Χρονοσειρές

Υπολογιστική εργασία

Ανάλυση χρονοσειράς και εύρεση σημείων σημαντικών αλλαγών

Γεώργιος Τσουμπλέκας, gktsoump@ece.auth.gr, ΑΕΜ: 9359

Νικόλαος Παπαγεωργίου, nikolaospg@ece.auth.gr, ΑΕΜ: 9425

---

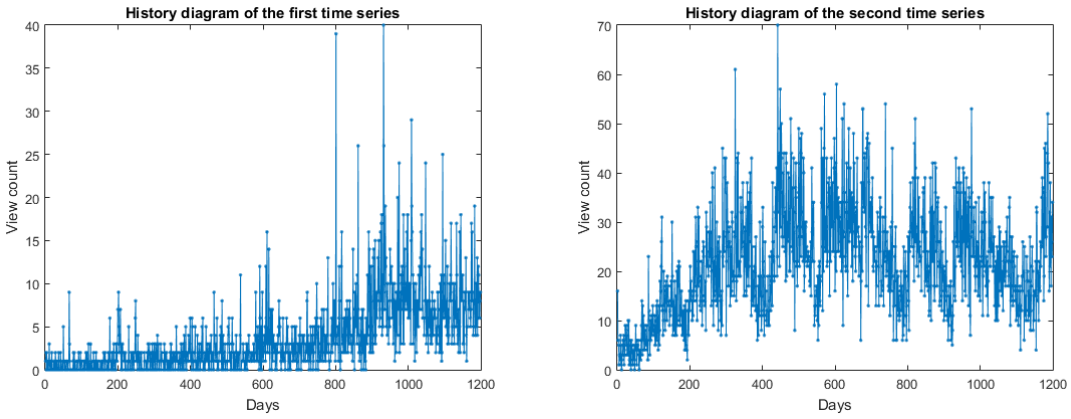
0 Εισαγωγή

Στην παρούσα εργασία καλούμαστε να αναλύσουμε 2 δοσμένες χρονοσειρές όπου καθεμιά αναπαριστά τον ημερήσιο αριθμό προβολών ενός βίντεο του Youtube. Πιο συγκεκριμένα, μας ενδιαφέρει να εντοπίσουμε σημαντικές αλλαγές στην ημερήσια ζήτηση προβολής ενός βίντεο, κάτι ιδιαίτερα χρήσιμο και πρακτικό μιας και το video content αποτελεί μεγάλο μέρος της κίνησης του διαδικτύου σήμερα. Έτσι, λοιπόν, η προσοχή μας θα επικεντρωθεί στο κομμάτι αυτό της εύρεσης τέτοιων σημαντικών αλλαγών στην χρονοσειρά. Αρχικά, θα ξεκινήσουμε αναλύοντας τις χρονοσειρές χρησιμοποιώντας γραμμικά μοντέλα (ARMA) και στην συνέχεια θα πραγματοποιηθεί και ανάλυση των χρονοσειρών με χρηση μη-γραμμικών δυναμικών χαοτικών συστημάτων. Τέλος, θα συγκρίνουμε τις δυο αυτές προσεγγίσεις και θα προέβουμε στην εξαγωγή κατάλληλων συμπερασμάτων.

1 Οπτικοποίηση χρονοσειρών

Σε πρώτη φάση, θα δημιουργήσουμε τα διαγράμματα των χρονοσειρών προκειμένου να δούμε αν μπορούμε να εξάγουμε κάποια γενικά συμπεράσματα για αυτές και να ξέρουμε προς ποιά κατεύθυνση πρέπει να κινηθούμε (πχ αν υπάρχει εμφανής τάση η περιοδικότητα).

Παρατηρούμε από τα διαγράμματα του Σχ.1 ότι και οι δύο χρονοσειρές εμφανίζουν τάση η οποία φαίνεται να μεταβάλλεται με τον χρόνο. Αυτό μας προϊδεάζει ότι κατά πάσα πιθανότητα οι δύο αυτές χρονοσειρές είναι μη-στάσιμες και θα χρειαστεί να τις μετατρέψουμε σε στάσιμες πρωτού προέβουμε στην προσαρμογή κάποιου μοντέλου σε αυτές. Επιπλέον, βλέπουμε ότι το εύρος τιμών που παίρνουν οι δύο χρονοσειρές διαφέρει, με αυτό της χρονοσειράς Β να είναι σχεδόν διπλάσιο της Α. Τέλος, παρατηρώντας πιο προσεκτικά την χρονοσειρά Α βλέπουμε ότι με την πάροδο του χρόνου η διακύμανση των τιμών γύρω από την μέση τιμή φαίνεται να αυξάνεται κάτι το οποίο μας υποδεικνύει την ύπαρξη μη στασιμότητας. Κάτι αντίστοιχο ισχύει σε μικρότερη κλίμακα και για την χρονοσειρά Β.



(a) Χρονοσειρά A (b) Χρονοσειρά B

Figure 1: Οι υπό μελέτη χρονοσειρές

2 Μετατροπή σε στάσιμες χρονοσειρές

2.1 Χρονοσειρά A

Αρχικά, θα σχηματίσουμε το διάγραμμα αυτοσυσχετίσεων της χρονοσειράς για να δούμε κατά πόσο υπάρχουν συσχετίσεις μεταξύ χρονικών στιγμών με υστέρηση μέχρι και T=100. Η τιμή αυτή του T είναι αρκετά μεγάλη και περιμένουμε σε μια στάσιμή χρονοσειρά η αυτοσυσχέτιση να γίνει στατιστικά ίση με το 0 πολύ πιο σύντομα. Παρόλα αυτά επιλέγουμε αυτή ώστε σε περίπτωση μη στάσιμης χρονοσειράς να μπορούμε να είμαστε πιο σίγουροι για την ύπαρξη ισχυρών συσχετίσεων που οφείλονται στην ύπαρξη τάσης. Κάτι τέτοιο, λοιπόν βλέπουμε να ισχύει και σε αυτή την περίπτωση (βλ. Σχ.2). Όπως παρατηρούμε στο Σχ.2, η αυτοσυσχέτιση εμφανίζει υψηλές τιμές και φθίνει με πολύ αργό ρυθμό (χαρακτηριστικό είναι το γεγονός ότι ακόμα και για υστέρηση 100 έχουμε μεγάλη τιμή αυτοσυσχέτισης). Τα παραπάνω είναι χαρακτηριστικά χρονοσειράς που εμφανίζει τάση για αυτό και το πρώτο βήμα μας θα είναι να την απαλείψουμε.

Αποφασίσαμε να μην χρησιμοποιήσουμε κάποια παραμετρική συνάρτηση f(t) του χρόνου μιας και δεν έχουμε κάποιο στοιχείο για το αν αυτή μπορεί να περιγραφεί καθοριστικά. Έτσι, η επιλογή ενός φίλτρου κινούμενου μέσου έμοιαζε πιο ταιριαστή. Γνωρίζουμε ότι η επιλογή μεγάλης τάξης απαλείφει κυρίως τις πιο αργές μεταβολές ενώ μικρή τάξη είναι ικανότερη στην απαλοιφή μεταβολών μικρότερης χρονικής κλίμακας. Κρίναμε, λοιπόν, πως θα ήταν πιο χρήσιμο να απαλείψουμε τις τάσεις σε πιο μικρή χρονική κλίμακα για να μην συνοπολογίζονται τυχόν spikes της τάσης μιας και μας ενδαφέρει η πραγματική μεταβολή των προβολών στα βίντεο σε ημερήσια βάση. Κατόπιν δοκίμής διαφόρων τιμών, τελικά η επιλογή φιλτρου κινούμενου μέσου τάξης 7 φαίνεται να είναι η πιο αποτελεσματική για αυτόν τον σκοπό οπότε και χρησιμοποιήθηκε. Στο Σχ.3 φαίνεται η χρονοσειρά και το διάγραμμα αυτοσυσχετίσεων αυτής αφού εφαρμόσουμε το φίλτρο κινούμενου μέσου τάξης 7.

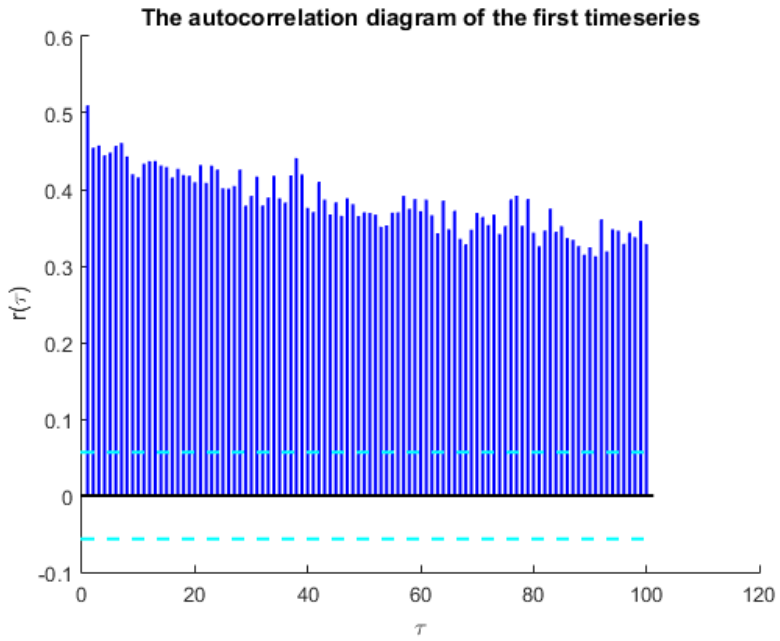
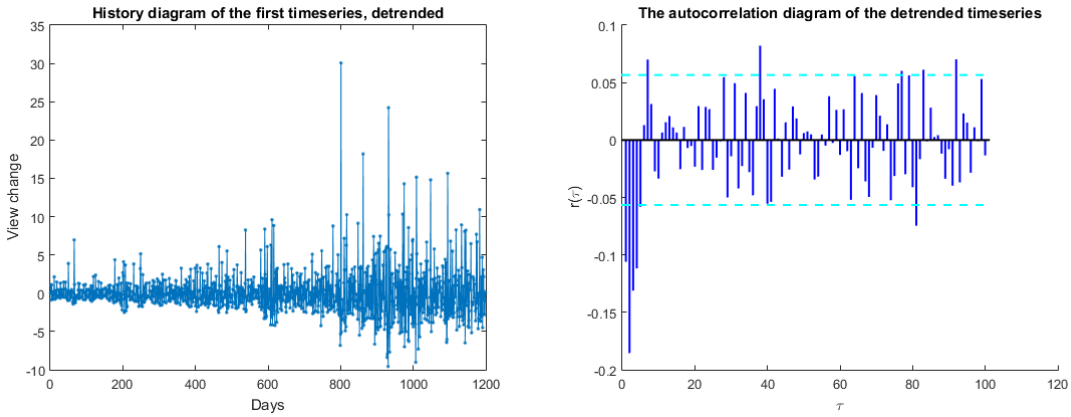


Figure 2: Διάγραμμα αυτοσυσχετίσεων της χρονοσειράς A

Όπως βλέπουμε στο Σχ.3α φαίνεται πως η τάση έχει απαλειφθεί με αποτελεσματικό τρόπο μιας και οι τιμές της νέας χρονοσειράς είναι κεντραρισμένες γύρω από το 0. Επιπλεόν, στο Σχ.3β φαίνεται ξεκάθαρα ότι έχουμε στατιστικά σημαντική αυτοσυσχέτιση για τις πρώτες 5 υστερήσεις άρα μπορούμε να αντλήσουμε κάποια πληροφορία από αυτές (η χρονοσειρά δεν είναι λευκός θόρυβος) ενώ από εκεί και μετά (με εξαίρεση κάποιες μεμονομένες τιμές που και πάλι όμως δεν είναι πολύ μεγάλες) είναι γενικά στατιστικά ίσες με το 0. Επιπλέον, δεν φαίνεται να υπάρχει κάποια περιοδικότητα στις τιμές των αυτοσυσχετίσεων οπότε από τα παραπάνω μπορούμε να συμπεράνουμε ότι η χρονοσειρά είναι πλέον όντως στάσιμη. Ο μόνος προβληματισμός μας είναι ότι η διασπορά των τιμών της χρονοσειράς ίσως μεταβάλλεται με τον χρόνο οπότε ίσως να χρειαζόταν να πάρουμε την διαφορά των λογαρίθμων της χρονοσειράς. Παρόλα αυτά θεωρούμε ότι ακόμα και αν ισχύει κάτι τέτοιο δεν φαίνεται να συμβαίνει σε μεγάλο βαθμό ώστε να καθιστά αδύνατη την προσαρμογή ενός ARMA μοντέλου στην χρονοσειρά. Έτσι, λοπόν, είμαστε πλέον σε θέση να εφαρμόσουμε την μέθοδο Box-Jenkins για την προσαρμογή γραμμικού μοντέλου στην χρονοσειρά A.

2.2 Χρονοσειρά B

Όπως και με την χρονοσειρά A, πρώτα θα σχηματίσουμε το διάγραμμα αυτοσυσχετίσεων της χρονοσειράς για να δούμε κατά πόσο υπάρχουν συσχετίσεις μεταξύ χρονικών στιγμών που απέχουν μέχρι και T=100. Ο λόγος που επιλέχθηκε αυτή η τιμή για το T είναι ο ίδιος που συζητήθηκε και πρηγουμένως στην χρονοσειρά A. Όπως βλέπουμε στο Σχ.4, η αυτσυσχέτιση εμφανίζει υψηλές τιμές και φθίνει με πολύ αργό ρυθμό (χαρακτηριστικό είναι το γεγονός ότι ακόμα και για υστέρηση 100 έχουμε μεγάλη τιμή αυτοσυσχέτισης). Επιπλέον, φαίνεται να



(a) Χρονοσειρά Α χωρίς τάση (b) Διάγραμμα αυτοσυσχετίσεων

Figure 3: Διαγράμματα για την χρονοσειρά Α αφού εφαρμόσουμε το φίλτρο κινούμενου μέσου

εμφανίζονται spikes (φθίνοντος μεγέθους) με περίοδο 7. Τα παραπάνω είναι χαρακτηριστικά χρονοσειράς που εμφανίζει τάση και περιοδικότητα για αυτό και το πρώτο βήμα μας θα είναι να τις απαλείψουμε.

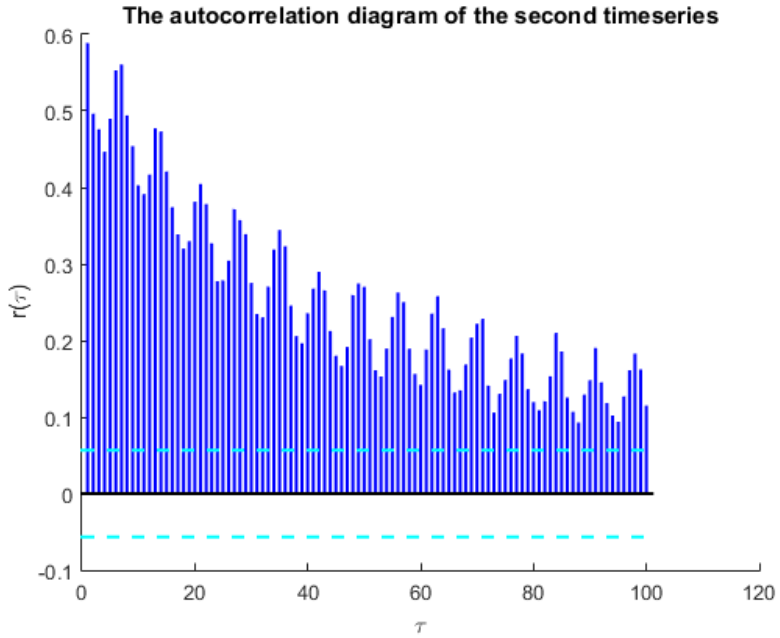
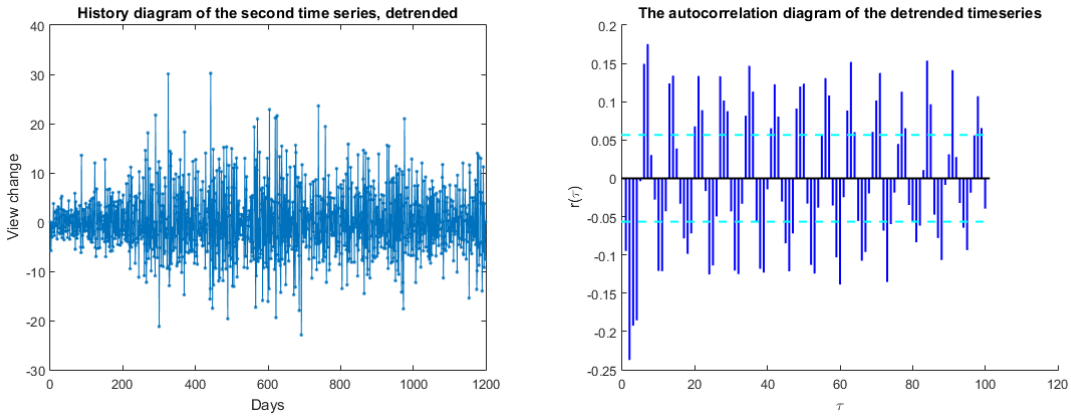


Figure 4: Διάγραμμα αυτοσυσχετίσεων της χρονοσειράς Β

Σε πρώτη φάση θα απαλείψουμε την τάση της χρονοσειράς Β. Όπως και με την χρονοσειρά Α, έτσι και εδώ, ακριβώς επειδή δεν γνωρίζουμε αν η τάση μπορεί να περιγραφεί με καθοριστικό τρόπο, αποφεύγουμε να χρησιμοποιήσουμε κάποια παραμετρική συνάρτηση f(t) του χρόνου και για αυτό χρησιμοποιούμε φίλτρο κινούμενου μέσου. Και πάλι μας ενδιαφέρει να απαλείψουμε κυρίως τις μεταβολές της τάσης σε μικρότερη κλίμακα για αυτό και η τάξη του μοντέλου που θα χρησιμοποιήσουμε θα είναι μικρή. Πιο συγκεκριμένα, πιο αποτελεσματική δουλειά φαίνεται



(a) Χρονοσειρά Β χωρίς τάση (b) Διάγραμμα αυτοσυσχετίσεων

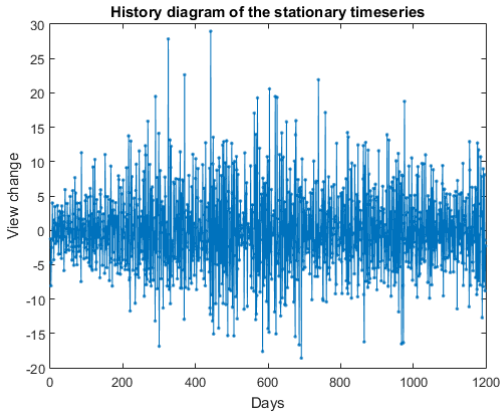
Figure 5: Διαγράμματα για την χρονοσειρά Β αφού εφαρμόσουμε το φίλτρο κινούμενου μέσου

να κάνει το μοντέλο κινούμενου μέσου τάξης 5 οπότε θα χρησιμοποιήσουμε αυτό. Στο Σχ.5 φαίνεται η χρονοσειρά και το διάγραμμα αυτοσυσχετίσεων αυτής αφού εφαρμόσουμε το φίλτρο κινούμενου μέσου τάξης 5.

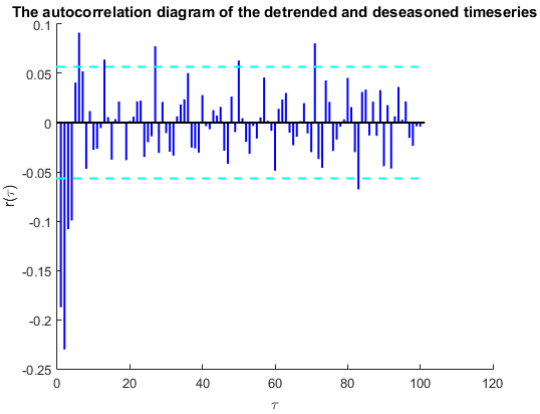
Από το Σχ.5α φαίνεται πως ο σκοπός μας απαλοιφής της τάση έχει επιτευχθεί μιας και οι τιμές της χρονοσειράς είναι κεντραρισμένες γύρω από το 0. Παρόλα αυτά, όμως, βλέπουμε ότι ακόμα εμφανίζεται περιοδικότητα στην χρονοσειρά, κάτι που γίνεται ξεκάθαρο στο διάγραμμα αυτοσυσχετίσεων στο Σχ.5β. Συγκεκριμένα, με εξαίρεση τις πρώτες 4 τιμές, από κει και πέρα φαίνεται να υπάρχει κάποια περιοδικότητα στις τιμές που εμφανίζονται με περίοδο 7.

Το επόμενο μας βήμα, λοιπόν, είναι η απαλοιφή της περιοδικότητας. Πιο συγκεκριμένα, δοκιμάστηκε απαλοιφή της περιοδικότητας και με φίλτρο κινούμενου μέσου και με seasonal components filter (με περίοδο 7 και στις 2 περιπτώσεις), με το δεύτερο να φαίνεται να κάνει καλύτερη δουλειά οπότε και προτιμήθηκε. Στο Σχ.6 φαίνεται η χρονοσειρά και το διάγραμμα αυτοσυσχετίσεων αυτής αφού εφαρμόσουμε το seasonal components filter τάξης 7.

Από το διάγραμμα της χρονοσειράς που προέκυψε (Σχ.6α) δεν μπορούμε να δούμε με το μάτι κάποια σημαντική διαφορά. Όμως, μελετώντας το διάγραμμα των αυτοσυσχετίσεων (Σχ.6β) είναι πλέον φανερό ότι έχουμε στατιστικά σημαντική αυτοσυσχέτιση για τις πρώτες 4 υστερήσεις άρα μπορούμε να αντλήσουμε κάποια πληροφορία από αυτές (η χρονοσειρά δεν είναι λευκός θόρυβος) ενώ από εκεί και μετά (με εξαίρεση κάποιες μεμονομένες τιμές που και πάλι όμως δεν είναι πολύ μεγάλες) είναι γενικά στατιστικά ίσες με το 0. Επιπλέον, δεν βλέπουμε να υπάρχει κάποια περιοδικότητα στις εμφανιζόμενες τιμές της αυτοσυσχέτισης κάτι που υποδεικνύει ότι η εποχικότητα απαλείφθηκε επιτυχώς. Επομένως, μπορούμε με ασφάλεια να ισχυριστούμε ότι η χρονοσειρά που έχει προκύψει είναι στάσιμη. Όπως και πριν, η μόνη μας αμφιβολία έχει να κάνει με την μεταβολή της διασποράς των τιμών με τον χρόνο αλλά και πάλι δεν θεωρούμε ότι είναι αρκετή για να μας δημιουργήσει πρόβλημα στην προσαρμογή κάποιου γραμμικού μοντέλου. Έτσι, λοιπόν, είμαστε πλέον σε θέση να εφαρμόσουμε την μέθοδο Box-Jenkins για την προσαρμογή γραμμικού μοντέλου στην χρονοσειρά Β.



(a) Χρονοσειρά B χωρίς τάση και περιοδικότητα



(b) Διάγραμμα αυτοσυσχετίσεων

Figure 6: Διαγράμματα για την χρονοσειρά B αφού απαλείψουμε την τάση και την εποχικότητα

3 Γραμμική ανάλυση

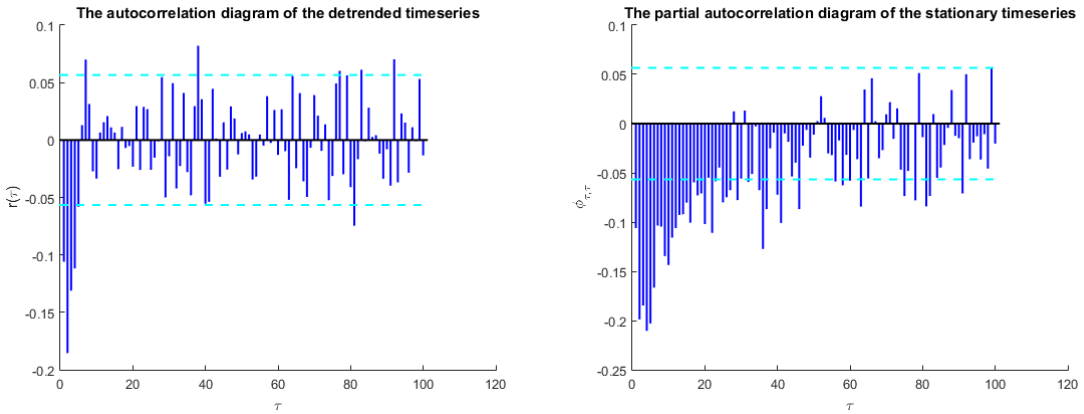
3.1 Χρονοσειρά A

3.1.1 Προσαρμογή γραμμικού μοντέλου

Γνωρίζουμε ότι αν θέλουμε να προσαρμόσουμε σε μια χρονοσειρά ένα μοντέλο AR(p) μπορούμε να επιλέξουμε το p ως την τιμή αυτή για την οποία οι μερικές αυτοσυσχετίσεις για υστερήσεις τ¿p γίνονται στατιστικά ίσες με το 0. Αντίστοιχα, αν θέλουμε να προσαρμόσουμε ένα MA(q) μοντέλο μπορούμε να επιλέξουμε το q ως την τιμή αυτή για την οποία για υστερήσεις τ¿q η συνάρτηση αυτοσυσχέτισης της χρονοσειράς γίνεται στατιστικά ίση με 0. Βέβαια, κατά την προσαρμογή ενός ARMA(p,q) μοντέλου τα κριτήρια αυτά δεν μπορούν να θεωρηθούν ασφαλή και έτσι καταφεύγουμε στην χρήση κριτηρίων πληροφορίας (πχ AIC,BIC,FPE) για την ορθή εκτίμηση των παραμέτρων p,q. Παρόλα αυτά, και καθαρά για λόγους σύγκρισης των μοντέλων, παραθέτουμε τα διαγράμματα αυτοσυσχέτισης και μερικής αυτοσυσχέτισης της χρονοσειράς στο Σχ.7

Όπως βλέπουμε από τα δυο αυτά διαγράμματα, αν χρησιμοποιούσαμε μοντέλο AR αυτό θα έπρεπε να είναι τουλάχιστον τάξης p=20 ενώ αν χρησιμοποιούσαμε MA μοντέλο αυτό θα έπρεπε να είναι τουλάχιστον τάξης q=5 (ίσως και q=7). Βλέπουμε ότι οι τιμές αυτές (ιδιαίτερα για το AR μοντέλο) είναι αρκετά μεγάλες και θα καθιστούσαν το γραμμικό μοντέλο που προσαρμόζουμε αρκετά πολύπλοκο. Αναμένουμε λοιπόν ένα κατάλληλο ARMA μοντέλο να δώσει καλύτερες τιμές για τα p,q.

Για τον προσδιορισμό των παραμέτρων p,q χρησιμοποιήσαμε το κριτήριο πληροφορίας AIC. Πιο συγκεκριμένα, υπολογίσαμε την τιμή AIC για κάθε δυνατό μοντέλο ARMA(p,q) για τιμές των p,q μέχρι και 10. Από αυτά τα 100 μοντέλα, τελικά επιλέχτηκε αυτό το οποίο επέστρεψε



(a) Διάγραμμα αυτοσυσχετίσεων (b) Διάγραμμα μερικών αυτοσυσχετίσεων

Figure 7: Χρονοσειρά A

την μικρότερη τιμή για το κριτήριο AIC. Πιο συγκεκριμένα, τρέχοντας το συγκεκριμένο κομμάτι του κώδικα στο Matlab προκύπτει ότι το μοντέλο με την καλύτερη προσαρμογή είναι το ARMA(2,3). Αυτό φαίνεται να συμφωνεί και με τις προβλέψεις μας για τιμές p¡20 και q¡7 οπότε φαίνεται να είναι κατάλληλο για να προχωρήσουμε με αυτό την ανάλυσή μας (δεν παρατίθενται εδώ διαγράμματα για τις τιμές του κριτηρίου AIC για τα διάφορα p,q μιας και θα ήταν δυσνόητη η ανάγνωση του προκύπτοντος διαγράμματος για τόσες τιμές των p,q). Αξίζει, επίσης, να σημειωθεί ότι για ορισμένους συνδυασμούς των p,q προέκυπταν μοντέλα τα οποία είτε δεν ήταν αντιστρέψιμα είτε δεν ήταν στάσιμα. Στις περιπτώσεις αυτές τα μοντέλα αυτά απορρίπτονταν κατευθείαν και κρατήθηκαν ως πιθανά μόνο όσα ήταν και αντιστρέψιμα και στάσιμα.

Τιμές Συντελεστών του ARMA(2,3)		
φ_0	φ_1	φ_2
-0.000307	1.293138	-0.466618
θ_1	θ_2	θ_3
1.796555	-0.676364	-0.122052
SD of noise: 2.331939		
AIC: 1.701783		
FPE: 5.483524		

Table 1: Στοιχεία του προσαρμοσμένου ARMA(2,3) μοντέλου

Στον πίνακα 1 φαίνονται διάφορα στοιχεία του μοντέλου που προσαρμόσαμε στην στάσιμη χρονοσειρά. Επιπλέον, στο Σχ.8 φαίνεται το NRMSE πρόβλεψης για βήμα μπροστά μέχρι και 5. Βλέπουμε ότι για πρόβλεψη ενός βήματος μπροστά το NRMSE είναι λίγο πάνω από 0.8 ενώ καθώς το βήμα αυξάνεται αυτό τείνει στο 1. Βλέπουμε, λοιπόν, ότι το NRMSE είναι σχετικά μεγάλο ακόμα και για πρόβλεψη ενός βήματος αλλά παρόλα αυτά το μοντέλο μπορεί να κάνει καλύτερη πρόβλεψη από το μοντέλο τυχαίου περιπάτου όπου απλά θα παίρναμε την μέση τιμή των εως τώρα παρατηρήσεων.

Τέλος, θα εφαρμόσουμε και τον έλεγχο Portmanteau στα υπόλοιπα του προσαρμοσμένου μον-

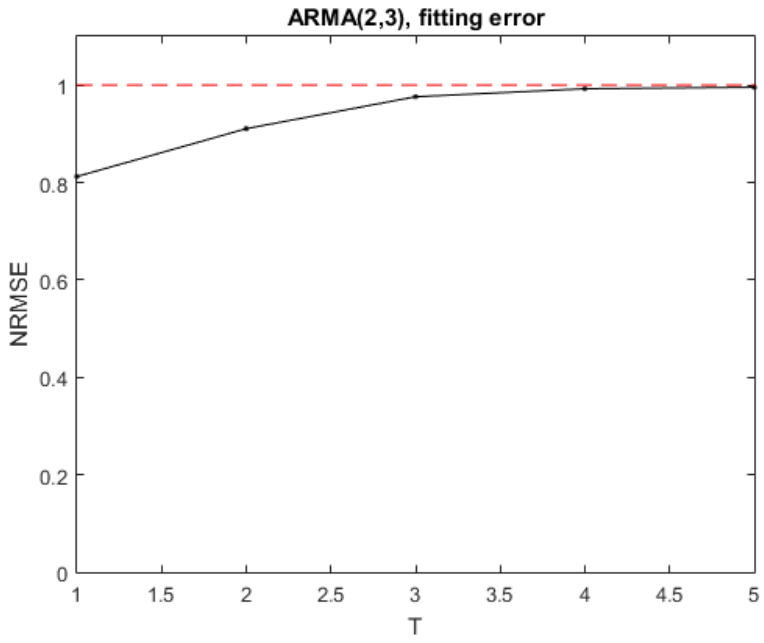
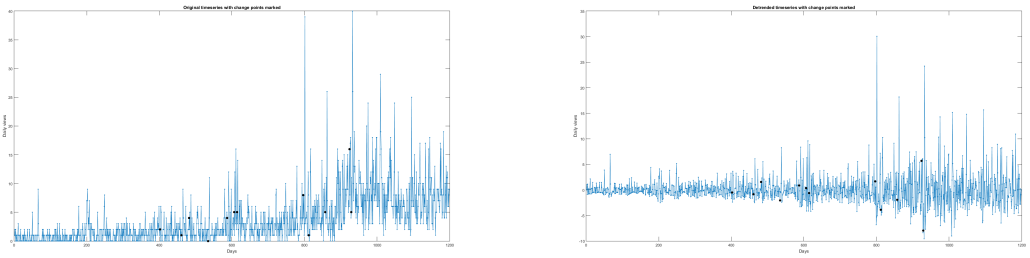


Figure 8: NRMSE για πρόβλεψη μέχρι και 5 βήματα μπροστά

τέλου για να βεβαιωθούμε ότι αυτά είναι λευκός θόρυβος και δεν "κρύβουν" κάποια πληροφορία που δεν μπορεί να ανιχνέυσει το μοντέλο. Από το Σχ.9 βλέπουμε ότι οι p-τιμές είναι αρκετά μεγάλες οπότε δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα υπόλοιπα είναι λευκός θόρυβος. Συνεπώς, από όλα τα παραπάνω, μπορούμε να συμπεράνουμε ότι έχουμε καλή προσαρμογή του μοντέλου στην χρονοσειρά.

Figure 9: Έλεγχος Portmanteau στα υπόλοιπα της χρονοσειράς



(a) Αρχική χρονοσειρά (b) Στάσιμη χρονοσειρά

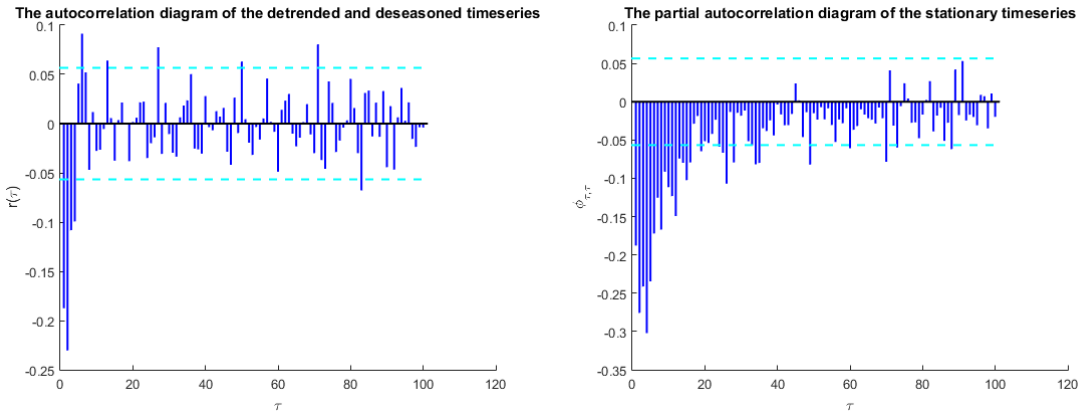
Figure 10: Σημεία αλλαγής στην χρονοσειρά A

3.1.2 Εντοπισμός σημείων αλλαγής

Όπως αναφέρθηκε και στην εισαγωγή, το βασικό κομμάτι αυτής της εργασίας αποτελεί η εύρεση σημείων αλλαγής στην χρονοσειρά. Πιο συγκεκριμένα, αναμένουμε τα σημεία αυτά που θα βρούμε να αντιστοιχίζονται σε στιγμές κατά τις οποίες έχουμε ραγδαία αύξηση η πτώση των ημερήσιων προβολών (έξω από τα προβλεπόμενα όρια) οι οποίες κατά πάσα πιθανότητα θα εμφανίζονται ως spikes στην χρονοσειρά. Για τον εντοπισμό τέτοιων σημείων χρησιμοποιήθηκε ως στατιστικό ο μέσος όρος των απόλυτων τιμών των σφαλμάτων πρόβλεψης έως και T βήματα μπροστά. Πιο συγκεκριμένα, το T επιλέχθηκε να είναι ίσο με 5 αφού η προβλεπτική ισχύς του μοντέλου δεν είναι πολύ μεγάλη οπότε δεν έχει ιδιαίτερο νήμα να προβλέπουμε για πολλά βήματα μπροστά. Επιπλέον, θεωρούμε ότι έχουμε σημαντική αλλαγή όταν η τιμή του στατιστικου μας ξεπερνάει την τιμή α όπου α=2s, με s την τυπική απόκλιση των παρατηρήσεων στο σύνολο εκμάθησης του μοντέλου.

Επιλέχθηκε το μοντέλο να χρησιμοποιεί κάθε φορά ως σύνολο εκμάθησης τις 400 τελευταίες παρατηρήσεις προκειμένου να μπορεί να προσαρμόζεται καλύτερα κάθε φορά στα νέα δεδομένα. Βλέπουμε ότι το σύνολο εκμάθησης είναι σχετικά μικρό (περίπου το 1/3 των συνολικών παρατηρήσεων) οπότε από την μια είμαστε σε μεγάλο βαθμό σίγουροι ότι δεν έχουμε overfitting στα συγκεκριμένα δεδομένα, από την άλλη ίσως δεν είναι και αρκετές οι παρατηρήσεις για να έχει ικανοποιητική προβλεπτική ισχύ το μοντέλο.

Βλέπουμε από τα διαγράμματα του Σχ.10 ότι ο αλγόριθμός μας έχει εντοπίσει 12 σημεία αλλαγής στην χρονοσειρά (μαύρες κουκίδες). Παρατηρώντας τα με το μάτι βλέπουμε ότι κάποια από αυτά βρίσκονται πριν από κάποια μεγάλη αύξηση ή πτώση της τιμής της χρονοσειράς (όπως αναμέναμε) ενώ άλλα φαίνεται να βρίσκονται σε θέσεις που δεν μπορούν να ερμηνευτούν με αυτόν τον τρόπο. Αντίστοιχα, υπάρχουν και spikes στα οποία όμως δεν έχουμε σημεία αλλαγής. Από τα παραπάνω, λοιπόν, γίνεται αντιληπτό ότι δεν μπορούμε να ισχυριστούμε ότι η προσέγγιση αυτή φαίνεται να δίνει χρήσιμα σημεία με αυτόματο τρόπο. Ίσως με κατάλληλη παραμετροποίηση των T και α να λαμβάναμε καλύτερα αποτελέσματα, όμως το βασικό πρόβλημα φαίνεται να είναι η αδυναμία του μοντέλου να κάνει πολύ καλές προβλέψεις. Αν ήμασταν σίγουροι ότι το μοντέλο μας έχει πολύ καλή προβλεπτική ικανότητα τότε θα μπορούσαμε να πούμε ότι τα σημεία αλλαγής που εντοπίζει είναι όντως σημεία που προκύπτουν από έντονες μεταβολές προερχόμενες από μη προβλεπόμενους (εξωτερικούς) παράγοντες. Τώρα όμως που



(a) Διάγραμμα αυτοσυσχετίσεων (b) Διάγραμμα μερικών αυτοσυσχετίσεων

Figure 11: Χρονοσειρά B

υπεσέρχεται και σε μέγαλο βαθμό το λάθος πρόβλεψης, δεν μπορούμε να ισχυριστούμε κάτι τέτοιο με ασφάλεια.

3.2 Χρονοσειρά B

3.2.1 Προσαρμογή γραμμικού μοντέλου

Τα βήματα που θα ακολουθήσουμε σε αυτήν την περίπτωση θα είναι τα ίδια που εφαρμόσαμε και για την χρονοσειρά A. Σε πρώτη φάση θα σχηματίσουμε τα διαγράμματα αυτοσυσχετίσεων και σχετικών αυτοσυσχετίσεων για να δούμε ποια θα ήταν τα κατάλληλα μοντέλα AR(p) και MA(q) για προσαρμοστούν στην χρονοσειρά.

Από το διάγραμμα μερικής αυτοσυσχέτισης βλέπουμε ότι αυτή γίνεται στατιστικά ίση με το 0 πρωτη φορά για τ=16 υποδεικνύοντας ότι με βάσει αυτού του κριτηρίου θα έπρεπε να επιλέξουμε για να προσαρμόσουμε ένα AR(16) μοντέλο. Αντίστοιχα, από το διάγραμμα των αυτοσυσχετίσεων, αυτές γίνονται για πρώτη φορά στατιστικά ίσες με το 0 για τ=4 (θα μπορούσαμε να επιλέξουμε πάντως και τ=6) υποδεικνύοντας ότι ικανοποιητικό θα ήταν ένα MA(4) (ή MA(6)) μοντέλο.

Στην συνέχεια, θα χρησιμοποιήσουμε το κριτηριο πληροφορίας AIC με αντίστοιχο τρόπο με πριν (για κάθε πιθανή τιμή των p,q από 0 ως 10) ευελπιστώντας να βρούμε ένα κατάλληλο ARMA μοντέλο με μικρότερα p,q από τα παραπάνω. Όπως και πριν, έτσι και τώρα μοντέλα τα οποία ήταν μη στάσιμα ή μη αντιστρέψιμα δεν λήφθηκαν υπόψιν. Τρέχοντας, λοιπόν το κομμάτι αυτό του προγράμματος μας βλέπουμε ότι το μοντέλο με την μικρότερη τιμή για το κριτήριο AIC είναι το ARMA(10,6). Βλέπουμε ότι οι τιμές των p,q είναι μικρότερες (ή ίσες) σε σύγκριση με αυτές των AR και MA μοντέλων που υπολογίστηκαν αλλά παρόλα αυτά παραμένουν αρκετά μεγάλες. Επιπλέον, διαπιστώσαμε ότι τα μοντέλα ARMA(3,3) και ARMA(4,7) έδωσαν αρκετά κοντινές τιμές για το κριτήριο AIC (ελαφρώς μεγαλύτερες). Αρχικά, θεωρήσαμε ότι θα ήταν

καλύτερα να χρησιμοποιούσαμε το μοντέλο ARMA(3,3) λόγω της μικρότερης πολυπλοκότητάς του. Παρόλα αυτά, όμως, είναι γεγονός ότι κατά τον υπολογισμό της τιμής AIC λαμβάνεται υπόψιν και η πολυπλοκότητα του μοντέλου (μεγάλη πολυπλοκότητα αυξάνει την τιμή) οπότε δεν υπάρχει σοβαρός λόγος για να μην την προτιμήσουμε. Άλλωστε, οι τιμές αυτές για τα p,q δεν είναι και απαγορευτικά μεγάλες, κάτι που παρατηρούμε και από το γεγονός ότι δεν έχουμε δραματική αύξηση του χρόνου εκτέλεσης για αύτο το ARMA μοντέλο. Επομένως, κρατάμε τελικά το ARMA(10,6) ως το καταλληλότερο.

Τιμές των συντελεστών φ του ARMA(10,6)											
Δείκτης	0	1	2	3	4	5	6	7	8	9	10
Τιμή	0	1.07	-0.104	0.113	0.384	-0.654	0.389	-0.093	-0.064	-0.001	-0.04

Table 2: Συντελεστές φ

Δείκτης	1	2	3	4	5	6
Τιμή	1.78	-0.412	-0.199	0.244	-1.13	0.718

Table 3: Συντελεστές ϑ

SD of noise	4.626837
AIC	3.077112
FPE	21.687942

Table 4: Χαρακτηριστικά του μοντέλου

Στους 3 παραπάνω πίνακες φαίνονται αναλυτικά οι τιμές των συντελεστων του μοντέλου ARMA(10,6) που προσαρμόσαμε στην χρονοσειρά καθώς και ορισμένες άλλες μετρικές όπως τα κριτήρια AIC και FPE και η διακύμανση των υπολοίπων. Μια παρατήρηση που μπορούμε να κάνουμε εδώ είναι ότι η διακύμανση του θορύβου φαίνεται να είναι μεγαλύτερη σε σχέση με την αντίστοιχη όταν προσαρμόσαμε το γραμμικό μας μοντέλο στην χρονοσειρά A κάτι που ίσως επηρεάσει στην συνέχεια την εύρεση των σημείων αλλαγής.

Επιπλέον, στο Σχ.12 φαίνεται το NRMSE πρόβλεψης για βήμα μπροστά μέχρι και 5. Παρατηρούμε πως για πρόβλεψη ενός βήματος μπροστά το NRMSE έχει τιμή 0.75 ενώ καθώς αυξάνεται το βήμα πρόβλεψης η τιμή του NRMSE τείνει στο 1. Σε σχέση με το μοντέλο που προσαρμόσαμε στην χρονοσειρά A βλέπουμε εδώ ότι το μοντέλο αυτό έχει κάπως καλύτερη ικανότητα πρόβλεψης. Βέβαια, οι τιμές του NRMSE παραμένουν αρκετά υψηλές ακόμα και για μικρά χρονικά βήματα, όμως είμαστε σίγουροι ότι οι προβλέψεις μας θα είναι κάπως καλύτερες από αυτές που θα παίρναμε αν χρησιμοποιούσαμε την μέση τιμή.

Τέλος, θα εφαρμόσουμε και τον έλεγχο Portmanteau στα υπόλοιπα του προσαρμοσμένου μοντέλου για να βεβαιωθούμε ότι αυτά είναι λευκός θόρυβος και δεν "κρύβουν" κάποια πληροφορία που δεν μπορεί να ανιχνέυσει το μοντέλο. Στο Σχ.13 βλέπουμε ότι οι p-τιμές είναι αρκετά μεγάλες (σχεδόν ίσες με 1) οπότε δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα υπόλοιπα είναι λευκός θόρυβος. Συνεπώς, από όλα τα παραπάνω μπορούμε να συμπεράνουμε ότι έχουμε καλή προσαρμογή του μοντέλου στην χρονοσειρά.

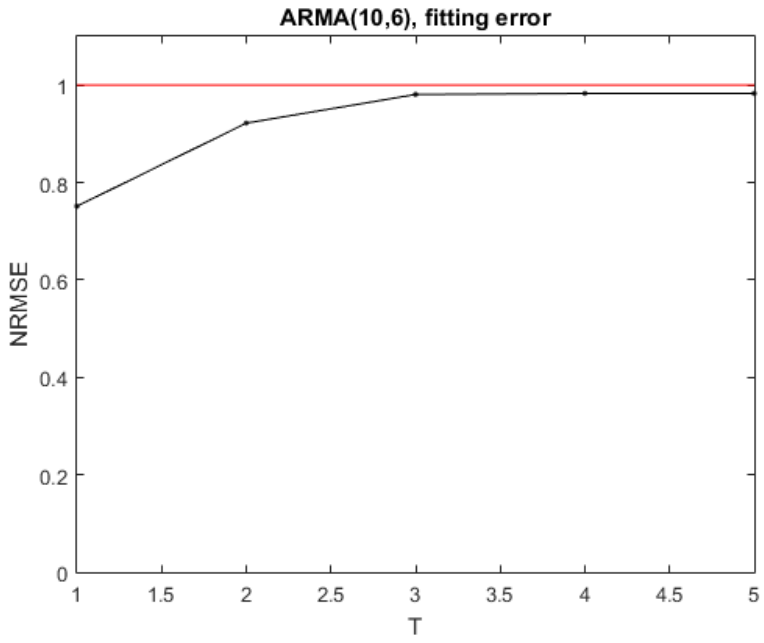


Figure 12: NRMSE για πρόβλεψη μέχρι και 5 βήματα μπροστά

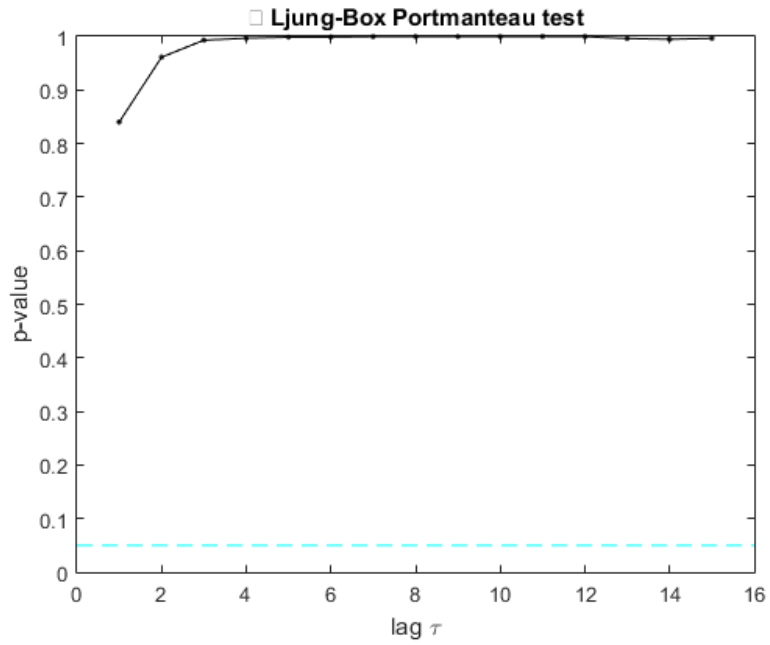
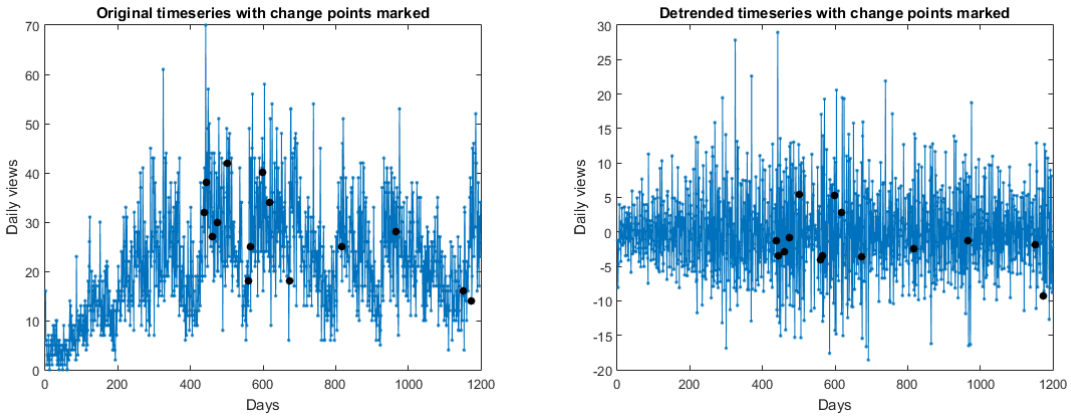


Figure 13: Έλεγχος Portmanteau στα υπόλοιπα της χρονοσειράς

3.2.2 Εντοπισμός σημείων αλλαγής

Για την εύρεση των σημείων αλλαγών θα εφαρμόσουμε την ίδια διαδικασία με αυτήν που χρησιμοποιήθηκε για την χρονοσειρά Α. Πιο συγκεκριμένα,θα υπολογίζουμε σε κάθε περίπτωση το ίδιο στατιστικό (μέσος όρος των απόλυτων τιμών των σφαλμάτων πρόβλεψης έως και T



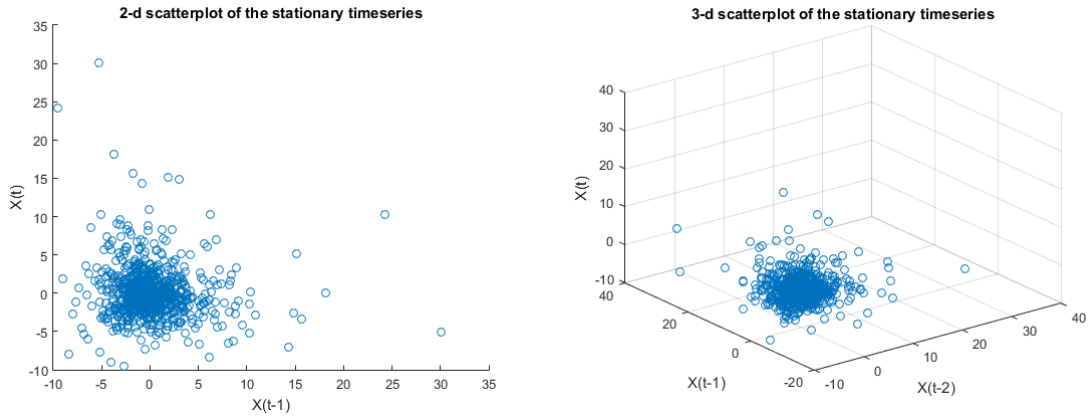
(a) Αρχική χρονοσειρά (b) Στάσιμη χρονοσειρά

Figure 14: Σημεία αλλαγής στην χρονοσειρά B

βήματα μπροστά). Εδώ, επιλέχθηκε και πάλι το T=5 γιατί ενώ έχουμε λίγο καλύτερη προβλεπτική ικανότητα σε σχέση με το μοντέλο της χρονοσειράς A, και πάλι δεν είναι αρκετά μεγάλη για να μας παρέχει αξιόπιστες προβλέψεις για μεγάλα T (το NRMSE τείνει γρήγορα στο 1). Όπως αναφέραμε στην αρχή της αναφοράς, οι παρατηρήσεις της χρονοσειράς αυτής εμφανίζουν μεγαλύτερη διακύμανση γύρω από την μέση τιμή σε σχέση με τις αντίστοιχες παρατηρήσεις της χρονοσειράς A. Επιπλέον, είναι γνωστό ότι ως μέτρο σύγκρισης για το στατιστικό μας (για να βρούμε τα σημεία αλλαγής) χρησιμοποιούμε την τιμή της τυπικής απόκλισης των παρατηρήσεων. Δεδομένου ότι το μοντέλο μας έχει καλύτερη προβλεπτική ικανότητα σε σχέση με αυτό του A (άρα θα έχουμε μικρότερα σφάλματα, άρα η τιμές του στατιστικού θα είναι μικρότερες) είναι λογικό η τιμή σύγκρισης α να είναι αναλογικά μικρότερη για αυτήν την περίπτωση. Έτσι, εν τέλει θέτουμε το α=1.5s (ενώ για την χρονοσειρά A είχαμε α=2s).

Ως σύνολο εκμάθησης χρησιμοποιήθηκαν πάλι 400 παρατηρήσεις οπότε ισχύουν και εδώ όσα αναφέραμε για την μέθοδο αυτή στην χρονοσειρά A. Συγκεκριμένα, επιλέχθηκε το μοντέλο να χρησιμοποιεί κάθε φορά ως σύνολο εκμάθησης τις 400 τελευταίες "γνωστές" παρατηρήσεις προκειμένου να μπορεί να προσαρμόζεται καλύτερα κάθε φορά στα νέα δεδομένα.

Στο Σχ.14 έχουν σημειωθεί με μαύρες κουκίδες τα σημεία αλλαγής πάνω στην αρχική χρονοσειρά και στην στάσιμη που προέκυψε από αυτήν (συνολικά 14). Για την χρονοσειρά A αναφέραμε ότι είναι δύσκολη η ακριβής εξαγωγή άμεσων συμπερασμάτων από τα σημεία που προέκυψαν μιας και το σφάλμα πρόβλεψης είναι αρκετά μεγάλο και δεν μπρούμε να εμπιστευτούμε την όλη διαδικασία. Κάτι αντίστοιχο ισχύει και για την χρονοσειρά B. Επιπλέον, ακριβώς επειδή είναι ούτως η άλλως μεγάλη η διακύμανση των παρατηρούμενων τιμών και υπάρχουν πολλά spikes, γίνεται ακόμα πιο δύσκολο να πούμε ξεκάθαρα ποια από αυτά είναι αναμενόμενα και ποιά όχι. Αν είχαμε προσαρμόσει στην χρονοσειρά ένα μοντέλο για το οποίο θα ήμασταν σίγουροι ότι έχει μεγάλη προβλεπτική ικανότητα (μικρό NRMSE) τότε θα μπορούσαμε να πούμε με μεγαλύτερη σιγουριά ότι τα σημεία που εντόπισε είναι όντως τα ζητούμενα σημεία αλλαγής. Τώρα, όμως που δεν ισχύει κάτι τέτοιο, δεν μπορούμε να είμαστε και τόσο σίγουροι για τα αποτελέσματά μας και τι συμπεράσματα μπορούν να εξαχθούν από αυτά.



(a) 2 διαστάσεων (b) 3 διαστάσεων

Figure 15: Διαγράμματα διασποράς χρονοσειράς A

4 Μη γραμμική ανάλυση

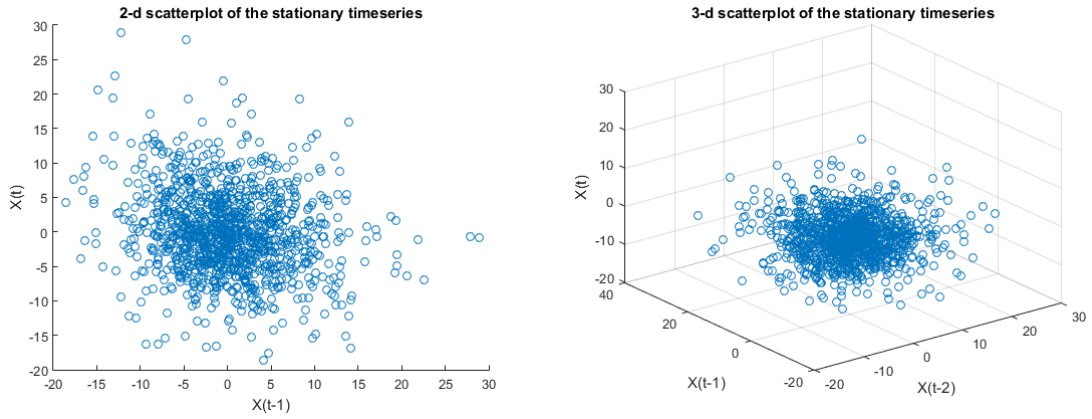
Εισαγωγική σημείωση: Και στις 2 περιπτώσεις χρησιμοποιούμε τις στάσιμες χρονοσειρές, όπως αυτές προέκυψαν από την ανάλυση παραπάνω.

4.1 Διαγράμματα διασπορών

Σε πρώτη φάση θα σχηματίσουμε τα διαγράμματα διασπορών (scatter plots) των χρονοσειρών για 2 και 3 διαστάσεις προκειμένου να διαπιστώσουμε αν μπορούμε να διακρίνουμε κάποια δομή σε αυτά που θα μας υποδεικνύει την ανάγκη προσαρμογής κάποιου μη γραμμικού χαοτικού μοντέλου.

4.1.1 Χρονοσειρά A

Όπως βλέπουμε από το Σχ.15α τα σημεία βρίσκονται κατανεμημένα κυρίως γύρω από το (0,0) και με τέτοιο τρόπο που φαίνεται να είναι τυχαία απλωμένα γύρω από αυτό. Κάτι αντίστοιχο βλέπουμε να ισχύει και στις 3 διαστάσεις επίσης. Επομένως, δεν μπορούμε να συμπεράνουμε κοιτώντας μόνο τα διαγράμματα διασπορών αν υπάρχει κάποιος χαοτικός μηχανισμός ο οποίος παράγει τη εν λόγω χρονοσειρά. Οι παρατηρήσεις προέρχονται είτε από κάποια αμιγώς στοχαστική διαδικασία η από κάποιο δυναμικό σύστημα του οποίου όμως ο ελκυστής είναι διάστασης ¿3 και για αυτό δεν μπορεί να περιγραφεί ορθά σε αυτά τα διαγράμματα (έχουμε αναδίπλωση).



(a) 2 διαστάσεων (b) 3 διαστάσεων

Figure 16: Διαγράμματα διασποράς χρονοσειράς B

4.1.2 Χρονοσειρά B

Σχηματίζοντας τα διαγράμματα διασπορών για 2 και 3 διαστάσεις (βλ. Σχ.16) βλέπουμε και πάλι ότι δεν μπορούμε να διακρίνουμε κάποια δομή που να παραπέμπει σε κάποιον παράξενο ελκυστή. Τα σημεία είναι και πάλι κατανεμημένα με φαινομενικά τυχαίο τρόπο γύρω από το 0 οπότε πάλι δεν μπορούμε να συμπεράνουμε μόνο από τα διαγράμματα διασπορών αν υπάρχει κάποιος χαοτικός μηχανισμός (με ελκυστη που μπορεί να περιγραφεί σε μέχρι 3 διαστάσεις) που να παράγει την χρονοσειρά B. Αξίζει να σημειωθεί ότι φαίνεται ξεκάθαρα ότι οι παρατηρήσεις της χρονοσειράς B παρουσιάζουν μεγαλύτερη διακύμανση σε σχέση με της A αφού τα σημεία είναι πιο απλωμένα σε σχέση με αυτά στα διαγράμματα διασποράς της A (το κέντρο όμως παραμένει το 0).

4.2 Επιλογή παραμέτρων

4.2.1 Χρονοσειρά A

α)Υστέρηση τ: Θεωρήσαμε ότι η χρονοσειρά A πρόκειται για χρονοσειρά από απεικονίσεις, δηλαδή από διακριτό σύστημα μιας και ο αριθμός προβολών μεταβάλλεται μέσα στην μέρα αλλά εμείς μελετάμε μόνο το σύνολο των προβολών σε μια μέρα. Για αυτό και επιλέγουμε να έχουμε τ=1.

β)Διάσταση εμβύθισης m: Για την επιλογή του m χρησιμοποιήθηκε η μέθοδος των ψευδών κοντινότερων γειτόνων (FNN). Στο Σχ.17, από το διάγραμμα με το ποσοστό ψευδών γειτόνων για διάφορες τιμες της διάστασης εμβύθισης, βλέπουμε ότι για τις τιμές του m που έχουμε αποτελέσματα (μέχρι m=5) καμία από αυτές δεν έχει επαρκώς μικρό ποσοστό ψευδών γειτόνων. Έτσι, λοιπόν, δεν μπορούμε να καταλήξουμε με σιγουριά σε κάποιο συμπέρασμα. Παρόλα αυτα, μπορούμε να θεωρήσουμε την τιμή m=5 ως μια αρκετά πιθανή υποψήφια για να

επιλεγεί. Πρωτού λάβουμε την τελική απόφαση θα εξετάσουμε και την διάσταση συσχέτισης ν.

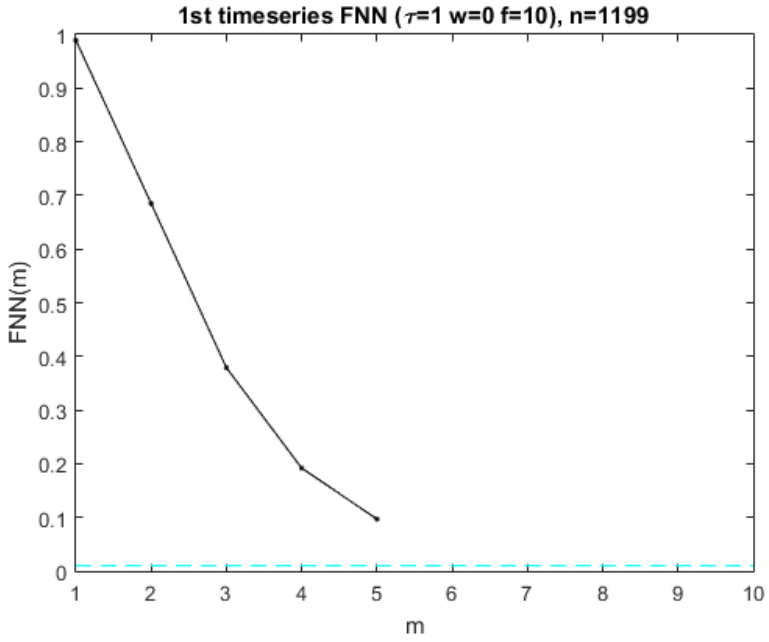
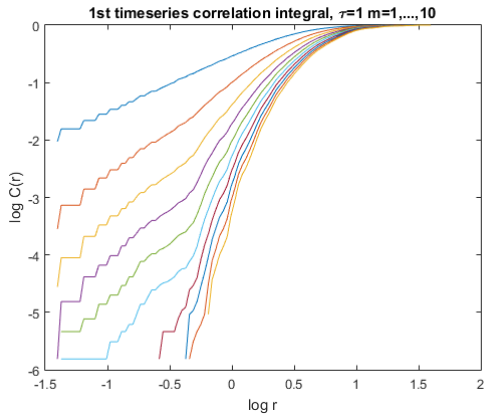


Figure 17: Διάγραμμα ποσοστού ψευδών γειτόνων για την χρονοσειρά Α

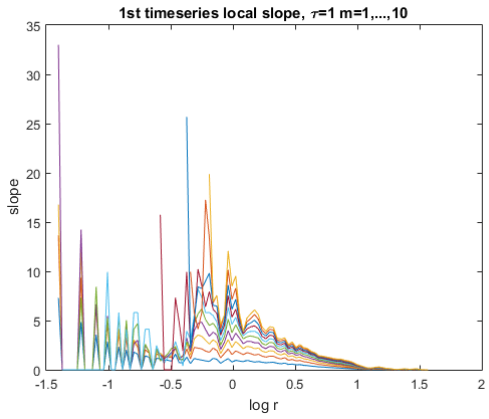
γ)Διάσταση συσχέτισης ν: Στο Σχ. 18 και 19 βλέπουμε τα διγράμματα που προέκυψαν απο την ανάλυση για την εύρεση της διάστασης συσχέτισης του ελκυστή. Από τα πρώτα διαγράμματα βλέπουμε πως είναι δύσκολο να ορίσουμε ξεκάθαρα μια κατάλληλη περιοχή κλιμάκωσης αφού αυτή έχει συρρικνωθεί σε μεγάλο βαθμό από τα αριστερά λογω της φτωχής στατιστικής (δεν έχουμε πολλές παρατηρήσεις) ενώ γενικά και η ύπαρξη θορύβου επιδρά αρνητικά. Από το Σχ.18α παρατηρούμε ότι η εκτίμηση της διάστασης συσχέτισης δεν ξεπερνάει σε καμία περίπτωση το 2 βάσει του οποίου μπρούμε να θεωρήσουμε ότι με μεγάλο ποσοστό βεβαιότητας μπορούμε να περιγράψουμε επαρκώς το σύστημα με 2 ανεξάρτητες μεταβλητές. Χαρακτηριστικό είναι επίσης και το γεγονός ότι για τιμές του m μεγαλύτερες του 2 η εκτίμηση του ν μας δίνει μικρότερες τιμές από ότι για m=1 και m=2. Κάτι τέτοιο ίσως οφείλεται στο γεγονός ότι λόγω της αύξησης διάστασης τα σημεία απλώνονται πολύ πιο αραιά στον χώρο με αποτέλεσμα να έχουμε τελικά υποεκτίμηση του ν. Χαρακτηριστικό είναι επίσης ότι οι περιοχές κλιμάκωσης για m=1 και m=2 βρίσκονται σε αντίστοιχες περιοχές ενώ για τιμές του m μεγαλύτερες του 2 βρίσκονται σε άλλη περιοχή τιμών του r (βλ. Σχ.18β). Από τα παραπάνω συμπεραίνουμε ότι οι εκτιμήσεις που έχουμε για $m \geq 2$ δεν είναι αξιόπιστες και μια πιθανή τιμή για την διάσταση εμβύθισης θα ήταν το m=2. Στον πίνακα 5 παρατίθενται και οι εκτιμώμενες τιμές για την διάσταση συσχέτισης που έδωσε το πρόγραμμα.

Εκτιμώμενες τιμές της διάστασης συσχέτισης ν										
m	1	2	3	4	5	6	7	8	9	10
ν	0.888	1.66	0.516	0.508	0.526	0.532	0.525	0.517	0.501	0.539

Table 5: Εκτίμηση διάστασης συσχέτισης χρονοσειράς Α

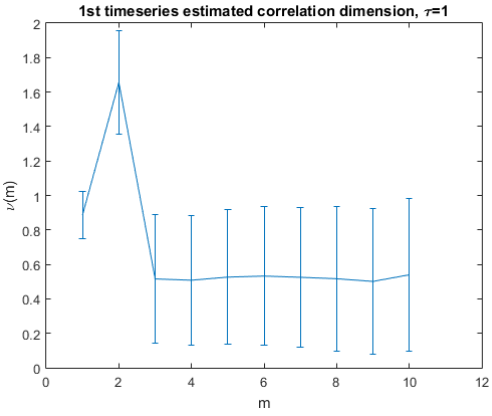


(a) logC(r)

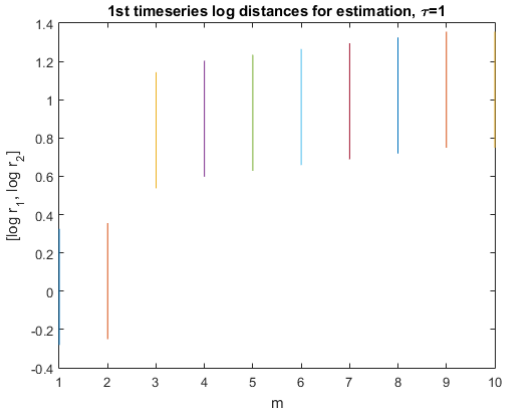


(b) Κλίση του logC(r)

Figure 18: Διαγράμματα για την διάσταση συσχέτισης της χρονοσειράς A

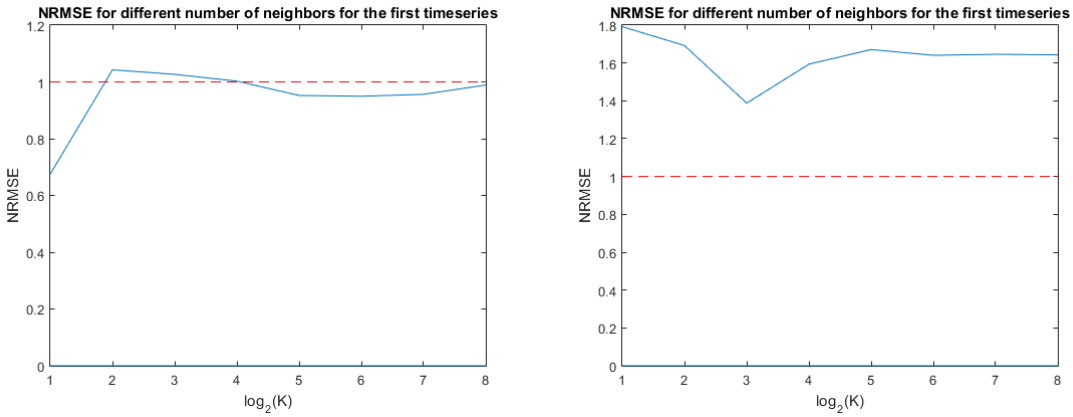


(a) Εκτίμηση της διάστασης συσχέτισης



(b) Περιοχές κλιμάκωσης

Figure 19: Διαγράμματα για την διάσταση συσχέτισης της χρονοσειράς A



(a) m=2 (b) m=5

Figure 20: NRMSE συναρτήσει του πλήθους γειτόνων K για την χρονοσειρά A

δ)Πλήθος γειτόνων K: Αφού δεν μπορέσαμε να εκτιμήσουμε με σιγουριά την διάσταση εμβύθισης m για το μοντέλο που θα χρησιμοποιήσουμε, ευελπιστούμε να το βρούμε σε συνδυασμό με το πλήθος των γειτόνων K για τα οποία θα προκύψει LAP μοντέλο με το μικρότερο NRMSE. Επειδή όλες αυτές τις παραμέτρους τις χρησιμοποιούμε για το μοντέλο που θα έχουμε για την εύρεση των σημείων αλλαγής, μας ενδιαφέρει το μοντέλο μας να δείχνει καλή προσαρμογή και προβλεπτική ικανότητα για συνθήκες ανάλογες με αυτές που θα έχουμε για την εύρεση των σημείων αλλαγής. Οι δυο υποψήφιες τιμές για το m ήταν 2 και 5 οπότε θα δοκιμάσουμε μια φορά για καθεμία από αυτές. Για τις τιμές αυτές του m θα δημιουργήσουμε διάφορα LAP μοντέλα καθένα από τα οποία έχει διαφορετικό αριθμό γειτόνων K. Θα προσαρμόσουμε τα μοντέλα αυτά στα πρώτα 400 σημεία του ελκυστή και θα συγκρίνουμε το NRMSE τους για βήμα πρόβλεψης έως T=5 (ίδιο με αυτό που είχαμε και στο ARMA μοντέλο, για λόγους σύγκρισης). Στο Σχ. 20 φαίνεται η τιμή NRMSE των μοντέλων αυτών για διάφορες τιμές του K.

Η καλύτερη τιμή του NRMSE προκύπτει για m=2,K=2 και είναι περίπου 0.7 (καλύτερη και από αυτήν που είχαμε με την προσαρμογή του ARMA μοντέλου). Παρότι, λοιπόν, το ποσοστό ψευδών γειτόνων για m=2 είναι αρκετά μεγάλο βλέπουμε ότι το μοντέλο αυτό μπορεί να κάνει αρκετά ικανοποιητικές προβλέψεις και μιας και αυτό είναι που μας ενδιαφέρει επλέγουμε να κρατήσουμε αυτό.

4.2.2 Χρονοσειρά B

α)Υστέρηση τ: Κατά αντιστοιχία με την χρονοσειρά A, θεωρήσαμε και εδώ ότι η χρονοσειρά προέρχεται από διακριτό σύστημα οπότε επιλέχθηκε και πάλι το τ=1.

β)Διάσταση εμβύθισης m: Στο Σχ.21 φαίνεται το ποσοστό ψευδών γειτόνων που έχουμε για διάφορες τιμές της διάστασης εμβύθισης m. Παρατηρούμε ότι για m¿4 δεν έχουμε αποτελέσματα αφού οι παρατηρήσεις δεν είναι αρκετές για να έχουμε αξιόπιστα στατιστικά σε τόσο μεγάλες διαστάσεις. Για τις τιμές που έχουμε αποτέλεσμα βλέπουμε ότι αυτές φθίνουν καθώς

αυξάνεται το m αλλά όχι σε σημείο που το ποσοστό να φτάσει να πέσει κάτω από το επιθυμητό. Βάσει των τιμών που έχουμε πάντως, το βέλτιστο m που θα μπορύσαμε να επιλέξουμε θα ήταν το m=4. Πρωτού λάβουμε την τελική μας απόφαση θα εξετάσουμε και την διάσταση συσχέτισης ν.

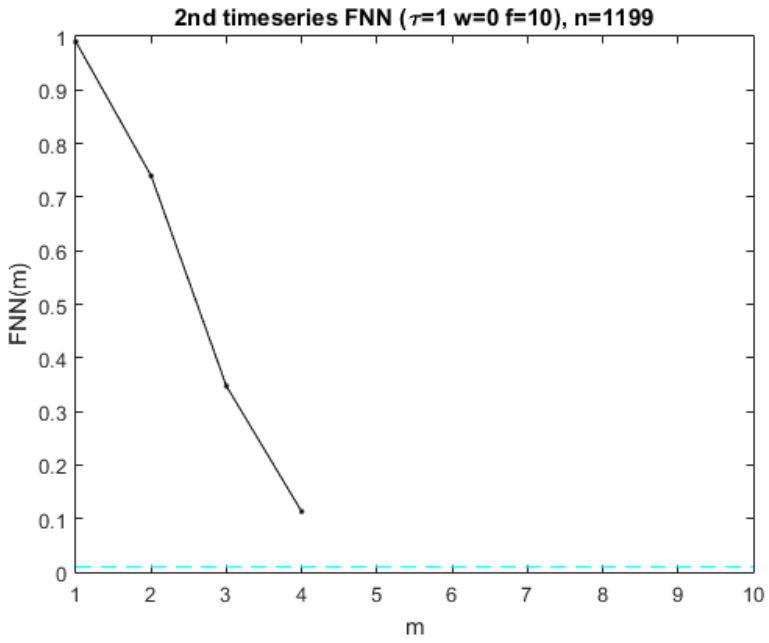


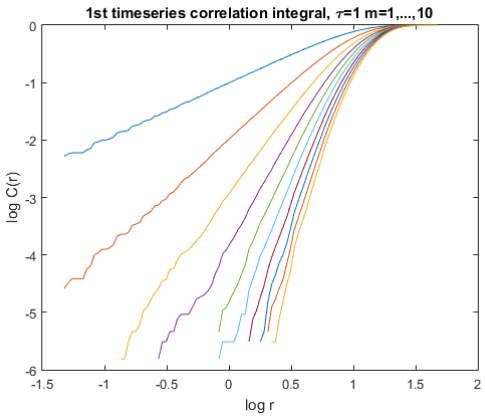
Figure 21: Διάγραμμα ποσοστού ψευδών γειτόνων για την χρονοσειρά Β

γ)Διάσταση συσχέτισης ν: Στο Σχ. 17 και 18 βλέπουμε τα διγράμματα που προέκυψαν απο την ανάλυση για την εύρεση της διάστασης συσχέτισης του ελκυστή. Για αυτήν την χρονοσειρά, βλέπουμε στα διαγράμματα του Σχ.22 ότι η επίδραση του θορύβου είναι ακόμα πιο έντονη με αποτέλεσμα να έχουμε ακόμα μεγαλύτερη μείωση της περιοχής κλιμάκωσης από τα αριστερά λόγω αλλοίωσης της στατιστικής. Ακόμα, παρατηρούμε ότι για τιμές $m \leq 4$ ισχύει ότι το m και το ν είναι περίπου ίσα. Από αυτό καταλαβαίνουμε ότι στις περιπτώσεις αυτές το m δεν είναι αρκετά μεγάλο για να περιγραφεί επαρκώς ο ελκυστής σε τόσες διαστάσεις. Αξιοσημείωτο είναι επίσης ότι για $m \geq 7$ φαίνεται να έχουμε πάλι υποεκτίμηση του ν για τους ίδιους λόγους που αναλύθηκαν και πριν για την χρονοσειρά Α. Μιας και το m δεν φαίνεται να ξεπερνάει την τιμή 5 μια πιθανή επιλογή για την διάσταση εμβύθισης θα ήταν m=5. Για την τιμή αυτή δεν έχουμε το ποσοστό ψευδών γειτόνων όμως για m=4 ήταν ήδη σχετικά μικρό οπότε αναμένουμε για μεγαλύτερες τιμές του m να είναι ακόμα μικρότερο. Επομένως, η τελική μας επιλογή θα είναι m=5.

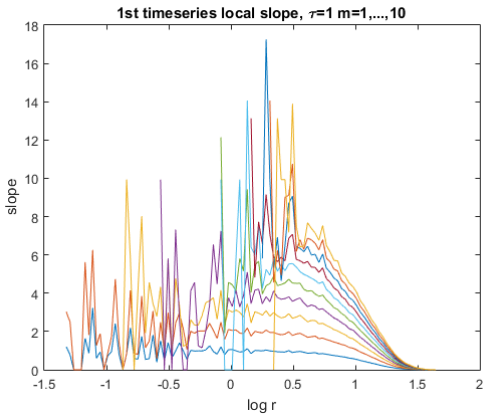
Εκτιμώμενες τιμές της διάστασης συσχέτισης ν										
m	1	2	3	4	5	6	7	8	9	10
ν	0.973	1.94	2.9	3.6	4.15	4.84	0.84	0.943	1.04	1.14

Table 6: Εκτίμηση διάστασης συσχέτισης χρονοσειράς Β

δ)Πλήθος γειτόνων Κ: Για m=5 θα βρούμε για ποιές τιμές του Κ το προσαρμοσμένο

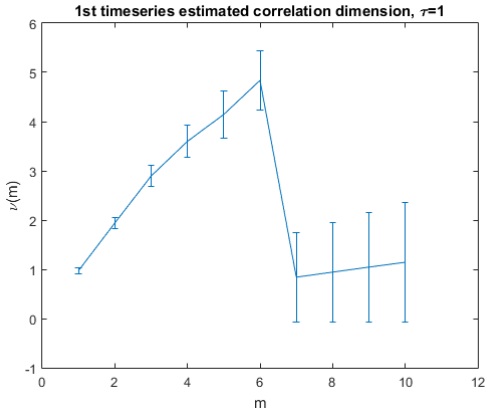


(a) logC(r)

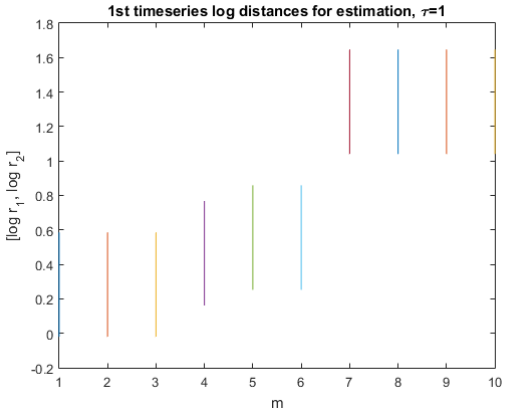


(b) Κλίση του logC(r)

Figure 22: Διαγράμματα για την διάσταση συσχέτισης της χρονοσειράς B



(a) Εκτίμηση της διάστασης συσχέτισης



(b) Περιοχές κλιμάκωσης

Figure 23: Διαγράμματα για την διάσταση συσχέτισης της χρονοσειράς B

κοντέλο έχει το μικρότερο NRMSE. Για λόγους συνέπειας θα χρησιμοποιήσουμε και εδώ τις ίδιες παραμέτρους που χρησιμοποιήσαμε και για την χρονοσειρά Α, δηλαδή βήμα πρόβλεψης έως και T=5 και σύνολο εκμάθησης τα 400 πρώτα σημεία του ελκυστή. Στο Σχ.24 φαίνονται τα σχετικά αποτελέσματα.

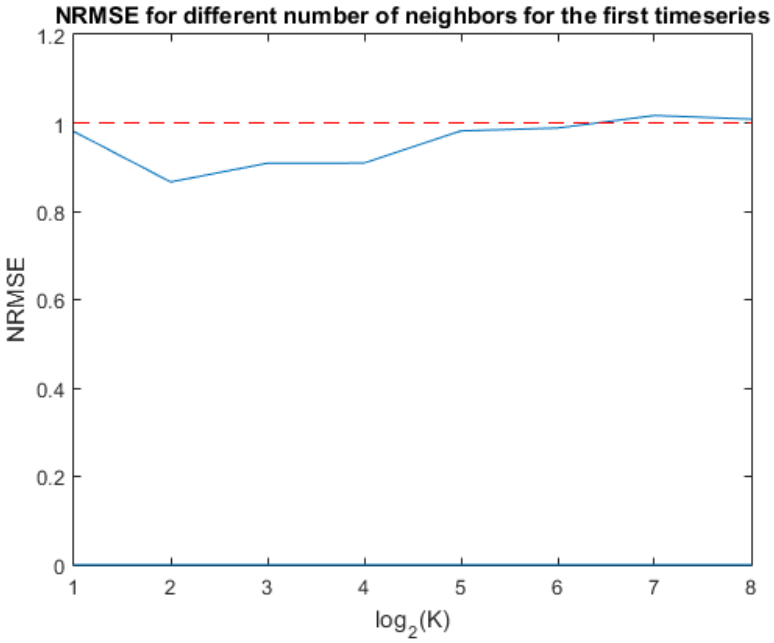


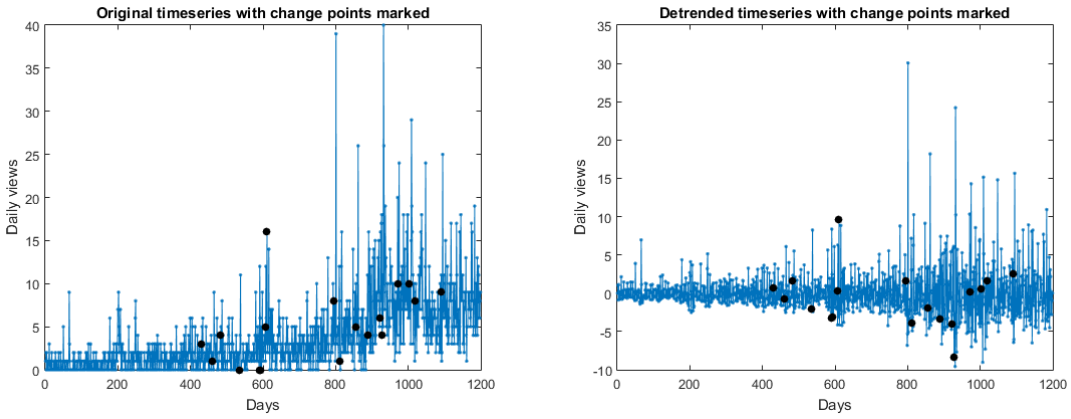
Figure 24: NRMSE συναρτήσει του πλήθους γειτόνων για την χρονοσειρά B με m=5

Παρατηρούμε ότι καλύτερα αποτελέσματα έχουμε για m=5,K=4 με το NRMSE να είναι περίπου 0.85 οπότε και το μοντέλο LAP που θα χρησιμοποιήσουμε θα έχει αυτές τις παραμέτρους. Αξίζει να σημειώσουμε, επίσης, ότι η προβλεπτική ικανότητα του μη γραμμικού μοντέλου για την χρονοσειρά B φαίνεται να μην είναι τόσο καλή όσο αυτής του ARMA μοντέλου που προσαρμόσαμε παραπάνω.

4.3 Εντοπισμός σημείων αλλαγής

4.3.1 Χρονοσειρά Α

Αφού έχουμε προσδιορίσει πλέον τις διάφορες παραμέτρους του μη-γραμμικού LAP μοντέλου θα το χρησιμοποιήσουμε για να κάνουμε με αυτό τις προβλέψεις μας και να βρούμε τα σημεία αλλαγής. Για λόγους συνέπειας, χρησιμοποιήθηκε και εδώ η ίδια ακριβώς διαδικασία που χρησιμοποιήσαμε και με την πρόβλεψη του γραμμικού ARMA μοντέλου. Πιο συγκεκριμένα, η χρονοσειρά προσαρμόζεται κάθε φορά στα τελευταία 400 "γνωστά" σημεία του ελκυστή ενώ το στατιστικό που υπολογίζεται είναι η μέση τιμή των απόλυτων τιμών σφαλμάτων πρόβλεψης για έως και 5 χρονικές στιγμές μπροστά. Όταν η τιμή του στατιστικού ξεπερνάει το a=2s (όπου s η τυπική απόκλιση των παρατηρήσεων του συνόλου εκμάθησης) τότε θεωρούμε ότι έχουμε βρεί ένα σημείο αλλαγής.



(a) Αρχική χρονοσειρά (b) Στάσιμη χρονοσειρά

Figure 25: Σημεία αλλαγής στην χρονοσειρά A

Στο Σχ.25 έχουν σημειωθεί με μαύρες κουκίδες τα σημεία αλλαγής πάνω στην αρχική χρονοσειρά και στην στάσιμη που προέκυψε από αυτήν (συνολικά 18). Από τις τιμές που προέκυψαν παρατηρούμε ότι το μοντέλο αυτό βρίσκει τα ίδια (ή αρκετά κοντινά) σημεία αλλαγής με αυτά που βρίσκει το ARMA μοντέλο καθώς και κάποια επιπλέον. Παρατηρούμε και πάλι ότι κάποια από αυτά τα σημεία βρίσκονται πάνω σε έντονα spikes ή αμέσως πριν από αυτά ενώ κάποια άλλα φαίνεται να βρίσκονται σε φαινομενικά "άσχετες" θέσεις. Έτσι, λοιπόν, είναι αρκετά πιθανό το μοντέλο, λόγω της έλλειψης μεγάλης προβλεπτικής ικανότητας, να θεωρεί ορισμένα σημεία ως σημεία αλλαγής ενώ στην πραγματικότητα το σφάλμα υπεσέρχεται από την αδυναμία πρόβλεψης του μοντέλου και όχι από κάποια μη αναμενόμενη αλλαγή στις τιμές της χρονοσειράς. Συμπεραίνουμε, λοιπόν, ότι δεν μπορούμε να εμπιστευτούμε απόλυτα τα αποτελέσματα που μας δίνει αυτή η μέθοδος ή να βγάλουμε κάποια ασφαλή συμπεράσματα από αυτά. Τέλος, στο Σχ.26 φαίνονται και οι προβλεπόμενες τιμές που επέστρεψε το μοντέλο μαζί με τις αντίστοιχες πραγματικές, όπου φαίνεται ξεκάθαρα η όχι και τόσο καλή προβλεπτική ικανότητα του μοντέλου.

4.3.2 Χρονοσειρά B

Όσον αφορά την χρονοσειρά B ,αφού επιλέξαμε τις παραμέτρους του LAP μοντέλου που θα προσαρμόσουμε, θα ακολουθήσουμε και πάλι τα ίδια βήματα όπως αυτά έχουν ήδη αναλυθεί παραπάνω. Όπως κάναμε και πριν για την χρονοσειρά B, μια χρονική στιγμή θα θεωρείται ως σημείο αλλαγής αν η τιμή του στατιστικού ξεπερνάει το α=1.5s (όπου s η τυπική απόκλιση των παρατηρήσεων του δείγματος εκμάθησης). Οι λόγοι που επιλέξαμε αυτή την τιμή για το α αναλύονται στην παράγραφο 3.2.2.

Στο Σχ.27 έχουν σημειωθεί με μαύρες κουκίδες τα σημεία αλλαγής πάνω στην αρχική χρονοσειρά και στην στάσιμη που προέκυψε από αυτήν (συνολικά 21). Παρατηρούμε και πάλι ότι το μοντέλο μας εντοπίζει τα ίδια (ή αρκετά κοντινά) σημεία αλλαγής με αυτά που βρίσκει το ARMA μοντέλο καθώς και κάποια επιπλέον. Κάποια από αυτά τα σημεία βρίσκονται πάνω σε έντονα spikes ή αμέσως πριν από αυτά ενώ κάποια άλλα φαίνεται να βρίσκονται σε φαινομενικά

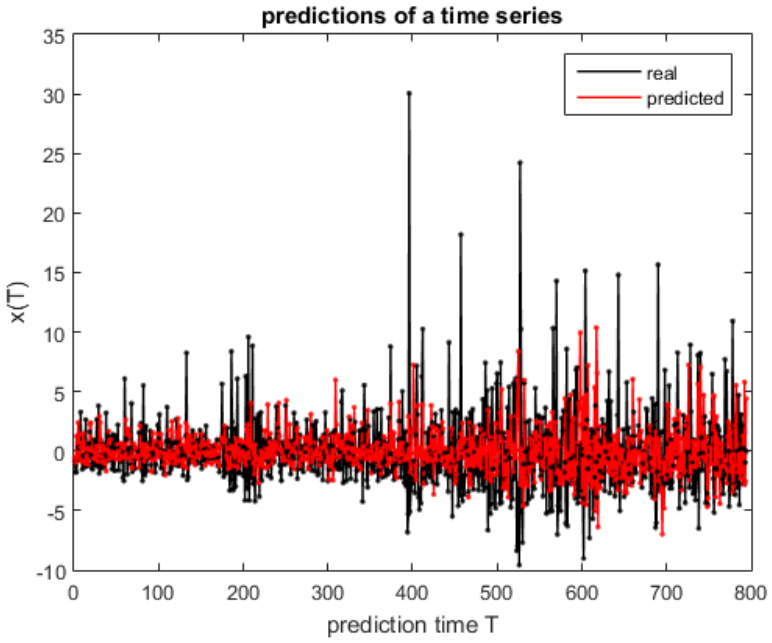


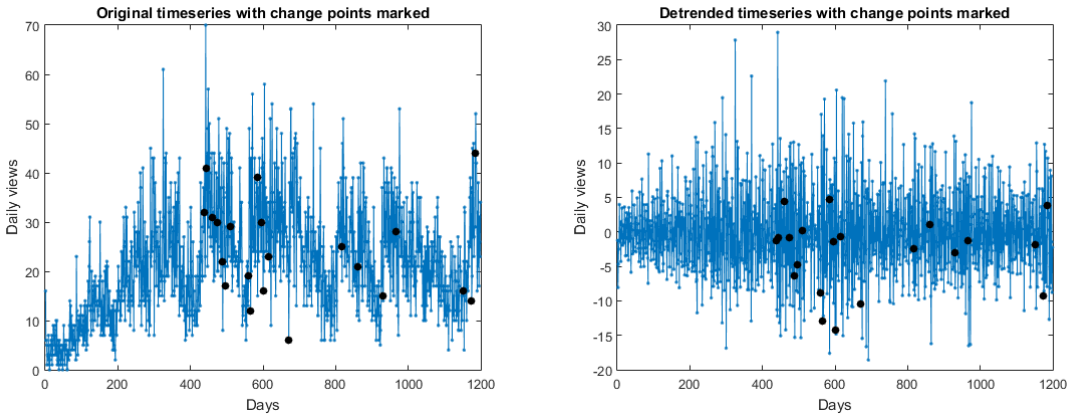
Figure 26: Προβλεπόμενες τιμές του LAP(2) για την χρονοσειρά A

"άσχετες" θέσεις. Γενικά, η συμπεριφορά που έχουμε με το μοντέλο αυτό για αυτήν την χρονοσειρά είναι όμοια με αυτήν που είχε το LAP μοντέλο που προσαρμόσαμε στην χρονοσειρά A. Επομένως, όσες παρατηρήσεις αναφέρθηκαν στην προηγούμενη παράγραφο ισχύουν και εδώ. Τέλος, στο Σχ.28 φαίνονται οι προβλεπόμενες τιμές που επέστρεψε το μοντέλο μαζί με τις αντίστοιχες πραγματικές, όπου φαίνεται ξεκάθαρα η όχι και τόσο καλή προβλεπτική ικανότητα του μοντέλου, ενώ χαρακτηριστικό αποτελεί και το γεγονός ότι το μοντέλο αδυνατεί να προβλέψει οποιοδήποτε spike.

5 Συμπεράσματα

Παρατηρώντας συνολικά τα αποτελέσματα που έχουν προκύψει για τα σημεία και λαμβάνοντας υπόψιν όλες τις παρατηρήσεις που αναφέρθηκαν ως τώρα, καταλαβαίνουμε ότι η μέθοδος αυτή δεν μπορεί να θεωρηθεί αξιόπιστη ούτε και τα αποτελέσματά της άμεσου πρακτικού ενδιαφέροντος. Η αδυναμία ικανοποιητικής πρόβλεψης για τα διάφορα μοντέλα που προσαρμόσαμε έχει ως αποτέλεσμα να υπεσέρχεται μεγάλο σφάλμα στις προβλέψεις μας με αποτέλεσμα να υποβαθμίζεται η αξιοπιστία του στατιστικού που χρησιμοποιήσαμε. Λόγω της αδυναμίας πρόβλεψης, προκύπτουν μεγάλα σφάλματα πρόβλεψης χωρίς όμως αυτά να προέρχονται από σημεία αλλαγής (πχ spikes στις τιμές της χρονοσειράς).

Μια πρώτη ιδέα θα ήταν να αυξάναμε την τιμή του ορίου α. Όμως με μεγάλη αύξηση του θα διατρέχαμε τον κίνδυνο να μην εντοπίσουμε σημεία τα οποία είναι όντως σημεία αλλαγής. Γενικά, η χρήση κάποιου άλλου στατιστικού ίσως αποδεικνυόταν πιο χρήσιμη δεδομένων αυτών των προσαρμοσμένων μοντέλων, παρόλα αυτά, όμως, δεν θεωρούμε το σκεπτικό πισω από την



(a) Αρχική χρονοσειρά (b) Στάσιμη χρονοσειρά

Figure 27: Σημεία αλλαγής στην χρονοσειρά B

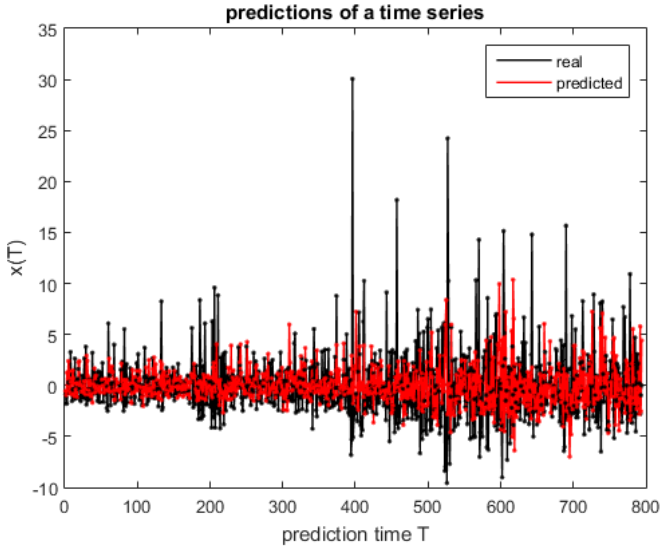


Figure 28: Προβλεπόμενες τιμές του LAP(4) για την χρονοσειρά B

εφαρμογή του συγκεκριμένου στατιστικού λανθασμένο ή κακό. Για να μπορέσει όμως να δώσει ορθά αποτελέσματα θα ήταν απαραίτητη η προσαρμογή ενός μοντέλου με μεγαλύτερη προβλεπτική ισχύ. Για παράδειγμα, η χρήση κάποιου τοπικά γραμμικού μοντέλου πρόβλεψης (LLP) ή ίσως και κάποιου νευρωνικού δικτύου (NN) τα οποία θα μπορούσαν να πραγματοποιήσουν καλύτερες προβλέψεις θα μας έδιναν και πιο αξιόπιστα αποτελέσματα για τα σημεία αλλαγής. Το παραπάνω, σε συνδυασμό με κατάλληλη επιλογή των διαφόρων παραμέτρων όπως το α, το T και το μέγεθος του συνόλου εκμάθησης είναι μια κατεύθυνση που θα άξιζε περαιτέρω μελλοντικής διερεύνησης για την αποδοτικότερη εύρεση σημείων αλλαγής σε τέτοιες χρονοσειρές.