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

Αυτοπαλινδρομικά Μοντέλα Κινητού Μέσου Όρου (ARIMA)

Σημειώσεις για το μάθημα 'Τεχνικές Προβλέψεων'

ΣΗΜΜΥ – 8^{ο} εξάμηνο

Περιεχόμενα

Εισαγωγή

Τα ολοκληρωμένα αυτοπαλινδρομικά μοντέλα κινητών μέσων όρων ARIMA (Auto Regressive Integrated Moving Average) είναι **στοχαστικά** μαθηματικά μοντέλα τα οποία μας βοηθάνε να αναλύσουμε και να προβλέψουμε την εξέλιξη μεγεθών. Σε αντίθεση με τα ντετερμινιστικά μοντέλα (βλ. γραμμική παλινδρόμηση), η χρήση των οποίων απαιτεί γνώση των παραγόντων από τις οποίες εξαρτάται το μέγεθος και όπου ο πλήρης εντοπισμός, μέτρηση και πρόβλεψή τους είναι πρακτικά αδύνατος, η εφαρμογή των μοντέλων ARIMA βασίζεται στον υπολογισμό της πιθανότητας για την οποία η τιμή του μεγέθους βρίσκεται εντός κάποιου διαστήματος. Τα μοντέλα ARIMA μελετήθηκαν εκτεταμένα από τους Box και Jenkins και συχνά αναφέρονται στη βιβλιογραφία με την ομώνυμη ονομασία.

Στην γενική τους μορφή τα μοντέλα ARIMA περιέχουν τον τυχαίο παράγοντα (σφάλμα πρόβλεψης), τιμές του μεγέθους που εμφανίστηκαν σε προηγούμενες περιόδους και σχετικούς στοχαστικούς παράγοντες. 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. Στην πράξη βέβαια δεν μπορούμε να είμαστε ποτέ σίγουροι για το ποιος είναι ο βέλτιστος γραμμικός συνδυασμός. Μπορούμε ωστόσο να τον προσεγγίσουμε ικανοποιητικά.

Η εφαρμογή των μοντέλων ARIMA προϋποθέτει να πληρούνται ορισμένες απαιτήσεις. Αρχικά, η χρονοσειρά πρέπει να είναι **διακριτή**, δηλαδή οι παρατηρήσεις της y_t να αναφέρονται σε ισαπέχουσες χρονικές στιγμές $y_t, y_{t+\tau}, y_{t+2\tau}$…. Εκτός αυτού, η χρονοσειρά πρέπει να είναι **στάσιμη**. Αυτό σημαίνει πως η μέση τιμή (μ), η διακύμανση (σ^2) και η συνάρτηση αυτοσυσχέτισης (ACF) της χρονοσειράς πρέπει να είναι σταθερές καθ' όλη τη διάρκεια του χρόνου. Έτσι, τα χαρακτηριστικά της δεν εξαρτώνται από τη χρονική στιγμή την οποία αυτή εξετάζεται (βλέπε λευκός θόρυβος) και οποιοδήποτε δείγμα της κατανομής της y_{t1}, y_{t2}…y_{tn} ταυτίζεται εν γένει με αυτό της $y_{t1+\tau}, y_{t2+\tau}$…$y_{tn+\tau}$. Με αυτό τον τρόπο η χρονοσειρά αποδεσμεύεται από την έννοια του χρόνου και μπορεί να μελετηθεί στοχαστικά. Τέλος, η εφαρμογή μοντέλων ARIMA προϋποθέτει την εξαγωγή **βραχυπρόθεσμων προβλέψεων**. Όπως αναφέρθηκε, τα συγκεκριμένα μοντέλα είναι γραμμικός συνδυασμός των παρελθοντικών τιμών της χρονοσειράς. Αυτό σημαίνει ότι αν θελήσουμε να προβλέψουμε την τιμή y_t, τότε απαιτείται γνώση των τιμών y_{t-1}, y_{t-2}…y_{t-n}. Για την πρόβλεψη της τιμής y_{t+1}, απαιτείται αντίστοιχα γνώση των τιμών y_t, y_{t-1}…y_{t-n-1}. Δεδομένου ωστόσο ότι η τιμή y_t δεν είναι διαθέσιμη από τα δεδομένα αλλά υπολογίστηκε νωρίτερα από το μοντέλο, αντιλαμβανόμαστε ότι είναι και λιγότερο αξιόπιστη. Συνεπώς για μακροπρόθεσμες προβλέψεις (όπου η τιμή τους θα εξαρτάται σχεδόν αποκλειστικά από τιμές που δεν ανήκουν στα δεδομένα καθ' αυτά) η αξιοπιστία θα είναι σημαντικά μικρότερη.

Επιλογή και εκτίμηση ενός μοντέλου ARIMA

Η επιλογή του καταλληλότερου μοντέλου ARIMA για μία χρονοσειρά δεν είναι πάντα προφανής. Συχνά περισσότερα από ένα μοντέλα μπορούν σχεδόν να ταυτίζονται, αφήνοντας την επιλογή στην κρίση μας. Επίσης, μπορεί κάποιο μοντέλο να προσαρμόζεται καλύτερα από ένα άλλο σε μία χρονοσειρά αλλά η πολυπλοκότητά του να είναι σημαντικά μεγαλύτερη και για αυτό το λόγο να απορριφτεί. Άλλωστε, μία καλύτερη προσαρμογή δεν μπορεί ποτέ να εγγυηθεί την επίτευξη μικρότερου σφάλματος πρόβλεψης.

Θέλοντας ωστόσο να αυτοματοποιήσουμε με κάποιον τρόπο τη διαδικασία εύρεσης του βέλτιστου μοντέλου ARIMA, ακολουθούμε την παρακάτω διαδικασία που περιλαμβάνει τρία στάδια: την αναγνώριση, την εκτίμηση και την διάγνωση.

1. Στο στάδιο της **αναγνώρισης** επιλέγονται ένα ή περισσότερα μοντέλα ARIMA τα οποία θεωρούμε βάση κάποιων στοιχείων (όπως οι γραφικές παραστάσεις της αυτοσυσχέτισης και της μερικής αυτοσυσχέτισης) ότι μπορούν να περιγράψουν ικανοποιητικά την χρονοσειρά. Θα αναφερθούμε αναλυτικά στο συγκεκριμένο στάδιο αργότερα.

2. Στο στάδιο της **εκτίμησης** υπολογίζουμε για κάθε ένα από τα υποψήφια μοντέλα τις παραμέτρους τους p, d, q και τα υλοποιούμε. Αυτό μπορεί να γίνει με αρκετούς τρόπους, ο πιο διαδεδομένος εκ των οποίων είναι υπολογίζοντας την *προσδοκώμενη πιθανοφάνεια* (Likelihood Estimation). Η προσδοκώμενη πιθανοφάνεια δείχνει επί της ουσίας κατά πόσο οι τιμές ενός μοντέλου με συγκεκριμένες παραμέτρους έχουν μεγάλη πιθανότητα να προσεγγίζουν τις πραγματικές τιμές της χρονοσειράς. Οι παράμετροι υπολογίζονται φυσικά με κριτήριο την μεγιστοποίηση της πιθανοφάνεια, ενώ συχνά χρησιμοποιείται και ο λογάριθμος αυτής.

$$L = \prod_{t=1}^{T} \left(\frac{1}{2\pi\sigma_t^2}\right)^{1/2} e^{-\sum_{t=1}^{T} \frac{(X_t - F(X_t))^2}{2\sigma t^2}}$$,ή εναλλακτικά

$$-2logL = \sum_{t=1}^{T} [\log(2\pi) + \log(\sigma_i^2) + \frac{\left(X_t - F(X_t)\right)^2}{\sigma_i^2}] \rightarrow$$

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

$$-2logL = n\,[\log(2\pi) + 1 + \log\left(\frac{RSS}{n}\right)]$$

,όπου L η προσδοκώμενη πιθανοφάνεια ταύτισης του μοντέλου με τα αρχικά δεδομένα, $F(X_t)$ η προβλεπόμενη από το μοντέλο τιμή τη περίοδο t, n ο αριθμός παρατηρήσεων, e_t το σφάλμα πρόβλεψης, σ^2 η διακύμανση των σφαλμάτων του μοντέλου και RSS (Round Sum of Squares) το άθροισμα των τετραγωνικών σφαλμάτων του μοντέλου.

Το κριτήριο αυτό λειτουργεί πρακτικά όπως η μέθοδος ελαχίστων τετραγώνων στην απλή γραμμική παλινδρόμηση για την επιλογή των παραμέτρων a και b, το οποίο ελαχιστοποιεί το άθροισμα των τετραγώνων των υπολειπόμενων σφαλμάτων μέσω της απαίτησης:

$$\min\left(\sum_{t=1}^{T} et^2\right), \text{όπου } e_t = y_t - f_t$$

Μάλιστα, στην περίπτωση της γραμμικής παλινδρόμησης το κριτήριο δίνει τις ίδιες παραμέτρους με αυτό των ελαχίστων τετραγώνων. Για την επιλογή των παραμέτρων συχνά συνδυάζεται η τεχνική των ελαχίστων τετραγώνων (Conditional Sum of Squares - CSS) με αυτή της προσδοκώμενης πιθανοφάνειας για καλύτερα αποτελέσματα. Φυσικά κανείς μπορεί κάλλιστα να χρησιμοποιήσει και άλλα κλασικά κριτήρια ελαχιστοποίησης σφαλμάτων για τον υπολογισμό των παραμέτρων (ME, MAPE, sMAPE κ.ο.κ.) όπως συμβαίνει για τον υπολογισμό των παραμέτρων των μοντέλων εκθετικής εξομάλυνσης.

3. Στο στάδιο του **διαγνωστικού ελέγχου** εφαρμόζουμε στατιστικά τεστ προκειμένου να ελεγχθεί αν τα μοντέλα που επιλέχθηκαν είναι στατιστικά σημαντικά. Ο διαγνωστικός έλεγχος προσαρμογής του μοντέλου πάνω στα αρχικά δεδομένα γίνεται μελετώντας την κατανομή των αντίστοιχων σφαλμάτων πρόβλεψης e_t. Αν το μοντέλο είναι αρκετά περιγραφικό, τότε τα σφάλματα που παράγει θα *πρέπει να οφείλονται αποκλειστικά στην τυχαιότητα* και συνεπώς να μην συσχετίζονται μεταξύ τους με κάποιον τρόπο, ή αλλιώς:

$$p_k(e) = \frac{\sum_{t=1}^{T}(e_{t-k} - \overline{e})(e_t - \overline{e})}{\sum_{t=k}^{T}(e_t - \overline{e})^2} = 0$$

, όπου p_k η συσχέτιση των σφαλμάτων για υστέρηση (διάστημα μεταξύ δύο παρατηρήσεων) k. Η συνάρτηση συσχέτισης θα μελετηθεί πιο αναλυτικά αργότερα.

Προφανώς τα σφάλματα ενός μοντέλου ARIMA ποτέ δεν είναι τελείως ασυσχέτιστα, όσο καλά και αν αυτό περιγράφει τη χρονοσειρά. Αναμένουμε λοιπόν για κάποιες υστερήσεις να βρούμε αρκετούς μη μηδενικούς δείκτες συσχέτισης. Για να ελέγχουμε αν αυτοί είναι σημαντικά διάφοροι του μηδενός, υπολογίζονται οι κατά προσέγγιση t-τιμές του τυπικού σφάλματός τους $S(r_k(e))$.

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

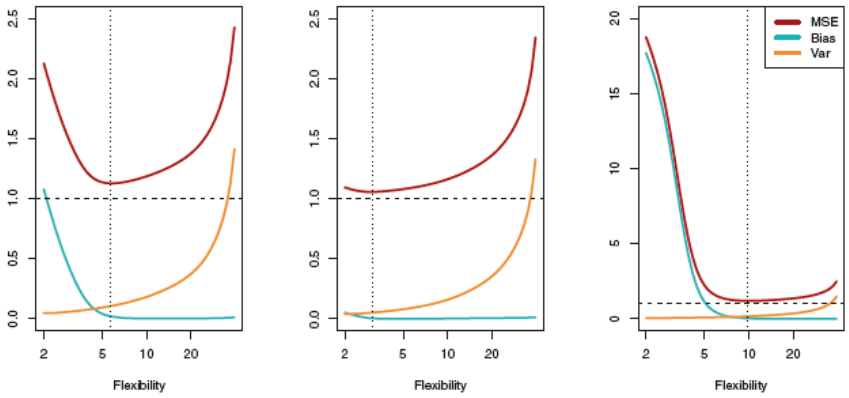
$$S(r_k(e)) = n^{-1/2}(1 + 2\sum_{j=1}^{k-1} {r_{j(e)}}^2)^{1/2}$$

Για να μην είναι σημαντική μία συσχέτιση (π.χ. να ανήκει στο 5% της κανονικής κατανομής πιθανότητας) η τιμή του t-δείκτη δεν πρέπει θεωρητικά να ξεπερνά την τιμή 2. Στην πράξη, για

υστέρηση 1,2 και 3 πρέπει να είναι μικρότερη του 1.25 και για μεγαλύτερη υστέρηση μικρότερη του 1.6. Το τεστ τύπου t θα μελετηθεί αναλυτικά αργότερα.

Άλλα κριτήρια που μπορούν να μας βοηθήσουν στην επιλογή του καταλληλότερου μοντέλου ARIMA είναι το Akaike's Information Criterion (AIC) και το Bayesian Information Criterion (BIC). Πρόκειται για κριτήρια τα οποία αξιολογούν κατά πόσο ταιριάζει το μοντέλο που εξετάζεται στην χρονοσειρά, συναρτήσει της πολυπλοκότητάς του. Φανερώνουν δηλαδή κατά πόσο αξίζει να γίνει το μοντέλο μας περισσότερο πολύπλοκο προκειμένου να αυξηθεί η πιθανότητα να ταυτίζονται οι παραγόμενες τιμές με τις πραγματικές. Η ποσοτικοποίηση της εν λόγω υπόθεσης είναι εξαιρετικά κρίσιμης σημασίας.

Σε όλα τα μοντέλα πρόβλεψης ισχύει ότι όσο αυξάνεται η πολυπλοκότητα (προσθέτονται παράγοντες) τόσο μειώνεται η προκατάληψή του (fitting). Ωστόσο η αύξηση των παραγόντων του οδηγεί σε συστηματική αύξηση της διακύμανσης των σφαλμάτων του και συνεπώς σε έλλειψη ακρίβειας πρόβλεψης (over-fitting). Στόχος μας είναι λοιπόν να καθορίσουμε πόσους παράγοντες θα πρέπει να έχει το μοντέλο προκειμένου να πετυχαίνουμε ταυτόχρονα μικρή προκατάληψη και υψηλή ακρίβεια. Χαρακτηριστικά παραδείγματα του εν λόγω φαινομένου δίνονται παρακάτω.



Αλληλεπίδραση προκατάληψης (bias) και ακρίβειας (variance) συναρτήσει της πολυπλοκότητας (flexibility) των μοντέλων. Η βέλτιστη επιλογή θα ήταν κατά σειρά από αριστερά προς τα δεξιά για τα τρία σετ δεδομένων η επιλογή ενός μοντέλου τεσσάρων, δύο και έξι παραγόντων. Αξίζει να σημειωθεί ότι το MSE δεν μπορεί να μας πληροφορήσει με αξιοπιστία για το ποιο μοντέλο είναι το πλέον κατάλληλο ανά περίπτωση αφού ελαχιστοποιείται για πολυπλοκότητα έξι, τρία και δέκα αντίστοιχα.

Ένα αρνητικό που εμφανίζουν ωστόσο τα δύο αυτά κριτήρια είναι ότι 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 ταιριάζει επαρκώς παρά μόνο για το ποιο είναι το καλύτερο από τα υπό εξέταση. Έτσι, η επιλογή του βέλτιστου μοντέλου γίνεται συγκρίνοντας την τιμή των κριτηρίων για όλα τα υποψήφια μοντέλα. Δεδομένου ότι τα

παραπάνω κριτήρια υπολογίζονται μέσω της μέγιστης πιθανοφάνειας θεωρούμε βέλτιστο εκείνο το μοντέλο που τα ελαχιστοποιεί. Αναλυτικά ο υπολογισμός τους δίνεται παρακάτω:

- **Akaike's Information Criterion (AIC):**

Η τιμή του κριτηρίου υπολογίζεται από τη σχέση

$$\mathrm{AIC} = -2\mathrm{logL} + 2(\mathrm{p}+\mathrm{q}+\mathrm{k}+1)$$

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

Σε περίπτωση που θέλουμε να δώσουμε μεγαλύτερο βάρος στην πολυπλοκότητα του μοντέλου χρησιμοποιούμε την παραλλαγή του κριτηρίου AICc:

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

,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 αποφευχθούν φαινόμενα υπερπροσαρμογής (over-fitting).

$$\mathrm{BIC} = \mathrm{AIC} + Log(n)(p+q+k+1)$$

Έχοντας λοιπόν κάποιος διαθέσιμο στα χέρια του ένα κριτήριο αξιολόγησης μοντέλων (AIC, AICc, BIC) και ένα κριτήριο βελτιστοποίησης παραμέτρων (L,CSS), μπορεί πλέον να επιλέξει και να εκτιμήσει το βέλτιστο μοντέλο ARIMA για κάθε χρονοσειρά. Αρχικά υπολογίζονται με το κριτήριο βελτιστοποίησης παραμέτρων οι βέλτιστες τιμές των παραμέτρων για όλα τα υπό εξέταση μοντέλα και στη συνέχεια από αυτά επιλέγεται το καλύτερο με τη βοήθεια των κριτηρίων αξιολόγησης. Για να εγκριθεί βέβαια το μοντέλο απαιτείται να ικανοποιεί τα τεστ του διαγνωστικού ελέγχου.

Αυτοσυσχέτιση, μερική αυτοσυσχέτιση και στασιμότητα

Όπως αναφέρθηκε νωρίτερα, στο στάδιο αναγνώρισης του καταλληλότερου μοντέλου ARIMA για μια χρονοσειρά απαιτείται η γνώση των τιμών των συντελεστών αυτοσυσχέτισης (Auto-Correlation Factor - ACF) και μερικής αυτοσυσχέτισης (Partial Auto-Correlation Factor - PACF).

Ο **συντελεστής αυτοσυσχέτισης** για υστέρηση k ορίζεται ως το πηλίκο της αυτοσυνδιακύμανσης γ_k προς τη διακύμανση $\gamma_0=\sigma^2$ της χρονοσειράς:

$$ACF(k) = pk = \frac{\gamma k}{\gamma o}$$

,όπου

$$\gamma_k = E[(y_{t-k} - \mu_{t-k})(y_t - \mu_t)] = E[(y_{t-k} - y_t) - \mu^2]$$

$$\gamma_0 = E[(y_t - \mu)^2] = E[y_t^2] - \mu^2 = \sigma^2$$

Εδώ Ε είναι η συνάρτηση προσδοκητής. Στην ουσία ο ACF μας δείχνει κατά πόσο η τιμή της χρονοσειράς σε μία περίοδο εξαρτάται από την τιμή της παρατήρησης k περιόδων πίσω. Παίρνει τιμές από +1 έως -1, οι οποίες δηλώνουν απόλυτα θετική ή αρνητική συσχέτιση αντίστοιχα, όπως ακριβώς ορίζει δηλαδή και η κλασική συσχέτιση κατά Pearson (R^2). Αν ο ACF ισούται με μηδέν τότε δεν υπάρχει καμία συσχέτιση μεταξύ των δύο τιμών.

Θεωρώντας στάσιμη χρονοσειρά η παραπάνω σχέση μπορεί να απλοποιηθεί ως:

$$p_k = \frac{\sum_{t=k+1}^{T}(y_t - \mu)(y_{t-k} - \mu)}{\sum_{t=1}^{T}(y_t - \mu)^2}$$

Ο **συντελεστής μερικής αυτοσυσχέτισης** 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 μπορούν ενδεχομένως να επιφέρουν οι ενδιάμεσες τιμές αυτής. Προφανώς για k=1 ο δείκτης ACF ταυτίζεται με αυτόν του PACF. Ο πιο ακριβής τρόπος υπολογισμού των PACF είναι ο υπολογισμός μιας σειράς συντελεστών παλινδρόμησης ελαχίστων τετραγώνων.

Θεωρώντας πάλι στάσιμη χρονοσειρά, οι συντελεστές PACF μπορούν να υπολογιστούν μέσω της σχέσης:

$$\varphi_{11} = \rho_1 \ , \varphi_{kk} = p_k - \sum_{j=1}^{k-1} \frac{\varphi_{k-1,j}\rho_{k-j}}{1 - \sum_{j=1}^{k-1}\varphi_{k-1,j}\rho_j} \gamma\iota\alpha\ k = 2,3 \ldots \kappa\alpha\iota$$

$$\varphi_{kj} = \varphi_{k-1,j} - \varphi_{kk}\varphi_{k-1,k-j}\ \gamma\iota\alpha\ k = 3,4,..\ j = 1,2,\ldots$$

Εδώ αξίζει να σημειωθεί ότι οι συντελεστές ACF και PACF των μοντέλων ARIMA αποτελούν στην πραγματικότητα εκτιμήσεις των πραγματικών αντίστοιχων συντελεστών της χρονοσειράς. Βρίσκονται αρκετά κοντά σε αυτές αλλά δεν ταυτίζονται. Αυτές οι εκτιμώμενες τιμές είναι οι συντελεστές των παραγόντων του γραμμικού συνδυασμού που συνθέτει το αντίστοιχο μοντέλο ARIMA και συμβολίζονται r_k και ϕ_k αντίστοιχα.

- Στατιστικός έλεγχος της ακρίβειας των συντελεστών ACF/PACF

Όπως έχουμε αναφέρει ήδη, επειδή οι παραγόμενες τιμές των μοντέλων ARIMA είναι αποτέλεσμα μιας στοχαστικής διαδικασίας και μόνο ένα μέρος των δυνατών διαδικασιών, οι τιμές των ACF και PACF είναι αντίστοιχα εκτιμήσεις των πραγματικών συντελεστών του μοντέλου.

Ισχύει ότι για p_k=0 οι τιμές των εκτιμώμενων συντελεστών r_k ακολουθούν κατά προσέγγιση κανονική κατανομή με τυπικό σφάλμα $S(r_k) = n^{-1/2}(1 + 2\sum_{j=1}^{k-1} {r_j}^2)^{1/2}$ για κ=2,3,.. και $S(\varphi_1) = n^{-1/2}$. Βασιζόμενη στην πρόταση αυτή υπολογίζουμε τα τυπικά σφάλματα των εκτιμηθέντων συντελεστών και ελέγχουμε πόσο απέχει κάθε τιμή r_k από την αρχική μας υπόθεση p_k=0. Αυτή η απόσταση δίνεται από το στατιστικό δείκτη t σαν έναν αριθμό εκτιμούμενων τυπικών σφαλμάτων:

$$t_{r_k} = \frac{r_k}{S(r_k)}$$

Γνωρίζοντας τώρα ότι σε μία κανονική κατανομή με κέντρο p_k το 95% των παρατηρήσεων r_k περιέχεται στο διάστημα $p_k \pm 2S(r_k)$ αντιλαμβανόμαστε ότι οποιοσδήποτε συντελεστής με απόλυτη τιμή t μεγαλύτερη του 2 είναι δείγμα έλλειψης σημαντικότητας βρισκόμενη σε ένα επίπεδο 5% περίπου. Για τους συντελεστές μερικής αυτοσυσχέτισης το τυπικό σφάλμα είναι αντίστοιχα $S(\varphi_k) = n^{-1/2}$.

- Στασιμότητα χρονοσειρών: Εξομάλυνση και Διαφόριση

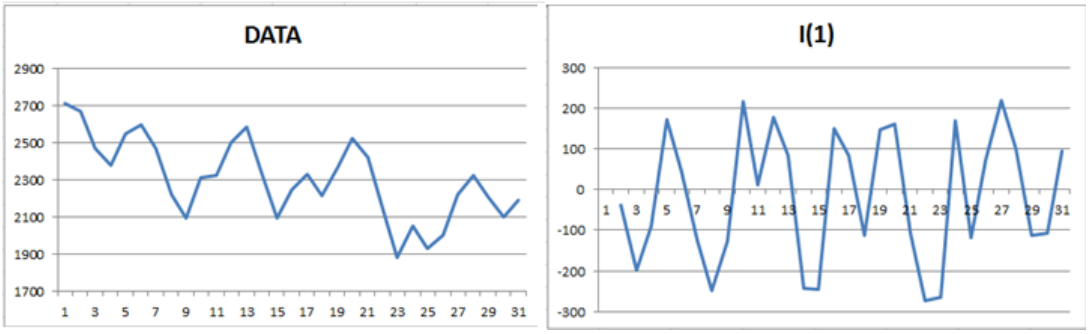
Όπως αναφέρθηκε νωρίτερα, για να είναι μία χρονοσειρά στάσιμη πρέπει η μέση τιμή, η διακύμανση και η συνάρτηση αυτοσυσχέτισής της να είναι σταθερές στην πάροδο του χρόνου. Αυτή η απαίτηση, η οποία είναι απαραίτητη στην αποτελεσματική εφαρμογή ενός μοντέλου ARIMA, σπανίως ικανοποιείται εξαρχής. Μπορούμε ωστόσο σχετικά εύκολα να εφαρμόσουμε μετασχηματισμούς στην χρονοσειρά και να το πετύχουμε.

1. Ένα πρώτο μέτρο που προτείνεται είναι η **λογαρίθμηση** της χρονοσειράς. Αυτό προσδίδει στη χρονοσειρά μικρότερη τιμή διακύμανσης και συνεπώς μεγαλύτερη σταθερότητα. Το πότε πρέπει να εφαρμόζεται η λογαρίθμηση δεν είναι σαφώς ορισμένο. Ωστόσο, αν η χρονοσειρά εμφανίζει έντονες διακυμάνσεις ή το εφαρμοζόμενο μοντέλο ARIMA παρουσιάζει συστηματικά μεγάλα σφάλματα ακρίβειας καλό θα ήταν να δοκιμαστεί η υιοθέτησή του. Εναλλακτικά μπορούν να εφαρμοστούν και άλλοι γενικότεροι τύποι μετασχηματισμών, όπως οι

μετασχηματισμοί Box-Cox. Απαραίτητη προϋπόθεση είναι φυσικά η θετική τιμή των παρατηρήσεων κατά μήκος όλου του σετ δεδομένων.

2. Αν η χρονοσειρά παρουσιάζει επιπλέον τάση ή εποχιακότητα, ένα μέτρο που προτείνεται είναι η διαφόρισή της (Differencing). Η **διαφόριση** περιορίζει τις διακυμάνσεις επιπέδου αφαιρώντας τάση και εποχιακότητα. Έτσι παράγεται μία χρονοσειρά σταθερού επιπέδου και διακύμανσης. Στην ουσία κατά τη διαφόριση μίας χρονοσειράς Τ περιόδων δημιουργείται μία νέα με στοιχεία της τις διαφορές των παρατηρήσεων της πρώτης. Ανάλογα την τάξη διαφόρισης έχουμε:

- $1^{η}$ τάξη: $y'_t=y_t-y_{t-1}$
- $2^{η}$ τάξη: $y''_t=y'_t-y'_{t-1}=y_t-2y_{t-1}+y_{t-2}$ κ.ο.κ.



Χρονοσειρά με τάση πριν και μετά τη διαφόριση πρώτης τάξης. Η νέα χρονοσειρά είναι εμφανώς πιο στάσιμη

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

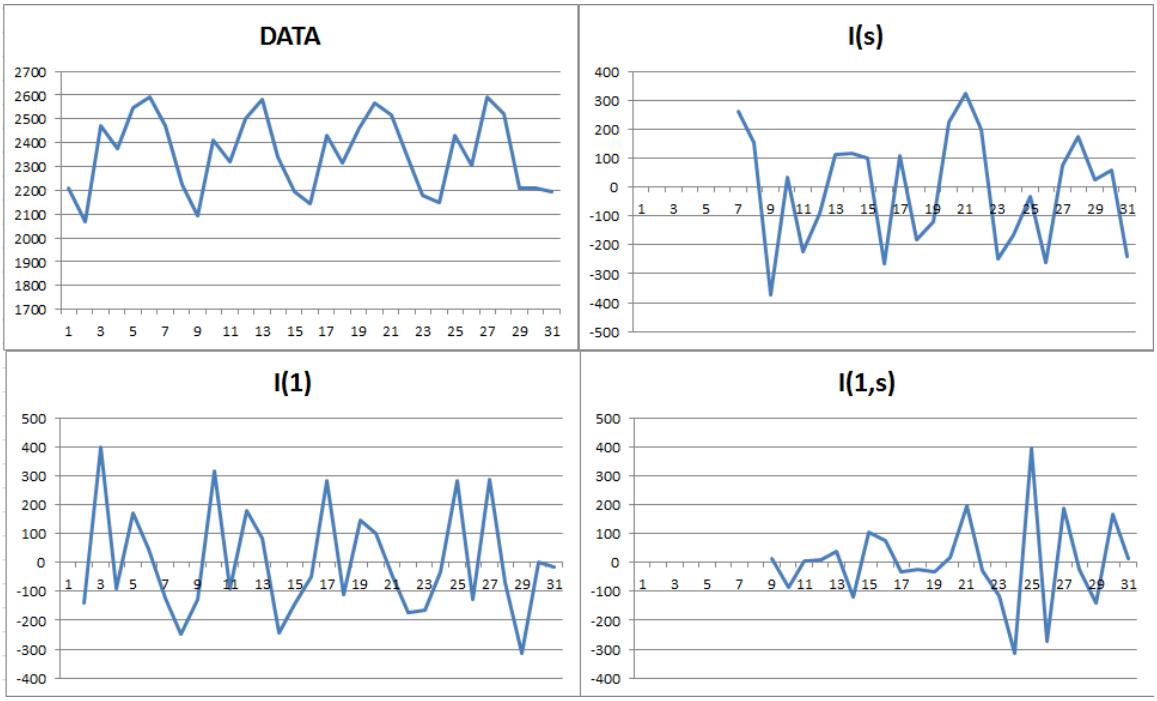
- Λευκός θόρυβος: $y_t-y_{t-1}=e_t$
- Τυχαίος Περίπατος: $y_t-y_{t-1}=c+e_t$

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 διαφόρισης μεταξύ των παρατηρήσεων της αρχικής χρονοσειράς και προηγούμενων αντίστοιχων περιόδων εποχιακότητας. Ανάλογα τη τάξη διαφόρισης έχουμε:

- $1^{η}$ τάξης $y'_t=y_t-y_{t-m}$
- $2^{η}$ τάξη: $y''_t=y'_t(t)-y'_{t-m}=y_t-2y_{t-m}+y_{t-2m}$, όπου m η περίοδος εποχιακότητας κ.ο.κ.

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 $\boldsymbol{(1-B)^n y_t}$ και την εποχιακή διαφόριση m τάξης N ως $\boldsymbol{(1-B^m)^N y_t}$.

Αν η εποχιακή διαφόριση δεν έχει αποδώσει επαρκή σταθερότητα, τότε ενδέχεται να πρέπει να συνδυαστεί με την απλή διαφόριση. Αυτό γίνεται αναπτύσσοντας τη σχέση $(1-B^m)^N(1-B)^n y_t$.



Εποχιακή εβδομαδιαία χρονοσειρά πριν και μετά τη συνδυαστική διαφόριση I(1,s) (πρώτης τάξης I(1)-πρώτης τάξης και εποχιακότητας 7 I(s)).

Αυτό που χρειάζεται να έχουμε κατά νου είναι ότι η διαφόριση δεν αποτελεί πανάκεια λύση και για αυτό δεν θα πρέπει να υπερβάλουμε με τη διαδικασία της διαφόρισης. Πρώτα απ' όλα 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 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 με τη σειρά του θα οδηγήσει αργότερα στην ανάγκη εφαρμογής μοντέλων ARMA υψηλής πολυπλοκότητας ικανών να αντιμετωπίσουν την τυχαιότητα. Το φαινόμενο αυτό ονομάζεται **υπερδιαφόριση** (over-differencing). *Στην πράξη δεν διαφορίζουμε για τιμές αυτοσυσχέτισης μικρότερες του 0.5 και σε καμία περίπτωση περισσότερες από δύο φορές.* Αν έχουμε έντονη αρνητική αυτοσυσχέτιση (<-0.5) είναι δείγμα υπερδιαφόρισης. *Επίσης, η εμπειρία δείχνει ότι ποτέ δεν χρησιμοποιείται εποχιακή διαφόριση μεγαλύτερη της πρώτης τάξης*.

Μοντέλα Αυτοπαλινδρόμησης - Autoregressive models- AR(p)

Τα συνήθη μοντέλα παλινδρόμησης θεωρούν σχέσεις (γραμμικές ή μη) μεταξύ των τιμών μίας χρονοσειράς και παραγόντων με τους οποίους συσχετίζεται (και από τους οποίους εξαρτώνται σε κάποιο βαθμό) για να την περιγράψουν στο χρόνο.

Τα μοντέλα αυτοπαλινδρόμησης θεωρούν από τη μεριά τους γραμμικές σχέσεις ανάμεσα στην παρατήρηση της χρονοσειράς που εξετάζεται και στις προηγούμενες τιμές αυτής. Ένα τέτοιο μοντέλο p τάξης αναπαριστάται αλγεβρικά ως εξής:

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

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

Στην ουσία δηλαδή 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:

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

, 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:

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

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

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

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

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

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

Μοντέλα Κινητού Μέσου Όρου - Moving Average 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:

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

,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 –προαιρετικά- την κατά μία σταθερά c.

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

$$\boldsymbol{\overline{y}_t = \left(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q\right) e_t}$$

Αν αναπτύξουμε τώρα την παραπάνω σχέση έχουμε:

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

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

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

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:

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

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

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

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

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

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

Δεδομένου ότι -1<ϕ_1<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):

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

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

Σύμφωνα με θεώρημα των γεωμετρικών σειρών αν -1<θ_1<1, τότε ο όρος $(1 - \theta_1 B)^{-1}$ μπορεί να γραφτεί ως ένα άθροισμα απείρων όρων μιας συγκλίνουσας σειράς $\left(1 + \theta_1 B + {\theta_1}^2 B^2 + {\theta_1}^3 B^3 + \cdots\right)$. Έτσι η αρχική σχέση μπορεί να γραφτεί ως ένα αμιγώς AR μοντέλο ως:

$$\left(1 + \theta_1 B + {\theta_1}^2 B^2 + {\theta_1}^3 B^3 + \cdots\right)\overline{y}_t = e_t$$

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<θ_1<1 για MA(1) και AR(1)), διασφαλίζει 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<θ_1<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.

Μοντέλα 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:

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

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:

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

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

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

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

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 εν τέλει να είναι ανεπαρκής. Η εισαγωγή λοιπόν μίας σταθεράς c μπορεί να βοηθήσει στον καλύτερο προσδιορισμό του επιπέδου της και σε γενικές γραμμές συστήνεται.
- Η διαφόριση πρώτης τάξης (d=1) συνεπάγεται ύπαρξη *σταθερής τάσης* στην αρχική χρονοσειρά. Αυτή θεωρητικά έχει εξαλειφτεί μετά τη διαφόριση, ωστόσο ενδέχεται να έχει υποπέσει στην από πάνω περίπτωση. Έτσι, η εισαγωγή μίας σταθεράς μπορεί να γίνει, έχοντας πρώτα βεβαιωθεί για την ανάγκη υπερτόνισης του επίπεδου της παραγόμενης χρονοσειράς. Συνήθως αποφεύγεται να γίνει κάτι τέτοιο ή αποφασίζουμε χρησιμοποιώντας όπως έχει αναφερθεί κάποιο από τα κριτήρια AIC, AICc και BIC για τα μοντέλα με και χωρίς σταθερά.
- 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.

➢ *Συνθήκες συντελεστών για τα μοντέλα ARIMA*

Όπως έχουμε ήδη αναφέρει, η εφαρμογή των μοντέλων ARIMA πρέπει να γίνεται μόνο σε στάσιμες χρονοσειρές. Αυτό εξασφαλίζει ότι μπορούμε να λάβουμε ικανοποιητικές εκτιμήσεις για τη μέση τιμή, τη διακύμανση και τη συνάρτηση αυτοσυσχέτισης της διαδικασίας από το δείγμα. Ο έλεγχος για το αν η χρονοσειρά είναι στάσιμη μπορεί σε μία απλή του μορφή να γίνει ελέγχοντας αν οι συντελεστές φi και θi του μοντέλου ικανοποιούν συγκεκριμένους περιορισμούς.

Παρακάτω δίνονται οι περιορισμοί που πρέπει να πληρούνται από τα πιο διαδεδομένα μοντέλα ARIMA. Για τα υπόλοιπα μοντέλα (που είναι και περισσότερο σύνθετα) οι αντίστοιχοι περιορισμοί γίνονται ιδιαίτερα σύνθετοι. Η παρουσίασή τους εδώ λοιπόν θεωρείται άσκοπη.

- 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 $\vartheta_1 - \vartheta_2 <1$
- Για μοντέλα AR(p) με p>2 πρέπει τουλάχιστον να ισχύει $\varphi_1 + \varphi_2 + \varphi_3 + \ldots + \varphi_p <1$

Επίσης, δεδομένων των τιμών αυτοσυσχέτισης μίας χρονοσειράς, για τους συντελεστές των μοντέλων ARIMA 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 $\rho_k = 0$ γιαk>q.

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

$$\rho_k = {\rho_1}^k$$ για k=1,2,..p

Αναγνώριση πιθανών μοντέλων ARIMA

Έχοντας δει ότι οι εκτιμώμενοι συντελεστές ACF και PACF μίας χρονοσειράς είναι επί τις ουσίας οι συντελεστές του μοντέλου ARIMA που την περιγράφει, μπορούμε παρατηρώντας την εξέλιξη των θεωρητικών συντελεστών ACF και PACF αυτής σε διαγράμματα να βγάλουμε συμπέρασμα για το ποια μοντέλα ARIMA είναι πιθανό να την περιγράφουν ικανοποιητικά. Συνοπτικά αναφέρουμε τα εξής:

Για ένα στάσιμο 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

Για ένα στάσιμο μοντέλο MA(q)

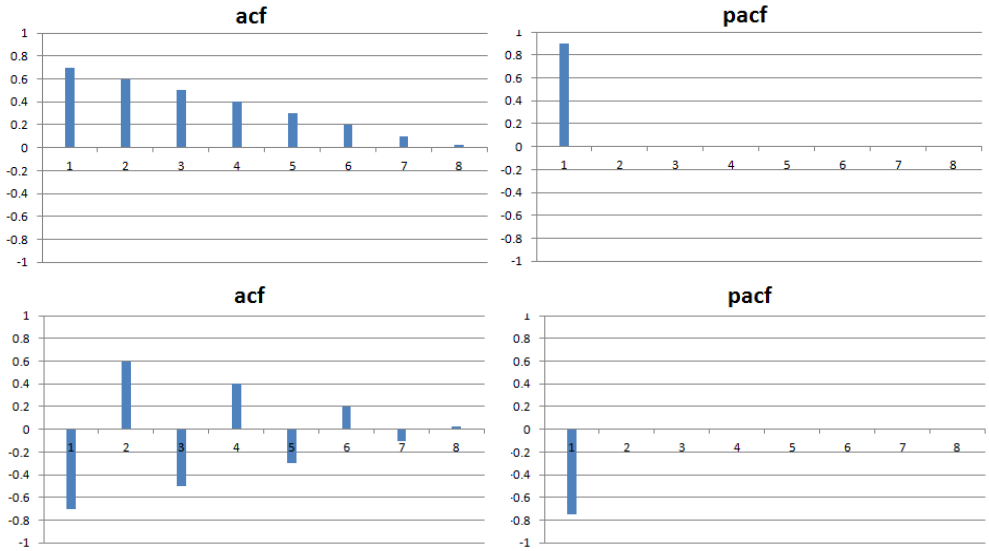
- Οι τιμές των συντελεστών ACF μηδενίζονται απότομα μετά από q περιόδους υστέρησης
- Οι τιμές των συντελεστών PACF φθίνουν προς το μηδέν ακολουθώντας εκθετική η ημιτονοειδή πορεία

Για ένα στάσιμο μοντέλο ARIMA(p,q)

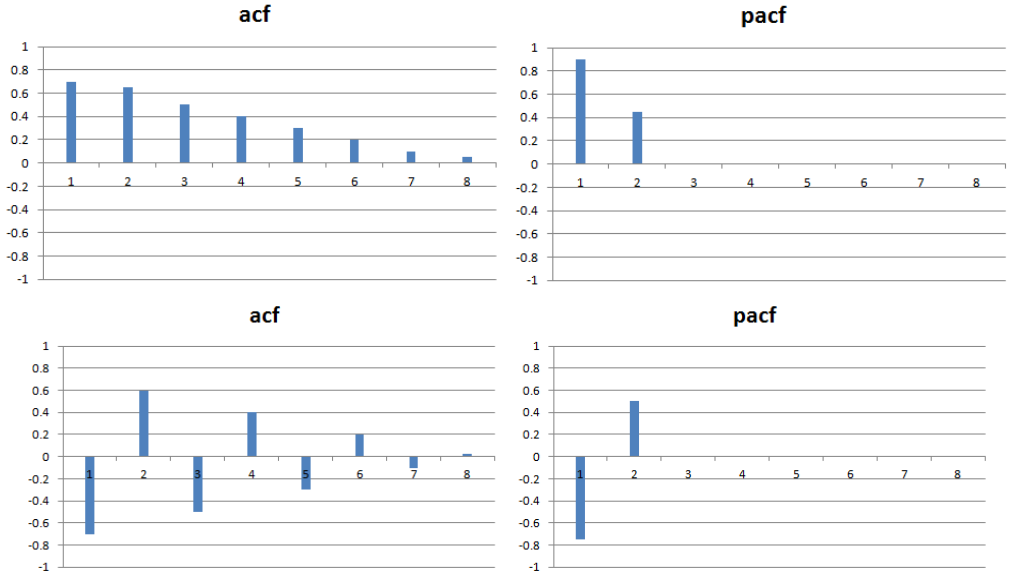
- Οι τιμές των συντελεστών ACF φθίνουν προς το μηδέν μετά από q-p περιόδους υστέρησης
- Οι τιμές των συντελεστών PACF φθίνουν προς το μηδέν μετά από p-q περιόδους υστέρησης

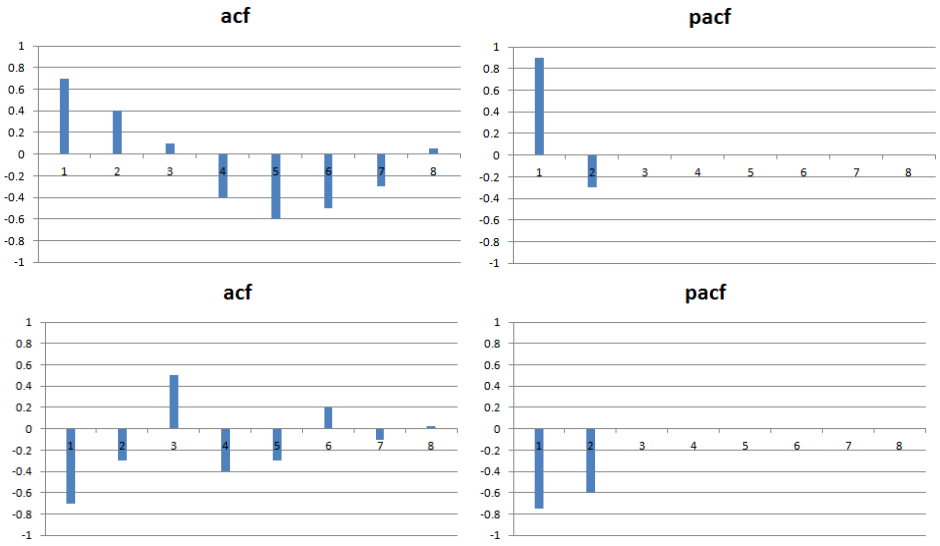
Παρακάτω δίνονται παραδείγματα αναγνώρισης συνήθων διαδικασιών ARIMA.

<u>Διαδικασίες AR(1):</u>

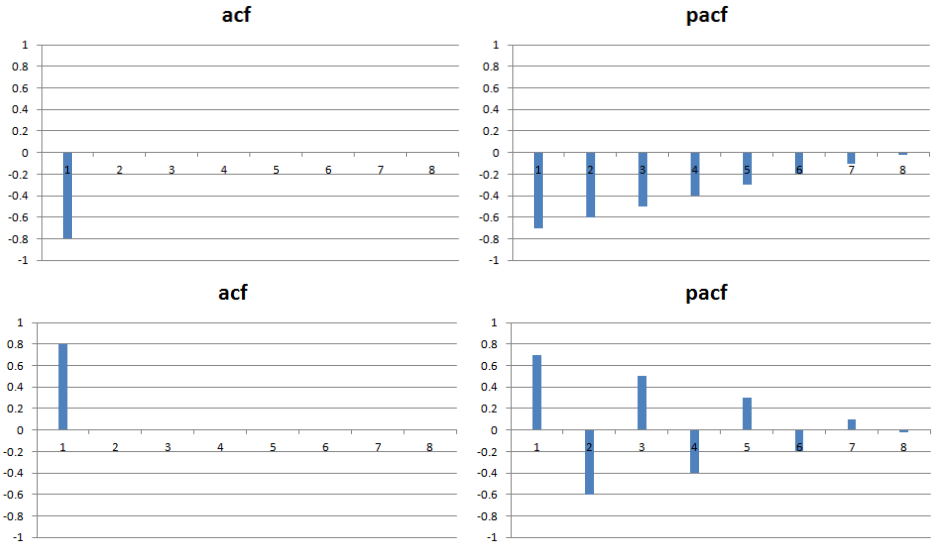


<u>Διαδικασίες AR(2):</u>

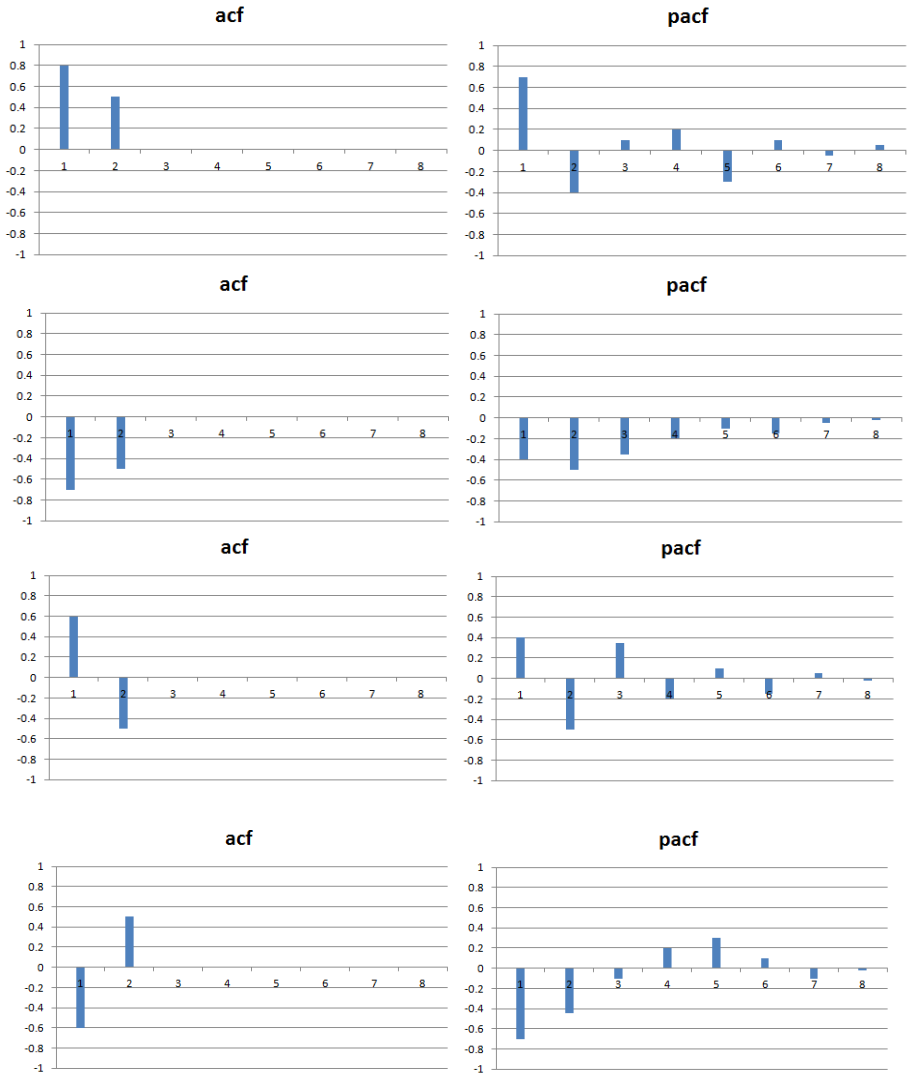




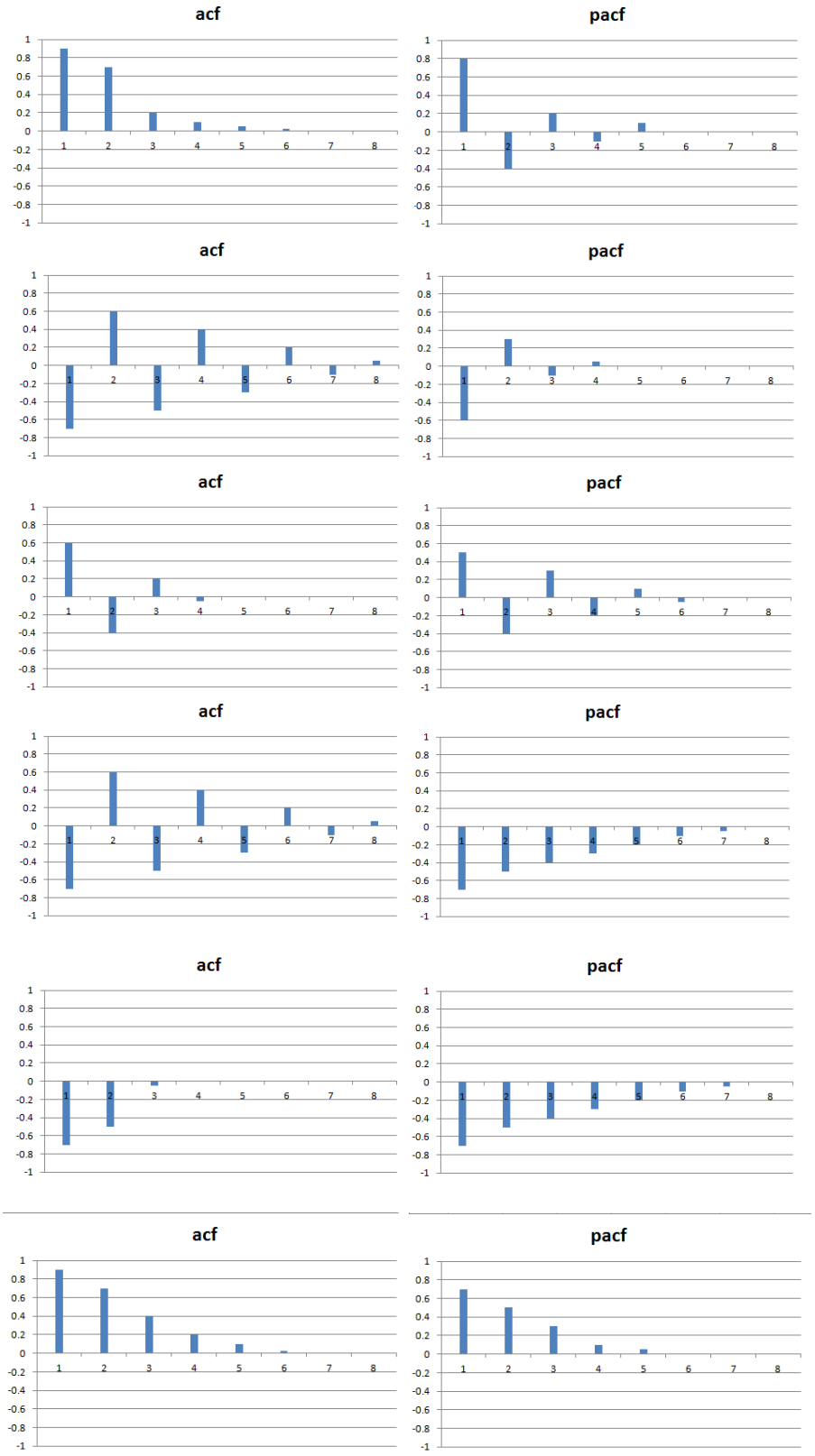
Διαδικασίες MA(1):



Διαδικασίες MA(2):



Διαδικασίες ARMA(1,1):



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 τιμές των δεδομένων. Αυτός είναι και ο λόγος που χρησιμοποιείται κυρίως για βραχυπρόθεσμες προβλέψεις.

Δεδομένου ότι η κατανομή 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 της με παρόμοια αποτελέσματα. Μία τέτοια διαδικασία (back-forecasting) 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-forecasting 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.

Επίλογος

Τα μοντέλα ARIMA μπορούν να αποτελέσουν εξαιρετικό εργαλείο στην κατανόηση της εξέλιξης των φυσικών μεγεθών αναλύοντάς και προεκτείνοντάς τα στο μέλλον. Αντιμετωπίζουν τις χρονοσειρές με μία στοχαστική ματιά βασιζόμενα αποκλειστικά στην κατανομή των παρελθοντικών τους τιμών και κυρίως στων πιο πρόσφατων. Αυτό τα καθιστά αποτελεσματικά κυρίως για βραχυπρόθεσμη πρόβλεψη.

Επειδή η διαδικασία που πρεσβεύει κάθε μοντέλο οδηγεί σε μερική αναπαράσταση της πραγματικής χρονοσειράς, το μόνο που μπορούμε να επιζητούμε είναι ο εντοπισμός εκείνου του μοντέλου που την περιγράφει αποτελεσματικά χωρίς να εισάγει περιττή πολυπλοκότητα στην όλη διαδικασία. Η φειδωλότητα σε ένα μοντέλο ARIMA, δηλαδή η επαρκής περιγραφή μίας χρονοσειράς χωρίς τη χρήση υπερβολικού αριθμού συντελεστών, είναι χαρακτηριστικό ζωτικής σημασίας. Αυτή μπορεί να εξασφαλιστεί χρησιμοποιώντας το σύνολο των μεθόδων και κριτηρίων που παρουσιάστηκαν.

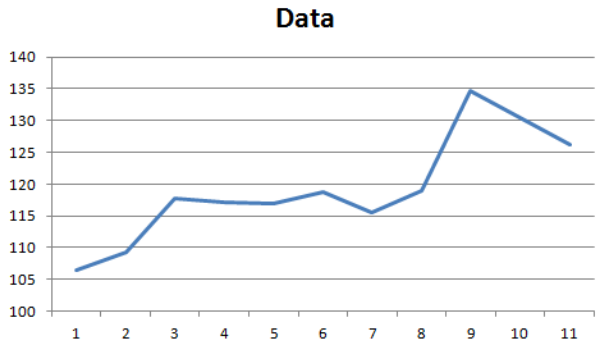
Σημαντικός παράγοντας στην αποτελεσματικότητα των μοντέλων είναι και η ύπαρξη στασιμότητας στην εξεταζόμενη χρονοσειρά. Νωρίτερα αναλύσαμε πώς η στασιμότητα και η αντιστρεψιμότητα εξασφαλίζουν την αύξηση της ακρίβειας προσαρμογής σε μοντέλα AR και MA αντίστοιχα, καθώς και μηχανισμούς με τους οποίους αυτή εξασφαλίζεται. Επιπλέον, με την προϋπόθεση της στασιμότητας εκτιμούνται συντελεστές υψηλής ποιότητας, στατιστικά δηλαδή σημαντικοί για το μοντέλο. Η συσχέτιση μεταξύ τους υποβιβάζεται και το μοντέλο αποκτά ευστάθεια στο χρόνο. Ένα μοντέλο ARIMA που τηρεί τις παραπάνω απαιτήσεις παράγει συνήθως και ακριβείς προβλέψεις σε βραχυπρόθεσμο πάντα επίπεδο.

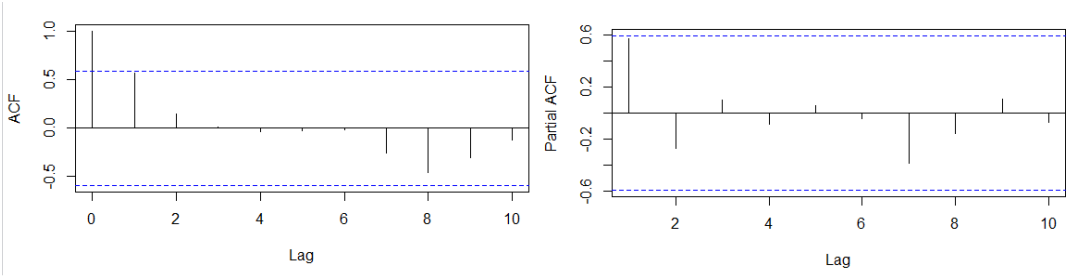
Εφαρμογές μοντέλων ARIMA

Παράδειγμα 1:

Δίνεται η ακόλουθη χρονοσειρά μήκους 10 παρατηρήσεων. Ζητείται πρόβλεψη για τη χρονική στιγμή 11 (θεωρητικά άγνωστης τιμής). Η επιλογή του κατάλληλου μοντέλου πρόβλεψης ARIMA να γίνει αξιοποιώντας τις γραφικές των συντελεστών ACF και PACF.

Period	Data
1	106.5
2	109.2
3	117.8
4	117.2
5	116.9
6	118.7
7	115.6
8	119
9	134.7
10	130.4
11	126.2

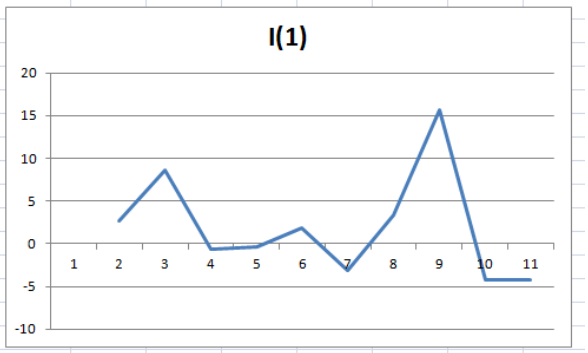


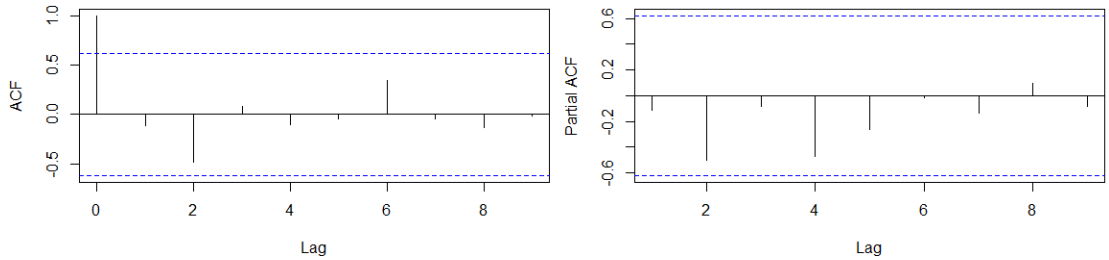


Παρατηρώντας τη γραφική παράσταση της χρονοσειράς βλέπουμε ότι αν και στο μεγαλύτερό της μέρος είναι σταθερή, στα δύο της άκρα εμφανίζει μία σχετική τάση. Αυτό αναδεικνύεται και μέσα από το διάγραμμα ACF καθώς εντοπίζεται κάποια ασθενής συσχέτιση για υστέρηση 1 (0.5), η οποία βέβαια σβήνει σχεδόν ακαριαία. Ταυτόχρονα βλέπουμε ότι στο διάγραμμα PACF υπάρχει σημαντική συσχέτιση για υστέρηση 1, πιθανό σημάδι ότι η εν λόγω χρονοσειρά μπορεί ενδεχομένως να περιγράφεται ικανοποιητικά από μία διαδικασία AR. Οι επιλογές μας σε αυτό το σημείο είναι δύο: Ή να δημιουργήσουμε μία στάσιμη χρονοσειρά και να προσπαθήσουμε να την προβλέψουμε μέσω ενός μοντέλου ARMA είτε να θεωρήσουμε επαρκή στασιμότητα στην αρχική χρονοσειρά και να την προεκτείνουμε υπολογίζοντας ένα μοντέλο AR.

Δεδομένης της τάσης της χρονοσειράς ας δοκιμάσουμε σαν πρώτο βήμα να εφαρμόσουμε διαφόριση πρώτης τάξης προκειμένου αυτή να γίνει επαρκώς στάσιμη. Εποχιακότητα δεν παρατηρείται οπότε δεν προβαίνουμε σε κάποια ανάλογη ενέργεια. Το αποτέλεσμα που προκύπτει είναι το παρακάτω:

Period	I(1)
1	
2	2.7
3	8.6
4	-0.6
5	-0.3
6	1.8
7	-3.1
8	3.4
9	15.7
10	-4.3
11	-4.2





Όπως βλέπουμε ο συντελεστής ACF είναι σημαντικός μόνο για υστέρηση 2 ενώ ο PACF για υστέρηση 2 και 4. Τα διαγράμματα ωστόσο δεν παρουσιάζουν κάποια ημιτονοειδή ή εκθετική μεταβολή πέρα από αυτά τα σημεία συνεπώς δεν υπάρχει εμφανής ένδειξη χρήσης κάποιου μοντέλου τύπου MA ή AR. Επίσης η εφαρμογή ενός πιο πολύπλοκου μοντέλου ARMA κρίνεται μάλλον υπερβολική. Έτσι, δεδομένης της διαφόρισης θα επιλέγαμε να εφαρμόσουμε στην αρχική μας χρονοσειρά ένα απλό μοντέλο ARIMA(0,1,0).

$$(1 - B)y_t = c + e_t$$

Για το μοντέλο αυτό επιλέγουμε σταθερά c=2.66, η οποία είναι ο αριθμητικός μέσος της χρονοσειράς διαφόρισης. Ο λόγος που το κάνουμε αυτό είναι ότι σε αντίθεση με το θεωρητικά αναμενόμενο αποτέλεσμα η διαφορισμένη χρονοσειρά που προκύπτει δεν είναι απολύτως στάσιμη, αλλά γίνεται επαρκώς αν συμπεριληφθεί σε αυτήν ο μέσος της c. Επίσης όπως έχει αναφερθεί, όταν το μοντέλο είναι σχετικά απλό συνίσταται η εισαγωγή σταθεράς, σε αντίθεση με τα πιο πολύπλοκα μοντέλα ARIMA όπου εξαιρείται. Συνεπώς έχουμε:

$$y_t = 2.66 + y_{t-1} + e_t$$

Observation	Data	ARIMA(0,1,0)	et
1	106.5	-	-
2	109.2	109.16	0.04
3	117.8	111.86	5.94
4	117.2	120.46	-3.26
5	116.9	119.86	-2.96
6	118.7	119.56	-0.86
7	115.6	121.36	-5.76
8	119	118.26	0.74
9	134.7	121.66	13.04
10	130.4	137.36	-6.96
11	***126.2***	**133.06**	**-6.86**

Το μοντέλο που επιλέχθηκε φαντάζει υπερβολικά απλό και αφελές, μπορεί ωστόσο να είναι αρκετά πιο αποδοτικό από άλλα πιο σύνθετα μοντέλα. Για να το εξακριβώσουμε αυτό θα υπολογίσουμε αρχικά την επίδοση του μοντέλου ARIMA(0,0,1).

$$y_t = c + (1 - \theta_1 B)e_t$$

Εδώ επιλέγουμε c=μ=118.6, ο οποίος είναι και ο αριθμητικός μέσος της αρχικής χρονοσειράς. Ο συντελεστής θ1 υπολογίζεται κατά τα γνωστά από την εξίσωση:

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

$$\theta_1 = \frac{-1 + \sqrt{1 - 4{\rho_1}^2}}{2\rho_1}$$

Η τιμή αυτοσυσχέτισης της αρχικής χρονοσειράς για υστέρηση 1 είναι 0.49. Συνεπώς έχουμε θ_1=-0.82 (η δεύτερη λύση -1.22 απορρίπτεται αφού πρέπει $-1 \leq \theta_1 \leq 1$).

Η εξίσωση του μοντέλου ARIMA(0,0,1) θα είναι λοιπόν:

$$y_t = 118.6 + 0.82e_{t-1} + e_t$$

Observation	Data	ARIMA(0,0,1)	et
1	106.5	106.50	0.00
2	109.2	118.60	-9.40
3	117.8	110.89	6.91
4	117.2	124.26	-7.06
5	116.9	112.81	4.09
6	118.7	121.96	-3.26
7	115.6	115.93	-0.33
8	119	118.33	0.67
9	134.7	119.15	15.55
10	130.4	131.35	-0.95
11	***126.2***	**117.82**	**8.38**

Υπενθυμίζουμε ότι επειδή η τιμή του μοντέλου δεν μπορεί να υπολογιστεί μέσω της εξίσωσης για τη χρονική στιγμή t=1, θεωρούμε τιμή ίση με αυτήν της αρχικής παρατήρησης της χρονοσειράς και μηδενικό σφάλμα. Εναλλακτικά θα μπορούσαμε να είχαμε χρησιμοποιήσει την τεχνική back-casting για πρόβλεψη στη στιγμή t=0 και να αξιοποιούσαμε την εκεί πρόβλεψή μας για την εκτίμηση του σφάλματος e_0.

Υπολογίζοντας τα σφάλματα MAPE για τα δύο μοντέλα βλέπουμε ότι το πρώτο εμφανίζει in sample σφάλμα 3.53% ενώ το δεύτερο 4.43%. Συνεπώς ορθά ακολουθήθηκε η πρώτη στρατηγική (αυτή που υπαγόρευσαν τα διαγράμματα ACF-PACF) και δεν υιοθετήθηκε κάποιο πιο πολύπλοκο μοντέλο. Μάλιστα αυτό μεταφράζεται σε καλύτερη ακρίβεια και στην τελική μας πρόβλεψη (6.86 αντί για 8.38 μονάδες απόλυτο σφάλμα).

Τέλος, δοκιμάζουμε το μοντέλο ARIMA(1,0,0) που τα διαγράμματα ACF και PACF μας είχαν υποδείξει εξαρχής, θεωρώντας μία αρχικά στάσιμη χρονοσειρά. Με αυτήν την υπόθεση, και δεδομένου ότι οι συντελεστές ACF ακολουθούν κάποια ημιτονοειδή κυμάτωση με τον συντελεστή PACF να είναι σημαντικός για υστέρηση 1, η εν λόγω επιλογή μπορεί να θεωρηθεί έγκυρη. Η σταθερά του μοντέλου αυτή τη φορά θα είναι ίση με c = μ(1-ρ1) = 60.49. Συνεπώς έχουμε:

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

$$y_t = 60.49 + 0.49 y_{t-1} + e_t$$

Observation	Data	ARIMA(1,0,0)	et
1	106.5	-	-
2	109.2	112.671	-3.47
3	117.8	113.994	3.81
4	117.2	118.208	-1.01
5	116.9	117.914	-1.01
6	118.7	117.767	0.93
7	115.6	118.649	-3.05
8	119	117.13	1.87
9	134.7	118.796	15.90
10	130.4	126.489	3.91
11	***126.2***	**124.382**	**1.82**

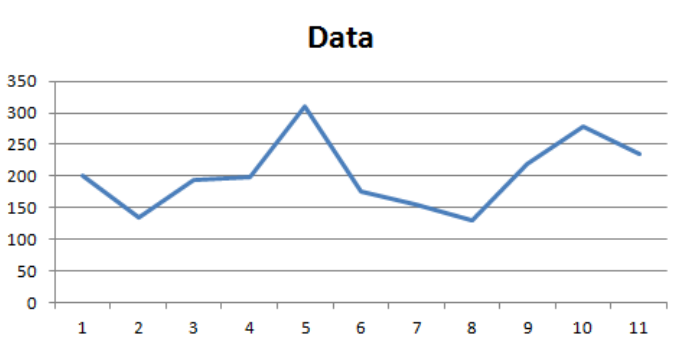
Το in sample σφάλμα εδώ είναι 3.10%, δηλαδή αρκετά καλύτερο από των δύο προηγούμενων μοντέλων. Επίσης η τελική μας πρόβλεψη είναι σημαντικά καλύτερη των προηγούμενων δύο μοντέλων με απόλυτη διαφορά μόλις 1.82 μονάδων.

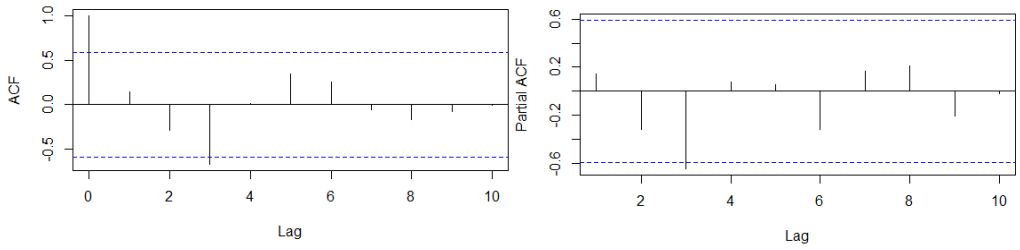
Όπως φάνηκε από το συγκεκριμένο παράδειγμα δεν θα μπορούσαμε να γνωρίζουμε με σιγουριά εξαρχής ποιο μοντέλο ARIMA θα έπρεπε να εφαρμόσουμε. Ενδείξεις για χρήση ενός μοντέλου AR προφανώς και υπήρχαν αλλά θα έπρεπε να είχαμε θεωρήσει καταχρηστικά στασιμότητα. Για αυτό και η χρήση των μοντέλων ARIMA γίνεται πάντα συμβουλευόμενοι κριτηρίων όπως το AIC και εκτελώντας τεστ σημαντικότητας. Εδώ για παράδειγμα, το μοντέλο (0,1,0) εμφάνιζε στα σφάλματά του συσχέτιση πρώτης τάξης 0.31, ενώ το (1,0,0) 0.12. Αυτό θα μας προβλημάτιζε και θα μας έκανε να αναζητήσουμε ένα καλύτερο μοντέλο για το σετ δεδομένων μας από την αρχική μας επιλογή.

Παράδειγμα 2:

Δίνεται η παρακάτω χρονοσειρά. Επιλέξτε κατάλληλο μοντέλο πρόβλεψης ARIMA βασιζόμενοι στις γραφικές παραστάσεις των συντελεστών ACF και PACF.

Period	Data
1	200
2	135
3	195
4	197.5
5	310
6	175
7	155
8	130
9	220
10	277.5
11	235





Όπως βλέπουμε οι συντελεστές ACF και PACF παρουσιάζουν υψηλές τιμές μόνο για υστέρηση 3 και στη συνέχεια γίνονται αρκετά ασήμαντες. Επίσης οι συντελεστές δεν παρουσιάζουν κάποια από τις γνωστές διακυμάνσεις. Επιλέγεται λοιπόν το μοντέλο ARIMA(0,0,0) με τη λογική ότι δεν υπάρχει καμία ένδειξη για το αν κάποιο άλλο μοντέλο πέραν του μέσου όρου της χρονοσειράς θα μπορούσε να προσδώσει κάτι περισσότερο προβλεπτικά.

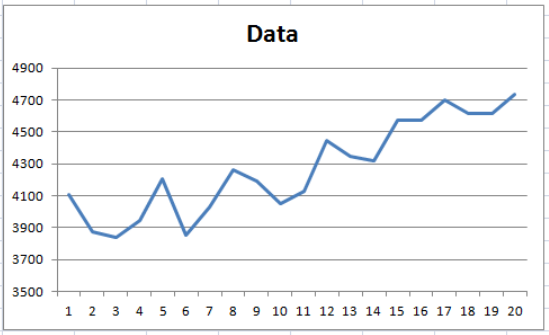
$$y_t = c + e_t$$

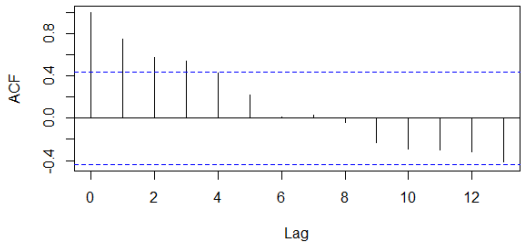
, όπου c=μ=202.73 ο μέσος όρος των παρατηρήσεων.

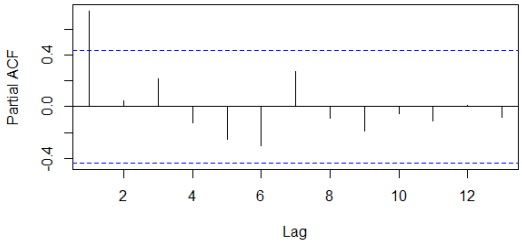
Παράδειγμα 3:

Δίνεται η παρακάτω χρονοσειρά. Επιλέξτε κατάλληλο μοντέλο πρόβλεψης ARIMA βασιζόμενοι στις γραφικές παραστάσεις των συντελεστών ACF και PACF και πραγματοποιείστε προβλέψεις για το in-sample δείγμα.

Period	Data
1	4109
2	3874
3	3842
4	3946
5	4207
6	3850
7	4030
8	4260
9	4193
10	4051
11	4126
12	4445
13	4344
14	4319
15	4571
16	4576
17	4699
18	4614
19	4613
20	4738

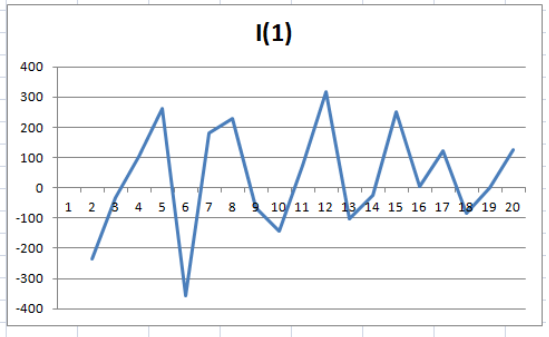


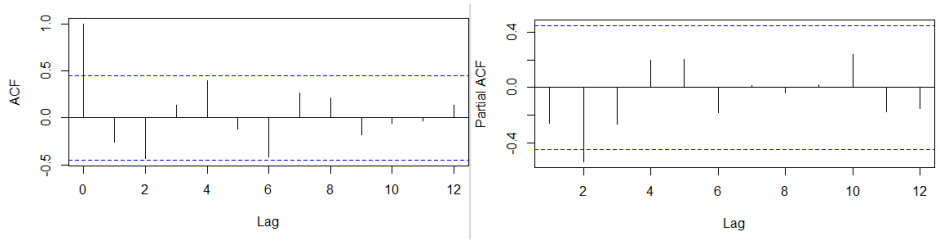




Όπως φαίνεται από το διάγραμμα η χρονοσειρά έχει εμφανή τάση. Αυτό φανερώνουν και οι για πολλές υστερήσεις υψηλές τιμές των συντελεστών ACF. Προχωράμε λοιπόν σε διαφόριση πρώτης τάξης.

Period	I(1)
1	
2	-235
3	-32
4	104
5	261
6	-357
7	180
8	230
9	-67
10	-142
11	75
12	319
13	-101
14	-25
15	252
16	5
17	123
18	-85
19	-1
20	125





Πλέον η χρονοσειρά μας έχει γίνει αρκετά στάσιμη και μπορούμε να επιλέξουμε μοντέλο πρόβλεψης. Δεδομένης της υψηλής τιμής του συντελεστή PACF για υστέρηση 2 και της ημιτονοειδούς κυμάτωσης των συντελεστών ACF καταλήγουμε στο μοντέλο ARIMA(2,1,0). Η γενική εξίσωση του μοντέλου θα είναι λοιπόν η ακόλουθη:

$$(1 - \varphi_1 B - \varphi_2 B^2)(1 - B)y_t \;\; = \;\; e_t$$

, όπου $\rho_1 = \frac{\varphi_1}{1-\varphi_2}$ και $\rho_2 = \frac{{\varphi_1}^2}{1-\varphi_2} + \varphi_2$. Σταθερά δεν εισάγουμε δεδομένης της ικανοποιητικής εικόνας που άφησε η διαφόριση.

Συνεπώς προκύπτει το σύστημα:

$$({\rho_1}^2 - 1){\varphi_2}^2 + (1 + \rho 2 - 2{\rho_1}^2)\varphi_2 + ({\rho_1}^2 - \rho_2) = 0$$

$$\varphi_1 = \rho_1(1 - \varphi_2)$$

Αυτό για ρ_1= -0.26 και ρ_2= -0.44 μας δίνει ϕ_1=-0.4 και ϕ2= -0.54. Παρατηρούμε ότι ισχύουν οι συνθήκες *-1<φ_2<1* , *φ_1 + φ_2 <1* και *φ_2 - φ_1 <1.*

Το τελικό μοντέλο θα έχει λοιπόν τη μορφή:

$$y_t = \; (1 + \varphi_1)y_{t-1} + (\varphi_2 - \varphi_1)y_{t-2} - \varphi_2 y_{t-3} + \; e_t \rightarrow$$

$$y_t = \; 0.6y_{t-1} - 0.14y_{t-2} + 0.54y_{t-3} + \; e_t$$

Observation	Data	I(1)	ARIMA(2,1,0)	et
1	4109	-	-	-
2	3874	-235	-	-
3	3842	-32	-	-
4	3946	104	3981.7	0.9
5	4207	261	3921.68	6.78
6	3850	-357	4046.44	5.1
7	4030	180	3851.86	4.42
8	4260	230	4150.78	2.56
9	4193	-67	4070.8	2.91
10	4051	-142	4095.6	1.1
11	4126	75	4143.98	0.44
12	4445	319	4172.68	6.13
13	4344	-101	4276.9	1.54
14	4319	-25	4212.14	2.47
15	4571	252	4383.54	4.1
16	4576	5	4483.7	2.02
17	4699	123	4437.92	5.56
18	4614	-85	4647.1	0.72
19	4613	-1	4581.58	0.68
20	4738	125	4659.3	1.66

Στην πραγματικότητα αν εφαρμοζόταν το μοντέλο ARIMA(3,1,0) οι προβλέψεις μας θα ήταν ακόμα καλύτερες. Ωστόσο αυτό δεν μπορεί να γίνει αντιληπτό μέσω της παρατήρησης των γραφημάτων παρά μόνο εφαρμόζοντας παραπλήσια μοντέλα με αυτό που επιλέξαμε αρχικά και ελέγχοντας τον δείκτη προσδοκώμενης πιθανοφάνειας αυτών. Η διαδικασία κατά την οποία προσθέτουμε μία παράμετρο παραπάνω σε ένα μοντέλο για να ελέγξουμε αν αυτή παράγει καλύτερα αποτελέσματα λέγεται **τεχνική υπερπροσαρμογή**. Η υπερπροσαρμογή πρέπει να γίνεται μόνο όταν τα συμπεράσματα που βγάζουμε από τις γραφικές των συντελεστών ACF και PACF είναι διφορούμενα ή όταν τα αποτελέσματα του t-test στον διαγνωστικό έλεγχο ξεπερνάνε τις επιθυμητές τιμές, δηλαδή υπάρχει σημαντική συσχέτιση μεταξύ των παραγόμενων σφαλμάτων.

Παράδειγμα 4:

Για την χρονοσειρά του παραδείγματος 3 να γίνει πρόβλεψη με ορίζοντα h=3 χρησιμοποιώντας μοντέλο πρόβλεψης ARIMA(1,1,1). Δίνονται οι συντελεστές φ_1=0.07 και ϑ_1=0.45. Στη συνέχεια να εξεταστεί η καταλληλότητα του μοντέλου.

Για την εξίσωση του μοντέλου ARIMA(1,1,1) έχουμε:

$$(1-\varphi_1 B)(1-B)y_t = (1-\theta_1 B)e_t \rightarrow$$

$$(1-B-\varphi_1 B+\varphi_1 B^2)y_t = (1-\theta_1 B)e_t \rightarrow$$

$$y_t = (1+\varphi_1)y_{t-1} - \varphi_1 y_{t-2} - \theta_1 e_{t-1} + e_t \rightarrow$$

$$y_t = 1.07y_{t-1} - 0.07y_{t-2} - 0.45\, e_{t-1} + e_t$$

Observation	Data	ARIMA(1,1,1)	et	APE (%)
1	4109	4109	0	
2	3874	3874	0	
3	3842	3857.55	-15.55	0.4
4	3946	3846.76	99.24	2.51
5	4207	3908.62	298.38	7.09
6	3850	4091	-241	6.26
7	4030	3933.46	96.54	2.4
8	4260	3999.16	260.84	6.12
9	4193	4158.72	34.28	0.82
10	4051	4172.88	-121.88	3.01
11	4126	4095.91	30.09	0.73
12	4445	4117.71	327.29	7.36
13	4344	4320.05	23.95	0.55
14	4319	4326.15	-7.15	0.17
15	4571	4320.47	250.53	5.48
16	4576	4475.9	100.1	2.19
17	4699	4531.31	167.69	3.57
18	4614	4632.15	-18.15	0.39
19	4613	4616.22	-3.22	0.07
20	4738	4614.38	123.62	2.61
21		4691.12		
22		4687.84		
23		4687.61		

Σε αυτό το σημείο υπενθυμίζουμε ότι για την πρόβλεψη τις χρονικές περιόδους 22 και 23 χρησιμοποιούμε ως δεδομένα τις ίδιες τις προβλέψεις του μοντέλου, δηλαδή τις τιμές 4691.12 και 4687.84 αντίστοιχα, με το σφάλμα εκεί να θεωρείται μηδενικό. Επίσης σημειώνουμε ότι όπως ήταν αναμενόμενο το προηγούμενο μοντέλο που μας το υπέδειξαν οι γραφικές των ACF και PACF σημείωσε στο κοινό δείγμα παρατηρήσεων 4-21 σαφώς καλύτερη προσαρμογή σε σχέση με το ARIMA(1,1,1) με MAPE 2.89% και 3.02% αντίστοιχα.

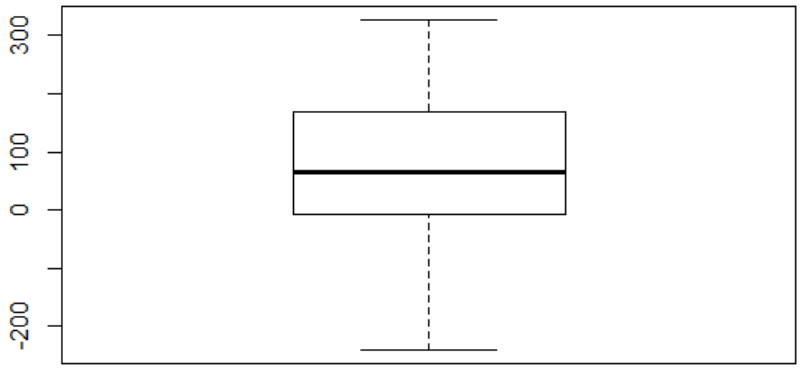
Για να εξακριβώσουμε την καταλληλότητα του μοντέλου μπορούμε να χρησιμοποιήσουμε πολλά στατιστικά εργαλεία. Αρχικά μπορούμε να ελέγξουμε τους ***συντελεστές αυτοσυσχέτισης των παραγόμενων σφαλμάτων*** για κάποιες από τις πιθανότερα ισχυρές υστερήσεις.

Υστέρηση	**ACF et**
1	-0.257
2	-0.533
3	0.262
4	0.295
5	-0.114

Παρατηρούμε ότι σε καμία υστέρηση ο συντελεστής δεν είναι αρκετά σημαντικός (>0.7). Αυτό είναι αρκετά καλό δείγμα για την προβλεπτική ικανότητα του μοντέλου αφού τα σφάλματα δεν συσχετίζονται μεταξύ τους, δηλαδή δεν ακολουθούν κάποιο συγκεκριμένο μοτίβο που θα δήλωνε ανικανότητα του μοντέλου να εκμεταλλευτεί την παρεχόμενη πληροφορία από τα δεδομένα.

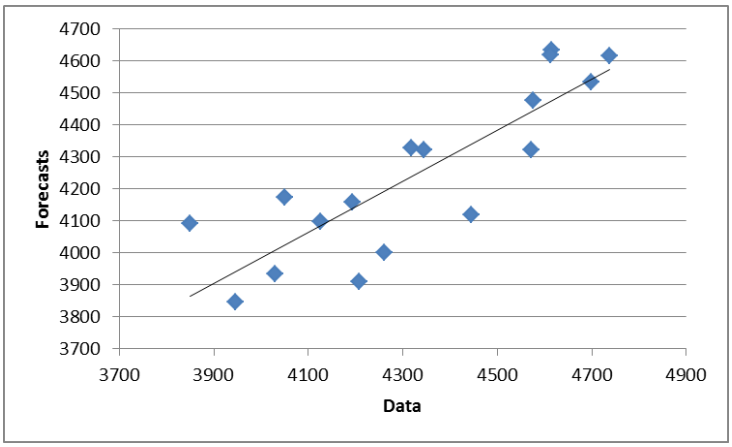
Οπτικά ο έλεγχος μπορεί να γίνει χρησιμοποιώντας διαγράμματα ***boxplot ή scatter***. Στα πρώτα, θεωρούμε ότι τα σφάλματά μας ακολουθούν μία κατανομή Poisson και συνεπώς οι τιμές τους μπορούν να χωρισθούν σε πέντε διακριτές περιοχές: Την ελάχιστη τιμή του δείγματος, το 25% των παρατηρήσεων, τον διάμεσο (median), το 75% των παρατηρήσεων και τη μέγιστη τιμή του δείγματος. Σε περίπτωση ύπαρξης outlier, δηλαδή ύπαρξη σφάλματος εκτός των ορίων της κατανομής, αυτό θα σημειωνόταν με μία κουκίδα εκτός των διαστημάτων που ορίζουν οι μέγιστες και ελάχιστες τιμές, ανάλογα τη φύση του (υπερβολικά απαισιόδοξη ή αισιόδοξη πρόβλεψη αντίστοιχα).

Όπως φαίνεται παρακάτω, οι αποστάσεις μεταξύ των τεσσάρων σημείων στην εν λόγω περίπτωση είναι σχεδόν ίσες (η κατανομή των σφαλμάτων προσεγγίζει την κανονική κατανομή) με τη διάμεσο να βρίσκεται αρκετά κοντά στο μηδέν. Με αυτό τον τρόπο δηλώνεται πρακτικά η αμεροληψία του μοντέλου αν και για να είμαστε απόλυτοι, στη πραγματικότητα βλέπουμε μέσω της διαμέσου ότι τα περισσότερα σφάλματα είναι ελαφρώς θετικά και συνεπώς το μοντέλο μας είναι σχετικά απαισιόδοξο. Γενικά όσο η διάμεσος πλησιάζει το 75% των τιμών (πάνω όριο στο κουτί), τόσα περισσότερα σφάλματα είναι θετικής τιμής και τόσο πιο απαισιόδοξο γίνεται το μοντέλο. Αντίθετα, απαισιόδοξο γίνεται όταν περισσότερα αρνητικά σφάλματα προκύπτουν με τα όρια του boxplot να μετατοπίζονται προς μικρότερες τιμές.



Boxplot για το μοντέλο ARIMA(1,1,1)

Στην περίπτωση του διαγράμματος scatter, αντιστοιχίζουμε τις τιμές των δεδομένων (άξονας x) με τις τιμές που προβλέψαμε (άξονας y). Αν οι τιμές βρίσκονται στο πάνω μισό του χώρου που θα όριζε η συνάρτηση y=x, τότε το μοντέλο είναι αισιόδοξο. Στην αντίθετη περίπτωση είναι απαισιόδοξο. Εδώ οι τιμές είναι σχετικά διαμοιρασμένες γύρω από τη νοητή ευθεία y=x, οπότε και το μοντέλο μας μπορεί να θεωρηθεί αμερόληπτο.



Scatter plot για το μοντέλο ARIMA(1,1,1)

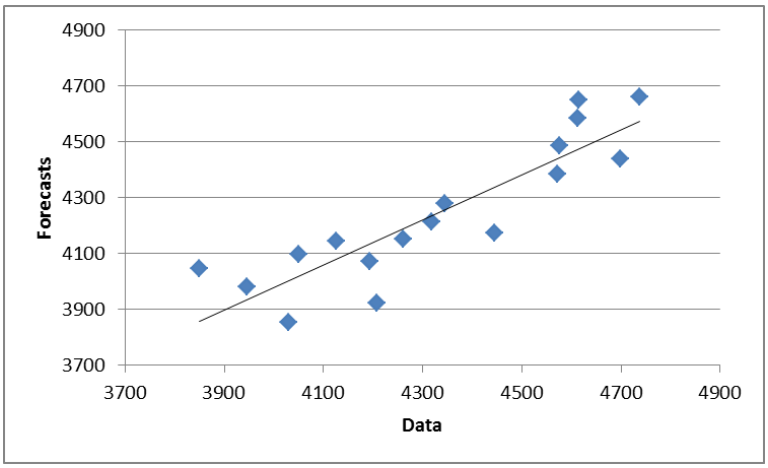
Τέλος, μπορούμε να ελέγξουμε τη ***στατιστική σημαντικότητα των υπολειπόμενων σφαλμάτων*** μέσω του δείκτη του t-test, όπου ιδανικά θα πρέπει να ισούται με μηδέν (πλήρως ασυσχέτιστα σφάλματα). Υπενθυμίζουμε ότι κανονικά η τιμή πρέπει να είναι μικρότερη του z=1.96 (ή 2 διακυμάνσεων πιο χοντρικά) για 95% διάστημα εμπιστοσύνης, αλλά στη πράξη πρέπει να είναι μικρότερη του 1.25 για υστέρηση 1, 2 και 3, και μικρότερη του 1.6 για μεγαλύτερες υστερήσεις. Παρακάτω υπολογίζουμε τον στατιστικό δείκτη για τις αντίστοιχες υστερήσεις. Όπως φαίνεται, η απαίτηση δεν ικανοποιείται παντού και μάλιστα βρισκόμαστε σημαντικά πάνω από τα όρια για υστέρηση 2. Υποψιαζόμαστε λοιπόν ότι ίσως ένα μοντέλο ARIMA (2,1,0) ή ARIMA (2,1,1) θα μπορούσε να δώσει καλύτερα αποτελέσματα από το ARIMA(1,1,1), όπως άλλωστε αποδείχτηκε και νωρίτερα μέσω του MAPE.

ARIMA(1,1,1)					
Lag	**ACFi**	**$ACFi^2$**	**$\Sigma ACFi^2$**	**$Sr_i(e)$**	***tr_i***
1	-0.257	0.066	0.000	0.236	-1.090
2	-0.533	0.284	0.066	0.251	-2.123
3	0.262	0.069	0.350	0.307	0.853
4	0.295	0.087	0.418	0.319	0.923
5	-0.114	0.013	0.505	0.334	-0.340

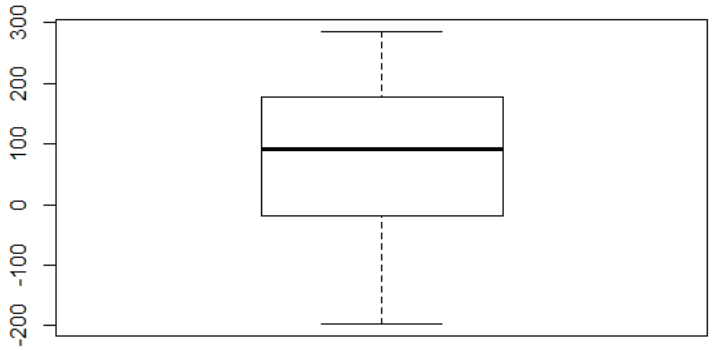
Πράγματι, στο Παράδειγμα 3 είχαμε επιλέξει βάση των διαγραμμάτων ACF και PACF το μοντέλο ARIMA (2,1,0). Ας δούμε την εικόνα που παρουσιάζει αυτό στατιστικά. Όπως φαίνεται παρακάτω, η σημαντικότητα του μοντέλου έχει βελτιωθεί για υστέρηση μεγαλύτερη του 2 αλλά είναι σημαντική για υστέρηση 1 παρόλο που οι δείκτες ACF διατηρούνται αυτοί καθαυτοί σε χαμηλά επίπεδα (<0.5). Βέβαια σαν συνολική εικόνα το εν λόγω μοντέλο είναι σαφώς καλύτερο από το ARIMA(1,1,1). Αν πραγματοποιούσαμε την ίδια διαδικασία για το μοντέλο ARIMA(3,1,0) (το οποίο όπως αναφέρθηκε στο παράδειγμα 3 θα ήταν και η βέλτιστη επιλογή μας με χρήση του κριτηρίου AIC) η εικόνα το σφαλμάτων θα γινόταν ακόμα καλύτερη.

ARIMA(2,1,0)					
Lag	**ACFi**	**$ACFi^2$**	**$\Sigma ACFi^2$**	**$Sr_i(e)$**	***tr_i***
1	-0.481	0.232	0.000	0.236	-2.042
2	0.116	0.013	0.232	0.285	0.405
3	-0.058	0.003	0.245	0.288	-0.203
4	0.044	0.002	0.248	0.288	0.154
5	0.133	0.018	0.250	0.289	0.462

Από άποψη αμεροληψίας, το διάγραμμα scatter για το μοντέλο ARIMA(2,1,0) εμφανίζει περίπου την ίδια εικόνα με πρίν, ωστόσο το διάγραμμα boxplot δηλώνει μία τάση μεγαλύτερης απαισιοδοξίας (η διάμεσος απέχει περισσότερο από το μηδέν).



Scatter plot για το μοντέλο ARIMA(2,1,0)



Boxplot για το μοντέλο ARIMA(2,1,0)

Πράγματι, το μέσο σφάλμα ME για το μοντέλο ARIMA (2,1,0) ισούται με 86.14, ενώ πριν (για τις ίδιες παρατηρήσεις) βρισκόταν στο 83.6. Ωστόσο έχει βελτιωθεί η ακρίβειά μας με το μέσο απόλυτο σφάλμα MAE να βρίσκεται στο 124.7 από 129.64.Βλέπουμε λοιπόν ότι ανάλογα με το κριτήριο που θα χρησιμοποιούσαμε και τον σκοπό μας (ελαχιστοποίηση ακρίβειας ή προκατάληψης) οι αποφάσεις μας θα ήταν εντελώς διαφορετικές.

Η αξιολόγηση των επιλογών μας και η τελική επιλογή μοντέλου θα μπορούσε να γίνει φυσικά όπως προείπαμε και με χρήση των κριτηρίων που έχουν αναλυθεί στη θεωρεία. Οι τιμές των κριτηρίων παρουσιάζονται παρακάτω.

Model	**LogLik**	**AIC**	**AICc**	**BIC**
ARIMA (2,1,0)	-109.51	227.03	228.88	235.53
ARIMA (1,1,1)	-117.18	240.35	242.07	249.03

Σε όλες τις περιπτώσεις, το ARIMA(2,1,0) φαίνεται να είναι η καλύτερη δυνατή επιλογή τόσο από άποψη προσαρμογής αλλά και πολυπλοκότητας.