Time series project

Panagiotis Souranis

August 6, 2020

Περίληψη

Στην παρούσα εργασία θα ασχοληθούμε με την ανάλυση ιστορικών δεδομένων (χρονοσειρών) ηλεκτρικής ενέργειας της Ιταλίας. Η επικράτεια της Ιταλίας χωρίζεται σε περιοχές / ζώνες ηλεκτρικής ενέργειας και σε κάθε περιοχή καταγράφεται η ζήτηση (demand) και η τιμή (price) της κιλοβατώρας σε ωριαία βάση. Για κάθε μέρα η τιμή καθορίζεται για όλες τις 24 ώρες από την προηγούμενη μέρα. Μας ενδιαφέρει να προβλέψουμε για κάθε ώρα 1,2,...,24, τη ζήτηση της επόμενης μέρας από τα δεδομένα ζήτησης ως και την προηγούμενη μέρα, και το ίδιο για την τιμή. Επίσης μας ενδιαφέρει να γνωρίζουμε τα χαρακτηριστικά του δυναμικού συστήματος ή στοχαστικής διαδικασίας που ορίζει την εξέλιξη τηςζήτησης και τιμής στην επικράτεια της Ιταλίας και σε κάθε περιοχή. Η ανάλυση θα γίνει σε 2 σκέλη. Στο 1ο σκέλος θα ασχοληθούμε με την γραμμική ανάλυση των χρονοσειρών που αντιστοιχούν στην ομάδα μας και στο 2ο σκέλος θα ακολουθήσει η μη γραμμική ανάλυση. Η περιοχή που αναλογεί στην ομάδα μας είναι η βόρεια περιοχή της Ιταλίας και η αντίστοιχη ώρα για είναι 7 το πρωί.

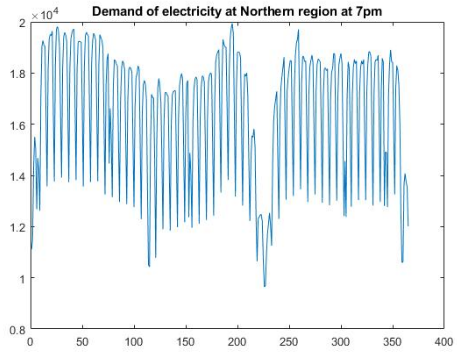
1 Γραμμική ανάλυση

1.1 Χρονοσειρά ζήτησης ηλεκτρικού ρεύματος

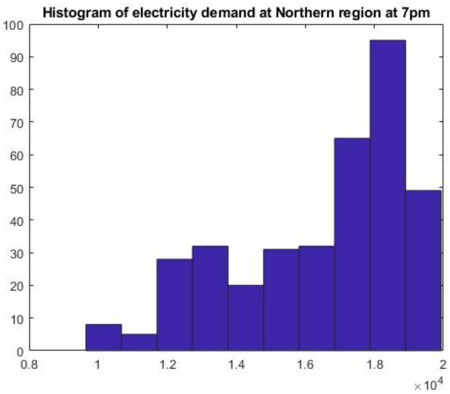
Ας ξεκινήσουμε με μια σύντομη παρουσίαση των δεδομένων που αντιστοιχούν στην ομάδα μας όπως φαίνεται στο Σχήμα 1.

Η χρονοσειρά των δεδομένων ζήτησης ορισμένη ώς στοχαστική διαδικασία έχει μέση τιμή $\mu = 16475.92$ και τυπική απόκλιση $\sigma = 2490.558$. Δεδομένου οτι οι συγκεκριμένες τιμές ειναι αρκετά υψηλές εφαρμόσαμε στα δεδομένα μας μετασχηματισμό λογαρίθμου προκειμένου να τις φέρουμε σε ενα μικρότερο εύρος τιμών και να μπορέσουμε να μειώσουμε την τυπική απόκλιση . Η μέση τιμή και η διακύμανση αφού έχουμε εφαρμόσει μετασχηματισμό λογαρίθμου κειμένονται πλέον στις τιμές $\mu = 9.69698$ και $\sigma = 0.1639382$. Ο πρώτος έλεγχος που διενεργείται κατα κανόνα είναι αυτός των Dickey-Fuller για την υπόθεση της στασιμότητας. Βέβαια απ όσο παρατηρούμε την χρονοσειρά των ιστορικών δεδομένων ζήτησης δε περιμένουμε η χρονοσειρά να είναι στάσιμη καθώς βλέπουμε οτι υπάρχει μια τάση στα δεδομένα. Οι υποθέσεις λοιπόν που εξετάζονται είναι οι παρακάτω:

$$\begin{aligned} H_0 &: \text{A unit root is present in a time series sample} \\ H_1 &: \text{The time series is stationary.} \end{aligned}$$



(α′) Χρονοσειρά ιστορικών τιμών ζήτησης

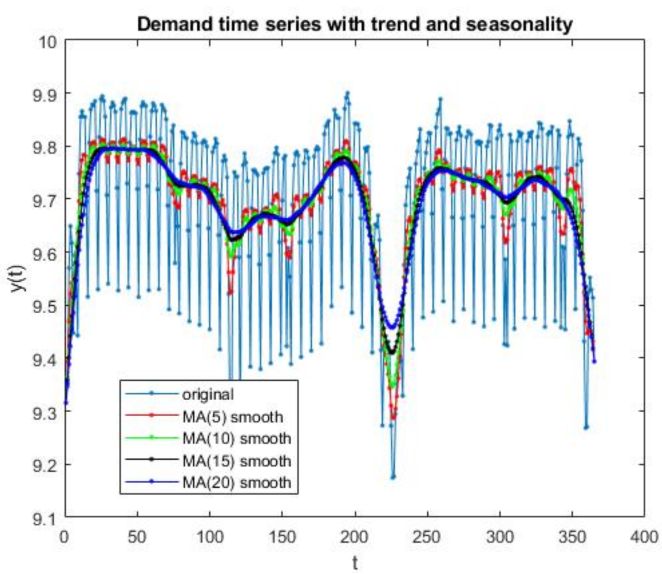


(β′) Ιστόγραμμα των δεδομένων της χρονοσειράς

Σχήμα 1: Περιγραφικά διαγράμματα ιστορικών δεδομένων ζήτησης

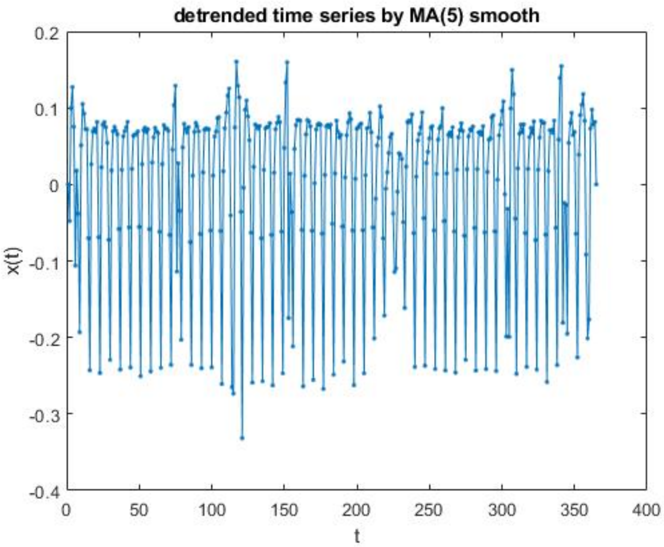
Η τιμή του $p - value$ που αντιστοιχεί στην τιμή του στατιστικού του ελέγχου είναι $p = 0.1920$ και συνεπώς η μηδενική υπόθεση γίνεται δεκτή. Άρα η αρχική χρονοσειρά <u>δεν</u> είναι στάσιμη. Το επόμενο μας βήμα λοιπόν είναι να απαλλέιψουμε την τάση και την περιοδικότητα απο τα δεδομένα.

Ας αρχίσουμε λοιπόν πρώτα απο την τάση. Προκειμένου να απαλείψουμε την τάση θα χρησιμοποιήσουμε 3 τρόπους. Απαλοιφή τάσης με την χρήση φίλτρου κινητού μέσου, απαλοιφή τάσης με την χρήση πολυωνύμου που προσδιορίζει την τάση και τέλος απαλοιφή της τάσης με την χρήση διαφορών k τάξης. Η προσαρμογή των φίλτρων κινούμενου μέσου στην αρχική χρονοσειρά φαίνεται στο Σχήμα 2.



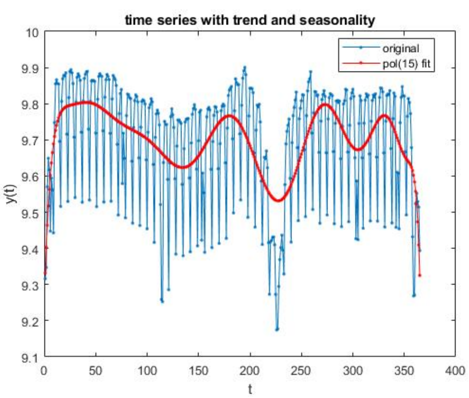
Σχήμα 2: Προσαρμογή φίλτρων κινούμενου μέσου στην αρχική χρονοσειρά

Επιλέχθηκε ως ιδανικό φίλτρο για να εκτιμήσουμε την τάση, το φίλτρο τάξης $q = 5$, καθώς παρατηρούμε όπως φαίνεται και στη χρονοσειρά των υπολοίπων στο Σχήμα 3, να έχει αφαιρεθεί οποιαδήποτε τάση στα δεδομένα.

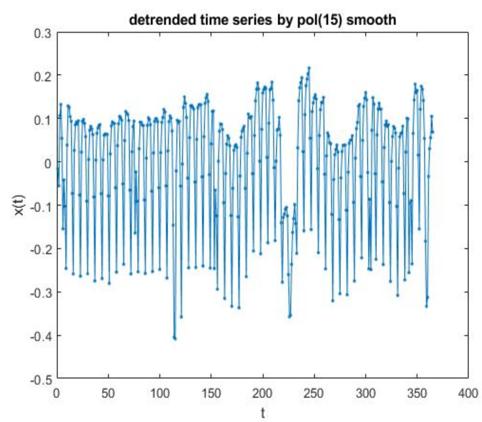


Σχήμα 3: Χρονοσειρά των υπολοίπων μετα απο προσαρμογή σε φίλτρο κινούμενου μέσου τάξης $q = 5$.

Στην συνέχεια έγινε μελέτη για την απαλοιφή της τάσης μέσω πολυωνύμου όπως φαίνεται και στο Σχήμα 4.



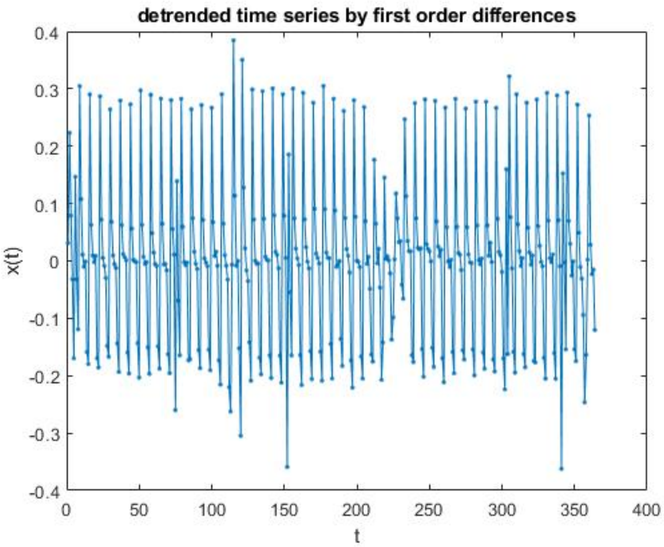
(α΄) Προσαρμογή πολυωνύμου τάξης 15 στα δεδομένα ζήτησης



(β΄) Χρονοσειρά απαλλαγμένη απο τάση μέσω πολυωνύμου

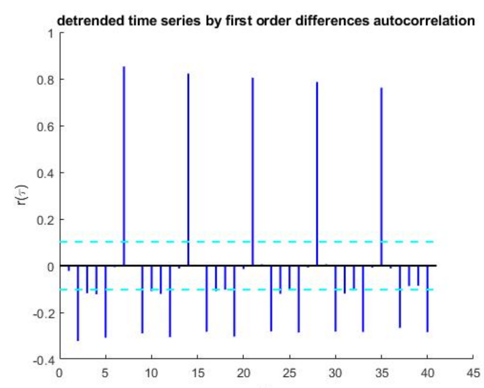
Σχήμα 4: Προσαρμογή και απαλοιφή τάσης μέσω πολυωνύμου τάξης 15

Για την εύρεση της κατάλληλης τάξης του πολυωνύμου έγινε αναζήτηση σε πλέγμα στο διάστημα τιμών $p \in [1, 17]$. Παρ΄ όλα αυτά παρατηρούμε οτι ακόμη και με πολυώνυμο μεγάλης τάξης δέν έχει γίνει καλή προσαρμογή και επομένως απορρίπτουμε την απαλοιφή της τάσης μέσω πολυωνύμου. Τέλος έγινε επίσης εξέταση της απαλοιφής της τάσης μέσω των διαφορών k τάξης, με τα αποτελέσματα να δίνονται παρακάτω στο Σχήμα 5.

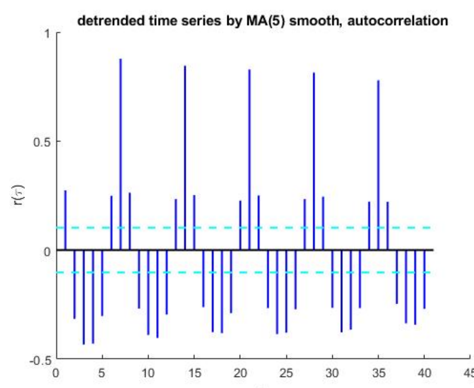


Σχήμα 5: Μετασχηματισμένη χρονοσειρά μέσω των διαφορών 1ης τάξης.

Παρατηρούμε οτι έχει απαλειφθεί ικανοποιητικά η τάση απο τα δεδομένα μας και ο έλεγχος. Έπειτα τα διαγράμματα των αυτοδιασπορών και για τις 2 περιπτώσεις παριστάνται στο Σχήμα 6.



(α΄) Διάγραμμα αυτοσυσχέτισης (πρώτες διαφορές)



(β΄) Διάγραμμα αυτοσυσχέτισης $MA(5)$

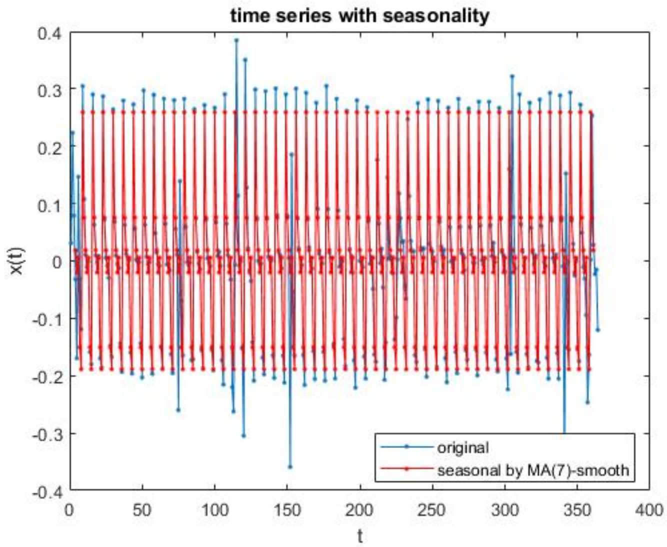
Σχήμα 6: Διάγραμμα αυτοσυσχετίσεων μετα την απαλοιφή της τάσης.

Όπως βλέπουμε η εποχικότητα είναι ξεκάθαρη στα διαγράμματα της αυτοσυσχέτισης καθώς παρουσιάζονται μέγιστα κάθε 7 χρονικές υστερήσεις, κάτι που είναι αναμενόμενο καθώς περιμέναμε να υπάρχει περιοδική συμπεριφορά κάθε εβδομάδα. Επιλέγουμε λοιπόν ώς βέλτιστο τρόπο απαλοιφή της τάσης τις διαφορές πρώτης τάξης και στην συνέχεια επιθυμούμε να απαλοίψουμε και την εποχική συμπεριφορά. Οι δύο προσεγγίσεις που θα ακολουθήσουμε είναι

1. Απαλοιφή περιοδικότητας μέσω φίλτρου κινούμενο μέσου

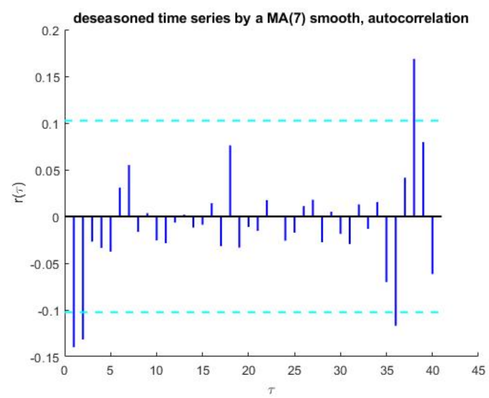
2. Απαλοιφή περιοδικότητας μέσω διαφορών k τάξης, όπου k η περιοδικότητα της χρονοσειράς.

Στο Σχήμα 7 δίνεται η προσαρμογή φίλτρου κινούμενο μέσου για την εύρεση του περιοδικού στοιχείου.

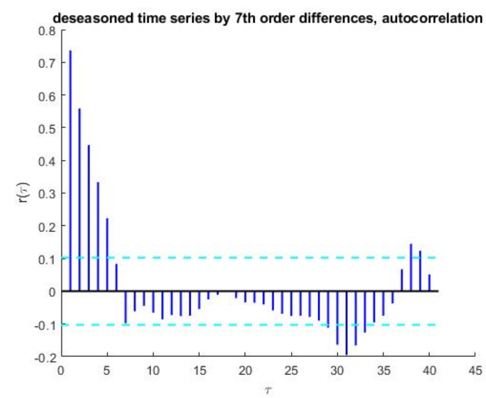


Σχήμα 7: Προσαρμογή φίλτρου κινούμενο μέσου για την εύρεση του περιοδικού στοιχείου.

Οι αυτοσυσχετίσεις που προκύπτουν για την χρονοσειρά απαλλαγμένη απο το περιοδικό στοιχείο μέσω των 2 τρόπων που αναφέραμε παραπάνω δίνονται στο Σχήμα 8.



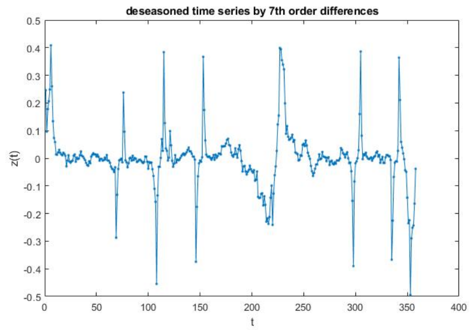
(α΄) Διάγραμμα αυτοσυσχέτισης (MA(7))



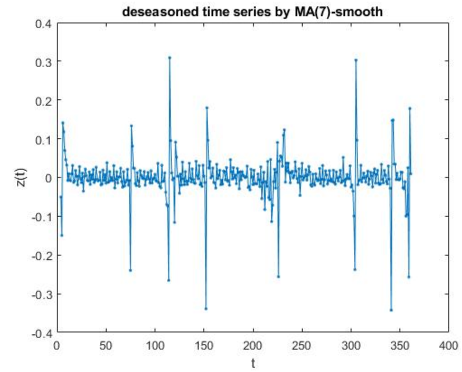
(β΄) Διάγραμμα αυτοσυσχέτισης (διαφορές 7ης τάξης)

Σχήμα 8: Διάγραμμα αυτοσυσχετίσεων μετα την απαλοιφή της περιοδικότητας.

Είναι σκόπιμο να τονίσουμε εδώ οτι οι διαφορές 7ης τάξης έγιναν στην αρχική χρονοσειρά και όχι σε αυτήν που προέκυψε απο την απαλοιφή της τάσης (σε αντίθεση με το φίλτρο κινούμενο μέσου που εφαρμόστηκε στην χρονοσειρά απο την οποία έχει απαλοιφθεί η τάση). Ο λόγος για τον οποίο έγινε αυτό είναι διότι αν η αρχική χρονοσειρά έχει κυρίαρχο το περιοδικό στοιχείο τότε η χρονοσειρά που θα προέκυπτε απο την απαλοιφή του περιοδικού στοιχείου με αυτόν τον τρόπο ενδεχομένως να ήταν στάσιμη.
Βέβαια αυτή η υπόθεση απορρίφθηκε καθώς όπως βλέπουμε στα παρακάτω διαγράμματα στο Σχήμα 9, μετα την απαλοιφή του περιοδικού στοιχείο, η χρονοσειρά που προκύπτει απο τις διαφορές 7ης τάξης φαίνεται να έχει τάση. Επομένως για να αφαιρεθεί και η τάση απο την χρονοσειρά που προέκυψε απο τις διαφορές 7ής τάξης, εφαρμόσαμε επιπλέον διαφορές πρώτης τάξης και τα αποτελέσματα για το Ljung-Box test για το εάν μπορούν οι μετασχηματισμένες χρονοσειρές να θεωρηθούν λευκός θόρυβος δίνονται στο Σχήμα 10.

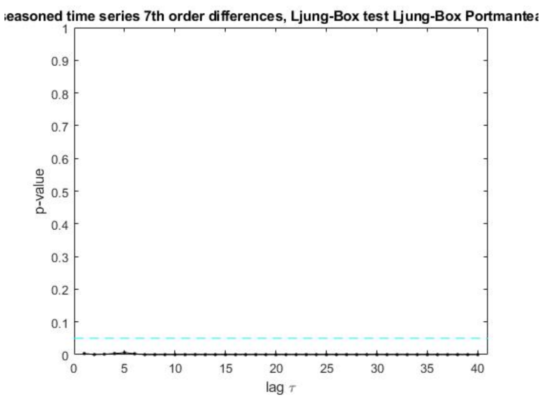


(α′) Χρονοσειρά που προέκυψε απο διαφορές 7ης τάξης

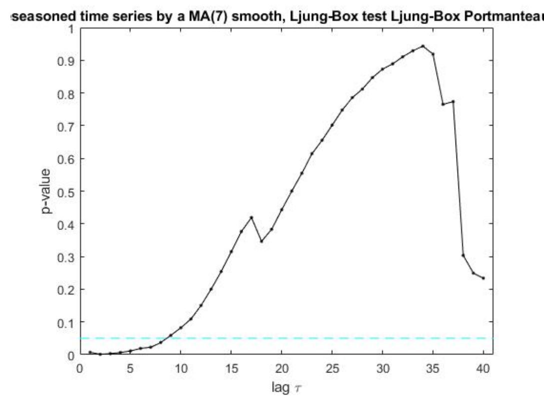


(β′) Χρονοσειρά που προέκυψε απο φίλτρου κινούμενου μέσου

Σχήμα 9: Χρονοσειρά απαλλαγμένη απο εποχικό στοιχείο.



(α′) Ljung-Box test for differenced time serie (1st order and 7th order differences)



(β′) Ljung-Box test for first order differences and MA(7) seasonal filter

Σχήμα 10: Ljung Box tests.

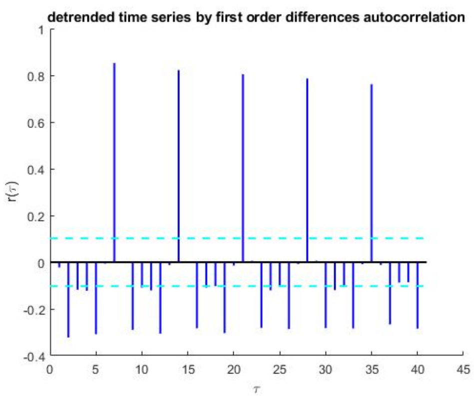
Το τέστ Ljung-Box διεξάγει τον παρακάτω έλεγχο:

$$H_0 \;:\; \text{The data are independently distributed.}$$
$$H_1 \;:\; \text{The data are not independently distributed; they exhibit serial correlation.}$$

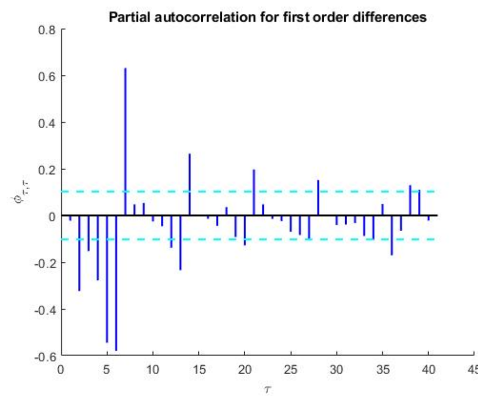
Είναι φανερό λοιπόν ότι στην πρώτη περίπτωση η χρονοσειρά που προκύπτει δεν μπορεί να θεωρηθεί λευκός θόρυβος καθώς για όλες τις υστερήσεις η τιμή του $p - value$ είναι κάτω απο το όριο σημαντικότητας 0.05 και συνεπώς η μηδενική υπόθεση απορρίπτεται. Αντιθέτως στην 2η περίπτωση βλέπουμε ότι για υστέρηση μεγαλύτερη του 10 βλέπουμε οτι τα δεδομένα μπορούν να θεωρηθούν ασυσχέτιστα, κατι το οποίο είναι εμφανές και στο διάγραμμα των αυτοσυσχετίσεων καθώς οι αυτοσυσχετίσεις που προκύπτουν ειναι οι περισσότερες κάτω απο το όριο σημαντικότητας. Βάση αυτου του αποτελέσματος, μπορούμε να συμπεράνουμε ότι μια στοχαστική διαδικασία που θα μπορούσε να περιγράψει την μετασχηματισμένη χρονοσειρά θα ήταν η $X_t = S_t + \varepsilon_t$, όπου $S_t = \dfrac{1}{k}\sum_{j=1}^{k} X_{t+jd}$, όπου $k = \left[n/d\right]$ ο αριθμός των περιόδων

στην χρονοσειρά. Εφόσον $X_t = \log(Y_t) - \log(Y_{t-1})$, αντικαθιστώντας θα έχουμε ότι $\log(Y_t) = \log(Y_{t-1}) + S'_t + \varepsilon_t$, όπου $S'_t = \dfrac{1}{k}\sum_{j=1}^{k}(\log(Y_{t+jd}) - \log(Y_{t+jd-1}))$.
Στην συνέχεια θα επιχειρήσουμε να βρούμε το βέλτιστο μοντέλο που να περιγράφει την χρονοσειρά. Όπως είδαμε στο Σχήμα 9 με την απαλοιφή της εποχικότητας με τις διαφορές 7ης τάξης, φαίνεται να έχει εξαλειφθεί η αργή τάση απο τα δεδομένα εκτός απο ένα κομμάτι μεταξύ των τιμών 150 και 250 το οποίο όπως είδαμε στην αρχική χρονοσειρά στο Σχήμα 1 φαίνεται να είναι μια απεριοδική ταλάντωση οπότε πιθανολογούμε οτι πρόκειται για μία κυκλική συμπεριφορά.
Εφόσον στο Σχήμα 9 φαίνεται ξεκάθαρα μια περιοδική συμπεριφορά πιθανολογούμε οτι ένα μοντέλο SARMA θα ήταν κατάλληλο για να περιγράψει τα δεδομένα. Επιστρέφουμε λοιπόν στην χρονοσειρά πρώτων διαφορών (εφόσον η χρονοσειρά των πρώτων διαφορών δεν παρουσιάζει τάση παρα μόνο εποχικότητα) που φαίνεται στο Σχήμα 5 για να διερευνήσουμε κατάλληλο μοντέλο.

Τα διαγράμματα αυτοσυσχέτισης και μερικής αυτοσυσχέτισης φαίνονται στο Σχήμα 11.



(α′) Διάγραμμα αυτοσυσχετίσεων

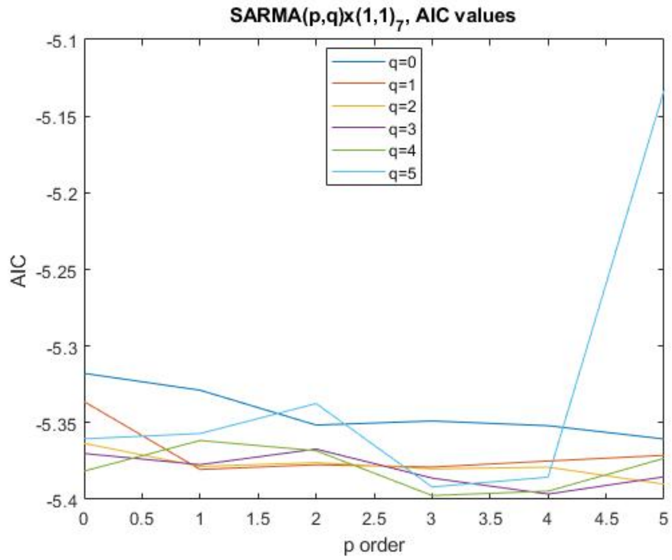


(β′) Διάγραμμα μερικών αυτοσυσχετίσεων

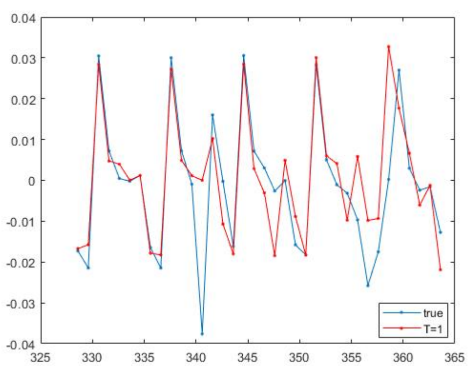
Σχήμα 11: Αυτοσυσχετίσεις

Συνεπώς θα εξετάσουμε όλους τους πιθανούς συνδυασμούς για $p, q \in [0, 5]$ και $P, Q \in [0, 5]$. Επιστρέφουμε λοιπόν στην χρονοσειρά των διαφορών 1ης τάξης για την εύρεση του καλύτερου μοντέλου. Για λόγους συντομίας παραθέτουμε τα καλύτερα αποτελέσματα που βρέθηκαν για $P = 1, Q = 1$ στο Σχήμα 12.

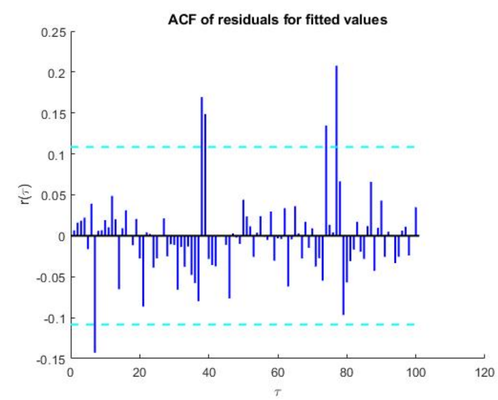
Όπως είναι φανερό οι βέλιστες τιμές AIC είναι όλες αρκετά κοντά και επιλέγουμε το απλούστερο μοντέλο που δίνεται για $p = 3, q = 4$. Προσαρμόζουμε λοιπόν το μοντέλο $\mathrm{SARMA}(3, 4)\times(1, 1)_7$ και η τιμή του AIC που προκύπτει είναι $\mathrm{AIC} = -5.3955$. Η τυπική απόκλιση του θορύβου στις fitted τιμές είναι $\sigma_\varepsilon = 0.006833$ ενώ οι τιμές τών NRMSE, FPE είναι $\mathrm{NRMSE} = 0.3959, \mathrm{FPE} = 0.000049$. Η τιμή του σφάλματος για το σύνολο αξιολόγησης το οποίο ορίστηκε να είναι το 10% της αρχικής χρονοσειράς προέκυψε να είναι $\mathrm{NRMSE} = 0.6360$. Στο Σχήμα 13 παριστόνται οι προβλέψεις στο σύνολο αξιολόγησης, τα υπόλοιπα που προέκυψαν για τις προβλέψεις στο σύνολο εκπαίδευσης, καθώς και το αντίστοιχο τους ιστόγραμμα, διάγραμμα αυτοσυσχέτισης και Quantile-Quantile (QQ)-plot.



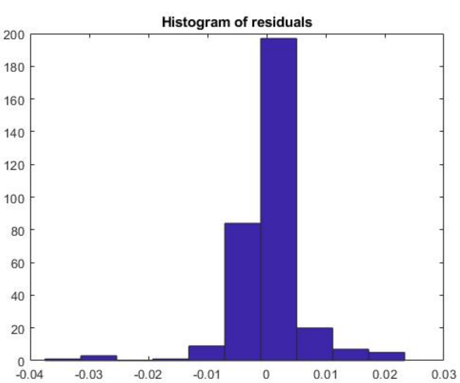
Σχήμα 12: Τιμές AIC για $P = 1, Q = 1$.



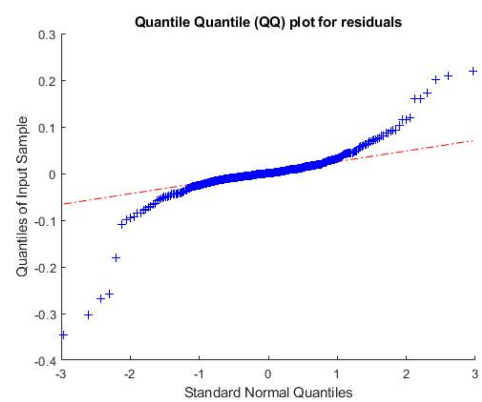
(α′) Forecasts on test set



(β′) ACF residuals on fitted values



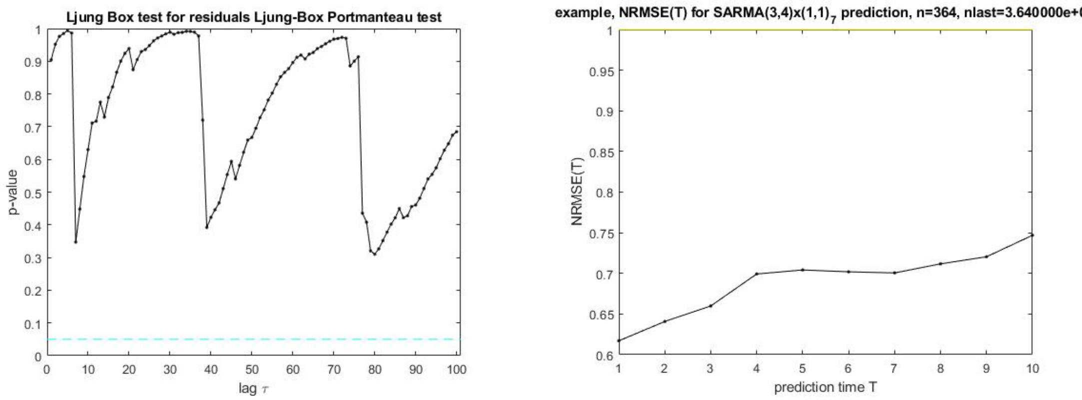
(γ′) Histogram residuals on fitted values



(δ′) QQ plot residuals on fitted values

Σχήμα 13: Diagnostics of residuals

Όπως ειναι φανερό όλες οι αυτοσυσχετίσεις εκτός απο 5 βρίσκονται μέσα στα όρια σημαντικότητας. Το τεστ Ljung-Box στα υπόλοιπα δίνει για όλες τις υστερήσεις μη σημαντικές αυτοσυσχετίσεις όπως φαίνεται στο Σχήμα 14. Ακόμη παρουσιάζονται και τα σφάλματα πρόβλεψης ανάλογα με τον ορίζοντα πρόβλεψης.



(α′) Ljung-Box test on residuals

(β′) Σφάλμα πρόβλεψης ανάλογα του χρόνου τ.

Σχήμα 14: Ljung-Box test and prediction error.

Μπορούμε να συμπεράνουμε λοιπόν οτι τα υπόλοιπα μας είναι λευκός θόρυβος και έχουμε κάνει μια καλή προσαρμογή με το μοντέλο μας. Τέλος οι συντελεστές του μοντέλο που περιγράφει την μετασχηματισμένη με διαφορές 7ης τάξης, χρονοσειρά X_t δίνονται στους Πίνακες 1-2.

ϕ_0	ϕ_1	ϕ_2	ϕ_3	ϕ_7	ϕ_8	ϕ_9	ϕ_{10}
0	2.1275	-1.4924	0.2733	0.9892	-2.1103	1.4744	-0.2703

Πίνακας 1: AR coefficients.

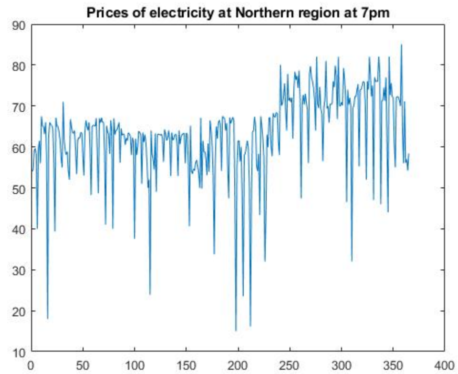
θ_1	θ_2	θ_3	θ_4	θ_7	θ_8	θ_9	θ_{10}	θ_{11}
2.3798	-1.8354	0.3172	0.0875	0.6997	-1.6198	1.2280	-0.1618	-0.0967

Πίνακας 2: MA coefficients.

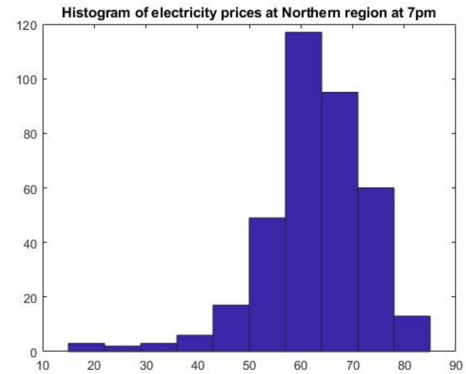
Τέλος το Lilliefors test το οποίο ελέγχει την μηδενική υπόθεση ότι τα δεδομένα προέρχονται απο κανονική κατανομή, δίνει για τα υπόλοιπα $p - value$, $p = 0.0001$ και συνεπώς απορρίπτεται η υπόθεση της κανονικής κατανομής των υπολοίπων στα fitted δεδομένα.

1.2 Χρονοσειρά τιμών ηλεκτρικού ρεύματος

Όπως και πριν, θα επαναλάβουμε την ίδια διαδικασία για την χρονοσειρά τιμής ηλεκτρικού ρεύματος αρχίζοντας με την επισκόπηση της χρονοσειράς στο Σχήμα 15.



(α′) Χρονοσειρά τιμών ηλεκτρικού ρεύματος.

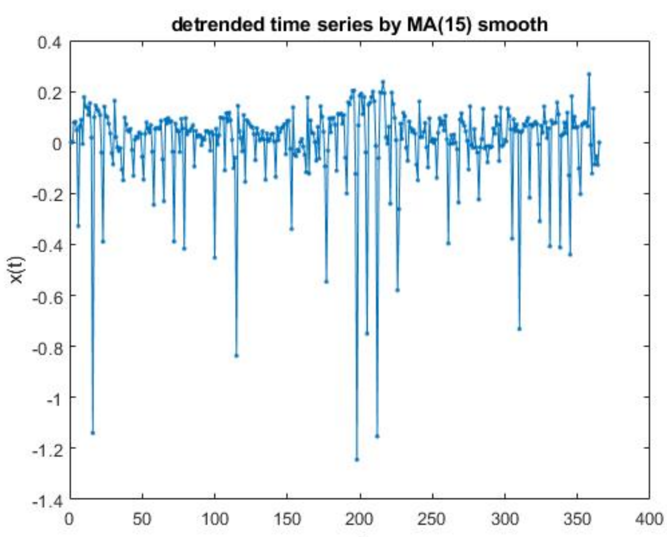


(β′) Ιστόγραμμα των δεδομένων της χρονοσειράς.

Σχήμα 15: Περιγραφικά διαγράμματα ιστορικών δεδομένων των τιμών

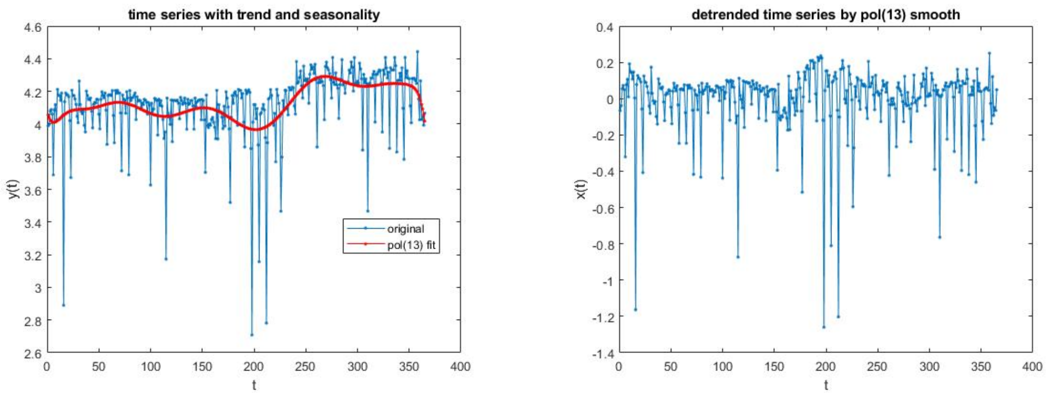
Όπως και πρίν, έτσι και τώρα διενεργούμε τον έλεγχο Dickey-Fuller για την υπόθεση της στασιμότητας. Η τιμή του $p - value$ αυτή τη φορά είναι $p = 0.0859$. Επομένως γίνεται δεκτή η μηδενική υπόθεση οτι σε μια αυτοπαλίνδρομη διαδικασία προσαρμοσμένη στα δεδομένα της χρονοσειράς βρίσκεται μοναδιαία ρίζα και συνεπώς η υπόθεση της στασιμότητας απορρίπτεται για επίπεδο σημαντικότητας $a = 0.05$. Δεδομένου οτι παρατηρούμε πολυ υψηλές εξάρσεις, χρησιμοποιούμε μετασχηματισμό λογαρίθμου.

Παρατηρούμε οτι υπάρχει τάση στα δεδομένα και θα ξεκινήσουμε λοιπόν πρώτα με την απαλοιφή της δοκιμάζοντας τις ίδιες μεθόδους με πρίν. Όπως και προηγουμένως θεωρήσαμε καλύτερη προσαρμογή αυτήν με φίλτρο κινούμενου μέσου τάξης 5 βάση των υπολοίπων που παρουσιάζονται στο Σχήμα 16.



Σχήμα 16: Χρονοσειρά των υπολοίπων μετα απο προσαρμογή σε φίλτρο κινούμενου μέσου τάξης $q = 5$.

Όπως και πριν έτσι και τώρα διερευνούμε την απαλοιφή της τάσης μέσω της χρήσης ενός πολυωνύμου. Η καλύτερη προσαρμογή έγινε για πολυώνυμο τάξης 13 και τα αποτελέσματα φαίνονται στο Σχήμα 17.

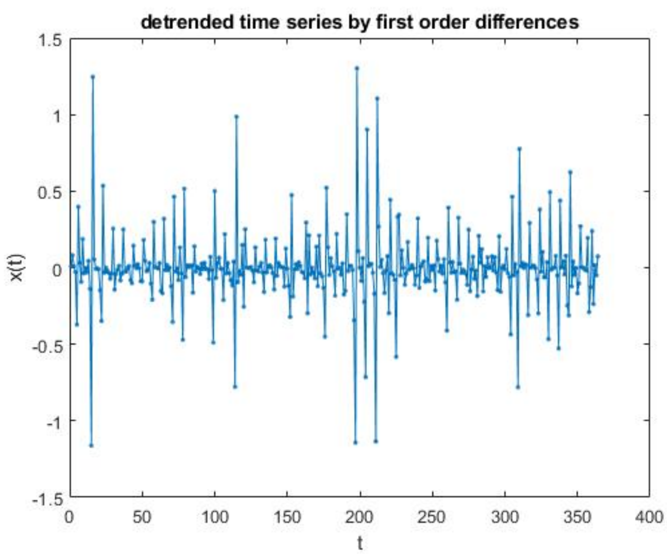


(α′) Προσαρμογή πολυωνύμου τάξης 13 στα δεδομένα τιμών

(β′) Χρονοσειρά απαλλαγμένη απο τάση μέσω πολυωνύμου

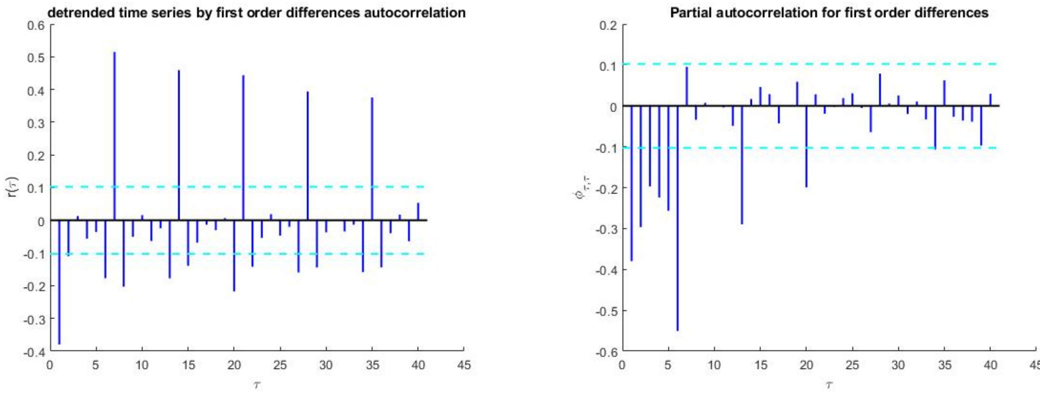
Σχήμα 17: Προσαρμογή και απαλοιφή τάσης μέσω πολυωνύμου τάξης 13.

Τέλος για την απαλοιφή της τάσης χρησιμοποιήθηκαν διαφορές πρώτης τάξης και τα αποτελέσματα της μετασχηματισμένης χρονοσειράς φαίνονται στο Σχήμα 18.



Σχήμα 18: Μετασχηματισμένη χρονοσειρά μέσω των διαφορών 1ης τάξης.

Όπως ειναι φανερό ο μετασχηματσμός με διαφορές πρώτης τάξης παρέχει τα καλύτερα αποτελέσματα για την απαλοιφή της τάσης. Ακόμη παρατηρούμε peaks που επαναλαμβάνονται ανα τακτά χρονικά διαστήματα, οπότε συμπαιρένουμε οτι επικρατεί εποχικότητα στην απαλλαγμένη απο την τάση χρονοσειρά κάτι που φαίνεται και ξεκάθαρα στο διάγραμμα των συσχετίσεων και αυτοσυσχετίσεων στο Σχήμα 19.



(α′) Διάγραμμα αυτοσυσχέτισης (πρώτες διαφορές)

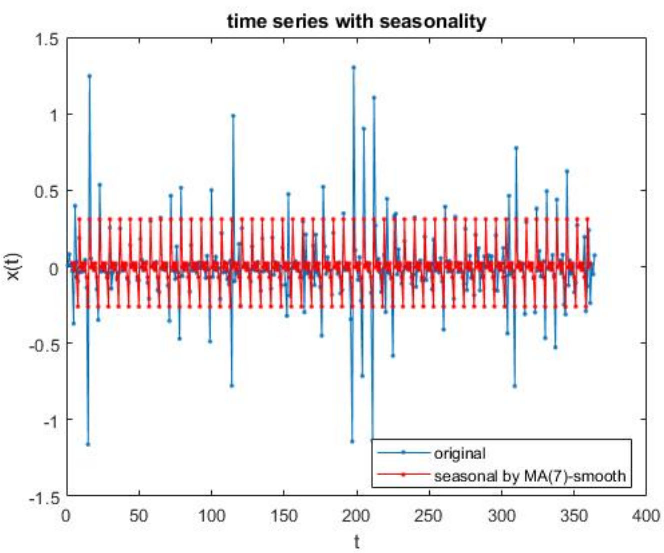
(β′) Διάγραμμα μερικής αυτοσυσχέτισης (πρώτες διαφορές)

Σχήμα 19: Διάγραμμα αυτοσυσχετίσεων μετα την απαλοιφή της τάσης.

Παρατηρούμε οτι έχουμε μέγιστα στο διάγραμμα αυτοσυσχετίσεων κάθε 7 χρονικές υστερήσεις (όπως ακριβώς και με την χρονοσειρά των τιμών ζήτησης). Αυτό βέβαια ήταν αναμενόμενο απο την φύση των δεδομένων.

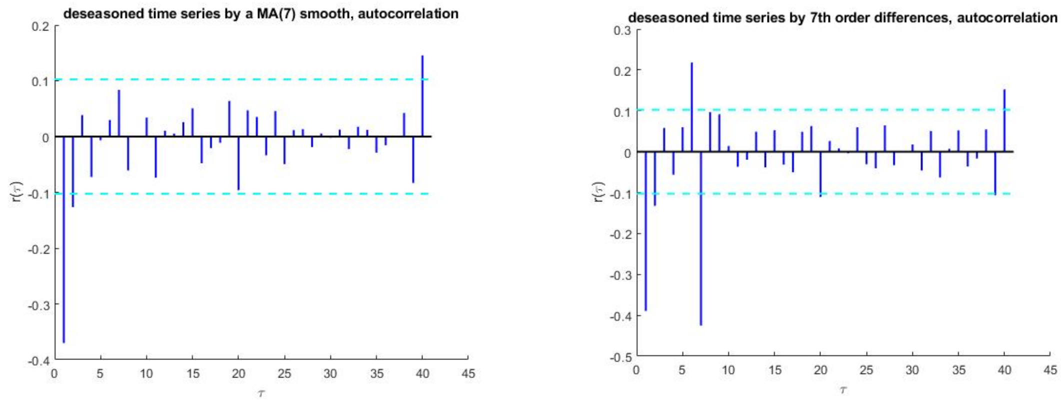
Στην συνέχεια θα επιχειρήσουμε να απαλείψουμε την περιοδικότητα με χρήση των διαφορών 7ης τάξης καθώς και με την χρήση φίλτρου κινούμενο μέσου για περιοδικότητες.

Η προσαρμογή φίλτρου κινούμενου μέσου τάξης 7 για εποχικότητες δίνεται στο Σχήμα 20.



Σχήμα 20: Προσαρμογή φίλτρου κινούμενου μέσου για την εύρεση του περιοδικού στοιχείου.

Παρακάτω παρατίθενται στο Σχήμα 21 η συνάρτηση αυτοσυσχέτισης για την απαλλαγμένη απο περιοδικότητα χρονοσειρά μέσω των διαφορών 7ης τάξης και μέσω της χρήσης του φίλτρου κινούμενου μέσου.

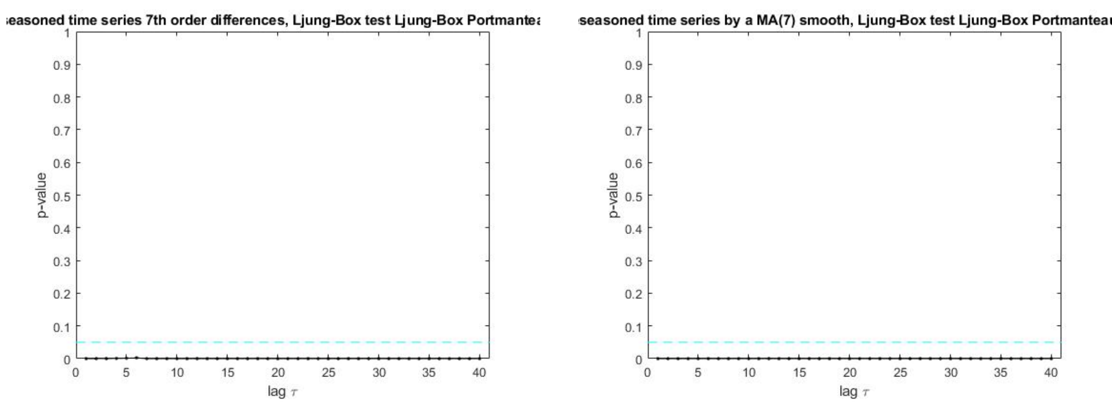


(α′) Διάγραμμα αυτοσυσχέτισης MA(7)

(β′) Διάγραμμα μερικής αυτοσυσχέτισης (διαφορές 7ης τάξης)

Σχήμα 21: Διάγραμμα αυτοσυσχετίσεων μετα την απαλοιφή της περιοδικότητας.

Όπως είναι φανερό η περιοδικότητα έχει εξαλειφθεί πλήρως μέσω της χρήσης του φίλτρου κινούμενο μέσου καθώς δεν παρατηρούμε πλέον μέγιστο στην έβδομη υστέρηση, κάτι που φαίνεται να υπάρχει μέσω της χρήσης των διαφορών 7ης τάξης. Όπως φαίνεται στα διαγράμματα αυτοσυσχετίσεων έχουμε υψηλές αυτοσυσχετίσεις στο αριστερά για υστέρηση 1 και στο δεξιά για υστέρηση 7 κάτι που οδηγεί στο συμπέρασμα οτι η χρονοσειρά που προκύπτει έπειτα και απο την απαλοιφή της περιοδικότητας δέν ειναι λευκός θόρυβος. Αυτο επιβεβαιώνεται και το τεστ Ljung-Box που απορρίπτει την μηδενική υπόθεση για όλες τις υστερήσεις και στις 2 περιπτώσεις όπως φαίνεται παρακάτω στο Σχήμα 22.

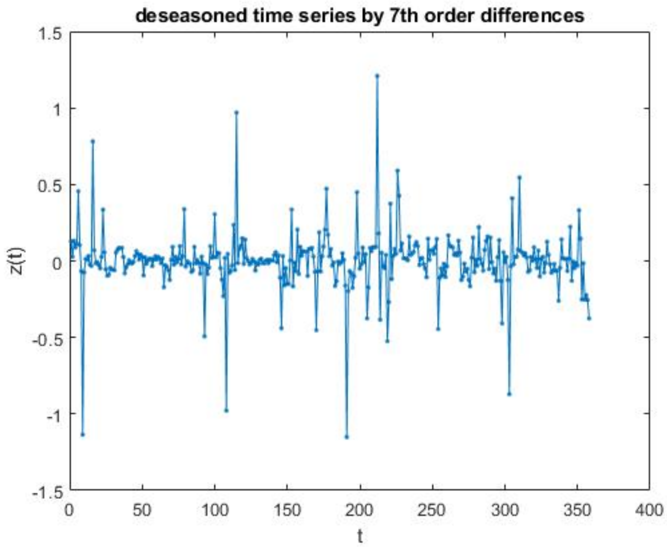


(α′) Ljung-Box test for differenced time serie (first and 7th order differences)

(β′) Ljung-Box test for deseasoned time serie with MA(7) filter

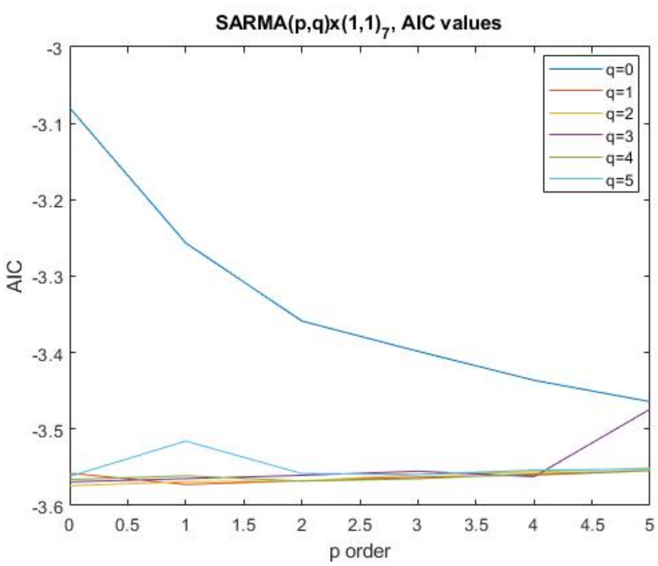
Σχήμα 22: Ljung-Box tests

Προκειμένου να μπορέσουμε να αποφανθούμε την μορφή του μοντέλου που θα χρησιμοποιήσουμε είναι σκόπιμο να δούμε την μετασχηματισμένη χρονοσειρά με τις διαφορές 7ης τάξης στο Σχήμα 23 για να συμπεράνουμε εάν έχουμε συσχέτιση μεταξύ των εποχικών κύκλων.



Σχήμα 23: Μετασχηματισμένη χρονοσειρά με διαφορές 7ης τάξης

Όπως βλέπουμε παρατηρείται έντονα το εποχικό στοιχείο, άρα απο αυτό συμπαιρένουμε ότι η μορφή του μοντέλου που θα περιέγραφε την χρονοσειρά, θα ήταν ένα μοντέλο τύπου SARMA. Επιστρέφουμε στην χρονοσειρά πρώτων διαφορών για να αναζητήσουμε για όλες τις πιθανές περιπτώσεις $p, q \in [0, 6]$ και $P, Q \in [0, 6]$ για την εύρεση του καλύτερου μοντέλου. Το βέλιστο μοντέλο δίνεται για εποχικές παραμέτρους $P = 1, Q = 1$. Παρακάτω στο Σχήμα 24 δίνεται το AIC διάγραμμα για αυτές τις τιμές μόνο για συντομία.



Σχήμα 24: Τιμές AIC για $P = 1, Q = 1$.

Βλέπουμε λοιπόν οτι το βέλτιστο μοντέλο κρίνοντας και απο τις παραμέτρους p, q επιτυγχάνεται για $p = 2, q = 4$. Προσαρμόζουμε λοιπόν το μοντέλο SARMA$(2, 4) \times (1, 1)_7$ και οι τιμές που προκύπτουν ειναι οι εξής: AIC $= -3.5680$, FPE $= 0.0282$, $\sigma_\varepsilon = 0.1635$. Η τιμή του σφάλματος για το σύνολο αξιολόγησης το οποίο ορίστηκε να ειναι το 10% της αρχικής χρονοσειράς προέκυψε να είναι NRMSE $= 0.5830$, ενώ για το σύνολο εκπαίδευσης NRMSE $= 0.6213$.

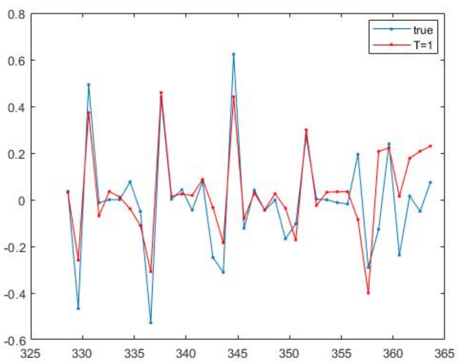
Στο Σχήμα 25 παριστόνται οι προβλέψεις στο σύνολο αξιολόγησης, τα υπόλοιπα που προέκυψαν για τις προβλέψεις στο σύνολο εκπαίδευσης, καθώς και το αντίστοιχο τους ιστόγραμμα, διάγραμμα αυτοσυσχέτισης και Quantile-Quantile (QQ)-plot. Οι συντελεστές του μοντέλου παραδίδονται στους Πίνακες 3-4.

ϕ_0	ϕ_1	ϕ_2	ϕ_7	ϕ_8	ϕ_9
0	0.1603	-0.9672	0.9207	-0.1613	0.8899

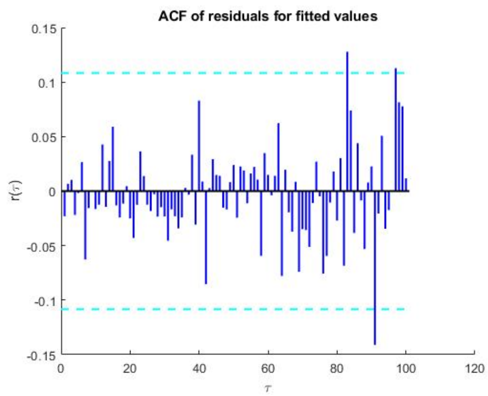
Πίνακας 3: AR coefficients.

θ_1	θ_2	θ_3	θ_4	θ_7	θ_8	θ_9	θ_{10}	θ_{11}
0.9522	-0.9858	0.7853	0.1578	0.6685	-0.6975	0.7153	-0.5482	-0.0502

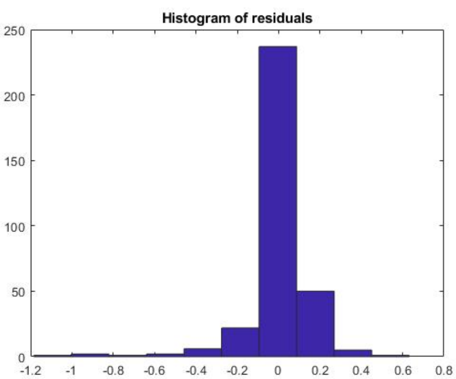
Πίνακας 4: MA coefficients.



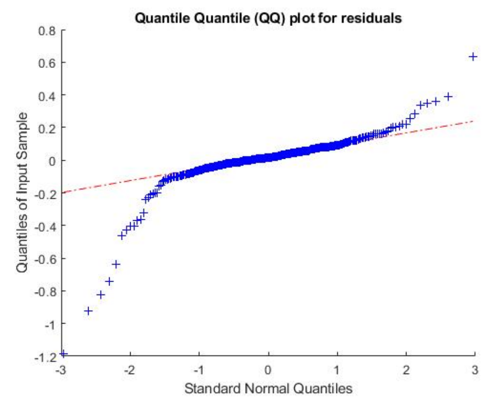
(α′) Forecasts on test set



(β′) ACF residuals on fitted values



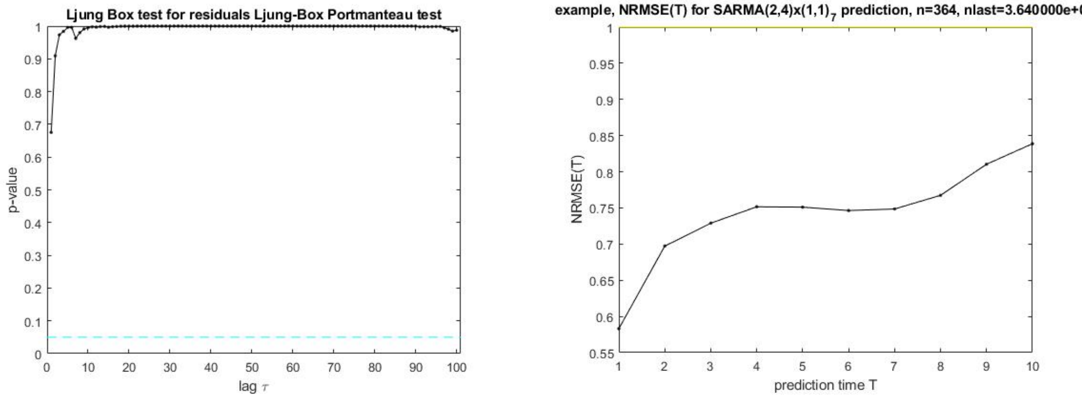
(γ′) Histogram residuals on fitted values



(δ′) QQ plot residuals on fitted values

Σχήμα 25: Diagnostics of residuals

Όπως ειναι φανερό οι αυτοσυσχετίσεις είναι όλες μη σημαντικές. Το τεστ Ljung-Box στα υπόλοιπα δίνει για όλες τις υστερήσεις μη σημαντικές αυτοσυσχετίσεις όπως φαίνεται στο Σχήμα 26. Ακόμη δίνονται τα σφάλματα πρόβλεψης ανάλογα με τον ορίζοντα πρόβλεψης στο ίδιο σχήμα. Τέλος το Lilliefors test το οποίο ελέγχει την μηδενική υπόθεση ότι τα δεδομένα προέρχονται απο κανονική κατανομή, δίνει για τα υπόλοιπα $p - value$, $p = 0.0001$ και συνεπώς απορρίπτεται η υπόθεση της κανονικής κατανομής των υπολοίπων στα fitted δεδομένα.



(α′) Ljung-Box test on residuals

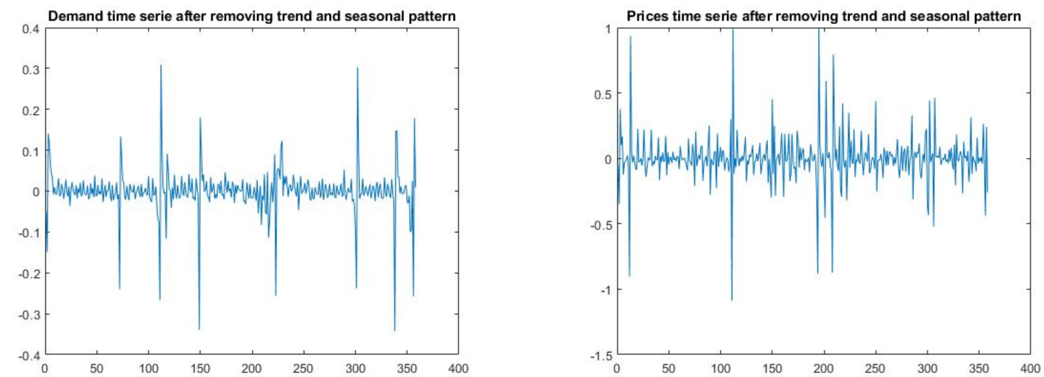
(β′) Σφάλμα πρόβλεψης ανάλογα του χρόνου τ.

Σχήμα 26: Ljung-Box test and prediction error.

Όπως βλέπουμε λοιπόν, τα αποτελέσματα που προέκυψαν για τις 2 χρονοσειρές φάνηκαν να έχουν αρκετά κοινά χαρακτηριστικά. Φαίνεται να υπάρχει κυριάρχο περιοδικό στοιχείο κάθε 7 υστερήσεις όπως φαίνεται απο τα διαγράμματα αυτοσυσχετίσεων για την μετασχηματισμένη με πρώτες διαφορές χρονοσειρά.
Επίσης παρουσιάζονται αρκετές ομοιότητες στο διάγραμμα αυτοσυσχετίσεων και μερικών αυτοσυσχετίσεων στην χρονοσειρά των πρώτων διαφορών και στις 2 περιπτώσεις. Τέλος ενδιαφέρον προκαλεί και η μορφή του μοντέλου που ειναι σχεδόν πανομοιότυπη εξαιρουμένων βέβαια των συντελεστών. Τέλος θα ήτανε σκόπιμο εφόσον θεωρούμε οτι η χρονοσειρά της τιμής και της ζήτησης θα συσχετίζονται με κάποιο τρόπο (περιμένουμε όταν ανεβαίνει η ζήτηση να ανεβαίνει και η τιμή) να υπολογίσουμε την γραμμική τους συσχέτιση. Πράγματι ο συντελεστής συσχέτισης Pearson δίνει συσχέτιση μεταξύ των δύο χρονοσειρών corr $= 0.5359$. Συνεπώς βάση της συσχέτισης που παρατηρούμε μπορούσαμε να υποθέσουμε οτι ένα καταλληλότερο μοντέλο θα μπορούσε να είναι SARMA με εξωγενής παράγοντες (SARMAX), όπου ο εξωγενής παράγοντας θα ήταν για την μέν χρονοσειρά ζήτησης, η χρονοσειρά τιμής και αντίστροφα.

1.3 Πρόβλεψη με AR(5) μοντέλο

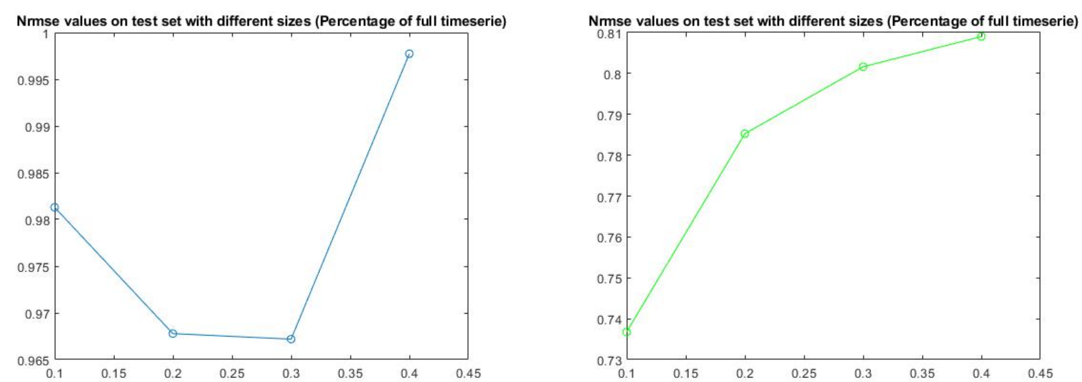
Παρακάτω θα δώσουμε τα αποτελέσματα για τις προβλέψεις με αυτοπαλίνδρομο μοντέλο AR(5). Προκειμένου να προσαρμόσουμε ενα AR(5) στα δεδομένα μας, θα πρέπει η χρονοσειρά μας να είναι στάσιμη, συνεπώς απαλλαγμένη απο την τάση αλλά και την εποχικότητα. Είδαμε προηγουμένως οτι για να απαλείψουμε εντελώς το εποχικό στοιχείο απο τα δεδομένα μας, ο καλύτερος τρόπος ειναι με προσαρμογή φίλτρου κινούμενου μέσου για την εποχικότητα. Οι χρονοσειρές της ζήτησης και της τιμής που προκύπτουν μετα την απαλοιφή και της περιοδικότητας δίνονται στο παρακάτω Σχήμα 27.



(α΄) Χρονοσειρά δεδομένων ζήτησης. (β΄) Χρονοσειρά δεδομένων τιμής.

Σχήμα 27: Χρονοσειρές απαλλαγμένες απο τάση και περιοδικότητα.

Οι τιμές του NRMSE για διαφορετικά σημεία στα οποία χωρίζεται το σύνολο αξιολόγησης δίνονται στο Σχήμα 28. Να σημειώσουμε εδώ οτι στον άξονα x παριστάται το ποσοστό των δεδομένων του συνόλου αξιολόγησης ως προς την αρχική χρονοσειρά.



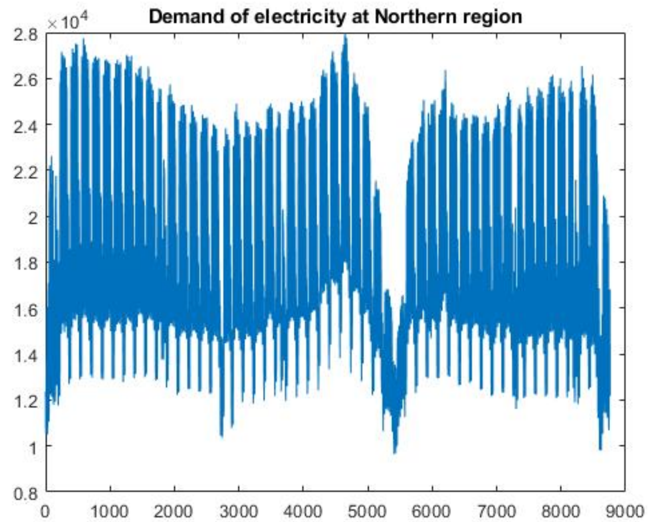
(α΄) Σφάλμα πρόβλεψης για την χρονοσειρά ζήτησης με διαφορετικά σημεία διαχωρισμού. (β΄) Σφάλμα πρόβλεψης για την χρονοσειρά των τιμών με διαφορετικά σημεία διαχωρισμού.

Σχήμα 28: Σφάλματα πρόβλεψης στις χρονοσειρές τιμής και ζήτησης.

Όπως είναι φανερό στα δεδομένα ζήτησης όπως είδαμε και στην ανάλυση που κάναμε προηγουμένως, αν αφαιρέσουμε την τάση και την εποχικότητα απο τα δεδομένα αυτο που μένει είναι λευκός θόρυβος και πράγματι το μοντέλο AR(5) έχει σφάλμα πρόβλεψης σε όλες τις περιπτώσεις κοντά στο 1 που πάει να πει ότι το μοντέλο προβλέπει ακριβώς όπως ένα naive μοντέλο με βάση την μέση τιμή.

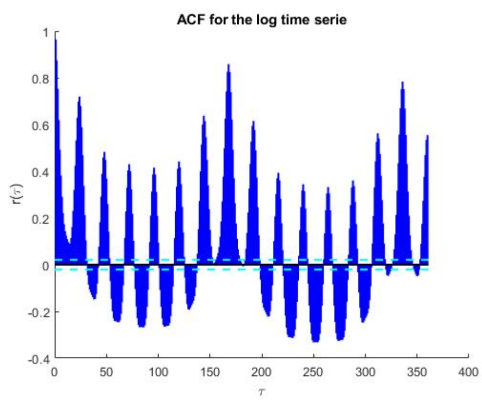
1.4 Πλήρης χρονοσειρά ζήτησης ηλεκτρικού ρεύματος

Σε αυτό το κομμάτι θα αναλύσουμε την πλήρη χρονοσειρά τόσο για την ζήτηση, όσο και για την τιμή του ηλεκτρικού ρεύματος. Θα αρχίσουμε όπως και προηγουμένως με την ζήτηση του ηλεκτρικού ρεύματος. Το διάγραμμα των ιστορικών τιμών δίνεται στο Σχήμα 29

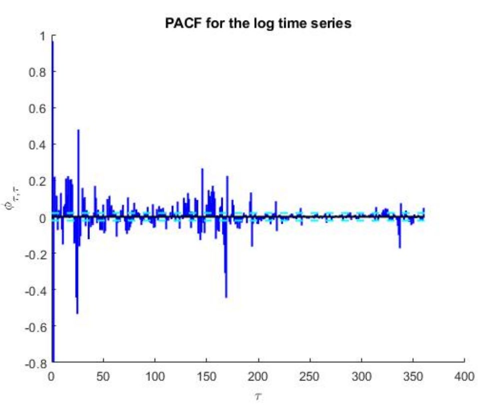


Σχήμα 29: Χρονοσειρά ζήτησης ηλεκτρικού ρεύματος.

Η χρονοσειρά των δεδομένων ζήτησης ορισμένη ώς στοχαστική διαδικασία έχει μέση τιμή $\mu = 18832$ και διακύμανση $\sigma = 4276.2132$. Δεδομένου οτι οι τιμές ειναι πολύ υψηλές και κατα κύριο λόγο η διακύμανση, εφαρμόσαμε μετασχηματισμό λογαρίθμου στην αρχική χρονοσειρά προκειμένου να φέρουμε τις τιμές σε ένα μικρότερο εύρος τιμών. Η νέα μέση τιμή που προκύπτει είναι $\mu = 9.8168$ και η νέα διακύμανση είναι $\sigma = 0.2321$.
Ο έλεγχος Dickey-Fuller για την υπόθεση της στασιμότητας δίνει $p - value$, $p = 0.5431$ οπότε συμπαιρένουμε οτι η χρονοσειρά δέν είναι στάσιμη, κάτι φυσικά που αναμέναμε να συμβεί. Όπως παρατηρούμε είναι εμφανής η τάση στα δεδομένα και όπως θα δούμε στο Σχήμα 30 στα διαγράμματα της αυτοσυσχέτισης και μερικής αυτοσυσχέτισης, υπάρχει έντονο το περιοδικό στοιχείο.



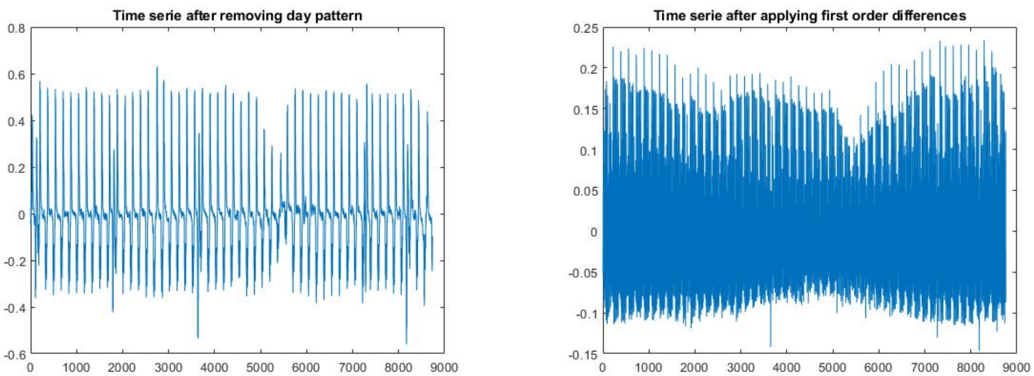
(α′) Διάγραμμα αυτοσυσχετίσεων για την αρχική χρονοσειρά



(β′) Διάγραμμα μερικών αυτοσυσχετίσεων για την αρχική χρονοσειρά

Σχήμα 30: Διαγράμματα συσχέτισης

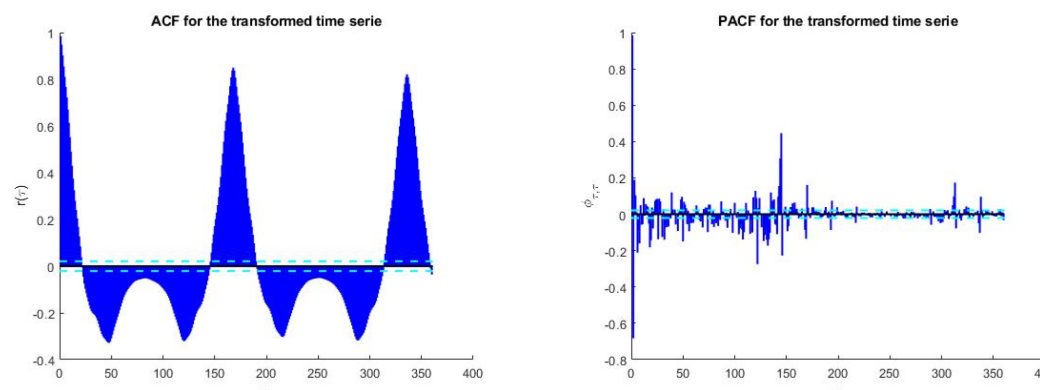
Όπως βλέπουμε στο Σχήμα 30, υπάρχει έντονο το περιοδικό στοιχείο το οποίο βέβαια παρατηρούμε οτι έχει σύνθετη συμπεριφορά, καθώς οι παρατηρήσεις μας είναι πλέον για ολόκληρο το 24ωρο οπότε ειναι λογικό να υποθέσουμε οτι τα δεδομένα μας θα έχουν 24ωρη περιοδικότητα αλλά ταυτοχρονα και εβδομαδιαια περιοδικότητα. Τέτοιου τύπου περιπτώσεις βέβαια είναι δύσκολο να αντιμετωπισθούν με κλασικά μοντέλα SARIMA και συνήθως επιστρατεύονται πιο σύγχρονα μοντέλα όπως τα μοντέλα TBATS [2] ή το μοντέλο Prophet [3] το οποίο αποτελεί την εξέλιξη των TBATS είναι σε θέση να συμπεριλάβει και ημέρες διακοπών στην μοντελοποίηση. Παρ'όλα αυτά θα δοκιμάσουμε να αντιμετωπίσουμε το πρόβλημα μας με κλασικά μοντέλα SARIMA.
Θα ξεκινήσουμε αφαιρώντας την ημερίσια περιοδική τάση απο την χρονοσειρά (καθώς είδαμε οτι η χρονοσειρά των ημερών παρουσιάζει τάση) παίρνοντας διαφορές τάξης 24 (σε αντιστοιχία με την αρχική μας ανάλυση όπου πήραμε διαφορές πρώτης τάξης). Η μετασχηματισμένη χρονοσειρά με διαφορές 24ης και 1ης τάξης δίνεται στο Σχήμα 31 όπου φαίνεται ξεκάθαρα οτι η μετασχηματισμένη χρονοσειρά των πρώτων διαφορών παρουσιάζει κάποια τάση η οποία πιθανολογούμε οτι ειναι περιοδική τάση (και για αυτον τον λόγο αφαιρούμε πρώτα την περιοδική τάση). Τα σχετικά διαγράμματα αυτοσυσχέτισης στο Σχήμα 32.



(α') Χρονοσειρά χωρίς το ημερίσιο περιοδικό στοιχείο.

(β') Μετασχηματισμένη χρονοσειρά με διαφορές πρώτης τάξης.

Σχήμα 31: Μετασχηματισμένες χρονοσειρές.

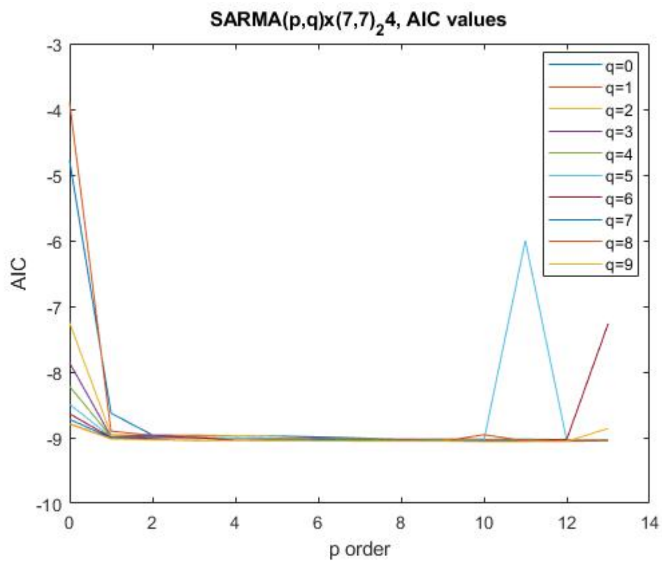


(α') Διάγραμμα αυτοσυσχετίσεων για την μετασχηματισμένη χρονοσειρά

(β') Διάγραμμα μερικών αυτοσυσχετίσεων για την μετασχηματισμένη χρονοσειρά

Σχήμα 32: Διαγράμματα συσχέτισης

Όπως βλέπουμε στο Σχήμα 31 φαίνεται να υπάρχει η συμπεριφορά που είχαμε στην αρχική χρονοσειρά με τα ημερίσια δεδομένα. Παρατηρούμε επίσης οτι στην τελική χρονοσειρά φαίνεται να έχει εξαλειφθεί οποιαδήποτε αργή μεταβολή υπήρχε στα δεδομένα και μένει να προσαρμόσουμε ενα εποχικό μοντέλο. Επιπλέον παρατηρούμε οτι επικρατεί έντονα το περιοδικό στοιχείο στην χρονοσειρά κάθε 24 ώρες αλλά και κάθε 7 ημέρες (αυτό φαίνεται και απο το διάγραμμα των αυτοσυσχετίσεων για υστέρηση 168 αλλά και εάν εστιάσουμε στην μετασχηματισμένη χρονοσειρά (α΄) στο Σχήμα 31). Προκειμένου να μπορέσουμε να συμπεριλάβουμε την περιοδικότητα της εβδομάδας που βλέπουμε οτι υπάρχει θα θεωρήσουμε $P, Q \in [0, 7]$. Παρ΄ όλα αυτά πιστέυουμε οτι η επιλογή $P = 7, Q = 7, s = 24$ είναι κατάλληλη καθώς στην προηγούμενη ανάλυση των ημερών το καλύτερο μοντέλο που προέκυψε ήταν για $P = 1, Q = 1, s = 7$. Πράγματι τα καλύτερα αποτελέσματα για το κριτήριο AIC προέκυψαν για $P = 7, Q = 7, s = 24$ και οι τιμές για το βέλτιστο AIC βάση των p, q δίνονται στο Σχήμα 33.

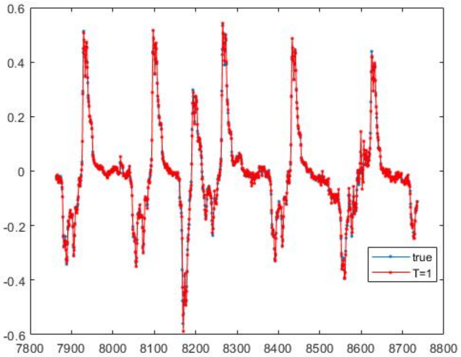


Σχήμα 33: Τιμές AIC για $P = 7, Q = 7$.

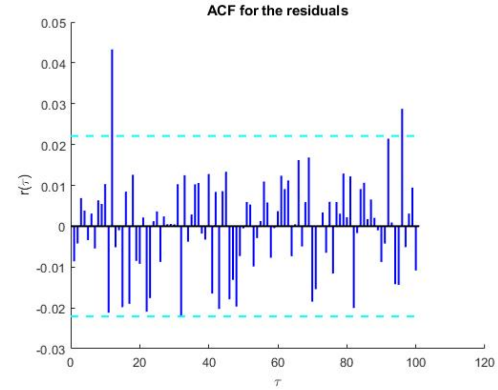
Βλέπουμε λοιπόν οτι το βέλτιστο μοντέλο κρίνοντας και απο τις παραμέτρους p, q επιτυχάνεται για $p = 10, q = 9$. Προσαρμόζουμε λοιπόν το μοντέλο $\text{SARMA}(10, 9) \times (7, 7)_{24}$ στήν μετασχηματισμένη χρονοσειρά (αντίστοιχα $\text{SARIMA}(10, 0, 9) \times (7, 1, 7)_{24}$ για την αρχική χρονοσειρά). Οι τιμές που προκύπτουν είναι οι εξής: $\text{AIC} = -9.0611, \text{FPE} = 0.000116, \sigma_\varepsilon = 0.010562$. Η τιμή του σφάλματος για το σύνολο αξιολόγησης το οποίο ορίστηκε να είναι το 10% της αρχικής χρονοσειράς προέκυψε να είναι $\text{NRMSE} = 0.0781$ ενώ για το σύνολο εκπαίδευσης $\text{NRMSE} = 0.0565$.

Στο Σχήμα 34 δίνονται οι προβλέψεις για το σύνολο αξιολόγησης, τα υπόλοιπα που προέκυψαν για τις προβλέψεις στο σύνολο εκπαίδευσης, καθώς και το αντίστοιχο τους ιστόγραμμα, διάγραμμα αυτοσυσχέτισης και Quantile-Quantile (QQ)-plot. Δέν θα παραθέσουμε τους συντελεστές του μοντέλου σε πίνακες, διότι το πλήθος των συντελεστών ειναι αρκετά μεγάλο. Ό έλεγχος για τις αυτοσυσχετίσεις στα υπόλοιπα μέσω του Ljung-Box test (ο οποίος δέχεται την μηδενική υπόθεση της μή ύπαρξης σημαντικών αυτοσυσχετίσεων στα υπόλοιπα) δίνεται στο Σχήμα 35 μαζί με τα σφάλματα πρόβλεψης ανάλογα του ορίζοντα πρόβλεψης.

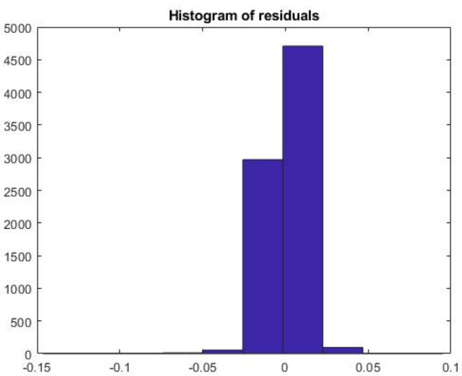
Τέλος το Lilliefors test το οποίο ελέγχει την μηδενική υπόθεση ότι τα δεδομένα προέρχονται απο κανονική κατανομή, δίνει για τα υπόλοιπα $p - value$, $p = 0.0001$ και συνεπώς απορρίπτεται η υπόθεση της κανονικής κατανομής των υπολοίπων στα fitted δεδομένα.



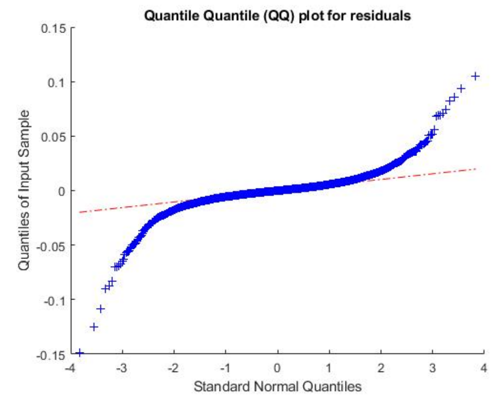
(α′) Forecasts on test set



(β′) ACF residuals on fitted values

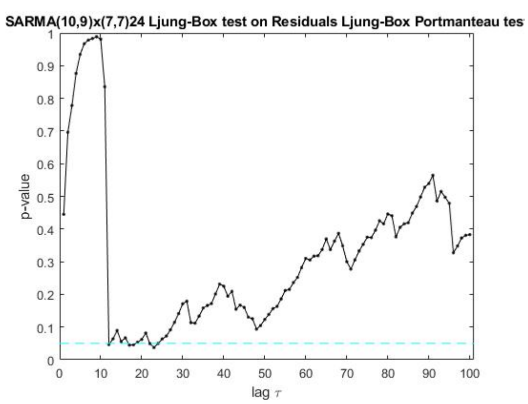


(γ′) Histogram residuals on fitted values

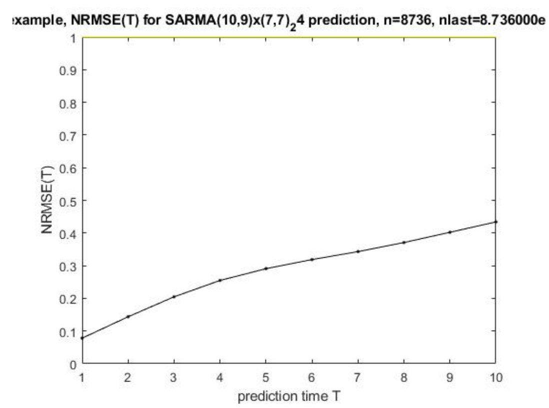


(δ′) QQ plot residuals on fitted values

Σχήμα 34: Diagnostics of residuals



(α′) Ljung-Box test on residuals.

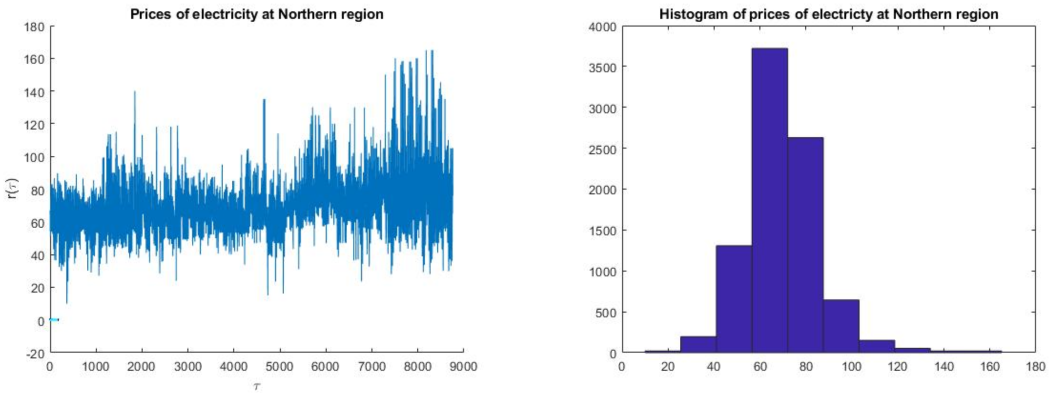


(β′) Σφάλμα πρόβλεψης ανάλογα του χρόνου τ.

Σχήμα 35: Ljung-Box test and prediction error

1.5 Πλήρης χρονοσειρά τιμής ηλεκτρικού ρεύματος

Αντίστοιχα για την χρονοσειρά τιμών θα ακολουθήσουμε την ίδια διαδικασία που ακολουθήσαμε προηγουμένως. Το διάγραμμα της χρονοσειράς και το αντίστοιχο ιστόγραμμα της παρουσιάζονται στο Σχήμα 36.

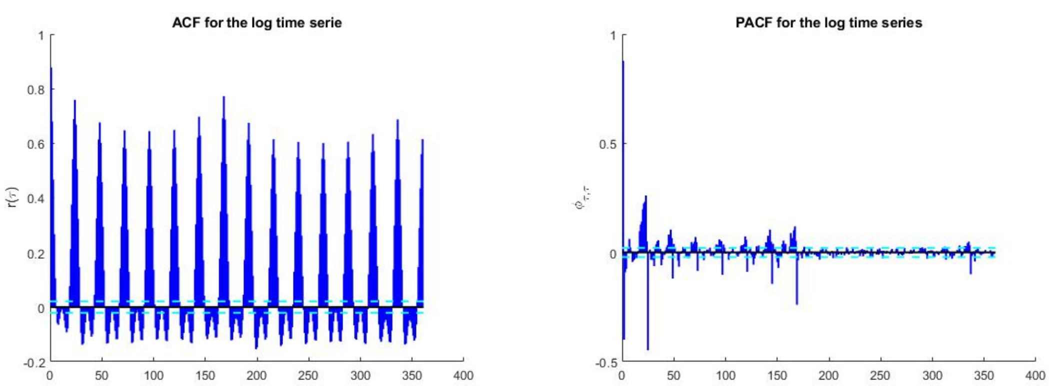


(α΄) Χρονοσειρά τιμών ηλεκτρικού ρεύματος.

(β΄) Ιστόγραμμα τιμών ηλεκτρικού ρεύματος.

Σχήμα 36: Περιγραφικά διαγράμματα ιστορικών δεδομένων τιμών.

Η χρονοσειρά των δεδομένων τιμής ηλεκτρικού ρεύματος ώς στοχαστική διαδικασία έχει μέση τιμή $\mu = 70.1760$ και διακύμανση $\sigma = 15.6654$. Ο έλεγχος Dickey-Fuller για την υπόθεση της στασιμότητας δίνει $p-value$, $p = 0.0001$ κάτι που πιθανών να ευθύνεται στις τιμές και το πλήθος των δεδομένων, αλλά παρ΄ όλα αυτά υποθέτουμε οτι θα υπάρχει περιοδική τάση (εφόσον υπήρχε τάση για την χρονοσειρά των ημερισίων τιμών του ρεύματος) και για αυτό θα προχωρήσουμε σε μετασχηματισμούς διαφορών k τάξης. Δεδομένου ότι παρατηρούνται αρκετές εξάρσεις στα δεδομένα, προκειμένου να σταθεροποιήσουμε τη διακύμανση εφαρμόσαμε μετασχηματισμό λογαρίθμου. Η νέα μέση τιμή που προκύπτει είναι $\mu = 4.2257$ και η νέα διακύμανση $\sigma = 0.2293$. Είναι σημαντικό να σημειώσουμε οτι επαναλβάνοντας το τέστ του ελέγχου της στασιμότητας προκύπτει νέο $p - value$, $p = 0.1981$ που απορρίπτει την υπόθεση της στασιμότητας. Τα διαγράμμα αυτοσυσχέτισης και μερικής αυτοσυσχέτισης για την λογαριθμημένη χρονοσειρά παρουσιάζονται στο Σχήμα 37.

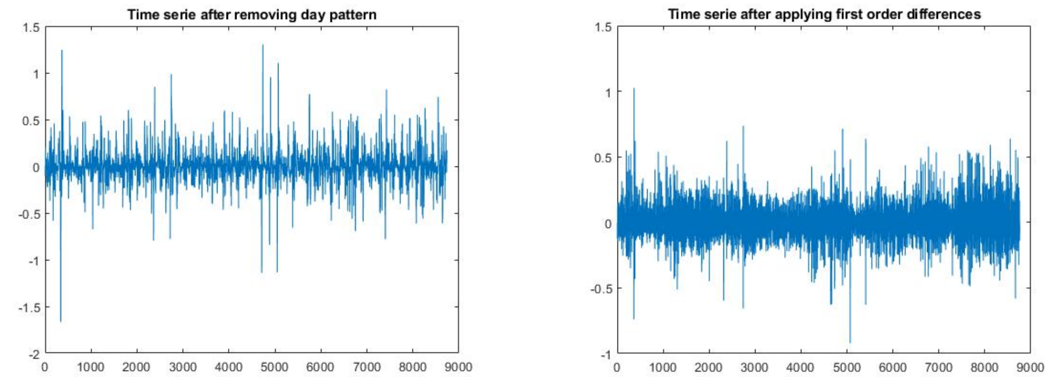


(α΄) Διάγραμμα συσχετίσεων.

(β΄) Διάγραμμα μερικών αυτοσυσχετίσεων.

Σχήμα 37: Διαγράμματα συσχέτισης.

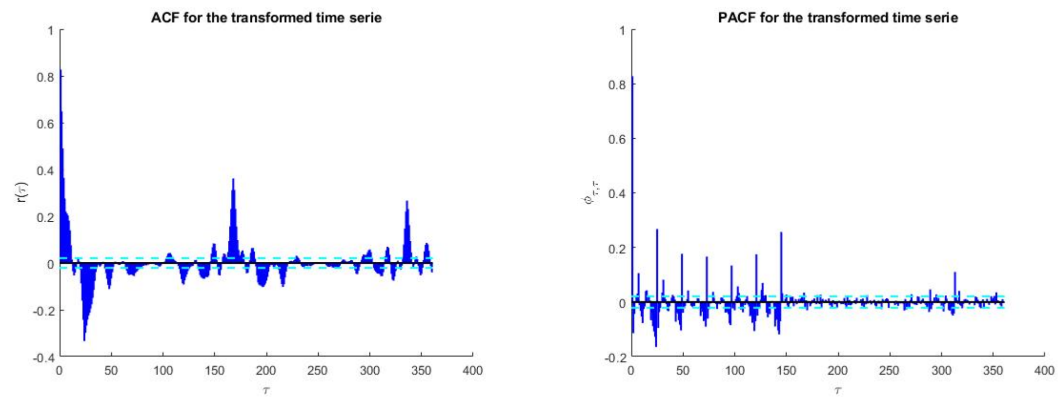
Όπως και προηγουμένως έτσι και τώρα ειναι εμφανής η ύπαρξη του περιοδικού στοιχείο της ημέρας αλλά και της εβδομάδας. Η μετασχηματισμένη χρονοσειρά με διαφορές 24ης τάξης (σε αντιστοιχία με τις διαφορές πρώτης τάξης για τις ημερίσιες τιμές) και με διαφορές πρώτης τάξης δίνεται στο Σχήμα 38.



(α΄) Διαφορές 24ης τάξης. (β΄) Διαφορές πρώτης τάξης.

Σχήμα 38: Μετασχηματισμένες χρονοσειρές.

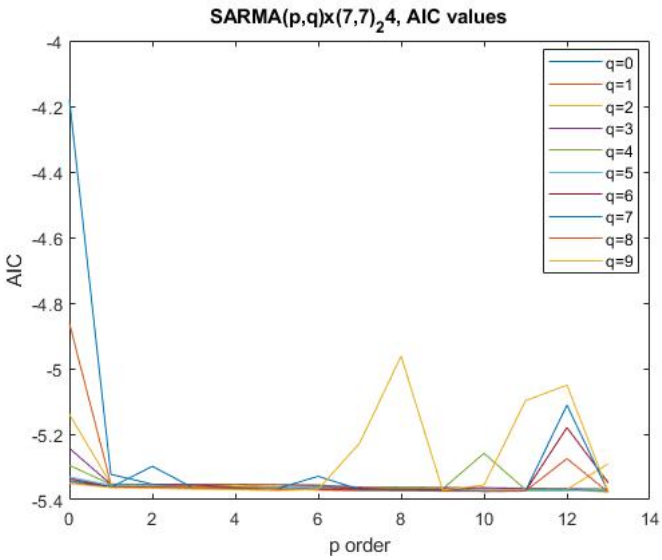
Όπως βλέπουμε στις διαφορές 24ης τάξης φαίνεται να έχουν απαλειφθεί αργές μεταβολές, κάτι που στις διαφορές πρώτης τάξης δεν φαίνεται να συμβαίνει. Τα διαγράμματα συσχέτισης και μερικής αυτοσυσχέτισης για την μετασχηματισμένη χρονοσειρά με διαφορές 24ης τάξης παρουσιάζονται στο Σχήμα 39.



(α΄) Διάγραμμα αυτοσυσχέτισης. (β΄) Διάγραμμα μερικής αυτοσυσχέτισης.

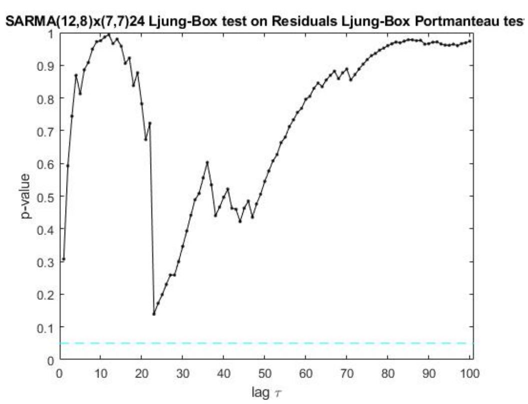
Σχήμα 39: Διαγράμματα συσχέτισης.

Υποθέτουμε όπως και πρίν ότι ένα κατάλληλο μοντέλο το οποίο θα μπορούσε να πιάσει την εβδομαδιαία περιοδικότητα θα ήταν για $P = 7, Q = 7, s = 24$ σε αντιστοιχία με τα εποχικά μέρη στις ημερίσιες χρονοσειρές ($P = 1, Q = 1, s = 7$). Παρ'όλα αυτά έγινε αναζήτηση σε όλες τις πιθανές τιμές $P, Q \in [0, 7]$. Η υπόθεση μας επιβεβαιώνεται και τώρα και το διάγραμμα με τις τιμές του AIC για τις διάφορες παραμέτρους p, q δίνεται στο Σχήμα 40.

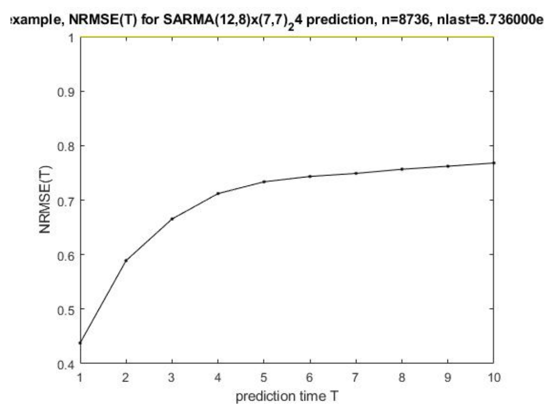


Σχήμα 40: Τιμές AIC για $P = 7, Q = 7$.

Επιλέγουμε λοιπόν $P = 7, Q = 7, s = 24, p = 12, q = 8$. Οι τιμές που προκύπτουν είναι οι εξής: AIC $= -5.2747$, FPE $= 0.005120$, $\sigma_\varepsilon = 0.070136$. Η τιμή του σφάλματος για το σύνολο αξιολόγησης το οποίο ορίστηκε να είναι το 10% της αρχικής χρονοσειράς προέκυψε να είναι NRMSE $= 0.4378$ ενώ για το σύνολο εκπαίδευσης NRMSE $= 0.4264$.
Στο Σχήμα 42 δίνονται οι προβλέψεις για το σύνολο αξιολόγησης, τα υπόλοιπα που προέκυψαν για τις προβλέψεις στο σύνολο εκπαίδευσης, καθώς και το αντίστοιχο τους ιστόγραμμα, διάγραμμα αυτοσυσχέτισης και Quantile-Quantile (QQ)-plot.
Δεν θα παραθέσουμε τους συντελεστές του μοντέλου σε πίνακες, διότι το πλήθος των συντελεστών ειναι αρκετά μεγάλο.
Ο έλεγχος για τις αυτοσυσχετίσεις στα υπόλοιπα μέσω του Ljung-Box test δίνεται στο Σχήμα 41 (ο οποίος δέχεται την μηδενική υπόθεση της μή ύπαρξης σημαντικών αυτοσυσχετίσεων στα υπόλοιπα) μαζί με το σφάλμα πρόβλεψης για διάφορες τιμές του ορίζοντα πρόβλεψης.

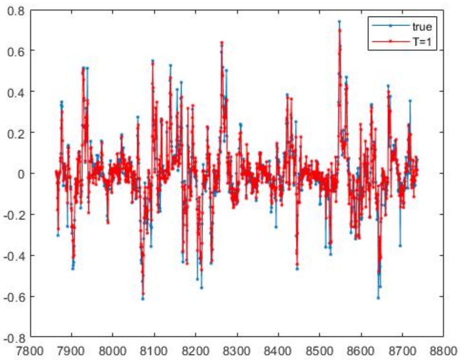


(α΄) Ljung-Box test on residuals.

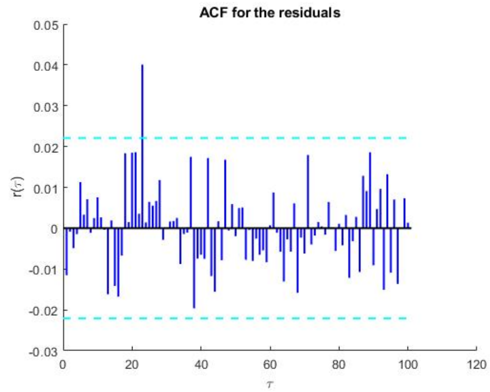


(β΄) Σφάλμα πρόβλεψης ανάλογα του χρόνου τ.

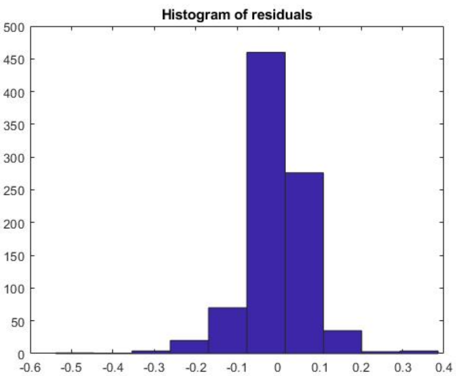
Σχήμα 41: Ljung-Box test and prediction error



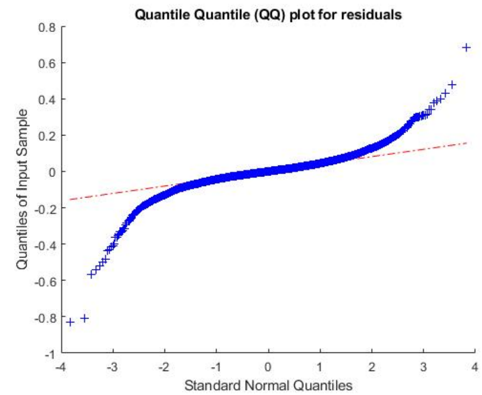
(α′) Forecasts on test set



(β′) ACF residuals on fitted values



(γ′) Histogram residuals on fitted values



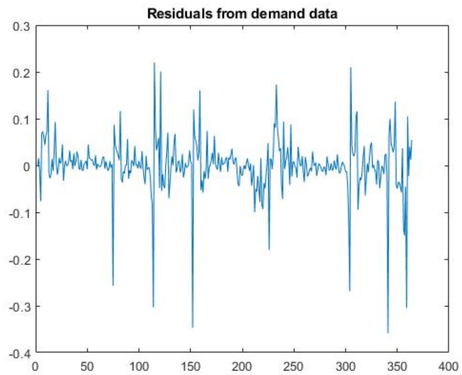
(δ′) QQ plot residuals on fitted values

Σχήμα 42: Diagnostics of residuals

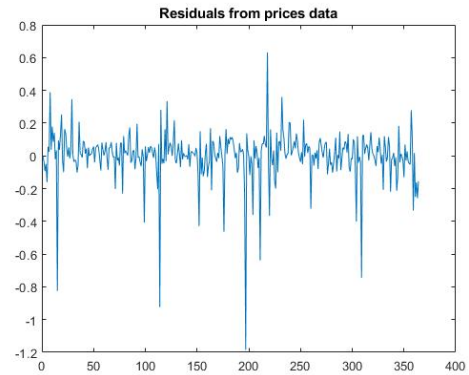
Τέλος το Lilliefors test το οποίο ελέγχει την μηδενική υπόθεση ότι τα δεδομένα προέρχονται απο κανονική κατανομή, δίνει για τα υπόλοιπα $p - value$, $p = 0.0001$ και συνεπώς απορρίπτεται η υπόθεση της κανονικής κατανομής των υπολοίπων στα fitted δεδομένα.

2 Μη γραμμική ανάλυση

Στο δεύτερο στάδιο ανάλυσης θέλουμε να διερευνήσουμε αν η κάθε μια από τις δύο χρονοσειρές για την περιοχή και ώρα αφού έχει απαλειφθεί η τάση και εποχικότητα έχει μη-γραμμικές αυτοσυσχετίσεις. Τα υπόλοιπα μετα την προσαρμογή των μοντέλων στις χρονοσειρές ζήτησης και τιμής παρατίθενται στο Σχήμα 43 και τα σχετικά τους διαγράμματα συσχέτισης στο Σχήμα 44.



(α΄) Χρονοσειρά υπολοίπων προσαρμογής στα δεδομένα ζήτησης.

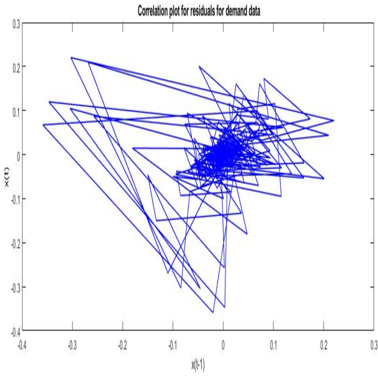


(β΄) Χρονοσειρά υπολοίπων προσαρμογής στα δεδομένα τιμής.

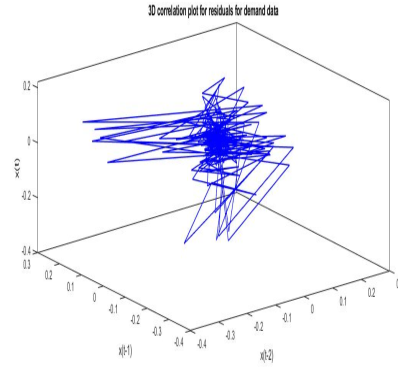
Σχήμα 43: Χρονοσειρές υπολοίπων

Εφόσον οι χρονοσειρές προέρχονται απο διακριτά συστήματα (στην προκειμένη περίπτωση ο χρόνος δειγματοληψίας είναι μεγάλος καθώς έχουμε μία παρατήρηση κάθε μέρα) θα θεωρήσουμε σαν υστέρηση στην ανάλυση που θα ακολουθήσουμε, $\tau = 1$ διότι θέτοντας μεγάλο τ σε συστήματα που προέρχονται απο μεγάλο χρόνο δειγματοληψίας κινδυνεύουμε να αποκόψουμε την δυναμική του συστήματος.
Θα αρχίσουμε την ανάλυση μας απο την χρονοσειρά ζήτησης του ηλεκτρικού ρεύματος. Τα πρώτα γραμμικά χαρακτηριστικά που θα υπολογίσουμε είναι η συνάρτηση αυτοσυσχέτισης και η συνάρτησης αμοιβαίας πληροφορίας για υστέρηση $\tau = 1$. Τα διαγράμματα που περιγράφουν την αυτοσυσχέτιση και αμοιβαία πληροφορία σε κάθε μία απο τις 21 χρονοσειρές (αρχική χρονοσειρά και 20 iid) παρουσιάζονται στο Σχήμα 45.
Απο τα μη γραμμικά χαρακτηριστικά επιλέξαμε να υπολογίσουμε τους ψευδούς κοντινότερους γείτονες για τον υπολογισμό της βέλτιστης διάστασης εμβύθισης και την διάσταση συσχέτισης για διάφορες τιμές του m και $\tau = 1$ που αναπαρίστανται στο Σχήμα 46.
Τέλος θα υπολογίσουμε τις τιμές του σφάλματος πρόβλεψης για ορίζοντα πρόβλεψης $h \in [1, 4]$ και διάσταση εμβύθισης $m \in [0, 30]$ και πλήθος γειτόνων $k \in [0, 30]$ στο Σχήμα 47.

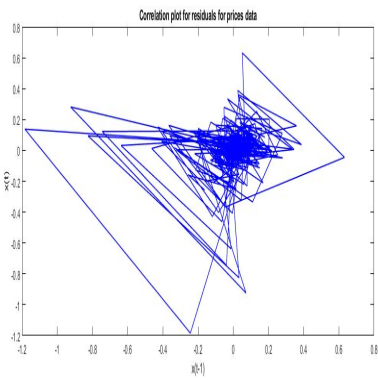
Όπως είναι φανερό και στις 4 περιπτώσεις το σφάλμα πρόβλεψης κυμαίνεται σε επίπεδα ανώτερα της μονάδας και επομένως συμπαιρένουμε οτι το μοντέλο μας δεν έχει καμία προβλεπτική ικανότητα αφου προβλέπει περίπου ίδια αλλά και χειρότερα απο την μέση τιμή.



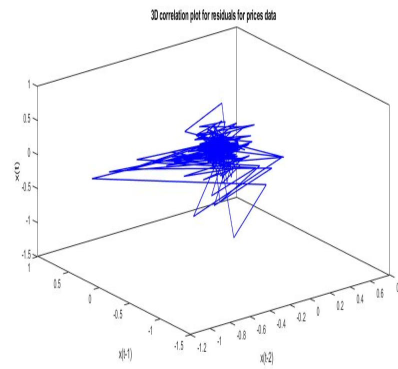
(α΄) Διάγραμμα συσχέτισης στα δεδομένα ζήτησης.



(β΄) Τρισδιάστατο διάγραμμα συσχέτισης στα δεδομένα ζήτησης.

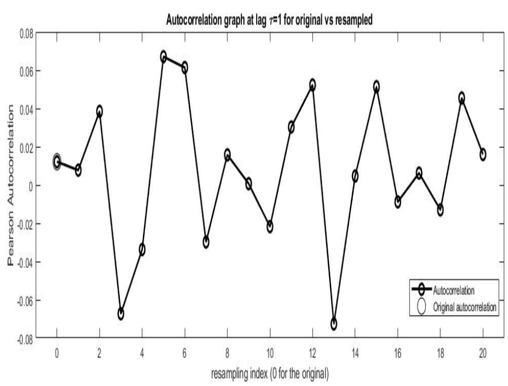


(γ΄) Διάγραμμα συσχέτισης στα δεδομένα τιμής.

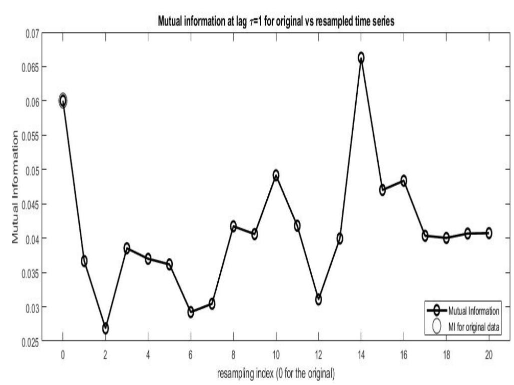


(δ΄) Τρισδιάστατο διάγραμμα συσχέτισης στα δεδομένα τιμής.

Σχήμα 44: Διαγράμματα συσχέτισης

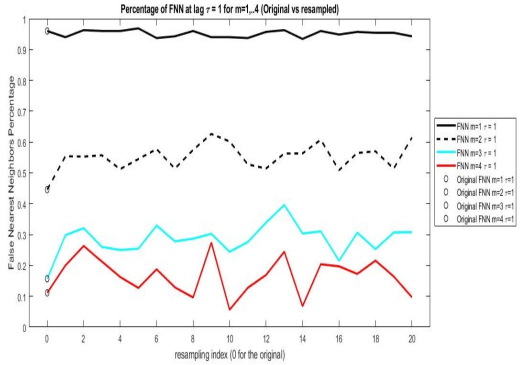


(α΄) Αυτοσυσχέτιση μεταξύ των 21 χρονοσειρών.

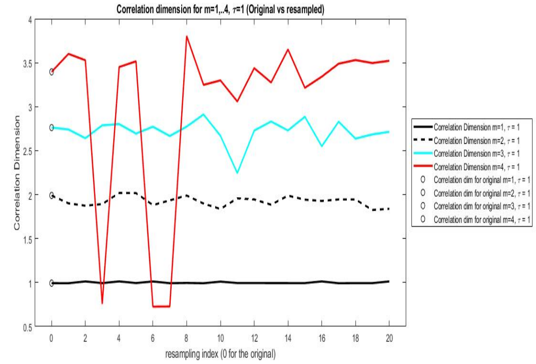


(β΄) Αμοιβαία πληροφορία για υστέρηση $\tau = 1$.

Σχήμα 45: Διαγράμματα συσχετίσεων και αμοιβαίας πληροφορίας στις 21 χρονοσειρές

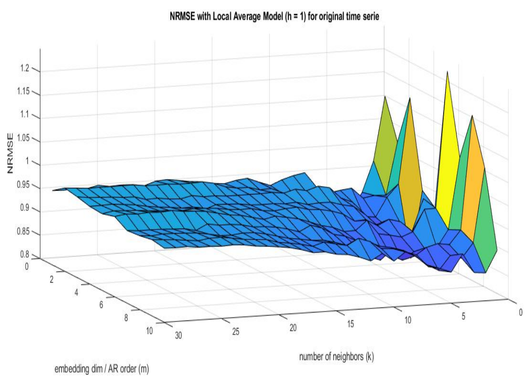


(α′) Ποσοστό ψευδών γειτόνων για τις 21 χρονοσειρές για $\tau = 1, m = 1, .., 4$.

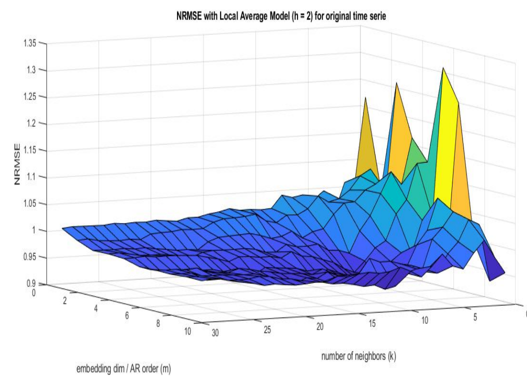


(β′) Διάσταση συσχέτισης για τις 21 χρονοσειρές για $\tau = 1, m = 1, .., 4$.

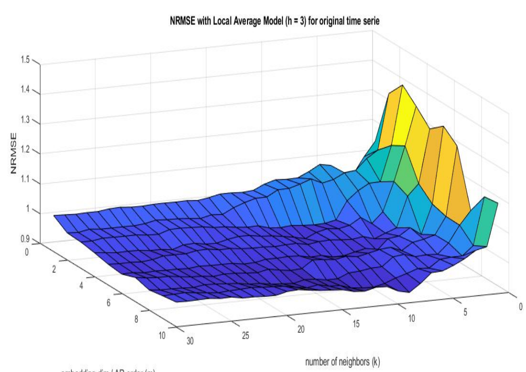
Σχήμα 46: Ποσοστό ψευδών γειτόνων και διάσταση συσχέτισης



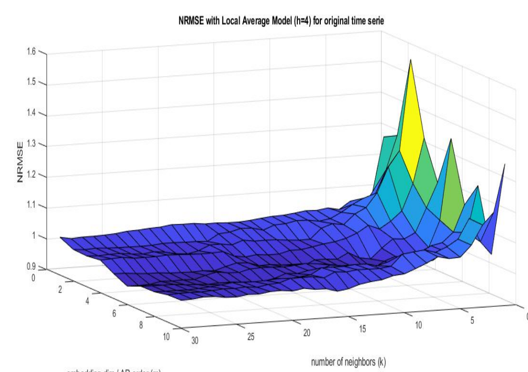
(α′) NRMSE vs number of neighbors vs embedding dimension (h=1).



(β′) NRMSE vs number of neighbors vs embedding dimension (h=2).



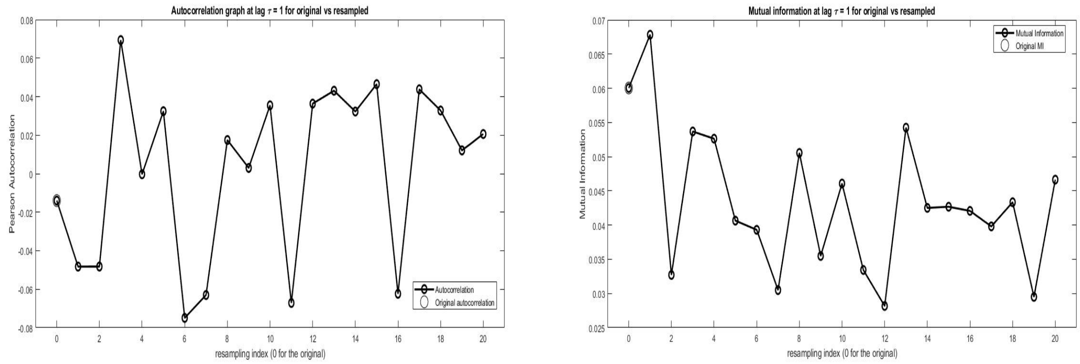
(γ′) NRMSE vs number of neighbors vs embedding dimension (h=3).



(δ′) NRMSE vs number of neighbors vs embedding dimension (h=4).

Σχήμα 47: Τιμές NRMSE για ορίζοντα πρόβλεψης $h = 1, .., 4$.

Θα ακολουθήσουμε την ίδια διαδικασία για την χρονοσειρά των υπολοίπων των τιμών του ηλεκτρικού ρεύματος. Τα διαγράμματα αυτοσυσχέτισης και αμοιβαίας πληροφορίας σε κάθε μια απο τις 21 χρονοσειρές (αρχική χρονοσειρά και 20 iid)παρουσιάζονται στο Σχήμα 48.

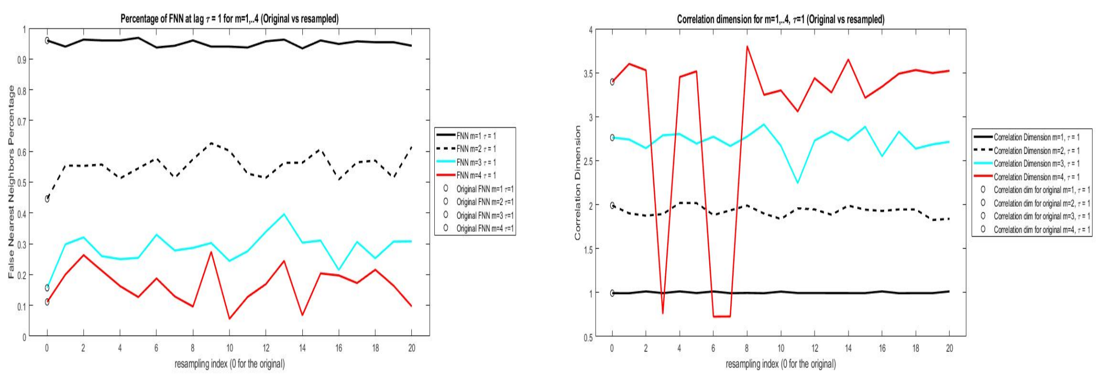


(α') Αυτοσυσχέτιση μεταξύ των 21 χρονοσειρών.

(β') Αμοιβαία πληροφορία για υστέρηση $\tau = 1$.

Σχήμα 48: Διαγράμματα αυτοσυσχετίσεων και αμοιβαίας πληροφορίας στις 21 χρονοσειρές

Απο τα μη γραμμικά χαρακτηριστικά επιλέξαμε όπως και πρίν να υπολογίσουμε τους ψευδούς κοντινότερους γείτονες για τον υπολογισμό της βέλτιστης διάστασης εμβύθισης και την διάσταση συσχέτισης για διάφορες τιμές του m και $\tau = 1$ που αναπαρίστανται στο Σχήμα 49.

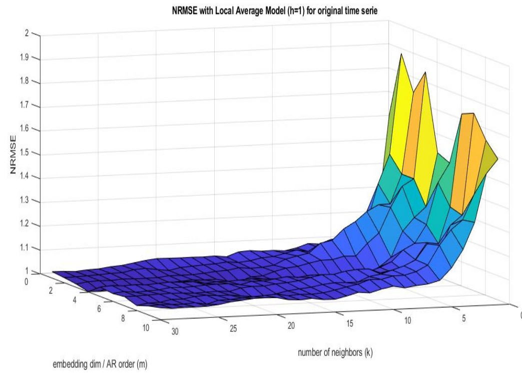


(α') Ποσοστό ψευδών γειτόνων για τις 21 χρονοσειρές για $\tau = 1, m = 1, .., 4$.

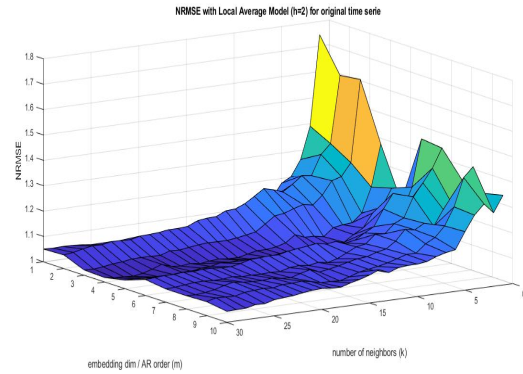
(β') Διάσταση συσχέτισης για τις 21 χρονοσειρές για $\tau = 1, m = 1, .., 4$.

Σχήμα 49: Ποσοστό ψευδών γειτόνων και διάσταση συσχέτισης

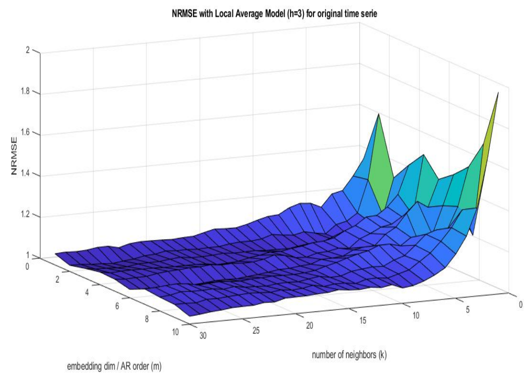
Τέλος θα υπολογίσουμε τις τιμές του σφάλματος πρόβλεψης για ορίζοντα πρόβλεψης $h \in [1, 4]$ και διάσταση εμβύθισης $m \in [0, 30]$ και πλήθος γειτόνων $k \in [0, 30]$ στο Σχήμα 50.



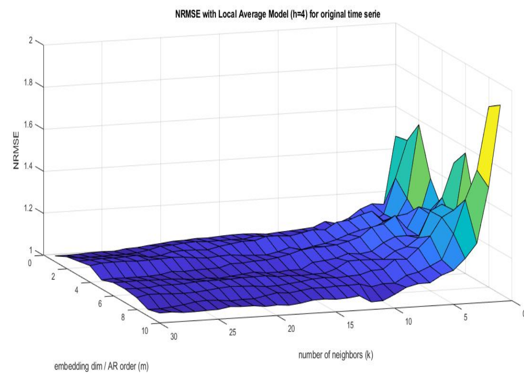
(α′) NRMSE vs number of neighbors vs embedding dimension (h=1).



(β′) NRMSE vs number of neighbors vs embedding dimension (h=2).



(γ′) NRMSE vs number of neighbors vs embedding dimension (h=3).



(δ′) NRMSE vs number of neighbors vs embedding dimension (h=4).

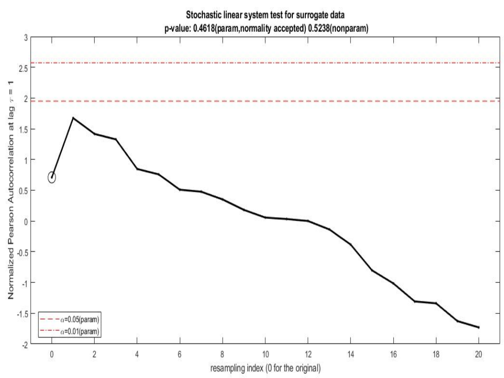
Σχήμα 50: Τιμές NRMSE για ορίζοντα πρόβλεψης $h = 1, .., 4$.

Όπως και προηγουμένως έτσι και τώρα και στις 4 περιπτώσεις το σφάλμα πρόβλεψης κυμαίνεται σε επίπεδα ανώτερα της μονάδας και επομένως συμπαιρένουμε οτι το μοντέλο μας δεν έχει καμία προβλεπτική ικανότητα αφου προβλέπει περίπου ίδια αλλά και χειρότερα απο την μέση τιμή.

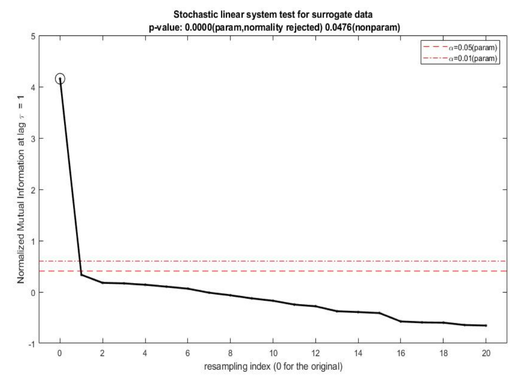
2.1 Συμπεράσματα μη γραμμικής ανάλυσης

Απο την μη γραμμική ανάλυση που διεξάγαμε και στις 2 χρονοσειρές (ζήτησης και τιμής) βρήκαμε τα παρακάτω συμπεράσματα.
Αρχίζοντας απο τα Σχήματα 45 και αντίστοιχα 48 βλέπουμε οτι τόσο η αμοιβαία πληροφορία αλλα και η αντίστοιχη συνάρτηση αυτοσυσχέτισης κυμαίνονται σε πολύ χαμηλά επίπεδα για την 1η υστέρηση (αυτό μας προδιαθέτει να υποθέσουμε οτι οι χρονοσειρές που δημιουργήσαμε είναι όλες θόρυβος).
Συνεχίζοντας με τα μη γραμμικά χαρακτηριστικά στα Σχήματα 46 και 49 αντίστοιχα βλέπουμε να παρουσιάζονται πάλι χαρακτηριστικά του θορύβου στα δεδομένα καθώς βλέπουμε οτι το ποσοστό των ψευδών γειτόνων οι οποίοι είναι πολύ ευαίσθητοι στην ύπαρξη θορύβου, δεν φτάνει ποτέ σε χαμηλά επίπεδα (κάτω του 1%). Επιπλέον όμως παρατηρούμε οτι και η αντίστοιχη διάσταση συσχέτισης και στις 2 περιπτώσεις παίρνει τιμή κοντά στο m και οταν το m αυξάνει, κάτι που είναι χαρακτηριστικό επίσης του θορύβου.
Τέλος βλέπουμε στα Σχήματα 47 και 50 οτι ενα μή γραμμικό μοντέλο δεν είναι σε καμία περίπτωση σε θέση να προβλέψει μελλοντικές στιγμές καθώς τα επίπεδα σφάλματος ειναι πολύ υψηλά.
Απο τα παραπάνω συμπαιρένουμε οτι η μορφή του συστήματος της αρχικής χρονοσειράς και

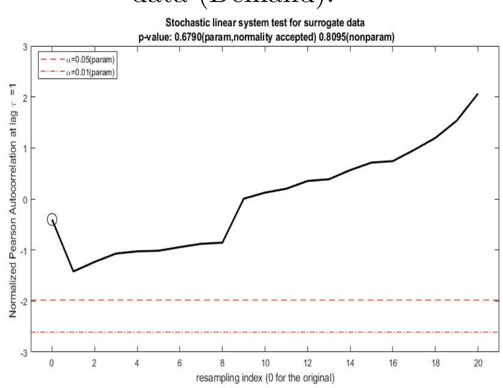
στις 2 περιπτώσεις. Θα μπορούσαμε βέβαια να υποθέσουμε οτι το σύστημα θα ήταν εφικτό να είναι ενα υψηλοδιάστατο χαοτικό σύστημα αλλά και αυτή η υπόθεση είναι δύσκολη να αποδειχτεί καθώς όπως βλέπουμε στα διαγράμματα των τιμών του NRMSE όσο αυξάνουμε το πλήθος των γειτόνων δεν παρατηρούμε κάποια αλλαγή στο σφάλμα πρόβλεψης και συνεπώς ειναι δύσκολο να υποθέσουμε υψηλοδιάστατο χαοτικό σύστημα.
Άρα συμπεραίνουμε λοιπόν οτι το σύστημα μας και στις 2 περιπτώσεις ειναι ενα στοχαστικό σύστημα με χαμηλή πολυπλοκότητα που προκύπτει απο το πλήθος των παραμέτρων των γραμμικών μοντέλων μας, και σχετικά μικρής μνήμης καθώς χρειαζόμαστε τις 7 προηγούμενες παρατηρήσεις για να προβλέψουμε την επόμενη.
Ολοκληρώνοντας θα εξετάσουμε μέσω της χρήσης των *Surrogate data* που δημιουργήσαμε την εγκυρότητα των αποτελεσμάτων μας. Τα δύο μέτρα που θα χρησιμοποιήσουμε για να διεξάγουμε τον έλεγχο μας είναι η αυτοσυσχέτιση των χρονοσειρών στην υστέρηση $\tau = 1$, και η αμοιβαία πληροφορία για την ίδια υστέρηση. Η μηδενική υπόθεση για τον έλεγχο που θα διενεργηθεί με είναι οτι τα δεδομένα μας προέρχονται απο γραμμική στοχαστική διαδικασία και συνεπώς δεν υπάρχει ενα δυναμικό μη γραμμικό σύστημα που να τα περιγράφει.



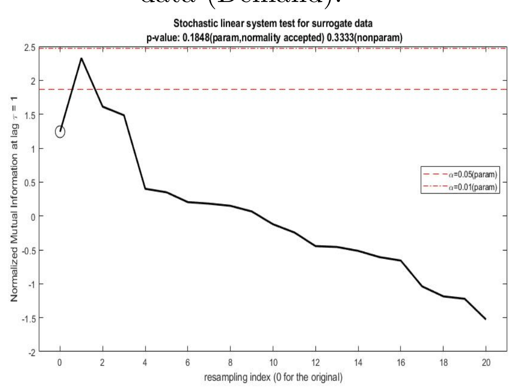
(α′) Stochastic linear test for Autocorrelation at lag $\tau = 1$ for surrogate data (Demand).



(β′) Stochastic linear test for Mutual Information at lag $\tau = 1$ for surrogate data (Demand).



(γ′) Stochastic linear test for autocorrelation at lag $\tau = 1$ for surrogate data (Price).



(δ′) Stochastic linear test for mutual information at lag $\tau = 1$ for surrogate data (Price).

Σχήμα 51: Surrogate data test.

Στο Σχήμα 51 βλέπουμε τους ελέγχους που διενεργήθηκαν σύμφωνα με τα στατιστικά που χρησιμοποιήσαμε για την υπόθεση του γραμμικου στοχαστικου συστήματος. Βλέπουμε οτι στην περίπτωση της αυτοσυσχέτισης για τα δεδομένα ζήτησης αποδεχομάστε την μηδενική υπόθεση και συνεπώς αυτο μας υποδυκνύει οτι απο πίσω κρύβεται ενα στοχαστικό σύστημα.

Για την περίπτωση της αμοιβαίας πληροφορίας βλέπουμε όμως οτι στην περίπτωση της ζήτησης απορρίπεται η μηδενική υπόθεση αλλα αυτό που μας προξενεί αμφιβολία είναι η απόρριψη του Kolmogorov-Smirnov test οτι οι τιμές των στατιστικών στα resampled data δεν ακολουθούν κανονική κατανομή και συνεπώς δεν μπορούμε να βασιστούμε στον παραμετρικό έλεγχο.
Στόν δε μή παραμετρικό έλεγχο μπορούμε να αποδεχτούμε την μηδενική υπόθεση καθώς το αντίστοιχο $p-value$ βρίσκεται στο όριο $p=0.0476$. Στην περίπτωση των δεδομένων ζήτησης βλέπουμε οτι αποδεχόμαστε και στις 2 περιπτώσεις την μηδενική υπόθεση καθώς και τα δύο τέστ που διενεργήσαμε δείχνουν στην ίδια κατεύθυνση.

Αναφορές

[1] Kugiumtzis, D., “Surrogate data test on time series,” In Modelling and Forecasting Financial Data, pp. 267-282. Springer, Boston, MA, 2002.

[2] Livera, A. and Hyndman, R. and Snyder, R., “Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing,” Journal of the American Statistical Association, vol. 106, no. 1, pp. 1513-1527, 2010.

[3] Taylor, S. J. and Letham, B., “Forecasting at scale,” The American Statistician, vol. 72, no. 1, pp. 37-45, 2018.