Στατιστική ανάλυση δεδομένων παραγωγής ηλεκτρικής ενέργειας

Διπλωματική Εργασία

Αθανάσιος Μανώλης

Πολυτεχνείο Κρήτης
Σχολή Μηχανικών Ορυκτών Πόρων
Χανιά,Ελλάδα

Επιβλέπων: Διονύσιος Χριστόπουλος, Καθηγητής, Τμήμα Μηχανικών Ορυκτών Πόρων, Πολυτεχνείου Κρήτης

Τριμελής Επιτροπή:
Διονύσιος Χριστόπουλος, Καθηγητής
Μιχαήλ Γαλετάκης, Καθηγητής, Τμήμα Μηχανικών Ορυκτών Πόρων
Ανδρέας Παυλίδης, Μετά-Διδάκτωρ Ερευνητής, Τμήμα Μηχανικών Ορυκτών Πόρων

Στον Κωνσταντίνο, τον Μιχάλη και την Κωνσταντίνα

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον κ. Διονύσιο Χριστόπουλο που με εμπιστεύτηκε για την εργασία και μου παρείχε τις πολύτιμες γνώσεις του και την βοήθειά του σε οποιοδήποτε πρόβλημα αντιμετώπισα. Η συνεργασία μας ήταν εξαιρετική και η καθοδήγηση του αξιέπαινη.

Ιδιαίτερα ευχαριστίες θα ήθελα να δώσω στον κ. Ανδρέα Παυλίδη για την άψογη καθοδήγηση που μου παρείχε και την αμέριστη βοήθεια που μου έδωσε για την καλύτερη δυνατή διεκπεραίωση της εργασίας αυτής, καθώς και τον κ. Μιχαήλ Γαλετάκη που δέχτηκε να είναι στην επιτροπή αξιολόγησης.

Τέλος θερμές ευχαριστίες και ευγνωμοσύνη εκφράζω στους γονείς μου και την αδερφή μου για την συμπαράσταση και εμψύχωση που μου παρείχαν όλα αυτά τα χρόνια και στους φίλους μου για την ανεκτίμητη στήριξη τους.

Περίληψη

Η παρούσα εργασία ασχολείται διεξοδικά με την πρόβλεψη χρονοσειρών. Οι χρονοσειρές αποτελούν ένα μεγάλο ποσοστό του συνολικού όγκου των δεδομένων που συλλέγουν εταιρίες και επιχειρηματικοί όμιλοι. Μία από τις πιο χρήσιμες εφαρμογές της ανάλυσης χρονοσειρών είναι η δημιουργία πρόβλεψης για μελλοντικές χρονικές στιγμές. Η διαδικασία αυτή έχει τυποποιηθεί μέσω μαθηματικών μοντέλων, τα όποια είναι γνωστά ως μοντέλα πρόβλεψης χρονοσειρών. Η εξαγωγή συμπερασμάτων από την πρόβλεψη αυτή έχει μεγάλη σημασία σε διάφορους τομείς. Για παράδειγμα γίνεται εφικτή η εξαγωγή χρήσιμων πληροφοριών σχετικά με την πιθανή ζήτηση σε ηλεκτρική ενέργεια.

Η παρούσα εργασία ασχολείται διεξοδικά με την περιγραφή βασικών μοντέλων χρονοσειρών και εν τέλει την δημιουργία πρόβλεψης με τα μοντέλα αυτά. Τα διαθέσιμα δεδομένα προς επεξεργασία είναι η παραγωγή ηλεκτρικής ισχύος στην χώρα του Βελγίου για την χρονική περίοδο μεταξύ της 1ης Ιανουαρίου του 2019 έως και τις 30 Σεπτεμβρίου του 2019. Στο πρώτο κεφάλαιο περιγράφονται τα βασικά χαρακτηριστικά και τα στατιστικά μεγέθη των χρονοσειρών. Επιπλέον γίνεται ανάλυση των εννοιών της τάσης και της περιοδικότητας. Στο δεύτερο κεφάλαιο πραγματοποιείται αναφορά στην έννοια της συσχέτισης στον χρόνο μεταξύ τυχαίων μεταβλητών και στην μέθοδο της γραμμικής παλινδρόμησης.

Στο τρίτο κεφάλαιο παρατίθενται οι μέθοδοι που θα χρησιμοποιηθούν για την ανάλυση των δεδομένων. Παρουσιάζεται το εποχιακό αυτοπαλινδρομούμενο μοντέλο κινούμενου μέσου όρου (SARIMA) καθώς και τα μοντέλα που υπάγονται στην οικογένεια του μοντέλου αυτού. Επιπρόσθετα παρατίθεται και η μέθοδος του εκθετικά σταθμισμένου κινούμενου μέσου όρου (EWMA). Τέλος, στο τέταρτο κεφάλαιο εξετάζουμε πως πραγματοποιείται η πρόβλεψη της παραγόμενης ηλεκτρικής ισχύος για καθένα από τα παραπάνω μοντέλα χρονοσειρών.

Στο πέμπτο κεφαλαίο πραγματοποιείται η επεξεργασία των δεδομένων. Τα δεδομένα παραγωγής έχουν χρονικό βήμα ανά 15 λεπτά. Για την επεξεργασία επιλέχθηκαν δύο διαφορετικές χρονικές περίοδοι, μία καλοκαιρινή (από 15 Ιουνίου έως 13 Ιουλίου) και μία χειμερινή (από 15 Ιανουαρίου έως 12 Φεβρουαρίου). Η επεξεργασία εφαρμόζεται και στην χρονοσειρά των μέσων όρων ανά εξάωρο. Επιπλέον γίνεται αξιολόγηση των αποτελεσμάτων των προβλέψεων μέσω στατιστικών μέτρων επιβεβαίωσης. Χρησιμοποιούνται συγκεκριμένα η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE), ο συντελεστής συσχέτισης του Pearson (RPe) και το μέσο σφάλμα (ME). Η αξιολόγηση γίνεται και μέσω διαγραμμάτων αυτοσυσχέτισης (ACF).

Πριν την πρόβλεψη, πραγματοποιείται χρονική παρεμβολή στα δεδομένα της κάθε χρονοσειράς με χρονικό βήμα ανά 15 λεπτά με την βοήθεια της μεθόδου EWMA. Στην συνέχεια, και στις ίδιες χρονοσειρές εφαρμόζεται η μέθοδος EWMA για μελλοντικές προβλέψεις. Για την πρόβλεψη με την μέθοδο EWMA, η επεξεργασία έγινε με χρονικό βήμα ανά 15 λεπτά και για τις δύο (καλοκαιρινή και χειμερινή) χρονικές περιόδους. Εκτιμήθηκε η κατανάλωση ανά ημέρα με βάση τον σταθμισμένο μέσο όρο της κατανάλωσης μιας περιόδου εύρους 192 λεπτών της ώρας νωρίτερα. Οι προβλέψεις με την μέθοδο EWMA οδήγησαν σε μέσο τετραγωνικό σφάλμα που κυμαίνεται από 13,45% έως 14,01% ως ποσοστό της αντίστοιχης μέσης στάθμης για τις δύο περιόδους.

Στην επεξεργασία με το μοντέλο SARIMA χρησιμοποιήθηκε τόσο η χρονοσειρά με βήμα ανά 15 λεπτά (και για τις δύο χρονικές περιόδους) όσο και η χρονοσειρά με βήμα ανά εξάωρο. Αρχικά αφαιρέθηκε το αιτιοκρατικό μέρος (τάση και περιοδικότητα). Βρέθηκαν δύο περιοδικότητες, μία ημερήσια και μία εβδομαδιαία. Για την απαλοιφή αυτών έγινε γραμμική παλινδρόμηση με την βοήθεια του λογισμικού Matlab. Για απαλλαγή εναπομεινάντων αυτοσυσχετίσεων εφαρμόστηκε ένα μοντέλο SARIMA(32, 0, 20)(60, 0, 20) για την χρονοσειρά ανά εξάωρο, ένα μοντέλο SARIMA(0, 2, 7)(104, 2, 7) για την καλοκαιρινή

περίοδο και ένα μοντέλο SARIMA(0,2,9)(106,2,9) για την χειμερινή. Έτσι μετά την απαλοιφή του αιτιοκρατικού και του στοχαστικού μέρους η χρονοσειρά αντιστοιχεί σε λευκό θόρυβο και επομένως είναι εφικτή η εκτίμηση μελλοντικών χρονικών στιγμών. Στην αξιολόγηση των εκτιμήσεων για την χρονοσειρά ανά εξάωρο προέκυψε πως το RPe κυμαίνεται από 97%–98% ενώ το ποσοστό της ρίζας του μέσου τετραγωνικού σφάλματος κυμαίνεται από 3,69% (για βήμα πρόβλεψης μίας ημέρας) έως 4,43% (για πρόβλεψη μία βδομάδα μετά). Στην χρονοσειρά με βήμα 15 λεπτών το RPe κυμαίνεται από 21% έως 76% και το ποσοστό της ρίζας του μέσου τετραγωνικού σφάλματος κυμαίνεται από 15,62% (για 12 ώρες στο μέλλον) έως 44,22% (για δύο ημέρες στο μέλλον) για την καλοκαιρινή περίοδο. Για την χειμερινή περίοδο το RPe κυμαίνεται από 9% έως 68% και ποσοστό της ρίζας του μέσου τετραγωνικού σφάλματος κυμαίνεται από 19,75% (για 12 ώρες στο μέλλον) έως 58,89% (για δύο ημέρες στο μέλλον).

Στο έκτο κεφάλαιο παρατίθενται τα γενικά συμπεράσματα καθώς και οι προτάσεις για μελλοντική έρευνα. Από την μελέτη των μέτρων επιβεβαίωσης προέκυψε πως στην χρονοσειρά ανά 15 λεπτά τόσο η μέθοδος EWMA όσο και τα μοντέλα SARIMA δίνουν αξιόπιστες προβλέψεις. Όμως για την μέθοδο EWMA προκύπτει πως σε εύρος τιμών 5570 (MWatt) για την χειμερινή περίοδο και 4545 (MWatt) για την καλοκαιρινή το ποσοστό της ρίζας του μέσου τετραγωνικού σφάλματος κυμαίνεται από 13,45% έως 14,01% (για μία μέρα μετά). Αντίστοιχα στο ίδιο εύρος τιμών για το μοντέλο SARIMA το ποσοστό του RMSE κυμαίνεται από 15,52%(για 12 ώρες μετά) έως 58,89% (για δύο μέρες μετά). Επομένως η αξιοπιστία της πρόβλεψης είναι μεγαλύτερη με την μέθοδο EWMA έναντι του μοντέλου SARIMA.

Τα ποσοστά σφάλματος που προκύπτουν από την πρόβλεψη με το μοντέλο SARIMA στις χρονοσειρές ανά 15 λεπτά προέρχονται από την απουσία κανονικής κατανομής, την έλλειψη στασιμότητας και την παρουσία αυτοσυσχετίσεων. Τέλος στην πρόβλεψη που έγινε για την χρονοσειρά ανά εξάωρο με το μο-

ντέλο SARIMA καταλήξαμε στο συμπέρασμα της αξιόπιστης πρόβλεψης. Τα ποσοστά των σφαλμάτων είναι μικρά. Συγκεκριμένα ένα ποσοστό της ρίζας του μέσου τετραγωνικού σφάλματος που κυμαίνεται από 3,69% (για μία μέρα μετά) έως 4,49% (για μία βδομάδα μετά) είναι ένα αξιόπιστο ποσοστό για την έκβαση εμπεριστατωμένων συμπερασμάτων.

Λέξεις κλειδιά

Χρονοσειρές, ανάλυση χρονοσειρών, πρόβλεψη χρονοσειρών, μοντέλα χρονοσειρών, παλινδρόμηση.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Εισαγωγή

Η εξέταση οικονομικών, μετεωρολογικών και άλλων δεδομένων είναι ένα κομμάτι της στατιστικής ανάλυσης που αφορά οποιαδήποτε εταιρία, κράτος ή ακόμα και ελεύθερο επαγγελματία προκειμένου να αποκτήσει μία σαφέστερη εικόνα σχετικά με την έρευνά που επιτελεί. Η ανάλυση αυτή γίνεται συνήθως σε δεδομένα τα όποια έχουν την μορφή χρονοσειρών (time series). Οι χρονοσειρές αφορούν την εξέλιξη στον χρόνο συγκεκριμένων μεταβλητών. Πρακτικά η ανάλυση χρονοσειρών είναι το κομμάτι της έρευνας που ασχολείται με συστήματα, διαδικασίες και πρότυπα που εξελίσσονται σε συνάρτηση με τον χρόνο.

Η ανάλυση των χρονοσειρών έχει δύο βασικούς στόχους:

α) Να μελετήσει και να αναγνωρίσει την φύση ενός φαινομένου, το οποίο αντιπροσωπεύει μια ακολουθία παρατηρήσεων

β) Να προβλέψει την μελλοντική εξέλιξη του φαινομένου.

Οι τεχνικές ανάλυσης των χρονοσειρών αναπτύχθηκαν κυρίως για να μπορέσουν να καλύψουν την ανάγκη των οικονομετρικών αναλύσεων για έγκυρες προβλέψεις των διάφορων οικονομετρικών παραγόντων. Ιδιαίτερη σημασία αξίζει να δώσει στο κομμάτι της πρόβλεψης. Στόχος των προβλέψεων με τις χρονοσειρές είναι να γίνουν όσο το δυνατόν πιο εύστοχες και να ελαχιστοποιήσουν τυχόν αποκλίσεις από τις πραγματικές μελλοντικές στιγμές που προσπαθούν να περιγράψουν. Έτσι μια μέθοδος πρόβλεψης συμβάλλει στην σωστή και έγκαιρη

λήψη αποφάσεων. Αν τα δεδομένα με τα οποία επιχειρούμε να κάνουμε την πρόβλεψη είναι ακριβή τότε η αποτελεσματικότητα αυτής μπορεί να είναι πολύ μεγάλη.

Υπάρχουν όμως αρκετή παράγοντες που μπορούν να εισάγουν προβλήματα στις προβλέψεις. Ένας τέτοιος σημαντικός παράγοντας είναι η αβεβαιότητα. Η αβεβαιότητα της εκτίμησης αναφέρεται σε απροσδόκητα γεγονότα τα οποία επηρέασαν την μελλοντική πορεία της χρονοσειράς και έχει σαν αποτέλεσμα τα διάφορα μοντέλα να εμφανίζουν αποκλίσεις από την πραγματική τιμή, αποκλίσεις οι οποίες μπορεί να είναι από μικρές μέχρι και ολικής αστοχίας της εκτίμησης. Ένας ακόμα σημαντικός παράγοντας που επηρεάζει την διαδικασία της πρόβλεψης είναι ο όγκος των δεδομένων. Όταν ο όγκος των δεδομένων δεν είναι μεγάλος, τότε μπορεί να μην υπάρχουν αρκετά δεδομένα για να μπορέσει να γίνει μια σωστή πρόβλεψη-εκτίμηση.

Υπάρχουν δύο βασικές κατηγορίες μεθόδων πρόβλεψης. Η πρώτη κατηγορία περιλαμβάνει τις ποσοτικές προβλέψεις, οι οποίες χρησιμοποιούνται όταν υπάρχει διαθέσιμη πληροφορία για το παρελθόν καθώς και όταν μπορεί να θεωρηθεί πως το πρότυπο συμπεριφοράς μπορεί να διατηρηθεί στο πέρασμα του χρόνου. Η δεύτερη κατηγορία περιλαμβάνει τις ποιοτικές μεθόδους, οι οποίες χρησιμοποιούνται όταν έχουμε διαθέσιμες λίγες ή καθόλου πληροφορίες για το παρελθόν. Στην περίπτωση αυτή απαιτείται προσωπική εμπειρία και γνώσεις για την επίλυση του προβλήματος.

Το κομμάτι της πρόβλεψης είναι ένα κομμάτι της στατιστικής ανάλυσης που εκτείνεται σε πολλά πεδία, όπως τις επιχειρήσεις, την βιομηχανία, την οικονομία, την ιατρική και τις περιβαλλοντολογικές επιστήμες. Τα προβλήματα της πρόβλεψης ταξινομούνται ως βραχυπρόθεσμα, μεσοπρόθεσμα και μακροπρόθεσμα. Τα βραχυπρόθεσμα προβλήματα πρόβλεψης περιλαμβάνουν πρόβλεψη των γεγονότων μόνο μικρών χρονικών περιόδων (ημέρες, εβδομάδες, μήνες) στο μέλλον. Τα μεσοπρόθεσμα προβλήματα εκτείνονται από ένα έως δύο χρόνια

στο μέλλον και τα μακροπρόθεσμα μπορούν να επεκταθούν σε περισσότερα από δύο χρόνια.

Κεφάλαιο 1

Εισαγωγή στις Χρονοσειρές

1.1 Χρονοσειρές

Ως χρονοσειρές (time series) ονομάζονται το σύνολο των παρατηρήσεων πάνω σε μια ποσοτική μεταβλητή που συγκεντρώνονται σε συγκεκριμένες χρονικές στιγμές. Πιο συγκεκριμένα, οι χρονοσειρές αποτελούνται από ένα σύνολο τιμών οι οποίες λαμβάνονται σε ίσες χρονικές περιόδους όπως για παράδειγμα ανά εβδομάδα, ανά έτος ή ανά ημέρα. Η χρονοσειρά είναι μια στοχαστική διαδικασία, αφού οι τιμές τις επηρεάζονται από τυχαίους παράγοντες και η τιμή κάθε χρονικής στιγμής αποτελεί και μια ξεχωριστή τυχαία μεταβλητή.

Ένα πολύ σημαντικό χαρακτηριστικό των χρονοσειρών είναι ότι συχνά αναφέρονται μόνο στο παρελθόν. Επομένως είναι κατάλληλες για εξαγωγή συμπερασμάτων από μια μεταβλητή αλλά κυρίως είναι κατάλληλες μέσω της πληροφορίας από το παρελθόν για προβλέψεις που σχετίζονται με μελλοντικές χρονικές στιγμές. Από μαθηματικής άποψης η χρονοσειρά είναι ένα σύνολο παρατηρήσεων $x_1, x_2, ..., x_n$ όπου ο δείκτης n παριστάνει ισαπέχοντα χρονικά διαστήματα. Οι παρατηρήσεις αυτές αποτελούν συγκεκριμένες τιμές κάποιον τυχαίων με-

ταβλητών $X_1, X_2, ..., X_n$ [19]. Η βασική διάκριση των χρονοσειρών είναι σε συνεχείς και διακριτές. Η πρώτη κατηγορία χρονοσειρών είναι αυτές που η τιμή του φαινομένου παρατηρείται συνεχώς. Ένα χαρακτηριστικό παράδειγμα συνεχών χρονοσειρών είναι η συνεχής παρακολούθηση των σεισμών [13]. Η δεύτερη κατηγορία χρονοσειρών είναι οι διακριτές και είναι αυτές όπου η τιμή του φαινομένου καταγράφεται σε ορισμένα χρονικά διαστήματα. Χαρακτηριστικό παράδειγμα τέτοιων χρονοσειρών είναι η τιμή μίας μετοχής ανά ημέρα ή ο αριθμός των ηλιακών κηλίδων ανά έτος [7]. Οι χρονοσειρές σήμερα εφαρμόζονται σε πολλά πεδία, όπως στην ιατρική [30] ή στα οικονομικά [34].

1.2 Βασικά χαρακτηριστικά χρονοσειρών

Οι βασικές έννοιες που χαρακτηρίζουν μια χρονοσειρά είναι η στασιμότητα, η τάση, η περιοδικότητα ή εποχικότητα και και οι ιδιόμορφες τιμές (outliers). Πριν από την οποιαδήποτε προσπάθεια μοντελοποίησης μια χρονοσειράς πρέπει να διερευνήσουμε εάν εμφανίζει κάποια από τις προαναφερόμενες συμπεριφορές. Η ανάλυση αυτή πραγματοποιείται μέσω γραφικών παραστάσεων και διαγραμμάτων. Η ανάλυση της χρονοσειράς μεσώ της απαραίτητης διερεύνησης μας δίνει την δυνατότητα χρησιμοποίησης αυτής σε ποικίλες εφαρμογές όπως στους επιστημονικούς τομείς. Σκοπός είναι η δημιουργία μοντέλων που θα μπορούν να προβλέψουν τις μελλοντικές τιμές ενός υπό εξέταση μεγέθους [28].

1.2.1 Στασιμότητα

Η στασιμότητα μιας χρονοσειράς διαδραματίζει καθοριστικό ρόλο στην προσπάθεια ανάλυσης της. Για να μπορέσει μια χρονοσειρά να θεωρηθεί στάσιμη πρέπει να μην υπάρχει συστηματική αλλαγή του μέσου όρου και της διασποράς. Αυτό σημαίνει ότι σε περίπτωση που υπάρχει τάση τότε η χρονοσειρά δεν

μπορεί να χαρακτηριστεί από στασιμότητα. Μια στοχαστική διαδικασία μπορεί να είναι πλήρως στάσιμη όταν δεν επηρεάζεται από κάποια μεταβολή στην αρχή μέτρησης του χρόνου. Οι διαδικασίες αυτές καλούνται και στάσιμες n-τάξης. Αυτό σημαίνει ότι η συνάρτηση κατανομής την χρονική στιγμή t είναι ακριβώς η ιδία με την συνάρτηση κατανομής την χρονική στιγμή $t + s$ (Εξ. 1.1).

$$F(x_t, x_{t+1}, ..., x_{t+T}) = F(x_{t+s}, x_{t+1+s}, ..., x_{t+T+s}) \tag{1.1}$$

Μια χρονοσειρά μπορεί να είναι είτε αυστηρώς είτε ασθενώς στάσιμη. Για να μπορέσει μια χρονοσειρά να θεωρηθεί ως αυστηρώς στάσιμη πρέπει η από κοινού κατανομή της πιθανότητας των $X_{t_1}, X_{t_2}, ..., X_t$ να είναι ίδια με την από κοινού κατανομή των $X_{t_1+s}, X_{t_2+s}, ..., X_{t_n+s}$. Δηλαδή, μια πιθανή μετατόπιση της χρονοσειράς κατά s δεν επηρεάζει την από κοινού κατανομή πιθανότητας, η όποια εξαρτάται από τα διαστήματα μεταξύ των $t_1, t_2, ..., t_n$ [24].

Λόγω της δυσκολίας στην εφαρμογή της στασιμότητας ο ορισμός αυτής μπορεί να γίνει πιο ευέλικτος προσδιορίζοντας αυτό το οποίο ονομάζεται ασθενής στάσιμη διαδικασία. Μια χρονοσειρά μπορεί να χαρακτηριστεί ως ασθενώς στάσιμη όταν οι ροπές πρώτης και δεύτερης τάξης είναι σταθερές στον χρόνο. Αυτό σημαίνει ότι μια χρονοσειρά καλείται με αυτόν τον τρόπο όταν ο μέσος όρος και η διασπορά της δεν μεταβάλλονται διαχρονικά. Απαιτείται επίσης η συνδιακύμανση των τιμών της σε δύο χρονικές υστερήσεις να εξαρτάται μόνο από τις χρονικές υστερήσεις και όχι από το χρονικό σημείο για το όποιο υπολογίζεται. Η έκφραση αυτή μπορεί να διατυπωθεί και με την βοήθεια των μαθηματικών σχέσεων (Εξ. 1.2)–(Εξ. 1.4).

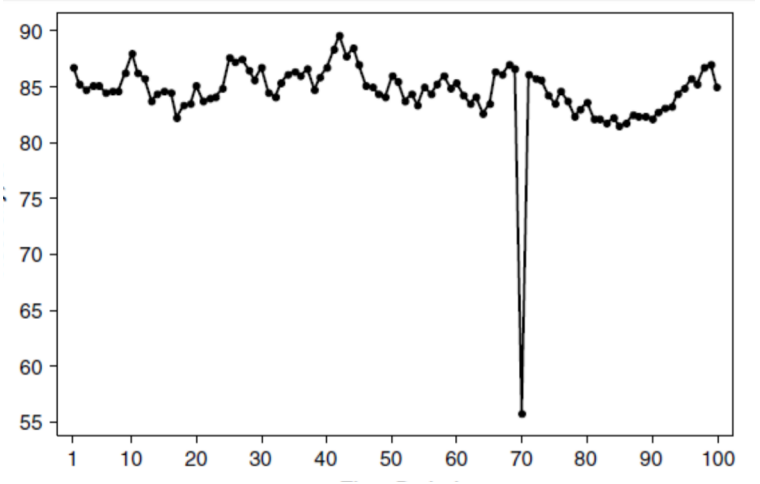
$$E(X_t) = \mu \tag{1.2}$$

$$\mathrm{Var}(X_t) = \sigma^2 \tag{1.3}$$

$$\mathrm{Cov}(X_t, X_{t+s}) = \mathrm{Cov}(X_{t+T}, X_{t+s+T}) = X_s \tag{1.4}$$

1.2.2 Ιδιόμορφες τιμές (Outliers)

Ακραίες τιμές χαρακτηρίζονται οι παρατηρήσεις που εμφανίζονται στην γραφική παράσταση κάποιας χρονοσειράς ως απότομες αλλαγές στο πρότυπο συμπεριφοράς της. Το πρόβλημα με τις ακραίες τιμές είναι ότι δεν μπορούν να προβλεφθούν αλλά η επίδραση τους έχει μικρή χρονική διάρκεια. Για να μπορέσει να γίνει σωστά η ερμηνεία των παρατηρήσεων αυτών απαιτείται ιδιαίτερη προσοχή, αφού χρειάζεται θεωρητική γνώση, κριτική ικανότητα και κατάλληλη εμπειρία. Μια ιδιόμορφη τιμή μπορεί να αναφέρεται σε μια ασυνήθιστη παρατήρηση που οφείλεται σε κάποιο μη αναμενόμενο γεγονός [31]. Ένα χαρακτηριστικό παράδειγμα ιδιόμορφης τιμής φαίνεται στο γράφημα (Σχ. 1.1).



Σχήμα 1.1: Διάγραμμα μεταβολής ιξώδους χημικής διεργασίας [28]

1.3 Τάση και περιοδικότητα

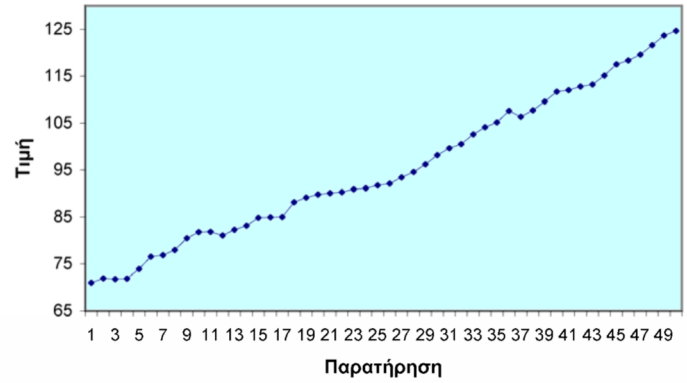
Ως τάση στις χρονοσειρές μπορεί να θεωρηθεί η συστηματική μεταβολή των τιμών ενός χαρακτηριστικού στην μονάδα του χρόνου. Η τάση όταν μπορεί να περιγραφεί από μια γνωστή συνάρτηση, $\mu_t = f(t)$, ονομάζεται αιτιοκρατική τάση. Η αιτιοκρατική τάση περιγράφεται συχνά από μικρού βαθμού πολυώνυμα και αντιστοιχεί σε μεταβολές που συμβαίνουν αργά. Σε περίπτωση που η τάση δεν μπορεί να περιγραφεί από μια γνωστή συνάρτηση αλλά παρουσιάζει αργές μεταβολές με τον χρόνο τότε λέγεται στοχαστική [24].

Η περιοδικότητα είναι ένα επαναλαμβανόμενο μοτίβο υψηλών ή χαμηλών τιμών ενός χαρακτηριστικού, το οποίο μεταβάλλεται ανά τακτά χρονικά διαστήματα. Η μέση τιμή των χρονοσειρών συχνά παρουσιάζει μια αυξητική ή φθίνουσα τάση ή εμφανίζει μια κυκλική επαναλαμβανόμενη συνέχεια σε χρονικά διαστήματα ή εποχές. Πριν την απαλοιφή των δυο αυτών φαινομένων πρέπει να γίνει περαιτέρω επεξεργασία των δεδομένων της χρονοσειράς. Μια χρονοσειρά X_t μπορεί σε κάθε χρονική στιγμή t να αναλυθεί στις συνιστώσες τάσης και περιοδικότητας. Αυτό γίνεται με χρήση του μοντέλου (Εξ. 1.5)

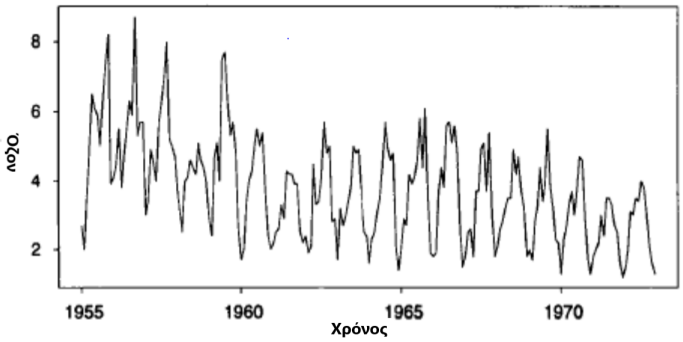
$$X_t = m_t + s_t + Y_t \tag{1.5}$$

όπου η συνιστώσα m_t εκφράζει την τάση ως συνάρτηση του χρόνου, η συνιστώσα s_t είναι μια περιοδική συνάρτηση η οποία εκφράζει την εποχικότητα και το Y_t αποτελεί τον θόρυβο του συστήματος (υπόλοιπα). Επισημαίνεται πως η απαλοιφή της τάσης και της περιοδικότητας γίνεται όταν δεν μας ενδιαφέρουν οι δυο αυτές έννοιες γιατί θεωρούμε ότι δημιουργούνται από άλλους παράγοντες και αιτιοκρατικά φαινόμενα. Σε περίπτωση πρόβλεψης, είτε συμπεριλαμβάνεται η τάση και η περιοδικότητα στο μοντέλο πρόβλεψης, είτε πραγματοποιείται εκτίμηση του μοντέλου της χρονοσειράς αφαιρώντας την τάση και την περιοδικότητα και στις προβλέψεις που προκύπτουν προστίθενται τα μεγέθη αυτά για να προκύψει η πρόβλεψη του παρατηρούμενου μεγέθους.

Στο πρώτο γράφημα (σχ. 1.2) απεικονίζεται ένα χαρακτηριστικό διάγραμμα ανοδικής τάσης [10], ενώ στο δεύτερο (σχ. 1.3) φαίνονται οι μέσες μηνιαίες τιμές του όζοντος στο κέντρο του Λος Άντζελες από το 1955 έως το 1972 [22]. Στο γράφημα αυτό είναι ευδιάκριτη η ύπαρξη εποχικότητας, η οποία είναι υψηλή το καλοκαίρι και χαμηλή τον χειμώνα.



Σχήμα 1.2: Διάγραμμα ανοδικής τάσης



Σχήμα 1.3: Μηνιαίες ενδείξεις όζοντος στο Λος Άντζελες

1.3.1 Απαλοιφή τάσης

Όταν η περιοδικότητα s_t είναι πολύ μικρή μπορούμε να θεωρήσουμε ότι έχουμε μόνο τάση m_t και επομένως ο τύπος της χρονοσειράς δίνεται από την σχέση (Εξ. 1.6).

$$X_t = m_t + Y_t \tag{1.6}$$

όπου το X_t είναι η χρονοσειρά και το Y_t αποτελεί τον θόρυβο του συστήματος (υπόλοιπα).

Σκοπός είναι ο προσδιορισμός της τάσης m_t προκειμένου να γίνει εφικτή η απαλοιφή της. Στην προηγούμενη υπό-ενότητα έγινε αναφορά σχετικά με τον διαχωρισμό της τάσης σε αιτιοκρατική και στοχαστική. Έτσι εάν η τάση είναι καθοριστική τότε προσδιορίζεται με την βοήθεια πολυωνύμου μικρού βαθμού όπως το πολυώνυμο (Εξ. 1.7), ενώ την στοχαστική την απαλείφουμε παίρνοντας πρώτες διαφορές [17].

$$m_t = f(t) = a_0 + a_1 t + ... + a_p t \tag{1.7}$$

Για να προσδιοριστούν οι παράμετροι $a_0, a_1, ..., a_p$ χρησιμοποιείται η μέθοδος των ελαχίστων τετραγώνων (ενότητα 2.3). Έχοντας μελετήσει την χρονοσειρά X_t σε εύρος χρόνου n μπορούμε να εκτιμήσουμε τους $(a_0, a_1, ..., a_p)$ συντελεστές με την βοήθεια της μεθόδου ελαχίστων τετραγώνων και κατ΄ επέκταση να απαλείψουμε την τάση [28].

Τρόποι αντιμετώπισης της στοχαστικής τάσης

Στην περίπτωση που η τάση έχει στοχαστικό χαρακτήρα τότε η απαλείφει πρέπει να πραγματοποιηθεί με την βοήθεια των πρώτων διαφορών και συγκεκριμένα μέσω του μετασχηματισμού

$$X_t = \nabla Y_t - Y_{t-1} = (1 - B)Y_t \tag{1.8}$$

όπου ∇ δηλώνει τον τελεστή της διαφοράς πρώτης τάξης και B είναι ο τελεστής υστέρησης (lag operator).

Σε περίπτωση που δεν αρκούν οι πρώτες διαφορές για να επιτευχθεί στασιμότητα, μπορούμε να εφαρμόσουμε μετασχηματισμό δεύτερης τάξης της μορφής

$$X_t = \nabla^2 Y_t = \nabla(\nabla Y_t) = (1-B)(1-B)Y_t = \\ = (1 - 2B + B^2)Y_t = Y_t - 2Y_{t-1} + Y_{t-2} \quad (1.9)$$

Τις περισσότερες φορές όμως χρησιμοποιούνται διαφορές πρώτης τάξης, διότι είναι αρκετές προκειμένου να εξαλείψουν την εμφάνιση τάσης στοχαστικού χαρακτήρα.

Απαλοιφή τάσης με την βοήθεια του κινούμενου μέσου

Ένας ακόμα τρόπος για τον προσδιορισμό της τάσης είναι με την χρήση του κινούμενου μέσου όρου τάξης $2q + 1$ (Moving average (MA) filter). Για κάθε χρονική στιγμή t, $q < t \leq n - q$ η τάση m_t της χρονοσειράς $(x_1, x_2...., x_n)$ προσδιορίζεται από τον μέσο των παρατηρήσεων στο διάστημα [t-q, t+q] [18]. Επομένως προκύπτει η σχέση (Εξ. 1.10).

$$m_t = \frac{1}{2q+1} \sum_{j=-q}^{q} (x_{t-j}) \quad (1.10)$$

Έτσι οι τιμές $(m_{q+1}, m_{q+2}..., m_{n-q+1})$ αφαιρούνται από τις αρχικές παρατηρήσεις και η χρονοσειρά που προκύπτει από την αφαίρεση αυτή είναι πλέον απαλλαγμένη από τάσης. Συνήθως οι πρώτες και οι τελευταίες παρατηρήσεις θέτονται ίσες με τις αρχικές ή δεν λαμβάνονται υπόψιν. Η επιλογή του q είναι πολύ σημαντική. Για παράδειγμα, εάν προσπαθήσουμε να απαλείψουμε μόνο

αργές μεταβολές της τάσης, τότε θα πρέπει να χρησιμοποιήσουμε μεγάλη τάξη, ενώ σε περίπτωση μεταβολών μικρότερης χρονικής κλίμακας η τάξη πρέπει να είναι μικρή [32]. Επιπλέον εάν δεν γνωρίζουμε την μορφή της υπάρχουσας τάσης, τότε δεν μπορούμε να γνωρίζουμε με βεβαιότητα ποία από τις προαναφερθείσες μεθόδους μπορούμε να χρησιμοποιήσουμε. Για παράδειγμα αν θέλουμε να μελετήσουμε τις πιθανές διακυμάνσεις ενός ημερήσιου δείκτη σε χρονικό ορίζοντα εβδομάδας ή μήνα, τότε είναι ασφαλέστερη η απαλοιφή της τάσης με κινούμενο μέσο τάξης 7 ή 30 [21].

Πιο διεξοδικά η μέθοδος του κινούμενου μέσου εφαρμόζεται σε χρονοσειρές στις οποίες δεν ευσταθεί η θεώρηση ότι η τάση m_t διατηρεί σταθερές τιμές καθ΄ όλη την διάρκεια του χρόνου. Στις χρονοσειρές αυτές εφαρμόζεται μια μέθοδος τριών βημάτων [18]. Στο πρώτο βήμα έχουμε την εξομάλυνση. Κατά την διαδικασία αυτή σε περίπτωση άρτιας περιόδου ο κινούμενος μέσος όρος είναι

$$\widehat{m}_t = \frac{0.5x_{t-q} + x_{t-q+1} + ... + x_{t+q-1} + 0.5x_{t+q}}{d} \tag{1.11}$$

ενώ σε περίπτωση περιττής περιόδου τότε ο κινούμενος μέσος δίνεται από την εξίσωση (Εξ. 1.12).

$$\widehat{m}_t = \frac{x_{t-q} + x_{t-q+1} + ... + x_{t+q-1} + x_{t+q}}{d} \tag{1.12}$$

όπου d εκφράζει την τιμή της περιόδου. Έτσι προκύπτει πως και στις δύο περιπτώσεις εξομάλυνση του κινούμενου μέσου με την μορφή

$$\widehat{m}_t = \sum_{j=-q} qa_j x_{t+j} \tag{1.13}$$

Με την παραπάνω διαδικασία προσδιορίζεται η τάση m_t και εξαλείφεται η δράση της περιοδικότητας s_t. Στην συνέχεια πηγαίνοντας στο δεύτερο βήμα λαμβάνονται υπόψιν οι μέσοι όροι

$$w_l = \frac{1}{k}\sum_{k=1}^{k}(x_{(k-1)d+l}) - \widehat{m}_{(k-1)d+l} \tag{1.14}$$

όπου $k = n/d$ ο αριθμών των περιόδων της χρονοσειράς. Οι μέσοι αυτοί εκφράζουν εποχικότητα. Επομένως αντί της προηγούμενης σχέσης χρησιμοποιούμε τις διαφορές $w_l - \overline{w}$ και προκύπτει πως η συνιστώσα της εποχικότητας δίνεται από την σχέση (Εξ. 1.15).

$$\widehat{s}_l = w_l - \overline{w} = w_l - \frac{1}{d}\sum_{l=1}^{d} w_l \tag{1.15}$$

όπου εν τέλει προκύπτει ότι $\bar{s}_t = \bar{s}_l$. Τέλος στο τρίτο βήμα έχοντας προσδιορίσει την εποχικότητα s_t και λαμβάνοντας υπόψιν την αρχική χρονοσειρά μπορούμε να απαλείψουμε την εποχικότητα μέσω των διαφορών της εξίσωσης (Εξ. 1.16).

$$d_t = x_t - \widehat{s}_t \tag{1.16}$$

Μετά την επιτέλεση των βημάτων αυτών γίνεται επαλήθευση γραφικά προκειμένου να εξακριβωθεί η στασιμότητα ή όχι της χρονοσειράς η οποία μετά την απαλοιφή της τάσης μπορεί να χαρακτηριστεί πλέον ως θόρυβος του συστήματος.

1.3.2 Απαλοιφή περιοδικότητας

Αν υποθέσουμε ότι η ίδια χρονοσειρά που μελετήθηκε προηγουμένους έχει μόνο περιοδικότητα τότε η εξίσωση που προκύπτει είναι η εξής

$$X_t = s_t + Y_t \tag{1.17}$$

όπου θεωρούμε ότι η περίοδος της χρονοσειράς είναι γνωστή και ίση με d. Σε περίπτωση που η περιοδικότητα είναι γνωστή και αντιστοιχεί σε κάποια περίοδο 24 ωρών ή μίας εβδομάδας τότε αναφέρεται ως εποχικότητα. Γενικά όταν η περίοδος είναι γνωστή τότε ένας καλός τρόπος εκτίμησης και απαλοιφής της s_t είναι με τους μέσους όρους των στοιχείων της περιοδικής συνάρτησης s_i (i=1 ,..., d) [24]. Εάν ονομάσουμε κ των αριθμό των περιόδων της χρονοσειράς

X_t τότε το κάθε δεδομένο της περιοδικής συνάρτησης προσδιορίζεται από την εξίσωση (Εξ. 1.18).

$$\widehat{s_i} = \frac{1}{k}\sum_{j=1}^{k}(x_{i+jd}) \tag{1.18}$$

Ο δεύτερος τρόπος προσδιορισμού της περιοδικότητας είναι με την βοήθεια του κινούμενου μέσου όρου θέτοντας την τάξη ίση με την περίοδο d. Με την χρήση του φίλτρου αυτού είναι εφικτή η απαλοιφή της περιοδικότητας. Για να γίνει η εκτίμηση της εποχικότητας με ακρίβεια θα πρέπει να υπολογιστεί η διαφορά της αρχικής χρονοσειράς $(x_{q+1}, x_{q+2}, ..., x_{n-q+1})$ και του κινούμενου μέσου $\widehat{\mu}_{q+1}, \widehat{\mu}_{q+2}, ..., \widehat{\mu}_{n-q+1}$. Η διαφορά αυτή είναι ίση με $w_t = w_{i+jd}$. Έπειτα παίρνουμε τον μέσο όρο των $w_t = w_{i+jd}$ και τον συμβολίζουμε με $\overline{w}_i$ [24]. Αν το άθροισμα των $\overline{w}_i$ δεν είναι 0 τότε αφαιρούμε την μέση τιμή και η περιοδική συνάρτηση δίνεται από την σχέση (Εξ. 1.19).

$$\widehat{s_i} = \overline{w}_i - \frac{1}{d}\sum_{j=1}^{d} w_j \tag{1.19}$$

1.3.3 Απαλοιφή τάσης και περιοδικότητας

Υπάρχουν πολλές περιπτώσεις χρονοσειρών όπου χαρακτηρίζονται και από τάση και από περιοδικότητα. Τότε συν διάζονται οι μέθοδοι απαλοιφής των δύο φαινομένων. Η σειρά εφαρμογής αυτών δεν είναι καθορισμένη.

Όταν στην γραφική παράσταση της χρονοσειράς υπάρχει εποχικότητα, τότε για την εκτίμηση της τάσης m_t θα πρέπει να γίνει εκτίμηση και της περιοδικότητας s_t. Η εκτίμηση των δυο αυτών συνιστωσών γίνεται μέσω επαναληπτικών διαδικασιών, με σκοπό τον διαδοχικό προσδιορισμό των παραμέτρων ή με μια διαδικασία ταυτόχρονου προσδιορισμού και των δύο. Στα πλαίσια της θεωρίας και για απλούστευση της διαδικασίας απαλοιφής θεωρούμε ότι η συνιστώσα s_t έχει μόνο μια περιοδικότητα σε συνάρτηση με τον χρόνο, με γνωστή περίοδο d,

για παράδειγμά εβδομαδιαία. Το φαινόμενο αυτό είναι πολύ σύνηθες σε μετεωρολογικά δεδομένα [5]. Έτσι αν έχουμε μια χρονοσειρά X_t η οποία παρουσιάζει εβδομαδιαία περιοδικότητα με περίοδο d=7 και ο χρόνος t είναι σε μέρες, απαραίτητη προϋπόθεσή είναι η εποχιακή συνιστώσα s_t να ικανοποιεί τις εξίσωσης $s_t = s_{t+d}$ και $\sum_{l=1}^{d} s_{t+1} = 0$.

Γενικά υπάρχουν αρκετές μέθοδοι για τον προσδιορισμό των φαινομένων αυτών όπως η μέθοδος χαμηλής τάσης, η μέθοδος κινούμενου μέσου και η μέθοδος διαφορών με υστέρηση d [7]. Στα πλαίσια την εργασίας αυτής έγινε μια σύντομη ανάπτυξη της μεθοδολογίας κινούμενου μέσου μίας και είναι η μέθοδος που χρησιμοποιήθηκε.

1.4 Στατιστικά μεγέθη χρονοσειράς

Όπως αναφέρθηκε προηγουμένως η χρονοσειρά είναι μια στοχαστική διαδικασία. Θα παρουσιαστούν επομένως τα βασικά στατιστικά μεγέθη μίας χρονοσειράς και συγκεκριμένα η μέση τιμή, η αυτοσυνδιακύμανση, η αυτοσυσχέτιση, ο λευκός θόρυβος και ο τυχαίος περίπατος [19].

1.4.1 Μέση τιμή

Η μέση τιμή μια χρονοσειράς προκύπτει από την σχέση (Εξ. 1.20)

$$\mu_t = E(X_t) = \int_{\infty}^{-\infty} x_t f_{X_t}(x_t) dx_t \tag{1.20}$$

Η μέση τιμή σχετίζεται άμεσα με την τάση της χρονοσειράς, αφού αν αυτή παρουσιάζει αυξητική ή πτωτική τάση σε ένα χρονικό διάστημα, αυτό θα φαίνεται στην μέση τιμή ως συνάρτηση του χρόνου.

1.4.2 Αυτοσυνδιακύμανση

Αν υποθέσουμε ότι έχουμε 2 τυχαίες μεταβλητές X_i και X_j. Η συνδιακύμανση των δύο αυτών μεταβλητών δίνεται από την σχέση

$$\mathrm{Cov}(X_i, X_j) = E(X - \mu_{X_i})(X_j - \mu_{X_j}) \tag{1.21}$$

Για την χρονοσειρά που μελετάμε στην θέση της μεταβλητής W μπορούμε να θέσουμε μία τιμή του X για διαφορετική χρονική περίοδο t και μέσω αυτού να προσδιορίσουμε την αυτοσυνδιακύμανση j-οστής τάξης γ_{jt} για την τυχαία μεταβλητή X_t σύμφωνα με την σχέση (Εξ. 1.22).

$$\gamma_{jt} = E(Y_t - \mu_t)(Y_{t-j} - \mu_{t-j}) = \int_{\infty}^{-\infty}\int_{\infty}^{-\infty}...\int_{\infty}^{-\infty}(x_t - \mu_t)(x_{t-j} - \mu_{t-j})x$$
$$f_{X_t, X_{t-1}, ..., X_{t-j}}(x_t, x_{t-1}, ..., x_{t-j})dx_t dx_{t-1}...dx_{t-j} \tag{1.22}$$

1.4.3 Συνάρτηση αυτοσυσχέτισης

Ο συντελεστής συσχέτισης ρ αποτελεί το μέτρο για τον βαθμό της σχέσης που υπάρχει μεταξύ δύο μεταβλητών. Ο συντελεστής αυτός παίρνει μόνο τιμές που βρίσκονται στο κλειστό διάστημα [-1, 1] και ανάλογα με την ακριβή τιμή που λαμβάνει προσδιορίζεται και η συσχέτιση του. Για τον συντελεστή συσχέτισης ισχύουν τα ακόλουθα [24]:

- Αν $\rho > 0$, τότε σημαίνει ότι έχουμε θετική συσχέτιση και συγκεκριμένα όσο πιο κοντά είναι το ρ στο 1 τόσο πιο ισχυρή συσχέτιση αναπτύσσεται.
- Αν $\rho < 0$, τότε έχουμε αρνητική συσχέτιση και μάλιστα όπως και πριν όσο πιο κοντά είναι η τιμή του ρ στο -1 τόσο πιο ισχυρή συσχέτιση θα παρουσιαστεί.

- Αν $\rho = 1$ ή $\rho = -1$, τότε υπάρχει η μέγιστη δυνατή συσχέτιση, η οποία είναι θετική όταν $\rho = 1$ και είναι αρνητική όταν $\rho = -1$.
- Εάν το $\rho = 0$, τότε δεν υπάρχει καθόλου συσχέτιση.

Όταν πραγματοποιείται ανάλυση χρονοσειρών, ο συντελεστής συσχέτισης μεταξύ δύο μεταβλητών ονομάζεται συντελεστής αυτοσυσχέτισης και είναι πρακτικά η κανονικοποίηση της συνδιακύμανσης με την διασπορά. Έτσι η αυτοσυσχέτιση με υστέρηση κ ορίζεται ως

$$\rho_k = \frac{\gamma_k}{\gamma_0} \tag{1.23}$$

Το διάγραμμα των αυτοσυσχετίσεων λέγεται ACF(autocorellation funcion) και αποτελεί βασικό εργαλείο για την αναγνώριση του μοντέλου. Επιπλέον η συνάρτηση αυτοσυσχέτισης είναι συμμετρική ως προς την αρχή των αξόνων και έτσι παίρνουμε μόνο το θετικό μισό της συνάρτησης, ενώ η γραφική παράσταση αυτής λέγεται διάγραμμα αυτοσυσχέτισης.

Πρακτικά, για να είναι μια χρονοσειρά στάσιμη θα πρέπει το διάγραμμα των αυτοσυσχετίσεων να φθίνει αμέσως μετά τις πρώτες υστερήσεις κ. Σε περίπτωσή που η χρονοσειρά παρουσιάζει εποχικότητα τότε υπάρχουν ισχυρές σχέσης αυτοσυσχέτισεις σε συγκεκριμένες υστερήσεις ανάλογα με την μορφή της περιοδικότητας (ημερήσια, μηνιαία, ετησια...). Για αυτόν τον λόγο υπάρχει ο συντελεστής μερικής αυτοσυσχέτισης (PACF), ο οποίος χρησιμοποιείται συνήθως σε χρονοσειρές οι οποίες χαρακτηρίζονται από περιοδικότητα και μετράει το βαθμό συσχέτισης μεταξύ των X_t και X_{t-k} [24], όταν η επίδραση των άλλων χρονικών υστερήσεων παραμένει σταθερή.

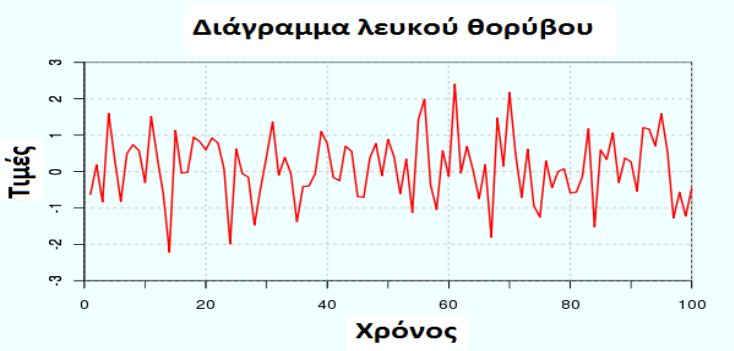
1.4.4 Λευκός θόρυβος

Ο λευκός θόρυβος αποτελεί ένα από τα βασικά στοιχεία κατά την μελέτη μιας χρονοσειράς, για αυτό και η μελέτη αυτού έχει σημαντικό ρόλο. Μπορεί μεταξύ

δύο τυχαίων μεταβλητών να υπάρχει μηδενική συσχέτιση, αυτό δεν σημαίνει όμως ότι οι μεταβλητές αυτές είναι και ανεξάρτητες. Μια χρονοσειρά μπορεί να μην περιέχει γραμμικές συσχετίσεις και τα στοιχεία της να μην είναι ανεξάρτητα μεταξύ τους. Έτσι ένα σύνολο ασυσχέτιστων τυχαίων μεταβλητών καλείται λευκός θόρυβος και συμβολίζεται με $\mathrm{WN}(0, \sigma_x^2)$, ενώ έχει μέση τιμή ίση με το 0 και διακύμανση σ_w^2. Ο μαθηματικός συμβολισμός του λευκού θορύβου είναι $E[X_i X_j] = \delta_{ij}\sigma_x^2$ και ισχύει για οποιεσδήποτε δύο τυχαίες μεταβλητές της χρονοσειράς X_t [16].

Η ονομασία του λευκού θορύβου οφείλεται στο ότι η χρονοσειρά αυτή χρησιμοποιείται σε ευρεία κλίμακα σε εφαρμογές μηχανικών και αναφέρεται σε μεταβολές θορύβου που παρομοιάζουν την ιδιότητα του λευκού φωτός. Αξίζει να σημειωθεί πως σε κάποια συγγράμματα ο όρος λευκός θόρυβος χρησιμοποιείται για χρονοσειρές ασυσχέτιστες μεταξύ τους, ενώ πολλές φορές ο όρος ταυτίζεται με την χρονοσειρά των ανεξάρτητων και ισόνομων τυχαίων μεταβλητών. Σε περίπτωση που τα στοιχεία της χρονοσειράς του λευκού θορύβου ακολουθούν την κανονική κατανομή τότε η χρονοσειρά λέγεται Γκαουσιανός λευκός θόρυβος [24].

Η απεικόνιση μιας γραφικής παράστασης λευκού θορύβου παρουσιάζεται παρακάτω μέσω του γραφήματος (Σχ. 1.4).



Σχήμα 1.4: Χρονοσειρά λευκού θορύβου [31]

Διαδικασίες λευκού θορύβου

Οι διαδικασίες του λευκού θορύβου χωρίζονται στις εξής κατηγορίες:

- Λευκός θόρυβος: Είναι μια χρονική σειρά ϵ_t με $E(\epsilon_t) = 0$, $E(\epsilon_t^2) = \sigma^2$ και ασυσχέτιστους μεταξύ τους όρους.
- Ανεξάρτητος λευκός θόρυβος: Ισχύει σχεδόν ότι και για τον λευκό θόρυβο, με την διαφορά ότι οι όροι της σειράς δεν είναι απλά ασυσχέτιστοι αλλά και ανεξάρτητοι μεταξύ τους, δηλαδή: $E(g(\epsilon_t)f(\epsilon_{t-k} = 0)), \forall\, k > 0$, $\forall\ f : R \to R, \forall\ g : R \to R$.
- Κανονικός ή Gaussian λευκός θόρυβος: Ισχύουν ότι και στις δύο παραπάνω περιπτώσεις μόνο που αυτήν την φορά η ϵ_t ακολουθεί κανονική κατανομή.

1.4.5 Τυχαίος περίπατος

Ο τυχαίος περίπατος είναι μια μη στάσιμη χρονοσειρά X_t για την οποία είναι γνωστό ότι κάθε όρος X_t προκύπτει από την αμέσως προηγούμενη μεταβλητή X_{t-1} στην οποία όμως προστίθεται ένα τυχαίο στοιχείο Y_t [19]. Η έκφραση αυτή διατυπώνεται καλύτερα μέσω της εξίσωσης (Εξ. 1.24).

$$X_t = X_{t-1} + Y_t + \delta \tag{1.24}$$

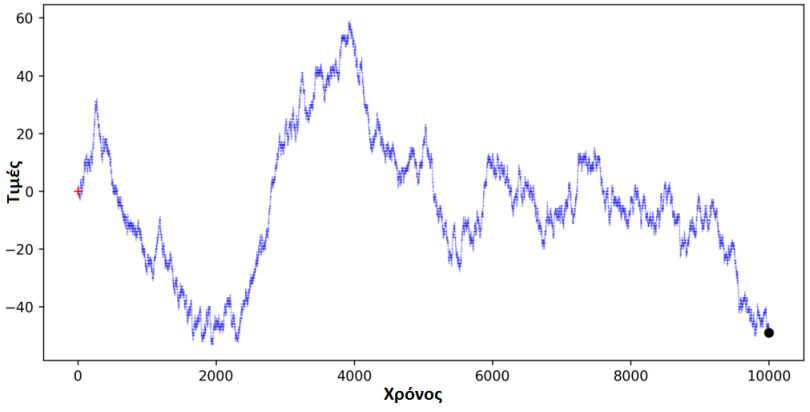
Το δ περιγράφει την τάση. Εάν το $\delta \neq 0$ τότε έχουμε έναν τυχαίο περίπατο με τάση. Από μαθηματικής άποψης ο τυχαίος περίπατος προέρχεται από το γεγονός ότι όταν $\delta = 0$ η τιμή της χρονοσειράς σε χρόνο t ισούται με την τιμή αυτής σε χρόνο $t-1$ συν τον λευκό θόρυβο X_t . Ο τυχαίος περίπατος αποτελεί μια ιδιαίτερη εφαρμογή του αυτοπαλινδρομούμενου μοντέλου πρώτης τάξης. Η θεωρία αυτού περιγράφει μια κατάσταση κατά την οποία η πιθανότητα

μετάβασης από μια κατάσταση σε μια άλλη είναι ακριβώς η ίδια για όλες τις πιθανές μεταβάσεις και δεν εξαρτάται σε καμία περίπτωση από παρελθοντικά δεδομένα. Αξίζει να αναφερθεί πως η χρησιμοποίηση του τυχαίου περιπάτου είναι ιδιαίτερα συχνή στο χρηματιστήριο [24].

Από τον ορισμό του τυχαίου προκύπτει ότι εάν αρχίσουμε από κάποια τιμή X_0 και αντικαθιστώντας επαναληπτικά τον ορισμό (Εξ. 1.24) για τυχαίους χρόνους t, ο ορισμός πλέον του τυχαίου περιπάτου θα γίνει

$$X_t = \sum_{k=0}^{t} Y_k \tag{1.25}$$

Η νέα αυτή αποτύπωση του τυχαίου περιπάτου δηλώνει πως μπορεί να προκύψει από το άθροισμα όλων των τυχαίων βημάτων ως προς t. Επιπλέον μέσω της σχέσης αυτή μπορεί να γίνει ξεκάθαρο πως η μέση τιμή είναι $E(X_t) = 0$ και η διασπορά είναι $\sigma_Y^2 = E[X_t^2] = t\sigma_x^2$. Η σχέση της διασποράς αυτής μας δίνει ότι η διασπορά του τυχαίου περιπάτου είναι ανάλογη του χρόνου και επομένως πρόκειται για μια μη στάσιμη χρονοσειρά. Ένα χαρακτηριστικό μοντέλο παρουσιάζεται στο γράφημα (1.5) [24].



Σχήμα 1.5: Χρονοσειρά τυχαίου περιπάτου

Κεφάλαιο 2

Γραμμική Παλινδρόμηση και Συσχέτιση

Μετά την ανάπτυξη των βασικών χαρακτηριστικών της χρονοσειράς θα μελετήσουμε τη συσχέτιση δύο μεταβλητών X και Y και θα εκτιμήσουμε τη συνάρτηση που δίνει τη Y γνωρίζοντας τη X. Επιπλέον θα μελετήσουμε εκτενέστερα την δυνατότητα εκτίμησης της μεταβλητής Y όταν εξαρτάται από περισσότερες από μία μεταβλητές.

2.1 Συσχέτιση δύο τιμών

Συχνά στη μελέτη ενός τεχνικού συστήματος ή φυσικού φαινομένου ενδιαφερόμαστε να προσδιορίσουμε τη σχέση μεταξύ δύο μεταβλητών του συστήματος. Για την μελέτη του φαινομένου της συσχέτισης υποθέτουμε πως έχουμε δύο τυχαίες μεταβλητές X, Y. Οι τυχαίες αυτές τιμές μπορούν να συσχετίζονται μεταξύ τους αφού η μία μπορεί να επηρεάζει την άλλη ή είναι εφικτή η εξάρτηση των δύο μεταβλητών από ένα τρίτο μέγεθος. Έτσι για παράδειγμα δύο

μεταβλητές οι οποίες επηρεάζουν η μία την άλλη είναι η τιμή ενός προϊόντος και η ζήτηση αυτού, αφού όσο μεγαλύτερη είναι η ζήτηση τόσο μεγαλύτερη θα είναι και η τιμή [25]. Επομένως η συσχέτιση δεν υποδηλώνει κάποια αιτιατική σχέση μεταξύ των δύο μεταβλητών (ρ=Corr(X, Y)). Ο συντελεστής που μας βοηθάει να εξακριβώσουμε το φαινόμενο αυτό μεταξύ των μεγεθών X, Y είναι ο συντελεστής αυτοσυσχέτισης που εκφράζει την γραμμική συσχέτιση μεταξύ δύο τιμών. Η παράμετρος αυτή χαρακτηρίζεται ως συντελεστής Pearson.

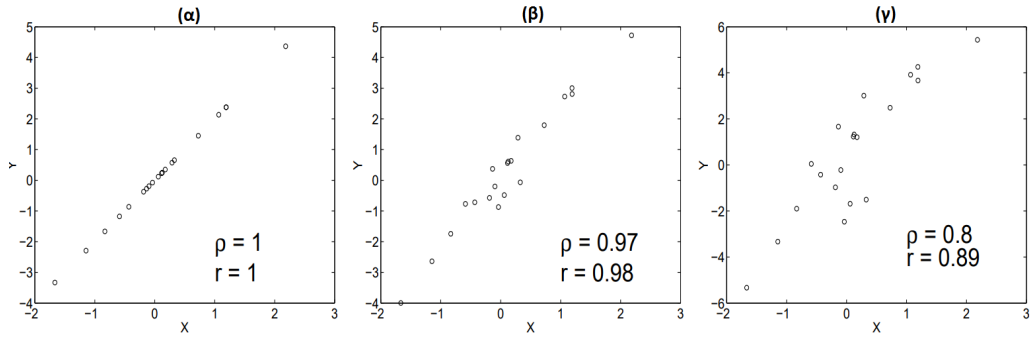
Δειγματικός συντελεστής συσχέτισης

Όταν οι παρατηρήσεις που διαθέτουμε προς επεξεργασία είναι κατανεμημένες κατά ζεύγη

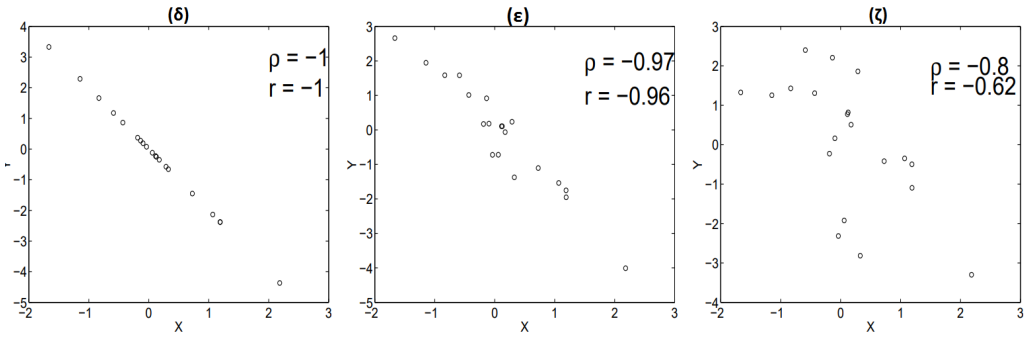
$$[(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)] \tag{2.1}$$

τότε μπορούμε να κάνουμε μια πρώτη εκτίμηση για την συσχέτιση των τιμών από το διάγραμμα διασποράς που απεικονίζει τα σημεία $(x_i, y_i, i = 1, ..., n)$ σε καρτεσιανό σύστημα συντεταγμένων.

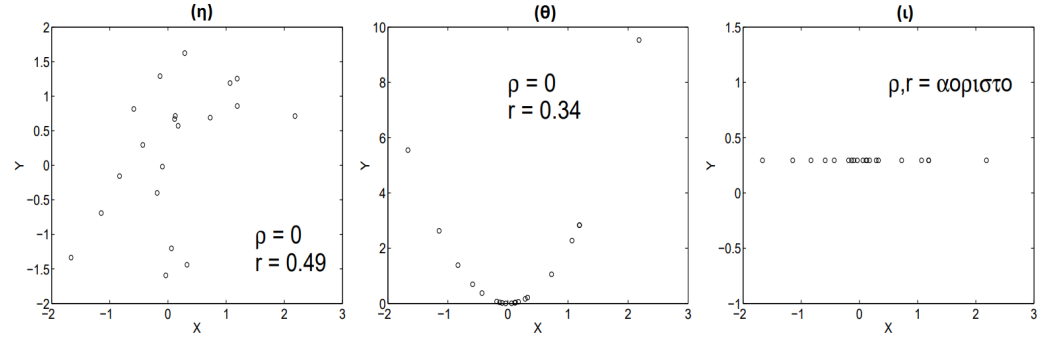
Στα παρακάτω σχήματα (Σχ. 2.1, Σχ. 2.2, Σχ. 2.3) απεικονίζονται τυπικά διαγράμματα διασποράς για ισχυρές και ασθενής συσχετίσεις μεταξύ δύο τιμών X και Y.



Σχήμα 2.1: Διάγραμμα διασποράς δύο τιμών X και Y με $n = 20$ παρατηρήσεις με θετική σχέση μεταξύ των τιμών. [25]



Σχήμα 2.2: Διάγραμμα διασποράς δύο τιμών X και Yμε $n = 20$ παρατηρήσεις με αρνητική σχέση μεταξύ των τιμών [25]



Σχήμα 2.3: Διάγραμμα διασποράς δύο τιμών X και Y με $n = 20$ παρατηρήσεις χωρίς καμία συσχέτιση μεταξύ των τιμών [25]

όπου συμπεραίνουμε πως στο γράφημα 2.1α και στο 2.2δ η αντισυσχέτιση μεταξύ των τιμών είναι εξαιρετική ($\rho = 1$ και $\rho = -1$). Στα διαγράμματα 2.1β και 2.2ε υπάρχει ισχυρή συσχέτιση (θετική με $\rho = 0.97$ και αρνητική με $\rho = -0.97$). Στα σχήματα 2.1γ και 2.2ζ η συσχέτιση είναι λιγότερο ισχυρή (θετική με $\rho = 0.8$ και αρνητική με $\rho = -0.8$ αντίστοιχα) ενώ στο 2.3η είναι $\rho = 0$ γιατί οι τιμές X και Y είναι ανεξάρτητες. Στο σχήμα 2.3θ είναι πάλι $\rho = 0$ αλλά οι X και Y δεν είναι ανεξάρτητες αλλά συσχετίζονται μη-γραμμικά. Τέλος στο γράφημα 2.3ι ο συντελεστής συσχέτισης δεν ορίζεται γιατί η Y είναι

σταθερή [23].

Το r λέγεται δειγματικός συντελεστής συσχέτισης και προσδιορίζεται μαθηματικά από την σχέση (Εξ. 2.2)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.2}$$

όπου το n εκφράζει το μέγεθος του δείγματος που διαθέτουμε, τα x_i, y_i είναι τιμές του σύνολο του δείγματος και το $\bar{x}, \bar{y}$ είναι η μέση τιμή του δείγματος για τις x, y παρατηρήσεις αντίστοιχα.

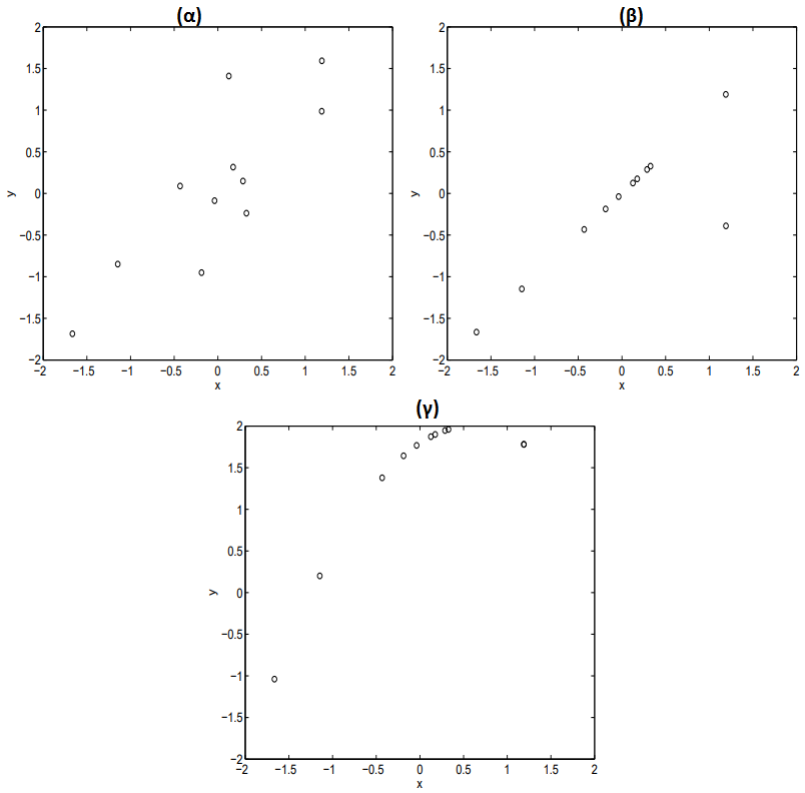
Καλύτερη φυσική ερμηνεία της συσχέτισης δύο μεγεθών επιτυγχάνεται με το r^2 που λέγεται συντελεστής προσδιορισμού, ο οποίος δίνει το ποσοστό μεταβολής των τιμών της Y που υπολογίζεται από τη X (και αντίστροφα) και είναι ένας χρήσιμος τρόπος να συνοψίσουμε τη σχέση δύο τιμών [25].

2.1.1 Συσχέτιση και γραμμικότητα

Ο συντελεστής ρ απευθύνεται στο μέγεθος της γραμμικής συσχέτισης μεταξύ δύο μεταβλητών Χ και Υ. Όμως για κάποιο δείγμα του ζεύγους τιμών των Χ και Υ η τιμή του r μπορεί να μην αποδίδει πάντα σωστά τη συσχέτιση, καθώς αυτή μπορεί να μην είναι γραμμική.

Στο σχήμα που ακολουθεί (Σχ. 2.4) δίνονται κάποια διαγράμματα διασποράς για τρία τυχαία δείγματα του ζεύγους (Χ, Υ). Και στα τρία δείγματα ο συντελεστής r είναι ίδιος και ίσος με 0.84. Συγκεκριμένα για το πρώτο δείγμα το ζεύγος είναι από κανονική κατανομή και άρα το r αποδίδει σωστά το μέγεθος της συσχέτισης τους. Αντίθετα στο δεύτερο και τρίτο δείγμα δεν αντιστοιχεί σε γραμμικά συσχετισμένα Χ και Υ. Πίο διεξοδικά στο δεύτερο δείγμα η σχέση των μεταβλητών είναι γραμμική και άρα (ρ=1) για όλα τα ζευγάρια τιμών εκτός από το απομακρυσμένο σημείο. Στο τρίτο η συσχέτιση των Χ και

Υ είναι απόλυτη αλλά μη-γραμμική με αποτέλεσμα να έχουμε υποεκτίμηση της αυτοσυσχέτισης [23].



Σχήμα 2.4: Διάγραμμα διασποράς δειγμάτων που δίνουν όλα τον ίδιο δειγματικό συντελεστή συσχέτισης $r = 0.84$ [34]

2.2 Ανάλυση παλινδρόμησης

Στην ανάλυση παλινδρόμησης εξετάζουμε την σχέση μεταξύ δυο ή περισσοτέρων ποσοτήτων με πιθανή αλληλεξάρτηση μεταξύ τους. Σε κάθε πρόβλημα παλινδρόμησης διακρίνουμε δύο είδη μεταβλητών: τις εξαρτημένες και τις ανεξάρτητες, όπου ως εξαρτημένες θεωρούνται οι μεταβλητές που μπορούμε να

ελέγξουμε και ανεξάρτητες αυτές που αντανακλάται το αποτέλεσμα των μεταβολών της στις εξαρτημένες μεταβλητές [24]. Το πρόβλημα της παλινδρόμησης αναφέρεται στην προσαρμογή ενός συνόλου πειραματικών δεδομένων (x(i), y(i)) σε ένα μαθηματικό πρότυπο $(y(x) = y_x; a_1, ..., a_k)$. Αν έχουμε δυο μεταβλητές X και Y για τις οποίες για κάθε τιμή της X μπορούμε να προβλέψουμε ακριβώς τις τιμή της Y τότε οι μεταβλητές αυτές συνδέονται με την συναρτησιακή σχέση $Y = f(X)$ [33]. Αν τα δεδομένα περιέχουν τυχαίες διακυμάνσεις μεταξύ τους τότε χαρακτηρίζονται από στοχαστικές σχέσεις και η γραμμική εξάρτηση επηρεασμένη από τον θόρυβο πλέον δίνεται από την εξίσωση (Εξ. 2.3).

$$y(i) = ax(i) + b + n(i) \tag{2.3}$$

όπου τα a, b είναι τυχαίες μεταβλητές και το n εκφράζει το μέγεθος της χρονοσειράς. Για να προσδιορίσουμε την στοχαστική εξάρτηση δυο μεταβλητών X και Y προσπαθούμε να βρούμε μια προσεγγιστική σχέση μεταξύ των μεταβλητών αυτών. Η συνηθέστερη μέθοδος η όποια χρησιμοποιείται για την περιγραφή της στοχαστικής εξάρτησης είναι η μέθοδος των ελαχίστων τετραγώνων. Ένα χαρακτηριστικό παράδειγμα για την μελέτη της συσχετίσεις που μπορεί να έχουν δύο μεταβλητές είναι η διάρκεια ζωής των ζωντανών οργανισμών σε μια περιοχή και τα επίπεδα μόλυνσης στην περιοχή αυτή. Οι μεταβλητές αυτές έχουν αρνητική εξάρτηση μεταξύ τους, αφού όσο μεγαλύτερη είναι η μόλυνση τόσο μικρότερη είναι η διάρκεια ζωής των οργανισμών [15]. Είναι επομένως εμφανής η σημασία της παλινδρόμησης σε επιστημονικούς και άλλους τομείς. Έτσι γίνεται διάκρισή του φαινομένου σε δύο βασικές κατηγορίες, την απλή γραμμική παλινδρόμηση και την πολλαπλή παλινδρόμηση.

2.2.1 Απλή γραμμική παλινδρόμηση

Ο πιο ενδεδειγμένος τρόπος διάκρισης του τύπου παλινδρόμησης είναι με την δημιουργία ενός διαγράμματος διασποράς μεταξύ των μεταβλητών ενδιαφέρο-

ντος. Εάν το διάγραμμα αυτό έχει μορφή επιμήκους κεκλιμένης έλλειψης, τότε η σχέση μεταξύ των μεταβλητών είναι πιθανότατα γραμμική και έχουμε πλέον να αντιμετωπίσουμε μια απλή γραμμική παλινδρόμηση όπου υπάρχει μόνο μια ανεξάρτητη μεταβλητή X και η εξαρτημένη μεταβλητή Y μπορεί να προσεγγισθεί ικανοποιητικά από μια γραμμική συνάρτηση του X. Στην απλή γραμμική παλινδρόμηση η συσχέτιση μεταξύ των μεταβλητών X και Y είναι συμμετρική αφού η εξαρτημένη μεταβλητή καθοδηγείται από την ανεξάρτητη. Η γραμμική σχέση $Y = \alpha + \beta X$ δε μπορεί, ασφαλώς, να περιγράψει τη γραμμική στοχαστική εξάρτηση των μεταβλητών X και Υ αφού αν, για παράδειγμα X είναι η τιμή ενός προϊόντος και Υ είναι η ζήτηση του προϊόντος αυτού, και διατηρήσουμε την τιμή X στο ίδιο επίπεδο X= x_1 τότε οι αντίστοιχες τιμές του Υ θα είναι διαφορετικές στις διάφορες επαναλήψεις [24]. Επομένως στην εξίσωση $Y = \alpha + \beta X$ πρέπει να προσθέσει και ο ορός του σφάλματος παλινδρόμησης ϵ_i για τον κατά υπόθεση προσδιορισμό πλέον της μεταβλητής Υ. Έτσι προκύπτει ένας ακόμα τρόπος απόδοσης της απλής γραμμικής παλινδρόμησης με την βοήθεια της σχέσης (Εξ. 2.4).

$$y = \beta_0 + \beta_1 x + \epsilon_i \tag{2.4}$$

όπου η β_0 είναι η τιμή του y όταν το $x = 0$ και λέγεται διαφορά ύψους και ο συντελεστής β_1 είναι η κλίση της ευθείας ή αλλιώς ο συντελεστής παλινδρόμησης.

Για την ανάλυση της γραμμικής παλινδρόμησης ακολουθούμε την εξής διαδικασία [25]

- Η μεταβλητή X πρέπει να ελεγχθεί για το πρόβλημα που μελετάμε, δηλαδή να γνωρίζουμε τις τιμές της χωρίς κανένα διάστημα αμφιβολίας.
- Πρέπει η σχέση $Y = \alpha + \beta X$ να ισχύει, άρα η εξάρτηση της Y από τη X να είναι γραμμική.

- Πρέπει να ισχύει, $E(\epsilon_i) = 0$ και $\mathrm{Var}(\epsilon_i) = \sigma^2$ για κάθε τιμή x_i της X, δηλαδή το σφάλμα παλινδρόμησης να έχει μέση τιμή μηδέν για κάθε τιμή της X και η διασπορά του είναι σταθερή και να μην εξαρτάται από την X.

2.2.2 Σημειακή εκτίμηση των παραμέτρων στην απλή γραμμική παλινδρόμηση

Για την επίλυση του προβλήματος της απλής γραμμικής παλινδρόμησης είναι απαραίτητη η εκτίμηση τριών βασικών παραμέτρων. Αυτές είναι:

- Η διαφορά ύψους της ευθείας παλινδρόμησης β_0
- Η κλίση της ευθείας παλινδρόμησης β_1
- Η διασπορά σφάλματος της παλινδρόμησης σ_e^2

Τα β_0 και β_1 προσδιορίζουν την ευθεία παλινδρόμησης και άρα καθορίζουν τη γραμμική σχέση εξάρτησης της τιμής Y από τη μεταβλητή X. Η παράμετρος σ_ϵ^2 προσδιορίζει τον βαθμό μεταβλητότητας γύρω από την ευθεία παλινδρόμησης και εκφράζει την αβεβαιότητα της γραμμικής σχέσης [24].

Εκτίμηση των παραμέτρων της ευθείας παλινδρόμησης

Η εκτίμηση των παραμέτρων β_0 και β_1 γίνεται με την μέθοδο των ελαχίστων τετραγώνων. Η μέθοδος λέγεται έτσι γιατί βρίσκει την ευθεία παλινδρόμησης με παραμέτρους b_0 και b_1. Οι εκτιμήσεις των β_0 και β_1 δίνονται από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων [33]

$$\min \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1)^2 \tag{2.5}$$

Για να γίνει επίλυση του προβλήματος αυτού χρειάζεται να θέσουμε τις μερικές παραγώγους των β_0 και β_1 ίσες με το μηδέν. Έτσι προκύπτει η εκτίμηση της κλίσης από την σχέση (Εξ. 2.7).

$$\beta_1 = b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.6}$$

Στην σχέση αυτή με αντικατάσταση του b_1 στην πρώτη εξίσωση του παραπάνω συστήματος παίρνουμε την εκτίμηση του σταθερού όρου b_0

$$\beta_0 = b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} \tag{2.7}$$

όπου το n εκφράζει το μέγεθος της χρονοσειράς, τα x_i, y_i είναι τιμές του συνόλου της χρονοσειράς και τα $\bar{x}, \bar{y}$ εκφράζουν τους μέσους όρους των τιμών αυτών. Τα b_0 και b_1 ορίζουν την ευθεία ελαχίστων τετραγώνων

$$y = b_0 + b_1 x \tag{2.8}$$

2.2.3 Διάστημα εμπιστοσύνης των παραμέτρων της απλής γραμμικής παλινδρόμησης

Όπως η μέση τιμή $\bar{x}$ και η τυπική απόκλιση s μπορούν να διαφέρουν από δείγμα σε δείγμα παρατηρήσεων μίας τιμής X, έτσι και η κλίση b_1 και η διαφορά ύψους b_0 μπορούν να διαφέρουν επίσης από δείγμα σε δείγμα ενός ζεύγους παρατηρήσεων (X, Υ). Για να μπορέσουμε να προσδιορίσουμε τα διαστήματα εμπιστοσύνης για τις παραμέτρους β_0 και β_1 θα μελετήσουμε την κατανομή των b_1 και b_0 αντίστοιχα [15].

Από την σχέση (Εξ. 2.5) προκύπτει πως ο εκτιμητής b_1 της κλίσης της ευθείας δίνεται ως γραμμικός συνδυασμός των $(y_1, ..., y_n)$. Επιπρόσθετα το b_1 έχει μέση τιμή

$$\mu_{b1} = E(b_1) = \beta_1 \tag{2.9}$$

και διασπορά

$$\sigma_{b1}^2 = Var(b_1) = \frac{\sigma_\epsilon^2}{S_{xx}} \tag{2.10}$$

όπου το S_{xx} είναι η διασπορά των τιμών του X. Η τυπική απόκλιση μπορεί να αποδοθεί και με την βοήθεια της σχέσης $\sigma_{b1} = \frac{\sigma_\epsilon}{\sqrt{S_{xx}}}$ και η διασπορά των σφαλμάτων μπορεί να εκτιμηθεί με την βοήθεια της ευθείας των ελαχίστων τετραγώνων ως s_ϵ^2 [24]. Επομένως η εκτίμηση της τυπικής απόκλισης θα γίνεται πλέον με την βοήθεια του τύπου (Εξ. 2.11).

$$s_{b1} = \frac{s_\epsilon}{\sqrt{S_{xx}}} \tag{2.11}$$

Τέλος θεωρώντας πως η Υ ακολουθεί την κανονική κατανομή και ο εκτιμητής της κλίσης b_1 ακολουθεί επίσης την κανονική κατανομή, το διάστημα εμπιστοσύνης της κλίσης β_1 δίνεται από την σχέση (Εξ. 2.12).

$$b_1 \pm t_{n-2,1-a/2}\frac{s_\epsilon}{\sqrt{S_{xx}}} \tag{2.12}$$

2.2.4 Πολλαπλή γραμμική παλινδρόμηση

Στην προηγούμενη ενότητα κάναμε μια αναφορά σχετικά με την γραμμική εξάρτηση μίας μεταβλητής Y από μια μόνο μεταβλητή X. Στην περίπτωση όμως που η ίδια μεταβλητή Υ εξαρτάται γραμμικά από μια ακόμα μεταβλητή X_2 ή και μία ακόμα X_3 και στο τέλος από ένα σύνολο n μεταβλητών τότε η σχέση της γραμμικής παλινδρόμησης διαμορφώνεται ως εξής

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \epsilon_i \tag{2.13}$$

οι συντελεστές $\beta_1, \beta_2, \beta_m$ λέγονται μερικοί συντελεστές παλινδρόμησης [11]. Έτσι προκύπτει πως το β_1 εκφράζει το μέγεθος της μεταβολής του Υ όταν υπάρχει αλλαγή στην τιμή της X_1 μεταβλητής κατά μια μονάδα αλλά οι υπόλοιπες μεταβλητές X_i παραμένουν σταθερές στην τιμή του μέσου όρου τους. Πολλές φορές αντί των μερικών συντελεστών παλινδρόμησης χρησιμοποιούνται και οι

τυποποιημένοι μερικοί συντελεστές β_i ή βήτα συντελεστές, οι οποίοι προέκυψαν από τον υπολογισμό της εξίσωσης της παλινδρόμησης, αφού προηγουμένως τροποποιήθηκαν όλες οι συμμετέχουσες μεταβλητές X_i και Υ με την αφαίρεση του μέσου όρου από όλες τις τιμές και διαιρώντας τη διαφορά με την τυπική απόκλιση των μεγεθών αυτών [11]. Το μεγαλύτερο πλεονέκτημα των συντελεστών αυτών έναντι των μερικών συντελεστών παλινδρόμησης είναι ότι τα μεγέθη τους μπορούν να συγκριθούν άμεσα μεταξύ τους, ως προς το μέγεθος της μεταβολής των ανεξάρτητων μεταβολών επί της ίδιας Υ. Εξαλείφεται δηλαδή το πρόβλημα της διαφορετικής κλίμακας των μετρήσεων που ελήφθησαν για κάθε τιμή του X.

2.2.5 Εκτίμηση μοντέλου πολλαπλής γραμμικής παλινδρόμησης

Όπως έγινε γνωστό κάθε μοντέλο πολλαπλής γραμμικής παλινδρόμησης μπορεί να περιγραφεί με την βοήθεια της σχέσης (Εξ. 2.12), όπου οι μεταβλητές $x_1, x_2, ..., x_m$ μπορεί να αντιπροσωπεύουν διαφορετικά μεγέθη.

Η εκτίμηση των παραμέτρων του μοντέλου της σχέσης (Εξ. 2.12) γίνεται με την μέθοδο των ελαχίστων τετραγώνων. Το άθροισμα των τετραγώνων των σφαλμάτων είναι

$$f(\beta_0, \beta_1, ..., \beta_m) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x1 + \beta_2 x_2 + ... + \beta_m x_m))^2 \qquad (2.14)$$

Το σύστημα των εξισώσεων το οποίο θα μας βοηθήσει στην εκτίμηση των παραμέτρων αυτών είναι μέσω των μερικών παραγώγων της συνάρτησης αυτής ως προς κάθε παράμετρο $(\beta_0, \beta_1, ..., \beta_m)$ και συγκεκριμένα

$$b_0 n + b_1 \sum x_{1i} + b_2 \sum x_{2i} + ... + b_m \sum x_{mi} = \sum y_i$$

$$b_0 \sum x_{1i} + b_1 \sum x_{1i}^2 + b_2 \sum x_{1i} x_{2i} + ... + b_m \sum x_{1i} x_{mi} = \sum x_{1i} y_i$$

και έτσι προκύπτει

$$b_0 \sum x_{mi} + b_1 \sum x_{1i} x_{mi} + b_2 \sum x_{mi} x_{2i} + ... + b_m \sum x_{mi}^2 = \sum x_{mi} y_i \quad (2.15)$$

από το οποίο προκύπτουν οι εκτιμήσεις b_0, b_1, b_m

Η εκτίμηση της εξαρτημένης μεταβλητής μέσω της πολλαπλής παλινδρόμησης που προσδιορίστηκε με τη μέθοδο ελαχίστων τετραγώνων δίνεται μέσω της εξίσωσης (Εξ. 2.16).

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_m x_{mi} \quad (2.16)$$

και τα σφάλματα του μοντέλου είναι $\epsilon_i = y_i - \bar{y}_i$.

2.3 Θεωρία ελαχίστων τετραγώνων

Όταν παρουσιάζεται ένα πρόβλημα παλινδρόμησης τότε η πιο συνηθισμένη μέθοδος εκτίμησης αποτελεί η μέθοδος ελαχίστων τετραγώνων method of ordinary least squares, OLS). Η πρώτη πιθανότητα φορά που εφαρμόστηκε η μέθοδος αυτή ήταν το 1805 σε μια εργασία του Γάλλου μαθηματικού Legendre αν και μάλλον ο Gauss ήταν αυτός που χρησιμοποίησε πρώτος την μέθοδο [24].

Αν η χρονοσειρά αποτελεί μια στοχαστική διαδικασία, δηλαδή είναι ένα σύστημα που οδηγείται από θόρυβο, τότε είναι εφικτή η εξέταση της εξάρτησης μεταξύ δύο μεταβλητών μέσω της προσαρμογής γραμμικών μοντέλων στην χρονοσειρά. Όταν γίνει εφικτός ο προσδιορισμός ενός πρότυπου εξάρτησης τότε η χρονοσειρά μπορεί να χρησιμοποιηθεί για πρόβλεψη. Η μέθοδος αυτή καλείται έτσι γιατί βρίσκει την ευθεία παλινδρόμησης ώστε το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων από την ευθεία να είναι το μικρότερο δυνατό. Έτσι αν έχουμε n_i σφάλματα, τα οποία αποτελούν τυχαίες μεταβλητές αλλά ακολουθούν την κανονική κατανομή με διασπορά σ^2 τότε το τυπικό μέγεθος των

σφαλμάτων προσδιορίζεται από το σ [24]. Πρακτικά η μέθοδος των ελαχίστων τετραγώνων επιδιώκει την ελαχιστοποίηση του αθροίσματος των τετραγωνικών σφαλμάτων. Επομένως ισχύει η σχέση (Εξ. 2.17).

$$\sum_{i=1}^{M} n_i^2 = \sum_{i=1}^{M} (y_i - ax_i - b)^2 \tag{2.17}$$

όπου τα σφάλματα n_i είναι τυχαίες μεταβλητές που ακολουθούν την κανονική κατανομή με μέση τιμή 0 και διασπορά σ^2. Σε περίπτωση που τα σφάλματα που μας ενδιαφέρουν διακρίνονται από στατιστική ανεξαρτησία και ακολουθούν κανονική κατανομή, τότε η μέθοδος των ελαχίστων τετραγώνων προκύπτει από την μέθοδο της μέγιστης πιθανοφάνειας [19]. Συγκεκριμένα, η πιθανότητα εμφάνισης του δείγματος δίνεται από την εξίσωση (Εξ. 2.18).

$$P = \prod_{i=1}^{M} \exp^{-(y_i - ab_i - b)^2 / 2\sigma^2} \tag{2.18}$$

Το M εκφράζει το μέγεθος της χρονοσειράς και τα a, b είναι τυχαίες μεταβλητές. Για γίνει εφικτή η ελαχιστοποίηση των τετραγωνικών σφαλμάτων ως προς τις μεταβλητές a, b πρέπει να γίνει μηδενισμός των μερικών παραγώγων, που σημαίνει ότι πρέπει να λυθεί η σχέση $f(x) = (y_i - ax_i - b)^2$ μέσω των εξισώσεων (Εξ. 2.19) και (Εξ. 2.20).

$$\frac{\partial f}{\partial a} = 0 \tag{2.19}$$

$$\frac{\partial f}{\partial b} = 0 \tag{2.20}$$

Με την επίλυση του συστήματος ελαχιστοποίησης των σφαλμάτων προκύπτουν οι εξισώσεις των τιμών των γραμμικών παραμέτρων

$$\alpha = r_{xy} \frac{\sigma_y}{\sigma_x} \tag{2.21}$$

$$b = \overline{y} - a\overline{x} \tag{2.22}$$

όπου το r_{xy} είναι ο δειγματικός συντελεστής συσχέτισης, ($\bar{x}$, $\bar{y}$) είναι οι δειγματικές μέσες τιμές και σ_x^2, σ_y^2 οι δειγματικές διασπορές.

2.4 Μέτρα επιβεβαίωσης και αβεβαιότητα εκτίμησης

2.4.1 Μέτρα επιβεβαίωσης για την εκτίμηση χρονοσειρών

Τα μέτρα επιβεβαίωσης αποτελούν ένα εξαιρετικά ικανό μέσο αξιολόγησης των αποτελεσμάτων που έχουν προκύψει από κάποια εκτίμηση ή κάποια πρόβλεψη που έχει πραγματοποιηθεί. Μας βοηθάνε έτσι ώστε να μπορέσουμε να κρίνουμε εάν και εφόσον η εκτίμηση που έχει επιτευχθεί έχει δώσει αξιόπιστα ή όχι αποτελέσματα. Τρία από τα βασικά μέτρα επιβεβαίωσης που χρησιμοποιούνται ευρέος είναι τα RMSE (Root mean square error), ME (mean error) και το RPe (Pearson's correlation coefficient) [24]. Τα μέτρα αυτά δεν είναι το μοναδικό μέσο αξιολόγησης των αποτελεσμάτων αφού υπάρχουν περίπτωσης όπου μπορεί να φαίνονται ικανοποιητικά και να κρίνουμε ότι το αποτέλεσμα είναι αξιόπιστο αλλά στην πραγματικότητα κάτι τέτοιο μπορεί να μην ισχύει. Για τον λόγο αυτό η σωστή αξιολόγηση των γραφικών παραστάσεων που προκύπτουν καθώς και η εμπειρία στην στατιστική ανάλυση παίζουν καθοριστικό ρόλο.

Το RMSE αποτελεί την τυπική απόκλιση των υπολοίπων. Τα residuals ή υπόλοιπα μεταξύ των διαθέσιμων δεδομένων και της τάσης είναι ένα μέσο μέτρησης που μας δείχνει πόσο έχουν εξαπλωθεί τα δεδομένα. Με αλλά λόγια μας δίνει μία εικόνα για το πόσο συγκεντρωμένα είναι τα στοιχεία που διαθέτουμε γύρω από την γραμμή που μας δείχνει την καλύτερη προσαρμογή των δεδομένων μας [3]. Το root mean square error χρησιμοποιείται κυρίως στην κλιματολογία, για πρόβλεψη και για αναλύσεις αυτοσυσχέτισης [5]. Μπορεί να πάρει είτε αρνητική είτε θετική τιμή. Ο τρόπος προσδιορισμού του συγκεκριμένου συντελεστή

δίνεται από την σχέση (Εξ. 2.23).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{n=1}^{n}[\hat{x_i} - x_i]^2} \tag{2.23}$$

όπου το x_i είναι οι τιμές για κάθε παρατήρηση και το $\hat{x_i}$ είναι οι τιμές που θα προβλεφθούν. Το RMSE υπολογίζει την ακρίβεια και την ορθότητα της εκτίμησης. Ένας άλλος τρόπος εκτίμησης του τετραγωνικού σφάλματος είναι ο εξής

- Αρχικά τετραγωνίζοντας τα υπόλοιπα
- Στην συνέχεια βρίσκοντας τον μέσο όρο των υπολοίπων
- Και τέλος παίρνοντας την τετραγωνική ρίζα του αποτελέσματος.

Τα τρία βήματα που ανεφέρθηκαν συνοψίζονται στην σχέση (Εξ. 2.24)

$$\mathrm{RMSE} = \sqrt{1 - r^2}\sigma_y \tag{2.24}$$

όπου το σ_y είναι η τυπική απόκλιση του y. Τέλος το εύρος τιμών του RMSE είναι πάντα μεταξύ του 0 και 1 ενώ του r είναι μεταξύ του 1 και -1 [24].

Το Mean error αναφέρεται στον μέσο όρο όλων των σφαλμάτων που υπάρχουν στα δεδομένα. Ως σφάλμα εκφράζεται μια αβεβαιότητα στην μέτρηση ή η διαφορά μεταξύ της μετρήσιμης τιμής και της προβλεπόμενης τιμής. Το μέσο σφάλμα προσδιορίζεται μέσω της σχέσης (Εξ. 2.25).

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}[\hat{x_i} - x] \tag{2.25}$$

Το μέσο σφάλμα υπολογίζει τη μεροληψία της εκτίμησης. Μεγάλες και θετικές ή αρνητικές τιμές του μέσου σφάλματος δηλώνουν την ύπαρξη αμεροληψίας

(bias). Η ύπαρξη αμεροληψίας (bias) είναι ένα χαρακτηριστικό της στατιστικής ανάλυσης. Η αμεροληψία ενός εκτιμητή είναι η διαφορά μεταξύ της τιμής που προσδοκάται να λάβει το προς εκτίμηση μέγεθος και της πραγματικής τιμής που εν τέλει θα προσδιοριστεί. Πρακτικά η αμεροληψία είναι η τάση που εμφανίζει ένα δείγμα τιμών για υπό-εκτίμηση ή υπέρ-εκτίμηση της τιμής σε ένα σύνολο δεδομένων. Δεν είναι λίγες όμως οι περιπτώσεις που το μέσο σφάλμα (ME) δίνει αποτελέσματα τα όποια δεν είναι πάντα ιδιαίτερα χρήσιμα γιατί οι θετικές και οι αρνητικές τιμές ακυρώνουν η μία την άλλη. Για την αντιμετώπιση αυτού πρέπει να χρησιμοποιούμε το απόλυτο μέσο σφάλμα ή mean absolute error. Το σφάλμα αυτό χρησιμοποιεί τις απόλυτες τιμές των σφαλμάτων στους υπολογισμούς [27], έτσι προκύπτει ένας μέσος όρος με περισσότερο νόημα, ο όποιος δίνεται από την σχέση (Εξ. 2.26).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{x_i} - x| \tag{2.26}$$

.

Το μέσο απόλυτο σφάλμα υπολογίζει την ορθότητα και την ακρίβεια της εκτίμησης. Και ενώ το μέσο σφάλμα συνήθως αναφέρεται στο μέσο απόλυτο σφάλμα πολλές φορές μπορεί να αναφερθεί ως

α) Μέση απόλυτη απόκλιση που μετράει την μέση τιμή της τυπικής απόκλισης, η οποία είναι η διάδοση των τιμών γύρω από το κέντρο του συνόλου των δεδομένων. Οι δύο όροι ακούγονται παρόμοιοι μόνο που η τυπική απόκλιση αναφέρεται στην μονάδα εξάπλωσης ενώ το σφάλμα στην διαφορά της μονάδας μέτρησης.

β)Μέσο τετραγωνικό σφάλμα, το οποίο χρησιμοποιείται στην ανάλυση αυτοσυσχέτισης για να δείξει πόσο κοντά είναι μια γραμμή αυτοσυσχέτισης στο σύνολο των δεδομένων [27].

Τέλος το RPe (υπό-ενότητα 1.4.3) χρησιμοποιείται στην στατιστική προκει-

μένου να μετρήσουμε πόσο ισχυρή είναι η σχέση μεταξύ δύο μεταβλητών [23]. Υπάρχουν διάφοροι τύποι του συντελεστή συσχέτισης (correlation coefficient). Ο πιο διαδεδομένος είναι αυτός του Pearson ο οποίος χρησιμοποιείται κυρίως στην γραμμική παλινδρόμηση [27].

Ο τύπος που χρησιμοποιείται συνήθως για την εφαρμογή του συντελεστή συσχέτισης του Pearson είναι ο εξής

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \tag{2.27}$$

Ο συντελεστής συσχέτισης του Pearson δίνει την τιμή της γραμμικής σχέσης μεταξύ δύο μεταβλητών. Το βασικό πρόβλημα του RPe είναι ότι δεν μπορεί να ξεχωρίσει την διαφορά μεταξύ μίας εξαρτημένης και μίας ανεξάρτητης μεταβλητής. Έτσι συμπεραίνουμε πως πρέπει να προσέχουμε τα δεδομένα που εισαγάγουμε στην συντελεστή αφού αυτός δεν θα μας δώσει την κλίση της ευθείας αλλά μόνο την σχέση μεταξύ των δεδομένων.

2.4.2 Αβεβαιότητα της εκτίμησης

Η εκτίμηση της αβεβαιότητας σε οποιαδήποτε τέτοιου είδους έρευνα έχει καθοριστικό ρόλο, αφού ο αριθμός των παραμέτρων είναι πολύ μεγάλος καθιστώντας την εγκυρότητά των αποτελεσμάτων αβέβαιη. Πιο συγκεκριμένα σε μια ευθεία ελαχίστων τετραγώνων τα σημεία (x_i, y_i) δεν προσαρμόζονται πάντα πάνω στην ευθεία με επιθυμητό τρόπο, αφού υπάρχει αβεβαιότητα ως προς τον προσδιορισμό των τιμών a, b . Η αβεβαιότητα των μεγεθών αυτών μπορεί να εκφραστεί με την βοήθεια της διασποράς. Τα μέτρα αβεβαιότητας επομένως δίνονται από τις μαθηματικές σχέσεις (Εξ. 2.28), (Εξ. 2.29).

$$\sigma_a^2 = \frac{\sigma^2}{M\sigma_x^2} \tag{2.28}$$

$$\sigma_a^2 = \frac{\sigma^2 \overline{x}^2}{M\sigma_x^2} \tag{2.29}$$

όπου σ_x^2 είναι η δειγματική διασπορά, το $\bar{x}$ είναι η δειγματική μέση τιμή και το M εκφράζει το μέγεθος της χρονοσειράς. Ας σημειωθεί ότι η διασπορά των παραμέτρων εξαρτάται από την διασπορά του σφάλματος σ^2, ενώ οι τιμές των παραμέτρων είναι ανεξάρτητες από το σ^2. Αν το πρόσημο του συντελεστή συσχέτισης p_{ab} είναι αρνητικό, τα σφάλματα εκτίμησης των a, b έχουν αντίθετη συσχέτιση, δηλαδή είναι πιθανό να διαφέρουν ως προς το πρόσημο.

Κεφάλαιο 3

Μοντέλα Χρονοσειρών

3.1 Μοντέλα SARIMA για πρόβλεψη χρονοσειρών

Τα μοντέλα SARIMA αποτελούν μία οικογένεια μοντέλων που συμπεριλαμβάνουν τα μοντέλα AR, MA, ARMA, ARIMA όπως θα δούμε και στην συνέχεια.

Έχοντας μια χρονοσειρά X_t και εφαρμόζοντας σε αυτήν την μέθοδο των πρώτων διαφορών τότε προκύπτει μια στάσιμη χρονοσειρά καθώς έχουν απαλειφθεί οι τάσεις. Θα θεωρηθεί ότι η X_t είναι μια στοχαστική διαδικασία, το οποίο σημαίνει ότι το σύστημα οδηγείται σε θόρυβο. Υπάρχουν πολύ μέθοδοι για την δημιουργία μίας εκτίμησης ή πρόβλεψης όταν έχουμε να αντιμετωπίσουμε τέτοιου είδους χρονοσειρές. Κάποιες από τις κυριότερες διαδικασίες για την επιτέλεση τέτοιων αναλύσεων παρατίθενται στην συνέχεια.

3.2 Αυτοπαλινδρομούμενη διαδικασία (AR)

Η αυτοπαλινδομούμενη διαδικασία τάξης p, AR(p) [autoregressive process of order p] ορίζει μια μεταβλητή ως συνάρτηση κάποιον άλλων ανεξάρτητων μεταβλητών. Συγκεκριμένα στα γραμμικά μοντέλα παλινδρόμησης η συνάρτηση είναι γραμμική, που σημαίνει ότι η εξαρτημένη μεταβλητή είναι γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών. Το AR(p) μοντέλο θεωρεί ως εξαρτημένες τις τιμές της χρονοσειράς σε κάποια τυχαία χρονική στιγμή t και σαν ανεξάρτητες θεωρεί τις τυχαίες μεταβλητές της χρονοσειράς σε προηγούμενους χρόνους [24]. Ως τάξη της αυτοπαλινδρομούμενης διαδικασίας θεωρείται ο αριθμός των υστερήσεων που συμπεριλαμβάνουμε. Ως υστερήσεις χαρακτηρίζονται οι τελεστές της προς τα πίσω μετατόπισης κατά την δημιουργία προβλέψεων. Ένα αυτοπαλινδρομούμενο μοντέλο τάξης p ορίζεται από την σχέση (Εξ. 3.1).

$$x_t = \phi_0 + \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + Z_t \tag{3.1}$$

όπου τα ϕ_0, ϕ_1, ..., ϕ_p είναι οι συντελεστές του μοντέλου και το Z_t ανήκει στην κατηγορία των ανεξάρτητων και ισόνομων μεταβλητών με μέση τιμή 0 [8]. Προκύπτει πως το AR(p) μοντέλο είναι γνωστό μόνο όταν η διασπορά και οι συντελεστές του λευκού θορύβου είναι γνωστοί. Επομένως οι συντελεστές της αυτοπαλινδρομούμενης διαδικασίας, όπως και η διασπορά του λευκού θορύβου είναι μεγέθη που εκτιμούνται από την χρονοσειρά και η χρησιμοποίηση αυτών έγκειται κυρίως στην πρόβλεψη της χρονοσειράς τις επόμενες χρονικές στιγμές. Βάση του μοντέλου AR(p) ο γραμμικός συνδυασμός της τάξης του μοντέλου και των τελευταίων τιμών της χρονοσειράς $x_{t-1}, x_{t-2}, ..., x_{t-p}$ μας δίνει την δυνατότητα να εξηγήσουμε ένα μέρος της μεταβλητής της χρονοσειράς την χρονική στιγμή t. Το κομμάτι της χρονοσειράς που δεν αναλύεται είναι ένα καθαρά στοχαστικό κομμάτι και δημιουργείτε λόγο εξωγενών επιδράσεων [24].

Κάνοντας χρήση του τελεστή υστέρησης η AR(p) μπορεί να πάρει την συμπαγή μορφή

$$\phi(B)X_t = Z_t \tag{3.2}$$

$$\phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i \tag{3.3}$$

όπου το B εκφράζει τον τελεστή υστέρησης του πολυωνύμου και το φ(B) αποτελεί τον τελεστή της αυτοπαλινδρομούμενης διαδικασίας.

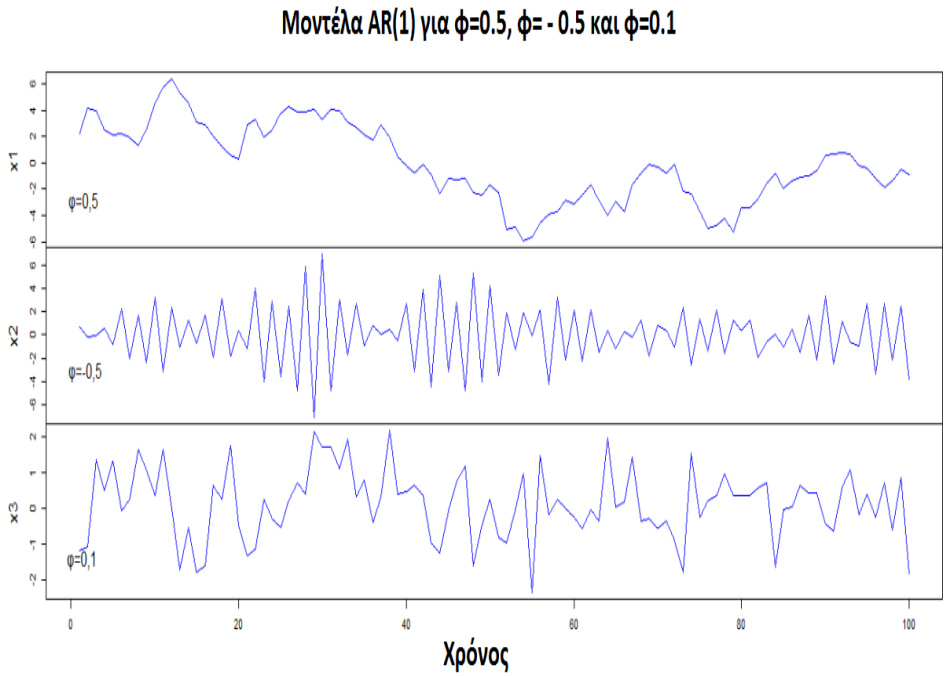
Το πολυώνυμο αυτό αποτελεί το χαρακτηριστικό πολυώνυμο της διαδικασίας. Τέλος μπορεί να θεωρηθεί πως η AR(p) μέθοδος είναι στάσιμη όταν οι ρίζες του χαρακτηριστικού πολυωνύμου είναι εκτός του μοναδιαίου κύκλου. Το πρώτο μέρος της αυτοπαλινδομούμενης διαδικασίας $X_{t-1}, ..., X_{t-p}$ ορίζεται ως καθοριστικό ή αιτιοκρατικό μοντέλο ενώ το δεύτερο μέρος Z_t ως το στοχαστικό [24].

3.2.1 Αυτοπαλινδρομούμενη διαδικασία πρώτης τάξης

Η διαδικασία αυτή είναι η πιο απλή διαδικασία παλινδρόμησης και συμβολίζεται ως AR(1),

$$x_t = \phi x_{t-1} + Z_t \tag{3.4}$$

με συνθήκη στασιμότητας $\phi < 1$. Σε περίπτωση που το $\phi \geq 1$ η διαδικασία είναι αυτή του τυχαίου περιπάτου [8]. Η γραφική απεικόνιση της αυτοπαλινδρομούμενης διαδικασίας πρώτης τάξης για διάφορες τυχαίες τιμές του ϕ (το ϕ είναι ο συντελεστής συσχέτισης) απεικονίζονται στο παρακάτω γράφημα (Σχ. 3.1).



Σχήμα 3.1: Αυτοπαλινδρομούμενα Μοντέλα πρώτης τάξης [4]

Επομένως με βάση τα παραπάνω μπορούμε να οδηγηθούμε στο συμπέρασμα ότι για την AR(1) διαδικασία ισχύουν τα εξής:

α) Εάν θεωρήσουμε ότι B είναι ο τελεστής της προς τα πίσω μετατόπισης και για τον οποίο ισχύει

$$B^k x_t = x_{t-k} \tag{3.5}$$

όπου το k αντιστοιχεί σε χρονικά βήματα. Τότε η εξίσωση (Εξ. 3.4) μπορεί να τροποποιηθεί ως

$$x_t = B\phi_1 x_t + Z_t = (1 - B\phi_1)x_t = Z_t = \sum_{i=0}^{\infty} \psi_i Z_{t-i} \tag{3.6}$$

και

$$\psi(B) = \sum_{i=0}^{\infty} B^i \phi_1^i < \infty \tag{3.7}$$

Για να μπορέσει να γίνει η σύγκλιση της παραπάνω χρονοσειράς πρέπει $|B| \leq 1$ ή αντίστοιχα θα πρέπει $\sum_{i=0}^{\infty} |\phi_1^i| < \infty$ και επομένως προκύπτει ότι $|\phi_1| < 1$ [24].

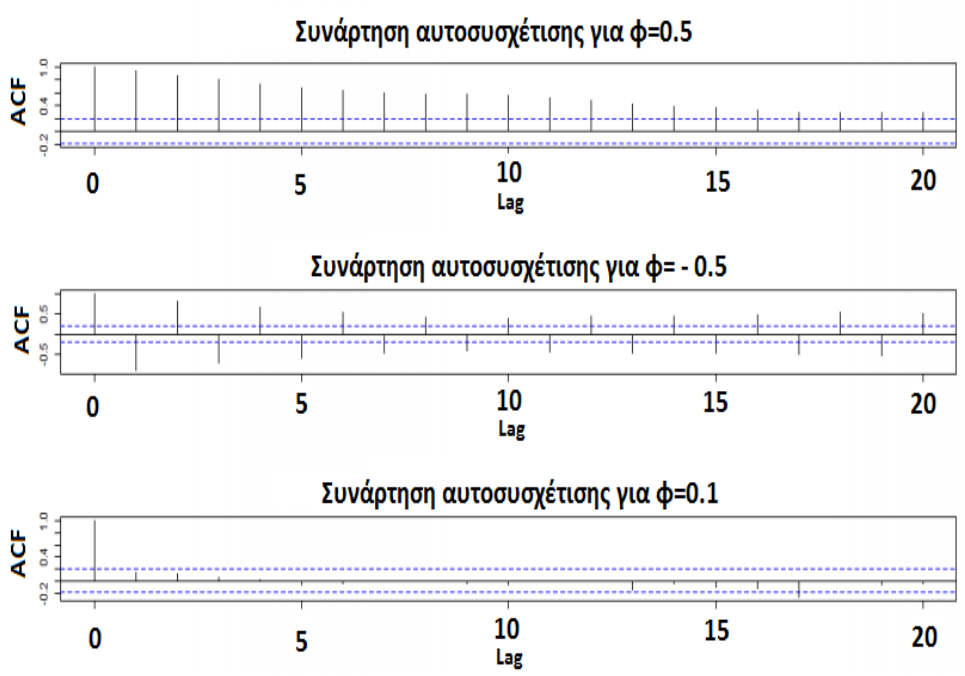
β) Τέλος πολύ σημαντικό ρόλο έχει η συνάρτηση αυτοσυσχέτισης. Εάν πολλαπλασιάσουμε την (Εξ. 3.4) για k χρονικά βήματα με x_{t-k}τότε προκύπτει ότι

$$x_{t-k}x_t = \phi_1 x_{t-k}x_{t-1} + x_{t-k}Z_t \tag{3.8}$$

για την όποια ξέρουμε ότι

$$E(x_{t-k}x_t) = E(\phi_1 x_{t-k}x_{t-1}) + E(x_{t-k}Z_t) = \phi_1\gamma(k-1) \tag{3.9}$$

Στα παρακάτω γραφήματα (Σχ. 3.2) αναγράφονται οι αυτοσυσχετίσεις των μοντέλων για τις τιμές του ϕ που χρησιμοποιήθηκε και προηγουμένως



Σχήμα 3.2: Συνάρτηση Αυτοσυσχέτισης για AR(1) [4]

3.2.2 Αυτοπαλινδομούμενη διαδικασία τάξης p

Η γενική μορφή του αυτοπαλινδρομούμενου μοντέλου τάξης p θεωρείται στάσιμη εάν και εφόσον οι ρίζες του χαρακτηριστικού πολυωνύμου

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p \tag{3.10}$$

είναι έκτος του μοναδιαίου κύκλου ή στην περίπτωση που οι ρίζες $\lambda_i, i = 1, 2, ..., p$ της εξίσωσης

$$\lambda^p - \phi_1 \lambda^{p-1} - ... - \phi_{p-1}\lambda - \phi_p = 0 \tag{3.11}$$

είναι μικρότερες της μονάδας.

Όταν μια χαρακτηριστική ρίζα είναι μεγαλύτερη της μονάδας τότε η αυτοπαλινδρομούμενη διαδικασία είναι ασταθής, ενώ η χρονοσειρά που δημιουργείται είναι εκρηκτικά μη-στάσιμη . Στην περίπτωση που η χαρακτηριστική ρίζα είναι ίση με την μονάδα τότε η χρονοσειρά μέσω μιας AR(p) διαδικασίας δεν είναι στάσιμη αλλά χαρακτηρίζεται ως μη-στάσιμη μοναδιαία ρίζα. Τέλος υπάρχει ακόμα μια περίπτωση χαρακτηριστικής ρίζας, αυτή της μικρότερης της μονάδας. Σε αυτήν την περίπτωση η πραγματική ρίζα υποδηλώνει εκθετική φθίνουσα αυτοσυσχέτιση, όπου αν είναι θετική γίνεται μονότονα, ενώ αν είναι αρνητική γίνεται εναλλασσόμενα ως προς το μηδέν. Μπορεί όμως να έχουμε δυο συζυγής μιγαδικές ρίζες οι οποίες δηλώνουν την ύπαρξη κύκλου στην διαδικασία [23].

3.2.3 Πρόβλεψη και προσαρμογή με AR μοντέλο

Πριν την οποιαδήποτε προσπάθεια πρόβλεψης μιας χρονοσειράς με ένα αυτοπαλινδρομούμενο μοντέλο θα πρέπει να γίνει η απαραίτητη διερεύνηση προκειμένου να γίνει η προσαρμογή του καταλληλότερου μοντέλου ανάλογα με την χρονοσειρά που διαθέτουμε. Έτσι η προσπάθεια προσαρμογής ενός αυτοπαλινδρομούμενου μοντέλου περιλαμβάνει την επιλογή της τάξης p του μοντέλου, την

εκτίμηση των παραμέτρων $\phi_1, ..., \phi_p$ του μοντέλου και τέλος το διαγνωστικό έλεγχο καταλληλότητας αυτού [24].

Στο πρώτο βήμα και στην επιλογή της τάξης p πολύ σημαντικό ρόλο έχει η χρησιμοποίηση δύο κριτηρίων. Το πρώτο αναφέρεται στην μερική αυτοσυσχέτιση (partial autocorrelation) . Η συνάρτηση μερικής αυτοσυσχέτισης όπως και αυτή της αυτοσυσχέτισης ορίζεται για κάθε υστέρηση τ την οποία θα την συμβολίσουμε εδώ ως p μίας και μελετάμε το AR μοντέλο. Εάν θεωρήσουμε τις μεταβλητές της χρονοσειράς X_t ως $x_t, x_{t-1}, ..., x_{t-p+1}, x_{t-p}$, η μερική αυτοσυσχέτιση μεταξύ των x_t και x_{t-p} μετράει την συσχέτιση τους, αλλά δεν λαμβάνει υπόψιν την επίδραση των ενδιάμεσων μεταβλητών. Ο προσδιορισμός της μερικής αυτοσυσχέτισης γίνεται από την εκτίμηση της τάξης του μοντέλου AR με αυξανόμενη τάξη [32]. Έτσι για το αυτοπαλινδρομούμενο μοντέλο πρώτης τάξης η μερική αυτοσυσχέτιση προσδιορίζεται με την βοήθεια της εξίσωσης (Εξ. 3.12).

$$x_t = \phi_{1,1} x_{t-1} + Z_t \tag{3.12}$$

όπου ο δεύτερος κατά σειρά δείκτης δείχνει την τάξη του μοντέλου, ενώ ο συντελεστής $\phi_{1,1}$ αποτελεί την μερική αυτοσυσχέτιση υστέρησης 1. Όσον αφορά το μοντέλο AR(2) η μερική αυτοσυσχέτιση προσδιορίζεται από την εξίσωση (Εξ. 3.13)

$$x_t = \phi_{1,2} x_{t-1} + \phi_{2,2} x_{t-2} + Z_t \tag{3.13}$$

Ο συντελεστής $\phi_{2,2}$ δηλώνει την προσφορά του x_{t-2} το οποίο προστίθεται σε αυτήν του αυτοπαλινδρομούμενου μοντέλου πρώτης τάξης για τον προσδιορισμό της χρονοσειράς X_t.

Αν η χρονοσειρά προέρχεται από μια στοχαστική διαδικασία τύπου AR(p) τότε ο συντελεστής $\phi_{p,p}$ δεν έχει μηδενική τιμή, ενώ όταν έχουμε υστερήσεις $k > p$ ο συντελεστής $\phi_{k,k}$ μπορεί να είναι ίσος με το μηδέν [24]. Το παραπάνω κριτήριο χρησιμοποιείται για τον προσδιορισμό της τάξης του μοντέλου, αν και υπάρχουν και άλλα κριτήρια για τον προσδιορισμό αυτής. Ως τάξη μο-

ντέλου θεωρείται το πλήθος των παραμέτρων που πρέπει να υπολογιστούν για να προσδιοριστεί το μοντέλο.

Τα κριτήρια αυτά βασίζονται στην πιθανοφάνεια των δεδομένων. Ως δείκτης πιθανοφάνειας θεωρείται η διασπορά των υπολοίπων (σφάλμα προσαρμογής) s_z^2 από την προσαρμογή του μοντέλου. Τα κριτήρια αυτά επιδιώκουν την εξισορρόπηση της μείωση του σφάλματος βάζοντας ποινή στην πολυπλοκότητα αυτού [32]. Αυτό μπορεί να γίνει με μια συνάρτηση κόστους της τάξης του μοντέλου που περιέχει το σφάλμα προσαρμογής και κάποιον όρο ποινής για την πολυπλοκότητα αυτού. Παρακάτω παρουσιάζονται δύο από τα πιο γνωστά κριτήρια για τον προσδιορισμό της τάξης της διαδικασίας τα οποία χρησιμοποιούν το σφάλμα προσαρμογής. Τα κριτήρια αυτά είναι

- Κριτήριο πληροφορίας Akaike

$$\mathrm{AIC}(p) = \ln(\sigma_z^2) + \frac{2p}{n} \tag{3.14}$$

- Κριτήριο Μπεϋζιανής πληροφορίας

$$\mathrm{BIC}(p) = \ln(\sigma_z^2) + \frac{p\ln(n)}{n} \tag{3.15}$$

όπου στις προηγούμενες μαθηματικές εκφράσεις το n εκφράζει το μήκος της χρονοσειράς, το σ_z^2 είναι η εκτιμώμενη διασπορά του σφάλματος και p είναι η τάξη του μοντέλου. Προκύπτει πως όσο μεγαλύτερη είναι η τάξη p τόσο το σφάλμα προσαρμογής γίνεται μικρότερο, ενώ για πολύ μεγάλες τάξεις το μοντέλο προσαρμόζεται περισσότερο στον λευκό θόρυβο παρά σε πραγματικές συσχετίσεις. Για αυτόν τον λόγο στην σχέση του κριτηρίου AIC υπάρχει ο δεύτερος όρος ο οποίος ονομάζεται όρος ποινής, και επιδρά αρνητικά αυξάνοντας την συνάρτηση AIC όταν και η τάξη του μοντέλου αυξάνεται [24]. Επομένως η τάξη του μοντέλου επιλέγεται μέσω της μικρότερης τιμής των κριτηρίων που προαναφέρθηκαν.

3.3 Μοντέλο κινούμενου μέσου MA(q)

Μια από τις βασικές στοχαστικές διαδικασίες είναι αυτή του κινούμενου μέσου τάξης q, που μαζί με την αυτοπαλινδρομούμενη διαδικασία έχουν καθοριστικό ρόλο στην ανάλυση των χρονοσειρών. Το μοντέλο αυτό προκύπτει από την σχέση (Εξ. 3.16).

$$x_t = Z_t - \theta_1 Z_{t-1} - \theta_2 Z_{t-2} - ... - \theta_q Z_{t-q} \tag{3.16}$$

όπου το Z_t είναι η μεταβλητή του λευκού θορύβου που ακολουθεί την κανονική κατανομή και τα $(\theta_1, \theta_2, ..., \theta_q)$ είναι οι παράμετροι του μοντέλου MA. Βασικό χαρακτηριστικό της MA(q) είναι η διατήρηση της στασιμότητας αφού δίνεται ως το άθροισμα των όρων του λευκού θορύβου. Η κύρια διαφορά του μοντέλου κινούμενου μέσου όρου με την αυτοπαλινδρομούμενη διαδικασία είναι ότι στην δεύτερη το καθοριστικό μέρος αντικαθίστανται από το στοχαστικό, και επομένως η μόνη πληροφορία που δίνεται για την X_t είναι από τις διαταράξεις στους q+1 πιο πρόσφατους χρόνους[8].

3.3.1 Στοχαστική διαδικασία πρώτης τάξης MA(1)

Το μοντέλο της σχέσης (Εξ. 3.16) για q=1, i.e. MA(1) δίνεται ως

$$x_t = Z_t - \theta Z_{t-1} = (1 - \theta B) Z_t \tag{3.17}$$

όπου το B εκφράζει τον τελεστή υστέρησης (lag operator). Για την διαδικασία αυτή ισχύουν τα εξής

- Διακύμανση: $\mathrm{Var}(x_t) = \sigma_x^2 = Var(Z_t + \theta_1 Z_{t-1}) = \sigma_Z^2 + \theta_1^2 \sigma_Z^2 = \sigma_Z^2(1+\theta_1^2)$
- Για να μπορέσουμε να προσδιορίσουμε την αυτοσυσχέτιση για κάποια υστέρηση τ πολλαπλασιάζουμε τον τύπο της MA(1) με την $X_{t-\tau}$. Έτσι

ξεκινώντας από την διασπορά X_t με μηδενική υστέρηση προκύπτει

$$X_t X_t = (Z_t - \theta Z_{t-1})(Z_t - \theta Z_{t-1}) \Rightarrow \sigma_x^2 = (1+\theta^2)\sigma_z^2 \tag{3.18}$$

ενώ για υστέρηση ένα έχουμε

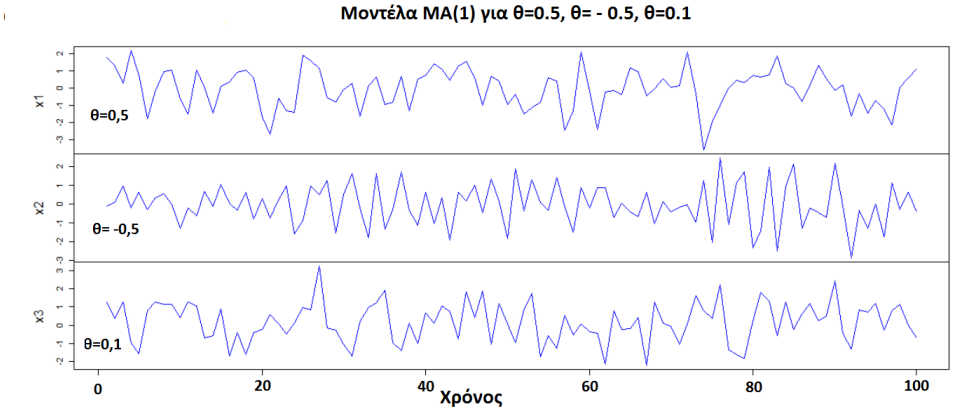
$$X_{t-1} X_t = (Z_{t-1} - \sigma Z_{t-2})(Z_t - \sigma_{t-1}) \Rightarrow r_1 = \frac{-\theta}{1+\theta^2} \tag{3.19}$$

Προκύπτει επομένως πως γενικά η συνάρτηση αυτοσυσχέτισης δίνεται από την σχέση (Εξ. 3.20)

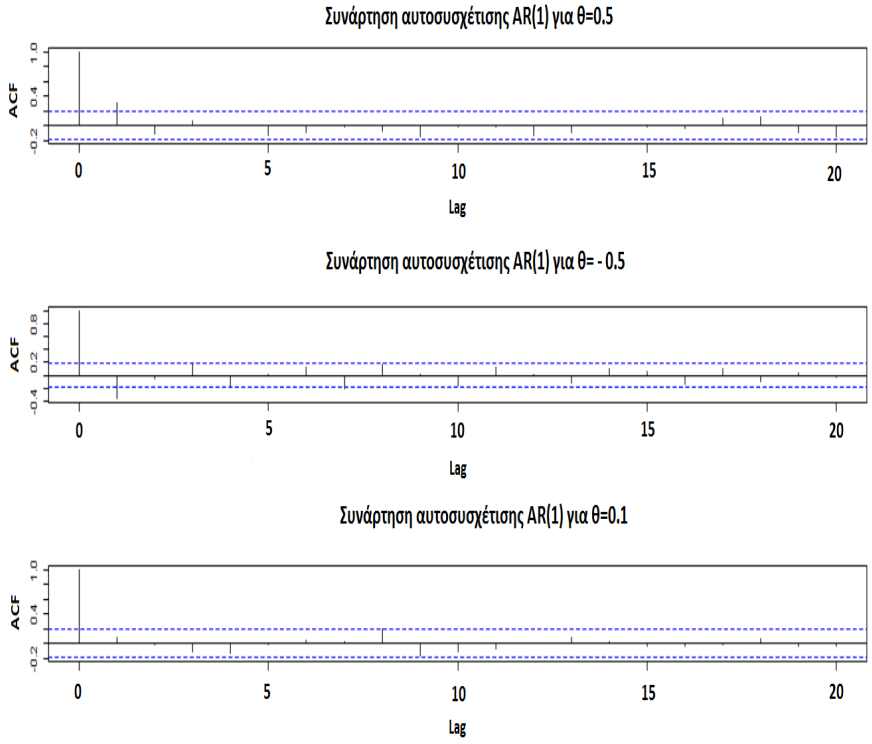
$$\begin{aligned} \rho_t &= \frac{\theta}{1+\theta^2}, \tau = 1 \\ \rho_\tau &= 0, \tau \geq 0 \end{aligned} \tag{3.20}$$

Από την προηγούμενη σχέση συμπεραίνουμε πως η μέγιστη τιμή είναι $\rho_1 = 0.5$ για $\theta = -1$ και η ελάχιστη $\rho_1 = -0.5$ για $\theta = 1$. Τέλος αξίζει να σημειωθεί ότι η σχέση που συνδέει την p_1 και το θ είναι γραμμική [24].

Η γραφική παράσταση και η συνάρτηση αυτοσυσχέτισης για τυχαίες τιμές του ϑ δίνονται παρακάτω στα γραφήματα (Σχ. 3.3, Σχ. 3.4).



Σχήμα 3.3: Μοντέλα MA για διαφορετικές τιμές του ϑ [14]



Σχήμα 3.4: Συνάρτηση Αυτοσυσχέτισης για διάφορες τιμές του ϑ [14]

3.3.2 Στοχαστική διαδικασία τάξης q

Αυτή η διαδικασία κινούμενου μέσου έχει αυτοσυσχέτιση η όποια είναι μη μηδενική μόνο για τις πρώτες q υστερήσεις και δίνεται από την σχέση (Εξ. 3.21)

$$\rho_\tau = \frac{-\theta_\tau + \theta_1 + \theta_{\tau+1} + ... + \theta_{q-\tau}\theta_q}{1 + \theta_1^2 + \theta_2^2 + ... + \theta_q^2}, \tau = 1, 2, ..., q \tag{3.21}$$

$$\rho_\tau = 0, t > q$$

όπου το τ εκφράζει την τιμή της υστέρησης ενώ η διασπορά του μοντέλου αυτού είναι

$$\sigma_x^2 = (1 + \theta_1^2 + ... + \theta_q^2)\sigma_z^2 \tag{3.22}$$

Η μερική αυτοσυσχέτιση φθίνει ανάλογα με με τις ρίζες του καθοριστικού πολυωνύμου. Τέλος γνωρίζουμε ότι για κάθε υστέρηση τ, η μερική αυτοσυσχέτιση μπορεί να υπολογιστεί από τις αυτοσυσχετίσεις $\rho_1, \rho_2, ..., \rho_q$ [23].

3.3.3 Προσδιορισμός της τάξης του MA μοντέλου

Για να μπορέσουμε να προσδιορίσουμε την τάξη q ενός μοντέλου κινούμενου μέσου πρέπει να χρησιμοποιήσουμε την παρατήρηση ότι η αυτοσυσχέτιση μηδενίζεται όταν πρόκειται για υστερήσεις οι οποίες είναι μεγαλύτερες από την τάξη q του μοντέλου που μελετάμε. Η τάξη του μοντέλου μπορεί να προσδιοριστεί μέσω της μεγαλύτερης υστέρησης που αναφέρεται σε μια στατιστικά μη μηδενική αυτοσυσχέτιση. Αξίζει να σημειωθεί ότι ο τρόπος αυτός προσδιορισμού της τάξης του MA μοντέλου είναι ο ίδιος με τον προσδιορισμό της τάξης του AR μοντέλου μέσω της μεγαλύτερης υστέρησης μη μηδενικής μερικής αυτοσυσχέτισης. Τέλος ο προσδιορισμός της τάξης αυτής μπορεί να γίνει,όπως και στο AR μοντέλο, μέσω των κριτηρίων AIC, BIC αλλά και με την βοήθεια των ACF, PACF όπως φαίνεται στον πίνακα που ακολουθεί

Διαδικασία	**ACF**	**PACF**
AR(1)	Εκθετική μείωση: θετικές τιμές αν ϕ>0, εναλλαγή προσήμου ξεκινώντας από αρνητική τιμή, αν ϕ<0	Απότομος μηδενισμός της περιόδου. Η τιμή την περίοδο 1 είναι: θετική αν ϕ>0 ή αρνητική αν ϕ<0
AR(p)	Οι τιμές των συντελεστών του ACF φθίνουν προς το μηδέν ακολουθώντας εκθετική ή ημιτονοειδή πορεία	Μη μηδενικές τιμές για τις πρώτες p περιόδους και στην συνέχεια απότομος μηδενισμός
MA(1)	Απότομος μηδενισμός μετά την περίοδο 1. Η τιμή της περιόδου 1 είναι : θετική αν ϑ<0 ή αρνητική αν ϑ>0	Εκθετική μείωση: εναλλαγή προσήμου ξεκινώντας από θετική τιμή αν ϑ>0 και αρνητικές τιμές αν ϑ<0
MA(q)	Μη μηδενικές τιμές για τις πρώτες χ περιόδους και στην συνέχεια απότομος μηδενισμός	Εκθετική μείωση ή φθίνουσα ημιτονοειδής συνάρτηση. Το ακριβές μέγεθος εξαρτάται από το πρόσημο και το μέγεθος του ϑ

Πίνακας 3.1: Εκτίμηση της τάξης των AR και MA με τη βοήθεια των ACF, PACF

3.4 Αυτοπαλινδρομούμενη διαδικασία κινούμενου μέσου (ARMA)

Τα AR, MA μοντέλα έχουν διαφορετικές ιδιότητες οι οποίες προκύπτουν από τον ορισμό τους. Όσον αφορά το αυτοπαλινδρομούμενο μοντέλο, εκφράζετε μέσω του χαρακτηριστικού πολυωνύμου υστέρησης $\phi(B)X_t = Z_t$, ενώ η διαδικασία του κινούμενου μέσου προκύπτει με την βοήθεια του δικού της χαρακτηριστικού πολυωνύμου υστέρησης $\theta(B)$ ως $X_t = \theta(B)Z_t$. Οι συσχέτιση των δύο αυτών διαφορετικών διαδικασιών γίνεται με την βοήθεια της αυτοπαληνδρομούμενης διαδικασίας κινούμενου μέσου, η οποία ονομάζεται και μικτή διαδικασία αφού συνδυάζει τις δυο προηγούμενες [24]. Η διαδικασία ARMA

δίνεται από το πολυώνυμο

$$X_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + Z_t - \theta_1 Z_{t-1} - \theta_2 Z_{t-2} - ... - \theta_q Z_{t-q} \quad (3.23)$$

ή με την βοήθεια του πολυωνύμου υστέρησης

$$\phi(B) X_t = \theta(B) Z_t \quad (3.24)$$

Η διαδικασία ARMA γράφεται και ως ARMA(p, q) αφού περιέχει την τάξη p του αυτοπαλινδρομούμενου μοντέλου καθώς και την τάξη q της διαδικασίας κινούμενου μέσου.

Σε ορισμένες περιπτώσεις προτιμάται η χρησιμοποίηση μεγάλης τάξης AR και MA μοντέλων αφού η εφαρμογή της μικτής διαδικασίας μπορεί να θεωρηθεί περιττή. Παρόλο αυτά η ARMA διαδικασία μπορεί να χρησιμοποιηθεί για πλήθος διαφορετικών περιστάσεων. Απαραίτητη προϋπόθεση είναι η στασιμότητα της χρονοσειράς [2]. Εάν η χρονοσειρά που διαθέτουμε δεν είναι στάσιμη τότε απαιτείται η μετατροπή της σε τέτοια. Η στασιμότητα της διαδικασίας ARMA ορίζεται από το AR μέρος, δηλαδή είναι στάσιμη μόνο όταν το χαρακτηριστικό πολυώνυμο $\phi(B)$ έχει ρίζες έκτος του μοναδιαίου κύκλου. Τότε η χρονοσειρά μπορεί να εκφραστεί ως

$$X_t = \frac{\theta(B)}{\phi(B)} Z_t \quad (3.25)$$

Σημαντικό ρόλο έχει και η αντιστρεψιμότητα της διαδικασίας η οποία εξαρτάται αντίστοιχα από την διαδικασία κινούμενου μέσου. Αυτό σημαίνει πως η διαδικασία είναι αντιστρέψιμη όταν το πολυώνυμο $\theta(B)$ έχει ρίζες εκτός του μοναδιαίου κύκλου και τότε εκφράζεται ως

$$\frac{\phi(B)}{\theta(B)} X_t = Z_t \quad (3.26)$$

Γενικά για όλες τις αυτοπαλινδρομούμενες διαδικασίες κινούμενου μέσου γνωρίζουμε τα εξής

- $E(X_{t-k}\epsilon_t) = 0 \ \forall k > 0$
- $E(X_t\epsilon_t) = \sigma^2$

Πρέπει όμως να ισχύει πως το $E(\epsilon_{t-k}\epsilon_t) = 0$ για $\forall k \neq 0$ και η συνδιακύμανση $E(X_{t-k}\epsilon_t)$ να περιέχει στοχαστικές διαταραχές μέχρι τη χρονική στιγμή $t - k$.

3.4.1 Αυτοπαλινδρομούμενη διαδικασία κινούμενου μέσου πρώτης τάξης

Η πιο απλή μικτή διαδικασία μερικής αυτοσυσχέτισης με τάξη ένα τόσο στο AR όσο και στο MA μοντέλο, είναι η ARMA(1, 1) . Έχει ιδιαίτερη πρακτική σημασία καθώς μέσω αυτής μπορούν να περιγράφουν στοχαστικές διαδικασίες που υπό άλλες προϋποθέσεις θα χρειαζόταν πολύ μεγάλης τάξης αυτοπαλινδρομούμενα και κινούμενου μέσου όρου μοντέλα. Η ARMA(1, 1) ορίζεται μέσω μίας γραμμικής εξίσωσης με έναν σταθερό συντελεστή. Έτσι προκύπτει η εξίσωση (Εξ. 3.27).

$$X_t = \phi X_{t-1} = Z_t + \theta Z_{t-1} \Rightarrow (1 - \phi B)X_t = (1 - \theta B)Z_t \tag{3.27}$$

Γνωρίζουμε πως το μοντέλο ARMA είναι ένας συνδυασμός των AR, MA μοντέλων. Ως εκ τούτου όταν ο συντελεστής ϑ είναι ίσος με το μηδέν τότε προκύπτει ότι ARMA(1, 1)=MA(1), επομένως η διαδικασία γίνεται ARMA(0,1). Αντίστοιχα όταν ο συντελεστής φ είναι ίσος με το μηδέν τότε προκύπτει ότι ARMA(1, 1)=AR(1) και επομένως η διαδικασία γίνεται πλέονARMA(1, 0) [29].

Για να μπορέσει επομένως να υπάρξει ανιστρεψιμότητα πρέπει $|\theta| < 1$, ενώ για την ύπαρξη στασιμότητας είναι απαραίτητη η συνθήκη $\phi < 1$. Όσον αφορά την αυτοσυσχέτιση της ARMA(ρ, q) μπορεί να υπολογισθεί με τον ίδιο

τρόπο, όπως για AR(p) και MA(q), πολλαπλασιάζοντας τα δύο μέρη της ισότητας (Εξ. 3.27) με $X_{t-\tau}$ για κάθε υστέρηση τ και εφαρμόζοντας τον τελεστή της μέσης τιμής όπου το γ_t εκφράζει την αυτοδιασπορά. Έτσι προκύπτει

$$X_{t-\tau}X_t = X_{t-\tau}(\phi X_{t-1} + Z_t - \theta Z_{t-1}) = \\ = \gamma_\tau = \phi\gamma_{\tau-1} + E[X_{t-\tau}Z_t] - \theta E[X_{t-\tau}Zt-1] \quad (3.28)$$

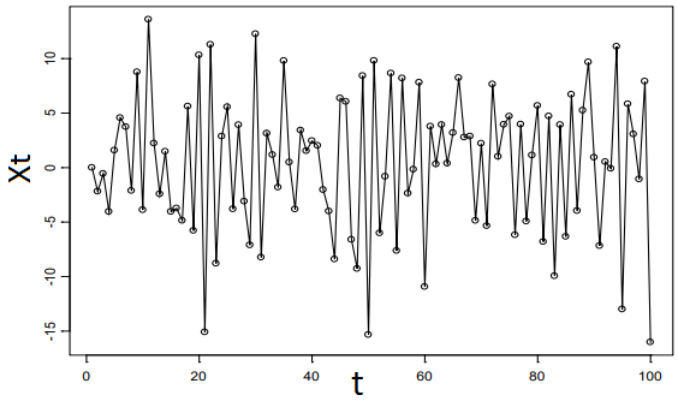
Η εξίσωση αυτή μας δείχνει πως η αυτοσυσχέτιση στην ARMA(1, 1) διαδικασία έχει την μορφή εκθετικής πτώσης, με την ιδιαιτερότητα ότι η πτώση ξεκινά από την ρ_1 και όχι από την ρ_0. Η σχέση της αυτοσυσχέτισης σε συνδυασμό με την σχέση αυτοδιασποράς για την ARMA(p, q) διαδικασία $\gamma_\tau = \phi_1\gamma_{\tau-1}+...+\phi_p\gamma_{\tau-p}$ για τ=0 και τ=1 μας δίνει τις σχέσης (Εξ. 3.29), (Εξ. 3.30).

$$\sigma_x^2 = \gamma_0 = \phi\gamma_1 + \sigma_z^2 - \theta(\phi - \theta)\sigma_z^2 \quad (3.29)$$

$$\gamma_1 = \phi\gamma_0 - \theta\sigma_z^2 \quad (3.30)$$

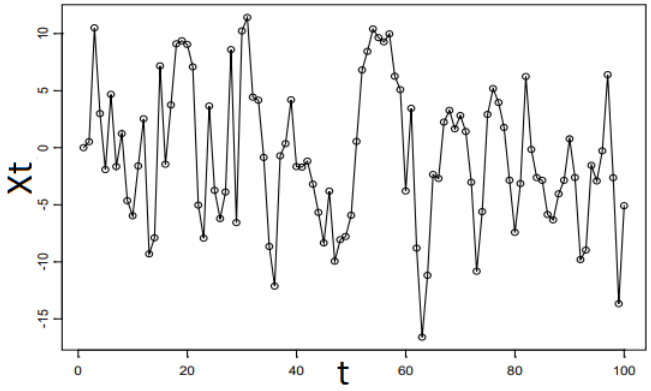
Γενικά μια ARMA(p, q) μπορεί και για μικρές τάξης των δεικτών της να παρουσιάσει κάποια μορφή αυτοσυσχέτισης και μερικής αυτοσυσχέτισης. Συγκεκριμένα η συνάρτηση της μερικής αυτοσυσχέτισης έχει την μορφή εκθετικής πτώσης, όπως συμβαίνει και με την συνάρτηση μερικής αυτοσυσχέτισης για την MA(1) διαδικασία. Παραδείγματα της εκθετικής αυτής πτώσης παρουσιάζονται στα παρακάτω σχήματα (Σχ. 3.5, Σχ. 3.6) όπου φαίνεται η μορφή των χρονολογικών σειρών και οι αντίστοιχες συναρτήσεις μερικής αυτοσυσχέτισης για τους τυχαίους συνδυασμούς τιμών $\theta_1 = 0.5$, $\phi_1 = -0.5$ και $\theta_1 = -0.5$, $\phi_1 = 0.5$. Επομένως ισχύει ότι

Α)Για $\theta_1 = 0.5$, $\phi_1 = -0.5$



Σχήμα 3.5: Συνάρτηση Μερικής Αυτοσυσχέτισης Α [16]

Β)Για $\theta_1 = -0.5$, $\phi_1 = 0.5$



Σχήμα 3.6: Συνάρτηση Μερικής Αυτοσυσχέτισης Β [16]

3.4.2 Χρήσιμες πληροφορίες για τα μοντέλα AR, MA και ARMA

Μετά την εκτενή ανάλυση των μοντέλων που ανήκουν στην οικογένεια των μοντέλων SARIMA, και συγκεκριμένα των AR, MA, ARMA θα γίνει παράθεση

κάποιων βασικών σχολιασμών που μπορούν να προκύψουν από την ανάλυση αυτή.

α) Παρατηρείται μια προτίμηση όσον αφορά τις εφαρμογές του μοντέλου AR έναντι του MA και αυτό γιατί μπορούν να αποτυπώσουν καλύτερη ερμηνεία του συστήματος επίλυσης μίας και προσδίδουν πιο βραχυπρόθεσμη μνήμη στο σύστημα, αλλά και έχει παρατηρηθεί πως είναι πιο ικανά στην πρόβλεψη μιας χρονοσειράς [24].

β) Δεν υπάρχει ένα μοντέλο που να δίνει βέλτιστα αποτελέσματα, αλλά μπορούν δύο ή και περισσότερα να δίνουν εξίσου ικανοποιητικά αποτελέσματα. Έτσι για παράδειγμα ένα μοντέλο ARMA(p, q) με μικρής τάξης συντελεστές μπορεί να δώσει τα ίδια αποτελέσματα με ένα μοντέλο AR(p) ή MA(q) μεγάλης τάξης.

γ) Δεν ενδείκνυται η προσπάθεια προσαρμογής μοντέλων μεγάλης τάξης γιατί ενώ θεωρητικά είναι καλύτερη στην ουσία μπορεί να μην αντιστοιχεί μόνο σε πληροφορίες του συστήματος αλλά και σε θόρυβο.

3.5 Ολοκληρωμένο αυτοπαλινδρομούμενο μοντέλο κινούμενου μέσου (ARIMA)

Το μοντέλο αυτό προκύπτει από τον συνδυασμό του μετασχηματισμού των πρώτων διαφορών και του μοντέλου ARMA(p, q) που προαναφέρθηκε. Βασική προϋπόθεση χρησιμοποίησης του μοντέλου αυτού είναι η στασιμότητα της χρονοσειράς, η οποία μπορεί να επιτευχθεί με την εφαρμογή d επαναλήψεων των πρώτων διαφορών και την προσαρμογή κάποιου ARMA μοντέλου για να προκύψει το μοντέλο ARIMA το οποίο συμβολίζεται ως ARIMA(p, d, q), όπου το p είναι ένα πολυώνυμο σύνθετου βαθμού του AR μοντέλου και συγκεκριμένα ένας μη αρνητικός ακέραιος αριθμός, το q είναι ένα πολυώνυμο σύνθετου

βαθμού του MA μοντέλου το οποίο είναι επίσης ένας μη αρνητικός ακέραιος αριθμός και το d είναι ένας μη αρνητικός ακέραιος αριθμός ο οποίος σχετίζεται με τον βαθμό μη περιοδικής διαφοροποίησης [24].

Για να μπορέσει μια στοχαστική διαδικασία Y_t να χαρακτηριστεί ως διαδικασία ARIMA(p, d, q) θεωρούμε πως η Y_t είναι μια μη στάσιμη διαδικασία και παρουσιάζει φαινόμενα τάσης. Πρέπει να εξετάσουμε αν η χρονοσειρά που προκύπτει από τις πρώτες διαφορές, $X_t = Y_t - Y_{t-1}$, μπορεί να θεωρηθεί στάσιμη. Εάν δεν είναι τότε εφαρμόζουμε διαφορές δεύτερης τάξης και προκύπτει $X'_t = Y_t - 2Y_{t-1} + Y_{t-2}$. Η διαδικασία αυτή θα συνεχιστεί μέχρι να δεχθούμε πως για κάποια τάξη d η χρονοσειρά X_t που ορίζεται ως

$$X_t = \nabla^d Y_t = (1 - B)^d Y_t \tag{3.31}$$

είναι στάσιμη. Μπορούμε να θεωρήσουμε πως η χρονοσειρά X_t περιγράφεται ως κάποια διαδικασία ARMA(p, q) όπου ισχύει πως $\phi(B)X_t = \theta(B)Z_t$ και η διαδικασία Y_t μπορεί να οριστεί ως $\phi(B)(1 - B)^d Y_t = \theta(B)Z_t$. Η σχέση αυτή δηλώνει πως η διαδικασία Y_t θεωρείτε ως μια ARMA διαδικασία που έχει μια μοναδιαία ρίζα με πολλαπλότητα όσο η τάξη d. Για αυτό ορίζεται ως διαδικασία ARIMA(p, d, q) [24].

3.5.1 Εποχιακό αυτοπαλινδρομούμενο μοντέλο κινούμενου μέσου (SARIMA)

Το αυτοπλανινδρομούμενο μοντέλο κινούμενου μέσου, ARMA είναι μια από τις πιο συχνά χρησιμοποιούμενες μεθόδους πρόβλεψης χρονοσειρών. Παρόλο αυτά το μεγάλο πρόβλημα που παρουσιάζει το μοντέλο αυτό έγκειται στο γεγονός ότι μπορεί να εφαρμοστεί μόνο σε χρονοσειρές οι οποίες έχουν τάση. Όταν όμως υπάρχει κάποια χρονοσειρά η οποία περιέχει και τάση και περιοδικότητα τότε το ARMA μοντέλο δεν είναι αποδοτικό και δεν μπορεί να εφαρμοστεί. Για αυτόν τον λόγο χρησιμοποιείται μια προέκταση του μοντέλου αυτού η οποία μπορεί

να διαχειριστεί χρονοσειρές με περιοδικότητα. Η προέκταση αυτή λέγεται εποχιακό αυτοπαλινδρομούμενο μοντέλο κινούμενου μέσου. Το μοντέλο SARIMA προσθέτει τρεις επιπλέον παραμέτρους για να προσδιορίσει την αυτοπαλινδρόμηση (AR), το κινούμενο μέσο όρο (MA) και την διαφοροποίηση l για το εποχιακό κομμάτι της χρονοσειράς καθώς και μία ακόμα παράμετρο για την περίοδο της εποχικότητας [24]. Έτσι προκύπτει πως το μοντέλο SARIMA συμβολίζεται ως SARIMA(p, d, q)(P, D, Q)m και προκύπτει από τον συνδυασμό του γενικού μοντέλου $\phi(B)(1-B)^d Y_t = \theta(B)Z_t$ για χρονοσειρά με τάση και του γενικού μοντέλου $\Phi(B^s)(1-B)^D Y_t = \Theta(B^s)Z_t$ για χρονοσειρά με εποχικότητα. Για το SARIMA(p, d, q)(P, D, Q)m ισχύει ότι

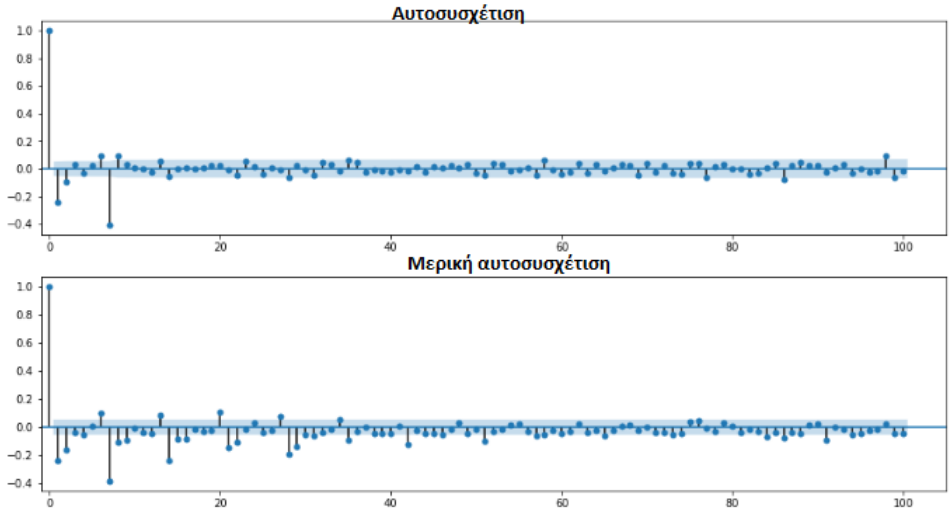
- $p =$ όρος της αυτοπαλινδρόμησης για την τάση
- $d =$ όρος της διαφοροποίησης για την τάση
- $q =$ όρος του κινούμενου μέσου όρου για την τάση
- $P =$ όρος της αυτοπαλινδρόμησης για την εποχικότητα
- $D =$ όρος της διαφοροποίησης για την εποχικότητα
- $Q =$ όρος του κινούμενου μέσου όρου για την εποχικότητα
- $m =$ ο αριθμός των βημάτων που ακολουθούνται για μία συγκεκριμένη περίοδο. Για παράδειγμα σε ένα SARIMA(3, 1, 0)(1, 1, 0)[12] το 12 υποδηλώνει εποχικότητα που αντιστοιχεί σε επαναληψιμότητα κάθε 12 βήματα μπροστά.

Παρόλο αυτά τις περισσότερες φορές τα μοντέλα αυτά έχουν πιο απλή μορφή. Αυτό οφείλεται στο γεγονός ότι η τάξη των πρώτων διαφορών είναι συνήθως ίση με ένα και η τάξη των s διαφορών είναι $\delta = 0$ δηλαδή προκύπτει ένα μοντέλο

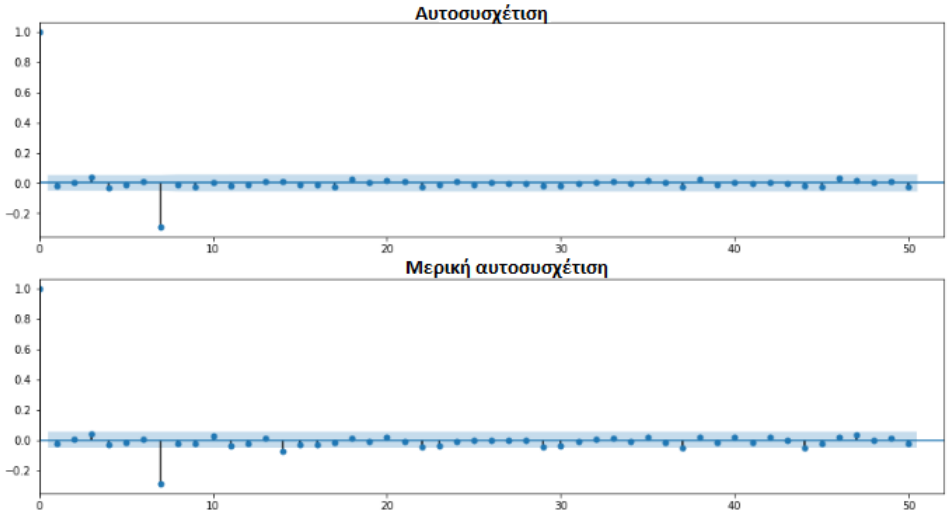
με τύπο SARIMA(p, 1, q)(P, D) και εκφράζεται μέσω της εξίσωσης (Εξ. 3.32) με χρονικό βήμα s

$$\phi(B)\Phi(B^s)(1-B)X_t = \theta(B)\Theta(B^s)Z_t \tag{3.32}$$

Ο πιο ενδεδειγμένος τρόπος για να μπορέσουμε να διακρίνουμε το καταλληλότερο μοντέλο είναι μέσω του διαγράμματος ACF(autocorelation function) και PACF(partial autocorelation function). Τα δύο αυτά διαγράμματα χρησιμοποιούνται ως ένα μέσο προσδιορισμού κάποιον παραμέτρων του μοντέλου SARIMA. Για παράδειγμα δύο ενδεικτικά διαγράμματα (Σχ. 3.7, Σχ. 3.8) για μια τυχαία χρονοσειρά αποτελούν τα ακόλουθα



Σχήμα 3.7: Διαγράμματα αυτοσυσχέτισης και μερικής αυτοσυσχέτισης [26]



Σχήμα 3.8: Διαγράμματα αυτοσυσχέτισης και μερικής αυτοσυσχέτισης [26]

Στα γραφήματα αυτά παρατηρούνται έντονες αρνητικές τιμές στα σημεία 1 και 7 όσον αφορά το διάγραμμα ACF, ενώ στο διάγραμμα του PACF οι τιμές αυτές παρατηρούνται στις θέσεις 1 και 2 όπως και στα σημεία 7, 14, 21 και ούτω καθεξής. Οι παρατηρήσεις αυτές μπορούν να μας οδηγήσουν στο συμπέρασμα ότι το μοντέλο μας παίρνει την μορφή SARIMA$(0, 1, 1)(0, 1, 1)[7]$. Εφαρμόζοντας το μοντέλο αυτό θα παρατηρήσουμε ότι οι αρνητικές τιμές θα εξαλειφθούν, καθιστώντας το μοντέλο μας αποδεκτό.

3.6 Εκθετικές μέθοδοι εξομάλυνσης

3.6.1 Εκθετικός κινούμενος μέσος, Exponational Moving Average (EMA)

Ο εκθετικός κινούμενος μέσος είναι ένας τύπος κινούμενου μέσου (υπό-εν. 1.3.1) ο οποίος δίνει επιπλέον σημασία στα πιο πρόσφατα δεδομένα. Πολλές φορές η

διεργασία EMA αναφέρεται και ως εκθετικά σταθμισμένος κινούμενος μέσος όρος ή exponentially weighted moving average (EWMA). Η βασική διαφορά του μοντέλου αυτού και του απλού κινούμενου μέσου όρου (simple moving average(SMA)) είναι ότι ο πρώτος δίνει μεγαλύτερη σημασία στις πρόσφατες διαφοροποίησης των τιμών και η στάθμιση που εφαρμόζει η μέθοδος μειώνεται εκθετικά όσο παλαιότερα είναι τα δεδομένα αλλά ποτέ δεν φτάνει στο 0. Η μέθοδος του εκθετικά σταθμισμένου κινούμενου μέσου όρου συνήθως χρησιμοποιείται για την δημιουργία σημάτων που απευθύνονται στον τομέα της αγοράς και των πωλήσεων [9] και εκφράζεται μαθηματικά μέσω της σχέσης (Εξ. 3.33)

$$\hat{X}_t = X_{t-1} + \alpha[\text{value}_t - X_{t-1}] \tag{3.33}$$

όπου ο συντελεστής α αντιπροσωπεύει τον βαθμό της μειώσεις του σταθμισμένου όρου και παίρνει τιμές μεταξύ του 0 και του 1. Το X_t εκφράζει την τιμή του μεγέθους την τωρινή χρονική στιγμή. Το X_{t-1} εκφράζει την τιμή του ίδιου μεγέθους μία προηγούμενη χρονική στιγμή.

Επεκτείνοντας τα αποτελέσματα της μεταβλητής X_{t-1} κάθε χρονική στιγμή, προκύπτει πως ο σταθμισμένος όρος κάθε παράγοντα $p_1, p_2, ..., p_n$ μειώνεται εκθετικά. Έτσι η (Εξ. 3.33) μπορεί να πάρει την μορφή

$$\hat{X}_t = \alpha[p_1 + (1-\alpha)p_2 + (1-\alpha)^2 p_3 + (1-\alpha)^3 p_4 + ...] \tag{3.34}$$

όπου p_1 είναι η τωρινή τιμή, p_2 είναι η τιμή της μεταβλητής κάποια προηγούμενη στιγμή και ούτω καθεξής. Υψηλότερη τιμή του συντελεστή υποδηλώνει γρηγορότερη μείωση των παλαιότερων τιμών. Τέλος ένας ακόμα τρόπος απόδοσης του σταθμισμένου εκθετικού κινούμενου μέσου όρου όταν $1/\alpha = 1 + (1-\alpha) + (1-\alpha)^2 + ...$, δίνεται μέσω της εξίσωσης (Εξ. 3.35)

$$\hat{X}_t = \frac{p_1 + (1-\alpha)p_2 + (1-\alpha)^2 p_3 + (1-\alpha)^3 p_4 + ...}{1 + (1-\alpha) + (1-\alpha)^2 + (1-\alpha)^3 + ...} \tag{3.35}$$

Τα τρία βασικά βήματα για τον υπολογισμό της EWMA είναι

- Ο προσδιορισμός του απλού κινούμενου μέσου
- Ο προσδιορισμός του πολλαπλασιαστή για το σταθμισμένο παράγοντα της προηγούμενης τιμής της EWMA
- Ο υπολογισμός την τρέχουσας τιμής της EWMA

Πιο αναλυτικά, πριν γίνει ο προσδιορισμός ενός EWMA μοντέλου πρέπει να υπολογιστεί ο απλός κινούμενος μέσος όρος (SMA) πάνω σε μία συγκεκριμένη χρονική περίοδο. Ο προσδιορισμός του μοντέλου αυτού είναι στην πραγματικότητα πολύ απλός. Για παράδειγμα προκύπτει από το άθροισμα των τιμών κλεισίματος του χρηματιστηρίου για μια συγκεκριμένη χρονική περίοδο διαιρεμένο με το σύνολο τους [9]. Στην συνέχεια πρέπει να προσδιοριστεί ο πολλαπλασιαστής του σταθμισμένου παράγοντα της EWMA, ο οποίος εάν έχουμε μια περίοδο n ημερών συνήθως είναι ίσος με $2/n + 1$ και έτσι είναι εφικτός ο προσδιορισμός της τιμής του EWMA.

3.6.2 Βασικά πλεονεκτήματα και περιορισμοί της EWMA

Τα βασικά πλεονεκτήματα ενός τέτοιου μοντέλου είναι τα εξής:

- Μπορεί να χρησιμοποιηθεί για να βρεθεί ο μέσος όρος ενός ολόκληρου ιστορικού δεδομένων.
- Ο χρήστης μπορεί να δώσει σημασία στα δεδομένα αυτά που τον διευκολύνουν περισσότερο.
- Κάθε δεδομένο στον εκθετικό σταθμισμένο κινούμενο μέσο όρο αντιπροσωπεύει έναν κινούμενο μέσο όρο από σημεία.

Όσον αφορά τους περιορισμούς ενός EWMA μοντέλου αυτοί είναι:

- Το μοντέλο μπορεί να χρησιμοποιηθεί μόνο όταν συνεχή δεδομένα κατά την διάρκεια μίας περιόδου είναι διαθέσιμα.
- Μπορεί να χρησιμοποιηθεί μόνο όταν θέλουμε να ανιχνεύσουμε μικρές μεταβολές στην διαδικασία.
- Η μέθοδος αυτή μπορεί να χρησιμοποιηθεί για να προσδιορίσουμε τον μέσο όρο. Για την παρακολούθησή της διακύμανσης των δεδομένων απαιτείται η χρησιμοποίηση άλλων μεθόδων, αφού η EWMA είναι ανεπαρκής για την διεργασία αυτή.

3.6.3 Διαφορές μεταξύ EWMA και SMA

Μία από τις βασικές διαφορές ενός εκθετικού σταθμισμένου κινούμενου μέσου και ενός απλού κινούμενου μέσου είναι ο διαφορετικός βαθμός ευαισθησίας στις αλλαγές που εμφανίζονται στα δεδομένα που χρησιμοποιήθηκαν στους υπολογισμούς. Αναλυτικότερα το μοντέλο EWMA λαμβάνει υπόψιν κυρίως τις πιο πρόσφατες τιμές των δεδομένων, ενώ το SMA δίνει ίσο βάρος σε όλες τις τιμές του συνόλου των στοιχείων που διαθέτουμε. Και οι δύο μέσοι όροι έχουν παραπλήσια χρήση, γιατί ερμηνεύονται με τον ίδιο τρόπο και χρησιμοποιούνται από αναλυτές για να εξομαλύνουν διακυμάνσεις στις τιμές. Το μοντέλο EWMA εφαρμόζει μεγαλύτερη στάθμιση σε πρόσφατα δεδομένα σε σχέση με τα παλαιότερα, είναι πιο αντιδραστικό στις τελευταίες αλλαγές των τιμών σε σχέση με το SMA μοντέλο, το οποίο δημιουργεί μεγαλύτερη συσχέτιση μεταξύ του χρόνου και των δεδομένων που διαθέτουμε. Έτσι μπορεί να δοθεί μια εξήγηση σχετικά με το γιατί η EWMA προτιμάται σε σχέση με άλλες μέθοδος προσδιορισμού του μέσου όρου. [12]

Κεφάλαιο 4

Πρόβλεψη Χρονοσειρών

4.1 Βασικά στάδια στην διαδικασία πρόβλεψης

Η πρόβλεψη χρονοσειρών είναι μια από τις βασικές διεργασίες που επιτελούνται στην στατιστική ανάλυση και έχει αποκτήσει ιδιαίτερο ενδιαφέρον τα τελευταία χρόνια αφού ο μεγάλος όγκος των δεδομένων κάνει εφικτή την επεξεργασία αυτών για την εξαγωγή αποτελεσμάτων και συμπερασμάτων για μελλοντικές χρονικές στιγμές. Έτσι αρχικά θα γίνει αναφορά στα 5 βασικά στάδια της διαδικασίας πρόβλεψης [17]

1^o **στάδιο: Καθορισμός προβλήματος** Το πιο σημαντικό πρόβλημα κατά την διαδικασία πρόβλεψης είναι ο καθορισμός του προβλήματος. Είναι έτσι απαραίτητη η διευκρίνηση ορισμένων παραμέτρων όπως για παράδειγμα του πως θα χρησιμοποιηθούν οι προβλέψεις και από ποίους.

2^o **στάδιο: Συγκέντρωση πληροφοριών** Σε αυτό το σημείο είναι απαραίτητη η διευκρίνηση δύο καθοριστικών παραμέτρων. Η πρώτη αναφέρεται στα στατιστικά δεδομένα και η δεύτερη αναφέρεται στην εμπειρία και τις γνώσεις του προσωπικού που ασχολείται με την συλλογή πληροφοριών που θα μας βοηθήσουν στην καλύτερη δυνατή διεκπεραίωση του προβλήματος.

3^o **στάδιο: Προκαταρτική ανάλυση** Στο στάδιο αυτό μας απασχολεί το είδος της πληροφορίας που αποκομίζουμε από τα ακατέργαστα δεδομένα. Έτσι αρχικά πρέπει να απεικονίσουμε την χρονοσειρά γραφικά και στην συνέχεια να υπολογίσουμε κάποιους βασικούς στατιστικούς δείκτες, όπως η μέση τιμή, η τυπική απόκλιση, η γραμμική τάση κ.α. Σκοπός της διαδικασίας αυτής είναι να αποκτηθεί μία πρώτη εικόνα των δεδομένων, αποκτώντας σημαντικές πληροφορίες σχετικά με την ύπαρξη ή όχι τάσης και εποχικότητας ή με την ύπαρξη ή όχι ιδιόμορφων τιμών (outliers). Το στάδιο αυτό μας δίνει σημαντικές πληροφορίες, οι οποίες θα μας οδηγήσουν στην επιλογή του καταλληλότερου μοντέλου πρόβλεψης.

4^o **στάδιο: Επιλογή και προσαρμογή του μοντέλου** Στο σημείο αυτό είναι εφικτή η επιλογή και ο καθορισμός των παραμέτρων διάφορων μοντέλων πρόβλεψης που έχουν ήδη επιλεγεί στο προηγούμενο βήμα.

5^o **στάδιο: Χρήση και αποτίμηση του μοντέλου πρόβλεψης** Αποτελεί το τελευταίο στάδιο της διαδικασίας αφού το μοντέλο έχει πλέον επιλεγεί και οι παράμετροι του έχουν χρησιμοποιηθεί ώστε να παραχθούν οι προβλέψεις. Κατά την πραγματοποίηση του βήματος αυτού γίνεται αποτίμηση του μοντέλου μέσω των πλεονεκτημάτων του και των μειονεκτημάτων του, και εφόσον χρειαστεί γίνεται επανάληψη κάποιον εκ των βημάτων που ανεφέρθηκαν προηγουμένως.

4.2 Πρόβλεψη χρονοσειρών με την χρήση μοντέλων SARIMA

Αφού πραγματοποιήθηκε η ανάλυση των βασικών μοντέλων SARIMA (κεφαλ. 3), θα χρησιμοποιήσουμε τα μοντέλα αυτά για την πραγματοποίηση πρόβλεψης θεωρώντας πως η χρονοσειρά είναι στάσιμη. Η επιλογή του μοντέλου είναι πολύ σημαντική αφού η καταλληλότητα αυτού θα κρίνει και την εγκυρότητά-αξιοπιστία της πρόβλεψης που θα πραγματοποιηθεί. Επιπλέον για το κεφάλαιο αυτό θα υποθέσουμε ότι η χρονοσειρά που μελετάμε θα έχει μηδενική μέση τιμή.

4.2.1 Πρόβλεψη με την βοήθεια της διαδικασίας Box-Jenkins

Το μοντέλο του ολοκληρωμένου αυτοπαλινδρομούμενου κινούμενου μέσου προτάθηκε αρχικά από τους Box και Jenkins και λόγο αυτού είναι συχνά γνωστό ως υπόδειγμα Box-Jenkins. Το βασικό χαρακτηριστικό των μοντέλων αυτών είναι ότι πρόκειται για εμπειρικά υποδείγματα, που σημαίνει ότι δημιουργούνται από δεδομένα και για αυτό για την κατασκευή τους είναι απαραίτητο να εφαρμοστεί η επαναληπτική διαδικασία που προτάθηκε από τους Box και Jenkins. Η διαδικασία αυτή περιλαμβάνει 4 βήματα τα οποία περιγράφονται παρακάτω [26]

Βήμα πρώτο: Ταυτοποίηση του υποδείγματος Με τον όρο ταυτοποίηση του υποδείγματος εννοούμε την προσπάθεια προσδιορισμού:

α) της τάξης της μη στασιμότητας

β) της τάξης των AR, MA πολυωνύμων

Για να γίνει εφικτός ο προσδιορισμός αυτός πρέπει να πραγματοποιηθε-

ί σύγκριση της μορφής των δειγματικών συναρτήσεων μερικής ή μη αυτοσυσχέτισης με τις συναρτήσεις των θεωρητικών συναρτήσεων μερικής ή μη αυτοσυσχέτισης οι οποίες αντιστοιχούν σε διαδικασίες με άπειρο πλήθος όρων.

Βήμα δεύτερο: Εκτίμηση του υποδείγματος Η πιο συνήθης μέθοδος για την εκτίμηση των παραμέτρων του υποδείγματος είναι η μέθοδος της μέγιστης πιθανοφάνειας. Για να μπορέσει να γίνει εφικτή η πραγματοποίηση της μεθόδου πρέπει οι εκτιμήσεις να είναι εντός των ορίων αντιστρεψιμότητας, στασιμότητας και να είναι στατιστικά σημαντικές

Βήμα τρίτο: Διάγνωση του υποδείγματος Στο βήμα αυτό πρέπει να ελέγξουμε την μηδενική υπόθεση η οποία μας δείχνει αν τα υπόλοιπα του υποδείγματος είναι λευκός θόρυβος και επομένως δεν περιέχουν χρήσιμες πληροφορίες.

Βήμα τέταρτο: Μεταδιάγνωση Εάν ένα δοκιμαστικό υπόδειγμα δεν απορριφθεί από τον διαγνωστικό έλεγχο δεν σημαίνει ότι μπορεί να γίνει αποδεκτό, αφού μπορεί να υπάρχουν και άλλα υποδείγματα τα οποία ανταποκρίνονται στις απαιτήσεις των βημάτων δύο και τρία. Στο τελικό στάδιο της μεταδιάγνωσης επιλέγεται τελικά το υπόδειγμα το οποίο εμφανίζει την καλύτερη δυνατή προσαρμογή. Για να γίνει η καλύτερη δυνατή επιλογή υπάρχει η δυνατότητα χρησιμοποίησης των (RMS, AIC, BIC) στατιστικών μέτρων (υπό-ενότητα 2.4.1, 3.2.3). Σε κάθε περίπτωση επιλέγεται το υπόδειγμα με την μικρότερη τιμή του στατιστικού βάσει του οποίου γίνεται η σύγκριση. Γενικά το BIC δίνει μεγαλύτερες τιμές σε σύγκριση με το AIC και επιβραβεύει τα οικονομικά υποδείγματα [6].

Η διαδικασία προσαρμογής ενός μοντέλου συνήθως καθοδηγείται από την αρχή της φειδούς, η οποία θεωρεί ότι το καλύτερο μοντέλο είναι αυτό με το

μικρότερο πλήθος παραμέτρων [32]. Έτσι προκύπτουν τα εξής

α) Αναφέρθηκε πως η καταλληλότητα ενός μοντέλου μπορεί να ελεγχθεί με την βοήθεια κάποιων βασικών στατιστικών μέτρων όπως του κριτηρίου πληροφορίας Akaike (AIC) ή το Μπαυζιανό κριτήριο (BIC) [6, 24]. Παρόλο αυτά η επιλογή αυτού είναι εφικτό να γίνει και με την βοήθεια των δειγματικών συναρτήσεων αυτοσυσχέτισης (ACF) καθώς και μερικής αυτοσυσχέτισης (PACF).

β) Οι παράμετροι των μοντέλων μπορούν να εκτιμηθούν με την βοήθεια της μεθόδου των ροπών ή με την μέθοδο των ελαχίστων τετραγώνων. Συγκεκριμένα στην μέθοδο των ελαχίστων τετραγώνων οι άγνωστοι παράμετροι εκτιμώνται από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων. Οι παραπάνω μέθοδοι θεωρούνται προσεγγίσεις της κλασικής μεθόδου εκτίμησης των παραμέτρων, η οποία είναι η μέθοδος της μέγιστης πιθανοφάνειας [24].

4.2.2 Πρόβλεψη με αυτοπαλινδρομούμενα μοντέλα AR(p)

Εάν η χρονοσειρά $x_1, x_2, ..., x_n$ αναλύεται με τον καλύτερο δυνατό τρόπο μέσω της διαδικασίας AR τότε το αντίστοιχο μοντέλο για x_{n+1} θα είναι

$$x_{n+1} = \phi_1 x_n + ... + \phi_p x_{n-p+1} + z_{n+1} \tag{4.1}$$

και η καλύτερη δυνατή πρόβλεψη για ένα χρονικό βήμα μπροστά δίνεται από την σχέση (Εξ. 4.2)

$$x_n(1) = \phi_1 x_n + ... + \phi_p x_{n-p+1} \tag{4.2}$$

Επομένως για k χρονικά βήματα μπροστά η πρόβλεψη θα μπορεί να προσδιοριστεί από τον τύπο (Εξ. 4.3)

$$x_n(k) = \phi_1 x_n(k-1) + ... + \phi_p x_n(k-p) \tag{4.3}$$

όπου κάθε τιμή του $x_n(k)$ είναι γνωστή από προηγούμενη πρόβλεψη ή είναι γνωστή καθώς δίνεται από την χρονοσειρά. Η πρόβλεψη που κάνουμε συνιστάται μέσω του αιτιοκρατικού μέρους του μοντέλου, όπου οι παρατηρήσεις οι οποίες δεν μας είναι γνωστές αντικαθιστώνται από τις προβλέψεις μας [23].

Κατά την διαδικασία πρόβλεψης υπάρχει ένα σφάλμα για τα ίδια χρονικά βήματα που εκφράζεται ως ο γραμμικός συνδυασμός των στοιχείων των λευκού θορύβου αλλά μόνο στις χρονικές στιγμές $n+1, ..., n+k$ [28] και δίνεται από την σχέση (Εξ. 4.4)

$$e_n(k) = \sum_{j=0}^{k-1} b_j z_{n+k-j} \tag{4.4}$$

Έτσι το μοντέλο θα έχει διασπορά

$$\mathrm{Var}[e_n(k)] = \sigma_z^2 \sum_{j=0}^{k-1} b_j^2 \tag{4.5}$$

4.2.3 Πρόβλεψη με μοντέλο κινούμενου μέσου όρου MA(q)

Για να μπορέσει να είναι εφικτή η πρόβλεψη με την βοήθεια του μοντέλου κινούμενου μέσου πρέπει αρχικά να προσδιορίσουμε το κατάλληλο μοντέλο για την χρονοσειρά $x_1, x_2, ..., x_n$ [24], έτσι η επόμενη χρονικά παρατήρηση θα δίνεται ως

$$x_{n+1} = z_{n+1} + \theta_1 z_n + ... + \theta_q z_{n-q+1} \tag{4.6}$$

και επομένως για ένα χρονικό βήμα μετά η καλύτερη δυνατή πρόβλεψη θα είναι

$$x_n(1) = \theta_1 z_n + ... + \theta_q z_{n-q+1} \tag{4.7}$$

. Γενικά για τυχαία k χρονικά βήματα η εκτίμηση δίνεται από την σχέση (Εξ. 4.8)

$$x_n(k) = \theta_k z_n + \theta_{k+1} z_{n-1} + ... + \theta_q z_{n-q+k}, k \leq 0 \tag{4.8}$$

με

$$x_n(k) = 0$$

Τα σφάλματα z_n, z_{n-1}, z_{n-2} που μπορεί να προκύψουν, μπορούν να υπολογιστούν με την βοήθεια των παρατηρήσεων x_n, x_{n-1}, x_{n-2} με την προϋπόθεση ότι οι αρχικές τους τιμές $z_1, z_2, ..., z_q$ είναι μηδέν. Πιο συγκεκριμένα ο υπολογισμός των $z_{q+1}, z_{q+2}, ..., z_q$ γίνεται μέσω της λύσης της εξίσωσης του MA(q), όπου θέτουμε $t = q$ και κάνουμε το ίδιο για τους χρόνους $t = q + 1, ..., n - 1$ [24].

4.2.4 Πρόβλεψη με αυτοπαλινδρομούμενα μοντέλα κινούμενου μέσου όρου ARMA

Το αυτοπαλινδρομούμενο μοντέλο κινούμενου μέσου συμβολίζεται ως ARMA(p, q). Θεωρώντας έτσι μια χρονοσειρά $x_1, x_2, ..., x_n$, τότε το αμέσως επόμενο χρονικά βήμα θα είναι

$$x_{n+1} = \phi_1 x_n + ... + \phi_p x_{n-p+1} + \theta_1 z_n + ... + \theta_q z_{n-q+1} \tag{4.9}$$

Έτσι για k χρονικά βήματα μπροστά ισχύει ότι εάν $k \leq q$ τότε

$$x_n(k) = \phi_1 x_n(k-1) + ... + \phi_p x_n(k-p) + \theta_k z_n + ... + \theta_q z_{n-q+1} \tag{4.10}$$

ενώ αν $k > q$ τότε

$$x_n(k) = \phi_1 x_n(k-1) + ... + \phi_p x_n(k-p) \tag{4.11}$$

Έτσι προκύπτει πως η πρόβλεψη με την βοήθεια του μοντέλου ARMA αποτελεί έναν συνδυασμό των προβλέψεων με τα μοντέλα MA και AR [2].

4.3 Διαγνωστικός έλεγχος

Ένα από τα βασικότερα στάδια στην διαδικασία πρόβλεψης των χρονοσειρών αποτελεί ο έλεγχος που πρέπει να πραγματοποιηθεί προκειμένου να γίνει επιλο-

γή του καλύτερου δυνατού μοντέλου το οποίο θα δίνει την καλύτερη πρόβλεψη. Αυτό γίνεται γιατί ενώ μπορεί να καταλήξουμε σε κάποιο μοντέλο, ένα άλλο ARIMA μοντέλο μπορεί να προσαρμόζεται καλύτερα. Ο διαγνωστικός έλεγχος αφορά τόσο τον στατιστικό έλεγχο για την σημαντικότητα των συντελεστών, τη συμπεριφορά των καταλοίπων όσο και την τάξη του υποδείγματος [24].

Προκύπτει έτσι από τον έλεγχο αυτό ότι αν κάποιο εκτιμώμενο μοντέλο καταφέρνει να εκφράσει ικανοποιητικά την διαδικασία από την οποία προέρχονται τα δεδομένα, τότε τα υπόλοιπα θα πρέπει να συμπεριφέρονται σαν λευκός θόρυβος, που σημαίνει ότι τα υπόλοιπα δεν πρέπει να αυτοσυσχετίζονται. Ο βασικός τρόπος μέσω του οποίου γίνεται ο έλεγχος των καταλοίπων είναι με την βοήθεια των διαγραμμάτων των ACF και PACF [20]. Παρόλο αυτά υπάρχουν και άλλα μέσα με τα οποία μπορεί να γίνει ο έλεγχος αυτός όπως για παράδειγμα μέσω του στατιστικού Q των Ljung και Box (LBQ), μέθοδος που δεν θα αναπτυχθεί στα πλαίσια της εργασίας αυτής.

Ένας ακόμα αποτελεσματικός τρόπος εκτίμησης της καταλληλότητας του μοντέλου είναι μέσω της σύγκρισης του με ένα άλλο υπόδειγμα μεγαλύτερης τάξης. Δηλαδή το υπόδειγμα ARMA(p, q) θα συγκριθεί με το υπόδειγμα ARMA(p+1, q) και ARMA(p, q+1). Έτσι αν το υπόδειγμα που εκτιμήθηκε είναι το κατάλληλο και περιγράφει την διαδικασία που παρήγαγε τα δεδομένα τότε οι επιπλέον συντελεστές που βρίσκονται στα μεγαλύτερα υποδείγματα δεν θα πρέπει να είναι στατιστικά διαφορετικοί από το μηδέν [1].

4.3.1 Κριτήρια επιλογής υποδείγματος

Πραγματοποιώντας αύξηση στην τάξη του υποδείγματος που επιλέχθηκε, δηλαδή προσθέτοντας υστερήσεις είτε για το αυτοπαλινδρομούμενο τμήμα είτε για το τμήμα του κινούμενου μέσου, παρατηρείται μια μείωση στο άθροισμα των τετραγώνων των καταλοίπων, αλλά θα υπάρξει και μια μείωση των βαθμών

ελευθερίας, γιατί εκτιμώνται περισσότερες παράμετροι. Αυτό σημαίνει ότι η προσθήκη μεταβλητών έχει οφέλη αλλά δημιουργεί και προβλήματα.

Για να μπορέσουμε να συγκρίνουμε υποδείγματα με διαφορετικό αριθμό παραμέτρων, πέραν του συντελεστή προσδιορισμού r^2 χρησιμοποιούνται και μερικά ακόμα κριτήρια. Στην ανάλυση χρονοσειρών χρησιμοποιούνται κυρίως το μέσο τετράγωνο των καταλοίπων (Residual Mean Square, RMS), το κριτήριο πληροφοριών Akaike (Akaike information Ctiterion, AIC) [6] και το Μπευσιανό κριτήριο του Schwartz (Schwartz Bayesian Criterion, BIC) [32] (υπόενότητα 3.5).

Κεφάλαιο 5

Ανάλυση δεδομένων παραγωγής ηλεκτρικής ενέργειας

Σκοπός της διπλωματικής εργασίας είναι η πρόβλεψη της παραγωγής ηλεκτρικής ενέργειας για μια ημέρα μετά με δεδομένα που αντιστοιχούν σε λεπτά ή ώρες της μέρας με την βοήθεια των μοντέλων SARIMA και της μεθόδου του εκθετικού σταθμισμένου μέσου όρου (exponational moving average). Επιπλέον πραγματοποιείται χρονική παρεμβολή στο σύνολο των δεδομένων με την βοήθεια της μεθόδου του κινούμενου σταθμισμένου μέσου όρου. Τα δεδομένα προέρχονται από την επίσημη ιστοσελίδα του ομίλου ELIA και πραγματοποιήθηκε σύγκριση των αποτελεσμάτων μεταξύ των μεθόδων που χρησιμοποιήθηκαν.

5.1 Περιοχή μελέτης και περιγραφή των διαθέσιμων δεδομένων

Η περιοχή μελέτης στην οποία απευθύνονται όλες οι αναλύσεις που πραγματοποιούνται στην εργασία αυτή είναι η χώρα του Βελγίου η οποία βρίσκεται στην βορειοδυτική Ευρώπη. Συνορεύει με την Ολλανδία, την Γερμανία, το Λουξεμβούργο και την Γαλλία. Ο πληθυσμός της ανέρχεται στους 11.52 εκατομμύρια κατοίκους και καταλαμβάνει μία έκταση της τάξης των 30.528 τετραγωνικών χιλιομέτρων. Η πρωτεύουσα της είναι οι Βρυξέλλες και είναι επισήμως δίγλωσση (Ολλανδικά, Γαλλικά) ενώ η πλειονότητα των κατοίκων μιλάει γαλλικά. Τέλος η πληθυσμιακή πυκνότητα της (371,9 κατ. ανά χλμ^2) είναι μία από τις μεγαλύτερες στην Ευρώπη.

Το σύνολο των δεδομένων αποκτήθηκε από τον όμιλο ELIA και αποτελείται από 26208 μετρήσεις από την 1η Ιανουαρίου του 2019 έως τις 30 Σεπτεμβρίου του 2019. Τα στοιχεία αυτά είναι διαχωρισμένα ανά 15 λεπτά της ώρας. Για την χρονοσειρά με χρονικό βήμα ανά 15 λεπτά επιλέγονται δύο χρονικές περίοδοι. Μία για την καλοκαιρινή περίοδο από 15 Ιουνίου έως 13 Ιουλίου και μία για την χειμερινή από 15 Ιανουαρίου έως 12 Φεβρουαρίου. Αυτό γίνεται για να αντιμετωπιστεί ο μεγάλος όγκος των δεδομένων της χρονοσειράς με χρονικό βήμα ανά 15 λεπτά αλλά και για να γίνει εφικτή η σύγκριση των αποτελεσμάτων της πρόβλεψης στις δύο περιόδους. Αναμένουμε διαφορετικά αποτελέσματα λόγω των διαφορετικών ενεργειακών αναγκών που επικρατούν στις δύο αυτές χρονικές περιόδους που χρησιμοποιήθηκαν. Η διαδικασία της πρόβλεψης θα γίνει για την χρονοσειρά με χρονικό βήμα ανά 15 λεπτά και για τις δύο χρονικές περιόδους.

Επιπρόσθετα γνωρίζουμε πως η μέρα αποτελείται από τέσσερα εξάωρα και κάθε εξάωρο περιλαμβάνει 24 τέταρτα της ώρας. Επιλέγεται μετασχηματισμός της χρονοσειράς από χρονικό βήμα ανά 15 λεπτά σε εξάωρα, με μέσους όρους

των 24 τετάρτων για κάθε ένα από τα εξάωρα αυτά. Η διαδικασία της πρόβλεψης θα εφαρμοστεί και στην χρονοσειρά με χρονικό βήμα ανά εξάωρο.

Η μετατροπή των δεδομένων από χρονικό βήμα ανά 15 λεπτά της ώρας σε μέσους όρους ανά εξάωρα έγινε για να μειωθεί ο όγκος των δεδομένων. Επιπλέον παρατηρήθηκε πως η χρονοσειρά με χρονικό βήμα ανά 15 λεπτά αποκλίνει από την κανονική κατανομή. Η κατανομή των δεδομένων της χρονοσειράς με μέσους όρους ανά εξάωρα πλησιάζει περισσότερο την κανονική (ενότητα 5.2).

Για την αξιολόγηση της μεθοδολογίας που ακολουθήθηκε με τα μοντέλα SARIMA και με τον σταθμισμένο μέσο όρο πριν την εκτέλεση της τελικής πρόβλεψης, δημιουργείται μία χρονοσειρά προσομοίωσης. Η δημιουργία της χρονοσειράς και η πρόβλεψη μελλοντικών χρονικών στιγμών με αυτήν αποσκοπεί στην αξιολόγηση εν τέλει της μεθοδολογίας που θα ακολουθηθεί για τον προσδιορισμό των εκτιμήσεων στην εργασία αυτή.

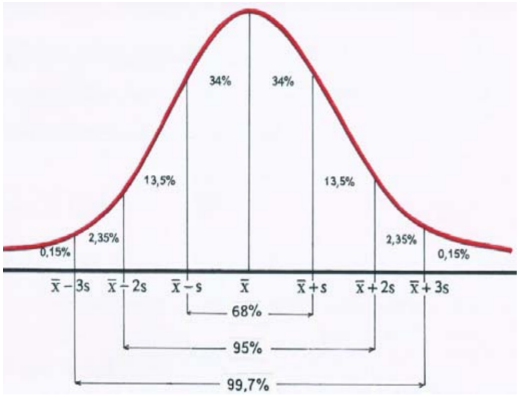
5.2 Στατιστική ανάλυση δεδομένων

Πριν την εφαρμογή των μεθόδων που χρησιμοποιήθηκαν για την δημιουργία πρόβλεψης, απαιτείται η στατιστική ανάλυση των δεδομένων που διαθέτουμε. Η ανάλυση αυτή γίνεται με την βοήθεια στατιστικών μέτρων, ιστογραμμάτων κ.α.

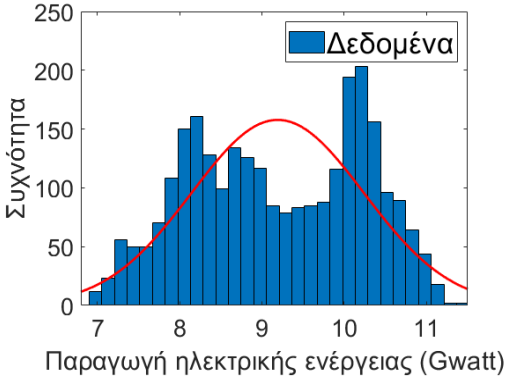
Στατιστική ανάλυση χρονοσειράς με χρονικό βήμα ανά 15 λεπτά για την καλοκαιρινή περίοδο

Με την βοήθεια του ιστογράμματος και του διαγράμματος κανονικής πιθανότητας (Σχ. 5.1) για την χρονοσειρά με χρονικό βήμα ανά 15 λεπτά για την περίοδο του καλοκαιριού μπορούμε να καταλήξουμε σε χρήσιμα συμπεράσματα σχετικά με την κατανομή των στοιχείων. Η μορφή της κανονικής κατανομής παρουσιάζεται στο σχήμα (Σχ. 5.1α΄). Το διάγραμμα κανονικής πιθανότητας δείχνει

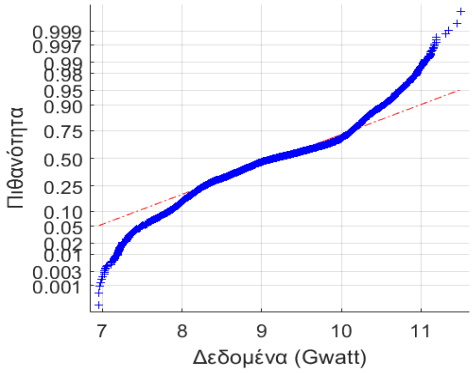
την κατανομή των δεδομένων σε σύγκριση με την κανονική κατανομή για την χρονοσειρά με χρονικό βήμα ανά 15 λεπτά της ώρας για την καλοκαιρινή περίοδο.



(α′) Διάγραμμα κανονικής πιθανότητας



(β′) Ιστόγραμμα χρονοσειράς ανά 15 λεπτά για την καλοκαιρινή περίοδο



(γ′) Διάγραμμα κανονικής πιθανότητας για χρονοσειρά ανά 15 λεπτά για την καλοκαιρινή περίοδο

Σχήμα 5.1: Ανάλυση κανονικότητας (ιστόγραμμα, διάγραμμα κανονικής πιθανότητας) της χρονοσειράς ανά 15 λεπτά για την καλοκαιρινή περίοδο

.

Μέσω των διαγραμμάτων που παρουσιάστηκαν παραπάνω γίνεται κατανοητό πως οι διαβαθμίσεις που παρουσιάζει το ιστόγραμμα της χρονοσειράς ανά 15 λεπτά μας οδηγούν στο συμπέρασμα της απόκλισης της χρονοσειράς από την κανονική κατανομή. Επιπλέον οι δύο κορυφές που εμφανίζονται στο ιστόγραμμα για την καλοκαιρινή περίοδο οφείλονται στις διαφορετικές ανάγκες για κατανάλωση ενέργειας μέσα σε 24 ώρες. Η μία κορυφή αντιστοιχεί στην μικρότερη ζήτηση για ενέργεια την νύχτα, ενώ η άλλη στην μεγαλύτερη ζήτηση την ημέρα. Όσον αφορά το διάγραμμα κανονικής πιθανότητας είναι εμφανής η μη γραμμικότητα των δεδομένων και κατ΄ επέκταση η απόκλιση που παρουσιάζουν αυτά από την κανονική κατανομή.

Στο τελευταίο κομμάτι της στατιστικής ανάλυσης της χρονοσειράς με χρονικό βήμα ανά 15 λεπτά για την καλοκαιρινή περίοδο παραθέτονται χαρακτηριστικά στατιστικά μέτρα στον πίνακα (Πιν. 5.1).

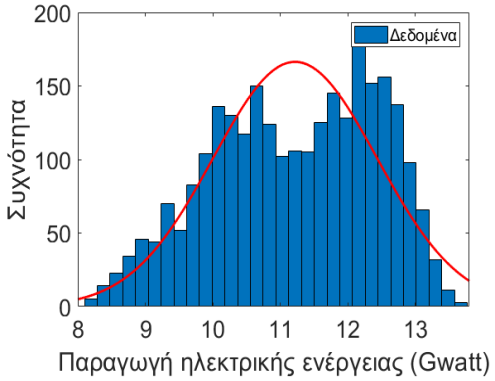
Στατιστικά Μέτρα	**Τιμές**
Μέση τιμή	9.19×10^3
Διασπορά	1.09×10^6
Κύρτωση	1.84
Ελάχιστη Τιμή	6.95×10^3
Μέγιστη τιμή	1.14×10^4
Εύρος τιμών	4.54×10^3

Πίνακας 5.1: Πίνακας στατιστικών μέτρων για την χρονοσειρά ανά 15 λεπτά για την καλοκαιρινή περίοδο

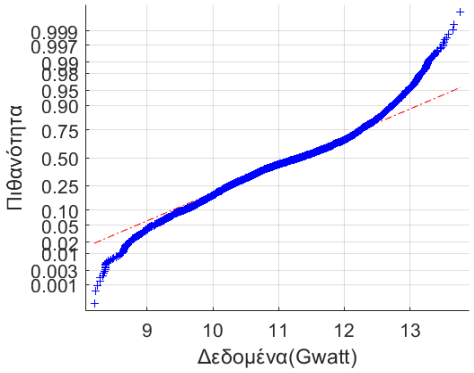
Από την παρατήρηση του πίνακα αυτού συμπεραίνουμε πως η τιμή της κύρτωσης δείχνει την απόκλιση της χρονοσειράς από την κανονική κατανομή.

Στατιστική ανάλυση χρονοσειράς με χρονικό βήμα ανά 15 λεπτά για την χειμερινή περίοδο

Με την βοήθεια του ιστογράμματος και του διαγράμματος κανονικής πιθανότητας (Σχ. 5.2) για την χρονοσειρά με χρονικό βήμα ανά 15 λεπτά για την περίοδο του χειμώνα μπορούμε να καταλήξουμε σε χρήσιμα συμπεράσματα σχετικά με την κατανομή των στοιχείων. Το διάγραμμα κανονικής πιθανότητας δείχνει την κατανομή των δεδομένων σε σύγκριση με την κανονική κατανομή για την χρονοσειρά με χρονικό βήμα ανά 15 λεπτά της ώρας.



(α′) Ιστόγραμμα χρονοσειράς ανά 15 λεπτά για την χειμερινή περίοδο



(β′) Διάγραμμα κανονικής πιθανότητας για χρονοσειρά ανά 15 λεπτά για την χειμερινή περίοδο

Σχήμα 5.2: Ανάλυση κανονικότητας (ιστόγραμμα, διάγραμμα κανονικής πιθανότητας) της χρονοσειράς ανά 15 λεπτά για την χειμερινή περίοδο.

Μέσω των διαγραμμάτων που παρουσιάστηκαν παραπάνω γίνεται κατανοητό πως οι διαβαθμίσεις που παρουσιάζει το ιστόγραμμα της χρονοσειράς ανά 15 λεπτά μας οδηγούν στο συμπέρασμα της απόκλισης της χρονοσειράς από την κανονική κατανομή. Επιπλέον οι δύο κορυφές που εμφανίζονται στο ιστόγραμμα για την καλοκαιρινή περίοδο οφείλονται στις διαφορετικές ανάγκες για κατανάλωση ενέργειας μέσα σε 24 ώρες. Η μία κορυφή αντιστοιχεί στην μικρότερη

ζήτηση για ενέργεια την νύχτα, ενώ η άλλη στην μεγαλύτερη ζήτηση την ημέρα. Όσον αφορά το διάγραμμα κανονικής πιθανότητας είναι εμφανής η μη γραμμικότητα των δεδομένων και κατ΄ επέκταση η απόκλιση που παρουσιάζουν αυτά από την κανονική κατανομή.

Στο τελευταίο κομμάτι της στατιστικής ανάλυσης της χρονοσειράς με χρονικό βήμα ανά 15 λεπτά για την χειμερινή περίοδο παραθέτονται κάποια χαρακτηριστικά στατιστικά μέτρα στον πίνακα (Πιν. 5.2). Η τιμή της κύρτωσης από την παρατήρηση του πίνακα αυτού δηλώνει την απόκλιση του συνόλου των δεδομένων από την κανονική κατανομή.

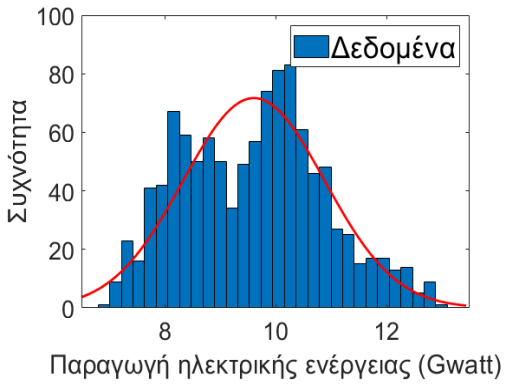
Στατιστικά Μέτρα	**Τιμές**
Μέση τιμή	1.12×10^4
Διασπορά	1.47×10^6
Κύρτωση	2.11
Ελάχιστη Τιμή	8.19×10^3
Μέγιστη τιμή	1.37×10^4
Εύρος τιμών	5.57×10^3

Πίνακας 5.2: Πίνακας στατιστικών μέτρων για την χρονοσειρά ανά 15 λεπτά για την χειμερινή περίοδο

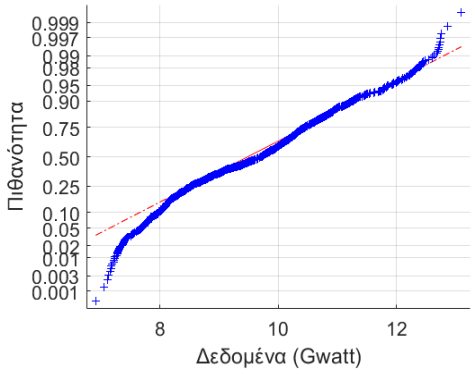
Στατιστική ανάλυση χρονοσειράς με χρονικό βήμα εξάωρο (μέσοι όροι ανά τέταρτα)

Για την αντιμετώπιση της κατανομής που αποκλίνει από την κανονική αλλά και για να μειώσουμε το μεγάλο μέγεθος της χρονοσειράς έγινε η μετατροπή αυτής σε μέσους όρους ανά εξάωρα της μέρας (υπό-εν. 5.2). Έτσι τα 26208 δεδομένα μειώνονται σε 1092. Όπως φαίνεται τόσο στο ιστόγραμμα όσο και στο διάγραμμα κανονικής πιθανότητας (Σχ. 5.3) για την χρονοσειρά μετά την

μετατροπή της σε μέσους όρους ανά εξάωρα, είναι εμφανές πως τα δεδομένα μας πλησιάζουν περισσότερο την κανονική κατανομή σε σύγκριση με το σύνολο των δεδομένων της χρονοσειράς ανά 15 λεπτά. Αυτό το συμπέρασμα προκύπτει λόγω των μικρότερων διαβαθμίσεων των τιμών στο ιστόγραμμα αλλά και από την μελέτη του διαγράμματος κανονικής πιθανότητας (Σχ. 5.3). Η ανάλυση αυτού μας οδηγεί στο συμπέρασμα της ομαλοποίησης των διακυμάνσεων των δεδομένων σε σχέση με τον χρόνο και της ύπαρξης κανονικής πλέον κατανομής. Όπως έχει αναφερθεί και προηγούμενες η παρουσία κανονικής κατανομής είναι απαραίτητο στοιχείο για την ορθή και εμπεριστατωμένη πρόβλεψη μελλοντικών χρονικών στιγμών.



(α′) Ιστόγραμμα χρονοσειράς ανά έξι ώρες



(β′) Διάγραμμα κανονικής πιθανότητας για χρονοσειρά με μέσους όρους εξάωρων

Σχήμα 5.3: Ανάλυση κανονικότητας (ιστόγραμμα, διάγραμμα κανονικής πιθανότητας) της χρονοσειράς με μέσους όρους ανά εξάωρα.

Τέλος παρουσιάζονται ορισμένα στατιστικά μέτρα για την χρονοσειρά με μέσους όρους ανά εξάωρα στον πίνακα (Πιν. 5.3). Η τιμή της κύρτωσης από την παρατήρηση του πίνακα αυτού δηλώνει πως το σύνολο των δεδομένων πλησιάζει περισσότερο την κανονική κατανομή στην χρονοσειρά ανά εξάωρο.

Στατιστικά Μέτρα	**Τιμές**
Μέση τιμή	9.60×10^3
Διασπορά	1.62×10^6
Κύρτωση	2.42
Ελάχιστη Τιμή	6.90×10^3
Μέγιστη τιμή	13.02×10^3
Εύρος τιμών	6.12×10^3

Πίνακας 5.3: Πίνακας στατιστικών μέτρων για την χρονοσειρά με μέσους όρους ανά εξάωρο

5.3 Προσομοίωση χρονοσειράς για κατασκευή συνθετικών δεδομένων

Πριν την εφαρμογή μίας μεθοδολογίας πάνω στο διαθέσιμο σύνολο δεδομένων, συνιστάται η δημιουργία μίας χρονοσειράς ως διαγνωστικό έλεγχο για την αποτελεσματικότητα της μεθοδολογίας αυτής. Στην προκειμένη περίπτωση θα γίνει εφαρμογή ενός μοντέλου SARIMA και της μεθόδου του EWMA πάνω στην διαθέσιμη χρονοσειρά. Οι συντελεστές για την δημιουργία αυτής παραθέτονται από τον χρήστη. Η χρονοσειρά θα είναι στάσιμη και έτσι θα μπορούμε να έχουμε μια αρχική εικόνα σχετικά με την αποτελεσματικότητα του κώδικα για την πρόβλεψη που θέλουμε να επιτελέσουμε. Τα συμπεράσματα αυτά θα βγουν με την βοήθεια των μέτρων επιβεβαίωσης καθώς και μέσω διαγραμμάτων τα οποία παρουσιάζουν το εύρος των τιμών των συντελεστών του μοντέλου που προσδιορίστηκαν μετά την πραγματοποίηση της πρόβλεψης.

Προσομοίωση χρονοσειράς για το μοντέλο SARIMA

Όσον αφορά το μοντέλο SARIMA η χρονοσειρά που δημιουργήθηκε αποτελείται από 3552 δεδομένα χωρισμένα ανά 15 λεπτά της ώρας. Για την δημιουργία της πρόβλεψης εκπαιδεύουμε την χρονοσειρά με το στοιχείο που αντιστοιχεί στην χρονική στιγμή t_1 έως το στοιχείο την χρονική στιγμή $t_{2687+i+t_1}$. Το i εκφράζει τον αριθμό των προβλέψεων, $i = 1, ..., 672$ που αντιστοιχεί σε μία βδομάδα. Στην συνέχεια γίνεται εκτίμηση της παραγωγής ηλεκτρικής ενέργειας για την χρονική στιγμή $t_{2687+i+96+t_1}$ για μία μέρα μέρα μετά (αντιστοιχεί σε 96 τέταρτα). Επιπλέον εκτιμάται η χρονική στιγμή $t_{2687+i+48+t_1}$ για 12 ώρες μετά και η $t_{2688+i+192+t_1}$ για δύο μέρες μετά.

Μέτρα επιβεβαίωσης	**Τιμές**
RMSE (MWatt) (12 ώρες μετά)	187
RMSE (MWatt) (μία μέρα μετά)	126.95
RMSE (MWatt) (δύο μέρες μετά)	122.62
ME (MWatt) (12 ώρες μετά)	11.19
ME (MWatt) (μία μέρα μετά)	14.77
ME (MWatt) (δύο μέρες μετά)	53.97
RPe (12 ώρες μετά)	0.77
RPe (μία μέρα μετά)	0.76
RPe (δύο μέρες μετά)	0.75

Πίνακας 5.4: Μέτρα επιβεβαίωσης για την χρονοσειρά προσομοίωσης ανά 15 λεπτά της ώρας. Το RMSE εκφράζει το μέσο τετραγωνικό σφάλμα. Το ME εκφράζει το μέσο σφάλμα και το RPe τον συντελεστή συσχέτισης του Pearson. Και οι τρεις συντελεστές χρησιμοποιήθηκαν για την αξιολόγηση των προβλέψεων για 12 ώρες μετά, για μία μέρα μετά και για δύο μέρες μετά.

Από την παρατήρηση του πίνακα (Πιν. 5.4) καταλήγουμε στο συμπέρασμα πως η εφαρμογή του κώδικα με τον οποίο θα κάνουμε εν τέλει την πρόβλεψη

σε μία στάσιμη χρονοσειρά γνωστών συντελεστών είναι ικανοποιητική.

Συγκεκριμένα από την παρατήρηση του μέτρου επιβεβαίωσης του συντελεστή του Pearson καταλήγουμε πως δίνει ικανοποιητικά αποτελέσματα αφού κυμαίνεται από 75%–77%. Εκφράζει επομένως ισχυρή συσχέτιση μεταξύ των γνωστών και των προβλέψιμων δεδομένων. Παρόλο αυτά όπως θα δούμε και στην συνέχεια επηρεάζεται από την παρουσία ιδιόμορφων τιμών (outliers) και αυτοσυσχετίσεων. Επομένως η χρήση αυτού δεν είναι αντιπροσωπευτική και δεν θα ληφθεί ιδιαίτερα υπόψιν στην συνέχεια της εκτίμησης των αποτελεσμάτων. Η παρατήρηση του RMSE μπορεί να οδηγήσει σε αξιόπιστα συμπεράσματα. Οι τιμές του μέτρου αυτού επιβεβαίωσης είναι ικανοποιητικές αφού το ποσοστό του μέσου τετραγωνικού σφάλματος κυμαίνεται από 1,98% (για δύο μέρες στο μέλλον) έως 3,03% (για 12 ώρες στο μέλλον) για εύρος 6166 (MWatt). Επομένως καταλήγουμε σε μικρά σφάλματα μεταξύ των πραγματικών δεδομένων και αυτών που προσδιορίστηκαν.

Για την δημιουργία της χρονοσειράς δώσαμε κάποιους συντελεστές στο μοντέλο μέσω του λογισμικού της Matlab και έγινε εκτίμηση των αντίστοιχων συντελεστών από την εκτέλεση της διαδικασίας πρόβλεψης. Οι συντελεστές που δόθηκαν αρχικά για το μοντέλο SARIMA φαίνονται στον πίνακα (Πιν. 5.5).

SAR	−0.71	0.03	−0.16	0.09	0.24	0.05	0.03	0.08	
MA	−1.50	−0.16	0.98	−0.47	0.13	0.16	−0.07	−0.19	0.12

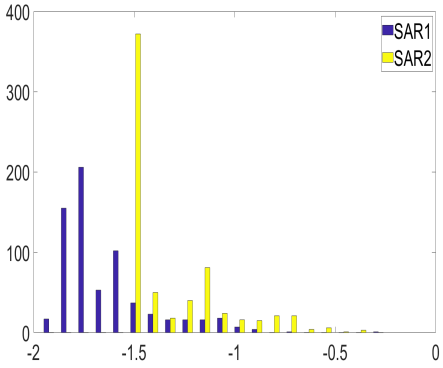
Πίνακας 5.5: Γνωστοί συντελεστές για την δημιουργία της χρονοσειράς.

Μετά την εφαρμογή της διαδικασίας πρόβλεψης εκτιμήθηκαν οι συντελεστές του μοντέλου SARIMA για κάθε προβλεπόμενη τιμή. Ενδεικτικές τιμές των συντελεστών φαίνονται στον πίνακα (Πιν. 5.6).

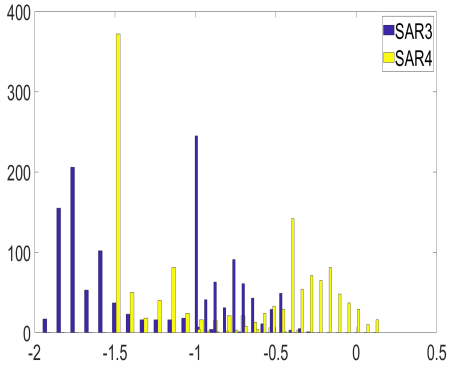
SAR	−1.62	−1.42	−0.99	−0.52	−0.5	−0.78	−0.27		
MA	−0.55	−0.77	−0.07	0.08	0.52	0.19	−0.96	0.31	0.25

Πίνακας 5.6: Πίνακας συντελεστών που προέκυψαν από την διαδικασία πρόβλεψης για χρονοσειρά ανά 15 λεπτά με το μοντέλο SARIMA. Ενδεικτικά μεγέθη από το τελευταίο μοντέλο που δημιουργήθηκε κατά την επιτέλεση της πρόβλεψης.

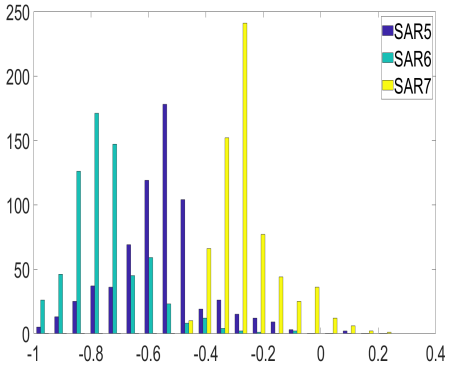
Έτσι προέκυψε το μοντέλο SARIMA(0, 2, 9)(105, 2, 9) με περιοδικότητα 96 που αντιστοιχεί σε ημερήσια περιοδικότητα. Το εύρος των συντελεστών που προέκυψαν φαίνεται στα σχήματα (σχ. 5.4, σχ. 5.5).



(α′) Συντελεστές (SAR1, SAR2)



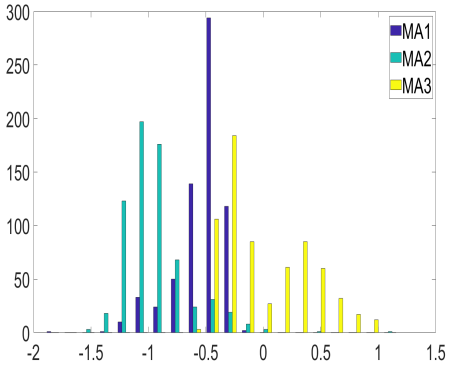
(β′) Συντελεστές (SAR3, SAR4)



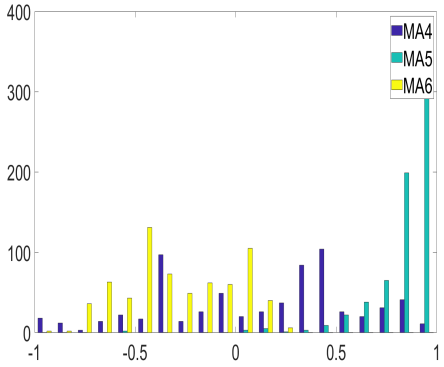
(γ′) Συντελεστές (SAR5, SAR6, SAR7)

Σχήμα 5.4: Συντελεστές του εποχιακού αυτοπαλινδρομούμενου συντελεστή που προσδιορίστηκαν μετά την διαδικασία της πρόβλεψης για χρονοσειρά ανά 15 λεπτά με το μοντέλο SARIMA

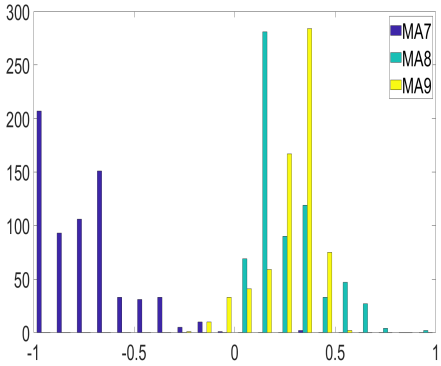
.



(α′) Συντελεστές (MA1, MA2, MA3)



(β′) Συντελεστές (MA4, MA5,MA6)

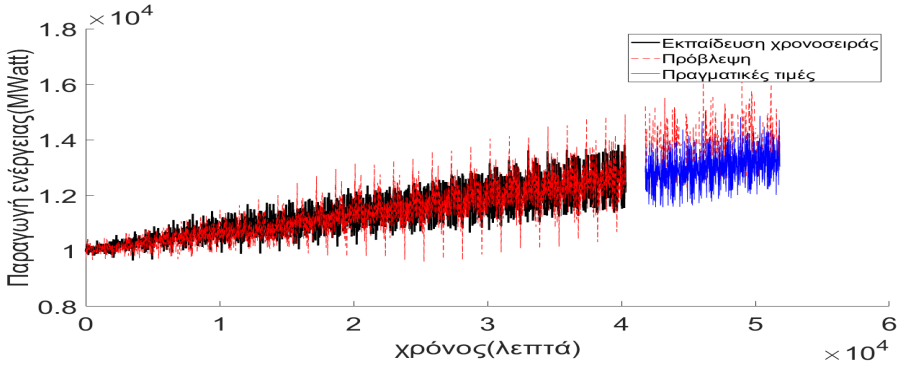


(γ′) Συντελεστές (MA7, MA8, MA9)

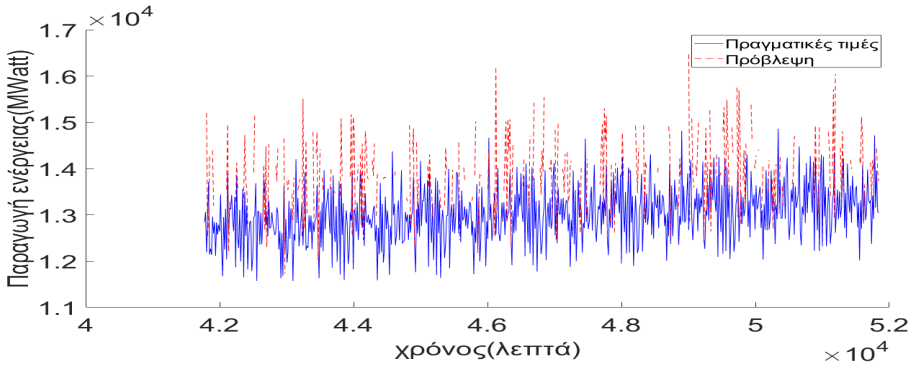
Σχήμα 5.5: Συντελεστές του κινούμενου μέσου όρου που προσδιορίστηκαν μετά την διαδικασία της πρόβλεψης για χρονοσειρά ανά 15 λεπτά με το μοντέλο SARIMA

.

Τέλος στο διάγραμμα (σχ. 5.6) παρουσιάζονται τα πραγματικά δεδομένα της χρονοσειράς προσομοίωσης και αυτά που προσδιορίστηκαν από την εκτέλεση της πρόβλεψης.



(α′) Διάγραμμα εκπαίδευσης χρονοσειράς, πραγματικών τιμών και πρόβλεψης



(β′) Διάγραμμα πραγματικών τιμών και πρόβλεψης

Σχήμα 5.6: Διάγραμμα γνωστών δεδομένων και πρόβλεψης για την χρονοσειρά προσομοίωσης με χρονικό βήμα ανά 15 λεπτά. Το κομμάτι εκπαίδευσης της χρονοσειράς φαίνεται με μαύρη συνεχή γραμμή. Το κομμάτι τις πρόβλεψης φαίνεται με κόκκινη διακεκομμένη. Οι πραγματικές τιμές για την περίοδο της πρόβλεψης φαίνονται με μπλε συνεχή γραμμή

Από την παρατήρηση των πινάκων (πιν. 5.5, πιν. 5.6) μπορούμε να οδηγηθούμε σε χρήσιμα συμπεράσματα. Οι πίνακες αυτοί απεικονίζουν τους συντελεστές του μοντέλου SARIMA που δόθηκαν από τον χρήστη για την δη-

μιουργία της χρονοσειράς καθώς και αυτούς που προέκυψαν από την διαδικασία της πρόβλεψης. Μέσω των πινάκων αυτών οδηγούμαστε στο συμπέρασμα της απόκλισης των συντελεστών που προσδιορίστηκαν σε σχέση με αυτούς που δόθηκαν αρχικά. Το συμπέρασμα αυτό μπορεί να προκύψει και μέσω των ιστογραμμάτων (σχ. 5.4, σχ. 5.5). Παρατηρούμε πως το εύρος των τιμών των συντελεστών που προέκυψαν δεν πλησιάζει την αρχική δοθείσα τιμή. Επιπλέον παρατηρείται μία απόκλιση μεταξύ των πραγματικών τιμών και των αντίστοιχων τιμών που προέκυψαν από την διαδικασία της πρόβλεψης στο σχήμα (σχ. 5.6). Συγκεκριμένα στο σχήμα (σχ. 5.6β′) υπάρχουν κάποιες τιμές πρόβλεψης που παρουσιάζουν μεγαλύτερη εκτίμηση σε σχέση με το σύνολο των δεδομένων. Η ύπαρξη των στοιχείων αυτών δηλώνει την παρουσία θετικής αμεροληψίας (bias). Η αμεροληψία αυτή δεν είναι μεγάλη γιατί οι εκτιμήσεις αυτές είναι αριθμητικά λίγες σε σύγκριση με το σύνολο των δεδομένων της πρόβλεψης. Επομένως δεν επηρεάζουν σε μεγάλο βαθμό την τιμή του μέσου σφάλματος.

Η απόκλιση των συντελεστών του μοντέλου καθώς και των πραγματικών τιμών και των αντίστοιχων της πρόβλεψης οφείλεται πιθανότατα στο ίδιο το μοντέλο SARIMA. Συγκεκριμένα αυτό προκύπτει από την δυνατότητα του μοντέλου SARIMA να εκτιμήσει με ακρίβεια τους συντελεστές p, q μόνο για χρονικές στιγμές συγκρίσιμες των συντελεστών αυτών. Επιπλέον το γεγονός ότι εισαγάγουμε ένα μοντέλο με τάξη $p = 8$ και $q = 9$ και η διαδικασία της πρόβλεψης σε μερικές περίπτωσης μας επιστρέφει ένα μοντέλο τάξης $p = 7$ και $q = 9$ προσδίδει ποσοστό σφάλματος στις εκτιμήσεις.

Προσομοίωση χρονοσειράς για την μέθοδο EWMA

Για την μέθοδο EWMA η χρονοσειρά που δημιουργήθηκε αποτελούνταν από 3552 στοιχεία με χρονικό βήμα ανά 15 λεπτά της ώρας. Εκτιμήθηκε η παραγωγή της κάθε μέρας με βάση τον σταθμισμένο μέσο όρο της παραγωγής μίας περιόδου εύρους 192 λεπτών νωρίτερα. Τα μέτρα επιβεβαίωσης τα οποία

προέκυψαν από τις προβλέψεις αυτές φαίνονται στον παρακάτω πίνακα.

Μέτρα επιβεβαίωσης	**Τιμές**
RMSE (MWatt)	403.75
RPe	0.27
ME (MWatt)	−1.22

Πίνακας 5.7: Μέτρα επιβεβαίωσης για πρόβλεψη για την χρονοσειρά προσομοίωσης με την μέθοδο EWMA . Το RMSE είναι το μέσο τετραγωνικό σφάλμα για την πρόβλεψη για μία μέρα μετά. Το RPe είναι ο συντελεστής αυτοσυσχέτισης μεταξύ των δεδομένων πρόβλεψης για μία μέρα μετά και των ήδη γνωστών δεδομένων. Το ME είναι το μέσο σφάλμα για πρόβλεψη για μία μέρα μετά.

Η παρατήρηση του πίνακα (Πιν. 5.7) οδηγεί στο συμπέρασμα πως, με εξαίρεση τον συντελεστή συσχέτισης ο οποίος επηρεάζεται από την παρουσία ιδιόμορφων τιμών (outlires) και αυτοσυσχετίσεων, οι υπόλοιποι δύο συντελεστές δίνουν αποτελέσματα τα οποία περιέχουν μικρά ποσοστά σφαλμάτων. Η τιμή του RMSE για παράδειγμα δείχνει την αξιοπιστία της πρόβλεψης. Το ποσοστά της ρίζας του μέσου τετραγωνικού σφάλματος είναι μικρό και είναι ίσο με 6,54% (για μία μέρα μετά) για εύρους 6166 (MWatt). Τέλος η τιμή του ME δηλώνει την απουσία αμεροληψίας (bias). Έτσι η χρονοσειρά της προσομοίωσης προσαρμόζεται ικανοποιητικά και στην περίπτωση του σταθμισμένου κινούμενου μέσου.

5.4 Χρονική παρεμβολή και πρόβλεψη με την βοήθεια του EWMA

Το μοντέλο του σταθμισμένου κινούμενου μέσου όρου αποτελεί μια σχετικά απλή διαδικασία. Μπορεί όμως να φανεί εξαιρετικά χρήσιμη όταν χρησιμοποιείται

για να πραγματοποιήσει χρονική παρεμβολή σε ένα σύνολο δεδομένων. Έκτος της εφαρμογής της μεθόδου για κάλυψη κενών (gap filling) είναι εφικτή η χρησιμοποίηση αυτής και για την δημιουργία πρόβλεψης.

Έτσι αρχικά εφαρμόζεται χρονική παρεμβολή στο σύνολο των δεδομένων για την κάλυψη τυχόν κενών θέσεων. Στην συνέχεια το σύνολο αυτό χωρίς την ύπαρξη πλέον κενών θα χρησιμοποιηθεί για την δημιουργία πρόβλεψης τόσο για την καλοκαιρινή περίοδο όσο και για την χειμερινή με την χρονοσειρά με χρονικό βήμα ανά 15 λεπτά αλλά και για την αντίστοιχη με χρονικό βήμα εξάωρο. Η εκτίμηση αυτή γίνεται για να εξακριβώσουμε την αποτελεσματικότητα της μεθόδου σε εφαρμογές πρόβλεψης αλλά και για να γίνει εφικτή η σύγκριση των αποτελεσμάτων μεταξύ της μεθόδου του σταθμισμένου κινούμενου μέσου όρου και των αντίστοιχων από την πρόβλεψη με μοντέλο SARIMA.

5.4.1 Χρονική παρεμβολή με EWMA

Ο σταθμισμένος κινούμενος μέσος όρος χρησιμοποιείται συχνά για την αντικατάσταση ήδη υπάρχοντων κενών στην χρονοσειρά. Ο τρόπος που χρησιμοποιείται η μέθοδος καθορίζει και τον τρόπο που θα εφαρμοστεί αυτή στα δεδομένα που μας ενδιαφέρουν. Η μέθοδος EWMA υπολογίζει τους σταθμισμένους όρους των δεδομένων της χρονοσειράς που βρίσκονται πριν την κενή θέση με την ιδιαιτερότητα ότι αυτά πρέπει να αθροίζουν στο ένα. Έπειτα για να προσδιορίσει την κενή θέση αθροίζει τα δεδομένα αυτά πολλαπλασιάζοντάς τα με τους αντίστοιχους σταθμισμένους όρους. Υπάρχουν πολύ τρόποι που μπορεί να εφαρμοστεί η μέθοδος του σταθμισμένου κινούμενου μέσου όρου.

Το διαθέσιμο σύνολο των δεδομένων μας περιλαμβάνει επτά συνεχή κενά. Για την εκτίμηση των στοιχείων αυτών εφαρμόστηκαν δύο διαφορετικές μεθοδολογίες του σταθμισμένο κινούμενου μέσου όρου σε τέσσερα τυχαία κομμάτια της χρονοσειράς τα όποια αποτελούνται από εννιά στοιχεία το κάθε ένα. Τα

δεδομένα αυτά είναι γνωστά αλλά αντικαθίστανται με κενά προκειμένου να γίνει η επανεκτίμηση τους με την βοήθεια του EWMA και στην συνέχεια η σύγκριση με τις πραγματικές τους τιμές. Έτσι θα αποφασιστεί ποία από τις δύο μεθοδολογίες είναι η αποτελεσματικότερη.

Πρώτη μεθοδολογία χρονικής παρεμβολής με EWMA

Η πρώτη μεθοδολογία κάνει κάλυψη κενών (gap filling) ανάλογα με το εύρος παραθύρου που έχει οριστεί (3, 8, 10, 15, 20). Υπολογίζει έναν αριθμό σταθμισμένων όρων αντίστοιχο του εύρους αυτού, δηλαδή από 3 έως 20 όρους. Στην συνέχεια σε συνάρτηση με τα στοιχεία που έχουν οριστεί τυχαία ως κενά η μεθοδολογία αντικαθιστά κάθε ένα από τα δεδομένα αυτά. Η κάλυψη αυτή γίνεται με το άθροισμα των στοιχείων που βρίσκονται μια χρονική στιγμή t που αντιστοιχεί σε μία μέρα πριν από το τυχαίο κενό στοιχείο έως και t αφαιρώντας όμως το εύρος παραθύρου που επιλέχθηκε. Το άθροισμα αυτό πολλαπλασιάζεται με τους αντίστοιχους σταθμισμένους όρους (3.6).

Δεύτερη μεθοδολογία χρονικής παρεμβολής με EWMA

Η δεύτερη μεθοδολογία προσδιορίζει τους σταθμισμένους όρους με τον ίδιο ακριβώς τρόπο όπως και η πρώτη. Σε αυτήν την περίπτωση τα στοιχεία που έχουν θεωρηθεί τυχαία ως κενά αντικαθίστανται με την τιμή που προκύπτει από το άθροισμα των δεδομένων που βρίσκονται μια χρονική στιγμή t_1 που αντιστοιχεί σε δύο ώρες πριν έως και t_1 αφαιρώντας όμως το εύρος παραθύρου που επιλέχθηκε. Το άθροισμα αυτό πολλαπλασιάζεται με τους αντίστοιχους συντελεστές στάθμισης. Οι πίνακες που ακολουθούν δείχνουν την αποτελεσματικότητα των δύο μεθόδων με την βοήθεια των μέτρων επιβεβαίωσης. Η αξιολόγηση γίνεται και για τις δύο περιόδους που έχουν επιλέγει με σύνολο 28 ημερών (για την καλοκαιρινή και την χειμερινή περίοδο).

N_b	**3**	**5**	**8**	**10**	**15**	**20**
RMSE (MWatt)	619.14	631.21	658.40	641.17	661.50	673.95
RPe	0.39	0.37	0.31	0.33	0.26	0.21
ME (MWatt)	7.05	2.88	−6.57	17.12	10.96	−1.99

Πίνακας 5.8: Πίνακας μέτρων επιβεβαίωσης για την χειμερινή περίοδο με εφαρμογή της πρώτης μεθοδολογίας. Το RMSE είναι το μέσο τετραγωνικό σφάλμα. Το RPe είναι ο συντελεστής συσχέτισης του Pearson. Το ME είναι το μέσο σφάλμα. Και τα τρία μέτρα προσδιορίζουν τα ποσοστά σφάλματος μεταξύ των γνωστών δεδομένων που έγιναν κενά και των δεδομένων που προσδιορίστηκαν για την κάλυψη των κενών αυτών.

N_b	**3**	**5**	**8**	**10**	**15**	**20**
RMSE (MWatt)	552.01	563.48	584.20	597.12	604.82	604.57
RPe	0.53	0.49	0.46	0.42	0.39	0.36
ME (MWatt)	−10.73	−0.83	−0.35	2.15	−1.88	6.26

Πίνακας 5.9: Πίνακας μέτρων επιβεβαίωσης για την χειμερινή περίοδο με εφαρμογή της δεύτερης μεθοδολογίας. Το RMSE είναι το μέσο τετραγωνικό σφάλμα. Το RPe είναι ο συντελεστής συσχέτισης του Pearson. Το ME είναι το μέσο σφάλμα. Και τα τρία μέτρα προσδιορίζουν τα ποσοστά σφάλματος μεταξύ των γνωστών δεδομένων που έγιναν κενά και των δεδομένων που προσδιορίστηκαν για την κάλυψη των κενών αυτών.

N_b	**3**	**5**	**8**	**10**	**15**	**20**
RMSE (MWatt)	627.15	636.56	643.39	641.77	671.27	676.07
RPe	0.38	0.36	0.33	0.33	0.25	0.20
ME (MWatt)	7.38	5.98	8.18	13.52	1.02	2.69

Πίνακας 5.10: Πίνακας μέτρων επιβεβαίωσης για την καλοκαιρινή περίοδο με εφαρμογή της πρώτης μεθοδολογίας. Το RMSE είναι το μέσο τετραγωνικό σφάλμα. Το RPe είναι ο συντελεστής συσχέτισης του Pearson. Το ME είναι το μέσο σφάλμα. Και τα τρία μέτρα προσδιορίζουν τα ποσοστά σφάλματος μεταξύ των γνωστών δεδομένων που έγιναν κενά και των δεδομένων που προσδιορίστηκαν για την κάλυψη των κενών αυτών.

N_b	**3**	**5**	**8**	**10**	**15**	**20**
RMSE (MWatt)	547.07	568.82	581	593.20	606.27	605.33
RPe	0.52	0.49	0.46	0.42	0.39	0.36
ME (MWatt)	3.87	3.20	7.01	2.29	6.75	0.90

Πίνακας 5.11: Πίνακας μέτρων επιβεβαίωσης για την καλοκαιρινή περίοδο με εφαρμογή της δεύτερης μεθοδολογίας. Το RMSE είναι το μέσο τετραγωνικό σφάλμα. Το RPe είναι ο συντελεστής συσχέτισης του Pearson. Το ME είναι το μέσο σφάλμα. Και τα τρία μέτρα προσδιορίζουν τα ποσοστά σφάλματος μεταξύ των γνωστών δεδομένων που έγιναν κενά και των δεδομένων που προσδιορίστηκαν για την κάλυψη των κενών αυτών.

Μετά την μελέτη των παραπάνω πινάκων προκύπτει πως η χαμηλή τιμή του μέσου σφάλματος (ME) δηλώνει αμεροληψία (bias). Επιπλέον μπορούμε να καταλήξουμε στο συμπέρασμα πως και οι δύο μεθοδολογίες δίνουν αποτελέσματα με μικρά αλλά μη αμελητέα σφάλματα. Συγκεκριμένα οι τιμές του συντελεστή συσχέτισης Pearson οι οποίες κυμαίνονται από 20%–38% για την καλοκαιρινή για παράδειγμα περίοδο με εφαρμογή της πρώτης μεθοδολογίας και από 36%–52% για την ίδια περίοδο μέσω της δεύτερης μεθοδολογίας δηλώνουν μέτρια

συσχέτιση μεταξύ των τιμών. Η έκβαση των συμπερασμάτων δεν στηρίζεται στον συντελεστή συσχέτισης λόγω την παρουσίας ιδιόμορφων τιμών και αυτοσυσχετίσεων, οι οποίες επηρεάζουν την τιμή αυτού.

Τα ποσοστά της ρίζας του μέσου τετραγωνικού σφάλματος δείχνουν την αξιοπιστία της πρόβλεψης. Συγκεκριμένα τα ποσοστά αυτά για την περίοδο του καλοκαιριού κυμαίνονται από 12,08% (για εύρος παραθύρου τρία) έως 13,80% (για εύρος παραθύρου 20) για την πρώτη μεθοδολογία και από 12,04% (για εύρος παραθύρου τρία) έως 13,31% (για εύρος παραθύρου 20) για την δεύτερη για εύρος τιμών 4545 (MWatt). Για την χειμερινή περίοδο τα ποσοστά κυμαίνονται από 11,11% (για εύρος παραθύρου τρία) έως 14,88% (για εύρος παραθύρου 20) για την πρώτη μεθοδολογία και από 9,91% (για εύρος παραθύρου τρία) έως 10,85% (για εύρος παραθύρου 20) για την δεύτερη για εύρος τιμών 5570 (MWatt). Επομένως από την συνδυαστική αξιολόγηση των μέτρων επιβεβαίωσης προκύπτει πως η ακρίβεια των εκτιμήσεων είναι ικανοποιητική αλλά περιέχουν μη αμελητέα ποσοστά σφάλματος. Έτσι μετά την σύγκριση των μέτρων επιβεβαίωσης προκύπτει πως η δεύτερη μεθοδολογία δίνει πιο ικανοποιητικά αποτελέσματα σε σύγκριση με την πρώτη. Επομένως θα προτιμηθεί προκειμένου να γίνει κάλυψη κενών στα δεδομένα της χρονοσειράς πριν την επιτέλεση της πρόβλεψης τόσο με το μοντέλο SARIMA όσο και με την μέθοδο του σταθμισμένου κινούμενου μέσου όρου.

Επιπρόσθετα από τους πίνακες αυτούς προκύπτει πως το εύρος παραθύρου (3, 5, 8, 10, ...) που επιλέχθηκε προκειμένου να γίνει η εκτίμηση των τιμών επηρεάζει την αποτελεσματικότητα της μεθόδου. Παρατηρούμε πως το εύρος αυτής επηρεάζει τις μετρήσεις. Όσο μικρότερο είναι το εύρος τόσο καλύτερα είναι τα αποτελέσματα που προκύπτουν, καθώς για περίοδο ίση με 3 και τα τρία κριτήρια είναι καλύτερα σε σύγκριση με την περίοδο ίση με 5 και ούτω καθεξής.

5.4.2 Εποχικές Προβλέψεις με EWMA για δεδομένα ανά 15 λεπτά

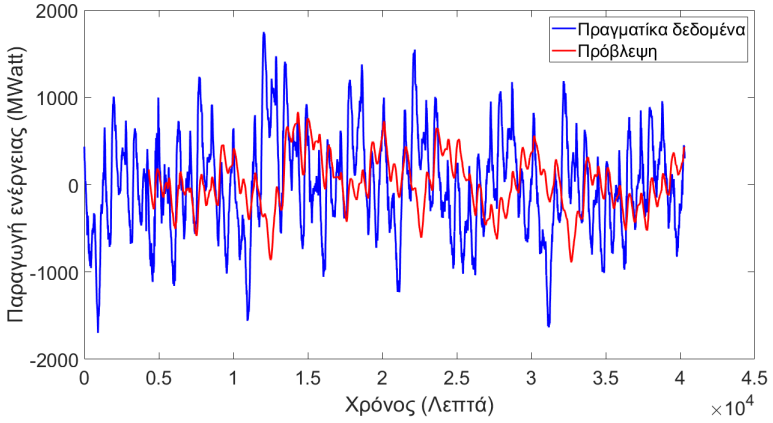
Η πρόβλεψη με τον σταθμισμένο κινούμενο μέσο όρο θα γίνει για μία μέρα μετά. Συγκεκριμένα ο τρόπος προσδιορισμού των μελλοντικών χρονικών στιγμών παρουσιάζεται μέσω της εξίσωσης (εξ. 5.1).

$$t_{i+96} = a_0 t_i + a_1 t_{i-1} + a_2 t_{i-2} \tag{5.1}$$

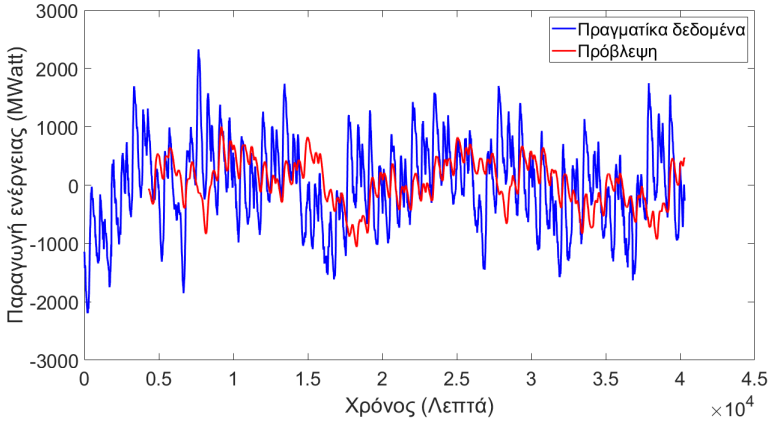
όπου τα a_0, a_1, a_2 εκφράζουν τους σταθμισμένους όρους που προσδιορίζονται με την βοήθεια του σταθμισμένου κινούμενου μέσου όρου (εξ. 3.34).

Η μέθοδος εφαρμόστηκε σε δύο διαφορετικά κομμάτια της χρονοσειράς με χρονικό βήμα ανά 15 λεπτά, ένα για την καλοκαιρινή περίοδο (από 15 Ιουνίου έως 13 Ιουλίου) και ένα για το χειμερινή περίοδο (από 15 Ιανουαρίου έως 12 Φεβρουαρίου) με 2688 σύνολο δεδομένων. Αυτό έγινε λόγω των διαφορετικών ενεργειακών αναγκών, του μεγάλου μεγέθους της χρονοσειράς εάν αυτή ήταν ολόκληρη και προκειμένου να γίνει και μία επιπλέον σύγκριση μεταξύ των προβλέψεων που έγιναν για τις δύο αυτές περιόδους. Έτσι και στις δύο περιπτώσεις εκτιμήθηκε η παραγωγή της κάθε μέρας με βάση τον σταθμισμένο μέσο όρο της παραγωγής μιας περιόδου εύρους 192 λεπτών νωρίτερα. Για την δημιουργία της πρόβλεψης αυτής επιλέγεται από την 2η μεθοδολογία της κάλυψης κενών το εύρος παραθύρου με τα μικρότερα ποσοστά σφάλματος. Έτσι επιλέγεται ένα εύρος παραθύρου ίσο με τρία. Μέσω του εύρους αυτού θα πραγματοποιήσουμε κάλυψη κενών στην χρονοσειρά. Η χρονοσειρά αυτή θα χρησιμοποιηθεί στην συνέχεια για την δημιουργία πρόβλεψης.

Τα αποτελέσματα με την μορφή διαγραμμάτων (σχ. 5.7, σχ. 5.8) καθώς και μέτρων επιβεβαίωσης (πιν. 5.12) και για τις δύο περιόδους παρουσιάζονται παρακάτω.



Σχήμα 5.7: Διάγραμμα πρόβλεψης για την περίοδο του καλοκαιριού με τον σταθμισμένο κινούμενο μέσο όρο (EWMA)



Σχήμα 5.8: Διάγραμμα πρόβλεψης για την περίοδο του χειμώνα με τον σταθμισμένο κινούμενο μέσο όρο (EWMA)

Μέτρα επιβεβαίωσης	**Καλοκαιρινή περίοδος**	**Χειμερινή περίοδος**
RMSE (Watt)	636.71	749.71
ME (MWatt)	0.21	−16.53
RPe	0.10	0.2

Πίνακας 5.12: Μέτρα επιβεβαίωσης της πρόβλεψης για την καλοκαιρινή και την χειμερινή περίοδο με την μέθοδο του EWMA. Το RMSE εκφράζει το μέσο τετραγωνικό σφάλμα. Το ME εκφράζει το μέσο σφάλμα. Το RPe εκφράζει τον συντελεστή συσχέτισης του Pearson. Και τα τρία μέτρα επιβεβαίωσης εκφράζουν τα ποσοστά σφάλματός μεταξύ των γνωστών και προβλεπόμενων δεδομένων.

Παρατηρούμε πως οι χαμηλές τιμές του μέσου σφάλματος δηλώνουν την απουσία αμεροληψίας (bias). Επιπλέον και στις δυο περιόδους ο συντελεστή συσχέτισης Peasron επηρεάζεται από την παρουσία ιδιόμορφων τιμών και επομένως η έκβαση των συμπερασμάτων δεν θα στηριχθεί στο στατιστικό μέτρο αυτό πρόβλεψης. Το μέσο τετραγωνικό σφάλμα κυμαίνεται από 13,45% έως 14,01% ως ποσοστό της αντίστοιχης μέσης στάθμης για την χειμερινή περίοδο (εύρος τιμών 5570 MWatt) και την καλοκαιρινή (εύρος τιμών 4545 MWatt) αντίστοιχα. Τα ποσοστά δηλώνουν την μικρή αλλά μη αμελητέα ύπαρξη σφαλμάτων μεταξύ των γνωστών και των προσδιορισμένων δεδομένων. Τέλος παρατηρούμε πως τα δεδομένα της πρόβλεψης παρουσιάζουν μια ομαλοποίηση σε σχέση με τα ήδη γνωστά δεδομένα. Αυτό το μικρότερο εύρος τιμών είναι αναμενόμενο. Η μέθοδος του σταθμισμένου κινούμενου μέσου όρου δημιουργεί ομαλοποίηση στα δεδομένα που εφαρμόζεται.

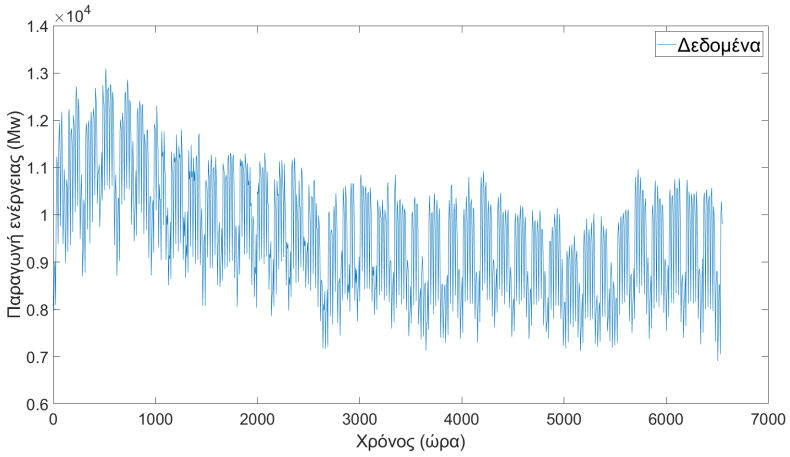
5.5 Πρόβλεψη με την βοήθεια των μοντέλων SARIMA

Το σύνολο των δεδομένων το οποίο χρησιμοποιήθηκε αποτελείται από 26208 τέταρτα την ώρας από την 1η Ιανουαρίου του 2019 έως τις 30 Σεπτεμβρίου του 2019. Η επεξεργασία των δεδομένων έγινε τόσο μέσω χρονοσειράς που αποτελούνταν από τέταρτα της ώρας και συνολικά 26208 δεδομένα όσο και από μέσους όρους εξαώρων με συνολικό μέγεθος χρονοσειράς τα 1092 στοιχεία.

5.5.1 Πρόβλεψη με μοντέλο SARIMA για μέσους όρους ανά εξάωρο

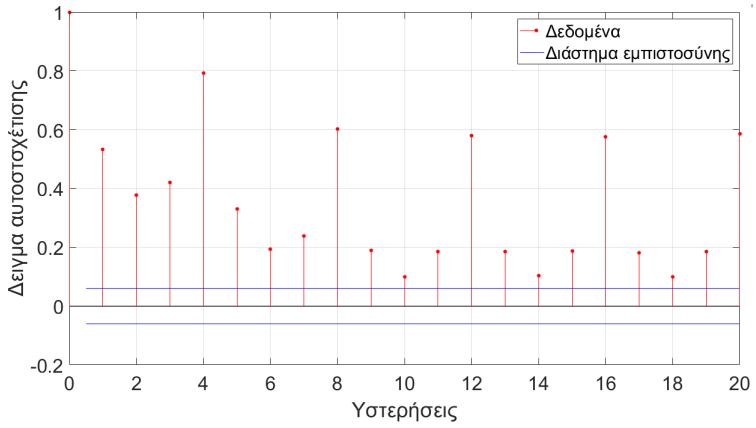
Η μετατροπή των δεδομένων από χρονικό βήμα ανά 15 λεπτά της ώρας σε μέσους όρους ανά εξάωρα έγινε για να μειωθεί ο όγκος των δεδομένων και γιατί παρατηρήθηκε πως η χρονοσειρά αποκλίνει από την κανονική κατανομή. Στην περίπτωση της χρονοσειράς με χρονικό βήμα ανά 15 λεπτά έχουμε να αντιμετωπίσουμε μια κατανομή που αποκλίνει από την κανονική. Έτσι επιλέχτηκε η επεξεργασία των δεδομένων και εν τέλει η προβλέψει να γίνει αρχικά για δεδομένα με χρονικό βήμα εξάωρο (μέσοι όροι ανά τέταρτο). Η κατανομή αυτών πλησιάζει την κανονική (ενότητα 5.2).

Η επιλογή των μοντέλων SARIMA έγινε γιατί είναι μοντέλο με ευρεία χρήση στην αγορά εργασίας. Επιπλέον είναι αποτελεσματικά στην αντιμετώπιση τόσο του αιτιοκρατικού όσο και του στοχαστικού μέρους της χρονοσειράς. Η αντιμετώπιση των οποίων είναι απαραίτητη για να γίνει εφικτή η όσο το δυνατόν πιο αξιόπιστη πρόβλεψη. Το διάγραμμα της χρονοσειρά με την οποία έγινε η επεξεργασία των δεδομένων παρουσιάζεται παρακάτω (Σχ. 5.9).



Σχήμα 5.9: Διάγραμμα Χρονοσειράς με μέσους όρους ανά εξάωρο

Με την βοήθεια του διαγράμματος αυτού μπορούμε να παρατηρήσουμε την παρουσία τάσης και περιοδικότητας αφού η γραφική παράσταση δεν είναι γραμμική αλλά παρουσιάζει μια φθίνουσα κλίση σε συνάρτηση με τον χρόνο. Η τάση και η περιοδικότητα της χρονοσειράς φαίνεται βέβαια και με την βοήθεια ενός διαγράμματος αυτοσυσχέτισης (ACF). Έτσι το διάγραμμα αυτό για την χρονοσειρά με μέσους όρους ανά εξάωρα δίνεται από το γράφημα (Σχ. 5.10).



Σχήμα 5.10: Διάγραμμα αυτοσυσχέτισης για χρονοσειρά με μέσους όρους ανά εξάωρο

Παρατηρούμε πως τα σημεία εξέρχονται από τα όρια εμπιστοσύνης και επομένως η χρονοσειρά δεν μπορεί να χαρακτηριστεί σαν διαδικασία λευκού θορύβου. Επομένως η χρονοσειρά παρουσιάζει τάση και περιοδικότητα, η αφαίρεση των οποίων είναι απαραίτητη. Με την βοήθεια του λογισμικού Matlab αφαιρέσαμε και τα δύο αυτά στοιχεία.

Απαλοιφή αιτιοκρατικού μέρους (τάση και περιοδικότητα)

Η απαλοιφή της τάσης και της περιοδικότητας έγινε με γραμμική παλινδρόμηση με την βοήθεια του λογισμικού της Matlab. Αρχικά έγινε η ταυτοποίηση των δύο στοιχείων. Καταλήξαμε έτσι σε δευτέρου βαθμού τάση και σε δύο περιοδικότητες. Μία ημερήσια και μία εβδομαδιαία. Η απαλοιφή της τάσης έγινε με την βοήθεια του πολυωνύμου (εξ. 5.2)

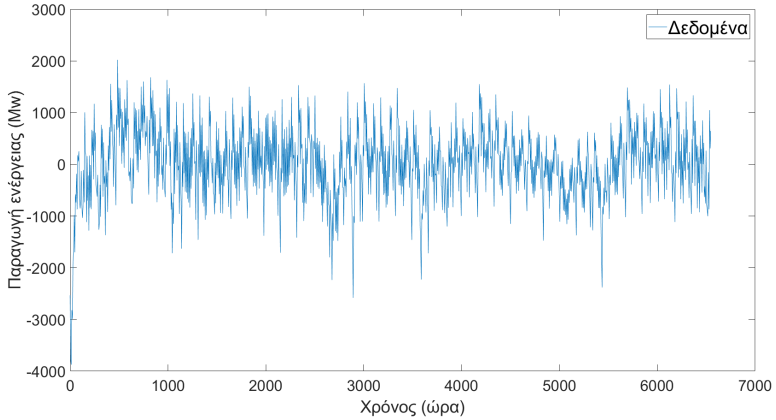
$$M(t_n) = \beta_1 \cos(k_1 t_n) + \beta_2 \sin(k_1 t_n) + \beta_3 \cos(k_2 t_n) + \beta_4 \sin(k_2 t_n) + \alpha_0 + \alpha_1 t_n + \alpha_2 t_n^2 \quad (5.2)$$

όπου τα $\beta_1, \beta_2, \beta_3, \beta_4$ είναι σταθεροί συντελεστές, τα $\alpha_0, \alpha_1, \alpha_2$ είναι οι συντελεστές της τάσης. Το $k_1 = 2\pi/4$ εκφράζει την ημερήσια περιοδικότητα και το $k_2 = 2\pi/4 \cdot 7$ εκφράζει την εβδομαδιαία. Το n εκφράζει το μέγεθος της χρονοσειράς, ενώ το t είναι μια τυχαία χρονική στιγμή. Οι συντελεστές της τάσης δευτέρου βαθμού φαίνονται στον πίνακα (πιν. 5.13).

Συντελεστές	**Τάση δευτέρου βαθμού**
β_1 (MWatt)	-3.54
β_2 (MWatt	0
β_3 (MWatt)	49.25
β_4 (MWatt)	1.60
α_0(MWatt)	1.14×10^4
α_1 (MWatt/sec)	-1
α_2(MWatt/sec^2)	1.02×10^{-4}

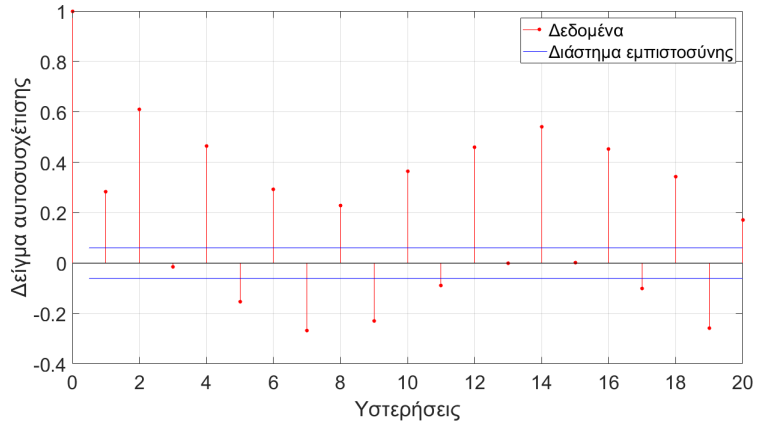
Πίνακας 5.13: Πίνακας συντελεστών δευτέρου βαθμού τάσης. Τα β_1, β_2, β_3, β_4 είναι οι σταθεροί συντελεστές της τάσης πρώτου και δευτέρου βαθμού. Τα $\alpha_0, \alpha_1, \alpha_2$ είναι οι συντελεστές της τάσης

Η χρονοσειρά που προέκυψε φαίνεται στο διάγραμμα (Σχ. 5.11). Διακρίνεται μία ομαλοποίηση της χρονοσειράς σε σχέση με τον χρόνο. Η ομαλοποίηση αυτή είναι ένα πρώτο δείγμα της πιθανής αποτελεσματικής αντιμετώπισης της τάσης και της περιοδικότητας σε ικανοποιητικό βαθμό.



Σχήμα 5.11: Διάγραμμα χρονοσειράς για μέσους όρους ανά εξάωρο μετά την αφαίρεση της τάσης και περιοδικότητας.

Γίνεται αντιληπτό πως παρόλο την αποτελεσματική αντιμετώπιση της τάσης και της περιοδικότητας η εμφάνιση ακραίων τιμών στο γράφημα (σχ. 5.11) παραπέμπει στην εμφάνιση αυτοσυσχετίσεων μεταξύ των τιμών. Το συμπέρασμα αυτό προκύπτει από την βοήθεια του διαγράμματος αυτοσυσχέτισης (Σχ. 5.12). Το διάγραμμα αυτό στην προκειμένη περίπτωση δείχνει την παρουσία αυτοσυσχετίσεων λόγω της ύπαρξης τιμών εκτός των ορίων εμπιστοσύνης. Επομένως η χρονοσειρά δεν μπορεί να χαρακτηριστεί ως λευκός θόρυβος.



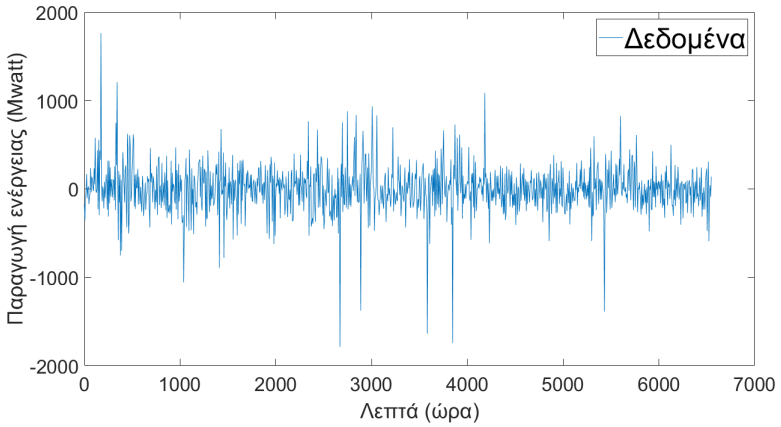
Σχήμα 5.12: Διάγραμμα αυτοσυσχέτισης για χρονοσειρά με μέσους όρους ανά εξάωρα μετά την αφαίρεση της τάσης και της περιοδικότητας

Απαλοιφή στοχαστικού μέρους με μοντέλο SARIMA(p,d,q)(P,D,Q)

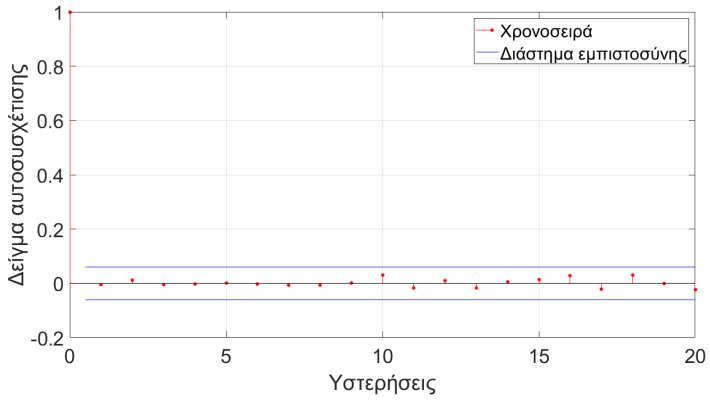
Για την αντιμετώπιση των αυτοσυσχετίσεων μεταξύ των τιμών που παρουσιάστηκαν στο διάγραμμα (σχ. 5.12) εφαρμόστηκε ένα εποχιακό αυτοπαλινδρομούμενο μοντέλο κινούμενου μέσου όρου SARIMA(p,d,q)(P,D,Q)m. Όπως και στην περίπτωση της εντολής regress έτσι και εδώ γίνεται αναγνώριση των περιοδικοτήτων και της τάσης δευτέρου βαθμού. Αυτό γίνεται για να γίνει εφικτή η εφαρμογή του μοντέλου SARIMA. Έτσι καταλήγουμε σε δύο περιοδικότητες μία ημερήσια και μία εβδομαδιαία. Θα μπορούσαν να υπάρχουν και άλλες εποχικότητες στο σύνολο των δεδομένων όπως ετήσια και μηνιαία. Στην περίπτωση της ετήσιας, αυτό είναι αδύνατο αφού το σύνολο των δεδομένων αναφέρεται σε μια περίοδο εννιά μηνών. Αξίζει επίσης να σημειωθεί πως η ύπαρξη μηνιαίας περιοδικότητας δεν είναι εφικτή. Αυτό οφείλεται στις διαφορετικές ενεργειακές ανάγκες που πιθανότατα παρουσιάζονται σε διαφορετικές χρονικές περιόδους του χρόνου. Έτσι για παράδειγμα διαφορετικές ενεργειακές ανάγκες αναμένουμε την πρωτοχρονιά σε σχέση με την πρώτη μέρα του Φεβρουαρίου.

Σαν αποτέλεσμα αυτού είναι η παρουσία προβλημάτων στασιμότητας στην χρονοσειρά ενδιαφέροντος.

Οι συντελεστές του μοντέλου SARIMA επιλέχθηκαν με την βοήθεια του κριτηρίου AIC, διαλέγοντας τον καλύτερο συντελεστή. Έτσι καταλήξαμε στο μοντέλο SARIMA(32, 0, 20)(60, 0, 20). Τα δυο διαγράμματα που ακολουθούν (Σχ. 5.13), (Σχ. 5.14) δείχνουν την απαλοιφή των αυτοσυσχετίσεων αφού ιδιαίτερα στην περίπτωση του διαγράμματος αυτοσυσχέτισης οι τιμές πλέον της χρονοσειράς βρίσκονται εντός των ορίων εμπιστοσύνης (confidence bounds).



Σχήμα 5.13: Διάγραμμα χρονοσειράς με μέσους όρους ανά εξάωρα μετά την αφαίρεση των αυτοσυσχετίσεων



Σχήμα 5.14: Διάγραμμα αυτοσυσχέτισεις με μέσους όρους ανά εξάωρα μετά την αφαίρεση των αυτοσυσχετίσεων

Δημιουργία πρόβλεψης με το μοντέλο SARIMA

Μετά την αφαίρεση του αιτιοκρατικού (τάση και εποχικότητα) και στοχαστικού (αυτοσυσχετίσεις) μέρους της χρονοσειράς, είναι πλέον εφικτή η δημιουργία πρόβλεψης και η αξιολόγηση των αποτελεσμάτων. Το σύνολο των στοιχείων της χρονοσειράς είναι 1092 δεδομένα. Για την δημιουργία της πρόβλεψης εκπαιδεύουμαι την χρονοσειρά με το στοιχείο που αντιστοιχεί στην χρονική στιγμή t_1 έως το στοιχείο την χρονική στιγμή $t_{1000+i-1}$. Το i εκφράζει τον αριθμό των προβλέψεων, $i = 1, ..., 64$. Οι προβλέψεις που πραγματοποιήθηκαν απευθύνονται σε ένα σύνολο 22 ημερών. Στην συνέχεια γίνεται εκτίμηση της παραγωγής ηλεκτρικής ενέργειας για την χρονική στιγμή $t_{1000+4+i-1}$ για μία μέρα στο μέλλον (αντιστοιχεί σε 4 εξάωρα). Επιπλέον εκτιμάται η χρονική στιγμή $t_{1000+8+i-1}$ για 2 μέρες στο μέλλον και η $t_{1000+28+i-1}$ για μια βδομάδα μετά. Παρακάτω φαίνονται τα αποτελέσματα που προέκυψαν από την πρόβλεψη που έγινε με την βοήθεια του εποχιακού αυτοπαλινδρομούμενου μοντέλου κινούμενου μέσου όρου. Η σύγκριση των αποτελεσμάτων έγινε με την βοήθεια των μέτρων επιβεβαίωσης και συγκεκριμένα των RMSE, RPe, MA.

Μέτρα επιβεβαίωσης	**Τιμές**
RMSE (MWatt) (μία μέρα μετά)	226.92
RMSE (MWatt) (δύο μέρες μετά)	225.43
RMSE (MWatt) (επτά μέρες μετά)	271.35
ME (MWatt) (μία μέρα μετά)	−58.02
ME (MWatt) (δύο μέρες μετά)	−90.27
ME (MWatt) (επτά μέρες μετά)	−195.56
RPe (μία μέρα μετά)	0.98
RPe (δύο μέρες μετά)	0.98
RPe (επτά μέρες μετά)	0.98

Πίνακας 5.14: Πίνακας μέτρων επιβεβαίωσης για πρόβλεψη με χρονοσειρά με μέσους όρους ανά εξάωρο. Το RMSE είναι το μέσο τετραγωνικό σφάλμα. Το ME είναι το μέσο σφάλμα. Το RPe είναι ο συντελεστής συσχέτισης του Pearson. Και τα τρία μέτρα επιβεβαίωσης αξιολογούν τις προβλέψεις που έγιναν για μία μέρα μετά, για δύο μέρες μετά και για μία βδομάδα μετά.

Χρησιμοποιώντας την μέθοδο SARIMA προκειμένου να πραγματοποιηθεί πρόβλεψη για την παραγωγή ηλεκτρικής ενέργειας στο Βέλγιο όταν η χρονοσειρά έχει τροποποιηθεί σε μέσους όρους ανά εξάωρα, τα αποτελέσματα είναι ικανοποιητικά καθώς τα μέτρα επιβεβαίωσης που χρησιμοποιήθηκαν περιέχουν μικρά ποσοστά σφάλματος. Για παράδειγμα στην περίπτωση του RPe ένα μέγεθος της τάξης του 98% είναι προφανώς ένα εξαιρετικά ικανοποιητικό ποσοστό αφού εκφράζει συσχέτιση μεταξύ των γνωστών δεδομένων και των δεδομένων πρόβλεψης. Στην περίπτωση του RMSE το ποσοστό του σφάλματος κυμαίνεται από 3,69% (για βήμα πρόβλεψης μίας ημέρας) έως 4,43% (για πρόβλεψη μία βδομάδα μετά) για εύρος τιμών 6120 (MWatt). Τέλος οι τιμές του μέσου σφάλματος δηλώνουν την παρουσία αμεροληψίας (bias).

5.5.2 Εποχικές προβλέψεις με μοντέλο SARIMA για δεδομένα με χρονικό βήμα ανά 15 λεπτά

Η ίδια διαδικασία που πραγματοποιήθηκε για την χρονοσειρά με μέσους όρους ανά εξάωρο έγινε και για την χρονοσειρά ανά τέταρτο την ώρας. Με την διαφορά ότι λόγω του μεγάλου όγκου των δεδομένων η επεξεργασία καθώς και η σύγκριση των δεδομένων και των προβλέψεων δεν έγινε για όλη την χρονοσειρά αλλά για κάποια κομμάτια αυτής και συγκεκριμένα επιλέχθηκε ένα κομμάτι για την καλοκαιρινή περίοδο από 15 Ιούνιο έως 13 Ιουλίου και ένα κομμάτι για την χειμερινή περίοδο από 15 Ιανουαρίου έως 12 Φεβρουαρίου. Το σύνολο των δεδομένων είναι 2688 στοιχεία.

Επιλογή βαθμού τάσης

Πριν γίνει η παράθεση και επεξήγηση των δυο αυτών αναλύσεων πρέπει να γίνει αναφορά στην ιδιαιτερότητα που παρουσιάζει η τάση σε σύγκριση με την τάση της χρονοσειράς με μέσους όρους ανά εξάωρα. Στην περίπτωση της χρονοσειράς αυτής καταλήξαμε στο συμπέρασμα πως έχουμε δεύτερου βαθμού τάση (υπό-ενότητα 5.5.1). Στην προκειμένη περίπτωση πρέπει να γίνει μια διευκρίνησή σχετικά με το τι τάση περιέχει η χρονοσειρά, αφού εύλογα μπορούμε να υποθέσουμε πως μέσα σε ένα τόσο μικρό χρονικό διάστημα αυτό του ενός περίπου μήνα ίσως να ήταν λάθος η εφαρμογή μιας τάσης δευτέρου βαθμού. Ο προσδιορισμός της τάσης δευτέρου βαθμού έγινε με την βοήθεια του πολυωνύμου (Εξ. 5.3)

$$M(t_n) = \beta_1 \cos(k_1 t_n) + \beta_2 \sin(k_1 t_n) + \beta_3 \cos(k_2 t_n) + \beta_4 \sin(k_2 t_n) + \alpha_0 + \alpha_1 t_n + \alpha_2 t_n^2 \quad (5.3)$$

όπου τα $\beta_1, \beta_2, \beta_3, \beta_4$ είναι σταθεροί συντελεστές, τα $\alpha_0, \alpha_1, \alpha_2$ είναι οι συντελεστές της τάσης. Το $k_1 = 2\pi/96$ εκφράζει την ημερήσια περιοδικότητα και

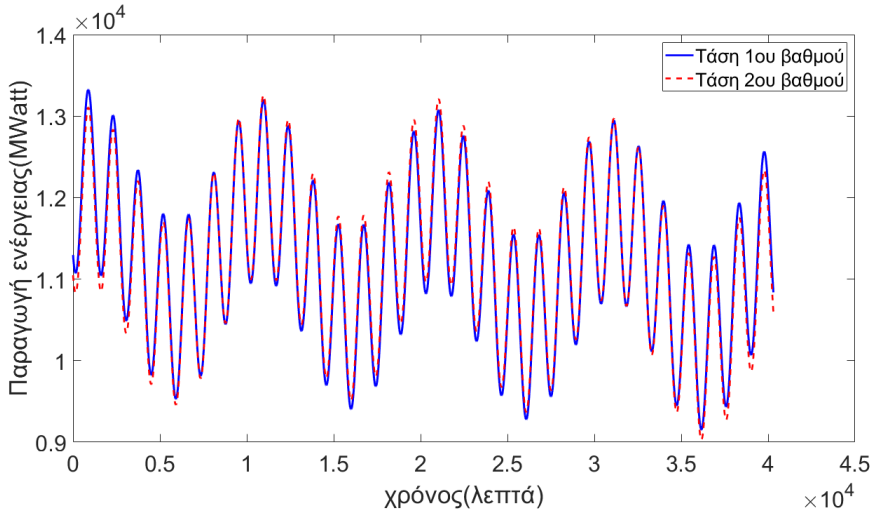
το $k_2 = 2\pi/(96 \cdot 7)$ εκφράζει την εβδομαδιαία. Το n εκφράζει το μέγεθος της χρονοσειράς, ενώ το t είναι μια τυχαία χρονική στιγμή.

Όσον αφορά τον προσδιορισμό της τάσης πρώτου βαθμού χρησιμοποιήθηκε το πολυώνυμο (5.4)

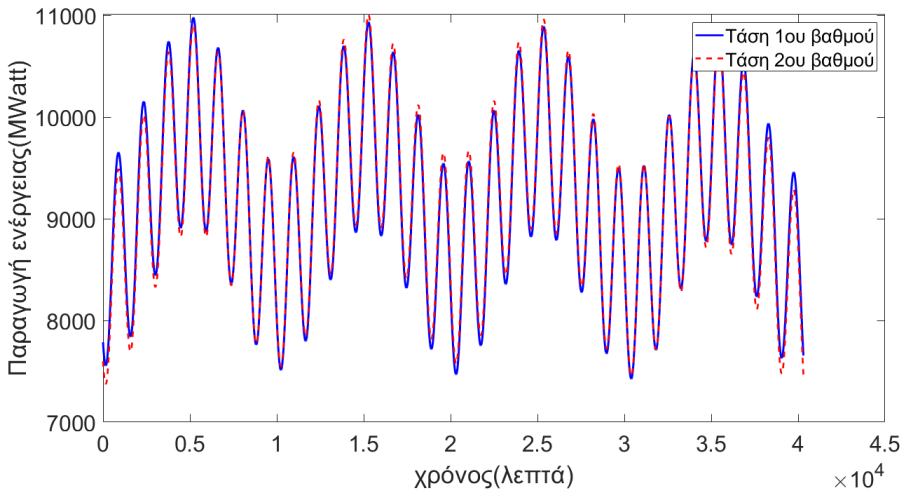
$$M(t_n) = \beta_1 \cos(k_1 t_n) + \beta_2 \sin(k_1 t_n) + \beta_3 \cos(k_2 t_n) + \beta_4 \sin(k_2 t_n) + \alpha_0 + \alpha_1 t_n \quad (5.4)$$

όπου τα $\beta_1, \beta_2, \beta_3, \beta_4$ είναι σταθεροί συντελεστές, τα α_0, α_1 είναι οι συντελεστές της τάσης. Το $k_1 = 2\pi/96$ εκφράζει την ημερήσια περιοδικότητα και το $k_2 = 2\pi/96 \cdot 7$ εκφράζει την εβδομαδιαία. Το n εκφράζει το μέγεθος της χρονοσειράς, ενώ το t είναι μία τυχαία χρονική στιγμή.

Ο λόγος που εν τέλει επιλέγεται η τάση πρώτου βαθμού μπορεί να φανεί καλύτερα μέσω των διαγραμμάτων που ακολουθούν (Σχ. 5.15), (Σχ. 5.16) και για τις δύο περιόδους που χρησιμοποιήθηκαν.



Σχήμα 5.15: Διάγραμμα διαφοράς τάσεων πρώτου και δευτέρου βαθμού για την χειμερινή περίοδο



Σχήμα 5.16: Διάγραμμα διαφοράς τάσεων πρώτου και δευτέρου βαθμού για την καλοκαιρινή περίοδο

Από τα διαγράμματα αυτά μπορούμε να συμπεράνουμε πως δεν υπάρχει ιδιαίτερα μεγάλη διαφορά μεταξύ των δύο τάσεων. Αυτό μπορεί να φανεί καλύτερα και από τους συντελεστές τους στον πίνακα (Πιν. 5.15)

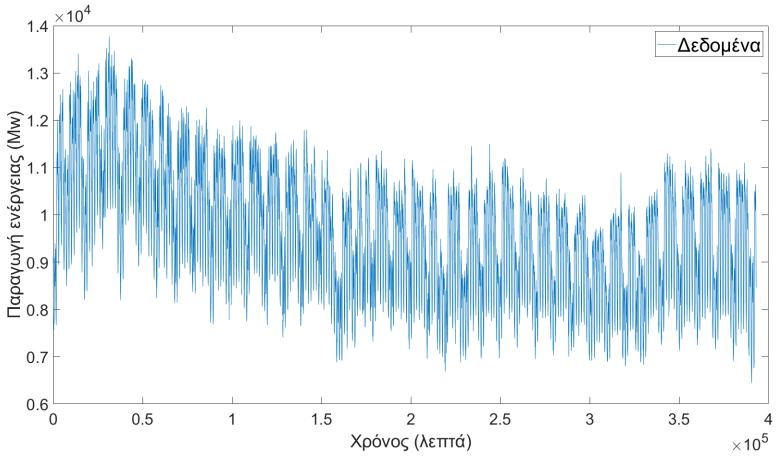
Συντελεστές	**Τάση πρώτου βαθμού**	**Τάση δευτέρου βαθμού**
β_1 (MWatt)	12.33	12.33
β_2 (MWatt)	2.15	2.15
β_3 (MWatt)	57.03	57.08
β_4 (MWatt)	30.01	30.01
α_0(MWatt)	9.24×10^3	8.98×10^3
α_1 (MWatt/sec)	-0.002	0.03
α_2(MWatt/sec^2)		-9.72×10^{-7}

Πίνακας 5.15: Πίνακας συντελεστών πρώτου και δευτέρου βαθμού τάσης. Τα β_1, β_2, β_3, β_4 είναι οι σταθεροί συντελεστές της τάσης πρώτου και δευτέρου βαθμού. Τα $\alpha_0, \alpha_1, \alpha_2$ είναι οι συντελεστές της τάσης

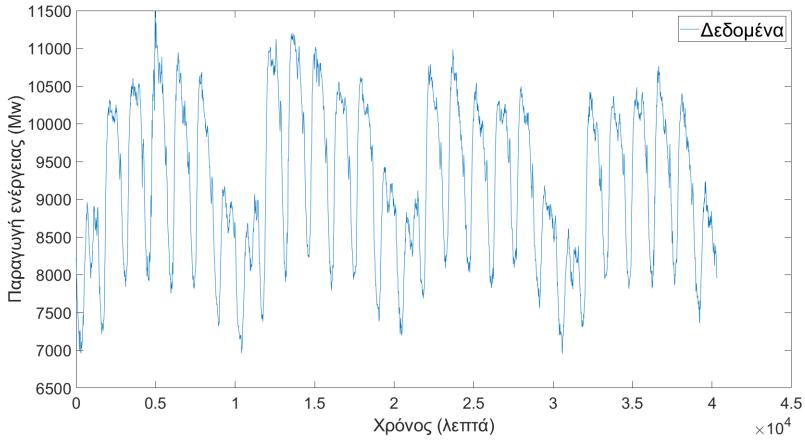
Ιδιαίτερα από την παρατήρηση του πίνακα αυτού μπορεί να γίνει αντιληπτή η ομοιότητα που παρουσιάζουν οι συντελεστές των δύο τάσεων. Παρόλο αυτά γίνεται επιλογή της τάσης πρώτου βαθμού. Αυτό συμβαίνει γιατί οι παράμετροι έχουν μικρή διαφορά και επομένως θα επιλεγεί η απλούστερη μορφής τάσης, η οποία μπορεί να μας προφυλάξει και από την εμφάνιση απότομων αλλαγών. Αυτό σημαίνει ότι μία τάση δευτέρου βαθμού μπορεί να προσαρμόζεται καλύτερα σε ήδη γνωστά δεδομένα αλλά υπάρχει η πιθανότητα να δημιουργεί μεγάλο σφάλμα όταν η πρόβλεψη δεν συμβαδίζει με τα ήδη γνωστά δεδομένα της χρονοσειρά.

5.5.3 Χρονοσειρά με χρονικό βήμα ανά 15 λεπτά για την καλοκαιρινή περίοδο

Για την καλοκαιρινή περίοδο έγινε επιλογή δεδομένων από τις 15 Ιουνίου έως 13 Ιουλίου. Η γραφική απεικόνιση της χρονοσειράς με χρονικό βήμα ανά 15 λεπτά της ώρας (Σχ. 5.17) για όλη την χρονοσειρά αλλά και για το συγκεκριμένο διάστημα (Σχ. 5.18) παρουσιάζονται ακολούθως.

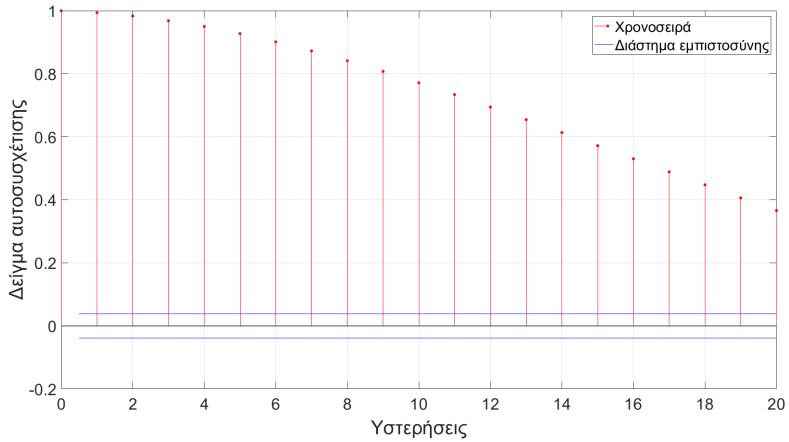


Σχήμα 5.17: Διάγραμμα χρονοσειράς ανά 15 λεπτά



Σχήμα 5.18: Διάγραμμα χρονοσειράς ανά 15 λεπτά για την καλοκαιρινή περίοδο

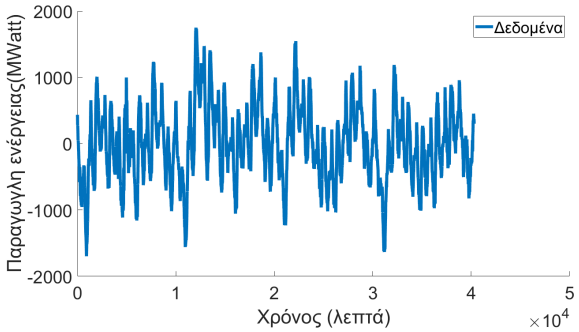
Τα διάγραμμα αυτά μας δίνουν ένα πρώτο συμπέρασμα σχετικά με την ύπαρξη τάσης και περιοδικότητας. Η παρουσία αυτών φαίνεται ακόμα καλύτερα όμως στο διάγραμμα αυτοσυσχέτισης (Σχ. 5.19) για την χρονοσειρά για το καλοκαίρι που ακολουθεί. Οι τιμές των δεδομένων βρίσκονται εκτός ορίων και επομένως είναι σίγουρη η ύπαρξη τάσης και περιοδικότητας στην χρονοσειρά παραγωγής ηλεκτρικής ενέργειας με χρονικό βήμα ανά 15 λεπτά.



Σχήμα 5.19: Διάγραμμα αυτοσυσχέτισης για την χρονοσειρά ανά 15 λεπτά για την καλοκαιρινή περίοδο

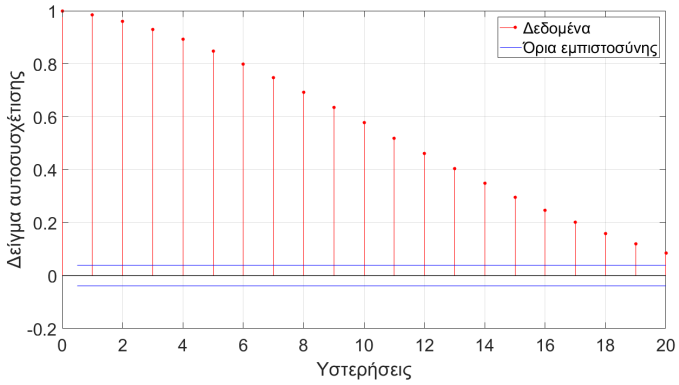
Απαλοιφή αιτιοκρατικού μέρους (τάση και περιοδικότητα)

Γίνεται αντιληπτή η αναγκαιότητα αντιμετώπισης του αιτιοκρατικού μέρους της χρονοσειράς προκείμενου να γίνει εφικτή η όσο τον δυνατόν πιο αξιόπιστη πρόβλεψη. Έτσι αρχικά με γραμμική παλινδρόμηση με την βοήθεια του λογισμικού της Matlab όπως και στην περίπτωση που η χρονοσειρά ήταν χωρισμένη σε μέσους όρους ανά εξάωρα θα γίνει απαλοιφή τόσο της τάσης όσο και της περιοδικότητας. Αρχικά πραγματοποιείται αναγνώριση των στοιχείων. Έχουμε καταλήξει σε πρώτου βαθμού τάση και σε δύο περιοδικότητες. Μία ημερήσια και μία εβδομαδιαία. Το αποτέλεσμα της εφαρμογής της εντολής στην χρονοσειρά φαίνεται στο ακόλουθο γράφημα (Σχ. 5.20)



Σχήμα 5.20: Διάγραμμα χρονοσειράς ανά 15 λεπτά για την καλοκαιρινή περίοδο μετά την αφαίρεση της τάσης και της περιοδικότητας

Η ομαλοποίηση αυτή που μπορεί να παρατηρηθεί δηλώνει πως η χρονοσειρά έχει απαλλαγεί από την τάση και την περιοδικότητα. Το πόσο ισχύει όμως το συμπέρασμα αυτό θα φανεί από το διάγραμμα αυτοσυσχέτισης (Σχ. 5.21)



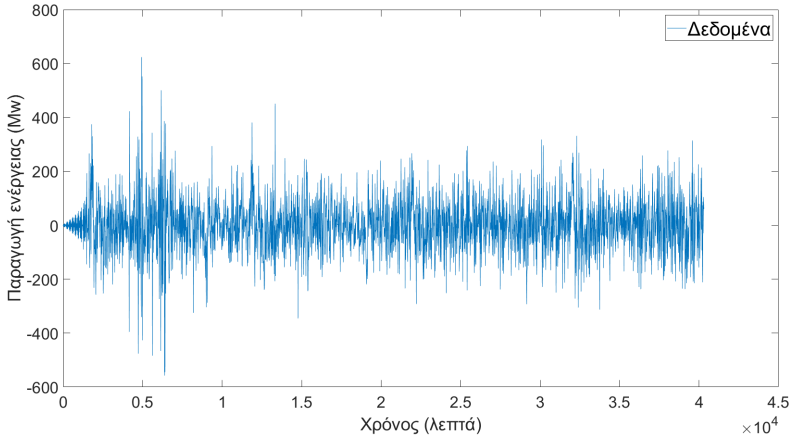
Σχήμα 5.21: Διάγραμμα αυτοσυσχέτισης για χρονοσειρά ανά 15 λεπτά μετά την αφαίρεση της τάσης και της περιοδικότητας για την περίοδο του καλοκαιριού

Είναι εμφανές πως μετά την αντιμετώπιση των στοιχείων αυτών, το μόνο που προκύπτει από το γράφημα, αφού τα σημεία του εξέρχονται σημαντικά από τα όρια εμπιστοσύνης, είναι η παρουσία αυτοσυσχετίσεων μεταξύ των τιμών. Η χρονοσειρά επομένως δεν μπορεί να χαρακτηριστεί ως λευκός θόρυβος.

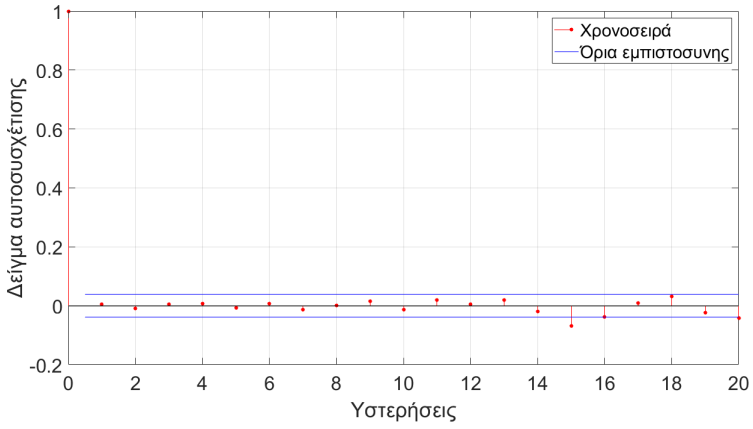
Απαλοιφή στοχαστικού μέρους με μοντέλο SARIMA(p,d,q)(P,D,Q)

Για να απαλλαγούμε από την παρουσία των αυτοσυσχετίσεων, όπως και στην περίπτωση που η χρονοσειρά ήταν χωρισμένοι σε μέσους όρους ανά εξάωρα, έτσι και εδώ θα εφαρμόσουμε ένα μοντέλο SARIMA(p,d,q)(P,D,Q). Αρχικά πρέπει να γίνει η αναγνώριση των περιοδικοτήτων που εμφανίζονται στην χρονοσειρά και η ο βαθμός της τάσης. Η αναγνώριση των στοιχείων αυτών πραγματοποιείται για να γίνει εφικτή η εφαρμογή του μοντέλου SARIMA. Καταλήγουμε στο συμπέρασμα πως λόγο του όγκου των δεδομένων που διαθέτουμε μπορούμε να έχουμε δύο περιοδικότητες, μία ημερησία και μία εβδομαδιαία. Η ύπαρξη άλλων περιοδικοτήτων δεν είναι εφικτή αφού ούτε μηνιαία μπορεί να υπάρχει μίας και το σύνολο των δεδομένων αποτελείται από δεδομένα που αντιστοιχούν σε 28 μέρες και προφανώς ούτε ετήσια. Επιπλέον καταλήξουμε σε πρώτου βαθμού τάση.

Η επιλογή των συντελεστών του μοντέλου έγινε με την βοήθεια του κριτηρίου AIC, διαλέγοντας τον καλύτερο συντελεστή. Έτσι καταλήξαμε στο μοντέλο SARIMA(0, 2, 7)(104, 2, 7). Τα δυο διαγράμματα που ακολουθούν (σχ. 5.22, σχ. 5.23) δείχνουν την μερική απαλοιφή των αυτοσυσχετίσεων μεταξύ των τιμών αφού ιδιαίτερα στην περίπτωση του διαγράμματος αυτοσυσχέτισης οι τιμές πλέον της χρονοσειράς βρίσκονται σχεδόν μέσα στα όρια εμπιστοσύνης (confidence bounds). Η αυτοσυσχετίσης δεν έχουν απαλοιφή πλήρως. Αυτό φαίνεται από το διάγραμμα αυτοσυσχέτισης (σχ. 5.23). Ακόμα και μετά την εφαρμογή του μοντέλου SARIMA είναι εμφανής η παρουσία τιμών εκτός των ορίων εμπιστοσύνης.



Σχήμα 5.22: Διάγραμμα χρονοσειράς ανά 15 λεπτά για την καλοκαιρινή περίοδο μετά την απαλοιφή των αυτοσυσχετίσεων



Σχήμα 5.23: Διάγραμμα αυτοσυσχέτισεις για χρονοσειρά ανά 15 λεπτά για την καλοκαιρινή περίοδο μετά την απαλοιφή των αυτοσυσχετίσεων

Δημιουργία πρόβλεψης με το μοντέλο SARIMA

Μετά την αφαίρεση του στοχαστικού (αυτοσυσχετίσης) και αιτιοκρατικού (τάση και περιοδικότητα) μέρους της χρονοσειράς είναι πλέον εφικτή η δημιουργία πρόβλεψης και η εξαγωγή συμπερασμάτων σχετικά με την αποτελεσματικότητα της μεθόδου που χρησιμοποιήθηκε για την πρόβλεψη στο τέταρτο της ώρας.

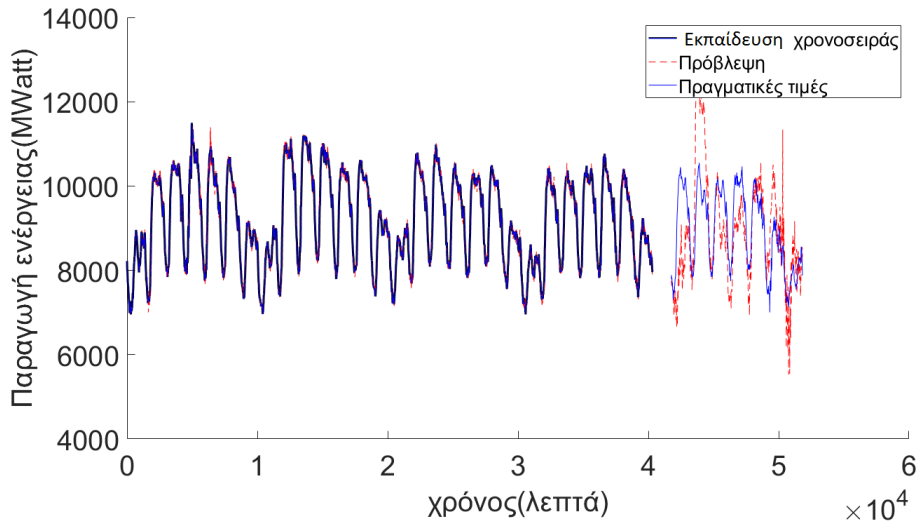
Για να μπορέσει να γίνει αυτή, εκπαιδεύουμε την χρονοσειρά των 2688 δεδομένων με το στοιχείο που αντιστοιχεί στην χρονική στιγμή t_1 έως το στοιχείο την χρονική στιγμή $t_{2687+i+t_1}$. Το i εκφράζει τον αριθμό των προβλέψεων, $i = 1, ..., 672$ που αντιστοιχεί σε μία βδομάδα. Στην συνέχεια γίνεται εκτίμηση της παραγωγής ηλεκτρικής ενέργειας για την χρονική στιγμή $t_{2687+i+96+t_1}$ για μία μέρα μετά (αντιστοιχεί σε 96 τέταρτα). Επιπλέον εκτιμάται η χρονική στιγμή $t_{2687+i+48+t_1}$ για 12 ώρες μετά και η $t_{2688+i+192+t_1}$ για δύο μέρες μετά. Οι συγκρίσεις των αποτελεσμάτων μεταξύ των προβλεπόμενων και τον πραγματικών τιμών γίνονται τόσο σε πίνακα (Πιν. 5.16) με την βοήθεια των μέτρων επιβεβαίωσης όσο και με την βοήθεια διαγράμματος (Σχ. 5.24). Ο πίνακας και το διάγραμμα ακολουθεί παρακάτω

Μέτρα επιβεβαίωσης	**12 ώρες μετά**	**1 μέρα μετά**	**2 μέρες μετά**
RMSE (MWatt)	0.71×10^3	1.03×10^3	2.01×10^3
ME (MWatt)	48.28	55.74	157.32
RPe	0.76	0.62	0.21

Πίνακας 5.16: Πίνακας μέτρων επιβεβαίωσης για την καλοκαιρινή περίοδο για την χρονοσειρά ανά 15 λεπτά. Το RMSE προσδιορίζει το μέσο τετραγωνικό σφάλμα. Το ME είναι το μέσο σφάλμα. Το RPe είναι ο συντελεστής συσχέτισης του Pearson. Και τα τρία μέτρα επιβεβαίωσης πραγματοποιούν αξιολόγηση μεταξύ των γνωστών δεδομένων και των δεδομένων πρόβλεψης για 12 ώρες μετά, 1 μέρα και 2 μέρες μετά.

Γίνεται έτσι αντιληπτό πως στην περίπτωση του μοντέλου SARIMA η πρόβλεψη που γίνεται για 12 ώρες στο μέλλον με ποσοστό της ρίζας του τετραγωνι-

κού σφάλματος 15,62% για εύρος τιμών 4545 (MWatt) είναι καλύτερη από τις αντίστοιχες για μία και δύο μέρες στο μέλλον με ποσοστά τετραγωνικού σφάλματος 22,66% και 44,22% αντίστοιχα. Επιπλέον ο συντελεστής συσχέτισης του Pearson ο οποίος κυμαίνεται από 21%–76% εκφράζει μία μέτρια προς ισχυρή συσχέτιση μεταξύ των τιμών. Η παρουσία ιδιόμορφων τιμών και αυτοσυσχετίσεων στο σύνολο των δεδομένων επηρεάζει την τιμή του συντελεστή συσχέτισης. Επομένως η έκβαση των συμπερασμάτων δεν θα στηριχθεί στο μέτρο αυτό επιβεβαίωσης. Οι τιμές του μέσου σφάλματος εκφράζουν την παρουσία αμεροληψίας. Από την συνολική αξιολόγηση των μέτρων επιβεβαίωσης καταλήγουμε πως και οι τρεις προβλέψεις περιέχουν μικρά αλλά μη αμελητέα ποσοστά σφάλματος. Η παρουσία των σφαλμάτων αυτών φαίνεται και στο γράφημα που ακολουθεί το οποίο κάνει μια σύγκριση μεταξύ των πραγματικών τιμών και τιμών που προβλέφθηκαν.



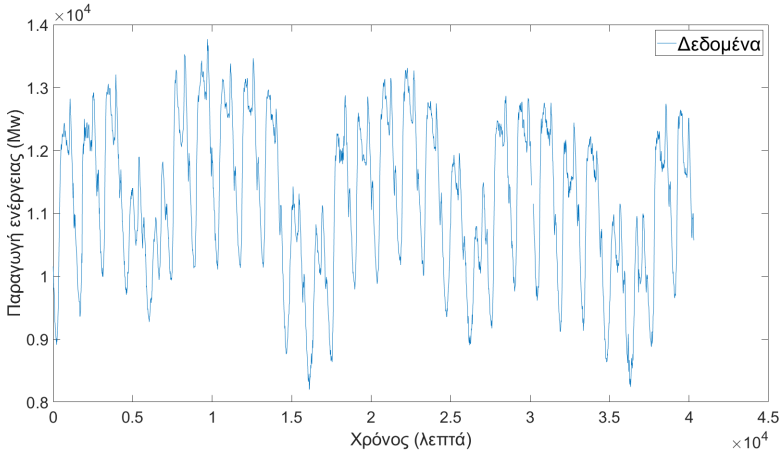
Σχήμα 5.24: Διάγραμμα γνωστών δεδομένων και πρόβλεψης για την καλοκαιρινή περίοδο για χρονοσειρά ανά 15 λεπτά. Το κομμάτι εκπαίδευσης της χρονοσειράς φαίνεται με μαύρη συνεχή γραμμή. Το κομμάτι τις πρόβλεψης φαίνεται με κόκκινη διακεκομμένη. Οι πραγματικές τιμές για την περίοδο της πρόβλεψης φαίνονται με μπλε συνεχή γραμμή

Είναι έτσι εμφανές πως το μοντέλο SARIMA πραγματοποίησε προβλέψεις με σημαντική παρουσία σφαλμάτων και για αυτό τα δεδομένα στο κομμάτι της εκτίμησης έχουν μεγάλες αποκλίσεις από τα ήδη γνωστά δεδομένα. Αυτό οφείλεται στην μη κανονική κατανομή των δεδομένων που διαθέτουμε. Επιπλέον ένας ακόμα παράγοντας είναι η απουσία στασιμότητας της χρονοσειράς με χρονικό βήμα ανά 15 λεπτά για την περίοδο του καλοκαιριού. Το συμπέρασμα αυτό προκύπτει μετά την μελέτη των διαγραμμάτων (σχ. 5.20, σχ. 5.22) όπου οι έντονες διακυμάνσεις και οι ακραίες τιμές που παρατηρούνται στα γραφήματα αυτά παραπέμπουν σε απουσία στασιμότητας και σε ύπαρξη αυτοσυσχετίσεων μεταξύ των τιμών.

5.5.4 Χρονοσειρά με χρονικό βήμα ανά 15 λεπτά για την χειμερινή περίοδο

Στο τελευταίο κομμάτι της ανάλυσης όσον αφορά την μέθοδο του εποχιακού αυτοπαλινδρομούμενου μοντέλου, αποσπάσαμε ένα ακόμα κομμάτι από το σύνολο της χρονοσειράς, το οποίο αναφέρεται στην περίοδο μεταξύ της 15 Ιανουαρίου του 2019 έως 12 Φεβρουαρίου του 2019 από την ημερομηνία αυτή.

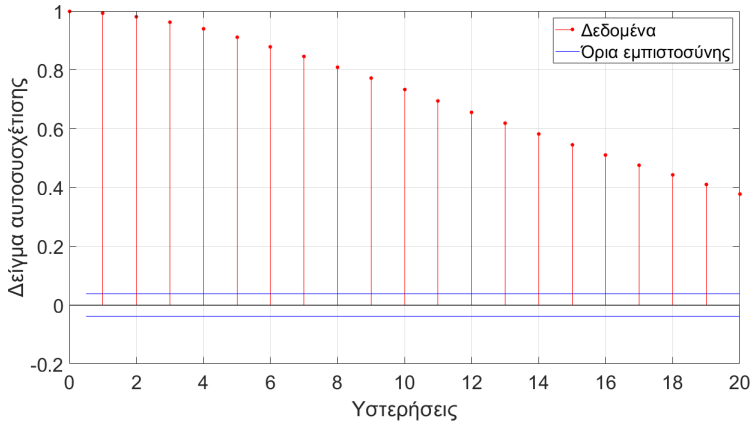
Η γραφική παράσταση για την περίοδο αυτή παρουσιάζεται παρακάτω στο διάγραμμα (Σχ. 5.25).



Σχήμα 5.25: Διάγραμμα χρονοσειράς ανά 15 λεπτά για την χειμερινή περίοδο

Το διάγραμμα αυτό αποτελεί ένα πρώτο δείγμα για την αναγνώριση φαινομένων όπως η τάση και η περιοδικότητα, αφού μπορούμε να παρατηρήσουμε τις διαβαθμίσεις που παρουσιάζει η χρονοσειρά σε σχέση με τον χρόνο. Έτσι συμπεραίνουμε πως η χρονοσειρά παρουσιάζει μεγέθη που πρέπει να εξαλείψουμε.

Αυτό θα φανεί ακόμα καλύτερα με την βοήθεια του διαγράμματος αυτοσυσχέτισης παρακάτω (Σχ. 5.26), μίας και αποτελεί ένα καλύτερο δείγμα για την αναγνώρισή τόσο της τάσης όσο και της περιοδικότητας.

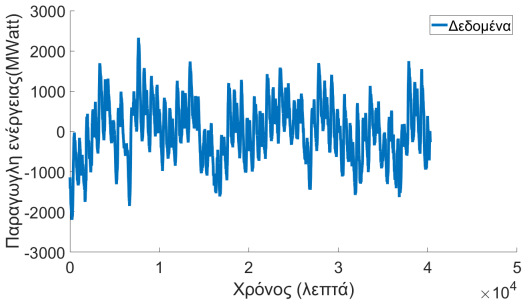


Σχήμα 5.26: Διάγραμμα αυτοσυσχέτισης για την χρονοσειρά ανά 15 λεπτά για την περίοδο του χειμώνα

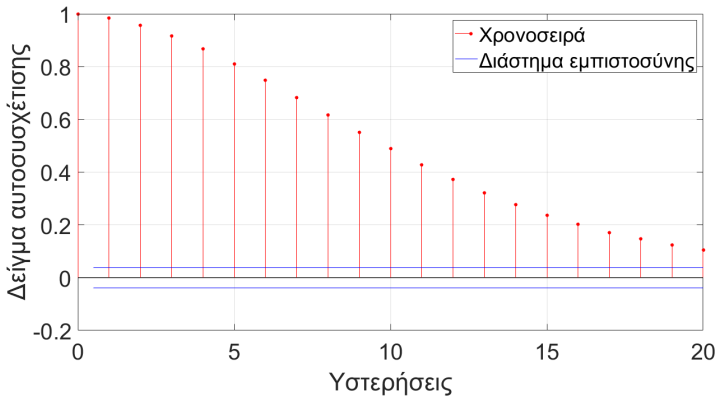
Όπως μπορούμε να διακρίνουμε τα στοιχεία της χρονοσειράς είναι εκτός των ορίων εμπιστοσύνης και επομένως η χρονοσειρά την δεδομένη στιγμή δεν μπορεί να χαρακτηριστεί ως λευκός θόρυβος. Θεωρείται έτσι δεδομένη η ύπαρξη τάσης και περιοδικότητας.

Απαλοιφή του αιτιοκρατικού μέρους (τάση και περιοδικότητα

Αρχικά όπως και στις προηγούμενες περιπτώσεις η απαλοιφή του αιτιοκρατικού μέρους της χρονοσειράς γίνεται με γραμμική παλινδρόμηση με την βοήθεια του λογισμικού της Matlab. Έχουμε καταλήξει σε πρώτου βαθμού τάση και σε δύο περιοδικότητες. Μία ημερησία και μία εβδομαδιαία. Η χρονοσειρά (σχ. 5.27) που προέκυψε μετά την εφαρμογή της εντολής καθώς και το αντίστοιχο διάγραμμα αυτοσυσχέτισης (σχ. 5.28) που θα μας βοηθήσουν να βγάλουμε τα απαραίτητα συμπεράσματα φαίνονται στην συνέχεια.



Σχήμα 5.27: Διάγραμμα χρονοσειράς ανά 15 λεπτά για την χειμερινή περίοδο μετά την απαλοιφή της τάσης και της περιοδικότητας

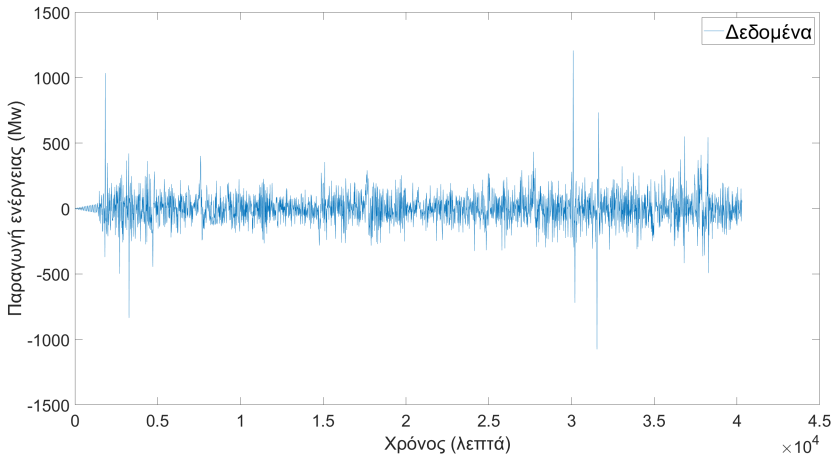


Σχήμα 5.28: Διάγραμμα αυτοσυσχέτισης για χρονοσειρά ανά 15 λεπτά για την χειμερινή περίοδο μετά την απαλοιφή της τάσης και της περιοδικότητας

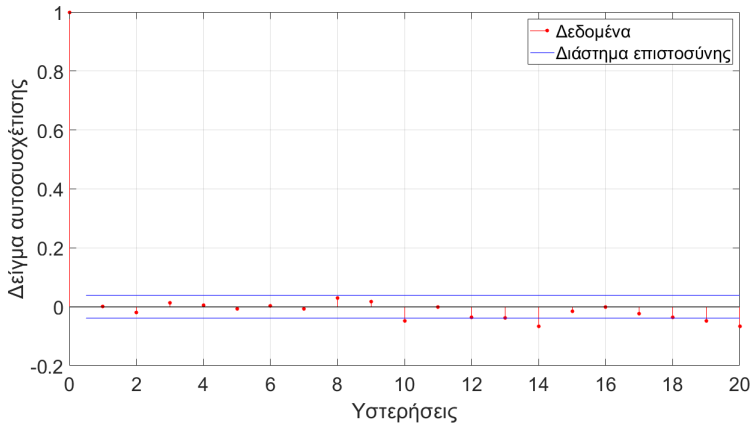
Από τα δύο διαγράμματα που προηγήθηκαν και ιδιαίτερα από το διάγραμμα αυτοσυσχέτισης (Σχ. 5.28), μπορούμε να καταλήξουμε στο συμπέρασμα πως δεδομένου του ότι η τάση και η εποχικότητα έχουν αφαιρεθεί, η χρονοσειρά έχοντας ακόμα τιμές εκτός των ορίων εμπιστοσύνης περιέχει αυτοσυσχετίσης μεταξύ των τιμών. Επομένως δεν μπορεί να χαρακτηρισθεί ως λευκός θόρυβος.

Απαλοιφή του στοχαστικού μέρους με εφαρμογή μοντέλου SARIMA(p,d,q)(P,D,Q)

Όπως και στην περίπτωση της καλοκαιρινής περιόδου έτσι και στην περίπτωση αυτή έχουμε πρώτου βαθμού τάση και δύο περιοδικότητες. Μία ημερήσια και μία εβδομαδιαία. Αυτό συμβαίνει γιατί τόσο μηνιαία όσο και ετήσια δεν μπορεί να υπάρξει λόγο του μικρού όγκου των δεδομένων. Η αναγνώριση των στοιχείων αυτών είναι απαραίτητη για την δυνατότητα εφαρμογής του μοντέλου SARIMA. Για την αντιμετώπιση των αυτοσυσχετίσεων χρησιμοποιείται ένα εποχιακό αυτοπαλινδρομούμενο μοντέλο SARIMA(p,d,q)(P,D,Q). Η επιλογή των συντελεστών και σε αυτήν την περίπτωση γίνεται με την βοήθεια των κριτηρίου AIC. Από την επιλογή του καλύτερου AIC προκύπτει το μοντέλο SARIMA(0, 2, 9)(106 , 2, 9). Τα δυο διαγράμματα που ακολουθούν (σχ. 5.29, σχ. 5.30) δείχνουν την μερική απαλοιφή των αυτοσυσχετίσεων μεταξύ των τιμών αφού ιδιαίτερα στην περίπτωση του διαγράμματος αυτοσυσχέτισης οι τιμές πλέον της χρονοσειράς βρίσκονται σχεδόν μέσα στα όρια εμπιστοσύνης (confidence bounds). Η αυτοσυσχετίσης δεν έχουν απαλοιφή πλήρως. Αυτό φαίνεται από το διάγραμμα αυτοσυσχέτισης (σχ. 5.30). Ακόμα και μετά την εφαρμογή του μοντέλου SARIMA είναι εμφανής η παρουσία τιμών εκτός των ορίων εμπιστοσύνης.



Σχήμα 5.29: Διάγραμμα χρονοσειράς ανά 15 λεπτά για την χειμερινή περίοδο μετά την μερική αφαίρεση των αυτοσυσχετίσεων



Σχήμα 5.30: Διάγραμμα αυτοσυσχέτισης για χρονοσειρά ανά 15 λεπτά για την χειμερινή περίοδο μετά την μερική αφαίρεση του στοχαστικού μέρους

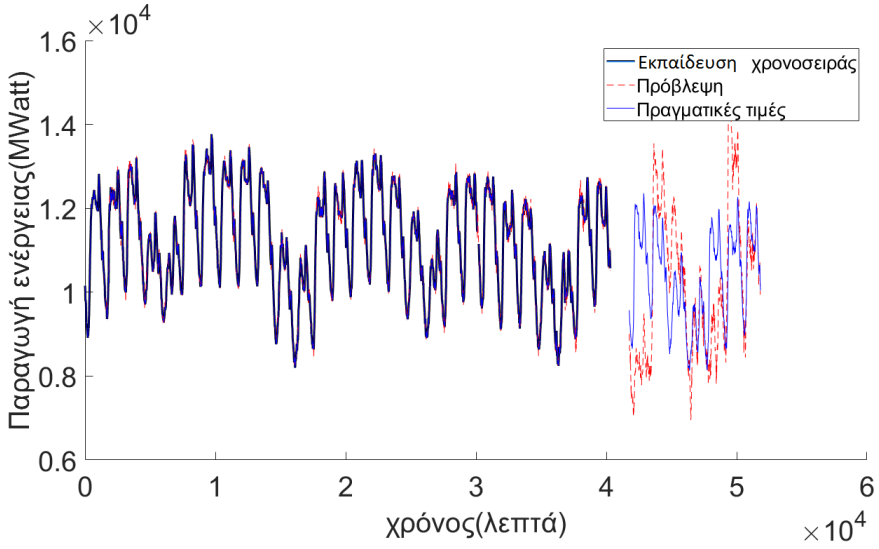
Δημιουργία πρόβλεψης με το μοντέλο SARIMA

Μετά την αφαίρεση του στοχαστικού (αυτοσυσχετίσης) και του αιτιοκρατικού (τάση και περιοδικότητα) μέρους είναι πλέον εφικτή η δημιουργία πρόβλεψης και η εξαγωγή συμπερασμάτων σχετικά με την αποτελεσματικότητα της μεθόδου που χρησιμοποιήθηκε για την πρόβλεψη στο τέταρτο της ώρας.

Για να μπορέσει να γίνει αυτή όπως και στις προηγούμενες περιπτώσεις εκπαιδεύουμε την χρονοσειρά των 2688 δεδομένων με το στοιχείο που αντιστοιχεί στην χρονική στιγμή t_1 έως το στοιχείο την χρονική στιγμή $t_{2687+i+t_1}$. Το i εκφράζει τον αριθμό των προβλέψεων, $i = 1, ..., 672$ που αντιστοιχεί σε μία βδομάδα. Στην συνέχεια γίνεται εκτίμηση της παραγωγής ηλεκτρικής ενέργειας για την χρονική στιγμή $t_{2687+i+96+t_1}$ για μία μέρα μετά (αντιστοιχεί σε 96 τέταρτα). Επιπλέον εκτιμάται η χρονική στιγμή $t_{2687+i+48+t_1}$ για 12 ώρες μετά και η $t_{2688+i+192+t_1}$ για δύο μέρες μετά. Οι συγκρίσεις των αποτελεσμάτων μεταξύ των προβλεπόμενων και τον πραγματικών τιμών γίνονται τόσο σε πίνακες με την βοήθεια των μέτρων επιβεβαίωσης όσο και με την βοήθεια διαγράμματος. Ο πίνακας (Πιν. 5.17) και το διάγραμμα (Σχ. 5.31) μέσω των οποίων έγινε η εξαγωγή των συμπερασμάτων ακολουθούν παρακάτω.

Μέτρα επιβεβαίωσης	**12 ώρες μετά**	**1 μέρα μετά**	**2 μέρες μετά**
RMSE (MWatt)	1.10×10^3	1.63×10^3	3.28×10^3
ME (MWatt)	91.83	156.20	163.73
RPe	0.68	0.53	0.09

Πίνακας 5.17: Πίνακας μέτρων επιβεβαίωσης για την χειμερινή περίοδο για χρονοσειρά ανά 15 λεπτά. Το RMSE είναι το μέσο τετραγωνικό σφάλμα. Το ME είναι το μέσο σφάλμα. Το RPe είναι ο συντελεστής συσχέτισης του Pearson. Και τα τρία μέτρα επιβεβαίωσης αξιολογούν τα αποτελέσματα που προέκυψαν μεταξύ των δεδομένων πρόβλεψης για 12 ώρες μετά, μία μέρα και δύο μέρες μετά με τα γνωστά δεδομένα που διαθέτουμε.



Σχήμα 5.31: Διάγραμμα γνωστών δεδομένων και πρόβλεψης για την χειμερινή περίοδο για χρονοσειρά ανά 15 λεπτά. Το κομμάτι εκπαίδευσης της χρονοσειράς φαίνεται με μαύρη συνεχή γραμμή. Το κομμάτι τις πρόβλεψης φαίνεται με κόκκινη διακεκομμένη. Οι πραγματικές τιμές για την περίοδο της πρόβλεψης φαίνονται με μπλε συνεχή γραμμή

.

Η μη αξιοπιστία των στατιστικών μέτρων φαίνεται τόσο από την μελέτη του πίνακα (Πιν. 5.17) όσο και από αυτή του γραφήματος (Σχ. 5.31). Γίνεται αντιληπτό πως στην περίπτωση του μοντέλου SARIMA η πρόβλεψη που γίνεται για 12 ώρες στο μέλλον με ποσοστό τετραγωνικού σφάλματος 19,75% για εύρος τιμών 5570 (MWatt) είναι σαφώς καλύτερη από τις αντίστοιχες για μία και δύο μέρες στο μέλλον με ποσοστά τετραγωνικού σφάλματος 29,26% και 58,89% αντίστοιχα. Και οι τρεις προβλέψεις που γίνονται με μοντέλο SARIMA για την χειμερινή περίοδο περιέχουν σημαντικά ποσοστά σφαλμάτων. Επομένως οι προβλέψεις που πραγματοποιήθηκαν δεν μπορούν να χαρακτηριστούν αξιόπιστες. Αυτό οφείλεται στην απόκλιση της χρονοσειράς με χρονικό βήμα ανά 15 λεπτά από την κανονική κατανομή όπως αναφέρθηκε και προηγουμένως (εν. 5.2).

Ένας ακόμα σημαντικός παράγοντας είναι η απουσία στασιμότητας από την χρονοσειρά. Η παρουσία ακραίων τιμών στο σχήμα (Σχ. 5.29) καθώς και η εμφανής απουσία ομαλοποίησης σε συνάρτηση με τον χρόνο της χρονοσειράς μετά την αφαίρεση της τάσης και της περιοδικότητας στο σχήμα (Σχ. 5.27) είναι χαρακτηριστικά μίας μη στάσιμης χρονοσειράς. Τέλος η απόκλιση αυτή οφείλεται και στην παρουσία αυτοσυσχετίσεων στην χρονοσειρά.

Κεφάλαιο 6

Συμπεράσματα

Στα πλαίσια της εργασίας αυτής πραγματοποιήθηκε εκτίμηση της παραγωγής ηλεκτρικής ισχύος στην χώρα του Βελγίου χρησιμοποιώντας μοντέλα SARIMA καθώς και την μέθοδο του σταθμισμένου κινούμενου μέσου όρου. Επιπλέον γίνεται προσπάθεια χρονικής παρεμβολής των ήδη υπάρχοντων κενών θέσεων που εμφανίζονται στο σύνολο των δεδομένων το οποίο χρησιμοποιείται για την επεξεργασία. Ο στόχος της μελέτης αυτής είναι η αποτελεσματικά κάλυψη των κενών στοιχείων καθώς και η δημιουργία αξιόπιστων προβλέψεων με μοντέλα SARIMA και με την μέθοδο του κινούμενου σταθμισμένου όρου και η σύγκριση των διαφορετικών μεθόδων που χρησιμοποιήθηκαν.

Το σύνολο των δεδομένων αποτελείται από 26208 στοιχεία τα οποία αντιστοιχούν σε πραγματικές μετρήσεις της παραγωγής απαιτούμενης ηλεκτρικής ισχύος με χρονικό βήμα ανά 15 λεπτά της ώρας και σε ένα χρονικό ορίζοντα 9 μηνών από την 1η Ιανουαρίου του 2019 έως τις 30 Σεπτεμβρίου του 2019 στην επικράτεια του Βελγίου. Η επεξεργασία των δεδομένων έγινε με χρονικό βήμα ανά 15 λεπτά καθώς και με χρονικό βήμα ανά εξάωρο (μέσοι όροι από τέταρτα) και συνολικά 1092 δεδομένα. Στην περίπτωση της ανάλυσης με την χρονοσειρά ανά 15 λεπτά έγινε επιλογή δύο διαφορετικών χρονικών περιόδων.

Έτσι επιλέχθηκε μια καλοκαιρινή περίοδος από 15 Ιουνίου έως 13 Ιουλίου και μία χειμερινή από 15 Ιανουαρίου έως 12 Φεβρουαρίου. Ο διαχωρισμός αυτός έγινε για να μειωθεί το μεγάλο μέγεθος της χρονοσειράς αλλά και λόγω των διαφορετικών ενεργειακών αναγκών που πιθανότατα αναμένουμε στις δύο αυτές διαφορετικές χρονικές περιόδους. Το βασικό πρόβλημα της ανάλυσης στην χρονοσειρά με χρονικό βήμα ανά 15 λεπτά ήταν η απόκλιση που εμφάνισαν τα δεδομένα από την κανονική κατανομή. Η αντιμετώπιση του προβλήματος αυτού έγινε με τον μετασχηματισμό της χρονοσειράς σε μέσους όρους ανά εξάωρα, όπου προέκυψε πως η χρονοσειρά πλέον ακολουθεί την κανονική κατανομή.

Στην επεξεργασία των αποτελεσμάτων που έγινε με την μέθοδο του σταθμισμένου κινούμενου μέσου όρου το σύνολο των δεδομένων που χρησιμοποιήθηκε αποτελούνταν από χρονικά βήματα ανά 15 λεπτά. Κρίθηκε απαραίτητη η χρονική παρεμβολή των δεδομένων, με σκοπό την δημιουργία πρόβλεψης τόσο με τον σταθμισμένο κινούμενο μέσο όρο όσο και με τα μοντέλα SARIMA με χρονοσειρά χωρίς την παρουσία κενών θέσεων. Έτσι επιλέχθηκε να γίνει εφαρμογή δύο διαφορετικών μεθοδολογιών και εν τέλει να επιλεγεί αυτή με τα πιο αξιόπιστα αποτελέσματα. Η αξιολόγηση των αποτελεσμάτων έγινε την βοήθεια τριών βασικών στατιστικών μέτρων επιβεβαίωσης και συγκεκριμένα της ρίζας του μέσου τετραγωνικού σφάλματος (RMSE), του μέσου σφάλματος (ME) και του συντελεστή συσχέτισης Pearson (RPe). Η εφαρμογή των μεθοδολογιών έγινε και για τις δύο περιόδους της χρονοσειράς με χρονικό βήμα ανά 15 λεπτά.

Με την βοήθεια των στατιστικών μέτρων επιβεβαίωσης προέκυψε πως τόσο στην καλοκαιρινή όσο και στην χειμερινή περίοδο οι εκτιμήσεις περιέχουν μη αμελητέα ποσοστά σφάλματος. Έτσι έγινε επιλογή της μεθοδολογίας με τα καλύτερα συγκριτικά μέτρα επιβεβαίωσης. Η πρώτη μεθοδολογία για την χειμερινή περίοδο για παράδειγμα έδωσε έναν συντελεστή συσχέτισης που κυμαίνεται από 21%–39% και ποσοστό του μέσου τετραγωνικού σφάλματος από 11,11% (για εύρος παραθύρου τρία) έως 12,08% (για εύρος παραθύρου 20) για εύρος τιμών

5570 (MWatt). Η δεύτερη μεθοδολογία έδωσε συντελεστή συσχέτισης από 36%–53% και ποσοστό του μέσου τετραγωνικού σφάλματος που κυμαίνεται από 9,91% (για εύρος παραθύρου τρία) έως 10,84% (για εύρος παραθύρου 20) για εύρος τιμών ίσο με 5570 (MWatt). Ο συντελεστής συσχέτισης επηρεάζεται από την παρουσία ιδιόμορφων τιμών. Επομένως δεν θα ληφθεί ιδιαίτερα υπόψιν στην εκπόνηση των συμπερασμάτων. Επομένως η χρονική παρεμβολή θα γίνει με την βοήθεια της δεύτερης μεθοδολογίας πριν την διαδικασία της πρόβλεψης τόσο με το μοντέλο SARIMA όσο και με την μέθοδο του σταθμισμένου κινούμενου μέσου όρου.

Στο κομμάτι της εκτίμησης των τιμών της παραγωγής ηλεκτρικής ισχύος για μελλοντικές χρονικές στιγμές με την μέθοδο του σταθμισμένου κινούμενου μέσου όρου χρησιμοποιήθηκε η χρονοσειρά με χρονικό βήμα ανά 15 λεπτά. Έγινε εκτίμηση των τιμών τόσο για την καλοκαιρινή όσο και για την χειμερινή περίοδο. Η χρονοσειρά αποτελείται από 28 μέρες συνολικά. Έτσι και στις δύο περιπτώσεις εκτιμήθηκε η κατανάλωση της κάθε μέρας με βάση τον σταθμισμένο μέσο όρο της κατανάλωσης μιας περιόδου εύρους 192 λεπτών νωρίτερα για μία βδομάδα μετά. Η αξιολόγηση των αποτελεσμάτων έγινε με την βοήθεια των στατιστικών μέτρων επιβεβαίωσης και συγκεκριμένα μέσω των RMSE, ME και RPe. Έτσι για την περίοδο του χειμώνα το ποσοστό του μέσου τετραγωνικού σφάλματος είναι 13,45% (για βήμα πρόβλεψης μίας ημέρας) του εύρους τιμών 5570 (MWatt), ενώ για την καλοκαιρινή περίοδο είναι 14,01% (για βήμα πρόβλεψης μίας ημέρας) του εύρους τιμών 4545 (MWatt) αντίστοιχα. Προέκυψε έτσι πως η μέθοδος EWMA δίνει αξιόπιστα αποτελέσματα, τα οποία όμως περιέχουν σημαντικά ποσοστά σφάλματος. Συγκεκριμένα το μέσο τετραγωνικό σφάλμα δηλώνει μικρά αλλά μη αμελητέα σφάλματα. Ο συντελεστής συσχέτισης του Pearson επηρεάζεται από την παρουσία ιδιόμορφων τιμών και δεν λαμβάνεται υπόψιν στην αξιολόγηση της μεθόδου.

Με την βοήθεια των μοντέλων SARIMA πραγματοποιήθηκε η διαδικασία

πρόβλεψης τόσο για την χρονοσειρά ανά 15 λεπτά όσο και για την αντίστοιχη με χρονικό βήμα εξάωρο. Όσον αφορά την χρονοσειρά με μέσους όρους ανά εξάωρα έγινε αναγνώριση των στοιχείων που περιείχε το σύνολο των δεδομένων μέσω διαγραμμάτων αυτοσυσχέτισης (ACF) όπως η τάση και η περιοδικότητα. Αρχικά καταλήξαμε σε 2ου βαθμού τάση και σε δύο περιοδικότητες. Μία ημερήσια και μία εβδομαδιαία. Η απαλοιφή αυτών έγινε με την βοήθεια του λογισμικού της Matlab μέσω γραμμικής παλινδρόμησης. Μετά και την απαλοιφή του αιτιοκρατικού μέρους της χρονοσειράς κρίθηκε απαραίτητη η αντιμετώπιση του στοχαστικού. Η απαλοιφή των αυτοσυσχετίσεων πραγματοποιήθηκε με την βοήθεια ενός μοντέλου SARIMA. Έτσι μετά και την αντιμετώπιση του αιτιοκρατικού και του στοχαστικού μέρους η χρονοσειρά μπορεί να χαρακτηριστεί ως λευκός θόρυβος. Γίνεται έτσι εφικτή η δημιουργία πρόβλεψης για μία μέρα μετά, για δύο μέρες μετά και για μία βδομάδα μετά και η αξιολόγηση των αποτελεσμάτων.

Η αξιολόγηση αυτή έγινε με την βοήθεια των RMSE, ME και RPe. Από την παρατήρηση των στατιστικών μέτρων τα οποία κυμαίνονται από 97%–98% για τον συντελεστή συσχέτισης και από 3,69% (για βήμα πρόβλεψης μίας ημέρας) έως 4,43% (για πρόβλεψη μία βδομάδα μετά) του εύρους τιμών 6120 (MWatt) για την ρίζα του μέσου τετραγωνικού σφάλματος καταλήξαμε στο συμπέρασμα πως τα μοντέλα SARIMA πραγματοποιούν αξιόπιστες εκτιμήσεις των μελλοντικών τιμών της παραγωγής ηλεκτρικής ενέργειας στην χρονοσειρά με μέσους όρους ανά εξάωρα.

Όσον αφορά τις εποχιακές προβλέψεις που πραγματοποιήθηκαν με την βοήθεια των μοντέλων SARIMA για την χρονοσειρά με χρονικό βήμα ανά 15 λεπτά ακολουθήθηκε η ίδια διαδικασία με την πρόβλεψη που έγινε στην χρονοσειρά με χρονικό βήμα εξάωρο. Στην προκειμένη περίπτωση παρουσιάστηκαν κάποιες διαφοροποιήσεις σε σχέση με τα προηγούμενα κομμάτια της επεξεργασίας. Αρχικά καταλήξαμε στο συμπέρασμα πως η χρονοσειρά έχει τάση πρώτου

βαθμού και όχι δευτέρου. Επιπλέον έχει δύο περιοδικότητες μία ημερήσια και μία εβδομαδιαία. Επομένως η εξάλειψη της τάσης και της περιοδικότητας έγινε με την βοήθεια της γραμμικής παλινδρόμησης. Μετά την αντιμετώπιση του αιτιοκρατικού μέρους της χρονοσειράς κρίθηκε απαραίτητη η εξάλειψη και του στοχαστικού μέρους αυτής. Αυτό έγινε με την βοήθεια ενός μοντέλου SARIMA τόσο για την καλοκαιρινή όσο και για την χειμερινή περίοδο. Έγινε έτσι εφικτή η δημιουργία πρόβλεψης και για τις δύο χρονικές περιόδους που επιλέχθηκαν.

Για την καλοκαιρινή περίοδο η αξιολόγηση των αποτελεσμάτων έγινε μέσω στατιστικών μέτρων επιβεβαίωσης που κυμαίνονται από 21%–76% για τον συντελεστή συσχέτισης και από 15,62% (για 12 ώρες στο μέλλον) έως 44,22% (για δύο μέρες στο μέλλον) για την ρίζα του μέσου τετραγωνικού σφάλματος για εύρος τιμών 4545 (MWatt). Για την χειμερινή περίοδο οι ίδιοι συντελεστές κυμαίνονται από 9%–68% για τον συντελεστή συσχέτισης και από 19,75% (για 12 ώρες στο μέλλον) έως 58,89% (για δύο μέρες στο μέλλον) για την ρίζα του μέσου τετραγωνικού σφάλματος για εύρος τιμών 5570 (MWatt). Η εκπόνηση των συμπερασμάτων δεν στηρίζεται στον συντελεστή συσχέτισης του Pearson καθόσον αυτός επηρεάζεται από την παρουσία ιδιόμορφων τιμών και αυτοσυσχετίσεων. Έτσι καταλήγουμε στο συμπέρασμα πως η εκτίμηση που έγινε για την καλοκαιρινή περίοδο είναι καλύτερη από την αντίστοιχη για την χειμερινή. Όμως και οι δύο προβλέψεις περιέχουν σημαντικά ποσοστά σφάλματος και επομένως δεν μπορούν να θεωρηθούν αξιόπιστες. Η απόκλιση των τιμών που προσδιορίστηκαν σε σύγκριση με τις πραγματικές τιμές στην χρονοσειρά με χρονικό βήμα ανά 15 λεπτά οφείλεται στην παρουσία μη κανονικής κατανομής των δεδομένων, στην απουσία στασιμότητας της χρονοσειράς που έγινε η επεξεργασία και στην εμφάνιση αυτοσυσχετίσεων.

Από την μελέτη των διάφορων μεθόδων επεξεργασίας που χρησιμοποιήθηκαν καταλήγουμε στο συμπέρασμα πως στην επεξεργασία των δεδομένων που

έγινε στην χρονοσειρά με χρονικό βήμα ανά 15 λεπτά, η μέθοδος του σταθμισμένου κινούμενου μέσου όρου έδωσε πιο αξιόπιστα αποτελέσματα από την πρόβλεψη με τα μοντέλα SARIMA. Επιπλέον η περίοδος του καλοκαιριού έδωσε την δυνατότητα καλύτερης εκτίμησης των δεδομένων σε σύγκριση με την περίοδο του χειμώνα στα μοντέλα SARIMA. Παρόλο αυτά οι εκτιμήσεις και για τις δύο περιόδους έδωσαν σημαντικά ποσοστά σφάλματος. Επιπλέον στην περίπτωση της εκτίμησης με μοντέλα SARIMA στην χρονοσειρά με μέσους όρους ανά εξάωρα οι προβλέψεις ήταν εξαιρετικές με μικρά ποσοστά σφαλμάτων.

6.1 Προτάσεις για μελλοντική έρευνα

Η δοκιμή πρόσθετων μεθοδολογιών της χρονικής παρεμβολής με σκοπό την κάλυψη κενών με μικρότερα ποσοστά σφάλματος παρουσιάζει ενδιαφέρον για μελλοντική έρευνα. Επιπλέον ενδιαφέρον παρουσιάζει η πραγματοποίηση μίας εκτενούς ανάλυσης της στασιμότητας της χρονοσειράς με χρονικό βήμα ανά 15 λεπτά. Η μετατροπή αυτής σε στάσιμη και η επανεκτίμηση των μελλοντικών τιμών με μοντέλα SARIMA. Επιπλέον για την διαδικασία αυτή μπορούν να χρησιμοποιηθούν μη-γραμμικά μοντέλα χρονοσειρών αλλά και νευρωνικά δίκτυα. Τα δεδομένα για την τέλεση της πρόβλεψης αποκτήθηκαν από την επίσημη ιστοσελίδα του ομίλου ELIA. Η πρόβλεψη που έγινε στην εργασία αυτή δεν συγκρίθηκε με την αντίστοιχη του ομίλου. Επομένως παρουσιάζει ιδιαίτερο ενδιαφέρον μια τέτοια πιθανή σύγκριση σε μια μελλοντική έρευνα.

Τέλος μια ενδιαφέρουσα εξέλιξη της έρευνας που πραγματοποιήθηκε στην εργασία αυτή είναι ο υπολογισμός των ποσοστιαίων σημείων P50, P90. Ο υπολογισμός αυτός πραγματοποιείται μέσω προσομοιώσεων. Τα σημεία αυτά εκφράζουν με ποσοστά πόση ενέργεια θα χρειαστεί να παραχθεί τον επόμενο χρόνο βάση των αντίστοιχων αναγκών για κατανάλωση. Συγκεκριμένα το ποσοστιαίο σημείο P50 χρησιμοποιείται για τον προσδιορισμό της ετήσιας μέσης

παραγωγής ηλεκτρικής ενέργειας. Το P50 υποδηλώνει πως η πιθανότητα για υπό-εκτίμηση ή υπέρ-εκτίμηση της προβλεπόμενης τιμής είναι 50% σε βάθος ενός χρόνου. Αντίστοιχα το P90 εκφράζει την ηλεκτρική ενέργεια που πρόκειται να παραχθεί με ποσοστό εμπιστοσύνης 90%. Αυτό σημαίνει πως υπάρχει 90% πιθανότητα για παραγωγή P90 ή περισσότερης ηλεκτρικής ενέργειας σε έναν χρόνο και μόλις 10% για παραγωγή λιγότερης.

Bibliography

[1] Nesreen K Ahmed, Amir F Atiya, Neamat El Gayar, and Hisham El-Shishiny. An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, 29(5-6):594–621, 2010.

[2] Donald WK Andrews and Werner Ploberger. Testing for serial correlation against an arma (1, 1) process. *Journal of the American Statistical Association*, 91(435):1331–1342, 1996.

[3] J Scott Armstrong and Fred Collopy. Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting*, 8(1):69–80, 1992.

[4] Petraki Eleutheria Aspirtakis George, Koumpoulis Spuridon. Linear regression fields of application. 2016.

[5] Charles Ernest Pelham Brooks, Nellie Carruthers, et al. Handbook of statistical methods in meteorology. *Handbook of statistical methods in meteorology.*, 1953.

[6] Joseph. E. Cavanaugh. Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters*, (2):201–208, 1997.

[7] Christopher Chatfield. *The analysis of time series: theory and practice*. Springer, 2013.

[8] Javier Contreras, Rosario Espinola, Francisco J Nogales, and Antonio J Conejo. Arima models to predict next-day electricity prices. *IEEE transactions on power systems*, 18(3):1014–1020, 2003.

[9] David R Cox. Prediction by exponentially weighted moving averages and related methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 23(2):414–422, 1961.

[10] Kouroutzioudi Despoina. Forecasting with costumize local model for time series with trend. Technical report, Aristotle University of Thessaloniki, 2017.

[11] Petridis Dimitrios. Multiple regression and correlation. 2015.

[12] Audrius Dzikevičius and Svetlana Šaranda. Ema versus sma usage to forecast stock markets: the case of s&p 500 and omx baltic benchmark. *Business: Theory and Practice*, 11(3):248–255, 2010.

[13] G.W. Ellis and A.S. Cakmark. Time series modelling of strong ground motion from multiple event earthquakes. *Soil Dynamics and Earthquake Engineering*, 10(1):42–54, January 1991.

[14] W.A. Fuller. *Introduction to Statistical Time Series*. John Wiley and Sons, 1995.

[15] Stauroula Gazi. *Linear time series models and Autoregression*. PhD thesis, 2015.

[16] Kalamvoki Georgia. *Forecasting Methods in Time series and Time series in greek economy*. PhD thesis, 2017.

[17] Margia Georgia. Forecasting and time series analysis. Technical report, Aristotle University of Thessaloniki, 2009.

[18] Seng Hansun. A new approach of moving average method in time series analysis. In *2013 conference on new media studies (CoNMedia)*, pages 1–4. IEEE, 2013.

[19] Dionissios T Hristopulos. *Introduction for probability and statistics for engineers*. Chania, 2016.

[20] R Hyndman. Better acf and pacf plots, but no optimal linear prediction. *Electronic Journal of Statistics [E]*, 8(2):2296–2300, 2014.

[21] Spiridon Katoikas. Statistical arbitrage and trading strategies. 2017.

[22] Terpenzikou Pasxalina Kotsani Olga-Konstantinos. Stohastic time series models: Basic meanings and tools. 2005.

[23] Dimitrios Kugiumtzis. *Data analysis, Computational Physics*. PhD thesis, 2011.

[24] Dimitrios Kugiumtzis. *Time series analysis, Modeling and Statistic*. PhD thesis, 2018.

[25] Nikoletou Kyriakh. Correlation and regression in business and financial sector. 1998.

[26] Ntentis Ioannis Markopoulos Aggelos-Konstantinos. Box-jenkins method in time series analysis. 2015.

[27] DG Mayer and DG Butler. Statistical validation. *Ecological modelling*, 68(1-2):21–32, 1993.

[28] D.C Montgomery, C.L Jennings, and K. Murat. *Introduction to Time Series Analysis and Forecasting*. Willey.

[29] S Sp Pappas, L Ekonomou, D Ch Karamousantas, GE Chatzarakis, SK Katsikas, and P Liatsis. Electricity demand loads modeling using autoregressive moving average (arma) models. *Energy*, 33(9):1353–1360, 2008.

[30] C-K Peng, Shlomo Havlin, H Eugene Stanley, and Ary L Goldberger. Quantification of scaling exponents and crossover phenomena in non-stationary heartbeat time series. *Chaos: an interdisciplinary journal of nonlinear science*, 5(1):82–87, 1995.

[31] J.Brockwell Peter and A.Davis Richard. *Introduction to Time Series and Forecasting*. Springer.

[32] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, (2):461–464, 1978.

[33] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

[34] Ruey S Tsay. *Analysis of financial time series*, volume 543. John wiley & sons, 2005.