Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχ. κ Μηχ. Υπολογιστών
Τομέας Ηλεκτρικών Βιομηχανικών Διατάξεων & Συστημάτων Αποφάσεων
Μονάδα Προβλέψεων & Στρατηγικής

Ολοκληρωμένα Αυτοπαλινδρομικά Μοντέλα Κινητού Μέσου Όρου (ARIMA)

Σημειώσεις για το μάθημα 'Τεχνικές Προβλέψεων'

ΣΗΜΜΥ ΕΜΠ – 8ο εξάμηνο

Ευάγγελος Σπηλιώτης

Διδάκτορας ΕΜΠ

Περιεχόμενα

1. Εισαγωγή

Τα ολοκληρωμένα αυτοπαλινδρομικά μοντέλα κινητών μέσων όρων **ARIMA** (**A**uto **R**egressive **I**ntegrated **M**oving **A**verage) είναι **στοχαστικά** μοντέλα τα οποία μας βοηθάνε να περιγράψουμε το μηχανισμό εξέλιξης ενός μεγέθους ενδιαφέροντος και ως εκ τούτου να προβλέψουμε την τιμή του στο μέλλον.

Όπως και άλλα **στατιστικά** μοντέλα πρόβλεψης χρονοσειρών, π.χ. αυτά της εκθετικής εξομάλυνσης, τα μοντέλα ARIMA παράγουν προβλέψεις βασιζόμενα αποκλειστικά στις ιστορικές παρατηρήσεις του υπό εξέταση μεγέθους. Αυτό αποτελεί από μόνο του σημαντικό πλεονέκτημα, ειδικά σε περιπτώσεις που δεν γνωρίζουμε τις εξωτερικές μεταβλητές που το επηρεάζουν ή δεν μπορούμε να τις αξιοποιήσουμε άμεσα (βλ. ντετερμινιστικά μοντέλα).

Η ειδοποιός διαφορά ωστόσο μεταξύ των μοντέλων ARIMA και των υπολοίπων μοντέλων χρονοσειρών, είναι ο ιδιαίτερος τρόπος με τον οποίο τα πρώτα διαχειρίζονται τα ιστορικά δεδομένα. Πιο συγκεκριμένα, τα μοντέλα ARIMA δεν υποθέτουν εξαρχής το μηχανισμό με τον οποίο εξελίσσεται το μέγεθος ενδιαφέροντος, αλλά αντίθετα επιλέγουν μέσα από μία ευρεία γκάμα μηχανισμών εκείνον που έχει τη μεγαλύτερη πιθανότητα να αποκαλύπτει τη σχέση που συνδέει την κάθε παρατήρηση με τις προηγούμενές της. Η εν λόγω **ευελιξία** είναι κρίσιμη για την επίδοση των μοντέλων ARIMA καθώς επιτρέπει τη μοντελοποίηση πολύπλοκων μηχανισμών που λαμβάνουν υπόψη τους ποικίλα χαρακτηριστικά χρονοσειρών.

Στην πραγματικότητα, το κάθε μοντέλο ARIMA εκφράζει ένα διαφορετικό μηχανισμό εξέλιξης και η επιλογή του καταλληλότερου για την προέκταση μιας χρονοσειράς γίνεται εξετάζοντας παράγοντες όπως η σχέση μεταξύ *k* διαδοχικών παρατηρήσεων (αυτοσυσχέτιση), η ύπαρξη τάσης, η ύπαρξη εποχιακότητας και το σφάλμα πρόβλεψης. Το γεγονός πως τα μοντέλα ARIMA μπορούν να αναπαράγουν ικανοποιητικά με αυτόματο τρόπο οποιοδήποτε μοτίβο ενδέχεται να κρύβεται στα δεδομένα χωρίς να απαιτείται κάποια ιδιαίτερη προ-επεξεργασία (π.χ. αποεποχικοποίηση ή αφαίρεση τάσης), διάδωσε σημαντικά τη χρήση τους σε εφαρμογές πρόβλεψης, καθιστώντας τα ολοκληρωμένες λύσεις. Τα μοντέλα ARIMA μελετήθηκαν εκτεταμένα από τους **Box και Jenkins** τη δεκαετία του ’70 και συχνά αναφέρονται στη βιβλιογραφία με το αντίστοιχο όνομα.

Στην γενική τους μορφή τα μοντέλα ARIMA αποτελούνται από τον τυχαίο παράγοντα (παράγοντας *MA*), τις τιμές που εμφανίστηκαν σε προηγούμενες περιόδους (παράγοντας *AR* και *I*), και άλλες στοχαστικές μεταβλητές. Πιο συγκεκριμένα, κάθε

μοντέλο ARIMA μπορεί να εκφραστεί ως **γραμμικός συνδυασμός** των παραπάνω παραγόντων και στόχος μας είναι να ανακαλύψουμε εκείνον που παράγει τις καλύτερες προβλέψεις. Έτσι, αν το μοντέλο περιλαμβάνει αποκλειστικά παράγοντες αυτοπαλινδρόμησης αναφέρεται ως *AR(p)*, αν περιλαμβάνει αποκλειστικά παράγοντες κινητών μέσω όρων ως *MA(q)*, και αν περιλαμβάνει και τους δύο ως *ARMA(p,q)*, όπου τα *p* και *q* δηλώνουν την τάξη του μοντέλου ανά παράγοντα. Ο παράγοντας *I(d)* αναφέρεται στη διαφόριση της χρονοσειράς πριν την εφαρμογή ενός μοντέλου *ARMA(p,q)* και έχει ως στόχο την αφαίρεση της τάσης από τα δεδομένα.

Σε περίπτωση που η χρονοσειρά είναι εποχιακή, τα μοντέλα ARIMA μπορούν να επεκταθούν κατάλληλα προκειμένου να προσομοιώσουν και την εποχιακή συμπεριφορά των δεδομένων. Σε αυτή την περίπτωση η έκφρασή τους έχει τη μορφή *ARIMA(p,d,q)(P,D,Q)*, όπου τα *P, D* και *Q* αναφέρονται αντίστοιχα στην τάξη των εποχιακών παραγόντων ARIMA.

Στην πράξη βέβαια δεν μπορούμε να είμαστε ποτέ σίγουροι για το ποιος είναι ο βέλτιστος συνδυασμός παραγόντων ή καλύτερα για το αν καταφέραμε να αποκαλύψουμε πλήρως το μοτίβο της χρονοσειράς. Μπορούμε ωστόσο να δίνουμε μία ικανοποιητική προσέγγιση στο εν λόγω ερώτημα εντοπίζοντας ένα μοντέλο που πληροί συγκεκριμένες προϋποθέσεις και παραμετροποιώντας το κατάλληλα.

Η εφαρμογή των μοντέλων ARMA προϋποθέτει επίσης ότι πληρούνται ορισμένες απαιτήσεις. Αρχικά, η χρονοσειρά μήκους *n* πρέπει να είναι **διακριτή**, δηλαδή οι παρατηρήσεις της Y_t να αναφέρονται σε ισαπέχουσες χρονικές στιγμές $Y_t, Y_{t+\tau}, \dots Y_{t+k\tau}$, όπου *τ* ακέραιος μεγαλύτερος του μηδέν. Η παραπάνω απαίτηση είναι απαραίτητη καθώς, δεδομένου ότι τα μοντέλα συσχετίζουν χρονικά τις παρατηρήσεις της χρονοσειράς, η μη χρονική συνέχειά τους αναιρεί οποιαδήποτε υπόθεση έχει γίνει για τις σχέσεις διασύνδεσής τους.

Εκτός αυτού, η χρονοσειρά οφείλει να είναι **στάσιμη**. Αυτό σημαίνει πως η μέση τιμή (μ), η διακύμανση (σ^2) και η συνάρτηση αυτοσυσχέτισης (*ACF*) πρέπει να διατηρούνται σχετικά σταθερές στο πέρασμα του χρόνου. Αν ισχύει η υπόθεση της στασιμότητας, τα χαρακτηριστικά της χρονοσειράς δεν εξαρτώνται από τη χρονική στιγμή στην οποία αυτή μελετάται (βλέπε λευκός θόρυβος) και κατά συνέπεια οποιοδήποτε τυχαίο δείγμα της $Y_{t1}, Y_{t2} \dots Y_{tn}$ ταυτίζεται εν γένει με οποιοδήποτε άλλο επιλεγεί $Y_{t1+\tau}, Y_{t2+\tau} \dots Y_{tn+\tau}$, όπου *τ* ακέραιος μεγαλύτερος του μηδέν. Έτσι, η χρονοσειρά αποδεσμεύεται από την έννοια του χρόνου και μπορεί να μελετηθεί στοχαστικά. Σε περίπτωση που η χρονοσειρά δεν είναι στάσιμη, κάτι τέτοιο μπορεί να επιτευχθεί με χρήση μετασχηματισμών ή/και

διαφόρισης, με εφαρμογή δηλαδή κάποιου ολοκληρωμένου εποχιακού ή μη εποχιακού μοντέλου ARIMA αντί για ARMA.

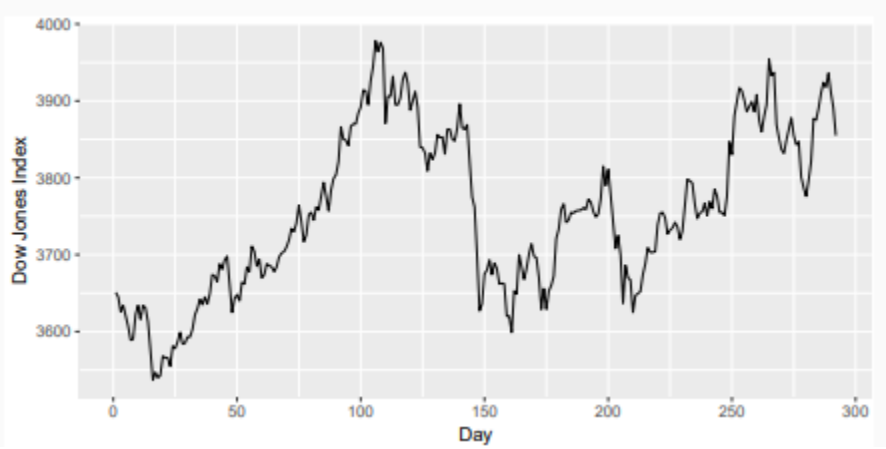
Τέλος, η εφαρμογή των μοντέλων ARIMA προϋποθέτει την παραγωγή **βραχυπρόθεσμων προβλέψεων**. Όπως αναφέρθηκε νωρίτερα, τα μοντέλα ARMA δεν είναι τίποτε άλλο από γραμμικοί συνδυασμοί ιστορικών παρατηρήσεων και στοχαστικών παραγόντων. Αυτό σημαίνει πως για να προβλέψουμε την τιμή Y_{n+1} απαιτείται γνώση των τιμών $Y_n, Y_{n-1} \ldots Y_{n-k}$. Αντίστοιχα, για την πρόβλεψη της τιμής Y_{n+2} απαιτείται γνώση των τιμών $Y_{n+1}, Y_n \ldots Y_{n-k-1}$, εκ των οποίων η Y_{n+1} δεν αποτελεί όμως δεδομένο αλλά μία πρόβλεψη που υπολογίστηκε νωρίτερα από το μοντέλο και που φυσικά εμπεριέχει κάποιο σχετικό σφάλμα. Αν επεκτείνουμε την εν λόγω διαδικασία στο μέλλον, γίνεται κατανοητό πως για μεγάλους ορίζοντες πρόβλεψης η τιμή των προβλέψεων θα βασίζεται αποκλειστικά σε προβλέψεις που πραγματοποιήθηκαν νωρίτερα. Έτσι, η αξιοπιστία και η ακρίβεια πρόβλεψής ενός μοντέλου ARIMA αναμένεται να μειώνεται σημαντικά καθώς αυξάνει ο ορίζοντας πρόβλεψης.

Στη συνέχεια παρουσιάζονται βασικά ζητήματα που αντιμετωπίζει κανείς κατά την παραγωγή προβλέψεων μέσω μοντέλων ARIMA, όπως είναι η επεξεργασία των δεδομένων, η αναγνώριση του καταλληλότερου μοντέλου, η εκτίμησή του και ο διαγνωστικός έλεγχος.

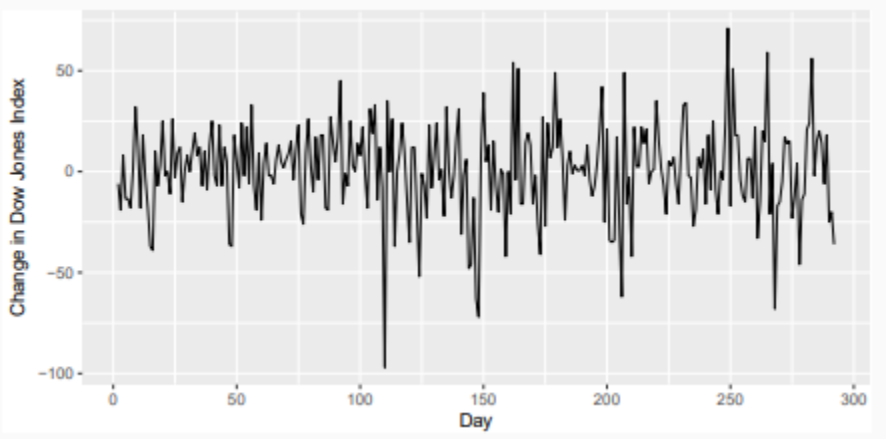
2. Επεξεργασία δεδομένων: Εξομάλυνση, Διαφόριση και Στασιμότητα

Όπως αναφέρθηκε, απαραίτητη προϋπόθεση για να εφαρμοστεί ένα μοντέλο ARMA είναι η χρονοσειρά που μελετάται να είναι στάσιμη, δηλαδή η μέση τιμή, η διακύμανση και η συνάρτηση αυτοσυσχέτισής της να είναι σταθερές στην πάροδο του χρόνου. Πρακτικά αυτό σημαίνει πως κάθε στάσιμη χρονοσειρά είναι σταθερού επιπέδου (οριζόντια) και μακροπρόθεσμα μη προβλέψιμη δεδομένης της έλλειψης προτύπων.

Η απαίτηση της στασιμότητας σπανίως ικανοποιείται εξαρχής καθώς οι περισσότερες χρονοσειρές χαρακτηρίζονται από τάση, εποχιακότητα και ασυνέχειες (π.χ. ειδικά γεγονότα). Μπορεί ωστόσο να ικανοποιηθεί σχετικά εύκολα μέσα από κάποιες τεχνικές επεξεργασίας δεδομένων.



Μία μη στάσιμη χρονοσειρά.



Μία στάσιμη χρονοσειρά.

2.1 Μετασχηματισμοί

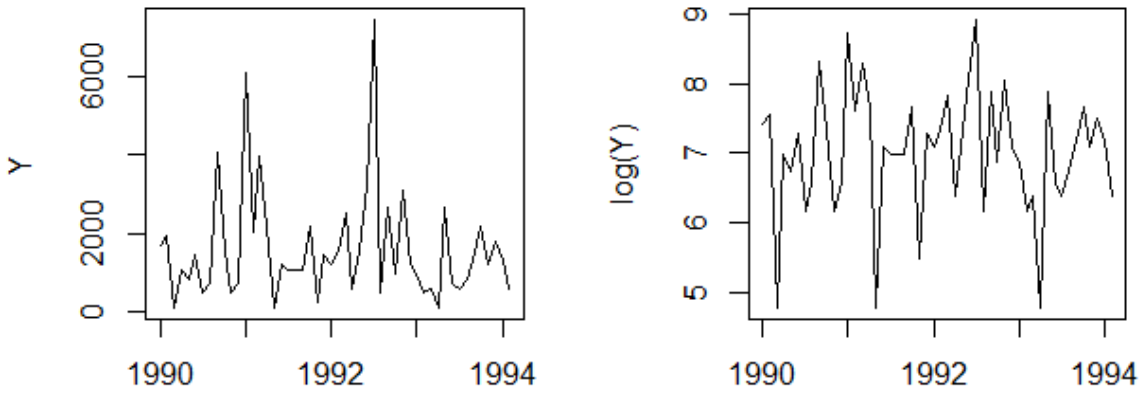
Ένα πρώτο μέτρο που προτείνεται για την εξομάλυνση της διακύμανσης της χρονοσειράς, είναι ο μετασχηματισμός των δεδομένων. Η εν λόγω διαδικασία περιορίζει την **τυχαιότητα και τις ακραίες τιμές** (ασυνέχειες) που τυχόν υπάρχουν, οδηγώντας έτσι σε μία νέα χρονοσειρά μικρότερης και σταθερότερης διακύμανσης. Έτσι, το επίπεδο της χρονοσειράς δεν μεταβάλλεται το ίδιο απότομα με πριν και μπορεί να υποτεθεί στασιμότητα.

Η **λογαρίθμηση** της χρονοσειράς αποτελεί την απλούστερη και πλέον διαδεδομένη μορφή μετασχηματισμού καθώς οδηγεί σε ικανοποιητικά αποτελέσματα με αρκετά απλό τρόπο και μικρό υπολογιστικό κόστος. Άλλη επιλογή είναι η χρήση μετασχηματισμών δυνάμεων, όπως π.χ. οι μετασχηματισμοί **Box-Cox**.

$$Y^{(\lambda)} \begin{cases} \frac{Y^{\lambda}-1}{\lambda}, \lambda \neq 0 \\ \ln(Y), \lambda = 0 \end{cases}$$

Απαραίτητη προϋπόθεση και στις δύο περιπτώσεις είναι φυσικά η χρονοσειρά να αποτελείται από γνησίως θετικές παρατηρήσεις.

Σημειώνεται πως δεν υπάρχει σαφές κριτήριο για το πότε οφείλει κανείς να εφαρμόζει μετασχηματισμούς. Ωστόσο, αν η χρονοσειρά εμφανίζει σαφείς διακυμάνσεις ή το εφαρμοζόμενο μοντέλο ARIMA παρουσιάζει συστηματικά μεγάλα σφάλματα ακρίβειας, καλό θα ήταν να δοκιμαστεί η υιοθέτησή τους. Τονίζεται επίσης πως, καθώς η χρήση μετασχηματισμών επηρεάζει την κλίμακα των δεδομένων, μετά την ολοκλήρωση της διαδικασίας πρόβλεψης απαιτείται η εφαρμογή του αντίστροφου μετασχηματισμού ούτως ώστε οι προβλέψεις να συμβαδίζουν με την αρχική κλίμακα της χρονοσειράς.



Χρονοσειρά με εμφανείς ασυνέχειες και αυξημένη τυχαιότητα πριν (αριστερά) και μετά (δεξιά) την εφαρμογή μετασχηματισμού λογαρίθμισης. Η νέα χρονοσειρά που προκύπτει είναι εμφανώς πιο στάσιμη.

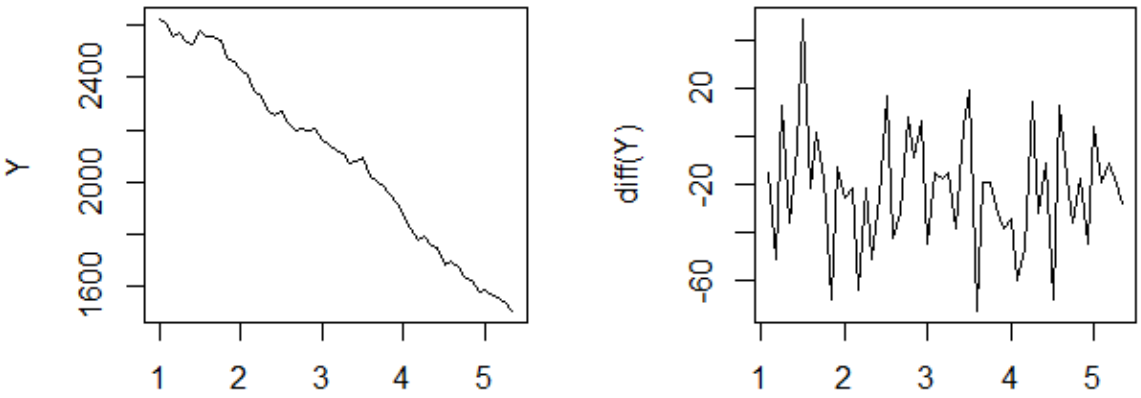
2.2 Διαφόριση

Οι μετασχηματισμοί περιορίζουν τις τυχαίες διακυμάνσεις αλλά αφήνουν ανεπηρέαστες τις συστηματικές διακυμάνσεις, εκείνες δηλαδή που οφείλονται στην ύπαρξη **τάσης** και **εποχιακότητας**. Σε τέτοιες περιπτώσεις, ένα μέτρο που προτείνεται είναι η διαφόριση (*differencing*) της χρονοσειράς.

Η διαφόριση, ανάλογα με την μορφή που λαμβάνει, περιορίζει τις διακυμάνσεις επιπέδου αφαιρώντας τάση και εποχιακότητα. Έτσι, παράγεται μία νέα χρονοσειρά σταθερότερου επιπέδου και διακύμανσης. Στην ουσία, κατά τη διαφόριση μίας χρονοσειράς *n* παρατηρήσεων, δημιουργείται μία νέα με στοιχεία της τις διαφορές των παρατηρήσεων της πρώτης.

Ανάλογα την τάξη διαφόρισης έχουμε λοιπόν:

- **1η τάξη:** $Y'_t=Y_t - Y_{t-1}$
- **2η τάξη:** $Y''_t = Y'_t - Y'_{t-1} = Y_t - 2Y_{t-1} + Y_{t-2}$, κ.ο.κ.



Χρονοσειρά με τάση πριν (αριστερά) και μετά (δεξιά) τη διαφόριση πρώτης τάξης. Η νέα χρονοσειρά που προκύπτει είναι εμφανώς πιο στάσιμη.

Προφανώς οι χρονοσειρές που προκύπτουν από τις διαφορίσεις 1ης και 2ης τάξης έχουν *n-1* και *n-2* παρατηρήσεις αντίστοιχα. Η διαφόριση μπορεί να είναι μέχρι και *n-1* τάξης αλλά στην πράξη, όπως θα δούμε και αργότερα, *χρησιμοποιούμε μόνο μέχρι 2ης*. Βάσει της διαφόρισης μπορούμε μάλιστα να ορίσουμε και να διαχωρίσουμε τις διαδικασίες του λευκού θορύβου (*white noise*) και του τυχαίου περιπάτου (*random walk*), οι οποίες αντιπροσωπεύουν τα μοντέλα ARIMA(0,0,0) και ARIMA(0,1,0), χωρίς και με σταθερά. Οι εν λόγω διαδικασίες είναι ιδιαίτερα σημαντικές καθώς συχνά χρησιμοποιούνται ως βάση αξιολόγησης πολυπλοκότερων μοντέλων.

- Λευκός θόρυβος: $Y_t = e_t$
- Τυχαίος Περίπατος: $Y_t = Y_{t-1} + c + e_t$

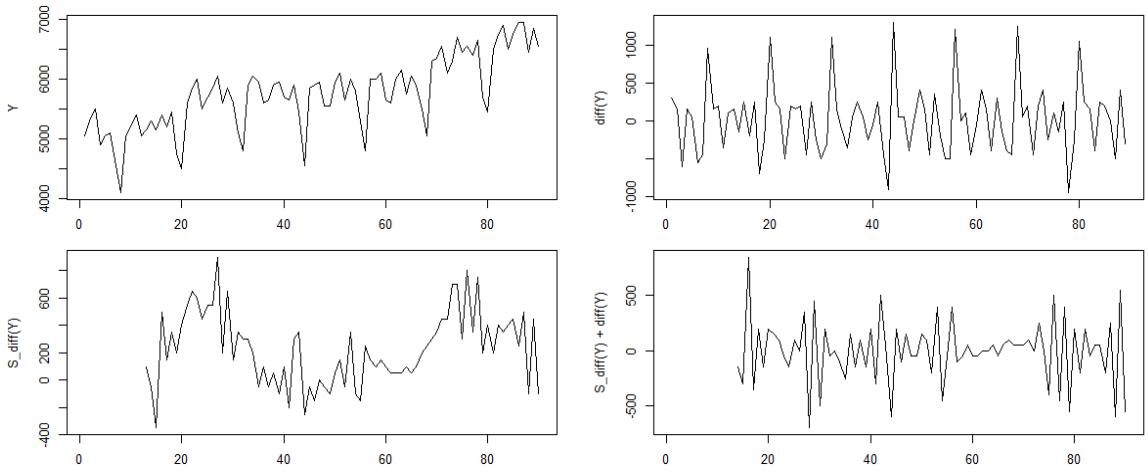
Κατ' αναλογία με την απλή διαφόριση μπορούμε να εφαρμόσουμε και **εποχιακή διαφόριση** σε περιπτώσεις χρονοσειρών έντονης εποχιακότητας. Εδώ η χρονοσειρά που παράγεται είναι αποτέλεσμα της διαφοράς μεταξύ των παρατηρήσεων της αρχικής χρονοσειράς και εκείνων των προηγούμενων αντίστοιχων εποχιακών περιόδων. Ανάλογα τη τάξη διαφόρισης έχουμε λοιπόν:

- **1η τάξη:** $Y'_t=Y_t - Y_{t-m}$
- **2η τάξη**: $Y''_t=Y'_t - Y'_{t-m}= Y_t - 2Y_{t-m} + Y_{t-2m}$

, όπου *m* η περίοδος εποχιακότητας.

Αν ορίσουμε τώρα ως **B** τον **τελεστή ολίσθησης**, ούτως ώστε $BY_t=Y_{t-1}$ και $B(BY_t)= B^2Y_t = Y_{t-2}$, τότε μπορούμε να αναπαραστήσουμε τη διαφόριση *n* τάξης ως $\boldsymbol{(1-B)^nY_t}$ και την εποχιακή διαφόριση περιόδου *m*, τάξης *N*, ως $\boldsymbol{(1-B^m)^NY_t}$.

Αν η εποχιακή διαφόριση δεν έχει αποδώσει επαρκή σταθερότητα λόγω ύπαρξης σχετικής τάσης, τότε ενδέχεται να πρέπει να συνδυαστεί με την απλή διαφόριση. Αυτό γίνεται αναπτύσσοντας τη σχέση $\boldsymbol{(1-B^m)^N(1-B)^ny_t}$. Προσέξτε ότι δεν έχει καμία σημασία η σειρά με την οποία εφαρμόζονται οι διαφορίσεις καθώς το τελικό αποτέλεσμα θα είναι το ίδιο και στις δύο περιπτώσεις. Ωστόσο, όταν εξετάζεται η ταυτόχρονη χρήση εποχιακής και μη διαφόρισης, είθισται να εφαρμόζεται πρώτα η εποχιακή καθώς ενδέχεται η χρονοσειρά που θα προκύψει να είναι επαρκώς στάσιμη και ως εκ τούτου να μην απαιτείται επιπλέον διαφόριση.



Εποχιακή μηνιαία χρονοσειρά (m=12) πριν (πάνω-αριστερά) και μετά τη χρήση διαφόρισης πρώτης τάξης (πάνω-δεξιά), εποχιακής διαφόρισης πρώτης τάξης (κάτω-αριστερά) και συνδυασμό τους (κάτω-δεξιά). Όπως φαίνεται, η νέα χρονοσειρά που προκύπτει από το συνδυασμό είναι εμφανώς πιο στάσιμη καθώς η απλή διαφόριση δεν απαλείφει τις εποχιακές διακυμάνσεις και η εποχιακή τις διακυμάνσεις του επιπέδου.

Ένα πράγμα που χρειάζεται να έχει κανείς κατά νου είναι ότι η διαφόριση δεν είναι πανάκεια και για αυτό το λόγο δεν θα πρέπει να υπερβάλει με την όλη διαδικασία. Για παράδειγμα, η διαφόριση οδηγεί σταδιακά σε μείωση του αριθμού των διαθέσιμων παρατηρήσεων. Αυτό μπορεί να μην αποτελεί ιδιαίτερο πρόβλημα για μεγάλες χρονοσειρές (περισσότερες από 100 παρατηρήσεις) ωστόσο μπορεί να αποβεί προβλεπτικά καταστροφικό σε περιπτώσεις μικρών χρονοσειρών (λιγότερες από 15-20 παρατηρήσεις). Η διαφόριση μειώνει επίσης σημαντικά την αυτοσυσχέτιση των χρονοσειρών αυξάνοντας την τυχαιότητα του μοντέλου. Αυτό το φαινόμενο μπορεί να οδηγήσει αργότερα με τη σειρά του στην ανάγκη εφαρμογής μοντέλων ARMA υψηλής

πολυπλοκότητας, τα οποία είναι ικανά να αντιμετωπίσουν την τυχαιότητα. Το εν λόγω φαινόμενο ονομάζεται **υπερδιαφόριση** (over-differencing).

Στην πράξη λοιπόν δεν πραγματοποιείται διαφόριση για τιμές αυτοσυσχέτισης (βλ. παράγραφο 4) μικρότερες του 0.5 και σε καμία περίπτωση περισσότερες από δύο φορές. Αν υπάρχει έντονη αρνητική αυτοσυσχέτιση (<-0.5), αυτό αποτελεί δείγμα υπερδιαφόρισης. Επίσης, η εμπειρία δείχνει ότι ποτέ δεν χρησιμοποιείται εποχιακή διαφόριση μεγαλύτερη της πρώτης τάξης.

3. Αναγνώριση, εκτίμηση και διάγνωση μοντέλων ARIMA

Η επιλογή του καταλληλότερου μοντέλου ARIMA δεν είναι πάντοτε προφανής. Συχνά, περισσότερα από ένα μοντέλα μπορεί να οδηγούν σε παραπλήσια αποτελέσματα, αφήνοντας την τελική επιλογή στην κρίση μας. Χαρακτηριστικό παράδειγμα αποτελεί η περίπτωση όπου, ενώ κάποιο μοντέλο ARIMA προσαρμόζεται καλά στη δοθείσα χρονοσειρά, μπορεί να απορριφθεί έναντι ενός άλλου μικρότερης πολυπλοκότητας και χειρότερης προσαρμογής.

Η παραπάνω πρακτική είναι αρκετά λογική στη βάση της αν αναλογιστεί κανείς το εξής: Όσο αυξάνει η πολυπλοκότητα ενός μοντέλου, τόσο βελτιώνεται η ικανότητά του να προσαρμόζεται στα δεδομένα. Ωστόσο, η τελική ακρίβεια πρόβλεψης δεν συμβαδίζει πάντοτε με το σφάλμα πρόβλεψης που εκτιμάται βάσει ιστορικών δεδομένων και μάλιστα είναι συνήθως αρκετά χειρότερη όταν το μοντέλο που έχει εκτιμηθεί αποτελεί αποτέλεσμα υπερπροσαρμογής. Έτσι, ο εντοπισμός μοντέλων ικανοποιητικής ακρίβειας και περιορισμένης πολυπλοκότητας αποτελεί βασική αρχή κατά τη διαδικασία επιλογής. Η εν λόγω πρακτική ενισχύεται από το μεγάλο πλήθος παραμέτρων που μπορεί να δεχτεί δυνητικά ως είσοδο ένα μοντέλο ARIMA, παράγοντας που θα οδηγούσε αντίστοιχα σε μακροσκελείς και χρονοβόρους ελέγχους.

Θέλοντας λοιπόν να αυτοματοποιήσουμε τη διαδικασία επιλογής, είθισται να ακολουθούμε την παρακάτω διαδικασία η οποία περιλαμβάνει τρία στάδια: την **Αναγνώριση**, την **Εκτίμηση** και τη **Διάγνωση**.

1. Στο στάδιο της αναγνώρισης επιλέγονται ένα ή περισσότερα μοντέλα ARIMA τα οποία θεωρούμε βάσει κάποιων ενδείξεων, όπως π.χ. οι γραφικές παραστάσεις της αυτοσυσχέτισης και της μερικής αυτοσυσχέτισης, ότι μπορούν να περιγράψουν ικανοποιητικά τη χρονοσειρά. Θα αναφερθούμε στη συνέχεια αναλυτικά στην εν λόγω διαδικασία.

2. Στο στάδιο της εκτίμησης εντοπίζονται για κάθε ένα από τα μοντέλα που αναγνωρίστηκαν οι παράμετροί *p*, *d*, *q*, *P*, *D* και *Q*, και βάσει αυτών παραμετροποιούνται κατάλληλα. Η εν λόγω διαδικασία μπορεί να γίνει με αρκετούς τρόπους, ο πιο διαδεδομένος εκ των οποίων είναι ο υπολογισμός της **προσδοκώμενης πιθανοφάνειας** (Likelihood Estimation). Η προσδοκώμενη πιθανοφάνεια δείχνει επί της ουσίας κατά πόσο το μοντέλο με τις παραμέτρους που επιλέχθηκαν έχει τη δυνατότητα να αναπαράγει τις πραγματικές τιμές της χρονοσειράς. Οι παράμετροι υπολογίζονται λοιπόν με κριτήριο την μεγιστοποίηση της πιθανοφάνειας ή την ελαχιστοποίηση του λογάριθμου αυτής ($-2logL$) ως εξής:

$$L = \prod_{t=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(Y_t - F_t)^2}{2\sigma^2}} \rightarrow$$

$$L = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(Y_t - F_t)^2} \rightarrow$$

$$-2logL = n\log(2\pi) + n\log(\sigma^2) + \frac{\sum_{t=1}^{n} {e_t}^2}{\sigma^2}$$

,όπου *L* η προσδοκώμενη πιθανοφάνεια ταύτισης του μοντέλου με τα αρχικά δεδομένα, F_t η προβλεπόμενη από το μοντέλο τιμή την περίοδο *t*, *n* ο αριθμός των ιστορικών δεδομένων, e_t το σφάλμα πρόβλεψης και σ^2 η διακύμανση των σφαλμάτων του μοντέλου. Σημειώνεται πως *η παραπάνω σχέση ισχύει αποκλειστικά για μοντέλα ARMA*, δηλαδή δεν λαμβάνει υπόψη της τυχόν διαφορίσεις.

Το κριτήριο της πιθανοφάνειας λειτουργεί πρακτικά όπως η μέθοδος ελαχίστων τετραγώνων στην περίπτωση της απλής γραμμικής παλινδρόμησης για την επιλογή των παραμέτρων *a* και *b*, το οποίο και ελαχιστοποιεί το άθροισμα των τετραγώνων των υπολειπόμενων σφαλμάτων μέσω της απαίτησης:

$$\min\left(\sum_{t=1}^{n} et^2\right), \text{όπου } e_t = Y_t - F_t$$

Μάλιστα, στην περίπτωση της γραμμικής παλινδρόμησης το κριτήριο της πιθανοφάνειας δίνει τις ίδιες παραμέτρους με αυτό των ελαχίστων τετραγώνων. Για την επιλογή των παραμέτρων συχνά συνδυάζεται η τεχνική των ελαχίστων τετραγώνων με αυτή της προσδοκώμενης πιθανοφάνειας για ακόμα καλύτερα αποτελέσματα. Εναλλακτικά, μπορεί κανείς να χρησιμοποιήσει άλλα κλασικά κριτήρια ελαχιστοποίησης σφαλμάτων (ME, MAPE, sMAPE κ.ο.κ.) όπως συμβαίνει π.χ. κατά τον υπολογισμό των παραμέτρων των μοντέλων εκθετικής εξομάλυνσης. Η πιθανοφάνεια αποτελεί ωστόσο όπως

αναφέρθηκε την πιο διαδεδομένη βιβλιογραφικά αντικειμενική συνάρτηση βελτιστοποίησης παραμέτρων.

Σε αυτό το σημείο αξίζει να σημειωθεί πως η πιθανοφάνεια δεν λαμβάνει καθόλου υπόψη της την πολυπλοκότητα του μοντέλου και *το αξιολογεί μόνο ως προς την ικανότητα προσαρμογής του*. Έτσι, ενώ απαντά στο ερώτημα *ποιες* θα πρέπει να είναι οι βέλτιστες παράμετροι του μοντέλου δεδομένης της τάξης του, δεν απαντά αποδοτικά στο *αν* κάποιο μοντέλο διαφορετικής πολυπλοκότητας είναι προτιμότερο προβλεπτικά. Επίσης, δεδομένου ότι το κριτήριο της πιθανοφάνειας δεν έχει όρια, δηλαδή μπορεί να λάβει οποιαδήποτε τιμή χωρίς να ξέρουμε ποια είναι η βέλτιστη (π.χ. μηδενικό σφάλμα), δεν μας δίνει καμία πληροφορία από μόνο του παρά μόνο όταν η τιμή του συγκριθεί με αυτές άλλων εναλλακτικών μοντέλων. Τέλος, αφού η ικανότητα προσαρμογής ενός μοντέλου είναι ανάλογη της πολυπλοκότητάς του, η πιθανοφάνεια έχει νόημα να *χρησιμοποιείται μόνο για τη σύγκριση μοντέλων ίδιας πολυπλοκότητας*.

3. Στο στάδιο του διαγνωστικού ελέγχου εφαρμόζονται στατιστικοί έλεγχοι προκειμένου να εξακριβωθεί αν τα μοντέλα που αναγνωρίστηκαν και εκτιμήθηκαν είναι προβλεπτικά άρτια. Ο διαγνωστικός έλεγχος γίνεται μελετώντας κυρίως την κατανομή των σφαλμάτων πρόβλεψης e_t των υποψήφιων μοντέλων. Αν το μοντέλο είναι άρτιο, τότε τα σφάλματα που αυτό παράγει θα πρέπει να οφείλονται αποκλειστικά στην τυχαιότητα της χρονοσειράς και συνεπώς να μην συσχετίζονται μεταξύ τους χρονικά.

Στη συνέχεια παρουσιάζονται οι βασικότερες αρχές των σταδίων αναγνώρισης, εκτίμησης και ελέγχου, καθώς και οι πλέον συνήθεις πρακτικές υλοποίησής τους.

4. Αναγνώριση μοντέλων ARIMA

Όπως αναφέρθηκε νωρίτερα, προκειμένου να αναγνωρίσει κανείς ποια μοντέλα ARIMA ενδέχεται να είναι κατάλληλα για την προέκταση μιας χρονοσειράς, απαιτούνται κάποιες ενδείξεις. Αυτές μπορεί να δοθούν είτε στατιστικά μέσω κατάλληλων τεστ υποθέσεων, είτε οπτικά μέσω διαγραμματικής ανάλυσης. Στη συγκεκριμένη παράγραφο θα αναφερθούμε αρχικά στις δεύτερες μεθόδους, που είναι πρακτικά εφαρμόσιμες για μικρό πλήθος χρονοσειρών, και στη συνέχεια θα δοθούν πληροφορίες για τις πρώτες που μπορούν να αξιοποιηθούν για την αυτοματοποίηση της όλης διαδικασίας σε εφαρμογές μεγάλου πλήθους δεδομένων.

4.1 Αναγνώριση με διαγραμματικές μεθόδους

Η γνώση των τιμών των συντελεστών αυτοσυσχέτισης (*Auto-Correlation Factor* - ***ACF***) και μερικής αυτοσυσχέτισης (*Partial Auto-Correlation Factor* - ***PACF***) μπορούν να αποτελέσουν σημαντικές ενδείξεις για το αν ένα μοντέλο ARIMA είναι κατάλληλο για την

προέκταση μιας χρονοσειράς. Αυτό συμβαίνει γιατί τα μοντέλα ARIMA βασίζονται ακριβώς στην υπόθεση ύπαρξης συσχετίσεων μεταξύ διαδοχικών παρατηρήσεων. Έτσι, αν οι εν λόγω συντελεστές σηματοδοτούν την ύπαρξη κάποιας στατιστικά σημαντικής συσχέτισης, γίνεται αντιληπτό πως αυτή μπορεί να αξιοποιηθεί άμεσα για την κατασκευή ενός αντίστοιχου μοντέλου.

Ο **συντελεστής αυτοσυσχέτισης** υστέρησης *k* μας δείχνει κατά πόσο η τιμή της χρονοσειράς σε μία περίοδο εξαρτάται στη γενική περίπτωση από την τιμή της παρατήρησης *k* περιόδων πίσω. Παίρνει τιμές από +1 έως -1, οι οποίες δηλώνουν απόλυτα θετική και αρνητική γραμμική συσχέτιση αντίστοιχα. Αν ο συντελεστής ισούται με μηδέν τότε δεν υπάρχει καμία συσχέτιση μεταξύ των δύο παρατηρήσεων. Η τιμή της αυτοσυσχέτισης δίνεται από την ακόλουθη σχέση:

$$p_k = \frac{\sum_{t=k+1}^{n}(Y_t - \mu)(Y_{t-k} - \mu)}{\sum_{t=1}^{n}(\mathrm{Y_t} - \mu)^2}$$

, όπου *μ* η μέση τιμή της χρονοσειράς.

Ο **συντελεστής μερικής αυτοσυσχέτισης** υστέρησης *k* δείχνει κατά πόσο η τιμή της χρονοσειράς σε μία περίοδο εξαρτάται από την τιμή της παρατήρησης *k* περιόδων πίσω, μη λαμβάνοντας υπόψη την επίδραση που μπορεί ενδεχομένως να επιφέρουν οι τιμές που παρεμβάλλονται. Προφανώς, για *k=1* ο δείκτης ACF ταυτίζεται με αυτόν του PACF. Και σε αυτήν την περίπτωση, ο συντελεστής παίρνει τιμές από +1 έως -1, οι οποίες δηλώνουν απόλυτα θετική και αρνητική γραμμική συσχέτιση αντίστοιχα. Η τιμή της μερικής αυτοσυσχέτισης δίνεται από τις ακόλουθες σχέσεις:

$$\varphi_{11} = \rho_1$$

$$\varphi_{22} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$$

$$\varphi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1}\varphi_{k-1,j}r_{k-j}}{1 - \sum_{j=1}^{k-1}\varphi_{k-1,j}r_j} \; \gamma\iota\alpha \; k = 3 \ldots$$

$$\varphi_{kj} = \varphi_{k-1,j} - \varphi_{kk}\varphi_{k-1,k-j} \; \gamma\iota\alpha \; k = 2 \ldots \; j = 1,2, \ldots k-1$$

Οι προσεγγίσεις των συντελεστών ACF και PACF αποτελούν στην πράξη τους συντελεστές των μοντέλων ARIMA και συμβολίζονται με φ_k και ϑ_k αντίστοιχα. Αν λοιπόν οι τιμές των συσχετίσεων είναι σημαντικές, τότε θα πρέπει να ληφθεί υπόψη και ο κατάλληλος παράγοντας ARIMA, σχηματίζοντας το αντίστοιχο μοντέλο. Πρακτικά, η σημαντικότητα των συντελεστών μπορεί να ελεγχθεί διαγραμματικά παρατηρώντας την εξέλιξη των

τιμών των συντελεστών ACF και PACF. Παρακάτω αναφέρονται ενδεικτικά κάποιες τέτοιες κλασικές περιπτώσεις.

Για ένα μοντέλο AR(p)

- Οι τιμές των συντελεστών ACF φθίνουν προς το μηδέν ακολουθώντας εκθετική η ημιτονοειδή πορεία
- Οι τιμές των συντελεστών PACF μηδενίζονται απότομα μετά από *p* περιόδους υστέρησης

Λεκτικά η συμπεριφορά που περιγράφεται από τα παραπάνω διαγράμματα μπορεί να εκφραστεί ως εξής: «*Όσο περνάει ο χρόνος τόσο λιγότερο συσχετίζεται η τιμή του μεγέθους με αυτές των προηγούμενων περιόδων, με τις p προηγούμενες παρατηρήσεις να την επηρεάζουν ωστόσο σημαντικά*».

Για ένα μοντέλο MA(q)

- Οι τιμές των συντελεστών ACF μηδενίζονται απότομα μετά από *q* περιόδους υστέρησης
- Οι τιμές των συντελεστών PACF φθίνουν προς το μηδέν ακολουθώντας εκθετική η ημιτονοειδή πορεία

Λεκτικά η συμπεριφορά που περιγράφεται από τα παραπάνω διαγράμματα μπορεί να εκφραστεί ως εξής: «*Η τιμή του μεγέθους παύει να συσχετίζεται σημαντικά με αυτές που είχε στο παρελθόν μετά από q περιόδους, με την επίδραση αυτών να φθίνει στο χρόνο*».

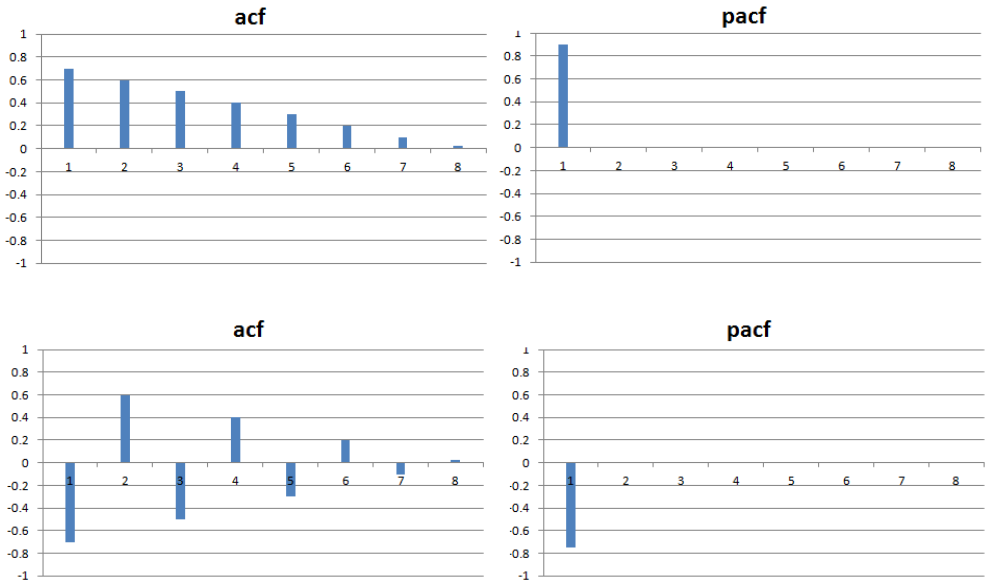
Για ένα μοντέλο ARIMA(p,q)

- Οι τιμές των συντελεστών ACF φθίνουν προς το μηδέν μετά από *q*-p περιόδους υστέρησης
- Οι τιμές των συντελεστών PACF φθίνουν προς το μηδέν μετά από *p-q* περιόδους υστέρησης

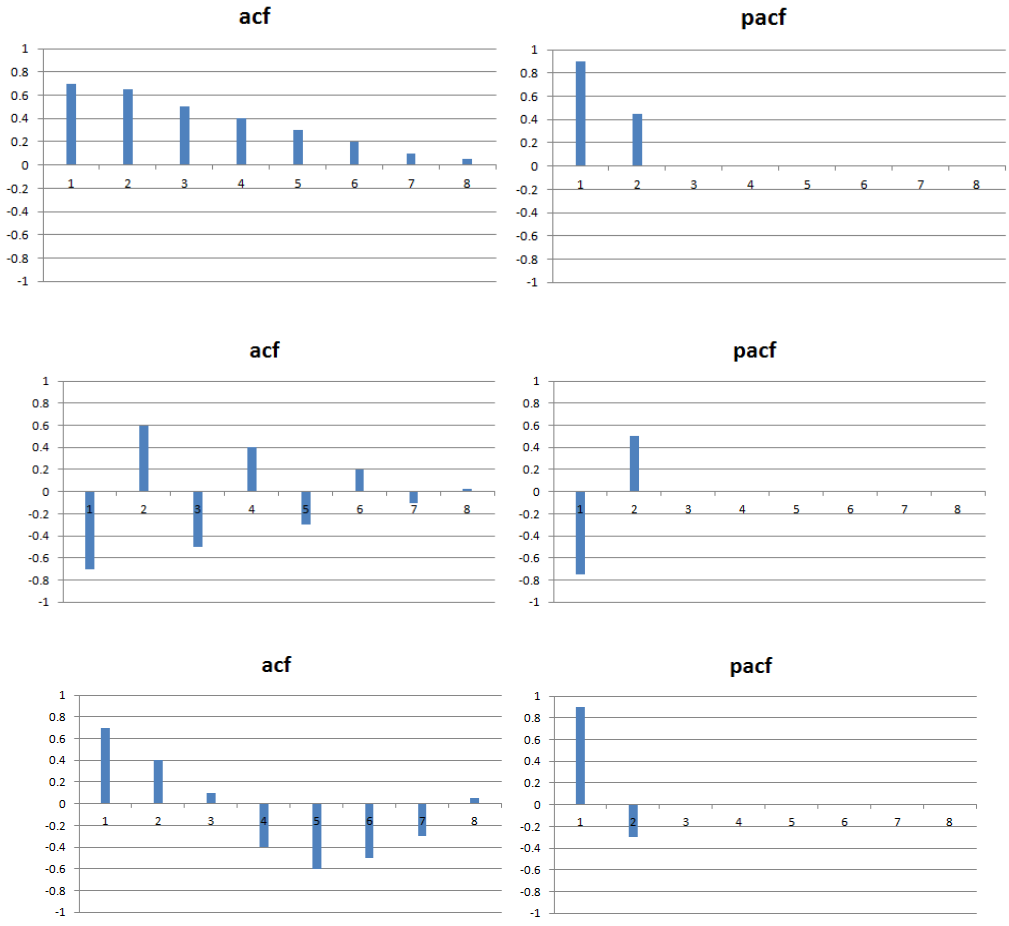
Λεκτικά η συμπεριφορά που περιγράφεται από τα παραπάνω διαγράμματα μπορεί να εκφραστεί ως εξής: «*Η τιμή του μεγέθους παύει να συσχετίζεται σημαντικά με αυτές που είχε στο παρελθόν μετά από q-q περιόδους, με τις q-q προηγούμενες να την επηρεάζουν ωστόσο σημαντικά*».

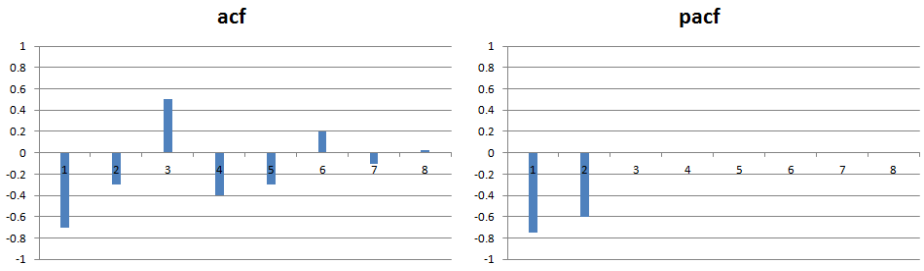
Ενδεικτικά παραδείγματα αναγνώρισης συνήθων διαδικασιών ARIMA παρουσιάζονται στα ακόλουθα διαγράμματα.

Διαδικασίες AR(1):

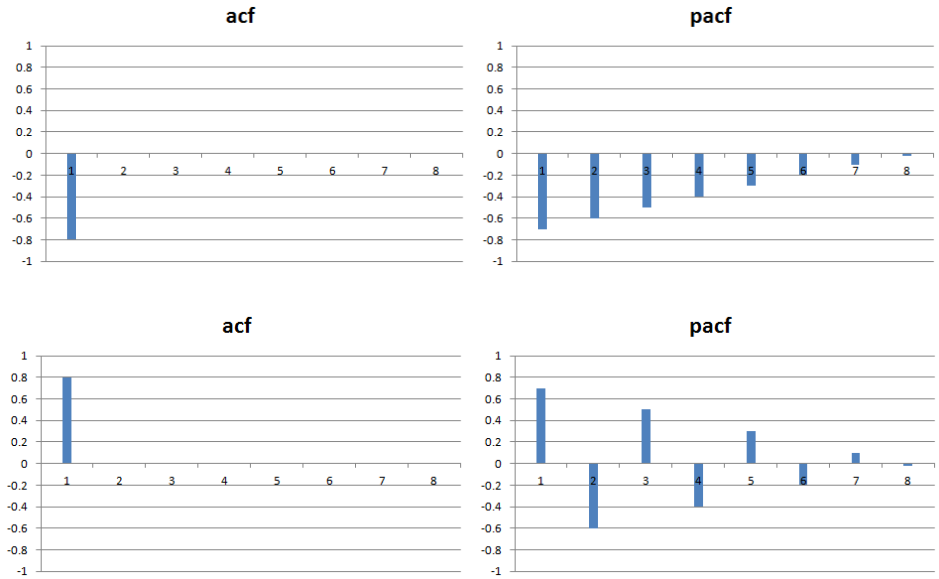


Διαδικασίες AR(2):

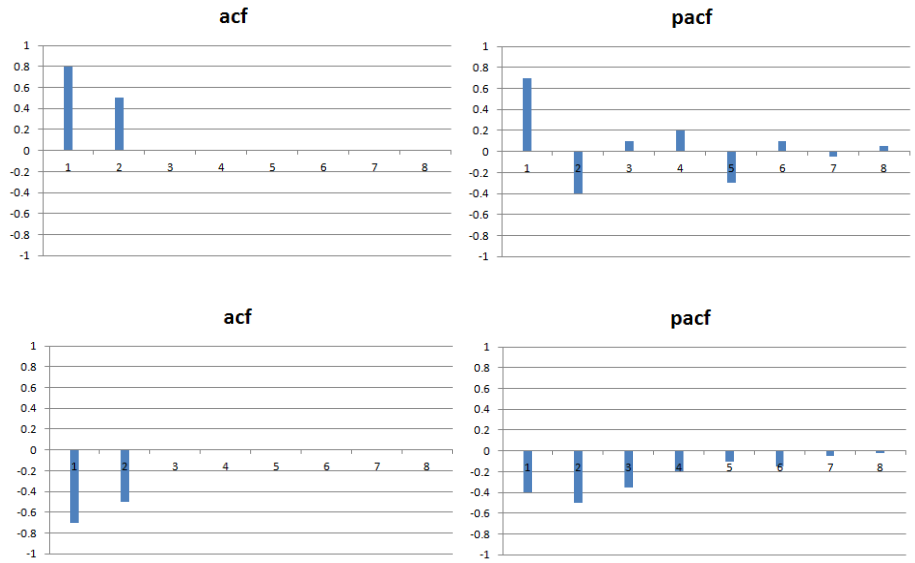


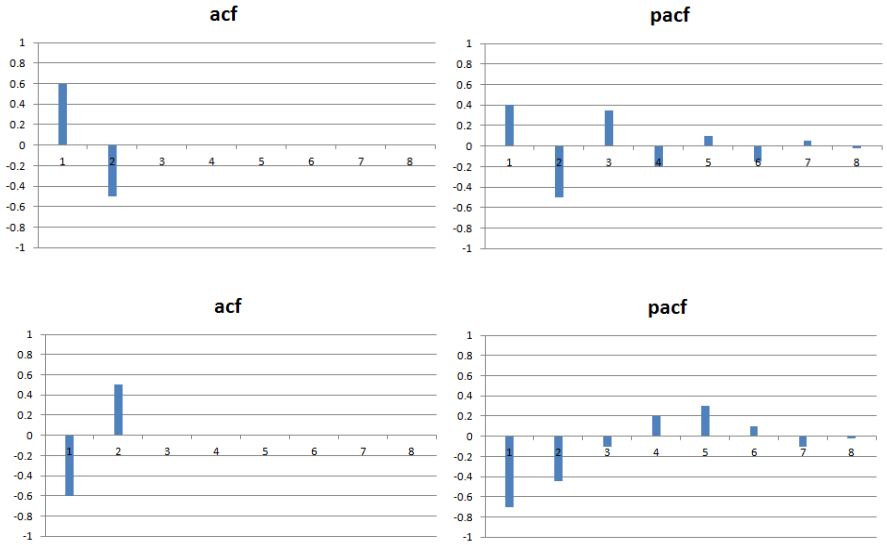


Διαδικασίες MA(1):

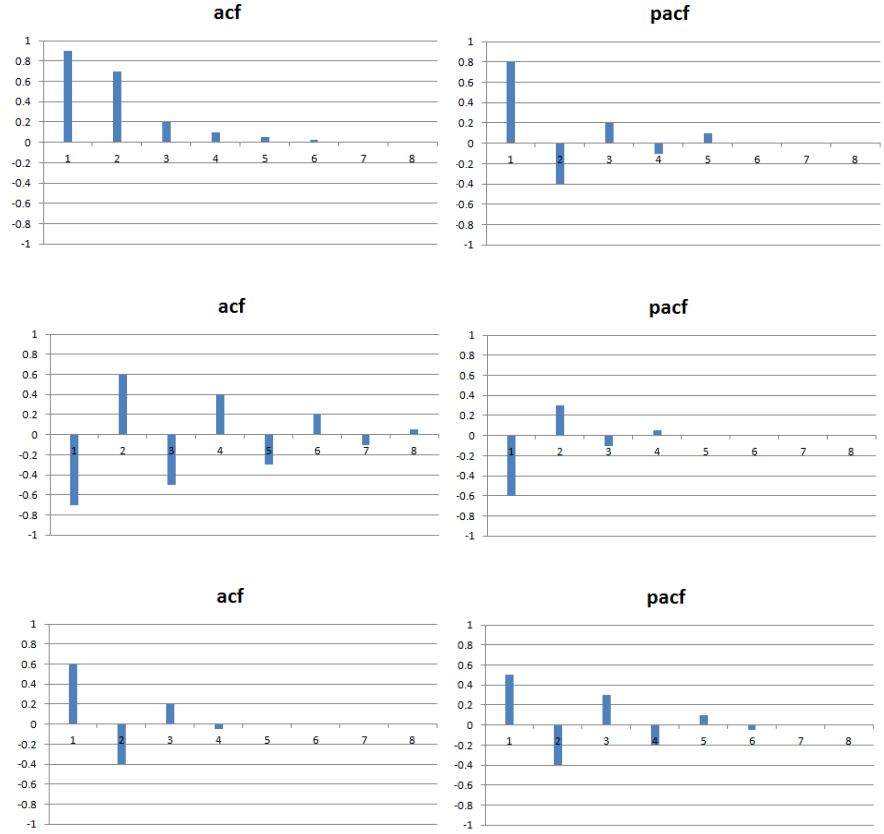


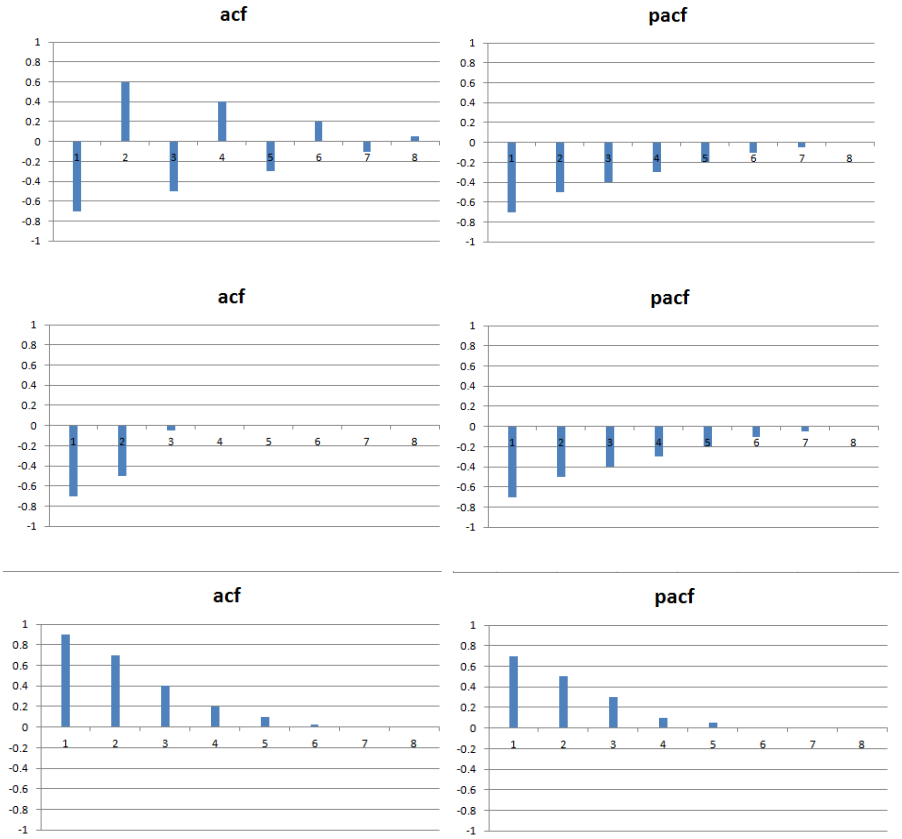
Διαδικασίες MA(2):





Διαδικασίες ARMA(1,1):





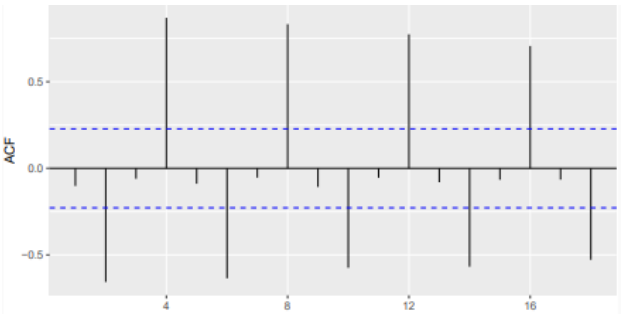
Σημειώνεται πως σε ρεαλιστικές συνθήκες τα μοτίβα που σκιαγραφήθηκαν και παρουσιάστηκαν νωρίτερα δεν είναι πάντοτε τόσο εμφανή. Για παράδειγμα, μπορεί λόγω τυχαιότητας ενώ οι συντελεστές ACF μιας χρονοσειράς μηδενίζουν πρακτικά μετά από *n* υστερήσεις, να παρουσιαστεί για μία μεμονωμένη υστέρηση *k>n* κάποια υψηλή τιμή συσχέτισης. Αυτός είναι άλλωστε και ο λόγος που σε τέτοιες περιπτώσεις συνίσταται η χρήση μετασχηματισμών. Αντίστοιχα, μπορεί λόγω τυχαιότητας να παραμορφωθεί για μεμονωμένες υστερήσεις η ημιτονοειδής ή η φθίνουσα συμπεριφορά ενός διαγράμματος PACF.

Έτσι, ποτέ δεν επιλέγουμε π.χ. ένα μοντέλο AR(p) αν δεν υπάρχει σημαντικότητα για όλες τις προηγούμενες υστερήσεις (*1,2,…p-1*). Με την ίδια λογική δεν απορρίπτουμε επίσης ποτέ π.χ. ένα μοντέλο MA(q) αν η ημιτονοειδής συμπεριφορά του διαγράμματος PACF διακόπτεται για μία μεμονωμένη υστέρηση. Συνοψίζοντας, προκειμένου να αναγνωρίζονται σωστά τα μοντέλα ARIMA μέσω διαγραμμάτων, πέρα από εξάσκηση, απαιτείται και αφαιρετική σκέψη: να μπορεί να εξάγει δηλαδή κανείς μέσα από αυτά τη γενικότερη συμπεριφορά των συσχετίσεων και των μερικών συσχετίσεων της χρονοσειράς.

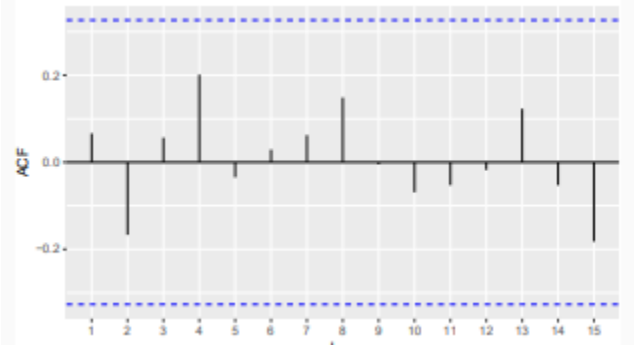
Δεδομένου του προβλήματος της υπερπροσαρμογής που αναφέρθηκε νωρίτερα, στην πράξη ποτέ δεν μελετάμε υστερήσεις μεγαλύτερες του πέντε (3) καθώς σε μία τέτοια περίπτωση το μοντέλο που θα αναγνωριστεί θα είναι ιδιαιτέρως πολύπλοκο και θα αποτελείται από μεγάλο πλήθος παραμέτρων. Αν υπάρχουν ενδείξεις για υστερήσεις μεγαλύτερες του 3, τότε το πιο πιθανό είναι ότι οφείλονται σε εποχιακούς και τυχαίους παράγοντες που δεν αντιμετωπίστηκαν αποδοτικά.

Σε αυτό το σημείο τονίζεται πως οι παραπάνω οπτικοί έλεγχοι μας βοηθούν να αναγνωρίσουμε μοντέλα ARMA και όχι μοντέλα ARIMA. Υποθέτουν δηλαδή πως η χρονοσειρά που μελετάται είναι στάσιμη και ότι, αν αυτή δεν ήταν στάσιμη αρχικά, έχουν γίνει όλες οι απαραίτητες ενέργειες για να προκύψει στασιμότητα. Η εν λόγω υπόθεση είναι καθοριστική καθώς η μη ικανοποίησή της θα έχει ως αποτέλεσμα τα διαγράμματα να παρουσιάζουν συσχετίσεις που μπορούν να αντιμετωπιστούν αποτελεσματικά με απλές διαφορίσεις και χωρίς να χρησιμοποιηθούν απαραίτητα παράγοντες AR και MA.

Ένας πρακτικός τρόπος για να καταλαβαίνουμε σε ποιες περιπτώσεις οι τιμές των αυτοσυσχετίσεων που παρουσιάζονται στα διαγράμματα ACF και PACF είναι στατιστικά σημαντικές, οπότε και απαιτούνται αντίστοιχες ενέργειες βάσει των μοτίβων τους, είναι να συγκρίνουμε το αποτέλεσμά τους με αυτό ενός λευκού θορύβου ίδιου πλήθους παρατηρήσεων. Για διάστημα εμπιστοσύνης 95%, οι αυτοσυσχετίσεις ενός λευκού θορύβου *N(0,1/n)* αναμένεται να λαμβάνουν τιμές στο διάστημα $[-1.96\sqrt{n}, 1.96\sqrt{n}]$. Αν η χρονοσειρά που μελετάται εμφανίζει παρόμοια συμπεριφορά, τότε δεν είναι στατιστικά διάφορη του λευκού θορύβου και συνεπώς είναι στάσιμη. Σε αντίθετη περίπτωση, είναι στατιστικά διάφορη και απαιτούνται αντίστοιχες ενέργειες. Παρακάτω παρουσιάζονται δύο διαγράμματα ACF για χρονοσειρές σημαντικά διάφορες και μη του λευκού θορύβου. Οι κρίσιμες τιμές επισημαίνονται μέσω των διακεκομμένων γραμμών.



Διάγραμμα ACF χρονοσειράς στατιστικά διάφορης του λευκού θορύβου.



Διάγραμμα ACF χρονοσειράς στατιστικά μη διάφορης του λευκού θορύβου.

4.2 Αναγνώριση με στατιστικές μεθόδους

Για να εξακριβωθεί με αυτόματο και στατιστικό τρόπο η ανάγκη διαφόρισης, υπάρχουν διάφορα τεστ υποθέσεων τα οποία μπορούν να φανούν ιδιαιτέρως χρήσιμα. Στην ουσία, τα εν λόγω τεστ κάνουν την αρχική υπόθεση ότι η χρονοσειρά είναι (ή δεν είναι) στάσιμη, εφαρμόζουν μία σειρά από ελέγχους και κριτήρια, και ανάλογα με τα αποτελέσματα απορρίπτουν ή δέχονται την αρχική υπόθεση προτείνοντας αντίστοιχες ενέργειες.

Το πλέον διαδεδομένο τεστ ελέγχου στασιμότητας είναι το **Augmented Dickey-Fuller (ADF)**, το οποίο και υποθέτει ότι η χρονοσειρά περιγράφεται από μία στοχαστική διαδικασία πρώτων διαφορών. Αν η εν λόγω διαδικασία έχει για ρίζα της το μηδέν (unit root), αυτό σημαίνει πως πιθανότατα δεν είναι στάσιμη και πως απαιτείται διαφόριση. Αντίθετα, αν δεν έχει για ρίζα της το μηδέν, τότε η χρονοσειρά είναι μάλλον στάσιμη και δεν απαιτείται διαφόριση. Το τεστ επιστρέφει ουσιαστικά την πιθανότητα (*p-value*) του παραπάνω σεναρίου, έχοντας ως αρχική υπόθεση πως η χρονοσειρά δεν είναι στάσιμη. Έτσι, υψηλές πιθανότητες σηματοδοτούν την έλλειψη στασιμότητας και μικρές την ύπαρξή της, ή καλύτερα τη μη ύπαρξη ενδείξεων για να υποστηριχθεί το αντίθετο. Συνήθως, ένα εμπειρικό κριτήριο που ακολουθείται για να υποστηρίξει κανείς την αρχική υπόθεση, είναι η πιθανότητα που υπολογίζεται να ξεπερνά το 0.05 (5%).

Αντίστοιχο τεστ είναι το **Kwiatkowski-Phillips-Schmidt-Shin (KPSS)**, με τη διαφορά ότι σε αυτήν την περίπτωση μικρές πιθανότητες συνιστούν την εφαρμογή διαφόρισης. Επίσης, για τον έλεγχο εποχιακής διαφόρισης, αρκετά χρήσιμο μπορεί να φανεί το τεστ **Canova-Hansen**.

Αξιοποιώντας λοιπόν τα παραπάνω τεστ, μπορεί κανείς αρκετά εύκολα και γρήγορα να επιλέξει τις τιμές των παραμέτρων *d* και *D* των μοντέλων ARIMA που θα αναγνωρίσει στη συνέχεια, δηλαδή την τάξη εποχιακής και μη διαφόρισής τους: Εφαρμόζει το τεστ της επιλογής του στα δεδομένα με και χωρίς της χρήση διαφόρισης (πρώτης τάξης, δεύτερης τάξης και πρώτης εποχιακής τάξης) και ανάλογα με τα αποτελέσματα προβαίνει σε

κατάλληλες ενέργειες για να οδηγηθεί σε στασιμότητα. Στη συνέχεια, εφαρμόζει επί της στάσιμης χρονοσειράς διάφορα αντιπροσωπευτικά μοντέλα ARMA και υπολογίζει για καθένα από αυτά την προσδοκώμενη πιθανοφάνεια. Για τα μοντέλα ίδιας πολυπλοκότητας επιλέγεται εκείνο που μεγιστοποιεί την πιθανοφάνεια, ενώ μεταξύ μοντέλων διαφορετικής πολυπλοκότητας επιλέγεται εκείνο που ελαχιστοποιεί την τιμή ενός σχετικού κριτηρίου (information criteria), όπως αυτά που παρουσιάζονται στην παράγραφο 6.

Αξίζει να σημειωθεί πως η στατιστική αναγνώριση των μοντέλων διαφέρει σημαντικά από τη διαγραμματική, με τις αποφάσεις στη δεύτερη περίπτωση να λαμβάνονται αρκετά πιο άμεσα και να έχουν μόνιμη ισχύ. Για παράδειγμα, η χρήση διαγραμμάτων μας βοηθά να αποφανθούμε αμέσως για το αν η χρονοσειρά χαρακτηρίζεται από τάση ή εποχιακότητα. Αντίστοιχα, επιλέγεται ένα και μόνο μοντέλο ARMA για την προέκταση της στάσιμης χρονοσειράς που έχει προκύψει. Αντίθετα, η στατιστική μέθοδος μελετά παράλληλα αρκετά διαφορετικά σενάρια και επιλέγει το καλύτερο βάσει της πιθανότητας που έχει το εκάστοτε μοντέλο να περιγράφει το μοτίβο της χρονοσειράς.

5. Εκτίμηση μοντέλων ARIMA

5.1 Μοντέλα αυτοπαλινδρόμησης - AR (p)

Τα συνήθη μοντέλα παλινδρόμησης θεωρούν γραμμικές ή μη σχέσεις μεταξύ της υπό εξέταση χρονοσειράς και των μεταβλητών από τις οποίες επηρεάζεται για να την προεκτείνουν στο μέλλον. Από τη μεριά τους τα μοντέλα αυτοπαλινδρόμησης θεωρούν γραμμικές σχέσεις ανάμεσα στις παρατηρήσεις της ίδιας της χρονοσειράς και τις αξιοποιούν για την περιγραφή της και την παραγωγή προβλέψεων. Ένα τέτοιο μοντέλο αναφέρεται ως *μοντέλο αυτοπαλινδρόμησης AR (p)*, όπου *p* η τάξη του, δηλαδή το πλήθος των παρελθοντικών παρατηρήσεων που αξιοποιούνται για την παραγωγή της πρόβλεψης. Αλγεβρικά αυτό αναπαρίσταται ως εξής:

$$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} \quad (1)$$

,όπου φ_i οι συντελεστές αυτοσυσχέτισης του μοντέλου για υστέρηση *i* και *c* μία σταθερά.

Στην ουσία το μοντέλο υποθέτει πως η τιμή της παρατήρησης y_t εξαρτάται κατά παράγοντα φ_1 από την προηγούμενη παρατήρηση, κατά παράγοντα φ_2 από την προ-προηγούμενη παρατήρηση … και κατά παράγοντα φ_p από την παρατήρηση που βρίσκεται *p* περιόδους πίσω. Υπολογίζεται ως γραμμικός συνδυασμός αυτών προσαυξάνοντάς την προαιρετικά κατά μία σταθερά *c*.

Χρησιμοποιώντας τον τελεστή ολίσθησης, ένα μοντέλο AR μπορεί να γραφτεί και ως:

$$(\mathbf{1} - \boldsymbol{\varphi_1 B} - \boldsymbol{\varphi_2 B^2} - \cdots - \boldsymbol{\varphi_p B^p})\overline{\boldsymbol{y}}_{\boldsymbol{t}} = \mathbf{0}$$

, όπου $\overline{y_t} = y_t - \mu$. Η χρονοσειρά $\overline{y_t}$ έχει τις ίδιες στατιστικές ιδιότητες με την αρχική χρονοσειρά με μηδενική μέση τιμή. Η χρήση της γίνεται προκειμένου να τονιστούν οι στοχαστικές συνιστώσες της χρονοσειράς. Αν αναπτύξουμε τώρα την παραπάνω σχέση έχουμε:

$$\overline{y}_t - \varphi_1 B\overline{y}_t - \varphi_2 B^2\overline{y}_t - \cdots - \varphi_p B^p\overline{y}_t = 0 \rightarrow$$

$$\overline{y}_t - \varphi_1\overline{y}_{t-1} - \varphi_2\overline{y}_{t-2} - \cdots - \varphi_p\overline{y}_{t-p} = 0 \rightarrow$$

$$y_t - \mu - \varphi_1(y_{t-1} - \mu) - \varphi_2(y_{t-2} - \mu) - \cdots - \varphi_p(y_{t-p} - \mu) = 0$$

$$y_t = \mu\left(1 - \varphi_1 - \varphi_2 - \cdots \varphi_p\right) + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} \ (2)$$

Από τις σχέσεις (1) και (2) οδηγούμαστε στο συμπέρασμα ότι για τη σταθερά *c* σε ένα μοντέλο *AR(p)* ισχύει

$$c = \mu\left(1 - \varphi_1 - \varphi_2 - \cdots \varphi_p\right)$$

Παρατηρήστε τέλος πως ένα μοντέλο AR(1),

- Ταυτίζεται με τον λευκό θόρυβο όταν $\varphi_1=0$
- Ταυτίζεται με τον τυχαίο περίπατο όταν $\varphi_1=1$ και $c=0$
- Ταυτίζεται με τον τυχαίο περίπατο με τάση όταν $\varphi_1=1$ και *c* διάφορο του μηδενός
- Ταλαντώνεται μεταξύ αρνητικών και θετικών τιμών όταν $\varphi_1<0$

5.2 Μοντέλα κινητού μέσου όρου – MA (q)

Τα μοντέλα κινητού μέσου όρου θεωρούν γραμμικές σχέσεις ανάμεσα στην παρατήρηση της χρονοσειράς που εξετάζεται και στα σφάλματα που εμφάνισε το μοντέλο MA σε προηγούμενες περιόδους. Ένα τέτοιο μοντέλο γράφεται αλγεβρικά ως εξής:

$$y_t = c + \theta_1 e_{t-1} + \ \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \ \ (3)$$

,όπου ϑ_i οι συντελεστές μερικής αυτοσυσχέτισης του μοντέλου MA για υστέρηση *i.*

Στην ουσία η τιμή της παρατήρησης y_t εξαρτάται κατά παράγοντα ϑ_1 από το σφάλμα που παρήγαγε το μοντέλο την προηγούμενη περίοδο, κατά παράγοντα ϑ_2 από το σφάλμα που παρήγαγε το μοντέλο την προ-προηγούμενη περίοδο … και κατά παράγοντα ϑ_q από το σφάλμα του μοντέλου *q* περιόδους πίσω. Υπολογίζεται ως γραμμικός συνδυασμός αυτών προσαυξάνοντάς την –προαιρετικά- κατά μία σταθερά *c*.

Χρησιμοποιώντας τον τελεστή ολίσθησης, ένα μοντέλο MA μπορεί να γραφτεί και ως:

$$\overline{\boldsymbol{y}}_{\boldsymbol{t}} = \left(\boldsymbol{\theta_1 B} + \boldsymbol{\theta_2 B^2} - \cdots + \boldsymbol{\theta_q B^q}\right)\boldsymbol{e_t}$$

Αν αναπτύξουμε τώρα την παραπάνω σχέση έχουμε:

$$\overline{y}_t = \theta_1 B e_t + \theta_2 B^2 e_t + \cdots + \theta_q B^q e_t \rightarrow$$

$$y_t = \mu + \theta_1 B e_t + \theta_2 B^2 e_t + \cdots + \theta_q B^q e_{t-q} \rightarrow$$

$$y_t = \mu + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \ (4)$$

Από τις σχέσεις (3) και (4) οδηγούμαστε στο συμπέρασμα ότι για τη σταθερά c σε ένα μοντέλο *MA(q)* ισχύει

$$c = \mu$$

Μία αρκετά ενδιαφέρουσα παρατήρηση είναι ότι *κάθε μοντέλο AR μπορεί να γραφτεί ως μοντέλο MA απείρων όρων*. Χωρίς βλάβη της γενικότητας αναλύουμε για παράδειγμα το μοντέλο AR(1) με τον εξής τρόπο:

$$y_t = c + \varphi_1 y_{t-1} \rightarrow$$

$$y_t = c + \varphi_1 (c + \varphi_1 y_{t-2} + \ e_{t-1}) \rightarrow$$

$$y_t = c(1 + \varphi_1) + {\varphi_1}^2 y_{t-2} + \ \varphi_1 e_{t-1} \rightarrow$$

$$y_t = c(1 + \varphi_1) + {\varphi_1}^2 (c + \varphi_1 y_{t-3} + \ e_{t-2}) + \ \varphi_1 e_{t-1} \rightarrow$$

$$y_t = c(1 + \varphi_1 + {\varphi_1}^2) + {\varphi_1}^3 y_{t-3} + {\varphi_1}^2 e_{t-2} + \varphi_1 e_{t-1}$$

$$y_t = c(1 + \varphi_1 + {\varphi_1}^2 + \cdots {\varphi_1}^n) + {\varphi_1}^n y_{t-n} + \cdots + {\varphi_1}^2 e_{t-2} + \varphi_1 e_{t-1}$$ κ.ο.κ.

Δεδομένου ότι *-1<φ₁<1*, μετά από άπειρες επαναλήψεις η της παραπάνω διαδικασίας ο όρος ${\varphi_1}^n y_{t-n}$ θα τείνει στο μηδέν. Έτσι η διαδικασία θα έχει μετατραπεί σε αμιγώς MA μοντέλο.

Για ένα καλώς ορισμένο μοντέλο MA ισχύει επίσης και η αντίστροφη διαδικασία. Έστω για παράδειγμα το μοντέλο MA(1):

$$\overline{y}_t = \theta_1 B e_t \rightarrow$$

$$(\theta_1 B)^{-1} \overline{y}_t = e_t$$

Σύμφωνα με θεώρημα των γεωμετρικών σειρών αν *-1<ϑ₁<1*, τότε ο όρος $(\theta_1 B)^{-1}$ μπορεί να γραφτεί ως ένα άθροισμα απείρων όρων μιας συγκλίνουσας σειράς $\left(\theta_1 B + {\theta_1}^2 B^2 + {\theta_1}^3 B^3 + \cdots\right)$. Έτσι η αρχική σχέση μπορεί να γραφτεί ως ένα αμιγώς AR μοντέλο.

Οι δύο παραπάνω μετασχηματισμοί βασίζονται σε μία ιδιότητα των μοντέλων AR και MA που ονομάζεται αντιστρεψιμότητα. Η αντιστρεψιμότητα, η οποία γίνεται υπό συγκεκριμένες συνθήκες για κάθε μοντέλο (βλ. εδώ *-1<ϑ₁<1* για MA(1) και AR(1)), διασφαλίζεται αν υπάρχει στασιμότητα στο μοντέλο και κατά συνέπεια μία καλή εκτίμηση της αρχικής χρονοσειράς. Τα μοντέλα AR είναι πάντα αντιστρέψιμα σε αντίθεση με τα MA. Η αντιστρεψιμότητα για τα μοντέλα MA μπορεί να γίνει και πρακτικά κατανοητή αν σκεφτούμε το εξής: Νωρίτερα θεωρώντας *-1<ϑ₁<1* αποδείξαμε ότι η διαδικασία MA(1) μπορεί να γραφτεί ως διαδικασία AR απείρων όρων, κάθε ένας εκ των οποίων έχει τη μορφή ${\theta_1}^i B^i \overline{y}_t$. Αν δεν ίσχυε η συνθήκη που θεωρήσαμε αντί ο συντελεστής ${\theta_1}^i$ να μικραίνει όσο αυξάνει το *i* θα αυξανόταν, δηλαδή οι πιο απομακρυσμένες παρατηρήσεις θα έπαιζαν μεγαλύτερο ρόλο στην πρόβλεψη των μελλοντικών τιμών. Αυτό βεβαίως είναι άτοπο. Παρακάτω θα δοθούν πιο αναλυτικά οι περιορισμοί παραμέτρων για κάθε μοντέλο ARIMA.

5.3 Ολοκληρωμένα αυτοπαλινδρομικά μοντέλα κινητού μέσου όρου - ARIMA (p,d,q)

Τα μοντέλα AR και MA μπορούν να συνδυαστούν αποδοτικά για την ανάλυση και πρόβλεψη στάσιμων χρονοσειρών. Έτσι, εισάγοντας στην εξίσωση και τα μοντέλα διαφόρισης για τη διασφάλιση της στασιμότητας, προκύπτουν τα μοντέλα ARIMA(p,d,q), όπου *p*, *d*, *q* η τάξη του αντίστοιχου μοντέλου. Το συνολικό μοντέλο αναπαρίσταται με τη χρήση του τελεστή ολίσθησης B ως εξής:

$$\begin{aligned}\left(\mathbf{1} - \boldsymbol{\varphi_1 B} - \boldsymbol{\varphi_2 B^2} - \cdots - \boldsymbol{\varphi_p B^p}\right)(\mathbf{1} - \boldsymbol{B})^{\boldsymbol{n}}(\mathbf{1} - \boldsymbol{B^m})^{\boldsymbol{N}}\boldsymbol{y_t} \\ = \boldsymbol{c} + \left(\boldsymbol{\theta_1 B} + \boldsymbol{\theta_2 B^2} + \cdots + \boldsymbol{\theta_q B^q}\right)\boldsymbol{e_t}\end{aligned}$$

Ο πρώτος όρος του πρώτου μέλους της εξίσωσης αναπαριστά το μοντέλο AR(p), ο δεύτερος την διαφόριση I(d), ενώ ο όρος στο δεύτερο μέλος της εξίσωσης το μοντέλο MA(q).

Θέλοντας να προσδιορίσουμε την σταθερά c, ακολουθούμε την ίδια διαδικασία με πριν. Εξισώνουμε δηλαδή τις δύο παρακάτω σχέσεις:

$$\begin{aligned}\left(1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p\right)(1 - B)^n(1 - B^m)^N \overline{y_t} \\ = \left(\theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right)e_t\end{aligned}$$

$$\overline{y}'_t = c + \varphi_1 y'_{t-1} + \varphi_2 y'_{t-2} + \cdots + \varphi_p y'_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}$$

,όπου $\overline{y}'_t$το προϊόν της εποχιακής και μη διαφόρισης της χρονοσειράς. Αναπτύσσοντας τώρα την πρώτη σχέση κατά τα γνωστά προκύπτει ότι:

$$\begin{aligned} y_t\left(1-\varphi_1-\varphi_2-\cdots-\varphi_p\right)(1-B)^n(1-B^m)^N \\ -\mu\left(1-\varphi_1-\varphi_2-\cdots-\varphi_p\right)(1-B)^n(1-B^m)^N \\ =\theta_1 B e_t+\theta_2 B^2 e_t+\cdots+\theta_q B^q e_{t-q} \end{aligned}$$

Η σταθερά *c* ισούται προφανώς με τον αντίθετο του δεύτερου όρου του πρώτου μέλους της παραπάνω εξίσωσης. Έτσι προκύπτει ότι για *n=N=0 για την σταθερά c θα ισχύει*:

$$c=\mu(1-\varphi_1-\varphi_2-\cdots-\varphi_p)$$

Σε οποιαδήποτε άλλη περίπτωση η σταθερά ισούται με μηδέν. Αρκεί να σκεφτούμε ότι για *n* ή *N* μεγαλύτερο του μηδέν ο δεύτερος όρος του πρώτου μέλους της εξίσωσης γράφεται ως:

$$\mu\left(\varphi_1+\varphi_2-\cdots+\varphi_p-1\right)(1-B)(1-B^m)(1-B)^{n-1}(1-B^m)^{N-1}=$$

$$K(1-B)(1-B^m)(1-B)^{n-1}(1-B^m)^{N-1}=$$

$$K(1-B^m-B+B^{2m})(1-B)^{n-1}(1-B^m)^{N-1}=$$

$$(K-K-K+K)(1-B)^{n-1}(1-B^m)^{N-1}=0$$

Αυτό ήταν αναμενόμενο δεδομένου ότι ιδανικά ο μέσος όρος μίας διαφορισμένης και άρα στάσιμης χρονοσειράς ισούται με μηδέν.

Πρακτικά η χρονοσειρά που προκύπτει από τη διαφόριση δεν είναι ποτέ απολύτως στάσιμη γύρω από το μηδέν. Έτσι, συνηθίζεται να προστίθεται σταθερά ακόμα και σε χρονοσειρές που έχουν υποστεί διαφόριση. Συνοπτικά αναφέρουμε τα εξής:

- Η μη διαφόριση (*d=0*) είναι δείγμα ύπαρξης *σταθερότητας* στην αρχική χρονοσειρά. Ωστόσο αυτή η σταθερότητα ενδέχεται να μην έχει για κέντρο της το μηδέν. Η εισαγωγή μίας σταθεράς *c* μπορεί να βοηθήσει στον καλύτερο προσδιορισμό του πραγματικού επιπέδου της χρονοσειράς και συνεπώς να βελτιώσει προβλεπτικά την επίδοσή του.
- Η διαφόριση πρώτης τάξης (*d=1*) είναι δείγμα ύπαρξης *τάσης* στην αρχική χρονοσειρά. Αυτή εξαλείφεται κατά τη διαφόριση, οπότε και οι προβλέψεις που παράγονται στη συνέχεια κινούνται γύρω από ένα σταθερό επίπεδο. Η εισαγωγή σταθεράς μπορεί να βοηθήσει στην επανένταξη της τάσης που αφαιρέθηκε κατά τη διαφόριση στις τελικές προβλέψεις και συνεπώς να βελτιώσει προβλεπτικά την επίδοση του μοντέλου.

- Η διαφόριση δεύτερης τάξης (*d=2*) συνεπάγεται ύπαρξη *χρονικά μεταβαλλόμενης τάσης* (τάση μέσα στην τάση) στην αρχική χρονοσειρά. Έτσι, η εισαγωγή σταθεράς σε αυτήν την περίπτωση θεωρείται άστοχη επιλογή, εκτός και αν υπάρχει σοβαρή ένδειξη ύπαρξης εκθετικού μοτίβου ανάπτυξης.

Προκειμένου να αντιλαμβάνεται κανείς στην πράξη την επίδραση της απλής διαφόρισης και της σταθεράς, αρκεί να θυμάται τα εξής:

- Αν c=0 και d=0, μακροπρόθεσμα οι προβλέψεις θα ισούνται με μηδέν
- Αν c=0 και d=1, μακροπρόθεσμα οι προβλέψεις θα ισούνται με μία μη μηδενική σταθερά
- Αν c=0 και d=2, μακροπρόθεσμα οι προβλέψεις θα ακολουθούν μία ευθεία γραμμή
- Αν c≠0 και d=0, μακροπρόθεσμα οι προβλέψεις θα ισούνται με τη μέση τιμή της χρονοσειράς
- Αν c≠0 και d=1, μακροπρόθεσμα οι προβλέψεις θα ακολουθούν μία ευθεία γραμμή
- Αν c≠0 και d=2, μακροπρόθεσμα οι προβλέψεις θα ακολουθούν μία εκθετική καμπύλη

5.4 Συνθήκες συντελεστών για τα μοντέλα ARMA

Η εφαρμογή των μοντέλων ARMA οφείλει να γίνεται μόνο επί στάσιμων χρονοσειρών καθώς κάτι τέτοιο εξασφαλίζει ικανοποιητικές εκτιμήσεις για τη μέση τιμή, τη διακύμανση και τη συνάρτηση αυτοσυσχέτισης της χρονοσειράς και συνεπώς ικανοποιητική παραμετροποίηση για τα υπό εξέταση μοντέλα. Προκειμένου να εξασφαλιστεί λοιπόν ότι οι συντελεστές *φi* και *ϑi* ενός μοντέλου ARMA έχουν εκτιμηθεί σωστά, καλούνται συνήθως να ικανοποιούν συγκεκριμένους περιορισμούς.

Παρακάτω δίνονται οι περιορισμοί που πρέπει να πληρούνται από τα πιο διαδεδομένα μοντέλα ARMA. Για τα υπόλοιπα μοντέλα που είναι και αρκετά πιο σύνθετα στη δομή τους, οι αντίστοιχοι περιορισμοί γίνονται ιδιαιτέρως πολύπλοκοι και για αυτό το λόγο δεν παρουσιάζονται.

- AR(1): $-1<\varphi_1<1$
- AR(2): $-1<\varphi_2<1$ και $\varphi_1 + \varphi_2 <1$ και $\varphi_2 - \varphi_1 <1$
- MA(1): $-1<\vartheta_1<1$
- MA(2): $-1<\vartheta_2<1$ και $\theta_1 + \vartheta_2 >-1$ και $\vartheta_1 - \vartheta_2 <1$

Επίσης, δεδομένων των τιμών αυτοσυσχέτισης μίας χρονοσειράς, για τους συντελεστές των μοντέλων ARMA ισχύει:

ARMA(p, q)	p_1	p_2
AR(1)	φ_1	-
MA(1)	$\frac{-\theta_1}{1+{\theta_1}^2}$	-
AR(2)	$\frac{\varphi_1}{1-\varphi_2}$	$\frac{{\varphi_1}^2}{1-\varphi_2}+\varphi_2$
MA(2)	$\frac{-\theta_1(1-\theta_2)}{1+{\theta_1}^2+{\theta_2}^2}$	$\frac{-\theta_2}{1+{\theta_1}^2+{\theta_2}^2}$
ARMA(1,1)	$\frac{(1-\varphi_1\theta_1)(\varphi_1-\theta_1)}{1+{\theta_1}^2-2\varphi_1\theta_1}$	$\rho_1\varphi_1$

Γενικά αποδεικνύεται ότι για **αμιγώς MA(q) διαδικασίες** ισχύει για τη συνάρτηση αυτοσυσχέτισης:

$$\rho_k = \frac{-\theta_k+\theta_{k+1}\theta_1+\cdots+\theta_{q-k}\theta_q}{1+{\theta_1}^2+{\theta_2}^2+\cdots+{\theta_q}^q}$$ για k=1,2,..q και $\rho_k = 0$ για $k>q$.

Αντίστοιχα για **αμιγώς AR(p) διαδικασίες** ισχύει για τη συνάρτηση αυτοσυσχέτισης:

$$\rho_k = {\rho_1}^k$$ για k=1,2,..p

6. Διάγνωση μοντέλων ARIMA

Ο διαγνωστικός έλεγχος έχει ως σκοπό να επιβεβαιώσει πως το μοντέλο το οποίο επιλέχθηκε και κατασκευάστηκε για να προεκτείνει τη χρονοσειρά ενδιαφέροντος είναι προβλεπτικά άρτιο. Νωρίτερα παρουσιάστηκαν κάποιες τεχνικές που βοηθούν σε αυτήν την κατεύθυνση, όπως:

- Η εφαρμογή μετασχηματισμών
- Η επιλογή βέλτιστης τάξης διαφόρισης, εποχιακής και μη
- Η μείωση της πολυπλοκότητας του μοντέλου
- Η εφαρμογή περιορισμών επί των παραμέτρων
- Η χρήση σταθεράς
- Η χρήση κριτηρίων για τη βελτίωση της προσαρμογής

Ωστόσο, η χρήση αυτών των τεχνικών είναι ικανή και όχι αναγκαία συνθήκη για την κατασκευή ακριβών και ευσταθών μοντέλων πρόβλεψης. Έτσι, μετά την ολοκλήρωση της όλης διαδικασίας, μπορεί κανείς να εφαρμόσει διάφορους διαγνωστικούς ελέγχους.

Μία πρώτη διάγνωση μπορεί να γίνει ελέγχοντας αν τα υπολειπόμενα σφάλματα εμφανίζουν κάποιου είδους συσχέτιση. Αυτό μπορεί εύκολα να ποσοτικοποιηθεί κατά τα γνωστά μέσα από τη σχέση

$$r_k(e) = \frac{\sum_{t=1+k}^{n}(e_t - \overline{e})(e_{t-k} - \overline{e})}{\sum_{t=1}^{n}(e_t - \overline{e})^2}$$

, όπου r_k η αυτοσυσχέτιση των σφαλμάτων για υστέρηση *k* ή πιο πρακτικά το ποσοστό διασύνδεσης δύο σφαλμάτων που παράχθηκαν με χρονική διαφορά *k* περιόδων. Είναι προφανές πως αν για κάποια υστέρηση υπάρχει σημαντική συσχέτιση, τότε το μοντέλο παράγει συστηματικά σφάλματα και συνεπώς δεν είναι προβλεπτικά άρτιο. Χαρακτηριστικό παράδειγμα αποτελεί η χρήση ενός μη εποχιακού μοντέλου για την πρόβλεψη μιας εποχιακής τριμηνιαίας χρονοσειράς, το οποίο σε αυτήν την περίπτωση θα παρουσιάζει παρόμοια σφάλματα ανά 4 περιόδους (*k=4*). Το ίδιο θα ισχύει και για τη χρήση ενός μοντέλου σταθερού επιπέδου σε χρονοσειρά με τάση, όπου με το πέρασμα του χρόνου (*k=1*) το παραγόμενο σφάλμα θα αυξάνεται γραμμικά.

Φυσικά τα σφάλματα ενός μοντέλου ARIMA ποτέ δεν είναι εντελώς ασυσχέτιστα στη πράξη, όσο καλά και αν αυτό περιγράφει τη χρονοσειρά. Αναμένουμε λοιπόν για κάποιες υστερήσεις να βρούμε μη μηδενικούς δείκτες συσχέτισης. Για να ελέγξουμε αν αυτοί είναι στατιστικά σημαντικά διάφοροι του μηδενός, υπολογίζονται οι κατά προσέγγιση *t-τιμές* του τυπικού σφάλματός τους.

$$t_{r_k} = \frac{r_k(e)}{S(r_{k(e)})}$$

$$S(r_k(e)) = n^{-1/2}(1 + 2\sum_{j=1}^{k-1} {r_{j(e)}}^2)^{1/2}$$

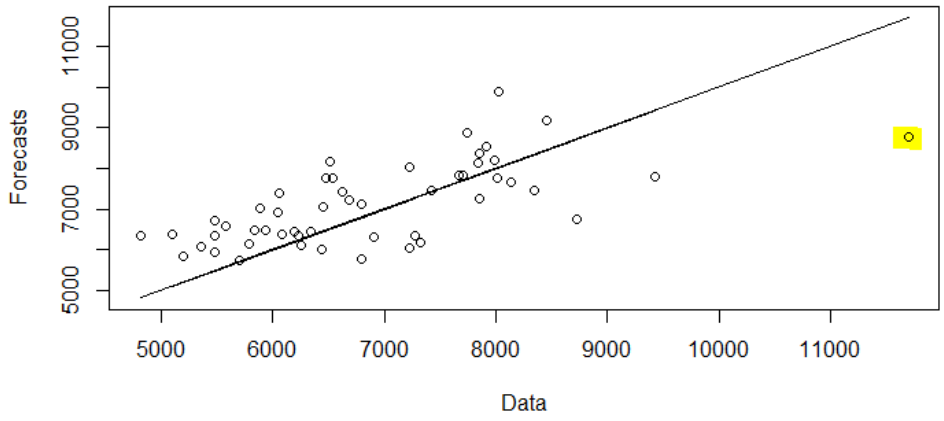
Για να μην είναι σημαντική μία αυτοσυσχέτιση (π.χ. με διάστημα εμπιστοσύνης 95% και κανονική κατανομής πιθανότητας) ο t-δείκτης δεν πρέπει θεωρητικά να ξεπερνά την τιμή *2*. Στην πράξη, για υστέρηση *k=1, 2 και 3* αυτή πρέπει να είναι μικρότερη του *1.25* και για μεγαλύτερη υστέρηση μικρότερη του *1.6*.

Σε αυτό το σημείο αναφέρεται πως ο διαγνωστικός έλεγχος, πέραν του στατιστικού τρόπου που αναφέρθηκε, μπορεί να γίνει και οπτικά μέσω της χρήσης διαγραμμάτων κατανομής προβλέψεων και υπολειπόμενων σφαλμάτων. Η εν λόγω εναλλακτική μπορεί να φανεί ιδιαίτερα χρήσιμη καθώς, όπως αναφέρθηκε νωρίτερα, όσο καλά και αν ένα μοντέλο προσαρμόζεται στα δεδομένα, λόγω τυχαιότητας, σπανίως θα καταφέρνει να παράγει αυστηρά ασυσχέτιστες προβλέψεις. Επίσης, ο εξαντλητικός έλεγχος, πέρα από

χρονοβόρος, μπορεί να περιορίσει σημαντικά το χώρο των λύσεων εξαιρώντας μοντέλα που στην πράξη είναι προβλεπτικά άρτια. Έτσι, ο ποιοτικός έλεγχος αποτελεί συνήθη πρακτική σε μη αυτοματοποιημένες εφαρμογές προβλέψεων. Ενδεικτικά αναφέρονται τα *scatterplots*, τα *boxplots* και τα *Q-Q plots*.

Scatterplot: Προβάλει τις προβλέψεις του μοντέλου έναντι των πραγματικών παρατηρήσεων. Βάσει αυτού προκύπτει πως αν οι προβλέψεις ταυτίζονται με την πραγματικότητα, τα σημεία του διαγράμματος θα τοποθετούνται στη διαγώνιο *F=Y*. Έτσι, όσο πιο ακριβές είναι το μοντέλο, τόσο περισσότερο προσεγγίζουν τα σημεία του διαγράμματος τη διαγώνιο και αντιστρόφως. Αντίστοιχα, όσο πιο ομοιόμορφα κατανέμονται οι παρατηρήσεις εκατέρωθεν της διαγωνίου, τόσο πιο αντικειμενικές είναι οι προβλέψεις. Η χρήση του scatterplot μπορεί να μας βοηθήσει να εντοπίσουμε εκτός των άλλων και outliers, μεμονωμένες περιπτώσεις δηλαδή που το μοντέλο δεν κατάφερε να αποδώσει ικανοποιητικά τις τιμές τις χρονοσειράς.

Στο ακόλουθο σχήμα, ένα scatterplot χρησιμοποιείται για το διαγνωστικό έλεγχο ενός μοντέλου ARIMA. Όπως παρατηρείται, το μοντέλο τείνει να παράγει αισιόδοξες προβλέψεις καθώς στις περισσότερες από τις περιπτώσεις οι προβλέψεις του είναι μεγαλύτερες από τις πραγματικές, πάνω δηλαδή από τη διαγώνιο που ορίζεται. Επίσης, υπάρχει μία περίπτωση για την οποία το μοντέλο παρήγαγε σημαντικά διαφορετική τιμή από την αναμενόμενη (outlier). Στο συγκεκριμένο παράδειγμα λοιπόν θα έπρεπε ίσως να απορρίψουμε το εν λόγω μοντέλο και να αναζητήσουμε κάποιο μικρότερης προκατάληψης.



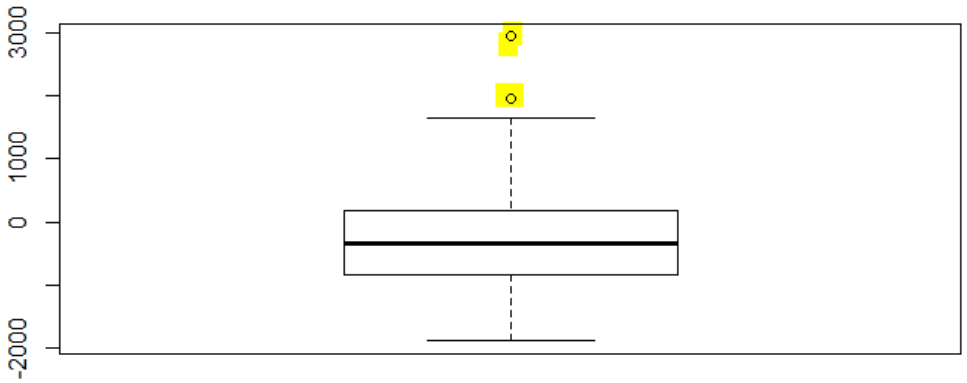
Οπτικός διαγνωστικός έλεγχος μοντέλου με χρήση scatterplot.

Boxplot: Παρουσιάζει την κατανομή των υπολειπόμενων σφαλμάτων με απτό και άμεσα αντιληπτό τρόπο. Τα δύο άκρα του διαγράμματος αναφέρονται στη μέγιστη και την ελάχιστη τιμή σφάλματος, τα περιθώρια του κουτιού που σχηματίζεται στα σημεία εντός

των οποίο τοποθετείται το 25% και το 75% των σφαλμάτων του δείγματος και η τονισμένη τιμή στο ενδιάμεσο σφάλμα. Από τα παραπάνω συμπεραίνουμε τα ακόλουθα:

- Αν ο ενδιάμεσος τοποθετείται στο μηδέν, τότε το μοντέλο είναι μη προκατειλημμένο. Αν ο ενδιάμεσος υπερβαίνει το μηδέν το μοντέλο είναι απαισιόδοξο, ενώ σε αντίθετη περίπτωση αισιόδοξο.
- Όσο περισσότερο προσεγγίζουν τα όρια του διαγράμματος τον ενδιάμεσο, τόσο πιο ακριβείς προβλέψεις παράγει το μοντέλο.
- Όσο λιγότερο ισαπέχουν τα όρια του διαγράμματος, με τόσο πιο ανομοιόμορφο τρόπο παράγονται τα σφάλματα. Το μοντέλο σε αυτήν την περίπτωση θεωρείται μη ευσταθές.

Βασικό πλεονέκτημα του boxplot έναντι του scatterplot είναι πως το πρώτο προσφέρει καλύτερη πληροφορία για τον πώς κατανέμονται τα σφάλματα σκιαγραφώντας τα όρια εντός των οποίων οι διακυμάνσεις μπορούν να θεωρηθούν ανεκτές.

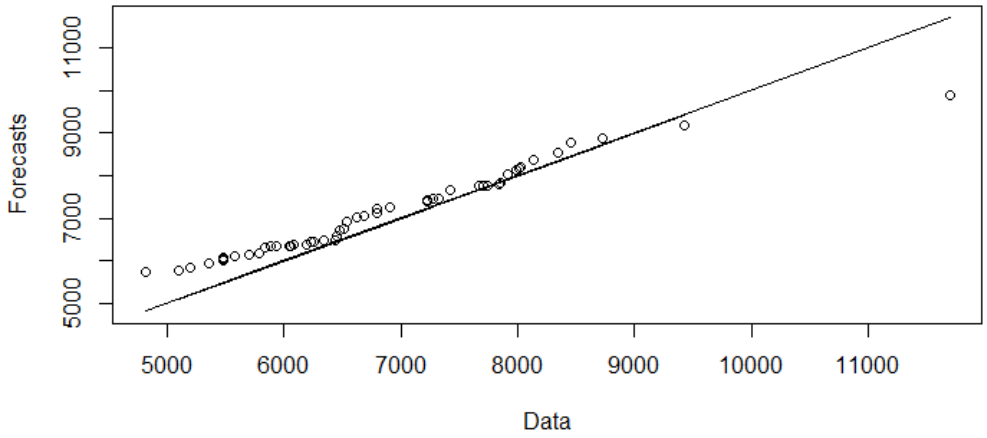


Οπτικός διαγνωστικός έλεγχος μοντέλου με χρήση boxplot.

Στο παραπάνω σχήμα χρησιμοποιείται ένα boxplot για το διαγνωστικό έλεγχο του ίδιου μοντέλου ARIMA που εφαρμόστηκε νωρίτερα στην περίπτωση του scatterplot. Όπως παρατηρείται, ο ενδιάμεσος τοποθετείται κάτω από το μηδέν, οπότε συμπεραίνουμε πως το μοντέλο τείνει να παράγει αισιόδοξες προβλέψεις. Από την άλλη μεριά τα όρια του διαγράμματος ισαπέχουν, κάτι που σηματοδοτεί ευστάθεια. Συνολικά, όπως και στην προηγούμενη περίπτωση, θα έπρεπε ίσως να απορρίψουμε το εν λόγω μοντέλο και να αναζητήσουμε κάποιο μικρότερης προκατάληψης. Σημειώνεται πως το διάγραμμα μας πληροφορεί και για την ύπαρξη δύο outliers τα οποία και σημειώνονται εκτός της κατανομής με τελείες.

Q-Q plot: Προβάλει την κατανομή των προβλέψεων του μοντέλου έναντι της κατανομής των παρατηρήσεων της χρονοσειράς. Το q-q plot προσεγγίζει αρκετά ως λογική αυτήν

του scatterplot έχει όμως μία βασική διαφορά: Τα σημεία του διαγράμματος δεν αναφέρονται στις πραγματικές τιμές του μοντέλου και της χρονοσειράς, αλλά στα τεταρτημόρια (quantiles) που αυτές ορίζουν. Έτσι, αν τα σημεία του διαγράμματος βρίσκονται πάνω στη διαγώνιο *Y=X*, αυτό σημαίνει πως οι προβλέψεις του μοντέλου κατανέμονται ανά διαστήματα με τον ίδιο τρόπο που κατανέμονται και οι παρατηρήσεις της χρονοσειράς. Αν για κάποιο διάστημα τα σημεία αποκλίνουν της διαγωνίου, αυτό αποτελεί ένδειξη πως για το συγκεκριμένο διάστημα το μοντέλο παράγει αισιόδοξες ή απαισιόδοξες προβλέψεις.



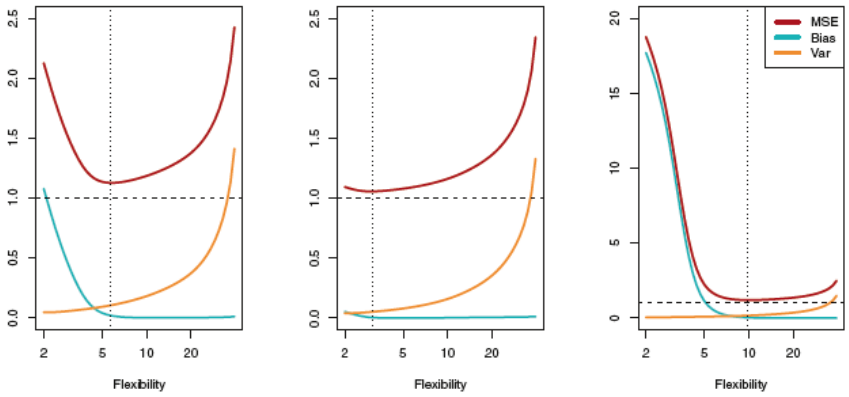
Οπτικός διαγνωστικός έλεγχος μοντέλου με χρήση q-q plot.

Το παραπάνω σχήμα χρησιμοποιείται για το διαγνωστικό έλεγχο του ίδιου μοντέλου ARIMA που εφαρμόστηκε στην περίπτωση των boxplot και scatterplot. Όπως παρατηρείται, ενώ για το διάστημα 6,000-9,500 η κατανομή των προβλέψεων ταυτίζεται με αυτή των παρατηρήσεων της χρονοσειράς, για το διάστημα 4,500-6,000 υπάρχει σημαντική διαφοροποίηση. Έτσι, συμπεραίνουμε πως το συγκεκριμένο μοντέλο δεν ανταποκρίνεται καλά μόνο για τις περιπτώσεις που το επίπεδο της χρονοσειράς είναι περιορισμένο, όπου και είναι αρκετά αισιόδοξο. Η παραπάνω πληροφορία είναι ιδιαίτερα σημαντική καθώς, ενώ μέσω του scatterplot είδαμε νωρίτερα ότι το μοντέλο είναι αισιόδοξο, δεν γνωρίζαμε αν υπάρχει κάποιο συγκεκριμένο διάστημα για το οποίο εντείνεται το εν λόγω φαινόμενο ή αν αυτό γενικεύεται για το σύνολο της κατανομής. Η ανάλυση της επίδοσης ενός μοντέλου ανά διαστήματα είναι υψίστης σημασίας σε αρκετές εφαρμογές καθώς το κόστος μιας λανθασμένης πρόβλεψης δεν είναι για όλα τα διαστήματα το ίδιο. Για παράδειγμα, σε μία εγκατάσταση παραγωγής ενέργειας από φωτοβολταϊκά μπορεί το μέγεθος του σφάλματος να είναι αδιάφορο τις απογευματινές ώρες όπου και η παραγωγή είναι περιορισμένη. Αντίθετα, τις μεσημεριανές ώρες που η παραγωγή είναι αυξημένη και καθορίζει τα έσοδα της εγκατάστασης, η ακρίβεια πρόβλεψης οφείλει να είναι ικανοποιητική.

Άλλα κριτήρια για τη διάγνωση των μοντέλων ARIMA, και συνεπώς και για την επιλογή του καταλληλότερου εξ αυτών, είναι το *Akaike's Information Criterion (AIC)* και το *Bayesian Information Criterion (BIC)*, τα λεγόμενα *information criteria*. Πρόκειται για κριτήρια τα οποία αξιολογούν την απόδοση του υπό εξέταση μοντέλου συναρτήσει της πολυπλοκότητάς του. Φανερώνουν δηλαδή πόσο πιο πολύπλοκο αξίζει να γίνει ένα μοντέλο προκειμένου αυτό να παράγει ακριβέστερες προβλέψεις, όπως αυτό προσδιορίζεται από την προσδοκώμενη πιθανοφάνεια. Η ποσοτικοποίηση του εν λόγω μεγέθους είναι όπως αναφέρθηκε νωρίτερα ιδιαιτέρως χρήσιμη.

Για όλα τα μοντέλα πρόβλεψης ισχύει ότι όσο αυξάνεται η πολυπλοκότητά τους (προστίθενται μεταβλητές *p*, *q*, *P*, *Q*, *c*) τόσο μειώνεται η προκατάληψη των παραγόμενων προβλέψεων (*bias*). Ωστόσο, η αύξηση των μεταβλητών (*flexibility-complexity*) οδηγεί σε αύξηση της διακύμανσης των παραγόμενων σφαλμάτων και συνεπώς σε μικρότερη ακρίβειας πρόβλεψης (*variance*). Το εν λόγω φαινόμενο είναι γνωστό ως *bias-variance trade-off* και κάνει την εμφάνισή του σε όλες τις διαδικασίες που καλούμαστε να επιλέξουμε μέθοδο πρόβλεψης.

Για να καθοριστεί λοιπόν το πλήθος των μεταβλητών του μοντέλου, να αποφευχθεί η υπερ-προσαρμογή (over-fitting) και να εξασφαλιστεί μικρή προκατάληψη και υψηλή ακρίβεια, μπορεί κανείς να χρησιμοποιήσει τα κριτήρια AIC και BIC. Χαρακτηριστικά παραδείγματα του εν λόγω φαινομένου δίνονται στο ακόλουθο σχήμα όπου παρουσιάζεται η μεταβολή του παραγόμενου τετραγωνικού σφάλματος, της διακύμανσης και της προκατάληψης καθώς αυξάνονται οι μεταβλητές ενός μοντέλου ARIMA.



Αλληλεπίδραση προκατάληψης (bias) και ακρίβειας (variance) συναρτήσει της πολυπλοκότητας (flexibility) των μοντέλων.

Όπως παρατηρείται, η βέλτιστη λύση ανά περίπτωση θα ήταν κατά σειρά από τα αριστερά προς τα δεξιά η επιλογή ενός μοντέλου τεσσάρων, δύο και επτά παραγόντων.

Αξίζει να σημειωθεί ότι το MSE δεν μπορεί να μας πληροφορήσει με αξιοπιστία για το ποιο μοντέλο είναι το πλέον κατάλληλο αφού ελαχιστοποιείται για πολυπλοκότητα έξι, τρία και δέκα αντίστοιχα. Δηλαδή, αν αντί να ελαχιστοποιούμε ταυτόχρονα τη διακύμανση και την προκατάληψη του μοντέλου, βελτιστοποιούσαμε την προσαρμογή κατά MSE, θα οδηγούμασταν σε υπερπροσαρμογή.

Το μόνο μειονέκτημα των εν λόγω κριτηρίων είναι πως επειδή δεν έχουν ως βάση τους κάποια συγκεκριμένη υπόθεση ακρίβειας (π.χ. επίτευξη μηδενικού σφάλματος), δεν μας πληροφορούν άμεσα για το αν το μοντέλο που επιλέχθηκε προσαρμόζεται επαρκώς στα δεδομένα παρά μόνο για το ποιο είναι το καλύτερο από αυτά που εξετάστηκαν. Έτσι, η επιλογή του βέλτιστου μοντέλου γίνεται συγκρίνοντας την τιμή του κριτηρίου για όλα τα υποψήφια μοντέλα. Δεδομένου μάλιστα ότι τα παραπάνω κριτήρια εκφράζονται συναρτήσει του λογαρίθμου της πιθανοφάνειας, θεωρούμε βέλτιστο εκείνο το μοντέλο που τα ελαχιστοποιεί.

Akaike's Information Criterion (AIC):

Η τιμή του κριτηρίου υπολογίζεται από τη σχέση

$$AIC = -2logL + 2(p + q + P + Q + k + 1)$$

, όπου *k=0* αν η σταθερά του μοντέλου c ισούται με μηδέν και *k=1* σε αντίθετη περίπτωση.

Αν θέλουμε να δώσουμε μεγαλύτερο βάρος στην πολυπλοκότητα του μοντέλου χρησιμοποιούμε τη διορθωμένη εκδοχή του κριτηρίου:

$$AIC_c = AIC + \frac{2(p + q + P + Q + k + 1)(p + q + P + Q + k + 2)}{n - p - q - P - Q - k - 2}$$

, όπου *n* το μέγεθος του δείγματος.

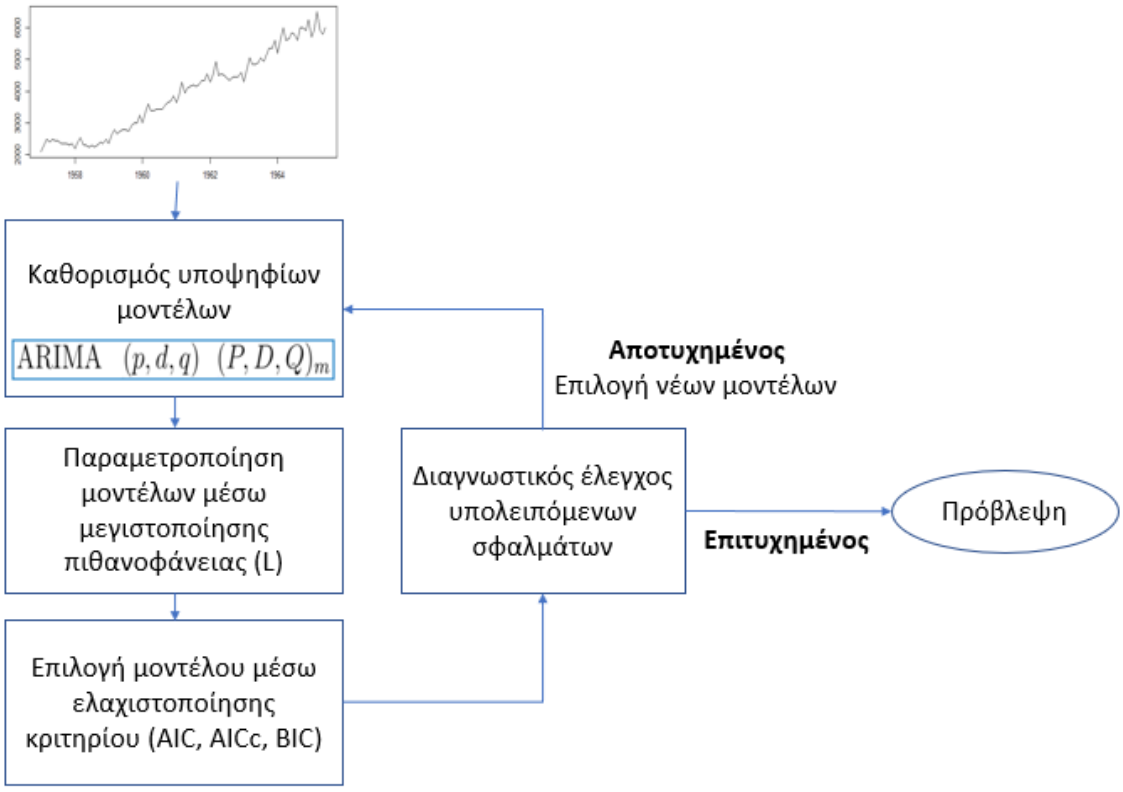
Bayesian Information Criterion (BIC):

Το BIC λειτουργεί σαν το AIC, δίνοντας όμως μεγάλο βάρος στην πολυπλοκότητα του μοντέλου για να αποφευχθεί πιο δραστικά το φαινόμενο της υπερπροσαρμογής.

$$BIC = AIC + (log(n) - 2)(p + q + P + Q + k - 1)$$

Τονίζεται πως οι τύποι των κριτηρίων AIC και BIC δεν λαμβάνουν όπως φαίνεται καθόλου υπόψη τους παράγοντες των διαφορίσεων *d* και *D*, όπως άλλωστε συμβαίνει και στην περίπτωση της πιθανοφάνειας. Έτσι, τα κριτήρια αυτά απαντούν στην ερώτηση ποιο μοντέλο ARIMA θα χρησιμοποιηθεί δεδομένης εποχιακής ή μη διαφόρισης.

Έχοντας λοιπόν κανείς διαθέσιμο στα χέρια του ένα κριτήριο αξιολόγησης μοντέλων (AIC, AICc, BIC) και ένα κριτήριο βελτιστοποίησης παραμέτρων (L), μπορεί πλέον να επιλέξει και να εκτιμήσει το βέλτιστο μοντέλο ARIMA για κάθε χρονοσειρά. Αρχικά, κάθε ένα από τα υποψήφια μοντέλα παραμετροποιείται βέλτιστα μέσω της μεγιστοποίησης της πιθανοφάνειας. Στη συνέχεια για κάθε ένα από αυτά υπολογίζεται η τιμή ενός εκ των AIC, AICc και BIC και επιλέγεται εκείνο που τα ελαχιστοποιεί. Στη συνέχεια οπτικοποιείται η εν λόγω διαδικασία η οποία και συνοψίζει τη μεθοδολογία αναγνώρισης, εκτίμησης και διάγνωσης.



Μεθοδολογία αναγνώρισης, εκτίμησης και διάγνωσης μοντέλων ARIMA

7. Πρόβλεψη με μοντέλα ARIMA

Όταν θέλουμε να υπολογίσουμε μέσω ενός μοντέλου ARIMA την τιμή της χρονοσειράς *y* την περίοδο *t*, τότε απαιτείται γνώση των τιμών y_{t-1}, y_{t-2} … y_{t-p} ή/και των τιμών e_{t-1}, e_{t-2} … e_{t-q}. Για την πρόβλεψη λοιπόν της τιμής $\widehat{y_{t+T}}$, απαιτείται αντίστοιχα γνώση των τιμών y_{t+T-1},y_{t+T-2}…y_{t+T-n} ή/και των τιμών e_{t+T-1},e_{t+T-2}…e_{t+T-n}, όπου *T* ο ορίζοντας πρόβλεψης. Για ένα μοντέλο ARMA(1,1) π.χ. αν θελήσουμε να προβλέψουμε 3 περιόδους μπροστά απαιτείται γνώση των τιμών y_t και e_t για τη πρώτη πρόβλεψη, των τιμών y_{t+1} και e_{t+1} για τη δεύτερη και των τιμών y_{t+2} και e_{t+2} για τη τρίτη.

Εδώ βλέπουμε ότι εμφανίζεται αμέσως ένα πρόβλημα: Οι τιμές y_t και e_t που πρέπει να χρησιμοποιηθούν για τη πρώτη πρόβλεψη είναι διαθέσιμες. Δεν ισχύει όμως το ίδιο και για τις y_{t+1} και e_{t+1} που απαιτούνται στην πρόβλεψη του δεύτερου διαστήματος. Για να γίνει λοιπόν πρόβλεψη θα πρέπει να θεωρήσουμε ως y_{t+1} την νωρίτερα εκτιμημένη τιμή της χρονοσειράς $\widehat{y_{t+1}}$ από το μοντέλο και μηδενικό σφάλμα. Αντίστοιχα για την τρίτη πρόβλεψη θεωρούμαι $y_{t+2}= \widehat{y_{t+2}}$ και μηδενικό σφάλμα. Μακροχρόνια το μοντέλο ARIMA εκφυλίζεται δηλαδή σε μοντέλο AR εξαρτώμενο μόνο από τις προβλέψεις που έχει κάνει το ίδιο και όχι από τις πραγματικές τιμές των δεδομένων. Αυτός είναι και ο λόγος που χρησιμοποιείται κυρίως για βραχυπρόθεσμες προβλέψεις.

Δεδομένου ότι η κατανομή των παρατηρήσεων σε μία απολύτως στάσιμη χρονοσειρά είναι παντού η ίδια, η πρόβλεψη μέσω ενός μοντέλου ARIMA μπορεί να γίνει και από το τέλος προς την αρχή με παρόμοια αποτελέσματα. Μία τέτοια διαδικασία (back-casting) φαντάζει αρχικά ανούσια, ωστόσο μπορεί να φανεί ιδιαίτερα χρήσιμη στην αρχική προσαρμογή του μοντέλου. Όπως είπαμε νωρίτερα, για τον υπολογισμό της τιμής y_t απαιτείται γνώση των τιμών y_{t-1}, y_{t-2} … y_{t-p} ή/και των τιμών e_{t-1}, e_{t-2} … e_{t-q}. Αυτό σημαίνει πως το αρχικά υπολογιζόμενο μοντέλο θα έχει *p* ή *q* κενές τιμές, αφού δεν υπάρχουν νωρίτερα δεδομένα για τον υπολογισμό τους. Εφαρμόζοντας την τεχνική του back-casting με ορίζοντα *p* ή *q* αντίστοιχα παρέχουμε στο μοντέλο τις απαιτούμενες αρχικές τιμές και το απαλλάσσουμε από τις μηδενικές. Συχνά, και όταν δεν ελέγχεται με αυτόν τον τρόπο η αποτελεσματικότητα του μοντέλου, οι αρχικές κενές τιμές αντικαθίστανται απλώς από τις αντίστοιχες της χρονοσειράς.

8. Επίλογος

Τα μοντέλα ARIMA αποτελούν ένα εξαιρετικό εργαλείο για την κατανόηση της εξέλιξης μεγεθών ενδιαφέροντος, αναλύοντάς και προεκτείνοντάς τα στο μέλλον. Αντιμετωπίζουν τις χρονοσειρές με μία στοχαστική ματιά βασιζόμενα αποκλειστικά στην κατανομή των παρελθοντικών τιμών και μάλιστα των πλέον πρόσφατων. Αυτό τα καθιστά αποτελεσματικά και ευέλικτα, κυρίως για την παραγωγή βραχυπρόθεσμων προβλέψεων.

Καθώς ο μηχανισμός εξέλιξης που περιγράφεται από κάθε μοντέλο ARIMA μπορεί να οδηγήσει μόλις σε μερική αναπαράσταση της πραγματικότητας, αυτό που πρέπει να επιζητούμε κατά την αναγνώριση, εκτίμηση και εφαρμογή τους είναι ο εντοπισμός εκείνου του μοντέλου που περιγράφει αποτελεσματικά τα αρχικά δεδομένα χωρίς να εισάγει περιττή πολυπλοκότητα στη διαδικασία πρόβλεψης. Η φειδωλότητα σε ένα μοντέλο ARIMA, δηλαδή η επαρκής περιγραφή της αρχικής χρονοσειράς χωρίς τη χρήση πολλαπλών μεταβλητών, είναι χαρακτηριστικό ζωτικής σημασίας και μπορεί να

εξασφαλιστεί χρησιμοποιώντας ένα σύνολο από κατάλληλες τεχνικές, κριτήρια και ελέγχους.

Σημαντικός παράγοντας για τη βελτίωση της αποτελεσματικότητας και της ακρίβειας των μοντέλων ARMA είναι η ύπαρξη στασιμότητας. Η στασιμότητα είναι εκείνη που εξασφαλίζει βελτιωμένη προσαρμογή και με την προϋπόθεσή της εκτιμώνται συντελεστές υψηλής ποιότητας με στατιστική σημαντικότητα. Έτσι, η ανάλυση των αρχικών δεδομένων, τόσο στατιστικά όσο και οπτικά, και η εφαρμογή διορθωτικών κινήσεων για τη σταθεροποίηση της διακύμανσής τους, είναι διαδικασίες καθοριστικής σημασίας.

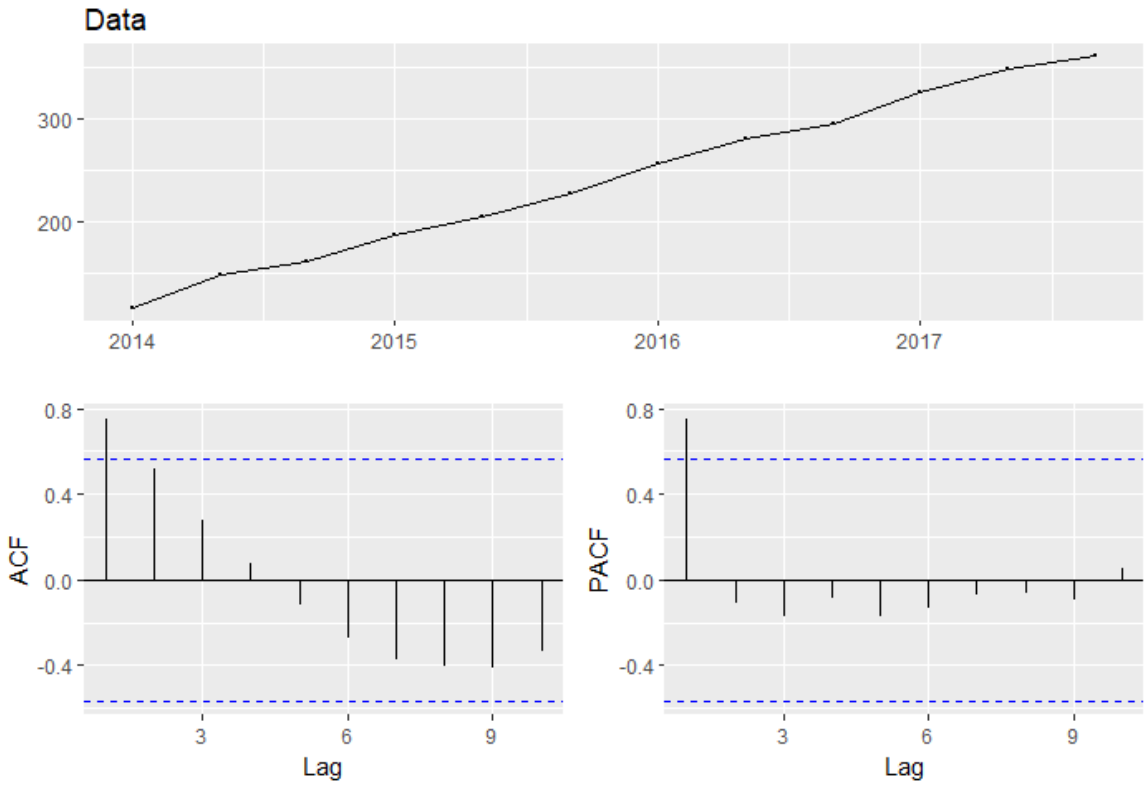
9. Εφαρμογή

Δίνεται η τετραμηνιαία χρονοσειρά Y_t.

2014 Q1	2014 Q2	2014 Q3	2015 Q1	2015 Q2	2015 Q3	2016 Q1	2016 Q2	2016 Q3	2017 Q1	2017 Q2	2017 Q3
116	149	161	187	205	228	256	281	295	326	348	361

Ζητείται να παραχθούν προβλέψεις για το επόμενο έτος χρησιμοποιώντας κατάλληλο μοντέλο ARIMA.

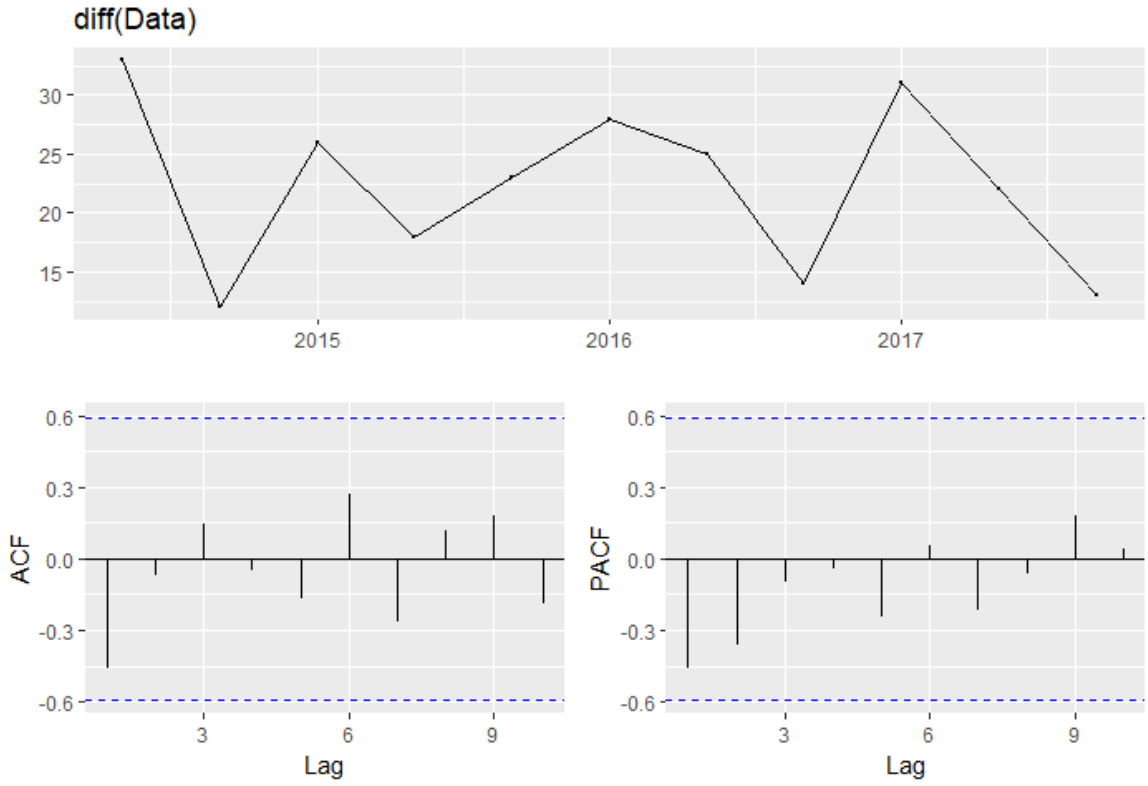
Για να αποφασίσουμε σχετικά με το ποιο μοντέλο ARIMA περιγράφει βέλτιστα τη χρονοσειρά Y_t, αρχικά απεικονίζουμε τις παρατηρήσεις της ούτως ώστε να αποφανθούμε σχετικά με την ύπαρξη ή όχι στασιμότητας.



Όπως παρατηρείται, η χρονοσειρά δεν παρουσιάζει εποχιακή συμπεριφορά, ωστόσο χαρακτηρίζεται από έντονη τάση. Αυτό επιβεβαιώνεται και από το διάγραμμα αυτοσυσχέτισης, καθώς οι τιμές ACF δεν μηδενίζουν απότομα, αλλά αντίθετα φθίνουν αργά με την πάροδο του χρόνου. Η εν λόγω συμπεριφορά αποτελεί ένδειξη πως η χρονοσειρά που μελετάται δεν είναι στάσιμη και πως απαιτείται διαφόριση προκειμένου

να αποκτήσει σταθερή διακύμανση. Δεδομένου μάλιστα ότι ο δείκτης μερικής αυτοσυσχέτισης για υστέρηση *k=1* είναι στατιστικά σημαντικός, η εν λόγω χρονοσειρά είναι πολύ πιθανό να περιγράφεται ικανοποιητικά από ένα μοντέλο ARIMA(1,1,0)

Πράγματι, μετά την εφαρμογή διαφόρισης 1ης τάξης, η χρονοσειρά που προκύπτει φαίνεται πως είναι επαρκώς στάσιμη καθώς κανείς από τους συντελεστές ACF και PACF δεν είναι πλέον στατιστικά σημαντικός. Επιπλέον, εφαρμόζοντας κανείς το τεστ στασιμότητας KPSS, λαμβάνει για την αρχική και τη διαφορισμένη χρονοσειρά τιμές test-statistic 0.51 και 0.22 αντίστοιχα. Για διάστημα εμπιστοσύνης 95%, η κρίσιμη τιμή t ισούται με 0.46, η οποία είναι μικρότερη του 0.51 και μεγαλύτερη του 0.21. Έτσι, συμπεραίνουμε πως η αρχική χρονοσειρά δεν ήταν στάσιμη, αλλά έγινε επαρκώς μετά την εφαρμογή της διαφόρισης.



Έχοντας εξασφαλίσει στασιμότητα προχωράμε στον υπολογισμό της χρονοσειράς των πρώτων διαφορών y_t και στην εφαρμογή επί αυτής ενός μοντέλου AR(1). Για να γίνει αυτό χρειάζεται να υπολογιστεί φυσικά η τιμή της αυτοσυσχέτισης για *k=1*, δηλαδή ο συντελεστής φ_1, ως εξής:

$$\varphi_1 = p_1 = \frac{\sum_{t=2}^{11}(y_t - \mu)(y_{t-1} - \mu)}{\sum_{t=1}^{11}(y_t - \mu)^2} = \frac{-242.35}{524.18} = -0.462$$

, όπου *μ=22.27* η μέση τιμή των πρώτων διαφορών.

t	Y_t	diff(Y_t)= y_t	y_t-μ	$(y_t-\mu)^2$	$(y_t-\mu)(y_{t-1}-\mu)$
2014-Q1	116				
2014-Q2	149	33	10.73	115.13	
2014-Q3	161	12	-10.27	105.47	-110.20
2015-Q1	187	26	3.73	13.91	-38.31
2015-Q2	205	18	-4.27	18.23	-15.93
2015-Q3	228	23	0.73	0.53	-3.12
2016-Q1	256	28	5.73	32.83	4.18
2016-Q2	281	25	2.73	7.45	15.64
2016-Q3	295	14	-8.27	68.39	-22.58
2017-Q1	326	31	8.73	76.21	-72.20
2017-Q2	348	22	-0.27	0.07	-2.36
2017-Q3	361	13	-9.27	85.93	2.50
Σύνολο				**524.18**	**-242.35**

Σειρά έχει η εφαρμογή του αντίστοιχου μοντέλου ARIMA με σταθερά (*n=1, m=3, N=0, ϑ_1 = ϑ_2= …= ϑ_q = 0, φ_1 = -0.462, φ_2 = φ_3 = … = φ_p = 0*)

$$\left(1-\varphi_1 B-\varphi_2 B^2-\cdots-\varphi_p B^p\right)(1-B)^n(1-B^m)^N\overline{Y_t} = \left(\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q\right)e_t \rightarrow$$

$$(1-\varphi_1 B)(1-B)\overline{Y_t} = 0 \rightarrow$$

$$\overline{Y_t} = (1+\varphi_1)\overline{Y_{t-1}} - \varphi_1\overline{Y_{t-2}} \rightarrow$$

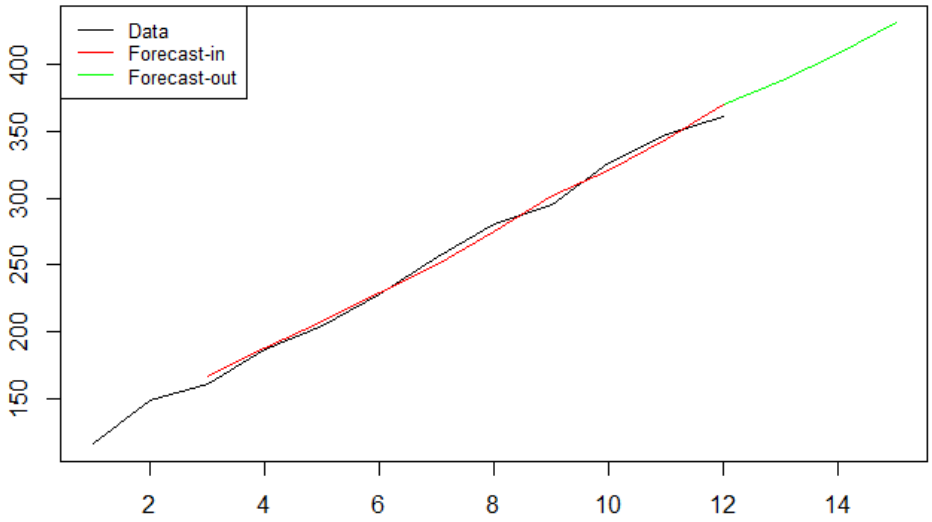
$$Y_t = (1+\varphi_1)Y_{t-1} - \varphi_1 Y_{t-2} + c$$

, όπου $c = \mu(1-\varphi_1) = 22.27(1+0.462) = 32.56$

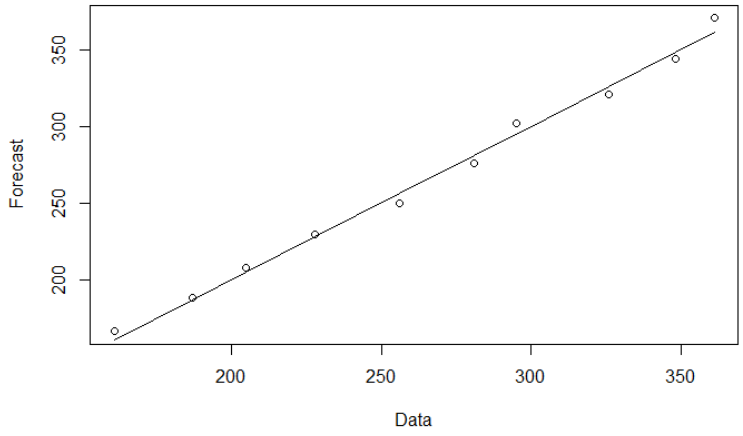
t	#	Y_t	ARIMA(1,1,0)	e_t
2014-Q1	1	116		
2014-Q2	2	149		
2014-Q3	3	161	166.31	-5.31
2015-Q1	4	187	188.02	-1.02
2015-Q2	5	205	207.55	-2.55
2015-Q3	6	228	229.24	-1.24
2016-Q1	7	256	249.93	6.07
2016-Q2	8	281	275.62	5.38
2016-Q3	9	295	302.01	-7.01
2017-Q1	10	326	321.09	4.91

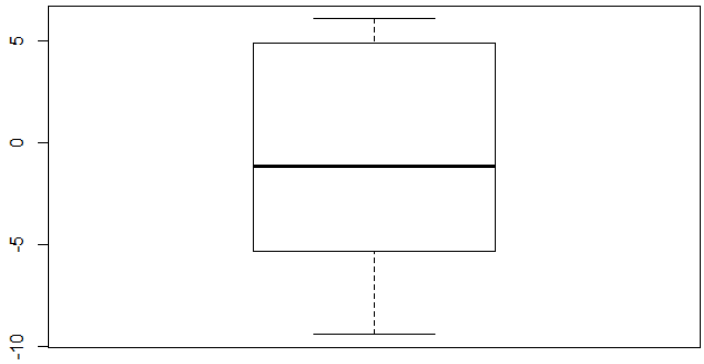
2017-Q2	11	348	344.24	3.76
2017-Q3	12	361	370.40	-9.40
2017-Q1	13		387.55	
2017-Q2	14		407.85	
2017-Q3	15		431.03	

Παρατηρείστε ότι για την παραγωγή προβλέψεων στις περιόδους (2017-Q2) και (2017-Q3) αξιοποιούνται λόγω έλλειψης πραγματικών δεδομένων (Y_{13} και Y_{14}) οι προηγούμενες προβλέψεις του μοντέλου ARIMA, δηλαδή οι τιμές Y_{12} και F_{13} για την παραγωγή της πρόβλεψης F_{14} και οι τιμές οι τιμές F_{13} και F_{14} για την παραγωγή της πρόβλεψης F_{15}. Το ακόλουθο διάγραμμα παρουσιάζει τις in-sample και out-of-sample προβλέψεις του μοντέλου που κατασκευάστηκε.

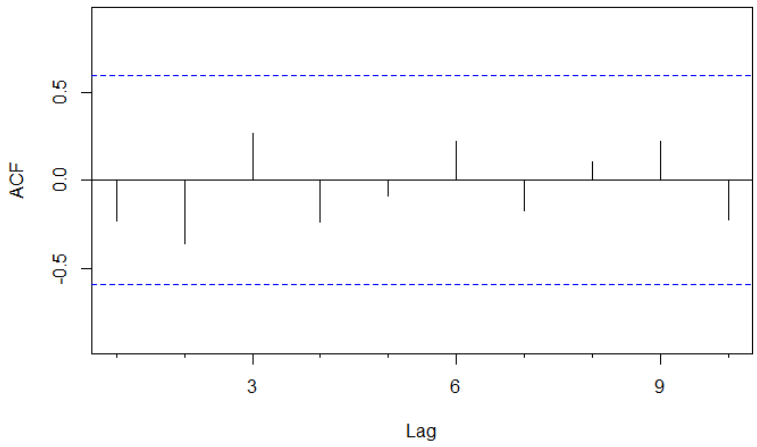


Προβλεπτικά, το μοντέλο φαίνεται πως είναι άρτιο καθώς το μέσο σφάλμα πρόβλεψης in-sample είναι ιδιαίτερα μικρό (ME=-0.64), δηλαδή δεν υπάρχει προκατάληψη. Αυτό μπορεί να επιβεβαιωθεί και οπτικά σχεδιάζοντας τα αντίστοιχα διαγράμματα *scatterplot* και *boxplot*. Όπως φαίνεται, στην πρώτη περίπτωση οι τιμές του διαγράμματος βρίσκονται πολύ κοντά στη διαγώνιο, ενώ στη δεύτερη τα σφάλματα κατανέμονται συμμετρικά γύρω από το μηδέν.





Επίσης, σχεδιάζοντας το διάγραμμα ACF των σφαλμάτων επιβεβαιώνεται πως τα παραγόμενα σφάλματα δεν παρουσιάζουν πρότυπα, δηλαδή είναι ανεξάρτητα μεταξύ τους και ισοδύναμα με το λευκό θόρυβο.



Τέλος, υπολογίζεται η τιμή πιθανοφάνειας του μοντέλου

$$-2logL = n\log(2\pi) + n\log(\sigma^2) + \frac{\sum_{t=1}^{n} {e_t}^2}{\sigma^2} = 69.19$$

η οποία αντιστοιχεί σε τιμή AIC

$$AIC = -2logL + 2(p + q + P + Q + k + 1) =$$

$$69.19 + 2 * (1 + 0 + 0 + 0 + 1 + 1) = 75.19$$

Δεδομένου ότι το μοντέλο που αναγνωρίστηκε, εκτιμήθηκε και διαγνώστηκε είναι προβλεπτικά άρτιο, θα μπορούσε κανείς να τερματίσει την όλη διαδικασία και να αξιοποιήσει τις προβλέψεις που παρήγαγε. Ωστόσο, προκειμένου να εξακριβωθεί περαιτέρω η χρησιμότητα του μοντέλου ARIMA(1,1,0) έναντι άλλων εναλλακτικών, μπορεί π.χ. κανείς να εκτιμήσει επιπρόσθετα το μοντέλο ARIMA(0,1,1) και να συγκρίνει την μεταξύ τους απόδοση.

Για το μοντέλο MA(1) ισχύει βάσει πρότερων υπολογισμών ότι $\rho_1 = -0.462$, και επιπλέον ότι

$$\rho_1 = \frac{-\theta_1}{1 + {\theta_1}^2}$$

Η παραπάνω εξίσωση έχει λύσεις τις τιμές θ_1=0.669 και θ_1=1.496. Ωστόσο βάσει γνωστών περιορισμών ισχύει ότι *-1<ϑ_1<1*, οπότε και θ_1=0.669.

Έτσι, μπορούμε πλέον να εφαρμόσουμε το αντίστοιχο μοντέλο ARIMA με σταθερά (*n=1, m=3, N=0, $\varphi_1 = \varphi_2 = \ldots = \varphi_q = 0$, $\vartheta_1 = 0.669$, $\vartheta_2 = \vartheta_3 = \ldots = \vartheta_p = 0$*)

$$\left(1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p\right)(1 - B)^n (1 - B^m)^N \overline{Y_t} = \left(\theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right) e_t \rightarrow$$

$$(1 - B)\overline{Y_t} = \theta_1 B e_t \rightarrow$$

$$\overline{Y_t} = \overline{Y_{t-1}} + \theta_1 e_{t-1} \rightarrow$$

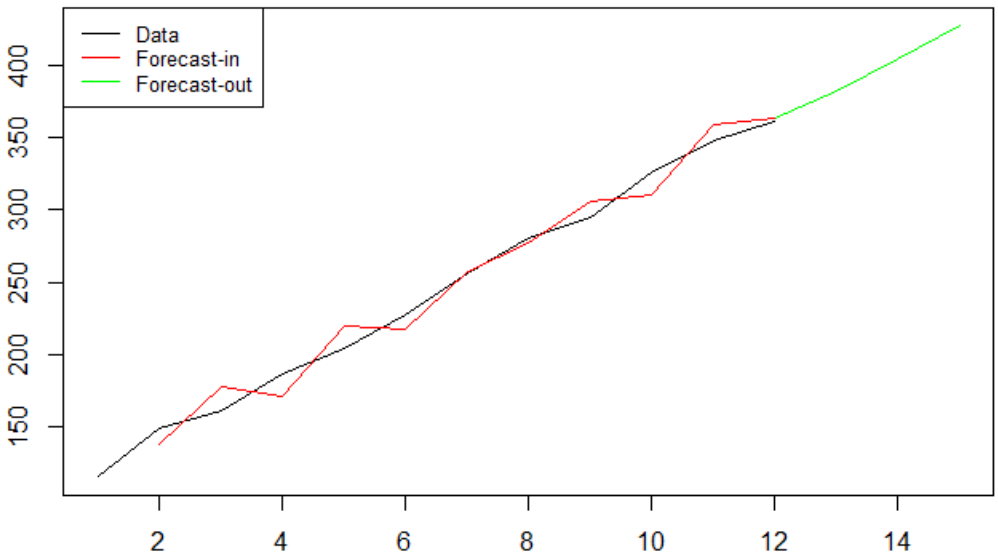
$$Y_t = Y_{t-1} + \theta_1 e_{t-1} + c$$

, όπου $c = \mu = 22.27$

t	#	Y_t	ARIMA(0,1,1)	e_t
2014-Q1	1	116		
2014-Q2	2	149	138.27	10.73
2014-Q3	3	161	178.45	-17.45
2015-Q1	4	187	171.60	15.40

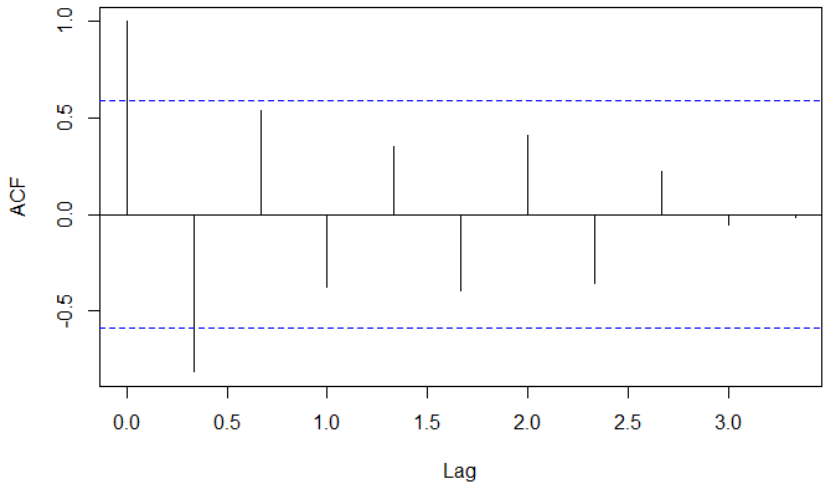
2015-Q2	5	205	219.57	-14.57
2015-Q3	6	228	217.52	10.48
2016-Q1	7	256	257.28	-1.28
2016-Q2	8	281	277.41	3.59
2016-Q3	9	295	305.67	-10.67
2017-Q1	10	326	310.13	15.87
2017-Q2	11	348	358.89	-10.89
2017-Q3	12	361	362.99	-1.99
2017-Q1	13		381.94	
2017-Q2	14		404.21	
2017-Q3	15		426.48	

Παρατηρείστε ότι για την παραγωγή προβλέψεων στις περιόδους (2017-Q2) και (2017-Q3) θεωρείται λόγω έλλειψης πραγματικών δεδομένων ότι τα αντίστοιχα σφάλματα (e_{13} και e_{14}) είναι μηδενικά. Έτσι, οι προβλέψεις F_{14} και F_{15} είναι ίδιες με αυτές της περιόδου (2017-Q1), προσαυξημένες κατά τη σταθερά c. Στο ακόλουθο διάγραμμα παρουσιάζονται οι in-sample και out-of-sample προβλέψεις του μοντέλου ARIMA(0,1,1) που κατασκευάστηκε.



Και σε αυτήν την περίπτωση το μοντέλο φαίνεται αρχικά πως είναι προβλεπτικά άρτιο καθώς το μέσο σφάλμα πρόβλεψης in-sample είναι ιδιαίτερα μικρό (ME=-0.07). Ωστόσο, αναλύοντας τα σφάλματά του μέσω του διαγράμματος ACF παρατηρεί κανείς εύκολα πως για υστέρηση k=1 και k=2 υπάρχουν στατιστικά σημαντικές συσχετίσεις μεταξύ των σφαλμάτων. Συνεπώς το μοντέλο ARIMA(0,1,1) αντιμετωπίζει πρόβλημα προσαρμογής

καθώς, σε αντίθεση με το μοντέλο ARIMA(1,1,0) δεν αποτυπώνει τη συσχέτιση που υπάρχει μεταξύ διαδοχικών παρατηρήσεων.



Για να αποφανθούμε μάλιστα με δομημένο τρόπο σχετικά με την απόδοσή του τελευταίου μοντέλου, υπολογίζουμε επιπλέον τη τιμή πιθανοφάνειάς του

$$-2logL = n\log(2\pi) + n\log(\sigma^2) + \frac{\sum_{t=1}^{n} {e_t}^2}{\sigma^2} = 85.20$$

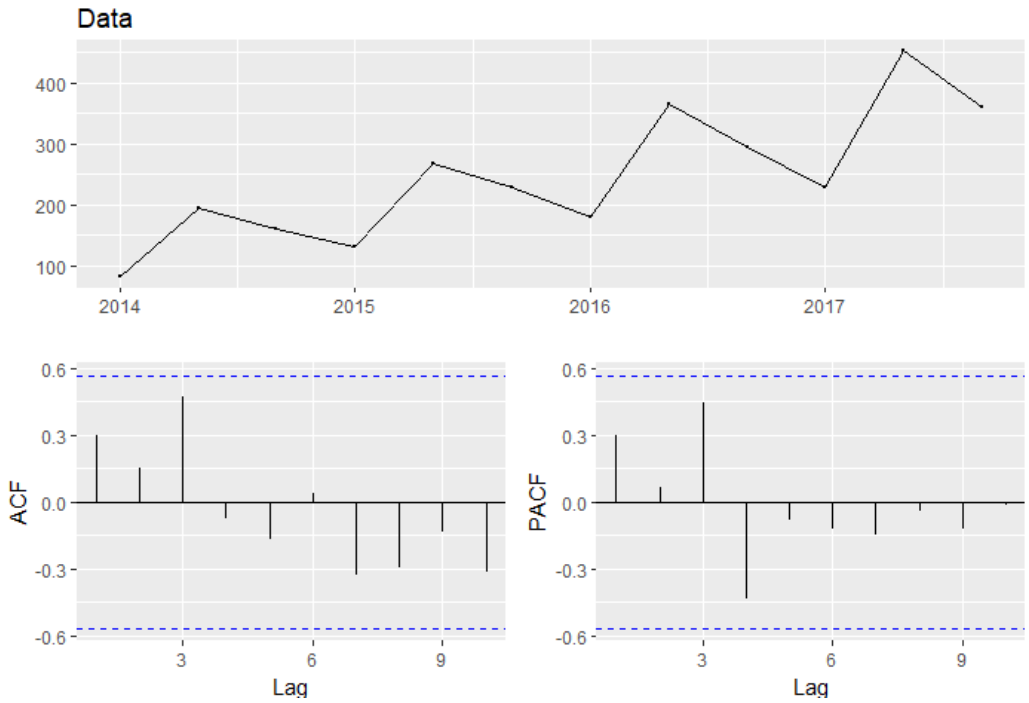
η οποία αντιστοιχεί σε τιμή AIC

$$AIC = -2logL + 2(p + q + P + Q + k + 1) =$$

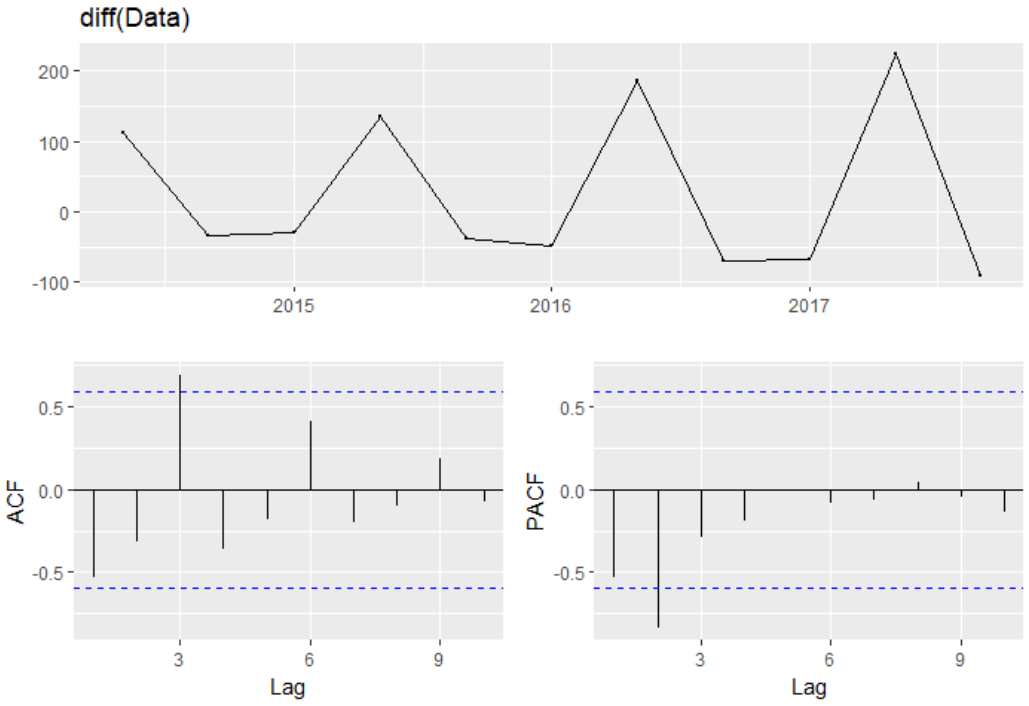
$$85.20 + 2 * (1 + 0 + 0 + 0 + 1 + 1) = 91.20$$

Παρατηρούμε πως το πρώτο μοντέλο έχει μικρότερη τιμή AIC από το δεύτερο, οπότε και αποκαλύπτεται ότι το μοντέλο ARIMA(1,1,0) προσαρμόζεται καλύτερα στα δεδομένα της χρονοσειράς. Μάλιστα, δεδομένου ότι η πολυπλοκότητα των δύο μοντέλων είναι ίδια, μπορούμε να συγκρίνουμε απευθείας τις τιμές πιθανοφάνειάς τους, εξάγοντας το ίδιο συμπέρασμα. Τονίζεται πως η τελευταία σύγκριση δεν θα επιτρεπόταν αν τα μοντέλα διέφεραν ως προς την πολυπλοκότητά τους, αν π.χ. η σύγκριση γινόταν με ένα από τα μοντέλα ARIMA(2,1,0), ARIMA(1,1,1) ή ARIMA(1,1,0) χωρίς τη χρήση σταθεράς. Για παράδειγμα, θα μπορούσε κανείς να εξετάσει αν τα ίδια μοντέλα ARIMA(1,1,0) και ARIMA(0,1,1) αποδίδουν καλύτερα χωρίς τη χρήση σταθεράς και βάσει του διαγνωστικού ελέγχου να προχωρούσε με την καλύτερη δυνατή λύση.

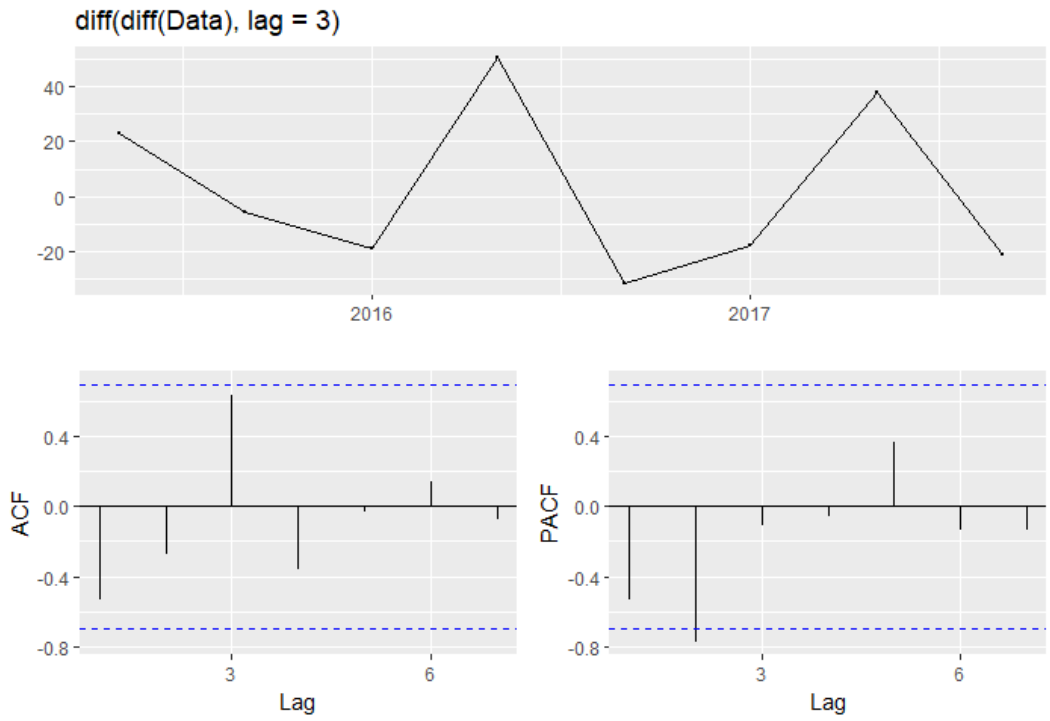
Θεωρήστε τώρα πως η αρχική χρονοσειρά ήταν εποχιακή με δείκτες εποχιακότητας 0.7, 1.3 και 1.0 αντίστοιχα. Η νέα χρονοσειρά που προκύπτει διαθέτει πλέον συστηματικές διακυμάνσεις εξαιτίας της ταυτόχρονης ύπαρξης εποχιακότητας και τάσης.



Το παραπάνω συμπέρασμα επιβεβαιώνεται οπτικά και από τα διαγράμματα ACF και PACF. Συγκεκριμένα, οι τιμές ACF δεν μηδενίζουν απότομα, αλλά αντίθετα φθίνουν με την πάροδο του χρόνου (τάση). Επιπλέον, για k=3 υπάρχει σημαντική συσχέτιση μεταξύ των παρατηρήσεων (εποχιακότητα). Έτσι, η εφαρμογή απλής διαφόρισης δεν αποτελεί λύση καθώς η χρονοσειρά που θα προκύψει θα συνεχίσει να είναι εποχιακή.



Αντίθετα, αν εφαρμοστεί επιπρόσθετα εποχιακή διαφόριση, τότε η χρονοσειρά που θα προκύψει θα είναι σχετικά στάσιμη και προβλέψιμη μέσω ενός μοντέλου ARMA.



Προβλεπτικά, το αποτέλεσμα σε αυτήν την περίπτωση αν δεν προχωρήσουμε σε διπλή εποχιακή διαφόριση θα είναι καταστροφικό και παρουσιάζεται παρακάτω. Όπως φαίνεται, τα μοντέλα χωρίς τη βοήθεια της διαφόρισης είναι αδύνατο να ακολουθήσουν την εποχιακότητα της χρονοσειράς και οδηγούνται συνεπώς σε σημαντικά και συστηματικά λάθη.