Πανεπιστήμιο Μακεδονίας

Δ.Π.Μ.Σ. Στα Πληροφοριακά Συστήματα

Ταξινόμηση και ομαδοποίηση χρηματοοικονομικών χρονικών σειρών με μοντέλα GARCH

Διπλωματική Εργασία του Βάιου Βίλλη

Επιβλέπων καθηγητής: Δ. Παπαναστασίου

Θεσσαλονίκη (Φεβρουάριος 2010)

Ευχαριστώ τον επιβλέποντα καθηγητή κ. Δημ. Παπαναστασίου για την βοήθειά του, καθώς και την οικογένεια μου για την στήριξη που μου παρείχε.

Περιεχόμενα

Περίληψη

Οι χρηματοοικονομικές χρονικές σειρές παρουσιάζουν συχνά παρόμοιες δομές αστάθειας. Η επιλογή τέτοιων σειρών ώστε να παρουσιάζουν παρόμοια συμπεριφορά μπορεί να είναι αρκετά σημαντική για την ανάλυση των μηχανισμών μεταφοράς της αστάθειας και στην πρόβλεψη των χρονοσειρών, χρησιμοποιώντας ως εργαλείο σειρές με παρόμοια δομή. Στο πρώτο κεφάλαιο αναπτύσσονται τα απαραίτητα θεωρητικά εργαλεία όπως τα μοντέλα ARMA και GARCH καθώς και το μέτρο της απόστασης μεταξύ δύο μοντέλων GARCH. Στο δεύτερο κεφάλαιο βρίσκεται η παρουσίαση των δεδομένων και της μεθοδολογίας που ακολουθήσαμε για να δημιουργήσουμε τις συστάδες καθώς και τα αποτελέσματα αυτής της ανάλυσης. Για την επεξεργασία των δεδομένων χρησιμοποιήσαμε το ελεύθερο λογισμικό της R. Στο τρίτο κεφάλαιο παρουσιάζονται τα συμπεράσματα και μια σύγκριση των αποτελεσμάτων με εργασία όπου επίσης έγινε ταξινόμηση χρηματοοικονομικών χρονικών σειρών αλλά με άλλο μέτρο απόστασης. Τέλος στο παράρτημα παρουσιάζονται κάποιες βασικές εντολές της γλώσσας R που χρησιμοποιήθηκαν και οι συναρτήσεις που ορίσαμε.

Abstract

Financial time series often present similar volatility structures. Selecting such series so that they have similar behavior can prove to be an important tool of analysis of the transmission mechanisms of volatility and to help forecast the time series, using the series with more similar structure. In the first chapter we develop the necessary tools such ARMA and GARCH models as well as the metrics needed to define the distance between two GARCH models. In the second chapter we present the data and the process we followed in order to exact the results and to create the clusters. For data processing we used the open – source programming language R. In the third chapter we show the conclusions reached and we compare our results with those of a similar paper. In the appendix some basic R commands and functions are shown along with the functions we defined.

Κεφάλαιο 1 Εισαγωγή – Στοιχεία Θεωρίας

1.1 Εισαγωγή

Οι χρηματοοικονομικές χρονοσειρές είναι αλληλένδετες και αλληλοεξαρτώμενες και υπόκεινται σε παρόμοιες δομές αστάθειας (**volatility**) εξαιτίας της δυνατής σχέσης και επίδρασης ανάμεσα στις αγορές. Γενικά περίοδοι ηρεμίας και αναταραχής ή ύφεσης μεταδίδονται από μια αγορά σε μια άλλη, αλλά κάποιες αγορές απορροφούν καλύτερα τέτοια φαινόμενα. Η κατάταξη τέτοιων σειρών σε ομογενείς συστάδες με παρόμοιες δομές αστάθειας είναι ένας πρακτικός στόχος για τους οικονομικούς αναλυτές καθώς κινήσεις σε δεδομένες χρονοσειρές θα μπορούσαν να χρησιμοποιηθούν ως βάση για την πρόβλεψη της κίνησης παρόμοιων χρονοσειρών.
Το θέμα της ταξινόμησης χρονοσειρών έχει ερευνηθεί αρκετά και υπάρχει πλήθος αναφορών στη διεθνή βιβλιογραφία, κυρίως στους τομείς της εξόρυξης δεδομένων στις χρονοσειρές (time series data mining) **[1]** Agrawal**,** στην επιστήμη των υπολογιστών και τις χρηματοοικονομικές χρονοσειρές. Ο Liao **[6]** δίνει μια εκτενή αναφορά όσον αφορά την συσταδοποίηση των χρονοσειρών και ξεχωρίζει τρεις κατηγορίες:

1. Προσέγγιση ακατέργαστων δεδομένων
2. Προσέγγιση σύμφωνα με χαρακτηριστικά των σειρών
3. Και μέθοδοι βασισμένες σε μοντέλα, όπου οι σειρές θεωρούνται όμοιες όταν τα μοντέλα που τις χαρακτηρίζουν είναι όμοια

Εδώ θα ασχοληθούμε με την τρίτη κατηγορία. Στην βιβλιογραφία χρησιμοποιείται συχνά η μέθοδος να συγκρίνονται δύο μοντέλα AR για να καταλήξουμε στην ομοιότητα δύο σειρών. Προσπαθούμε να προσδιορίσουμε τη δομή του μέσου της ανέλιξης υποθέτοντας ότι είναι η γεννήτρια των δεδομένων και συνήθως θεωρούμε την διακύμανση σταθερή. Αυτή είναι μια σωστή προσέγγιση όταν κάνουμε ταξινόμηση βασισμένοι σε μοντέλα ARMA και δεδομένης της ομοσκεδαστικής διακύμανσης **[12]**Piccolo **,** στην οποία όμως περίπτωση η διακύμανση είναι συνάρτηση των παραμέτρων του μοντέλου και έτσι συμπεριλαμβάνεται έμμεσα στην ταξινόμηση. Στην περίπτωση όμως ετεροσκεδαστικών χρονοσειρών στις οποίες η

(δεσμευμένη) διακύμανση ακολουθεί μια στοχαστική ανέλιξη (συνήθως GARCH ,**[3]**) η σύγκριση της συμπεριφοράς της διακύμανσης παίζει σημαντικό ρόλο.

Αυτό είναι αρκετά σημαντικό όταν ερευνούμε χρηματοοικονομικές σειρές, όπου ο επενδυτής έχει μεγάλο εύρος επενδυτικού χαρτοφυλακίου (εκατοντάδες μετοχές), και θέλει να έχει ομάδες σειρών με παρόμοια χαρακτηριστικά (όμοια διακύμανση, όμοια συμπεριφορά κτλ). Επιπλέον η αστάθεια της απόδοσης, θεωρείται γενικά σαν ένας «μετρητής» του ρίσκου της συγκεκριμένης απόδοσης, με άλλα λόγια η ταξινόμηση των αποδόσεων διαφόρων επενδύσεων σε συστάδες είναι ισοδύναμο με την ταξινόμηση των επενδύσεων σε συστάδες παρόμοιου ρίσκου. Επιπλέον κινήσεις σε μια χρονική σειρά μπορούν να βοηθήσουν να προβλέψουμε την κίνηση σε μια παρόμοια χρονική σειρά. Οι χρηματοοικονομικές σειρές υπόκεινται σε αλληλομετακινήσεις και σε παρόμοιες δομές αστάθειας, εξαιτίας της αμοιβαίας εξάρτησης ανάμεσα στις αγορές και την αυξανόμενη ενοποίηση των αγορών σε παγκόσμιο επίπεδο. Γενικά περίοδοι ταραχής μεταδίδονται από την μια αγορά στην άλλη. Η ταξινόμηση των χρηματοοικονομικών σειρών σε ομογενείς συστάδες με παρόμοιες δομές αστάθειας είναι ένας σημαντικός στόχος για τους οικονομικούς αναλυτές.

Σε αυτή τη εργασία θα χρησιμοποιήσουμε ένα μέτρο απόστασης που αρχικά είχε εισαχθεί στην εργασία του Piccolo **[12]** για μοντέλα AR και στην συνέχεια επεκτάθηκε και στα μοντέλα GARCH (Generalized AutoRegressive Conditional Heteroskedasticity) από τον Otrando **[9]**. Η απόσταση αυτή συγκρίνει τις στοχαστικές ιδιότητες των ζευγαριών σειρών, τις διαφορές δηλαδή μεταξύ δύο ανελίξεων που παράγουν τέτοια δεδομένα. Στην πράξη αυτό που γίνεται είναι ότι η εκτίμηση μοντέλων GARCH δίνει τη στατιστική δομή των χρονοσειρών , έτσι ώστε η σύγκριση των μοντέλων που δημιουργούν τις διαδικασίες παραγωγής δεδομένων είναι ισοδύναμη με την σύγκριση των δομών αστάθειας κάθε χρονοσειράς. Η επέκταση αυτή στα μοντέλα GARCH είναι σχετικά απλή δεδομένου της ομοιότητας των δομών μεταξύ των ARMA και των GARCH μοντέλων. Στην πράξη αναπαριστούμε τα κατάλοιπα (residuals) ενός GARCH μοντέλου σε μορφή ARMA και χρησιμοποιώντας την έκφραση της απόστασης για ARMA μοντέλα **[12]** μπορούμε να καταλήξουμε στην έκφραση της απόστασης μεταξύ δυο GARCH μοντέλων. Αυτή η αναπαράσταση μας δίνει μια διατύπωση της απόστασης σαν συνάρτηση των παραμέτρων GARCH. Έχοντας αυτή την απόσταση δημιουργούμε έναν πίνακα με τις αποστάσεις των χρονοσειρών μεταξύ τους και πάνω σε αυτό τον πίνακα εφαρμόζουμε

έναν αλγόριθμο δημιουργίας συστάδων (clustering) για να πετύχουμε την ταξινόμηση, και θα χρησιμοποιήσουμε μη- ιεραρχικές μεθόδους και πιο συγκεκριμένα την μέθοδο K-means clustering. Η ανάλυση θα γίνει μέσω του προγράμματος ελεύθερου λογισμικού της R και η διαδικασία που θα ακολουθηθεί θα ταξινομήσει τις αποδόσεις των σειρών σε συστάδες όμοιου ρίσκου.

Στη συνέχεια του κεφαλαίου θα αναφέρουμε τα απαραίτητα θεωρητικά εργαλεία ώστε να μπορέσουμε να εκφράσουμε την απόσταση, καθώς και πως καταλήγουμε στην διατύπωση αυτής, με ιδιαίτερη έμφαση στο μοντέλο GARCH(1,1) που θα είναι και το βασικό μας μοντέλο. Στη συνέχεια αναφέρουμε και την μέθοδο K-means που θα χρησιμοποιήσουμε για την συσταδοποίηση (clustering) των σειρών. Στο κεφάλαιο δύο γίνεται η παρουσίαση της μεθοδολογίας που ακολουθήθηκε για την ταξινόμηση οκτώ χρονικών σειρών αποδόσεων μετοχών του Χ.Α.Α. της περιόδου 1998 – 2002 καθώς και σύγκριση αποτελεσμάτων με αντίστοιχα άλλης μεθόδου υπολογισμού του μέτρου της απόστασης. Στο παράρτημα παρουσιάζονται εντολές της R που χρησιμοποιήσαμε κατά την ανάλυση των δεδομένων καθώς και ο κώδικας των συναρτήσεων που ορίσαμε.

1.2 Αποδόσεις

Στα οικονομικά ο βαθμός απόδοσης (Rate of Return), ή αλλιώς απόδοση επενδύσεων (Return on Investment – ROI), ή βαθμός κέρδους ή απλά **απόδοση** είναι ο λόγος του πλούτου που κερδίζεται ή χάνεται (είτε πραγματοποιήσιμος είτε όχι) σε μια επένδυση ως προς το ποσό που επενδύθηκε αρχικά. Ο πλούτος που κερδήθηκε ή χάθηκε λέγεται τόκος, είτε κέρδος/ απώλεια , είτε καθαρό κέρδος/ απώλεια και ο πλούτος που επενδύθηκε ονομάζεται κεφάλαιο, περιουσιακό στοιχείο ή το αρχικό κόστος της επένδυσης. Η απόδοση εκφράζεται συνήθως ως ποσοστό παρά ως κλάσμα. Υπάρχουν πολλοί τρόποι υπολογισμού της απόδοσης, αλλά αυτός που θα χρησιμοποιήσουμε είναι η **λογαριθμική απόδοση** (logarithmic or continuously compounded return) η οποία δίνεται από τον τύπο :

$$r_t = \ln\left(\frac{p_t}{p_{t-1}}\right) = \ln p_t - \ln p_{t-1} \qquad \textbf{(1)}$$

όπου p_t η τιμή της σειράς τη χρονική στιγμή t. Η λογαριθμική απόδοση διευκολύνει μαθηματικούς υπολογισμούς και κατά προσέγγιση αντιστοιχεί στο ρυθμό μεταβολής, δηλαδή

$$r_t = \ln\left(\frac{p_t}{p_{t-1}}\right) \approx \frac{p_t - p_{t-1}}{p_{t-1}}$$. Βλέπε Tsay (**[14])**

1.3 Γραμμικά μοντέλα ARMA

Αντίθετα με την υπόθεση του τυχαίου περιπάτου, το γεγονός ότι εμπειρικά παρατηρούμε σημαντική αυτοσυσχέτιση στις αποδόσεις για μία χρονική υστέρηση μας οδηγεί να θεωρήσουμε ότι η τιμή των αποδόσεων τη στιγμή t εξαρτάται από την τιμή τη στιγμή t-1 **[14] Tsay.** Το απλό αυτοπαλινδρομούμενο μοντέλο τάξεως p, (AR – Autoregressive) δίνεται από τον τύπο :

$$y_t = \mu + \sum_{i=1}^{p} \varphi_i y_{t-i} + \varepsilon_t \qquad \textbf{(2)}$$

όπου μ η προσδοκώμενη τιμή του $\mathbf{y_t}$ **,** ε_t είναι λευκός θόρυβος μηδενικού μέσου και διακύμανσης σ_α^2 και $\boldsymbol{\varphi_i}$ οι συντελεστές αυτοπαλινδρόμησης.

Το μοντέλο AR είναι της ίδιας μορφής με το μοντέλο της απλής γραμμικής παλινδρόμησης. Το μοντέλο της (2) είναι το AR(p). Για να έχει το AR(p) την επιθυμητή ιδιότητα της στασιμότητας πρέπει να ισχύει ο περιορισμός

$$\sum_{i=1}^{p} \varphi_i < 1 .$$

Το μοντέλο κινητού μέσου ***MA(q)*** τάξεως q είναι:

$$y_t = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \qquad \textbf{(3)}$$

οι θ_i οι συντελεστές του μοντέλου, μ η προσδοκώμενη τιμή του y_t και ε_{t-i} λευκός θόρυβος.

Η σύνθεση των δύο προηγούμενων δίνει το μεικτό μοντέλο ***ARMA(p,q)***

$$y_t = \mu + \sum_{i=1}^{p} \varphi_i y_{t-1} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \quad \textbf{(4)}$$

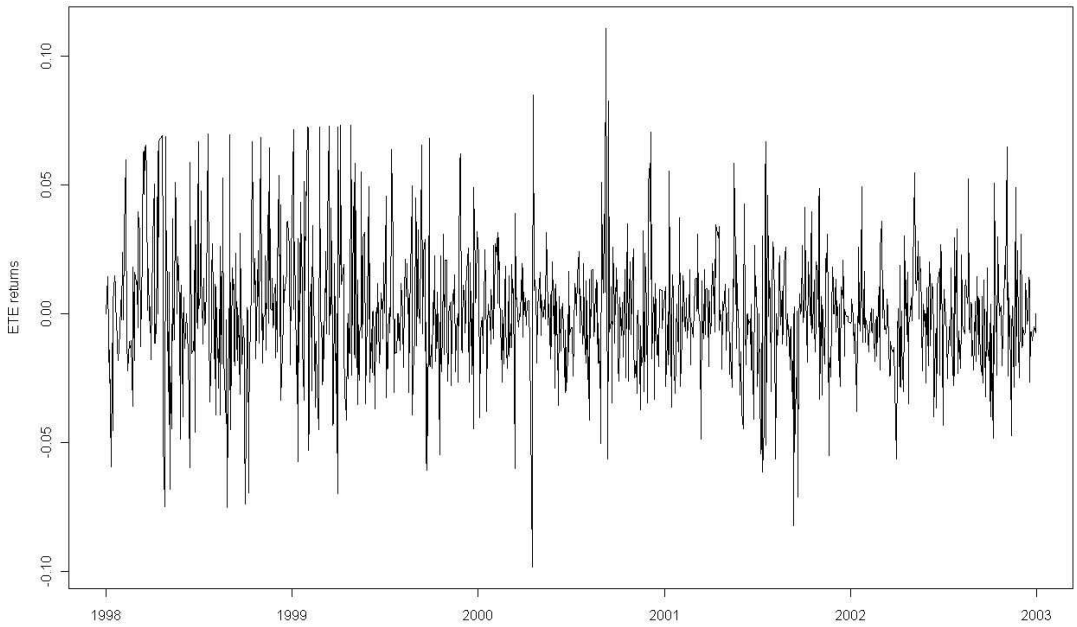
Το ARMA(p,q) περιλαμβάνει τους p αυτοπαλινδρομούμενους συντελεστές, και τους q συντελεστές κινητού μέσου και βοηθούν σε μια πιο λιτή μοντελοποίηση μιας σειράς δεδομένων. Γενικά τα ε_t είναι λευκός θόρυβος δευτέρου βαθμού.

1.4 Μοντέλα GARCH

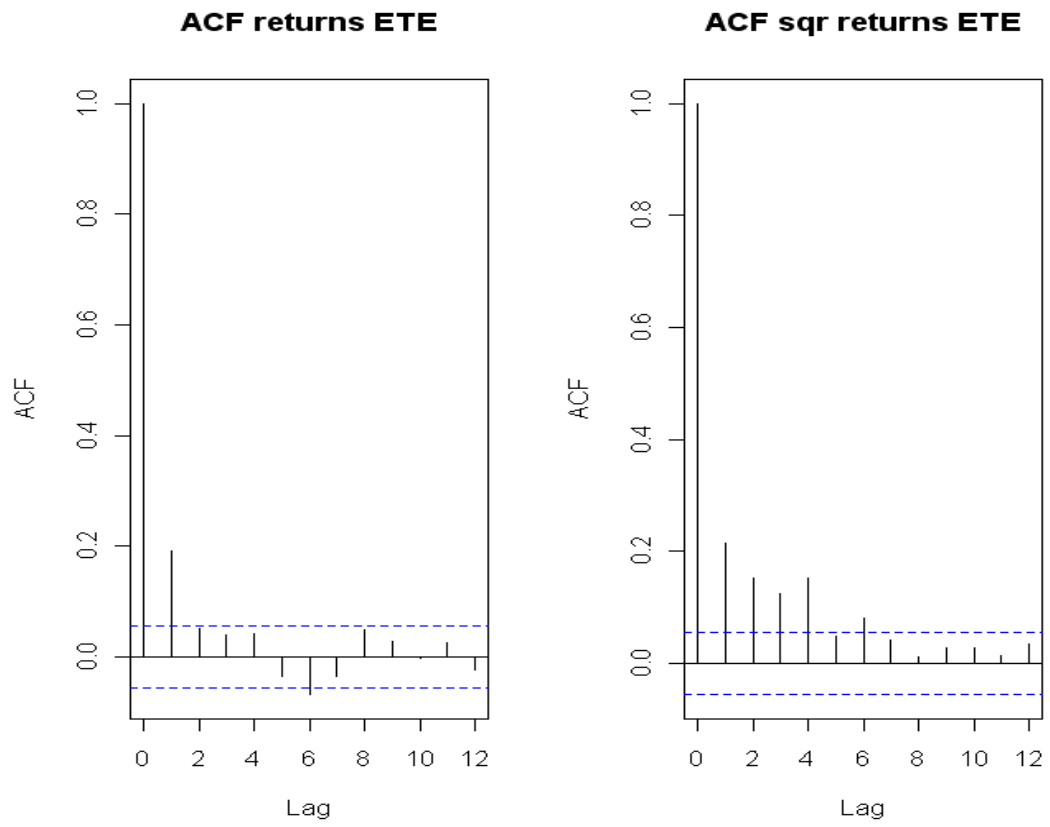
1.4.1 Εισαγωγή – Μοντέλα ARCH

Η ασταθής συμπεριφορά των αγορών αναφέρεται γενικά στην βιβλιογραφία ως αστάθεια (volatility). Η αστάθεια έχει καταστεί μια πολύ σημαντική έννοια στην οικονομική θεωρία και στις εφαρμογές αυτής, όπως η διαχείριση κινδύνων (Risk management), διαχείριση χαρτοφυλακίου, παράγωγες επενδύσεις κ.α. Με στατιστικούς όρους η αστάθεια μετράται με την διακύμανση ή την τυπική απόκλιση. Τα μοντέλα GARCH εισήχθησαν πρώτη φορά από τον Engle **[3]** και επεκτάθηκαν από τον Bollerslev **[2]** και τον Nelson **[8]** και μπορούμε με αυτά να μοντελοποιήσουμε αστάθεια μεταβλητή με τον χρόνο και να παρατηρήσουμε πολλά από τα χαρακτηριστικά που συναντώνται σε οικονομικές χρονοσειρές. Υπάρχουν διάφορες παραλλαγές των μοντέλων GARCH, όπως τα EGARCH που εισάγουν και επιδράσεις της μόχλευσης (leverage), TGARCH παρόμοιο με το EGARCH αλλά με άλλη μορφή, το PGARCH κ.α. **[15]** Zivot.

Αν πάρουμε τη γραφική παράσταση μιας τυπικής οικονομικής σειράς αποδόσεων, **εικ.1** για παράδειγμα εδώ των αποδόσεων της μετοχής της ΕΤΕ, θα παρατηρήσουμε από τη συνάρτηση αυτοσυσχετίσεων ότι ενώ στην ίδια τη σειρά δεν έχουμε ισχυρή συσχέτιση **εικ.2** , στο τετράγωνο των αποδόσεων υπάρχει συσχέτιση και από την στιγμή που οι τετραγωνισμένες αποδόσεις μετρούν την δευτεροβάθμια ροπή της αρχικής χρονοσειράς αυτό μας δείχνει ότι η διακύμανση εξαρτάται από τις προηγούμενες τιμές της ή με άλλα λόγια ότι οι αρχική σειρά έχει μεταβλητή ως προς το χρόνο δεσμευμένη ετεροσκεδαστικότητα ή συσταδοποίηση αστάθειας (volatility clustering).



Εικόνα 1 Χρονοσειρά αποδόσεων της ETE



Εικόνα 2 Συναρτήσεις αυτοσυσχετίσεων για αποδόσεις και τετράγωνα των αποδόσεων

Η γραμμική συσχέτιση στα τετράγωνα των αποδόσεων μπορεί να μοντελοποιηθεί χρησιμοποιώντας ένα μοντέλο AR για τα τετράγωνα των καταλοίπων. Καθώς όμως οι αποδόσεις είναι μια χρονική σειρά μηδενικού μέσου γραμμικά ασυσχέτιστη αλλά με τα τετράγωνα τους να συσχετίζονται, τότε μπορούμε να εφαρμόσουμε σε αυτά ένα μοντέλο GARCH. Το GARCH είναι ένα μοντέλο για τα τετράγωνα μιας μηδενικού μέσου σειράς, δηλαδή για την υπό συνθήκη δεσμευμένη διακύμανση (conditional variance). Αυτή η διακύμανση για μια χρηματοοικονομική σειρά παριστά αστάθεια ή κίνδυνο.

Έχουμε λοιπόν: $\mathbf{y_t = e_t}$ **(5)**

όπου $\boldsymbol{\varepsilon_t}$ είναι λευκός θόρυβος δευτέρου βαθμού.

Για να μπορέσουμε να μοντελοποιήσουμε για μεταβλητή ως προς το χρόνο δεσμευμένη ετεροσκεδαστικότητα (time – varying conditional heteroskedasticity) υποθέτουμε ότι $Var_{t-1}(\varepsilon_t) = h_t$ με $\mathrm{Var}_{t-1}(\cdot)$ να είναι η διακύμανση δεσμευμένη στην πληροφορία την χρονική στιγμή t-1 και τελικά έχουμε:

$$h_t = \gamma + a_1\varepsilon_{t-1}^2 + \ldots + a_p\varepsilon_{t-p}^2 \qquad \textbf{(6)}$$

Αφού το ε_t έχει μέσο μηδέν, $Var_{t-1}(\varepsilon) = E_{t-1}\left(\varepsilon_t^2\right) = h_t$ και μπορούμε να γράψουμε την (6) ως

$$\varepsilon_t^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \ldots + \alpha_p\varepsilon_{t-p}^2 + u_t \qquad \textbf{(7)}$$

με $u_t = \varepsilon_t^2 - E_{t-1}\left(\varepsilon_t^2\right)$ ή $u_t = \varepsilon_t^2 - h_t$ να είναι λευκός θόρυβος μηδενικού μέσου.

Η (7) αναπαριστά μια AR(p) διαδικασία ως προς το ε_t^2 και το μοντέλο των (5) και (6) είναι γνωστό ως ARCH (autoregressive conditional heteroskedasticity) και αναφέρεται ως ARCH(p) μοντέλο.
Μια διαφορετική διατύπωση αυτού του μοντέλου είναι

$$u_t = \varepsilon_t^2 - E_{t-1}\left(\varepsilon_t^2\right)$$
$$y_t = c + \varepsilon_t$$
$$\varepsilon_t = z_t\sigma_t$$

όπου z_t είναι μια ανεξάρτητη και όμοια κατανεμημένη τυχαία μεταβλητή με ορισμένη κατανομή.

Στο βασικό μοντέλο ARCH η z_t θεωρούμε ότι ακολουθεί την τυποποιημένη κανονική κατανομή. Η παραπάνω διατύπωση είναι χρήσιμη για να εξάγουμε τις ιδιότητες του μοντέλου καθώς και για τον προσδιορισμό της συνάρτησης πιθανοφάνειας που χρησιμοποιείται για την συμπερασματολογία.

1.4.2 Έλεγχος για φαινόμενα ARCH

Πριν τον προσδιορισμό ενός πλήρους μοντέλου ARCH για οικονομικές χρονοσειρές ελέγχουμε για ARCH επιπτώσεις στα κατάλοιπα. Αν δεν υπάρχουν ARCH επιπτώσεις στα κατάλοιπα τότε το μοντέλο μας είναι άκυρο και δεν εκφράζει ορθά την οικονομική σειρά. Επειδή ένα μοντέλο ARCH μπορεί να γραφεί ως AR ως προς τα τετράγωνα των καταλοίπων στην (7) ,ένας έλεγχος πολλαπλασιαστή Lagrange (LM) μπορεί να κατασκευαστεί στην βοηθητική παλινδρόμηση (7). Αν η μηδενική υπόθεση είναι ότι δεν υπάρχουν επιπτώσεις ARCH $H_0:\alpha_1=\alpha_2=\ldots=\alpha_p=0$ η στατιστική ελέγχου είναι

$$LM = T \bullet R^2 \sim \chi^2(p)$$

όπου T είναι το μέγεθος του δείγματος και ο συντελεστής προσδιορισμού R^2 υπολογίζεται από την βοηθητική παλινδρόμηση (7) χρησιμοποιώντας τα εκτιμώμενα κατάλοιπα.

1.4.3 Μοντέλα GARCH

Αν το τεστ LM είναι στατιστικά σημαντικό μπορούμε να υπολογίσουμε το μοντέλο ARCH και να εκτιμήσουμε την χρονικά μεταβαλλόμενη αστάθεια $\mathbf{h_t}$ από τις προηγούμενες χρονικές στιγμές. Ωστόσο στην πράξη αποδεικνύεται ότι χρειάζεται ένας μεγάλος αριθμός υστερήσεων p και συνεπώς ένας μεγάλος αριθμός παραμέτρων για να έχουμε ένα ικανοποιητικό μοντέλο. Ένα αποδοτικότερο μοντέλο προτάθηκε

από τον Bollerslev **[2]** και αντικαθιστά το μοντέλο στην (6) με την παρακάτω διατύπωση:

$$h_t = \gamma + \sum_{i=1}^{p} a_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} b_j h_{t-j} \qquad \textbf{(8)}$$

όπου οι συντελεστές α_i (i=0,…,p) και b_j (j=1,…,q) είναι θετικοί έτσι ώστε η διακύμανση h_t να είναι πάντα θετική.

Το μοντέλο της **(8)** μαζί με την **(5)** είναι το γενικευμένο ARCH , ή GARCH(p,q) μοντέλο. Όταν q=0 το GARCH μοντέλο γίνεται το ARCH(p).
Στο μοντέλο GARCH(p,q) η δεσμευμένη διακύμανση του ε_t , h_t ,εξαρτάται από τα τετράγωνα των καταλοίπων των p προηγούμενων περιόδων καθώς και από την δεσμευμένη διακύμανση των q προηγούμενων περιόδων. Συνήθως αρκεί ένα GARCH(1,1) με μόνο τρείς παραμέτρους για να έχουμε ένα καλό μοντέλο για την χρονική σειρά που μελετούμε. **[14]**

Ιδιότητες των μοντέλων GARCH

Δεδομένης της ARMA αναπαράστασης του GARCH μοντέλου από τις (9) και (10) πολλές ιδιότητες των μοντέλων αυτών μπορούν να εξαχθούν από αυτές των αντίστοιχων ARMA διαδικασιών για το ε_t^2. Για παράδειγμα για να είναι στάσιμο το μοντέλο GARCH(1,1) πρέπει να ισχύει $\alpha_1 + \beta_1 < 1$ όπως στην (10). Δεδομένης της στασιμότητας του GARCH(1,1), η διακύμανση του ε_t μπορεί να δειχθεί ότι είναι ίση με

$$Var(\varepsilon_t) = E(\varepsilon_t^2) = \gamma_0 / (1 - a_1 - b_1)$$

καθώς από την (10) έχουμε:

$$E(\varepsilon_t^2) = \gamma_0 + (a_1 + b_1) E(\varepsilon_{t-1}^2)$$

όποτε εξαιτίας της στασιμότητας

$$E(\varepsilon_t^2) = \gamma_0 + (a_1 + b_1) E(\varepsilon_t^2) .$$

Για την γενική περίπτωση του GARCH(p,q) τα τετράγωνα των καταλοίπων ακολουθούν ένα ARMA(max(p,q),q).
Η στασιμότητα της συνδιακύμανσης επιβάλλει

$\sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_i < 1$ και η διακύμανση της ε_t είναι : $Var(\varepsilon_t) = \frac{\gamma_0}{1 - \left(\sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_i \right)}$.

1.4.4 Μοντέλα GARCH και παρατηρήσεις

Στην πράξη οι ερευνητές έχουν ανακαλύψει πολλά αποτελέσματα όσον αφορά την αστάθεια των οικονομικών χρονικών σειρών. Πολλά από αυτά αναφέρονται από τον Bollerslev **[2]** και Engle **[3].** Χρησιμοποιώντας την ARMA αναπαράσταση των GARCH μοντέλων, αποδεικνύεται ότι τα GARCH μοντέλα μπορούν να εξηγήσουν πολλά από αυτά τα αποτελέσματα. Κάποια από αυτά είναι η συσταδοποίηση αστάθειας και οι παχιές ουρές.

Συσταδοποίηση αστάθειας (volatility clustering)

Θεωρούμε το μοντέλο GARCH(1,1) από την (9). Συνήθως ο συντελεστής b_1 κυμαίνεται γύρω από το 0.9 για πολλές εβδομαδιαίες και ημερήσιες οικονομικές χρονικές σειρές. Δεδομένης αυτής της τιμής του b_1 συμπεραίνουμε ότι μεγάλες τιμές του σ_{t-1}^2 θα ακολουθούνται από μεγάλες τιμές σ_t^2 και μικρές τιμές σ_{t-1}^2 θα ακολουθούνται από μικρές τιμές του σ_t^2. Το ίδιο μπορεί να εξαχθεί και από την ARMA αναπαράσταση (10), όπου μεγάλες/μικρές αλλαγές στο ε_{t-1}^2 ακολουθούνται από μεγάλες/μικρές αλλαγές στο ε_t^2.

Παχιές ουρές

Είναι γνωστό ότι η κατανομή πολλών χρονοσειρών υψηλής συχνότητας, έχει συνήθως πιο παχιές ουρές από την κανονική κατανομή. Αυτό σημαίνει ότι μεγάλες αλλαγές είναι πιο πιθανό να συμβούν απ' ότι στην κανονική κατανομή. Ο Bollerslev **[2]** δίνει την συνθήκη που πρέπει να ισχύει για την ύπαρξη τεταρτοβάθμιας ροπής ενός μοντέλου GARCH(1,1). Θεωρώντας ότι υπάρχει η τετάρτου βαθμού ροπή ο Bollerslev **[2]** αποδεικνύει ότι η κύρτωση ενός GARCH(1,1) είναι μεγαλύτερη από τρία την κύρτωση της κανονικής κατανομής. Αυτό το αποτέλεσμα μπορεί να

αποδειχθεί και για την γενική περίπτωση του GARCH(p,q) **[13]**Terasvirta 99. Έτσι ένα μοντέλο GARCH μπορεί να μοντελοποιήσει επαρκώς τις παχιές ουρές που συνήθως παρατηρούνται σε οικονομικές χρονοσειρές.

1.4.5 ARMA αναπαράσταση, ορισμός της απόστασης και ιδιότητες

Η εξίσωση (8) για την περίπτωση GARCH(1,1) και θεωρώντας δύο χρονοσειρές $\mathbf{y_{1,t}} = \boldsymbol{\varepsilon_{1,t}}$ και $\mathbf{y_{2,t}} = \boldsymbol{\varepsilon_{2,t}}$ γίνεται ως εξής:

$$\begin{aligned} h_{1,t} &= \gamma_1 + \alpha_1 \varepsilon_{1,t-1}^2 + \beta_1 h_{1,t-1} \\ h_{2,t} &= \gamma_2 + \alpha_2 \varepsilon_{2,t-1}^2 + \beta_2 h_{2,t-1} \end{aligned} \qquad \textbf{(9)}$$

όπου υπενθυμίζουμε ότι h_t είναι η διακύμανση δεσμευμένη στην πληροφορία την χρονική στιγμή t-1.
Οι περιορισμοί που ισχύουν είναι : $\boldsymbol{\gamma_i > 0, 0 < \alpha_i < 1,\ 0 < \beta_i < 1,\ 0 < \alpha_i + \beta_i < 1 (i=1,2)}$.

Ξαναγράφοντας την (9) μπορούμε να συμπεράνουμε ότι τα τετράγωνα των καταλοίπων ακολουθούν ένα ARMA(1,1).

$$\varepsilon_{i,t}^2 = \gamma_i + (a_i + \beta_i)\varepsilon_{i,t-1}^2 - \beta_i(\varepsilon_{i,t-1}^2 - h_{i,t-1}) + (\varepsilon_{i,t}^2 - h_{i,t}), i = 1,2 \qquad \textbf{(10)}$$

όπου $\varepsilon_{i,t}^2 - h_{i,t}$ είναι σφάλματα μηδενικού μέσου.

Αντικαθιστώντας στην (7) τα σφάλματα με την ARMA(1,1) αναπαράσταση τους, έχουμε την AR(∞) έκφραση:

$$\varepsilon_{i,t}^2 = \frac{\gamma_i}{1-\beta_i} + a_i \sum_{j=1}^{\infty} \beta_i^{j-1} \varepsilon_{i,t-j}^2 + (\varepsilon_{i,t}^2 - h_{i,t}) \qquad \textbf{(11)}$$

Σε αυτή τη μορφή τα δύο μοντέλα GARCH(1,1) μπορούν να συγκριθούν με το μέτρο της απόστασης που πρότεινε ο Piccolo **[12].** Στην [12] η γενική μορφή του μέτρου είναι:

$$\left[\sum_{j=0}^{\infty} \left(\pi_{1,j} - \pi_{2,j} \right)^2 \right]^{1/2} \qquad \textbf{(12)}$$

όπου οι σταθερές $\pi_{1,j}$,και $\pi_{2,j}$ είναι οι αυτοπαλινδρομούμενοι συντελεστές των δύο AR(∞) διαδικασιών.

Χρησιμοποιώντας την (11) μπορούμε να πάρουμε ένα αντίστοιχο μέτρο απόστασης για μοντέλα GARCH(1,1) το οποίο θα έχει την μορφή:

$$d = \left[\sum_{j=0}^{\infty} \left(\alpha_1 \beta_1^j - \alpha_2 \beta_2^j \right)^2 \right]^{1/2} .$$

Αναπτύσσοντας την ταυτότητα στις αγκύλες παίρνουμε:

$$d = \left[\alpha \sum_{j=0}^{\infty} \beta_1^{2j} + \alpha_2^2 \sum_{j=0}^{\infty} \beta_2^{2j} - 2\alpha_1 \alpha_2 \sum_{j=0}^{\infty} \left(\beta_1 \beta_2 \right)^j \right]^{1/2} = \left[\frac{\alpha_1^2}{1-\beta_1^2} + \frac{\alpha_2^2}{1-\beta_2^2} - \frac{2\alpha_1 \alpha_2}{1-\beta_1 \beta_2} \right]^{1/2} \textbf{(13)}$$

από **[9]** Otrando.

Σημειώνεται ότι στις προηγούμενες αναπτύξεις η σταθερά $\frac{\gamma_i}{1-\beta_i}$ δεν λήφθηκε υπ όψιν στους υπολογισμούς. Στην πράξη δεν επηρεάζει την δυναμική της αστάθειας των 2 σειρών όπως εκφράζονται από τους αυτοπαλινδρομούμενους συντελεστές.
Μπορεί επίσης να επεκταθεί η έννοια της απόστασης και για την γενική περίπτωση των GARCH(p,q) μοντέλων. Σε αυτή την εργασία θα χρησιμοποιήσουμε την απλή μορφή πού δίνεται από την (13) και αντιστοιχεί στο GARCH(1,1).

1.5 Συσταδοποίηση – Αλγόριθμοι

Η συσταδοποίηση είναι μια μέθοδος ανάθεσης των στοιχείων ενός συνόλου σε υποσύνολα (συστάδες) ωστέ οι συστάδες που θα δημιουργηθούν να είναι παρόμοιες ως προς κάποιο κριτήριο. Συνήθως δεν υπάρχει καμιά πρότερη γνώση σχετικά με το πόσες ομάδες (συστάδες – clusters) θα δημιουργηθούν ή ποια θα είναι η δομή των συστάδων αλλά όλα αποφασίζονται στην πορεία από αποφάσεις που παίρνονται κατά την εκτέλεση του αλγορίθμου βάσει κάποιων παραμέτρων. Ουσιαστικά ο κύριος στόχος την ανάλυσης συστάδων (cluster analysis) είναι να αναδείξει τις ομάδες που

ανήκει καλύτερα κάποιο αντικείμενο ή μεταβλητή. Αρχικά θα πρέπει να αναπτυχθεί μια κλίμακα με την οποία να μετράται η ομοιότητα ή σχέση μεταξύ των αντικειμένων. Ως τέτοια κλίμακα συνήθως μπορούν να χρησιμοποιηθούν διάφορα μέτρα από τα οποία μερικά αναφέρονται παρακάτω ως τα πιο συνηθισμένα.

1.5.1 Αποστάσεις

Ευκλείδεια απόσταση

Αν έχουμε δύο παρατηρήσεις p- διαστάσεων $\mathbf{x}=[x_1, x_2, ..., x_p]$ και $\mathbf{y}=[y_1, y_2, ..., y_p]$

Τότε η ευκλείδεια απόσταση είναι :

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+...+(x_p-y_p)^2} = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$$

Στατιστική απόσταση

Η στατιστική απόσταση μεταξύ 2 παρατηρήσεων όπως παραπάνω είναι της μορφής :

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'\,\mathbf{A}\,(\mathbf{x}-\mathbf{y})}$$

όπου $\mathbf{A=S^{-1}}$ με το **S** να είναι ο πίνακας της δειγματικής διακύμανσης και συνδιακύμανσης.

Απόσταση Minkowski

Η απόσταση Minkowski δίνεται από το τύπο: $d(\mathbf{x},\mathbf{y}) = \left[\sum_{i=1}^{p}|x_i-y_i|^m\right]^{1/m}$

Για m=1 η απόσταση είναι η απόσταση Manhattan., για m=2 έχουμε την ευκλείδεια απόσταση. Γενικά για διάφορες τιμές τις m αλλάζει το βάρος που δίνεται σε μεγαλύτερες και μικρότερες διαφορές.

Μέτρο Canberra και συντελεστής Czekanowski

Και τα δύο μέτρα αυτά ορίζονται για θετικές μεταβλητές

Canberra $$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{(x_i + y_i)}$$

Czekanowski $$d(\mathbf{x},\mathbf{y}) = 1 - \frac{2\sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p} (x_i + y_i)}$$

1.5.2 Αλγόριθμοι

Υπάρχουν δύο βασικές ομάδες αλγορίθμων : οι ιεραρχικές μέθοδοι και οι μη – ιεραρχικές. **[4] Johnson**

Ιεραρχικές μέθοδοι

Οι ιεραρχικές μέθοδοι προχωρούν είτε με μια σειρά διαδοχικών συγχωνεύσεων είτε με σειρά διαδοχικών διαιρέσεων των αρχικών μεταβλητών ή αντικειμένων.

Οι **συναθροιστικές ιεραρχικές μέθοδοι** ξεκινούν με τα ίδια τα αντικείμενα να αποτελούν μία συστάδα το καθένα. Τα πιο όμοια αντικείμενα ομαδοποιούνται αρχικά και στη συνέχεια αυτές οι αρχικές συστάδες ομαδοποιούνται και αυτές μέχρι να καταλήξουμε να έχουμε μια μόνο συστάδα.

Οι **διαιρετικές ιεραρχικές μέθοδοι** ακολουθούν την ακριβώς αντίθετη διαδικασία. Ξεκινάει έχοντας μόνο μια συστάδα που περιέχει όλα τα αντικείμενα (μεταβλητές) και στη συνέχεια διαιρεί την αρχική ομάδα και κάθε αντικείμενο τοποθετείται στην συστάδα με την οποία απέχει όσο το δυνατόν περισσότερο από τα αντικείμενα της άλλης συστάδας. Η διαδικασία συνεχίζει μέχρι να καταλήξουμε να έχουμε τόσες συστάδες όσα και αντικείμενα.

Για τις ***συναθροιστικές ιεραρχικές μεθόδους*** τα βήματα που ακολουθεί ο αλγόριθμος για να τοποθετήσει σε συστάδες ***N*** αντικείμενα (μεταβλητές ή στοιχεία) είναι:

1. Ξεκινάμε με N συστάδες η καθεμία να περιέχει από ένα στοιχείο και ένα συμμετρικό πίνακα αποστάσεων **D={d_{ik}}**
2. Ψάχνουμε τον πίνακα D για το κοντινότερο (πιο όμοιο) ζευγάρι αποστάσεων, έστω η απόσταση μεταξύ των δύο συστάδων U και V να είναι d_{uv}.

3. Συγχωνεύουμε τις δύο $_{\text{συστάδες}}$, ονομάζουμε την καινούρια συστάδα (UV) και ενημερώνουμε τον πίνακα αποστάσεων διαγράφοντας τις γραμμές και στήλες που περιέχουν τις U και V και προσθέτοντας μια γραμμή και μια στήλη για την καινούρια συστάδα (UV) και υπολογίζοντας τις αποστάσεις μεταξύ των υπολοίπων συστάδων και της (UV).
4. Επαναλαμβάνουμε τα βήματα 2 και 3 ***N-1*** φορές συνολικά.

Δυο σημαντικές παραλλαγές του παραπάνω αλγορίθμου είναι οι μέθοδοι ***linkage.***

Για τον ***single – linkage*** αλγόριθμο οι συστάδες δημιουργούνται από την κάθε οντότητα συγχωνεύοντας τους «κοντινότερους γείτονες» όπου ο όρος «κοντινότερος γείτονας» υποδηλώνει την μικρότερη απόσταση ή μεγαλύτερη ομοιότητα. Αρχικά εντοπίζουμε στον πίνακα των αποστάσεων την μικρότερη απόσταση στον πίνακα **D** και συγχωνεύουμε τις συστάδες σύμφωνα με το βήμα 2 παραπάνω. Στο βήμα 3 οι αποστάσεις μεταξύ της συστάδας (UV) και κάποιας άλλης συστάδας w, είναι υπολογίζεται από : $d_{(UV)W} = min\{d_{UW}, d_{VW}\}$

Το αποτέλεσμα του αλγορίθμου μπορεί να αναπαρασταθεί με ένα δενδρόγραμμα.

Για τον ***complete – linkage*** αλγόριθμο οι συστάδες δημιουργούνται με παρόμοιο τρόπο με τον single – linkage με την μόνη διαφορά ότι σε κάθε βήμα η απόσταση μεταξύ των συστάδων προσδιορίζεται από την απόσταση των δύο στοιχείων ένα από κάθε συστάδα που είναι *πιο μακρινά.* Η διαφορά βρίσκεται στο βήμα 3 όπου οι αποστάσεις υπολογίζονται από τον τύπο : $d_{(UV)W} = max\{d_{UW}, d_{VW}\}$

Και στις δυο παραπάνω ιεραρχικές μεθόδους δεν μας προσφέρει ιδιαίτερα να αφήσουμε τον αλγόριθμο να τρέξει ως το τέλος και να έχουμε στην πρώτη περίπτωση μια τελική συστάδα και στην άλλη τόσες συστάδες όσες και αντικείμενα. Αυτό που γίνεται είναι να σταματούμε σε ένα συγκεκριμένο βάθος ανάλογα με αυτό που μας ενδιαφέρει και στο πόσες συστάδες αναμένουμε να έχουμε.

Όπως και με τις περισσότερες μεθόδους, πηγές σφαλμάτων και μεταβολών δεν λαμβάνονται υπ όψιν στις ιεραρχικές μεθόδους. Αυτό σημαίνει ότι μια μέθοδος συσταδοποίησης θα είναι ευαίσθητη σε «σημεία θορύβου». Στις ιεραρχικές μεθόδους συσταδοποίησης δεν υπάρχει πρόβλεψη για ανακατανομή των αντικειμένων που μπορεί να είχαν τοποθετηθεί λανθασμένα σε αρχικά στάδια του αλγορίθμου. Συνεπώς

η τελική διάταξη των συστάδων πρέπει να ελέγχεται για να δούμε αν είναι λογική. Γενικά για ένα δεδομένο πρόβλημα καλό είναι να δοκιμάζονται πολλές μέθοδοι συσταδοποίησης και μέσα σε μια συγκεκριμένη μέθοδο διαφορετικές αποστάσεις που θα χρησιμοποιεί ο αλγόριθμος. Αν τα αποτελέσματα από αυτές τις περιπτώσεις ταυτίζονται μπορούμε πιο εύκολα να καταλήξουμε στο συμπέρασμα ότι η ομαδοποίηση είναι ορθή.

Η ευστάθεια μιας λύσης προερχόμενης από ιεραρχική μέθοδο μπορεί να ελεγχθεί εφαρμόζοντας μικρές διαταραχές στα δεδομένα και παρατηρώντας πως αντιδρά σε αυτές. Αν οι ομάδες είναι σωστά διαχωρισμένες, τότε τα αποτελέσματα πριν και μετά τις διαταραχές, θα πρέπει να είναι τα ίδια.

Αν στα πρώτα βήματα εκτέλεσης του αλγορίθμου έχουμε ίσες τιμές στον πίνακα των αποστάσεων , μπορεί να έχουμε πολλαπλές λύσεις. Τα δενδρογράμματα δηλαδή που αντιστοιχούν σε διαφορετική αντιμετώπιση των ίσων αποστάσεων μπορεί να είναι διαφορετικά, ιδιαίτερα στα χαμηλότερα επίπεδα. Αυτό δεν είναι ένα πρόβλημα που προκύπτει απ' την μέθοδο αλλά απλά υπάρχουν πολλαπλές λύσεις για δεδομένη ομάδα δεδομένων. Οι πολλαπλές λύσεις δεν είναι απαραίτητα σημάδι ότι δεν έγινε σωστή ανάλυση των δεδομένων, αλλά ο χρήστης πρέπει να γνωρίζει την ύπαρξή τους ώστε τα δενδρογράμματα να μπορούν να ερμηνευτούν ορθά και συγκρινόμενα με διαφορετικά δενδρογράμματα να διατηρούν την ισχύ τους επί αυτών.

Μη – ιεραρχικές μέθοδοι

Οι μη – ιεραρχικές μέθοδοι δημιουργίας συστάδων είναι σχεδιασμένες ώστε να ομαδοποιούν αντικείμενα σε μια συλλογή Κ συστάδων. Ο αριθμός συστάδων Κ, μπορεί είτε να είναι ορισμένος εκ των προτέρων είτε να προσδιορίζεται σαν μέρος του αλγορίθμου. Επειδή δεν χρειάζεται να προσδιοριστεί πίνακας αποστάσεων μεταξύ των αντικειμένων που θέλουμε να οργανώσουμε σε συστάδες, και τα βασικά δεδομένα δεν χρειάζεται να αποθηκευτούν στον υπολογιστή κατά το τρέξιμο του αλγορίθμου, οι μη ιεραρχικές μέθοδοι μπορούν να εφαρμοστούν σε πολύ μεγαλύτερο όγκο δεδομένων από ότι οι ιεραρχικές.

Οι μη ιεραρχικές μέθοδοι ξεκινούν είτε από μια αρχική διαίρεση των αντικειμένων σε ομάδες ή από ένα αρχικό σύνολο κομβικών σημείων (seed points) τα οποία θα σχηματίσουν τον πυρήνα των συστάδων. Καλές επιλογές για τις αρχικές συνθήκες θα πρέπει να είναι αμερόληπτες έτσι μια καλή επιλογή μπορεί να είναι μια τυχαία

επιλογή αρχικών κομβικών σημείων ή ένας τυχαίος διαχωρισμός σε αρχικές ομάδες. Παρακάτω ακολουθεί η πιο γνωστή μη ιεραρχική μέθοδος η K-means.

K - Means μέθοδος

Ο McQueen **[7]** προτείνει τον όρο K-means για να περιγράψει έναν αλγόριθμο ο οποίος αναθέτει κάθε αντικείμενο, στην συστάδα που έχει τον κοντινότερο σε αυτό μέσο. Η διαδικασία που ακολουθείται για την πιο απλή μορφή αποτελείται από τα παρακάτω βήματα:

1. Χωρίζουμε τα αντικείμενα σε ***K*** αρχικές συστάδες και υπολογίζουμε τον μέσο(κέντρο) της συστάδας
2. Ανατρέχουμε όλα τα αντικείμενα, αναθέτοντας καθένα από αυτά στη συστάδα με της οποίας το μέσο είναι πιο κοντά. Ο υπολογισμός αυτής της απόστασης μεταξύ κάθε αντικειμένου και του κέντρου της συστάδας είναι συνήθως η Ευκλείδεια απόσταση όπως ορίστηκε παραπάνω. Αφού γίνουν οι νέες αναθέσεις των αντικειμένων σε συστάδες σύμφωνα με αυτό τον κανόνα, υπολογίζουμε εκ νέου το κέντρο κάθε συστάδας.
3. Επαναλαμβάνουμε τη διαδικασία έως ότου να μην μπορούν να γίνουν περαιτέρω ανακατατάξεις.

Αντί να ξεκινάμε με μία διαίρεση όλων των αντικειμένων σε K αρχικές συστάδες θα μπορούσαμε να ορίσουμε ***K*** αρχικά κομβικά σημεία και να προχωρήσουμε από εκεί στο δεύτερο βήμα.

Η τελική ανάθεση των αντικειμένων σε συστάδες θα εξαρτάται σε κάποιο βαθμό από τον αρχικό διαχωρισμό των ομάδων ή την επιλογή των σημείων. Στην πράξη οι περισσότερες αλλαγές στην ανάθεση των αντικειμένων συμβαίνουν κατά το πρώτο βήμα ανακατανομής.

Λόγοι για να μην έχουμε σταθερό αριθμό συστάδων ***K*** κατά την εκτέλεση του αλγορίθμου

- Αν δύο ή περισσότερα κομβικά σημεία βρίσκονται σε μία συστάδα οι παραγόμενες συστάδες θα είναι ανεπαρκώς διαφοροποιημένες
- Η ύπαρξη μιας ακραίας τιμής (outlier) μπορεί να δώσει τουλάχιστον μια συστάδα με πολύ διασκορπισμένα αντικείμενα.

- Ακόμη και αν ο πληθυσμός αποτελείται από ***K*** ομάδες, η δειγματοληπτική μέθοδος μπορεί να είναι τέτοια που δεδομένα από την πιο σπάνια ομάδα να μην εμφανίζονται στο δείγμα. Το να εξαναγκάζουμε τα δεδομένα σε ***K*** ομάδες μπορεί να οδηγήσει στη δημιουργία συστάδων που δεν έχουν νόημα.

Αυτό που συνιστάται αν ο αλγόριθμος απαιτεί να οριστεί ο αριθμός των συστάδων είναι να τρέχουμε τον αλγόριθμο για αρκετές τιμές του ***K*** ώστε να καταλήξουμε σε ένα ασφαλές συμπέρασμα όσον αφορά την τελική σύνθεση των συστάδων.

Κεφάλαιο 2 Μοντελοποίηση δεδομένων – Ομαδοποίηση

2.1 Εισαγωγή – παρουσίαση χρονοσειρών

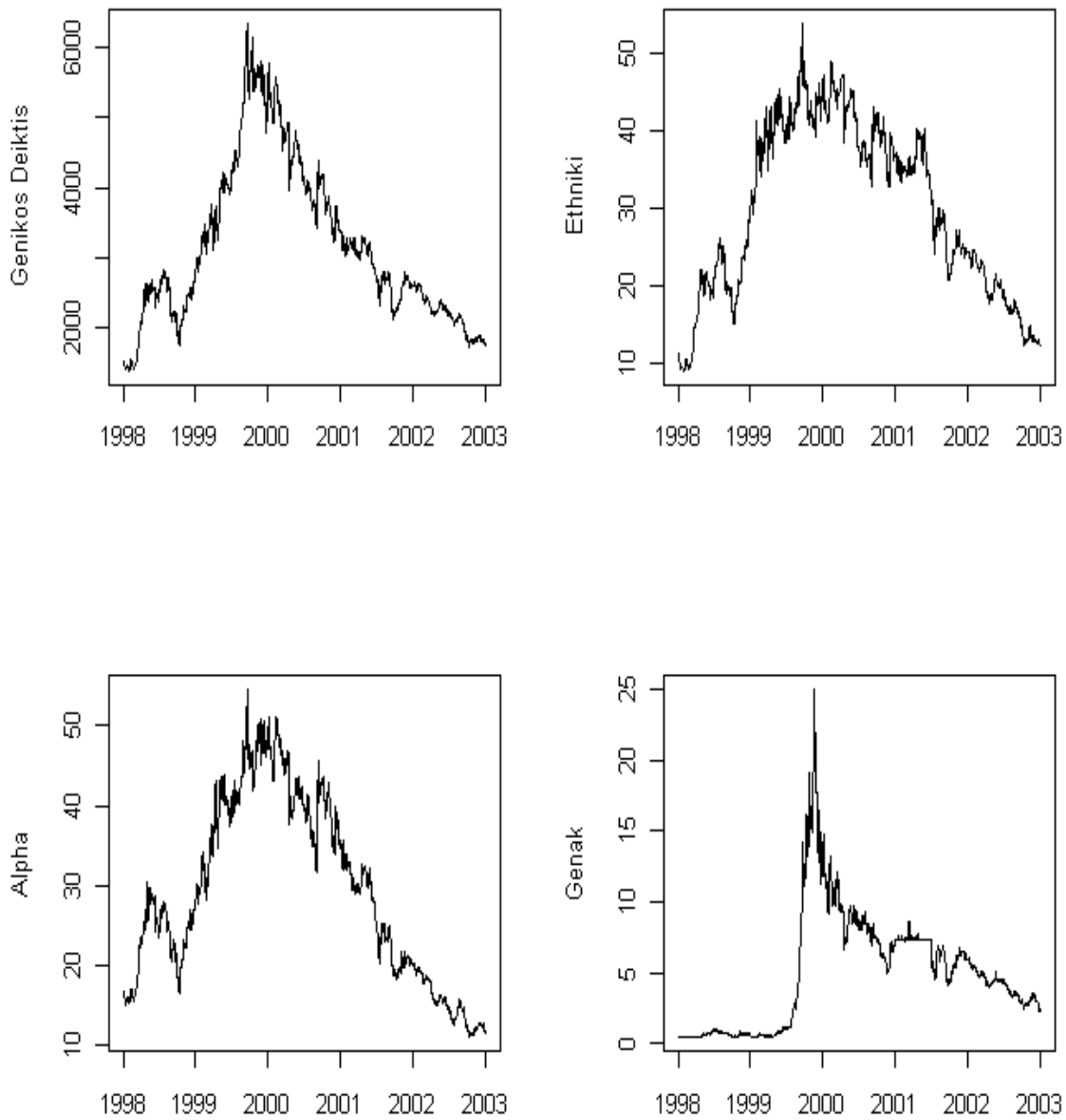
Σε αυτή την εργασία οι χρονοσειρές που θα αναλύσουμε είναι τιμές μετοχών του Χ.Α.Α. οι οποίες αφορούν την πενταετία από 2/1/1998 έως 31/12/2002 καλύπτοντας έτσι την περίοδο εξαιρετικής ανόδου και πτώσης του Χ.Α.Α. με κορυφή το 1999. Στόχος μας είναι να ομαδοποιήσουμε σε συστάδες τις μετοχές βάσει του μέτρου της απόστασης που αναφέρθηκε παραπάνω, ώστε να καταλήξουμε να έχουμε ομάδες που να περιλαμβάνουν τις μετοχές που είναι αξιόπιστες, τις μετοχές υψηλού ρίσκου και μετοχές που ανήκουν σε ενδιάμεσες κατηγορίες κάνοντας έτσι μια αξιολόγηση των μετοχών.

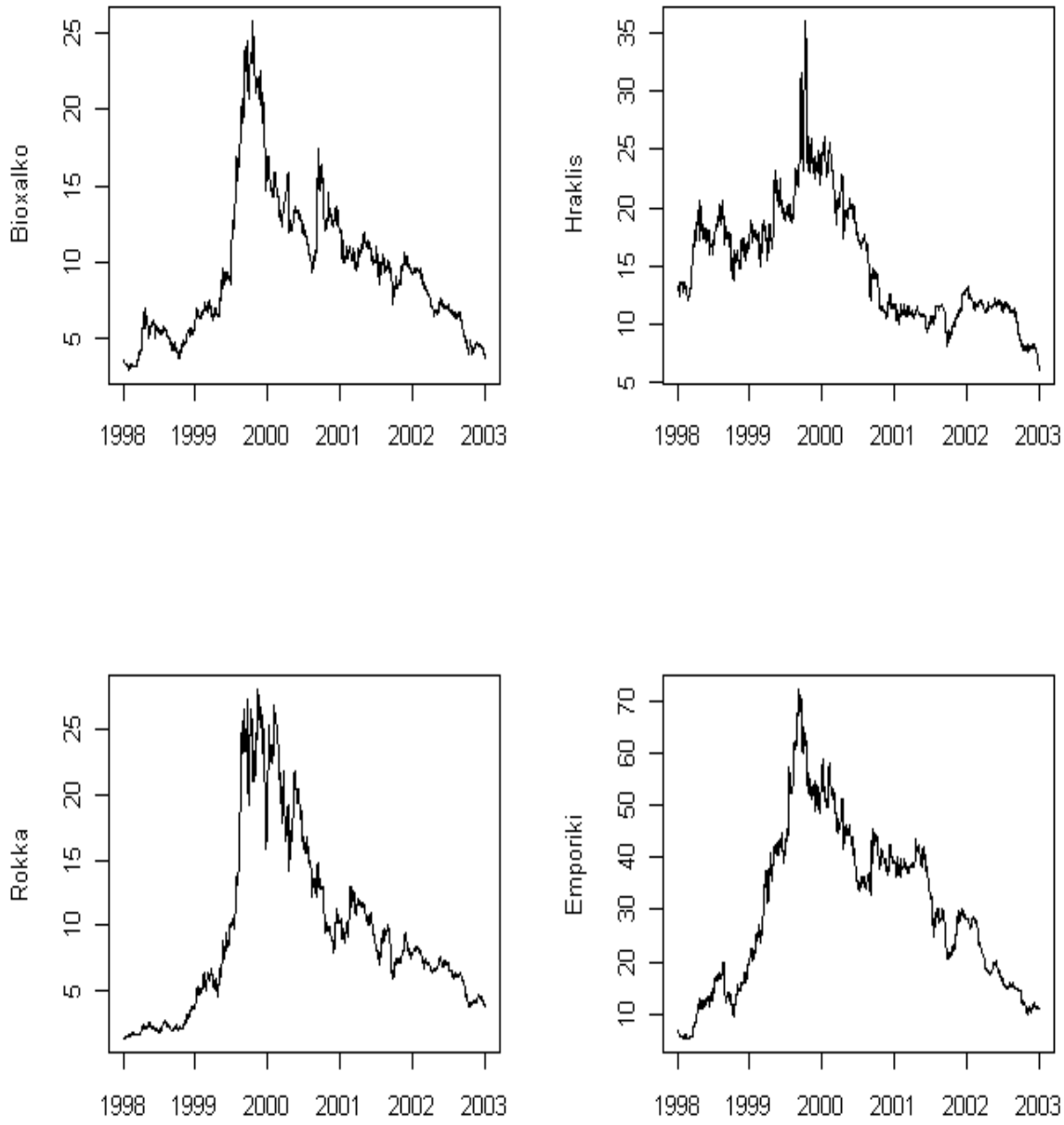
Οι μετοχές που θα εξετάσουμε είναι ο **Γενικός Δείκτης** του Χ.Α.Α. η μετοχή της **Εθνικής Τράπεζας**, της **Alpha Bank**, της **ΓΕΝΑΚ** (Εθνική Ακινήτων), της **Βιοχάλκο**, της **Ηρακλής**, της **ΡΟΚΚΑ** κατασκευαστικής και της **Εμπορικής** τράπεζας.

Η διαδικασία που θα ακολουθηθεί για να φτάσουμε να εφαρμόσουμε τον αλγόριθμο συσταδοποίησης θα έχει ως εξής: Αφού υπολογίσουμε αρχικά τις αποδόσεις των μετοχών από τις αρχικές τιμές, θα αφαιρέσουμε τον μέσο ώστε να γίνουν στάσιμες οι σειρές. Εφαρμόζουμε στη συνέχεια ένα **AR(1)** στις προσαρμοσμένες αποδόσεις. Από το μοντέλο αυτό παίρνουμε τα κατάλοιπα και ελέγχουμε τον μέσο ο οποίος θεωρητικά είναι μηδενικός, στην πράξη αποδεικνύεται να είναι της τάξης 10^{-8} με 10^{-7}. Η διαδικασία για να καταλήξουμε να έχουμε τις αποδόσεις λευκασμένες (whitened) και διορθωμένες για μέσο μηδέν καλείται **προλεύκανση.** Σε αυτά τα κατάλοιπα εφαρμόζουμε ένα GARCH(1,1) από όπου υπολογίζουμε τους συντελεστές $\mathbf{a_1}$ και $\mathbf{b_1}$ του μοντέλου της **(9)** και εφαρμόζουμε για αυτούς την σχέση **(13)** για να πάρουμε τον πίνακα των αποστάσεων. Πάνω στον πίνακα των αποστάσεων εφαρμόζουμε την μέθοδο K-means, για αριθμό ***K*** των συστάδων ίσο με τρείς και τέσσερις, με δύο διαφορετικές υλοποιήσεις του αλγορίθμου με τις συναρτήσεις **kmeans** και **pam** της **R** (παράρτημα) για τις μη – ιεραρχικές μεθόδους και με την συνάρτηση **hclust** της **R** για τις ιεραρχικές μεθόδους.

2.2 Γραφικές παραστάσεις χρονοσειρών

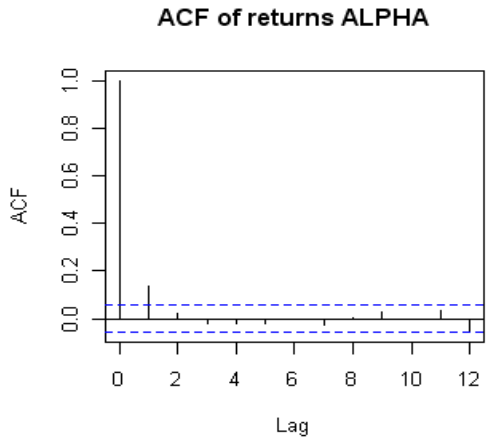
Στην παρακάτω εικόνα υπάρχουν οι γραφικές παραστάσεις των αρχικών δεδομένων συναρτήσει του χρόνου για τις σειρές που μελετούμε.

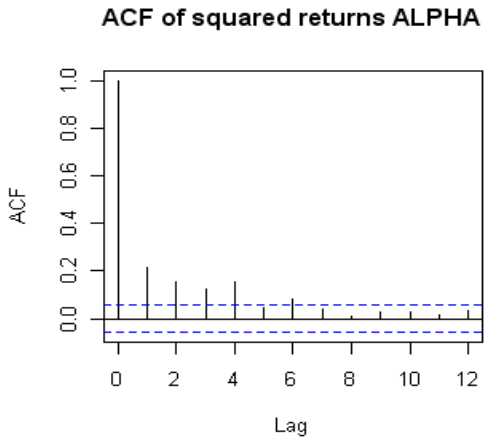


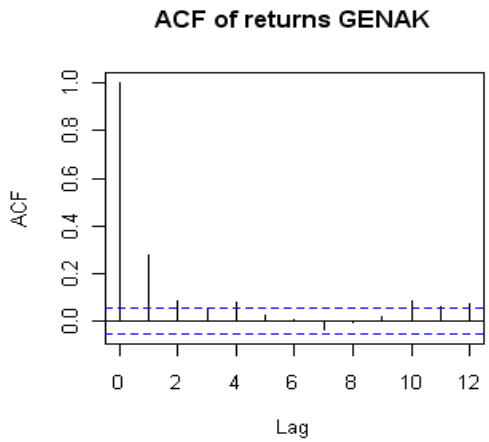


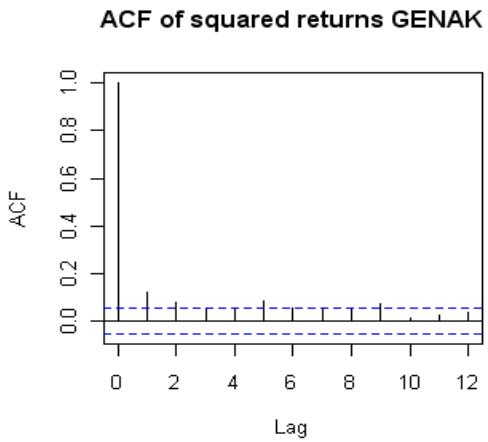
Εικόνα 3 Time series plots

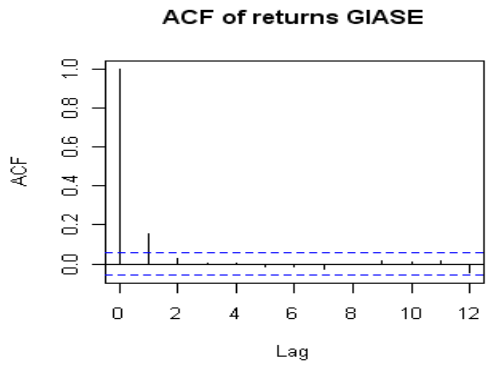
Παρακάτω ακολουθούν οι συναρτήσεις αυτοσυσχετίσεων για τις αποδόσεις και για τα τετράγωνα των αποδόσεων.

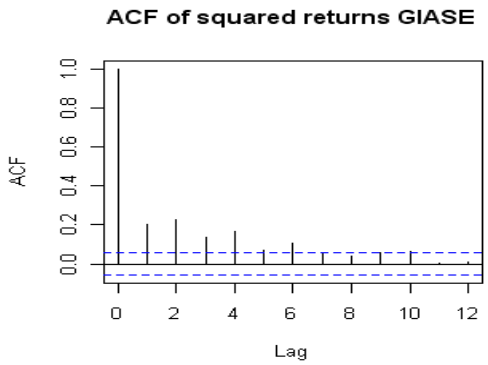

ACF of returns ALPHA
ACF
Lag

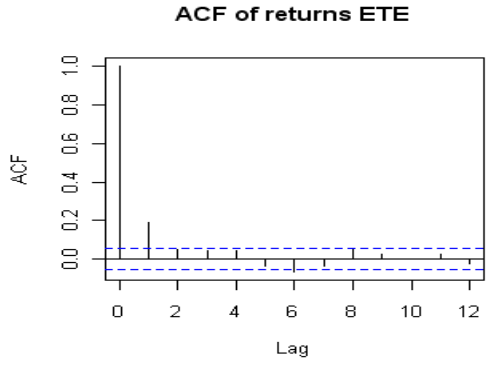

ACF of squared returns ALPHA
ACF
Lag

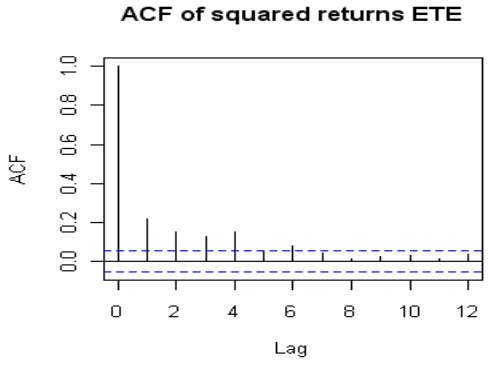

ACF of returns GENAK
ACF
Lag

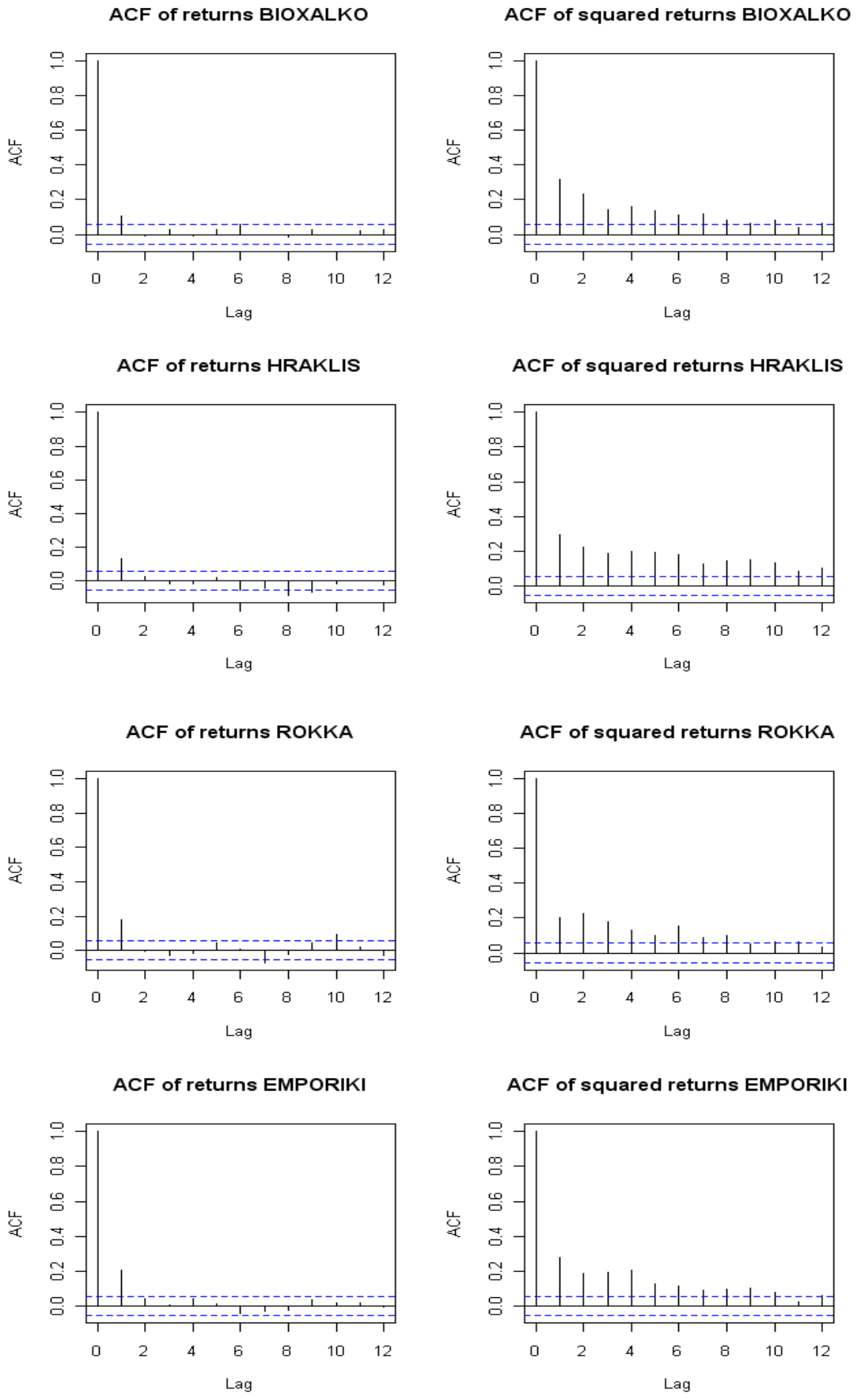

ACF of squared returns GENAK
ACF
Lag


ACF of returns GIASE
ACF
Lag


ACF of squared returns GIASE
ACF
Lag


ACF of returns ETE
ACF
Lag


ACF of squared returns ETE
ACF
Lag



Εικόνα 4 Συναρτήσεις αυτοσυσχετίσεων για αποδόσεις και τετράγωνα των αποδόσεων

2.3 Μεθοδολογία

2.3.1 Υπολογισμός παραμέτρων GARCH - Αποστάσεων

Όπως παρατηρούμε από τις συναρτήσεις αυτοσυσχετίσεων, και έχουμε αναφέρει παραπάνω, στα τετράγωνα των αποδόσεων (διορθωμένες για μέσο) όλων των σειρών έχουμε ισχυρή συσχέτιση ως προς τις προηγούμενες σειρές κάτι που μας οδηγεί να συμπεράνουμε ότι έχουμε μεταβλητή ως προς το χρόνο δεσμευμένη ετεροσκεδαστικότητα.

Εφαρμόζοντας ένα **AR(1)** με την συνάρτηση R, στις διορθωμένες αποδόσεις, έχουμε αφαιρέσει τον μέσο, παίρνουμε τους παρακάτω συντελεστές $\boldsymbol{\varphi_i}$ από την εξίσωση (9) για τις σειρές που εξετάζουμε:

Πίνακας 1 – Συντελεστές μοντέλου AR(1)

Σειρές	***Τιμή***
Γενικός δείκτης	0.156
Εθνική Τράπεζα	0.191
Alpha Bank	0.134
ΓΕΝΑΚ	0.278
Βιοχαλκο	0.105
Ηρακλής	0.134
ΡΟΚΚΑ	0.179
Εμπορική Τράπεζα	0.208

Από το τρέξιμο του **AR(1)** παραπάνω αποθηκεύουμε τα κατάλοιπα αφαιρούμε τις πρώτες τιμές που είναι μηδενικές. και προσαρμόζουμε ένα **GARCH(1,1)** με την εντολή garch() της R για κάθε σειρά δεδομένων.

Έχουμε τα παρακάτω αποτελέσματα από την εντολή **summary()** :

Πίνακας 2 – Αποτελέσματα μοντέλων GARCH

Χρονικές σειρές	coefficients				Jarque - Bera (Residuals)		Box - Ljung Test (Sqr Residuals)	
		a0	a1	b1	X-squared	p-value	X-squared	p-value
GIASE		3.00E-05	**1.88E-01**	**7.45E-01**	163.1867	< 2.2e-16	0.3334	0.5637
	t-value	4.839	8.461	25.451				
	Pr (>|t|)	1.31E-06	< 2e-16	< 2e-16				
ETE		8.34E-05	**1.74E-01**	**6.87E-01**	79.5044	< 2.2e-16	0.0778	0.7803
	t-value	4.905	6.644	15.535				
	Pr (>|t|)	9.32E-07	3.06E-11	< 2e-16				
ALPHA		7.63E-05	**1.67E-01**	**7.02E-01**	155.3369	< 2.2e-16	0.3639	0.5463
	t-value	4.336	6.788	15.687				
	Pr (>|t|)	1.45E-05	1.14E-11	< 2e-16				
GENAK		1.04E-05	**2.04E-01**	**8.26E-01**	725.6033	< 2.2e-16	3.6795	0.05509
	t-value	13.33	19.29	123.96				
	Pr (>|t|)	<2e-16	<2e-16	<2e-16				
BIOXALKO		1.39E-04	**2.24E-01**	**6.19E-01**	41.8692	8.10E-10	0.4847	0.4863
	t-value	5.067	5.768	10.984				
	Pr (>|t|)	4.04E-07	8.05E-09	< 2e-16				
Hraklis		3.97E-05	**1.71E-01**	**7.80E-01**	91.8239	< 2.2e-16	0.9517	0.3293
	t-value	5.464	8.13	36.02				
	Pr (>|t|)	4.66E-08	4.44E-16	< 2e-16				
Rokka		6.86E-05	**1.70E-01**	**7.82E-01**	4.5673	0.1019	0.2016	0.6535
	t-value	4.22	6.903	29.361				
	Pr (>|t|)	2.44E-05	5.09E-12	< 2e-16				
Emporiki		5.24E-05	**1.62E-01**	**7.73E-01**	187.1168	< 2.2e-16	0.0365	0.8484
	t-value	7.249	8.247	33.871				
	Pr (>|t|)	4.18E-13	2.22E-16	< 2e-16				

Έχοντας τις τιμές των a_1 και b_1 μπορούμε να χρησιμοποιήσουμε τη σχέση **(13)** για να δημιουργήσουμε τον πίνακα των αποστάσεων. Για τη δημιουργία του πίνακα κατασκευάσαμε μια συνάρτηση στην R που έχει σαν είσοδο δύο διανύσματα που περιέχουν τις τιμές των συντελεστών ***a*** και ***b*** του μοντέλου GARCH(1,1) και η οποία μας επιστρέφει έναν πίνακα των αποστάσεων τάξεως mxm όπου m το πλήθος των σειρών που θέλουμε να ομαδοποιήσουμε. Ο κώδικας της συνάρτησης υπολογισμού της απόστασης φαίνεται στο παράρτημα.

Από την συνάρτηση distestimate() (παράρτημα) που αποτελεί την εφαρμογή της σχέσης **(13)** παίρνουμε τον παρακάτω πίνακα αποστάσεων:

Πίνακας 3 – Αποστάσεις των σειρών

	Giase	ete	alpha	genak	bioxk	irak	rokka	emp
giase	0	0.053215	0.053509	0.104844	0.066409	0.025299	0.026824	0.032787
ete	0.053215	0	0.008363	0.156149	0.055542	0.06215	0.062877	0.048509
alpha	0.053509	0.008363	0	0.154863	0.063558	0.059053	0.059614	0.043732
genak	0.104844	0.156149	0.154863	0	0.158998	0.097619	0.097466	0.115972
bioxk	0.066409	0.055542	0.063558	0.158998	0	0.088779	0.090115	0.086022
irak	0.025299	0.06215	0.059053	0.097619	0.088779	0	0.001534	0.0189
rokka	0.026824	0.062877	0.059614	0.097466	0.090115	0.001534	0	0.018737
emp	0.032787	0.048509	0.043732	0.115972	0.086022	0.0189	0.018737	0

2.3.2 Υπολογισμός συστάδων

Εφαρμόζοντας στον παραπάνω πίνακα την συνάρτηση ***kmeans()*** της ***R*** για clustering με την μη – ιεραρχική μέθοδο kmeans παίρνουμε τα παρακάτω αποτελέσματα από την R.

	Giase	Ete	alpha	genak	bioxk	irak	rokka	emp
1	0.05336161	0.004181292	0.004181292	0.1555058	0.05955017	0.06060160	0.06124541	0.04612043
2	0.06640879	0.055542241	0.063558100	0.1589979	0.00000000	0.08877868	0.09011539	0.08602208
3	0.02122748	0.056687641	0.053976879	0.1039754	0.08283124	0.01143314	0.01177367	0.01760574
4	0.10484433	0.156148545	0.154862980	0.0000000	0.15899792	0.09761939	0.09746574	0.11597207

Clustering vector:

giase ete alpha genak bioxk irak rokka emp

3 1 1 4 2 3 3 3

Βλέπουμε ότι το clustering vector για αριθμό συστάδων ίσο με τέσσερα μας δίνει τις εξής συστάδες:

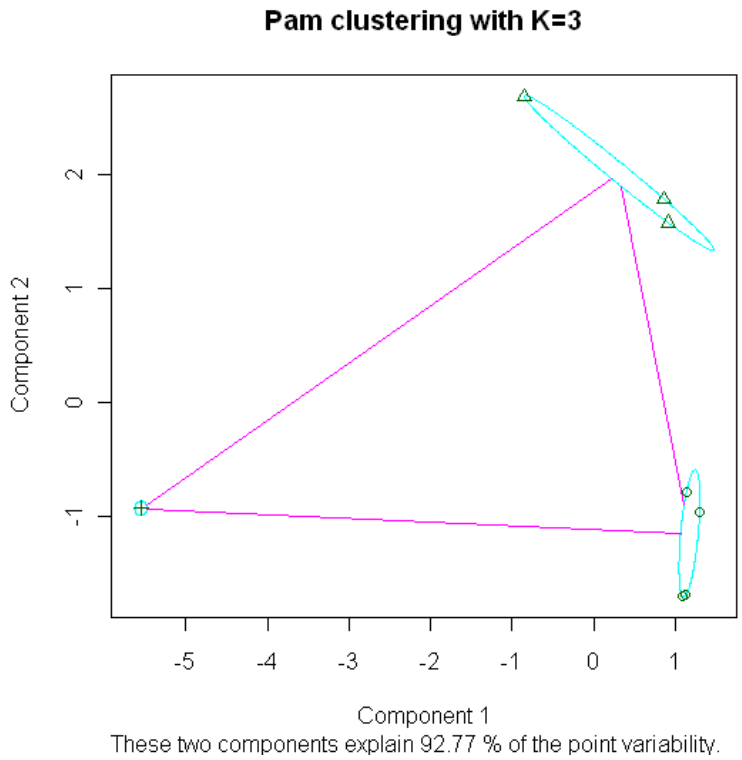
1. **Εθνική τράπεζα,alpha bank**
2. **Βιοχάλκο**
3. **Ηρακλής, Ροκκα, Γενικός δείκτης, Εμπορική τράπεζα**
4. **Γενακ**

Για αριθμό συστάδων ίσο με 3 με την ***kmeans*** παίρνουμε την συσταδοποίηση:

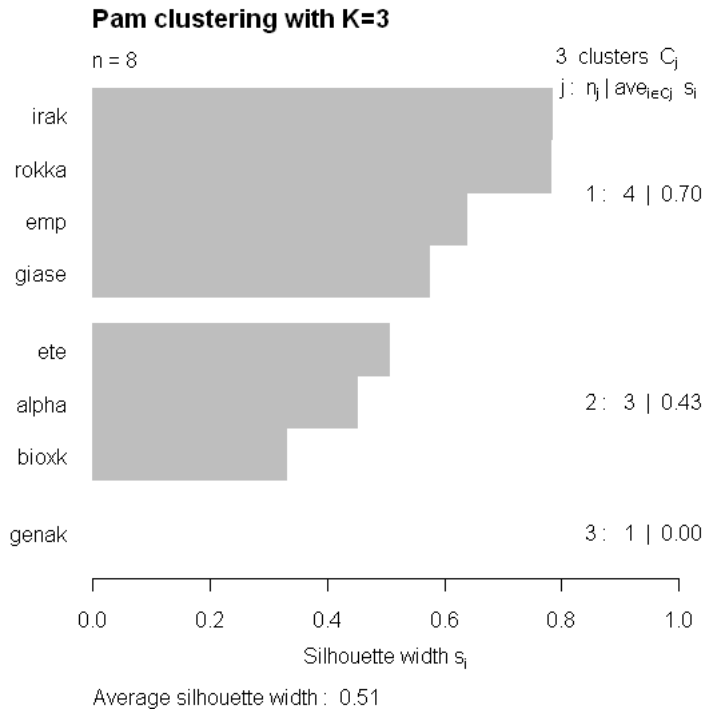
1. **Εθνική τράπεζα,alpha bank, Βιοχάλκο**
2. **Ηρακλής, Ροκκα, Γενικός δείκτης, Εμπορική τράπεζα**
3. **Γενακ**

Η διαφορά όπως βλέπουμε ανάμεσα στα δύο τρεξίματα είναι η προσθήκη της Βιοχάλκο στην συστάδα της Εθνικής και της alpha. Μπορούμε να υποθέσουμε αρχικά ότι η Εθνική με την Alpha αποτελούν έναν ισχυρό κόμβο καθώς και η Εμπορική με τον γενικό δείκτη, την Ροκκα και τον Ηρακλή έναν άλλο. Η Γενακ αποτελεί μόνη της συστάδα και στα δύο τρεξίματα κάτι που φανερώνει και την ιδιαιτερότητα στην συμπεριφορά της μετοχής.

Μια άλλη συνάρτηση που κάνει k-means clustering στην R είναι και pam(). Εξαιτίας της δυνατότητας της **pam()** δίνει και γραφικές παραστάσεις καθώς και περισσότερες δυνατότητες επεξεργασίας θα μας δώσει μια καλύτερη εποπτεία των αποτελεσμάτων που πήραμε από την k-means. Η pam() κάνει την συσταδοποίηση ελαχιστοποιώντας το άθροισμα των διαφορών (dissimilarity matrix) και όχι με το άθροισμα των τετραγώνων των ευκλείδειων αποστάσεων όπως η kmeans. Το ποιο χρήσιμο χαρακτηριστικό της είναι όμως οι γραφικές της δυνατότητες. Παρακάτω βλέπουμε τα αποτελέσματα της pam() για αριθμό συστάδων τρία:



Εικόνα 5 Clustering graph με την Pam() για K=3

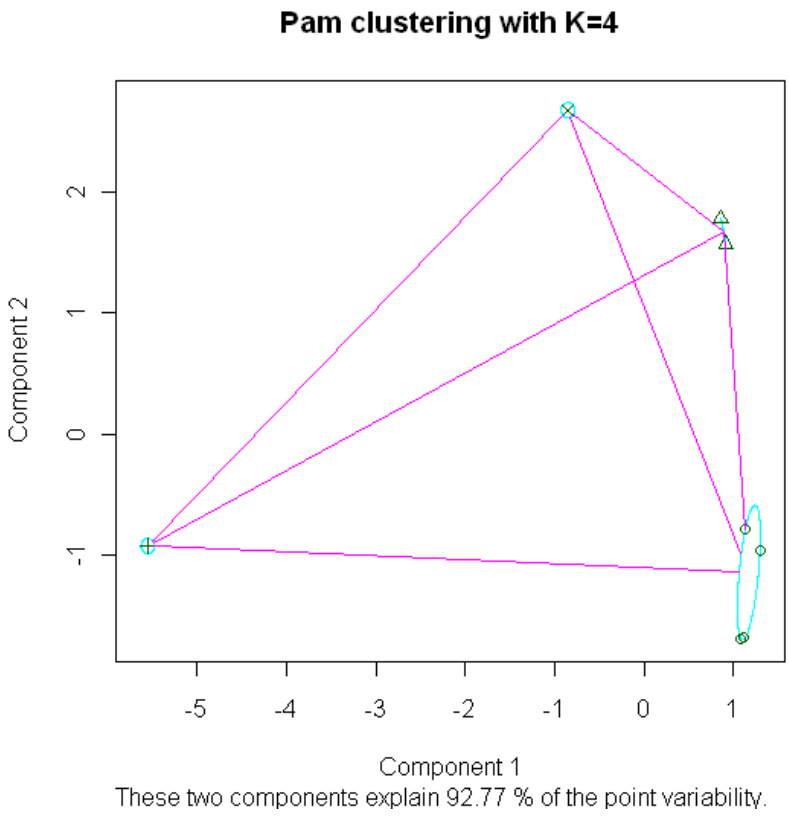


Εικόνα 6 Συστάδες και πλάτη με την Pam() για K=3

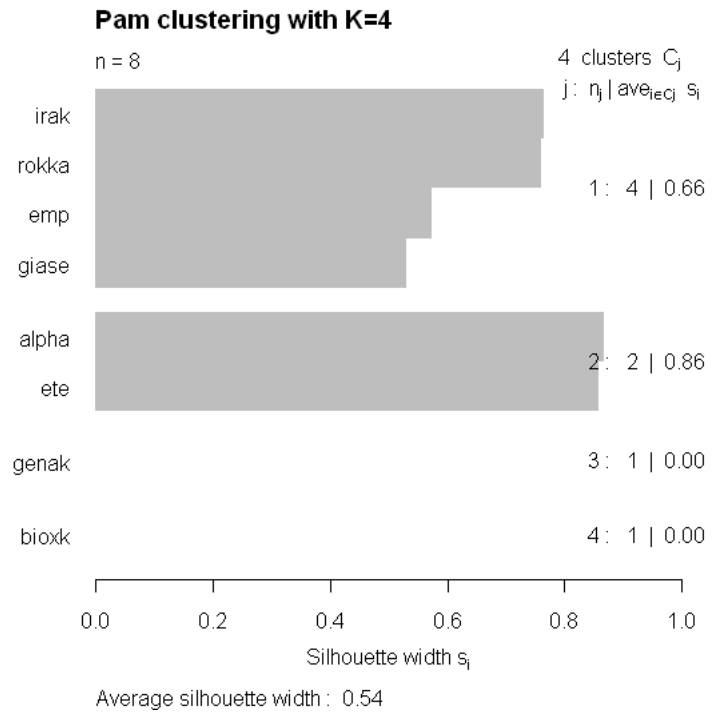
Στην εικόνα 6 βλέπουμε ένα είδος ραβδογράμματος όπου οι ράβδοι είναι ομαδοποιημένες ανά συστάδα και σε κάθε συστάδα η κάθε μετοχή που μελετούμε παίρνει μια τιμή s(i) (πλάτος) από 0 έως 1. Τιμές s(i) κοντά στη μονάδα σημαίνει ότι είναι πολύ καλά ομαδοποιημένες , τιμές κοντά στο 0 ότι η παρατήρηση βρίσκεται μεταξύ δύο συστάδων και αρνητικές τιμές ότι η παρατήρηση είναι τοποθετημένη σε λάθος συστάδα.

Παρατηρούμε ότι η πρώτη συστάδα με μέσο όρο 0.7 αποτελεί μια καλή εκτίμηση της συστάδας, ενώ για την δεύτερη συστάδα έχουμε ένα μέσο όρο 0.43 που δεν είναι αρκετά ικανοποιητικός κάτι που μας οδηγεί να υποθέσουμε ότι κάποια μετοχή δεν ανήκει σε αυτή την συστάδα.

Για αριθμό συστάδων ίσο με τέσσερα βλέπουμε τα παρακάτω:



Εικόνα 7 Clustering graph με την Pam() για K=4

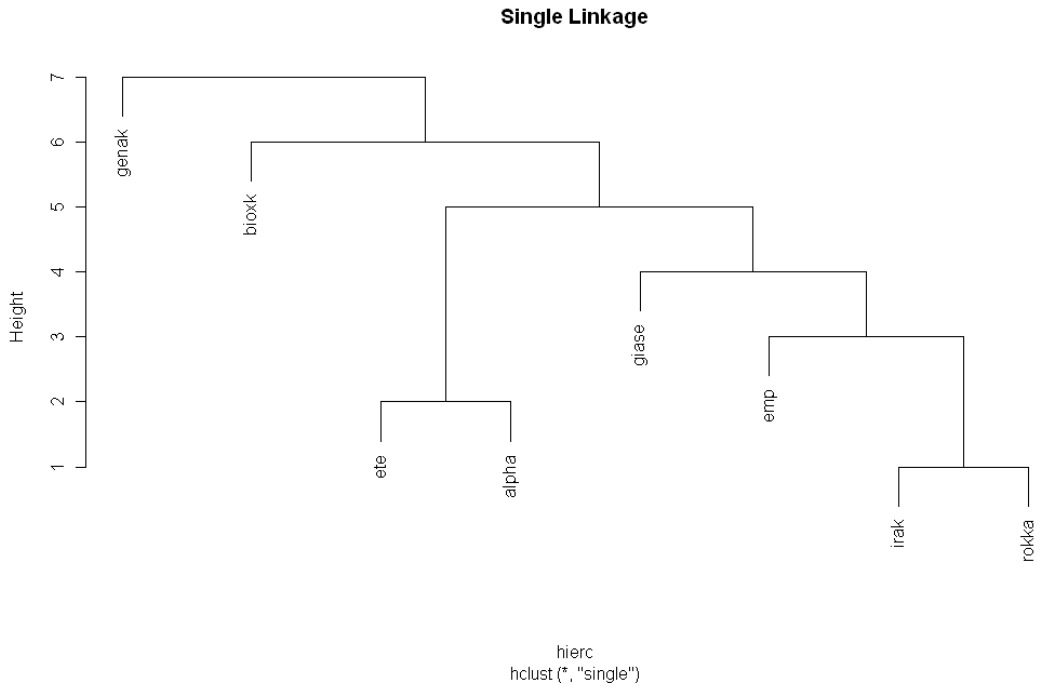


Εικόνα 8 Συστάδες και πλάτη με την Pam() για K=4

Για αριθμό συστάδων ίσο με τέσσερα ο αλγόριθμος μας δίνει ελαφρώς καλύτερα αποτελέσματα, καθώς έχουμε για την πρώτη συστάδα πλάτος **s(i) 0.66** από **0.70** που είχαμε για K=3, και έχουμε για την δεύτερη συστάδα. **0.86** πλάτος από **0.43** για K=3, αλλά με την προσθήκη της Βιοχάλκο στη συστάδα σε αυτή την περίπτωση.

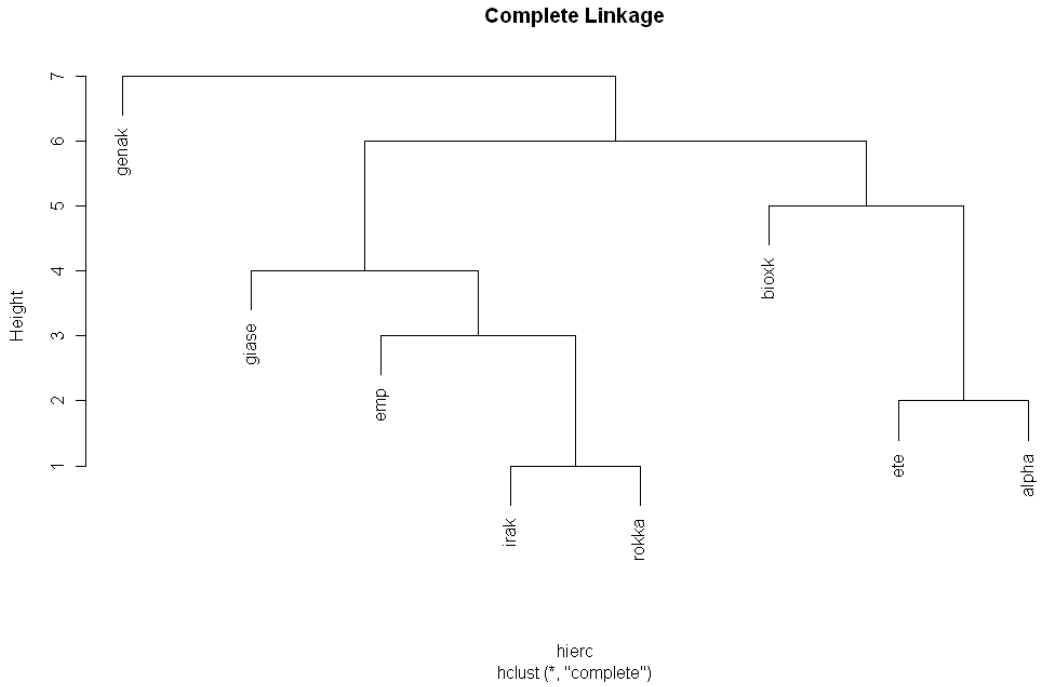
Θα εξετάσουμε τα αποτελέσματα που πήραμε για τρέξιμο με την συνάρτηση **hclust()** της ***R*** για υπολογισμό των συστάδων με ιεραρχικούς αλγορίθμους. Εξετάσαμε τα δεδομένα για δυο περιπτώσεις του αλγορίθμου για single – linkage και complete – linkage. Για να χρησιμοποιήσουμε την συνάρτηση hclust πρέπει πρώτα να μετατρέψουμε τον πίνακα των αποστάσεων σε αντικείμενο της μορφής dist με τη συνάρτηση dist της R. Η συνάρτηση dist υπολογίζει τις αποστάσεις μεταξύ των γραμμών ενός πίνακα δεδομένων.

Single – linkage



Εικόνα 9 Clustering graph με τον Single-linkage αλγόριθμο

Complete Linkage



Εικόνα 10 Clustering graph με τον Complete-linkage αλγόριθμο

Για τον ***single – linkage*** αλγόριθμο έχουμε:

Στο πρώτο βήμα του αλγορίθμου η Ηρακλής πηγαίνει με την Ροκκα, στο $2^{ο}$ η ΕΤΕ με την άλφα, στο $3^{ο}$ και στο $4^{ο}$ βήμα μπαίνουν η Εμπορική και ο γενικός δείκτης στην συστάδα του Ηρακλή και τέλος προστίθεται η Βιοχάλκο και η Γενάκ στο σύνολο των αντικειμένων και όχι σε κάποια συγκεκριμένη συστάδα. Παρατηρούμε εδώ δηλαδή ότι αν σταματήσουμε τον αλγόριθμο σε ένα βάθος 4 και πάρουμε τα αποτελέσματα έχουμε την ίδια διαρρύθμιση των συστάδων όπως και με την kmeans και pam για αριθμό συστάδων ίσο με 4. Η Βιοχάλκο όμως, ακόμα και βάθος 5 να επιλέγαμε, δεν θα πήγαινε στη συστάδα της Εθνικής με της Alpha αλλά θα άνηκε στη συνολική συστάδα με όλες τις παρατηρήσεις.

Για τον ***complete - linkage*** αλγόριθμο έχουμε:

Παρατηρούμε ότι ακολουθούνται ακριβώς τα ίδια βήματα με παραπάνω με την διαφορά ότι στο $5^{ο}$ βήμα η Βιοχάλκο πηγαίνει στη συστάδα της Εθνικής με την Alpha. Για αυτή την περίπτωση τα αποτελέσματα ταυτίζονται με αυτά που πήραμε από τους αλγόριθμους με την K-means για αριθμό συστάδων 3 και 4.

Κεφάλαιο 3

Συμπεράσματα – Σύγκριση αποτελεσμάτων

Αναλύσαμε παραπάνω την μεθοδολογία που ακολουθήσαμε ώστε να φτάσουμε σε ένα συμπέρασμα όσον αφορά την τελική δομή των συστάδων. Ορίστηκε ένα μέτρο απόστασης **(13)** και βάσει αυτού του μέτρου και μέσω των μοντέλων GARCH(1,1) πήραμε έναν πίνακα αποστάσεων των μετοχών που μελετήσαμε. Εφαρμόσαμε σε αυτόν τον πίνακα τέσσερις διαφορετικούς αλγορίθμους , δύο από hierarchical clustering με τους single – linkage και complete – linkage αλγόριθμους και δύο από non- hierarchical με δυο παραλλαγές του k-means αλγορίθμου για συστάδες μεγέθους 3 και 4. Μπορούμε να πούμε όσον αφορά τις τελικές συστάδες και την σύνθεσή τους, ότι υπάρχουν δυο ισχυρές ομάδες η μία που περιέχει την Εθνική και την Alpha και άλλη μια που περιέχει την Ηρακλής, Ρόκκα, Εμπορική και Γενικό Δείκτη. Η Γενακ σε όλες τις εκτελέσεις του αλγορίθμου αποτελεί μια συστάδα μόνη της και η Βιοχάλκο ανάλογα με το πόσες συστάδες θέλουμε να σχηματίσουμε είναι είτε μόνη της μια συστάδα ή ανήκει στην συστάδα της Εθνικής με την Alpha για τους περισσότερους αλγόριθμους. Συνολικά έχουμε τον παρακάτω πίνακα:

Πίνακας 4 – Περίληψη αποτελεσμάτων αλγορίθμων

Αριθμός cluster	***K-means***		***Pam***		***Single-Linkage***		***Complete-linkage***	
	3	***4***	***3***	***4***	***3****	***4***	***3***	***4***
1	ETE,alpha, βιοχάλκο	ETE,alpha	ETE,alpha, βιοχάλκο	ETE,alpha	-	ETE,alpha	ETE,alpha, βιοχάλκο	ETE,alpha
2	Γεν.δείκτης, ηρακ,ροκκα εμπορική	Γεν.δείκτης, ηρακ,ροκκα εμπορική	Γεν.δείκτης, ηρακ,ροκκα εμπορική	Γεν.δείκτης, ηρακ,ροκκα εμπορική	-	Γεν.δείκτης, ηρακ,ροκκα εμπορική	Γεν.δείκτης, ηρακ,ροκκα εμπορική	Γεν.δείκτης, ηρακ,ροκκα εμπορική
3	Γενακ	Βιοχάλκο	Γενακ	Βιοχάλκο	-	Βιοχάλκο	Γενακ	Βιοχάλκο
4	-	Γενακ	-	Γενακ	-	Γενακ	-	Γενακ

**Με τον single – linkage δεν μπορούμε να εξάγουμε 3 συστάδες. Αν θέλουμε να έχουμε πιο λίγες συστάδες μπορούμε να επιλέξουμε 2 όπου η μια είναι η γενακ και η δεύτερη όλες οι υπόλοιπες σειρές.*

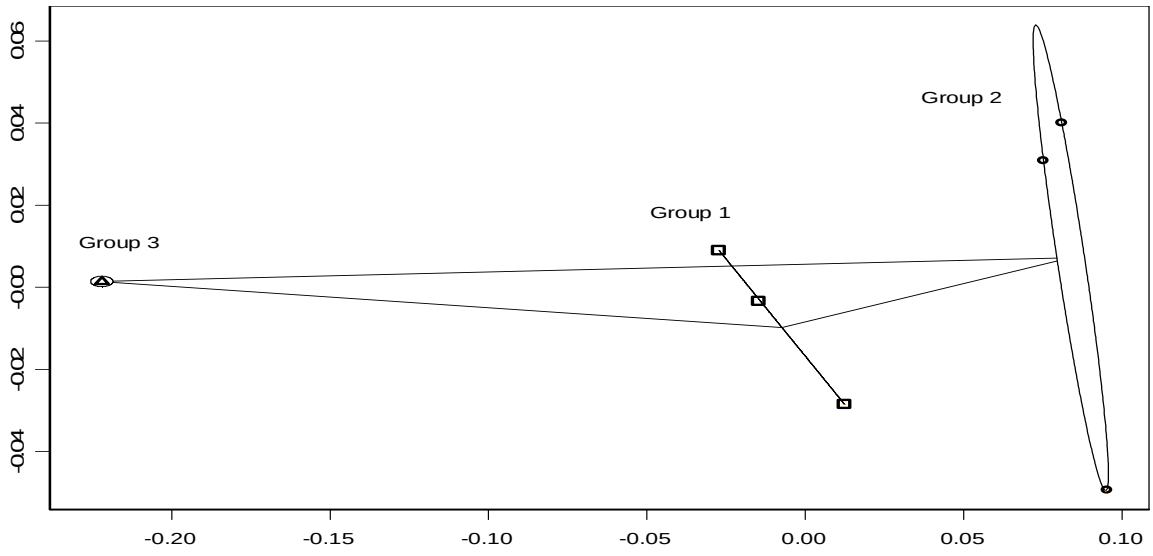
Η διαδικασία που ακολουθήθηκε παραπάνω είχε ως στόχο να ομαδοποιηθούν χρονικές σειρές ανάλογα με τις δομές αστάθειας που παρουσιάζουν και έτσι ώστε τελικά να έχουμε όσο το δυνατό πιο ομογενείς συστάδες.

Συγκρίναμε τα αποτελέσματα με εργασία **[10] Παπαναστασίου**, όπου έγινε επίσης ταξινόμηση και συσταδοποίηση για χρονικές σειρές που παρουσιάζουν δομές αστάθειας. Σε αυτή χρησιμοποιήθηκε ένα μέτρο, **[5] Kakizawa,** το Kullback – Leibler μέτρο, το οποίο στη θεωρία πιθανοτήτων και τη θεωρία πληροφοριών είναι ένα μη-συμμετρικό μέτρο της διαφοράς μεταξύ δύο κατανομών P και Q. Αυτό το μέτρο εφαρμόστηκε για μοντέλα GARCH και παρατηρήσαμε ότι:

Για τις σειρές Γενακ, Γενικός δείκτης, Ηρακλής, Βιοχάλκο, ΕΤΕ , alpha και Ροκκα οι συστάδες που μας έδωσε ο αλγόριθμος ***k-means*** με την συνάρτηση ***pam*** ήταν

- Γενικός δείκτης , ηρακλής και βιοχάλκο
- ΕΤΕ, alpha και Ροκκα
- Γενακ

Η γραφική παράσταση φαίνεται παρακάτω:



Παρατηρούμε ότι οι συστάδες έχουν σχεδόν τα ίδια κέντρα, η πρώτη τον γενικό δείκτη και η δεύτερη την ΕΤΕ , και οι αλλαγές είναι η προσθήκη της Βιοχάλκο από την συστάδα της ΕΤΕ σε αυτή του Γενικού δείκτη και η προσθήκη της Ροκκα από τον γενικό δείκτη στην ομάδα της ΕΤΕ. Τα κέντρα αποτελούν τις ισχυρότερες μετοχές από τις συστάδες που δημιουργούνται και μπορούμε να τις κατατάξουμε σε συστάδες ρίσκου, ανάγοντας την συστάδα της ΕΤΕ σε χαμηλού ρίσκου, την συστάδα του Γεν. Δείκτη σε μέσου ρίσκου και την ομάδα της ΓΕΝΑΚ σε υψηλού ρίσκου.

Παράρτημα – Γλώσσα R

Η γλώσσα R είναι μια γλώσσα προγραμματισμού βασισμένη στο λογισμικό S-Plus, που αναπτύχθηκε με κύριο στόχο να γίνει ένα εργαλείο στατιστικής ανάλυσης και επεξεργασίας δεδομένων. Είναι μια γλώσσα ελεύθερου λογισμικού με πολλά πακέτα διαθέσιμα στους χρήστες για την εκτέλεση πολλών στατιστικών (στατιστικά τέστ, γραμμική και μη – γραμμική ανάλυση, ανάλυση χρονοσειρών, ταξινόμηση και συσταδοποίηση) και γραφικών τεχνικών. Στη βιβλιογραφία υπάρχουν πολλά αξιόλογα εγχειρίδια εκμάθησης της γλώσσας. **[11] Paradis.** Μπορεί να βρεθεί στην σελίδα **www.cran.r-project.org**

Η R εκτελείται μέσω command line αν και έχουν αναπτυχθεί και gui. Στο command line μπορούμε να γράφουμε τις εντολές, πιο αποδοτικό όμως είναι να γράφουμε όλες μαζί τις εντολές σε ένα script, και να τις εκτελούμε. Επίσης υπάρχει η δυνατότητα να γράψουμε τις δικιές μας συναρτήσεις δίνοντας μας μεγαλύτερη ευελιξία και περισσότερες δυνατότητες.

Κάποιες βασικές εντολές που χρησιμοποιήσαμε είναι:

Για την εισαγωγή αρχείων άλλης μορφής (.xls,.csv,doc) είναι η εντολή **read.csv()** ή **read()**, ή **read.table()**. Όλες είναι πλήρως παραμετροποιημένες και η πλήρης σύνταξη τους μπορεί να βρεθεί με το help της R, γράφοντας help(read.csv) ή οποιαδήποτε εντολή θέλουμε να μάθουμε την σύνταξή της. Για να χρησιμοποιήσουμε κάποια εντολή ή συνάρτηση που δεν βρίσκεται στο βασικό πακέτο της R πρέπει πρώτα να την φορτώσουμε στο workspace με την εντολή **library()** βάζοντας το όνομα του πακέτου στις παρενθέσεις. Με την συνάρτηση **objects()** βλέπουμε ποια αντικείμενα υπάρχουν αποθηκευμένα στην μνήμη. Με την **rm()** διαγράφουμε κάποιο αντικείμενο. Με την ***names(object_name)*** παίρνουμε ονομαστικά τα στοιχεία ενός αντικειμένου. Με την ***summary(object_name)*** παίρνουμε μια περίληψη του τι περιέχει ένα αντικείμενο, συνήθως στατιστικά για το ελάχιστο, μέγιστο και τα τεταρτημόρια αν είναι πίνακας ή διάνυσμα ή αν το αντικείμενο έχει προέλθει από στατιστική ανάλυση, garch() π.χ., περισσότερες λεπτομέρειες σχετικά με την εκτέλεση της μεθόδου (στατιστικά τεστ, p- values, τυπικές αποκλίσεις.

Για τη δημιουργία λίστας αναφέρονται ενδεικτικά δύο τρόποι:
Με την εντολη ***a<-c(1,2,3,4)*** όπου δημιουργείται μια λίστα με 4 στοιχεία.
Ή με την εντολή ***a<-list(A=x,B=y)*** όπου δίνουμε ονόματα στα στοιχεία της λίστας για να μας διευκολύνουν στην αναζήτηση καθώς και να είναι πιο ξεκάθαρο τι ακριβώς αντιπροσωπεύει κάθε στοιχείο. Κάθε λίστα μπορεί να έχει ως στοιχεία άλλες λίστες, διανύσματα ή πίνακες.
Για την δημιουργία πίνακα χρησιμοποιήσαμε την εντολή matrix().

Η R περιέχει δομές ελέγχου και επαναλήψεων όπως:
if (condition) {expression1} **else {**expression2} : Αν ικανοποιηθεί η συνθήκη εκτελείται η σχέση 1 αλλιώς η σχέση 2.
for (x in expression1) {expression2} : Επαναλαμβάνει την σχέση 2 όσο ισχύει η σχέση 1. Η σχέση 1 είναι της μορφής 1:n συνήθως δίνοντας έτσι ένα διάστημα που θα τρέξει το loop. Αντίστοιχα ορίζονται και οι δομές με while, switch, repeat. Χρήσιμες είναι και οι εντολές apply και sapply για αποδοτικότερη διαχείριση διανυσμάτων και πινάκων, οι οποίες εφαρμόζουν σε όλα τα στοιχεία ενός πίνακα μια συνάρτηση.

Χρησιμοποιούμενες συναρτήσεις

Μια από τις συναρτήσεις που χρησιμοποιήσαμε αρχικά ήταν η **as.POSIXct()** που παίρνει ως δεδομένο ένα αρχείο τύπου factor που στην περίπτωσή μας ήταν η σειρά του χρόνου των αρχικών δεδομένων, και την μετατρέπει σε μορφής date. Αυτό το κάναμε για να μπορούμε να παίρνουμε την γραφική παράσταση κάποιας χρονοσειράς έχοντας και την χρονική κλίμακα στον άξονα των x.
Μια συνάρτηση που δημιουργήσαμε ήταν η arestimate() για να υπολογίσουμε τους συντελεστές που προέκυψαν από το AR(1) καθώς και τα Residuals.

```
arestimate<-function(X){
  a<-ar(X,aic=FALSE,order.max=1)
  coef<-a$ar
  al<-list(Coefficient=coef,Residuals=na.omit(a$resid))
}
```

Αυτό που επιστρέφεται στο τέλος είναι μια λίστα που έχει για στοιχεία τον συντελεστή του AR και τα κατάλοιπα.

Η εντολή **ar()** υπολογίζει ένα ar μοντέλο προσφέροντας την κατάλληλη παραμετροποίηση. Επίσης οι **acf()** και **pacf()** υπολογίζουν τις συναρτήσεις αυτοσυσχετίσεων και τις μερικές συναρτήσεις αυτοσυσχετίσεων αντίστοιχα εκτυπώνοντας και το αντίστοιχο γράφημα.
Για να τρέξουμε ένα garch(p,q) σε μια σειρά x δίνουμε την εντολή ***garch(x, order = c(1, 1) .***

Για clustering χρησιμοποιήσαμε τις συναρτήσεις :

- ***kmeans():*** που εφαρμόζεται πάνω σε ένα πίνακα, και ορίζουμε επίσης και τον αριθμό των συστάδων.
- ***pam():*** Επίσης συνάρτηση με τον αλγόριθμο kmeans(), με παρόμοια σύνταξη αλλά περισσότερες δυνατότητες, όπως γραφική αναπαράσταση των συστάδων.
- ***hclust():*** Συνάρτηση που υπολογίζει το clustering vector βάσει αλγορίθμου hierarchical clustering , single-linkage ή complete – linkage. Προσφέρει επίσης δυνατότητες γραφικής αναπαράστασης. Για να μπορέσουμε να πάρουμε το γράφημα εφαρμόζουμε αρχικά την συνάρτηση ***dist()*** στον πίνακα των αποστάσεων, έπειτα την συνάρτηση
  hclust(d,method=”single or complete”) όπου d είναι το αποτέλεσμα της dist() και τέλος για να πάρουμε το γράφημα την συνάρτηση plclust(h) όπου h το αποτέλεσμα από την hclust().

Βιβλιογραφία

[1] Agrawal R., Faloutsos C. and Swami A. (1994). Efficient similarity search in sequence databases. Lecture notes in Computer Science, 69-84

[2] Bollerslev T. (1986). Generalized Autoregressive Conditional Heteroskedasticity, Journal of Econometrics, 31, 307-321

[3] Engle R.F. (1982). Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation, Econometrica, 50, 987 – 1008

[4] Johnson R.A., Wichern D.W, (1992). Applied Multivariate Statistical Analysis, Prentice Hall, 738-760

[5] Kakizawa Y., Shumway R.H., Taniguchi M. (1988). Discrimination and Clustering for Multivariate Time Series[, Journal of the American Statistical Association 93, 328 – 340

[6] Liao T. (2005). Clustering Time Series Data: A Survey, Pattern Recognition, 38,1857-1874

[7] MacQueen J.B.(1967) Some Methods for Classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability,1,Berkeley,CA: University of California Press,281-297

[8] Nelson D.B. (1991). Conditional Heteroskedasticity in Asset Returns : A New Approach, Econometrica, 59(2), 347-370

[9] Otrando E. (2004). Classifying the Markets Volatility with ARMA Distance Measures, Quaderni di Statistica. V6. 1-19

[10]Papanastasiou D. (2009). Classification and Clustering of GARCH Time Series, ASMDA

[11] Paradis E. (2002) R for Beginners, Université Montpellier II

[12] Piccolo D. (1990). A Distance Measure for Classifying ARIMA Models, Journal of time Series Analysis, 11, 153-64.

[13] Teräsvirta T.,He . (1999a). Properties of Moments of a Family of GARCH processes, Journal of Econometrics, 92, 173-192

[14] Tsay R.S. (2002). Analysis of Financial Time Series, JohnWiley and sons

[15] Zivot E., Wang J. (2003). Modeling Financial time Series with S-Plus, Springer Verlag