ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΝΑΛΥΣΗ ΧΡΟΝΟΣΕΙΡΩΝ ΜΕ ΕΦΑΡΜΟΓΗ ΣΕ ΒΙΟΪΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ

ΕΛΠΙΔΑ ΛΟΪΖΟΥ

ΕΠΙΒΛΕΠΟΥΣΑ: ΒΟΝΤΑ ΦΙΛΙΑ

ΑΝΑΠΛΗΡΩΤΡΙΑ ΚΑΘΗΓΗΤΡΙΑ Ε.Μ.Π.

Επιτροπή Καθηγητών: Βόντα Φιλία, Αλέξανδρος Καραγρηγορίου, Καρώνη Χρυσηίς

Αθήνα, Φεβρουάριος 2020

ΠΕΡΙΕΧΟΜΕΝΑ

II. Μέρος II - Εφαρμογή

Κατάλογος Διαγραμμάτων

Κατάλογος Πινάκων

Περίληψη

Η παρούσα διπλωματική εργασία μελετά την ανάλυση και πρόβλεψη χρονοσειρών. Στα πρώτα κεφάλαια γίνεται ενδελεχής αναφορά στο θεωρητικό κομμάτι με σκοπό την κατανόηση βασικών εννοιών και την κατηγοριοποίηση των χρονοσειρών για περαιτέρω ανάλυσή τους. Έπεται η θεωρητική μελέτη των πιθανών υποδειγμάτων μίας χρονολογικής σειράς, με τη βοήθεια των οποίων παράγονται οι εκτιμούμενες τιμές, δηλαδή γίνεται η πρόβλεψη των χρονοσειρών. Τέλος, παρουσιάζεται μία εφαρμογή στην R σε πραγματικά δεδομένα επιδημιολογικής φύσης τα οποία έχουν διατεθεί από το Τμήμα Επιδημιολογικής Επιτήρησης και Παρέμβασης του ΚΕ.ΕΛ.Π.ΝΟ.

Abstract

This dissertation deals with the analysis and forecasting of time series and focuses on the understanding of the past as long as investigating the future. In the first chapters we investigate theoretically the subject with the goal of understading the basic definitions as wel as the different cathegories of time series in order to analyse them further. The main characteristics of a time series such as trend, seasonality, variation etc. are examined as well. In the sequel, we perform an extensive study of the time series modeling and their forecasting, with the main goal being to predict the future values of a time series. The final chapter of the thesis is devoted to the analysis of a real data set based on R. The data set consists of a series of epidemiological data which was made available by the Center for Disease Control and Prevention KEELPNO.

ΜΕΡΟΣ Ι

Θεωρητικό Μερος

Ε Υ Χ Α Ρ Ι Σ Τ Ι Ε Σ

Η παρούσα εργασία με τίτλο «Ανάλυση χρονοσειρών με εφαρμογή σε βιοϊατρικά δεδομένα» αποτελεί διπλωματική εργασία στο πλαίσιο του προπτυχιακού προγράμματος της σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών του Εθνικού Μετσόβιου Πολυτεχνείου.

Στο σημείο αυτό θα ήθελα να εκφράσω τις ειλικρινείς και θερμές ευχαριστίες μου για την επιβλέπουσα καθηγήτριά μου Κυρία Βόντα Φιλία που με εμπιστεύτηκε για την εργασία και μου παρείχε όλη την απαραίτητη βοήθεια για την εκπόνηση αυτής.

Ευχαριστώ επίσης τον κ. Α. Καραγρηγορίου, Καθηγητή στο Πανεπιστήμιο Αιγαίου και την κ. Χ. Καρώνη, Καθηγήτρια στο Εθνικό Μετσόβιο Πολυτεχνείο για την συμμετοχή τους στην τριμελή επιτροπή.

Σε συνέχεια θα ήθελα να αναφέρω ότι η πηγή προέλευσης των δεδομένων είναι το Σύστημα Παρατηρητών Νοσηρότητας Πρωτοβάθμιας Φροντίδας Υγείας (Sentinel) του Τμήματος Επιδημιολογικής Επιτήρησης και Παρέμβασης του ΚΕ.ΕΛ.Π.ΝΟ. Εκφράζουμε τις ευχαριστίες μας προς το ΚΕ.ΕΛ.Π.ΝΟ για την παραχώρηση των δεδομένων καθώς και τη Δρ. Χριστίνα Παρπούλα για την πολύτιμη βοήθειά της σχετικά με την παραχώρηση των δεδομένων.

Τέλος από αυτές τις ευχαριστίες δε θα μπορούσαν να λείπουν οι γονείς μου, Αγγελική και Παναγιώτης, για την υποστήριξη τους καθ'όλη τη διάρκεια των σπουδών μου.

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Κύριος Σκοπός της Εργασίας

Η πρόβλεψη είναι μία από τις σημαντικότερες προκλήσεις σε μία επιχείρηση και εν γένει για οποιονδήποτε επιστήμονα που καλείται να εξάγει συμπεράσματα για το μέλλον βάσει δεδομένα από το παρελθόν με σκοπό τη σωστή και έγκαιρη λήψη αποφάσεων. Αυτή η δυνατότητα προσφέρεται μέσα από την ανάλυση των χρονολογικών σειρών και την πρόβλεψή τους μέσω κατάλληλων μοντέλων, ένα αντικείμενο που θα αναλυθεί και θα εξεταστεί πλήρως στην παρούσα εργασία.

Οι χρονολογικές σειρές αποτελούνται από μία σειρά από παρατηρήσεις που λαμβάνονται σε διαδοχικά ισαπέχουσες ή μη χρονικές τιμές. Η ανάλυση των χρονολογικών σειρών περιλαμβάνει μεθόδους για την ανάλυση δεδομένων χρονοσειρών προκειμένου να εξαχθούν σημαντικά στατιστικά στοιχεία και άλλα χαρακτηριστικά των δεδομένων. Η πρόβλεψη χρονολογικών σειρών γίνεται μέσω της χρήσης ενός μοντέλου το οποίο κάνει δυνατή την πρόβλεψη μελλοντικών τιμών βάσει προηγούμενων τιμών.

Κύριος σκοπός της εργασίας είναι ο αναγνώστης να κατανοήσει τη χρησιμότητα των χρονολογικών σειρών και να είναι σε θέση να ανταποκριθεί στη μελέτη αυτών. Η εργασία προσφέρει μία μελέτη και την εις βάθος επεξεργασία των δεδομένων που συνιστούν μία χρονολογική σειρά. Αναλύονται και περιγράφονται ενδελεχώς τα κύρια χαρακτηριστικά μίας χρονοσειράς, όπως η τάση, η εποχικότητα, η κυκλικότητα, με σκοπό την κατανόησή της, για περαιτέρω επεξεργασία. Γίνεται μία αναλυτική αναφορά στα βασικά υποδείγματα χρονολογικών σειρών AR, MA, ARMA, ARIMA, SARIMA, με τη βοήθεια των οποίων διευρενάται η πρόβλεψη των μελλοντικών τιμών μίας χρονοσειράς. Τέλος, αναλύεται στην R ένα παράδειγμα επιδημιολογικών δεδομένων ως εφαρμογή του θεωρητικού μέρους που προηγήθηκε και εξάγονται συμπεράσματα.

ΚΕΦΑΛΑΙΟ 2

Ανάλυση Χρονοσειρών

2.1 Εισαγωγή στις Χρονοσειρές

Η ακολουθία των παρατηρήσεων που αλλάζει τιμές σε ορισμένες χρονικές στιγμές ή περιόδους που ισαπέχουν μεταξύ τους ονομάζεται *Χρονοσειρά* ή *Χρονολογική Σειρά (Time Series)*. Η επιλογή των παρατηρήσεων ενός μεγέθους γίνεται συνήθως με συγκεκριμένο χρονικό βήμα που ονομάζεται *χρόνος δειγματοληψίας (Sampling Time)*. Για παράδειγμα, έχουμε τη μέτρηση της θερμοκρασίας κάθε ώρα ή τη μέση θερμοκρασία κάθε ημέρας, τη μεταβολή της συναλλαγματικής αξίας ανά λεπτό ή την τιμή μιας μετοχής στο κλείσιμο της ημέρας.

Τα μοντέλα χρονοσειρών που χρησιμοποιούνται είναι κατά βάση στοχαστικά μοντέλα. Σε πολλά προβλήματα καλούμαστε να αναλύσουμε ένα χρονοεξαρτώμενο φαινόμενο στο οποίο υπάρχουν πολλοί άγνωστοι παράγοντες που αποτρέπουν τον ορισμό ενός προσδιοριστικού μοντέλου (βλέπε [1]). Συστήματα τα οποία εξελίσσονται χρονικά κατά τρόπο που περιέχει, σε μικρό ή μεγάλο βαθμό, *τυχαιότητα (stochasticity, randomness)* και όχι κατά τρόπο *προσδιοριστικό (deterministic)* αποτελούν μοντέλα *Στοχαστικών Διαδικασιών (Stochastic Processes)*.

Με τον όρο χρονοσειρά λοιπόν, εννοούμε μία ακολουθία $\{X_{t,}\ t\ \in T\}$ όπου Τ είναι η χρονική περίοδος που εξελίσσεται το φαινόμενο που μελετάμε ή $\{X_{t=1}^{n}\} = \{X_1, X_2, .., X_n\}$ για κάποια χρονική περίοδο n σε μονάδες δειγματοληψίας. Για κάθε χρονική στιγμή t θεωρούμε τις παρατηρήσεις $x_1, x_2, \dots, x_T$ οτι είναι συγκεκριμένες τιμές ή συγκεκριμένες πραγματοποιήσεις των τυχαίων μεταβλητών $X_1, X_2, .., X_T$ και ότι επιπλέον οι τυχαίες μεταβλητές αυτές $X_1, X_2, .., X_T$ είναι μέρος μιας άπειρης σειράς τυχαίων μεταβλητών η οποία ονομάζεται στοχαστική διαδικασία.

Παραδείγματα τέτοιων χρονοσειρών είναι:

(i) Οι ημερήσιες, αεροπορικές και οδικές, αφίξεις τουριστών στη χώρα μας $\{x_t\}$ με $t = 1,2, \dots$

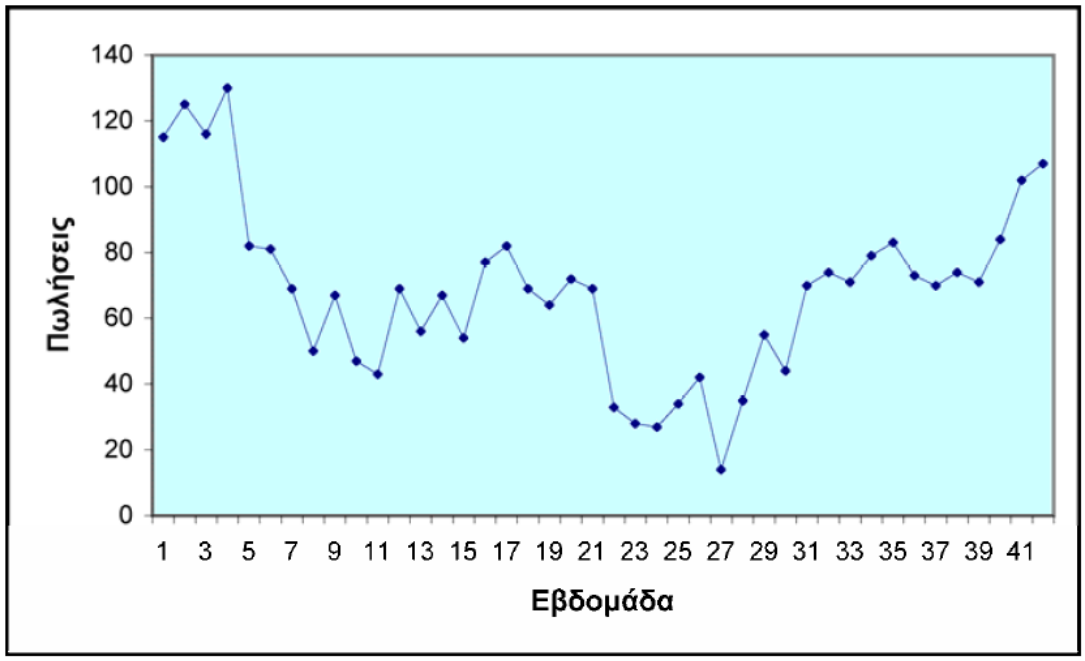
(ii) Ο αριθμός $\{x_t\}$ πελατών μέσα σε ένα πολυκατάστημα κατά τη χρονική στιγμή *t,* με *t* ∈ [0,*T*]

(iii) Οι εβδομαδιαίες πωλήσεις $\{x_t\}$ ενός προϊόντος στο χρονικό διάστημα [0,t] με t ≥ 0, (Διάγραμμα 2.1)

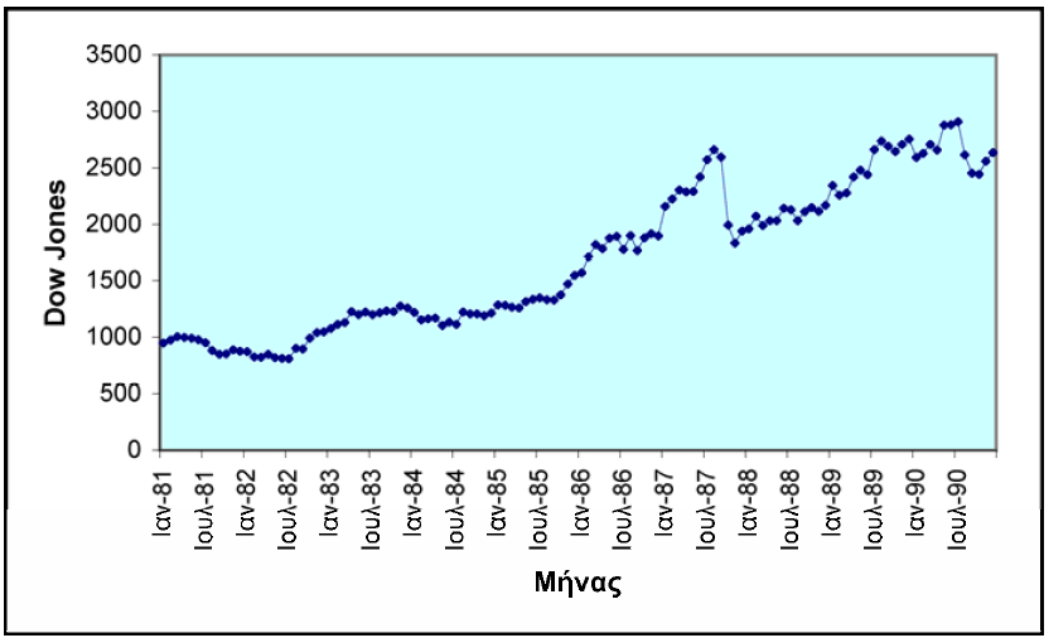
(iv) Η ημερήσια κατανάλωση ηλεκτρικού ρεύματος καθώς και η ημερήσια κατανάλωση ύδατος , $\{x_t\}$ και $\{y_t\}$ αντίστοιχα, σε μία μεγάλη γεωγραφική περιοχή της χώρας με $t = 1,2, \ldots$

(v) Οι οικονομικές χρονοσειρές, όπως οι μέσες μηνιαίες τιμές κλεισίματος του δείκτη Dow Jones ή το ετήσιο ακαθάριστο εθνικό προϊόν, $\{x_t\}$ και $\{y_t\}$ αντίστοιχα, με $t = 1,2, \ldots.$, (Διάγραμμα 2.2)

(vi) Οι μετεωρολογικές χρονοσειρές, όπως η θερμοκρασία περιβάλλοντος και ατμοσφαιρική πίεση, $\{x_t\}$ και $\{y_t\}$ αντίστοιχα, σε συγκεκριμένη γεωγραφική περιοχή με γεωγραφικές συντεταγμένες *(l,a,h)* κατά τη χρονική στιγμή *t.* Εδώ η χρησιμοποιούμενη παράμετρος *t* είναι περισσότερο σύνθετη και πιο συγκεκριμένα *t=(l,a,h).*



Διάγραμμα 2.1 Η χρονοσειρά των πωλήσεων ενός προϊόντος τις τελευταίες 42 εβδομάδες.



Διάγραμμα 2.2 Μέσες μηνιαίες τιμές του δείκτη Dow Jones για τα έτη 1982 έως 1990.

Δεδομένων των παραπάνω παραδειγμάτων, διαπιστώνεται ότι οι χρονοσειρές μπορούν να αφορούν διακριτά μεγέθη $\{x_t\}$ σε διακριτό χρόνο t , παράδειγμα (i) , διακριτά μεγέθη $\{x_t\}$ σε συνεχή χρόνο t, παραδείγματα (ii) και (iii), συνεχή μεγέθη $\{x_t\}$ σε διακριτό χρόνο t, παραδείγματα (iv) και (v) και συνεχή μεγέθη $\{x_t\}$ σε συνεχή χρόνο t, παράδειγμα (vi).

Οι χρονοσειρές βρίσκουν εφαρμογή σε πολλούς τομείς, μεταξύ άλλων, τις Οικονομικές Επιστήμες, την Ιατρική, την Κοινωνιολογία κ.α. διότι η ανάλυση των χρονοσειρών επιτρέπει την πρόβλεψη των τιμών του μεγέθους που παρατηρείται. Οι χρονοσειρές καταγράφουν τις παρελθοντικές τιμές ενός μεγέθους επιτρέποντας έτσι την εκτίμηση των μελλοντικών τιμών. Βέβαια, οι χρονολογικές σειρές δεν χρησιμοποιούνται αποκλειστικά και μόνο ως εργαλείο πρόγνωσης, άλλα είναι ένα πολύ χρήσιμο εργαλείο για την ανάλυση και καλύτερη κατανόηση της συμπεριφοράς του ίδιου του φαινομένου, αφού καταγράφεται η ιστορία του.

2.2 Στόχοι της Ανάλυσης Χρονοσειρών

Ο κυριότερος στόχος της Ανάλυσης Χρονοσειρών είναι η ανεύρεση των βασικών χαρακτηριστικών της χρονοσειράς και η περιγραφή της εσωτερικής δομής της.

Το πρώτο στάδιο για την ανάλυση μίας χρονοσειράς είναι η γραφική απεικόνιση των τιμών της σε συνάρτηση με το χρόνο. Με τον τρόπο αυτό τα βασικά χαρακτηριστικά της χρονοσειράς εμφανίζονται ως γραφικά μοτίβα όπως η τάση, η κυκλικότητα, η

εποχικότητα και οι ακραίες τιμές. Η αναγνώριση των χαρακτηριστικών αυτών καθορίζει και το είδος της ανάλυσης που θα ακολουθηθεί, επιτρέποντας την επιλογή κατάλληλου μοντέλου.

Τέλος στόχος της Ανάλυσης Χρονοσειρών αποτελεί η πρόβλεψη των μελλοντικών τιμών της χρονοσειράς και ο προσδιορισμός της αβεβαιότητας αυτών των προβλέψεων. Στόχος της πρόβλεψης των χρονοσειρών είναι όσο το δυνατό περισσότερη ακρίβεια στις προβλεπόμενες τιμές συγκριτικά με τις πραγματικές μελλοντικές τιμές έτσι ώστε να συμβάλλουν στη σωστή και έγκαιρη λήψη αποφάσεων.

2.3 Βασικά Χαρακτηριστικά και Κατηγορίες Χρονοσειρών

Όπως αναφέρθηκε και παραπάνω μία χρονολογική σειρά αναλύεται στα επιμέρους χαρακτηριστικά της. Με βάση αυτά τα χαρακτηριστικά οι χρονοσειρές κατηγοριοποιούνται ανάλογα.

Για να μελετήσει κανείς μία χρονοσειρά πρέπει να φτιάξει το γράφημα των τιμών της στο πεδίο του χρόνου. Είναι το πρώτο βήμα στην ανάλυση της χρονοσειράς, καθώς έτσι παρατηρούνται τα βασικά χαρακτηριστικά της με «γυμνό μάτι», όπως η τάση, η κυκλικότητα, η εποχικότητα και οι ακραίες τιμές.

2.3.1 Μονοδιάστατες και Πολυδιάστατες Χρονοσειρές

Οι χρονοσειρές σύμφωνα με το πλήθος των μεγεθών που καταγράφουν διαχωρίζονται σε μονοδιάστατες και πολυδιάστατες χρονοσειρές.

Μονοδιάστες Χρονοσειρές (Univariate Time Series)

Οι μονοδιάστατες Χρονοσειρές καταγράφουν τιμές ενός μεγέθους. Μια μονοδιάστατη χρονοσειρά αποτελεί μία ακολουθία τιμών της ίδιας μεταβλητής στο πέρασμα του χρόνου. Μία τυπική μορφή μίας μονοδιάστατης μεταβλητής είναι ως εξής:

$$\{(t_1, data\ value_1)\ , (t_2, data\ value_2)\ , \dots\ , (t_n, data\ value_n)\}$$

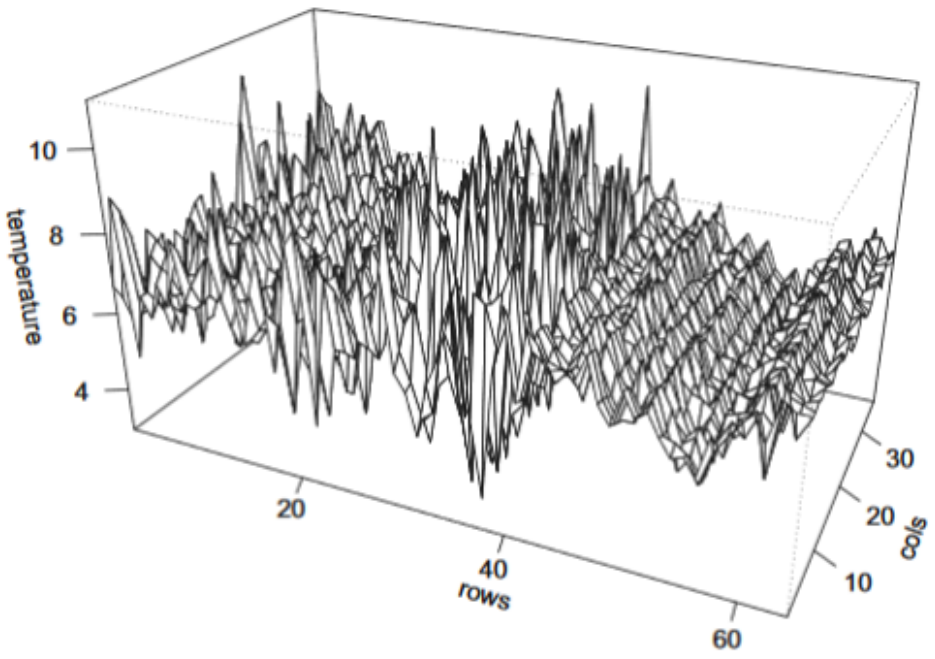
για $i = 1,2, \dots n$ όπου $data\ value$ είναι η τιμή που αντιστοιχεί σε κάθε χρονική στιγμή t_i.

Πολυδιάστατες Χρονοσειρές (Multivariate Time Series)

Όταν μια χρονοσειρά περιέχει παραπάνω από μία μεταβλητές, τότε πρόκειται για πολυδιάστατη χρονοσειρά. Με τη χρήση πολυδιάσταστων χρονοσειρών έχουμε τη

δυνατότητα ταυτόχρονης παρατήρησης πολλών μεγεθών για το ίδιο σύστημα. Σε αυτή την περίπτωση, είναι σύνηθες, οι μεταβλητές μεταξύ τους να αλληλοσυσχετίζονται με την πάροδο του χρόνου. Αν μία μεταβλητή Χ είναι χρήσιμη για την πρόγνωση μελλοντικών τιμών της μεταβλητής Υ τότε η πολυδιάστατη μεταβλητή είναι ομογενής *(homogeneous)*, αλλιώς είναι ετερογενής *(heterogeneous)*. Στις ομογενείς πολυδιάστες χρονοσειρές, όποια αλλαγή προκληθεί σε ένα στοιχείο των παρατηρήσεων της μίας μεταβλητής προκύπτει αντίστοιχη αλλαγή στις παρατηρήσεις των άλλων μεταβλητών που σχετίζονται με το φαινόμενο που παρατηρούμε.

Στην περίπτωση των πολυδιάστατων χρονοσειρών είναι χρήσιμο να εισάγουμε την έννοια των διανυσματικών χρονοσειρών $x_t = (x_{t1}, x_{t2}, \dots, x_{tp})'$, που περιέχει ως συστατικά p μονοδιάστατες χρονοσειρές. Θεωρούμε το διάνυσμα στήλη $p \times 1$ των παρατηρούμενων p χρονοσειρών ως x_t. Το διάνυσμα γραμμή x_t' είναι ο ανάστροφός του (βλέπε [1])



Διάγραμμα 2.3 Δισδιάστατη χρονοσειρά μέτρησης της θερμοκρασίας. *Πηγή: [1] Robert H. Shumway, David S. Stoffer: Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics) Fourth (4th) Edition*

2.3.2 Στατιστικά Μεγέθη Χρονοσειράς

Μία χρονοσειρά, όπως έχει αναφερθεί παραπάνω, είναι μία στοχαστική διαδικασία, δηλαδή κάθε παρατήρησή της αποτελεί μία τυχαία μεταβλητή. Τα βασικότερα χαρακτηριστικά μίας τυχαίας μεταβλητής X είναι η μέση τιμή $\mu = E[X]$, η διασπορά

$\sigma^2 = V[X]$, και όταν υπάρχουν ζεύγη τυχαίων μεταβλητών, η μικτή ροπή 2ας τάξης, δηλαδή η συνδιακύμανση $\sigma_{XY} = Cov(X, Y)$. Παρακάτω παρουσιάζονται τα κυριότερα στατιστικά μεγέθη μίας στατιστικής διαδικασίας.

Μέση Τιμή

Η μέση τιμή *(mean value)* μίας χρονοσειράς $\{X_t : t \in T\}$ δίνεται από την ακόλουθη συνάρτηση (βλέπε [1]) για όλη την παράγραφο :

$$\mu = E[X_t]$$

Η μέση τιμή δηλώνει την αναμενόμενη τιμή ή αλλιώς την προσδοκώμενη τιμή μίας διαδικασίας σε χρόνο t.

Συνδιακύμανση

Η συνάρτηση συνδιακύμανσης είναι μέτρο του βαθμού συσχετίσεως δύο μεταβλητών και ορίζεται ως

$$\gamma_{s,t} = Cov(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)]$$

Η συνάρτηση συνδιακύμανσης υπολογίζει τη γραμμική εξάρτηση μεταξύ δύο τιμών της χρονοσειράς που παρατηρούνται σε διαφορετικούς χρόνους.

Διακύμανση

Στην περίπτωση όπου $s = t$ τότε πρόκειται για τη διακύμανση που ορίζεται ως

$$\sigma^2 = Var(X_t) = E[(x_t - \mu_t)^2]$$

Οι χρονοσειρές λοιπόν σύμφωνα με τα στατιστικά μεγέθη τους διαχωρίζονται σε στάσιμες ή μη στάσιμες.

2.4 Στασιμότητα

Στην ανάλυση των χρονοσειρών η στασιμότητα αποτελεί μία βασική προϋπόθεση. Πρώτος στόχος είναι η αναγνώριση μίας χρονοσειράς ως προς την στασιμότητα έτσι ώστε να ακολουθήσει σχετική διερεύνηση , όπως θα αναλύσουμε παρακάτω.

2.4.1 Στάσιμες Χρονοσειρές

Στάσιμες Χρονοσειρές *(Stationary Time Series)* είναι οι χρονοσειρές στις οποίες τα στατιστικά μέτρα όπως η μέση τιμή, η διασπορά, η μικτή ροπή 2ας τάξης, δηλαδή η συνδιακύμανση μένουν αναλλοίωτα στο χρόνο.

Όταν όλα τα στατιστικά μέτρα παραμένουν αναλλοίωτα στο χρόνο τότε μιλάμε για *αυστηρή στασιμότητα.*

Ορισμός Αυστηρής Στασιμότητας

Η χρονοσειρά $\{X_t: t \in T\}$ ονομάζεται αυστηρά στάσιμη όταν $\forall\, n \in N, t_i \in T\ (i = 1, \dots, n)$ και $h \in T$ ισχύει η παρακάτω σχέση ισοδυναμίας :

$(X_{t_1}, \dots, X_{t_n})$ έχει την ίδια κατανομή με $(X_{t_1+h}, \dots, X_{t_n+h})$ (2.4.1)

Συνεπώς για μία αυστηρά στάσιμη χρονοσειρά $\{X_t: t \in T\}$ ισχύει :

i. $\mu = E[X_t],\ t \in T$, ανεξάρτητη του t
ii. $\gamma(h) = Cov(X_t, X_{t+h}),\ t, h \in T$, ανεξάρτητη του t
iii. Ολες οι ροπές μεγαλύτερης τάξης είναι αναλλοίωτες ως προς το χρόνο t

Οι παραπάνω δύο συνθήκες μόνο είναι προφανώς ασθενέστερες από τη συνθήκη (2.4.1). Η απαίτηση να ισχύουν οι ως άνω δύο πρώτες συνθήκες, μαζί με την απαίτηση της πεπερασμένης διασποράς σ^2, μας δίνουν την ιδιότητα της ασθενούς, ή υπό ευρεία έννοια στασιμότητας. Επομένως έχουμε τον ορισμό:

Ορισμός Ασθενούς Στασιμότητας

Η χρονοσειρά $\{X_t: t \in T\}$ ονομάζεται ασθενώς, ή υπό ευρεία έννοια, στάσιμη όταν ικανοποιούνται και οι τρεις παρακάτω συνθήκες:

i. $\mu = E[X_t],\ t \in T$, ανεξάρτητη του t
ii. $\gamma(h) = Cov(X_t, X_{t+h}),\ t, h \in T$, ανεξάρτητη του t
iii. $\sigma^2 = Var(X_t)$, ανεξάρτητη του t (ειδική περίπτωση της συνδιασποράς)

Για τις στάσιμες χρονοσειρές ισχύει ότι η διασπορά $\sigma^2 = V[X_t] = \gamma(0) \geq |\gamma(h)| \quad \forall\, h \in T.$

(βλέπε[18]).

2.4.2 Μη Στάσιμες Χρονοσειρές

Η μη-στασιμότητα αποτελεί σοβαρό πρόβλημα στην ανάλυση χρονοσειρών και ειδικότερα όταν προσπαθούμε να κάνουμε προβλέψεις. Σαν στοιχεία ύπαρξης μη-στασιμότητας είναι κυρίως η ύπαρξη τάσης, εποχικότητας, κυκλικότητας, και ακραίων τιμών.

Τάση (Trend)

Η μακροχρόνια ομαλή κεντρική κίνηση, την οποία ακολουθεί η χρονολογική σειρά κατά τη διάρκεια ολόκληρης της χρονικής περιόδου ονομάζεται *Τάση (Trend).* Έτσι η τάση μπορεί να είναι ανοδική, καθοδική ή σύνθετη. Η τάση θεωρείται ανύπαρκτη εάν η κεντρική ομαλή κίνηση της χρονοσειράς ακολουθεί νοητή ευθεία παράλληλη στον άξονα του χρόνου (βλέπε [13]). Η τάση μπορεί να παρασταθεί ως μία απλή γραμμική συνάρτηση με το χρόνο ή ίσως και πολυωνυμική συνάρτηση του χρόνου ή εκθετική.

Γενικά όταν η τάση σε μια χρονοσειρά μπορεί να περιγραφεί από κάποια γνωστή ή εκτιμώμενη συνάρτηση του χρόνου, $\mu(t) = f(t)$, ονομάζεται *καθοριστική τάση (deterministic trend)*. Μπορεί όμως η τάση σε μια χρονοσειρά να μην είναι δυνατόν να περιγραφεί από μια γνωστή (παραμετρική) συνάρτηση του χρόνου, να παρουσιάζει δηλαδή αργές μεταβολές με το χρόνο αλλά όχι με κάποιο καθοριστικό τρόπο. Αυτή η τάση λέγεται *στοχαστική (stochastic trend)* (βλέπε [19]).

Στο παρακάτω γράφημα παρουσιάζεται η ετήσια παραγωγή των Η.Π.Α. για το μπλε και gorgonzola τυρί, όπου παρατηρείται ο τετραπλασιασμός της παραγωγής από το 1950 έως το 1997. Η γραμμική τάση σ' αυτό το γράφημα απεικονίζεται με σταθερή θετική κλίση παρά την μεταβολή της παραγωγής χρόνο με το χρόνο (βλέπε [5]).

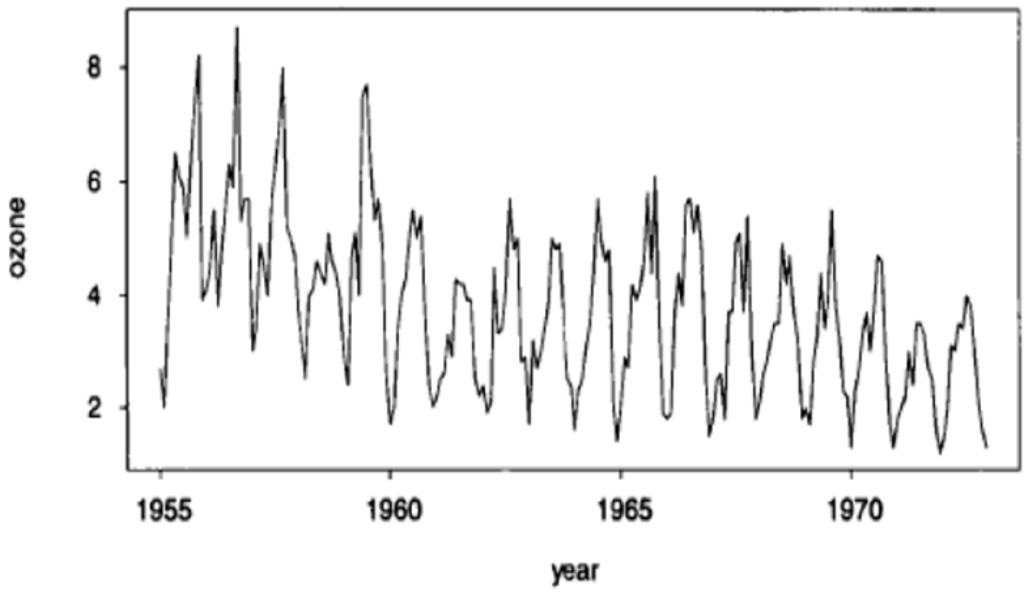
Διάγραμμα 2.4 Η ετήσια παραγωγή των Η.Π.Α. για το μπλε και gorgonzola τυρί. *Πηγή: [5] Douglas C. Montgomery, "Introduction to Time Series Analysis and Forecasting", 2008, Willey*

Εποχικότητα (Seasonality)

Οι χρονοσειρές των οποίων τα δεδομένα επαναλαμβάνονται με τον ίδιο περίπου τρόπο σε ορισμένα χρονικά διαστήματα παρουσιάζουν εποχικότητα. Είναι μια

περιοδική διακύμανση η οποία έχει σταθερό και μικρότερο ή ίσο μήκος ενός έτους. Εφόσον, η εποχική διακύμανση παρουσιάζεται με συστηματικό τρόπο, είναι ένα χαρακτηριστικό εύκολα οπτικά αναγνωρίσιμο που μπορεί να μετρηθεί και να απομονωθεί, ώστε να μην επηρεάζει τα δεδομένα μας. Η νέα χρονοσειρά που προκύπτει ονομάζεται αποεποχικοποιημένη χρονοσειρά.

Στο επόμενο σχήμα, όπως λήφθηκε από το βιβλίο των Daniel Pena, George Tiao και Ruey Tsay (βλέπε [6]), φαίνονται οι μέσες μηνιαίες τιμές του όζοντος στο κέντρο του Λος Άντζελες από το 1955 μέχρι το 1972. Παρατηρείται ότι, το ατμοσφαιρικό όζον που είναι ένας δείκτης της ατμοσφαιρικής ρύπανσης παρουσιάζει έντονη εποχικότητα, η οποία είναι υψηλή κατά τους καλοκαιρινούς μήνες και χαμηλή τον χειμώνα.



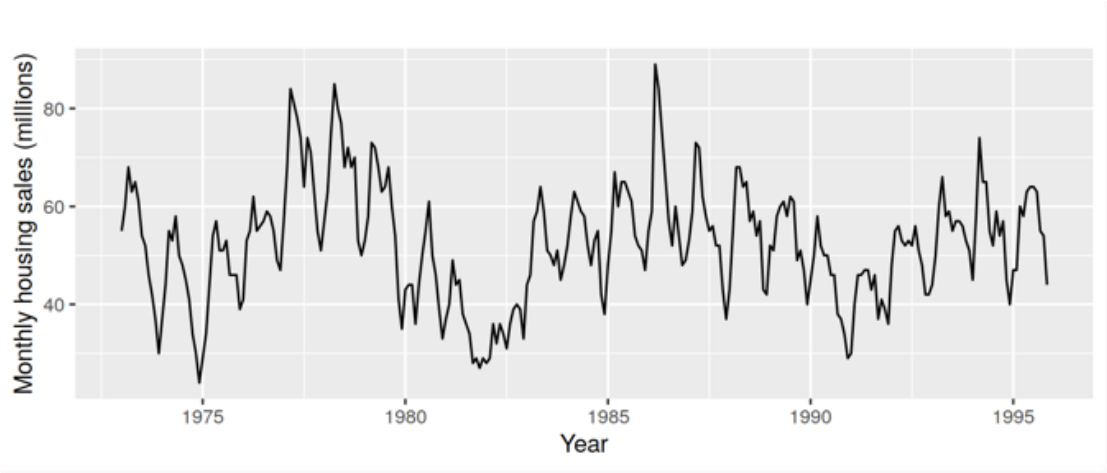
Διάγραμμα 2.5 Μηνιαίες ενδείξεις όζοντος στο Λος Άντζελες. *Πηγή: [6] A Course in Time Series Analysis,: Daniel Pena, George C. Tiao and Ruey S. Tsay (Eds.), John Wiley, New York, 2001*

Κυκλικότητα (Cyclic)

Μία χρονοσειρά παρουσιάζει κυκλικότητα στη συμπεριφορά των τιμών της όταν παρατηρούνται διακυμάνσεις με ανοδικές και καθοδικές φάσεις που επαναλαμβάνονται διαδοχικά γύρω από τη γραμμή τάσης. Η κυκλική συμπεριφορά ορίζεται από δύο κάτω σημεία καμπής (trough) και ένα άνω σημείο καμπής (peak) το οποίο παρεμβάλλεται μεταξύ αυτών. Οι κυκλικές μεταβολές δεν επαναλαμβάνονται σε κανονικά χρονικά διαστήματα με αποτέλεσμα η συχνότητα μεταξύ δύο “peak” ή δύο “trough” να μην είναι σταθερή, επομένως δεν υπάρχει καθορισμένη, με σταθερό μήκος, περίοδος. Η κυκλικότητα ως χαρακτηριστικό μίας χρονοσειράς απεικονίζεται

γραφικά ως μία μεταβολή που κυμαίνεται από μία χαμηλή στάθμη σε μία πιο υψηλή και παρατηρείται ολοκλήρωση του κύκλου σε χρονικό διάστημα μεγαλύτερου του ενός έτους και συνήθως της τάξεως της πενταετίας και δεκαετίας. Η κυκλική κίνηση δεν ακολουθεί κανένα κανονικό μοντέλο αλλά κινείται απρόβλεπτα, για αυτό το λόγο στην πράξη οι κυκλικές αυξομειώσεις είναι οι πλέον δύσκολες να αντιμετωπιστούν. Η κυκλικότητα εμφανίζεται κυρίως σε οικονομικές χρονοσειρές, όπως το Ακαθάριστο Εθνικό Προϊόν, λόγω των ανόδων και των υφέσεων που παρουσιάζουν οι οικονομίες.

Στο παρακάτω διάγραμμα φαίνονται οι μηνιαίες πωλήσεις νεόκτιστων μονοκατοικιών στις Ηνωμένες Πολιτείες της Αμερικής κατά την περίοδο 1973-1995. Παρατηρείται έντονη κυκλικότητα ανά έτος καθώς επίσης έντονη εποχικότητα με περίοδο 6-10 χρόνια.



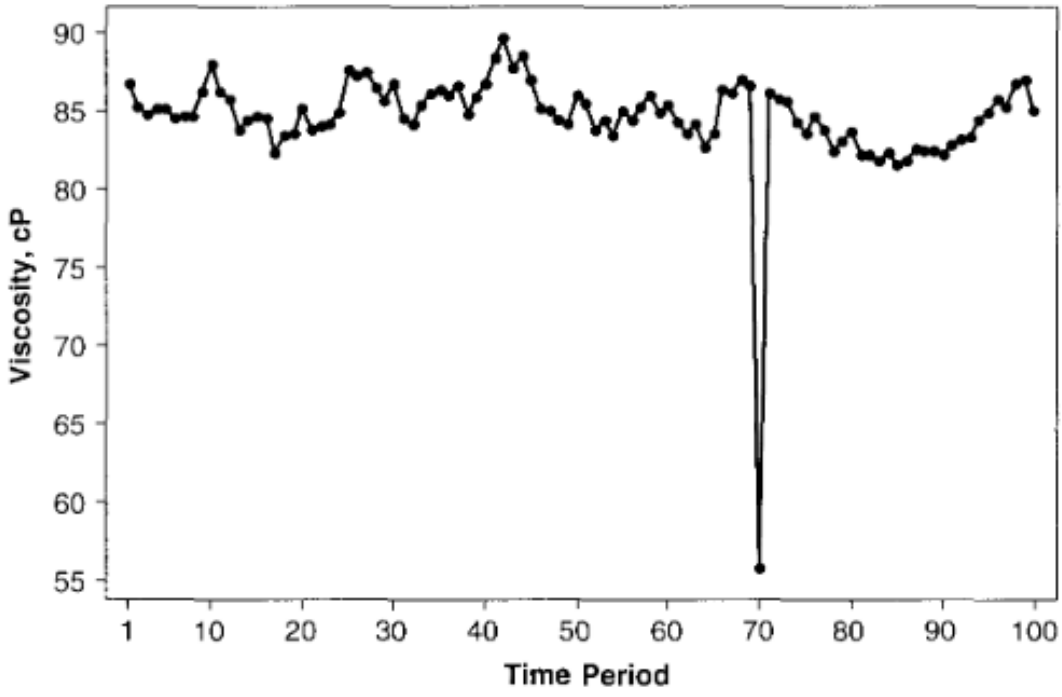
Διάγραμμα 2.6 Μηνιαίες πωλήσεις νεόκτιστων μονοκατοικιών στις Ηνωμένες Πολιτείες της Αμερικής (1973-1995). *Πηγή: [20] Cyclic and seasonal time series (https://robjhyndman.com/hyndsight/cyclicts/)*

Ιδιάζοντα Σημεία (Outliers)

Τα ιδιάζοντα σημεία (outliers) είναι οι απομονωμένες παρατηρήσεις, ακραίες τιμές που εμφανίζονται στο γράφημα κάποιας χρονοσειράς ως απότομες αλλαγές στο πρότυπο συμπεριφοράς της. Τα ιδιάζοντα σημεία είναι μη προβλέψιμα και η επίδρασή τους στην χρονοσειρά έχει μικρή χρονική διάρκεια. Η ερμηνεία τέτοιων παρατηρήσεων χρειάζεται ιδιαίτερη προσοχή, διότι απαιτείται θεωρητική γνώση, κριτική ικανότητα και κοινή λογική. Ένα outlier μπορεί να αντιπροσωπεύει μια ασυνήθιστη παρατήρηση που οφείλεται σε κάποιο απρόβλεπτο γεγονός ή κάποια βλάβη στο σύστημα καταγραφής των παρατηρήσεων. Για παράδειγμα, μια απεργία μπορεί να προκαλέσει μεγάλη πτώση στην παραγωγή μιας βιοτεχνίας.

Στο επόμενο διάγραμμα φαίνεται η χρονοσειρά καταγραφής ιξώδους κατά τη χρονική διάρκεια μίας χημικής διαδικασίας και παρατηρείται μία απομονωμένη τιμή τη

χρονική στιγμή 70, κατά την οποία το σύστημα καταγραφής ελέγχθηκε και διαπιστώθηκε κάποια δυσλειτουργία στον αισθητήρα καταγραφής.



Διάγραμμα 2.7 Καταγραφή ιξώδους κατά τη χρονική διάρκεια μίας χημικής διαδικασίας. *Πηγή: [5] Douglas C. Montgomery, Cheryl L. Jennings, Murat Kulahci, "Introduction to Time Series Analysis and Forecasting", 2008, Wiley*

2.5 Συσχέτιση Χρονοσειράς

Με τον όρο συσχέτιση αναφερόμαστε στη σχέση μεταξύ δύο τυχαίων μεταβλητών. Πιο ειδικά, με τον όρο συσχέτιση μίας χρονοσειράς εννοούμε την ύπαρξη εξάρτησης μεταξύ μίας τιμής της χρονοσειράς x_t τη χρονική στιγμη t και μίας άλλης τιμής x_{t+h} με χρονική υστέρηση h (lag= h). Πρακτικά αυτό σημαίνει ότι η μεταβολή μίας τιμής οφείλεται στη συμπεριφορά της προηγούμενης από αυτή τιμής αν $h = 1$ ή της h - υστέρησής της .

Πριν διερευνήσουμε τις συσχετίσεις σε μία χρονοσειρά, παρακάτω παρουσιάζεται η βασική χρονοσειρά μηδενικής συσχέτισης.

2.5.1 Λευκός Θόρυβος

Το βασικό δομικό στοιχείο για όλες τις χρονοσειρές ειναι ο Λευκός Θόρυβος *(White Noise)* $\{\varepsilon_t\}$. Αν θεωρήσουμε διαδοχικά στοιχεία της χρονοσειράς ως τυχαίες

μεταβλητές, τότε αυτές αποτελούν ανεξάρτητες τυχαίες μεταβλητές με ίδια κατανομή *(independent and identically distributed, iid)* όταν οι $\varepsilon_t, \varepsilon_{t+1}, \dots, \varepsilon_{t+\tau}$ τυχαίες μεταβλητές για τ > 1 έχουν την ίδια κατανομή και είναι ανεξάρτητες μεταξύ τους. Για την ανεξαρτησία ισχύει:

- $E(\varepsilon_t) = 0$
- $\gamma_\varepsilon(t,\tau) = cov(\varepsilon_t, \varepsilon_\tau) = \sigma_\varepsilon^2\,,\ \ t = \tau$
- $\gamma_\varepsilon(t,\tau) = cov(\varepsilon_t, \varepsilon_\tau) = 0\,,\ \ t \neq \tau$

Μια iid χρονοσειρά είναι εντελώς τυχαία και δεν περιέχει αυτοσυσχετίσεις (γραμμικές ή μη-γραμμικές), δηλαδή δεν υπάρχουν συσχετίσεις μεταξύ των τυχαίων μεταβλητών της χρονοσειράς. Μια iid χρονοσειρά λέγεται και λευκός θόρυβος *(white noise)* και η κατανομή της συμβολίζεται ως $WN(0, \sigma_\varepsilon^2)$ με μέση τιμή 0 και διασπορά σ_ε^2. Αν επιπλέον οι τυχαίες μεταβλητές της χρονοσειράς λευκού θορύβου ακολουθούν κανονική (Γκαουσιανή) κατανομή, τότε η χρονοσειρά λέγεται Γκαουσιανός λευκός θόρυβος *(Gaussian white noise)*. (βλέπε [15]).

2.5.2 Εκτίμηση Συσχέτισης

Για να διαπιστώσουμε το βαθμό συσχέτισης μεταξύ των τιμών ενός συνόλου παρατηρήσεων με σκοπό την επιλογή ενός κατάλληλου μοντέλου για τα δεδομένα, ένα σημαντικό εργαλείο που χρησιμοποιείται είναι η δειγματική συνάρτηση αυτοσυσχέτισης *(sample autocorrelation function = sample ACF)* των δεδομένων.

Έστω $x_1, x_2, \dots, x_n$ οι παρατηρήσεις μίας χρονοσειράς. Ο δειγματικός μέσος των $x_1, x_2, \dots, x_n$ είναι :

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

Η εκτίμηση της δειγματικής αυτοδιασποράς *(sample autocovariance)* με υστέρηση h δίνεται ως:

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

Η εκτίμηση της δειγματικής αυτοσυσχέτισης *(sample autocorrelation function)* είναι:

$$\hat{\rho}(h) = r_h = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad -n < h < n$$

Για $h = 0$ είναι $\hat{\rho}(h) = 1$.

Σύμφωνα με τους Douglas, Montgomery, Jennings και Kulahci (βλέπε [5]), ως εμπειρικός κανόνας προτείνεται ότι χρειάζονται τουλάχιστον 50 παρατηρήσεις για την ύπαρξη αξιόπιστης εκτίμησης της ACF, και οι μεμονωμένες δειγματικές αυτοσυσχετίσεις πρέπει να υπολογίζονται μέχρι την υστέρηση h, η οποία είναι περίπου $T/4$.

Στην περίπτωση χρονοσειράς λευκού θορύβου, δηλαδή όταν οι παρατηρούμενες τιμές της χρονοσειράς είναι ασυσχέτιστες μεταξύ τους, τότε θεωρητικά θα ισχύει $\hat{\rho}(h) = 0,\ \forall h \neq 0$. Αποδεικνύεται πως η εκτιμούμενη αυτοσυσχέτιση της χρονοσειράς λευκού θορύβου ακολουθεί κανονική κατανομή, $r_h \sim N(0, {}^{1}/_{n})$. Για τον λόγο αυτό θεωρούμε ότι η αυτοσυσχέτιση για κάποιο h είναι 'στατιστικά μηδενική' αν $r_h \in [-1{,}96/\sqrt{n}\,, 1{,}96/\sqrt{n}]$.

Η παραπάνω προσέγγιση για την αυτοσυσχέτιση μας επιτρέπει να σχεδιάσουμε (παραμετρικό) έλεγχο σημαντικότητας για την αυτοσυσχέτιση, δηλαδή $\mathrm{H}_0: \rho_h = 0$ και $\mathrm{H}_1: \rho_h \neq 0$. Θεωρώντας ως στατιστικό ελέγχου το r_h, η απορριπτική περιοχή είναι $R = \left\{\left|\frac{r_h}{\sqrt{1/n}}\right| > z_{1-a/2}\right\}$ σε επίπεδο σημαντικότητας a. (Βλέπε [19] σε αυτή την παράγραφο)

Πρακτικά λοιπόν ορίζουμε ως σημαντική αυτοσυσχέτιση για κάποια υστέρηση h, όταν η δειγματική αυτοσυσχέτιση r_h είναι έξω από το όριο $\pm z_{1-a/2}/\sqrt{n}$, που για επίπεδο σημαντικότητας $a = 5\%$ το όριο προσεγγιστικά είναι $\pm 1{,}96/\sqrt{n}$.

Γενικότερα, ο συντελεστής αυτοσυσχέτισης ρ, όπως παρουσιάστηκε και παραπάνω, εκφράζει το βαθμό και τον τρόπο με τον οποίο δύο μεταβλητές συσχετίζονται, δηλαδή πώς μία τυχαία μεταβλητή επηρεάζεται από την άλλη. Ο συντελεστής αυτοσυσχέτισης ρ παίρνει τιμές στο διάστημα $[-1,1]$. Οι χαρακτηριστικές τιμές του ρ ερμηνεύονται ως εξής:

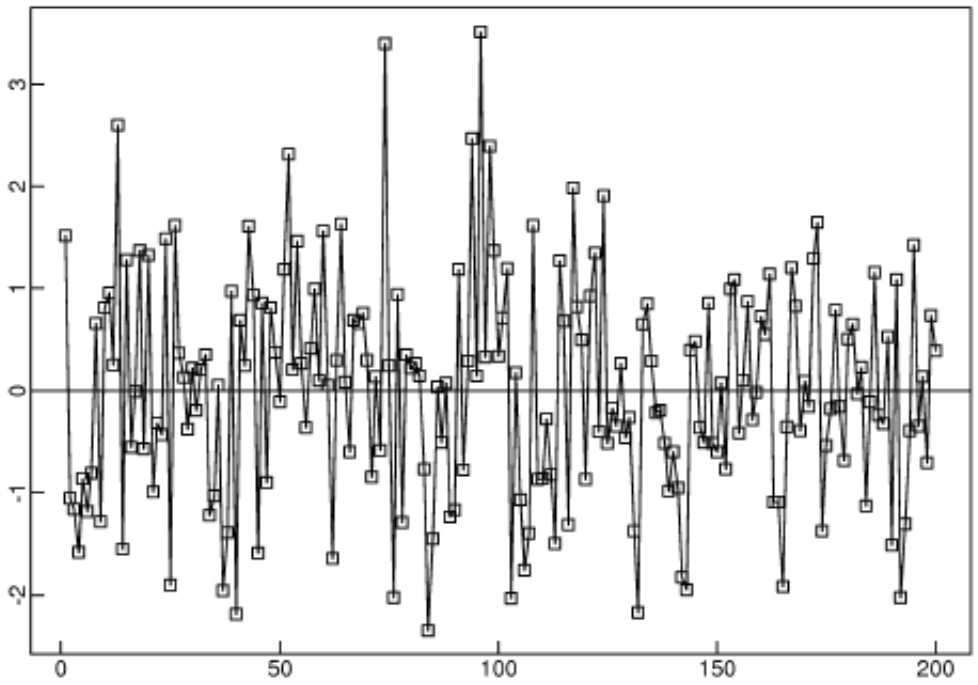
- $\rho = 1$ τότε υπάρχει τέλεια θετική συσχέτιση μεταξύ δύο τυχαίων μεταβλητών
- $\rho = 0$ τότε δεν υπάρχει καμιά (γραμμική) συσχέτιση μεταξύ δύο τυχαίων μεταβλητών
- $\rho = 1$ τότε υπάρχει τέλεια αρνητική συσχέτιση μεταξύ δύο τυχαίων μεταβλητών

Οταν $\rho = \pm 1$ η σχέση είναι αιτιοκρατική κι όχι πιθανοκρατική γιατί γνωρίζοντας την τιμή της μιας τυχαίας μεταβλητής γνωρίζουμε και την τιμή της άλλης τυχαίας μεταβλητής ακριβώς. ΄Οταν ο συνετελεστής συσχέτισης είναι κοντά στο -1 ή 1 , η γραμμική συσχέτιση των δύο τυχαίων μεταβλητών είναι ισχυρή (συνήθως χαρακτηρίζονται ισχυρές οι συσχετίσεις όταν $|\rho| > 0{,}9$) ενώ όταν είναι κοντά στο 0 οι τυχαίες μεταβλητές είναι πρακτικά ασυσχέτιστες.

Επιπλέον, ο συντελεστής αυτοσυσχέτισης ρ δεν εξαρτάται από τη μονάδα μέτρησης των τυχαίων μεταβλητών, καθώς επίσης είναι συμμετρικός γύρω από το 0, δηλαδή $\rho_h = \rho_{-h}$. Για αυτόν τον λόγο έχει νόημα να μελετάμε μόνο το θετικό ή το αρνητικό μέρος.

Για ένα οποιοδήποτε πεπερασμένο σύνολο παρατηρήσεων $\{x_1, x_2, \dots, x_n\}$ μπορεί να υπολογιστεί η δειγματική αυτοδιασπορά και η δειγματική αυτοσυσχέτιση. Αυτό σημαίνει ότι ο υπολογισμός τους δεν περιορίζεται μόνο σε παρατηρήσεις που αφορούν στάσιμες χρονοσειρές. Το ερώτημα είναι πώς αναπαρίσταται γραφικά η *ACF* σε κάθε περίπτωση χρονοσειράς, στάσιμης ή μη-στάσιμης.

Για στάσιμες χρονοσειρές, που περιέχουν τάση, η $|\hat{\rho}(h)|$ εμφανίζει πολύ υψηλές τιμές και φθίνει αργά με τις υστερήσεις h. Αντίστοιχα η αυτοσυσχέτιση μιας (μη-στάσιμης) χρονοσειράς με έντονη περιοδικότητα ή εποχικότητα θα παρουσιάσει ταλαντώσεις με κορυφές σε υστερήσεις που είναι πολλαπλάσια της περιοδικότητας.

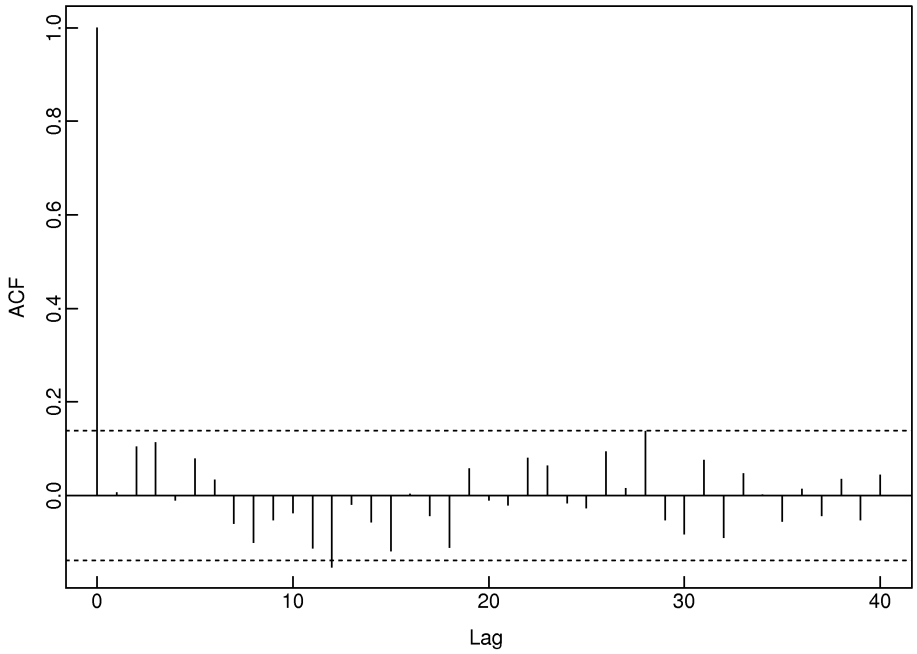


Διάγραμμα 2.8 Απεικόνιση 200 τυχαίων τιμών κανονικής κατανομής iid (0,1), λευκού θορύβου. *Πηγή: [8] Peter J. Brockwell, Richard A. Davis - Introduction to Time Series and Forecasting (2002, Springer)*

Τα επόμενα παραδείγματα αντλούνται από το βιβλίο « Introduction to Time Series and Forecasting » (βλέπε [8]) .

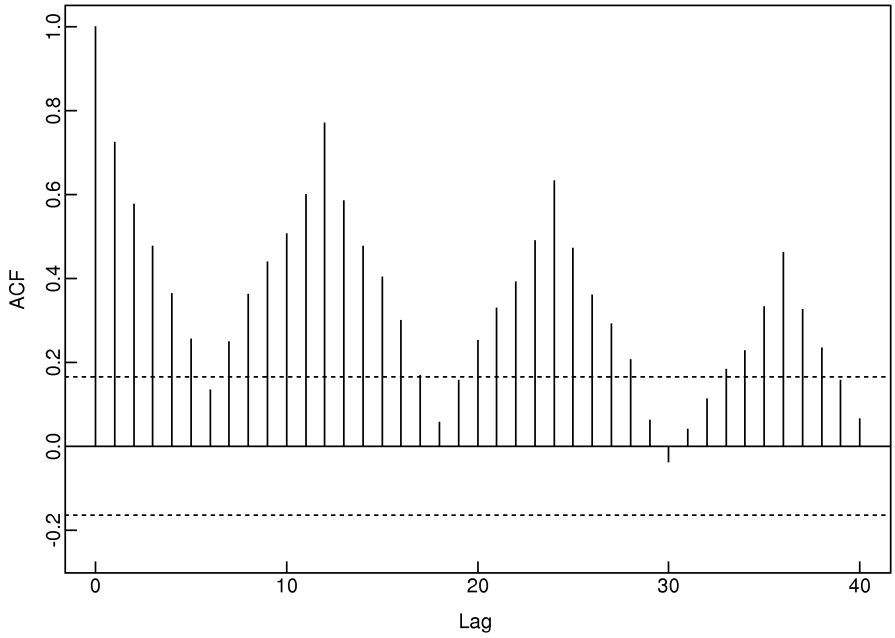
Το διάγραμμα 2.8 απεικονίζει 200 τυχαίες τιμές από μία κανονική κατανομή iid (0,1), λευκού θορύβου. Στο διάγραμμα 2.9 φαίνεται το αντίστοιχο γράφημα της δειγματικής αυτοσυσχέτισης με υστερήσεις $0,1 \dots ,40$. Θα περίμενε κανείς 2 μόνο τιμές να πέφτουν έξω από τα όρια, όπως και φαίνεται, δηλαδή μόνο το 5% των τιμών της δειγματικής αυτοσυσχέτισης να βρίσκεται εκτός του διαστήματος $\pm 1{,}96/\sqrt{n}$. Οι

διακεκομμένες οριζόντιες γραμμές απεικονίζουν τα όρια στατιστικής σημαντικότητας.



Διάγραμμα 2.9 Η συνάρτηση δειγματικής αυτοσυσχέτισης ACF για τα δεδομένα του γραφήματος 2.8. *Πηγή: [8] Peter J. Brockwell, Richard A. Davis - Introduction to Time Series and Forecasting (2002, Springer)*

Αντίστοιχα στο διάγραμμα 2.10 φαίνεται η συνάρτηση δειγματικής αυτοσυσχέτισης για μία μη-στάσιμη χρονοσειρά που περιέχει τάση και εποχικότητα.



Διάγραμμα 2.10 Συνάρτηση δειγματικής αυτοσυσχέτισης μη-στάσιμης χρονοσειράς με τάση και εποχικότητα. *Πηγή: [8] Peter J. Brockwell, Richard A. Davis - Introduction to Time Series and Forecasting (2002, Springer)*

Η εφαρμογή αυτής της ιδιότητας είναι αρκετά σημαντική γιατί η διαδικασία επιλογής στατιστικού μοντέλου βασίζεται στην ελάττωση των μη-στάσιμων χρονοσειρών σε χρονοσειρές λευκού θορύβου χρησιμοποιώντας κατάλληλες μετατροπές. Μετά την εφαρμογή αυτής της διαδικασίας, το γράφημα της συνάρτησης της δειγματικής αυτοσυσχέτισης των υπολοίπων θα πρέπει να βρίσκεται μεταξύ των ορίων που δόθηκαν παραπάνω.

Επιπρόσθετα για τον έλεγχο της συσχέτισης μεταξύ δύο τιμών x_t και x_{t+h} όταν έχουν αφαιρεθεί οι γραμμικές επιδράσεις των ενδιάμεσων τιμών $\{x_{t+1}, \dots, x_{t+h-1}\}$ χρησιμοποιείται η συνάρτηση μερικής αυτοσυσχετίσεως *(partial autocorrelation function = PACF)*.

Σε επόμενο κεφάλαιο θα γίνει περαιτέρω ανάλυση στην χρήση των συναρτήσεων της δειγματικής αυτοσυσχέτισης *ACF* και της μερικής αυτοσυσχετίσης *PACF* για τον εντοπισμό της τάξης μοντέλου που θα χρησιμοποιηθεί για την ανάλυση της χρονοσειράς.

2.5.3 Έλεγχος Στασιμότητας

Όπως αναφέρθηκε στην προηγούμενη παράγραφο, ο έλεγχος της στασιμότητας μπορεί να επιτευχθεί μέσω των γραφικών παραστάσεων της αυτοσυσχέτισης και της μερικής αυτοσυσχέτισης. Επιπρόσθετα μπορούμε να ελέγξουμε τη στασιμότητα ελέγχοντας την ύπαρξη μοναδιαίων ριζών *(unit roots)*.

Στις χρονικές σειρές η ύπαρξη μοναδιαίας ρίζας σημαίνει ότι κάποια ρίζα της εξίσωσης $f(x) = 0$ του πολυωνύμου n–οστού βαθμού γενικά, $f(x) = 1 - \rho_1 x^1 - \rho_2 x^2 - \rho_3 x^3 - \cdots - \rho_n x^n$ ισούται με το ένα, δηλαδή βρίσκεται πάνω στον μοναδιαίο κύκλο (βλέπε [17]).

Ενα αυτοπαλινδρομούμενο μοντέλο πρώτης τάξης AR(1), δηλαδή όταν υπάρχει συσχέτιση μιας τ.μ. X σε χρόνο t με την προηγούμενη ή την επόμενή της, του οποίου η ανάλυση θα γίνει σε επόμενο κεφάλαιο, θα μπορούσε να έχει μία ρίζα που ισούται με το 1. Το μοντέλο πρώτης τάξης με συντελεστή αυτοσυσχέτισης ρ κοντά στη μονάδα και λευκό θόρυβο u_t γράφεται ως

$$Y_t = \rho Y_{t-1} + u_t$$

όπου u_t η διαδικασία λευκού θορύβου *(white noise)* με μέσο μηδέν και σταθερή διακύμανση.

Έχει αποδειχθεί ότι σε αυτό το μοντέλο ο εκτιμητής $\hat{\rho}$ είναι μεροληπτικός και υποεκτιμά την παράμετρο ρ. Στην περίπτωση όμως για $|\rho| < 1$ ο εκτιμητής $\hat{\rho}$ είναι συνεπής.

Στην περίπτωση που ο συντελεστής ισούται με τη μονάδα ($\rho = 1$) , δηλαδή έχει μοναδιαία ρίζα, τότε το υπόδειγμα είναι μία διαδικασία μη στάσιμη. Αν ο συντελεστής είναι μικρότερος της μονάδας ($|\rho| < 1$) το υπόδειγμα είναι μία διαδικασία στάσιμη.

Συνοψίζοντας, έχουμε τις παρακάτω υποθέσεις:

$H_0: \rho = 1$ η διαδικασία Y_t είναι μη στάσιμη (υπάρχει μοναδιαία ρίζα).

$H_a: \rho < 1$ η διαδικασία Y_t είναι στάσιμη (δεν υπάρχει μοναδιαία ρίζα).

Στην περίπτωση που ισχύει η H_0, δηλαδή έχουμε μοναδιαία ρίζα, τότε έχουμε τη διαδικασία του τυχαίου περίπατου, δηλαδή έχουμε μία μη στάσιμη διαδικασία.

Οι έλεγχοι αυτοί που καλούνται έλεγχοι μοναδιαίας ρίζας *(unit root tests)* αντιστοιχούν στην υπόθεση $H_0: \rho = 1$ για την εξίσωση αυτοπαλινδρόμησης. Εκτιμώντας την εξίσωση $Y_t = \rho Y_{t-1} + u_t$ με τη μέθοδο των ελαχίστων τετραγώνων είναι λογικό να κάνουμε τον έλεγχο της $H_0: \rho = 1$ με την κατανομή t - Student. Ο εκτιμητής όμως μπορεί να είναι μεροληπτικός, οπότε η κατανομή t - Student (λόγω συμμετρίας) να μην είναι η κατάλληλη για τον έλεγχο της χρονοσειράς που εξετάζουμε, πολύ δε περισσότερο όταν η διαδικασία είναι και μη στάσιμη.

[17] - Νικόλαος Δριτσάκης. Σημειώσεις "Υπολογιστικές Τεχνικές Εκτιμητικής"/ μάθημα 4 (Καθηγητής Τμήμα Εφαρμοσμένης Πληροφορικής Πανεπιστήμιο Μακεδονίας) (http://users.uom.gr/~drits/lessons/Lesson%204(MSc%20Inf).pdf)

2.5.3.1 Έλεγχος Dickey-Fuller (DF) και Augmented Dickey-Fuller (ADF)

Οι Dickey-Fuller μέσω των πειραμάτων Monte-Carlo ανέπτυξαν μία κατάλληλη διαδικασία για τον έλεγχο $H_0: \rho = 1$. Ο έλεγχος Dickey-Fuller εξετάζει αν μία χρονοσειρά έχει μοναδιαία ρίζα ή αντίστοιχα αν η χρονοσειρά ακολουθεί τη διαδικασία του τυχαίου περίπατου. Έστω το υπόδειγμα :

$$\Delta X_t = \delta_0 + \delta_1 t + \delta X_{t-1} + e_t$$

όπου έχουμε υποθέσει ότι η χρονοσειρά έχει μοναδιαία ρίζα με μετατόπιση και ντετερμινιστική τάση και e_t είναι μία ανεξάρτητη και στάσιμη διαδικασία.

Άρα ο έλεγχος Dickey-Fuller εξετάζει αν μία χρονοσειρά έχει μοναδιαία ρίζα $H_0: \delta = 0$ και αν οι πρώτες διαφορές βοηθούν στην απομάκρυνση της ρίζας αυτής.

Στην περίπτωση που ο όρος e_t δεν είναι ανεξάρτητος, λόγω συσχετίσεων στη χρονική σειρά, τότε χρησιμοποιείται ο επαυξημένος έλεγχος δημιουργώντας το υπόδειγμα:

$$\Delta X_t = \delta_0 + \delta_1 t + \delta X_{t-1} + \sum_{i=1}^{p} \beta_i \Delta X_{t-i} + e_i \qquad (2.5.3.1)$$

όπου p, η τάξη της αυτοπαλίνδρομης διαδικασίας. Στο βιβλίο του Hamilton (1994) (βλέπε [13]) παρουσιάζονται 4 διαφορετικές περιπτώσεις στις οποίες ο επαυξημένος έλεγχος Dickey-Fuller (Augmented Dickey-Fuller) μπορεί να εφαρμοστεί.

Μία πρώτη περίπτωση που μπορούμε να εξετάσουμε αντιστοιχεί στην μηδενική υπόθεση ότι η X_t ακολουθεί τυχαίο περίπατο χωρίς μετατόπιση *(drift)* ή τάση και από την (2.5.3.1) αφαιρείται η σταθερά και η τάση ($\delta_1 t$).

$$\Delta X_t = \delta X_{t-1} + \sum_{i=1}^{p} \beta_i \Delta X_{t-i} + e_i \qquad (2.5.3.2)$$

- H_0: $\delta = 0$ $(\eta\ X_t\ \epsilon\acute{\iota}\nu\alpha\iota\ \tau\upsilon\chi\alpha\acute{\iota}\circ\varsigma\ \pi\epsilon\rho\acute{\iota}\pi\alpha\tau\circ\varsigma,\ \ \epsilon\acute{\iota}\nu\alpha\iota\ \mu\eta - \sigma\tau\acute{\alpha}\sigma\iota\mu\eta)$
- H_α: $\delta < 0$ $(\delta\epsilon\nu\ \iota\sigma\chi\acute{\upsilon}\epsilon\iota\ \eta\ H_0)$

Σε δεύτερη περίπτωση η μηδενική υπόθεση παραμένει η ίδια όπως στην πρώτη περίπτωση αλλά προσθέτουμε τη σταθερά.

$$\Delta X_t = \delta_0 + \delta X_{t-1} + \sum_{i=1}^{p} \beta_i \Delta X_{t-i} + e_i \qquad (2.5.3.3)$$

- H_0: $\delta = 0$ $(\eta\ \sigma\epsilon\iota\rho\acute{\alpha}\ X_t\ \epsilon\acute{\iota}\nu\alpha\iota\ \tau\upsilon\chi\alpha\acute{\iota}\circ\varsigma\ \pi\epsilon\rho\acute{\iota}\pi\alpha\tau\circ\varsigma\ \mu\epsilon\ \pi\epsilon\rho\iota\pi\lambda\acute{\alpha}\nu\eta\sigma\eta, \delta\eta\lambda\alpha\delta\acute{\eta}$ $\pi\epsilon\rho\iota\acute{\epsilon}\chi\epsilon\iota\ \mu\acute{\iota}\alpha\ \mu\circ\nu\alpha\delta\iota\alpha\acute{\iota}\alpha\ \rho\acute{\iota}\zeta\alpha\ \acute{\alpha}\rho\alpha\ \epsilon\acute{\iota}\nu\alpha\iota\ \mu\eta - \sigma\tau\acute{\alpha}\sigma\iota\mu\eta)$
- H_α: $\delta < 0$ $(\delta\epsilon\nu\ \iota\sigma\chi\acute{\upsilon}\epsilon\iota\ \eta\ H_0)$

Στις δύο πρώτες περιπτώσεις η πληθυσμιακή τιμή της σταθεράς θεωρείται μηδενική.

Στην τρίτη περίπτωση υποθέτουμε ότι η X_t είναι τυχαίος περίπατος με περιπλάνηση, επομένως η πληθυσμιακή τιμή της σταθεράς είναι διάφορη του μηδενός.

$$\Delta X_t = \delta_0 + \delta X_{t-1} + \sum_{i=1}^{p} \beta_i \Delta X_{t-i} + e_i \qquad (2.5.3.4)$$

- H_0: $\delta = 0$ (η $\sigma\epsilon\iota\rho\acute{\alpha}$ X_t $\epsilon\acute{\iota}\nu\alpha\iota$ $\tau\upsilon\chi\alpha\acute{\iota}o\varsigma$ $\pi\epsilon\rho\acute{\iota}\pi\alpha\tau o\varsigma$ $\mu\epsilon$ $\pi\epsilon\rho\iota\pi\lambda\acute{\alpha}\nu\eta\sigma\eta$, $\delta\eta\lambda\alpha\delta\acute{\eta}$ $\pi\epsilon\rho\iota\acute{\epsilon}\chi\epsilon\iota$ $\mu\acute{\iota}\alpha$ $\mu o\nu\alpha\delta\iota\alpha\acute{\iota}\alpha$ $\rho\acute{\iota}\zeta\alpha$ $\acute{\alpha}\rho\alpha$ $\epsilon\acute{\iota}\nu\alpha\iota$ $\mu\eta - \sigma\tau\acute{\alpha}\sigma\iota\mu\eta$)
- H_α: $\delta < 0$ ($\delta\epsilon\nu$ $\iota\sigma\chi\acute{\upsilon}\epsilon\iota$ η H_0)

Στην τέταρτη περίπτωση η μηδενική υπόθεση είναι ότι η X_t ακολουθεί τυχαίο περίπατο με περιπλάνηση γύρω από μία στοχαστική τάση.

$$\Delta X_t = \delta_0 + \delta_1 t + \delta X_{t-1} + \sum_{i=1}^{p} \beta_i \Delta X_{t-i} + e_i \qquad (2.5.3.5)$$

- H_0: $\delta = 0$ (η $\sigma\epsilon\iota\rho\acute{\alpha}$ X_t $\epsilon\acute{\iota}\nu\alpha\iota$ $\tau\upsilon\chi\alpha\acute{\iota}o\varsigma$ $\pi\epsilon\rho\acute{\iota}\pi\alpha\tau o\varsigma$ $\mu\epsilon$ $\pi\epsilon\rho\iota\pi\lambda\acute{\alpha}\nu\eta\sigma\eta$ $\gamma\acute{\upsilon}\rho\omega$ $\alpha\pi\acute{o}$ $\mu\acute{\iota}\alpha$ $\sigma\tau o\chi\alpha\sigma\tau\iota\kappa\acute{\eta}$ $\tau\acute{\alpha}\sigma\eta$, $\delta\eta\lambda\alpha\delta\acute{\eta}$ $\pi\epsilon\rho\iota\acute{\epsilon}\chi\epsilon\iota$ $\mu\acute{\iota}\alpha$ $\mu o\nu\alpha\delta\iota\alpha\acute{\iota}\alpha$ $\rho\acute{\iota}\zeta\alpha$ $\acute{\alpha}\rho\alpha$ $\epsilon\acute{\iota}\nu\alpha\iota$ $\mu\eta - \sigma\tau\acute{\alpha}\sigma\iota\mu\eta$)
- H_α: $\delta < 0$ ($\delta\epsilon\nu$ $\iota\sigma\chi\acute{\upsilon}\epsilon\iota$ η H_0)

Ο έλεγχος των Dickey-Fuller γίνεται με την κατανομή Dickey-Fuller t-Student ενώ η σύγκριση για την αποδοχή ή την απόρριψη της H_0 γίνεται από τις κριτικές τιμές του MaKinnon των πινάκων Dickey-Fuller (1979). (βλέπε [22])

Η μηδενική υπόθεση απορρίπτεται όταν το στατιστικό t-Student του συντελεστή δ είναι μικρότερο από την εκάστοτε κριτική τιμή των πινάκων Dickey-Fuller (1979).

Ο έλεγχος Dickey-Fuller αφορά έλεγχο ύπαρξης μοναδιαίας ρίζας σε αυτοπαλίνδρομο μοντέλο πρώτης τάξης AR(1), ενώ ο επαυξημένος έλεγχος Dickey-Fuller αφορά έλεγχο μοναδιαίας ρίζας σε ένα υπόδειγμα τάξης μεγαλύτερης της πρώτης.

2.5.3.2 Έλεγχος Kwiatkowski, Phillips, Schmidt and Shin (KPSS)

Ο παραπάνω έλεγχος εξετάζει την μηδενική υπόθεση ότι η χρονοσειρά X_t είναι μη στάσιμη. Την αντίθετη περίπτωση, δηλαδή η περίπτωση όπου η μηδενική υπόθεση εξετάζει αν η χρονοσειρά είναι στάσιμη ως προς την τάση, έναντι της εναλλακτικής ότι υπάρχει μοναδιαία ρίζα, περιγράφεται από τον KPSS έλεγχο (Kwiatkowski, Phillips, Schmidt and Shin, 1992).

Οι Kwiatkowski, Phillips, Schmidt και Shin (βλέπε [9]) βασίστηκαν πάνω στην ιδέα ότι η χρονοσειρά είναι στάσιμη ως προς την τάση, και υπολογίζεται ως το άθροισμα της τάσης, ενός υποδείγματος τυχαίου περίπατου και ενός στάσιμου κατάλοιπου. Το μοντέλο είναι ως εξής:

$$Y_t = d_t + r_t + \varepsilon_t,$$

$$r_t = r_{t-1} + u_t$$

όπου $d_t = \sum_{i=0}^{p} \beta_i t^i$, για $p = 0, 1$, περιλαμβάνει αιτιοκρατικά μέρη του μοντέλου (σταθερά, αιτιοκρατική τάση), ε_t είναι *iid* N(0, ${\sigma_\varepsilon}^2$), r_t είναι τυχαίος περίπατος με διακύμανση ${\sigma_u}^2$ και u_t είναι *iid* N(0, ${\sigma_\varepsilon}^2$).

Ο έλεγχος *KPSS* βασίζεται στην έλεγχοσυνάρτηση *LM* με την υπόθεση ότι ο τυχαίος περίπατος έχει μηδενική διακύμανση , δηλαδή H_0: ${\sigma_u}^2 = 0$, το οποίο σημαίνει ότι r_t είναι σταθερό, με εναλλακτική υπόθεση H_1: ${\sigma_u}^2 > 0$. Η ελεγχοσυνάρτηση *KPSS* ορίζεται ως εξής:

$$LM = \sum_{t=1}^{T} {s_t}^2 \Big/ {\hat{\sigma}_\varepsilon}^2$$

όπου $s_t = \sum_{t=1}^{T} \hat{\varepsilon}_t \,, t = 1,2, \dots, T$ και ${\hat{\sigma}_\varepsilon}^2$ είναι η εκτίμηση της διακύμανσης των καταλοίπων ε_t.

Σημειώνεται, ότι για να γίνει αποδεκτή η μηδενική υπόθεση της στασιμότητας πρέπει η τιμή της ελεγχοσυνάρτησης *LM* να είναι μικρότερη από την κρίσιμη τιμή για συγκεκριμένο επίπεδο σημαντικότητας.

2.6 Μετασχηματισμός μη-στάσιμης σε στάσιμη χρονοσειρά

Όπως είδαμε και παραπάνω, τις περισσότερες φορές για μία χρονοσειρά θα απαιτείται κατάλληλος μετασχηματισμός έτσι ώστε να μετατραπεί σε στάσιμη και να προχωρήσουμε στην ανάλυση της μετασχηματισμένης χρονοσειράς.

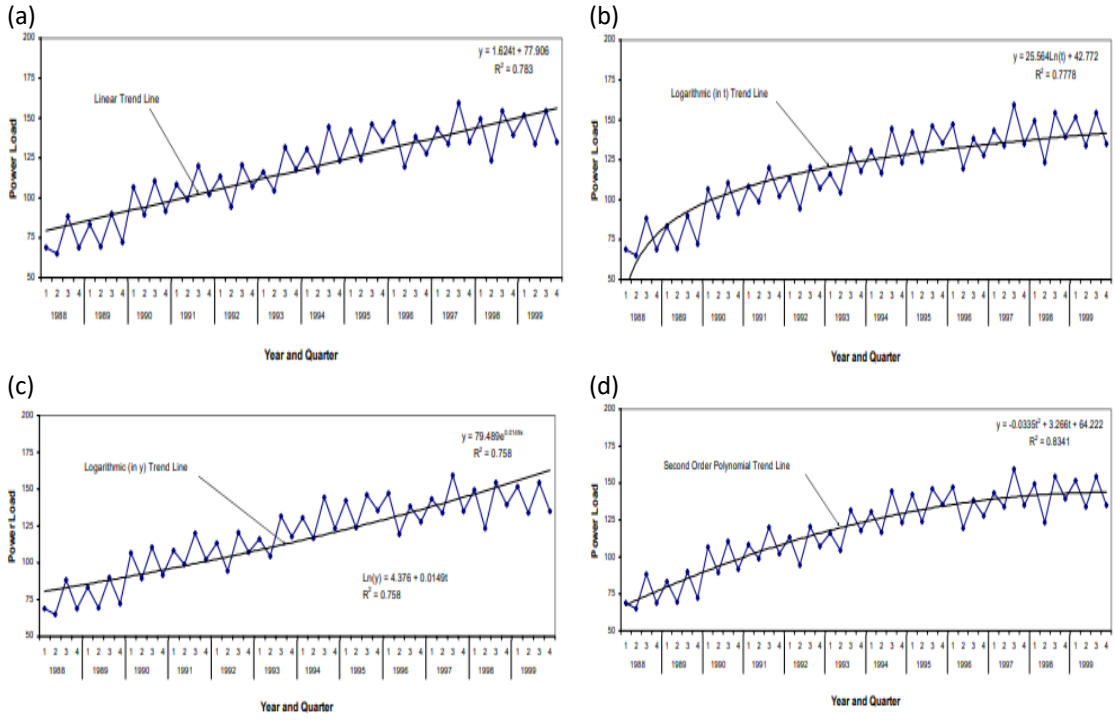
Υπάρχουν περιπτώσεις που το γράφημα της αρχικής χρονοσειράς παρουσιάζει μεγάλες διακυμάνσεις τόσο περιέχοντας συνιστώσες όπως τάση ή/και εποχικότητα καθώς επίσης παρουσιάζοντας αστάθεια στη διασπορά.

Γενικά η απαλοιφή της τάσης ή της εποχικότητας (γενικά περιοδικότητας) γίνεται όταν δεν μας ενδιαφέρει να μελετήσουμε τις μεταβολές στη χρονοσειρά που οφείλονται σε τάσεις ή περιοδικότητα γιατί θεωρούμε ότι δημιουργούνται από άλλους παράγοντες που δε σχετίζονται με το σύστημα που θέλουμε να διερευνήσουμε ή να περιγράψουμε. Για προβλέψεις, είτε συμπεριλαμβάνουμε την τάση και περιοδικότητα στο μοντέλο πρόβλεψης, είτε εκτιμούμε το μοντέλο στη χρονοσειρά που προκύπτει αφαιρώντας την τάση ή περιοδικότητα και στις προβλέψεις του μοντέλου αυτού προσθέτουμε την τάση και την περιοδικότητα για να πάρουμε την πρόβλεψη του παρατηρούμενου μεγέθους.

Υπάρχουν περιπτώσεις, όπως παρουσιάζονται παρακάτω, όπου η τάση σε μια χρονοσειρά μπορεί να περιγραφεί από κάποια γνωστή ή εκτιμώμενη συνάρτηση του χρόνου, $\mu_t = f(t)$, και τότε ονομάζεται *ντιτερμινιστική (καθοριστική)τάση (deterministic trend)*. Μπορεί όμως η τάση σε μια χρονοσειρά να μην είναι δυνατόν να περιγραφεί από μια γνωστή (παραμετρική) συνάρτηση του χρόνου, να παρουσιάζει δηλαδή αργές μεταβολές με το χρόνο αλλά όχι με κάποιο συγκεκριμένο τρόπο. Αυτή η τάση λέγεται *στοχαστική (stochastic trend)*. Στα χρηματο-οικονομικά, τυπικά οι διάφοροι δείκτες παρουσιάζουν στοχαστική τάση.

Κάποιες από τις γνωστές συναρτήσεις – υποδείγματα καθορισμού της τάσης είναι τα εξής:

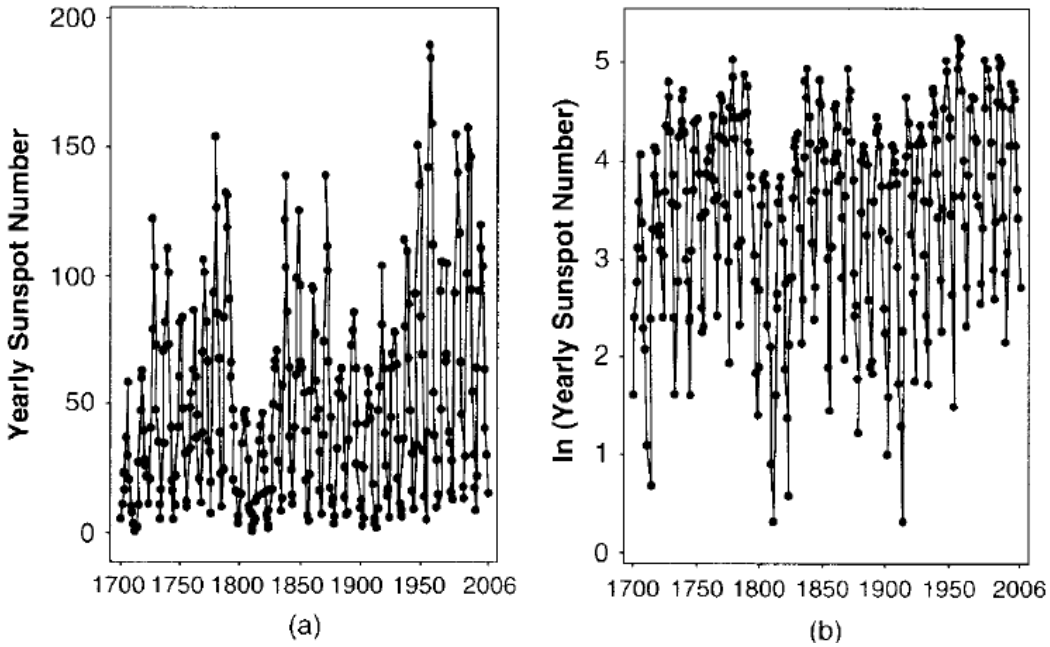
- Γραμμική Τάση (Διάγραμμα 2.11a): $\mu_t = a + bt + e_t$
- Λογαριθμική Τάση (Διάγραμμα 2.11b): $\mu_t = a + b\ ln(t) + e_t$
- Εκθετική Τάση (Διάγραμμα 2.11c): $\mu_t = exp\{a + bt + e_t\}$
- Πολυωνυμική Τάση, π.χ. 2ου βαθμού (Διάγραμμα 2.11d): $\mu_t = a + b_1 t + b_2 t^2 + e_t$



Διάγραμμα 2.11 (a) Χρονοσειρά επηρεασμένη από γραμμική τάση. (b) Χρονοσειρά επηρεασμένη από λογαριθμική τάση. (c) Χρονοσειρά επηρεασμένη από εκθετική τάση. (d) Χρονοσειρά επηρεασμένη από πολυωνυμική τάση 2ου βαθμού.

2.6.1 Σταθεροποίηση διασποράς

Συχνό φαινόμενο στα δεδομένα χρονοσειρών είναι η ύπαρξη αστάθειας στη διακύμανση. Ο μετασχηματισμός των δεδομένων είναι χρήσιμος συχνά στην σταθεροποίηση της διακύμανσης. Για παράδειγμα, όπως φαίνεται στο παρακάτω γράφημα, για τον αριθμό των ηλιακών κηλίδων η μεταβλητότητα από το 1800 έως το 1830 είναι μικρότερη συγκριτικά με το διάστημα 1830 έως 1880. (βλέπε [5])



Διάγραμμα 2.12 Ετήσιος αριθμός ηλιακών κηλίδων (a) πριν το μετασχηματισμό δεδομένων (b) μετά τον λογαριθμικό μετασχηματισμό των δεδομένων. *Πηγή: [5] Douglas C. Montgomery, Cheryl L. Jennings, Murat Kulahci, "Introduction to Time Series Analysis and Forecasting", 2008, Wiley*

Ένας πολύ γνωστός τρόπος μετασχηματισμού δεδομένων που αντιμετωπίζει το πρόβλημα της αστάθειας στη διακύμανση όταν αυτή είναι συνάρτηση της μέσης τιμής, δηλαδή $Var(X_t) = g(\mu_X)$, έχει προταθεί από τους Box και Cox το 1964 και είναι ο μετασχηματισμός δύναμης, ως εξής:

$$y = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ lnx, & \lambda = 0 \end{cases}$$

Ο παραπάνω τρόπος χρησιμοποιείται υπό την υπόθεση ότι η διακύμανση της αρχικής χρονοσειράς μεταβάλλεται ως συνάρτηση της τάσης. Ο μετασχηματισμός Box και Cox αντιμετωπίζει το πρόβλημα της σταθεροποίησης της διακύμανσης με παράμετρο λ που πρέπει να εκτιμηθεί από τα δεδομένα. Για συγκεκριμένες μορφές της διακύμανσης, δίνονται στον Πίνακα 2.1 ο μετασχηματισμός της χρονοσειράς για τη

σταθεροποίηση της διασποράς καθώς και η αντίστοιχη τιμή της παραμέτρου λ του μετασχηματισμού δύναμης.

Για $\lambda = 0$ έχουμε τον απλό λογαριθμικό μετασχηματισμό, ο οποίος χρησιμοποιείται συχνά σε περιπτώσεις όπου η μεταβλητότητα στην αρχική χρονοσειρά αυξάνεται με το μέσο επίπεδο της χρονοσειράς. Όταν η τυπική απόκλιση της αρχικής χρονοσειράς αυξάνεται γραμμικά με τη μέση τιμή, τότε ο απλός λογαριθμικός μετασχηματισμός είναι ο βέλτιστος τρόπος σταθεροποίησης της διακύμανσης.

λ	Μετασχηματισμός ισοδύναμος	$VarX_t$
-1	$\frac{1}{X_t}$	$c \cdot \mu^4$
-0.5	$\frac{1}{\sqrt{X_t}}$	$c \cdot \mu^3$
0	$\ln X_t$	$c \cdot \mu^2$
0.5	$\sqrt{X_t}$	$c \cdot \mu$

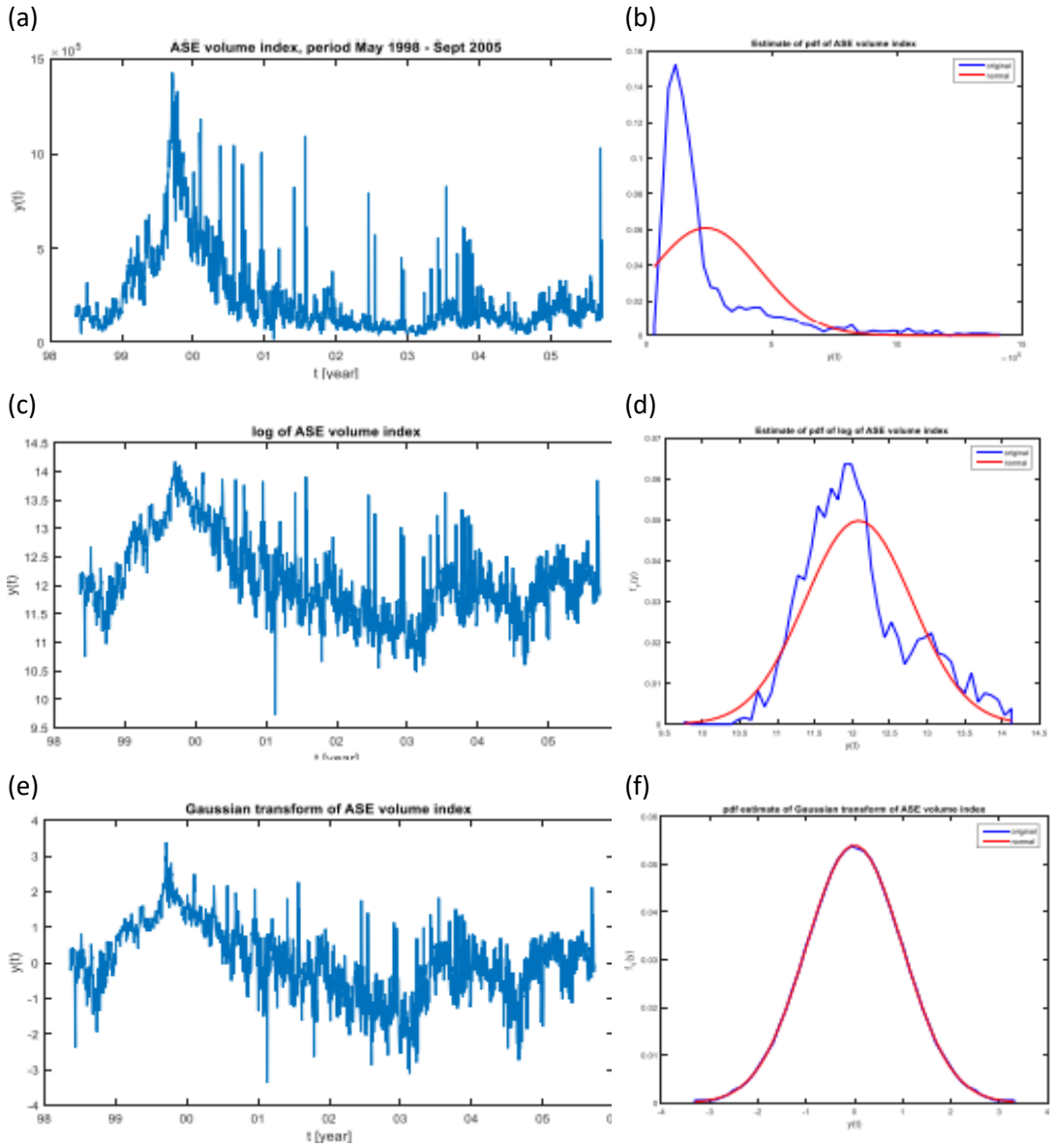
Πίνακας 2.1 Μετασχηματισμοί σταθεροποίησης διακύμανσης (Στήλη 2) για συγκεκριμένες συναρτήσεις της διακύμανσης ως προς την τάση (Στήλη 3, c είναι σταθερά) και η αντίστοιχη τιμή της παραμέτρου λ του μετασχηματισμού της δύναμης. *Πηγή: [15] Κουγιουμτζής Δημήτρης. Σημειώσεις «Ανάλυση Χρονοσειρών» (Αν. Καθηγητής Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ, Πολυτεχνική Σχολή, ΑΠΘ)*

Ο μετασχηματισμός της δύναμης εκτός από το να σταθεροποιεί τη διακύμανση της αρχικής χρονοσειράς, διορθώνει την περιθώρια κατανομή της προς την κατεύθυνση της κανονικής (Γκαουσιανής) κατανομής.

Σημειώνεται πως ο πιο κατάλληλος μετασχηματισμός για να πετύχουμε η περιθώρια κατανομή να είναι ακριβώς Γκαουσιανή είναι

$$x_t = \Phi^{-1}\big(F_Y(y_t)\big)$$

όπου $F_Y(y_t)$ είναι η περιθώρια αθροιστική κατανομή της αρχικής χρονοσειράς και $\Phi^{-1}(u)$ είναι η αντίστροφη αθροιστική συνάρτηση της τυπικής Γκαουσιανής κατανομής. Σημειώνεται ότι η τιμή u περιγράφει τυχαία μεταβλητή $U \sim U[0,1]$, δηλαδή ομοιόμορφη κατανομή στο $[0,1]$. Είναι γνωστό ότι σαν τυχαία μεταβλητή η αθροιστική συνάρτηση κατανομής (είτε η $F_Y(y)$, είτε η $\Phi(x)$) ακολουθεί ομοιόμορφη κατανομή στο $[0,1]$. (βλέπε [19]).



Διάγραμμα 2.13 Σταθεροποίηση διακύμανσης
(a) Το γράφημα της αρχικής χρονοσειράς, (b) Το γράφημα της περιθώριας κατανομής της (επίσης στο σχήμα δίνεται το γράφημα της κανονικής κατανομής), (c) το γράφημα της χρονοσειράς που προκύπτει από το μετασχηματισμό του λογαρίθμου, (d) το γράφημα της περιθώριας κατανομής της μετασχηματισμένης χρονοσειράς καθώς και της κανονικής κατανομής, (e) το γράφημα της χρονοσειράς που προκύπτει από το μετασχηματισμό $x_t = \Phi^{-1}\left(F_Y(y_t)\right)$ και (f) το γράφημα της περιθώριας κατανομής της x_t που μετατράπηκε σε κανονική κατανομή. *Πηγή: [21] http://users.auth.gr/dkugiu/Teach/TimeSeries/Lec2.pdf)*

2.6.2 Απαλοιφή Τάσης και Περιοδικότητας

Εν συνεχεία του μετασχηματισμού των δεδομένων σταθεροποιώντας τη διασπορά στην χρονοσειρά, χρησιμοποιούνται διάφοροι τρόποι που στοχεύουν στη διόρθωση της χρονοσειράς απαλοίφοντας την συνιστώσα της τάσης, της εποχικότητας ή των κυκλικών κυμάνσεων.

Μετά από τη γραφική αναπαράσταση της αρχικής χρονοσειράς και κατ επέκταση της μετασχηματισμένης χρονοσειράς, και αφού παρατηρηθεί ότι τα δεδομένα επηρεάζονται από μία τουλάχιστον από τις παραπάνω συνιστώσες, τότε πρόκειται για μία μη-στάσιμη χρονοσειρά. Η ανάλυση και η πρόγνωση μίας μη-στάσιμης χρονοσειράς διευκολύνεται όταν απαλοίφουμε τη συνιστώσα της τάσης ή/και της περιοδικότητας και συνήθως αυτή η διαδικασία ονομάζεται *κλασικό μοντέλο διαμέρισης (classical decomposition model)*.

Είναι χρήσιμο λοιπόν να αναπαριστούμε τα δεδομένα ως μία πραγματοποίηση από τις δύο παρακάτω διαδικασίες που φανερώνουν τον τρόπο με τον οποίο οι συνιστώσες των χρονολογικών σειρών συνδέονται μεταξύ τους.

Πρόκειται για το προσθετικό μοντέλο (additive model):

$$X_t = m_t + s_t + Y_t$$

και το πολλαπλασιαστικό μοντέλο (multiplicative model):

$$X_t = m_t \times s_t \times Y_t$$

όπου οι συνιστώσες της χρονοσειράς έχουν τους ακόλουθους συμβολισμούς:

$m_t =$ η συνιστώσα της Τάσης

$s_t =$ η συνιστώσα της Εποχικότητας και

$Y_t =$ είναι το υπόλοιπο απαλλαγμένο από τάση και περιοδικότητα.

Το κύριο χαρακτηριστικό του προσθετικού μοντέλου είναι ότι όλες οι συνιστώσες είναι ανεξάρτητες μεταξύ τους και εκφράζονται στην ίδια μονάδα μέτρησης με εκείνη των παρατηρήσεων της χρονοσειράς. Αντίθετα, στο πολλαπλασιαστικό μοντέλο, μόνο η τάση εκφράζεται στην ίδια μονάδα με εκείνη της χρονονειράς, ενώ τα υπόλοιπα στοιχεία είναι δείκτες ανεξάρτητοι από μονάδες μέτρησης. Επίσης, το προσθετικό μοντέλο είναι κατάλληλο για χρονοσειρές στις οποίες το εύρος της εποχικής συνιστώσας παραμένει σταθερό. Αντίστοιχα, το πολλαπλασιαστικό μοντέλο είναι περισσότερο κατάλληλο αν οι εποχικές επιδράσεις αυξάνονται ή μειώνονται με σταθερό ποσοστό.

Αν μία χρονοσειρά χρήζει αναπαράστασης κατά το πολλαπλασιαστικό μοντέλο, τότε εύκολα μπορεί να τροποποιηθεί σε προσθετικό μοντέλο χρησιμοποιώντας λογαρίθμους, είτε φυσικό λογάριθμο είτε κοινό.

$$\log(X_t) = \log(m_t \times s_t \times Y_t)$$

$$\log X_t = \log m_t + \log s_t + \log Y_t$$

Μετά τη χρήση λογαρίθμων οι συνιστώσες της τάσης, της εποχικότητας και το κατάλοιπο δρουν και πάλι στη χρονοσειρά προσθετικά, οπότε ισχύουν οι κανόνες για το προσθετικό μοντέλο.

2.6.2.1 Εκτίμηση και Απαλοιφή Τάσης εν απουσία Περιοδικότητας

Αφού γίνει γραφική αναπαράσταση των δεδομένων μίας χρονοσειράς σε συνάρτηση με το χρόνο και διαπιστωθεί ότι η χρονοσειρά επηρεάζεται μόνο από τη συνιστώσα της τάσης τότε το μοντέλο γράφεται ως εξής:

$$X_t = m_t + Y_t \qquad t = 1,2, \dots, n$$

όπου $EY_t = 0$.

Γενικά αν παρατηρηθεί καθοριστική (ντιτερμινιστική) τάση στα δεδομένα της χρονοσειράς, τότε η τάση μπορεί εύκολα να εκτιμηθεί και στη συνέχεια να απαλειφθεί. Για παράδειγμα, ο πιο απλός τρόπος εμφάνισης της τάσης στα δεδομένα ειναι ως μία γραμμική συνάρτηση με το χρόνο $m_t = \beta_0 + \beta_1 t$.

Σε αυτή την περίπτωση η εφαρμογή της μεθόδου των ελαχίστων τετραγώνων μπορεί να προσδιορίσει την τάση αυτή. Η μέθοδος ελαχιστοποιεί το άθροισμα:

$$\sum_{t=1}^{n} (x_t - \alpha - \beta t)^2$$

δίνοντας

$$\widehat{m}_t = \hat{\alpha} + \widehat{\beta_t}$$

με

$$\hat{\alpha} = \bar{x} - \hat{\beta} t$$

$$\hat{\beta} = C_{tx}/C_{tt}$$

όπου

$$\bar{x} = n^{-1} \sum_t x_t$$

$$\bar{t} = n^{-1} \sum_t t = (n+1)/2$$

και

$$C_{tx} = \sum_t (t - \bar{t})(x_t - \bar{x}),$$

$$C_{tt} = \sum_t (t - \bar{t})^2 = n\,(n^2 - 1)/12$$

Οι ίδιοι τύποι ισχύουν για το λογάριθμο της εκθετικής τάσης. Επίσης, ανάλογα αποτελέσματα προκύπτουν για πολυωνυμικές τάσεις από την παρακάτω ελαχιστοποίηση:

$$\min_{\{\alpha_j : j=0,\dots,k\}} \sum_{t=1}^{n} (x_t - \alpha_0 - \alpha_1 t - \alpha_2 t^2 - \dots - \alpha_k t^k)^2.$$

Εναλλακτικός τρόπος εκτίμησης της τάσης m_t είναι με τη μέθοδο *(αμφίπλευρης) εξομάλυνσης κινητού μέσου (moving average smoothing method)*. Συγκεκριμένα με $1 \le q \ll n$ εξομαλύνουμε την $X_t = m_t + Y_t$, $t = 1,2, \dots, n$, από την

$$w_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j} = \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t+j} + \frac{1}{2q+1} \sum_{j=-q}^{q} Y_{t+j} \ , q+1 \le t \le n-q$$

Επειδή έχουμε $EY_t = 0$, συμπεραίνουμε ότι το δεύτερο άθροισμα στα δεξιά της παραπάνω σχέσης είναι πολύ κοντά στο μηδέν και συνεπώς μπορούμε να δεχθούμε ότι τα m_t

ικανοποιούν το σύστημα εξισώσεων:

$$\frac{1}{2q+1} \sum_{t=-q}^{q} m_{t+j} = w_t = \frac{1}{2q+1} \sum_{t=-q}^{q} X_{t+j} \ , q+1 \le t \le n-q$$

Αν τώρα, με t σταθερό, τα m_{t+j}, $-q \le j \le q$, συνδέονται γραμμικά μεταξύ τους, τότε έχουμε τις παρακάτω εκτιμήτριες κινητού μέσου:

$$\widehat{m}_t = \frac{1}{2q+1} \sum_{t=-q}^{q} X_{t+j} \ , q+1 \le t \le n-q \ .$$

Παραλλαγή της παραπάνω μεθόδου είναι η μέθοδος της μονόπλευρης εξομάλυνσης η οποία είναι γνωστή επίσης και ως εκθετική εξομάλυνση. Με $\alpha \in [0,1]$ θεωρούμε ότι οι εκτιμήτριες $\widehat{m}_t$ ικανοποιούν το παρακάτω σύστημα εξισώσεων:

$\widehat{m}_t = \alpha x_1 + (1-a)\widehat{m}_{t-1}, t = 2, \dots, n$ με $\widehat{m}_1 = x_1$

Με διαδοχική αντικατάσταση εύκολα διαπιστώνεται ότι

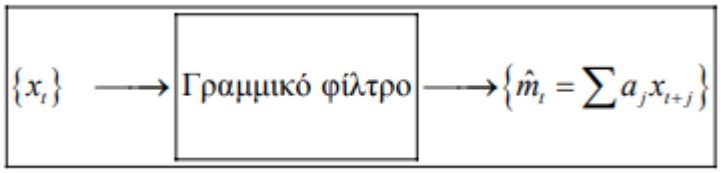
$\widehat{m}_t = \alpha \sum_{j=0}^{t-2} (1-\alpha)^j x_{t-j} + (1-\alpha)^{t-1} x_1$ για $t \ge 2$,

όπου το άθροισμα συντελεστών δίνει τη μονάδα όπως και στην περίπτωση του αμφίπλευρου κινητού μέσου.

Με βάση τα παραπάνω μπορούμε να θεωρήσουμε ότι και στις δύο περιπτώσεις η εκτιμηθείσα τάση $\widehat{m}_t$ είναι το αποτέλεσμα της δράσης ενός γραμμικού φίλτρου πάνω στην $\{X_t\}$. Τούτο διότι έχουμε:

$\widehat{m}_t = \sum_{j=-\infty}^{\infty} \alpha_j x_{t-j}$, $t \in \mathbb{Z}$ με $\sum_j |\alpha_j| < \infty$

Το αποτέλεσμα αυτής της δράσης πάνω στην $\{X_t\}$ είναι απαλλαγμένο από τον θόρυβο υψηλής συχνότητας $\{Y_t\}$ ενώ διατηρεί τις χαμηλές συχνότητες *(φίλτρο χαμηλών συχνοτήτων).*



Σημειώνεται επίσης ότι όταν η τάση m_t είναι ένα χαμηλού βαθμού πολυώνυμο του t είναι δυνατόν, με κατάλληλη επιλογή του q και των βαρών α_j, με $|j| \leq q$, και $\alpha_j = 0$ για $|j| > q$ να απαλλαγεί από τον θόρυβο υψηλών συχνοτήτων χωρίς να αλλοιωθεί η υπάρχουσα πολυωνυμική σχέση μεταξύ της τάσης m_t και του χρόνου t.

Στην περίπτωση που η τάση είναι στοχαστική, δηλαδή δεν φαίνεται να είναι κάποια γνωστή συνάρτηση με το χρόνο, τότε ο πιο κατάλληλος τρόπος απαλοιφής της τάσης από τη χρονοσειρά είναι η *μέθοδος των διαφορών (differencing)*. Αυτή η μέθοδος υπερτερεί συγκριτικά με τις προηγούμενες σε δύο σημεία. Αρχικά δεν απαιτεί την εκτίμηση της συνιστώσας της τάσης, δηλαδή δεν χρειάζεται να εκτιμηθούν παράμετροι, άρα ειναι πιο απλή μέθοδος, καθώς απλά απαλείφει την τάση από τη χρονοσειρά. Αυτό ταυτόχρονα αποτελεί μειονέκτημα αν είναι αναγκαία η εκτίμηση της τάσης m_t, οπότε η μέθοδος της διαμέρισης είναι καταλληλότερη. Κατ'επέκταση, καθώς γίνεται η προσαρμογή του μοντέλου υποθέτει ότι η ύπαρξη της τάσης στη χρονοσειρά ειναι σταθερή καθ'όλο το μήκος της χρονοσειράς.

Με τη μέθοδο των διαφορών επιδιώκεται η εξάλειψη της τάσης m_t που υπάρχει στη χρονοσειρά $\{X_t\}$ δημιουργώντας μία νέα χρονοσειρά από τις διαφορές μεταξύ διαδοχικών όρων $x_t - x_{t-1}$ για $t = 2,3, \dots, n$. Για παράδειγμα, αν στην αρχική χρονοσειρά η τάση ειναι γραμμική, η νέα χρονοσειρά $w_t = x_t - x_{t-1}$ που παράγεται έχει μηδενική τάση.

Για την παραπάνω διαδικασία διαφορών βοηθά ιδιαίτερα η χρήση του τελεστή $\mathcal{B}$, γνωστός ως ο *οπισθοδρομικός τελεστής (backward operator)* και ο τελεστής ∇ γνωστός ως ο *τελεστής διαφοράς (difference operator)*.

Οπισθοδρομικός τελεστής: $\mathcal{B}x_t = x_{t-1}$

Τελεστής Διαφοράς: $\nabla x_t = x_t - x_{t-1} = (1 - \mathcal{B})x_t$

Εφαρμόζοντας τους τελεστές αυτούς k φορές λαμβάνουμε αντίστοιχα:

$$\mathcal{B}^k x_t = x_{t-k}$$

και

$\nabla^k x_t = \nabla(\nabla^{k-1} x_t)$ με $\nabla^0 = I x_t = x_t$

όπου με I συμβολίζεται ο ταυτοτικός τελεστής. Για $k = 2$ η τελευταία δίνει:

$$\nabla^2 x_t = \nabla(\nabla x_t) = \nabla(x_t - x_{t-1}) = x_t - 2x_{t-1} + x_{t-2} = (I - 2\mathcal{B} + \mathcal{B}^2)x_t,$$

καθώς διαπιστώνεται εύκολα ότι ισχύει ο Διωνυμικός τύπος

$$\nabla^k = (I - \mathcal{B})^k = \sum_{l=0}^{k} \binom{k}{l} (-1)^l \mathcal{B}^l.$$

Συνεπώς

$$\nabla^k x_t = (I - \mathcal{B})^k x_t = \sum_{l=0}^{k} \binom{k}{l} (-1)^l x_{t-l}.$$

Έτσι όταν η τάση m_t είναι πολυώνυμο βαθμού k, όταν δηλαδή έχουμε

$$X_t = m_t + Y_t = \sum_{j=0}^{k} \alpha_j t^j + Y_t \,, t \in \mathbb{Z}$$

τότε, εφαρμόζοντας τον τελεστή ∇^k εξαλείφεται η πολυωνυμική τάση και προκύπτει η στάσιμη χρονοσειρά

$$X_t^{(k)} \equiv \nabla^k X_k = k!\, \alpha_k + \nabla^k Y_t, \qquad t \in \mathbb{Z}.$$

Στην πραγματικότητα η παραπάνω μέθοδος εφαρμόζεται σταδιακά. Αυτό γιατί για κάθε τιμή του k που εφαρμόζεται ο αναδρομικός τύπος, χρειάζεται να γίνει η γραφική αναπαράσταση της κάθε νέας χρονοσειράς $\left\{X_t^{(k)}, t = 1, \dots, n - k\right\}$, από την οποία προκύπτει αν έχει επιτευχθεί απαλοιφή της τάσης m_t δηλαδή στάσιμη χρονοσειρά. Αν όχι, ο τελεστής ∇ εφαρμόζεται άλλη μία φορά και συνεχίζεται η

διαδικασία. Συνήθως δεν χρειάζεται η διαδικασία αυτή να επαναληφθεί πάνω από δύο ή το πολύ τρεις φορές.

Αν η τάση είναι στοχαστική, τότε η συνιστώσα της τάσης m_t μπορεί να περιγραφεί ως μία στοχαστική διαδικασία χρησιμοποιώντας το υπόδειγμα του τυχαίου περίπατου με περιπλάνηση.

$$m_t = \delta + m_{t-1} + z_t$$

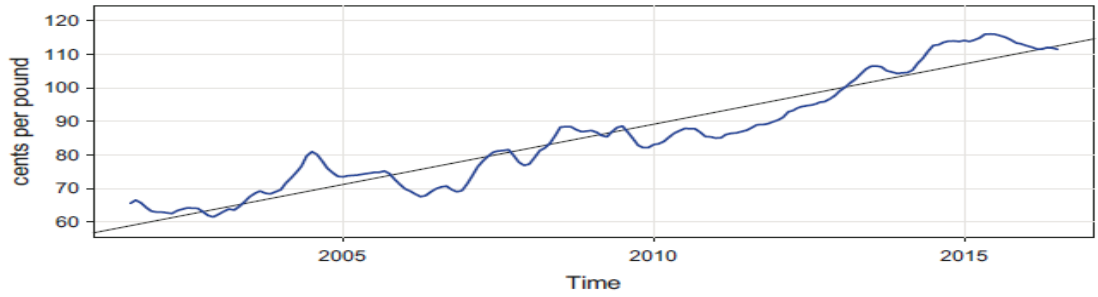
όπου z_t είναι λευκός θόρυβος ανεξάρτητος του Y_t.

Εφαρμόζοντας τις πρώτες διαφορές η χρονοσειρά γίνεται:

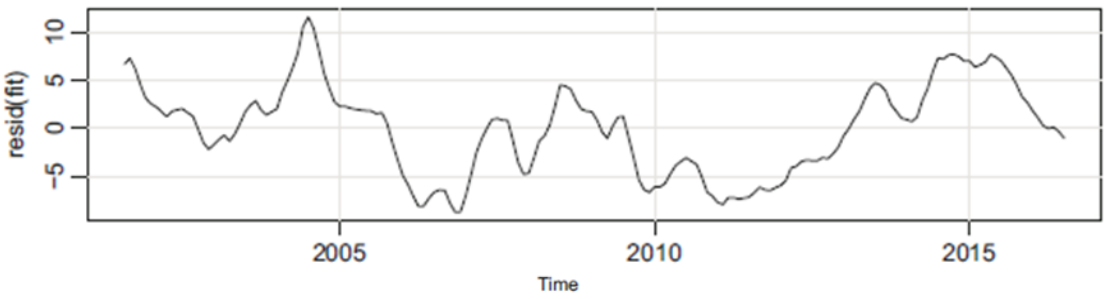
$$X_t - X_{t-1} = (m_t + Y_t) - (m_{t-1} + Y_{t-1})$$
$$= \delta + z_t + Y_t - Y_{t-1}.$$

Το παρακάτω παράδειγμα δίνεται από το βιβλίο «Time Series Analysis and Its Applications: With R Examples» [1] και δείχνει τη χρονοσειρά για τις μηνιαίες τιμές των κοτόπουλων στις Ηνωμένες Πολιτείες από τα μέσα του 2001 μέχρι τα μέσα του 2016, καθώς και το διάγραμμα της νέας, στάσιμης χρονοσειράς με τη μέθοδο της διαμέρισης και των διαφορών.

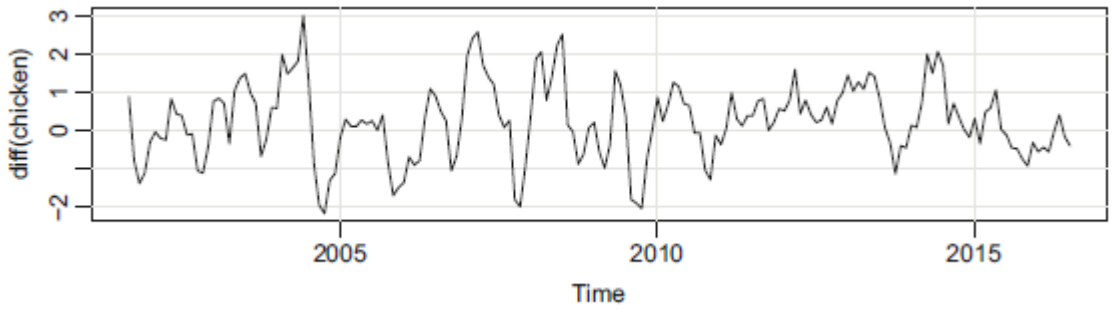
(a)



(b)



(c)



Διάγραμμα 2.14 Απαλοιφή Τάσης

(a) Διάγραμμα αρχικής χρονοσειράς: Μηνιαία τιμή κοτόπουλων στις Ηνωμένες Πολιτείες της Αμερικής κατά την περίοδο Αύγουστος 2001 – Ιούλιος 2016 υπό τάση. (b) Δίαγραμμα χρονοσειράς απαλλαγμένη από τη συνιστώσα της τάσης με τη μέθοδο της διαμέρισης. (c) Δίαγραμμα χρονοσειράς απαλλαγμένη από τη συνιστώσα της τάσης με εφαρμογή των πρώτων διαφορών. *Πηγή: [1] Robert H. Shumway, David S. Stoffer: Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics) Fourth (4th) Edition*

2.6.2.2 Εκτίμηση και Απαλοιφή Περιοδικότητας εν απουσία Τάσης

Υποθέτουμε τώρα ότι τα δεδομένα στη χρονοσειρά επηρεάζονται μόνο από τη συνιστώσα της εποχικότητας οπότε το μοντέλο γράφεται:

$$X_t = s_t + Y_t \qquad t = 1,2, \dots, n$$

όπου $EY_t = 0$.

Επιπλέον θεωρούμε ότι d είναι η περιόδος της περιοδικής συνάρτησης s_t και είναι γνωστή και $s_{t+d} = s_t \; \forall t$. Αν η περιοδικότητα αντιστοιχεί σε κάποια γνωστή περίοδο, όπως 24 ώρες ή 12 μήνες, αναφέρεται ως εποχικότητα. Όπως και με τη συνιστώσα της τάσης, έτσι και με την περιοδική συνιστώσα, αν επιτρέπει την προσέγγιση μίας γνωστής, παραμετρικής, περοδικής συνάρτησης $s_t = f(t)$, όπως την ημιτονονοειδή συνάρτηση, τότε η απαλοιφή της συνιστώσας από το μοντέλο είναι εύκολη.

Συχνά όμως δε συναντάμε γνωστή συνάρτηση για την εκτίμηση της περιοδικής συνιστώσας και ένας απλός τρόπος εκτίμησης της s_t, για γνωστή περίοδο d, είναι από τους μέσους όρους των στοιχείων της περιοδικής συνάρτησης, $s_i \; i = 1, .., d$. Αν $k = {}^{n}/_{d}$ είναι ο αριθμός των περιόδων στη χρονοσειρά $\{X_1, X_2, \dots, X_n\}$, τότε το κάθε στοιχείο της περιοδικής συνάρτησης s_i μπορεί να εκτιμηθεί ως

$$\hat{s}_i = \frac{1}{k}\sum_{j=1}^{k} X_{i+jd}.$$

Ένας δεύτερος τρόπος εκτίμησης της εποχικής συνιστώσας είναι ο κινούμενος μέσος όρος θέτοντας την τάξη του ίση με την περίοδο d. Χρησιμοποιώντας αυτό το φίλτρο εξομάλυνσης επιτυγχάνεται η απαλοιφή της εποχικότητας, δηλαδή παίρνοντας τον κινούμενο μέσο όρο τάξης d σε μεγάλο βαθμό απαλοίφουμε το περιοδικό στοιχείο περιόδου d.

Για να εκτιμήσουμε το s_t με ακρίβεια, ώστε στη συνέχεια να το απαλοίψουμε, θα πρέπει να πάρουμε πρώτα τη διαφορά της αρχικής χρονοσειράς $\{X_{q+1}, X_{q+2}, \dots, X_{n-q+1}\}$ και του κινούμενου μέσου $\{\hat{\mu}_{q+1}, \hat{\mu}_{q+2}, \dots, \hat{\mu}_{n-q+1}\}$, έστω $W_t = X_t - \mu_t$. Στη συνέχεια παίρνουμε το μέσο όρο των $W_t = W_{i+jd}$ ως προς κάθε στοιχείο i για $i = 1, \dots, d$, έστω $\overline{W}_i$. Αν τα $\overline{W}_i$ για $i = 1, \dots, d$ δεν αθροίζονται στο 0, τότε αφαιρούμε τη μέση τιμή τους και η εκτίμηση της περοδικής συνάρτησης είναι

$$\hat{s}_i = \overline{W}_i - \frac{1}{d}\sum_{j=1}^{d} W_j\,, \qquad j = 1, \dots, d$$

όπου $\hat{s}_{t-d} = \hat{s}_t$.

Αν θέλουμε απλά να απαλείψουμε το περιοδικό στοιχείο με περίοδο d μπορούμε απλά να πάρουμε τις διαφορές υστέρησης d, ή d-διαφορές

$$\nabla_d X_t = X_t - X_{t-d} = (1 - B^d)X_t.$$

(βλέπε [19] - Κουγιουμτζής Δημήτρης. Σημειώσεις «Ανάλυση Χρονοσειρών»)

2.6.2.3 Εκτίμηση και Απαλοιφή Τάσης και Περιοδικότητας

Στην περίπτωση που μία χρονοσειρά επηρεάζεται από τη συνιστώσα της τάσης και της περιοδικότητας, τότε συνδυάζονται οι παραπάνω μέθοδοι για την απαλοιφή και των δύο. Η σειρά που εφαρμόζεται κάθε φορά στην απαλοιφή δεν είναι πάντα καθορισμένη. Θα χρησιμοποιήσουμε ένα παράδειγμα από τις σημειώσεις «Ανάλυση Χρονοσειρών» (βλέπε [19]) για να παρουσιάσουμε την συνδυαστική μέθοδο.

Παράδειγμα

Στο σχήμα 2.15a, δίνεται ο γενικός δείκτης τιμών καταναλωτή (general index for consumer price, GICP) σε μηνιαίες τιμές από τον Ιανουάριο 2001 ως τον Αύγουστο 2005. Η χρονοσειρά έχει λοιπόν μήκος $n = 56$.
Παρατηρείται ότι το διάγραμμα GICP παρουσιάζει σταθερή τάση για όλο την περίοδο αλλά και ασθενή ετήσια περιοδικότητα (εποχικότητα). Η τάση μπορεί να αποδοθεί ικανοποιητικά ως μία γραμμική συνάρτηση m_t του χρόνου t (μήνα).

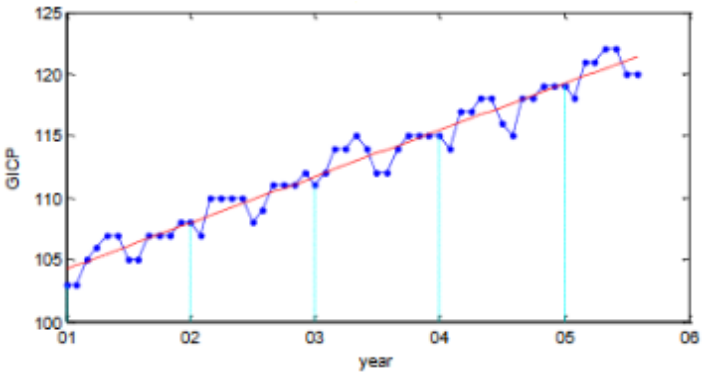
Θεωρούμε το μοντέλο :

$$X_t = m_t + s_t + Y_t \quad t = 1,2,\dots,56$$

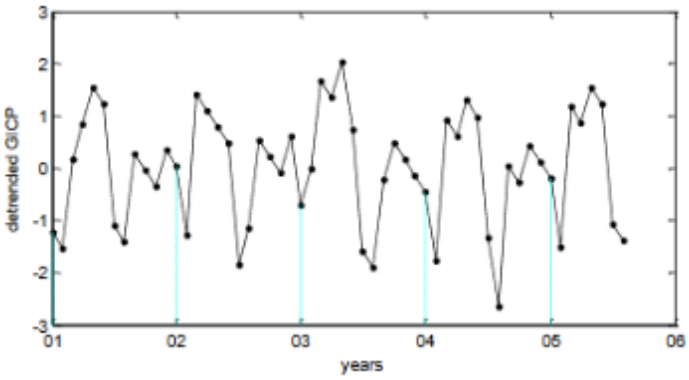
όπου $EY_t = 0$.

Η προσαρμογή απλού γραμμικού μοντέλου παλινδρόμησης του m_t ως προς το t, θεωρώντας τις παρατηρήσεις $\{X_t\}_{t=1}^{56}$ ως τιμές του m_t, έδωσε $m_t = 103{,}9 + 0{,}31t$ και φαίνεται με κόκκινη γραμμή στο σχήμα 9a.

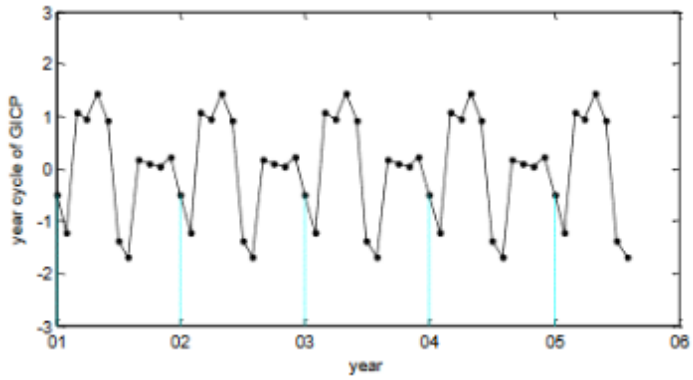
(a)



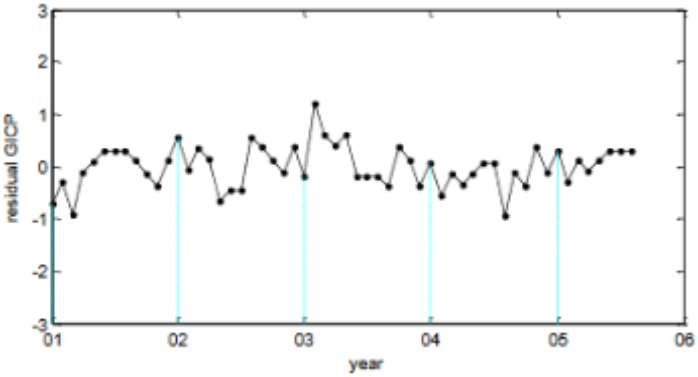
(b)



(c)



(d)



Διάγραμμα 2.15 Απαλοιφή Τάσης και Περιοδικότητας (Εποχικότητας) για το γενικό δείκτη τιμών καταναλωτή (general index for consumer price, GICP)
(a) Διάγραμμα αρχικής χρονοσειράς: Μηνιαίες τιμές γενικού δείκτη τιμών καταναλωτή (GICP) την περίοδο Ιανουάριος 2001 – Αύγουστος 2005 (η κάθετη γραμμή δηλώνει την αρχή του έτους). Στο διάγραμμα φαίνεται η προσαρμογή γραμμικού μοντέλου τάσης. Η ευθεία γραμμή δηλώνει την προσαρμοσμένη γραμμική τάση. (b) Η χρονοσειρά που προκύπτει από την αφαίρεση της γραμμικής τάσης στη χρονοσειρά GICP. (c) Ο εκτιμούμενος ετήσιος κύκλος για την GICP. (d) Η χρονοσειρά που προκύπτει από την αφαίρεση του εκτιμούμενου ετήσιου κύκλου από τη χρονοσειρά στο (b). *Πηγή: [19] Κουγιουμτζής Δημήτρης. Σημειώσεις "Ανάλυση Χρονοσειρών" (Αν. Καθηγητής Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ, Πολυτεχνική Σχολή, ΑΠΘ) (http://users.auth.gr/dkugiu/Teach/TimeSeries/TimeSeries.pdf)*

Στο σχήμα 2.15b αποδίδεται η χρονοσειρά GICP απαλλαγμένη από τη συνιστώσα της γραμμικής τάσης $X'_t = X_t - m_t$. Αυτή είναι η κατάλληλη χρονοσειρά αν θέλουμε να αναλύσουμε τη μεταβολή του GICP απαλλαγμένη από τον πληθωρισμό (πληθωριστική τάση). Σε αυτό το διάγραμμα γίνεται καλύτερα εμφανής η ετήσια περιοδικότητα του GICP, παρόλα αυτά δεν είναι προφανές με ποια περιοδική συνάρτηση του χρόνου μπορούμε να εκτιμήσουμε τον ετήσιο κύκλο [annual cycle]. Μπορούμε να εκτιμήσουμε τον ετήσιο κύκλο του GICP από τους μέσους όρους των στοιχείων της περιοδικής συνάρτησης, δηλαδή τις μέσες τιμές του κάθε μήνα για τα έτη 2001 – 2005 (ως το 2004 για τους μήνες Σεπτέμβριο – Δεκέμβριο). Η περιοδική

χρονοσειρά $\{\widehat{s_t}\}_{t=1}^{n}$ (του επαναλαμβανόμενου ετήσιου κύκλου) δίνεται στο Σχήμα 2.15c, όπου η περίοδος είναι $d = 12$. Τη μέση τιμή για τον Ιανουάριο την αφαιρούμε από τις 5 παρατηρήσεις του Ιανουαρίου (για τα έτη 2001 – 2005). Το ίδιο κάνουμε και για τους άλλους μήνες. Με αυτόν τον τρόπο παίρνουμε τη χρονοσειρά των υπολοίπων $\{Y_t\}_{t=1}^{n}$ (Σχήμα 2.15d)

$$Y_t = X'_t - \widehat{s_t} = X_t - \widehat{m_t} - \widehat{s_t}.$$

Παρατηρούμε ότι αυτή η χρονοσειρά είναι απαλλαγμένη από τάση και περιοδικότητα και δε φαίνεται να έχει κάποια κανονικότητα ή δομή.

ΚΕΦΑΛΑΙΟ 3

Υποδείγματα Χρονολογικών Σειρών

3.1 Στάσιμο Υπόδειγμα Χρονολογικής Σειράς

Όπως αναφέρθηκε στις προηγούμενες παραγράφους, κάθε χρονολογική σειρά (time series) αποτελείται από ένα σύνολο τιμών $x_1, x_2, \dots, x_T$, που αποτελούν συγκεκριμένες τιμές μιας τυχαίας μεταβλητής $X_1, X_2, .., X_T$ (ή περισσοτέρων μεταβλητών), δηλαδή αποτελούν μια σειρά πραγματοποίησεων της τυχαίας μεταβλητής X_t. Επομένως, μια χρονολογική σειρά X_t είναι μια στοχαστική διαδικασία και θα μπορούσε να περιγραφεί από μια συνάρτηση πιθανότητας της μορφής :

$$f(x_1, x_2, \dots, x_n)$$

Η παραπάνω υποθετική συνάρτηση πιθανότητας δεν είναι όμως γνωστή καθώς το σύνολο των παραγόντων που επιδρούν στη διαμόρφωση των τιμών της τυχαίας μεταβλητής $X_t, t = 1, 2, \dots, n$ δεν μπορούν να βρεθούν πλήρως και να υπολογιστούν στη συνέχεια. Επομένως, όπως αναφέρθηκε στις προηγούμενες παραγράφους, στα πλαίσια της εφαρμογής τεχνικών ανάλυσης χρονολογικών σειρών, σκοπός είναι η εύρεση ενός υποδείγματος χρονολογικής σειράς που να παράγει εκτιμούμενες τιμές που να τείνουν στις πραγματικές τιμές (βλέπε [10]). Για την εύρεση ενός τέτοιου υποδείγματος, σημαντικό ρόλο έχει η ιδιότητα της στασιμότητας για την διερευνούμενη χρονολογική σειρά.

Όπως αναφέρθηκε, μια χρονολογική σειρά είναι στάσιμη αν η μέση τιμή της και η διακύμανση της είναι σταθερές ως προς τη μεταβολή του χρόνου, καθώς επίσης η συνδιακύμανση (αυτοσυνδιακύμανση) μεταξύ των τιμών x_t και x_{t+k} εξαρτάται από την απόσταση k (αλλιώς η χρονική απόσταση k ονομάζεται χρονική υστέρηση) που απέχουν και όχι από τον χρόνο t (βλέπε [11]).

Η συνάρτηση αυτοσυσχέτισης (*ACF – Autocorellation Function*) και η συνάρτηση μερικής αυτοσυσχέτισης (*PACF – Partial Autocorellation Function*), όπως αναφέρθηκε αναλυτικά στην αντίστοιχη παράγραφο, είναι ιδιαίτερα χρήσιμες για την εξειδίκευση της μορφής της στοχαστικής διαδικασίας που «γέννησε» τις τιμές της

διερευνούμενης χρονολογικής σειράς και γίνεται χρήση των αντίστοιχων διαγραμμάτων τους ως προς το χρόνο.

3.1.1 Αυτοπαλίνδρομα Υποδείγματα AR(p)

Αυτοπαλίνδρομο υπόδειγμα τάξης p, που συμβολίζεται με *AR(p) (Autoregressive)*, ονομάζεται το υπόδειγμα που εκφράζεται από την ακόλουθη γραμμική σχέση (3.1.1.1):

$$X_t = a_0 + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + \varepsilon_t \qquad (3.1.1.1)$$

Όπου:

- Η τάξη p εκφράζει το μέγεθος της υστέρησης και η μεταβλητή ε_t εκφράζει τον λευκό θόρυβο.

- Η ορολογία «αυτοπαλίνδρομο» πηγάζει από το ότι η σχέση (3.1.1.1) εκφράζει ένα υπόδειγμα γραμμικής παλινδρόμησης όπου οι ερμηνευτικές (ανεξάρτητες) μεταβλητές ($X_{t-1}, X_{t-2}, .., X_{t-p}$) εκφράζουν τιμές της εξαρτημένης μταβλητής X_t σε προηγούμενες χρονικές στιγμές (με υστέρηση εως p).

- Με χρήση τελεστού υστέρησης L *(Lag)* το αυτοπαλίνδρομο υπόδειγμα τάξης p μπορεί να γραφεί στη μορφή της σχέσης (3.1.1.2):

$$\left(1 - \alpha_1 L + \alpha_2 L^2 + \ldots + \alpha_p L^{t-p}\right) X_t = \varepsilon_t \qquad (3.1.1.2)$$

Σε ένα αυτοπαλίνδρομο υπόδειγμα τάξης p, για τις αυτοσυνδιακυμάνσεις ισχύει η σχέση (3.1.1.3):

$$\gamma_0 = \sigma^2 + \alpha_1 \gamma_1 + \alpha_2 \gamma_2 + \ldots + \alpha_p \gamma_p \qquad (3.1.1.3)$$

Επίσης, για τις αυτοσυνδιακυμάνσεις τάξης s και τους αντίστοιχους συντελεστές αυτοσυσχέτισης ισχύουν οι σχέσεις (3.1.1.4) και (3.1.1.5):

$$\gamma_s = \alpha_1 \gamma_{s-1} + \alpha_2 \gamma_{s-2} + \ldots + \alpha_p \gamma_{s-p} \quad \gamma\iota\alpha\ s > 0 \qquad (3.1.1.4)$$

και

$$\rho_s = \alpha_1 \rho_{s-1} + \alpha_2 \rho_{s-2} + \ldots + \alpha_p \rho_{s-p} \quad \gamma\iota\alpha\ s > 0 \qquad (3.1.1.5)$$

Από την τελευταία σχέση προκύπτουν οι (p στο πλήθος) εξισώσεις *Yule – Walker* που παρατίθενται στη σχέση (3.1.1.6):

$$\begin{cases} \rho_1 = \alpha_1 + \alpha_2\rho_1 + \alpha_2\rho_2 \ldots + \alpha_p\rho_{p-1} \\ \rho_2 = \alpha_1\rho_1 + \alpha_2 + \alpha_3\rho_1 \ldots + \alpha_p\rho_{p-1} \\ \vdots \;=\; \vdots \;+\; \vdots \;+\; \ldots \;+\; \vdots \\ \rho_p = \alpha_1\rho_{p-1} + \alpha_2\rho_{p-2} + \alpha_3\rho_{p-3} \ldots + \alpha_p \end{cases} \quad (3.1.1.6)$$

Από τη λύση του $p \times p$ συστήματος της σχέσης (3.1.1.6) προκύπτουν οι τιμές των παραμέτρων του AR(p) υποδείγματος. Το παραπάνω σύστημα δύναται να γραφεί σε πινακοποιημένη μορφή όπως ακολούθως (3.1.1.7):

$$P_{1\times p} = P_{p\times p}A_{p\times 1} \quad (3.1.1.7)$$

όπου

$$P_{1\times p} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{bmatrix}, P_{p\times p} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \ldots & \rho_{p-1} \\ \rho_1 & 1 & \rho_3 & \ldots & \rho_{p-2} \\ \vdots & \vdots & \ldots & \vdots & \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \ldots & 1 \end{bmatrix} \text{και } A_{p\times 1} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix}$$

Ο υπολογισμός των παραμέτρων ενός AR(p) υποδείγματος, δηλαδή η εκτίμηση του υποδείγματος AR(p), εφόσον η τάξη αυτού έχει οριστεί με χρήση της δειγματικής συνάρτησης μερικής αυτοσυσχέτισης, δύναται να πραγματοποιηθεί με χρήση της σχέσης (3.1.1.6). Συγκεκριμένα, οι αυτοσυσχετίσεις $\rho_i, i = 1, 2, \ldots, p$ υπολογίζονται με χρήση του δείγματος (δηλαδή αντικαθιστώνται από τις δειγματικές αυτοσυσχετίσεις $\hat{\rho}_s$). Άλλες μέθοδοι υπολογισμού των παραμέτρων ενός στάσιμου αυτοπαλίνδρομου υποδείγματος είναι η μέθοδος ελαχίστων τετραγώνων με ή χωρίς συνθήκη και η βελτιωμένη μέθοδος των ροπών.

Οι μέθοδοι ελαχίστων τετραγώνων ασυμπτωτικά συμπίπτουν με τη μέθοδο μέγιστης πιθανοφάνειας. Όπως αναφέρθηκε, το υπόδειγμα AR(p) δίνεται από τη σχέση (3.1.1.1) και δύναται να θεωρηθεί ως ένα γραμμικό μοντέλο παλινδρόμησης με $p - 1$ ερμηνευτικές (στοχαστικές) μεταβλητές. Με χρήση της μεθόδου ελαχίστων τετραγώνων παράγονται εκτιμήσεις των παραμέτρων που έχουν την ιδιότητα των μεγάλων δειγμάτων και ακολουθούν προσεγγιστικά την κανονική κατανομή. Στην περίπτωση ενός δείγματος μεγέθους n παράγεται το σύστημα $n - p$ παρατηρήσεων (3.1.1.8):

$$\begin{cases} X_{p+1} = \alpha_0 + \alpha_1 X_p + \alpha_2 X_{p-1} + \alpha_3 X_{p-2} \ldots + \alpha_p X_1 + \varepsilon_{p+1} \\ X_{p+2} = \alpha_0 + \alpha_1 X_{p+1} + \alpha_2 X_p + \alpha_3 X_{p-1} \ldots + \alpha_p X_2 + \varepsilon_{p+2} \\ \vdots \quad = \quad \vdots \quad + \quad \vdots \quad + \quad \ldots \quad + \quad \vdots \\ X_n = \alpha_0 + \alpha_1 X_{n-1} + \alpha_2 X_{n-2} + \alpha_3 X_{n-3} \ldots + \alpha_p X_{n-p} + \varepsilon_n \end{cases} \quad (3.1.1.8)$$

Με χρήση πινάκων το παραπάνω σύστημα γράφεται ως ακολούθως (3.1.1.9):

$$X_{(n-p)\times 1} = X_{(n-p)\times(p+1)} A_{(p+1)\times 1} + E_{(n-p)\times 1} \quad (3.1.1.9)$$

όπου

$$X_{(n-p)\times 1} = \begin{bmatrix} X_{p+1} \\ X_{p+2} \\ \vdots \\ X_n \end{bmatrix}, \; X_{(n-p)\times(p+1)} = \begin{bmatrix} 1 & X_p & X_{p-1} & \ldots & X_1 \\ 1 & X_{p+1} & X_p & \ldots & X_2 \\ \vdots & \vdots & \ldots & \vdots & \\ 1 & X_{n-1} & X_{n-2} & .. & X_{n-p} \end{bmatrix},$$

$$E_{(n-p)\times 1} = \begin{bmatrix} \varepsilon_{p+1} \\ \varepsilon_{p+2} \\ \varepsilon_{p+3} \\ \vdots \\ \varepsilon_n \end{bmatrix} \text{ και } A_{(p+1)\times 1} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix}$$

Οι εκτιμήτριες ελαχίστων τετραγώνων δίνονται από τη γνωστή σχέση που παρατίθεται ακολούθως (3.1.1.10) (βλέπε [10]):

$$\hat{A} = \left(X'_{p\times p} X_{p\times p}\right)^{-1} X'_{p\times p} X_{1\times p} \quad (3.1.1.10)$$

Όλες οι παραπάνω μέθοδοι υπολογισμού των παραμέτρων βασίστηκαν στην γνώση της τάξης p. Αν υποτεθεί ότι η τάξη του AR(p) υποδείγματος, δεν είναι γνωστή, τότε το υπόδειγμα προσαρμόζεται στην αρχική χρονολογική σειρά και υπολογίζονται οι δειγματικοί συντελεστές μερικής αυτοσυσχέτισης τάξης p (αφού οι πληθυσμιακοί συντελεστές μερικής αυτοσυσχέτισης είναι άγνωστοι, χρησιμοποιούμε τους δειγματικούς συντελεστές μερικής αυτοσυσχέτισης). Ο συντελεστής μερικής αυτοσυσχέτισης θεωρούμενος ως ο συντελεστής παλινδρόμησης της τελευταίας μεταβλητής που υπεισέρχεται στο υπόδειγμα AR(p), εκφράζει τη στατιστική σημαντικότητα της τελευταίας μεταβλητής μέσα από τον αντίστοιχο έλεγχο υποθέσεων. Επομένως, αν ο συντελεστής μερικής αυτοσυσχέτισης είναι σημαντικά διάφορος του μηδενός, τότε το υπόδειγμα p τάξεως, είναι προτιμότερο από το υπόδειγμα τάξης $p-1$. Σε διαφορετική περίπτωση, αν δηλαδή ο συντελεστής

μερικής αυτοσυσχέτισης δεν είναι σημαντικά διάφορος του μηδενός, τότε το υπόδειγμα $p-1$ τάξεως, είναι προτιμότερο από το υπόδειγμα τάξης p. Επομένως ο υπολογισμός των συντελεστών μερικής αυτοσυσχέτισης έχει ιδιαίτερη πρακτική σημασία στην εκτίμηση της τάξης p του υπό διερεύνηση υποδείγματος (βλέπε [12]).

3.1.2 Υποδείγματα Κινητού Μέσου MA(q)

Κινητού μέσου υπόδειγμα τάξης q, που συμβολίζεται ως *MA(q) (Moving Average)*, ονομάζεται το υπόδειγμα που εκφράζεται από την ακόλουθη γραμμική σχέση (3.1.2.1):

$$X_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q} \tag{3.1.2.1}$$

όπου:

• Η τάξη q εκφράζει το μέγεθος της υστέρησης της μεταβλητής ε_t που εκφράζει τον λευκό θόρυβο.

• Η ορολογία «κινητός μέσος» πηγάζει από το ότι η σχέση (3.1.2.1) εκφράζει ένα σταθμισμένο άθροισμα των τιμών της ε_t.

Για ένα υπόδειγμα κινητών μέσων ισχύουν οι ακόλουθες σχέσεις (βλέπε [3]):

$$E(X_t) = \mu \tag{3.1.2.2}$$

$$Var(X_t) = \gamma_0 = \left(1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2\right)\sigma^2, \mu\varepsilon\ \theta_0 = 1 \tag{3.1.2.3}$$

$$\gamma_s = Cov(X_t, X_{t-s}) = \begin{cases} \left(\theta_s + \theta_{s+1}\theta_1 + \theta_{s+2}\theta_2 \ldots + \theta_q\theta_{q+s}\right)\sigma^2, & s = 1,2,\ldots,q \\ 0, & , \quad s > q \end{cases} \tag{3.1.2.4}$$

Επομένως, η μέση τιμή του υποδείγματος είναι σταθερή ως προς το χρόνο (βλέπε [11]). Η σχέση (3.1.2.1), με τη βοήθεια του τελεστή κινητού μέσου τάξης q γράφεται ως εξής (3.1.2.5):

$$X_t = (1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q)\varepsilon_t \tag{3.1.2.5}$$

ή ισοδύναμα

$$X_t = \Theta(L)\varepsilon_t \tag{3.1.2.6}$$

όπου $\Theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q$

Από την σχέση (3.1.2.6) είναι εμφανές ότι το MA(q) υπόδειγμα έχει μορφή γραμμικού μοντέλου με πεπερασμένη τάξη q. Για το υπόδειγμα MA(q) δεν

απαιτούνται περιορισμοί για τα θ_i ώστε να εξασφαλίζεται η στασιμότητα της χρονολογικής σειράς, αλλά απαιτούνται περιορισμοί για να εξασφαλισθεί η αντιστρεψιμότητα της χρονολογικής σειράς. Η αντιστρεψιμότητα μιας χρονολογικής σειράς αναφέρεται στο ότι η συνάρτηση αυτοσυσχέτισης δεν ορίζει μονοσήμαντα το MA(q) υπόδειγμα της χρονολογικής σειράς. Αποδεικνύεται ότι το MA(q) υπόδειγμα της χρονολογικής σειράς είναι αντιστρέψιμο, δηλαδή ότι το υπόδειγμα MA(q) μετατρέπεται σε υπόδειγμα AR(∞) (άπειρης τάξης), όταν οι ρίζες της εξίσωσης $\Theta(\omega) = 0, \omega \in C$ (από τη σχέση 3.1.2.6) βρίσκονται έξω από το μοναδιαίο κύκλο (δηλαδή οι πραγματικές ή φανταστικές λύσεις να έχουν μέτρο μεγαλύτερο της μονάδας).

Ο υπολογισμός των παραμέτρων ενός MA(q) υποδείγματος, δηλαδή η εκτίμηση του υποδείγματος συνολικά πραγματοποιείται με διάφορες μεθόδους που αναφέρονται ακολούθως. Σε ένα υπόδειγμα MA(q) η συνάρτηση αυτοσυσχέτισης τείνει να μηδενιστεί έπειτά από αριθμό υστερήσεων q. Αυτό, από στατιστικής σκοπιάς σημαίνει ότι οι συντελεστές αυτοσυσχέτισης κρίνονται στατιστικώς σημαντικοί για $s \leq q$ και μη στατιστικώς σημαντικοί για $s > q$. Ο έλεγχος σημαντικότητας των αυτοσυσχετίσεων πραγματοποιείται με όμοιο τρόπο με αυτόν που ακολουθείται στην περίπτωση των αυτοπαλίνδρομων υποδειγμάτων, δηλαδή με την αντίστοιχη μηδενική υπόθεση (Η μηδενική υπόθεση εκφράζει μηδενική αυτοσυσχέτιση) και την αντίστοιχη εναλλακτική υπόθεση (Η εναλλακτική υπόθεση εκφράζει μη μηδενική αυτοσυσχέτιση). Επομένως, στην περίπτωση που έχει καθοριστεί η τάξη q, οι παράμετροι του MA(q) υποδείγματος υπολογίζονται με χρήση της σχέσης (3.1.1.6) αντικαθιστώντας κατάλληλα τις δειγματικές αυτοσυσχετίσεις.

Στο σημείο αυτό πρέπει να τονιστεί ότι η μέθοδος ελαχίστων τετραγώνων δεν εφαρμόζεται στην περίπτωση των υποδειγμάτων κινητών μέσων, όπως στην περίπτωση των αυτοπαλίνδρομων υποδειγμάτων, καθώς η συνάρτηση στην οποία καταλήγει η αντίστοιχη διαδικασία με σκοπό την ελαχιστοποίηση της τιμής της δεν είναι γραμμική ως προς τις παραμέτρους του MA(q) υποδείγματος (βλέπε [11]).

Συγκεκριμένα, η συνάρτηση που περιγράφει τα τετράγωνα των σφαλμάτων δίνεται ακολούθως (3.1.2.7):

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(X_t - \mu - \theta_1 \varepsilon_{i-1} - \theta_2 \varepsilon_{i-2} - \ldots - \theta_q \varepsilon_{i-q}\right)^2 \qquad (3.1.2.7)$$

που δεν είναι γραμμική ως προς τις παραμέτρους $\theta_i, i = 1,2, \ldots, q$. Το ότι η παραπάνω συνάρτηση δεν είναι γραμμική ως προς τις παραμέτρους, γίνεται εύκολα αντιληπτό αν υποθέσουμε ότι ο αριθμός των υστερήσεων είναι $q = 1$, οπότε η σχέση (3.1.2.1) γίνεται ως εξής (3.1.2.8):

$$X_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} \qquad (3.1.2.8)$$

Στην περίπτωση ενός υποδείγματος κινητών μέσων MA(1) η σχέση (3.1.2.7) γίνεται (3.1.2.9):

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (X_t - \theta_1 \varepsilon_{i-1})^2 \quad (3.1.2.9)$$

Όπως είναι γνωστό, κάθε στάσιμη χρονολογική σειρά που περιγράφεται από ένα MA(1) υπόδειγμα δύναται να εκφραστεί και μέσω ενός AR(∞) υποδείγματος με χρήση της σχέσης (3.1.2.10):

$$(1 - \theta_1 L + \theta_1^2 L^2 + \theta_1^3 L^3 + \ldots) X_t = \varepsilon_t \quad (3.1.2.10)$$

Από τη σχέση (3.1.2.10) αντικαθιστώντας τον τελεστή υστέρησης έχουμε την ακόλουθη σχέση (3.1.2.11):

$$\varepsilon_t = X_t - \theta_1 X_{t-1} + \theta_1^2 X_{t-2} + \theta_1^3 X_{t-3} + \ldots \quad (3.1.2.10)$$

Άρα, η συνάρτηση που περιγράφει τα τετράγωνα των σφαλμάτων δίνεται ακολούθως (3.1.2.11):

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (\varepsilon_t = X_t - \theta_1 X_{t-1} + \theta_1^2 X_{t-2} + \theta_1^3 X_{t-3} + \ldots)^2$$

(3.1.2.11)

Από τη σχέση (3.1.2.11) είναι εμφανές ότι η μέθοδος καταλήγει σε μη γραμμική εξίσωση ως προς την παράμετρο θ, κάτι το οποίο σημαίνει ότι απαιτείται χρήση μη γραμμικών μεθόδων (βλέπε [11]).

3.2 Μεικτό Υπόδειγμα Χρονολογικής Σειράς

3.2.1 Το Μεικτό Υπόδειγμα ARMA (p,q)

Το μεικτό υπόδειγμα τάξης p, q (ARMA (p,q)) ορίζεται ως εξής από την ακόλουθη σχέση (3.2.1.1):

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

(3.2.1.1)

Ένα ARMA(p, q) υπόδειγμα χρονολογικής σειράς, μπορεί να θεωρηθεί:

α) Ως υπόδειγμα χρονολογικής σειράς AR(p) με σφάλματα που περιγράφονται από ένα υπόδειγμα χρονολογικής σειράς MA(q).

β) Ως υπόδειγμα χρονολογικής σειράς MA(q), με σφάλματα που περιγράφονται από ένα υπόδειγμα χρονολογικής σειράς AR(p).

Αφού το μεικτό υπόδειγμα ARMA (p,q) αποτελεί συνδυασμό των βασικών υποδειγμάτων AR(p) και MA(q), για την εκτίμηση των παραμέτρων ενός υποδείγματος ARMA(p,q), γίνεται συνδυασμός των μεθόδων εκτίμησης των παραμέτρων στα υποδείγματα AR(p) και MA(q) που αναφέρθηκαν στις προηγούμενες παραγράφους. Τα χαρακτηριστικά των ARMA(p,q) υποδειγμάτων που σχετίζονται με την συνάρτηση αυτοσυσχέτισης και με τη συνάρτηση μερικής αυτοσυσχέτισης περιγράφονται αναλυτικά στον ακόλουθο πίνακα 3.1.

Υπόδειγμα	Συνάρτηση Αυτοσυσχετίσης ACF	Συνάρτηση Μερικής Αυτοσυσχετίσης PACF
Λευκός θόρυβος	Μηδέν	Μηδέν
AR(p)	Φθίνουν προς το μηδέν από ρ_p	Μηδέν μετά το φ_{pp}
MA(q)	Μηδέν μετά το ρ_q	Φθίνει γεωμετρικά ή κυματιστά από το φ_{qq}
ARMA(p, q)	Φθίνει γεωμετρικά από το ρ_q	Φθίνει γεωμετρικά ή κυματιστά από το φ_{pp}

Πίνακας 3.1 Συνάρτηση αυτοσυσχέτισης και μερικής αυτοσυσχέτισης των ARMA(p, q) υποδειγμάτων *(Πηγή: [14] Χάλκος Γ. 2004. Σημειώσεις μαθήματος "Χρονολογικές σειρές και προβλέψεις")*

Ο πίνακας 3.1 παρέχει τη δυνατότητα μιας σύντομης εκτίμησης του κατάλληλου υποδείγματος που περιγράφει μια χρονολογική σειρά έχοντας υπόψη τις συναρτήσεις αυτοσυσχέτισης και μερικής αυτοσυσχέτισης, αλλά είναι απαραίτητοι οι έλεγχοι που παρουσιάζονται σε επόμενη παράγραφο για την ταυτοποίηση και εξειδίκευση του κατάλληλου υποδείγματος.

Σε πραγματικά δεδομένα χρονολογικών σειρών, η εκτίμηση των παραμέτρων ενός υποδείγματος ARMA(p, q) είναι εξαιρετικά επίπονη διαδικασία για να γίνει με πράξεις δίχως Η/Υ και γι'αυτό το λόγο γίνεται χρήση λογισμικών όπως είναι το E-

VIEWS και το MINITAB. Η εύρεση της τάξης ενός γραμμικού στάσιμου υποδείγματος ARMA (p, q) γίνεται με χρήση διαφόρων κριτηρίων, όπως του Akaike με τις παραλλαγές του, του Quenouille, κλπ. Το κριτήριο του Akaike υποθέτει την τάξη του υποδείγματος ίση με $i = 1, 2, 3, \dots, m$ και για κάθε διακριτή τάξη, υπολογίζεται η ποσότητα που δίνεται από τη σχέση (3.2.1.2):

$$AIC(K) = n \log {\sigma_\varepsilon}^2 + K \qquad \text{(3.2.1.2):}$$

όπου: n το πλήθος των παρατηρήσεων της χρονολογικής σειράς, ${\sigma_\varepsilon}^2$ η διακύμανση των σφαλμάτων και $K = t + 1$, όπου t οι παράμετροι του υποδείγματος συν ένα για τη μέση τιμή.

Ως τάξη του υπό διερεύνηση υποδείγματος ARMA (p, q) επιλέγεται εκείνο το i για το οποίο η ποσότητα AIC(K) λαμβάνει την ελάχιστη τιμή. Όταν το υπόδειγμα είναι της μορφής ARMA(p,q), οπότε ισχύει $K = p + q + 1$, τότε για κάθε συνδυασμό των τάξεων p και q υπολογίζεται η ποσότητα AIC και τελικά ως τάξη του υποδείγματος θεωρούνται, εκείνα τα p, q που ελαχιστοποιούν την ποσότητα AIC. Οπως γίνεται αντιληπτό, η διαδικασία επιλογής των κατάλληλων p και q είναι μια διαδικασία επίπονη που παίρνει αρκετό χρόνο, ειδικά όταν οι τάξεις p και q παρουσιάζουν υψηλές τιμές, ώστε να καλυφθούν όλοι οι δυνατοί συνδυασμοί.

Στην περίπτωση που ένα υπόδειγμα ARMA (p, q) είναι στάσιμο, τότε για την εξειδίκευση των τάξεων p και q ακολουθείται η διαδικασία που περιγράφηκε παραπάνω. Στην περίπτωση που δεν επιτυγχάνεται στασιμότητα της χρονολογικής σειράς, τότε ακολουθείται μια διαδικασία δημιουργίας μιας δευτερογενούς χρονολογικής σειράς που προκύπτει από τον κατάλληλο αριθμό διαφορών των αρχικών τιμών. Αυτή η διαδικασία λήψης κατάλληλου αριθμού διαφορών περιγράφεται αναλυτικά στην επόμενη παράγραφο.

3.2.2 Το Μεικτό Υπόδειγμα Πρώτων Διαφορών ARIMA (p,d,q)

Στην περίπτωση του μεικτού υποδείγματος ARMA(p,q) της προηγούμενης παραγράφου που εφαρμόζεται στις πρώτες διαφορές ($X_t - X_{t-1}$) ή σε μεγαλύτερης τάξης διαφορές της εξαρτημένης μεταβλητής, τότε το μεικτό υπόδειγμα ονομάζεται ARIMA (p,d,q). Οι πρώτες διαφορές ($X_t - X_{t-1}$) ή μεγαλύτερης τάξης διαφορές της εξαρτημένης μεταβλητής λαμβάνονται στην περίπτωση που η αρχική σειρά είναι μη στάσιμη λόγω μέσης τιμής που εξαρτάται από το χρόνο (δηλαδή όταν ισχύει $E(X_t) = t\mu$) ή λόγω διακύμανσης που εξαρτάται από το χρόνο (δηλαδή όταν ισχύει $V(X_t) = t\sigma^2$). Για παράδειγμα, η περίπτωση του τυχαίου περίπατου (Random Walk) δεν είναι στάσιμη χρονολογική σειρά:

$$X_t = X_{t-1} + e_t \qquad \text{(3.2.2.1)}$$

Όπως επίσης και στην περίπτωση που υπάρχει και σταθερός όρος, που τότε ονομάζεται τυχαίος περίπατος με μετατόπιση *(Random Walk with drift),* επίσης δεν είναι στάσιμη χρονολογική σειρά:

$$X_t = a + X_{t-1} + e_t \qquad (3.2.2.2)$$

Στις δύο προαναφερθείσες περιπτώσεις αν ληφθούν οι πρώτες διαφορές ($X_t - X_{t-1}$) τότε η σειρά που προκύπτει είναι στάσιμη καθώς από τη σχέση (3.2.2.1) λαμβάνεται η σχέση $X_t - X_{t-1} = e_t$, δοθέντος ότι e_t είναι λευκός θόρυβος, δηλαδή στάσιμη χρονολογική σειρά. Η διαφορά $X_t - X_{t-1}$ μπορεί να γραφεί με τη βοήθεια κατάλληλου τελεστή που εκφράζει τον αριθμό των διαφορών. Για παράδειγμα ισχύει:

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

και

$$X_t - X_{t-2} = (1 - L)^2 X_t = \nabla^2 X_t$$

και γενικότερα $\nabla^d = (1 - L)^d$ είναι ο συντελεστής διαφορών d τάξης.

Η γενική μορφή ενός μεικτού υποδείγματος διαφορών ARIMA (p,d,q) δίνεται από τη σχέση, αν στη θέση των X_t τοποθετηθούν οι αντίστοιχες διαφορές $X_t - X_{t-d} = (1 - L)^d X_t = \nabla^d X_t$. Επομένως η γενική μορφή ενός μεικτού υποδείγματος διαφορών ARIMA (p,d,q) δίνεται από τη σχέση:

$$X_t = (1 + \alpha_1)X_{t-1} + (\alpha_2 - \alpha_1)X_{t-2} + \ldots + (\alpha_p - \alpha_{p-1})X_{t-p} - \alpha_p X_{t-p-1} + u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2} - \ldots - \theta_q u_{t-q}$$

όπου u_t είναι ο διαταρακτικός όρος της χρονολογικής σειράς των αντίστοιχων διαφορών.

3.2.3 Το Μεικτό Εποχικό Υπόδειγμα SARIMA (p,d,q) (P,D,Q)[s]

Αν το μεικτό υπόδειγμα της προηγούμενης παραγράφου εφαρμόζεται στην εποχική συνιστώσα, τότε το νέο υπόδειγμα ονομάζεται Μεικτό Εποχικό Υπόδειγμα και συμβολίζεται ως *SARIMA (p,d,q)×(P,D,Q)[s].* Οι ιδιότητες ενός τέτοιου υποδείγματος είναι όμοιες με αυτές του απλού ARIMA και βασική επιδίωξη είναι η επίτευξη στασιμότητας μιας τέτοιας σειράς.

Συχνά, οι χρονολογικές σειρές πραγματικών δεδομένων εκφράζονται μέσα από μια εποχική συνιστώσα που επαναλαμβάνεται ανά συγκεκριμένο αριθμό παρατηρήσεων. Για παράδειγμα, στην περίπτωση μηνιαίων παρατηρήσεων ισχύει

ότι $s = 12$ (12 μήνες σε 1 έτος), για τις τριμηνιαίες παρατηρήσεις ισχύει $s = 4$ (4 τρίμηνα σε 1 έτος), κτλ. Προκειμένου να αντιμετωπιστεί αυτού του είδους η εποχικότητα, τα υποδείγματα ARIMA έχουν γενικευθεί κατάλληλα και κατασκευάστηκαν μοντέλα SARIMA *(Seasonal Autoregressive Integrated Moving Average Model).*

Επομένως, το μοντέλο SARIMA (p,d,q)×(P,D,Q)[s] ή διαφορετικά το εποχικό ARIMA εκφράζει ένα μοντέλο ARIMA, όπως αυτό περιγράφεται στην προηγούμενη παράγραφο με την εποχική συνιστώσα να χαρακτηρίζεται από τις παραμέτρους με τα κεφαλαία γράμματα (P,D,Q) καθώς το *P* αναφέρεται στο αυτοπαλίνδρομο μέρος της εποχικής συνιστώσας, το *Q* αναφέρεται στο μέρος κινητών μέσων της εποχικής συνιστώσας και το *D* αναφέρεται στον κατάλληλο αριθμό των διαφορών ώστε η εποχική συνιστώσα να εκφράζεται από μια στάσιμη χρονολογική σειρά.

3.2.4 Το Μεικτό Υπόδειγμα ARMAX (p,q)

Η εισαγωγή εξωγενών μεταβλητών $Z_i, i = 1,2, \dots, k$ σε ένα υπόδειγμα ARMA (p,q), παράγει το υπόδειγμα ARMAX(p,q), το οποίο, συμπεριλαμβάνει την επίδραση από μία ή περισσότερες εξωγενείς μεταβλητές. Η γενική μορφή του μοντέλου ARMAX (p,q) είναι:

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + \sum_{k=1}^{r} \beta_k Z_{tk} + \varepsilon_t \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

Οι συντελεστές $\beta_i, i = 1,2, \dots, r$ έχουν ουσιαστικά την ίδια ερμηνεία με εκείνη που συναντάται στην πολλαπλή γραμμική παλινδρόμηση, δηλαδή, η αναμενόμενη επίδραση που έχει η αύξηση μιας μονάδας σε κάποια ερμηνευτική μεταβλητή από τις $Z_i, i = 1,2, \dots, r$, όταν οι υπόλοιπες παραμένουν σταθερές, είναι του μεγέθους $\beta i, i = 1,2, \dots, r$. Τα υποδείγματα ARMAX παρουσιάζουν τις ίδιες απαιτήσεις στασιμότητας όπως και τα υποδείγματα ARMA. Στην περίπτωση που η χρονολογική σειρά δεν είναι στάσιμη, τότε μετατρέπεται σε στάσιμο υπόδειγμα ARIMA χρησιμοποιώντας τις κατάλληλες διαφορές. Στην περίπτωση αυτή, οι εξωγενείς μεταβλητές $Z_i, i = 1,2, \dots, r$ εισέρχονται σε ένα στάσιμο μοντέλο ARIMAX (p, D, q) και η ερμηνεία των συντελεστών $\beta_i, i = 1,2, \dots, r$ ισχύει και για τις αντίστοιχες διαφορές μεταξύ της τρέχουσας ερμηνευτικής μεταβλητής και της χρονικά υστερημένης ερμηνευτικής μεταβλητής. Επίσης, στην περίπτωση χρήσης εξωγενών μεταβλητών στα υποδείγματα ARIMAX (p, D, q), δύναται να γίνει χρήση της εποχικής συνιστώσας

όπως στα υποδείγματα SARIMA, έτσι δημιουργούνται τα υποδείγματα της μορφής SARIMAX (p, D, q) (p, d, q)s (βλέπε [7]). Αν μια χρονολογική σειρά X_t είναι στάσιμη, τότε το υπόδειγμα SARIMAX (p, D, q) (p, d, q)s έχει τη μορφή που δίνεται από τη σχέση

$$a(L)A(L)x_t = z_t'\beta + \theta(L)\Theta(L)\varepsilon_t$$

όπου $\alpha(L)$ και $\theta(L)$ είναι οι περιγραφές του αυτοπαλίνδρομου μέρους και του μέρους κινητών μέσων αντιστοίχως, καθώς επίσης $A(L)$ και $\Theta(L)$ είναι πολυώνυμες εποχιακές συναρτήσεις υστέρησης των προαναφερθέντων στοιχείων του υποδείγματος.

3.3 Αξιολόγηση Υποδείγματος Χρονολογικής Σειράς

Για την ταυτοποίηση του κατάλληλου αριθμού υστερήσεων *(lags)* του υπό διερεύνηση υποδείγματος καθώς επίσης και για την επιλογή των αριθμών p και q ώστε η σειρά να είναι στάσιμη, έγινε αναφορά σε προηγούμενες παραγράφους. Η στασιμότητα του υπό διερεύνηση υποδείγματος αποτελεί απαραίτητη και αναγκαία προϋπόθεση ώστε ένα υπόδειγμα να είναι δεκτό για πρόβλεψη μελλοντικών τιμών. Το κριτήριο στασιμότητας της χρονολογικής σειράς «Stationary R^2» εκφράζει το αν είναι στάσιμη ή όχι η χρονολογική σειρά και χρησιμοποιείται για τον αντίστοιχο έλεγχο. Οι τιμές του κριτηρίου που είναι κοντά στην μονάδα εκφράζουν στάσιμη χρονολογική σειρά.

Στην διαδικασία διερεύνησης και ταυτοποίησης του κατάλληλου υποδείγματος για την περιγραφή μιας χρονολογικής σειράς, υπάρχει η περίπτωση να βρεθούν περισσότερα του ενός υποδείγματος που να είναι στάσιμα και να περιγράφουν τις τιμές μιας χρονολογικής σειράς. Σε αυτή την περίπτωση το βέλτιστο υπόδειγμα επιλέγεται έχοντας ως βασικό κριτήριο τον βαθμό προσαρμογής του υποδείγματος στα πραγματικά δεδομένα της αρχικής χρονολογικής σειράς. Ο βαθμός προσαρμογής ενός οποιουδήποτε υποδείγματος στα πραγματικά δεδομένα αξιολογείται με τη χρήση στατιστικών κριτηρίων όπως είναι τα ακόλουθα:

1. Το στατιστικό κριτήριο «R^2» που εκφράζει το ποσοστό της άγνωστης διασποράς της εξαρτημένης μεταβλητής που εξηγείται από τη γνώση των ερμηνευτικών μεταβλητών και χρησιμοποιείται στην γραμμική

παλινδρόμηση. Το γεγονός ότι οι ερμηνευτικές μεταβλητές είναι τιμές της εξαρτημένες σε προηγούμενες χρονικές στιγμές, δεν επηρεάζει τη λειτουργία του κριτηρίου. Τιμές κοντά στην μονάδα εκφράζουν υπόδειγμα με ικανοποιητική προσαρμογή.

2. Το στατιστικό κριτήριο *«Root Mean Squared Error» (RMSE)* εκφράζει την τυπική απόκλιση των σφαλμάτων μεταξύ πραγματικών και εκτιμώμενων τιμών και ως βέλτιστο θεωρείται το υπόδειγμα που αντιστοιχεί στην ελάχιστη τιμή. Το στατιστικό κριτήριο *«Root Mean Squared Error» (RMSE)* υπολογίζεται από τον τύπο:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}$$

3. Το στατιστικό κριτήριο *«Mean Absolute Error» (MAE)* που εκφράζει την μέση τιμή των απόλυτων τιμών των σφαλμάτων μεταξύ πραγματικών και εκτιμώμενων τιμών και ως βέλτιστο θεωρείται το υπόδειγμα που αντιστοιχεί στην ελάχιστη τιμή. Το στατιστικό κριτήριο *«Mean Absolute Error» (MAE)* υπολογίζεται από τον τύπο:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i|$$

Ο έλεγχος Ljung-Box Q εκφράζει τον βαθμό καταλληλότητας του υποδείγματος ανάμεσα σε πολλά που επιτυγχάνουν ικανοποιητικές τιμές στα παραπάνω στατιστικά κριτήρια καθώς ελέγχει αν κάποια ομάδα αυτοσυσχετίσεων της χρονολογικής σειράς των καταλοίπων είναι διαφορετική από μηδέν (βλέπε [4]).

ΚΕΦΑΛΑΙΟ 4

Πρόβλεψη

Όπως αναφέρθηκε και αρχικά, τελικός και κύριος στόχος της ανάλυσης μίας χρονοσειράς είναι η πραγματοποίηση προβλέψεων (*Forecasting*) για τις μελλοντικές τιμές της, καθώς και ο προσδιορισμός της αβεβαιότητας αυτών των προβλέψεων.

Η ανάλυση μιας χρονολογικής σειράς δεδομένων μιας μεταβλητής στοχεύει στην πρόβλεψη των μελλοντικών τιμών της αντίστοιχης μεταβλητής, με βάση, κυρίως τις παρελθούσες τιμές της μεταβλητής αυτής. Επομένως, γίνεται η υπόθεση ότι η παρελθοντική συμπεριφορά της χρονολογικής σειράς θα είναι όμοια στο παρόν και στο μέλλον χωρίς ιδιαίτερες διαφοροποιήσεις. Αυτή η υπόθεση αποτελεί μειονέκτημα της προβλεπτικής ικανότητας των χρονολογικών σειρών διότι άλλοτε είναι αληθής και άλλοτε δεν ισχύει. Η χρήση ερμηνευτικών μεταβλητών στις χρονολογικές σειρές επιλύει το πρόβλημα αυτό μερικώς, καθώς υπάρχουν παράγοντες που επιδρούν στη διαμόρφωση των τιμών μιας χρονολογικής σειράς από την χρονική στιγμή κατασκευής του υποδείγματος αλλά υπάρχουν και άλλοι που πιθανόν να μην έχουν συμπεριληφθεί στο υπόδειγμα αυτό. Επομένως, ίσως η βασικότερη υπόθεση που γίνεται στο πλαίσιο της ανάλυσης των χρονολογικών σειρών είναι ότι το πρότυπο συμπεριφοράς της, θα είναι όμοιο στο μέλλον. Αυτό σημαίνει ότι αν οι εξωτερικοί παράγοντες που διαμορφώνουν σημαντικά τις τιμές μιας χρονολογικής σειράς παραμένουν σταθεροί, τότε η χρονολογική σειρά δεν θα παρουσιάζει έντονες διαφοροποιήσεις στις μελλοντικές τιμές της και η πρόβλεψη των μελλοντικών τιμών, μέσω του υποδείγματος αυτού, θα είναι ικανοποιητική.

Ένα άλλο θέμα που πρέπει να λαμβάνεται υπόψιν στις προβλέψεις που βασίζονται στις χρονολογικές σειρές είναι αυτού του χρονικού μήκους των προβλεπόμενων τιμών, δηλαδή το πόσο απέχει χρονικά η ταυτοποίηση του υποδείγματος από τις προβλεπόμενες τιμές που παράγει. Ένα χαρακτηριστικό των χρονολογικών σειρών είναι ότι οι προβλέψεις αυτών είναι ακριβείς για σύντομο χρονικό διάστημα στο μέλλον. Ουσιαστικά, οι προβλεπόμενες μελλοντικές τιμές μιας χρονολογικής σειράς αποτελούν σύνθεση των προβλέψεων των κύριων συνιστωσών αυτής, δηλαδή της τάσης, της εποχικότητας, της κυκλικότητας και της άρρυθμης μεταβολής. Οι τρεις πρώτες κύριες συνιστώσες δύναται να εντοπιστούν ως μέρη του συνολικού υποδείγματος, επομένως να μοντελοποιηθούν ώστε να παράγουν προβλέψεις. Αντίθετα από τις τρεις αναφερθείσες κύριες συνιστώσες, η άρρυθμη μεταβολή είναι

εκείνη η συνιστώσα που δεν είναι ουσιαστικά δυνατό να προβλεφθεί. Επομένως, αν η συνεισφορά της άρρυθμης συνιστώσας είναι υψηλή στο τελικό υπόδειγμα της χρονολογικής σειράς, τότε μειώνεται η προβλεπτική ικανότητα του υποδείγματος.

Στο κεφάλαιο αυτό θεωρείται ότι τυχόν συντελεστές τάσης-εποχικότητας είναι γνωστοί, και επιλεγμένο υπόδειγμα ARIMA για την ερμηνεία μιας χρονοσειράς είναι σωστό και με γνωστές παραμέτρους. Στην πραγματικότητα κάτι τέτοιο δεν ισχύει απόλυτα, αφού πάντα υπάρχει αβεβαιότητα για το επιλεχθέν μοντέλο καθώς με κατάλληλα κριτήρια επιτυγχάνεται η αξιοπιστία του μοντέλου. Επίσης οι παράμετροι της τάσης, της εποχικότητας και του ARIMA μοντέλου δεν είναι γνωστές, αλλά εκτιμημένες. Για μεγάλα δείγματα χρονοσειρών όμως προσεγγιστικά θεωρείται ότι οι εκτιμήσεις διαφέρουν ελάχιστα από τις πραγματικές τιμές των παραμέτρων και η αβεβαιότητα στις εκτιμήσεις τους συνεισφέρει ελάχιστα στην αβεβαιότητα των προβλέψεων.

4.1 Πρόβλεψη Ελαχίστου Μέσου Τετραγωνικού Σφάλματος

Έστω μία χρονοσειρά $\{X_n\}$ όπου t η χρονική περίοδος κατά τη διάρκεια της οποίας έχουμε την πραγματοποίηση των τιμών $X_1, X_2, X_3, \dots, X_{n-1}, X_n$. Στόχος είναι η πρόβλεψη των τιμών της X_{n+l} που θα συμβούν σε l χρονικές στιγμές στο μέλλον. Θα χρησιμοποιηθεί ο συμβολισμός για την πρόβλεψη της χρονοσειράς $X_n(l)$ όπου η πρόβλεψη γίνεται σε χρόνο n για μετά από l χρονικές στιγμές *(lead l)*. Το πρόβλημα είναι ο προσδιορισμός της βέλτιστης πρόβλεψης, δεδομένων των τ.μ. $X_1, X_2, \dots, X_n$.

Η εκτίμηση της πρόβλεψης ορίζεται ως

$$\widehat{X_n}(l) = E(X_{n+l}|X_1, X_2, \dots, X_n)$$

Επιθυμητές ιδιότητες για καλή πρόβλεψη αποτελούν η αμεροληψία *(unbiasedness)* και αποδοτικότητα *(efficiency)*. Από την υπόθεση της αμεροληψίας έχουμε ότι :

$$E(X_{n+l} - \hat{X}_n(l)) = 0$$

ενώ η διασπορά ορίζεται ως

$$Var\big(e_t(l)\big) = Var\big(X_{n+l} - \hat{X}_n(l)\big)$$

Συνδυάζοντας τα παραπάνω, ως βέλτιστη πρόβλεψη θεωρείται εκείνη που ελαχιστοποιεί το μέσο τετραγωνικό σφάλμα πρόβλεψης *(Minimum MSE Forecast)* για κάθε βήμα πρόβλεψης l.

$$E\left(\left(X_{n+l} - \hat{X}_n(l)\right)^2\right)$$

4.1.1 Απλές Τεχνικές Πρόβλεψης – Αιτιοκρατική Τάση (Deterministic Trend)

Η πιο απλή μέθοδος πρόβλεψης αφορά τη χρονική σειρά που αποτελείται μόνο από χρονικές τάσεις *(trends),* που είτε είναι γνωστές είτε χρειάζεται να εκτιμηθούν, δηλαδή,

$$X_t = m_t + Y_t$$

όπου m_t είναι μία αιτιοκρατική τάση σε συνάρτηση με τον χρόνο t και Y_t είναι ο λευκός θόρυβος.

Η πρόβλεψη γίνεται με την επέκταση (extrapolation) του αιτιοκρατικού όρου σε μελλοντικούς χρόνους, δηλαδή η πρόβλεψη του X_{t+l} είναι

$$X_t(l) = E(m_{t+l} + Y_{t+l}|X_t, X_{t-1}, \dots, X_1) = m_{t+l}$$

Το σφάλμα της πρόβλεψης είναι

$$e_t(l) = Y_{t+l}$$

Άρα το $e_t(l)$ είναι λευκός θόρυβος με διασπορά σ_Y^2.

(βλέπε [2]- Jonathan D. Cryer, Kung-Sik Chan – Time Series Analysis: With Applications in R (Springer Texts in Statistics) 2nd Edition)

4.2 Πρόβλεψη Στάσιμων Χρονολογικών Σειρών με Γραμμικά Μοντέλα

Σε αυτό το κεφάλαιο η χρονική σειρά για την οποία θα γίνουν προβλέψεις θεωρείται στάσιμη ή ότι έχει μετασχηματιστεί σε στάσιμη με κάποια από τις μεθόδους που αναλύθηκαν σε προηγούμενο κεφάλαιο. Τα γραμμικά μοντέλα στάσιμων χρονολογικών σειρών που μελετήθηκαν είναι τα αυτοπαλίνδρομα AR, του κινητού μέσου MA και τα μεικτά ARMA. Αυτά τα μοντέλα θα χρησημοποιηθούν για να γίνουν οι προβλέψεις.

4.2.1 Πρόβλεψη με Αυτοπαλίνδρομα Μοντέλα

<u>AR(1) Μοντέλο</u>

Έστω η στάσιμη χρονοσειρά AR(1), δηλαδή, $X_t = \alpha_1 X_{t-1} + \varepsilon_t$ και έχουμε τις τιμές $(X_1, X_2, \dots, X_n)$ μέχρι και το χρόνο n. Οι προβλέψεις για τους επόμενους χρόνους θα είναι:

$$\hat{X}_n(1) = E(X_{n+1}|X_1, X_2, \dots, X_n) = E(\alpha_1 X_n + \varepsilon_{n+1}|X_1, X_2, \dots, X_n) = \alpha_1 X_n + E(\varepsilon_{n+1})$$
$$= \alpha_1 X_n$$

$$\hat{X}_n(2) = E(X_{n+2}|X_1, X_2, \dots, X_n) = E(\alpha_1 X_{n+1} + \varepsilon_{n+2}|X_1, X_2, \dots, X_n)$$
$$= \alpha_1 E(X_{n+1}|X_1, X_2, \dots, X_n) + E(\varepsilon_{n+2}) = \alpha_1 \hat{X}_n(1) = a_1^2 X_n$$

…

$$\hat{X}_n(l) = E(X_{n+l}|X_1, X_2, \dots, X_n) = E(\alpha_1 X_{n+l-1} + \varepsilon_{n+l-1}|X_1, X_2, \dots, X_n)$$
$$= \alpha_1 \hat{X}_n(l-1) = a_1^l X_n$$

Αν γενικότερα έχουμε $Y_t = d_t + X_t$ όπου X_t είναι μία στάσιμη AR(1) χρονοσειρά και d_t είναι η τάση, τότε προκύπτει $X_t = Y_t - d_t$, επομένως

$$\hat{Y}_n(l) = E(Y_{n+l}|X_1, X_2, \dots, X_n) = d_{n+l} + \hat{X}_n(l) = d_{n+l} + a_1^l X_n$$
$$= d_{n+l} + a_1^l (Y_n - d_n)$$

Παρατηρούμαι ότι όσο αυξάνεται το l η πρόβλεψη $\hat{Y}_n(l)$ συγκλίνει στην τάση d_{n+l}.

Το σφάλμα της πρόβλεψης για προήγηση (lead) l είναι

$$\hat{e}_n(l) = \sum_{i=0}^{n-1} a_1^i \varepsilon_{n+l-i}$$

Ως γραμμική συνάρτηση του θορύβου, η $\hat{e}_n(l)$ ακολουθεί κανονική κατανομή $E\left(\hat{e}_n(l)\right) = 0$ και με διασπορά

$$\sigma_e^2(l) = \sum_{i=0}^{l-1} a_1^{2i} Var(\varepsilon_{n+l-i}) = \sigma^2 \sum_{i=0}^{l-1} (a_1^2)^i = \sigma^2 \frac{1 - a_1^{2l}}{1 - a_1^2} \to \frac{\sigma^2}{1 - a_1^2} = \gamma(0)$$
$$= Var(X_t)$$

Δηλαδή αρχικά η διασπορά είναι μικρή $\left(Var\left(\hat{e}_n(1)\right) = \sigma^2\right)$ και αυξάνεται όσο απομακρυνόμαστε στο μέλλον συγκλίνοντας στην στάσιμη διασπορά της χρονοσειράς X. Το διάστημα πρόβλεψης συντελεστού $1 - a$, για την Y_{n+l}, είναι

$$\left(d_{n+l} + a_1^l X_n - \sigma \sqrt{\frac{1 - a_1^{2l}}{1 - a_1^2}} z_{\{a/2\}}, d_{n+l} + a_1^l X_n + \sigma \sqrt{\frac{1 - a_1^{2l}}{1 - a_1^2}} z_{\{a/2\}} \right)$$

Για όλο το υπόλοιπο κεφάλαιο βλέπε *[16]- Μπούτσικας Μιχαήλ. Σημειώσεις "Ανάλυση Χρονολογικών Σειρών"*

4.2.2 Πρόβλεψη με Μοντέλα Μέσου Όρου

MA(1) Μοντέλο

Έστω η στάσιμη χρονοσειρά κινητού μέσου MA(1) , $X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$ και έχουμε τις τιμές $(X_1, X_2, \dots, X_n)$ μέχρι και το χρόνο n. Οι προβλέψεις για τους επόμενους χρόνους θα είναι:

$$\hat{X}_n(1) = E(X_{n+1}|X_1, X_2, \dots, X_n) = E(\varepsilon_{n+1} + \theta_1 \varepsilon_n|X_1, X_2, \dots, X_n)$$
$$= E(\varepsilon_{n+1}) + \theta_1 E(\varepsilon_n|X_1, X_2, \dots, X_n) = \theta_1 \varepsilon_n$$

διότι $E(\varepsilon_n|X_1, X_2, \dots, X_n) = \varepsilon_n$ αφού

$$\varepsilon_n = (1 - \theta_1 B)^{-1} X_n = \sum_{i=0}^{\infty} (-\theta_1)^i X_{n-i} \approx \sum_{i=0}^{n-1} (-\theta_1)^i X_{n-i}$$

(κρατάμε τους n πρώτους όρους, οι υπόλοιποι έχουν σχεδόν μηδενική συνεισφορά). Οι προβλέψεις μεγαλύτερης προήγησης είναι μηδενικές διότι

$$\hat{X}_n(l) = E(X_{n+l}|X_1, X_2, \dots, X_n) = E(\varepsilon_{n+l} + \theta_1 \varepsilon_{n+l-1}|X_1, X_2, \dots, X_n) = 0, l \geq 2$$

Αν πάλι θεωρήσουμε γενικότερα $Y_t = d_t + X_t$ όπου X_t είναι μία στάσιμη MA(1) χρονοσειρά και d_t είναι η τάση, τότε $X_t = Y_t - d_t$, και

$$\hat{Y}_n(l) = E(Y_{n+l}|X_1, X_2, \dots, X_n) = d_{n+l} + \hat{X}_n(l) = \begin{cases} d_{n+1} + \theta_1 \varepsilon_n, & l = 1 \\ d_{n+l}, & l \geq 2 \end{cases}$$

Το σφάλμα της πρόβλεψης προήγησης l είναι

$$\hat{e}_n(1) = Y_{n+1} - \hat{Y}_n(l) = Y_{n+l} - d_{n+l} = X_{n+l}, \qquad l \geq 2$$

Τα σφάλματα των προβλέψεων είναι κανονικά με μέσες τιμές 0 και διασπορές ${\sigma_\varepsilon}^2(1) = \sigma^2$ και ${\sigma_\varepsilon}^2(l) = \sigma^2\left(1 + {\theta_1}^2\right), l \geq 1$. Εδώ η αύξηση της διασποράς γίνεται απότομα, ενώ στην περίπτωση της AR(1) ήταν σταδιακή. Το διάστημα πρόβλεψης συντελεστού $1 - a$, για την Y_{n+l} , είναι

$$\hat{Y}_n(l) - \sigma_\varepsilon(l) z_{\{a/2\}}, \hat{Y}_n(l) + \sigma_\varepsilon(l) z_{\{a/2\}}$$

Όμοια γίνεται η πρόβλεψη για τα μεγαλύτερα τάξης MA υποδείγματα.

4.2.3 Πρόβλεψη σε τυχαίο περίπατο

Εδώ ισχύει ότι $X_t = X_{t-1} + \varepsilon_t$ άρα

$$\hat{X}_n(1) = E(X_{n+1}|X_1, X_2, \dots, X_n) = E(X_n + \varepsilon_{n+1}|X_1, X_2, \dots, X_n) = X_n$$

για $l \geq 2$:

$$\hat{X}_n(l) = E(X_{n+l}|X_1, X_2, \dots, X_n) = E(X_{n+l-1} + \varepsilon_{n+l}|X_1, X_2, \dots, X_n) = \hat{X}_n(l-1) = \cdots$$
$$= \hat{X}_n(1) = X_n$$

Αν $Y_t = d_t + X_t$ τότε

$$\hat{Y}_n(l) = E(d_{n+l} + X_{n+l}|X_1, X_2, \dots, X_n) = d_{n+l} + X_n$$

με σφάλμα πρόβλεψης

$$\hat{e}_n(l) = Y_{n+l} - \hat{Y}_n(l) = Y_{n+l} - d_{n+l} - X_n = X_{n+l} - X_n = \varepsilon_{n+l} + \cdots + \varepsilon_{n+1}$$

το οποίο έχει διασπορά $Var\big(\hat{e}_n(l)\big) = l\sigma^2$ και επομένως, αντίθετα με τα AR(1) και MA(1) η αβεβαιότητα στην πρόβλεψη μεγαλώνει όσο απομακρυνόμαστε στο μέλλον. Το διάστημα πρόβλεψης συντελεστού $1-\alpha$, για την Y_{n+l} , είναι

$$\left(d_{n+l} + X_n - \sqrt{l}\sigma z_{\{\alpha/2\}}\,, d_{n+l} + X_n + \sqrt{l}\sigma z_{\{\alpha/2\}}\ \right)$$

4.2.4 Πρόβλεψη με βάση το γενικό υπόδειγμα ARMA(p,q)

Για τη στάσιμη χρονοσειρά ARMA(p,q) ισχύει:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_p X_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \dots - \theta_q\varepsilon_{t-q}$$

άρα η πρόβλεψη της X για το χρόνο $n+l$ όταν είμαστε στο χρόνο n, έχοντας γνωστές τις τιμές $(X_1, X_2, \dots, X_n)$ είναι

$$\hat{X}_n(l) = E\big(X_{n+l}\big|(X_1, X_2, \dots, X_n)\big)$$

$$= E\left(\alpha_1 X_{n+l-1} + \cdots + \alpha_p X_{n+l-p} + \varepsilon_{n+l} + \theta_1\varepsilon_{n+l-1} + \cdots + \theta_q\varepsilon_{n+l-q}\Big|(X_1, X_2, \dots, X_n)\right)$$
$$= E(\alpha_1 X_{n+l-1}|\alpha_1 X_{n+l-1}) + \cdots + E\left(\alpha_p X_{n+l-p}\Big|(X_1, X_2, \dots, X_n)\right)$$
$$+ E\big(\varepsilon_{n+l}\big|(X_1, X_2, \dots, X_n)\big) + E\big(\theta_1\varepsilon_{n+l-1}\big|(X_1, X_2, \dots, X_n)\big) + \cdots$$
$$+ E\left(\theta_q\varepsilon_{n+l-q}\Big|(X_1, X_2, \dots, X_n)\right)$$

Όμως επειδή $E\big(\varepsilon_t\big|(X_1, X_2, \dots, X_n)\big) = \varepsilon_t I\ \ (t \leq n)$, δηλαδή $E\big(\varepsilon_t\big|(X_1, X_2, \dots, X_n)\big) = 0$ για $t > n$ και $E\big(\varepsilon_t\big|(X_1, X_2, \dots, X_n)\big) = \varepsilon_t$ για $t \leq n$, τότε η παραπάνω σχέση γίνεται

$$\hat{X}_n(l) = \alpha_1\hat{X}_n(l-1) + \cdots + \alpha_p\hat{X}_n(l-p) + \theta_1\varepsilon_{n+l-1}I(l \leq 1) + \cdots + \theta_q\varepsilon_{n+l-q}I(l \leq q)$$

όπου $\hat{X}_n(s) = X_{n+s}$ αν $s \leq 0$ και $I(\alpha \leq \beta) = 1$ ή 0 ανάλογα αν $\alpha \leq \beta$ ή όχι. Άρα οι προβλέψεις υπολογίζονται αναδρομικά για $l = 1,2,\dots$ αφού πρώτα

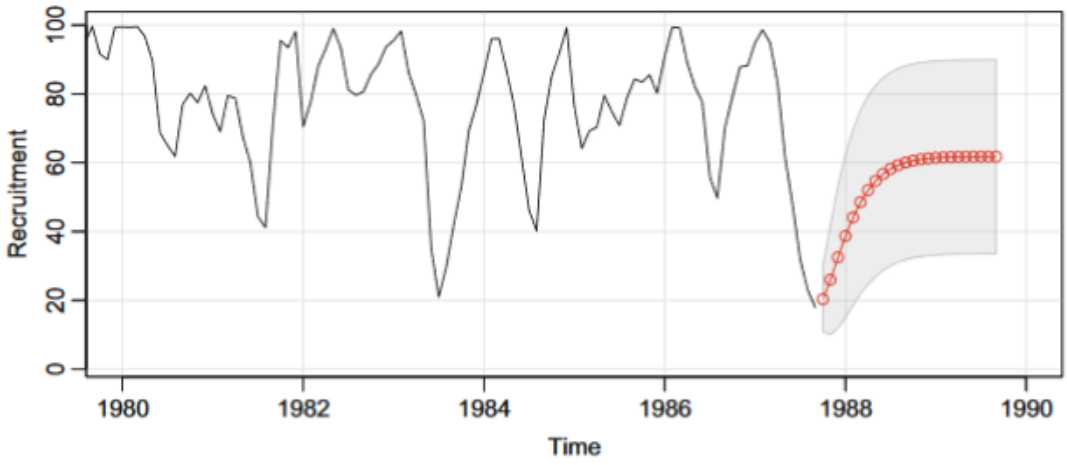
υπολογιστούν αναδρομικά όπως και σε προηγούμενη περίπτωση (κατά τις εκτιμήσεις ελαχίστων τετραγώνων), οι τιμές του θορύβου από τις

$$\varepsilon_t = X_t - \left(\alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}\right), \qquad t = 1,2,\dots,n$$

(θεωρώντας π.χ. ότι $X_t = \bar{X}, t \leq 0$ και $\varepsilon_t = 0, \ \ t \leq 0$). Σημειώνεται ότι για $l > q$ οι όροι που περιέχουν τα ε_i μηδενίζονται στις παραπάνω αναδρομικές σχέσεις και έτσι γράφονται απλούστερα

$$\hat{X}_n(l) = a_1 \hat{X}_n(l-1) + \cdots + a_p \hat{X}_n(l-p), \qquad l > q$$

οι οποίες είναι οι εξισώσεις Yule-Walker κατά τον υπολογισμό των αυτοσυσχετίσεων ενός ARMA(p,q) υποδείγματος. Επομένως, για προηγήσεις μεγαλύτερες του q, οι προβλέψεις $\hat{X}_n(l)$ θα έχουν την ίδια συμπεριφορά με τις αυτοσυσχετίσεις $\rho(s)$. Δηλαδή θα γράφονται ως γραμμικός συνδυασμός από δυνάμεις των χαρακτηριστικών ριζών $r_1, \dots, r_p$ του AR(p) μέρους του ARMA(p,q). Οι όροι της $\hat{X}_n(l)$ που αντιστοιχούν σε $r_i \in R$ θα μειώνονται εκθετικά ενώ τα ζεύγη των συζυγών μιγαδικών ριζών θα σχηματίζουν ημιτονοειδείς όρους με εκθετικά μειούμενο πλάτος. Η διαφορά με τους τύπους των $\rho(s)$ είναι ότι τώρα οι συντελεστές των ${r_i}^2$ στο ανάπτυγμα της $\hat{X}_n(l)$ εξαρτώνται από τα $\varepsilon_1, \ \varepsilon_2, \dots, \ \varepsilon_q$. Σύμφωνα με τα παραπάνω, για μεγάλες προηγήσεις l η πρόβλεψη $\hat{X}_n(l) \to 0$.



Διάγραμμα 4.1 Απεικόνιση 24 – μήνες πρόγνωση για την ARMA(1,1) χρονοσειρά προσλήψεων. Τα αρχικά δεδομένα αφορούν την περίοδο Ιανουάριος 1980 έως Σεπτέμβριος 1987, και μετά φαίνονται η πρόγνωση μαζί με το σφάλμα. *Πηγή: [1] Robert H. Shumway, David S. Stoffer: Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics) Fourth (4th) Edition*

Αν όμοια με παραπάνω $Y_t = d_t + X_t$ όπου d_t είναι ο συντελεστής τάσης (ή και εποχικότητας) και X είναι στάσιμη ARMA(p,q) χρονοσειρά τότε:

$$\hat{Y}_n(l) = E\big(d_{n+l} + X_{n+l}\big|(X_1, X_2, \dots, X_n)\big) = d_{n+l} + E\big(X_{n+l}\big|(X_1, X_2, \dots, X_n)\big)$$
$$= d_{n+l} + \hat{X}_n(l), \quad l = 1,2, \dots$$

Το σφάλμα πρόβλεψης προήγησης l είναι

$$\hat{e}_n(l) = Y_{n+l} - \hat{Y}_n(l) = d_{n+l} + X_{n+l} - d_{n+l} - \hat{X}_n(l) = X_{n+l} - \hat{X}_n(l) \to X_{n+l}$$

και ακολουθεί κανονική κατανομή με

$$E\big(\hat{e}_n(l)\big) = E\left(X_{n+l} - \hat{X}_n(l)\right) = E\left(X_{n+l} - E\big(X_{n+l}\big|(X_1, X_2, \dots, X_n)\big)\right)$$
$$= E(X_{n+l}) - E\left(E\big(X_{n+l}\big|(X_1, X_2, \dots, X_n)\big)\right) = 0$$

και

$$\sigma_e{}^2(l) = Var\big(\hat{e}_n(l)\big) = Var\left(X_{n+l} - \hat{X}_n(l)\right) \to Var(X_{n+l}) = \gamma(0)$$

το διάστημα πρόβλεψης συντελεστού $1 - a$, για την Y_{n+l} , είναι

$$\left(d_{n+l} + \hat{X}_n(l) - \sigma_e(l) z_{\{a/2\}}, d_{n+l} + \hat{X}_n(l) + \sigma_e(l) z_{\{a/2\}}\right)$$
$$\to \left(d_{n+l} - \sqrt{\gamma(0)} z_{\{a/2\}}, d_{n+l} + \hat{X}_n(l) + \sqrt{\gamma(0)} z_{\{a/2\}}\right)$$

4.3 Πρόβλεψη Μη Στάσιμων Χρονολογικών Σειρών με Γραμμικά Μοντέλα

ARIMA(p,1,q) Μοντέλο

Όταν η χρονοσειρά δεν είναι στάσιμη τότε η πρόβλεψη γίνεται με τα μοντέλα ARIMA ή SARIMA. Οι γενικοί τύποι για τις προβλέψεις με αυτά τα μοντέλα είναι αρκετά περίπλοκοι, για αυτό θα εξετάσουμε το μη στάσιμο υπόδειγμα ARIMA(p,1,q) το οποίο έχει μία μοναδιαία ρίζα.

Όπως έχει ξανααναφερθεί οι πρώτες διαφορές μίας ARIMA(p,1,q) χρονοσειράς Y_t είναι $X_t = (1 - B)Y_t = Y_t - Y_{t-1}$, τότε η X_t είναι στάσιμη ARMA(p,q) χρονοσειρά. Προκύπτει εύκολα ότι

$$Y_{n+l} = Y_{n+l-1} + X_{n+l} = \cdots = \sum_{i=1}^{l} X_{n+i} + Y_n$$

και επομένως,

$$\hat{Y}_n(l) = E(Y_{n+l}|(X_1, X_2, \dots, X_n)) = E\left(\sum_{i=1}^{l} X_{n+i} + Y_n \middle| (X_1, X_2, \dots, X_n)\right)$$
$$= \hat{X}_n(l) + \cdots + \hat{X}_n(1) + Y_n$$

όπου το σφάλμα της πρόβλεψης της Y_t με προήγηση l θα είναι

$$\hat{e}_n^Y(l) = Y_{n+l} - \hat{Y}_n(l) = \left(\sum_{i=1}^{l} X_{n+i} + Y_n\right) - \left(\sum_{i=1}^{l} \hat{X}_n(i) + Y_n\right) = \hat{e}_n^X(l) + \cdots + \hat{e}_n^X(1)$$

δηλαδή ισούται με το άθροισμα των σφαλμάτων πρόβλεψης της X_t με προηγήσεις $1,2,\dots,l$. Συνεπώς το σφάλμα $\hat{e}_n^Y(l)$ θα έχει μέση τιμή 0 (έχουμε αμερόληπτες προβλέψεις) με διασπορά όμως που μεγαλώνει απεριόριστα διότι $Var\left(\hat{e}_n^X(i)\right) \to \gamma(0)$.

Βλέπε [18] - Μπούτσικας Μιχαήλ. Σημειώσεις "Ανάλυση Χρονολογικών Σειρών"

ΜΕΡΟΣ II

Εφαρμογή

ΚΕΦΑΛΑΙΟ 5

Ανάλυση Χρονοσειρών με Εφαρμογή στην R

5.1 Παρουσίαση Δεδομένων

Τα δεδομένα που χρησιμοποιούνται για τη διεξαγωγή της εφαρμογής της ανάλυσης των χρονοσειρών στην R, έχουν αντληθεί από το Κέντρο Ελέγχου και Πρόληψης Νοσημάτων (ΚΕ.ΕΛ.Π.ΝΟ.) και αφορούν επιδημιολογικά δεδομένα.

Η χρονοσειρά που εξετάζεται, η οποία θα παρουσιαστεί αναλυτικά παρακάτω, αντλείται από ένα σύστημα παρατηρητών νοσηρότητας στην Πρωτοβάθμια Φροντίδα Υγείας (σύστημα "sentinel") και τη βελτιωμένη εκδοχή του. Το σύστημα sentinel αποτελείται από ιατρούς-παρατηρητές στην Πρωτοβάθμια Φροντίδα Υγείας (ιδιώτες, ΚΥ, ΠΕΔΥ), γεωγραφικά διασκορπισμένους, οι οποίοι δηλώνουν κάθε εβδομάδα στο ΚΕ.ΕΛ.Π.ΝΟ., μεταξύ άλλων, τον αριθμό ασθενών με γριπώδη συνδρομή (Influenza-like Illness – ILI) που βλέπουν στα ιατρεία τους, και το συνολικό αριθμό επισκέψεων. Από τις δηλώσεις αυτές των παρατηρητών, υπολογίζεται κάθε εβδομάδα ο αριθμός γριπωδών συνδρομών ανά 1000 επισκέψεις (ILI rate), ο οποίος και αντανακλά τη δραστηριότητα της γρίπης στην κοινότητα.

Αναλυτικά η χρονοσειρά:

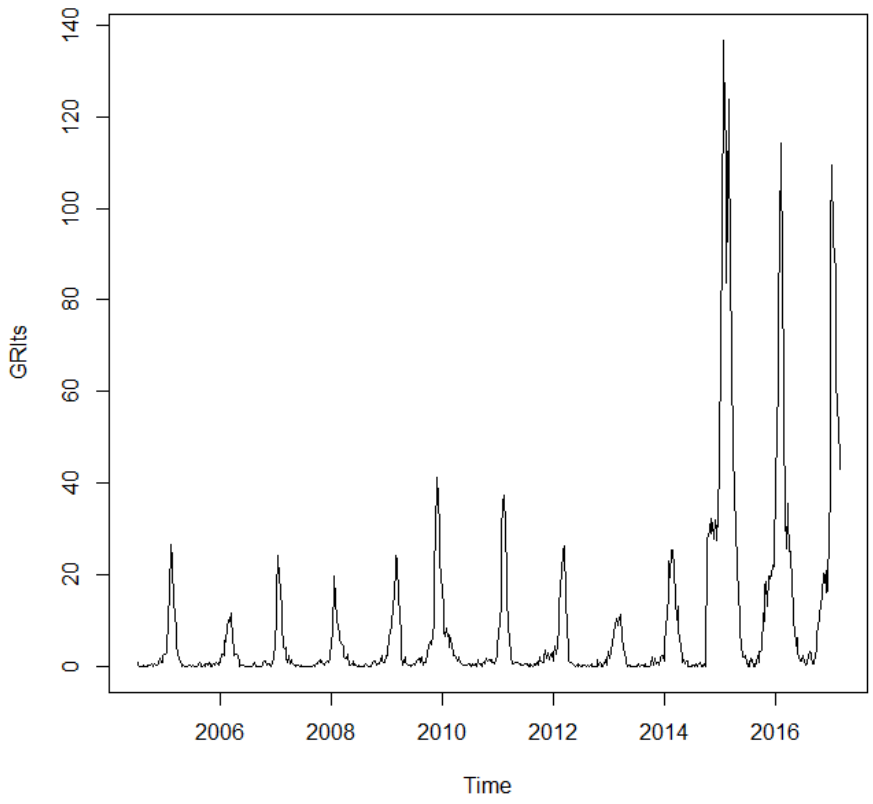
GRI : ο συνολικός αριθμός γριπωδών συνδρομών ανά 1000 επισκέψεις (Συνολικό ILI rate / 1000 επισκέψεις) όπου οι παρατηρήσεις της χρονοσειράς συγκεντρώνονται ανά εβδομάδα. Οι παρατηρήσεις ξεκινάνε από την 28^{η} βδομάδα του 2004 και εξελίσσονται έως και την 8^{η} βδομάδα του 2017. Συνολικά το μέγεθος της χρονοσειράς αποτελείται από 660 παρατηρήσεις.

Παρακάτω η χρονοσειρά θα αναλυθεί στην R.

5.2 Απεικόνιση Χρονοσειρών στην R

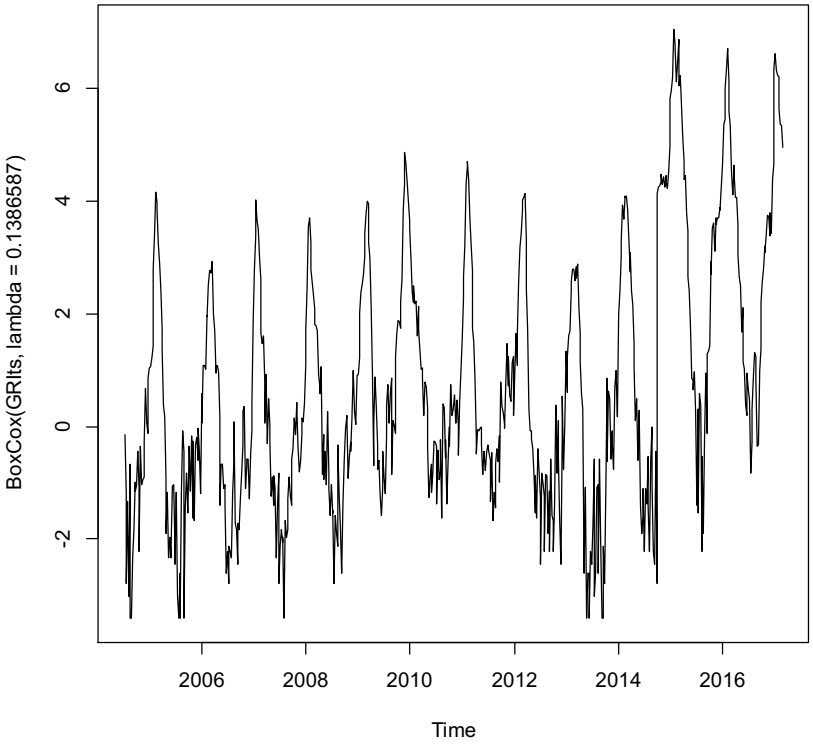
Παρακάτω ακολουθούν οι αντίστοιχες εντολές στην R και η ανάλυση της χρονοσειράς. Για να εισάγουμε τα δεδομένα στην R, χρησημοποιούμε την εντολή *read.csv2()*. Η εντολή *ts()* χρησιμοποιείται για να δημιουργήσει ένα αντικείμενο χρονοσειράς.

```
> GRI <- read.csv2(file.choose(),header=TRUE)[,1]

> length(GRI)

[1] 660

> GRIts <-ts(GRI, frequency=365.25/7,start=c(2004,28))

> plot.ts(GRIts)
```



Στο παραπάνω γράφημα η διακύμανση φαίνεται να αυξάνεται με το χρόνο. Για αυτό το λόγο οι χρονοσειρά χρήζει μετασχηματισμού για τη σταθεροποίηση της διακύμανσης. Ο μετασχηματισμός της χρονοσειράς θα πραγματοποιηθεί με τη βοήθεια του μετασχηματισμού Box Cox με βέλτιστη παράμετρο λ χρησιμοποιώντας το αντιστοιχο πακέτο στην R.

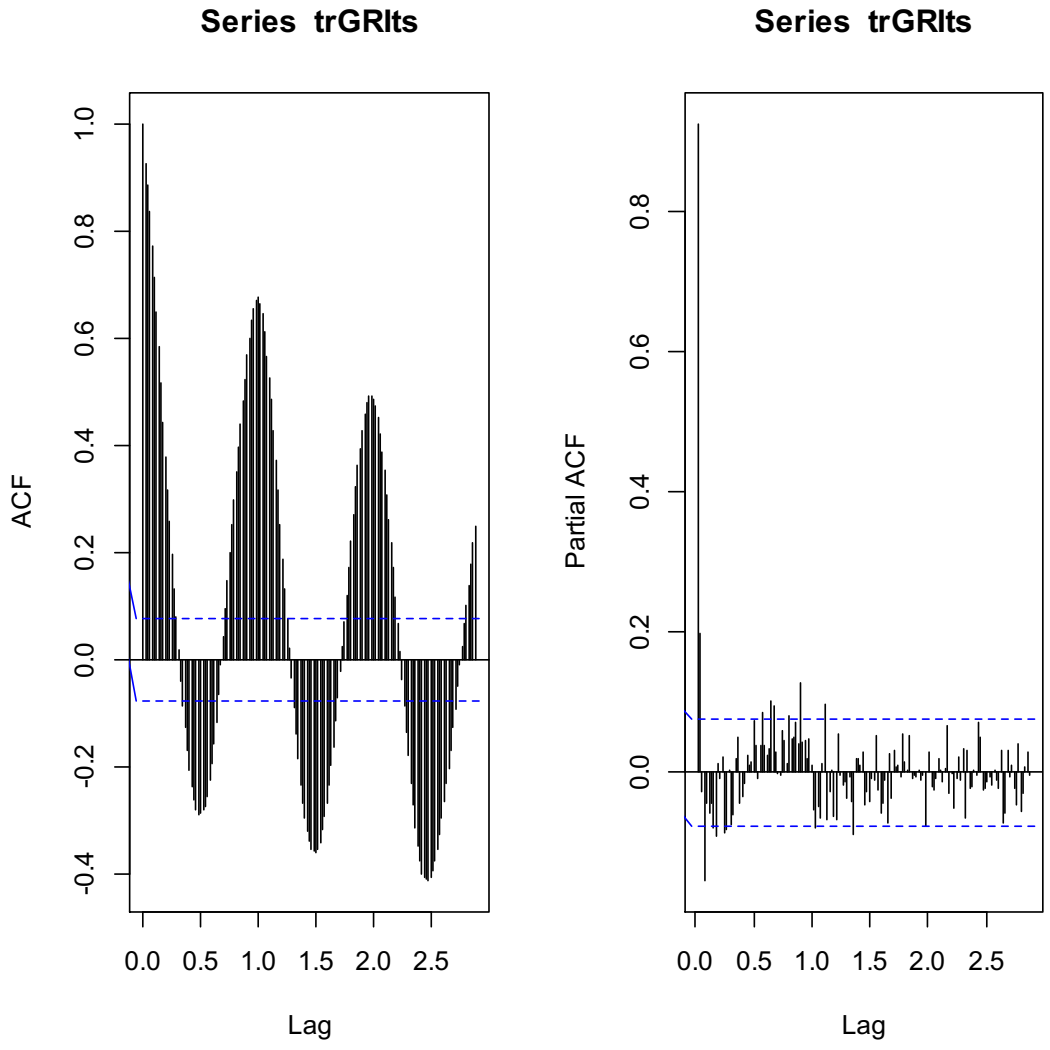
```
> library(forecast)

> lambda <- BoxCox.lambda(GRIts)

> lambda

[1] 0.1386587

> plot.ts(BoxCox(GRIts, lambda = lambda), ylab= "BoxCox(GRIts, lambda =
0.1386587)")
```



Από το μετασχηματισμένο διάγραμμα καταλαβαίνουμε ότι οι χρονοσειρά δεν είναι στάσιμη, φαίνεται να περιέχει τη συνιστώσα της εποχικότητας και της τάσης. Το διάγραμμα παρουσιάζει μία εποχική διακύμανση ανά τους μήνες για τον αριθμό τών γριπωδών συνδρομών με κορύφωση τους χειμερινούς μήνες και ελάττωση τους καλοκαιρινούς, όπως θα ήταν αναμενόμενο. Για περαιτέρω έλεγχο της στασιμότητας πραγματοποιούμε τα διαγράμματα αυτοσυσχέτισης (ACF) και μερικής

αυτοσυσχέτισης (PACF) της χρονοσειράς (μετασχηματισμένης) και επιβεβαιώνουμε τη μη-στασιμότητα με κατάλληλους ελέγχους.

```
> trGRIts <- BoxCox(GRIts, lambda = lambda)    #transformed GRIts
> par(mfrow=c(1,2))
> acftrGRIts <- acf(trGRIts, lag.max=150)
> pacftrGRIts <- pacf(trGRIts, lag.max=150)
```

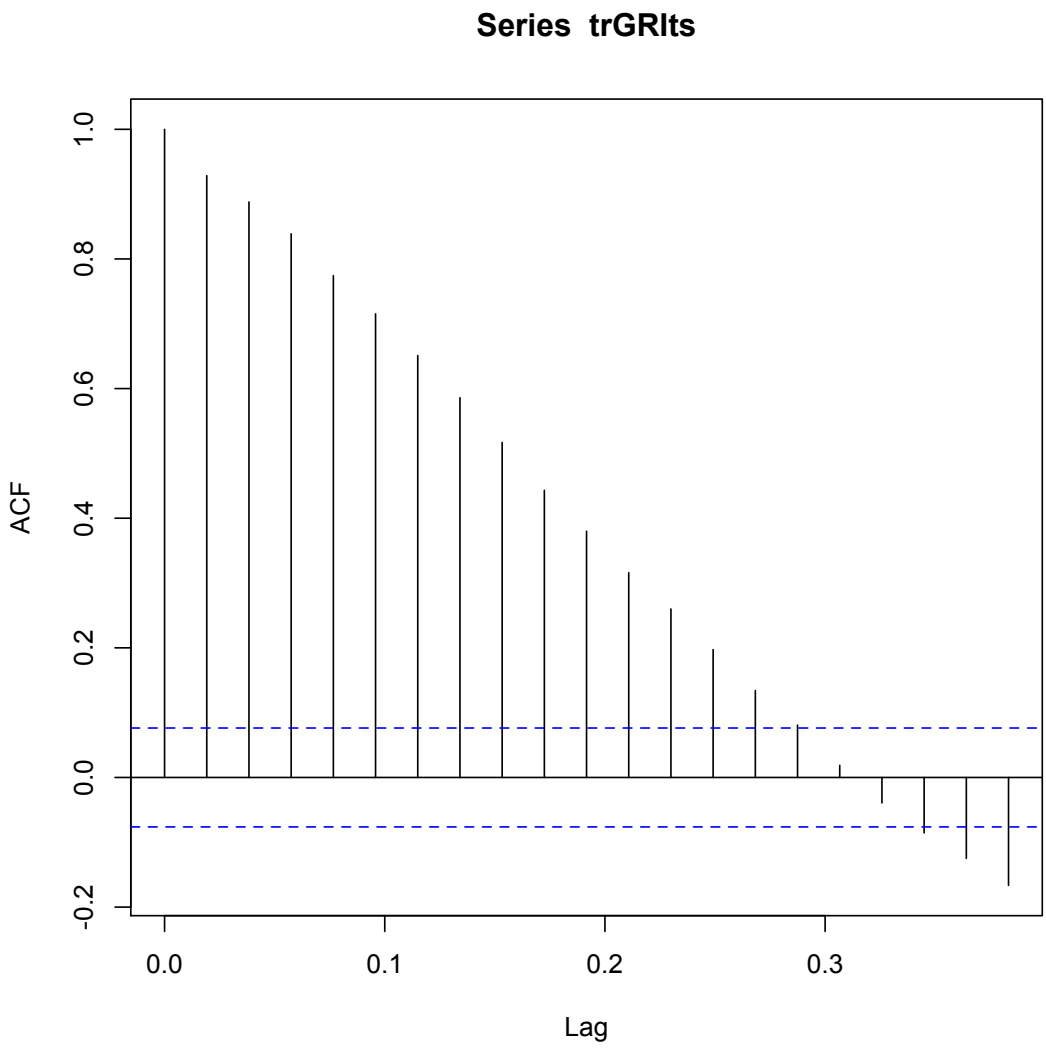


Το διάγραμμα αυτοσυσχέτισης της μετασχηματισμένης χρονοσειράς με μέγιστο αριθμό υστέρησης max.lag=150, παρουσιάζει όπως περιμέναμε μη-στασιμότητα , καθώς όλες οι τιμές πέφτουν έξω από τα όρια. Επιπρόσθετα παρατηρείται

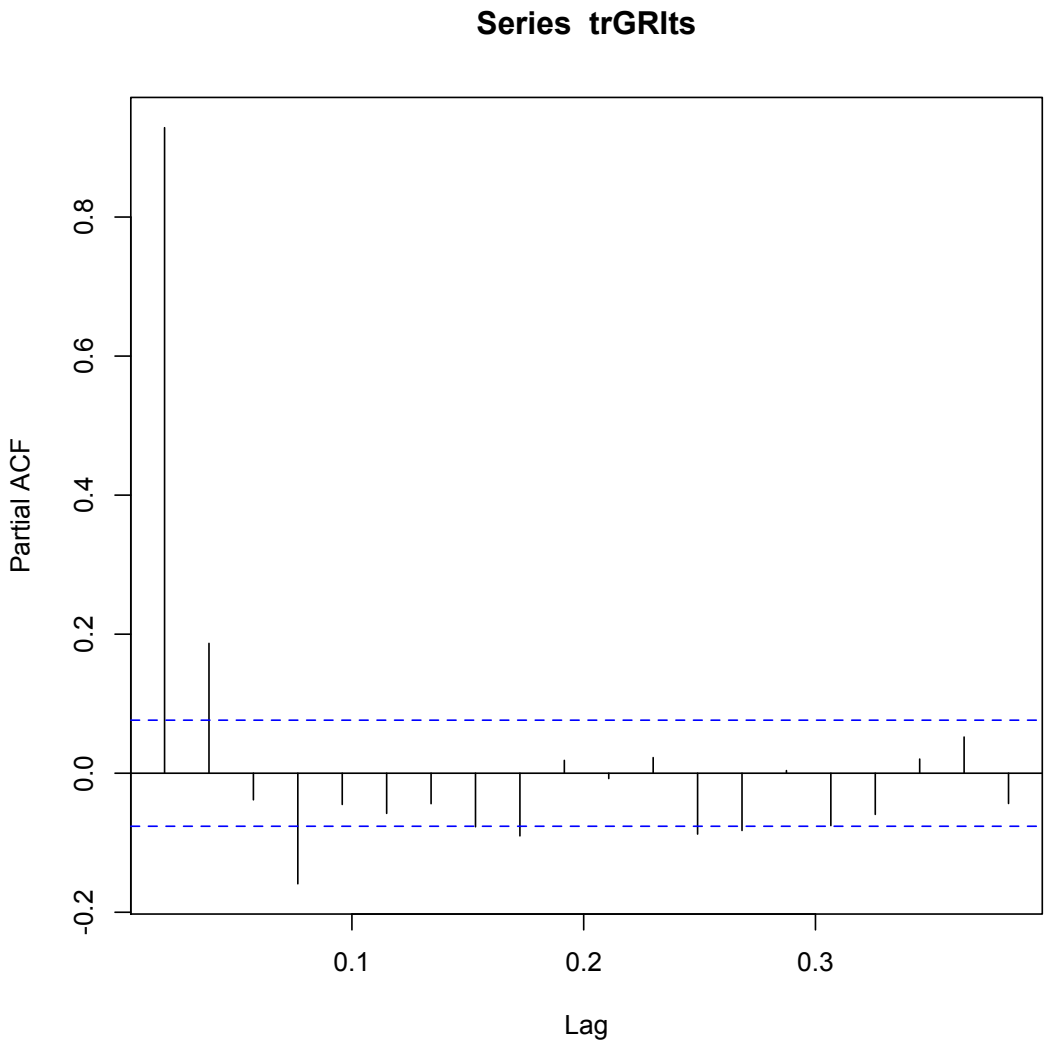
περιοδικότητα σε κάθε διάγραμμα το οποίο αποτελεί ένδειξη ότι η χρονοσειρά περιέχει εποχική συνιστώσα.

Για μεγαλύτερη ευκρίνεια δηλώνουμε max.lag=20 στο διάγραμμα αυτοσυσχέτισης όπου διαπιστώνουμε ότι η αυτοσυσχέτιση φθίνει αργά με την υστέρηση το οποίο είναι ένδειξη μη στασιμότητας που οφείλεται στην ύπαρξη τάσης.

```
> acftrGRIts <- acf(trGRIts, lag.max=20)
```



> pacftrGRIts <- pacf(trGRIts, lag.max=20)



Για να επιβεβαιώσουμε τη μη-στασιμότητα εφαρμόζουμε έλεγχο Augmented Dickey-Fuller στην R.

> library(tseries)

> adf.test(trGRIts)

Augmented Dickey-Fuller Test

data: trGRIts

Dickey-Fuller = -6.3382, Lag order = 8, p-value = 0.01

alternative hypothesis: stationary

Warning message:

In adf.test(trGRIts) : p-value smaller than printed p-value

Παρατηρούμε ότι το p-value του ελέγχου Augmented Dickey-Fuller είναι μικρότερο από 0.05 που σημαίνει ότι η μηδενική υπόθεση, H_0: μη-στάσιμη χρονοσειρά, απορρίπτεται, δηλαδή η χρονοσειρά είναι στάσιμη ή δεν έχει μοναδιαία ρίζα.

Για περαιτέρω έλεγχο, χρησιμοποιούμε KPSS test.

```
> kpss.test(trGRIts)

        KPSS Test for Level Stationarity

data:  trGRIts
KPSS Level = 1.7734, Truncation lag parameter = 6, p-value = 0.01

Warning message:
In kpss.test(trGRIts) : p-value smaller than printed p-value
```

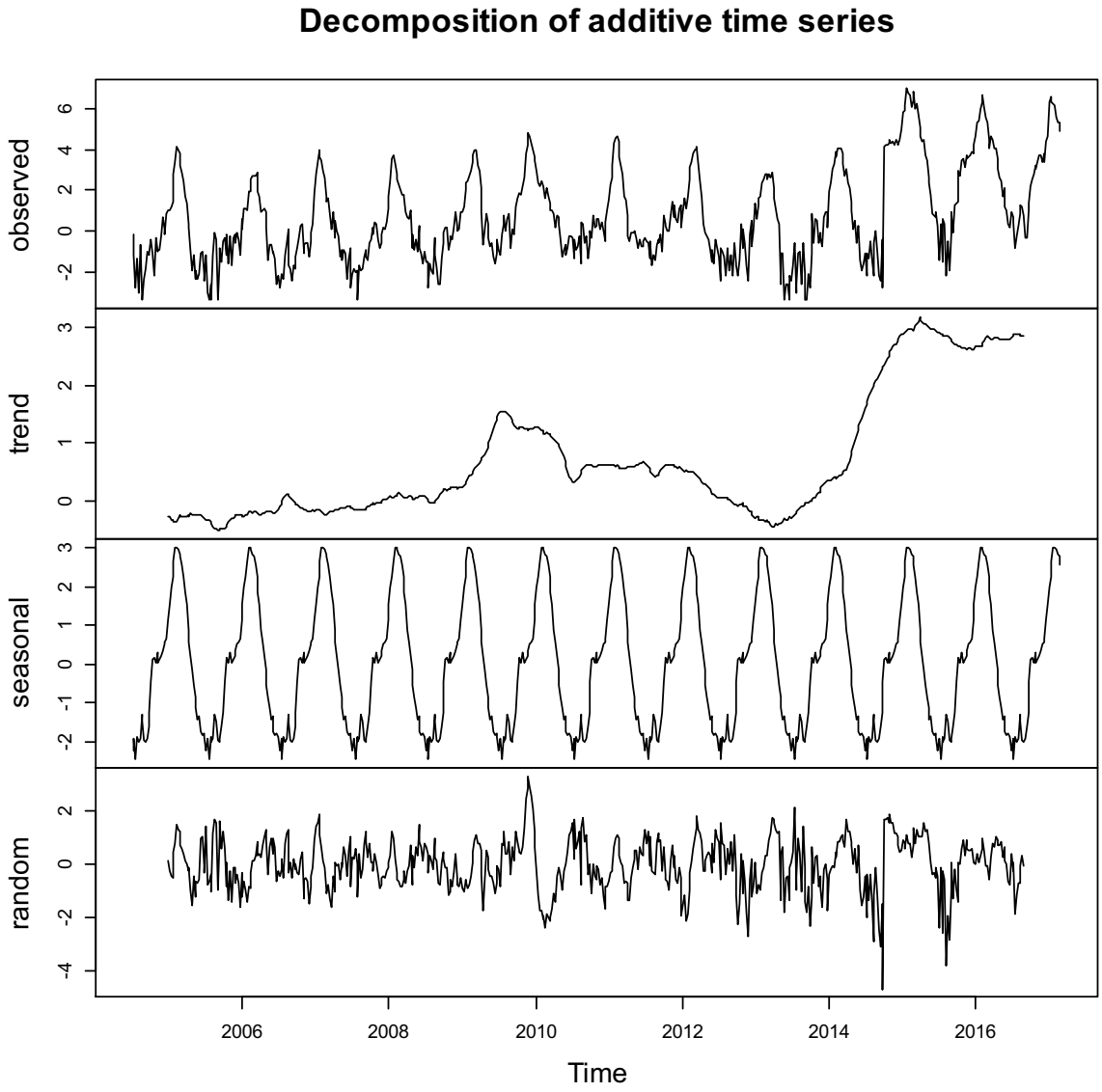
Πράγματι ο KPSS έλεγχος δείχνει μη στασιμότητα της χρονοσειράς, αφού απορρίπτεται η μηδενική υπόθεση, H_0: στάσιμη χρονοσειρά, p-value=0.01<0.05.

Η διαφορά ανάμεσα στους δύο ελέγχους έγκειται στο ότι η χρονοσειρά δεν παρουσιάζει μοναδιαία ρίζα αλλά η μη-στασιμότητα προέρχεται από την συνιστώσα της τάσης.

Για να ελέγξουμε τις συνιστώσες της τάσης, της εποχικότητας και το υπόλοιπο, είτε με το προσθετικό μοντέλο είτε με το πολλαπλασιαστικό, διαχωρίζουμε τις συνιστώσες χρησιμοποιώντας την εντολή *decompose()* στην R:

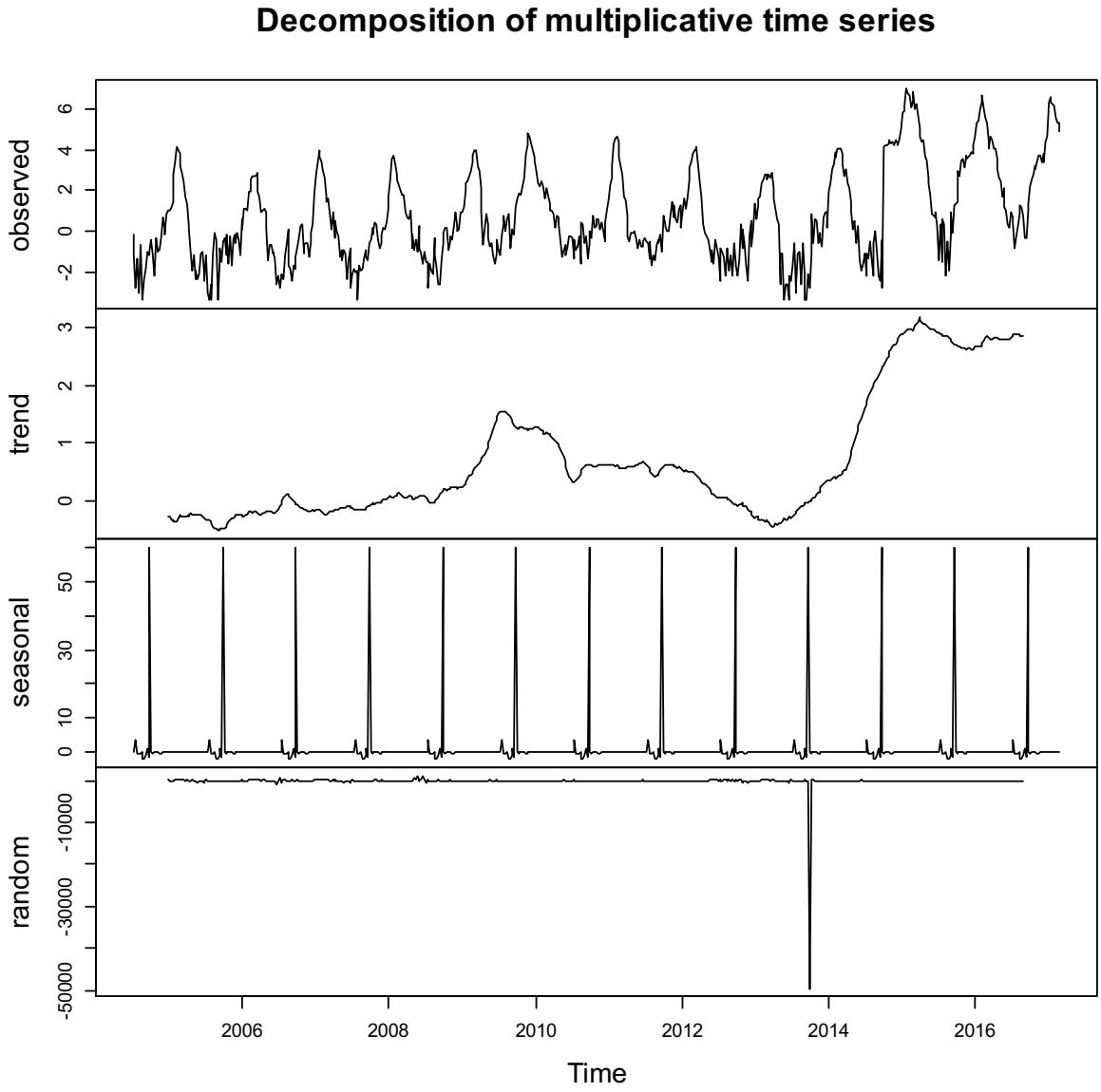
> trGRItscomponentsA <- decompose(trGRIts, "additive")#decomposition as per additive model

> plot(trGRItscomponentsA)



> trGRItscomponentsM <- decompose(trGRIts, "multiplicative")#decomposition as per multiplicative model

> plot(trGRItscomponentsM)



Στα παραπάνω γραφήματα, είτε με το προσθετικό μοντέλο , είτε με το πολλαπλασιαστικό, διακρίνονται και οι δύο συνιστώσες , της τάσης και της εποχικότητας.

Σκοπός είναι η GRI χρονοσειρά να μετατραπεί σε στάσιμη ώστε να γίνει επιλογή μοντέλου ARIMA.

Επομένως για να αφαιρέσουμε τη συνιστώσα της εποχικότητας από τα δεδομένα γράφουμε στην R:

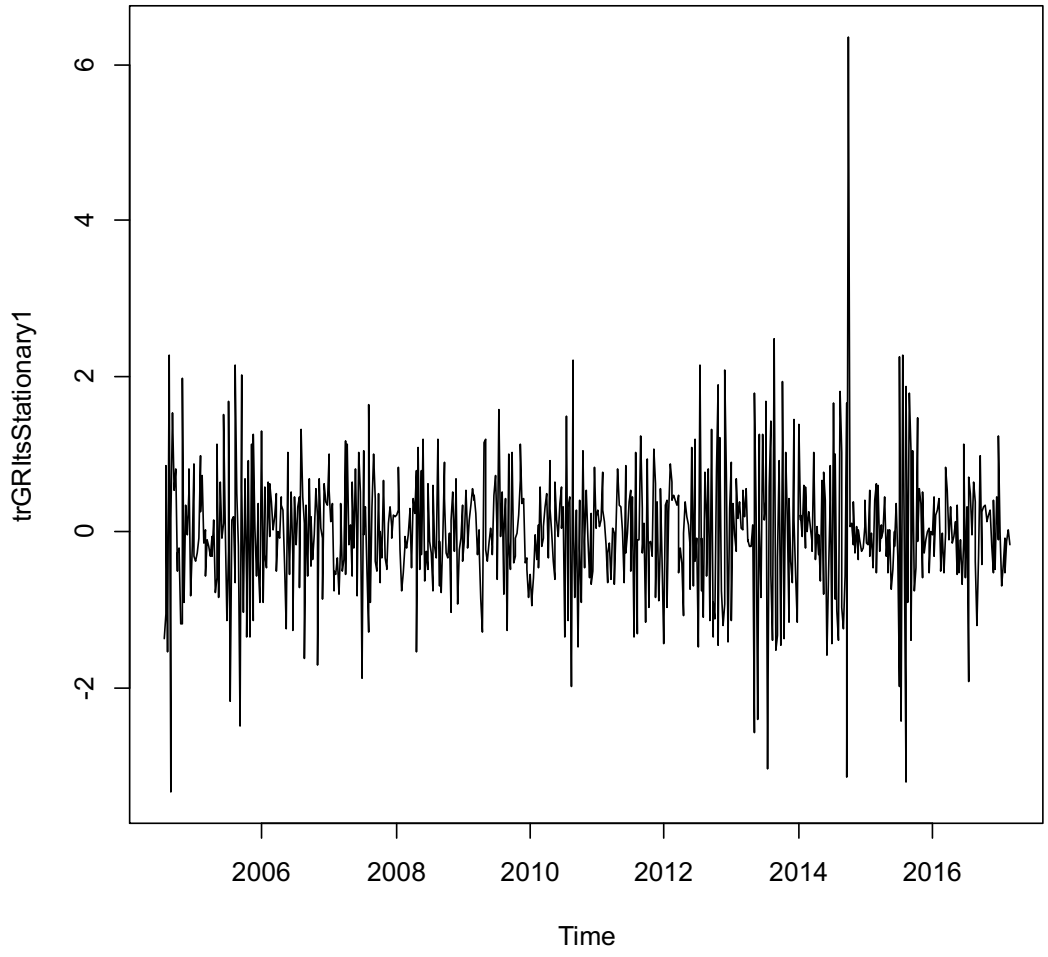
>trGRItsSeasonallyAdjustedA <- seasadj(trGRItscomponentsA) #deseasonalise the additive model

>trGRItsSeasonallyAdjustedM <- seasadj(trGRItscomponentsM) #deseasonalise the multiplicative model

Στη συνέχεια για την απαλοιφή της τάσης χρησιμοποιούμε τη μέθοδο των διαφορών, εφόσον η τάση δεν φαίνεται να είναι κάποια γνωστή συνάρτηση με το χρόνο.

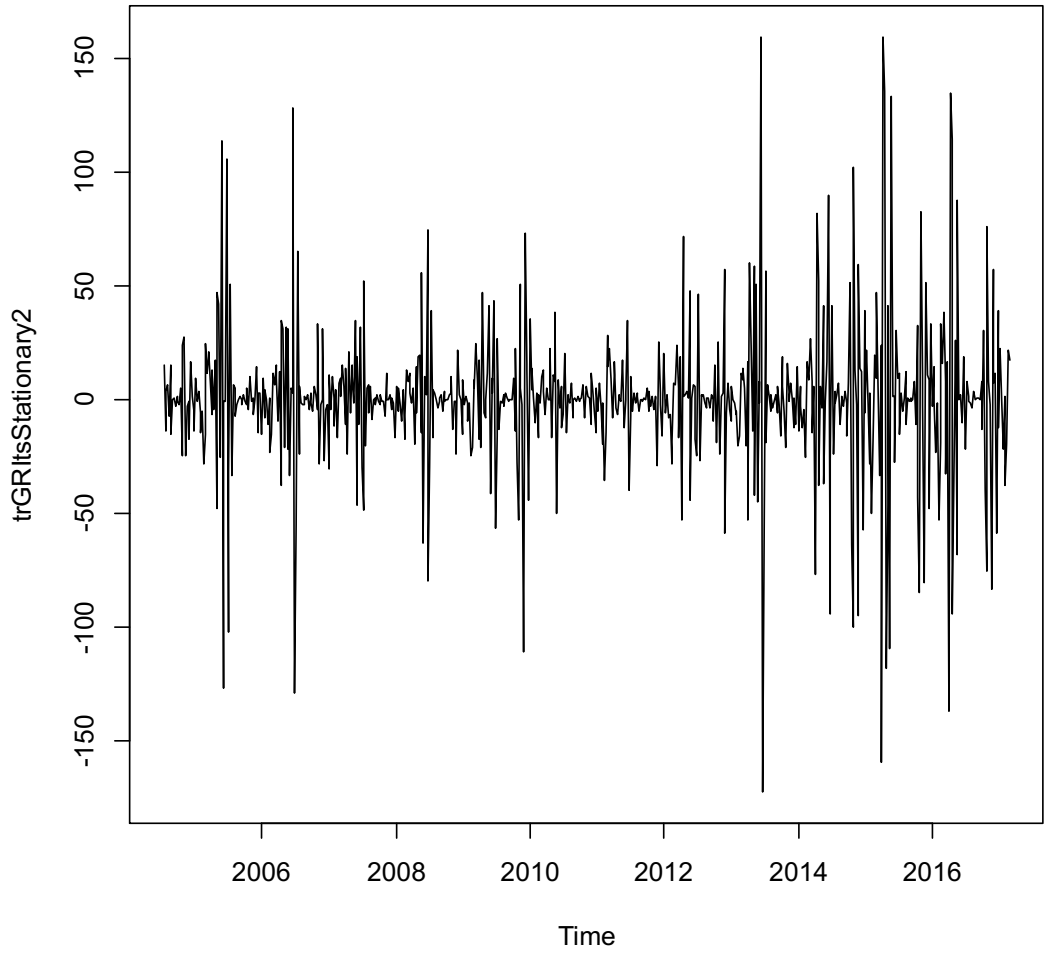
>trGRItsStationary1 <- diff(trGRItsSeasonallyAdjustedA, differences=1) #as per additive

> plot(trGRItsStationary1)



>trGRItsStationary2 <- diff(trGRItsSeasonallyAdjustedM, differences=1) #as per multiplicative

> plot(trGRItsStationary2)



Οι παραπάνω εντολές στην R εκτελούν τη μέθοδο των διαφορών τόσο στο προσθετικό μοντέλο όσο και στο πολλαπλασιαστικό παίρνοντας μία διαφορά.

Για να διαπιστώσουμε αν η χρονοσειρά έχει μετατραπεί σε στάσιμη εκτελούμε πάλι KPSS test στην R:

> kpss.test(trGRItsStationary1)#as per additive

KPSS Test for Level Stationarity

data: trGRItsStationary1

KPSS Level = 0.02587, Truncation lag parameter = 6, p-value = 0.1

Warning message:

In kpss.test(trGRItsStationary1) : p-value greater than printed p-value

> kpss.test(trGRItsStationary2)#as per multiplicative

KPSS Test for Level Stationarity

data: trGRItsStationary2

KPSS Level = 0.011608, Truncation lag parameter = 6, p-value = 0.1

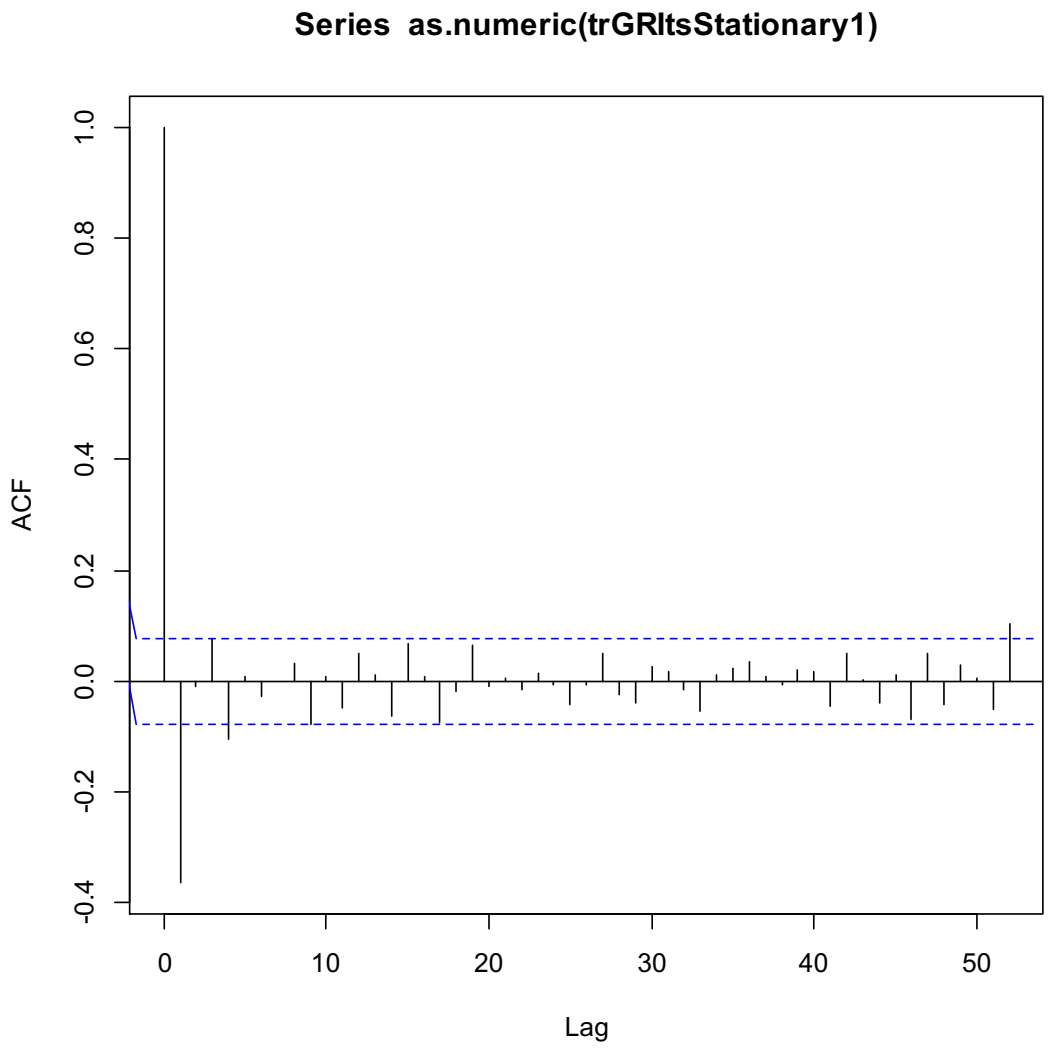
Warning message:

In kpss.test(trGRItsStationary2) : p-value greater than printed p-value

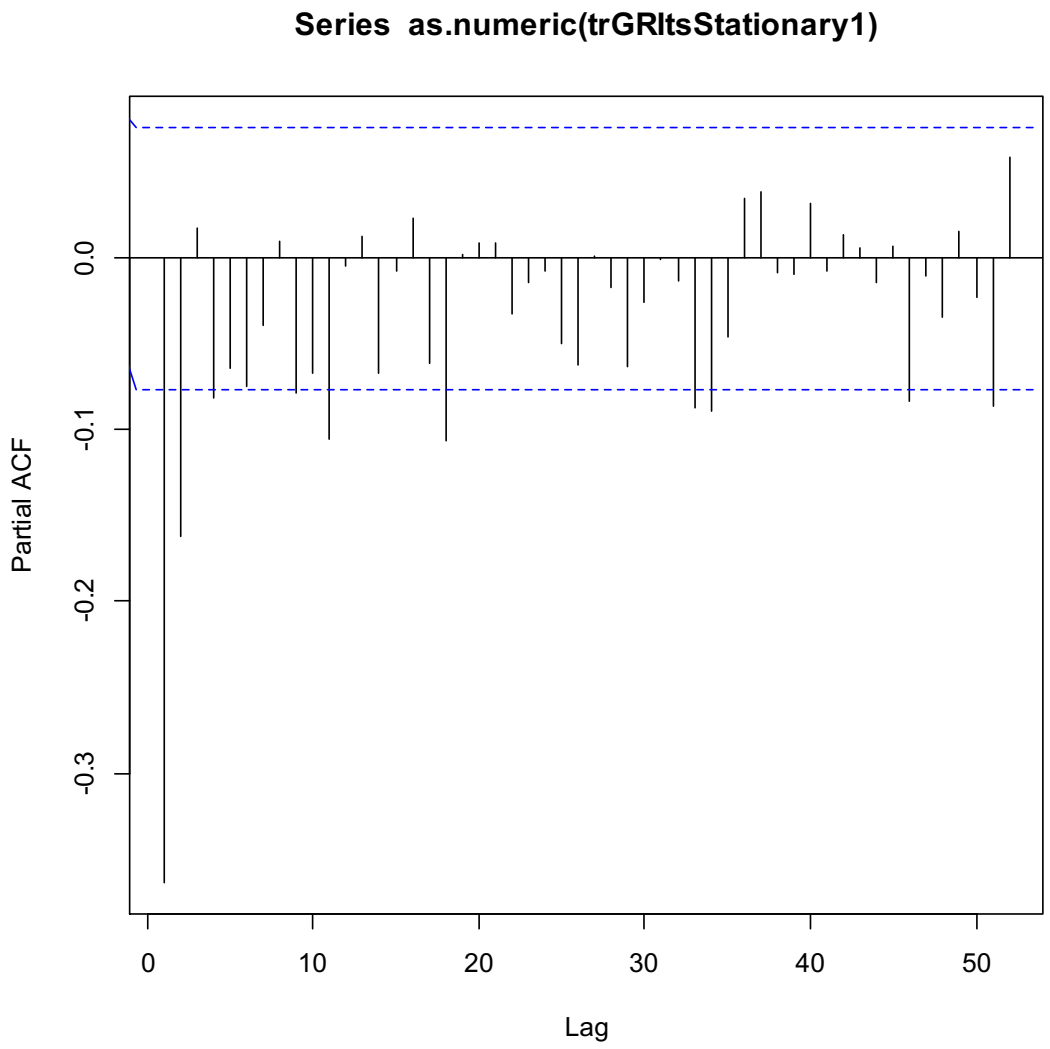
Παρατηρούμε ότι και οι δύο έλεγχοι εμφανίζουν p-value > 0.05 , που σημαίνει ότι η μηδενική υπόθεση δεν απορρίπτεται άρα η χρονοσειρά είναι στάσιμη.

Εφαρμόζουμε ξανά τα διαγράμματα αυτοσυσχέτισης και μερικής αυτοσυσχέτισης και διαπιστώνουμε, όπως περιμέναμε, στασιμότητα της μετασχηματισμένης χρονοσειράς. Από το διάγραμμα αυτοσυσχέτισης παρατηρούμε ότι η χρονοσειρά των πρώτων διαφορών που περιγράφει τις εβδομαδιαίες μεταβολές του αριθμού των γριπωδών συνδρομών δεν παρουσιάζει τάσεις και η αυτοσυσχέτισή της τίνει γρήγορα στο 0. Το γεγονός ότι κάποιες αυτοσυσχετίσεις για κάποιες υστερήσεις είναι στατιστικά σημαντικές σημαίνει ότι η χρονοσειρά των διαφορών δεν είναι λευκός θόρυβος αλλά στάσιμη χρονοσειρά με ασθενείς αυτοσυσχετίσεις. Επίσης παρατηρούμε ότι το q είναι τουλάχιστον 1 και το πολύ 3. Από το διάγραμμα μερικής αυτοσυσχέτισης παρατηρούμε ότι έχουμε 2 στατιστικά σημαντικές μερικές αυτοσυσχετίσεις που ίσως είναι ένδειξη ότι το p=2.

> acf(as.numeric(trGRItsStationary1), lag.max=52)



> pacf(as.numeric(trGRItsStationary1), lag.max=52)



Σε συνέχεια της ανάλυσης των δεδομένων μπορούμε να καταλήξουμε στο μεικτό εποχικό υπόδειγμα, δεδομένης τάσης και εποχικότητας στα δεδομένα, τη στασιμότητα των οποίων επιτύχαμε με τη μέθοδο των πρώτων διαφορών. Δοκιμάζουμε διάφορα SARIMA(p,1,q)(0,0,1)[s] όπου s=52 εβδομάδες μέχρι να βρούμε αυτό με το ελάχιστο AIC.

> arima111<-arima(GRIts, order=c(1,1,1),seasonal = list(order = c(0,0,1), period = 52),method="ML", optim.method = "BFGS")

> AIC111<-arima111$aic

> AIC111

[1] 4137.989

> BIC111<-AIC(arima111,k=log(length(GRIts)))

> BIC111

[1] 4155.958

> arima112<-arima(GRIts, order=c(1,1,2),seasonal = list(order = c(0,0,1), period = 52),method="ML", optim.method = "BFGS")

> AIC112<-arima112$aic

> AIC112

[1] 4136.8

> BIC112<-AIC(arima112,k=log(length(GRIts)))

> BIC112

[1] 4159.261

> arima113<-arima(GRIts, order=c(1,1,3),seasonal = list(order = c(0,0,1), period = 52),method="ML", optim.method = "BFGS")

> AIC113<-arima113$aic

> AIC113

[1] 4130.594

> BIC113<-AIC(arima113,k=log(length(GRIts)))

> BIC113

[1] 4157.547

> arima211<-arima(GRIts, order=c(2,1,1),seasonal = list(order = c(0,0,1), period = 52),method="ML", optim.method = "BFGS")

> AIC211<-arima211$aic

> AIC211

[1] 4137.832

> BIC211<-AIC(arima211,k=log(length(GRIts)))

> BIC211

[1] 4160.294

> arima212<-arima(GRIts, order=c(2,1,2),seasonal = list(order = c(0,0,1), period = 52),method="ML", optim.method = "BFGS")

> AIC212<-arima212$aic

> AIC212

[1] 4129.537

> BIC212<-AIC(arima212,k=log(length(GRIts)))

> BIC212

[1] 4156.49

> arima213<-arima(GRIts, order=c(2,1,3),seasonal = list(order = c(0,0,1), period = 52),method="ML", optim.method = "BFGS")

> AIC213<-arima213$aic

> AIC213

[1] 4132.16

> BIC213<-AIC(arima213,k=log(length(GRIts)))

> BIC213

[1] 4163.605

> arima311<-arima(GRIts, order=c(3,1,1),seasonal = list(order = c(0,0,1), period = 52),method="ML", optim.method = "BFGS")

> AIC311<-arima311$aic

> AIC311

[1] 4130.336

> BIC311<-AIC(arima311,k=log(length(GRIts)))

> BIC311

[1] 4157.289

> arima312<-arima(GRIts, order=c(3,1,2),seasonal = list(order = c(0,0,1), period = 52),method="ML", optim.method = "BFGS")

Warning messages:

1: In log(s2) :

2: In log(s2) : NaNs produced

3: In log(s2) : NaNs produced

> AIC312<-arima312$aic

> AIC312

[1] 4121.291

> BIC312<-AIC(arima312,k=log(length(GRIts)))

> BIC312

[1] 4152.737

> arima313<-arima(GRIts, order=c(3,1,3),seasonal = list(order = c(0,0,1), period = 52),method="ML", optim.method = "BFGS")

> AIC313<-arima313$aic

> AIC313

[1] 4128.059

> BIC313<-AIC(arima313,k=log(length(GRIts)))

> BIC313

[1] 4163.997

Παρακάτω παρατίθεται πίνακας συνοψίζοντας όλα τα παραπάνω αποτελέσματα.

p,d,q,P,D,Q	AIC	BIC
1,1,1,0,0,1	4137.989	4155.958
1,1,2,0,0,1	4136.8	4159.261
1,1,3,0,0,1	4130.594	4157.547
2,1,1,0,0,1	4137.832	4160.294
2,1,2,0,0,1	4129.537	4156.49
2,1,3,0,0,1	4132.16	4163.605
3,1,1,0,0,1	4130.336	4157.289
3,1,2,0,0,1	4121.291	4152.737
3,1,3,0,0,1	4128.059	4163.997

Μετά από τις παραπάνω δοκιμές παρατηρούμε ότι το μικρότερο AIC και BIC αντιστοιχεί στο μοντέλο SARIMA(3,1,2)(0,0,1)[52]. Για να διαπιστώσουμε αν η επιλογή είναι σωστή χρησιμοποιούμε την συνάρτηση της R *auto.arima()* που επιστρέφει το καλύτερο δυνατό ARIMA μοντέλο για τα δεδομένα σύμφωνα με το AIC κριτήριο.

```
> arimaGRI <- auto.arima(GRIts)

> arimaGRI

Series: GRIts

ARIMA(3,1,2)(0,0,1)[52]

Coefficients:
         ar1      ar2     ar3      ma1     ma2     sma1
      0.5125  -0.9062  0.1535  -0.3950  0.9524  0.1476
s.e.  0.0519   0.0417  0.0429   0.0353  0.0283  0.0429

sigma^2 estimated as 29.99:  log likelihood=-2053.65
AIC=4121.29   AICc=4121.46   BIC=4152.73
```

Πράγματι, η συνάρτηση επέλεξε ως βέλτιστο μοντέλο το SARIMA(3,1,2)(0,0,1).

Αρα το εκτιμώμενο μοντέλο γράφεται ως

$$\hat{x}_t = 0.5125x_{t-1} - 0.9062x_{t-2} + 0.1535x_{t-3} + u_t - 0.3950u_{t-1} + 0.9524u_{t-2} + 0.1476u_{t-52}$$

για *t=1,…,660* (5.1.1)

Με βάση το παραπάνω μοντέλο που υπέδειξε η R συνεχίζουμε σε πρόγνωση των δεδομένων για τις επόμενες 52 εβδομάδες/ 1 χρόνο γράφοντας στην R:

```
> forecast_GRI<- forecast(arimaGRI, h=52)

> forecast_GRI
```

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2017.166	35.87704	28.859017	42.89506	25.1438999	46.61018
2017.185	32.55607	22.031601	43.08054	16.4602840	48.65185
2017.205	36.62698	23.041924	50.21203	15.8504311	57.40352
2017.224	38.73251	22.267520	55.19750	13.5514805	63.91354
2017.243	36.62883	17.807882	55.44978	7.8446732	65.41299
2017.262	34.63474	13.986850	55.28262	3.0565193	66.21295
2017.281	33.79458	11.485345	56.10381	-0.3244485	67.91360
2017.300	35.13849	11.095555	59.18143	-1.6320086	71.90900
2017.320	35.19285	9.459581	60.92613	-4.1627910	74.54850
2017.339	33.71608	6.518242	60.91391	-7.8794223	75.31158
2017.358	32.31857	3.821503	60.81564	-11.2639330	75.90107
2017.377	33.64128	3.840132	63.44243	-11.9356449	79.21821
2017.396	34.21867	3.076003	65.36134	-13.4099294	81.84727
2017.415	33.68293	1.274523	66.09133	-15.8814483	83.24730
2017.435	32.48727	-1.061186	66.03573	-18.8206682	83.79522
2017.454	32.50843	-2.133480	67.15035	-20.4718017	85.48867
2017.473	33.60868	-2.151888	69.36924	-21.0823883	88.29974
2017.492	33.79784	-3.074121	70.66980	-22.5929584	90.18864
2017.511	33.04548	-4.866357	70.95731	-24.9356688	91.02662
2017.530	32.46928	-6.421338	71.35990	-27.0087874	91.94735
2017.550	32.91777	-6.947610	72.78315	-28.0510669	93.88660
2017.569	33.51614	-7.334437	74.36672	-28.9594269	95.99171
2017.588	33.48243	-8.324635	75.28949	-30.4559574	97.42081
2017.607	33.14054	-9.572631	75.85372	-32.1836201	98.46471
2017.626	33.12311	-10.472426	76.71864	-33.5505072	99.79672

2017.645 33.38084 -11.101387 77.86306 -34.6488547 101.41053

2017.665 33.26723 -12.095582 78.63004 -36.1092035 102.64366

2017.684 32.99240 -13.219931 79.20473 -37.6832612 103.66806

2017.703 32.99580 -14.037033 80.02863 -38.9347092 104.92631

2017.722 33.32977 -14.516609 81.17614 -39.8449511 106.50449

2017.741 34.13416 -14.525736 82.79406 -40.2847297 108.55305

2017.760 34.20523 -15.253359 83.66382 -41.4351559 109.84562

2017.780 34.25948 -15.973520 84.49248 -42.5652645 111.08423

2017.799 34.60383 -16.389229 85.59690 -43.3833266 112.59100

2017.818 34.60844 -17.142175 86.35906 -44.5372962 113.75418

2017.837 35.23542 -17.266795 87.73763 -45.0597884 115.53062

2017.856 35.60878 -17.628901 88.84646 -45.8112271 117.02879

2017.875 35.45905 -18.498200 89.41630 -47.0614414 117.97954

2017.895 34.95156 -19.718390 89.62152 -48.6589157 118.56204

2017.914 35.80023 -19.578965 91.17943 -48.8949407 120.49540

2017.933 35.04757 -21.031613 91.12674 -50.7181371 120.81327

2017.952 37.02177 -19.743952 93.78749 -49.7939102 123.83745

2017.971 37.76929 -19.673444 95.21202 -50.0817889 125.62036

2017.990 46.29465 -11.820734 104.41004 -42.5851606 135.17447

2018.010 48.17436 -10.608249 106.95696 -41.7258810 138.07460

2018.029 45.78097 -13.659177 105.22112 -45.1248914 136.68684

2018.048 44.76707 -15.320844 104.85498 -47.1294638 136.66360

2018.067 44.15745 -16.572281 104.88718 -48.7206578 137.03556

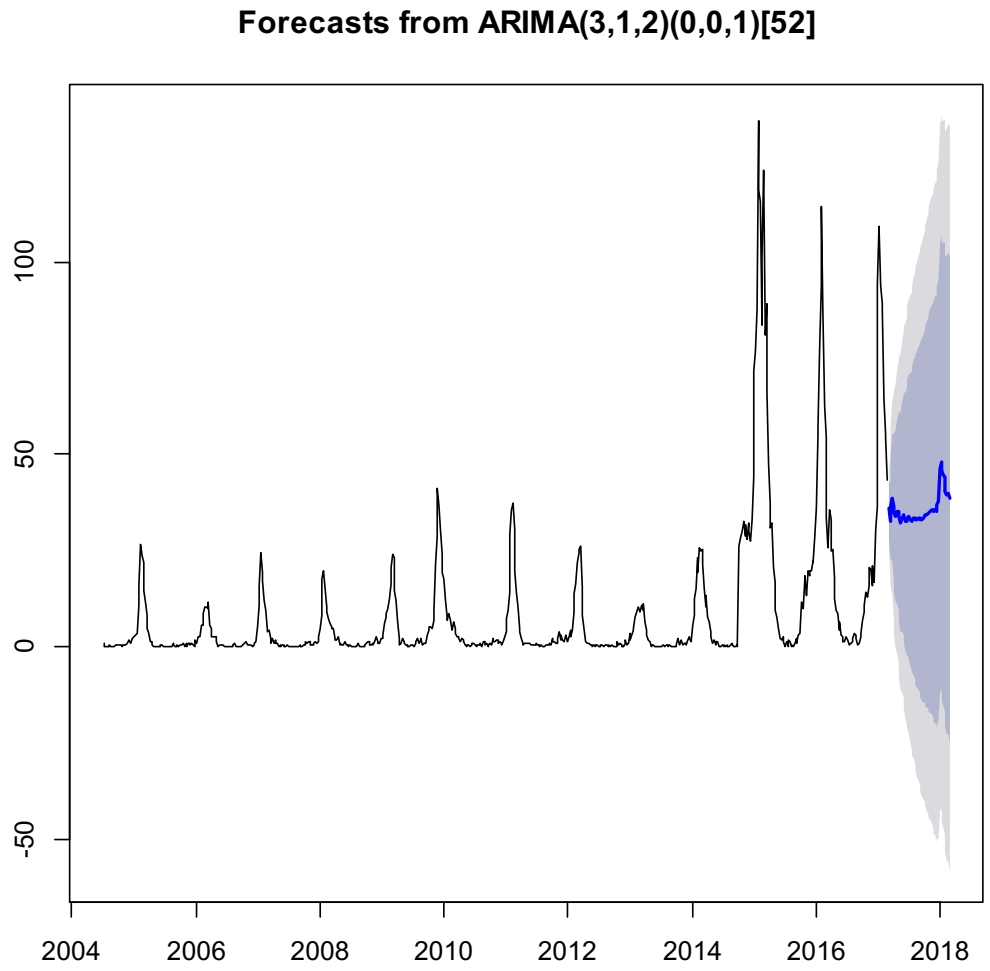
2018.086 40.33719 -21.030098 101.70449 -53.5159802 134.19037

2018.105 39.50564 -22.492665 101.50395 -55.3125863 134.32387

2018.125 39.88966 -22.731270 102.51058 -55.8807860 135.66010

2018.144 38.46848 -24.768264 101.70522 -58.2437733 135.18073

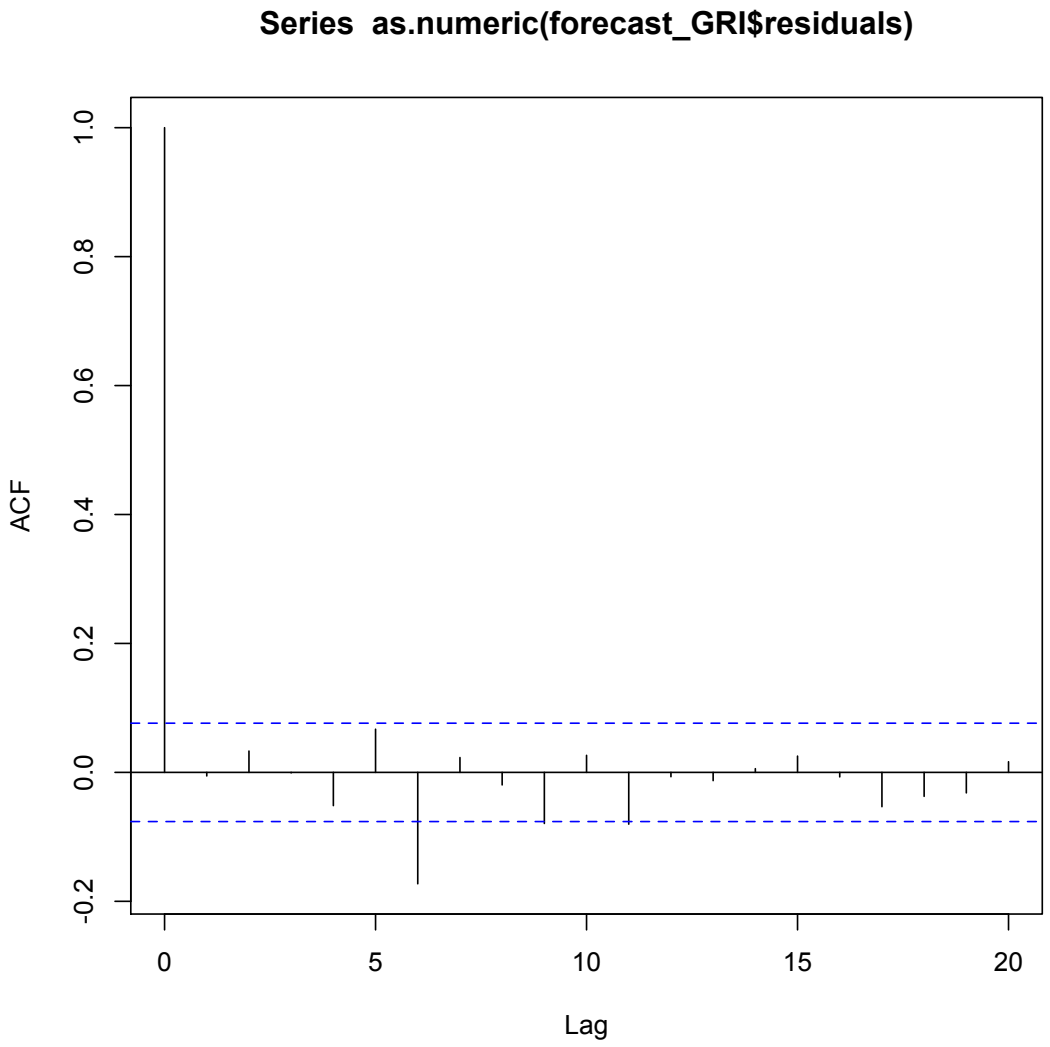
> plot(forecast_GRI)



Οι προγνωστικές τιμές που δίνει η R είναι αποδεκτές αλλά τα διαστήματα εμπιστοσύνης περιλαμβάνουν αρνητικές τιμές το οποίο θα πρέπει να αξιολογηθεί με τη δέουσα προσοχή.

Στη συνέχεια πρέπει να γίνει έλεγχος ανεξαρτησίας και κανονικότητας των προγνωστικών σφαλμάτων. Για να ελέγξουμε την ανεξαρτησία των καταλοίπων φτιάχνουμε το γράφημα αυτοσυσχέτισης και χρησιμοποιούμε Ljung-Box test.

> acf(forecast_GRI$residuals, lag.max=20)



Το διάγραμμα αυτοσυσχέτισης των υπολοίπων δείχνει για την τιμή στην υστέρηση lag=6 να πέφτει έξω από τα όρια εμπιστοσύνης και δύο τιμές να βρίσκονται πάνω στο όριο.

Θα εξετάσουμε με το στατιστικό κριτήριο Ljung – Box, για κάθε υστέρηση μέχρι lag=10, αν υπάρχουν συσχετίσεις μεταξύ των καταλοίπων και αν η τιμή που πέφτει εκτός ορίων στο διάγραμμα των αυτοσυσχετίσεων αποτελεί σημαντική ένδειξη απόρριψης της μηδενικής υπόθεσης, ότι τα κατάλοιπα έχουν συμπεριφορά λευκού θορύβου.

```
> Box.test(forecast_GRI$residuals, lag=1, type="Ljung-Box")

        Box-Ljung test

data:  forecast_GRI$residuals
X-squared = 0.020415, df = 1, p-value = 0.8864

> Box.test(forecast_GRI$residuals, lag=2, type="Ljung-Box")

        Box-Ljung test

data:  forecast_GRI$residuals
X-squared = 0.74769, df = 2, p-value = 0.6881

> Box.test(forecast_GRI$residuals, lag=3, type="Ljung-Box")

        Box-Ljung test

data:  forecast_GRI$residuals
X-squared = 0.74812, df = 3, p-value = 0.8618
```

```
> Box.test(forecast_GRI$residuals, lag=4, type="Ljung-Box")

        Box-Ljung test

data:  forecast_GRI$residuals
X-squared = 2.5192, df = 4, p-value = 0.6412

> Box.test(forecast_GRI$residuals, lag=5, type="Ljung-Box")

        Box-Ljung test

data:  forecast_GRI$residuals
X-squared = 5.5112, df = 5, p-value = 0.3567

> Box.test(forecast_GRI$residuals, lag=6, type="Ljung-Box")

        Box-Ljung test

data:  forecast_GRI$residuals
X-squared = 25.464, df = 6, p-value = 0.0002801
```

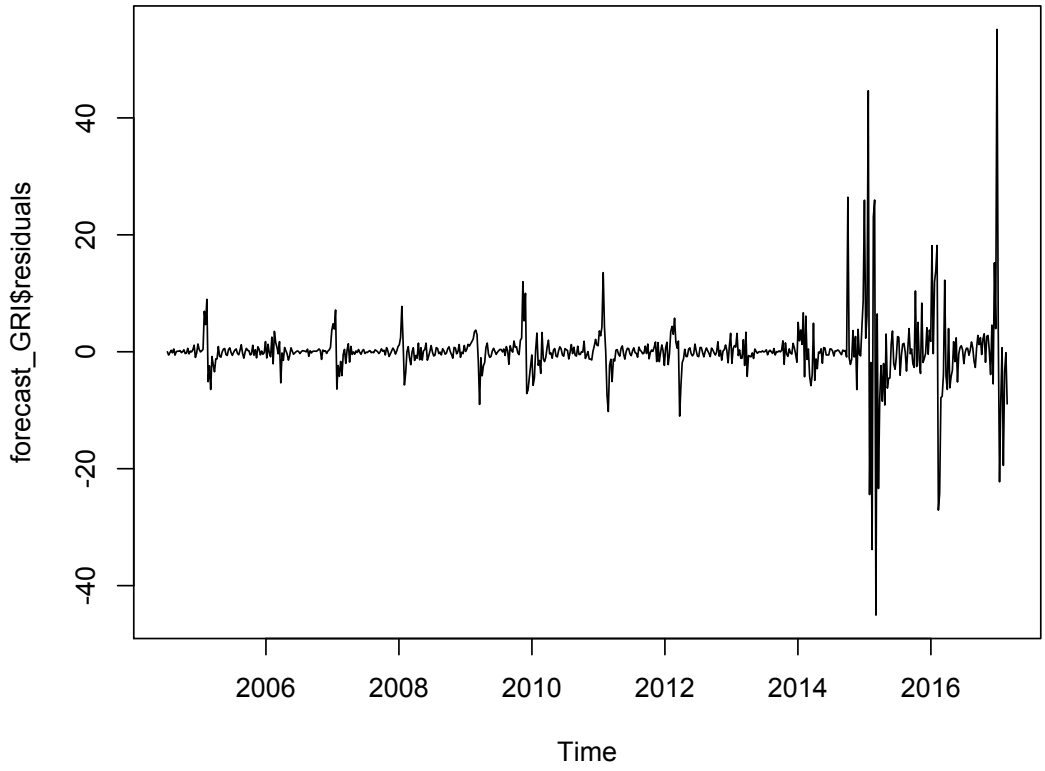
```
> Box.test(forecast_GRI$residuals, lag=7, type="Ljung-Box")

        Box-Ljung test

data:  forecast_GRI$residuals
X-squared = 25.818, df = 7, p-value = 0.0005428

> Box.test(forecast_GRI$residuals, lag=8, type="Ljung-Box")

        Box-Ljung test

data:  forecast_GRI$residuals
X-squared = 26.072, df = 8, p-value = 0.001021

> Box.test(forecast_GRI$residuals, lag=9, type="Ljung-Box")

        Box-Ljung test

data:  forecast_GRI$residuals
X-squared = 30.287, df = 9, p-value = 0.0003921
```

```
> Box.test(forecast_GRI$residuals, lag=10, type="Ljung-Box")

        Box-Ljung test

data:  forecast_GRI$residuals
X-squared = 30.756, df = 10, p-value = 0.0006437
```

Όπως φαίνεται από τα Box-Ljung tests, η μηδενική υπόθεση δεν μπορεί να απορριφθεί μέχρι και για την υστέρηση lag=5, όπου για κάθε υστέρηση μέχρι lag=5 το p-value > 0.05. Το p-value για την υστέρηση lag=6 είναι 0.0002801 < 0.05 που δηλώνει μία ένδειξη συσχέτισης. Λόγω αυτού, για τις υστερήσεις lags = 7, 8, 9, 10 τα p-value < 0.05 παρόλο που οι αντίστοιχες αυτοσυσχετίσεις στο διάγραμμα μένουν εντός του διαστήματος εμπιστοσύνης. Επομένως δεν υπάρχει ισχυρή ένδειξη για συσχέτιση των καταλοίπων και κατ'επέκταση δεν υπάρχουν σοβαρά τεκμήρια για την αντίθετη υπόθεση, δηλαδή ότι τα σφάλματα ακολουθούν συμπεριφορά λευκού θορύβου.

Περαιτέρω μελέτη των καταλοίπων γίνεται εξετάζοντας αν τα σφάλματα πρόγνωσης είναι κανονικά κατανεμημένα με μέση τιμή μηδέν και σταθερή διασπορά.

> plot.ts(forecast_GRI$residuals)   # make time plot of forecast errors



Η μέση τιμή των σφαλμάτων φαίνεται από το διάγραμμα των καταλοίπων να είναι μηδενική. Ελέγχουμε με τη σχετική συνάρτηση στην R.

> mean(forecast_GRI$residuals)

[1] 0.03883779

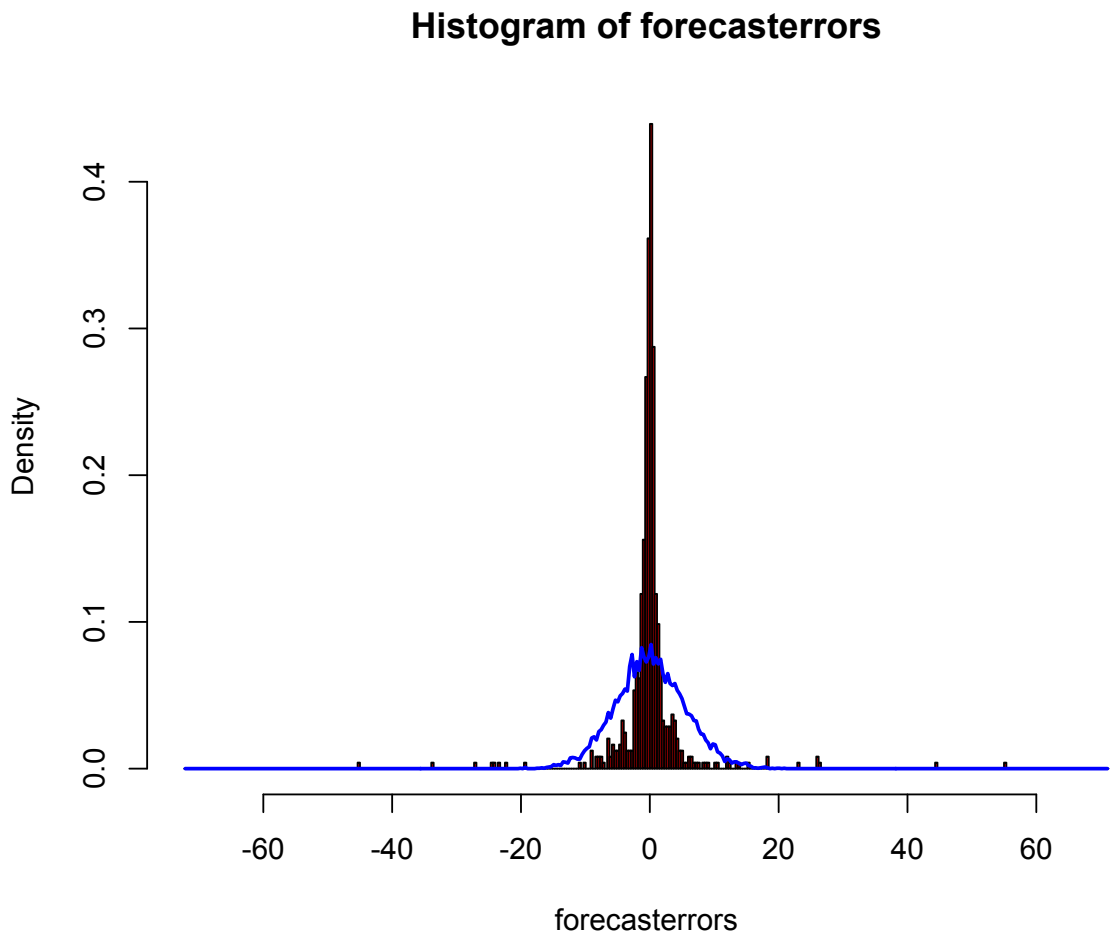
Αν και η μέση τιμή των σφαλμάτων δεν ειναι ακριβώς 0 είναι πάρα πολύ κοντά στο 0 όπως θα έπρεπε να είναι και όπως φαίνεται στο διάγραμμα των σφαλμάτων. Ελέγχουμε και για την κανονικότητα.

Στον παρακάτω κώδικα η διασπορά που υπολογίζεται είναι:

> sd(forecast_GRI$residuals) #mysd

[1] 5.451067

```
> plotForecastErrors <- function(forecasterrors)
+ {
+   # make a histogram of the forecast errors:
+   mybinsize <- IQR(forecasterrors,na.rm=TRUE)/4
+   mysd   <- sd(forecasterrors,na.rm=TRUE)
+   mymin  <- min(forecasterrors,na.rm=TRUE) - mysd*5
+   mymax  <- max(forecasterrors,na.rm=TRUE) + mysd*3
+   # generate normally distributed data with mean 0 and standard deviation mysd
+    mynorm <- rnorm(10000, mean=0, sd=mysd) ποια η τιμη mysd? -την ειχα ορισει πιο πανω   να δωσεις τον αριθμο εννοω
+   mymin2 <- min(mynorm)
+   mymax2 <- max(mynorm)
+   if (mymin2 < mymin) { mymin <- mymin2 }
+   if (mymax2 > mymax) { mymax <- mymax2 }
+    # make a red histogram of the forecast errors, with the normally distributed data overlaid:
+   mybins <- seq(mymin, mymax, mybinsize)
+   hist(forecasterrors, col="red", freq=FALSE, breaks=mybins)
+   # freq=FALSE ensures the area under the histogram = 1
+   # generate normally distributed data with mean 0 and standard deviation mysd
+   myhist <- hist(mynorm, plot=FALSE, breaks=mybins)
+   # plot the normal curve as a blue line on top of the histogram of forecast errors:
+   points(myhist$mids, myhist$density, type="l", col="blue", lwd=2)
+ }
> plotForecastErrors(forecast_GRI$residuals)
```

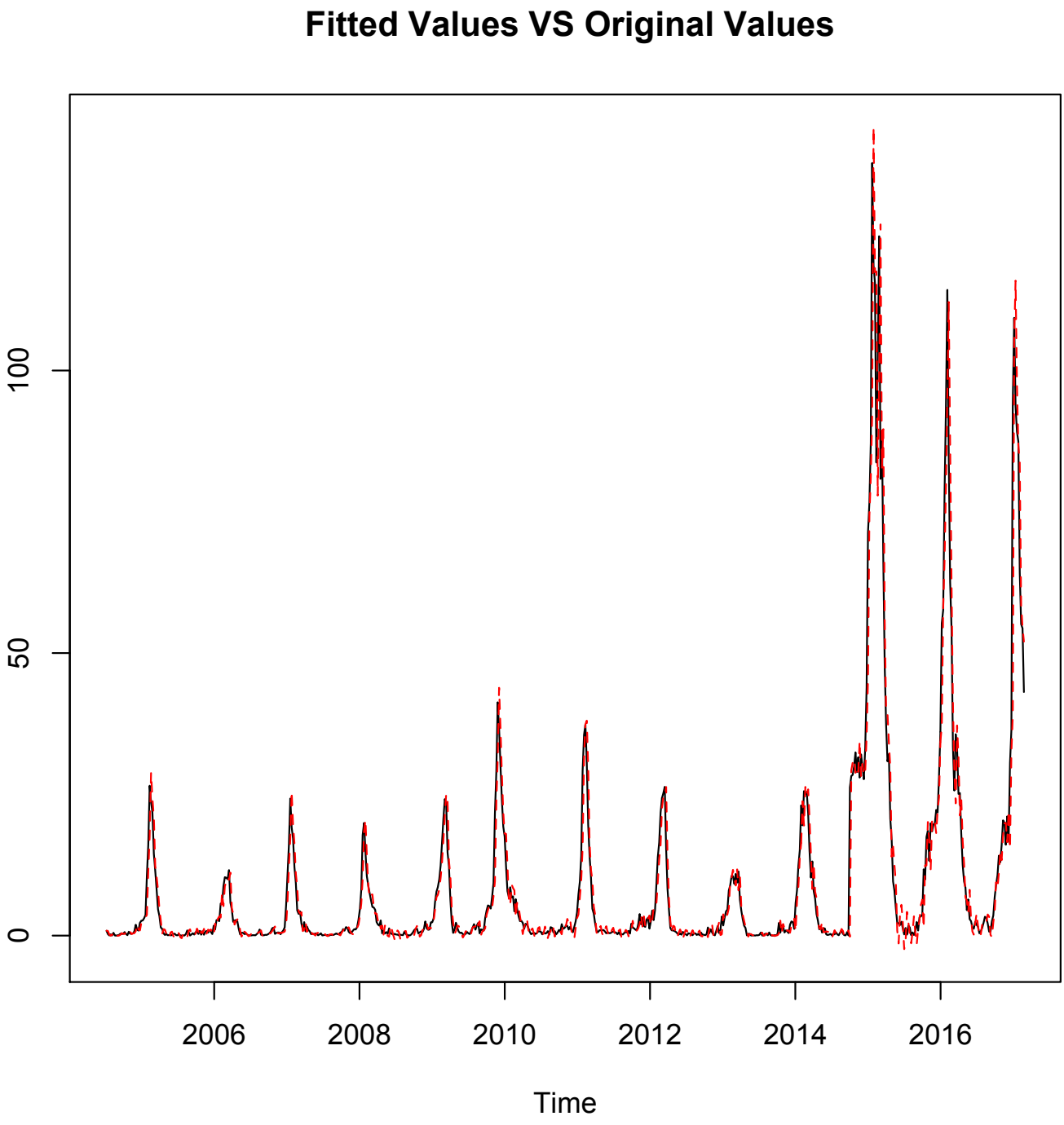


Η γραφική παράσταση των σφαλμάτων πρόγνωσης δείχνει ότι τα σφάλματα πρόγνωσης έχουν σχεδόν σταθερή διακύμανση στο πέρασμα του χρόνου με εξαίρεση κάποιες απομακρυσμένες τιμές και προς τα τελευταία χρόνια που φαίνεται να υπάρχει πιο έντονη αλλαγή.

Από το ιστόγραμμα των σφαλμάτων πρόγνωσης, εκτός από κάποιες απομακρυσμένες τιμές που παρατηρούνται, παρατηρείται ότι η μέση τιμή βρίσκεται γύρω από το μηδέν, όπως περιμέναμε, και έτσι μπορεί να βγει το σμπέρασμα ότι τα σφάλματα πρόγνωσης κατανέμονται σύμφωνα με την κανονική κατανομή με μέση τιμή μηδέν.

Επομένως, μπορούμε να καταλήξουμε ότι το επιλεχθέν μοντέλο SARIMA(3,1,2)(0,0,1)[52] αποτελεί ένα ικανοποιητικό μοντέλο πρόβλεψης για το συνολικό αριθμό γριπωδών συνδρομών. Για την επιβεβαίωση αυτού του συμπεράσματος κατασκευάζουμε το γράφημα των πραγματικών τιμών μαζί με τις εκτιμηθείσες τιμές από το μοντέλο (5.1.1) .

```
> x<-GRI

> y<- fitted.values(forecast_GRI)

> ts.plot(x, y, gpars = list(col = c("black", "red")), lty = 1:2, main= " Fitted Values VS
Original Values")
```



Παρατηρούμε ότι το μοντέλο μας δίνει εκτιμήσεις πολύ κοντά στις πραγματικές τιμές, το οποίο επιβεβαιώνει το παραπάνω συμπέρασμα.

Βιβλιογραφία

Αγγλική

[1] Robert H. Shumway, David S. Stoffer: Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics) Fourth (4th) Edition

[2] Jonathan D. Cryer, Kung-Sik Chan – Time Series Analysis: With Applications in R (Springer Texts in Statistics) 2nd Edition

[3] Hamilton J.D. 1994 “Time Series Analysis”, Princeton

[4] Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. Biometrika, 65(2), 297-303.

[5] Douglas C. Montgomery, Cheryl L. Jennings, Murat Kulahci, “Introduction to Time Series Analysis and Forecasting”, 2008, Wiley

[6] A Course in Time Series Analysis,: Daniel Pena, George C. Tiao and Ruey S. Tsay (Eds.), John Wiley, New York, 2001

[7] Cools, M., Moons, E., & Wets, G. (2009). Investigating the variability in daily traffic counts through use of ARIMAX and SARIMAX models: assessing the effect of holidays on two site locations. Transportation Research Record, 2136(1), 57-66.

[8] Peter J. Brockwell, Richard A. Davis - Introduction to Time Series and Forecasting (2002, Springer)

[9] Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., Shin, Y. (1992). "Testing the null hypothesis of stationarity against the alternative of a unit root". Journal of Econometrics

Ελληνική

[10] Χρήστου Γ. 2002. “Εισαγωγή στην οικονομετρία” Τόμος Β', Εκδόσεις Gutenberg

[11] Δημέλη Σ. 2003. “Σύγχρονες Μέθοδοι Ανάλυσης Χρονολογικών Σειρών”, Εκδόσεις Κριτική

[12] Θαλασσινός, Λ. (1991). Ανάλυση Χρονολογικών Σειρών: Box–Jenkins. Εκδόσεις Σταμούλης.

[13] Π. Θ. Τζωρτζόπουλου (Καθηγητής Στατιστικής Ανώτατης Σχολής Οικονομικών και Εμπορικών Επιστημών) "Ανάλυση Χρονολογικών Σειρών", Εκδόσεις Σμπίλιας, Αθήνα 1991.

[14] Χάλκος Γ. 2004. Σημειώσεις μαθήματος “Χρονολογικές σειρές και προβλέψεις”

[15] Κουγιουμτζής Δημήτρης. Σημειώσεις “Ανάλυση Δεδομένων” (Αν. Καθηγητής Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ, Πολυτεχνική Σχολή, ΑΠΘ)

[16] Μπούτσικας Μιχαήλ. Σημειώσεις “Ανάλυση Χρονολογικών Σειρών”(Αναπληρωτής Καθηγητής Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης Πανεπιστήμιο Πειραιώς)

Ηλεκτρονική

[17] Νικόλαος Δριτσάκης. Σημειώσεις “Υπολογιστικές Τεχνικές Εκτιμητικής”/ μάθημα 4 (Καθηγητής Τμήμα Εφαρμοσμένης Πληροφορικής Πανεπιστήμιο Μακεδονίας) (http://users.uom.gr/~drits/lessons/Lesson%204(MSc%20Inf).pdf)

[18] Γ. Ε. Κοκολάκης. Σημειώσεις Ανάλυσης Χρονοσειρών , Τομέας Μαθηματικών , Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών (http://www.math.ntua.gr/~kokolakis/SEMFE/TimeSeries_Ch_1.pdf)

[19] Κουγιουμτζής Δημήτρης. Σημειώσεις “Ανάλυση Χρονοσειρών” (Αν. Καθηγητής Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ, Πολυτεχνική Σχολή, ΑΠΘ) (http://users.auth.gr/dkugiu/Teach/TimeSeries/TimeSeries.pdf)

[20] Cyclic and seasonal time series (https://robjhyndman.com/hyndsight/cyclicts/)

[21] *http://users.auth.gr/dkugiu/Teach/TimeSeries/Lec2.pdf*

[22] *http://users.uom.gr/~drits/lessons/Lesson%205(MSc%20Inf).pdf*