ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ & ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Ομαδοποίηση Χρονοσειρών Βάσει Ποιοτικών Χαρακτηριστικών και Μοτίβων και Αξιοποίηση Ευρημάτων για τη Βελτίωση της Ακρίβειας Πρόβλεψης Κλασικών Μεθόδων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Νικόλαος Κ. Αθηνιώτης

Επιβλέπων: Ασημακόπουλος Βασίλειος

Καθηγητής Ε.Μ.Π

Υπεύθυνος: Σπηλιώτης Ευάγγελος

Διδάκτωρ Ε.Μ.Π

ΕΡΓΑΣΤΗΡΙΟ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Αθήνα, Φεβρουάριος 2019

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ & ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Ομαδοποίηση Χρονοσειρών Βάσει Ποιοτικών Χαρακτηριστικών και Μοτίβων και Αξιοποίηση Ευρημάτων για τη Βελτίωση της Ακρίβειας Πρόβλεψης Κλασικών Μεθόδων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Νικόλαος Κ. Αθηνιώτης

Επιβλέπων: Ασημακόπουλος Βασίλειος

Καθηγητής, Ε.Μ.Π

Υπεύθυνος: Σπηλιώτης Ευάγγελος

Διδάκτωρ, Ε.Μ.Π

Εγκρίθηκε από την τριμελή επιτροπή την 28η Φεβρουαρίου 2019

…….	……..	……..
Βασίλειος Ασημακόπουλος	Ιωάννης Ψαρράς	Δημήτριος Ασκούνης
Καθηγητής, Ε.Μ.Π	Καθηγητής, Ε.Μ.Π	Καθηγητής, Ε.Μ.Π

ΕΡΓΑΣΤΗΡΙΟ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Αθήνα, Φεβρουάριος 2019

(Υπογραφή)

……

ΝΙΚΟΛΑΟΣ Κ. ΑΘΗΝΙΩΤΗΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών, Ε.Μ.Π.

© 2019 – All rights reserved

Copyright © Νικόλαος Κ. Αθηνιώτης, 2019

Με επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τους συγγραφείς.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τους συγγραφείς και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

ΠΡΟΛΟΓΟΣ

Η διπλωματική αυτή εργασία εκπονήθηκε στα πλαίσια των ερευνητικών δραστηριοτήτων της Μονάδας Προβλέψεων και Στρατηγικής κατά το ακαδημαϊκό έτος 2018-2019. Η μονάδα υπάγεται στον Τομέα Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων της Σχολής Ηλεκτρολόγων Μηχανικών & Μηχανικών Η/Υ, του Εθνικού Μετσόβιου Πολυτεχνείου.

Αρχικά, θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Βασίλη Ασημακόπουλο για την ευκαιρία που μου έδωσε, αναθέτοντας μου τη συγκεκριμένη διπλωματική εργασία, ώστε να ασχοληθώ με το αντικείμενο των προβλέψεων, το οποίο αποτελεί πλέον ένα από τα βασικά μου επιστημονικά και επαγγελματικά ενδιαφέροντα. Επίσης, θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Ιωάννη Ψαρρά και τον Καθηγητή κ. Δημήτριο Ασκούνη για την τιμή που μου έκαναν να συμμετάσχουν στην τριμελή εξεταστική επιτροπή της εργασίας.

Θα ήθελα να δώσω θερμές ευχαριστίες στο Διδάκτορα κ. Ευάγγελο Σπηλιώτη. Η συνεισφορά του ήταν καθοριστική για την ολοκλήρωση της παρούσας εργασίας, λόγω της διαρκούς παρακολούθησης του, καθώς και των πολύτιμων υποδείξεων, συμβουλών και παρατηρήσεων του καθ' όλη τη διάρκεια της συνεργασίας μας. Επίσης, θα ήθελα να ευχαριστήσω και τα υπόλοιπα μέλη της Μονάδας Προβλέψεων και Στρατηγικής.

Αφιερώνω αυτήν την εργασία στους γονείς μου, Κωνσταντίνο Αθηνιώτη και Ειρήνη Χατζέλλη και στον αδερφό μου, Ιωάννη Αθηνιώτη, για την ολόψυχη αγάπη και υποστήριξη τους όλα αυτά τα χρόνια.

Τέλος, θα ήθελα να ευχαριστήσω όλους τους φίλους μου για τη συνεχή υποστήριξη τους και τα υπέροχα φοιτητικά χρόνια που περάσαμε μαζί.

Νικόλαος Αθηνιώτης

Αθήνα, Φεβρουάριος 2019

ΠΕΡΙΛΗΨΗ

Σκοπός της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη μίας μεθοδολογίας μέσω της οποίας θα επιτευχθεί μείωση του μέσου σφάλματος που παράγει ένα σύνολο μεθόδων πρόβλεψης όταν αυτό καλείται να προεκτείνει μια συγκεκριμένη ομάδα χρονοσειρών. Η μείωση αυτή πραγματοποιείται μέσω της επιλογής της καταλληλότερης μεθόδου πρόβλεψης για κάθε χρονοσειρά, έχοντας ως γνώμονα την απόδοση των εξεταζόμενων μεθόδων σε αντίστοιχες χρονοσειρές ίδιου μοτίβου και ποιοτικών χαρακτηριστικών. Για να γίνει αυτό, οι χρονοσειρές συγκρίνονται αρχικά ως προς την ομοιότητά τους και στη συνέχεια ομαδοποιούνται ούτως ώστε να διαπιστωθεί βάσει ελέγχων ποιες μέθοδοι πρόβλεψης παράγουν το μικρότερο δυνατό σφάλμα ανά περίπτωση. Η ύπαρξη μιας τέτοιας μεθοδολογίας είναι ιδιαιτέρως χρήσιμη στις μέρες μας, καθώς τα μεγάλα δεδομένα (Big Data) καθιστούν την πρόβλεψη χρονοσειρών μέσω κλασικών διαγωνισμών μια επίπονη και χρονοβόρα διαδικασία.

Αρχικά παρουσιάζεται το αντικείμενο της διπλωματικής, η περιγραφή του προβλήματος και οι πρακτικές εφαρμογές της. Στη συνέχεια, δίνεται ο γενικός ορισμός της χρονοσειράς και όλα τα βασικά χαρακτηριστικά της, καθώς και μια μικρή ιστορική ανασκόπηση του τομέα των προβλέψεων. Έπειτα, αναλύονται όλοι οι μέθοδοι πρόβλεψης που χρησιμοποιήθηκαν στην διπλωματική εργασία για την παραγωγή των προβλέψεων και γίνεται διεξοδική βιβλιογραφική επισκόπηση στις μεθόδους σύγκρισης χρονοσειρών.

Στη συνέχεια της διπλωματικής παρουσιάζεται η γενική ιδέα της μεθοδολογίας που αναπτύχθηκε και το προγραμματιστικό περιβάλλον RStudio μαζί με τη γλώσσα προγραμματισμού R, τα οποία ήταν τα εργαλεία για την ανάλυση των δεδομένων και την εξαγωγή των αποτελεσμάτων. Έπειτα, αναλύεται όλη η διαδικασία και όλα τα πειράματα που εκτελέστηκαν, μαζί με τα αντίστοιχα αριθμητικά και γραφικά αποτελέσματα, ώστε τελικώς να καταλήξουμε στην προτεινόμενη μεθοδολογία.

Στο τελευταίο κομμάτι της διπλωματικής εργασίας, συνοψίζονται τα αποτελέσματα όλων των πειραμάτων και βγαίνει ένα τελικό συμπέρασμα για την απόδοση της μεθοδολογίας. Επίσης προτείνονται μελλοντικές προεκτάσεις της συγκεκριμένης εργασίας και θα παρουσιαστεί όλη η βιβλιογραφία που χρησιμοποιήθηκε.

***Λέξεις Κλειδιά*:** Τεχνικές Προβλέψεων, Ομοιότητα Χρονοσειρών, Επιλογή Κατάλληλης Μεθόδου Πρόβλεψης, Λήψη Αποφάσεων

ABSTRACT

The aim of the diploma thesis is to develop a methodology that improves the accuracy achieved by various forecasting methods utilized to extrapolate a set of time series. This becomes possible by selecting the most appropriate forecasting method per series, using as input information about the historical performance of the examined methods across series of similar patterns and characteristics. More specifically, time series are first grouped based on their similarity and then, forecasts are generated according to the accuracy reported for each one of the examined methods per group. The existence of such a methodology is particularly useful nowadays, since Big Data make traditional time series forecasting a strenuous and time intensive process.

The thesis begins by presenting its subject, a description of the examined problem and a recommendation about where the proposed solution could be implemented. Then, an introduction in time series analysis is made, as well as a small literature review of relevant approaches. The forecasting methods used in this thesis are also presented and analyzed along with a thorough bibliographic overview on techniques used for comparing the similarity of times series.

Afterwards, the proposed methodological framework is presented together with the RStudio programming environment and the programming language R, which were utilized for analyzing the data and exporting the results. Next, the process followed and the experiments performed are described, including numeric and graphical results that formulated the final proposal.

In the last part of the diploma thesis, the results of all experiments are summarized and a final conclusion is drawn about the performance of the methodology. After that, future extensions of the thesis and the literature are proposed.

Keywords**:** Forecasting Techniques, Similarity of Time Series, Selection of Appropriate Forecasting Method, Decision Making.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Πίνακας Εικόνων

Κεφάλαιο 1: Εισαγωγή

1.1 Αντικείμενο της εργασίας

Πως γίνεται να προβλέψω με μεγάλη ακρίβεια το μέλλον; Η παραγωγή ορθών προβλέψεων αποτελούσε και ακόμα αποτελεί ένα άλυτο πρόβλημα. Σε αντίθεση με τα παλαιά χρόνια, όπου χρησιμοποιούνταν κυρίως η διαίσθηση για να παραχθούν προβλέψεις, εδώ και πολλά χρόνια έχει αναπτυχθεί ραγδαία η επιστήμη των τεχνικών προβλέψεων ή αλλιώς “forecasting”, λόγω της τεχνολογικής εξέλιξης. Το “forecasting” είναι η διαδικασία παραγωγής προβλέψεων στο μέλλον που βασίζονται στις παρελθοντικές τιμές του υπό εξέταση μεγέθους. Οι χρονοσειρές είναι ακολουθίες τιμών, οι οποίες αναλύονται και προεκτείνονται στο μέλλον. Έχουν αναπτυχθεί πάρα πολλές μέθοδοι, καθώς και διάφοροι απλοί και πολύπλοκοι αλγόριθμοι με σκοπό την παραγωγή ακριβών προβλέψεων, δηλαδή τη μηδενική διαφορά πρόβλεψης και πραγματικής τιμής στο μέλλον.

Ένα από τα κύρια ζητούμενα της επιστήμης των προβλέψεων είναι η επιλογή της κατάλληλης μεθόδου πρόβλεψης που θα επιφέρει το ελάχιστο σφάλμα. Λόγω της εξέλιξης της τεχνολογίας και των υπολογιστών, έχει αυξηθεί ραγδαία ο όγκος των καθημερινών δεδομένων που λαμβάνονται. Ως απόρροια αυτού, είναι απαγορευτική η σύγκριση όλων των μεθόδων για ένα τόσο υπερογκώδης πλήθος δεδομένων, καθώς ο χρόνος εκτέλεσης των συγκρίσεων είναι εξαντλητικός. Επίσης, σε ένα τόσο μεγάλο πλήθος δεδομένων, ο παράγοντας της τυχαιότητας είναι ιδιαίτερα αυξημένος, οπότε ένας εναλλακτικός τρόπος σύγκρισης των μεθόδων πρόβλεψης είναι επιτακτική ανάγκη.

Η συγκεκριμένη διπλωματική εργασία στοχεύει στην ανάπτυξη μιας μεθοδολογίας, όπου η επιλογή μιας μεθόδου πρόβλεψης για μια οποιαδήποτε χρονοσειρά θα γίνεται μέσω ανάλυσης και πρόβλεψης των όμοιων της από ένα σύνολο χρονοσειρών. Πιο συγκεκριμένα, ομαδοποιούνται οι χρονοσειρές βάσει της ομοιότητας τους και τα ευρήματα αξιοποιούνται, ώστε να επιλέγεται σε κάθε περίπτωση χρονοσειράς η κατάλληλη μέθοδος πρόβλεψης. Με αυτόν τον τρόπο, επιτυγχάνεται μείωση του σφάλματος πρόβλεψης των κλασικών μεθόδων.

1.2 Δομή της εργασίας

Στο δεύτερο κεφάλαιο της παρούσας εργασίας πραγματοποιείται μια εισαγωγή στην επιστήμη των προβλέψεων και των χρονοσειρών. Αρχικά παρουσιάζονται οι προβλέψεις και οι κατηγορίες τους με λίγα λόγια. Στη συνέχεια δίνεται ο ορισμός της χρονοσειράς, οι διάφορες κατηγορίες στις οποίες μπορούν να διαχωριστούν, τα χαρακτηριστικά και οι στόχοι τους. Έπειτα αναλύεται η γενική διαδικασία παραγωγής προβλέψεων μέσω χρονοσειρών, τα ποιοτικά τους χαρακτηριστικά και το κεφάλαιο τελειώνει με μια ιστορική ανασκόπηση του κόσμου των προβλέψεων.

Στο τρίτο κεφάλαιο, αρχικά, παρουσιάζονται επιγραμματικά τα γενικά μοντέλα πρόβλεψης και στη συνέχεια αναλύονται εκτενώς όλοι οι μέθοδοι πρόβλεψης που χρησιμοποιούνται στην εργασία. Τέλος, αναφέρονται τα διάφορα σφάλματα που χρησιμοποιούνται για την εκτίμηση των προβλέψεων.

Το τέταρτο κεφάλαιο είναι αφιερωμένο σε ένα πολύ σημαντικό και ιδιαίτερο κομμάτι της εργασίας, που είναι η παρουσίαση και η ανάλυση των μεθόδων που χρησιμοποιήθηκαν για τη σύγκριση και τελικώς την ομαδοποίηση των χρονοσειρών. Σε κάθε μέθοδο παρουσιάζεται ο ορισμός, ο μαθηματικός τύπος και ο αλγόριθμος της. Επίσης γίνεται αναφορά και σε κάποια παραδείγματα με εικόνες, ώστε να γίνει πιο κατανοητή η φύση της μεθόδου.

Μετά τη θεωρητική παρουσίαση όλων των μοντέλων πρόβλεψης και σύγκρισης χρονοσειρών, στο πέμπτο κεφάλαιο γίνεται μια μικρή αναφορά στην γλώσσα προγραμματισμού R και στο RStudio, τα οποία είναι τα εργαλεία που χρησιμοποιήθηκαν, και στη συνέχεια περιγράφεται αναλυτικώς όλη η μεθοδολογία που αναπτύχθηκε, μαζί με τα αριθμητικά και γραφικά αποτελέσματα.

Στο τελευταίο κεφάλαιο της εργασίας γίνεται μια συνοπτική παρουσίαση όλων των αποτελεσμάτων, εξάγονται συμπεράσματα και προτείνονται θέματα για μελέτη και ανάλυση στο μέλλον.

Κεφάλαιο 2: Προβλέψεις και Χρονοσειρές

2.1 Γενικά για τις προβλέψεις

Οι προβλέψεις είναι απαραίτητες για ένα μεγάλο αριθμό αποφάσεων σχεδιασμού και προγραμματισμού. Υπάρχουν οι αντικειμενικές προβλέψεις (forecasting), όπου έχουν επιστημονική βάση και υπάρχει η δυνατότητα ανάλυσης του σφάλματος τους, και οι υποκειμενικές προβλέψεις (prediction), όπου δεν στηρίζουν επιστημονικά τα αποτελέσματά τους και έτσι υπάρχει περιορισμένη έως και μηδενική ανάλυση των σφαλμάτων τους.

Με βάση το χρονικό ορίζοντα, οι προβλέψεις χωρίζονται σε τρεις κατηγορίες: τις βραχυπρόθεσμες, τις μεσοπρόθεσμες και τις μακροπρόθεσμες. Χρησιμοποιώντας τον εκάστοτε ορίζοντα πρόβλεψης, μπαίνουν συγκεκριμένοι στόχοι, γίνονται συγκεκριμένες ενέργειες και λαμβάνονται κατάλληλες αποφάσεις. Παραδείγματος χάριν, για τον προγραμματισμό παραγωγής ενός εργοστασίου και το χρονικό προγραμματισμό εντολών παραγωγής χρησιμοποιούνται οι βραχυπρόθεσμες αποφάσεις, ενώ για το συγκεντρωτικό προγραμματισμό παραγωγής, τον προγραμματισμό απαιτούμενου προσωπικού και την πολιτική διαχείρισης αποθεμάτων χρησιμοποιούνται οι μεσοπρόθεσμες αποφάσεις. Τέλος, οι μακροπρόθεσμες αποφάσεις χρησιμοποιούνται στην περίπτωση της εισαγωγής νέου προϊόντος ή στην επέκταση εργοστασίου, καθώς απαιτείται η βαθιά γνώση της ζήτησης του προϊόντος.

Οι τεχνικές και οι μέθοδοι προβλέψεων έχουν αναπτυχθεί σε μεγάλο βαθμό τα τελευταία χρόνια, επειδή γίνονται αναγκαίες όλο και περισσότερο στην καθημερινότητά μας, όπως π.χ. στη λειτουργία ενός εργοστασίου. Η αβεβαιότητα που πολλές φορές χαρακτηρίζει τη ζήτηση προϊόντων ή υπηρεσιών και, συνεπώς, τις απαιτήσεις σε μηχανές, υλικά, κεφάλαια, ανθρώπινο δυναμικό και γενικά δυναμικότητα, κατέστησε αναγκαία την ανάπτυξη μεθόδων πρόβλεψης. Στη ραγδαία ανάπτυξη των προβλέψεων παίζουν καθοριστικό ρόλο οι λεγόμενες χρονοσειρές.

2.2 Χαρακτηριστικά των χρονοσειρών

2.2.1 Ορισμός της χρονοσειράς

Με τον όρο χρονοσειρά εννοείται συνήθως μια ακολουθία $\{x_t : t = 0,1,2,...\}$, όπου κάθε x_t εκφράζει την κατάσταση ενός συστήματος κατά την χρονική στιγμή t, δηλαδή διαδοχικές παρατηρήσεις οι οποίες εξελίσσονται στο χρόνο. Παραδείγματα χρονοσειρών είναι:

i) Οι ημερήσιες, αεροπορικές και οδικές, αφίξεις τουριστών στην χώρα μας x_t, με $t = 1,2,3,....$

ii) Ο αριθμός x_t πελατών μέσα σε ένα πολυκατάστημα κατά τη χρονική στιγμή t με $t \in [0,T]$.

iii) Ο συνολικός αριθμός τροχαίων ατυχημάτων x_t κατά μήκος μιας οδικής αρτηρίας στο χρονικό διάστημα $[0,t]$ με $t \geq 0$.

iv) Η ημερήσια κατανάλωση ηλεκτρικού ρεύματος καθώς και η ημερήσια κατανάλωση ύδατος, x_t και y_t αντίστοιχα, σε μια μεγάλη γεωγραφική περιοχή της χώρας με $t = 1,2,3,....$

v) Οι οικονομικές χρονοσειρές, όπως το ετήσιο ακαθάριστο εθνικό προϊόν και το ετήσιο ισοζύγιο εξωτερικών συναλλαγών, x_t και y_t αντίστοιχα, με $t = 1,2,3,....$

vi) Οι μετεωρολογικές χρονοσειρές, όπως η θερμοκρασία περιβάλλοντος και η ατμοσφαιρική πίεση, x_t και y_t αντίστοιχα, σε συγκεκριμένη γεωγραφική περιοχή με γεωγραφικές συντεταγμένες (l,a,h) κατά την χρονική στιγμή t. Εδώ η χρησιμοποιούμενη παράμετρος **t** είναι περισσότερη σύνθετη και συγκεκριμένα **t** = (l,a,h,t)

2.2.2 Κατηγορίες χρονοσειρών

Οι χρονοσειρές διακρίνονται σε συνεχείς και σε διακριτές. Συνεχείς χρονοσειρές είναι αυτές που η τιμή του φαινομένου παρατηρείται συνεχώς, ενώ διακριτές χρονοσειρές είναι αυτές όπου η τιμή του φαινομένου καταγράφεται σε ορισμένα χρονικά διαστήματα. Όπως διαπιστώνει κανείς από τα παραπάνω παραδείγματα, οι χρονοσειρές μπορούν να αφορούν διακριτά μεγέθη x_t σε διακριτό χρόνο t (περίπτωση (i)), διακριτά μεγέθη x_t σε συνεχή χρόνο t (περιπτώσεις (ii) και (iii)), συνεχή μεγέθη x_t σε διακριτό χρόνο (περιπτώσεις (iv) και (v)) και συνεχή μεγέθη x_t σε συνεχή χρόνο t, περίπτωση (vi). Το πρόβλημα είναι η "πρόβλεψη" μελλοντικών τιμών της χρονοσειράς με βάση τις μέχρι σήμερα τιμές της ίδιας χρονοσειράς (περιπτώσεις (i)-(iii)), είτε ακόμα και σε συνδυασμό με τις μέχρι σήμερα τιμές μιας άλλης χρονοσειράς η οποία εξελίσσεται παράλληλα με την πρώτη και επιδρά πάνω σ΄ αυτή (περιπτώσεις (iv)-(vi)), οπότε γίνεται λόγος για πολυμεταβλητές χρονοσειρές. Το σύνολο των δυνατών καταστάσεων ονομάζεται χώρος καταστάσεων και συμβολίζεται με S, ένα (μονοδιάστατο) υποσύνολο του R ή γενικότερα ένα πολυδιάστατο υποσύνολο του R^d, ενώ το σύνολο τιμών του t ονομάζεται παραμετρικός χώρος. Αυτός ο χώρος συμβολίζεται με T και μπορεί επίσης να είναι υποσύνολο του R^k όταν χρειάζεται ένα πολυδιάστατο t για να καθορίσουμε πέραν του χρόνου t και π.χ. γεωγραφικές συντεταγμένες σε χωρο-χρονοσειρές (spatial time series) (βλ. παράδειγμα (vi) παραπάνω). Σημειώνεται ότι οι όροι διακριτά και συνεχή μεγέθη είναι σε αντιστοιχία με τους όρους διακριτές και συνεχείς τυχαίες μεταβλητές.

Μια άλλη κατηγορία που μπορούν να καταταχθούν οι χρονοσειρές είναι ο τρόπος προσδιορισμού των μελλοντικών δεδομένων, δηλαδή η εξάρτηση των διαδοχικών παρατηρήσεων που τις απαρτίζουν. Όταν οι παρατηρήσεις δεν είναι ανεξάρτητες μεταξύ τους, τότε οι μελλοντικές τιμές μπορούν να προσδιοριστούν από τις προηγούμενες. Ένα τέτοιο σύστημα ονομάζεται ντετερμινιστικό. Παρ' όλα αυτά, στις περισσότερες χρονοσειρές στον πραγματικό κόσμο, οι μελλοντικές τιμές καθορίζονται μερικώς από το παρελθόν, επειδή εμπλέκεται και ο "τυχαίος παράγοντας", άρα σε αυτήν την περίπτωση έχουμε τις λεγόμενες στοχαστικές χρονοσειρές.

2.2.3 Στόχοι της ανάλυσης χρονοσειρών

Σε κάθε περίπτωση πάντως, οι χρονοσειρές περιέχουν τις παρελθοντικές τιμές της μεταβλητής των διαδοχικών καταστάσεων στο χρόνο. Έτσι, μπορούν να αναλυθούν ώστε να βγουν συμπεράσματα για την συμπεριφορά της μεταβλητής. Με βάση την πληροφορία από το παρελθόν, επιτρέπεται να προβλεφθούν οι τιμές της στο μέλλον. Η πρόβλεψη, δηλαδή το πώς η ακολουθία των παρατηρήσεων θα συνεχιστεί στο μέλλον, είναι η μεγαλύτερη πρόκληση στην ανάλυση χρονοσειρών. Το ζητούμενο είναι να ακολουθείται μια διαδικασία που θα εξασφαλίσει ότι θα παραχθούν όσο το δυνατόν πιο ακριβείς προβλέψεις, αξιοποιώντας στο έπακρο όλη την διαθέσιμη ιστορική πληροφορία.

Υπάρχουν τρεις κύριοι στόχοι στην ανάλυση χρονοσειρών (Forecasting, Modeling, System Characterization):

- Ο σκοπός του Forecasting είναι να προβλέψει με ακρίβεια την βραχυπρόθεσμη εξέλιξη ενός συστήματος.
- Το Modeling διακρίνει εκείνα τα χαρακτηριστικά που μας δίνουν ξεκάθαρη εικόνα για τη μακροπρόθεσμη συμπεριφορά ενός συστήματος.
- Ο σκοπός του System Characterization είναι να προσπαθήσει με λίγη ή μηδενική πρωτύτερη γνώση να καθορίσει θεμελιώδη χαρακτηριστικά, όπως ο βαθμός ελευθερίας ενός συστήματος ή το σύνολο της τυχαιότητας.

Και οι τρεις στόχοι απαιτούν τα στοιχεία των χρονολογικών σειρών να είναι όσο περισσότερο ακριβή γίνεται.

2.2.4 Διαδικασία για την παραγωγή προβλέψεων μέσω χρονοσειρών

Σύμφωνα με τους Μακρυδάκη, Wheelright και Hyndman (1998), η διαδικασία πρόβλεψης απαρτίζεται από πέντε βασικά βήματα. Το πρώτο βήμα για την παραγωγή προβλέψεων είναι να καθοριστεί πλήρως η φύση του προβλήματος και η χρησιμότητα των προβλέψεων, έτσι ώστε από την αρχή να είναι ξεκάθαρος ο στόχος της όλης διαδικασίας. Αυτό θα βοηθήσει στο να παραχθεί πιο έγκυρο και ακριβές αποτέλεσμα. Το επόμενο και ένα από τα πιο σημαντικά βήματα της διαδικασίας είναι η συλλογή όλων των ιστορικών στοιχείων που χαρακτηρίζουν το πρόβλημα, καθώς και η συλλογή όλων των δεδομένων κάποιου άλλου συστήματος που μπορεί έμμεσα να βοηθήσει στο χαρακτηρισμό του κύριου προβλήματος. Το τρίτο βήμα είναι η ανάλυση της χρονοσειράς που δημιουργείται από τα ιστορικά δεδομένα, δηλαδή η μελέτη του γραφήματος της χρονοσειράς, ώστε να γίνουν αντιληπτά ορισμένα βασικά χαρακτηριστικά της, όπως τα ποιοτικά χαρακτηριστικά ή η μορφή της. Με αυτόν τον τρόπο θα είναι ξεκάθαρο αν υπάρχει σημαντική τάση, εποχικότητα, κυκλικότητα ή ορισμένες ασυνήθιστες τιμές, ενέργεια που οδηγεί στην προσαρμογή των χρονοσειρών, όπως στην αποεποχικοποίηση (απομόνωση εποχικών συνιστωσών), στην απομόνωση όλων των συνιστωσών ή στην αναγνώριση ασυνήθιστων τιμών. Έτσι θα μπορεί να προετοιμαστεί καταλλήλως η χρονοσειρά για τη διαδικασία της πρόβλεψης. Η επιλογή της κατάλληλης μεθόδου πρόβλεψης αποτελεί το επόμενο βήμα και είναι ένα από τα ζητήματα που απασχολεί συνεχώς την επιστημονική κοινότητα. Οι βασικοί παράγοντες που καθορίζουν την επιλογή της μεθόδου είναι ο σκοπός πρόβλεψης, η περίοδος και ο ορίζοντας πρόβλεψης (βραχυπρόθεσμος, μεσοπρόθεσμος, μακροπρόθεσμος), το κόστος της μεθόδου (απαιτήσεις της μεθόδου, χρήση ειδικού εξοπλισμού), η επιζητούμενη ακρίβεια, η απλότητα, η ευκολία εφαρμογής και τα διαθέσιμα στοιχεία. Το πέμπτο και τελευταίο στάδιο της διαδικασίας πρόβλεψης είναι η χρήση και η αξιολόγηση των μοντέλων που επιλέχτηκαν να χρησιμοποιηθούν.

2.2.5 Ποιοτικά χαρακτηριστικά χρονοσειρών

Η μελέτη του γραφήματος μιας χρονοσειράς οδηγεί στην αναγνώριση των βασικών ποιοτικών χαρακτηριστικών της, τα οποία είναι η τάση, η εποχικότητα, η κυκλικότητα και η τυχαιότητα. Αυτά τα χαρακτηριστικά παίζουν καθοριστικό ρόλο στη μετέπειτα πορεία της διαδικασίας και της παραγωγής προβλέψεων.

Η **τάση** μπορεί να είναι ανοδική, φθίνουσα ή μηδενική και δηλώνει την μακροπρόθεσμη μεταβολή του μέσου επιπέδου των τιμών της χρονοσειράς. Βέβαια για να εκτιμηθεί με ορθό τρόπο αν κάποια μεταβολή είναι μακροπρόθεσμη, πρέπει να υπάρχει ένας ικανός αριθμός παρατηρήσεων, ώστε να μην μπορεί να παρερμηνευτεί το κατάλληλο μήκος της περιόδου που εμφανίζεται η τάση.

Η **εποχικότητα** είναι ένα μοτίβο διακύμανσης των τιμών της χρονοσειράς που εμφανίζεται σε σταθερά διαστήματα και μικρότερα του ενός έτους (βδομάδα, μήνας, τρίμηνο, κ.α.). Η διακύμανση αυτή είναι περιοδική και συνήθως ορατή και κατανοητή, οπότε είναι σχετικά εύκολο να μετρηθεί. Έτσι με ορισμένες τεχνικές μπορούν να υπολογιστούν οι δείκτες εποχικότητας, να απομονωθούν και να προκύψει η τελική αποεποχικοποιημένη χρονοσειρά.

Η **κυκλικότητα** είναι ένα χαρακτηριστικό παρόμοιο με την εποχικότητα, με τη διαφορά ότι η περιοδικότητα και η διάρκεια κύκλου δεν εμφανίζουν σταθερότητα. Πιο συγκεκριμένα, η κυκλικότητα εμφανίζεται κατά περιόδους, το μήκος των οποίων συνήθως ξεπερνά το ένα έτος, και χαρακτηρίζεται ως μια κυματοειδή μεταβολή λόγω ειδικών εξωγενών συνθηκών. Μερικά παραδείγματα είναι το Ακαθάριστο Εθνικό Προϊόν, οι τιμές των μετοχών κ.α., λόγω της διακύμανσης της οικονομίας κάθε χώρας ανά διαστήματα.

Η **τυχαιότητα** (ή αλλιώς **μη κανονικές διακυμάνσεις)** αποτελεί τον υπολειπόμενο παράγοντα μετά την απομόνωση των τριών πρώτων παραγόντων. Χαρακτηρίζεται ως το τυχαίο γεγονός κατά την εξέλιξη μιας χρονοσειράς.

Οι **ασυνέχειες** αποτελούν υποσύνολο των μη κανονικών διακυμάνσεων και είναι οι απομονωμένες και απότομες αλλαγές που παρατηρούνται σε μια χρονοσειρά. Αυτές οι αλλαγές είναι δύσκολο να προβλεφθούν από τα ιστορικά δεδομένα και συνήθως χαρακτηρίζουν έναν απρόβλεπτο παράγοντα. Ανάλογα με το αν έχουν περιοδική ή μόνιμη συμπεριφορά, χωρίζονται σε outliers (ή special events) και σε level-shifts, αντίστοιχα.

- Τα outliers (ή special events) αντιπροσωπεύουν ασυνήθιστες παρατηρήσεις που οφείλονται σε τυχαία ή απρόβλεπτα γεγονότα, όπως για παράδειγμα μια απρόβλεπτη καταιγίδα στα μέσα του καλοκαιριού που οδηγεί στην καταστροφή καλλιεργειών. Χρειάζεται ιδιαίτερη προσοχή στην ανάλυση και στην ερμηνεία τους, ώστε να μπορέσουν να αντιμετωπιστούν αποτελεσματικά.
- Τα level-shifts αποτελούν τις απότομες αλλαγές στο μέσο επίπεδο των χρονοσειρών, όπως για παράδειγμα στη μείωση των πωλήσεων μιας επιχείρησης με την εμφάνιση ενός ανταγωνιστή της.

2.2.6 Ιστορική ανασκόπηση χρονοσειρών

Η τεχνική των προβλέψεων έχει εμφανιστεί από τα πολύ παλιά χρόνια. Βέβαια τις περισσότερες φορές, οι προβλέψεις ήταν ανακριβείς, καθώς δε στηρίζονταν σε ιστορικά γεγονότα, αλλά παράγονταν κυρίως με τη τεχνική της κριτικής πρόβλεψης, χωρίς να κρατείται αμερόληπτη στάση. Κατά τη διάρκεια του 17ου αιώνα, η ανάπτυξη των επιχειρησιακών προβλέψεων θεωρείται από τις σπουδαιότερες καινοτομίες στον χώρο αυτό. Τα επόμενα 300 χρόνια, σημειώθηκε ιδιαίτερα σημαντική και ουσιώδη πρόοδος στις μεθόδους πρόβλεψης που βασίζονται σε ιστορικά δεδομένα. Με την πάροδο των χρόνων και την εξέλιξη της τεχνολογίας και ιδιαίτερα της πληροφορικής, καθώς επίσης και την αναγκαιότητα να παράγονται ορθές προβλέψεις, ιδιαίτερα στον κλάδο των επιχειρήσεων, οι τεχνικές που χρησιμοποιούνται για να παραχθεί η πρόβλεψη έχουν αλλάξει ριζικά. Η στατιστική ανάλυση των δεδομένων των χρονοσειρών και οι βάσεις για την πρόβλεψη τους τέθηκαν κυρίως κατά τη διάρκεια του μεσοπολέμου (εργασίες των Yule, Wald, κλπ). Οι επίσημες τεχνικές προβλέψεων παράχθηκαν λίγο μετά το 2ο Παγκόσμιο Πόλεμο στις σκανδιναβικές χώρες και σιγά σιγά εξαπλώθηκαν στο Ηνωμένο Βασίλειο στις αρχές της δεκαετίας του ‘50 και σε άλλες αναπτυγμένες οικονομικά χώρες τη δεκαετία του ‘60. Κατά τη δεκαετία του 1950, κυρίως από τους ερευνητές του Cowles Foundation Group, αναπτύχθηκαν τα οικονομετρικά υποδείγματα ταυτόχρονων αλληλεξαρτημένων εξισώσεων για την εκτίμηση των οποίων χρησιμοποιήθηκαν μέθοδοι μέγιστης πιθανοφάνειας. Τα συγκεκριμένα υποδείγματα βρήκαν εφαρμογή κυρίως στη μακροοικονομία. Τη δεκαετία του 1960, αναπτύχθηκαν οι τεχνικές εκθετικής εξομάλυνσης χρονοσειρών από επιχειρησιακούς ερευνητές (Holt, Winters) με σκοπό κυρίως την πρόβλεψη. Τη δεκαετία του 1970, δόθηκε μεγάλη ώθηση σε εφαρμογές από τους Box και Jenkins με την ανάπτυξη της μεθοδολογίας για τη δημιουργία εμπειρικών υποδειγμάτων χρονοσειρών. Σύμφωνα με τη μεθοδολογία αυτή, η αναζήτηση των πληροφοριών για τη δημιουργία των υποδειγμάτων κατευθύνεται στα ίδια τα διαθέσιμα δεδομένα, δηλαδή για τη δημιουργία των υποδειγμάτων δεν είναι απαραίτητο να στηριχθούμε σε κάποια προϋπάρχουσα θεωρία. Οι προβλέψεις με υποδείγματα Box-Jenkins ήταν τις περισσότερες φορές καλύτερες από αυτές των μεγάλης κλίμακας μακροοικονομικών υποδειγμάτων. Όμως οι μακροοικονομολόγοι τα χαρακτήρισαν ακατάλληλα για οικονομική πολιτική, επειδή ήταν εμπειρικά και κατά συνέπεια στερούνταν θεωρητικής βάσης. Μέχρι και τη δεκαετία του 1980 παρατηρείται έντονη αντιπαλότητα μεταξύ των ερευνητών, όπως π.χ. αυτής των Johnston και McGraw Hill το 1986. Όμως από τότε και έπειτα, παρατηρείται σύγκλιση απόψεων και παύση αντιπαλότητας. Τα τελευταία 25 χρόνια έχει παρατηρηθεί ραγδαία ανάπτυξη στον τομέα των στατιστικών

προβλέψεων, χάρη στην εξέλιξη της τεχνολογίας και των λογισμικών, και στις κριτικές προβλέψεις. Αυτή η κατηγορία προβλέψεων εξαρτάται αποκλειστικά από τον ανθρώπινο παράγοντα και τα δεδομένα που απαιτεί είναι η διαίσθηση, η κρίση και η συσσωρευμένη γνώση και εμπειρία.

Κεφάλαιο 3: Τεχνικές Προβλέψεων

3.1 Γενικά μοντέλα πρόβλεψης

Η πρόβλεψη χρονοσειρών αποτελεί ένα από τα καυτά ζητήματα της επιστήμης των προβλέψεων. Ο σκοπός είναι η ελαχιστοποίηση του σφάλματος της πρόβλεψης και της πραγματικής τιμής. Έχουν αναπτυχθεί πολλά μοντέλα που παράγουν προβλέψεις, όπου το καθένα απαρτίζεται από το δικό του αλγόριθμο. Αρχικά θα παρουσιαστούν τα γενικά μοντέλα προβλέψεων και έπειτα θα αναλυθούν πιο συγκεκριμένα οι μέθοδοι πρόβλεψης που χρησιμοποιούνται στην παρούσα διπλωματική εργασία.

Τα γενικά μοντέλα πρόβλεψης που χρησιμοποιούνται για την παραγωγή των προβλέψεων είναι τα εξής:

- Υποκειμενικές ή Διαισθητικές Μέθοδοι:
  – Συνεντεύξεις, δημοσκοπήσεις, κ.α.
  – Μέθοδος DELPHI
- Μέθοδοι βασισμένες στο μέσο όρο παλαιότερων δεδομένων:
  – Κινούμενοι μέσοι (moving averages)
  – Εκθετική εξομάλυνση (exponential smoothing)
- Μοντέλα παλινδρόμησης σε ιστορικά δεδομένα:
  – Προεκβολή τάσης (trend extrapolation)
- Αιτιοκρατικά (causal) ή οικονομετρικά μοντέλα
- Ανάλυση χρονοσειρών με χρήση στοχαστικών δεδομένων:
  – Ανάλυση Box-Jenkins

3.2 Μέθοδοι πρόβλεψης

3.2.1 Μέθοδος Naïve

Η μέθοδος Naive, ή αλλιώς απλοϊκή ή αφελής μέθοδος, αποτελεί την απλούστερη στατιστική μέθοδο, καθώς θεωρεί ότι η πρόβλεψη για την κάθε χρονική περίοδο t ισούται με την τιμή της προηγούμενης περιόδου t-1 των δεδομένων, δηλαδή:

$$F_t = Y_{t-1}$$

όπου, t η χρονική περίοδος

Y_{t-1} η πραγματική τιμή των δεδομένων της προηγούμενης περιόδου t-1

F_t η πρόβλεψη των δεδομένων τη χρονική περίοδο t

Τα πλεονεκτήματα της μεθόδου είναι η απλότητα και το χαμηλός κόστος. Λόγω της τόσο απλοποιημένης φύσης της μεθόδου, είναι λογικό η Naive να μην παράγει ακριβείς προβλέψεις. Παρ' όλα αυτά μπορεί να χρησιμοποιηθεί για την πρόβλεψη έως και μιας περιόδου στο μέλλον ή ως σημείο αναφοράς απόδοσης (benchmark) για άλλες πιο πολύπλοκες μεθόδους. Η μέθοδος είναι αποδοτική όταν ο μέσος όρος, η τάση ή τα εποχιακά φαινόμενα είναι σταθερά και η τυχαιότητα είναι μικρή. Εάν τα τυχαία λάθη είναι μεγάλα, τότε χρησιμοποιώντας την τιμή της τελευταίας περιόδου για την πρόβλεψη της επόμενης περιόδου μπορεί να οδηγήσει σε υψηλές διακυμάνσεις οι οποίες θα φέρουν λανθασμένα αποτελέσματα.

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση naive() της βιβλιοθήκης "forecast" της γλώσσας R, όπου λαμβάνει ως είσοδο την αποεποχικοποιημένη χρονοσειρά προς προέκταση και τον ορίζοντα πρόβλεψης h, και επιστρέφει τις προβλέψεις.

3.2.2 Seasonal Naïve

Η μέθοδος Seasonal Naive αποτελεί μια προσαρμογή της απλοϊκής μεθόδου Naive, καθώς λαμβάνει υπόψη την εποχιακότητα των δεδομένων. Πιο συγκεκριμένα, θεωρεί ότι η πρόβλεψη για την κάθε χρονική περίοδο t ισούται με την τιμή της αντίστοιχης περιόδου του προηγούμενου κύκλου εποχιακότητας, δηλαδή:

$$F_t = Y_{t-pos}$$

όπου, t η χρονική περίοδος

pos ο αριθμός των περιόδων ενός κύκλου εποχιακότητας

Y_{t-pos} η πραγματική τιμή των δεδομένων της αντίστοιχης περιόδου του προηγούμενου

F_t η πρόβλεψη των δεδομένων τη χρονική περίοδο t

Η μέθοδος αυτή είναι ιδιαίτερα χρήσιμη σε δεδομένα με πολύ έντονη εποχιακότητα. Παρ' όλα αυτά, δε μπορεί να αναγνωρίσει την τάση στα δεδομένα, οπότε επιλέγεται να χρησιμοποιείται μόνο σε περιπτώσεις έντονης εποχιακότητας ή σαν σημείο αναφοράς απόδοσης άλλων μεθόδων.

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση snaive() της βιβλιοθήκης "forecast" της γλώσσας R, όπου λαμβάνει ως είσοδο τη χρονοσειρά προς προέκταση και τον ορίζοντα πρόβλεψης h και επιστρέφει τις προβλέψεις.

3.2.3 Random Walk with Drift

Η μέθοδος Random Walk with Drift αποτελεί και αυτή μια προσαρμογή της απλοϊκής μεθόδου Naive, όπως η Seasonal Naive. Η διαφορά τους είναι ότι η μέθοδος αυτή αντιλαμβάνεται την τάση στα δεδομένα, αλλά δε μπορεί να εντοπίσει την εποχιακότητα. Πιο συγκεκριμένα, θεωρεί ότι η πρόβλεψη για την κάθε χρονική περίοδο t ισούται με το άθροισμα της τιμής της προηγούμενης περιόδου t-1 των δεδομένων, του όρου σφάλματος (γνωστός ως white noise) και μιας παραμέτρου drift, δηλαδή:

$$F_t = Y_{t-1} + Z_t + c$$

όπου, t η χρονική περίοδος

c η παράμετρος drift

Y_{t-1} η πραγματική τιμή των δεδομένων της προηγούμενης περιόδου t-1

Z_t ο όρος σφάλματος (white noise)

F_t η πρόβλεψη των δεδομένων τη χρονική περίοδο t

Ο όρος σφάλματος (white noise) είναι μια τυπική μεταβλητή με μέσο όρο 0 και διακύμανση 1. Το drift συμπεριφέρεται όπως η τάση, καθώς όταν c>0, η μέθοδος θα εμφανίσει μια ανοδική τάση. Σύμφωνα με τον Rowland, η μέθοδος αυτή παρουσιάζει μια ντετερμινιστική, καθώς και μια στοχαστική τάση. Το μοντέλο Random Walk with Drift χρησιμοποιείται σαν ένα γραμμικό μοντέλο πρόβλεψης. Παρατηρούμε ότι όταν c=0 και μηδενικό όρο σφάλματος, η μέθοδος ταυτίζεται πλήρως με το απλοϊκό μοντέλο Naive.

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση rwf() της βιβλιοθήκης “forecast” της γλώσσας R, όπου λαμβάνει ως είσοδο την αποεποχικοποιημένη χρονοσειρά προς προέκταση και τον ορίζοντα πρόβλεψης h και επιστρέφει τις προβλέψεις.

3.2.4 Απλή Εκθετική Εξομάλυνση - Μοντέλο Σταθερού Επιπέδου (Simple Exponential Smoothing)

Το μοντέλο σταθερού επιπέδου, ή αλλιώς απλής εκθετικής εξομάλυνσης, ανήκει στις μεθόδους εξομάλυνσης και περιγράφεται από τις παρακάτω εξισώσεις:

$$e_t = Y_t - F_t$$

$$S_t = S_{t-1} + a * e_t$$

$$F_{t+1} = S_t$$

όπου, t η χρονική περίοδος

Y_t η πραγματική τιμή των δεδομένων τη χρονική περίοδο t

F_t η πρόβλεψη των δεδομένων τη χρονική περίοδο t

e_t το σφάλμα (απόκλιση πραγματικής τιμής και πρόβλεψης) τη χρονική περίοδο t

S_t το επίπεδο της χρονοσειράς τη χρονική περίοδο t

a ο συντελεστής εξομάλυνσης που λαμβάνει τιμές στο διάστημα [0,1]

Το συγκεκριμένο μοντέλο υποθέτει ότι τα δεδομένα χαρακτηρίζονται από σταθερό μέσο όρο και κατ' επέκταση από απουσία τάσης. Χρησιμοποιείται κυρίως σε χρονοσειρές που παρουσιάζουν υψηλή τυχαιότητα ή θόρυβο. Χαρακτηριστικά της μεθόδου είναι ότι η πρόβλεψη για κάθε χρονική περίοδο είναι ίση με το επίπεδο της χρονοσειράς και ότι το επίπεδο της επόμενης περιόδου ισούται με το άθροισμα του επιπέδου της προηγούμενης περιόδου και του γινομένου του σφάλματος επί το συντελεστή εξομάλυνσης a.

Τα δύο προβλήματα που αντιμετωπίζει κάποιος στην εφαρμογή της μεθόδου της απλής εκθετικής εξομάλυνσης είναι η επιλογή του αρχικού επιπέδου S_0 και η επιλογή του κατάλληλου συντελεστή εξομάλυνσης a. Αντιλαμβάνεται κανείς ότι η κακή αρχικοποίηση του αρχικού επιπέδου μπορεί να οδηγήσει σε ανακριβή αποτελέσματα. Συνήθως ως αρχικό επίπεδο S_0 χρησιμοποιείται ένα από τα εξής:

- Ο μέσος όρος όλων των παρατηρήσεων
- Ο μέσος όρος των v πρώτων παρατηρήσεων
- Η πρώτη παρατήρηση
- Το σταθερό επίπεδο από τη γραμμή του μοντέλου της απλής γραμμικής παλινδρόμησης

Ο συντελεστής εξομάλυνσης παίρνει τιμές στο διάστημα [0,1] και μπορεί να παρατηρηθεί ότι όταν a=0, τότε η κάθε πρόβλεψη ισούται με το αρχικό επίπεδο S_0 και όταν a=1, τότε το μοντέλο ταυτίζεται με την απλοϊκή μέθοδο Naive και ότι η πρόβλεψη για την κάθε χρονική περίοδο t ισούται με την τιμή της προηγούμενης περιόδου t-1 των δεδομένων. Οπότε κρίνεται αναγκαίο να επιλέγεται κάθε φορά η βέλτιστη τιμή αυτού του συντελεστή. Οι δυο παράγοντες που καθιστούν βέλτιστο το συντελεστή είναι οι εξής:

- Ο θόρυβος στα δεδομένα της χρονοσειράς. Η έντονη ύπαρξη θορύβου συνεπάγεται την αλλοίωση των προβλέψεων, οπότε χρησιμοποιείται μικρή τιμή του συντελεστή, για να μην υπάρχει υπερβολική αντίδραση στο θόρυβο.
- Η σταθερότητα του μέσου όρου της χρονοσειράς. Η μεταβολή του μέσου όρου απαιτεί μεγάλη τιμή του συντελεστή, ώστε οι προβλέψεις να παρακολουθούν τις μεταβολές των δεδομένων, ενώ αντιθέτως η σταθερότητα του μέσου όρου απαιτεί μικρή τιμή του συντελεστή.

Παρ' όλα αυτά, τα υπολογιστικά συστήματα βοηθούν στην επιλογή του κατάλληλου συντελεστή εξομάλυνσης. Το πιο διαδεδομένο κριτήριο για τη βελτιστοποίηση του είναι η ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος (MSE). Οπότε με την εφαρμογή μιας επαναληπτικής διαδικασίας στον υπολογιστή, μπορεί εύκολα να καθοριστεί ο βέλτιστος συντελεστής.

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση ses() της βιβλιοθήκης "forecast" της γλώσσας R, όπου λαμβάνει ως είσοδο την αποεποχικοποιημένη χρονοσειρά προς προέκταση και τον ορίζοντα πρόβλεψης h και επιστρέφει τις προβλέψεις.

3.2.5 Εκθετική Εξομάλυνση Γραμμικής Τάσης – Μοντέλου Γραμμικής Τάσης (Holt Exponential Smoothing)

Το μοντέλο γραμμικής τάσης, ή αλλιώς εκθετικής εξομάλυνσης γραμμικής τάσης, ανήκει και αυτό στις μεθόδους εξομάλυνσης και περιγράφεται από τις παρακάτω εξισώσεις:

$$e_t = Y_t - F_t$$

$$S_t = S_{t-1} + T_{t-1} + a * e_t$$

$$T_t = T_{t-1} + b * e_t$$

$$F_{t+m} = S_t + m * T_t$$

όπου, t η χρονική περίοδος

Y_t η πραγματική τιμή των δεδομένων τη χρονική περίοδο t

F_t η πρόβλεψη των δεδομένων τη χρονική περίοδο t

T_t η τάση των δεδομένων τη χρονική περίοδο t

e_t το σφάλμα (απόκλιση πραγματικής τιμής και πρόβλεψης) τη χρονική περίοδο t

S_t το επίπεδο της χρονοσειράς τη χρονική περίοδο t

m ο χρονικός ορίζοντας της πρόβλεψης

a ο συντελεστής εξομάλυνσης επιπέδου που λαμβάνει τιμές στο διάστημα [0,1]

b ο συντελεστής εξομάλυνσης τάσης που λαμβάνει τιμές στο διάστημα [0,1]

Το μοντέλο εκθετικής εξομάλυνσης γραμμικής τάσης εισήχθη από τον Holt το 1957 και αποτελεί μια επέκταση του μοντέλου απλής εκθετικής εξομάλυνσης, με τη διαφορά ότι θεωρεί την ύπαρξη και της τάσης. Χρησιμοποιείται σε χρονοσειρές που χαρακτηρίζονται από τη συνιστώσα της τάσης και καθώς η συνιστώσα αυτή παρατηρείται συχνά στα δεδομένα, μπορεί να βγει το συμπέρασμα ότι είναι πιο αποτελεσματικό σε σχέση με το μοντέλο σταθερού επιπέδου. Χαρακτηριστικά της μεθόδου είναι ότι η πρόβλεψη για κάθε χρονική περίοδο είναι ίση με το άθροισμα του επιπέδου της χρονοσειράς και του γινομένου της τάσης επί το χρονικό ορίζοντα. Επίσης ότι το επίπεδο της επόμενης περιόδου ισούται με το άθροισμα του επιπέδου

της προηγούμενης περιόδου, της τάσης της προηγούμενης περιόδου του γινομένου του σφάλματος επί το συντελεστή εξομάλυνσης a.

Το πρόβλημα που αντιμετωπίζει κάποιος στην εφαρμογή της μεθόδου της εκθετικής εξομάλυνσης γραμμικής τάσης είναι η επιλογή του αρχικού επιπέδου S_0, η επιλογή της αρχικής τάσης T_0 και οι επιλογές των κατάλληλων συντελεστών εξομάλυνσης a, b. Αντιλαμβάνεται κανείς ότι η κακή αρχικοποίηση των αρχικών τιμών μπορεί να οδηγήσει σε ανακριβή αποτελέσματα. Η επιλογή του S_0 γίνεται ομοίως με το μοντέλο σταθερού επιπέδου, ενώ ως αρχική τάση T_0 χρησιμοποιείται ένα από τα εξής:

- Διαφορά δεύτερης και πρώτης παρατήρησης ($Y_2 – Y_1$).
- Διαφορά v-στής και πρώτης παρατήρησης διαιρεμένης με v-1 ($(Y_v – Y_1) / (v-1)$).
- Η σταθερά της κλίσης από τη γραμμή του μοντέλου της απλής γραμμικής παλινδρόμησης

Οι συντελεστές εξομάλυνσης παίρνουν τιμές στο διάστημα [0,1] και κρίνεται αναγκαίο να επιλέγονται κάθε φορά οι βέλτιστες τιμές για αυτούς τους συντελεστές. Εδώ είναι σημαντικό να επισημανθεί ότι η βέλτιστη τιμή του συντελεστή του επιπέδου (a) είναι μεγαλύτερη από τη βέλτιστη τιμή του συντελεστή της τάσης (b) και ο λόγος είναι ότι η τιμή του επιπέδου είναι συνήθως πολύ μεγαλύτερη από την τιμή της τάσης.

Τα υπολογιστικά συστήματα βοηθούν στην επιλογή των κατάλληλων συντελεστών εξομάλυνσης. Το πιο διαδεδομένο κριτήριο για τη βελτιστοποίηση τους είναι η ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος (MSE). Οπότε με την εφαρμογή της γραμμικής αναζήτησης υπολογίζεται ο καλύτερος συνδυασμός των συντελεστών εξομάλυνσης, δηλαδή ο συνδυασμός αυτός που παράγει το μικρότερο σφάλμα.

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση holt (damped = F) της βιβλιοθήκης “forecast” της γλώσσας R, όπου λαμβάνει ως είσοδο την αποεποχικοποιημένη χρονοσειρά προς προέκταση και τον ορίζοντα πρόβλεψης h και επιστρέφει τις προβλέψεις.

3.2.6 Εκθετική Εξομάλυνση Μη Γραμμικής Τάσης – Μοντέλο Μη Γραμμικής Τάσης (Damped Exponential Smoothing)

Το μοντέλο μη γραμμικής τάσης, ή αλλιώς εκθετικής εξομάλυνσης μη γραμμικής τάσης, ανήκει και αυτό στις μεθόδους εξομάλυνσης και περιγράφεται από τις παρακάτω εξισώσεις:

$$e_t = Y_t - F_t$$

$$S_t = S_{t-1} + \varphi * T_{t-1} + a * e_t$$

$$T_t = \varphi * T_{t-1} + b * e_t$$

$$F_{t+m} = S_t + \sum_{i=1}^{m} \varphi^i * T_t$$

όπου, t η χρονική περίοδος

Y_t η πραγματική τιμή των δεδομένων τη χρονική περίοδο t

F_t η πρόβλεψη των δεδομένων τη χρονική περίοδο t

T_t η τάση των δεδομένων τη χρονική περίοδο t

e_t το σφάλμα (απόκλιση πραγματικής τιμής και πρόβλεψης) τη χρονική περίοδο t

S_t το επίπεδο της χρονοσειράς τη χρονική περίοδο t

m ο χρονικός ορίζοντας της πρόβλεψης

a ο συντελεστής εξομάλυνσης επιπέδου που λαμβάνει τιμές στο διάστημα [0,1]

b ο συντελεστής εξομάλυνσης τάσης που λαμβάνει τιμές στο διάστημα [0,1]

φ η παράμετρος διόρθωσης της τάσης που λαμβάνει τιμές στο διάστημα [0,+οο)

Το μοντέλο εκθετικής εξομάλυνσης μη γραμμικής τάσης δημιουργήθηκε από τους Gardner και McKenzieto το 1985 και αποτελεί προέκταση του μοντέλου της εκθετικής εξομάλυνσης γραμμικής τάσης. Η διαφοροποίηση τους παρατηρείται στο γεγονός ότι το μοντέλο μη γραμμικής τάσης έχει τη δυνατότητα, μέσω της προσθήκης της παραμέτρου διόρθωσης της τάσης (φ), να προσαρμόζεται και σε δεδομένα με μη γραμμική τάση. Λόγω αυτής της προσαρμογής, το μοντέλο αυτό είναι ιδιαιτέρως αποτελεσματικό συγκριτικά με τις

άλλες μεθόδους που έχουν αναφερθεί και ειδικά σε προβλέψεις με μεγάλο χρονικά ορίζοντα ή σε περιπτώσεις όπου η επιλογή κάποιου συγκεκριμένου μοντέλου για την περιγραφή κάθε χρονοσειράς δεδομένων είναι αδύνατη. Χαρακτηριστικό της μεθόδου είναι ότι η πρόβλεψη για κάθε χρονική περίοδο είναι ένας μη γραμμικός συνδυασμός του επιπέδου και της τάσης των δεδομένων λόγω της παραμέτρου διόρθωσης της τάσης.

Η παράμετρος αυτή, εν γένει, δεν έχει άνω και κάτω όριο, παρ' όλα αυτά κρίνεται αναγκαίο να επιλέγεται η κατάλληλη τιμή της για κάθε διαφορετική περίπτωση δεδομένων. Το μοντέλο μη γραμμικής τάσης μπορεί να πάρει πολλές μορφές, ανάλογα με την τιμή της παραμέτρου φ.

- Αν φ = 0, τότε προκύπτει το μοντέλο της απλής εκθετικής εξομάλυνσης (Ses).
- Αν φ = 1, τότε προκύπτει το μοντέλο της γραμμικής τάσης (Holt).
- Αν 0 < φ < 1, τότε προκύπτει το μοντέλο της φθίνουσας τάσης (Damped).
- Αν φ > 1, τότε προκύπτει το μοντέλο της εκθετικής τάσης.

Τα κύρια προβλήματα που αντιμετωπίζει κάποιος στην εφαρμογή της μεθόδου της εκθετικής εξομάλυνσης μη γραμμικής τάσης είναι δύο. Καταρχάς, η εσφαλμένη βελτιστοποίηση της παραμέτρου φ μπορεί να οδηγήσει σε παραγωγή προκατειλημμένων προβλέψεων. Έτσι συνηθίζεται ο περιορισμός αυτής της παραμέτρου στο διάστημα [0,1], ώστε να αποφεύγονται οι υπεραισιόδοξες προβλέψεις που παράγει το μοντέλο εκθετικής τάσης. Το δεύτερο πρόβλημα είναι η επιλογή του αρχικού επιπέδου S_0, η επιλογή της τάσης T_0 και οι επιλογές των κατάλληλων συντελεστών εξομάλυνσης a, b. Η επικρατέστερη μέθοδος για την επιλογή του S_0 και T_0 είναι η εφαρμογή της γραμμικής παλινδρόμησης, ενώ το πιο διαδεδομένο κριτήριο για τη βελτιστοποίηση των συντελεστών εξομάλυνσης a, b είναι η ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος (MSE).

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση holt (damped = T) της βιβλιοθήκης "forecast" της γλώσσας R, όπου λαμβάνει ως είσοδο την αποεποχικοποιημένη χρονοσειρά προς προέκταση και τον ορίζοντα πρόβλεψης h και επιστρέφει τις προβλέψεις.

3.2.7 Μέθοδος Theta

Η μέθοδος Theta αναπτύχθηκε από τους Ασημακόπουλο Β. και Νικολόπουλο Κ. το 2000 και αποτελεί μια μονοδιάστατη μέθοδο πρόβλεψης, που προσεγγίζει την αποσύνθεση με καινοτόμο τρόπο για την παραγωγή προβλέψεων με μεγαλύτερη ακρίβεια συγκριτικά με τις άλλες μεθόδους. Πιο συγκεκριμένα, βασίζεται στη μεταβολή των τοπικών καμπυλοτήτων μιας χρονοσειράς με τη βοήθεια της παραμέτρου θ. Με αυτόν τον τρόπο, η αρχική χρονοσειρά αποσυντίθεται σε τουλάχιστον δύο καινούργιες χρονοσειρές (γραμμές Theta), οι οποίες προεκτείνονται στο μέλλον σαν ξεχωριστές χρονοσειρές και έπειτα προβλέπεται η τελική πρόβλεψη ως ο συνδυασμός αυτών των παραγόμενων προβλέψεων. Οι γραμμές Theta που παράγονται έχουν την ιδιότητα να διατηρούν τη μέση τιμή και την κλίση της αρχικής χρονοσειράς, αλλά όχι τις τοπικές καμπυλότητες και τη διακύμανση. Για διάφορες τιμές της θ, μπορεί να γίνει καλύτερη προσέγγιση της μακροπρόθεσμης ή βραχυπρόθεσμης συμπεριφοράς των δεδομένων. Όσο μικρότερη η τιμή της, τόσο μεγαλύτερος ο βαθμός μείωσης των καμπυλοτήτων. Πιο συγκεκριμένα έχουμε τις εξής περιπτώσεις:

- Για θ < 1, τονίζονται περισσότερο τα μακροπρόθεσμα χαρακτηριστικά της χρονοσειράς.
- Για θ > 1, τονίζονται περισσότερο τα βραχυπρόθεσμα χαρακτηριστικά της χρονοσειράς.
- Για θ = 0, η χρονοσειρά ταυτίζεται με την ευθεία της απλής γραμμικής παλινδρόμησης.
- Για θ = -1, η χρονοσειρά αντιστοιχεί με τη συμμετρική της αρχικής χρονοσειράς ως προς την ευθεία της γραμμικής παλινδρόμησης.

Το μοντέλο Theta που αποσυνθέτει τη χρονοσειρά σε δύο γραμμές Theta με παραμέτρους θ=0 και θ=2 ονομάζεται κλασική μέθοδος Theta και εφαρμόστηκε στο διαγωνισμό προβλέψεων Μ3 για να δώσει προβλέψεις για τις 3003 χρονοσειρές του διαγωνισμού, όπου παρήγαγε εντυπωσιακά αποτελέσματα. Ο μαθηματικός τύπος που περιγράφει τη κλασική μέθοδο Theta φαίνεται παρακάτω:

$$Y_t = \frac{1}{2} * (Y_t^{\theta=0} + Y_t^{\theta=2})$$

Τα βήματα που ακολουθούνται για την εφαρμογή της κλασικής μεθόδου Theta είναι τα εξής:

- Έλεγχος της εκάστοτε χρονοσειράς για τυχόν στατιστικά σημαντική εποχιακότητα.
- Αποεποχικοποίηση μέσω της κλασικής μεθόδου πολλαπλασιαστικής αποσύνθεσης, εφόσον αποδειχθεί εποχιακότητα.
- Αποσύνθεση της εκάστοτε χρονοσειράς σε δύο γραμμές Theta, τη θ=0 όπου ταυτίζεται με την ευθεία γραμμικής παλινδρόμησης και τη θ=2.
- Προέκταση της γραμμής θ=0 με τον κλασικό τρόπο και της γραμμής θ=2 με τη μέθοδο της απλής εκθετικής εξομάλυνσης.
- Συνδυασμός των προβλέψεων αυτών των γραμμών με ίσα βάρη
- Εποχικοποίηση των παραγόμενων προβλέψεων με τους δείκτες εποχιακότητας που υπολογίστηκαν κατά την αποεποχικοποίηση.

Εδώ είναι σημαντικό να αναφερθεί ότι με βάση τον χρονικό ορίζοντα πρόβλεψης, επιλέγονται και τα κατάλληλα βάρη, ώστε να συνδυαστούν οι γραμμές Theta και να παραχθεί η τελική πρόβλεψη.

Το 2008, ο Νικολακόπουλος Κ. και οι συνεργάτες του παρουσίασαν μια απλοποιημένη διαδικασία υπολογισμού των γραμμών Theta, όπου ο υπολογισμός πραγματοποιείται με το άθροισμα της γραμμής Theta με θ=0 (ευθεία γραμμικής παλινδρόμησης) και θ-φορές το σφάλμα της αρχικής χρονοσειράς από τη γραμμή Theta με θ=0. Ο μαθηματικός τύπος φαίνεται παρακάτω:

$$e_t = Y_t - LRL_t$$

$$\text{Theta Line}(\theta)_t = Y_t^{\theta} = LRL_t + \theta * e_t$$

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση thetaf() της βιβλιοθήκης “forecast” της γλώσσας R, όπου λαμβάνει ως είσοδο την αποεποχικοποιημένη χρονοσειρά προς προέκταση και τον ορίζοντα πρόβλεψης h και επιστρέφει τις προβλέψεις.

3.2.8 Ολοκληρωμένα Αυτοπαλινδρομικά Μοντέλα Κινητού Μέσου Όρου – ARIMA (AutoRegressive Integrated-Moving Average)

Τα ολοκληρωμένα αυτοπαλινδρομικά μοντέλα κινητού μέσου όρου (ARIMA) είναι στοχαστικά μαθηματικά μοντέλα που περιγράφουν, αναλύουν και προβλέπουν τη διαχρονική εξέλιξη κάποιου φυσικού μεγέθους που εξαρτάται από μη ντετερμινιστικούς παράγοντες. Για να εφαρμοστούν σε μια χρονοσειρά, πρέπει εκείνη να είναι στάσιμη και διακριτή. Χρησιμοποιούνται για προβλέψεις των μελλοντικών τιμών, λαμβάνοντας υπόψη τις παρελθοντικές τιμές. Για παράδειγμα, για μια τυχαία χρονοσειρά, τα μοντέλα ARIMA βασίζονται στην παραδοχή της αλληλεξάρτησης μεταξύ των τιμών που λαμβάνει η χρονοσειρά για τις διάφορες χρονικές στιγμές. Εξάλλου, η πλειοψηφία των φυσικών μεγεθών δε δίνει τη δυνατότητα της πλήρης επίγνωσης όλων των παραγόντων που επηρεάζουν την εξέλιξη τους στο χρόνο, οπότε είναι ιδιαίτερα δύσκολη η περιγραφή τους με ένα ντετερμινιστικό μοντέλο. Από την άλλη μεριά, τα μη ντετερμινιστικά μοντέλα έχουν τη δυνατότητα να περιγράφουν την εξέλιξη ενός μεγέθους, υπολογίζοντας τη πιθανότητα η τιμή του να βρίσκεται σε κάποιο διάστημα. Εδώ είναι ιδιαίτερα σημαντικό να αναφερθεί, ότι τα μοντέλα ARIMA μπορούν να μοντελοποιήσουν ένα μεγάλο εύρος εποχιακών δεδομένων.

Τα μοντέλα ARIMA δύνανται να εκφραστούν σαν ένας γραμμικός συνδυασμός του τυχαίου παράγοντα (τυχαίο σφάλμα ή σφάλμα πρόβλεψης), των τιμών του μεγέθους σε προηγούμενες χρονικές στιγμές και κάποιων άλλων στοχαστικών παραγόντων. Βέβαια σε πραγματικά δεδομένα είναι ιδιαίτερα δύσκολο να εντοπιστούν αυτοί οι τρεις παράγοντες, αλλά μπορούν να προσεγγιστούν αποτελεσματικά.

Τα συγκεκριμένα μοντέλα έχουν μελετηθεί εκτεταμένα από τους Box και Jenkins (1970), οι οποίοι έχουν προτείνει μια μεγάλη ομάδα αλγεβρικών μοντέλων πρόβλεψης, όπου το κάθε ένα μπορεί να εφαρμοστεί σε κατάλληλη χρονοσειρά ώστε να βελτιστοποιείται η τελική πρόβλεψη. Οι Box και Jenkins έχουν παράγει μια μεθοδολογία, που περιλαμβάνει τα εξής τρία στάδια:

Ταυτοποίηση (Identification): Περιλαμβάνει τον καθορισμό του αριθμού d (διαφορές που χρησιμοποιούνται για την μετατροπή μιας διαδικασίας σε στάσιμη σε περίπτωση που δεν είναι), του αριθμού p (τάξη της αυτοπαλίνδρομης διαδικασίας AR) και του αριθμού q (τάξη της διαδικασίας κινητού μέσου MA).

Εκτίμηση (Estimation): Περιλαμβάνει τον καθορισμό των p παραμέτρων της αυτοπαλίνδρομης διαδικασίας και των q παραμέτρων της διαδικασίας κινητού μέσου. Επίσης εκτελείται έλεγχος της τάξεως συγκρίνοντας το υπόδειγμα ARIMA με άλλο μεγαλύτερης τάξεως (Κριτήρια AIC,SBC).

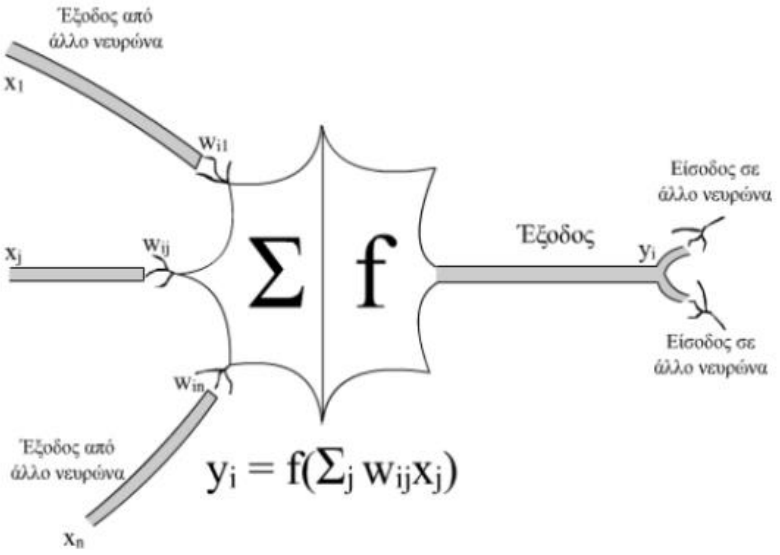
Προβλέψεις (Forecasting): Γνωρίζοντας το εκτιμώμενο υπόδειγμα και τις υπάρχουσες πληροφορίες για μια χρονοσειρά, μπορεί να παραχθεί η πρόβλεψη για τις επόμενες ζητούμενες περιόδους.

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση forecast(auto.arima(), h) της βιβλιοθήκης “forecast” της γλώσσας R, όπου λαμβάνει ως είσοδο τη χρονοσειρά προς προέκταση και τον ορίζοντα πρόβλεψης h και επιστρέφει τις προβλέψεις.

3.2.9 Μέθοδος Neural Network Time Series Forecasts

Η μέθοδος πρόβλεψης χρονοσειρών μέσω τεχνητών νευρωνικών δικτύων εφαρμόζεται και χρησιμοποιείται ευρέως στην παραγωγή προβλέψεων και είναι μια διαφορετική οπτική σε σύγκριση με τις άλλες μεθόδους. Αυτός ο τρόπος πρόβλεψης έχει μελετηθεί από πολλούς ερευνητές όπως τους Hornik et al. (1989), Cybenko (1989), Zhang et al. (1998) και τους Balkin and Ord (2000), σύμφωνα με τους οποίους, τα νευρωνικά δίκτυα αποτελούν έναν ισχυρό ανταγωνιστή των παραδοσιακών μεθόδων πρόβλεψης και μια σημαντική προσθήκη στην εργαλειοθήκη της πρόβλεψης χρονοσειρών.

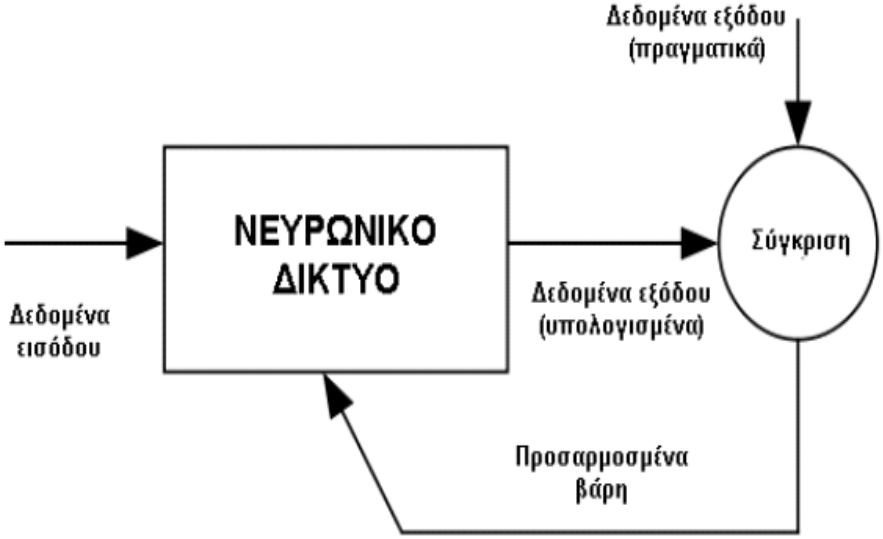
Τεχνητό νευρωνικό δίκτυο ονομάζεται ένα κύκλωμα διασυνδεδεμένων νευρώνων, το οποίο αποτελεί ένα αλγοριθμικό κατασκεύασμα και έχει ως σκοπό την επίλυση κάποιου μαθηματικού προβλήματος. Πιο συγκεκριμένα, είναι ένα δίκτυο που περιλαμβάνεται από υπολογιστικούς κόμβους (νευρώνες) διασυνδεδεμένους μεταξύ τους. Το σχηματικό διάγραμμα ενός νευρωνικού δικτύου φαίνεται παρακάτω:



Εικόνα 3.2.9.1: Σχηματικό διάγραμμα ενός νευρωνικού δικτύου

Υπάρχουν τρεις τύποι νευρώνων: οι νευρώνες εισόδου, οι νευρώνες εξόδου και οι υπολογιστικοί νευρώνες (ή κρυμμένοι νευρώνες). Οι νευρώνες εισόδου λαμβάνουν ένα σύνολο εισόδων (ή από το περιβάλλον είτε από άλλους νευρώνες). Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με το αντίστοιχο συνοπτικό βάρος και υπολογίζουν το ολικό άθροισμα των γινομένων. Το άθροισμα αυτό τροφοδοτείται ως όρισμα στη συνάρτηση μεταφοράς, η οποία μεταφέρει το αποτέλεσμα στο νευρώνα εξόδου (περιβάλλον ή είσοδος σε άλλους νευρώνες του δικτύου.

Για να μπορέσει να χρησιμοποιηθεί ένα δίκτυο, πρέπει πρώτα να περάσει τη διαδικασία εκπαίδευσης, δηλαδή τη συνεχή τροποποίηση των βαρών, έως ότου επιτευχθεί η βέλτιστη σύγκλιση μεταξύ των πραγματικών δεδομένων εξόδου και αυτών που προβλέπει (παράγει) το δίκτυο. Το σχηματικό διάγραμμα της διαδικασίας εκπαίδευσης ενός τεχνητού νευρωνικού δικτύου φαίνεται παρακάτω:



Εικόνα 3.2.9.2: Διαδικασία εκπαίδευσης ενός νευρωνικού δικτύου

Σε αυτό το σημείο είναι σημαντικό να αναφερθούν μερικά πλεονεκτήματα και μειονεκτήματα των Νευρωνικών Δικτύων.

Μερικά από τα πλεονεκτήματα είναι:

- Η μη γραμμική δομή τους, που τους δίνει την δυνατότητα να αντιλαμβάνονται τόσο γραμμικές όσο και μη γραμμικές συσχετίσεις μεταξύ των δεδομένων.
- Η προσαρμοστικότητα τους σε διαφορετικά περιβάλλοντα, καθώς διαθέτουν την ικανότητα να μαθαίνουν μέσω παραδειγμάτων.
- Η ταχύτητα επεξεργασίας, που οφείλεται στην παράλληλη επεξεργασία των κόμβων.
- Η ικανότητα τους να εντοπίζουν την εποχιακότητα και την τάση μέσα στα δεδομένα, που συνεπάγεται τη μη αναγκαιότητα για αποεποχικοποίηση πριν παραχθούν οι προβλέψεις.

Ενώ μερικά από τα μειονεκτήματα τους είναι:

- Βαθιά γνώση και εμπειρία του μελετητή, καθώς στη μοντελοποίηση του προβλήματος προσφέρεται υψηλός βαθμός ελευθερίας.
- Ιδιαίτερα δύσκολη η ερμηνεία των αποτελεσμάτων, επειδή τα τεχνητά νευρωνικά δίκτυα λειτουργούν ουσιαστικά ως μαύρα κουτιά, που δέχονται τις εισόδους, τις επεξεργάζονται και παράγουν τις εξόδους.

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση forecast(nnetar(), h) της βιβλιοθήκης “forecast” της γλώσσας R, όπου λαμβάνει ως είσοδο τη χρονοσειρά προς προέκταση και τον ορίζοντα πρόβλεψης h και επιστρέφει τις προβλέψεις.

3.2.10 Μοντέλο Χώρου Καταστάσεων των Μεθόδων Εκθετικής Εξομάλυνσης (Exponential Smoothing State Space Model)

Το μοντέλο του χώρου καταστάσεων των μεθόδων της εκθετικής εξομάλυνσης αποτελεί μια ευρεία οικογένεια των μεθόδων της εκθετικής εξομάλυνσης, όπως για παράδειγμα η απλή εκθετική εξομάλυνση ή η εκθετική εξομάλυνση γραμμικής τάσης. Η επιλογή της κάθε μεθόδου πραγματοποιείται με βάση την ταξινόμηση που δημιουργήθηκε από τους Hyndman et al. (2008). Η ταξινόμηση αυτή βασίζεται σε τρεις μεταβλητές που χαρακτηρίζουν την κάθε μέθοδο (σφάλμα, τάση, εποχιακότητα), όπου κάθε μια από αυτές τις τρεις μεταβλητές μπορεί να χαρακτηριστεί ως αθροιστική (additive), πολλαπλασιαστική (multiplicative), ουδείς (none) και αυτόματη (automatically). Πιο συγκεκριμένα ισχύει ότι “A”=additive, “M”=multiplicative, “N”=none, “Z”=automatically. Η ταξινόμηση πραγματοποιείται σύμφωνα με μια συμβολοσειρά τριών γραμμάτων, όπου το πρώτο γράμμα περιγράφει τον τύπο του σφάλματος, το δεύτερο γράμμα περιγράφει τον τύπο της τάσης και το τρίτο γράμμα περιγράφει τον τύπο της εποχιακότητας. Μερικά χαρακτηριστικά παραδείγματα είναι:

- η συμβολοσειρά “ANN” συμβολίζει την απλή εκθετική εξομάλυνση με προσθετικό τύπο σφαλμάτων.
- η συμβολοσειρά “MAM” συμβολίζει την πολλαπλασιαστική μέθοδο Holt-Winters με πολλαπλασιαστικό τύπο σφαλμάτων.
- η συμβολοσειρά “AAN” συμβολίζει την εκθετική εξομάλυνση γραμμικής τάσης με αθροιστικό τύπο σφαλμάτων.

Ο πίνακας των διάφορων συνδυασμών συμβολοσειρών φαίνεται παρακάτω:

Trend component	**Seasonal component**		
	N (None)	**A (additive)**	**M (Multiplicative)**
N (None)	NN	NA	NM
A (additive)	AN	AA	AM
M (Multiplicative)	MN	MA	MM
A_d(Additive Damped)	A_dN	A_dA	A_dM
M_d(Multiplicative damped)	M_dN	M_dA	M_dM

Εικόνα 3.2.10.1: Οι διάφοροι συνδυασμοί συμβολοσειρών για τη μέθοδο πρόβλεψης ETS

Η μέθοδος αυτή είναι ιδιαίτερα χρήσιμη και αποτελεσματική, καθώς μπορεί να αντιμετωπίσει κάθε διαφορετικό συνδυασμό τάσης και εποχιακότητας. Πιο συγκεκριμένα, με έναν ελεγχόμενο τρόπο, μπορεί με βάση το μοτίβο της χρονοσειράς που προβλέπει, να αλλάζει τη δομή της μεθόδου που χρησιμοποιεί για να παράγει τις προβλέψεις, αφού από τη φύση της μπορεί να μετατραπεί σε διάφορες μεθόδους πρόβλεψης. Οι προβλέψεις που παράγει είναι πολύ ακριβείς συγκριτικά με τις παραδοσιακές μεθόδους, και αυτό επαληθεύεται από το γεγονός της πολύ καλής απόδοσης που είχε στο διαγωνισμό προβλέψεων M3-Competition.

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση forecast(ets(), h) της βιβλιοθήκης “forecast” της γλώσσας R, όπου λαμβάνει ως είσοδο τη χρονοσειρά προς προέκταση και τον ορίζοντα πρόβλεψης h και επιστρέφει τις προβλέψεις.

3.2.11 Μοντέλο Χώρου Καταστάσεων των Μεθόδων Εκθετικής Εξομάλυνσης με Μετασχηματισμό Box-Cox, Σφάλματα ARMA και Δείκτες Τάσης και Εποχιακότητας (Exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal components)

Ο μετασχηματισμός Box-Cox δημιουργήθηκε από τους Box και Cox (1964) και είναι ικανός να αντιμετωπίσει μη γραμμικά δεδομένα. Σε συνδυασμό με το μοντέλο ARMA (AutoRegressive Moving Average model), μπορούν να δημιουργήσουν ανεξαρτησία μεταξύ των δεδομένων μιας χρονοσειράς. Έπειτα, οι De Livera, Hyndman, Snyder (2010) παρουσίασαν μια εναλλακτική μέθοδο, ώστε να παράγουν προβλέψεις για πολύπλοκες εποχιακές χρονοσειρές και αποδείξαν ότι είναι ιδιαίτερα αποτελεσματική. Αυτή η μέθοδος είναι γνωστή ως BATS (Exponential Smoothing State Space model with Box-Cox transformation, ARMA errors, Trend and Seasonal Components). Παρ' όλα αυτά, η μέθοδος BATS δεν τα πήγαινε καλά όταν η εποχιακότητα είχε μεγαλύτερη συχνότητα και ήταν ιδιαίτερα πολύπλοκη. Οπότε το 2011, προτάθηκε η μέθοδος TBATS (Trigonometric Exponential Smoothing State Space model with Box-Cox transformation, ARMA errors, Trend and Seasonal Components), όπου η τριγωνομετρική έκφραση της εποχιακότητας μπορεί να δώσει στη χρονοσειρά την ευελιξία να αντιμετωπίζει σύνθετες εποχικότητες και επίσης μειώνει τις παραμέτρους του μοντέλου όταν η συχνότητα είναι πολύ υψηλή. Επιπροσθέτως, η TBATS είναι ικανή να χειρίζεται μη ακέραιες εποχιακές συχνότητες και τα ληφθέντα δεδομένα δε χρειάζονται κανονικοποίηση, συγκριτικά με την BATS.

Η μέθοδος TBATS χρησιμοποιεί έναν συνδυασμό μεταξύ όρων Fourier, το μοντέλο χώρου καταστάσεων των μεθόδων εκθετικής εξομάλυνσης και το μετασχηματισμό Box-Cox με έναν τελειώς αυτοματοποιημένο τρόπο. Παρ' όλα αυτά, λόγω του τρόπου λειτουργίας της, είναι πιθανόν σε μερικές περιπτώσεις να εξάγει ανακριβή αποτελέσματα. Οι παράμετροι που ζητά είναι οι (omega, p, q, phi, <m1, k1>…<mj, kj>), όπου omega είναι η παράμετρος Box-Cox, phi η παράμετρος απόσβεσης (damping parameter), p,q είναι οι παράμετροι του ARMA, $m_1 \ldots m_j$ οι εποχιακές περίοδοι που χρησιμοποιούνται στο μοντέλο και $k_1 \ldots kj$ οι αντίστοιχοι όροι Fourier.

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση forecast(tbats(), h) της βιβλιοθήκης "forecast" της γλώσσας R, όπου λαμβάνει ως είσοδο τη χρονοσειρά προς προέκταση και τον ορίζοντα πρόβλεψης h και επιστρέφει τις προβλέψεις.

3.2.12 Πρόβλεψη Μέσω Αποσύνθεσης STL (STL Decomposition)

Το μοντέλο πρόβλεψης, μέσω της αποσύνθεσης STL, χρησιμοποιεί μια οποιαδήποτε μη εποχιακή μέθοδο για να παράγει προβλέψεις στα αποεποχικοποιημένα δεδομένα, που έχουν προκύψει μέσω της αποσύνθεσης STL, και στη συνέχεια τα ξανα-εποχικοποιεί με τους κατάλληλους εποχιακούς δείκτες.

Η αποσύνθεση STL είναι μια πολύπλευρη και πολυχρησιμοποιούμενη μέθοδος, η οποία αποσυνθέτει μια χρονοσειρά στα δομικά της στοιχεία. Η μέθοδος αυτή αναπτύχθηκε από τους R.B. Cleveland, W.S. Cleveland, J.E. McRae και I. Terpenning (1990). Τα αρχικά STL αποτελούν ένα ακρωνύμιο της πρότασης “Seasonal and Trend decomposition using Loess”, όπου η Loess είναι μια μέθοδος για να εκτιμούμε μη γραμμικές σχέσεις. Η μέθοδος αποσύνθεσης STL διαχωρίζει τη χρονοσειρά σε δείκτες τάσης, δείκτες εποχιακότητας και υπολειπόμενους δείκτες (θόρυβος), ενώ η κυκλικότητα, εάν υπάρχει, εμπεριέχεται στην τάση. Η μέθοδος χρησιμοποιείται μόνο σε αθροιστικές αποσυνθέσεις (additive decompositions), οπότε θεωρείται ότι:

$$Y_t = S_t + T_t + E_t$$

όπου, t η χρονική περίοδος

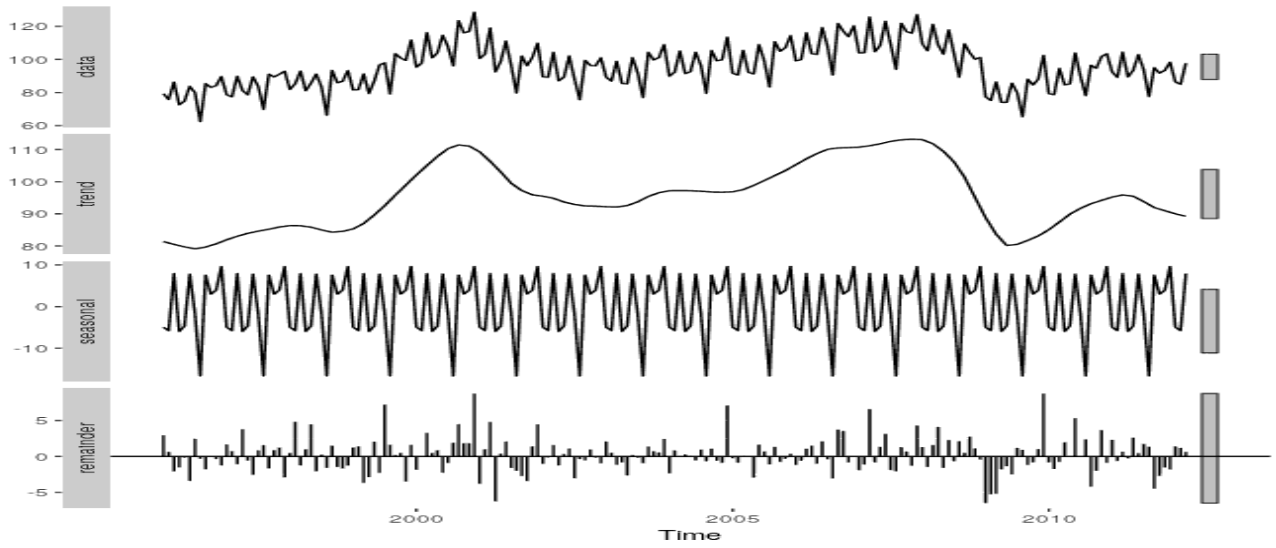
Y_t η πραγματική τιμή των δεδομένων τη χρονική περίοδο t

S_t ο δείκτης εποχιακότητας τη χρονική περίοδο t

T_t ο δείκτης τάσης-κύκλου τη χρονική περίοδο t

E_t ο υπολειπόμενος δείκτης (θόρυβος) τη χρονική περίοδο t

Παρακάτω φαίνεται ένα παράδειγμα συνδυαστικού γραφήματος τεσσάρων χρονοσειρών που αποτυπώνει την αποσύνθεση STL στα δεδομένα (data), δηλαδή την εξαγωγή των δεικτών τάσης-κύκλου, εποχιακότητας και του υπολοίπου (remainder):



Εικόνα 3.2.12.1: Παραγόμενες χρονοσειρές από την αποσύνθεση STL

Σε αντίθεση με το μοναδικό μειονέκτημα της μεθόδου STL (μόνο αθροιστικές αποσυνθέσεις), έχει αρκετά πλεονεκτήματα συγκριτικά με άλλες μεθόδους, τα οποία είναι τα εξής:

- Μπορεί να αντιμετωπίσει κάθε τύπο εποχιακότητας, ενώ άλλες μέθοδοι αρκούνται στα μηνιαία και στα τριμηνιαία δεδομένα.
- Η ομαλότητα των δεικτών τάσης-κύκλου μπορεί να ελέγχεται από το χρήστη.
- Οι εποχιακοί δείκτες μπορούν να αλλάζουν σταδιακά και ο ρυθμός αλλαγής μπορεί να ρυθμίζεται από το χρήστη.
- Είναι ανθεκτική σε απομακρυσμένα δεδομένα (outliers), δηλαδή δε μπορούν να επηρεάσουν τις εκτιμήσεις των δεικτών τάσης-κύκλου και εποχιακότητας, αλλά μπορούν να επηρεάσουν τους υπολειπόμενους δείκτες (remainder component).

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση stlm() της βιβλιοθήκης “forecast” της γλώσσας R, όπου λαμβάνει ως είσοδο τη χρονοσειρά προς προέκταση και επιστρέφει τα αποεποχικοποιημένα δεδομένα εφαρμόζοντας την αποσύνθεση STL. Έπειτα παράγει και ξανα-εποχικοποιεί τις προβλέψεις με τους κατάλληλους εποχιακούς δείκτες.

3.3 Σφάλματα Προβλέψεων

Η αξιολόγηση μιας μεθόδου πρόβλεψης επιτυγχάνεται μέσω της μέτρησης της ακρίβειας των προβλέψεων που παράγει. Η ακρίβεια υπολογίζεται μέσω ορισμένων τύπων σφάλματος, οι οποίοι θα παρουσιαστούν παρακάτω. Το σφάλμα ορίζεται ως η διαφορά μεταξύ της πραγματικής τιμής και της πρόβλεψης για μια περίοδο και ο βασικός στόχος του είναι να τείνει στο μηδέν. Θεωρώντας ως Y_i την πραγματική τιμή και ως F_i την τιμή πρόβλεψης, παρατίθενται οι βασικότεροι τύποι σφαλμάτων:

- Βασικός Τύπος Σφάλματος (Error)

$$e_i = Y_i - F_i$$

- Μέσο Σφάλμα (Mean Error)

$$ME = \frac{1}{n}\sum_{i=1}^{n}(Y_i - F_i)$$

- Μέσο Απόλυτο Σφάλμα (Mean Absolute Error)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - F_i|$$

- Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - F_i)^2$$

- Ρίζα Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - F_i)^2}$$

- Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (Mean Absolute Percentage Error)

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - F_i}{Y_i}\right| \cdot 100\%$$

- Συμμετρικό Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (symmetric Mean Absolute Percentage Error)

$$\mathrm{sMAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{2 \cdot (Y_i - F_i)}{Y_i + F_i}\right| \cdot 100\%$$

Στην παρούσα διπλωματική εργασία χρησιμοποιείται μόνο το συμμετρικό μέσο απόλυτο ποσοστιαίο σφάλμα (sMAPE) για τον υπολογισμό των σφαλμάτων, καθώς είναι ιδιαίτερα χρήσιμο για χρονοσειρές με διαφορετικό επίπεδο μέσης τιμής και επίσης δεν είναι τόσο ευαίσθητο σε outliers. Ο συγκεκριμένος τύπος σφάλματος είναι και ο καθοριστικός παράγοντας επιλογής μεθόδου για το διαγωνισμό που εκτελείται μεταξύ των μεθόδων που αναφέρθηκαν προηγουμένως στο κεφάλαιο, αφού κάθε φορά επιλέγεται εκείνη η μέθοδος που παράγει το μικρότερο σφάλμα sMAPE. Χρησιμοποιείται η συνάρτηση smape() της βιβλιοθήκης “Metrics” της γλώσσας R, όπου λαμβάνει ως ορίσματα τις πραγματικές μελλοντικές τιμές και τις παραγόμενες προβλέψεις και επιστρέφει το σφάλμα.

Κεφάλαιο 4: Μέθοδοι Σύγκρισης Χρονοσειρών

4.1 Εισαγωγή στη σύγκριση χρονοσειρών

Στο κεφάλαιο αυτό αναλύεται και εξηγείται ένα από τα βασικότερα κομμάτια της διπλωματικής εργασίας, που είναι η σύγκριση και εν τέλει η ομαδοποίηση των χρονοσειρών βάσει της ομοιότητας τους.

Κατά καιρούς έχουν αναπτυχθεί πολλές μέθοδοι που συγκρίνουν την ομοιότητα διαφόρων χρονοσειρών. Για παράδειγμα, από την πιο απλή που είναι μέσω της απόστασης των ένα-προς-ένα σημείων των χρονοσειρών με χρήση της Ευκλείδειας απόστασης, έως μέσω της ανάλυσης της εικόνας της εκάστοτε χρονοσειράς με αλγορίθμους μηχανικής μάθησης, ώστε να παραχθεί μια αριθμητική ακολουθία ή ένα διάγραμμα που να αποτυπώνει τη μορφή της στο χρόνο.

Στη συγκεκριμένη διπλωματική εργασία χρησιμοποιήθηκαν τέσσερις μέθοδοι σύγκρισης χρονοσειρών, η LCSS, EDR, ERP, DTW. Η κάθε μέθοδος χρησιμοποιεί τον δικό της ατομικό αλγόριθμο και παράγει έναν αριθμό που αποτυπώνει μια εκτίμηση για το πόσο όμοιες είναι οι χρονοσειρές που εξετάζονται. Η πρώτη μέθοδος, λόγω της φύσης του αλγορίθμου της, όσο πιο μεγάλο αριθμητικό αποτέλεσμα δώσει, τόσο πιο όμοιες είναι οι χρονοσειρές. Αντιθέτως στις άλλες τρεις μεθόδους, το μικρότερο αποτέλεσμα δίνει και τη μεγαλύτερη ομοιότητα των χρονοσειρών που συγκρίνονται.

Στη συνέχεια του κεφαλαίου θα αναλυθούν εκτενέστερα οι αλγόριθμοι των τεσσάρων μεθόδων, ώστε να γίνει σαφές πως εξήλθε το συμπέρασμα για το βαθμό ομοιότητας των χρονοσειρών και τελικώς για την ομαδοποίηση τους.

4.2 Longest Common SubSequence (LCSS)

Η μέθοδος Longest Common SubSequence (Das et al. - 1997, Vlachos et al. - 2003) εντοπίζει τη μέγιστη κοινή υποακολουθία που εμπεριέχεται στις ακολουθίες προς σύγκριση (συνήθως δύο). Υποακολουθία είναι κάθε μικρότερη ακολουθία που βρίσκεται μέσα στην αρχική, οπότε για μια ακολουθία μήκους n, υπάρχουν 2^n πιθανές υποακολουθίες. Για παράδειγμα, οι ακολουθίες “abc”, “abd”, “bce” είναι υποακολουθίες της “abcde”. Η μέθοδος LCSS χρησιμοποιείται ευρέως στην επιστήμη των υπολογιστών, όπως για παράδειγμα στη σύγκριση δεδομένων, στην υπολογιστική γλωσσολογία και στη βιοφαρμακευτική. Το αποτέλεσμα της μεθόδου αυτής, δηλαδή η μέγιστη κοινή υποακολουθία χρησιμοποιείται συνήθως ως συγκριτικό μέτρο. Τα σημεία που απαρτίζουν την υποακολουθία δε χρειάζεται να βρίσκονται σε συνεχόμενες θέσεις, αλλά να έχουν ίδια σχετική θέση στην αρχική ακολουθία. Για παράδειγμα, μεταξύ των ακολουθιών “ABCDOIT” και “ACQRIEOT”, η μέγιστη υποακολουθία είναι η “ACIOT” με μήκος 5. Το μήκος είναι ο καθοριστικός παράγοντας στη σύγκριση ακολουθιών, καθώς δίνει μια ξεκάθαρη εικόνα για το αν μια ακολουθία μοιάζει περισσότερο με μια άλλη συγκριτικά με μια τρίτη.

Το μοντέλο LCSS υπολογίζεται με τη βοήθεια του δυναμικού προγραμματισμού, καθώς η δομή του του επιτρέπει να “σπάει” συνεχώς σε μικρότερα απλούστερα υποπροβλήματα, μέχρις ότου δημιουργηθεί ένα αρκετά απλό πρόβλημα. Ο δυναμικός προγραμματισμός δίνει τη δυνατότητα να απομνημονεύονται οι λύσεις των υποπροβλημάτων σε έναν πίνακα, ώστε να μην υπολογίζονται ξανά και ξανά κάθε φορά που περνάμε στο επόμενο υποπρόβλημα. Έτσι καταφέρνουμε να μειώσουμε πολύ την πολυπλοκότητα της μεθόδου, η οποία είναι O(M*N), όπου M, N τα μήκη των δύο ακολουθιών που θα συγκριθούν.

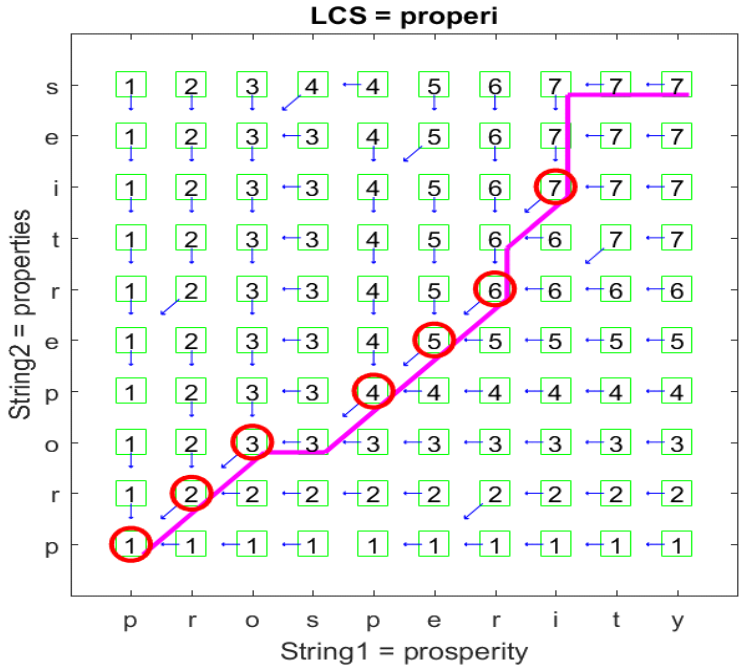
Στην παρούσα διπλωματική επεξεργαζόμαστε αριθμητικά δεδομένα και όχι συμβολοσειρές, οπότε είναι αναγκαίο να εισαχθεί ένα κατώφλι που θα καθορίζει αν οι αντίστοιχοι αριθμοί που συγκρίνονται είναι ίσοι ή όχι. Άρα, έχοντας ως δεδομένα δύο χρονοσειρές **X**, **Y**, όπου $X=(x_0,\ldots,x_{M-1})$ και $Y=(y_0,\ldots,y_{N-1})$ και το κατώφλι **ε**, ο αναδρομικός τύπος υπολογισμού της μέγιστης κοινής υποακολουθίας φαίνεται παρακάτω:

$$LCSS(X,Y) = \begin{cases} 0 & \text{if } M-1=0 \text{ or } N-1=0 \\ LCSS(Rest(X), Rest(Y)) + 1 & \text{if } |x_0 - y_0| \leq \epsilon \\ max\{LCSS(Rest(X), Y), LCSS(X, Rest(Y))\} & \text{otherwise} \end{cases}$$

Όπου Rest(X) = x_{i-1}, Rest(Y) = y_{i-1} (το i εκφράζει το αριθμό της εκάστοτε επανάληψης).

Παρατηρείται ότι το αποτέλεσμα υπολογίζεται με δυναμικό προγραμματισμό, καθώς σε κάθε επανάληψη η συνάρτηση LCSS καλεί τον εαυτό της. Αρχικά, στην πρώτη γραμμή και στην πρώτη στήλη συμπληρώνονται μηδενικά και έπειτα ξεκινάει να τρέχει επαναληπτικά η μέθοδος, κρατώντας στον πίνακα LCSS όλες τις τιμές που έχουν υπολογιστεί, ώστε να ξαναχρησιμοποιηθούν στην επόμενη επανάληψη. Επίσης είναι σημαντικό να αναφερθεί ότι δε μετριέται η ευκλείδεια απόσταση μεταξύ των εκάστοτε τιμών (όπως κάνουν άλλες μέθοδοι, που θα αναφερθούν στη συνέχεια), αλλά καθορίζεται άμα αυτές οι τιμές θεωρούνται ίσες σύμφωνα με το κατώφλι. Αν ναι, τότε το μήκος της υποακολουθίας αυξάνεται κατά ένα. Έτσι δημιουργείται ο πίνακας LCSS, οποίος στη θέση (M, N) δείχνει ποιο είναι το μήκος της μέγιστης κοινής υποακολουθίας μεταξύ δύο χρονοσειρών.

Στην παρακάτω εικόνα, φαίνεται ένα πολύ ξεκάθαρο παράδειγμα σύγκρισης των συμβολοσειρών "prosperity" και "properties", όπου η μέγιστη κοινή υποακολουθία είναι η "properi" με μήκος=7.



Εικόνα 4.1.1: Παράδειγμα σύγκρισης χρονοσειρών μέσω της μεθόδου LCSS

Ομοίως εφαρμόζεται και σε αριθμητικές ακολουθίες, έχοντας το κατώφλι **ε** ως καθοριστικό παράγοντα ισότητας των εκάστοτε στοιχείων.

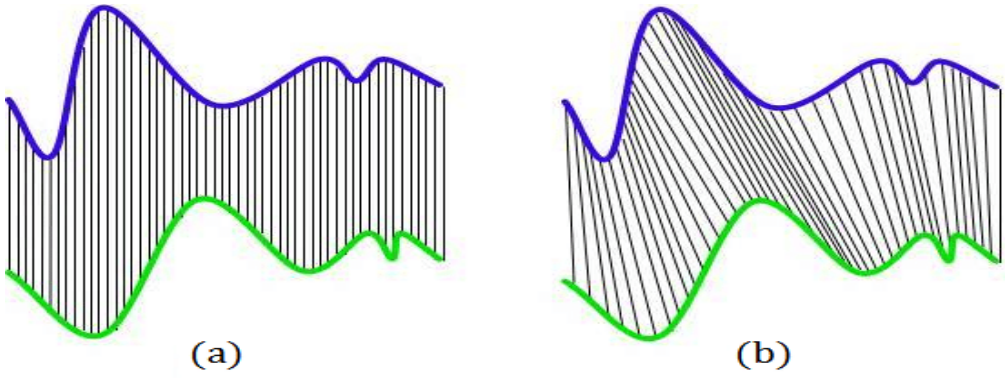
Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση LCSSDistance() της βιβλιοθήκης “TSdist” της γλώσσας R, όπου λαμβάνει ως είσοδο τις δύο χρονοσειρές που θα συγκριθούν και το κατώφλι **ε** (το οποίο το θέσαμε ίσο με **ε** = 0.1) κάτω από το οποίο θεωρούνται ίσες οι δύο εκάστοτε τιμές που συγκρίνονται και επιστρέφει το μήκος της μέγιστης κοινής υποακολουθίας. Προαιρετικά, μπορεί να εισαχθεί ως είσοδος και ένας περιορισμός, ώστε να μην αντιστοιχίζονται σημεία που βρίσκονται χρονικά μακριά.

4.3 Dynamic Time Warping (DTW)

Η μέθοδος DTW (Velichko and Zagoruyko - 1970, Sakoe and Chiba - 1971) είναι ένας αλγόριθμος μέτρησης της ομοιότητας μεταξύ δύο ετερογενών χρονοσειρών. Η μέθοδος υπολογίζει μια απόσταση που αντικατοπτρίζει σε μεγάλο βαθμό τη σχέση μεταξύ τους, γι' αυτό συνήθως αποκαλείται «απόσταση DTW». Χρησιμοποιείται κατά κόρον σε διάφορες επιστήμες, όπως η ρομποτική, η φαρμακευτική, η αναγνώριση προτύπων, ήχων και εικόνων, η σεισμολογία, η ανθρωπολογία, τα οικονομικά κ.α.

Η απόσταση DTW μεταξύ δύο ακολουθιών είναι το άθροισμα των αποστάσεων μεταξύ εκείνων των ζευγών σημείων τους, ούτως ώστε η συνολική απόσταση να είναι η ελάχιστη. Χρησιμοποιείται ο δυναμικός προγραμματισμός για να βρεθούν τα συγκεκριμένα αντίστοιχα σημεία που ελαχιστοποιούν την απόσταση, οπότε η πολυπλοκότητα της μεθόδου έχει μειωθεί σε O(M*N), όπου M, N τα μήκη των δύο ακολουθιών που συγκρίνονται. Στη μέθοδο DTW, όλα τα σημεία της μιας ακολουθίας πρέπει να αντιστοιχηθούν με τουλάχιστον ένα σημείο από την άλλη ακολουθία και επίσης το πρώτο και το τελευταίο σημείο της μιας ακολουθίας πρέπει να «ματσάρει» με το πρώτο και το τελευταίο σημείο της άλλης ακολουθίας, αντίστοιχα. Σε αντίθεση με την Ευκλείδεια Απόσταση που απαιτεί ένα-προς-ένα αντιστοίχιση, η DTW επιτρέπει την οποιαδήποτε αντιστοίχιση, με την προϋπόθεση ότι η αντιστοίχιση θα είναι μονοτονικά αυξανόμενη. Ωστόσο, η DTW είναι ιδιαίτερα ευαίσθητη στα απομακρυσμένα σημεία (outliers).

Στο παρακάτω σχήμα φαίνεται η αντιστοίχιση δύο τυχαίων χρονοσειρών με χρήση a) της Ευκλείδειας Απόστασης και b) απόστασης DTW:



Εικόνα 4.2.1: Παράδειγμα αντιστοίχισης δύο χρονοσειρών μέσω της μεθόδου a) Ευκλείδειας Απόστασης b) DTW

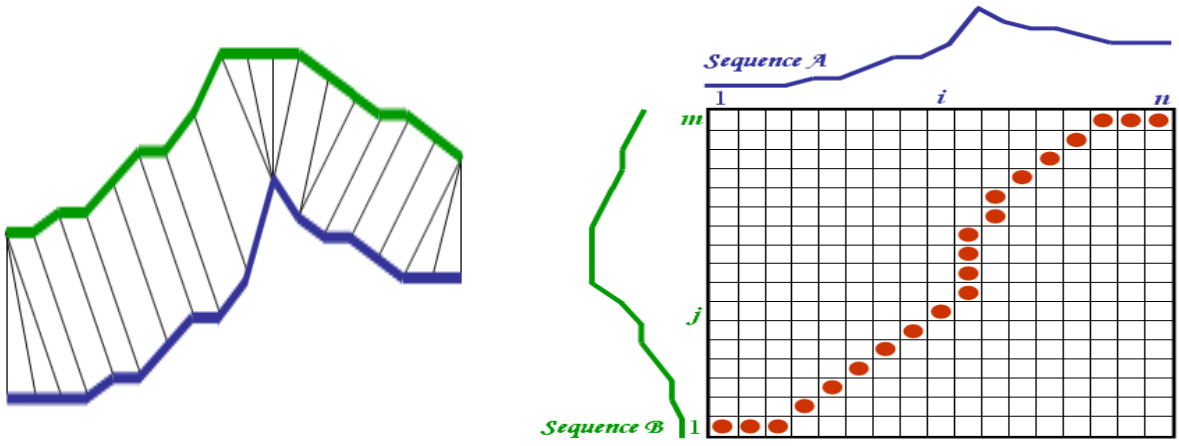
Η αντιστοίχιση απεικονίζεται με ευθείες γραμμές και ενώ η Ευκλείδεια Απόσταση ζητά ένα-προς-ένα αντιστοίχιση, στην περίπτωση του DTW ένα σημείο της μιας χρονοσειράς μπορεί να αντιστοιχηθεί με περισσότερα από ένα σημεία της άλλης χρονοσειράς.

Άρα, έχοντας ως δεδομένα δύο χρονοσειρές **X**, **Y,** όπου $X=(x_0,\ldots,x_{M-1})$ και $Y=(y_0,\ldots,y_{N-1})$, ο αναδρομικός τύπος υπολογισμού της απόστασης DTW είναι:

$$DTW(X,Y) = \begin{cases} 0 & \text{if } M-1 = N-1 = 0 \\ \inf & \text{if } M-1 = 0 \text{ or } N-1 = 0 \\ d(x_0,y_0) + min\{DTW(Rest(X),Rest(Y)), \\ DTW(Rest(X),Y), DTW(X,Rest(Y))\} & \text{otherwise} \end{cases}$$

Όπου Rest(X) = x_{i-1}, Rest(Y) = y_{i-1} (το i εκφράζει το αριθμό της εκάστοτε επανάληψης) και d(x,y) η Ευκλείδεια Απόσταση των σημείων x και y.

Αρχικά, στην πρώτη γραμμή και στην πρώτη στήλη συμπληρώνονται μηδενικά και έπειτα ξεκινάει η διαδικασία του αλγορίθμου. Στον πίνακα DTW αποθηκεύονται οι αποστάσεις μεταξύ όλων των σημείων των δύο χρονοσειρών, σύμφωνα με τον παραπάνω τύπο και έπειτα ξεκινώντας από το σημείο (0,0) του πίνακα, βρίσκεται το βέλτιστο μονοπάτι μέχρι το σημείο (M, N). Ένα τέτοιο παράδειγμα για δύο τυχαίες χρονοσειρές παρουσιάζεται παρακάτω, όπου αρχικά φαίνεται η αντιστοίχιση με τη μέθοδο DTW και έπειτα το βέλτιστο μονοπάτι, το οποίο απεικονίζεται με κόκκινες βούλες:



Εικόνα 4.2.2: Αντιστοίχιση δυο χρονοσειρών μέσω της μεθόδου DTW και το αντίστοιχο βέλτιστο μονοπάτι

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση dtw() της βιβλιοθήκης “dtw” της γλώσσας R, όπου λαμβάνει ως είσοδο τις δύο χρονοσειρές που θα συγκριθούν και επιστρέφει διάφορες τιμές που αφοράνε τον αλγόριθμο. Εδώ επιλέχθηκε να επιστρέφει τη συνολική τους απόσταση, καθώς αυτό είναι που ενδιαφέρει άμεσα.

4.4 Edit Distance on Real sequence (EDR)

Η μέθοδος Edit Distance on Real sequence (Chen et. al. - 2005) είναι ένα μέτρο σύγκρισης ακολουθιών, που βασίζεται στη μέτρηση του ελάχιστου πλήθους των αναγκαίων μεμονωμένων αλλαγών που πρέπει να γίνουν στη μια ακολουθία, ώστε να ταυτιστεί με την άλλη. Οι αλλαγές αυτές μπορεί να είναι εισαγωγή, διαγραφή ή αντικατάσταση ενός ή περισσότερων στοιχείων μιας ακολουθίας. Για παράδειγμα, μεταξύ των ακολουθιών “ABCDOIT” και “ABQDOIKT”, ο ελάχιστος αριθμός αλλαγών είναι 2, καθώς αν στη δεύτερη ακολουθία αντικαταστήσουμε το "Q" με το "C" και αφαιρέσουμε το "K", θα ταυτιστεί με την πρώτη ακολουθία.

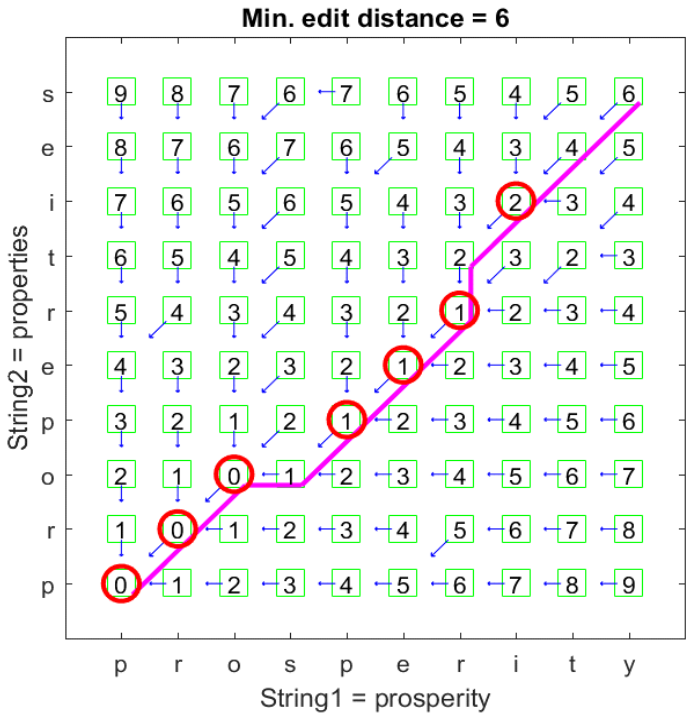
Όπως και η μέθοδος Longest Common SubSequence, που αναλύθηκε προηγουμένως, έτσι και η μέθοδος EDR βασίστηκε στην επεξεργασία και τη σύγκριση συμβολοσειρών, όμως έπειτα προσαρμόστηκε για να εφαρμόζεται και σε αριθμητικές ακολουθίες. Για να πραγματοποιηθεί αυτό, είναι αναγκαίο και εδώ να εισαχθεί ένα κατώφλι που θα καθορίζει αν οι αντίστοιχοι αριθμοί που συγκρίνονται είναι ίσοι ή όχι. Σε αντίθεση με την LCSS, η EDR επιβάλλει "penalties" στα στοιχεία των ακολουθιών που δεν ταυτίζονται, δηλαδή "τιμωρεί" τις χρονοσειρές κάθε φορά που απαιτείται μια αλλαγή (εισαγωγή, διαγραφή ή αντικατάσταση), γεγονός που βελτιώνει την ακρίβεια της. Η δημοφιλέστερη προσέγγιση αυτής της μεθόδου είναι μέσω του δυναμικού προγραμματισμού, καθώς βελτιώνει αισθητά την πολυπλοκότητα, η οποία είναι O(M*N), όπου M, N τα μήκη των δύο ακολουθιών που θα συγκριθούν.

Στην παρούσα διπλωματική επεξεργαζόμαστε αριθμητικά δεδομένα και όχι συμβολοσειρές. Άρα, έχοντας ως δεδομένα δύο χρονοσειρές **X**, **Y**, όπου $X=(x_0,\ldots,x_{M-1})$ και $Y=(y_0,\ldots,y_{N-1})$ και το κατώφλι **ε**, ο αναδρομικός τύπος υπολογισμού του ελάχιστου πλήθος αλλαγών που πρέπει να γίνουν είναι:

$$EDR(X,Y) = \begin{cases} N & \text{if } M-1=0 \\ M & \text{if } N-1=0 \\ min\{EDR(Rest(X),Rest(Y)) + d_{edr}(x_0,y_0), \\ EDR(Rest(X),Y)+1, EDR(X,Rest(Y))+1\} & \text{otherwise} \end{cases}$$

Όπου Rest(X) = x_{i-1}, Rest(Y) = y_{i-1} (το i εκφράζει το αριθμό της εκάστοτε επανάληψης). Επίσης, άμα οι εκάστοτε παρατηρήσεις που συγκρίνονται θεωρούνται ίσες, τότε δεν υπάρχει "penalty", οπότε $d_{edr}=0$, ενώ στην αντίθετη περίπτωση, θέτουμε $d_{edr}=1$.

Το αποτέλεσμα υπολογίζεται με δυναμικό προγραμματισμό, καθώς σε κάθε επανάληψη η συνάρτηση EDR καλεί τον εαυτό της. Με αυτόν τον τρόπο, τρέχει επαναληπτικά η μέθοδος, κρατώντας στον πίνακα EDR όλες τις τιμές που έχουν υπολογιστεί, ώστε να ξαναχρησιμοποιηθούν στην επόμενη επανάληψη. Αρχικά, στην πρώτη γραμμή και στην πρώτη στήλη συμπληρώνονται οι αριθμοί από το ένα έως το πλήθος της δεύτερης και πρώτης χρονοσειράς, αντίστοιχα, και έπειτα ξεκινάει η διαδικασία του αλγορίθμου. Σε κάθε επανάληψη, ελέγχεται αν η διαφορά των δύο εκάστοτε στοιχείων είναι μικρότερη από το κατώφλι και ορίζεται το ανάλογο subcost. Έπειτα, έχοντας αποθηκευμένες τις προηγούμενες τιμές, ελέγχεται ποια τιμή από τις τρείς που φαίνονται στην παραπάνω μαθηματική σχέση είναι μικρότερη. Με τη συνεχή επανάληψη της ίδιας διαδικασίας, δημιουργείται ο πίνακας EDR, ο οποίος στη θέση (M, N) δείχνει ποιο είναι το ελάχιστο πλήθος αλλαγών που απαιτούνται, ώστε να ταυτιστούν οι δύο χρονοσειρές. Η μέθοδος έχει εξαιρετικά αποτελέσματα και έχει αποδειχθεί πιο ακριβείς σε χρονοσειρές που χαρακτηρίζονται από έντονο θόρυβο συγκριτικά με τις μεθόδους που αναφέρθηκαν προηγουμένως. Στην παρακάτω εικόνα, φαίνεται ένα πολύ ξεκάθαρο παράδειγμα σύγκρισης των συμβολοσειρών "prosperity" και "properties", όπου το ελάχιστο πλήθος αναγκαίων αλλαγών αποδεικνύεται ίσο με 6.



Εικόνα 4.3.1: Παράδειγμα σύγκρισης χρονοσειρών μέσω της μεθόδου EDR

Ομοίως εφαρμόζεται και σε αριθμητικές ακολουθίες, έχοντας το κατώφλι **ε** ως καθοριστικό παράγοντα ισότητας των εκάστοτε στοιχείων.

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση EDRDistance() της βιβλιοθήκης “TSdist” της γλώσσας R, όπου λαμβάνει ως είσοδο τις δύο χρονοσειρές που θα συγκριθούν και το κατώφλι **ε** (το οποίο το θέσαμε ίσο με **ε** = 0.1) κάτω από το οποίο θεωρούνται ίσες οι δύο εκάστοτε τιμές που συγκρίνονται και επιστρέφει το ελάχιστο πλήθος των αναγκαίων αλλαγών που πρέπει να γίνουν. Προαιρετικά, μπορεί να εισαχθεί ως είσοδος και ένας περιορισμός, ώστε να μην αντιστοιχίζονται σημεία που βρίσκονται χρονικά μακριά.

4.5 Edit distance with Real Penalty (ERP)

Η μέθοδος Edit distance with Real Penalty (Chen and Ng - 2004) είναι ένας αλγόριθμος μέτρησης της ομοιότητας μεταξύ δύο χρονοσειρών και αποτελεί ένα συνδυασμό των μεθόδων DTW και EDR, οι οποίες αναφέρθηκαν προηγουμένως. Η ομοιότητα της με τη DTW είναι ο τρόπος αντιστοίχισης των σημείων των εκάστοτε χρονοσειρών που συγκρίνονται, ενώ διαφέρουν στο γεγονός ότι η ERP δεν επαναλαμβάνει τα προηγούμενα στοιχεία. Το κοινό χαρακτηριστικό που έχει με την EDR είναι ότι χρησιμοποιεί την ανεκτικότητα της τελευταίας στα απόντα σημεία και στα σημεία που δεν αντιστοιχίζονται με κανένα άλλο σημείο της άλλης χρονοσειράς. Βέβαια, σε αντίθεση με την EDR, το "penalty" υλοποιείται θέτοντας μια παράμετρο g και χρησιμοποιώντας την ευκλείδεια απόσταση και όχι μέσω της δυαδικής τιμής ανάλογα με την ικανοποίηση του κατωφλίου **ε**.

Άρα, έχοντας ως δεδομένα δύο χρονοσειρές **X**, **Y**, όπου $X=(x_0,\ldots,x_{M-1})$ και $Y=(y_0,\ldots,y_{N-1})$ και τη παράμετρο g (gap), ο αναδρομικός τύπος υπολογισμού της απόστασης ERP είναι:

$$ERP(X,Y) = \begin{cases} \sum_{i=0}^{N-1} |y_i - g| & \text{if } M-1=0 \\ \sum_{i=0}^{M-1} |x_i - g| & \text{if } N-1=0 \\ min\{ERP(Rest(X),Rest(Y)) + d(x_0,y_0), \\ ERP(Rest(X),Y) + d(x_0,g), ERP(X,Rest(Y)) + d(g,y_0)\} & \text{otherwise} \end{cases}$$

Όπου Rest(X) = x_{i-1}, Rest(Y) = y_{i-1} (το i εκφράζει το αριθμό της εκάστοτε επανάληψης), g (gap) εκφράζει το “penalty” για την απουσία ή τη μη αντιστοίχιση ενός στοιχείου από μια χρονοσειρά και d(x,y) η Ευκλείδεια Απόσταση των σημείων x και y.

Οι αποστάσεις στον πίνακα ERP υλοποιούνται μέσω του γραμμικού προγραμματισμού και υπολογίζονται μέσω της ευκλείδειας απόστασης. Η κύρια ιδέα είναι να βρεθεί το ελάχιστο μονοπάτι του πίνακα που περιγράφει πλήρως την αντιστοίχιση μεταξύ των σημείων των χρονοσειρών. Αρχικά, στην πρώτη γραμμή και στην πρώτη στήλη συμπληρώνονται οι αποστάσεις των κάθε σημείων της δεύτερης και πρώτης χρονοσειράς από τη παράμετρο g, αντίστοιχα. Έπειτα ξεκινάει η διαδικασία του αλγορίθμου και υπολογίζονται όλες τις αποστάσεις των σημείων των χρονοσειρών μέσω του παραπάνω τύπου. Όταν κάποιο σημείο δεν έχει "ταίρι", τότε προστίθεται ένα αριθμός gap με τιμή g στην ελλιπή χρονοσειρά, ώστε να μπορεί να πραγματοποιηθεί η ευκλείδεια απόσταση. Πολλοί μελετητές αυτής της μεθόδου

έχουν προτείνει ότι η καλύτερη τιμή για την παράμετρο g είναι το μηδέν (αυτή η τιμή χρησιμοποιήθηκε και στην παρούσα διπλωματική εργασία) για τους εξής λόγους:

- Η απόσταση μεταξύ των ακολουθιών R και S αντιστοιχεί στην διαφορά μεταξύ της περιοχής κάτω από την καμπύλη της R και της περιοχής κάτω από την καμπύλη της S.
- Διατηρείται η μέση τιμή της αλλαγμένης χρονοσειράς μετά την προσθήκη των κενών σημείων (gap).

Στην παρούσα διπλωματική εργασία χρησιμοποιείται η συνάρτηση ERPDistance() της βιβλιοθήκης “TSdist” της γλώσσας R, όπου λαμβάνει ως είσοδο τις δύο χρονοσειρές που θα συγκριθούν και την τιμή του "penalty" g που επιβάλλεται στα απόντα σημεία ή στην ακολουθία σημείων που δεν έχουν αντιστοιχηθεί με κανένα σημείο της άλλης χρονοσειράς. Προαιρετικά, μπορεί να εισαχθεί ως είσοδος και ένας περιορισμός, ώστε να μην αντιστοιχίζονται σημεία που βρίσκονται χρονικά μακριά.

Κεφάλαιο 5: Μεθοδολογία Πρόβλεψης Χρονοσειρών

5.1 Ιδέα της μεθόδου

Η ιδέα της παρούσας διπλωματικής εργασίας πηγάζει από το σκεπτικό πως θα μπορούσε κανείς να επιλέξει την κατάλληλη μέθοδο πρόβλεψης για μια συγκεκριμένη χρονοσειρά, έχοντας αναλύσει και προβλέψει χρονοσειρές ίδιου μοτίβου με ίδια ποιοτικά χαρακτηριστικά, δηλαδή όμοιες με την αρχική. Με άλλα λόγια, απαντά στο πως θα βελτιωθεί η ακρίβεια πρόβλεψης των κλασικών μεθόδων για μια χρονοσειρά, γνωρίζοντας ποια μέθοδος επιλέχθηκε για την πρόβλεψη των ομοίων της.

Πιο συγκεκριμένα, η ιδέα της διπλωματικής εργασίας είναι αρχικά η ομαδοποίηση πληθώρας χρονοσειρών με γνώμονα την ομοιότητα τους και με χρήση των μεθόδων σύγκρισης που αναλύθηκαν στο Κεφάλαιο 5. Στη συνέχεια, να γίνει αξιοποίηση των ευρημάτων, ώστε να επιτυγχάνεται η αυτόματη επιλογή της κατάλληλης μεθόδου πρόβλεψης για την εκάστοτε χρονοσειρά, βάσει της απόδοσης τους στις υπόλοιπες που βρίσκονται στην ίδια ομάδα.

Αυτή η μέθοδος είναι ιδιαίτερα σημαντική και χρήσιμη και θα μπορεί να χρησιμοποιηθεί σε πολλές εφαρμογές, όπως σε μεγάλες βάσεις δεδομένων (Big Data), όπου το πλήθος των δεδομένων καθιστά τη σύγκριση και την πρόβλεψη τους μια αρκετά εξαντλητική και αργή διαδικασία. Επιπροσθέτως βρίσκει εφαρμογή σε περιπτώσεις αυξημένης τυχαιότητας, όπου η αβεβαιότητα της επιλογής της κατάλληλης μεθόδου πρόβλεψης και της παραμετροποίησης είναι αυξημένη.

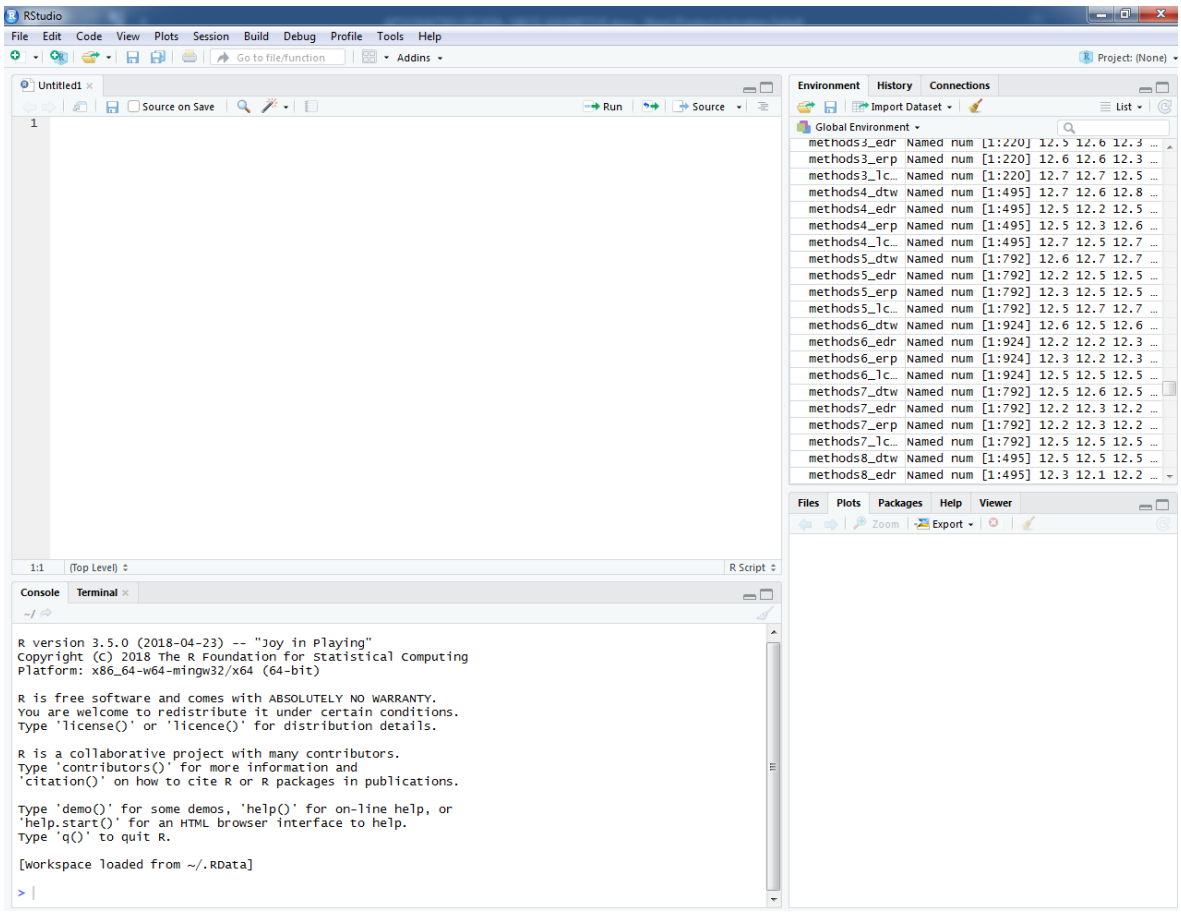
Τα δεδομένα που χρησιμοποιήθηκαν είναι οι 3003 χρονοσειρές του M3-Competition. Ο M3-Competition είναι ο μεγαλύτερος διαγωνισμός πρόβλεψης που οργανώθηκε ποτέ, καθώς ζητούμενο ήταν να δοθούν προβλέψεις για 3003 διαφορετικές χρονοσειρές. Οι 3003 χρονοσειρές συμπεριλαμβάνουν διάφορους τύπους δεδομένων (micro, industry, macro, economics, demographic, other) και με διάφορες συχνότητες παρατηρήσεων (ετήσιες, τριμηνιαίες, μηνιαίες και άλλες). Το ελάχιστο μήκος παρατηρήσεων για κάθε τύπο δεδομένων είναι 14 και το μέγιστο μήκος είναι 126.

5.2 Το RStudio ως εργαλείο προβλέψεων

Η παρούσα διπλωματική εργασία δημιουργήθηκε και ελέγχθηκε εξ' ολοκλήρου με τη γλώσσα προγραμματισμού R. Η R είναι ένα ανοιχτό και ελεύθερο λογισμικό και αποτελεί ταυτόχρονα μια ισχυρή γλώσσα προγραμματισμού και ένα προγραμματιστικό περιβάλλον που προτείνεται για πολλές εργασίες, όπως στατιστικούς υπολογισμούς, αναλύσεις, γραφικές απεικονίσεις κ.α. Δημιουργήθηκε από τον John Chambers (1976) στα εργαστήρια Bell και είναι παρόμοια με τη γλώσσα S. Χρησιμοποιείται ευρέως ως εκπαιδευτική γλώσσα και ως ερευνητικό εργαλείο και μπορεί να μεταγλωττιστεί και να εκτελεστεί σε μεγάλη γκάμα λειτουργικών συστημάτων, όπως τα Windows της Microsoft, το MacOS της Apple και τα Ubuntu του Linux.

Οι δυνατότητες της R είναι τεράστιες και είναι μια από τις δημοφιλέστερες γλώσσες για την ανάλυση και τον υπολογισμό διαφόρων στατιστικών εργασιών, όπως τα γραμμικά και μη γραμμικά μοντέλα, οι χρονοσειρές, η ταξινόμηση, η ομαδοποίηση και η μηχανική μάθηση, καθώς επίσης και την μεγάλη γκάμα τεχνικών δημιουργίας γραφημάτων που διαθέτει. Οι χρήστες της γλώσσας R, εκτός το ότι έχουν την δυνατότητα απέραντης βοήθειας από το διαδίκτυο, λαμβάνουν και εξαιρετική βοήθεια από το ίδιο το λογισμικό, καθώς περιέχει μια τεράστια συλλογή βιβλιοθηκών-πακέτων (packages). Πιο συγκεκριμένα, σύμφωνα με το δίκτυο CRAN (Comprehensive R Archive Network), αυτή τη στιγμή υπάρχουν 13285 διαθέσιμα πακέτα.

Στη συγκεκριμένη διπλωματική εργασία θα χρησιμοποιηθεί το πρόγραμμα RStudio, ένα ελεύθερο προγραμματιστικό περιβάλλον για την γλώσσα R, το οποίο παρέχει μια οργανωμένη διάταξη και διάφορες πρόσθετες επιλογές. Το πρόγραμμα αυτό παρέχει μια ευρεία γκάμα δυνατοτήτων για την δημιουργία, ανάλυση, πρόβλεψη και απεικόνιση χρονοσειρών. Το RStudio είναι ένα εξαιρετικά χρήσιμο εργαλείο, καθώς δίνει τη δυνατότητα γραφής και εκτέλεσης του κώδικα, καθώς και εξαγωγής αποτελεσμάτων και γραφημάτων. Στην παρακάτω εικόνα φαίνεται το προγραμματιστικό περιβάλλον RStudio, το οποίο είναι χωρισμένο σε τέσσερα τμήματα:



Εικόνα 5.2.1: Προγραμματιστικό περιβάλλον RStudio

Στο πάνω αριστερά παράθυρο φαίνεται το script του κώδικα, στο κάτω αριστερά φαίνεται η κονσόλα της R, στο πάνω δεξιά φαίνονται όλες οι αποθηκευμένες μεταβλητές και συναρτήσεις, καθώς και το ιστορικό των πρόσφατων εντολών που έχουν χρησιμοποιηθεί και στο κάτω δεξιά φαίνεται ένα σύνολο καρτελών, όπως τα αρχεία, τα γραφήματα, τα πακέτα (packages) και η βοήθεια (διάφορες θεωρητικές πληροφορίες για τα πακέτα, τις συναρτήσεις και γενικά για το RStudio).

5.3 Διαδικασία παραγωγής προβλέψεων και αντίστοιχες εντολές στην R

5.3.1 Αποεποχικοποίηση

Ο γενικός σκοπός της διπλωματικής εργασίας είναι η ομαδοποίηση χρονοσειρών με βάση την ομοιότητα τους και η επιλογή της κατάλληλης μεθόδου πρόβλεψης, ώστε να μειωθεί το σφάλμα των κλασσικών μεθόδων. Αρχικά, είναι απαραίτητο να παραχθούν οι προβλέψεις όλων των μεθόδων πρόβλεψης για τις 3003 χρονοσειρές του M3. Στην παρούσα διπλωματική εργασία θα χρησιμοποιηθούν 12 μέθοδοι πρόβλεψης, οι οποίες αναλύθηκαν στο Κεφάλαιο 3 και είναι οι εξής:

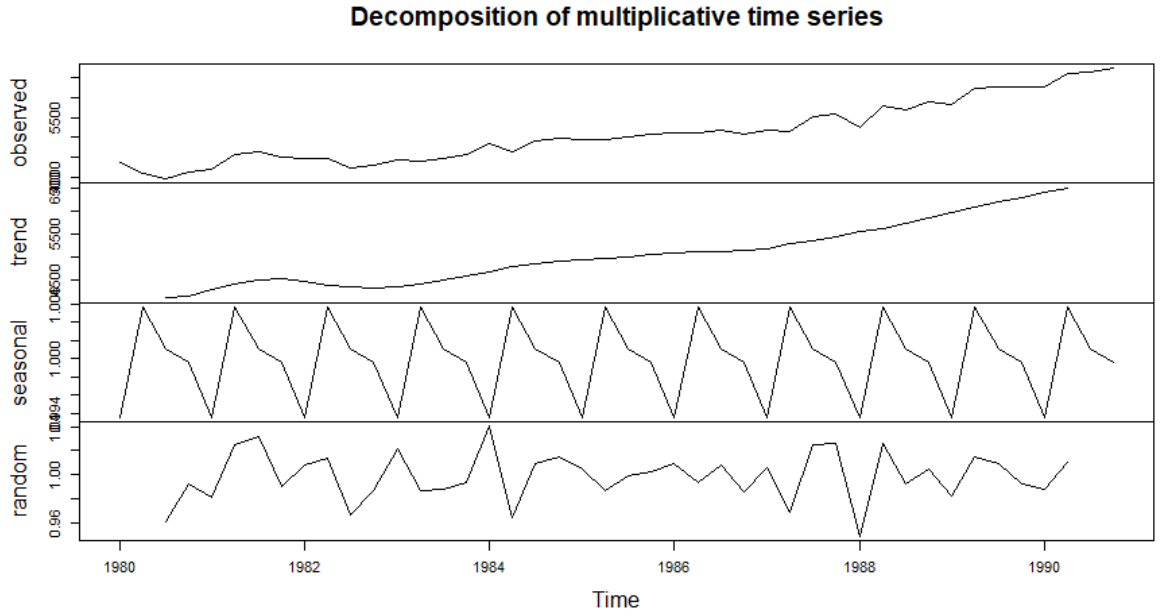
- Arima
- Simple Exponential Smoothing
- Damped Trend Exponential Smoothing
- Holt Exponential Smoothing
- Theta
- Naive
- Seasonal Naive
- Neural Network Time Series Forecasts
- Random Walk with Drift
- Exponential Smoothing state space model
- Exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal components)
- STL Decomposition and Forecasting

Μεγάλο μέρος αυτών των χρονοσειρών χαρακτηρίζεται από εποχικότητα, η οποία αναλύθηκε στο Κεφάλαιο 2 και είναι ένα μοτίβο περιοδικής διακύμανσης των τιμών της χρονοσειράς που εμφανίζεται σε σταθερά διαστήματα και μικρότερα του ενός έτους. Οι μέθοδοι πρόβλεψης Simple Exponential Smoothing, Damped Trend Exponential Smoothing, Holt Exponential Smoothing, Theta, Naive και Random Walk with Drift απαιτούν αποεποχικοποιημένες χρονοσειρές για να παράγουν τις ακριβείς προβλέψεις τους σε, οπότε κρίνεται απαραίτητο αρχικά να γίνει η αποεποχικοποίηση. Σε αντίθεση με τις παραπάνω, οι υπόλοιπες μέθοδοι χρησιμοποιούν τα αρχικά κανονικοποιημένα δεδομένα. Η αποεποχικοποιημένη χρονοσειρά προκύπτει μέσω του λόγου των πραγματικών τιμών της με τον αντίστοιχο δείκτη εποχικότητας. Οι δείκτες αυτοί υπολογίζονται μέσω της Κλασικής Μεθόδου Αποσύνθεσης που καλείται μέσω της εξής συνάρτησης της γλώσσας R:

ts.decompose <- decompose (time series, type = "multiplicative", filter = NULL)

Στη μεταβλητή “ts.decompose” αποθηκεύεται όλη η αποσύνθεση, ενώ στις παραμέτρους της συνάρτησης έχουμε το “time series”, όπου εισάγεται η χρονοσειρά προς αποσύνθεση, το type που δηλώνει αν η αποεποχικοποίηση θα γίνει με το προσθετικό (“additive”) ή το πολλαπλασιαστικό μοντέλο (“multiplicative”) και το filter που ορίζει την τεχνική με την οποία θα προσδιοριστούν οι δείκτες εποχικότητας (το NULL δηλώνει ότι θα γίνει χρήση των κινητών και κεντρικών μέσων όρων). Με αυτήν την τεχνική, μπορούν να υπολογιστούν οι δείκτες εποχικότητας, να απομονωθούν και να προκύψει η τελική αποεποχικοποιημένη χρονοσειρά. Παρακάτω φαίνεται ένα γραφικό παράδειγμα της Κλασικής Μεθόδου Αποσύνθεσης, που εκτελέστηκε στην 1000η χρονοσειρά του M3 dataset, χρησιμοποιώντας την εντολή:

plot (*decompose (data[[1000]]$x, type = "multiplicative", filter = NULL))*

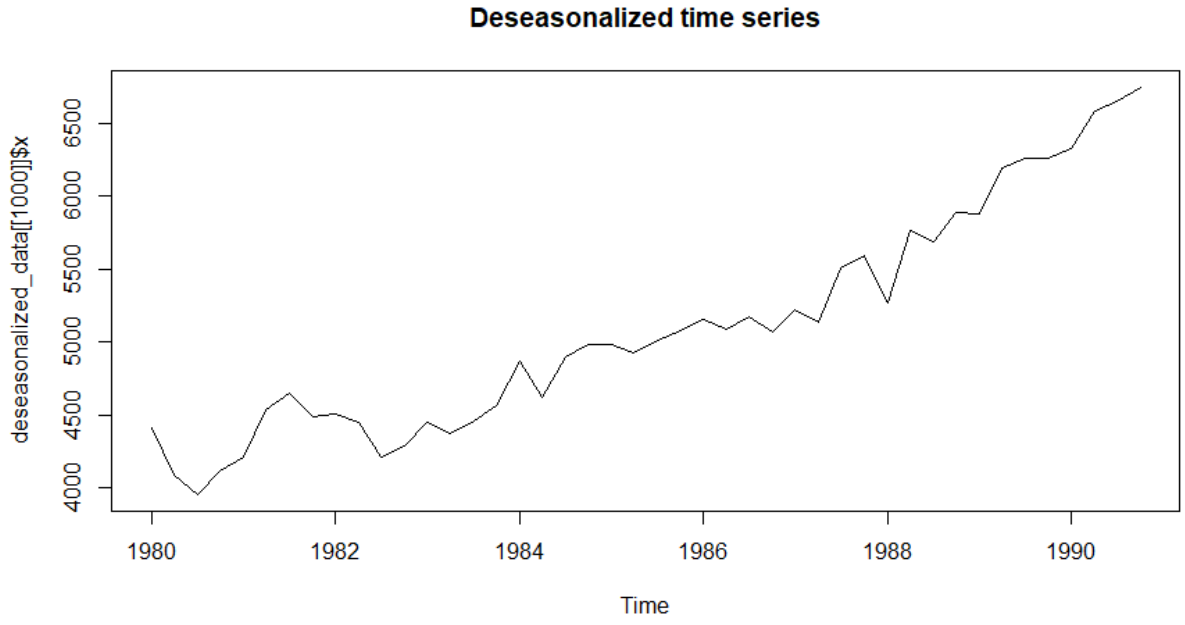


Εικόνα 5.3.1.1: Κλασσική Μέθοδος Αποσύνθεσης στη 1000η χρονοσειρά του M3

Παρατηρείται ότι ουσιαστικά η αποσύνθεση χωρίζει την αρχική χρονοσειρά σε 4 διαφορετικές χρονοσειρές, την αρχική (observed), την τάση (trend), την εποχικότητα (seasonal) και την τυχαιότητα (random). Έχοντας τους δείκτες εποχικότητας, μπορούν πλέον να υπολογιστούν τα αποεποχικοποιημένα δεδομένα “deseasonalized_ts”, διαιρώντας τα αρχικά δεδομένα με τον αντίστοιχο δείκτη εποχικότητας (καθώς χρησιμοποιείται το πολλαπλασιαστικό μοντέλο), το οποίο πραγματοποιείται με την παρακάτω εντολή στη γλώσσα R:

deseasonalized_ts <- time_series / ts.decompose$seasonal

Μέσω του συμβόλου “$” απομονώνεται μια συγκεκριμένη “υποχρονοσειρά” της αποσύνθεσης, όπου στη συγκεκριμένη περίπτωση επιλέχτηκαν οι δείκτες εποχικότητας. Στη συνέχεια του παραδείγματος για την 1000η χρονοσειρά του M3, παρακάτω φαίνεται η αποεποχικοποιημένη χρονοσειρά:



Εικόνα 5.3.1.2: Αποεποχικοποιημένη χρονοσειρά της 1000ης χρονοσειράς του M3

5.3.2 Τεστ εποχικότητας

Παρ' όλη την επιθυμία κάποιων μεθόδων να παράγουν ακριβείς προβλέψεις σε αποεποχικοποιημένες χρονοσειρές, πρέπει να τονιστεί ότι δεν είναι εποχιακές όλες οι χρονοσειρές του Μ3. Οπότε, οφείλει να γίνεται ένας έλεγχος εποχιακής συμπεριφοράς. Ο έλεγχος αυτός πραγματοποιείται μέσω της σύγκρισης της αυτοσυσχέτισης των δεδομένων με περίοδο καθυστέρησης k (ίση με τον αριθμό των περιόδων ενός κύκλου εποχιακότητας pos) με τις αυτοσυσχετίσεις περιόδου καθυστέρησης έως και μιας μονάδας μικρότερης από τον αριθμό περιόδων ενός κύκλου εποχιακότητας. Δηλαδή πρέπει να ισχύει η εξής ανισότητα:

$$\mathrm{ACF_{pos}} > \mathrm{Limit}$$

$$\mathrm{ACF_k} = \frac{\sum_{i=1+k}^{n}[(Y_i-\overline{Y})*(Y_{i-k}-\overline{Y})]}{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}$$

$$\mathrm{Limit} = \mathrm{t_{critical}} * \sqrt{\frac{1+2*(\mathrm{ACF_1}+\sum_{i=2}^{pos-1} ACF_i^2}{n}}$$

όπου, Y οι αρχικές παρατηρήσεις

$\overline{Y}$ ο μέσος όρος των αρχικών παρατηρήσεων

n το πλήθος των αρχικών παρατηρήσεων

$\mathrm{t_{critical}}$ συντελεστής που καθορίζεται από το επίπεδο εμπιστοσύνης

Στη συγκεκριμένη περίπτωση, θεωρήθηκε διάστημα εμπιστοσύνης ίσο με 90%, οπότε $\mathrm{t_{critical}} = 1{,}645$.

Όσες χρονοσειρές περνάνε από αυτόν τον έλεγχο αποεποχικοποιούνται, αλλιώς προβλέπονται κανονικά με τα αρχικά δεδομένα.

5.3.3 Τρόπος παραγωγής των προβλέψεων

Όσον αφορά τη διαδικασία της πρόβλεψης, κάθε χρονοσειρά προβλέπεται μέσω των 12 μεθόδων, όπου για κάθε μέθοδο χρησιμοποιείται αντίστοιχη συνάρτηση της γλώσσας R, που αναφέρεται στο Κεφάλαιο 3 μετά τη θεωρητική επεξήγηση της κάθε μεθόδου. Για παράδειγμα, η πρόβλεψη της απλής εκθετικής εξομάλυνσης γίνεται με την εξής συνάρτηση:

forecast_ses <- ses (deseasonalized_ts, h=length(real_future_values))

Όπου η πρώτη παράμετρος είναι η αποεποχικοποιημένη χρονοσειρά και η δεύτερη είναι το πλήθος των περιόδων πρόβλεψης, το οποίο είναι ίσο με το πλήθος των πραγματικών μελλοντικών τιμών, ώστε έπειτα να συγκριθούν και να υπολογιστεί το αντίστοιχο σφάλμα sMAPE για κάθε χρονοσειρά. Ομοίως με αντίστοιχες συναρτήσεις, υπολογίζονται οι προβλέψεις για τις υπόλοιπες μεθόδους.

Μετά την πρόβλεψη των χρονοσειρών, είναι αναγκαία η ξανα-εποχικοποίηση των προβλέψεων που παράχθηκαν από αποεποχικοποιημένες χρονοσειρές. Η διαδικασία αυτή απαιτεί τον πολλαπλασιασμό (καθώς χρησιμοποιείται το πολλαπλασιαστικό μοντέλο) της κάθε παρατήρησης με τον αντίστοιχο δείκτη εποχικότητας που προκύπτει από την Κλασική Μέθοδο Αποσύνθεσης.

Αφού έχουν παραχθεί οι προβλέψεις, μπορεί πλέον να υπολογιστεί το Συμμετρικό Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (sMAPE) της κάθε χρονοσειράς, χρησιμοποιώντας την εξής συνάρτηση της γλώσσας R και αποθηκεύοντας το στη μεταβλητή smape_ts:

smape_ts* <- smape (*real_future_values, forecast_future_values)*100

Όπου η πρώτη παράμετρος είναι οι πραγματικές μελλοντικές τιμές και η δεύτερη παράμετρος είναι τιμές που προβλέφθηκαν. Έπειτα πολλαπλασιάζεται το σφάλμα με το 100, ώστε η τελική τιμή να ναι εκφρασμένη επί τοις εκατό.

5.4 Σύγκριση χρονοσειρών

5.4.1 Ομαδοποίηση χρονοσειρών βάσει της ομοιότητας τους

Η ομαδοποίηση ή συσταδοποίηση (clustering) δεδομένων είναι η διαδικασία, στην οποία τα δεδομένα χωρίζονται σε διάφορες λογικές ομάδες. Στην παρούσα διπλωματική δημιουργούνται ομάδες χρονοσειρών με κριτήριο την ομοιότητα τους, δηλαδή για κάθε μια από αυτές βρίσκεται το σύνολο των χρονοσειρών από τις υπολειπόμενες 3002 που είναι πολύ όμοιες με την αρχική. Η σύγκριση των χρονοσειρών πραγματοποιείται με τη χρήση των τεσσάρων μεθόδων, που παρουσιάστηκαν εκτενώς στο Κεφάλαιο 4, οι οποίες είναι η Longest Common Subsequence (LCSS), η Dynamic Time Warping (DTW), η Edit Distance for Real sequences (EDR) και η Edit distance with Real Penalty (ERP).

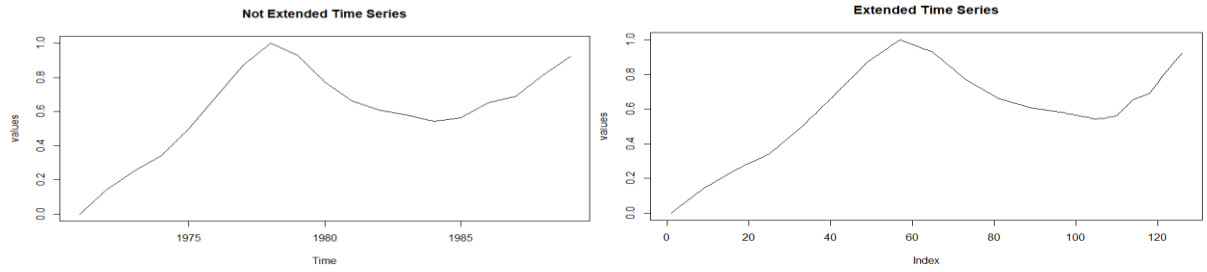
Οι μέθοδοι LCSS και EDR παράγουν ένα θετικό ακέραιο νούμερο, που στην πρώτη μέθοδο αντιπροσωπεύει τη μέγιστη κοινή υποακολουθία μεταξύ δύο χρονοσειρών και στη δεύτερη μέθοδο αντιπροσωπεύει τις ελάχιστες δυνατές αλλαγές (προσθήκη, διαγραφή, αντικατάσταση) που πρέπει να εκτελεστούν στις παρατηρήσεις της μιας χρονοσειράς, ώστε να ταυτιστεί απόλυτα με την άλλη. Αντιθέτως, οι μέθοδοι DTW και ERP υπολογίζουν απόσταση μεταξύ των δύο χρονοσειρών, με αποτέλεσμα να μπορεί να είναι οποιοσδήποτε θετικός αριθμός. Οπότε η διαδικασία που ακολουθήθηκε, ώστε να δημιουργηθούν οι διάφορες ομάδες, είναι αρχικά να κανονικοποιηθούν όλες οι χρονοσειρές, ώστε να μην υπάρχει διαφορά στη κλίμακα, το οποίο υλοποιήθηκε με τον τύπο:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

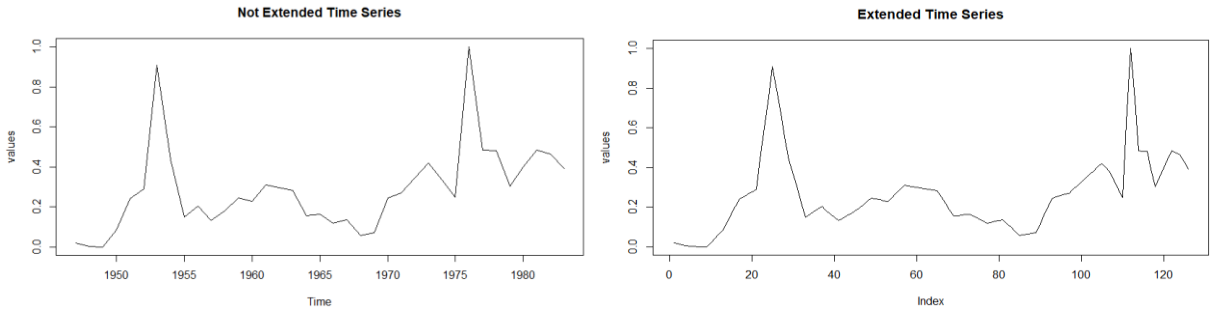
Όπου X είναι οι κανονικοποιημένες τιμές. Έπειτα ορίστηκε ως αναφορά (reference) μια προς μια κάθε χρονοσειρά από τις 3003 και συγκρινόταν με τις υπόλοιπες 3002 ώστε κάθε φορά να παράγονται 4 διανύσματα (ένα για κάθε μέθοδο σύγκρισης) μεγέθους 3003 που να περιέχουν τον αριθμητικό συγκριτικό συσχετισμό (ακέραιος αριθμός ή απόσταση) της reference με τη κάθε μια από τις υπόλοιπες. Στη συνέχεια, βρίσκοντας τη μέγιστη τιμή για την LCSS και την ελάχιστη τιμή για τις υπόλοιπες, καθώς και ένα ποσοστό, καθορίστηκε τελικώς ποιες χρονοσειρές θα θεωρηθούν όμοιες με τη reference. Για παράδειγμα για τη μεθόδους EDR, ERP, DTW, αν χρησιμοποιηθεί το ποσοστό των 50%, τότε όσες χρονοσειρές έχουν απόσταση από τη reference μικρότερη ίση από το min(distance)+min(distance)*50%, τότε θεωρούνται ίσες με αυτή. Αντίστοιχα για την LCSS, με τη διαφορά ότι ο έλεγχος θα είναι max(distance)-max(distance)*50%.

5.4.2 Μέθοδος Διαστολής Μήκους Χρονοσειρών

Οι 3003 χρονοσειρές του διαγωνισμού M3 ποικίλλουν στο μέγεθος, με αποτέλεσμα να παρατηρηθεί ότι κυρίως η μέθοδος LCSS δεν παράγει λογικά αποτελέσματα. Αυτό σημαίνει ότι θεωρεί όμοιες 2 χρονοσειρές, οι οποίες στην πραγματικότητα δεν μοιάζουν καθόλου και αυτό συμβαίνει λόγω του διαφορετικού μήκους χρονοσειρών και της φύσης του αλγορίθμου της να βρίσκει μέγιστες κοινές υποακολουθίες. Το πρόβλημα αυτό λύθηκε μεταβάλλοντας τα μήκη των χρονοσειρών και θέτοντας τα ίσα με το μήκος της μεγαλύτερης χρονοσειράς, το οποίο είναι 126 παρατηρήσεις. Για να επιτευχθεί αυτό, δημιουργήθηκε ένας αλγόριθμος, ο οποίος ξεκινώντας από τη αρχή της χρονοσειράς, εισάγει ανάμεσα στις παρατηρήσεις της το μέσο όρο της προηγούμενης και της επόμενης τιμής. Όταν τελειώσει αυτή τη διαδικασία και το μήκος της χρονοσειράς δεν είναι 126, ξαναρχίζει από την αρχή εκτελώντας την ίδια διαδικασία στη χρονοσειρά που έχει προκύψει από την προηγούμενη επανάληψη. Μέσω αυτής της μεθοδολογίας, παρατηρήθηκε ότι οι τέσσερις μέθοδοι σύγκρισης δημιουργούσαν ομάδες με πραγματικά όμοιες χρονοσειρές, και επίσης δεν επηρεάστηκε καθόλου η μορφή και το σχήμα της κάθε χρονοσειράς. Παρ' όλα αυτά, παρατηρείται μια πολύ μικρή οριζόντια μετατόπιση, γεγονός που δεν επηρεάζει τη σύγκριση των χρονοσειρών και αυτό επιβεβαιώνεται και από τη μείωση του σφάλματος sMAPE, το οποίο θα παρουσιαστεί στη συνέχεια. Η αλλαγή στις τιμές του x-άξονα δεν επηρεάζει καθόλου τις συγκρίσεις, απλά δείχνει ποιο είναι το index της κάθε παρατήρησης. Παρακάτω φαίνονται δύο παραδείγματα γραφημάτων για την 500^{η} και 200^{η} χρονοσειρά του M3, αντίστοιχα, οι οποίες έχουν τελείως διαφορετική μορφή:



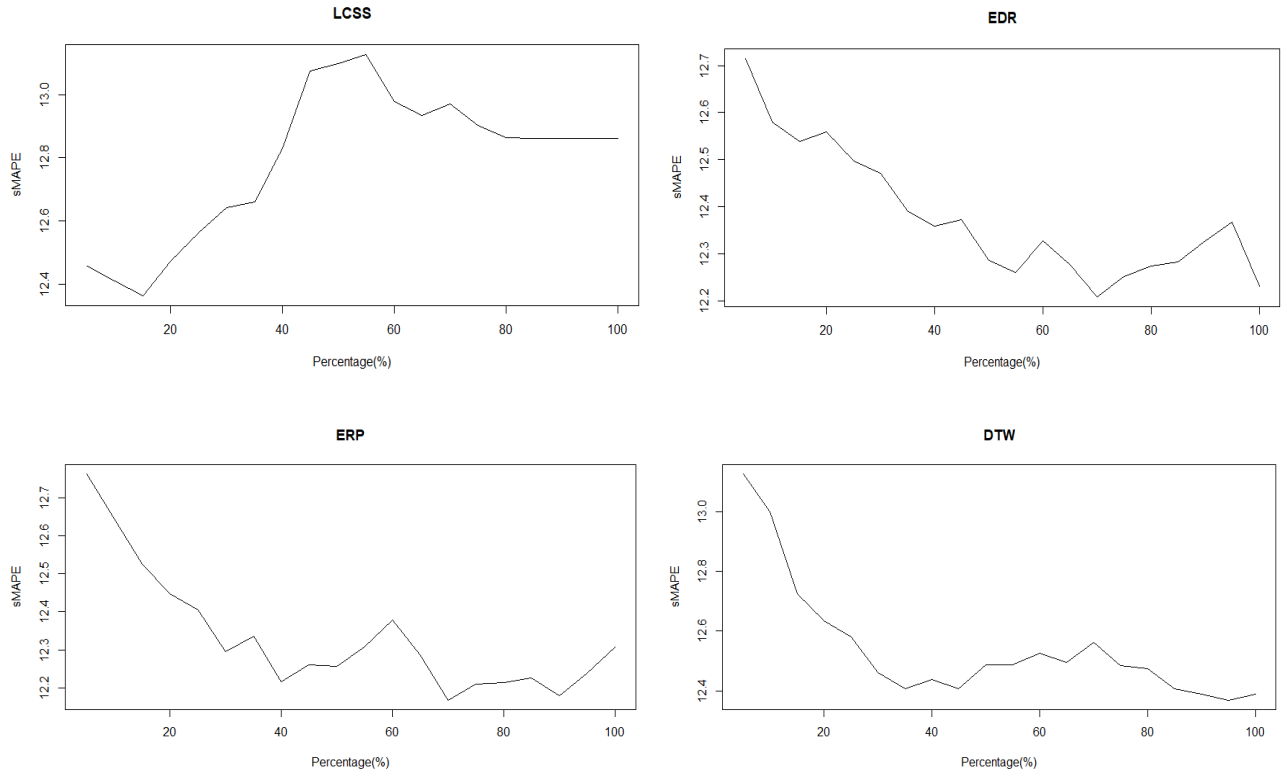
Εικόνα 5.4.2.1: Παράδειγμα μη διαστολής και διαστολής μήκους της 500^{ης} χρονοσειράς του M3, αντίστοιχα



Εικόνα 5.4.2.2: Παράδειγμα μη διαστολής και διαστολής μήκους της 200^{ης} χρονοσειράς του M3, αντίστοιχα

5.4.3 Εύρεση βέλτιστου κατωφλίου και παρουσίαση των προβλέψεων

Σε αυτό το σημείο, αφού δημιουργούνται ομάδες με πραγματικά όμοιες χρονοσειρές, σκοπός είναι να βρεθεί το ποσοστό εκείνο που πρέπει να χρησιμοποιηθεί σε κάθε μέθοδο σύγκρισης, ώστε να παράγει ομάδες χρονοσειρών που εξάγουν τελικώς το μικρότερο δυνατό σφάλμα. Για αυτό το λόγο δημιουργήθηκε μια επαναληπτική διαδικασία, η οποία ξεκινάει από το ποσοστό των 5% μέχρι το ποσοστό των 100% με βήμα 5% και κάθε φορά υπολογίζει το τελικό σφάλμα. Εκτελώντας αυτόν τον αλγόριθμο, προκύπτουν οι εξής τέσσερις γραφικές παραστάσεις:



Εικόνα 5.4.3.1: Οι γραφικές παραστάσεις για τον εντοπισμό του βέλτιστου κατωφλίου μέσω των μεθόδων LCSS, EDR, ERP, DTW, αντίστοιχα

Παρατηρείται ότι οι τιμές των ποσοστών που παράγουν τελικώς το ελάχιστο σφάλμα, καθώς και τα αντίστοιχα σφάλματα είναι τα εξής:

	Percentage (%)	***Minimum sMape (%)***
LCSS	15	12.36
EDR	70	12.21
ERP	70	12.17
DTW	95	12.37

Εικόνα 5.4.3.2: Αποτελέσματα των ποσοστών που παράγουν το ελάχιστο σφάλμα και το αντίστοιχο ελάχιστο σφάλμα

Με σκοπό την επιβεβαίωση ότι η μέθοδος διαστολής του μήκους των χρονοσειρών είναι ορθή και αποτελεσματική, συγκρίνονται τα σφάλματα που παράγονται από τις ομάδες των όμοιων χρονοσειρών που δημιουργούν οι τέσσερις μέθοδοι σύγκρισης πριν τη διαστολή και μετά τη διαστολή των χρονοσειρών, καθώς και σε ποιο ποσοστό επιτυγχάνονται. Στον παρακάτω πίνακα φαίνεται ότι, εκτός της εξαγωγής πραγματικά όμοιων χρονοσειρών, η μέθοδος που εφαρμόστηκε παράγει και μικρότερα σφάλματα:

	EXTENDED		***NOT EXTENDED***	
	Percentage (%)	***Minimum sMape (%)***	***Percentage (%)***	***Minimum sMape (%)***
LCSS	15	12.36	40	12.66
EDR	70	12.21	100	12.21
ERP	70	12.17	70	12.23
DTW	95	12.37	60	12.46

Εικόνα 5.4.3.3: Αποτελέσματα με και χωρίς τη μέθοδο διαστολής του μήκους των χρονοσειρών

Παρατηρείται ότι με τη χρήση της μεθόδου διαστολής του μήκους των χρονοσειρών, επετεύχθη μείωση των σφαλμάτων για όλες τις μεθόδους σύγκρισης. Η μεγαλύτερη μείωση διακρίνεται στη μέθοδο LCSS, αποτέλεσμα απολύτως αναμενόμενο, καθώς ήταν η μέθοδος που είχε το κύριο πρόβλημα στην παραγωγή ομάδων με πραγματικά όμοιες χρονοσειρές.

5.5 Μεθοδολογία κεντρικής ιδέας

5.5.1 Παραγωγή των προβλέψεων μέσω των κλασικών μεθόδων

Στα προηγούμενα αποτελέσματα δεν αναφέρθηκε ο τρόπος με τον οποίον οι ομάδες που δημιουργούνται παράγουν τα συγκεκριμένα σφάλματα. Αρχικά θα παρουσιαστεί ο μέσος όρος σφαλμάτων sMAPE που παράγει κάθε μέθοδος πρόβλεψης από τις 12 που χρησιμοποιήθηκαν για κάθε χρονοσειρά από τις 3003 του M3. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα:

Method	*sMape (%)*
ARIMA	13.59
SES	13.50
DAMPED	13.09
HOLT	14.94
THETA	12.86
NAÏVE	14.77
sNAÏVE	15.19
NEURAL NETWORK	15.54
RANDOM WALK	14.81
ETS	13.07
TBATS	13.13
STLM	13.23

Εικόνα 5.5.1.1: Αποτελέσματα του μέσου όρου σφαλμάτων της κάθε μεθόδου πρόβλεψης για τις 3003 χρονοσειρές του M3

Σε αυτό το σημείο είναι ιδιαίτερα σημαντικό να επισημανθεί ότι κατά μέσο όρο στις 3003 χρονοσειρές του M3, η καλύτερη μέθοδος είναι η Theta, η οποία αναπτύχθηκε εξ' ολοκλήρου από την Μονάδα Προβλέψεων και Στρατηγικής του Εργαστηρίου Συστημάτων Αποφάσεων και Διοίκησης του Εθνικού Μετσόβιου Πολυτεχνείου.

5.5.2 Διαδικασία παραγωγής τελικών προβλέψεων

Κάθε μέθοδος πρόβλεψης παράγει ένα διαφορετικό σφάλμα (στην περίπτωση μας μετρούμενο μέσω του sMAPE) για κάθε χρονοσειρά και κατά συνέπεια ένα διαφορετικό μέσο σφάλμα για ένα σύνολο χρονοσειρών (στην περίπτωση μας τα δεδομένα του M3). Σκοπός της παρούσας διπλωματικής εργασίας είναι η μείωση ενός τέτοιου μέσου σφάλματος επιλέγοντας για κάθε χρονοσειρά την ακριβέστερη προβλεπτική μέθοδο.

Προκειμένου να καταστεί δυνατή μία τέτοια επιλογή, για κάθε χρονοσειρά δημιουργούνται τέσσερις ομάδες όμοιων χρονοσειρών με την αρχική από το dataset του M3 (μια από κάθε μέθοδο σύγκρισης, δηλαδή τις LCSS, EDR, ERP, DTW). Στη συνέχεια, για κάθε όμοια χρονοσειρά, υπολογίστηκε το σφάλμα που παράγει κάθε μια από τις 12 μεθόδους πρόβλεψης. Έπειτα, επιλέγοντας κάθε φορά τη μέθοδο που παράγει το μικρότερο μέσο σφάλμα στις όμοιες χρονοσειρές, τη χρησιμοποιούσαμε ώστε να επιλέξουμε το σφάλμα εκείνης της μεθόδου στην αρχική χρονοσειρά. Με αυτή τη μεθοδολογία, δημιουργήθηκαν 4 διανύσματα μεγέθους 3003 (ένα διάνυσμα για κάθε μέθοδο σύγκρισης), όπου σε κάθε θέση υπήρχε το σφάλμα της μεθόδου που υποδείξαν οι όμοιες χρονοσειρές της αντίστοιχης αρχικής χρονοσειράς. Τέλος, υπολογίζοντας το μέσο όρο για αυτές τις 3003 τιμές, παρήχθησαν τα τελικά αποτελέσματα, τα οποία παρουσιάζονται στον παρακάτω πίνακα:

Method	*sMape (%)*
ARIMA	13.59
SES	13.50
DAMPED	13.09
HOLT	14.94
THETA	12.86
NAÏVE	14.77
sNAÏVE	15.19
NEURAL NETWORK	15.54
RANDOM WALK	14.81
ETS	13.07
TBATS	13.13
STLM	13.23
LCSS	12.36
EDR	12.21
ERP	12.17
DTW	12.37

Εικόνα 5.5.2.1: Αποτελέσματα του μέσου όρου σφαλμάτων της κάθε μεθόδου πρόβλεψης για τις 3003 χρονοσειρές του M3 και της επιλογής κατάλληλης μεθόδου σύμφωνα με τις ομάδες που δημιούργησαν οι μέθοδοι LCSS, EDR, ERP, DTW

Παρατηρείται εμφανής μείωση των σφαλμάτων χρησιμοποιώντας σε κάθε χρονοσειρά τη μέθοδο που υποδεικνύαν οι όμοιες της. Το ελάχιστο σφάλμα είναι **12.17%** και το παρήγαγε η μέθοδος σύγκρισης **ERP** (Edit distance with Real Penalty).

5.6 Παραγωγή προβλέψεων μέσω χρησιμοποίησης διαφορετικού πλήθους μεθόδων πρόβλεψης

Τα αποτελέσματα της προηγούμενης διαδικασίας δικαιώνουν την ιδέα του πειράματος, καθώς επιτεύχθηκε η μείωση του σφάλματος sMAPE σε 12.17%, χρησιμοποιώντας τη μέθοδο σύγκρισης ERP. Παρ' όλα αυτά, στην παραπάνω διαδικασία χρησιμοποιήθηκαν και οι 12 μέθοδοι πρόβλεψης. Σε αυτό το σημείο, η ίδια διαδικασία επαναλαμβάνεται, χρησιμοποιώντας κάθε φορά διαφορετικό πλήθος μεθόδων, από μία μέχρι έντεκα μεθόδους. Για κάθε διαφορετικό πλήθος μεθόδων, υπολογίστηκε, με την παραπάνω διαδικασία, το σφάλμα που παράγεται, χρησιμοποιώντας κάθε φορά διαφορετικό συνδυασμό μεθόδων. Για παράδειγμα, αν έχουμε 12 μεθόδους και θέλουμε να χρησιμοποιήσουμε 6 από αυτές, τότε οι διαφορετικοί συνδυασμοί μεθόδων που προκύπτουν είναι 924, ενώ αν θέλουμε να χρησιμοποιήσουμε 3 από αυτούς, τότε οι συνδυασμοί είναι 220. Παρακάτω παρουσιάζεται ένας πίνακας που δείχνει το πλήθος των διαφορετικών συνδυασμών που προκύπτουν, εάν χρησιμοποιηθεί ποικίλο πλήθος μεθόδων:

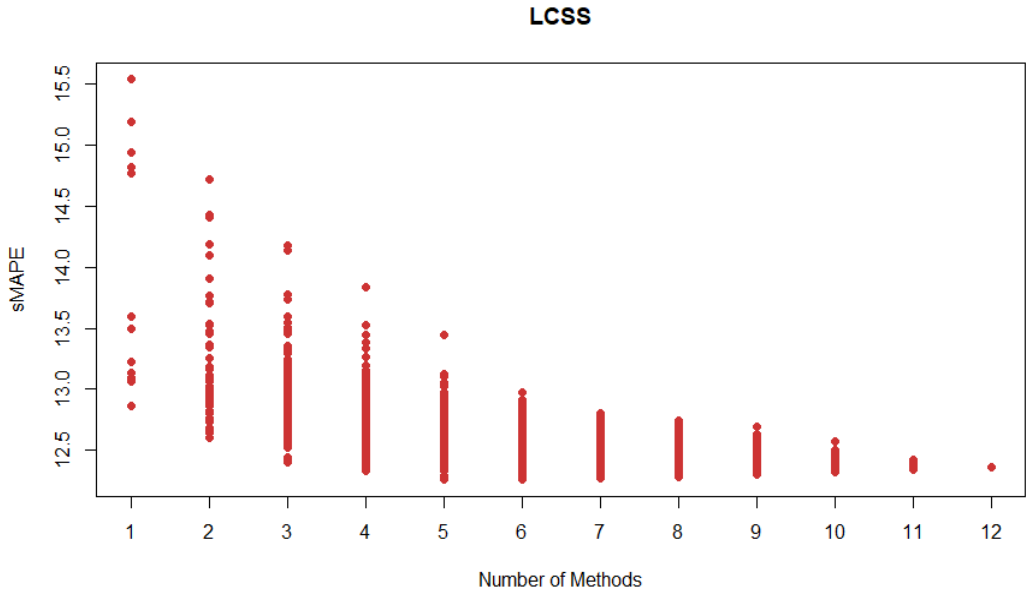
Number of Methods	Number of Combinations
1	12
2	66
3	220
4	495
5	792
6	924
7	792
8	495
9	220
10	66
11	12
12	1

Εικόνα 5.6.1: Πλήθος διαφορετικών συνδυασμών, βάσει του αριθμού των μεθόδων που χρησιμοποιούνται

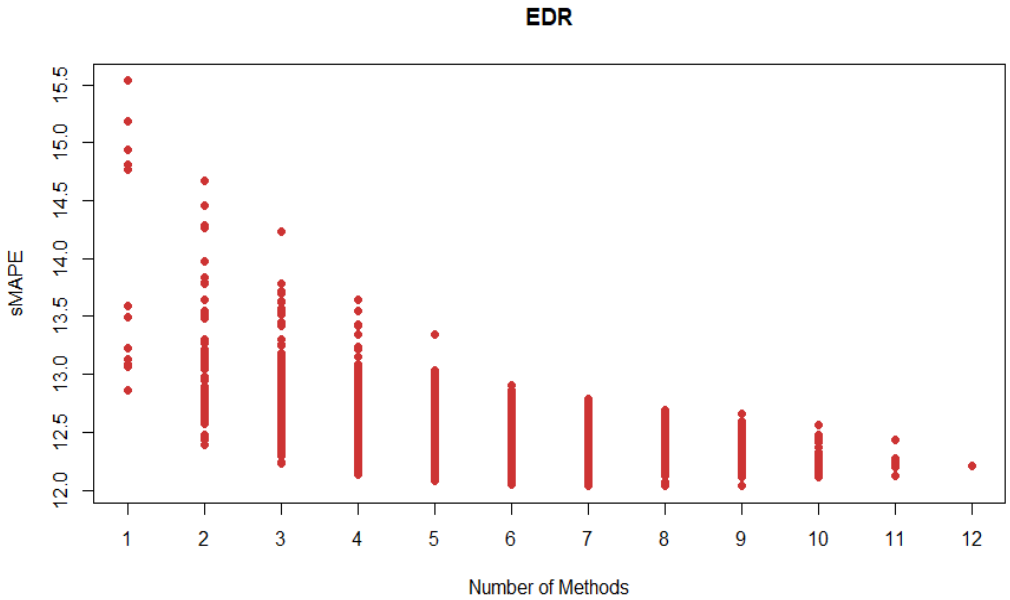
Ο μεγαλύτερος αριθμός συνδυασμών παρατηρείται στις 6 μεθόδους, γεγονός αναμενόμενο καθώς υπάρχουν 12 διαθέσιμες μέθοδοι. Για παράδειγμα, άμα χρησιμοποιούνται 6 μέθοδοι, δύο από όλους τους πιθανούς συνδυασμούς θα είναι οι εξής:

- ARIMA, SES, DAMPED, THETA, ETS, TBATS
- HOLT, THETA, NAÏVE, NEURAL NETWORK, sNAIVE, STLM

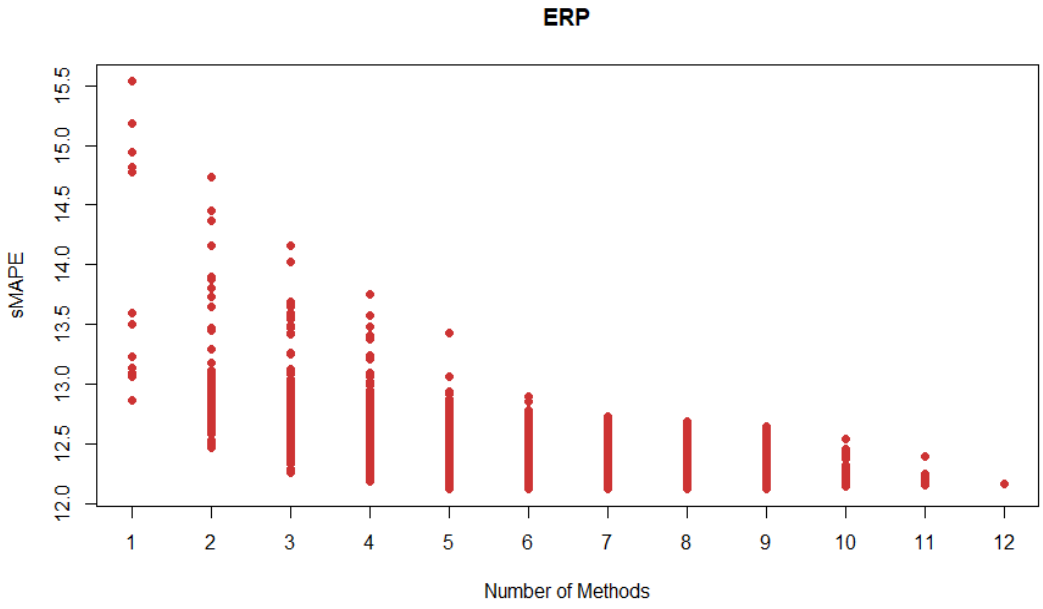
Η παραπάνω διαδικασία επαναλαμβάνεται τέσσερις φορές, όπου κάθε φορά χρησιμοποιείται η ομάδα όμοιων χρονοσειρών που παρήγαγε κάθε μέθοδος σύγκρισης (LCSS, EDR, ERP, DTW). Έχοντας πλέον υπολογίσει το σφάλμα κάθε συνδυασμού για κάθε διαφορετικό πλήθος μεθόδων, δημιουργήθηκε ένα διάγραμμα για κάθε μέθοδο σύγκρισης, όπου δείχνει με κάθετες γραμμές το εύρος των σφαλμάτων. Ο x-άξονας απεικονίζει το πόσες μέθοδοι χρησιμοποιούνται κάθε φορά και ο y-άξονας δείχνει τα σφάλματα. Τα διαγράμματα αυτά είναι τα εξής:



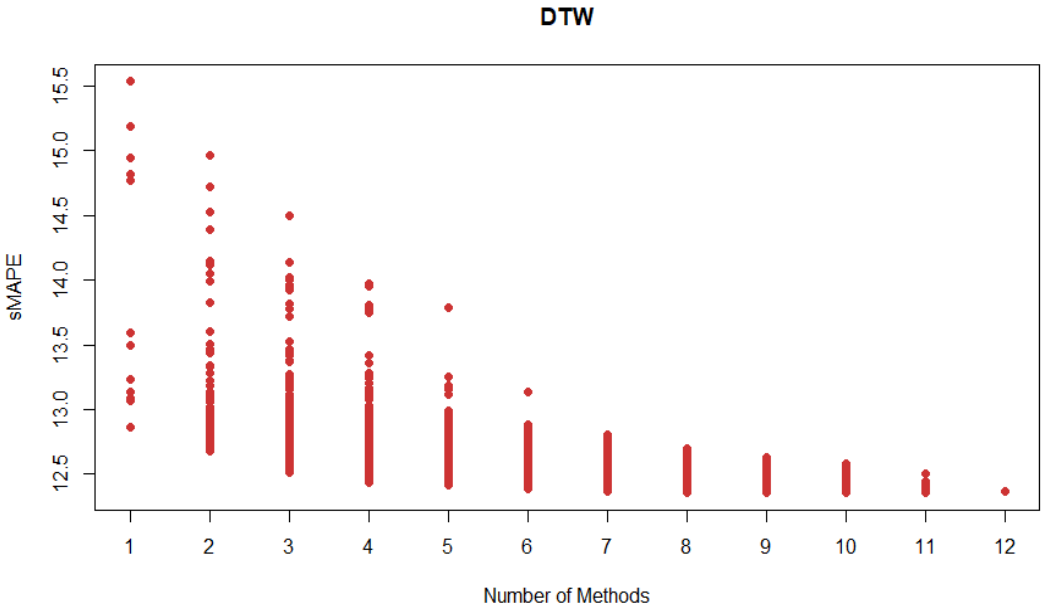
Εικόνα 5.6.2: Εύρος σφαλμάτων για πλήθος μεθόδων που χρησιμοποιούνται από 1 έως 12, μέσω της μεθόδου LCSS



Εικόνα 5.6.3: Εύρος σφαλμάτων για πλήθος μεθόδων που χρησιμοποιούνται από 1 έως 12, μέσω της μεθόδου EDR



Εικόνα 5.6.4: Εύρος σφαλμάτων για πλήθος μεθόδων που χρησιμοποιούνται από 1 έως 12, μέσω της μεθόδου ERP



Εικόνα 5.6.5: Εύρος σφαλμάτων για πλήθος μεθόδων που χρησιμοποιούνται από 1 έως 12, μέσω της μεθόδου DTW

Παρατηρείται ότι καθώς αυξάνεται ο αριθμός των μεθόδων που χρησιμοποιούνται, αυξάνονται και οι επιλογές μας στο σύνολο των σφαλμάτων, ώστε τελικώς να επιλεγεί εκείνη η μέθοδος που παράγει το μικρότερο σφάλμα. Παρ΄ όλο που σε καμία από τις τέσσερις περιπτώσεις το ελάχιστο σφάλμα δεν εμφανίζεται στο μέγιστο πλήθος μεθόδων (δηλαδή στις 12), είναι ξεκάθαρο ότι επιλέγοντας και τις 12 μεθόδους, είμαστε πιο ασφαλείς στην ακρίβεια

των προβλέψεων, με την έννοια ότι το σφάλμα των 12 μεθόδων μπορεί να μην είναι το ελάχιστο, αλλά είναι καλύτερο από τα υπόλοιπα σφάλματα που παράγουν τα άλλη πλήθη μεθόδων. Παρακάτω φαίνονται 4 πίνακες (ένας για κάθε μέθοδο σύγκρισης) με τα ελάχιστα και τα μέγιστα σφάλματα που παράγει κάθε διαφορετικό πλήθος μεθόδων:

Αναλύοντας τα παραπάνω διαγράμματα και τις αντίστοιχες τιμές, το ελάχιστο sMAPE είναι 12.03% και το παράγουν οι ομάδες χρονοσειρών που δημιούργησε η EDR, όταν χρησιμοποιούνται **7 μέθοδοι** και πιο συγκεκριμένα οι ARIMA, SES, DAMPED, THETA, sNAIVE, ETS, TBATS. Συμπερασματικά, παρατηρείται ότι, με τη διαδικασία αυτή, επιτεύχθηκε η εκ νέου μείωση του σφάλματος sMAPE.

5.7 Εναλλακτικές προβλέψεις μέσω ειδικών βαρών

Η επιλογή της κατάλληλης μεθόδου πρόβλεψης για μια χρονοσειρά, με βάση την απόδοση της στις όμοιες χρονοσειρές της, αποδείχθηκε ιδιαίτερα αποτελεσματική, καθώς επιτεύχθηκε η αύξηση της ακρίβειας πρόβλεψης. Στην παρακάτω διαδικασία, θα υλοποιηθεί ένας άλλος τρόπος εύρεσης του σφάλματος των προβλέψεων.

Οι αρχικές προβλέψεις που παρήγαγε η κάθε μέθοδος πρόβλεψης για κάθε χρονοσειρά του M3 θα συνδυαστούν με κατάλληλα βάρη που προκύπτουν βάσει της απόδοσης των μεθόδων στις όμοιες της κάθε χρονοσειράς. Υπολογίστηκε διαφορετικό βάρος ανά μέθοδο πρόβλεψης και ανά χρονοσειρά. Ενδεικτικά στους παρακάτω πίνακες παρουσιάζονται, για κάθε μέθοδο σύγκρισης, τα ειδικά βάρη που προκύπτουν για τις πέντε πρώτες χρονοσειρές απ' όλες τις μεθόδους πρόβλεψης:

LCSS					
Method	**Time Series**				
	1	2	3	4	5
ARIMA	0.09	0.08	0.06	0.09	0.08
SES	0.69	0.08	0.10	0.07	0.11
DAMPED	0.95	0.08	0.08	0.09	0.09
HOLT	0.09	0.09	0.06	0.09	0.05
THETA	0.09	0.09	0.12	0.08	0.09
NAÏVE	0.07	0.09	0.10	0.07	0.09
sNAIVE	0.06	0.08	0.09	0.06	0.09
NEURAL NETWORK	0.07	0.06	0.05	0.07	0.06
RANDOM WALK	0.10	0.09	0.10	0.10	0.08
ETS	0.09	0.08	0.07	0.09	0.09
TBATS	0.08	0.10	0.09	0.10	0.10
STLM	0.09	0.08	0.07	0.09	0.08

Εικόνα 5.7.1: Ειδικά βάρη για τις πέντε πρώτες χρονοσειρές απ' όλες τις μεθόδους πρόβλεψης μέσω της μεθόδου LCSS

EDR					
Method	Time Series				
	1	2	3	4	5
ARIMA	0.11	0.10	0.10	0.09	0.08
SES	0.05	0.10	0.10	0.07	0.10
DAMPED	0.10	0.06	0.06	0.08	0.09
HOLT	0.10	0.09	0.09	0.10	0.06
THETA	0.07	0.05	0.05	0.08	0.09
NAÏVE	0.05	0.10	0.10	0.07	0.08
sNAIVE	0.04	0.10	0.10	0.07	0.09
NEURAL NETWORK	0.09	0.02	0.02	0.07	0.07
RANDOM WALK	0.10	0.03	0.03	0.10	0.07
ETS	0.10	0.10	0.10	0.09	0.09
TBATS	0.09	0.12	0.12	0.10	0.09
STLM	0.10	0.10	0.10	0.09	0.08

Εικόνα 5.7.2: Ειδικά βάρη για τις πέντε πρώτες χρονοσειρές απ' όλες τις μεθόδους πρόβλεψης μέσω της μεθόδου EDR

ERP					
Method	Time Series				
	1	2	3	4	5
ARIMA	0.09	0.11	0.10	0.08	0.09
SES	0.06	0.09	0.10	0.07	0.10
DAMPED	0.09	0.07	0.06	0.08	0.08
HOLT	0.09	0.13	0.09	0.10	0.07
THETA	0.07	0.07	0.05	0.09	0.09
NAÏVE	0.06	0.09	0.10	0.07	0.10
sNAIVE	0.05	0.01	0.10	0.07	0.09
NEURAL NETWORK	0.10	0.03	0.02	0.07	0.07
RANDOM WALK	0.09	0.06	0.03	0.11	0.09
ETS	0.09	0.09	0.10	0.08	0.08
TBATS	0.11	0.11	0.12	0.09	0.09
STLM	0.09	0.09	0.10	0.08	0.07

Εικόνα 5.7.3: Ειδικά βάρη για τις πέντε πρώτες χρονοσειρές απ' όλες τις μεθόδους πρόβλεψης μέσω της μεθόδου ERP

DTW					
Method	Time Series				
	1	2	3	4	5
ARIMA	0.09	0.08	0.10	0.08	0.08
SES	0.07	0.09	0.10	0.08	0.10
DAMPED	0.09	0.08	0.06	0.08	0.09
HOLT	0.09	0.07	0.09	0.09	0.07
THETA	0.09	0.10	0.05	0.09	0.08
NAÏVE	0.07	0.09	0.10	0.08	0.09
sNAIVE	0.06	0.09	0.10	0.08	0.08
NEURAL NETWORK	0.08	0.06	0.02	0.07	0.05
RANDOM WALK	0.10	0.10	0.03	0.10	0.09
ETS	0.09	0.08	0.10	0.09	0.09
TBATS	0.09	0.08	0.12	0.08	0.09
STLM	0.09	0.08	0.10	0.09	0.09

Εικόνα 5.7.4: Ειδικά βάρη για τις πέντε πρώτες χρονοσειρές απ' όλες τις μεθόδους πρόβλεψης μέσω της μεθόδου DTW

Τα βάρη αυτά είναι σταθερές τιμές και συνδυάζονται σε κάθε παρατήρηση που προβλέπει κάθε μέθοδος πρόβλεψης για κάθε χρονοσειρά, ώστε τελικώς να προκύψει η τελική πρόβλεψη για κάθε χρονοσειρά. Έπειτα για να συγκριθεί αυτή η μέθοδος με την προηγούμενη, υπολογίστηκε το μέσο σφάλμα για τις 3003 χρονοσειρές, με αποτέλεσμα να προκύψουν τέσσερα σφάλματα, ένα για κάθε μέθοδο σύγκρισης. Τα συγκεκριμένα σφάλματα φαίνονται στον παρακάτω πίνακα:

	sMape (%)
LCSS	13.01
EDR	12.92
ERP	12.93
DTW	13.59

Εικόνα 5.7.5: Αποτελέσματα του μέσου όρου σφαλμάτων από τη μέθοδο προσθήκης ειδικών βαρών στις προβλέψεις

Παρατηρείται ότι τα αποτελέσματα είναι ιδιαίτερα αυξημένα συγκριτικά με τα αντίστοιχα σφάλματα της διαδικασίας επιλογής κατάλληλης μεθόδου που παρουσιάστηκε προηγουμένως, γεγονός που δικαιώνει την κύρια ιδέα της διπλωματικής εργασίας. Πιο συγκεκριμένα, το να προβλέπεται κάθε μεμονωμένη χρονοσειρά με τη μέθοδο που έχει την αποδοτικότερη συμπεριφορά στις όμοιες της είναι ένα εξαιρετικά αποτελεσματικό μοντέλο, και

μάλιστα καλύτερο από την περίπτωση που υπολογίζονται κατάλληλα βάρη τα οποία έχουν τη δυνατότητα να αντισταθμίζουν τις προβλέψεις της κάθε μεθόδου.

Κεφάλαιο 6: Αποτελέσματα και Προεκτάσεις

6.1 Σύνοψη αποτελεσμάτων και συμπεράσματα

Συμπερασματικά, η παρούσα διπλωματική εργασία χωρίζεται σε δύο κομμάτια. Αρχικά γίνεται μια αναλυτική βιβλιογραφική επισκόπηση των μεθόδων πρόβλεψης και σύγκρισης χρονοσειρών που χρησιμοποιήθηκαν στο πειραματικό κομμάτι, με στόχο την πλήρη ενημέρωση του αναγνώστη για τα εργαλεία και τις συναρτήσεις που χρησιμοποιήθηκαν. Το κύριο μέρος και ο βασικός σκοπός της εργασίας είναι η ανάπτυξη ενός μοντέλου, όπου για μια χρονοσειρά θα μπορεί να επιλέγεται αυτόματα η κατάλληλη μέθοδος πρόβλεψης ανάμεσα από ένα πλήθος μεθόδων, γνωρίζοντας την απόδοση τους σε ένα σύνολο με όμοιες χρονοσειρές με αυτήν. Το μοντέλο αποδείχθηκε πολύ αποδοτικό, αφού επιτεύχθηκε η μείωση του σφάλματος πρόβλεψης, συγκριτικά με το μέσο σφάλμα που παράγει μεμονωμένα κάθε μια μέθοδος πρόβλεψης για τις χρονοσειρές του M3. Υπενθυμίζεται ότι στη παρούσα διπλωματική το σφάλμα μετρήθηκε μέσω του sMAPE.

Κατά τη διάρκεια της εργασίας εκτελέστηκαν αρκετά επιμέρους πειράματα, ώστε τελικώς να δημιουργηθεί το πλήρες μοντέλο. Αρχικά, αποεποχικοποιήθηκαν όσες χρονοσειρές έδειξε το ειδικό τεστ εποχικότητας ότι χαρακτηρίζονται από έντονη εποχικότητα. Έπειτα, μετά το πέρας της μεθόδου διαστολής του μήκους κάθε χρονοσειράς για να είναι αποδοτικό το μοντέλο, κάθε μια χρονοσειρά από τις 3003 του M3 συγκρίθηκε με τις υπόλοιπες, με σκοπό να ομαδοποιηθούν οι όμοιες της. Παρ' όλα αυτά το κριτήριο για να είναι μια χρονοσειρά όμοια με κάποια άλλη είναι σχετικό, οπότε τέθηκε ένα συγκεκριμένο κατώφλι. Αυτό το κατώφλι προέρχεται μετά από μια επαναληπτική μέθοδο, όπου για κάθε κατώφλι (μέσα σε ένα εύρος τιμών) μετρήθηκε το τελικό σφάλμα, με σκοπό να βρεθεί ένα συγκεκριμένο κατώφλι, το οποίο ομαδοποιεί τις χρονοσειρές με τέτοιο τρόπο, ώστε τελικώς να παράγεται το ελάχιστο σφάλμα.

Στη συνέχεια, αφού έχουν προετοιμαστεί οι ομάδες χρονοσειρών, ήταν η ώρα για την παραγωγή των προβλέψεων. Αρχικά, παράχθηκε το μέσο σφάλμα που παράγει κάθε μέθοδος πρόβλεψης για τις 3003 χρονοσειρές του M3, με σκοπό τη σύγκρισης τους με τις προβλέψεις του τελικού μοντέλου. Το ελάχιστο σφάλμα το παράγει η μέθοδος THETA και είναι ίσο με 12.86%. Οπότε ο σκοπός του μοντέλου που αναπτύχθηκε ήταν η μείωση αυτού του σφάλματος, το οποίο και επιτεύχθηκε και από τις τέσσερις μεθόδους σύγκρισης που ομαδοποιήσαν τις χρονοσειρές. Πιο συγκεκριμένα, από την ομαδοποίηση μέσω της LCSS υπήρξε σφάλμα 12.36%, μέσω της EDR υπήρξε σφάλμα 12.21%, μέσω της ERP υπήρξε

σφάλμα 12.17% και μέσω της DTW υπήρξε σφάλμα 12.37%. Παρατηρούμε ότι το καλύτερο σφάλμα το παράγουν οι ομάδες που δημιούργησε η μέθοδος ERP.

Παρ' όλη την επίτευξη του κύριου στόχου της εργασίας, τα συγκεκριμένα αποτελέσματα παράχθηκαν χρησιμοποιώντας και τις 12 μεθόδους πρόβλεψης, οπότε το επόμενο πείραμα ήταν η εκ νέου παραγωγή προβλέψεων, αυτή τη φορά χρησιμοποιώντας κάθε φορά διαφορετικό πλήθος και συνδυασμό μεθόδων πρόβλεψης. Αφού εκτελέστηκε το παραπάνω πείραμα και υπολογίστηκε ένας μεγάλος αριθμός προβλέψεων και σφαλμάτων, βγήκε το συμπέρασμα ότι το ελάχιστο σφάλμα το δίνει η EDR όταν χρησιμοποιούνται 7 μέθοδοι πρόβλεψης και πιο συγκεκριμένα οι ARIMA, SES, DAMPED, THETA, sNAIVE, ETS, TBATS. Το σφάλμα αυτό είναι 12.03%, άρα είναι ξεκάθαρη η αύξηση της ακρίβειας, συγκριτικά με τα προηγούμενα αποτελέσματα.

Τέλος, υπολογίστηκαν ξανά οι προβλέψεις και τα αντίστοιχα σφάλματα, χρησιμοποιώντας έναν εναλλακτικό τρόπο. Πιο συγκεκριμένα, συνδυάστηκαν οι προβλέψεις κάθε μεθόδου για κάθε χρονοσειρά, χρησιμοποιώντας σε κάθε πρόβλεψη ένα ειδικό βάρος που προέρχεται από την απόδοση της αντίστοιχης μεθόδου πρόβλεψης στις όμοιες χρονοσειρές με αυτήν. Τα τελικά σφάλματα ήταν ανεβασμένα συγκριτικά με τα αντίστοιχα από το κύριο μοντέλο που αναπτύχθηκε, γεγονός που αποδεικνύει την υψηλή αποδοτικότητα και αποτελεσματικότητα της κύριας ιδέας της διπλωματική εργασίας.

Εν κατακλείδι, τα τελικά συμπεράσματα που εξάγονται από το πειραματικό κομμάτι της διπλωματικής εργασίας είναι τα εξής:

- Οι μέθοδοι σύγκρισης χρονοσειρών που χρησιμοποιήθηκαν αποδείχθηκαν άκρως αποτελεσματικές, καθώς δημιούργησαν ομάδες πραγματικά όμοιων χρονοσειρών, όπως αποδεικνύεται από το αποτέλεσμα.
- Οι ομάδες χρονοσειρών που δημιουργήθηκαν και από τις τέσσερις μεθόδους σύγκρισης παρήγαγαν μικρότερο μέσο σφάλμα συγκριτικά με τη μέθοδο THETA. Η καλύτερη μέθοδος σύγκρισης αποδείχθηκε η ERP.
- Από τα διαγράμματα που δημιουργήθηκαν, είναι ξεκάθαρο ότι όσο αυξάνεται το πλήθος των μεθόδων που συμμετέχουν στο διαγωνισμό επιλογής της καταλληλότερης μεθόδου πρόβλεψης, τόσο μειώνεται και το μέσο σφάλμα.
- Η χρήση ειδικών βαρών στις προβλέψεις κάθε μεθόδου για μια χρονοσειρά, ανάλογα με την απόδοση τους στις όμοιες της, δεν ενδείκνυται για την εκ νέου μείωση του μέσου σφάλματος.

6.2 Μελλοντικές προεκτάσεις

Το μοντέλο που αναπτύχθηκε είχε ως σκοπό την αύξηση της ακρίβειας πρόβλεψης των κλασικών μεθόδων για μια οποιαδήποτε χρονοσειρά του M3, έχοντας την πληροφορία της απόδοσης τους στις όμοιες χρονοσειρές με αυτήν, και έπειτα επιλέγοντας αυτόματα για κάθε χρονοσειρά την κατάλληλη μέθοδο πρόβλεψης. Αυτό επιτεύχθηκε ομαδοποιώντας τις χρονοσειρές με κριτήριο την ομοιότητα τους και εξάγοντας την πληροφορία για το ποια μέθοδος πρόβλεψης δίνει το μικρότερο σφάλμα στις όμοιες χρονοσειρές με αυτήν. Η διαδικασία αυτή, εν τέλει, έδωσε πολύ ικανοποιητικά αποτελέσματα, καθώς επιτεύχθηκε ο βασικός στόχος της, δηλαδή η μείωση του σφάλματος, συγκριτικά με τα σφάλματα που δίνει κατά μέσο όρο η κάθε μέθοδος πρόβλεψης στις χρονοσειρές του M3. Παρά την επιτυχία του βασικού στόχου και των πειραμάτων που εκτελέστηκαν, είναι δυνατή η περαιτέρω βελτίωση του μοντέλου αυτού.

Μια πρώτη ενέργεια που προτείνεται είναι η προσθήκη περισσότερων μοντέλων πρόβλεψης στη διαδικασία παραγωγής των προβλέψεων. Με αυτόν τον τρόπο διευρύνονται οι διαθέσιμες επιλογές για την κατάλληλη μέθοδο πρόβλεψης που θα χρησιμοποιηθεί στην εκάστοτε χρονοσειρά.

Η ομαδοποίηση των χρονοσειρών επιτεύχθηκε με τη βοήθεια τεσσάρων μεθόδων σύγκρισης, της LCSS, της EDR, της ERP και της DTW. Κάθε μέθοδος απ' αυτές έδωσε τις δικές της ομάδες χρονοσειρών, οπότε και το δικό της τελικό αποτέλεσμα. Η χρήση περισσότερων μεθόδων σύγκρισης, όπως η Time Warp Edit Distance (TWED) και η Optimal Subsequence Bijection (OSB) είναι ένα σημαντικό βήμα για τη βελτίωση της παρούσας διπλωματικής εργασίας, καθώς θα υπάρχουν περισσότερες ομάδες χρονοσειρών και κατ' επέκταση περισσότερα αποτελέσματα, γεγονός που αυξάνει την πιθανότητα περαιτέρω αύξησης της ακρίβειας πρόβλεψης.

Η ταχύτητα εκτέλεσης του μοντέλου εξαρτάται από την αποδοτικότητα του κώδικα, καθώς και από το πλήθος και το μήκος των χρονοσειρών. Η βελτίωση του κώδικα είναι ικανή να δώσει ταχύτατα αποτελέσματα και σε σύνολα με πολύ μεγαλύτερο πλήθος και μήκος χρονοσειρών. Οπότε, τέλος, προτείνεται η προσπάθεια βελτίωσης της απόδοσής του, καθώς θα ήταν ιδιαίτερα ενδιαφέρουσα η εφαρμογή του μοντέλου σε περισσότερες και μεγαλύτερες χρονοσειρές από αυτές του M3, όπως είναι το σύνολο των χρονοσειρών του M4 που απαρτίζεται από 100.000 διαφορετικές χρονοσειρές. Με αυτόν τον τρόπο, η κάθε χρονοσειρά θα ψάχνει τις όμοιες της από ένα πολύ μεγαλύτερο πλήθος χρονοσειρών, οπότε είναι υψηλή η πιθανότητα να δημιουργηθούν ομάδες με χρονοσειρές που είναι ακόμα πιο όμοιες.

Βιβλιογραφία/Αναφορές

- Assimakopoulos, V. and Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting* **16**, 521-530.
- Augustine D. Pwasong, Saratha A/P. Sathasivam (2015). “Forecasting Performance of Random Walk with Drift and Feed Forward Neural Network Models”, Published online in MECS, I.J. Intelligent Systems and Applications, 2015, 09, 49-56
- Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole.
- Box, G.E.P., Jenkins, G.M., and Reinsel G.C. (1994) "Time Series Analysis: Forecasting and Control", 3rd Edition, Holden-Day.
- Carlien Fick (2017). “Modifying and generalizing the Radon transform for improved curve-sensitive feature extraction”, Stellenbosch Universtiy
- Chein L. (2005): "Similarity Search Over Time Series and Trajectory Data". University of Waterloo
- Chen, L., & Ng, R. (2004). On The Marriage of Lp-norms and Edit Distance. In Proceedings of the Thirtieth International Conference on Very Large Data Bases (pp. 792–803).
- Chen, L., Ozsu, M. T., & Oria, V. (2005).*Robust and Fast Similarity Search for Moving Object Trajectories*. In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data (pp. 491-502).
- Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 6(1), 3-73.
- Cuturi, M. (2011). Fast Global Alignment Kernels. In Proceedings of the 28th International Conference on Machine Learning (pp. 929–936).
- David Meyer and Christian Buchta (2013). proxy: Distance and Similarity Measures. R package version 0.4-10.

- De Livera, A.M., Hyndman, R.J., & Snyder, R. D. (2011), Forecasting time series with complex seasonal patterns using exponential smoothing, *Journal of the American Statistical Association*, **106**(496), 1513-1527.
- Dynamic time warping, Wikipedia
- Esling, P., & Agon, C. (2012). *Time-series data mining.* ACM Computing Surveys, 45(1), 1-34.
- Forecasting, Wikipedia
- Gaidon, A., Harchaoui, Z., & Schmid, C. (2011). A time series kernel for action recognition. In BMVC 2011 - British Machine Vision Conference (pp. 63.1–63.11).
- GenTxWarper: “Mining of gene expression time series with dynamic time warping techniques” (web-site)
- Hawkins John. “Economic forecasting: history and procedures”
- Hyndman and Athanasopoulos (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia.
- Hyndman J. Rob (2008). “Time series and forecasting in R”, Monash University
- Hyndman J. Rob, Athanasopoulos George (2018). “Forecasting: Principles and Practice”, Monash University, Australia
- Hyndman, R.J., Akram, Md., and Archibald, B. (2008) "The admissible parameter space for exponential smoothing models". *Annals of Statistical Mathematics*, **60**(2), 407–426.
- Hyndman, R.J., and Billah, B. (2003) Unmasking the Theta method. *International J. Forecasting*, **19**, 287-290.
- Hyndman, R.J., Koehler, A.B., Ord, J.K., and Snyder, R.D. (2008) *Forecasting with exponential smoothing: the state space approach*, Springer-Verlag.
- Hyndman, R.J., Koehler, A.B., Snyder, R.D., and Grose, S. (2002) "A state space framework for automatic forecasting using exponential smoothing methods", *International J. Forecasting*, **18**(3), 439–454.
- Hyndman, RJ and Khandakar, Y (2008) "Automatic time series forecasting: The forecast package for R", *Journal of Statistical Software*, **26**(3).
- J. Scott Armstrong (2011). “Evaluating Forecasting Methods”, Principles of Forecasting, pp 443-472

- Köknar-Tezel S. (2010). “Two Sides of Outliers: Optimal Subsequence Bijection and Classification of Imbalanced Datasets” Ph.D. Dissertation Defense, TempleUniversity-Department of Computer and Information Sciences.
- Lei, H., & Sun, B. (2007). A Study on the Dynamic Time Warping in Kernel Machines. In 2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System (pp. 839–845).
- Liao, T. W. (2005). *Clustering of time series data-a survey*. Pattern Recognition, 38(11), 1857-1874.
- Longest common subsequence problem, Wikipedia
- Longin Jan Latecki, Qiang Wang, Suzan Köknar-Tezel, Vasileios Megalooikonomou (2007). “Optimal Subsequence Bijection”, Temple University-Department of Computer and Information Sciences.
- Makridakis and Hibon (2000) The M3-competition: results, conclusions and implications. *International Journal of Forecasting*, **16**, 451-476.
- Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler (1982) The accuracy of extrapolation (time series) methods: results of a forecasting competition.*Journal of Forecasting*, **1**, 111–153.
- Marteau, P.-F., & Gibet, S. (2014). On Recursive Edit Distance Kernels With Applications To Time Series Classification. IEEE Transactions on Neural Networks and Learning Systems, PP(6), 1–13.
- Mori U., Mendiburu A., Lozano A. J. "Distance Measures for Time Series in R: The TSdist Package". University of the Basque Country UPV/EHU
- Mori, A.; Uchida, S.; Kurazume, R.; Taniguchi, R.; Hasegawa, T. & Sakoe, H. *Early Recognition and Prediction of Gestures* Proc. 18th International Conference on Pattern Recognition ICPR 2006, 2006, 3, 560-563
- Muller M. *Dynamic Time Warping* in*Information Retrieval for Music and Motion*. Springer Berlin Heidelberg; 2007. p. 69-84.
- Neil A. Gershenfeld and Andreas S. Weigend (1992). “Time Series Prediction(Forecasting The Future And Understanding The Past)”, 1st Edition, pp 2.
- Peter Ellis (2016). “Error, Trend, Seasonality-ets and its forecast model friends”, in his personal blog “free range statistics”

- Pinheiro, J.C., and Bates, D.M. (2000) "Mixed-Effects Models in S and S-PLUS", Springer
- Pree, H., Herwig, B., Gruber, T., Sick, B., David, K., & Lukowicz, P. (2014). On general purpose time series similarity measures and their use as kernel functions in support vector machines. Information Sciences, 281, 478–495.
- R. B. Cleveland, W. S. Cleveland, J.E. McRae, and I. Terpenning (1990) STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, **6**, 3–73.
- Rabiner L, Rosenberg A, Levinson S (1978).*Considerations in dynamic time warping algorithms for discrete word recognition.* IEEE Trans. Acoust., Speech, Signal Process., 26(6), 575-582. ISSN 0096-3518.
- Ramasubramanian V., Martin Xavier, K.A. and Anathan, P.S. (2014). “Shrimp Data Modelling using Statistical Tools: State Space based Exponential Smoothing & Response Surface Methodology”
- Roger Jang, "Data Clustering and Pattern Recognition" (website)
- Ruey S. Tsay (2000). “Time Series and Forecasting: Brief History and Future Research Journal of the American Statistical Association Vol. 95, No. 450 (Jun., 2000), pp. 638-643
- Sakoe, H. *Two-level DP-matching–A dynamic programming-based pattern matching algorithm for connected word recognition*Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on, 1979, 27, 588-595
- Sakoe, H.; Chiba, S., *Dynamic programming algorithm optimization for spoken word recognition,* Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on , vol.26, no.1, pp. 43-49, Feb 1978.
- Toni Giorgino. *Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package.* Journal of Statistical Software, 31(7), 1-24
- Tormene, P.; Giorgino, T.; Quaglini, S. & Stefanelli, M. *Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation.*Artif Intell Med, 2009, 45, 11-34.
- Vaida Pilinkienė (2008). “Selection of Market Demand Forecast Methods: Criteria and Application”, ISSN 1392-2785 ENGINEERING ECONOMICS. 2008. No 3 (58).
- Vlachos, M., Kollios, G., & Gunopulos, D. (2002). *Discovering similar multidimensional trajectories*. In Proceedings 18th International Conference on Data Engineering (pp. 673-684). IEEE Comput. Soc. doi:10.1109/ICDE.2002.994784

- Wang, X, Smith, KA, Hyndman, RJ (2006) "Characteristic-based clustering for time series data", *Data Mining and Knowledge Discovery*, **13**(3), 335-364.
- Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2012).*Experimental comparison of representation methods and distance measures for time series data.* Data Mining and Knowledge Discovery, 26(2), 275-309.
- Νευρωνικό δίκτυο, ΒΙΚΙΠΑΙΔΕΙΑ

Διπλωματικές Εργασίες-Βιβλία-Σημειώσεις

- Διοίκηση Επιχειρήσεων Πάτρα – ΤΕΙ Δυτικής Ελλάδας (Τεχνικές Προβλέψεων & Ελέγχου) "ΤΕΧΝΙΚΕΣ ΠΡΟΒΛΕΨΕΩΝ & ΕΛΕΓΧΟΥ ΜΑΘΗΜΑ ΘΕΩΡΙΑΣ-ΣΤΑΣΙΜΕΣ ΔΙΑΔΙΚΑΣΙΕΣ-ΥΠΟΔΕΙΓΜΑΤΑ ARIMA (p,d,q)"
- Δρ. Καμπουρίδης Γεώργιος "Συστήματα Πρόβλεψης"
- Εμίρης Δ.(2012) "ΠΡΟΒΛΕΨΕΙΣ"
- Καλαμβόκη Γ.(2017). "Μέθοδοι πρόβλεψης χρονοσειρών:Χρονοσειρές στην ελληνική οικονομιία"
- Κοκολάκης Γ. Ε. Math.ntua.gr (2010). "ΑΝΑΛΥΣΗ ΧΡΟΝΟΣΕΙΡΩΝ"
- Κουμεντάκος Ι. Αθανάσιος - Γεωργακάκος (2017). "Ανάπτυξη Διαδικτυακής Εφαρμογής Προβλέψεων Μεθόδων Χρονοσειρών "
- Λαδάς Γ. Παναγιώτης (2014). "Βραχυπρόθεσμη πρόβλεψη ενεργειακής ζήτησης-Προσεγγίσεις βασισμένες στη Μηχανική Μάθηση"
- Λεγάκη Ι. Νικολέττα-Ζαμπέτα (2012). "Μελέτη Εναλλακτικών Προσεγγίσεων της Μεθοδολογίας Croston μέσω Εμπειρικής Αξιολόγησης."
- Μηλιώνης Α. (2015) "ΧΡΟΝΟΣΕΙΡΕΣ", Σημειώσεις Πανεπιστημιακών Παραδόσεων-Πανεπιστήμιο Αιγαίου.
- Μπροκαλάκης Ι. (2014). "ΠΡΟΒΛΕΨΗ ΦΑΡΜΑΚΕΥΤΙΚΩΝ ΠΩΛΗΣΕΩΝ ΜΕ ΤΗΝ ΧΡΗΣΗ ΤΕΧΝΗΤΩΝ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ ΚΑΙ ΝΕΥΡΟ-ΑΣΑΦΩΝ ΣΥΣΤΗΜΑΤΩΝ"
- Νικολάου Ι. Ευθύμιος (2007). "ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ ΚΑΙ ΕΦΑΡΜΟΓΗ ΓΡΑΜΜΙΚΩΝ, ΜΗ-ΓΡΑΜΜΙΚΩΝ ΚΑΙ ΝΕΥΡΟ-ΑΣΑΦΩΝ ΜΕΘΟΔΩΝ, ΓΙΑ ΤΗ ΒΡΑΧΥΠΡΟΘΕΣΜΗ ΠΡΟΒΛΕΨΗ ΠΑΡΑΓΩΓΗΣ ΕΝΕΡΓΕΙΑΣ ΑΠΟ ΑΙΟΛΙΚΑ ΠΑΡΚΑ"

- Νταβέλης Ε.(2017). “Βελτίωση ακρίβειας στατιστικών μεθόδων πρόβλεψης σε χρονοσειρές μικρού ιστορικού με χρήση τεχνικών συσταδοποίησης εποχιακών δεικτών συναφών δεδομένων.”
- Πετρόπουλος Φ., Ασημακόπουλος Β., (2011). Επιχειρησιακές Προβλέψεις, εκδόσεις συμμετρία, Αθήνα.
- Σπυλιώτης Ε. (2013). “Πρόβλεψη ενεργειακής κατανάλωσης κτηρίων και εντάσεων χρήσης ηλεκτρικής ενέργειας με χρήση δεικτών ενεργειακής κατανάλωσης.”
- Τιτομιχελάκη Κ. Μαρία (2009). “ΠΡΟΒΛΕΨΗ ΜΕΤΟΧΩΝ ΜΕ ΤΗ ΧΡΗΣΗ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ”