ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Δ.Π.Μ.Σ: ΜΑΘΗΜΑΤΙΚΗ ΠΡΟΤΥΠΟΠΟΙΗΣΗ ΣΕ ΣΥΓΧΡΟΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΚΑΙ ΤΗ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ

ΑΝΑΛΥΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ ΧΡΟΝΟΣΕΙΡΩΝ ΜΕ ΜΟΝΤΕΛΑ ARIMA ΚΑΙ ΕΦΑΡΜΟΓΕΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χασιρτζόγλου Μάρκος

Επιβλέπουσα καθηγήτρια: Καρώνη Χρυσηίς, Καθηγήτρια, Τομέας Μαθηματικών, Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών, Εθνικό Μετσόβιο Πολυτεχνείο.

ΑΘΗΝΑ
ΙΟΥΝΙΟΣ 2020

Ευχαριστίες

Σε αυτό το σημείο, θα ήθελα να ευχαριστήσω ιδιαιτέρως την καθηγήτρια μου κύρια Χρυσηίς Καρώνη, Καθηγήτρια του Εθνικού Μετσόβιου Πολυτεχνείου, για την εμπιστοσύνη που μου έδειξε αλλά και για τη βοήθεια της σε κάθε στάδιο της παρούσας διπλωματικής εργασίας. Επίσης θα ήθελα να ευχαριστήσω θερμά τον κύριο Αλέξανδρο Ρήγα, Ομότιμο Καθηγητή του Δημοκριτείου Πανεπιστημίου Θράκης, για την πολύτιμη συμβολή του στην ολοκλήρωση και στη διεκπεραίωση της παρούσας διπλωματικής εργασίας.

Περίληψη

Στην εποχή μας, το αντικείμενο των προβλέψεων σημειώνει ραγδαία εξέλιξη σε πληθώρα διαφορετικών επιστημονικών πεδίων. Η ανάπτυξη και διάθεση ελευθέρων λογισμικών, όπως η προγραμματιστική γλώσσα R, ενισχύει ιδιαίτερα τη δραστηριότητα αυτή. Σε αυτά τα πλαίσια, σκοπός της παρούσας διπλωματικής εργασίας είναι η βασική κατανόηση και εκμάθηση της επιστήμης των προβλέψεων γενικότερα, καθώς και η χρήση των γραμμικών αυτοπαλίνδρομων μοντέλων ARIMA όπως και η εφαρμογή αυτών με δεδομένα από τρείς διαφορετικούς επιστημονικούς κλάδους.

Αρχικά στο πρώτο κεφάλαιο, γίνεται μια εισαγωγική αναφορά γύρω από την έννοια της χρονοσειράς περιλαμβάνοντας τον ορισμό, τον διαχωρισμό των χρονοσειρών και παραδείγματα πραγματικών χρονοσειρών.

Έπειτα στο δεύτερο κεφάλαιο , μελετώνται βιβλιογραφικά τα χαρακτηριστικά των χρονοσειρών. Εκεί, γίνεται εκτενής ανάλυση για τα κυριά δομικά στοιχεία μιας χρονοσειράς όπως η τάση , η κυκλικότητα, η εποχικότητα και οι ακραίες τιμές. Επιπλέον, παρουσιάζονται οι βασικοί στατιστικοί δείκτες για την μελέτη της χρονοσειράς και γίνεται αναφορά για τη βασικότερη προϋπόθεση στην ανάλυση χρονοσειρών, την στασιμότητα.

Στην συνέχεια της θεωρητικής προσέγγισης των χρονοσειρών, στο τρίτο κεφάλαιο, γίνεται μια κατηγοριοποίηση των προβλέψεων και ακολουθούν τα διαφορά είδη απλών χρονοσειρών που στην τελική σύνθεση τους δημιουργούν τα μοντέλα ARIMA.

Στο τελευταίο κομμάτι της παρούσας εργασίας, στο τέταρτο κεφάλαιο, γίνεται εφαρμογή των μοντέλων για τρεις διαφορετικές χρονοσειρές και παρατίθενται αναλυτικά όλα τα εργαλεία που χρησιμοποιούνται μαζί με τα αποτελέσματα. Ακολουθούν τα συμπεράσματα από την εφαρμογή των μοντέλων.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Χρονοσειρές, Προβλέψεις, Αυτοπαλίνδρομα μοντέλα, Εφαρμογή μοντέλων ARIMA

Abstract

Nowadays, the issue of predictions is evolving rapidly in a variety of different scientific fields. The development and availability of free software such as the R programming language greatly enhances the activity. In this context, the purpose of this postgraduate thesis is the basic understanding and learning of predictions in general, and further, the use of linear autoregressive ARIMA models and their application with data from three different scientific fields.

In the first chapter, it is made an introductory reference around the concept of time series including the definition, the separation of time series and examples of real time series.

Then in the second chapter, the characteristics of time series are studied using the available bibliography. Here, an extensive analysis is made of main structural elements of a time series such as trend, cyclic, seasonality and extreme values. In addition, the basic statistical indicators for the study of time series are presented and reference is made to the most basic condition in time series analysis, stationarity.

In the third chapter, a categorization of predictions is made and there are different types of simple time series which in their final composition are creating ARIMA models.

In the last part of the present work, in the fourth chapter, there are three applications using ARIMA models for three different time series. Here, also, are presented in detail all the tools used together with the results of modeling. At last, there are the conclusions from the applications of modeling.

KEY WORDS: Time series, Forecasting, Autoregressive models, ARIMA applications

Εισαγωγή

Στο αντικείμενο των προβλέψεων κύριος στόχος είναι η κατά το δυνατόν μεγαλύτερη ευστοχία και η ελαχιστοποίηση της απόκλισή τους από τις πραγματικές μελλοντικές τιμές. Με αυτόν τον τρόπο, οι μέθοδοι πρόβλεψης συμβάλλουν στη σωστή και έγκαιρη λήψη αποφάσεων.

Τα τελευταία χρόνια, υπάρχει μεγάλο ενδιαφέρον ως προς την εξόρυξη δεδομένων και την μελέτη τους. Ο μεγαλύτερος όγκος δεδομένων δίνεται στη μορφή χρονοσειρών. Η πρόβλεψη μέσω των χρονοσειρών χρησιμοποιείται ευρέως και συχνά σε πλήθος επιστημονικών κλάδων. Παραδείγματος χάρη, στην στατιστική, στην οικονομετρία (οικονομετρικά μοντέλα), στην μετεωρολογία (πρόγνωση του καιρού), στην σεισμολογία, στην ιατρική (ηλεκτροεγκεφαλογράφημα), στην αστρονομία, στη μηχανική επικοινωνιών και δικτύων. Δηλαδή σε μεγάλο βαθμό σε οποιοδήποτε τομέα των εφαρμοσμένων επιστημών και της μηχανικής που περιλαμβάνονται χρονικές μετρήσεις.

Τα δεδομένα, λοιπόν, περιέχουν όλη τη πληροφορία αλλά χρειάζεται και η κριτική ικανότητα του επιστήμονα, για να επιτευχθούν ικανοποιητικά ακριβείς προβλέψεις. Στην πράξη, εγκυμονούν παράγοντες που εισάγουν σφάλματα στην παραπάνω διαδικασία. Ένας τέτοιος παράγοντας είναι η αβεβαιότητα. Η αβεβαιότητα αναφέρεται σε απρόσμενα γεγονότα, που επηρεάζουν την πραγματική μελλοντική πορεία της χρονοσειράς υπό εξέταση, και έχει σαν αποτέλεσμα τα διάφορα μοντέλα να εμφανίζουν από μικρή απόκλιση, έως και ολική αστοχία. Επίσης, ένας επιπλέον παράγοντας είναι ότι τα δεδομένα μπορεί να μην είναι αρκετά σε αριθμό για τη διενέργεια μιας σωστής εκτίμησης. Σε τέτοιες περιπτώσεις, η εμπειρία, η ευρηματικότητα και η κρίση του επιστήμονα θα προβάλλει λύσεις.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Κεφάλαιο 1: Εισαγωγή στις χρονοσειρές

1.1 Ορισμός χρονοσειράς

Ως χρονοσειρά ή χρονολογική σειρά (time series στην αγγλική γλώσσα) ορίζεται ένα σύνολο δεδομένων, τα οποία συλλέγονται διαχρονικά και εκφράζουν την εξέλιξη των τιμών μιας μεταβλητής κατά τη διάρκεια διαδοχικών χρονικών περιόδων. Ειδικότερα, η χρονοσειρά αποτελείται από ένα σύνολο παρατηρήσεων μιας μεταβλητής, οι τιμές της οποίας είναι ιεραρχημένες με βάση τη χρονική περίοδο στην οποία αναφέρονται, π.χ. έτος, τρίμηνο, μήνας κ.α. Βασική προϋπόθεση για την μελέτη χρονοσειρών αποτελεί ο τρόπος μέτρησης των τιμών της χρονοσειράς ώστε να δημιουργείται η δυνατότητα σύγκρισης των δεδομένων. Εννοώντας ότι, αν η χρονοσειρά αποτελείται από λόγου χάρη ημερήσια δεδομένα, τότε η μέτρηση της κάθε τιμής θα πρέπει να γίνεται την ιδιά χρονική στιγμή μέσα στην μέρα και στην ιδιά τοποθεσία, αν αυτό επηρεάζει την ιδιά την τιμή.

Σύμφωνα με τα Μαθηματικά, μια χρονοσειρά ορίζεται από τις τιμές $Y_1, Y_2, \ldots Y_n$ κάποιας μεταβλητής Y κατά τις χρονικές στιγμές $t_1, t_2, \ldots t_n$. Επομένως το Y είναι μια συνάρτηση του t, και συμβολίζεται ως Y=F(t). Η γραφική παράσταση της συνάρτησης Y=F(t) παρουσιάζει την εξέλιξη της μεταβλητής Y στο χρόνο. Ενδιαφέρον παρουσιάζει η γραφική παράσταση μίας χρονοσειράς που περιγράφεται σαν την κίνηση ενός σημείου καθώς κυλάει ο χρόνος. Η βιβλιογραφική εμπειρία αποκαλύπτει κάποιες μορφές χαρακτηριστικών κινήσεων (characteristic movements), που εμφανίζονται συνηθέστερα σε κάποιο βαθμό. Η μετέπειτα ανάλυση και μελέτη αυτών των κινήσεων είναι η κινητήριος δύναμη που σε πολλές περιπτώσεις μας ωθεί στην ορθή πρόβλεψη (forecasting) των μελλοντικών τιμών της χρονοσειράς.

1.2 Γενικός διαχωρισμός χρονοσειρών

Αρχικά, οι χρονοσειρές διακρίνονται σε συνεχείς και σε διακριτές ανάλογα τον χρόνο λήψης δεδομένων για την εξέλιξη του φαινομένου. Συνεχείς χρονοσειρές είναι

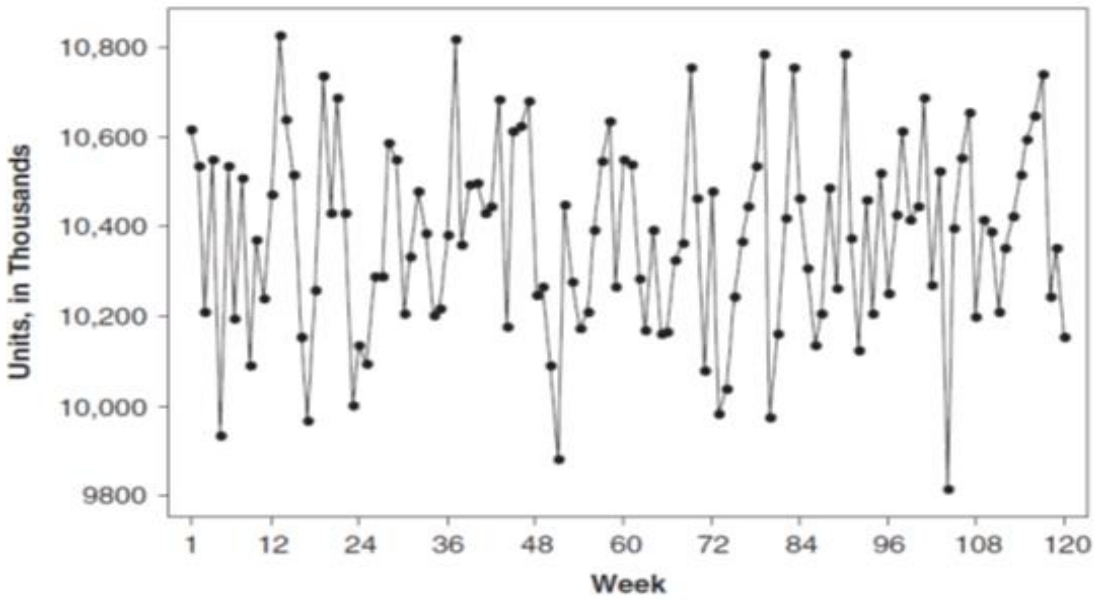
αυτές που η τιμή του φαινομένου παρατηρείται συνεχώς. Διακριτές χρονοσειρές είναι αυτές όπου η τιμή του φαινομένου καταγράφεται σε ορισμένα χρονικά διαστήματα.

Σε δεύτερο στάδιο, θα μπορούσε να γίνει διαχωρισμός των χρονοσειρών σε δύο βασικές κατηγορίες βάσει των διαδικασιών που καθορίζουν τις επόμενες τιμές των χρονοσειρών. Συνεπώς θα μπορούσαν να κατηγοριοποιηθούν σε ντετερμινιστικές, οι οποίες έχουν ως κύριο χαρακτηριστικό ότι οι διαδοχικές αυτές παρατηρήσεις δεν είναι ανεξάρτητες μεταξύ τους, αλλά οι μελλοντικές τιμές μπορούν να προσδιοριστούν από τις προηγούμενες και τις στοχαστικές, στις οποίες οι τιμές των μελλοντικών παρατηρήσεων προκύπτουν από μια στοχαστική διαδικασία και δεν περιγράφονται πλήρως από το παρελθόν των αντίστοιχων τιμών. Ωστόσο η πραγματικότητα διαφέρει πάντα της θεωρίας, οπότε για την μελέτη πραγματικών χρονοσειρών, δηλαδή δεδομένων που αντιπροσωπεύουν ένα πραγματικό μέγεθος, η εξέλιξη αυτών των μεγεθών είναι εν γένει άγνωστη και χρήζει προσοχής. Δηλαδή, η πλειοψηφία των πραγματικών χρονοσειρών επηρεάζονται από πολλαπλούς τυχαίους παράγοντες και έτσι το μέλλον καθορίζεται μόνο μερικώς από το παρελθόν. Γενικεύοντας λοιπόν, οι χρονοσειρές αντιπροσωπεύουν στοχαστικές διαδικασίες και πρέπει να γίνει αντιληπτή η σημασία του πλήρους προσδιορισμού των παραγόντων αν είναι τυχαίοι.

1.3 Εφαρμογές και παραδείγματα χρονοσειρών

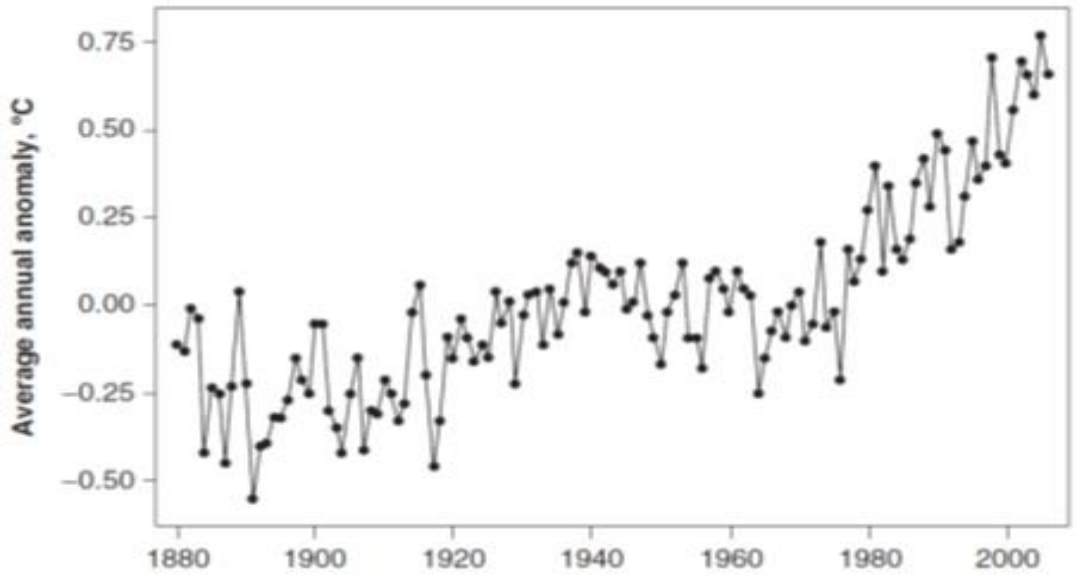
Καθώς τα φαινόμενα, στην πλειοψηφία τους, είναι δυναμικά με το χρόνο, το πεδίο εφαρμογών και χρήσης των χρονοσειρών είναι εξαιρετικά ευρύ. Οι χρονοσειρές βρίσκουν εφαρμογές σε διάφορα επιστημονικά πεδία, όπως τα Οικονομικά, την Ιατρική, τη Χημεία, την Εγκληματολογία, την Κοινωνιολογία κ.ά. Παραδείγματος χάρη: οι ημερήσιες αεροπορικές και οδικές αφίξεις τουριστών στην χώρα, ο αριθμός πελατών μέσα σε ένα πολυκατάστημα, ο συνολικός αριθμός τροχαίων ατυχημάτων κατά μήκος μιας οδικής αρτηρίας, η ημερήσια κατανάλωση ηλεκτρικού ρεύματος, η ημερήσια κατανάλωση ύδατος σε μια μεγάλη γεωγραφική περιοχή και η πορεία μια χημικής αντίδρασης. Σαν οικονομικές χρονοσειρές παρατηρούνται: το ετήσιο ακαθάριστο εθνικό προϊόν, το ετήσιο ισοζύγιο εξωτερικών συναλλαγών και οι προβλέψεις ζήτησης-προσφοράς ενός προϊόντος. Δίνονται παραδειγματικά χρονοσειρές με τις γραφικές τους παραστάσεις παρακάτω. Στην Εικόνα 1.1 του

Κεφαλαίου 1 φαίνονται οι εβδομαδιαίες πωλήσεις φαρμακευτικών προϊόντων σε χιλιάδες μονάδες για 120 εβδομάδες. (Montgomery, Jennings & Kulahci, 2015).



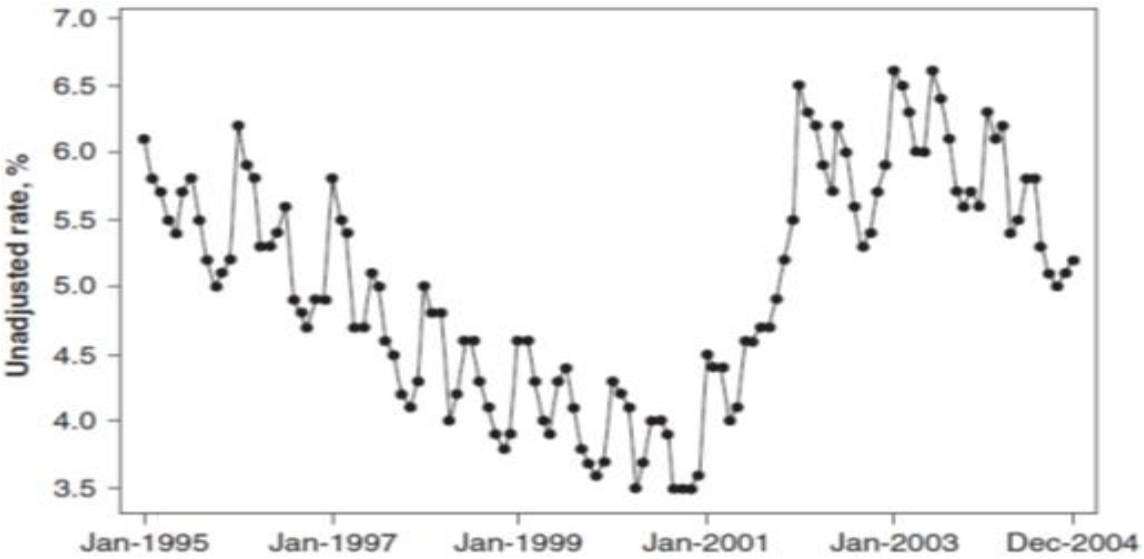
Εικόνα 1.1
Γράφημα χρονοσειράς των εβδομαδιαίων πωλήσεων φαρμακευτικών προϊόντων σε χιλιάδες μονάδες.

Η Εικόνα 1.2 αφορά τη παγκόσμια μέση θερμοκρασία επιφάνειας αέρα ανά έτος μετρημένη σε βαθμούς Κελσίου (1880-2000) (Montgomery, Jennings & Kulahci, 2015).



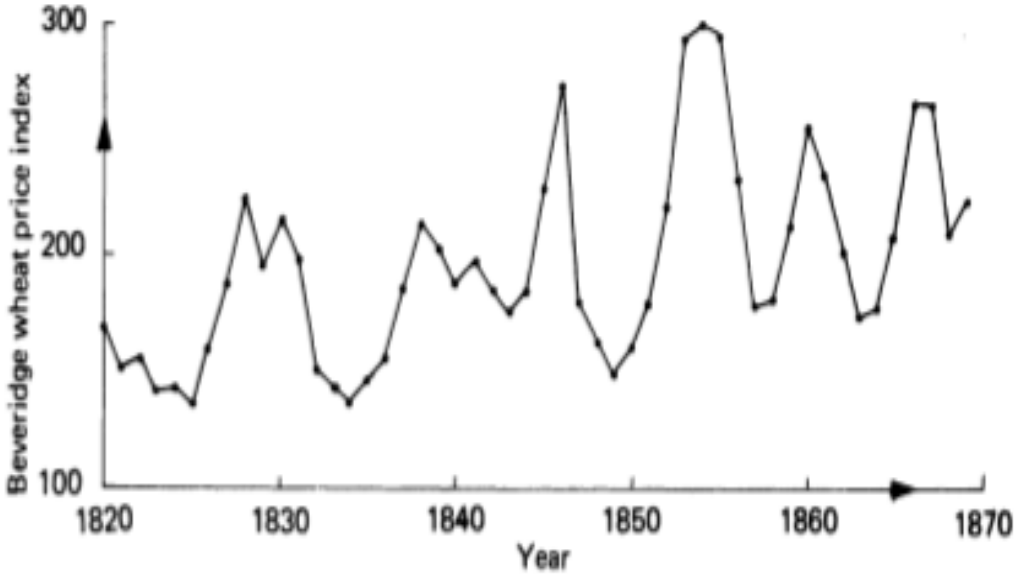
Εικόνα 1.2
Γράφημα χρονοσειράς της παγκόσμιας μέσης θερμοκρασίας επιφάνειας αέρα ανά έτος.

Η Εικόνα 1.3, απεικονίζεται το μηνιαίο ποσοστό ανεργίας σε εργατικό δυναμικό από το 1995 έως το 2004 (Montgomery, Jennings & Kulahci, 2015).



Εικόνα 1.3
Γράφημα χρονοσειράς του μηνιαίου ποσοστού ανεργίας σε εργατικό δυναμικό.

Στο τελευταίο παράδειγμα, Εικόνα 1.4, παρουσιάζεται η τιμή του σιταριού στο Beveridge από το 1820 έως το 1870. (Chatfield, 2003)



Εικόνα 1.4
Γράφημα χρονοσειράς της ετήσιας τιμής του σιταριού στο Beveridge.

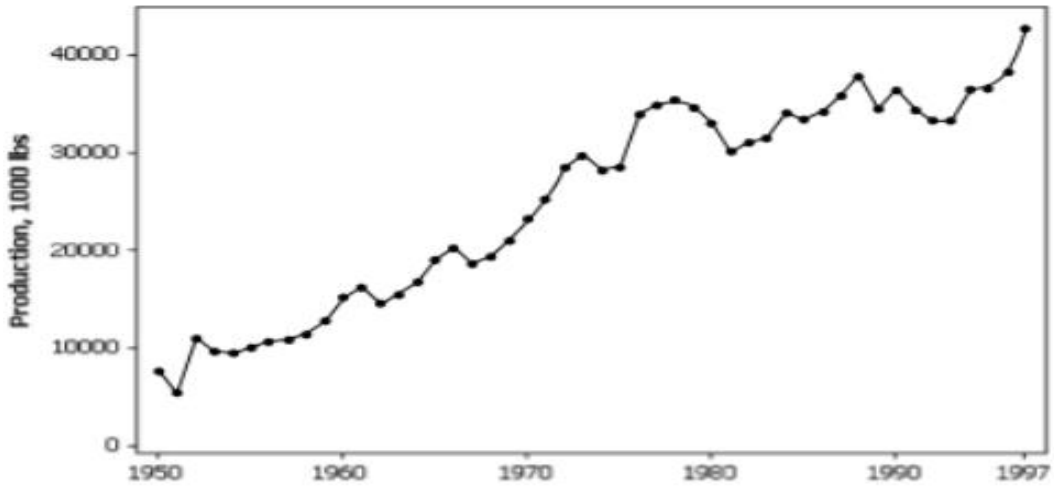
Κεφάλαιο 2: Γενικά χαρακτηριστικά χρονοσειρών

2.1 Βασικά χαρακτηριστικά και ανάλυση χρονοσειρών

Για την ύπαρξη μιας χρονοσειράς είναι απαραίτητο να αναφερθούν οι παρελθοντικές τιμές που περιγράφουν το αντίστοιχο μέγεθος. Οι τιμές του μεγέθους αποτελούν και την ιστορία του ή επικρατέστερα ονομάζονται ιστορικά δεδομένα. Συμβαίνει όμως να εμφανίζεται μεγάλος όγκος δεδομένων για την περιγραφή απλά και μόνο ενός μεγέθους. Έτσι, για την κατανόηση των δεδομένων και στη συνέχεια για την διάκριση των βασικών χαρακτηριστικών, βασική ανάγκη είναι η αναπαράσταση των ιστορικών δεδομένων. Ως αποτέλεσμα, η δισδιάστατη γραφική απεικόνιση των πραγματικών τιμών των διαθέσιμων δεδομένων, σα συνάρτηση του χρόνου, αποτελεί ένα πολύ σημαντικό εργαλείο τόσο για την ανάλυση της αντίστοιχης χρονοσειράς όσο και για τη διαδικασία πρόβλεψης. Από την οπτικοποίηση, λοιπόν, των παρατηρήσεων καθίσταται ευκολότερη η διαδικασία αναγνώρισης των ποιοτικών χαρακτηριστικών της χρονοσειράς όπως η τάση, η κυκλικότητα, η εποχικότητα και οι ακανόνιστες διακυμάνσεις που θα αναλυθούν εκτενέστερα παρακάτω.

- Τάση

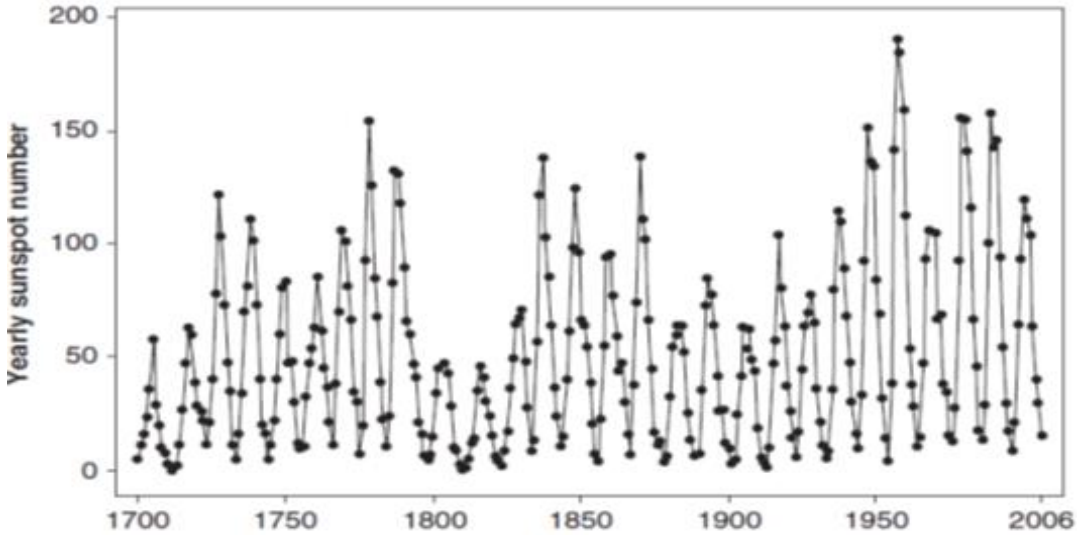
Η τάση (trend) θα μπορούσε να ορισθεί ως μια μακροπρόθεσμη μεταβολή του μέσου επιπέδου των τιμών μιας χρονοσειράς. Δηλαδή, η τάση των τιμών ενδέχεται να είναι ανοδική ή πτωτική ή σταθερή συνεχόμενα για ένα συγκεκριμένο χρονικό διάστημα. Κάποιες φόρες δύναται να εκτιμηθεί από διάφορες οικογένειες καμπυλών, όπως μια ευθεία γραμμή, ένα πολυώνυμο ανώτερης τάξης ή μια εκθετική καμπύλη. Για να γίνει έλεγχος για το αν μια σειρά παρουσιάζει τάση θα πρέπει να υπάρχει ένας ικανός αριθμός παρατηρήσεων και να εκτιμηθεί σε ένα κατάλληλο χρονικό διάστημα. Η Εικόνα 2.1 δείχνει την ετήσια παραγωγή τυριού gorgonzola στις Η.Π.Α. από το 1950 έως το 1997 μετρημένη σε λίβρες (Montgomery, Jennings & Kulahci, 2015). Διαπιστώνεται εύκολα στην παρακάτω εικόνα ότι η χρονοσειρά παρουσιάζει ανοδική-αυξητική τάση.



Εικόνα 2.1
Γράφημα χρονοσειράς της ετήσιας παραγωγής τυριού gorgonzola στις Η.Π.Α.

- Κυκλικότητα

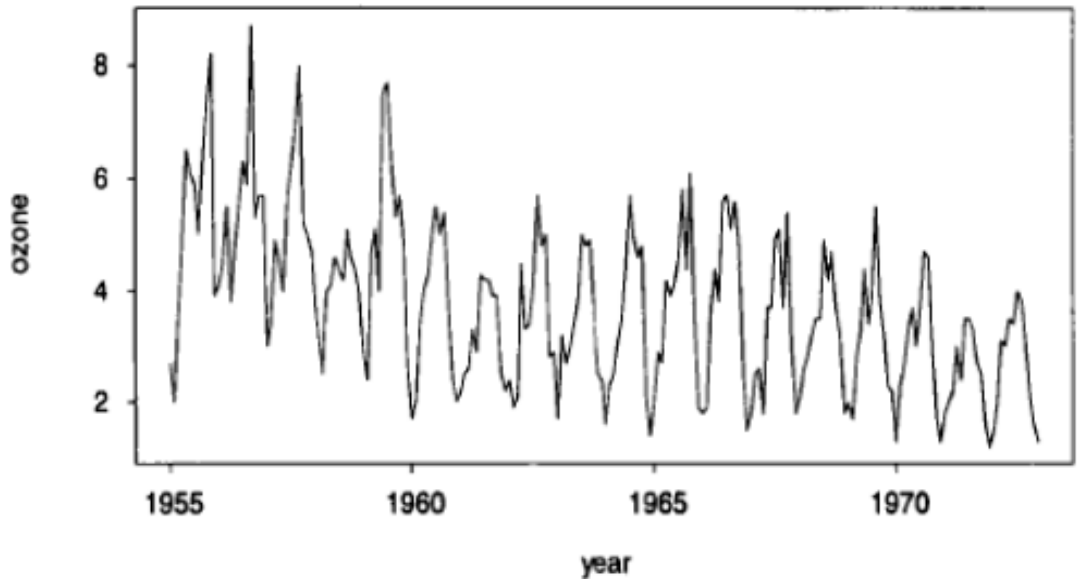
Η κυκλικότητα (cyclic) αναφέρεται σε ταλαντώσεις γύρω από τη γραμμή ή καμπύλη τάσης. Αντικατοπτρίζει μια μεταβολή που εμφανίζεται λόγω εξωγενών παραγόντων, όπως οικονομικές και τεχνολογικές εξελίξεις, κατά τη διάρκεια μεγάλων χρονικών περιόδων. Συνήθως, οι περίοδοι έχουν χρονική διάρκεια μεγαλύτερη του έτους, χωρίς όμως αυτό να σημαίνει ότι είναι σταθερού χρονικού βήματος (Brillinger, 1981). Παρουσιάζεται στις γραφικές παραστάσεις των χρονοσειρών ως μια κυματοειδής γραμμή που κινείται ανάμεσα στην υψηλότερη και χαμηλότερη στάθμη. Στην Εικόνα 2.2 παρατηρείται ο αριθμός των ηλιακών κηλίδων από το 1700 έως το 2006 και διαφαίνονται κυκλικά μοτίβο διάφορων μεγεθών που επαναλαμβάνονται κάθε περίπου δέκα χρόνια (Montgomery, Jennings & Kulahci, 2015).



Εικόνα 2.2
Γράφημα χρονοσειράς του ετήσιου αριθμού ηλιακών κηλίδων.

- Εποχικότητα

Η εποχικότητα (seasonal) μπορεί να εκφραστεί σαν μια περιοδική διακύμανση η οποία έχει σταθερό, μικρότερο ή ίσο μήκος ενός έτους. Η διακύμανση αυτή είναι άμεσα κατανοητή και διακριτή με το μάτι, διότι τα δεδομένα ορισμένων χρονοσειρών επαναλαμβάνονται ιδιοτρόπως σε τακτά χρονικά διαστήματα. Είναι εύκολα αντιληπτό σε χρονοσειρές, όπως η ποσότητα κατανάλωσης του πετρελαίου θέρμανσης, η οποία είναι μεγαλύτερη κατά τους χειμερινούς μήνες κάθε έτους και όπως η μηνιαία κατανάλωση παγωτού η οποία είναι μεγαλύτερη κατά την καλοκαιρινή περίοδο σε σχέση με την χειμερινή (Brillinger, 1981). Εφόσον, η εποχική διακύμανση είναι ένα χαρακτηριστικό εύκολα οπτικά αναγνωρίσιμο που μπορεί να απομονωθεί, δεν επηρεάζει τα δεδομένα μας. Η Εικόνα 2.3 δείχνει τις μηνιαίες μετρήσεις ατμοσφαιρικού όζοντος στο Λος Άντζελες από το 1955 έως το 1975(Peña, Tiao & Tsay, 2001). Παρατηρείται, λοιπόν, ότι το ατμοσφαιρικό όζον, που αποτελεί δείκτη της ατμοσφαιρικής ρύπανσης, παρουσιάζει έντονη εποχικότητα, η οποία είναι υψηλή κατά τους καλοκαιρινούς μήνες και χαμηλή τον χειμώνα.



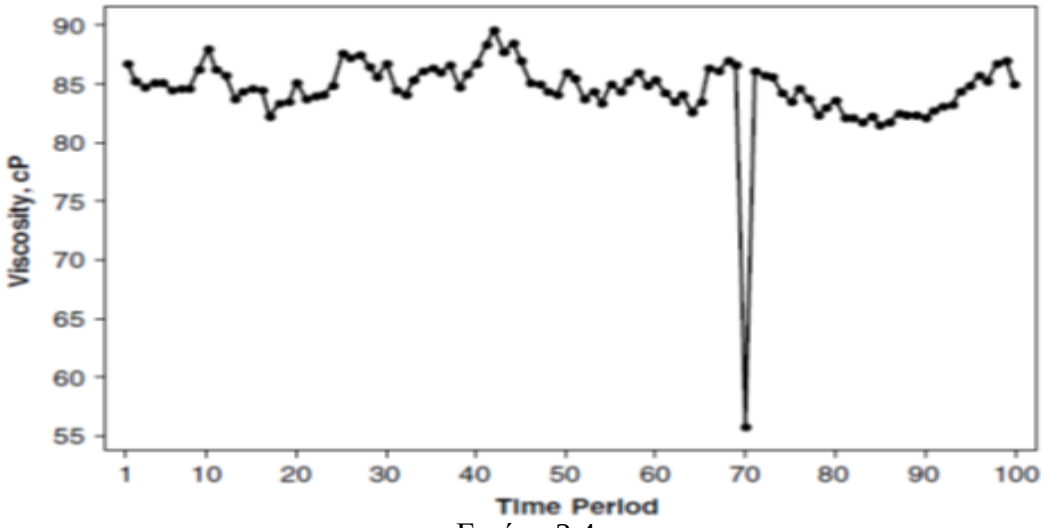
Εικόνα 2.3
Γράφημα χρονοσειράς των μηνιαίων μετρήσεων όζοντος στο Λος Άντζελες.

- Μη κανονικές διακυμάνσεις ή τυχαιότητα

Οι μη κανονικές διακυμάνσεις είναι κύριο χαρακτηριστικό των περισσότερων χρονοσειρών πού αντιμετωπίζεται αρκετά δύσκολα. Συνήθως, αυτές οι διακυμάνσεις αντιπροσωπεύουν την επιρροή μιας στοχαστικής διαδικασίας στην εξέλιξη του υπό μελέτη μεγέθους ή κάποια ασυνέχεια που συνδέεται με κάποιο εξαίρετο γεγονός.

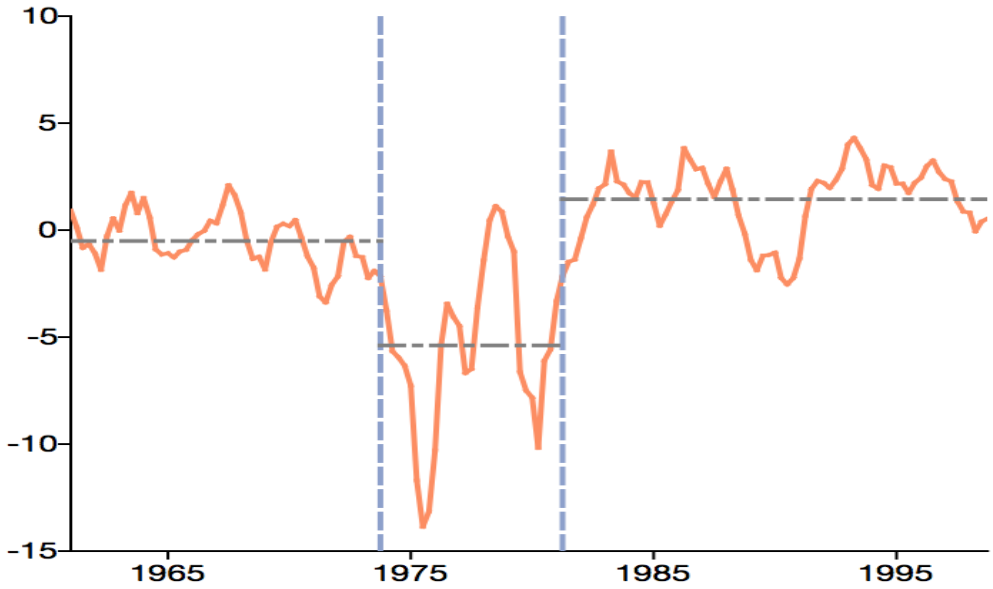
Ακριβώς λόγω της στοχαστικής φύσης της εμφάνισης και της μεταβολής των παραμέτρων που προκαλούν αυτές τις διακυμάνσεις, εν γένει θεωρούνται ως εκείνες που απομένουν όταν η τάση, η κυκλικότητα και η εποχικότητα έχουν απομακρυνθεί.

Οι ασυνέχειες είναι απότομες αλλαγές που εμφανίζονται στην εξέλιξη της χρονοσειράς και δε δύναται να προβλεφθούν από τα ιστορικά δεδομένα της αντίστοιχης χρονοσειράς. Αυτές οι απότομες αλλαγές μπορεί να έχουν περιοδικό ή μόνιμο χαρακτήρα και ανάλογα ονομάζονται. Οι ασυνέχειες με παροδικό χαρακτήρα, έχει επικρατήσει να λέγονται, σύμφωνα με την αγγλική ορολογία, ως special events ή outliers και βασικό χαρακτηριστικό τους είναι η μικρής χρονικής διάρκειας επίδρασή τους. Η αναγνώριση τους δεν είναι τόσο εύκολη διαδικασία είτε εξαιτίας της σύντομης διάρκειάς τους είτε λόγω των διάφορων άλλων παραγόντων που επηρεάζουν την ίδια χρονοσειρά τις ίδιες χρονικές περιόδους. Συνεπώς, η κατάταξη κάποιων τιμών ως special events απαιτεί εξίσου θεωρητική γνώση, κριτική ικανότητα και εμπειρία από την πλευρά του ερευνητή. Χαρακτηριστικό παράδειγμα ενός outlier είναι η δραματική πτώση της παραγωγής ενός εργοστασίου λόγω απεργίας του προσωπικού. Το outlier, γενικότερα, αντιπροσωπεύει κάποιο εξαίρετο και απρόβλεπτο γεγονός όπως είναι η απεργία η κάποια άλλη θεομηνία, όπως ακραία καιρικά φαινόμενα (σεισμός, πλημμύρες). Δεν είναι προγραμματισμένο γεγονός, όπου κάποιος θα μπορούσε να προβλέψει μήτε την διάρκεια του μήτε την ένταση του, παρ' όλα αυτά το γεγονός επιδρά με έντονο τρόπο στο αποτέλεσμα. Δίνεται παράδειγμα με την Εικόνα 2.4 όπου φαίνεται το γράφημα αναγνώσεων ιξώδους χημικής διεργασίας με αισθητήρα δυσλειτουργίας (Montgomery, Jennings & Kulahci, 2015).



Εικόνα 2.4
Γράφημα ιξώδους χημικής διεργασίας με αισθητήρα δυσλειτουργίας.

Στον αντίποδα, οι ασυνέχειες με πιο μόνιμο χαρακτήρα ονομάζονται level-shifts. Οι ασυνέχειες αυτές εμφανίζονται ομοίως ως απότομες αλλαγές αλλά η επιρροή τους έγκειται σε αλλαγή του μέσου επιπέδου τιμών της χρονοσειράς. Δηλαδή, ένα παράδειγμα ασυνέχειας level – shift, είναι η πτώση των πωλήσεων μίας εταιρείας λόγω εισαγωγής στην αγορά μιας ανταγωνίστριας εταιρείας. Συνηθέστερα, μετά από την απότομη μείωση του αριθμού των πωλήσεων την χρονική στιγμή εισόδου της ανταγωνίστριας εταιρείας, έπεται απλά η σταθεροποίηση του αριθμού των πωλήσεων, σταθερά σε χαμηλότερο επίπεδο. Τροχοπέδη αποτελεί η κατηγοριοποίηση των διαφόρων γεγονότων και συχνά δημιουργούνται προβληματισμοί και σύγχυση στο διαχωρισμό τους, εξού και απαιτείται υψηλό γνωστικό επίπεδο όσο και κοινή λογική. Σύμφωνα με το Aptech Systems (2020) παρατίθεται παρακάτω παράδειγμα ασυνέχειας level-shift. Η Εικόνα 2.5 απεικονίζει γραφικά τα φορολογικά προσαρμοσμένα πραγματικά επιτόκια στο Ηνωμένο Βασίλειο από το 1960 έως το 2000 (Aptech Systems, 2020).



Εικόνα 2.5
Γράφημα των φορολογικά προσαρμοσμένων πραγματικών επιτοκίων στο Ηνωμένο Βασίλειο.

Αντικείμενο της ανάλυσης χρονοσειρών είναι ο διαχωρισμός των κύριων χαρακτηριστικών των χρονοσειρών και η απομόνωσή τους. Τα κύρια αυτά χαρακτηριστικά των χρονοσειρών όπως έχουν ήδη αναφερθεί είναι: η τάση, η κυκλικότητα, η εποχικότητα και η τυχαιότητα. Για την ανάλυση των χρονοσειρών χρησιμοποιούνται οι ακόλουθοι συμβολισμοί, όπου t = 1,2,3...,n.:

- Y_t = Πραγματική τιμή της χρονοσειράς
- T_t = Τάση
- S_t = Εποχικότητα
- C_t = Κυκλικότητα
- I_t = Τυχαίες κινήσεις

Η μαθηματική περιγραφή μιας χρονοσειράς, που φανερώνει τον τρόπο με τον οποίο οι παρατηρήσεις της χρονοσειράς προσδιορίζονται από τις τέσσερις συνιστώσες της χρονοσειράς, παρουσιάζεται από δύο μοντέλα. Τα χρησιμοποιούμενα μοντέλα είναι το προσθετικό μοντέλο (additive model) και το πολλαπλασιαστικό μοντέλο (multiplicative model).

Στο προσθετικό μοντέλο οι πραγματικές τιμές της χρονοσειράς για κάθε περίοδο θεωρούνται ως το άθροισμα των τεσσάρων συνιστωσών στο ίδιο σύστημα μονάδων μέτρησης και δημιουργούνται με τον ακόλουθο τρόπο:

$$Y_t = T_t + S_t + C_t + I_t$$

Στο πολλαπλασιαστικό μοντέλο οι πραγματικές τιμές της χρονοσειράς προσδιορίζονται από το γινόμενο των τεσσάρων συνιστωσών ανεξαρτήτου συστήματος μονάδων μέτρησης ως ακολούθως:

$$Y_t = T_t \bullet S_t \bullet C_t \bullet I_t$$

Όσον αφορά την επιλογή μοντέλου, από τα δύο παραπάνω, το πολλαπλασιαστικό μοντέλο υπερτερεί έναντι του προσθετικού μοντέλου, καθώς χρησιμοποιείται στην πληθώρα των περιπτώσεων επειδή έχει λιγότερο υπολογιστικό κόστος. Ένα άλλο βασικό μειονέκτημα του αθροιστικού είναι ότι βασίζεται στην υπόθεση ότι οι συνιστώσες της χρονοσειράς είναι ανεξάρτητες μεταξύ τους. Τις περισσότερες φορές είναι άτοπη η υπόθεση γιατί οι συνιστώσες αλληλεξαρτώνται και κυρίως αυτό παρατηρείται στις οικονομικές χρονοσειρές. Επιπλέον, υποθετικά αν είναι αποδεκτό ότι η τάση δεν επηρεάζει τον εποχικό συντελεστή θα εμφανιστεί σχεδόν βέβαια πρόβλημα στις μακροχρόνιες προβλέψεις. Οι συνηθέστερες εξαιρέσεις είναι τα φυσικά φαινόμενα όπου η τάση συνήθως δεν επηρεάζει μεταξύ των άλλων τις εποχικές

διακυμάνσεις. Καταλήγοντας, προτιμάται βιβλιογραφικά το πολλαπλασιαστικό μοντέλο.

2.2 Στατιστικοί δείκτες χρονοσειρών

Είναι γνωστό ότι μια χρονοσειρά είναι η πραγματοποίηση μιας στοχαστικής διαδικασίας, δηλαδή αποτελείται από ένα σύνολο παρατηρήσεων-τυχαίων μεταβλητών. Οι βασικοί στατιστικοί δείκτες των χρονοσειρών, που είναι άρρηκτα συνδεδεμένοι με τα βασικά χαρακτηριστικά των χρονοσειρών, είναι: η μέση τιμή, η αυτοσυνδιακύμανση και η αυτοσυσχέτιση. Οι ορισμοί και οι συμβολισμοί αναφέρονται παρακάτω (Hamilton, 1994).

- Μέση τιμή

Η μέση τιμή ή αναμενόμενη τιμή μιας χρονοσειράς Y ορίζεται ως:

$$\mu_t = E(Y_t) = \int_{-\infty}^{+\infty} y_t f_{Y_t}(y_t) dy_t$$

Η μέση τιμή μ_t σχετίζεται άμεσα με την έννοια της τάσης της χρονοσειράς, εφόσον είναι εκφρασμένη συναρτήσει της κάθε χρονικής στιγμής t της κάθε παρατήρησης Y_t. Συγκεκριμένα, αν σε μια χρονοσειρά παρουσιάζεται αυξητική ή πτωτική τάση αντίστοιχα σε ένα χρονικό διάστημα, αυτό θα αποτυπώνεται και στη μέση τιμή.

- Αυτοσυνδιακύμανση

Έστω δύο τυχαίες μεταβλητές *X* και *W*. Η συνδιακύμανση (covariance) των εν λόγω τυχαίων μεταβλητών δίνεται από την σχέση:

$$Cov(X, W) = E(X - \mu_x)(W - \mu_w)$$

Για την προς μελέτη χρονοσειρά, ορίζεται η αυτοσυνδιακύμανση (autocovariance) j-οστης τάξης γ_{tj} της τυχαίας μεταβλητής Y_t με μια προηγούμενη χρονική τυχαία μεταβλητή Y_{t-j} ως:

$$\gamma_{jt} = E(Y_t - \mu_t)(Y_{t-j} - \mu_{t-j}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty}(y_t - \mu_t)(y_{t-j} - \mu_{t-j})x$$

$$f_{Y_t, Y_{t-1},\dots, Y_{t-j}}(y_t, y_{t-1},\dots, y_{t-j})dy_t dy_{t-1}\dots dy_{t-j}$$

με από κοινού συνάρτηση πιθανότητας για κάθε τυχαία μεταβλητή(Y_t, Y_{t-1},…, Y_{t-j}) ως:

$$f_{Y_t, Y_{t-1},\dots, Y_{t-j}}(y_t, y_{t-1},\dots, y_{t-j})$$

Στην περίπτωση που j = 0, προκύπτει η διακύμανση της τυχαίας μεταβλητής Y_t ως:

$$\gamma_{0t} = E(Y_t - \mu_t)^2 = \int_{-\infty}^{+\infty}(y_t - \mu_t)^2 f_{Y_t}(y_t)dy_t$$

- Αυτοσυσχέτιση

Ο συντελεστής αυτοσυσχέτισης είναι ένας από τους πιο σημαντικούς στατιστικούς δείκτες ο οποίος χρησιμοποιείται ευρέως στην ανάλυση χρονοσειρών για την διαπίστωση ύπαρξης τυχαιότητας ή μη στην χρονοσειρά. Η αυτοσυσχέτιση (autocorrelation) j-οστής τάξης ρ_{jt} της τυχαίας μεταβλητής Y_t με μια προηγούμενη χρονική εκδοχή της Y_{t-j} ορίζεται για κάθε t ως εξής:

$$\rho_{jt} = \frac{\gamma_{jt}}{\gamma_{0t}} = \frac{E(Y_t - \mu_t)(Y_{t-j} - \mu_{t-j})}{E(Y_t - \mu_t)^2}$$

Το σύνολο τιμών της αυτοσυσχέτισης ρ_{jt} είναι το [-1,1]. Η αυτοσυσχέτιση είναι ένας καταλυτικής σημασίας στατιστικός δείκτης στη μελέτη χρονοσειρών, διότι έτσι δίνεται ένα μέτρο για τον βαθμό της σχέσης μεταξύ των δύο μεταβλητών. Ειδικότερα, φτιάχνοντας το γράφημα της αυτοσυσχέτισης συναρτήσει της καθυστέρησης j, το οποίο ονομάζεται συνάρτηση αυτοσυσχέτισης (autocorrelation function- ACF) δύναται να αποσαφηνιστούν στοιχεία για τα χαρακτηριστικά της χρονοσειράς. Αν η τιμή του ρ_{jt} ισούται περίπου με +1 ή -1, τότε οι παρατηρήσεις Y_t και Y_{t-j} είναι ισχυρά συσχετισμένες. Ενώ, αν ρ_{jt} ισούται με 0, τότε εύκολα προκύπτει ότι οι παρατηρήσεις Y_t και Y_{t-j} είναι ασυσχέτιστες. Συνεπώς, με γνώμονα την αυτοσυσχέτιση μπορούν να μελετηθούν τα ποιοτικά χαρακτηριστικά των χρονοσειρών, όπως η εποχικότητα και η στασιμότητα. Στην επόμενη ενότητα λοιπόν θα γίνει ενδελεχέστερη μελέτη στην έννοιας της στασιμότητας και ό,τι συνεπάγεται αυτή.

Όπως και η συνάρτηση αυτοσυσχέτισης, η συνάρτηση μερικής αυτοσυσχέτισης (partial autocorrelation function-PACF) αποτελεί πολύτιμη πηγή πληροφόρησης σχετικά με τα χαρακτηριστικά της αλληλεξάρτησης που δημιουργεί μια στοχαστική διαδικασία. Οι συντελεστές μερικής αυτοσυσχέτισης μετρούν το βαθμό της σχέσης μεταξύ των Y_t και Y_{t-k} όταν οι επιδράσεις όλων των άλλων χρονικών υστερήσεων 1,2,3, ...,k-1 έχουν αφαιρεθεί. Ο συντελεστής μερικής αυτοσυσχέτισης τάξης k συμβολίζεται με α_k και μπορεί να υπολογισθεί εφαρμόζοντας τη μέθοδο της πολλαπλής γραμμικής παλινδρόμησης με εξαρτημένη μεταβλητή την Y_t και ανεξάρτητες μεταβλητές τις Y_{t-1}, ... , Y_{t-k}:

$$Y_t = b_0 + b_1 Y_{t-1} + \ldots + b_k Y_{t-k}$$

Ο συντελεστής α_k ισούται με τον συντελεστή b_k. Ο πρώτος συντελεστής μερικής αυτοσυσχέτισης α_1 είναι πάντα ίσος με τον πρώτο συντελεστή αυτοσυσχέτισης r_1.

2.3 Στασιμότητα

Γενικά, μια χρονοσειρά λέγεται στάσιμη αν θεωρηθεί ότι οι στατιστικές της ιδιότητες παραμένουν αμετάβλητες στο χρόνο. Απλοποιημένα, όταν δεν υφίσταται συστηματική αλλαγή του μέσου όρου και της διασποράς της χρονοσειράς στο χρόνο (Yao & Herbert, 2009).

Μία χρονοσειρά λέγεται αυστηρώς στάσιμη (strictly stationary) ή πρώτης τάξης αν η από κοινού κατανομή (joint distribution) των τυχαίων μεταβλητών $Yt_1,\ldots,Yt_n$ είναι ίδια με την από κοινού κατανομή των $Yt_{1+\tau}, \ldots Yt_{n+\tau}$ για όλα τα $t_1,...,t_n$. Με άλλα λόγια, μετατοπίζοντας την αρχή του χρόνου κατά ένα χρονικό διάστημα, έστω τ, δεν μεταβάλλεται η από κοινού κατανομή. Αυτή πρέπει να εξαρτάται μόνο από τα διαστήματα μεταξύ των $t_1,...,t_n$. Ο παραπάνω ορισμός ισχύει για οποιαδήποτε τιμή του n. Για τον ορισμό της στασιμότητας μιας χρονοσειράς ένας άλλος τρόπος είναι ο υπολογισμός των ροπών της χρονοσειράς. Ειδικότερα, εάν n=1, συνεπάγεται ότι η κατανομή της Y_t πρέπει να είναι ίδια για όλα τα t, έτσι ώστε: $\mu(t) = \mu$ και $\sigma^2(t) = \sigma^2$. Δηλαδή η μέση τιμή και η διασπορά είναι σταθερές και δεν εξαρτώνται από την τιμή του t.

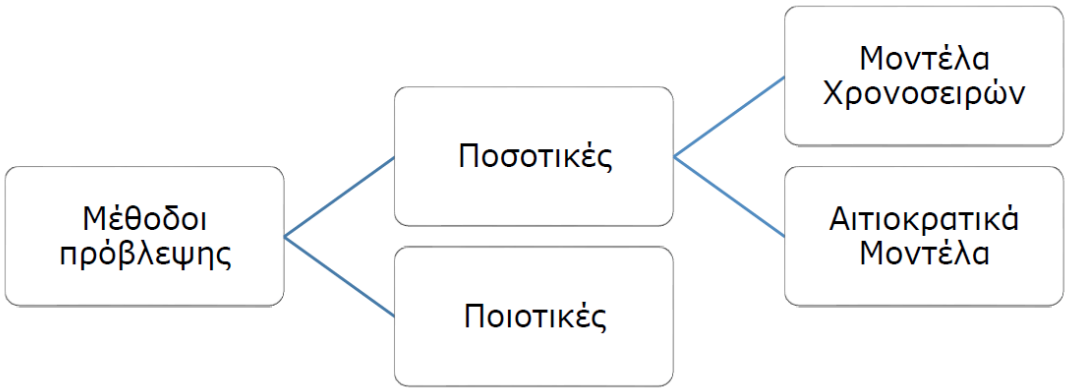
Επί του πρακτέου οφείλει να οριστεί η στασιμότητα μ' ένα λιγότερο περιοριστικό τρόπο από τον παραπάνω. Μια χρονοσειρά καλείται ασθενώς στάσιμη ή δεύτερης τάξης, αν η μέση τιμή της είναι σταθερή και αν η αυτοδιασπορά εξαρτάται μόνο από την καθυστέρηση έτσι ώστε: $\mu(t) = \mu$ και $cov\{Y_t, Y_{t+\tau}\} = \gamma(\tau)$. Προϋποθέσεις για ροπές μεγαλύτερης τάξης απ' αυτές της δευτέρας τάξης δε γίνονται. Επιπρόσθετα, πρέπει να σημειωθεί ότι η μέση τιμή και η διασπορά είναι πεπερασμένες.

Η ύπαρξη μη στασιμότητας είναι ένα από τα βασικά προβλήματα στην ανάλυση χρονοσειρών και από τα πρώτα που είναι επιτακτική ανάγκη να λυθεί. Οι περισσότερες χρονοσειρές είναι μη στάσιμες αφού εμπεριέχουν τάση, εποχικότητα και κυκλικές κυμάνσεις. Η μη στασιμότητα αυτών των χρονοσειρών αποτελεί εμπόδιο για την ανάλυση τους και περαιτέρω την πρόβλεψη, αλλά μπορούν με κατάλληλες τεχνικές και μαθηματικά εργαλεία να μετατραπούν σε στάσιμες.

Κεφάλαιο 3: Πρόβλεψη και μέθοδοι πρόβλεψης χρονοσειρών

3.1 Γενικά περί πρόβλεψης-Κατηγορίες προβλέψεων

Οι μέθοδοι πρόβλεψης συμβάλλουν στην λογικά ορθή και έγκαιρη λήψη αποφάσεων καθώς και στον πετυχημένο σχεδιασμό. Κατά κύριο λόγο, αν τα δεδομένα, που χρησιμοποιούνται, είναι υψηλής ποιότητας, δύναται να υπάρχουν ικανοποιητικά ακριβείς προβλέψεις. Επειδή η πράξη υπολείπεται της θεωρίας γενικά, οι προβλέψεις επηρεάζονται από ποικίλους παράγοντες. Για παράδειγμα, τα δεδομένα δεν είναι πάντα ακριβή και αξιόπιστα. Μπορεί επιπλέον, να μην είναι αρκετά σε αριθμό για τη διενέργεια μιας σωστής εκτίμησης. Ακόμη, είναι ευρέως γνωστό πως το παρελθόν δεν είναι πάντα ο σωστός οδηγός εκτίμησης για το μέλλον. Επίσης σημαντική επιρροή προς το υπό μελέτη φαινόμενο μπορεί να ασκούνται από απρόβλεπτους εξωτερικούς παράγοντες. Έτσι λοιπόν, διακρίνονται δύο κατηγορίες προβλέψεων που περιγράφονται σχηματικά παρακάτω με το Σχήμα 3.1.



Σχήμα 3.1
Κατηγοριοποίηση μεθόδων προβλέψεων.

Για τις ποσοτικές μεθόδους προβλέψεων, ισχύει ότι πρέπει να τηρούνται κάποιες βασικές προϋποθέσεις. Εννοώντας, να είναι διαθέσιμη και επαρκής η πληροφορία για το παρελθόν, η πληροφορία να μπορεί να ποσοτικοποιηθεί και τέλος να ισχύει η υπόθεση σταθερότητας. Δηλαδή ότι τουλάχιστον κάποιοι παράγοντες και εκφάνσεις του παρελθόντος θα επαναληφθούν μελλοντικά.

Όσον αφορά τα μοντέλα των χρονοσειρών, σε αντίθεση με το αιτιοκρατικό μοντέλο δεν αναλύουν τη σχέση που υπάρχει μεταξύ της υπό εξέταση μεταβλητής και άλλων ανεξάρτητων μεταβλητών αλλά θεωρούν το συνολικό σύστημα ως ένα «μαύρο» κουτί και δεν ασχολούνται καθόλου με τους παράγοντες που πιθανόν να το επηρεάζουν. Έτσι, η πρόβλεψη για το μέλλον στηρίζεται αποκλειστικά σε ιστορικά δεδομένα του παρελθόντος της υπό εξέτασης μεταβλητής και σε σφάλματα που προκύπτουν από αυτά. Ο σκοπός της χρήσης χρονοσειρών ως προβλεπτικά μοντέλα είναι η εύρεση ενός μοντέλου που ακολουθεί τις αντίστοιχες παρατηρήσεις και μελλοντικά η προέκτασή του. Η πρόβλεψη λοιπόν με χρονοσειρές αποτελεί αναμφίβολα πρόκληση για να βρεθεί ο τρόπος με τον οποίο η ακολουθία των παρατηρήσεων θα συνεχιστεί στο μέλλον, με απώτερο στόχο να ακολουθηθεί μια διαδικασία που θα εξασφαλίσει ότι θα παραχθούν όσο το δυνατόν ακριβέστερες προβλέψεις, εκμεταλλευόμενες σε υψηλό βαθμό όλη την διαθέσιμη ιστορική πληροφορία. Η ανάπτυξη των μοντέλων χρονοσειρών και η εφαρμογή τους αποτελούν και το αντικείμενο της παρούσας εργασίας.

3.2 Ο αλγόριθμος της πρόβλεψης

Βιβλιογραφικά, τα βασικά στάδια σε μια διαδικασία πρόβλεψης είναι τα εξής (Αγιακόγλου & Οικονόμου, 2004):

I. Καθορισμός του προβλήματος

Επί το πλείστον, αποτελεί το πιο δύσκολο μέρος στη διαδικασία πρόβλεψης και ταυτόχρονα είναι μέγιστης σημασίας. Αυτό συμβαίνει διότι θα πρέπει να καταστεί σαφής και κατανοητός ο τρόπος με τον οποίο θα χρησιμοποιηθούν οι προβλέψεις και από ποιους.

II. Συγκέντρωση πληροφοριών

Σε αυτό το βήμα απαιτούνται τουλάχιστον δύο είδη πληροφοριών. Αρχικά, είναι οι πληροφορίες που αφορούν τα στατιστικά (αριθμητικά) δεδομένα και δεύτερον η κριτική ικανότητα και η εμπειρία του προσωπικού με αυτή τη συλλογή δεδομένων. Τέλος, οι παραπάνω πληροφορίες πρέπει να συλλεχθούν πριν ξεκινήσει η διαδικασία της πρόβλεψης.

III. Προκαταρτική ανάλυση

Στο τρίτο στάδιο γίνεται αναφορά για το είδος της πληροφορίας που αποκομίζεται από τα ακατέργαστα ιστορικά δεδομένα. Εν αρχή, αναπαρίστανται γραφικά τα δεδομένα και στη συνέχεια, υπολογίζονται οι βασικοί στατιστικοί δείκτες, όπως τυπική απόκλιση, μέση τιμή, ελάχιστη τιμή, μέγιστη τιμή και γραμμική τάση. Οι παραπάνω δείκτες αναδεικνύουν κάποια δευτερεύοντα χαρακτηριστικά της χρονοσειράς. Με αυτόν τον τρόπο δημιουργείται μία «προαίσθηση» για τα δεδομένα, δίνοντας απαντήσεις σε ερωτήματα όπως αν υπάρχουν πρότυπα κυκλικότητας (cyclic patterns), αν υπάρχει σημαντική τάση ή εποχικότητα και τέλος, αν υπάρχουν ασυνήθιστες τιμές (outliers). Η ανάλυση ετούτη είναι ο οδηγός για την επιλογή οικογένειας μοντέλων πρόβλεψης που ορθολογικά αναμένεται να προσδώσει ικανοποιητικά αποτελέσματα.

IV. Επιλογή και προσαρμογή μοντέλου

Εδώ γίνεται η επιλογή και ο καθορισμός των παραμέτρων διάφορων ποσοτικών μοντέλων πρόβλεψης. Συμβαίνει συχνά να είναι απαραίτητες διάφορες προσομοιώσεις.

V. Χρήση και αποτίμηση του μοντέλου πρόβλεψης

Στο τελευταίο στάδιο, εφόσον ένα μοντέλο έχει επιλεγεί υποκειμενικά και οι παράμετροι του έχουν προηγουμένως καθοριστεί, χρησιμοποιείται ώστε να πραγματωθούν προβλέψεις. Κατά την εξέλιξη της διαδικασίας, γίνεται συνεχώς αποτίμηση των πλεονεκτημάτων και μειονεκτημάτων του μοντέλου και, εφόσον κριθεί αναγκαίο, επαναλαμβάνονται ορισμένα βήματα.

3.3 Λευκός θόρυβος

Ξεκινώντας με τα πιο απλά παραδείγματα χρονοσειρών, η απλούστερη στάσιμη χρονοσειρά ονομάζεται λευκός θόρυβος (white noise). Αποτελεί δομική λίθο για όλες τις υπόλοιπες χρονοσειρές (Engle, 1982) και ορίζεται ως:

$$Y_t = \varepsilon_t$$

Οι παρατηρήσεις μιας χρονοσειράς λευκού θορύβου ε_t έχουν μέση τιμή ίση με μηδέν και σταθερή διακύμανση όλες τις χρονικές στιγμές. Επιπλέον, όλες οι παρατηρήσεις είναι ασυσχέτιστες μεταξύ τους. Δηλαδή ισχύουν:

- $\mu_t = E(\varepsilon_t) = 0$, για κάθε t.
- $\gamma_{0t} = E(Y_t - \mu_t)^2 = E(\varepsilon_t^2) = \sigma^2$, για κάθε t.
- $\gamma_{jt} = E(Y_t - \mu_t)(Y_{t-j} - \mu_{t-j}) = E(\varepsilon_t \varepsilon_{t-j}) = 0$, για κάθε t, $j \neq 0$

Αν αντικατασταθεί η δεύτερη σχέση με την ισχυρότερη υπόθεση της ανεξαρτησίας των παρατηρήσεων ε_t, τότε προκύπτει η ανεξάρτητη (independent) χρονοσειρά λευκού θορύβου (Hamilton, 1994). Αν υποθετικά για τις παρατηρήσεις της χρονοσειράς λευκού θορύβου ισχύει και η συνθήκη:

$$\varepsilon_t \sim N\left(0, \sigma^2\right)$$

δηλαδή ακολουθούν κανονική Γκαουσιανή κατανομή, τότε η χρονοσειρά λέγεται Γκαουσιανός λευκός θόρυβος (Gaussian white noise).

3.4 Τυχαίος περίπατος

Ο τυχαίος περίπατος (random walk) είναι μια μη-στάσιμη χρονοσειρά, όπου κάθε στοιχείο της y_t προκύπτει όταν στο προηγούμενο στοιχείο της y_{t-1} προστεθεί μια τυχαία μεταβλητή ε_t. Δηλαδή η χρονοσειρά αποτελεί τυχαίο περίπατο (Pfaff, 2008) αν ισχύει η συνθήκη:

$$y_t = y_{t-1} + \varepsilon_t = y_0 + \sum_{s=1}^{t} \varepsilon_t$$

όπου ε_t είναι ο λευκός θόρυβος.

Αν η τιμή y_0 ισούται με μηδέν, τότε η χρονοσειρά λευκού θορύβου είναι ανεξάρτητη από το y_0. Έτσι, γίνεται σαφές ότι η μέση τιμή του τυχαίου περιπάτου είναι μηδέν και η διακύμανσή του είναι $\gamma_{0t} = E(y_t^2) = t\sigma^2$. Εύλογα, η διακύμανση αυξάνεται με το χρόνο οπότε επιβεβαιώνεται ότι ο τυχαίος περίπατος είναι μια μη-στάσιμη χρονοσειρά.

3.5 Απλή γραμμική παλινδρόμηση

Η απλή γραμμική παλινδρόμηση (simple linear regression) ορίζεται σαν μια συνάρτηση ανάμεσα σε μια εξαρτημένη μεταβλητή *Y* και σε μια ανεξάρτητη μεταβλητή *X* . Όσον αφορά τις χρονοσειρές, η εξαρτημένη μεταβλητή είναι η μεταβλητή πρόβλεψης και ως ανεξάρτητη είναι ο χρόνος. Σε αυτήν την περίπτωση υποτίθεται ότι η μεταξύ τους σχέση είναι γραμμική, παρόλο που γενικώς η υπόθεση αυτή ανταποκρίνεται σπάνια στην πραγματικότητα. Κάλο είναι για να εφαρμοστεί η μέθοδος της απλής γραμμικής παλινδρόμησης, πρώτα να γίνει μετασχηματισμός της σχέσης των δύο μεταβλητών σε γραμμική (γραμμικοποίηση). Η μέθοδος της γραμμικής παλινδρόμησης χρησιμοποιείται για την κατανόηση των σχέσεων μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών, δηλαδή ο βαθμός συσχέτισής τους. Η πρόβλεψη με την χρήση της απλής γραμμικής παλινδρόμησης δίνει μια άρτια εικόνα της μέσης και μακροπρόθεσμης συμπεριφοράς της υπό μελέτη μεταβλητής. Το μοντέλο της απλής γραμμικής παλινδρόμησης ορίζεται από την σχέση:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

όπου β_0, β_1 οι άγνωστοι συντελεστές του μοντέλου και ε το τυχαίο σφάλμα. Οι τιμές των συντελεστών β_0 και β_1 εκτιμώνται με τα b_0 και b_1 αντίστοιχα και υπολογίζονται με βάση την αρχή των ελαχίστων τετραγώνων. Εννοώντας, ότι επιλέγονται οι συντελεστές που ελαχιστοποιούν το άθροισμα των τετραγώνων των πραγματικών τιμών από τις προβλεπόμενες σε κάθε χρονική περίοδο. Οι συντελεστές b_0 και b_1 δίνονται από τους παρακάτω μαθηματικούς τύπους:

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$$b_1 = \frac{\sum_{i=1}^{n} [(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Οπότε προκύπτει το πραγματικό μοντέλο γραμμικής παλινδρόμησης για δοθέν δείγμα n ως:

$$Y_i = b_0 + b_1 X_i + e_i \quad \text{για } i = 1, 2, \ldots, n.$$

όπου e_i το υπόλοιπο (residual).

3.6 Αυτοπαλινδρούμενη χρονοσειρά (AR)

Αυτοπαλινδρούμενη χρονοσειρά (autoregressive time series) τάξης p (AR(p)) λέγεται μια χρονοσειρά όταν κάθε παρατήρηση y_t εκφράζεται ως ένα σταθμισμένο άθροισμα μιας σταθεράς *δ*, *p* καθυστερημένων εκδοχών της χρονοσειράς *y* και μιας χρονοσειράς λευκού θορύβου. Ορίζεται από την παρακάτω σχέση:

$$y_t = \delta + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \varepsilon_t$$

όπου *δ* η παράμετρος που σχετίζεται με τη μέση τιμή της χρονοσειράς, ε_t η χρονοσειρά του λευκού θορύβου και $(\varphi_1,\varphi_2,\ldots,\varphi_p)$ οι αυτοπαλινδρούμενοι παράμετροι. Εννοιολογικά, ο όρος αυτοπαλίνδρομο αναφέρεται στο ότι η σχέση αυτή αποτελεί ένα υπόδειγμα παλινδρόμησης όπου η εξαρτημένη μεταβλητή y_t παλινδρομείτε στις προηγούμενες τιμές της ίδιας της μεταβλητής y_t. Το p υποδηλώνει την τάξη του αυτοπαλίνδρομου μοντέλου και αφορά το μήκος της υστερήσεως. Εάν χρησιμοποιηθεί ο τελεστής υστέρησης *B* που ορίζεται ως:

$$B^k y_t = y_{t-k}$$

το υπόδειγμα γράφεται ως:

$$(1 - \varphi_1 B - \ldots - \varphi_p B^p) y_t = \delta + \varepsilon_t$$

ή διαφορετικά ως:

$$\Phi(B) y_t = \delta + \varepsilon_t$$

όπου Φ(*B*) το χαρακτηριστικό πολυώνυμο της AR(p). Σύμφωνα με την θεωρία, για τον έλεγχο της στασιμότητας της χρονοσειράς AR(p) είναι αποδεκτό ότι μια AR(p) είναι στάσιμη εάν οι ρίζες του χαρακτηριστικού πολυωνύμου είναι εκτός του μοναδιαίου κύκλου.

Στην περίπτωση που η τάξη του αυτοπαλίνδρομου μοντέλου ισούται με ένα (p=1), η χρονοσειρά συμβολίζεται με AR(1) και δίνεται από τον τύπο:

$$y_t = \delta + \varphi_1 y_{t-1} + \varepsilon_t$$

Ο συντελεστής φ_1 κυμαίνεται ανάμεσα στις τιμές ανάμεσα στο -1 και στο 1. Τέλος, χρησιμοποιείται ένα AR(p) μοντέλο όταν οι συντελεστές αυτοσυσχέτισης φθίνουν στο μηδέν και ενώ συγχρόνως υπάρχουν *p* σημαντικοί στατιστικά συντελεστές μερικής αυτοσυσχέτισης.

3.7 Χρονοσειρές Κινητού Μέσου (MA)

Ξεκινώντας, οι χρονοσειρές κινητού μέσου χρησιμοποιούνται για περιγραφή φαινομένων στα οποία οποιοδήποτε γεγονός δημιουργεί ένα άμεσο αποτέλεσμα που η επίδραση του διαρκεί, αν και το ίδιο το γεγονός έχει διακοπεί . Οι διαδικασίες κινητού μέσου παρατηρούνται σε διάφορα επιστημονικά πεδία και ιδιαίτερα στην οικονομετρία. Λόγου χάρη, μια απεργία επηρεάζει την οικονομία όχι μόνο την στιγμή που πραγματοποιείται, αλλά και την επόμενη χρονική περίοδο (Chatfield, 2003).

Μια χρονοσειρά *Y* καλείται χρονοσειρά κινητού μέσου τάξης q (MA(q)), όταν κάθε παρατήρηση y_t εκφράζεται ως ένα σταθμισμένο άθροισμα μιας σταθεράς μ, μιας χρονοσειράς λευκού θορύβου και q καθυστερημένων εκδοχών της χρονοσειράς λευκού θορύβου. Ορίζεται από την παρακάτω σχέση (Montgomery, Jennings & Kulahci, 2015):

$$y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}$$

όπου ε_t είναι η χρονοσειρά λευκού θορύβου, q η υψηλότερη υστέρηση των διαταραχών και οι παράμετροι μ και (θ_1, θ_2, …, θ_q), που ανήκουν στο σύνολο των πραγματικών αριθμών. Χρησιμοποιώντας τον τελεστή υστέρησης *B* το υπόδειγμα γράφεται ως:

$$y_t = \mu + (1 - \theta_1 B - \ldots - \theta_q B^q)\varepsilon_t = \mu + \Theta(B)\varepsilon_t$$

όπου Θ(*B*) το χαρακτηριστικό πολυώνυμο της MA(q). Η MA(q) ανήκει στην ομάδα των στάσιμων χρονοσειρών αφού αποτελείται από ένα πεπερασμένο άθροισμα όρων λευκού θορύβου. Γενικά, μια γραμμική διαδικασία $\{X_t\}$ είναι αυστηρά αντιστρέψιμη ως προς την μεταβλητή $\{W_t\}$ αν και μόνο αν υπάρχει πολυώνυμο απείρων όρων που ορίζεται ως:

$$\pi(B) = \pi_0 + \pi_1 B + \pi_2 B^2 + \ldots \text{ με } \sum_{j=0}^{\infty} |\pi_j| < \infty$$

και έτσι δύναται να γραφεί ως:

$$W_t = \pi(B) X_t$$

Αντίστοιχα, μια MA διαδικασία τάξης q είναι αντιστρέψιμη, όταν οι ρίζες του χαρακτηριστικού πολυωνύμου Θ(*B*) βρίσκονται εκτός του μοναδιαίου κύκλου. Εν τέλει, μια MA διαδικασία είναι αντιστρέψιμη αν δύναται να εκφραστεί ως μια

αυτοπαλίνδρομη διαδικασία με άπειρους όρους. Εάν η τάξη του μοντέλου κινητού μέσου όρου ισούται με ένα (q=1), τότε συμβολίζεται με MA(1) και διατυπώνεται από την σχέση:

$$y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1}$$

Η παρατήρηση y_t εξαρτάται από τον όρο του σφάλματος ε_t και το προηγούμενο χρονικά σφάλμα ε_{t-1}, ενώ ο συντελεστής θ_1 λαμβάνει τιμές από -1 έως 1. Σε γενικές γραμμές ένα MA(q) μοντέλο χρησιμοποιείται όταν οι συντελεστές μερικής αυτοσυσχέτισης φθίνουν εκθετικά στο μηδέν και ταυτόχρονα υπάρχουν q στατιστικά σημαντικοί συντελεστές αυτοσυσχέτισης.

3.8 Αυτοπαλίνδρομα μοντέλα κινητού μέσου όρου (ARMA)

Ορισμένες στάσιμες χρονοσειρές δεν μπορούν να μοντελοποιηθούν εξ' ολοκλήρου ως AR ή MA χρονοσειρές, αφού υπάρχει περίπτωση να παρουσιάζουν χαρακτηριστικά και από τις δύο κατηγορίες μοντέλων. Έτσι, ένα πιο γενικό μοντέλο είναι συνδυασμός ενός AR(p) μοντέλου και ενός MA(q) μοντέλου, το οποίο ονομάζεται αυτοπαλίνδρομο μοντέλο κινητού μέσου όρου τάξης (p, q). Κάθε παρατήρηση y_t μιας ARMA(p, q) χρονοσειράς Y διατυπώνεται παρακάτω ως εξής (Montgomery, Jennings & Kulahci, 2015):

$$y_t = \delta + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}$$

Εύλογα από τα προηγούμενα προκύπτει ότι οι παράμετροι φ, θ υπόκεινται στους περιορισμούς: $-1 < \varphi_i < 1$, $-1 < \theta_i < 1$. Συνεχίζοντας, με χρήση του τελεστή υστέρησης B το μοντέλο δίνεται ως εξής:

$$(1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p) y_t = \delta + (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q) \varepsilon_t$$

ή διαφορετικά ως:

$$\Phi(B) y_t = \delta + \Theta(B) \varepsilon_t$$

Η στασιμότητα της χρονοσειράς ARMA επιβεβαιώνεται από το AR μέρος, δηλαδή είναι στάσιμη αν το χαρακτηριστικό πολυώνυμο $\Phi(B)$ έχει ρίζες εκτός του μοναδιαίου κύκλου. Αντιστοίχως, η αντιστρεψιμότητα της χρονοσειράς ARMA επιβεβαιώνεται από το MA μέρος, δηλαδή είναι αντιστρέψιμη αν το χαρακτηριστικό

πολυώνυμο Θ(B) έχει ρίζες εκτός του μοναδιαίου κύκλου. Τέλος, παραδειγματικά ένα μοντέλο ARMA(1,1) δίνεται απλά από την σχέση:

$$y_t = \delta + \varphi_1 y_{t-1} + \varepsilon_t - \theta_1 \varepsilon_{t-1}$$

Για την ορθή επιλογή παραμέτρων για μια ARMA διαδικασία ακολουθεί παρακάτω ο επεξηγηματικός Πίνακας 3.1.

Πίνακας 3.1
Εκτίμηση της τάξης των AR και MA με τη βοήθεια των ACF-PACF.

Διαδικασία	ACF	PACF
AR(1)	Εκθετική μείωση : θετικές τιμές αν φ1 >0 εναλλαγή πρόσημου ξεκινώντας από αρνητική τιμή, αν φ1 < 0	Απότομος μηδενισμός μετά την περίοδο 1. Η τιμή την περίοδο 1 είναι : θετική αν φ1 > 0 αρνητική αν φ1 < 0
AR(p)	Εκθετική μείωση ή πρότυπο φθίνουσας ημιτονοειδούς συνάρτησης. Το ακριβές πρότυπο εξαρτάται από το πρόσημο και το μέγεθος των φ1 φ2 , ... , φp.	Μη μηδενικές τιμές για τις πρώτες p περιόδους και στη συνέχεια απότομος μηδενισμός.
MA(1)	Απότομος μηδενισμός μετά την περίοδο 1. Η τιμή την περίοδο 1 είναι : θετική αν θ1 < 0 αρνητική αν θ1 > 0	Εκθετική μείωση : εναλλαγή πρόσημου ξεκινώντας από θετική τιμή αν θ1 < 0 αρνητικές τιμές αν θ1 >0
MA(q)	Μη μηδενικές τιμές για τις πρώτες q περιόδους και στη συνέχεια απότομος μηδενισμός.	Εκθετική μείωση ή πρότυπο φθίνουσας ημιτονοειδούς συνάρτησης. Το ακριβές πρότυπο εξαρτάται από το πρόσημο και το μέγεθος των θ1 θ2 , ... , θq.

3.9 Μικτό ολοκληρωμένο μοντέλο ARIMA

Στην πλειονότητά τους οι χρονοσειρές δεν χαρακτηρίζονται ως στάσιμες διαδικασίες. Οι χρονοσειρές βολεύει να είναι στάσιμες διότι είναι εύκολα επεξεργάσιμες. Όταν μια χρονοσειρά μετατρέπεται σε στάσιμη, χρησιμοποιώντας τις πρώτες διαφορές η σειρά ονομάζεται ολοκληρώσιμη πρώτης τάξης και συμβολίζεται

με I(1). Εάν η χρονοσειρά μετατρέπεται σε στάσιμη χρησιμοποιώντας τις δεύτερες διαφορές, είναι ολοκληρώσιμη δεύτερης τάξης και συμβολίζεται με I(2). Γενικεύοντας, εάν d είναι ο αριθμός των διαφορών που μετατρέπει μια σειρά σε στάσιμη, η σειρά ονομάζεται ολοκληρώσιμη d τάξεως και συμβολίζεται με I(d). Με χρήση του τελεστή ολίσθησης οι πρώτες διαφορές ορίζονται ως:

$$y_t - y_{t-1} = (1 - B)y_t = \Delta y_t$$

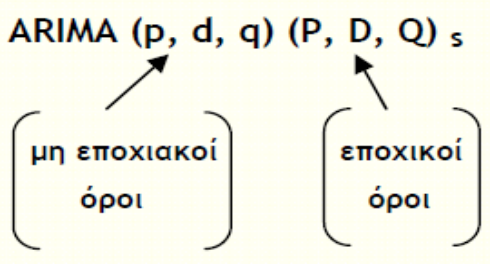
Τα στοχαστικά μοντέλα ARIMA είναι εύχρηστα καθώς ξεπερνούν το εμπόδιο της στασιμότητας που αντιμετώπιζαν τα μοντέλα ARMA, καθώς με την διαφόριση γίνεται προσπάθεια να εξαλειφθεί η στασιμότητα. Η λέξη ARIMA (Autoregressive Integrated Moving Average) ελληνιστί μεταφράζεται ως αυτοπαλίνδρομα μοντέλα κινητού μέσου και ανάλογα την τάξη, ARIMA(p, d, q) (Yao & Herbert, 2009). Επομένως, μια ARIMA(p, d, q) διαδικασία, είναι μια διαδικασία η οποία «διαφορίζεται» d φορές εξάγει μια ARMA(p, q) διαδικασία. Για ένα ολοκληρωμένο μοντελο ARIMA(p, d, q) ισχύει ότι p είναι η τάξη του αυτοπαλίνδρομου μοντέλο, d η τάξη της διαφόρισης για την επίτευξη της στασιμότητας και q η τάξη του κινητού μέσου όρου μοντέλου. Ορίζεται μαθηματικά ως (Pfaff, 2008):

$$\Phi(B)\left(1-B\right)^{d} y_t = \delta + \Theta\left(B\right)\varepsilon_t$$

Το πολυώνυμο $\Phi(B)(1-B)^d$ έχει μια ρίζα ίση με την μονάδα, τάξης d, και όλες τις άλλες εκτός μοναδιαίου κύκλου.

3.10 Εποχικό μοντέλο ARIMA

Ένα εποχικό (seasonal) μοντέλο ARIMA η αλλιώς σπανίως SARIMA, διαμορφώνεται με την εισαγωγή πρόσθετων εποχιακών όρων στο γνωστό μοντέλο ARIMA. Αυτά τα μοντέλα συμβολίζονται στο Σχήμα 3.2.



Σχήμα 3.2
Απεικόνιση εποχικού μοντέλου.

Ο όρος s απεικονίζει την εποχικότητα. Το εποχιακό μέρος του μοντέλου αποτελείται από παρόμοιους όρους με αυτούς από τις μη εποχικές συνιστώσες του μοντέλου, αλλά περιλαμβάνει μετατοπίσεις της εποχικής περιόδου. Παραδείγματος χάρη, ένα μοντέλο ARIMA (χωρίς σταθερά) που αφορά τα τριμηνιαία στοιχεία (m=4) και μπορεί να διατυπωθεί ως:

$$\left(1-\varphi_1 B\right)\left(1-\ \Phi_1 B^4\right)\left(1-B\right)\left(1-B^4\right)y_t = \ \left(1+\theta_1 B\right)\left(1+\ \Theta_1 B^4\right)e_t$$

Εύκολα εννοείται ότι η εποχικότητα του μοντέλου αποτυπώνεται στην συνάρτηση αυτοσυσχέτισης (Autocorrelation function – ACF) και στην συνάρτηση μερικής αυτοσυσχέτισης (Partial autocorrelation function – PACF). Παραδειγματικά ένα μοντέλο ARIMA$(0, 0, 0)(0, 0, 1)_{12}$ θα δείξει:

- Μια σημαντική τιμή στην καθυστέρηση 12 στην ACF αλλά όχι άλλες σημαντικές τιμές.
- Εκθετική πτώση στις εποχιακές καθυστερήσεις του PACF (π.χ. στις καθυστερήσεις (12,24,36,...).

ενώ ένα μοντελο ARIMA $(0,0,0)(1,0,0)_{12}$ θα υποδείξει:

- Εκθετική πτώση στις εποχιακές καθυστερήσεις του ACF.
- Μια μοναδική σημαντική τιμή στην καθυστέρηση 12 στην PACF .

3.11 Εκτίμηση παραμέτρων για μοντέλα ARIMA

Για την εκτίμηση των παραμέτρων ενός ARIMA μοντέλου, η μέθοδος ελαχίστων τετραγώνων είναι μια επιλογή αλλά η εφαρμογή της δυσκολεύει όταν στο μοντέλο συμπεριλαμβάνονται MA όροι ($q > 0$). Αντί αυτού, μπορεί να χρησιμοποιηθεί μια επαναληπτική μέθοδος. Αφού επιλεχθούν αρχικές τιμές για τις παραμέτρους, στη συνέχεια βελτιώνονται μέσω μιας επαναληπτικής διαδικασίας μέχρι να ελαχιστοποιηθεί το άθροισμα των τετραγώνων των σφαλμάτων. Ως επαναληπτική μέθοδος προτείνεται συχνά η μέθοδος μεγίστης πιθανοφάνειας.

Κάθε εκτιμώμενος συντελεστής έχει ένα τυπικό σφάλμα γιατί είναι μια στατιστική τιμή που βασίζεται σε πληροφορία από ένα μόνο δείγμα. Ένα διαφορετικό δείγμα πιθανόν να έδινε διαφορετικές εκτιμήσεις για τους συντελεστές του μοντέλου. Δηλαδή στην πράξη απορρίπτεται κάθε εκτιμώμενη τιμή συντελεστή με απόλυτη t-τιμή

μικρότερη του δύο. Κάθε συντελεστής με απόλυτη τιμή t < 2 δεν είναι σημαντικά διάφορος του μηδενός για επίπεδο σημαντικότητας 5% και οδηγεί στη δημιουργία μη αξιοποιήσιμων μοντέλων και επομένως σε λιγότερης ακριβείας προβλέψεις. Τέλος, αξίζει να αναφερθεί ότι το επίπεδο σημαντικότητας 5% δεν αποτελεί κάποιο μαθηματικό κανόνα αλλά επιλέγεται συνηθέστερα.

Στην πλειοψηφία τους τα στατιστικά υπολογιστικά πακέτα προσαρμόζουν αυτόματα ένα μοντέλο ARIMA στη χρονοσειρά, εκτελούν όλους τους αναγκαίους στατιστικούς ελέγχους και παραδίδουν στον χρήστη εκθέσεις σχετικά με τις τιμές όλων των στατιστικών δεικτών που χρησιμοποιούνται για τον έλεγχο της καταλληλόλητας του μοντέλου (Κουγιουμτζής, 2005).

3.12 Επιλογή του καλύτερου μοντέλου ARIMA

Εφόσον έχουν εκτιμηθεί οι παράμετροι ενός μοντέλου ARIMA είναι ανάγκη μια εκ νέου διερεύνηση, ώστε να διαπιστωθεί, εάν το επιλεγμένο μοντέλο έχει τη δυνατότητα να βελτιωθεί. Συγκεκριμένα σε αυτό το στάδιο της μοντελοποίησης πρέπει:

- Εάν προκύψουν συντελεστές στατιστικά μη σημαντικοί, οι αντίστοιχοι όροι να αφαιρεθούν από το μοντέλο. Οι ACF και PACF παρέχουν καθοδήγηση στην επιλογή ενός απλού AR ή MA μοντέλου.
- Εάν το καταλληλότερο μοντέλο είναι ένα σύνθετο ARMA μοντέλο, υπάρχει δυσκολία στο να αναγνωρισθεί από τις συναρτήσεις ACF και PACF. Αρχικά, μετά την επιλογή ενός απλού μοντέλου πρέπει να μελετηθεί εάν αυτό δύναται να επεκταθεί σε ARMA μοντέλο.
- Εάν έχουν βρεθεί περισσότερα από ένα «καλά» μοντέλα, πρέπει να εφαρμοσθεί μια μέθοδος επιλογής του καλύτερου από τα υπόλοιπα.

Ένα ικανοποιητικό κριτήριο για την επιλογή του βέλτιστου μοντέλου είναι το Akaike's Information Criterion ή AIC, το οποίο αποθαρρύνει την εισαγωγή πρόσθετων όρων στο προϋπάρχουν μοντέλο. Αν k = p + q + P + Q είναι το πλήθος των όρων, τότε επιλέγονται οι p, q, P, Q που ελαχιστοποιούν το AIC:

$$AIC = -2logL + 2k$$ όπου L η πιθανοφάνεια.

Μια άλλη εκδοχή του AIC είναι το AICc(Corrected Akaike's Information Criterion) που χρησιμοποιείται συνήθως σε μικρότερα δείγματα (ο αριθμός των παρατηρήσεων μικρότερος των σαράντα παρατηρήσεων) και ορίζεται ως:

$$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1}$$ όπου n ο αριθμός των παρατηρήσεων.

Διαφορά στις τιμές του AIC μικρότερη του δύο δεν θεωρείται σημαντική και επιλέγεται το απλούστερο μοντέλο. Χρησιμοποιούνται ευρέως επίσης αρκετές παραλλαγές του AIC όπως το BIC (Bayesian Information Criterion) και το FPE (Final Prediction Error) (Κουγιουμτζής, 2005).

3.13 Διαγνωστικός έλεγχος

Σε τελικό στάδιο εφόσον το επιλεγμένο μοντέλο θεωρείται το βέλτιστο, είναι επιτακτική ανάγκη να επιβεβαιωθεί η επάρκεια του μοντέλου. Αυτό επιτυγχάνεται με την εξέταση των υπολοίπων (σφαλμάτων) ώστε να διαπιστωθεί εάν ακολουθούν αυτά κάποιο πρότυπο.

Τα υπόλοιπα, λοιπόν, ενός καλού μοντέλου πρόβλεψης πρέπει να αποτελούν έναν «λευκό θόρυβο» ενώ συγχρόνως οι ACF και PACF των υπολοίπων δεν πρέπει να παρουσιάζουν στατιστικά σημαντικές αυτοσυσχετίσεις και αντίστοιχα μερικές αυτοσυσχετίσεις. Για να εξετασθούν συνολικά οι συντελεστές αυτοσυσχέτισης των υπολοίπων χρησιμοποιούνται διάφοροι στατιστικοί και εμπειρικοί έλεγχοι. Αναφορικά κάποια από αυτά είναι ο στατιστικός δείκτης Q * (Ljung-Box), το γράφημα Q-Q plot και ο έλεγχος Anderson-Darling. Όσον αφορά το δείκτη Q * , αν η τιμή του δεν είναι στατιστικά σημαντική τα υπόλοιπα θεωρούνται μια σειρά λευκού θορύβου. Σε περίπτωση που τα υπόλοιπα δεν αποτελούν σειρά λευκού θορύβου τότε το μοντέλο είναι ανεπαρκές και πρέπει να πραγματοποιηθεί περαιτέρω εξέταση.

Γενικεύοντας, το πρότυπο που ακολουθούν οι στατιστικά σημαντικοί συντελεστές αυτοσυσχέτισης και μερικής αυτοσυσχέτισης των υπολοίπων, υποδεικνύουν άμεσα τον τρόπο βελτίωσης του μοντέλου. Παραδείγματος χάρη, για στατιστικά σημαντικές τιμές, για εποχιακές καθυστερήσεις, προτείνεται η προσθήκη μιας εποχικής συνιστώσας. Επιπλέον, στατιστικά σημαντικές τιμές για μικρές καθυστερήσεις υποδεικνύουν κατά κανόνα την αύξηση των μη εποχιακών AR ή MA

συνιστωσών του μοντέλου. Είθισται τα μοντέλα με τις μικρότερες AIC τιμές να έχουν υπόλοιπα λευκού θορύβου. Παραδόξως αλλά περιστασιακά, συμβαίνει να υιοθετούνται μοντέλα όχι με την μικρότερη AIC τιμή αλλά αυτά με τα «καλύτερα» υπόλοιπα (Κουγιουμτζής, 2005).

Κεφάλαιο 4: Εφαρμογές

Με τη βοήθεια του επιβλέποντος, μελετήθηκαν περισσότερες από 15 χρονοσειρές και καταλήξαμε ότι, με τρεις εφαρμογές θα ασχολείται η παρούσα διπλωματική εργασία. Παρατίθεται στο παράρτημα ο τρόπος εύρεσης των δεδομένων.

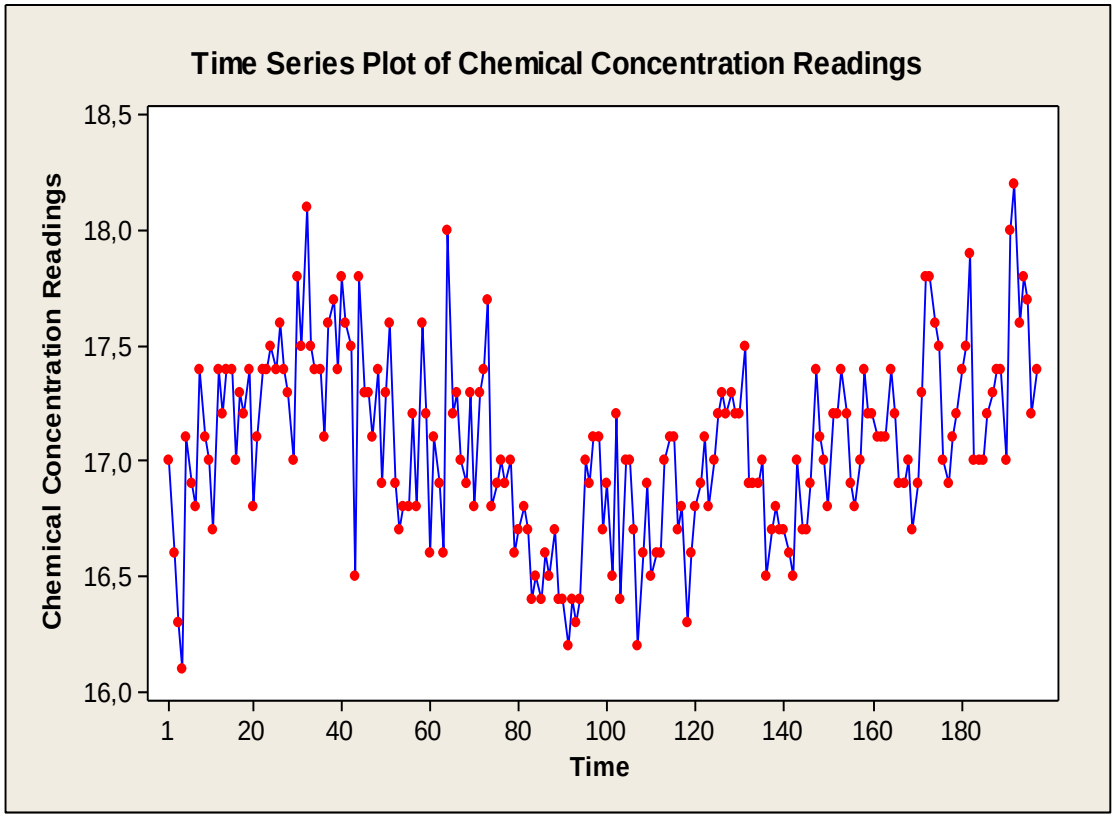
Εφαρμογή 1

4.1.1 Προκαταρτική ανάλυση για την πρώτη χρονολογική σειρά

Η πρώτη εφαρμογή αφορά την χρονολογική σειρά που αποτελείται από δεδομένα μετρήσεων χημικής συγκέντρωσης (Box & Jenkins, 1976). Ξεκινώντας, παρουσιάζονται παρακάτω τα αριθμητικά δεδομένα στον Πίνακα 4.1.1 και η γραφική αναπαράσταση της χρονοσειράς στο Διάγραμμα 4.1.1. Επιπλέον, παρουσιάζεται η στατιστική ανάλυση των δεδομένων της χρονοσειράς. Περιλαμβάνονται εκεί τα μέτρα θέσης (μέση τιμή και διάμεσος), τα μέτρα διασποράς (διακύμανση και ενδοτεταρτημοριακό εύρος Q) και τα μέτρα που περιγράφουν το σχήμα της κατανομής τους (κύρτωση και λοξότητα).

Πίνακας 4.1.1
Δεδομένα μετρήσεων χημικής συγκέντρωσης.

```
> tsdl[[162]]
Time Series:
Start = 1
End = 197
Frequency = 1
  [1] 17.0 16.6 16.3 16.1 17.1 16.9 16.8 17.4 17.1 17.0 16.7 17.4 17.2 17.4 17.4 17.0 17.3 17.
2 17.4 16.8 17.1 17.4 17.4 17.5 17.4 17.6 17.4 17.3 17.0 17.8 17.5 18.1 17.5 17.4 17.4 17.1
 [37] 17.6 17.7 17.4 17.8 17.6 17.5 16.5 17.8 17.3 17.3 17.1 17.4 16.9 17.3 17.6 16.9 16.7 16.
8 16.8 17.2 16.8 17.6 17.2 16.6 17.1 16.9 16.6 18.0 17.2 17.3 17.0 16.9 17.3 16.8 17.3 17.4
 [73] 17.7 16.8 16.9 17.0 16.9 17.0 16.6 16.7 16.8 16.7 16.4 16.5 16.4 16.6 16.5 16.7 16.4 16.
4 16.2 16.4 16.3 16.4 17.0 16.9 17.1 17.1 16.7 16.9 16.5 17.2 16.4 17.0 17.0 16.7 16.2 16.6
[109] 16.9 16.5 16.6 16.6 17.0 17.1 17.1 16.7 16.8 16.3 16.6 16.8 16.9 17.1 16.8 17.0 17.2 17.
3 17.2 17.3 17.2 17.2 17.5 16.9 16.9 16.9 17.0 16.5 16.7 16.8 16.7 16.7 16.6 16.5 17.0 16.7
[145] 16.7 16.9 17.4 17.1 17.0 16.8 17.2 17.2 17.4 17.2 16.9 16.8 17.0 17.4 17.2 17.2 17.1 17.
1 17.1 17.4 17.2 16.9 16.9 17.0 16.7 16.9 17.3 17.8 17.8 17.6 17.5 17.0 16.9 17.1 17.2 17.4
[181] 17.5 17.9 17.0 17.0 17.0 17.2 17.3 17.4 17.4 17.0 18.0 18.2 17.6 17.8 17.7 17.2 17.4
attr(,"source")
[1] Box & Jenkins (1976)
attr(,"description")
[1] Chemical concentration readings
```



Διάγραμμα 4.1.1
Γραφική παράσταση των δεδομένων των μετρήσεων χημικής συγκέντρωσης.

Για τον υπολογισμό της περιγραφικής στατιστικής χρησιμοποιείται το στατιστικό πρόγραμμα Minitab 14 και δίνονται τα αποτελέσματα στην Εικόνα 4.1.1. Ακολουθεί συγκεντρωτικά η στατιστική περιγραφή στον Πίνακα 4.1.2.

Descriptive Statistics: C1

Variable	Mean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum
C1	17,062	0,399	0,159	16,100	16,800	17,000	17,400	18,200

Variable	Skewness	Kurtosis
C1	0,16	-0,12

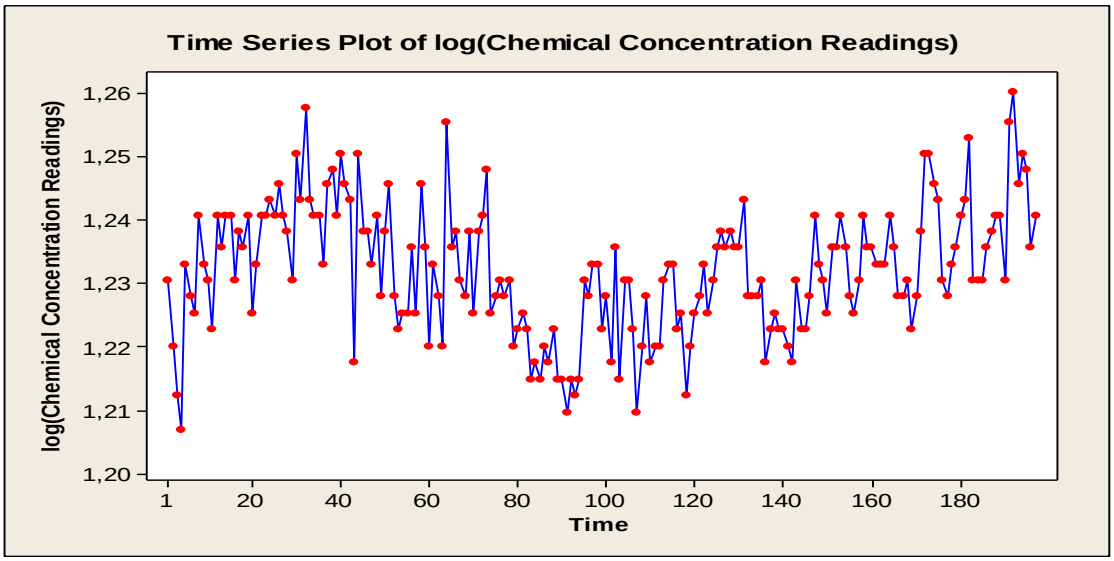
Εικόνα 4.1.1
Αποτελέσματα Minitab 14 περιγραφικής στατιστικής για τα δεδομένα των μετρήσεων χημικής συγκέντρωσης.

Πίνακας 4.1.2
Συγκεντρωτικός πίνακας στατιστικών ιδιοτήτων των δεδομένων των μετρήσεων χημικής συγκέντρωσης.

Περιγραφική στατιστική για μετρήσεις συγκέντρωσης χημικού διαλύματος	
Ελάχιστη τιμή	16,1
1ο ενδοτεταρτημοριακό εύρος Q	16,8
Διάμεσος	17
Μέση τιμή	17,062
3ο ενδοτεταρτημοριακό εύρος Q	17,4
Μέγιστη τιμή	18,2
Διακύμανση	0,159
Τυπική απόκλιση	0,399
Λοξότητα	0,16
Κύρτωση	-0,12

Τα στατιστικά στοιχεία δεν λαμβάνουν υπόψιν το χρόνο και επομένως όλες οι τιμές θεωρούνται ισοδύναμες. Έτσι, η χρονολογική σειρά αποτελείται από ένα κοινό σύνολο δεδομένων, το οποίο εύκολα περιγράφεται από την μέση τιμή και την διακύμανση. Ωστόσο, εφόσον η μέση τιμή και η διακύμανση αυξομειώνονται με την πάροδο του χρόνου, δεν έχει νόημα η περεταίρω μελέτη τους.

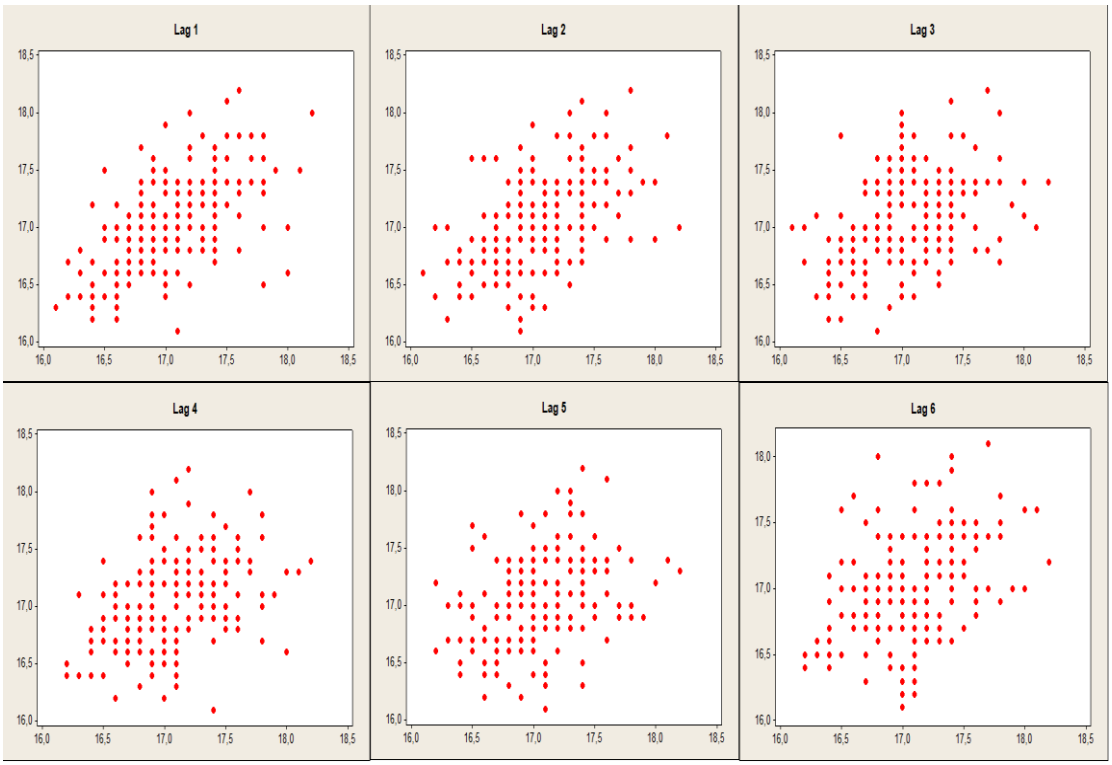
Μια συνήθης προσέγγιση που βρίσκει εφαρμογή στην οικονομετρία είναι να μετασχηματίζονται τα αρχικά ακατέργαστα δεδομένα στους αντίστοιχους λογαρίθμους τους. Με τον μετασχηματισμό αυτό, πετυχαίνεται αφενός να γίνονται πιο ξεκάθαρα τα μοτίβα που εμφανίζονται στο γράφημα και αφετέρου να μεταφέρεται όλη η πληροφορία για το ιστορικό των δεδομένων μας. Η λογαριθμημένη εκδοχή των δεδομένων φαίνεται στο Διάγραμμα 4.1.2.



Διάγραμμα 4.1.2
Γραφική παράσταση με λογαριθμημένα τα δεδομένα των μετρήσεων χημικής συγκέντρωσης.

Δεν παρατηρείται καμία διαφορά στα μοτίβα από το Διάγραμμα 4.1 και 4.2, οπότε και δεν προτιμάται η λογαριθμημένη εκδοχή στην συνέχεια.

Η δημιουργία των διαγραμμάτων διασποράς (scatter plots) βοηθάει στην αναγνώριση της συναρτησιακής μορφής της προς μελέτη μεταβλητής σε προηγούμενες χρονικές περιόδους. Κάθε γράφημα του Διαγράμματος 4.1.3 δείχνει τα δεδομένα x_t σε σχέση με τα x_{t-h} για διάφορες τιμές της υστέρησης *h*. Συγκεκριμένα, το πρώτο γράφημα δείχνει την συναρτησιακή μορφή που συνδέει κάθε παρατήρηση των δεδομένων μας με την προηγούμενη παρατήρηση. Δηλαδή, τα σημεία του πρώτου γραφήματος είναι τα $(x_2, x_1),(x_3, x_2),\ldots(x_{197}, x_{196})$.Κατά συνέπεια το δεύτερου είναι τα $(x_3, x_1)(x_4, x_2)\ldots(x_{197}, x_{195})$. Ιδιοτρόπως, κατασκευάζονται και τα υπόλοιπα γραφήματα. Στο Διάγραμμα 4.1.3 είναι ξεκάθαρη η γραμμική συσχέτιση των δεδομένων μέχρι την τρίτη υστέρηση(lag=3). Δηλαδή κάθε τιμή των μετρήσεων χημικής συγκέντρωσης είναι άμεσα και γραμμικά συνδεδεμένη με τις δύο προηγούμενες τιμές. Επιπροσθέτως, είναι εμφανές ότι η γραμμική αυτή σχέση χάνεται όσο μεγαλώνουν οι υστερήσεις. Πιο συγκεκριμένα, από την υστέρηση h=3 και μετά παρατηρείται μεγάλη εξασθένιση της γραμμικής συσχέτισης. Για αυτόν το λόγο, σε οποιαδήποτε περίπτωση προσαρμογής μοντέλου, οποιαδήποτε υστέρηση h>3 παρότι είναι στατιστικά σημαντική δεν θα συμπεριληφθεί.



Διάγραμμα 4.1.3
Διαγράμματα διασποράς των δεδομένων των μετρήσεων χημικής συγκέντρωσης σε υστερήσεις h=1,2…6.

4.1.2 Έλεγχος στασιμότητας με χρήση του επαυξημένου ελέγχου Dickey-Fuller και του ελέγχου Kwiatkowski-Phillips-Schmidt-Shin

Κατ' αρχήν, πριν γίνει κάποια επεξεργασία των αριθμητικών δεδομένών της χρονοσειράς, είναι ανάγκη να μελετηθεί η στασιμότητα της. Ένας τέτοιος έλεγχος είναι ο έλεγχος Dickey-Fuller (DF test) (Dickey & Fuller, 1997). Συγκεκριμένα, ο έλεγχος Dickey-Fuller ελέγχει αν το φ ισούται με μηδέν (φ=0) στο μοντελο των δεδομένων:

$$y_t = \alpha + \beta_t + \varphi y_{t-1} + e_t$$

το οποίο γράφεται ως:

$$\Delta y_t = y_t - y_{t-1} = \alpha + \beta_t + \gamma y_{t-1} + e_t$$

όπου y_t είναι η χρονοσειρά.

Γράφεται με αυτό το τρόπο ώστε να μπορεί να γίνει μια γραμμική παλινδρόμηση του Δy_t ως προς το χρόνο(t) και το Δy_{t-1} και να αποσαφηνιστεί αν το γ είναι διάφορο του μηδενός ($\gamma \neq 0$). Εάν το γ ισούται με το μηδέν ($\gamma = 0$) τότε η χρονοσειρά είναι μια ακολουθία τυχαίου περιπάτου. Εάν όχι, και ισχύει ότι:

$$-1 < 1 + \gamma < 1$$

τότε η χρονοσειρά είναι στάσιμη. Αντίστοιχα, χρησιμοποιείται ο επαυξημένος έλεγχος Dickey-Fuller (ADF test) ο οποίος επιτρέπει τον έλεγχο στασιμότητας σε υψηλότερης τάξης αυτοπαλίνδρομα μοντέλα περιέχοντας τον ορό Δy_{t-p}. Και σε αυτήν την περίπτωση, γίνεται εξέταση αν το γ ισούται με μηδέν ($\gamma = 0$).

$$\Delta y_t = \alpha + \beta_t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \dots$$

Η μηδενική υπόθεση και για τους δυο ελέγχους DF και ADF είναι ότι η χρονοσειρά είναι μη στάσιμη. Για να απορριφθεί η μηδενική υπόθεση πρέπει το p-value να είναι μικρότερο του 0.05 (p-value < 0.005) (Holmes, Scheuerell & Ward).

Ένας επιπλέον έλεγχος στασιμότητας είναι ο Kwiatkowski-Phillips-Schmidt-Shin (KPSS test). Ο έλεγχος KPSS έχει ως μηδενική υπόθεση ότι τα δεδομένα είναι στάσιμα γύρω από μια συνιστώσα τάσης. Συνεπώς, η απόρριψη της μηδενικής υπόθεσης σε επίπεδο σημαντικότητας 5%, υποδηλώνει ότι τα δεδομένα μας δεν παρουσιάζουν στασιμότητα γύρω από μια συνιστώσα τάσης (Kwiatkowski et al., 1992). Χρησιμοποιώντας τις εντολές της R γλώσσας adf.test και kpss.test, που βρίσκονται στη βιβλιοθήκη tseries, λαμβάνονται τα αποτελέσματα των ελέγχων ADF και KPSS στην Εικόνα 4.1.2.

```
> adf.test(tsdl[[162]])

        Augmented Dickey-Fuller Test

data:  tsdl[[162]]
Dickey-Fuller = -2.6562, Lag order = 5, p-value = 0.3014
alternative hypothesis: stationary

> kpss.test(tsdl[[162]], null = "Trend")

        KPSS Test for Trend Stationarity

data:  tsdl[[162]]
KPSS Trend = 0.48572, Truncation lag parameter = 4, p-value = 0.01

Warning message:
In kpss.test(tsdl[[162]], null = "Trend") :
  p-value smaller than printed p-value
>
```

Εικόνα 4.1.2
Αποτελέσματα ελέγχων ADF και KPSS για τα δεδομένα των μετρήσεων χημικής συγκέντρωσης.

Το p-value στην περίπτωση του ελέγχου ADF ισούται με 0,3014 οπότε και δεν απορρίπτεται η μηδενική υπόθεση ενώ στην περίπτωση του KPSS το p-value ισούται με τιμή μικρότερη του 0,01 οπότε και απορρίπτεται η μηδενική υπόθεση. Εύλογα, η χρονολογική σειρά περιέχει συνιστώσα τάσης και δεν είναι στάσιμη.

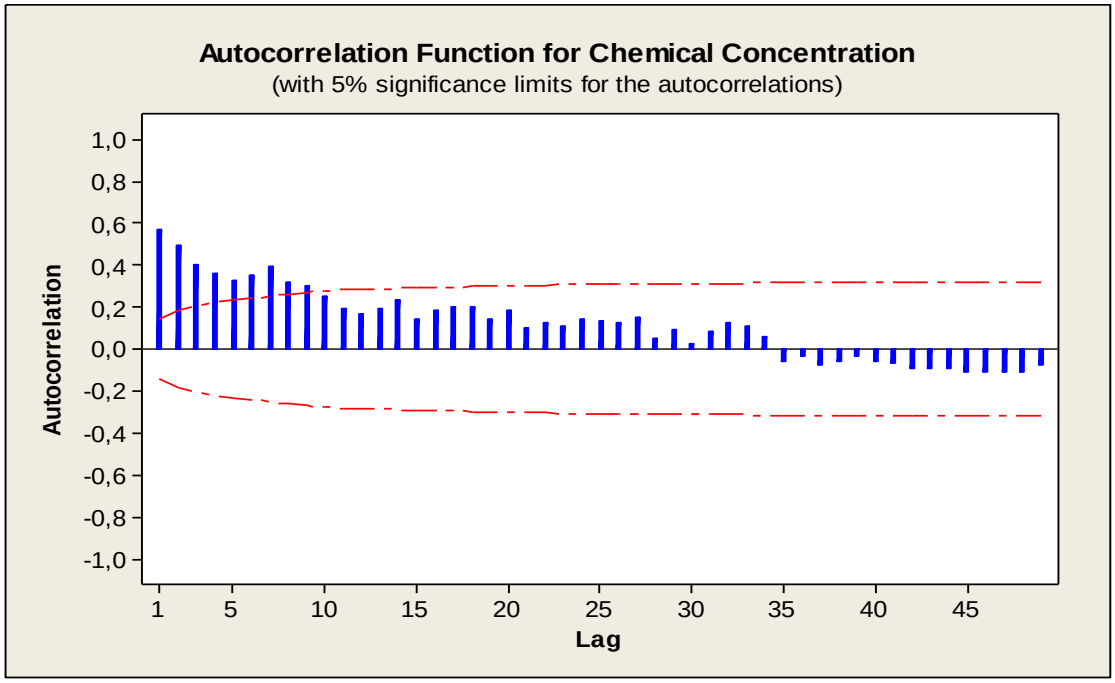
4.1.3 Υπολογισμός της συνάρτησης αυτοσυσχέτισης (ACF) και της συνάρτησης μερικής αυτοσυσχέτισης (PACF)

Ο υπολογισμός των συντελεστών αυτοσυσχέτισης, για διάφορες χρονικές υστερήσεις(lags) μιας χρονοσειράς, βοηθά στη αποτίμηση μιας χρονοσειράς ως προς την τυχαιότητα και την στασιμότητα. Περαιτέρω, σε βαθύτερη ανάλυση, μέσω των συναρτήσεων αυτοσυσχέτισης και μερικής αυτοσυσχέτισης επιλέγονται οι διάφοροι παράμετροι μοντέλων.

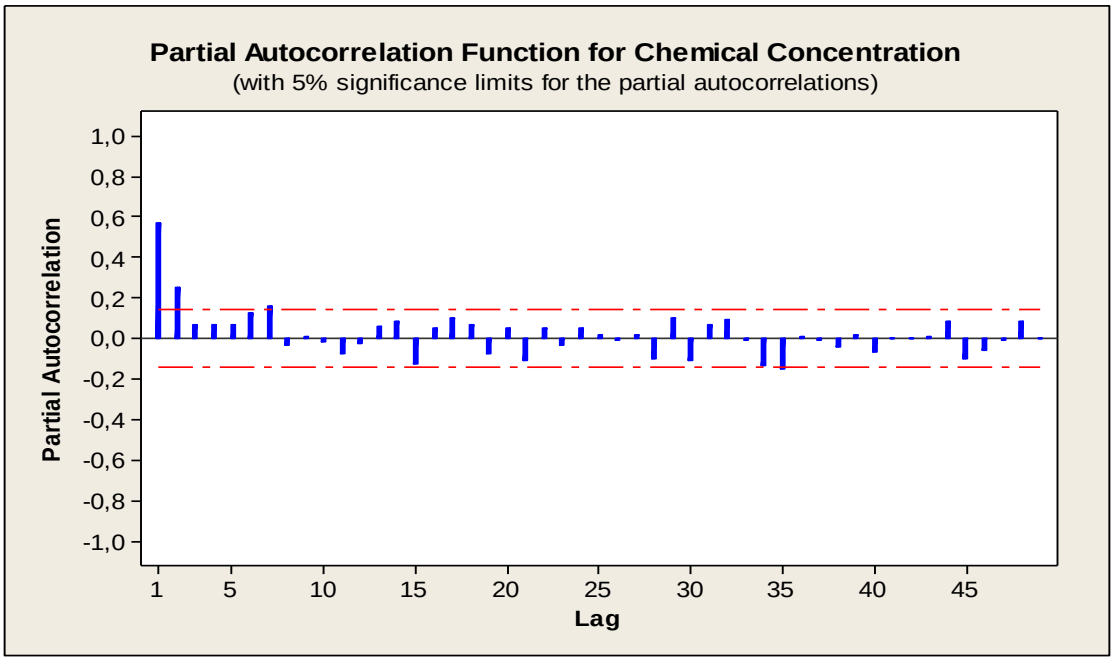
Τυχαία χρονοσειρά θεωρείται η χρονοσειρά στην οποία κάθε παρατήρηση είναι ανεξάρτητη από οποιαδήποτε άλλη παρατήρηση. Σε μία τυχαία χρονοσειρά το 95% των συντελεστών αυτοσυσχέτισης βρίσκονται στο διάστημα που ορίζεται από τις τιμές $\pm 1.96/\sqrt{n}$ όπου n είναι ο αριθμός των παρατηρήσεων. Εάν οι συντελεστές αυτοσυσχέτισης βρίσκονται εκτός των παραπάνω ορίων τότε υπάρχει συσχέτιση ανάμεσα στις παρατηρήσεις και άρα η χρονοσειρά δεν είναι τυχαία. Όσον αφορά τη στασιμότητα, για μία μη στάσιμη χρονοσειρά, οι συντελεστές αυτοσυσχέτισης είναι

διάφοροι του μηδενός για αρκετές επαναλήψεις από τις πρώτες χρονικές υστερήσεις και αργά, προσεγγίζουν το μηδέν.

Για τον υπολογισμό της συνάρτησης ACF και PACF χρησιμοποιείται το υπολογιστικό πρόγραμμα Minitab 14. Ακολουθεί το Διάγραμμα 4.1.4 και 4.1.5 με τα γραφήματα των συναρτήσεων αυτοσυσχέτισης και μερικής αυτοσυσχέτισης.



Διάγραμμα 4.1.4
Συνάρτηση αυτοσυσχέτισης (ACF) των δεδομένων των μετρήσεων χημικής συγκέντρωσης.



Διάγραμμα 4.1.5
Συνάρτηση μερικής αυτοσυσχέτισης (PACF) των δεδομένων των μετρήσεων χημικής συγκέντρωσης.

Από το διάγραμμα της συνάρτησης αυτοσυσχέτισης(Διάγραμμα 4.1.4) γίνεται οπτικά αντιληπτό ότι η χρονοσειρά είναι τυχαία και μη στάσιμη. Εάν τα δεδομένα προέρχονται από στάσιμη διαδικασία, το γράφημα της ACF έπρεπε να φθίνει εκθετικά στο μηδέν. Στο Διάγραμμα 4.1.4 φαίνεται ότι η ACF ακολουθεί φθίνουσα πορεία αλλά αυτό γίνεται μετά από πολλές επαναλήψεις και με αργό ρυθμό. Το υπολογιστικό αποτέλεσμα (έλεγχος ADF) καθώς και το γραφικό (συνάρτησης αυτοσυσχέτισης) ήταν αναμενόμενα, διότι ευκολά διαφαίνεται και από τη γραφική παράσταση (Διάγραμμα 4.1.1) ότι η χρονοσειρά δεν διατηρεί σταθερές τις στατιστικές της ιδιότητες στο πέρας του χρόνου, οπότε και δεν είναι στάσιμη.

4.1.4 Αντιμετώπιση της στασιμότητας

Η απαλοιφή της τάσης από τα δεδομένα της χρονοσειράς πετυχαίνετε με τη χρήση της μεθόδου της διαφόρισης. Δηλαδή, οι σειρές διαφορών πρώτης τάξης προκύπτουν από τη διαφορά δύο διαδοχικών παρατηρήσεων:

$$Y'_t = Y_t - Y_{t-1}$$

Οι σειρές διαφορών πρώτης τάξης έχουν n-1 δεδομένα εάν οι παρατηρήσεις της χρονοσειράς είναι n. Εάν γίνει έλεγχος και συνεχίζει η νέα χρονοσειρά να μην είναι στάσιμη, προτείνεται η διαφόριση των δεδομένων για δεύτερη φορά.

$$Y''_t = Y'_t - Y'_{t-1} = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$$

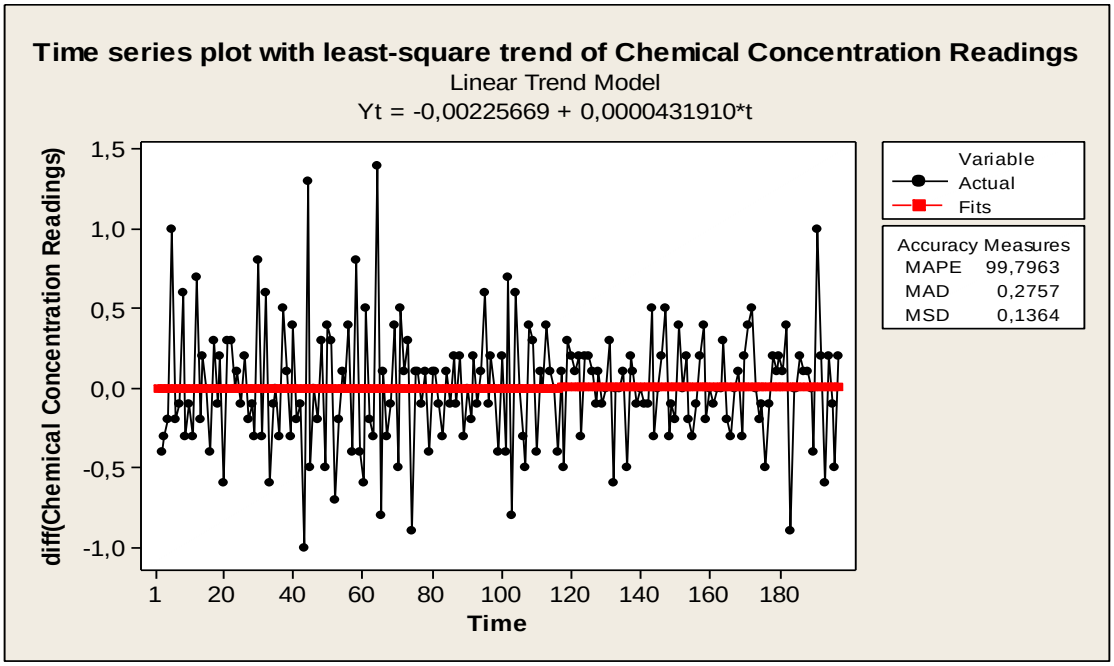
Με αυτόν τον τρόπο προκύπτει η σειρά διαφορών δεύτερης τάξης. Οι σειρές διαφορών δεύτερης τάξης περιέχουν n-2 δεδομένα.

Χρησιμοποιώντας, μέσω της γλώσσας R, την εντολή ndiffs() από το πακέτο forecast δίνεται η δυνατότητα να βρεθεί η τάξη των διαφορών που χρειάζεται ώστε η χρονοσειρά να «περνάει» τους ελέγχους στασιμότητας ADF και KPSS. Τα αποτελέσματα βρίσκονται στην Εικόνα 4.1.3.

```
> ndiffs(tsdl[[162]], test = "adf")
[1] 0
> ndiffs(tsdl[[162]], test = "kpss")
[1] 1
```

Εικόνα 4.1.3
Αποτελέσματα για το βαθμό διαφόρισης των δεδομένων των μετρήσεων χημικής συγκέντρωσης.

Τα αποτελέσματα υποδεικνύουν ότι η χρονολογική σειρά για να «περνάει» τον έλεγχο ADF δεν χρειάζεται διαφόριση, το οποίο είναι άτοπο. Όσον αφορά τον έλεγχο KPSS απαιτείται να γίνει διαφόριση μια φορά. Για να διευκρινιστεί, γίνεται η διαφόριση των δεδομένων και παρατίθεται η ολοκληρωμένη I(1) χρονοσειρά μαζί με την ευθεία ελάχιστων τετραγώνων στο Διάγραμμα 4.1.6.



Διάγραμμα 4.1.6
Γράφημα μετασχηματισμένων με πρώτες διαφορές δεδομένων των μετρήσεων χημικής συγκέντρωσης μαζί με την ευθεία ελαχίστων τετραγώνων.

Παρατηρώντας το Διάγραμμα 4.1.6 δημιουργούνται εμφανείς υπόνοιες για στασιμότητα ως προς τη μέση τιμή και για επιβεβαίωση πραγματοποιούνται εκ νέου οι έλεγχοι στασιμότητας ADF και KPSS. Τα αποτελέσματα των ελέγχων φαίνονται στην Εικόνα 4.1.4.

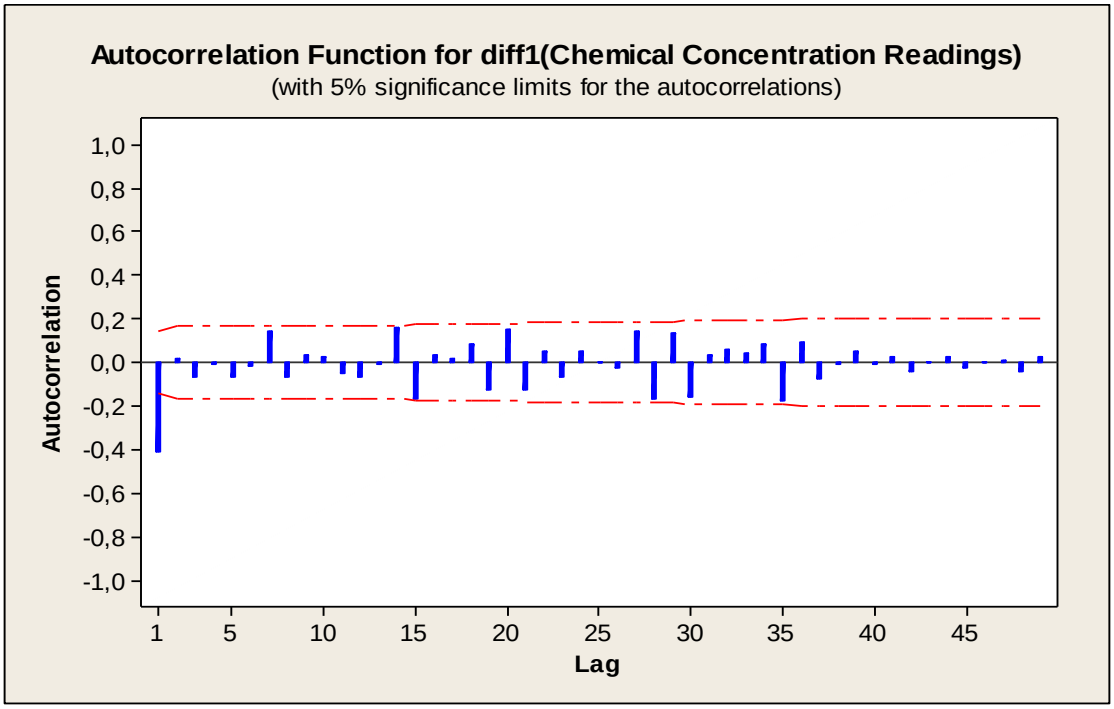
```
> adf.test(diff(tsdl[[162]]))

        Augmented Dickey-Fuller Test

data:  diff(tsdl[[162]])
Dickey-Fuller = -9.9271, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(diff(tsdl[[162]])) : p-value smaller than printed p-value
> kpss.test(diff(tsdl[[162]]))

        KPSS Test for Level Stationarity

data:  diff(tsdl[[162]])
KPSS Level = 0.02313, Truncation lag parameter = 4, p-value = 0.1

Warning message:
In kpss.test(diff(tsdl[[162]])) : p-value greater than printed p-value
>
```

Εικόνα 4.1.4
Αποτελέσματα ελέγχου ADF και KPSS για τα μετασχηματισμένα με πρώτες διαφορές δεδομένα των μετρήσεων χημικής συγκέντρωσης.

Η μικρή p-value του ADF test 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, οδηγεί στο συμπέρασμα ότι δεν μπορεί να απορριφθεί η μηδενική υπόθεση περί στασιμότητας των δεδομένων γύρω από μια συνιστώσα τάσης, πάλι σε επίπεδο σημαντικότητας 5%. Συμπερασματικά, οι δυο αυτοί έλεγχοι καταλήγουν ότι τα δεδομένα που έχουν προκύψει από τις πρώτες διαφορές είναι πλέον στάσιμα. Άξιο αναφοράς είναι 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. Κατά συνέπεια, δεν περιλαμβάνεται περιοδικότητα στο μοντέλο.

4.1.5 Επιλογή παραμέτρων για μοντελοποίηση

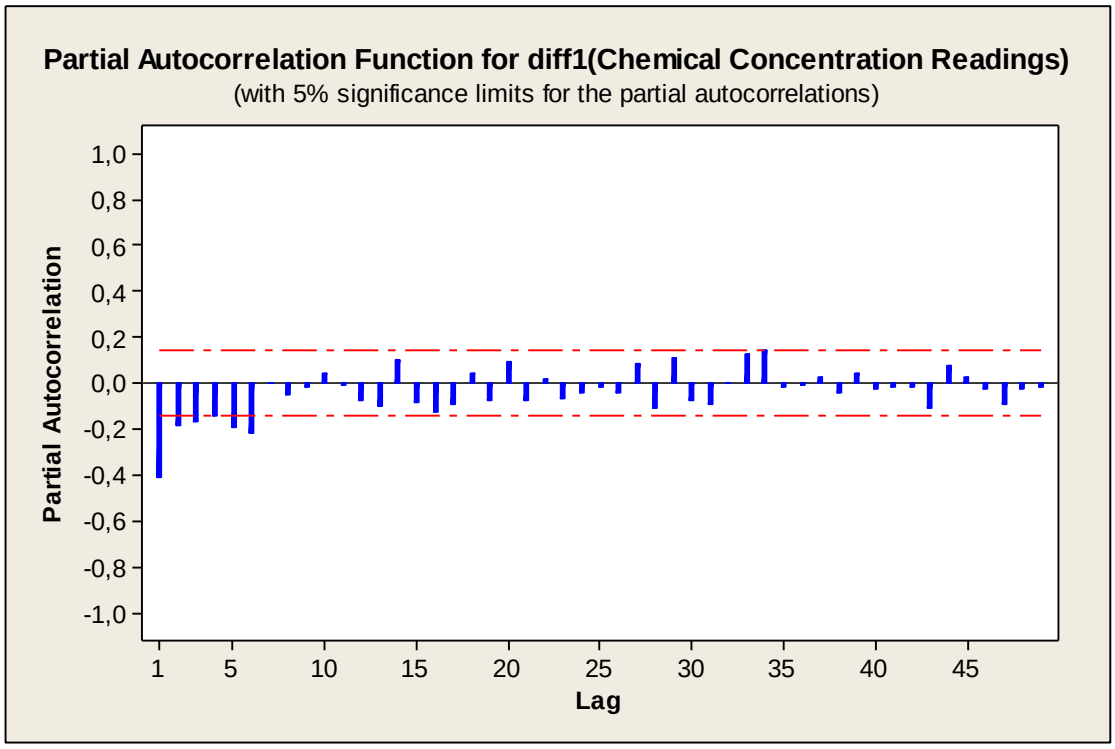
Το επόμενο βήμα περιλαμβάνει μια πρώτη εκτίμηση για τις τάξεις των AR(p) και MA(q) για το μοντέλο. Έχει γίνει αναφορά σε προηγούμενες ενότητες (Πίνακας 3.1) ότι αυτό γίνεται με τη βοήθεια των γραφημάτων των συναρτήσεων ACF και PACF. Για την εκτίμηση των συντελεστών κατασκευάζονται τα γραφήματα της ACF και της PACF που απεικονίζονται κατά αντιστοιχία στο Διάγραμμα 4.1.7 και 4.1.8.



Διάγραμμα 4.1.7
Συνάρτηση αυτοσυσχέτισης (ACF) των μετασχηματισμένων με πρώτες διαφορές δεδομένων των μετρήσεων χημικής συγκέντρωσης.

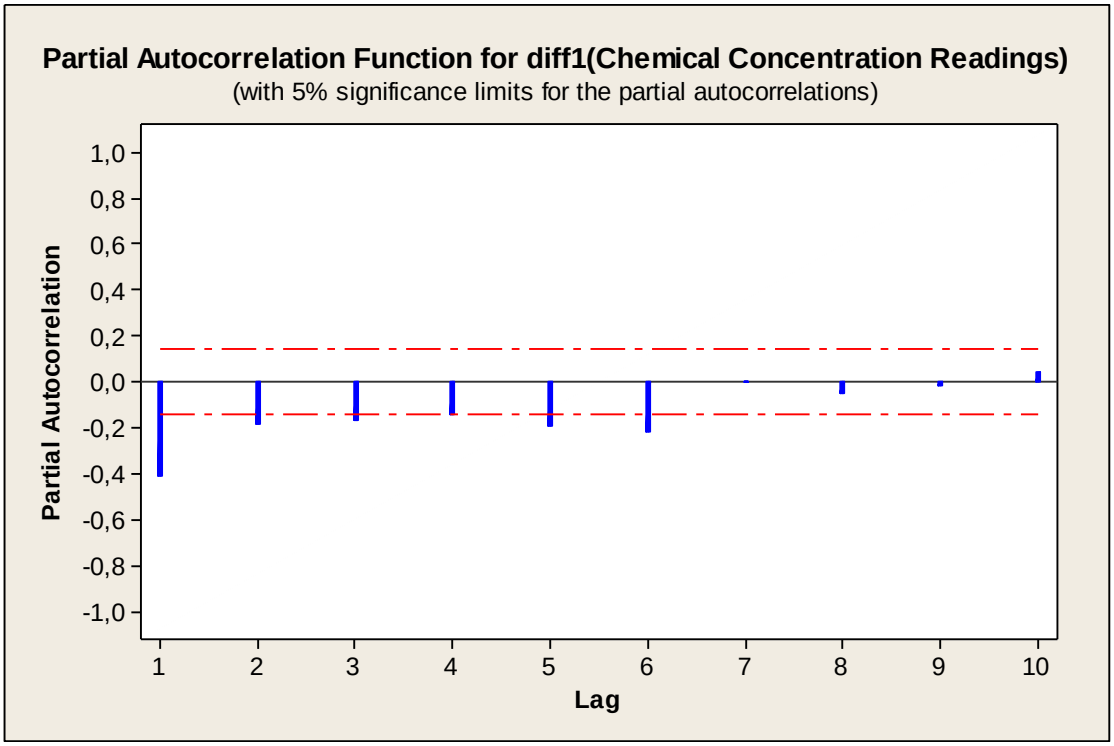
,

Το Διάγραμμα 4.1.7 δείχνει μόνο τον πρώτο όρο της ACF να πέφτει εκτός των ορίων ενώ οι υπόλοιποι όροι να βρίσκονται εντός ορίων με 5% επίπεδο σημαντικότητας. Με βάση αυτές τις παρατηρήσεις και τον Πίνακα 3.1 η πιο πιθανή τιμή στην τάξη της MA διαδικασίας είναι η πρώτη. Δηλαδή, επιλέγεται η τιμή q = 1.



Διάγραμμα 4.1.8
Συνάρτηση μερικής αυτοσυσχέτισης (PACF) των μετασχηματισμένων με πρώτες διαφορές δεδομένων των μετρήσεων χημικής συγκέντρωσης.

Στο Διάγραμμα 4.1.8 δεν διαφαίνεται καθαρά η συμπεριφορά της συνάρτησης μερικής αυτοσυσχέτισης (PACF) στις αρχικές υστερήσεις. Πιθανώς να ακολουθεί κάποιο εκθετικό πρότυπο συνάρτησης ή διαφορετικά κάποιο πρότυπο φθίνουσας ημιτονοειδούς συνάρτησης. Αυτό συμβαίνει έως περίπου την όγδοη χρονική υστέρηση (h = 8) ενώ στην συνέχεια οι υπόλοιπες τιμές της PACF βρίσκεται εντός ορίων για επίπεδο σημαντικότητας 5%. Οπότε δημιουργείται για χάρη ευκολίας το Διάγραμμα 4.1.6 όπου φαίνεται η συνάρτηση PACF για τις πρώτες δέκα υστερήσεις.



Διάγραμμα 4.1.9
Συνάρτησης μερικής αυτοσυσχέτισης (PACF) των μετασχηματισμένων με πρώτες διαφορές δεδομένων των μετρήσεων χημικής συγκέντρωσης για 10 χρονικές υστερήσεις.

Είναι άξιο παρατήρησης, λοιπόν, ότι ο πρώτος, ο δεύτερος και ο εκτός όρος βρίσκονται εκτός ορίων και στη συνέχεια ελάχιστα ο τρίτος και ο πέμπτος όρος. Ωστόσο, εξαιτίας του γεγονότος ότι για κάθε υστέρηση χάνεται η γραμμική εξάρτηση των δεδομένων, δε θα θεωρηθεί στατιστικά σημαντικός ο έκτος και ο πέμπτος όρος (h>3). Πιο πιθανές τιμές είναι p=1 και έπειτα p=2 και p=3.

4.1.6 Επιλογή μοντέλου ARIMA

Έχοντας επιλέξει τις παραμέτρους, τα μοντέλα ARIMA που πιθανώς να περιγράφουν καλυτέρα τα δεδομένα μετρήσεων χημικής συγκέντρωσης είναι το μοντέλο ARIMA(1,1,1), το μοντελο ARIMA(2,1,1) και το μοντέλο ARIMA (3,1,1). Προσαρμόζοντας τα δεδομένα και με τη βοήθεια του Minitab 14 παρουσιάζονται τα μοντέλα στην Εικόνα 4.1.5.,4.1.6 και 4.1.7.

```
Final Estimates of Parameters

Type      Coef  SE Coef      T      P
AR   1  0,2187   0,0940   2,33  0,021
MA   1  0,8253   0,0539  15,31  0,000


Differencing: 1 regular difference
Number of observations:  Original series 197, after differencing 196
Residuals:    SS =  19,2792 (backforecasts excluded)
              MS =  0,0994  DF = 194


Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   11,3   25,1   46,7   51,1
DF             10     22     34     46
P-Value     0,335  0,294  0,071  0,281
```

Εικόνα 4.1.5
Αποτελέσματα μοντέλου ARIMA(1,1,1).

Βλέποντας τα p-value, όλοι οι οροί του μοντέλου είναι στατιστικά σημαντικοί και τα τυπικά σφάλματα είναι μικρά.

```
Final Estimates of Parameters

Type       Coef  SE Coef         T      P
AR   1  -1,3923   0,0657    -21,19  0,000
AR   2  -0,4002   0,0657     -6,09  0,000
MA   1  -0,9844   0,0002  -4120,24  0,000


Differencing: 1 regular difference
Number of observations:  Original series 197, after differencing 196
Residuals:    SS =  22,1315 (backforecasts excluded)
              MS =  0,1147  DF = 193


Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   25,5   41,0   64,3   67,8
DF              9     21     33     45
P-Value     0,002  0,006  0,001  0,016
```

Εικόνα 4.1.6
Αποτελέσματα μοντέλου ARIMA(2,1,1).

Επίσης, όλοι οι οροί του μοντέλου είναι στατιστικά σημαντικοί και τα τυπικά σφάλματα λαμβάνουν μικρές τιμές.

```
Final Estimates of Parameters

Type       Coef  SE Coef      T      P
AR   1   0,2334   0,0982   2,38  0,018
AR   2   0,0930   0,0865   1,08  0,284
AR   3  -0,0513   0,0828  -0,62  0,537
MA   1   0,8489   0,0682  12,44  0,000


Differencing: 1 regular difference
Number of observations:  Original series 197, after differencing 196
Residuals:    SS =  19,1329 (backforecasts excluded)
              MS =  0,0997  DF = 192


Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   10,1   22,9   42,6   46,5
DF              8     20     32     44
P-Value     0,256  0,292  0,100  0,369
```

Εικόνα 4.1.7
Αποτελέσματα μοντέλου ARIMA(3,1,1).

Παρατηρείται ότι οι όροι AR(2) και AR(3) του μοντέλου δεν είναι στατιστικά σημαντικοί. Τα τυπικά σφάλματα παραμένουν μικρά. Οπότε, το μοντελο ARIMA(3,1,1) απορρίπτεται με ευκολία. Για την επιλογή αναμεσά στα αλλά δύο μοντέλα χρησιμοποιείται η γλώσσα R που παρέχει τη δυνατότητα υπολογισμού του κριτήριου AIC, AICc και BIC. Η γλώσσα R δίνει τα αποτελέσματα για το AIC, το AICc και το BIC των μοντέλων ARIMA(1,1,1) και ARIMA(2,1,1) στην Εικόνα 4.1.8.

```
> Arima(tsdl[[162]],order=c(1,1,1))
Series: tsdl[[162]]
ARIMA(1,1,1)

Coefficients:
          ar1      ma1
      0.2155  -0.8193
s.e.  0.1011   0.0631

sigma^2 estimated as 0.09952:  log likelihood=-51.37
AIC=108.74   AICc=108.87   BIC=118.58
> Arima(tsdl[[162]],order=c(2,1,1))
Series: tsdl[[162]]
ARIMA(2,1,1)

Coefficients:
          ar1     ar2      ma1
      0.2469  0.0946  -0.8639
s.e.  0.0978  0.0881   0.0636

sigma^2 estimated as 0.09944:  log likelihood=-50.8
AIC=109.6   AICc=109.81   BIC=122.71
```

Εικόνα 4.1.8,
Υπολογισμός κριτήριου AIC, AICc και BIC για ARIMA(1,1,1) και ARIMA(2,1,1).

Σύμφωνα με το κριτήριο AIC, AICc και το κριτήριο BIC διακρίνεται το μοντελο ARIMA(1,1,1) που έχει τις μικρότερες τιμές (AIC=108,74, AICc=109,81 και BIC=118,58) και είναι και το απλούστερο. Επιπροσθέτως, το πακέτο forecast της R διαθέτει εντολή(auto.arima) για την βέλτιστη επιλογή μοντέλου ARIMA σύμφωνα με το κριτήριο AICc . Τα αποτελέσματα φαίνονται στην Εικόνα 4.1.9 και συμφωνούν απολυτά με την επιλογή μας.

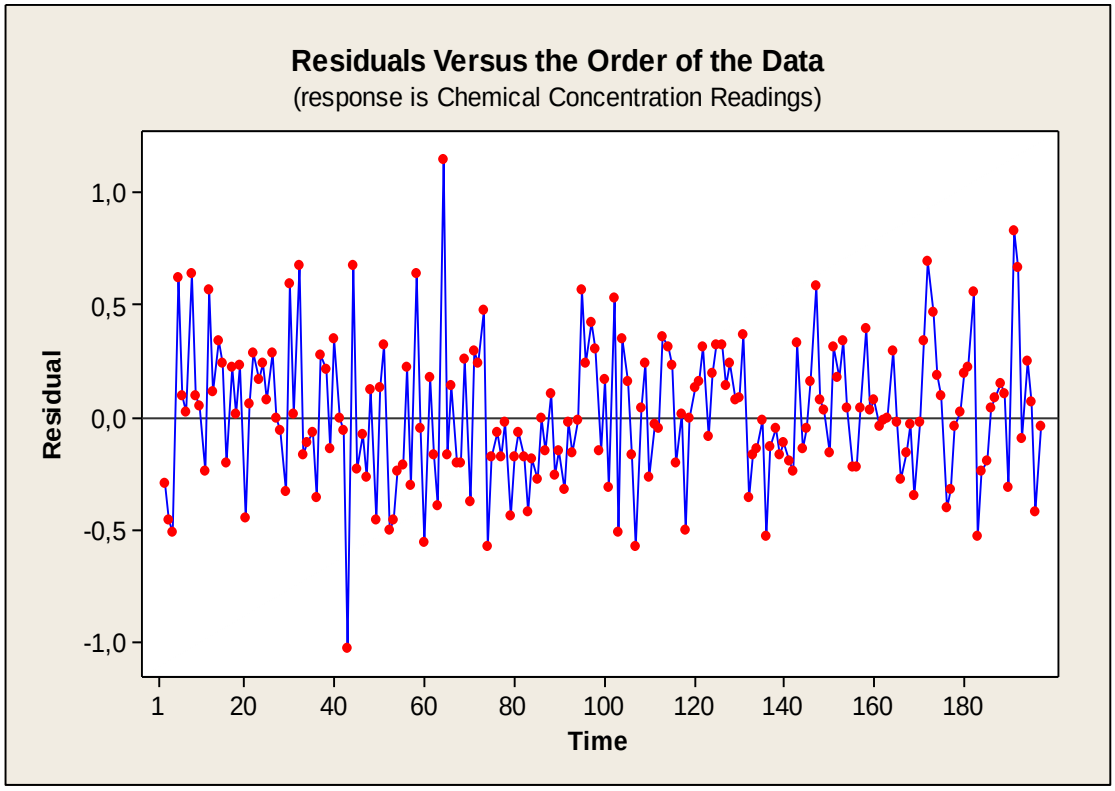
```
> auto.arima(tsdl[[162]],trace=FALSE,stepwise =FALSE,approximation=FALSE)
Series: tsdl[[162]]
ARIMA(1,1,1)

Coefficients:
         ar1      ma1
      0.2155  -0.8193
s.e.  0.1011   0.0631

sigma^2 estimated as 0.09952:  log likelihood=-51.37
AIC=108.74   AICc=108.87   BIC=118.58
```

Εικόνα 4.1.9

Αποτελέσματα auto.arima για τη χρονοσειρά με τις μετρήσεις χημικής συγκέντρωσης.

4.1.7 Διαγνωστικοί έλεγχοι υπολοίπων

Στο τελευταίο βήμα διαπιστώνεται εάν τα υπόλοιπα του μοντέλου ικανοποιούν τους διαγνωστικούς ελέγχους. Τα υπόλοιπα δείχνουν πόσο απέχει η κάθε εκτίμηση έπειτα από την προσαρμογή του εκάστοτε μοντέλου ARIMA σε σχέση με τις πραγματικές τιμές. Στην περίπτωση μας το μοντελο που μελετάται είναι το ARIMA(1,1,1). Σύμφωνα με τη θεωρία, τα υπόλοιπα πρέπει να έχουν ιδιότητες λευκού θορύβου. Μια πιο ιδανική εκδοχή είναι να είναι εκτός από ανεξάρτητα και ισόνομα, και κανονικά κατανεμημένα. Με τη βοήθεια του Minitab 14 κατασκευάζεται το Διάγραμμα 4.1.10 που απεικονίζει τα υπόλοιπα του μοντέλου ARIMA(1,1,1) στο χρόνο και η Εικόνα 4.1.10 που δείχνει την περιγραφική στατιστική των υπολοίπων.



Διάγραμμα 4.1.10
Υπόλοιπα μοντέλου ARIMA(1,1,1).

Descriptive Statistics: RESI1

Variable	Mean	Variance	Minimum	Q1	Q3	Maximum
RESI1	0,0144	0,0987	-1,0261	-0,1919	0,2301	1,1410

Εικόνα 4.1.10
Περιγραφική στατιστική των υπολοίπων του μοντέλου ARIMA(1,1,1).

Η μέση τιμή ισούται με 0,014 και η διασπορά με 0,098. Για να αποτελούν τα σφάλματα σειρά λευκού θορύβου πρέπει η μέση τιμή να είναι μηδενική, που θεωρείται στην περίπτωσή μας, και η τιμή της διασπορά να διαμένει σταθερή στο χρόνο. Εννοώντας, ότι για να είναι η διασπορά σταθερή στο χρόνο πρέπει τα σφάλματα να είναι ομοσκεδαστικά. Γραφικά λοιπόν είναι δύσκολο να ελεγχθεί αν υπάρχει ετεροσκεδαστικότητα η ομοσκεδαστικότητα στα σφάλματα. Ο έλεγχος ετεροσκεδαστικότητας γίνεται με την εντολή arch.test της βιβλιοθήκης (library) aTSA της R και παρακάτω στην Εικόνα 4.1.11 φαίνονται τα αποτελέσματα για το μοντελο μας.

```
> arch.test(fit)
ARCH heteroscedasticity test for residuals
alternative: heteroscedastic

Portmanteau-Q test:
     order    PQ p.value
[1,]     4  4.56   0.335
[2,]     8  7.44   0.491
[3,]    12 11.22   0.510
[4,]    16 14.03   0.597
[5,]    20 16.90   0.660
[6,]    24 28.64   0.234
Lagrange-Multiplier test:
     order    LM  p.value
[1,]     4 74.11 5.55e-16
[2,]     8 32.72 2.99e-05
[3,]    12 18.12 7.89e-02
[4,]    16 12.99 6.03e-01
[5,]    20  9.12 9.71e-01
[6,]    24  5.93 1.00e+00
```

Εικόνα 4.1.11
Αποτελέσματα ελέγχου ομοσκεδαστικότητας υπολοίπων για το μοντελο ARIMA(1,1,1).

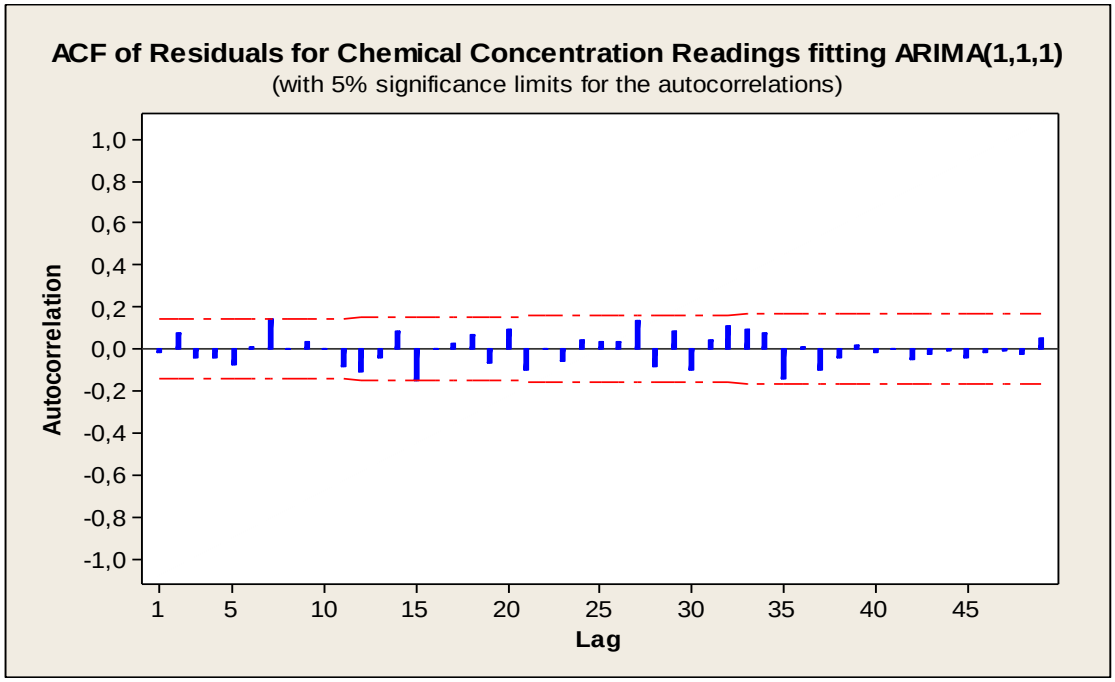
Ο πρώτος τύπος ελέγχου εξετάζει αν τα τετράγωνα των υπολοίπων είναι μια ακολουθία λευκού θορύβου, που ονομάζεται έλεγχος Portmanteau Q. Ο δεύτερος τύπος ελέγχου που προτείνεται από τον Engle (1982) είναι η δοκιμή Lagrange Multiplier, η οποία ταιριάζει σε ένα μοντέλο γραμμικής παλινδρόμησης τα τετράγωνα των υπολοίπων και εξετάζει εάν το προσαρμοσμένο μοντέλο είναι στατιστικά σημαντικό. Επομένως, η μηδενική υπόθεση είναι ότι τα τετράγωνα των σφαλμάτων είναι μια ακολουθία λευκού θορύβου, δηλαδή, τα υπόλοιπα είναι ομοσκεδαστικά. Από τις τιμές των p-value δεν απορρίπτεται η υπόθεση οπότε τα υπόλοιπα αποτελούν μια σειρά λευκού θορύβου.

Προκειμένου να γίνει έλεγχος κατά πόσο τα υπόλοιπα του μοντέλου ARIMA(1,1,1) είναι ισόνομα και ανεξάρτητα, χρησιμοποιείται υπολογιστικά με την R ο έλεγχος Box-Pierce (Box.test) και ο έλεγχος Ljung-Box μέσω της εντολής checkresiduals. Οι παραπάνω εντολές βρίσκονται στην βιβλιοθήκη (library) forecast. Στην Εικόνα 4.1.12 φαίνονται τα αποτελέσματα των ελέγχων Box-Pierce και Ljung-Box.

```
> res <- resid(fit)
> Box.test(res, lag = 12, fitdf = 2)

        Box-Pierce test

data:  res
X-squared = 10.927, df = 10, p-value = 0.3632

> checkresiduals(fit)

        Ljung-Box test

data:  Residuals from ARIMA(1,1,1)
Q* = 7.6399, df = 8, p-value = 0.4694

Model df: 2.   Total lags used: 10
```

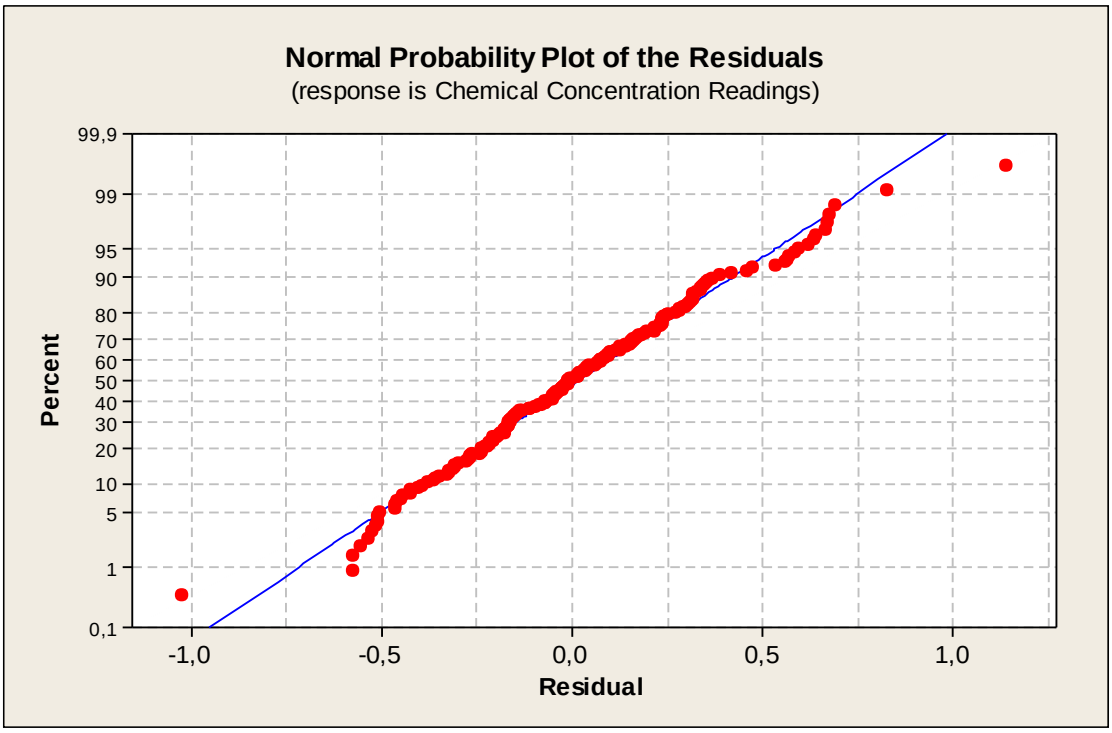
Εικόνα 4.1.12
Αποτελέσματα ελέγχων Box-Pierce και Ljung-Box για το μοντελο ARIMA(1,1,1).

Για τον υπολογισμό των στατιστικών, η μηδενική υπόθεση είναι ότι τα υπόλοιπα είναι ανεξάρτητα και ισόνομα. Από την Εικόνα 4.1.12, όλες οι τιμές p-value>0.05 και συνεπώς δεν είναι αποδεκτή η απόρριψη της μηδενικής υπόθεσης σε επίπεδο σημαντικότητας 5%. (Box & Pierce, 1970). Επιπροσθέτως, με τη βοήθεια του γραφήματος των δειγματικών ACF των υπολοίπων γίνεται έλεγχος της ανεξαρτησίας των υπολοίπων και απορρίπτεται η μηδενική υπόθεση της ανεξαρτησίας, εάν περισσότερες από δυο (40(lags)×0,05(significance level)=2) αυτοσυσχετίσεις πέσουν εκτός ορίων. Ακολουθεί το Διάγραμμα 4.1.11 με την συνάρτηση αυτοσυσχέτισης ACF των υπολοίπων.

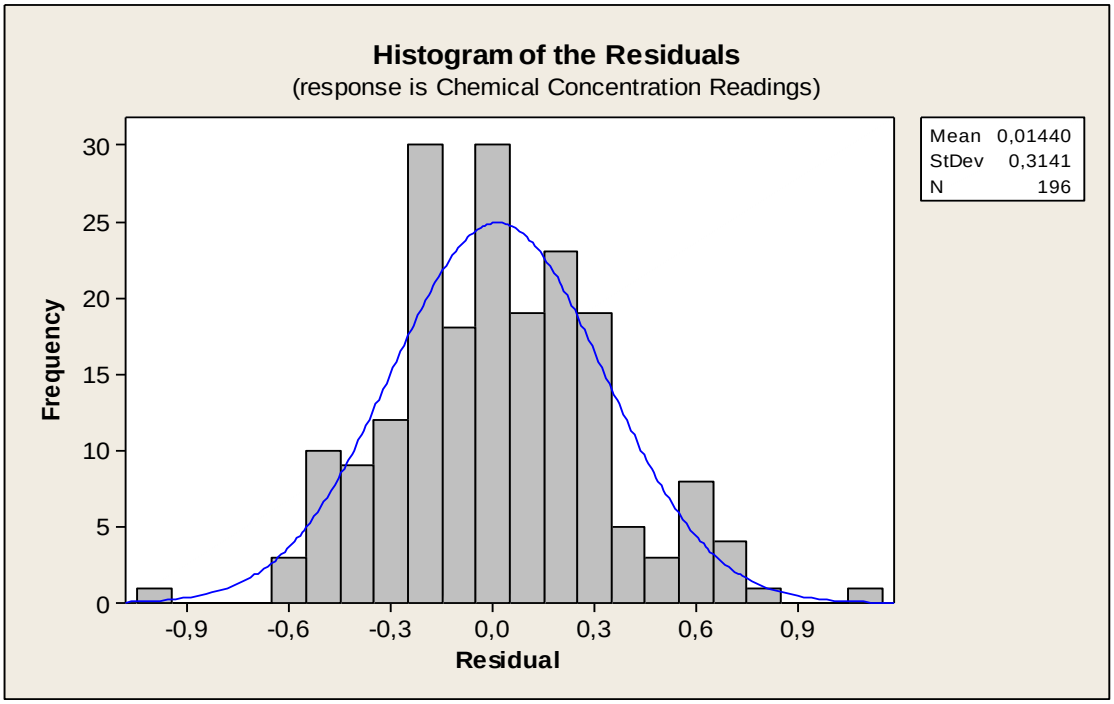


Διάγραμμα 4.1.11
Συνάρτηση αυτοσυσχέτισης (ACF) των υπολοίπων του μοντέλου ARIMA(1,1,1).

Κανένας όρος δεν πέφτει εκτός ορίων οπότε εύλογα τα υπόλοιπα είναι ανεξάρτητα. Επιπλέον, γίνεται έλεγχος εάν τα σφάλματα ακολουθούν την κανονική κατανομή. Γραφικά, κατασκευάζονται με το Minitab 14 το ιστόγραμμα των υπολοίπων και το Q-Q plot στο Διάγραμμα 4.1.12 και 4.1.13.



Διάγραμμα 4.1.12
Q-Q plot υπολοίπων του μοντελο ARIMA(1,1,1).



Διάγραμμα 4.1.13
Ιστόγραμμα υπολοίπων μαζί με τη καμπύλη κανονικής κατανομής για το μοντελο ARIMA(1,1,1).

Το γραφικό αποτέλεσμα δείχνει ότι τα υπόλοιπα μάλλον δεν προέρχονται από την κανονική κατανομή αλλά αυτό επιβεβαιώνεται και με τους ελέγχους Jarque–Bera και Anderson-Darling που τα εμφανίζουν τα αποτελέσματα της Εικόνας 4.1.13.

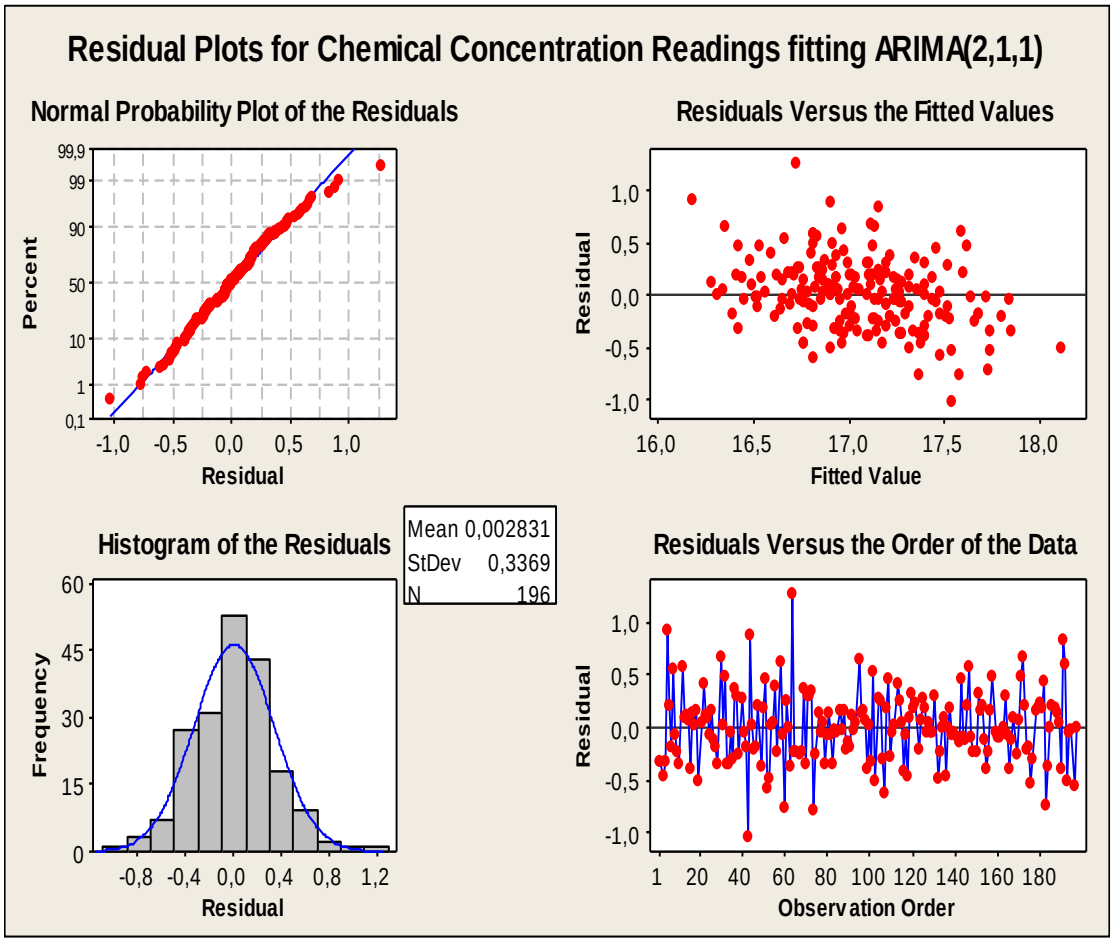
```
> fit<-arima(tsdl[[162]],order=c(1,1,1))
> res <- resid(fit)
> jarque.bera.test(res)

        Jarque Bera Test

data:  res
X-squared = 6.0729, df = 2, p-value = 0.048

> ad.test(res)

        Anderson-Darling test of goodness-of-fit
        Null hypothesis: uniform distribution
        Parameters assumed to be fixed

data:  res
An = Inf, p-value = 3.046e-06
```

Εικόνα 4.1.13
Αποτελέσματα ελέγχων κανονικότητας υπολοίπων Jarque–Bera και Anderson-Darling για το μοντέλο ARIMA(1,1,1).

Ο έλεγχος Anderson-Darling (Anderson & Darling, 1954) πραγματοποιεί στατιστικούς ελέγχους και έχει ως μηδενική υπόθεση ότι τα υπόλοιπα προέρχονται από την κανονική κατανομή. Επιπροσθέτως, 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 ισούται με τρία ενώ η λοξότητα ισούται με μηδέν. Και από τα δύο ελέγχους απορρίπτεται η μηδενική υπόθεση και επιβεβαιώνεται ότι τα υπόλοιπα δεν ακολουθούν κανονική κατανομή.

Εάν δεν θεωρούνταν η μέση τιμή(mean) των υπολοίπων του μοντέλου ARIMA(1,1,1) μηδενική, καθώς ισούται με 0,01 , στις προβλέψεις του μοντέλου θα έπρεπε να προστίθεται η μέση τιμή για να μην δημιουργείται πρόβλημα. Το αμέσως καλύτερο μοντελο, σύμφωνα με το AIC και την προηγουμένη ενδελεχή μελέτη, είναι το ARIMA(2,1,1). Επιπλέον η διαφορά τους στο κριτήριο είναι μικρότερή του δυο που θεωρείται βιβλιογραφικά αμελητέα. Ξεκινώντας, κατασκευάζονται περιληπτικά στο Διάγραμμα 4.1.14 το γράφημα των υπολοίπων στο χρόνο, το Q-Q plot, το ιστόγραμμα

των υπολοίπων και το γράφημα των υπολοίπων σε σχέση με τις προσαρμοσμένες τιμές για το μοντελο ARIMA(2,1,1).

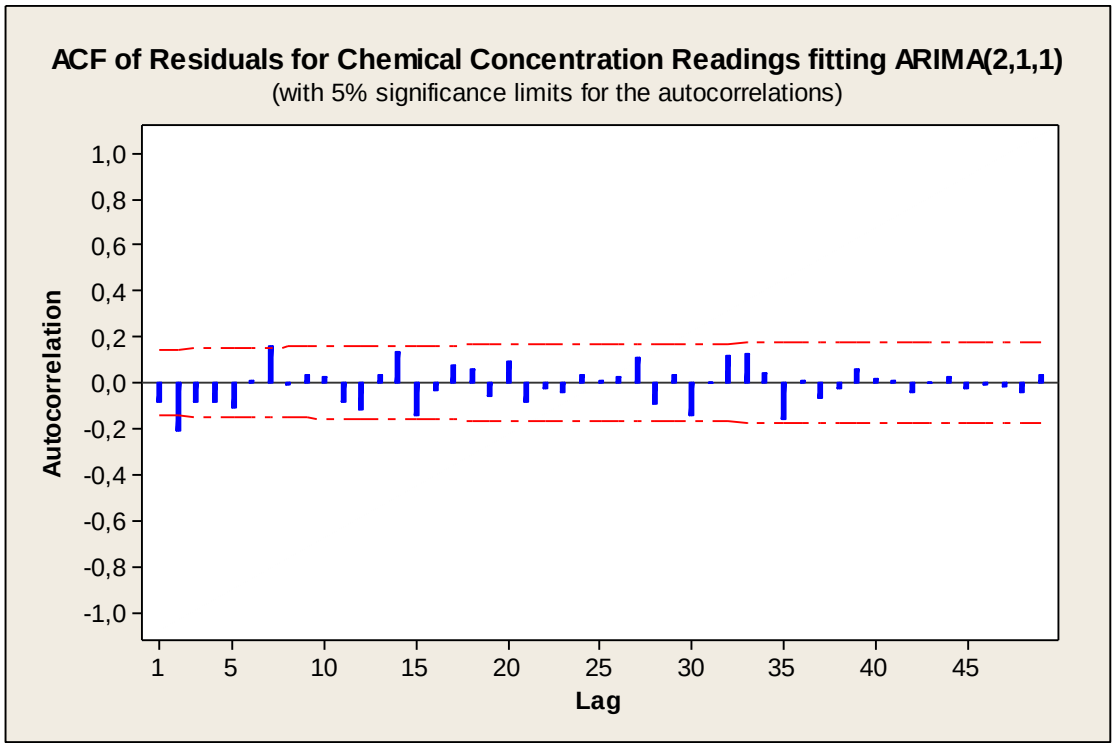


Διάγραμμα 4.1.14
Γράφημα των υπολοίπων στο χρόνο, Q-Q plot, ιστόγραμμα των υπολοίπων και το γράφημα των υπολοίπων σε σχέση με τις προσαρμοσμένες τιμές για το μοντελο ARIMA(2,1,1).

Η μέση τιμή για τα υπόλοιπα ισούται με 0,002 οπότε και θεωρείται ευκολότερά μηδενική. Από το γράφημα των υπολοίπων σε σχέση με τις προσαρμοσμένες τιμές παρατηρείται ομοσκεδαστικότητα που επιβεβαιώνεται με το arch.test. που παρουσιάζεται στην Εικόνα 4.1.14. . Παρατηρώντας την συνάρτηση αυτοσυσχέτισης (ACF) των υπολοίπων, που φαίνεται στο Διάγραμμα 4.1.15, τα υπόλοιπα είναι ανεξάρτητα καθώς μόνο δύο οροί βρίσκονται εκτός ορίων. Επιπλέον, από το ιστόγραμμα των υπολοίπων και το Q-Q plot διαφαίνεται μια «καλύτερη» εικόνα κανονικότητας σε σχέση με τα υπόλοιπα του μοντέλου ARIMA(1,1,1). Γίνεται διάγνωση κανονικότητας με τους ελέγχους Jarque–Bera και Anderson-Darling και παρατίθενται τα αποτελέσματα στην Εικόνα 4.1.15.

```
> fit <- arima(tsdl[[162]],order=c(2,1,1))
> arch.test(fit)
ARCH heteroscedasticity test for residuals
alternative: heteroscedastic

Portmanteau-Q test:
     order    PQ p.value
[1,]     4  5.61   0.230
[2,]     8  7.53   0.481
[3,]    12 12.40   0.414
[4,]    16 14.29   0.577
[5,]    20 17.88   0.596
[6,]    24 29.94   0.187
Lagrange-Multiplier test:
     order    LM  p.value
[1,]     4 72.22 1.44e-15
[2,]     8 32.35 3.50e-05
[3,]    12 17.47 9.48e-02
[4,]    16 12.68 6.27e-01
[5,]    20  8.92 9.75e-01
[6,]    24  5.73 1.00e+00
```

Εικόνα 4.1.14
Αποτελέσματα ελέγχων ομοσκεδαστικότητας υπολοίπων για το μοντέλο ARIMA(2,1,1).



Διάγραμμα 4.1.15
Συνάρτηση αυτοσυσχέτισης (ACF) των υπολοίπων του μοντέλου ARIMA(2,1,1).

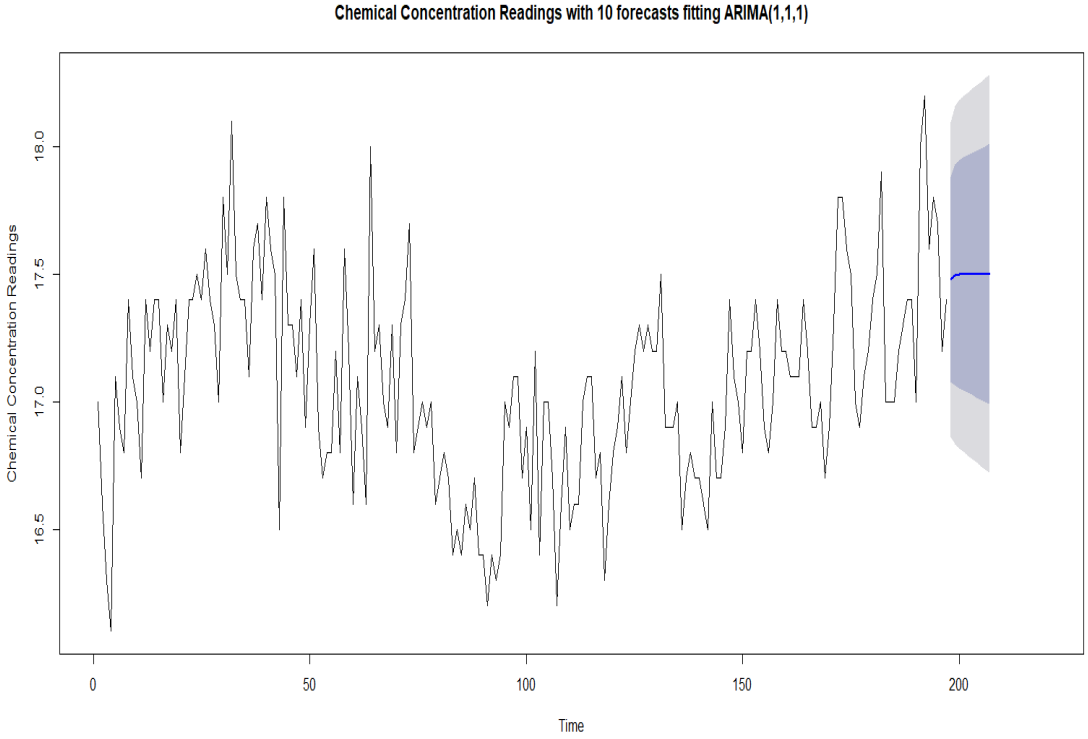
```
> fit <- arima(tsdl[[162]],order=c(2,1,1))
> jarque.bera.test(residuals(fit))

        Jarque Bera Test

data:  residuals(fit)
X-squared = 5.5811, df = 2, p-value = 0.06139

> ad.test(residuals(fit))

        Anderson-Darling test of goodness-of-fit
        Null hypothesis: uniform distribution
        Parameters assumed to be fixed

data:  residuals(fit)
An = Inf, p-value = 3.046e-06
```

Εικόνα 4.1.15
Έλεγχοι κανονικότητας υπολοίπων Jarque–Bera και Anderson-Darling για το μοντέλο ARIMA(2,1,1).

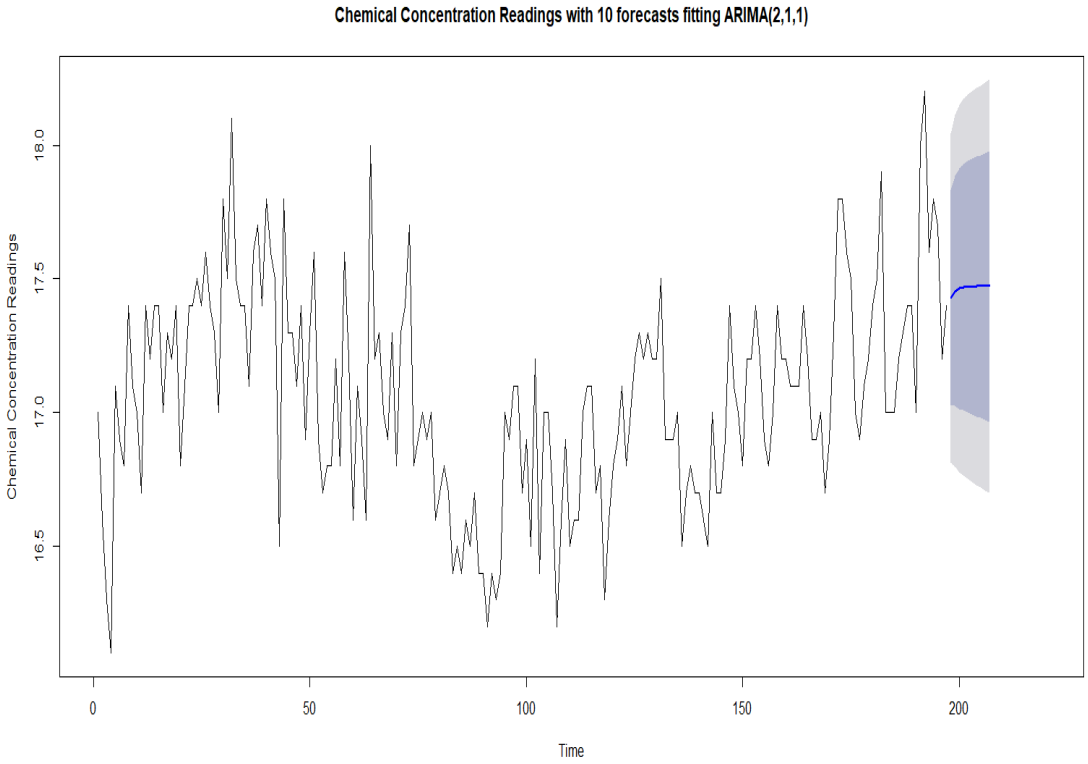
Από την Εικόνα 4.1.15 παρατηρείται ότι το p-value του ελέγχου Jarque–Bera δεν απορρίπτει την μηδενική υπόθεση περί κανονικότητας υπολοίπων ενώ το p-value του ελέγχου Anderson-Darling την απορρίπτει. Μια πιθανή εξήγηση είναι ότι ο έλεγχος Anderson-Darling δίνει μεγαλύτερο βάρος στις ουρές της κατανομής των υπολοίπων που και γραφικά (Διάγραμμα 4.1.11-QQ plot) φαίνεται ότι στις ουρές τα δεδομένα απέχουν από την ευθεία της κανονικής κατανομής.

4.1.8 Προβλεπτική ικανότητα μοντέλου

Χρησιμοποιώντας την R κατασκευάζονται τα γραφήματα πραγματικών τιμών των μετρήσεων χημικής συγκέντρωσης μαζί με τα μοντέλα ARIMA(1,1,1) και ARIMA(2,1,1) με τις επόμενες 10 προβλέψεις στο Διάγραμμα 4.1.16 και 4.1.17. Η σκούρα μπλε σκίαση δείχνει το 80% διάστημα εμπιστοσύνης για προβλέψεις του μοντέλου ενώ η ανοιχτή μπλε σκίαση δείχνει το 95% διάστημα εμπιστοσύνης για τις προβλέψεις του μοντέλου. Δεν παρατηρούνται γραφικά, σημαντικές διαφορές στις προβλέψεις των δύο μοντέλων. Εν τέλει, από τα μοντέλα ARIMA(1,1,1) και ARIMA(2,1,1) θα προτιμούνταν αυτό με τις καλύτερες προβλέψεις. Η σύγκριση θα μπορούσε να γίνει αν διατίθονταν οι πραγματικές μελλοντικές τιμές.



Διάγραμμα 4.1.16
Γραφική παράστασή των δεδομένων των μετρήσεων χημικής συγκέντρωσης μαζί με 10 προβλέψεις από το μοντελο ARIMA(1,1,1).



Διάγραμμα 4.1.17
Γραφική παράστασή των δεδομένων των μετρήσεων χημικής συγκέντρωσης μαζί με 10 προβλέψεις από το μοντελο ARIMA(2,1,1).

Εφαρμογή 2

4.2.1 Προκαταρτική ανάλυση για την δεύτερη χρονολογική σειρά

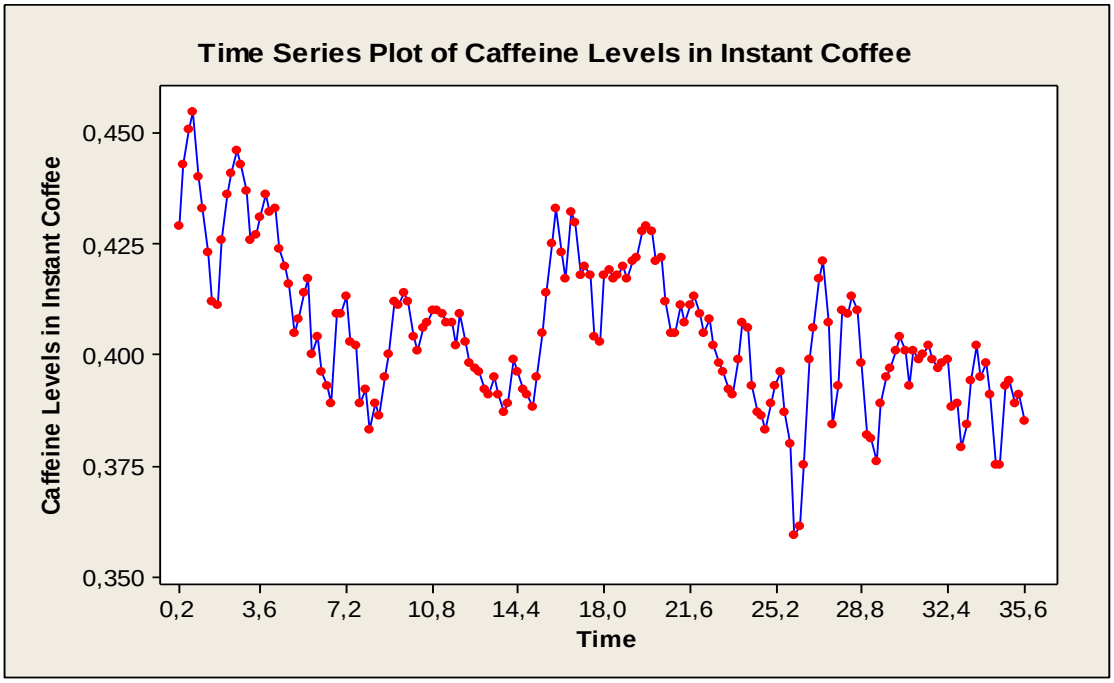
Η δεύτερη εφαρμογή αφορά την χρονολογική σειρά που αποτελείται από δεδομένα επιπέδου καφεΐνης στο στιγμιαίο καφέ (Hipel et al.,1994). Στην αρχή, παρουσιάζονται τα αριθμητικά δεδομένα στον Πίνακα 4.2.1 και η γραφική αναπαράσταση της χρονοσειράς στο Διάγραμμα 4.2.1. Επιπλέον, παρουσιάζεται η στατιστική ανάλυση των δεδομένων της χρονοσειράς στον Πίνακα 4.2.2.

Πίνακας 4.2.1
Δεδομένα του επιπέδου καφεΐνης στο στιγμιαίο καφέ.

```
> tsdl[[238]]
Time Series:
Start = c(1, 1)
End = c(36, 3)
Frequency = 5
  [1] 0.429 0.443 0.451 0.455 0.440 0.433 0.423 0.412 0.411 0.426 0.436 0.441 0.446 0.443 0.43
7 0.426 0.427 0.431 0.436 0.432 0.433 0.424 0.420 0.416 0.405 0.408 0.414 0.417 0.400 0.404
 [31] 0.396 0.393 0.389 0.409 0.409 0.413 0.403 0.402 0.389 0.392 0.383 0.389 0.386 0.395 0.40
0 0.412 0.411 0.414 0.412 0.404 0.401 0.406 0.407 0.410 0.410 0.409 0.407 0.407 0.402 0.409
 [61] 0.403 0.398 0.397 0.396 0.392 0.391 0.395 0.391 0.387 0.389 0.399 0.396 0.392 0.391 0.38
8 0.395 0.405 0.414 0.425 0.433 0.423 0.417 0.432 0.430 0.418 0.420 0.418 0.404 0.403 0.418
 [91] 0.419 0.417 0.418 0.420 0.417 0.421 0.422 0.428 0.429 0.428 0.421 0.422 0.412 0.405 0.40
5 0.411 0.407 0.411 0.413 0.409 0.405 0.408 0.402 0.398 0.396 0.392 0.391 0.399 0.407 0.406
[121] 0.393 0.387 0.386 0.383 0.389 0.393 0.396 0.387 0.380 0.359 0.361 0.375 0.399 0.406 0.41
7 0.421 0.407 0.384 0.393 0.410 0.409 0.413 0.410 0.398 0.382 0.381 0.376 0.389 0.395 0.397
[151] 0.401 0.404 0.401 0.393 0.401 0.399 0.400 0.402 0.399 0.397 0.398 0.399 0.388 0.389 0.37
9 0.384 0.394 0.402 0.395 0.398 0.391 0.375 0.375 0.393 0.394 0.389 0.391 0.385
attr(,"source")
[1] Hipel and McLeod (1994)
attr(,"description")
[1] Caffeine levels in instant coffee (seasonal period=5)
```

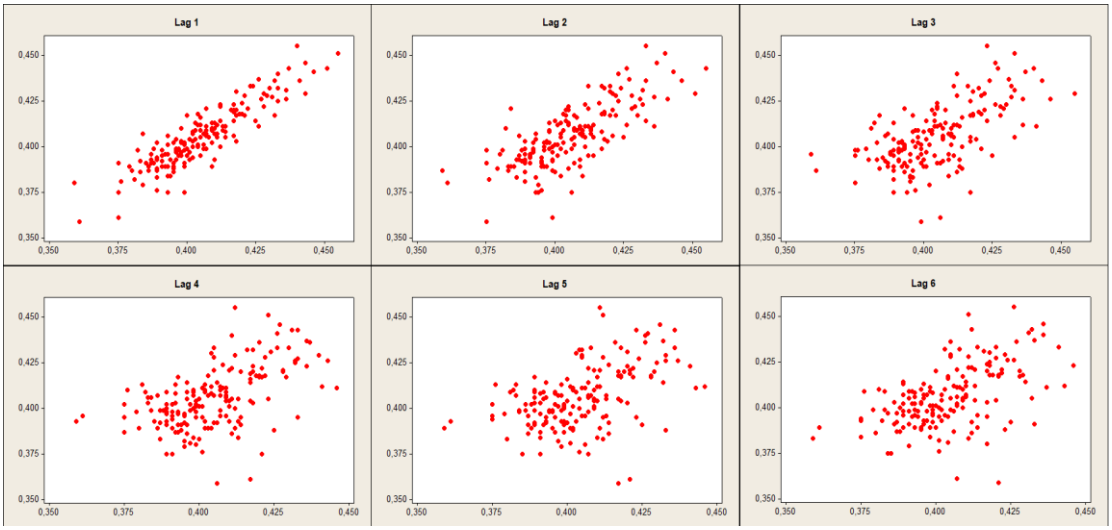
Πίνακας 4.2.2
Συγκεντρωτικός πίνακας περιγραφικής στατιστικής των δεδομένων του επιπέδου καφεΐνης στο στιγμιαίο καφέ.

Ελάχιστη τιμή	0,359
1ο ενδοτεταρτημοριακό εύρος Q	0,393
Διάμεσος	0,404
Μέση τιμή	0,4055
3ο ενδοτεταρτημοριακό εύρος Q	0,417
Μέγιστη τιμή	0,455
Διακύμανση	0,000287
Τυπική απόκλιση	0,01693
Λοξότητα	0,34
Κύρτωση	0,21



Διάγραμμα 4.2.1
Γραφική παράσταση δεδομένων του επιπέδου καφεΐνης στο στιγμιαίο καφέ.

Δημιουργούνται τα διαγράμματα διασποράς (scatter plots) στο Διάγραμμα 4.2.2 που βοηθάνε στην αναγνώριση της συναρτησιακής μορφής της υπό μελέτη μεταβλητής σε προηγούμενες χρονικές περιόδους. Στις αρχικές υστερήσεις φαίνεται η έντονη γραμμική σχέση μεταξύ των τιμών που στη συνέχεια χάνεται όσο μεγαλώνουν οι υστερήσεις. Πιο συγκεκριμένα, από την υστέρηση h=5 και μετά παρατηρείται μεγάλη εξασθένιση της γραμμικής συσχέτισης και για αυτόν το λόγο, σε οποιαδήποτε περίπτωση προσαρμογής μοντέλου, οποιαδήποτε υστέρηση μεγαλύτερη του πέντε (h>5) είναι στατιστικά σημαντική δεν θα ληφθεί υπόψιν.



Διάγραμμα 4.2.2
Διαγράμματα διασποράς των δεδομένων των μετρήσεων χημικής συγκέντρωσης σε υστερήσεις h=1,2…6.

4.2.2 Έλεγχος στασιμότητας με χρήση του επαυξημένου ελέγχου Dickey-Fuller και του ελέγχου Kwiatkowski-Phillips-Schmidt-Shin

Όπως και στη προηγουμένη εφαρμογή, γίνεται έλεγχος της στασιμότητας της χρονολογικής σειράς με τους ελέγχους ADF και KPSS. Χρησιμοποιώντας τις εντολές της R γλώσσας adf.test και kpss.test λαμβάνονται τα αποτελέσματα των ελέγχων ADF και KPSS στην Εικόνα 4.2.1.

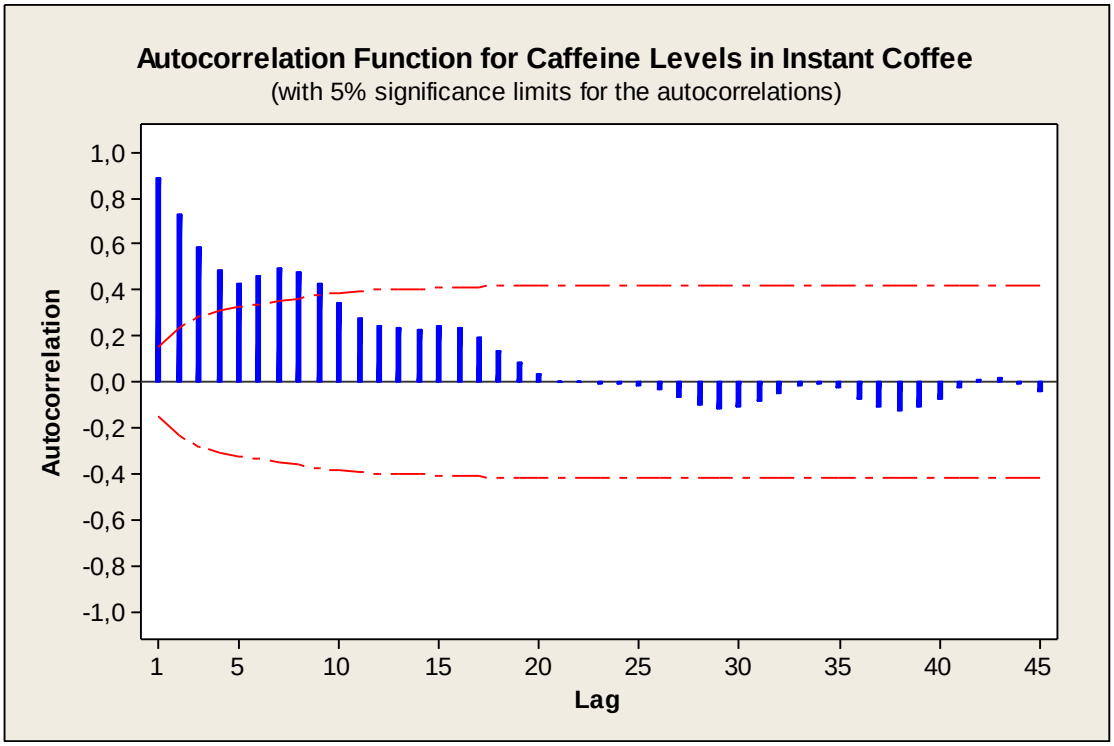
```
> adf.test(tsdl[[238]])

        Augmented Dickey-Fuller Test

data:  tsdl[[238]]
Dickey-Fuller = -2.2021, Lag order = 5, p-value = 0.4918
alternative hypothesis: stationary

> kpss.test(tsdl[[238]],null="Trend")

        KPSS Test for Trend Stationarity

data:  tsdl[[238]]
KPSS Trend = 0.16593, Truncation lag parameter = 4, p-value = 0.03339
```

Εικόνα 4.2.1
Αποτελέσματα ελέγχων ADF και KPSS για τα δεδομένα του επιπέδου καφεΐνης στο στιγμιαίο καφέ.

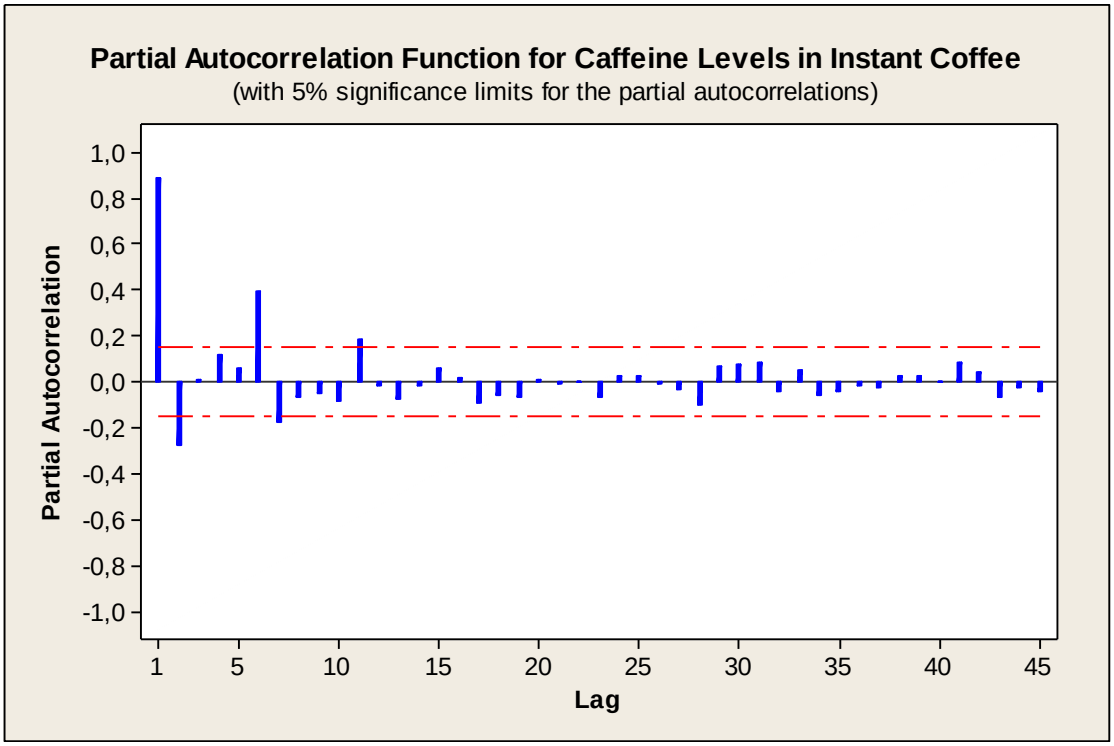
Από τις τιμές των p-value και των δύο ελέγχων, συμπεραίνεται εύκολα ότι η χρονολογική σειρά δεν είναι στάσιμη.

4.2.3 Υπολογισμός της συνάρτησης αυτοσυσχέτισης (ACF) και της συνάρτησης μερικής αυτοσυσχέτισης (PACF)

Για τον υπολογισμό της συνάρτησης αυτοσυσχέτισης (ACF) και μερικής αυτοσυσχέτισης (PACF) χρησιμοποιείται το υπολογιστικό πρόγραμμα Minitab 14. Ακολουθεί το Διάγραμμα 4.2.3 και το Διάγραμμα 4.2.4 με τα γραφήματα αυτών.



Διάγραμμα 4.2.3
Συνάρτηση αυτοσυσχέτισης (ACF) των δεδομένων του επιπέδου καφεΐνης στο στιγμιαίο καφέ.



Διάγραμμα 4.2.4
Συνάρτηση μερικής αυτοσυσχέτισης (PACF) των δεδομένων του επιπέδου καφεΐνης στο στιγμιαίο καφέ.

Με γραφικό τρόπο επιβεβαιώνεται το γεγονός ότι η χρονοσειρά με το επίπεδο καφεΐνης στο στιγμιαίο καφέ είναι μη στάσιμη.

4.2.4 Αντιμετώπιση της στασιμότητας

Για την αντιμετώπιση της στασιμότητας, προτείνεται η διαφόριση των δεδομένων του επιπέδου καφεΐνης στο στιγμιαίο καφέ. Ξαναυπολογίζονται επιπλέον και οι έλεγχοι ADF και KPSS για τα νέα μετασχηματισμένα δεδομένα με πρώτες διάφορες και τα αποτελέσματα αυτών φαίνονται στην Εικόνα 4.2.2.

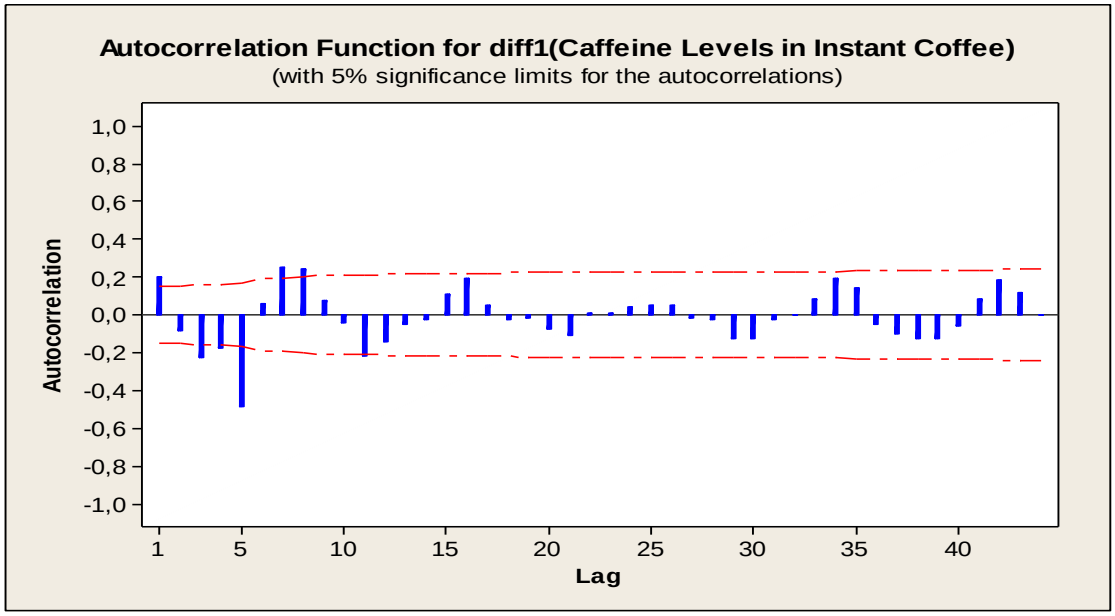
```
> adf.test(diff(tsdl[[238]]))

        Augmented Dickey-Fuller Test

data:  diff(tsdl[[238]])
Dickey-Fuller = -7.0074, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(diff(tsdl[[238]])) : p-value smaller than printed p-value
> kpss.test(diff(tsdl[[238]]))

        KPSS Test for Level Stationarity

data:  diff(tsdl[[238]])
KPSS Level = 0.0208, Truncation lag parameter = 4, p-value = 0.1

Warning message:
In kpss.test(diff(tsdl[[238]])) : p-value greater than printed p-value
```

Εικόνα 4.2.2
Έλεγχοι ADF και KPSS για τα μετασχηματισμένα δεδομένα με πρώτες διάφορες του επιπέδου καφεΐνης στο στιγμιαίο καφέ.

Με την χρήση των ελέγχων συμπεραίνεται ότι η χρονοσειρά παρουσιάζει πλέον στασιμότητα.

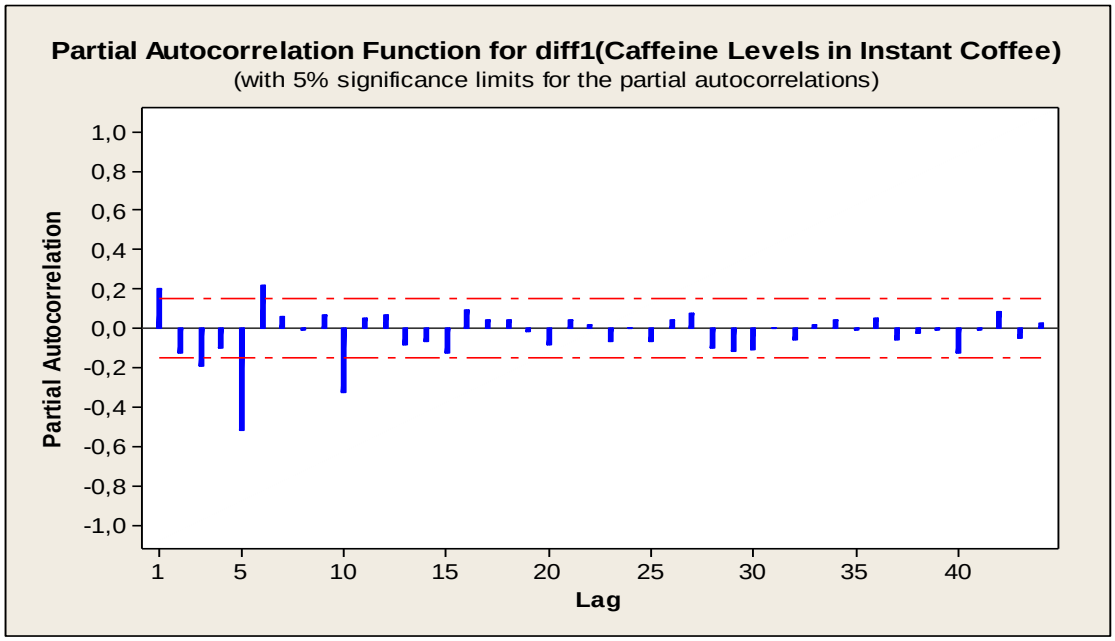
4.2.5 Επιλογή παραμέτρων για μοντελοποίηση

Το επόμενο βήμα περιλαμβάνει μια πρώτη εκτίμηση για τις τάξεις των p, q για το μοντέλο. Έχει γίνει αναφορά σε προηγούμενες ενότητες (Πίνακας 3.1) ότι αυτό γίνεται με τη βοήθεια των γραφημάτων των συναρτήσεων ACF και PACF . Για την εκτίμηση των συντελεστών κατασκευάζονται τα γραφήματα των ACF και PACF που απεικονίζονται κατά αντιστοιχία στο Διάγραμμα 4.2.5 και στο Διάγραμμα 4.2.6.



Διάγραμμα 4.2.5
Συνάρτηση αυτοσυσχέτισης (ACF) των μετασχηματισμένων με πρώτες διαφορές δεδομένων του επιπέδου καφεΐνης στο στιγμιαίο καφέ.

Το Διάγραμμα 4.2.4 δείχνει πολλούς όρους της ACF να πέφτουν εκτός ορίων με 5% επίπεδο σημαντικότητας. Η εντονότερη απόκλιση από τα όρια βρίσκεται στον πέμπτο ορό της ACF οπότε προτείνεται ένα εποχιακό μοντελο ARIMA η αλλιώς SARIMA(p,d,q)(P,D,Q)$_5$. Οι πιθανές τιμές για το q είναι μεγαλύτερες του 1 και για το Q προτείνεται η τιμή 1 εφόσον ο πέμπτος όρος είναι εκτός ορίων.



Διάγραμμα 4.2.6
Συνάρτηση μερικής αυτοσυσχέτισης (ACF) των μετασχηματισμένων με πρώτες διαφορές δεδομένων του επιπέδου καφεΐνης στο στιγμιαίο καφέ

Στο Διάγραμμα 4.2.5 φαίνονται έντονες αποκλίσεις για το πέμπτο και το δέκατο όρο της PACF οπότε ενισχύεται η άποψη για την χρήση ενός SARIMA μοντέλου. Στην περίπτωση εδώ εκτός ορίων βρίσκονται ο πρώτος και ο τρίτος όρος οπότε πιθανές τιμές του p είναι 1 και μεγαλύτερες του 1. Όσον αφορά τον εποχικό όρο P προτεινόμενη τιμή είναι το 1 καθότι ο πέμπτος ορός βρίσκεται εκτός ορίων.

4.2.6 Επιλογή μοντέλου ARIMA

Καθότι τα πιθανά μοντέλα είναι πολλά σε πληθώρα χρησιμοποιείται η εντολή auto.arima με τη βοήθεια της R. Στην περίπτωση που δεν υπήρχε αυτή η δυνατότητα προτείνεται ο παρακάτω τρόπος για τη διαλογή του μοντέλου. Αρχικά, επιλέγεται το μοντελο με τις μέγιστες αποδέκτες, σύμφωνα με τις συναρτήσεις αυτοσυσχέτισης (ACF) και μερικής αυτοσυσχέτισης (PACF), τάξεις των AR και MA παραμέτρων . Σε αντιστοιχία αυτό γίνεται και με το εποχιακό κομμάτι. Στην συνέχεια μελετώνται όλα τα εμφωλευμένα σε αυτό πιθανά μοντέλα ARIMA μικρότερης τάξης Τελικά, επιλέγεται αυτό με το μικρότερο AIC, AICc ή BIC που είναι στατιστικά σημαντικό και περιέχει μικρά σφάλματα. Τα αποτελέσματα του auto.arima παρατάσσονται στην Εικόνα 4.2.3.

```
> auto.arima(tsdl[[238]],approximation=FALSE,stepwise=FALSE
Series: tsdl[[238]]
ARIMA(2,1,2)(0,0,1)[5]

Coefficients:
          ar1      ar2     ma1     ma2     sma1
      -0.7549  -0.3378  1.1411  0.6708  -0.8371
s.e.   0.1600   0.1750  0.1382  0.1231   0.0550

sigma^2 estimated as 2.951e-05:  log likelihood=670.96
AIC=-1329.92   AICc=-1329.42   BIC=-1310.86
```

Εικόνα 4.2.3
Αποτελέσματα auto.arima για τη χρονοσειρά επιπέδου καφεΐνης στο στιγμιαίο καφέ.

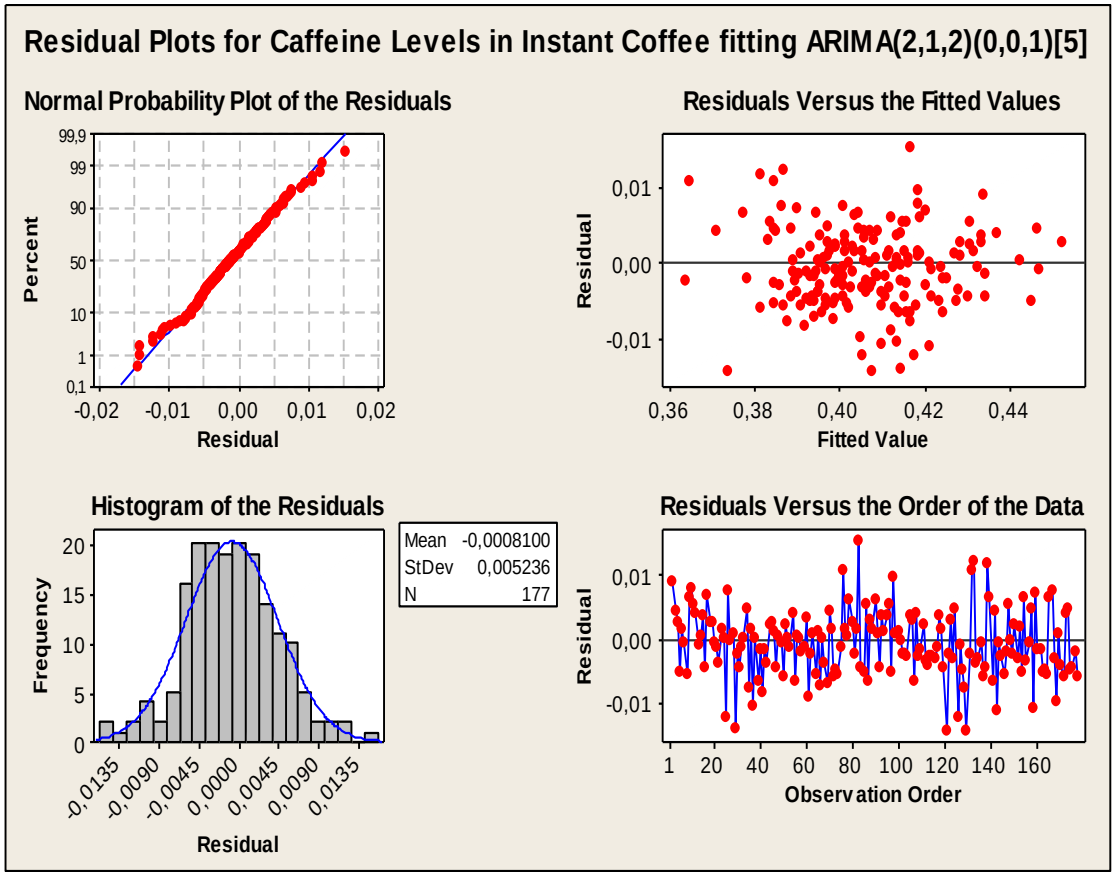
Σύμφωνα λοιπόν με το κριτήριο AICc το βέλτιστο μοντελο είναι το ARIMA(2,1,2)$(0,0,1)_5$. Χρησιμοποιείται το Minitab 14 για να ελεγχθεί αν οι παράμετροι του μοντέλου είναι στατιστικά σημαντικοί. Κρίνοντας από τα p-value όλοι οι παράμετροι είναι στατιστικά σημαντικοί. Παρατίθεται το μοντελο ARIMA(2,1,2)$(0,0,1)_5$ στην Εικόνα 4.2.4.

```
Final Estimates of Parameters

Type       Coef  SE Coef      T      P
AR   1  -0,7210   0,1962  -3,67  0,000
AR   2  -0,3343   0,1517  -2,20  0,029
MA   1  -1,1083   0,1714  -6,46  0,000
MA   2  -0,6632   0,1175  -5,65  0,000
SMA  5   0,8943   0,0380  23,54  0,000


Differencing: 1 regular difference
Number of observations:  Original series 178, after differencing 177
Residuals:    SS =  0,00494073 (backforecasts excluded)
              MS =  0,00002873  DF = 172


Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   15,1   21,0   32,5   43,6
DF              7     19     31     43
P-Value     0,035  0,335  0,393  0,448
```

Εικόνα 4.2.4.
Αποτελέσματα μοντέλου ARIMA(2,1,2)(0,0,1)$_5$.

4.2.7 Διαγνωστικοί έλεγχοι υπολοίπων

Τα υπολείμματα είναι χρήσιμα στο να ελέγχουν εάν ένα μοντέλο έχει καταγράψει επαρκώς τις πληροφορίες στα δεδομένα. Μια καλή μέθοδος πρόβλεψης προϋποθέτει ότι δεν υπάρχουν συσχετίσεις μεταξύ υπολειμμάτων επειδή τότε δεν υπάρχουν πληροφορίες που απομένουν στα υπολείμματα. Επιπλέον, εάν τα υπολείμματα έχουν μέσο όρο διαφορετικό από το μηδέν τότε οι προβλέψεις είναι μεροληπτικές. Το βασικότερό λοιπόν για να θεωρείται αξιόπιστο το μοντελο πρόβλεψης είναι τα υπόλοιπα να έχουν ιδιότητες λευκού θορύβου. Αυτό περιλαμβάνει ότι η μέση τιμή των υπολοίπων είναι μηδενική και η διακύμανση τους είναι σταθερή στο πέρας του χρόνου. Ξεκινώντας, λοιπόν, κατασκευάζονται περιληπτικά στο Διάγραμμα 4.2.7 το γράφημα των υπολοίπων στο χρόνο, το Q-Q plot, το ιστόγραμμα των υπολοίπων και το γράφημα των υπολοίπων σε σχέση με τις προσαρμοσμένες τιμές για το μοντελο ARIMA(2,1,2)(0,0,1)$_5$.

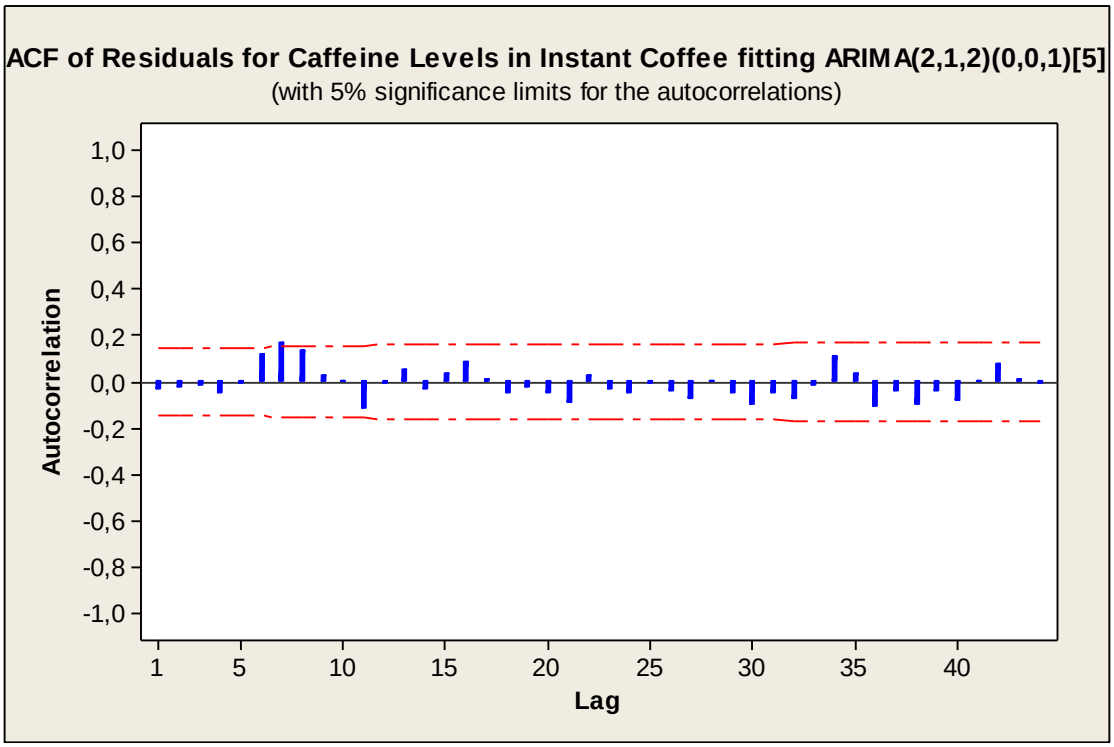


Διάγραμμα 4.2.7
Γράφημα των υπολοίπων στο χρόνο, Q-Q plot, ιστόγραμμα των υπολοίπων και το γράφημα των υπολοίπων σε σχέση με τις προσαρμοσμένες τιμές για το μοντελο ARIMA$(2,1,2)(0,0,1)_5$.

Αρχικά, η μέση τιμή των υπολοίπων είναι -0,0008 που θεωρείται μηδενική. Όσον αφορά τη διακύμανση στο πέρας του χρόνου από το γράφημα των υπολοίπων σε σχέση με τις προσαρμοσμένες τιμές παρατηρείται ομοσκεδαστικότητα που επιβεβαιώνεται με το arch.test. που παρουσιάζεται στην Εικόνα 4.2.5. Παρατηρώντας την συνάρτηση αυτοσυσχέτισης (ACF) των υπολοίπων, που φαίνεται στο Διάγραμμα 4.2.8 παρακάτω, τα υπόλοιπα είναι ανεξάρτητα καθώς μόνο ένας όρος βρίσκονται εκτός ορίων. Αυτό ενισχύεται και με τον έλεγχο Ljung-Box, με μηδενική υπόθεση ότι τα υπόλοιπα είναι ισόνομα και ανεξάρτητα, που φαίνονται στην Εικόνα 4.2.6. Επιπλέον, από το ιστόγραμμα των υπολοίπων και το Q-Q plot διαφαίνεται μια αποδεκτή εικόνα κανονικότητας. Επιβεβαιώνεται η υπόθεσης της κανονικότητας των υπολοίπων με τον έλεγχο Jarque–Bera και όχι με τον Anderson-Darling, λόγω των ουρών της κατανομής, από τα παρατιθέμενα αποτελέσματα στην Εικόνα 4.2.7.

```
> arch.test(fit)
ARCH heteroscedasticity test for residuals
alternative: heteroscedastic

Portmanteau-Q test:
      order    PQ p.value
[1,]     4  6.55   0.162
[2,]     8  8.34   0.401
[3,]    12 10.59   0.564
[4,]    16 12.42   0.715
[5,]    20 15.30   0.759
[6,]    24 17.27   0.837
Lagrange-Multiplier test:
      order    LM  p.value
[1,]     4 66.47 2.43e-14
[2,]     8 26.42 4.24e-04
[3,]    12 15.96 1.43e-01
[4,]    16 10.30 8.00e-01
[5,]    20  6.68 9.96e-01
[6,]    24  5.02 1.00e+00
```

Εικόνα 4.2.5
Αποτελέσματα ελέγχων ομοσκεδαστικότητας υπολοίπων για το μοντελο ARIMA(2,1,2)(0,0,1)$_5$.



Διάγραμμα 4.2.8
Συνάρτηση αυτοσυσχέτισης (ACF) των υπολοίπων για το μοντελο ARIMA(2,1,2)(0,0,1)$_5$.

```
> checkresiduals(fit)

        Ljung-Box test

data:  Residuals from ARIMA(2,1,2)(0,0,1)[5]
Q* = 12.291, df = 5, p-value = 0.03102

Model df: 5.   Total lags used: 10
```

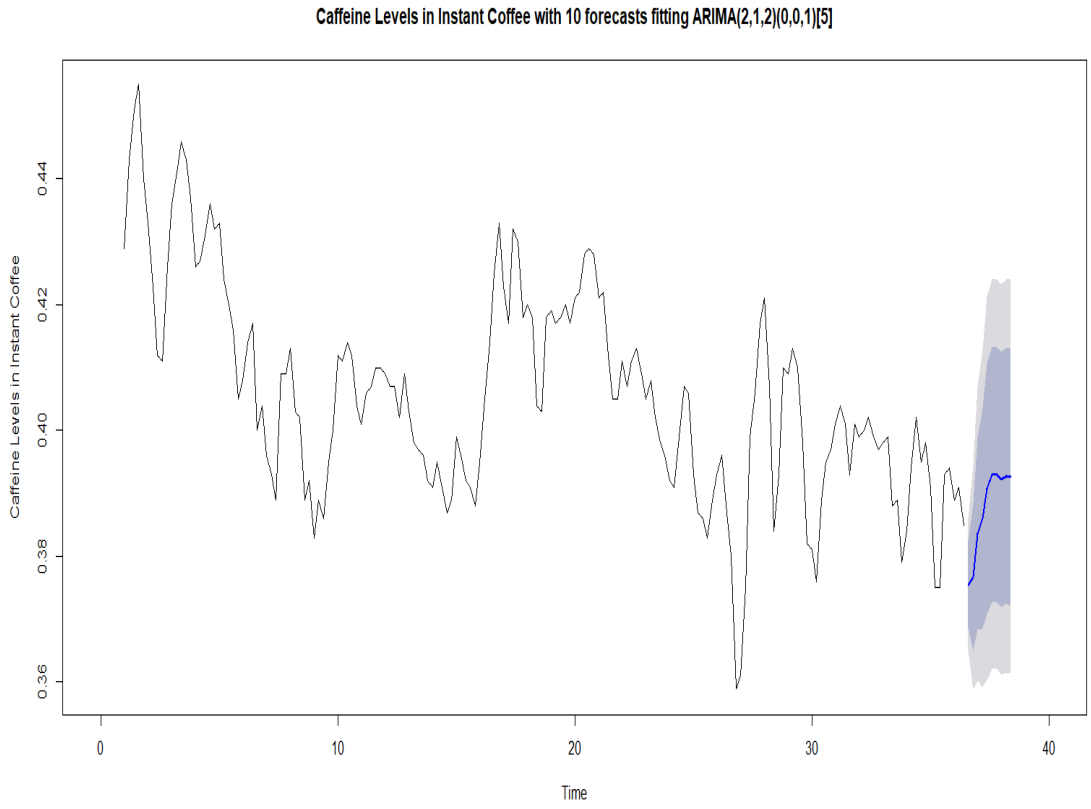
Εικόνα 4.2.6
Έλεγχος Ljung-Box για το μοντελο ARIMA(2,1,2)$(0,0,1)_5$.

```
> ad.test(res)

        Anderson-Darling test of goodness-of-fit
        Null hypothesis: uniform distribution
        Parameters assumed to be fixed

data:  res
An = Inf, p-value = 3.371e-06

> jarque.bera.test(res)

        Jarque Bera Test

data:  res
X-squared = 3.0057, df = 2, p-value = 0.2225
```

Εικόνα 4.2.7
Έλεγχοι κανονικότητας υπολοίπων Jarque–Bera και Anderson-Darling για το μοντελο ARIMA(2,1,2)$(0,0,1)_5$.

4.2.8 Προβλεπτική ικανότητα μοντέλου

Χρησιμοποιώντας την προγραμματιστική γλώσσα R και την εντολή forecast από την βιβλιοθήκη forecast κατασκευάζεται το γράφημα των πραγματικών τιμών του επιπέδου καφεΐνης στον στιγμιαίο καφέ μαζί με το εποχικό μοντέλο ARIMA(2,1,2)$(0,0,1)_5$ με τις επόμενες 10 προβλέψεις στο Διάγραμμα 4.2.9. Η σκούρα μπλε σκίαση υποδεικνύει το 80% διάστημα εμπιστοσύνης για προβλέψεις του μοντέλου ενώ η ανοιχτή μπλε σκίαση υποδεικνύει το 95% διάστημα εμπιστοσύνης για τις προβλέψεις του μοντέλου.



Διάγραμμα 4.2.9
Γραφική παράστασή των δεδομένων του επιπέδου καφεΐνης στον στιγμιαίο καφέ μαζί με 10 προβλέψεις από το μοντελο ARIMA(2,1,2)$(0,0,1)_5$.

Εφαρμογή 3

4.3.1 Προκαταρτική ανάλυση για την τρίτη χρονολογική σειρά

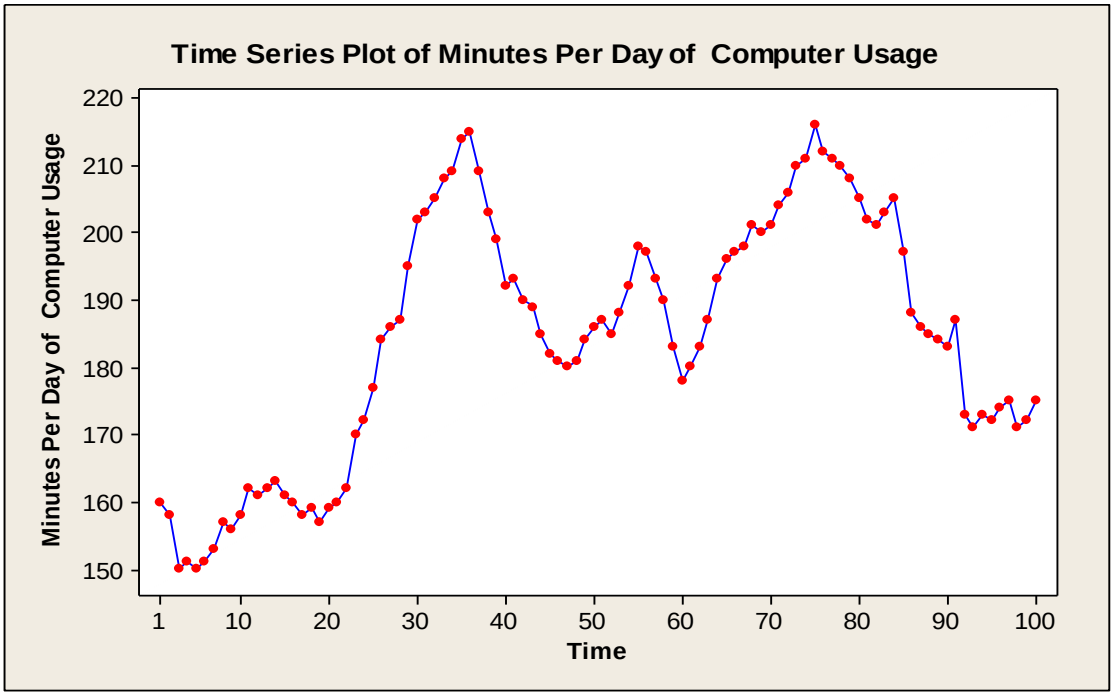
Η τρίτη και τελευταία εφαρμογή αφορά την χρονολογική σειρά που αποτελείται από δεδομένα καθημερινής χρήσης ηλεκτρονικού υπολογιστή σε λεπτά (Montgomery, 1976). Στην αρχή, παρουσιάζονται τα αριθμητικά δεδομένα στον Πίνακα 4.3.1 και η γραφική αναπαράσταση της χρονοσειράς στο Διάγραμμα 4.3.1. Επιπλέον, παρουσιάζεται η στατιστική ανάλυση των δεδομένων της χρονοσειράς στον Πίνακα 4.3.2.

Πίνακας 4.3.1
Δεδομένα καθημερινής χρήσης ηλεκτρονικού υπολογιστή σε λεπτά.

```
> > > tsdl[[505]]
Time Series:
Start = 1
End = 100
Frequency = 1
  [1] 160 158 150 151 150 151 153 157 156 158 162 161 162 163 161 160 158 159 157 159 160 162
170 172 177 184 186 187 195 202 203 205 208 209 214 215 209 203 199 192 193 190 189 185 182
 [46] 181 180 181 184 186 187 185 188 192 198 197 193 190 183 178 180 183 187 193 196 197 198
201 200 201 204 206 210 211 216 212 211 210 208 205 202 201 203 205 197 188 186 185 184 183
 [91] 187 173 171 173 172 174 175 171 172 175
attr(,"source")
[1] Montgomery          (1976)
attr(,"description")
[1] Minutes of usage per day of a computer terminal (p.270: Montgomery)
```

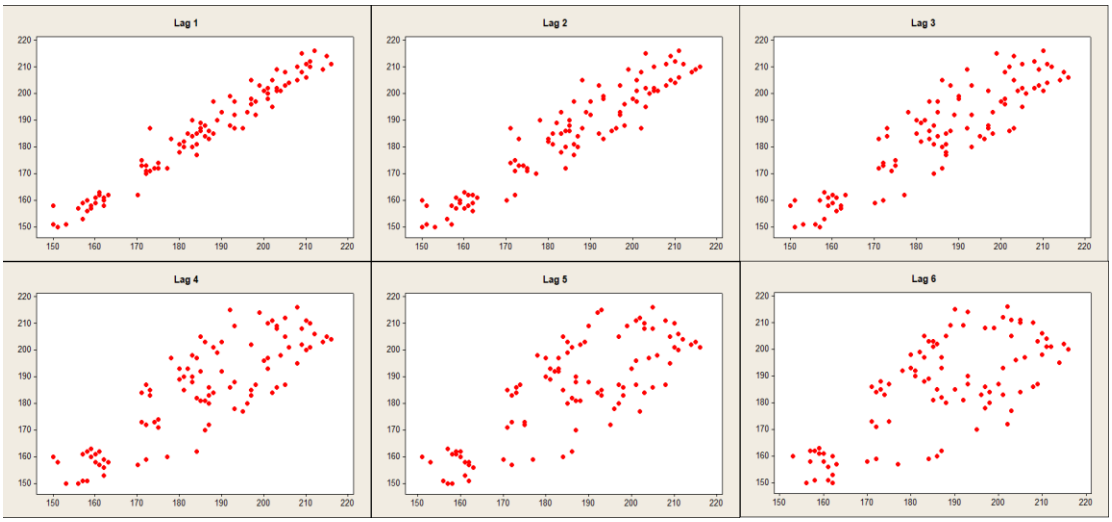
Πίνακας 4.3.2
Περιγραφική στατιστική για δεδομένα καθημερινής χρήσης ηλεκτρονικού υπολογιστή σε λεπτά.

Ελάχιστη τιμή	150
1ο ενδοτεταρτημοριακό εύρος Q	171,25
Διάμεσος	186
Μέση τιμή	184,46
3ο ενδοτεταρτημοριακό εύρος Q	200,75
Μέγιστη τιμή	216
Διακύμανση	334,57
Τυπική απόκλιση	18,29
Λοξότητα	-0,22
Κύρτωση	-1,02



Διάγραμμα 4.3.1
Γραφική παράσταση των δεδομένων καθημερινής χρήσης ηλεκτρονικού υπολογιστή σε λεπτά.

Δημιουργούνται τα διαγράμματα διασποράς στο Διάγραμμα 4.3.2 που κύριο στόχο έχουν την αναγνώριση της συναρτησιακής μορφής της προς μελέτης μεταβλητής σε προηγούμενες χρονικές περιόδους. Στις αρχικές υστερήσεις φαίνεται η έντονη γραμμική σχέση μεταξύ των τιμών που στη συνέχεια χάνεται. Συγκεκριμένα, από την υστέρηση h=4 και μετά, παρατηρείται μεγάλη εξασθένιση της γραμμικής συσχέτισης και έτσι σε οποιαδήποτε περίπτωση προσαρμογής μοντέλου, οποιαδήποτε υστέρηση μεγαλύτερη του τέσσερα (h>4) δεν θα ληφθεί υπόψιν.



Διάγραμμα 4.3.2
Διαγράμματα διασποράς των δεδομένων καθημερινής χρήσης ηλεκτρονικού υπολογιστή σε λεπτά σε υστερήσεις h=1,2…6.

4.3.2 Έλεγχος στασιμότητας με χρήση του επαυξημένου ελέγχου Dickey-Fuller και του ελέγχου Kwiatkowski-Phillips-Schmidt-Shin

Σε πρώτο βήμα, γίνεται έλεγχος της στασιμότητας της χρονολογικής σειράς με τους ελέγχους ADF και KPSS. Χρησιμοποιώντας τις εντολές της R adf.test και kpss.test λαμβάνονται τα αποτελέσματα των ελέγχων ADF και KPSS στην Εικόνα 4.3.1.

```
> adf.test(tsdl[[505]])

        Augmented Dickey-Fuller Test

data:  tsdl[[505]]
Dickey-Fuller = -1.9071, Lag order = 4, p-value = 0.6151
alternative hypothesis: stationary

> kpss.test(tsdl[[505]])

        KPSS Test for Level Stationarity

data:  tsdl[[505]]
KPSS Level = 0.77212, Truncation lag parameter = 4, p-value = 0.01
```

Εικόνα 4.3.1.
Αποτελέσματα ελέγχου ADF και KPSS για τα δεδομένα καθημερινής χρήσης ηλεκτρονικού υπολογιστή σε λεπτά.

Κρίνοντας από τις τιμές των p-value η χρονολογική σειρά δεν είναι στάσιμη. Αυτό θα μπορούσε να επιβεβαιωθεί και με τα γραφήματα των συναρτήσεων αυτοσυσχέτισης (ACF) και μερικής αυτοσυσχέτισης (PACF) ,όπως στις προηγούμενες δυο εφαρμογές, αλλά παραλείπεται.

4.3.3 Αντιμετώπιση της στασιμότητας

Για την αντιμετώπιση της στασιμότητας, προτείνεται ο μετασχηματισμός με πρώτες διάφορες των δεδομένων καθημερινής χρήσης ηλεκτρονικού υπολογιστή σε λεπτά. Ξαναυπολογίζονται επιπλέον και οι έλεγχοι ADF και KPSS για τα νέα μετασχηματισμένα με πρώτες διάφορες δεδομένα καθημερινής χρήσης ηλεκτρονικού υπολογιστή σε λεπτά και τα αποτελέσματα αυτών φαίνονται στην Εικόνα 4.3.2.

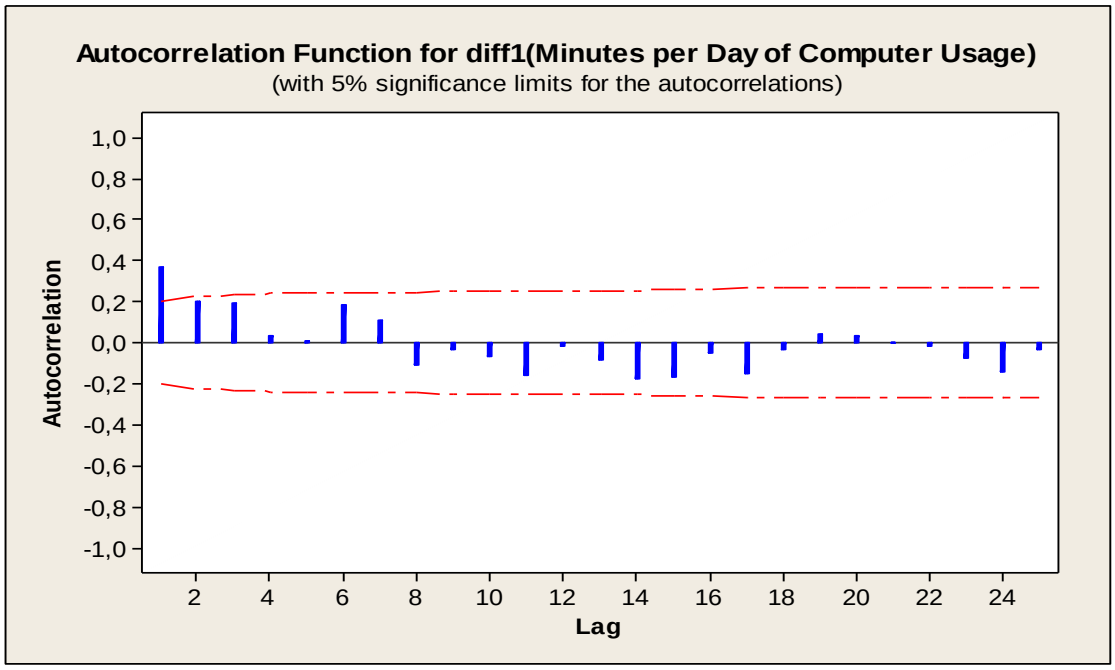
```
> adf.test(diff(tsdl[[505]]))

        Augmented Dickey-Fuller Test

data:  diff(tsdl[[505]])
Dickey-Fuller = -3.8765, Lag order = 4, p-value = 0.01808
alternative hypothesis: stationary

> kpss.test(diff(tsdl[[505]]))

        KPSS Test for Level Stationarity

data:  diff(tsdl[[505]])
KPSS Level = 0.21966, Truncation lag parameter = 3, p-value = 0.1
```

Εικόνα 4.3.2
Αποτελέσματα ελέγχου ADF και KPSS για τα μετασχηματισμένα με πρώτες διαφορές δεδομένα καθημερινής χρήσης ηλεκτρονικού υπολογιστή σε λεπτά.

Η χρονολογική σειρά μετά τη χρήση πρώτων διαφορών παρουσιάζει στασιμότητα.

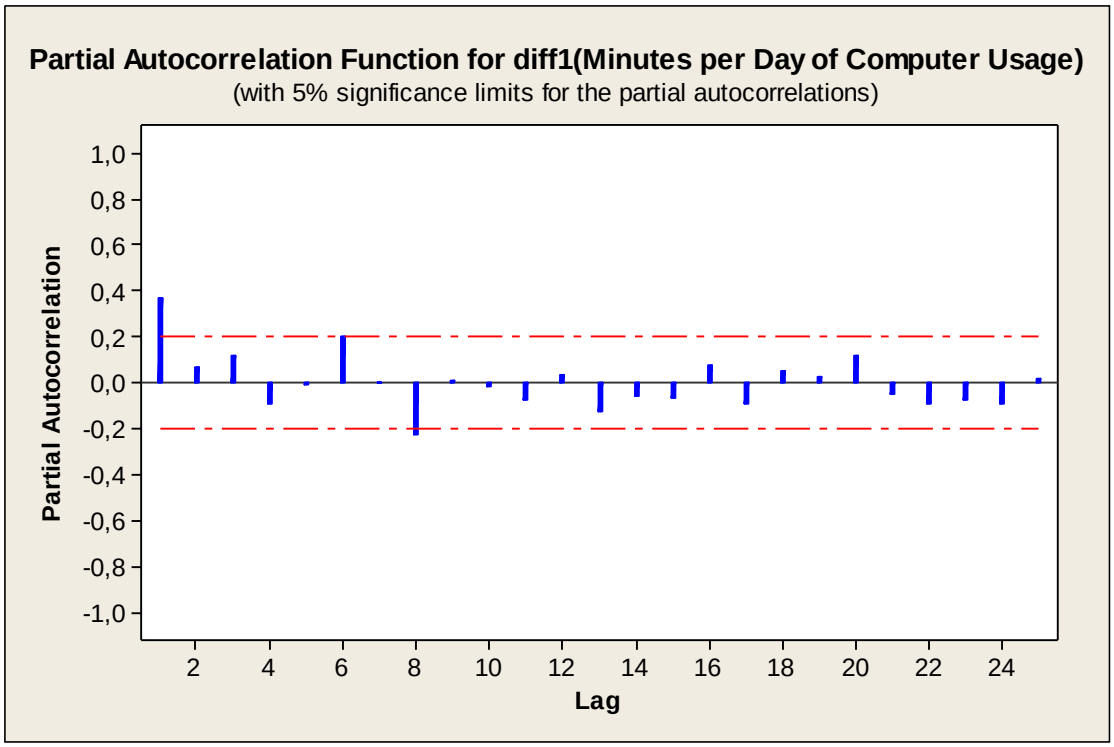
4.3.4 Επιλογή παραμέτρων για μοντελοποίηση

Το επόμενο βήμα της ανάλυσης χρονοσειρών περιλαμβάνει μια πρώτη εκτίμηση για τις τάξεις των p, q για το μοντέλο. Σε προηγούμενες ενότητες γίνεται αναφορά ότι αυτό επιτυγχάνεται με τη βοήθεια των γραφημάτων των συναρτήσεων ACF και PACF. Κατασκευάζονται τα γραφήματα των ACF και PACF που απεικονίζονται κατά αντιστοιχία στο Διάγραμμα 4.3.3 και 4.3.4.



Διάγραμμα 4.3.3
Συνάρτηση αυτοσυσχέτισης (ACF) των μετασχηματισμένων με πρώτες διαφορές δεδομένων της καθημερινής χρήσης ηλεκτρονικού υπολογιστή σε λεπτά.

Στο Διάγραμμα 4.3.3 παρατηρείται ότι μόνο ο πρώτος όρος της ACF είναι εκτός ορίων για επίπεδο σημαντικότητας 5%. Σύμφωνα με τον Πίνακα 3.1 η πιθανότερη τιμή για την τάξη του MA(q) είναι η πρώτη (q=1).



Διάγραμμα 4.3.4
Συνάρτηση μερικής αυτοσυσχέτισης (PACF) των μετασχηματισμένων με πρώτες διαφορές δεδομένων της καθημερινής χρήσης ηλεκτρονικού υπολογιστή σε λεπτά.

Στο Διάγραμμα 4.3.4 παρατηρείται ότι μόνο ο πρώτος όρος και ο όγδοος της ACF είναι εκτός ορίων για επίπεδο σημαντικότητας 5%. Ωστόσο, εξαιτίας του γεγονότος ότι για κάθε υστέρηση φθίνει η γραμμική εξάρτηση των δεδομένων, δε θα θεωρηθεί στατιστικά σημαντικός ο όγδοος όρος. Σύμφωνα με τον Πίνακα 3.1 η πιθανότερη τιμή για την τάξη του AR(p) είναι η πρώτη (p=1).

4.3.5 Επιλογή μοντέλου ARIMA

Το πιθανότερο μοντελο ARIMA, σύμφωνα με τη μελέτη των συναρτήσεων αυτοσυσχέτισης και μερικής αυτοσυσχέτισης, είναι το ARIMA(1,1,1). Χρησιμοποιείται η εντολή auto.arima με τη βοήθεια της R. Τα λεπτομερή αποτελέσματα του auto.arima φαίνονται στην Εικόνα 4.3.3.

```
> auto.arima(tsdl[[505]],approximation=FALSE,stepwise=FALSE)
Series: tsdl[[505]]
ARIMA(3,1,2)

Coefficients:
          ar1      ar2     ar3     ma1     ma2
      -0.2441  -0.6036  0.4806  0.6254  0.9404
s.e.   0.0982   0.0891  0.0898  0.0611  0.1471

sigma^2 estimated as 11.07:  log likelihood=-258.06
AIC=528.12   AICc=529.03   BIC=543.69
```

Εικόνα 4.3.3.
Αποτελέσματα auto.arima για τη χρονοσειρά καθημερινής χρήσης ηλεκτρονικού υπολογιστή σε λεπτά.

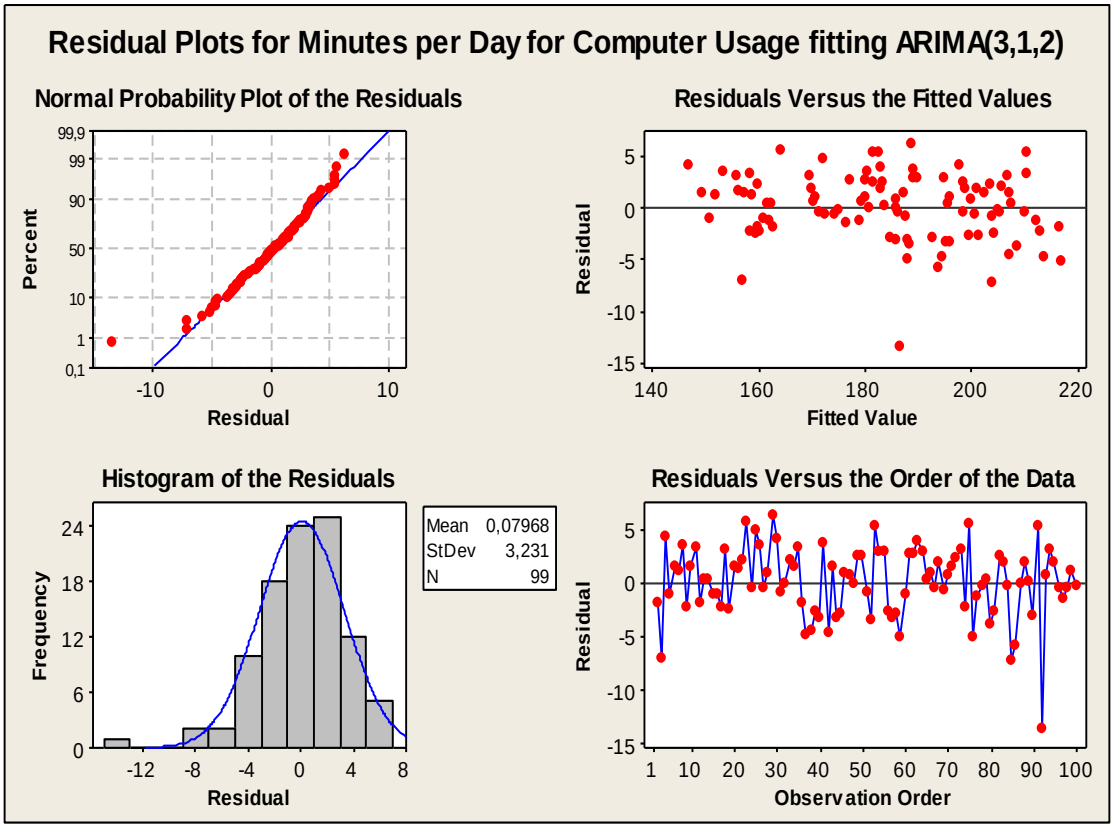
Προτείνεται το μοντελο ARIMA(3,1,2) βάσει του κριτηρίου AICc. Παραδόξως, μικρότερη τιμή στο κριτήριο AICc παρουσιάζει ένα μοντελο υψηλότερης τάξης. Γίνεται έλεγχος από το Minitab 14 και οι παράμετροι από το μοντέλο είναι στατιστικά σημαντικοί, κρίνοντας από τις τιμές των p-value. Τα αποτελέσματα του μοντέλου ARIMA(3,1,2) φαίνεται στην Εικόνα 4.3.4

```
Final Estimates of Parameters

Type       Coef  SE Coef       T      P
AR   1  -0,3288   0,1020   -3,22  0,002
AR   2  -0,6625   0,0777   -8,52  0,000
AR   3   0,4568   0,0943    4,84  0,000
MA   1  -0,6894   0,0694   -9,93  0,000
MA   2  -0,9809   0,0391  -25,12  0,000


Differencing: 1 regular difference
Number of observations:  Original series 100, after differencing 99
Residuals:    SS =  1023,86 (backforecasts excluded)
              MS =  10,89  DF = 94


Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square    5,7   14,8   20,0   30,3
DF              7     19     31     43
P-Value     0,577  0,732  0,936  0,928
```

Εικόνα 4.3.4
Αποτελέσματα μοντέλου ARIMA(3,1,2).

4.3.6 Διαγνωστικοί έλεγχοι υπολοίπων

Η αξιοπιστία του μοντέλου βασίζεται στο να έχουν τα υπόλοιπα ιδιότητες λευκού θορύβου. Δηλαδή, η μέση τιμή των υπολοίπων να είναι μηδενική και η διακύμανση σταθερή στο πέρας του χρόνου. Ξεκινώντας, λοιπόν, κατασκευάζονται περιληπτικά στο Διάγραμμα 4.3.5 το γράφημα των υπολοίπων στο χρόνο, το Q-Q plot, το ιστόγραμμα των υπολοίπων και το γράφημα των υπολοίπων σε σχέση με τις προσαρμοσμένες τιμές για το μοντελο ARIMA(3,1,2).



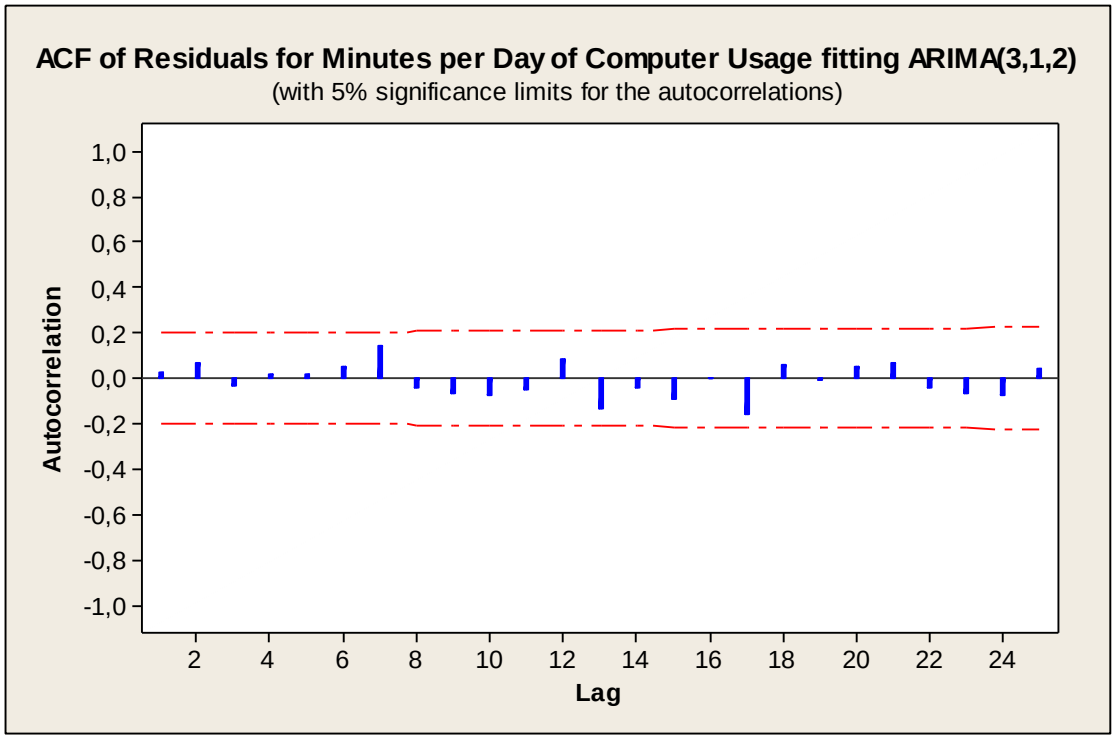
Διάγραμμα 4.3.5
Γράφημα των υπολοίπων στο χρόνο, Q-Q plot, ιστόγραμμα των υπολοίπων και το γράφημα των υπολοίπων σε σχέση με τις προσαρμοσμένες τιμές για το μοντελο ARIMA(3,1,2).

Γραφικά, όσον αφορά τη διακύμανση χρονικά από το γράφημα των υπολοίπων σε σχέση με τις προσαρμοσμένες τιμές παρατηρείται ομοσκεδαστικότητα, που επιβεβαιώνεται με το arch.test. που παρουσιάζεται στην Εικόνα 4.3.5. Παρατηρώντας την συνάρτηση αυτοσυσχέτισης (ACF) των υπολοίπων, που φαίνεται στο Διάγραμμα 4.3.6 παρακάτω, τα υπόλοιπα είναι ανεξάρτητα καθώς κανένας όρος δεν βρίσκονται εκτός ορίων. Αυτό αποδυναμώνεται με τους ελέγχους Box-Pierce και Ljung-Box, με

μηδενική υπόθεση ότι τα υπόλοιπα είναι ισόνομα και ανεξάρτητα, που φαίνονται στην Εικόνα 4.3.6. Επιπλέον, από το ιστόγραμμα των υπολοίπων και το Q-Q plot δεν φαίνεται καθαρά μια αποδεκτή εικόνα κανονικότητας. Δεν επιβεβαιώνεται ο ισχυρισμός της κανονικότητας των υπολοίπων με τους ελέγχους Jarque–Bera και Anderson-Darling από τα παρατιθέμενα αποτελέσματα στην Εικόνα 4.3.7.

```
> arch.test(fit)
ARCH heteroscedasticity test for residuals
alternative: heteroscedastic

Portmanteau-Q test:
     order   PQ p.value
[1,]     4 1.70   0.791
[2,]     8 5.73   0.678
[3,]    12 5.95   0.919
[4,]    16 7.58   0.960
[5,]    20 8.27   0.990
[6,]    24 8.61   0.998
Lagrange-Multiplier test:
     order    LM  p.value
[1,]     4 80.19 0.00e+00
[2,]     8 31.05 6.09e-05
[3,]    12 16.36 1.28e-01
[4,]    16  6.62 9.67e-01
[5,]    20  2.97 1.00e+00
[6,]    24  1.56 1.00e+00
```

Εικόνα 4.3.5
Αποτελέσματα ελέγχου ομοσκεδαστικότητας υπολοίπων για το μοντέλο ARIMA(3,1,2).



Διάγραμμα 4.3.6
Συνάρτηση αυτοσυσχέτισης (ACF) των υπολοίπων για το μοντελο ARIMA(3,1,2).

```
> Box.test(res, lag = 12, fitdf = 5)

        Box-Pierce test

data:  res
X-squared = 5.129, df = 7, p-value = 0.6442

> checkresiduals(fit)

        Ljung-Box test

data:  Residuals from ARIMA(3,1,2)
Q* = 4.8193, df = 5, p-value = 0.4383

Model df: 5.   Total lags used: 10
```

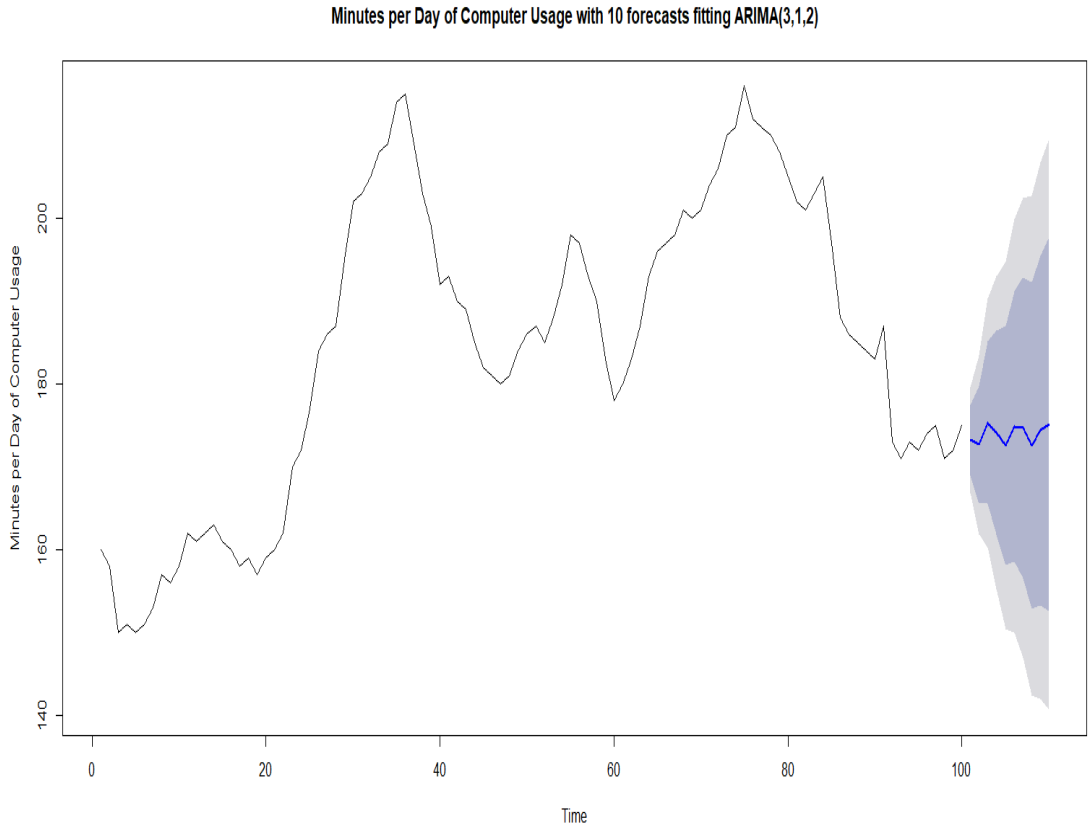
Εικόνα 4.3.6
Αποτελέσματα ελέγχων υπολοίπων Box-Pierce και Ljung-Box για το μοντελο ARIMA(3,1,2).

```
> ad.test(res)

        Anderson-Darling test of goodness-of-fit
        Null hypothesis: uniform distribution
        Parameters assumed to be fixed

data:  res
An = Inf, p-value = 6e-06

> jarque.bera.test(res)

        Jarque Bera Test

data:  res
X-squared = 46.808, df = 2, p-value = 6.851e-11
```

Εικόνα 4.3.7
Αποτελέσματα ελέγχων κανονικότητας υπολοίπων Jarque–Bera και Anderson-Darling για το μοντέλο ARIMA(3,1,2).

4.3.7 Προβλεπτική ικανότητα μοντέλου

Χρησιμοποιώντας την προγραμματιστική γλώσσα R και την εντολή forecast από την βιβλιοθήκη forecast κατασκευάζεται το γράφημα των πραγματικών τιμών της καθημερινής χρήσης ηλεκτρονικού υπολογιστή σε λεπτά μαζί με μοντέλο ARIMA(3,1,2) με τις επόμενες 10 προβλέψεις στο Διάγραμμα 4.3.7. Η σκούρα μπλε σκίαση υποδεικνύει το 80% διάστημα εμπιστοσύνης για προβλέψεις του μοντέλου ενώ η ανοιχτή μπλε σκίαση υποδεικνύει το 95% διάστημα εμπιστοσύνης για τις προβλέψεις του μοντέλου.



Διάγραμμα 4.3.7
Γραφική παράστασή των δεδομένων καθημερινής χρήσης ηλεκτρονικού υπολογιστή σε λεπτά μαζί με 10 προβλέψεις από το μοντελο ARIMA(3,1,2).

Συμπεράσματα

Σκοπός της παρούσας διπλωματικής είναι η χρήση των αυτοπαλίνδρομων μοντέλων ARIMA για την πρόβλεψη των μελλοντικών τιμών για τρεις διαφορετικές χρονοσειρές. Μέσω της ανάλυσης και της εφαρμογής τους εξάγονται κάποια ενδιαφέροντα συμπεράσματα.

- Και για τις τρεις εφαρμογές βρέθηκαν καταλληλά μοντέλα πρόβλεψης που τηρούν θεωρητικά, όλες τις προϋποθέσεις των μοντέλων ARIMA. Πρακτικό αποτέλεσμα θα υπήρχε, αν διατίθονταν οι πραγματικές μελλοντικές τιμές και υπολογίζονταν οι αποκλίσεις.
- Τα μοντέλα ARIMA προϋποθέτουν η μεταβλητή που μελετάται να παρουσιάζει σειριακή εξάρτηση μεταξύ των παρελθοντικών τιμών και μάλιστα γραμμική. Όσον αφορά τις μετρήσεις χημικής συγκέντρωσης τα μοντέλα ARIMA προτείνονται για προβλέψεις γιατί συνηθέστερα οι χημικές αντιδράσεις ακολουθούν γραμμικά μοντέλα.
- Κάθε χρονοσειρά μπορεί να περιγράφεται από ένα η περισσότερα μοντέλα ARIMA και να μην είναι εύκολη η διάκριση του βέλτιστου μοντέλου.
- Εάν τα υπόλοιπα των μοντέλων ARIMA δεν έχουν ιδιότητες λευκού θορύβου τα μοντέλα χάνουν την προβλεπτική τους ικανότητα. Αδυνατούν έτσι να προβλέψουν κάποια ακραία τιμή της μεταβλητής μελλοντικά καθώς δεν εμπεριέχουν την πληροφορία των σφαλμάτων.
- Επιπλέον, τα μοντέλα ARIMA μπορούν να δώσουν ικανοποιητικά αποτελέσματα σε περίπτωση μακροπροθέσμων μελλοντικών προβλέψεων, ενώ αδυνατούν να κάνουν βραχυπρόθεσμες προβλέψεις λόγω του ότι υπολογίζουν δεσμευμένες μέσες τιμές.

Μια μεγάλη ποικιλία μοντέλων υπάρχει για να ξεπεράσει κάποια από τα εμπόδια των μοντέλων ARIMA. Τα μοντέλα GARCH (Generalized Autoregressive Conditional Heteroskedasticity) είναι ικανά να περιγράψουν την μεταβλητότητα που συναντάται σε μια χρονοσειρά υπολογίζοντας δεσμευμένες διασπορές. Επίσης, τα μοντέλα VAR (Vector Autoregressive) είναι σε θέση να χρησιμοποιήσουν πληροφορία όχι μόνο από τις παρελθοντικές τιμές της μεταβλητής αλλά και από άλλες χρονοσειρές που δρουν παράλληλα. Τέλος, τα Τεχνητά Νευρωνικά Δίκτυα χρησιμοποιούν πληροφορία όχι μόνο από τις παρελθοντικές τιμές της μεταβλητής αλλά ταυτόχρονα από ενδογενείς και εξωγενείς μεταβλητές.

ΠΑΡΑΡΤΗΜΑ

Εύρεση δεδομένων

Για την εύρεση δεδομένων, διεξήχθη ενδελεχής μελέτη σε διάφορα βιβλία καθώς και σε βιβλιοθήκες του διαδικτύου. Τελικά, καταλήξαμε στο πακέτο της R γλώσσας προγραμματισμού Time Series Data Library (TSDL) που δημιουργήθηκε το 2018 από τον Rob Hyndman, Professor of Statistics at Monash University, Australia. Συγκεκριμένα, αποτελεί ένα δωρεάν πακέτο λογισμικού ανοιχτού κώδικα, με άδεια χρήσης GPL-3, οπότε είναι ελεύθερο προς ακαδημαϊκή και ερευνητική χρήση. Ακόμη, είναι ένα λογισμικό πακέτο που περιέχει 648 πραγματικές χρονοσειρές με πολυπληθή δεδομένα που αφορούν επιστήμες όπως: Οικολογία, Υδρολογία, Χημεία, Οικονομικά, Μετεωρολογία, Υπολογιστική Μηχανική κ.α.

```
> library(tsdl)
> library(forecast)
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
> tsdl
Time Series Data Library: 648 time series

                     Frequency
Subject               0.1 0.25   1   4   5   6  12  13  52 365 Total
  Agriculture           0    0  37   0   0   0   3   0   0   0    40
  Chemistry             0    0   8   0   0   0   0   0   0   0     8
  Computing             0    0   6   0   0   0   0   0   0   0     6
  Crime                 0    0   1   0   0   0   2   1   0   0     4
  Demography            1    0   9   2   0   0   3   0   0   2    17
  Ecology               0    0  23   0   0   0   0   0   0   0    23
  Finance               0    0  23   5   0   0  20   0   2   1    51
  Health                0    0   8   0   0   0   6   0   1   0    15
  Hydrology             0    0  42   0   0   0  78   1   0   6   127
  Industry              0    0   9   0   0   0   2   0   1   0    12
  Labour market         0    0   3   4   0   0  17   0   0   0    24
  Macroeconomic         0    0  18  33   0   0   5   0   0   0    56
  Meteorology           0    0  18   0   0   0  17   0   0  12    47
  Microeconomic         0    0  27   1   0   0   7   0   1   0    36
  Miscellaneous         0    0   4   0   1   1   3   0   1   0    10
  Physics               0    0  12   0   0   0   4   0   0   0    16
  Production            0    0   4  14   0   0  28   1   1   0    48
  Sales                 0    0  10   3   0   0  24   0   9   0    46
  Sport                 0    1   1   0   0   0   0   0   0   0     2
  Transport and tourism 0    0   1   1   0   0  12   0   0   0    14
  Tree-rings            0    0  34   0   0   0   1   0   0   0    35
  Utilities             0    0   2   1   0   0   8   0   0   0    11
  Total                 1    1 300  64   1   1 240   3  16  21   648
>
```

Για να εντοπιστούν οι χρονοσειρές αναλόγως την θεματολογία που αφορούν, δηλαδή μια από τις είκοσι τρεις κατηγορίες, χρησιμοποιείται η εντολή:

```
> for (i in 1:length(tsdl)) {
+   if ((attr(tsdl[[i]], "subject")) == "Chemistry") {
+     print(i)
+   }
+ }
[1] 162
[1] 165
[1] 166
[1] 167
[1] 171
[1] 172
[1] 495
[1] 508
>
```

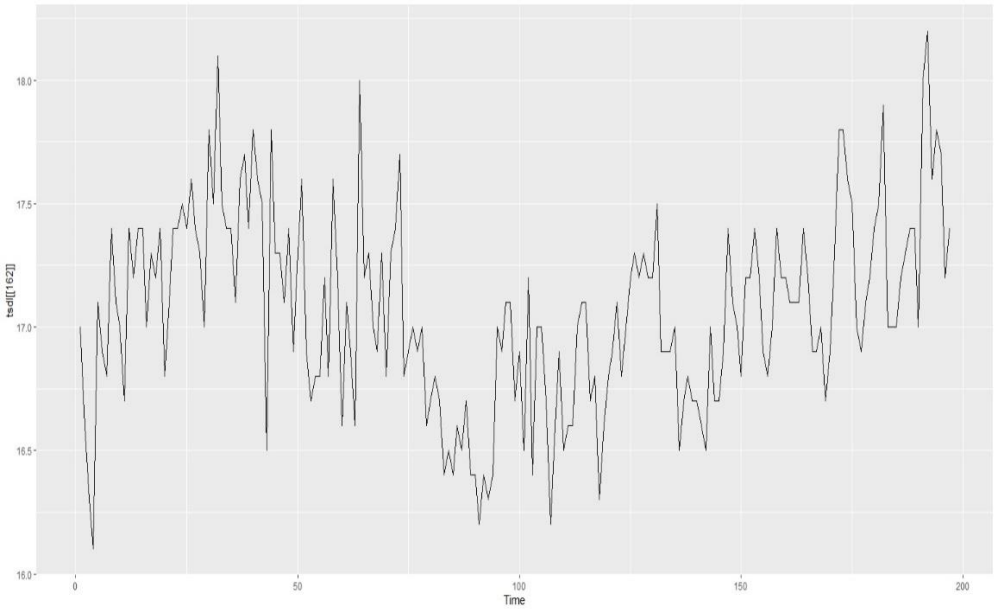
Και όπως φαίνεται παραπάνω παρουσιάζονται οι χρονοσειρές κατά αύξοντα αριθμό που αφορούν λόγω χάρη, την Χημεία. Έπειτα επιλέγεται μια από αυτές για να μελετηθεί η πηγή της και περισσότερες πληροφορίες για την περιγραφή της. Δηλαδή:

```
> attr(tsdl[[162]],'description')
[1] "Chemical concentration readings"
> attr(tsdl[[162]],'source')
[1] "Box & Jenkins (1976)"
>
```

Στην συνέχεια, για να εμφανιστεί το διάγραμμα της χρονοσειράς που επιλέχθηκε χρησιμοποιείται η εντολή:

```
> autoplot(tsdl[[162]])
```

Έτσι λοιπόν, εμφανίζεται παρακάτω και η γραφική παράσταση:



Έτσι λοιπόν χειροκίνητα, μετά από παρατήρηση των γραφικών παραστάσεων και μελέτη πολλών διαφορετικών χρονοσειρών, έγινε η επιλογή αυτών με κριτήριο την εντελώς διαφορετική θεματολογία μεταξύ τους και το δικό μας επιστημονικό ενδιαφέρον.

ΒΙΒΛΙΟΓΡΑΦΙΑ

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

Aptech.com., 2020. *Introduction To The Fundamentals Of Time Series Data And Analysis Aptech*. [online] Available at: https://www.aptech.com/blog/introduction-to-the-fundamentals-of-time-series-data-and-analysis/.

Box, G. & Jenkins G., 1976. *Time Series Analysis*. Englewood Cliffs, N.J.: Prentice Hall.

Box, G. & Pierce, D., 1970. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *Journal of the American Statistical Association*, 65(332), pp.1509-1526.

Brillinger, D.R., 1981. *Time Series: Data Analysis and Theory*. Siam.

Chatfield, C., 2003. *The Analysis of Time Series*. 6th ed. London: Chapman and Hall.

Caroni, C. & Karioti, V., 2004. Detecting an innovative outlier in a set of time series. *Computational Statistics and Data Analysis*, 46, pp. 561-570.

Dickey, D. A. & Fuller, W.A, 1997. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366): pp. 427-431.

Engle, R.F., 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation, *Econometrica*, 50, pp.987-1008.

Hamilton, J.D., 1994. *Time Series Analysis*, Princeton University Press.

Holmes, E., Scheuerell, M. & Ward E., 2020. *Applied time series analysis for fisheries and environmental data*. NOAA Fisheries, Northwest Fisheries Science Center, Seattle.

Hipel, K., McLeod, A., Panu, U. & Singh, V., 1994. *Stochastic And Statistical Methods In Hydrology And Environmental Engineering*. Dordrecht: Springer Netherlands.

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

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.

Montgomery, D., 1976. Computers: Reassuring, but Dispensable. *Science*, 193(4256), pp.270-271.

Montgomery, D., Jennings, C. & Kulahci, M., 2015. *Introduction to Time Series Analysis and Forecasting*. New York, NY: John Wiley & Sons.

Peña, D., Tiao, G. & Tsay, R., 2001. *A Course in Time Series Analysis*. New York: J.

Pfaff, B., 2008. *Analysis of Integrated and Cointegrated Time Series with R*. Kronberg im Taunus: Springer.

Yao, J. & Herbert, J., 2009. Financial time-series analysis with rough sets. *Applied Soft Computing*, 9(3), pp.1000-1007.

Αγιακόγλου, Χ. & Οικονόμου, Γ., 2004. *Μέθοδοι προβλέψεων και ανάλυσης αποφάσεων*, Β' έκδοση, εκδόσεις Γ. Μπένου, Αθήνα.

Καρώνη Χ. & Οικονόμου, Π., 2017. Στατιστικά Μοντέλα Παλινδρόμησης, 2η έκδοση, Αθήνα: Συμεών.

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

Κουγιουμτζής, Δ., 2005. *Γραμμική ανάλυση χρονοσειρών*, πανεπιστημιακές σημειώσεις.