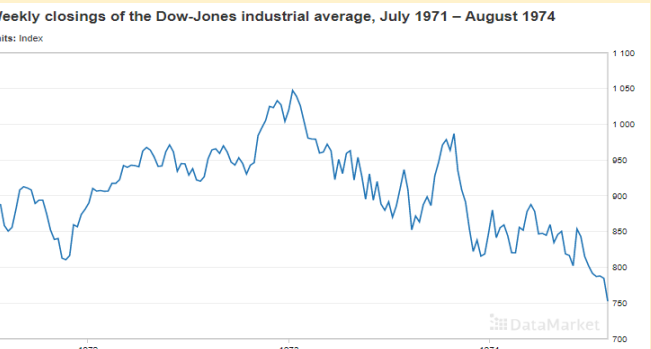




Εισαγωγή στις χρονοσειρές

Ος **χρονοσειρά (time series)** ορίζεται μια σειρά από παρατηρήσεις ενός φαινομένου οι οποίες γίνονται σε συγκεκριμένες χρονικές στιγμές (π.χ. ανά έτος, τρίμηνο, ημέρα) οι οποίες ισαπέχουν μεταξύ τους και εκφράζουν τη εξέλιξη του φαινομένου στο χρόνο.

Μια χρονοσειρά αποτελείται από ένα σύνολο παρατηρήσεων $y_1,y_2,...,y_n$, όπου n το πλήθος των μετρήσεων στα αντίστοιχα ισαπέχοντα χρονικά διαστήματα. Οι παρατηρήσεις $y_1,y_2,...,y_n$ είναι συγκεκριμένες τιμές των τυχαίων μεταβλητών $Y_1, Y_2 ,...,Y_n$. Η ακολουθία αυτή τυχαίων μεταβλητών ονομάζεται **στοχαστική διαδικασία** και συμβολίζεται με $\{Y_n\}$.



Σχήμα 1: Χρονοσειρά δείκτη Dow-Jones ανά εβδομάδα (Roberts, 1992)

Βασικά στατιστικά μεγέθη

- Η ροπή 1ης τάξης ή αλλιώς **μέση τιμή**:

$$\mu=E[Y]=\int_{-\infty}^{\infty} y_t f_Y(y,t)dy$$

- Η **αυτοσυνδιακύμανση** (autocovariance) :

$$\gamma(t_1,t_2)=E[(Y_{t_1}-\mu_{t_1})(Y_{t_2}-\mu_{t_2})]=E[(Y_{t_1}-Y_{t_2})-\mu_{t_1}\mu_{t_2}]=\kappa(t_1,t_2)-\mu_{t_1}\mu_{t_2}$$

όπου $\kappa(t_1,t_2)=E[Y_{t_1}Y_{t_2}]=\int_{-\infty}^{\infty} y_1y_2f_Y(y_1,y_2,t_1,t_2)dy_1dy_2$.

- Η **διασπορά**:

για $t=t_1=t_2$ ορίζεται η διασπορά $\sigma_t^2\equiv\sigma_{t,t}^2=E[(Y_t-\mu_t)^2]$.

- Η **αυτοσυσχέτιση** (autocorrelation):

$$\rho_t\equiv\rho(t)=\frac{\gamma(t)}{\gamma(0)},\quad\text{όπου }\gamma(0)=\sigma_t^2$$

Χαρακτηριστικά μεγέθη των χρονοσειρών

- Τάση**: Η μακροχρόνια αύξηση ή μείωση των τιμών μιας μεταβλητής που παρατηρείται κατά τη διάρκεια μιας χρονικής περιόδου.

- Εποχικότητα**: Η χρονοσειρά παρουσιάζει ίδια συμπεριφορά ανά περίοδο d π.χ. μήνας, τρίμηνο, χρόνος.

- Ασταθής διασπορά**: Η διασπορά της χρονοσειράς δεν είναι σταθερή για κάθε t.

- Ακραίες τιμές (outliers)**: Τιμές οι οποίες απέχουν σημαντικά από τις υπόλοιπες παρατηρήσεις. Συνήθως οφείλονται σε κάποιο απόρρολεπτο παράγοντα και δημιουργούν πρόβλημα στη μοντελοποίηση.

Γενικό προσθετικό μοντέλο χρονοσειράς με τάση και εποχικότητα:

$$Y_t=\mu_t+s_t+X_t$$

Βασικά μοντέλα γραμμικών χρονοσειρών

Τα βασικά μοντέλα που χρησιμοποιούνται στην ανάλυση χρονοσειρών, οι οποίες παρουσιάζουν γραμμικές συσχετίσεις μεταξύ των τιμών τους είναι το μοντέλο αυτοπαλινδρόμησης (AR), το μοντέλο κινητού μέσου (MA), ο συνδυασμός αυτών των δυο μοντέλων (ARMA) και τα εποχικά ολοκληρωμένα αυτοπαλινδρόμενα μοντέλα (SARIMA).

- Μοντέλα αυτοπαλινδρόμησης AR(p)**:

$$Y_t=\varphi_1Y_{t-1}+\varphi_2Y_{t-2}+...+\varphi_pY_{t-p}+Z_t$$

όπου $Z_t\sim\text{WN}(0,\sigma_z^2)$.

- Μοντέλα κινητού μέσου MA(q)**:

$$Y_t=Z_t-\theta_1Z_{t-1}-\theta_2Z_{t-2}-\dots-\theta_qZ_{t-q}$$

όπου $Z_t\sim\text{WN}(0,\sigma_z^2)$.

- Μοντέλα ARMA(p,q)**:

$$Y_t=\varphi_1Y_{t-1}+\varphi_2Y_{t-2}+...+\varphi_pY_{t-p}+Z_t-\theta_1Z_{t-1}-\theta_2Z_{t-2}-\dots-\theta_qZ_{t-q}$$

όπου $Z_t\sim\text{WN}(0,\sigma_z^2)$.

- Εποχικά ολοκληρωμένα αυτοπαλινδρομούμενα μοντέλα SARIMA(p,d,q)×(P,D,Q)ₛ**

$$\varphi(B)[\Phi(B)(1-B)^D](1-B)^dY_t=\theta(B)\Theta(B)^DZ_t$$

όπου $Z_t\sim\text{WN}(0,\sigma_z^2)$.

Μεθοδολογία Box-Jenkins

- Ταυτοποίηση μοντέλου**

Από το διάγραμμα της PACF προσδιορίζουμε το πλήθος των AR όρων του μοντέλου, ενώ από το διάγραμμα της ACF προσδιορίζουμε το πλήθος των MA όρων του μοντέλου. Τα διαγράμματα γίνονται αφού έχει μετασχηματιστεί η χρονοσειρά από μη στάσιμη σε στάσιμη.

- Εκτίμηση των παραμέτρων**

Η εκτίμηση των παραμέτρων στα ARMA μοντέλα γίνεται είτε με τη μέθοδο των ροπών, είτε με τη μέθοδο των ελαχίστων τετραγώνων (least square method). Η πρώτη γίνεται με βάση την εκτίμηση της αυτοσυσχέτισης, ενώ η δεύτερη με την ελαχιστοποίηση του σφάλματος προσαρμογής.

- Διανυστικός έλεγχος**

Εξετάζουμε την καταλληλότητα του προτεινόμενου μοντέλου, ελέγχοντας την ανεξαρτησία των υπολοίπων (residuals) και την κανονικότητά τους.

Στην περίπτωση που δύο ή περισσότερα μοντέλα είναι ικανοποιητικά, επιλέγουμε εκείνο που ελαχιστοποιεί την τιμή του κριτηρίου AIC ή BIC.

$$AIC(p,q)=N\ln(\hat{\sigma}_e^2)+2(p+q)$$

$$BIC(p,q)=\log(\hat{\sigma}_e^2)+\frac{(p+q)\ln(N)}{N}$$

όπου N: διαθέσιμος αριθμός δεδομένων

ΠΡΟΗΓΜΕΝΕΣ ΜΕΘΟΔΟΙ ΠΡΟΒΛΕΨΗΣ ΧΡΟΝΟΣΕΙΡΩΝ

Χ. ΣΤΑΜΑΤΟΠΟΥΛΟΣ, Σ. ΜΑΛΕΦΑΚΗ

Department of Mechanical Engineering and Aeronautics, University of Patras, Greece

p: η τάξη του AR μοντέλου
q: η τάξη του MA μοντέλου
 $\hat{\sigma}_e^2$:η εκτίμηση της διασποράς λευκού θορύβου

- Πρόβλεψη και σφάλμα πρόβλεψης**

Το τελικό βήμα που είναι και ο στόχος της όλης ανάλυσης, είναι η πρόβλεψη κάποιων μελλοντικών τιμών της χρονοσειράς (forecasting). Η πρόβλεψη για κ χρονικές στιγμές συμβολίζεται με $y_n(k)$ ενώ η πραγματική τιμής της χρονοσειράς εκείνη τη χρονική στιγμή συμβολίζεται με y_{n+k} . Το σφάλμα της πρόβλεψης (prediction error) θα είναι:

$$\varepsilon_n(k)=Y_{n+k}-Y_n(k)$$

Για να διακρίνουμε μεταξύ δύο μοντέλων ποιο δίνει την καλύτερη πρόβλεψη μπορούμε να χρησιμοποιήσουμε το μέσο τετραγωνικό σφάλμα (MSE), το μέσο απόλυτο σφάλμα (MAD) ή το μέσο ποσοστιαίο απόλυτο σφάλμα (MAPE).

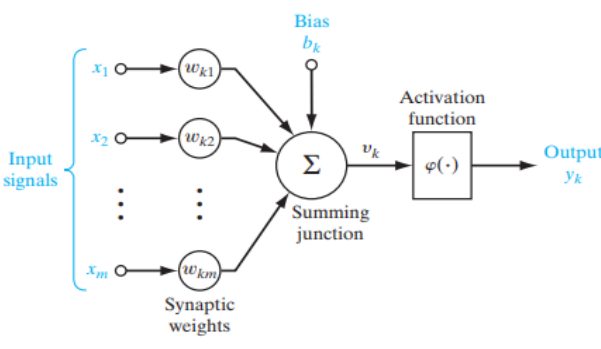
$$MAD=\frac{1}{n}\sum_{i=1}^n|e_i|$$

$$MSE=\frac{1}{n}\sum_{i=1}^n(e_i)^2$$

$$MAPE=\frac{1}{n}\sum_{i=1}^n|\frac{e_i}{Y_i}|100\%$$

Μοντελοποίηση μη γραμμικών χρονοσειρών μέσω τεχνητών νευρωνικών δικτύων (artificial neural netwroks)

Τα νευρωνικά δίκτυα αποτελούν ένα δίκτυο νευρώνων σε αντιποichία με τη λειτουργία του ανθρώπινου εγκεφάλου. Ένας τυπικός τεχνητός νευρώνας αποτελείται από τα σήματα εισόδου (input layer) $x_1, x_2, \dots, x_m$. Κάθε μια από αυτές τις εισόδους πολλαπλασιάζεται με ένα συντελεστή, ο οποίος συμβολίζεται με w_{jk} (weight) και ονομάζεται συναπτικό βάρος. Το κ συμβολίζει τον νευρώνα του δικτύου. Όλες οι εισοδοι πολλαπλασιάζονται με τα αντίστοιχα συναπτικά τους βάρη και εν συνεχεία αθροίζονται όλα μαζί σε έναν γραμμικό αθροιστή (linear combiner). Σε αυτό το άθροισμα προστίθεται και η μεροληψία (bias), η οποία συμβολίζεται με w_{jk} και έχει τιμή εισόδου ίση με τη μονάδα και συναπτικό βάρος b. Το συνολικό αυτό άθροισμα ονομάζεται τοπικό πεδίο (net input) και συμβολίζεται με v_k . Το τοπικό πεδίο εισέρχεται σε μια μη γραμμική συνάρτηση, η οποία ονομάζεται συνάρτηση ενεργοποίησης (activation function), και έτσι προκύπτει η έξοδος y_k του νευρώνα k. Η παραπάνω τυπική διαδικασία για ένα τεχνητό νευρώνα παρουσιάζεται γραφικά στο Σχήμα 2.



Σχήμα 2: Γραφική απεικόνιση τεχνητού νευρώνα k.

Η συνάρτηση ενεργοποίησης είναι συνήθως η σιγμοειδής συνάρτηση (sigmoid function). Τα βασικά παραδείγματα σιγμοειδών συναρτήσεων είναι η λογιστική συνάρτηση (logistic function) και η υπερβολική εφαπτομενική (hyperbolic tangent). Η λογιστική συνάρτηση ορίζεται ως:

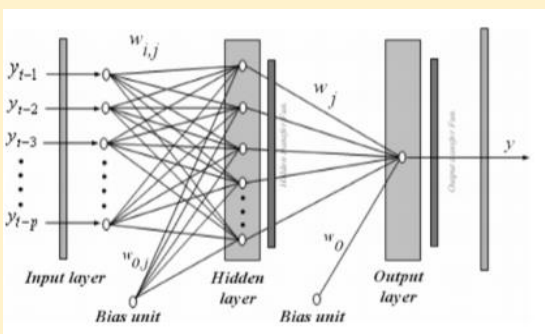
$$\Phi(v)=\frac{1}{1+\exp(-av)},$$

όπου a είναι η παράμετρος κλίσης της σιγμοειδής συνάρτησης (slope parameter) και η οποία έχει ως σύνολο τιμών το [0,∞). Το σύνολο τιμών της λογιστικής είναι το [0,1], ενώ της υπερβολικής εφαπτομενικής το [-1,1].

Μερικές αρχιτεκτονικές δικτύων:

- δίκτυα εμπρόσθιας τροφοδότησης ενός επιπέδου (Single-Layer Feedforward Networks)
- τα δίκτυα εμπρόσθιας τροφοδότησης πολλαπλών επιπέδων (Multilayer Feedforward Networks)
- τα αναδρομικά δίκτυα (Recurrent Networks).

Ένα δίκτυο πολυεπίπεδου αισθητήρα με 1 κρυφό επίπεδο παρουσιάζεται στο Σχήμα 3.



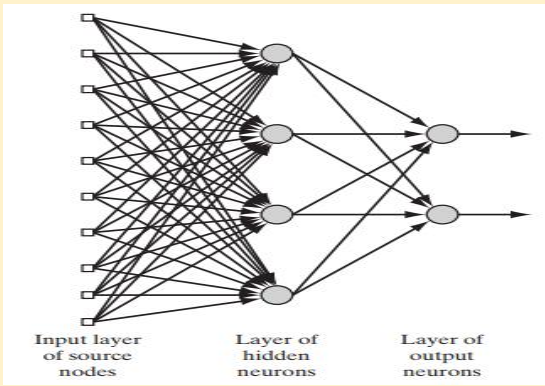
Σχήμα 3: Γραφική απεικόνιση δικτύου με ένα κρυφό επίπεδο, 4 νευρώνες στο κρυφό επίπεδο, 10 εισόδους και 2 εξόδους (10x4x2)

Στη μοντελοποίηση και πρόβλεψη χρονοσειρών τα πιο ευρέως διαδοθέντα δίκτυα είναι τα δίκτυα ενός κρυφού επιπέδου και εμπρόσθιας τροφοδότησης (Single hidden layer feedforward network). Η σχέση μεταξύ εισόδου (input layer) και εξόδου (output layer) παρουσιάζεται γραφικά στο Σχήμα 3 και δίνεται από τι σχέση :

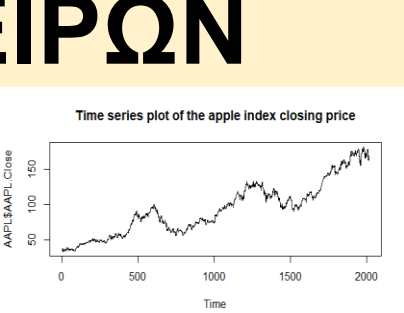
$$Y_t=w_0+\sum_{j=1}^q\varphi_jg(w_0+\sum_{i=1}^p\varphi_iY_{t-i}^p+Y_{t-i})+v_t,$$

όπου w_j για $j=0, 1,..., q$ και w_{ij} για $i=0, 1,...,p$ και $j=1, 2,..., q$, είναι οι παράμετροι του μοντέλου, δηλαδή τα συναπτικά βάρη. Με ρ συμβολίζονται οι κόμβοι εισόδου και q το πλήθος των κρυφών επιπέδων.

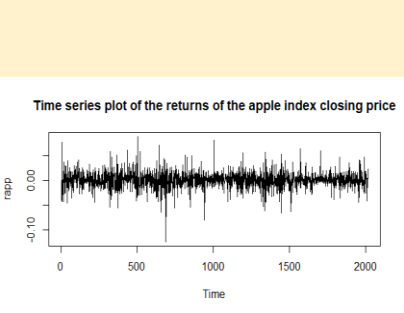
Στην πρόβλεψη είναι πολύ σημαντικό να καθορίσουμε ένα σύνολο από τις πιο πρόσφατες τιμές της χρονοσειράς, οι οποίες θα επιηρεάσουν την πρόβλεψη και από αυτές αυτές από τις οποίες θα καθορισθεί ο αριθμός των κόμβων εισόδου, δηλαδή το input layer. Εκτός από το επίπεδο εισόδου, πρέπει να ορισθεί ο αλγόριθμος μάθησης, ο οποίος στην παρούσα εργασία είναι ο αλγόριθμος οπισθοδιάδοσης του σφάλματος (error backpropagation algorithm). Τέλος, πρέπει να καταλήξουμε στο βέλτιστο αριθμό κόμβων στο κρυφό επίπεδο. Η συνάρτηση ενεργοποίησης του κάθε νευρώνα είναι η σιγμοειδής συνάρτηση, ενώ το δίκτυο δίνει μια τιμή απόκρισης, δηλαδή κάνει πρόβλεψη 1-βήματος μπροστά.



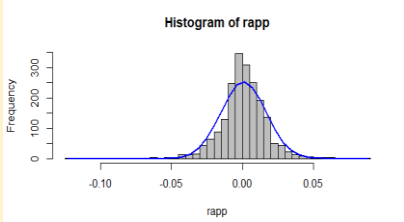
Σχήμα 4: Δίκτυο ενός κρυφού επιπέδου, εμπρόσθιας τροφοδότησης



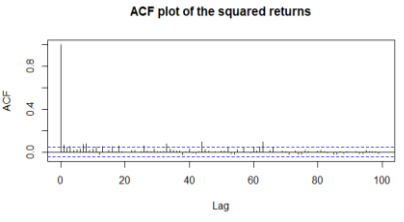
Σχήμα 5: Διάγραμμα χρονοσειράς τιμής κλεισίματος της APPLE από την 01-05-2010 έως και την 2-05-2018



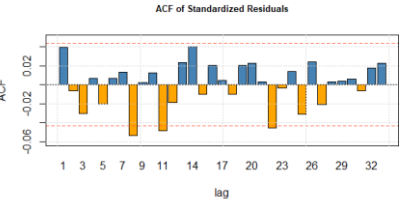
Σχήμα 6: Διάγραμμα χρονοσειράς των αποδόσεων.



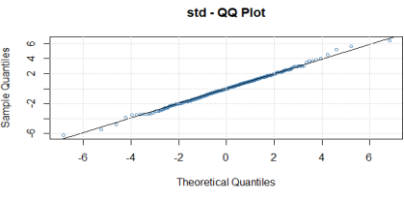
Σχήμα 7: Κατανομή των αποδόσεων σε σύγκριση με την κανονική κατανομή.



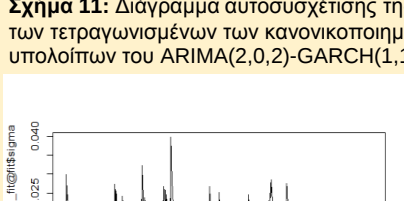
Σχήμα 8: Διάγραμμα αυτοσυσχέτισης της χρονοσειράς των τετραγώνων των αποδόσεων.



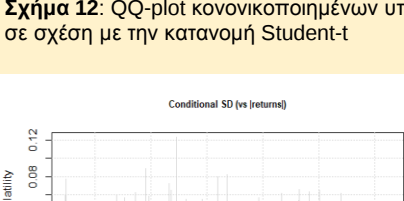
Σχήμα 9: Διάγραμμα μερικής αυτοσυσχέτισης της χρονοσειράς των τετραγώνων των αποδόσεων.



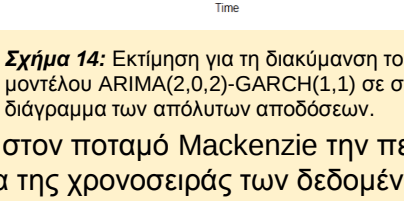
Σχήμα 10: Διάγραμμα αυτοσυσχέτισης της χρονοσειράς των κανονικοποιημένων υπολοίπων του ARIMA(2,0,2)-GARCH(1,1).



Σχήμα 11: Διάγραμμα αυτοσυσχέτισης της χρονοσειράς των τετραγώνων των κανονικοποιημένων υπολοίπων του ARIMA(2,0,2)-GARCH(1,1).

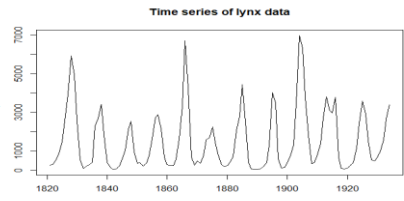


Σχήμα 12: QQ-plot κανονικοποιημένων υπολοίπων σε σχέση με την κατανομή Student-t

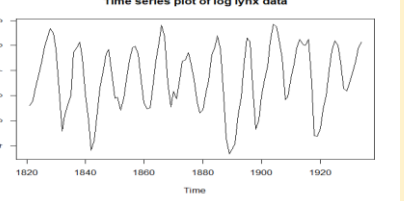


Σχήμα 13: Εκτίμηση για τη διακύμανση του υβριδικού μοντέλου ARIMA(2,0,2)-GARCH(1,1).

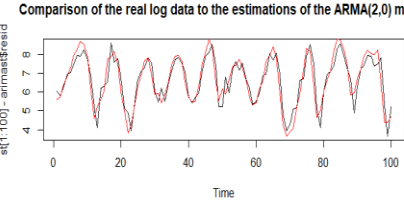
- Εφαρμογή για τον ετήσιο αριθμό του канаδικού λούκου στον ποταμό Mackenzie την περίοδο 1821-1934. Στο Σχήμα 15 παρουσιάζεται το διάγραμμα της χρονοσειράς των δεδομένων, ενώ στο Σχήμα 16 το διάγραμμα των λογαριθμισμένων δεδομένων. Ένα AR(2) μοντέλο είναι το καταλληλότερο για τη μοντελοποίηση του γραμμικού μέρους των δεδομένων. Στο Σχήμα 17 παρουσιάζεται η εκτίμησης του AR(2) μοντέλου για το σύνολο εκπαίδευσης (training data set), ενώ στο Σχήμα 18 οι προβλέψεις του μοντέλου σε σύγκριση με το σύνολο ελέγχου (test data set). Για τη μοντελοποίηση του μη γραμμικού μέρους χρησιμοποιήθηκε ένα νευρωνικό δίκτυο με 5 κόμβους εισόδου, 5 νευρώνες στο κρυφό επίπεδο και 1 έξοδο ANN(5x5x1). Οι εκτιμήσεις του νευρωνικού δικτύου φαίνονται στο Σχήμα 19 ενώ οι εκτιμήσεις για το σύνολο ελέγχου στο Σχήμα 20. Οι εκτιμήσεις του υβριδικού μοντέλου ARIMA(2,0,0)-ANN(4x4x1) για το σύνολο εκπαίδευσης και οι εκτιμήσεις για το σύνολο ελέγχου παρουσιάζονται στα Σχήματα 21,22. Συμπερασματικά το υβριδικό μοντέλο αποδείχθηκε το αποτελεσματικότερο, αφού εκμεταλλεύτηκε τα πλεονεκτήματα τόσο των ARIMA μοντέλων για τη μοντελοποίηση της γραμμικής συνιστώσας, όσο και των νευρωνικών δικτύων για τη μοντελοποίηση των εναπομεινανών μη γραμμικοτήτων των δεδομένων.



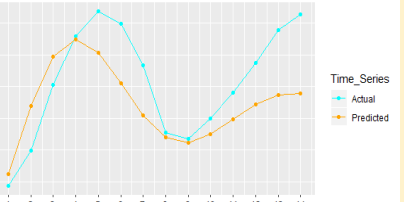
Σχήμα 15: Διάγραμμα χρονοσειράς канаδικού λούκου.



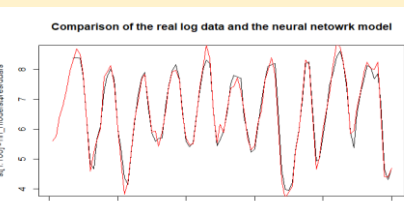
Σχήμα 16: Διάγραμμα χρονοσειράς λογαριθμισμένων δεδομένων.



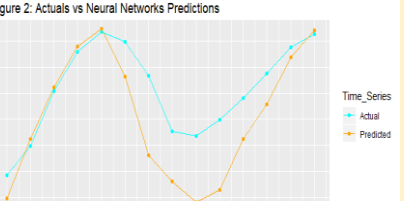
Σχήμα 17: Εκτίμηση του AR(2) μοντέλου για το σύνολο εκπαίδευσης.



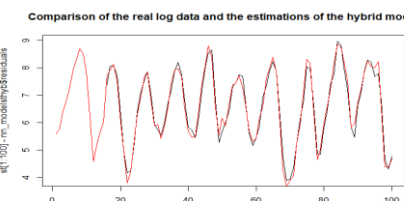
Σχήμα 18: Εκτίμηση του AR(2) μοντέλου για το σύνολο ελέγχου.



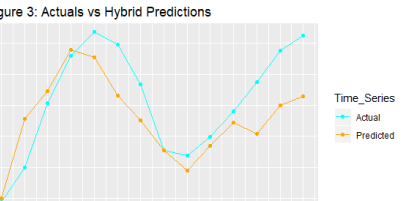
Σχήμα 19: Εκτίμηση του νευρωνικού (5x5x1) για το σύνολο εκπαίδευσης.



Σχήμα 20: Εκτίμηση του νευρωνικού (5x5x1) για το σύνολο ελέγχου.

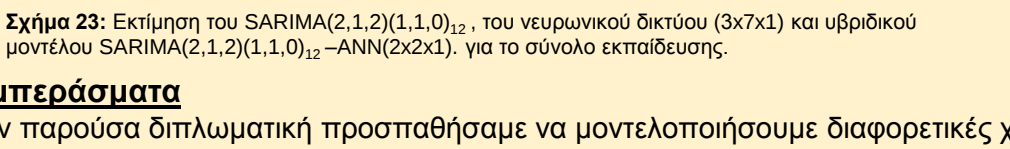
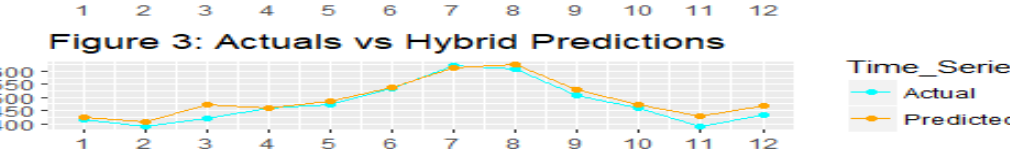
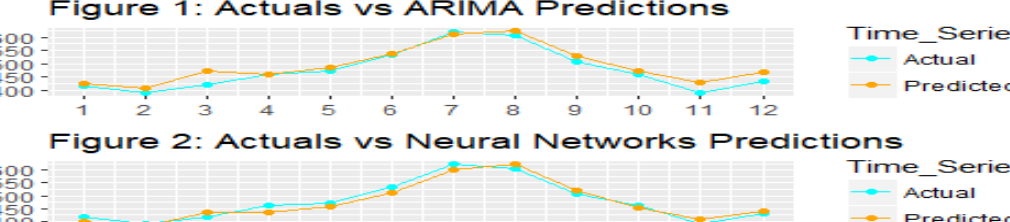


Σχήμα 21: Εκτίμηση του υβριδικού μοντέλου ARIMA(2,0,0)-ANN (4x4x1) για το σύνολο εκπαίδευσης.



Σχήμα 22: Εκτίμηση του υβριδικού μοντέλου ARIMA(2,0,0)-ANN (4x4x1) για το σύνολο ελέγχου.

- Εφαρμογή για τα συνολικά αριθμό των επιβατών των αεροπορικών εταιρειών παγκοσμίως σε μηνιαία βάση από το 1949 έως και το 1960. Για πρόβλεψη αρχικά χρησιμοποιήθηκε το μοντέλο SARIMA(2,1,2)(1,1,0)12. Για την αρχιτεκτονική του νευρωνικού δικτύου δοκιμάσθηκαν διάφοροι συνδυασμοί (cross validation) και βέλτιστη αρχιτεκτονική σύμφωνα με τα κριτήρια RMSE, MAE και MAPE είχε το νευρωνικό (3x7x1), το οποίο χρησιμοποιήθηκε για πρόβλεψη. Τέλος, χρησιμοποιήθηκε και το υβριδικό μοντέλο SARIMA(2,1,2)(1,1,0)12-ANN(2x2x1). Σύμφωνα με τα κριτήρια RMSE, MAE και MAPE, η μικρότερη τιμή των στατιστικών σφαλμάτων προκύπτει για το απλό νευρωνικό δίκτυο. Το υβριδικό μοντέλο για αυτό το σύνολο δεδομένων δεν είναι αποδοτικότερο από το απλό νευρωνικό δίκτυο, αντίθετα δίνει λιγότερο αποτελεσματικές προβλέψεις. Οπότε συμπεραίνουμε ότι τα υδρικά μοντέλα δεν είναι πάντα η καλύτερη επιλογή. Στο Σχήμα 23 παρουσιάζονται γραφικά οι εκτιμήσεις για το σύνολο ελέγχου του κάθε μοντέλου.



Σχήμα 23: Εκτίμηση του SARIMA(2,1,2)(1,1,0)12, του νευρωνικού δικτύου (3x7x1) και υβριδικού μοντέλου SARIMA(2,1,2)(1,1,0)12-ANN(2x2x1). για το σύνολο εκπαίδευσης.

Συμπεράσματα

Στην παρούσα διπλωματική προσπάθησαμε να μοντελοποιήσουμε διαφορετικές χρονοσειρές, η κάθε μια εκ των οποίων είχε τα δικά της χαρακτηριστικά. Παρατηρήσαμε ότι όταν υπάρχουν γραμμικές μόνο εξαρτήσεις, ένα τυπικό μοντέλο ARIMA/SARIMA μπορεί να δώσει ικανοποιητικά αποτελέσματα. Στη περίπτωση που τα δεδομένα μας παρουσιάζουν μη γραμμικά στοιχεία, τα νευρωνικά δίκτυα μπορεί να αποδείχθούν αρκετά αποδοτικά. Στη γενικότερη περίπτωση που τα δεδομένα αποτελούνται τόσο από γραμμικά όσο και μη γραμμικά στοιχεία, ενδέχεται ένα υβριδικό μοντέλο ARIMA-ANN να δίνει τα καλύτερα αποτελέσματα. Τέλος, στις περιπτώσεις που υπάρχει η ειδική μορφή μη γραμμικότητας η οποία αναφέρεται στη βιβλιογραφία ως volatility clustering και παρουσιάζεται κυρίως σε χρηματοοικονομικές εφαρμογές, τα μοντέλα ARCH-GARCH είναι τα πλέον κατάλληλα, λόγω των ιδιαίτερητων που παρουσιάζουν αυτού του είδους τα δεδομένα ως προς την συμμετρία και την κώντωση της κατανομής τους.

Βιβλιογραφία

- [1] C. Chatfield (1984), "The Analysis of time series: An Introduction (Sixth Edition)", Chapman & Hall
- [2] Spyros Makridakis, Michele Hibon (1997), "ARMA models and the Box-Jenkins methodology", Journal of forecasting, Vol. 16, 147-163
- [3] Δημήτρης Κουνιουμήζης (2014), "ΑΝΑΛΥΣΗ ΧΡΟΝΟΣΕΙΡΩΝ", Αριστοτέλειο Παν/μιο Θεσσαλονίκης
- [4] G. P. Zhang (2003), "Time series forecasting using a hybrid ARIMA and neural network model", Neurocomputing 159 – 175
- [5]Simon Haykin, "Neural Networks and Learning Machines", Third Edition, McMaster University Hamilton, Ontario, Canada
- [6] G. Peter Zhang (2003), "Time series forecasting using a hybrid ARIMA and neural network model", Department of Management, J. Mack Robinson College of Business, Georgia State University, University Plaza, Atlanta, Neurocomputing 50, pp. 159 – 175.
- [7] Tim Bollerslev (1986), "GENERALIZED AUTOREGRESSIVE CONDITIONAL HETEROSKEDASTICITY", Journal of Econometrics 31 307-327.
- [8] Ruey R. Tsay, "Analysis of Financial Time Series", University of Chicago