Πανεπιστήμιο Πειραιώς

Τμήμα Ψηφιακών Συστημάτων

Π.Μ.Σ. Πληροφοριακά Συστήματα & Υπηρεσίες

Πρόγραμμα Μεταπτυχιακών Σπουδών: Big Data & Analytics

Θέμα Διπλωματικής

Πρόβλεψη Ταχύτητας Πλοίου με Χρήση Χρονοσειρών

Ζερβούδης Στέφανος – Α.Μ. : ΜΕ2009

Επιβλέπων Καθηγητής: Φιλιππάκης Μιχαήλ

Πειραιάς, 2021/22

Περίληψη

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια της φοίτησής μου στο Πανεπιστήμιο Πειραιώς στο Τμήμα Ψηφιακών Συστημάτων, και έχει ως αντικείμενο την πρόβλεψη της ταχύτητας ενός πλοίου βάσει χρονοσειρών. Πιο συγκεκριμένα τα δεδομένα τα οποία χρησιμοποιήθηκαν προήλθαν από τον σύνδεσμο: Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance | Zenodo [16], τα οποία περιλαμβάνουν δεδομένα πλοίων όπου έχουν συλλεχθεί από τον δυτικό κόλπο της Γαλλίας στην περιοχή του Βισκαϊκού κόλπου, από τους δέκτες του AIS για χρονικό διάστημα έξι μηνών, από 01/10/2015 έως 31/03/2016. Τα προβλεπτικά μοντέλα τα οποία χρησιμοποιήθηκαν ήταν το ARIMA και VAR. Για τις προβλέψεις των μοντέλων κατασκευάστηκε ένα σύνολο δεδομένων για το μοντέλο ARIMA όπου περιέχει την ταχύτητα του πλοίου βάσει της ημερομηνίας και ένα σύνολο δεδομένων για το μοντέλο VAR όπου περιέχει δεδομένα πλοίου, καιρού, ανέμου και ωκεανού βάσει της ημερομηνίας. Για την αξιολόγηση των προβλέψεων των μοντέλων επιλέχθηκαν τα κριτήρια, AIC,BIC, RMSE και MAE. Μετά την ανάλυση και σύγκριση των δύο μοντέλων, το VAR μοντέλο αποδίδει καλύτερες προβλέψεις από το μοντέλο ARIMA με RMSE 2.67 knots.

Λέξεις Κλειδιά: Χωροχρονικά Δεδομένα, Χρονοσειρές, Πρόβλεψη Ταχύτητας Πλοίου, ARIMA Μοντέλο, VAR Μοντέλο

Abstract

This research has been conducted during my postgraduate degree at the University of Piraeus, Department of Digital Systems and the main goal is to predict vessel's speed with timeseries. More specifically, the data that have been used was downloaded from the following link: Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance | Zenodo [16], and contain data that were collected from the West coast of France, the Bay of Biscay, with AIS sensors, for time period of six months, from October 1st, 2015 to March 31st, 2016. ARIMA and VAR models have been used for our predictions. For the timeseries analysis we created one dataset for ARIMA model that contains vessel speed by datetime and one dataset for VAR model that contains ship, weather, wind and ocean data by datetime. The criteria that have been selected to evaluate the predictions of models are AIC, BIC, RMSE and MAE. After analyzing and comparing these two models, we conclude that the VAR model performs better than ARIMA with RMSE score of 2.67 knots.

Keywords: Spatiotemporal Data, Time Series, Vessel Speed Prediction, ARIMA Model, VAR Model

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Φιλιππάκη Μιχαήλ για τις σαφείς οδηγίες, την πολύτιμη εποπτεία και καθοδήγησή του κατά την διάρκεια της διπλωματικής και κυρίως για την εμπιστοσύνη και ενθάρρυνση που μου έδωσε ώστε να ασχοληθώ με αυτό το αντικείμενο. Επιπλέον ένα μεγάλο ευχαριστώ σε όλους τους καθηγητές του μεταπτυχιακού προγράμματος «Μεγάλα Δεδομένα και Αναλυτική» για όλα τα εφόδια που μας έδωσαν.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου για όλη την δύναμη και την στήριξη που μου προσέφεραν όλα αυτά τα χρόνια, καθώς αποτελούν παράδειγμα και έμπνευση στην ζωή μου.

Πίνακας Περιεχομένων

Περιεχόμενα Εικόνων

Περιεχόμενα Πινάκων

Πίνακας Συντομεύσεων – Ακρωνυμίων

Ακρωνύμιο	Επεξήγηση
IMO	International Maritime Organization
SSMS	SQL Server Management Studio
SSIS	SQL Server Integration Services
ACVF	Autocovariance Function
ACF	Autocorrelation Function
IID	Independently Identically Distributed
WN	White Noise
ADF	Augmented Dickey Fuller
KPSS	Kwiatkowski-Philips-Schmidt-Shin
PP	Phillips-Perron
AR	Autoregressive Model
MA	Moving Average Model
ARMA	Autoregressive Moving Average Model
ARIMA	Autoregressive Integrated Moving Average Model
AIC	Akaike's Information Criterion
BIC	Bayesian Information Criterion
VAR	Vector Auto Regression Model
RMSE	Root Mean Square Error
MAD	Mean Absolute Deviation
MSE	Mean Square Error
MAPE	Mean Absolute Percentage Error
MPE	Mean Percentage Error

“Don’t never prophesy: If you prophesies right, ain’t nobody going to remember and if you prophesies wrong, ain’t nobody going to let you forget.”

– Mark Twain

Κεφάλαιο 1ο

1.1 Εισαγωγή

Τα τελευταία χρόνια υπάρχει αυξημένο ενδιαφέρον για την βελτιστοποίηση της λειτουργίας του πλοίου για καλύτερη βιωσιμότητα και κερδοφορία καθώς απαιτεί ακριβείς προβλέψεις της ταχύτητας. Πλέον η αυξημένη πρόοδος στην ισχύ των υπολογιστών καθώς και τα αυξημένα διαθέσιμα δεδομένα, ωθούν τις ναυτιλιακές επιχειρήσεις να βασίζονται σε μοντέλα Μηχανικής Μάθησης ώστε να αποδίδονται προβλέψεις με μεγαλύτερη ακρίβεια και να εξάγεται πληροφορία η οποία μπορεί να θεωρηθεί σημαντική στην αποδοτικότητα των πλοίων.

Το τέταρτο τρίμηνο του 2021 αποτέλεσε μια νέα παγκόσμια ενεργειακή κρίση, όπου προκάλεσε πρόσφατα σημαντικό ερευνητικό ενδιαφέρον για την μείωση των καυσίμων κατανάλωσης, το σχετικό κόστος και τις εκπομπές ρύπων. Κύριος παράγοντας για την βελτιστοποίηση της λειτουργίας των πλοίων είναι μια ακριβής πρόβλεψη της ταχύτητας του πλοίου. Οι παραδοσιακές μέθοδοι εκτίμησης της ταχύτητας βασίζονται κυρίως σε θεωρητικούς υπολογισμούς, αριθμητική μοντελοποίηση, προσομοίωση, ή πειραματική εργασία που μπορεί να είναι δαπανηρή, χρονοβόρα, να έχει περιορισμούς αβεβαιότητες ή να μην μπορεί να χρήζει εφαρμογής σε διαφορετικά είδη πλοίων και συνθήκες λειτουργίας τους.

Στην μελέτη [18] έγινε προσέγγιση της ταχύτητας του πλοίου με την χρήση γραμμικής παλινδρόμισης, παλινδρόμηση με δέντρα διαφορετικού μεγέθους και παλινδρούμενες διαδικασίες Gauss, κτλ. Παρατηρήθηκε ότι η εκθετική GPR μέθοδος παρέχει την πιο ακριβή πρόβλεψη, με ποσοστό 91.00%, της ταχύτητας του πλοίου υπό πραγματικές συνθήκες λειτουργίας.

Την 1η Ιανουαρίου του 2020 ο Διεθνής Ναυτιλιακός Οργανισμός (International Maritime Organization – IMO) ανακοίνωσε ότι απαιτείται από τα πλοία να μειώσουν τις εκπομπές θείου από το 3.5% σε 0.5%. Η υιοθέτηση του παραπάνω κανόνα από τις Ναυτιλιακές εταιρείες τις ανάγκασε να στραφούν σε εναλλακτικό τύπο καυσίμου ο οποίος ήταν πιο και πιο ακριβός. Επομένως, για να αντισταθμίσουν τα παραπάνω μέτρα πρέπει είτε να ανανεώσουν τον στόλο τους με νέα πλοία, τα οποία να είναι πιο φιλικά στο περιβάλλον είτε να προσπαθήσουν να βελτιώσουν την απόδοση του υπάρχοντος πλοίου. Στην έρευνα [19] ο συγγραφέας συγκρίνει την απόδοση ενός γνωστού λογισμικού του Simcenter Amesim με αλγόριθμους Μηχανικής Μάθησης για την πρόβλεψη της ταχύτητας του πλοίου για ένα διάστημα έξι ημερών από πραγματικά δεδομένα. Η πρόβλεψη έγινε με το Νευρωνικό μοντέλο ANN και το Vector Linear Regression και τα αποτελέσματα ήταν αρκετά κοντά με τα πραγματικά δεδομένα, με Μέσο Σφάλμα 89 %.

Ένα μοντέλο που είναι ικανό να προβλέψει την ταχύτητα ενός πλοίου μπορεί να διασφαλίσει επίσης την αναμενόμενη ώρα άφιξης του φορτίου [20]. Για παράδειγμα η διαδρομή ενός ταξιδιού μπορεί να βελτιστοποιηθεί σχεδιάζοντας την πορεία του πλοίου με το ελάχιστο κόστος καυσίμων, δηλαδή πρόκειται για την πρόβλεψη της ταχύτητας του πλοίου σε σχέση με τις καιρικές συνθήκες που συναντά το πλοίο, τα χαρακτηριστικά των κινητήρων και την ώθηση της κύριας μηχανής. Η μελέτη έδειξε ότι με την χρήση AR(1) μοντέλου όπου δεν χρησιμοποιήθηκαν οι μεταβλητές καιρού δεν απόδωσαν καλά αποτελέσματα ενώ η ανάλυση παλινδρόμησης όπου συμπεριλάβανε τα δεδομένα καιρού απέδωσε καλύτερα αποτελέσματα με σφάλμα στην αναμενόμενη ώρα άφιξης να είναι περίπου στην 1 ώρα.

Η ταχύτητα του πλοίου είναι μία από τις σημαντικές παραμέτρους που διέπουν την ασφάλεια, την έκτακτη ανάγκη και τον προγραμματισμό μεταφορών στην Αρκτική [21]. Για την πρόβλεψη της ταχύτητας του πλοίου σε νερά τα οποία είναι καλυμμένα με πάγο, άλλες μελέτες κάνουν προσομοιώσεις βασισμένες στην φυσική. Τα δεδομένα μελέτης επιλέχθηκαν από το

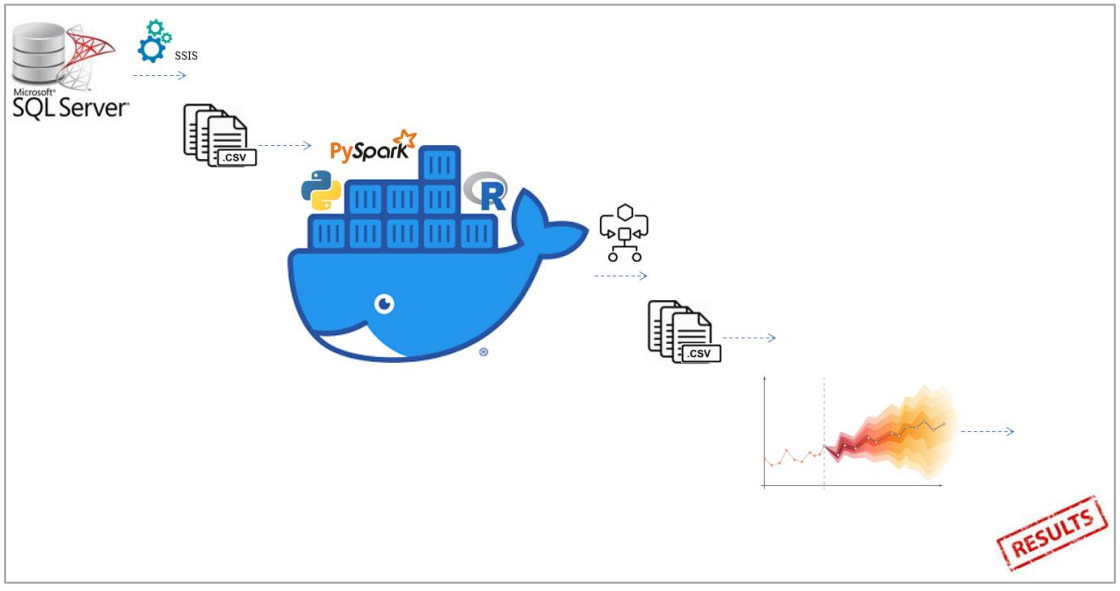
AIS για την περιοχή της Ανατολικής θάλασσας Μπάρεντς και της Νότιας θάλασσας Καρά, χωρίς την ρητή γνώση των τοπικών συνθηκών πάγου γύρω από το σκάφος. Στην συγκεκριμένη έρευνα έγινε η χρήση τριών μοντέλων Μηχανικής Μάθησης (Random Forest, XGBoost και LigthGBM) και πέτυχαν πρόβλεψη με μέσο απόλυτο σφάλμα 3,5 κόμβων.

1.2 Δομή Έρευνας

Στο 1ο κεφάλαιο γίνεται μια συνοπτική εισαγωγή σχετικά με το πώς έχουν βοηθήσει οι προβλέψεις της ταχύτητας του πλοίου με την Μηχανική Μάθηση και Χρονοσειρών και αναφέρεται σχετική αλληλογραφία με εφαρμογές αυτών που αφορούν είτε οικονομικούς είτε περιβαλλοντικούς είτε σε θέματα ασφάλειας. Στο 2ο κεφάλαιο γίνεται μια εισαγωγή στις χρονοσειρές, παρουσιάζονται οι περιγραφικές τεχνικές, μοντέλα προβλέψεων, κριτήρια επιλογής χαρακτηριστικών καθώς και τα κριτήρια με τα οποία θα γίνει η αξιολόγηση της πρόβλεψης. Στο 3ο κεφάλαιο γίνεται η εκτενής ανάλυση των δεδομένων, ο τρόπος με τον οποίο προ επεξεργαστήκαμε τα δεδομένα και παρουσιάζεται ο αλγόριθμος ο οποίος χρησιμοποιήθηκε ώστε να γίνει η εξαγωγή των τελικών αρχείων ανάλυσης. Τέλος, στο 4ο κεφάλαιο παρουσιάζονται οι αναλύσεις, αποτελέσματα των μοντέλων χρονοσειρών που έγιναν στα δύο σύνολα δεδομένων καθώς και η σύγκριση μεταξύ τους.

1.3 Τρόπος Υλοποίησης της Έρευνας

Για την υλοποίησή μας χρησιμοποιήθηκαν διαφορετικά προγράμματα και γλώσσες προγραμματισμού για την εξυπηρέτηση των σκοπών ανάλυσης. Αρχικά τα δεδομένα από [16] έγιναν εισαγωγή τοπικά στον SQL Server Management – SSMS και στην συνέχεια εφόσον πραγματοποιήθηκαν όλες οι απαιτούμενες αλλαγές, έγιναν εξαγωγή τα απαιτούμενα αρχεία με την χρήση του SQL Server Integration Services – SSIS. Στην συνέχεια κατασκευάστηκε ένας Docker Container που περιλαμβάνει τις Pyspark, Python και R γλώσσες προγραμματισμού. Με την χρήση της Pyspark για να αποκτήσουμε την πληροφορία που θέλουμε εφαρμόσαμε τον αλγόριθμο που θα μας διαμερίσει τον χώρο και στο τέλος γίνεται η εξαγωγή των δύο συνόλων δεδομένων που θα χρησιμοποιηθούν στην ανάλυσή μας. Με την χρήση της Python έγινε η ανάλυση των χρονοσειρών μας και εξάχθηκαν τα τελικά αποτελέσματα. Τέλος, έγινε χρήση της γλώσσας R σε παραδείγματα χρονοσειρών που χρειάστηκε στην 2η ενότητα.



2.1.1-1 Υλοποίηση Έρευνας

Κεφάλαιο 2ο

Εισαγωγικές Έννοιες Χρονοσειρών

Η ανάλυση χρονοσειρών δίνει την δυνατότητα για μια σειρά δεδομένων να απαντήσει σε ερωτήματα όπως ποια είναι η αιτιατή επίδραση μιας μεταβλητής σε μια άλλη κατά την πάροδο του χρόνου. Αυτό έχει σαν αποτέλεσμα να εξάγουμε συμπεράσματα για την συμπεριφορά των μεταβλητών. Με βάση την πληροφορία από το παρελθόν, δημιουργούνται μέθοδοι που στοχεύουν είτε να προσδιορίσουν την φύση του προβλήματος είτε να δημιουργήσουν χρήσιμα προβλεπτικά μοντέλα ώστε να βοηθήσουν στην λήψη αποφάσεων. Σε αυτό το κεφάλαιο θα περιγραφούν βασικές έννοιες και ορισμοί χρονοσειρών, θα αναλυθούν τα κύρια χαρακτηριστικά και οι συνιστώσες μιας χρονοσειράς.

2.1 Χρονοσειρές

Μια χρονοσειρά είναι ένα σύνολο παρατηρήσεων πάνω σε μια ποσοτική μεταβλητή που συγκεντρώνονται με το πέρας του χρόνου. Δηλαδή πρόκειται για δεδομένα πάνω στην συμπεριφορά μίας ή περισσότερων μεταβλητών σε διαδοχικές ισαπέχουσες χρονικές περιόδους. Μια χρονοσειρά συνήθως εννοούμε μια ακολουθία παρατηρήσεων $\{X_t : t = 0,1,2,...,T\}$ όπου κάθε X_t εκφράζει την κατάσταση ενός συστήματος κατά την χρονική στιγμή t το οποίο εξελίσσεται στον χρόνο κατά τυχαίο εν γένει τρόπο.

2.1.1 Ορισμός Χρονοσειράς

Το σύνολο δεδομένων το οποίο συλλέγεται διαχρονικά και εκφράζει την εξέλιξη των τιμών μιας μεταβλητής κατά την διάρκεια ίσων διαχρονικών χρονικών περιόδων, ονομάζεται χρονοσειρά ή χρονολογική σειρά ή χρονική σειρά. Σύμφωνα με τον μαθηματικό ορισμό, "χρονοσειρά ορίζεται το σύνολο των παρατηρήσεων $x_1, x_2, ..., x_n$ από τις τιμές $X_1, X_2, ..., X_n$ μιας τυχαίας μεταβλητής X κατά τις ισαπέχουσες χρονικές στιγμές $t_1, t_2, ..., t_n$. Επομένως η ακολουθία αυτή των τυχαίων μεταβλητών ονομάζεται στοχαστική διαδικασία και συμβολίζεται με $X(t)$ [1].

2.1.2 Τύποι Χρονοσειρών

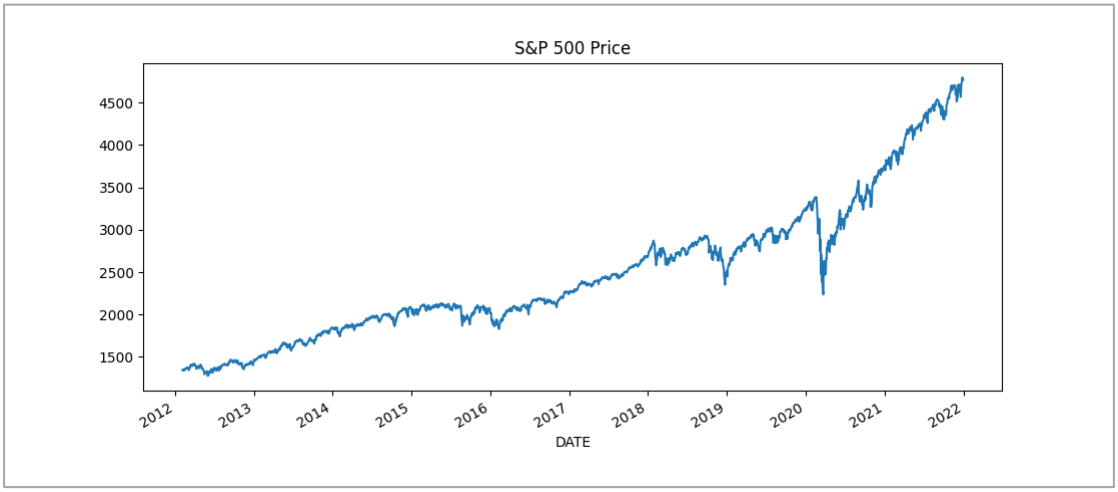
Μια χρονοσειρά είναι ένα σύνολο από παρατηρήσεις που καταγράφονται διαχρονικά μέσα στον χρόνο. Οι καταγραφές αυτές μπορεί να λαμβάνονται σε διακριτό ή και συνεχή χρόνο. Έτσι οι χρονοσειρές διακρίνονται σε δύο βασικές κατηγορίες, βάσει του χρόνου δειγματοληψίας. Οι διακριτές χρονοσειρές (*discrete time series*) είναι αυτές όπου η τιμή του φαινομένου καταγράφεται σε ορισμένα χρονικά διαστήματα, όπως για παράδειγμα η τιμή μιας μετοχής ημερησίως. Οι συνεχείς χρονοσειρές (*continuous time series*) είναι αυτές όπου η τιμή ενός φαινομένου παρατηρείται συνεχώς, όπως για παράδειγμα η συνεχής παρακολούθηση της θερμοκρασίας, αέρα ή ακόμη και η συνεχής παρακολούθηση σεισμών. Επιπλέον οι διακριτές χρονοσειρές θα μπορούσαν επιπλέον να κατηγοριοποιηθούν:

i. Σε στιγμιαίες χρονοσειρές, όπου επιλέγεται ένα δείγμα από μια συνεχή χρονοσειρά, όπως για παράδειγμα ο υπολογισμός της θερμοκρασίας σε ωριαία διαστήματα
ii. Σε συσσωρευμένες χρονοσειρές, όπου αποτελούν συσσώρευση πληροφορίας για κάποια χρονική στιγμή, όπως για παράδειγμα οι πωλήσεις ενός προϊόντος που πραγματοποιούνται στο τέλος κάθε μήνα ή τριμήνου. Οι παρατηρήσεις αυτές που προέρχονται από την ίδια χρονική περίοδο ονομάζονται ομαδοποιημένα διαστρωματικά δεδομένα.

Επίσης, το χρονικό διάστημα μιας χρονοσειράς μπορεί να είναι σταθερό (ημερήσιο, μηνιαίο, τρίμηνο, χρόνου, κτλ.). Επιπλέον, υπάρχουν περιπτώσεις που ο φυσικός χρόνος δεν είναι σταθερός, όπως για παράδειγμα οι τιμές του χρηματιστηρίου στις μη εργάσιμες μέρες (Σαββατοκύριακο, γιορτές, κτλ.) που παραμένει κλειστό, και στις συγκεκριμένες περιπτώσεις πρέπει να ακολουθηθούν διαφορετικές τεχνικές.

2.1.3 Εφαρμογές Χρονοσειρών

Υπάρχουν πολλοί τομείς και πεδία από ερευνητικής πλευράς όπου είναι χρήσιμες οι χρονοσειρές όπως, οικονομικοί, μηχανικοί, περιβαλλοντικοί, πολιτικοί, ιατρικοί, κοινωνικές επιστήμες και πολλοί άλλοι [3]. Ένα χαρακτηριστικό παράδειγμα είναι η τιμή κλεισίματος μιας μετοχής, καθώς στην 2.1.3-1 απεικονίζονται οι μεταβολές των τιμών του χρηματιστηριακού δείκτη S&P500 για το χρονικό διάστημα από 2012-01-01 έως 2022-01-01.



2.1.3-1 Ημερήσια μεταβολή τιμών του Χρηματιστηριακού δείκτη S&P500

2.1.4 Χρησιμότητα Χρονοσειρών

Ο κύριος σκοπός στην ανάλυση χρονοσειρών είναι ο προσδιορισμός ενός μοντέλου ώστε να περιγράφει το μοτίβο μιας χρονοσειράς. Οι βασικές χρήσεις ενός τέτοιου μοντέλου είναι οι παρακάτω [2]:

i. Περιγραφή (Describe): Τα πρώτα βήματα για την ανάλυση είναι η απεικόνιση των δεδομένων σε γράφημα, και η χρήση στατιστικών μεθόδων για την περιγραφή των δεδομένων.
ii. Εξήγηση (Explanatory): Η εξήγηση χρονοσειράς με μια μεταβλητή, βασίζεται μόνο σε παρελθοντικές τιμές της μεταβλητής αυτής. Ενώ, για να εξηγήσουμε την συμπεριφορά μιας χρονοσειράς με πολλές μεταβλητές, μπορεί να χρησιμοποιηθεί η συμπεριφορά μιας μεταβλητής ώστε να εξηγηθεί η συμπεριφορά μιας άλλης ή ακόμα και να γίνουν μοντέλα εκπαίδευσης πάνω στις μεταβλητές αυτές.
iii. Πρόβλεψη (Forecasting): Είναι η πρόβλεψη ή εκτίμηση μελλοντικών τιμών μιας χρονοσειράς. Δηλαδή χρησιμοποιούνται οι υπάρχουσες παρατηρήσεις ώστε να γίνει πρόβλεψη μελλοντικών τιμών.
iv. Έλεγχος (Control): Είναι η διαδικασία με την οποία μια πρόβλεψη επιτρέπει στον αναλυτή να προσδιορίσει διορθωτικές κινήσεις που θα μπορούσαν να δημιουργήσουν προβλήματα στην ανάλυση.

2.2 Περιγραφικές Τεχνικές

Για την στατιστική επεξεργασία των χρονοσειρών είναι ιδιαίτερα χρήσιμο να διακρίνουμε τις τέσσερις συνιστώσες της σύμφωνα με την παρακάτω σχέση [4]:

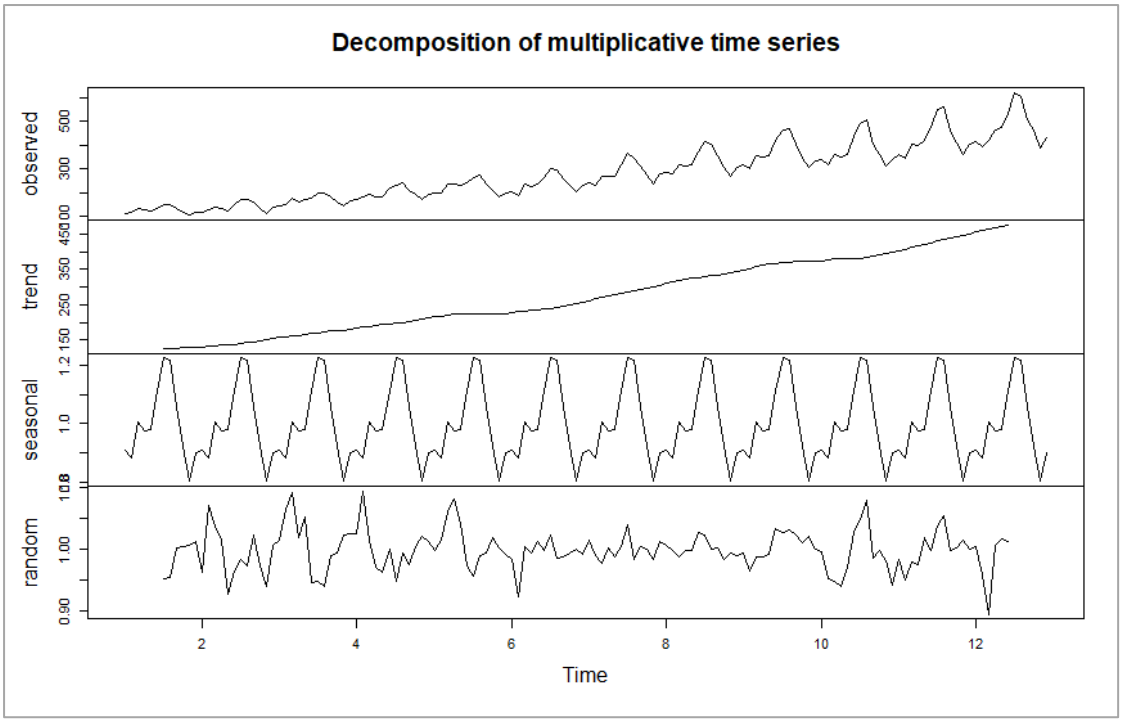
i. Τάση (Trend): Τάση είναι η μακροχρόνια κίνηση που ακολουθεί η χρονοσειρά όταν παρατηρείται σε μια εκτεταμένη περίοδο του χρόνου και υπάρχουν αυξήσεις ή μειώσεις των τιμών της. Η τάση μπορεί να θεωρηθεί και ανύπαρκτη όταν η κεντρική κίνησή της είναι παράλληλη με τον άξονα του χρόνου, χωρίς να υπάρχουν αυξομειώσεις. Οι πιο συνήθεις μέθοδοι προσδιορισμού της τάσης είναι μέσω της μεθόδου κινητών μέσων, ελαχίστων τετραγώνων και άλλοι.

ii. Εποχικότητα(Seasonality): Μια χρονοσειρά παρουσιάζει εποχικότητα όταν η διασπορά της παρουσιάζει ίδια συμπεριφορά ανά χρονικές περιόδους t. Συνήθως οι περιοδικές διακυμάνσεις αναφέρονται σε χρονικά διαστήματα μικρότερα του έτους. Ένα παράδειγμα εποχικότητας είναι η αύξηση των πωλήσεων στο παγωτό κατά τους καλοκαιρινούς μήνες.

iii. Κυκλικότητα (Cyclicity): Η κυκλικότητα εκφράζει τις κυκλικές διακυμάνσεις για περιόδους μεγαλύτερες του έτους που επαναλαμβάνονται σε ίσα χρονικά διαστήματα και οφείλονται σε εξωτερικούς παράγοντες.

iv. Τυχαίες Κυμάνσεις (Irregular Fluctuations / Outliers): Οι ακανόνιστες ή ακραίες ή τυχαίες κινήσεις σε μια χρονοσειρά αποτελούν τον θόρυβό της και χαρακτηρίζονται από μικρές χρονικές περιόδους. Είναι δηλαδή τιμές οι οποίες απέχουν σημαντικά από τις υπόλοιπες παρατηρήσεις, όπου συνήθως οφείλονται σε κάποιο απρόβλεπτο παράγοντα και δημιουργούν πρόβλημα στην μοντελοποίηση και για αυτό χρήζουν σημαντική προσοχή.

Στην παρακάτω εικόνα 2.2-1 γίνεται απεικόνιση των συνιστωσών μιας χρονοσειράς.



2.2-1 Συνιστώσες Χρονοσειράς

2.3 Ανάλυση Χρονοσειρών

Για την ανάλυση των χρονοσειρών χρησιμοποιείται είτε το προσθετικό μοντέλο είτε πολλαπλασιαστικό μοντέλο. Το Προσθετικό Μοντέλο (1) θεωρεί ότι όλες οι συνιστώσες είναι ανεξάρτητες μεταξύ τους και εκφράζονται στην ίδια μονάδα μέτρησης της χρονοσειράς ενώ το Πολλαπλασιαστικό μοντέλο (2) θεωρεί ότι μόνο η τάση εκφράζεται στην ίδια μονάδα μέτρησης αυτής της χρονοσειράς.

$$Y_t = T_t + S_t + C_t + I_t \qquad t = 1, 2, ..., n \tag{2.3-1}$$

$$Y_t = T_t \cdot S_t \cdot C_t \cdot I_t \qquad t = 1, 2, ..., n \tag{2.3-2}$$

Όπου:

Y_t : η παρατηρούμενη χρονοσειρά, ή κάποιος (συνήθως ο λογαριθμικός) μετασχηματισμός της

T_t : η μακροχρόνια τάση

S_t : η εποχική συνιστώσα

C_t : η κυκλική συνιστώσα

I_t : η άρρυθμη συνιστώσα

Επιπλέον, για την σωστή ανάλυση μιας χρονοσειράς πρέπει να τηρούνται τα παρακάτω βήματα:

i. Με την γραφική απεικόνιση της χρονοσειράς, πρέπει να ελεγχθούν τα παρακάτω:
α) η τάση

β) η εποχικότητα

γ) τυχών απότομες ανωμαλίες στην συμπεριφορά της χρονοσειράς ή αποκλίνουσων τιμών

ii. Να γίνει απαλοιφή της τάσης και εποχικότητας με σκοπό να έχουν στάσιμα κατάλοιπα.

Για να επιτευχθεί αυτό πρέπει να γίνουν οι παρακάτω μετατροπές στα δεδομένα:

α) Εάν οι διακυμάνσεις φαίνονται να αυξάνουν γραμμικά με τον χρόνο η χρονοσειρά μπορεί να μετατραπεί σε λογαριθμική $\{\ln X_1, ..., \ln X_n\}$ ώστε οι διακυμάνσεις να γίνουν πιο σταθερές. Επιπλέον, πρέπει να διασφαλιστεί ότι όλα τα δεδομένα είναι θετικά, στην αντίθετη περίπτωση, πρέπει να προστεθεί μια θετική σταθερά πριν μετατραπεί σε λογαριθμική εξίσωση η χρονοσειρά.

β) Με χρήση υστερήσεων, δηλαδή να γίνει αντικατάσταση μιας χρονοσειράς $\{X_t\}$ σε $\{Y_t : X_t - X_{t-d}\}$ με $d > 0$, με σκοπό να μετατραπεί σε στάσιμη.

iii. Να εξασφαλιστεί ότι έχει επιλεγεί το κατάλληλο μοντέλο ώστε να προσαρμόζεται στα δεδομένα μας, βάσει των αυτοσυσχετίσεων ώστε να χρησιμοποιηθεί στην συνέχεια το τελικό μοντέλο για τις προβλέψεις.

iv. Τέλος μια επιπλέον χρήσιμη ενναλακτική είναι να εκφραστεί η χρονοσειρά σε όρους Fourier. Η μετατροπή αυτή είναι πολύ σημαντική σε αρκετούς τομείς, όπως η επεξεργασία σήματος.

2.3.1 Στάσιμα Μοντέλα & Συνάρτηση Αυτοσυσχέτισης

Για μια χρονοσειρά $\{X_t, t=0,\pm1,....\}$ από μια χρονολογική περίοδο t έως την παρούσα χρονική περίοδο $t=h$ και θέλουμε να προβλέψουμε τις μελλοντικές τιμές της, πρέπει να βασιστούμε στις ήδη υπάρχουσες γνωστές τιμές και στην εξάρτησή τους που ενδέχεται να υπάρχει μεταξύ των παρελθοντικών με των μελλοντικών τιμών της χρονοσειράς, υποθέτοντας ότι παραμένουν αναλλοίωτες μέσα στο πέρας του χρόνου. Πιο συγκεκριμένα, αν η μέση τιμή μ_t, η διακύμανση γ_{0t} και οι αυτοσυνδιακυμάσνσεις $\gamma_{jt}, j=1,2,...$ δεν εξαρτώνται από την χρονική στιγμή t τότε η στοχαστική ανέλιξη ονομάζεται ασθενώς στάσιμη ή στασιμότητα β΄ τάξης (Weak Stationarity or Second order Stationarity). Δηλαδή μια στοχαστική διαδικασία λέγεται ασθενώς στάσιμη όταν για κάθε χρονική στιγμή έχει τις κάτωθι ιδιότητες [5]:

Έχει σταθερή μέση τιμή σε όλες τις χρονικές περιόδους.

$$E(Z_t) = \mu_t = \int Z_t f(z_t) dz \tag{2.3.1-1}$$

Έχει σταθερή πεπερασμένη διακύμανση σε όλες τις χρονικές περιόδους.

$$Var(Z_t) = \sigma_t^2 = E(Z_t - \mu_t)^2 = \int (Z_t - \mu_t)^2 f(z_t) \tag{2.3.1-2}$$

Οι συνδιακυμάσεις (ACVF) της χρονοσειράς μεταξύ των τιμών σε οποιαδήποτε χρονικά διαστήματα $t_1 - t_2$ εξαρτώνται μόνο από το k, δηλαδή στην διαφορά μεταξύ των δύο χρονικών περιόδων και όχι από την θέση των σημείων στους άξονες.

$$Cov(Z_{t_1}, Z_{t_2}) = E[(Z_{t_1} - \mu_{t_1})(Z_{t2} - \mu_{t_2})] = \gamma(t_1 - t_2) \tag{2.3.1-3}$$

Έχει συντελεστή αυτοσυσχέτισης (ACF)

$$\rho(t_1, t_2) = \frac{\gamma(t_1 - t_2)}{\sqrt{\sigma_{t_1}^2} \cdot \sqrt{\sigma_{t2}^2}} = \frac{\gamma_k}{\gamma_0} \tag{2.3.1-4}$$

Να σημειωθεί ότι η συνάρτηση $\gamma(t_1 - t_2)$ ονομάζεται αυτοσυσχέτιση με $k = (t_1 - t_2)$ και γ(0) διακύμανση. Επίσης ο συντελεστής συσχέτισης παίρνει τιμές: $-1 < \rho < 1$, όπου:

$\rho = -1$: τέλεια αρνητική συσχέτιση

$\rho = 0$: οι μεταβλητές δεν συσχετίζονται

$\rho = 1$: τέλεια θετική συσχέτιση

Αυστηρή στασιμότητα (Strict Stationarity), είναι μια χρονοσειρά $\{X_t, t=0,\pm1,....\}$ με την προϋπόθεση ότι $\{X_1,....,X_n\} = \{X_{1+h},....,X_{n+h}\}$ είναι ισοκατανεμημένη για όλα τα χρονικά

διαστήματα h με $n>0$ ή διαφορετικά συγκλίνει ομοιόμορφα στην μακροχρόνια ισορροπία της. Επομένως, μια αυστηρά στάσιμη διαδικασία είναι υποχρεωτικά και ασθενώς στάσιμη, το αντίθετο όμως δεν ισχύει. Εξαίρεση αποτελεί η κανονική κατανομή, όπου η αυστηρή στασιμότητα ταυτίζεται με την ασθενή στασιμότητα.

2.3.2 Λευκός Θόρυβος & IID Θόρυβος

Ένα στοχαστικό υπόδειγμα $\{X_t : t \in \mathbb{Z}\}$ με $t = 0 \pm 1, \pm 2, \ldots$, με τις παρακάτω ιδιότητες:

$$\begin{cases} E(X_t) = 0 \quad \forall t & (1) \\ Var(X_t) = \sigma^2 < \infty \quad \forall t & (2) \\ Cov(X_t, X_{t-k}) = 0 \quad \forall t, \forall k & (3) \end{cases} \qquad (2.3.2\text{-}1)$$

ονομάζεται Λευκός Θόρυβος (White Noise) με μέση τιμή μηδέν και σταθερή διακύμανση και συμβολίζεται:

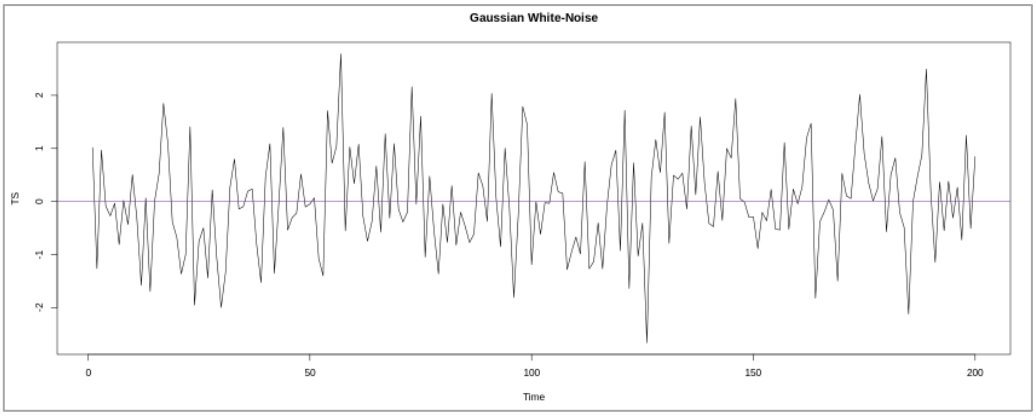
$$\{X_t\} \sim WN(0, \sigma^2) \qquad (2.3.2\text{-}2)$$

Η πρώτη συνθήκη αναφέρει ότι η αναμενόμενη τιμή θα είναι πάντα σταθερή και ίση με το μηδέν. Στην δεύτερη συνθήκη αναφέρεται ότι η διακύμανση θα είναι σταθερή. Τέλος, η τρίτη συνθήκη αναφέρει ότι οι μεταβλητές είναι ασυσχέτιστες για όλες τις υστερήσεις. Εάν, το στοχαστικό υπόδειγμα είναι ομοιόμορφα κατανεμημένο με μέση τιμή μηδέν και σταθερή διακύμανση τότε ονομάζεται IID Θόρυβος (Independently Identically Distributed) και συμβολίζεται ως:

$$\{X_t\} \sim IID(0, \sigma^2) \qquad (2.3.2\text{-}3)$$

Να σημειωθεί ότι μια $ID(0, \sigma^2)$ συχνότητα είναι και $WN(0, \sigma^2)$ αλλά το αντίστροφο δεν ισχύει.

Επιπλέον, υπάρχει και ο Gaussian Λευκός Θόρυβος, όπου η $\mu = 0$ και η $\sigma^2 = 1$ και απεικονίζεται στην παρακάτω εικόνα.



2.3.2-1 Gaussian White Noise

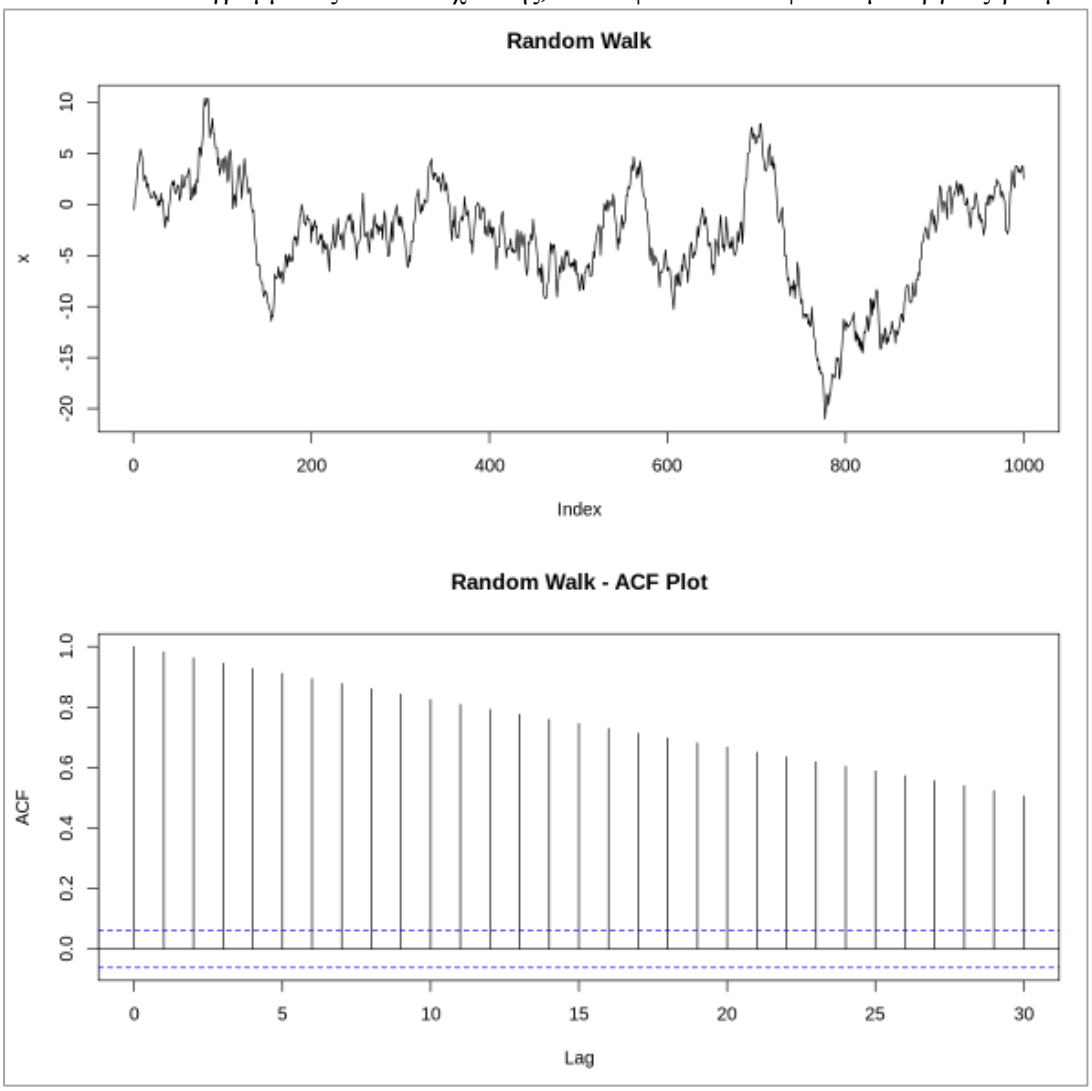
2.3.3 Τυχαίος Περίπατος – Random Walk

Ο τυχαίος περίπατος αποτελεί ένα επιπλέον είδος στοχαστικών διαδικασιών, καθώς κατατάσσονται στις κατηγορίες των μη – στάσιμων χρονοσειρών. Εάν υποθέσουμε ότι ένα στοχαστικό υπόδειγμα $\{X_t : t = 0,1,2,...\}$ με τις παρακάτω συνθήκες:

$$\begin{cases} X_0 = \delta < \infty \\ X_t = X_{t-1} + w_t, \ \forall \ t = 1,2,... \\ w_t \sim WN(0, \sigma_u^2) \end{cases} \qquad (2.3.3\text{-}1)$$

οονομάζεται Τυχαίος Περίπατος (RandomWalk). Δηλαδή, η κάθε παρατήρηση X_t προκύπτει από την παρατήρηση της προηγούμενης χρονικής περιόδου X_{t-1}, αυξημένη κατά τον Gaussian θόρυβο w_t. Στην εικόνα 2.3.3-1 απεικονίζεται ένα παράδειγμα τυχαίου περίπατου με την

συνοδεία του διαγράμματος αυτοσυσχέτισης, όπου φαίνεται να φθίνει με αργούς ρυθμούς.



2.3.3-1 Random Walk - ACF plot

2.3.4 Έλεγχοι Στασιμότητας

Η στασιμότητα μιας χρονοσειράς είναι ένα πολύ σημαντικό κριτήριο, ώστε να διεξαχθεί ένα προβλεπτικό μοντέλο. Οι τρόποι με τους οποίους μπορούμε να ελέγξουμε την στασιμότητα μιας χρονοσειράς είναι με οπτικούς τρόπους και οι έλεγχοι μοναδικών ριζών.

Ο οπτικός έλεγχος στασιμότητας αναφέρεται στον τρόπο με το οποίο μια χρονοσειρά αποτυπώνεται σε μια γραφική παράσταση συναρτήσει του χρόνου. Εάν διαπιστώσουμε ότι η χρονοσειρά παρουσιάζει τάση, εποχικότητα, κυκλική διακύμανση ή ακανόνιστες μεταβολές, τότε ότι η χρονοσειρά δεν είναι στάσιμη. Επιπλέον, από το διάγραμμα της αυτοσυσχέτισης μπορεί να διαπιστωθεί ότι εάν ο συντελεστής αυτοσυσχέτισης ξεκινάει από υψηλές τιμές και φθίνει με αργό ρυθμό, τότε η χρονοσειρά δεν είναι στάσιμη. Σε αυτές τις περιπτώσεις καταφεύγουμε στην λύση των πρώτων διαφορών όπου στην συνέχεια η εξεταζόμενη μεταβλητή της χρονοσειράς γίνεται στάσιμη. Επομένως, χρησιμοποιούνται οι παρακάτω έλεγχοι στασιμότητας:

H_0: Εάν ο συντελεστής αυτοσυσχέτισης $\rho_k = 0$, τότε δεν υπάρχει συσχέτιση και η χρονοσειρά είναι στάσιμη.

H_1: Εάν ο συντελεστής αυτοσυσχέτισης $\rho_k \neq 0$, τότε υπάρχει συσχέτιση και η χρονοσειρά δεν είναι στάσιμη.

2.3.4.1 Box – Pierce Test

Σύμφωνα με τον έλεγχο Box – Pierce, οι συντελεστές συσχέτισης είναι μηδέν και ορίζεται ως εξής [10]:

$$Q = n\sum_{k-1}^{m} \hat{p}_k^2 \sim X^2(m) \qquad (2.3.4.1\text{-}1)$$

Όπου:

Q: Είναι τιμή ελέγχου Box – Pierce

n: Αριθμός Παρατηρήσεων

$\hat{p}$: Αυτοσυσχέτιση

m : Μέγεθος Υστερήσεων

Επειδή όμως το στατιστικό Q δεν είναι αξιόπιστο για μικρά δείγματα, οι Ljung – Box [11] πρότειναν μια παραλλαγή του ελέγχου (2.3.4.1-1) ώστε η στατιστική τιμή του LB_Q να είναι ισχυρή τόσο σε μικρά δείγματα όσο και σε μεγάλα δείγματα και δίνεται από τον παρακάτω τύπο:

$$LB\_Q = n(n+2)\sum_{k-1}^{m}\left(\frac{\hat{p}_k^2}{n-k}\right) \sim X^2(m) \qquad (2.3.4.1.2)$$

Με υποθέσεις ελέγχου:

H_0: Εάν ο συντελεστής αυτοσυσχέτισης $\rho_k = 0$ και $LB\_Q < X_a^2(\mu)$ τότε η χρονοσειρά είναι στάσιμη.

H_1: Εάν ο συντελεστής αυτοσυσχέτισης $\rho_k \neq 0$ και $LB\_Q > X_a^2(\mu)$ τότε η χρονοσειρά δεν είναι στάσιμη.

2.3.4.2 ADF Test

Έλεγχοι όπως Augmented Dickey Fuller Test (ADF Test) και Kwiatkowski-Phillips-Schmidt-Shin (KPSS Test) ανήκουν στους ελέγχους που ονομάζονται μοναδικών ριζών [6], [7]].

Εάν υποθέσουμε για τον έλεγχο ADF Test μια στοχαστική διαδικασία

$$Y_t = \varphi Y_{t-1} + \varepsilon_t \qquad (2.3.4.2\ \text{-}1)$$

όπου $|\varphi| \le 1$ και ε_t είναι ο Λευκός Θόρυβος.

Εάν το $|\varphi| = 1$ τότε λέμε ότι έχουμε μοναδική ρίζα. Πιο συγκεκριμένα εάν το $|\varphi| = 1$ είναι υπόδειγμα τυχαίου περίπατου (χωρίς drift) και είναι μη στάσιμη χρονοσειρά. Διαφορετικά εάν το $|\varphi| < 1$ τότε η χρονοσειρά είναι στάσιμη. Τέλος, εάν το $|\varphi| > 1$ τότε η χρονοσειρά αυξάνεται συνεχώς με την πάροδο του χρόνου. Ο έλεγχος ADF αρχικά υπολογίζει την πρώτη διαφορά:

$$Y_t - Y_{t-1} = \varphi Y_{t-1} + \varepsilon_t - Y_{t-1} \Rightarrow Y_t - Y_{t-1} = Y_{t-1}(\varphi - 1) + \varepsilon_t \qquad (2.3.4.2\text{-}2)$$

Εάν υποθέσουμε ότι $\Delta Y_t = Y_t - Y_{t-1}$ και ότι $\beta = \varphi - 1$ τότε η παραπάνω εξίσωση διαμορφώνεται ως εξής:

$$\Delta Y_t = \beta Y_t + \varepsilon_t \qquad (2.3.4.2\text{-}3)$$

Εφαρμόζεται η μέθοδος των ελαχίστων τετραγώνων στην εξίσωσή παλινδρόμησης (2.3.4.2-3) και με βάση τον λόγο $t_\beta = \dfrac{\varphi - 1}{s_{\hat{\varphi}}} = \dfrac{\hat{\beta}}{s_{\hat{\beta}}}$, όπου $s_{\hat{\varphi}}$ και $s_{\hat{\beta}}$ αποτελούν τα εκτιμώμενα τυπικά σφάλματα των εκτιμώμενων παραμέτρων.

Έτσι διαμορφώνονται οι παρακάτω έλεγχοι υποθέσεων:

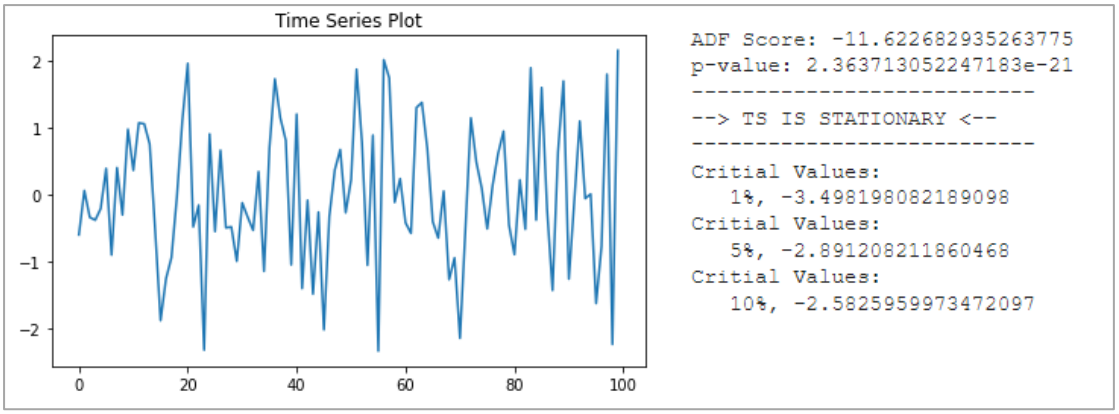
H_0: Εάν $\beta = 0 \to \varphi = 1$ και $t_\beta > \text{στατιστικά } \tau$, τότε η χρονοσειρά δεν είναι στάσιμη.

H_1: Εάν $\beta < 0 \to \varphi < 1$ και $t_\beta < \text{στατιστικά } \tau$, τότε η χρονοσειρά είναι στάσιμη.

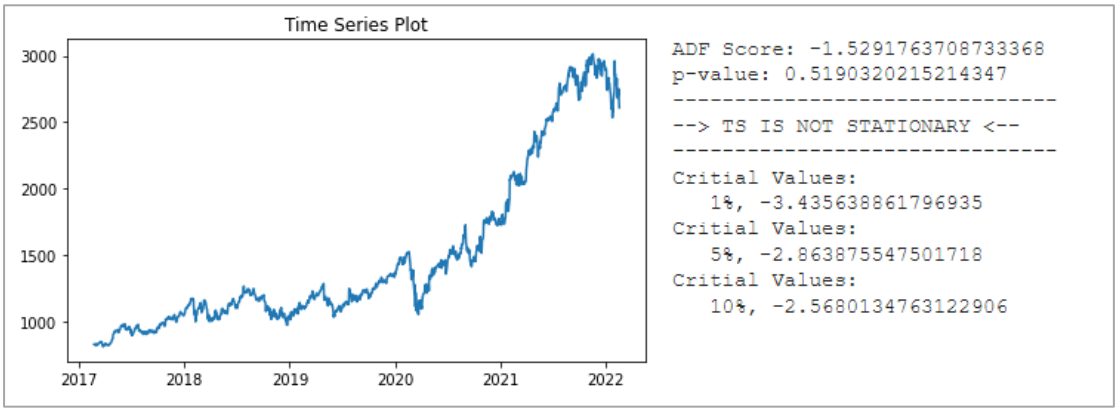
Επίσης για τον χαρακτηρισμό της στασιμότητας με επίπεδα σημαντικότητας $\alpha = 1\%$, $\alpha = 5\%$ και $\alpha = 10\%$ συγκρίνουμε τη τιμή t-Statistic ADF με την τιμή Mackinnon σε κάθε επίπεδο σημαντικότητας $(1\%, 5\%, 10\%)$. Εάν η τιμή t-Statistic ADF είναι μικρότερη της τιμής Mackinnon σε επίπεδο $(1\%, 5\%, 10\%)$ τότε η χρονοσειρά είναι στάσιμη.

Επιπλέον, εάν η στασιμότητα επιτυγχάνεται σε επίπεδο 1% τότε λέμε ότι έχουμε ισχυρή στασιμότητα, εάν η στασιμότητα επιτυγχάνεται σε επίπεδο 5% τότε λέμε ότι έχουμε μέτρια στασιμότητα και εάν η στασιμότητα επιτυγχάνεται σε επίπεδο 10% τότε λέμε ότι έχουμε ασθενής στασιμότητα.

Στις παρακάτω εικόνες 2.3.4.2-1, 2.3.4.2-2 αναφέρονται δύο παραδείγματα όπου με βάση τον έλεγχο ADF καταλήγουμε εάν μια χρονοσειρά είναι στάσιμη ή όχι. Παρατηρούμε ότι στην εικόνα 2.3.4.2-1 η $p - value \simeq 2.363 \cdot 10^{-21} < 0.05$, και με τιμή ADF Score $\simeq -11.62 < (1\% \simeq -3.498, 5\% \simeq -2.891, 10\% \simeq -2.582)$ επομένως, απορρίπτουμε την μηδενική υπόθεση H_0 που σημαίνει ότι η χρονοσειρά έχει ισχυρή στασιμότητα, ενώ στην εικόνα 2.3.4.2-2 παρατηρούμε ότι $p - value \simeq 0.519 > 0.05$ με αποτέλεσμα να δεχόμαστε την μηδενική υπόθεση H_0 που σημαίνει ότι η χρονοσειρά δεν είναι στάσιμη.



2.3.4.2-1 Augmented Dickey Fuller (ADF) - Στάσιμη Χρονοσειρά



2.3.4.2-2 Augmented Dickey Fuller (ADF) - Μη Στάσιμη Χρονοσειρά

2.3.4.3 KPSS Test

Οι Kwiatkowski-Phillips-Schmidt-Shin πραγματοποίησαν έναν έλεγχο στασιμότητας με την διαφορά ότι η μηδενική υπόθεση υποθέτει ότι δεν υπάρχει μοναδική ρίζα και η χρονοσειρά θεωρείται στάσιμη. Υποθέτουμε για τον έλεγχο KPSS Test μια στοχαστική διαδικασία

$$Y_t = \varphi_0 + \varphi_2 Y_{t-1} + \varepsilon_t \qquad (2.3.4.3\text{-}1)$$

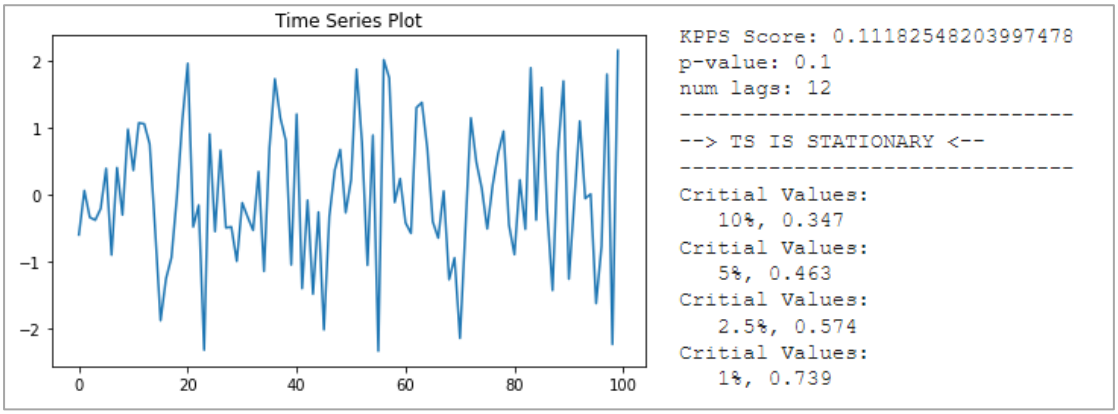
με τις παρακάτω υποθέσεις ελέγχου:

H_0: Εάν $\varphi_2 = 0$, και t-Statistics KPPS $>$ κρίσιμη τιμή τότε η χρονοσειρά είναι στάσιμη

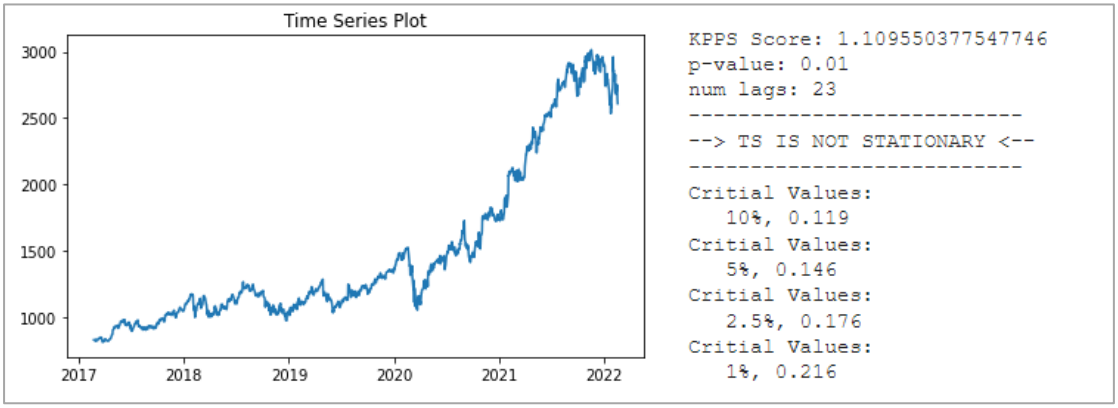
H_1: Εάν $\varphi_2 < 0$, και t-Statistics KPPS $<$ κρίσιμη τιμή τότε η χρονοσειρά δεν είναι στάσιμη

Στις παρακάτω εικόνες 2.3.4.3-3, 2.3.4.3-4 αναφέρονται δύο παραδείγματα όπου με βάση τον έλεγχο KPSS καταλήγουμε εάν μια χρονοσειρά είναι στάσιμη ή όχι. Παρατηρούμε ότι στην εικόνα 2.3.4.3-3 η $p-value \simeq 0.1 > 0.05$, και με τιμή KPSS Score $\simeq 0.111 > (1\% \simeq 0.739\%, 2.5\% \simeq 0.574, 5\% \simeq 0.463, 10\% \simeq 0.347)$, δεχόμαστε την μηδενική υπόθεση H_0 που σημαίνει ότι η χρονοσειρά έχει ισχυρή στασιμότητα, ενώ στην

εικόνα 2.3.4.3-4 παρατηρούμε ότι $p-value \simeq 0.01 < 0.05$ με αποτέλεσμα να δεχόμαστε την μηδενική υπόθεση H_0 που σημαίνει ότι η χρονοσειρά δεν είναι στάσιμη.



2.3.4.3-3 KPSS Test - Στάσιμη Χρονοσειρά



2.3.4.3-4 KPSS Test - Μη Στάσιμη Χρονοσειρά

2.4 Μετασχηματισμός μη-στάσιμης σε στάσιμη Χρονοσειρά

Όπως έχει αναφερθεί και στο κεφάλαιο 2.3 για την ανάλυση μιας χρονοσειράς, θα πρέπει να εξασφαλίσουμε την στασιμότητά της. Σε περίπτωση που η χρονοσειρά δεν είναι στάσιμη θα πρέπει να απαλειφθούν τυχόν μεταβολές της τάσης ή και της εποχικότητας. Πιο συγκεκριμένα η απαλοιφή της τάσης ή της εποχικότητας (ή γενικά της περιοδικότητας) εφαρμόζεται ώστε να μελετηθούν οι μεταβολές της χρονοσειράς που οφείλονται σε τάσεις ή περιοδικότητα διότι θεωρούμε ότι δεν σχετίζονται με το σύστημα. Αντιθέτως, όταν θέλουμε να αντλήσουμε πληροφορία από την χρονοσειρά σχετικά με την τάση ή εποχικότητα ή και τα δύο, τότε η ανάλυσή μας περιορίζεται στην εκτίμηση της τάσης ή της εποχικότητας και η υπόλοιπη πληροφορία δεν λαμβάνεται υπόψη. Για τις προβλέψεις, είτε απαλείφουμε την τάση ή την εποχικότητα ώστε να γίνει η εκτίμηση είτε την συμπεριλαμβάνουμε για να πάρουμε την πρόβλεψη του πραγματικού μεγέθους [5].

2.4.1 Απαλοιφή της Τάσης

Έστω ότι θεωρούμε ότι μια μη στάσιμη χρονοσειρά οφείλεται στην ύπαρξη της τάσης τότε η χρονοσειρά Y_t θα έχει την παρακάτω μορφή:

$$Y_t = \mu_t + X_t \tag{2.4.1-1}$$

Υπάρχουν τρείς διαφορετικοί τρόποι απαλοιφής της τάσης [5]. Όταν η τάση μεταβάλλεται συναρτήσει του χρόνου τότε σε αυτή την περίπτωση μπορεί να εφαρμοστεί μια παραμετρική συνάρτηση $f(t)$, βαθμού p

$$\mu_t = f(t) = a_0 + a_1 t + \ldots + a_t t^p \tag{2.4.1-2}$$

ώστε να γίνει η εκτίμηση και στην συνέχεια να γίνει απαλοιφή. Ένας δεύτερος τρόπος είναι όταν η τάση μπορεί να απαλειφθεί από τις πρώτες διαφορές (ή την πρώτη υστέρηση), καθώς χρησιμοποιείται ο παρακάτω μετασχηματισμός:

$$X_t = \nabla Y_t = Y_t - Y_{t-1} = (1-B)Y_t \tag{2.4.1-3}$$

Στην περίπτωση που δεν γίνει στάσιμη η χρονοσειρά μπορούμε να πάρουμε πάλι τις πρώτες διαφορές και ονομάζεται μετασχηματισμός δεύτερης υστέρησης και δίνεται από τον παρακάτω μετασχηματισμό:

$$X_t = \nabla^2 Y_t = \nabla\left(\nabla Y_t\right) = (1-B)(1-B)Y_t = (1-2B+B^2)Y_t = Y_t - 2Y_{t-1} + Y_{t-2} \tag{2.4.1-4}$$

Γενικότερα εάν η τάση εκφράζεται από πολυώνυμο βαθμού p τότε η απαλοιφή γίνεται με την χρήση των p υστερήσεων ως εξής:

$$X_t = \nabla^P Y_t \tag{2.4.1-5}$$

Ένας τρίτος τρόπος απαλοιφής της τάσης είναι η χρήση του κινητού μέσου τάξης $2q+1$. Δηλαδή για κάθε χρονικής στιγμή t με $q < t \le n-q$, γίνεται εκτίμηση της μ_t της χρονοσειράς από τον μέσο των παρατηρήσεων στο χρονικό διάστημα $\left[t-q, t+q\right]$ δηλαδή:

$$\hat{\mu}_t = \frac{1}{2q+1}\sum_{j=-p}^{q} Y_{t-j} \tag{2.4.1-6}$$

2.4.2 Απαλοιφή Εποχικότητας ή Περιοδικότητας

Εάν υποθέσουμε ότι η χρονοσειρά έχει εποχικότητα ή περιοδικότητα, τότε η χρονοσειρά δίνεται από τον παρακάτω τύπο:

$$Y_t = s_t + X_t \tag{2.4.2-1}$$

Παρομοίως όπως και στην τάση, υπάρχουν τρείς διαφορετικοί τρόποι όπου μπορεί να γίνει η απαλοιφή της εποχικότητας [5]. Όταν η εποχικότητα ή περιοδικότητα μεταβάλλεται συναρτήσει του χρόνου μπορεί να προσεγγιστεί από κάποια παραμετρική συνάρτηση

$s_t = f(t)$ όπως για παράδειγμα μια ημιτονοειδής συνάρτηση. Ο δεύτερος τρόπος είναι η απαλοιφή του περιοδικού στοιχείου με βάση τις υστερήσεις d ή d διαφορές. Τέλος ο τρίτος τρόπος είναι η εκτίμηση της περιοδικότητας ή εποχικότητας σε γνωστό χρονικό διάστημα με τον κινούμενο μέσο όρο θέτοντας την τάξη ίση με την περίοδο d .

2.4.3 Απαλοιφή Τάσης και Εποχικότητας ή Περιοδικότητας

Εάν μια χρονοσειρά παρουσιάζει τάση και περιοδικότητα, τότε συνδυάζονται οι παραπάνω μέθοδοι που αναλύθηκαν στο κεφάλαιο 2.4.1 και 2.4.2 για την απαλοιφή τους ώστε η χρονοσειρά να μετατραπεί από μη στάσιμη σε στάσιμη χρονοσειρά.

2.5 Μοντέλα Πρόβλεψης Χρονοσειρών

Στην συγκεκριμένη ενότητα θα αναφερθούν μέθοδοι πρόβλεψης που χρησιμοποιούνται στις μονομεταβλητές χρονοσειρές, όπως είναι κάποια μοντέλα στοχαστικών διαδικασιών (AR, MA, ARMA, ARIMA) και το VAR μοντέλο για τις πολυμεταβλητές χρονοσειρές καθώς και το κριτήρια με τα οποία αξιολογείται μια πρόβλεψη.

2.5.1 Γραμμική Στοχαστική Διαδικασία

Μία γραμμική στοχαστική διαδικασία για κάθε χρονική στιγμή t ορίζεται ως το άθροισμα ασυσχέτιστων τυχαίων μεταβλητών (λευκός θόρυβος) όπως φαίνεται στην παρακάτω σχέση:

$$X_t = \sum_{i=-\infty}^{\infty} \varphi_i Z_{t-i} + Z_t \quad \text{όπου} \quad Z_t \sim WN(0, \sigma^2) \tag{2.5.1-1}$$

Να σημειωθεί ότι για την μελέτη των γραμμικών χρονοσειρών είναι σημαντική η συσχέτιση και όχι η εξάρτησή τους και δεν απαιτείται οι τυχαίες μεταβλητές να ακολουθούν τον IID θόρυβο αλλά απλώς τον λευκό θόρυβο (WN). Επίσης για να είναι στάσιμη χρονοσειρά θα πρέπει το άθροισμα των συντελεστών του φ_i να μην απειρίζεται, δηλαδή να ισχύει:

$$\sum_{i=-\infty}^{\infty} |\varphi_i| < \infty \tag{2.5.1-2}$$

2.5.2 Autoregressive Processes AR(p)

Μια αυτοπαλινδρομούμενη διαδικασία τάξης p ή $AR(p)$ (autoregressive process of order p) ορίζεται σύμφωνα με την παρακάτω εξίσωση [14]:

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \ldots \varphi_p X_{t-p} + Z_t \text{ όπου } Z_t \sim WN(0, \sigma^2) \tag{2.5.2-3}$$

,όπου $\varphi = (\varphi_0, \varphi_1, \varphi_2, \ldots, \varphi_p)$ είναι τα διανύσματα των συντελεστών του μοντέλου και p θετικός ακέραιος αριθμός και Z_t να ακολουθεί την κανονική κατανομή του λευκού θορύβου $Z_t \sim WN(0, \sigma^2)$. Ο κύριος στόχος των αυτοπαλινδρομούμενων διαδικασιών είναι να μπορεί

να δοθεί εξήγηση στις παρούσες τιμές μιας χρονοσειράς X_t, χρησιμοποιώντας παρελθοντικές τιμές προηγούμενων περιόδων $X_{t-1}, X_{t-2}, \ldots X_{t-p}$. Υποθέτοντας ότι η μέση τιμή $EX_t \neq 0$ τότε αντικαθιστώντας στην σχέση (2.5.2-3) όπου X_t με $X_t - \mu$ προκύπτει ότι:

$$\begin{cases} X_t - \mu = \varphi_1 \left(X_{t-1} - \mu \right) + \varphi_2 \left(X_{t-2} - \mu \right) + \ldots \varphi_p \left(X_{t-p} - \mu \right) + Z_t \\ X_t = \alpha + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \ldots \varphi_p X_{t-p} + Z_t \end{cases} \quad (2.5.2\text{-}4)$$

, όπου:

$$\alpha = \mu\left(1 - \varphi_1 - \ldots \varphi_p\right) \quad (2.5.2\text{-}5)$$

Κάνοντας χρήση του τελεστή υστέρησης η $AR(p)$ έχουμε την μορφή του χαρακτηριστικού πολυωνύμου:

$$X_t - \varphi_1 X_{t-1} - \varphi_2 X_{t-2} - \ldots - \varphi_p X_{t-p} = Z_t \quad (2.5.2\text{-}6)$$

με $BX_t = X_{t-1}$ (2.5.2-7)

$$\left(1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p\right) X_t = Z_t \quad (2.5.2\text{-}8)$$

και χρησιμοποιώντας $\varphi(B) X_t = Z_t$ (2.5.2-9)

με $\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p$ (2.5.2-10)

προκύπτει η εξίσωση της προς λύση της $AR(p)$ διαδικασίας:

$$X_t = \frac{1}{\varphi(B)} Z_t \quad (2.5.2\text{-}11)$$

Όπου στην συνέχεια εάν οι ρίζες του χαρακτηριστικού πολυωνύμου είναι εκτός του μοναδιαίου κύκλου ή αντίστοιχα οι ρίζες το πολυωνύμου

$$\lambda^p - \varphi_1 \lambda^{p^{-1}} - \ldots - \varphi_{p-1} \lambda - \varphi_p = 0 \quad (2.5.2\text{-}12)$$

είναι εντός του μοναδιαίου κύκλου τότε η διαδικασία $AR(p)$ είναι στάσιμη.

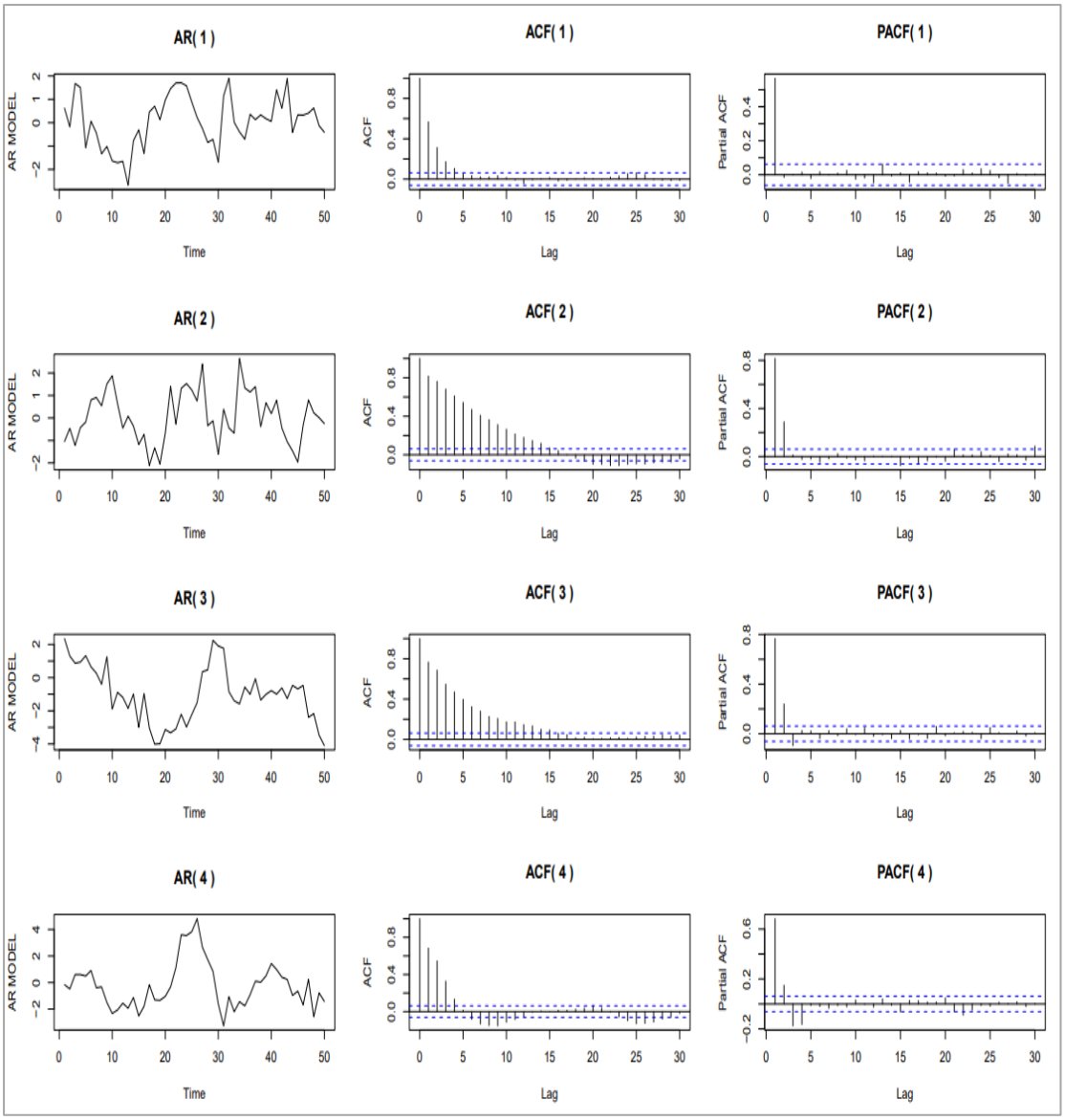
Μια αυτοπαλινδρομούμενη διαδικασία $1^{ης}$ τάξης $AR(1)$ έχει μορφή:

$$X_t = \varphi_0 + \varphi_1 X_{t-1} + \omega_t \qquad (2.5.2\text{-}13)$$

και ισχύουν τα παρακάτω:

- Όταν $-1 < \varphi_1 < 1$ τότε η χρονοσειρά είναι στάσιμη
- Όταν $\varphi_1 = 0$ τότε η χρονοσειρά είναι ισοδύναμη του λευκού θορύβου
- Όταν $\varphi_1 < 0$ τότε η χρονοσειρά φθίνει προς την μέση τιμή
- Όταν $\varphi_1 = 1$ και $\varphi_0 = 0$ τότε η χρονοσειρά είναι ισοδύναμη με τον τυχαίο περίπατο

Παρακάτω απεικονίζονται αυτοπαλινδρομούμενες διαδικασίες $1^{ης}$ τάξης $AR(1)$ με διαφορετικές κλήσεις καθώς και τα αντίστοιχα διαγράμματα αυτοσυσχετίσεων, μερικών αυτοσυσχετίσεων.



2.5.2-1 Προσομοίωση AR(q) διαδικασίας

Στα γραφήματα 2.5.2-1 έχουν προσομοιωθεί $AR(1)$ διαδικασίες με τέσσερις διαφορετικές, θετικές και αρνητικές κλίσεις με $\varphi = \{0.6,\ 0.3,\ -0.1,\ -0.2\}$ καθώς και τα αντίστοιχα διαγράμματα αυτοσυσχετίσεων.

Παρατηρείται ότι όταν η διαδικασία έχει κλίση $0 < \varphi < 1$ τότε το διάγραμμα των αυτοσυχετίσεων ACF φθίνει εκθετικά στο 0, όσο οι υστερήσεις αυξάνουν. Επίσης όταν έχει κλίσεις $-1 < \varphi < 0$ τότε το διάγραμμα των αυτοσυχετίσεων ACF μειώνεται επίσης εκθετικά στο 0 καθώς αυξάνονται οι υστερήσεις αλλά τα αλγεβρικά πρόσημα εναλλάσσονται μεταξύ θετικού και αρνητικού. Επιπλέον το διάγραμμα PACF τείνει προς το μηδέν όσο οι υστερήσεις είναι $lags > p$.

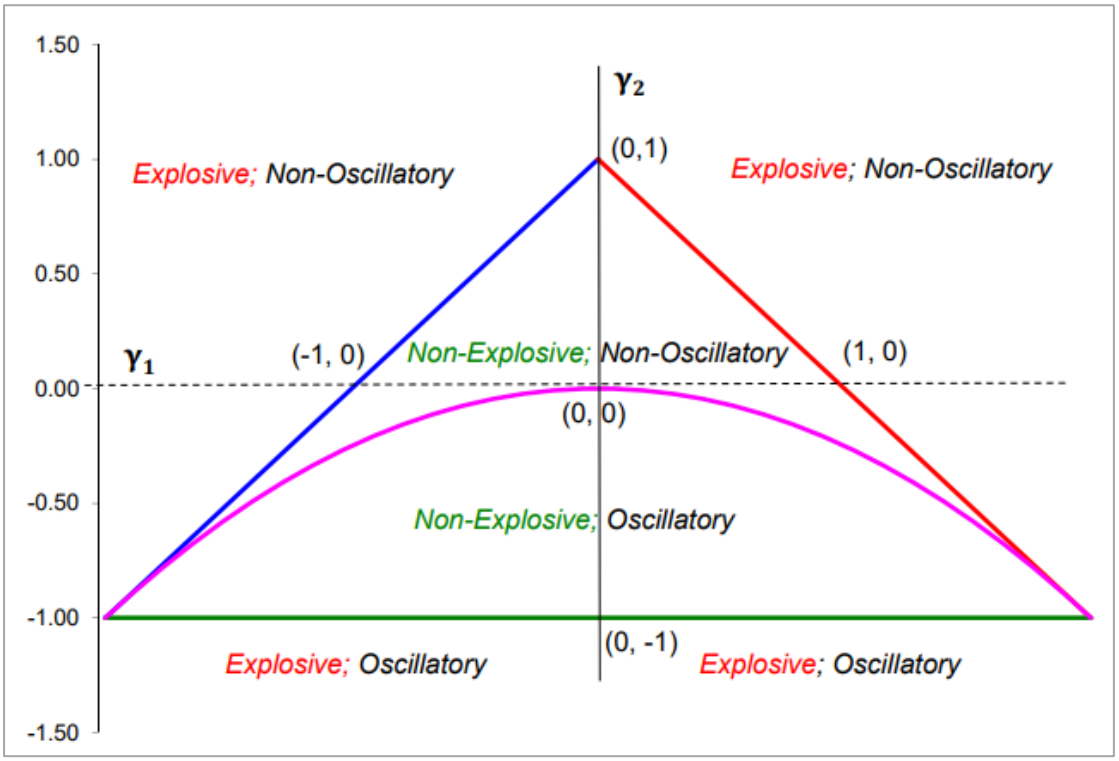
Μια αυτοπαλινδρομούμενη διαδικασία 2ης τάξης $AR(2)$ έχει μορφή [13]:

$$X_t = \varphi_0 + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \omega_t \qquad (2.5.2\text{-}7)$$

και ισχύουν τα παρακάτω:

- Όταν $\begin{cases} \varphi_2 + \varphi_1 < 1 \\ \varphi_2 - \varphi_1 < 1 \\ |\varphi_2| < 1 \end{cases}$ τότε η χρονοσειρά είναι στάσιμη. Ειδικότερα μια διαδικασία $AR(2)$ μπορεί να είναι στάσιμη με :
  - Δύο πραγματικές ρίζες αν $\varphi_1^2 + 4\varphi_2 > 0$
  - Μία διπλή πραγματική ρίζα αν $\varphi_1^2 + 4\varphi_2 = 0$
  - Δύο συζυγείς μιγαδικές ρίζες αν $\varphi_1^2 + 4\varphi_2 < 0$

Στην παρακάτω εικόνα γίνεται μια αναπαράσταση των παραπάνω, όπου η στασιμότητα μιας χρονοσειράς προσδιορίζεται όταν έχει ρίζες εντός του τριγώνου με την κόκκινη, μπλε και πράσινη γραμμή.



2.5.2-2 Διάγραμμα Στασιμότητας [13]

2.5.3 Moving Average Process MA(q)

Μια αυτοπαλινδρομούμενη διαδικασία τάξης q ή $MA(q)$ (moving average of order q) ορίζεται σύμφωνα με την παρακάτω εξίσωση [14]:

$$X_t = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q} \text{ όπου } Z_t \sim WN(0, \sigma^2) \qquad (2.5.3\text{-}1)$$

Όπου $\theta_1, \theta_2, \ldots, \theta_q$ είναι οποιοιδήποτε πραγματικοί αριθμοί, q είναι ο αριθμός υστερήσεων και Z_t να ακολουθεί την κανονική κατανομή του λευκού θορύβου $Z_t \sim WN(0, \sigma^2)$. Τα

μοντέλα Κινητού Μέσου είναι παρόμοια μοντέλα όπως τα AR με την εξαίρεση ότι κάθε τιμή συγκρίνεται με τον θόρυβο ή σφάλμα των προηγούμενων παρατηρήσεων.

Κάνοντας χρήση του τελεστή υστέρησης η $MA(q)$ έχουμε την μορφή του χαρακτηριστικού πολυωνύμου:

$$X_t = \theta(B) Z_t \tag{2.5.3-2}$$

με τελεστή υστέρησης:

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q \tag{2.5.3-3}$$

προκύπτει ο γραμμικός συνδυασμός:

$$B^k Z_t = Z_{t-k} \tag{2.5.3-4}$$

Μια αυτοπαλινδρομούμενη διαδικασία $1^{ης}$ τάξης $MA(1)$ έχει μορφή:

$$X_t = Z_t + \theta_1 Z_{t-1} \tag{2.5.3-5}$$

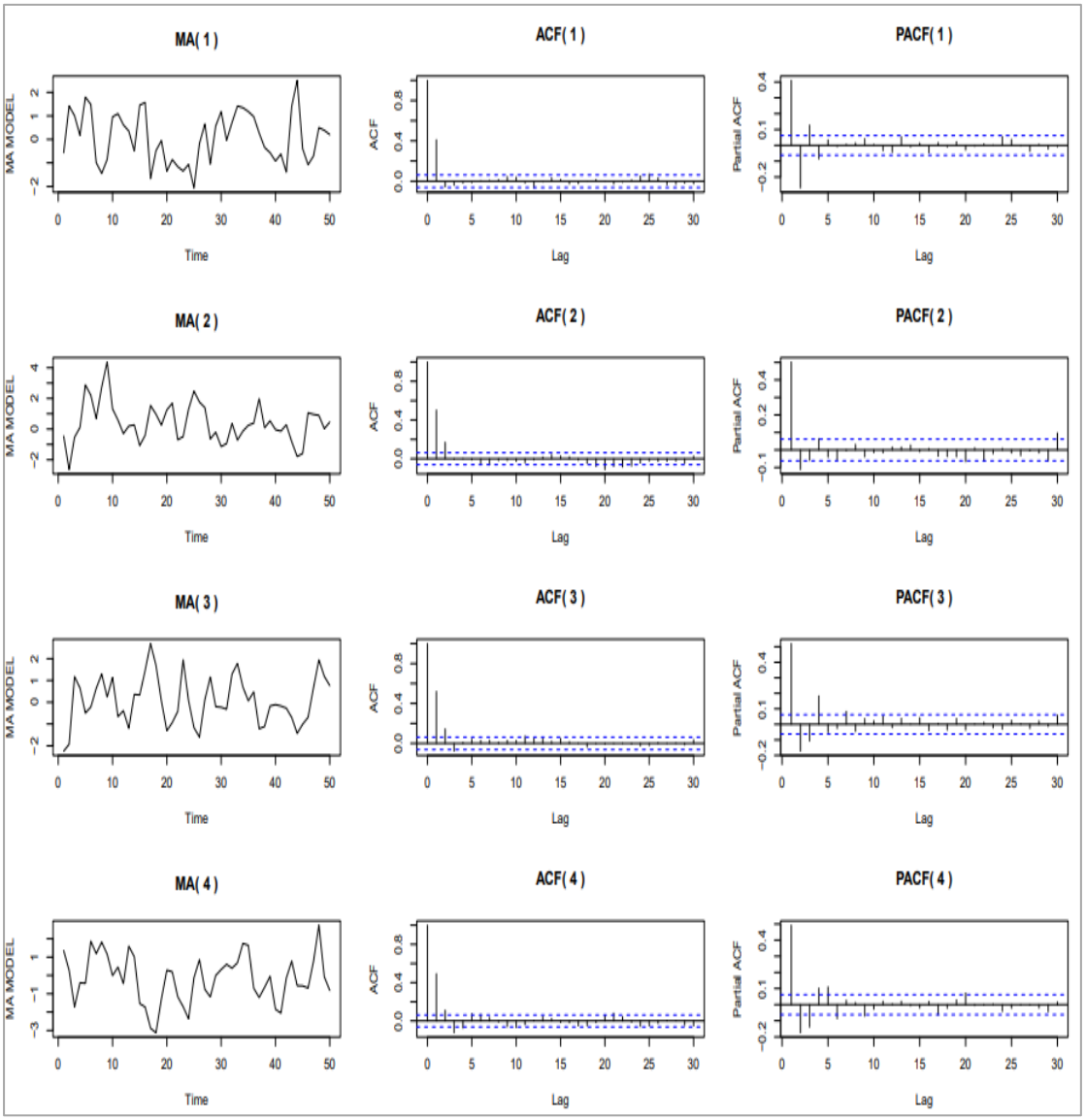
και για να είναι στάσιμη θα πρέπει $-1 < \theta_1 < 1$

Μια αυτοπαλινδρομούμενη διαδικασία $2^{ης}$ τάξης $MA(2)$ έχει μορφή:

$$X_t = Z_t + \theta_1 Z_{t-1} + + \theta_2 Z_{t-2} \tag{2.5.3-6}$$

και για να είναι στάσιμη θα πρέπει να ισχύει $\begin{cases} \theta_2 + \theta_1 < 1 \\ \theta_2 - \theta_1 < 1 \\ |\theta_2| < 1 \end{cases}$

Παρακάτω απεικονίζονται αυτοπαλινδρομούμενες διαδικασίες $1^{ης}$ τάξης $MA(1)$ με διαφορετικές κλήσεις καθώς και τα αντίστοιχα διαγράμματα αυτοσυσχετίσεων και μερικών αυτοσυσχετίσεων.



2.5.3-1 Προσομοίωση MA(q) διαδικασίας

Στα γραφήματα 2.5.3-1 έχουν προσομοιωθεί $MA(1)$ διαδικασίες με τέσσερις διαφορετικές, θετικές και αρνητικές κλίσεις με $\varphi = \{0.6,\ 0.3,\ -0.1,\ -0.2\}$ καθώς και τα αντίστοιχα διαγράμματα αυτοσυσχετίσεων.

Παρατηρείται ότι το διάγραμμα των αυτοσυχετίσεων ACF φθίνει αρκετά γρήγορα στο 0, για υστερήσεις $lags > q$, ενώ τα διαγράμματα PACF τείνουν πολύ αργά προς το μηδέν.

2.5.4 Autoregressive Moving Average (ARMA) Processes

Μια διαδικασία $ARMA$ είναι ένα μοντέλο πρόβλεψης στο οποίο συνδυάζονται οι μέθοδοι της αυτοπαλινδρομούμενης διαδικασίας $AR(p)$ και αυτοπαλινδρομούμενης διαδικασίας $MA(q)$ υποθέτοντας ότι η χρονοσειρά είναι στάσιμη και όταν υπάρχουν διακυμάνσεις, κυμαίνονται γύρω από ένα συγκεκριμένο χρονικό διάστημα [14].

Μια χρονοσειρά X_t είναι $ARMA(p,q)$ διαδικασία εάν η χρονοσειρά X_t είναι στάσιμη, δηλαδή ($-1<\varphi(z)<1$ και $-1<\theta(z)<1$) και για κάθε χρονική στιγμή t ισχύει:

$$X_t-\varphi_1X_{t-1}-\ldots-\varphi_pX_{t-p}=Z_t+\theta_1Z_{t-1}+\ldots+\theta_qZ_{t-q} \text{ με } Z_t\sim WN(0,\sigma^2) \quad (2.5.4\text{-}1)$$

με πολυώνυμα

$$\begin{cases}\varphi(z)=1-\varphi_1z-\ldots-\varphi_pz^p\\ \theta(z)=1+\theta_1z+\ldots+\theta_qz^q\end{cases} \quad (2.5.4\text{-}2)$$

,όπου δεν έχουν κοινούς παράγοντες.

Όταν τα πολυώνυμα της σχέσης (2.5.4-2) έχουν κοινούς παράγοντες τότε υπάρχουν περιττές παράμετροι με αποτέλεσμα να περιπλέκουν την ανάλυση του μοντέλου και θα πρέπει να απλοποιηθούν.

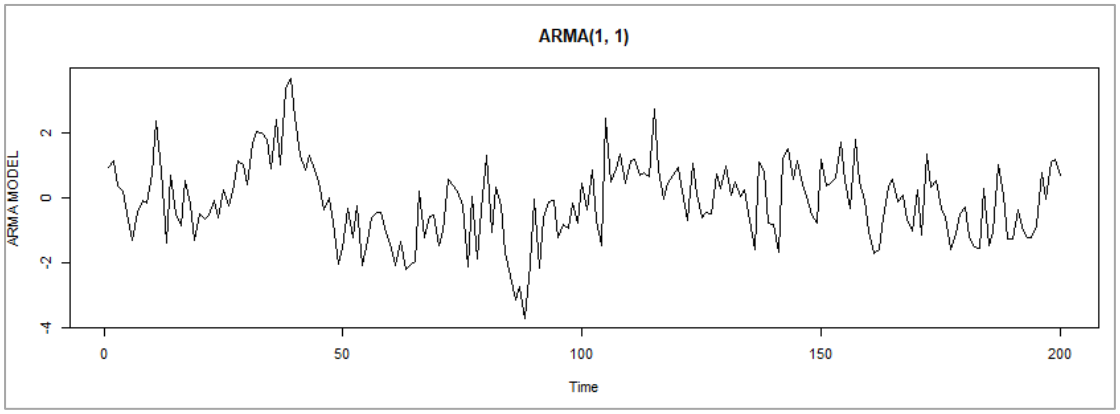
Κάνοντας χρήση του συντελεστή υστέρησης στην σχέση (2.5.4-1) προκύπτει η σχέση:

$$\varphi(B)X_t=\theta(B)Z_t \quad (2.5.4\text{-}3)$$

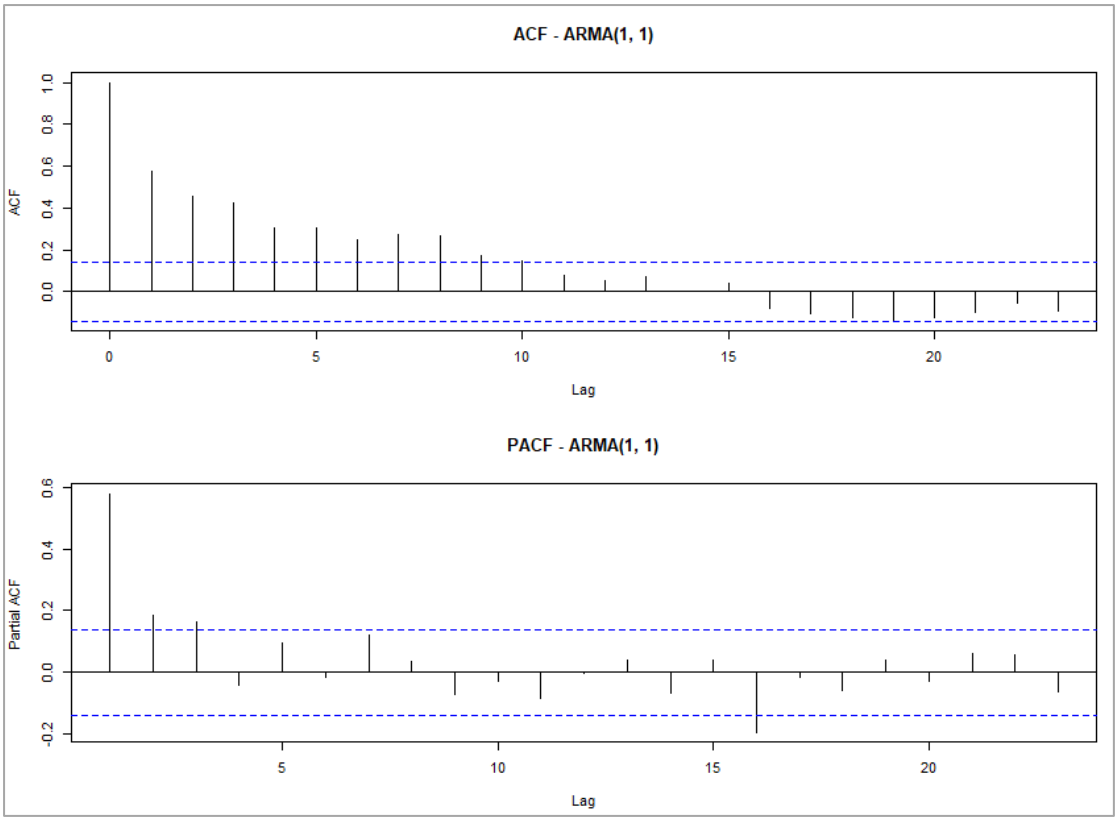
όπου, $\varphi(B)$ είναι ο αυτοπαλινδρομούμενος τελεστής και $\theta(B)$ είναι ο τελεστής τους κινούμενου μέσου της σχέσης (2.5.4-2) αντίστοιχα. Να σημειωθεί ότι όταν ο τελεστής $\varphi(B)=1$ τότε το μοντέλο $ARMA(p,q)$ είναι το ισοδύναμο με το $MA(q)$ μοντέλο, ενώ όταν ο τελεστής $\theta(B)=1$ τότε το μοντέλο $ARMA(p,q)$ είναι το ισοδύναμο με το $AR(p)$ μοντέλο. Στις παρακάτω εικόνες 2.5.4-1, 2.5.4-2 απεικονίζεται ένα $ARMA(1,1)$ μοντέλο με εξίσωση

$$X_t=0.9X_{t-1}+Z_t-0.4Z_{t-1} \quad (2.5.4\text{-}4)$$

με τα αντίστοιχα διαγράμματα αυτοσυσχετίσεων, μερικών αυτοσυσχετίσεων.



2.5.4-1 Προσομοίωση Μοντέλου ARMA(1,1)



2.5.4-2 Απεικόνιση ACF, PACF Μοντέλου ARMA(1,1)

2.5.5 Autoregressive Moving Integrated Average (ARIMA) Processes

Μια γενίκευση των $ARMA$ μοντέλων περιλαμβάνει μια κατηγορία μη στάσιμων χρονοσειρών ενσωματώνοντας την υστέρηση στο μοντέλο [14]. Ένα τέτοιο παράδειγμα μιας μη στάσιμης διαδικασίας είναι ο τυχαίος περίπατος Random Walk. Όπως έχει αναφερθεί στο κεφάλαιο 2.3.3 ο τυχαίος περίπατος Random Walk είναι μια μη στάσιμη $AR(1)$ διαδικασία με παράμετρο $\varphi = 1$ και έχει την παρακάτω μορφή:

$$X_t = X_{t-1} + Z_t \quad \text{με } Z_t \sim WN(0, \sigma^2) \tag{2.5.5-1}$$

Παρόλα αυτά η πρώτη υστέρηση

$$\nabla X_t = X_t - X_{t-1} \tag{2.5.5-2}$$

είναι μια στάσιμη διαδικασία καθώς είναι ο White Noise Z_t. Επομένως εάν συμπεριληφθεί ο WN τότε αναφερόμαστε για μια διαδικασία $ARMA(0,0)$ ή διαφορετικά για μια διαδικασία $ARIMA(0,1,0)$ όπως προκύπτει μετά την πρώτη υστέρηση της X_t.

Μια διαδικασία X_t λέμε ότι είναι $ARIMA(p,d,q)$ μοντέλο εάν:

$$\nabla^d X_t = (1-B)^d X_t \tag{2.5.5-3}$$

και το μοντέλο μπορεί να γραφεί ως

$$\varphi(B)(1-B)^d X_t = \theta(B)Z_t \text{ με } Z_t \sim WN(0,\sigma^2) \text{ και } d \geq 0 \qquad (2.5.5\text{-}4)$$

2.5.6 Vector Auto Regression (VAR) Processes

Ένα μοντέλο πολυμεταβλητής χρονοσειράς αποτελείται από πολλές εξαρτημένες μεταβλητές από τον χρόνο, σε αντίθεση με τα μοντέλα μονομεταβλητών χρονοσειρών που αποτελούνται από μια μοναδική μεταβλητή που είναι εξαρτημένη από τον χρόνο. Το VAR μοντέλο κυρίως χρησιμοποιείται για χρονοσειρές που σχετίζονται με οικονομικά, χρηματοοικονομικά δεδομένα αλλά και για προβλέψεις. Συχνά παρέχει καλύτερες προβλέψεις από μοντέλα μονομεταβλητών χρονοσειρών και από μοντέλα σύνθετων εξισώσεων. Η υπόθεση που ισχύει για τα μοντέλα πολυμεταβλητών χρονοσειρών είναι ότι οι εξαρτημένες μεταβλητές όχι μόνο εξαρτώνται από τις παρελθοντικές τιμές του χρόνου αλλά και από την μεταξύ τους εξάρτηση. Επομένως ένα VAR μοντέλο ορίζεται ως εξής [15]:

$$Y_t = c + \sum_{j=1}^{p} \Pi_j Y_{t-f} + \varepsilon_t \qquad (2.5.6\text{-}1)$$

Όπου:

$Y_t = \begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{nt} \end{bmatrix}$ είναι ένα $(n \times 1)$ διάνυσμα της χρονοσειράς.

$t = 1, 2, \ldots, T$

c : είναι οι n σταθεροί παράγοντες από τα διανύσματα των χρονοσειρών

Π_j : Είναι ο τετραγωνικός πίνακας $(n \times n)$ των μεταβλητών από $j = 1, 2, \ldots, p$

ε_t : είναι μια διαδικασία WN , με μηδενική μέση τιμή και ανεξάρτητα μεταξύ τους

Ένα παράδειγμα ενός $VAR(2)$ δίνεται από την παρακάτω μορφή:

$$\underset{(n\times 1)}{\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix}} = \underset{(n\times 1)}{\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}} + \underset{(n\times n)}{\begin{pmatrix} \pi_{11}^1 & \pi_{12}^1 \\ \pi_{21}^1 & \pi_{22}^1 \end{pmatrix}} \underset{(n\times 1)}{\begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix}} + \underset{(n\times n)}{\begin{pmatrix} \pi_{11}^2 & \pi_{12}^2 \\ \pi_{21}^2 & \pi_{22}^2 \end{pmatrix}} \underset{(n\times 1)}{\begin{pmatrix} y_{1t-2} \\ y_{2t-2} \end{pmatrix}} + \underset{(n\times 1)}{\begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}} \qquad (2.5.6\text{-}2)$$

$$\Rightarrow \begin{cases} y_{1t} = c_1 + \pi_{11}^1 y_{1t-1} + \pi_{12}^1 y_{2t-1} + \pi_{11}^2 y_{1t-2} + \pi_{12}^2 y_{2t-2} + \varepsilon_{1t} \\ y_{2t} = c_2 + \pi_{21}^1 y_{1t-1} + \pi_{22}^1 y_{2t-1} + \pi_{21}^2 y_{1t-2} + \pi_{22}^2 y_{2t-2} + \varepsilon_{2t} \end{cases} \qquad (2.5.6\text{-}3)$$

2.6 LASSO Regression – Feature Selection

Η LASSO (*Least Absolute Shrinkage and Selection Operator*) Regression είναι μία μέθοδος συρρίκνωσης καθώς είναι βασισμένη στην γραμμική παλινδρόμηση. Η Lasso παλινδρόμηση ονομάζεται και *L1 Regularization,* η οποία προσθέτει μια ποινή *(penalty level)* στην συνάρτηση κόστους. Η Lasso Regression δίνεται από τον παρακάτω τύπο [17]:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{k=1}^{p}\beta_k x_{ik} \right)^2 + \lambda \sum_{k=1}^{p}|\beta_k| \tag{2.6-1}$$

όπου, $\lambda \geq 0$ είναι μια ρυθμιστική παράμετρος.

Ορισμένοι συντελεστές της μπορούν να μηδενιστούν και να εξαλειφθούν από το μοντέλο. Αυτό έχει σαν αποτέλεσμα να δημιουργούνται απλούστερα μοντέλα εφόσον οι συντελεστές με την μεγαλύτερη τιμή *(penalty level*) να γίνονται ίσοι με το μηδέν. Από την άλλη πλευρά η Ridge Regression ή διαφορετικά *L2 Regularization,* είναι επίσης μια τεχνική συρρίκνωσης. Η κύρια διαφορά μεταξύ των δύο είναι ότι η Lasso συρρικνώνει τον συντελεστή του λιγότερο σημαντικού χαρακτηριστικού στο μηδέν, αφαιρώντας έτσι εντελώς κάποια χαρακτηριστικά. Έτσι, η μέθοδος αυτή λειτουργεί καλά για την επιλογή χαρακτηριστικών (στην περίπτωση που υπάρχει μεγάλος αριθμός χαρακτηριστικών). Ενώ η μέθοδος Ridge από την άλλη, δεν μηδενίζει τους συντελεστές αλλά μειώνει την τιμή τους, ώστε τα λιγότερο σημαντικά χαρακτηριστικά να μην έχουν μεγάλη επίδραση στο μοντέλο.

Είναι εύκολο να παρατηρήσουμε ότι όταν η παράμετρος $\lambda = 0$, τότε η συνάρτηση κόστους μετατρέπεται στην συνάρτηση ελαχίστων τετραγώνων. Εάν ο συντελεστής λ είναι αρκετά μεγάλος τότε όλο και περισσότεροι συντελεστές μηδενίζονται και επίσης έχει σαν αποτέλεσμα να οδηγούμαστε σε λάθος συμπεράσματα καθώς όταν υπάρχει υψηλή συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών, θα επιλέξει αυθαίρετα μία και θα αγνοήσει όλες τις υπόλοιπες.

2.7 Κριτήρια Αξιολόγησης της Πρόβλεψης

Για την επιλογή του καταλληλότερου προβλεπτικού μοντέλου χρησιμοποιήθηκαν τα κριτήρια Akaike's Information Criterion (AIC) και το Bayesian Information Criterion (BIC). Τα συγκεκριμένα κριτήρια αξιολογούν κατά πόσο ταιριάζει το μοντέλο που εξετάζεται συναρτήσει της πολυπλοκότητάς του. Πιο συγκεκριμένα στα μοντέλα πρόβλεψης όσο αυξάνεται η πολυπλοκότητα, δηλαδή όσο προστίθενται παράγοντες, τόσο μειώνεται και η πιθανότητα να έχουμε καλό fitting. Επιπλέον η αύξηση των παραγόντων πολλές φορές οδηγεί σε αύξηση στην διακύμανση των σφαλμάτων και συνεπώς σε έλλειψη ακρίβειας. Το μειονέκτημα που εμφανίζουν τα παραπάνω δύο κριτήρια είναι ότι δεν βασίζονται σε κάποια υπόθεση, όπως για παράδειγμα την επίτευξη μηδενικού σφάλματος παρά μόνο για το ποιο μοντέλο είναι καλύτερο προς εξέταση. Δεδομένου ότι τα κριτήρια αξιολόγησης υπολογίζονται μέσω της μέγιστης πιθανοφάνειας, θεωρείται βέλτιστο εκείνο το μοντέλο που τα ελαχιστοποιεί [12].

Το κριτήριο Akaike's Information Criterion (AIC) υπολογίζεται μαθηματικά σύμφωνα με τον παρακάτω τύπο:

$$AIC = -2\log L + 2(p+q+k+1) \tag{2.7-1}$$

,όπου στην περίπτωση που $c = 0$ τότε το $k = 0$ και όταν $c \neq 0$ τότε το $k = 1$.

Στην περίπτωση που θέλουμε να δώσουμε μεγαλύτερο βάρος στην πολυπλοκότητα χρησιμοποιείται μια παραλλαγή του παραπάνω κριτηρίου, το AIC_c και δίνεται σύμφωνα με το παρακάτω τύπο:

$$AIC_c = AIC = \frac{2 \cdot (p+q+k+1) \cdot (p+q+k+2)}{n-p-q-k-2} \tag{2.7-2}$$

,όπου n είναι το μέγεθος του δείγματος.

Το κριτήριο Bayesian Information Criterion (BIC) υπολογίζεται σύμφωνα με τον παρακάτω τύπο:

$$BIC = AIC + \log(n) \cdot (p+q+k+1) \tag{2.7-3}$$

Συμπεραίνοντας, για να επιτευχθεί η καλύτερη πρόβλεψη, πρέπει να ελαχιστοποιούνται οι τιμές των κριτηρίων AIC και BIC.

Επιπλέον σημαντικό ρόλο στην μελέτη των χρονοσειρών αποτελούν τα κατάλοιπα (residuals), όπου συγκρίνουν την πραγματικές με τις προβλεπόμενες τιμές. Παρακάτω αναφέρονται τα κυριότερα κριτήρια που έχουν προταθεί για την αξιολόγηση της προβλεπτικής ικανότητας:

Ρίζα του Μέσου Τετραγωνικού Σφάλματος (Root Mean Square Error - RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2} \tag{2.7-4}$$

Μέση Απόλυτη Απόκλιση (Mean Absolute Deviation - MAD):

$$MAD = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right| \tag{2.7-5}$$

Μέσο Τετραγωνικό Σφάλμα (Mean Square Error - MSE):

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2 \tag{2.7-6}$$

Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (Mean Absolute Percentage Error - MAPE):

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{\left(Y_t - \hat{Y}_t\right)}{Y_t} \tag{2.7-7}$$

Μέσο Ποσοστιαίο Σφάλμα (Mean Percentage Error - MPE):

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\frac{\left|Y_t - \hat{Y}_t\right|}{Y_t} \tag{2.7-8}$$

Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.

Κεφάλαιο 3ο

3.1 Δεδομένα & Προ-επεξεργασία Δεδομένων

Στο συγκεκριμένο κεφάλαιο θα γίνει εκτενής ανάλυση των δεδομένων που χρησιμοποιήσαμε, θα παρουσιαστεί η μεθοδολογία και οι αλγόριθμοι που χρησιμοποιήσαμε για την εξόρυξή των δεδομένων καθώς επίσης θα αναλυθούν τα τελικά αρχεία που εξάχθηκαν για την ανάλυση που θα πραγματοποιηθεί στο 4ο κεφάλαιο.

Τα δεδομένα άντλησης προήλθαν από το link: Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance | Zenodo [16]. Χρησιμοποιήθηκαν διάφορα σύνολα δεδομένων ώστε με την κατάλληλη προ εξεργασία και μετασχηματισμούς να καταλήξουμε στα τελικά αρχεία της ανάλυσή μας.

Πιο συγκεκριμένα τα δεδομένα που θα χρησιμοποιήσουμε από τον παραπάνω σύνδεσμο περιέχουν δεδομένα πλοίων που έχει συλλέξει το σύστημα Automatic Identification System – AIS από δέκτες, για χρονικό διάστημα έξι μηνών, από 01/10/2015 έως 31/03/2016, στην περιοχή Bay of Biscay (Βισκαϊκός κόλπος) που βρίσκεται στο δυτικό τμήμα της Γαλλίας και χωρίζονται σε τέσσερις κατηγορίες δεδομένων:

i. Navigation Data
ii. Vessel-oriented Data
iii. Geographic Data
iv. Environmental Data

3.1.1 Δεδομένα Μελέτης

Για την ανάλυσή μας θα χρησιμοποιηθούν τα παρακάτω σύνολα δεδομένων:

- [P1] AIS Data.zip
- [P1] AIS Status, Codes and Types.zip
- [E1] Ocean Conditions.zip
- [E2] Weather Conditions.zip

Να σημειωθεί ότι στην περίπτωση που υπάρχουν διπλοεγγραφές έχουν αφαιρεθεί και στα δεδομένα [nary_dynamic, oc_month, table_weather_Observation] έχει προστεθεί μια επιπλέον στήλη (id) αντίστοιχα, ώστε να μοναδικοποιείται η πληροφορία.

Από το [P1] AIS Data.zip επιλέχθηκαν τα αρχεία nari_dynamic.csv και nari_static.csv τα οποία περιέχουν τις θέσης, χαρακτηριστικά και προορισμούς των πλοίων:

nari_dynamic.csv	Row Data:18.495.677
Attribute	**Description**
* mmsi	MMSI identifier for vessel
* status	Navigational status | ** Status linked to "Navigational Status.csv"
* turn	Rate of turn, right or left, 0 to 720 degrees per minute
* speed	Speed over ground in knots (allowed values: 0-102.2 knots)
* course	Course over ground (allowed values: 0-359.9 degrees)
* heading	True heading in degrees (0-359), relative to true north
* lon	Longitude (georeference: WGS 1984)
* lat	Latitude (georeference: WGS 1984)
* t	Timestamp in UNIX epochs

3.1-1 Δεδομένα - nari_dynamic.csv

nary_static.csv	Row Data:1.029.992
Attribute	**Description**
* sourcemmsi	** corresponds to "mmsi" field in nary_dynamic.csv
* imo	IMO ship identification number (7 digits);
* callsign	International radio call sign (max 7 characters), assigned to the vessel by its country of registry
* shipname	Name of the vessel (max 20 characters)
* shiptype	Code for the type of the vessel | ** Shiptype linked to "Ship Types List.csv"
* to_bow	Distance (meters) to Bow
* to_stern	Distance (meters) to Stern --> to_bow + to_stern = LENGTH of the vessel
* to_starboard	Distance (meters) to Starboard, i.e., right side of the vessel --> to_port + to_starboard = BEAM at the vessel's nominal waterline
* to_port	Distance (meters) to Port, i.e., left side of the vessel (meters)
* eta	ETA (estimated time of arrival) in format dd-mm hh:mm (day, month, hour, minute) – UTC time zone
* draught	Allowed values: 0.1-25.5 meters
* destination	Destination of this trip (manually entered)
* mothershipmmsi	Dimensions of ship in meters and reference point for reported position
* t	timestamp in UNIX epochs

3.1-2 Δεδομένα - nary_static.csv

Μια αναπαράσταση των δεδομένων nari_dynamic φαίνεται στην παρακάτω εικόνα:

3.1.1-1 Δεδομένα nari_dynamic.csv

Από το αρχείο AIS Status, Codes and Types.zip επιλέχθηκαν τα αρχεία Navigational_Status.csv, Ship_Types_List.csv και Ship_Types_Detailled_List.csv και περιέχουν δεδομένα αναφορικά με το status και τον τύπο των πλοίων:

Navigational Status.csv	Row Data:16
Attribute	**Description**
* code	Unique identifier of the navigational status | ** code is linked to "status" field in nari_dynamic.csv
* status	Detailed description of the status

3.1-3 Δεδομένα - Navigational_Status.csv

Ship_Types_List.csv	Row Data:38
Attribute	**Description**
* id_shiptype	Unique identifier of the record | ** id_shiptype is linked to id_shiptype field in Ship_Types_Detailled_List.csv
* shiptype_min	min value of the given shiptype
* shiptype_max	max value of the given shiptype
* type_name	type name
* ais_type_summary	AIS summary of the given type

3.1-4 Δεδομένα - Ship_Types_List.csv

Ship_Types_Detailled List.csv	Row Data:233
Attribute	**Description**
* id_detailedtype	Unique identifier of the record
* detailed_type	Detailed description of the given type
* id_shiptype	** corresponds to "id_shiptype" field in Ship Types List.csv

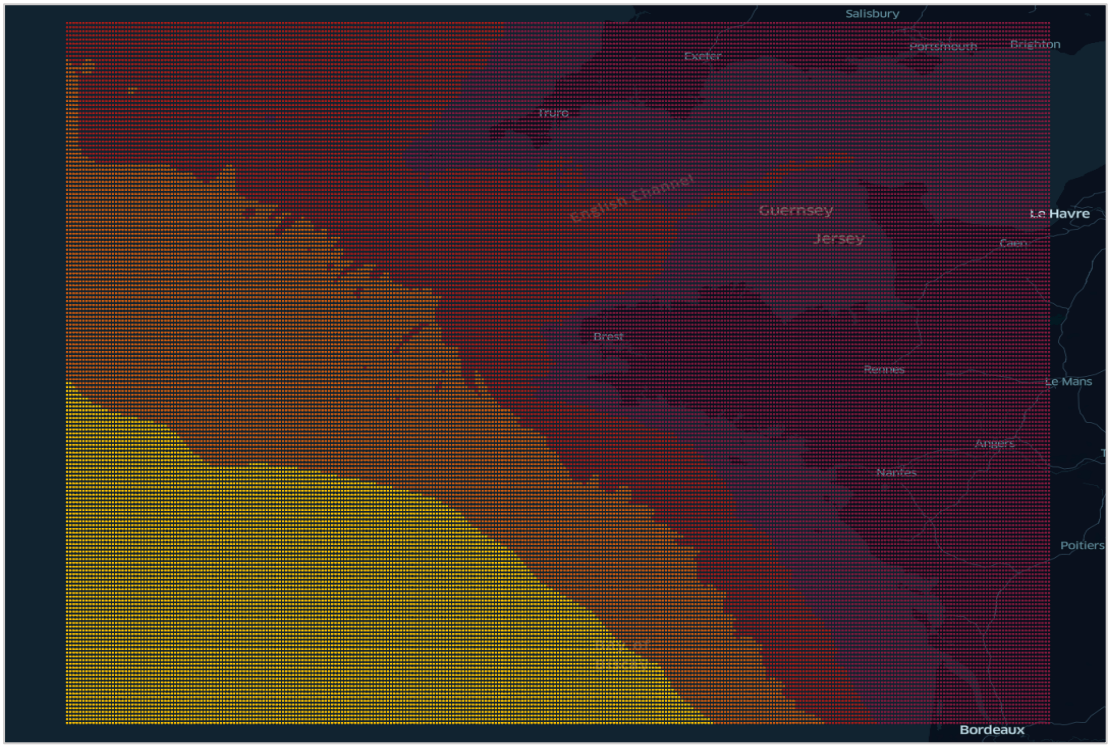
3.1-5 Δεδομένα - Ship_Types_Detailled_List.csv

Από το [E1] Ocean Conditions.zip επιλέχθηκαν τα αρχεία oc_october.csv, oc_november.csv, oc_december.csv, oc_january.csv, oc_february.csv, oc_march.csv και oc_january.csv τα οποία περιέχουν δεδομένα ωκεανού που καλύπτουν από -10.00 έως 0.00 γεωγραφικό μήκος και από 45.00 έως 51.00 γεωγραφικό πλάτος. Κάθε παρατήρηση με την κοντινότερή της απέχει 0.03330 γεωγραφικό πλάτος και 0.03340 γεωγραφικό μήκος και προκύπτει καινούργια παρατήρηση ανά τρείς ώρες.

oc_month.csv	Row Data:66.194.415
Attribute	**Description**
* **lat**	longitude in degrees (-180.0 to 180.0)
* **lon**	latitude in degrees (-90.0 to 90.0)
* **dpt**	bottom depth in meters. (Undefined value=-16384)
* **wlv**	sea surface height above sea level in meters (=>tidal effect). (Undefined value=-327.67)
* **hs**	significant height of wind and swell waves (=>see state) (Undefined value=--65.534)
* **lm**	mean wave length in meters. (Undefined value=-32767)
* **dir**	wave mean direction ("from direction") (Undefined value=-3276.7)
* **ts**	unix timestamp - number of seconds since 1970/01/01

3.1-6 Δεδομένα - oc_month.csv

Μια αναπαράσταση των δεδομένων oc_month.csv απεικονίζονται στην παρακάτω εικόνα, όπου όσο πιο ανοικτό είναι το χρώμα τόσο μεγαλύτερο είναι το βάθος του ωκεανού:



3.1.1-2 Δεδομένα oc_month.csv

Τέλος, από το [E2] Weather Conditions.zip επιλέχθηκαν τα αρχεία table_wheatherObservation.csv, table_weatherStation.csv και table_windDirection.csv, τα οποία περιέχουν δεδομένα καιρού και ανέμου και καλύπτουν από -10.00 έως 0.00 γεωγραφικό μήκος και από 45.00 έως 51.00 γεωγραφικό πλάτος. Κάθε καινούργια παρατήρηση προκύπτει ανά μια ώρα για τον κάθε ένα από τους 16 μετεωρολογικούς σταθμούς:

table_weatherObservation.csv	Row Data:71.515
Attribute	**Description**
* id_station	id of the stations | ** id_station is linked to table_weatherStation.csv
* local_time	unix timestamp - number of seconds since "01/01/1970"
* T	Air temperature (°C) at 2-meter height above the earth's surface
* Tn	Minimum air temperature (°C) during the past period (not exceeding 12 hours)
* Tx	Maximum air temperature (°C) during the past period (not exceeding 12 hours)
* P	Atmospheric pressure reduced to mean sea level (mm of mercury)
* U	Relative humidity (%) at a height of 2 meters above the earth's surface
* id_windDirection	Mean wind direction (compass point) at a height of 10-12 meters above the earths surface over the 10-minute period immediately preceding the observation | ** id_windDirection is linked to table_weatherStation.csv
* Ff	Mean wind speed at a height of 10-12 meters above the earth’s surface over the 10-minute period immediately preceding the observation (meters per seconds)

* ff10	Maximum gust value at a height of 10-12 meters above the earth's surface over the 10-minute period immediately preceding the observation (meters per seconds)
* ff3	Maximum gust value at a height of 10-12 meters above the earth's surface between the periods of observations (meters per seconds)
* VV	Horizontal visibility (km)
* Td	Dewpoint temperature at a height of 2 meters above the earth's surface (°C)
* RRR	Amount of precipitation (mm)
* tR	The period of time during which the specified amount of precipitation was accumulated

3.1-7 Δεδομένα - table_weatherObservation.csv

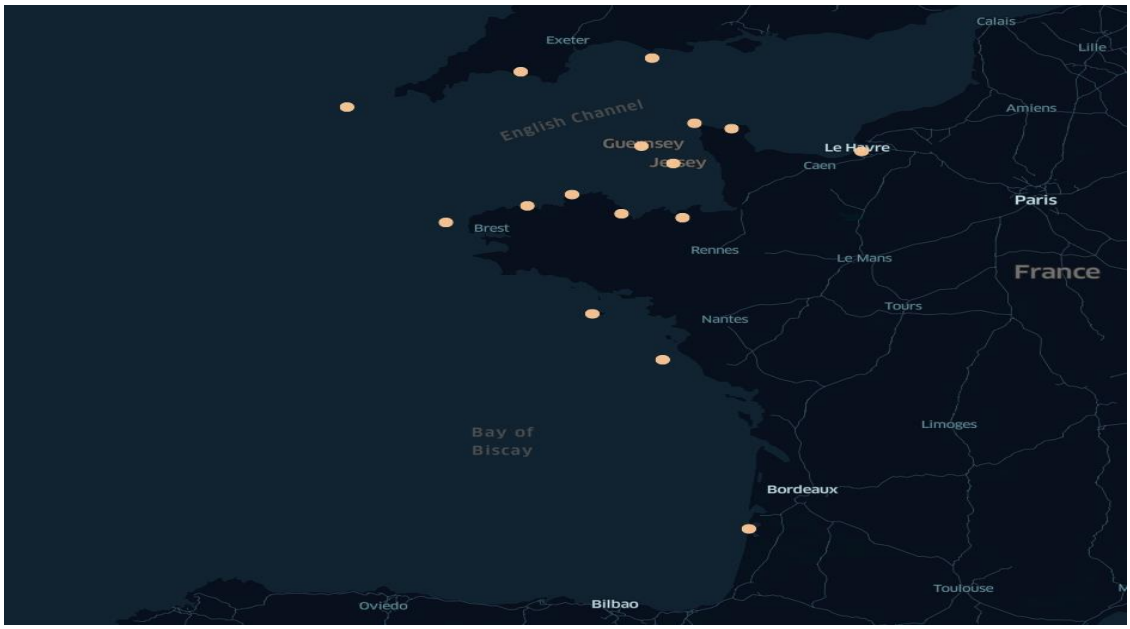
table_weatherStation.csv	Row Data:16
Attribute	**Description**
* id_station	** corresponds to "id_station" field in table_weatherObservation.csv
* station_name	weather station name
* latitude	latitude of weather station
* longitude	longitude of weather station
* elevation	elevation

3.1-8 Δεδομένα - table_weatherStation.csv

table_windDirection.csv	Row Data:19
Attribute	**Description**
* id_windDirection	** corresponds to "id_ windDirection " field in table_weatherObservation.csv
* DD_num	value of wind direction
* DD_plainText	long description of wind direction
* DD_shortText	short description of wind direction

3.1-9 Δεδομένα - table_windDirection.csv

Μια απεικόνιση των μετεωρολογικών σταθμών αποτυπώνεται στην παρακάτω εικόνα:



3.1.1-3 Τοποθεσίες Μετεωρολογικών Σταθμών

3.1.2 Μεθοδολογία Επεξεργασία Δεδομένων

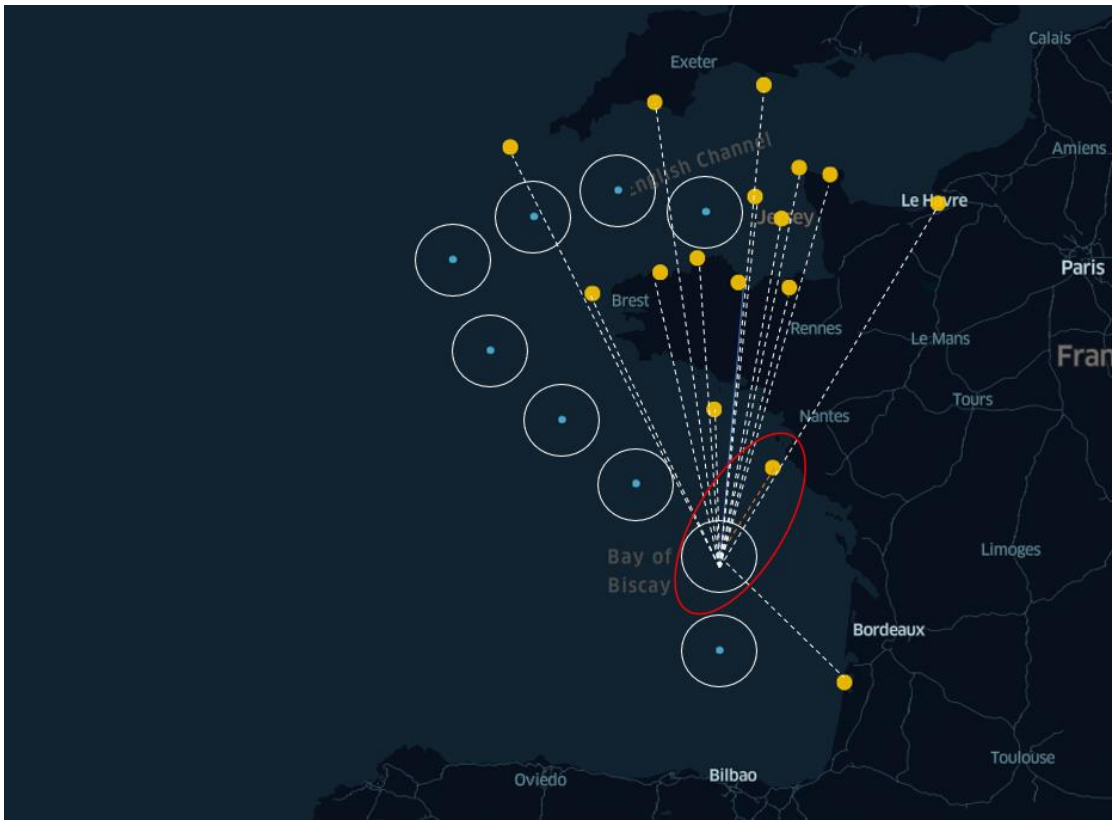
Με βάση τα παραπάνω σύνολα δεδομένων που αναλύθηκαν στο κεφάλαιο 3.1.1 θα δημιουργηθούν δύο αρχεία για τους σκοπούς της ανάλυσής μας τα οποία θα είναι τα παρακάτω:

i. Το 1ο σύνολο δεδομένων που θα εξαχθεί, θα περιλαμβάνει δεδομένα ταχύτητας ενός πλοίου που θα επιλεχθεί για το χρονικό διάστημα από 2015-10-01 έως 2016-03-31.

ii. Το 2ο σύνολο δεδομένων που θα εξαχθεί, θα περιλαμβάνει δεδομένα ταχύτητας ενός πλοίου που θα επιλεχθεί για το χρονικό διάστημα από 2015-10-01 έως 2016-03-31 καθώς επίσης και δεδομένα όπου βάσει της πορείας του πλοίου θα επιλέγονται τα κοντινότερα διαθέσιμα δεδομένα καιρού από τους 16 σταθμούς της περιοχής και επίσης θα επιλέγονται τα κοντινότερα δεδομένα ωκεανού βάσει της γεωγραφικής τοποθεσίας του πλοίου.

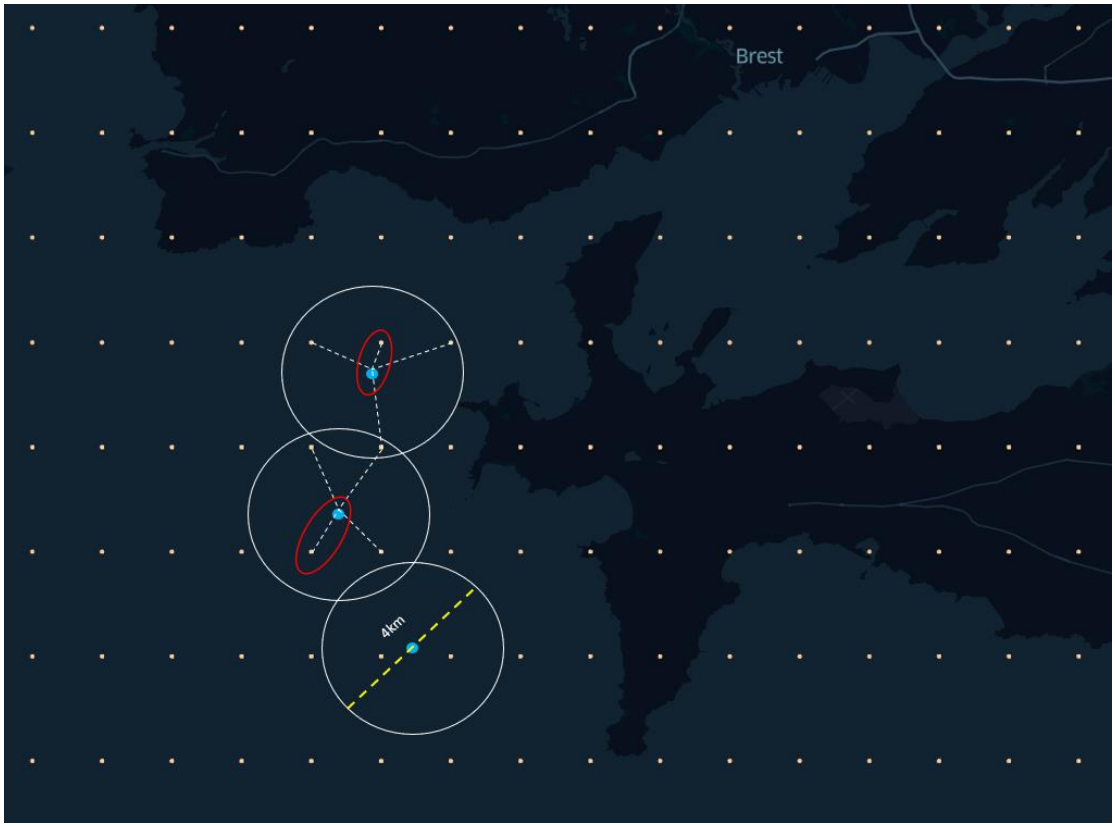
Για την επιλογή των κοντινότερων δεδομένων καιρού ή ωκεανού, ακολουθήθηκαν δύο διαφορετικές τεχνικές. Για την κάθε τοποθεσία του πλοίου επιλέγεται η κοντινότερη απόσταση της τοποθεσίας πλοίου από όλους τους μετεωρολογικούς σταθμούς.

Στην εικόνα 3.1.2-1 απεικονίζεται ο τρόπος με τον οποίο γίνεται η συλλογή των δεδομένων καιρού. Κατά την πορεία του πλοίου (μπλε κουκίδες) υπολογίζονται οι κοντινότερες αποστάσεις από όλους τους μετεωρολογικούς σταθμούς (κίτρινες κουκίδες), και λαμβάνουμε την τελευταία χρονικά διαθέσιμη την πληροφορία καιρού. Επομένως, το πλοίο στην παρακάτω εικόνα θα λάβει την πληροφορία από τον σταθμό που βρίσκεται στο κόκκινο περίγραμμα.



3.1.2-1 Επιλογή Δεδομένων από Μετεωρολογικούς Σταθμούς

Στην εικόνα 3.1.2-2 απεικονίζεται ο τρόπος με τον οποίο γίνεται η συλλογή των δεδομένων ωκεανού. Κατά την πορεία του πλοίου (μπλε κουκίδες) επιλέγεται μια ακτίνα 4km διότι οι παρατηρήσεις (κίτρινες κουκίδες) απέχουν 0.03330 γεωγραφικό πλάτος και 0.03340 γεωγραφικό μήκος. Επομένως λαμβάνουμε την πιο πρόσφατη πληροφορία εντός της ακτίνας με την κοντινότερη απόσταση όπως φαίνεται με το κόκκινο περίγραμμα.



3.1.2-2 Επιλογή Δεδομένων Ωκεανού

3.1.3 Αλγόριθμος Διαχωρισμού Χώρου

Για τον υπολογισμό της κοντινότερης διαθέσιμης πληροφορίας από τα δεδομένα ωκεανού, κατά την πορεία του πλοίου, χρησιμοποιήθηκαν τα παρακάτω σύνολα δεδομένων:

i. Dataset_A: από το σύνολο δεδομένων 3.1-1 Δεδομένα - nari_dynamic.csv έγινε επιλογή των πεδίων (lon, lat, t) για ένα συγκεκριμένο πλοίο και γίνεται προσθήκη στήλης με ονομασία (id) όπου είναι ένας αύξων αριθμός και της στήλης (dataset_name) όπου περιέχει το όνομα του συνόλου δεδομένων.

ii. Dataset_B: από τα σύνολα δεδομένων 3.1-6 Δεδομένα - oc_month.csv έγινε επιλογή των πεδίων (lon, lat, ts) και γίνεται προσθήκη στήλης με ονομασία (id) όπου είναι ένας αύξων αριθμός και η στήλη (dataset_name) όπου περιέχει το όνομα του συνόλου δεδομένων.

Επομένως προκύπτουν τα παρακάτω σύνολα δεδομένων:

DATASET - A		DATASET - B	
id	ID δεδομένων Πλοίου	id	ID δεδομένων Ωκεανού
lon	Γεωγραφικό Μήκος Πλοίου	lon	Γεωγραφικό Μήκος Ωκεανού
lat	Γεωγραφικό Πλάτος Πλοίου	lat	Γεωγραφικό Πλάτος Ωκεανού
ts	timestamp Πλοίου	ts	timestamp Ωκεανού
datset_name	Ονομασία Dataset Πλοίου	datset_name	Ονομασία Dataset Ωκεανού

3.1-10 Δεδομένα Επιλογής Αλγόριθμου

Ο σκοπός του αλγόριθμου είναι να βοηθήσει στην μείωση των υπολογισμών απόστασης μεταξύ του (lon,lat) του πλοίου με το (lon,lat) πληροφορίας του ωκεανού και παρουσιάζεται παρακάτω:

ALGORITHM: HORIZONTAL PARTITIONING

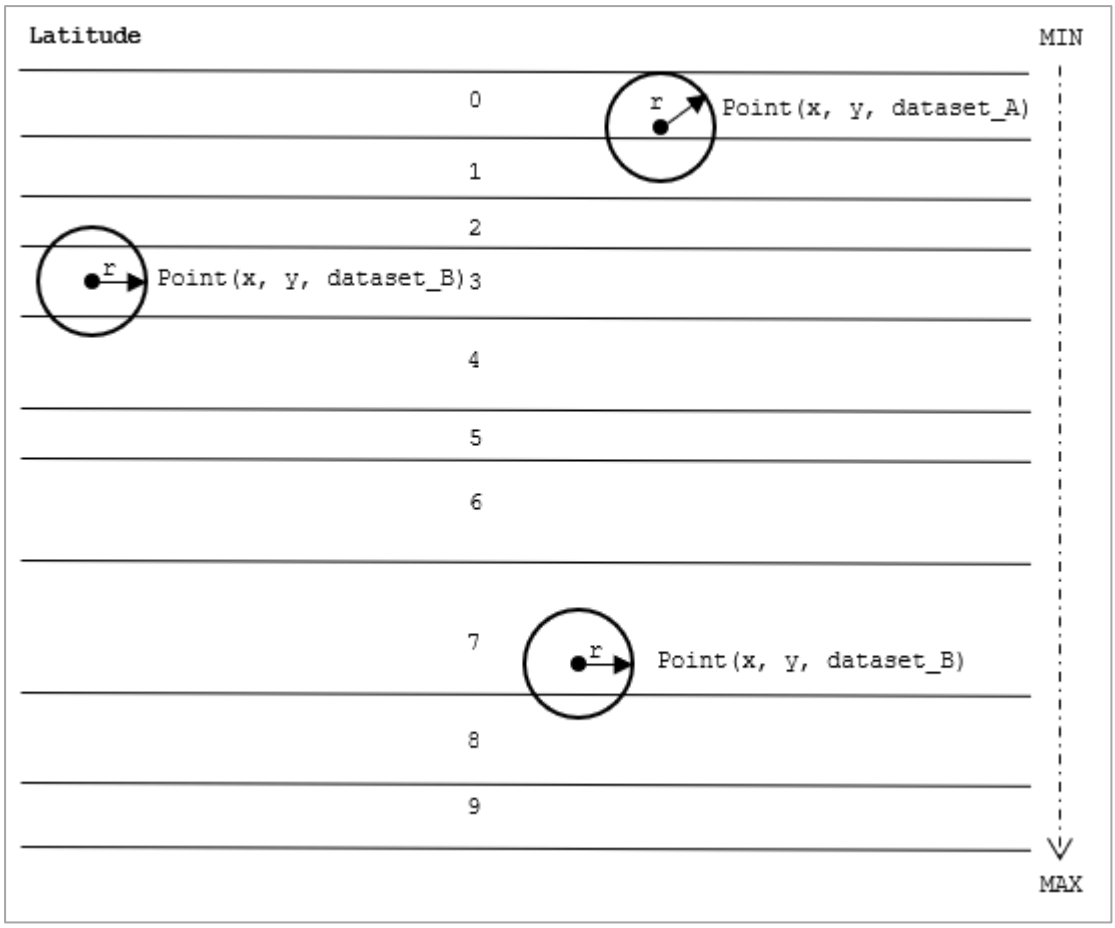
```
1   Input (dataset_A, dataset_B, r, splits)
2   Function Horizontal_Partitioning:
3       union dataset_A, dataset_B as dataset
4       sort dataset(lat)  // sort data by latitude ascending
5       size ← int(count(df) / number_of _partitions)
6       for x in dataset do
7           partition_id = int(row_number(x) / size)
8       end for
9       // Duplicate points of dataset B to the nearest horizontal cells
10      maxlatitude ← max_latitude(partition_id)
11      for x in dataset do
12          if x in dataset ← B  then
13              if x(lat) – r < maxlatitude or x(lat)+r > maxlatitide then
14                  Duplicate points to the horizontal cells
15              end if
16          end if
17      end for
18  end Function
```

3.1.3-1 Αλγόριθμος Οριζόντιας Διαμέρισης Χώρου

Αρχικά γίνεται ένωση των δύο συνόλων δεδομένων και τοποθετούνται οι εγγραφές σε αύξουσα σειρά με βάση το γεωγραφικό πλάτος (lat). Στην συνέχεια γίνεται η διαμέριση του χώρου σε οριζόντιες γραμμές, όπου στην συγκεκριμένη περίπτωση χωρίστηκε ο χώρος σε 10 επίπεδα. Για την κατάταξη του κάθε σημείου στον χώρο, δημιουργήσαμε μια επιπλέον στήλη (increasing_id) όπου είναι μια αύξουσα αρίθμηση από το 0 έως το μέγεθος του συνόλου δεδομένων και στην συνέχεια διαιρώντας την κάθε τιμή (increasing_id) με το σύνολο των γραμμών, μπορούμε να αποφανθούμε σε ποια χωρική διαμέριση (horizontal_cell) ανήκει το κάθε σημείο.

Σε κάποιες διαμερίσεις του χώρου πιθανότατα να έχουμε κάποια σημεία τα οποία να είναι κοντά μεταξύ τους σύμφωνα με την συγκεκριμένη ακτίνα που έχει δοθεί. Επομένως, σε αυτές τις περιπτώσεις θα δημιουργηθούν διπλότυπες εγγραφές και θα αντιγραφούν στις κοντινότερες διαμερίσεις του χώρου.

Σύμφωνα με τα παραπάνω ο χώρος έχει διαμεριστεί σε 10 επίπεδα. Στην εικόνα 3.1.3-2 παρατηρούμε ότι σημείο που βρίσκεται στο 3ο επίπεδο πρέπει να αντιγραφεί και στο 2ο, 4ο επίπεδο αντίστοιχα, διότι χρειαζόμαστε τα ζευγάρια των σημείων από το Dataset_A, εάν υπάρχουν εντός του δοθέντος κύκλου.



3.1.3-2 Τρόπος Δημιουργίας Διπλότυπων

Σύμφωνα με τον αλγόριθμο για την δημιουργία των διπλότυπων, αρχικά πρέπει να υπολογιστούν οι μεγαλύτερες τιμές γεωγραφικού πλάτους (lat) για κάθε διαμέριση χώρου. Στην συνέχεια γίνεται ο έλεγχος για κάθε σημείο από το Dataset_B, εάν η απόστασή του από την μεγαλύτερη ή μικρότερη τιμή του γεωγραφικού πλάτους της διαμέρισης που ανήκει το σημείο είναι εντός του κύκλου, τότε θα δημιουργείται διπλή εγγραφή στις κοντινότερες διαμερίσεις, διαφορετικά δεν θα δημιουργείται διπλότυπη εγγραφή.

Επίσης είναι σημαντικό να αναφέρουμε ότι στο επίπεδο 0, δεν θα πραγματοποιηθεί κάποια διπλοεγγραφή στο επίπεδο 1 διότι μόνο τα δεδομένα από το Dataset_B μπορούν να αντιγραφούν.

Τέλος, γνωρίζοντας πλέον το κάθε σημείο σε ποια διαμέριση χώρου ανήκει, διαχωρίζουμε τα δύο σύνολα δεδομένων βάσει του (dataset_name) και τα ενώνουμε με το (horizontal_cell) ώστε να βγάλουμε την τελική απαιτούμενη πληροφορία που θέλουμε για την ανάλυσή μας.

3.2 Τελικά Δεδομένα Ανάλυσης

Μετά την προ-επεξεργασία των δεδομένων που ακολουθήσαμε στην ενότητα 3 καταλήγουμε στα τελικά σύνολα δεδομένων που θα έχει η ανάλυσή μας.

DATASET - I		DATASET - II		Features No.
ts	Vessel timestamp	vsl_ts	Vessel timestamp	1
vld_speed	Speed of Vessel	vsl_status	Status of Vessel	2
		vsl_turn	Rate of turn, right or left, 0 to 720 degrees per minute	3
		vsl_course	Course over ground (allowed values: 0-359.9 degrees)	4
		vsl_heading	True heading in degrees (0-359), relative to true north	5
		wthr_air_temp	Air temperature (°C) at 2-meter height above the earth's surface	6
		wthr_min_air_temp	Minimum air temperature (°C) during the past period (not exceeding 12 hours)	7
		wthr_max_air_temp	Maximum air temperature (°C) during the past period (not exceeding 12 hours)	8
		wthr_atm_pres	Atmospheric pressure reduced to mean sea level (mm of mercury)	9
		wthr_humidity	Relative humidity (%) at a height of 2 meters above the earth's surface	10
		wthr_min_speed	Mean wind speed at a height of 10-12 meters above the earth’s surface over the 10-minute period immediately preceding the observation (meters per seconds)	11
		wthr_max_gus_vl1	Maximum gust value at a height of 10-12 meters above the earth’s surface over the 10-minute period immediately preceding the observation (meters per seconds)	12
		wthr_max_gus_vl2	Maximum gust value at a height of 10-12 meters above the earth’s surface between the periods of observations (meters per seconds)	13
		wthr_hor_vsblty	Horizontal visibility (km)	14
		wthr_dewpoint_temp	Dewpoint temperature at a height of 2 meters above the earth's surface (°C)	15
		wthr_amt_prcptn	Amount of precipitation (mm)	16
		wthr_tm_prcptn	The period of time during which the specified amount of precipitation was accumulated	17
		wind_value	value of wind direction	18
		ocn_depth	bottom depth in meters. (Undefined value=-16384)	19
		ocn_srf_hght	sea surface height above sea level in meters (=>tidal effect). (Undefined value=-327.67)	20
		ocn_waves_hght	significant height of wind and swell waves (=>see state) (Undefined value=--65.534)	21
		ocn_wave_len	mean wave length in meters. (Undefined value=-32767)	22
		ocn_wave_dir	wave mean direction ("from direction") (Undefined value=-3276.7)	23
		vsl_speed	Speed of Vessel	24

3.2-1 Σύνολα Δεδομένων Timeseries Analysis

Κεφάλαιο 4ο

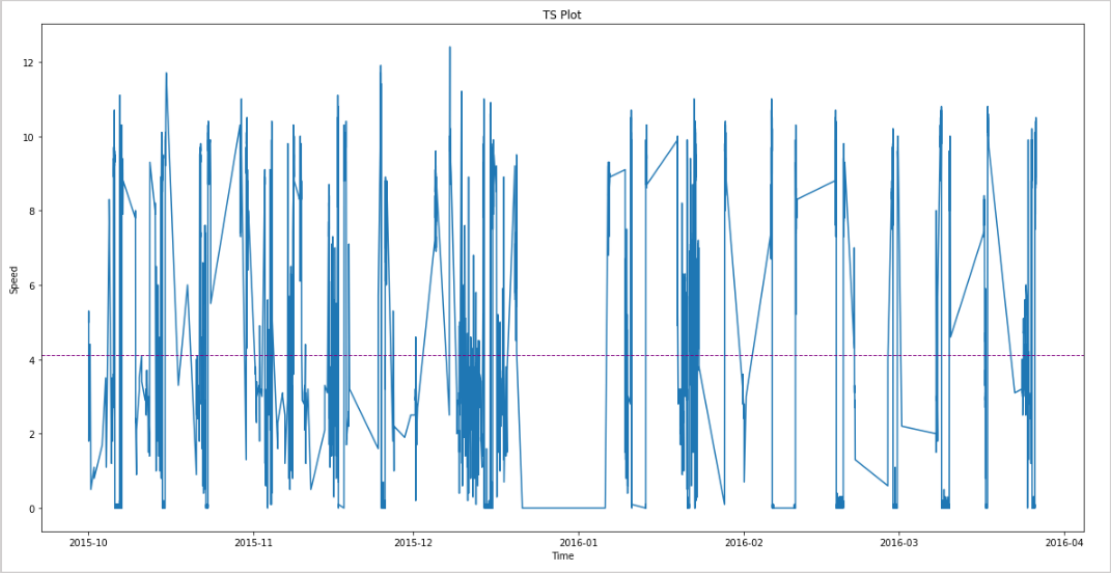
Στο κεφάλαιο αυτό θα γίνει και ανάλυση των αποτελεσμάτων για το ARIMA μοντέλο που εφαρμόστηκε με το σύνολο δεδομένων «DATASET I» και το VAR μοντέλου που εφαρμόστηκε για το σύνολο δεδομένων «DATASET II». Τέλος, θα αναλυθούν τα συμπεράσματα και θα γίνει η εκτίμηση του καλύτερου μοντέλου.

4.1 ARIMA Model - DATASET I

Το σύνολο δεδομένων το οποίο χρησιμοποιήθηκε για την ανάλυση του ARIMA μοντέλου ήταν το «DATASET I» και αποτελούνταν από 98.384 εγγραφές. Τα μέτρα θέσης και διασποράς θα αποτυπωθούν παρακάτω καθώς επίσης και το διάγραμμα της χρονοσειράς.

	speed
count	98384.000000
mean	4.120351
std	4.341863
min	0.000000
25%	0.000000
50%	1.700000
75%	9.000000
max	12.400000

4.1-1 Μέτρα Θέσης - Διασποράς | Dataset I - Before Resampling



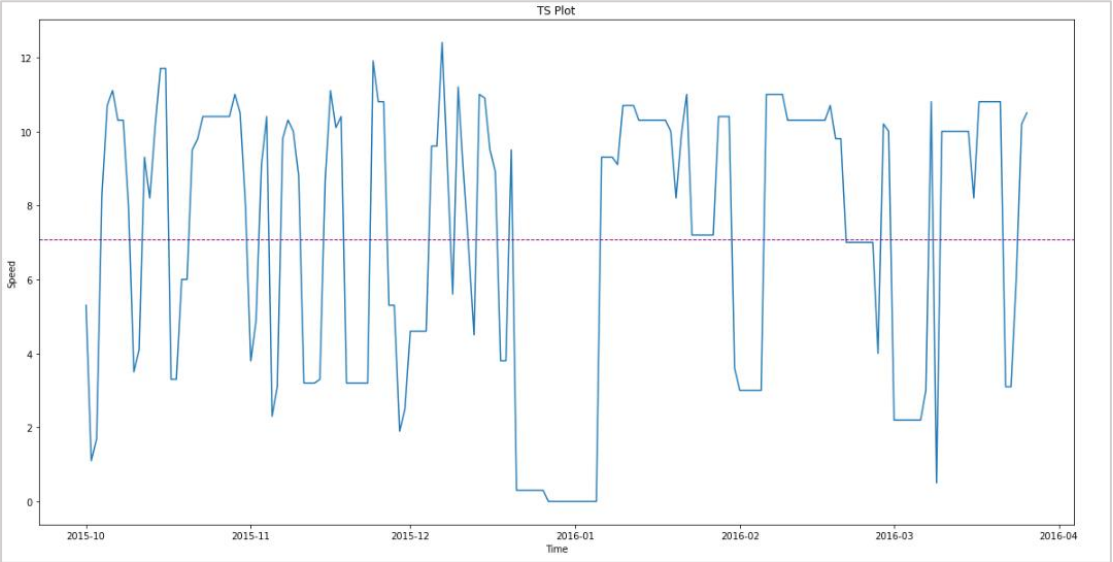
4.1-2 TS Plot | Dataset I - Before Resampling

Άρα τα δεδομένα έχουν μέση τιμή ίση με 4.12, διακύμανση 4.34, μέγιστη τιμή ταχύτητας 12.4 knots και ελάχιστη τιμή ταχύτητας 0 knots. Η διάμεσος είναι ίση με 1.7 knots και το 75% των παρατηρήσεων είναι μεγαλύτερες από 0 knots και το 25% των παρατηρήσεων είναι μεγαλύτερες από 9 knots.

Παρατηρήθηκε ότι τα δεδομένα δεν είχαν μια συγκεκριμένη συχνότητα (δηλαδή σε μια ημέρα πιθανότατα να έχουμε μια παρατήρηση αλλά μπορεί να έχουμε και περισσότερες), έγινε αναγωγή των δεδομένων στην ημέρα ώστε να λαμβάνουμε την μεγαλύτερη ταχύτητα που είχε εντός μιας ημέρας με αποτέλεσμα να είναι εφικτή η ανάλυση της χρονοσειράς. Μετά από την αναπροσαρμογή των δεδομένων, προκύπτει ένα σύνολο δεδομένων από 178 εγγραφές, δηλαδή μια για κάθε ημέρα και αυτές οι εγγραφές θα είναι οι τελικές που θα χρησιμοποιηθούν στο μοντέλο. Επομένως, τα μέτρα θέσης και διασποράς του νέου συνόλου δεδομένων αποτυπώνονται παρακάτω καθώς επίσης και το διάγραμμα της χρονοσειράς:

	speed
	max
count	178.000000
mean	7.089326
std	3.782243
min	0.000000
25%	3.225000
50%	8.900000
75%	10.300000
max	12.400000

4.1-3 Μέτρα Θέσης - Διασποράς | Dataset I - After Resampling



4.1-4 TS Plot | Dataset I - After Resampling

Άρα τα δεδομένα έχουν μέση τιμή ίση με 7.08, διακύμανση 3.78, μέγιστη τιμή ταχύτητας 12.4 knots και ελάχιστη τιμή ταχύτητας 0 knots. Η διάμεσος είναι ίση με 8.9 knots και το 75% των παρατηρήσεων είναι μεγαλύτερες από 3.2 knots και το 25% των παρατηρήσεων είναι μεγαλύτερες από 10.3 knots.

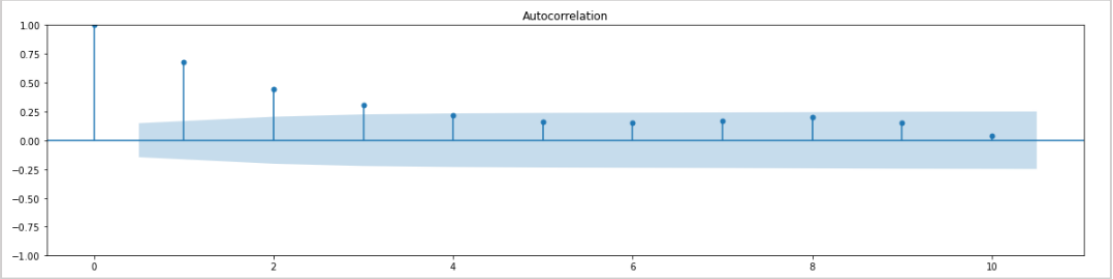
Στην συνέχεια πραγματοποιήθηκε έλεγχος στασιμότητας της χρονοσειράς με βάση το ADF test.

```
   Augmented Dickey-Fuller Test on "('speed', 'max')"
   -----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level    = 0.05
Test Statistic        = -5.7362
No. Lags Chosen       = 0
Critical value 1%     = -3.468
Critical value 5%     = -2.878
Critical value 10%    = -2.576
=> P-Value = 0.0. Rejecting Null Hypothesis.
=> Series is Stationary.
```

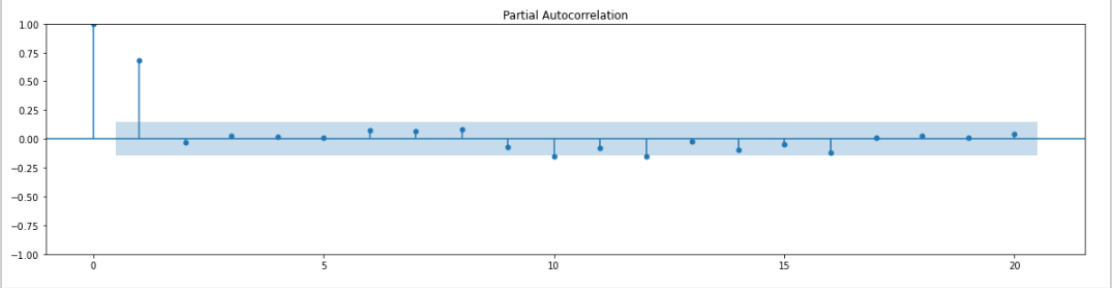
4.1-5 ADF Test

Σύμφωνα με την εικόνα 4.1-5 επειδή το $p-value<0.05$ η χρονοσειρά είναι στάσιμη.

Επιπροσθέτως έγινε υπολογισμός των συντελεστών αυτοσυσχέτισης και μερικής αυτοσυσχέτισης καθώς βοηθάει στην αποτίμηση μιας χρονοσειράς ως προς την τυχαιότητα και την στασιμότητα. Σε μια τυχαία χρονοσειρά το 95% των συντελεστών αυτοσυσχέτισης ανήκουν στο διάστημα $\pm 1.96\sqrt{n}$, όπου n είναι ο αριθμός των παρατηρήσεων. Εάν οι συντελεστές βρίσκονται εντός των ορίων τότε υπάρχει συσχέτιση ανάμεσα στις παρατηρήσεις και η χρονοσειρά δεν είναι τυχαία. Οι γραφικές αναπαραστάσεις της συνάρτησης αυτοσυσχέτισης και μερικής αυτοσυσχέτισης εμφανίζονται στα γραφήματα 4.1-6, 4.1-7



4.1-6 Διάγραμμα ACF της Χρονοσειράς



4.1-7 Διάγραμμα PACF της Χρονοσειράς

Εφόσον το ACF διάγραμμα μειώνεται σταδιακά και το PACF διάγραμμα έχει απότομη πτώση, αυτό σημαίνει ότι αναφερόμαστε σε ένα AR μοντέλο. Ωστόσο η μετάφραση των γραφικών αποτελεί πρόκληση μερικές φορές και για αυτό τον λόγο θα γίνει χρήση ενός πλέγματος ώστε να βρεθούν οι βέλτιστες τιμές για το μοντέλο. Να σημειωθεί ότι επειδή μετά την 3η υστέρηση γίνεται στατιστικά σημαντικό το διάγραμμα ACF μπορούμε να θεωρήσουμε ένα AR(3)

μοντέλο και επειδή το διάγραμμα PACF διάγραμμα γίνεται στατιστικά σημαντικό μετά την 1η υστέρηση μπορούμε να θεωρήσουμε ένα MA(1) μοντέλο.

Στην παρακάτω εικόνα 4.1-88 παρουσιάζονται οι πιθανοί συνδυασμοί $ARIMA(p,d,q)$ και αποτυπώνονται το AIC, BIC και RMSE, σε φθίνουσα σειρά με βάση την τιμή του AIC.

	AIC	BIC	RMSE
(1, 1, 1)	823.672017	833.025999	3.344321
(1, 0, 0)	824.708084	834.079976	3.350109
(1, 1, 2)	825.671887	838.143862	3.343666
(2, 1, 1)	825.671889	838.143865	3.343699
(1, 0, 1)	826.703377	839.199233	3.354090
...	...	...	...
(3, 3, 0)	980.189209	992.612990	20.633009
(0, 3, 1)	981.071104	987.282995	11.169333
(2, 3, 0)	1012.708243	1022.026079	27.916596
(1, 3, 0)	1061.665246	1067.877137	39.701775
(0, 3, 0)	1155.654058	1158.760004	61.673357

[64 rows x 3 columns]

4.1-8 Συνδυασμοί ARIMA(p,d,q)

Επομένως θα χρησιμοποιήσουμε τον συνδυασμό $ARIMA(1,1,1)$ ώστε να γίνει fit του μοντέλου. Τα αποτελέσματα που προκύπτουν είναι τα παρακάτω:

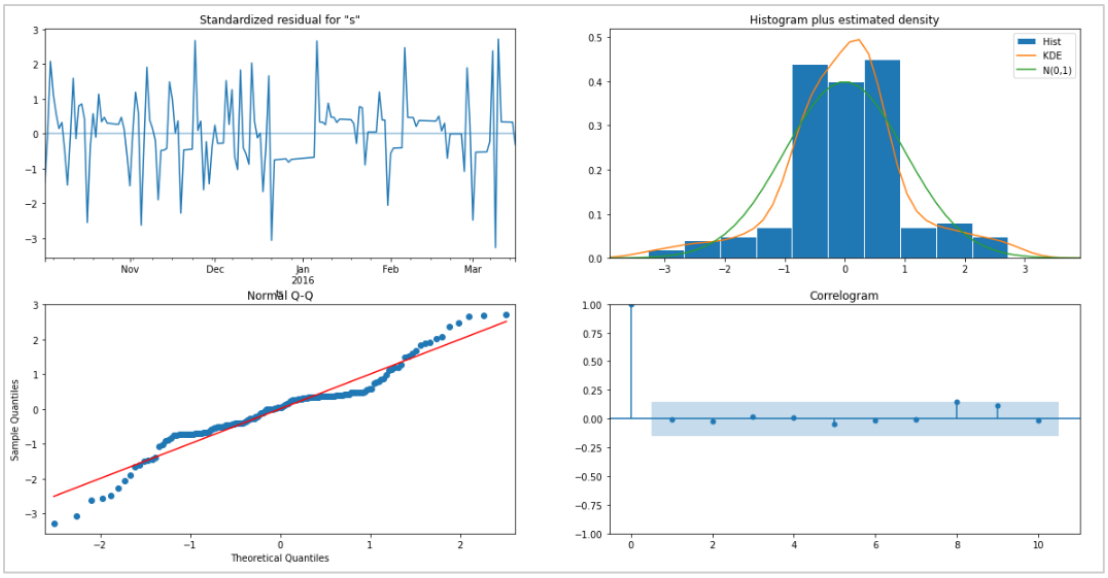
SARIMAX Results

Dep. Variable:	speed_max	No. Observations:	168
Model:	ARIMA(1, 1, 1)	Log Likelihood	-408.836
Date:	Wed, 22 Jun 2022	AIC	823.672
Time:	19:32:35	BIC	833.026
Sample:	10-01-2015 - 03-16-2016	HQIC	827.469
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6905	0.113	6.102	0.000	0.469	0.912
ma.L1	-0.9999	3.178	-0.315	0.753	-7.229	5.230
sigma2	7.6737	24.127	0.318	0.750	-39.614	54.962

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	18.99
Prob(Q):	0.97	Prob(JB):	0.00
Heteroskedasticity (H):	0.91	Skew:	-0.14
Prob(H) (two-sided):	0.72	Kurtosis:	4.63

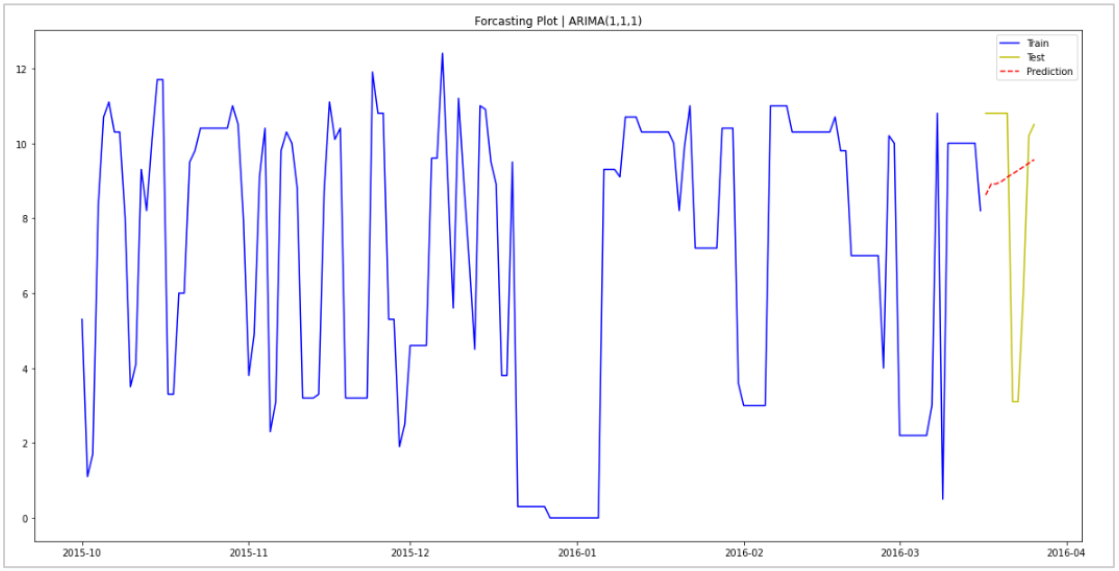
4.1-9 Αποτελέσματα ARIMA(1,1,1) μοντέλου



4.1-10 Διαγνωστικά Αποτελέσματα Μοντέλου

Παρατηρείται ότι τα κατάλοιπα κυμαίνονται κοντά στον μέσο μηδέν και έχουν ομοιόμορφη διακύμανση. Το ιστόγραμμα υποδεικνύει κανονική κατανομή με μέση τιμή μηδέν. Σχεδόν όλες οι τελείες πέφτουν πάνω στην κόκκινη γραμμή που σημαίνει ότι ακολουθούν την κανονική κατανομή και οποιεσδήποτε σημαντικές αποκλίσεις σημαίνουν ότι η κατανομή είναι ανομοιόμορφη. Σύμφωνα με το διάγραμμα ACF παρατηρείται ότι τα κατάλοιπα δεν συσχετίζονται.

Τέλος, στο παρακάτω διάγραμμα παρουσιάζεται η πρόβλεψη του μοντέλου $ARIMA(1,1,1)$ για διάστημα 10 ημερών με RMSE = 3,34 knots. Με την μπλε γραμμή είναι τα δεδομένα που έχει εκπαιδευτεί το μοντέλο, με κίτρινο χρώμα είναι τα τεστ δεδομένα και η κόκκινη διακεκομμένη γραμμή είναι η πρόβλεψη της ταχύτητας.



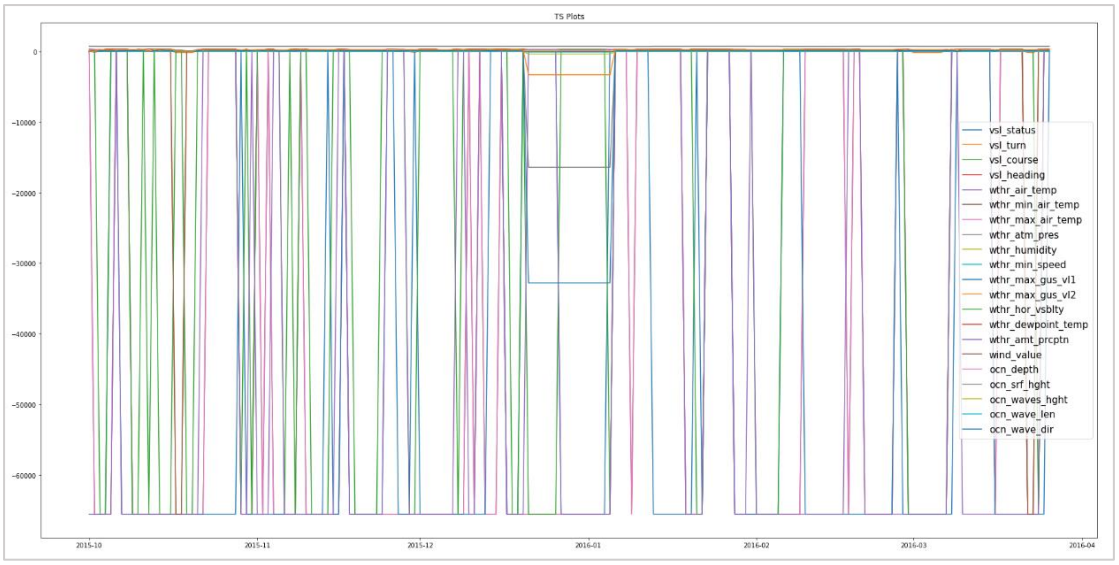
4.1-11 ARIMA(1,1,1) - Πρόβλεψη Ταχύτητας για διάστημα 10 ημερών

4.2 VAR Model – DATASET II

Το σύνολο δεδομένων το οποίο χρησιμοποιήθηκε για την ανάλυση του VAR μοντέλου ήταν το «DATASET II» και αποτελούνταν από 98.384 εγγραφές. Ακολουθήσαμε την ίδια διαδικασία όπως με το «DATASET I» κάνοντας δηλαδή αναγωγή των δεδομένων στην ημέρα ώστε να λαμβάνουμε την μεγαλύτερη ταχύτητα που είχε εντός μιας ημέρας με αποτέλεσμα να είναι εφικτή η ανάλυση της χρονοσειράς. Μετά από την αναπροσαρμογή των δεδομένων, προκύπτει ένα σύνολο δεδομένων από 178 εγγραφές και τα μέτρα θέσης και διασποράς του νέου συνόλου δεδομένων αποτυπώνονται παρακάτω καθώς επίσης και το διάγραμμα της χρονοσειράς:

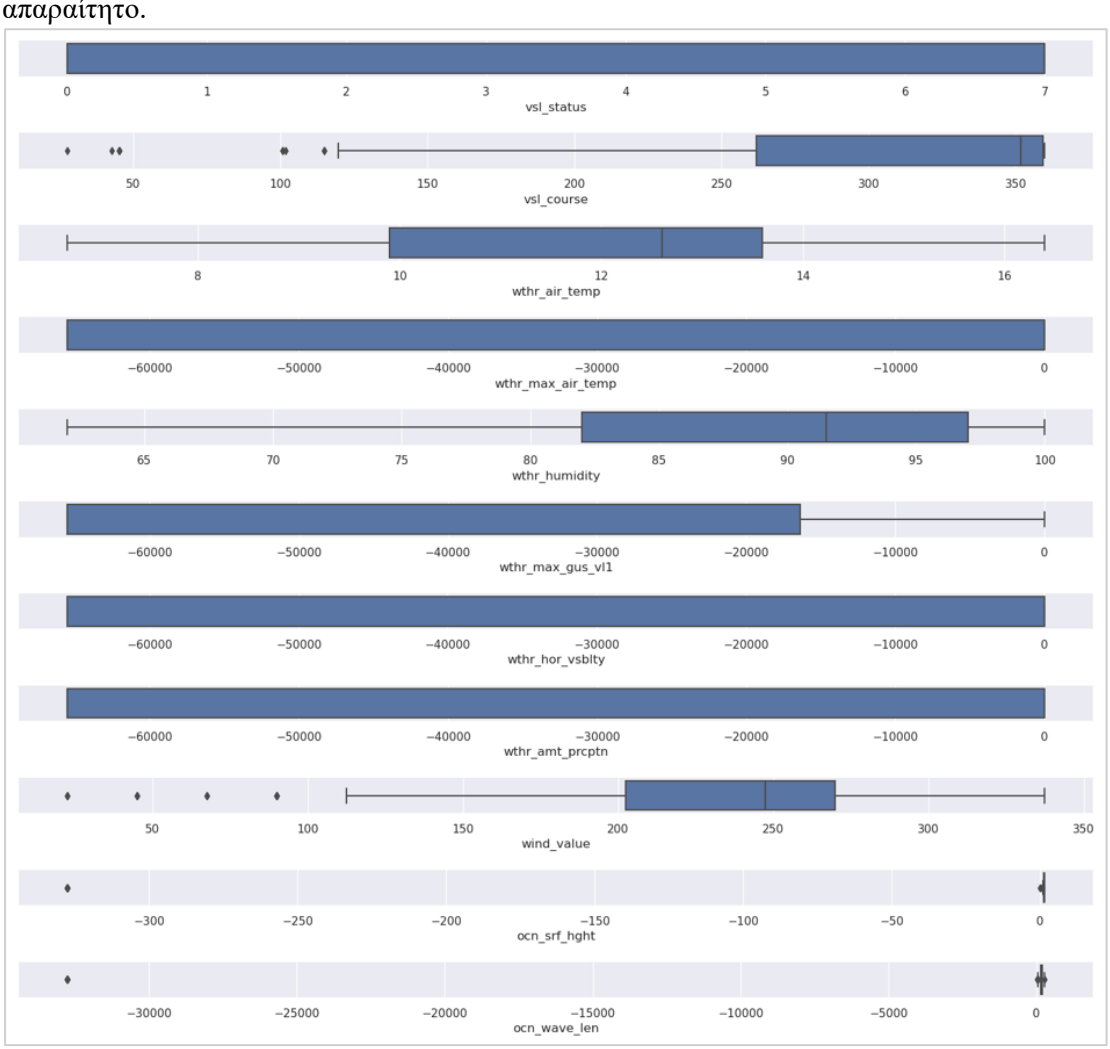
	count	mean	std	min	25%	50%	75%	max
vsl_status	178.0	3.539326	3.509652	0.00	0.0000	7.00	7.000	7.00
vsl_turn	178.0	94.971910	74.701926	-126.00	127.0000	127.00	127.000	127.00
vsl_course	178.0	298.425843	91.223494	27.50	261.9000	351.70	359.300	359.90
vsl_heading	178.0	278.595506	93.825143	23.00	235.0000	323.50	357.000	359.00
wthr_air_temp	178.0	11.985393	2.356362	6.70	9.9000	12.60	13.600	16.40
wthr_min_air_temp	178.0	-43074.046629	31195.999392	-65536.00	-65536.0000	-65536.00	6.400	13.70
wthr_max_air_temp	178.0	-43073.034831	31197.404612	-65536.00	-65536.0000	-65536.00	9.600	17.10
wthr_atm_pres	178.0	764.032584	6.620935	741.60	761.2500	765.00	768.700	776.20
wthr_humidity	178.0	89.247191	9.278224	62.00	82.0000	91.50	97.000	100.00
wthr_min_speed	178.0	10.876404	3.718393	2.00	8.0000	10.50	13.000	25.00
wthr_max_gus_vl1	178.0	-48962.528090	28573.038693	-65536.00	-65536.0000	-65536.00	-16372.750	33.00
wthr_max_gus_vl2	178.0	-1456.533708	9743.111964	-65536.00	13.0000	16.00	20.000	33.00
wthr_hor_vsblty	178.0	-18763.808989	29723.115744	-65536.00	-65536.0000	12.00	20.000	50.00
wthr_dewpoint_temp	178.0	9.635393	3.629732	2.10	6.9000	10.65	12.200	15.90
wthr_amt_prcptn	178.0	-44917.594944	30518.581033	-65536.00	-65536.0000	-65536.00	0.200	5.00
wthr_tm_prcptn	178.0	-1469.797753	9741.093895	-65536.00	1.0000	3.00	3.000	12.00
wind_value	178.0	227.907303	77.316457	22.50	202.5000	247.50	270.000	337.50
ocn_depth	178.0	-1369.682584	4731.719214	-16383.50	111.0000	113.25	118.875	131.00
ocn_srf_hght	178.0	-28.325169	94.340495	-327.67	1.1025	1.26	1.310	1.54
ocn_waves_hght	178.0	-2.583539	19.873224	-65.53	2.3300	3.79	4.280	6.26
ocn_wave_len	178.0	-2785.376404	9449.004333	-32767.00	140.0000	182.50	198.000	286.00
ocn_wave_dir	178.0	-41.404494	1020.817442	-3276.70	266.8750	279.70	298.900	354.50
vsl_speed	178.0	7.089326	3.782243	0.00	3.2250	8.90	10.300	12.40

4.2-1 Μέτρα Θέσης - Διασποράς | Dataset II - Before Removing Outliers

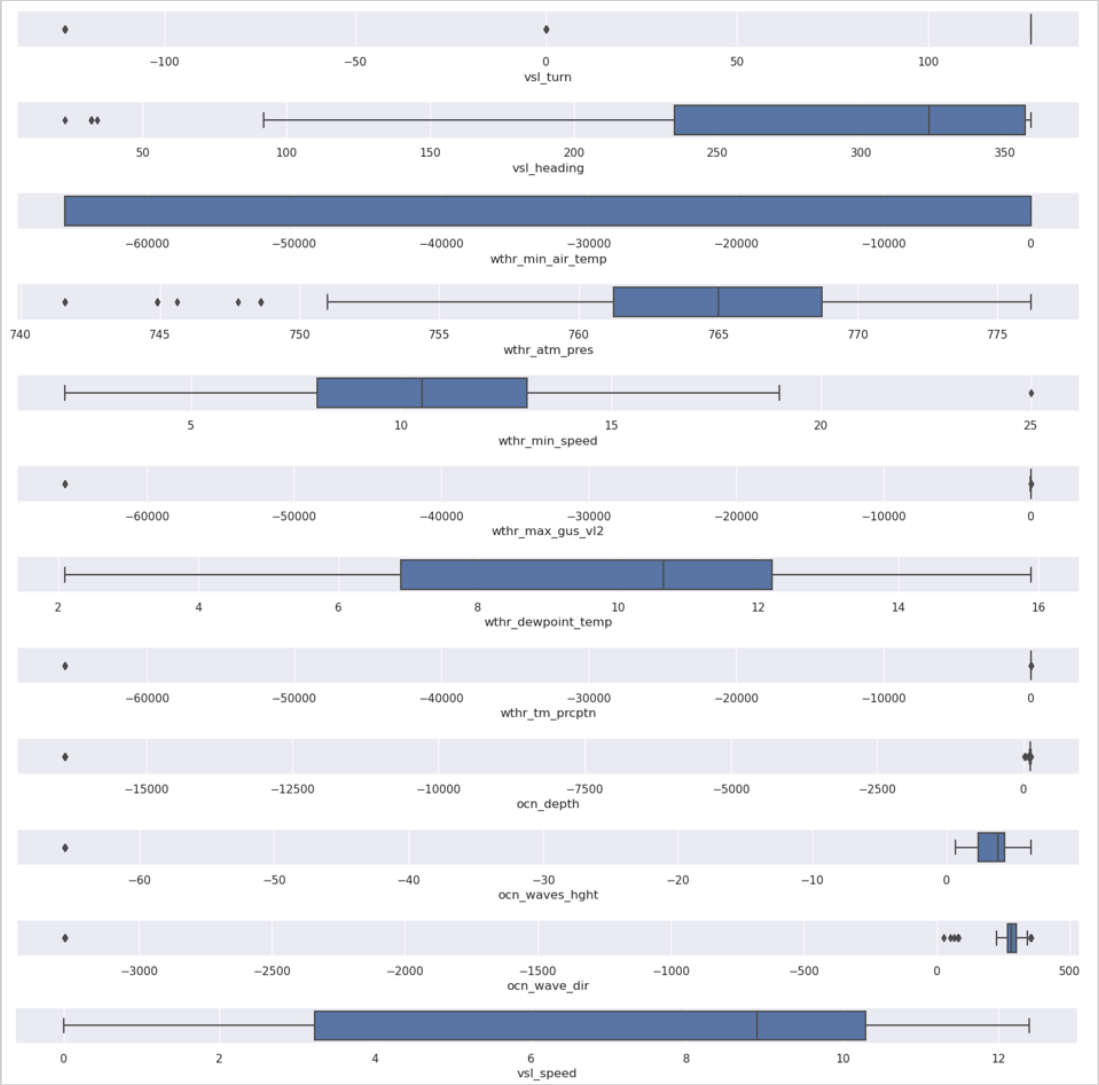


4.2-2 TS Plot | Dataset II - Before Removing Outliers

Σύμφωνα με τις παραπάνω εικόνες 4.2-1, 4.2-2 παρατηρείται ότι υπάρχουν ακραίες τιμές και για την επαλήθευση θα χρησιμοποιήσουμε διάγραμμα boxplot ώστε να βεβαιωθούμε ποια χαρακτηριστικά έχουν ακραίες τιμές και στην συνέχεια να αφαιρεθούν εφόσον κρίνεται απαραίτητο.



4.2-3 Box Plot_1 - Dataset II



4.2-4 Box Plot_2 - Dataset II

Σύμφωνα με την περιγραφή των δεδομένων που πραγματοποιήθηκε στο κεφάλαιο 3.1.1 έγινε αντικατάσταση των ακραίων τιμών με NULL, καθώς υποδηλώνουν ότι δεν υπάρχει διαθέσιμη πληροφορία. Στην συνέχεια έγινε αντικατάσταση με την διάμεσο του κάθε χαρακτηριστικού. Οι τιμές των αντικαταστάσεων αποτυπώνονται στον ακόλουθο πίνακα:

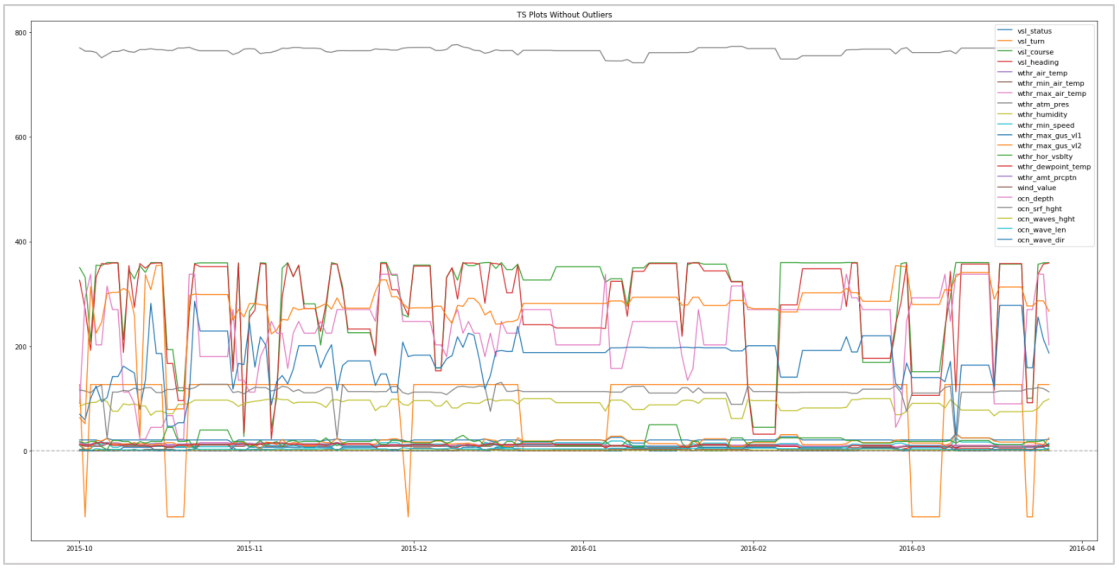
Χαρακτηριστικό	**Ακραία Τιμή Αντικατάστασης**
wthr_min_air_temp wthr_max_air_temp wthr_max_gus_vl1 wthr_max_gus_vl2 wthr_hor_vsblty wthr_amt_prcptn wthr_tm_prcptn	-65536.0
ocn_depth	-16383.5
ocn_srf_hght	-327.67
ocn_waves_hght	-65.53
ocn_wave_len	-32767.0
ocn_wave_dir	-3276.7

4.2-1 Ακραίες Τιμές Χαρακτηριστικών

Εφόσον έγιναν οι απαλοιφές των ακραίων τιμών αποτυπώνονται στις παρακάτω εικόνες τα νέα μέτρα θέσης και διασποράς του νέου συνόλου δεδομένων καθώς επίσης και το διάγραμμα της χρονοσειράς:

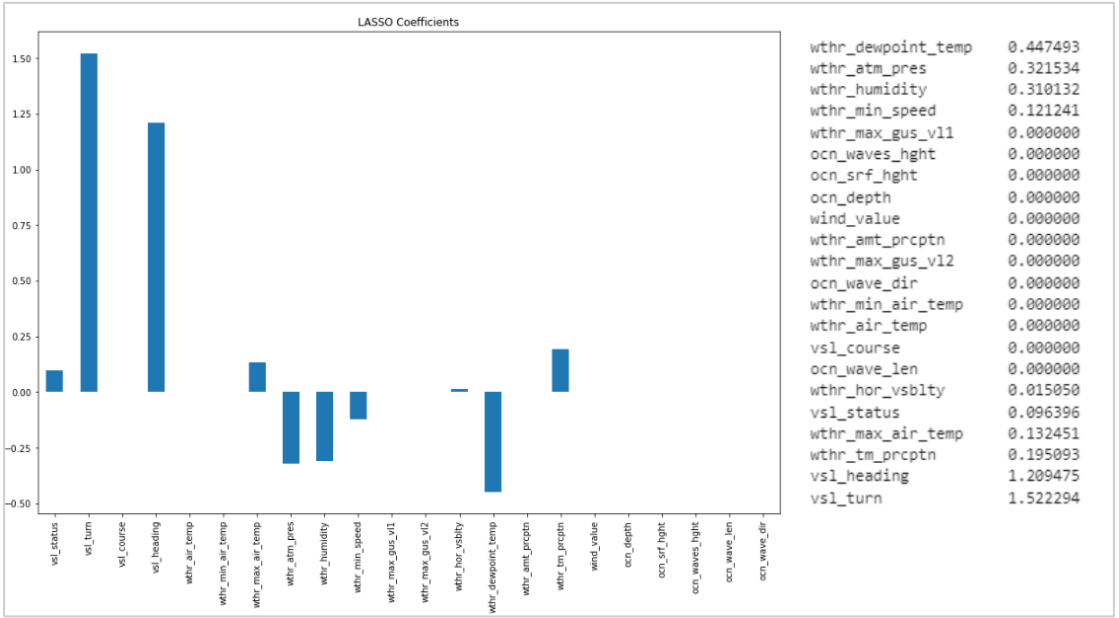
	count	mean	std	min	25%	50%	75%	max
vsl_status	178.0	3.539326	3.509652	0.00	0.0000	7.000	7.000	7.00
vsl_turn	178.0	94.971910	74.701926	-126.00	127.0000	127.000	127.000	127.00
vsl_course	178.0	298.425843	91.223494	27.50	261.9000	351.700	359.300	359.90
vsl_heading	178.0	278.595506	93.825143	23.00	235.0000	323.500	357.000	359.00
wthr_air_temp	178.0	11.985393	2.356362	6.70	9.9000	12.600	13.600	16.40
wthr_min_air_temp	178.0	8.705618	1.509507	4.80	8.7000	8.700	8.700	13.70
wthr_max_air_temp	178.0	11.097753	1.526340	8.10	10.8000	10.800	10.800	17.10
wthr_atm_pres	178.0	764.032584	6.620935	741.60	761.2500	765.000	768.700	776.20
wthr_humidity	178.0	89.247191	9.278224	62.00	82.0000	91.500	97.000	100.00
wthr_min_speed	178.0	10.876404	3.718393	2.00	8.0000	10.500	13.000	25.00
wthr_max_gus_vl1	178.0	21.073034	2.099301	15.00	21.0000	21.000	21.000	33.00
wthr_max_gus_vl2	178.0	16.544944	5.449377	3.00	13.0000	16.000	20.000	33.00
wthr_hor_vsblty	178.0	18.516854	9.576085	0.30	15.0000	18.000	20.000	50.00
wthr_dewpoint_temp	178.0	9.635393	3.629732	2.10	6.9000	10.650	12.200	15.90
wthr_amt_prcptn	178.0	0.680337	0.692499	0.20	0.5000	0.500	0.500	5.00
wthr_tm_prcptn	178.0	2.988764	2.055690	1.00	3.0000	3.000	3.000	12.00
wind_value	178.0	227.907303	77.316457	22.50	202.5000	247.500	270.000	337.50
ocn_depth	178.0	113.193820	14.299276	11.00	111.5000	113.500	118.875	131.00
ocn_srf_hght	178.0	1.242472	0.148162	0.00	1.2000	1.270	1.310	1.54
ocn_waves_hght	178.0	3.653315	1.184967	0.63	2.8325	3.855	4.280	6.26
ocn_wave_len	178.0	176.870787	45.695753	30.00	155.2500	188.000	198.000	286.00
ocn_wave_dir	178.0	278.451685	49.369622	27.00	272.7000	281.700	298.900	354.50
vsl_speed	178.0	7.089326	3.782243	0.00	3.2250	8.900	10.300	12.40

4.2-5 Μέτρα Θέσης - Διασποράς | Dataset II - After Removing Outliers



4.2-6 TS Plot | Dataset II - After Removing Outliers

Έπειτα έγινε επιλογή των χαρακτηριστικών που θα χρησιμοποιηθούν για την πρόβλεψη της ταχύτητας του πλοίου, vsl_speed, χρησιμοποιώντας την μέθοδο LASSO Regression. Στην παρακάτω εικόνα 4.2-7 όσοι συντελεστές των χαρακτηριστικών είναι ίσοι με το μηδέν τότε απορρίπτουμε τα χαρακτηριστικά αυτά και δεχόμαστε αυτά που η απόλυτη τιμή τους είναι θετική.



4.2-7 Αποδοχή - Απόρριψη Χαρακτηριστικών

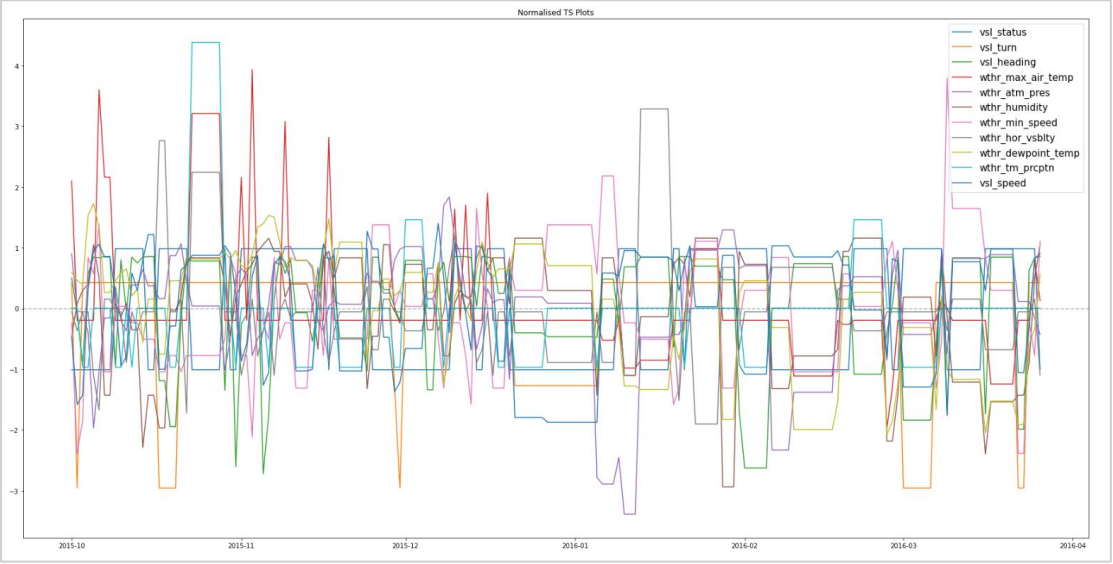
Επομένως τα χαρακτηριστικά που θα αποδεχθούμε για το μοντέλο μας είναι τα εξής:

```
' vsl_status ',' vsl_turn ',' vsl_heading ',' wthr_max_air_temp ',' wthr_atm_pres ',' wthr_humidity ',
 'wthr_min_speed','wthr_hor_vsblty','wthr_dewpoint_temp','wthr_tm_prcptn','vsl_speed'
```

Ύστερα από κανονικοποίηση των παραπάνω χαρακτηριστικών τα νέα μέτρα θέσης, διασποράς και το διάγραμμα της χρονοσειράς αποτυπώνονται παρακάτω.

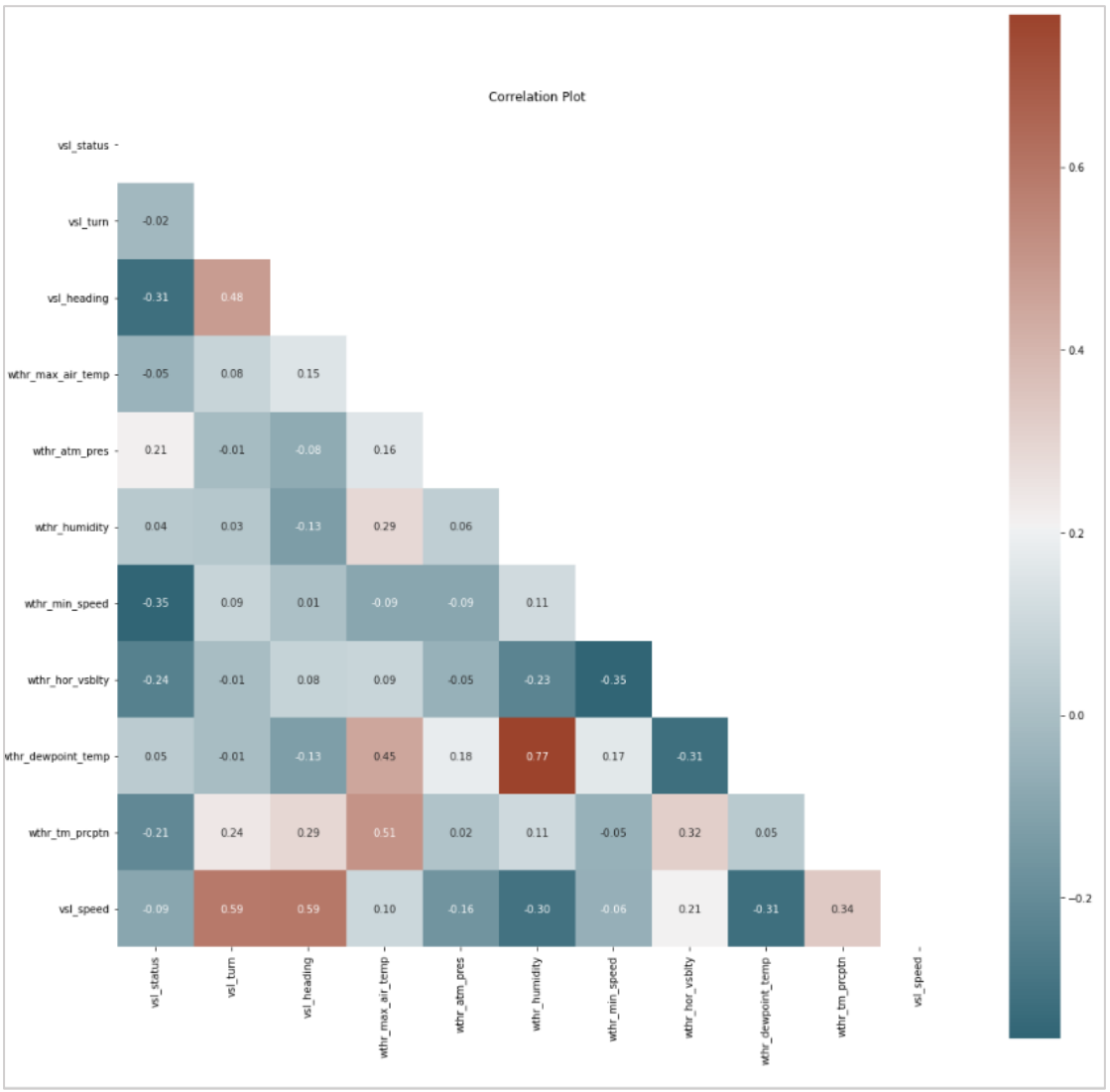
	count	mean	std	min	25%	50%	75%	max
vsl_status	178.0	-1.609200e-16	1.0	-1.008455	-1.008455	0.986045	0.986045	0.986045
vsl_turn	178.0	-3.368092e-17	1.0	-2.958048	0.428745	0.428745	0.428745	0.428745
vsl_heading	178.0	-1.372186e-16	1.0	-2.724169	-0.464646	0.478598	0.835645	0.856961
wthr_max_air_temp	178.0	1.728252e-14	1.0	-1.964014	-0.195076	-0.195076	-0.195076	3.932446
wthr_atm_pres	178.0	8.119847e-14	1.0	-3.388129	-0.420271	0.146115	0.704948	1.837719
wthr_humidity	178.0	1.933534e-16	1.0	-2.936682	-0.781097	0.242806	0.835592	1.158930
wthr_min_speed	178.0	1.621674e-16	1.0	-2.387162	-0.773561	-0.101228	0.571106	3.798306
wthr_hor_vsblty	178.0	-6.860929e-16	1.0	-1.902328	-0.367254	-0.053973	0.154880	3.287684
wthr_dewpoint_temp	178.0	-4.347334e-16	1.0	-2.076019	-0.753607	0.279527	0.706555	1.725914
wthr_tm_prcptn	178.0	-8.449468e-18	1.0	-0.967444	0.005466	0.005466	0.005466	4.383558
vsl_speed	178.0	1.109599e-15	1.0	-1.874371	-1.021702	0.478730	0.848881	1.404107

4.2-8 Μέτρα Θέσης - Διασποράς | Dataset II - Κανονικοποιημένα Χαρακτηριστικά



4.2-9 TS Plot | Dataset II - Κανονικοποιημένα Χαρακτηριστικά

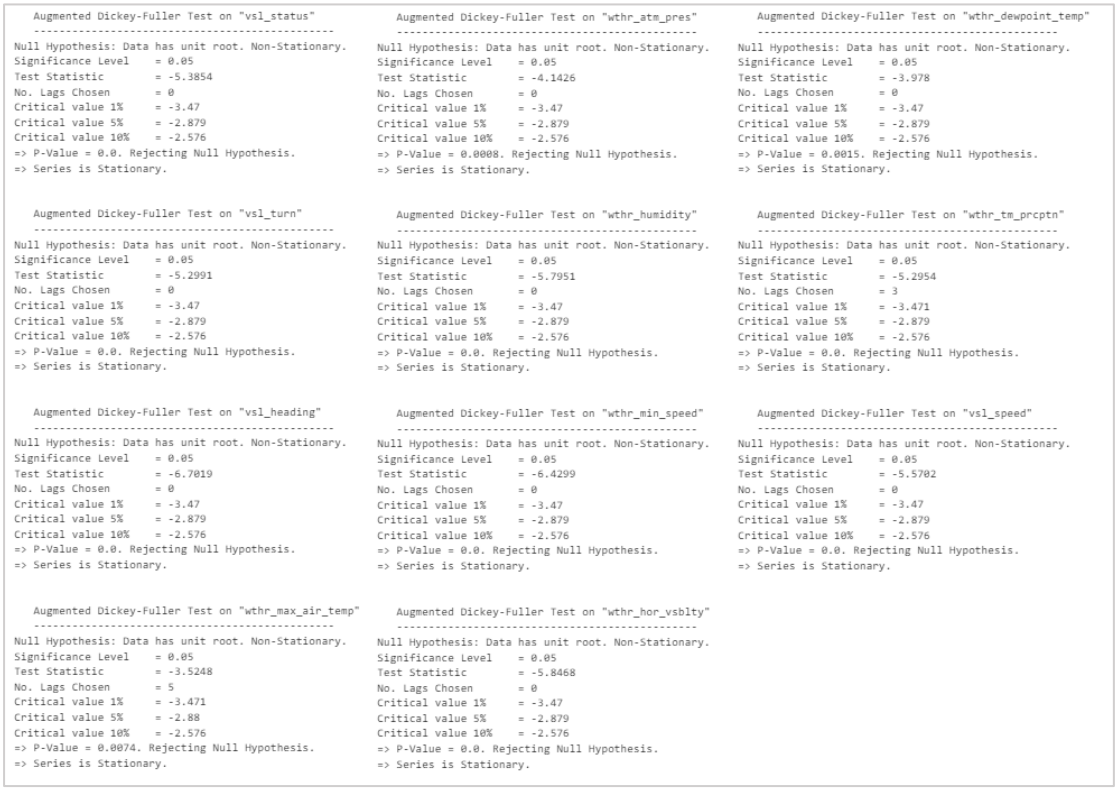
Με βάση το διάγραμμα των συσχετίσεων παρατηρούμε ότι τα χαρακτηριστικά vls_turn, vls_heading έχουν την μεγαλύτερη συσχέτιση με την ταχύτητα του πλοίου της τάξης του 59% σε σχέση με τα υπόλοιπα χαρακτηριστικά.



4.2-10 Correlation Plot

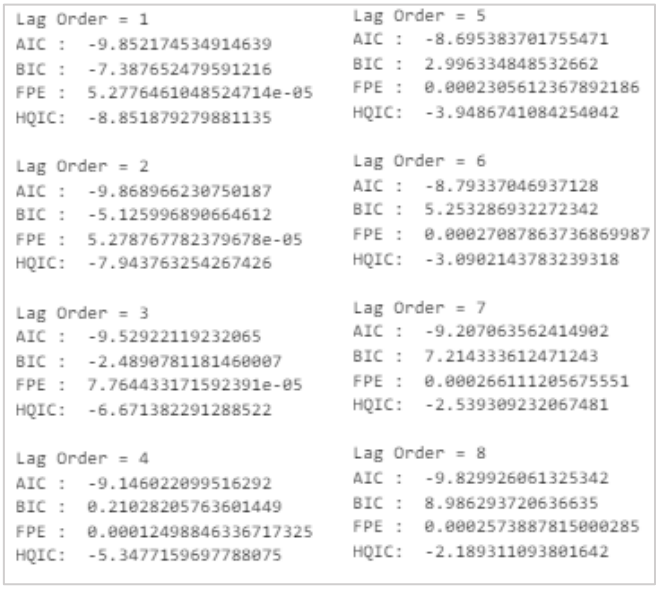
Για την εκπαίδευση του VAR μοντέλου θα πρέπει όλες οι χρονοσειρές να είναι στάσιμες και στην συνέχεια βρίσκουμε την υστέρηση που θα γίνει fit το μοντέλο ανάλογα την τιμή του AIC.

Εκτελώντας το ADF test για την κάθε χρονοσειρά βλέπουμε ότι όλες οι χρονοσειρές είναι στάσιμες και έχουν τα παρακάτω αποτελέσματα:



4.2-11 ADF test για όλες τις χρονοσειρές

Επιπλέον, με βάση την εικόνα 4.2-12 θα γίνει επιλογή της υστέρησης που εκπαιδευτεί το VAR μοντέλο. Επειδή μετά την 3η υστέρηση το AIC μειώνεται σημαντικά, θα γίνουν δοκιμές μεταξύ των πιο χαμηλών τιμών.



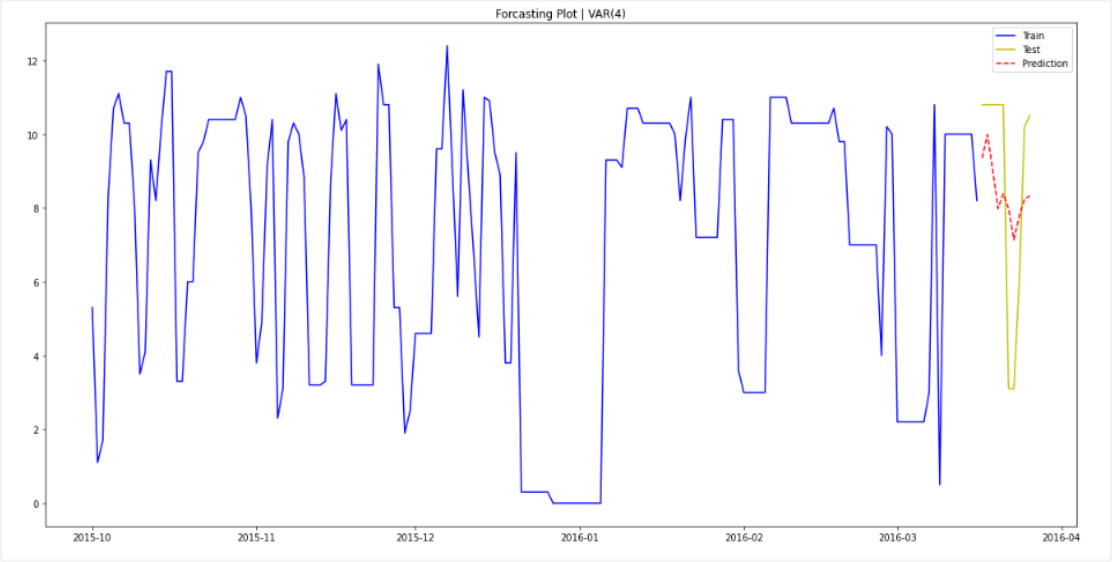
4.2-12 VAR Model Lags

Μετά από δοκιμές στην προβλεπτική ικανότητα του μοντέλου, αποδείχθηκε ότι η 4η υστέρηση απέδωσε τα καλύτερα αποτελέσματα όπως φαίνεται στον πίνακα 4.2-2:

Lag Order	RMSE
1	2,8814
2	3,1429
3	3,2574
4	2,6743
5	3,4054
6	3,7984
7	7,168
8	8,2342

4.2-2 Αποτελέσματα VAR μοντέλου

Τέλος, στο παρακάτω διάγραμμα παρουσιάζεται η πρόβλεψη του μοντέλου $VAR(4)$ για διάστημα 10 ημερών με RMSE = 2,67 knots. Με την μπλε γραμμή είναι τα δεδομένα που έχει εκπαιδευτεί το μοντέλο, με κίτρινο χρώμα είναι τα τεστ δεδομένα και η κόκκινη διακεκομμένη γραμμή είναι η πρόβλεψη της ταχύτητας.



4.2-13 VAR(4) - Πρόβλεψη Ταχύτητας για διάστημα 10 ημερών

4.3 Συμπεράσματα

Στην παρούσα έρευνα μελετήθηκαν τεχνικές πρόβλεψης χρονοσειρών και εφαρμόστηκαν στα σύνολα δεδομένων που είναι διαθέσιμα στο link Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance | Zenodo [16]. Τα δεδομένα αφορούσαν δεδομένα πλοίου για την χρονική περίοδο 2015-10-01 έως 2016-03-31. Πραγματοποιήθηκαν διάφοροι μετασχηματισμοί των δεδομένων ώστε να καταλήξουν σε δύο σύνολο δεδομένων, ένα που θα έχει ως δεδομένα τον χρόνο, ταχύτητα και ένα δεύτερο που θα περιέχει μεταβλητές χρόνου, καιρού, ωκεανού, ανέμου που ενδέχεται να επηρεάζουν την ταχύτητα. Ο τρόπος με τον οποίο δημιουργήθηκε το συνδυαστικό σύνολο δεδομένων ήταν με βάση την μεθοδολογία, αλγόριθμο της διαμέρισης του χώρου που βοήθησε ώστε να λαμβάνουμε την κοντινότερη και πιο πρόσφατη πληροφορία κοντά στην τοποθεσία του πλοίου που είχε την αντίστοιχη χρονική στιγμή.

Τέλος, για την πρόβλεψη χρησιμοποιήθηκαν δύο μοντέλα χρονοσειρών για το ARIMA και το VAR, για το μονομετάβλητο και πολυμετάβλητο σύνολο δεδομένων αντίστοιχα. Το χρονικό διάστημα πρόβλεψης και για τα δύο μοντέλα ήταν δέκα ημερών. Συνοψίζοντας, τα αποτελέσματα του ARIMA(1,1,1) μοντέλου έχουν προβλεπτικό σφάλμα 3.34 knots, ενώ το VAR(4) είχε προβλεπτικό σφάλμα της τάξης των 2.67 knots. Άρα καταλήγουμε στο συμπέρασμα ότι το πολυμετάβλητο VAR μοντέλο που εξετάστηκε στο σύνολο δεδομένων «DATASET II» αποδίδει την καλύτερη πρόβλεψη.

Βιβλιογραφία

[1] Spiegel, R.M., Stephens, L.J., *"Στατιστική 3η Έκδοση.Εκδόσεις Τζίολα."*, 2000.

[2] George E. P.Box, Gwilym M.Jenkins, Gregory C.Reinsel, Greta M. Ljung, *"Time Series Analysis: Forecasting and Control"*, vol. 5, 2015.

[3] Ivanovic, M., & Kurbalija, V., *"Time series analysis and possible applications."* 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2016.
DOI: https://doi.org/10.1109/MIPRO.2016.7522190

[4] Douglas C. Montgomery, CheryL L. Jennings, Murat Kulahci, *"Time Series Analysis and Forecasting"*, vol.2, 2015.

[5] Peter J.Brockwell, Richard A.Davis, *"Intoduction to Time Series and Forecasting"*, Springer, vol. 3, 2016.

[6] Dickey A. Dickey & Wayne A. Fuller, *"Distribution of the Estimators for Autoregressive Time Series with a Unit Root"*, 2012, vol. 74, no.366a, pp.427 – 431.
DOI: https://doi.org/10.1080/01621459.1979.10482531

[7] Graham Elliott,Thomas J. Rothenberg, James H. Stock, *"Efficient Test for an Autoregressive Unit Root"*, 1996, vol.64, no.4, pp.813 – 836
DOI: https://doi.org/10.2307/2171846

[8] Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y., *"Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics"*, 1992, vol. 54, no. (1-3), pp. 159–178.
DOI: https://doi.org/10.1016/0304-4076(92)90104-y

[9] Attila Hornok,Rolf Larson, *"The finite sample distribution of the KPSS Test"*, 2000, vol.3, no.1, pp.108 – 121.
DOI: https://doi.org/10.1111/1368-423X.00041

[10] G. E. P. Box, David A. Pierce, *"Distribution of Residual Autocorrelations in Autoregressive – Integrated Moving Average Time Series Models"*, 2012, vol. 65, no. 332, pp. 1509 – 1526.
DOI: http://dx.doi.org/10.1080/01621459.1970.10481180

[11] G. M. Ljung, G. E. P. Box, *"On a measure of lack of fit in time series models"*, Journal of American Statistical Association, 1978, vol. 65, no. 2, pp.297 – 303.
DOI: https://doi.org/10.1093/biomet/65.2.297

[12] Triyani Hendrawati, I Made Sumertajaya, Aji Wigena, Bagus Sartono, *"Performance Evaluation of AIC and BIC in Time Series Clustering with Picollo Method"*, Proceedings of the 1st International Conference on Statistics and Analytics, ICSA 2019, 2020.
DOI: http://dx.doi.org/10.4108/eai.2-8-2019.2290340

[13] Arnold Zellner, *"An Introduction to Bayesian Inference in Econometrics"*, Wiley, 1996.

[14] Robert H. Shumway, David S.Stoffer, "Time Series Analysis and Its Applications", Springer Texts in Statistics, 2017, vol.4.
DOI: https://doi.org/10.1007/978-3-319-52452-8

[15] Eric Zivot, Jiahui Wang, *"Vector Autoregressive Models for Multivariate Time Series"*, Modeling Time Series with S-Plus®. Springer, 2003, pp.369 – 413
DOI: https://doi.org/10.1007/978-0-387-21763-5

[16] Cyril Ray, Richard Dreo, Elena Camossi, Anne-Laure Jousselme, Clement Iphar, *"Heterogeneous integrated dataset for Maritime Intelligence, surveillance, as reconnaissance"*, Data in brief 104141, 2019.
DOI: https://doi.org/10.1016/j.dib.2019.104141

[17] Tibshirani, R, *"Regression Shrinkage and Selection via the Lasso"*, Journal of the Royal Statistical Society: Series B, 1995, vol.56, no.1, pp. 267 – 288.

[18] Ameen M. Bassam, Alexander B. Phillips, Stephen R. Turnock, Philip A. Wilson, *"Spip speed prediction based on machine learning for efficient shipping operation"*,Ocean Engineering, 2022, vol. 245.
DOI: https://doi.org/10.1016/j.oceaneng.2021.110449

[19] Kiriakos Alexiou, Efthimios Pariotis, Theodoros Zannis, Stylianos Polyzos, Helen Leligou, *"Comparative evaluation of Machine Learning algorithms and Physical based models for the prediction of Vessel Speed in real life applications"*, 25th Pan-Hellenic Conference of Informatics, 2021, pp. 442 – 447.
DOI: https://doi.org/10.1145/3503823.3503904

[20] Wengang Mao, Igor Rychlik, Jonas Wallin, Gaute Storhaug, *"Statistical models for the speed prediction of a container ship"*, Ocean Engineering, 2016, vol. 126, pp. 152 – 162.
DOI: https://doi.org/10.1016/j.oceaneng.2016.08.033

[21] Prithvi S Rao, Ekaterina Kim, Bjornar Brende Smestad, Bjon Egil Asbjornslett, Anirban Bhattacharyya, *"Predicting vessel speed in the Artic without knowing ice conditions using AIS data and decision tress"*, Meritime Transport Research, vol. 2, 2021.
DOI: https://doi.org/10.1016/j.martra.2021.100024