ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ & ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Ανάλυση και πρόβλεψη χρονοσειρών με εφαρμογές στην ναυλαγορά Tanker

Χοντζοπούλου Νίκη

Επιβλέπουσα Καθηγήτρια: Καρώνη Χρυσηίς, Καθηγήτρια Ε.Μ.Π

Αθήνα

Οκτώβρης 2018

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ & ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Ανάλυση και πρόβλεψη χρονοσειρών με εφαρμογές στην ναυλαγορά Tanker

Χοντζοπούλου Νίκη

Τριμελής Επιτροπή

……..	……..	……..
Καρώνη Χρυσηίς	Λυρίδης Β. Δημήτριος	Παπανικολάου Βασίλης
Καθηγήτρια	Αναπλ. Καθηγητής	Καθηγητής
Ε.Μ.Π	Ε.Μ.Π	Ε.Μ.Π

Αθήνα

2018

……

Χοντζοπούλου Νίκη, 2018

Διπλωματούχος Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών Ε.Μ.Π

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω ιδιαίτερα την επιβλέπουσα της διπλωματικής μου κ. Καρώνη, για την ευκαιρία που μου έδωσε με την ανάθεσή της και την πολύτιμη βοήθεια που μου παρείχε τόσο κατά τη διάρκεια της εργασίας μου, όσο και συνολικά κατά τη διάρκεια των σπουδών μου. Τίποτα ωστόσο δεν θα είχε πραγματοποιηθεί χωρίς τη συμβολή του κ. Λυρίδη, τον οποίο ευχαριστώ για τις κατευθύνσεις του ως προς την προσέγγιση του θέματος της έρευνάς μου και το υλικό που μου συνέστησε γύρω από τα ζητήματα που αφορούσαν την εργασία μου. Επιπρόσθετα, θα ήθελα να ευχαριστήσω ξεχωριστά και το τρίτο μέλος της επιτροπής, κ. Παπανικολάου, για την σημαντική συμβολή του στην περάτωση της διπλωματικής εργασίας. Τέλος, θέλω να πω ένα μεγάλο ευχαριστώ στην οικογένειά μου και τους φίλους μου, για την στήριξη και την υπομονή τους όλα αυτά τα χρόνια.

Περίληψη

Σκοπός της παρούσας διπλωματικής εργασίας είναι η πρόβλεψη του δείκτη Worldscale της Ναυλαγοράς Tanker για πλοία Very Large Crude Carriers (VLCC) στη γραμμή μεταφοράς αργού πετρελαίου Ras Tanura-Rotterdam. Για την πρόβλεψη των μελλοντικών τιμών του δείκτη Worldscale θα χρησιμοποιήσουμε τα αυτοπαλινδρομικά μοντέλα της οικογένειας ARIMA που αναπτύχθηκαν από τους Box και Jenkins. Τα μοντέλα αυτά εν γένει αποτελούν ένα σημαντικό εργαλείο για την προσομοίωση διαφόρων χρονοσειρών.

Το κείμενο είναι πρακτικά χωρισμένο σε τρία μέρη. Αρχικά, παρατίθενται οι βασικές έννοιες, ο σκοπός για την ανάλυση χρονοσειρών και η ναυλαγορά από την οποία έχει αντληθεί ο δείκτης Worldscale. Στην συνέχεια παρουσιάζονται τα μοντέλα τα οποία χρησιμοποιούνται για την περιγραφή και πρόβλεψη χρονοσειρών καθώς επίσης και η εφαρμογή των παραπάνω μεθόδων στον προς πρόβλεψη δείκτη Worldscale.

Στο πρώτο κεφάλαιο παρουσιάζονται κάποιες βασικές έννοιες για τα είδη των χρονοσειρών και τα προβλήματα πρόβλεψης και γίνεται μια αναφορά στη ναυλαγορά Tanker για τα δεξαμενόπλοια υγρού φορτίου. Στο δεύτερο κεφάλαιο αναλύονται τα βασικά μονοδιάστατα μοντέλα πρόβλεψης χρονοσειρών καθώς επίσης και ο ρόλος των συναρτήσεων αυτοσυσχέτισης (ACF) και μερικής αυτοσυσχέτισης (PACF). Εν συνεχεία, εστιάζουμε στις μεθόδους εκτιμήσεων των συντελεστών του μοντέλου ARIMA για τα δεδομένα μας και στους διαγνωστικούς ελέγχους για καλή προσαρμογή του, με τη χρήση της γλώσσας προγραμματισμού R.

Στο τρίτο κεφάλαιο παρουσιάζεται η εφαρμογή και τα αποτελέσματα που πήραμε από την R κατά την ανάλυση του δείκτη Worldscale. Πιο συγκεκριμένα, χρησιμοποιείται όλη η θεωρία που αναφέρεται στα προηγούμενα δυο κεφάλαια με σκοπό να δημιουργηθεί το καταλληλότερο αυτοπαλινδρομικό μοντέλο που θα περιγράψει ικανοποιητικά τον δείκτη Worldscale. Κατά τη διάρκεια αυτής της διαδικασίας, παρουσιάζεται αναλυτικά η μεθοδολογία που ακολουθήθηκε για να καταλήξουμε στο καλύτερο μοντέλο έτσι ώστε να γίνει κατανοητός ο τρόπος κατασκευής ενός μοντέλου. Στο τέλος της εφαρμογής, παρατίθενται οι μελλοντικές προβλέψεις που έδωσε το επιλεγμένο μοντέλο οι οποίες συγκρίνονται με τις πραγματικές τιμές του δείκτη και ελέγχεται η ακρίβεια των αποτελεσμάτων με τη βοήθεια του μέσου απόλυτου σφάλματος (MAE) και της ρίζας του τετραγωνικού απόλυτου σφάλματος (RMSE).

Abstract

This thesis focuses on analyzing and forecasting the Worldscale index for Very Large Crude Carriers (VLCC) vessels in Tanker market, for the route Ras Tanurra-Rotterdam. Box and Jenkins methods were applied, which are widely used in simulation and forecasting of various time series. The first two parts are theoretical whereas the last part is computational. For the computational part, R programming language was employed.

In the first section, some basic ideas of time series objects and the tanker market are introduced. In the second section, we discuss about several autoregressive models and the purpose of the autocorrelation (ACF) and partial autocorrelation (PACF) functions. Subsequently, we focus on model identification, estimation of the coefficients of autoregressive models and the goodness of model's fit through diagnostic checking.

The third section consists of application in Worldscale index of the theory that was presented at the two previews sections. The analytical methodology is presented in order to understand how to build the most suitable model for Worldscale index. Furthermore, forecast is performed for the values of Worldscale index for each month of the year 2017 and comparison is made between the forecast values from the model and the actual values of the Worldscale index, for the same time period. Finally, we examined if the selected model performs better in comparison to all other candidate models, using forecasting accuracy criteria such as Mean Absolute Error(MAE) and Mean Squared Error (RMSE).

Περιεχόμενα

Κεφάλαιο 1

1. Ανάλυση χρονοσειρών και είδη Ναυλαγορών

1.1 Εισαγωγή στην Ανάλυση Χρονοσειρών

Πολλές φορές στην καθημερινή ζωή, προκύπτει η ανάγκη της πρόβλεψης της συμπεριφοράς μιας μεταβλητής στο μέλλον. Τέτοιου είδους πειράματα μπορούν να ταξινομηθούν σε δύο κατηγορίες, τα προσδιοριστικά (ή ντετερμινιστικά) και τα τυχαία. Στην πρώτη κατηγορία τα αποτελέσματα του πειράματος είναι τελείως προβλέψιμα. Για παράδειγμα, αν σε σώμα μάζας m, εφαρμόσουμε δύναμη F τότε αυτό θα αποκτήσει επιτάχυνση $\alpha = F / m$ όσες φορές και αν επαναλάβω το πείραμα κάτω από τις ίδιες φαινομενικά συνθήκες. Σε τέτοιου είδους προβλήματα που η μεταβλητή διέπεται από συγκεκριμένους νόμους, η χρήση της στατιστικής είναι άσκοπη.

Από την άλλη μεριά υπάρχουν πειράματα τα οποία δεν είναι ντετερμινιστικά, δεν υπάρχουν αυστηροί νόμοι που τα διέπουν και οι παράγοντες που τα επηρεάζουν είναι τόσοι πολλοί που μπορούν να θεωρηθούν τυχαία γεγονότα. Σε αυτή την περίπτωση η στατιστική μας βοηθάει να ποσοτικοποιήσουμε με κάποιον τρόπο την τυχαιότητα τους και να δώσουμε πιθανές εκτιμήσεις περιορίζοντας την αβεβαιότητα των προβλημάτων όσο περισσότερο γίνεται. Σε τέτοιου είδους προβλήματα κατατάσσεται η πρόβλεψη της τιμής μιας μετοχής σε μια εβδομάδα, οι πωλήσεις μιας εταιρίας τον ερχόμενο μήνα κ.ο.κ .

Στις περιπτώσεις όπου υπάρχουν ιστορικά δεδομένα για τις μεταβλητές που μελετάμε τότε λαμβάνοντας υπόψιν τα ιστορικά αυτά δεδομένα, μας επιτρέπεται να κάνουμε προβλέψεις που αν και μη ακριβείς μας βοηθούν να πάρουμε σημαντικές αποφάσεις για το μέλλον. Μια μέθοδος για να ποσοτικοποιήσουμε τις προβλέψεις για το μέλλον πραγματοποιείται μέσω της ανάλυσης χρονοσειρών . Το πρόβλημα που προκύπτει είναι η πρόβλεψη μελλοντικών τιμών με βάση τις μέχρι σήμερα τιμές της ίδιας χρονοσειράς, είτε ακόμα σε συνδυασμό με τις μέχρι σήμερα τιμές μιας άλλης χρονοσειράς η οποία εξελίσσεται παράλληλα με την πρώτη και επιδρά πάνω σε αυτήν. Όσον αφορά τα προβλήματα πρόβλεψης μπορούν να ταξινομηθούν σε τρείς κατηγορίες:

1. Βραχυπρόθεσμα
2. Μεσοπρόθεσμα
3. Μακροπρόθεσμα

Τα βραχυπρόθεσμα προβλήματα πρόβλεψης περιλαμβάνουν την πρόβλεψη γεγονότων μόνο μερικές χρονικές περιόδους (ημέρες, εβδομάδες, μήνες) στο μέλλον. Από την άλλη πλευρά οι μεσοπρόθεσμες προβλέψεις παρατείνονται από ένα έως δυο χρόνια στο μέλλον και τέλος οι μακροπρόθεσμες μπορούν να επεκταθούν στην πρόβλεψη πολλών χρόνων μετά στο μέλλον. Οι βραχυπρόθεσμες και μεσοπρόθεσμες προβλέψεις βασίζονται συνήθως στην αναγνώριση, των μοτίβων που υπάρχουν στα ιστορικά δεδομένα και έπειτα στη μοντελοποίηση τους. Στην παρούσα διπλωματική θα επικεντρωθούμε στις βραχυπρόθεσμες προβλέπεις μιας χρονοσειράς μερικών μηνών στο μέλλον. Η μορφή της πρόβλεψης παίζει πολύ σημαντικό ρόλο. Συνήθως σκεφτόμαστε μια πρόβλεψη ως έναν αριθμό που αντιπροσωπεύει την καλύτερη εκτίμηση της μελλοντικής τιμής της μεταβλητής που μας ενδιαφέρει, δηλαδή μια σημειακή εκτίμηση (ή πρόβλεψη).

Τώρα αυτές οι προβλέψεις είναι σχεδόν πάντα λάθος και συνεπώς χρειάζεται να ορίσουμε και ένα σφάλμα πρόβλεψης. Επομένως, μια καλή πρακτική είναι να συνοδεύσουμε μια πρόβλεψη με μια εκτίμηση για το πόσο μεγάλο ενδέχεται να είναι ένα σφάλμα πρόβλεψης.

Άλλα σημαντικά χαρακτηριστικά του προβλήματος πρόβλεψης είναι ο χρονικός ορίζοντας της πρόβλεψης, δηλαδή ο αριθμός των μελλοντικών περιόδων για τις οποίες γίνεται η πρόβλεψη και συχνά υπαγορεύεται από τη φύση του προβλήματος.

Γενικά, θα πρέπει να κάνουμε μια διάκριση όταν αναφερόμαστε στην πρόβλεψη (ή στην προβλεπόμενη τιμή) του X_t που έγινε σε κάποια προηγούμενη χρονική περίοδο $t-h$, και στην εκτίμηση της τιμής της X_t που προέκυψε από την εκτίμηση των παραμέτρων σε μοντέλο χρονοσειράς βασισμένο σε ιστορικά δεδομένα. Ας σημειώσουμε εδώ ότι h είναι ο χρόνος πρόβλεψης. Η πρόβλεψη που έγινε κατά την χρονική περίοδο $t+h$ συμβολίζεται με $\hat{x}_{t+h}$.

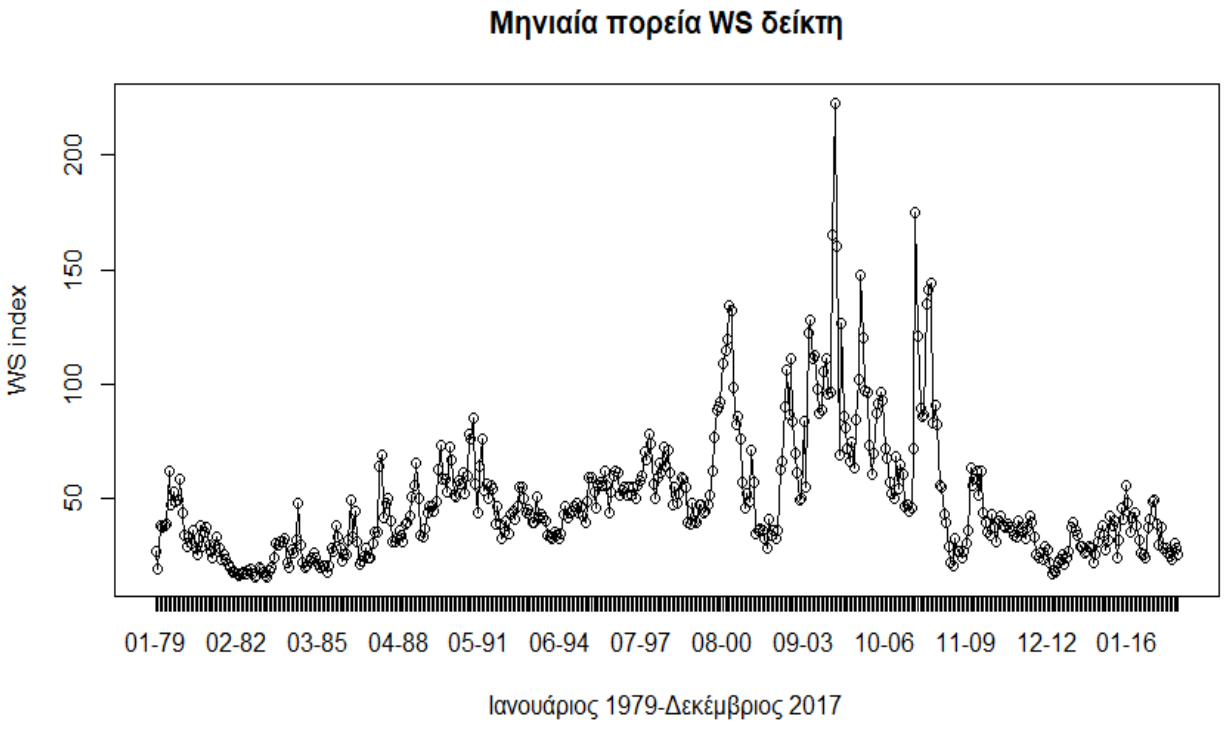
Υπάρχει πολύ μεγάλο ενδιαφέρον για την πρόβλεψη με lag=1, η οποία είναι η πρόβλεψη της παρατήρησης κατά την περίοδο t, x_t , που προέκυψε από την αμέσως προηγούμενη x_{t-1}. Θα συμβολίζουμε την προσαρμοσμένη τιμή της x_t ως $\widehat{x_t}$.

Μια σημαντική κατηγορία μοντέλων της ανάλυσης χρονοσειρών είναι τα μονοπαραμετρικά μοντέλα λόγω του ότι δε χρειαζόμαστε καμιά άλλη πληροφορία πέρα από το ιστορικό της μεταβλητής που μελετάμε. Σε πολλές περιπτώσεις, η γνώση των ιστορικών τιμών είναι αρκετά ικανοποιητική για να μας δώσει πολύ καλές εκτιμήσεις, καθώς μέσα στη ίδια τη χρονοσειρά μπορεί να υποκρύπτονται σχεδόν όλες οι απαραίτητες πληροφορίες που χρειαζόμαστε. Πριν προχωρήσουμε την ανάλυση μας είναι αναγκαίο να ορίσουμε τι ακριβώς εννοούμε όταν αναφερόμαστε σε μια χρονοσειρά.

Με τον όρο χρονοσειρά εννοούμε μια ακολουθία $\{x_t : t = 1, 2, 3, ...\}$ όπου κάθε x_t εκφράζει την κατά τη χρονική στιγμή t κατάσταση ενός συστήματος το οποίο εξελίσσεται κατά τυχαία εν γένει τρόπο (stochastic system) (Κοκολάκης, 2010). Δηλαδή, είναι ένα σύνολο παρατηρήσεων $x_1, x_2, ..., x_t$ καθεμιά από τις οποίες έχει παρατηρηθεί σε συγκεκριμένο χρόνο t.

Μια χρονοσειρά διακριτού χρόνου είναι εκείνη για την οποία το σύνολο T_0 του χρόνου που έχουν παρατηρηθεί οι παρατηρήσεις είναι ένα διακριτό σύνολο όπως για παράδειγμα οι ημερήσιες, αεροπορικές και οδικές αφίξεις τουριστών στη χώρα μας x_t με t=1,2,3,…

Ένα άλλο παράδειγμα χρονοσειράς στη ναυτιλία είναι οι μηνιαίες τιμές των διακυμάνσεων των ναύλων. Συμβολίζουμε συνήθως τις τιμές μιας τέτοιας χρονοσειράς με τον όρο x_i όπου το i παίρνει τιμές από το 1 μέχρι το Τ. Δηλαδή η τιμή x_2 είναι η τιμή της μεταβλητής Χ που μελετάμε, την χρονική περίοδο 2. Στο Διάγραμμα 1.1 παριστάνονται γραφικά οι τιμές x_i των ναύλων για πλοία VLCC (Very Large Crude Carriers) μεταφορικής ικανότητας 280.000 DWT στη γραμμή μεταφοράς αργού πετρελαίου Ρας Τανούρα- Ρότερνταμ, ως προς την χρονική περίοδο παρατήρησης i. Στον άξονα τον x έχουμε το έτος και τον μήνα για 39 χρόνια από τον Ιανουάριο 1979 έως τον Δεκέμβριο του 2017. Εδώ να αναφέρουμε ότι τα δεδομένα αντλήθηκαν από την βάση δεδομένων Clarkson's.



<u>Διάγραμμα 1.1</u>: Μηνιαίες τιμές ναύλων μεταφοράς πετρελαίου με πλοία VLCC .

1.2 Η ναυλαγορά Charter

1.2.1 Η δομή της ναυλαγοράς Charter

Υπάρχουν δύο είδη ναυλαγορών στις οποίες διαιρούνται οι θαλάσσιες μεταφορές. Η πρώτη είναι η ναυλαγορά charter που θα εξεταστεί στην παρούσα διπλωματική και η δεύτερη είναι η ναυλαγορά liner. Η μεγαλύτερη διαφορά μεταξύ των δύο ναυλαγορών είναι ότι στην πρώτη παρουσιάζεται αυτό που ονομάζουμε **«τέλειος ανταγωνισμός»** ενώ στη δεύτερη οι πωλητές της υπηρεσίας οργανώνονται σε καρτέλ, που ονομάζονται κοινοπραξίες, οι οποίες καθορίζουν τον ναύλο για κάθε είδους εμπορεύματος σε συγκεκριμένη διαδρομή και συνεπώς δεν υπάρχει τέλειος ανταγωνισμός.

Θα λέμε ότι υπάρχει ***τέλειος ανταγωνισμός*** σε μια αγορά προϊόντων ή υπηρεσιών αν η τιμή στην οποία προσφέρεται το προϊόν ή η υπηρεσία δεν μπορεί να επηρεαστεί ή να ελεγχθεί ούτε από έναν μεμονωμένο αγοραστή, ούτε από έναν μεμονωμένο πωλητή του προϊόντος ή της υπηρεσίας (Ψαραύτης, 2005).

Με τον όρο υπηρεσία στη ναυλαγορά charter εννοούμε τη διάθεση μεταφορικής ικανότητας. Ο πωλητής είναι ο πλοιοκτήτης που προσφέρει το πλοίο του για ναύλωση και ο αγοραστής είναι ο ναυλωτής. Έτσι, εφόσον επικρατεί τέλειος ανταγωνισμός οι τιμές των ναύλων δε μπορούν να επηρεαστούν ούτε από έναν μεμονωμένο πλοιοκτήτη αλλά ούτε και από έναν μεμονωμένο ναυλωτή. Συνεπώς, οι ναύλοι προσδιορίζονται καθαρά από το ισοζύγιο προσφοράς και ζήτησης μεταφορικής ικανότητας και ο πλοιοκτήτης ναυλώνει ολόκληρο το πλοίο κάτω από αμοιβαία αποδεκτούς όρους μεταξύ εκείνου και του ναυλωτή. Το πλοίο μπορεί να ταξιδέψει οπουδήποτε και για οπουδήποτε χρονικό διάστημα και ο πλοιοκτήτης μπορεί να ναυλώσει το πλοίο του είτε για ένα μόνο ταξίδι είτε ακόμα και για δέκα με δεκαπέντε χρόνια, ανάλογα με το συμβόλαιό του. Έτσι, ο ναυλωτής έχει στην κατοχή του ολόκληρο το πλοίο και μπορεί να κάνει οτιδήποτε με αυτό, ακόμα και να το ναυλώσει σε κάποιον άλλον. Από άποψη του είδους του εμπορεύματος, στη ναυλαγορά charter συνήθως μεταφέρονται μεγάλες ομοιογενείς ποσότητες εμπορευμάτων και το πλοίο μπορεί να είναι εντελώς γεμάτο στη μία κατεύθυνση και άδειο στην επιστροφή. Τα προϊόντα που μεταφέρονται είναι κατά κύριο λόγο χύδην, όπως πετρέλαιο, μεταλλεύματα, κάρβουνο, σιτηρά κ.ο.κ. Το είδος των πλοίων που ανήκουν στην ναυλαγορά charter μπορεί να είναι είτε δεξαμενόπλοια, όπως τα πλοία VLCC που μελετώνται στην παρούσα διπλωματική, είτε bulk carriers, πλοία μεταλλευμάτων, OBO’s κ.ο.κ.

1.2.2 Είδη Ναύλων και Συμβολαίων

Υπάρχουν διάφορα είδη ναύλων και συμβολαίων που πραγματοποιούνται μεταξύ πλοιοκτήτη και ναυλωτή στη ναυλαγορά charter. Καθένα από αυτά κατανέμει με διαφορετικό τρόπο τα κόστη και τις ευθύνες μεταξύ ναυλωτή και πλοιοκτήτη, εκ των οποίων τα βασικότερα είναι τα εξής (Stopford, 2003):

- *Ναύλωση μονού ταξιδιού (voyage charter) :*
  Στη ναύλωση μονού ταξιδιού ο πλοιοκτήτης παραχωρεί το πλοίο του στον ναυλωτή προκειμένου να μεταφερθεί μια συγκεκριμένη ποσότητα ενός εμπορεύματος με ένα προκαθορισμένο πλοίο, από ένα δεδομένο λιμάνι Α, σε ένα δεδομένο λιμάνι Β και μέσα σε δεδομένο χρονικό διάστημα. Για παράδειγμα, έστω ότι ένας έμπορος σιτηρών επιθυμεί να μεταφέρει 25.000 τόνους σιτηρών από το Port Cartier του Καναδά στο Tilbury στο Ηνωμένο Βασίλειο. Αφού καλέσει τον διαπραγματευτή για να τον ενημερώσει ότι χρειάζεται μεταφορά για το φορτίο του, εκείνος με τη σειρά του κανονίζει ένα πλοίο για μεταφορά σε διαπραγματεύσιμη τιμή ναύλου για κάθε τόνο μεταφερόμενου φορτίου (π.χ. 5.20 $ / ton). Εφόσον καθοριστούν οι όροι μεταξύ ναυλωτή και πλοιοκτήτη, το πλοίο αναμένεται να καταφτάσει στην συμφωνημένη ημερομηνία, φορτώνει το εμπόρευμα, το μεταφέρει στο Tilbury, έπειτα ξεφορτώνει και η συμφωνία ολοκληρώνεται.
  Εδώ ο πλοιοκτήτης πληρώνει όλα τα έξοδα που απαιτούνται για τη λειτουργία του πλοίου όπως τα καύσιμα, το πλήρωμα κ.ο.κ με πιθανή εξαίρεση ίσως τα έξοδα φορτοεκφόρτωσης. Για τη ναύλωση μονού ταξιδιού υπάρχουν τρεις κατηγορίες :

  a) **Άμεση:** Εκτελείται μέσα σε μερικές εβδομάδες από την υπογραφή του συμβολαίου και ο αντίστοιχος ναύλος λέγεται στιγμιαίος ναύλος (spot rate).
  b) **Μελλοντική (forward charter):** Εκτελείται κάποια χρονική στιγμή στο μέλλον π.χ. σε δύο μήνες.
  c) **Επαναληπτική (consecutive):** όταν αφορά έναν αριθμό από όμοια επαναληπτικά ταξίδια.

- *Χρονοναύλωση (term charter) ή Ναύλωση προθεσμίας:*
  Στην περίπτωση της χρονοναύλωσης νοικιάζεται το πλοίο μαζί με το πλήρωμά του για ένα προκαθορισμένο χρονικό διάστημα. Ο πλοιοκτήτης παρέχει στον ναυλωτή το πλήρωμα καθώς και τη συντήρησή του και εγγυάται ότι το πλοίο ικανοποιεί διάφορα κριτήρια απόδοσης (ταχύτητα, κατανάλωση κ.ο.κ). Η τιμή εδώ ορίζεται σε $/ton DWT/μήνα. Η διάρκεια ναύλωσης μπορεί να είναι ο χρόνος που απαιτείται για την ολοκλήρωση ενός μονού ταξιδιού (trip charter) ή μιας περιόδου μερικών μηνών ή ακόμα και ετών (period charter). Σε αντίθεση με τη ναύλωση μονού ταξιδιού ο ναυλωτής πληρώνει ξεχωριστά τα καύσιμα, τα έξοδα λειτουργίας του σκάφους, λιμενικά τέλη και έξοδα

φορτοεκφόρτωσης. Όσο διαρκεί η χρονοναύλωση, ο ναυλωτής έχει τη δυνατότητα να χειριστεί το πλοίο όπως εκείνος θέλει, ακόμα και να το ναυλώσει σε κάποιον άλλον. Όπως και στη ναύλωση μονού ταξιδιού και εδώ έχουμε διάφορες κατηγορίες ναυλώσεων όπως άμεση μελλοντική και «bareboat». Στην τελευταία περίπτωση ο ναυλωτής παρέχει και το πλήρωμα.

- *Συμβόλαιο φόρτωσης (contact of affreightment) :*
  Το συμβόλαιο φόρτωσης είναι λίγο πιο περίπλοκο και παρουσιάζει ομοιότητες με την επαναληπτική ναύλωση. Η διαφορά εδώ είναι ότι το όνομα του πλοίου δεν προκαθορίζεται. Στο συμβόλαιο φόρτωσης, ο πλοιοκτήτης συμφωνεί να μεταφέρει μια σειρά φορτίων με σταθερή τιμή ανά τόνο. Για παράδειγμα, ο φορτωτής μπορεί να έχει μια σύμβαση για να προμηθευτεί 50.000 τόνους άνθρακα από την Κολομβία στο Ρότερνταμ σε διάστημα δύο μηνών. Θα επιθυμούσε να κανονίσει τη σύμβασή του σε μια ενιαία σύμβαση με συμφωνημένη τιμή ανά τόνο και να αφήσει τις υπόλοιπες λεπτομέρειες κάθε ταξιδιού στον πλοιοκτήτη. Αυτό επιτρέπει στον πλοιοκτήτη να έχει την ελευθερία να χρησιμοποιήσει οποιοδήποτε πλοίο θελήσει προκειμένου να εκπληρώσει τις υποχρεώσεις του σύμφωνα με το συμβόλαιο.

1.2.3 Η Ναυλαγορά Tankers

Ένα πολύ σημαντικό μέρος της ναυλαγοράς charter είναι αυτή των δεξαμενοπλοίων, η οποία μελετάται και στην παρούσα διπλωματική εργασία. Πρόκειται για τη ναυλαγορά Tankers και αυτό που την καθιστά ενδιαφέρουσα είναι οι διακυμάνσεις των ναύλων. Στην προκειμένη περίπτωση, μπορούν να παρουσιαστούν και διακυμάνσεις του στιγμιαίου ναύλου της τάξης άνω του 500% αφήνοντας τεράστια περιθώρια κέρδους και ζημίας τόσο για τον πλοιοκτήτη όσο και για τον ναυλωτή.

Στον κύκλο της ναυτιλίας ένας πλοιοκτήτης έρχεται αντιμέτωπος με διάφορα διλήμματα σχετικά με το ποια είναι η πιο συμφέρουσα επιλογή για εκείνον να κάνει με το πλοίο του. Από τις επιλογές που έχει στην κατοχή του, οι κυριότερες είναι οι εξής :

- Να διαθέσει το πλοίο του στη Στιγμιαία ναυλαγορά (Spot market).
- Να χρονοναυλώσει το πλοίο του (Charter market).
- Να παροπλίσει το πλοίο του σε κάποιο λιμάνι.
- Να αποσύρει το πλοίο του, καταστρέφοντάς το.

Στη στιγμιαία ναυλαγορά, ο πλοιοκτήτης από τη στιγμή που κλείσει το συμβόλαιό του με τον ναυλωτή, τον ενημερώνει για το ποσό εκείνο των χρημάτων που θα πρέπει να του διαθέσει ανάλογα με τις τρέχουσες τιμές των ναύλων και το συμφωνημένο ταξίδι εκτελείτε άμεσα. Από την άλλη μεριά, στην περίπτωση της χρονοναύλωσης, ο

πλοιοκτήτης συμφωνεί μια σταθερή τιμή που θα πρέπει ο ναυλωτής να καταβάλλει σε αυτόν, για ένα μεγαλύτερο χρονικό διάστημα, ρισκάροντας σε περίπτωση όπου τα ναύλα παρουσιάσουν ανοδική πορεία τις επόμενες χρονικές περιόδους, να έχει δεσμευτεί σε σταθερά έσοδα. Στην περίπτωση που αποφασίσει να παροπλίσει το πλοίο του σημαίνει ότι σταματά να το χρησιμοποιεί για ένα συγκεκριμένο χρονικό διάστημα και είναι απλώς αγκυροβολημένο σε κάποιο λιμάνι. Ο λόγος για να παροπλίσει ένας πλοιοκτήτης το πλοίο του μπορεί να είναι είτε επειδή περιμένει σε μια καλύτερη τιμή διάλυσης, είτε ακόμα επειδή μπορεί να παρουσιάζεται υπερβολική προσφορά μεταφορικής ικανότητας σε μια συγκεκριμένη περίοδο και αναμένει να το διαθέσει στην αγορά κάποια στιγμή στο μέλλον. Οι περίοδοι παροπλισμού ενός πλοίου μπορεί να είναι σύντομοι, όπως μερικές εβδομάδες αλλά και μακροπρόθεσμοι όπως πέντε ή και περισσότερα χρόνια. Προφανώς, τα πάντα εξαρτώνται από την ικανότητα να προβλέψει κανείς τις διακυμάνσεις των ναύλων οι οποίες παίζουν κεντρικό ρόλο για τη λήψη αποφάσεων στη ναυτιλία.

1.2.4 Ο δείκτης Worldscale

Ο δείκτης Worldscale είναι ένας σχετικά εύκολος μηχανισμός να περιγράψει κανείς τις διακυμάνσεις του στιγμιαίου ναύλου. Η βιομηχανία δεξαμενόπλοιων χρησιμοποιεί τον συγκεκριμένο δείκτη των ναύλων ως ένα βολικό τρόπο διαπραγμάτευσης του ναύλου ανά βαρέλι μεταφερόμενου πετρελαίου σε διαφορετικές διαδρομές. Για να περιγράψουμε συστηματικά την αγορά υπάρχουν δύο τρόποι:

a) Να θεωρήσουμε μια αντιπροσωπευτική διαδρομή και να χρησιμοποιήσουμε τον στιγμιαίο ναύλο στη διαδρομή αυτή για να περιγράψουμε την αγορά.
b) Να θεωρήσουμε έναν αντιπροσωπευτικό πλοιοκτήτη και να υπολογίσουμε τον στιγμιαίο ναύλο που θα έκανε τον πλοιοκτήτη να δεχτεί τον ναύλο αυτό αντί να παροπλίσει το πλοίο του.

Ο δείκτης Worldscale δημοσιεύεται σε ένα βιβλίο που χρησιμοποιείται ως βάση για τον υπολογισμό του στιγμιαίου ναύλου των δεξαμενοπλοίων. Το βιβλίο δείχνει για κάθε διαδρομή των δεξαμενοπλοίων, το κόστος μεταφοράς ενός τόνου φορτίου, χρησιμοποιώντας τα δεδομένα ενός πρότυπου πλοίου σε ένα γύρο ταξιδιού. Αυτό το κόστος είναι γνωστό ως Worldscale 100 και η υπόθεση για το πρότυπο πλοίο που χρησιμοποιείται ανανεώνεται ανά χρονικά διαστήματα. Στον Πίνακα 1 παρουσιάζεται η υπόθεση του πρότυπου πλοίου που έγινε το 2007. Συνεπώς, για την περιγραφή της ναυλαγοράς tankers στην περίπτωση b) όπου θεωρούμε έναν αντιπροσωπευτικό πλοιοκτήτη, γίνεται υπόθεση για τον πλοιοκτήτη αυτόν να διαθέτει ένα tanker που ικανοποιεί τα χαρακτηριστικά του Πίνακα 1.

Πίνακας 1 : Πρότυπο πλοίο Tanker	
Συνολική χωρητικότητα:	75.000 τόνοι DWT
Βύθισμα (θαλάσσιο νερό):	30,5 πόδια
Μέση οικονομική ταχύτητα:	14,5 κόμβοι
Κατανάλωση καυσίμου:	
• Εν πλω	55 τόνοι/ ημέρα
• Λιμάνι	110 τόνοι/ ταξίδι
Χρόνος στο λιμάνι:	96 ώρες
Fixed Hire Element:	12.000 $ / ημέρα
Μεσιτικά έξοδα:	2,5 %

Εδώ να σημειώσουμε ότι το « fixed hire element » είναι καθαρά εικονική τιμή που δείχνει πόσο χάνει ο πλοιοκτήτης ανά ημέρα αν διαθέσει το πλοίο του στη ναυλαγορά αντί να το παροπλίσει.

Ο δείκτης Worldscale, τον οποίο εν συντομία συμβολίζουμε WS, διευκολύνει τους ναυλωτές και τους πλοιοκτήτες να συγκρίνουν τα κέρδη των πλοίων τους σε διαφορετικές διαδρομές. Ας υποθέσουμε ότι ένα δεξαμενόπλοιο είναι διαθέσιμο στη στιγμιαία ναυλαγορά, δηλαδή περιμένει διαθέσιμο φορτίο για να μεταφερθεί και ο ιδιοκτήτης συμφωνεί ένα κόστος WS 50 για ένα ταξίδι από το Jubail στο Rotterdam. Για να υπολογίσει πόσα χρήματα θα κερδίσει, πρώτα κοιτάει πόση είναι η τιμή του ναύλου ανά τόνο για τον δείκτη WS 100 από το Jubail στο Rotterdam. Εάν υποθέσουμε ότι η τιμή του δείκτη WS 100 είναι 17.30 $/ τόνο, για τη συγκεκριμένη διαδρομή, τότε εφόσον ο ναύλος έχει κανονιστεί να είναι WS 50 θα λάβει το μισά χρήματα από αυτό το ποσό, δηλαδή 8.65 $/ τόνο. Επομένως, αν το πλοίο του μπορεί να μεταφέρει φορτίο 250.000 τόνων τότε τα έσοδα του πλοιοκτήτη για την εκπλήρωση αυτού του ταξιδιού θα ανέρχονται στα $8,65 \times 250.000 = 2.162.500\ \$$.

Ο δείκτης Worldscale για μια συγκεκριμένη διαδρομή υπολογίζεται ως εξής :

$$WS = \frac{\Sigma\tau\iota\gamma\mu\iota\alpha\acute{\iota}ος\ \text{N}\alpha\acute{\upsilon}λος\ \sigma\tau\eta\ \Delta\iota\alpha\delta\rho\omega\mu\acute{\eta}}{\text{B}\alpha\sigma\iota\kappa\acute{ο}ς\ \text{N}\alpha\acute{\upsilon}λος\ \sigma\tau\eta\ \Delta\iota\alpha\delta\rho\omega\mu\acute{\eta}} \times 100\ ,$$

όπου,

- Ο στιγμιαίος ναύλος είναι ο ναύλος που έχει προκύψει από τον ελεύθερο ανταγωνισμό της αγοράς (μετριέται σε $/ton ωφέλιμου φορτίου).
- Ο βασικός ναύλος είναι εκείνος ο ναύλος ο οποίος μόλις θα καλύπτει τα έξοδα του ταξιδιού (μετριέται σε $/ton ωφέλιμου φορτίου).

Συνεπώς, ο βασικός ναύλος για μια συγκεκριμένη διαδρομή και για τις τρέχουσες τιμές καυσίμων και λοιπών εξόδων, υπολογίζεται να είναι ο ναύλος εκείνος που μόλις θα καλύπτει τα έξοδα του ταξιδιού (με επιστροφή) που αφορούν καύσιμα, λιμενικά τέλη, διόδια καναλιών, συν κάποια δολάρια τη μέρα για το tanker αυτό των 75.000 τόνων DWT.

Κεφάλαιο 2

2. Βασικά μονοπαραμετρικά μοντέλα

2.1 Στάσιμες και μη Στάσιμες Χρονοσειρές

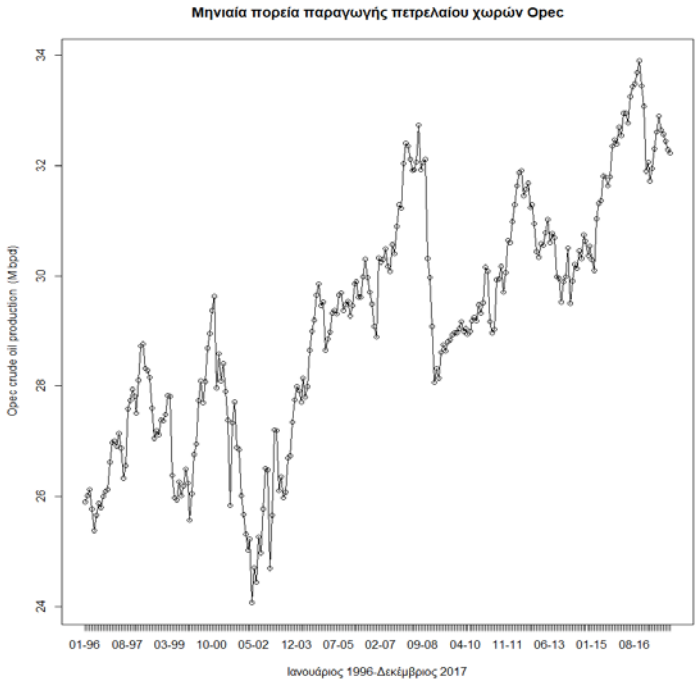
Ένα βασικό ρόλο στην ανάλυση χρονοσειρών αντιπροσωπεύουν οι διαδικασίες των οποίων οι ιδιότητες τους ή τουλάχιστον μερικές από αυτές δεν αλλάζουν με τον χρόνο. Εάν θέλουμε να κάνουμε προβλέψεις τότε το πρώτο πράγμα που θα πρέπει να υποθέσουμε είναι ότι «κάτι» δεν αλλάζει με τον χρόνο και συνεπώς είναι αναγκαίο να ορίσουμε ένα πολύ σημαντικό είδος χρονοσειρών, που είναι οι στάσιμες χρονοσειρές. Δεδομένου ότι η στασιμότητα είναι μια υπόθεση στην οποία βασίζονται πολλές στατιστικές διαδικασίες που χρησιμοποιούνται στη ανάλυση χρονοσειρών, τα μη-στάσιμα δεδομένα συνήθως μετασχηματίζονται έτσι ώστε να γίνουν στάσιμα.

Μια χρονοσειρά $\{X_t, t = 0,1,2,...\}$ θα λέγεται στάσιμη αν έχει στατιστικές ιδιότητες (συνάρτηση μέση τιμής, διασποράς, αυτοσυνδιακύμανσης κ.τ.λ) ίδιες με αυτές της χρονικά μετατοπισμένης σειράς $\{X_{t+h}, t = 0,1,2,...\}$ για κάθε ακέραιο *h*. Αυτό συμβαίνει όταν η από κοινού κατανομή πιθανότητας των παρατηρήσεων $x_1, x_2, ..., x_n$ είναι ακριβώς η ίδια με την από κοινού κατανομή πιθανότητας των παρατηρήσεων $x_{1+h}, x_{2+h}, ..., x_{n+h}$ για κάθε $h \in \mathbb{Z}$ και $n > 0$. Τότε λέμε ότι η χρονοσειρά είναι ***αυστηρώς*** στάσιμη. Αργότερα, θα μιλήσουμε και για την ***ασθενή*** στασιμότητα.

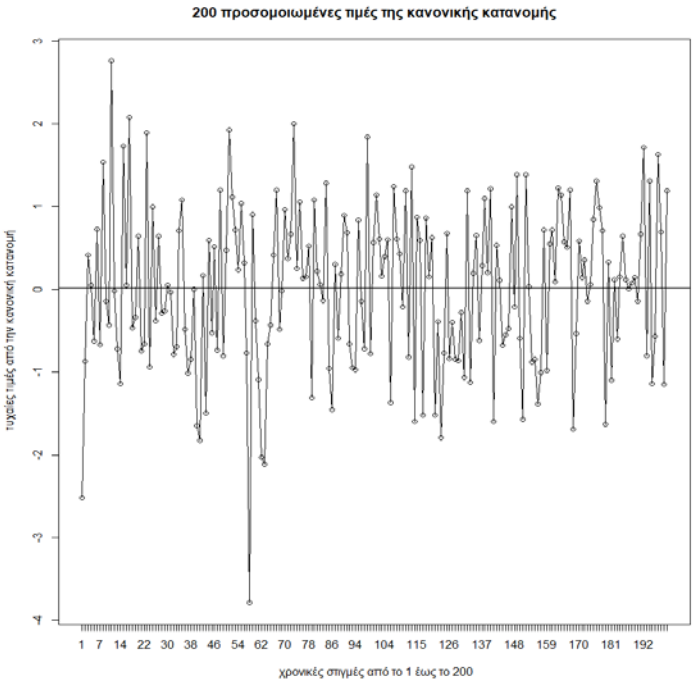
Η πιο συνηθισμένη αιτία παράβασης της στασιμότητας είναι μια τάση στον μέσο όρο. Αυτή η αλλαγή της μέσης τιμής με την πάροδο του χρόνου μπορεί να προκαλέσει υποτιμημένες προβλέψεις για το μέλλον.

Στα Διαγράμματα 2.1.1 και 2.1.2 παρουσιάζονται γραφικά δύο παραδείγματα χρονοσειρών που προέρχονται, κατ' αντιστοιχία, από στάσιμες και μη-στάσιμες διαδικασίες. Πιο συγκεκριμένα, στο Διάγραμμα 2.1.1 παρουσιάζεται η χρονική εξέλιξη της παραγωγής πετρελαίου των χωρών OPEC, οι τιμές της οποίας αντλήθηκαν από την βάση δεδομένων Clarkson's. Στον άξονα x έχουμε το έτος και τον μήνα για 22 χρόνια από τον Ιανουάριο του 1996 έως τον Δεκέμβριο του 2017. Στον άξονα y έχουμε τις αντίστοιχες μηνιαίες ποσότητες παραγωγής πετρελαίου των χωρών OPEC, σε εκατομμύρια βαρέλια. Η παραγωγή του πετρελαίου μπορεί να δώσει μια εικόνα της ζήτησης μεταφορικής ικανότητας για τα πλοία VLCC αφού τα πλοία αυτά μεταφέρουν πετρέλαιο και συνεπώς η ζήτησή τους για μεταφορά πετρελαίου ενδέχεται να επηρεάζεται από αυτή τη μεταβλητή. Με μια πρώτη ματιά στο Διάγραμμα 2.1.1, είναι εμφανής μια ανοδική τάση στην παραγωγή του πετρελαίου των χωρών OPEC για τα 22 αυτά χρόνια, γεγονός που φανερώνει ότι τα δεδομένα μας δεν προέρχονται από μια στάσιμη διαδικασία.

Από την άλλη μεριά, στο Διάγραμμα 2.1.2, έχουμε 200 τυχαίες προσομοιωμένες τιμές από την κανονική κατανομή $N(0,1)$, με μέση τιμή μηδέν και διασπορά ένα. Εφόσον η μέση τιμή είναι σταθερά ίση με μηδέν, δηλαδή δεν αποτελεί συνάρτηση του χρόνου t, και επιπλέον τα δεδομένα μας δεν φαίνεται να παρουσιάζουν μη σταθερές αποκλίσεις από τη μέση τιμή, μπορούμε να υποθέσουμε ότι ενδέχεται να προέρχονται από μια στάσιμη διαδικασία. Εναλλακτικά, θα λέμε ότι τα δεδομένα μας είναι μια «πραγματοποίηση» μιας στάσιμης διαδικασίας.



<u>Διάγραμμα 2.1.1</u>: Χρονοσειρά παραγωγής πετρελαίου χωρών Opec



<u>Διάγραμμα 2.1.2</u>: Χρονοσειρά 200 προσομοιωμένων τιμών από την $N(0,1)$

Πριν ορίσουμε την ασθενή στασιμότητα, πρέπει να πούμε ότι για μια χρονοσειρά $\{X_t\}$ με $E(X_t^2)<\infty$, η συνάρτηση μέσης τιμής της $\{X_t\}$ ορίζεται να είναι: $\mu_X(t)=E(X_t)$. Ακόμη η συνάρτηση συνδιακύμανσης της $\{X_t\}$ ορίζεται να είναι :

$\gamma_X(r,s)=Cov(X_r,X_s)=E\left[\left(X_r-\mu_X(r)\right)\left(X_s-\mu_X(s)\right)\right]$ για κάθε ακέραιο r και s.

Θα λέμε ότι η $\{X_t\}$ είναι ***(ασθενώς)*** στάσιμη χρονοσειρά εάν :

i. Η συνάρτηση μέσης τιμής $\mu_X(t)$ είναι ανεξάρτητη από το *t*.

ii. Η συνάρτηση αυτοσυνδιακύμανσης $\gamma_X(t+h,t) = Cov(X_{t+h}, X_t) = E\left[\left(X_{t+h} - \mu_X(t)\right)\left(X_t - \mu_X(t)\right)\right]$ είναι ανεξάρτητη του *t* για κάθε *h*.

Θεωρούμε ότι εάν η $\{X_t\}$ είναι *αυστηρώς* στάσιμη και ισχύει ότι $E\left(X_t^2\right) < \infty$ για όλα τα t, τότε η $\{X_t\}$ είναι επίσης *ασθενώς* στάσιμη (Brockwell & Davis, 2006, p. 13). Στο εξής όταν αναφερόμαστε στην στασιμότητα θα εννοούμε την *ασθενή* στασιμότητα όπως ορίστηκε παραπάνω. Μερικές ιδιότητες που θα φανούν χρήσιμες σε παρακάτω υπολογισμούς και αφορούν τη συνάρτηση αυτοσυνδιακύμανσης $\gamma(\cdot)$ είναι οι εξής:

$$\gamma(0) \geq 0$$
$$|\gamma(h)| \leq \gamma(0) \text{ για όλα τα } h$$
$$\text{Η } \gamma(\cdot) \text{ είναι άρτια, δηλαδή}$$
$$\gamma(h) = Cov(X_{t+h}, X_t) = Cov(X_t, X_{t+h}) = \gamma(-h)$$

2.2 Βασικοί Τελεστές

Όπως αναφέραμε και προηγουμένως, εάν τα δεδομένα μας προέρχονται από μη-στάσιμη διαδικασία, τότε ένας μετασχηματισμός μπορεί εύκολα να τα μετατρέψει σε στάσιμα έτσι ώστε να μπορέσουμε να τα μοντελοποιήσουμε και να κάνουμε μελλοντικές προβλέψεις. Μια προσέγγιση για να επιτευχθεί αυτό, αναπτύχθηκε από τους Box και Jenkins (1976), οι οποίοι προτείνουν να εφαρμόσουμε ένα τελεστή διαφοράς, επαναλαμβανόμενα στην αρχική σειρά $\{X_t\}$ μέχρι οι διαφορές των παρατηρήσεων να μοιάζουν με μια πραγματοποίηση μιας στάσιμης διαδικασίας $\{Y_t\}$. Αμέσως μετά, μπορούμε να χρησιμοποιήσουμε τη θεωρία των στάσιμων διαδικασιών για να μοντελοποιήσουμε, αναλύσουμε και προβλέψουμε την $\{Y_t\}$ και έπειτα την αρχική διαδικασία.

Ορίζουμε τον τελεστή πρώτης διαφοράς (lag-1 difference operator) ∇ ως εξής :

$$\nabla X_t = X_t - X_{t-1} = (1-B)X_t,$$

όπου ο B είναι ο τελεστής οπίσθιας μετάθεσης (backward shift operator) τέτοιος ώστε:

$$BX_t = X_{t-1}\ .$$

Με παρόμοιο τρόπο ορίζονται και οι τελεστές B και ∇ μεγαλύτερης διαφοράς ως εξής:

$$B^j\left(X_t\right)=X_{t-j} \text{ και } \nabla^j\left(X_t\right)=\nabla\left(\nabla^{j-1}\left(X_t\right)\right),\ \mathrm{j}\geq 1 \text{ με } \nabla^0\left(X_t\right)=X_t\ .$$

Τα πολυώνυμα που προκύπτουν εφαρμόζοντας τους τελεστές B και ∇ χειρίζονται ακριβώς με τον ίδιο τρόπο όπως οι πολυωνυμικές συναρτήσεις πραγματικών μεταβλητών. Για παράδειγμα,

$$\nabla^2\left(X_t\right)=\nabla\left(\nabla\left(X_t\right)\right)=\left(1-B\right)\left(1-B\right)X_t=\left(1-2B+B^2\right)X_t=X_t-2X_{t-1}+X_{t-2}\ .$$

Αν υποθέσουμε ότι έχουμε δεδομένα των οποίων η τάση μπορεί να εκτιμηθεί προσαρμόζοντας μια γραμμική συνάρτηση της μορφής $m_t=c_0+c_1t$, όπου $c_0,c_1\in\mathbb{R}$, τότε εφαρμόζοντας τον τελεστή ∇ έχουμε:

$\nabla m_t=m_t-m_{t-1}=c_0+c_1t-\left(c_0+c_1\left(t-1\right)\right)=c_1$, δηλαδή μια σταθερά.

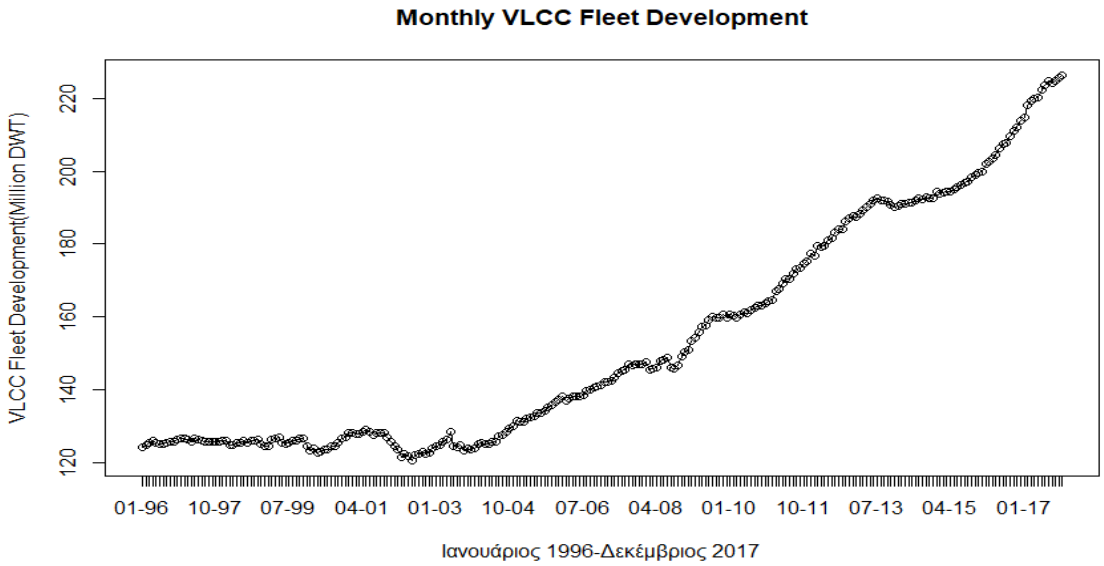
Το Διάγραμμα 2.2.1 δείχνει την χρονοσειρά παγκόσμιας ανάπτυξης στόλου VLCC πλοίων (fleet development) σε εκατομμύρια τόνους μεταφορικής ικανότητας DWT για την περίοδο από τον Ιανουάριο του 1996 έως τον Δεκέμβριο του 2017. Τα δεδομένα αντλήθηκαν από την βάση δεδομένων Clarkson's. Από το γράφημα φαίνεται να ταιριάζει να προσαρμόσουμε στα δεδομένα μας είτε μια εκθετική τάση, είτε μια πολυωνυμική τάση δευτέρου βαθμού της μορφής $m_t=\alpha_0+\alpha_1t+a_2t^2$, όπου $\alpha_0,\alpha_1,\alpha_2\in\mathbb{R}$. Με την εφαρμογή του τελεστή ∇^2 στην πολυωνυμική εξίσωση τάσης δευτέρου βαθμού έχουμε:

$$\begin{aligned}\nabla^2m_t&=\nabla\left(\nabla m_t\right)=\nabla\left(c_0+c_1t+c_2t^2-\left(c_0+c_1(t-1)+c_2(t-1)^2\right)\right)=\nabla\left(c_1+2c_2t-c_2\right)\\&=c_1+2c_2t-c_2-\left(c_1+2c_2(t-1)-c_2\right)=2c_2\end{aligned}$$

Δηλαδή ο τελεστής ∇^2 έδωσε μια τάση σταθερή και ανεξάρτητη του χρόνου t.

Κατά τον ίδιο τρόπο, οποιαδήποτε πολυωνυμική τάση βαθμού k, μπορεί να μειωθεί σε μια σταθερά με την εφαρμογή του τελεστή ∇^k. Για παράδειγμα, έστω ότι έχουμε δεδομένα x_t τα οποία υποθέτουμε ότι ικανοποιούν την εξίσωση $X_t=m_t+e_t$, όπου m_t μια πολυωνυμική συνάρτηση τάσης της μορφής $m_t=\sum_{j=0}^{k}c_jt^j$ με $c_j\in\mathbb{R}$ και e_t το σφάλμα. Εάν υποθέσουμε ότι το σφάλμα είναι μια στάσιμη διαδικασία με μέση τιμή μηδέν, τότε η εφαρμογή του τελεστή ∇^k δίνει : $\nabla^kX_t=k!c_k+\nabla^ke_t$, δηλαδή μια στάσιμη διαδικασία με μέση τιμή σταθερή και ίση με $k!c_k$ (Brockwell & Davis, 2002, p. 30). Με βάση αυτά τα συμπεράσματα, μπορούμε να πούμε ότι είναι πιθανό σε οποιοδήποτε ακολουθία δεδομένων $\left\{x_t\right\}$ να εφαρμόσουμε τον τελεστή ∇ επανειλημμένως έως ότου βρούμε μια ακολουθία $\left\{\nabla^kx_t\right\}$ που μπορεί ευλόγως να μοντελοποιηθεί σαν μια πραγματοποίηση μιας στάσιμης διαδικασίας. Εδώ είναι

σημαντικό να πούμε ότι στην πράξη, η τάξη k των διαφορών που απαιτούνται να εφαρμόσουμε, προκειμένου να προκύψει μια στάσιμη διαδικασία, είναι μικρή με πιο συνηθισμένες τις τιμές 1 ή 2. Αυτό προκύπτει από το γεγονός ότι πολλές συναρτήσεις μπορούν να προσεγγιστούν σε ένα διάστημα πεπερασμένου μήκους από ένα πολυώνυμο πολύ μικρού βαθμού.



<u>Διάγραμμα 2.2.1</u>: Χρονοσειρά παγκόσμιας ανάπτυξης στόλου VLCC πλοίων

Μια χρήσιμη μέθοδος εκτίμησης της τάσης m_t είναι η μέθοδος ελαχίστων τετραγώνων (least squares method). Η μέθοδος ελαχίστων τετραγώνων, στην περίπτωση που υποθέσουμε ότι η τάση των δεδομένων μας περιγράφεται καλύτερα από ένα πολυώνυμο δευτέρου βαθμού, προσαρμόζει μια παραμετρική οικογένεια συναρτήσεων της μορφής $m_t = \alpha_0 + \alpha_1 t + a_2 t^2$ στα δεδομένα $\{x_1, x_2, ..., x_n\}$ επιλέγοντας ως παραμέτρους $\alpha_0, \alpha_1, \alpha_2$ να είναι πραγματικοί αριθμοί τέτοιοι ώστε να ελαχιστοποιείται το $\sum_{t=1}^{n} (x_t - m_t)^2$. Με αντίστοιχο τρόπο η μέθοδος χρησιμοποιείται και για μεγαλύτερης τάξης πολυωνυμικές τάσεις.

2.3 Κλασική Αποσύνθεση (Classical Decomposition)

Το πρώτο βήμα για να αναλύσουμε οποιαδήποτε χρονοσειρά είναι να δούμε το γράφημα των παρατηρήσεων. Αν δεν υπάρχουν εμφανείς ασυνέχειες, όπως για παράδειγμα μια απότομη αλλαγή του επιπέδου, μπορεί να είναι ενδεδειγμένο να αναλύσουμε τη χρονοσειρά πρώτα σπάζοντας την σε ομογενή τμήματα. Αν υπάρχουν οποιεσδήποτε ακραίες παρατηρήσεις πρέπει να μελετηθούν προσεκτικά έτσι ώστε να μπορούμε να δικαιολογήσουμε την ύπαρξη τους ή να τα απορρίψουμε. Μια απλή μέθοδος για να περιγράψουμε μια σειρά είναι αυτή της κλασικής αποσύνθεσης.

Η ιδέα είναι ότι η σειρά μπορεί να αναλυθεί σε τρείς συνιστώσες :

m_t : Μια αργά μεταβαλλόμενη συνάρτηση την οποία ονομάζουμε συνιστώσα τάσης. Υποδηλώνει μακράς διάρκειας αλλαγές στη μέση τιμή.

s_t : Μια συνάρτηση με γνωστή περίοδο *d* την οποία ονομάζουμε εποχιακή συνιστώσα. Υποδηλώνει κυκλικές διακυμάνσεις που σχετίζονται με τους μήνες.

e_t : Είναι μια τυχαία συνιστώσα, την οποία ονομάζουμε θόρυβο και την θεωρούμε στάσιμη με την έννοια της ασθενούς στασιμότητας.

Σκοπός είναι να δημιουργήσουμε ξεχωριστά μοντέλα για αυτές τις 4 συνιστώσες και μετά να τις συνδέσουμε είτε προσθετικά (the classical decomposition model) :

$$X_t = m_t + s_t + e_t \qquad (2.3.1)$$

Είτε πολλαπλασιαστικά

$$X_t = m_t \bullet s_t \bullet e_t \qquad (2.3.2)$$

Μερικά μοντέλα λαμβάνουν υπόψιν και μια επιπλέον συνιστώσα, την οποία ονομάζουν κυκλική συνιστώσα. Ωστόσο, επειδή πολύ συχνά θεωρείται ότι η κυκλική συνιστώσα αποτελεί μέρος της γενικής τάσης, δεν λαμβάνεται υπόψιν σαν ξεχωριστή συνιστώσα. Σε αντίθεση με την περιοδική συνιστώσα των μοντέλων (2.3.1) και (2.3.2), ένας κύκλος δεν έχει ένα σταθερό διάστημα που παρατηρείται και συνεπώς σε ένα σύνολο δεδομένων μπορεί ο πρώτος κύκλος να είναι τέσσερις μήνες και ο επόμενος δύο χρόνια.

Η επιλογή για το πιο από τα δύο μοντέλα είναι καταλληλότερο να αναπαραστήσει τα δεδομένα μας, δεν είναι πάντα εμφανής απλά παρατηρώντας το γράφημα των δεδομένων μιας χρονοσειράς. Ένας γενικός κανόνας είναι ότι σε περίπτωση που βλέπουμε τα δεδομένα μιας χρονοσειράς να παρουσιάζουν δραστική μείωση ή αύξηση στο πλάτος της χρονοσειράς προς τα τελευταία παρατηρούμενα χρόνια, τότε ενδείκνυται να δοκιμάσουμε έναν μετασχηματισμό πρώτα στα δεδομένα μας, όπως για παράδειγμα να τα λογαριθμήσουμε ή το πολλαπλασιαστικό μοντέλο της σχέσης (2.3.2). Αντίστοιχα, όταν βλέπουμε ότι η εποχικότητα και ο θόρυβος εμφανίζει να

αυξάνεται σε συνάρτηση με τον χρόνο, πάλι ένας πρωταρχικός μετασχηματισμός με λογαρίθμους θα κάνει πιο συμβατό το μοντέλο μας στη σχέση (2.3.1).

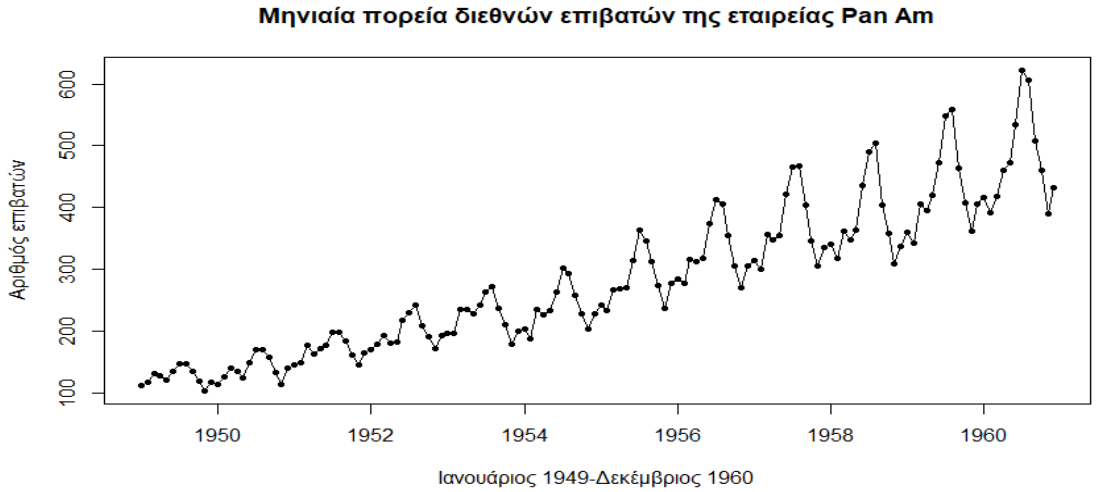
$\ln(X_t) = \ln(m_t \cdot s_t \cdot e_t) = \ln(m_t) + \ln(s_t) + \ln(e_t)$.

Βλέπουμε λοιπόν ότι σε περίπτωση που θεωρήσουμε ότι η χρονοσειρά μας περιγράφεται από την σχέση $X_t = m_t \bullet s_t \bullet e_t$, εύκολα μετασχηματίζεται στο προσθετικό μοντέλο παίρνοντας λογαρίθμους. Σκοπός μας είναι να εκτιμήσουμε και να αποσπάσουμε τις ντετερμινιστικές συνιστώσες m_t και s_t με την ελπίδα ότι τα υπόλοιπα (η συνιστώσα θορύβου) e_t θα είναι μια στάσιμη τυχαία διαδικασία. Έπειτα μπορούμε να χρησιμοποιήσουμε τη θεωρία μιας τέτοιας διαδικασίας για να βρούμε ένα ικανοποιητικό μοντέλο πιθανότητας για τη διαδικασία e_t, να αναλύσουμε τις ιδιότητες της και σε συνδυασμό με τις συνιστώσες m_t και s_t να προβλέψουμε την προσομοίωση της $\{X_t\}$.

Γενικά, η μέθοδος της κλασικής αποσύνθεσης παρότι είναι εύκολη και απλή στη χρήση παρουσιάζει αρκετά μειονεκτήματα. Ένα βασικό μειονέκτημά της είναι ότι υποθέτει πως η εποχιακή συνιστώσα μένει σταθερή σε διάφορες χρονικές περιόδους, γεγονός που στην πράξη πολλές φορές δε συναντάται. Ωστόσο, το πρόβλημα της σταθερής περιόδου λύνεται μέσω της χρήσης πιο προχωρημένων μεθόδων που παρέχει η R, όπως την STL αποσύνθεση που πραγματοποιείται μέσω της εντολής stl του πακέτου forecast (Hyndman, et al., 2018) και επιτρέπει την εποχιακή συνιστώσα να προσαρμόζεται κατάλληλα σε διάφορα χρονικά διαστήματα. Επιπροσθέτως, η κλασική αποσύνθεση δεν μπορεί να διαχειριστεί την περίπτωση όπου τα δεδομένα έχουν κάποιες αγνοούμενες τιμές και είναι αργή στο να πιάσει τις απότομες αλλαγές στο γράφημα μιας χρονοσειράς. Τέλος, το γεγονός ότι τα μοντέλα (2.3.1) και (2.3.2) θεωρούν πως οι συνιστώσες της τάσης, της περιοδικότητας και του θορύβου συμβάλουν στα αρχικά δεδομένα είτε προσθετικά είτε πολλαπλασιαστικά, είναι μια υπόθεση που μπορεί να αποδειχθεί εσφαλμένη. Παρ' όλα αυτά με την κλασική αποσύνθεση μπορούμε να πάρουμε μια γρήγορη ιδέα για τα δεδομένα μας με έναν εύκολο τρόπο.

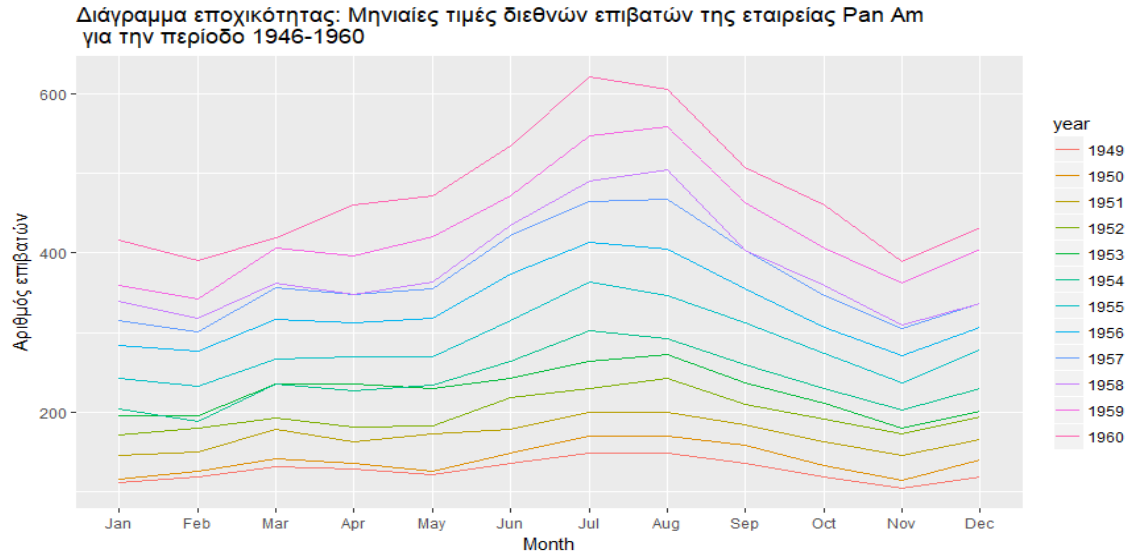
Στο Διάγραμμα 2.3.1 απεικονίζεται ο αριθμός των διεθνών επιβατών ανά μήνα της αεροπορικής εταιρείας Pan Am των Ηνωμένων Πολιτειών για την περίοδο 1949-1960. Τα δεδομένα βρίσκονται ελεύθερα στη γλώσσα παραγραμματισμού R τα οποία φορτώνουμε μέσω της εντολής AirPassengers.

Η εταιρεία Pan Am χρησιμοποίησε τα δεδομένα για να προβλέψει τη μελλοντική ζήτηση προτού παραγγείλει νέα αεροσκάφη.



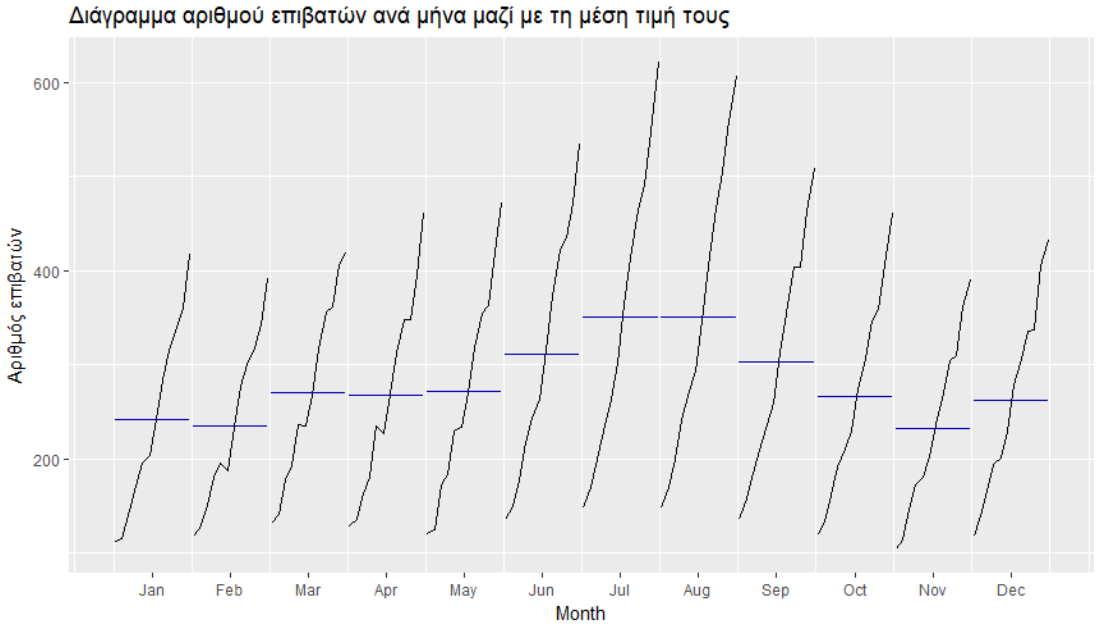
<u>Διάγραμμα 2.3.1</u>: Χρονοσειρά αριθμού διεθνών επιβατών της αεροπορικής εταιρείας Pan Am.

Παρατηρώντας το Διάγραμμα 2.3.1 φαίνεται μια αυξανόμενη τάση στα δεδομένα μας με την πάροδο του χρόνου καθώς και φαινόμενα εποχικότητας αφού το συγκεκριμένο μοτίβο επαναλαμβάνεται μετά από συγκεκριμένα χρονικά διαστήματα. Η περίοδος που εμφανίζει ένα σύνολο δεδομένων, γίνεται εύκολα αντιληπτή με ένα διάγραμμα περιοδικότητας όπως απεικονίζεται στο σχήμα 2.3.2 αλλά και με τη βοήθεια της R μέσω της εντολής frequency. Σύμφωνα με το Διάγραμμα 2.3.2 είναι εμφανής μια σταθερή περίοδος 12 μηνών με μέγιστες τιμές τον μήνα Ιούλιο και ελάχιστες τον μήνα Νοέμβριο. Αυτό δικαιολογείται και λογικά αφού τους μήνες Ιούλιο και Αύγουστο που παρουσιάζεται μεγάλος αριθμός επιβατών για αεροπορικές πτήσεις είναι και η περίοδος θερινών διακοπών. Το διάγραμμα εποχικότητας λαμβάνεται αφού κατεβάσουμε τις βιβλιοθήκες forecast και ggplot2 (Wickham, et al., 2018) στην R, μέσω της εντολής ggseasonplot.



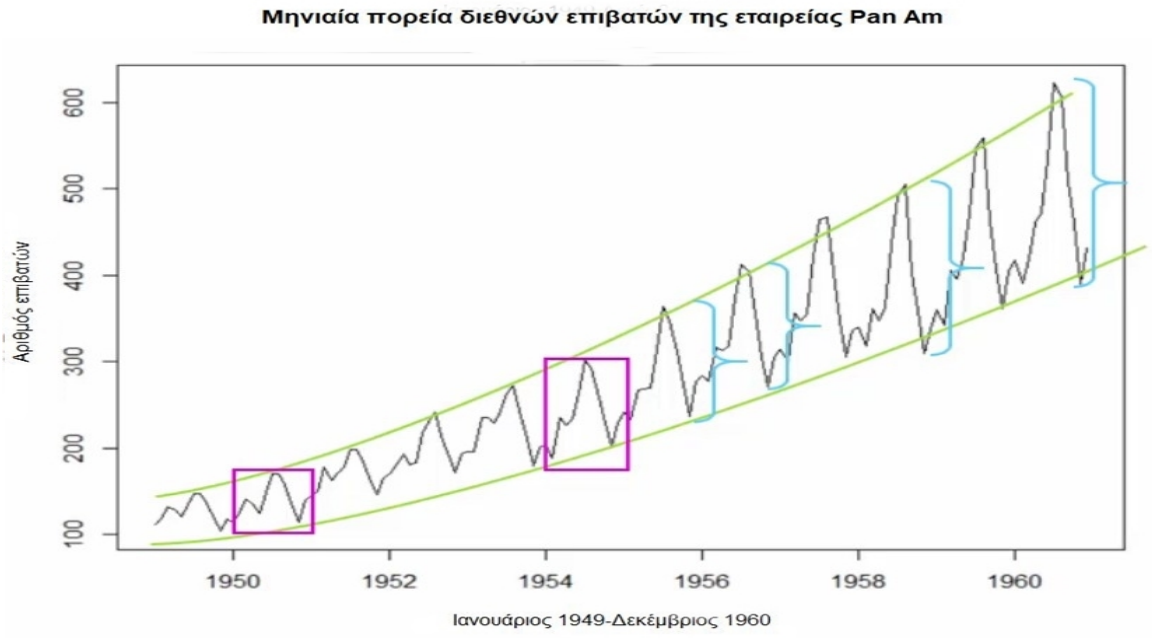
<u>Διάγραμμα 2.3.2</u>: Διάγραμμα εποχικότητας αριθμού διεθνών επιβατών της αεροπορικής εταιρείας Pan Am.

Το Διάγραμμα 2.3.3 δείχνει στον άξονα τον x τους μήνες από τον Ιανουάριο έως τον Δεκέμβριο και στον άξονα τον y το πλήθος των επιβατών της αεροπορικής εταιρείας Pan Am, για κάθε μήνα ξεχωριστά. Δηλαδή στον μήνα Ιανουάριο φαίνονται με την μαύρη γραμμή, η εξέλιξη του πλήθους των αεροεπιβατών συνολικά για τα έτη 1949-1960. Αντίστοιχα στον μήνα Φεβρουάριο κ.ο.κ. Οι οριζόντιες γραμμές μας δείχνουν τον μέσο όρο των αεροεπιβατών για τους εκάστοτε μήνες. Αυτό που παρατηρούμε είναι ότι τους μήνες των θερινών διακοπών Ιούλιο-Αύγουστο έχουμε τους περισσότερους επιβάτες κατά μέσο όρο, ενώ τους μήνες Φεβρουάριο-Νοέμβριο τους λιγότερους. Το Διάγραμμα 2.3.3 λαμβάνεται αφού κατεβάσουμε τις βιβλιοθήκες forecast και ggplot2 στην R, μέσω της εντολής ggmonthplot.



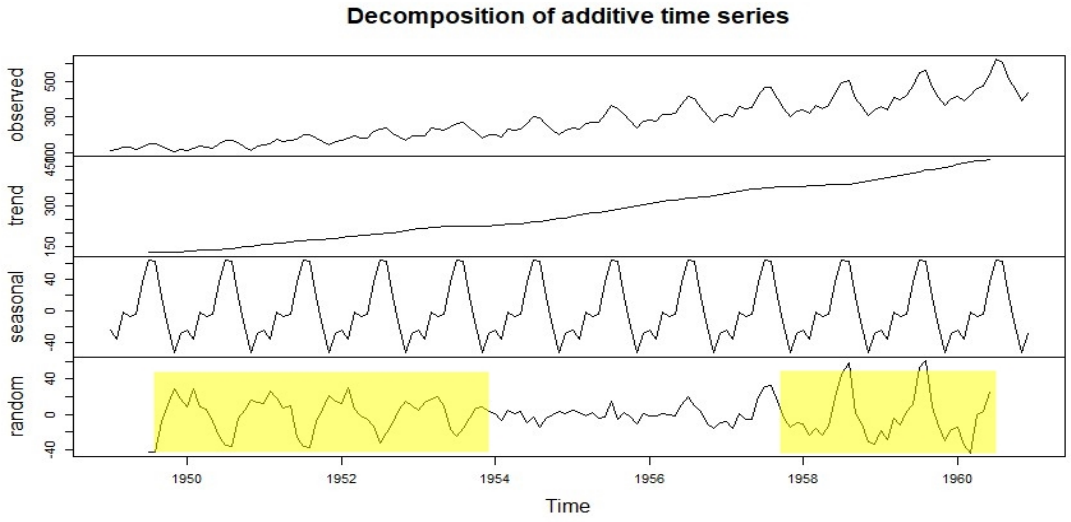
<u>Διάγραμμα 2.3.3</u>: Διάγραμμα αριθμού επιβατών ανά μήνα για τα έτη 1949 έως 1960, μαζί με τις μέσες τιμές τους.

Το γεγονός ότι έχουμε μια αύξηση στο πλάτος των δεδομένων μας με την πάροδο του χρόνου όπως σημειώνεται με τις μπλε αγκύλες του Διαγράμματος 2.3.4, σε συνδυασμό με την εμφανή τάση και το επιλαμβανόμενο μοτίβο που σημειώνεται με στο μωβ πλαίσιο του ίδιου διαγράμματος, υποδηλώνει την προσαρμογή του πολλαπλασιαστικού μοντέλου, ως καταλληλότερο να περιγράψει τα δεδομένα μας. Σε περίπτωση που έχουμε αμφιβολίες για το πιο μοντέλο ενδείκνυται μπορούμε να δοκιμάσουμε και τα δύο μοντέλα και να συγκρίνουμε τα αποτελέσματα



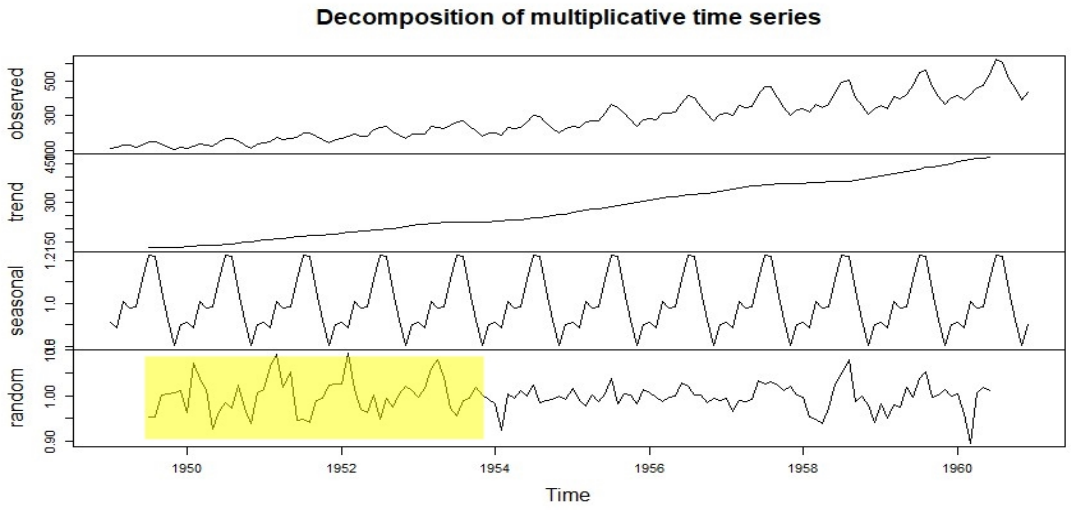
Διάγραμμα 2.3.4

Στην προκειμένη περίπτωση θα δοκιμάσουμε και τα δύο μοντέλα και θα συγκρίνουμε τα αποτελέσματα. Στο Διάγραμμα 2.3.5 έπειτα από την προσαρμογή των δεδομένων μας στη σχέση (2.3.1) παίρνουμε τα γραφήματα των συνιστωσών m_t, s_t και e_t ξεχωριστά, η πρόσθεση των οποίων μας δίνει τα αρχικά μας δεδομένα. Οι τιμές που δίνονται στις συνιστώσες m_t, s_t και e_t στην R, βρίσκονται μέσω της εντολής decompose πληκτρολογώντας την κατάλληλη παράμετρο, additive, στην περίπτωση του προσθετικού μοντέλου και αντίστοιχα στου πολλαπλασιαστικού το multiplicative. Έπειτα με την εντολή plot μπορούμε να σχεδιάσουμε το Διάγραμμα 2.3.5 που απεικονίζεται παρακάτω. Η τάση που παρουσιάζουν τα δεδομένα μας, όπως απεικονίζεται στο δεύτερο παράθυρο του Διαγράμματος 2.3.5 είναι εμφανής και μάλιστα προτείνει μια γραμμική συνάρτηση της μορφής $m_t = c_0 + c_1 t$ για την εκτίμηση της. Ακόμη, στην εποχιακή συνιστώσα που απεικονίζεται στο τρίτο παράθυρο του ίδιου διαγράμματος είναι εμφανή τα μέγιστα και τα ελάχιστα που παρουσιάζονται στους μήνες του Ιουλίου και Νοέμβριο αντίστοιχα. Τέλος στο τρίτο παράθυρο ο εικονιζόμενος όρος του σφάλματος e_t δεν μοιάζει να είναι τελείως τυχαίος αφού όπως φαίνεται στο κίτρινο πλαίσιο, οι μήνες από το 1949-1953 και από το 1958-1960 φαίνεται να εξακολουθούν να παρουσιάζουν ένα συγκεκριμένο μοτίβο. Μόνο η περίοδος από 1954 έως το 1957 μοιάζει να είναι τυχαία στον όρο e_t. Αυτό το γεγονός υποδηλώνει ότι το μοντέλο επιδέχεται περεταίρω βελτίωση αφού όπως είπαμε στην περίπτωση της αποσύνθεσης θέλουμε όλη η πληροφορία να βρίσκεται στις συνιστώσες m_t, s_t και μόνο η τυχαιότητα να είναι στη συνιστώσα e_t.



<u>Διάγραμμα 2.3.5</u>: Κλασική αποσύνθεση στην περίπτωση του προσθετικού μοντέλου της σχέσης (2.3.1)

Στο Διάγραμμα 2.3.6 αντίστοιχα έπειτα από την προσαρμογή των δεδομένων μας στη σχέση (2.3.2) παίρνουμε τα γραφήματα των συνιστωσών m_t, s_t και e_t ξεχωριστά, ο πολλαπλασιασμός των οποίων μας δίνει τα αρχικά μας δεδομένα. Όπως και στο Διάγραμμα 2.3.5 είναι εμφανής η τάση και η περιοδικότητα και παρουσιάζονται με παρόμοιο τρόπο όπως στο προηγούμενο μοντέλο με τη διαφορά να βρίσκεται στον όρο του σφάλματος e_t. Στην περίπτωση του πολλαπλασιαστικού μοντέλου ο όρος e_t φαίνεται να είναι τυχαίος για τα έτη από το 1954 έως το 1960, ωστόσο για τα έτη 1949 έως 1953 φαίνεται να επικρατεί ακόμα ένα μοτίβο. Σε γενικές γραμμές ο όρος e_t μοιάζει να είναι τυχαίος, αλλά το μοτίβο που παρουσιάζεται στα οχτώ πρώτα χρόνια φανερώνει ότι θα μπορούσαμε να βρούμε ένα καλύτερο μοντέλο για την περιγραφή των δεδομένων μας.



<u>Διάγραμμα 2.3.6</u>: Κλασική αποσύνθεση στην περίπτωση του πολλαπλασιαστικού μοντέλου της σχέσης (2.3.2)

2.4 Μοντέλα με μηδενική μέση τιμή

Πριν προχωρήσουμε την ανάλυσή μας, είναι σημαντικό να δούμε μερικά από τα πιο γνωστά παραδείγματα χρονοσειρών που θα μας χρησιμεύσουν σε επόμενες ενότητες και να διερευνήσουμε εάν αυτές προέρχονται από στάσιμες ή μη στάσιμες διαδικασίες σύμφωνα με τον ορισμό της ασθενούς στασιμότητας.

2.4.1 iid θόρυβος (iid noise)

Το πιο απλό μοντέλο χρονοσειράς, είναι αυτό που δεν έχει τάση ή εποχιακή συνιστώσα και στο οποίο οι παρατηρήσεις είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με μηδενική μέση τιμή. Καλούμε μια τέτοια ακολουθία τυχαίων μεταβλητών $X_1, X_2, X_3, X_4, \ldots$ ως iid θόρυβο. Από τον ορισμό μπορούμε να γράψουμε για οποιονδήποτε θετικό ακέραιο *n* και πραγματικούς αριθμούς $x_1, x_2, \ldots, x_n$,

$$P[X_1 \le x_1, \ldots X_n \le x_n] \overset{\text{ιδιοτ.ανεξαρτ.}}{=} P[X_1 \le x_1] \bullet \bullet \bullet P[X_n \le x_n] = F(x_1) \bullet \bullet \bullet F(x_n) \ ,$$

όπου F(•) είναι η αθροιστική συνάρτηση κατανομής για καθεμιά από τις ισόνομα κατανεμημένες τυχαίες μεταβλητές $X_1, X_2, X_3, X_4, \ldots$. Σε αυτό το μοντέλο δεν υπάρχει εξάρτηση μεταξύ των παρατηρήσεων. Πιο συγκεκριμένα, για όλα τα $h \ge 1$ και όλα τα $x, x_1, x_2, \ldots, x_n$,

$$P[X_{n+h} \le x \mid X_1 = x_1, \ldots X_n = x_n] = P[X_{n+h} \le x] .$$

Έτσι, η γνώση των $X_1, X_2, X_3, X_4, \ldots$ δεν έχει καμία σημασία στην πρόβλεψη της συμπεριφοράς της X_{n+h}. Αν η $\{X_t\}$ είναι iid θόρυβος και επιπλέον ισχύει ότι $\mathrm{E}\left(X_t^2\right) = \sigma^2 < \infty$ τότε είναι φανερό ότι ο ορισμός της ασθενούς στασιμότητας ικανοποιείται αφού :

1. $\mathrm{E}\left(X_t\right) = 0$ για όλα τα t .
2. Υπό την υπόθεση της ανεξαρτησίας, η συνάρτηση συνδιακύμανσης είναι:

Για $h \ne 0$: $\gamma_X\left(t+h,t\right) \overset{h\ne 0}{=} \mathrm{Cov}\left(X_{t+h}, X_t\right) \overset{X_{t+h}, X_t \ \text{ανεξ.}}{=} 0$ και

Για $h = 0$:

$$\begin{aligned}\gamma_X\left(t+h,t\right) \overset{h=0}{=} \gamma_X\left(t,t\right) &= \mathrm{Cov}\left(X_t, X_t\right) = \mathrm{E}\left[\left(X_t - \mu_X\left(t\right)\right)\left(X_t - \mu_X\left(t\right)\right)\right] \\ &= \mathrm{E}\left[\left(X_t - \mu_X\left(t\right)\right)^2\right] = \mathrm{Var}\left(X_t\right) = \mathrm{E}\left(X_t^2\right) - E\left(X_t\right)^2 \\ &= \sigma^2 - 0 = \sigma^2\end{aligned}$$

Συνεπώς,

$$\gamma_X(t+h,t)=\begin{cases}\sigma^2, & \text{αν h=0}\\ 0, & \text{αν h}\neq 0\end{cases}$$

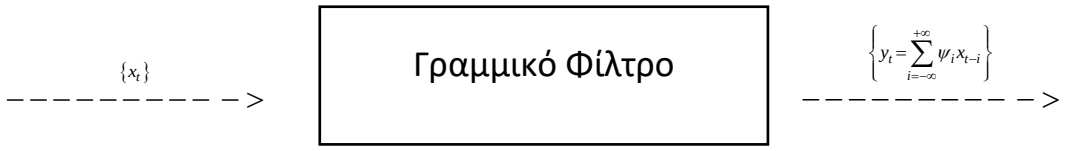
Όπου φαίνεται ότι η συνάρτηση αυτοσυνδιακύμανσης είναι ανεξάρτητη του t.

Έτσι μια iid διαδικασία με πεπερασμένη ροπή 2ης τάξης $\left(\mathrm{E}\left(X_t^2\right)=\sigma^2<\infty\right)$ είναι στάσιμη διαδικασία. Θα συμβολίζουμε την $\{X_t\}$ με μέση τιμή 0 και διασπορά σ^2 ως εξής : $\{X_t\}\sim IID\left(0,\sigma^2\right)$.

2.4.2 Λευκός θόρυβος (White noise)

Εάν $\{X_t\}$ είναι μια ακολουθία από ασυσχέτιστες τυχαίες μεταβλητές κάθε μια από τις οποίες έχει μηδενική μέση τιμή και διασπορά σ^2 τότε η $\{X_t\}$ είναι στάσιμη με την ίδια συνάρτηση συνδιακύμανσης του iid θορύβου. Μια τέτοια ακολουθία τυχαίων μεταβλητών καλείται ακολουθία λευκού θορύβου με μέση τιμή 0 και διασπορά σ^2. Θα συμβολίζουμε την $\{X_t\}$ ως εξής : $\{X_t\}\sim WN\left(0,\sigma^2\right)$. Ο λευκός θόρυβος είναι μια iid διαδικασία αν επιπλέον οι παρατηρήσεις είναι και ανεξάρτητες. Συνεπώς κάθε $IID\left(0,\sigma^2\right)$ ακολουθία είναι $WN\left(0,\sigma^2\right)$ αλλά όχι αντιστρόφως.

2.5 Γραμμικό φίλτρο



Στην στατιστική μοντελοποίηση συχνά ασχολούμαστε με τον εντοπισμό της πραγματικής σχέσης μεταξύ δύο χρονοσειρών $\{X_t\}$ και $\{Y_t\}$. Μια σημαντική υπόθεση που μας παρέχει ευκολία στη μοντελοποίηση είναι αυτή της γραμμικότητας (Montgomery, et al., 2016). Ένα γραμμικό φίλτρο για παράδειγμα, είναι μια γραμμική «λειτουργία» από μια σειρά $\{X_t\}$ σε μια άλλη $\{Y_t\}$ τέτοια ώστε:

$$Y_t = L(X_t) = \sum_{i=-\infty}^{+\infty} \psi_i X_{t-i} \quad \text{με t} = \cdots,-2,-1,0,1,2,\cdots \qquad (2.5.1)$$

Με βάση την σχέση (2.5.1) ένα γραμμικό φίλτρο είναι μια διαδικασία που μετατρέπει την αρχική χρονοσειρά $\{X_t\}$ σε μια άλλη $\{Y_t\}$ η οποία σχετίζεται γραμμικά με την πρώτη και περιλαμβάνει όλες τις παρελθοντικές, παροντικές και μελλοντικές τιμές της X_t σε μια μορφή αθροίσματος με διαφορετικά βάρη $\{\psi_t\}$ για κάθε παρατήρηση X_t. Για το γραμμικό φίλτρο της εξίσωσης (2.5.1) υποθέτουμε τις εξής ιδιότητες:

- Οι συντελεστές ψ_i είναι σταθερές και ανεξάρτητες του χρόνου t.
- Εάν $\psi_i = 0$ για $i < 0$, τότε η Y_t είναι μια γραμμική συνάρτηση των τωρινών και παρελθοντικών τιμών της X_t: $Y_t = \psi_0 X_t + \psi_1 X_{t-1} + \ldots$
- Το φίλτρο λέγεται σταθερό (stable) και η χρονοσειρά $\{Y_t\}$ στάσιμη εάν $\sum_{i=-\infty}^{+\infty} |\psi_i| < +\infty$.

Μπορούμε να θεωρήσουμε ότι για μια χρονοσειρά $\{Y_t\}$ στην οποία παρουσιάζεται μεγάλη εξάρτηση μεταξύ των διαδοχικών τιμών της, παράγεται από το άθροισμα τυχαίων μεταβλητών e_t μιας διαδικασίας λευκού θορύβου με μηδενική μέση τιμή και διασπορά σ_α^2 $\left(\{e_t\} \sim WN\left(0, \sigma_\alpha^2\right)\right)$.

Ένα γραμμικό φίλτρο $L(\cdot)$ που επιδρά πάνω στην διαδικασία λευκού θορύβου $\{e_t\}$, t>0 δίνει μια νέα χρονοσειρά της μορφής:

$$Y_t = L(e_t) = \mu + \sum_{i=0}^{\infty} \psi_i e_{t-i} = \mu + \psi_0 e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \cdots \qquad (2.5.2)$$

Η διαδικασία λευκού θορύβου είναι στάσιμη και το μ εκφράζει την μέση τιμή γύρω από την οποία οι τιμές της Y_t πάλλονται. Στην περίπτωση αυτή όπου η αρχική χρονοσειρά $\{e_t\}$, t>0 είναι στάσιμη και το φίλτρο σταθερό, δηλαδή $\sum_{i=0}^{\infty}|\psi_i| < +\infty$, τότε μετά την εφαρμογή του φίλτρου παίρνουμε επίσης μια στάσιμη διαδικασία $\{Y_t\}$. Χρησιμοποιώντας τον τελεστή οπίσθιας μετάθεσης B η σχέση (2.5.2) γίνεται :

$$Y_t = \mu + \psi_0 e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \cdots = \mu + \psi_0 B^0 e_t + \psi_1 B^1 e_t + \psi_2 B^2 e_t + \cdots$$

$$= \mu + \sum_{i=0}^{\infty} \psi_i B^i e_t = \mu + \underbrace{\left(\sum_{i=0}^{\infty} \psi_i B^i \right)}_{\Psi(B)} e_t = \mu + \Psi(B) e_t$$

$$\Rightarrow Y_t = \mu + \Psi(B) e_t \qquad (2.5.3)$$

Μια χρονοσειρά $\{Y_t\}$ που ικανοποιεί την σχέση (2.5.3) καλείται άπειρης τάξης κινητός μέσος όρος και μας δίνει μια γενική κλάση μοντέλων για οποιαδήποτε στάσιμη σειρά.

2.6 Μοντέλο κινητού μέσου όρου πεπερασμένης τάξης q MA(q)

Στην μοντελοποίηση στάσιμων χρονοσειρών όπως η $\{Y_t\}$ που ικανοποιεί την εξίσωση (2.5.3), είναι φανερό ότι δεν είναι πρακτικό να προσπαθήσουμε να εκτιμήσουμε τις άπειρες τιμές των βαρών $\{\psi_t\}$ προκειμένου να καθοριστεί το μοντέλο. Συνεπώς είναι αναγκαίο να ορίσουμε τον πεπερασμένης τάξης κινητό μέσο όρο που χρησιμοποιεί πεπερασμένο αριθμό βαρών $\{\psi_t\}$. Η εξίσωση (2.6.1) ορίζει ένα μοντέλο κινητού μέσου όρου τάξης q :

$$X_t = \mu + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \qquad (2.6.1)$$

Θεωρούμε ότι $\{e_t\} \sim WN\left(0, \sigma_e^2\right)$, δηλαδή είναι διαδικασία λευκού θορύβου μέσης τιμής 0 και διασποράς σ_e^2. Εδώ να αναφέρουμε ότι η τάξη q του μοντέλου λέγεται και υστέρηση (lag). Αφού ορίσαμε την διαδικασία κινητού μέσου όρου τάξης q, το επόμενη βήμα είναι να ελέγξουμε αν πρόκειται για μια στάσιμη διαδικασία.

Υπολογίζουμε την αναμενόμενη τιμή της X σε χρόνο t :

$$E[X_t] = E\left[\mu + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_p e_{t-p}\right] = E[\mu] + E[e_t] + E[\theta_1 e_{t-1}] + E[\theta_2 e_{t-2}] + \ldots + E[\theta_p e_{t-p}]$$

$$\underset{\theta_i \in \mathbb{R}}{\overset{\mu \in \mathbb{R}}{=}} \mu + E[e_t] + \theta_1 E[e_{t-1}] + \theta_2 E[e_{t-2}] + \ldots + \theta_p E[e_{t-p}] \underset{E[e_t] = \ldots = E[e_{t-p}] = 0}{\overset{\{e_t\} \sim WN(0, \sigma_e^2)}{=}} \mu$$

Όπως φαίνεται $E[X_t] = \mu = \sigma\tau\alpha\theta.$ ανεξάρτητη του t.

Συνεχίζουμε υπολογίζοντας την ροπή 2ης τάξης της $\{X_t\}$:

$$\begin{aligned}
\mathrm{E}\left(X_t^2\right) &= Var\left(X_t\right)+E\left(X_t\right)^2 \\
&= Var\left(\mu+e_t+\theta_1 e_{t-1}+\theta_2 e_{t-2}+\cdots+\theta_q e_{t-q}\right)+\mu^2 \\
&\underset{\substack{\chi\rho\eta\sigma\iota\mu o\pi o\iota\acute{\omega}\\ \text{ιδιοτ. διασποράς}}}{\overset{Var(\mu+e_t)=Var(e_t)}{=}} Var\left(e_t\right)+\theta_1^2 Var\left(e_{t-1}\right)+\theta_2^2 Var\left(e_{t-2}\right)+\ldots+\theta_q^2 Var\left(e_{t-q}\right)+\mu^2 \\
&\overset{\{e_t\}\sim WN(0,\sigma_e^2)}{=} \sigma_e^2+\theta_1^2\sigma_e^2+\theta_2^2\sigma_e^2+\ldots+\theta_q^2\sigma_e^2+\mu^2 \\
&= \sigma_e^2\left(1+\theta_1^2+\theta_2^2+\ldots+\theta_q^2\right)+\mu^2 \ <+\infty
\end{aligned}$$

Άρα, $\mathrm{E}\left(X_t^2\right) = Var\left(X_t\right)+\mu^2 = \sigma_e^2\left(1+\theta_1^2+\theta_2^2+\ldots+\theta_q^2\right)+\mu^2 \ <+\infty$ (2.6.2)

Όπως βλέπουμε η ροπή 2ης τάξης $\mathrm{E}\left(X_t^2\right)$ είναι πεπερασμένη, ανεξάρτητη του χρόνου t και η διασπορά της X_t είναι $Var\left(X_t\right)= \sigma_e^2\left(1+\theta_1^2+\theta_2^2+\ldots+\theta_q^2\right)$. Αν επίσης υπολογίσουμε και τη συνάρτηση αυτοσυνδιακύμανσης και βγει και αυτή ανεξάρτητη του t, τότε η διαδικασία μας θα είναι στάσιμη. Για τον υπολογισμό της συνάρτησης αυτοσυνδιακύμανσης, για ευκολία πράξεων, θα αποδείξουμε μόνο την ειδική περίπτωση της 1ης τάξης MA(1) διαδικασίας και έπειτα θα γενικεύσουμε για την τάξη q.

Ας υποθέσουμε λοιπόν ότι η χρονοσειρά $\{X_t\}$ ορίζεται από την εξίσωση (2.6.1) αντικαθιστώντας το q=1. Δηλαδή έχουμε το 1ης τάξης μοντέλο κινητού μέσου όρου MA(1) το οποίο ορίζεται ως:

$$X_t = \mu+e_t+\theta_1 e_{t-1}\ ,\ \ \mathrm{t}=1,2,\ldots,$$

Για να ελέγξουμε την συνάρτηση αυτοσυνδιακύμανσης διακρίνουμε τις εξής τέσσερεις περιπτώσεις :

$$\Gamma\iota\alpha\ \mathrm{h}=0:\ \gamma_X\left(t,t-h\right)\overset{h=0}{=}\gamma_X\left(t,t\right)=Cov\left(X_t,X_t\right)=Var\left(X_t\right)\underset{\gamma\iota\alpha\ \mathrm{q}=1}{\overset{\alpha\pi\acute{o}\ \tau\eta\nu\ (2.6.2)}{=}}\sigma_e^2\left(1+\theta_1^2\right)$$

$$\begin{aligned}
\Gamma\iota\alpha\ \mathrm{h}=-1:\ \gamma_X\left(t,t-h\right)\overset{h=-1}{=}\gamma_X\left(t,t-1\right)&=Cov\left(X_t,X_{t-1}\right)=Cov\left(\mu+e_t+\theta_1 e_{t-1},\mu+e_{t-1}+\theta_1 e_{t-2}\right) \\
&=Cov\left(\mu,\mu\right)+Cov\left(\mu,e_{t-1}\right)+\theta_1 Cov\left(\mu,e_{t-2}\right)+Cov\left(e_t,\mu\right)+Cov\left(e_t,e_{t-1}\right)+ \\
&\quad+\theta_1 Cov\left(e_t,e_{t-2}\right)+\theta_1 Cov\left(e_{t-1},\mu\right)+\theta_1 Cov\left(e_{t-1},e_{t-1}\right)+\theta_1^2 Cov\left(e_{t-1},e_{t-2}\right) \\
&\overset{e_t\ \alpha\sigma\upsilon\sigma\chi.}{=}\theta_1 Cov\left(e_{t-1},e_{t-1}\right)=\theta_1 Var\left(e_{t-1}\right)=\theta_1\sigma_e^2
\end{aligned}$$

$$\Gamma\iota\alpha\ \text{h= } 1: \gamma_X(t,t-h) \overset{h=-1}{=} \gamma_X(t,t+1) = Cov(X_t, X_{t+1}) = Cov(\mu + e_t + \theta_1 e_{t-1}, \mu + e_{t+1} + \theta_1 e_t)$$
$$= Cov(\mu,\mu) + Cov(\mu, e_{t+1}) + \theta_1 Cov(\mu, e_t) + Cov(e_t, \mu) + Cov(e_t, e_{t+1}) +$$
$$+ \theta_1 Cov(e_t, e_t) + \theta_1 Cov(e_{t-1}, \mu) + \theta_1 Cov(e_{t-1}, e_{t+1}) + \theta_1^2 Cov(e_{t-1}, e_t)$$
$$\overset{e_t\ \alpha\sigma\upsilon\sigma\chi.}{=} \theta_1 Cov(e_t, e_t) = \theta_1 Var(e_t) = \theta_1 \sigma_e^2$$

$$\Gamma\iota\alpha\ |\text{h}|>1: \gamma_X(t,t-h) \overset{|h|>1}{=} Cov(X_t, X_{t-h}) = Cov(\mu + e_t + \theta_1 e_{t-1}, \mu + e_{t-h} + \theta_1 e_{t-h-1})$$
$$= Cov(\mu,\mu) + Cov(\mu, e_{t-h}) + \theta_1 Cov(\mu, e_{t-h-1}) + Cov(e_t, \mu) + Cov(e_t, e_{t-h}) +$$
$$+ \theta_1 Cov(e_t, e_{t-h-1}) + \theta_1 Cov(e_{t-1}, \mu) + \theta_1 Cov(e_{t-1}, e_{t-h}) + \theta_1^2 Cov(e_{t-1}, e_{t-h-1})$$
$$\underset{|h|>1}{\overset{e_t\ \alpha\sigma\upsilon\sigma\chi.}{=}} 0+0+\cdots+0 = 0$$

Συνεπώς,

$$\gamma_X(t,t-h) = \begin{cases} \sigma_e^2(1+\theta_1^2) & \alpha\nu\ \text{h=}0 \\ \sigma_e^2\theta_1 & \alpha\nu\ \text{h=}\pm 1 \\ 0 & \alpha\nu\ |\text{h}|>1 \end{cases}$$

Όπως βλέπουμε η συνάρτηση αυτοσυνδιακύμανσης είναι ανεξάρτητη από το t και επομένως ικανοποιείται ο ορισμός της ασθενούς στασιμότητας και η διαδικασία MA(1) είναι στάσιμη. Με αντίστοιχο τρόπο στην γενική περίπτωση όπου η τάξη του μοντέλου MA είναι q, η συνάρτηση αυτοσυνδιακύμανσης βγαίνει :

$$\gamma_X(t,t-h) = \begin{cases} \sigma_e^2 \sum_{i=0}^{q-|h|} \theta_i \theta_{i+|h|} & \alpha\nu\ |\text{h}| \le q \quad \text{όπου } \theta_0 = 1 \\ 0 & \alpha\nu\ |\text{h}|>\text{q} \end{cases}$$

Δηλαδή ανεξάρτητη του χρόνου t .

Με βάση τα παραπάνω κάθε διαδικασία που ικανοποιεί την εξίσωση (2.6.1) είναι στάσιμη και το μοντέλου του κινητού μέσου όρου τάξης q είναι κατάλληλο για να περιγράψει μόνο στάσιμες διαδικασίες. Στο αυτό το μοντέλο έχουμε $q+2$ αγνώστους, τους συντελεστές ή αλλιώς βάρη $\theta_1, \theta_2, ..., \theta_q$, τον μέσο μ και τη διασπορά σ_e^2 του λευκού θορύβου. Συνεπώς για να καθοριστεί το μοντέλο πλήρως και να μπορεί να εφαρμοστεί, αρκεί να εκτιμήσουμε αυτές τις άγνωστες παραμέτρους με βάση τα δεδομένα μας. Άρα, για ένα σύνολο δεδομένων $\{x_1, x_{2,}, ..., x_n\}$, όπου $n > q$, δηλαδή το πλήθος των παρατηρήσεων μας να είναι μεγαλύτερο από την υστέρηση q που κοιτάμε πίσω, έχοντας εκτιμήσει τις άγνωστες

αυτές παραμέτρους του μοντέλου MA(q) μπορούμε να προβλέψουμε την τιμή της μεταβλητής που μελετάμε την επόμενη χρονική στιγμή $n+1$, δηλαδή την X_{n+1}. Αντίστοιχα μπορούμε να προβλέψουμε και τις επόμενες χρονικές στιγμές $n+2, n+3, \ldots, n+q$ της μεταβλητής που μελετάμε. Βέβαια είναι φυσικό ότι όσο μακρύτερα στο μέλλον κάνουμε μια πρόβλεψη τόσο λιγότερο ακριβής γινόμαστε.

Αν θεωρήσουμε ότι η Y_t είναι μια καινούρια χρονοσειρά, τέτοια ώστε $Y_t = X_t - \mu$, τότε η εξίσωση της μορφής (2.6.1) μπορεί εύκολα να γραφτεί σε μια ισοδύναμη της μορφής :

$$Y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} = \sum_{i=0}^{q} \theta_i e_{t-i} \ , \text{ όπου } \theta_0 = 1 \quad (2.6.2)$$

Ουσιαστικά η Y_t εκφράζει πόσο απέχει η αρχική μας χρονοσειρά X_t από την σταθερή μέση τιμή της, δηλαδή οι τιμές της Y_t μας δείχνουν τις αποκλίσεις από τον μέσο όρο.

Εφαρμόζοντας τον τελεστή οπίσθιας μετάθεσης B η σχέση (2.6.2) γίνεται :

$$\begin{aligned} Y_t &= e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} = \mathrm{B}^0 e_t + \theta_1 B^1 e_t + \theta_2 B^2 e_t + \cdots + \theta_q B^q e_t \\ &= \left(1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right) e_t = \underbrace{\left(\sum_{i=0}^{q} \theta_i B^i\right)}_{\theta(\mathrm{B})} e_t = \theta(\mathrm{B}) e_t \end{aligned}$$

Άρα καταλήγουμε στην σχέση :

$$Y_t = \theta(\mathrm{B}) e_t \ ,$$

όπου $\theta(B)$ ορίζουμε να είναι το MA πολυώνυμο τάξης q της μορφής:

$$\theta(\mathrm{B}) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$$

2.7 Αυτοπαλινδρομικό μοντέλο AR(p)

Ένα πολύ βασικό μοντέλο το οποίο θα μελετηθεί είναι το Αυτοπαλινδρομικό μοντέλο AR (Autoregressive Model). Ας υποθέσουμε μια χρονοσειρά $\{X_t\}$ η οποία περιγράφεται από την εξίσωση :

$$X_t = \mu + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + e_t, \quad (2.7.1)$$

όπου μ εκφράζει την σταθερή μέση τιμή της χρονοσειράς $\{X_t\}$ και φ_i τον συντελεστή που δείχνει την επιρροή της X_{t-i} μεταβλητής του παρελθόντος στην τρέχουσα X_t. Ο όρος e_t είναι ο όρος του σφάλματος ο οποίος και σε αυτή την περίπτωση υποθέτουμε ότι αποτελείται από ασυσχέτιστες και ισόνομες τυχαίες μεταβλητές με μηδενική μέση τιμή και σταθερή διασπορά σ_e^2, δηλαδή είναι μια ακολουθία λευκού θορύβου $\{e_t\} \sim WN(0, \sigma_e^2)$. Ένα μοντέλο που ικανοποιεί την εξίσωση (2.7.1) λέγεται αυτοπαλινδρομικό μοντέλο τάξης p $(AR(p))$ και όπως φαίνεται και από την ονομασία του συσχετίζει με γραμμικό τρόπο την τρέχουσα τιμή της μεταβλητής, X_t, με τις p παρελθοντικές τιμές της ίδια μεταβλητής. Αντίστοιχα και εδώ η τάξη p του μοντέλου λέγεται και υστέρηση (lag). Με βάση την εξίσωση (2.7.1) μπορούμε να βρούμε την αναμενόμενη τιμή της μεταβλητής X την χρονική στιγμή *t*, δοθέντος *p* αρχικές τιμές της ίδιας μεταβλητής. Πιο συγκεκριμένα, έχουμε:

$$E[X_t] = E\left[\mu + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + e_t\right]$$

$$\underset{\substack{\text{της συνάρτησης E}}}{\overset{\text{λόγω γραμμικοτητας}}{=}} E[\mu] + E[\varphi_1 X_{t-1}] + E[\varphi_2 X_{t-2}] + \cdots + E[\varphi_p X_{t-p}] + E[e_t]$$

$$\underset{E(e_t)=0}{\overset{\mu\in\mathbb{R},\varphi_i\in\mathbb{R}}{=}} \mu + \varphi_1 E(X_{t-1}) + \varphi_2 E(X_{t-2}) + \cdots + \varphi_p E(X_{t-p})$$

Άρα,

$$E[X_t] = \mu + \varphi_1 E(X_{t-1}) + \varphi_2 E(X_{t-2}) + \cdots + \varphi_p E(X_{t-p})$$

Αντίστοιχα και εδώ μπορούμε να θεωρήσουμε ότι η Y_t είναι μια καινούρια χρονοσειρά, τέτοια ώστε $Y_t = X_t - \mu$, τότε η εξίσωση της μορφής (2.7.1) μπορεί εύκολα να γραφτεί σε μια ισοδύναμη της μορφής :

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + e_t = \sum_{i=1}^{p} \varphi_i Y_{t-i} + e_t, \quad (2.7.2)$$

όπου η Y_t εκφράζει πόσο απέχει η αρχική μας χρονοσειρά X_t από την σταθερή μέση τιμή της. Φυσικά, στην ίδια μορφή της εξίσωσης (2.7.2) καταλήγουμε και στην περίπτωση όπου η μέση τιμή της αρχικής χρονοσειράς X_t είναι μηδέν.

Συνεπώς με τη βοήθεια του τελεστή οπίσθιας μετάθεσης B η σχέση (2.7.1) γίνεται :

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + e_t = \varphi_1 B^1 Y_t + \varphi_2 B^2 Y_t + \cdots + \varphi_p B^p Y_t + e_t = \sum_{i=1}^{p} \varphi_i B^i Y_t \; + \; e_t$$

$$\Rightarrow Y_t - \sum_{i=1}^{p} \varphi_i B^i Y_t = e_t \;\Rightarrow\; \underbrace{\left(1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p\right)}_{\varphi(B)} Y_t = e_t$$

Άρα καταλήγουμε στην σχέση :

$$\varphi\left(\mathrm{B}\right) Y_t = e_t \,,$$

όπου $\varphi\left(B\right)$ ορίζουμε να είναι το AR πολυώνυμο τάξης p της μορφής:

$$\varphi\left(\mathrm{B}\right) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p \quad (2.7.3)$$

Στο αυτό το μοντέλο έχουμε $p+2$ αγνώστους, τους συντελεστές $\varphi_1, \varphi_2, ..., \varphi_p$, τον μέσο μ και τη διασπορά σ_e^2 του λευκού θορύβου. Συνεπώς για να καθοριστεί το μοντέλο πλήρως και να μπορεί να εφαρμοστεί, αρκεί να εκτιμήσουμε αυτές τις άγνωστες παραμέτρους με βάση τα δεδομένα μας.

Μεγάλο ενδιαφέρον παρουσιάζει η πρόβλεψη της παρατήρησης κατά την χρονική περίοδο *t*, X_t, που προέκυψε από την αμέσως προηγούμενη *t-1*, δηλαδή σε υστέρηση 1 (lag=1). Ένα τέτοιο παράδειγμα είναι αυτό της 1ης τάξης αυτοπαλινδρομικής $AR\left(1\right)$ διαδικασίας που περιγράφεται από την εξίσωση (2.7.2) αν αντικαταστήσουμε το $p=1$. Τότε θα πάρουμε μια εξίσωση της μορφής :

$$Y_t = \varphi_1 Y_{t-1} + e_t \;,\;\; \text{t=1,2,...,} \quad \left\{e_t\right\} \sim WN(0, \sigma_e^2) \qquad (2.7.4)$$

Αποδεικνύεται ότι η $AR(p)$ διαδικασία είναι μια ειδική περίπτωση του γραμμικού φίλτρου και για ευκολία πράξεων θα το αποδείξουμε μόνο για την $AR\left(1\right)$ διαδικασία. Για να το πετύχουμε αυτό θα αντικαταστήσουμε στην εξίσωση (2.7.4) την σχέση $Y_{t-1} = \varphi_1 Y_{t-2} + e_{t-1}$ και ούτω καθεξής μέχρι s τέτοιες αντικαταστάσεις. Δηλαδή,

$$\begin{aligned}
Y_t &= \varphi_1 Y_{t-1} + e_t \\
&= \varphi_1 \overbrace{\left(\varphi_1 Y_{t-2} + e_{t-1}\right)}^{Y_{t-1}} + e_t \\
&= \varphi_1^2 \overbrace{\left(\varphi_1 Y_{t-3} + e_{t-2}\right)}^{Y_{t-2}} + \varphi_1 e_{t-1} + e_t \\
&\quad\vdots \\
&= \varphi_1^{s+1} Y_{t-(s+1)} + \varphi_1^{s} e_{t-s} + \varphi_1^{s-1} e_{t-(s-1)} + \ldots + \varphi_1^{2} e_{t-2} + \varphi_1 e_{t-1} + e_t
\end{aligned}$$

Συνεπώς στην περίπτωση του $AR(1)$ μοντέλου καταλήγουμε στη σχέση :

$$Y_t = \varphi_1^{s+1} Y_{t-(s+1)} + e_t + \varphi_1 e_{t-1} + \varphi_1^2 e_{t-2} + \ldots + \varphi_1^s e_{t-s} \qquad (2.7.6)$$

Τώρα για $s \to +\infty$ η (2.7.6) γίνεται :

$$Y_t \overset{s\to+\infty}{=} e_t + \varphi_1 e_{t-1} + \varphi_1^2 e_{t-2} + \ldots + \varphi_1^s e_{t-s} = \sum_{i=0}^{\infty} \varphi_1^i e_{t-i} \quad \Rightarrow \quad Y_t = \sum_{i=0}^{\infty} \varphi_1^i e_{t-i} \ .$$

Δηλαδή ένα μοντέλο γραμμικού φίλτρου με βάρη $\psi_i = \varphi_1^i$ και η $\sum_{i=0}^{\infty} \varphi_1^i e_{t-i}$ συγκλίνει αν και μόνο αν $|\varphi_1| < 1$. Δηλαδή, το φίλτρο είναι σταθερό και η χρονοσειρά $\{Y_t\}$ στάσιμη αν και μόνο αν $|\varphi_1| < 1$.

Με αντίστοιχες διαδικασίες αποδεικνύεται ότι και το $AR(p)$ μοντέλο αποτελεί μια ειδική περίπτωση του γραμμικού φίλτρου εάν αντιστρέψουμε τον γραμμικό τελεστή $\varphi(\mathrm{B})$ της σχέσης $\varphi(\mathrm{B}) Y_t = e_t$. Δηλαδή, θα πάρουμε τη σχέση:

$$Y_t = \varphi^{-1}(B) e_t = \psi(B) e_t, \text{ όπου } \psi(B) = \sum_{i=0}^{+\infty} \psi_i B^i$$

Συνοψίζοντας, μια Αυτοπαλινδρομική διαδικασία όπως είδαμε είναι μια περίπτωση γραμμικού φίλτρου και συνεπώς μπορεί να είναι είτε στάσιμη διαδικασία, είτε μη στάσιμη. Όπως είδαμε και στην ενότητα 2.5 απαραίτητη προϋπόθεση για να έχουμε στάσιμη διαδικασία είναι $\sum_{i=-\infty}^{+\infty} |\psi_i| < +\infty$. Αυτό σημαίνει ότι οι συντελεστές φ_t του AR(p) πρέπει να είναι τέτοιοι ώστε τα βάρη ψ_i να δημιουργούν μια συγκλίνουσα σειρά. Αυτό συμβαίνει όταν οι ρίζες του AR πολυωνύμου τάξης p της σχέσης (2.7.3) είναι κατά απόλυτη τιμή μεγαλύτερες από την μονάδα.

Για παράδειγμα , στο $AR(1)$ μοντέλο το πολυώνυμο τάξης 1 είναι : $\varphi(B) = 1 - \varphi_1 B$ με ρίζα $B \overset{\varphi_1 \neq 0}{=} \frac{1}{\varphi_1}$ και συνεπώς για να είναι η ρίζα μεγαλύτερη από την μονάδα κατά απόλυτη τιμή, πρέπει $|\varphi_1| < 1$ γεγονός που μας επιβεβαιώνει τα παραπάνω αποτελέσματα .

Συνεχίζουμε υπολογίζοντας την αναμενόμενη τιμή μιας στάσιμης διαδικασίας $\{Y_t\}$ που ικανοποιεί την εξίσωση του $AR(1)$ μοντέλου:

$$Y_t = \varphi_1 Y_{t-1} + e_t \ , \text{ όπου } \{e_t\} \sim WN(0, \sigma_e^2) \text{ και } |\varphi_1| < 1 \ .$$

Επίσης τα e_t θεωρούμε ότι είναι ασυσχέτιστα με τα Y_s για κάθε $s < t$. Χρησιμοποιώντας ότι $\mathrm{E}(e_t) = 0$ έχουμε :

$$\mathrm{E}(Y_t)=\mathrm{E}(\varphi_1 Y_{t-1}+e_t)=\varphi_1\mathrm{E}(Y_{t-1})+\overbrace{\mathrm{E}(e_t)}^{=0}=\varphi_1\mathrm{E}(\varphi Y_{t-2}+e_{t-1})$$

$$=\varphi_1^2\mathrm{E}(Y_{t-2})+\overbrace{\mathrm{E}(e_{t-1})}^{=0}$$

$$=\varphi_1^2\mathrm{E}(\varphi_1 Y_{t-3}+e_{t-2})$$

$$=\varphi_1^3\mathrm{E}(Y_{t-3})+\overbrace{\mathrm{E}(e_{t-2})}^{=0}$$

$$\vdots$$

$$=\varphi_1^t\mathrm{E}(Y_0)+\overbrace{\mathrm{E}(e_1)}^{=0}$$

$$=\varphi_1^t\mathrm{E}(Y_0)\overset{t\to+\infty}{\to}0 \quad \text{αφού } |\varphi_1|<1$$

Άρα, $\mathrm{E}(Y_t)=0$.

Για να βρούμε την συνάρτηση αυτοσυνδιακύμανσης της $\{Y_t\}$ πολλαπλασιάζουμε κάθε πλευρά της εξίσωσης $Y_t=\varphi_1 Y_{t-1}+e_t$ με το X_{t-h}, όπου h>0 και έπειτα παίρνουμε σε κάθε μέλος τις μέσες τιμές. Έχουμε,

$$Y_t\cdot Y_{t-h}=(\varphi_1 Y_{t-1}+e_t)\cdot Y_{t-h}=(\varphi_1 Y_{t-1}\cdot Y_{t-h}+e_t\cdot Y_{t-h})$$

$$\Rightarrow \mathrm{E}(Y_t\cdot Y_{t-h})=\mathrm{E}(\varphi_1 Y_{t-1}\cdot Y_{t-h}+e_t\cdot Y_{t-h})$$

$$=\mathrm{E}(\varphi_1 Y_{t-1}\cdot Y_{t-h})+\mathrm{E}(e_t\cdot Y_{t-h})$$

$$\underset{\substack{\alpha\varphi o\acute{\upsilon}\ \mathrm{h}>0\\ \mathrm{t\text{-}h}<\mathrm{t}}}{\overset{e_t,X_{t-h}\ \alpha\sigma\upsilon\sigma\chi.}{=}}\varphi_1\mathrm{E}(Y_{t-1}\cdot Y_{t-h})+\mathrm{E}(e_t\cdot Y_{t-h})$$

Από τις ιδιότητες της συνάρτησης συνδιακύμανσης γνωρίζουμε ότι : $Cov(Y_t,Y_{t-h})=E(Y_t\cdot Y_{t-h})-E(Y_t)E(Y_{t-h})$ και συνεπώς αντικαθιστώντας στην παραπάνω σχέση το $E(Y_t\cdot Y_{t+h})$ με το $Cov(Y_t,Y_{t-h})+E(Y_t)E(Y_{t-h})$ έχουμε:

$$\Rightarrow Cov\left(Y_t,Y_{t-h}\right)+\overbrace{E\left(Y_t\right)E\left(Y_{t-h}\right)}^{=0}=\varphi_1\left[Cov\left(Y_{t-1},Y_{t-h}\right)+\overbrace{\mathrm{E}\left(Y_{t-1}\right)\cdot\mathrm{E}\left(Y_{t-h}\right)}^{=0}\right]+Cov\left(e_t,Y_{t-h}\right)+\overbrace{E\left(e_t\right)E\left(Y_{t-h}\right)}^{=0}$$

$$\Rightarrow Cov\left(Y_t,Y_{t-h}\right)=\varphi_1 Cov\left(Y_{t-1},Y_{t-h}\right)+Cov\left(e_t,Y_{t-h}\right)$$

$$\underset{\substack{\alpha\varphi o\acute{\upsilon}\ \mathrm{h>0}\\ \mathrm{t-h<t}}}{\overset{e_t,X_{t-h}\ \alpha\sigma\upsilon\sigma\chi.}{=}}\varphi_1 Cov\left(Y_{t-1},Y_{t-h}\right)+0$$

$$=\varphi_1 Cov\left(\varphi_1 Y_{t-2},Y_{t-h}\right)+\overbrace{Cov\left(e_{t-1},Y_{t-h}\right)}^{=0}$$

$$=\varphi_1^2 Cov\left(Y_{t-2},Y_{t-h}\right)$$

$$\vdots$$

$$=\varphi_1^h Cov\left(Y_{t-h},Y_{t-h}\right)$$

$$=\varphi_1^h Cov\left(Y_t,Y_t\right)\ \text{αφού}\ \{Y_t\}\ \text{στάσιμη διαδικασία άρα}\ Cov\left(Y_{t-h},Y_{t-h}\right)=Cov\left(Y_t,Y_t\right)$$

$$=\varphi_1^h\gamma_Y\left(0\right)$$

$$\Rightarrow\ Cov\left(Y_t,Y_{t-h}\right)=\varphi_1^h\gamma_Y\left(0\right)\qquad(2.7.8)$$

Μένει να υπολογίσουμε την διασπορά της διαδικασίας $\gamma_Y\left(0\right)=Var(Y_t)$.

Από το γεγονός ότι η συνάρτηση αυτοσυνδιακύμανσης είναι μια γραμμική συνάρτηση και ότι τα e_t είναι ασυσχέτιστα με τα Y_{t-1} έχουμε :

$$\gamma_Y\left(0\right)=Cov\left(Y_t,Y_t\right)=Cov\left(\varphi_1Y_{t-1}+e_t,\varphi_1Y_{t-1}+e_t\right)=Cov\left(\varphi_1Y_{t-1},\varphi_1Y_{t-1}\right)+Cov\left(\varphi_1Y_{t-1},e_t\right)+Cov\left(e_t,\varphi_1Y_{t-1}\right)+Cov\left(e_t,e_t\right)$$

$$=\varphi_1^2Cov\left(Y_{t-1},Y_{t-1}\right)+\varphi_1\overbrace{Cov\left(Y_{t-1},e_t\right)}^{=0}+\varphi_1\overbrace{Cov\left(E_t,Y_{t-1}\right)}^{=0}+Cov\left(e_t,e_t\right)$$

$$=\varphi_1^2Cov\left(Y_{t-1},Y_{t-1}\right)+Var\left(e_t\right)$$

$$\underset{\alpha\varphi o\acute{\upsilon}\ \{Y_t\}\ \text{στάσιμη}}{\overset{Cov\left(Y_{t-1},Y_{t-1}\right)=Cov\left(Y_t,Y_t\right)}{=}}\varphi_1^2Cov\left(Y_{t-1},Y_{t-1}\right)+Var\left(e_t\right)$$

$$=\varphi_1^2Cov\left(Y_t,Y_t\right)+\sigma_e^2$$

$$=\varphi_1^2\gamma_Y\left(0\right)+\sigma_e^2$$

$$\Rightarrow\gamma_Y\left(0\right)=\frac{\sigma_e^2}{\left(1-\varphi_e^2\right)}$$

Άρα η συνδιακύμανση μιας στάσιμης διαδικασίας $\left\{Y_t\right\}$ που ικανοποιεί την εξίσωση του $AR(1)$ μοντέλου είναι : $\gamma_Y\left(t,t-h\right)=Cov\left(Y_t,Y_{t-h}\right)=\varphi_1^h\frac{\sigma_e^2}{1-\varphi_e^2}$.

2.8 Αυτοπαλινδρομικά μοντέλα κινητού μέσου όρου ARMA(p,q)

Σε αυτή την ενότητα θα εισάγουμε ένα παραμετρικό μοντέλο στάσιμων χρονοσειρων, το αυτοπαλινδρομικό μοντέλο κινητού μέσου όρου $ARMA$ με παραμέτρους $p,\ q$, το οποίο αποτελεί ένα συνδιασμό των παραπάνω μοντέλων AR και MA.

Για μια μεγάλη κλάση δειγματικών συναρτήσεων αυτοσυνδιακύμανσης είναι πιθανό να βρούμε μια $ARMA$ διαδικασία $\{Y_t\}$ με συνάρτηση αυτοσυνδιακύμανσης $\gamma_Y(\cdot)$ όπου προσεγγίζεται πολύ καλά από την δειγματική συνάρτησης αυτοσυνδιακύμανσης $\gamma(\cdot)$ την οποία θα ορίσουμε στην ενότητα 2.11. Πιο συγκεκριμένα, για οποιονδήποτε θετικό ακέραιο k , υπάρχει μια $ARMA$ διαδικασία $\{Y_t\}$ τέτοια ώστε : $\gamma_Y(h)=\gamma(h)$ για h=0,1,…,k .

Για τον λόγο αυτό και για άλλους λόγους η γραμμική οικογένεια $ARMA$ διαδικασιών παίζει ένα πολύ σημαντικό ρόλο στην μοντελοποίηση μονοπαραμετρικών χρονοσειρών. Η διαδικασία $ARMA$ ορίζεται από γραμμικες εξισώσεις διαφορών με σταθερούς συντελεστές και περιέχει όρους της μορφής $\varphi_i Y_{t-i}$ και $\theta_i e_{t-i}$. Ο συνδιασμός $AR(p)$ και $MA(q)$ μοντέλων σε ένα μας δίνει το μοντέλο μηδενικής μέσης τιμής $ARMA(p,q)$, με p αυτοπαλινδρομικούς όρους και q όρους κινητού μέσου όρου που ορίζεται ως εξής :

$$\begin{aligned} & Y_t-\varphi_1 Y_{t-1}-\ldots-\varphi_p Y_{t-p}=e_t+\theta_1 e_{t-1}+\ldots+\theta_q e_{t-q} \\ \Rightarrow\ & Y_t=\sum_{i=1}^{p}\varphi_i Y_{t-i}+\sum_{i=1}^{q}\theta_i e_{t-i} \end{aligned} \qquad (2.8.1)$$

Εδώ θεωρούμε ότι τα πολυώνυμα $\left(1-\varphi_1\lambda-\ldots-\varphi_p\lambda^p\right)$ και $\left(1+\theta_1\lambda+\ldots+\theta_p\lambda^p\right)$ δεν έχουν κοινούς παράγοντες και οτι $\{e_t\}\sim WN\left(0,\sigma_e^2\right)$ όπως ορίστηκε στη διαδικασία $MA(q)$. Σε αυτό το μοντέλο, όπως φαίνεται και από την εξίσωση (2.8.1) θα υποτεθεί για ένα αρχικό σύνολο δεδομένων $\{x_1,x_2,\ldots,x_n\}$ είτε ότι η μέση τιμή τους είναι μηδέν, είτε ότι τα δεδομένα μας έχουν υποστεί δίορθωση της μέσης τιμής αφαιρώντας από αυτά τη δειγματική μέση τιμή τους (διότι εάν το δείγμα μας είναι ικανοποιητικά μεγάλο, τότε από κεντρικό οριακό θεώρημα η μέση τιμή της διαδικασίας θα προσεγγίζεται πολύ καλά από τη δειγματική). Έτσι, αν θεωρήσουμε ότι $\{Y_t\}$ είναι η διαδικασία που προέκυψε μετά τον μετασχηματισμό αυτό, το αντίστοιχο μοντέλο για την αρχική μας χρονοσειρά $\{X_t\}$ μπορεί να βρεθεί αντικαθιστώντας το Y_t, για κάθε t, με το $X_t-\bar{x}$.

Ο δειγματικός μέσος για ένα σύνολο δεδομένων $x_1, x_2, \ldots, x_n$ ορίζεται ως:

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t \ .$$

Με βάση τις παραπάνω παρατηρήσεις, τα μετασχηματισμένα δεδομένα μας $\{y_1, y_2, \ldots, y_n\}$ θα είναι κατάλληλα να προσαρμοστούν σε ένα μηδενικής μέσης τιμής $ARMA$ μοντέλο της εξίσωσης (2.8.1).

Χρησιμοποιώντας και εδώ τον τελεστή οπίσθιας μετάθεσης και σύμφωνα με τους συμβολισμούς των ενοτήτων 2.6 και 2.7 η εξίσωση (2.8.1) γράφεται εν συντομία :

$$\varphi(B)Y_t = \theta(B)e_t, \qquad \{e_t\} \sim WN\left(0, \sigma_e^2\right) \qquad (2.8.2)$$

όπου τα ποπλυώνυμα $\left(1-\varphi_1 B - \ldots - \varphi_p B^p\right)$, $\left(1+\theta_1 B + \ldots + \theta_q B^q\right)$ δεν έχουν κοινούς παράγοντες. Το μοντέλο $ARMA$ είναι κατάλληλα μόνο στην περίπτωση όπου η $\{Y_t\}$ είναι στάσιμη χρονοσειρά και σε περίπτωση που δεν παρουσιάζει στασιμότητα μπορούμε να δοκιμάσουμε να την μετασχηματίσουμε έτσι ώστε να γίνει στάσιμη και έπειτα να εφορμόσουμε το μοντέλο.

Προκειμένου να προσδιοριστεί το μοντέλο $ARMA(p,q)$ της εξίσωσης (2.8.2) είναι αναγκαίο να καθορίσουμε πρώτα $\text{τις p+q+1 άγνωστες παραμέτρους}$, δηλαδή τους συντελεστές $\varphi_1, \varphi_2, \ldots, \varphi_p$, $\theta_1, \theta_2, \ldots, \theta_q$ και την διασπορά του λευκού θορύβου σ_e^2.

Στο σημείο αυτό πρέπει να πούμε ότι για την εξίσωση (2.8.1) μια στάσιμη λύση $\{Y_t\}$ υπάρχει και είναι μοναδική, αν και μόνο αν, για το AR πολυώνυμο ισχύει ότι είναι διάφορο του μηδενός για όλα τα σημεία που έχουν μέτρο ένα, δηλαδή

$$\varphi(\lambda) = 1 - \varphi_1\lambda - \ldots - \varphi_p\lambda^p \neq 0 \text{ για όλα τα} |\lambda| = 1 \quad (2.8.3)$$

Όπως φαίνεται από την εξίσωση (2.8.1) στην $ARMA$ διαδικασία η Y_t μπορεί να εκφραστεί σε όρους e_t. Παρακάτω θα δούμε μερικές ιδιότητες για τη σχεση των δύο αυτών διαδικασιών $\{Y_t\}$ και $\{e_t\}$ που εμφανίζονται στον ορισμό της $ARMA$ διαδικασίας της (2.8.1).

Θα λέμε ότι μια $ARMA$ διαδικασία $\{Y_t\}$ είναι αιτιώδης (ή αιτιώδης συνάρτηση της $\{e_t\}$) εαν υπάρχουν σταθερές $\{\psi_j\}$ τέτοιες ώστε $\sum_{j=0}^{\infty}|\psi_j| < \infty$ και

$$Y_t = \sum_{j=0}^{\infty}\psi_j e_{t-j} \text{ για όλα τα } t. \qquad (2.8.4)$$

Η αιτιότητα είναι ισοδύναμη με τη συνθήκη :

$$\varphi(\lambda) = 1 - \varphi_1\lambda - \ldots - \varphi_p\lambda^p \neq 0 \text{ για όλα τα } |\lambda| \leq 1 \qquad (2.8.5)$$

Αυτό σημαίνει οτι οι ρίζες του AR πολυωνύμου βρίσκονται εκτός του μοναδιαίου κύκλου, γεγονός που εξασφαλίζει την στασιμότητα για το AR μέρος της διαδικασίας $ARMA$. Η ακολουθία $\{\psi_j\}$ στην (2.8.4) υπολογίζεται από τη σχέση $\psi(\lambda)=\sum_{j=0}^{\infty}\psi_j\lambda^j=\frac{\theta(\lambda)}{\varphi(\lambda)}$ που ισοδύναμα γίνεται:

$$\left(1-\varphi_1\lambda-\ldots-\varphi_p\lambda^p\right)\left(\psi_0+\psi_1\lambda+\ldots\right)=1+\theta_1\lambda+\ldots+\theta_q\lambda^q.$$

Εξισώνοντας τους συντελεστές $\lambda^j,\ j=0,1,\ldots,$ βρίσκουμε ότι :

$$\begin{aligned}1&=\psi_0\\ \theta_1&=\psi_1-\psi_0\varphi_1\\ \theta_2&=\psi_2-\psi_1\varphi_1-\psi_0\varphi_2,\\ &\vdots\end{aligned}$$

ή ισοδύναμα,

$$\psi_j-\sum_{k=1}^{p}\varphi_k\psi_{j-k}=\theta_j,\ \ j=0,1,\ldots,\ \text{όπου}\ \theta_0=1,\ \theta_j=0\ \text{για}\ j>q\ \text{και}\ \psi_j=0\ \text{για}\ \ j<0\quad(2.8.6).$$

Αντίθετα, η ένοια της αντιστρεψιμότητας επιτρέπει την e_t να εκφραστεί σε όρους $Y_s,\ s<t$ και καθορίζεται από το MA μέρος της διαδικασίας $ARMA$. Συνεπώς, θα λέμε ότι μια $ARMA$ διαδικασία $\{Y_t\}$ είναι αντιστέψιμη εάν υπάρχουν σταθερές $\{\pi_j\}$ τέτοιες ώστε $\sum_{j=0}^{\infty}|\pi_j|<\infty$ και

$$e_t=\sum_{j=0}^{\infty}\pi_jY_{t-j}\ \text{για όλα τα}\ t$$

Η αντιστρεψιμότητα είναι ισοδύναμη με τη συνθήκη:

$$\theta(\lambda)=1+\theta_1\lambda+\ldots+\theta_q\lambda^q\neq 0\ \text{για όλα τα}\ |\lambda|\leq 1$$

Αλλάζοντας το AR πολυώνυμο στην εξίσωση (2.8.6) με το MA πολυώνυμο βρίσκουμε ότι η ακολουθία $\{\pi_j\}$ καθορίζεται από τις εξισώσεις :

$$\pi_j-\sum_{k=1}^{q}\theta_k\pi_{j-k}=-\varphi_j,\ \ j=0,1,\ldots,\ \text{όπου}\ \varphi_0=-1,\ \varphi_j=0\ \text{για}\ j>p\ \text{και}\ \pi_j=0\ \text{για}\ \ j<0\quad(2.8.7)$$

Ας σκεφτούμε για παράδειγμα την $ARMA(1,1)$ διαδικασία Y_t που ικανοποιεί την εξίσωση :

$$Y_t-0.5Y_{t-1}=e_t-0.4e_{t-1}\ ,\quad \{e_t\}\sim WN\left(0,\sigma_e^2\right)\quad(2.8.9)$$

Εφόσον το AR πολυώνυμο $\varphi(\lambda)=1-0.5\lambda$ μηδενίζεται για $\lambda=2$ που πέφτει εκτός του μοναδιαίου κύκλου από την (2.8.3) και την (2.8.5) συμπεραίνουμε ότι υπάρχει μοναδική ARMA διαδικασία που ικανοποιεί την (2.8.9) που είναι επίσης αιτιώδης. Οι συντελεστές του $MA(\infty)$ αναπαράστασης της $\{Y_t\}$ βρίσκονται απευθείας από την σχέση (2.8.6) η οποία δίνει :

$$\begin{aligned}
&\psi_0=1,\\
&\psi_1=0.4+0.5,\\
&\psi_2=0.5(0.4+0.5),\\
&\psi_j=0.5^{j-1}(0.4+0.5),\quad j=1,2,...
\end{aligned}$$

Αντίστοιχα, το MA πολυώνυμο $\theta(\lambda)=1+0.4\lambda$ μηδενίζεται για $\lambda=-2.5$ που πέφτει επίσης εκτός του μοναδιαίου κύκλου συνεπώς συμπεραίνουμε ότι η $\{Y_t\}$ είναι αντιστρέψιμη με συντελεστές $\{\pi_j\}$ που δίνονται από την (2.8.7) :

$$\begin{aligned}
&\pi_0=1,\\
&\pi_1=-(0.4+0.5),\\
&\pi_2=-(0.4+0.5)(-0.4),\\
&\pi_j=-(0.4+0.5)(-0.4)^{j-1},\quad j=1,2,...
\end{aligned}$$

Κλείνοντας θα σημειώσουμε ότι εάν η $\{Y_t\}$ είναι μια ARMA διαδικασία που ορίζεται από την $\varphi(B)Y_t=\theta(B)e_t$, όπου $\theta(\lambda)\neq 0$ για $|\lambda|=1$, τότε είναι πάντα πιθανό να βρούμε πολυώνυμα $\hat{\varphi}(\lambda)$,$\hat{\theta}(\lambda)$ και μια ακολουθία λευκού θορύβου $\{W_t\}$ τέτοια ώστε $\hat{\varphi}(B)Y_t=\hat{\theta}(B)W_t$ με $\hat{\theta}(\lambda)$ και $\hat{\varphi}(\lambda)$ να είναι διάφορα του μηδενός για $|\lambda|\leq 1$. Ωστόσο, εάν η αρχική ακολουθία λευκού θορύβου των δεδομένων μας $\{e_t\}$ είναι iid, τότε η νέα ακολουθία λευκού θορύβου δεν συνεπάγεται ότι θα είναι και αυτή iid εκτός αν η $\{e_t\}$ είναι γκαουσιανός λευκός θόρυβος. Όπως θα δούμε παρακάτω, ο ορισμός της αιτιότητας παίζει σημαντικό ρόλο στον καθορισμό των συναρτήσεων ACF και PACF που θα ορίσουμε στην ενότητα 2.11.

2.9 Ολοκληρώσιμα αυτοπαλινδρομικά μοντέλα κινητού μέσου όρου ARIMA(p,d,q):

Σε αυτή την ενότητα θα εξετάσουμε την περίπτωση εύρεσης ενός κατάλληλου μοντέλου για ένα δοθέν σύνολο δεδομένων $\{x_1, x_2, ..., x_n\}$ που δεν δημιουργήθηκε απαραίτητα από μια στάσιμη διαδικασία, όπως τα δεδομένα που παρουσιάστηκαν στα Διαγράμματα 2.1.1 και 2.3.1. Στην ενότητα 2.2 είδαμε ότι μια προσέγγιση για να επιτευχθεί αυτό είναι να εφαρμόσουμε τον τελεστή διαφοράς ∇, επαναλαμβανόμενα στη αρχική μη στάσιμη σειρά από την οποία προέρχονται τα δεδομένα μας, μέχρι οι διαφορές των παρατηρήσεων να μοιάζουν με μια πραγματοποίηση μιας στάσιμης διαδικασίας. Συνεπώς, εάν $\{X_t\}$ είναι η αρχική μας μη στάσιμη διαδικασία περιμένουμε έπειτα από την εφαρμογή του τελεστή ∇, d φορές, η διαδικασία $\{Y_t\}$: $Y_t = (1-B)^d X_t = \nabla^d (X_t)$ να είναι στάσιμη.

Για παράδειγμα εάν δοκιμάσουμε να πάρουμε τις πρώτες διαφορές $(d=1)$ για την αρχική διαδικασία $\{X_t\}$ έχουμε :

$$Y_t = (1-B)^1 X_t = X_t - BX_t = X_t - X_{t-1}$$

Εάν δεν προκύψει στάσιμη διαδικασία με τις πρώτες διαφορές, δοκιμάζουμε τις δεύτερες $(d=2)$ και έχουμε :

$$Y_t = (1-B)^2 X_t = (1-2B+B^2) X_t = X_t - 2BX_t + B^2 X_t = X_t - 2X_{t-1} + X_{t-2}\,.$$ κ.ο.κ

Με βάση αυτή την παρατήρηση ορίζουμε το Ολοκληρώσιμο (Integrated) μοντέλο τάξης d, για μια αρχική μη στάσιμη διαδικασία $\{X_t\}$ να είναι :

$$Y_t = \nabla^d (X_t) = \mu + \varepsilon_t \quad ,$$

όπου d μη αρνητικός ακέραιος, μ η σταθερή πλέον μέση τιμή της $\{Y_t\}$ και ε_t ένα τυχαίο σφάλμα.

Ο συνδυασμός του Ολοκληρώσιμου μοντέλου με το $ARMA$ μοντέλο που αναλύσαμε στην προηγούμενη ενότητα ενσωματώνουν μια μεγάλη ποικιλία μη στάσιμων χρονοσειρών που μας παρέχονται από την $ARIMA$ διαδικασία, η οποία μειώνεται σε μια $ARMA$ διαδικασία έπειτα από την εφαρμογή του τελεστή διαφοράς ∇, d πεπερασμένες φορές. Έτσι, το $ARIMA(p,d,q)$ μοντέλο τάξης p,d,q , το οποίο περιλαμβάνει p όρους του AR μοντέλου , d-διαφορές και q όρους του MA μοντέλου ορίζεται να είναι :

$$Y_t = \varphi_1 Y_{t-1} + ... + \varphi_p Y_{t-p} + e_t + \theta_1 e_{t-1} + ... + \theta_q e_{t-q} \text{ , όπου } Y_t = \nabla^d X_{t-1} \quad (2.9.1)$$

Χρησιμοποιώντας τα πολυώνυμα $\varphi(B)$ και $\theta(B)$ που ορίσαμε στις προηγούμενες ενότητες η (2.9.1) γίνεται :

$$\underbrace{\left(1-\varphi_1 B-\varphi_2 B^2-\cdots-\varphi_q B^q\right)}_{\varphi(B)} Y_t = \underbrace{\left(1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q\right)}_{\theta(B)} e_t$$

$$\Rightarrow \varphi(B) Y_t = \theta(B) e_t \quad , \text{ όπου } Y_t = \nabla^d (X_t) = (1-B)^d X_t$$

Αν θέλουμε να γράψουμε την παραπάνω σχέση χρησιμοποιώντας την αρχική χρονοσειρά $\{X_t\}$, μπορούμε να ορίσουμε $\varphi^*(B) = \varphi(B)(1-B)^d$ και συνεπώς θα έχουμε:

$$\varphi^*(B) X_t \equiv \varphi(B)(1-B)^d X_t = \theta(B) e_t, \qquad \mathrm{e}_t \sim WN\left(0, \sigma_e^2\right)$$

όπου τα πολυώνυμα $\varphi(B)$ και $\theta(B)$ είναι βαθμών p και q αντίστοιχα με $\varphi(B) \neq 0$ για όλα τα $|B| \leq 1$. Δηλαδή το AR πολυώνυμο, $\varphi(B)$, δε πρέπει να μηδενίζεται για τις τιμές του B που είναι κατά απόλυτη τιμή μικρότερες από τη μονάδα και αυτό συμβαίνει διότι έχουμε θεωρήσει πως έπειτα από d το πλήθος διαφορές, η χρονοσειρά $\{Y_t\}$ που προέκυψε από αυτές, $Y_t = \nabla^d (X_t)$, είναι στάσιμη. Για να συμβαίνει αυτό, όπως είπαμε και στην ενότητα 2.7, πρέπει οι ρίζες του AR πολυωνύμου να βρίσκονται εκτός του μοναδιαίου κύκλου και συνεπώς, το πολυώνυμο $\varphi(B)$ να μη μηδενίζεται για τις τιμές του B που βρίσκονται εντός του μοναδιαίου κύκλου. Το πολυώνυμο $\varphi^*(B) = \varphi(B)(1-B)^d$ έχει ακριβώς d ρίζες ίσες με την μονάδα, δηλαδή οι ρίζες του βρίσκονται πάνω στην περιφέρεια του μοναδιαίου κύκλου. Η αρχική διαδικασία $\{X_t\}$ είναι στάσιμη αν και μόνο αν $d = 0$, που σε αυτή την περίπτωση μειώνεται σε μια $ARMA(p,q)$ διαδικασία.

Η R μας δίνει την δυνατότητα να υπολογίσουμε τις αντίστροφες ρίζες των AR και MA πολυωνύμου αντιστοίχως για να ελέγξουμε την στασιμότητα. Αυτό γίνεται ως εξής:

Ας υποθέσουμε ότι έχουμε μια $ARIMA(p,d,q)$ διαδικασία :

$$\varphi(B) Y_t = \theta(B) e_t \quad , \text{ όπου } Y_t = \nabla^d (X_t) = (1-B)^d X_t \text{ και}$$

$$\varphi(B) = 1-\varphi_1 B-\varphi_2 B^2-\cdots-\varphi_q B^q$$
$$\theta(B) = 1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q$$

Τότε η R για να υπολογίσει τις αντίστροφες ρίζες των AR και MA πολυωνύμων αντιστοίχως, δημιουργεί τους τετραγωνικούς πίνακες (companion matrices):

$$F(\varphi)=\begin{bmatrix} \varphi_1 & \varphi_2 & . & . & . & \varphi_{p-1} & \varphi_p \\ 1 & 0 & . & . & . & 0 & 0 \\ 0 & 1 & . & . & . & 0 & 0 \\ . & . & . & & & . & . \\ . & . & & . & & . & . \\ . & . & & & . & . & . \\ 0 & 0 & . & . & . & 1 & 0 \end{bmatrix}, \; F(\theta)=\begin{bmatrix} \theta_1 & \theta_2 & . & . & . & \theta_{q-1} & \theta_q \\ 1 & 0 & . & . & . & 0 & 0 \\ 0 & 1 & . & . & . & 0 & 0 \\ . & . & . & & & . & . \\ . & . & & . & & . & . \\ . & . & & & . & . & . \\ 0 & 0 & . & . & . & 1 & 0 \end{bmatrix}$$

Και έπειτα βρίσκει τις ιδιοτιμές τους. Τα μέτρα των ιδιοτιμών του $F(\varphi)$ είναι οι αντίστροφες ρίζες του AR πολυωνύμου και αντίστοιχα τα μέτρα των ιδιοτιμών του $F(\theta)$ είναι οι αντίστροφες ρίζες του MA πολυωνύμου. Σε περίπτωση μιγαδικής ιδιοτιμής $r+ci$ το μέτρο είναι $\sqrt{r^2+c^2}$ και όπως έχει δειχτεί από Hamilton (1994, κεφ.1) η διαδικασία $\{Y_t\}$ είναι στάσιμη και αντιστρέψιμη εάν τα μέτρα κάθε ιδιοτιμής του F είναι αυστηρά μικρότερα του 1, δηλαδή πέφτουν εντός του μοναδιαίου κύκλου.

Συνοψίζοντας, εάν d είναι ένας μη αρνητικός ακέραιος, τότε η $\{X_t\}$ είναι μια αιτιώδης $ARIMA(p,d,q)$ διαδικασία εάν η $Y_t=(1-B)^d X_t$ είναι μια $ARMA(p,q)$ διαδικασία. Η εξίσωση (2.9.1) της $ARIMA$ διαδικασίας είναι χρήσιμη για να αναπαραστήσουμε δεδομένα και στην περίπτωση που παρουσιάζουν τάση αλλά και στην περίπτωση όπου δεν παρουσιάζουν τάση.

2.10 Συγκεντρωτικός Πίνακας

Με βάση όλα τα παραπάνω, το σημαντικότερο βήμα προκειμένου να επιλέξουμε ποιό μοντέλο ενδείκνυται να αναπαραστήσει ένα δοθέν σύνολο δεδομένων, είναι να αναγνωρίσουμε εάν τα δεδομένα μας είναι στάσιμα ή όχι. Ανάλογα με το εάν έχουμε ένα σύνολο δεδομένων που προέρχεται από μια στάσιμη ή όχι διαδικασία, επιλέγουμε τα αντίστοιχα μοντέλα που αναπαριστούν στάσιμες και μη στάσιμες διαδικασίες. Όπως είδαμε στο Κεφάλαιο 2 κάποια μοντέλα μπορεί να είναι είτε στάσιμα είτε μη στάσιμα ανάλογα με συγκεκριμένες συνθήκες που πρέπει να ικανοποιούνται. Στον Πίνακα 2.10.1 παρουσιάζονται συνοπτικά όλα τα μοντέλα που συζητήθηκαν στο Κεφάλαιο 2 και κατηγοριοποιούνται ανάλογα με την συνθήκη στασιμότητας.

Πίνακας 2.10.1:

Μοντέλο	**Στάσιμο**	**Μη-Στάσιμο**
MA(q)	Εκ κατασκευής στάσιμο	—
AR(p)	Όταν οι ρίζες του AR πολυωνύμου, $\varphi(\mathrm{B}) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_q B^q$ πέφτουν εκτός του μοναδιαίου κύκλου	Όταν τουλάχιστον μια ρίζα του AR πολυωνύμου, $\varphi(\mathrm{B}) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_q B^q$ πέφτει εντός του μοναδιαίου κύκλου
ARMA(p,q)	Καθορίζεται από το AR μέρος του μοντέλου, δηλαδή όταν οι ρίζες του πολυωνύμου πέφτουν εκτός του μοναδιαίου κύκλου.	Όταν τουλάχιστον μια ρίζα του AR πολυωνύμου, $\varphi(\mathrm{B}) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_q B^q$ πέφτει εντός του μοναδιαίου κύκλου
ARIMA(p,q,d)	όταν d=0 και ταυτίζεται με το ARMA(p,q) μοντέλο	όταν $d \neq 0$ είναι από την κατασκευή του μη στάσιμο

2.11 Συνάρτηση Αυτοσυσχέτισης ACF και Μερικής Αυτοσυσχέτισης PACF

Οι συναρτήσεις αυτοσυσχέτισης ACF και μερικής αυτοσυσχέτισης PACF είναι δύο πολύ χρήσιμα εργαλεία που μας βοηθάνε αφενός να ανιχνεύσουμε συσχέτιση στα δεδομένα μας και αφετέρου να επιλέξουμε πιο είναι το κατάλληλο μοντέλο για να αναπαραστήσουμε τα δεδομένα μας. Για μια στάσιμη χρονοσειρά $\{X_t\}$ όπως ορίσαμε παραπάνω, η συνάρτηση αυτοσυσχέτισης, σε υστέρηση h, θα δίνεται από τον τύπο:

$$\rho(h)=\frac{\gamma(h)}{\gamma(0)}=\frac{Cov(X_{t+h},X_t)}{Cov(X_t,X_t)}=\frac{E\left[\left(X_{t+h}-\mu_X(t+h)\right)\left(X_t-\mu_X(t)\right)\right]}{Var(X_t)}$$

$$\underset{\text{λόγω στασιμότητας}}{\overset{\mu_X(t+h)=\mu_X(t)=\mu_X}{=}}\frac{E\left[\left(X_{t+h}-\mu_X\right)\left(X_t-\mu_X\right)\right]}{Var(X_t)}=\frac{E\left[\left(X_{t+h}-\mu_X\right)\left(X_t-\mu_X\right)\right]}{E\left[\left(X_t-\mu_X\right)^2\right]}$$

Δηλαδή,

$$\rho(h)=\frac{\gamma(h)}{\gamma(0)}=\frac{E\left[\left(X_{t+h}-\mu_X\right)\left(X_t-\mu_X\right)\right]}{E\left[\left(X_t-\mu_X\right)^2\right]}=Cor(X_{t+h},X_t)$$

Στον αριθμητή έχουμε την συνδιακύμανση μεταξύ δύο παρατηρήσεων που απέχουν h χρονικές στιγμές μεταξύ τους, η οποία είναι ίδια για κάθε *t* , και στον παρονομαστή έχουμε την κοινή διασπορά όλων των παρατηρήσεων την οποία συμβολίζουμε με $\gamma(0)$. Οι τιμές που λαμβάνει η συνάρτηση ρ κυμαίνονται από -1 έως 1 με την τιμή 1 να δείχνει τέλεια θετική γραμμική συσχέτιση, την τιμή -1 τέλεια γραμμική αρνητική συσχέτιση και την τιμή 0 καθόλου συσχέτιση. Από τα παραπάνω έπεται ότι η τιμή της συνάρτησης αυτοσυσχέτισης σε υστέρηση $h=0$ είναι ίση με τη μονάδα.

Ωστόσο, επειδή στην πράξη έχουμε πεπερασμένες παρατηρήσεις της διαδικασίας X_t, εκτιμούμε την μέση τιμή, την διασπορά, την αυτοσυνδιακύμανση και την αυτοσυσχέτιση με τις αντίστοιχες δειγματικές συναρτήσεις.

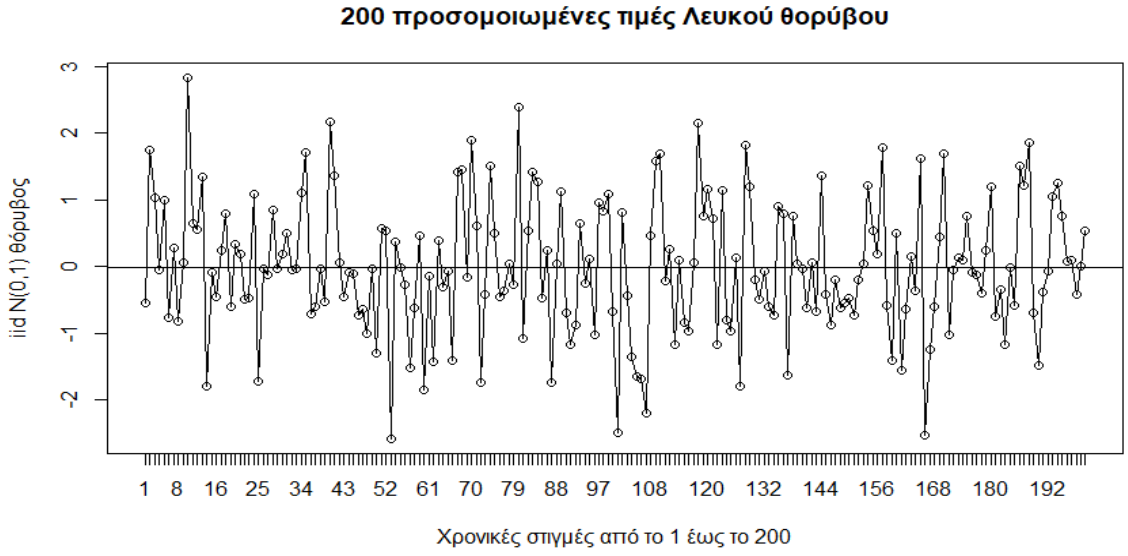
Συνεπώς, για ένα δείγμα $x_1, x_2, \ldots, x_n$ παρατηρήσεων μιας χρονοσειράς, η δειγματική μέση τιμή των $x_1, x_2, \ldots, x_n$ όπως ορίσαμε και παραπάνω δίνεται από τον τύπο: $\bar{x}=\frac{1}{n}\sum_{t=1}^{n}x_t$. Αντίστοιχα, η δειγματική διασπορά δίνεται από τον τύπο: $\hat{\sigma}_X^2=\frac{1}{n}\sum_{t=1}^{n}\left(x_t-\bar{x}\right)^2$ και η δειγματική συνάρτηση αυτοσυνδιακύμανσης ορίζεται ως:

$$\hat{\gamma}(h)=\frac{1}{n}\sum_{t=1}^{n-|h|}\left(x_{t+|h|}-\bar{x}\right)\left(x_t-\bar{x}\right),\quad \text{-n < h < n .}$$

Δηλαδή, για $h \geq 0$ η $\hat{\gamma}(h)$ είναι περίπου ίση με την δειγματική συνδιακύμανση των $n-h$ ζευγαριών παρατηρήσεων $(x_1, x_{1+h}), (x_2, x_{2+h}), \ldots, (x_{n-h}, x_n)$. Τέλος, η δειγματική συνάρτηση αυτοσυσχέτισης είναι:

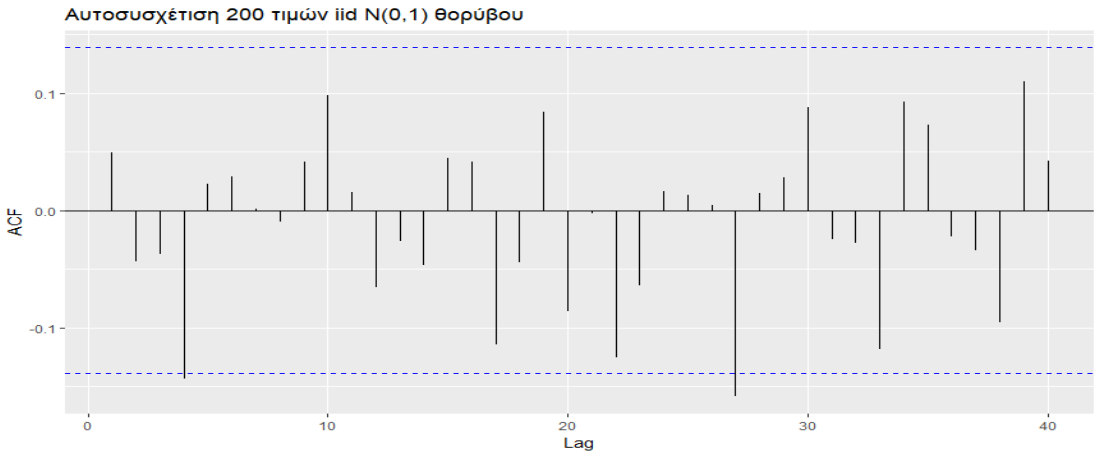
$$r_h = \hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \frac{\sum_{t=1}^{n-|h|} \left(x_{t+|h|} - \bar{x}\right)\left(x_t - \bar{x}\right)}{\sum_{t=1}^{n} \left(x_t - \bar{x}\right)^2}, \quad \text{-n < h < n .}$$

Εφόσον οι πραγματικές τιμές της ρ_h δεν μας είναι γνωστές, ούτε μπορούν να υπολογιστούν από ένα δείγμα πεπερασμένων δεδομένων $x_1, x_2, \ldots, x_n$, δημιουργείται η ανάγκη να τις εκτιμήσουμε επιλέγοντας κατάλληλες εκτιμήτριες. Για παράδειγμα εάν υποθέσουμε ότι ένα σύνολο δεδομένων $x_1, x_2, \ldots, x_n$ μπορεί να είναι οι πραγματικές τιμές μιας στάσιμης διαδικασίας $\{X_t\}$ τότε η δειγματική συνάρτηση αυτοσυσχέτισης r_h μας παρέχει μια εκτίμηση για την άγνωστη συνάρτηση αυτοσυσχέτισης ρ_h της $\{X_t\}$. Αυτή η εκτίμηση μπορεί να προτείνει πιο από τα υποψήφια στάσιμα μοντέλα χρονοσειράς είναι πιθανόν να αναπαραστήσει την εξάρτηση που συναντούμε στα δεδομένα μας. Πιο συγκεκριμένα, μια δειγματική συνάρτηση αυτοσυσχέτισης r_h που είναι κοντά στο μηδέν για όλες τις μη μηδενικές υστερήσεις h , προτείνει ότι ένα κατάλληλο μοντέλο αναπαράστασης των δεδομένων μας είναι αυτό που αποτελείται από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με μηδενική μέση τιμή. Το Διάγραμμα 2.11.1 δείχνει 200 προσομοιωμένες τιμές που προέρχονται από ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση 1. Μάλιστα, ένα τέτοιο μοντέλο, καλείται γκαουσιανός λευκός θόρυβος και συμβολίζεται $IID\ \mathrm{N}(0,1)$.



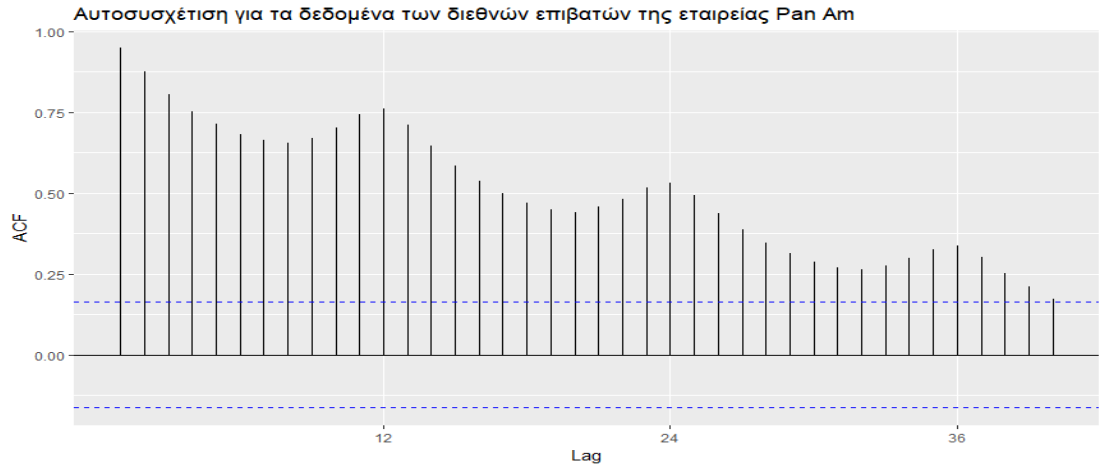
<u>Διάγραμμα 2.11.1</u>: Χρονοσειρά 200 προσομοιωμένων τιμών iid N(0,1) θορύβου

Το Διάγραμμα 2.11.2 δείχνει τις αντίστοιχες δειγματικές αυτοσυσχετίσεις r_h σε υστερήσεις h=1,2,…,40 και ονομάζεται διάγραμμα αυτοσυσχέτισης (ACF). Για ανεξάρτητες τυχαίες μεταβλητές, δεν υπάρχει συσχέτιση και συνεπώς $\rho(h)=0$ για $h>0$. Επομένως, θα περιμένουμε και οι αντίστοιχες δειγματικές αυτοσυσχετίσεις r_h να είναι κοντά στο 0. Πιο συγκεκριμένα, μπορεί να δειχτεί ότι για μια ακολουθία από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με πεπερασμένη διακύμανση, η δειγματική συνάρτηση αυτοσυσχέτισης r_h, για μεγάλες τιμές του n, είναι τυχαία μεταβλητή iid θορύβου από την κανονική κατανομή με μέση τιμή $\mu=0$ και διασπορά $\sigma^2=1/n$. Συμβολίζουμε με $IID\ \mathrm{N}\left(0,\frac{1}{n}\right)$. Ως εκ τούτου περιμένουμε περίπου το 95% των δειγματικών αυτοσυσχετίσεων r_h, για $h>0$, να είναι εντός των ορίων $\pm 1.96/\sqrt{n}$. Αυτό γιατί για την τυχαία μεταβλητή $r_h \sim \mathrm{N}\left(0,\frac{1}{n}\right)$, η τιμή 1,96 είναι το Z=0,95 ποσοστημόριο της τυπικής κανονικής κατανομής και συνεπώς : $r_h=\mu+Z\cdot\sigma \underset{\sigma=\sqrt{1/n}}{\overset{\mu=0}{=}} 0+1.96\cdot\sqrt{\frac{1}{n}}=\frac{1.96}{\sqrt{n}}$. Αφού περιμένουμε το 95% των τιμών της δειγματικής αυτοσυσχέτισης r_h να είναι εντός των ορίων $\pm 1.96/\sqrt{n}$, το υπόλοιπο 5% πρέπει να πέφτει εκτός. Δηλαδή από το σύνολο των 40 τιμών r_h, h=1,2,…,40 περιμένουμε δύο τιμές $\left(\frac{5}{100}\cdot 40=2\right)$ να πέφτουν εκτός των ορίων $\pm 1.96/\sqrt{n}$. Στο Διάγραμμα 2.11.2 φαίνονται οι τιμές της δειγματικής συνάρτησης αυτοσυσχέτισης r_h για τα δεδομένα του γκαουσιανού λευκού θορύβου του Διαγράμματος 2.11.1 σε υστερήσεις h=1,2,…,40. Οι οριζόντιες μπλε διακεκομμένες γραμμές παριστάνουν τα όρια $\pm 1.96/\sqrt{n}$. Όπως αναμέναμε $r_h \approx 0$ για όλες τις υστερήσεις h, εκτός από τις 2 τιμές στις υστερήσεις $h=4$ και $h=27$ οι οποίες πέφτουν εκτός των ορίων $\pm 1.96/\sqrt{n}$.



<u>Διάγραμμα 2.11.2</u>: Οι τιμές της δειγματικής συνάρτησης αυτοσυσχέτισης για τα δεδομένα του Διαγράμματος 2.11.1 σε υστερήσεις h=1,2,…,40 μαζί με τα όρια $\pm 1.96/\sqrt{n}$.

Όπως φαίνεται, η δειγματική αυτοσυνδιακύμανση και αυτοσυσχέτιση μπορούν να υπολογιστούν για κάθε σύνολο δεδομένων $x_1, x_2, \ldots, x_n$ και δεν περιορίζονται μόνο σε παρατηρήσεις από στάσιμες χρονοσειρές. Για δεδομένα που περιέχουν τάση, η $\left|\hat{\rho}(h)\right|$ θα παρουσιάζει αργή πτώση καθώς η υστέρηση h αυξάνει και για δεδομένα με σημαντική ντετερμινιστική περιοδική συνιστώσα η $\left|\hat{\rho}(h)\right|$ θα παρουσιάζει ίδια συμπεριφορά με την ίδια περιοδικότητα. Στο Διάγραμμα 2.11.3 φαίνονται οι τιμές της δειγματικής συνάρτησης αυτοσυσχέτισης που αντιστοιχούν στα δεδομένα του Διαγράμματος 2.3.1 των διεθνών επιβατών της αεροπορικής εταιρείας Pan Am. Είναι εμφανή από το διάγραμμα αυτοσυσχέτισης ACF, η τάση και η σταθερή περιοδικότητα των 12 μηνών όπως αναμέναμε.



<u>Διάγραμμα 2.11.3</u>: Οι τιμές της δειγματικής συνάρτησης αυτοσυσχέτισης για τα δεδομένα των διεθνών επιβατών της αεροπορικής εταιρείας Pan Am σε υστερήσεις h=1,2,…,40.

Όπως βλέπουμε η συνάρτηση αυτοσυσχέτισης (ACF) μιας στάσιμης διαδικασίας $\{X_t\}$, σε υστέρηση h, είναι ένα μέτρο που δείχνει την γραμμική εξάρτηση μεταξύ των παρατηρήσεων της ίδιας μεταβλητής που επέχουν h χρονικές στιγμές μεταξύ τους.

Επειδή η συνάρτηση αυτοσυσχέτισης $\rho(\cdot)$ είναι μια άρτια συνάρτηση προκύπτει ότι:

$$\rho(h) = \rho(-h)$$

Δηλαδή, η αυτοσυσχέτιση μεταξύ της X_t και X_{t+h} είναι η ίδια με την αυτοσυσχέτιση X_t και X_{t-h}. Ωστόσο, η εξάρτηση μεταξύ των ενδιάμεσων μεταβλητών X_s, $t-h<s<t$ παίζει επίσης σημαντικό ρόλο.

Για παράδειγμα, στο $AR(1)$ μοντέλο της εξίσωσης (2.7.4) : $Y_t = \varphi_1 Y_{t-1} + e_t$, $|\varphi_1| < 1$ και $e_t \sim WN(0, \sigma_e^2)$, έχουμε βρει ήδη την συνάρτηση αυτοσυνδιακύμανσης από τη σχέση (2.7.8) η οποία ισούται με $\gamma(h) = Cov(Y_t, Y_{t-h}) = \varphi_1^h \gamma(0)$. Εύκολα προκύπτει για το $AR(1)$ μοντέλο ότι η συνάρτηση αυτοσυσχέτισης είναι:

$$\rho(h) = Cor(Y_t, Y_{t-h}) = \frac{\gamma(h)}{\gamma(0)} = \frac{\varphi_1^h \gamma(0)}{\gamma(0)} = \varphi_1^h, \text{ για } h>0$$

$$\Rightarrow \rho(h) = \varphi_1^h, \quad h > 0. \qquad (2.11.1)$$

Άρα για $h = 2$ έχουμε: $\rho(2) = Cor(Y_t, Y_{t-2}) = \varphi_1^2 > 0$ το οποίο σημαίνει ότι οι Y_t, Y_{t-2} είναι συσχετισμένες. Ωστόσο, οι Y_t, Y_{t-2} δεν είναι απευθείας συσχετισμένες. Ουσιαστικά, η αυτοσυσχέτιση $Cor(Y_t, Y_{t-2}) = \varphi_1^2$ δρα έμμεσα, αφού υπολογίζεται η συσχέτιση Y_t, Y_{t-1} και έπειτα την συσχέτιση Y_{t-1}, Y_{t-2}, από το οποίο συνεπάγεται ότι και η Y_t σχετίζεται με την Y_{t-2}.

Η διαφορά με την συνάρτηση μερικής αυτοσυσχέτισης (PACF) είναι ότι η τελευταία δίνει μόνο την απευθείας, την άμεση δηλαδή συσχέτιση, μεταξύ Y_t και Y_{t-h} (Cowpertwait & Metcalfe, 2009).

Συνεπώς, εάν $\{X_t\}$ είναι μια στάσιμη διαδικασία, η συνάρτηση μερικής αυτοσυσχέτισης σε υστέρηση h, για $h \geq 2$, ορίζεται ως η άμεση συσχέτιση μεταξύ X_t και X_{t-h} με την γραμμική εξάρτηση μεταξύ των ενδιάμεσων μεταβλητών X_s με $t-h < s < t$ να έχει αφαιρεθεί. Συμβολίζουμε με $a(h)$ και θεωρούμε ότι :

$$a(0) = 1$$
$$a(1) = \rho(1)$$
$$a(h) = a(-h), \text{ για } h<0 \text{ και άρα είναι άρτια συνάρτηση.}$$

Παρακάτω θα προσπαθήσουμε να υπολογίσουμε τις συναρτήσεις αυτοσυσχέτισης και μερικής αυτοσυσχέτισης για τα μοντέλα που αναλύσαμε έτσι ώστε όταν μας δίνεται ένα σύνολο δεδομένων να μπορούμε με βάση αυτές να κατανοήσουμε πιο μοντέλο ενδείκνυται για να τα αναπαραστήσουμε.

2.11.1 Υπολογισμός συναρτήσεων αυτοσυσχέτισης (ACF) και μερικής αυτοσυσχέτισης (PACF) για το μοντέλο κινητού μέσου όρου MA

Υπενθυμίζουμε ότι η εξίσωση που περιγράφει το στάσιμο μοντέλο MA, τάξης p, δίνεται από τη σχέση:

$$X_t = \mu + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}\,, \text{ όπου } \{e_t\} \sim WN\left(0, \sigma_e^2\right)$$

Όπως αποδείξαμε στην ενότητα 2.6, η μέση τιμή, η διασπορά και η αυτοσυνδιακύμανση μιας τέτοιας διαδικασίας $\{X_t\}$ δίνονται κατ' αντιστοιχία από τις εξής σχέσεις:

$$E\left(X_t\right) = \mu$$

$$Var\left(X_t\right) = \gamma\left(0\right) = \sigma_e^2\left(1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2\right)$$

$$\gamma\left(h\right) = \begin{cases} \sigma_e^2 \sum_{i=0}^{q-|h|} \theta_i \theta_{i+|h|} & , \text{ αν } |\mathrm{h}| \le q \\ 0 & , \text{ αν } |\mathrm{h}| > \mathrm{q} \end{cases}, \quad \text{όπου } \theta_0 = 1$$

Εφόσον έχουμε βρει την συνάρτηση αυτοσυνδιακύμανσης, σε υστέρηση h, διαιρώντας την με τον όρο $\gamma\left(0\right)$, μπορούμε εύκολα να βρούμε την συνάρτηση αυτοσυσχέτισης. Συνεπώς,

$$\rho\left(h\right) = \frac{\gamma\left(h\right)}{\gamma\left(0\right)} = \begin{cases} \dfrac{\sigma_e^2 \sum_{i=0}^{q-|h|} \theta_i \theta_{i+|h|}}{\sigma_e^2 \sum_{i=0}^{q} \theta_i^2} & , \text{ αν } |\mathrm{h}| \le q \\ 0 & , \text{ αν } |\mathrm{h}| > \mathrm{q} \end{cases}, \quad \text{όπου } \theta_0 = 1$$

Τελικά, η συνάρτηση αυτοσυσχέτισης (ACF) μιας διαδικασίας $\{X_t\}$ που ικανοποιεί το μοντέλο $MA(q)$ δίνεται από τον τύπο:

$$\rho\left(h\right) = \begin{cases} \dfrac{\theta_h + \theta_1\theta_{1+|h|} + \theta_2\theta_{2+|h|} + \ldots + \theta_q\theta_{q+|h|}}{1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2} & , \text{ αν } |\mathrm{h}| \le q \\ 0 & , \text{ αν } |\mathrm{h}| > \mathrm{q} \end{cases}$$

Παρατηρούμε ότι οι τιμές που παίρνει η ACF αποκόπτονται μετά την υστέρηση q. Αυτό το χαρακτηριστικό της ACF είναι πολύ χρήσιμο για να προσδιορίσουμε την τάξη του μοντέλου MA, μετά από την οποία οι τιμές της ACF αποκόπτονται.

Ωστόσο, στις εφαρμογές δεδομένων στην πραγματική ζωή, η δειγματική συνάρτησης αυτοσυσχέτισης r_h, δεν είναι απαραίτητα ίση με το μηδέν έπειτα από την υστέρηση q. Αναμένουμε όμως οι τιμές της να γίνονται πολύ μικρές, κατά απόλυτη τιμή, μετά την υστέρηση q. Για ένα σύνολο δεδομένων n παρατηρήσεων αυτό συχνά τεστάρεται ανάμεσα στα όρια $\pm 1.96/\sqrt{n}$, όπου $1/\sqrt{n}$ είναι περίπου η τιμή της τυπικής απόκλισης της ACF, για οποιαδήποτε υστέρηση, κάτω από την υπόθεση της ανεξαρτησίας όπως αναφέραμε στην ενότητα 2.11. Εδώ να σημειώσουμε ότι μια πιο ακριβής εκτίμηση για το τυπικό σφάλμα της δειγματικής συνάρτησης αυτοσυσχέτισης r_h, παρέχεται από τον (Bartlett, 1946) με βάση τον τύπο:

$$S\left(r_h\right) = n^{-1/2}\left(1 + 2\sum_{j=1}^{h-1} r_j^{*2}\right)^{1/2}, \quad \text{όπου } \mathrm{r}_j^* = \begin{cases} r_j, \text{ για } \rho_j \neq 0 \\ 0, \text{ για } \rho_j = 0 \end{cases}$$

Ειδική περίπτωση του παραπάνω τύπου είναι τα δεδομένα λευκού θορύβου, όπως αυτά του Διαγράμματος 2.11.1 για τα οποία ισχύει ότι $\rho_j = 0$ για όλα τα j. Έτσι, για τον λευκό θόρυβο, για τα οποία δεν υπάρχει αυτοσυσχέτιση, ένα λογικό διάστημα μέσα στο οποίο θα πρέπει να πέφτουν οι τιμές της δειγματικής συνάρτηση αυτοσυσχέτισης θα έπρεπε να είναι το $\pm 1.96/\sqrt{n}$. Οποιαδήποτε άλλη ένδειξη θα μπορούσε να θεωρηθεί ως απόδειξη για σειριακή εξάρτηση της διαδικασίας.

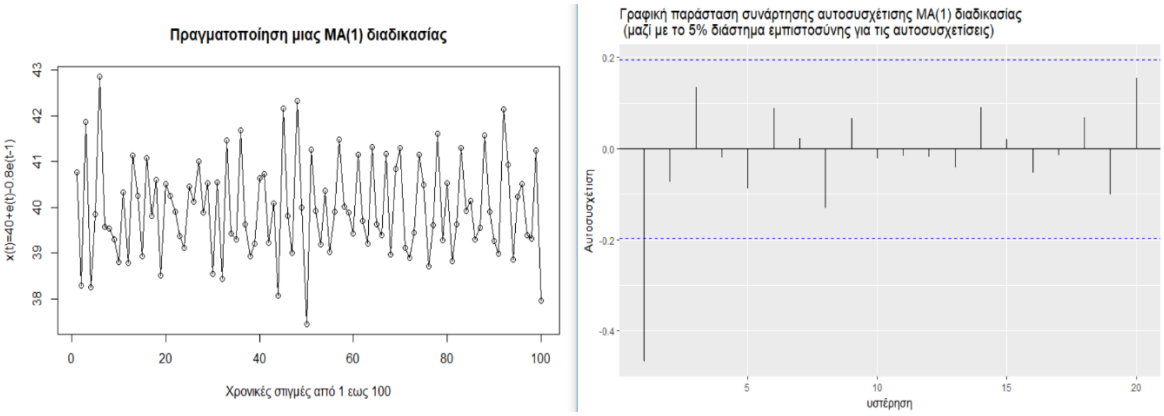
Ο υπολογισμός της συνάρτησης PACF για το $MA(q)$ μοντέλο υπολογίζεται λύνοντας τις εξισώσεις Yule-Walker, που ορίζουμε στην αμέσως επόμενη ενότητα 2.11.2, παρουσιάζει μια αρκετά περίπλοκη μορφή και για τον σκοπό της συγκεκριμένης διπλωματικής θα αναφέρουμε μόνο ότι είτε ακολουθεί μια φθίνουσα εκθετική πορεία στο μηδέν είτε έχει την μορφή φθίνουσας ημιτονοειδούς συνάρτησης που τείνει στο μηδέν. Ως εκ τούτου, για την εύρεση κατάλληλου μοντέλου για τα δεδομένα μας, συνίσταται να χρησιμοποιούνται ταυτοχρόνως και η δειγματική ACF και η δειγματική PACF.

Το Διάγραμμα 2.11.4 δείχνει μια πραγματοποίηση 100 τιμών του $MA(1)$ μοντέλου της μορφής:

$$x_t = 40 + e_t - 0.8e_{t-1}$$

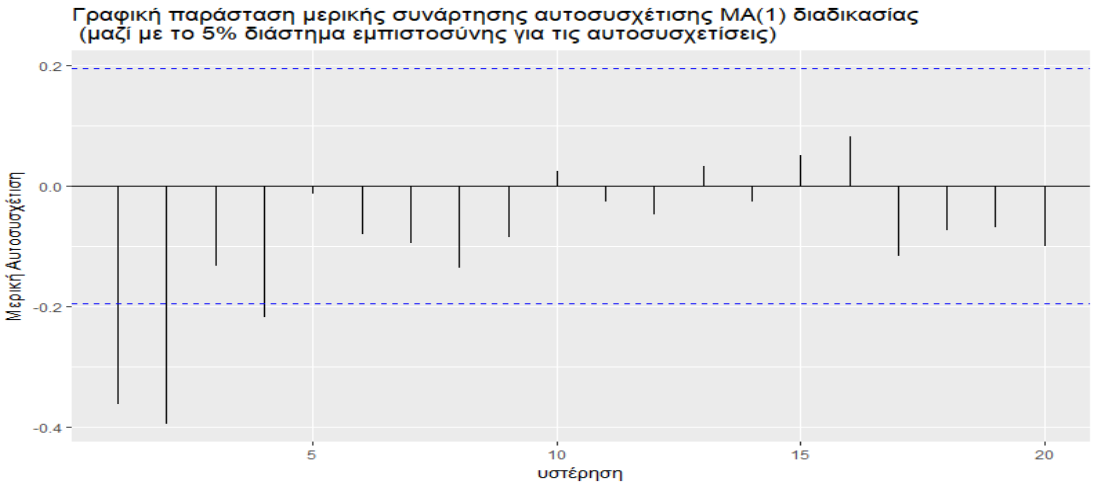
μαζί με τις τιμές της δειγματικής συνάρτησης αυτοσυσχέτισης. Παρατηρώντας το γράφημα βλέπουμε ότι η μέση τιμή και η διασπορά των παρατηρήσεων παραμένουν σταθερές με την πάροδο του χρόνου. Ακόμη, είναι εμφανές ότι οι τιμές $\{x_t\}_{t=1}^{100}$ ταλαντεύονται διαδοχικά πάνω και κάτω από τη μέση τιμή 40, προϊδεάζοντάς μας για μια αρνητική αυτοσυσχέτιση στα δεδομένα μας, γεγονός που επιβεβαιώνεται και

από το γράφημα ACF του ίδιου διαγράμματος με την αρνητική συσχέτιση στην υστέρηση $q=1$.



<u>Διάγραμμα 2.11.1.1</u>: Μια πραγματοποίηση της MA(1) διαδικασίας : $x_t=40+e_t-0.8e_{t-1}$ μαζί με τις τιμές της δειγματικής συνάρτησης αυτοσυσχέτισης.

Η δειγματική μερική συνάρτηση αυτοσυσχέτισης, PACF, του παραπάνω μοντέλου φαίνεται στο Διάγραμμα 2.11.1.2. Παρουσιάζει εκθετική πτώση που τείνει στο μηδέν, δηλαδή όλες οι τιμές πέφτουν μέσα στις μπλε διακεκομμένες μετά την υστέρηση 4.

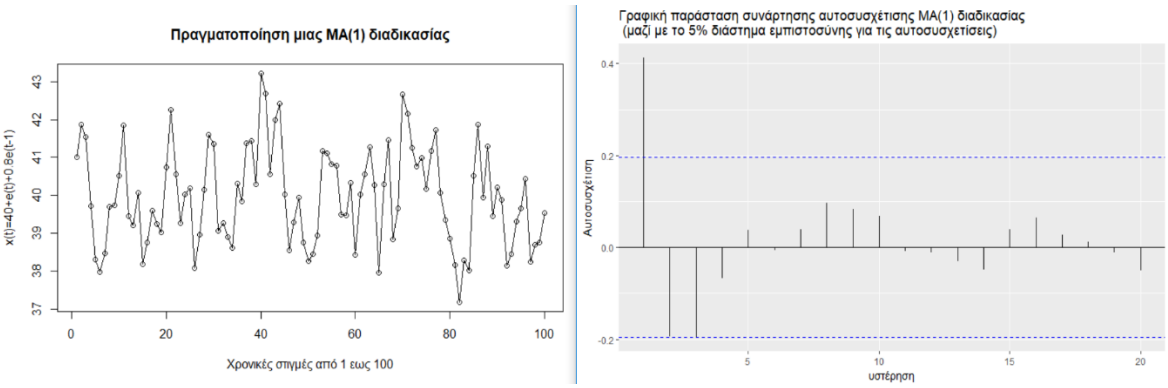


<u>Διάγραμμα 2.11.1.2</u>: Οι τιμές της δειγματικής συνάρτησης μερικής αυτοσυσχέτισης της MA(1) διαδικασίας : $x_t=40+e_t-0.8e_{t-1}$.

Αντίστοιχα, το Διάγραμμα 2.11.1.3 δείχνει μια πραγματοποίηση 100 τιμών του $MA(1)$ μοντέλου της μορφής:

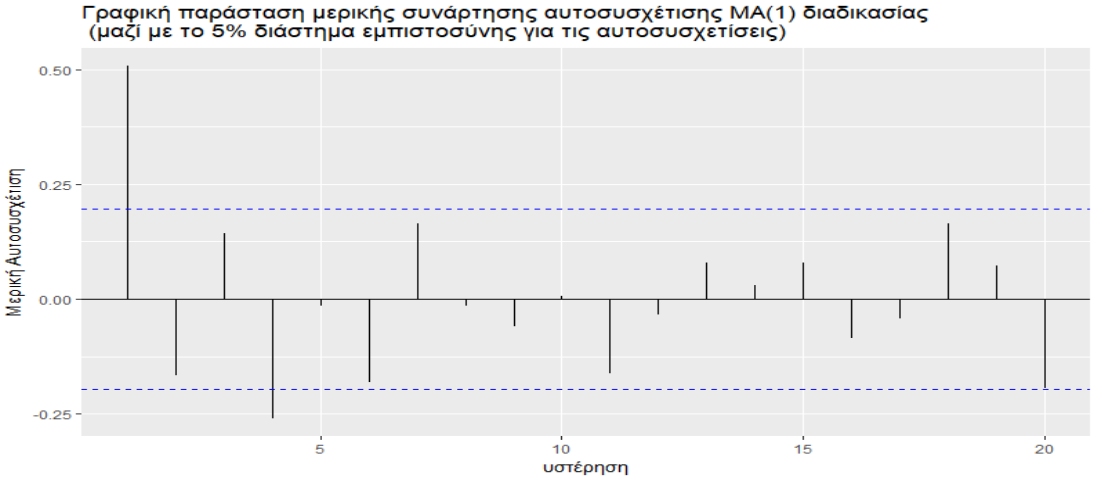
$$x_t = 40 + e_t + 0.8e_{t-1}$$

μαζί με τις τιμές της δειγματικής συνάρτησης αυτοσυσχέτισης. Εδώ παρατηρούμε τις διαδοχικές παρατηρήσεις $\{x_t\}_{t=1}^{100}$ να είναι πολύ κοντά η μια με την άλλη, ακολουθώντας την ίδια πορεία, δηλαδή αν η x_{t-1} είναι πάνω από τη μέση τιμή 40 (η αντίστοιχα κάτω από τη μέση τιμή 40) τότε και η x_t είναι πάνω από 40 (ή αντίστοιχα κάτω), τις περισσότερες φορές. Το γεγονός αυτό μας προϊδεάζει για μια θετική αυτοσυσχέτιση στα δεδομένα μας, πράγμα που επιβεβαιώνεται και από το γράφημα ACF του ίδιου διαγράμματος με την θετική συσχέτιση στην υστέρηση $q=1$.



<u>Διάγραμμα 2.11.1.3</u>: Μια πραγματοποίηση της MA(1) διαδικασίας : $x_t=40+e_t+0.8e_{t-1}$ μαζί με τις τιμές της δειγματικής συνάρτησης αυτοσυσχέτισης.

Η δειγματική μερική συνάρτηση αυτοσυσχέτισης, PACF, του παραπάνω μοντέλου φαίνεται στο Διάγραμμα 2.11.1.4. Εδώ φαίνεται ένα μοτίβο εκθετικής πτώσης, κατά απόλυτη τιμή που τείνει στο μηδέν.



<u>Διάγραμμα 2.11.1.4</u>: Οι τιμές της δειγματικής συνάρτησης μερικής αυτοσυσχέτισης της MA(1) διαδικασίας : $x_t=40+e_t+0.8e_{t-1}$.

2.11.2 Υπολογισμός συναρτήσεων αυτοσυσχέτισης (ACF) και μερικής αυτοσυσχέτισης (PACF) για το αυτοπαλίνδρομικό μοντέλο AR

Από την ενότητα 2.7, είδαμε ότι το γενικό μοντέλο $AR(p)$ δίνεται από την εξίσωση:

$$X_t = \mu + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + e_t,$$ όπου e_t είναι λευκός θόρυβος.

Η παραπάνω εξίσωση είδαμε ότι γράφεται στην σύντομη μορφή:

$$\varphi(B) X_t = \mu + e_t, \quad \text{όπου} \quad \varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p$$

Η συνθήκη στασιμότητας ικανοποιείται όταν οι ρίζες του AR πολυωνύμου, $\varphi(B)$, βρίσκονται εκτός του μοναδιαίου κύκλου. Ισοδύναμη συνθήκη στασιμότητας είναι οι ρίζες του πολυωνύμου

$$\lambda^p - \varphi_1 \lambda^{p-1} - \ldots - \varphi_{p-1}\lambda - \varphi_p = 0 \quad (2.11.2.1)$$

να βρίσκονται εντός του μοναδιαίου κύκλου, δηλαδή να είναι μικρότερες του ένα κατά απόλυτη τιμή. Κάτω από αυτή την συνθήκη, μια χρονοσειρά $\{X_t\}$ που ικανοποιεί την εξίσωση του $AR(p)$ μοντέλου θεωρείται επίσης ότι έχει την μορφή άπειρου απόλυτου αθροίσματος μιας MA διαδικασίας. Δηλαδή,

$$X_t = \delta + \Psi(B) e_t = \delta + \sum_{i=0}^{\infty} \psi_i e_{t-i} \quad (2.11.2.2)$$

Το $\Psi(B)$ λαμβάνεται εάν αντιστρέψουμε το $\varphi(B)$, δηλαδή $\Psi(B) = \varphi^{-1}(B)$. Επίσης ισχύει $\sum_{i=0}^{\infty} |\psi_i| < \infty$ λόγω συνθήκης στασιμότητας.

Τα βάρη των τυχαίων όρων e_t της εξίσωσης (2.11.2.2) μπορούν να υπολογιστούν από τον τύπο : $\varphi(B)\Psi(B) = 1$ ως εξής:

$$\psi_j = 0, \quad j < 0$$
$$\psi_0 = 1$$
$$\psi_j - \varphi_1 \psi_{j-1} - \varphi_2 \psi_{j-2} - \ldots - \varphi_p \psi_{j-p} = 0 \text{ για } j = 1, 2, \ldots$$

Αποδεικνύεται ότι η αναμενόμενη τιμή της $\{X_t\}$ διαδικασίας που ικανοποιεί την εξίσωση (2.11.2.2) είναι ίση με:

$$\mathrm{E}(X_t) = \delta = \frac{\mu}{1 - \varphi_1 - \varphi_2 - \ldots - \varphi_p}$$

Αντίστοιχα η αυτοσυνδιακύμανση υπολογίζεται να είναι:

$$\gamma(h) = Cov(X_t, X_{t-h}) = Cov\left(\mu + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \ldots + \varphi_p X_{t-p} + e_t, X_{t-h}\right)$$
$$= \sum_{i=1}^{p} \varphi_i Cov(X_{t-i}, X_{t-h}) + Cov(e_t, X_{t-h})$$
$$= \sum_{i=1}^{p} \varphi_i \gamma_X(h-i) + \begin{cases} \sigma_e^2, & \text{αν h=0} \\ 0, & \text{αν h>0} \end{cases} \qquad (2.11.2.3)$$

Αντικαθιστώντας στον παραπάνω τύπο $h=0$ παίρνουμε τη διασπορά η οποία ισούται με:

$$\gamma(0) = \sum_{i=1}^{p} \varphi_i \gamma_X(i) + \sigma_e^2$$
$$\Rightarrow \gamma(0)\left[1 - \sum_{i=1}^{p} \varphi_i \rho_X(i)\right] = \sigma_e^2$$

Διαιρώντας την (2.11.2.3) με $\gamma(0)$ για $h>0$ παίρνουμε την συνάρτηση αυτοσυσχέτισης ACF της $AR(p)$ διαδικασίας, η οποία ικανοποιεί τις παρακάτω εξισώσεις που είναι γνωστές με την ονομασία Yule-Walker :

$$\rho(h) = \sum_{i=1}^{p} \varphi_i Cor(X_{t-i}, X_{t-h})$$
$$\Rightarrow \rho(h) = \sum_{i=1}^{p} \varphi_i \rho(h-i)$$
$$\Rightarrow \rho(h) = \varphi_1 \rho(h-1) + \varphi_2 \rho(h-2) + \ldots + \varphi_p \rho(h-p), \quad \text{h=1,2,...} \qquad (2.11.2.4)$$

Οι εξισώσεις της σχέσης (2.11.2.4) είναι p τάξης γραμμικές εξισώσεις διαφορών που σημαίνει ότι η ACF μιας $AR(p)$ διαδικασίας μπορεί να βρεθεί από τις p ρίζες του πολυωνύμου (2.11.2.1). Για παράδειγμα εάν οι ρίζες του $\lambda^p - \varphi_1 \lambda^{p-1} - \ldots - \varphi_{p-1}\lambda - \varphi_p = 0$ είναι όλες διακριτές και πραγματικές έχουμε:

$\rho(h) = c_1 \lambda_1^h + c_2 \lambda_2^h + \ldots + c_p \lambda_p^h$, h=1,2,... , όπου $c_1, c_2, \ldots, c_p$ είναι συγκεκριμένες σταθερές.

Ωστόσο, γενικά οι ρίζες του πολυωνύμου (2.11.2.1) δεν είναι πάντα όλες διακριτές και πραγματικές. Έτσι, η ACF μιας $AR(p)$ διαδικασίας μπορεί να είναι μια μίξη εκθετικής πτώσης στο μηδέν και φθίνουσας ημιτονοειδής συνάρτησης, ανάλογα με τις ρίζες της εξίσωσης (2.11.2.1).

Για τον υπολογισμό της μερικής συνάρτησης αυτοσυσχέτισης PACF θα χρησιμοποιήσουμε τις Yule-Walker εξισώσεις όπως δίνονται από την σχέση (2.11.2.4).

Συνεπώς, για οποιαδήποτε τιμή της h, θεωρούμε:

$$\rho(j) = \sum_{i=1}^{h} \varphi_{ih} Cor\left(X_{t-i}, X_{t-j}\right)$$
$$= \sum_{i=1}^{h} \varphi_{ih} \rho(j-i), \qquad j = 1,2,...,h \qquad (2.11.2.5)$$

Πιο αναλυτικά μπορούμε να γράψουμε:

$$\text{Για } j=1 \ : \ \rho(1) = \varphi_{1h} + \varphi_{2h}\rho(1) + ... + \varphi_{hh}\rho(h-1)$$

$$\text{Για } j=2 \ : \ \rho(2) = \varphi_{1h}\rho(1) + \varphi_{2h} + ... + \varphi_{hh}\rho(h-2)$$

$$\vdots$$

$$\text{Για } j=p \ : \ \rho(h) = \varphi_{1h}\rho(h-1) + \varphi_{2h}\rho(h-2) + ... + \varphi_{hh}$$

Οι εξισώσεις (2.11.5) με τη μορφή πίνακα μπορούν να γραφούν :

$$\begin{bmatrix} 1 & \rho(1) & \rho(2) & ... & \rho(h-1) \\ \rho(1) & 1 & \rho(3) & ... & \rho(h-2) \\ \rho(2) & \rho(1) & 1 & ... & \rho(h-3) \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \rho(h-1) & \rho(h-2) & \rho(h-3) & ... & 1 \end{bmatrix} \begin{bmatrix} \varphi_{1h} \\ \varphi_{2h} \\ \varphi_{3h} \\ . \\ . \\ . \\ \varphi_{hh} \end{bmatrix} = \begin{bmatrix} \rho(1) \\ \rho(2) \\ \rho(3) \\ . \\ . \\ . \\ \rho(h) \end{bmatrix}$$

$$\Rightarrow \mathrm{P}_h \varphi_h = \rho_h ,$$

όπου,

$$\mathrm{P}_h = \begin{bmatrix} 1 & \rho(1) & \rho(2) & ... & \rho(h-1) \\ \rho(1) & 1 & \rho(3) & ... & \rho(h-2) \\ \rho(2) & \rho(1) & 1 & ... & \rho(h-3) \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \rho(h-1) & \rho(h-2) & \rho(h-3) & ... & 1 \end{bmatrix}, \ \varphi_h = \begin{bmatrix} \varphi_{1h} \\ \varphi_{2h} \\ \varphi_{3h} \\ . \\ . \\ . \\ \varphi_{hh} \end{bmatrix} \text{ και } \rho_h = \begin{bmatrix} \rho(1) \\ \rho(2) \\ \rho(3) \\ . \\ . \\ . \\ \rho(h) \end{bmatrix}$$

Έτσι, προκειμένου να λύσουμε ως προς φ_h, έχουμε :

$$\varphi_h = \mathrm{P}_h^{-1} \rho_h \qquad (2.11.2.6)$$

Για οποιοδήποτε h, $h=1,2,...,$ ο τελευταίος συντελεστής φ_{hh} είναι η μερική αυτοσυσχέτιση της διαδικασίας σε υστέρηση h.

Για παράδειγμα για τις διάφορες τιμές της h έχουμε :

$$\text{Για } h=1: \ \varphi_{11}=\rho_1$$

$$\text{Για } h=2: \ \varphi_{22}=\frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}}=\frac{\rho_2-\rho_1^2}{1-\rho_1^2}$$

$$\text{Για } h=3: \ \varphi_{33}=\frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}}$$

. . .

$$\text{Για } h: \quad \varphi_{hh}=\frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{h-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{h-3} & \rho_2 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \rho_{h-1} & \rho_{h-2} & \rho_{h-3} & \cdots & \rho_1 & \rho_h \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{h-2} & \rho_{h-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{h-3} & \rho_{h-2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \rho_{h-1} & \rho_{h-2} & \rho_{h-3} & \cdots & \rho_1 & 1 \end{vmatrix}}$$

Εδώ πρέπει να πούμε ότι για την $AR(p)$ διαδικασία ισχύει $\varphi_{hh}=0$ για $h>p$. Συνεπώς, η PACF τείνει στο μηδέν μετά την υστέρηση p για την $AR(p)$ διαδικασία.

Συμπεραίνουμε λοιπόν ότι η PACF μπορεί να χρησιμοποιηθεί για την εύρεση της τάξης μιας AR διαδικασίας με τον ίδιο τρόπο που η ACF χρησιμοποιείται για την εύρεση της τάξης μιας MA διαδικασίας.

Η δειγματική εκτίμηση της PACF, $\hat{\varphi}_{hh}$, υπολογίζεται χρησιμοποιώντας την δειγματική ACF, r_h. Δηλαδή, αντικαθιστούμε στις εξισώσεις Yule-Walker τις τιμές των αυτοσυσχετίσεων ρ_j με τις αντίστοιχες εκτιμήσεις τους r_j και έπειτα λύνουμε το σύστημα για $h=1,2,...$.

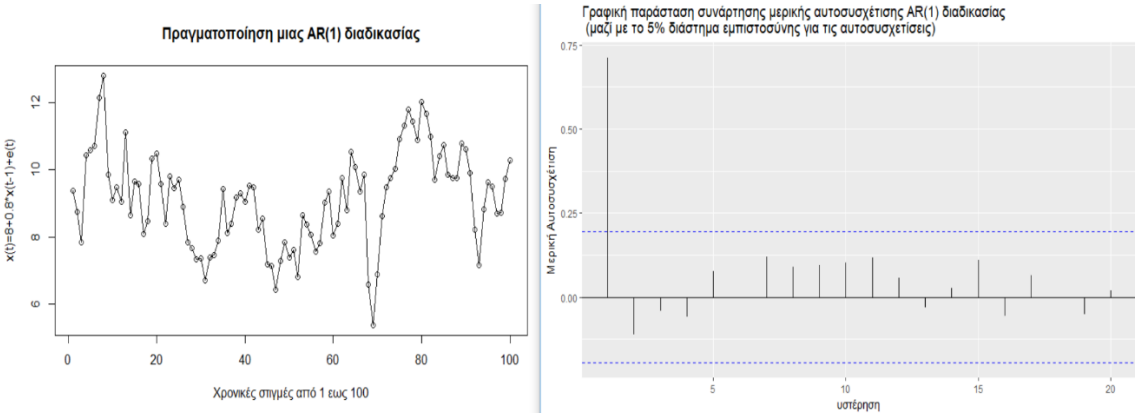
Επιπροσθέτως, για ένα δείγμα n παρατηρήσεων από μια $AR(p)$ διαδικασία, η δειγματική PACF $\hat{\varphi}_{hh}$ για $h > p$ θεωρείται ότι είναι κατά προσέγγιση κανονικά κατανεμημένη με μέση τιμή $E\left(\hat{\varphi}_{hh}\right) \approx 0$ και διασπορά $Var\left(\hat{\varphi}_{hh}\right) \approx \frac{1}{n}$ (Shumway & Stoffer, 2011, p. 121) .

Συνεπώς, το 95% διάστημα εμπιστοσύνης για να κρίνουμε εάν η $\hat{\varphi}_{hh}$ είναι στατιστικά σημαντική, δηλαδή διάφορη του μηδενός, δίνεται και σε αυτή την περίπτωση από τα όρια $\pm 1.96 / \sqrt{n}$.

Το Διάγραμμα 2.11.2.8 δείχνει μια πραγματοποίηση 100 τιμών του $AR(1)$ μοντέλου της μορφής:

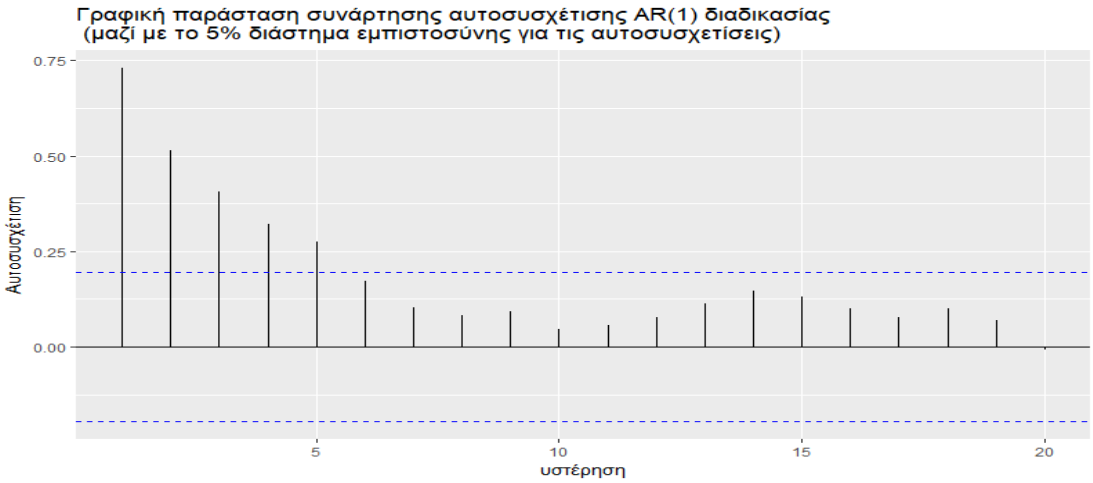
$$x_t = 8 + 0.8x_{t-1} + e_t$$

μαζί με τις τιμές της δειγματικής συνάρτησης μερικής αυτοσυσχέτισης. Παρατηρώντας το γράφημα βλέπουμε ότι η μέση τιμή και η διασπορά των παρατηρήσεων παραμένουν σταθερές με την πάροδο του χρόνου ενώ η PACF τείνει στο μηδέν (πέφτει μέσα στις μπλε διακεκομμένες) μετά την υστέρηση 1. Συνεπώς, η μόνη στατιστικά σημαντική τιμή της PACF είναι αυτή στην υστέρηση 1, επιβεβαιώνοντας ότι το $AR(1)$ είναι αυτό που ταιριάζει στα δεδομένα μας.



<u>Διάγραμμα 2.11.2.8</u>: Μια πραγματοποίηση της AR(1) *διαδικασίας : x_t=8+0.8x_t +e_t μαζί με τις τιμές της δειγματικής συνάρτησης μερικής αυτοσυσχέτισης.*

Η δειγματική συνάρτηση αυτοσυσχέτισης, ACF, του παραπάνω μοντέλου φαίνεται στο Διάγραμμα 2.11.2.9. Από τη σχέση (2.11.2.1) έχουμε δείξει ότι η συνάρτηση αυτοσυσχέτισης ενός στάσιμου $AR(1)$ μοντέλου είναι $\rho(h) = \varphi_1^h$, $h = 1, 2, \ldots$, δηλαδή έχει ένα μοτίβο εκθετικής πτώσης. Αυτό φαίνεται και από το παρακάτω διάγραμμα που η ACF φθίνει εκθετικά στο μηδέν.



<u>Διάγραμμα 2.11.2.9</u>: Οι τιμές της δειγματικής συνάρτησης μερικής αυτοσυσχέτισης της της AR(1) *διαδικασίας :* $x_t=8+0.8x_t+e_t$.

2.11.3 Υπολογισμός συναρτήσεων αυτοσυσχέτισης (ACF) και μερικής αυτοσυσχέτισης (PACF) για το αυτοπαλινδρομικό μοντέλο κινητού μέσου όρου ARMA

Τώρα θα μελετήσουμε πως υπολογίζονται οι συναρτήσεις ACF και PACF σε μοντέλα ARMA. Οι δύο αυτές συναρτήσεις υπολογίζονται εφόσον έχουμε προσδιορίσει την συνάρτηση αυτοσυνδιακύμανσης $\gamma(\cdot)$ μιας ARMA διαδικασίας $\{Y_t\}$ την οποία θα υποθέσουμε ότι είναι αιτιώδης και έπειτα διαιρώντας με $\gamma(0)$ παίρνουμε τη συνάρτηση αυτοσυσχέτισης. Αντίστοιχα, η συνάρτηση μερικής αυτοσυσχέτισης PACF υπολογίζεται και εκείνη μέσω της $\gamma(\cdot)$.

Αρχικά λοιπόν καθορίζουμε την συνάρτηση αυτοσυνδιακύμανσης $\gamma(\cdot)$ μιας αιτιώδους $ARMA(p,q)$ διαδικασίας που ορίζεται από τη σχέση :

$$\varphi(B)Y_t=\theta(B)e_t, \qquad \{e_t\}\sim WN(0,\sigma_e^2)$$

Με την υπόθεση της αιτιότητας συνεπάγεται ότι

$$Y_t=\sum_{j=0}^{\infty}\psi_j e_{t-j}\,, \qquad (2.11.3.1)$$

όπου $\sum_{j=0}^{\infty}\psi_j\lambda^j=\dfrac{\theta(\lambda)}{\varphi(\lambda)}$, $|\lambda|\le 1$ και τα $\{\psi_j\}$ υπολογίζονται όπως τα υπολογίσαμε στην ενότητα 2.8.

Αν πολλαπλασιάσουμε και τα δυο μέλη της $ARMA(p,q)$ διαδικασίας που περιγράφεται από την εξίσωση (2.8.1) :

$$Y_t - \varphi_1 Y_{t-1} - \ldots - \varphi_p Y_{t-p} = e_t + \theta_1 e_{t-1} + \ldots + \theta_p e_{t-p},$$

με το Y_{t-h}, $h = 0,1,2,\ldots,$ και πάρουμε τις αναμενόμενες τιμές δεξιά και αριστερά καταλήγουμε στην εξίσωση:

$$\gamma(h) - \varphi_1 \gamma(h-1) - \ldots - \varphi_p \gamma(h-p) = \sigma_e^2 \sum_{j=0}^{\infty} \theta_{h+j} \psi_j, \quad 0 \le h \le m, \qquad (2.11.3.2)$$

και

$$\gamma(h) - \varphi_1 \gamma(h-1) - \ldots - \varphi_p \gamma(h-p) = 0, \quad h \ge m, \qquad (2.11.3.3)$$

όπου $m = \max(p, q+1)$, $\psi_j = 0$ για $j < 0$, $\theta_0 = 1$ και $\theta_j = 0$ για $j \notin \{0,\ldots,q\}$.

Υπενθυμίζουμε την ιδιότητα της συνάρτησης συνδιακύμανσης $Cov(Y_t, Y_{t+h}) = E(Y_{t+h}Y_t) - E(Y_{t+h})E(Y_t)$ και από το γεγονός ότι έχουμε υποθέσει για την ARMA διαδικασία $\{Y_t\}$ ότι είναι μηδενικής μέσης τιμής και στάσιμη βγαίνει ότι $E(Y_{t+h}) = E(Y_t) = 0$ και συνεπώς η συνάρτηση αυτοσυνδιακύμανσης $\gamma(\cdot)$ ισούται με τη μέση τιμή του γινομένου $E(Y_{t+h}Y_t)$. Για τον υπολογισμό του δεξί μέλους της σχέσης (2.11.3.2) έχουμε αντικαταστήσει το $Y_{t+h} = \sum_{j=0}^{\infty} \psi_j e_{t+h-j}$ από τη σχέση (2.11.3.1) όπου χρειάζεται και έχουμε κάνει χρήση της ιδιότητας $\sigma_e^2 = Var(e_t) = E(e_t^2) - \overbrace{E(e_t)^2}^{=0}$. Η εξίσωση (2.11.3.3) είναι ένα σύνολο από γραμμικές εξισώσεις διαφορών με σταθερούς συντελεστές, για την οποία είναι γνωστό ότι η γενική λύση είναι της μορφής:

$$\gamma(h) = a_1 \xi_1^{-h} + a_2 \xi_2^{-h} + \ldots + a_p \xi_p^{-h}, \quad h \ge m - p, \quad (2.11.3.4)$$

όπου $\xi_1, \ldots \xi_p$ είναι οι ρίζες (θεωρούνται ότι είναι διαφορετικές μεταξύ τους) της εξίσωσης $\varphi(\lambda) = 0$ και $\alpha_1, \alpha_2, \ldots, \alpha_p$ είναι αυθαίρετες σταθερές. Φυσικά ψάχνουμε για μια λύση της (2.11.3.3) που επίσης ικανοποιεί και την (2.11.3.2). Συνεπώς, αντικαθιστούμε τη λύση (2.11.3.4) στην (2.11.3.2) και παίρνουμε ένα σύνολο από m γραμμικές εξισώσεις που καθορίζουν τις μοναδικές σταθερές $\alpha_1, \alpha_2, \ldots, \alpha_p$ και τις $m - p$ το πλήθος αυτοσυνδιακυμάνσεις $\gamma(h)$, $0 \le h < m - p$.

Για παράδειγμα στην αιτιώδη $ARMA(1,1)$ διαδικασία της μορφής :

$Y_t - \varphi_1 Y_{t-1} = e_t + \theta_1 e_{t-1}$, $\{e_t\} \sim WN(0, \sigma_e^2)$ έχουμε $p = q = 1$ άρα $m = \max(1,2) = 2$.

Από την εξίσωση (2.11.3.2) παίρνουμε:

$$\overset{\gamma(-1)=\gamma(1)}{\Rightarrow} \quad \gamma(0)-\varphi_1\gamma(1)=\sigma_e^2\left(\theta_0\psi_0+\theta_1\psi_1+\theta_2\psi_2+\ldots\right), \text{ για h=0} \quad (2.11.3.5)$$

και

$$\gamma(1)-\varphi_1\gamma(0)=\sigma_e^2\left(\theta_1\psi_0+\theta_2\psi_1+\theta_3\psi_2+\ldots\right), \quad \text{για h=1} \quad (2.11.3.6)$$

$$\overset{2.11.3.5}{\Rightarrow} \gamma(0)-\varphi_1\gamma(1)=\sigma_e^2\left(\overbrace{\theta_0}^{=1}\psi_0+\theta_1\psi_1+\overbrace{\theta_2\psi_2+\ldots}^{\theta_j=0,\ \text{για } j\notin\{0,1\}}\right)$$

$$\Rightarrow \gamma(0)-\varphi_1\gamma(1)=\sigma_e^2\left(\psi_0+\theta_1\psi_1\right)$$

Τα βάρη ψ_0,ψ_1 υπολογίζονται από την (2.11.2.6) λόγω της υπόθεσης αιτιότητας με τον ίδιο τρόπο που τα υπολογίσαμε στην ενότητα 2.8 λύνοντας την (2.8.6). Συνεπώς, καταλήγουμε στις τιμές $\psi_0=1$ και $\psi_1=\theta_1+\varphi_1$, οπότε τελικά έχουμε :

$$\gamma(0)-\varphi_1\gamma(1)=\sigma_e^2\left(1+\theta_1\left(\theta_1+\varphi_1\right)\right) \qquad (2.11.3.7)$$

και

$$\overset{2.11.3.6}{\Rightarrow} \quad \gamma(1)-\varphi_1\gamma(0)=\sigma_e^2\left(\theta_1\overbrace{\psi_0}^{=1}+\overbrace{\theta_2\psi_1+\theta_3\psi_2+\ldots}^{=0}\right)$$

$$\Rightarrow \quad \gamma(1)-\varphi_1\gamma(0)=\sigma_e^2\theta_1 \qquad (2.11.3.8)$$

Αντίστοιχα, από την εξίσωση (2.11.8) παίρνουμε:

$$\gamma(h)-\varphi_1\gamma(h-1)=0, \quad h\geq 2$$

Η λύση της τελευταίας είναι: $\gamma(h)=\alpha\varphi_1^{\,h}, \ h\geq 1$. Αντικαθιστώντας την τιμή $\gamma(1)=\alpha\varphi_1$ από την λύση που βρήκαμε στις εξισώσεις (2.11.3.7) και (2.11.3.8) παίρνουμε δυο γραμμικές εξισώσεις με δύο αγνώστους την σταθερά α και την άγνωστη διασπορά $\gamma(0)$. Αυτές οι εξισώσεις λύνονται με πολύ εύκολο τρόπο και μας δίνουν για $|\varphi_1|<1$, $\gamma(0)=\sigma_e^2\left[1+\dfrac{\left(\theta_1+\varphi_1\right)^2}{1-\varphi_1^2}\right]$, $\gamma(1)=\sigma_e^2\left[\theta_1+\varphi_1+\dfrac{\left(\theta_1+\varphi_1\right)^2\varphi_1}{1-\varphi_1^2}\right]$.

Συνεπώς, η ACF της ARMA διαδικασίας βρίσκεται εύκολα απλά διαιρώντας με το $\gamma(0)$ αφού ισχύει $\rho(h)=\dfrac{\gamma(h)}{\gamma(0)}$. Παρομοίως, για οποιοδήποτε σύνολο παρατηρήσεων $\{y_1,y_2,\ldots,y_n\}$ η δειγματική ACF $\hat{\rho}(\cdot)$ υπολογίζεται ως: $\hat{\rho}(h)=\dfrac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$.

Η συνάρτηση μερικής αυτοσυσχέτισης PACF της ARMA διαδικασίας $\{Y_t\}$ είναι η συνάρτηση $\alpha(\cdot)$ που ορίζεται από τις εξισώσεις:

$$\alpha(0)=1$$
$$\kappa\alpha\iota$$
$$\alpha(h)=\varphi_{hh},\ \ h\geq 1,$$

όπου φ_{hh} είναι η τελευταία συνιστώσα της 2.11.2.6. Μπορούμε να αποφύγουμε τον υπολογισμό της αυτοσυσχέτισης ρ_h για τον υπολογισμό της PACF και να την υπολογίσουμε κατευθείαν προσδιορίζοντας μόνο την αυτοσυνδιακύμανση από τον τύπο : $\varphi_h=\Gamma_h^{-1}\gamma_h$, όπου $\Gamma_h=\left[\gamma(i-j)\right]_{i,j=1}^{h}$ και $\gamma_h=\left[\gamma(1),\gamma(2),...,\gamma(h),\right]'$. Όπως βλέπουμε ο παραπάνω τύπος ισούται με τον αρχικό μας τύπο της σχέσης 2.11.2.6 αν διαιρέσουμε κάθε συνιστώσα του Γ_h και γ_h με $\gamma(0)$. Έτσι, για οποιοδήποτε σύνολο παρατηρήσεων $\{y_1,y_2,...,y_n\}$ με $y_i\neq y_j$ η δειγματική PACF $\hat{a}(\cdot)$ υπολογίζεται ως:

$$\hat{\alpha}(0)=1$$
$$\kappa\alpha\iota$$
$$\hat{\alpha}(h)=\hat{\varphi}_{hh},\ \ h\geq 1,\ \text{με}\ \hat{\varphi}_{hh}\ \text{την τελευταία συνιστώσα του}\ \hat{\varphi}_h=\Gamma_h^{-1}\hat{\gamma}_h.$$

Εδώ να σημειώσουμε ότι οι συναρτήσεις ACF και PACF της διαδικασίας ARMA μπορούν και οι δύο να εμφανίζουν φθίνουσα εκθετική πορεία σε συνδυασμό ή χωρίς φθίνουσας ημιτονοειδούς συνάρτησης, γεγονός που δυσχεραίνει τον εντοπισμό της τάξης της ARMA(p,q) διαδικασίας. Για τον λόγο αυτό μια συνήθης τακτική είναι να δοκιμάσουμε διαφόρων τάξης μοντέλα και να επιλέξουμε τα καλύτερα σύμφωνα με τα κριτήρια AIC και BIC που θα αναλύσουμε παρακάτω.

2.11.4 Συγκεντρωτικός πίνακας συμπερασμάτων συναρτήσεων ACF, PACF

Μέχρι στιγμής έχουμε κατηγοριοποιήσει τα μοντέλα μας σε στάσιμα και μη στάσιμα. Το επόμενο βήμα ήταν να μελετήσουμε το είδος της εξάρτησης που συναντάται μεταξύ των παρατηρήσεων έτσι ώστε να είμαστε σε θέση να επιλέξουμε την τάξη του μοντέλου. Στον Πίνακα 2.11.4.1 συνοψίζονται οι θεωρητικές τιμές των ACF και PACF για στάσιμες διαδικασίες των μοντέλων που αναλύσαμε παραπάνω.

Πίνακας 2.11.4.1 :

Θεωρητική συμπεριφορά ACF και PACF στάσιμων διαδικασιών		
Μοντέλο	**ACF**	**PACF**
MA(q)	Οι τιμές της αποκόπτονται μετά την υστέρηση q	Φθίνει εκθετικά στο μηδέν και/ή παρουσιάζει σταδιακή απόσβεση με ή χωρίς ημιτονοειδή συμπεριφορά που τείνει στο μηδέν
AR(p)	Φθίνει εκθετικά στο μηδέν και/ή παρουσιάζει σταδιακή απόσβεση με ή χωρίς ημιτονοειδή συμπεριφορά που τείνει στο μηδέν	Οι τιμές τις αποκόπτονται μετά την υστέρηση p
ARMA(p,q)	Φθίνει εκθετικά στο μηδέν και/ή παρουσιάζει σταδιακή απόσβεση με ή χωρίς ημιτονοειδή συμπεριφορά που τείνει στο μηδέν	Φθίνει εκθετικά στο μηδέν και/ή παρουσιάζει σταδιακή απόσβεση με ή χωρίς ημιτονοειδή συμπεριφορά που τείνει στο μηδέν

2.12 Κατασκευή και εξακρίβωση γενικής μορφής μοντέλου σε προβλήματα χρονοσειρών

Σε αυτήν την ενότητα θα περιγράψουμε πέντε βασικά βήματα που χρειάζονται προκειμένου να βρούμε πιο μοντέλο ενδείκνυται να περιγράψει ικανοποιητικά τα δεδομένα μας έτσι ώστε να οδηγηθούμε σε έγκυρες προβλέψεις για το μέλλον. Το πρώτο βήμα είναι να βρούμε τη γενική μορφή του μοντέλου και να διαπιστώσουμε κατά πόσο μπορεί να προέρχεται από μια στάσιμη η όχι διαδικασία. Αυτό το βήμα περιλαμβάνει κατάλληλους ελέγχους στασιμότητας καθώς και εξέταση των δειγματικών συναρτήσεων αυτοσυσχέτισης ACF και μερικής αυτοσυσχέτισης PACF. Τα αποτελέσματα της PACF δίνουν μια εκτίμηση για την τάξη του μοντέλου AR ενώ η ACF αντίστοιχα δίνει μια εκτίμηση για την τάξη του μοντέλου MA. Επιπροσθέτως, προσδιορίζεται η βέλτιστη υστέρηση χρησιμοποιώντας τα κριτήρια AIC (Akaike Information Criterion) και BIC (Bayesian Information Criterion). Τα κριτήρια αυτά είναι ένα μέτρο καλής προσαρμογής του μοντέλου καθώς ποινικοποιούν τον υπερβολικό αριθμό παραμέτρων. Όσο μικρότερες είναι οι τιμές του κριτηρίου, τόσο καλύτερη είναι η προδιαγραφή του μοντέλου. Το δεύτερο βήμα, αφού έχουμε βρει πιο μοντέλο θα χρησιμοποιήσουμε ελέγχουμε την σημαντικότητα των εκτιμημένων συντελεστών με z-test. Αν κάποιος βγει στατιστικά ασήμαντος, δηλαδή δεχτούμε την υπόθεση ότι είναι μηδενικός, τότε αφαιρούμε αυτήν την παράμετρο μειώνοντας την τάξη του μοντέλου και ελέγχουμε αν τα κριτήρια AIC και BIC δίνουν καλύτερες τιμές. Το τρίτο βήμα εξετάζει την έκταση στην οποία τα μοντέλα είναι έγκυρα και περιλαμβάνει έλεγχο υπολοίπων (σφαλμάτων) όπως η κατανομή, η ανεξαρτησία των σφαλμάτων και αν υπάρχουν φαινόμενα ετεροσκεδαστικότητας (δηλαδή εάν η διασπορά των σφαλμάτων δεν είναι σταθερή). Το τέταρτο βήμα περιλαμβάνει την πρόβλεψη των μελλοντικών τιμών της χρονοσειράς και στο πέμπτο και τελευταίο βήμα θα γίνει η αξιολόγηση της ακρίβειας των προβλέψεων μέσω διάφορων κριτηρίων όπως το Root Mean Squared Error (RMSE) και Mean Absolute Error (MAE).

2.12.1 Ταυτοποίηση Μοντέλου (Model Identification)

Οι προσπάθειες για τον προσδιορισμό μοντέλου πρέπει να ξεκινήσουν με προκαταρκτικές προσπάθειες για την κατανόηση της διαδικασίας από την οποία τα δεδομένα προέρχονται και να ελέγξουμε κατά πόσο το δείγμα μας είναι αρκετά μεγάλο για να δώσει ικανοποιητικά αποτελέσματα. Είναι αναγκαίο το δείγμα μας να αποτελείται από περισσότερες από 50 παρατηρήσεις για να θεωρηθεί αξιόπιστη η μελέτη του και όσο περισσότερες παρατηρήσεις έχουμε τόσο πιο αξιόπιστα συμπεράσματα βγάζουμε. Επιπροσθέτως, προτού προβούμε σε οποιαδήποτε διαδικασία κατασκευής μοντέλου προτείνεται, εκτός από το διάγραμμα που απεικονίζει τα δεδομένα μας σε σχέση με την χρονική περίοδο παρατήρησης , να δημιουργήσουμε διαγράμματα διασποράς (scatter plots) για τα δεδομένα μας τα οποία δείχνουν την συναρτησιακή μορφή της προς μελέτης μεταβλητής σε προηγούμενες χρονικές περιόδους. Σε όλα τα προαναφερθέντα μοντέλα για ένα σύνολο δεδομένων $x_1, x_2, ..., x_n$ έχουμε κάνει υπόθεση περί γραμμικής σχέσης της x_t με προηγούμενες χρονικές περιόδους της $x_{t-1}, x_{t-2}, ...$ Αντίστοιχα, η αυτοσυσχέτιση και αυτοσυνδιακύμανση είναι μέτρα που ανιχνεύουν εάν υπάρχει μια τέτοια γραμμική εξάρτηση αλλά αδυνατούν να προσδιορίσουν εάν υπάρχει οποιαδήποτε άλλη εξάρτηση, όπως για παράδειγμα μια πολυωνυμική. Επομένως, εάν έχουμε υποθέσει ότι η μεταβλητή μας εξαρτάται γραμμικά μόνο από την αμέσως προηγούμενη τιμή της, προτού εφαρμόσουμε το AR(1) μοντέλο, ένα διάγραμμα με συντεταγμένες τα σημεία $(x_{t-1}, x_t),\ t = 2, ..., n$ θα φανέρωνε αμέσως την συνάρτηση που συνδέει την x_t με την x_{t-1}.

Για να αποφανθούμε εάν τα δεδομένα μας προέρχονται από στάσιμες ή όχι διαδικασίες εκτός του γραφήματος συναρτήσει του χρόνου, όπως αυτό του Διαγράμματος 1.1 που μπορεί να φανερώνει την τάση των δεδομένων μας και κατά πόσο η διασπορά τους τείνει να αυξάνει ή να μειώνει σε βάθος χρόνου, μπορούμε να κάνουμε διάφορους ελέγχους στασιμότητας. Τέτοιοι έλεγχοι πραγματοποιούνται με το **Augmented Dickey-Fuller (ADF) t-statistic test** και το **Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test,** μέσω της R.

Ο ADF έλεγχος πραγματοποιεί τις υποθέσεις:

H_0 : Υπάρχει ρίζα του AR πολυωνύμου που βρίσκεται στο σύνορο ή εντός του μοναδιαίου κύκλου, δηλαδή η χρονοσειρά δεν είναι στάσιμη.

H_1 : Όλες οι ρίζες του AR πολυωνύμου βρίσκονται εκτός του μοναδιαίου κύκλου, δηλαδή η χρονοσειρά παρουσιάζει στασιμότητα.

Ο έλεγχος αυτός, ανιχνεύει την αυτοσυσχέτιση στα δεδομένα μας, την αφαιρεί και έπειτα ελέγχει την στασιμότητα. Μικρές τιμές της p-value, δηλαδή μικρότερες του 0.05, μας οδηγούν στην απόρριψη της μηδενικής υπόθεσης για μη στασιμότητα σε επίπεδο σημαντικότητας 5%.

Ο έλεγχος KPSS (Kwiatkowski, et al., 1992) σε αντίθεση με τον έλεγχο ADF (Said & Dickey, 1984) έχει ως μηδενική υπόθεση ότι η χρονοσειρά είναι στάσιμη γύρω από μια ντετερμινιστική συνιστώσα τάσης (trend-stationaity). Αυτό σημαίνει ότι η χρονοσειρά που μελετάται έχει μια τάση (η οποία είναι συνάρτηση του χρόνου t, όχι όμως απαραίτητα γραμμική) η οποία μπορεί να αφαιρεθεί και να αφήσει μια στάσιμη διαδικασία. Δηλαδή στο KPSS test, η ύπαρξη μοναδιαίας ρίζας στο AR πολυώνυμο είναι η ενναλακτική υπόθεση. Συνεπώς, μικρές τιμές της p-value απορρίπτουν την μηδενική υπόθεση και προτείνουν ότι η σειρά δεν είναι στάσιμη, συμπεραίνοντας ότι χρειάζεται να πάρουμε διαφορές μέχρι να καταλήξουμε σε στάσιμη διαδικασία.

Οι έλεγχοι που εφαρμόζονται σε ένα σύνολο δεδομένων για ύπαρξη μοναδιαίας ρίζας δεν πρέπει να συγχέονται με του ελέγχους για ύπαρξη ντετερμινιστικής συνιστώσας τάσης. Παρόλου που μοιράζονται πολλές ιδιότητες, είναι διαφορετικοί από πολλές απόψεις. Εδώ να πούμε ότι μερικές φορές είναι πιθανό για μια χρονοσειρά να μην είναι στάσιμη όμως να μην έχει μοναδιαία ρίζα και η τάση της να είναι σταθερή και ανεξάρτητη του χρόνου t (trend stationarity). Τόσο στον ADF (έλεγχος για ύπαρξη μοναδιαίας ρίζας) έλεγχο, όσο και στον KPSS (έλεγχος ύπαρξης ντετερμινιστικής συνιστώσας τάσης) έλεγχο, η μέση τιμή των δεδομένων που μελετώνται μπορεί είτε να αυξάνεται είτε να μειώνεται με την πάροδο του χρόνου. Ωστόσο, όταν υπάρχει ένα σοκ σε κάποια χρονική περίοδο των δεδομένων μας, η μέση τιμή τους επηρεάζεται από αυτό το σοκ αλλά όταν αυτό περάσει, επανέρχεται στην αρχική μέση τιμή που είχαν τα δεδομένα πριν το σοκ. Δηλαδή, μετά την επίδραση του σοκ η μέση τιμή των δεδομένων μας συγκλίνει στην μέση τιμή που είχε πριν συμβεί το σοκ. Κατά συνέπεια, όταν ελέγχεται αν τα δεδομένα μας είναι στάσιμα γύρω από μια ντετερμινιστική τάση, σε μια τέτοια περίπτωση δείχνει ότι το σοκ ήταν μια κατάσταση, το οποίο επηρέασε μόνο παροδικά την συνολική τάση των δεδομένων μας και δεν λαμβάνεται υπόψιν σαν παράγοντας που κάνει τα δεδομένα μας μη στάσιμα. Αντίθετα, στον έλεγχο για ύπαρξη μοναδιαίας ρίζας, το σοκ έχει μόνιμο αντίκτυπο στην τάση (Nielsen, 2005).

Επιπροσθέτως, οι δύο αυτοί έλεγχοι στασιμότητας εξετάζουν την στασιμότητα υποθέτοντας ότι τα δεδομένα μας μπορούν να προσαρμοστούν σε τριών ειδών γραμμικά μοντέλα. Το πρώτο είναι ένα μοντέλο με μηδενική μέση τιμή και ύπαρξης γραμμικής συνάρτησης τάσης σε σχέση με τον χρόνο t, το δεύτερο είναι ένα μοντέλο με μέση τιμή αλλά χωρίς γραμμική τάση και το τελευταίο και με μέση τιμή και με γραμμική τάση. Οι έλεγχοι αυτοί, έχουν διαφορετικούς μαθηματικούς τύπους για την περιγραφή των παραπάνω μοντέλων και επομένως είναι καλό να εφαρμόζουμε και τους δυο για να καταλήξουμε σε ακριβέστερα συμπεράσματα. Ωστόσο, το μειονέκτημά τους είναι ότι και οι δύο αυτοί έλεγχοι βασίζονται σε πολύ συγκεκριμένα είδη στασιμότητας και δεν εξετάζουν την στασιμότητα με την ευρύτερη έννοια (Dettling, 2013, p. 13). Συνεπώς έχουν μικρή ισχύ στο να εντοπίσουν όλων των ειδών μη στασιμότητας σε δεδομένα και στην πράξη συνήθως αποτυγχάνουν σε τέτοιες περιπτώσεις. Για τον λόγο αυτό συνίσταται να

έχουμε πάντα οπτική εικόνα του γραφήματος μιας χρονοσειράς αλλά και από των γραφημάτων ACF, PACF.

Σε περίπτωση που έχουμε ενδείξεις για μη στασιμότητα, δημιουργούμε μια νέα διαδικασία με τις διαφορές πρώτης τάξης και επαναλαμβάνουμε τα παραπάνω βήματα. Εάν πάλι έχουμε υπόνοιες για μη στασιμότητα παίρνουμε δεύτερες διαφορές και συνεχίζουμε την ίδια διαδικασία μέχρι να καταλήξουμε σε στασιμότητα. Ο αριθμός που δείχνει το πλήθος των διαφορών που χρειαστήκαμε για να καταλήξουμε σε στασιμότητα μας δίνει την τιμή d του γενικού μοντέλου $ARIMA(p,d,q)$.

Μια πρωταρχική επιλογή των παραμέτρων p,q για την νέα πλέον στάσιμη διαδικασία που έχει προκύψει από τις d διαφορές μπορεί να καθοριστεί παρατηρώντας τα γραφήματα ACF, PACF. Πιο συγκεκριμένα, η παράμετρος q καθορίζεται να είναι το σημείο μετά από το οποίο οι τιμές της PACF αποκόπτονται, δηλαδή πέφτουν εντός των ορίων που ορίζουν οι μπλε διακεκομμένες. Αντίστοιχα, η παράμετρος p καθορίζεται να είναι το σημείο μετά από το οποίο οι τιμές της ACF αποκόπτονται. Σε περίπτωση που δεν έχουμε ξεκάθαρη εικόνα για τις τιμές των παραμέτρων p,q, μια μέθοδος είναι να επιλέξουμε αυθαίρετες αρχικές τιμές σχετικά μεγάλες και έπειτα συγκρίνουμε όλα τα μοντέλα που προκύπτουν επιλέγοντας μικρότερες τιμές.

Έπειτα, μόλις καταλήξουμε στις επιλογή των παραμέτρων p,q, προσαρμόζουμε το μοντέλο στα δεδομένα μας και ελέγχουμε την σημαντικότητα των εκτιμημένων συντελεστών με z-test. Εάν τα αποτελέσματά μας δείχνουν ότι κάποιος συντελεστής δεν είναι στατιστικά σημαντικός τότε τροποποιούμε το μοντέλο αφαιρώντας αυτή την παράμετρο. Δηλαδή, εάν έχουμε προσαρμόσει τα δεδομένα μας το μοντέλο $\mathrm{ARMA}(p,q)$, και βρεθεί ότι ο AR συντελεστής δεν είναι στατιστικά σημαντικός την $\chi\rho\text{ονική περίοδο t-p}$, τότε επιλέγουμε να προσαρμόσουμε το πιο οικονομικό μοντέλο $\mathrm{ARMA}(p\text{-}1,q)$. Σε αυτό το σημείο πρέπει να είμαστε προσεκτικοί και να επαληθεύσουμε τα αποτελέσματά μας συγκρίνοντας το αρχικό μοντέλο που είχαμε με όλα τα απλούστερα μοντέλα που προέκυψαν στην περίπτωση που αφαιρέσαμε τους στατιστικά ασήμαντους συντελεστές.

2.12.2 Εκτίμηση Παραμέτρων

Σε αυτήν την ενότητα θα επικεντρωθούμε στην εκτίμηση των άγνωστων παραμέτρων $\varphi=\left(\varphi_1,...,\varphi_p\right)'$, $\theta=\left(\theta_1,...,\theta_p\right)'$ και σ_e^2 για ένα δοθέν σύνολο δεδομένων $x_1,x_2,...,x_{\mathrm{N}}$ που προέρχεται από μια αρχική $\{X_t\}$ διαδικασία. Με βάση τα παραπάνω, όλα τα μοντέλα που αναλύθηκαν στις προηγούμενες ενότητες, AR, MA, $ARMA$, αποτελούν υποκατηγορίες του $ARIMA$ μοντέλου αφού είναι προφανές ότι ισχύουν τα εξής: $ARIMA(p,0,0)\equiv AR(p)$, $ARIMA(0,0,q)\equiv MA(q)$ και $ARIMA(p,0,q)\equiv ARMA(p,q)$. Συνεπώς, εξετάζοντας την περίπτωση της εκτίμησης των παραμέτρων του γενικού $ARIMA$ μοντέλου, έχουμε καλύψει όλες τις περιπτώσεις.

Σε ένα δοθέν σύνολο δεδομένων με $N(N=n+d)$ στο πλήθος γνωστές διαδοχικές παρατηρήσεις της αρχικής διαδικασίας $\{X_t\}$ που επιθυμούμε να μελετήσουμε, συμβολίζουμε τις παρατηρήσεις αυτές ως $X_{-d+1},X_{-d+2},X_0,X_1,...,X_n$. Με την βοήθεια αυτών των παρατηρήσεων, κατασκευάζουμε μια νέα σειρά $Y_t=\nabla^d(X_t)$, πλήθους n=N-d παρατηρήσεων, παίρνοντας τις διαφορές τάξης d. Έτσι δημιουργούμε την νέα στάσιμη πλέον σειρά Y_t με όρους $Y_1,Y_2,...,Y_n$. Συνεπώς, το πρόβλημά μας για την εκτίμηση των συντελεστών του αρχικού μη στάσιμου $ARIMA(p,q,d)$ μοντέλου, μειώνεται στην εκτίμηση των συντελεστών μιας αιτιώδους $ARMA(p,q)$ διαδικασίας, όπως ορίστηκε στην σχέση 2.8.2.

2.12.3 Εκτίμηση μέγιστης πιθανοφάνειας

Στην ενότητα αυτή θα συζητηθεί το πρόβλημα εκτίμησης των παραμέτρων $\varphi=\left(\varphi_1,...,\varphi_p\right)$, $\theta=\left(\theta_1,...,\theta_q\right)$ και σ_e^2 στην περίπτωση όπου τα p και q θεωρούνται γνωστά. Ακόμη, θα εστιάσουμε και στο ζήτημα της κατάλληλης επιλογής της τάξης των p, q, εκτός από τα διαγράμματα ACF/PACF που έχουμε ήδη δει.

Όταν οι παράμετροι p, q είναι γνωστοί, καλοί εκτιμητές των φ,θ μπορούν να βρεθούν αν φανταστούμε ότι θα δεδομένα μας αποτελούν παρατηρήσεις από μια στάσιμη Γκαουσιανή χρονοσειρά. Αυτό επιτυγχάνεται μεγιστοποιώντας την

πιθανοφάνεια ως προς τις $p+q+1$ άγνωστες παραμέτρους $\varphi=(\varphi_1,...,\varphi_p)$, $\theta=(\theta_1,...,\theta_q)$ και σ_e^2. Οι εκτιμητές αυτοί που λαμβάνονται από την συγκεκριμένη διαδικασία είναι γνωστοί ως εκτιμητές μέγιστης πιθανοφάνειας. Η μεγιστοποίηση της συνάρτησης πιθανοφάνειας δεν είναι γραμμική με την έννοια ότι η συνάρτηση που πρέπει να μεγιστοποιηθεί δεν είναι μια γραμμική συνάρτηση που θα μπορούσε να λυθεί λύνοντας ένα σύστημα από γραμμικές εξισώσεις. Συνεπώς, ο αλγόριθμος απαιτεί αρχικές παραμέτρους για τα p, q έτσι ώστε να κάνει διάφορες δοκιμές, να αναζητήσει και στο τέλος να καταλήξει στις τιμές εκείνες των p, q που μεγιστοποιούν την συνάρτηση. Είναι φυσικό ότι όσο πιο κοντά στις πραγματικές τιμές των p, q δώσουμε στον αλγόριθμο για αρχικές τιμές, τόσο πιο γρήγορη θα γίνει η αναζήτηση.

Για την εκτίμηση με την μέθοδο της μέγιστης πιθανοφάνειας, υποθέτουμε ότι η $\{Y_t\}$ που δημιουργήσαμε είναι μια Γκαουσιανή χρονοσειρά, δηλαδή οι παρατηρήσεις τις προέρχονται από την κανονική κατανομή, με μέση τιμή μηδέν και αυτοσυνδιακύμανση $Cov(Y_i,Y_j)=E(Y_iY_j)-\overbrace{E(Y_i)E(Y_j)}^{=0}=E(Y_iY_j)$. Ακόμη και αν η σειρά $\{Y_t\}$ που προκύπτει δεν έχει μέση τιμή $\mu=\mathrm{E}(Y_t)$ μηδέν, αν το δείγμα μας είναι αρκετά μεγάλο τότε η μέση τιμή της $\{Y_t\}$ μπορεί να εκτιμηθεί από τον δειγματικό μέσο και να δουλέψουμε με την σειρά που προκύπτει αν αφαιρέσουμε από κάθε παρατήρηση τον δειγματικό μέσο και εργαστούμε ακριβώς με τον ίδιο τρόπο.

Για τις n παρατηρήσεις $Y_n=(Y_1,Y_2,...,Y_n)'$ και τις εκτιμήσεις τους, $\hat{Y}_n=(\hat{Y}_1,\hat{Y}_2,...,\hat{Y}_n)'$, όπου $\hat{Y}_1=0$ και $\hat{Y}_j=E(Y_j/Y_1,...,Y_{j-1})$, $j\geq 2$, συμβολίζουμε με Γ_n τον πίνακα συνδιακύμανσης ο οποίος ισούται με $\Gamma_n=\mathrm{E}(Y_nY_n')$ και υποθέτουμε ότι είναι αντιστρέψιμος.

Η πιθανοφάνεια της Y_n ορίζεται να είναι:

$$L(\Gamma_n)=(2\pi)^{-n/2}(\det\Gamma_n)^{-1/2}\exp\left(-\frac{1}{2}Y_n'\Gamma_n^{-1}Y_n\right) \quad (2.12.3.1)$$

Ο απευθείας υπολογισμός των $\det(\Gamma_n)$ και του Γ_n^{-1} είναι δύσχρηστος και μπορεί να αποφευχθεί εκφράζοντας την (2.12.3.1) σε όρους των σφαλμάτων πρόβλεψης $Y_j-\hat{Y}_j$ και τις διακυμάνσεις τους υ_{j-1}, $j=1,...,n$.

Με βάση αυτό, η πιθανοφάνεια του διανύσματος Y_n, της σχέσης 2.12.3.1, μπορεί να μειωθεί στην σχέση:

$$L(\Gamma_n) = \frac{1}{\sqrt{(2\pi)^n \upsilon_0 ... \upsilon_{n-1}}} \exp\left(-\frac{1}{2} \sum_{j=1}^{n} \left(Y_j - \hat{Y}_j \right)^2 / \upsilon_{j-1} \right) \quad (2.12.3.2)$$

Εάν ο πίνακας διακύμανσης Γ_n μπορεί να εκφραστεί σε πεπερασμένο αριθμό άγνωστων παραμέτρων $\beta_1, ..., \beta_r$, όπου $\beta = (\varphi', \theta')' = (\varphi_1, ..., \varphi_p, \theta_1, ..., \theta_q)'$ (όπως στην περίπτωση που η $\{Y_t\}$ είναι μια $ARMA(p,q)$ διαδικασία), οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων είναι εκείνες οι τιμές που ελαχιστοποιούν την L για το δοθέν σύνολο δεδομένων.

Όταν οι $Y_1, ..., Y_n$ είναι iid θόρυβος, είναι γνωστό (Lehmann & Casella, 1983) ότι για μεγάλα n, οι εκτιμητές μέγιστης πιθανοφάνειας είναι κατά προσέγγιση κανονικά κατανεμημένοι.

Εδώ είναι σημαντικό να πούμε ότι, όπως αναφέρεται και από τους (Brockwell & Davis, 2002, p. 159), ακόμα και αν η $\{Y_t\}$ δεν είναι Γκαουσιανή, δηλαδή οι παρατηρήσεις της δεν προέρχονται από την κανονική κατανομή, μπορούμε να θεωρήσουμε την (2.12.3.2) σαν ένα μέτρο καλής προσαρμογής του μοντέλου στα δεδομένα μας. Έτσι, μπορούν να επιλεχθούν οι παράμετροι $\beta_1, \beta_2, ..., \beta_r$ με τέτοιο τρόπο έτσι ώστε να μεγιστοποιείται η (2.12.3.2) χωρίς να είναι προυπόθεση η $\{Y_t\}$ να είναι γκαουσιανή. Αυτή η παρατήρηση προκύπτει από το γεγονός ότι η κατανομή των εκτιμήσεων $\left(\hat{\varphi}, \hat{\theta}, \hat{\sigma}_e^2 \right)$ πλησιάζει ασυμπτωτικά την κανονική κατανομή (Kedem & Fokianos, 2002, p. 19) για ανεξάρτητες και ισόνομες τυχαίες μεταβλητής μηδενικής μέσης τιμής, $\{e_t\} \sim IID\left(0, \sigma_e^2\right)$, ανεξάρτητα από το αν η κατανομή των $\{e_t\}$ είναι κανονική. Τα μειονεκτήματα που έχει η μέθοδος της μέγιστης πιθανοφάνειας είναι αφενός η πολυπλοκότητα του προβλήματος μεγιστοποίησης της συνάρτησης που συναντάμε σε μερικές περιπτώσεις και αφετέρου η ανάγκη για καλή αρχικοποίηση των τιμών p, q έτσι ώστε να ξεκινήσει ο αλγόριθμος.

Με βάση τις παραπάνω παρατηρήσεις, θα καλούμε τους εκτιμητές $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_r$ ως εκτιμητές μέγιστης πιθανοφάνειας ακόμα και αν η $\{Y_t\}$ δεν είναι Γκαουσιανή. Έτσι, ανεξάρτητα από την από κοινού κατανομή των $Y_1, ..., Y_n$, θα αναφερόμαστε στην (2.12.3.1) και στην αλγεβρικά ισοδύναμή της (2.12.3.2) ως πιθανοφάνεια (ή Γκαουσιανή πιθανοφάνεια) των $Y_1, ..., Y_n$.

Η τελική μορφή που παίρνει η Γκαουσιανή πιθανοφάνεια για μια ARMA(p,q) διαδικασία είναι :

$$L(\varphi,\theta,\sigma_e)=\frac{1}{\sqrt{\left(2\pi\sigma_e^2\right)^n r_0 \ldots r_{n-1}}}\exp\left\{-\frac{1}{2\sigma_e^2}\sum_{j=1}^{n}\frac{\left(Y_j-\hat{Y}_j\right)^2}{r_{j-1}}\right\}, \quad (2.12.3.3)$$

όπου $r_n=\sigma_e^{-2}E\left(Y_{n+1}-\hat{Y}_{n+1}\right)$.

Παίρνοντας την μερική παράγωγο ως προς σ_e^2 και τονίζοντας ότι οι $\widehat{Y_j}, r_j$ δεν εξαρτώνται από την σ_e^2 στην (2.12.3.3), βρίσκουμε ότι οι εκτιμητές μέγιστης πιθανοφάνειας $\hat{\varphi},\hat{\theta}$ και $\widehat{\sigma_e^2}$ ικανοποιούν τις εξισώσεις:

$$\widehat{\sigma_e^2}=\frac{S\left(\hat{\varphi},\hat{\theta}\right)}{n}, \quad \text{όπου } S\left(\hat{\varphi},\hat{\theta}\right)=\sum_{j=1}^{n}\frac{\left(Y_j-\hat{Y}_j\right)^2}{r_{j-1}} \quad (2.12.3.4)$$

και οι εκτιμήσεις $\hat{\varphi},\hat{\theta}$ είναι οι τιμές εκείνες των φ,θ που ελαχιστοποιούν την

$$l(\varphi,\theta)=\ln\left(\frac{S(\varphi,\theta)}{n}\right)+\frac{\sum_{j=1}^{n}\ln r_{j-1}}{n} \quad (2.12.3.5)$$

2.12.4 Μέθοδος ελαχίστων τετραγώνων

Εκτός από την μέθοδο της μέγιστης πιθανοφάνειας, άλλη μια μέθοδος για την εκτίμηση των αγνώστων παραμέτρων είναι η μέθοδος ελαχίστων τετραγώνων. Οι εκτιμητές ελαχίστων τετραγώνων $\tilde{\varphi}$ και $\tilde{\theta}$ των φ,θ για το $ARMA(p,q)$ μοντέλο λαμβάνονται από την ελαχιστοποίηση της συνάρτησης S της σχέσης (2.12.3.4) και όχι της l της σχέσης (2.12.3.5). Αυτό γίνεται πάντα με τον περιορισμό ότι το μοντέλο είναι αντιστρέψιμο και αιτιώδες. Ο εκτιμητής της διασποράς σ_e^2 με την εκτίμηση ελαχίστων τετραγώνων είναι:

$$\tilde{\sigma}_e^2=\frac{S\left(\tilde{\varphi},\tilde{\theta}\right)}{n-p-q} \quad (2.12.4.1)$$

2.12.5 Επιλογή τάξης των p,q

Από τη στιγμή που τα δεδομένα μας έχουν μετασχηματιστεί έτσι ώστε να γίνουν στάσιμα και να μπορούν να προσαρμοστούν σε ένα $ARMA(p,q)$ μοντέλο, ερχόμαστε αντιμέτωποι με το πρόβλημα της επιλογής κατάλληλων τιμών για τις τάξεις p και q. Έχουμε ήδη αναφέρει ότι τα διαγράμματα ACF / $PACF$ μπορούν να μας δώσουν μια αρχική ένδειξη για τις υποψήφιες τιμές των p,q. Ωστόσο, τα διαγράμματα αυτά δεν μας δίνουν πάντα μια ξεκάθαρη εικόνα και συνεπώς είναι εύλογο να συμβουλευόμαστε και άλλα κριτήρια πέραν αυτών. Στην ενότητα αυτή θα ασχοληθούμε με τέτοιου είδους κριτήρια τα οποία αποτελούν μέτρα καλής προσαρμογής του μοντέλου καθώς ποινικοποιούν τον υπερβολικό αριθμό παραμέτρων.

Σε προβλήματα πρόβλεψης, όπως αυτό που ασχολείται η παρούσα διπλωματική δεν θεωρείται πλεονέκτημα να επιλέγονται οι τιμές των p,q να είναι αυθαίρετα πολύ μεγάλες. Το να προσαρμόζουμε ένα μοντέλο με πολύ μεγάλες τάξεις των παραμέτρων p,q έχει ως αποτέλεσμα να οδηγούμαστε σε υποεκτίμηση της διασποράς του λευκού θορύβου. Έτσι, όταν χρησιμοποιήσουμε το μοντέλο για να κάνουμε μια μελλοντική πρόβλεψη, το μέσο τετραγωνικό σφάλμα που θα προκύψει από αυτές, δεν εξαρτάται μόνο από τη διακύμανση του λευκού θορύβου του μοντέλου αλλά και από τα σφάλματα που προκύπτουν από την εκτίμηση των παραμέτρων του μοντέλου. Κατά συνέπεια, τα σφάλματα αυτά θα είναι πολύ μεγαλύτερα για τα μοντέλα εκείνα στα οποία έχουν επιλεγεί αυθαίρετα μεγάλες τιμές των p,q. Για τον λόγο αυτό χρειαζόμαστε να ορίσουμε έναν «όρο ποινής» για να αποτρέψουμε την προσαρμογή των μοντέλων σε πολύ μεγάλο αριθμό παραμέτρων. Τέτοια κριτήρια που βασίζονται σε τέτοιου είδους ποινικοποιημένες συναρτήσεις είναι τα AIC $\left(\text{Akaike Information Criterion}\right)$, BIC $\left(\text{Bayesian Information Criterion}\right)$ και η διορθωμένη έκδοση του κριτηρίου πληροφορίας AIC (1973), που προτάθηκε από τους (Hurvich & Tsai, 1989) και είναι γνωστό με την ονομασία $AICC$ (Bias-Corrected Akaike Information Criterion) . Όσο μικρότερες είναι οι τιμές των κριτηρίων αυτών, τόσο καλύτερη είναι η προδιαγραφή του μοντέλου.

Για το κριτήριο AIC επιλέγονται τα p,q,φ_p,θ_p έτσι ώστε να ελαχιστοποιηθεί η συνάρτηση της σχέσης (2.12.5.1):

$$AIC=-2\ln L\left(\varphi_p,\theta_q,S\left(\varphi_p,\theta_q\right)/n\right)+2\left(p+q+1\right) \quad (2.12.5.1)$$

Το προτιμότερο μοντέλο με βάση αυτό το κριτήριο είναι εκείνο με το μικρότερο AIC.

Θα έλεγε κανείς ότι η εισαγωγή μεγαλύτερης τάξης παραμέτρων *p,q* στο μοντέλο βελτιώνει σε κάθε περίπτωση την προσαρμογή του μοντέλου ανεξάρτητα αν αυτές είναι στατιστικά σημαντικές ή όχι. Αυτό θα συνέβαινε γιατί η πρόσθεση μεγαλύτερης τάξης παραμέτρων *p,q* αυξάνει το $\ln L\left(\varphi_p, \theta_q, S\left(\varphi_p, \theta_q\right)/n\right)$, άρα ο πρώτος όρος του AIC μειώνεται. Από την άλλη αυξάνεται ο δεύτερος όρος του AIC. Τελικά η εισαγωγή μεγαλύτερης τάξης παραμέτρων *p,q* στο μοντέλο μειώνει την τιμή του AIC μόνο αν αυτές βελτιώνουν την προσαρμογή του μοντέλου σε βαθμό που υπερβαίνει το αυξημένο αντίβαρο του δεύτερου όρου *2(p+q+1)* (Καρώνη & Οικονόμου, 2017, p. 189).

Τα αντίστοιχα ισχύουν και για το διορθωμένο κριτήριο AICC το οποίο ορίζεται από τη σχέση (2.12.5.2) να είναι:

$$AICC = -2\ln L\left(\varphi_p, \theta_q, S\left(\varphi_p, \theta_q\right)/n\right) + 2\left(p+q+1\right)n/\left(n-p-q-2\right) \qquad (2.12.5.2)$$

Για την προσαρμογή αυτοπαλινδρομικών μοντέλων σε ένα σύνολο δεδομένων, οι έρευνες Monte Carlo από τους (Jones, 1975) και (Shibata, 1976), έδειξαν ότι το κριτήριο AIC έχει την τάση να υπερεκτιμά την τάξη p με αποτέλεσμα να οδηγούμαστε σε «υπερπροσαρμογή» μοντέλου. Εδώ, να αναφέρουμε ότι οι όροι ποινής $2\left(p+q+1\right)$ και $2\left(p+q+1\right)n/\left(n-p-q-2\right)$ από τα στατιστικά AIC και AICC αντίστοιχα, είναι ασυμπτωτικά ισοδύναμοι καθώς $n \to \infty$. Ωστόσο, ο όρος ποινής από το στατιστικό AICC είναι πιο ακριβής για μεγάλης τάξης μοντέλα, πράγμα που αντισταθμίζει την τάση του μη διορθωμένου στατιστικού AIC για υπερπροσαρμογή.

Το στατιστικό BIC είναι επίσης ένα άλλο κριτήριο που προσπαθεί να διορθώσει αυτό το μειονέκτημα του AIC για υπερπροσαρμογή. Για μια μηδενικής τάξης, αιτιώδη και αντιστρέψιμη $ARMA\left(p,q\right)$ διαδικασία, το κριτήριο BIC, το οποίο ανακαλύφθηκε από τον (Schwert, 1978) ορίζεται να είναι:

$$BIC = \left(n-p-q\right)\ln\left[n\hat{\sigma}_e^2/\left(n-p-q\right)\right] + n\left(1+\ln\sqrt{2\pi}\right) + \left(p+q\right)\ln\left[\left(\sum_{t=1}^{n} X_t^2 - n\hat{\sigma}_e^2\right)/\left(p+q\right)\right] \quad (2.12.5.3),$$

όπου $\hat{\sigma}_e^2$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας της διασποράς του λευκού θορύβου.

Το κριτήριο BIC είναι ένα συνεπές κριτήριο επιλογής τάξης με την έννοια ότι εάν τα δεδομένα μας $\left\{X_1, \ldots, X_n\right\}$ είναι πράγματι παρατηρήσεις μιας $ARMA\left(p,q\right)$ διαδικασίας και αν τα $\hat{p}$ και $\hat{q}$ είναι οι εκτιμήσεις των τάξεων που έχουν βρεθεί με την ελαχιστοποίηση του BIC, τότε $\hat{p} \to p$ και $\hat{q} \to q$ με πιθανότητα 1 όσο το $n \to \infty$ (Hannan, 1980). Εδώ να πούμε ότι αυτή η ιδιότητα δεν ισχύει για τα κριτήρια AIC και AICC.

2.12.6 Διαγνωστικός έλεγχος

Μετά την εφαρμογή του προτεινόμενου μοντέλου στα δεδομένα μας, πρέπει να εξετάσουμε κατά πόσο είναι ικανοποιητικό, εάν τηρούνται οι προϋποθέσεις της θεωρίας και κατά πόσο ενδέχεται βελτιώσεις. Αυτό μπορεί να γίνει με τον διαγνωστικό έλεγχο των υπολοίπων του μοντέλου που έχουμε εκτιμήσει. Συγκρίνοντας τις παρατηρούμενες τιμές με τις εκτιμήσεις των τιμών που προκύπτουν έπειτα από την προσαρμογή του μοντέλου μας, μπορούμε να αποφανθούμε για την καταλληλόλητα του.

Για παράδειγμα, όταν προσαρμόζουμε ένα $ARMA(p,q)$ μοντέλο σε ένα δοσμένο σύνολο δεδομένων, προσδιορίζουμε τους εκτιμητές μέγιστης πιθανοφάνειας $\hat{\varphi},\hat{\theta}$ και $\hat{\sigma}_e^2$ των παραμέτρων φ,θ και σ_e^2. Κατά τη διάρκεια αυτής της διαδικασίας, έχουν υπολογιστεί, από το προσαρμοσμένο μοντέλο, οι εκτιμώμενες τιμές, $\hat{Y}_t\left(\hat{\varphi},\hat{\theta}\right)$, της Y_t που βασίζονται στις $Y_1,\ldots,Y_{t-1}$.

Τα υπόλοιπα που παίρνουμε από το προσαρμοσμένο μοντέλο, ορίζονται να είναι:

$$\hat{e}_t = Y_t - \hat{Y}_t\left(\hat{\varphi},\hat{\theta}\right), \quad t=1,\ldots,n \qquad (2.12.6.1).$$

Συνεπώς, εάν υποθέσουμε ότι το μοντέλο $ARMA(p,q)$ που δημιουργήθηκε με τη μέγιστη πιθανοφάνεια είναι η πραγματική διαδικασία που παράγει την $\{Y_t\}$, τότε μπορούμε να πούμε ότι $\left\{\hat{e}_t\right\} \sim WN\left(0,\hat{\sigma}_e^2\right)$.

Εν τούτοις, όταν έχουμε ένα δοθέν σύνολο δεδομένων $\{Y_1,\ldots,Y_n\}$ και επιλέγουμε να προσαρμόσουμε ένα $ARMA(p,q)$ μοντέλο, υποθέτουμε μόνο ότι η $Y_1,\ldots,Y_n$ έχει δημιουργηθεί από μια $ARMA(p,q)$ διαδικασία με άγνωστες παραμέτρους φ,θ και σ_e^2, των οποίων οι εκτιμητές μέγιστης πιθανοφάνειας είναι τα $\hat{\varphi},\hat{\theta}$ και $\hat{\sigma}_e^2$ αντίστοιχα. Με αυτές τις υποθέσεις σε περίπτωση που το μοντέλο μας είναι κατάλληλο, η $\hat{e}_t$, $t=1,\ldots,n$ πρέπει να έχει ιδιότητες, που είναι παρόμοιες με αυτές μιας διαδικασίας λευκού θορύβου. Αυτό έγκειται στο γεγονός ότι $E\left(\hat{e}_t\left(\varphi,\theta\right)-e_t\right)^2 \to 0$ καθώς $t\to\infty$. Αυτό σημαίνει ότι όσο μεγαλώνει το δείγμα μας τότε οι τιμές των υπολοίπων πρέπει να συγκλίνουν στις τυχαίες διακυμάνσεις που όπως έχουμε αναφέρει είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές, μηδενικής μέσης τιμής και διασποράς σ_e^2.

Κατά συνέπεια, οι ιδιότητες των υπολοίπων $\left\{\hat{e}_t\right\}$ θα πρέπει να αντικατοπτρίζουν τις ιδιότητες της ακολουθίας $\{e_t\}$ που παράχθηκε από την συγκεκριμένη

$ARMA(p,q)$ διαδικασία. Πιο συγκεκριμένα, η διαδικασία $\{\hat{e}_t\}$ θα πρέπει κατά προσέγγιση να είναι :

i. ασυσχέτιστη εάν $\{e_t\} \sim WN(0,\sigma_e^2)$

ii. ανεξάρτητη εάν $\{e_t\} \sim IID(0,\sigma_e^2)$

iii. κανονικά κατανεμημένη εάν $\{e_t\} \sim N(0,\sigma_e^2)$

Αν διαιρέσουμε τα υπόλοιπα $\hat{e}_t$, $t = 1,2,...,n$ με την θεωρητική εκτίμηση της ρίζας της διασπορά τους, $\hat{\sigma}_e = \sqrt{\left(\sum_{t=1}^{n} e_t^2\right)/n}$, λαμβάνουμε τα προσαρμοσμένα υπόλοιπα (rescaled residuals) $\hat{R}_t$, $t = 1,2,...,n$ τα οποία είναι:

$$\hat{R}_t = \hat{e}_t / \hat{\sigma}_e^2 \qquad (2.12.6.2)$$

Εάν το προσαρμοσμένο μοντέλο είναι κατάλληλο τότε τα προσαρμοσμένα υπόλοιπα πρέπει να έχουν ιδιότητες αντίστοιχες με αυτές μιας ακολουθίας λευκού θορύβου με μέση τιμή μηδέν και διασπορά ίση με 1 ($WN(0,1)$).

Οι παρακάτω διαγνωστικοί έλεγχοι είναι όλοι βασισμένοι στις ιδιότητες που αναμένουμε από τα υπόλοιπα ή τα προσαρμοσμένα υπόλοιπα, κάτω από την υπόθεση ότι το μοντέλο μας είναι σωστό και ότι $\{e_t\} \sim IID(0,\sigma_e^2)$.

2.12.6.1 Το γράφημα των $\{\hat{R}t, t=1,...,n\}$

Εάν το προσαρμοσμένο μοντέλο είναι κατάλληλο, τότε όπως αναφέραμε και παραπάνω, το γράφημα των προσαρμοσμένων υπολοίπων $\{\hat{R}_t,\ t = 1,...,n\}$ πρέπει να μοιάζει με αυτό ενός λευκού θορύβου με διακύμανση ίση με 1. Ενώ είναι δύσκολο να καταλάβουμε την συσχέτιση που ενδέχεται να υπάρχει μεταξύ των παρατηρήσεων του $\{\hat{R}_t,\ t = 1,...,n\}$ αποκλειστικά και μόνο από το γράφημά τους, οι αποκλίσεις από τη μέση τιμή μηδέν, είναι μερικές φορές εμφανής. Τέτοιο παράδειγμα είναι όταν στο γράφημα των $\hat{R}_t$, συναρτήσει του t, μπορεί να εμφανίζεται μια τάση ή εάν η διασπορά φαίνεται να είναι ασταθής με έντονες διακυμάνσεις. Δηλαδή όταν το μέγεθός τους εξαρτάται σε μεγάλο βαθμό από τις χρονικές στιγμές t.

Έπειτα από τον γραφικό έλεγχο, το επόμενο βήμα που μπορούμε να κάνουμε είναι να διαπιστώσουμε εάν η δειγματική συνάρτηση αυτοσυσχέτισης των $\left\{\hat{e}_t\right\}$ $\left(\text{ή ισοδύναμα των } \left\{\widehat{R}_t\right\}\right)$,συμπεριφέρεται όπως θα έπρεπε κάτω από την υπόθεση ότι το μοντέλο που προσαρμόσαμε είναι κατάλληλο.

Όπως έχουμε αναφέρει σε προηγούμενες ενότητες, είναι γνωστό ότι για μεγάλες τιμές του n, οι δειγματικές αυτοσυσχετίσεις από μια iid ακολουθία $Y_1,...,Y_n$ με πεπερασμένη διακύμανση, είναι κατά προσέγγιση iid ακολουθία που ακολουθεί την κανονική κατανομή $N\left(0, 1/n\right)$. Συνεπώς, ένας τρόπος για να εξετάσουμε κατά πόσο τα παρατηρούμενα υπόλοιπα είναι συνεπή με τον iid θόρυβο, είναι να εξετάσουμε τις δειγματικές αυτοσυσχετίσεις των υπολοίπων και να απορρίψουμε την υπόθεση του iid θορύβου εάν περισσότερες από το 5% των υστερήσεων πέφτουν εκτός των ορίων $\pm 1.96/\sqrt{n}$. Εδώ να αναφέρουμε ότι τα εκτιμημένα υπόλοιπα δεν είναι εφικτό να είναι ακριβώς iid ακόμα και αν το πραγματικό μοντέλο που παράγει τα δεδομένα μας είναι όπως υποτέθηκε.

2.12.6.2 Έλεγχοι Box-Pierce, Ljung-Box και McLeod-Li

Έπειτα από αρκετές εφαρμογές διαφόρων ερευνητών, παρατηρήθηκε ότι μόνο ένας τέτοιου είδους έλεγχος για τις αυτοσυσχετίσεις, δεν είναι καλή τεχνική και αρκετές φορές μπορεί να ελλοχεύει κινδύνους. Προκειμένου να διορθωθεί κάτι τέτοιο, τα όρια $\pm 1.96/\sqrt{n}$ θα έπρεπε να τροποποιηθούν για να δώσουν έναν πιο ακριβή έλεγχο. Τέτοιοι έλεγχοι $\left(\text{portmanteau test}\right)$ δημιουργήθηκαν από τους (Box & Pierce, 1970) και (Ljung & Box, 1978) με τα οποία ελέγχεται η υπόθεση της ανεξαρτησίας των σφαλμάτων εξετάζοντας την γενική εικόνα των h πρώτων αυτοσυσχετίσεων χωρίς να λαμβάνεται υπόψιν η κάθε τιμή της αυτοσυσχέτισης ξεχωριστά.

Συνεπώς, αντί για να ελέγξουμε πότε οι δειγματικές αυτοσυσχετίσεις $\widehat{\rho}(j)$ πέφτουν ανάμεσα στα όρια όπως αναφέρθηκε παραπάνω μπορούμε να θεωρήσουμε το στατιστικό

$$Q = n\sum_{j=1}^{h}\widehat{\rho}^{\,2}(j)$$

Η τυχαία μεταβλητή Q ακολουθεί κατά προσέγγιση την χ^2 $(chi-squared)$ με h βαθμούς ελευθερίας. Απορρίπτουμε την μηδενική υπόθεση της ανεξαρτησίας σε επίπεδο α, εάν $Q > \chi^2_{1-\alpha}(h)$, όπου $\chi^2_{1-\alpha}(h)$ είναι το $1-\alpha$ ποσοστημόριο της χ^2 (chi-squared) κατανομής με h βαθμούς ελευθερίας. Μια μεγάλη τιμή του Q θα σήμαινε ότι οι δειγματικές αυτοσυσχετίσεις των δεδομένων μας, είναι πολύ μεγάλες για να είναι πράγματι τα δεδομένα μας ένα δείγμα από μια iid ακολουθία. Μετά την εφαρμογή του στατιστικού αυτού, παρατηρήθηκε από τους Ljung-Box, ότι δεν έδινε πάντα έγκυρα αποτελέσματα. Κατά συνέπεια, οι ίδιοι κατέληξαν ότι ένα τροποποιημένο στατιστικό σε σχέση με το παλιό Q, το οποίο συμβόλισαν Q_{LB} και ακολουθεί την ίδια κατανομή χ^2 αλλά δίνει πιο ακριβή αποτελέσματα. Το στατιστικό αυτό ορίζεται να είναι:

$$Q_{LB} = n(n+2)\sum_{j=1}^{h} \hat{\rho}^2(j)/(n-j)$$

και όπως αναφέρθηκε ακολουθεί και αυτό, προσεγγιστικά, την χ^2 με h βαθμούς ελευθερίας.

Ένα ακόμη portmanteau test, το οποία ανακαλύφθηκε από τους (McLeod & Li, 1983), μπορεί να χρησιμοποιηθεί σαν ένας επιπλέον έλεγχος καθώς εάν τα υπόλοιπα μας είναι iid, τότε και τα τετράγωνά τους είναι επίσης iid. Βασίζεται στο ίδιο στατιστικό όπως και των Ljung-Box, όμως, αντί για τις δειγματικές αυτοσυσχετίσεις των δεδομένων μας έχουμε τις δειγματικές αυτοσυσχετίσεις των τετραγώνων των δεδομένων, $\hat{\rho}_{ww}(h)$. Το στατιστικό αυτό δίνεται από τη σχέση:

$$Q_{ML} = n(n+2)\sum_{k=1}^{h} \hat{\rho}^2_{ww}(k)/(n-k)$$

Η μηδενική υπόθεση περί ανεξαρτησίας των δεδομένων (δηλαδή ότι είναι iid) απορρίπτεται σε επίπεδο σημαντικότητας a εάν η παρατηρούμενη τιμή του Q_{ML} είναι μεγαλύτερη από το $1-\alpha$ ποσοστημόριο της $\chi^2(h)$ κατανομής.

Παρότι οι παραπάνω έλεγχοι θεωρούνται αρκετά ικανοποιητικοί και αποτελούν βασικότερα εργαλεία ελέγχου καλής προσαρμογής του μοντέλου, παρακάτω παραθέτουμε και μερικούς εναλλακτικούς ελέγχους τυχαιότητας των υπολοίπων.

2.12.6.3 Έλεγχος σημειων αλλαγής προσήμου(The turning point test)

Ας υποθέσουμε ότι έχουμε μια ακολουθία από παρατηρήσεις, την $y_1, y_2, ..., y_n$. Τότε λέμε ότι υπάρχει σημείο αλλαγής προσήμου σε χρόνο $i,\ 1 < i < n,$ εάν $y_{i-1} < y_i$ και $y_i > y_{i+1}$ ή εάν $y_{i-1} > y_i$ και $y_i < y_{i+1}$. Αν επιπλέον υποθέσουμε ότι οι τιμές μας $y_1, y_2, ..., y_n$ είναι $IID\left(0, \sigma^2\right)$, η πιθανότητα να έχουμε σημείο αλλαγής προσήμου σε χρόνο i είναι $2/3$. Αυτό συμβαίνει διότι για την τριάδα από τους τυχαίους πραγματικούς αριθμούς y_{i-1}, y_i, y_{i+1} θα έχουμε πιθανότητα $\frac{1}{3}$ κάποιος να είναι ο μικρότερος και $\frac{1}{3}$ κάποιος να είναι ο μεγαλύτερος. Άρα, η πιθανότητα ύπαρξης σημείου αλλαγής προσήμου στον χρόνο i είναι ίση με $\frac{2}{3}$ $\left(\text{η πιθανότητα ο } y_i \text{ να είναι μέγιστος} + \text{η πιθανότητα ο } y_i \text{ να είναι ελάχιστος}\right)$. Αν συμβολίσουμε με T το πλήθος των σημείων αλλαγής προσήμου μιας $IID\left(0, \sigma^2\right)$ ακολουθίας μεγέθους n, τότε έχουμε n-2 πιθανά σημεία αλλαγής προσήμου. Εδώ να αναφέρουμε ότι έχουμε αποκλείσει από πιθανά σημεία αλλαγής προσήμου την πρώτη παρατήρηση y_1, καθώς και την τελευταία y_n.

Συνεπώς, η αναμενόμενη τιμή του T είναι:

$$\mu_{\mathrm{T}} = \mathrm{E}\left(\mathrm{T}\right) = 2\left(n-2\right)/3$$

Ακόμη, αποδεικνύεται ότι η διασπορά του T δίνεται από τον τύπο :

$$\sigma_{\mathrm{T}}^2 = Var\left[T\right] = \frac{16n-29}{90}$$

Μια μεγάλη τιμή της ποσότητας $\mathrm{T} - \mu_{\mathrm{T}}$ θα φανέρωνε ότι η σειρά αποκλίνει πολύ περισσότερο από όσο θα αναμέναμε από μια iid ακολουθία. Από την άλλη μεριά, μια τιμή της ποσότητας $\mathrm{T} - \mu_{\mathrm{T}}$ που θα ήταν αρκετά κάτω από το μηδέν θα φανέρωνε συσχέτιση μεταξύ των γειτονικών παρατηρήσεων.

Για μια iid ακολουθία με μεγάλη τιμή του n, από το Κεντρικό Οριακό Θεώρημα (Κ.Ο.Θ), αποδεικνύεται ότι $\mathrm{T} \sim \mathrm{N}\left(\mu_{\mathrm{T}}, \sigma_{\mathrm{T}}^2\right)$. Συνεπώς, με την εφαρμογή του ελέγχου των σημείων αλλαγής προσήμου, μπορούμε να απορρίψουμε την μηδενική υπόθεση περί ανεξαρτησίας σε επίπεδο σημαντικότητας α, εάν $\left|\mathrm{T} - \mu_{\mathrm{T}}\right| / \sigma_{\mathrm{T}} > \Phi_{1-\alpha/2}$, όπου $\Phi_{1-\alpha/2}$ είναι το $1-\alpha/2$ ποσοστημόριο της τυπικής κανονικής κατανομής. (Μια συνήθης τιμή του α είναι η 0.05, για την οποία η αντίστοιχη τιμή του $\Phi_{1-\alpha/2}$ είναι η 1.96).

2.12.6.4 Έλεγχος προσήμου διαφορών (Difference - Sign test)

Στον εν λόγω έλεγχο μετράμε τον αριθμό S των τιμών για τον χρόνο i, για τον οποίο ισχύει $y_i > y_{i-1}$, $i = 2,...,n$ ή ισοδύναμα τον αριθμό των φορών όπου οι διαφορές $y_i - y_{i-1}$ είναι θετικές. Ο έλεγχος προσήμου διαφορών είναι μια αντίστοιχη διαδικασία όπως ο έλεγχος αλλαγής προσήμου που αναφέραμε παραπάνω.

Είναι εμφανές ότι για μια iid διαδικασία, η αναμενόμενη τιμή του S θα είναι:

$$\mu_S = E(S) = \frac{n-1}{2}$$

Ακόμη, μπορεί να δειχθεί ότι κάτω από την ίδια υπόθεση της iid διαδικασίας, η διασπορά του S είναι: $\sigma_S^2 = Var(S) = (n+1)/12$.

Εάν η ποσότητα $S - \mu_S$ μας δώσει μια θετική (ή αντίστοιχα αρνητική) τιμή, πολύ μεγαλύτερη του μηδενός, θα μας φανέρωνε την παρουσία τάσης στα δεδομένα (είτε αυξανόμενης, είτε μειούμενης). Αυτό θα ερχόταν σε αντίθεση με την υπόθεση της ανεξαρτησίας και θα μας οδηγούσε στη απόρριψη της. Με βάση τον έλεγχο αυτό, απορρίπτουμε την υπόθεση ότι δεν υπάρχει τάση στα δεδομένα μας, εάν $|S - \mu_S| / \sigma_S > \Phi_{1-\alpha/2}$.

Ο έλεγχος προσήμου διαφορών πρέπει να χειρίζεται με προσοχή διότι σε ένα σύνολο από παρατηρήσεις με έντονη εποχιακή συνιστώσα, ο έλεγχος θα αποτύγχανε. Αυτό γιατί το σύνολο δεδομένων μας θα έδειχνε ότι έχει περάσει τον έλεγχο προσήμου διαφορών για την τυχαιότητα, αφού χονδρικά οι μισές από τις παρατηρήσεις θα είναι σημεία που αυξάνουν και συνεπώς θα βγάζαμε εσφαλμένα συμπεράσματα.

2.12.6.5 Έλεγχος Κανονικότητας (Checking for Normality)

Στην περίπτωση όπου τα υπόλοιπα εκτός από ανεξάρτητα είναι κανονικά κατανεμημένα, ένας εύκολος αρχικός έλεγχος για να διαπιστωθεί αυτό είναι με τη βοήθεια ενός Q-Q plot (Quantile-Quantile plot). Ο έλεγχος αυτός λειτουργεί ως εξής: Ας υποθέσουμε ότι έχουμε ένα τυχαίο δείγμα , μεγέθους n, από την κανονική κατανομή $\mathrm{N}\left(\mu,\sigma^2\right)$ το οποίο συμβολίζουμε με $e_1,e_2,...,e_n$. Αρχικά, τοποθετούμε τις παρατηρήσεις μας σε αύξουσα σειρά οι οποίες συμβολίζονται ως $e_{(1)},e_{(2)},...,e_{(n)}$. Εάν $\mathrm{X}_{(1)},\mathrm{X}_{(2)},...,\mathrm{X}_{(n)}$ είναι οι παρατηρήσεις τοποθετημένες σε αύξουσα σειρά από ένα δείγμα μεγέθους n της τυπικής κανονικής κατανομής $\mathrm{N}\left(0,1\right)$ τότε:

$$Ee_{(j)}=\mu+\sigma m_j,\ \text{όπου}\ \ m_j=EX_{(j)},\ \ j=1,...,n$$

Το γράφημα των σημείων $\left(m_1,e_{(1)}\right),...,\left(m_n,e_{(n)}\right)$ ονομάζεται Q-Q plot και μπορεί εύκολα να κατασκευαστεί με την βοήθεια της R, μέσω της εντολής qqplot. Εάν η υπόθεση της κανονικότητας είναι σωστή, τότε το γράφημα θα πρέπει να είναι κατά προσέγγιση γραμμικό. Κατά συνέπεια, η τετραγωνική συσχέτιση των σημείων $\left(m_i,e_{(i)}\right),\ \ i=1,...,n$ πρέπει να είναι κοντά στο 1. Επομένως, η υπόθεση της κανονικότητας απορρίπτεται αν ο συντελεστής συσχέτισης R^2 είναι σημαντικά μικρός.

Ο συντελεστής συσχέτισης R^2 ορίζεται να είναι:

$$R^2=\frac{\left(\sum_{i=1}^{n}\left(e_{(i)}-\bar{e}\right)\Phi^{-1}\left(\frac{i-0.5}{n}\right)\right)^2}{\sum_{i=1}^{n}\left(e_{(i)}-\bar{e}\right)^2\sum_{i=1}^{n}\left(\Phi^{-1}\left(\frac{i-0.5}{n}\right)\right)^n},\ \text{όπου}\ \bar{e}=n^{-1}\left(e_1+...+e_n\right)$$

Πέραν του γραφικού αυτού ελέγχου, συνηθισμένοι έλεγχοι κανονικότητας των υπολοίπων είναι οι έλεγχοι Kolmogorov-Smirnov , Anderson-Darling και Jarque-Bera.

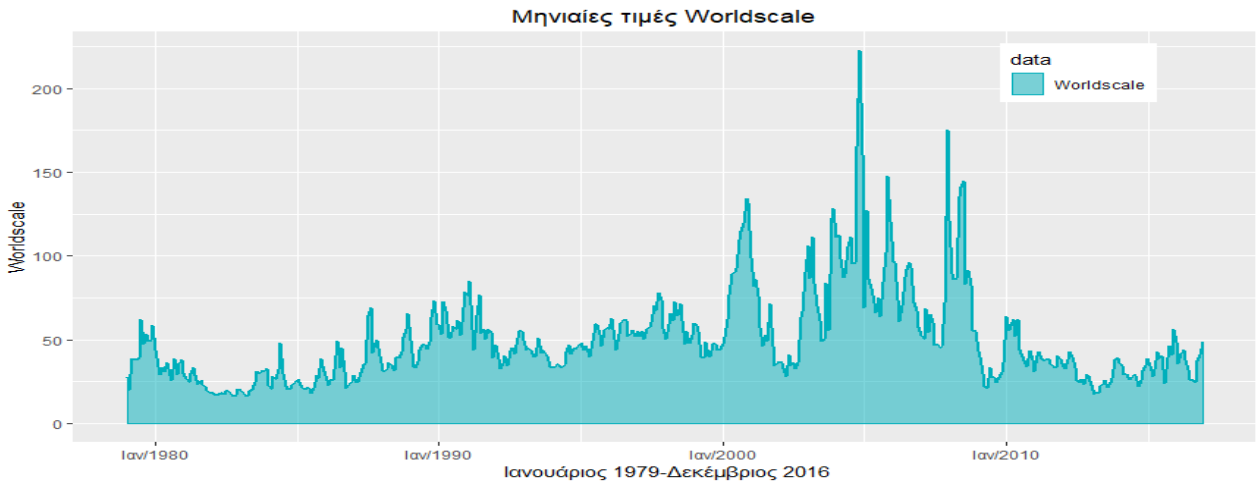
3. Εφαρμογές στον δείκτη Worldscale

Στο κεφάλαιο αυτό θα εφαρμόσουμε τις τεχνικές που συζητήθηκαν σε όλες τις προηγούμενες ενότητες για να προβλέψουμε τις μελλοντικές τιμές του δείκτη Worldscale του Διαγράμματος 3.1.

Τα δεδομένα μας αφορούν τον δείκτη Worldscale που αναλύθηκε στην ενότητα 1.2.4 και περιλαμβάνουν μηνιαίες τιμές στιγμιαίου ναύλου 39 χρόνων, από τον Ιανουάριο του 1979 έως τον Δεκέμβριο του 2017, για πλοία VLCC μεταφορικής ικανότητας 280.000 DWT στη γραμμή μεταφοράς αργού πετρελαίου από το Ras Tanura της Αραβικής Χερσονήσου προς το Rotterdam της Ολλανδίας. Τα επίπεδα των ναύλων (freight rates) είναι οι πιο χρήσιμες περιγραφικές μεταβλητές αφού εμπεριέχουν σημαντική πληροφορία ικανή να περιγράψει την κατάσταση της ναυτιλίας (Ζαχαριουδάκης, 2007). Είναι αναγκαίο να δίνεται πληροφορία για τιμές ναύλων Spot ή Period , για διάφορους τύπους πλοίων, γραμμές ,φορτία αλλά και χωρητικότητα (DWT). Η πηγή από όπου αντλήθηκαν όλες οι ναυτιλιακές χρονοσειρές είναι η βάση δεδομένων Clarkson's.

Στην ανάλυση μας προκειμένου να ελέγξουμε κατά πόσο οι προβλέψεις μας είναι ακριβής και ανταποκρίνονται στα πραγματικά δεδομένα, αφαιρέσαμε τις τιμές του τελευταίου έτους, δηλαδή τις 12 τιμές του 2017 τις οποίες θα θεωρούμε άγνωστες μέχρι το τέλος της ανάλυσης μας όπου θα προσπαθήσουμε να τις προβλέψουμε και να συγκρίνουμε τα αποτελέσματά μας με τις πραγματικές τιμές.

Σε όλο το κεφάλαιο θα συμβολίζουμε με $x_1, x_2, \ldots, x_{456}$ το δείγμα των παρατηρήσεων μας με τις μηνιαίες τιμές του Worldscale δείκτη από τον Ιανουάριο του 1979 έως τον Δεκέμβριο του 2016. Το Διάγραμμα 3.1 δείχνει στον άξονα των x, τους μήνες από τον Ιανουάριο του 1979 έως τον Δεκέμβριο του 2016 και στον άξονα των y τις τιμές του δείκτη Worldscale τους αντίστοιχους αυτούς μήνες.



<u>Διάγραμμα 3.1</u>: Μηνιαίες τιμές spot ναύλων μεταφοράς πετρελαίου με πλοία VLCC στη διαδρομή Ras Tanura-Rotterdam.

Πριν αρχίσουμε τη διαδικασία που απαιτείται για τη μοντελοποίηση του δείκτη Worldscale είναι σύνηθες να ξεκινήσουμε με την εξέταση των χαρακτηριστικών της κατανομής των δεδομένων μας. Αυτή η στατιστική ανάλυση των δεδομένων ονομάζεται περιγραφική στατιστική και περιλαμβάνει μέτρα θέσης (μέση τιμή, διάμεσος), μέτρα διασποράς (διακύμανση, ενδοτεταρτημοριακό Εύρος Q) και μέτρα που περιγράφουν το σχήμα της κατανομής τους (κύρτωση, λοξότητα).
Επομένως, σε αυτό το βήμα υπολογίζεται ένα σύνολο περιγραφικών στατιστικών για τις μηνιαίες τιμές των ναύλων. Αυτά τα στατιστικά στοιχεία δεν λαμβάνουν υπόψη το χρόνο και επομένως όλα τα σημεία με τις τιμές των ναύλων θεωρούνται ισοδύναμα. Κατά συνέπεια, η χρονολογική σειρά μειώνεται σε ένα κοινό σύνολο δεδομένων, το οποίο μπορεί εύκολα να περιγραφεί από την κεντρική τάση (μέση τιμή) και την εξάπλωση (διακύμανση). Όταν οι δύο αυτές ποσότητες δεν μεταβάλλονται κατά τη διάρκεια του χρόνου (στασιμότητα), θα μπορούσαν να χρησιμοποιηθούν ως περιγραφικά χαρακτηριστικά των επιπέδων των ναύλων (μέσες τιμές ναύλων και μέγεθος των διακυμάνσεων τους). Ωστόσο, εφόσον η κεντρική τάση και η διακύμανση αυξομειώνονται με την πάροδο του χρόνου, αυτές οι ποσότητες δεν έχουν νόημα και είναι απαραίτητη η εκτίμηση της αύξησης ή της μείωσης τους στον χρόνο. Ο Πίνακας 3.1 μας δίνει τα περιγραφικά χαρακτηριστικά του δείκτη Worldscale.

Πίνακας 3.1:

Summary Statistics	
Minimum	16,17
1st Quantile	31,07
Median	42,83
Mean	49,34
3rd Quantile	58,75
Maximum	222,5
Variance	767,06
Standard deviation	27,70
Skewness	1,93
Kurtosis	5,55

Ο δειγματικός μέσος, η τυπική απόκλιση, η διακύμανση, η λοξότητα και η κύρτωση δίνονται αντίστοιχα από τις σχέσεις:

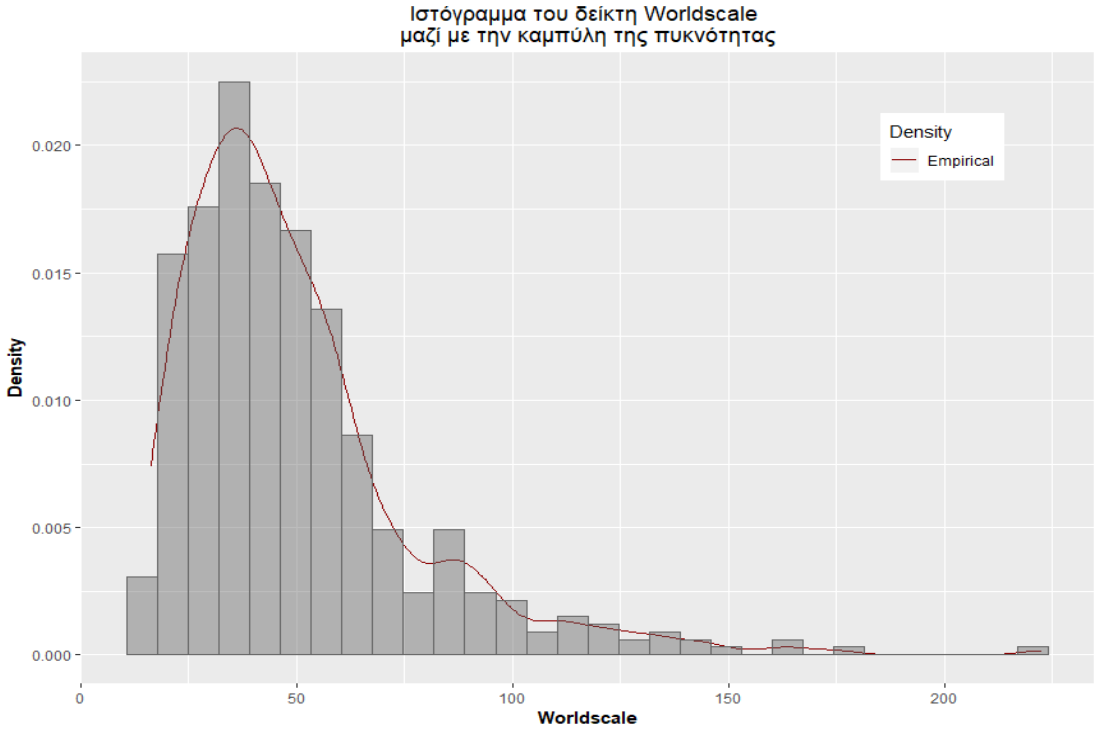
$$\text{mean: } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$

$$\text{standard deviation: } s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}, \qquad \text{variance: } \text{var} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1},$$

$$\text{skewness} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3 \Big/ n}{s^3}, \qquad \text{kurtosis} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4 \Big/ n}{s^4} - 3$$

Εδώ να σημειώσουμε ότι το πρώτο ποσοστημόριο (1st Quantile) που ισούται με 31,07 είναι η τιμή κάτω από την οποία βρίσκεται το 25% των ναύλων εφόσον αυτά τοποθετηθούν σε αύξουσα σειρά. Αντίστοιχα το 3ο ποσοστημόριο που ισούται με 58,75 είναι η τιμή κάτω από την οποία βρίσκεται το 75% των ναύλων εφόσον αυτά τοποθετηθούν σε αύξουσα σειρά. Η λοξότητα (skewness) είναι ένα μέτρο που δείχνει τον βαθμό ασσυμετρίας της κατανομής των ναύλων γύρω από τη μέση τιμή. Μια κατανομή η οποία είναι συμμετρική ως προς τη μέση τιμή (π.χ η κανονική κατανομή) έχει λοξότητα μηδέν. Συνεπώς, η λοξότητα που μας δίνει ο Πίνακας 3.1 για τα ναύλα είναι κοντά στο 2 που δηλώνει θετική ασσυμετρία. Αντίστοιχα, η κύρτωση είναι ένα μέτρο που περιγράφει το βαθμό κυρτότητας της κατανομής των ναύλων και συγκρίνεται με την τιμή 3 που είναι ο συντελεστής κυρτότητας α της κανονικής κατανομής, όπου $\alpha = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4 / n}{s^4}$. Η διαφορά α-3, για λεπτόκυρτες κατανομές παίρνει θετικές τιμές (θετική κύρτωση), ενώ για πλατύκυρτες κατανομές γίνεται αρνητική (αρνητική κύρτωση). Η τιμή 5,55 που δίνει η κύρτωση της κατανομής των ναύλων δηλώνει ότι ανήκει στις λεπτόκυρτες κατανομές.

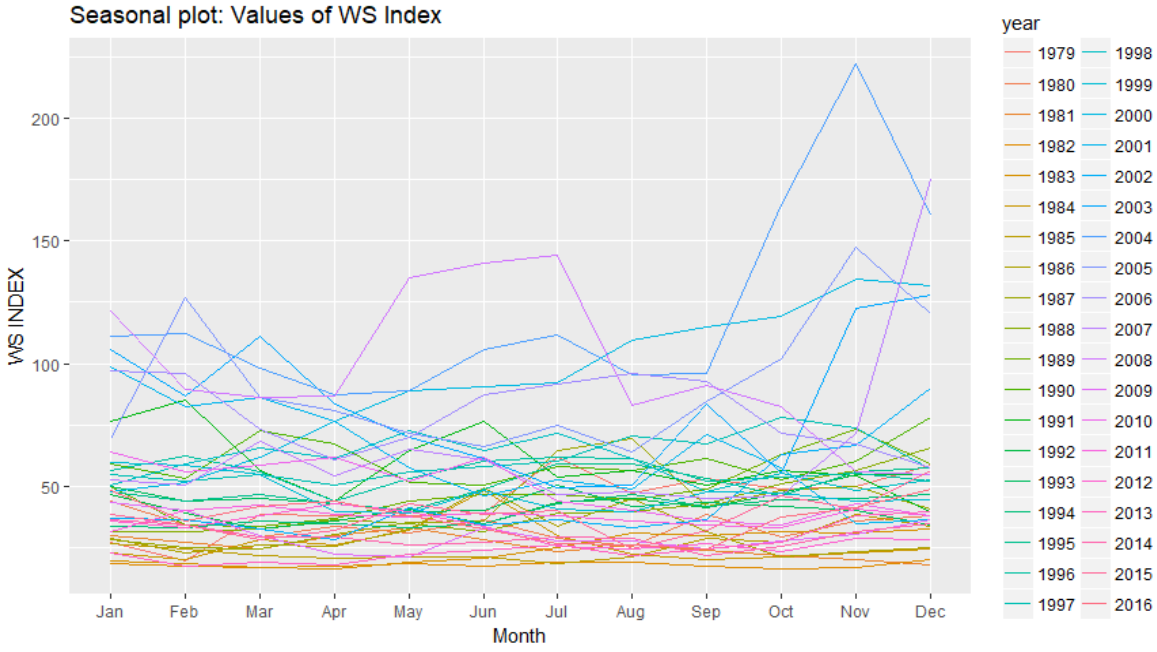
Το Διάγραμμα 3.2 δείχνει την εμπειρική κατανομή των ναύλων μαζί με το ιστόγραμμα τους. Στη περίπτωση αυτή, όπου δεν γνωρίζουμε εξαρχής την συναρτησιακή μορφή της κατανομής του υπο μελέτη χαρακτηριστικού, εφαρμόζουμε την λεγόμενη μη παραμετρική εκτιμήτρια της κατανομής, η οποία μας δίνει μια οπτική εικόνα της κατανομής του υπό μελέτη χαρακτηριστικού και μας επιτρέπει να πάρουμε μια ιδέα για το ποια ενδέχεται να είναι η κατανομή που ακολουθεί (Κοκολάκης & Φουσκάκης, 2009).



<u>Διάγραμμα 3.2</u>: Ιστόγραμμα των ναύλων μαζί με την μη-παραμετρική εκτιμήτρια της κατανομής των ναύλων.

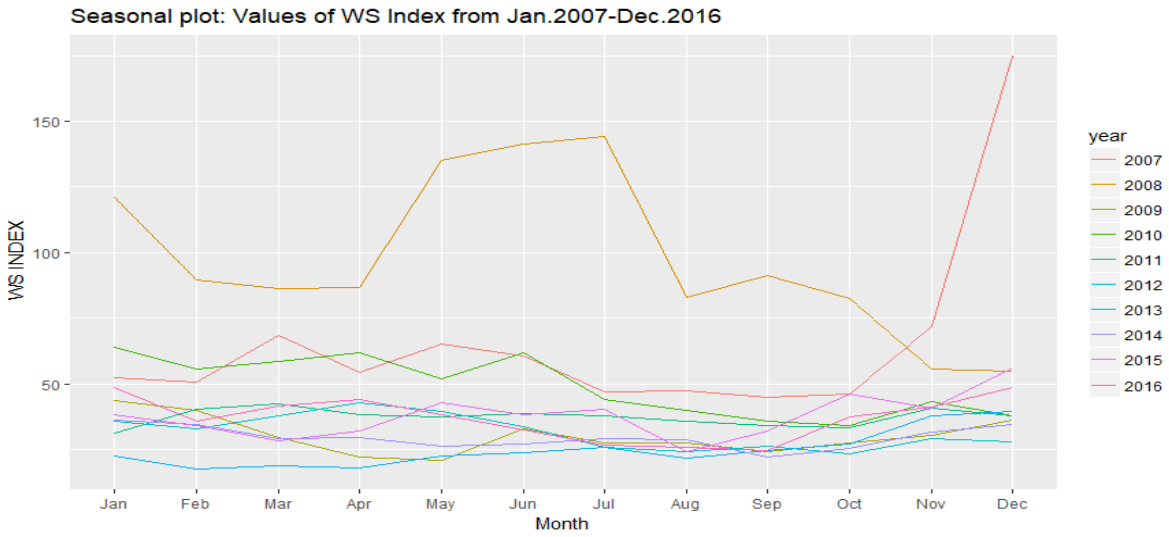
Συνεχίζοντας την ανάλυση μας, θα προσπαθήσουμε να δούμε αν εμφανίζονται φαινόμενα περιοδικότητας στον δείκτη Worldscale. Από τη γραφική παράσταση του Διαγράμματος 3.1, δεν έχουμε εμφανείς υπόνοιες για περιοδικά φαινόμενα όπως αυτά των δεδομένων των διεθνών επιβατών της αεροπορικής εταιρείας Pan Am (βλ. Διάγραμμα 2.3.1). Ωστόσο, για να επιβεβαιώσουμε την παρατήρηση αυτή θα κάνουμε τρία διαγράμματα εποχικότητας. Το πρώτο θα δείχνει συνολικά για τα έτη από το 1979 έως το 2016 εάν εμφανίζονται φαινόμενα εποχικότητας και τα αλλά δυο θα εξετάζουν την περίπτωση που δεν υπάρχει σταθερή εποχικότητα συνολικά για όλα τα χρόνια αλλά εμφανίζεται τα τελευταία 10 και 5 χρόνια αντιστοίχως.

Το Διάγραμμα 3.3 δείχνει στον άξονα τον x τους μήνες από τον Ιανουάριο έως τον Δεκέμβριο και στον άξονα τον y τις τιμές του δείκτη WS, όπου για κάθε έτος απεικονίζονται με διαφορετικό χρώμα. Είναι εμφανές ότι τα δεδομένα μας δεν φαίνεται να παρουσιάζουν συγκεκριμένη και σταθερή περιοδικότητα συνολικά για όλα τα έτη και δεν διακρίνονται ελάχιστες ή μέγιστες τιμές του δείκτη σε συγκεκριμένους μήνες. Αυτό που κεντρίζει την προσοχή είναι η μεγάλη άνοδος των ναύλων για τα έτη 2003-2005 από τον Σεπτέμβριο στον Νοέμβριο καθώς επίσης η άνοδος από τον Οκτώβρη στον Νοέμβρη του 2003 και αντίστοιχα η τεράστια άνοδος από τον Νοέμβρη στον Δεκέμβρη του 2007.



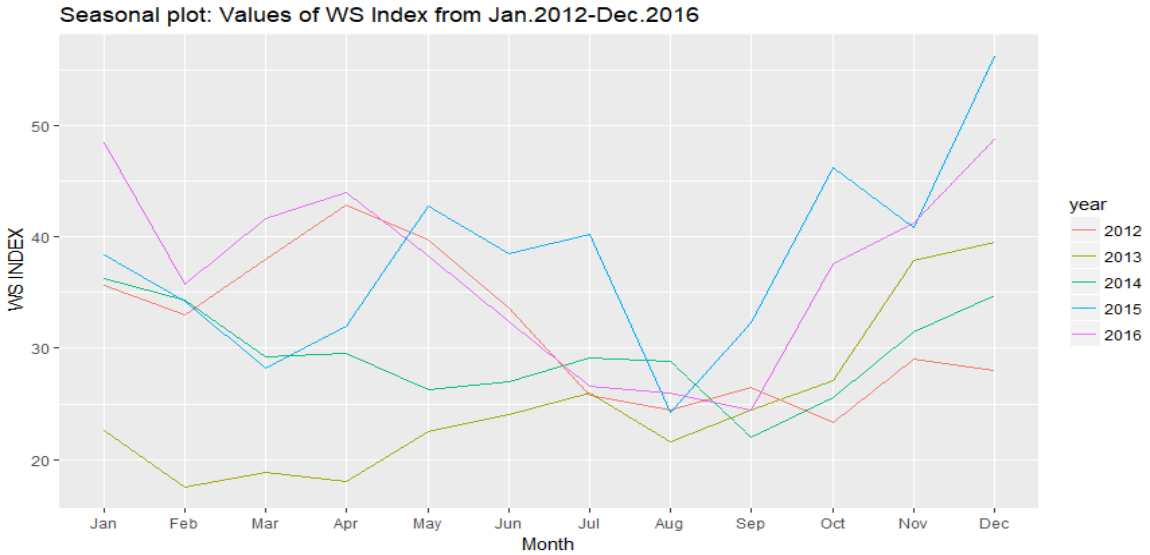
<u>Διάγραμμα 3.3</u>: Διάγραμμα εποχικότητας ναύλων για πλοία VLCC 280.00 DWT για τα έτη 1979 έως 2016.

Το Διάγραμμα 3.4 δείχνει στον άξονα τον x τους μήνες από τον Ιανουάριο έως τον Δεκέμβριο και στον άξονα τον y τις τιμές του δείκτη WS, για την δεκαετία από το 2007 έως το 2016. Αντίστοιχα συμπεράσματα βγάζουμε και σε αυτό το διάγραμμα όπου τα δεδομένα μας δεν φαίνεται να ακολουθούν παρόμοιες συμπεριφορές σε συγκεκριμένους μήνες και συνεπώς δεν μπορούμε να λάβουμε υπόψιν ούτε σε αυτή την περίπτωση περιοδική συνιστώσα.



<u>Διάγραμμα 3.4</u>: Διάγραμμα εποχικότητας ναύλων για πλοία VLCC 280.00 DWT για τα έτη 2007 έως 2016.

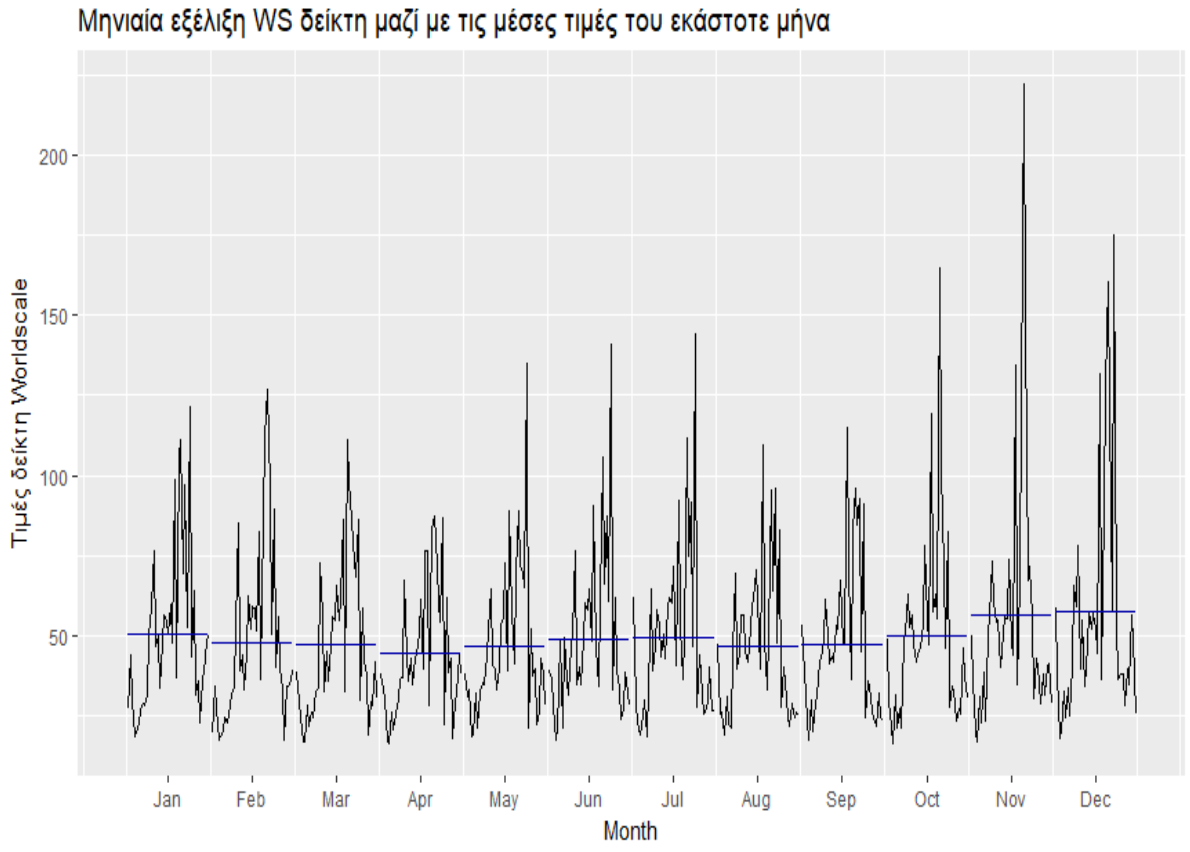
Το Διάγραμμα 3.5 δείχνει στον άξονα τον x τους μήνες από τον Ιανουάριο έως τον Δεκέμβριο και στον άξονα τον y τις τιμές του δείκτη WS, για την πενταετία από το 2012 έως το 2016. Αυτό που παρατηρούμε είναι ότι για την τετραετία από το 2013-2016 έχουμε πάντα άνοδο των ναύλων τον μήνα Οκτώβρη σε σχέση με τις τιμές που είχαμε τον Σεπτέμβρη. Αντίστοιχα, για την ίδια τετραετία οι τιμές των ναύλων τον Δεκέμβρη παρουσιάζουν αύξηση συγκριτικά με τις τιμές τον ναύλων τον προηγούμενο μήνα, Νοέμβρη. Εξαίρεση αποτελεί το έτος 2012 στο οποίο παρουσιάζεται πτώση από τον Σεπτέμβρη στον Οκτώβρη και αντίστοιχα από τον Νοέμβρη στον Δεκέμβρη . Με βάση αυτές τις παρατηρήσεις, ενδέχεται να πρέπει να λάβουμε υπόψιν ότι ο δείκτης ακολουθεί μια περιοδική συμπεριφορά 12 μηνών για τους μήνες Σεπτέμβρη-Οκτώβρη και Νοέμβρη-Δεκέμβρη για τα τελευταία τέσσερα χρόνια.



<u>Διάγραμμα 3.5</u>: Διάγραμμα εποχικότητας ναύλων για πλοία VLCC 280.00 DWT για τα έτη 2012 έως 2016.

Εκτός από τα διαγράμματα εποχικότητας, σημαντική πληροφορία μπορεί να αντληθεί και με τη βοήθεια του διαγράμματος monthplot. Με το διάγραμμα αυτό μπορούμε να δούμε ξεχωριστά για κάθε μήνα την πορεία των τιμών του δείκτη Worldscale. Τέτοιο παράδειγμα είναι η περίπτωση όπου ενδιαφερόμαστε να παρατηρήσουμε την πορεία των τιμών του δείκτη μόνο τον μήνα Ιανουάριο αλλά για όλα τα έτη από το 1979 έως το 2016.

Το Διάγραμμα 3.6 δείχνει στον άξονα τον x τους μήνες από τον Ιανουάριο έως τον Δεκέμβριο και στον άξονα τον y τις τιμές του δείκτη WS, για κάθε μήνα ξεχωριστά. Δηλαδή στον μήνα Ιανουάριο φαίνονται όλες οι τιμές που έχει πάρει ο δείκτης WS συνολικά για τα έτη 1979-2016 (μαύρη γραμμή). Αντίστοιχα στον μήνα Φεβρουάριο κ.ο.κ. Οι οριζόντιες γραμμές μας δείχνουν τον μέσο όρο του WS δείκτη τους εκάστοτε μήνες. Αυτό που παρατηρούμε είναι ότι γενικά δεν έχουμε μεγάλες διαφορές στις τιμές των ναύλων κατά μέσο όρο ανά μήνα (αυτό εξηγεί και το φαινόμενο των κύκλων που υπάρχει στη ναυτιλία όπου όταν παρατηρείται μεγάλη άνοδος των ναύλων για κάποια περίοδο, μετά υπάρχει κάθοδος των ναύλων έτσι ώστε να λειτουργεί το σύστημα εξισορροπητικά). Τέλος, βλέπουμε ότι οι μήνες Νοέμβρης-Δεκέμβρης παρουσιάζουν κατά μέσο όρο τις μεγαλύτερες τιμές των ναύλων σε σχέση με τους υπόλοιπους μήνες.

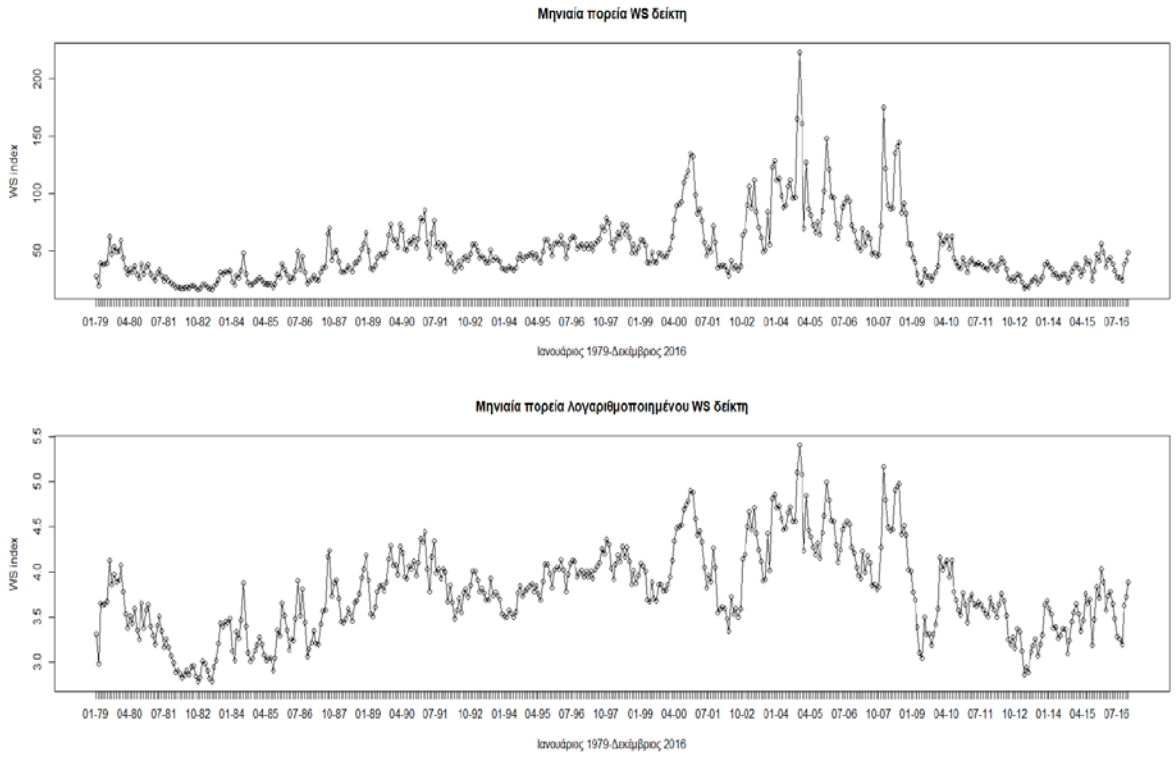


<u>Διάγραμμα 3.6</u>: Διάγραμμα εξέλιξης ναύλων ανά μήνα για πλοία VLCC 280.00 DWT για τα έτη 1979 έως 2016, μαζί με τις μέσες τιμές τους.

Η τάση των ναύλων του Διαγράμματος 3.1 μπορεί να μη δείχνει ξεκάθαρη ανοδική η καθοδική πορεία ωστόσο μοιάζει ασταθής κατά την πάροδο του χρόνου με σταθερή καθοδική πορεία στην τάση από το 2010 και μετά σε σχέση με την περίοδο 2000-2009. Η ασταθής συμπεριφορά τόσο της τάσης όσο και της διακύμανσης των δεδομένων μας, μας προϊδεάζουν για μη στασιμότητα.

Μια συνηθισμένη προσέγγιση που χρησιμοποιείται στην οικονομετρία όταν αναλύονται και μοντελοποιούνται μεγέθη που περιγράφουν τιμές, όπως για παράδειγμα χρηματικά ποσά, είναι να μετασχηματίζονται τα αρχικά ακατέργαστα δεδομένα παίρνοντας λογαρίθμους. Με τον μετασχηματισμό αυτό αφενός γίνονται πιο ξεκάθαρα τα μοτίβα που εμφανίζονται σε ένα γράφημα χρονοσειράς όπως αυτό του Διαγράμματος 3.1 και αφετέρου μεταφέρεται όλη την πληροφορία για το ιστορικό των δεδομένων μας καθώς μας δίνονται τα ίδια συμπεράσματα που θα μας έδιναν τα δεδομένα μας πριν πάρουμε λογαρίθμους.

Με βάση αυτήν την παρατήρηση θα συνεχίσουμε από εδώ και πέρα την ανάλυση μας για τις λογαριθμοποιημένες τιμές των δεδομένων μας του δείκτη Worldscale $\log(x_1), \log(x_2), \ldots, \log(x_{456})$. Το πρώτο γράφημα του Διαγράμματος 3.7 δείχνει τις τιμές του WS ανά μήνα για τα έτη 1979-2016 πριν την λογαρίθμηση ενώ το δεύτερο γράφημα του Διαγράμματος 3.7 δείχνει τις λογαριθμοποιημένες μηνιαίες τιμές του WS για τα αντίστοιχα έτη .



<u>Διάγραμμα 3.7</u>: Σύγκριση διαγραμμάτων των αρχικών τιμών ναύλων ανά μήνα με τις λογαριθμοποιημένες *τιμές των ναύλων ανά μήνα, για πλοία VLCC 280.00 DWT στα έτη 1979 έως 2016.*

Είναι εμφανές από το Διάγραμμα 3.7 ότι ενώ στο πρώτο γράφημα οι διαδοχικές μηνιαίες τιμές είναι δυσδιάκριτες (φαίνονται αρκετά κοντά η μια στην άλλη), στο δεύτερο γράφημα του ίδιου διαγράμματος είναι πιο ευδιάκριτα τα μοτίβα που επικρατούν μεταξύ των διαδοχικών μηνιαίων τιμών του δείκτη.

Εκτός από τα Διαγράμματα 3.1 και 3.7 που μας προϊδεάζουν για μη στασιμότητα στα δεδομένα μας, θα προσπαθήσουμε να ενισχύσουμε την αρχική μας εντύπωση ελέγχοντας την στασιμότητα στα δεδομένα μας χρησιμοποιώντας τα **Augmented Dickey-Fuller (ADF) t-statistic test** και **Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test**.

Οι συγκεκριμένοι έλεγχοι βρίσκονται στην βιβλιοθήκη tseries (Trapletti, et al., 2018) της R και εκτελούνται μέσω των εντολών adf.test και kpss.test αντίστοιχα.

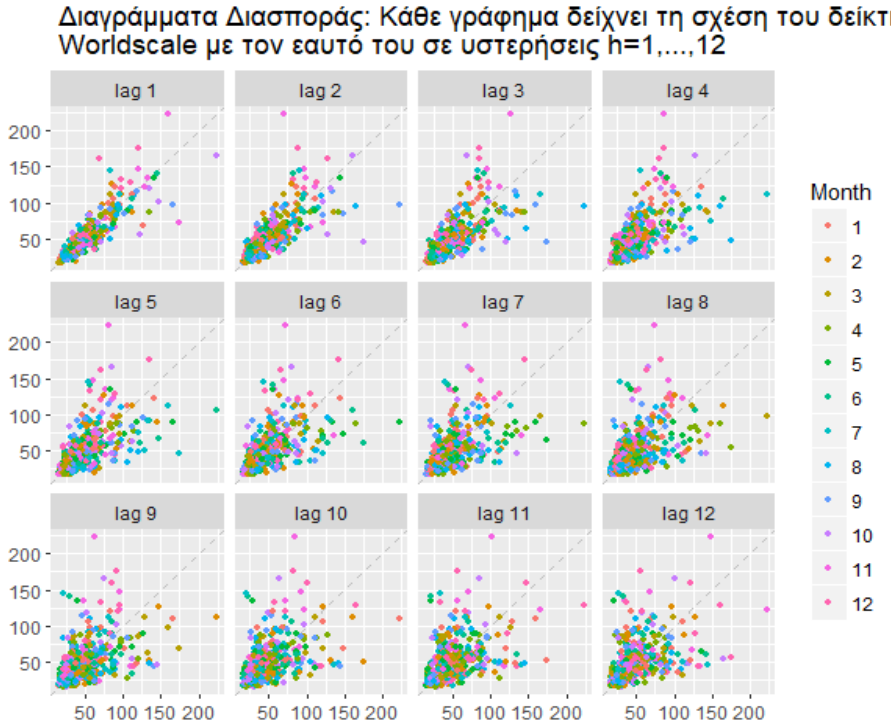
Ο έλεγχος ADF έχει ως μηδενική υπόθεση την ύπαρξη τουλάχιστον μιας ρίζας στο AR πολυώνυμο που βρίσκεται στο σύνορο ή εντός του μοναδιαίου κύκλου (που θα μας υποδείκνυε μη-στασιμότητα). Συνεπώς, αν θέλουμε να κάνουμε αυτόν τον έλεγχο σε επίπεδο σημαντικότητας 5%, μια μικρή τιμή της p-value θα μας οδηγούσε στην απόρριψη της αρχικής μας υπόθεσης, που θα σήμαινε ότι τα δεδομένα μας παρουσιάζουν στασιμότητα. Σε αντίθεση με τον ADF έλεγχο, ο έλεγχος KPSS έχει ως μηδενική υπόθεση ότι τα δεδομένα είναι στάσιμα γύρω από μια ντετερμινιστική συνιστώσα τάσης. Κατά συνέπεια, η απόρριψη της μηδενικής υπόθεσης σε επίπεδο σημαντικότητας 5%, θα υποδήλωνε ότι τα δεδομένα μας δεν παρουσιάζουν στασιμότητα γύρω από μια ντετερμινιστική τάση.

Ο Πίνακας 3.2 μας δείχνει τα αποτελέσματα που πήραμε από την R πραγματοποιώντας τους παραπάνω ελέγχους στα δεδομένα μας. Η μεγάλη p-value που μας δίνει το ADF test του Πίνακα 3.2 μας οδηγεί στο συμπέρασμα ότι δεν έχουμε αρκετές υπόνοιες για να απορρίψουμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας 5%. Όσο αφορά το KPSS test, η p-value (p-value<0.05) που πήραμε, μας οδηγεί στην απόρριψη της μηδενικής υπόθεσης ότι τα δεδομένα μας είναι στάσιμα γύρω από μια ντετερμινιστική συνιστώσα τάσης. Όπως φαίνεται και οι δύο αυτοί έλεγχοι καταλήγουν στο συμπέρασμα ότι τα δεδομένα μας δεν προέρχονται από στάσιμη διαδικασία.

Πίνακας 3.2

Έλεγχος στασιμότητας των δεδομένων			
Τεστ για στασιμότητα	**Μηδενική υπόθεση**	**test-value**	**p-value**
Augmented Dickey-Fuller Test	H_0: Υπάρχει ρίζα του AR πολυωνύμου που βρίσκεται στο σύνορο ή εντός του μοναδιαίου κύκλου (υποδεικνύει μη-στασιμότητα)	-3,29	0,073
Kwiatkowski-Phillips-Schmidt-Shin Test	H_0: Η χρονοσειρά είναι στάσιμη γύρω από μια ντετερμινιστική συνιστώσα τάσης (υποδεικνύει στασιμότητα)	1,93	0,01

Όπως αναφέραμε και στην ενότητα 2.12.1 η δημιουργία των διαγραμμάτων διασποράς (scatter plots) μας βοηθάει να αναγνωρίσουμε την συναρτησιακή μορφή της προς μελέτης μεταβλητής σε προηγούμενες χρονικές περιόδους. Κάθε γράφημα του Διαγράμματος 3.8 δείχνει τα x_t σε σχέση με τα x_{t-h} για διάφορες τιμές της υστέρησης h. Πιο συγκεκριμένα, το πρώτο γράφημα του Διαγράμματος 3.8 δείχνει την συναρτησιακή μορφή που συνδέει κάθε παρατήρηση των δεδομένων μας x_t με την παρατήρηση x_{t-1} του προηγούμενου μήνα. Δηλαδή απεικονίζονται όλα τα σημεία $(x_2, x_1), (x_3, x_2), \ldots, (x_{456}, x_{455})$. Αντίστοιχα, το δεύτερο γράφημα του Διαγράμματος 3.8 δείχνει την συναρτησιακή μορφή της x_t με την παρατήρηση x_{t-2} των δύο προηγούμενων μηνών. Δηλαδή απεικονίζονται όλα τα σημεία $(x_3, x_1), (x_4, x_2), \ldots, (x_{456}, x_{454})$. Η διαδικασία αυτή συνεχίζεται μέχρι 12 υστερήσεις, δηλαδή μέχρι τη διάρκεια ενός έτους. Παρατηρώντας το Διάγραμμα 3.8 είναι ξεκάθαρη η πολύ έντονη γραμμική συσχέτιση των δεδομένων μας σε υστέρηση 1 (lag=1). Δηλαδή κάθε τιμή του δείκτη WS είναι άμεσα και γραμμικά συνδεδεμένη με την τιμή του ακριβώς προηγούμενου μήνα. Επιπροσθέτως, είναι εμφανές ότι η γραμμική αυτή σχέση χάνεται όσο μεγαλώνουν οι υστερήσεις, δηλαδή όσο μεγαλώνουν οι χρονικές στιγμές του παρελθόντος που κοιτάμε πίσω. Πιο συγκεκριμένα, από την υστέρηση h=5 και μετά παρατηρούμε την μεγάλη εξασθένιση της γραμμικής συσχέτισης που συναντούμε στα δεδομένα μας. Για τον λόγο αυτό, σε οποιαδήποτε προσαρμογή μοντέλου προχωρήσουμε παρακάτω, σε περίπτωση που καταλήξουμε στο συμπέρασμα ότι οποιαδήποτε υστέρηση $h \geq 5$ είναι στατιστικά σημαντική για τις μελλοντικές προβλέψεις μας, δεν θα τις συμπεριλάβουμε υπόψιν.



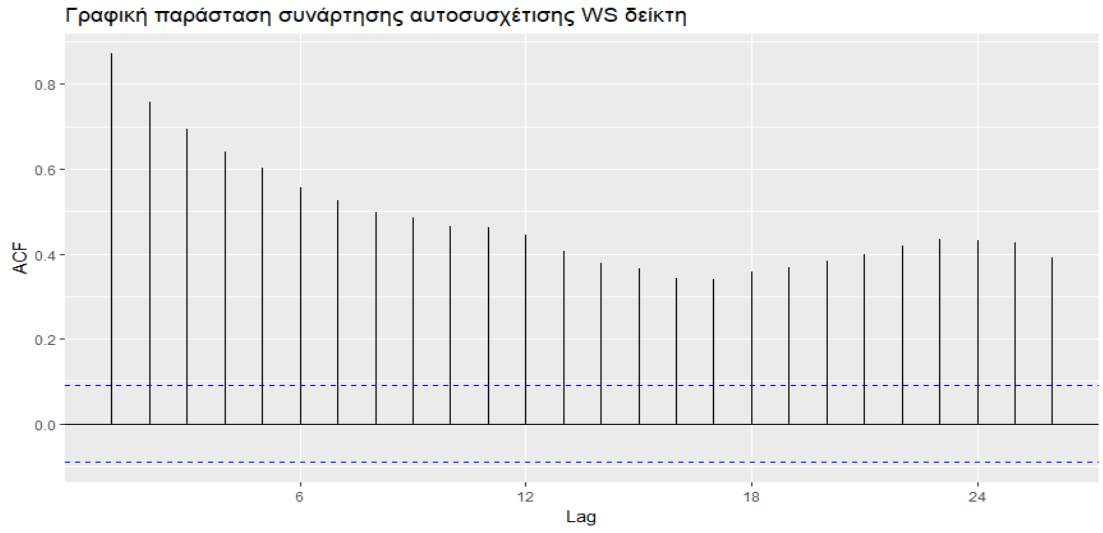
<u>Διάγραμμα 3.8</u>: Διάγραμμα διασποράς (scatter plot) των ναύλων για πλοία VLCC 280.00 DWT για τα έτη 1979 έως 2016 σε υστερήσεις h=1,2,…,12.

Εκτός από τους στατιστικούς ελέγχους ADF και KPSS που πραγματοποιήσαμε προκειμένου να ελέγξουμε την στασιμότητα και τα Διαγράμματα 3.1-3.5, έχουμε αναφέρει στο Κεφάλαιο 2 εξίσου την σημαντικότητα των διαγραμμάτων ACF και $PACF$. Με τη βοήθεια των δειγματικών ACF, $PACF$ θα μπορούσαμε να πάρουμε επίσης μια εικόνα για το αν υπάρχει τάση ή εποχικότητα στα δεδομένα μας και κατά συνέπεια αν προέρχονται από μη στάσιμη διαδικασία.

Το Διάγραμμα 3.9 δείχνει τις τιμές των δειγματικών αυτοσυσχετίσεων για τα αρχικά λογαριθμημένα δεδομένα του WS δείκτη σε υστερήσεις (lags) 0,1,…,26. Οι δύο διακεκομμένες δείχνουν το 95% διάστημα εμπιστοσύνης για τον έλεγχο:

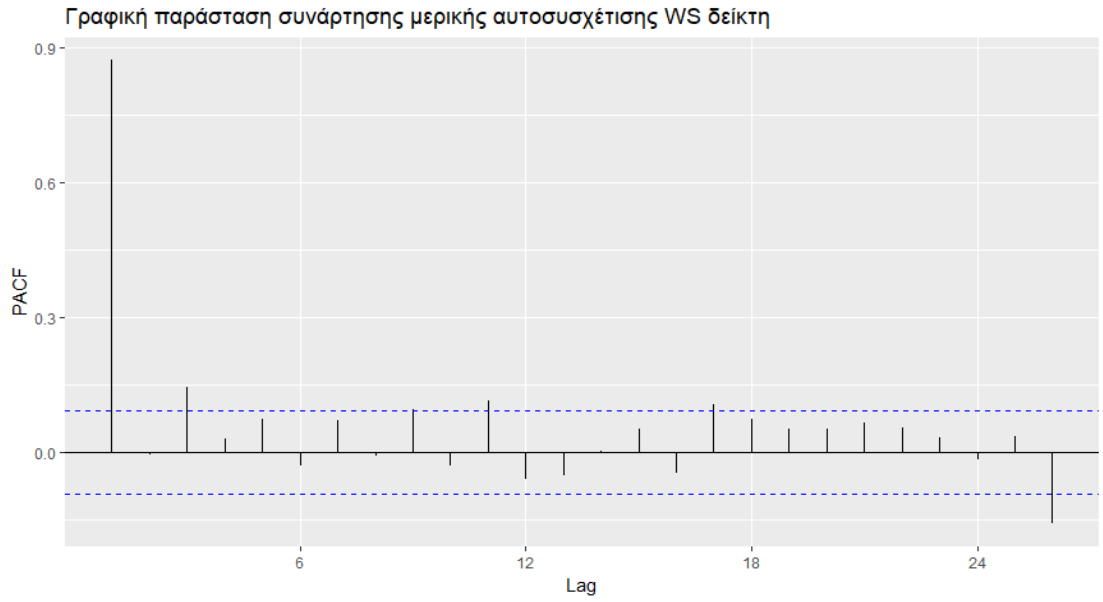
$$H_0 \ : \ \rho = 0$$
$$H_1 \ : \ \rho \neq 0$$

Σε περίπτωση που τα δεδομένα μας προέρχονταν από στάσιμη διαδικασία, το γράφημα ACF θα έπρεπε να φθίνει εκθετικά στο μηδέν κάτι που δεν συμβαίνει στην προκειμένη περίπτωση. Στο Διάγραμμα 3.9 φαίνεται ότι η ACF ακολουθεί φθίνουσα πορεία αλλά αυτό γίνεται με πολύ αργό ρυθμό. Ακόμη είναι εμφανής μια ασταθής αλλά περιοδική συμπεριφορά 12 μηνών (βλ. lag=12 και lag=24), δηλαδή κατά τη διάρκεια ενός έτους, γεγονός που επιβεβαιώνει και την αρχική μας παρατήρηση για τους μήνες Οκτώβρη-Σεπτέμβρη και Νοέμβρη-Δεκέμβρη από το γράφημα εποχικότητας του Διαγράμματος 3.5.



<u>Διάγραμμα 3.9</u>: Διάγραμμα δειγματικής αυτοσυσχέτισης ACF του δείκτη log(WS) για τα έτη 1979 έως 2016.

Το Διάγραμμα 3.10 δείχνει τις τιμές των δειγματικών μερικών αυτοσυσχετίσεων για τα αρχικά δεδομένα του λογαριθμημένου WS δείκτη σε υστερήσεις (lags) 0,1,…,26. Από το Διάγραμμα 3.10 φαίνεται ότι τα δεδομένα μας παρουσιάζουν μεγάλη συσχέτιση για την υστέρηση 1 γεγονός που μας επιβεβαιώνει το συμπέρασμα που βγάλαμε και από το Διάγραμμα διασποράς 3.8.



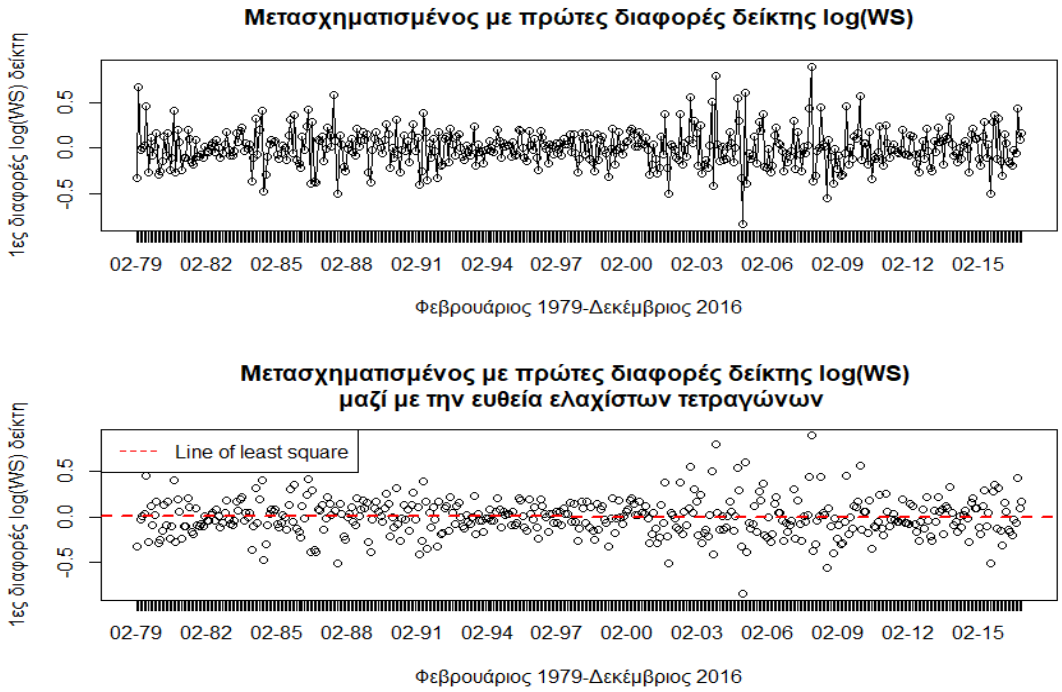
<u>Διάγραμμα 3.10</u>: Διάγραμμα δειγματικής μερικής αυτοσυσχέτισης PACF του δείκτη log(WS) για τα έτη 1979 έως 2016.

Συγκρίνοντας και τα δύο διαγράμματα ενισχύουμε την αρχική μας άποψη ότι τα δεδομένα μας δεν παρουσιάζουν στασιμότητα.

Έχοντας λάβει υπόψιν τα παραπάνω, αρχικά δοκιμάζουμε να μετασχηματίσουμε τα δεδομένα μας παίρνοντας πρώτες διαφορές. Να αναφέρουμε ότι κατά τον μετασχηματισμό των δεδομένων μας σε πρώτες διαφορές, το πλήθος των παρατηρήσεων μας μειώνεται κατά ένα και χάνεται η πρώτη παρατήρηση. Γενικά, στις d φορές που θα δημιουργήσουμε διαφορές χάνονται οι d πρώτες παρατηρήσεις.

Όπως ήδη έχουμε αναφέρει στην αρχή του κεφαλαίου έχουμε συμβολίσει με $x_1, x_2, ..., x_{456}$ τα δεδομένα μας. Μετασχηματίζοντας όμως τα δεδομένα μας με λογαρίθμους και δουλεύοντας πλέον με το σύνολο $\log(x_1), \log(x_2), ..., \log(x_{456})$ οι πρώτες διαφορές που παίρνουμε είναι οι εξής: $\log(x_i) - \log(x_{i-1}), \ \forall\, i > 1.$

Το Διάγραμμα 3.11 δείχνει τα νέα δεδομένα που προκύπτουν έπειτα από τον μετασχηματισμό με πρώτες διαφορές και η κόκκινη διακεκομμένη γραμμή είναι η ευθεία ελαχίστων τετραγώνων που δείχνει την κατεύθυνσή στην οποία κυμαίνονται έτσι ώστε να αποφανθούμε γραφικά για το αν υπάρχουν τάσεις ή καταφέραμε να τα κάνουμε στάσιμα. Εκ πρώτης όψεως τα δεδομένα δείχνουν να ταλαντεύονται γύρω από το μηδέν, χωρίς να παρουσιάζουν αύξηση ή μείωση στην τάση τους. Ωστόσο, η διασπορά τους δεν μοιάζει σταθερή κατά την πάροδο του χρόνου αφού φαίνεται να παρουσιάζει αυξομειώσεις (την περίοδο 1993-2000 έχουμε μικρότερη διασπορά σε σχέση με την περίοδο 2002-2008).



<u>Διάγραμμα 3.11</u>: Μετασχηματισμένος τιμές ναύλων με πρώτες διαφορές μαζί με την ευθεία ελαχίστων τετραγώνων.

Έχοντας μετασχηματίσει τη σειρά μας παίρνοντας πρώτες διαφορές και παρατηρήσει από το Διάγραμμα 3.11 ότι υπάρχουν υπόνοιες για στασιμότητα ως προς τη μέση τιμή, πραγματοποιούμε εκ νέου τους ελέγχους στασιμότητας ADF και KPSS έτσι ώστε να επιβεβαιώσουμε αυτό το συμπέρασμα.

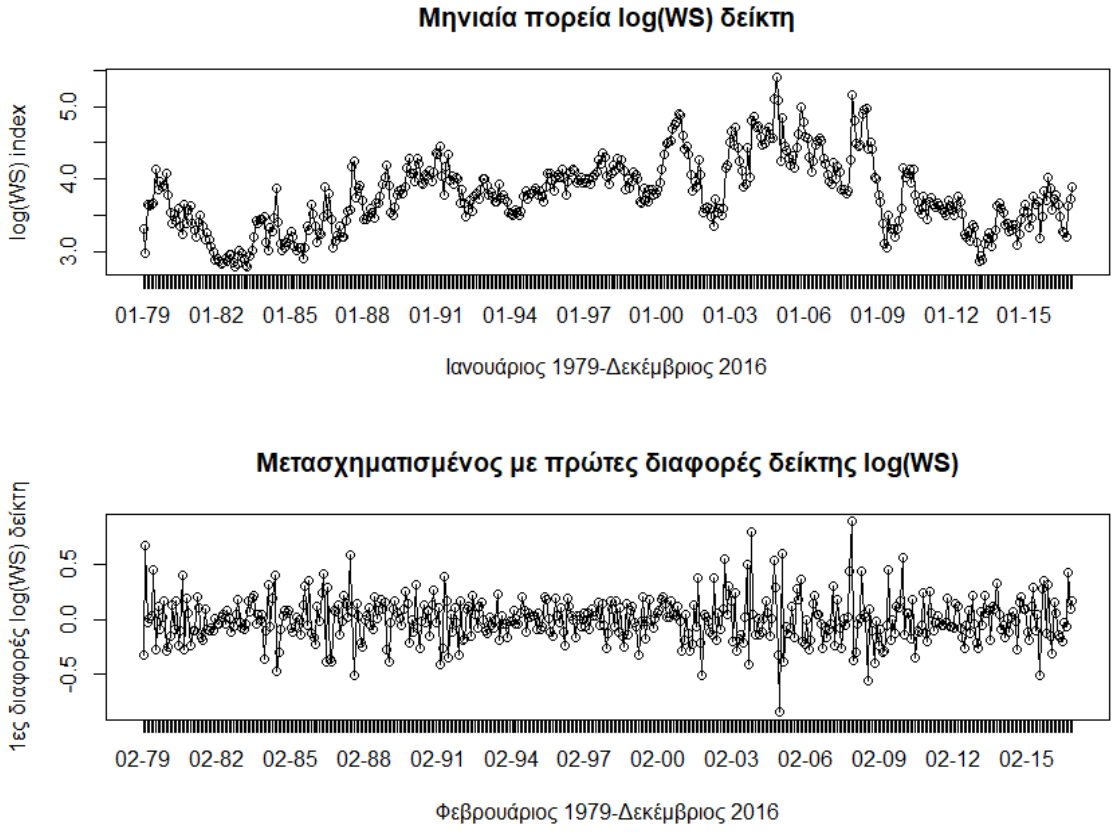
Ο Πίνακας 3.3 μας δείχνει τα αποτελέσματα που πήραμε από την R πραγματοποιώντας τους παραπάνω ελέγχους στασιμότητας (ADF, KPSS) για τα νέα μετασχηματισμένα δεδομένα μας με πρώτες διαφορές.

Πίνακας 3.3

Έλεγχος στασιμότητας του μετασχηματισμένου log(WS) δείκτη με πρώτες διαφορές			
Τεστ για στασιμότητα	**Μηδενική υπόθεση**	**test-value**	**p-value**
Augmented Dickey-Fuller Test	H_0: Υπάρχει ρίζα του AR πολυωνύμου που βρίσκεται στο σύνορο ή εντός του μοναδιαίου κύκλου (υποδεικνύει μη-στασιμότητα)	-9,18	0,01
Kwiatkowski-Phillips-Schmidt-Shin Test	H_0: Η χρονοσειρά είναι στάσιμη γύρω από μια ντετερμινιστική συνιστώσα τάσης (υποδεικνύει στασιμότητα)	0,02	0,1

Η μικρή p-value που μας δίνει το ADF test του Πίνακα 3.3 μας οδηγεί στην απόρριψη της μηδενικής υπόθεσης για μη-στασιμότητα σε επίπεδο σημαντικότητας 5%. Όσο αφορά το KPSS test, η τόσο μεγάλη τιμή της p-value που πήραμε, μας οδηγεί στο συμπέρασμα ότι δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση περί στασιμότητας των δεδομένων γύρω από μια ντετερμινιστική τάση, πάλι σε επίπεδο σημαντικότητας 5%. Συμπεραίνουμε λοιπόν ότι και οι δυο αυτοί έλεγχοι καταλήγουν στο συμπέρασμα ότι τα δεδομένα μας που έχουν προκύψει από τις πρώτες διαφορές είναι πλέον στάσιμα. Εδώ είναι σημαντικό να αναφέρουμε ότι η στασιμότητα προέκυψε χωρίς την συμβολή περιοδικής συνιστώσας. Σε περίπτωση που η περιοδική συνιστώσα ήταν απαραίτητη, τότε η στασιμότητα δεν θα μπορούσε να προκύψει με τις πρώτες διαφορές. Συνεπώς, δεν χρειάζεται να συμπεριλάβουμε περιοδικότητα στο μοντέλο μας. Ωστόσο, όπως είδαμε από το Διάγραμμα 3.11 η διασπορά των δεδομένων μας δεν μοιάζει σταθερή κάτι που αδυνατούν να υπολογίσουν στην συγκεκριμένη περίπτωση οι έλεγχοι ADF και KPSS.

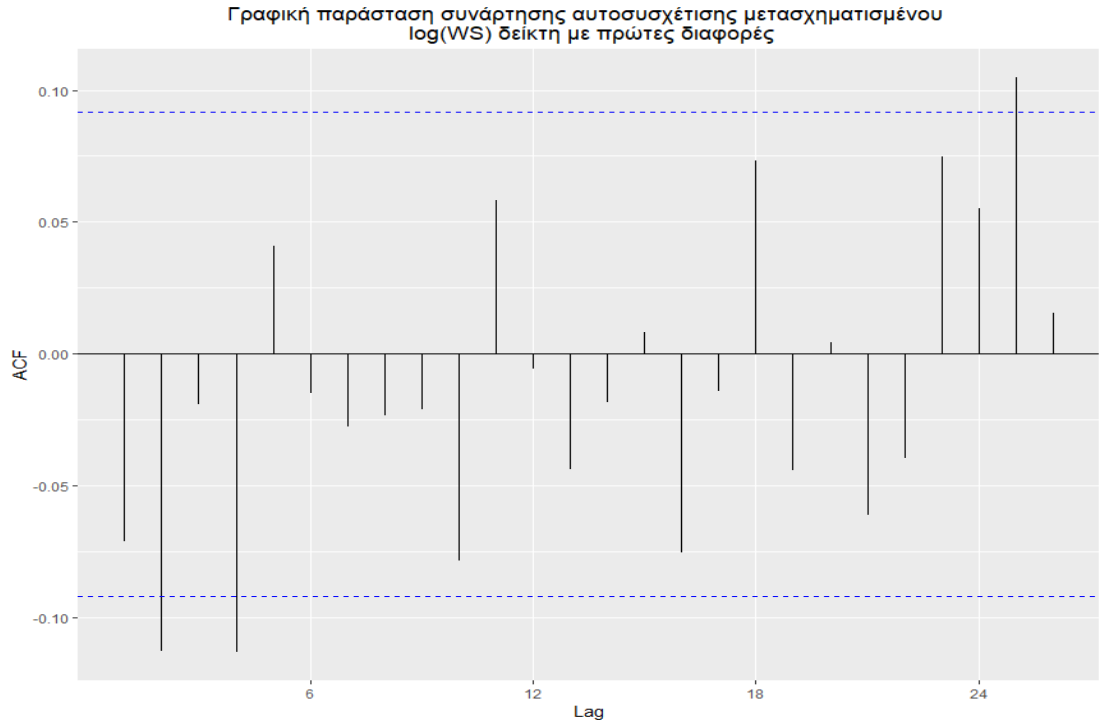
Το πρώτο σχήμα του Διαγράμματος 3.12 δείχνει τα αρχικά δεδομένα του λογαριθμοποιημένου δείκτη WS και το δεύτερο σχήμα τα δεδομένα που πήραμε έπειτα από τον μετασχηματισμό των πρώτων διαφορών. Με αυτή την σύγκριση μπορούμε να δούμε πόσο ξεκάθαρα φαίνεται στο δεύτερο διάγραμμα ότι τα δεδομένα μας μοιάζουν να είναι στάσιμα γύρω από το μηδέν.



<u>Διάγραμμα 3.12</u>: Σύγκριση γραφημάτων αρχικών μηνιαίων τιμών δείκτη log(WS) με τις τιμές του μετασχηματισμένου με πρώτες διαφορές δείκτη log(WS).

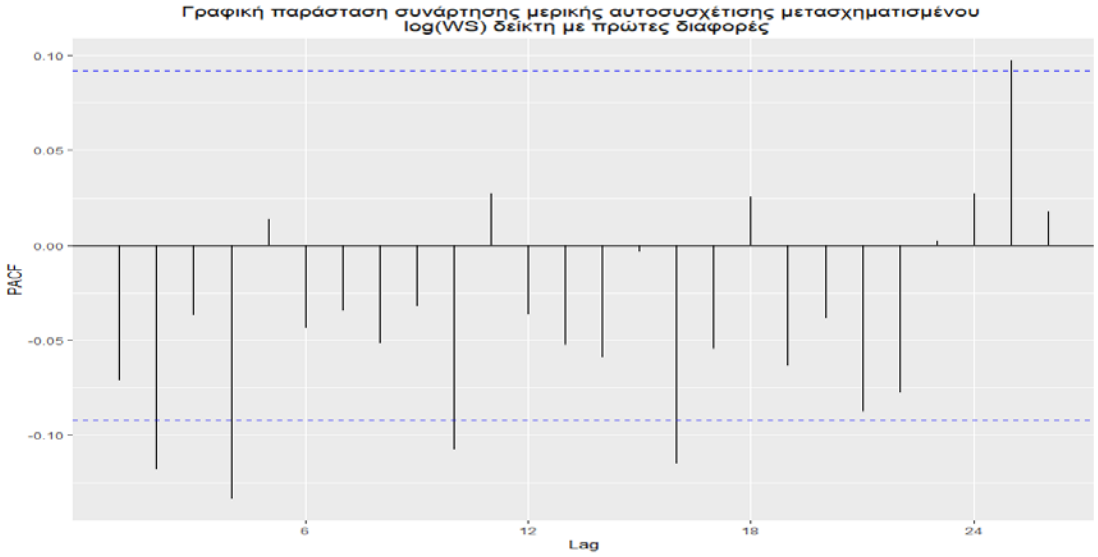
Το επόμενο βήμα που θέλουμε να κάνουμε, είναι να πάρουμε μια πρώτη εκτίμηση για τις τάξεις των p, q που θα χρησιμοποιήσουμε. Όπως έχουμε αναφέρει σε προηγούμενες ενότητες αυτό γίνεται με τη βοήθεια των γραφημάτων ACF, $PACF$. Για την εκτίμηση του πλήθους των συντελεστών p, q που θα λάβουμε υπόψιν φτιάχνουμε τα γραφήματα ACF, $PACF$ που απεικονίζονται κατά αντιστοιχία στο Διάγραμμα 3.13 και 3.14.

Το Διάγραμμα 3.13 δείχνει τον δεύτερο και τον τέταρτο όρο της ACF να πέφτει εκτός των ορίων και συνεπώς να θεωρούνται διάφοροι του μηδενός ενώ όλοι οι επόμενοι όροι στις υστερήσεις $h = 5, \ldots, 26$ μπορούν να θεωρηθούν στατιστικά ασήμαντοι. Με βάση αυτές τις παρατηρήσεις συμπεραίνουμε ότι η αρχικές πιθανές τιμές που θα βάζαμε στην τάξη q είναι ή το 2 ή το 4.



<u>Διάγραμμα 3.13</u>: Διάγραμμα δειγματικής αυτοσυσχέτισης ACF του μετασχηματισμένου με πρώτες διαφορές δείκτη log(WS).

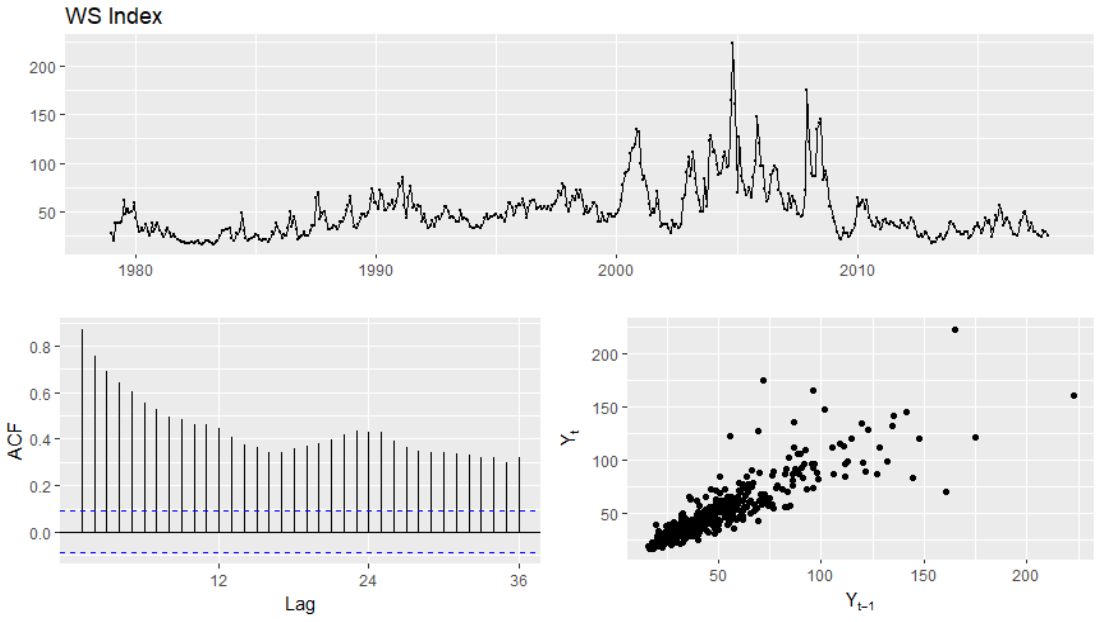
Από το Διάγραμμα 3.14 των δειγματικών μερικών αυτοσυσχετίσεων η εικόνα που παίρνουμε είναι λίγο ασαφής. Βλέπουμε ότι ο δεύτερος όρος της PACF και ο τέταρτος όρος είναι στατιστικά σημαντικοί. Ακόμη βλέπουμε μερικούς όρους μετά την υστέρηση 9 να πέφτουν εκτός ορίων και πιθανότατα να θεωρούνται και αυτοί στατιστικά σημαντικοί. Ωστόσο, όπως αναφέραμε και προηγουμένως, εξαιτίας του γεγονότος ότι για κάθε υστέρηση $h \geq 5$ χάνεται η γραμμική εξάρτηση των δεδομένων μας, δε θα την συμπεριλάβουμε υπόψιν. Με βάση αυτές τις παρατηρήσεις περιμένουμε ότι η τάξη p θα είναι ή το 2 ή το 4.



<u>Διάγραμμα 3.14</u>: Διάγραμμα δειγματικής μερικής αυτοσυσχέτισης PACF του μετασχηματισμένου με πρώτες διαφορές δείκτη log(WS).

Επειδή δεν έχουμε ξεκάθαρη εικόνα για την επιλογή των παραμέτρων p, q προτιμούμε να ξεκινήσουμε με το πιο γενικό μοντέλο $ARIMA(4,1,4)$ και στη συνέχεια να ελέγξουμε όλα τα εμφωλευμένα σε αυτό μοντέλα και να τα συγκρίνουμε μεταξύ τους.

Το Διάγραμμα 3.15 δείχνει συγκεντρωτικά το αρχικό σύνολο δεδομένων μας με τις τιμές του δείκτη WS μαζί με την δειγματική ACF και το διάγραμμα διασποράς σε υστέρηση h=1. Το παρακάτω διάγραμμα λαμβάνεται από την βιβλιοθήκη forecast μέσω της εντολής ggtsdisplay, επιλέγοντας ως όρισμα plot.type='scatter'.



<u>Διάγραμμα 3.15</u>: Συνδυασμένο διάγραμμα που περιλαμβάνει τις αρχικές μηνιαίες τιμές του WS δείκτη μαζί με την δειγματική αυτοσυσχέτιση ACF και το διάγραμμα διασποράς σε υστέρηση 1.

Συνεπώς, προσαρμόζοντας όλα τα προτεινόμενα μοντέλα με την βοήθεια της R, λαμβάνουμε τα εξής αποτελέσματα:

$ARIMA(4,1,4)$:

```
> Arima(Worldscale,order = c(4,1,4))
Series: Worldscale
ARIMA(4,1,4)

Coefficients:
          ar1      ar2     ar3     ar4     ma1     ma2      ma3      ma4
      -0.7287  -0.3275  0.4107  0.5921  0.6608  0.1246  -0.6455  -0.8405
s.e.   0.0983   0.0793  0.0829  0.0894  0.0800  0.0491   0.0505   0.0783

sigma^2 estimated as 0.03749:  log likelihood=103.08
AIC=-188.16   AICc=-187.76   BIC=-151.08
```

Το μοντέλο $ARIMA(4,1,4)$ δίνει μικρά τυπικά σφάλματα, με όλους τους συντελεστές να βγαίνουν στατιστικά σημαντικοί. Παρότι έχουμε επιθυμητά αποτελέσματα και είναι υποψήφιο μοντέλο για προσαρμογή, θα προχωρήσουμε και στην προσαρμογή των υπόλοιπων μοντέλων.

Η προσαρμογή του $ARIMA(3,1,4)$ δίνει:

```
> Arima(Worldscale,order = c(3,1,4))
Series: Worldscale
ARIMA(3,1,4)

Coefficients:
          ar1     ar2     ar3     ma1      ma2      ma3      ma4
      -0.2504  0.6295  0.1709  0.1447  -0.8273  -0.1719  -0.0232
s.e.   0.8920  0.1271  0.6070  0.8911   0.1573   0.7705   0.0840

sigma^2 estimated as 0.03857:  log likelihood=98.18
AIC=-180.36   AICc=-180.04   BIC=-147.4
```

Για το μοντέλο $ARIMA(3,1,4)$ έχουμε αρκετά μεγάλα τυπικά σφάλματα στον πρώτο και τρίτο όρο των AR και MA αντίστοιχα. Ταυτόχρονα οι τιμές των κριτηρίων AIC, AICc και BIC μεγάλωσαν σε σχέση με το προηγούμενο. Συνεπώς, η ελάττωση της τάξης p κατά μια μονάδα, δεν οδήγησε σε ικανοποιητικότερα αποτελέσματα.

Η προσαρμογή του $ARIMA(2,1,4)$ δίνει:

```
> Arima(Worldscale,order = c(2,1,4))
Series: Worldscale
ARIMA(2,1,4)

Coefficients:
          ar1     ar2      ma1      ma2     ma3     ma4
      -0.0027  0.6440  -0.1029  -0.8190  0.0454  -0.026
s.e.   0.1248  0.1239   0.1321   0.1375  0.0694   0.072

sigma^2 estimated as 0.0385:  log likelihood=98.13
AIC=-182.26   AICc=-182.01   BIC=-153.42
```

Εδώ φαίνεται ότι έχουμε καλή προσαρμογή, όπως και στο πρώτο μοντέλο. Ωστόσο, ενώ υπάρχει αύξηση των κριτηρίων AIC και AICc σε σχέση με το αμέσως προηγούμενο μοντέλο, το κριτήριο BIC έχει ελαττωθεί ελάχιστα.

Συνεχίζουμε με την προσαρμογή του μοντέλου $ARIMA(1,1,4)$:

```
> Arima(worldscale,order = c(1,1,4))
Series: Worldscale
ARIMA(1,1,4)

Coefficients:
          ar1      ma1      ma2      ma3      ma4
      -0.4367   0.3466  -0.2010  -0.1104  -0.1709
s.e.   0.3995   0.3995   0.0604   0.0831   0.0558

sigma^2 estimated as 0.03925:  log likelihood=93.4
AIC=-174.8   AICc=-174.61   BIC=-150.07
```

Το μοντέλο $ARIMA(1,1,4)$ μας δίνει μεγάλα τυπικά σφάλματα στον όρο AR και στον πρώτο όρο MA ενώ ταυτόχρονα οι τιμές των κριτηρίων AIC, AICc και BIC μεγάλωσαν σε σχέση με όλα τα προηγούμενα μοντέλα. Συνεπώς, η ελάττωση της τάξης p κατά μια μονάδα, δεν οδήγησε σε ικανοποιητικότερα αποτελέσματα.

Η προσαρμογή του $ARIMA(4,1,3)$ δίνει:

```
> Arima(Worldscale,order = c(4,1,3))
Series: Worldscale
ARIMA(4,1,3)

Coefficients:
         ar1     ar2      ar3      ar4      ma1      ma2     ma3
      0.1658  0.6849  -0.1109  -0.0210  -0.2722  -0.8448  0.1929
s.e.     NaN  0.0237      NaN   0.0539      NaN   0.0797     NaN

sigma^2 estimated as 0.03859:  log likelihood=98.11
AIC=-180.21   AICc=-179.89   BIC=-147.25
Warning message:
In sqrt(diag(x$var.coef)) : NaNs produced
```

Παρατηρούμε ότι υπάρχει κάποιο πρόβλημα υπολογισμού στα τυπικά σφάλματα του πρώτου και του τρίτου όρου των AR και MA. Συνεπώς, δεν μπορούμε να έχουμε μια σαφή εικόνα και να βγάλουμε συμπεράσματα.

Συνεχίζουμε με την προσαρμογή του $ARIMA(4,1,2)$:

```
> Arima(Worldscale,order = c(4,1,2))
Series: Worldscale
ARIMA(4,1,2)

Coefficients:
          ar1     ar2     ar3      ar4      ma1      ma2
      -0.0552  0.6698  0.0396  -0.0203  -0.0512  -0.8494
s.e.   0.0905  0.0846  0.0538   0.0545   0.0779   0.0768

sigma^2 estimated as 0.03849:  log likelihood=98.15
AIC=-182.3   AICc=-182.05   BIC=-153.45
```

Το μοντέλο $ARIMA(4,1,2)$ μας δίνει όμοια αποτελέσματα με το μοντέλο $ARIMA(2,1,4)$ με μικρότερα τυπικά σφάλματα. Οι τιμές των κριτηρίων AIC, AICc και BIC έχουν ελαττωθεί ελάχιστα σε σχέση με το $ARIMA(2,1,4)$.

Η προσαρμογή του $ARIMA(4,1,1)$ δίνει:

```
> Arima(Worldscale,order = c(4,1,1))
Series: Worldscale
ARIMA(4,1,1)

Coefficients:
          ar1      ar2      ar3      ar4     ma1
      -0.8782  -0.2018  -0.1415  -0.1129  0.8077
s.e.   0.2095   0.0635   0.0666   0.0560  0.2094

sigma^2 estimated as 0.03965:  log likelihood=91.12
AIC=-170.24   AICc=-170.05   BIC=-145.52
```

Για το μοντέλο αυτό έχουμε τις μεγαλύτερες τιμές των κριτηρίων AIC, AICc και BIC σε σχέση με όλα τα προηγούμενα μοντέλα, ενώ ταυτόχρονα παρουσιάζει μεγαλύτερα σφάλματα στον πρώτο όρο του AR.

Η προσαρμογή του $ARIMA(3,1,3)$ δίνει:

```
> Arima(Worldscale,order = c(3,1,3))
Series: Worldscale
ARIMA(3,1,3)

Coefficients:
          ar1     ar2     ar3     ma1      ma2      ma3
      -0.5114  0.6404  0.3470  0.4045  -0.8700  -0.3932
s.e.   0.9132  0.0961  0.6013  0.9249   0.0557   0.7975

sigma^2 estimated as 0.03849:  log likelihood=98.17
AIC=-182.34   AICc=-182.08   BIC=-153.49
```

Σε αυτό το μοντέλο έχουμε πολύ μεγάλα τυπικά σφάλματα στον πρώτο και τον τρίτο όρο του AR και MA μέρους αντίστοιχα παρόλο που πετυχαίνουμε μικρές τιμές των κριτηρίων AIC, AICc και BIC.

Η προσαρμογή του $ARIMA(2,1,3)$ δίνει:

```
> Arima(Worldscale,order = c(2,1,3))
Series: Worldscale
ARIMA(2,1,3)

Coefficients:
          ar1     ar2      ma1      ma2     ma3
      -0.0075  0.6753  -0.0985  -0.8613  0.0483
s.e.   0.1150  0.0804   0.1229   0.0641  0.0668

sigma^2 estimated as 0.03842:  log likelihood=98.06
AIC=-184.13   AICc=-183.94   BIC=-159.41
```

Έχουμε αρκετά ικανοποιητική προσαρμογή με μικρά τυπικά σφάλματα και τις μικρότερες τιμές των κριτηρίων AIC, AICc και BIC μετά από το μοντέλο $ARIMA(4,1,4)$.

Η προσαρμογή του $ARIMA(1,1,3)$ δίνει:

```
> Arima(Worldscale,order = c(1,1,3))
Series: Worldscale
ARIMA(1,1,3)

Coefficients:
          ar1     ma1      ma2      ma3
      -0.8582  0.7743  -0.2434  -0.1001
s.e.   0.0907  0.1001   0.0691   0.0618

sigma^2 estimated as 0.03972:  log likelihood=90.19
AIC=-170.38   AICc=-170.25   BIC=-149.78
```

Παρόμοια σχόλια με το μοντέλο $ARIMA(4,1,1)$. Έχουμε μικρά τυπικά σφάλματα αλλά οι τιμές των κριτηρίων AIC, AICc και BIC έχουν αυξηθεί σε σχέση με τα προηγούμενα μοντέλα.

Η προσαρμογή του $ARIMA(3,1,2)$ δίνει:

```
> Arima(Worldscale,order = c(3,1,2))
Series: Worldscale
ARIMA(3,1,2)

Coefficients:
          ar1     ar2     ar3      ma1      ma2
      -0.0631  0.6717  0.0396  -0.0440  -0.8622
s.e.   0.0828  0.0798  0.0535   0.0687   0.0621

sigma^2 estimated as 0.03842:  log likelihood=98.08
AIC=-184.16   AICc=-183.97   BIC=-159.44
```

Έχουμε όμοια σχόλια με το μοντέλο $ARIMA(2,1,3)$.

Η προσαρμογή του $ARIMA(3,1,1)$ δίνει:

```
> Arima(Worldscale,order = c(3,1,1))
Series: Worldscale
ARIMA(3,1,1)

Coefficients:
         ar1      ar2     ar3      ma1
      0.8309  -0.0535  0.0429  -0.9530
s.e.  0.0538   0.0617  0.0501   0.0263

sigma^2 estimated as 0.03886:  log likelihood=95.03
AIC=-180.06   AICc=-179.92   BIC=-159.45
```

Έχουμε μικρά τυπικά σφάλματα για όλους τους όρους αλλά οι τιμές των κριτηρίων AIC, AICc και BIC δεν είναι αρκετά χαμηλές.

Η προσαρμογή του $ARIMA(2,1,2)$ δίνει:

```
> Arima(Worldscale,order = c(2,1,2))
Series: Worldscale
ARIMA(2,1,2)

Coefficients:
          ar1     ar2      ma1      ma2
      -0.0632  0.6769  -0.0231  -0.8701
s.e.   0.0769  0.0748   0.0552   0.0544

sigma^2 estimated as 0.03838:  log likelihood=97.81
AIC=-185.61   AICc=-185.48   BIC=-165.01
```

Σε αυτό το μοντέλο έχουμε μικρά τυπικά σφάλματα και τις αμέσως μικρότερες τιμές των κριτηρίων AIC και AIC μετά το μοντέλο $ARIMA(4,1,4)$. Ωστόσο, οι συντελεστές του πρώτου όρου AR και του πρώτου όρου MA βγήκαν στατιστικά ασήμαντοι. Το κριτήριο

BIC δείχνει την μικρότερη μέχρι στιγμής τιμή σε σχέση με όλα τα προηγούμενα μοντέλα.

Η προσαρμογή του $ARIMA(1,1,2)$ δίνει:

```
> Arima(Worldscale,order = c(1,1,2))
Series: Worldscale
ARIMA(1,1,2)

Coefficients:
          ar1     ma1      ma2
      -0.7573  0.6790  -0.1733
s.e.   0.1226  0.1328   0.0576

sigma^2 estimated as 0.03987:  log likelihood=88.86
AIC=-169.72   AICc=-169.63   BIC=-153.24
```

Σε αυτό το μοντέλο παρότι έχουμε μικρά τυπικά σφάλματα, έχουμε τις μεγαλύτερες τιμές των κριτηρίων AIC, AICc και BIC από όλα τα προηγούμενα μοντέλα.

Η προσαρμογή του $ARIMA(2,1,1)$ δίνει:

```
> Arima(Worldscale,order = c(2,1,1))
Series: Worldscale
ARIMA(2,1,1)

Coefficients:
          ar1      ar2      ma1
      0.8225  -0.0221  -0.9457
s.e.  0.0537   0.0498   0.0271

sigma^2 estimated as 0.03884:  log likelihood=94.66
AIC=-181.32   AICc=-181.23   BIC=-164.84
```

Εδώ έχουμε μικρά τυπικά σφάλματα και οι τιμές των κριτηρίων AIC, AIc και BIC είναι από τις μικρότερες παρατηρούμενες. Ωστόσο, ο δεύτερος όρος του AR βγαίνει στατιστικά ασήμαντος.

Η προσαρμογή του $ARIMA(1,1,1)$ δίνει:

```
> Arima(Worldscale,order = c(1,1,1))
Series: Worldscale
ARIMA(1,1,1)

Coefficients:
          ar1     ma1
      -0.8882  0.9362
s.e.   0.0579  0.0426

sigma^2 estimated as 0.04056:  log likelihood=84.45
AIC=-162.9   AICc=-162.85   BIC=-150.54
```

Στο μοντέλο αυτό αν και έχουμε μικρά τυπικά σφάλματα με όλους τους συντελεστές στατιστικά σημαντικούς ενώ ταυτόχρονα παρατηρούνται οι μεγαλύτερες τιμές των κριτηρίων AIC, AICc και BIC.

Στον Πίνακα 3.4 παρουσιάζονται συγκεντρωτικά οι τιμές των κριτηρίων AIC, AICc και BIC έπειτα από την προσαρμογή όλων των μοντέλων.

Πίνακας 3.4:

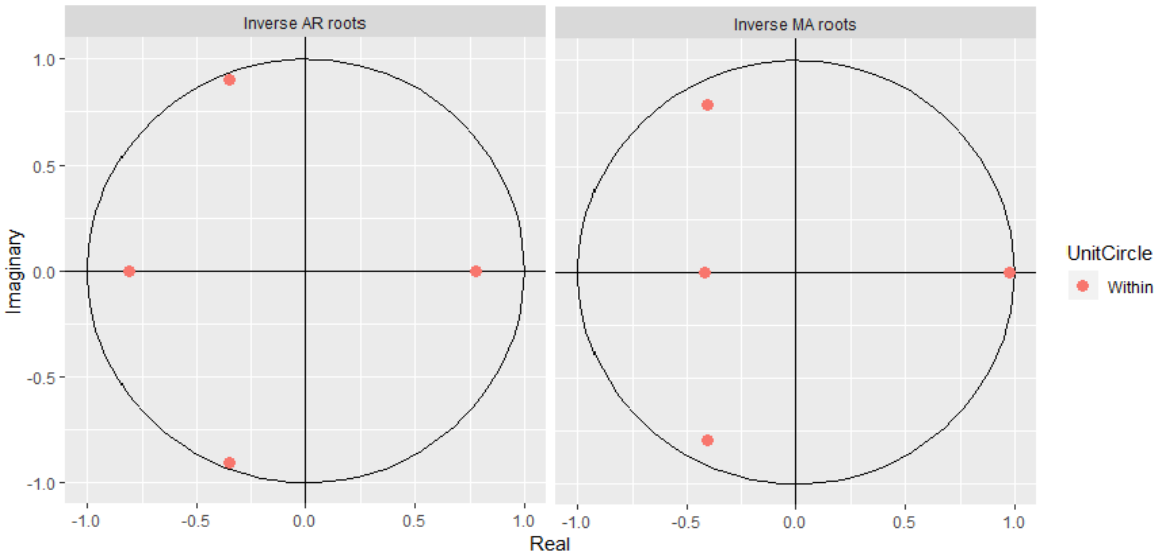
Μοντέλο	AIC	AICc	BIC
ARIMA(4,1,4)	-188.16	-187.76	-151.08
ARIMA(3,1,4)	-180.36	-180.04	-147.4
ARIMA(2,1,4)	-182.26	-182.01	-153.42
ARIMA(1,1,4)	-174.8	-174.61	-150.07
ARIMA(4,1,3)	-180.21	-179.89	-147.25
ARIMA(4,1,2)	-182.3	-182.05	-153.45
ARIMA(4,1,1)	-170.24	-170.05	-145.52
ARIMA(3,1,3)	-182.34	-182.08	-153.49
ARIMA(2,1,3)	-184.13	-183.94	-159.41
ARIMA(1,1,3)	-170.38	-170.25	-149.78
ARIMA(3,1,2)	-184.16	-183.97	-159.44
ARIMA(3,1,1)	-180.06	-179.92	-159.45
ARIMA(2,1,2)	-185.61	-185.48	-165.01
ARIMA(1,1,2)	-169.72	-169.63	-153.24
ARIMA(2,1,1)	-181.32	-181.23	-164.84
ARIMA(1,1,1)	-162.9	-162.85	-150.54

Μέχρι στιγμής το πρώτο υποψήφιο μοντέλο που προτείνεται με βάση τις τιμές AIC και AICc είναι το $ARIMA(4,1,4)$ ενώ το κριτήριο BIC δείχνει ως καλύτερο μοντέλο το $ARIMA(2,1,2)$. Έπειτα από τον έλεγχο για την σημαντικότητα των συντελεστών των δύο αυτών μοντέλων, το μοντέλο $ARIMA(4,1,4)$ έδειξε ότι έχει όλους τους συντελεστές του στατιστικά σημαντικούς ενώ το εμφωλευμένο του μοντέλο $ARIMA(2,1,2)$ έχει στατιστικά σημαντικούς συντελεστές μόνο τον δεύτερο όρο AR και τον δεύτερο όρο MA. Συνεπώς, θα επιλέξουμε το μοντέλο $ARIMA(4,1,4)$ που έχει όλους τους συντελεστές στατιστικά σημαντικούς και τις μικρότερες τιμές των κριτηρίων AIC και AICc.

Θα ξεκινήσουμε επομένως, με το μοντέλο $ARIMA(4,1,4)$ το οποίο σύμφωνα με τα αποτελέσματα που πήραμε από την R περιγράφεται από την εξίσωση:

$$Y_t = -0.7287Y_{t-1} - 0.3275Y_{t-2} + 0.4107Y_{t-3} + 0.5921Y_{t-4} + e_t + 0.6608e_{t-1} + 0.1246e_{t-2} - \\ - 0.6455e_{t-3} - 0.8405e_{t-4}\ ,\ \text{όπου}\ \left(Y_t = \nabla \log\left(X_t\right)\right) \qquad (3.1)$$

Μια αιτιώδης και αντιστρέψιμη διαδικασία ARMA όπως αναλύσαμε στην ενότητα 2.8, πρέπει να έχει τις ρίζες του AR και του MA πολυωνύμου εκτός του μοναδιαίου κύκλου. Αυτό είναι ισοδύναμο με το να έχει όλες τις αντίστροφες ρίζες (inverse roots) εντός του μοναδιαίου κύκλου. Εάν ικανοποιείται αυτή η συνθήκη η ARMA διαδικασία είναι στάσιμη και επίσης εξασφαλίζεται η μοναδικότητα της λύσης της, δίνοντας μοναδικές τιμές για τις μελλοντικές χρονικές στιγμές. Το μετασχηματισμένα πλέον δεδομένα με πρώτες διαφορές του WS δείκτη, σύμφωνα με την θεωρία ταυτίζονται με το μοντέλο ARMA(4,4). Το Διάγραμμα 3.16 δείχνει τις αντίστροφες ρίζες των πολυωνύμων AR και MA κατ' αντιστοιχία για το μοντέλο μας, οι οποίες πέφτουν όλες εντός του μοναδιαίου κύκλου εξασφαλίζοντας τη στασιμότητα και την μοναδικότητα της λύσης. Στο Διάγραμμα 3.16 κάποιες ρίζες φαίνονται ξεκάθαρα ότι βρίσκονται εντός του μοναδιαίου κύκλου ενώ κάποιες άλλες μπορεί να φαίνεται ότι πλησιάζουν πολύ το σύνορο του μοναδιαίου κύκλου, ωστόσο καμία δεν ακουμπάει στο σύνορο και είναι όλες αυστηρά μικρότερες της μονάδας.

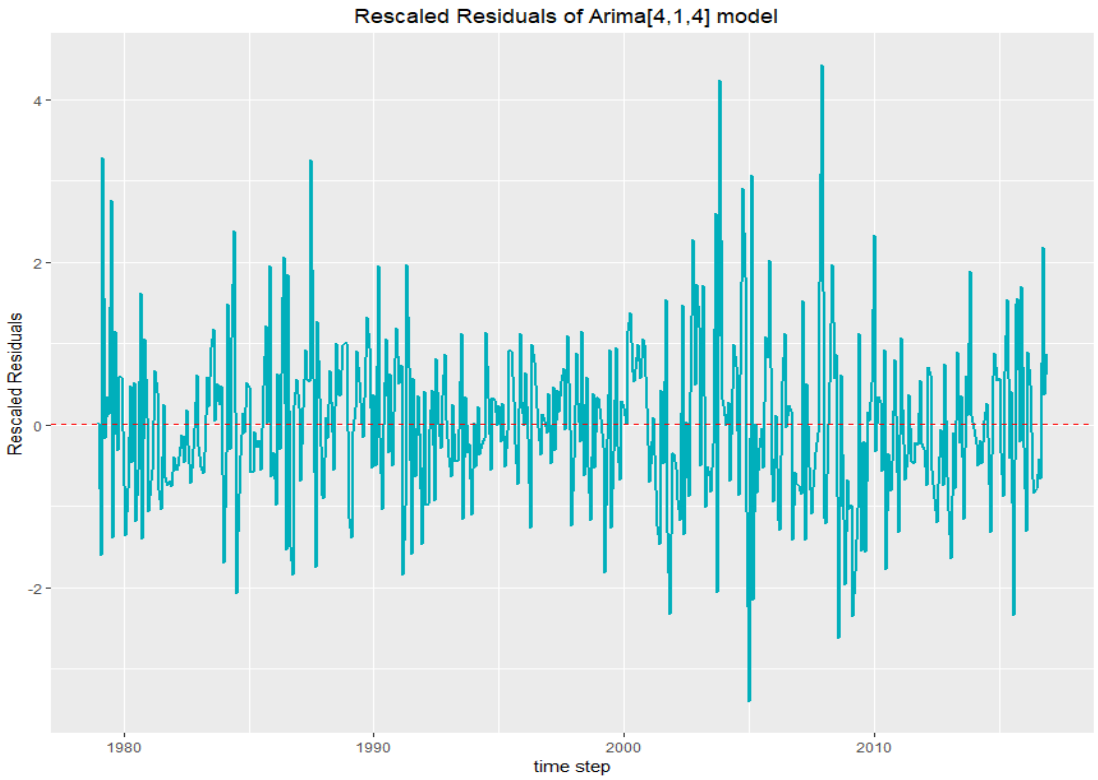


<u>Διάγραμμα 3.16:</u> Οι αντίστροφες ρίζες των AR και MA πολυωνύμων.

Το επόμενο βήμα είναι να διαπιστώσουμε εάν τα υπόλοιπα του μοντέλου ικανοποιούν τους διαγνωστικούς ελέγχους της ενότητας 2.12.6. Υπενθυμίζουμε ότι τα υπόλοιπα του $ARIMA(4,1,4)$ δείχνουν πόσο απέχει η κάθε εκτίμηση $\hat{Y}_t$ που λάβαμε έπειτα από την προσαρμογή του μοντέλου σε σχέση με τις πραγματικές τιμές $Y_t = \log(X_t)$, του δείκτη WS. Σύμφωνα με τη θεωρία τα υπόλοιπα πρέπει να έχουν ίδιες ιδιότητες με αυτές του λευκού θορύβου και ιδανικά να είναι, εκτός από ανεξάρτητα, και κανονικά κατανεμημένα.

Το Διάγραμμα 3.17 δείχνει το γράφημα των προσαρμοσμένων υπολοίπων $\hat{R}_t$, όπως ορίστηκε στην ενότητα 2.12.6.1 μαζί με την μέση τιμή τους. Όπως έχουμε πει στην

ενότητα 2.12.6 εάν το μοντέλο ARIMA(4,1,4) είναι κατάλληλο τότε τα προσασμοσμένα υπόλοιπα $\widehat{R}_t$ πρέπει να έχουν ιδιότητες ίδιες με αυτές του λευκού θορύβου, μέσης τιμής μηδέν και διασποράς ένα.



<u>Διάγραμμα 3.17:</u> Προσαρμοσμένα υπόλοιπα $\left\{\widehat{R}_t, t=1,\ldots,456\right\}$ *του μοντέλου ARIMA(4,1,4) .*

Στο Διάγραμμα 3.17, τα προσαρμοσμένα υπόλοιπα μοιάζουν να είναι στάσιμα γύρω από το μηδέν. Ωστόσο, η διασπορά τους μοιάζει να είναι αυξημένη στα έτη 2003-2008 σε σχέση με τα έτη 19930-2000. Συνεπώς, βλέπουμε ότι παρουσιάζονται φαινόμενα ετεροσκεδαστικότητας στα υπόλοιπα, με την διασπορά να μοιάζει ασταθή κατά την πάροδο του χρόνου. Ο Πίνακας 3.5 έχει ληφθεί με την εντολή summary της R, και μας δείχνει κάποια από τα χαρακτηριστικά των προσαρμοσμένων υπολοίπων $\widehat{R}_t$.

Πίνακας 3.5

summary	
	Rescaled Residuals $\widehat{R}_t$
Min	-3,40
1st Qu.	-0,60
Median	-0,05
Mean	0,00
3rd Qu.	0,57
Max	4,42
Variance	1,00

Όπως φαίνεται από τον Πίνακα 3.5 η δειγματική μέση τιμή των υπολοίπων $\widehat{R}_t$ είναι μηδέν και η διασπορά τους ένα. Ωστόσο, τα υπόλοιπα δεν συμβαδίζουν με την θεωρία που τα θέλει να είναι λευκός θόρυβος εξαιτίας του γεγονότος ότι υπάρχει ετεροσκεδαστικότητα στα σφάλματα και η διασπορά τους δεν είναι σταθερά ίση με ένα. Το συμπέρασμα αυτό επιβεβαιώνεται και από τον έλεγχο ετεροσκεδαστικότητας που μας δίνει η R με την βοήθεια της εντολής arch.test που βρίσκεται στη βιβλιοθήκη 'aTSA'. Ο έλεγχος αυτός έχει ως μηδενική υπόθεση ότι τα υπόλοιπα του ARIMA(4,1,4) μοντέλου είναι ομοσκεδαστικά με εναλλακτική ότι δεν είναι. Οι πολύ μικρές τιμές της p-value που πήραμε με την εφαρμογή του μας οδηγούν στην απόρριψη της μηδενικής υπόθεσης ότι τα σφάλματα του ARIMA(4,1,4) μοντέλου είναι ομοσκεδαστικά.

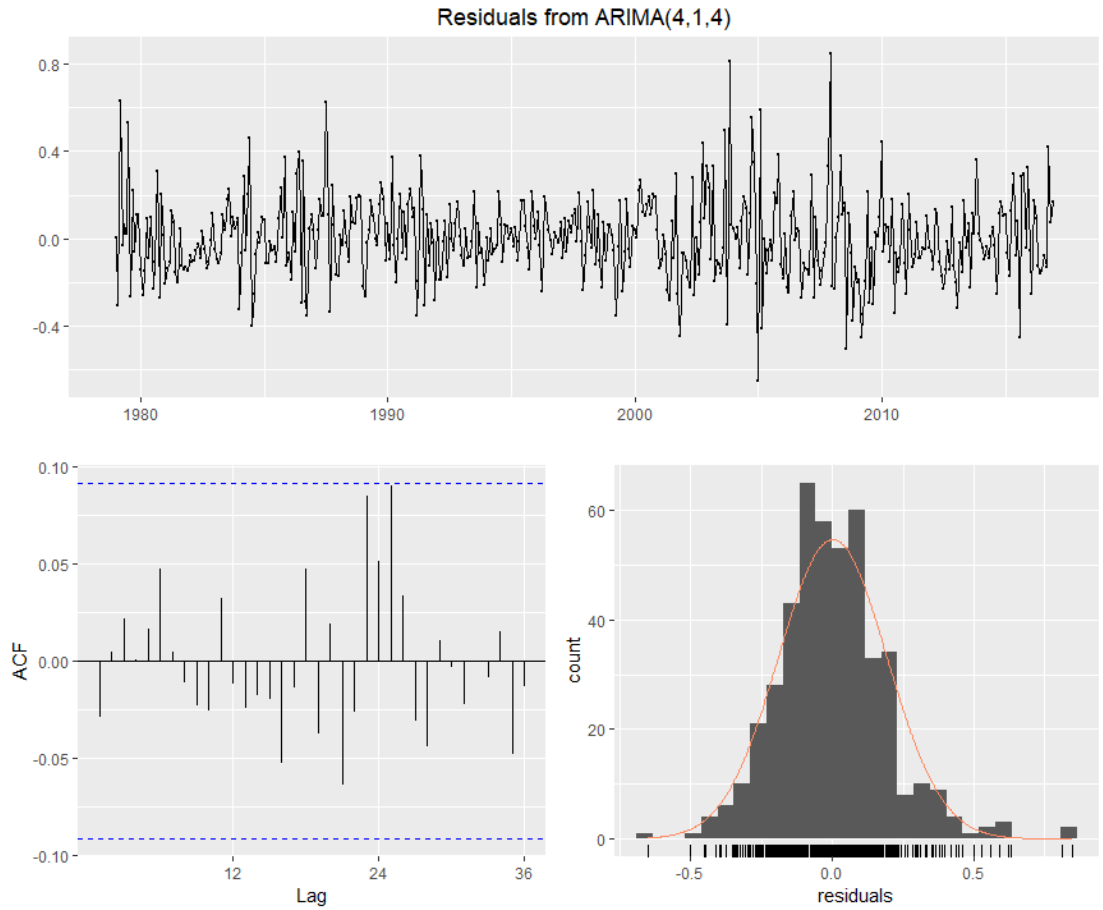
Ο Πίνακας 3.6 δείχνει τις τιμές των στατιστικών που αναφέραμε στις ενότητες 2.4.6.2-2.4.6.4 προκειμένου να ελέγξουμε κατά πόσο τα υπόλοιπα του μοντέλου ARIMA(4,1,4) είναι ισόνομα και ανεξάρτητα. Για τον υπολογισμό των στατιστικών και των p-value η μηδενική υπόθεση είναι ότι τα υπόλοιπα είναι ανεξάρτητα και ισόνομα με εναλλακτική ότι δεν είναι. Από τον Πίνακα 3.6 βλέπουμε ότι όλες οι τιμές p-value>0.05 και συνεπώς δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας 5%.

Πίνακας 3.6

Randomness Test Statistics			
Μηδενική Υπόθεση: Τα υπόλοιπα είναι ανεξάρτητα και ισόνομα			
Test	**Distribution**	**Statistic**	**p-value**
Box-Pierce test	Q ~ chisq(10)	2.36	0.993
Ljung-Box Q	Q ~ chisq(20)	6.94	0.997
McLeod-Li Q	Q ~ chisq(20)	11.75	0.756
Turning points T	(T-302.7)/9 ~ N(0,1)	299	0.683
Diff signs S	(S-227.5)/6.2 ~ N(0,1)	216	0.063
Rank P	(P-51870)/1625.6 ~ N(0,1)	50681	0.465

Επιπροσθέτως, σύμφωνα με όσα έχουμε πει στην ενότητα 2.12.6.1, με τη βοήθεια του γραφήματος των δειγματικών ACF των υπολοίπων μπορούμε να ελέγξουμε την ανεξαρτησία των υπολοίπων και να απορρίψουμε την μηδενική υπόθεση της ανεξαρτησίας, εάν περισσότερες από δύο αυτοσυσχετίσεις $\left(36\times 5\% \simeq 2\right)$ πέσουν εκτός των ορίων $\pm 1.96/\sqrt{n}$. Οι τιμές των δειγματικών ACF των υπολοίπων φαίνονται στο δεύτερο γράφημα του Διαγράμματος 3.18 και παρατηρούμε ότι δεν υπάρχουν αυτοσυσχετίσεις στα υπόλοιπα και συνεπώς δεν έχουμε αρκετές ενδείξεις για να απορρίψουμε την μηδενική υπόθεση περί ανεξαρτησίας των υπολοίπων.

Στο πρώτο γράφημα του Διαγράμματος 3.18 παίρνουμε τα υπόλοιπα του μοντέλου $ARIMA(4,1,4)$ της εξίσωσης 3.1 για όλες τις χρονικές στιγμές από τον Ιανουάριο του 1979 έως τον Δεκέμβριο του 2016. Η διαφορά από το Διάγραμμα 3.17 είναι ότι δεν τα έχουμε διαιρέσει με την τυπική τους απόκλιση. Το τελευταίο και τρίτο γράφημα του Διαγράμματος 3.18 δείχνει το ιστόγραμμα των υπολοίπων μαζί με την καμπύλη της κανονικής κατανομής. Εκ πρώτης όψεως τα υπόλοιπα δεν δείχνουν να έχουν αποκλίσεις από την κανονική κατανομή, ωστόσο ο έλεγχος της κανονικότητας θα εξεταστεί πιο λεπτομερώς παρακάτω. Το Διάγραμμα 3.18 λαμβάνεται με την εντολή ggtsdisplay που βρίσκεται στο πακέτο 'forecast' της R.



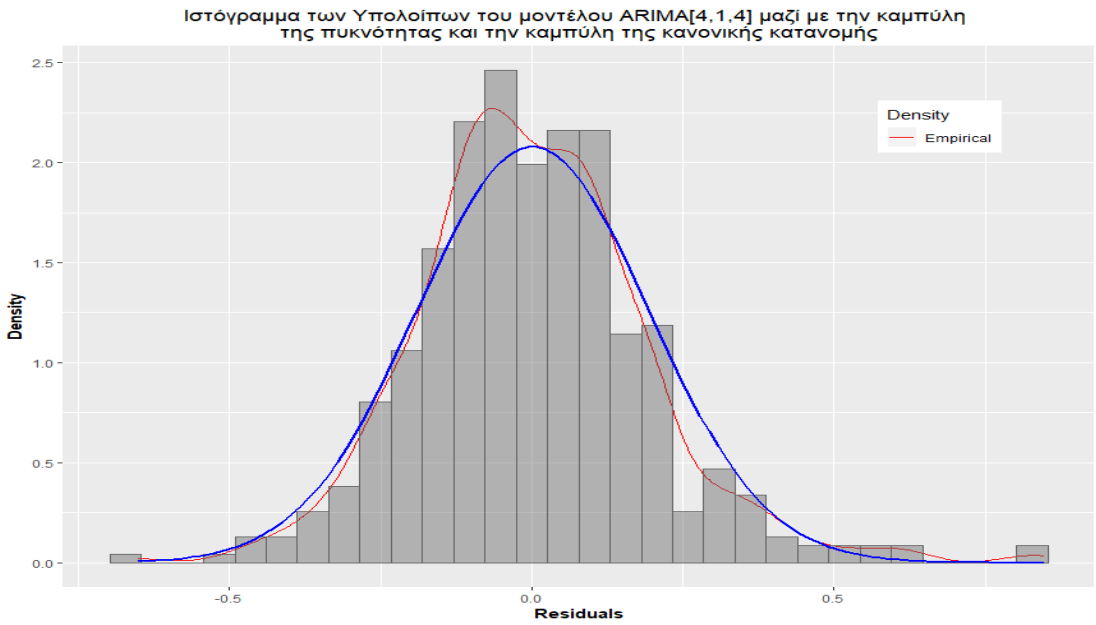
<u>Διάγραμμα 3.18:</u> Συνδυασμένο διάγραμμα των υπολοίπων $\{\hat{e}_t\}$ του μοντέλου ARIMA(4,1,4), το οποίο περιλαμβάνει: α) το γράφημα υπολοίπων $\{\hat{e}_t, t=1,\ldots,456\}$, β) τις δειγματικές αυτοσυσχετίσεις και γ) το ιστόγραμμα των υπολοίπων μαζί με την καμπύλη κανονικότητας .

Όπως έχουμε αναφέρει και στην αρχή του κεφαλαίου η δημιουργία των διαγραμμάτων διασποράς (scatter plots) μας βοηθάει να αναγνωρίσουμε την συναρτησιακή μορφή της προς μελέτης μεταβλητής σε προηγούμενες χρονικές

περιόδους. Το γράφημα με τις δειγματικές ACF ανιχνεύει μόνο γραμμική εξάρτηση και συνεπώς τα διαγράμματα διασποράς μας βοηθούν να αποφανθούμε για το αν τα υπόλοιπα μας παρουσιάζουν οποιαδήποτε άλλη εξάρτηση μεταξύ τους. Υπενθυμίζουμε ότι κάθε γράφημα του Διαγράμματος 3.19 δείχνει τα $\hat{e}_t$ σε σχέση με τα $\hat{e}_{t-h}$ για διάφορες τιμές της υστέρησης h. Έτσι, το πρώτο γράφημα απεικονίζει όλα τα σημεία $(\hat{e}_2, \hat{e}_1), (\hat{e}_3, \hat{e}_2), \ldots, (\hat{e}_{456}, \hat{e}_{455})$ και η διαδικασία αυτή συνεχίζεται μέχρι 12 υστερήσεις. Από το Διάγραμμα 3.19 φαίνεται ότι τα υπόλοιπα δεν σχετίζονται με κάποιο τρόπο μεταξύ τους, γεγονός που επιβεβαιώνει όλα τα προηγούμενα συμπεράσματα που βγάλαμε από τον Πίνακα 3.6

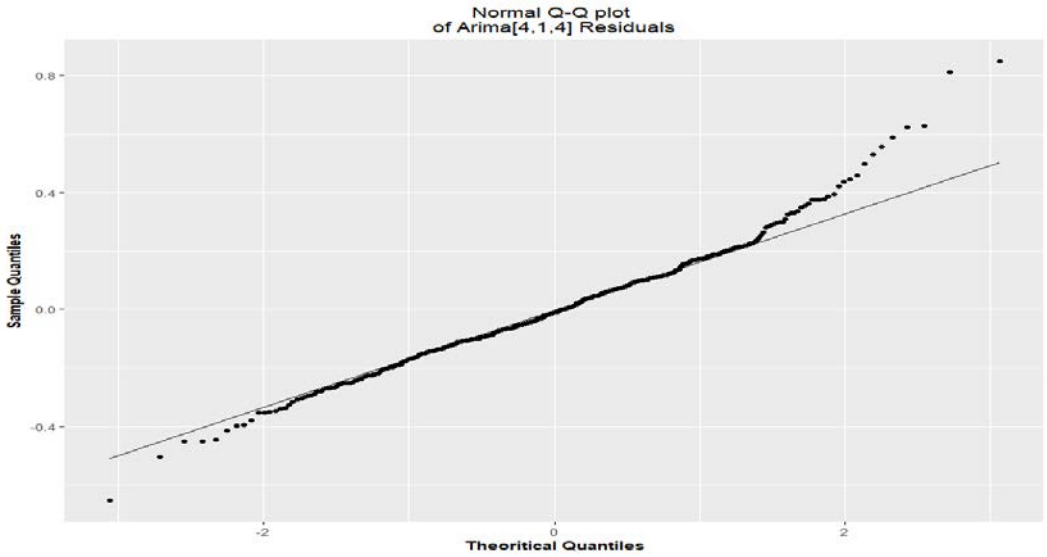
<u>Διάγραμμα 3.19</u>: Διάγραμμα διασποράς (scatter plot) των υπολοίπων του μοντέλου ARIMA(4,1,4) σε υστερήσεις h=1,2,…,12.

Το Διάγραμμα 3.20 δείχνει το ιστόγραμμα των υπολοίπων σε συνδυασμό με την μη παραμετρική εκτιμήτρια (κόκκινη γραμμή) για την κατανομή των υπολοίπων του μοντέλου ARIMA(4,1,4) και την καμπύλη της κανονικής κατανομής (μπλε γραμμή). Η μη-παραμετρική εκτιμήτρια της κατανομής φαίνεται να έχει δύο κορυφές, γεγονός που δημιουργεί πρόβλημα και δείχνει να αποκλίνει ελάχιστα από την καμπύλη της κανονικής κατανομής (μπλε γραμμή). Πολύ πιθανόν, για ένα λίγο μεγαλύτερο δείγμα από αυτό που ήδη έχουμε, τα υπόλοιπα να προσέγγιζαν καλύτερα την κανονική κατανομή.



<u>Διάγραμμα 3.20:</u> Ιστόγραμμα των υπολοίπων του ARIMA(4,1,4) μοντέλου μαζί με την μη-παραμετρική εκτιμήτρια της κατανομής των υπολοίπων και την καμπύλη της κανονικής κατανομής .

Για ένα σύνολο δεδομένων που προέρχεται από την κανονική κατανομή, όπως αναλύσαμε στην ενότητα 2.12.6.5 , θα πρέπει τα σημεία του Q-Q plot να βρίσκονται πάνω στην ευθεία που ορίζει το διάγραμμα χωρίς αποκλίσεις στις ουρές. Το Q-Q plot του Διαγράμματος 3.21 δείχνει ότι υπάρχουν αποκλίσεις από την κανονική κατανομή στην δεξιά ουρά των υπολοίπων, γεγονός που δημιουργεί πρόβλημα. Αυτό μπορεί να οφείλεται στις πολύ έντονες διακυμάνσεις των ναύλων που παρατηρήθηκαν την περίοδο 2002-2008, τα οποία επηρεάζουν και τα σφάλματα. Αν εξαιρέσουμε εκείνη την περίοδο των έντονων διακυμάνσεων, τα σφάλματα μοιάζουν κανονικά κατανεμημένα.



<u>Διάγραμμα 3.21:</u> Q-Q plot, αναπαριστώντας τα ποσοστημόρια των υπολοίπων (y-άξονας) σε σχέση με τα ποσοστημόρια ενός δείγματος από την κανονική κατανομή με μέση τιμή και τυπική απόκλιση ίδια με αυτή των υπολοίπων (x-άξονας).

Εκτός από τους γραφικούς αυτούς ελέγχους, σημαντικοί στατιστικοί έλεγχοι για κανονικότητα των υπολοίπων είναι οι Kolmogorov-Smirnov, Anderson-Darling και Jarque–Bera .

Μέσω του ελέγχου Kolmogorov-Smirnov (Conover, 1999) υπολογίζεται πόσο απέχει η εμπειρική αθροιστική κατανομή των υπολοίπων από την αθροιστική συνάρτηση της κανονικής κατανομής και ελέγχονται οι υποθέσεις:

$\mathrm{H}_0:$ Το δείγμα προέρχεται από την κανονική κατανομή

$\mathrm{H}_1:$ Το δείγμα δεν προέρχεται από την κανονική κατανομή

Αντίστοιχα, ο Anderson-Darling (Anderson & Darling, 1954) πραγματοποιεί τους ίδιους στατιστικούς ελέγχους με τον Kolmogorov-Smirnov και έχει και αυτό ως μηδενική υπόθεση ότι τα υπόλοιπα προέρχονται από την κανονική κατανομή. Η διαφορά των δυο αυτών ελέγχων είναι ότι ο Anderson-Darling δίνει μεγαλύτερο βάρος στις ουρές της κατανομής των υπολοίπων ενώ ο Kolmogorov-Smirnov υπολογίζει τις αποκλίσεις της εμπειρικής κατανομής από την κανονική κατανομή, στο κέντρο.

Ο έλεγχος Jarque–Bera (Jarque & Bera, 1980) έχει και αυτός ως μηδενική υπόθεση την κανονικότητα των δεδομένων και συγκρίνει την κύρτωση και την λοξότητα των υπολοίπων σε σχέση με την κύρτωση και την λοξότητα της κανονικής κατανομής για την οποία ισχύει ότι η κύρτωση είναι ίση με τρία ενώ η λοξότητα ίση με μηδέν. Το στατιστικό Jarque-Bera, $n\left[m_3^2 / \left(6m_2^3\right) + \left(m_4 / m_2^3 - 3\right)^2 / 24 \right]$, όπου $m_r = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^r \Big/ n$, είναι ασυμπτωτικά κατανεμημένο όπως η $\chi^2(2)$ εάν $\{e_t\} \sim iidN\left(\mu, \sigma^2\right)$. Η μηδενική υπόθεση περί κανονικότητας των δεδομένων απορρίπτεται εάν η τιμή του στατιστικού είναι αρκετά μεγάλη (σε επίπεδο σημαντικότητας 5% εάν η p-value του ελέγχου είναι μικρότερη από 5%).

Ο Πίνακας 3.7 δείχνει τα αποτελέσματα που λάβαμε έπειτα από την πραγματοποίηση των ελέγχων κανονικότητας. Παρατηρούμε ότι οι έλεγχοι Anderson- Darling και Jarque-Bera συμφωνούν και απορρίπτουν την μηδενική υπόθεση περί κανονικότητας των δεδομένων σε επίπεδο σημαντικότητας 1%, ενώ ο έλεγχος Kolmogorov-Smirnov δείχνει ότι δεν έχουμε αρκετές ενδείξεις για να απορρίψουμε την υπόθεση της κανονικότητας των υπολοίπων. Το αποτέλεσμα αυτό ήταν αναμενόμενο αφού όπως είδαμε στο Q-Q plot του Διαγράμματος 3.21, τα υπόλοιπα προσαρμόζονται πολύ καλά στο κέντρο της κανονικής κατανομής ενώ έχουμε αποκλίσεις μόνο στη δεξιά ουρά.

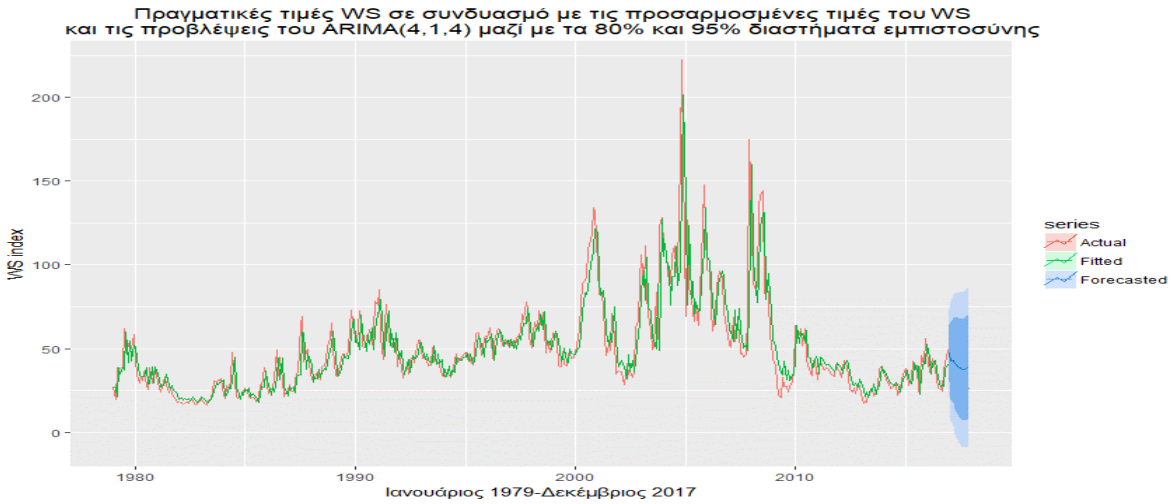
Πίνακας 3.7

Testing Normality		
Μηδενική Υπόθεση: Τα υπόλοιπα ακολουθούν την κανονική κατανομή		
Test	**Statistic**	**p-value**
Anderson-Darling	1,77	0,0002
Kolmogorov-Smirnov	0,05	0,1657
Jarque Bera	80	<2,2e-16

Με βάση όλα τα παραπάνω, δεν μπορούμε να θεωρήσουμε ότι τα υπόλοιπα του μοντέλου ARIMA(4,1,4) προέρχονται από την κανονική κατανομή. Ωστόσο, όπως διαπιστώσαμε από τους ελέγχους τυχαιότητας, τα υπόλοιπα είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητής, με μηδενική μέση τιμή. Η διασπορά τους αυξομειώνεται με την πάροδο του χρόνου γεγονός που δεν τα κάνει συμβατά με την θεωρία που τα θέλει να είναι λευκός θόρυβος με σταθερή διασπορά. Εδώ να σημειώσουμε ότι κανένα από τα εμφωλευμένα μοντέλα δεν μπορούσε να αντιμετωπίσει το φαινόμενο της ετεροσκεδαστικότητας στα υπόλοιπα και να ξεπεράσει αυτό το πρόβλημα.

Παρόλα αυτά, επειδή το μοντέλο ARIMA είναι κατάλληλο για να υπολογίζει την αναμενόμενη τιμή της προ μελέτη μεταβλητής, δεδομένου τις τιμές που είχε σε προηγούμενες χρονικές στιγμές, θα συνεχίσουμε προσαρμόζοντάς το στα δεδομένα μας. Φυσικά, θα πρέπει να έχουμε υπόψιν ότι δεν μπορεί να μοντελοποιήσει την μεταβαλλόμενη με το χρόνο διασπορά και σε περίπτωση που ενδέχεται να συμβεί κάποιο σοκ το 2017, θα αποτύχει να το προβλέψει.

Με βάση τα παραπάνω, το επόμενο βήμα μας είναι να χρησιμοποιήσουμε το μοντέλο $ARIMA(4,1,4)$ για να προβλέψουμε τις μηνιαίες τιμές του δείκτη Worldscale για το έτος 2017. Στο Διάγραμμα 3.22 αναπαρίστανται με την κόκκινη γραμμή οι πραγματικές τιμές του δείκτη Worldscale την χρονική περίοδο από τον Ιανουάριο του 1979 έως τον Δεκέμβριο του 2016 και με την πράσινη γραμμή οι προσαρμοσμένες τιμές του δείκτη του μοντέλου ARIMA(4,1,4) την αντίστοιχη χρονική περίοδο. Η μπλε γραμμή, μας δείχνει τις προβλέψεις που πήραμε μετά την εφαρμογή του μοντέλου ARIMA(4,1,4) για όλους τους μήνες του τελευταίου έτους 2017. Τα 80% και 95% διαστήματα εμπιστοσύνης για της προβλέψεις που μας έδωσε το μοντέλο ARIMA(4,1,4) φαίνονται κατά αντιστοιχία με σκούρα μπλε σκίαση και ανοιχτή μπλε σκίαση.



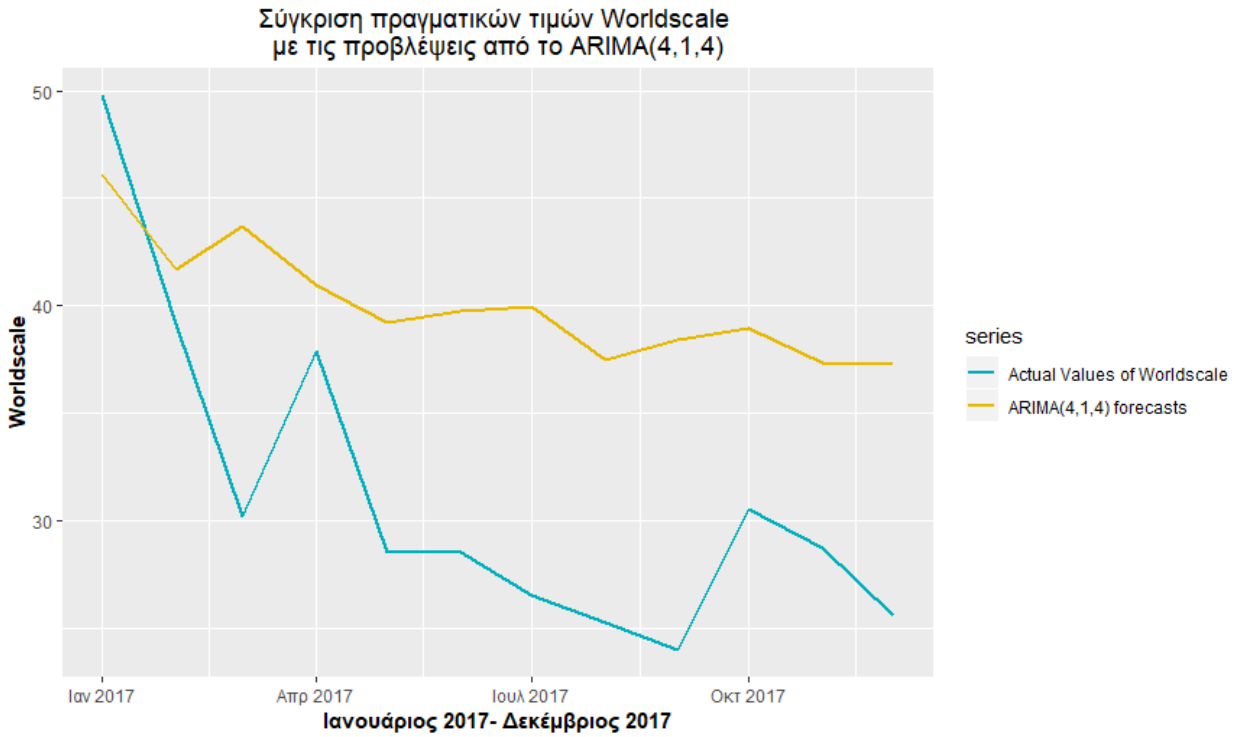
<u>Διάγραμμα 3.22</u>: Μηνιαίες τιμές ναύλων μεταφοράς πετρελαίου με πλοία VLCC στη διαδρομή Ras Tanura-Rotterdam (κόκκινη γραμμή) την περίοδο Ιαν.1979-Δεκ.2016 μαζί με τις προσαρμοσμένες τιμές του μοντέλου ARIMA(4,1,4) (πράσινη γραμμή) την ίδια χρονική περίοδο και τις προβλέψεις του μοντέλου για όλους τους μήνες του έτους 2017 (μπλε γραμμή). Τα 80 και 95% διαστήματα εμπιστοσύνης για τις προβλεπόμενες τιμές του 2017 φαίνονται κατ' αντιστοιχία με σκούρα μπλε σκίαση και ανοιχτή μπλε σκίαση.

Στο Διάγραμμα 3.23 αναπαρίστανται με τη μαύρη γραμμή οι πραγματικές μηνιαίες τιμές του δείκτη για όλη την περίοδο από τον Ιανουάριο του 1979 έως τον Δεκέμβριο του 2017 και με κόκκινη γραμμή μόνο οι προβλέψεις που πήραμε από το μοντέλο $ARIMA(4,1,4)$ για όλους τους μήνες του έτους 2017. Το διάγραμμα αυτό είναι πολύ χρήσιμο προκειμένου να συγκρίνουμε γραφικά τις προβλέψεις που πήραμε από το μοντέλο μας με τις πραγματικές τιμές του δείκτη. Κατά αντιστοιχία με το Διάγραμμα 3.22, η σκούρα μπλε σκίαση δείχνει το 80% διάστημα εμπιστοσύνης για τις μηνιαίες προβλέψεις του μοντέλου ARIMA(4,1,4) του έτους 2017 ενώ η ανοιχτή μπλε σκίαση δείχνει το 95% διάστημα εμπιστοσύνης. Από το Διάγραμμα 3.23, παρατηρούμε ότι το μοντέλο ARIMA(4,1,4) είναι ικανό να συλληφθεί την πτώση των ναύλων που παρατηρήθηκε το τελευταίο έτος του 2017 σε σχέση με την τιμή που είχε τον Δεκέμβριο του 2016.

<u>Διάγραμμα 3.23</u>: Μηνιαίες τιμές ναύλων μεταφοράς πετρελαίου με πλοία VLCC στη διαδρομή Ras Tanura-Rotterdam μαζί με τις μηνιαίες προβλέψεις του μοντέλου ARIMA(4,1,4) για το έτος 2017 και τα 80% (σκούρο γκρι), 95%(ανοιχτό γκρι) διαστήματα εμπιστοσύνης για τις προβλεπόμενες τιμές.

Προκειμένου να έχουμε μια καλύτερη εικόνα για την σύγκριση των μηνιαίων προβλέψεων του μοντέλου ARIMA(4,1,4) σε σχέση με τις πραγματικές τιμές του δείκτη WS, την χρονική περίοδο από τον Ιανουάριο του 2017 έως τον Δεκέμβριο του 2017, πραγματοποιούμε τον γραφικό έλεγχο του Διαγράμματος 3.24. Το Διάγραμμα 3.24 συγκρίνει τις μηνιαίες προβλέψεις του μοντέλου ARIMA(4,1,4)(πορτοκαλί γραμμή) με τις πραγματικές τιμές που είχε ο δείκτης WS (μπλε γραμμή) την αντίστοιχη χρονική περίοδο. Από το διάγραμμα αυτό παρατηρούμε ότι όσο ο χρονικός ορίζοντας της πρόβλεψης μεγαλώνει, τόσο μεγαλώνουν οι αποστάσεις μεταξύ των πραγματικών

τιμών του δείκτη και των προβλέψεων. Ακόμη, όπως και στο προηγούμενο Διάγραμμα 3.23, είναι εμφανές ότι το μοντέλο ARIMA(4,1,4) μπορεί να αντιληφθεί την πτώση των ναύλων ακόμα και αν αυτό δεν επιτυγχάνεται με ακρίβεια.



<u>Διάγραμμα 3.24</u>: Σύγκριση πραγματικών μηνιαίων τιμών δείκτη Worldscale με τις προβλέψεις που έδωσε το μοντέλο ARIMA(4,1,4) για την περίοδο από τον Ιανουάριο 2017-Δεκέμβριο 2017.

Αφού πραγματοποιήσαμε όλους τους απαραίτητους γραφικούς ελέγχους, στον Πίνακα 3.8 λαμβάνουμε τις 12 σημειακές προβλέψεις του δείκτη WS με βάση το μοντέλο ARIMA(4,1,4) για το έτος 2017. Πιο συγκεκριμένα, στον Πίνακα 3.8, εκτός από τις σημειακές προβλέψεις έχουμε τα 80% και 95% διαστήματα εμπιστοσύνης για τις σημειακές εκτιμήσεις του μοντέλου ARIMA(4,1,4). Στην τελευταία στήλη του πίνακα βλέπουμε τις πραγματικές τιμές του δείκτη WS για το έτος 2017 έτσι ώστε να έχουμε μια πιο ολοκληρωμένη άποψη για την ακρίβεια των προβλέψεων. Στο σημείο αυτό να αναφέρουμε ότι όλες οι τιμές που παρουσιάζονται στον Πίνακα 3.8 λαμβάνονται μέσω της εντολής forecast που βρίσκεται στο πακέτο 'forecast' της R.

Αυτό που παρατηρούμε από τον Πίνακα 3.8 είναι ότι τα αυτοπαλινδρομικά μοντέλα ARIMA δεν μπορούν να προβλέψουν τις μελλοντικές τιμές του δείκτη Worldscale με μεγάλη ακρίβεια. Επίσης, βλέπουμε ότι όσο μεγαλώνει ο χρονικός ορίζοντας της πρόβλεψης, τόσο αποκλίνουν οι πραγματικές τιμές από τις προβλέψεις. Παρόλα αυτά, καμία πραγματική τιμή του δείκτη δεν πέφτει εκτός από τα 80% και 95% διαστήματα εμπιστοσύνης που λάβαμε για της σημειακές εκτιμήσεις του μοντέλου

ARIMA(4,1,4). Συνεπώς, όλες οι τιμές του δείκτη πέφτουν εντός των ορίων αυτών και όπως είδαμε και γραφικά το μοντέλο ARIMA(4,1,4) είναι ικανό να δώσει μια γενική εικόνα για την άνοδο ή την κάθοδο των ναύλων συνολικά για όλους τους μήνες του 2017.

Πίνακας 3.8

Date	ARIMA(4,1,4) point forecast	80% Low	80% High	95% Low	95% High	Actual values of WS
Jan 2017	47,071	35,931	59,074	31,500	67,383	49,750
Feb 2017	41,686	29,684	58,544	24,798	70,074	39,125
Mar 2017	43,715	29,586	64,593	24,062	79,422	30,200
Apr 2017	40,956	26,771	62,658	21,375	78,474	37,875
May 2017	39,224	25,140	61,199	19,866	77,448	28,500
Jun 2017	39,743	24,931	63,355	19,477	81,095	28,500
Jul 2017	39,980	24,745	64,594	19,195	83,271	26,500
Aug 2017	37,464	22,951	61,155	17,707	79,266	25,250
Sep 2017	38,421	23,303	63,348	17,884	82,545	24,000
Oct 2017	38,929	23,386	64,801	17,857	84,866	30,500
Nov 2017	37,365	22,315	62,567	16,986	82,197	28,750
Dec 2017	37,270	22,102	62,847	16,761	82,873	25,600

Προκειμένου να ελέγξουμε κατά πόσο το μοντέλο ARIMA(4,1,4) που διαλέξαμε είναι πράγματι το καλύτερο δυνατό μοντέλο σε σχέση με όλα τα υπόλοιπα θα εξετάσουμε το μέσο απόλυτο σφάλμα (MAE) και την τετραγωνική ρίζα του μέσου απόλυτου σφάλματος (RMSE). Τα δύο αυτά σφάλματα πρόβλεψης είναι ενδεικτικά για την ακρίβεια των προβλέψεων και μετράνε πόσο απέχουν οι σημειακές εκτιμήσεις που λαμβάνονται μέσω ενός ARIMA μοντέλου από τις πραγματικές. Στην πραγματικότητα, κάθε είδους πρόβλεψη έχει κάποιο επίπεδο σφάλματος. Ένας κύριος λόγος που συμβαίνει αυτό είναι ότι η τυχαιότητα που παρουσιάζεται στα σφάλματα της προ μελέτης μεταβλητής δεν μπορεί να εξηγηθεί πλήρως από το μοντέλο που χρησιμοποιείται. Δηλαδή, τα υπόλοιπα του προσαρμοσμένου μοντέλου δεν αποτελούν μια πραγματοποίηση της διαδικασίας των τυχαίων διακυμάνσεων $\{e_t\}$ από την οποία έχει προέλθει η προς μελέτη μεταβλητή.

Τα κριτήρια που θα αξιολογήσουν την προβλεπτική ισχύ των ARIMA μοντέλων ορίζονται ως εξής:

$MAE = \frac{\sum_{i=1}^{n}|e_i|}{n} = \frac{\sum_{i=1}^{n}|x_i - \hat{x}_i|}{n}$, $RMSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}|x_i - \hat{x}_i|^2}{n}}$, όπου $\hat{x}_i$ είναι η πρόβλεψη την χρονική στιγμή i .

Το μέσο απόλυτο σφάλμα (MAE) είναι ο μέσος όρος των απόλυτων σφαλμάτων του μοντέλου και είναι ένα μέτρο που δείχνει πόσο προσεγγίζουν οι εκτιμήσεις του μοντέλου, τις πραγματικές τιμές της προ μελέτης μεταβλητής. Συνεπώς, όσο μικρότερη η τιμή του MAE τόσο καλύτερη η προσαρμογή του μοντέλου.

Αντίστοιχα με το μέσο απόλυτο σφάλμα, η τετραγωνική ρίζα του μέσου απόλυτου σφάλματος (RMSE) είναι ο μέσος όρος των διαφορών των τετραγώνων μεταξύ των εκτιμώμενων τιμών και των πραγματικών. Το σφάλμα αυτό αντικατοπτρίζει την διακύμανση και συχνά αναφέρεται ως μέτρο ρίσκου. Όσο μικρότερη η τιμή του RMSE τόσο καλύτερη η προσαρμογή του μοντέλου.

Υπενθυμίζουμε ότι έχουμε χωρίσει τα δεδομένα μας σε ένα σύνολο εκπαίδευσης (training set) το οποίο περιλαμβάνει όλες τις πραγματικές τιμές του δείκτη WS από τον Ιανουάριο του 1979 έως τον Δεκέμβριο του 2016 και σε ένα σύνολο δοκιμασίας (test set) το οποίο περιλαμβάνει όλες τις πραγματικές μηνιαίες τιμές του δείκτη WS για το έτος 2017. Αυτό το κάναμε διότι δεν έχει νόημα να εξετάζουμε την εγκυρότητα ενός μοντέλου στο δείγμα το οποίο το έχουμε βάλει να προσαρμοστεί, αλλά πρέπει να εξετάσουμε την συμπεριφορά του σε ένα καινούριο σύνολο δεδομένων. Με βάση την παρατήρηση αυτή, θα υπολογίσουμε για όλα τα εμφωλευμένα μοντέλα του ARIMA(4,1,4) που αναφέραμε στο κεφάλαιο αυτό, το MAE και το RMSE, τόσο για το σύνολο εκπαίδευσης όσο και για το σύνολο δοκιμασίας των καινούριων δεδομένων. Ο Πίνακας 3.8 δείχνει όλα τα αποτελέσματα που λάβαμε για το training set και το test set για κάθε εμφωλευμένο μοντέλο. Οι τιμές του Πίνακα 3.9 λαμβάνονται με την εντολή accuracy που βρίσκεται στο πακέτο 'forecast' της R.

Training set: είναι όλες οι πραγματικές τιμές του λογαριθμημένου δείκτη WS από Ιαν.1979-Δεκεμβ.2016

Test set: είναι το σύνολο καινούριων δεδομένων που ελέγχω. Ουσιαστικά κάνει τις συγκρίσεις μεταξύ των πραγματικών δεδομένων του λογαριθμημένου δείκτη WS από τον Ιαν.2017-Δεκ.2017 με τις λογαριθμοποιημένες τιμές που έχει προβλέψει για την περίοδο αυτή.

Πίνακας 3.9

MODEL	MAE		RMSE	
	training set	test set	training set	test set
ARIMA(4,1,4)	0,1459434	0,2831895	0,1916937	0,3137072
ARIMA(3,1,4)	0,1492223	0,3031277	0,1946728	0,3345163
ARIMA(2,1,4)	0,1492538	0,3037729	0,1946946	0,3349692
ARIMA(1,1,4)	0,149334	0,3343342	0,1968029	0,3728212
ARIMA(4,1,3)	0,1492692	0,3048691	0,1947044	0,3360107
ARIMA(4,1,2)	0,1492477	0,3038713	0,1946863	0,3351417
ARIMA(4,1,1)	0,1489219	0,4045671	0,1978058	0,4449058
ARIMA(3,1,3)	0,1492345	0,3042407	0,1946776	0,3356767
ARIMA(2,1,3)	0,1492637	0,3057613	0,1947226	0,3368725
ARIMA(1,1,3)	0,1495498	0,4160654	0,1982107	0,4554036
ARIMA(3,1,2)	0,1492666	0,3059598	0,1947162	0,3371083
ARIMA(3,1,1)	0,1494525	0,3037088	0,196049	0,3319294
ARIMA(2,1,2)	0,1492501	0,2990142	0,1948353	0,3296065
ARIMA(1,1,2)	0,1493986	0,4552806	0,1987957	0,4973537
ARIMA(2,1,1)	0,149638	0,2974931	0,1962084	0,3251905
ARIMA(1,1,1)	0,1498739	0,4807865	0,2007422	0,5226464

Με βάση τα αποτελέσματα του Πίνακα 3.9 βλέπουμε ότι πράγματι το μοντέλο ARIMA(4,1,4) στο training set στο οποίο το έχουμε προσαρμόσει, είναι το μοντέλο με τα μικρότερα παρατηρούμενα σφάλματα σε σχέση με όλα τα εμφωλευμένα του. Συνεπώς επιβεβαιώνεται ότι η αρχική μας επιλογή ήταν η ορθότερη. Τα αποτελέσματά της συνάδουν στα εντελώς καινούρια δεδομένα (test set) που επαληθεύουν ότι το ARIMA(4,1,4) είναι πράγματι το καλύτερο μοντέλο αφού και εκεί παρατηρούνται οι μικρότερες τιμές των κριτηρίων. Εδώ αξίζει να παρατηρήσουμε πόσο έχουν μεγαλώσει οι τιμές των MAE και RMSE στο test set του κάθε μοντέλου σε σχέση με τις τιμές που είχαν στο training set. Συμπεραίνουμε λοιπόν ότι σε ένα καινούριο σύνολο δεδομένων τα μοντέλα ARIMA δεν μπορούν να συμπεριφερθούν με την ίδια ακρίβεια.

Στον Πίνακα 3.10 παραθέτω τις σημειακές εκτιμήσεις που δώσανε όλα τα εμφωλευμένα μοντέλα που προαναφέραμε για την περίοδο Ιανουάριος 2017-Δεκέμβριος 2017. Στην τελευταία στήλη έχουμε τις πραγματικές τιμές που έλαβε ο δείκτης Worldscale την περίοδο αυτή. Στα κόκκινα πλαίσια βρίσκονται οι εκτιμημένες τιμές που προσέγγιζαν καλύτερα τις πραγματικές. Βλέπουμε ότι πράγματι το μοντέλο ARIMA(4,1,4) που χρησιμοποιήσαμε έχει τις περισσότερες τιμές του πιο κοντά στις πραγματικές σε σχέση με τα άλλα μοντέλα. Το αμέσως επόμενο καλύτερο μοντέλο είναι το ARIMA(2,1,1). Τα μοντέλα αυτά δεν φαίνεται να ενδείκνυνται για μελλοντικές προβλέψεις 12 μηνών, βλέπουμε ότι δίνουν καλά αποτελέσματα μόνο για να προβλέψουμε 1-4 μήνες μπροστά.

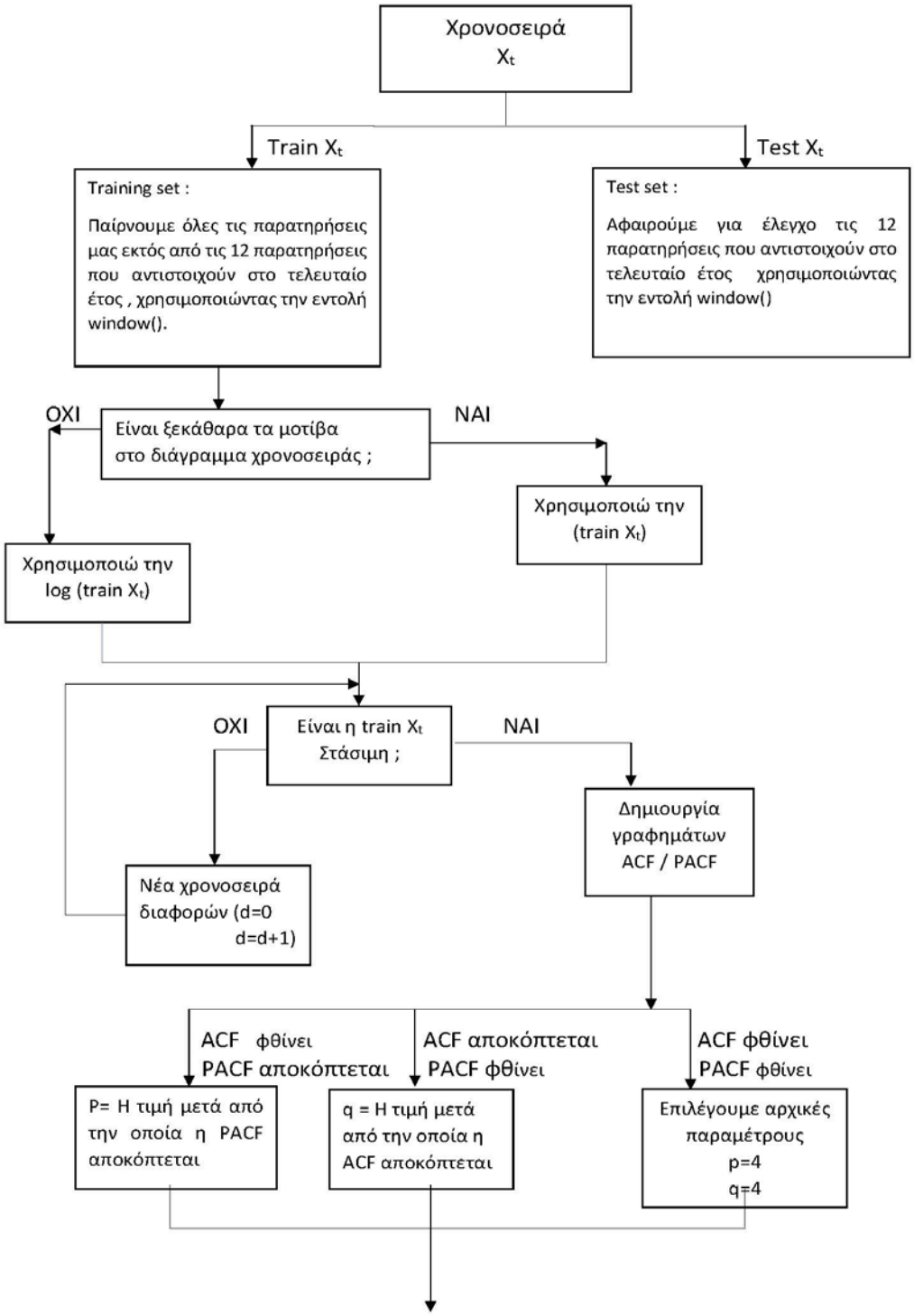
Πίνακας 3.10

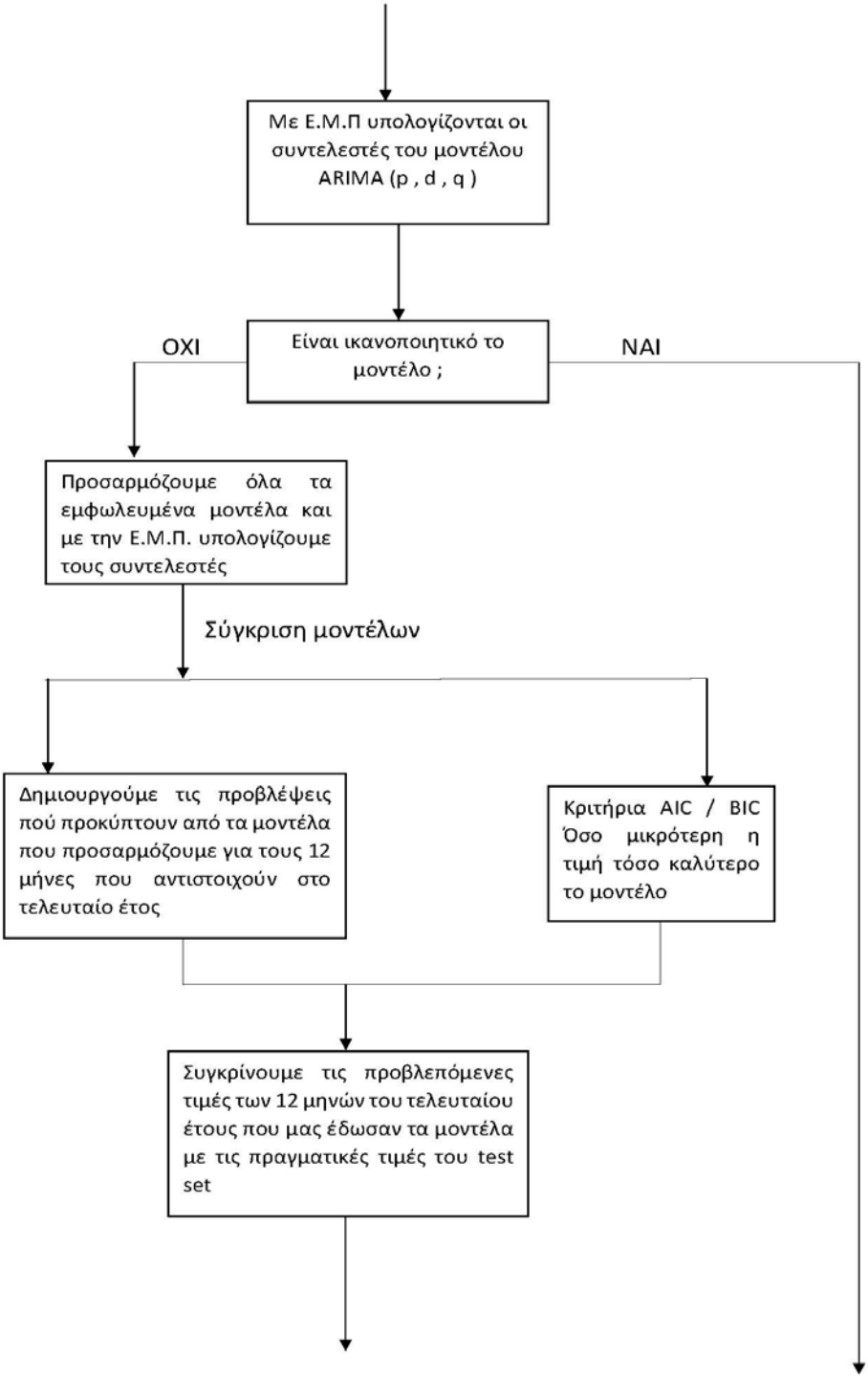
Date	ARIMA(4,1,4)	ARIMA(3,1,4)	ARIMA(2,1,4)	ARIMA(1,1,4)	ARIMA(4,1,3)	ARIMA(4,1,2)	ARIMA(4,1,1)	ARIMA(3,1,3)	Actual values of WS
Jan 2017	47,071	47,956	47,935	47,383	47,897	47,939	47,568	47,855	49,750
Feb 2017	41,686	44,270	44,342	42,611	44,352	44,320	44,224	44,204	39,125
Mar 2017	43,715	44,180	44,180	42,937	44,205	44,196	45,800	44,166	30,200
Apr 2017	40,956	41,695	41,779	41,155	41,853	41,766	44,384	41,712	37,875
May 2017	39,224	41,676	41,686	41,924	41,752	41,704	45,906	41,760	28,500
Jun 2017	39,743	40,175	40,213	41,586	40,281	40,217	44,997	40,223	28,500
Jul 2017	39,980	40,135	40,160	41,734	40,222	40,170	45,506	40,226	26,500
Aug 2017	37,464	39,226	39,240	41,669	39,292	39,250	45,186	39,286	25,250
Sep 2017	38,421	39,181	39,209	41,697	39,258	39,214	45,319	39,252	24,000
Oct 2017	38,929	38,624	38,628	41,685	38,664	38,638	45,297	38,680	30,500
Nov 2017	37,365	38,583	38,610	41,690	38,645	38,612	45,277	38,631	28,750
Dec 2017	37,270	38,240	38,241	41,688	38,263	38,250	45,316	38,283	25,600

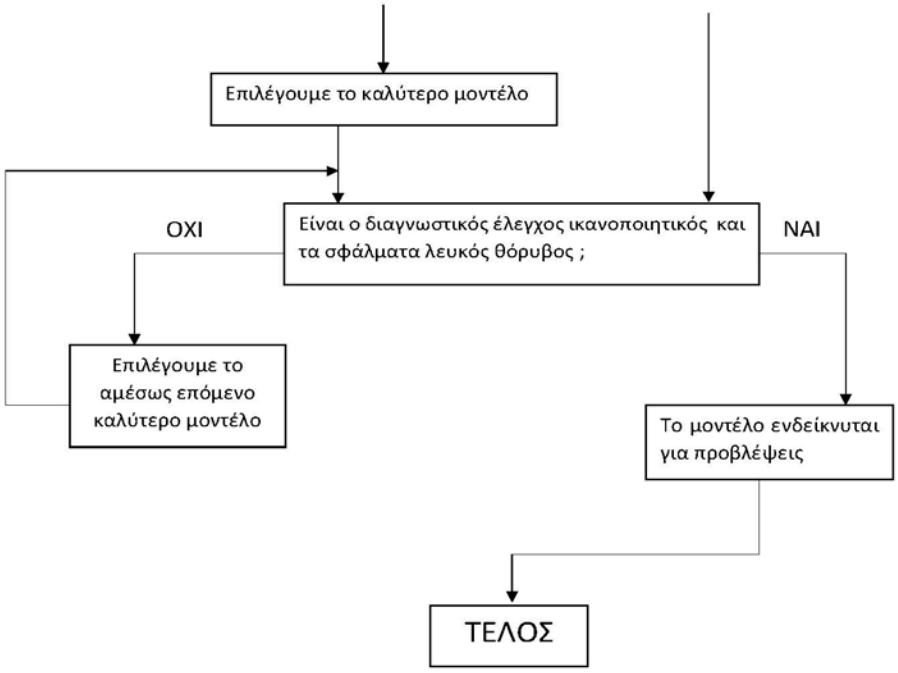
Date	ARIMA(2,1,3)	ARIMA(1,1,3)	ARIMA(3,1,2)	ARIMA(3,1,1)	ARIMA(2,1,2)	ARIMA(1,1,2)	ARIMA(2,1,1)	ARIMA(1,1,1)	Actual values of WS
Jan 2017	47,711	47,835	47,739	46,399	47,579	49,318	46,451	49,820	49,750
Feb 2017	44,225	45,280	44,239	44,271	44,180	46,992	44,438	48,913	39,125
Mar 2017	44,058	46,296	44,093	43,001	43,633	48,743	42,895	49,718	30,200
Apr 2017	41,858	45,423	41,868	41,988	41,530	47,411	41,707	49,002	37,875
May 2017	41,767	46,171	41,785	41,146	41,310	48,416	40,786	49,637	28,500
Jun 2017	40,348	45,528	40,356	40,459	39,965	47,653	40,069	49,073	28,500
Jul 2017	40,299	46,079	40,308	39,901	39,905	48,230	39,508	49,574	26,500
Aug 2017	39,370	45,606	39,377	39,443	39,024	47,792	39,068	49,129	25,250
Sep 2017	39,345	46,012	39,350	39,068	39,039	48,123	38,722	49,524	24,000
Oct 2017	38,730	45,663	38,737	38,760	38,453	47,872	38,449	49,173	30,500
Nov 2017	38,718	45,962	38,721	38,506	38,500	48,062	38,234	49,485	28,750
Dec 2017	38,308	45,705	38,314	38,297	38,105	47,918	38,063	49,207	25,600

Στο Διάγραμμα 3.25 παρουσιάζονται τα βήματα που ακολουθήθηκαν για την μελέτη του δείκτη Worldscale.

Διάγραμμα 3.25




Με Ε.Μ.Π υπολογίζονται οι συντελεστές του μοντέλου ARIMA (p , d , q)
Είναι ικανοποιητικό το μοντέλο ;
ΟΧΙ
ΝΑΙ
Προσαρμόζουμε όλα τα εμφωλευμένα μοντέλα και με την Ε.Μ.Π. υπολογίζουμε τους συντελεστές
Σύγκριση μοντέλων
Δημιουργούμε τις προβλέψεις πού προκύπτουν από τα μοντέλα που προσαρμόζουμε για τους 12 μήνες που αντιστοιχούν στο τελευταίο έτος
Κριτήρια AIC / BIC Όσο μικρότερη η τιμή τόσο καλύτερο το μοντέλο
Συγκρίνουμε τις προβλεπόμενες τιμές των 12 μηνών του τελευταίου έτους που μας έδωσαν τα μοντέλα με τις πραγματικές τιμές του test set


Επιλέγουμε το καλύτερο μοντέλο
ΟΧΙ
Είναι ο διαγνωστικός έλεγχος ικανοποιητικός και τα σφάλματα λευκός θόρυβος ;
ΝΑΙ
Επιλέγουμε το αμέσως επόμενο καλύτερο μοντέλο
Το μοντέλο ενδείκνυται για προβλέψεις
ΤΕΛΟΣ

Συμπεράσματα

Σκοπός της παρούσας διπλωματικής ήταν η χρήση των αυτοπαλινδρομικών μοντέλων ARIMA για την πρόβλεψη του δείκτη Worldscale για πλοία VLCC μεταφορικής ικανότητας 280.000 DWT. Οι προαναφερθείσες μέθοδοι μελετήθηκαν τόσο σε θεωρητικό όσο και σε πρακτικό επίπεδο. Μέσω της ανάλυσης και της εφαρμογής τους εξήγαμε κάποια πολύ ενδιαφέρονται αποτελέσματα.

Αρχικά είδαμε ότι με την βοήθεια των μοντέλων αυτών μπορούμε να πάρουμε μια γενική εικόνα για το αν θα υπάρχει άνοδος η κάθοδος των ναύλων τους επόμενους μήνες. Ωστόσο, δεν ήταν ικανά να δώσουν καλές εκτιμήσεις των μελλοντικών χρονικών στιγμών και όσο μεγάλωνε ο χρονικός ορίζοντας της πρόβλεψης, τόσο χανόταν η ακρίβεια των προβλέψεων. Συνεπώς, βλέπουμε ότι τα μοντέλα ARIMA αδυνατούν να μελοποιήσουν περίπλοκους δείκτες όπως αυτός του Worldscale και ένας λόγος που συμβαίνει αυτό είναι ότι αποτυγχάνουν να συμπεριλάβουν υπόψιν τους τη μεταβλητότητα που παρουσιάζεται σε μια χρονοσειρά. Ο λόγος που μερικές από τις εκτιμήσεις που μας έδωσαν τα μοντέλα ARIMA ήταν ικανοποιητικές, είναι γιατί τα τελευταία χρόνια ο δείκτης δεν παρουσιάζει έντονες διακυμάνσεις και φαίνεται να κυμαίνεται σε σταθερά επίπεδα. Συνεπώς, αφού δεν είχαμε τον παράγοντα της έντονης μεταβλητότητας για τα τελευταία χρόνια, δεν είχαμε πολύ μεγάλες αποκλίσεις στις προβλέψεις μας από τις πραγματικές τιμές του δείκτη. Σε περίπτωση που θέλαμε να χρησιμοποιήσουμε τα μοντέλα αυτά στην περίοδο των έντονων διακυμάνσεων των ναύλων (από το 2003- 2008), η εφαρμογή τους θα αποτύγχανε πλήρως.

Μια μεγάλη ποικιλία μοντέλων ικανά να περιγράψουν την μεταβλητότητα που συναντάται σε μια χρονοσειρά είναι τα GARCH (Generalized Autoregressive Conditional Heteroskedasticity). Τα μοντέλα αυτά, ορίζονται να είναι ετεροσκεδαστικά, δηλαδή ικανά να περιγράψουν χρονοσειρές στις οποίες η διακύμανση τους μεταβάλλεται σε συνάρτηση με τον χρόνο. Κατά συνέπεια, τα μοντέλα αυτά υπολογίζουν την διασπορά της προ μελέτης μεταβλητής δεδομένου της διασποράς της ίδιας μεταβλητής σε προηγούμενες χρονικές στιγμές (δηλαδή υπολογίζουν δεσμευμένες διασπορές). Από την άλλη πλευρά, τα μοντέλα ARIMA υπολογίζουν την αναμενόμενη τιμή της σειράς, δεδομένου τις τιμές που είχε τις προηγούμενες χρονικές στιγμές (δηλαδή υπολογίζουν δεσμευμένες μέσες τιμές).

Ακόμη, σε περίπτωση που η προς μελέτη μεταβλητή παρουσιάζει ενδείξεις εποχικότητας όπως αυτά που είχε το σύνολο δεδομένων των αεροεπιβατών της Pan Am, θα έπρεπε να χρησιμοποιηθεί μια άλλη κατηγορία μοντέλων τα οποία ονομάζονται SARIMA (Seasonal Autoregressive Integrated Moving Average) και είναι ικανά να μοντελοποιήσουν σειρές που παρουσιάζουν παρόμοιες συμπεριφορές σε συγκεκριμένα χρονικά διαστήματα.

Εκτός από τα αυτοπαλινδρομικά αυτά μοντέλα, υπάρχει και μια ακόμα κατηγορία μοντέλων, που ονομάζονται VAR (Vector Autoregresive) και τα οποία είναι σε θέση να χρησιμοποιήσουν πληροφορία όχι μόνο από τις παρελθοντικές τιμές της προ μελέτης μεταβλητής αλλά και από άλλες χρονοσειρές που δρουν παράλληλα με αυτήν που μελετάμε. Στο σύστημα της ναυτιλίας λόγω τόσο των εξωσυστημικών παραμέτρων που δρουν στη διαμόρφωση των ναύλων (όπως για παράδειγμα η ζήτηση για μεταφορική ικανότητα) όσο και στους εσωσυστημικούς παραμέτρους (όπως για παράδειγμα η προσφορά για μεταφορική ικανότητα) που τους επηρεάζουν, είναι αναγκαίο να λαμβάνεται πληροφορία από πολλές άλλες χρονοσειρές ταυτόχρονα.

Ωστόσο, όλες οι προαναφερθέντες κατηγορίες μοντέλων μπορούν να δώσουν ικανοποιητικά αποτελέσματα σε περίπτωση που θέλουμε να κάνουμε βραχυπρόθεσμες προβλέψεις στο μέλλον ενώ αδυνατούν να κάνουν μακροπρόθεσμες προβλέψεις, ακόμα και αν το σύνολο δεδομένων που χρησιμοποιούμε για προσαρμογή είναι πολύ μεγάλο.

Επιπροσθέτως, τα μοντέλα ARIMA προϋποθέτουν η μεταβλητή που μελετάμε να παρουσιάζει σειριακή εξάρτηση μεταξύ των παρελθοντικών τιμών της και μάλιστα γραμμική. Τέλος, απαιτείται, έπειτα από την προσαρμογή του μοντέλου τα υπόλοιπα να είναι λευκός θόρυβος και να μην έχει μείνει άλλη πληροφορία σε αυτά που δεν έχει χρησιμοποιηθεί στο μοντέλο.

Τα προβλήματα που αναφέραμε παραπάνω, λύνονται με την βοήθεια μοντέλων με τη χρήση μηχανικής μάθησης (Machine Learning Time Series Analysis) τα οποία βασίζονται σε περίπλοκους αλγορίθμους. Πιο συγκεκριμένα, το μοντέλο LSTM (Long Short-Term Memory) που ανήκει στην κατηγορία των επαναλαμβανόμενων νευρωνικών δικτύων είναι ένα ευρέως διαδεδομένο μοντέλο ικανό να λύσει το πρόβλημα της μη γραμμικής σειριακής εξάρτησης και προτιμάται στην περίπτωση όπου έχουμε ένα πολύ μεγάλο σύνολο δεδομένων για επεξεργασία και θέλουμε να πραγματοποιήσουμε μακροπρόθεσμες προβλέψεις. Ωστόσο, το μειονέκτημά τους είναι ότι είναι αρκετά χρονοβόρα στη εκτέλεση τους και περίπλοκα ως προς την κατανόηση τους (Namin & Namin, 2018).

Μια τεχνική που αναπτύχθηκε το 2004 με Τεχνητά Νευρωνικά Δίκτυα και ήταν ικανή να προβλέψει τις τιμές του δείκτη Worldscale για τρείς, έξι και εννέα μήνες στο μέλλον, χρησιμοποιώντας πληροφορία όχι μόνο από τις παρελθοντικές τιμές της προ μελέτης μεταβλητής αλλά ταυτόχρονα και από ενδογενείς μεταβλητές (προσφορά μεταφορικής ικανότητα, τιμές νεότευκτων πλοίων, νέες παραγγελίες κ.λ.π) αλλά και εξωγενείς μεταβλητές (ζήτηση μεταφορικής ικανότητας, παραγωγή πετρελαίων χωρών OPEC κ.λ.π) , ήταν η τεχνική FORESIM (Lyridis, et al., 2004). Σε αντίθεση με τα ARIMA μοντέλα, αυτή η τεχνική έδειξε ότι τα τεχνητά νευρωνικά δίκτυα ξεπερνούν την αδυναμία των ARIMA μοντέλων για βραχυπρόθεσμες προβλέψεις.

Βιβλιογραφία

Anderson, T. W. & Darling, D. A., 1954. A Test of Goodness of Fit. *Journal of the American Statistical Association*, **49**, pp. 765-769.

Bartlett, M. S., 1946. On the Theoretical Specification and Sampling Properties of Autocorrelated Time-Series. *Supplement to the Journal of the Royal Statistical Society*, **8**, pp. 27-41.

Box, G. E. P., Jenkins, G. M. & Reinsen, G. C., 2008. *Time Series Analysis: Forecasting and Control.* 4th ed. New Jersey: Wiley & Hoboken.

Box, G. E. P. & Pierce, D. A., 1970. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series, **65**, pp. 1509-1526.

Brockwell, P. J. & Davis, R. A., 2002. *Introduction to Time Series and Forecasting.* 2nd ed. New York: Springer.

Brockwell, P. J. & Davis, R. A., 2006. *Time Series: Theory and Methods.* 2nd ed. New York: Springer Science & Business Media.

Clarkson Research: The Shipping Intelligence Network. [Online]
Available at: www.clarksons.net

Conover, W. J., 1999. Practical Nonparametric Statistics. *John Wiley & Sons*, pp. 428-433.

Cowpertwait, P. S. P. & Metcalfe, A. V., 2009. *Introductory Time Series with R.* New York: Springer Science & Business Media.

Dettling, M., 2013. *Applied Time Series Analysis.* Zurich: University of Applied Sciences.

Hamilton, J. D., 1994. *Time Series Analysis.* New Jersey: Prinston University Press.

Hannan, . E. . J., 1980. The Estimation of the Order of an ARMA Process. *The Annals of Statistics*, **8**, pp. 1071-1081.

Hurvich, C. M. & Tsai, C.-L., 1989. Regression and time series model selection in small samples. *Biometrika*, **76**, pp. 297-307.

Hyndman, R., Athanasopoulos, G., Bergmeir, C. & Caceres, G., 2018. *Forecasting Functions for Time Series and Linear Models.* [Online]
Available at: https://cran.r-project.org/web/packages/forecast/forecast.pdf

Jarque, C. M. & Bera, A. K., 1980. Efficient tests for normality, heteroscedasticity. *Economics Letters*, **6**, pp. 255-259.

Jones, R. H., 1975. Fitting Autoregressions. *Journal of the American Statistical Association*, **70**, pp. 590-592.

Kedem, B. & Fokianos, K., 2002. *Regression Models for Time Series Analysis.* New Jersey: Wiley & Hoboken.

Kwiatkowski, D., Phillips, P. C., Schmidt, P. & Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. *Journal of Econometrics*, **54**, pp. 159-178.

Lehmann, E. L. & Casella, G., 1983. *Theory of Point Estimation.* New York: Springer.

Ljung , G. M. & Box, G. E. P., 1978. On a Measure of Lack of Fit in Time Series Models. *Biometrica*, **65**, pp. 297-303.

Lyridis, D., Zacharioudakis, P., Mitrou, P. & Mylonas, A., 2004. Forecasting Tanker Market Using Artificial Neural Networks. *Maritime Economics and Logistics,* **6**(2), pp. 93-108.

McLeod, A. I. & Li, W. K., 1983. Diagnostic checking ARMA time series models. *Journal of Time Series Analysis*, **4**, pp. 269-273.

Montgomery, D. J., Jennings, C. L. & Kulahki, M., 2016. *Introduction to Time Series Analysis and Forecasting.* 4th ed. New Jersey: Wiley & Hoboken.

Namin, A. S. & Namin, S. S., 2018. *Forecasting Economic And Financial Time Series: ARIMA vs LSTM,* Texas: Texas Tech University.

Nielsen, H. B., 2005. *Non-Stationary Time Series and Unit Root Tests.* [Online]
Available at:
https://pdfs.semanticscholar.org/2d0e/62db75bdeafd2277fab1039f4866aef642b7.pdf

Said, S. E. & Dickey, D. A., 1984. Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika*, **71**, pp. 599-607.

Schwert, W. G., 1978. Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, pp. 461-464.

Shibata, R., 1976. Selection of the Order of an Autoregressive Model by Akaike's Information Criterion. *Biometrika*, **63**, pp. 117-126.

Shumway, R. H. & Stoffer, D. S., 2011. *Time Series Analysis and Its Applications-With R Examples.* 3rd ed. New York: Springer Science & Business Media.

Stopford, M., 2003. Supply, Demand and Freight Rates. In: *Maritime Economics.* London and New York: Taylor & Francis Ltd, pp. 135-209.

Trapletti, A., Hornik, K. & LeBaron, B., 2018. *Time series Analysis and Computational Finance.* [Online]
Available at: https://cran.r-project.org/web/packages/tseries/tseries.pdf

Wickham, H., Chang, W. & Henry , L., 2018. *Create Elegant Data Visualisations Using the Grammar of Graphics.* [Online]
Available at: https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf

Ζαχαριουδάκης, Π., 2007. *Ανάπτυξη Εργαλείων Λήψης Αποφάσεων στη Ναυτιλία.* Αθήνα: Εθνικό Μετσόβιο Πολυτεχνείο.

Καρώνη, Χ. & Οικονόμου, Π., 2017. *Στατιστικά Μοντέλα Παλινδρόμησης.* 2 εκδ. Αθήνα: Συμεών.

Κοκολάκης, Γ., 2010. *Ανάλυση χρονοσειρών.* Αθήνα: Εθνικό Μετσόβιο Πολυτεχνείο.

Κοκολάκης, Γ. & Φουσκάκης, Δ., 2009. *Στατιστική Θεωρία & Εφαρμογές.* Αθήνα: Συμεών.

Ψαραύτης, Χ., 2005. *Οικονομική Θαλάσσιων μεταφορών Ι.* Αθήνα: Εθνικό Μετσόβιο Πολυτεχνείο.