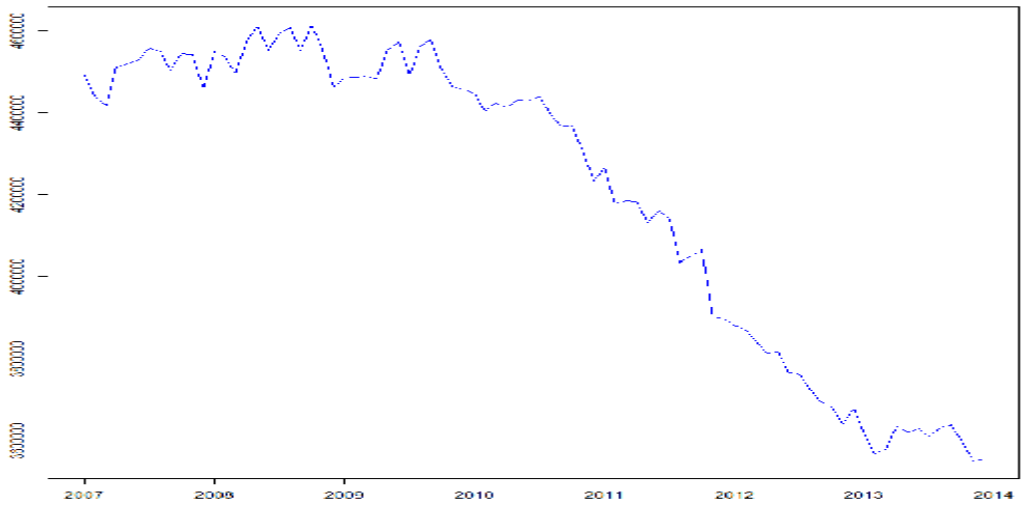
ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

(ΠΡΩΗΝ)ΤΜΗΜΑ ΕΦΑΡΜΟΓΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΔΙΟΙΚΗΣΗ ΚΑΙ ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ
ΤΜΗΜΑ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ

«Εφαρμογή τεχνικών Εξόρυξης Δεδομένων στην ανάλυση δεδομένων χρονικών σειρών»



ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΤΩΝ:

Γκέλιου Αντιγόνη Νικιέλ Μαγκνταλένα Φύτρος Σαμψών

ΕΠΟΠΤΗΣ ΚΑΘΗΓΗΤΗΣ:

Αντζουλάτος Γεράσιμος

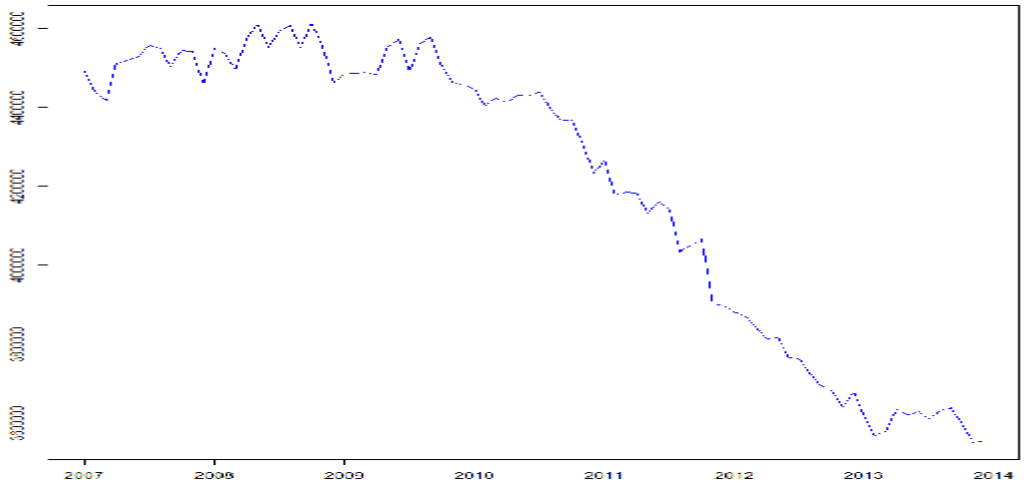
ΠΑΤΡΑ, ΙΟΥΝΙΟΣ 2015

TECHNOLOGICAL EDUCATIONAL INSTITUTION OF WESTERN GREECE

TECHNOLOGICAL INSTITUTE OF MANAGEMENT AND ECONOMICS

DEPARTMENT OF APPLIED COMPUTING ON MANAGEMENT AND THE ECONOMY

«Application of Data Mining techniques in time series data analysis»



DIPLOMA THESIS:

Geliou Antigoni **Nikiel Magdalena** **Fytros Sampson**

12674 **12673** **12671**

SUPERVISED PROFESSOR:

Antzoulatos Gerasimos

PATRA, JUNE 2015

ΕΥΧΑΡΙΣΤΙΕΣ

Θα θέλαμε να ευχαριστήσουμε θερμά τον κ. Γεράσιμο Αντζουλάτο, καθηγητή του τμήματος Διοίκησης Επιχειρήσεων, για την καθοδήγηση και υποστήριξη κατά την υλοποίηση της παρούσας εργασίας, αλλά και για την πολύ καλή συνεργασία που είχαμε όλο αυτό το διάστημα. Επίσης, θέλουμε να ευχαριστήσουμε τις οικογένειες μας για την στήριξη έως την ολοκλήρωση των σπουδών μας.

ΠΕΡΙΛΗΨΗ

Η ανάλυση χρονοσειρών αποσκοπεί στην ανίχνευση εκείνων των χαρακτηριστικών που συμβάλουν στην κατανόηση της ιστορικής συμπεριφοράς μιας μεταβλητής και επιτρέπουν την πρόβλεψη μελλοντικών τιμών της. Τα δεδομένα χρονικών σειρών είναι ένας ειδικός τύπος ακολουθιακών δεδομένων στα οποία η κάθε εγγραφή είναι μια χρονική σειρά, δηλαδή μια σειρά μετρήσεων οι οποίες γίνονται με τη πάροδο του χρόνου. Χαρακτηριστικά παραδείγματα δεδομένων χρονικών σειρών είναι οι τιμές κλεισίματος των μετοχών στα χρηματιστήρια, τα δεδομένα του ρυθμού μεταβολής του ΑΕΠ κτλ. Η ανάγκη πρόβλεψης εμφανίζεται σε πολλά προβλήματα λήψης αποφάσεων. Χαρακτηριστικά παραδείγματα είναι ο προγραμματισμός παραγγελιών μιας εταιρείας που εμπορεύεται ένα προϊόν στηρίζεται σε προβλέψεις της ζήτησης του προϊόντος, η επένδυση σε μία ή περισσότερες μετοχές ενός ιδιώτη ή μιας επιχείρησης στηρίζεται σε προβλέψεις των μελλοντικών τιμών των αξιών των μετοχών και των επιτοκίων. Η πρόβλεψη μελλοντικών συμπεριφορών στηρίζεται στην ανάλυση παρατηρήσεων που αναφέρονται στο παρελθόν (ιστορικά δεδομένα). Στην παρούσα πτυχιακή εργασία θα αναλυθεί ο τρόπος εφαρμογής τεχνικών Εξόρυξης Δεδομένων στην ανάλυση και πρόβλεψη χρονοσειρών με σκοπό την υποβοήθηση της διαδικασίας λήψης αποφάσεων. Για την υλοποίηση της εφαρμογής των τεχνικών ανάλυσης και πρόβλεψης χρονοσειρών θα χρησιμοποιηθεί το προγραμματιστικό πακέτο R.

ABSTRACT

Time series analysis is intended to detect those characteristics which contribute to the understanding of the historical behavior of one variable and allow future values of prediction. The time series data is a special type of sequential data in which each entry is a time series, that means is a series of measurements which are made over time. Examples of time series data are the closing prices on the stock markets, the GDP rate of change data etc. The need to provide appears in many decision-making problems. Typical examples are scheduling orders a company that sells a product based on forecasts of demand for the product, investing in one or more shares of a private person or a company based on forecasts of future prices, securities, shares and interest rates. The future behavior prediction based on analysis of the observations provided in the past (historical data). In this diploma thesis will analyze how to implement Data Mining techniques in analyzing and forecasting time series in order to assist the decision making process. To implement the application of technical analysis and forecasting of time series will use the programming package R.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ vi

ΕΥΡΕΤΗΡΙΟ ΔΙΑΓΡΑΜΜΑΤΩΝ vii

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ viii

ΕΙΣΑΓΩΓΗ 1

ΚΕΦΑΛΑΙΟ 1^{ο} – ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ 2

1.1. Ορισμός εξόρυξης δεδομένων 2

1.2. Το περιεχόμενο της εξόρυξης δεδομένων 2

1.3. Τύποι δεδομένων 3

1.4. Εξόρυξη Δεδομένων και Ανακάλυψη της Γνώσης 4

1.5. Μεθοδολογίες Εξόρυξης Δεδομένων 7

1.5.1. Προγνωστικού Τύπου 7

1.5.2. Περιγραφικού Τύπου 10

1.6. Γιατί είναι σημαντική η εξόρυξη δεδομένων 12

ΚΕΦΑΛΑΙΟ 2^{ο} – ΑΝΑΛΥΣΗ ΧΡΟΝΟΣΕΙΡΩΝ ΜΕ ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ 14

2.1. Τι είναι οι χρονικές σειρές και που εφαρμόζονται 14

2.2. Στόχοι της ανάλυσης χρονοσειρών 16

2.3. Διάφοροι τύποι χρονοσειρών 17

2.4. Εξόρυξη δεδομένων των χρονικών σειρών 18

2.4.1. Συσταδοποίηση χρονοσειρών (Clustering) 18

2.4.2. Κατηγοριοποίηση - Ταξινόμηση χρονοσειρών (Classification) 28

2.4.3. Πρόβλεψη χρονοσειρών 32

2.4.4. Τμηματοποίηση χρονοσειρών (Segmentation) 43

2.4.5. Σύνοψη χρονοσειρών (Summerization) 44

2.4.6. Ανίχνευση ανωμαλιών χρονοσειρών (Anomaly Detection) 48

2.4.7. Ευρετηριοποίηση χρονοσειρών (Indexing) 49

ΚΕΦΑΛΑΙΟ 3^{ο} - ΧΡΗΣΗ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΟΥ ΠΑΚΕΤΟΥ R ΣΤΗΝ ΑΝΑΛΥΣΗ ΧΡΟΝΟΣΕΙΡΩΝ 52

3.1. Λόγοι χρήσης της R 52

3.2. Εφαρμογή της συσταδοποίησης 53

3.3. Εφαρμογή της πρόβλεψης σε χρονοσειρά 65

3.3.1. Πρόβλεψη με μοντέλο SARIMA 67

3.3.2. Πρόβλεψη με Νευρωνικά Δίκτυα (NN) 71

3.3.3. Σύγκριση του μοντέλου SARIMA με τα Νευρωνικά Δίκτυα 74

ΣΥΜΠΕΡΑΣΜΑΤΑ 76

ΒΙΒΛΙΟΓΡΑΦΙΑ 78

ΠΑΡΑΡΤΗΜΑ I 79

ΠΑΡΑΡΤΗΜΑ II 90

ΠΑΡΑΡΤΗΜΑ III 96

ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ

Εικόνα 1 Τα βήματα της διαδικασίας ανεύρεσης γνώσης σε βάσεις δεδομένων 6

Εικόνα 2 Γραμμική παλινδρόμηση 8

Εικόνα 3 Μη-γραμμική παλινδρόμηση 9

Εικόνα 4 Εσωτερική δομή μιας χρονοσειράς 16

Εικόνα 5 Ελάχιστη απόσταση μεταξύ συστάδων 22

Εικόνα 6 Μέγιστη απόσταση μεταξύ συστάδων 23

Εικόνα 7 Μέση απόσταση μεταξύ των συστάδων.....23

Εικόνα 8 Απόσταση κέντρων βάρους δύο συστάδων.....24

Εικόνα 9 Παράδειγμα χωρισμού των δεδομένων σε συστάδες/ομάδες.....26

Εικόνα 10 Παράδειγμα τυπικού δέντρου απόφασης 29

Εικόνα 11 Παράδειγμα κοντινότερων γειτόνων ενός σημείου.....30

Εικόνα 12 Αρχιτεκτονική ενός νευρωνικού δικτύου για την πρόβλεψη χρονοσειράς....40

Εικόνα 13 Παράδειγμα τμηματοποίησης χρονοσειράς 44

Εικόνα 14 Παράδειγμα προσέγγισης Αναζήτησης σε Χρόνο για ένα οπτικό ερώτημα..45

Εικόνα 15 Παράδειγμα ενός συστήματος απεικόνισης 46

Εικόνα 16 Η προσέγγιση οπτικοποίησης Σπιράλ 47

Εικόνα 17 Παράδειγμα προσέγγισης με το Δέντρο Viz 48

Εικόνα 18 Παράδειγμα Ανίχνευσης ανωμαλιών 48

Εικόνα 19 Παράδειγμα μείωσης διαστάσεων των χρονοσειρών σε δύο διαστάσεις.....50

Εικόνα 20 Παράδειγμα ιεραρχικής οργάνωσης χρησιμοποιώντας ένα R-tree 51

ΕΥΡΕΤΗΡΙΟ ΔΙΑΓΡΑΜΜΑΤΩΝ

Διάγραμμα 1 Σύγκριση Ελλάδα - Ουγγαρία.....54

Διάγραμμα 2 Σύγκριση Ελλάδα – Ηνωμένο Βασίλειο.....54

Διάγραμμα 3 Συσταδοποιήση Ελλάδα - Πορτογαλία.....56

Διάγραμμα 4 Συσταδοποίηση με 10 χώρες της Ευρωπαϊκής Ένωσης.....57

Διάγραμμα 5 Συντελεστής περιγράμματος με την Ευκλείδεια απόσταση.....58

Διάγραμμα 6 Συντελεστής περιγράμματος με την απόσταση Manhattan.....59

Διάγραμμα 7 Δενδρόγραμμα ιεραρχικής συσταδοποίησης απλού συνδέσμου.....60

Διάγραμμα 8 Εκτίμηση της ιεραρχικής συσταδοποίησης απλού συνδέσμου.....61

Διάγραμμα 9 Δενδρόγραμμα ιεραρχικής συσταδοποίησης πλήρους συνδέσμου.....61

Διάγραμμα 10 Εκτίμηση της ιεραρχικής συσταδοποίησης πλήρους συνδέσμου.....62

Διάγραμμα 11 Δενδρόγραμμα ιεραρχικής συσταδοποίησης κέντρων βάρους.....63

Διάγραμμα 12 Εκτίμηση της ιεραρχικής συσταδοποίησης με μέθοδο κέντρων βάρους.63

Διάγραμμα 13 Απεικόνιση χρονοσειράς για τα έτη 2002-2013.....65

Διάγραμμα 14 Εσωτερική δομή χρονοσειράς.....66

Διάγραμμα 15 Πρόβλεψη για το έτος 2013 με την μέθοδο SARIMA.....68

Διάγραμμα 16 Απασχολούμενοι στην Ελλάδα μέχρι το έτος 2016- μέθοδος SARIMA.70

Διάγραμμα 17 Απεικόνιση χρονοσειράς για τα έτη 2012-2012.....71

Διάγραμμα 18 Πρόβλεψη για το έτος 2013 με την μέθοδο NN.....72

Διάγραμμα 19 Απασχολούμενοι στην Ελλάδα μέχρι το 2016 με την μέθοδο NN.....73

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

Πίνακας 1 Παράδειγμα συναλλαγών καλαθιού αγοράς 11

Πίνακας 2 Χωρισμός δεδομένων σε ομάδες με τον αλγόριθμο k-μέσων 55

Πίνακας 3 Σύγκριση της Ευκλείδειας απόστασης με την City-block 59

Πίνακας 4 Σύγκριση μεθόδων ιεραρχικής συσταδοποίησης 64

Πίνακας 5 Προβλεπόμενη τιμή απασχολούμενων με την μέθοδο SARIMA 68

Πίνακας 6 Προβλεπόμενη τιμή και Διαστήματα Εμπιστοσύνης πρόβλεψης 68

Πίνακας 7 Απασχολούμενοι στην Ελλάδα για τα έτη 2014-2016 - μέθοδος SARIMA .. 70

Πίνακας 8 Πρόβλεψη για το έτος 2013 με την μέθοδο NN 72

Πίνακας 9 Απασχολούμενοι στην Ελλάδα για τα έτη 2014-2016 με την μέθοδο NN 74

Πίνακας 10 Σύγκριση πρόβλεψης για το έτος 2013 74

Πίνακας 11 Έλεγχος πρόβλεψης με NN μέχρι το έτος 2016 75

ΕΙΣΑΓΩΓΗ

Τις τελευταίες δεκαετίες έχει αναπτυχθεί ραγδαία η ανάγκη για εξόρυξη δεδομένων, για το λόγο ότι ο όγκος των πληροφοριών αυξάνεται συνεχώς, όπως και η ανάγκη για εξερεύνηση μεγάλων βάσεων δεδομένων, για την εύρεση καινούργιων προτύπων, αλλά και για την κατανόηση τους. Πέρα από αυτό, οι τεχνικές εξόρυξης δεδομένων χρησιμοποιούνται για να προβλέψουν μία μελλοντική παρατήρηση, αρκεί να γνωρίζουν τις προηγούμενες τιμές των παρατηρήσεων. Σαφώς η πρόβλεψη είναι πολύ αποδοτική και μπορεί να βοηθήσει σε πολλούς τομείς. Ένας τομέας που τα τελευταία χρόνια χρησιμοποιεί πολύ συχνά την εξόρυξη δεδομένων είναι οι χρονικές σειρές, δηλαδή παρατηρήσεις που συλλέγονται ανά τακτά χρονικά διαστήματα. Οι τεχνικές εξόρυξης των χρονοσειρών χρησιμοποιούν αλγορίθμους συσταδοποίησης, κατηγοριοποίησης, πρόβλεψης, τμηματοποίησης, σύνοψης, ανίχνευσης ανωμαλιών και ευρετηριοποίησης. Η παρούσα πτυχιακή εργασία αναφέρεται σε θεωρητικό υπόβαθρο αλλά και πρακτικό, καθώς και στους αλγόριθμους που χρησιμοποιούν οι χρονοσειρές για εξόρυξη δεδομένων. Αρχικά θα οριστούν και εξηγηθούν έννοιες, ορισμοί, η ανάλυση και μεθοδολογίες τις εξόρυξης δεδομένων. Στην συνέχεια θα αναφερθούμε στις χρονοσειρές, περιγράφοντας τους τύπους και τους στόχους της χρησιμοποίησης των χρονοσειρών. Επίσης θα αναλυθούν αναλυτικά οι αλγόριθμοι που χρησιμοποιούνται στην εξόρυξη δεδομένων των χρονοσειρών. Τέλος θα αναφέρουμε για ποιους λόγους πρέπει να χρησιμοποιείται το προγραμματιστικό περιβάλλον R και έπειτα θα υλοποιηθεί πρόγραμμα εφαρμογής της συσταδοποίησης και της πρόβλεψης με δύο διαφορετικούς τρόπους ώστε να βρεθεί ο βέλτιστος. Αφού υλοποιηθούν τα παραπάνω παραδείγματα εμφανίζονται τα συμπεράσματα από τα πειράματα, όπως επίσης στο παράρτημα εμφανίζονται οι κώδικες που χρησιμοποιήθηκαν με σχόλια για τυχόν περαιτέρω κατανόηση.

ΚΕΦΑΛΑΙΟ 1ο – ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

1.1. Ορισμός εξόρυξης δεδομένων

Ως Εξόρυξη Δεδομένων (Data Mining) θεωρείται η ανακάλυψη γνώσης από μια βάση δεδομένων. Οι τεχνικές εξόρυξης δεδομένων δεν εφαρμόζονται μόνο για να ερευνήσουν τις μεγάλες βάσεις δεδομένων, με σκοπό να βρεθούν νέα και χρήσιμα πρότυπα αλλά και παρέχουν την δυνατότητα πρόβλεψης του αποτελέσματος μιας μελλοντικής παρατήρησης, όπως για παράδειγμα η πρόβλεψη των δρομολογίων μιας αεροπορικής εταιρείας και του καταμερισμού του προσωπικού της που στηρίζεται σε προβλέψεις της ζήτησης θέσεων σε συγκεκριμένες πτήσεις. Η εξόρυξη δεδομένων βοηθάει στην ανάλυση, στην ταξινόμηση και στην κατανόηση μεγάλου όγκου πληροφοριών, πράγμα που είναι εξαιρετικά χρήσιμο ώστε να απαντηθεί κάποιο ερώτημα που τέθηκε ή για να γίνει κατάληξη σε κάποιο συμπέρασμα (1) (2).

1.2. Το περιεχόμενο της εξόρυξης δεδομένων

Στην εξόρυξη δεδομένων εντάσσονται οι σχεσιακές βάσεις δεδομένων, αποθήκες δεδομένων, συναλλακτικές βάσεις δεδομένων και προηγμένα συστήματα δεδομένων και πληροφοριών.

Σχεσιακές βάσεις δεδομένων (Relational Databases): είναι μία συλλογή από πίνακες όπου ο κάθε πίνακας αποτελείται από ένα σύνολο χαρακτηριστικών, δηλαδή πεδία, και αποθηκεύει ένα μεγάλο σύνολο εγγραφών. Κάθε εγγραφή σε έναν σχεσιακό πίνακα αντιπροσωπεύει ένα αντικείμενο που έχει εντοπιστεί από ένα μοναδικό κλειδί και περιγράφεται από ένα σύνολο τιμών του χαρακτηριστικού.

Αποθήκες δεδομένων (Data warehouses): μια αποθήκη δεδομένων αποτελείται από πληροφορίες που συλλέγονται από πολλαπλές πηγές. Οι αποθήκες δεδομένων κατασκευάζονται μέσω της διαδικασίας ενοποίησης, μετασχηματισμού και φόρτωση δεδομένων.

Βάσεις δεδομένων συναλλαγών (Transactional Databases): μια βάση δεδομένων συναλλαγής αποτελείται από ένα αρχείο όπου η κάθε εγγραφή αντιπροσωπεύει μία

συναλλαγή. Μια συναλλαγή περιλαμβάνει έναν μοναδικό αριθμό συναλλαγής και μία λίστα με τα στοιχεία που συνθέτουν την συναλλαγή.

Προηγμένα συστήματα δεδομένων και πληροφοριών (Advanced data and information systems and advanced applications): τα προηγμένα συστήματα δεδομένων περιλαμβάνουν νέες εφαρμογές για τον χειρισμό χωρικών δεδομένων, στοιχεία εφαρμοζομένου μηχανικού σχεδίου, υπερκείμενα και πολυμέσα δεδομένων, στοιχεία σχετικά με τον χρόνο, ροή δεδομένων και τον παγκόσμιο ιστό. Οι εφαρμογές αυτές απαιτούν αξιόπιστες δομές δεδομένων και μεθόδων για τον χειρισμό σύνθετων αντικειμένων (1).

1.3. Τύποι δεδομένων

Ως ένα σύνολο δεδομένων μπορεί να θεωρηθεί μια συλλογή αντικειμένων. Τα αντικείμενα αυτά περιγράφονται από ένα πλήθος χαρακτηριστικών. Για τον διαχωρισμό των χαρακτηριστικών χρησιμοποιούνται τρείς τρόποι. Ο πρώτος είναι με βάση των τύπων των τιμών που δέχεται, ο δεύτερος με βάση τις ιδιότητες που έχουν και τέλος με βάση το πλήθος των τιμών. Σύμφωνα με αυτά, οι τύποι δεδομένων χωρίζονται σε δυο μεγάλες κατηγορίες:

Κατηγορικά (Categorical) ή Ποιοτικά (Qualitative) χαρακτηριστικά: είναι εκείνα που έχουν ένα πεπερασμένο αριθμό διακριτών τιμών, συνήθως μικρότερο του εκατό. Τα ποιοτικά χαρακτηριστικά στερούνται τις περισσότερες από τις ιδιότητες των αριθμών, χρησιμοποιούν μόνο ευκρίνεια $(=, \neq)$ και διάταξη $(<, \leq, >, \geq)$. Ακόμη και αν αναπαρίστανται από αριθμούς (ακέραιους) πρέπει να αντιμετωπίζονται ως σύμβολα. Παράδειγμα κατηγορικών μεταβλητών είναι το «φύλο», το «βάρος», το «ύψος». Τα κατηγορικά χαρακτηριστικά χωρίζονται σε **ονομαστικά (nominal)**, όπου οι τιμές μιας μεταβλητής είναι απλώς ονόματα και δεν υποδηλώνουν τίποτα, για παράδειγμα «ταχυδρομικοί κώδικες», «χρώμα μαλλιών» και οι πληροφορίες που παρέχουν είναι αρκετές μόνο για το διαχωρισμό μεταξύ των αντικειμένων $(=, \neq)$, ενώ τα **τακτικά (ordinal)** παρέχουν πληροφορίες για την ταξινόμηση των αντικειμένων $(<, >)$, για παράδειγμα «βαθμοί», «αριθμοί οδών» ή σκληρότητα ορυκτών $\{καλή, καλύτερη, κάλλιστη\}$, όπου μπορεί να αναπαρασταθεί επίσης σωστά με τις τιμές $\{1, 2, 3\}$. Επίσης μία ειδική περίπτωση των κατηγορικών χαρακτηριστικών είναι

τα **δυαδικά χαρακτηριστικά (binary),** τα οποία λαμβάνουν μόνο δύο τιμές όπως 0 ή 1, «σωστό» ή «λάθος».

Αριθμητικά (Numerical) ή Ποσοτικά (Quantitative) χαρακτηριστικά: είναι εκείνα που έχουν πεπερασμένο ή άπειρο πλήθος αριθμητικών τιμών και χρησιμοποιούν τις πιο πολλές από τις ιδιότητες των αριθμών όπως η πρόσθεση $(+,-)$ και ο πολλαπλασιασμός $(*,/)$, πράγμα που στα κατηγορικά χαρακτηριστικά είναι αδύνατον. Τα αριθμητικά χαρακτηριστικά χωρίζονται σε **διαστημάτων (interval),** όπου οι διαφορές μεταξύ των τιμών έχουν σημασία, για παράδειγμα «ημερομηνίες ημερολογίων» και σε **αναλογιών (ratio)**, όπου τόσο οι διαφορές όσο και οι αναλογίες έχουν σημασία, για παράδειγμα «νομισματικές ποσότητες», «ηλικία», «ηλεκτρικό ρεύμα» (2).

1.4. Εξόρυξη Δεδομένων και Ανακάλυψη της Γνώσης

Η εξόρυξη δεδομένων αποτελεί ένα βήμα της Ανακάλυψης Γνώσης από τις Βάσεις Δεδομένων (Knowledge Discovery in Databases – KDD) η οποία αποτελεί τη συνολική διεργασία της μετατροπής ακατέργαστων δεδομένων σε σημαντικές πληροφορίες. Η KDD είναι μη-τετριμένη διαδικασία ανακάλυψης έγκυρων, νέων και ενδεχομένως χρήσιμων και τελικά κατανοητών προτύπων στα δεδομένα (3).

Η διαδικασία ανακάλυψης της γνώσης είναι επαναληπτική, και αποτελείται από εννέα βήματα. Σε κάθε βήμα η διαδικασία επαναλαμβάνεται, επομένως πρέπει να γίνεται κατανόηση σε βάθος της διαδικασίας, όπως επίσης και των διαφορετικών αναγκών και δυνατοτήτων σε κάθε βήμα.

Ένα σύστημα ανακάλυψης γνώσης περιλαμβάνει:

1. Ανάπτυξη της κατανόησης στον τομέα εφαρμογής (Developing an understanding of the application domain)

   Η κατανόηση του πεδίου εφαρμογής είναι το προπαρασκευαστικό στάδιο, δηλαδή η προετοιμασία του εδάφους ώστε να γίνει καλύτερη κατανόηση για τις πολλές αποφάσεις που πρέπει να ληφθούν (μετατροπή, επιλογή αλγορίθμου). Οι άνθρωποι που είναι υπεύθυνοι για ένα έργο KDD πρέπει να κατανοήσουν και

να καθορίσουν τους στόχους του τελικού χρήστη και το περιβάλλον στο οποίο η διαδικασία ανακάλυψης της γνώσης θα λάβει μέρος.

2. Επιλογή και δημιουργία ενός συνόλου δεδομένων στην οποία θα πραγματοποιηθεί η ανακάλυψη (Selecting and creating a data set on which discovery will be performed)

   Εφόσον έχουν οριστεί οι στόχοι του τελικού χρήστη, τα δεδομένα που θα χρησιμοποιηθούν για την ανακάλυψη της γνώσης θα πρέπει να έχουν προσδιοριστεί. Τα διαθέσιμα στοιχεία και η απόκτηση πρόσθετων απαραίτητων στοιχείων θα ενσωματωθούν σε ένα σύνολο δεδομένων, έτσι ώστε να υπάρχουν όλα τα στοιχεία για την επιτυχή ανακάλυψη της γνώσης. Αν λείπουν κάποιες σημαντικές ιδιότητες, τότε η μελέτη πιθανόν θα αποτύχει.

3. Προεπεξεργασία και καθορισμός (Preprocessing and cleansing)

   Σε αυτό το στάδιο εντοπίζονται και διαχειρίζονται οι ελλιπείς τιμές, απομακρύνεται ο θόρυβος και οι ακραίες τιμές, αποσκοπώντας στην όσο το δυνατό μεγαλύτερη αξιοπιστία των δεδομένων. Για την προεπεξεργασία των δεδομένων χρησιμοποιούνται είτε στατιστικές μέθοδοι είτε αλγόριθμοι από το πεδίο της εξόρυξης δεδομένων.

4. Μετασχηματισμός δεδομένων (Data transformation)

   Τα καλύτερα στοιχεία για την εξόρυξη δεδομένων είναι προετοιμασμένα και έχουν ήδη αναπτυχθεί. Αυτό το βήμα είναι πολύ σημαντικό για την επιτυχία όλου του έργου ανακάλυψης της γνώσης (Knowledge Discovery Process), διότι δίνει εκπληκτικά αποτελέσματα και γίνεται κατανόηση του μετασχηματισμού.

5. Επιλογή της κατάλληλης εργασίας εξόρυξης δεδομένων (Choosing the appropriate Data Mining task)

   Στο στάδιο αυτό γίνεται η επιλογή του τύπου της εξόρυξης δεδομένων, που εξαρτάται από τους στόχους του KDD και τα προηγούμενα βήματα. Οι δύο σημαντικοί στόχοι της εξόρυξης δεδομένων είναι η Πρόβλεψη και η Περιγραφή των δεδομένων

6. Επιλογή του αλγορίθμου εξόρυξης δεδομένων (Choosing the Data Mining algorithm)

   Έχοντας την στρατηγική, τώρα γίνεται η επιλογή της τακτικής. Σε αυτό το στάδιο περιλαμβάνεται η επιλογή της συγκεκριμένης μεθόδου που θα χρησιμοποιηθεί για την αναζήτηση προτύπων. Η Μετα-Μάθηση (Meta-

Learning) επικεντρώνεται στην εξήγηση του αλγορίθμου της εξόρυξης δεδομένων όταν είναι επιτυχής ή ανεπιτυχής.

7. Εφαρμογή του αλγορίθμου εξόρυξης δεδομένων (Employing the Data Mining algorithm)

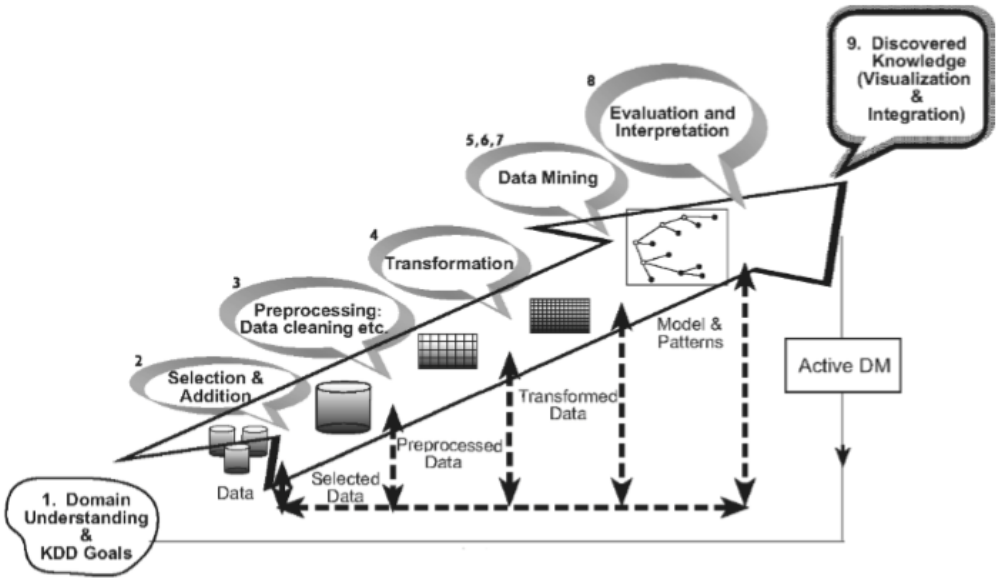
   Γίνεται η εφαρμογή του αλγορίθμου της εξόρυξης δεδομένων που επιλέχτηκε προηγούμενος. Σε αυτό το βήμα μπορεί να χρειαστεί να χρησιμοποιηθεί ο αλγόριθμος αρκετές φορές, έως ότου να ληφθεί ένα ικανοποιημένο αποτέλεσμα.

8. Αξιολόγηση και ερμηνεία των αποτελεσμάτων της ανάλυσης (Evaluation)

   Σε αυτό το στάδιο αξιολογείται και ερμηνεύεται η εξόρυξη προτύπων (κανόνες, αξιοπιστία), σε σχέση με τους στόχους που έχουν οριστεί από το πρώτο βήμα. Το βήμα προεπεξεργασίας αξιολογείται σε σχέση με την επίδρασή του στα αποτελέσματα του αλγορίθμου της εξόρυξης δεδομένων. Η σαφήνεια και η χρησιμότητα του μοντέλου επικεντρώνεται στο συγκεκριμένο βήμα, όπως επίσης και η ανακάλυψη γνώσης τεκμηριώνεται για περαιτέρω χρήση.

9. Χρησιμοποίηση των γνώσεων που ανακαλύφθηκαν (Using the discovered knowledge)

   Η γνώση είναι πλέον έτοιμη να ενσωματωθεί σε άλλο σύστημα για περαιτέρω δράση και είναι ενεργοποιημένη, με την έννοια ότι μπορούν να γίνουν αλλαγές στο σύστημα, όπως επίσης και η μέτρηση των αποτελεσμάτων. Η επιτυχία αυτού του βήματος καθορίζει την αποτελεσματικότητα όλης της διαδικασίας KDD (3).



Εικόνα 1 Τα βήματα της διαδικασίας ανεύρεσης γνώσης σε βάσεις δεδομένων

1.5. Μεθοδολογίες Εξόρυξης Δεδομένων

1.5.1. Προγνωστικού Τύπου

Ο στόχος του προγνωστικού **τύπου (predictive type)** είναι η δυνατότητα πρόγνωσης μιας τιμής ενός χαρακτηριστικού μέσα από άλλες τιμές χαρακτηριστικών οι οποίες είναι γνωστές. Η τιμή του προβλεπόμενου χαρακτηριστικού είναι γνωστή ως **εξαρτημένη μεταβλητή (dependent variable)**, ενώ οι τιμές των χαρακτηριστικών που χρησιμοποιούνται για την πρόβλεψη είναι γνωστές ως **ανεξάρτητες μεταβλητές (independent variable)**. Τα βασικά μοντέλα που χρησιμοποιούνται στον προγνωστικό τύπο είναι η κατηγοριοποίηση-ταξινόμηση, η παλινδρόμηση και η πρόβλεψη (2).

Ταξινόμηση-Κατηγοριοποίηση (Classification)

Η κατηγοριοποίηση είναι η διαδικασία εκμάθησης μιας **συνάρτησης - στόχου (target function)** f, η οποία απεικονίζει κάθε σύνολο χαρακτηριστικών x σε μια από τις προκαθορισμένες ετικέτες κατηγορίας y. Η συνάρτηση-στόχος, είναι γνωστή και ως **μοντέλο κατηγοριοποίησης (descriptive model)**.

Ένα μοντέλο κατηγοριοποίησης χρησιμοποιείται ως ένα επεξηγηματικό εργαλείο για την ταξινόμηση των αντικειμένων σε διαφορετικές κατηγορίες και ονομάζεται **περιγραφική μοντελοποίηση** . Επίσης εφαρμόζεται στην πρόβλεψη της ετικέτας της κατηγορίας μη γνωστών εγγραφών και καλείται **προβλεπτική μοντελοποίηση** (2).

Για την κατηγοριοποίηση - ταξινόμηση των δεδομένων χρησιμοποιούνται τα νευρωνικά δίκτυα, οι απλοϊκοί κατηγοριοποιητές Bayes, οι βασισμένοι σε κανόνες κατηγοριοποιητές, οι μηχανές διανυσμάτων υποστήριξης και δέντρα απόφασης, τα οποία ταξινομούν ένα μεγάλο πλήθος δεδομένων σε ομάδες, σύμφωνα με κάποιο στόχο. Η ταξινόμηση αυτή γίνεται σύμφωνα με κάποιους κανόνες. Τα δέντρα απόφασης είναι εύκολα κατανοητά, διότι γίνεται γραφική απεικόνιση των κανόνων, όμως εάν πρόκειται για αριθμητικά δεδομένα, το δέντρο μπορεί να γίνει πάρα πολύ μεγάλο και έτσι θα δυσκολευτεί η κατανόηση του (2) (3).

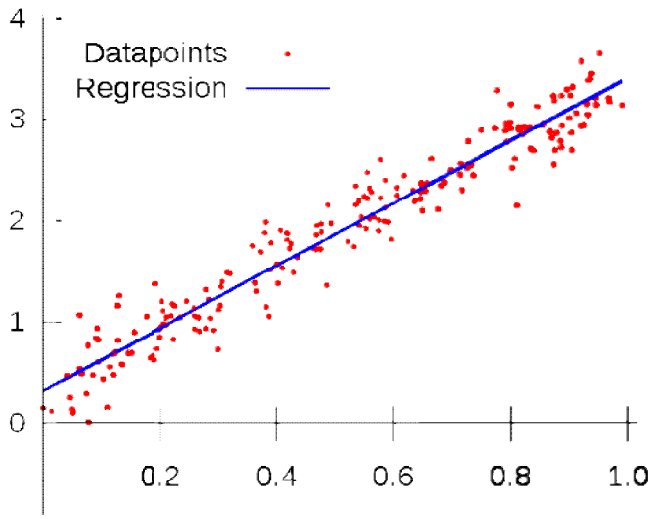
Η ταξινόμηση χρησιμοποιείται πολύ στην πρόβλεψη των χρονικών σειρών, με στόχο την δημιουργία ενός μοντέλου που βασίζεται σε δεδομένα που έχουν ήδη

γνωστά χαρακτηριστικά-τιμές, και στη συνέχεια χρησιμοποιείται το μοντέλο αυτό ώστε να γίνει η πρόβλεψη και στα δεδομένα που είναι άγνωστα. Τα νέα χαρακτηριστικά γνωρίσματα που προκύπτουν από τις χρονικές σειρές μπορεί να βελτιώσουν την απόδοση της ταξινόμησης των μοντέλων (4).

Παλινδρόμηση (Regression)

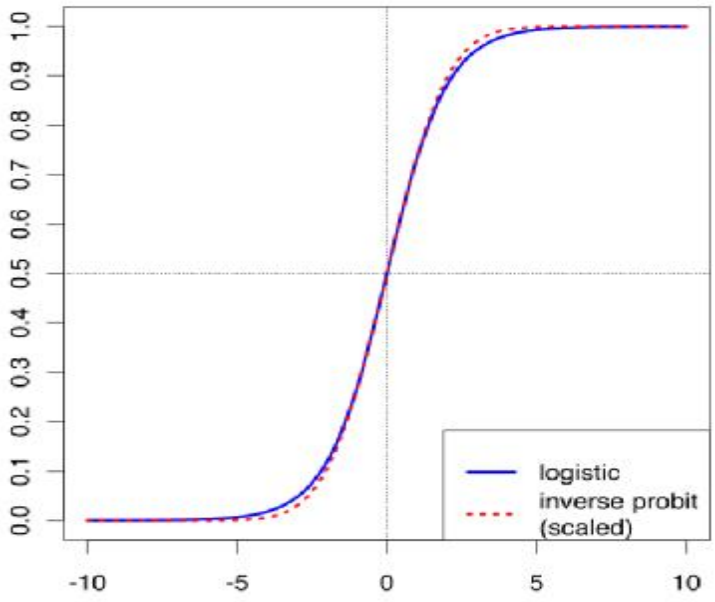
Η παλινδρόμηση είναι η δημιουργία μιας συνάρτησης των **ανεξάρτητων μεταβλητών (independent variables)**, οι οποίοι είναι γνωστοί επίσης ως προγνωστικοί δείκτες, για να γίνει η πρόβλεψη μιας **εξαρτημένης μεταβλητής (dependent variable)**, που ονομάζεται επίσης απάντηση (response). Για παράδειγμα, οι τράπεζες αξιολογούν τον κίνδυνο των δανείων με βάση την ηλικία, το εισόδημα, το επάγγελμα (2).

Οι μέθοδοι της παλινδρόμησης είναι γραμμική, μη-γραμμική και η λογιστική παλινδρόμηση. Επίσης συχνά χρησιμοποιούνται τα δέντρα παλινδρόμησης και τα νευρωνικά δίκτυα. Η **γραμμική παλινδρόμηση (linear regression)** μελετά την σχέση που υπάρχει ανάμεσα στην εξαρτημένη μεταβλητή, δηλαδή η τιμή του χαρακτηριστικού που γίνεται η πρόβλεψη, και των ανεξάρτητων μεταβλητών, δηλαδή τα χαρακτηριστικά που χρησιμοποιούνται για να γίνει η πρόβλεψη. Οι ανεξάρτητες μεταβλητές είναι οι παράγοντες που επηρεάζουν την εξαρτημένη μεταβλητή. Επειδή μελετά την σχέση μεταξύ ακριβώς δύο τιμών, η γραμμή παλινδρόμησης περνάει από το μέσο του νέφους των τιμών, όπως φαίνεται στην Εικόνα 2.



Εικόνα 2 Γραμμική παλινδρόμηση

Στην **μη-γραμμική παλινδρόμηση (non-linear regression)**, σε αντίθεση με την γραμμική παλινδρόμηση, η γραμμή παλινδρόμησης που περνάει από το μέσο του νέφους των τιμών δεν είναι ευθεία, όπως φαίνεται στην Εικόνα 3.



Εικόνα 3 Μη-γραμμική παλινδρόμηση

Στην κατηγορία αυτή ανήκει η **λογιστική παλινδρόμηση (logistic regression)** και χρησιμοποιείται για να προβλέψει την πιθανότητα εμφάνισης ενός γεγονότος, καθώς επίσης σε περιπτώσεις στις οποίες γίνεται η πρόβλεψη της απουσίας ή παρουσίας ενός χαρακτηριστικού με την τοποθέτηση των δεδομένων σε μια λογιστική καμπύλη, όπως στην Εικόνα 3.. Η λογιστική παλινδρόμηση χρησιμοποιεί αριθμητικά και κατηγορικά δεδομένα, σε αντίθεση με την γραμμική παλινδρόμηση η οποία χρησιμοποιεί μόνο αριθμητικές τιμές (5).

Πρόβλεψη (Forecasting)

Η πρόβλεψη χρονοσειρών είναι η πρόβλεψη μελλοντικών γεγονότων που βασίζονται σε ιστορικά δεδομένα. Ένα παράδειγμα είναι η πρόβλεψη της τιμής μιας μετοχής με βάση τις προηγούμενες τιμές της (4).

Οι μέθοδοι πρόβλεψης μπορούν να ταξινομηθούν σε τρείς τύπους, την **κριτική πρόβλεψη (judgemental forecasts)**, που βασίζεται σε υποκειμενική κρίση, διαίσθηση, καθώς και σε κάθε άλλη σχετική πληροφορία, την **μονομεταβλητή μέθοδος (univariate methods)**, όπου οι προβλέψεις εξαρτώνται μόνο από την υπάρχουσα και την προηγούμενη τιμή της σειράς που προβλέπονται, ενδεχομένως

αυξημένες από την συνάρτηση του χρόνου, όπως μια γραμμική τάση, και τέλος η **πολυμετάβλητη μέθοδος (multivariate methods)**, όπου οι προβλέψεις μιας συγκεκριμένης μεταβλητής εξαρτώνται από την τιμή ενός ή περισσοτέρων πρόσθετων μεταβλητών χρονοσειρών, που ονομάζονται προγνωστικοί δείκτες ή επεξηγηματικές μεταβλητές. Η πολυμετάβλητη πρόβλεψη μπορεί να εξαρτάται από ένα πολυμετάβλητο μοντέλο που αφορά περισσότερες από μια εξίσωση, εφόσον οι μεταβλητές είναι εξαρτημένες.

Η μέθοδος πρόβλεψης μπορεί να συνδυάσει περισσότερους από έναν από τους παραπάνω τύπους (6).

1.5.2. Περιγραφικού Τύπου

Ο στόχος του **περιγραφικού τύπου (descriptive type)** είναι η δυνατότητα εξαγωγής πρότυπων όπως οι τάσεις, οι συσχετίσεις και οι ανωμαλίες που περιγράφουν τις βασικές σχέσεις που υπάρχουν στα δεδομένα. Τα μοντέλα που χρησιμοποιούνται στον περιγραφικό τύπο είναι η συσταδοποίηση και ο συσχετισμός (2).

Συσταδοποίηση (Clustering)

Η συσταδοποίηση χωρίζει τα δεδομένα σε ομάδες, με βάση τις πληροφορίες και τις σχέσεις μεταξύ των δεδομένων που ομαδοποιεί. Ο στόχος της συσταδοποίησης είναι η κατάταξη των δεδομένων με τέτοιο τρόπο ώστε τα δεδομένα μιας ομάδας να είναι όμοια ή συσχετιζόμενα μεταξύ τους και διαφορετικά ή μη συσχετιζόμενα με τα δεδομένα των άλλων ομάδων. Όσο πιο μεγάλη είναι η ομοιότητα των δεδομένων σε μία ομάδα και όσο πιο μεγάλη είναι η διαφορά ανάμεσα στις ομάδες, τόσο πιο καλή είναι η συσταδοποίηση.

Η συσταδοποίηση χρησιμοποιείται σε πολλές εφαρμογές όπως η ψυχολογία, η βιολογία, η στατιστική, η εξόρυξη δεδομένων και μπορεί να χωριστεί σε δύο κατηγορίες, ανάλογα με τον σκοπό για τον οποίο γίνεται. Η κατάταξη των διάφορων αντικειμένων, με βάση τα κοινά τους χαρακτηριστικά, σε ομάδες ή κατηγορίες είναι πολύ σημαντική για τους ανθρώπους, διότι αυτός είναι ένας τρόπος με τον οποίο περιγράφουν τον κόσμο. Στο στάδιο αυτό ο σκοπός της συσταδοποίησης είναι η κατανόηση των δεδομένων, δηλαδή να γίνει η **Συσταδοποίηση για Κατανόηση** και

εφαρμόζεται στην βιολογία, όπου έχει χρησιμοποιηθεί ώστε να βρεθούν ομάδες γονιδίων με παρόμοιες λειτουργίες, στις επιχειρήσεις, όπου μπορεί να χρησιμοποιηθεί στην ομαδοποίηση των πελατών, στην ανάκτηση πληροφοριών, στο κλίμα, στην αναγνώριση προτύπων. Κάποιες τεχνικές συσταδοποίησης χαρακτηρίζουν κάθε συστάδα σε σχέση με ένα πρότυπο συστάδας, δηλαδή ένα αντικείμενο, το οποίο είναι αντιπροσωπευτικό των άλλων αντικειμένων στην συστάδα. Αυτά τα πρότυπα συστάδων μπορούν να χρησιμοποιηθούν ως βάση για ένα πλήθος αναλύσεων δεδομένων. Επειδή στο στάδιο αυτό ο σκοπός της συσταδοποίησης είναι η μελέτη των τεχνικών εύρεσης των πιο αντιπροσωπευτικών προτύπων συστάδας, το στάδιο αυτό ονομάζεται **Συσταδοποίηση για Χρησιμότητα** και εφαρμόζεται στην περίληψη, στην συμπίεση και στην αποτελεσματική εύρεση των πλησιέστερων γειτόνων (2).

Συσχετισμός (Association)

Η ανάλυση συσχέτισης (association analysis) χρησιμοποιείται για την ανακάλυψη σχέσεων που είναι κρυμμένες σε μεγάλες βάσεις δεδομένων. Οι σχέσεις που αποκαλύπτονται εκφράζονται με τη μορφή των **κανόνων συσχέτισης (association rules)**. Για παράδειγμα ο κανόνας $\{Πάνες\ μωρού \rightarrow Μπύρα\}$ μπορεί να προκύψει από τον Πίνακας 1:

Κωδικός Συναλλαγής (TID)	Αντικείμενα
1	{Ψωμί, Γάλα}
2	{Ψωμί, Πάνες μωρού, Μπύρα, Αυγά}
3	{Γάλα, Πάνες μωρού, Μπύρα, Cola}
4	{Ψωμί, Γάλα, Πάνες μωρού, Μπύρα}
5	{Ψωμί, Γάλα, Πάνες μωρού, Cola}

Πίνακας 1 Παράδειγμα συναλλαγών καλαθιού αγοράς

Ο κανόνας δείχνει την ισχυρή σχέση ανάμεσα στην πώληση πάνων μωρού και της μπύρας, επειδή πολλοί πελάτες που αγοράζουν πάνες μωρού αγοράζουν και μπύρα. Οι έμποροι μπορούν να χρησιμοποιήσουν τέτοιους κανόνες για να ανακαλύψουν την αγοραστική συμπεριφορά, τον τρόπο με τον οποίο συνίσταται η τοποθέτηση των

προϊόντων στα ράφια, την καλύτερη διαχείριση των αποθεμάτων της αποθήκης και για την διαχείριση πελατειακών σχέσεων.

Ένας κανόνας συσχέτισης είναι μια πρόταση συνεπαγωγής της μορφής $X \rightarrow Y$, όπου τα X και Y είναι ξένα μεταξύ τους στοιχειοσύνολα, δηλαδή $X \cap Y = \emptyset$. Η ισχύς του κανόνα συσχέτισης μετριέται με βάση την **υποστήριξη (support)**, η οποία καθορίζει την συχνότητα με την οποία εφαρμόζεται ο κανόνας σε ένα σύνολο δεδομένων, και την **εμπιστοσύνη (confidence)**, η οποία καθορίζει πόσο συχνά τα αντικείμενα στο στοιχειοσύνολο Y εμφανίζονται σε συναλλαγές που περιέχουν το X. Οι ορισμοί των μέτρων αυτών είναι: (2)

$$Support, s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N},$$

$$Confidence, c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

1.6. Γιατί είναι σημαντική η εξόρυξη δεδομένων

Η εξόρυξη δεδομένων έχει προκαλέσει μεγάλη προσοχή στον τομέα των πληροφοριών, λόγω του ότι μπορεί να διαχειριστεί δεδομένα τεράστιων όγκων, μετατρέποντας αυτά τα δεδομένα σε χρήσιμες πληροφορίες και γνώσεις. Οι πληροφορίες και η γνώση που αποκτάται από την εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί για εφαρμογές που άπτονται της ανάλυση της αγοράς όπως τη διατήρηση των πελατών, την ανίχνευση απάτης, τον έλεγχο της παραγωγής ακόμα και στην επιστήμη και στην ιατρική για νέες ανακαλύψεις.

Η εξέλιξη της εξόρυξης δεδομένων από την αρχική της μορφή μέχρι τη σημερινή απέχει αρκετά, καθώς το 1960 από την απλή επεξεργασία αρχείων εξελίχθηκε σε δημιουργία ισχυρών βάσεων δεδομένων, πράγμα που στην συνέχεια οδήγησε σε δικτυωτό και ιεραρχικό σύστημα βάσης δεδομένων. Η ερευνά και ανάπτυξη των παραπάνω δημιούργησε το 1970 τις σχεσιακές βάσεις δεδομένων. Εν συνεχεία το 1980 δημιουργήθηκαν προηγμένα μοντέλα συστημάτων διαχείρισης βάσεων δεδομένων όπως extended-relational, object-oriented και object-relational και συστήματα για συγκεκριμένες εφαρμογές (π.χ., χωρικές, χωροχρονικές, πολυμέσα, επιστημονικές και μηχανικές βάσεις δεδομένων). Ακόμα τα ετερογενή συστήματα

βάσεων δεδομένων στο Διαδίκτυο με παγκόσμια συστήματα πληροφοριών, όπως ο παγκόσμιος ιστός, έχουν επίσης προκύψει και διαδραματίζουν σημαντικό ρόλο στη βιομηχανία πληροφοριών.

Σήμερα, λοιπόν, τα δεδομένα αποθηκεύονται σε πολλά διαφορετικά είδη βάσεων δεδομένων. Ωστόσο, η ταχέως αναπτυσσόμενη, τεραστία ποσότητα δεδομένων, που συλλέγονται και αποθηκεύονται σε μεγάλα σύνολα βάσεων δεδομένων, έχει ξεπεράσει κατά πολύ την ανθρώπινη ικανότητα για την επεξεργασία και κατανόηση, επομένως απαιτούνται ισχυρά εργαλεία ανάλυσης. Συνεπώς, τα εργαλεία εξόρυξης δεδομένων πραγματοποιούν την ανάλυση των δεδομένων και μπορεί να ανακαλύψουν σημαντικά πρότυπα δεδομένων, συμβάλλοντας σε μεγάλο βαθμό στις επιχειρηματικές στρατηγικές, επιστημονικής και ιατρικής έρευνας (1).

ΚΕΦΑΛΑΙΟ 2ο – ΑΝΑΛΥΣΗ ΧΡΟΝΟΣΕΙΡΩΝ ΜΕ ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

2.1. Τι είναι οι χρονικές σειρές και που εφαρμόζονται

Χρονοσειρά είναι ένα σύνολο από παρατηρήσεις που συλλέγονται σε ορισμένες χρονικές στιγμές που ισαπέχουν μεταξύ τους. Οι τιμές αυτές μπορεί να είναι ημερήσιες, εβδομαδιαίες, ωριαίες και διακρίνονται σε στάσιμες και μη στάσιμες. Μία χρονοσειρά λέγεται στάσιμη όταν οι διακυμάνσεις των τιμών της δεν μεταβάλλονται με τον χρόνο, ενώ μία χρονοσειρά που δεν είναι στάσιμη μπορεί να έχει τάση (ανοδική, καθοδική ή και τα δύο), να εμφανίζει περιοδικότητα ή να παρουσιάζει εποχικότητα (7).

Τα δύο βασικά χαρακτηριστικά των χρονοσειρών που μελετώνται συνήθως είναι η **στασιμότητα (stagnation)**, όπου οι διακυμάνσεις των τιμών της χρονοσειράς δεν αλλάζουν με το χρόνο, που σημαίνει ότι η χρονοσειρά είναι στάσιμη, και η **τάση (trends)**, η οποία μελετάται μόνο σε μη στάσιμη χρονοσειρά όπου μπορεί να έχει τάσεις, όπως αλλαγές στη μέση τιμή της με το χρόνο, δηλαδή όταν μια χρονοσειρά παρουσιάζει σταθερά ανοδική ή καθοδική πορεία, για πολλές διαδοχικές χρονικές περιόδους. Ένα παράδειγμα είναι η συμπεριφορά του δείκτη τιμών λιανικής πώλησης του Ηνωμένου Βασιλείου που έχει δείξει μια αύξηση κάθε χρόνο εδώ και πολλά χρόνια. Επίσης η ανάλυση της τάσης εξαρτάται, εν μέρει, από το μήκος της παρατηρούμενης σειράς (6).

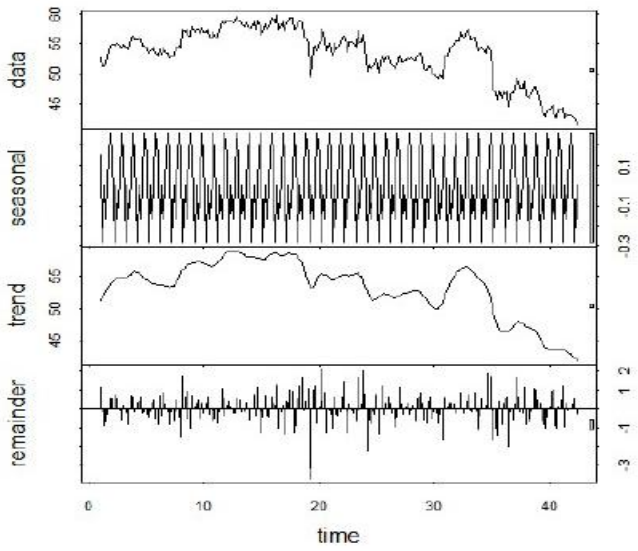
Άλλα χαρακτηριστικά που μπορούν να μελετηθούν σε μη στάσιμες χρονοσειρές είναι **περιοδικότητα (periodicity)**, όπου οι αυξομειώσεις των τιμών των χρονοσειρών αλλάζουν σε συγκεκριμένες περιόδους σε μία μη στάσιμη χρονοσειρά, και η **εποχικότητα (seasonality)**, στην οποία οι τιμές των χρονοσειρών μεταβάλλονται ανάλογα με την εποχή του χρόνου. Αν μια χρονοσειρά μετριέται μόνο σε ετήσια βάση (δηλαδή μία φορά ανά έτος), τότε δεν είναι δυνατόν να υπάρχει εποχιακή διακύμανση. Ένα παράδειγμα είναι το μοτίβο πωλήσεων για παγωτό η οποία είναι πάντα υψηλή το καλοκαίρι. Χρησιμοποιείται για πολλές σειρές, εάν μετρηθούν εβδομαδιαία, μηνιαία ή τριμηνιαία, όταν παρόμοια πρότυπα

συμπεριφοράς παρατηρούνται σε συγκεκριμένες περιόδους του έτους (6). **Κύκλοι (Cyclic)** είναι παρόμοιο χαρακτηριστικό με την εποχικότητα με διαφορά ότι οι αυξομειώσεις των τιμών των χαρακτηριστικών δεν είναι συστηματικές ούτε σε καθορισμένη περίοδο. Επίσης μελετάται η **αυτοσυσχέτιση (autocorrelation)**, δηλαδή η γραμμική αυτοσυσχέτιση (linear autocorrelation), όπου μπορεί να χρησιμοποιηθεί εύκολα όταν η χρονοσειρά είναι στάσιμη, διότι τότε παράγει καθαρά και σωστά αποτελέσματα, ενώ όταν η χρονοσειρά είναι μη στάσιμη συνήθως παρατηρείται ότι έχει πολύ υψηλές τιμές. Η συσχέτιση μεταξύ των διαδοχικών τιμών των ίδιων χρονοσειρών καλείται **αυτοσυσχέτιση**.

Άλλα χαρακτηριστικά που ερευνάται σε μη στάσιμες χρονοσειρές είναι η **αιτιοκρατία (determinism)** και η **στοχαστικότητα (stochasticity)**. Επειδή όλες οι χρονοσειρές που προέρχονται από πραγματικά μεγέθη περιέχουν θόρυβο, αυτό έχει ως συνέπεια όλες οι πραγματικές χρονοσειρές να είναι στοχαστικές. Για να γίνει αποτελεσματική ανάλυση των χρονοσειρών πρέπει να βρεθεί το αιτιοκρατικό (σταθερό) μέρος που παράγει τη χρονοσειρά και να αναλυθεί. Στην περίπτωση που το αιτιοκρατικό μέρος είναι κρυμμένο μέσα σε θόρυβο ή δεν αποτελεί μεγάλο μέρος της χρονοσειράς, τότε η χρονοσειρά θεωρείται στοχαστική και η ανάλυση της γίνεται με στατιστική περιγραφή. Τέλος, τα **κατάλοιπα (residuals)** υποδηλώνουν εάν οι διακυμάνσεις είναι τυχαίες ή συστηματικές (7).

Οι κλασσικές μέθοδοι λειτουργούν αρκετά καλά όταν η μεταβολή κυριαρχείται από μια κανονική γραμμική τάση ή / και τακτική εποχικότητα. Ωστόσο, δεν λειτουργούν πολύ καλά, όταν η τάση ή / και εποχικότητα αλλάζουν στο πέρασμα του χρόνου ή όταν οι διαδοχικές τιμές των ακανόνιστων διακυμάνσεων συσχετίζονται.

Η Εικόνα 4 δείχνει την εσωτερική δομή μιας χρονοσειράς, δηλαδή την αποσύνθεση της ως προς την εποχικότητα, την τάση και τα κατάλοιπα.



Εικόνα 4 Εσωτερική δομή μιας χρονοσειράς

Η χρήση και η εφαρμογή με ανάλυση χρονοσειρών έχουν βοηθήσει σε διαφόρους τομείς αφού χρησιμοποιούνται στη στατιστική, στη διαχείριση σημάτων, στην αναγνώριση προτύπων, στην οικονομετρία, στα οικονομικά μαθηματικά, στην πρόγνωση καιρού, στην πρόβλεψη του σεισμού, στα ηλεκτροεγκεφαλογραφήματα, στον μηχανικό έλεγχο, στη μηχανική επικοινωνιών κ.ά. (7).

2.2. Στόχοι της ανάλυσης χρονοσειρών

Χρησιμοποιώντας τα παραπάνω χαρακτηριστικά των χρονοσειρών μελετάται η συμπεριφορά ενός χαρακτηριστικού στη πάροδο του χρόνου με στόχο την εύρεση των προτύπων στα δεδομένα και την πρόβλεψη των μελλοντικών τιμών. Το ιδιαίτερο χαρακτηριστικό των στοιχείων χρονολογικών σειρών είναι ότι οι διαδοχικές τιμές δεν είναι συνήθως ανεξάρτητες και έτσι η ανάλυση θα πρέπει να λάβει υπόψη του τη σειρά με την οποία συλλέγονται οι παρατηρήσεις. Για την ακρίβεια, μελετάται που υπάρχει ομοιότητα μεταξύ των χρονοσειρών, εξετάζεται η δομή των χρονοσειρών και κατηγοριοποιείται η συμπεριφορά τους. Κύριοι στόχοι κατά την ανάλυση χρονοσειρών είναι (6):

Περιγραφή: περιγραφή των δεδομένων χρησιμοποιώντας συνοπτικά στατιστικά στοιχεία ή και γραφικές μεθόδους. Μία γραφική παράσταση χρόνου (time plot) των δεδομένων είναι αρκετά πολύτιμη.

Μοντελοποίηση: βρίσκει ένα κατάλληλο στατιστικό μοντέλο για να περιγράψει τη διαδικασία παραγωγής των τιμών της χρονοσειράς. Ένα στατιστικό μοντέλο μονομεταβλητής/μονοπαραγοντικής (univariate model) για μια συγκεκριμένη τιμή βασίζεται μόνο σε προηγούμενες τιμές της μεταβλητής, ενώ ένα πολυμεταβλητό/πολυπαραγοντικό στατιστικό μοντέλο (multivariate model) για μια συγκεκριμένη τιμή μπορεί να βασίζεται όχι μόνο στις προηγούμενες τιμές της, αλλά και για το παρόν και το παρελθόν, των άλλων μεταβλητών της χρονοσειράς.

Πρόβλεψη: όπου προσδοκούμε το μέλλον να είναι σαν το παρελθόν, και What-if πρόβλεψης όπου χρησιμοποιείται ένα πολυπαραγοντικό μοντέλο για να κατανοήσει το αποτέλεσμα (effect) στην αλλαγή κατανομής των τιμών.

Έλεγχος: οι καλές προβλέψεις επιτρέπουν στον αναλυτή να αναλάβει δράση, ώστε να ελέγχει την εν λόγω διαδικασία, εάν αυτό είναι μια βιομηχανική διεργασία, ή μια οικονομία ή οτιδήποτε άλλο.

2.3. Διάφοροι τύποι χρονοσειρών

Μια χρονοσειρά είναι μια σειρά από παρατηρήσεις που μετρήθηκαν διαδοχικά μέσα στο χρόνο. Οι μετρήσεις αυτές γίνονται συνεχώς μέσα στον χρόνο ή λαμβάνονται ως ένα διακριτό σύνολο χρονικών σημείων. Με αυτόν τον τρόπο οι χρονοσειρές χωρίζονται σε δύο τύπους, τις συνεχείς και τις διακριτές χρονοσειρές αντίστοιχα.

Για μια συνεχή χρονοσειρά, η παρατηρούμενη μεταβλητή είναι ουσιαστικά μια μεταβλητή που καταγράφεται συνεχώς. Η μέθοδος για την ανάλυση τέτοιας χρονοσειράς είναι η διαίρεση της σε ίσα χρονικά διαστήματα ώστε να προκύψει μια σειρά διακριτού τύπου. Μια χρονοσειρά διακριτού τύπου προκύπτει όταν λαμβάνονται δείγματα από μια συνεχή σειρά, όταν συγκεντρώνονται τα δείγματα κατά την διάρκεια μιας χρονικής περιόδου και τέλος όταν η σειρά είναι εκ φύσεως διακριτή. Όταν η χρονοσειρά είναι διακριτή, τα δεδομένα καταγράφονται σε ίσα χρονικά διαστήματα.

Τα δεδομένα συγκεντρώνονται είτε σε όλο το χρόνο είτε σε όλη την σειρά και έτσι προκύπτουν δύο διαφορετικές καταστάσεις. Όταν τα δεδομένα βρίσκονται σε όλο το

χρόνο τότε η κατάσταση αυτή ονομάζεται χρονική συνάθροιση, ενώ όταν αθροίζονται σε όλη την σειρά τότε ονομάζεται ταυτόχρονη συσσωμάτωση (6).

2.4. Εξόρυξη δεδομένων των χρονικών σειρών

Τα τελευταία χρόνια έχει αυξηθεί ραγδαία η ανάγκη της εξόρυξης δεδομένων των χρονικών σειρών, όχι μόνο επειδή οι αλγόριθμοι της εξόρυξης δεδομένων έχουν την δυνατότητα να επεξεργάζονται δεδομένα με τεράστιο μέγεθος, αλλά και επειδή η ανάγκη των αλγόριθμων παρουσιάζεται πλέον στην στατιστική, στην μηχανική μάθηση, στην επεξεργασία σήματος, στην ανάκτηση των πληροφοριών και στα μαθηματικά. Οι κλασσικοί αλγόριθμοι αναλαμβάνουν δεδομένα χαμηλών διαστάσεων σε αντίθεση με τους αλγόριθμους της εξόρυξης δεδομένων χρονοσειρών, διότι είναι σε θέση να ασχοληθούν με δεδομένα εκατοντάδων και χιλιάδων διαστάσεων. Όταν τα δεδομένα είναι υψηλών διαστάσεων ο χρόνος εκτίμησης και ο χρόνος υπολογισμού των δεδομένων αυξάνεται, κατά συνέπεια να αλλοιωθεί το αποτέλεσμα. Στην περίπτωση αυτή η συνήθης προσέγγιση των αλγόριθμων κατηγοριοποίησης και ομαδοποίησης χάνει το νόημα τους. Για το λόγο αυτό, σχεδόν όλοι οι αλγόριθμοι χρονικών σειρών στο πεδίο της εξόρυξης δεδομένων αποφεύγουν να χρησιμοποιούν τα ακατέργαστα δεδομένα (raw data). Σε πιο πολύπλοκα δεδομένα, όπως είναι τα δεδομένα υψηλών διατάσεων, οι προσεγγίσεις Εξόρυξης Χρονολογικών Δεδομένων που έχουν προταθεί στη διεθνή βιβλιογραφία κατατάσσονται ανάλογα με το σκοπό της ανάλυσης σε αλγόριθμους συσταδοποίησης, κατηγοριοποίησης, πρόβλεψης, τμηματοποίησης, σύνοψης, ανίχνευσης ανωμαλιών και ευρετηριοποίησης (1) (3) (5) (6) (7).

2.4.1. Συσταδοποίηση χρονοσειρών (Clustering)

Η συσταδοποίηση έχει στόχο τον χωρισμό των δεδομένων της χρονοσειράς σε ομάδες, ώστε κάθε δεδομένο να κατατάσσεται στην ομάδα εκείνη όπου ανήκει, σύμφωνα με τα κριτήρια ή τα χαρακτηριστικά των δεδομένων που προσθέτουμε κάθε φορά και να απέχει όσο το δυνατόν περισσότερο από τα δεδομένα άλλων ομάδων. Για παράδειγμα, στα νοσοκομεία η συσταδοποίηση δεδομένων που μπορεί να χρησιμοποιηθεί είναι να ομαδοποιήσουν τους ασθενείς με βάση τα χαρακτηριστικά

τους. Αντίστοιχα σε μία εταιρία μπορεί να γίνει ομαδοποίηση πελατών με παρόμοια συμπεριφορά και ακόμα ομαδοποίηση μετοχών με παρόμοια διακύμανση τιμών. Η εφαρμογή συσταδοποίησης σε αυτούς αλλά και σε περισσότερους τομείς χρησιμοποιείται για την διευκόλυνση του ανθρώπου αλλά και στην κατανόηση κάποιων προτύπων για την περαιτέρω επεξεργασία.

Η ολοκλήρωση της διαδικασίας της συσταδοποίησης επιτυγχάνεται μετά από τα παρακάτω τέσσερα βήματα (7):

1. Επιλογή των χαρακτηριστικών γνωρισμάτων: αποτελεί μέρος της προεπεξεργασίας των δεδομένων και είναι αναγκαίο βήμα για την επίτευξη της καλύτερης ομοιογένειας των ομάδων με τα δεδομένα.
2. Επιλογή του αλγορίθμου συσταδοποίησης: η επιλογή του αλγορίθμου προσδιορίζεται από το μέτρο γειτνίασης και το κριτήριο συσταδοποίησης. Όσο πιο σωστός είναι ο αλγόριθμος τόσο πιο σωστό το σχήμα συσταδοποίησης.
3. Επικύρωση των αποτελεσμάτων: πρόκειται για την αξιολόγηση των αποτελεσμάτων όπου πραγματοποιείται, συγκρίνοντας τα με τα αποτελέσματα που ήδη γνωρίζουμε .
4. Ερμηνεία των αποτελεσμάτων: αναλύονται τα αποτελέσματα από ειδικούς. Τα αποτελέσματα αυτά σε συνδυασμό με άλλα στοιχεία βοηθούν στην εξαγωγή γνώσης και κατανόηση της κατανομής των δεδομένων.

Οι αλγόριθμοι που χρησιμοποιούνται για τη συσταδοποίηση εξετάζουν ομάδες με παρόμοιες εγγραφές σύμφωνα με τα χαρακτηριστικά των δεδομένων. Εξετάζουν τη συσχέτιση που έχουν τα δεδομένα μεταξύ τους, ώστε τα δεδομένα εκείνα με τη μεγαλύτερη συσχέτιση να είναι σε μία ομάδα και να απέχουν από τα δεδομένα που έχουν μικρή συσχέτιση. Μέθοδοι που χρησιμοποιούνται για τη συσταδοποίηση είναι πολλοί, όμως για την επιλογή της κατάλληλης μεθόδου εξετάζουμε το πρόβλημα που θέλουμε να λύσουμε. Οι τρεις κυριότερες κατηγορίες αλγόριθμων συσταδοποίησης είναι η ιεραρχική (hierarchical) συσταδοποίηση, η διαμεριστική (partitioning), και η συσταδοποίηση βασισμένη στην πυκνότητα (density based) (1) (2) (3) (5).

Ένα πρόβλημα της συσταδοποίησης, είναι η επιλογή του πλήθους των συστάδων όπου δε πρέπει να προσδιορίζεται τυχαία ή αυθαίρετα. Δεν είναι εύκολο να προσδιοριστεί το ακριβές πλήθος των συστάδων που χρειάζεται για την καλύτερη

ομαδοποίηση. Μία αλλαγή στο πλήθος των συστάδων θα φέρει αλλαγές στην ομαδοποίηση των δεδομένων, με αποτέλεσμα τα δεδομένα να καταταχθούν σε διαφορετική ομάδα. Επίσης είναι γνωστό ότι δεν υπάρχει μόνο μία σωστή λύση σε ένα πρόβλημα συσταδοποίησης.

Τέλος, τα αποτελέσματα της συσταδοποίησης μπορούν να βοηθήσουν στον ορισμό της κατηγοριοποίησης (όπως η ταξινόμηση των μετοχών που σχετίζονται με κάποια ομάδα) ή να βοηθήσουν στην κατασκευή των στατιστικών μοντέλων με τα οποία είναι εφικτό να γίνει η περιγραφή του πληθυσμού. Η συσταδοποίηση δημιουργεί κανόνες για ανάθεση νέων εγγραφών σε κλάσεις και χρησιμεύει για αναγνώριση και διάγνωση (7).

Μέτρα Απόστασης: Ευκλείδεια, City-block, DTW

Ο υπολογισμός των ομοιοτήτων των αντικειμένων στην συσταδοποίηση επιτυγχάνεται με βάση τα μέτρα ομοιότητας που χρησιμοποιούν μετρικές, δηλαδή την απόσταση μεταξύ δύο σημείων. Οι αποστάσεις **Minkowski** χρησιμοποιούνται στα περισσότερα προβλήματα και υπολογίζονται από τον τύπο:

$$d(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{1/p}$$

Βασισμένο στον παραπάνω τύπο έχουν δημιουργηθεί νέα μέτρα απόστασης αλλάζοντας την τιμή του p. Για $p = 2$ η απόσταση ονομάζεται **Ευκλείδεια** και δίνεται από τον τύπο:

$$d(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^2\right)^{1/2}$$

Για $p = 1$ η απόσταση ονομάζεται **City Block (Manhattan)** και ο τύπος της είναι:

$$d(x,y) = \sum_{i=1}^{d} |x_i - y_i|$$

Η ευκλείδεια απόσταση δεν είναι τόσο κατατοπιστική όταν οι τιμές των σημείων αλλάζουν με την πάροδο του χρόνου. Σε αυτή την περίπτωση χρησιμοποιείται μία άλλη μετρική, η **Dynamic Time Warping (DTW)**.

Η μετρική **Dynamic Time Warping** είναι μία μέθοδος υπολογισμού της απόστασης μεταξύ σημείων, που σκοπός της είναι να υπολογίζει τη βέλτιστη αντιστοιχία μεταξύ δύο χρονοσειρών που είναι διαφορετικές ως προς τον χρόνο ή την ταχύτητα, μέσω κάποιων περιορισμών (7) (8). Αρχικό βήμα για τον υπολογισμό της απόστασης αυτής είναι να οριστεί το μονοπάτι στρέβλωσης (warping path), ένα μονοπάτι που αποτελείται από συνεχόμενα στοιχεία ενός πίνακα και ορίζει μια μη γραμμική απεικόνιση μεταξύ δυο χρονοσειρών αποσκοπώντας στην ελαχιστοποίηση της απόστασης μεταξύ των χρονοσειρών. Έστω δυο χρονοσειρές $Q = q_1, q_2, \cdots, q_n$ και $C = c_1, c_2, \cdots, c_m$ με μήκος n και m αντίστοιχα. Για να οριστεί το μονοπάτι στρέβλωσης χρησιμοποιείται ο παρακάτω τύπος:

$$w(i,j) = d(q_i, c_j) + \min\{w(i-1, j-1), w(i-1, j), w(i, j-1)\}$$

όπου $i = 1, \cdots, n$ και $j = 1, \cdots, m$.

Αφού υπολογιστεί το μονοπάτι στρέβλωσης χρησιμοποιείται ο παρακάτω τύπος για τον υπολογισμό της απόστασης:

$$DTW(Q, C) = \min\left\{\frac{\sqrt{\sum_{k=1}^{K} W_k}}{K}\right\}$$

όπου k χρησιμοποιείται για να αντιμετωπίσει το διαφορετικό μήκος των χρονοσειρών.

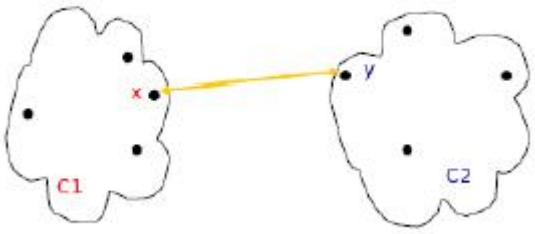
Ιεραρχική συσταδοποίηση

Σε αυτή τη κατηγορία μεθόδων δημιουργείται ένα σύνολο από εμφωλευμένες ομάδες (υπο-ομάδες) στην αρχική και μεγάλη ομάδα που οργανώνονται σε ένα ιεραρχικό δέντρο που μπορεί να παρασταθεί για την κατανόηση του με ένα δενδρόγραμμα. Σε κάθε επίπεδο της ιεραρχίας υπάρχουν υπο-ομάδες με πολλά δεδομένα (αντικείμενα) ενώ στο τελευταίο επίπεδο κάθε αντικείμενο ξεχωριστά αποτελεί μια ομάδα. Οι ιεραρχικοί αλγόριθμοι διαφέρουν ως προς τον τρόπο που δημιουργούνται τα σύνολα των συστάδων. Η ιεραρχική συσταδοποίηση διακρίνεται σε δύο μεγάλες υποκατηγορίες διαιρετικούς αλγόριθμους και τους συσσωρευτικούς. Ο **Διαιρετικός Αλγόριθμος (Divisive)** ξεκινάει από τα πρώτα επίπεδα και κατεβαίνει προς τα κάτω (top-down). Αρχικά όλα τα σημεία θεωρούνται μέλη της ίδιας ομάδας, προχωρά προς τα κάτω, διαιρεί τις μεγάλες ομάδες σε μικρότερες και τοποθετεί εκεί

τα σημεία με τη μεγαλύτερη ομοιογένεια μέχρι να φτάσει στο επιθυμητό αποτέλεσμα. Ο **Συσσωρευτικός Αλγόριθμος (Agglomerative)** λειτουργεί αντίθετα από τον διαιρετικό, καθώς ξεκινά τη διαδικασία από κάτω προς τα πάνω (bottom-up). Έχοντας αρχικά το κάθε σημείο μόνο του σε μία ομάδα, προχωρώντας προς τα πάνω επίπεδα, εξετάζει και συγχωνεύει σε κάθε βήμα τα πιο όμοια σημεία σε ζεύγη ομάδων. Η διαδικασία της συγχώνευσης επαναλαμβάνεται μέχρι να μείνουν k συστάδες (ή μία συστάδα), είτε μέχρι έναν προκαθορισμένο αριθμό συστάδων, αρκεί το κάθε σημείο να έχει ενταχθεί σε μία ομάδα.

Στην συσσωρευτική ιεραρχική συσταδοποίηση η ένωση ή ο διαχωρισμός των συστάδων εκτελείται σύμφωνα με κάποια κριτήρια σύνδεσης. Τα κριτήρια αυτά είναι:

Απόσταση απλού συνδέσμου (single-link clustering): βρίσκει τα δύο σημεία διαφορετικών ομάδων με την μικρότερη απόσταση και αύτη την απόσταση τη θεωρεί και ως την απόσταση των συστάδων.

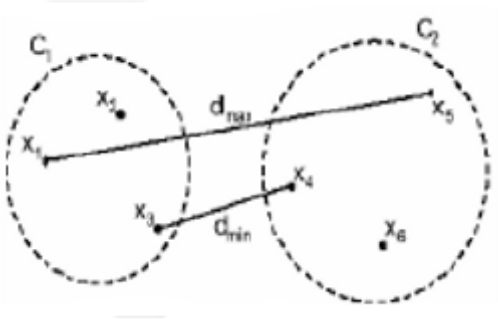


Εικόνα 5 Ελάχιστη απόσταση μεταξύ συστάδων

$$d_{min}(C_1, C_2) = \min_{x \in C_1, y \in C_2} \|x - y\|,$$

όπου $\|\cdot\|$ είναι οποιαδήποτε μέτρική απόστασης.

Απόσταση πλήρους συνδέσμου (Complete-link clustering): αντίθετη από την παραπάνω, αφού βρίσκει τα δύο σημεία διαφορετικών ομάδων που απέχουν περισσότερο από οποιοδήποτε άλλο και αυτή θεωρείται η απόσταση μεταξύ των συστάδων.

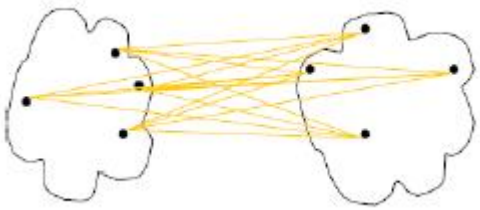


Εικόνα 6 Μέγιστη απόσταση μεταξύ συστάδων

$$d_{max}(C_1, C_2) = \max_{x \in C_1, y \in C_2} \|x - y\|,$$

όπου $\|\cdot\|$ είναι οποιαδήποτε μέτρική απόστασης.

Απόσταση μέσου συνδέσμου (Average-link clustering): εξετάζει όλα τα σημεία της μίας συστάδας με όλα τα σημεία της άλλης, βρίσκει τη μέση απόσταση και αυτή θεωρεί και ως απόσταση των δύο συστάδων.

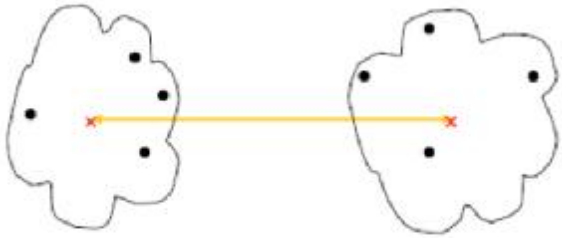


Εικόνα 7 Μέση απόσταση μεταξύ των συστάδων

$$proximity\left(C_i, C_j\right) = \frac{\sum_{\substack{x \in C_i \\ y \in C_j}} proximity(x, y)}{m_i \times m_j},$$

όπου m_i και m_j εκφράζουν το πλήθος των σημείων που κείνται εντός των συστάδων C_i και C_j αντίστοιχα

Απόσταση κέντρων βάρους (Centroid-link clustering): βρίσκει την ομοιότητα μεταξύ δύο συστάδων, δηλαδή την απόσταση ανάμεσα στα κέντρα βάρους των δύο συστάδων.



Εικόνα 8 Απόσταση κέντρων βάρους δύο συστάδων

Απόσταση με τη μέθοδο του Ward (Ward-link clustering): η ομοιότητα μεταξύ δύο συστάδων ορίζεται ως η αύξηση του τετραγωνικού σφάλματος που προκύπτει όταν συγχωνεύονται δύο συστάδες.

$$D_w(C_i, C_j) = \sum_{x \in C_i} (x - r_i)^2 + \sum_{x \in C_j} (x - r_j)^2 - \sum_{x \in C_{ij}} (x - r_{ij})^2$$

όπου D_w είναι η απόσταση του Ward μεταξύ των συστάδων C_i και C_j, δηλαδή είναι η διαφορά μεταξύ του ολικού λάθους των δύο συστάδων και του ολικού λάθους αν ενώσουμε τις 2 συστάδες σε μια συστάδα, έστω την C_{ij} (2) (7).

Βασικό πλεονέκτημα της ιεραρχικής συσταδοποίησης είναι πως δεν χρειάζεται η εισαγωγή του αριθμού των ομάδων, αρκεί να κοπεί το δενδρόγραμμα στο κατάλληλο επίπεδο. Επίσης περιέχει εμφωλευμένες ομάδες που αυτό επιτρέπει στους χρήστες να κάνουν διάφορους διαχωρισμούς των σημείων ανάλογα με το επιθυμητό επίπεδο ομοιότητας. Ένα από τα μειονεκτήματα είναι πως ο αλγόριθμος ιεραρχικής συσταδοποίησης δεν έχει δυνατότητα οπισθοδρόμησης, δεν μπορεί να αναιρέσει, να παραβλέψει ότι έχει ήδη δημιουργηθεί και να πάει πίσω. Επίσης η μέθοδος απλού συνδέσμου μεταξύ των συστάδων είναι πιο ευέλικτη, διότι ενώνει λίγα σημεία μεταξύ των συστάδων, με αποτέλεσμα μερικά σημεία που σχηματίζουν την απόσταση ανάμεσα σε δύο συστάδες προκαλέσουν την ένωση των δύο αυτών συστάδων (chaining effect). Στον μέσο σύνδεσμο, μπορεί να προκαλέσει επιμήκυνση των συστάδων για τον διαχωρισμό και να ενοποιήσει τμήματα επιμηκών γειτονικών συστάδων. Τέλος η μέθοδος απόσταση πλήρους συνδέσμου δημιουργεί πιο συμπαγείς συστάδες και πιο χρήσιμες ιεραρχίες από ότι η μέθοδος απλού συνδέσμου (7).

Διαμεριστική συσταδοποίηση

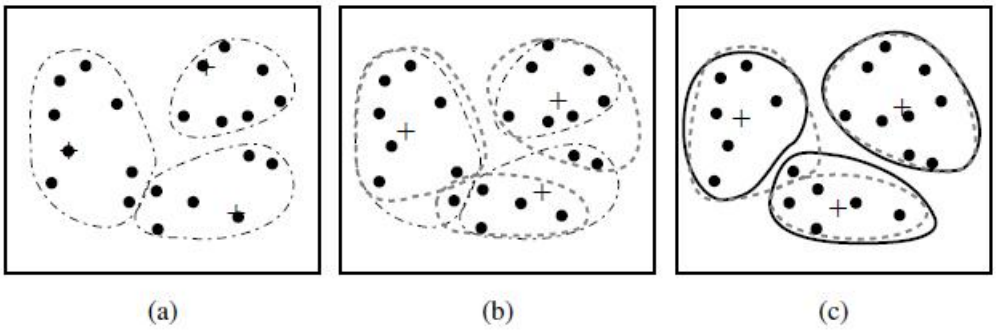
Αντίθετα από την ιεραρχική συσταδοποίηση, η διαμεριστική δεν σχηματίζει κάποια ιεραρχική δομή, ο διαχωρισμός των σημείων σε συστάδες πραγματοποιείται σε ένα βήμα και ως αποτέλεσμα λαμβάνεται μόνο ένα σύνολο συστάδων, παρά το γεγονός ότι εσωτερικά μπορεί να δημιουργηθούν αρκετά διαφορετικά σύνολα συστάδων (1) (2) (3) (5) (7). Αφού δημιουργήσει μία αρχική τμηματοποίηση στη συνέχεια επαναλαμβάνει τη διαδικασία διαχωρισμού των σημείων με σκοπό να βελτιώσει την τμηματοποίηση. Κάθε τμηματοποίηση αποτελείται από μία συστάδα όπου κάθε ομάδα έχει τουλάχιστον ένα δεδομένο και κάθε ένα δεδομένο βρίσκεται σε μόνο μία συστάδα. Επ (8) (8)αναλαμβάνει τη διαδικασία με κριτήριο τα δεδομένα μίας συστάδας να είναι «κοντά» το ένα με το άλλο, ενώ αντικείμενα από άλλες συστάδες να είναι «μακριά». Για να συμβεί αυτό και να ικανοποιηθεί αυτό το κριτήριο πρέπει να υπολογισθούν όλες οι πιθανές διαχωρίσεις δεδομένων (n - στοιχείων σε k-συστάδες) όπου είναι εξαντλητικό και αυξάνει την πολυπλοκότητά του, έτσι η αναζήτηση βέλτιστης λύσης περιορίζεται σε ένα μικρό υποσύνολο των πιθανών λύσεων. Επίσης μειονέκτημα αυτού του αλγόριθμου είναι πως απαιτείται η εισαγωγή του επιθυμητού πλήθους των συστάδων σε αντίθεση με την ιεραρχική συσταδοποίηση.

Μία από τις πιο γνωστές και ευρέως χρησιμοποιούμενη μέθοδος διαμεριστικής συσταδοποίησης, είναι ο αλγόριθμος k-μέσων (k–means), που το 1967 υλοποιήθηκε από τον Mac Queen (1) (2) (3) (5) (7). Αρχικά ο αλγόριθμος k-means λαμβάνει την παράμετρο εισόδου k, και χωρίζει ένα σύνολο n αντικειμένων σε k ομάδες, έτσι ώστε η ομοιότητα της ομάδας που προκύπτει να είναι υψηλή, αλλά η ομοιότητα με τις άλλες ομάδες να είναι χαμηλή. Η ομοιότητα της ομάδας μετράται σε σχέση με τη μέση τιμή των αντικειμένων σε μία ομάδα, το οποίο μπορεί να θεωρηθεί ως το κέντρο βάρους του συνόλου της ομάδας/συστάδας. Πρώτον, επιλέγει τυχαία k ομάδες που περιλαμβάνουν n αντικείμενα, καθένα από τα οποία τα χωρίζει σε μία ομάδα έτσι ώστε σε κάθε ομάδα να έχει εκχωρηθεί εκείνο το αντικείμενο με την μεγαλύτερη ομοιότητα, με βάση την απόσταση μεταξύ του αντικειμένου και το κέντρο βάρους της ομάδας. Στη συνέχεια υπολογίζει το νέο μέσο για την κάθε συστάδα και κάνει ότι αλλαγές χρειάζεται ώστε να φέρει το καλύτερο αποτέλεσμα. Για να βρει την ελάχιστη τιμή του αθροίσματος του τετραγωνικού σφάλματος χρησιμοποιεί τη παρακάτω σχέση:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 ,$$

όπου E είναι το άθροισμα του τετραγωνικού σφάλματος για όλα τα αντικείμενα στο σύνολο δεδομένων, p είναι το σημείο που βρίσκεται ένα συγκεκριμένο αντικείμενο στο χώρο και m_i είναι ο μέσος όρος συστάδας C_i (και το p και το C_i είναι πολυδιάστατα). Για κάθε αντικείμενο σε κάθε συστάδα, η απόσταση από το αντικείμενο προς το κέντρο βάρους της συστάδας είναι στο τετράγωνο και οι αποστάσεις αθροίζονται. Ο αλγόριθμος προσπαθεί να κάνει τις συστάδες k συμπαγής και όσο το δυνατόν λιγότερες. Ως είσοδο ο k-μέσων δέχεται τον αριθμό συστάδων (k) και ένα σύνολο δεδομένων που περιέχει n αντικείμενα (D), και ως έξοδο δίνει ένα σύνολο από k ομάδες σωστά διαχωρισμένες. Η μέθοδος του αλγορίθμου είναι η εξής:

1. Επιλέγει αυθαίρετα k αντικείμενα από D σύνολο δεδομένων ως αρχικά κέντρα βάρους της συστάδας
2. Σχηματίζει k συστάδες αποδίδοντας κάθε σημείο στο πλησιέστερο κέντρο βάρους του
3. Υπολογίζει ξανά το κέντρο βάρους κάθε συστάδας
4. Επαναλαμβάνει μέχρι να μην μπορεί να κάνει άλλη αλλαγή στα κέντρα βάρους



Εικόνα 9 Παράδειγμα χωρισμού των δεδομένων σε συστάδες/ομάδες

Το γεγονός ότι ζητά από τον χρήστη να εισάγει ως είσοδο τον αριθμό k των συστάδων, μπορεί να θεωρηθεί ως μειονέκτημα.

Άλλοι διαμεριστικοί αλγόριθμοι που χρησιμοποιούνται ευρέως ο αλγόριθμος PAM, CLARA KAI CLARANS (1) (2) (3) (5) (7).

Εκτίμηση συσταδοποίησης

Για την εκτίμηση της συσταδοποίησης συχνά χρησιμοποιείται η μέθοδος του **Συντελεστή Περιγράμματος** (Silhouette Coefficient) (5). Αυτή η μέθοδος συνδυάζει τον υπολογισμό τόσο της συνοχής που υπάρχει σε μία συστάδα αλλά και τον διαχωρισμό της συστάδας με τις υπόλοιπες. Ο υπολογισμός του συντελεστή περιγράμματος ολοκληρώνεται με τα παρακάτω βήματα:

για κάθε σημείο i υπολόγισε

1. τη μέση απόσταση του σημείου i από τα υπόλοιπα σημεία της συστάδας του, απόσταση a.
2. τη μέση απόσταση του σημείου i από όλα τα σημεία των άλλων συστάδων και επέλεξε την μικρότερη μέση απόσταση, απόσταση b.

Ο υπολογισμός του Συντελεστή Περιγράμματος προκύπτει από τον τύπο:

$$s_i = \frac{(b-a)}{\max(a,b)}$$

Η τιμή του συντελεστή περιγράμματος είναι μεταξύ του -1 και του 1. Οι αρνητικές τιμές δεν είναι επιθυμητές καθώς προκύπτει ότι η απόσταση α είναι μεγαλύτερη από την απόσταση b με αποτέλεσμα ότι ένα σημείο έχει ενταχθεί σε λάθος ομάδα. Οπωσδήποτε πρέπει ο συντελεστής περιγράμματος να είναι θετικός που προκύπτει ότι το σημείο έχει ενταχθεί στη συστάδα με τα πιο όμοια του σημεία.

Για τον υπολογισμό όλων των σημείων των συστάδων για την εκτίμηση πόσο καλή είναι η συσταδοποίηση χρησιμοποιείται ο τύπος του Μέσου Συντελεστή Περιγράμματος :

$$SC = \frac{1}{N}\sum_{i=1}^{N} s_i$$

2.4.2. Κατηγοριοποίηση - Ταξινόμηση χρονοσειρών (Classification)

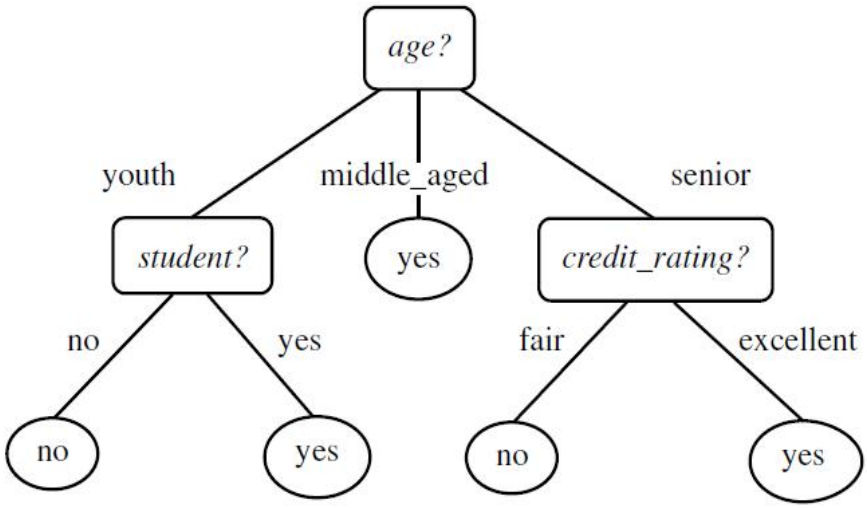
Η κατηγοριοποίηση είναι μία από τις πιο γνωστές τεχνικές εξόρυξης δεδομένων. Ορίζεται ως εποπτευόμενη μάθηση και δείχνει τα δεδομένα που εισέρχονται σε μία από τις είδη καθορισμένες κλάσεις (1) (2) (3) (5). Η αναγνώριση προτύπων μπορεί να θεωρηθεί ως ένα είδος κατηγοριοποίησης αφού έχει ως στόχο την ταξινόμηση δεδομένων με βάση την ομοιότητα. Η κατηγοριοποίηση είναι χρήσιμη για την πρόβλεψη ή την περιγραφή ενός συνόλου δεδομένων τα οποία έχουν δυαδικές ή ονομαστικές κατηγορίες. Είναι λιγότερο αποτελεσματικές για κατηγορίες τακτικών χαρακτηριστικών, όπως για παράδειγμα η κατηγοριοποίηση ενός ατόμου σε μία ομάδα ανάλογα με το εισόδημά του (υψηλό, χαμηλό, μεσαίο), διότι δεν λαμβάνουν υπόψη την υπονοούμενη σειρά μεταξύ των κατηγοριών (2).

Έτσι η κατηγοριοποίηση έχοντας μία βάση δεδομένων και κάποιες από τις κατηγορίες που αναφέρονται παραπάνω, μπορεί να δημιουργήσει ένα μοντέλο το οποίο είναι ικανό να καταχωρεί με επιτυχία κάθε αντικείμενο της βάσης σε μία μόνο από τις κατηγορίες αυτές (3) (7). Η μέθοδος του κοντινότερου γείτονα (k nearest neighbor) και τα δέντρα αποφάσεων (decision trees) είναι οι δύο πιο δημοφιλείς μέθοδοι που χρησιμοποιούνται στην κατηγοριοποίηση. Οι κατηγοριοποιητές δένδρων απόφασης παρομοιάζονται με τους **πρόθυμους μαθητές (eager leaners)**, δηλαδή αυτός που όταν του δίνεται ένα σύνολο από πλειάδες εκπαίδευσης, θα κατασκευάσει μια γενίκευση (μοντέλο ταξινόμησης) πριν την λήψη νέων (π.χ. δοκιμή) πλειάδων για ταξινόμηση (2).

Κατηγοριοποίηση με επαγωγή δέντρου απόφασης (Classification by Decision Tree Induction): το δέντρο απόφασης είναι μια απλή, ευρέως χρησιμοποιούμενη τεχνική κατηγοριοποίησης και δημιουργείται από μια βάση δεδομένων $D = \{t_1, \cdots, t_n\}$, όπου $t_i = [t_{i1}, \cdots t_{ih}]$ είναι μια περίπτωση η οποία χαρακτηρίζεται από τα γνωρίσματα $\{A_1, A_2, \cdots, A_h\}$. Επίσης δίνεται και ένα σύνολο από κατηγορίες $C = \{C_1, \cdots, C_m\}$. Ένα δέντρο απόφασης έχει τέσσερις ιδιότητες: ο **κόμβος ρίζα** δεν έχει εισερχόμενες ακμές και μηδέν ή περισσότερες εξερχόμενες, ο **εσωτερικός κόμβος** παίρνει το όνομά του από ένα γνώρισμα A_i, και έχει ακριβώς μια εισερχόμενη ακμή και δύο ή πιο περισσότερες εξερχόμενες, τα **φύλλα** ή **τερματικοί κόμβοι** έχουν ως όνομα μια κατηγορία C_i και έχει ακριβώς μια εισερχόμενη ακμή και μια εξερχόμενη και τέλος η

ακμή παίρνει το όνομά της από ένα κατηγόρημα-τιμή, το οποίο εφαρμόζεται στο γνώρισμα που συνδέεται με τον κόμβο ρίζα. Τα γνωρίσματα που χρησιμοποιούνται για να ονοματίσουν τους κόμβους του δέντρου και γύρω από τα οποία θα λάβουν χώρα οι διαιρέσεις ονομάζονται **γνωρίσματα διαχωρισμού** (splitting attributes) και τα γνωρίσματα που χρησιμοποιούνται για να ονοματίσουν τις ακμές του δέντρου ονομάζονται **κατηγορήματα διαχωρισμού** (splitting predicates) (1) (2) (3) (5).

Ένα παράδειγμα τυπικού δέντρου απόφασης απεικονίζεται στην παρακάτω εικόνα και προβλέπει αν ένας πελάτης του *AllElectronics* ενδέχεται να αγοράσει έναν υπολογιστή. Οι εσωτερικοί κόμβοι συμβολίζονται με ορθογώνιο και αντιπροσωπεύουν μια δοκιμή σε ένα γνώρισμα, ενώ οι κόμβοι φύλλα συμβολίζονται με οβάλ και αντιπροσωπεύουν μια κλάση, δηλαδή το *ενδεχόμενο (πιθανότητα) της αγοράς του υπολογιστή = ναι* (*either buys_computer = yes*) και την *αγορά του υπολογιστή = όχι* (*buys_computer = no*) (1) (2).

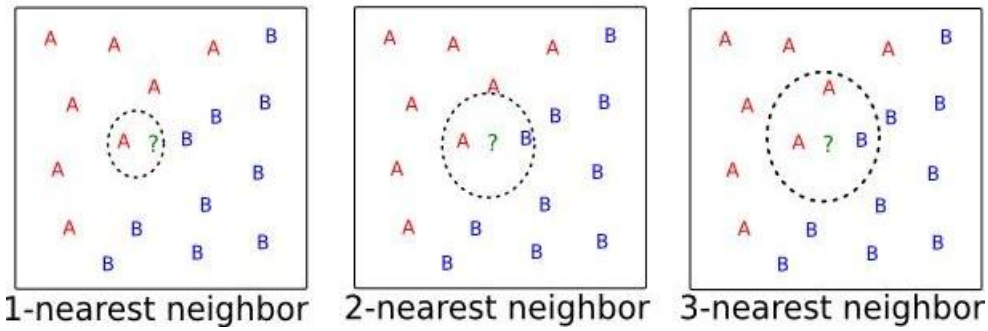


Εικόνα 10 Παράδειγμα τυπικού δέντρου απόφασης

Αντίθετη περίπτωση από αυτή του πρόθυμου μαθητή είναι ο **τεμπέλης μαθητής (lazy leaners)**, όπου περιμένει μέχρι τα τελευταία λεπτά πριν κάνει οποιαδήποτε κατασκευή του μοντέλου, προκειμένου να χαρακτηρίσει ένα συγκεκριμένο έλεγχο πλειάδας. Όταν δίνεται μια πλειάδα εκπαίδευσης, ένας τεμπέλης μαθητής αποθηκεύει και περιμένει μέχρι να δοθεί μια δοκιμή πλειάδα . Μόνο όταν βλέπει το τεστ κάνει γενίκευση, προκειμένου να χαρακτηρίσει την πλειάδα με βάση την ομοιότητά της με

την αποθηκευμένη πλειάδα εκπαίδευσης. Στην κατηγορία αυτή ανήκει η μέθοδος κοντινότερου γείτονα (2).

k-κοντινότερου γείτονα (k-Nearest-Neighbor): οι ταξινομητές κοντινότερου γείτονα που βασίζονται στην εκμάθηση κατ 'αναλογία, δηλαδή συγκρίνοντας δοσμένες πλειάδες με πλειάδες εκπαίδευσης. Οι πλειάδες εκπαίδευσης περιγράφονται από n χαρακτηριστικά. Κάθε πλειάδα αντιπροσωπεύει ένα σημείο σε ένα n-διάστατο χώρο. Με αυτόν τον τρόπο, όλες οι πλειάδες εκπαίδευσης αποθηκεύονται σε ένα χώρο n-διαστάσεων. Όταν δίνεται μια άγνωστη πλειάδα z, ο κατηγοριοποιητής αναζητά στο σύνολο εκπαίδευσης ακριβώς k πλειάδες που βρίσκονται πλησιέστερα προς την άγνωστη πλειάδα z. Η Εικόνα 11 απεικονίζει τους 1-,2-, και 3- κοντινότερους γείτονες ενός σημείου. Το σημείο δεδομένων κατηγοριοποιείται με βάση τις ετικέτες κατηγορίας των γειτόνων. Σε περίπτωση που οι γείτονες έχουν περισσότερες από μια ετικέτες, το σημείο δεδομένων αποδίδεται στην κατηγορία πλειοψηφίας των πλησιέστερων γειτόνων. Στο πρώτο σχήμα της Εικόνας 11 ο 1-πλησιέστερος γείτονας του σημείου δεδομένων είναι το Α. Αν το πλήθος των πλησιέστερων γειτόνων είναι 3, όπως φαίνεται στο τρίτο σχήμα, τότε η γειτονιά περιλαμβάνει δύο Α και ένα Β. Χρησιμοποιώντας τη στρατηγική της ψήφου πλειοψηφίας, το σημείο δεδομένων αποδίδεται στην Α κατηγορία. Στην περίπτωση που υπάρχει ισοψηφία μεταξύ των κατηγοριών, όπως στο δεύτερο σχήμα, η γειτονιά περιλαμβάνει ένα Α και ένα Β, τότε επιλέγεται τυχαία ένα από αυτά για να κατηγοριοποιηθεί το σημείο δεδομένων (2) (7).



Εικόνα 11 Παράδειγμα κοντινότερων γειτόνων ενός σημείου

Μπεϋζιανή Κατηγοριοποίηση (Bayesian classification): σε πολλές εφαρμογές η ετικέτα κατηγορίας μιας εγγραφής ελέγχου δεν μπορεί να προβλεφθεί με βεβαιότητα, ακόμη και αν το σύνολο χαρακτηριστικών της είναι όμοιο με ένα από τα δείγματα εκπαίδευσης (1) (2). Η κατάσταση αυτή προκαλείται συνήθως εξαιτίας δεδομένων με θόρυβο ή ύπαρξης συγκεκριμένων παραγόντων οι οποίοι επηρεάζουν την κατηγοριοποίηση και δεν περιλαμβάνονται στην ανάλυση. Για παράδειγμα ο προσδιορισμός του αν η διατροφή ενός ατόμου είναι υγιεινή ή αν η συχνότητα άσκησης είναι επαρκής αποτελεί θέμα ερμηνείας, το οποίο μπορεί να εισάγει αβεβαιότητες στο πρόβλημα εκπαίδευσης. Η μπεζϋζιανή κατηγοριοποίηση βασίζεται στον κανόνα του Bayes και χρησιμοποιείται για να επιλύει τα προβλήματα κατηγοριοποίησης. Ένας κατηγοριοποιητής Bayes εκτιμά την εξαρτώμενη από την κατηγορία πιθανότητα υποθέτοντας ότι τα χαρακτηριστικά είναι υπό συνθήκη ανεξάρτητα, δεδομένης μιας ετικέτας κατηγορίας y. Η υπόθεση της υπό συνθήκη ανεξαρτησίας μπορεί να εκφραστεί με τον τύπο:

$$P(X|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y)$$

όπου κάθε σύνολο χαρακτηριστικών $X = \{X_1, X_2, \cdots, X_d\}$ αποτελείται από d χαρακτηριστικά. Με την συνθήκη ανεξαρτησίας, αντί να υπολογίζεται η εξαρτώμενη από την κατηγορία πιθανότητα για κάθε συνδυασμό του X, αρκεί να εκτιμηθεί η υπό συνθήκη πιθανότητα για κάθε X_i δοθέντος του Y. Η προσέγγιση αυτή είναι πιο πρακτική επειδή δεν απαιτεί ένα πολύ μεγάλο σύνολο εκπαίδευσης για να υπολογιστεί μία καλή εκτίμηση της πιθανότητας. Για να κατηγοριοποίηση μια εγγραφή ελέγχου, ο απλοϊκός κατά Bayes κατηγοριοποιητής υπολογίζει την εκ των υστέρων πιθανότητα για κάθε κατηγορία Y:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y)}{P(X)}$$

Δεδομένου ότι η τιμή $P(X)$ είναι σταθερή για κάθε Y, αρκεί να επιλεγεί η κατηγορία που μεγιστοποιεί τον αριθμητή $P(Y)\prod_{i=1}^{d} P(X_i|Y)$ (2).

2.4.3. Πρόβλεψη χρονοσειρών

Τα Προγνωστικά μοντέλα αναλύουν τις παρατηρήσεις μιας χρονοσειράς για να επισημάνουν αυτά που έχουν συμβεί στο παρελθόν, προκειμένου να προβλέψουν την πιο πιθανή συμπεριφορά των χρονοσειρών στον μέλλον. Η f_t είναι η πρόβλεψη της τυχαίας τιμής Y_t για την περίοδο t. Το t είναι ο δείκτης της τρέχουσας περιόδου και οι τιμές $\{y_t, y_{t-1}, \cdots, y_{t-k+1}\}$ είναι ήδη γνωστές τιμές της χρονοσειράς για k περιόδους στο παρελθόν. Η γενική δομή ενός μοντέλου πρόβλεψης είναι (5):

$$f_{t+1} = F(y_t, y_{t-1}, \cdots, y_{t-k+1})$$

Για να αναπτυχθεί ένα προγνωστικό μοντέλο είναι απαραίτητο να επιλεχθεί η μορφή της συνάρτησης F που αντιπροσωπεύει την συγκεκριμένη χρονική σειρά που γίνεται η ανάλυση. Εφόσον γίνει αυτό, τότε μπορεί κανείς να προσπαθήσει να βρει τιμές για τις παραμέτρους της συνάρτησης F.

Σε ορισμένες περιπτώσεις θα μπορούσε κανείς να δημιουργήσει τις προβλέψεις για ένα δεδομένο αριθμό περιόδων στο μέλλον και όχι μόνο για την επόμενη περίοδο $t + 1$. Οι προβλέψεις έγιναν τη χρονική στιγμή t για περιόδους πέραν τις $t + 1$ εφαρμόζοντας το μοντέλο με τις τιμές που είναι ήδη γνωστές μέχρι τη χρονική στιγμή t, καθώς και με τις προβλέψεις που έγιναν χρησιμοποιώντας το ίδιο μοντέλο για τις επόμενες περιόδους, δηλαδή:

$$f_{t+h} = F(f_{t+h-1}, f_{t+h-2}, \cdots, f_{t+1}, y_t, y_{t-1}, \cdots, y_{t-k+1})$$

Αναμένεται πως οι προβλέψεις θα είναι όλο και λιγότερο ακριβείς όσο η περίοδος πρόβλεψης $t + h$ απομακρύνεται περισσότερο από τον χρονικό ορίζοντα t. Η φορά των δεικτών θα αντιστοιχεί στην ακολουθία των φυσικών αριθμών $\{1,2,3, \cdots\}$, έτσι η πρώτη περίοδος της χρονοσειράς αριθμείται με 1 και ούτω καθεξής (5).

Για την πρόβλεψη χρονοσειρών υπάρχουν βασικά στατιστικά μοντέλα όπως γραμμική παλινδρόμηση (linear regression), αυτοπαλινδρόμηση (Autoregression - AR), κινούμενοι μέσοι όροι (Moving Average-MA), αυτοπαλίνδομος κινητός μέσος όρος (autoregressive moving average -ARMA) και ολοκληρωμένος αυτοπαλίνδρομος κινητός μέσος όρος (autoregressive integrated moving average – ARIMA). Εκτός από τα βασικά μοντέλα αυτά, η πρόβλεψη χρονοσειρών μπορεί να γίνει και με τα νευρωνικά δίκτυα (neural networks) (6) (7) (9) (10).

Γραμμική παλινδρόμηση (Linear Regression)

Η γραμμική παλινδρόμηση προβλέπει με μια γραμμική συνάρτηση των δεικτών πρόβλεψης ως εξής:

$$y = c_0 + c_1 x_1 + c_2 x_2 + \cdots + c_k x_k \qquad (1)$$

όπου $x_1, x_2, \cdots, x_k$ είναι οι παράγοντες πρόβλεψης και y η απάντηση της πρόβλεψης (4).

Αυτοπαλινδρόμηση (Autoregression - AR)

Η αυτοπαλινδρόμηση είναι ένα μοντέλο της γραμμικής παλινδρόμησης, όπου η εξαρτημένη μεταβλητή θεωρείται η τυχαία μεταβλητή της χρονοσειράς σε μια χρονική στιγμή t, x_t, και ως ανεξάρτητες μεταβλητές θεωρούνται οι τυχαίες μεταβλητές της χρονοσειράς σε προηγούμενους χρόνους, δηλαδή $x_{t-1}, \cdots, x_{t-p}$. Ο αριθμός των υστερήσεων που συμπεριλαμβάνεται καλείται τάξη (order). Ένα **αυτοπαλινδρούμενο μοντέλο τάξης p** συμβολίζεται $AR(p)$ και ορίζεται ως (6) (9):

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \cdots + \varphi_p x_{t-p} + z_t \qquad (2)$$

όπου z_t υποδηλώνει μια τυχαία διαδικασία με μηδενική μέση τιμή και διασπορά σ_z^2 και ονομάζεται λευκός θόρυβος[1], και $\varphi_1, \varphi_2, \cdots, \varphi_p$ είναι οι συντελεστές του μοντέλου. Για την απλοποίηση του τύπου *(2)* μπορεί να χρησιμοποιηθεί ένας τελεστής B, όπου $x_{t-1} = Bx_t$, και έτσι το μοντέλο $AR(p)$ θα γραφεί συνοπτικά με τη μορφή (9) (6):

$$\varphi(B)x_t = z_t \qquad (3)$$

όπου $\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$ είναι ένα πολυώνυμο B τάξης p.

[1] Λευκός θόρυβος (white noise): ο λευκός θόρυβος είναι μία στάσιμη χρονοσειρά. Η χρονοσειρά αυτή αποτελείται από ανεξάρτητες τυχαίες μεταβλητές με την ίδια κατανομή (independent and identically distributed, idd) . Μία idd χρονοσειρά είναι εντελώς τυχαία και δεν έχει συσχετίσεις μεταξύ στοιχείων της χρονοσειράς. Η κατανομή συμβολίζεται ως $WN\ (0, \sigma_\varepsilon^2)$ με μηδενική μέση τιμή και διασπορά σ_ε^2.

Οι ιδιότητες του μοντέλου AR, από τον τύπο *(2)*, μπορούν να εξεταστούν εξετάζοντας τις ιδιότητες της συνάρτησης φ. Καθώς ο B είναι ένας τελεστής, οι αλγεβρικές ιδιότητες της φ θα πρέπει να διερευνηθούν από την εξέταση των ιδιοτήτων της $\varphi(x)$, όπου x συμβολίζει μια μεταβλητή, παρά από την $\varphi(B)$ (6). Με την προϋπόθεση ότι οι ρίζες του $\varphi(x) = 0$ το μοντέλο μπορεί να γραφεί:

$$x_t = \sum_{j\geq 0}^{\infty} \psi_j \, z_{t-j} \qquad (4)$$

για ορισμένες σταθερές ψ_j τέτοιες ώστε $\sum|\psi_j| < \infty$.

Ένα παράδειγμα μιας διαδικασίας AR, είναι η περίπτωση «πρώτης τάξης» (first-order), και δίνεται από τον τύπο:

$$X_t = \varphi X_{t-1} + Z_t \qquad (5)$$

Υπάρχει μια μοναδική στάσιμη λύση του τύπου *(5)*, υπό την προϋπόθεση ότι $|\varphi| < 1$. Η συνάρτηση αυτοσυσχέτισης από μια στάσιμη $AR(1)$ διαδικασία δίνεται από $\rho_k = \varphi^k$ για $k = 0,1,2, \ldots$. Η συνάρτηση αυτοσυσχέτισης μπορεί να βρεθεί από την επίλυση μιας σειράς διαφορετικών εξισώσεων που ονομάζονται Yule – Walker εξισώσεις, και υπολογίζεται από τον τύπο (6) (7):

$$\rho_k = \varphi_1\rho_{k-1} + \varphi_2\rho_{k-2} + \cdots + \varphi_p\rho_{k-p} \qquad (6)$$

για $k = 0,1,2, \ldots$, όπου $\rho_0 = 0$.

Κινούμενοι Μέσοι Όροι (Moving Average - MA)

Μια χρονοσειρά x_t λέγεται ότι είναι **κινητός μέσος όρος** της τάξης q, συμβολίζεται $MA(q)$, αν είναι γραμμικό άθροισμα των τελευταίων q τυχαίων δονήσεων (random shocks) έτσι ώστε (6):

$$x_t = z_t + \theta_1 z_{t-1} + \cdots + \theta_q z_{t-q} \qquad (7)$$

όπου z_t υποδηλώνει μια τυχαία διαδικασία με μηδενική μέση τιμή και διασπορά σ_z^2 και ονομάζεται λευκός θόρυβος και θ_t είναι οι παράμετροι. Ο τύπος *(7)* μπορεί να γραφτεί συνοπτικά:

$$x_t = \theta(B) z_t \qquad (8)$$

όπου $\theta(B) = 1 - \theta_1 B + \cdots + \theta_q B^q$ είναι ένα πολυώνυμο B τάξης q.

Μπορεί να αποδειχθεί ότι μια διαδικασία του κινούμενου μέσου όρου (MA) τάξης f είναι στάσιμη για όλες τις τιμές των παραμέτρων. Ωστόσο συχνά επιβάλλεται μια συνθήκη στις τιμές των παραμέτρων ενός μοντέλου MA και ονομάζεται «αντιστρέψιμη συνθήκη» (invertibility condition), προκειμένου να διασφαλιστεί ότι υπάρχει ένα μοναδικό μοντέλο MA για μία συνάρτηση αυτοσυσχέτισης (autocorrelation function). Αυτή η συνθήκη μπορεί να εξηγηθεί ως εξής:

Ας υποθέσουμε ότι Z_t και Z'_t είναι ανεξάρτητες τυχαίες διαδικασίες και ότι $\theta \in (-1,1)$. Έτσι είναι εύκολο να αποδειχθεί ότι δύο MA(1) διαδικασίες οι οποίες ορίζονται από $x_t = Z_t + \theta Z_{t-1}$ και $x_t = Z'_t + \theta^{-1} Z'_{t-1}$ έχουν την ίδια συνάρτηση αυτοσυσχέτισης. Έτσι το πολυώνυμο $\theta(B)$ δεν καθορίζεται μόνο από τη συνάρτηση αυτοσυσχέτισης, με συνέπεια λαμβάνοντας υπόψη ένα δείγμα από τη συνάρτηση αυτοσυσχέτισης δεν είναι δυνατόν να εκτιμηθεί μια μοναδική διαδικασία MA από συγκεκριμένο σύνολο δεδομένων χωρίς να έχουν οριστεί περιορισμοί σχετικά με το τι επιτρέπεται. Για να επιλυθεί το πρόβλημα αυτό, συνήθως απαιτείται το πολυώνυμο $\theta(x)$ να έχει όλες τις ρίζες του εκτός του μοναδιαίου κύκλου. Με αυτό τον τρόπο ο πρώτος τύπος μπορεί να γραφτεί με τη μορφή:

$$X_t - \sum_{j \geq 1} \pi_j X_{t-j} = Z_t \qquad (9)$$

για ορισμένες σταθερές π_j τέτοιες ώστε $\sum |\pi_j| < \infty$.

Η διαδικασία MA σχετίζεται με ένα μαθηματικό θεώρημα που ονομάζεται Wold (θεώρημα της διάσπασης – decomposition theorem), που δείχνει ότι κάθε σταθερή διαδικασία μπορεί να εκφραστεί ως το άθροισμα των δύο τύπων των διαδικασιών, ένα από τα οποία είναι μη-αιτιοκρατικό (non-deterministic), ενώ το άλλο είναι γραμμικά αιτιοκρατικό (linearly deterministic). Αυτοί οι όροι ορίζονται ως ακολούθως. Εάν η διαδικασία μπορεί να προβλεφθεί ακριβώς από την γραμμική παλινδρόμηση από τις προηγούμενες τιμές-δεδομένα, ακόμα και όταν οι καινούργιες – πρόσφατες τιμές-δεδομένα δεν είναι διαθέσιμες, τότε η διαδικασία ονομάζεται αιτιοκρατική (deterministic). Αν όμως οι προηγούμενες τιμές-δεδομένα δεν είναι αναγκαίες για την πρόβλεψη της διαδικασίας από την γραμμική παλινδρόμηση, τότε η διαδικασία ονομάζεται μη-αιτιοκρατική ή τακτική ή στοχαστική (non-deterministic). Το θεώρημα αυτό, και συγκεκριμένα η μη-αιτιοκρατική διαδικασία,

μπορεί να εκφραστεί ως μια διαδικασία MA με την προϋπόθεση ότι οι διαδοχικές τιμές της ακολουθίας Z_t είναι ασυσχέτιστες. Κάθε σταθερή non-deterministic χρονική σειρά μπορεί να εκφραστεί με την μορφή:

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \qquad (10)$$

όπου $\psi_0 = 1$ και $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, και Z_t υποδηλώνει μια τυχαία διαδικασία με μηδενική μέση τιμή και διασπορά σ_z^2 και ονομάζεται λευκός θόρυβος. Τα Z_t μερικές φορές αποκαλούνται καινοτομίες, όπως είναι το σφάλμα της πρόβλεψης, καθώς είναι ένα βήμα μπροστά, όταν η γραμμική πρόβλεψη χρησιμοποιείται ώστε να προχωρήσει στην πρόβλεψη. Ο παραπάνω τύπος *(10)* είναι μια μέθοδος της $MA(\infty)$, και ονομάζεται «Αναπαράσταση Wold» (Wold representation) της μεθόδου. Μερικές φορές όμως ονομάζεται «γραμμική μέθοδος» (linear process). Όταν οι μεταβλητές Z_t διανέμονται κανονικά, τότε έχουμε μηδενική συσχέτιση, δηλαδή ανεξαρτησία, πράγμα που ονομάζεται «Gaussian γραμμική μέθοδος» (Gaussian Linear Process) (6) (7).

Αυτοπαλίνδομος κινητός μέσος όρος (Autoregressive Moving Average-ARMA)

Το μικτό αποτέλεσμα της αυτοπαλινδρόμησης (AR) της τάξης p και του κινούμενου μέσου όρου (MA) της τάξης q, ορίζουν το μοντέλο $ARMA(p, q)$. Το μοντέλο αυτό θεωρείται ως το γενικό μοντέλο για την πρόβλεψη στάσιμης χρονοσειράς. Εάν B είναι τελεστής τέτοιο ώστε $x_{t-1} = Bx_t$, τότε το μοντέλο $ARMA$ μπορεί να γραφτεί ως (6):

$$\varphi(B)x_t = \theta(B)z_t \qquad (11)$$

όπου $\varphi(B)$, $\theta(B)$ είναι πολυώνυμα του B της τάξης p και q αντίστοιχα.

Ο τύπος αυτός αποτελεί μια στάσιμη λύση υπό την προϋπόθεση ότι οι ρίζες του $\varphi(x) = 0$ βρίσκονται εκτός του μοναδιαίου κύκλου. Η διαδικασία αυτή είναι αντιστρέψιμη με την προϋπόθεση ότι οι ρίζες του $\theta(x) = 0$ βρίσκονται εκτός του μοναδιαίου κύκλου.

Η σπουδαιότητα της $ARMA$ διαδικασίας είναι ότι πολλά σύνολα πραγματικών δεδομένων μπορεί να προσεγγιστούν με πιο ήπιο (parsimonious) τρόπο, δηλαδή με

λιγότερους παραμέτρους, με ένα μικτό μοντέλο $ARMA$ και όχι από την απλή διαδικασία AR ή MA. Κάθε σταθερή μέθοδος μπορεί να αναπαρασταθεί ως μοντέλο $MA(\infty)$, συμφώνα με το θεώρημα Wold και τον τύπο *(10)*, όμως αυτό μπορεί να συνεπάγεται με άπειρους παραμέτρους και έτσι δεν βοηθά καθόλου στην μοντελοποίηση. Το μοντέλο $ARMA$ μπορεί να θεωρηθεί ως ένα μοντέλο από μόνο του ή ως προσέγγιση της αναπαράστασης Wold, και σε αυτήν την περίπτωση, το πολυώνυμο B, που δίνετε από τον τύπο *(10)*, δίνει:

$$\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j \quad (12)$$

και μπορεί να φτάνει στο άπειρο, με σκοπό την προσέγγιση πολυώνυμου της μορφής:

$$\psi(B) = \frac{\theta(B)}{\varphi(B)} \quad (13)$$

που δίνει τελικά το μοντέλο ARMA.

Η βασική προϋπόθεση για την σωστή εκτίμηση των μοντέλων AR, MA και ARMA είναι πως η χρονοσειρά πρέπει να είναι στάσιμη και η αυτοσυσχέτιση (ACF) και η μερική αυτοσυσχέτιση (PACF) πρέπει να είναι χρονικά ανεξάρτητες (6) (7).

Ως αυτοσυσχέτιση ορίζεται η συσχέτιση μεταξύ δύο διαδοχικών τιμών της ίδιας χρονοσειράς (6). Πιο συγκεκριμένα ορίζεται ο συντελεστής αυτοσυσχέτισης και η μερική αυτοσυσχέτιση ως ακολούθως (5) (6) (7) (9) (10):

Ο **συντελεστής αυτοσυσχέτισης (autocorrelation ACF)** για υστέρηση h δείχνει τον βαθμό συσχέτισης μεταξύ των τιμών χρονοσειράς $\{Y_t\}$ και των τιμών της μετατοπισμένης κατά τάξη h σειράς $\{Y_{t-h}\}$ και ορίζεται ως:

$$ACF_h = Corr(Y_t, Y_{t-h}) = \frac{Cov(Y_t, Y_{t-h})}{\sqrt{Var(Y_t)}\sqrt{Var(Y_{t-h})}} \quad (14)$$

όπου $Cov(Y_t, Y_{t-h}) = E\big[\big(Y_t - \mu_Y(t)\big)\big(Y_{t-h} - \mu_Y(t-h)\big)\big]$ είναι η συνάρτηση συνδιακύμανσης της χρονοσειράς $\{Y_t\}$ με μέση τιμή $\mu_Y(t) = E(Y_t)$.

Η **μερική αυτοσυσχέτιση (partial autocorrelation PACF)** εκφράζει την συσχέτιση μεταξύ $\{Y_t\}$ και $\{Y_{t-h}\}$ που δεν υπολογίζεται για μικρότερες καθυστερήσεις από 1 έως $h-1$ και ορίζεται ως εξής:

$$PACF_h = corr(Y_t, Y_{t-h} | Y_{t-1}, Y_{t-2}, \cdots, Y_{t-h+1}) \quad (15)$$

Ολοκληρωμένος αυτοπαλίνδρομος κινητός μέσος όρος (Autoregressive Integrated Moving Average – ARIMA)

Υπάρχουν όμως και χρονοσειρές οι οποίες δεν είναι στάσιμες. Σε αυτήν την περίπτωση χρησιμοποιείται το **ολοκληρωμένο αυτοπαλινδρομούμενο μοντέλο κινούμενου μέσου** ή **ολοκληρωμένο μικτό μοντέλο** (**autoregressive integrated moving average model, ARIMA**), το οποίο αποτελεί επέκταση του μοντέλου $ARMA$ και εκφράζεται μέσω του μαθηματικού τύπου (6) (10):

$$\varphi_p(B)(1-B)^d x_t = \theta_q(B) z_t \Leftrightarrow$$
$$\underbrace{(1-\varphi_1 B-\cdots-\varphi_p B^p)}_{\text{AR(p)}}\underbrace{(1-B)^d x_t}_{\text{d διαφορές}} = \underbrace{(1-\theta_1 B-\cdots-\theta_q B^q)}_{\text{MA(q)}} z_t + c \quad (16)$$

όπου $\varphi_p(B)$ είναι ο μη-εποχικός τελεστής αυτοπαλινδρόμησης τάξης p, x_t είναι η τιμή της χρονοσειράς τη χρονική στιγμή t, $\theta_q(B)$ είναι ο μη-εποχικός τελεστής κινητού μέσου τάξης q και c σταθερά η οποία δίνεται από τη σχέση: $c = \mu(1-\varphi_1-\cdots-\varphi_p)$, όπου μ η μέση τιμή.

Το μοντέλο $ARIMA$, επειδή είναι επέκταση του μοντέλου $ARMA$, χρησιμοποιεί τον δείκτη p, που είναι η τάξη του αυτοπαλινδρούμενου μοντέλου AR, τον δείκτη q, που είναι η τάξη του κινούμενου μέσου όρου από το μοντέλο MA, και τον δείκτη d που είναι ο βαθμός της πρώτης διαφοράς που υφίσταται. Έτσι το μοντέλο $ARIMA$ γράφεται ως εξής:

$$ARIMA(p,d,q)$$

Η σταθερά c επηρεάζει αρκετά την μακροχρόνια πρόβλεψη που λαμβάνεται από αυτά τα υποδείγματα. Χαρακτηριστικά (10):

- Αν $c = 0$ $και$ $d = 0$, ο όρος της μακροχρόνιας πρόβλεψης τείνει στο 0.
- Αν $c = 0$ $και$ $d = 1$, ο όρος της μακροχρόνιας πρόβλεψης τείνει σε μη μηδενική σταθερά.
- Αν $c = 0$ $και$ $d = 2$, ο όρος της μακροχρόνιας πρόβλεψης ακολουθεί ευθεία γραμμή.

- Αν $c \neq 0$ *και* $d = 0$, ο όρος της μακροχρόνιας πρόβλεψης τείνει στη μέση τιμή των δεδομένων.
- Αν $c \neq 0$ *και* $d = 1$, ο όρος της μακροχρόνιας πρόβλεψης ακολουθεί ευθεία γραμμή.
- Αν $c \neq 0$ *και* $d = 2$, ο όρος της μακροχρόνιας πρόβλεψης παρουσιάζει τετραγωνική τάση.

Η τιμή του d επίσης έχει επίδραση στα διαστήματα πρόβλεψης – όσο υψηλότερη είναι η τιμή του d τόσο γρηγορότερα τα διαστήματα πρόβλεψης αυξάνουν σε μέγεθος. Για $d = 0$ η τυπική απόκλιση της μακροχρόνιας πρόβλεψης θα τείνει στην τυπική απόκλιση των ιστορικών δεδομένων, συνεπώς τα διαστήματα πρόβλεψης θα είναι ουσιαστικά τα ίδια (10).

Συχνά οι χρονοσειρές διαθέτουν μια εποχική συνιστώσα που επαναλαμβάνεται κάθε s παρατηρήσεις. Για μηνιαίες παρατηρήσεις το $s = 12$, για τριμηνιαίες το $s = 4$ και ούτω καθεξής. Προκειμένου να ασχοληθούν με την εποχικότητα, τα μοντέλα $ARIMA$ έχουν γενικευτεί σε μοντέλα $SARIMA$, τα οποία χρησιμοποιούνται όταν μια μη-στάσιμη χρονοσειρά περιέχει εποχικότητα (9) (6) (10). Έστω B^s τέτοιο ώστε $B^s X_t = X_{t-s}$. Έτσι η εποχική διαφορά μπορεί να γραφτεί ως $(X_t - X_{t-1}) = (1 - B^s)X_t$. Ένα μοντέλο $SARIMA$ με μη εποχικούς όρους της τάξης (p, d, q) και εποχικότητα της τάξης (P, D, Q) είναι συντομογραφία του $SARIMA(p, d, q) \times (P, D, Q)_s$ μοντέλου και γράφεται ως (9) (6) (10):

$$\varphi(B)\Phi(B^s)(1 - B)^d(1 - B^s)^D X_t = \theta(B)\Theta(B^s)Z_t \qquad (17)$$

Όπου Φ, Θ υποδηλώνουν πολυώνυμα B^s της τάξης P, Q αντίστοιχα.

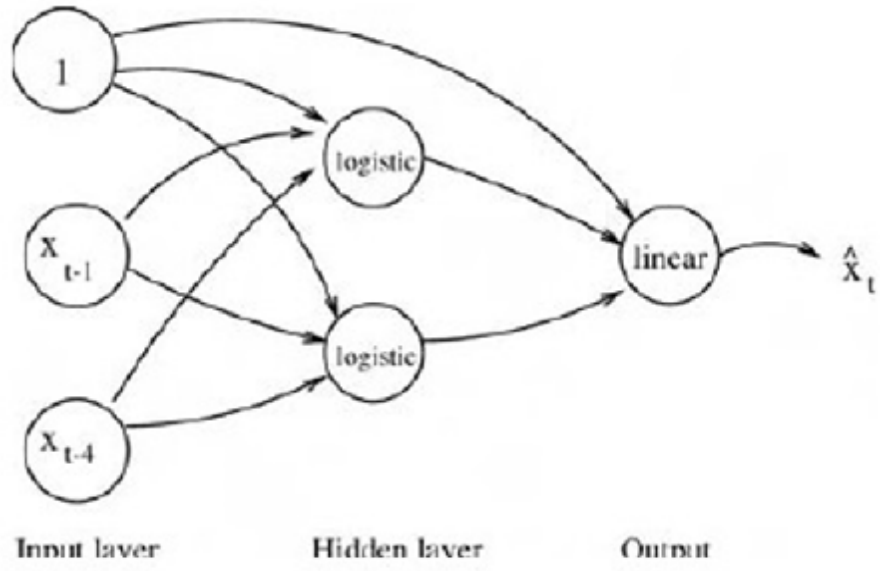
Το μοντέλο SARIMA γράφεται ως εξής (6) :

$$ARIMA(p, d. q) \times (P, D, Q)_s$$

Νευρωνικά Δίκτυα (Neural Networks - NN)

Τα νευρωνικά δίκτυα είναι ένας διαφορετικός τρόπος για την πρόβλεψη των μη-γραμμικών μοντέλων και έχουν εφαρμοστεί ήδη σε πάρα πολλά επιστημονικά πεδία αντιμετωπίζοντας με πολύ μεγάλη επιτυχία προβλήματα πρόβλεψης και βελτιστοποίησης.

Ένα νευρωνικό δίκτυο μπορεί να θεωρηθεί ως ένα σύστημα το οποίο συνδέει ένα σύνολο εισόδων και ένα σύνολο εξόδων με ένα μη-γραμμικό τρόπο. Στις χρονοσειρές η έξοδος μπορεί να θεωρηθεί ως μια τιμή της χρονοσειράς που θα προβλεφθεί και ως είσοδος μπορούν να θεωρηθούν οι τιμές της χρονοσειράς σε παρελθοντικές χρονικές στιγμές. Οι συνδέσεις μεταξύ των εισόδων και των εξόδων κατασκευάζονται μέσω ενός ή περισσότερων επίπεδων ή στρωμάτων από νευρώνες που καλούνται κρυφά επίπεδα ή στρώματα. Κύριο χαρακτηριστικό των Νευρωνικών Δικτύων είναι η αρχιτεκτονική τους, η οποία καθορίζει τον αριθμό των κρυφών επιπέδων, τον αριθμό των νευρώνων σε κάθε επίπεδο και τον τρόπο που συνδέονται οι είσοδοι και οι έξοδοι στα επίπεδα (6).



Εικόνα 12 Αρχιτεκτονική ενός νευρωνικού δικτύου για την πρόβλεψη χρονοσειράς

Η Εικόνα 12 απεικονίζει το συνηθισμένο τύπο τροφοδοσίας προς τα εμπρός καθώς δεν υπάρχουν βρόχοι ανάδρασης. Μια κατάλληλη αρχιτεκτονική για ένα συγκεκριμένο πρόβλημα πρέπει να καθορίζεται από τα συμφραζόμενα, ίσως με την βοήθεια εξωτερικών εκτιμήσεων ή με την χρήση ιδιοτήτων των δεδομένων. Πολλές

φορές η διαδικασία δοκιμής και λάθους είναι απαραίτητη, όπως για παράδειγμα η επιλογή του πλήθους των κρυφών νευρώνων. Έχει αποδειχθεί πως τα Νευρωνικά Δίκτυα Εμπρόσθιας Τροφοδότησης, όπως αυτό του παραπάνω σχήματος (Εικόνα 12), είναι πάρα πολύ αποτελεσματικά στην προσέγγιση συναρτήσεων καθώς και στην πρόβλεψη και είναι παρόμοια με ένα είδος μη-γραμμικού μοντέλου παλινδρόμησης (1) (2).

Στην Εικόνα 12 κάθε είσοδος συνδέεται με δύο νευρώνες, και οι νευρώνες συνδέονται με την έξοδο. Υπάρχει επίσης μία σύνδεση από την είσοδο προς την έξοδο. Η «δύναμη» της κάθε σύνδεσης μετράται με μία παράμετρο που ονομάζεται βάρος. Υπάρχει η πιθανότητα να υπάρχει ένας μεγάλος αριθμός των παραμέτρων για την εκτίμηση. Μια αριθμητική τιμή υπολογίζεται για κάθε νευρώνα σε χρονική περίοδο t ως εξής:

Έστω $y_{i,t}$ δηλώνει την τιμή της εισόδου i στο χρόνο t. Ας υποθέσουμε, όπως φαίνεται και στην Εικόνα 12, ότι οι τιμές των εισόδων είναι $y_{1,t} = 1$, είναι η είσοδος της πόλωσης κάθε νευρώνα, $y_{2,t} = x_{t-1}$ και $y_{3,t} = x_{t-4}$. Το W_{ij} υποδηλώνει το βάρος της σύνδεσης μεταξύ της εισόδου y_i και του νευρώνα j στο κρυφό επίπεδο. Αυτό παραμένει σταθερό με την πάροδο του χρόνου. Για κάθε νευρώνα υπολογίζεται ένα γραμμικό άθροισμα των εισόδων με τύπο $\sum W_{ij}\, y_{i,t} = v_{j,t}$, όπου $j = 1,2$. Στη συνέχεια επιλέγεται μια συνάρτηση που ονομάζεται συνάρτηση ενεργοποίησης, ώστε να γίνει η μετατροπή των τιμών v_j στην τελική τιμή για το νευρώνα. Η συνάρτηση αυτή είναι μη-γραμμική. Συχνά χρησιμοποιείται η λογιστική συνάρτηση $z = \frac{1}{(1+e^{-v})}$, η οποία δίνει τιμές εντός της περιοχής $(0,1)$. Στο συγκεκριμένο παράδειγμα, αυτό δίνει τις τιμές $z_{1,t}$ και $z_{2,t}$ για τους δύο νευρώνες σε κάθε χρονική στιγμή t. Μια τέτοια συνάρτηση μπορεί στη συνέχεια να εφαρμοστεί στις τιμές $z_{1,t}$, $z_{2,t}$ και στη συνεχή είσοδο, προκειμένου να πάρει την προβλεπόμενη έξοδο. Ωστόσο, η λογιστική συνάρτηση δεν πρέπει να χρησιμοποιείται στο στάδιο της εξόδου σε πρόβλεψη χρονοσειρών, εκτός αν τα δεδομένα βρίσκονται εντός της περιοχής $(0,1)$, διότι τότε οι προβλέψεις θα είναι λανθασμένες. Αντιθέτως, μια γραμμική συνάρτηση των τιμών των νευρώνων μπορεί να χρησιμοποιηθεί, η οποία συνεπάγεται με την συνάρτηση ενεργοποίησης στο στάδιο της εξόδου.

Η εισαγωγή μιας σταθερής μονάδας εισόδου, που συνδέεται σε κάθε νευρώνα στο κρυφό στρώμα, αλλά και στη έξοδο, αποφεύγοντας έτσι την ανάγκη της εισαγωγής μεμονωμένων τιμών στον υπολογιστή, ονομάζεται «κατώφλι» ή «πόλωση» (bias). Ουσιαστικά τα «κατώφλια» ενσωματώνονται στα υπόλοιπα βάρη, τα οποία μετρούν την δύναμη κάθε σύνδεσης από την είσοδο της μονάδας και έτσι γίνεται μέρος του συνόλου των βαρών, που είναι οι παράμετροι του μοντέλου, τα οποία μπορούν να αντιμετωπίζονται με τον ίδιο τρόπο.

Για ένα μοντέλο νευρωνικού δικτύου, με ένα κρυφό επίπεδο H νευρώνων, η γενική εξίσωση πρόβλεψης για τον υπολογισμό της πρόγνωσης του $x_t(έξοδος)$ χρησιμοποιώντας επιλεγμένες παρατηρήσεις του παρελθόντος, $x_{t-j_1}, \cdots, x_{t-j_k}$ $(είσοδος)$, μπορεί να γραφτεί ως (6):

$$\widehat{x_t} = \varphi_0 \left(W_{co} + \sum_{h=1}^{H} W_{ho}\, \varphi_h \left(W_{ch} + \sum_{i=1}^{k} W_{ih}\, x_{t-j_i} \right) \right) \qquad (18)$$

όπου W_{ch} δηλώνει τα βάρη για τις συνδέσεις μεταξύ της εισόδου και των κρυφών νευρώνων, $h = 1, \cdots, H$, και W_{co} δηλώνουν το βάρος της πόλωσης του κάθε νευρώνα εξόδου, τα βάρη W_{ih}και W_{ho}δηλώνουν τα βάρη για τις άλλες συνδέσεις μεταξύ των εισόδων και των κρυφών νευρώνων και μεταξύ κρυφών νευρώνων και εξόδων, αντίστοιχα. Οι δύο συναρτήσεις φ_h και φ_0 δηλώνουν τις συναρτήσεις ενεργοποίησης που χρησιμοποιούνται στους νευρώνες των κρυφών στρωμάτων και του στρώματος εξόδου αντίστοιχα.

Τα βάρη που χρησιμοποιούνται στα μοντέλα Νευρωνικών Δικτύων υπολογίζονται από τα δεδομένα εκπαίδευσης με την ελαχιστοποίηση του αθροίσματος των τετραγώνων των προβλέψεων ενός δείγματος, δηλαδή με τον τύπο $S = \sum_t (\hat{x}_{t-1}(1) - x_t)^2$. Τα Νευρωνικά Δίκτυα ορίζουν ένα σύνολο δεδομένων για εκπαίδευση (training set). Συνήθως γίνεται τυχαία επιλογή του 70% με 80% των εγγραφών του συνόλου δεδομένων που θα χρησιμοποιηθεί για την εκπαίδευση. Για τη δημιουργία του συνόλου δεδομένων για έλεγχο (testing set) θα χρησιμοποιηθούν οι υπόλοιπες 20% με 30% εγγραφές που δεν χρησιμοποιήθηκαν για εκπαίδευση. Στον έλεγχο θα γίνει η εκτίμηση της ικανότητας του εκπαιδευμένου Νευρωνικού Δικτύου να γενικεύει (γενίκευση), δηλαδή να δίνει αξιόπιστες προβλέψεις σε νέα δεδομένα με τα οποία δεν έχει εκπαιδευτεί (6).

Αρχικά το δίκτυο τροφοδοτείται με παραδείγματα εκπαίδευσης, τα οποία αποτελούνται από ένα σύνολο προτύπων εισόδου και με επιθυμητές εξόδους. Στην συνέχεια, για κάθε πρότυπο εκπαίδευσης, οι τιμές εισόδου σταθμίζονται και αθροίζονται σε κάθε νευρώνα του κρυμμένου στρώματος. Έπειτα, το σταθμισμένο άθροισμα αφού περάσει από τη συνάρτηση ενεργοποίησης του κρυφού νευρώνα, δίνει την κρυμμένη αξία εξόδου του νευρώνα, η οποία γίνεται η είσοδος στους νευρώνες του στρώματος εξόδου. Οι τιμές εξόδου υπολογίζονται και συγκρίνονται με τις επιθυμητές τιμές, έτσι ώστε να προσδιοριστεί το πόσο κοντά αντιστοιχούν οι πραγματικοί έξοδοι του δικτύου με τις επιθυμητές. Τέλος, τα βάρη ενημερώνονται και το δίκτυο μπορεί να έχει μια καλύτερη προσέγγιση για την επιθυμητή έξοδο. Αυτή η διαδικασία επαναλαμβάνεται πολλές φορές έως ότου οι διαφορές μεταξύ των τιμών εξόδου του δικτύου και τις γνωστές τιμές-στόχου να είναι όσο το δυνατόν μικρότερη (3).

Ο πιο δημοφιλής αλγόριθμος που χρησιμοποιείτε για τον υπολογισμό παραγώγων της αντικειμενικής συνάρτησης S, έτσι ώστε να μπορεί να ελαχιστοποιηθεί είναι ο αλγόριθμος που ονομάζεται **ανάστροφη διάδοση (back propagation)**. Η ανάστροφη διάδοση μαθαίνει από επαναληπτική επεξεργασία του συνόλου εκπαίδευσης, συγκρίνοντας την πρόβλεψη του δικτύου για κάθε πλειάδα με την πραγματική γνωστή τιμή-στόχο. Η τιμή-στόχος μπορεί να είναι η γνωστή ετικέτα τάξη των πλειάδων εκπαίδευσης (για προβλήματα ταξινόμησης) ή μια συνεχή τιμή (για την πρόβλεψη). Για κάθε πλειάδα εκπαίδευσης, τα βάρη τροποποιούνται έτσι ώστε να ελαχιστοποιείται το μέσο τετραγωνικό σφάλμα μεταξύ της πρόβλεψης του δικτύου και της πραγματικής τιμής-στόχου. Αυτές οι τροποποιήσεις έχουν γίνει στην "προς τα πίσω" κατεύθυνση, η οποία διασχίζει το Νευρωνικό Δίκτυο από το στρώμα εξόδου μέσα από κάθε νευρώνα του κρυμμένου στρώματος και φτάνοντας στους κόμβους εισόδου. Στην "προς τα πίσω" κατεύθυνση τα βάρη των συνάψεων ενημερώνονται. Αν και η διαδικασία δεν είναι εγγυημένη, τα βάρη τελικά θα συγκλίνουν και η μαθησιακή διαδικασία σταματά (1) (6).

2.4.4. Τμηματοποίηση χρονοσειρών (Segmentation)

Η τμηματοποίηση σε χρονοσειρές αναφέρεται ως ένας αλγόριθμος μειωμένων διαστάσεων. Παρά το γεγονός ότι τα τμήματα που δημιουργούνται θα μπορούσαν να

θεωρηθούν πολυώνυμα, η αναπαράσταση των τμημάτων είναι μια γραμμική συνάρτηση. Μια Τμηματική Γραμμική Παράσταση (Piecewise Linear Representation-PLR), αναφέρει την προσέγγιση μιας χρονοσειράς Q, μήκους n, με K ευθείες γραμμές όπως στην Εικόνα 13:

Εικόνα 13 Παράδειγμα τμηματοποίησης χρονοσειράς

Οι περισσότεροι αλγόριθμοι τμηματοποίησης χρονοσειρών μπορούν να ομαδοποιηθούν σε τρείς κατηγορίες, παρά την διαφορετική εφαρμογή τους:

Sliding-Windows (SW): ένα τμήμα αναπτύσσεται μέχρι να ξεπεραστεί κάποιο όριο σφάλματος. Η διαδικασία επαναλαμβάνεται με τα υπόλοιπα δεδομένα.

Top-Down (TD): η χρονοσειρά αναδρομικά διαμοιράζεται μέχρι την στιγμή που δεν πληρούνται τα κριτήρια.

Bottom-Up (BU): τα τμήματα συγχωνεύονται με την καλύτερη δυνατή προσέγγιση, μέχρι την στιγμή που δεν θα πληρούνται τα κριτήρια.

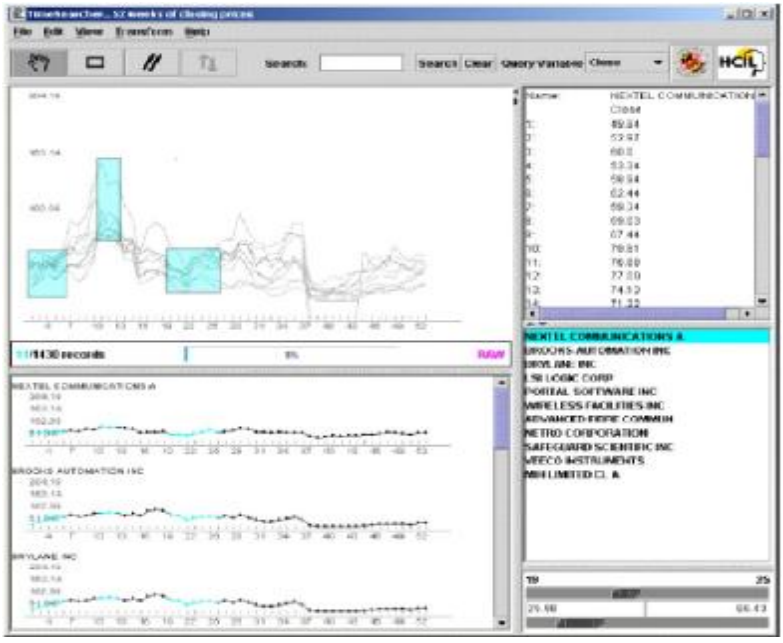
Η ποιότητα ενός αλγορίθμου τμηματοποίησης μπορεί να μετρηθεί με το σφάλμα ανασυγκρότησης για ένα σταθερό αριθμό τμημάτων. Το σφάλμα ανασυγκρότησης είναι η Ευκλείδεια απόσταση μεταξύ των αρχικών δεδομένων και της τμηματικής αναπαράστασης (segment representation) (3).

2.4.5. Σύνοψη χρονοσειρών (Summerization)

Επειδή πολλές φορές το μήκος των δεδομένων χρονοσειρών είναι πολύ μεγάλο, είναι απαραίτητο και πολύ χρήσιμο να γίνει μια στατιστική σύνοψη των δεδομένων. Επίσης χρησιμοποιείται η οπτικοποίηση, δηλαδή η γραφική παρουσίαση της

πληροφορίας για την εξαγωγή της ουσίας από τα δεδομένα. Ο εντοπισμός ανωμαλιών είναι μια ειδική περίπτωση, όπου τα ανώμαλα ή/και επαναλαμβανόμενα μοτίβα παρουσιάζουν ενδιαφέρον. Η σύνοψη μπορεί να θεωρηθεί ως ένας ειδικός τύπος ομαδοποίησης, όπου τα δεδομένα χωρίζονται σε υποσύνολα, με απλές περιγραφές, ώστε να παραχθεί ένα υψηλότερο επίπεδο στην προβολή των δεδομένων. Οι πιο γνωστές προσεγγίσεις για την απεικόνιση των συνόλων δεδομένων περιλαμβάνουν την Αναζήτηση σε χρόνο (Time Searcher), την Ημερολογιακή οπτικοποίηση (Calendar-Based Visualization), το Σπιράλ (Spiral) και το Δένδρο Viz (VizTree).

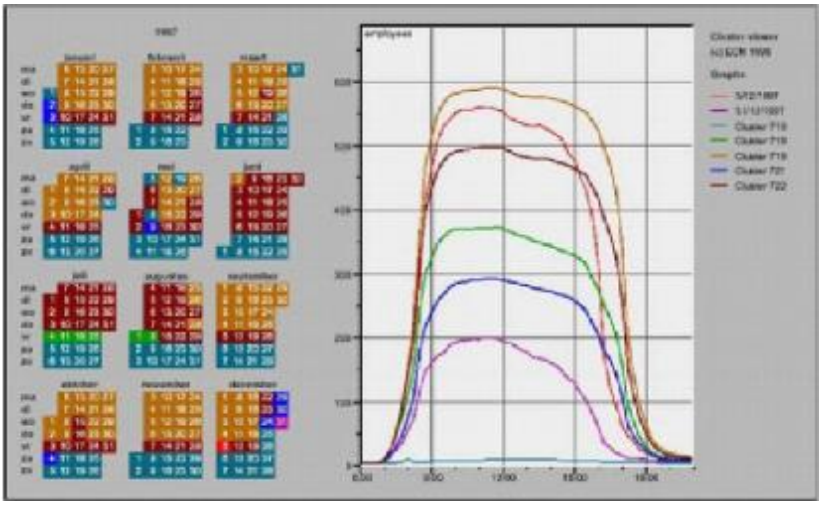
Αναζήτηση σε χρόνο (Time Searcher): είναι ένα διερευνητικό και οπτικό εργαλείο που επιτρέπει στον χρήστη την ανάκτηση χρονοσειρών με την δημιουργία ερωτημάτων, τα λεγόμενα Time Boxes. Η Εικόνα 14 δείχνει τρία Time Boxes στο στάδιο της επεξεργασίας, ώστε να καθοριστούν οι χρονικές σειρές. Η προσέγγιση αυτή της απεικόνισης των συνόλων δεδομένων προϋποθέτει γνώσεις σχετικά με τα σύνολα δεδομένων και οι χρήστες πρέπει να έχουν μια γενική ιδέα για το τι αναζητούν ή το τι τους ενδιαφέρει.



Εικόνα 14 Παράδειγμα προσέγγισης Αναζήτησης σε Χρόνο για ένα οπτικό ερώτημα

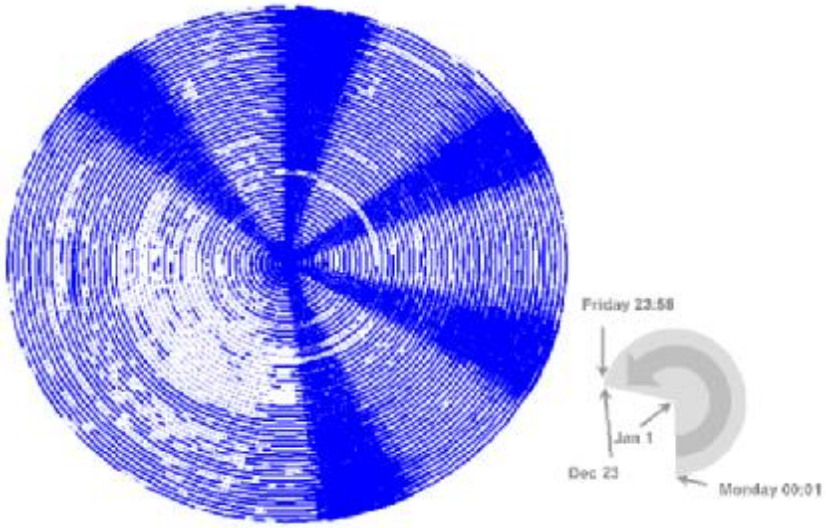
Ημερολογιακή οπτικοποίηση (Calendar-Based Visualization): είναι ένα σύστημα οπτικοποίησης, το οποίο συγκεντρώνει τις χρονοσειρές, χρησιμοποιώντας έναν bottom-up αλγόριθμο ομαδοποίησης. Το σύστημα εμφανίζει σχέδια που αντιπροσωπεύουν τον μέσο όρο συστάδας, μαζί με ένα ημερολόγιο στο οποίο η κάθε

μέρα έχει διαφορετικό χρωματικό κώδικα από την συστάδα στην οποία ανήκει. Η Εικόνα 15 δείχνει ένα παράδειγμα του συστήματος απεικόνισης, συγκεκριμένα παρουσιάζεται παράδειγμα συστάδας και του ημερολογίου με την οπτικοποίηση σε υπάλληλο που εργάζεται σε χρονικά δεδομένα και δείχνει έξι ομάδες, που αντιπροσωπεύουν διαφορετικό μοτίβο εργάσιμων ημερών Με την χρήση αυτής της προσέγγισης προκύπτουν απλοί κανόνες, όπως: «Κατά τους χειμερινούς μήνες η κατανάλωση ενέργειας είναι μεγαλύτερη από ότι τους καλοκαιρινούς μήνες».



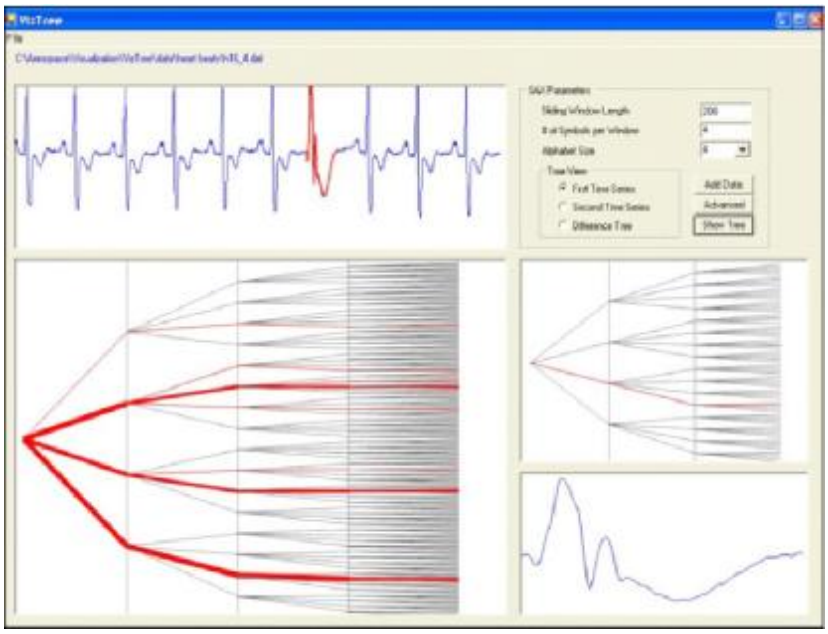
Εικόνα 15 Παράδειγμα ενός συστήματος απεικόνισης

Σπιράλ (Spiral): απεικονίζει κάθε περιοδικό τμήμα της χρονοσειράς σε ένα «δαχτυλίδι» (ring) και τα χαρακτηριστικά όπως το χρώμα και το πάχος της γραμμής που χρησιμοποιείται χαρακτηρίζει τις τιμές των δεδομένων. Η χρήση της προσέγγισης αυτής είναι η αναγνώριση των περιοδικών δομών στα δεδομένα. Η Εικόνα 16 εμφανίζει την ετήσια κατανάλωση ενέργειας. Η χρησιμότητα αυτού του εργαλείου είναι περιορισμένη για χρονοσειρές όταν η περίοδος είναι άγνωστη ή όταν δεν παρουσιάζουν περιοδικές συμπεριφορές.



Εικόνα 16 Η προσέγγιση οπτικοποίησης Σπιράλ

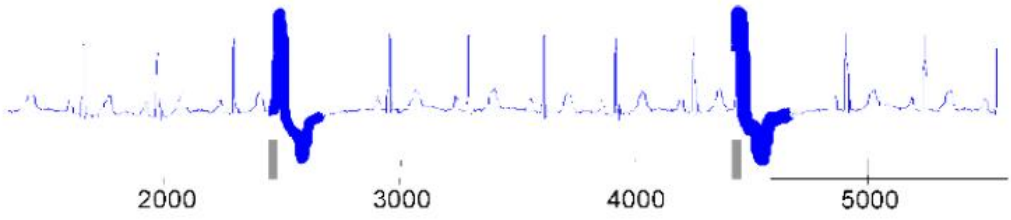
Δέντρο Viz (VizTree): είναι μια καινούργια τεχνική που έχει στόχο να ανακαλύψει τα άγνωστα σχέδια με ελάχιστη ή καμία γνώση σχετικά με τα δεδομένα. Παρέχει μια συνολική οπτική περίληψη και πιθανώς μπορεί να αποκαλύψει κρυμμένες δομές στα δεδομένα. Η προσέγγιση αυτή πρώτα μετασχηματίζει τις χρονοσειρές σε μια συμβολική αναπαράσταση, και κωδικοποιεί τα δεδομένα σε τροποποιημένο δέντρο, στο οποίο η συχνότητα και οι άλλες ιδιότητες των μοτίβων αντιστοιχίζονται σε χρώματα και άλλες οπτικές ιδιότητες. Η δομή του δέντρου χρειάζεται να έχει διακριτά δεδομένα, όμως επειδή οι αρχικές χρονοσειρές δεν είναι διακριτές, χρησιμοποιώντας μια διακριτοποίηση χρονοσειρών, τα συνεχή δεδομένα μπορούν να μετατραπούν σε διακριτά δεδομένα. Παρακάτω παρουσιάζεται ένα παράδειγμα προσέγγισης με το Δέντρο Viz στα δεδομένα με ανωμαλίες ενός ECG (electrocardiography – ηλεκτροκαρδιογράφημα) (3).



Εικόνα 17 **Παράδειγμα προσέγγισης με το Δέντρο** Viz

2.4.6. Ανίχνευση ανωμαλιών χρονοσειρών (Anomaly Detection)

Η ανίχνευση ανωμαλιών θεωρείται ο προσδιορισμός των άγνωστων προτύπων. Ως μια ανωμαλία εννοείται μια συμπεριφορά που αποκλίνει από την «κανονική». Το πρόβλημα της Ανίχνευσης ανωμαλιών σε χρονοσειρές είναι ότι περιλαμβάνει ενδιαφέροντα σχέδια, τα οποία δεν είναι απαραίτητα ανωμαλίες. Ο εντοπισμός των ανωμαλιών συνδέεται στενά με την Σύνοψη, η οποία αναφέρθηκε στην προηγούμενη ενότητα. Η Εικόνα 18 απεικονίζει την ιδέα ανίχνευσης ανωμαλιών (3).



Εικόνα 18 **Παράδειγμα Ανίχνευσης ανωμαλιών**

2.4.7. Ευρετηριοποίηση χρονοσειρών (Indexing)

Η αναζήτηση με βάση το περιεχόμενο σε βάσεις δεδομένων χρονοσειρών είναι πολύ σημαντική και περιλαμβάνει μια ταιριαστή ακολουθία που χωρίζεται σε δύο κατηγορίες, την ολόκληρη αντιστοίχηση και την αντιστοίχηση υποακολουθίας.

Ολόκληρη αντιστοίχηση (whole matching): μια ερώτηση της χρονοσειράς συγκρίνεται με μια βάση δεδομένων των επιμέρους χρονοσειρών για να εντοπίσει εκείνες με παρόμοιο ερώτημα.

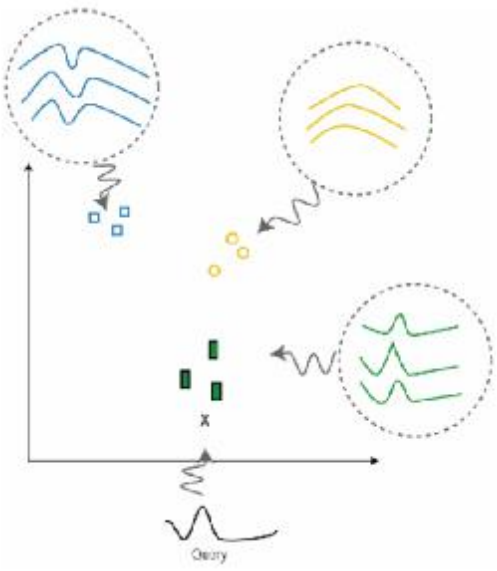
Αντιστοίχηση Υποακολουθίας (subsequence matching): ένα σύντομο ερώτημα υποακολουθίας χρονοσειράς συγκρίνεται με μεγαλύτερες χρονοσειρές σύροντας κατά μήκος την μακρύτερη ακολουθία για να βρεθεί η καλύτερη θέση που να ταιριάζει.

Η εφαρμογή της μεθόδου αυτής είναι περιορισμένη σε περιπτώσεις κατά τις οποίες ορισμένες πληροφορίες σχετικά με τα δεδομένα είναι γνωστά εκ των προτέρων. Η αντιστοίχηση υποακολουθίας μπορεί να γενικευτεί σε όλη την αντιστοίχηση με την διαίρεση ακολουθιών σε μη-επικαλυπτόμενα τμήματα είτε σε συγκεκριμένες περιόδους είτε από το σχήμα τους. Για παράδειγμα μπορεί να θέλουμε να εξάγουμε από το ηλεκτροκαρδιογράφημα μόνο τους μεμονωμένους χτύπους παλμών.

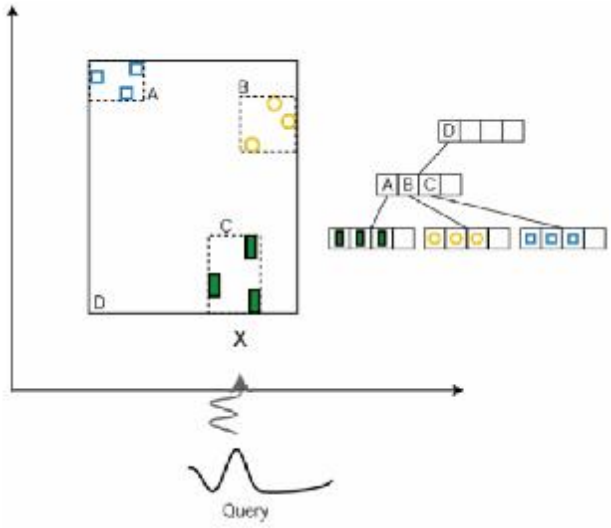
Λαμβάνοντας υπόψη μια βάση δεδομένων ακολουθιών, ο απλούστερος τρόπος για να βρεθεί το πιο κοντινό ταίριασμα σε μια δοσμένη ακολουθία Q, είναι να γίνει εκτέλεση μιας γραμμικής ή σειριακής σάρωσης των δεδομένων. Κάθε ακολουθία ανακτάται από το δίσκο και η απόσταση της προς την Q υπολογίζεται σύμφωνα με την προεπιλεγμένη απόσταση που έχει οριστεί. Το πιο κοντινό ταίριασμα επιστρέφεται στον χρήστη ως η πλησιέστερη αντιστοιχία. Η τεχνική γραμμικής σάρωσης όμως είναι πολύ δαπανηρή για να εφαρμοστεί, διότι απαιτεί πολλές προσβάσεις στο δίσκο και λειτουργεί από τις πρώτες σειρές, η οποία μπορεί αν είναι αρκετά μεγάλη. Μια καλύτερη εναλλακτική για την γραμμική σάρωση, θα ήταν να αποθηκεύει σε δύο επίπεδα προσέγγισης δεδομένων, τα πρωτογενή δεδομένα και η συμπιεσμένη έκδοση τους. Έτσι η γραμμική σάρωση εκτελείται επί των συμπιεσμένων ακολουθιών και το κάτω όριο για την αρχική απόσταση υπολογίζεται για όλες τις ακολουθίες. Η μικρότερη απόσταση από την Q ενημερώνεται μετά από κάθε σειρά που ανακτάται. Η αναζήτηση μπορεί να τερματιστεί όταν το κατώτερο όριο που εξετάζεται υπερβαίνει την ελάχιστη απόσταση που βρέθηκε μέχρι τώρα.

Ένας πιο αποδοτικός τρόπος για να εκτελεστεί η Ευρετηριοποίηση είναι η χρησιμοποίηση μιας δομής ευρετηρίου που θα συγκεντρώνει παρόμοιες ακολουθίες στην ίδια ομάδα, με αποτέλεσμα να παρέχει ταχύτερη πρόσβαση στις πιο υποσχόμενες ακολουθίες. Χρησιμοποιώντας διάφορες τεχνικές κλαδέματος και δομές ευρετηριοποίησης αποφεύγετε η εξέταση μεγάλων τμημάτων του συνόλου δεδομένων, ενώ εξακολουθεί να διασφαλίζετε ότι τα αποτελέσματα θα είναι τα ίδια με τα αποτελέσματα της γραμμικής σάρωσης. Οι δομές ευρετηριοποίησης μπορούν να χωριστούν σε δύο κατηγορίες, διανυσματική βάση και μετρική βάση.

Διανυσματική βάση δομής ευρετηριοποίησης (vector based indexing structures): οι διανυσματικοί δείκτες λειτουργούν στα συμπιεσμένα δεδομένα. Οι αρχικές ακολουθίες συμπιέζονται χρησιμοποιώντας μια μέθοδο μείωσης διαστάσεων, και τα πολυδιάστατα διανύσματα που προκύπτουν μπορούν να ομαδοποιηθούν σε παρόμοιες συστάδες, χρησιμοποιώντας διανυσματική βάση δομής ευρετηριοποίησης, όπως φαίνετε στην Εικόνα 19. Η δομή αυτή μπορεί να χωριστεί σε δύο κατηγορίες: την ιεραρχική (hierarchical) ή την μη-ιεραρχική (non-hierarchical). Η πιο κοινή ιεραρχική διανυσματική βάση ευρετηριοποίησης είναι το R-δέντρο (R-tree) ή κάποια παραλλαγή του. Το R-δέντρο αποτελείται από πολυδιάστατα διανύσματα για τα επίπεδα των φύλλων, τα οποία οργανώνονται στο δέντρο χρησιμοποιώντας ορθογώνια που μπορεί να επικαλύπτονται, όπως φαίνεται στην Εικόνα 20.



Εικόνα 19 Παράδειγμα μείωσης διαστάσεων των χρονοσειρών σε δύο διαστάσεις



Εικόνα 20 **Παράδειγμα ιεραρχικής οργάνωσης χρησιμοποιώντας ένα** R-tree

Μετρική βάση δομής ευρετηριοποίησης (metric based indexing structures): τυπικά η τεχνική αυτή αποδίδει καλύτερα από την διανυσματική βάση, και προτείνεται οι διαστάσεις να είναι περισσότερες από 5. Εδώ τα αντικείμενα δεν συγκεντρώνονται με βάση τα συμπιεσμένα χαρακτηριστικά τους, αλλά με τις σχετικές αποστάσεις του αντικειμένου. Η επιλογή αντικειμένου αναφοράς από το οποίο θα υπολογιστούν όλες οι αποστάσεις αντικειμένων, μπορεί να ποικίλει σε διαφορετικές προσεγγίσεις. Παραδείγματα των μετρικών δέντρων (metric trees) περιλαμβάνουν το Vantage Point Tree, M-tree και GNAT. Όλες οι παραλλαγές αυτών των δέντρων, εκμεταλλεύονται τις αποστάσεις των σημείων αναφοράς σε συνδυασμό με την τριγωνική ανισότητα για να κλαδευτούν τα μέρη του δέντρου, όπου το ταίριασμα δεν είναι κοντά με αυτά που έχουν ήδη ανακαλυφθεί (3).

ΚΕΦΑΛΑΙΟ 3ο - ΧΡΗΣΗ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΟΥ ΠΑΚΕΤΟΥ R ΣΤΗΝ ΑΝΑΛΥΣΗ ΧΡΟΝΟΣΕΙΡΩΝ

3.1. Λόγοι χρήσης της R

Η R (http://www.r-project.org/) υπάρχει τις δύο τελευταίες δεκαετίες και είναι μια συνεχώς αυξανόμενη βιβλιοθήκη εξειδικευμένων δεδομένων οπτικοποίησης, ανάλυσης και διαχείρισης πακέτων δεδομένων. Με περίπου δύο εκατομμύρια χρήστες, η R κατέχει μία από τις μεγαλύτερες βιβλιοθήκες με στατιστικούς αλγόριθμους και με στατιστικά πακέτα. Εξελίχθηκε από μια απλή στατιστική γλώσσα της πληροφορικής σε ένα πλήρες αναλυτικό περιβάλλον.

Η R χαρακτηρίζεται ως πλήρες αναλυτικό περιβάλλον επειδή τα δεδομένα από διάφορες βάσεις δεδομένων μπορούν να χρησιμοποιηθούν άμεσα. Επιπλέον θεωρείται ως η μεγαλύτερη και ταχύτερη αναπτυσσόμενη στατιστική βιβλιοθήκη ανοιχτού κώδικα, διότι ο τρέχων αριθμός των στατιστικών πακέτων και ο ρυθμός ανάπτυξης κατά την οποία συνεχίζουν τα νέα πακέτα να αναβαθμίζονται, εξασφαλίζει την συνέχεια της R ως μια μακροπρόθεσμη λύση για τα αναλυτικά προβλήματα. Τέλος παρέχει ένα ευρύ φάσμα των λύσεων από τις βιβλιοθήκες των πακέτων που μπορούν να χρησιμοποιηθούν στην στατιστική, στην εξόρυξη και στην οπτικοποίηση των δεδομένων, στις εφαρμογές στο διαδίκτυο και σε άλλες εφαρμογές.

Ο πηγαίος κώδικας της R έχει σχεδιαστεί για να εξασφαλίζει τις λύσεις του προβλήματος και να ενσωματώνει μια συγκεκριμένη εφαρμογή. Ο ανοιχτός πηγαίος κώδικας έχει το πλεονέκτημα ότι είναι ευρέως αξιολογημένος από επιστημονικές βιβλιογραφίες. Αυτό σημαίνει ότι τα σφάλματα θα βρεθούν, οι πληροφορίες θα κοινοποιηθούν και τα προβλήματα θα λυθούν εύκολα. Για την αναλυτική πλατφόρμα R υπάρχει μεγάλος αριθμός εκπαιδευτικού υλικού με τη μορφή βιβλίων που είναι διαθέσιμος και στο διαδίκτυο. Επίσης προσφέρει τα καλύτερα εργαλεία οπτικοποίησης δεδομένων στο λογισμικό ανάλυσης. Ο κύριος λόγος για τον οποίο το λογισμικό τρίτων (third-party) ξεκίνησε να δημιουργεί διασυνδέσεις με την R είναι επειδή η γραφική βιβλιοθήκη των πακέτων στην R είναι πιο προηγμένη και συνεχίζει να αποκτάει περισσότερες δυνατότητες μέχρι και σήμερα. Επιπλέον η πλατφόρμα

είναι δωρεάν για όποιον θέλει να την αποκτήσει μέσω διαδικτύου, και για μικρές αλλά και για μεγάλες αναλυτικές ομάδες. Η R προσφέρει ευέλικτο προγραμματισμό για το περιβάλλον με τα δεδομένα του χρήστη. Περιλαμβάνει τα πακέτα που εξασφαλίζουν τη συμβατότητα με τη Java, Python και της C++. Τέλος είναι για κάποιον εύκολο να αλλάξει από άλλη αναλυτική πλατφόρμα και να μεταβεί στην πλατφόρμα της R .

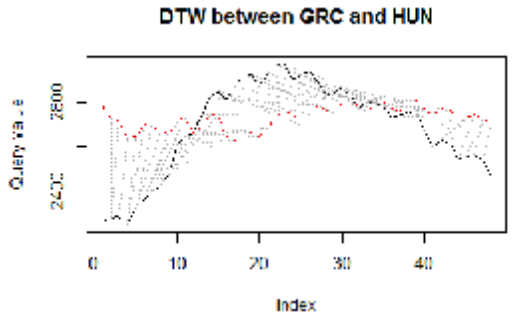
Ως υπολογιστική πλατφόρμα, η R είναι πολύ καλά προσαρμοσμένη στις ανάγκες της εξόρυξης δεδομένων. Παρέχει μεγάλη σειρά από πακέτα που καλύπτουν τα πρότυπα παλινδρόμησης, τα δέντρα αποφάσεων, τους κανόνες συσχέτισης, την συσταδοποίηση, την μηχανική μάθηση, τα νευρωνικά δίκτυα, την κατηγοριοποίηση, την πρόβλεψη και πολλά ακόμα (11).

3.2. Εφαρμογή της συσταδοποίησης

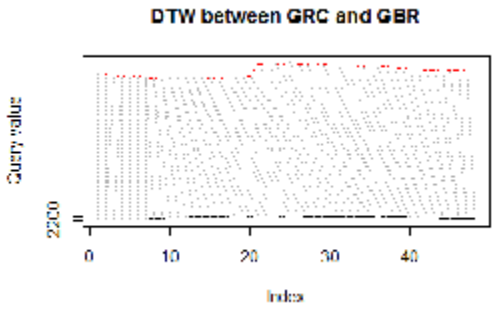
Στην παρούσα πτυχιακή εργασία θα χρησιμοποιήσουμε βάσεις δεδομένων οι οποίες προέρχεται από την ιστοσελίδα www.quandl.com (12) και περιγράφουν τους απασχολούμενους (σε χιλιάδες) 22 χωρών της Ευρωπαϊκής Ένωσης για τα έτη 2002 έως 2013. Οι εγγραφές είναι ανά τρίμηνο, η πρώτη εγγραφή έγινε το Φεβρουάριο, η δεύτερη το Μάιο, στην συνέχεια τον Αύγουστο και τέλος τον Νοέμβριο, πράγμα που σημαίνει ότι το σύνολο των εγγραφών για κάθε χώρα είναι 48. Οι χώρες που θα χρησιμοποιήσουμε στη συσταδοποίηση είναι η Ελλάδα (grc), Βουλγαρία (bgr), Αυστρία (aut), Βέλγιο (bel), Κροατία (hrv), Δανία (dnk), Εσθονία (est), Φινλανδία (fin), Ουγγαρία (hun) Ιρλανδία (irl), Ιταλία (ita), Λιθουανία (ltu), Μολδαβία (mda), Ολλανδία (nld), Πολωνία (pol), Πορτογαλία (prt), Ρουμανία (rou), Σλοβενία (svn), Σλοβακία (svk), Ισπανία (esp), Σουηδία (swe) και το Ηνωμένο Βασίλειο (gbr).

Σε αυτό το σημείο πρέπει να σημειωθεί, πως στα δεδομένα η πρώτη εγγραφή, δηλαδή η πρώτη σειρά, ξεκινά το 2013-12-31 και η τελευταία εγγραφή τελειώνει το 2002-03-31. Στόχος αυτής της προσπάθειας είναι να δούμε πόσο καλά μπορούν να ομαδοποιηθούν οι χώρες και με ποιο τρόπο, θα χρησιμοποιηθεί ο αλγόριθμος k-μέσων και η ιεραρχική συσταδοποίηση προσπαθώντας πάντα να βρεθεί το βέλτιστο αποτέλεσμα, αλλά και το μέτρο ομοιότητας της μετρικής Dynamic Time Warping για να δούμε πόσο όμοια είναι η Ελλάδα με τις υπόλοιπες χώρες.

Αρχικά με χρήση της βιβλιοθήκης ***Quandl*** δημιουργήσαμε ένα πίνακα που περιέχει όλες τις χώρες και όλες τις εγγραφές και αποτελείται από 22 στήλες, δηλαδή τις χώρες και 48 εγγραφές, δηλαδή το κάθε τρίμηνο για τα έτη 2002-2013. Έπειτα χρησιμοποιήσαμε τη μετρική απόστασης Dynamic Time Warping (DTW) που υλοποιείται με την βιβλιοθήκη ***dtw*** ώστε να δούμε την ομοιότητα της χρονοσειράς της Ελλάδας με τις υπόλοιπες χώρες τις Ευρωπαϊκής Ένωσης. Με την εντολή ***dtw()*** δημιουργήθηκε ο πίνακας για τον υπολογισμό της βέλτιστης απόστασης. Παρακάτω δίνονται τα διαγράμματα που προέκυψαν από την σύγκριση Ελλάδας-Ουγγαρίας και Ελλάδας-Ηνωμένο Βασίλειο. Επιλέχθηκαν αυτά τα διαγράμματα επειδή είναι τελείως αντίθετα μεταξύ τους και μπορούμε να δούμε και τις δύο περιπτώσεις.



Διάγραμμα 1 Σύγκριση Ελλάδα - Ουγγαρία



Διάγραμμα 2 Σύγκριση Ελλάδα – Ηνωμένο Βασίλειο

Η Ελλάδα είναι η χρονοσειρά που απεικονίζεται με το μαύρο χρώμα και όλες οι υπόλοιπες χώρες με το κόκκινο. Το Διάγραμμα 1 είναι πολύ ενδιαφέρον καθώς στην αρχή των μετρήσεων φαίνεται πως υπάρχει κάποια διαφορά μεταξύ τους, η Ελλάδα βρίσκεται στους 2.265.000 εργαζομένους ενώ η Ουγγαρία στους 2.790.000. Στην συνέχεια όμως για κάποια χρονική στιγμή οι χρονοσειρές είναι σχεδόν ίδιες αφού αυξάνεται κατά πολύ ο αριθμός εργαζομένων στην Ελλάδα. Τέλος βλέπουμε πως

αποκτάται ξανά μία απόσταση μεταξύ τους καθώς οι εργαζόμενοι στην Ελλάδα μειώνονται ξανά.

Στο Διάγραμμα 2 η διαφορά είναι εμφανής, δεν υπάρχει καμία ομοιότητα ή κανένα σημείο που να συμπίπτουν οι δύο χρονοσειρές. Αυτό βέβαια μπορεί να επιβεβαιωθεί από τον πίνακα df.xwres (ΠΑΡΑΡΤΗΜΑ Ι) με τους εργαζόμενους των χωρών, καθώς οι απασχολούμενοι της Ελλάδας κυμαίνονται στους 2.265.000 εργαζομένους, ενώ του Ηνωμένου Βασιλείου στους 25.545.000 εργαζομένους, η διαφορά είναι πολύ μεγάλη.

Έπειτα χρησιμοποιήθηκε ο αλγόριθμος k-μέσων για τη διαμεριστική συσταδοποίηση των εγγραφών. Χρησιμοποιήθηκε η βιβλιοθήκη ***cluster*** και η συνάρτηση ***kmeans()*** του προγράμματος R, η οποία χώρισε τα δεδομένα σε τέσσερις ομάδες. Ως αποτέλεσμα προκύπτει ο Πίνακας 2 που δείχνει πόσα στοιχεία έχουν ενταχθεί σε κάθε ομάδα.

1η συστάδα	2η συστάδα	3η συστάδα	4η συστάδα
13	15	12	8

Πίνακας 2 Χωρισμός δεδομένων σε ομάδες με τον αλγόριθμο k-μέσων

Από την συσταδοποίηση όμως μπορούμε να αποσπάσουμε επιπλέον πληροφορίες, όπως για τον τρόπο με τον οποίο έχουν χωριστεί τα δεδομένα με βάση τα έτη που βρίσκονται:

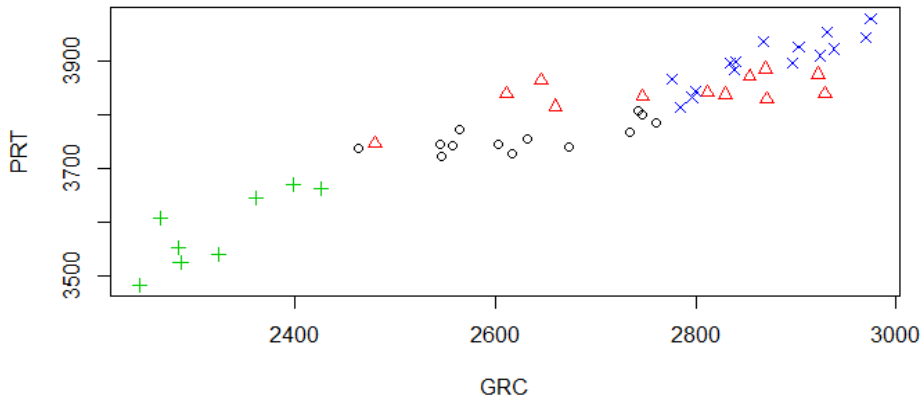
Clustering vector:

[1] 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4

[43] 4 4 4 4 4 4

Σύμφωνα με τα αποτελέσματα της συσταδοποίησης προκύπτει ότι τα έτη 2013 και 2012 έχουν ενταχθεί στην 2η ομάδα, έτη 2011, 2010 και 2009 βρίσκονται στην ομάδα 1η και τα έτη 2008, 2007, 2006, 2005 έχουν ενταχθεί στην 3η ομάδα εκτός από τον 3ο μήνα του 2005 που έχει ενταχθεί στην ομάδα 4. Τα υπόλοιπα έτη, δηλαδή 2004, 2003 και 2002 είναι ενταγμένες στην 4η ομάδα. Από τα παραπάνω μπορούμε να συμπεράνουμε πως από το 2002 μέχρι το 2011, ανά 3 ή 4 έτη υπήρχε σημαντική αλλαγή στον αριθμό των απασχολούμενων, επειδή αλλάζουν οι ομάδες.

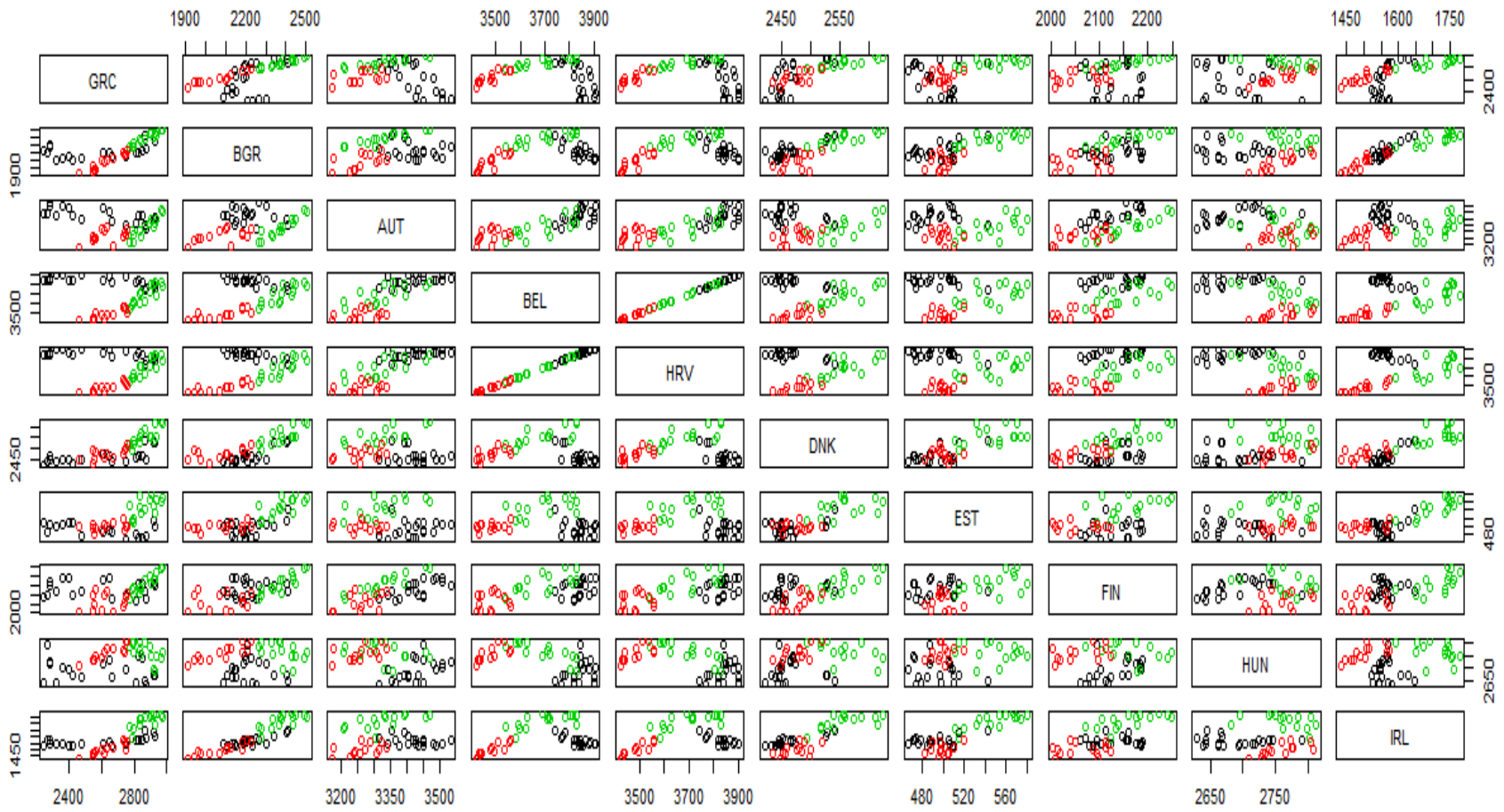
Έπειτα δημιουργήσαμε ένα διάγραμμα συσταδοποίησης της Ελλάδας και της Πορτογαλίας. Επιλέχθηκε η Πορτογαλία διότι είναι η πιο κοντινή χώρα σε πληθυσμό με την Ελλάδα.



Διάγραμμα 3 Συσταδοποιήση Ελλάδα - Πορτογαλία

Στο Διάγραμμα 3 η πράσινη ομάδα είναι η $4^{η}$ συστάδα, καθώς έχει 8 στοιχεία. Αντίστοιχα, η λευκή ομάδα είναι η $1^{η}$ συστάδα, η κόκκινη ομάδα είναι η $3^{η}$ ομάδα και η μπλε ομάδα είναι η $2^{η}$ συστάδα. Παρατηρούμε πώς η συσταδοποίηση δεν είναι ακριβώς αυτή που θα θέλαμε, για παράδειγμα η κόκκινη ομάδα απλώνεται πολύ στο διάγραμμα και τα σημεία της είναι πολύ αραιά ενώ θα μπορούσαν να ενταχθούν σε μίαν άλλη κοντινή ομάδα. Αυτό όμως συμβαίνει γιατί ο αλγόριθμος k-μέσων δεν είναι εξειδικευμένος σε ομαδοποίηση δεδομένων με τέτοια μορφή, βλέπουμε πως τα δεδομένα διατάσσονται σαν μία ευθεία γραμμή, λόγω ότι το πρώτο βήμα του αλγόριθμου αυτού είναι ο υπολογισμός του κέντρου βάρους, γίνεται πιο αποτελεσματικός όταν τα δεδομένα έχουν σφαιρική μορφή.

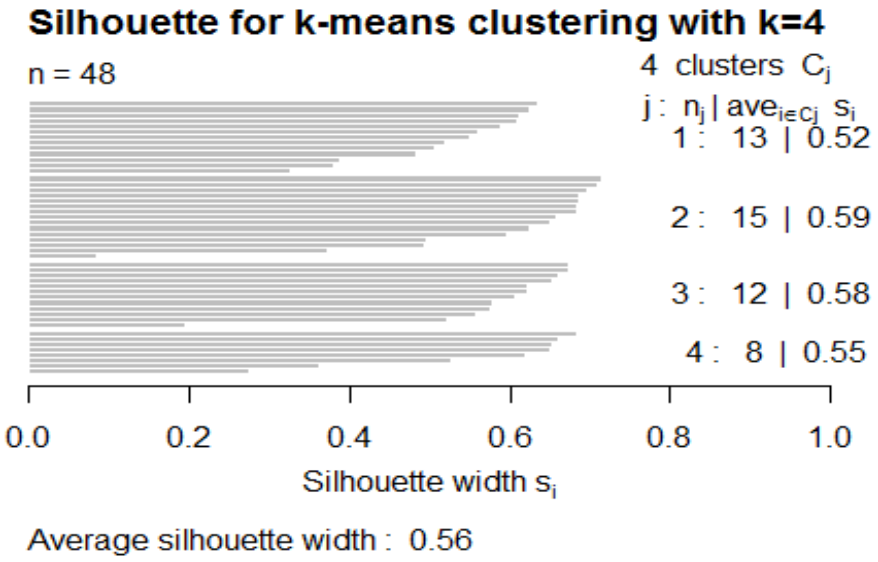
Παρακάτω εμφανίζεται διάγραμμα συσταδοποίησης με 10 χώρες της Ευρωπαϊκής Ένωσης:



Διάγραμμα 4 Συσταδοποίηση με 10 χώρες της Ευρωπαϊκής Ένωσης

Μία σχέση που αξίζει να αναφερθεί είναι η συσταδοποίηση μεταξύ του Βελγίου (BEL) και της Κροατίας (HRV), που τα δεδομένα είναι κατανεμημένα σε μία ευθεία γραμμή.

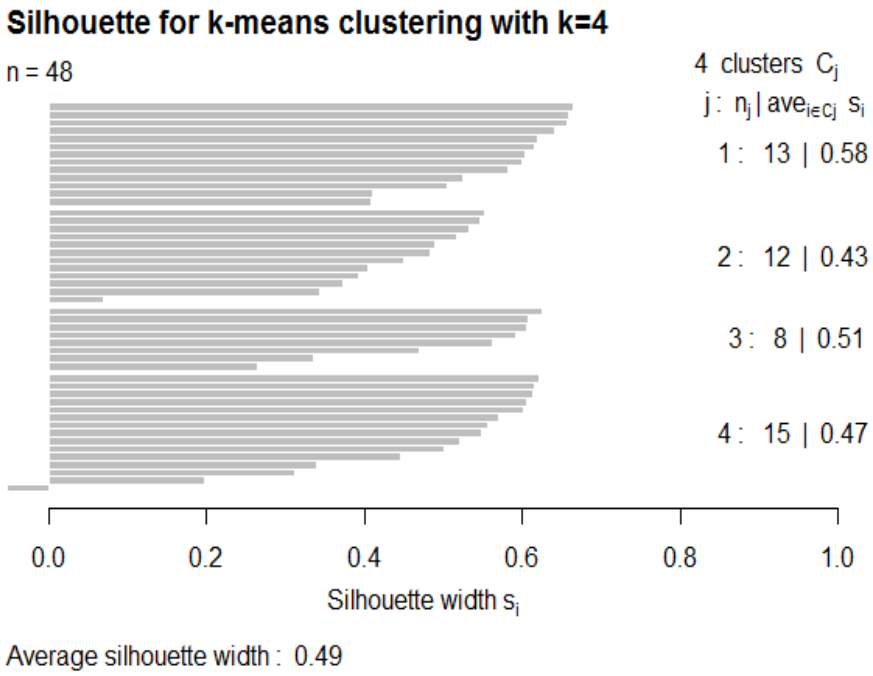
Στη συνέχεια με την συνάρτηση ***silhouette***() και παραμέτρους το αποτέλεσμα της ομαδοποίησης με τον k-μέσων και τη μήτρα απόστασης που υπολογίστηκε με τη μέθοδο της ευκλείδειας απόστασης δημιουργήθηκε, με τη συνάρτηση ***plot***(), το διάγραμμα του silhouette (συντελεστή περιγράμματος) για την εκτίμηση της συσταδοποίησης:



Διάγραμμα 5 Συντελεστής περιγράμματος με την Ευκλείδεια απόσταση

Το Διάγραμμα 5 δείχνει που έχει ενταχθεί κάθε στοιχείο και τον συντελεστή περιγράμματος για κάθε ομάδα. Ο καλύτερος συντελεστή περιγράμματος είναι στις ομάδες 2 και 3 που έχουν 0,59 και 0,58 αντίστοιχα. Η τιμή του μέσου συντελεστή περιγράμματος είναι 0,56 που σημαίνει η συσταδοποίηση που έχει γίνει είναι σχετικά σωστή αφού είναι τουλάχιστον πάνω από 0,50 αν και το ιδανικό είναι ο συντελεστής να είναι όσο πιο κοντά στο 1.

Στη συνέχεια υλοποιήθηκε η ίδια διαδικασία όπως και παραπάνω, όμως για τον υπολογισμό της μήτρας απόστασης χρησιμοποιήθηκε η απόσταση City-block (Manhattan). Έτσι ο συντελεστής περιγράμματος προκύπτει ως εξής:



Διάγραμμα 6 Συντελεστής περιγράμματος με την απόσταση Manhattan

Η διαφορά είναι εμφανή αφού ο συντελεστής περιγράμματος είναι 0,49 αρκετά χαμηλός και επίσης παρατηρείται πως μία από τα τις 15 εγγραφές της 4ης συστάδας είναι κάτω από το μηδέν, που σημαίνει ότι έχει ενταχθεί σε λάθος ομάδα.

	Ευκλείδεια	City-block (Manhattan)
Silouette 1ης ομάδας	0.52	0.58
Silouette 2ης ομάδας	0.59	0.43
Silouette 3ης ομάδας	0.58	0.51
Silouette 4ης ομάδας	0.55	0.47
Μέσος Συντελεστής περιγράμματος (silhouette)	**0,56**	**0,49**

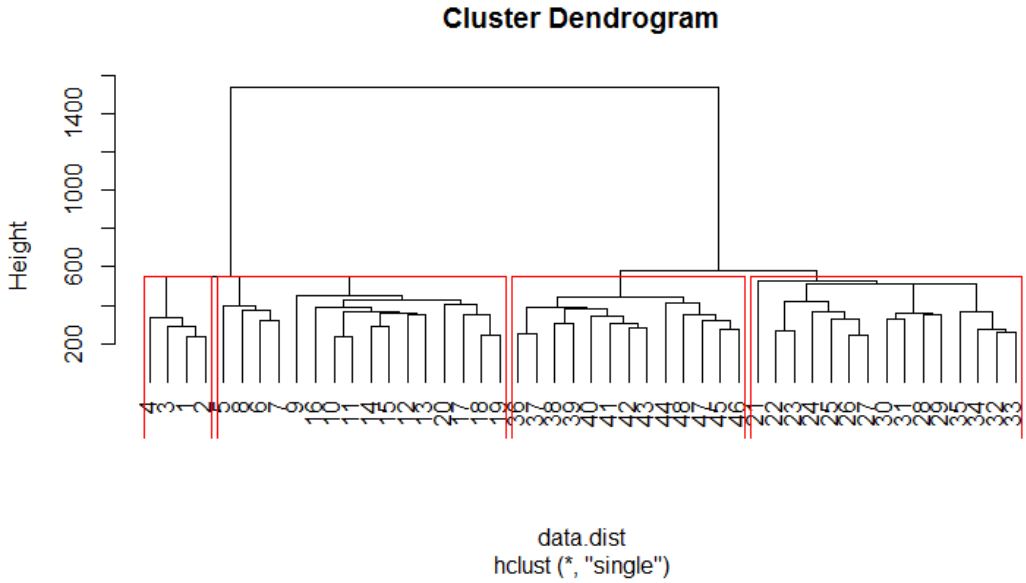
Πίνακας 3 Σύγκριση της Ευκλείδειας απόστασης με την City-block

Σύμφωνα με τα παραπάνω αποτελέσματα η Ευκλείδεια απόσταση έδωσε καλύτερη συσταδοποίηση από την απόσταση City block. Στην Ευκλείδεια απόσταση

ο συντελεστής περιγράμματος των ομάδων είναι πάντα αρκετά καλός, στην City-block υπάρχουν ομάδες όπως η ομάδα 2 και η ομάδα 4 με αρκετά χαμηλό συντελεστή.

Σε αυτό το σημείο θα εκτελεστεί ομαδοποίηση των δεδομένων με ιεραρχική συσταδοποίηση χρησιμοποιώντας τρείς διαφορετικούς μεθόδους. Σε όλες τις μεθόδους ο ιεραρχικός αλγόριθμος ομαδοποιεί τα δεδομένα σε 4 ομάδες όπου στα φύλλα του δένδρου είναι η ένδειξη της αντίστοιχης εγγραφής του συνόλου δεδομένων.

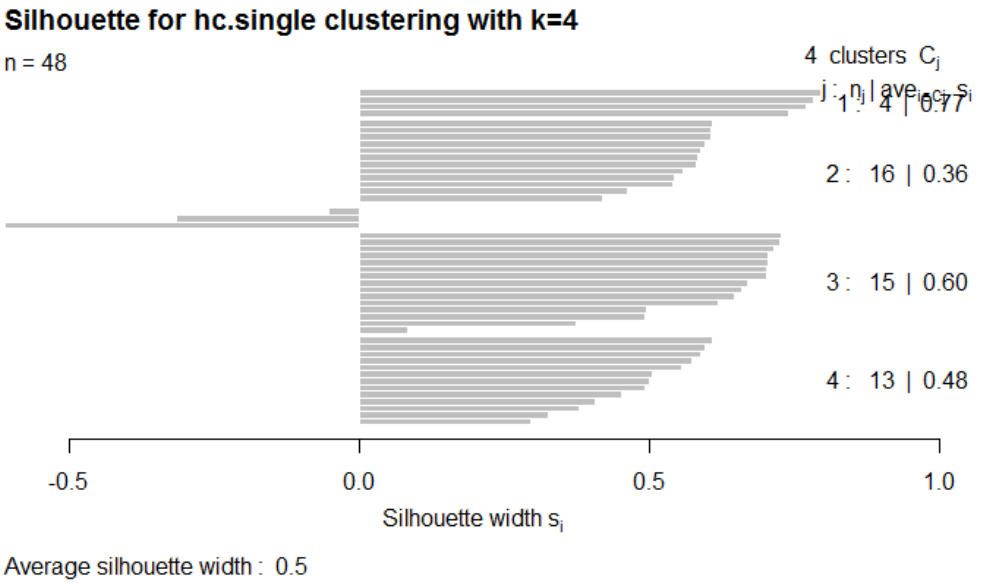
Με την συνάρτηση *hclust()* και παραμέτρους τη μήτρα απόστασης των δεδομένων με ευκλείδεια απόσταση και τη μέθοδο ιεραρχικής συσταδοποίησης απλού συνδέσμου (single), προκύπτει το παρακάτω δενδρόγραμμα.



Διάγραμμα 7 Δενδρόγραμμα ιεραρχικής συσταδοποίησης απλού συνδέσμου

Από το δενδρόγραμμα παρατηρούμε πως στην 1η συστάδα εμφανίζονται τα 4 τρίμηνα του 2013 (εγγραφές 1 έως 4), στη 2η ομάδα οι εγγραφές των ετών 2012, 2011, 2010, 2009 (εγγραφές 5 έως 20), στην 3η ομάδα οι εγγραφές των ετών 2008, 2007, 2006 και το τελευταίο τρίμηνο της χρονιάς 2005 (εγγραφές 36 έως 48), και στην 4η ομάδα οι εγγραφές των τριών τρίμηνων του 2005 του 2004, 2003 και 2002 (εγγραφές 21 έως 35).

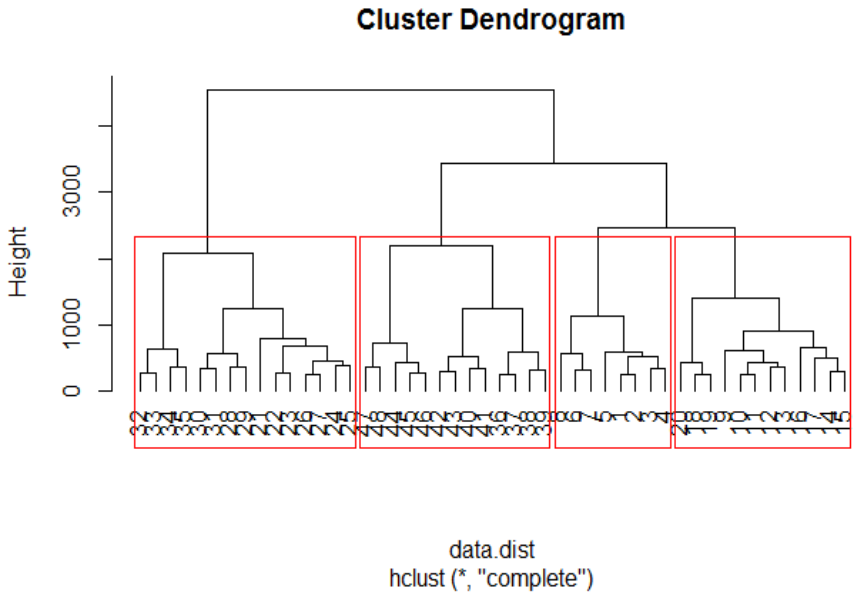
Έπειτα για την εκτίμηση της ιεραρχικής συσταδοποίησης χρησιμοποιήθηκε η συνάρτηση *silhouette()* .



Διάγραμμα 8 Εκτίμηση της ιεραρχικής συσταδοποίησης απλού συνδέσμου

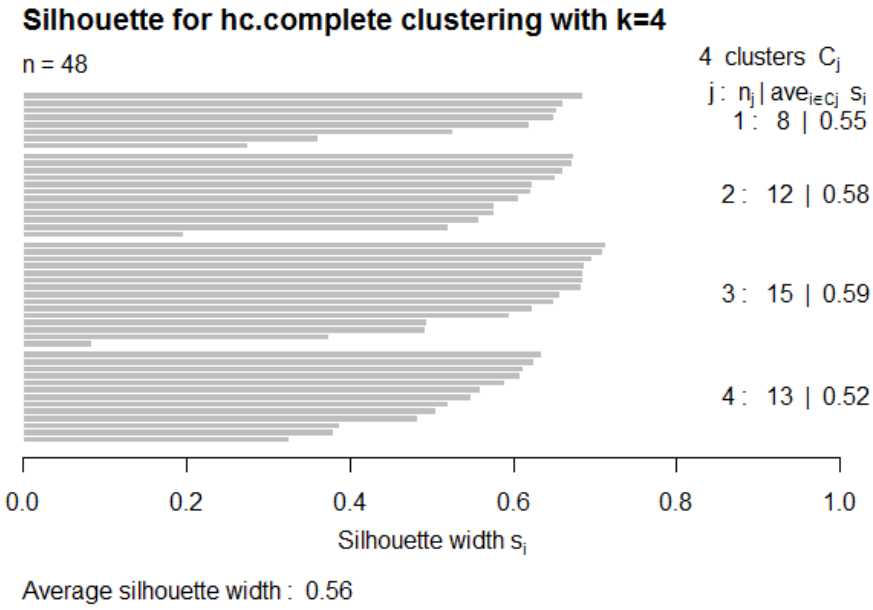
Η συσταδοποίηση είναι μέτρια καθώς ο συντελεστής είναι 0,50. Υπάρχουν διαφορές μεταξύ των ομάδων, καθώς η $1^{η}$ ομάδα είναι πολύ σωστά ομαδοποιημένη με συντελεστή περιγράμματος 0,77, αντίθετα η ομάδα 2 δεν είναι τόσο επιτυχημένη αφού ο συντελεστής είναι 0,36, αρκετά χαμηλός, και 3 εγγραφές έχουν ομαδοποιηθεί λάθος.

Η ιεραρχική συσταδοποίηση με τη μέθοδο πλήρους συνδέσμου (complete) δίνει το παρακάτω διάγραμμα:



Διάγραμμα 9 Δενδρόγραμμα ιεραρχικής συσταδοποίησης πλήρους συνδέσμου

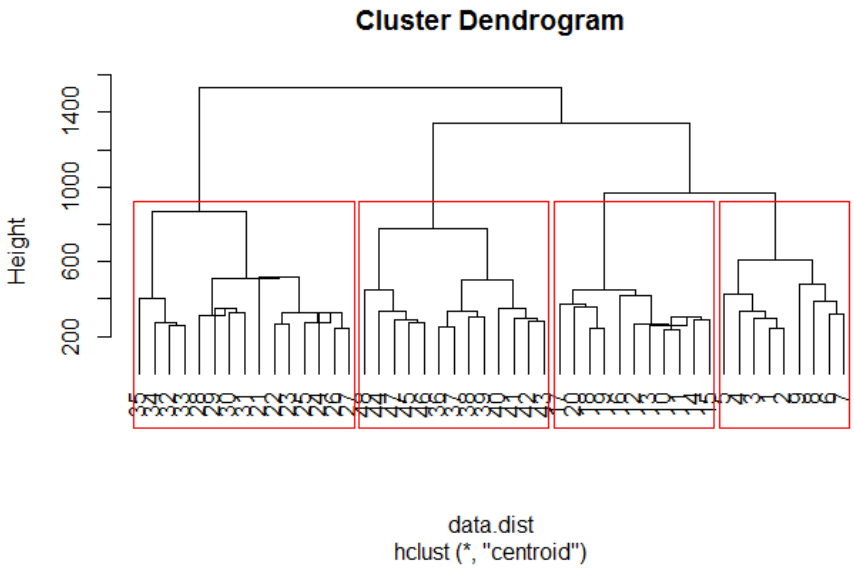
Από το δενδρόγραμμα παρατηρούμε πως στην 1η συστάδα εμφανίζονται οι εγγραφές 21 έως 35 στα έτη 2013, 2012, 2011και στα 3 τρίμηνα του έτους 2010. Στη 2η ομάδα εμφανίζονται οι εγγραφές 36 έως 48 οι οποίες αντιστοιχούν στο πρώτο τρίμηνο του 2010 και σε όλα τα τρίμηνα των ετών 2009 και 2008 και 2007, στην 3η ομάδα βρίσκονται οι εγγραφές 1 έως 8 που αντιστοιχούν στα έτη 2006 και 2005 και στην τελευταία ομάδα οι εγγραφές 9 έως 20 στα έτη 2004, 2003 και 2002.



Διάγραμμα 10 **Εκτίμηση της ιεραρχικής συσταδοποίησης πλήρους συνδέσμου**

Βλέπουμε πως η ιεραρχική συσταδοποίηση πλήρους συνδέσμου είναι πιο αποτελεσματική από την ιεραρχική συσταδοποίηση απλού συνδέσμου για το λόγο ότι ο συντελεστής περιγράμματος για το σύνολο της ομαδοποίησης είναι 0,56 αρκετά καλός, και όλες οι εγγραφές έχουν ενταχθεί στην σωστή ομάδα.

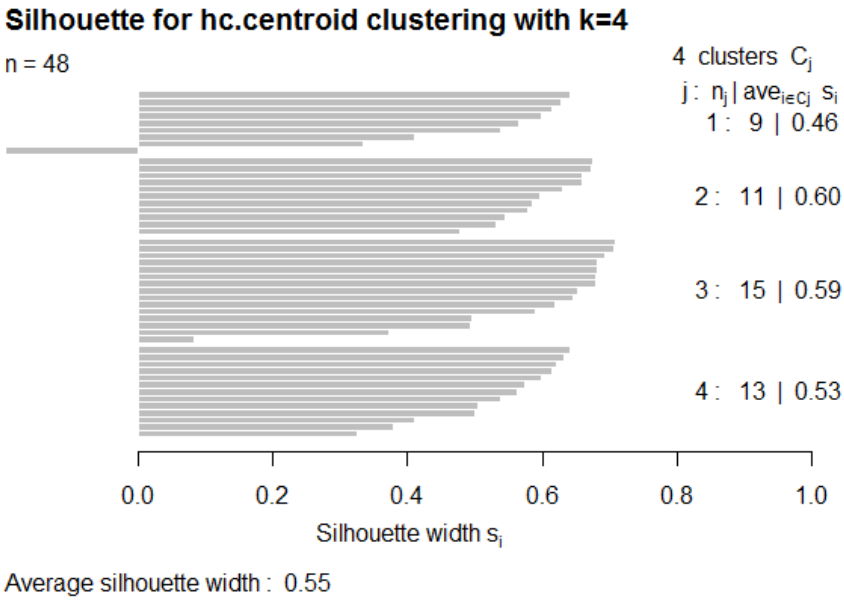
Ιεραρχική συσταδοποίηση με τη μέθοδο κέντρων βάρους(centroid) δίνει το παρακάτω δενδρόγραμμα:



Διάγραμμα 11 Δενδρόγραμμα ιεραρχικής συσταδοποίησης κέντρων βάρους

Παρατηρούμε ότι η στην 1η ομάδα έχουν ενταχθεί ο εγγραφές 21 έως 35 στα έτη 2013,2012 ,2011 και τα 3 τρίμηνα του 2010 , στην 2η ομάδα βρίσκονται οι εγγραφές 36 έως 48 στο πρώτο τριμήνου του 2010 και στα έτη 2009,2008,2007 , στην 3η ομάδα βρίσκονται οι εγγραφές 10 έως 20 στα έτη 2006,2005 και στα 3 τρίμηνα του έτους 2004. Τέλος στην 4η ομάδα είναι ενταγμένες οι εγγραφές 1 έως 9 στο πρώτου τριμήνου του 2004 και στα έτη 2003 και 2002.

Εκτίμηση της ιεραρχικής συσταδοποίησης με τη μέθοδο κέντρων βάρους.



Διάγραμμα 12 Εκτίμηση της ιεραρχικής συσταδοποίησης με μέθοδο κέντρων βάρους

Ο συντελεστής περιγράμματος είναι 0,55 έχει ελάχιστη διαφορά με την μέθοδο πλήρους συνδέσμου, μόνο μία εγγραφή έχει ενταχθεί σε λάθος ομάδα.

Ο

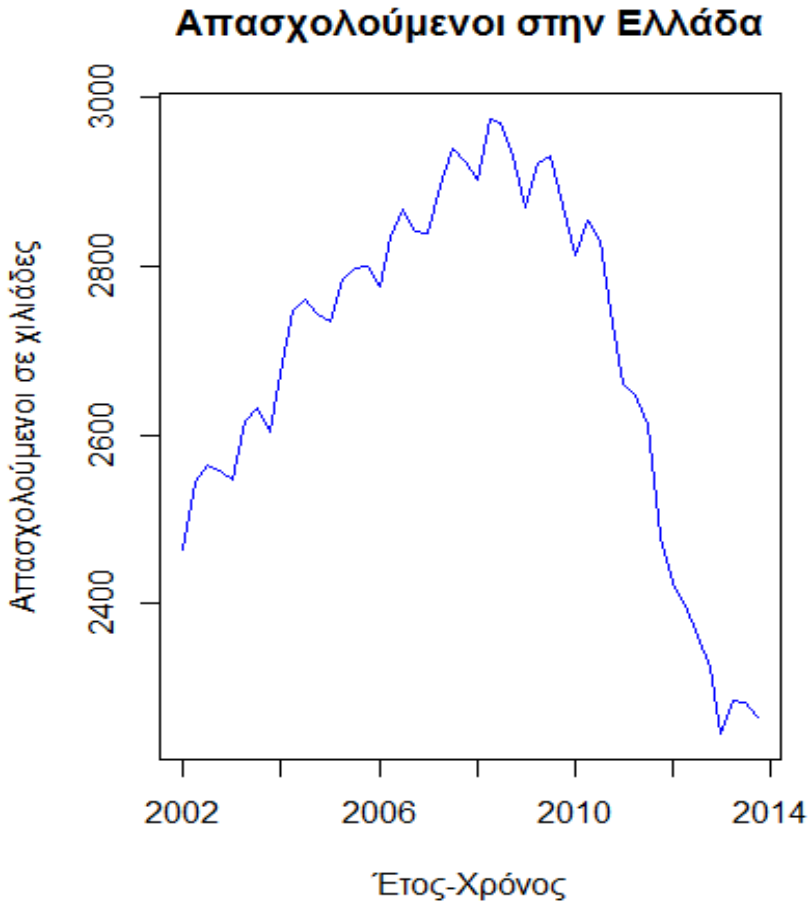
Πίνακας 4 δείχνει την σύγκριση των μεθόδων που χρησιμοποιήθηκαν με βάση τον συντελεστή περιγράμματος. Η ιεραρχική συσταδοποίηση πλήρους συνδέσμου ήταν πιο σωστή και αποτελεσματική από τις άλλες δύο μεθόδους.

Ιεραρχική συσταδοποίηση			
	Απλού συνδέσμου	**Πλήρους συνδέσμου**	**Κέντρων βάρους**
Silhouette 1ης ομάδας	0.77	0.55	0.46
Silhouette 2ης ομάδας	0.36	0.58	0.60
Silhouette 3ης ομάδας	0.60	0.59	0.59
Silhouette 4ης ομάδας	0.48	0.52	0.53
Μέσος Συντελεστής περιγράμματος (silhouette)	**0.50**	**0.56**	**0.55**

Πίνακας 4 Σύγκριση μεθόδων ιεραρχικής συσταδοποίησης

3.3. Εφαρμογή της πρόβλεψης σε χρονοσειρά

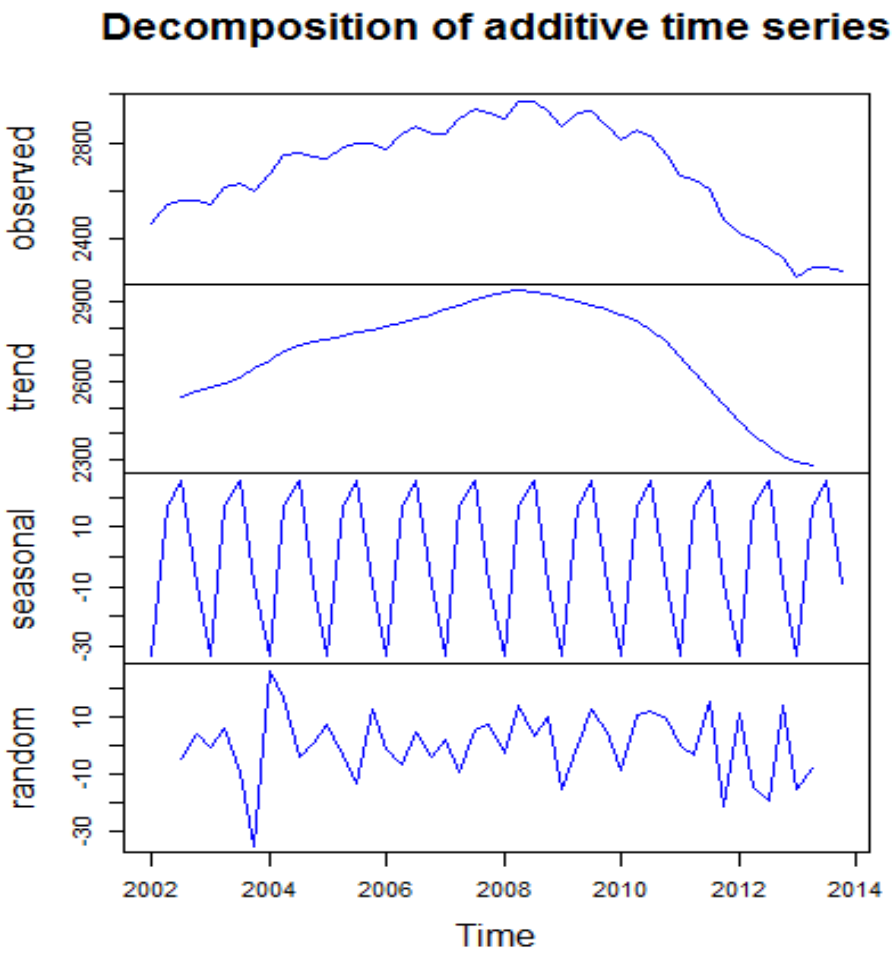
Για την πρόβλεψη θα χρησιμοποιήσουμε μόνο την βάση δεδομένων που αναφέρεται στους απασχολούμενους στην Ελλάδα για τα έτη 2002 έως 2013, ανά τρίμηνο. Έτσι θα έχουμε 48 εγγραφές. Αρχικά ορίζουμε την χρονοσειρά, η οποία ξεκινάει από το έτος 2002 από το πρώτο μήνα του τρίμηνου, χρησιμοποιώντας την συνάρτηση *ts()* και δημιουργούμε το διάγραμμα της χρονοσειράς.



Διάγραμμα 13 Απεικόνιση χρονοσειράς για τα έτη 2002-2013

Βλέπουμε από το Διάγραμμα 13 πως από το έτος 2002 μέχρι περίπου το έτος 2009 οι απασχολούμενοι στην Ελλάδα αυξαίνονται συνεχώς, ξεπερνούν τις 2900 χιλιάδες. Από το 2009 μέχρι το 2012 έχουμε ραγδαία μείωση των εργαζομένων στην χώρα μας, οι οποίοι το 2012 φτάνουν τις 2300 χιλιάδες.

Στην συνέχεια με την χρήση της συνάρτησης ***decompose*** δημιουργούμε το διάγραμμα της εσωτερικής δομής της χρονοσειράς ώστε να δούμε αν είναι στάσιμη ή μη-στάσιμη.



Διάγραμμα 14 Εσωτερική δομή χρονοσειράς

Από το Διάγραμμα 14 βλέπουμε πως η χρονοσειρά μας παρουσιάζει τάση (trend), διότι μέχρι το έτος 2009 έχει σταθερά ανοδική πορεία, σε αντίθεση με τα έτη από το 2009 έως 2012, όπου έχουμε έντονη καθοδική πορεία. Επίσης παρατηρούμε ότι παρουσιάζει και εποχικότητα (seasonal), επειδή τον πρώτο μήνα (Φεβρουάριος) και τον τελευταίο (Νοέμβριος) που έγινε η καταγραφή έχουμε αρνητική εποχικότητα, σε αντίθεση με τους μήνες Μάιος και Αύγουστος, όπου έχουμε θετική εποχικότητα, πράγμα που σημαίνει πως τους καλοκαιρινούς μήνες η απασχόληση στην Ελλάδα αυξάνεται σε σχέση με τους χειμερινούς μήνες. Επειδή η χρονοσειρά μας παρουσιάζει τάση και εποχικότητα θεωρείτε μη-στάσιμη χρονοσειρά.

3.3.1. Πρόβλεψη με μοντέλο SARIMA

Έστω ότι θέλουμε να προβλέψουμε πόσοι απασχολούμενοι θα είναι στην Ελλάδα το έτος 2013, για κάθε τρίμηνο. Θα δημιουργήσουμε καινούργια χρονοσειρά με τα ίδια δεδομένα, όμως θα αφαιρέσουμε τις 4 τελευταίες εγγραφές που αναφέρονται για το έτος 2013, επειδή όταν θα κάνουμε την πρόβλεψη για το έτος 2013 θα την συγκρίνουμε με το πραγματικό αριθμό. Με τον τρόπο αυτό θα έχουμε 44 εγγραφές.

Επειδή η χρονοσειρά είναι μη-στάσιμη και περιέχει εποχικότητα, το κατάλληλο μοντέλο για την πρόβλεψη θα είναι το μοντέλο $SARIMA(p,d,q) \times (P,D,Q)_s$. Χρησιμοποιώντας την βιβλιοθήκη ***forecast*** καλούμε το μοντέλο ***SARIMA*** και ύστερα από πειράματα στα οποία αλλάζαμε το p, d, q, P, D, Q βρίσκουμε την κατάλληλη τάξη, δηλαδή αυτή που ελαχιστοποιεί τον κανόνα του Akaike (AIC).

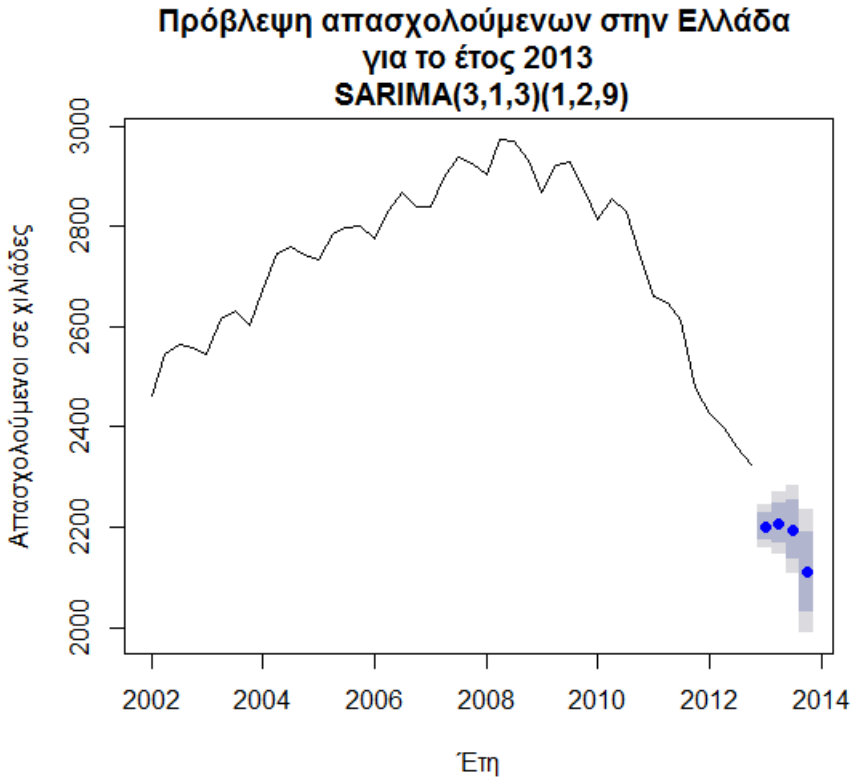
Ο κανόνας Akaike Information Criterion (AIC) δείχνει το μοντέλο που ελαχιστοποιεί:

$$AIC = -2\log(maximum\ likelihood) + 2k$$

όπου $k = p + q + 1$, εάν το μοντέλο περιέχει ένα σημείο τομής ή σταθερό όρο, και $k = p + q$ όταν περιέχει κάτι διαφορετικό. Η προσθήκη του όρου $2(p + q + 1)$ ή $2(p + q)$ χρησιμεύει ως συνάρτηση ποινής (penalty function) ώστε να αποφεύγονται μοντέλα με πάρα πολλές παραμέτρους (13).

Προκύπτει ότι η καλύτερη τάξη του μοντέλου ***SARIMA*** για να γίνει η πρόβλεψη χρονοσειρών είναι το μοντέλο $SARIMA(3,1,3) \times (1,2,9)$ το οποίο έχει AIC=372,8

Με την βιβλιοθήκη ***forecast*** και το μοντέλο ***SARIMA*** που βρήκαμε πριν προβλέπουμε την απασχόληση στην Ελλάδα για τα τρίμηνα του έτους 2013 και δημιουργούμε το διάγραμμα της πρόβλεψης.



Διάγραμμα 15 Πρόβλεψη για το έτος 2013 με την μέθοδο SARIMA

Από το Διάγραμμα 15 φαίνεται ότι οι απασχολούμενοι στην Ελλάδα συνεχίζονται να μειώνονται, εκτός από ένα μικρό διάστημα ανάμεσα στο 1ο και 2ο τρίμηνο του 2013, όπου για ελάχιστη τιμή οι απασχολούμενοι αυξάνονται. Το ίδιο μπορούμε να διαπιστώσουμε από τον Πίνακας 5 ο οποίος δείχνει πόσοι απασχολούμενοι θα είναι στην χώρα μας για τα τέσσερα τρίμηνα του 2013 σύμφωνα με την πρόβλεψη, όπως επίσης και την διαφορά της πραγματικής τιμής από την προβλεπόμενη:

Έτος	Τρίμηνο	Προβλεπόμενη Τιμή	Πραγματική Τιμή	Διαφορά
2013	**Q1**	2201,326	2245,3	43,974
2013	**Q2**	2207,436	2285,7	78,264
2013	**Q3**	2194,972	2283,5	88,528
2013	**Q4**	2111,171	2265,8	154,629

Πίνακας 5 Προβλεπόμενη τιμή απασχολούμενων με την μέθοδο SARIMA

Στον Πίνακας 6 εμφανίζονται οι τιμές πρόβλεψης με βάσει το μοντέλο SARIMA καθώς και τα διαστήματα εμπιστοσύνης με επίπεδο σημαντικότητας 80% και 95% αντίστοιχα.

		Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2013	**Q1**	2201.326	2173.729	2228.922	2159.120	2243.531
2013	**Q2**	2207.436	2167.347	2247.524	2146.126	2268.746
2013	**Q3**	2194.972	2136.639	2253.305	2105.760	2284.184
2013	**Q4**	2111.171	2030.867	2191.476	1988.356	2233.986

Πίνακας 6 Προβλεπόμενη τιμή και Διαστήματα Εμπιστοσύνης πρόβλεψης

Χρησιμοποιώντας την εντολή *accuracy()* εμφανίζονται έξι μέτρα για την ακρίβεια πρόβλεψης. Μερικά από αυτά είναι το Μέσο Σφάλμα (Mean Error – ME), η Ρίζα του Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error - RMSE) και το Μέσο Απόλυτο Ποσοστό Σφάλματος (Mean Absolute Percentage Error – MAPE).

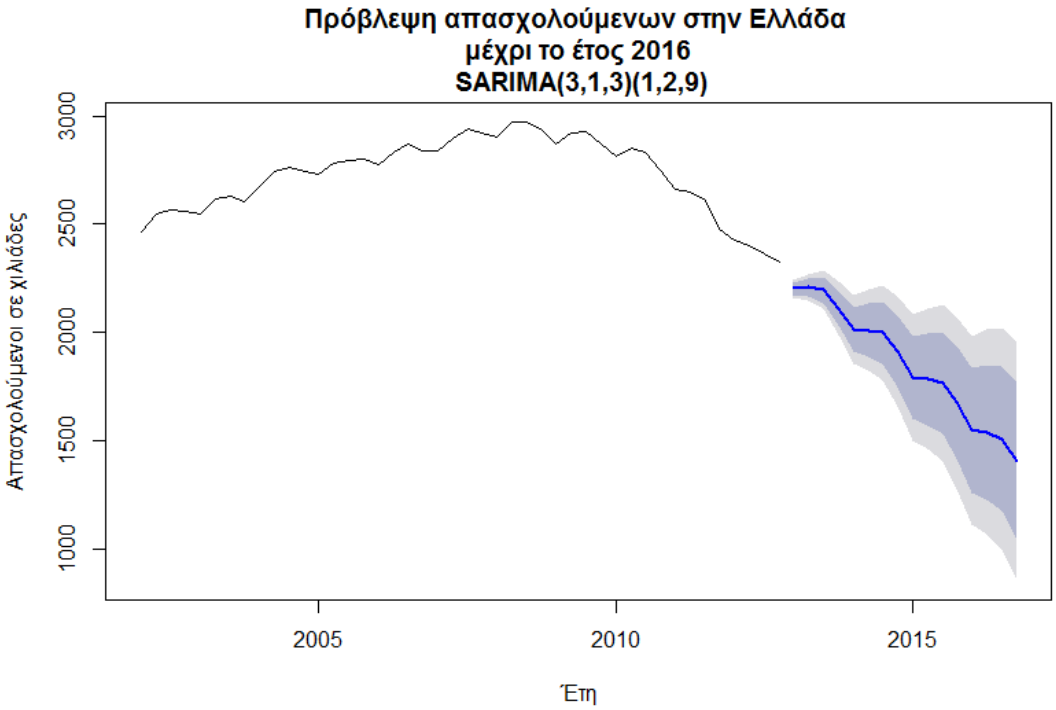
Το Μέσο Σφάλμα υπολογίζεται από τον τύπο $ME = \sum_{t=N-m+1}^{N} \frac{e_t}{m}$, και δείχνει το μέσο όρο του βάρους της πρόβλεψης. Το Σφάλμα της Πρόβλεψης (e_t) είναι η διαφορά μεταξύ της πραγματικής τιμής από την προβλεπόμενη και δίνεται από τον τύπο: $e_t = x_t - \hat{x}_{t-1}$. Από τους παραπάνω τύπους προκύπτει πως όταν το $ME > 0$ αυτό σημαίνει ότι οι προβλέψεις έχουν υποτιμηθεί, όπως συμβαίνει και στην περίπτωση μας όπου το $ME = 91{,}35$, ενώ όταν $ME < 0$ τότε οι προβλέψεις έχουν υπερεκτιμηθεί (6).

Η Ρίζα του Μέσου Τετραγωνικού Σφάλματος υπολογίζεται από τον τύπο: $RMSE = \frac{1}{m}\sqrt{(\sum_{t=N-m+1}^{N} e_t^2)}$ και μας δείχνει πόση διαφορά θα έχουν οι προβλεπόμενες τιμές από την πραγματική τιμή, στην περίπτωση μας $RMSE = 99{,}75$ (6).

Τέλος το Μέσο Απόλυτο Ποσοστό Σφάλματος υπολογίζεται από τον τύπο: $MAPE = \sum_{t=N-m+1}^{N} |e_t / x_t| / m$ και δείχνει το ποσοστό που διαφέρει η πρόβλεψη από την πραγματική τιμή, εδώ $MAPE = 4{,}02\%$ (6).

Με τον ίδιο τρόπο μπορούμε να προβλέψουμε τους απασχολούμενους στην Ελλάδα μέχρι το έτος 2016, εφαρμόζοντας το ίδιο μοντέλο SARIMA, στο ίδιο

σύνολο δεδομένων. Όπως φαίνεται στο Διάγραμμα 16 και στον Πίνακας 7 οι εργαζόμενοι στην Ελλάδα θα συνεχίζουν να μειώνονται ραγδαία.



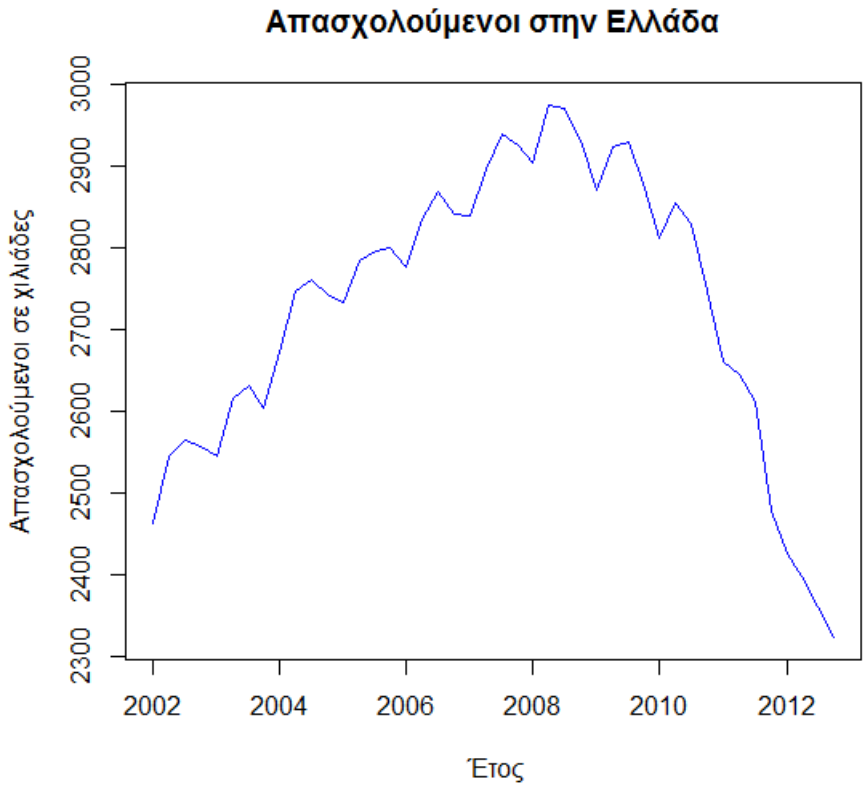
Διάγραμμα 16 Απασχολούμενοι στην Ελλάδα μέχρι το έτος 2016- μέθοδος SARIMA

Έτος	Τρίμηνο	Προβλεπόμενη Τιμή
2014	**Q1**	2012,700
2014	**Q2**	2009,058
2014	**Q3**	1999,628
2014	**Q4**	1911,762
2015	**Q1**	1792,826
2015	**Q2**	1784,144
2015	**Q3**	1763,821
2015	**Q4**	1675,252
2016	**Q1**	1546,653
2016	**Q2**	1539,054
2016	**Q3**	1508,168
2016	**Q4**	1401,928

Πίνακας 7 Απασχολούμενοι στην Ελλάδα για τα έτη 2014-2016 - μέθοδος SARIMA

3.3.2. Πρόβλεψη με Νευρωνικά Δίκτυα (NN)

Για την πρόβλεψη με νευρωνικά δίκτυα θα χρησιμοποιήσουμε την βάση δεδομένων που αναφέρεται στους απασχολούμενους στην Ελλάδα για τα έτη 2002 έως 2013, ανά τρίμηνο, με τον ίδιο ακριβώς τρόπο όπως κάναμε και στην πρόβλεψη προηγουμένως. Αρχικά έχουμε 48 εγγραφές, από τις οποίες οι 44 θα οριστούν ως εγγραφές για εκπαίδευση (train) και οι υπόλοιπες 4 θα οριστούν για έλεγχο (test), δηλαδή θα γίνει εκπαίδευση των νευρώνων από το 2002 μέχρι το 2012 και με βάση αυτά τα δεδομένα θα γίνει η πρόβλεψη για το έτος 2013. Στην συνέχεια θα ορίσουμε ως χρονοσειρά τις 44 εγγραφές που είναι προς εκπαίδευση. Το διάγραμμα της χρονοσειράς για τα έτη 2002 έως 2012 είναι το εξής:



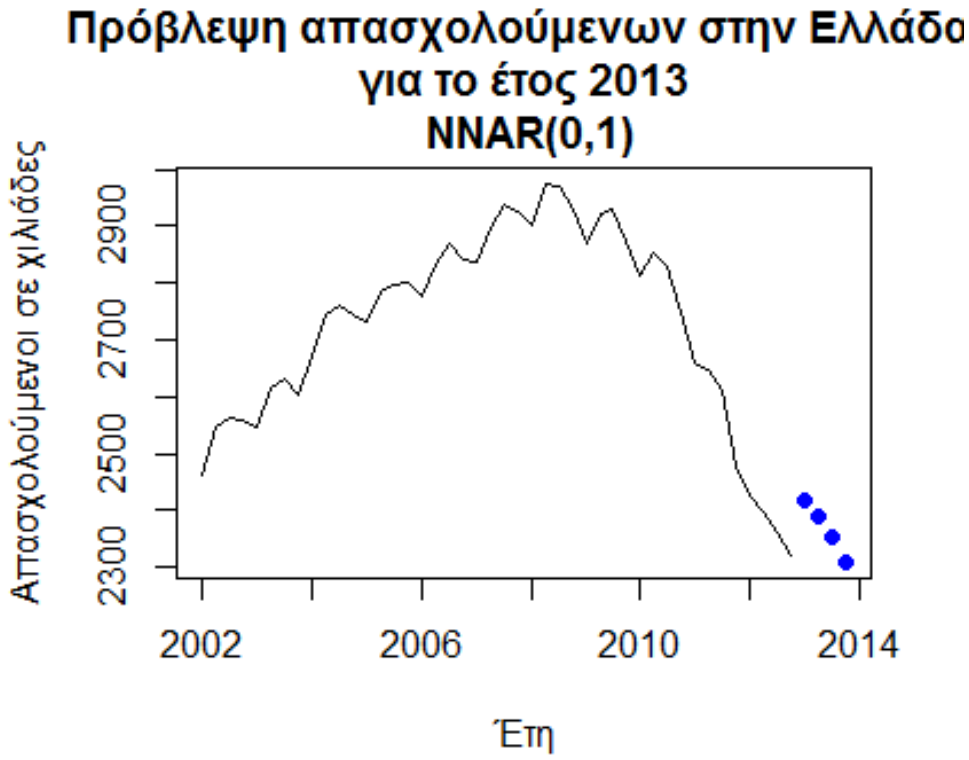
Διάγραμμα 17 Απεικόνιση χρονοσειράς για τα έτη 2012-2012

Με την χρήση της συνάρτησης *nnetar()* που ανήκει στην βιβλιοθήκη *forecast* θα κάνουμε πρόβλεψη των απασχολούμενων στην Ελλάδα για το έτος 2012 και έπειτα θα συγκρίνουμε την προβλεπόμενη τιμή με την πραγματική τιμή του έτους 2013 για να δούμε πόσο επιτυχής είναι η πρόβλεψη. Το μοντέλο του νευρωνικού δικτύου είναι $nnetar(p, P, K)$, όπου p είναι η τάξη του αυτοπαλινδρούμενου μοντέλου AR, P η

εποχικότητα, και K ο αριθμός των κρυφών στρωμάτων των νευρώνων. Στην περίπτωση μας το κατάλληλο μοντέλο είναι $nnetar(0,1,3)$. Η πρόβλεψη των απασχολούμενων για το έτος 2013, η πραγματική τιμή και η διαφορά τους φαίνονται στον Πίνακας 8.

Έτος	Τρίμηνο	Προβλεπόμενη Τιμή	Πραγματική Τιμή	Διαφορά
2013	**Q1**	2418,779	2245,3	173,479
2013	**Q2**	2390,011	2285,7	104,311
2013	**Q3**	2352,274	2283,5	68,774
2013	**Q4**	2309,456	2265,8	43,656

Πίνακας 8 Πρόβλεψη για το έτος 2013 με την μέθοδο NN

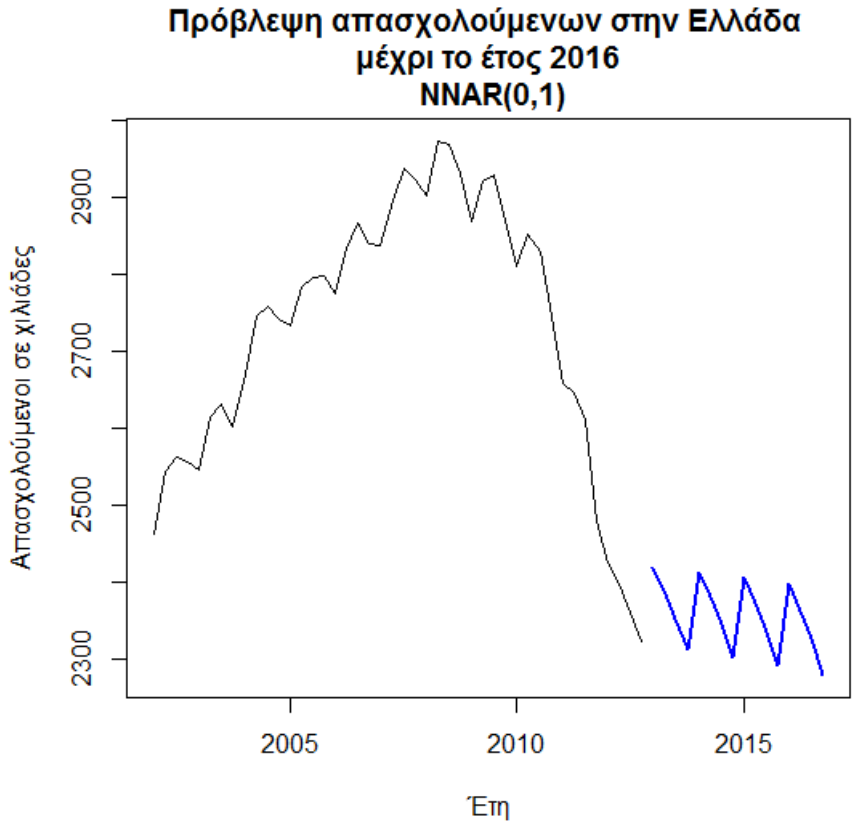


Διάγραμμα 18 Πρόβλεψη για το έτος 2013 με την μέθοδο NN

Ο Πίνακας 8 δείχνει ότι η προβλεπόμενη τιμή στο πρώτο τρίμηνο του 2013 είναι μεγαλύτερη από την πραγματική τιμή των απασχολούμενων και στην συνέχεια μειώνεται, ενώ η πραγματική τιμή αυξάνετε στο δεύτερο τρίμηνο και μειώνεται αργότερα. Το ίδιο βλέπουμε και στο Διάγραμμα 18. Η απόκλιση των προβλεπόμενων

τιμών με τις πραγματικές δεν απέχει πολύ, πράγμα που αποδεικνύεται χρησιμοποιώντας την συνάρτηση *accuracy()*, η οποία εμφανίζει την ακρίβεια της πρόβλεψης. Εφαρμόζοντας τους ίδιους δείκτες που χρησιμοποιήσαμε στην πρόβλεψη με το μοντέλο $SARIMA$, προκύπτει ότι το Μέσο Σφάλμα είναι $ME = -97{,}97$ που σημαίνει ότι οι προβλέψεις μας έχουν υπερεκτιμηθεί. Η Ρίζα του Μέσου Τετραγωνικού Σφάλματος είναι $RMSE = 109{,}30$ και δείχνει την διαφορά της πραγματικής τιμής από την προβλεπόμενη. Τέλος το Μέσο Απόλυτο Ποσοστό Σφάλματος στην περίπτωσή μας είναι $MAPE = 4{,}32$, δηλαδή κατά μέσο όρο το ποσοστό που διαφέρει η πρόβλεψη μας από την πραγματική τιμή είναι 4,32%.

Με το ίδιο μοντέλο νευρωνικών δικτύων και χρησιμοποιώντας το ίδιο σύνολο εκπαίδευσης, μπορούμε να κάνουμε την πρόβλεψη των απασχολούμενων στην Ελλάδα μέχρι το έτος 2016. Το διάγραμμα της πρόβλεψης φαίνεται στο Διάγραμμα 19.



Διάγραμμα 19 Απασχολούμενοι στην Ελλάδα μέχρι το 2016 με την μέθοδο NN

Στον Πίνακας 9 φαίνεται ότι οι απασχολούμενοι στην Ελλάδα θα συνεχίζουν να μειώνονται συνεχώς, φτάνοντας το τέταρτο τρίμηνο του 2016 τους 2292,680 χιλιάδες.

Έτος	Τρίμηνο	Προβλεπόμενη Τιμή
2014	**Q1**	2412,595
2014	**Q2**	2382,014
2014	**Q3**	2344,888
2014	**Q4**	2309,083
2015	**Q1**	2405,537
2015	**Q2**	2372,898
2015	**Q3**	2336,028
2015	**Q4**	2301,164
2016	**Q1**	2397,996
2016	**Q2**	2363,243
2016	**Q3**	2326,642
2016	**Q4**	**2292,680**

Πίνακας 9 Απασχολούμενοι στην Ελλάδα για τα έτη 2014-2016 με την μέθοδο NN

3.3.3. Σύγκριση του μοντέλου SARIMA με τα Νευρωνικά Δίκτυα

Ύστερα από την εκτέλεση της πρόβλεψης με την μέθοδο SARIMA και με τα Νευρωνικά Δίκτυα στην ίδια βάση δεδομένων μπορούμε να συγκρίνουμε τα αποτελέσματα με βάση τα μέτρα για την ακρίβεια της πρόβλεψης. Στον Πίνακας 10 βλέπουμε την σύγκριση των δύο μεθόδων για το έτος 2013:

Μέθοδος	**SARIMA**			**NEURAL NETWORKS**		
Έλεγχος	**ME**	**RMSE**	**MAPE(%)**	**ME**	**RMSE**	**MAPE(%)**
2013	91.35	99,75	4,02	-97.97	109.30	4.32

Πίνακας 10 Σύγκριση πρόβλεψης για το έτος 2013

Βλέπουμε ότι η πρόβλεψη με το μοντέλο SARIMA είναι καλύτερη σε σχέση με τα Νευρωνικά Δίκτυα, διότι και τα τρία μέτρα της ακρίβειας έχουν καλύτερα αποτελέσματα. Το Μέσο Σφάλμα (ME) της πρόβλεψης είναι καλύτερο με την μέθοδο SARIMA σε σχέση με τα Νευρωνικά Δίκτυα, διότι ο αριθμός είναι πιο μικρός. Στην πρώτη περίπτωση οι προβλέψεις έχουν υποτιμηθεί, ενώ στην δεύτερη έχουν υπερεκτιμηθεί. Η Ρίζα του Μέσου Τετραγωνικού Σφάλματος (RMSE) πάλι είναι μικρότερη στη μέθοδο SARIMA από ότι στα Νευρωνικά Δίκτυα, το ίδιο ισχύει και για το Μέσο Απόλυτο Ποσοστό Σφάλματος (MAPE), το οποίο έχει ελάχιστη διαφορά ανάμεσα στις δύο μεθόδους. Επειδή το MAPE δείχνει το ποσοστό που διαφέρει η

πρόβλεψη από την πραγματική τιμή, όσο πιο μικρό είναι, τόσο καλύτερη είναι η πρόβλεψη.

Συμπερασματικά, όσον αφορά τον τρόπο της απεικόνισης της πρόβλεψης τα διαγράμματα του μοντέλου SARIMA είναι πιο κατανοητά από τα διαγράμματα των νευρωνικών δικτύων. Επίσης στο διάγραμμα με το μοντέλο SARIMA απεικονίζεται και το διάστημα εμπιστοσύνης τιμών που μπορούν να πάρουν οι προβλεπόμενες τιμές. Ανακεφαλαιώνοντας, η πρόβλεψη με χρήση της μεθόδου SARIMA είναι πιο αποτελεσματική

Στον Πίνακας 11 παρουσιάζονται οι εκτιμώμενες τιμές πρόβλεψης των Απασχολούμενων ανά τρίμηνο για τα έτη 2014-2016 με το Μοντέλο $SARIMA(3,1,3) \times (1,2,9)$ και των Νευρωνικών Δικτύων:

Έτος	Τρίμηνο	SARIMA	Νευρωνικά Δίκτυα
2014	**Q1**	2012,700	2412,595
2014	**Q2**	2009,058	2382,014
2014	**Q3**	1999,628	2344,888
2014	**Q4**	1911,762	2309,083
2015	**Q1**	1792,826	2405,537
2015	**Q2**	1784,144	2372,898
2015	**Q3**	1763,821	2336,028
2015	**Q4**	1675,252	2301,164
2016	**Q1**	1546,653	2397,996
2016	**Q2**	1539,054	2363,243
2016	**Q3**	1508,168	2326,642
2016	**Q4**	1401,928	2292,680

Πίνακας 11 Προβλέψεις με SARIMA και ΝΝ από το 2014 - 2016

ΣΥΜΠΕΡΑΣΜΑΤΑ

Η Εξόρυξη Δεδομένων είναι πολύ σημαντική διότι με αυτήν μπορούμε να επεξεργαστούμε σημαντικές πληροφορίες από μεγάλες βάσεις δεδομένων, να πραγματοποιήσουμε την ανάλυση των δεδομένων αυτών, και να καταλήξουμε σε σημαντικά συμπεράσματα. Εκτός αυτού, τα τελευταία χρόνια χρησιμοποιείται σε πολλά επιστημονικά πεδία όπως αυτό των θετικών επιστημών, της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης, αλλά και της ιατρικής, βιολογίας, του μάρκετινγκ για την προσέλκυση νέων πελατών αλλά και την διατήρηση των ήδη υπαρχόντων πελατών τους και αλλού.

Σημαντική είναι η συμβολή της Εξόρυξης Δεδομένων στην ανάλυση χρονοσειρών, όχι μόνο επειδή έχει την δυνατότητα να περιγράφει δεδομένα χρησιμοποιώντας συνοπτικά στατιστικά στοιχεία και γραφικές μεθόδους, αλλά παρέχει και την δυνατότητα της συσταδοποίησης, της πρόβλεψης, της κατηγοριοποίησης χρονοσειρών με συνέπεια την εξαγωγή πολύτιμης γνώσης. Για παράδειγμα όταν το μήκος των χρονοσειρών είναι πολύ μεγάλο είναι απαραίτητο και πολύ χρήσιμο να γίνει μια στατιστική σύνοψη των δεδομένων.

Ειδικότερα με την συσταδοποίηση των χρονικών σειρών μπορούμε να εξετάσουμε με πιο τρόπο ομαδοποιούνται τα δεδομένα που χρησιμοποιούμε κάθε φορά και να αντλήσουμε ενδιαφέρουσες πληροφορίες και συμπεράσματα. Κύριος στόχος όμως την συσταδοποίησης είναι η κατανόηση των δεδομένων για δημιουργία προτύπων. Βασικό κριτήριο για τη χρησιμοποίηση της συσταδοποίησης είναι η επιλογή του αλγορίθμου που θα χρησιμοποιηθεί να μην είναι τυχαία αλλά να επιλέγεται με βάση τα δεδομένα, για παράδειγμα ο k-μέσων λειτουργεί πολύ πιο αποτελεσματικά σε δεδομένα σφαιρικής κατανομής. Επίσης μία άλλη παράμετρος που θα πρέπει να προσεχθεί είναι ο αριθμός των συστάδων που θα ζητηθούν από έναν αλγόριθμο συσταδοποίησης. Επειδή δεν υπάρχει εκ των προτέρων γνώση για το πλήθος των συστάδων που σχηματίζονται στα δεδομένα, θα πρέπει να γίνουν πολλές δοκιμές μέχρι να βρεθεί η βέλτιστη συσταδοποίηση. Τέλος για δεδομένα χρονοσειρών, η μετρική απόστασης Dynamic Time Warping είναι πολύ κατατοπιστική μέσω των

διαγραμμάτων της, οι ομοιότητες και η απόσταση των χρονοσειρών είναι εμφανή και η κατανόηση τους αρκετά εύκολη.

Η πρόβλεψη χρονοσειρών, από την πλευρά της, μπορεί να βοηθήσει να παρθούν αποφάσεις που να είναι χρήσιμες για το μέλλον, αρκεί να έχουμε τα προηγούμενα δεδομένα / εγγραφές, γιατί σε αυτά γίνεται η πρόβλεψη. Εάν η χρονοσειρά που διαθέτουμε είναι στάσιμη, η πρόβλεψη των μελλοντικών τιμών είναι απλή, και γίνεται με την χρησιμοποίηση των μοντέλων AR, MA ή / και τον συνδυασμό τους (ARIMA). Όταν όμως μετά της ανάλυση της χρονοσειράς θα καταλήξουμε ότι η χρονοσειρά που μελετάμε είναι μη-στάσιμη τότε η πρόβλεψη της είναι λίγο πιο δύσκολη, όμως υπάρχουν μοντέλα όπως ARIMA, ή SARIMA αν η χρονοσειρά περιέχει εποχικότητα, τα Νευρωνικά Δίκτυα και οι Μηχανές Διανυσματικής Υποστήριξης. Συγκεκριμένα, στην ανάλυση που επιχειρήθηκε στην παρούσα πτυχιακή η πρόβλεψη με χρήση του μοντέλου SARIMA έχει καλύτερα αποτελέσματα.

Τέλος, το προγραμματιστικό πακέτο R είναι πολύ χρήσιμο διότι παρέχει πάρα πολλές δυνατότητες στους αναλυτές, όχι μόνο επειδή περιέχει αλγόριθμους, βιβλιοθήκες και συναρτήσεις έτοιμα για χρήση, αλλά και επειδή παρέχει και πάρα πολλές μορφές οπτικοποίησης των δεδομένων, με διαγράμματα, εικόνες και πίνακες. Είναι γεγονός ότι αντιμετωπίσαμε αρκετές δυσκολίες κατά την εφαρμογή των μεθόδων στο προγραμματιστικό περιβάλλον R, καθώς ήταν η πρώτη φορά που χρησιμοποιήσαμε το συγκεκριμένο περιβάλλον, ωστόσο επειδή έχουν γραφτεί πολλά βιβλία για το λογισμικό αυτό και επειδή η R από μόνη της παρέχει ένα πλαίσιο βοήθειας, όπου παρουσιάζονται λεπτομερώς όλες οι βιβλιοθήκες, οι συναρτήσεις, τα διαγράμματα που παρέχει, και ο τρόπος χρήσης τους, καταφέραμε να ξεπεράσουμε τις δυσκολίες μας και να ασχοληθούμε με το εντυπωσιακό πρόγραμμα αυτό.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. **Han, Jiawei και Kamber, Micheline.** *Data Mining Concepts and Techniques.* San Francisco : Diane Cerra, 2006. 978-1-55860-901-3.

2. **Tan, Pang-Ning, Steinbach, Michael και Kumar, Vipin.** *Εισαγωγή στην Εξόρυξη Δεδομένων* . Θεσσαλονίκη : ΤΖΙΟΛΑ, 2010. 978-960-418-162-9.

3. **Maimon, Oded και Rokach, Lior.** *Data Mining and Knowledge Discovery Handbook.* London : Springer, 2005. 978-0-387-09822-7.

4. **Yanchang, Zhao.** *R and Data Mining: Examples and Case Studies.* s.l. : Elsevier, 2012.

5. **Vercellis, Carlo.** *Business Intelligence Data Mining and Optimization for Decision Making.* s.l. : WILEY, 2009. 978-0-470-51138-1.

6. **Chatfield, Chris.** *Time-Series Forecasting.* s.l. : Chapman, 2000. 1-58488-063-5.

7. **Μπεγκόμ, Τζαχίντα.** Ανάλυση των Χρηματιστηριακών Δεδομένων με χρήση των Αλγόριθμων Εξόρυξης. *Μεταπτυχιακή Διπλωματική Εργασία.* Πάτρα : Πανεπιστήμιο Πατρών, 2013.

8. *Exact indexing of dynamic time warping.* **Keogh, E. και Ratanamahatana, C. A.** 3, 2005, Knowledge and Information Systems, Τόμ. 7, σσ. 358-386.

9. **Brockwell, Peter J. και Davis, Richard A.** *Introduction to Time Series and Forecasting.* 2η. New York : Springer-Verlag, 2002.

10. **Hyndman, Rob J και Athanasopoulos, George.** Forecasting: principles and practice. [Ηλεκτρονικό] https://www.otexts.org/book/fpp.

11. **Ohri, A.** *R for Business Analytics.* New York : Springer, 2012. 978-1-4614-4342-1.

12. www.quandl.com. [Ηλεκτρονικό]

13. **Cryer, Jonathan D. και Chan, Kung-Sik.** *Time Series Analysis With Application in R.* s.l. : Springer, 2008. 978-0-387-75958-6.

ΠΑΡΑΡΤΗΜΑ Ι

Στο Παράρτημα αυτό παραθέτουμε τον κώδικα σε R για τη Συσταδοποίηση καθώς και τα αποτελέσματα των εντολών.

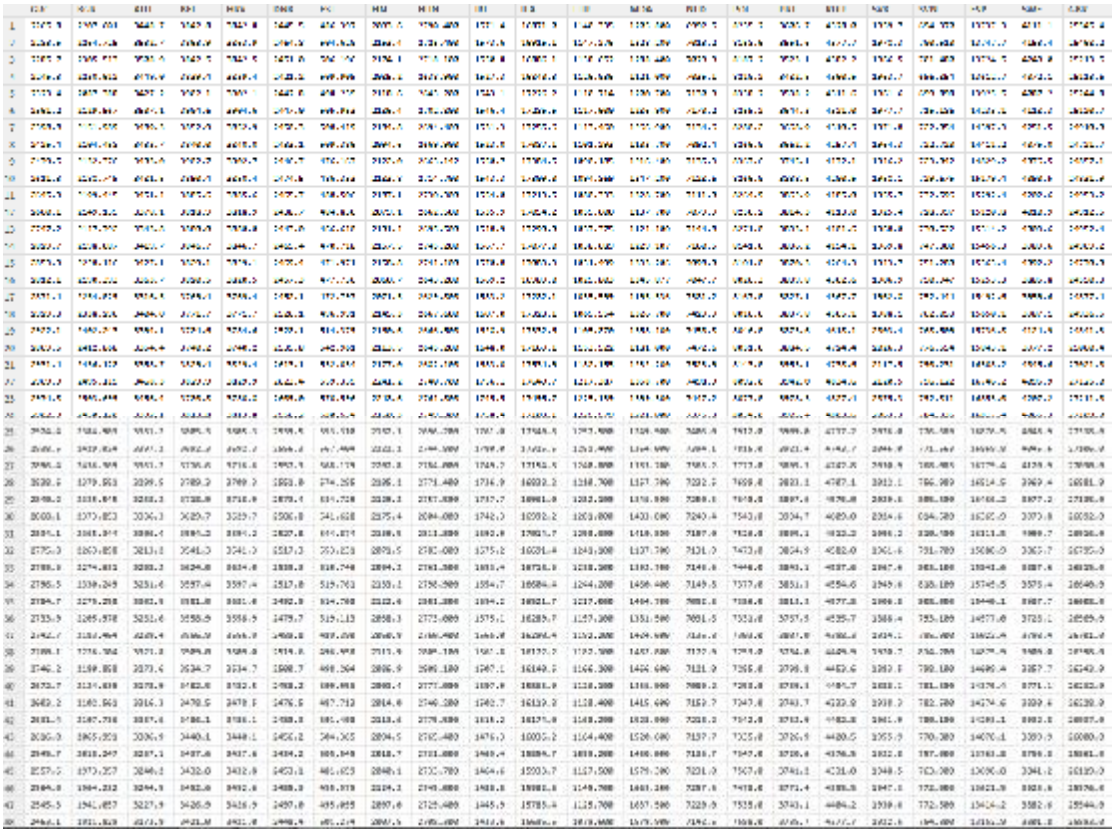
Συλλογή και δημιουργία συνόλου δεδομένων

```
# βιβλιοθήκη ώστε να μπορούμε να χρησιμοποιήσουμε τις βάσεις δεδομένων απευθείας από το site

library ("Quandl", lib.loc="~/R/win-library/3.2")

# εισαγωγή χωρών από το Quandl όπου οι εγγραφές ξεκινούν το 2002 και σταματούν το 2013 με μετρήσεις ανά 3 μήνες

ts.bgr <- Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_BGR", authcode = "Ae4sHKADqFeL42iKJ1Va", collapse = "quarterly", trim_start ="2002-01-31", trim_end = "2013-12-31")

ts.grc<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_GRC", authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31", trim_end="2013-12-31")

ts.aut<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_AUT", authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31", trim_end="2013-12-31")

ts.bel<-
Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_BEL",authcode="Ae4sHKADqFeL42iKJ1Va",collapse="quarterly",trim_start="2002-01-31", trim_end="2013-12-31")

ts.hrv<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_HRV", authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31", trim_end="2013-12-31")

ts.dnk<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_DNK", authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31", trim_end="2013-12-31")
```

```
ts.est<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_EST",
authcode="Ae4sHKADqFeL42iKJ1Va",collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

ts.fin<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_FIN",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

ts.hun<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_HUN",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

ts.irl<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_IRL",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

ts.ita<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_ITA",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

ts.ltu<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_LTU",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

ts.mda<-Quandl("ILOSTAT/EAP_TEAP_NB_SEX_T_M_MDA",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

ts.nld<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_NLD",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

ts.pol<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_POL",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

ts.prt<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_PRT",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")
```

```
ts.rou<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_ROU",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

ts.svk<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_SVK",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

ts.svn<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_SVN",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

ts.esp<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_ESP",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

ts.swe<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_SWE",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

ts.gbr<-Quandl("ILOSTAT/EES_TEES_NB_SEX_T_ECO_SECTOR_TOTAL_M_GBR",
authcode="Ae4sHKADqFeL42iKJ1Va", collapse="quarterly",trim_start="2002-01-31",
trim_end="2013-12-31")

# δημιουργία μιας ενιαίας βάσης δεδομένων με όλες τις χώρες και τις τιμές τους

df.xwres <- data.frame( cbind('GRC'=ts.grc$Value, 'BGR'= ts.bgr$Value, 'AUT'=
ts.aut$Value),'BEL'=ts.bel$Value,'HRV'=ts.hrv$Value,'DNK'=ts.dnk$Value,'EST'=ts.est$Value,'FI
N'=ts.fin$Value,'HUN'=ts.hun$Value,'IRL'=ts.irl$Value,'ITA'=ts.ita$Value,'LTU'=ts.ltu$Value,'MD
A'=ts.mda$Value,'NLD'=ts.nld$Value,'POL'=ts.pol$Value,'PRT'=ts.prt$Value,
'ROU'=ts.rou$Value, 'SVK'=ts.svk$Value, 'SVN'=ts.svn$Value, 'ESP'=ts.esp$Value,
'SWE'=ts.swe$Value, 'GBR'=ts.gbr$Value)

View(df.xwres)
```

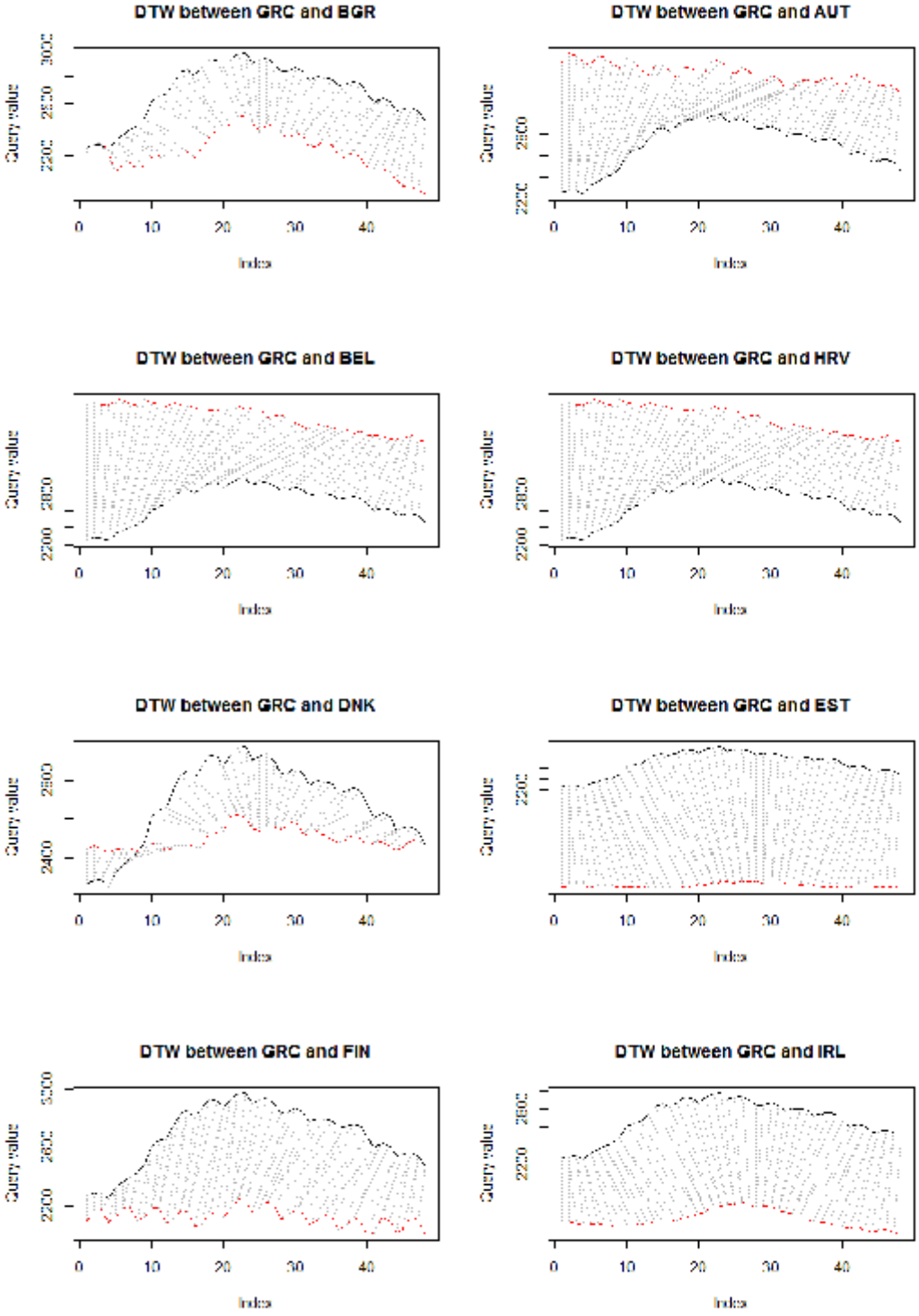
Το αποτέλεσμα της τελευταίας εντολής είναι:

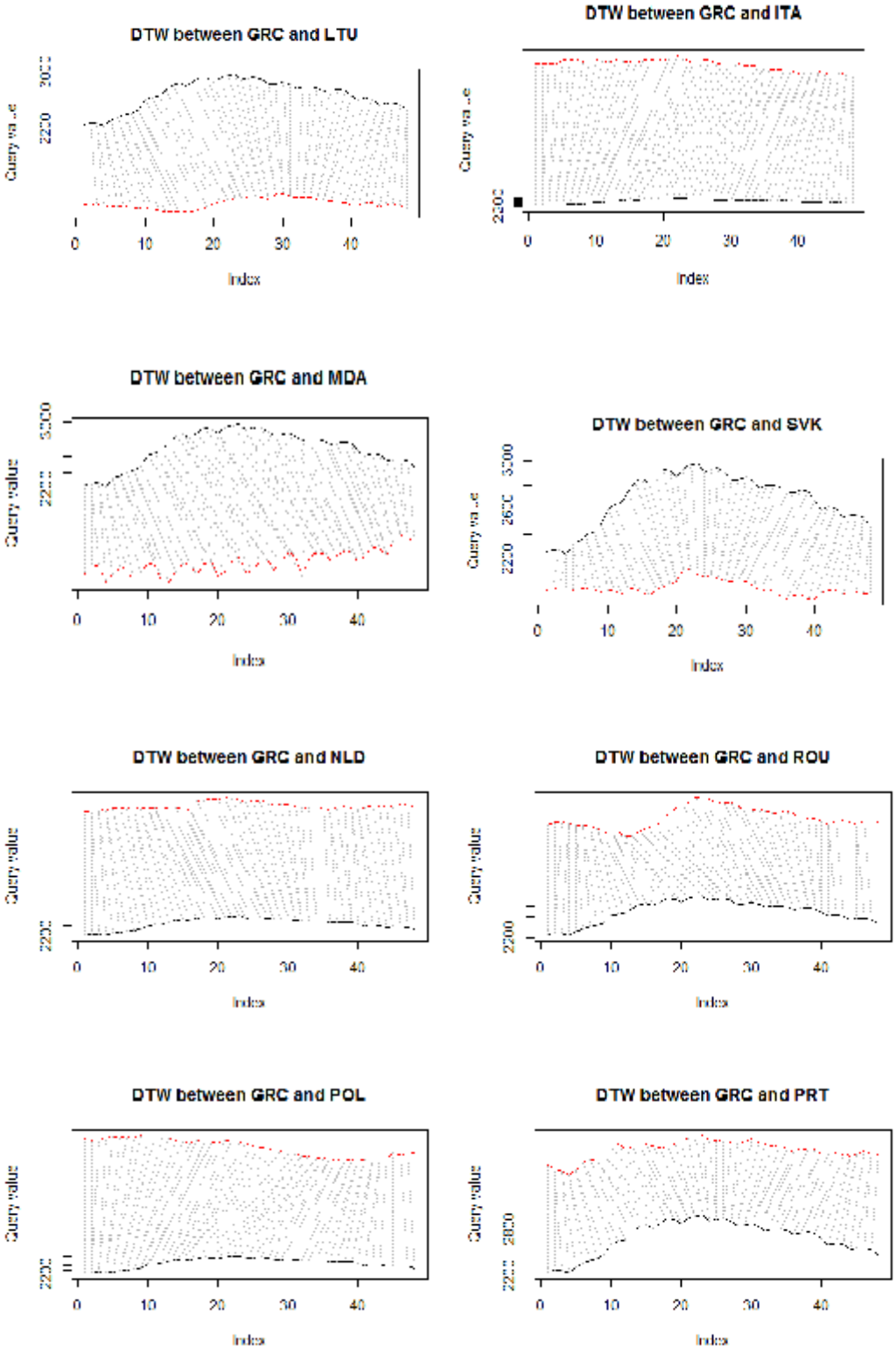


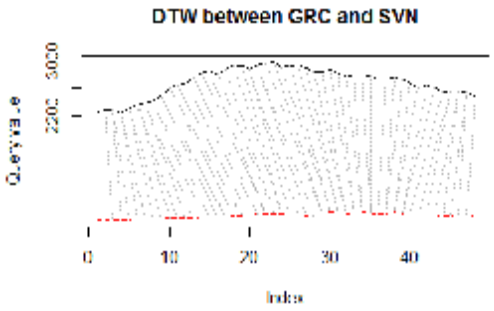
Μέτρηση Απόστασης Χρονοσειρών με την DTW

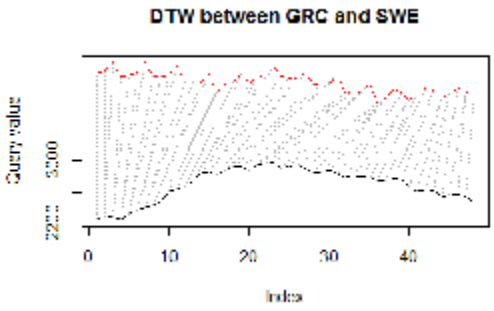
```
library (dtw)

# μια επαναληπτική for ώστε να παίρνει την Ελλάδα με όλες τις άλλες χώρες για τον
υπολογισμό της ομοιότητας .και στο τέλος γίνονται τα διαγράμματα

countries <- names(df.xwres)

i=1

for (j in 2:length(countries))  {

  titlefig <- paste("DTW between", countries[i],"and", countries[j], sep=" " )

  align<-dtw (df.xwres[,countries[i]], df.xwres [,countries[j]], step.pattern =asymmetricP2,
keep=TRUE,  main=titlefig)

  dtwPlotTwoWay(align,main=titlefig)

}

# διαγράμματα dtw
```

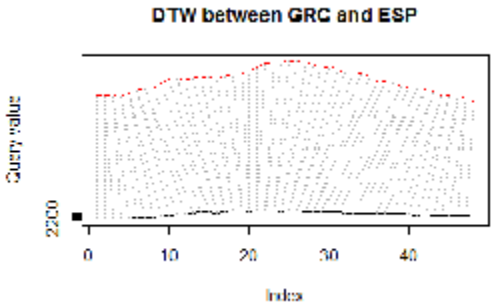
Τα αποτελέσματα των παραπάνω εντολών είναι τα σχεδιαγράμματα που παρουσιάζονται αμέσως μετά και στα οποία η Ελλάδα εμφανίζεται με μαύρη γραμμή.




DTW between GRC and LTU
DTW between GRC and ITA
DTW between GRC and MDA
DTW between GRC and SVK
DTW between GRC and NLD
DTW between GRC and ROU
DTW between GRC and POL
DTW between GRC and PRT
Query value
Index







Συσταδοποίηση

```
# για την ομαδοποίηση με κ-μεσων χρειαζόμαστε τη βιβλιοθήκη cluster

library (cluster)

# χρήση K-means αλγόριθμου στο σύνολο των χωρών με 4 ομάδες

kmeans.result <- kmeans(df.xwres,4)

 kmeans.result
```

Αποτελέσματα συσταδοποίησης με τον k-means:

k-means clustering with 4 clusters of sizes 15, 12, 13, 8

Cluster means:	GRC	BGR	AUT	BEL	HRV
1	2871.720	2382.246	3331.400	3697.593	3697.593
2	2769.333	2236.794	3394.775	3824.983	3824.983
3	2629.377	2077.388	3263.900	3481.354	3481.354
4	2323.637	2185.232	3474.575	3861.025	3861.025

Cluster means:	DNK	EST	FIN	HUN	IRL
1	2557.353	548.3579	2160.347	2763.307	1714.700
2	2475.900	487.4999	2128.442	2683.728	1575.633
3	2475.308	498.9738	2061.615	2757.854	1501.03
4	2444.462	503.6399	2136.337	2697.888	1545.862

Cluster means:	ITA	LTU	MDA	NLD	POL
1	17063.71	1237.651	1346.873	7297.207	7729.333
2	17208.95	1068.908	1260.302	7220.025	8164.000
3	16026.20	1146.338	1497.354	7162.062	7381.846
4	17045.86	1124.421	1225.200	7074.850	8241.850

Cluster means:	PRT	ROU	SVK	SVN	ESP
1	3899.180	4677.547	2025.840	792.5415	16375.25
2	3838.625	4312.375	1962.867	746.1599	15377.63
3	3756.315	4406.477	1924.815	779.6846	14173.62
4	3584.688	4331.113	1967.950	706.6340	13973.66

Cluster means:	SWE	GBR
1	4017.793	26973.47
2	4033.375	24876.02
3	3843.969	26177.00
4	4139.700	25167.24

Αξιολόγηση αποτελεσμάτων συσταδοποίησης k-means

```
# χρήση συνάρτησης dist() που υπολογίζει τη μήτρα απόστασης στο σύνολο των χωρών με τη μέθοδο της ευκλείδειας απόστασης

data.dist<- dist(df.xwres, method='euclidean')

# χρησιμοποίηση της συνάρτησης silhouette με ορίσματα το αποτέλεσμα του k-means και τη μήτρα απόστασης

si <- silhouette (kmeans.result$cluster, data.dist )

# δημιουργία διαγράμματος του συντελεστή περιγράμματος

plot(si, col=c("grey"), main=paste("Silhouette for k-means clustering with k=4"))
```

```
# αποτελέσματα

si
```

	cluster	neighbor	sil_width		cluster	neighbor	sil_width
[1,]	3	2	0.6196641	[25,]	1	3	0.6844790
[2,]	3	2	0.6531012	[26,]	1	3	0.6969283
[3,]	3	1	0.6842458	[27,]	1	3	0.7096040
[4,]	3	1	0.6599841	[28,]	1	3	0.6849978
[5,]	3	1	0.6498461	[29,]	1	3	0.7132857
[6,]	3	1	0.5267067	[30,]	1	3	0.6867658
[7,]	3	1	0.3611157	[31,]	1	3	0.6570702
[8,]	3	1	0.2756921	[32,]	1	2	0.4930619
[9,]	1	3	0.1957420	[33,]	1	2	0.4948053
[10,]	1	3	0.5755611	[34,]	1	2	0.3733041
[11,]	1	3	0.6061202	[35,]	1	2	0.0835448
[12,]	1	3	0.5201235	[36,]	2	1	0.3800853
[13,]	1	3	0.6228135	[37,]	2	1	0.3258613
[14,]	1	3	0.6723363	[38,]	2	1	0.5054315
[15,]	1	3	0.6611396	[39,]	2	4	0.6088396
[16,]	1	3	0.5575186	[40,]	2	4	0.6345561
[17,]	1	3	0.6732953	[41,]	2	4	0.6244363
[18,]	1	3	0.6512426	[42,]	2	4	0.5886458
[19,]	1	3	0.6221615	[43,]	2	4	0.6110760
[20,]	1	3	0.5768104	[44,]	2	4	0.5487640
[21,]	4	1	0.5950983	[45,]	2	4	0.5587459
[22,]	4	1	0.6235489	[46,]	2	4	0.5194695
[23,]	4	1	0.6496162	[47,]	2	4	0.4821214
[24,]	4	1	0.6832510	[48,]	2	4	0.3870755

Παρακάτω εμφανίζεται ο κώδικας των εντολών για συσταδοποίηση με k-means, αλλά χρησιμοποιώντας τη μετρική απόστασης manhattan, καθώς και οι εντολές συσταδοποίησης με Ιεραρχικά σχήματα. Τα αποτελέσματα αξιολογούνται με χρήση του Συντελεστή Περιγράμματος.

```
data.dist<- dist(df.xwres, method='manhattan') # μήτρα απόστασης με μέθοδο

Manhattan

si<- silhouette(kmeans.result$cluster, data.dist ) # σε ποια ομάδα είναι το καθένα και ο

γείτονας

plot(si, col=c("grey"), main=paste("Silhouette for k-means clustering with k=4"))

# βιβλιοθήκη ώστε να χρησιμοποιήσουμε συνάρτηση dtw

#hc.single

data.dist<- dist(df.xwres, method='euclidean')

hc.single <- hclust(data.dist, method="single") # ιεραρχική συσταδοποίηση με μέθοδο απλού

συνδέσμου

plot(hc.single, hang = -1)  #δενδόγραμμα

rect.hclust(hc.single, k=4)  #κόβει το δενδρόγραμμα σε 4 ομάδες

groups <- cutree(hc.single, k=4) # χωρίζει σε 4 ομάδες

 si.single <- silhouette(dist=data.dist, groups)

plot(si.single, col=c("grey") ,main=paste("Silhouette for hc.single clustering with k=4", sep=" ")

)

#hc.complete

hc.complete<-hclust(data.dist, method="complete") # ιεραρχική μέθοδος (dist υπολογίζει

μήτρα απόστασης με μέθοδο πλήρους συνδέσμου)

plot(hc.complete, hang = -1)

rect.hclust(hc.complete, k=4) #κόβει το δενδρόγραμμα σε 4 ομάδες
```

```
groups2<- cutree(hc.complete, k=4) #χωρίζει σε 4 ομάδες

si.complete <- silhouette(dist=data.dist, groups2)

plot(si.complete, col=c("grey") ,main=paste("Silhouette for hc.complete clustering with k=4",
sep=" ") )

#hc.centroid

hc.centroid<- hclust(data.dist, method="centroid")

plot(hc.centroid, hang = -1)# rect.hclust(hc.centroid, k=4)

groups3<- cutree(hc.centroid, k=4)

si.centroid <- silhouette(dist=data.dist, groups3)

plot(si.centroid, col=c("grey") ,main=paste("Silhouette for hc.centroid clustering with
k=4",sep=" "))
```

ΠΑΡΑΡΤΗΜΑ ΙΙ

Στο Παράρτημα αυτό παραθέτουμε τον κώδικα σε R για την Πρόβλεψη με την μέθοδο SARIMA καθώς και τα αποτελέσματα των εντολών.

Δημιουργία συνόλου δεδομένων και χρονοσειράς για τα έτη 2002-2013

```
# καλούμε την βάση δεδομένων από το Excel

library (readxl)

pathname = paste (getwd(), "Ergazomenoi.xlsx", sep="/")

Apasxoloumenoi <- read_excel (pathname, sheet = 1, col_names = TRUE, na='na')

# ορίζουμε την χρονοσειρά με την συνάρτηση ts(), η οποία ξεκινά από το 2002 με συχνότητα
4 (τριμηνιαία)

Apasxoloumenoi <- ts (Apasxoloumenoi$Ergazomenoi,start=c(2002,1),frequency=4)

View(Apasxoloumenoi)
```

Το αποτέλεσμα της τελευταίας εντολής είναι:

```
> Apasxoloumenoi
```

	Qtr1	Qtr2	Qtr3	Qtr4
2002	2463.1	2545.3	2564.0	2557.6
2003	2545.7	2616.0	2631.4	2603.2
2004	2672.7	2746.2	2760.1	2742.7
2005	2733.9	2784.7	2796.5	2799.5
2006	2775.8	2834.1	2868.1	2840.2
2007	2838.6	2896.4	2938.5	2924.4
2008	2902.9	2974.8	2969.9	2931.4
2009	2869.5	2922.1	2929.3	2871.4
2010	2812.1	2853.9	2829.7	2747.2

2011	2660.1	2645.9	2611.3	2479.5
2012	2425.4	2398.8	2361.2	2323.4
2013	2245.3	2285.7	2283.5	2265.8

Διάγραμμα χρονοσειράς και η εσωτερική της δομή

```
# το διάγραμμα της χρονοσειράς

plot (Apasxoloumenoi, main= "Απασχολούμενοι στην Ελλάδα", xlab= "Έτος-Χρόνος", ylab= "Απασχολούμενοι σε χιλιάδες", col= "blue")

# εσωτερική δομής της χρονοσειράς  με χρήση της συνάρτησης decompose()

Apasxoloumenoi.dekomp <- decompose (Apasxoloumenoi)

# διάγραμμα εσωτερικής δομής της χρονοσειράς

plot (Apasxoloumenoi.dekomp, col= "blue")
```

Δημιουργία καινούργιας χρονοσειράς για τα έτη 2002-2012

Κάνουμε την πρόβλεψη για το έτος 2013 χρησιμοποιώντας την βιβλιοθήκη forecast αφού έχουμε καλέσει άλλο Excel στο οποίο έχουμε αφαιρέσει τις 4 τελευταίες εγγραφές που αναφέρονται στο έτος 2013 και ξαναορίζουμε την χρονοσειρά

```
library (forecast)

pathname = paste (getwd(), "Ergazomenoi1.xlsx", sep="/")

Apasxoloumenoi1 <- read_excel (pathname, sheet = 1 ,col_names = TRUE, na='na')

Apasxoloumenoi1<-ts(Apasxoloumenoi1$Ergazomenoi,start=c(2002,1),frequency=4)

View(Apasxoloumenoi1)
```

Το αποτέλεσμα της τελευταίας εντολής είναι:

```
> Apasxoloumenoi1
```

	Qtr1	Qtr2	Qtr3	Qtr4
2002	2463.1	2545.3	2564.0	2557.6
2003	2545.7	2616.0	2631.4	2603.2
2004	2672.7	2746.2	2760.1	2742.7
2005	2733.9	2784.7	2796.5	2799.5
2006	2775.8	2834.1	2868.1	2840.2
2007	2838.6	2896.4	2938.5	2924.4
2008	2902.9	2974.8	2969.9	2931.4
2009	2869.5	2922.1	2929.3	2871.4
2010	2812.1	2853.9	2829.7	2747.2
2011	2660.1	2645.9	2611.3	2479.5
2012	2425.4	2398.8	2361.2	2323.4

Με χρήση της συνάρτησης SARIMA, που είναι επέκταση της ARIMA, και με παραμέτρους p=3, d=1, q=3, P=1, D=2, Q=9 προκύπτει το καλύτερο μοντέλο πρόβλεψης με κριτήριο τον κανόνα AIC

```
arima.1 <- arima(Apasxoloumenoi1,order=c(3,1,3),seasonal = list(order=c(1,2,9)))

arima.1
```

Αποτέλεσμα του μοντέλου:

```
> arima.1

Call:

arima (x = Apasxoloumenoi1, order = c(3, 1, 3), seasonal = list(order = c(1, 2, 9)))

Coefficients:
```

ar1	ar2	ar3	ma1	ma2	ma3	sar1	sma2
-0.5439	-0.1563	-0.0904	0.6837	0.6852	0.997	-0.8040	-0.6476

s.e.	0.2483	0.2578	0.2820	0.2721	0.2074	0.299	0.3984	0.9738

	sma3	sma4	sma5	sma6	sma7	sma8	sma9
	0.4137	0.0989	0.0526	0.0195	0.2589	0.3388	-0.6110
s.e.	0.7028	0.5613	0.5411	0.6755	0.7037	0.7603	0.6501

sigma^2 estimated as 297.5: log likelihood = -169.4, aic = 372.8

Πρόβλεψη για το έτος 2013

```
predict <- forecast(arima.1,h=4)

predict
```

Το αποτέλεσμα της πρόβλεψης με την ελάχιστη και την μέγιστη πρόβλεψη

> predict

		Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2013	Q1	2201.326	2173.729	2228.922	2159.120	2243.531
2013	Q2	2207.436	2167.347	2247.524	2146.126	2268.746
2013	Q3	2194.972	2136.639	2253.305	2105.760	2284.184
2013	Q4	2111.171	2030.867	2191.476	1988.356	2233.986

Διάγραμμα της πρόβλεψης

```
plot (predict, main= "Πρόβλεψη απασχολούμενων στην Ελλάδα \n για το έτος 2013 \n
SARIMA(3,1,3)(1,2,9)", xlab= "Ετη", ylab= "Απασχολούμενοι σε χιλιάδες")
```

Έλεγχος της πρόβλεψης χρησιμοποιώντας την εντολή accuracy()

```
accuracy (predict, Apasxoloumenoi [45:48])
```

```
> accuracy (predict, Apasxoloumenoi [45:48])
```

	ME	RMSE	MAE	MPE	MAPE
Test set	91.34	99.75	91.34	4.02	4.02

	MASE	ACF1
Test set	2.32	NA

Πρόβλεψη μέχρι το έτος 2016

```
predict1<-forecast(arima.1,h=16)
```

Το αποτέλεσμα της πρόβλεψης

		Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2013	Q1	2201.326	2173.729	2228.922	2159.120	2243.531
2013	Q2	2207.436	2167.347	2247.524	2146.126	2268.746
2013	Q3	2194.972	2136.639	2253.305	2105.760	2284.184
2013	Q4	2111.171	2030.867	2191.476	1988.356	2233.986
2014	Q1	2012.700	1909.946	2115.453	1855.552	2169.848
2014	Q2	2009.058	1886.731	2131.386	1821.975	2196.142
2014	Q3	1999.628	1856.331	2142.924	1780.475	2218.781
2014	Q4	1911.762	1746.072	2077.452	1658.360	2165.164
2015	Q1	1792.826	1602.320	1983.333	1501.472	2084.181
2015	Q2	1784.144	1571.435	1996.853	1458.834	2109.455

2015	Q3	1763.821	1527.845	1999.798	1402.926	2124.717
2015	Q4	1675.252	1415.180	1935.323	1277.506	2072.997
2016	Q1	1546.653	1260.172	1833.135	1108.517	1984.789
2016	Q2	1539.054	1228.447	1849.662	1064.021	2014.088
2016	Q3	1508.168	1172.582	1843.754	994.933	2021.402
2016	Q4	1401.928	1041.066	1762.791	850.037	1953.820

Διάγραμμα πρόβλεψης

```
plot (predict1, main= "Πρόβλεψη απασχολούμενων στην Ελλάδα \n μέχρι το έτος 2016 \n SARIMA(3,1,3)(1,2,9)", xlab= "Ετη", ylab= "Απασχολούμενοι σε χιλιάδες")
```

ΠΑΡΑΡΤΗΜΑ III

Στο Παράρτημα αυτό παραθέτουμε τον κώδικα σε R για την Πρόβλεψη με την μέθοδο Νευρωνικών Δικτύων (NN) καθώς και τα αποτελέσματα των εντολών.

Δημιουργία συνόλου δεδομένων, ορισμός συνόλου εκπαίδευσης και ελέγχου, και ορισμός της χρονοσειράς για τα έτη 2002-2012

```
# βιβλιοθήκη για να καλέσουμε δεδομένα από excel στην R

library (readxl)

# βιβλιοθήκη που βοηθάει στην πρόβλεψη και τα νευρωνικά δίκτυα

library (forecast)

# το μονοπάτι για να περάσουμε δεδομένα από excel στην R

pathname = paste (getwd(), "Ergazomenoi.xlsx", sep="/")

dset.Ergazomenoi <-  read_excel(pathname, sheet = 1 ,col_names = TRUE, na='na')

# ορισμός συνόλου εκπαίδευσης και συνόλου για έλεγχο

dset.Ergazomenoi.train <- dset.Ergazomenoi [1:44,]

dset.Ergazomenoi.test <- dset.Ergazomenoi [45:48,]

# εμφανίζει τα δεδομένα που του έχουμε ορίσει

View (dset.Ergazomenoi)

# ορίζουμε χρονοσειρά από το 2002 μέχρι το 2012 με συχνότητα 4 (τριμηνιαία)

ts.Ergazomenoi.train <- ts(dset.Ergazomenoi.train, start=c(2002,1),frequency=4)

View(ts.Ergazomenoi.train)
```

Το αποτέλεσμα της τελευταίας εντολής είναι:

ts.Ergazomenoi.train

	Qtr1	Qtr2	Qtr3	Qtr4
2002	2463.1	2545.3	2564.0	2557.6
2003	2545.7	2616.0	2631.4	2603.2
2004	2672.7	2746.2	2760.1	2742.7
2005	2733.9	2784.7	2796.5	2799.5
2006	2775.8	2834.1	2868.1	2840.2
2007	2838.6	2896.4	2938.5	2924.4
2008	2902.9	2974.8	2969.9	2931.4
2009	2869.5	2922.1	2929.3	2871.4
2010	2812.1	2853.9	2829.7	2747.2
2011	2660.1	2645.9	2611.3	2479.5
2012	2425.4	2398.8	2361.2	2323.4

Διάγραμμα χρονοσειράς

```
plot (ts.Ergazomenoi.train,main= "Απασχολούμενοι στην Ελλάδα", xlab= " Έτος", ylab=
"Απασχολούμενοι σε χιλιάδες",col="blue")
```

Πρόβλεψη για το έτος 2013

Η πρόβλεψη του 2013 θα γίνει με την κλήση της συνάρτησης nnetar() με p=0 , P=1 και size=3, για h=4 περιόδους, δηλαδή για τα 4 τρίμηνα του 2013

```
forecastnn <-  nnetar(ts.Ergazomenoi.train, p=0, P=1, size=3)

nnfcast <- forecast (forecastnn,h=4)

nnfcast
```

Το αποτέλεσμα της τελευταίας εντολής είναι:

	Qtr1	Qtr2	Qtr3	Qtr4
2013	2418.779	2390.011	2352.274	2309.456

Διάγραμμα της πρόβλεψης για το έτος 2013

```
plot (nnfcast,main= "Πρόβλεψη απασχολούμενων στην Ελλάδα \n για το έτος 2014 \n
NNAR(0,1)", xlab="Έτη", ylab="Απασχολούμενοι σε χιλιάδες")
```

Κάνουμε έλεγχο της πρόβλεψης

```
accuracy(nnfcast)
```

```
> accuracy(nnfcast)
```

	ME	RMSE	MAE	MPE	MAPE
Test set	-97.97	109.30	-97.97	4.32	4.32

	MASE	ACF1
Test set	2.47	NA

Πρόβλεψη για μέχρι το έτος 2016

```
forecastnn1 <- nnetar(ts.Ergazomenoi.train, p=0, P=1, size=3)

forecastnn1

nnfcast1<- forecast(forecastnn1,h=16)

nnfcast1
```

Το αποτέλεσμα της τελευταίας εντολής είναι:

	Qtr1	Qtr2	Qtr3	Qtr4
2013	2419.205	2390.635	2353.266	2316.482
2014	2412.595	2382.014	2344.888	2309.083
2015	2405.537	2372.898	2336.028	2301.164
2016	2397.996	2363.243	2326.642	2292.680

Δημιουργία διαγράμματος της πρόβλεψης

```
plot(nnfcast1,main= "Πρόβλεψη απασχολούμενων στην Ελλάδα \n μέχρι το έτος 2016 \n NNAR(0,1)", xlab= "Έτη", ylab= "Απασχολούμενοι σε χιλιάδες")
```