.0000001154	0.0000001254	0.0000001495
ExtraTreesRegressor	0.0000000448	0.0000000284	0.0000000302	0.0000000353
KNeighborsRegressor	0.0000002176	0.0000002452	0.0000003233	0.0000003758
Ridge Regressor	0.0000010274	0.0000003802	0.0000004778	0.0000010425
Huber Regressor	0.0000001001	0.0000001089	0.0000002603	0.0000004913
Linear Regressor	0.0000000584	0.0000000181	0.0000000135	0.0000000133
Bayesian Ridge	0.0000000584	0.0000000181	0.0000000135	0.0000000134
SGDR Regressor	0.0000007650	0.0000003061	0.0000004912	0.0000015252

Πίνακας 5.5: Mean Squared Error - Αποτελέσματα για Γεωγραφικό μήκος

Μέθοδος Εκτίμησης	Horizon 1	Horizon 2	Horizon 5	Horizon 10
XGBRegressor	0.0000000868	0.0000000818	0.0000000649	0.0000000618
RandomForestRegressor	0.0000000688	0.0000000492	off limits	off limits
DecisionTreeRegressor	0.0000001320	0.0000001042	0.0000000944	0.0000001051
ExtraTreesRegressor	0.0000000557	0.0000000326	0.0000000283	0.0000000289
KNeighborsRegressor	0.0000001633	0.0000002111	0.0000002947	0.0000003319
Ridge Regressor	0.0000002897	0.0000001862	0.0000004758	0.0000009838
Huber Regressor	0.0000000595	0.0000001294	0.0000004893	0.0000016019
Linear Regressor	0.0000000592	0.0000000081	0.0000000060	0.0000000058
Bayesian Ridge	0.0000000593	0.0000000082	0.0000000060	0.0000000058
SGDR Regressor	0.0000004870	0.0000002971	0.0000004895	0.0000014723

Πίνακας 5.6: Mean Squared Error - Αποτελέσματα για Γεωγραφικό πλάτος

Μέθοδος Εκτίμησης	Horizon 1	Horizon 2	Horizon 5	Horizon 10
XGBRegressor	0.0000295149	0.0000288372	0.0000264458	0.0000256383
RandomForestRegressor	0.0000214668	0.0000185053	off limits	off limits
DecisionTreeRegressor	0.0000288612	0.0000251035	0.0000266105	0.0000285279
ExtraTreesRegressor	0.0000190718	0.0000150451	0.0000163586	0.0000175914
KNeighborsRegressor	0.0000443971	0.0000478453	0.0000519255	0.0000582925
Ridge Regressor	0.0001086764	0.0000677153	0.0000772539	0.0001094290
Huber Regressor	0.0000301136	0.0000320785	0.0000441401	0.0000577944
Linear Regressor	0.0000238933	0.0000079596	0.0000072533	0.0000073681
Bayesian Ridge	0.0000237606	0.0000078792	0.0000071977	0.0000073013
SGDR Regressor	0.0001074612	0.0000703632	0.0000861141	0.0001519653

Πίνακας 5.7: Mean Absolute Percentage Error - Αποτελέσματα για Γεωγραφικό μήκος

Μέθοδος Εκτίμησης	Horizon 1	Horizon 2	Horizon 5	Horizon 10
XGBRegressor	0.0000035018	0.0000034236	0.0000029865	0.0000028700
RandomForestRegressor	0.0000025569	0.0000019868	off limits	off limits
DecisionTreeRegressor	0.0000031896	0.0000026797	0.0000025285	0.0000026602
ExtraTreesRegressor	0.0000023277	0.0000016013	0.0000014655	0.0000015214
KNeighborsRegressor	0.0000039926	0.0000047220	0.0000049339	0.0000046571
Ridge Regressor	0.0000091158	0.0000070023	0.0000093346	0.0000128200
Huber Regressor	0.0000029927	0.0000044326	0.0000085616	0.0000154600
Linear Regressor	0.0000029549	0.0000006397	0.0000005346	0.0000005376
Bayesian Ridge	0.0000029557	0.0000006338	0.0000005338	0.0000005372
SGDR Regressor	0.0000074612	0.0000692332	0.0000854532	0.0001452768

Πίνακας 5.8: Mean Absolute Percentage Error - Αποτελέσματα για Γεωγραφικό πλάτος

Οι πίνακες 5.7, 5.8 παρουσιάζουν τα αποτελέσματα για το Μέσο απόλυτο ποσοστό σφάλματος (Mean Absolute Percentage Error- Mape) στην εκτίμηση του γεωγραφικού μήκους και πλάτους. Το MAPE είναι το ποσοστό που μετρά τη μέση απόλυτη ποσοστιαία διαφορά μεταξύ προβλεπόμενων και πραγματικών τιμών. Πιο απλά, μετρά τον μέσο όρο του μεγέθους των σφαλμάτων ως ποσοστό των πραγματικών τιμών.

Πέραν της παρόμοιας απόδοσης των μοντέλων και σε αυτό το κριτήριο, γίνεται αντιληπτό πως το MAPE συμπεριφέρεται παραλληλικά με τις τάσεις των MAE και MSE. Η κύρια διαφορά μεταξύ MAPE, MSE και MAE είναι ότι το MAPE εκφράζει τα σφάλματα ως ποσοστό των πραγματικών τιμών. Αντίθετα, τα MAE και MSE βρίσκονται στις ίδιες μονάδες με τις μεταβλητές προς πρόβλεψη. Ωστόσο, επειδή το MAPE βασίζεται σε απόλυτα ποσοστά, εξακολουθεί να λαμβάνει υπόψη τη σχετική κλίμακα σφαλμάτων. Μεγαλύτερα σφάλματα, σε ποσοστιαίες τιμές, θα συνεισφέρουν περισσότερο στο MAPE, όπως π.χ. μεγαλύτερα σφάλματα συνεισφέρουν περισσότερο στο MAE.

Μέθοδος Εκτίμησης	Horizon 1	Horizon 2	Horizon 5	Horizon 10
XGBRegressor	0.0000448051	0.0000409333	0.0000368004	0.0000379654
RandomForestRegressor	0.0000055817	0.0000056670	off limits	off limits
DecisionTreeRegressor	0.0000065000	0.0000067000	0.0000070000	0.0000071813
ExtraTreesRegressor	0.0000067029	0.0000058480	0.0000059058	0.0000060498
KNeighborsRegressor	0.0000074070	0.0000082730	0.0000096210	0.0000120830
Ridge Regressor	0.0000795933	0.0000434718	0.0000270602	0.0000261010
Huber Regressor	0.0000050021	0.0000058835	0.0000078211	0.0000114779
Linear Regressor	0.0000065367	0.0000046676	0.0000054574	0.0000056551
Bayesian Ridge	0.0000049608	0.0000045359	0.0000054535	0.0000056682
SGDR Regressor	0.0001464147	0.0001493787	0.0001407592	0.0002649395

Πίνακας 5.9: Median Absolute Error - Αποτελέσματα για Γεωγραφικό μήκος

Μέθοδος Εκτίμησης	Horizon 1	Horizon 2	Horizon 5	Horizon 10
XGBRegressor	0.0000517029	0.0000546210	0.0000533274	0.0000499844
RandomForestRegressor	0.0000069470	0.0000067038	off limits	off limits
DecisionTreeRegressor	0.0000080000	0.0000080000	0.0000089413	0.0000084519
ExtraTreesRegressor	0.0000077032	0.0000066774	0.0000066712	0.0000066688
KNeighborsRegressor	0.0000085800	0.0000089100	0.0000097200	0.0000108300
Ridge Regressor	0.0002997184	0.0001558805	0.0000712328	0.0000504294
Huber Regressor	0.0000089560	0.0000102892	0.0000102543	0.0000122951
Linear Regressor	0.0000068927	0.0000069800	0.0000061659	0.0000064454
Bayesian Ridge	0.0000074020	0.0000068884	0.0000061397	0.0000064296
SGDR Regressor	0.0002425369	0.0001378544	0.0001332001	0.0002545223

Πίνακας 5.10: Median Absolute Error - Αποτελέσματα για Γεωγραφικό πλάτος

Στους πίνακες 5.9, 5.10 παρουσιάζονται τα αποτελέσματα για το Διάμεσο Απόλυτο Σφάλμα (Median Absolute Error- MedAE) κατά την πρόβλεψη του γεωγραφικού μήκους και γεωγραφικού πλάτους. Το MedAE υπολογίζει τη μέση απόλυτη διαφορά μεταξύ των πραγματικών τιμών και των προβλέψεων που παρέχει το μοντέλο. Ένα χαμηλό MedAE υποδεικνύει ότι το μοντέλο έχει καλή ακρίβεια στις προβλέψεις του, καθώς οι απόλυτες αποκλίσεις μεταξύ των προβλέψεων και των πραγματικών τιμών είναι μικρές. Αντίστροφα, ένα υψηλό MedAE υποδηλώνει πως το μοντέλο αδυνατεί να παράγει ακριβείς προβλέψεις, καθώς οι απόλυτες διαφορές είναι μεγαλύτερες.

Παρατηρείται πως και σε αυτήν την περίπτωση, τα μοντέλα συμπεριφέρονται με παρόμοιο τρόπο και τα αποτελέσματα των μετρήσεων είναι συναφή με τα αποτελέσματα των MAE, MSE και Mape. Αυτό αιτιολογείται καθώς το MedAE φέρει ομοιότητες με αυτές των παραπάνω κριτηρίων, αλλά η έμφασή του στην διάμεσο και η ανθεκτικότητά του στις ακραίες τιμές το καθιστούν ένα διακριτικό μέτρο για την αξιολόγηση της απόδοσης του μοντέλου.

Μέθοδος Εκτίμησης	Horizon 1	Horizon 2	Horizon 5	Horizon 10
XGBRegressor	0.0046000162	0.0046448389	0.0045008341	0.0038890520
RandomForestRegressor	0.0053255500	0.0059680610	off limits	off limits
DecisionTreeRegressor	0.0099815000	0.0068450000	0.0062850000	0.0068450000
ExtraTreesRegressor	0.0040399340	0.0038297340	0.0032825840	0.0029432500
KNeighborsRegressor	0.0039501640	0.0039416980	0.0042198340	0.0044067340
Ridge Regressor	0.0048079683	0.0033465514	0.0038830092	0.0042824799
Huber Regressor	0.0042558663	0.0041532055	0.0042211681	0.0047051769
Linear Regressor	0.0041357947	0.0040709229	0.0040459768	0.0040310704
Bayesian Ridge	0.0041213527	0.0040671449	0.0040683144	0.0040543363
SGDR Regressor	0.0043204393	0.0036696515	0.0042063541	0.0052591019

Πίνακας 5.11: Max Error - Αποτελέσματα για Γεωγραφικό μήκος

Μέθοδος Εκτίμησης	Horizon 1	Horizon 2	Horizon 5	Horizon 10
XGBRegressor	0.0029566926	0.0030482454	0.0029643220	0.0029986543
RandomForestRegressor	0.0029679600	0.0029173300	off limits	off limits
DecisionTreeRegressor	0.0058890000	0.0047330000	0.0055580000	0.0072170000
ExtraTreesRegressor	0.0029941600	0.0027773200	0.0029380200	0.0028557400
KNeighborsRegressor	0.0029781800	0.0029743000	0.0032447200	0.0039945600
Ridge Regressor	0.0027195534	0.0030213215	0.0032283061	0.0033910308
Huber Regressor	0.0029653234	0.0030835979	0.0031335668	0.0035205839
Linear Regressor	0.0030603799	0.0028213229	0.0029594811	0.0029552838
Bayesian Ridge	0.0030179796	0.0028198094	0.0029565173	0.0029515710
SGDR Regressor	0.0033302156	0.0034785663	0.0041025361	0.0050741125

Πίνακας 5.12: Max Error - Αποτελέσματα για Γεωγραφικό πλάτος

Οι δυο τελευταίοι πίνακες απεικονίζουν τα αποτελέσματα του Μέγιστου Σφάλματος (Max Error- MaxE) κατά την διεξαγωγή των ίδιων πειραμάτων, όπως περιγράφηκαν παραπάνω. Το MaxE είναι ένας όρος που χρησιμοποιείται για να μετρήσει το μέγιστο απόλυτο σφάλμα που παρατηρείται σε προβλέψεις που πραγματοποιεί ένα μοντέλο. Αντιπροσωπεύει, δηλαδή, τη μεγαλύτερη απόκλιση μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών σε ένα σύνολο δεδομένων, αναδεικνύοντας το μέγιστο σφάλμα που το μοντέλο έχει καταγράψει.

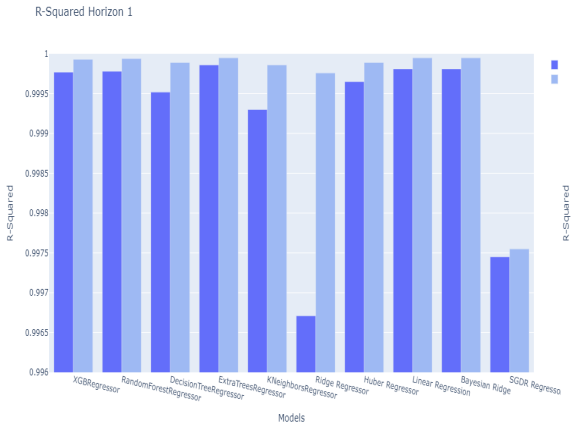
Τα μοντέλα εξακολουθούν να εμφανίζουν συνεπή συμπεριφορά και στις καταγεγραμμένες τιμές του Max Error.

5.2 Γραφική Αναπαράσταση Αποτελεσμάτων

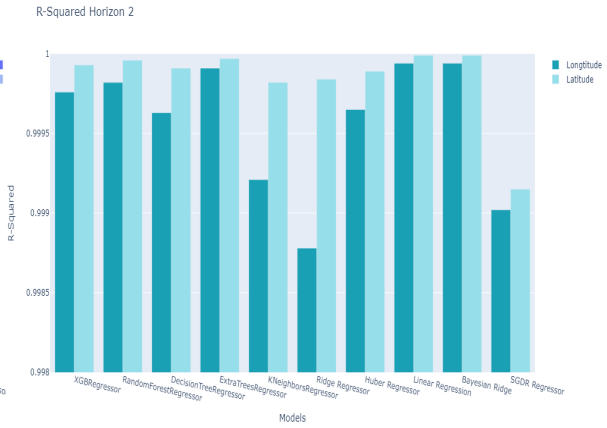
Μετά τη λεπτομερή ανάλυση των αποτελεσμάτων που παρουσιάζονται στους πίνακες 5.3 - 5.11, προχωρούμε στην παρουσίασή τους σε μορφή διαγραμμάτων (bar charts) για να διευκολύνουμε την γρήγορη και εύκολη σύγκριση των διαφόρων μεθόδων.

Κάθε σχήμα αντιπροσωπεύει ένα συγκεκριμένο κριτήριο απόδοσης και αποτελείται από τέσσερα διαγράμματα που αναπαριστούν τους τέσσερις ορίζοντες ανάλυσης αντίστοιχα. Αυτή η γραφική αναπαράσταση επιτρέπει την γρήγορη εξέταση της απόδοση κάθε μεθόδου σε σχέση με τους διάφορους ορίζοντες, παρέχοντας μια πλήρη εικόνα της απόδοσης του συστήματος. Με αυτόν τον τρόπο, μπορούμε εύκολα να αναγνωρίσουμε ποιες μέθοδοι έχουν την καλύτερη απόδοση για κάθε κριτήριο αξιολόγησης και πως η απόδοση μεταβάλλεται ανάλογα με τον ορίζοντα της ανάλυσης.

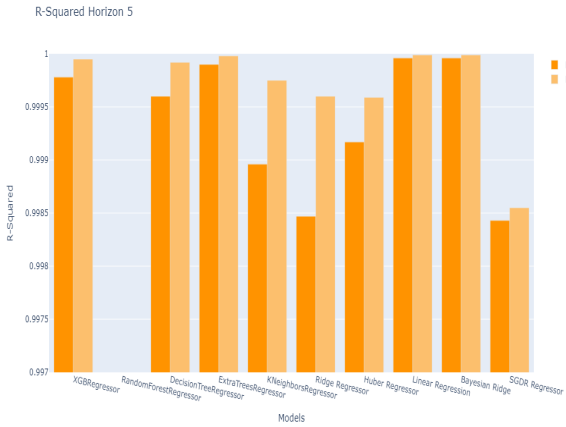
Τα διαγράμματα αυτού του είδους αποτελούν ένα ισχυρό εργαλείο για την ανάλυση και την παρουσίαση ποσοτικών αποτελεσμάτων. Η σύγχρονη επιστήμη των δεδομένων(data science) δίνει ιδιαίτερη έμφαση στην αποτελεσματική οπτικοποίηση τους καθώς αυτή μπορεί να απλοποιεί πολύπλοκες πληροφορίες αλλά και να αποκαλύπτει κρυφά πρότυπα και τάσεις εντός των συνόλων δεδομένων.



(a) `Horizon 1`



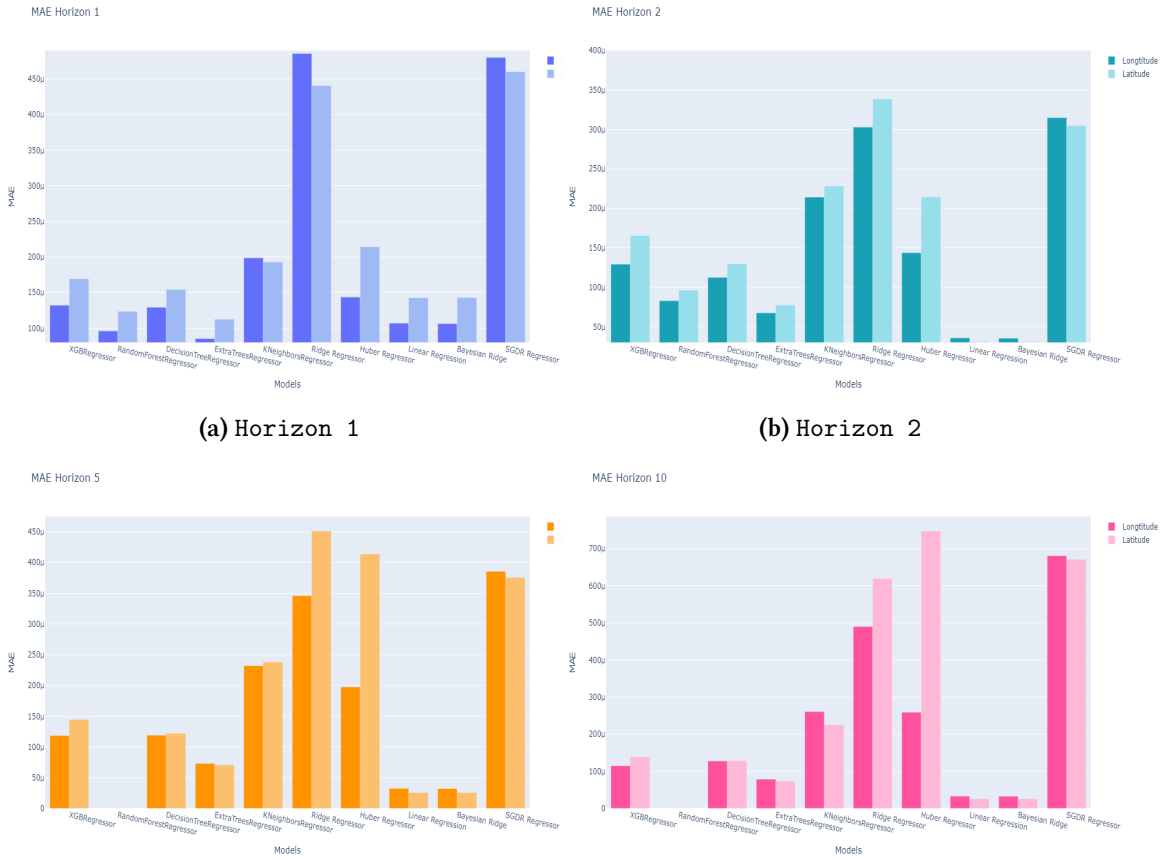
(b) `Horizon 2`



(c) `Horizon 5`

(d) `Horizon 10`

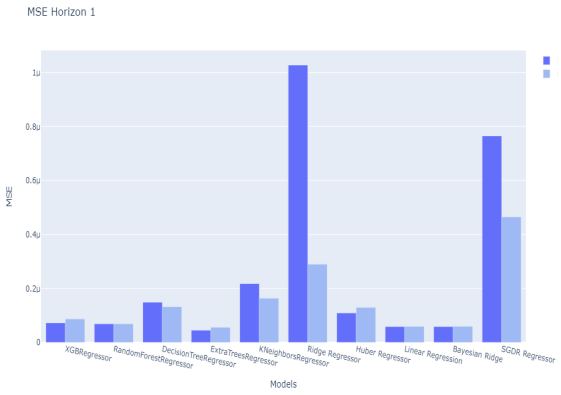
Σχήμα 5.1: Σύγκριση των τιμών R^2 για τους Ορίζοντες 1, 2, 5, 10



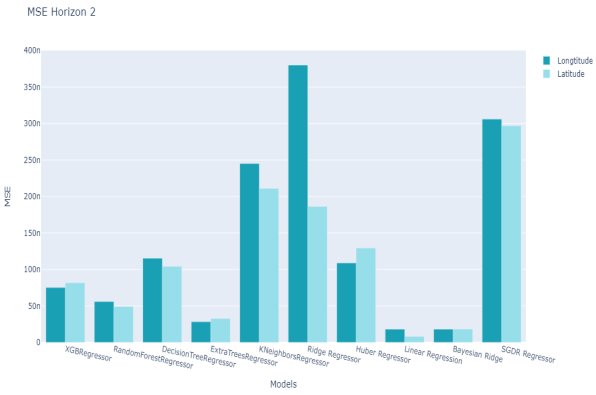
(a) `Horizon 1` **(b)** `Horizon 2`

(c) `Horizon 5` **(d)** `Horizon 10`

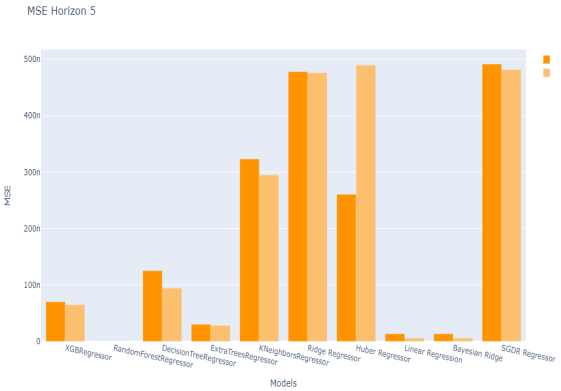
Σχήμα 5.2: Σύγκριση των τιμών MAE για τους Ορίζοντες 1, 2, 5, 10



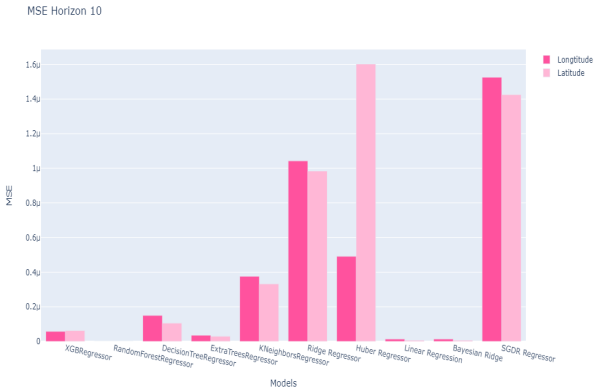
(a) `Horizon 1`



(b) `Horizon 2`

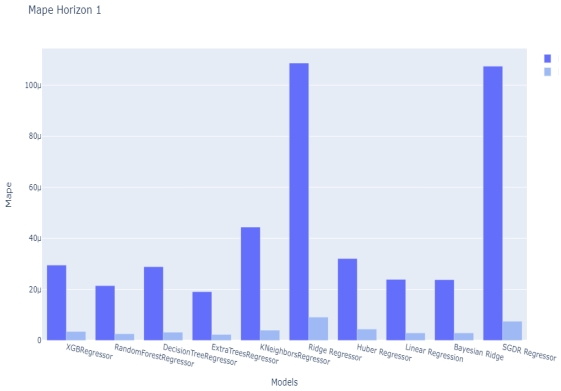


(c) `Horizon 5`

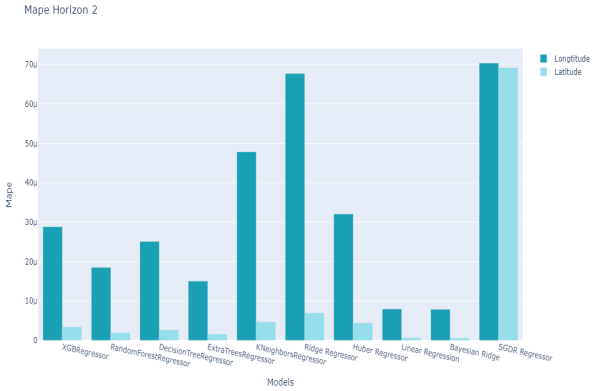


(d) `Horizon 10`

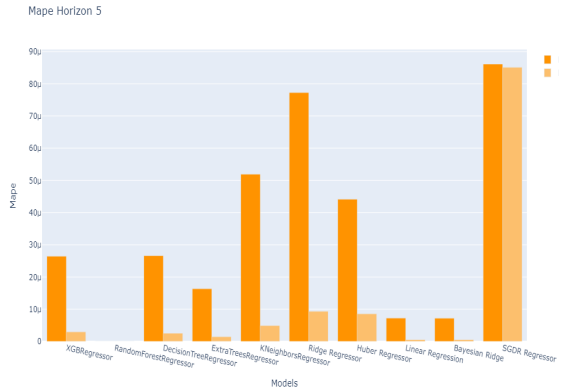
Σχήμα 5.3: Σύγκριση των τιμών MSE για τους Ορίζοντες 1, 2, 5, 10



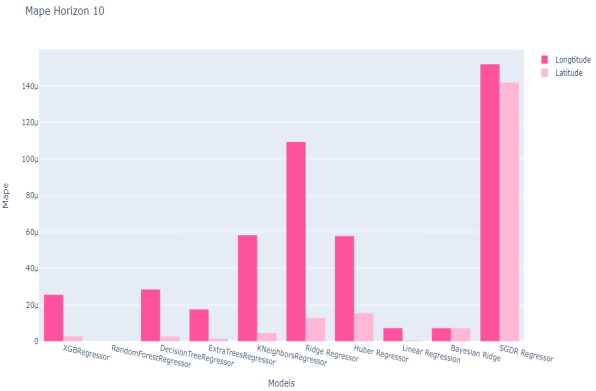
(a) `Horizon 1`



(b) `Horizon 2`

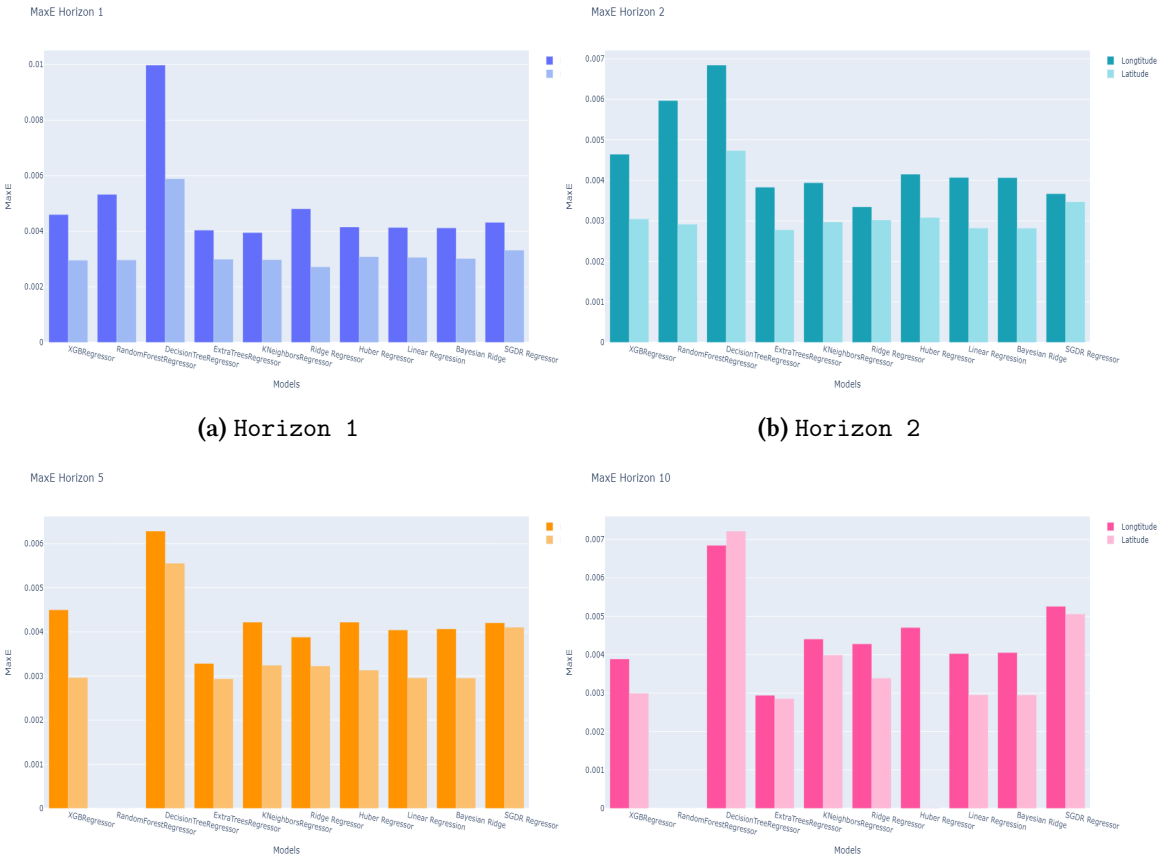


(c) `Horizon 5`



(d) `Horizon 10`

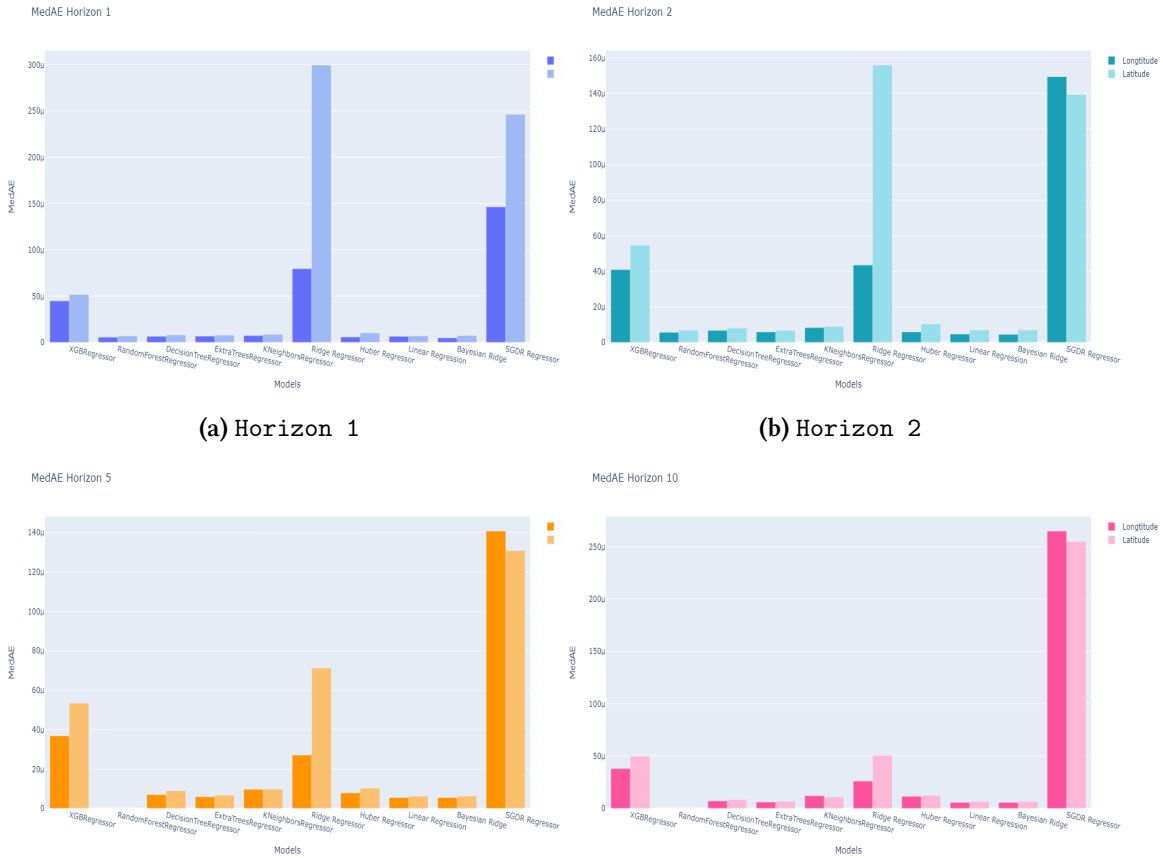
Σχήμα 5.4: Σύγκριση των τιμών Mape για τους Ορίζοντες 1, 2, 5, 10



(a) `Horizon 1` **(b)** `Horizon 2`

(c) `Horizon 5` **(d)** `Horizon 10`

Σχήμα 5.5: Σύγκριση των τιμών Max Error για τους Ορίζοντες 1, 2, 5, 10



(a) `Horizon 1` **(b)** `Horizon 2`

(c) `Horizon 5` **(d)** `Horizon 10`

Σχήμα 5.6: Σύγκριση των τιμών Median Absolute Error για τους Ορίζοντες 1, 2, 5, 10

Κεφάλαιο: Συμπεράσματα & Μελλοντικές Επεκτάσεις

Από τα παραπάνω αποτελέσματα, προκύπτει ότι τόσο τα γραμμικά μοντέλα, όπως το Linear Regressor και το Bayesian Ridge αλλά και το XGBRegressor εμφανίζουν συνεχώς υψηλή απόδοση σε όλα τα κριτήρια για τις γεωγραφικές διαστάσεις, καθιστώντας τα ικανά να ανταποκρίνονται σε διάφορους ορίζοντες πρόβλεψης και να παράγουν προβλέψεις υψηλής ακρίβειας. Αντίθετα, ο RandomForestRegressor είναι αποτελεσματικός για μικρούς ορίζοντες, αλλά αντιμετωπίζει προβλήματα με αυξημένο χρόνο πρόβλεψης. Η μειωμένη απόδοση ορισμένων μοντέλων με μεγαλύτερους ορίζοντες πρόβλεψης μπορεί να οφείλεται σε προβλήματα υπερεκπαίδευσης (overfitting), υπολογιστικής πολυπλοκότητας ή και ποιότητας δεδομένων.

Το συνολικό συμπέρασμα είναι ότι η επιλογή του μοντέλου εξαρτάται από τη φύση των δεδομένων και την ανάγκη για ακρίβεια ή ταχύτητα πρόβλεψης. Η φύση του παρόντος πειράματος σχετίζεται με δεδομένα χαμηλής διάστασης (low dimensional) αλλά υψηλής συχνότητας (high frequency). Με βάση την υπολογιστική ισχύ που είναι διαθέσιμη σήμερα, παραδοσιακές στατιστικές μέθοδοι όπως τα μοντέλα γραμμικής παλινδρόμησης μπορούν να παράξουν προβλέψεις υψηλής ακρίβειας χωρίς την πολύπλοκη δομή των υποδειγμάτων μηχανικής μάθησης.

Η επέκταση του πειράματος με χρήση δεδομένων υψηλής διάστασης (high-dimensional) αποτελεί ένα ακόμα σημαντικό βήμα στον τομέα της μηχανικής μάθησης και της στατιστικής. Σε αντίθεση με τα δεδομένα χαμηλής διάστασης που χρησιμοποιήθηκαν, τα υψηλής διάστασης δεδομένα αντιμετωπίζουν μια σειρά προκλήσεων που σχετίζονται με την αύξηση του υπολογιστικού φόρτου και τη δυσκολία στην εξαγωγή σημαντικών πληροφοριών. Για την αντιμετώπιση του προβλήματος των υψηλών διαστάσεων, υπάρχουν αρκετές τεχνικές που μπορούν να εφαρμοστούν:

- Επιλογή Χαρακτηριστικών (Feature Selection): Επιλέγοντας μόνο τα σημαντικότερα χαρακτηριστικά, μπορεί να μειωθεί η διάσταση των δεδομένων. Αυτό μπορεί να γίνει βάσει της συσχέτισης με τη μεταβλητή εξόδου ή με άλλα κριτήρια.

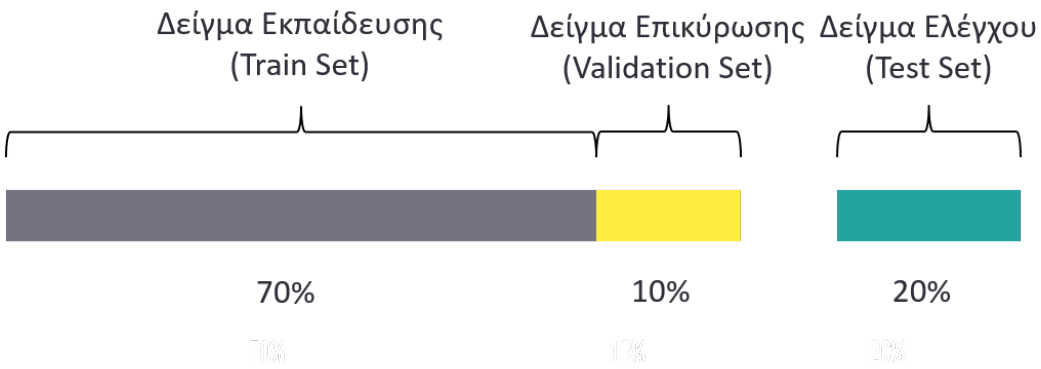
- Μείωση Διάστασης (Dimensionality Reduction): Μεθόδοι όπως η Ανάλυση σε Κύριες Συνιστώσες (PCA) ή η Ανάλυση σε Υποχώρους (Manifold Learning) μειώνουν τον αριθμό των διαστάσεων διατηρώντας τα σημαντικά χαρακτηριστικά.

- Ρύθμιση Υπερ-Παραμέτρων (Hyperparameter Tuning): Οι αλγόριθμοι ML που λειτουρ-

γούν σε υψηλές διαστάσεις συνήθως έχουν πολλές υπερ-παραμέτρους. Η ρύθμισή τους με κατάλληλο τρόπο είναι σημαντική για την επίτευξη καλών επιδόσεων.

Συνολικά, η ανάλυση των υψηλών διαστάσεων δεδομένων είναι σημαντική σήμερα για πολλούς επιστημονικούς τομείς, όπως επίσης και για τη ναυτιλία. Στη ναυτιλία, υψηλές διαστάσεις μπορεί να αναφέρονται σε μεγάλα και πολύπλοκα σύνολα δεδομένων (big data) αισθητήρων που παρακολουθούν την κατάσταση των πλοίων ή των φορτίων, όπως η θέση, η ταχύτητα, η κατεύθυνση, η θερμοκρασία και άλλες παράμετροι, οι οποίες, αν αναλυθούν με τον κατάλληλο τρόπο, μπορούν να συμβάλουν στην αποδοτικότητα και την ασφάλεια των ναυτιλιακών διαδικασιών.

Κεφάλαιο: Appendix



Σχήμα 7.1: `Split Dataset`

Στη μηχανική μάθηση, τα δεδομένα χωρίζονται συνήθως σε τρία κύρια σύνολα: το σύνολο εκπαίδευσης(training Set), το σύνολο επικύρωσης(validation Set) και το σύνολο ελέγχου(test Set). Το σύνολο εκπαίδευσης χρησιμοποιείται για την εκπαίδευση του μοντέλου μηχανικής μάθησης και αποτελεί τη βάση ώστε το μοντέλο να μάθει τα πρότυπα και τις σχέσεις των δεδομένων. Το σύνολο επικύρωσης είναι ξεχωριστό από το σύνολο εκπαίδευσης και χρησιμοποιείται για τη βελτιστοποίηση των ρυθμίσεων του μοντέλου και για την πρόληψη της υπερπροσαρμογής(overfitting) κατά τη διάρκεια της εκπαίδευσης. Οι ρυθμίσεις αυτές ονομάζονται υπερπαράμετροι (hyperparameters), και η διαδικασία επικύρωσης βοηθάει στη βελτιστοποίησή τους. Διάφορες τιμές υπερπαραμέτρων μπορούν να οδηγήσουν σε καλύτερη ή χειρότερη απόδοση των μοντέλων. Τέλος. το σύνολο ελέγχου περιέχει δεδομένα που δεν έχουν τροφοδοτηθεί προηγουμένως στο μοντέλο και χρησιμοποιούνται για να παράξουν μια αντικειμενική αξιολόγηση της απόδοσης του τελικού μοντέλου.

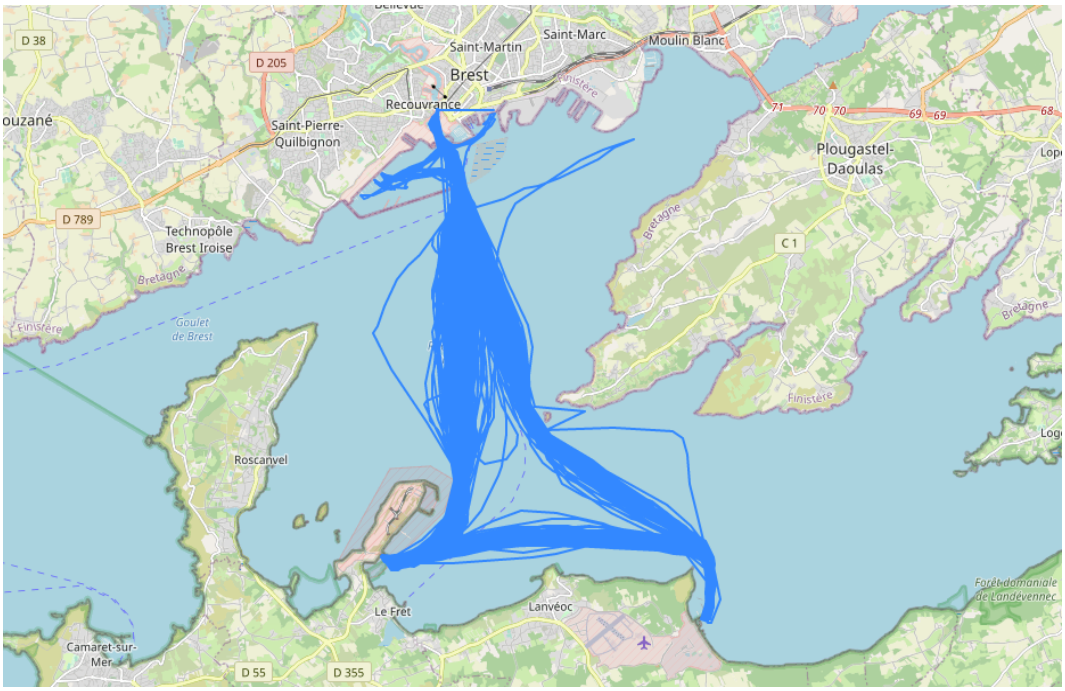
Οι τυπικές αναλογίες διαίρεσης των δεδομένων μπορεί να ποικίλλουν, αλλά συνήθως τα δεδομένα διαχωρίζονται ως εξής: το 70-80% για εκπαίδευση, 10-15% για επικύρωση και το υπόλοιπο 10-15% για ελέγχο.

	Y	t+1	t+2	t+3	t+4	t+5	t+6
Train	0,196	0,196	0,196	0,196	0,196	0,196	0,196
	-0,706	-0,706	-0,706	-0,706	-0,706	-0,706	-0,706
	1,710	1,710	1,710	1,710	1,710	1,710	1,710
	1,149	1,149	1,149	1,149	1,149	1,149	1,149
	0,747	0,747	0,747	0,747	0,747	0,747	0,747
	-1,369	-1,369	-1,369	-1,369	-1,369	-1,369	-1,369
Test	-0,629	Forecast	-0,629	-0,629	-0,629	-0,629	-0,629
	-1,184		Forecast	-1,184	-1,184	-1,184	-1,184
	-0,863			Forecast	-0,863	-0,863	-0,863
	-0,677				Forecast	-0,677	-0,677
	-1,753					Forecast	-1,753
	-2,411						Forecast

Σχήμα 7.2: Αναδρομική (Recursive) Μέθοδος

Η αναδρομική πρόβλεψη είναι μια μέθοδος που χρησιμοποιείται στην πρόβλεψη χρονοσειρών, όπου οι προβλέψεις γίνονται αναδρομικά, ενημερώνοντας τις εισόδους του μοντέλου με τις πραγματικές τιμές από τα προηγούμενα χρονικά βήματα. Η αναδρομική πρόβλεψη χρησιμοποιείται συχνά για να προσομοιώσει σενάρια πραγματικού χρόνου, όπου το μοντέλο πρέπει να πραγματοποιεί προβλέψεις καθώς γίνονται διαθέσιμα νέα δεδομένα. Αυτό επιτρέπει έναν συνεχή κύκλο ανατροφοδότησης, όπου οι προβλέψεις του μοντέλου βελτιώνονται καθώς συλλέγονται περισσότερα πραγματικά δεδομένα.

Τα παρακάτω στιγμιότυπα απεικονίζουν τις διαδρομές του πλοίου, που χρησιμοποιείται για πρόβλεψη, στο δείγμα εκπαίδευσης και στο δείγμα ελέγχου αντιστοίχως:



Σχήμα 7.3: Διαδρομή πλοίου- Δείγμα εκπαίδευσης

Σχήμα 7.4: Διαδρομή πλοίου- Δείγμα ελέγχου

Βιβλιογραφία

[1] M. Grote, N. Mazurek, C. Gräbsch, J. Zeilinger, S. Le Floch, D.-S. Wahrendorf, and T. Höfer, "Dry bulk cargo shipping—an overlooked threat to the marine environment?" *Marine pollution bulletin*, vol. 110, no. 1, pp. 511–519, 2016.

[2] T.-P. Liang and Y.-H. Liu, "Research landscape of business intelligence and big data analytics: A bibliometrics study," *Expert Systems with Applications*, vol. 111, pp. 2–10, 2018.

[3] Z. H. Munim, M. Dushenko, V. J. Jimenez, M. H. Shakil, and M. Imset, "Big data and artificial intelligence in the maritime industry: a bibliometric review and future research directions," *Maritime Policy and Management*, vol. 47, no. 5, pp. 577–597, 2020. [Online]. Available: https://doi.org/10.1080/03088839.2020.1788731

[4] G. Ç. Ceyhun, "Recent developments of artificial intelligence in business logistics: A maritime industry case," *Digital Business Strategies in Blockchain Ecosystems: Transformational Design and Future of Global Business*, pp. 343–353, 2020.

[5] U. of Münster, "Automatic Identification System (AIS) data based Ship-Supply Forecasting," no. November, pp. 3–24, 2019. [Online]. Available: https://www.econstor.eu/handle/10419/209386

[6] K. F. Yuen, G. Xu, and J. S. L. Lam, "Special issue on 'artificial intelligence & big data in shipping'," pp. 575–576, 2020.

[7] K. Kulkarni, F. Goerlandt, J. Li, O. V. Banda, and P. Kujala, "Preventing shipping accidents: Past, present, and future of waterway risk management with Baltic Sea focus," *Safety Science*, vol. 129, no. May, p. 104798, 2020. [Online]. Available: https://doi.org/10.1016/j.ssci.2020.104798

[8] EASO, "Consolidated Annual Activity Report 2018," *European Asylum Support Office*, no. June, 2017.

[9] EMSA, "Annual Overview of Marine Casualties and Incidents 2018 - EMSA - European Maritime Safety Agency," *European Maritime Safety Agency*, no. October, p. 175, 2018. [Online]. Available: http://www.emsa.europa.eu/news-a-press-centre/external-news/item/3406-annual-overview-of-marine-casualties-and-incidents-2018.html

[10] O. Vinogradov and O. Morozova, "Aspects of neural networks use for predicting emergency situations," *Civil Security Technology*, vol. 18, no. 1, pp. 52–59, 2021.

[11] P. Antao and C. G. Soares, "Causal factors in accidents of high-speed craft and conventional ocean-going vessels," *Reliability Engineering & System Safety*, vol. 93, no. 9, pp. 1292–1304, 2008.

[12] E. Kulbiej and P. Wołejsza, "Naval artificial intelligence," in *12th International Conference on Marine Navigation and Safety of Sea Transportation TransNav*, 2017, pp. 21–23.

[13] E. Gaponenko, D. Malyshev, S. Y. Misyurin, Y. A. Semenov, and E. Semenova, "On a possibility of using vibrational displacement in artificial intelligence based ship accident predictive models," *Procedia Computer Science*, vol. 213, pp. 696–702, 2022.

[14] P. Chen, Y. Huang, J. Mou, and P. Van Gelder, "Probabilistic risk analysis for ship-ship collision: State-of-the-art," *Safety science*, vol. 117, pp. 108–122, 2019.

[15] Z. Wang and J. Yin, "Risk assessment of inland waterborne transportation using data mining," *Maritime Policy & Management*, vol. 47, no. 5, pp. 633–648, 2020.

[16] Y. Ma, Y. Zhao, Y. Wang, L. Gan, and Y. Zheng, "Collision-avoidance under colregs for unmanned surface vehicles via deep reinforcement learning," *Maritime Policy & Management*, vol. 47, no. 5, pp. 665–686, 2020.

[17] Z. Yan, Y. Xiao, L. Cheng, R. He, X. Ruan, X. Zhou, M. Li, and R. Bin, "Exploring ais data for intelligent maritime routes extraction," *Applied Ocean Research*, vol. 101, p. 102271, 2020.

[18] W. Xing, J. Wang, K. Zhou, H. Li, Y. Li, and Z. Yang, "A hierarchical methodology for vessel traffic flow prediction using bayesian tensor decomposition and similarity grouping," *Ocean Engineering*, vol. 286, p. 115687, 2023.

[19] J. Chen, H. Chen, Y. Zhao, and X. Li, "Fb-bigru: a deep learning model for ais-based vessel trajectory curve fitting and analysis," *Ocean Engineering*, vol. 266, p. 112898, 2022.

[20] E. Chondrodima, P. Mandalis, N. Pelekis, and Y. Theodoridis, "Machine Learning Models for Vessel Route Forecasting: An Experimental Comparison," *Proceedings - IEEE International Conference on Mobile Data Management*, vol. 2022-June, no. Mdm, pp. 262–269, 2022.

[21] Z. Dong, "Prediction of ship traffic flow based on wavelet decomposition and lstm," in *2022 7th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*. IEEE, 2022, pp. 88–93.

[22] H. Li, H. Jiao, and Z. Yang, "Ais data-driven ship trajectory prediction modelling and analysis based on machine learning and deep learning methods," *Transportation Research Part E: Logistics and Transportation Review*, vol. 175, p. 103152, 2023.

[23] A. Coraddu, L. Oneto, F. Baldi, and D. Anguita, "Vessels fuel consumption forecast and trim optimisation: A data analytics perspective," *Ocean Engineering*, vol. 130, no. September 2015, pp. 351–370, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.oceaneng.2016.11.058

[24] T. Anan, H. Higuchi, and N. Hamada, "New artificial intelligence technology improving fuel efficiency and reducing co2 emissions of ships through use of operational big data," *Fujitsu Sci. Tech. J*, vol. 53, no. 6, pp. 23–28, 2017.

[25] T. Uyanık, Ç. Karatuğ, and Y. Arslanoğlu, "Machine learning approach to ship fuel consumption: A case of container vessel," *Transportation Research Part D: Transport and Environment*, vol. 84, no. May, 2020.

[26] T. Uyanik, Y. Arslanoglu, and O. Kalenderli, "Ship fuel consumption prediction with machine learning," in *Proceedings of the 4th International Mediterranean Science and Engineering Congress, Antalya, Turkey*, 2019, pp. 25–27.

[27] C. Gkerekos, I. Lazakis, and G. Theotokatos, "Machine learning models for predicting ship main engine fuel oil consumption: A comparative study," *Ocean Engineering*, vol. 188, p. 106282, 2019.

[28] E. Tzannatos, "A decision support system for the promotion of security in shipping," *Disaster Prevention and Management: An International Journal*, vol. 12, no. 3, pp. 222–229, 2003.

[29] H. Wan, Y. Xiao, W. Xu, S. Fu, and P. Gao, "A cognitive approach for identification of ship illegal behaviors by using knowledge graph," in *2023 7th International Conference on Transportation Information and Safety (ICTIS)*. IEEE, 2023, pp. 732–737.

[30] K. D. Schwehr and P. A. McGillivary, "Marine ship automatic identification system (ais) for enhanced coastal security capabilities: an oil spill tracking application," in *OCEANS 2007*. IEEE, 2007, pp. 1–9.

[31] M. Abouheaf, S. Qu, W. Gueaieb, R. Abielmona, and M. Harb, "Responding to illegal activities along the canadian coastlines using reinforcement learning," *IEEE Instrumentation & Measurement Magazine*, vol. 24, no. 2, pp. 118–126, 2021.

[32] C. Zhang, J. Bin, W. Wang, X. Peng, R. Wang, R. Halldearn, and Z. Liu, "Ais data driven general vessel destination prediction: A random forest based approach," *Transportation Research Part C: Emerging Technologies*, vol. 118, p. 102729, 2020.

[33] O. Bodunov, F. Schmidt, A. Martin, A. Brito, and C. Fetzer, "Real-time destination and eta prediction for maritime traffic," in *Proceedings of the 12th ACM international conference on distributed and event-based systems*, 2018, pp. 198–201.

[34] R. J. Hyndman, "George athanasopoulos," *Forecasting: Principles and Practice. Monash University, Australia*, vol. 2, p. 23, 2018.

[35] E. Tu, G. Zhang, L. Rachmawati, E. Rajabally, and G.-B. Huang, "Exploiting ais data for intelligent maritime navigation: A comprehensive survey from data to methodology," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1559–1582, 2017.

[36] Z. Yan, L. Cheng, R. He, and H. Yang, "Extracting ship stopping information from ais data," *Ocean Engineering*, vol. 250, p. 111004, 2022.

[37] L. Cazzanti and G. Pallotta, "Mining maritime vessel traffic: Promises, challenges, techniques," *MTS/IEEE OCEANS 2015 - Genova: Discovering Sustainable Ocean Energy for a New World*, 2015.

[38] [Online]. Available: https://www.postgresql.org/about/

[39] [Online]. Available: https://postgis.net/docs/manual-1.5/ch08.html

[40] [Online]. Available: https://www.pgadmin.org/

[41] [Online]. Available: https://jupyter-notebook.readthedocs.io/en/latest/

[42] C. Ray, R. Dréo, E. Camossi, A.-L. Jousselme, and C. Iphar, "Heterogeneous integrated dataset for maritime intelligence, surveillance, and reconnaissance," *Data in brief*, vol. 25, p. 104141, 2019.

[43] A. Artikis, *Guide to Maritime Informatics*, 2021.

[44] [Online]. Available: https://chorochronos.datastories.org/?q=node/9

[45] [Online]. Available: https://zenodo.org/record/2563256

[46] [Online]. Available: https://zenodo.org/record/1167595

[47] [Online]. Available: https://www.marineregions.org/downloads.php

[48] G. Kumar, S. Basri, A. A. Imam, S. A. Khowaja, L. F. Capretz, and A. O. Balogun, "Data harmonization for heterogeneous datasets: A systematic literature review," *Applied Sciences*, vol. 11, no. 17, p. 8275, 2021.

[49] [Online]. Available: https://github.com/faypanou/Thesis.git

[50] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Statistical learning," in *An Introduction to Statistical Learning: with Applications in Python*. Springer, 2023, pp. 15–67.