



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ
ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ & ΣΥΣΤΗΜΑΤΩΝ
ΑΠΟΦΑΣΕΩΝ

**Ανάπτυξη πλαισίου αυτοματοποιημένης
προέκτασης χρονοσειρών μέσω της γενίκευσης
της μεθόδου πρόβλεψης θ**

Διδακτορική Διατριβή

Ευάγγελος Ν. Σπηλιώτης

Αθήνα, Σεπτέμβριος 2017



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ
ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ & ΣΥΣΤΗΜΑΤΩΝ
ΑΠΟΦΑΣΕΩΝ

Ανάπτυξη πλαισίου αυτοματοποιημένης προέκτασης χρονοσειρών μέσω της γενίκευσης της μεθόδου πρόβλεψης θ

Διδακτορική Διατριβή

Ευάγγελος Ν. Σπηλιώτης

Συμβουλευτική Επιτροπή: Βασίλειος Ασημακόπουλος, Καθηγητής Ε.Μ.Π
Ιωάννης Ψαρράς, Καθηγητής Ε.Μ.Π
Δημήτριος Ασκούνης, Καθηγητής Ε.Μ.Π

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 22^η Σεπτεμβρίου 2017.

Βασίλειος Ασημακόπουλος
Καθηγητής Ε.Μ.Π

Ιωάννης Ψαρράς
Καθηγητής Ε.Μ.Π

Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π

Χρυσόστομος Δούκας
Επίκουρος Καθηγητής
Ε.Μ.Π.

Κωνσταντίνος Νικολόπουλος
Καθηγητής Bangor
University

Φώτιος Πετρόπουλος
Αναπληρωτής Καθηγητής
University of Bath

Σπύρος Μακριδάκης Πρύτανης
Νεάπολις Πάφου & Ομότιμος
Καθηγητής INSEAD

Αθήνα, Σεπτέμβριος 2017

Ευάγγελος Ν. Σπηλιώτης

Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © **Ευάγγελος Ν. Σπηλιώτης**, 2017

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

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 τον συγγραφέα.

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

*Ο σκοπός των προβλέψεων δεν είναι η πρόβλεψη του μέλλοντος αλλά η κατανόηση του
τι χρειάζεσαι για να πραγματοποιήσεις λογικές ενέργειες στο παρόν*

Paul Saffo

Πρόλογος

Η παρούσα διδακτορική διατριβή εκπονήθηκε στα πλαίσια των ερευνητικών δραστηριοτήτων της Μονάδας Προβλέψεων και Στρατηγικής του Εργαστηρίου Αποφάσεων, Συστημάτων και Διοίκησης κατά τα ακαδημαϊκά έτη 2014-2017. Η μονάδα υπάγεται στον Τομέα Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων της Σχολής Ηλεκτρολόγων Μηχανικών & Μηχανικών Η/Υ του Εθνικού Μετσόβιου Πολυτεχνείου στην οποία και αποτελώ ενεργό μέλος ως ερευνητικός συνεργάτης από τον Σεπτέμβριο του 2013. Η ενασχόλησή μου με τις προβλέψεις ξεκίνησε σε προπτυχιακό επίπεδο, όταν μέσα από το μάθημα των Τεχνικών Προβλέψεων ήρθα σε επαφή με τους ανθρώπους του εργαστηρίου και ξεκίνησα να ασχολούμαι με τις ενεργειακές προβλέψεις στα πλαίσια της διπλωματικής μου εργασίας. Το πέρας των προπτυχιακών μου σπουδών ακολούθησε η εγγραφή μου ως Υποψήφιος Διδάκτορας του ΕΜΠ τον Φεβρουάριο του 2014 και η συστηματική μου ενασχόληση με το ευρύτερο επιστημονικό πεδίο των προβλέψεων, τόσο σε θεωρητικό όσο και σε πρακτικό επίπεδο. Η ολοκλήρωση της παρούσας διατριβής αποτελεί για μένα σταθμό στην πορεία μου ως ερευνητής και ταυτόχρονα αφετηρία για μία νέα στον τομέα της επιχειρησιακής έρευνας.

Αρχικά, θα ήθελα να ευχαριστήσω τον Καθηγητή μου κ. Βασίλειο Ασημακόπουλο για την ευκαιρία που μου έδωσε να ασχοληθώ με το αντικείμενο των προβλέψεων, την εμπιστοσύνη που μου έδειξε καθ' όλη τη διάρκεια της στενής μας συνεργασίας, αλλά και για τις πολύτιμες συμβουλές και προτάσεις του σε κάθε ζήτημα που προέκυψε κατά την περαίωση της διδακτορικής μου διατριβής. Ακόμα, θα ήθελα να ευχαριστήσω τα άλλα δύο μέλη της συμβουλευτικής μου επιτροπής, τον Καθηγητή κ. Ιωάννη Ψαρρά και τον Καθηγητή κ. Δημήτριο Ασκούνη για την σημαντική υποστήριξή τους, όπως και τον Καθηγητή Κωνσταντίνο Νικολόπουλο και τον Επίκουρο Καθηγητή Χρυσόστομο Δούκα τόσο για τη συμμετοχή τους στην επιτροπή εξέτασης, όσο και για την ευρύτερη ερευνητική μας συνεργασία. Για την τιμή που μου έκανε να συμμετάσχει στην επιτροπή εξέτασης θα ήθελα να ευχαριστήσω ακόμα τον Ομότιμο Καθηγητή Σπύρο Μακριδάκη. Επίσης, ιδιαίτερος θα ήθελα να ευχαριστήσω τον Αναπληρωτή Καθηγητή Φώτιο Πετρόπουλο καθώς ήταν εκείνος που μου κίνησε το ενδιαφέρον να ασχοληθώ με τις προβλέψεις και με έφερε σε επαφή με τον κλάδο της επιχειρησιακής έρευνας. Το ενδιαφέρον που συνεχίζει να δείχνει ήταν καθοριστικό στην εξέλιξη της διατριβής μου και στη γενικότερη πορεία μου.

Θα ήθελα ακόμα να ευχαριστήσω τους συνεργάτες μου στη Μονάδα Αχιλλέα Ράππη, Νικολέττα-Ζαμπέτα Λεγάκη, Φωτεινή Σκιαδά, Χριστίνα Κωνσταντινίδου, Βαγγέλη Νταβέλη, Νίκη Κοντοέ, Αρτέμη Σεμένογλου, Νίκο Σίμο, Ηλέκτρα Σκεπετάρη, Κατερίνα Κουτσούρη, Αντρέα Κούλουμο, Χρήστο Σεραφίδη και Άντρη Τόσκα για τη συνεργασία μας και τη συμβολή τους όλα αυτά τα χρόνια καθώς και την κα. Λένα Μαυροδοπούλου για τη συνεχή βοήθειά της. Ένα ξεχωριστό ευχαριστώ σε όλους τους διπλωματικούς και μεταπτυχιακούς φοιτητές τους οποίους επέβλεψα με αμοιβαία οφέλη.

Τέλος, θα ήθελα να ευχαριστήσω τους δικούς μου ανθρώπους, συγγενείς και φίλους, για τη στήριξή τους όλα αυτά τα χρόνια: Τους γονείς μου Νίκο και Έφη, την αδερφή μου Ελένη, τη φίλη μου Αφροδίτη και τους φίλους μου Κώστα, Λευτέρη, Παύλο, Δημήτρη και Κατερίνα.

Ευάγγελος Ν. Σπηλιώτης

Αθήνα, Σεπτέμβριος 2017

Περίληψη

Αντικείμενο της παρούσας διδακτορικής διατριβής είναι η ανάπτυξη μίας ολοκληρωμένης μεθοδολογίας παραγωγής προβλέψεων που θα καθιστά αποδοτικότερη τη διαδικασία της προέκτασης χρονοσειρών. Στόχος είναι η εν λόγω μεθοδολογία να επιτρέπει την αποτελεσματική και αυτόματη προέκταση μεγάλου πλήθους χρονοσειρών, ανεξαρτήτως τύπου και χαρακτηριστικών, με έναν ενιαίο τρόπο. Έτσι, η διαδικασία της πρόβλεψης θα απλοποιηθεί σημαντικά για επιχειρήσεις και οργανισμούς, παρέχοντας ακριβείς προβλέψεις και μάλιστα χωρίς να αυξάνεται σημαντικά το υπολογιστικό κόστος.

Προκειμένου να επιτευχθεί ο παραπάνω στόχος κρίνεται απαραίτητος αρχικά ο εντοπισμός και η αξιοποίηση όλων εκείνων των μεθόδων που οδηγούν σε συστηματική βελτίωση της προβλεπτικής ακρίβειας και ο συνδυασμός τους εντός ενός ενιαίου πλαισίου. Η εν λόγω διαδικασία έγινε μέσω μίας εκτενούς βιβλιογραφικής επισκόπησης η οποία αφορούσε ευρεία γκάμα μοντέλων και τεχνικών προβλέψεων. Ερευνήθηκαν τα μειονεκτήματα και τα πλεονεκτήματα της κάθε μεθόδου, αξιοποιήθηκαν οι πιο αποδοτικές και προτάθηκαν νέες που συνδυάζουν τα θετικά τους στοιχεία.

Πιο συγκεκριμένα, σε πρώτη φάση εντοπίστηκαν οι παράγοντες που καθιστούν τη μέθοδο πρόβλεψης θ τόσο επιτυχημένη καθώς και τα σημεία στα οποία αυτή επιδέχεται βελτιώσεις. Τα συμπεράσματα της ανάλυσης αξιοποιήθηκαν για την επέκταση της μεθόδου προκειμένου αυτή να μπορεί να μοντελοποιεί αποδοτικά μη γραμμικά μοτίβα τάσης και να συνδυάζει τις συνιστώσες τάσης και επιπέδου είτε με προσθετικό, είτε με πολλαπλασιαστικό τρόπο. Η εν λόγω επέκταση γενικεύει ουσιαστικά την κλασική προσέγγιση θ μέσω της δημιουργίας ποικιλίας μοντέλων διαφορετικών ιδιοτήτων. Έτσι, επιλέγοντας για κάθε χρονοσειρά το καταλληλότερο εξ αυτών, δύναται να βελτιωθεί σημαντικά η μέση προβλεπτική ικανότητα της μεθόδου.

Προκειμένου να αυξηθεί περαιτέρω η προβλεπτική ακρίβεια της γενικευμένης μεθόδου θ , αναπτύχθηκε επιπλέον μία νέα τεχνική εξομάλυνσης η οποία εξαλείφει με μη παραμετρικό τρόπο το θόρυβο που πιθανώς εμπεριέχεται στα δεδομένα, αποκαλύπτοντας έτσι το πραγματικό σήμα της χρονοσειράς. Η τεχνική αυτή μπορεί να φανεί ιδιαίτερα χρήσιμη καθώς, σε αντίθεση με άλλες μεθόδους εξομάλυνσης, αξιοποιεί μία μη γραμμική τεχνική για τη συρρίκνωση των τοπικών διακυμάνσεων οδηγώντας σε σχεδόν απόλυτα λείες χρονοσειρές, η πρόβλεψη επί των οποίων γίνεται σαφώς πιο αποτελεσματικά.

Η γενικευμένη μέθοδος θ συνδυάζεται κατάλληλα με την μέθοδο της μη γραμμικής εξομάλυνσης που προτείνεται και άλλες υπάρχουσες τεχνικές για τη διαμόρφωση της τελικής μεθοδολογίας πρόβλεψης. Η αποτελεσματικότητά της εξακριβώθηκε χρησιμοποιώντας σετ από πραγματικά και τεχνητά δεδομένα. Στην πρώτη περίπτωση χρησιμοποιήθηκαν χρονοσειρές από τους διαγωνισμούς προβλέψεων M , ενώ στη δεύτερη από γεννήτρια παραγωγής δεδομένων που κατασκευάστηκε. Τα αποτελέσματα είναι ιδιαίτερος ενθαρρυντικά και αποδεικνύουν την ανωτερότητα της προσέγγισης.

Λέξεις Κλειδιά: Ακρίβεια πρόβλεψης, Εξομάλυνση χρονοσειρών, Μέθοδος θ , Πρόβλεψη χρονοσειρών, Τεχνικές προβλέψεων

Abstract

The current doctoral thesis focuses on the development of an integrated forecasting methodology, able to extrapolate time series in a more efficient way. The main aim of the methodology is therefore to contribute towards automatically predicting large volume of series of various characteristics and types in a consolidated way, leading simultaneously to a robust forecasting process of high accuracy and low computational cost.

In order to achieve such an objective, one must first identify and select the methods which systematically improve forecasting performance and combine them appropriately to form an integrated methodological framework. The literature indicates a variety of forecasting models and techniques for enhancing forecasting accuracy. The pros and cons of many approaches were examined, the best ones were exploited, while new ones were proposed.

More specifically, the reasons which make the *Theta* method so effective in time series extrapolation were first detected and some expansions were suggested to improve it. This led to an expanded version of the method which enables us to accurately forecast time series of non-linear trend patterns, as well as to combine the components of trend and level either additively or multiplicatively. The proposed model generalizes in fact the classic approach of the method by introducing multiple models of different principles. Thus, if the best model is selected per case to extrapolate a series, the average performance of the method is expected to be significantly improved.

To further improve the performance of the proposed forecasting method, a new technique for smoothing time series was introduced. The technique is based on a non-parametric approach for mitigating the effect of the noise component of the data which enables us to extract the real signal of the series and extrapolate it. This is of vital importance since, in contrast to typical approaches reported in the literature, the technique exploits a nonlinear algorithm to shrink local variances and curvatures leading to a far more smoothed series. The smoothed series can then be easily extrapolated by any forecasting model, leading to better forecasts.

The generalized form of the *Theta* method is combined with the smoothing technique introduced and many other algorithms found in the literature to develop a complete forecasting methodology. The effectiveness of the methodology is evaluated using both real and generated data. The real data refer to the time series of the M competitions, while the generated data are simulated using a process which was developed for this purpose. That way the conclusions extracted are representative and based on multiple series of various characteristics. The results of the thesis are more than encouraging, reporting the superiority of the proposed approach.

Key Words: Forecasting Accuracy, Smoothing, Theta method, Time series forecasting, Forecasting techniques

Περιεχόμενα

Πρόλογος	7
Περίληψη	9
Abstract	10
Περιεχόμενα	11
Περιεχόμενα Εικόνων	14
Περιεχόμενα Πινάκων	17
Ευρεία Περίληψη	21
Α. Εισαγωγή	21
Β. Χρονοσειρές και Ανάλυση	27
Γ. Μοντέλα πρόβλεψης	30
Δ. Μέθοδοι εξομάλυνσης	33
Ε. Γεννήτρια Τεχνητών δεδομένων	35
ΣΤ. Ανάλυση και εξέταση αποτελεσμάτων	37
Ζ. Συμπεράσματα και προοπτικές	38
Κεφάλαιο 1: Εισαγωγή	43
1.1 Γενικά για τις προβλέψεις	43
1.2 Αντικείμενο και στόχος της διατριβής	45
1.3 Συμβολή της διατριβής	49
1.4 Δομή της διατριβής	50
1.5 Ερευνητικά αποτελέσματα της διατριβής	52
Κεφάλαιο 2: Ανάλυση χρονοσειρών και απόδοσης μοντέλων πρόβλεψης	57
2.1 Ποιοτικά χαρακτηριστικά χρονοσειρών	57
2.2 Αποσύνθεση χρονοσειρών	60
2.2.1 Κλασική μέθοδος αποσύνθεσης	62
2.2.2 Μέθοδοι συρρίκνωσης συντελεστών εποχιακότητας	65
2.3 Ανάλυση χρονοσειρών	69
2.3.1 Βασικοί στατιστικοί δείκτες	69
2.3.2 Στασιμότητα χρονοσειρών	71
2.3.3 Στατιστικοί Έλεγχοι	73
2.3.3.2 Έλεγχος κανονικότητας	75
2.3.3.3 Έλεγχος στασιμότητας	78

2.3.3.4 Έλεγχος εποχιακότητας.....	79
2.4 Διαφόριση και μετασχηματισμοί χρονοσειρών	80
2.4.1 Διαφόριση.....	81
2.4.2 Μετασχηματισμοί	82
2.5 Στατιστική ανάλυση απόδοσης μοντέλων πρόβλεψης.....	88
2.5.1 Ορισμοί και μέθοδοι αξιολόγησης	88
2.5.2 Στατιστικοί δείκτες αξιολόγησης.....	92
Κεφάλαιο 3: Μοντέλα πρόβλεψης	99
3.1 Γενικά για τα μοντέλα πρόβλεψης	99
3.2 Κατηγορίες μοντέλων πρόβλεψης.....	99
3.3 Μοντέλο χρονοσειρών.....	100
3.3.1 Απλοϊκή μέθοδος (Naïve)	101
3.3.2 Μέθοδος αποσύνθεσης	101
3.3.3 Μοντέλα εκθετικής εξομάλυνσης	102
3.3.4 Αυτοπαλινδρομούμενα μοντέλα κινητού μέσου όρου (ARMA).....	107
3.4 Η μέθοδος θ (Theta method).....	113
3.4.1 Συνοπτική παρουσίαση της μεθόδου και βιβλιογραφική επισκόπηση	113
3.4.2 Η μέθοδος θ και κάποιες ενδιαφέρουσες διαπιστώσεις.....	116
3.4.3 Επέκταση της μεθόδου θ για χρονοσειρές μη γραμμικής τάσης.....	120
3.4.4 Αθροιστικό και πολλαπλασιαστικό μοντέλο θ	124
Κεφάλαιο 4: Εξομάλυνση χρονοσειρών	131
4.1 Προετοιμασία χρονοσειρών: Αποκάλυψη μοτίβων και μείωση της αβεβαιότητας κατά την παραμετροποίηση των μοντέλων πρόβλεψης	132
4.2 Εξομάλυνση χρονοσειρών με χρήση κινητών μέσων όρων.....	136
4.3 Εξομάλυνση χρονοσειρών με χρήση Kernel.....	145
4.4 Εξομάλυνση χρονοσειρών μέσω συρρίκνωσης τοπικών καμπυλοτήτων	154
4.4.1 Ο κλασικός μετασχηματισμός θ	155
4.4.2 Ο μη γραμμικός μετασχηματισμός θ	156
4.4.5 Μετασχηματισμός εξισορρόπησης τοπικών διακυμάνσεων	159
Κεφάλαιο 5: Παραγωγή τεχνητών χρονοσειρών μέσω προσομοίωσης.....	171
5.1 Διαγωνισμοί προβλέψεων.....	171
5.1.1 Διαγωνισμοί M.....	172

5.1.2 Διαγωνισμοί Προβλέψεων Neutral Networks	177
5.1.3 Διαγωνισμός T	177
5.1.4 Προβληματισμός πάνω στα δεδομένα των διαγωνισμών προβλέψεων	178
5.2 Γεννήτρια παραγωγής τεχνητών χρονοσειρών	181
5.2.1 Ανάλυση χαρακτηριστικών χρονοσειράς	182
5.2.2 Μεθοδολογία παραγωγής τεχνητών χρονοσειρών	190
5.2.3 Υλοποίηση γεννήτριας και παραγωγή χρονοσειρών	194
5.3 Υποστήριξη της πληρότητας των παραγόμενων χρονοσειρών	199
Κεφάλαιο 6: Πειραματική διάταξη - Ανάλυση και εξέταση των αποτελεσμάτων	211
6.1 Παρουσίαση της προτεινόμενης μεθοδολογίας	212
6.2 Πειραματική διάταξη	218
6.2.1 Αποτελέσματα μεθοδολογίας για τα τεχνητά δεδομένα της γεννήτριας	220
6.2.2 Αποτελέσματα μεθοδολογίας για τα δεδομένα των διαγωνισμών M	230
Κεφάλαιο 7: Συμπεράσματα και προεκτάσεις	247
7.1 Ανασκόπηση	247
7.1 Συμπεράσματα διατριβής	250
7.2 Προεκτάσεις διατριβής	254
Βιβλιογραφία	257
Παράρτημα	263
Λίστα Δημοσιεύσεων	263
Π1. Σε επιστημονικά περιοδικά	263
Π2. Σε πρακτικά επιστημονικών συνεδρίων	263
Π3. Σε βιβλία	264
Π4. Σε κλαδικά περιοδικά	264
Π5. Σε διεθνή επιστημονικά συνέδρια	264
Π6. Σε φοιτητικά συνέδρια:	265

Περιεχόμενα Εικόνων

Εικόνα 2.1: Μηνιαίες χρονοσειρές του διαγωνισμού M3 με έντονο το συστατικό της τάσης, εκθετικής (αριστερά) και γραμμικής (δεξιά).	58
Εικόνα 2.2: Μηνιαία χρονοσειρά του διαγωνισμού M3 με έντονο το συστατικό της κυκλικότητας.	59
Εικόνα 2.3: Μηνιαία χρονοσειρά του διαγωνισμού M3 με έντονο το συστατικό της εποχιακότητας.	59
Εικόνα 2.4: Μηνιαία χρονοσειρά του διαγωνισμού M3 με έντονο το συστατικό της τυχαιότητας.	60
Εικόνα 2.5: Μηνιαίες χρονοσειρές του διαγωνισμού M3 με (αριστερά) additive outliers και (δεξιά) level shifts.	60
Εικόνα 2.6: Ανάλυση χρονοσειράς με κλασική αποσύνθεση στα συστατικά trend (T), circle (C), seasonality (S) και randomness (R).	65
Εικόνα 2.7: Υπολογισμός δεικτών εποχιακότητας για μηνιαία χρονοσειρά του διαγωνισμού M3 με και χωρίς συρρίκνωση συντελεστών.	68
Εικόνα 2.8: Γραφική απεικόνιση της κατανομής των παρατηρήσεων μίας χρονοσειράς με σχετική κυρτότητα.	76
Εικόνα 2.9: Χρήση Q-Q plot για τον έλεγχο της κανονικότητας των δεδομένων μίας χρονοσειράς με παρατηρήσεις κατανομημένες στο δείγμα της ανομοιόμορφα (αριστερά) και ομοιόμορφα (δεξιά), αντίστοιχα.	76
Εικόνα 2.10: Εποχιακή μηνιαία χρονοσειρά (Data) πριν και μετά τη συνδυαστική διαφορίση πρώτης τάξης (Diff) και εποχιακής διαφορίσης (Seasonal Diff).	82
Εικόνα 2.11: Σύγκριση γραφήματος της περιθώριας κατανομής μίας μη κανονικής χρονοσειράς πριν (αριστερά) και μετά (δεξιά) την εφαρμογή του μετασχηματισμού Box-Cox.	84
Εικόνα 2.12: Παράδειγμα επιλογής βέλτιστης παραμέτρου λ με χρήση της μεθόδου των Box-Cox.	85
Εικόνα 2.13: Αποτέλεσμα της εφαρμογής μετασχηματισμών σε μία χρονοσειρά που ακολουθεί μη κανονική κατανομή (p-value=0).	86
Εικόνα 2.14: Αλληλεπίδραση προκατάληψης (bias) και ακρίβειας (variance) συναρτήσει της πολυπλοκότητας (flexibility) των μοντέλων.	89
Εικόνα 2.15: Παρουσίαση μεθόδου rolling origin με ανανέωση του fit period σε κάθε επανάληψη κατά w.	91
Εικόνα 3.1: Μεταβολή της τιμής των βαρών συνδυασμού δύο γραμμών θ όταν η μία εξ αυτών είναι η μηδενική.	118
Εικόνα 3.2: Απόδοση της μεθόδου θ για μία χρονοσειρά γραμμικής και μη γραμμικής τάσης.	120
Εικόνα 3.3: Χρήση εναλλακτικών καμπύλων ως Y_i^0 στην κλασική μέθοδο θ	122
Εικόνα 3.4: Χρήση γραμμικού προσθετικού (πράσινο) και πολλαπλασιαστικού (κόκκινο) μοντέλου θ για την προέκταση δύο ετήσιων (πάνω) και μηνιαίων (κάτω) χρονοσειρών.	126
Εικόνα 4.1: Εφαρμογή του μετασχηματισμού λογαρίθμου για περιορισμό της διακύμανσης σε δύο χρονοσειρές με ακραίες τιμές θετικής και αρνητικής κατεύθυνσης.	134
Εικόνα 4.2: Εφαρμογή ΚΜΟ διαφόρων τάξεων σε χρονοσειρά υψηλής τυχαιότητας.	137
Εικόνα 4.3: Εφαρμογή απλού και σταθμισμένου ΚΜΟ(5) με βάρη Henderson.	139
Εικόνα 4.4: Εφαρμογή της τεχνικής backcasting για τη συμπλήρωση των κενών τιμών κατά την εφαρμογή ενός ΚΚΜΟ(12).	141
Εικόνα 4.5: Εξομάλυνση μηνιαίας χρονοσειράς μέσω kernel με χρήση κοινού πυρήνα αλλά διαφορετικού εύρους (5, 7, 9 και 11).	146
Εικόνα 4.6: Καμπύλες βασικών συναρτήσεων kernel.	147
Εικόνα 4.7: Εφαρμογή εξομάλυνσης μέσω kernel εύρους 11 σε μηνιαία χρονοσειρά για χρήση διαφορετικών πυρήνων.	148

Εικόνα 4.8: Ο μετασχηματισμός θ για τιμές παραμέτρου 1, 0.8, 0.6, 0.4 και 0.1.	155
Εικόνα 4.9: Εξομάλυνση τριάδας παρατηρήσεων με χρήση του μη γραμμικού μετασχηματισμού θ για παράμετρο εξομάλυνσης c ίση με 0 (αρχικά δεδομένα), 0.5 και 1.	157
Εικόνα 4.10: Εξομάλυνση μηνιαίας χρονοσειράς με χρήση του μη γραμμικού μετασχηματισμού θ για παράμετρο εξομάλυνσης c ίση με 0.1, 0.5 και 1 και συνθήκη τερματισμού την επίτευξη διακύμανσης μικρότερης του 5%.	158
Εικόνα 4.11: Εξομάλυνση μηνιαίας χρονοσειράς με χρήση του μη γραμμικού μετασχηματισμού θ για παράμετρο εξομάλυνσης c ίση με 0.1 και συνθήκη τερματισμού την επίτευξη διακύμανσης μικρότερης του 50%, 25%, 10% και 1%.	159
Εικόνα 4.12: Εξομάλυνση μίας σειράς τεσσάρων σημείων σε περίπτωση ετερόσημων (αριστερά) και ομόσημων (δεξιά) τοπικών διακυμάνσεων.	161
Εικόνα 4.13: Αποτέλεσμα εξομάλυνσης μηνιαίας χρονοσειράς υψηλής διακύμανσης (μαύρο) με χρήση της τεχνικής εξισορρόπησης τοπικών διακυμάνσεων μετά από 10 (κόκκινο), 50 (μπλε), 100 (πράσινο) και 200 (γαλάζιο) επαναλήψεις.	162
Εικόνα 4.14: Εξομάλυνση μέσω της τεχνικής εξισορρόπησης τοπικών διακυμάνσεων με συνθήκη τερματισμού 0.01% (αριστερά) και 0.001% σε μηνιαία χρονοσειρά του διαγωνισμού M3.	163
Εικόνα 4.15: Διάγραμμα ροής για την ολοκληρωμένη εφαρμογή της προτεινόμενης τεχνικής εξομάλυνσης μέσω της εξισορρόπησης τοπικών διακυμάνσεων.	168
Εικόνα 5.1: Δυσδιάστατη παρουσίαση των δεδομένων του διαγωνισμού M3 βάσει χαρακτηριστικών τους μέσω της μεθοδολογίας των Kang et al., 2017.	180
Εικόνα 5.2: Εφαρμογή διαδικασίας συρρίκνωσης συντελεστών εποχιακότητας Miller-Williams (κόκκινο) για την αποτελεσματικότερη αποτύπωση των χαρακτηριστικών της χρονοσειράς σε σχέση με την κλασική προσέγγιση (μαύρο).	184
Εικόνα 5.3: Κατανομή των χρονοσειρών του διαγωνισμού M3 βάσει έντασης εποχιακότητας.	186
Εικόνα 5.4: Κατανομή των χρονοσειρών του διαγωνισμού M3 βάσει έντασης τυχαιότητας.	187
Εικόνα 5.5: Κατανομή κλίσης χρονοσειρών γραμμικού μοτίβου στα μηνιαία δεδομένα του διαγωνισμού M3.	187
Εικόνα 5.6: Κατανομή κλίσης χρονοσειρών εκθετικού μοτίβου στα μηνιαία δεδομένα του διαγωνισμού M3.	188
Εικόνα 5.7: Κατανομή κλίσης χρονοσειρών λογαριθμικού μοτίβου στα μηνιαία δεδομένα του διαγωνισμού M3.	188
Εικόνα 5.8: Κατανομή κλίσης χρονοσειρών αντίστροφου μοτίβου στα μηνιαία δεδομένα του διαγωνισμού M3.	189
Εικόνα 5.9: Κατανομή κλίσης χρονοσειρών μοτίβου δύναμης στα μηνιαία δεδομένα του διαγωνισμού M3.	189
Εικόνα 5.10: Παραγωγή μίας τυχαίας χρονοσειράς EXP(H,M,L) μέσω της προτεινόμενης γεννήτριας.	192
Εικόνα 5.11: Παραγωγή μίας τυχαίας χρονοσειράς LIN(L,M,H) μέσω της προτεινόμενης γεννήτριας.	192
Εικόνα 5.12: Παραγωγή μίας τυχαίας χρονοσειράς LOG(L,M,L) μέσω της προτεινόμενης γεννήτριας.	193
Εικόνα 5.13: Παραγωγή μίας τυχαίας χρονοσειράς INV(H,M,M) μέσω της προτεινόμενης γεννήτριας. ..	193
Εικόνα 5.14: Παραγωγή μίας τυχαίας χρονοσειράς POW(M,L,H) μέσω της προτεινόμενης γεννήτριας. ...	194
Εικόνα 5.15: Rstudio - Το γραφικό περιβάλλον της R.	195
Εικόνα 5.16: Πλήθος παραγόμενων χρονοσειρών από τη γεννήτρια ανά κατηγορία χρονοσειράς.	198
Εικόνα 5.17: Απεικόνιση ποιοτικών χαρακτηριστικών των μηνιαίων χρονοσειρών του M3 στο χώρο – Γραμμική τάση.	200
Εικόνα 5.18: Απεικόνιση ποιοτικών χαρακτηριστικών των μηνιαίων χρονοσειρών του M3 στο χώρο – Εκθετική τάση.	201

<i>Εικόνα 5.19: Απεικόνιση ποιοτικών χαρακτηριστικών χρονοσειρών γεννήτριας στον χώρο – Γραμμική τάση.....</i>	<i>202</i>
<i>Εικόνα 5.20: Απεικόνιση ποιοτικών χαρακτηριστικών χρονοσειρών γεννήτριας στον χώρο – Εκθετική τάση.....</i>	<i>202</i>
<i>Εικόνα 5.21: Απεικόνιση της έντασης των ποιοτικών χαρακτηριστικών των μηνιαίων χρονοσειρών του M3 στον χώρο με χρήση PCA.....</i>	<i>205</i>
<i>Εικόνα 5.22: Απεικόνιση των χρονοσειρών του M3 (μαύρο) και των μηνιαίων τεχνητών της γεννήτριας (κόκκινο) στο χώρο που ορίστηκε μέσω της PCA και των μεταβλητών των Kang et al., 2017.....</i>	<i>206</i>
<i>Εικόνα 5.23: Η διαδικτυακή εφαρμογή που αναπτύχθηκε για την παροχή των τεχνητών δεδομένων στην ευρύτερη επιστημονική κοινότητα.....</i>	<i>208</i>
<i>Εικόνα 6.1: Διάγραμμα ροής για την ολοκληρωμένη εφαρμογή της προτεινόμενης μεθοδολογίας πρόβλεψης χρονοσειρών.....</i>	<i>217</i>

Περιεχόμενα Πινάκων

Πίνακας 2.1: Προτεινόμενη μέθοδος συρρίκνωσης εποχιακών δεικτών κατά Miller-Williams.....	67
Πίνακας 2.2: Προβλεπτική ακρίβεια κλασικών μοντέλων πρόβλεψης για αποσύνθεση των εποχιακών δεδομένων των διαγωνισμών M3 και M1 μέσω τις κλασικής μεθόδου και εναλλακτικών μεθόδων συρρίκνωσης εποχιακών δεικτών.	68
Πίνακας 2.3: Κατάσταση σφάλματος και έλεγχος υποθέσεως.	74
Πίνακας 2.4: Τιμές του δείκτη t-critical σε συνάρτηση με το διαστήματα εμπιστοσύνης.	80
Πίνακας 2.5: Προβλεπτική ακρίβεια κλασικών μοντέλων πρόβλεψης (sMAPE) για κλασική αποεποχικοποίηση των δεδομένων των διαγωνισμών M1 και M3 και εφαρμογή μετασχηματισμών.	87
Πίνακας 2.6: Παρατηρήσεις και προβλέψεις χρονοσειράς προς στατιστική ανάλυση απόδοσης προβλέψεων.	92
Πίνακας 3.1 Μοντέλα εκθετικής εξομάλυνσης ανάλογα με τον τύπο σφάλματος, εποχιακότητας και τάσης.	103
Πίνακας 3.2: Υπολογισμός μοντέλων εκθετικής εξομάλυνσης προσθετικού σφάλματος. Για τα μοντέλα πολλαπλασιαστικού σφάλματος αντικαθιστούμε τον παράγοντα ϵ_t με $\mu\epsilon\epsilon_t$	104
Πίνακας 3.3: Εκτίμηση παραμέτρων μοντέλων ARIMA μέσω της χρήσης των συντελεστών αυτοσυσχέτισης.....	112
Πίνακας 3.4: Προβλεπτική ακρίβεια μεθόδου θ (sMAPE) για κλασική αποεποχικοποίηση των δεδομένων των διαγωνισμών M1 και M3 και χρήση εναλλακτικών μοτίβων τάσης.	123
Πίνακας 3.5: Προβλεπτική ακρίβεια των προσθετικών και πολλαπλασιαστικών μοντέλων θ (sMAPE) για κλασική πολλαπλασιαστική αποεποχικοποίηση των δεδομένων των διαγωνισμών M1 και M3 και χρήση εναλλακτικών μοτίβων τάσης.	127
Πίνακας 4.1: Τιμές βαρών σταθμισμένου κινητού μέσου όρου για κλασικές εκδοχές ΚΜΟ.	139
Πίνακας 4.2: Απόδοση κλασικών μεθόδων πρόβλεψης χρονοσειρών βάσει sMAPE στα δεδομένα του διαγωνισμού M3 για χρήση διαφόρων μορφών ΚΜΟ – εφαρμογή backcasting μέσω damped.	142
Πίνακας 4.3: Απόδοση κλασικών μεθόδων πρόβλεψης χρονοσειρών βάσει sMAPE στα δεδομένα του διαγωνισμού M3 για εφαρμογή ΚΜΟ(3) και χρήση διαφορετικών μοντέλων πρόβλεψης για εφαρμογή της τεχνικής backcasting.	144
Πίνακας 4.4: Απόδοση συναρτήσεων kernel ως προς τη θεωρητικά βέλτιστη Epanechnikov.....	148
Πίνακας 4.5: Απόδοση κλασικών μεθόδων πρόβλεψης χρονοσειρών βάσει sMAPE στα εξομαλυμένα δεδομένα του διαγωνισμού M3 μέσω kernel - Εύρος 9, backcasting μέσω Theta και χρήση διαφόρων πυρήνων.	149
Πίνακας 4.6: Απόδοση κλασικών μεθόδων πρόβλεψης χρονοσειρών βάσει sMAPE στα εξομαλυμένα δεδομένα του διαγωνισμού M3 μέσω kernel - πυρήνας Triweight, εύρος 9 και backcasting μέσω κλασικών μεθόδων πρόβλεψης.	152
Πίνακας 4.7: Απόδοση κλασικών μεθόδων πρόβλεψης χρονοσειρών βάσει sMAPE στα εξομαλυμένα δεδομένα του διαγωνισμού M3 μέσω kernel - πυρήνας Triweight, backcasting μέσω Theta και χρήση εύρους 5, 7, 9 και 11.	153
Πίνακας 4.8: Απόδοση κλασικών μεθόδων πρόβλεψης χρονοσειρών βάσει sMAPE στα δεδομένα του διαγωνισμού M3 για εξομάλυνση με χρήση της τεχνικής εξισορρόπησης τοπικών διακυμάνσεων με συνθήκη τερματισμού 0.1%, 0.01% και 0.001% αντίστοιχα.	163
Πίνακας 4.9: Σύγκριση μεθόδων εξομάλυνσης ως προς τα χαρακτηριστικά και τις απαιτήσεις τους	164
Πίνακας 5.1: Πλήθος χρονοσειρών διαγωνισμού M1 ανά συχνότητα και κατηγορία.	173
Πίνακας 5.2: Πλήθος χρονοσειρών διαγωνισμού M3 ανά συχνότητα και κατηγορία.	174

Πίνακας 5.3: Αποτελέσματα διαγωνισμού M3 για το σύνολο των δεδομένων (sMAPE) ανά ορίζοντα πρόβλεψης.	175
Πίνακας 5.4: Οι μέθοδοι που δίνουν τα καλύτερα αποτελέσματα στο διαγωνισμό M3 για το σύνολο των δεδομένων ανά κατηγορία.	176
Πίνακας 5.5: Πλήθος χρονοσειρών διαγωνισμού T ανά συχνότητα και κατηγορία.	178
Πίνακας 5.6: Όρια δεικτών για ποιοτικό διαχωρισμό των χρονοσειρών βάσει της έντασης των δομικών τους χαρακτηριστικών.	186
Πίνακας 5.7: Πίνακας συσχετίσεων έντασης των ποιοτικών χαρακτηριστικών των μηνιαίων δεδομένων του M3.	199
Πίνακας 5.8: Συντελεστές δόμησης τεχνητών μεταβλητών PCA για την απεικόνιση των χρονοσειρών του M3 στο χώρο.	204
Πίνακας 6.1: Κατηγοριοποίηση τεχνητών δεδομένων γεννήτριας βάσει έντασης χαρακτηριστικών.	221
Πίνακας 6.2: Επίδοση προτεινόμενης μεθοδολογίας σε σχέση με άλλες απλούστερες προσεγγίσεις στα τεχνητά δεδομένα της γεννήτριας – Αναλυτικά αποτελέσματα.	222
Πίνακας 6.3: Επίδοση προτεινόμενης μεθοδολογίας σε σχέση με άλλες απλούστερες προσεγγίσεις στα τεχνητά δεδομένα της γεννήτριας – Συνοπτικά αποτελέσματα.	223
Πίνακας 6.4: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση εποχιακότητας.	225
Πίνακας 6.5: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση τυχαιότητας.	226
Πίνακας 6.6: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση γραμμικής τάσης.	226
Πίνακας 6.7: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση εκθετικής τάσης.	227
Πίνακας 6.8: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση λογαριθμικής τάσης.	227
Πίνακας 6.9: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση αντίστροφης τάσης.	227
Πίνακας 6.10: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση τάσης σε μορφή δύναμης.	228
Πίνακας 6.11: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση τάσης σε μορφή δύναμης.	229
Πίνακας 6.12: Κατηγοριοποίηση δεδομένων διαγωνισμού M βάσει συχνότητας.	230
Πίνακας 6.13: Κατηγοριοποίηση δεδομένων διαγωνισμού M βάσει είδους εφαρμογής.	231
Πίνακας 6.14: Κατηγοριοποίηση δεδομένων διαγωνισμού M3 βάσει έντασης χαρακτηριστικών.	231
Πίνακας 6.15: Κατηγοριοποίηση δεδομένων διαγωνισμού M1 βάσει έντασης χαρακτηριστικών.	232
Πίνακας 6.16: Επίδοση προτεινόμενης μεθοδολογίας σε σχέση με άλλες απλούστερες προσεγγίσεις στα δεδομένα των διαγωνισμών M – Αναλυτικά αποτελέσματα.	233
Πίνακας 6.17: Επίδοση προτεινόμενης μεθοδολογίας σε σχέση με άλλες απλούστερες προσεγγίσεις στα δεδομένα των διαγωνισμών – Συνοπτικά αποτελέσματα.	234
Πίνακας 6.18: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά ένταση εποχιακότητας.	236
Πίνακας 6.19: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά ένταση τυχαιότητας.	237
Πίνακας 6.20: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά ένταση γραμμικής τάσης.	238

Πίνακας 6.21: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά ένταση εκθετικής τάσης.....	238
Πίνακα 6.22: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά ένταση λογαριθμικής τάσης.	238
Πίνακα 6.23: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά ένταση αντίστροφης τάσης.....	239
Πίνακα 6.24: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά ένταση τάσης σε μορφή δύναμης.	239
Πίνακας 6.25: Επίδοση προτεινόμενης μεθοδολογίας σε σχέση με άλλες απλούστερες προσεγγίσεις στα δεδομένα των διαγωνισμών M ανά διαγωνισμό και συχνότητα.....	240
Πίνακας 6.26: Επίδοση προτεινόμενης μεθοδολογίας σε σχέση με άλλες απλούστερες προσεγγίσεις στα δεδομένα των διαγωνισμών M ανά συχνότητα.....	241
Πίνακας 6.27: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά είδος εφαρμογής.	241

Ευρεία Περίληψη

A. Εισαγωγή

Οι προβλέψεις αποτελούν αναπόσπαστο κομμάτι της καθημερινότητάς μας και βρίσκονται στο επίκεντρο οποιασδήποτε διαδικασίας σχεδιασμού και προγραμματισμού. Αυτό μπορεί να μην γίνεται άμεσα αντιληπτό για διεργασίες μικρής σημασίας, όπως π.χ. το πότε θα πρέπει να ξυπνήσει κάποιος για να φτάσει έγκαιρα στην εργασία του, ωστόσο πραγματοποιείται με συστηματικό και μεθοδικό τρόπο όταν πρόκειται για διεργασίες που αφορούν την οργάνωση και τη λειτουργία επιχειρήσεων.

Το βασικότερο πρόβλημα που σχετίζεται με τις προβλέψεις είναι αυτό της αβεβαιότητας. Κανείς δεν είναι σε θέση να προσδιορίσει επακριβώς το μέλλον ενώ οι διαφοροποιήσεις μεταξύ πρόβλεψης και αλήθειας εξαρτώνται πέραν της προβλεπτικής ικανότητας της εκάστοτε μεθόδου από μία πληθώρα παραγόντων που συχνά είναι αδύνατο να εντοπιστούν και να ληφθούν αποτελεσματικά υπόψη εκ των προτέρων. Αυτή η αβεβαιότητα είναι το στοιχείο που, αν και συχνά κάνει το κοινό να αντιμετωπίζει την επιστήμη των προβλέψεων με ιδιαίτερη καχυποψία, καθιστά τη διαδικασία της πρόβλεψης απαραίτητη. Δεν είναι τυχαίο άλλωστε το γεγονός ότι η ζήτηση σε προβλέψεις αυξάνεται δραματικά σε περιόδους αστάθειας, εκεί δηλαδή που οι μέθοδοι είναι πιθανότερο να αποτύχουν.

Στις μέρες μας η καχυποψία που συνοδεύει τις προβλέψεις έχει ευτυχώς περιοριστεί σημαντικά, κυρίως λόγω της μακροχρόνιας προώθησής τους στον κόσμο των μεγάλων επιχειρήσεων οι οποίες είναι πλέον σε θέση να αντιληφθούν την αξία τους. Τυπικά, η διαδικασία της πρόβλεψης σε μία επιχείρηση διαμορφώνεται ως εξής: Ένας αναλυτής ή ένα εμπορικό πακέτο παράγει στατιστικές προβλέψεις για ένα μέγεθος ενδιαφέροντος όπως π.χ. οι πωλήσεις της επιχείρησης. Στη συνέχεια ένας εμπειρογνώμονας έχοντας υπόψη την κατάσταση της αγοράς και άλλες σχετικές μεταβλητές, όπως π.χ. είναι η προώθηση ενός προϊόντος, αυξομειώνει τις στατιστικές προβλέψεις δίνοντας την τελική του εκτίμηση. Οι εν λόγω προβλέψεις μπορούν κάλλιστα να μεταβληθούν εκ νέου από την ανώτερη διοίκηση της επιχείρησης ή άλλα τμήματά της βάσει στόχων που έχουν τεθεί για την αντίστοιχη περίοδο και άλλων σχετικών παραγόντων για τους οποίους υπάρχει πληροφόρηση. Όπως γίνεται κατανοητό, η όλη διαδικασία που παρουσιάστηκε είναι ιδιαίτερα πολύπλοκη και χρονοβόρα ενώ απαιτεί την περιοδική και μεθοδική ενασχόληση αρκετών προσώπων. Έτσι, η εφαρμογή της καθίσταται πρακτικά εφικτή μόνο για περιορισμένο πλήθος νευραλγικών μεγεθών, όπως π.χ. οι συνολικές πωλήσεις της επιχείρησης. Τι γίνεται ωστόσο όταν μία επιχείρηση πρέπει να αποφανθεί σχετικά με τα αποθέματά της και τις πωλήσεις της σε πληθώρα κωδικών προϊόντων;

Η πρόβλεψη μεγάλου πλήθους χρονοσειρών είναι ένα υπαρκτό πρόβλημα σε επιχειρήσεις οι οποίες καλούνται μέσα σε σύντομο χρονικό διάστημα να καθορίσουν με σχετική ακρίβεια ποιες θα είναι συνολικά και ανά κωδικό προϊόντος οι πωλήσεις τους, η παραγωγή τους, τα έσοδά τους κτλ. Η πρόβλεψη βρίσκεται οργανωτικά στη βάση της επιχείρησης και καθορίζει σε μεγάλο βαθμό τις κινήσεις της, τόσο μικρού βεληνεκούς,

όπως είναι π.χ. η αύξηση της παραγωγής ενός προϊόντος, όσο και μεγάλου, όπως είναι π.χ. οι επενδυτικές της κινήσεις. Βιβλιογραφικά το εν λόγω πρόβλημα έχει απασχολήσει πολλούς ερευνητές οι οποίοι προσπάθησαν μέσα από απλές και πολύπλοκες προσεγγίσεις να δώσουν λύσεις.

Το βασικότερο πρόβλημα μίας διαδικασίας παραγωγής προβλέψεων που αφορά μεγάλο πλήθος χρονοσειρών, είναι όπως αναφέρθηκε ο χρόνος. Προκειμένου να μπορεί να ολοκληρώνεται εντός ικανοποιητικού διαστήματος θα πρέπει αρχικά να είναι πλήρως αυτοματοποιημένη, δηλαδή να μην εμπλέκει πουθενά τον ανθρώπινο παράγοντα, εκτός και αν βέβαια κάτι τέτοιο κρίνεται επιθυμητό εκ των υστέρων. Από τη στιγμή ωστόσο που δεν εμπλέκεται καταρχήν η ανθρώπινη εμπειρία στη διαδικασία της πρόβλεψης, γίνεται κατανοητό πως κανείς δεν μπορεί πλέον να εντοπίσει πιθανές αστοχίες των στατιστικών μεθόδων ούτε και να τις αυξομειώσει αντίστοιχα. Μία τέτοια διαδικασία θα πρέπει λοιπόν να είναι εφοδιασμένη με τέτοια εργαλεία, τεχνικές και μεθόδους που να εξασφαλίζει τη μέγιστη δυνατή ακρίβεια, χωρίς φυσικά να την επιβαρύνει με υπέρογκο υπολογιστικό κόστος. Ταυτόχρονα, δεδομένου ότι η διαδικασία καλείται να αντιμετωπίσει χρονοσειρές οποιασδήποτε συχνότητας, είδους εφαρμογής και χαρακτηριστικών, η επιλογή των παραπάνω οφείλει να γίνει με τέτοιον τρόπο που να εξασφαλίζονται ικανοποιητικά αποτελέσματα σε κάθε περίπτωση.

Το βασικότερο πρόβλημα που καλείται να αντιμετωπίσει κανείς εδώ είναι η επιλογή του βέλτιστου μοντέλου πρόβλεψης βάσει των χαρακτηριστικών της χρονοσειράς (π.χ. συχνότητα, επίπεδο θορύβου, ρυθμός ανάπτυξης). Η επιλογή αυτή είναι μία ιδιαίτερα απαιτητική διαδικασία καθώς εμπεριέχει δύο είδη αβεβαιότητας: εκείνη που σχετίζεται με την επιλογή της μεθόδου αυτής καθαυτής (*model selection*), και εκείνη που σχετίζεται με την παραμετροποίησή της (*model estimation*). Και οι δύο μορφές αβεβαιότητας συνδέονται κυρίως με την τυχαιότητα των δεδομένων που χρησιμοποιούνται για την εκτίμησή τους και την αξιολόγησή τους, όπως επίσης και με το κριτήριο αξιολόγησης. Η μελέτη της καταλληλότητας δύο μεθόδων για την πρόβλεψη π.χ. της ζήτησης πετρελαίου θέρμανσης, θα επηρεαστεί σημαντικά από τη θερμοκρασία που επικρατεί και συνεπώς αν αυτή γίνει σε περίοδο παγετού είναι πιθανό να προταθεί η χρήση ενός μοντέλου που σε φυσιολογικές συνθήκες δεν αποδίδει το ίδιο ικανοποιητικά. Για τον ίδιο λόγο θα έχουν εκτιμηθεί εσφαλμένα και οι παράμετροι του μοντέλου, μειώνοντας έτσι τη μέση απόδοσή του. Προκειμένου να περιοριστεί η επίδραση της αβεβαιότητας που σκιαγραφήθηκε νωρίτερα έχουν προταθεί αρκετές τεχνικές όπως ο συνδυασμός προβλέψεων (*forecast combination*) (Hibon M. & Evgeniou T., 2005) και η συνάθροιση δεδομένων σε διαφορετικές συχνότητες (*temporal aggregation*) (Silvestrini, A., Veredas, D. 2008).

Στην πρώτη περίπτωση πολλές διαφορετικές μέθοδοι πρόβλεψης χρησιμοποιούνται ξεχωριστά για την προέκταση της χρονοσειράς και στη συνέχεια συνδυάζονται κατάλληλα προκειμένου να οδηγήσουν σε μία πιο αντιπροσωπευτική πρόβλεψη. Η παραπάνω προσέγγιση αν και φαίνεται αποδοτική εκ πρώτης όψεως γεννά αρκετά ερωτήματα όπως το πόσες και ποιες μέθοδοι θα εφαρμοστούν και με ποια βάρη θα συνδυαστούν. Αν π.χ. η χρονοσειρά που καλούμαστε να προβλέψουμε είναι εποχιακή και εμφανίζει ταυτόχρονα σημαντική τάση αλλά κάποια από τα διαθέσιμα μοντέλα πρόβλεψης είναι μη εποχιακά ή

δεν αποδίδουν ικανοποιητικά το μοτίβο ανάπτυξης της χρονοσειράς, η συμμετοχή τους θα πρέπει να είναι αν όχι μηδενική τουλάχιστον αμελητέα ή εναλλακτικά να αντικατασταθούν από άλλα καταλληλότερα μοντέλα πρόβλεψης με μη μηδενικά βάρη. Η επιλογή των μεθόδων και ο υπολογισμός των βαρών σε κάθε περίπτωση εισάγουν από μόνα τους μία πρόσθετη αβεβαιότητα στη διαδικασία της πρόβλεψης που, όπως έχει αποδειχτεί, συνήθως έχει αρνητική επίπτωση στην ακρίβεια πρόβλεψης ή αν μη τι άλλο δεν οδηγεί σε συστηματική βελτίωση των προβλέψεων σε σύγκριση με άλλες απλούστερες προσεγγίσεις. Το εν λόγω πρόβλημα είναι ευρέως γνωστό ως 'forecast combination ruzzle' και μεγάλο πλήθος μελετών έχει πραγματοποιηθεί στην κατεύθυνση της επίλυσής του χωρίς όμως να υπάρχει μέχρι σήμερα κάποια βέλτιστη προσέγγιση. Έτσι, περιοριζόμαστε συνήθως εμπειρικά στο συνδυασμό μεθόδων που παράγουν σημαντικά διαφορετικές προβλέψεις με ίσα βάρη.

Στη δεύτερη προσέγγιση τα δεδομένα της αρχικής συχνότητας (π.χ. μηνιαία) συναθροίζονται σε μία χαμηλότερης τιμής (π.χ. ετήσια) με σκοπό η προέκταση να πραγματοποιηθεί επί μίας χρονοσειράς μικρότερης διακύμανσης που αναμένεται να οδηγήσει σε πιο στιβαρές και αξιόπιστες προβλέψεις. Αν και η προσέγγιση φαίνεται ιδιαίτερα υποσχόμενη, εμπεριέχει και αυτή σημαντική αβεβαιότητα όπως π.χ. τον εντοπισμό του βέλτιστου επιπέδου συνάθροισης. Επίσης, αν κανείς προβλέψει μία εποχιακή χρονοσειρά σε συχνότητα όπου η εποχιακότητα έχει εξαλειφθεί πλήρως, πόσο ακριβής αναμένεται να είναι η πρόβλεψη; Προκειμένου να αντιμετωπιστούν τα παραπάνω ζητήματα συχνά πραγματοποιείται συνδυασμός των δύο τεχνικών που αναφέρθηκαν (συνδυασμός προβλέψεων από όλες τις δυνατές συχνότητες) με αντίστοιχα πλεονεκτήματα, γνωστός και ως *multiple temporal aggregation*. Μειονέκτημα βέβαια της προσέγγισης αυτής είναι το αυξημένο υπολογιστικό κόστος καθώς απαιτείται η εκτίμηση και η εφαρμογή πολλαπλών μοντέλων πρόβλεψης και μάλιστα χωρίς να εγγυάται κανείς τη βελτίωση της προβλεπτικής ακρίβειας της μεθόδου. Τέλος, μεγάλο ζήτημα σε αυτήν την προσέγγιση αποτελεί ο βέλτιστος συνδυασμός εποχιακών και μη εποχιακών προβλέψεων καθώς και η επιλογή του βέλτιστου μοντέλου πρόβλεψης ανά συχνότητα. Πολλές μελέτες υποστηρίζουν ότι λόγω του συνδυασμού των μεθόδων τα εν λόγω προβλήματα δεν είναι μείζονος σημασίας, ωστόσο ένας τέτοιος ισχυρισμός είναι μάλλον καταχρηστικός και γενικευμένος.

Μία απλούστερη λύση για την επιλογή του καταλληλότερου μοντέλου πρόβλεψης είναι η χρήση διαγνωστικών ελέγχων και συγκεκριμένα η παρακολούθηση του *out-of-sample* σφάλματός των μοντέλων, του σφάλματος δηλαδή που αυτά εμφανίζουν σε δείγμα δεδομένων το οποίο δεν είχε χρησιμοποιηθεί κατά την παραμετροποίησή τους (*Tashman, 2000*). Οι εν λόγω έλεγχοι θεωρούνται αρκετά πιο αντιπροσωπευτικοί από άλλους που βασίζονται στο *in-sample* σφάλμα των μοντέλων, και συνεπώς συνίσταται συχνά η απόκρυψη δεδομένων και η μετέπειτα αξιοποίησή τους τους προκειμένου να επιλεγεί το πλέον ακριβές. Φυσικά και εδώ η τυχαιότητα μπορεί να παραπλανήσει σημαντικά και δεν είναι καθόλου σπάνιο το φαινόμενο προσωρινές και τυχαίες ιδιαιτερότητες στο δείγμα που χρησιμοποιείται για το διαγνωστικό έλεγχο να οδηγούν σε εσφαλμένα συμπεράσματα. Επαναλαμβάνοντας την εν λόγω διαδικασία αρκετές φορές (*rolling-origin evaluation*) το πρόβλημα της τυχαιότητας μπορεί να περιοριστεί, ωστόσο νέα προβλήματα κάνουν την

εμφάνισή τους. Αυτά σχετίζονται κυρίως με το υπολογιστικό κόστος το οποίο αυξάνεται δραματικά σε τέτοιες διαδικασίες, καθώς και με το διαθέσιμο πλήθος δεδομένων. Πρακτικά, οι χρονοσειρές που καλείται κανείς να προβλέψει μπορεί να διαθέτουν ελάχιστη ιστορία (π.χ. νέα προϊόντα). Έτσι, η απόκρυψη δεδομένων και μάλιστα τόσων ώστε η διαδικασία να επαναληφθεί αρκετές φορές, είναι μάλλον εξιδανικευμένη.

Με βάση τα παραπάνω καταλήγει κανείς εύκολα στο συμπέρασμα ότι σε πραγματικές συνθήκες η επιτυχής επιλογή του βέλτιστου μοντέλου πρόβλεψης ανά περίπτωση και μάλιστα σε περιορισμένο χρόνο είναι μάλλον αδύνατη. Έτσι, το ενδιαφέρον μετατοπίζεται στον εντοπισμό ενός μοντέλου πρόβλεψης η καθολική εφαρμογή του οποίου μπορεί μεν να μην οδηγήσει μεμονωμένα στο καλύτερο δυνατό αποτέλεσμα, παράγει όμως στιβαρές προβλέψεις ανεξαρτήτως εφαρμογής, τύπου και χαρακτηριστικών χρονοσειράς, οδηγώντας τελικά σε βελτιωμένη μέση απόδοση. Η ευκολία παραμετροποίησης και εφαρμογής του αποτελεί και πάλι βέβαια σημαντική προϋπόθεση.

Ανατρέχοντας κανείς στη βιβλιογραφία, το πρόβλημα που παρουσιάστηκε παραπάνω σχετικά με τον εντοπισμό της καταλληλότερης μεθόδου πρόβλεψης μπορεί να περιορίσει στα μοντέλα χρονοσειρών, όπως είναι π.χ. τα μοντέλα εκθετικής εξομάλυνσης, τα μοντέλα *ARIMA* και η μέθοδος θ . Τα εν λόγω μοντέλα βασίζονται αποκλειστικά στα ιστορικά δεδομένα του μεγέθους για το οποίο επιθυμείται πρόβλεψη, εκτιμώντας αρχικά τη σχέση που συνδέει τις ιστορικές παρατηρήσεις μεταξύ τους και αξιοποιώντας τη στη συνέχεια για την επέκτασή τους στο μέλλον. Μέθοδοι τεχνικής νοημοσύνης, όπως π.χ. τα νευρωνικά δίκτυα απορρίφθηκαν ως επιλογές αφενός λόγω της αυξημένης πολυπλοκότητάς τους, και αφετέρου του όγκου δεδομένων που απαιτούν για την αποτελεσματική εκπαίδευσή τους. Άλλωστε, η πείρα έχει δείξει πως οι στατιστικές μέθοδοι δεν υστερούν εν τέλει σε ακρίβεια, ειδικά όταν πρόκειται για μέσο- και μακροπρόθεσμες προβλέψεις.

Συνεχίζοντας την παραπάνω αναζήτηση, οι οικογένειες μοντέλων εκθετικής εξομάλυνσης και *ARIMA* φαίνεται να αποτελούν τις πλέον δημοφιλείς λύσεις, ενώ πληθώρα αλγορίθμων υπάρχει διαθέσιμη για την επιλογή, εκτίμηση και εφαρμογή τους. Τα παραπάνω μοντέλα είναι αποτελεσματικά ιδιαίτερα για μικρούς ορίζοντες πρόβλεψης και εμφανίζουν ικανοποιητικά επίπεδα ακρίβειας ανεξαρτήτως των χαρακτηριστικών της χρονοσειράς, ενώ δεν συνεπάγονται ιδιαίτερα υψηλού υπολογιστικού κόστους για ένα σύστημα. Ωστόσο, παρά τη διαρκή εξέλιξή τους, μία σημαντικά πιο απλοϊκή μέθοδος που βασίζεται αποκλειστικά στη μεταβολή των καμπυλοτήτων της χρονοσειράς μέσω γραμμικών μετασχηματισμών, η μέθοδος Θ (θ), συνεχίζει σήμερα, 20 χρόνια από την εμφάνισή της, να αποτελεί την ακριβέστερη επιλογή όταν αναφερόμαστε σε πρόβλεψη κατά μήκος μεγάλου όγκου χρονοσειρών. Ελάχιστες μέθοδοι έχουν καταφέρει να την ξεπεράσουν σε αντίστοιχες μελέτες και μάλιστα μετά από εφαρμογή πολύπλοκων αλγορίθμων προετοιμασίας χρονοσειρών και συνδυασμού τεχνικών προβλέψεων.

Η μέθοδος θ έγινε γνωστή κυρίως λόγω της απόδοσής της στον διαγωνισμό M3 (*Makridakis & Hibon, 2000*), τον μεγαλύτερο διαγωνισμό προβλέψεων που διοργανώθηκε ποτέ, όπου και αναδείχθηκε νικήτρια έναντι σαφώς πιο πολύπλοκων μεθόδων και ολοκληρωμένων συστημάτων πρόβλεψης. Αυτό που προκαλεί επίσης εντύπωση είναι ότι,

αν και από τότε έχουν γίνει κάποιες προσπάθειες για τη βελτίωση της απόδοσης της μεθόδου και την εξακρίβωση της απόδοσής της σε ποικίλες εφαρμογές και σετ δεδομένων με ενθαρρυντικά αποτελέσματα, λίγες είναι εκείνες οι μελέτες που συνεισφέρουν πραγματικά στην εξέλιξή της.

Πρώτος στόχος της παρούσας διατριβής είναι λοιπόν να εντοπιστούν οι παράγοντες που καθιστούν τη μέθοδο *θ* τόσο πετυχημένη και στη συνέχεια να αξιοποιηθούν τα ευρήματα ούτως ώστε να προκύψει μία νέα μέθοδος με βελτιωμένα χαρακτηριστικά και δυνατότητες. Σημεία στα οποία υστερεί η κλασική μέθοδος θα λάβουν επίσης βελτιώσεις. Η μέθοδος θα αξιοποιηθεί στη συνέχεια στη γενικευμένη μεθοδολογία πρόβλεψης που αναπτύσσεται παράγοντας αξιόπιστες και στιβαρές προβλέψεις κατά μήκος μεγάλου πλήθους χρονοσειρών.

Βέβαια, όσο και αν μία μέθοδος θεωρείται γενικευμένη και στιβαρή, σε καμία περίπτωση δεν μπορεί κανείς να περιμένει πως θα πετυχαίνει σταθερά από μόνη της ικανοποιητική ακρίβεια, ιδιαίτερα σε χρονοσειρές με μικρό πλήθος δεδομένων, υψηλά επίπεδα θορύβου και ειδικά γεγονότα (*special events*). Τα παραπάνω χαρακτηριστικά συνδέονται άμεσα με την αύξηση του σφάλματος πρόβλεψης και προκειμένου αυτό να διατηρηθεί σε σχετικά χαμηλά επίπεδα θα πρέπει να εφαρμοστούν ειδικές τεχνικές που περιορίσουν την επίδρασή τους. Η εν λόγω διαδικασία αναφέρεται συνήθως με τον όρο προετοιμασία ή προ-επεξεργασία χρονοσειράς και περιλαμβάνει την εφαρμογή διαγνωστικών ελέγχων και μεθόδων προκειμένου να εξακριβωθεί αφενός αν η χρονοσειρά είναι επαρκώς προβλέψιμη και αν όχι να γίνει τροποποιώντας τα δεδομένα κατάλληλα.

Το χαρακτηριστικότερο παράδειγμα προετοιμασίας χρονοσειρών είναι η χρήση μετασχηματισμών οι οποίοι περιορίζουν σημαντικά τη διακύμανση των δεδομένων και οδηγούν σε κανονικότητα (*Proietti & Lütkepohl, 2013*). Περιοριστικός είναι βέβαια ο παράγοντας πως οι μετασχηματισμοί στην πλειοψηφία τους δεν μπορούν να εφαρμοστούν σε μη θετικά δεδομένα, ενώ δεν εγγυόνται ότι θα εξαλείψουν την επίδραση ειδικών γεγονότων και ακραίων τιμών (*outliers*). Αυτό συμβαίνει καθώς οι μετασχηματισμοί αντιμετωπίζουν συνήθως με μη γραμμικό τρόπο τις παρατηρήσεις και συνεπώς, αν μία χρονοσειρά έχει σημαντικά μικρότερες και μεγαλύτερες τιμές σε σύγκριση με το συνηθισμένο, κάποιες από αυτές θα μείνουν ανεπηρέαστες ή ακόμα χειρότερα θα υπερτονιστούν. Οι μετασχηματισμοί αποδίδουν συνεπώς καλά μόνο όταν η διακύμανση είναι ομοιόμορφη κατά μήκος της χρονοσειράς, όταν τα *outliers* είναι μίας μόνο κατεύθυνσης και, συνήθως, όταν αυτή δεν χαρακτηρίζεται από τάση (στάσιμη χρονοσειρά).

Για να ξεπεράσει κανείς τους παραπάνω περιορισμούς, μπορεί να στραφεί σε μεθόδους εντοπισμού και περιορισμού *outliers* (*Ledolter, 1989*), τα οποία εκτός από ακραίες τιμές μπορεί να αναφέρονται σε αλλαγές επιπέδου (*level change*) ή στιγμιαίες μεταβολές (*temporal change*). Αυτή η προσέγγιση έχει ωστόσο το μειονέκτημα ότι δεν βελτιώνει την απόδοση του μοντέλου σε όλες τις χρονοσειρές αλλά μόνο σε αυτές που εμφανίζουν ανωμαλίες, ενώ παράλληλα αυξάνει σημαντικά το υπολογιστικό κόστος της διαδικασίας της πρόβλεψης καθώς απαιτεί αρκετούς και απαιτητικούς διαγνωστικούς ελέγχους.

Μία λύση σε αυτά τα προβλήματα μπορεί να είναι η εξομάλυνση της χρονοσειράς με χρήση κατάλληλων τεχνικών, όπως είναι π.χ. οι κινητοί μέσοι όροι και τα φίλτρα *kernel*. Το πλεονέκτημά των εν λόγω τεχνικών είναι ότι περιορίζουν αποτελεσματικά τη διακύμανση των χρονοσειρών, οδηγώντας σε ομαλές και σαφώς πιο προβλέψιμες χρονοσειρές. Ο βαθμός εξομάλυνσης και ο τρόπος εφαρμογής της είναι βέβαια παράμετροι που χρίζουν μελέτης, ωστόσο συνεπάγονται σίγουρα πολύ μικρότερη επιβάρυνση σε υπολογιστικό κόστος σε σχέση με άλλες τεχνικές που αναφέρθηκαν και μπορούν να οδηγήσουν σε εξίσου καλά αποτελέσματα.

Έτσι, δεύτερος στόχος της παρούσας διατριβής είναι η επιλογή μίας τεχνικής εξομάλυνσης που σε συνδυασμό με το μοντέλο πρόβλεψης που επιλέχθηκε νωρίτερα θα βοηθά στη γενικότερη βελτίωση της προβλεπτικής ακρίβειας της μεθοδολογίας παράγοντας στιβαρές προβλέψεις ακόμα και σε περιπτώσεις χρονοσειρών που λόγω ιδιαιτεροτήτων είναι δύσκολο να προβλεφθούν με συμβατικά μοντέλα. Για αυτό το σκοπό θα μελετηθούν οι μέθοδοι εξομάλυνσης που συναντώνται στη βιβλιογραφία, θα εντοπιστούν τα πλεονεκτήματα και μειονεκτήματά τους, και μαθαίνοντας από αυτά θα αναπτυχθεί μία νέα μέθοδος με βελτιωμένα χαρακτηριστικά. Τέτοια μπορεί να είναι η εξομάλυνση χωρίς παραμέτρους, η επίτευξη μεγαλύτερου βαθμού εξομάλυνσης, η αποτελεσματικότερη αντιμετώπιση ακραίων τιμών κ.ο.κ.

Σε αυτό το σημείο αναφέρεται επίσης πως η μεθοδολογία θα συνοδεύεται και από άλλες τεχνικές απαραίτητες για την αποτελεσματική εφαρμογή της μέθοδου πρόβλεψης, όπως είναι π.χ. η αποεποχικοποίηση. Αυτό συμβαίνει καθώς η εξομάλυνση βοηθάει μόνο στην απαλοιφή τυχαιών διακυμάνσεων. Όταν αυτές είναι συστηματικές και περιοδικές μπορούν να αντιμετωπιστούν ξεχωριστά και να αξιοποιηθούν για τη βελτίωση της ακρίβειας πρόβλεψης. Η εποχιακότητα στα δεδομένα είναι ένα τέτοιο περιοδικό φαινόμενο το οποίο μπορεί να μοντελοποιηθεί σχετικά εύκολα με ειδικές τεχνικές και να χρησιμοποιηθεί για να εξαλείψει την επίδραση της εποχής στα δεδομένα πριν της εφαρμογή οποιαδήποτε άλλης επεξεργασίας. Η επιλογή της καλύτερης δυνατής τεχνικής για να γίνει αυτό θα αποτελέσει άλλο ένα αντικείμενο μελέτης στην εργασία.

Έχοντας καθορίσει πλέον τη μεθοδολογία πρόβλεψης, απαραίτητη προϋπόθεση για την εξακρίβωση της αποτελεσματικότητάς της είναι η δοκιμή της σε μεγάλο πλήθος δεδομένων, σε εφαρμογές δηλαδή σαν αυτές για τις οποίες πραγματικά προορίζεται. Για να γίνει αυτό απαιτείται αρχικά η συλλογή δεδομένων τα οποία διαθέτουν διαφορετικά χαρακτηριστικά (π.χ. εποχιακότητα, τάση, κύκλο και τυχαιότητα) και ιδιαιτερότητες (π.χ. ακραίες τιμές, μικρό ή μεγάλο μήκος), ενώ αναφέρονται παράλληλά σε ένα ευρύ φάσμα εφαρμογών (π.χ. δημογραφικά, οικονομικά ή ενεργειακά δεδομένα). Τέτοια δεδομένα χρησιμοποιούνται συχνά στη επιστημονική κοινότητα των προβλέψεων (κυρίως από αντίστοιχους διαγωνισμούς που έχουν γίνει ώστε να υπάρχει ένα προφανές μέτρο σύγκρισης) ωστόσο το πλήθος τους είναι αρκετά περιορισμένο (μερικές χιλιάδες χρονοσειρές). Ένα άλλο πρόβλημα είναι ότι ακόμα και αυτές οι χρονοσειρές δεν κατανέμονται ομοιόμορφα ως προς τα χαρακτηριστικά, τις ιδιαιτερότητες και τον τύπο τους, οδηγώντας έτσι πιθανώς σε εσφαλμένα συμπεράσματα. Μέθοδοι δηλαδή που ανταποκρίνονται καλύτερα στα χαρακτηριστικά χρονοσειρών που εμφανίζονται

συχνότερα στο συγκεκριμένο σετ δεδομένων έχουν καλύτερη πιθανότητα να αναδειχτούν ως οι πλέον αποτελεσματικές.

Το παραπάνω πρόβλημα είναι δύσκολο να επιλυθεί καθώς αφενός η συλλογή μεγάλου δείγματος χρονοσειρών είναι από μόνη της μία εξαιρετικά χρονοβόρα εργασία, και αφετέρου η δημιουργία ενός ομοιόμορφου δείγματος ανά είδος χρονοσειράς είναι μία ιδιαίτερος απαιτητική διαδικασία που απαιτεί τον καθορισμό των πιθανών τύπων χρονοσειρών, των επιμέρους χαρακτηριστικών και ιδιοτεροτήτων, και επιπλέον τη βαθμονόμηση και μέτρησή τους. Πρακτικά, ακόμα και αν μιλήσουμε για πολύ μεγάλο όγκο δεδομένων, κανείς δεν μπορεί να εγγηθεί λοιπόν ότι το δείγμα θα είναι πλήρως αντιπροσωπευτικό, αν και σίγουρα θα είναι ενδεικτικό της πραγματικότητας.

Σε αυτήν την κατεύθυνση, αν και η μεθοδολογία θα αξιολογηθεί αρχικά σε ικανοποιητικό πλήθος πραγματικών δεδομένων, αντίστοιχη ανάλυση θα πραγματοποιηθεί και για μεγάλο πλήθος τεχνητών δεδομένων. Τα τεχνητά δεδομένα θα προκύψουν από μία γεννήτρια χρονοσειρών που θα κατασκευαστεί για αυτόν τον σκοπό και η οποία θα δίνει τη δυνατότητα να παραχθεί ένα προκαθορισμένο πλήθος χρονοσειρών με συγκεκριμένα χαρακτηριστικά δεδομένης έντασης. Με αυτόν τον τρόπο θα δημιουργηθούν ομοιόμορφα δείγματα δεδομένων ανά είδος χρονοσειράς και η αξιολόγηση της μεθοδολογίας θα είναι σαφώς πιο αναλυτική και αντιπροσωπευτική.

Στις ακόλουθες παραγράφους συνοψίζεται το περιεχόμενο καθενός από τα κεφάλαια της παρούσας διατριβής: Στο δεύτερο κεφάλαιο περιγράφονται οι τεχνικές που συναντώνται στη βιβλιογραφία για την ανάλυση και την επεξεργασία των χρονοσειρών, ενώ στο τρίτο κεφάλαιο παρουσιάζονται οι βασικότερες μέθοδοι παραγωγής στατιστικών προβλέψεων με χρήση μοντέλων χρονοσειρών. Εκεί παρουσιάζεται και η γενικευμένη μέθοδος του μοντέλου πρόβλεψης θ . Το κεφάλαιο που ακολουθεί ασχολείται με τις μεθόδους εξομάλυνσης που χρησιμοποιούνται συνήθως για την ομαλοποίηση των δεδομένων και προτείνεται μία νέα μη παραμετρική μέθοδος εξομάλυνσης. Στο πέμπτο κεφάλαιο αναπτύσσεται η γεννήτρια χρονοσειρών για την παραγωγή τεχνητών δεδομένων, ενώ στο έκτο κεφάλαιο αναλύεται η επίδοση της μεθοδολογίας, τόσο σε πραγματικά όσο και σε τεχνητά δεδομένα. Τέλος, στο τελευταίο κεφάλαιο της διατριβής συνοψίζονται τα συμπεράσματα της εργασίας και τίθενται συγκεκριμένοι στόχοι ως μελλοντικές προεκτάσεις της μεθοδολογίας προκειμένου αυτή αφενός να βελτιωθεί και αφετέρου να υιοθετηθεί από επιχειρήσεις και οργανισμούς.

B. Χρονοσειρές και Ανάλυση

Η ανάλυση των δεδομένων μιας χρονοσειράς είναι μία διαδικασία που συντελείται καθ' όλη τη διαδικασία της πρόβλεψης και για αυτό το λόγο μας απασχολεί ιδιαίτερα. Η μελέτη των χαρακτηριστικών των χρονοσειρών, η προετοιμασία τους βάσει αυτών και η τελική αξιολόγηση των αποτελεσμάτων είναι κάποια χαρακτηριστικά παραδείγματα ανάλυσης που εφαρμόζονται πριν, κατά τη διάρκεια και μετά το πέρας της πρόβλεψης.

Προκειμένου να αντιληφθεί κανείς τη σημασία της ανάλυσης αρκεί να αναλογιστεί πόσες διαφορετικές χρονοσειρές ενδέχεται να εμφανιστούν και πώς μπορεί να τις αντιμετωπίσει αποτελεσματικά με έναν αυτοματοποιημένο τρόπο. Χρονοσειρές με τάση εμφανίζουν

διαφορετική συμπεριφορά από τις στάσιμες, ενώ χρονοσειρές με σημαντική διακύμανση απαιτούν ειδική αντιμετώπιση προκειμένου να προβλεφθούν αποδοτικά. Το ίδιο ισχύει και για εποχιακές χρονοσειρές τις οποίες στην περίπτωση χρήσης μη εποχιακών μοντέλων πρόβλεψης πρέπει να τις διαχειριζόμαστε ξεχωριστά. Κάθε μοντέλο πρόβλεψης έχει συγκεκριμένες ιδιότητες και δυνατότητες και ανταποκρίνεται περισσότερο αποδοτικά σε συγκεκριμένους τύπους χρονοσειρών. Έτσι, όταν καλούμαστε να επιλέξουμε ένα μοντέλο πρόβλεψης θα πρέπει πρώτα απ' όλα να έχουμε ενημερωθεί σχετικά με τον τύπο της χρονοσειράς, τα χαρακτηριστικά της δηλαδή και την ένταση αυτών. Ακόμα όμως και όταν αναφερόμαστε στην πρόβλεψη μέσω ενός καθολικού μοντέλου πρόβλεψης (όπως προτείνεται την παρούσα διατριβή) η εφαρμογή σχετικών διαγνωστικών ελέγχων και ο έλεγχος της αποτελεσματικότητας των μεθόδων είναι καθοριστικής σημασίας.

Πρώτο βήμα για την ανάλυση μια χρονοσειράς είναι η αναγνώριση των ποιοτικών της χαρακτηριστικών: Τάση, Εποχιακότητα, Τυχειότητα και Κύκλος (*Makridakis et al., 1983*). Τα παραπάνω χαρακτηριστικά, αν και αναφέρονται σε ποιοτικές έννοιες που έχει επινοήσει ο άνθρωπος για να αντιλαμβάνεται και να περιγράφει την πραγματικότητα, έχουν και πρακτικά ιδιαίτερη σημασία για την αποτελεσματική προέκταση των χρονοσειρών. Μέσω της διαδικασίας της αποσύνθεσης μπορεί μάλιστα κανείς να αναλύσει την χρονοσειρά στις εν λόγω συνιστώσες και να τις αξιοποιήσει κατά την πρόβλεψη. Χαρακτηριστικό είναι το παράδειγμα της αποεποχικοποίησης κατά την οποία η συνιστώσα της εποχιακότητας εντοπίζεται και απομακρύνεται από τα δεδομένα προκειμένου να απαλλαχθούν από συστηματικές και περιοδικές διακυμάνσεις που οφείλονται σε εποχιακούς παράγοντες.

Η διαδικασία με την οποία εφαρμόζεται η αποσύνθεση και συνεπώς η αποεποχικοποίηση δεν είναι μοναδική και πλήθος διαφορετικών μεθόδων έχουν προταθεί για την υλοποίησή της. Άλλες μέθοδοι υποθέτουν ότι οι συνιστώσες συνδέονται μεταξύ τους πολλαπλασιαστικά και άλλες προσθετικά. Έτσι, στην πρώτη περίπτωση π.χ. υποθέτουμε ότι η εποχιακότητα μεταβάλλεται καθώς αλλάζει το επίπεδο της χρονοσειράς, ενώ στη δεύτερη ότι η επίδραση της εποχής είναι ανεξάρτητη. Επίσης, άλλες τεχνικές θεωρούν ότι η τάση είναι συγκεκριμένης μορφής για την μοντελοποίησύν (π.χ. γραμμική) ενώ άλλες ότι αποτελεί μία μη προκαθορισμένη καμπύλη. Έτσι, στην πρώτη περίπτωση η τάση δίνεται από συγκεκριμένη μαθηματική σχέση και μπορεί να προεκταθεί με καθορισμένο τρόπο στο μέλλον, ενώ στη δεύτερη όχι. Μάλιστα, δεδομένου ότι στη δεύτερη περίπτωση η τάση δίνεται σε μορφή ακαθόριστης καμπύλης συχνό είναι το φαινόμενο αυτή να συμπεριλαμβάνει τη συνιστώσα του κύκλου, της εποχιακής δηλαδή διακύμανσης για την οποία όμως δεν γνωρίζουμε ούτε την περίοδο ούτε την έντασή της. Στην ίδια λογική, ότι ο κύκλος είναι δηλαδή επί της ουσίας μη προβλέψιμος από μόνος του, κάποιες τεχνικές επιλέγουν να συμπεριλάβουν την επίδρασή του στη συνιστώσα της τυχειότητας. Οι επιλογές όπως φαίνεται είναι αρκετές και στο σημείο της ανάλυσης καλούμαστε να τις αξιολογήσουμε και να επιλέξουμε τις πλέον αποδοτικές.

Η πιο διαδεδομένη τεχνική αποσύνθεσης είναι η κλασική μέθοδος αποσύνθεσης (*Kendall & Stuart, 1983*) κατά την οποία οι συνιστώσες τάσης, εποχιακότητας, τυχειότητας και κύκλου συνδέονται μεταξύ τους με πολλαπλασιαστικό τρόπο και κινητοί μέσοι όροι

εφαρμόζονται για τον διαχωρισμό τους. Η μέθοδος έχει το πλεονέκτημα ότι με σχετικά απλοποιημένο τρόπο καταφέρνει να εντοπίσει τα επιμέρους χαρακτηριστικά με υψηλή ακρίβεια, κάτι το οποίο έχει διαπιστωθεί σε πληθώρα δημοσιευμένων εργασιών. Βέβαια, δεδομένης της τυχαιότητας στα δεδομένα, κανείς δεν μπορεί να ισχυριστεί ότι η συγκεκριμένη διαδικασία διαχωρίζει τα αρχικά δεδομένα με απόλυτη ακρίβεια καθώς είναι πιθανό π.χ. μέρος της εποχιακής επίδρασης να μοντελοποιηθεί ως θόρυβος και το αντίστροφο. Σημαντικό ρόλο σε αυτό παίζει προφανώς και η αρχική υπόθεση της μεθόδου ότι η τάση της χρονοσειράς είναι γραμμική, κάτι το οποίο φυσικά δεν ισχύει πάντοτε. Έτσι, μέρος της τάσης μπορεί να μοντελοποιηθεί αντίστοιχα ως κύκλος ή ως τυχαιότητα. Προκειμένου να αντιμετωπιστεί το παραπάνω πρόβλημα, και δεδομένου μάλιστα ότι η αποσύνθεση χρησιμοποιείται κυρίως για την αποεποχικοποίηση των δεδομένων, προτείνεται η χρήση τεχνικών συρρίκνωσης συντελεστών εποχιακότητας (*Miller & Williams, 2003*).

Η πλέον διαδεδομένη προσέγγιση πρόβλεψης είναι η αποεποχικοποίηση των δεδομένων αφαιρώντας την εποχιακή συνιστώσα (δείκτες εποχιακότητας), η πρόβλεψη με χρήση κατάλληλων μεθόδων και η εποχικοποίηση των προβλέψεων ούτως ώστε να συμπεριληφθεί σε αυτές η εποχική συμπεριφορά. Η αποεποχικοποιημένη χρονοσειρά αποτελείται συνεπώς από τις συνιστώσες τάσης, τυχαιότητας και κύκλου και προβλέπεται ξεχωριστά. Αν η διαδικασία της αποσύνθεσης έχει στεφθεί με απόλυτη επιτυχία τότε η εποχιακότητα έχει εξαιρεθεί πλήρως από τη χρονοσειρά και μπορούμε να την προβλέψουμε με ασφάλεια. Δεδομένης ωστόσο της τυχαιότητας που κάνει την εποχιακότητα να μην εμφανίζεται με πανομοιότυπο τρόπο σε κάθε περίοδο (κάτι το οποίο υποθέτουμε) κάτι τέτοιο είναι πρακτικά αδύνατο. Για να περιοριστούν τα σφάλματα στην εκτίμηση των εποχιακών συντελεστών μπορεί κανείς να μελετήσει τη συστηματικότητα των λόγων εποχιακότητας, του κατά πόσο ταυτίζεται δηλαδή η εκάστοτε εποχιακή επίδραση με την μέση που υπολογίστηκε νωρίτερα. Βάσει των αποτελεσμάτων διορθώνονται οι αρχικοί δείκτες οδηγώντας σε συρρικνωμένους συντελεστές. Έτσι, οι εποχικοποιημένες προβλέψεις είναι πιο πιθανό να βρεθούν κοντά στην πραγματικότητα και το σφάλμα πρόβλεψης μειώνεται αντίστοιχα. Στην παρούσα εργασία παρουσιάζονται τέτοιες μέθοδοι για τη συρρίκνωση των συντελεστών εποχιακότητας και επιλέγεται βάσει βιβλιογραφίας η πλέον αποδοτική για να συμπεριληφθεί στην προτεινόμενη μεθοδολογία.

Φυσικά, για να λάβει χώρα η διαδικασία της αποεποχικοποίησης, βασική προϋπόθεση είναι η χρονοσειρά να είναι πράγματι εποχιακή. Ο παραπάνω διαγνωστικός έλεγχος είναι καθοριστικός καθώς σε αντίθετη περίπτωση υποθέτουμε την ύπαρξη εποχιακής διακύμανσης που στην πράξη δεν υπάρχει και την εφαρμόζουμε καταχρηστικά επί των παραγόμενων προβλέψεων μειώνοντας έτσι την τελική ακρίβεια των προβλέψεων. Η αποτελεσματική αναγνώριση εποχιακής συμπεριφοράς μπορεί να γίνει μέσω κατάλληλων στατιστικών ελέγχων γνωστών ως τεστ εποχιακότητας. Στην προτεινόμενη μεθοδολογία συμπεριλαμβάνονται τέτοιοι έλεγχοι και οι χρονοσειρές μπορούν έτσι να αποεποχικοποιούνται επιλεκτικά.

Στην ίδια φιλοσοφία μπορεί κανείς να εφαρμόσει διαγνωστικούς ελέγχους για να εξακριβώσει αν μία χρονοσειρά είναι στάσιμη, δηλαδή έχει σταθερή διακύμανση και μέση

τιμή κατά μήκος των δεδομένων της, ή αν ακολουθεί την κανονική κατανομή. Τα παραπάνω μπορούν να αποτελέσουν χρήσιμες πληροφορίες όταν κανείς καλείται να επεξεργαστεί τα δεδομένα προκειμένου να τα ομαλοποιήσει ή να περιορίσει τη διακύμανσή τους με σχετικές τεχνικές. Τέτοιες είναι η τεχνική της διαφορίσης και του μετασχηματισμού, οι οποίες επίσης εμφανίζονται με ποικίλους τρόπους στη βιβλιογραφία. Στην παρούσα διατριβή αξιοποιούνται πιο συγκεκριμένα οι μετασχηματισμοί *Box-Cox* οι οποίοι με μη γραμμικό τρόπο μετατρέπουν με παραμετρικό τρόπο πιθανώς ανομοιόμορφες χρονοσειρές σε νέες που ακολουθούν την κανονική κατανομή. Ο ακριβής τρόπος με τον οποίο ελέγχεται η κανονικότητα των δεδομένων και γίνεται η παραμετροποίηση της διαδικασίας παρουσιάζεται αναλυτικά σε αντίστοιχη παράγραφο.

Τέλος, η ανάλυση των χρονοσειρών περιλαμβάνει μεταξύ άλλων την καταγραφή και αξιολόγηση των αποτελεσμάτων, της επίδοσης δηλαδή της μεθόδου που χρησιμοποιήθηκε για την πρόβλεψή τους. Η αξιολόγηση μπορεί να γίνει έχοντας ως γνώμονα την ακρίβεια της πρόβλεψης ή την προκατάληψή της. Στην πρώτη περίπτωση, η οποία είναι και η πλέον συνήθης, ελέγχεται η μέση διαφορά της πρόβλεψης από την πραγματικότητα, το πόσο κοντά βρίσκεται δηλαδή η πρόβλεψη στο επιθυμητό. Αντίθετα, στη δεύτερη περίπτωση ελέγχεται κατά πόσο η μέθοδος έχει την τάση να προβλέπει ανομοιόμορφα ως προς την πραγματικότητα, αν δηλαδή οι προβλέψεις είναι συστηματικά μεγαλύτερες ή μικρότερες από το επιθυμητό. Για να θεωρηθεί μία πρόβλεψη ικανοποιητική θα πρέπει τα επιμέρους σφάλματα να εμφανίζουν μικρή μη συστηματική διακύμανση και να κατανέμονται ομοιόμορφα γύρω από την μέση τους τιμή. Σε αντίθετη περίπτωση είτε η μέθοδος είναι ανακριβείς, είτε είναι προκατειλημμένη, είτε εμφανίζει π.χ. λόγω της αποτυχημένης αποεποχικοποίησής της συστηματικά λάθη.

Στην εργασία αυτή παρουσιάζονται τα περισσότερα μέτρα αξιολόγησης προβλέψεων που έχουν προταθεί στη βιβλιογραφία, γνωστά και ως δείκτες σφάλματος, και προτείνεται ως βασικό κριτήριο το *sMAPE*, ένας ποσοστιαίος δείκτης σφάλματος ο οποίος έχει διαπιστωθεί ότι μπορεί να χρησιμοποιηθεί με σχετική ασφάλεια για την αξιολόγηση πλήθους χρονοσειρών διαφορετικής κλίμακας ως προς την ακρίβεια που πετυχαίνουν σε αυτές οι μέθοδοι πρόβλεψης.

Γ. Μοντέλα πρόβλεψης

Ο πυρήνας της διαδικασίας της πρόβλεψης είναι το μοντέλο πρόβλεψης καθώς, άσχετα με το ποιες τεχνικές έχουν εφαρμοστεί για την προετοιμασία της χρονοσειράς, το μοντέλο είναι εκείνο που θα χρησιμοποιηθεί εν τέλει για την προέκτασή της και συνεπώς θα καθορίσει σε μεγάλο βαθμό την ακρίβεια που θα επιτευχθεί. Αν και τα μοντέλα πρόβλεψης μπορεί να αναφέρονται σε διάφορες μεθόδους (π.χ. τεχνικής νοημοσύνης) στην παρούσα εργασία για λόγους που αναφέρθηκαν νωρίτερα περιοριζόμαστε στις στατιστικές μεθόδους και συγκεκριμένα στα μοντέλα πρόβλεψης χρονοσειρών.

Τα μοντέλα πρόβλεψης χρονοσειρών είναι διαδικασίες οι οποίες προσπαθούν να εντοπίσουν και να περιγράψουν τον τρόπο με τον οποίο συνδέονται μεταξύ τους χρονικά τα δεδομένα μίας χρονοσειράς. Στη συνέχεια χρησιμοποιούν την παραπάνω σχέση για να την προεκτείνουν στο μέλλον για τον ζητούμενο ορίζοντα. Φυσικά κάθε μοντέλο κάνει

διαφορετικές υποθέσεις σχετικά με τον τύπο της χρονοσειράς που δέχεται ως είσοδο και συνεπώς παράγει διαφορετικά αποτελέσματα. Για παράδειγμα, η πλειοψηφία των μοντέλων πρόβλεψης είναι μη εποχιακά, κάτι το οποίο σημαίνει πως αν η χρονοσειρά είναι εποχιακή τα συγκεκριμένα μοντέλα θα αντιμετωπίσουν την εποχιακότητα ως θόρυβο. Άλλα μοντέλα υποθέτουν πως η χρονοσειρά είναι στάσιμη. Έτσι, πέρα από την εποχιακότητα δεν μπορούν να εντοπίσουν τάση στα δεδομένα και η πρόβλεψη αντιπροσωπεύει το τρέχον επίπεδο της χρονοσειράς.

Το απλούστερο μοντέλο πρόβλεψης που μπορεί να χρησιμοποιήσει κανείς είναι αυτό της απλοϊκής μεθόδου (*Naïve*). Η μέθοδος υποθέτει πως ότι συνέβη στην τελευταία γνωστή περίοδο θα ξανασυμβεί στο μέλλον. Συνεπώς αφενός η πρόβλεψη δεν μπορεί να εμπεριέχει τάση ή εποχιακότητα και αφετέρου επηρεάζεται σημαντικά από τυχαίες διακυμάνσεις στη χρονοσειρά. Έτσι, συνηθίζεται να χρησιμοποιείται ως μέτρο σύγκρισης ακρίβειας μεθόδων και όχι σαν μέθοδος πρόβλεψης. Η χρήση δεικτών εποχιακότητας και η προσθήκη τάσης με πρόσθετες τεχνικές μπορούν βέβαια να βελτιώσουν σημαντικά την ακρίβεια της μεθόδου και να την καταστήσουν μία πρώτη λύση.

Άλλο απλοϊκό μοντέλο πρόβλεψης είναι αυτό της αποσύνθεσης (*decomposition*). Το μοντέλο θεωρεί ότι η χρονοσειρά μπορεί να προβλεφθεί προεκτείνοντας τα επιμέρους ποιοτικά χαρακτηριστικά που υπολογίστηκαν μέσω της διαδικασίας της αποσύνθεσης και συνδυάζοντάς τα κατάλληλα. Πρακτικά η προσέγγιση δεν οδηγεί ωστόσο σε ακριβείς προβλέψεις καθώς αφενός ο κύκλος και η τυχαιότητα δεν μπορούν να προβλεφθούν, ενώ η συνιστώσα της τάσης προεκτείνεται μέσω μίας προκαθορισμένης συνάρτησης που δεν είναι η πλέον αντιπροσωπευτική. Έτσι, η τελική πρόβλεψη είναι επί της ουσίας μία γραμμική και εποχιακή προέκταση της χρονοσειράς, χωρίς όμως να έχει εξεταστεί προσεκτικά για αυτήν το τρέχον επίπεδο το οποίο και συνεισφέρει καθοριστικά στην ακρίβεια που θα επιτευχθεί.

Οι μέθοδοι εκθετικής εξομάλυνσης (*exponential smoothing*) είναι μία οικογένεια μοντέλων πρόβλεψης χρονοσειρών τα οποία απαντούν σε αρκετά από τα προβλήματα που τέθηκαν νωρίτερα. Βασίζονται στη λογική ότι τόσο η τάση, όσο και η εποχιακότητα και το επίπεδο της χρονοσειράς ενδέχεται να αλλάζουν με την πάροδο του χρόνου και συνεπώς πρέπει κανείς να τα επαναπροσδιορίζει κατάλληλα. Αυτό γίνεται εξομαλύνοντας τη χρονοσειρά και διορθώνοντας τις εν λόγω συνιστώσες βάσει του τρέχοντος σφάλματος του μοντέλου. Συνολικά ορίζονται 30 διαφορετικά μοντέλα εκθετικής εξομάλυνσης (*Hyndman et al., 2008*), κάθε ένα από τα οποία θεωρεί είτε προσθετική είτε πολλαπλασιαστική σχέση μεταξύ των συνιστωσών. Επιπλέον, κάποια από αυτά περιορίζουν τους υπολογισμούς τους στην εκτίμηση του επιπέδου (*simple exponential smoothing*), ή της τάσης και του επιπέδου (*Holt* και *Damped*) της χρονοσειράς χωρίς να υποθέτουν εποχιακότητα, ενώ άλλα αποτελούν ολοκληρωμένες λύσεις ικανές να μοντελοποιήσουν οποιοδήποτε μοτίβο χρονοσειράς. Μειονέκτημά τους είναι κυρίως η αβεβαιότητα που υπάρχει κατά την επιλογή του βέλτιστου μοντέλου και την παραμετροποίησή του, το αυξημένο υπολογιστικό κόστος για να γίνει κάτι τέτοιο και η σχετικά μειωμένη ακρίβεια σε μέσο- και μακροπρόθεσμες προβλέψεις.

Σε παρόμοια λογική κινούνται και τα μοντέλα παλινδρόμησης κινητών μέσων όρων (*ARIMA*) τα οποία προσπαθούν να εξάγουν χρονικές σχέσεις μεταξύ των παρατηρήσεων τις οποίες και τις αξιοποιούν στη συνέχεια για την προέκτασή τους. Αν και τα συγκεκριμένα μοντέλα αρχικά κατασκευάστηκαν για στάσιμες χρονοσειρές, δύναται να χρησιμοποιηθούν για την προέκταση και μη στάσιμων χρονοσειρών (*Box & Jenkins, 1970*). Αποτελούν συνεπώς γενικευμένες και ολοκληρωμένες λύσεις με τα ίδια ωστόσο μειονεκτήματα που έχουν και τα μοντέλα εκθετικής εξομάλυνσης. Μάλιστα, δεδομένου ότι τα μοντέλα *ARIMA* δεν έχουν συγκεκριμένο πλήθος, η διαδικασία εντοπισμού του καταλληλότερου είναι αρκετά πιο αβέβαιη, ενώ για πολύπλοκα μοντέλα η παραμετροποίηση γίνεται μόνο προσεγγιστικά, κάτι που επίσης εμπεριέχει σημαντικό κίνδυνο. Τέλος, επειδή η πρόβλεψη των μοντέλων βασίζεται αποκλειστικά στις πιο πρόσφατες παρατηρήσεις, είτε πρόκειται για πραγματικά δεδομένα είτε για προβλέψεις που παρήχθησαν νωρίτερα, η χρήση τους οδηγεί σε ανακριβείς προβλέψεις για μακρινούς ορίζοντες πρόβλεψης.

Ένα άλλο πολύ γνωστό μοντέλο πρόβλεψης είναι η μέθοδος θ (*Assimakopoulos & Nikolopoulos, 2000*). Πρόκειται για ένα μοντέλο πρόβλεψης το οποίο βασίζεται στην μεταβολή των τοπικών καμπυλοτήτων της χρονοσειράς μέσω ενός γραμμικού μετασχηματισμού. Όταν κανείς θέλει να προεκτείνει τη χρονοσειρά στο μακρινό μέλλον συμφέρει να περιορίσει τις καμπυλοτήτες της για να αναδειχτούν τα μακροχρόνια χαρακτηριστικά της (τάση), ενώ αντίστοιχα όταν θέλει να την προεκτείνει στο κοντινό μέλλον να τις τονίσει για να προσδιοριστεί καλύτερα το τρέχον επίπεδό της. Έτσι η ιδέα της μεθόδου είναι η χρονοσειρά να αποσυντίθεται σε επιμέρους γραμμές, γνωστές ως γραμμές θ , κάθε μία από τις οποίες έχει την ίδια μέση τιμή και κλίση με την αρχική χρονοσειρά αλλά διαφορετική καμπυλότητα. Η πρόβλεψη κάθε μίας εξ αυτών ξεχωριστά και ο κατάλληλος συνδυασμός τους οδηγεί λοιπόν σε μία τελική πρόβλεψη η οποία εμπεριέχει πληροφορία τόσο για τα μακροπρόθεσμα όσο και βραχυπρόθεσμα χαρακτηριστικά των δεδομένων. Γραμμές με τιμές θ μικρότερες της μονάδας τονίζουν τα μακροχρόνια χαρακτηριστικά και το αντίστροφο.

Η μέθοδος θ έγινε ιδιαίτερα γνωστή μέσα από τα αποτελέσματα του διαγωνισμού *M3* (*Makridakis et al., 2000*), του μεγαλύτερου διαγωνισμού προβλέψεων που πραγματοποιήθηκε ποτέ και στον οποία συμμετείχαν μοντέλα, μέθοδοι και υπολογιστικά πακέτα προβλέψεων από πολλά πανεπιστήμια και εταιρίες του κόσμου. Η μέθοδος, αν και από τις απλούστερες που συμμετείχαν στο διαγωνισμό, κατάφερε να πρωτεύσει αναδεικνύοντας την αξία της τεχνικής αποσύνθεσης στην οποία βασίζεται. Παραδόξως η μέθοδος, αν και μελετήθηκε αρκετές φορές ως προς την ισχύ των αποτελεσμάτων του διαγωνισμού, ποτέ δεν αποτέλεσε ιδιαίτερο αντικείμενο μελέτης. Στην παρούσα εργασία, αναγνωρίζοντας την ανωτερότητα της μεθόδου και την απλότητά της, μελετώνται τα χαρακτηριστικά της και γίνεται προσπάθεια να αξιοποιηθούν περαιτέρω τα προτερήματά της και αντίστοιχα να επιλυθούν τα μειονεκτήματά της και οι περιορισμοί τους οποίους θέτει.

Ο βασικότερος περιορισμός της μεθόδου φαίνεται να είναι το γραμμικό μοτίβο τάσης το οποίο αυτή υποθέτει. Η κλασική προσέγγιση της μεθόδου θ αποτελείται από δύο γραμμές,

την 0 και τη 2. Η γραμμή μηδέν ισοδυναμεί με την ευθεία απλής γραμμικής παλινδρόμησης και είναι εκείνη που προσδίδει την τάση στις προβλέψεις. Αντίθετα, η γραμμή δύο προεκτείνεται μέσω απλής εκθετικής εξομάλυνσης και βοηθά στον προσδιορισμό του επιπέδου. Έτσι, γίνεται αμέσως αντιληπτό ότι χρονοσειρές οι οποίες χαρακτηρίζονται από μη γραμμικά μοτίβα τάσης, όπως π.χ. εκθετικό ή λογαριθμικό, δεν μπορούν να προεκταθούν ικανοποιητικά με τη χρήση της συγκεκριμένης μεθόδου. Η αντικατάσταση της γραμμής 0 με άλλη καμπύλη αντίστοιχου μοτίβου με αυτό της τάσης της φαντάζει λοιπόν μία ιδιαίτερος υποσχόμενη μετατροπή.

Μία επιπλέον επέκταση που εξετάζεται στα πλαίσια της παρούσας διατριβής είναι η έκφραση της μεθόδου θ τόσο με προσθετικό (κλασική προσέγγιση) όσο και με πολλαπλασιαστικό τρόπο. Αυτό έχει ως αποτέλεσμα οι συνιστώσες της τάσης και του επιπέδου να μπορούν να συνδέονται είτε προσθετικά είτε πολλαπλασιαστικά, αυξάνοντας έτσι την ευελιξία του μοντέλου και την ικανότητα προσαρμογής του. Αντίστοιχες εκφράσεις εμφανίζονται στη βιβλιογραφία για την περίπτωση των μοντέλων εκθετικής εξομάλυνσης με σημαντικά πλεονεκτήματα, τα οποία και ενέπνευσαν επί της ουσίας την εν λόγω προσθήκη. Χαρακτηριστικά αναφέρεται η παραγωγή θετικών προβλέψεων ανεξαρτήτως εισόδου και η αυξημένη ευαισθησία στις μεταβολές του επιπέδου.

Η γενίκευση της μεθόδου θ για μη γραμμικά μοτίβα τάσης πολλαπλασιαστικής και προσθετικής μορφής παρουσιάζεται αναλυτικά και η βελτιωμένη της ακρίβεια αποδεικνύεται μέσω της εφαρμογής των μοντέλων θ που προκύπτουν στα δεδομένα των διαγωνισμών M και σε μεγάλο πλήθος τεχνητών χρονοσειρών, και την επιτυχημένη επιλογή του επικρατέστερου για την τελική προέκτασή τους.

Δ. Μέθοδοι εξομάλυνσης

Όπως αναφέρθηκε νωρίτερα, η μεθοδολογία πρόβλεψης που προτείνεται στην παρούσα διατριβή βασίζεται στη χρήση της μεθόδου πρόβλεψης θ πάνω σε χρονοσειρές οι οποίες έχουν επιδεχθεί εξομάλυνση. Στο αντίστοιχο κεφάλαιο περιγράφονται με μεγαλύτερη σαφήνεια οι λόγοι για τους οποίους η εξομάλυνση φαίνεται να αποτελεί την αποδοτικότερη λύση για την παρούσα εφαρμογή έναντι άλλων εναλλακτικών και στη συνέχεια γίνεται μία προσπάθεια για τον καθορισμό και υιοθέτηση της πλέον αποτελεσματικής τεχνικής.

Η βιβλιογραφία παρουσιάζει ως τη συνηθέστερη μορφή εξομάλυνσης αυτή των κινητών μέσων όρων. Σύμφωνα με αυτήν κάθε παρατήρηση τη χρονοσειράς αντικαθίσταται από το σταθμισμένο άθροισμα των γειτονικών της. Έτσι, η νέα χρονοσειρά που προκύπτει είναι σαφώς πιο ομαλή και το τρέχον επίπεδο της πιο κοντά στο πραγματικό. Ανάλογα με το πλήθος των παρατηρήσεων που συναθροίζονται (εύρος) αυξάνει ή μειώνεται αντίστοιχα και ο βαθμός εξομάλυνσης, ενώ η επιλογή βαρών στη συνάθροιση καθορίζει το βαθμό στον οποίο συνεισφέρει κάθε παρατήρηση στο τελικό αποτέλεσμα δίνοντας έτσι περισσότερο ή λιγότερο βάρος στις πιο πρόσφατες παρατηρήσεις.

Μειονέκτημα της εν λόγω μεθόδου είναι αφενός ότι η εξομάλυνση δεν είναι δυνατό να πραγματοποιηθεί στα δύο άκρα της χρονοσειράς λόγω έλλειψης δεδομένων, και αφετέρου ότι η ύπαρξη μίας ακραίας τιμής στη χρονοσειρά θα επηρεάσει όλες τις γειτονικές της οδηγώντας μεν σε πιο ομαλό αποτέλεσμα από το αρχικό, όχι όμως και στο επιθυμητό. Το

πρώτο πρόβλημα περιορίζεται σε κάποιο βαθμό με χρήση της τεχνικής *backcasting*, της προέκτασης δηλαδή των δύο άκρων της χρονοσειράς με χρήση τεχνικών προβλέψεων. Ωστόσο, για μεγάλο βαθμό εξομάλυνσης ο ορίζοντας πρόβλεψης μεγαλώνει σημαντικά εισάγοντας έτσι σημαντική αβεβαιότητα στην όλη διαδικασία. Συνεπώς η τεχνική των κινητών μέσων όρων δεν μπορεί να θεωρηθεί ιδιαίτερα επιτυχημένη επιλογή.

Αντίστοιχα συμπεράσματα εξάγονται και για την περίπτωση συναφών μεθόδων όπως είναι η εξομάλυνση με χρήση *kernel* καθώς αποτελούν γενίκευση των κινητών μέσων όρων και συνεπώς χαρακτηρίζονται από τα ίδια μειονεκτήματα. Σε αυτήν την κατεύθυνση εξετάζεται λοιπόν ως εναλλακτική η χρήση του μετασχηματισμού θ καθώς αυτός μπορεί να συρρικνώσει αποτελεσματικά την καμπυλότητα της χρονοσειράς προσδίδοντας έτσι ομαλότερο αποτέλεσμα. Δεδομένου μάλιστα ότι ο μετασχηματισμός συρρικνώνει τις καμπύλες ανάλογα με την απόσταση της εκάστοτε παρατήρησης από την ευθεία της απλής γραμμικής παλινδρόμησης που ορίζεται, καταλαβαίνει κανείς πως η εν λόγω μέθοδος μπορεί να είναι περισσότερο αποτελεσματική από τους απλούς μέσους όρους σε περιπτώσεις ακραίων τιμών. Επίσης, καθώς ο μετασχηματισμός εφαρμόζεται απευθείας στις παρατηρήσεις της χρονοσειράς μία προς μία, δεν παράγει κενές παρατηρήσεις και συνεπώς λύνει το πρόβλημα που αντιμετωπίστηκε νωρίτερα. Μοναδικό αλλά καθοριστικό μειονέκτημα της μεθόδου αποτελεί η γραμμικότητα με την οποία θα χαρακτηρίζεται η νέα ομαλή χρονοσειρά που θα προκύψει καθώς όσο οι καμπυλότητες συρρικνώνονται, τόσο αυτή προσεγγίζει την ευθεία της απλής γραμμικής παλινδρόμησης οδηγώντας στη γενική περίπτωση σε προβλέψεις χαμηλής ακρίβειας.

Η εφαρμογή του μη γραμμικού μετασχηματισμού θ μπορεί να αποτελέσει μία βελτιωμένη λύση στο πρόβλημα της εξομάλυνσης καθώς συρρικνώνει σταδιακά τις καμπυλότητες της χρονοσειράς προσδίδοντας ένα μη γραμμικό αποτέλεσμα (*Assimakopoulos, 1995*). Κατά την εφαρμογή του μετασχηματισμού υπολογίζεται αρχικά η τοπική διακύμανση σε κάθε σημείο της χρονοσειράς και στη συνέχεια, μετακινώντας κατάλληλα το εκάστοτε σημείο, αυτή περιορίζεται επιλεκτικά κατά ένα προκαθορισμένο ποσοστό ξεκινώντας από τα σημεία με την υψηλότερη τιμή και προχωρώντας σειριακά προς αυτά με την μικρότερη. Αυτό έχει ως αποτέλεσμα αφενός πιθανές ακραίες τιμές να αντιμετωπίζονται με μεγαλύτερη αποτελεσματικότητα και αφετέρου τα βασικά μοτίβα της αρχικής χρονοσειράς να διατηρούνται σε ικανοποιητικό βαθμό. Μειονέκτημα της μεθόδου είναι το υψηλό υπολογιστικό κόστος και η χρήση κάποιων παραμέτρων όπως είναι ο βαθμός εξομάλυνσης και ο καθορισμός του σημείου στο οποίο τερματίζει η όλη διαδικασία της εξομάλυνσης. Επίσης ο μετασχηματισμός εισάγει δύο κενές τιμές στην αρχή και στο τέλος της χρονοσειράς, κάτι που όμως αντιμετωπίζεται εύκολα όπως και νωρίτερα με χρήση της τεχνικής *backcasting*.

Χρησιμοποιώντας ως εφαλτήριο τα παραπάνω ευρήματα, και τροποποιώντας κατάλληλα το μη γραμμικό μετασχηματισμό θ , εισάγεται μία νέα τεχνική εξομάλυνσης η οποία αναφέρεται ως μετασχηματισμός εξισορρόπησης τοπικών διακυμάνσεων. Η εν λόγω μη παραμετρική τεχνική περιορίζει τις τοπικές διακυμάνσεις της χρονοσειράς με τον ίδιο τρόπο που το κάνει και ο μη γραμμικός μετασχηματισμός θ , αντί ωστόσο η μείωση να γίνεται με προκαθορισμένο ποσοστιαίο τρόπο, αυτή καθορίζεται κάθε φορά από την τιμή των

γειτονικών παρατηρήσεων και έχει ως σκοπό η διακύμανση να μηδενίζεται τοπικά. Επειδή μάλιστα ο μετασχηματισμός εφαρμόζεται σε τετράδες σημείων δεν οδηγεί ποτέ σε απόλυτη ευθεία, διατηρώντας έτσι τα αρχικά μοτίβα της χρονοσειράς.

Τα παραπάνω χαρακτηριστικά του μετασχηματισμού εξισορρόπησης τοπικών διακυμάνσεων έχουν ως αποτέλεσμα λοιπόν αφενός να μην χρειάζεται να οριστεί κάπου ο βαθμός εξομάλυνσης της χρονοσειράς, και αφετέρου να μην καθορίζεται πρακτικά κάποιιο όριο ομαλότητας για τον τερματισμό της όλης διαδικασίας. Η επιλεκτικότητα της μεθόδου διατηρείται ενώ εισάγονται όπως και πριν δύο και μόνο κενές τιμές στα δύο άκρα της χρονοσειράς που μπορούν εύκολα να συμπληρωθούν. Το αποτέλεσμα όπως αποδεικνύεται είναι μία εξαιρετικά ομαλή χρονοσειρά με βελτιωμένες προβλεπτικές ιδιότητες που σχετίζονται τόσο με την ανάδειξη ενός αντιπροσωπευτικότερου τρέχοντος επιπέδου για την χρονοσειρά, όσο και με τη βελτιωμένη εκτίμηση της μακροχρόνιας τάσης της. Μάλιστα, καθώς η χρονοσειρά που προκύπτει είναι εντελώς ομαλή, το τρέχον επίπεδο της χρονοσειράς ταυτίζεται με την τελευταία γνωστή παρατήρησή της, γεγονός που απλοποιεί σημαντικά την πολυπλοκότητα οποιασδήποτε μεθόδου που θα χρησιμοποιούταν για την προέκτασή της. Στην περίπτωση έτσι της γενικευμένης μεθόδου θ , η πρόβλεψη ισοδυναμεί με μία καμπύλη η οποία ξεκινάει από την τελευταία γνωστή τιμή της εξομαλυμένης χρονοσειράς και προεκτείνεται βάσει του μοτίβου τάσης που έχουμε υποθέσει. Αυτή η απλοποιημένη τεχνική προέκτασης χρονοσειρών συνοψίζει ουσιαστικά την προτεινόμενη μεθοδολογία πρόβλεψης της παρούσας διατριβής.

E. Γεννήτρια Τεχνητών δεδομένων

Προκειμένου να εξακριβωθεί η αποτελεσματικότητα της προτεινόμενης μεθοδολογίας απαιτείται η εφαρμογή της και η αξιολόγησή της σε σχέση με άλλες μεθόδους για μεγάλο όγκο δεδομένων τα οποία διαθέτουν διαφορετικά χαρακτηριστικά (π.χ. εποχιακότητα, τάση, κύκλο και τυχαιότητα). Η παραπάνω διαδικασία αξιολόγησης συναντάται βιβλιογραφικά στους λεγόμενους διαγωνισμούς προβλέψεων κατά τους οποίους πολλές διαφορετικές μέθοδοι και πακέτα προβλέψεων χρησιμοποιούνται για την πρόβλεψη ενός σετ δεδομένων και στη συνέχεια αξιολογείται η επίδοσή τους τόσο σε απόλυτες τιμές σφάλματος όσο και συγκριτικά σε σχέση με το τι πέτυχαν οι υπόλοιπες μέθοδοι που συμμετείχαν.

Το πλεονέκτημα τέτοιων συγκρίσεων είναι ότι μπορούμε να κατατάξουμε με σχετική αξιοπιστία ως προς την ακρίβειά τους προσεγγίσεις για την παραγωγή προβλέψεων ανάλογα με το είδος των δεδομένων (π.χ. δημογραφικά, οικονομικά, βιομηχανικά κτλ.) και τα χαρακτηριστικά τους. Έτσι εξάγονται χρήσιμα συμπεράσματα και εμπειρικοί κανόνες οι οποίοι μπορεί να φανούν ιδιαίτερα χρήσιμοι μελλοντικά κατά τη διαδικασία της πρόβλεψης. Για παράδειγμα, βασικά συμπεράσματα που έχουν εξαχθεί μέσα από τους διαγωνισμούς προβλέψεων M είναι ότι (α) οι πιο απλές μέθοδοι είναι πιθανό να εμφανίσουν καλύτερη απόδοση προβλεπτικά από πιο σύνθετες, (β) με την αύξηση του ορίζοντα πρόβλεψης μειώνεται η ακρίβεια πρόβλεψης, (γ) ανάλογα με το κριτήριο αξιολόγησης που χρησιμοποιείται ενδέχεται να αλλάξει η κατάταξη μεταξύ των μεθόδων που συμμετέχουν, και (γ) ο συνδυασμός απλών μεθόδων αποτελεί συνήθως καλύτερη

επιλογή από την πρόβλεψη με τις επιμέρους μεθόδους ξεχωριστά. Άλλοι διαγωνισμοί προβλέψεων μικρότερου βεληνεκούς είναι ο *T* και αυτός των *Neutral Networks*.

Δυστυχώς βέβαια, όπως αναφέρθηκε νωρίτερα στην εισαγωγή, το να μπορέσει κανείς να αξιολογήσει αντικειμενικά τις μεθόδους προβλέψεων με χρήση πραγματικών δεδομένων είναι πρακτικά αδύνατο καθώς, ακόμα και να καταφέρει να συλλέξει ικανοποιητικό πλήθος χρονοσειρών, αυτές θα κατανέμονται ανομοιόμορφα ως προς τα χαρακτηριστικά τους και τις εφαρμογές στις οποίες αναφέρονται. Για παράδειγμα, τα δεδομένα του διαγωνισμού *M3*, τα οποία χρησιμοποιούνται πολύ συχνά για την αξιολόγηση νέων μεθόδων πρόβλεψης, είναι αφενός λίγα σε αριθμό (3003 χρονοσειρές) και αφετέρου έχει διαπιστωθεί πως η πλειοψηφία τους διαθέτει συγκεκριμένα χαρακτηριστικά. Έτσι, μέθοδοι οι οποίες ανταποκρίνονται σε αυτά αποδοτικότερα έχουν καλύτερες πιθανότητες να αποδώσουν καλά στο συγκεκριμένο διαγωνισμό. Ο εν λόγω προβληματισμός επεκτείνεται και σε άλλα σχετικά ζητήματα σε αντίστοιχη παράγραφο της εργασίας και το βασικό συμπέρασμα είναι πως για την αντικειμενικότερη αξιολόγηση των μεθόδων απαραίτητη προϋπόθεση είναι η κατασκευή και η αξιοποίηση τεχνητών χρονοσειρών ομοιόμορφα κατανεμημένων ως προς τα χαρακτηριστικά τους.

Για να παραχθούν χρονοσειρές με τεχνητό τρόπο απαιτείται πρώτα απ' όλα ο καθορισμός της διαδικασίας μέσω της οποίας αυτές θα γεννιούνται, του μηχανισμού δηλαδή που θα καθορίζει τα επιμέρους τους χαρακτηριστικά και θα τα συνθέτει. Στην παρούσα διατριβή υποθέτουμε πως κάθε χρονοσειρά μπορεί να αναλυθεί σύμφωνα με τη μέθοδο της πολλαπλασιαστικής αποσύνθεσης σε τρεις συνιστώσες: Τάση, Εποχιακότητα και Τυχαιότητα. Ο κύκλος απορρίφθηκε ως χαρακτηριστικό καθώς για να έχει αυτός νόημα οι χρονοσειρές που θα παραχθούν θα πρέπει να έχουν ιδιαίτερα μεγάλο μήκος, κάτι που εδώ δεν κρίνεται απαραίτητο. Έτσι, για κάθε χρονοσειρά παράγονται δείκτες εποχιακότητας και ένα μοτίβο τάσης, και στο γινόμενο αυτών εφαρμόζεται λευκός θόρυβος. Το αποτέλεσμα είναι μία νέα τυχαία χρονοσειρά με συγκεκριμένο επίπεδο εποχιακότητας, τάσης και θορύβου.

Προκειμένου βέβαια να μπορέσει κανείς να αξιολογήσει αποδοτικότερα και αναλυτικότερα την επίδοση των μεθόδων ανά είδος χρονοσειράς, ορίζονται επιπλέον τρία επίπεδα έντασης για κάθε χαρακτηριστικό. Έτσι, κάθε χρονοσειρά μπορεί να είναι Χαμηλής, Μεσαίας ή Υψηλής έντασης ως προς την εποχιακότητα, την τάση και την τυχαιότητα της. Ο τρόπος με τον οποίο μετράτε το κάθε χαρακτηριστικό και τα όρια κάθε κατηγορίας έντασης καθορίζονται εντός της εργασίας χρησιμοποιώντας ως βασική αναφορά πραγματικά δεδομένα από τον διαγωνισμό *M3*. Οι χρονοσειρές που παράγονται ανά κατηγορία είναι 1,000, έχουν μηνιαία εποχιακότητα και διαθέτουν συνολικά 100 παρατηρήσεις. Από αυτές, οι 82 χρησιμοποιούνται για την εκπαίδευση των μοντέλων και οι 18 για την αξιολόγησή του, όπως γίνεται δηλαδή και στην περίπτωση του διαγωνισμού *M3*. Επιπλέον, προκειμένου να εξακριβωθεί κατά πόσο η γενικευμένη μέθοδος θ μπορεί να οδηγήσει σε βελτιωμένη προβλεπτική ακρίβεια, παράγονται πέρα από χρονοσειρές με γραμμικό μοτίβο τάσης, και άλλες με εκθετική, λογαριθμική, αντίστροφη ή σε μορφή δύναμης τάση.

Η γεννήτρια υλοποιήθηκε με χρήση της γλώσσας προγραμματισμού R , μίας γλώσσας ελεύθερου λογισμικού η οποία αναπτύσσεται ραγδαία τα τελευταία χρόνια λόγω των ισχυρών βιβλιοθηκών που διαθέτει. Ο χρήστης της γεννήτριας θέτει ως μοναδική είσοδο του προγράμματος το πλήθος χρονοσειρών που επιθυμεί να δημιουργήσει και στη συνέχεια αυτό παράγει και αποθηκεύει τις νέες χρονοσειρές σε αντίστοιχα αρχεία .csv. Πέρα από τις χρονοσειρές αυτές καθαυτές, το πρόγραμμα μπορεί να υπολογίσει απευθείας σφάλματα πρόβλεψης για τις επιμέρους χρονοσειρές και κατηγορίες χαρακτηριστικών που ορίστηκαν νωρίτερα. Η γεννήτρια είναι διαθέσιμη και στη διεύθυνση <http://fsudataset.com/>.

ΣΤ. Ανάλυση και εξέταση αποτελεσμάτων

Στο συγκεκριμένο κεφάλαιο αναλύονται τα αποτελέσματα της προτεινόμενης μεθοδολογίας προκειμένου να αναδειχτεί η ανωτερότητά της έναντι άλλων κλασικών μεθόδων πρόβλεψης. Η αξιολόγηση αφορά δύο διαφορετικά σετ δεδομένων τα οποία και εξετάζονται ξεχωριστά: Τα δεδομένα των διαγωνισμών M και τα τεχνητά δεδομένα που παράχθηκαν μέσω της γεννήτριας που κατασκευάστηκε νωρίτερα. Σκοπός είναι να εξαχθούν αντικειμενικά συμπεράσματα δοκιμάζοντας την προτεινόμενη μεθοδολογία τόσο σε πραγματικές συνθήκες, όσο και σε τεχνητές που όμως χαρακτηρίζονται από μεγάλη ποικιλία και ανομοιομορφία.

Πιο συγκεκριμένα, η μεθοδολογία εφαρμόζεται αρχικά επί 135.000 μηνιαίων χρονοσειρών, παραχθειςών από τη γεννήτρια με τεχνητό τρόπο, οι οποίες χαρακτηρίζονται από διαφορετικά μοτίβα και επίπεδα έντασης τάσης, εποχιακότητας και τυχαιότητας. Στη συνέχεια, η μεθοδολογία εφαρμόζεται επί των 4004 πραγματικών χρονοσειρών των διαγωνισμών M για την επιβεβαίωση των αποτελεσμάτων.

Δεδομένου ότι οι χρονοσειρές της γεννήτριας είναι προκαθορισμένων χαρακτηριστικών, γίνεται αντιληπτό πως η πρώτη αξιολόγηση θα προσδώσει μία αναλυτική εικόνα για τον τρόπο με τον οποίο συμπεριφέρεται η μεθοδολογία ανά τύπο χρονοσειράς, και μάλιστα για μεγάλο δείγμα αυτού. Από την άλλη μεριά, τα δεδομένα των διαγωνισμών χαρακτηρίζονται επιπλέον από διαφορετικές συχνότητες (ετήσιες, τριμηνιαίες, μηνιαίες, άλλα) και είδη εφαρμογής (δημογραφικά, μικροοικονομικά, μακροοικονομικά, βιομηχανικά και άλλα), στοιχεία απόντα στην περίπτωση των τεχνητών δεδομένων. Έτσι, η δεύτερη αξιολόγηση θα αξιοποιηθεί προκειμένου να επιβεβαιώσει τα συμπεράσματα που εξάχθηκαν μέσω της γεννήτριας, να ερευνήσει τυχόν αποκλίσεις, να μελετήσει την επίδοσή της μεθοδολογίας ανά κατηγορία εφαρμογής και να εκτιμήσει την μέση της απόδοση στον πραγματικό κόσμο.

Τα αποτελέσματα και στις δύο περιπτώσεις δείχνουν πως η προτεινόμενη μέθοδος μπορεί να βελτιώσει σημαντικά την ακρίβεια πρόβλεψης και μάλιστα ανεξαρτήτως των χαρακτηριστικών της χρονοσειράς. Για λόγους πληρότητας εξετάζεται βέβαια η συμπεριφορά της μεθόδου ανά ένταση χαρακτηριστικού και εξάγονται αντίστοιχα συμπεράσματα. Συνοπτικά, η μέθοδος φαίνεται να αποδίδει καλύτερα από τις κλασικές προσεγγίσεις σε θορυβώδεις, εποχιακές χρονοσειρές μεσαίας ή χαμηλής έντασης τάσης, ανεξαρτήτως μοτίβου. Αυτό οφείλεται αφενός στις συνδυαστικές επιπτώσεις των τεχνικών

συρρίκνωσης εποχιακών δεικτών και εξομάλυνσης, και αφετέρου στην επιλεκτικότητα της μεθοδολογίας η οποία και καταφέρνει μέσω των διαγνωστικών της ελέγχων να εντοπίσει με επιτυχία τα πιο αποδοτικά μοντέλα πρόβλεψης θ για την ακριβέστερη προέκταση των χρονοσειρών.

Ζ. Συμπεράσματα και προοπτικές

Η εφαρμογή της προτεινόμενης μεθοδολογίας σε πλήθος πραγματικών και τεχνητών δεδομένων αποδεικνύει όπως αναφέρθηκε νωρίτερα την ανωτερότητά της έναντι άλλων κλασικών προσεγγίσεων, οδηγώντας σε σημαντικά βελτιωμένες προβλέψεις. Ωστόσο, πέραν της γενικότερης βελτιωμένης απόδοσης της μεθόδου, υπάρχουν αρκετά χρήσιμα συμπεράσματα που εξάγονται μέσω της παρούσας διατριβής.

Πιο συγκεκριμένα, η συρρίκνωση των εποχιακών δεικτών φαίνεται να βελτιώνει συστηματικά τις προβλέψεις των εποχιακών χρονοσειρών, ενώ όσο αυξάνεται η εποχιακότητα, τόσο φαίνεται να αυξάνει και το κέρδος της προτεινόμενης προσέγγισης. Επιπλέον, στη γενική περίπτωση η εξομάλυνση των χρονοσειρών οδηγεί σε βελτιωμένη προβλεπτική ακρίβεια. Εξαιρέσεις ενδέχεται να υπάρξουν κυρίως σε χρονοσειρές που είναι ήδη επαρκώς ομαλές, χωρίς όμως αυτό να αποτελεί ιδιαίτερο πρόβλημα για τη μέθοδο δεδομένου ότι σε ομαλές χρονοσειρές τα σφάλματα είναι ήδη περιορισμένα. Σημειώνεται βέβαια πως από ένα επίπεδο θορύβου και μετά η προτεινόμενη προσέγγιση ενδέχεται να αδυνατεί να προσδώσει σημαντικά κέρδη καθώς η τυχαιότητα υπερέρχει του σήματος. Εκτός αυτού, φαίνεται πως όσο αυξάνεται η τάση της χρονοσειράς, τόσο μειώνονται τα προσδοκόμενα οφέλη της μεθόδου. Ωστόσο, για λογικά επίπεδα τάσης και τυχαιότητας η προτεινόμενη προσέγγιση οδηγεί σε σημαντικά καλύτερες προβλέψεις.

Τα αποτελέσματα δείχνουν επίσης πως τα οφέλη συρρίκνωσης εποχιακών δεικτών και εξομάλυνσης είναι ανεξάρτητα μεταξύ τους, πράγμα που σημαίνει πως αυτά μπορούν να απομονωθούν όποτε αυτό κρίνεται απαραίτητο, όπως π.χ. σε περιπτώσεις ομαλών χρονοσειρών. Ακόμα η σωστή επιλογή μοτίβου τάσης, όπως π.χ. το εκθετικό, φαίνεται πως μπορεί να βελτιώσει σημαντικά την ακρίβεια των προβλέψεων ανεξαρτήτως συχνότητας χρονοσειράς και είδους εφαρμογής. Έτσι, ανάλογα με το μοτίβο τάσης της χρονοσειράς, διαφορετικά μοντέλα μπορούν να αξιοποιηθούν και να οδηγήσουν σε βελτιωμένες προβλέψεις: Εκθετικό για γραμμική τάση, Εκθετικό για εκθετική τάση, Λογαριθμικό για Λογαριθμική Τάση, Αντίστροφο για αντίστροφη τάση και Γραμμικό για τάση σε μορφή δύναμης.

Ένα άλλο χρήσιμο συμπέρασμα είναι πως τα προσθετικά μοντέλα θ είναι στη γενική περίπτωση πιο ακριβή από τα πολλαπλασιαστικά. Ωστόσο οι διαφορές τους είναι σχετικά μικρές, ειδικά για μη γραμμικά μοτίβα τάσης, και μεμονωμένα και οι δύο εκφράσεις της μεθόδου μπορούν να προσδώσουν οφέλη. Αυτός είναι και ο λόγος που η επιλογή του βέλτιστου μοντέλου θ βάσει κατάλληλων διαγνωστικών ελέγχων οδηγεί συνολικά σε καλύτερες προβλέψεις, χωρίς αυτό να σημαίνει βέβαια πως επιλέγεται πάντα η βέλτιστη δυνατή λύση. Επίσης, μετά την επεξεργασία των δεδομένων, οι διαφορές γίνονται ακόμα μικρότερες.

Σχετικά με τις επιλογές μεταξύ των μοντέλων θ αναφέρεται πως η επιλογή του βέλτιστου μοτίβου τάσης είναι σημαντικότερη της έκφρασης του μοντέλου καθώς, ειδικά μετά την επεξεργασία των δεδομένων, οι διαφορές μεταξύ πολλαπλασιαστικών και προσθετικών μοντέλων ίδιας τάσης είναι αμελητέες ενώ οι διαφορές μεταξύ διαφορετικών μοντέλων τάσης διατηρούνται μεγάλες. Βέβαια όλα τα μοντέλα παραμένουν εξαιρετικά χρήσιμα και αυτό φαίνεται από το γεγονός πως στα πειράματα που πραγματοποιήθηκαν αυτά επιλέγονται σε σεβαστό ποσοστό από τον αλγόριθμο που κατασκευάστηκε οδηγώντας σε αύξηση της ακρίβειας της μεθόδου. Σημειώνεται επίσης πως η επιλογή του βέλτιστου μοντέλου πρόβλεψης είναι ακριβέστερη για ομαλές χρονοσειρές, δηλαδή για χρονοσειρές περιορισμένης εποχιακότητας και τυχαιότητας, καθώς και για χρονοσειρές αυξημένης τάσης. Αυτό συμβαίνει αφενός γιατί οι διακυμάνσεις, είτε τυχαίες είτε συστηματικές, αλλοιώνουν σε κάποιο βαθμό το μοτίβο της χρονοσειράς, και αφετέρου γιατί όταν είναι έντονη η μορφή της τάσης, αυτή είναι ευκολότερο να εντοπιστεί από τον αλγόριθμο επιλογής.

Συνοπτικά, με την εφαρμογή της προτεινόμενης μεθοδολογίας μπορεί κανείς να επιλέξει με σχετική ασφάλεια το βέλτιστο μοντέλο θ για κάθε χρονοσειρά και να αποκομίσει σημαντικά οφέλη σε ακρίβεια και υπολογιστικό κόστος. Αν η ένταση της τυχαιότητας είναι ιδιαίτερα χαμηλή, τότε ενδέχεται να πρέπει να παρακαμφθεί το βήμα της εξομάλυνσης, ενώ αν η χρονοσειρά είναι στάσιμη με σημαντικές διακυμάνσεις, τότε ενδέχεται η επιλογή μοντέλου να μην είναι η βέλτιστη.

Ως προς τις προεκτάσεις της παρούσας διατριβής, πρωταρχικό μέλημα θα ήταν η ενσωμάτωση στην προτεινόμενη προσέγγιση επιπλέον τεχνικών προετοιμασίας δεδομένων προκειμένου η γενικευμένη μέθοδος θ να είναι σε θέση να οδηγήσει σε ακόμα καλύτερες προβλέψεις. Όπως φάνηκε από τα αποτελέσματα της εργασίας, τα οφέλη που προέρχονται από την προ-επεξεργασία των δεδομένων είναι ιδιαίτερα σημαντικά και ως εκ τούτου αξίζει να μελετηθεί περαιτέρω. Έτσι, η εισαγωγή π.χ. μίας τεχνικής για την αναγνώριση και την αφαίρεση ειδικών γεγονότων που αλλοιώνουν τα μοτίβα της χρονοσειράς και αποπροσανατολίζουν το μοντέλο πρόβλεψης, θα ήταν μία πρώτη χρήσιμη προέκταση. Μάλιστα, δεδομένου ότι η προτεινόμενη εξομάλυνση αφαιρεί ακραίες τιμές (additive outliers) αλλά δυσκολεύεται εκ φύσεως να απομακρύνει άλλα είδη ειδικών γεγονότων, όπως π.χ. μεταβολές επιπέδου (level shift), η παραπάνω διαδικασία θα μπορούσε να επικεντρωθεί σε συγκεκριμένες κατηγορίες outlier, περιορίζοντας έτσι ταυτόχρονα το επιπλέον υπολογιστικό κόστος.

Άλλη προέκταση της παρούσας διατριβής θα ήταν ο αναλυτικότερος έλεγχος του τρόπου με τον οποίο αλληλοεπιδρά η προτεινόμενη προσέγγιση επεξεργασίας δεδομένων με άλλα μοντέλα πρόβλεψης χρονοσειρών εκτός του θ . Για παράδειγμα, στο κεφάλαιο που παρουσιάστηκε η τεχνική εξομάλυνσης φάνηκε πως τα μοντέλα εκθετικής εξομάλυνσης με τάση, όπως η *holt*, δεν συνεργάζονται αποδοτικά με την προτεινόμενη τεχνική οδηγώντας στη γενική περίπτωση σε χειρότερες προβλέψεις. Αυτό οφείλεται όπως αναφέρθηκε στο γεγονός ότι η εξομάλυνση υπερτονίζει την πρόσφατη τάση στα δεδομένα, την οποία και τα μοντέλα εκθετικής εξομάλυνσης την εφαρμόζουν αυτούσια στην πρόβλεψη. Έτσι, σε αντίθεση με την μέθοδο θ που στηρίζεται στην μακροχρόνια και στιβαρή τάση της

χρονοσειράς, αυτά οδηγούνται σε προκατειλημμένες προβλέψεις υψηλού σφάλματος. Ως εκ τούτου, θα είχε ιδιαίτερο ενδιαφέρον να μελετηθεί ποιοι μέθοδοι μπορούν να αξιοποιηθούν εναλλακτικά για τη βελτίωση της απόδοσής τους, ή αντίστοιχα ποιες παραμετροποιήσεις πρέπει να γίνουν στην προτεινόμενη τεχνική εξομάλυνσης για την παραγωγή ακριβέστερων προβλέψεων. Για παράδειγμα, οι μετασχηματισμοί δεδομένων έχει αποδειχτεί στη βιβλιογραφία πως συμβάλουν στη βελτίωση της προβλεπτικής ικανότητας των εν λόγω μοντέλων, σε περιορισμένο όμως βαθμό. Ίσως η μελέτη άλλων τεχνικών πέραν των μετασχηματισμών δεδομένων να κρύβουν σημαντικότερα οφέλη.

Άλλη προέκταση της εργασίας θα μπορούσε να αφορά τη βελτίωση ή την αντικατάσταση του αλγορίθμου επιλογής βέλτιστου μοντέλου θ με άλλη αποδοτικότερη προσέγγιση. Για παράδειγμα, θα μπορούσε να εξεταστεί η πιθανότητα του συνδυασμού των προβλέψεων που προέρχονται από τα επικρατέστερα μοντέλα θ , ή της εξαίρεσης των χειρότερων εξ αυτών. Έτσι, τα οφέλη του συνδυασμού προβλέψεων θα έβρισκαν άμεσα εφαρμογή στην παρούσα μεθοδολογία και θα οδηγούσαν πιθανώς σε καλύτερες προβλέψεις. Τέτοιες πρακτικές ερευνήθηκαν σε κάποιο βαθμό ωστόσο τα αποτελέσματα ήταν διφορούμενα και χρήζουν περισσότερης μελέτης.

Μεγάλο ενδιαφέρον θα εμφάνιζε φυσικά η εφαρμογή και η αξιολόγηση της προτεινόμενης προσέγγισης σε πραγματικό περιβάλλον. Μπορεί η μεθοδολογία που αναπτύχθηκε να εφαρμόστηκε με επιτυχία σε μεγάλο πλήθος χρονοσειρών πληθώρας χαρακτηριστικών, επιβεβαιώνοντας την προβλεπτική της ανωτερότητα, ωστόσο η χρήση της για παραγωγή προβλέψεων στο πραγματικό περιβάλλον μιας επιχείρησης θα μπορούσε να οδηγήσει σε επιπλέον χρήσιμα συμπεράσματα. Για παράδειγμα, ίσως να υπάρχουν τύποι δεδομένων που δεν εξετάστηκαν και για τους οποίους η μεθοδολογία δεν αποδίδει ιδιαίτερα καλά. Επίσης, είναι πιθανό βάσει χρόνου εκτέλεσης και ακρίβειας που επιθυμείται από την επιχείρηση να μην κρίνεται σε όλες της περιπτώσεις απαραίτητη η χρήση της μεθοδολογίας έναντι άλλων πιο απλοϊκών προσεγγίσεων. Ο εντοπισμός τέτοιων περιπτώσεων και η εύρεση πρακτικά αποδοτικών λύσεων είναι νευραλγική για την ουσιαστική προσφορά αντίστοιχων μεθόδων στις επιχειρήσεις.

Άλλη χρήσιμη επέκταση της παρούσας διατριβής είναι η ενσωμάτωσή της εντός ενός ολοκληρωμένου συστήματος υποστήριξης προβλέψεων προκειμένου αυτή να γίνει άμεσα και εύκολα διαθέσιμη σε επιχειρήσεις και οργανισμού. Η εν λόγω επέκταση έχει ήδη ξεκινήσει υπό την ονομασία *OMEN* και αυτή τη στιγμή βρίσκεται στο στάδιο ελέγχου ποιότητας λειτουργίας και προσθήκης επιπλέον δυνατοτήτων. Το *OMEN* είναι ένα διαδικτυακό εργαλείο παραγωγής προβλέψεων το οποίο κάνει χρήση των βιβλιοθηκών της γλώσσας προγραμματισμού R - στην οποία και αναπτύχθηκαν όλες οι μέθοδοι και οι τεχνικές προβλέψεων που παρουσιάζονται στην εργασία – για την ανάλυση και την προέκταση των χρονοσειρών, ενώ επιπλέον χρησιμοποιεί διεπαφές ούτως ώστε οι χρήστες να μπορούν να αξιοποιούν τις τεχνικές που παρέχονται μέσω εύχρηστων οθονών και εργαλείων. Η εφαρμογή είναι διαθέσιμη στη διεύθυνση <http://omen-project.eu/> και για λόγους πληρότητας, πέραν της παρούσας μεθοδολογίας, ενσωματώνει και μεγάλο πλήθος άλλων κλασικών και προηγμένων τεχνικών και μοντέλων προβλέψεων.

Η επέκταση της παρούσας μεθοδολογίας για χρονοσειρές διακοπτόμενης ζήτησης θα είχε και αυτή με τη σειρά της αρκετό ενδιαφέρον, κυρίως λόγου του ότι όταν κανείς αναφέρεται σε μεγάλο όγκο δεδομένων σε επιχειρήσεις και οργανισμού, έρχεται αντιμέτωπος συνήθως με την παραγωγή προβλέψεων σε επίπεδο καταστήματος και κωδικού προϊόντων. Σε αυτές τις περιπτώσεις τα δεδομένα χαρακτηρίζονται από διακοπές, και συνεπώς χρήζουν επιπλέον επεξεργασίας προκειμένου κοινά μοντέλα πρόβλεψης, όπως η μέθοδος θ , να μπορούν να τα διαχειριστούν αποδοτικά.

Τέλος, δεδομένου ότι η μέθοδος θ , αν απλοϊκή στη βάση της, αποδεικνύεται μία ευέλικτη και ιδιαίτερος ακριβής τεχνική προέκτασης χρονοσειρών, γίνεται κατανοητό πως η περαιτέρω έρευνα γύρω από της ιδιότητές της έχει ακόμα πολλά να προσφέρει. Για παράδειγμα, θα είχε ενδιαφέρον να ενσωματωθεί στη μεθοδολογία μία τεχνική για την αυτόματη επιλογή της βέλτιστης παραμέτρου θ , η οποία και όπως παρουσιάστηκε μπορεί να επηρεάσει σημαντικά την προβλεπτική ικανότητα των μοντέλων θ που προκύπτουν. Επιπλέον, θα μπορούσε να εξεταστεί η χρήση πολλαπλών γραμμών θ , η εισαγωγή νέων γραμμών μηδέν, ο συνδυασμός πολλών διαφορετικών μοτίβων τάσης ταυτοχρόνως και η επιλογή διαφορετικών μοντέλων πρόβλεψης πέραν της SES για την προέκταση καθεμίας εκ των γραμμών θ που προκύπτουν.

Όσο πιο πίσω κοιτάζεις στο παρελθόν, τόσο πιο μπροστά στο μέλλον βλέπεις.

Winston Churchill

1.1 Γενικά για τις προβλέψεις

Καθημερινά καλούμαστε να λάβουμε αποφάσεις για πράγματα που θα συμβούν στο μέλλον. Προκειμένου να το κάνουμε αυτό, είτε συνειδητά είτε ασυνείδητα πραγματοποιούμε εκτιμήσεις σχετικά με την εξέλιξη ποικίλων μεγεθών. Για παράδειγμα, ο καθορισμός της ώρας που θα πρέπει να ξεκινήσουμε από το σπίτι προκειμένου να είμαστε εγκαίρως σε μία συνάντηση εμπεριέχει μία ασυνείδητη εκτίμηση της κίνησης που θα συναντήσουμε στο δρόμο. Αντίστοιχα, ο καθορισμός των φιαλών μπύρας που πρέπει να εμφιαλώσει μία ζυθοποιία εμπεριέχει μία συνειδητή εκτίμηση των πωλήσεων της για το επόμενο χρονικό διάστημα. Όπως γίνεται αντιληπτό, ακόμα και αν συχνά δεν το αντιλαμβανόμαστε, οι προβλέψεις αποτελούν αναπόσπαστο κομμάτι της ζωής μας και καθορίζουν καθημερινά σημαντικές και ασήμαντες αποφάσεις μας.

Φυσικά, όσο και αν μία ασήμαντη απόφαση για μας, όπως το αν θα πάρουμε ομπρέλα, μπορεί να ληφθεί αρκετά εύκολα με την απλή λογική, η πρόβλεψη του μεγέθους με το οποίο αυτή σχετίζεται (εδώ της βροχόπτωσης) μπορεί να αποτελεί πληροφορία μείζονος σημασίας για κάποιο άλλο πρόσωπο. Σε αυτήν την περίπτωση ο αποφασίζων χρειάζεται μία σαφώς πιο στιβαρή πρόβλεψη από αυτήν της απλής ανθρώπινης λογικής. Έτσι, στο παραπάνω παράδειγμα ένας γεωργός θα προτιμήσει να ενημερωθεί π.χ. μέσω του δελτίου καιρού σχετικά με την πιθανότητα βροχόπτωσης στην περιοχή του από το να την εκτιμήσει ο ίδιος βάσει προσωπικής εμπειρίας. Βασικοί παράγοντες συνεπώς για την σημασία της κάθε πρόβλεψης στη ζωή μας είναι αφενός ο βαθμός αβεβαιότητας που χαρακτηρίζει το μέγεθος που καλούμαστε να προβλέψουμε και αφετέρου η σημασία του για τον αποφασίζων.

Δεδομένης λοιπόν της αβεβαιότητας και της χρησιμότητας που θα είχαν τεκμηριωμένες προβλέψεις σε σημαντικές εφαρμογές, όπως είναι ο σχεδιασμός της στρατηγικής για μία επιχείρηση ή ένα κράτος, οι προβλέψεις άρχισαν να λαμβάνουν σημαντικό ερευνητικό ενδιαφέρον κυρίως στα τέλη της δεκαετίας του '80. Καταλυτικός παράγοντας σε αυτήν την εξέλιξη ήταν η προώθηση των ηλεκτρονικών μέσων τα οποία εξασφάλισαν τη συλλογή επαρκών ιστορικών δεδομένων για τη μελέτη του εκάστοτε μεγέθους ενδιαφέροντος και την ευκολότερη εφαρμογή μεθόδων πρόβλεψης επί αυτών για την εξαγωγή προβλέψεων σε ικανοποιητικό χρόνο. Η μελέτη των τεχνικών προβλέψεων πραγματοποιήθηκε κυρίως από ακαδημαϊκούς, αν και σε μερικές περιπτώσεις πρακτικά ασκούμενοι συνέβαλαν και αυτοί στην εξέλιξή τους.

Σήμερα, έπειτα από σχεδόν 40 χρόνια από την εμφάνιση των πρώτων σχετικών μελετών, οι τεχνικές προβλέψεων έχουν εξελιχθεί σημαντικά και αποτελούν αναπόσπαστο κομμάτι της καθημερινής λειτουργίας μεγάλων και μεσαίων επιχειρήσεων. Με τη διαθέσιμη υπολογιστική ισχύ χιλιάδες χρονοσειρές είναι δυνατόν να προβλεφθούν άμεσα με κατάλληλα μοντέλα και στη συνέχεια να αξιοποιηθούν από τα αντίστοιχα τμήματα

παραγωγής, πωλήσεων και στρατηγικού σχεδιασμού. Φυσικά η έρευνα δεν σταματά ποτέ και επικεντρώνεται κάθε φορά σε επιστημονικές περιοχές και εφαρμογές που χρήζουν περαιτέρω μελέτης και ανάπτυξης.

Σε αυτό το σημείο αξίζει να σημειωθεί φυσικά πως όσο και αν εξελιχθούν τα μοντέλα πρόβλεψης, ποτέ δεν θα καταστεί δυνατός ο επακριβής προσδιορισμός του μέλλοντος. Τα μοντέλα πρόβλεψης δεν είναι προφήτες και συνεπώς πάντα οι εκτιμήσεις τους θα οδηγούν σε μικρά ή μεγάλα σφάλματα. Αυτό προκύπτει όπως αναφέρθηκε νωρίτερα από την αβεβαιότητα που χαρακτηρίζει τη ζωή μας και την οποία θα πρέπει για αυτό το λόγο να λαμβάνουμε σοβαρά υπόψη πριν βιαστούμε να ασκήσουμε αυστηρή κριτική στις προβλέψεις. Σημασία έχει να είμαστε σε θέση να εντοπίζουμε και να εκμεταλλευόμαστε στο μέγιστο βαθμό την πληροφορία που κρύβεται στο παρελθόν και λαμβάνοντας υπόψη σχετικές μεταβλητές να προσφέρουμε λογικές και όσον το δυνατόν πιο αξιόπιστες προβλέψεις. Αστοχίες θα υπάρχουν πάντα, ιδιαίτερα σε περιόδους μεγάλης αβεβαιότητας. Όπως συνηθίζεται να λέγεται, όλα τα μοντέλα είναι λάθος, κάποια όμως είναι χρήσιμα.

Προς υποστήριξη των παραπάνω ισχυρισμών, ο καθηγητής Μακρυδάκης στο βιβλίο του *Χορεύοντας με την Τύχη* αναφέρει πως η αβεβαιότητα διακρίνεται σε αυτήν του 'μετρώ' και σε αυτήν της 'καρύδας'. Μπορούμε να πέσουμε έξω στην πρόβλεψή μας για το πότε θα φτάσουμε στη δουλειά μας κατά 2 ή 3 λεπτά λόγω κάποιας καθυστέρησης χωρίς κάτι τέτοιο να υποδηλώνει ότι το μοντέλο πρόβλεψης που χρησιμοποιήσαμε είναι προβληματικό. Οι προβλέψεις μας ήταν ρεαλιστικές και θα πρέπει να τις εμπιστευτούμε ξανά για τον προγραμματισμό μας καθώς διακυμάνσεις στους χρόνους άφιξης είναι αναμενόμενες. Αντίθετα, ποτέ δεν θα μπορούσαμε να προβλέψουμε αν περπατώντας σε ένα νησί θα πέσει μία καρύδα στο κεφάλι μας. Το γεγονός αυτό είναι εντελώς τυχαίο, σπάνιο και κανένα μοντέλο πρόβλεψης δεν θα μπορούσε ποτέ να μας απαντήσει για το αν θα συνέβαινε.

Τραγική ειρωνεία αποτελεί βέβαια το γεγονός ότι οι τεχνικές προβλέψεων αποκτούν αυξημένη ζήτηση σε περιόδους αστάθειας και κρίσης, εκεί δηλαδή όπου και οι παραγόμενες προβλέψεις αναμένεται να εμφανίζουν μεγαλύτερη απόκλιση από την πραγματικότητα, να υπάρξουν δηλαδή αρκετές 'καρύδες'. Αντίθετα περιφρονούνται ή λαμβάνουν αυστηρότατη κριτική σε περιόδους σταθερότητας όπου και μπορούν να επιφέρουν σημαντικά οφέλη σε επιχειρήσεις και οργανισμούς. Χαρακτηριστικό είναι επίσης το φαινόμενο κατά το οποίο κανείς θα βιαστεί να απορρίψει ένα αξιόπιστο μοντέλο που σε μία μεμονωμένη στιγμή απέτυχε να προβλέψει με ακρίβεια, αλλά με έντονο ενδιαφέρον θα υιοθετήσει τις προβλέψεις κάποιου άλλου που αποτύγχανε συστηματικά αλλά πρόβλεψε με ακρίβεια ένα αναπάντεχο γεγονός, προφανώς βασιζόμενο σε μεγάλο βαθμό στη τύχη. Και τα δύο φαινόμενα σχετίζονται στενά με την ανθρώπινη φύση και το πώς αυτή επηρεάζεται από συναισθήματα όπως ο φόβος για το μέλλον και ο εντυπωσιασμός.

Προς περαιτέρω υποστήριξη των τεχνικών προβλέψεων και της συνεχούς μελέτης τους αρκεί να αναλογιστούμε τη διαφωνία των θεωρητικών στατιστικολόγων με αυτούς που μελετούν τις τεχνικές προβλέψεων. Οι στατιστικολόγοι υποστηρίζουν ότι τα

συμπεράσματα των μελετών που σχετίζονται με τις προβλέψεις είναι αυθαίρετα καθώς δεν αποδεικνύονται θεωρητικά αλλά υποστηρίζονται μόνο εμπειρικά μέσω πειραμάτων και δοκιμών. Σύμφωνα με αυτούς το βέλτιστο μοντέλο πρόβλεψης είναι εκείνο που προσαρμόζεται καλύτερα στα ιστορικά δεδομένα. Αυτός ο ισχυρισμός φυσικά δεν ισχύει σε καμία περίπτωση καθώς τα αποτελέσματα πληθώρας μελετών και διαγωνισμών μοντέλων πρόβλεψης έχουν αποδείξει εδώ και αρκετά χρόνια ότι το σφάλμα προσαρμογής δεν συσχετίζεται με αυτό των προβλέψεων καθώς και ότι η χρήση πολύπλοκων μεθόδων δεν εγγυάται σε καμία περίπτωση καλύτερες προβλέψεις σε σχέση με ένα απλούστερο μοντέλο. Η σωστή προσέγγιση είναι να αντιληφθούμε ότι οι προβλέψεις παράγονται για να εξυπηρετήσουν επιχειρήσεις και να λύσουν προβλήματα της καθημερινής μας ζωής και συνεπώς, ανεξάρτητα από την θεωρία, η μελέτη τους συνάδει με τα εμπειρικά αποτελέσματα και την επίδοση των μοντέλων στην πράξη.

Συμπερασματικά, πρωταρχικό μας μέλημα καθώς εξελίσσεται η επιστήμη των προβλέψεων είναι η σωστή πληροφόρηση εκείνων που αξιοποιούν τις προβλέψεις προκειμένου να είναι σε θέση να αντιληφθούν τα όρια της προβλεπτικής ακρίβειας ενός μοντέλου, να κατανοούν την αστοχία του εξαιτίας ενός αναπάντεχου γεγονότος και να αξιοποιούν τις προβλέψεις του κατάλληλα με κριτική σκέψη για το σχεδιασμό, τη χάραξη στρατηγικής, τον προγραμματισμό και τη λήψη αποφάσεων στην επιχείρηση ή στον οργανισμό τους. Σε δεύτερη φάση αυτό που πρέπει να απασχολεί τον κλάδο των προβλέψεων είναι η γενικότερη βελτίωση της απόδοσης της διαδικασίας πρόβλεψης. Όταν μία επιχείρηση καλείται να προβλέψει καθημερινά χιλιάδες χρονοσειρές γίνεται αντιληπτό ότι η διαδικασία της πρόβλεψης θα πρέπει να είναι όσο τον δυνατόν πιο αποδοτική εφαρμόζοντας κατάλληλα μοντέλα πρόβλεψης γενικής φύσης και εξάγοντας ακριβείς και αμερόληπτες προβλέψεις σε αποδεκτό χρόνο. Θα πρέπει δηλαδή όπως αναφέρει ο *Steve Jobs* η όλη διαδικασία πρόβλεψης να γίνεται 'thinner, lighter & faster'. Σε οποιαδήποτε άλλη περίπτωση οι προβλέψεις παύουν να εξυπηρετούν τον άμεσα ενδιαφερόμενο και χάνουν την πρακτική τους αξία. Στην παρούσα διατριβή ασχολούμαστε με το δεύτερο στόχο, την ανάπτυξη δηλαδή αποδοτικότερων λύσεων για την παραγωγή προβλέψεων σε μαζική κλίμακα.

1.2 Αντικείμενο και στόχος της διατριβής

Όπως αναφέρθηκε οι προβλέψεις βρίσκονται είτε το αντιλαμβανόμαστε είτε όχι στην καθημερινότητάς μας ενώ αποτελούν αναπόσπαστο κομμάτι του στρατηγικού σχεδιασμού, του προγραμματισμού και της οργάνωσης των επιχειρήσεων και των οργανισμών. Ωστόσο, προκειμένου οι προβλέψεις να αποκτήσουν πραγματική αξία, θα πρέπει αυτές να εξάγονται μέσω μίας αποδοτικής διαδικασίας, μίας γενικευμένης δηλαδή μεθοδολογίας που παράγει σε αποδεκτό για την επιχείρηση χρόνο αξιόπιστες προβλέψεις υψηλής ακρίβειας. Βασική αιτία αποτελεί ο μεγάλος όγκος δεδομένων που συλλέγεται καθημερινά από τις επιχειρήσεις και την οποία καλούνται να αξιοποιήσουν στο μέγιστο βαθμό προκειμένου να βελτιώσουν τις διαδικασίες τους και να μεγιστοποιήσουν τα κέρδη τους. Αντικείμενο και στόχος της παρούσας διατριβής είναι λοιπόν η σχεδίαση και η ανάπτυξη μίας τέτοιας μεθοδολογίας αξιοποιώντας κατάλληλα τα εργαλεία προβλέψεων που υπάρχουν ήδη διαθέσιμα και προτείνοντας νέα ή βελτιώσεις τους όπου αυτό απαιτείται.

Στην κατεύθυνση δημιουργίας μίας αυτοματοποιημένης και γενικευμένης μεθοδολογίας παραγωγής προβλέψεων κρίνεται σκόπιμο λοιπόν να εγκαταλειφθούν σε πρώτη φάση προσεγγίσεις κατά τις οποίες οι χρονοσειρές αναλύονται και επεξεργάζονται μεμονωμένα. Αν και οι εν λόγω πρακτικές οδηγούν συνήθως στην επιλογή και την εφαρμογή του καταλληλότερου μοντέλου πρόβλεψης ικανού να παράγει ακριβείς προβλέψεις, η όλη διαδικασία είναι ιδιαίτερα χρονοβόρα και συχνά απαιτεί για την ολοκλήρωσή της την εμπλοκή του ανθρώπινου παράγοντα. Χαρακτηριστικά παραδείγματα αποτελούν η επιλεκτική εφαρμογή μεθόδων προετοιμασίας χρονοσειρών και η αναθεώρηση των προβλέψεων που προέρχονται από στατιστικά μοντέλα μέσω κριτικών μεθόδων προκειμένου να ληφθούν υπόψη πρόσθετες μεταβλητές και η ανθρώπινη εμπειρία. Δεδομένης μάλιστα της αβεβαιότητας στις προβλέψεις, η οποία θα μπορούσε να οδηγήσει σε άστοχες προβλέψεις ακόμα και κατά την εφαρμογή μίας ιδιαίτερα λεπτομερούς ανάλυσης, η επένδυση σημαντικών ανθρώπινων και υπολογιστικών πόρων στη διαδικασία της πρόβλεψης είναι μάλλον ανούσια.

Σε μία πλήρως αυτοματοποιημένη διαδικασία πρόβλεψης δύο βασικά εμπόδια που καλείται κανείς να ξεπεράσει είναι αυτά της επιλογής του βέλτιστου μοντέλου πρόβλεψης (*model selection*) και της αποτελεσματικής παραμετροποίησής του (*model estimation*). Οι επιλογές σε μοντέλα πρόβλεψης και μεθόδους για την εκτίμησή τους είναι αμέτρητες καθιστώντας τον έλεγχο της καταλληλότητάς τους πρακτικά αδύνατο, ακόμα και στην περίπτωση που η βέλτιστη τεχνική για την αξιολόγηση της απόδοσής τους έχει καθορισθεί με επιτυχία. Αρκεί να σημειωθεί ότι από τα μέσα της δεκαετίας του '70 οι ερευνητές κατασκευάζουν διαρκώς νέα μοντέλα πρόβλεψης και μεθόδους για την βελτίωσή τους και εντοπίζουν καινούργιες τεχνικές για την επιλογή του πλέον αποδοτικού ανά περίπτωση βάσει διαφόρων κριτηρίων.

Ανάλογα με τα χαρακτηριστικά της κάθε χρονοσειράς (τυχειότητα, εποχιακότητα, τάση, στασιμότητα, κανονικότητα κτλ.), τον τύπο των δεδομένων (μικροοικονομικά, μακροοικονομικά, ενεργειακά, βιομηχανικά, δημογραφικά κτλ.), τη συχνότητα δειγματοληψίας (ετήσια, τριμηνιαία, μηνιαία, εβδομαδιαία, ωριαία, ανά λεπτό κτλ.), το μήκος της (πλήθος παρατηρήσεων σε σχέση με τον ορίζοντα πρόβλεψης) και την εφαρμογή στην οποία θα αξιοποιηθούν οι προβλέψεις, μπορούν να προταθούν λοιπόν διαφορετικά μοντέλα πρόβλεψης άγνωστης εξ αρχής καταλληλότητας. Φυσικά κανένα σύστημα πρόβλεψης δεν είναι δυνατό ούτε να διαθέτει όλα τα πιθανά μοντέλα αλλά ούτε και να τα εκτιμάει σε ικανοποιητικό χρόνο κάθε φορά που απαιτείται πρόβλεψη. Επιπλέον, κάποια μοντέλα είναι εξαιρετικά πολύπλοκα ή εξειδικευμένα για συγκεκριμένες εφαρμογές και συνεπώς θα ήταν ανούσιο να ελέγχονται συστηματικά.

Ανατρέχοντας κανείς στη βιβλιογραφία μπορεί να περιορίσει το πρόβλημα εντοπισμού κατάλληλης μεθόδου πρόβλεψης που σκιαγραφήθηκε παραπάνω χρησιμοποιώντας μοντέλα πρόβλεψης χρονοσειρών γενικής εφαρμογής, όπως είναι τα μοντέλα εκθετικής εξομάλυνσης, τα μοντέλα *ARIMA* και η μέθοδος θ . Τα εν λόγω μοντέλα βασίζονται αποκλειστικά στα ιστορικά δεδομένα του μεγέθους για το οποίο επιθυμείται πρόβλεψη η οποία και εξάγεται εκτιμώντας αρχικά τη σχέση που συνδέει τα ιστορικά δεδομένα μεταξύ τους και αξιοποιώντας τη στη συνέχεια κατάλληλα για να επεκταθούν στο μέλλον. Οι

οικογένειες μοντέλων εκθετικής εξομάλυνσης και *ARIMA*, αν και αποτελούν προϊόντα της δεκαετίας του '70, συνεχίζουν να χρησιμοποιούνται ευρέως έως τις μέρες μας για την παραγωγή προβλέψεων σε χιλιάδες επιχειρήσεις. Αυτό συμβαίνει καθώς, ειδικά για μικρούς ορίζοντες πρόβλεψης, εμφανίζουν ικανοποιητικά επίπεδα ακρίβειας ανεξαρτήτως των χαρακτηριστικών της χρονοσειράς, ενώ δεν συνεπάγονται υψηλό υπολογιστικό κόστος για την επιχείρηση. Έτσι, τα εν λόγω μοντέλα συνεχίζουν να αποτελούν αντικείμενο μελέτης για τους ερευνητές των προβλέψεων με χαρακτηριστικό παράδειγμα την πρόσφατη πρόταση ενός νέου αλγορίθμου ικανού να επιλέγει το επικρατέστερο μοντέλο ανά χρονοσειρά και να το εκτιμά αποδοτικά.

Ωστόσο, παρά τη διαρκή εξέλιξή τους, η μέθοδος θ συνεχίζει σήμερα έπειτα από σχεδόν 20 χρόνια από την εμφάνισή της να αποτελεί την ακριβέστερη επιλογή όταν αναφερόμαστε σε πρόβλεψη κατά μήκος μεγάλου όγκου χρονοσειρών. Η μέθοδος έγινε γνωστή κυρίως λόγω της απόδοσής της στον διαγωνισμό *M3*, τον μεγαλύτερο διαγωνισμό προβλέψεων που διοργανώθηκε ποτέ, όπου και αναδείχτηκε νικήτρια έναντι σαφώς πιο πολύπλοκων μεθόδων και ολοκληρωμένων συστημάτων πρόβλεψης. Αυτό που προκαλεί εντύπωση είναι ότι, αν και από τότε έχουν γίνει κάποιες προσπάθειες για τη βελτίωση της απόδοσης της μεθόδου και την εξακρίβωση της απόδοσής της σε ποικίλες εφαρμογές και σετ δεδομένων με ενθαρρυντικά αποτελέσματα, λίγες είναι εκείνες οι μελέτες που συνεισφέρουν πραγματικά στην εξέλιξή της.

Πρώτος στόχος της παρούσας διατριβής είναι λοιπόν να εντοπιστούν οι παράγοντες που καθιστούν τη μέθοδο θ τόσο πετυχημένη και στη συνέχεια να αξιοποιηθούν στο μέγιστο βαθμό ούτως ώστε να προκύψει μία νέα πιο ευέλικτη μέθοδος με βελτιωμένα χαρακτηριστικά και δυνατότητες. Αυτή θα μπορεί να αξιοποιηθεί στη συνέχεια σε μία γενικευμένη μεθοδολογία πρόβλεψης παράγοντας αξιόπιστες και στιβαρές προβλέψεις.

Όσον αφορά τώρα την εκτίμηση του μοντέλου πρόβλεψης, αρκετές είναι οι μέθοδοι οι οποίες μπορούν να χρησιμοποιηθούν για την βελτίωση του υπολογισμού της τιμής των παραμέτρων του. Κατά κανόνα, προκειμένου να υπολογίσουμε τις παραμέτρους ενός μοντέλου θέτουμε αρχικά ένα κριτήριο βελτιστοποίησης, όπως η ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος ή η μεγιστοποίηση της πιθανοφάνειας, και στη συνέχεια εντοπίζουμε εκείνο το συνδυασμό παραμέτρων που τον ικανοποιούν. Βρίσκουμε δηλαδή εκείνες τις τιμές παραμέτρων για τις οποίες επιτυγχάνεται η καλύτερη δυνατή προσαρμογή του μοντέλου στα δοθέντα δεδομένα. Βασικό πρόβλημα σε αυτή τη διαδικασία είναι ότι πιθανές σποραδικές ανομοιομορφίες στη χρονοσειρά (π.χ. κάποια ακραία τιμή) θα επηρεάσουν σημαντικά την παραμετροποίηση καθώς το μοντέλο θα προσπαθήσει να προσαρμοστεί σε ένα λανθάνον πρότυπο που δεν συνάδει αναγκαστικά με την πραγματικότητα. Οι ανομοιομορφίες θα οδηγήσουν λοιπόν σε ανεπαρκή περιγραφή των μοτίβων της χρονοσειράς και πιθανώς σε ανακριβείς προβλέψεις. Παρόμοια αποτελέσματα ενδέχεται να υπάρξουν σε περιπτώσεις χρονοσειρών με αυξημένη την παρουσία θορύβου.

Προκειμένου να περιοριστεί η επίδραση τέτοιων φαινομένων έχουν προταθεί αρκετές τεχνικές προ-επεξεργασίας χρονοσειρών. Χαρακτηριστικά παραδείγματα είναι η χρήση

μετασχηματισμών οι οποίοι περιορίζουν τη διακύμανση των δεδομένων και οδηγούν σε κανονικότητα, αλλά και οι μέθοδοι εξομάλυνσης οι οποίες περιορίζουν δραστικά το θόρυβο που εμπεριέχεται στις χρονοσειρές. Άλλες τεχνικές που περιορίζουν έμμεσα την επίδραση της λανθασμένης παραμετροποίησης είναι η χρήση μεθόδων συνάθροισης δεδομένων σε διαφορετικές συχνότητες (*temporal aggregation*) καθώς και ο συνδυασμός προβλέψεων (*forecast combination*). Στην πρώτη περίπτωση τα δεδομένα μίας υψηλής συχνότητας (π.χ. μηνιαία) αθροίζονται σε μία χαμηλότερης (π.χ. ετήσια) και πραγματοποιείται πρόβλεψη επί αυτών με την πεποίθηση ότι, αφού στη νέα συχνότητα η διακύμανση είναι μικρότερη, η παραμετροποίηση του αντίστοιχου μοντέλου θα γίνει με μεγαλύτερη επιτυχία. Στη δεύτερη περίπτωση υποθέτουμε ότι μεμονωμένα κάποια μοντέλα πρόβλεψης είναι πιθανό να επηρεαστούν σημαντικά από ανομοιομορφίες στη χρονοσειρά, ωστόσο αυτό δεν δύναται να συμβεί για όλα, δηλαδή στη μέση περίπτωση. Έτσι, αν συνδυάσουμε κατάλληλα τις προβλέψεις πολλών διαφορετικών μοντέλων η τελική μέση πρόβλεψη αναμένεται να επηρεαστεί λιγότερο από τα μειονεκτήματα των επιμέρους μεθόδων. Μάλιστα, συχνά πραγματοποιείται συνδυασμός των παραπάνω τεχνικών (συνδυασμός προβλέψεων από όλες τις δυνατές συχνότητες) με αντίστοιχα πλεονεκτήματα, γνωστό και ως *multiple temporal aggregation*.

Ένας επιπρόσθετος στόχος λοιπόν της παρούσας διδακτορικής διατριβής είναι η εύρεση της πλέον συμφέρουσας τεχνικής για την αποδοτική επεξεργασία μεγάλου πλήθους χρονοσειρών πριν την πρόβλεψη, καθώς και η βελτίωσή της προκειμένου να εξασφαλιστεί μια πιο ικανοποιητική παραμετροποίηση. Η εν λόγω μελέτη περιλαμβάνει λοιπόν σε πρώτη φάση τη βιβλιογραφική επισκόπηση των εν λόγω τεχνικών προετοιμασίας χρονοσειρών, τον εντοπισμό των πλεονεκτημάτων και των μειονεκτημάτων τους και τον καθορισμό της καταλληλότερης. Στη συνέχεια, έχοντας εξετάζοντας τα χαρακτηριστικά των μεθόδων που έχουν προταθεί για την επιλεγθείσα τεχνική, θα προταθεί μία νέα βελτιωμένη προσέγγιση, ικανή να προσδώσει υψηλότερη προβλεπτική ακρίβεια.

Φυσικά, προκειμένου η όλη διαδικασία πρόβλεψης να καταστεί αποδοτική, ενδέχεται να να συμπεριληφθούν σε αυτή και άλλες μέθοδοι απαραίτητες για την προετοιμασία της χρονοσειράς και τη βελτίωση της προβλεπτικής ικανότητας του μοντέλου πρόβλεψης. Χαρακτηριστικό παράδειγμα αποτελεί η αποεποχικοποίηση των χρονοσειρών. Για πολλά χρόνια κάποιες από τις μεθόδους αποεποχικοποίησης κυριαρχούν στη βιβλιογραφία παρόλο που υπάρχουν μελέτες που αποδεικνύουν ότι εναλλακτικές, και μάλιστα σχετικά απλές μέθοδοι, μπορούν να οδηγήσουν σε καλύτερα αποτελέσματα. Η ενσωμάτωση τέτοιων εργαλείων σε μία ολοκληρωμένη μεθοδολογία πρόβλεψης είναι καθοριστικής σημασίας καθώς ενδέχεται να επιφέρει σημαντικές βελτιώσεις στην απόδοσή της με αμελητέο επιπλέον υπολογιστικό κόστος.

Ένας άλλος στόχος της παρούσας διατριβής είναι λοιπόν ο συνδυασμός των επιμέρους μεθόδων, τεχνικών και μοντέλων που επιλέχθηκαν να χρησιμοποιηθούν στη διαδικασία πρόβλεψης σε ένα ενιαίο και εναρμονισμένο μεθοδολογικό πλαίσιο ούτως ώστε η αλληλεπίδρασή τους να προσδώσει το βέλτιστο προβλεπτικά αποτέλεσμα. Προκειμένου μάλιστα να εξακριβωθεί η αποτελεσματικότητα της μεθοδολογίας θα αξιολογηθεί η απόδοσή της τόσο σε πραγματικά όσο και σε τεχνικά δεδομένα. Στην πρώτη περίπτωση

θα αξιοποιηθούν οι χρονοσειρές από γνωστούς διαγωνισμούς προβλέψεων οι οποίοι και αποτελούν σημείο αναφοράς για τον κλάδο, ενώ στη δεύτερη θα κατασκευαστούν αλγόριθμοι για την παραγωγή νέων με τεχνητό τρόπο. Αυτό μπορεί να γίνει είτε μέσω αναδειγματοληψίας, της παραγωγής δηλαδή πολλαπλών παρεμφερών αντιτύπων από μία δοσμένη χρονοσειρά, είτε μέσω μίας γεννήτρια χρονοσειρών, την παραγωγής δηλαδή τυχαίων χρονοσειρών με ποικίλα χαρακτηριστικά διαφορετικής έντασης χωρίς τη χρήση κάποιας αναφοράς. Η δεύτερη προσέγγιση είναι αυτή που θα μας απασχολήσει για λόγους που θα παρουσιαστούν σε αντίστοιχο κεφάλαιο.

1.3 Συμβολή της διατριβής

Η συμβολή της παρούσας διατριβής αφορά το ευρύτερο επιστημονικό πεδίο της επιχειρησιακής έρευνας και συγκεκριμένα αυτό της αποδοτικότερης παραγωγής προβλέψεων για επιχειρήσεις και οργανισμούς. Δεδομένης της αβεβαιότητας του μέλλοντος και της σημασίας των προβλέψεων στις καθημερινές αποφάσεις, τον προγραμματισμό και το σχεδιασμό μίας επιχείρησης, το εν λόγω αντικείμενο μελέτης χρίζει ιδιαίτερου ενδιαφέροντος. Τέσσερεις είναι οι άξονες καινοτομίας της εν λόγω διατριβής:

- Στα πλαίσια εύρεσης μίας γενικευμένης μεθοδολογίας για την πρόβλεψη πληθώρας χρονοσειρών διαφορετικών τύπων και χαρακτηριστικών, εντοπίστηκαν οι παράγοντες που καθιστούν τη μέθοδο θ τόσο επιτυχημένη καθώς και τα σημεία στα οποία αυτή επιδέχεται βελτιώσεις. Τα συμπεράσματα της παραπάνω ανάλυσης αξιοποιήθηκαν για τη δημιουργία μίας νέας μεθόδου η οποία, σε αντίθεση με την κλασική μέθοδο θ , είναι ικανή να μοντελοποιεί αποδοτικά μη γραμμικά μοτίβα τάσης όπως π.χ. εκθετική ή λογαριθμική τάση. Έτσι, αν και η νέα μέθοδος συνεχίζει να βασίζεται στη χρήση πολλαπλών γραμμών θ διαφορετικής καμπυλότητας για την εκτίμηση των μακροπρόθεσμων και βραχυπρόθεσμων χαρακτηριστικών της χρονοσειράς, αυτές υπολογίζονται λαμβάνοντας υπόψη είτε γραμμικές είτε μη γραμμικές μεταβολές στο επίπεδό της. Επιπλέον, η κλασική προσθετική μέθοδος θ εκφράστηκε μέσω μίας πολλαπλασιαστικής σχέσης με στόχο οι συνιστώσες τάσης και επιπέδου που αυτή εκτιμά να μπορούν να αλληλοεπιδρούν με αντίστοιχο τρόπο. Οι εν λόγω επεκτάσεις οδήγησαν ουσιαστικά στη δημιουργία πολλών μοντέλων θ διαφορετικών ιδιοτήτων και αποτελούν κατά συνέπεια μία γενίκευση της κλασικής μεθόδου, αναδεικνύοντας πώς μπορεί κανείς να αντλήσει από μία χρονοσειρά μεγάλο όγκο πληροφορίας αποσυνθέτοντάς τη σε κατάλληλες γραμμές θ για την παραγωγή αξιόπιστων και στιβαρών προβλέψεων.
- Αναπτύχθηκε μία νέα τεχνική εξομάλυνσης η οποία περιορίζει τις τοπικές διακυμάνσεις και εξαλείφει το θόρυβο που εμπεριέχεται στα δεδομένα αποκαλύπτοντας έτσι το πραγματικό σήμα της χρονοσειράς. Η εν λόγω μέθοδος μπορεί να φανεί ιδιαίτερα χρήσιμη καθώς, σε αντίθεση με μεθόδους εξομάλυνσης που εντοπίζονται στη βιβλιογραφία, αξιοποιεί μία μη γραμμική τεχνική για τη συρρίκνωση των τοπικών διακυμάνσεων οδηγώντας σε σχεδόν απόλυτα λείες χρονοσειρές η πρόβλεψη επί των οποίων γίνεται σαφώς πιο αποτελεσματικά από οποιοδήποτε μοντέλο πρόβλεψης. Βασικό πλεονέκτημα της μεθόδου είναι ότι δεν

χρησιμοποιεί όπως είθισται κάποιο σταθμισμένο μέσο όρο επί των δεδομένων για την επίτευξη της εξομάλυνσης. Έτσι, τυχόν ακραίες τιμές απομακρύνονται πλήρως από τα δεδομένα και δεν εμπεριέχονται στη νέα χρονοσειρά που προκύπτει σε εξασθενημένη μορφή. Αυτό έχει ως επιπλέον πλεονέκτημα τη μη δημιουργία κενών τιμών στην αρχή και στο τέλος της χρονοσειράς ανάλογα με τον βαθμό εξομάλυνσης, όπως συμβαίνει π.χ. στη περίπτωση των κινητών μέσων όρων και των μεθόδων kernel, φαινόμενο το οποίο απαιτεί τη μετέπειτα χρήση τεχνικών συμπλήρωσης δεδομένων με αμφίβολα αποτελέσματα. Μία τέτοια αποτελεσματική τεχνική εξομάλυνσης μπορεί να προσδώσει αξία στη διαδικασία της πρόβλεψης διευκολύνοντας την προσαρμογή των μοντέλων στη χρονοσειρά και μειώνοντας την αβεβαιότητα κατά την παραμετροποίηση.

- Η εν λόγω διατριβή οδήγησε στο σχεδιασμό και την ανάπτυξη μίας ενιαίας μεθοδολογίας η οποία συνδυάζει αποτελεσματικά υπάρχουσες και νέες τεχνικές για την προετοιμασία και την πρόβλεψη χρονοσειρών. Η εν λόγω μεθοδολογία, η οποία έχει γενική χρήση και μπορεί να χρησιμοποιηθεί για την αποδοτική πρόβλεψη μεγάλου πλήθους χρονοσειρών σε επιχειρήσεις, εξασφαλίζει βελτιωμένη ακρίβεια με ικανοποιητικό υπολογιστικό κόστος χωρίς να απαιτείται η εμπλοκή του ανθρώπινου παράγοντα. Η χρήση μίας τέτοιας μεθοδολογίας έχει ως αποτέλεσμα λοιπόν την πλήρη αυτοματοποίηση της διαδικασίας της πρόβλεψης αξιοποιώντας κατάλληλα τις βέλτιστες πρακτικές που η επιστήμη των προβλέψεων έχει αναδείξει. Μάλιστα, προκειμένου η πρόβλεψη να γίνεται με αυτόματο τρόπο, αναπτύχθηκε επιπλέον μία τεχνική επιλογής του καταλληλότερου μοντέλου πρόβλεψης θ εξ αυτών που κατασκευάστηκαν νωρίτερα, η οποία αξιοποιεί το σφάλμα προσαρμογής του εκάστοτε μοντέλου για να επιλέξει το καταλληλότερο και να οδηγήσει μέσω απλών κριτηρίων στο καλύτερο δυνατό αποτέλεσμα.
- Η αποτελεσματικότητα της προτεινόμενης μεθοδολογίας εξακριβώθηκε χρησιμοποιώντας σετ από πραγματικά και τεχνικά δεδομένα. Στην πρώτη περίπτωση χρησιμοποιήθηκαν χρονοσειρές από διαγωνισμούς προβλέψεων, ενώ στη δεύτερη αναπτύχθηκε κατάλληλη γεννήτρια για την παραγωγή νέων ποικίλων ιδιοτήτων. Και στις δύο περιπτώσεις εξάχθηκαν χρήσιμα συμπεράσματα για τον τρόπο με τον οποίο συμπεριφέρεται η μέθοδος θ ανάλογα με τα χαρακτηριστικά της εκάστοτε χρονοσειράς και της έντασής τους, τα οποία και μπορούν να αξιοποιηθούν σε μελλοντικές μελέτες για την περαιτέρω βελτίωσή της. Επίσης, η γεννήτρια μπορεί να χρησιμοποιηθεί για την αξιολόγηση οποιαδήποτε άλλης μεθόδου σε χρονοσειρές προκαθορισμένου πλήθους και χαρακτηριστικών.

1.4 Δομή της διατριβής

Η παρούσα διατριβή οργανώνεται σε επτά κεφάλαια κατά μήκος των οποίων παρουσιάζονται οι τεχνικές προετοιμασίας και πρόβλεψης χρονοσειρών που αξιοποιήθηκαν από τη βιβλιογραφία ή αναπτύχθηκαν στα πλαίσια της εργασίας, οι μέθοδοι αξιολόγησης προβλέψεων, οι τεχνικές δημιουργίας τεχνητών χρονοσειρών, η

σχεδίαση της προτεινόμενης μεθοδολογίας, τα πειραματικά αποτελέσματα, τα συμπεράσματα και οι προεκτάσεις της.

Πιο αναλυτικά, στο *Κεφάλαιο 2* πραγματοποιείται μία διεξοδική περιγραφή των ποιοτικών χαρακτηριστικών των χρονοσειρών και παρουσιάζονται μέθοδοι αποσύνθεσής τους σε αυτά. Δίνεται έμφαση στην κλασική μέθοδο αποσύνθεσης, η οποία και θα χρησιμοποιηθεί αργότερα στην προτεινόμενη μεθοδολογία, ενώ παρατίθενται τεχνικές συρρίκνωσης εποχιακών δεικτών για τον καλύτερο προσδιορισμό τους και την έμμεση βελτίωση της ακρίβειας των μοντέλων πρόβλεψης. Στη συνέχεια αναλύεται η σημασία των στατιστικών ελέγχων για την εξακρίβωση στασιμότητας και κανονικότητας στις χρονοσειρές και παρουσιάζονται λύσεις, όπως είναι η εφαρμογή μετασχηματισμών, για την επίτευξή τους. Τέλος παρουσιάζονται διάφοροι δείκτες μέτρησης της απόδοσης των μοντέλων πρόβλεψης σε σχέση με την ακρίβεια και την προκατάληψή τους και καθορίζονται τεχνικές για την αξιολόγησή τους.

Στο τρίτο κεφάλαιο της εργασίας παρουσιάζονται οι μέθοδοι προβλέψεων και συγκεκριμένα τα στατιστικά μοντέλα πρόβλεψης χρονοσειρών. Αφού γίνει μία σύντομη περιγραφή των βασικότερων μοντέλων και κυρίως αυτών που θα αξιοποιηθούν στην παρούσα εργασία, παρουσιάζεται διεξοδικά το μοντέλο πρόβλεψης θ και αναλύεται η όλη λογική και συμπεριφορά του. Αφού αναγνωριστούν τα πλεονεκτήματά του παρουσιάζονται επεκτάσεις του προκειμένου να γενικευτεί η χρήση του για αποτελεσματικότερη πρόβλεψη χρονοσειρών με μη γραμμικά μοτίβα τάσης.

Στο τέταρτο κεφάλαιο παρουσιάζονται διεξοδικά οι βασικότερες τεχνικές εξομάλυνσης χρονοσειρών και αναδεικνύονται μελανά τους σημεία τα οποία επιδέχονται βελτιώσεων. Βάσει αυτών αναπτύσσεται μία νέα μη γραμμική μέθοδος εξομάλυνσης η οποία και θα χρησιμοποιηθεί για την καλύτερη παραμετροποίηση των μοντέλων πρόβλεψης και τη βελτίωση της προβλεπτικής τους ακρίβειας.

Στο *Κεφάλαιο 5* παρουσιάζονται τα αποτελέσματα από τους βασικότερους διαγωνισμούς προβλέψεων που έχουν διεξαχθεί μέχρι σήμερα και αναδεικνύεται η ανάγκη για παραγωγή τεχνητών χρονοσειρών προκειμένου να ελεγχθεί πιο αντικειμενικά η απόδοση των διαφόρων μοντέλων πρόβλεψης βάσει ιδιοτήτων δεδομένων. Σε αυτή την κατεύθυνση αναπτύσσεται μία γεννήτρια τεχνητών χρονοσειρών η οποία είναι σε θέση να παράγει μεγάλο πλήθος χρονοσειρών με χαρακτηριστικά (τάση, εποχιακότητα και θόρυβος) ποικίλης έντασης.

Η μεθοδολογία η οποία προτείνεται σχεδιάζεται και περιγράφεται αναλυτικά στο έκτο κεφάλαιο της εργασίας όπου και αξιολογείται κατά μήκος των σετ δεδομένων που ορίστηκαν στο προηγούμενο κεφάλαιο. Πραγματοποιούνται συγκρίσεις της μεθοδολογίας με άλλες επιλογές και εναλλακτικές τεχνικές ανάλυσης και πρόβλεψης χρονοσειρών και αναδεικνύεται η καταλληλότητά της για την αυτόματη πρόβλεψη μεγάλου όγκου χρονοσειρών μέσω μίας ενιαίας και γενικευμένης διαδικασίας. Τα συμπεράσματα και οι προεκτάσεις της εργασίας παρατίθενται στο *Κεφάλαιο 7* ύστερα από διεξοδική ανάλυση των πειραματικών αποτελεσμάτων.

1.5 Ερευνητικά αποτελέσματα της διατριβής

Η διατριβή ολοκληρώθηκε μέσα σε περίπου 3.5 χρόνια (Φεβρουάριος 2014-Σεπτέμβριος 2017) και οδήγησε σε:

- ✓ **5 δημοσιεύσεις σε διεθνή επιστημονικά περιοδικά**
 - Spiliotis, E., Assimakopoulos, V., Nikolopoulos, K. "Thinner, lighter, faster..." and the way forward for time series extrapolation methods: superimposing smoothing, a variation of the Theta method and shrinkage of seasonal factors, *International Journal of Production Economics - accepted*
 - Sarris, D., Spiliotis, E., Assimakopoulos, V. Exploiting resampling techniques for model selection in forecasting: An empirical evaluation using out-of-sample tests, *Operational Research: An International Journal - in print*
 - Marinakis, V., Doukas, H., Spiliotis, E., Papastamatiou, I., 2017. Decision Support for Intelligent Energy Management in Buildings Using the Thermal Comfort Model, *International Journal of Computational Intelligence Systems*, 10 (1), 882–893
 - Papastamatiou, I., Doukas, H., Spiliotis, E., Psarras, J., 2016. How “OPTIMUS” is a city in terms of energy optimization? e-SCEAF: A web based decision support tool for local authorities, *Information Fusion*, 29, 149-161
 - Spiliotis E., Raptis A., Legaki Z.N., Assimakopoulos V., 2014. Forecasting electrical consumption of commercial buildings using energy performance indicators. *International Journal of Decision Support Systems*, 1(2), 164–182
- ✓ **4 δημοσιεύσεις σε πρακτικά διεθνών συνεδρίων**
 - Spiliotis, E., Marinakis, V., Doukas, H., Psarras, J., 2016. Integrating a Decision Support System with smart grid infrastructures and ICT solutions towards energy cost reduction. 7th International Conference on Information, Intelligence, Systems and Applications (IISA 2016), 1-6
 - Doukas, H., Marinakis, V., Spiliotis, E., Psarras, J., 2016. OPTIMUS Decision Support Tools: Transforming Multidisciplinary Data to Energy Management Action Plans. 7th International Conference on Information, Intelligence, Systems and Applications (IISA 2016), 1-5
 - Spiliotis, E., Anastasopoulos, G., Dede, P., Marinakis, V. Doukas, H, 2015. A framework for integrating user experience in action plan evaluation through social media: Transforming user generated content into knowledge to optimise energy use in buildings. 6th International Conference on Information, Intelligence, Systems and Applications (IISA 2015), 1-6
 - Androulaki, S., Spiliotis, E., Doukas, H., Papastamatiou, I., Psarras, J., 2014. Proposing a Smart City Energy Assessment Framework: Linking local vision with data sets. 5th International Conference on Information, Intelligence, Systems and Applications (IISA 2014), 50-56

✓ **12 ανακοινώσεις σε διεθνή επιστημονικά συνέδρια**

- Spiliotis, E., Patikos, A., Assimakopoulos V., Kouloumos, A., 2017. Data as a service: Providing new datasets to the forecasting community for time series analysis. 37th International Symposium on Forecasting (ISF 2017), June 25 – 28, 2017, Cairns, Australia
- Skepetari, I., Spiliotis, E., Raptis, A., Assimakopoulos, V., 2017. Assisting judgmental adjustments in hierarchical forecasts through interactive forecasting support systems. 37th International Symposium on Forecasting (ISF 2017), June 25 – 28, 2017, Cairns, Australia
- Spiliotis, E., Assimakopoulos, V., Nikolopoulos, K., 2016. "Thinner, lighter, faster..." and the way forward for time series extrapolation methods. 19th International Symposium on Inventories, August 22-26, 2016, Hilton Budapest Hotel, Budapest, Hungary
- Spiliotis, E., Assimakopoulos V., 2016. Transforming the theta model into a flexible decomposition method by considering nonlinear trends. 36th International Symposium on Forecasting (ISF 2016), June 19 – 22, 2016, Santander, Spain
- Skepetari, I., Spiliotis, E., Raptis, A., Assimakopoulos V., 2016. OMEN: Promoting forecasting support systems. 36th International Symposium on Forecasting (ISF 2016), June 19 – 22, 2016, Santander, Spain
- Spiliotis, E., Marinakis, V., Tsapelas, G., Skondrogianni, C., Doukas, H., 2016. Reducing Energy and CO2 Emissions at the City Premises: A Decision Support System for Local Authorities. 5th International Symposium and 27th National Conference on Operational Research (HELORS 2016), June 9-11, 2016, Aigaleo, Athens
- Spiliotis, E., Raptis, A., Skepetari, E., Assimakopoulos, V., 2015. Exploiting business intelligence of water companies. ForWarD: an online water demand forecasting tool. 35th International Symposium on Forecasting (ISF 2015), June 21 – 24, 2015, Riverside, California, USA
- Spiliotis, E., Petropoulos, F., Kourentzes, N., Assimakopoulos, V., 2014. Examining the effect of temporal aggregation on forecasting accuracy for hierarchical energy consumption time series. 34th International Symposium on Forecasting (ISF 2014), June 29 – July 2, 2014, Rotterdam, The Netherlands
- Paterakis, S., Spiliotis, E., Assimakopoulos, V., 2014. Forecasting Electrical Demand in Commercial Buildings through Energy Performance Indicators using Time Series Methods. 20th Conference of the International Federation of Operational Research Societies EURO (IFORS), July 13-18, 2014, Barcelona, Spain
- Spiliotis, E., Raptis, A., Legaki, N. Z., Assimakopoulos, V., 2013. A method for predicting electrical consumption in energy-intensive buildings through energy performance indicators. 2nd International Symposium and 24th National Conference on Operational Research (HELORS 2013), 26 - 28 September, 2013, Athens, Greece

- Gkolemis, N., Legaki, N.Z., Assimakopoulos, V., Spiliotis, E., 2013. Modeling electrical energy operations of buildings to forecast and achieve energy efficiency. 26th European Conference on Operational Research EURO (INFORMS) MMXIII, July 1-4, 2013, Rome, Italy
- Spiliotis, E., Raptis, A., Legaki, N.Z., Assimakopoulos, V., 2013. Forecasting energy consumption of buildings using performance indicators. 26th European Conference on Operational Research EURO (INFORMS) MMXIII, July 1-4, 2013, Rome, Italy
- ✓ **1 δημοσίευση σε κεφάλαιο βιβλίου**
 - Androutaki, S., Doukas, H., Spiliotis, E., Papastamatiou, I., Psarras, J., 2016. A framework to assess the behavior and performance of a city towards energy optimization. Chapter in book: Intelligent Computing Systems, Edition: 2016, Series Volume 627, Chapter: 9, Publisher: Springer-Verlag Berlin Heidelberg, Editors: George A. Tsihrintzis, Maria Virvou, Lakhmi C. Jain, DOI: 10.1007/978-3-662-49179-9
- ✓ **1 δημοσίευση σε κλαδικά περιοδικά**
 - Spiliotis, E., Raptis, A., Assimakopoulos V., 2016. Off-the-Shelf vs. Customized Forecasting Support Systems. Foresight: The International Journal of Applied Forecasting, Fall 2016 (Issue 43), 42-48
- ✓ **2 ανακοινώσεις σε φοιτητικά συνέδρια**
 - Σίμος, Ν., Σπηλιώτης, Ε., Ασημακόπουλος Β., 2015. Ανάπτυξη Μεθοδολογίας για τη Μείωση των Ουρών Αναμονής στα Υποκαταστήματα Τραπεζής. 4ο Φοιτητικό Συνέδριο Ελληνικής Εταιρίας Επιχειρησιακών Ερευνών - "Επιχειρησιακή Έρευνα - Ευκαιρίες και Προκλήσεις Μέσα στην Κρίση", 17-18 Δεκ. 2015
 - Φακίνος, Γ., Σπηλιώτης, Ε., Ασημακόπουλος Β., 2015. Πρόβλεψη ηλεκτρικής κατανάλωσης σε σύμπλεγμα εμπορικών κτιρίων με χρήση μεθόδων χρονικής συνάθροισης. 4ο Φοιτητικό Συνέδριο Ελληνικής Εταιρίας Επιχειρησιακών Ερευνών - "Επιχειρησιακή Έρευνα - Ευκαιρίες και Προκλήσεις Μέσα στην Κρίση", 17-18 Δεκ. 2015
- ✓ **10 διπλωματικές εργασίες προπτυχιακού επιπέδου**
 - Νταβέλης, Ε., 2017. Βελτίωση ακρίβειας στατιστικών μεθόδων πρόβλεψης σε χρονοσειρές μικρού ιστορικού με χρήση τεχνικών συσταδοποίησης εποχιακών δεικτών συναφών δεδομένων
 - Σίμος, Ν., 2016. Ανάπτυξη μεθοδολογίας για τον Περιορισμό των Ουρών Αναμονής στα Καταστήματα Τραπεζής
 - Σαρρής, Δ., 2016. Έλεγχος ευστάθειας απόδοσης μοντέλων πρόβλεψης με χρήση τεχνικών αναδειγματοληψίας
 - Φακίνος, Γ., 2016. Πρόβλεψη ηλεκτρικής κατανάλωσης σε σύμπλεγμα εμπορικών κτιρίων με χρήση μεθόδων χρονικής συνάθροισης δεδομένων
 - Αλμπανούδης, Ο., 2016. Σύγκριση Απόδοσης Συμβατικών Παλινδρομικών Μοντέλων και Μοντέλων Πρόβλεψης Χρονοσειρών με Χρονική Συνάθροιση σε Δεδομένα Ενεργειακής Κατανάλωσης Σ.Η.Ε.

- Συμβουλάκης, Ν.Ε., 2015. Ανάπτυξη μεθοδολογίας για την αξιολόγηση αποφάσεων επενδύσεων στηριζόμενοι σε τεχνικές Sentiment Analysis και στατιστικές μεθόδους
- Γαρδέρης, Γ., 2015. Μελέτη κριτηρίων λογαρίθμησης προς Βελτίωση της Προβλεπτικής Ικανότητας της Μεθόδου Theta (Θ)
- Κελαϊδίτης, Γ., 2014. Υλοποίηση Διαδικτυακής Πλατφόρμας Διαχείρισης Ενεργειακών Δεδομένων Οικιακών Καταναλωτών
- Μηλάκας Τσαμουράς, Ι., 2014. Σχεδιασμός Πλαισίου Ψηφιακής Ανάπτυξης και Εφαρμογή στην Ελλάδα
- Πατεράκης, Σ., 2014. Πρόβλεψη Ηλεκτρικής Κατανάλωσης Κτιρίων μέσω Ενεργειακών Δεικτών και Μεθόδων Χρονοσειρών
- ✓ **1 διπλωματική εργασία μεταπτυχιακού επιπέδου:**
 - Ροδόπουλος, Δ., 2017. Πρόβλεψη κατανάλωσης νερού για οικιακούς χρήστες με αξιοποίηση δημογραφικών στοιχείων

Η πρόβλεψη είναι πολύ δύσκολη, ιδιαίτερα όταν αφορά το μέλλον

Yogi Berra

Κεφάλαιο 2: Ανάλυση χρονοσειρών και απόδοσης μοντέλων πρόβλεψης

Οι χρονοσειρές, ή χρονολογικές σειρές, αποτελούν ένα σύνολο διαδοχικών παρατηρήσεων της τιμής ενός μεγέθους. Δεδομένου ότι οι επιμέρους παρατηρήσεις είναι συνεχείς και διατεταγμένες στον άξονα του χρόνου, γίνεται αμέσως αντιληπτό ότι παρέχουν σημαντική πληροφορία σχετικά με τα χαρακτηριστικά του εκάστοτε μεγέθους και τον τρόπο με τον οποίο αυτό εξελίσσεται. Έτσι, η ανάλυσή τους, η οποία μπορεί να γίνει μέσω στατιστικών δεικτών και μεθόδων, έχει ιδιαίτερη αξία.

Στην πιο απλή τους περίπτωση οι χρονοσειρές είναι ντετερμινιστικές, δηλαδή η κάθε παρατήρηση εξαρτάται από τις προηγούμενες και παράγεται μέσω μίας συγκεκριμένης σχέσης. Αν αυτή η σχέση προσδιοριστεί ικανοποιητικά, τότε το μέλλον μπορεί να προβλεφθεί με αξιολογη ακρίβεια. Φυσικά, κάτι τέτοιο δεν είναι δυνατό πάντοτε εφικτό σε πραγματικά δεδομένα καθώς, πέραν της σχέσης που ενδέχεται να συνδέει τις επιμέρους παρατηρήσεις μίας χρονοσειράς, ο θόρυβος ή τυχαιότητα την επηρεάζουν σημαντικά. Οι χρονοσειρές και τα μοντέλα που περιέχουν τον στοχαστικό παράγοντα της τυχαιότητας χαρακτηρίζονται ως στοχαστικά.

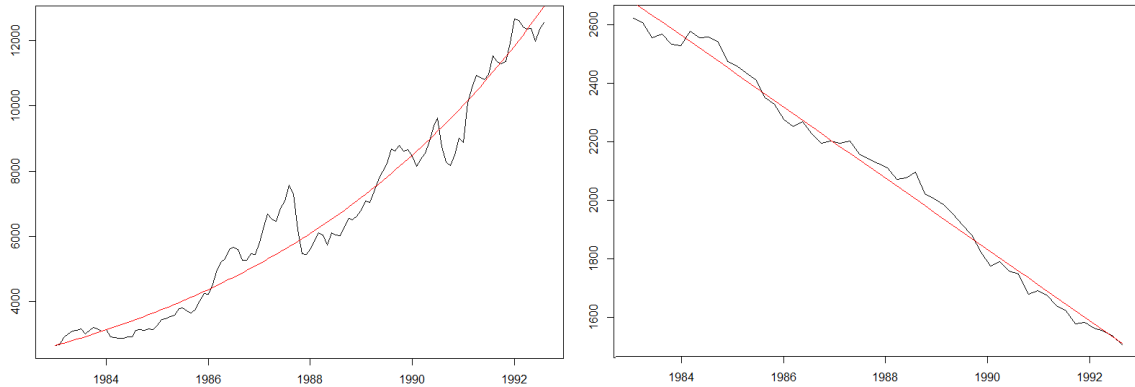
Σε κάθε περίπτωση, ντετερμινιστικής ή στοχαστικής χρονοσειράς, κάποια από τα βασικά χαρακτηριστικά των χρονοσειρών μπορούν να απομονωθούν, να εκτιμηθούν και να μετρηθούν με ικανοποιητική ακρίβεια. Στο παρόν κεφάλαιο παρουσιάζονται λοιπόν τα ποιοτικά χαρακτηριστικά των χρονοσειρών, στατιστικοί δείκτες οι οποίοι είναι απαραίτητοι για την μελέτη τους και τέλος δείκτες σφαλμάτων που μπορούν να αξιοποιηθούν κατά τη διαδικασία της παραμετροποίησης ενός μοντέλου πρόβλεψης και την αξιολόγηση της απόδοσής του.

2.1 Ποιοτικά χαρακτηριστικά χρονοσειρών

Ως ποιοτικά χαρακτηριστικά χρονοσειρών αναφέρουμε εκείνα τα μεγέθη τα οποία ένας παρατηρητής μπορεί να εντοπίσει κατά την οπτική μελέτη μίας χρονοσειράς. Στην πραγματικότητα τα εν λόγω μεγέθη έχουν ορισθεί καταχρηστικά προκειμένου τα μοντέλα πρόβλεψης που σχεδιάζουμε και χρησιμοποιούμε να προσαρμόζονται στα μεγέθη με τα οποία είναι εξοικειωμένος ο άνθρωπος και η λειτουργίας τους να γίνεται πιο εύκολα αντιληπτή. Ωστόσο, η διαδικασία αποσύνθεσης της χρονοσειράς στα ποιοτικά της χαρακτηριστικά δεν παύει να είναι ιδιαίτερη χρήσιμη καθώς μας δίνει τη δυνατότητα να απομονώνουμε συστατικά της και να εφαρμόζουμε κατάλληλα μοντέλα πρόβλεψης σε αυτά προκειμένου να βελτιώσουμε την προβλεπτική μας ικανότητα.

Τα ποιοτικά χαρακτηριστικά τα οποία διακρίνουμε (*Makridakis et al., 1983*) είναι η τάση, η κυκλικότητα (κύκλος), η εποχιακότητα και η τυχαιότητα (θόρυβος/διακύμανση). Παρακάτω δίνεται μία σύντομη περιγραφή τους.

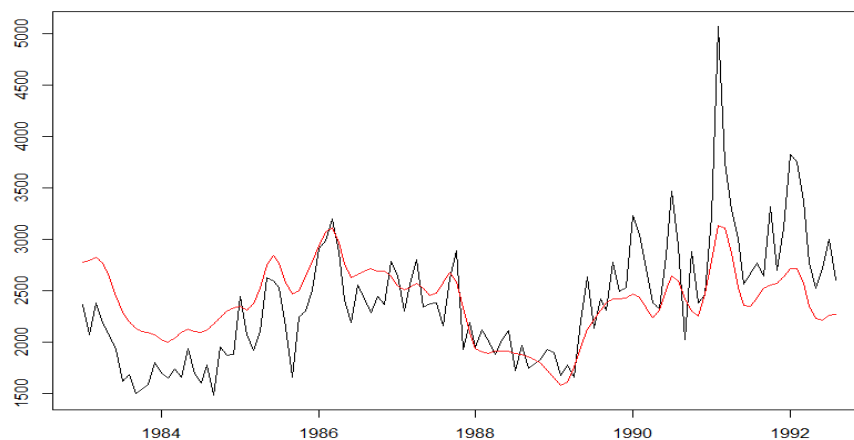
Τάση: Ως τάση ορίζεται η μακροπρόθεσμη μεταβολή του μέσου επιπέδου των τιμών της χρονοσειράς. Προκειμένου να γίνει αντιληπτή θα πρέπει να είναι διαθέσιμος ένας ικανός αριθμός παρατηρήσεων, σε σχέση βέβαια και με τη συχνότητα των δεδομένων. Για παράδειγμα, σε μία ετήσια χρονοσειρά τάση μπορεί να εντοπιστεί ακόμα και μέσα σε ελάχιστες παρατηρήσεις. Στην περίπτωση ωστόσο ωριαίων ή εβδομαδιαίων δεδομένων το αντίστοιχο δείγμα θα πρέπει να είναι σαφώς μεγαλύτερο.



Εικόνα 2.1: Μηνιαίες χρονοσειρές του διαγωνισμού M3 με έντονο το συστατικό της τάσης, εκθετικής (αριστερά) και γραμμικής (δεξιά).

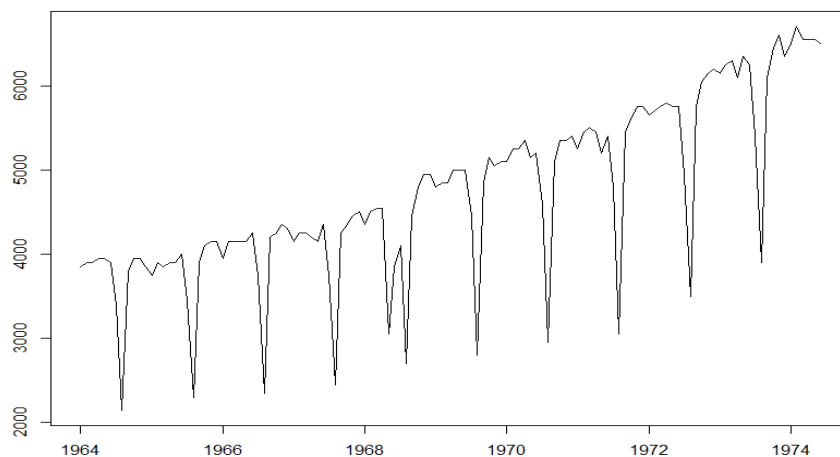
Σε αυτό το σημείο αναφέρεται ότι η τάση δεν έχει πάντα ένα συγκεκριμένο μοτίβο. Στην πιο απλή της περίπτωση η τάση είναι γραμμική (σταθερή μεταβολή επιπέδου), μπορεί ωστόσο να είναι και εκθετικής (μη γραμμική μεταβολή επιπέδου), λογαριθμική ή άλλης μορφής. Συμβατικά, και δεδομένου ότι συνήθως δεν υπάρχουν ιδιαίτερες ενδείξεις περί του τύπου της τάσης, θεωρούμε καταχρηστικά ότι η τάση δεν είναι μεταβαλλόμενη και την εκτιμάμε γραμμικά.

Κυκλικότητα: Αναφέρεται στις μακροχρόνιες και κυματοειδής μεταβολές που συμβαίνουν σε ένα μέγεθος, κυρίως λόγω ειδικών εξωγενών συνθηκών. Ο κύκλος δεν είναι ένα μέγεθος περιοδικό, με την έννοια ότι δεν εμφανίζεται συστηματικά και σε συγκεκριμένα χρονικά διαστήματα, ενώ η έντασή του δεν είναι ομοιόμορφη σε όλη την έκταση της χρονοσειράς. Πρόκειται συνεπώς για ένα μέγεθος παρατηρήσιμο αλλά μη προβλέψιμο. Χαρακτηριστικά παραδείγματα κυκλικότητας εμφανίζονται σε χρονοσειρές οικονομικών μεγεθών όπως το Α.Ε.Π. και παρουσιάζουν τις αυξομειώσεις της οικονομίας μεταξύ κρίσεων.



Εικόνα 2.2: Μηνιαία χρονοσειρά του διαγωνισμού M3 με έντονο το συστατικό της κυκλικότητας.

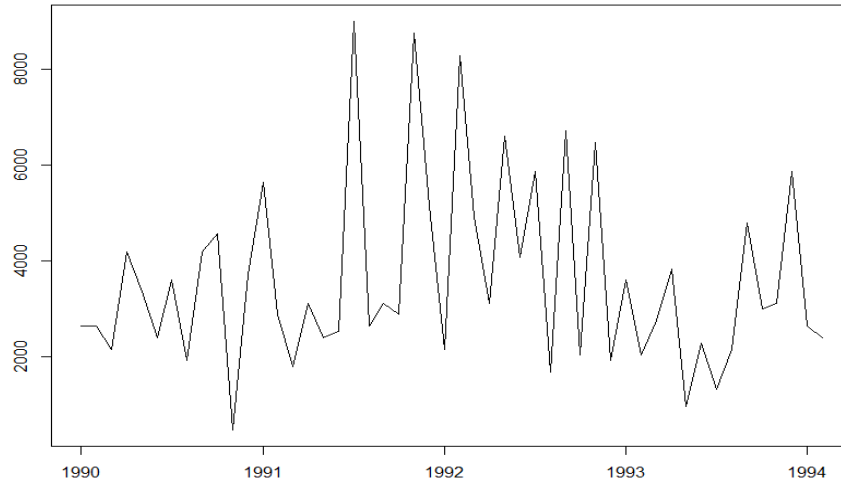
Εποχιακότητα: Αναφέρεται σε περιοδικές διακυμάνσεις που χαρακτηρίζουν το μέγεθος εντός διαστημάτων σταθερού μεγέθους και μήκους μικρότερου του ενός έτους. Η διακύμανση αυτή μπορεί εύκολα να μετρηθεί και συνεπώς, αν δεν υπάρξει κάποια απρόσμενη εξωτερική επίδραση, είναι ιδιαίτερα προβλέψιμη. Δεδομένης μάλιστα της ευκολίας να την απομονώσουμε, συνηθίζεται να προβλέπουμε επί των αποεποχικοποιημένων χρονοσειρών (*seasonally adjusted time series*) και να εποχικοποιούμε τις προβλέψεις μας στη συνέχεια.



Εικόνα 2.3: Μηνιαία χρονοσειρά του διαγωνισμού M3 με έντονο το συστατικό της εποχιακότητας.

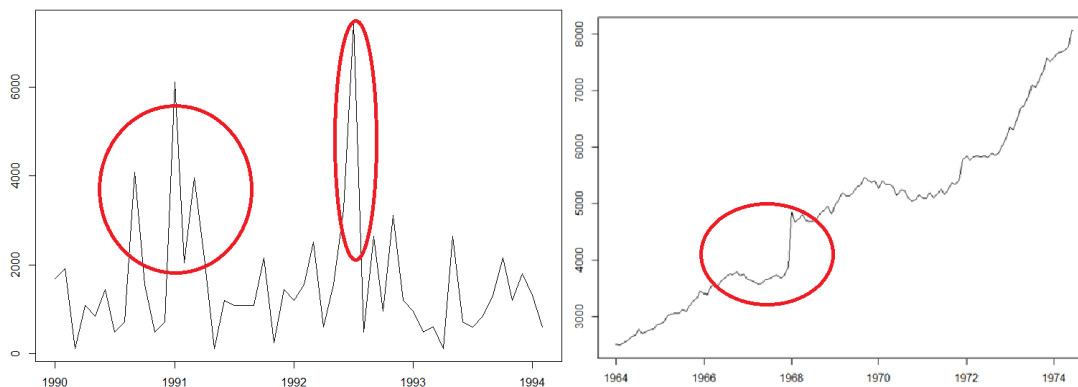
Τυχαιότητα: Πρόκειται για τις μεταβολές οι οποίες δεν μπορούν να ερμηνευθούν μέσω των υπολοίπων ποιοτικών χαρακτηριστικών της χρονοσειράς (*residuals*). Συνήθως οι χρονοσειρές δεν χαρακτηρίζονται από ακραίες μεταβολές, μπορούν όμως κάλλιστα να προκύψουν και να έχουν είτε παροδικό (απτόμενες αλλαγές στο πρότυπο συμπεριφοράς – ακραίες τιμές) είτε μόνιμο (μεταβολή επιπέδου) χαρακτήρα. Δεδομένου ότι η τυχαιότητα δεν μπορεί να μοντελοποιηθεί, το εν λόγω μέγεθος είναι μη προβλέψιμο και καθιστά χρονοσειρές με έντονο το χαρακτηριστικό της αρκετά πιο δύσκολο να προβλεφθούν.

Κεφάλαιο 2: Ανάλυση χρονοσειρών και απόδοσης μοντέλων πρόβλεψης



Εικόνα 2.4: Μηνιαία χρονοσειρά του διαγωνισμού M3 με έντονο το συστατικό της τυχαιότητας.

Όπως αναφέρθηκε νωρίτερα, δεδομένου ότι όλες οι χρονοσειρές έχουν στοχαστικό χαρακτήρα, η τυχαιότητα μπορεί να εντοπιστεί σε κάθε χρονοσειρά. Έτσι, συνηθίζεται να αναφερόμαστε σε αυτή μόνο όταν υπάρχουν συγκεκριμένοι και έντονοι παράγοντες τυχαιότητας, γνωστά και ως *outliers*. Τα *outliers* (Ledolter, 1989) συνήθως διακρίνονται σε μεταβολές επιπέδου (*level change*), ακραίες τιμές (*additive outliers - extreme values*) και προσωρινές μεταβολές (*temporal changes*). Ο εντοπισμός και η προσαρμογή τους είναι εφικτή και πραγματοποιείται μέσω κατάλληλων μεθόδων. Ωστόσο βρίσκεται έξω από τα όρια ενδιαφέροντος της παρούσας διατριβής και για αυτό δεν μελετάται περεταίρω. Αναφέρεται βέβαια ότι σε περιπτώσεις ακραίων μεταβολών συνίσταται η απαλοιφή τους προκειμένου να μην αλλοιώνονται τα βασικά μοτίβα της χρονοσειράς.



Εικόνα 2.5: Μηνιαίες χρονοσειρές του διαγωνισμού M3 με (αριστερά) *additive outliers* και (δεξιά) *level shifts*.

2.2 Αποσύνθεση χρονοσειρών

Στις τεχνικές προβλέψεων, ως αποσύνθεση χαρακτηρίζεται συνήθως η διαδικασία διαχωρισμού μίας χρονοσειράς στα τέσσερα επιμέρους συστατικά που αναφέρθηκαν

νωρίτερα, δηλαδή στην τάση (**T-trend**), στην κυκλικότητα (**C-circle**), στην εποχιακότητα (**S-seasonality**) και στην τυχαιότητα (**R-randomness**). Η εν λόγω διαδικασία είναι ιδιαίτερα σημαντική καθώς, πέρα από το ότι μας βοηθά να κατανοήσουμε τις ιδιότητες της χρονοσειράς, μπορεί δυνητικά να συνεισφέρει στη βελτίωση της απόδοσης των μοντέλων πρόβλεψης. Αυτό συμβαίνει καθώς κάποια από τα συστατικά της χρονοσειράς, όπως η εποχιακότητα, είναι αρκετά προβλέψιμα από μόνα τους και συνεπώς μπορούν να εκτιμηθούν ξεχωριστά και να συμπεριληφθούν εκ των υστέρων στις παραγόμενες προβλέψεις. Μάλιστα, για το παράδειγμα της εποχιακότητας, το εν λόγω συστατικό μπορεί να μοντελοποιηθεί αποτελεσματικά ακόμα και χωρίς τη διαδικασία της αποσύνθεσης μέσω εποχιακών μοντέλων πρόβλεψης, όπως το *Taylor* και το *Holt-Winters*. Σε αντίθετη περίπτωση η εποχιακότητα εισάγει μία επιπρόσθετη διακύμανση στο μοντέλο παραπλανώντας το για το μοτίβο συμπεριφοράς που ακολουθεί η χρονοσειρά.

Ο πιο διαδεδομένος τρόπος αποσύνθεσης στη βιβλιογραφία είναι η κλασική μέθοδος αποσύνθεσης (*Kendall & Stuart, 1983*), γνωστή και ως πολλαπλασιαστική αποσύνθεση κινητών μέσων όρων (*Classical Multiplicative Decomposition by Moving Averages*). Η εν λόγω μέθοδος θεωρεί ότι κάθε χρονοσειρά αποτελείται από τα τέσσερα συστατικά που αναφέρθηκαν και την παράγουν μέσω του γινομένου τους, δηλαδή:

$$Y_i = T_i * C_i * S_i * R_i$$

όπου Y_i είναι η παρατήρηση της χρονοσειράς Y την χρονική περίοδο i .

Προφανώς υπάρχουν και άλλες διαδικασίες αποσύνθεσης χρονοσειρών. Αναφορικά αναφέρουμε τις εξής:

- Classical Multiplicative Decomposition by Moving Averages – κλασική πολλαπλασιαστική μέθοδος
- Classical Additive Decomposition by Moving Averages – κλασική προσθετική μέθοδος. Ίδιας φιλοσοφίας με την κλασική μέθοδο αλλά θεωρεί προσθετική σχέση μεταξύ των συστατικών αποσύνθεσης
- Seasonal decomposition by Loess (*Cleveland et al. , 1990*)
- Zaycoff's Method (*Zaycoff, 1936*)
- Μέθοδος Census II: X-11 και X-12 (*McKenzie, 1984*)
- Central Planning Bureau Method - CPB (*Wald, 1936*)
- SABL Method (*Cleveland et al., 1978*)

Βασική διαφορά μεταξύ των προσθετικών και πολλαπλασιαστικών μεθόδων αποσύνθεσης αποτελεί το γεγονός ότι στην πρώτη περίπτωση οι μεταβολές που οφείλονται στο κάθε συστατικό αποσύνθεσης εφαρμόζονται ανεξάρτητα για να παράγουν την αρχική χρονοσειρά, σε αντίθεση με τη δεύτερη όπου αλληλοεπιδρούν. Έτσι, η εποχιακότητα σε μία χρονοσειρά με ανοδική τάση θα γίνει εντονότερη με την πάροδο του χρόνου μέσω του πολλαπλασιαστικού μοντέλου (αύξηση εποχιακής διακύμανσης κατά τη μεταβολή του επιπέδου της χρονοσειράς), ενώ θα παραμείνει ομοιόμορφη στη περίπτωση του προσθετικού (προσθήκη ίδιας εποχιακής διακύμανσης ανεξαρτήτως επιπέδου). Ένας εύκολος τρόπος για να αποφασίσει κανείς ποιο μοντέλο από τα δύο θα πρέπει να

εφαρμοστεί είναι να αποεποχικοποιήσει τη χρονοσειρά με το πολλαπλασιαστικό μοντέλο και να περάσει από τη χρονοσειρά που προκύπτει την ευθεία της απλής γραμμικής παλινδρόμησης. Αν η ευθεία που προκύπτει διαθέτει κλίση παραμέτρου με στατιστική σημαντικότητα, τότε η επιλογή αποσύνθεσης ήταν σωστή. Σε αντίθετη περίπτωση πρέπει να εφαρμοστεί το προσθετικό μοντέλο (Kuiper, 1979).

Σημειώνεται ότι η διαδικασία αποσύνθεσης είναι ανεξάρτητη από το μοντέλο πρόβλεψης που θα εφαρμοστεί και μπορεί να χρησιμοποιείται στο στάδιο επεξεργασίας της χρονοσειράς είτε για διαγνωστικούς ελέγχους, είτε για την προετοιμασία της πρόβλεψης. Επίσης σημειώνεται ότι οι μέθοδοι που αναφέρθηκαν παραπάνω δεν είναι αυστηρά ορισμένες με την έννοια ότι εμφανίζονται σε αρκετές εκδόσεις με αξιόλογες διαφοροποιήσεις, χαρακτηριστικές εκ των οποίων είναι η εξαίρεση ακραίων τιμών κατά τον υπολογισμό των επιμέρους συστατικών αποσύνθεσης και η εφαρμογή μεθόδων συρρίκνωσης εποχιακών δεικτών (*shrinkage methods*).

2.2.1 Κλασική μέθοδος αποσύνθεσης

Σε αυτό το σημείο παρουσιάζεται η διαδικασία εφαρμογής της κλασικής αποσύνθεσης, η οποία και θα χρησιμοποιηθεί στην παρούσα διατριβή για την αποεποχικοποίηση των χρονοσειρών. Τα βήματα της μεθόδου είναι τα εξής (Makridakis et al., 1983):

Βήμα 1^ο : Βάσει της εποχιακότητας της χρονοσειράς εφαρμόζεται στα δεδομένα κινητός μέσος όρος (*Moving Average - MA*) αντίστοιχου βαθμού (*Order - O*). Για παράδειγμα, σε μηνιαία χρονοσειρά εφαρμόζεται κεντρικός κινητός μέσος όρος (12), ενώ για ημερήσια απλός κινητός μέσος όρος (7). Δεδομένου ότι οι κινητοί μέσοι όροι εξομαλύνουν τη χρονοσειρά συναθροίζοντας τις επιμέρους παρατηρήσεις της βάσει του βαθμού τους, οδηγούν στον προσδιορισμό του τρέχοντος επιπέδου της χρονοσειράς απομακρύνοντας ταυτόχρονα εποχιακότητα και σχετικές διακυμάνσεις. Συνεπώς η εφαρμογή τους οδηγεί στο εξής αποτέλεσμα:

$$MA(o) = T * C$$

Σημειώνουμε ότι ο κινητός μέσος όρος εφαρμόζεται σε χρονοσειρές περιττής εποχιακότητας s και ισούται στο σημείο t με

$$MA(s)_t = \frac{1}{s} \sum_{i=-s \text{ mod } 2}^{s \text{ mod } 2} Y_{t+i}$$

ενώ ο κεντρικός κινητός μέσος όρος σε χρονοσειρές άρτιας εποχιακότητας s και ισούται με το μέσο όρο των κινητών μέσων όρων της χρονοσειράς στα σημεία που προηγούνται και έπονται από το ζητούμενο t . Οι κινητοί μέσοι όροι παρουσιάζονται αναλυτικά στο [Κεφάλαιο 4](#). Η παραπάνω διαδικασία γίνεται δεδομένου ότι όταν έχουμε άρτιο βαθμό δεν μπορούμε να αντιστοιχίσουμε την τιμή του κινητού μέσου όρου σε μία παρατήρηση της χρονοσειράς. Έτσι, αντιστοιχούμε τον μέσο όρο των δύο τιμών που την περικλείουν. Τέλος, σημειώνουμε ότι για τις τιμές της χρονοσειράς για τις οποίες δεν υπάρχει δείγμα παρατηρήσεων για τον υπολογισμό των κινητών μέσων όρων (αρχή και τέλος χρονοσειράς) εΐθισται η χρήση τεχνικών *backcasting* για τον υπολογισμό τους, της

κατάλληλης προέκτασης δηλαδή της αρχικής χρονοσειράς με χρήση μίας μεθόδου πρόβλεψης η οποία συνήθως για λόγους απλότητας είναι η απλοϊκή μέθοδος.

Βήμα 2°: Διαιρείται η αρχική χρονοσειρά με τη σειρά των κινητών μέσων όρων. Αυτή η διαδικασία παράγει τους λόγους εποχιακότητας οι οποίοι όμως εμπεριέχουν τον παράγοντα της τυχαιότητας.

$$\frac{Y}{MA(o)} = \frac{T * C * S * R}{T * C} = S * R$$

Βήμα 3°: Σε αυτό το βήμα υπολογίζονται οι δείκτες εποχιακότητας, οι μέσοι δηλαδή λόγοι εποχιακότητας της χρονοσειράς. Στόχος είναι να καθοριστεί το εποχιακό μοτίβο της χρονοσειράς έχοντας απαλείψει όσον το δυνατόν γίνεται τον παράγοντα της τυχαιότητας. Για να γίνει αυτό υπολογίζεται ανά περίοδο συχνότητας ενδιαφέροντος ο μέσος όρος των αντίστοιχων λόγων εποχιακότητας. Έτσι, για την περίπτωση μίας εβδομαδιαίας χρονοσειράς ο δείκτης εποχιακότητας της Δευτέρας υπολογίζεται ως ο μέσος όρος των λόγων εποχιακότητας που αντιστοιχούν σε όλες τις Δευτέρες της χρονοσειράς.

Προκειμένου τα αποτελέσματα να είναι όσο τον δυνατό απαλλαγμένα από την τυχαιότητα, είθισται να εξαιρούνται κατά τον υπολογισμό οι δύο ακραίες τιμές του δείγματος, δηλαδή η μέγιστη και η ελάχιστη τιμή των λόγων εποχιακότητας. Σε αυτήν την περίπτωση προκειμένου να γίνει η αποεποχικοποίηση των δεδομένων θα πρέπει βέβαια να διαθέτουμε χρονοσειρά μήκους τουλάχιστον τριών περιόδων. Η εν λόγω απαίτηση δεν αποτελεί ωστόσο ουσιαστικό περιορισμό καθώς σε αντίθετη περίπτωση αφενός οι δείκτες που θα υπολογιστούν δε θα είναι αντιπροσωπευτικοί, και αφετέρου όποια μέθοδος πρόβλεψης και αν εφαρμοστεί στη συνέχεια δεν θα διαθέτει ικανό πλήθος παρατηρήσεων για την επίτευξη καλής εφαρμογής, οπότε και η όλη διαδικασία πρόβλεψης θεωρείται εξ αρχής ριψοκίνδυνη.

Βήμα 4°: Παράγεται η αποεποχικοποιημένη χρονοσειρά διαιρώντας την αρχική χρονοσειρά με τους δείκτες εποχιακότητας που προέκυψαν. Προφανώς οι δείκτες εποχιακότητας είναι όσοι και η συχνότητα της χρονοσειράς, οπότε και επαναλαμβάνονται ανά περίοδο. Η σειρά που προκύπτει εμπεριέχει την τάση και τον κύκλο, είναι απαλλαγμένη δηλαδή από την εποχιακότητα αλλά εμπεριέχει τυχαιότητα.

$$\frac{Y}{S} = \frac{T * C * S * R}{S} = T * C * R$$

Βήμα 5°: Όπως και στο πρώτο βήμα στόχος είναι η απομάκρυνση της διακύμανσης της χρονοσειράς με χρήση κινητών μέσων όρων, η εφαρμογή των οποίων παράγει το γινόμενο της τάσης-κύκλου. Αυτή τη φορά ωστόσο, επειδή έχει απαλειφθεί ήδη η εποχιακότητα, η όλη διακύμανση που θα υπολογιστεί θα είναι τυχαία και θα ισούται με τον παράγοντα τυχαιότητας της χρονοσειράς. Για να γίνει αυτό μπορεί να υπολογιστεί ο κινητός μέσος όρος 3, 5 ή 6 και να διαιρεθεί με την αποεποχικοποιημένη χρονοσειρά. Για ακόμα πιο αποτελεσματική εξομάλυνση προτείνεται συχνά η εφαρμογή του διπλού

Κεφάλαιο 2: Ανάλυση χρονοσειρών και απόδοσης μοντέλων πρόβλεψης

κινητού μέσου όρου 3×3 , η εφαρμογή δηλαδή ενός κινητού μέσου όρου τάξης 3 και η συνάθροιση των παραγόμενων τιμών ανά τρεις μεταξύ τους.

$$\frac{T * C * R}{MA(o)} = R$$

Βήμα 6^ο: Έχοντας διαθέσιμη τη σειρά τάσης-κύκλου είναι πλέον δυνατός ο υπολογισμός της τάσης και του κύκλου της χρονοσειράς. Για να γίνει αυτό θα πρέπει να αρχικά να επιλεγεί ένα συγκεκριμένο μοτίβο τάσης και να αφαιρεθεί από αυτήν.

Οι επιλογές εδώ είναι αρκετές και εναπόκεινται στην κρίση του αποφασίζων. Στην κλασική περίπτωση θεωρούμε γραμμικό μοντέλο τάσης το οποίο δεν είναι άλλο από την ευθεία της απλής γραμμικής παλινδρόμησης. Θα μπορούσε κανείς ωστόσο να επιλέξει μία εκθετική ή λογαριθμική γραμμή. Για την περίπτωση της απλής γραμμικής παλινδρόμησης αρκεί ο υπολογισμός των συντελεστών γραμμικής παρεμβολής ως εξής:

$$b = \frac{\frac{\sum_{i=1}^n (t_i T C_i)}{n} - \bar{t} \bar{T} \bar{C}}{\frac{\sum_{i=1}^n t_i^2}{n} - \bar{t}^2}$$

$$a = \bar{T} \bar{C} - b \bar{t}$$

$$T = a + bt$$

όπου $\bar{T} \bar{C}$ και \bar{t} ο μέσος όρος των παρατηρήσεων της σειράς τάσης-κύκλου και παρατηρήσεων, αντίστοιχα. Έτσι,

$$\frac{T * C}{T} = \frac{T * C}{a + bt} = C$$

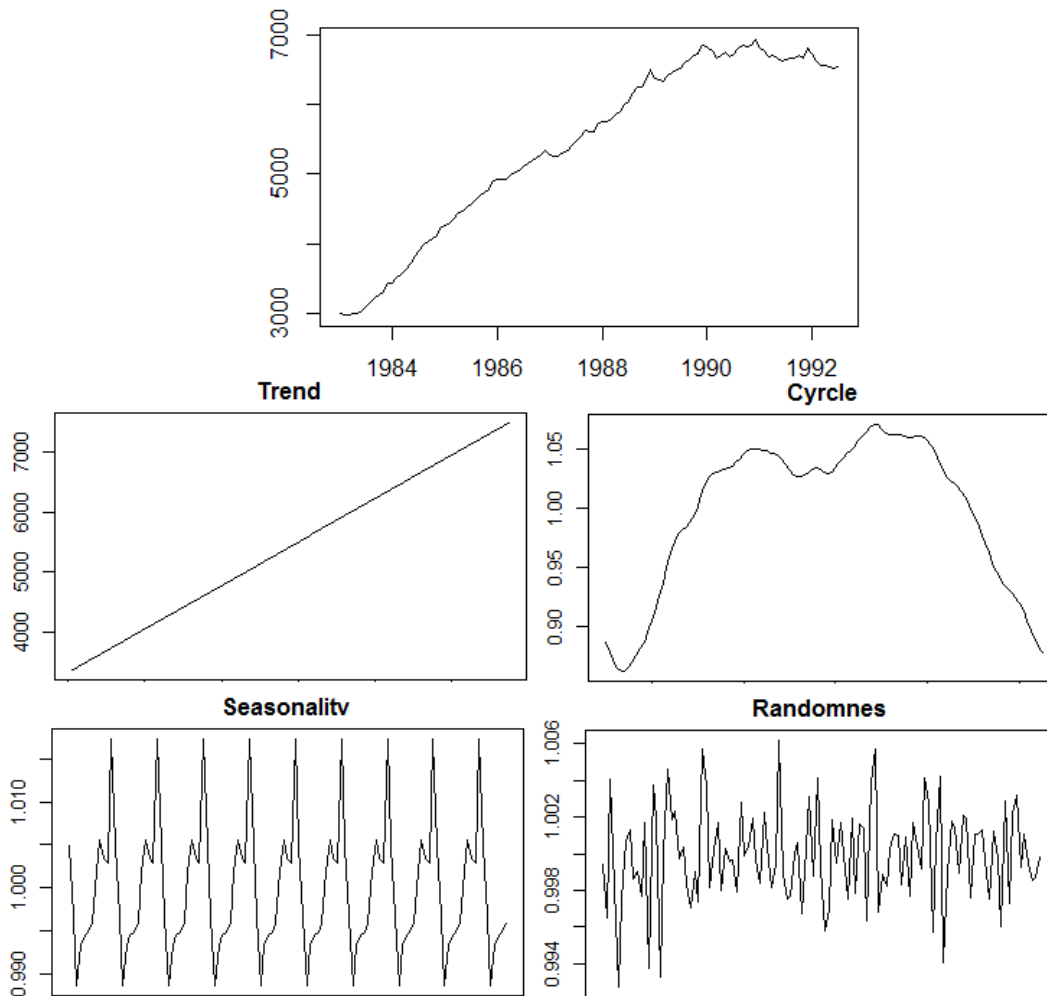
Σημειώνεται ότι η μέθοδος αποσύνθεσης, η οποία συνήθως χρησιμοποιείται για παραγωγή αποεποχικοποιημένων χρονοσειρών και την εφαρμογή μοντέλων πρόβλεψης επί αυτών, μπορεί κάλλιστα να αξιοποιηθεί από μόνη της για την παραγωγή προβλέψεων. Δεδομένου ότι η τυχαιότητα είναι μη προβλέψιμο μέγεθος η πρόβλεψη μέσω της κλασική αποσύνθεσης δίνεται ως εξής:

$$F_i = T_i * S_i * C_{i-1}, i > n$$

Στην παραπάνω σχέση βλέπουμε δηλαδή ότι κατά την προέκταση έχει εξαιρεθεί η τυχαιότητα, οι δείκτες εποχιακότητας είναι ίδιοι με αυτούς που υπολογίστηκαν στο βήμα 3, η τάση εκτιμάται επεκτείνοντας το μοντέλο τάσης που έχει επιλεγεί και ο κύκλος, επειδή δεν υπάρχει ένδειξη για το αντίθετο, παίρνει τιμή ίση με την τελευταία γνωστή.

Στην παρούσα διατριβή όταν θα γίνεται αναφορά σε αποεποχικοποίηση ή στην αποεποχικοποιημένη χρονοσειρά θα εννοείται πάντοτε η κλασική μέθοδος αποσύνθεσης και τα παράγοντά της. Η εν λόγω διαδικασία εφαρμόζεται εύκολα μέσω της γλώσσας

προγραμματισμού R και με τη χρήση κατάλληλης συνάρτησης που γράφτηκε για αυτό τον σκοπό. Στο ακόλουθο παράδειγμα φαίνεται το αποτέλεσμα της αποσύνθεσης για μία μηνιαία χρονοσειρά του M3.



Εικόνα 2.6: Ανάλυση χρονοσειράς με κλασική αποσύνθεση στα συστατικά trend (T), cycle (C), seasonality (S) και randomness (R).

2.2.2 Μέθοδοι συρρίκνωσης συντελεστών εποχιακότητας

Ο επακριβής προσδιορισμός των συντελεστών εποχιακότητας είναι καθοριστικός για την επίτευξη υψηλής ακρίβειας στις προβλέψεις. Αυτό συμβαίνει καθώς σε περίπτωση που εκτιμήσουμε λανθασμένα τις τιμές των συντελεστών, η αποεποχικοποιημένη χρονοσειρά θα εμπεριέχει μία επιπρόσθετη εποχιακότητα η οποία θα γίνεται αντιληπτή από τα μοντέλα πρόβλεψης ως θόρυβος, μειώνοντας έτσι την προβλεπτική τους ικανότητα. Αντίστοιχα, κατά την εποχικοποίηση των προβλέψεων με ακατάλληλους δείκτες θα οδηγηθούμε σε συστηματικά αισιόδοξες ή απαισιόδοξες προβλέψεις ανά περίοδο εποχιακότητας.

Προκειμένου να αποφευχθούν όσο το δυνατόν τέτοια φαινόμενα συχνά εφαρμόζεται μετά την αποσύνθεση της χρονοσειράς μία μέθοδος συρρίκνωσης συντελεστών

εποχιακότητας. Επί της ουσίας οι εν λόγω μέθοδοι εξετάζουν τη σημαντικότητα των δεικτών εποχιακότητας αξιολογώντας τη διακύμανση των λόγων εποχιακότητας από τους οποίους παράγονται. Έτσι, οι τιμές τους αναπροσαρμόζονται αναλόγως. Αν και υπάρχουν αρκετές μέθοδοι που πραγματοποιούν μία τέτοια διαδικασία οι πλέον γνωστές και αποτελεσματικές είναι αυτές των *James-Stein*, *Lemon-Krutchkoff* και *Miller-Williams* (*Miller & Williams, 2003*).

2.2.2.1 Μέθοδος συρρίκνωσης James-Stein

Έστω S οι τιμές των εποχιακών δεικτών που παράχθηκαν από τη κλασική μέθοδο αποσύνθεσης. Οι *James* και *Stein* προτείνουν την επανεξέτασή τους λαμβάνοντας τις τιμές S^{JS} ως εξής:

$$S_j^{JS} = W^{JS} + (1 - W^{JS})S_j$$

$$W^{JS} = \frac{pos - 3}{pos - 1} \frac{V}{V + A}$$

$$V = \frac{1}{pos} \sum_{j=1}^{pos} \frac{\sum_{k=1}^{K_j} (S_{jk} - S_j)^2}{K_j(K_j - 1)}$$

$$A = \max\left(\frac{\sum_{j=1}^{pos} (S_j - 1)^2}{pos - 1} - V, 0\right)$$

, όπου pos είναι οι περίοδοι της χρονοσειράς σε ένα πλήρη κύκλο εποχιακότητας, S_{jk} είναι ο λόγος εποχιακότητας για την εποχή j στο έτος k και K_j είναι το πλήθος των λόγων εποχιακότητας που υπολογίστηκαν μέσω της αποσύνθεσης για την εποχή j .

Οι παράγοντες A και V δηλώνουν τις διαφορές που εντοπίζονται στους λόγους εποχιακότητας λόγω δειγματοληπτικών λαθών και άλλων παραγόντων και βάσει αυτών καθορίζεται αντίστοιχα ο βαθμός συρρίκνωσης των δεικτών.

2.2.2.2 Μέθοδος συρρίκνωσης Lemon-Krutchkoff

Βασιζόμενοι στη δουλειά των *James-Stein*, οι *Lemon* και *Krutchkoff* προτείνουν την επανεξέτασή των δεικτών λαμβάνοντας για την περίοδο j την τιμή $S_{j^*}^{LK}$ ως εξής:

$$S_{j^*}^{LK} = \sum_{j=1}^{pos} W_{j^*,j} S_j$$

$$W_{j^*,j} = \frac{L_{j^*,j}}{\sum_{j=1}^{pos} L_{j^*,j}}$$

$$L_{i,j} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{S_i - S_j}{\sigma}\right)^2}$$

$$\sigma = \sqrt{V}$$

Στην εν λόγω περίπτωση οι δείκτες εποχιακότητας υπολογίζονται ως ένας σταθμισμένος μέσος όρος των δεικτών που προκύπτουν από την κλασική μέθοδο αποσύνθεσης. Τα βάρη φανερώνουν την πιθανότητα να παρατηρηθεί εποχιακότητα S_j όταν ο πραγματικός δείκτης είναι S_j . Σημειώνεται ότι κατά τον υπολογισμό των δεικτών υποθέτουμε ότι η κατανομή δειγματοληψίας είναι κανονική, κάτι το οποίο δεν απέχει πολύ από την πραγματικότητα δεδομένης της στάθμισης των λόγων εποχιακότητας που έχουν προϋπολογιστεί.

2.2.2.3 Μέθοδος συρρίκνωσης Miller-Williams

Η συρρίκνωση συντελεστών κατά *Miller* και *Williams* περιλαμβάνει μία διαδικασία κατά την οποία βάσει της κυρτότητας των δεικτών εποχιακότητας και της τιμής του δείκτη W^{JS} επιλέγεται η καταλληλότερη από τις μεθόδους συρρίκνωσης *James-Stein* και *Lemon-Krutchkoff* που παρουσιάστηκαν νωρίτερα. Δεδομένης της κυρτότητας (*skewness*), οι επιλογές παρουσιάζονται στον *Πίνακα 2.1*.

$$skewness = \frac{\frac{1}{pos} \sum_{i=1}^{pos} (S_i - \bar{S})^3}{\left(\frac{1}{pos} \sum_{i=1}^{pos} (S_i - \bar{S})^2\right)^{\frac{3}{2}}} \begin{cases} < 0.5 \text{ Συμμετρικοί παράγοντες} \\ \geq 0.5 \text{ Ασύμμετροι παράγοντες} \end{cases}$$

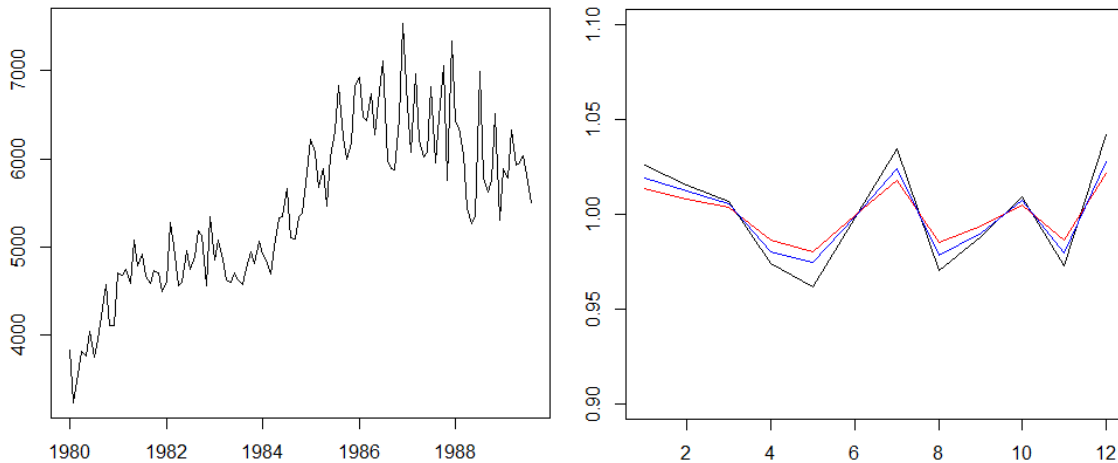
Πίνακας 2.1: Προτεινόμενη μέθοδος συρρίκνωσης εποχιακών δεικτών κατά Miller-Williams.

W^{JS}	Συμμετρικοί παράγοντες	Ασύμμετροι παράγοντες
$W^{JS} < 0.2$	Οποιαδήποτε επιλογή	Lemon-Krutchkoff
$0.2 < W^{JS} < 0.5$	James-Stein	Lemon-Krutchkoff
$W^{JS} > 0.5$	James-Stein	James-Stein

Στην *Εικόνα 2.7* παρουσιάζεται ένα χαρακτηριστικό παράδειγμα για τους δείκτες εποχιακότητας που παράγονται ανά μέθοδο. Όπως παρατηρείται, και οι δύο μέθοδοι συρρίκνωσης έχουν περιορίσει σημαντικά την εποχιακή διακύμανση που υπολογίστηκε αρχικά βάσει της κλασικής μεθόδου αποσύνθεσης, με την μέθοδο *James-Stein* να τις ελαχιστοποιεί.

Τα προβλεπτικά οφέλη της χρήσης των μεθόδων συρρίκνωσης εποχιακών δεικτών αποδεικνύονται εμπράκτως μέσα από τα αποτελέσματα του *Πίνακα 2.2*, όπου η κλασική μέθοδος αποσύνθεσης και οι τρεις εναλλακτικές μέθοδοι συρρίκνωσης συντελεστών που παρουσιάστηκαν νωρίτερα χρησιμοποιούνται για να αποεποχικοποιήσουν τα εποχιακά δεδομένα των διαγωνισμών *M3* και *M1* (4004 χρονοσειρές). Τα αποτελέσματα αφορούν πέντε κλασικές μεθόδους πρόβλεψης που περιγράφονται αναλυτικά στο *Κεφάλαιο 3*.

Κεφάλαιο 2: Ανάλυση χρονοσειρών και απόδοσης μοντέλων πρόβλεψης



Εικόνα 2.7: Υπολογισμός δεικτών εποχιακότητας για μηνιαία χρονοσειρά του διαγωνισμού M3 με και χωρίς συρρίκνωση συντελεστών.

Δεξιά οι τιμές των δεικτών για εφαρμογή της κλασικής μεθόδου αποσύνθεσης (μαύρο), της μεθόδου *James-Stein* (κόκκινο) και της μεθόδου *Lemon-Krutchkoff* (μπλε).

Συνοπτικά, η μέθοδος *James-Stein* βελτιώνει ανάλογα το μοντέλο την ακρίβεια πρόβλεψης από 0.21% έως και 0.43%, ποσοστό διόλου ευκαταφρόνητο δεδομένης της απλότητας της διαδικασίας μέσω της οποίας επιτυγχάνεται και του γεγονότος ότι μόνο ένα μέρος των εποχιακών χρονοσειρών του δείγματος επιδέχεται αποεποχικοποίηση. Η αντίστοιχη βελτίωση κυμαίνεται από 0.04% έως 0.27% για την περίπτωση της *Lemon-Krutchkoff* και από 0.15% έως 0.29% για την μέθοδο *Miller-Williams*. Σημειώνεται επίσης ότι λόγω της ανωτερότητας του μεθοδολογικού πλαισίου της προσέγγισης *Miller-Williams* και της γενίκευσης που αυτή προσφέρει προτείνεται βιβλιογραφικά η υιοθέτησή της έναντι αυτής των *Lemon-Krutchkoff* και *James-Stein*, ακόμα και αν στο συγκεκριμένο δείγμα δεδομένων η τελευταία φαίνεται να αποδίδει ελαφρώς καλύτερα.

Πίνακας 2.2: Προβλεπτική ακρίβεια κλασικών μοντέλων πρόβλεψης για αποσύνθεση των εποχιακών δεδομένων των διαγωνισμών M3 και M1 μέσω τις κλασικής μεθόδου και εναλλακτικών μεθόδων συρρίκνωσης εποχιακών δεικτών.

Δεδομένα	Naive	SES	Holt	Damped	Theta
Κλασική αποεποχικοποίηση					
M3 Τριμηνιαία	9.97	9.79	11.23	9.74	9.29
M3 Μηνιαία	16.71	14.26	15.82	13.95	13.87
M1 Τριμηνιαία	17.90	17.69	19.40	18.37	16.30
M1 Μηνιαία	16.53	14.99	18.75	15.82	16.52
Σύνολο	15.06	13.52	15.51	13.57	13.42
James-Stein					
M3 Τριμηνιαία	9.96	9.78	11.27	9.73	9.28
M3 Μηνιαία	16.65	14.20	15.71	13.90	13.77
M1 Τριμηνιαία	17.89	17.68	19.27	18.31	16.29

M1 Μηνιαία	16.42	14.94	18.76	15.80	16.48
Σύνολο	15.00	13.48	15.46	13.54	13.37
Lemon-Krutchkoff					
M3 Τριμηνιαία	9.95	9.77	11.17	9.73	9.28
M3 Μηνιαία	16.69	14.22	15.83	13.94	13.80
M1 Τριμηνιαία	17.89	17.68	19.28	18.21	16.28
M1 Μηνιαία	16.46	14.98	18.79	15.77	16.51
Σύνολο	15.03	13.49	15.50	13.55	13.39
Miller-Williams					
M3 Τριμηνιαία	9.95	9.78	11.23	9.72	9.28
M3 Μηνιαία	16.67	14.21	15.74	13.92	13.79
M1 Τριμηνιαία	17.89	17.68	19.30	18.29	16.29
M1 Μηνιαία	16.51	14.99	18.77	15.82	16.51
Σύνολο	15.03	13.49	15.47	13.55	13.38

2.3 Ανάλυση χρονοσειρών

Η βασική στατιστική ανάλυση των χρονοσειρών πραγματοποιείται μέσω της χρήσης στατιστικών δεικτών και μεθόδων. Στο παρόν κεφάλαιο παρουσιάζονται κάποιες από αυτές και δίνεται ιδιαίτερη βαρύτητα σε όσες χρησιμοποιούνται για την υλοποίηση και αξιολόγηση των τεχνικών της παρούσας διατριβής.

2.3.1 Βασικοί στατιστικοί δείκτες

Η **μέση τιμή** (ροπή πρώτης τάξης ή *average* ή *mean*) μίας χρονοσειράς, η οποία μας δείχνει το επίπεδο γύρω από το οποίο μεταβάλλονται μία χρονοσειρά, είναι ίση με:

$$\forall t \in Z, \mu_t = E[Y_t] = \int_{-\infty}^{\infty} y f_Y(y, t) dy$$

αν προσεγγίσουμε την χρονοσειρά σαν στοχαστική διαδικασία, ή διαφορετικά $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_t$ για το σύνολο των διακριτών παρατηρήσεων της χρονοσειράς.

Η ροπή δεύτερης τάξης(τάση), δηλαδή ο τρόπος με τον οποίο μεταβάλλεται το επίπεδο της χρονοσειράς, ορίζεται αντίστοιχα ως:

$$\forall t_1, t_2 \in Z \quad \kappa(t_1, t_2) = E[Y_{t_1} Y_{t_2}] = \iint_{-\infty}^{\infty} y_1 y_2 f_Y(y_1, y_2, t_1, t_2) dy$$

Η κεντρική ροπή δεύτερης τάξης, που ονομάζεται αυτοδιασπορά (*autocovariance*), είναι

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

Για $t_1 = t_2 = t$ ορίζεται λοιπόν η **διακύμανση** της χρονοσειράς (*variance*) ως:

$$\sigma_t^2 = \sigma_{Y_t}^2 = E[(Y_t - \mu_t)^2] = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}$$

Η ρίζα της διακύμανσης μας δίνει τη **διασπορά** (τυπική απόκλιση ή *standard deviation*) της χρονοσειράς. Και τα δύο μεγέθη χρησιμοποιούνται για να δείξουν το βαθμό συγκέντρωσης των παρατηρήσεων γύρω από τη μέση τιμή. Αντίστοιχα ορίζονται οι ροπές και οι κεντρικές ροπές μεγαλύτερης τάξης για δύο μεταβλητές και οι ροπές γενικεύονται για περισσότερες μεταβλητές. Σημειώνεται ότι μέσα από τις ροπές κάθε τάξης μπορεί να οριστεί για δύο ή περισσότερες μεταβλητές η αντίστοιχη κοινή κατανομή. Σε αυτήν τη γενική περιγραφή της στοχαστικής διαδικασίας οι κατανομές και οι ροπές είναι συναρτήσεις των χρονικών στιγμών, δηλαδή μπορούν να μεταβάλλονται με το χρόνο.

Άλλοι βασικοί στατιστικοί δείκτες είναι:

Μέγιστη και ελάχιστη τιμή: Οι ακραίες τιμές της χρονοσειράς βοηθούν στην αντίληψη της διακύμανσης των δεδομένων της, καθώς και της τυχαιότητας αυτών.

Ενδιάμεση και πλέον εμφανιζόμενη τιμή: Η ενδιάμεση τιμή (*median*) είναι η μέση παρατήρηση της χρονοσειράς όταν οι τιμές αυτής καταταχτούν σε αύξουσα σειρά. Σε περίπτωση που το πλήθος τους είναι άρτιο, ως ενδιάμεση τιμή ορίζεται ο μέσος όρος των δύο μεσαίων παρατηρήσεων της. Αντίθετα, ως πλέον εμφανιζόμενη τιμή (*mode*) ορίζεται εκείνη η τιμή της χρονοσειράς η οποία εμφανίζεται συχνότερα. Σε περίπτωση που οι παρατηρήσεις διακρίνονται από συνέχεια, μπορεί κανείς να εφαρμόσει τεχνικές kernel για τον προσδιορισμό της κατανομής τους, της ενδιάμεσης και της πλέον εμφανιζόμενης τιμής. τους Στόχος και στις δύο περιπτώσεις είναι ένας πιο στιβαρός προσδιορισμός του επιπέδου γύρω από το οποίο κινείται η χρονοσειρά.

Συνδιακύμανση: Ο δείκτης αυτός εμφανίζει τη σχέση ανάμεσα σε δύο μεταβλητές, συγκρίνοντας τη φορά μεταβολής τους. Δείχνει δηλαδή αν μεταβάλλονται προς την ίδια (θετική συνδιακύμανση) ή προς την αντίθετη κατεύθυνση (αρνητική συνδιακύμανση) και τη συσχέτιση ή μη των συνόλων των δύο μεταβλητών (όταν δεν υπάρχει συσχέτιση, πρόκειται για μηδενική συνδιακύμανση). Στην περίπτωση της μελέτης χρονοσειράς γίνεται υπολογισμός της συνδιακύμανσης των παρατηρήσεων της συγκριτικά με τον αύξοντα αριθμό της χρονικής περιόδου.

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^n [(X_i - \bar{X})(Y_i - \bar{Y})]$$

Συντελεστής γραμμικής συσχέτισης: Είναι το μέτρο της συγκέντρωσης των σημείων ενός διαγράμματος διασποράς γύρω από την ευθεία παλινδρόμησης, που δηλώνει τη γραμμική συσχέτιση μεταξύ δύο συνόλων δεδομένων (κατεύθυνση και βαθμό συσχέτισης).

- Για $0 < r < 1$ οι δύο μεταβλητές X και Y είναι θετικά και γραμμικά συσχετισμένες.
- Για $-1 < r < 0$ οι δύο μεταβλητές είναι αρνητικά και γραμμικά συσχετισμένες.
- Για $r = \pm 1$ υπάρχει τέλεια θετική γραμμική συσχέτιση, οπότε όλα τα σημεία βρίσκονται επί της ευθείας παλινδρόμησης.
- Για $r = 0$ δεν υπάρχει καθόλου συσχέτιση μεταξύ των δύο μεταβλητών.

Να σημειώσουμε ότι όσο η τιμή του δείκτη r τείνει στις ακρινές τιμές του διαστήματος $[-1, 1]$ στο οποίο κινείται, τόσο πιο ισχυρή είναι η συσχέτιση μεταξύ των δύο μεταβλητών. Για απόλυτη τιμή μεγαλύτερη του 0.75 θεωρούμε πολύ ισχυρή συσχέτιση και για μικρότερη του 0.35 ανύπαρκτη συσχέτιση. Ο δείκτης μπορεί να χρησιμοποιηθεί για την εξακρίβωση ύπαρξης γραμμικής τάσης σε μία χρονοσειρά.

$$r_{xy} = \frac{\sum_{i=1}^n [(X_i - \bar{X})(Y_i - \bar{Y})]}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Συντελεστής αυτοσυσχέτισης: Ο δείκτης αυτός εκφράζει τη συσχέτιση μεταξύ παρατηρήσεων της ίδιας πλέον μεταβλητής, οι οποίες έχουν χρονική διαφορά k περιόδους (*lag* ή *υστέρηση*). Η τιμή του κινείται στο διάστημα $[0, 1]$ και όταν είναι κοντά στο άνω όριο της φανερώνει μεγάλη συσχέτιση, ενώ το αντίθετο συμβαίνει για τιμή κοντά στο μηδέν. Ο δείκτης χρησιμοποιείται στην περίπτωση που μας ενδιαφέρει η μελέτη της εποχικής συμπεριφοράς μιας χρονοσειράς και μπορεί να λειτουργήσει ως ένδειξη για τον έλεγχο ύπαρξης της (υπολογισμός με k ίσο με τη συχνότητα της χρονοσειράς). Επίσης μπορεί να αξιοποιηθεί για την εξακρίβωση ύπαρξης συστηματικής γραμμικής τάσης (υπολογισμός με $k=1$).

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

Συντελεστής μεταβλητότητας: Πρόκειται για κανονικοποιημένο μέτρο της διασποράς των παρατηρήσεων δείγματος ή πληθυσμού ως προς τη μέση τιμή τους και πλεονεκτεί έναντι της απλής τυπικής απόκλισης λόγω της μη συμπερίληψης της επίδρασης του επιπέδου των παρατηρήσεων. Δε γίνεται όμως να υπολογιστεί σε περίπτωση μηδενικής μέσης τιμής. Αποτελεί ένδειξη για την τυχαιότητα που διακρίνει τη χρονοσειρά.

Ρυθμός ανάπτυξης: Πρόκειται για ένα μέτρο της πορείας των δεδομένων μιας χρονοσειράς, είτε αύξουσα ή φθίνουσα, για ορισμένο χρονικό διάστημα. Η τιμή του παρουσιάζεται σε ποσοστιαία μορφή και προκύπτει ως η σύγκριση των δεδομένων του τελευταίου έτους με τα δεδομένα της υπόλοιπης χρονικής περιόδου που διατίθεται. Η μαθηματική έκφραση του ρυθμού ανάπτυξης έχει ως εξής:

$$Growth Rate = \frac{\frac{1}{ppy} \sum_{i=n-ppy+1}^n Y_i - \frac{1}{n-ppy} \sum_{i=1}^{n-ppy} Y_i}{\frac{1}{n-ppy} \sum_{i=1}^{n-ppy} Y_i} 100\%$$

,όπου Y το διάνυσμα των n παρατηρήσεων και ppy το πλήθος των περιόδων στο μήκος ενός έτους. Αποτελεί ένδειξη για το μέγεθος της τάσης της χρονοσειράς.

2.3.2 Στασιμότητα χρονοσειρών

Η στατιστική περιγραφή μίας χρονοσειράς απλουστεύεται σημαντικά αν θεωρήσουμε ότι οι στατιστικές της ιδιότητες (*ροπές*) παραμένουν σταθερές στο χρόνο. Μία τέτοια

χρονοσειρά ορίζεται ως στάσιμη και βασικό χαρακτηριστικό της αποτελεί η διατήρηση της μέσης τιμής και της διακύμανσής της κατά μήκος του χρόνου.

Η στασιμότητα δύσκολα παρατηρείται σε πραγματικά προβλήματα, ωστόσο χρησιμοποιείται συχνά ως υπόθεση εργασίας για την εξαγωγή χρήσιμων συμπερασμάτων. Συχνά μετατρέπουμε μάλιστα με κατάλληλες μεθόδους μη στάσιμες χρονοσειρές σε στάσιμες ούτως ώστε να τις μελετήσουμε πιο αποτελεσματικά. Στην πράξη, μία χρονοσειρά η οποία προσεγγίζει τις ιδιότητες του λευκού θορύβου μπορεί να θεωρηθεί απόλυτα στάσιμη, ενώ μία χρονοσειρά με έντονη εποχιακότητα ή τάση μη στάσιμη. Ειδικότερα ορίζονται δύο μορφές στασιμότητας: η *αυστηρή* και η *ασθενής στασιμότητα*.

Μία στοχαστική διαδικασία (ή χρονοσειρά) $\{Y_t\}_{t=-\infty}^{\infty}$ είναι αυστηρά στάσιμη (*strict-sense stationary*) όταν οι κατανομές της για κάθε τάξη (ή ισοδύναμα όλες οι ροπές) είναι σταθερές στο χρόνο, δηλαδή ισχύει:

$$\forall t \in \mathbb{Z}, f_{Y_t}(y) = f_Y(y, t) = f_Y(y, t) \text{ και}$$

$$\forall t_1, t_2 \in \mathbb{Z}, f_{Y_{t_1} Y_{t_2}}(y_1, y_2) = f_{Y_{t_1-t_2}}(y_1, y_2)$$

Αντίστοιχα ισχύουν για κατανομές μεγαλύτερης τάξης. Για ροπές τάξης μεγαλύτερης του ένα, οι κατανομές δίνονται ως συνάρτηση όχι των χρονικών στιγμών, π.χ. t_1, t_2 , αλλά της υστέρησης μεταξύ των χρονικών στιγμών, π.χ. $\tau = t_1 - t_2$, δηλαδή για οποιοσδήποτε δύο χρονικές στιγμές που απέχουν μεταξύ τους τ χρονικά βήματα. Ο έλεγχος της αυστηρής στασιμότητας απαιτεί τη διερεύνηση κοινών κατανομών ή ροπών όλων των τάξεων και δεν αποτελεί μια πρακτικά χρήσιμη ιδιότητα. Για αυτό συχνά η συνθήκη στασιμότητας περιορίζεται στις δύο πρώτες ροπές.

Μια χρονοσειρά είναι ασθενώς στάσιμη (*weak ή wide-sense stationary*) όταν οι ροπές πρώτης και δεύτερης τάξης είναι σταθερές στο χρόνο, δηλαδή:

α) η μέση τιμή είναι σταθερή : $\forall t \in \mathbb{Z}, E[Y_t] = \mu$

β) η αυτοδιασπορά ορίζεται μόνο ως προς την υστέρηση και όχι τις χρονικές στιγμές

$$\forall t_1, t_2 \in \mathbb{Z} \quad \gamma(t, t - \tau) = \gamma(t) = \gamma(\tau) = \gamma_\tau = \frac{1}{n} \sum_{t=\tau+1}^n (y_t y_{t-\tau} - \bar{y}^2), \quad \tau = 0, 1, \dots, n - 1$$

και για $\tau=0$ η διασπορά προκύπτει ως:

$$s_x^2 \equiv \gamma(0) \equiv \sigma_x^2 = \frac{1}{n} \sum_{t=1}^n (y_t^2 - \bar{y}^2)$$

Για τη μελέτη συσχετίσεων σε στάσιμες χρονοσειρές χρησιμοποιείται η αυτοσυσχέτιση. Στη γενική της μορφή η αυτοσυσχέτιση για υστέρηση τ ορίζεται ως:

$$\rho_\tau \equiv \rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}$$

και μετράει τη συσχέτιση των παρατηρήσεων της χρονοσειράς που βρίσκονται σε χρονική υστέρηση τ . Μπορεί επίσης να δειχθεί πως $|\gamma_\tau| \leq \gamma_0$ και άρα $\rho_\tau \leq 1$ για κάθε υστέρηση τ . Αν θέλαμε να περιορίσουμε τη συσχέτιση των y_t και $y_{t-\tau}$ μόνο στην απευθείας συσχέτισή τους, αναφερόμαστε στην μερική αυτοσυσχέτιση, η οποία ορίζεται ως $Corr(y_t, y_{t-\tau} \mid y_{t-1}, \dots, y_{t-\tau+1})$.

Παρακάτω θα παρουσιαστούν αναλυτικά στατιστικοί έλεγχοι για την εξακρίβωση της στασιμότητας σε μία χρονοσειρά. Ωστόσο αυτή μπορεί να ελεγχθεί άμεσα στην πράξη μέσω των δεικτών αυτοσυσχέτισης και μερικής αυτοσυσχέτισης μελετώντας της εξέλιξή τους για διάφορες τιμές υστέρησης τ . Για να είναι μία χρονοσειρά στάσιμη θα πρέπει λοιπόν:

- Οι τιμές και των δύο δεικτών να μην ακολουθούν κάποιο μοτίβο καθώς αυξάνεται η τιμή της υστέρησης (ημιτονοειδές ή εκθετικό)
- Οι τιμές των δεικτών να μην εμφανίζουν υψηλές τιμές (π.χ. μεγαλύτερες από 0.5) για καμία τιμή υστέρησης.

2.3.3 Στατιστικοί Έλεγχοι

Οι στατιστικοί έλεγχοι χρησιμοποιούνται προκειμένου να διαπιστωθεί αν μία υπόθεση που γίνεται ισχύει, αν δηλαδή υπάρχει στατιστικά σημαντική ένδειξη που να την υποστηρίζει. Τέτοιοι έλεγχοι είναι ιδιαίτερα χρήσιμοι καθώς μας πληροφορούν π.χ. για το αν μία χρονοσειρά είναι εποχιακή, στάσιμη ή οι προβλέψεις μίας μεθόδου διαφέρουν από μίας άλλης.

Πριν προχωρήσουμε στην παρουσίαση των διάφορων τύπων στατιστικών ελέγχων θα χρειαστεί να ορίσουμε δύο βασικές στοχαστικές διαδικασίες, οι οποίες θα μας χρησιμεύσουν ως αναφορές στη συνέχεια. Οι διαδικασίες είναι αυτές του *λευκού θορύβου* και του *τυχαίου περιπάτου*.

Λευκός θόρυβος (white noise): Μια στοχαστική διαδικασία η οποία μπορεί να μην έχει γραμμικές συσχετίσεις αλλά τα στοιχεία της ενδέχεται να μην είναι ανεξάρτητα. Συμβολίζεται ως $WN(0, \sigma_z^2)$ με μέση τιμή 0 και διασπορά σ_z^2 . Μαθηματικά αυτό περιγράφεται από τη σχέση $E[X_i X_j] = \delta_{ij} \sigma_x^2$ για οποιεσδήποτε δύο τυχαίες μεταβλητές της χρονοσειράς.

Τυχαίος Περίπατος (random walk): Είναι μια μη-στάσιμη χρονοσειρά $\{r_t\}_{t=-\infty}^{\infty}$, όπου η κάθε τυχαία μεταβλητή r_t για χρόνο t προκύπτει όταν στην προηγούμενη τυχαία μεταβλητή r_{t-1} προστεθεί ένα τυχαίο βήμα u_t . Η χρονοσειρά περιγράφεται σε αυτήν την περίπτωση από τη σχέση $r_t = r_{t-1} + u_t$.

2.3.3.1 Μηδενική και Εναλλακτική υπόθεση - πορεία ελέγχου

Δυο στατιστικές υποθέσεις εμπλέκονται στον έλεγχο υποθέσεων: Η *μηδενική* και η *εναλλακτική* υπόθεση. Η πρώτη είναι αυτή που ελέγχεται με σκοπό να αμφισβητηθεί και συμβολίζεται συνήθως ως H_0 . Μερικές φορές, η μηδενική υπόθεση λέγεται και υπόθεση μη διαφοράς επειδή είναι μια άποψη που συμφωνεί με τις συνθήκες που θεωρούμε ότι αληθεύουν για τον πληθυσμό που μελετούμε. Κατά τον έλεγχο της λοιπόν, η μηδενική

Κεφάλαιο 2: Ανάλυση χρονοσειρών και απόδοσης μοντέλων πρόβλεψης

υπόθεση είτε απορρίπτεται, είτε όχι. Σημειώνεται ότι σε περίπτωση που δεν απορριφθεί δεν σημαίνει αναγκαστικά ότι ισχύει. Για αυτό το λόγο συχνά αναφέρουμε κατά την μη απόρριψή της ότι 'τα δεδομένα πάνω στα οποία στηρίζεται ο έλεγχος δεν επαρκούν για την απόρριψή της' ή εν συντομία ότι 'ενδέχεται να είναι αληθινή'. Εάν πάλι ο έλεγχος οδηγήσει στην απόρριψή της, τότε συμπεραίνουμε ότι τα δεδομένα είναι συμβατά με κάποια άλλη υπόθεση, την εναλλακτική H_1 .

Στην συνέχεια κατασκευάζεται μια στατιστική η οποία ονομάζεται κριτήριο ελέγχου και η τιμή της οποίας υπολογίζεται από τα δεδομένα του δείγματος. Το μέγεθος της τιμής του κριτηρίου είναι εκείνο που καθορίζει το αν θα απορρίψουμε ή όχι τη μηδενική υπόθεση ενώ η πρόταση στην οποία καταλήγει βασίζεται στην πιθανότητα αυτή να ικανοποιείται στο διαθέσιμο δείγμα.

Η στατιστική σημαντικότητα ενός αποτελέσματος είναι η πιθανότητα η υποφαινόμενη σχέση σε ένα δείγμα να εμφανίστηκε κατά τύχη, δηλαδή στον πληθυσμό από τον οποίο το δείγμα προήλθε να μην υπάρχει καμία τέτοια συσχέτιση. Το κριτήριο αυτό είναι η τιμή p -value, ή υπολογισθείσα πιθανότητα, και εκφράζει την εκτιμώμενη πιθανότητα να απορριφθεί η μηδενική υπόθεση ενός συμπεράσματος όταν η υπόθεση αυτή είναι αληθής. Ο όρος επίπεδο σημαντικότητας ($alpha$), χρησιμοποιείται για να αναφερθούμε στην πιθανότητα που προαποφασίστηκε πριν ελεγχθεί η υπόθεση. Αυτή αντιστοιχεί στην πιθανότητα του σφάλματος τύπου I, ενώ η p -value αναφέρεται στην πιθανότητα που υπολογίστηκε μετά τη συγκεκριμένη μελέτη. Το σφάλμα τύπου I είναι η εσφαλμένη απόρριψη της μηδενικής υπόθεσης, ενώ το σφάλμα τύπου II είναι η εσφαλμένη αποδοχή της.

Η ισχύς ενός στατιστικού ελέγχου ορίζεται ως η διαφορά της μονάδας και της πιθανότητας σφάλματος τύπου II ($beta$) και συνεπώς επιθυμούμε τη μεγιστοποίησή της. Ο ακόλουθος πίνακας δείχνει τη σχέση μεταξύ κατάστασης και σφάλματος στον έλεγχο υποθέσεως.

Πίνακας 2.3: Κατάσταση σφάλματος και έλεγχος υποθέσεως.

Στατιστική Απόφαση	Πραγματική κατάσταση	
	H_0 αληθεύει	H_0 εσφαλμένη
Απόρριψη H_0	Σφάλμα τύπου I	Σωστή απόφαση
Αποδοχή H_0	Σωστή απόφαση	Σφάλμα τύπου II

Σφάλμα τύπου I: Είναι η εσφαλμένη απόρριψη της μηδενικής υπόθεσης. Ως μέγιστη πιθανότητα προκαθορίζεται η τιμή $alpha$. Δεν επηρεάζεται από το μέγεθος του δείγματος καθώς καθορίζεται εκ των προτέρων.

Σφάλμα τύπου II: Είναι η εσφαλμένη αποδοχή της μηδενικής υπόθεσης. Ως μέγιστη πιθανότητα καθορίζεται η τιμή $beta$. Αυτή εξαρτάται από το μέγεθος του δείγματος καθώς

και την τιμή α . Δεν μπορεί να εκτιμηθεί παρά μόνο ως μια συνάρτηση της επίδρασης επί του πραγματικού πληθυσμού. Μειώνεται καθώς αυξάνεται το μέγεθος του δείγματος

Η επιλογή του επιπέδου σημαντικότητας στο οποίο απορρίπτεται η H_0 είναι αυθαίρετη. Τα πλέον συνηθισμένα επίπεδα που χρησιμοποιούνται στη βιβλιογραφία (Johnson, 2013) είναι τα $\{p < 0.05, p < 0.01, p < 0.001\}$. Οι περισσότεροι ερευνητές αναφέρονται στους ελέγχους ως «στατιστικά σημαντικό» για $p < 0.05$ και ως «στατιστικά πολύ σημαντικό» για $p < 0.001$. Συνήθως η τιμή 0.05 ορίζεται ως όριο αποκοπής.

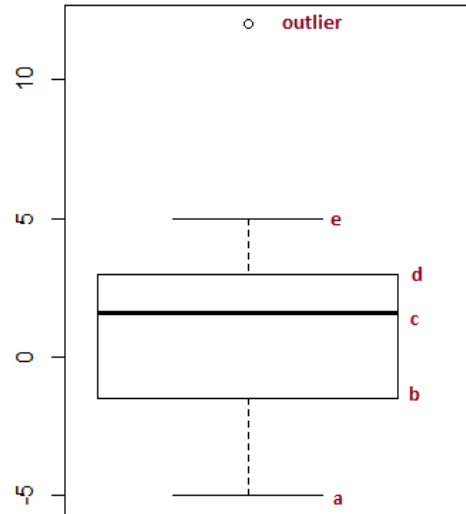
2.3.3.2 Έλεγχος κανονικότητας

Η σημασία της κανονικής κατανομής στη στατιστική ανάλυση είναι αδιαμφισβήτητη αφού αποτελεί την υποκείμενη υπόθεση πολλών στατιστικών διαδικασιών, όπως τα t-τεστ, η γραμμική ανάλυση παλινδρόμησης και η ανάλυση διασποράς (ANOVA). Όταν η υπόθεση ύπαρξης κανονικότητας παραβιάζεται, η ερμηνεία των αντίστοιχων διαδικασιών και τα συμπεράσματά τους μπορεί να μην είναι αξιόπιστα και έγκυρα. Επομένως, είναι σημαντικό να πραγματοποιείται πάντα έλεγχος της κανονικότητας πριν τις σχετικές στατιστικές διαδικασίες. Οι τρεις συνήθεις μέθοδοι για να αξιολογήσουμε εάν ένα δείγμα τυχαίων παρατηρήσεων προέρχεται από έναν πληθυσμό που ακολουθεί κανονική κατανομή είναι οι γραφικές, οι αριθμητικές και οι έλεγχοι υποθέσεων.

Γραφικές μέθοδοι: Το γράφημα κανονικού τεταρτημορίου (Q-Q plot) είναι το επικρατέστερο και πλέον αποτελεσματικό μέσο για να ελεγχθεί γραφικά η κανονικότητα των δεδομένων. Άλλες γραφικές μέθοδοι είναι τα ιστογράμματα και τα box-plots.

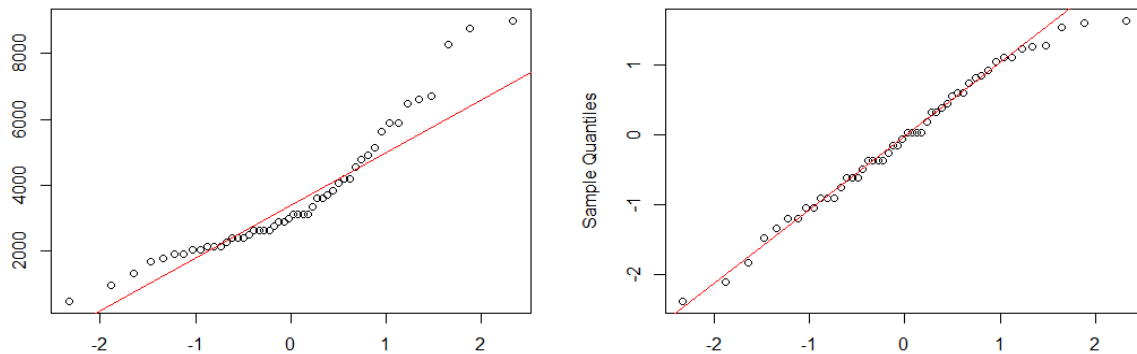
Στην περίπτωση π.χ. των box-plots απεικονίζεται κατά τμήματα η κατανομή των παρατηρήσεων της χρονοσειράς πληροφορώντας μας έτσι σχετικά με την ύπαρξη κάποιας ακραίας τιμής ή κυρτότητας. Για να ακολουθούν τα δεδομένα κανονική κατανομή θα πρέπει λοιπόν οι αποστάσεις $a-b$, $b-c$, $c-d$ και $d-e$ που φαίνονται στην [Εικόνα 2.8](#) να είναι ίσες μεταξύ τους. Σημειώνεται ότι οι τιμές a , b , c , d και e σηματοδοτούν την ελάχιστη τιμή του δείγματος, το 25% των παρατηρήσεων, την ενδιάμεση τιμή, το 75% των παρατηρήσεων και τη μέγιστη τιμή του δείγματος αντίστοιχα. Ως outlier χαρακτηρίζεται οποιαδήποτε παρατήρηση που δεν ανήκει στην κατανομή που ορίζει το υπόλοιπο δείγμα.

Κεφάλαιο 2: Ανάλυση χρονοσειρών και απόδοσης μοντέλων πρόβλεψης



Εικόνα 2.8: Γραφική απεικόνιση της κατανομής των παρατηρήσεων μίας χρονοσειράς με σχετική κυρτότητα.

Αν και οι γραφικές μέθοδοι αποτελούν χρήσιμα εργαλεία για την εποπτεία και την μελέτη των χρονοσειρών, δεν επαρκούν σε καμία περίπτωση όταν επιθυμούμε να οδηγηθούμε σε ασφαλή και τεκμηριωμένα συμπεράσματα περί κανονικότητας. Προς υποστήριξη των γραφικών μεθόδων, πραγματοποιούνται λοιπόν αυστηρότεροι έλεγχοι.



Εικόνα 2.9: Χρήση Q-Q plot για τον έλεγχο της κανονικότητας των δεδομένων μίας χρονοσειράς με παρατηρήσεις κατανεμημένες στο δείγμα της ανομοιόμορφα (αριστερά) και ομοιόμορφα (δεξιά), αντίστοιχα.

Αριθμητικές μέθοδοι: Βάσει αυτών των μεθόδων υπολογίζονται οι δείκτες καμπυλότητας και κυρτότητας και ελέγχεται βάσει κριτηρίων πλέον εάν το δείγμα δεδομένων της χρονοσειράς ακολουθεί την κανονική κατανομή.

Δύο ή περισσότερες κατανομές συχνοτήτων με τον ίδιο αριθμό παρατηρήσεων είναι δυνατό να διαφέρουν μεταξύ τους, παρά το γεγονός ότι έχουν τον ίδιο μέσο, την ίδια

διακύμανση και τον ίδιο συντελεστή ασυμμετρίας. Η αιτία μιας τέτοιας διαφοράς βρίσκεται στον βαθμό συγκέντρωσης των τιμών της μεταβλητής γύρω από τον μέσο και τα άκρα της κατανομής. Για την περιγραφή και την μέτρηση αυτού του χαρακτηριστικού χρησιμοποιούμε τον όρο κύρτωση της κατανομής.

Για παράδειγμα, μια κατανομή μπορεί να είναι συγκεντρωμένη κοντά στη μέση τιμή και κατά συνέπεια η καμπύλη που την απεικονίζει γραφικά να έχει έντονη κύρτωση. Μπορεί όμως να είναι συγκεντρωμένη με τέτοιο τρόπο ώστε η καμπύλη που την παριστάνει να μην παρουσιάζει σαφή κύρτωση. Η κυρτότητα της καμπύλης ορίζεται από τη σχέση:

$$\alpha_4 = \frac{E[(X - \mu)^4]}{\sigma^4} = \frac{\mu_4}{\sigma^4}$$

, όπου α_4 ο συντελεστής κυρτότητας, μ η μέση τιμή της χρονοσειράς και σ η τυπική της απόκλιση. Η κανονική κατανομή χρησιμοποιείται ως αναφορά, με συντελεστή κύρτωσης ίσο με 3. Έτσι, αν

- $\beta_2 > 3$ τότε η κατανομή είναι λεπτόκυρτη
- $\beta_2 = 3$ τότε η κατανομή είναι μεσόκυρτη
- $\beta_2 < 3$ τότε η κατανομή είναι πλατύκυρτη

Προφανώς για να είναι μία καμπύλη κανονική θα πρέπει να είναι και μεσόκυρτη.

Έλεγχοι υποθέσεων: Τα τεστ ή αλλιώς έλεγχοι υποθέσεων χρησιμοποιούνται προκειμένου να ελεγχθεί με στατιστική σημαντικότητα η πιθανότητα η χρονοσειρά που εξετάζεται να ικανοποιεί μία υπόθεση και στη συγκεκριμένη περίπτωση να είναι κανονική. Ακολουθούν την λογική των υποθέσεων που παρουσιάστηκε νωρίτερα και από τα πλέον γνωστά είναι αυτά των *Shapiro-Wilk*, *Kolmogorov-Smirnoff*, *Anderson-Darling* και του *Lilliefors*.

Σημειώνεται ότι οι παραπάνω έλεγχοι κανονικότητας συχνά παράγουν διαφορετικά συμπεράσματα αφού για το ίδιο δείγμα δεδομένων ενδέχεται κάποιο τεστ να απορρίψει τη μηδενική υπόθεση ενώ ένα άλλο να την επιβεβαιώσει. Έτσι, οι *Razali και Wah (2011)* πραγματοποίησαν μέσω προσομοιώσεων Μόντε Κάρλο μια συγκριτική μελέτη ισχύος για τους τέσσερις ελέγχους για διάφορες τυποποιημένες κατανομές προκειμένου να εξακριβώσουν τις αποτελεσματικότητά τους. Η μελέτη πραγματοποιήθηκε έχοντας ως μηδενική υπόθεση ότι η κατανομή είναι κανονική και τα αποτελέσματα έδειξαν ότι το τεστ των *Shapiro & Wilk (1965)* έχει τη μεγαλύτερη ισχύ ανεξαρτήτως του μεγέθους του δείγματος. Το εν λόγω συμπέρασμα συμφωνεί με τα αποτελέσματα αντίστοιχων μελετών των *Mendes Pala (2003)*, *Keskin (2006)* και *Farrell & Stewart (2006)* τα οποία επίσης εμφανίζουν τον έλεγχο *Shapiro-Wilk* ως τον πλέον ισχυρό για έλεγχο της υπόθεσης κανονικότητας.

Το τεστ Shapiro-Wilk: Έστω Y_1, Y_2, \dots, Y_n δείγμα n παρατηρήσεων πάνω στην τυχαία μεταβλητή Y , της οποίας η άγνωστη συνάρτηση κατανομής είναι $F_Y(Y)$, $Y \in \mathbb{R}$. Οι προς έλεγχο υποθέσεις είναι οι εξής:

- H_0 : η $F_Y(Y)$ είναι η συνάρτηση κατανομής της κανονικής κατανομής με άγνωστη μέση τιμή και άγνωστη διασπορά
- H_1 : η $F_Y(Y)$ είναι η συνάρτηση κατανομής μίας μη κανονικής κατανομής

Η στατιστική συνάρτηση για τον έλεγχο των υποθέσεων αυτών είναι η

$$W = \frac{[\sum_{i=1}^k a_i (Y^{(n-i+1)} - Y^{(i)})]^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

,όπου $Y^{(i)}$ είναι η i παρατήρηση του διατεταγμένου κατά αύξουσα τάξη μεγέθους δείγματος, k είναι ένας ακέραιος αριθμός (συνήθως ίσος με $n/2$) και a_i με $i = 1, 2, \dots, k$ είναι σταθεροί συντελεστές. Παρατηρούμε ότι οι μικρές τιμές της στατιστικής συνάρτησης W αποτελούν ένδειξη ότι η μηδενική υπόθεση δεν είναι αληθής.

Επομένως, ο κανόνας απόφασης είναι ο εξής: Η μηδενική υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α εάν η τιμή της στατιστικής συνάρτησης W είναι μικρότερη από το α -ποσοστιαίο σημείο της κατανομής της. Ένα πολύ χρήσιμο χαρακτηριστικό του ελέγχου *Shapiro-Wilk* είναι ότι αρκετοί ανεξάρτητοι έλεγχοι καλής προσαρμογής μπορούν να συνδυασθούν (ενοποιηθούν) σε έναν ενιαίο έλεγχο κανονικότητας. Αυτό βοηθά πολύ στην περίπτωση όπου αρκετά μικρά δείγματα από, ενδεχομένως, διαφορετικούς πληθυσμούς είναι ανεπαρκή από μόνα τους να οδηγήσουν σε απόρριψη της υπόθεσης της κανονικότητας, αλλά συνδυαζόμενα παρέχουν ενδείξεις που είναι αρκετές για την απόρριψη της υπόθεσης της κανονικότητας.

2.3.3.3 Έλεγχος στασιμότητας

Σύμφωνα με τους *Nelson & Plosser (1982)* πολλές χρονικές σειρές περιέχουν μοναδιαίες ρίζες που χαρακτηρίζονται από την ύπαρξη στοχαστικών τάσεων. Έτσι, οι έλεγχοι μοναδιαίας ρίζας καθίστανται σημαντικοί στην εξέταση της στασιμότητας των χρονοσειρών. Η μοναδιαία ρίζα μπορεί να ελεγχθεί χρησιμοποιώντας τον επαυξημένο έλεγχο *Dickey-Fuller*, γνωστό και ως *ADF test* (*Said & Dickey, 1984*).

Το **τεστ ADF** αναφέρεται στο στατιστικό κριτήριο t του συντελεστή δ_2 της παρακάτω παλινδρόμησης

$$\Delta Y_t = \delta_0 + \delta_1 t + \delta_2 Y_{t-1} + \sum_{i=1}^k a_i \Delta Y_{t-1} + u_t$$

και ελέγχει για ύπαρξη μοναδιαίας ρίζας στην Y_t , δηλαδή στο λογάριθμο όλων των μεταβλητών του υποσυνόλου της χρονοσειράς Y τη χρονική στιγμή t . Η μεταβλητή ΔY_{t-1} εκφράζει τις πρώτες διαφορές με k χρονική υστέρηση ενώ η μεταβλητή u_t προσαρμόζει τα σφάλματα της αυτοσυσχέτισης. Οι συντελεστές δ_0 , δ_1 , δ_2 και a_i είναι προς εκτίμηση. Σημειώνεται ότι για να καθοριστεί η καλύτερη δυνατή εξίσωση μπορεί να χρησιμοποιηθεί η πρόταση των *Engle and Yoo (1987)* και το κριτήριο του Akaike (*AIC*). Η κατανομή του ελέγχου του επαυξημένου *Dickey-Fuller* είναι μη κανονική και οι κρίσιμες τιμές που

χρησιμοποιούνται προκύπτουν από τους πίνακες του Mackinnon. Οι προς έλεγχο υποθέσεις είναι οι εξής:

- H_0 : Ο συντελεστής παλινδρόμησης δ_2 έχει μηδενική τιμή
- H_1 : Ο συντελεστής παλινδρόμησης δ_2 έχει αρνητική τιμή

Αφού η μηδενική υπόθεση στον επαυξημένο έλεγχο *Dickey-Fuller* είναι ότι μια χρονική σειρά περιέχει μοναδιαία ρίζα (έχει τάση και δεν είναι στάσιμη), αυτή γίνεται αποδεκτή μόνο όταν δεν υπάρχει μια δυναμική απόδειξη εναντίον της. Έτσι, η εν λόγω προσέγγιση έχει χαμηλότερη ισχύ έναντι των σχεδόν στάσιμων διαδικασιών μοναδιαίας ρίζας, ελέγχων δηλαδή που εξετάζουν αν ισχύει ακριβώς το αντίθετο. Προς αυτή την κατεύθυνση οι *Kwiatkowski, Phillips, Schmidt και Shin (1992)* παρουσίασαν έναν εναλλακτικό έλεγχο σύμφωνα με τον οποίο η μηδενική υπόθεση αναφέρει ότι η χρονοσειρά είναι στάσιμη, γνωστό και ως *KPSS test*.

Το **τεστ KPSS** ελέγχει λοιπόν τη στασιμότητα μιας χρονοσειράς είτε ως προς το επίπεδο είτε ως προς την τάση της. Υποθέτουμε δηλαδή ότι μια χρονική σειρά μπορεί να αποτελείται από μια προσδιοριστική χρονική τάση, τον τυχαίο περίπατο, και το σφάλμα στασιμότητας ως εξής:

$$Y_t = \delta_t + r_t + \varepsilon_t$$

,όπου r_t είναι ο τυχαίος περίπατος.

Οι προς έλεγχο υποθέσεις είναι οι εξής:

- H_0 : Η χρονοσειρά είναι στάσιμη γύρω από μια σταθερά ($\delta=0$)
- H_1 : Η χρονοσειρά εμφανίζει χρονική τάση ($\delta \neq 0$)

2.3.3.4 Έλεγχος εποχιακότητας

Ο έλεγχος εποχιακότητας είναι ιδιαίτερα σημαντικός κατά τη πρόβλεψη των χρονοσειρών καθώς καθορίζει αν θα υποθέσουμε κάποιο εποχιακό μοτίβο. Αυτό είναι ιδιαίτερα σημαντικό καθώς αν δεν λάβουμε υπόψη την εποχιακότητα σε μία εποχιακή χρονοσειρά, το μοντέλο πρόβλεψης θα αντιμετωπίσει την συστηματικά εποχιακή διακύμανση ως θόρυβο μη αξιοποιώντας έτσι τη διαθέσιμη πληροφορία. Σε περίπτωση πάλι που υποθέσουμε εποχιακότητα για μία μη εποχιακή χρονοσειρά, το μοντέλο θα εισάγει εποχιακότητα στις προβλέψεις, κάτι που δεν θα συνάδει με την πραγματικότητα. Σε οποιαδήποτε από τις δύο περιπτώσεις αναμένεται λοιπόν να μειωθεί σημαντικά η ακρίβεια πρόβλεψης.

Ο έλεγχος εποχιακής συμπεριφοράς εφαρμόζεται μελετώντας την αυτοσυσχέτιση των δεδομένων με περίοδο υστέρησης k ίση με τον αριθμό των περιόδων ενός κύκλου εποχιακότητας p σε σύγκριση με τις αυτοσυσχετίσεις περιόδου υστέρησης έως και μιας μικρότερης από τον αριθμό των περιόδων ενός κύκλου εποχιακότητας (*Πετρόπουλος & Ασημακόπουλος, 2011*). Μια χρονοσειρά θεωρείται λοιπόν εποχιακή αν ισχύει:

$$|ACF_{pos}| > Limit \text{ όπου}$$

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

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

, όπου Y είναι το διάνυσμα των αρχικών παρατηρήσεων της χρονοσειράς, \bar{Y} είναι η μέση τιμή της και n είναι το μήκος της. Η τιμή του συντελεστή t -critical καθορίζεται από το επίπεδο εμπιστοσύνης που επιθυμούμε για τον έλεγχο εποχιακής συμπεριφοράς σύμφωνα με τον πίνακα που ακολουθεί. Προφανώς, όσο αυξάνει το επίπεδο εμπιστοσύνης τόσο μεγαλύτερη θα πρέπει να είναι η συσχέτιση των παρατηρήσεων με υστέρηση k προκειμένου η χρονοσειρά να κριθεί ως εποχιακή. Συνήθως επιλέγεται το επίπεδο εμπιστοσύνης 90% ούτως ώστε να εξασφαλίζεται επαρκής σημαντικότητα στα συμπεράσματά μας χωρίς ωστόσο να γίνεται απαγορευτικός ο χαρακτηρισμός μίας χρονοσειράς ως εποχιακή.

Με τον έλεγχο εποχιακότητας μπορούμε στη συνέχεια να επιλέξουμε αν θα πρέπει να αποσυνθέσουμε τη χρονοσειρά ούτως ώστε να την αποεποχικοποιήσουμε αποτελεσματικά, απαλείφοντας το συντελεστή εποχιακότητας.

Πίνακας 2.4: Τιμές του δείκτη t -critical σε συνάρτηση με το διαστήματα εμπιστοσύνης.

Επίπεδο εμπιστοσύνης	$t_{critical}$
80%	1.280
90%	1.645
95%	1.960
98%	2.330
99%	2.580

2.4 Διαφόριση και μετασχηματισμοί χρονοσειρών

Η ύπαρξη στασιμότητας σε μία χρονοσειρά, δηλαδή η διατήρηση στην ασθενή της μορφή των τιμών των ροπών πρώτης και δεύτερης τάξης, αποτελεί σημαντικό πλεονέκτημα κατά τη διαδικασία της πρόβλεψης. Αυτό συμβαίνει καθώς μία στάσιμη χρονοσειρά δεν διαθέτει ούτε τάση ούτε εποχιακότητα και συνεπώς καλούμαστε να προσδιορίσουμε μόνο το τρέχον επίπεδό της μειώνοντας έτσι την πολυπλοκότητά του μοντέλου πρόβλεψης και περιορίζοντας την αβεβαιότητα κατά την παραμετροποίησή του. Μάλιστα, για συγκεκριμένες οικογένειες μοντέλων πρόβλεψης όπως τα $ARMA$ η στασιμότητα αποτελεί απαραίτητη προϋπόθεση.

Το ίδιο ισχύει και για την περίπτωση της κανονικότητας καθώς αυτή εξασφαλίζει ότι καμία από τις παρατηρήσεις της χρονοσειράς δεν αποτελεί *outlier* και συνεπώς το μοντέλο είναι σχετικά απίθανο να επηρεαστεί σημαντικά από παροδικές διακυμάνσεις και σποραδικά ακραίες τιμές. Η έλλειψη στασιμότητας και κανονικότητας αποτελούν λοιπόν από τα πρώτα προβλήματα που πρέπει να αντιμετωπιστούν κατά την προετοιμασία των χρονοσειρών για την εξασφάλιση των οποίων μπορούν να εφαρμοστούν αντίστοιχα *διαφόριση και μετασχηματισμοί*.

2.4.1 Διαφόριση

Ως διαφόριση (*differencing*) ορίζεται η διαδικασία κατά την οποία, χρησιμοποιώντας τα αρχικά δεδομένα μίας χρονοσειράς μήκους n , παράγεται μία νέα με στοιχεία της τις διαδοχικές μεταβολές των αρχικών παρατηρήσεων. Η διαφόριση έχει ως αποτέλεσμα τον περιορισμό των διακυμάνσεων επιπέδου αφαιρώντας τάση ή/και εποχιακότητα και την παραγωγή μίας χρονοσειράς σταθερού επιπέδου και διακύμανσης. Αν η διαδικασία της διαφόρισης στεφθεί με επιτυχία θα πρέπει λοιπόν η χρονοσειρά που προκύπτει να είναι στάσιμη και, πέραν της σχέσης που συνδέει τις επιμέρους παρατηρήσεις, να χαρακτηρίζεται αποκλειστικά από τυχαιότητα. Ανάλογα την τάξη διαφόρισης έχουμε:

- 1^η τάξη: $y'_t = y_t - y_{t-1}$
- 2^η τάξη: $y''_t = y'_t - y'_{t-1} = y_t - 2y_{t-1} + y_{t-2}$
- n ^η τάξη: $y^{(n)}_t = y^{(n-1)}_t - y^{(n-1)}_{t-1}$

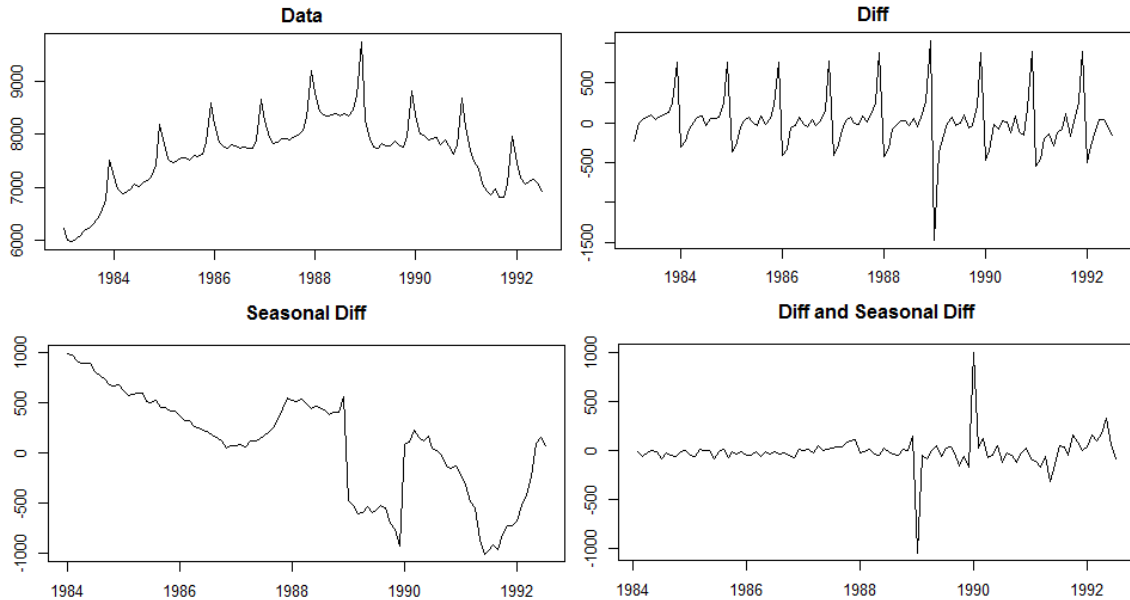
Από τις παραπάνω σχέσεις φαίνεται ότι οι χρονοσειρές που προκύπτουν από μία διαφόριση n ^{ης} τάξης θα έχουν $n-n$ παρατηρήσεις. Προκειμένου λοιπόν να υπάρχει ικανός αριθμός παρατηρήσεων για την παραγωγή αξιόπιστων προβλέψεων η διαφόριση θα πρέπει να γίνεται με σύνεση. Συνεπώς αν και θεωρητικά μπορεί να επιλεγθεί διαφόριση μέχρι και $n-1$ τάξης, πρακτικά χρησιμοποιείται διαφόριση μέχρι δεύτερου βαθμού. Αυτό συμβαίνει καθώς η διαφόριση πρώτης τάξης επαρκεί για την απομάκρυνση γραμμικής τάσης, ενώ η δεύτερη για την εξάλειψη εκθετικής (τάση εντός τάσης). Αν η τάση έχει απομακρυνθεί και η χρονοσειρά δεν είναι εποχιακή, τότε το πιο πιθανό είναι ότι η χρονοσειρά θα έχει αποκτήσει επαρκή στασιμότητα.

Σε περίπτωση εποχιακής χρονοσειράς κατ' αναλογία με την απλή διαφόριση μπορεί να εφαρμοστεί και εποχιακή διαφόριση διαφορίζοντας κάθε παρατήρηση με αυτήν της προηγούμενης περιόδου ως εξής:

$$y'_t = y_t - y_{t-m}$$

, όπου m η συχνότητα της χρονοσειράς.

Κεφάλαιο 2: Ανάλυση χρονοσειρών και απόδοσης μοντέλων πρόβλεψης



Εικόνα 2.10: Εποχιακή μηνιαία χρονοσειρά (Data) πριν και μετά τη συνδυαστική διαφορίση πρώτης τάξης (Diff) και εποχιακής διαφορίσης (Seasonal Diff).

Όπως φαίνεται η απλή διαφορίση διατηρεί την εποχιακότητα, η εποχιακή διατηρεί την τάση και ο συνδυασμός τους οδηγεί σε ικανοποιητική στασιμότητα.

Αν οριστεί ως B ο τελεστής ολίσθησης ούτως ώστε $By_t = y_{t-1}$ και $B(By_t) = B^2y_t = y_{t-2}$, τότε η διαφορίση n τάξης παρουσιάζεται ως $(1-B)^n y_t$ και η εποχιακή διαφορίση m τάξης N ως $(1-B^m)^N y_t$. Σε περίπτωση συνδυασμού των δύο τύπων διαφορίσεων λαμβάνουμε αντίστοιχα $(1-B^m)^N (1-B)^n y_t$. Σημειώνεται ότι λόγω της φύσης της εποχιακής διαφορίσης, νόημα διαθέτει αποκλειστικά αυτή της πρώτης τάξης. Στην [Εικόνα 2.10](#) δίνεται ένα χαρακτηριστικό παράδειγμα διαφορίσης μίας εποχιακής χρονοσειράς με τάση.

Για την ενίσχυση των ισχυρισμών σχετικά με την εγκρατή εφαρμογή της διαδικασίας της διαφορίσης αναφέρεται επιπρόσθετα ότι η διαφορίση μειώνει σημαντικά την αυτοσυσχέτιση των χρονοσειρών αυξάνοντας την τυχαιότητα του μοντέλου. Έτσι, σε περιπτώσεις υπερδιαφορίσης (*over-differencing*) προκειμένου να επιτευχθούν ακριβείς προβλέψεις θα απαιτηθεί η χρήση πολύπλοκων μεθόδων, κάτι το οποίο βέβαια δεν είναι επιθυμητό.

2.4.2 Μετασχηματισμοί

Προκειμένου να περιοριστεί η διασπορά και η τυχαιότητα μίας χρονοσειράς και να διευκολυνθεί η πρόβλεψή της, συνίσταται συχνά η εφαρμογή κάποιου μετασχηματισμού που θα οδηγήσει σε κανονικότητα. Το αν θα πρέπει βέβαια να εφαρμοστεί ένας μετασχηματισμός, καθώς και το ποιος, δεν είναι σε καμία περίπτωση σαφώς ορισμένο. Έτσι, συνήθως επιλέγουμε να εφαρμόσουμε κάποιον έλεγχο κανονικότητας (π.χ. τεστ *Shapiro-Wilk*) και ανάλογα το αποτέλεσμα του προχωρούμε ή όχι στην εφαρμογή μίας οικογένειας μετασχηματισμών που έχουμε προκαθορίσει. Ο μετασχηματισμός μπορεί να είναι ένας απλός *λογάριθμος* επί των δεδομένων, ενώ άλλες φορές μπορεί να επεκταθεί σε αυτούς των *Box-Cox (1964)* και *Johnson (1949)*. Άλλες οικογένειες μετασχηματισμών

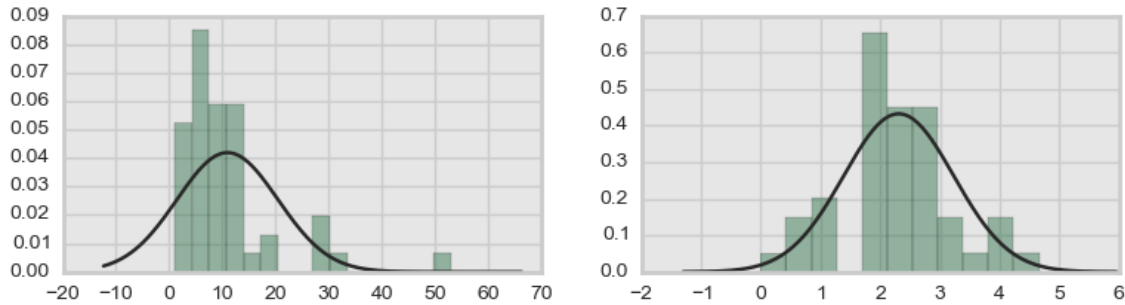
σχετίζονται με τα σφάλματα που παράγουν τα μοντέλα πρόβλεψης όταν εφαρμοστούν επί των δεδομένων, ωστόσο επειδή αυτά συνδέονται άμεσα με το εκάστοτε μοντέλο πρόβλεψης και δεν αποτελούν γενικές λύσεις δεν θα παρουσιαστούν στην παρούσα εργασία. Σημειώνεται επίσης ότι σε αρκετές περιπτώσεις προκειμένου να περιοριστεί η διασπορά των δεδομένων εφαρμόζονται μετασχηματισμοί ανεξαρτήτως κανονικότητας.

Λογαρίθμηση: Η απλούστερη μορφή μετασχηματισμού είναι η λογαρίθμηση της χρονοσειράς, η εφαρμογή δηλαδή λογαρίθμων επί των αρχικών δεδομένων. Λόγω της μη γραμμικής συμπεριφοράς της συνάρτησης του λογαρίθμου ο μετασχηματισμός έχει ως αποτέλεσμα να περιορίζονται οι ακραίες τιμές της χρονοσειράς και να ορίζεται ένα σαφέστερο επίπεδο. Φυσικά η ύπαρξη τιμών μικρότερων ή ίσων του μηδέν αποτελεί έναν περιορισμό στην εν λόγω προσέγγιση, ενώ σε περίπτωση ασυνήθιστα μικρών τιμών σε σχέση με το υπόλοιπο δείγμα η λογαρίθμηση είναι πολύ πιθανό να τις υπερτονίσει αντί να τις εξαλείψει. Για αυτόν το σκοπό μπορεί κανείς να επιλέξει άλλους καταλληλότερους μη γραμμικούς μετασχηματισμούς, όπως είναι οι μετασχηματισμοί Box-Cox.

Μετασχηματισμός Box-Cox: Οι μετασχηματισμοί των Box-Cox, γνωστοί και ως *power transformation*, βασίζονται στην ύψωση της χρονοσειράς σε μία δύναμη τέτοια ώστε να επιτευχθεί κανονικότητα. Έτσι, περιορίζεται όσο είναι εφικτό η διασπορά των παρατηρήσεων και αυτές συγκεντρώνονται ομοιόμορφα γύρω από μία μέση τιμή. Η ζητούμενη δύναμη είναι γνωστή ως λ (*lambda*) και ο μετασχηματισμός δίνεται από τη σχέση:

$$x_t = \frac{y_t^\lambda - 1}{\lambda}$$

, όπου y_t και x_t είναι η τιμή της χρονοσειράς τη χρονική στιγμή t πριν και μετά τον μετασχηματισμό. Βάσει του ορισμού του φαίνεται ότι πέρα από πιο πολύπλοκες επιλογές ο μετασχηματισμός μπορεί να οδηγήσει και σε απλές τροποποιήσεις των δεδομένων όπως η χρήση τετραγωνικής ή κυβικής ρίζας ($\lambda=0.5$ και $\lambda=1/3$), η αντιστροφή της χρονοσειράς ($\lambda=-1$) ή ακόμα και στη διατήρηση της αρχικής χρονοσειράς ($\lambda=1$). Σημειώνεται επίσης ότι για $\lambda=0$ ο μετασχηματισμός ταυτίζεται με τη λογαρίθμηση των δεδομένων. Απαραίτητη προϋπόθεση και σε αυτήν την περίπτωση είναι η θετική τιμή των παρατηρήσεων κατά μήκος όλου του σετ δεδομένων. Στην *Εικόνα 2.11* φαίνεται ένα παράδειγμα κανονικοποίησης των δεδομένων μίας μη κανονικής χρονοσειράς με υψηλή διασπορά.



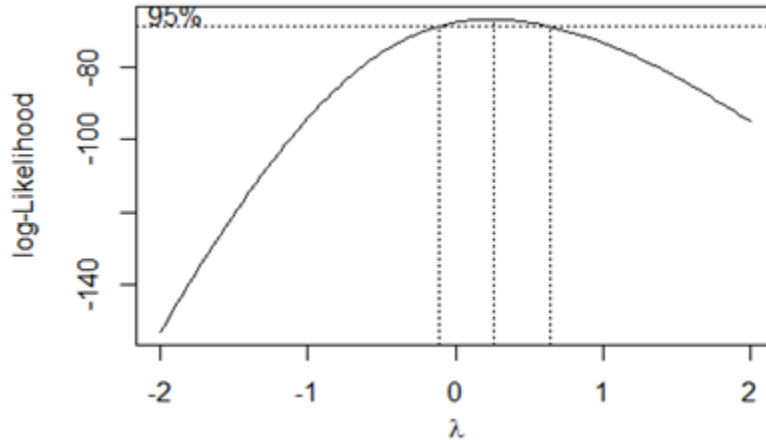
Εικόνα 2.11: Σύγκριση γραφήματος της περιθώριας κατανομής μίας μη κανονικής χρονοσειράς πριν (αριστερά) και μετά (δεξιά) την εφαρμογή του μετασχηματισμού Box-Cox.

Σημειώνεται ότι η κανονικοποίηση μέσω ενός μετασχηματισμού δεν είναι απαραίτητο ότι θα είναι πάντα πετυχημένη, δηλαδή ο μετασχηματισμός ενδέχεται να μη μετατρέψει οποιαδήποτε περιθώρια κατανομή σε κανονική. Ακόμα όμως και αν γίνει κάτι τέτοιο ενδέχεται επίσης να μην οδηγήσει σε καλύτερες προβλέψεις. Ο πρώτος ισχυρισμός σχετίζεται με τους στατιστικούς ελέγχους κανονικότητας που εφαρμόζονται για τον καθορισμό της τιμής λ , τα συμπεράσματα των οποίων δεν είναι πάντα αληθή και επηρεάζονται από αρκετούς παράγοντες. Επίσης, δεδομένου ότι οι τιμές λ που εξετάζονται εκτείνονται συνήθως στο διάστημα $\{-2,2\}$, μπορεί να μην υπάρχει ικανό λ εντός διαστήματος που να οδηγεί σε ικανοποιητική κανονικοποίηση. Ο δεύτερος ισχυρισμός σχετίζεται με την τυχαιότητα της χρονοσειράς, η οποία και για να επιχειρείται μετασχηματισμός αναμένεται να είναι αρκετά υψηλή. Έτσι, ακόμα και αν οι προβλέψεις που παράγει κανείς επί της μετασχηματισμένης χρονοσειράς είναι πιο σταθερές, το μέλλον μπορεί να είναι τέτοιο που να τις αναδειξει ατυχείς.

Πέρα από τους ελέγχους κανονικότητας μπορεί κανείς να χρησιμοποιήσει και άλλες μεθόδους για τον καθορισμό της βέλτιστης παραμέτρου λ . Χαρακτηριστικό παραδείγματα αποτελεί η μεθοδολογία του *Guerrero (1993)* που χρησιμοποιεί ως κριτήριο την ελαχιστοποίηση του συντελεστή μεταβλητότητας. Άλλη μέθοδος είναι να ευρεθεί εκείνο το λ για το οποίο ένα γραμμικό μοντέλο περνάει βέλτιστα από τη χρονοσειρά που προκύπτει, δηλαδή ελαχιστοποιεί τα τετραγωνικά σφάλματα γραμμής-δεδομένων. Οι ίδιοι *Box και Cox (Venables & Ripley, 2002)* προτείνουν ωστόσο την επιλογή του λ μέσω της ελαχιστοποίησης της συνάρτησης πιθανοφάνειας (*log-likelihood*) η οποία δίνεται από τον τύπο:

$$\widehat{L}(\lambda) = \bar{y} - \frac{n}{2} \log RSS \left(\frac{y^\lambda}{\hat{y}^{\lambda-1}} \right), \text{ όπου } \hat{y} = \sqrt[n]{y_1 y_2 \dots y_n}$$

, όπου y_i η παρατήρηση της υπό εξέταση χρονοσειράς τη στιγμή i , n το μήκος της, \bar{y} η μέση της τιμή και RSS (*Residual Sum of Squares*) το άθροισμα των τετραγωνικών σφαλμάτων γραμμής παλινδρόμησης-δεδομένων. Εφαρμόζοντας την εν λόγω συνάρτηση για διάφορες τιμές λ από ένα διάστημα γίνεται εφικτή λοιπόν η επιλογή της καταλληλότερης που μεγιστοποιεί και την εν λόγω πιθανοφάνεια. Στην *Εικόνα 2.12* φαίνεται ένα παράδειγμα επιλογής παραμέτρου λ στο διάστημα $\{-2,2\}$.



Εικόνα 2.12: Παράδειγμα επιλογής βέλτιστης παραμέτρου λ με χρήση της μεθόδους των Box-Cox.

Αν και βέλτιστη τιμή θεωρείται η 0.25, με διάστημα εμπιστοσύνης 95% τιμές από -0.1 έως 0.6 μπορούν να ικανοποιήσουν τη συνθήκη κανονικότητας.

Σημειώνεται επίσης ότι το πρόβλημα της απροσδιοριστίας της γενικής μορφής του μετασχηματισμού Box-Cox μπορεί να λυθεί μέσω της επέκτασης της εξίσωσής του ως εξής:

$$x_t = \begin{cases} \frac{(y_t + 1)^\lambda - 1}{\lambda} & \lambda \neq 0, y \geq 0 \\ \log(y_t + 1) & \lambda = 0, y \geq 0 \\ -[(-y_t + 1)^{2-\lambda} - 1]/(2 - \lambda) & \lambda \neq 2, y < 0 \\ -\log(-y_t + 1) & \lambda = 2, y < 0 \end{cases}$$

Μετασχηματισμός Johnson: Ένας άλλος ιδιαίτερα ισχυρός μετασχηματισμός είναι αυτός του Johnson. Βασίζεται στη χρήση τριών εξισώσεων κατανομών και τον έλεγχο των τιμών που πρέπει να πάρουν οι παράμετροι που τις ορίζουν προκειμένου η νέα χρονοσειρά που προκύπτει να ακολουθεί την κανονική κατανομή. Αυτές είναι οι εξής:

$$x_t = \gamma + \eta \ln\left(\frac{y_t - \varepsilon}{\lambda + \varepsilon - \chi}\right), \text{Bounded system (SB)}$$

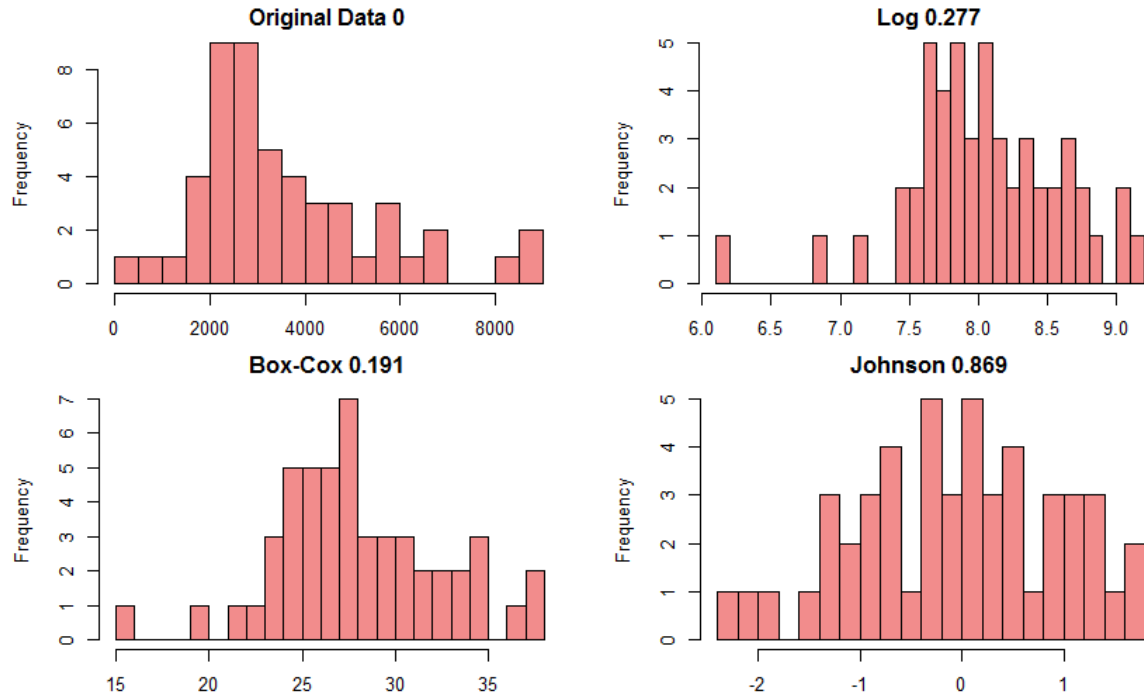
$$x_t = \gamma + \eta \ln(y_t - \varepsilon), \text{Log - normal system (SL)}$$

$$x_t = \gamma + \eta \sinh^{-1}\left(\frac{y_t - \varepsilon}{\lambda}\right), \text{Unbounded system (SU)}$$

Ο έλεγχος της κανονικότητας γίνεται συνήθως μέσω του τεστ *Shapiro-Wilk*, αν και συχνά αντικαθίσταται από αυτό του *Anderson-Darling*. Στόχος είναι να πληροφορηθούμε πρώτον για το αν η χρονοσειρά χρειάζεται να μετασχηματιστεί (δεν ακολουθεί την κανονική κατανομή), και δεύτερον για το ποια από τις τρεις εξισώσεις πρέπει να χρησιμοποιηθεί αν δεν ισχύει η υπόθεση κανονικότητας. Ο μετασχηματισμός δεν είναι ιδιαίτερα διαδεδομένος λόγω των πολλών παραμέτρων που πρέπει να υπολογιστούν για

Κεφάλαιο 2: Ανάλυση χρονοσειρών και απόδοσης μοντέλων πρόβλεψης

την εφαρμογή του, ωστόσο οδηγεί σε ιδιαίτερα καλά αποτελέσματα και σε αντίθεση με τη γενική μορφή των μετασχηματισμών Box-Cox, ορίζεται για οποιαδήποτε τιμή παρατήρησης. Στην *Εικόνα 2.13* αντιπαραβάλλεται το αποτέλεσμα της μεθόδου Johnson με αυτό της απλής λογαρίθμησης και του βέλτιστου μετασχηματισμού Box-Cox.



Εικόνα 2.13: Αποτέλεσμα της εφαρμογής μετασχηματισμών σε μία χρονοσειρά που ακολουθεί μη κανονική κατανομή ($p\text{-value}=0$).

Όπως φαίνεται, ο μετασχηματισμός Johnson καθιστά τη χρονοσειρά σχεδόν κανονική ($p\text{-value}=0.869$) και ακολουθεί ο μετασχηματισμός του λογαρίθμου και των Box-Cox.

Προκειμένου να αναδειχτεί η σημασία του μετασχηματισμού δεδομένων στη βελτίωση της προβλεπτικής ακρίβειας, στον *Πίνακα 2.5* παρουσιάζεται η επίδοση πέντε κλασικών μεθόδων πρόβλεψης στα δεδομένα των διαγωνισμών *M1* και *M3* (4004 χρονοσειρές) με και χωρίς την εφαρμογή κάποιας επεξεργασίας για τον περιορισμό και την σταθεροποίηση της διακύμανσης. Πιο συγκεκριμένα, ελέγχεται η επίδοση των μεθόδων *Naive*, *Ses*, *Holt*, *Damped* και *Theta* σε δεδομένα τα οποία, αφού έχουν αποεποχικοποιηθεί, στη συνέχεια είτε δεν υπόκεινται σε επεξεργασία, είτε λογαριθμούνται, είτε εφαρμόζεται ο βέλτιστος μετασχηματισμός Box-Cox με τιμή λ στο διάστημα $[-1, 1]$ βάσει της μεθοδολογίας επιλογής του *Guerrero*.

Όπως παρατηρείται, πέραν των ετήσιων χρονοσειρών, όπου η μη επεξεργασία των δεδομένων φαίνεται να αποτελεί την πλέον αποδοτική λύση για τις περισσότερες μεθόδους πρόβλεψης, σε όλες τις άλλες συχνότητες η χρήση μετασχηματισμών επιδρά ιδιαίτερα θετικά στην βελτίωση της ακρίβειας πρόβλεψης, η οποία και εδώ μετράτε μέσω του δείκτη σφάλματος *sMAPE*. Ιδιαίτερο ενδιαφέρον έχει επίσης το γεγονός ότι η απλοποιημένη προσέγγιση της λογαρίθμησης οδηγεί σε καλύτερα αποτελέσματα από το

γενικευμένο μετασχηματισμό του Box-Cox, ο οποίος και θεωρητικά επιλέγει τη βέλτιστη παράμετρο μετασχηματισμού.

Πίνακας 2.5: Προβλεπτική ακρίβεια κλασικών μοντέλων πρόβλεψης (sMAPE) για κλασική αποεποχικοποίηση των δεδομένων των διαγωνισμών M1 και M3 και εφαρμογή μετασχηματισμών.

Δεδομένα	Naive	SES	Holt	Damped	Theta
Αρχικά δεδομένα ($\lambda = 1$)					
M3 Ετήσια	17.88	17.77	19.25	17.26	16.73
M3 Τριμηνιαία	9.97	9.78	11.28	9.75	9.28
M3 Μηνιαία	16.71	14.26	15.78	13.96	13.86
M3 Άλλα	6.30	6.28	4.69	4.28	4.92
M1 Ετήσια	22.43	23.10	17.00	18.02	20.18
M1 Τριμηνιαία	17.90	17.69	19.13	18.45	16.28
M1 Μηνιαία	16.53	14.99	18.80	15.83	16.52
Σύνολο	15.46	14.32	15.70	13.97	13.89
Λογαρίθμηση ($\lambda = 0$)					
M3 Ετήσια	17.88	17.96	19.61	17.35	16.11
M3 Τριμηνιαία	9.97	9.99	10.95	9.77	9.42
M3 Μηνιαία	16.71	14.01	14.74	13.78	13.47
M3 Άλλα	6.30	6.22	4.45	4.20	5.30
M1 Ετήσια	22.43	23.38	18.90	17.57	19.25
M1 Τριμηνιαία	17.90	17.75	19.55	16.73	16.06
M1 Μηνιαία	16.53	14.98	17.52	15.69	14.69
Σύνολο	15.46	14.31	15.22	13.80	13.36
Μετασχηματισμός Box-Cox ($\lambda \in [-1, 1]$)					
M3 Ετήσια	17.88	17.78	20.98	18.26	16.77
M3 Τριμηνιαία	9.97	9.74	11.36	9.98	9.23
M3 Μηνιαία	16.71	14.04	14.84	13.80	13.49
M3 Άλλα	6.30	6.27	4.51	4.33	5.02
M1 Ετήσια	22.43	23.17	18.52	17.42	18.12
M1 Τριμηνιαία	17.90	17.71	21.28	18.49	16.07
M1 Μηνιαία	16.53	14.95	17.55	15.70	14.80
Σύνολο	15.46	14.23	15.63	14.08	13.39

Το παραπάνω πείραμα επιβεβαιώνει γενικεύσεις που θέλουν απλούστερες μορφές μετασχηματισμών να αποδίδουν το ίδιο ή και καλύτερα σε σχέση με πιο πολύπλοκες λύσεις, και τους μετασχηματισμούς να βοηθάνε συνολικά στη μείωση ή έστω στη διατήρηση του σφάλματος σε ικανοποιητικά επίπεδα. Κοινώς, κανείς δεν μπορεί να γνωρίζει εκ των προτέρων αν ένας μετασχηματισμός θα οδηγήσει μεμονωμένα σε αύξηση της προβλεπτικής ακρίβειας, σίγουρα μπορεί ωστόσο να επιφέρει βελτιώσεις συνολικά και μάλιστα μέσω χρήσης απλοϊκών μεθόδων. Αυτό πρακτικά ερμηνεύεται αφενός από την αβεβαιότητα που χαρακτηρίζει τις χρονοσειρές γενικότερα και ειδικότερα αυτές που

πραγματικά χρήζουν μετασχηματισμού για τη συρρίκνωση της διακύμανσής τους, και αφετέρου από τις ιδιαιτερότητες των χρονοσειρών που περιλαμβάνονται στο δείγμα (π.χ. πολλαπλασιαστικό μοτίβο τάσης).

2.5 Στατιστική ανάλυση απόδοσης μοντέλων πρόβλεψης

Αντίθετα με τη βασική στατιστική ανάλυση, η οποία πραγματοποιείται επί της ίδιας της χρονοσειράς, στη στατιστική ανάλυση απόδοσης μοντέλων πρόβλεψης απαιτείται η σύγκριση μεταξύ των πραγματικών και των προβλεπόμενων παρατηρήσεων, όπως αυτές έχουν προκύψει από την εφαρμογή κάποιας μεθόδου πρόβλεψης. Έτσι, δεδομένης μίας χρονοσειράς μήκους n , η παρατήρησή της στο σημείο i , Y_i , θα συγκριθεί με την πρόβλεψη στο αντίστοιχο σημείο F_i .

2.5.1 Ορισμοί και μέθοδοι αξιολόγησης

Σκοπός της εν λόγω ανάλυσης είναι να εκτιμήσουμε την απόδοση ενός μοντέλου πρόβλεψης, είτε ως προς την ακρίβεια των προβλέψεων που παράγει (*accuracy*), είτε ως προς την προκατάληψή του (*bias*). Στην πρώτη περίπτωση μας ενδιαφέρει η μέση διακύμανση των προβλέψεων γύρω από τις πραγματικές τιμές, ενώ στη δεύτερη αν οι προβλέψεις του αστοχούν συστηματικά προς μία κατεύθυνση, δηλαδή αν το μοντέλο παράγει σταθερά αισιόδοξες (μεγαλύτερες από τις πραγματικές) ή απαισιόδοξες (μικρότερες από τις πραγματικές) προβλέψεις. Προφανώς ένα μοντέλο θεωρείται κατάλληλο για πρόβλεψη όταν είναι ταυτόχρονα αμερόληπτο και ακριβές. Σε αυτή την περίπτωση η μέση τιμή των προβλέψεων θα ταυτίζεται με την μέση τιμή των δεδομένων (αμεροληψία) και το εν λόγω αποτέλεσμα θα προκύπτει από προβλέψεις που βρίσκονται πολύ κοντά στις εκάστοτε πραγματικές παρατηρήσεις (ακρίβεια). Δεδομένου ότι οι δύο αυτές συνιστώσες προβλεπτικής ικανότητας μπορεί να είναι ανεξάρτητες μεταξύ τους, γίνεται αντιληπτό ότι δύο μοντέλα μπορούν να έχουν ίδια ακρίβεια με διαφορετική προκατάληψη και ίδια προκατάληψη με διαφορετική ακρίβεια.

Θεωρητικά μπορούμε να αναλύσουμε το σφάλμα πρόβλεψης στις δύο συνιστώσες του ως εξής (*Syntetos & Katsikopoulos, 2016*):

$$MSE = bias^2 + variance + unexplained\ variance$$

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

$$variance = \frac{1}{n} \sum_{i=1}^n \left(F_i - \frac{1}{n} \sum_{j=1}^n F_j \right)^2$$

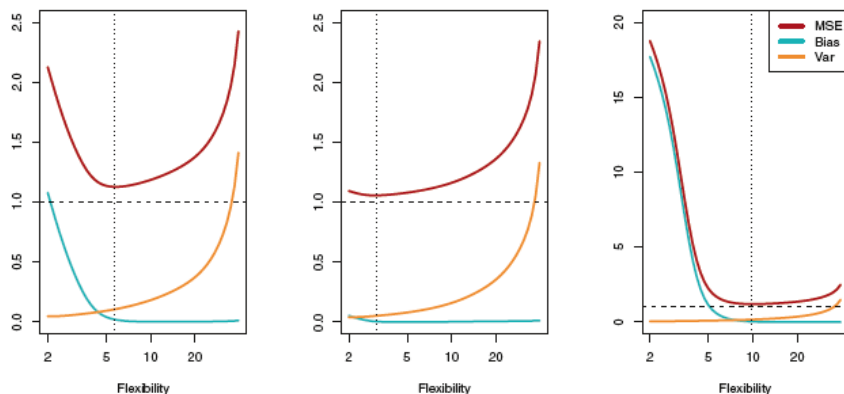
$$unexplained\ variance = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \frac{1}{n} \sum_{j=1}^n Y_j \right)^2$$

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

, όπου MSE το μέσο τετραγωνικό σφάλμα πρόβλεψης, $bias$ η προκατάληψη του μοντέλου, $variance$ η ακρίβειά του και $unexplained\ variance$ η διακύμανση η οποία οφείλεται στη διακύμανση των δεδομένων και δεν μπορεί να μοντελοποιηθεί με κάποιον τρόπο.

Δυστυχώς στην πράξη η συμφιλίωση των προβλέψεων ούτως ώστε να είναι ταυτόχρονα και αμερόληπτες και ακριβείς είναι μία ιδιαίτερα απαιτητική διαδικασία καθώς η βελτίωση του ενός οδηγεί σε χειροτέρευση του άλλου. Στη γενική περίπτωση, η όσο αυξάνεται η πολυπλοκότητα ενός μοντέλου πρόβλεψης, δηλαδή αυτό μοντελοποιεί μέσω επιπρόσθετων παραμέτρων περισσότερες σχέσεις μεταξύ των δεδομένων της χρονοσειράς, τόσο μειώνεται η προκατάληψή του ($fitting$). Ωστόσο η αύξηση των παραγόντων του οδηγεί σε συστηματική αύξηση της διακύμανσης των σφαλμάτων του και συνεπώς σε έλλειψη ακρίβειας πρόβλεψης ($over-fitting$). Το εν λόγω φαινόμενο είναι γνωστό ως $bias-variance\ trade-off$ και στόχος μας είναι να επιλέξουμε ένα μοντέλο με τέτοια πολυπλοκότητα ούτως ώστε να συμφιλιώνει όσο το δυνατόν τα δύο ζητούμενα (*James et al., 2013*).

Χαρακτηριστικά παραδείγματα του εν λόγω φαινομένου δίνονται στην *Εικόνα 2.14* όπου το μέσο τετραγωνικό σφάλμα πληθώρας μοντέλων διαφορετικής πολυπλοκότητας αναλύεται στις συνιστώσες προκατάληψης και ακρίβειας. Όπως παρουσιάζεται, η βέλτιστη επιλογή θα ήταν κατά σειρά από αριστερά προς τα δεξιά για τα τρία σετ δεδομένων η χρήση ενός μοντέλου τεσσάρων, δύο και έξι μεταβλητών καθώς σε εκείνα τα σημεία ελαχιστοποιούνται ταυτόχρονα οι δύο παράγοντες. Αξίζει να σημειωθεί ότι το MSE από μόνο του δεν μπορεί να μας πληροφορήσει με αξιοπιστία για το ποιο μοντέλο είναι το πλέον κατάλληλο ανά περίπτωση αφού ελαχιστοποιείται για πολυπλοκότητα έξι, τρία και δέκα αντίστοιχα.



Εικόνα 2.14: Αλληλεπίδραση προκατάληψης ($bias$) και ακρίβειας ($variance$) συναρτήσει της πολυπλοκότητας ($flexibility$) των μοντέλων.

Σε αυτό το σημείο σημειώνεται ότι επειδή δεν είναι καθόλου εύκολο να αναλύσει κανείς την απόδοση των μεθόδων πρόβλεψης ταυτόχρονα και ως προς τις δύο συνιστώσες, είθισται στην πράξη να επιλέγεται ένα και μόνο κριτήριο αξιολόγησης (σφάλμα) και να το αξιοποιούμε για την αξιολόγησή τους. Προφανώς με αυτόν τον τρόπο θα ελεγχθεί

ανομοιόμορφα η απόδοση του μοντέλου ως προς την αμεροληψία και την ακρίβειά του, ωστόσο είναι ο μόνος τρόπος για να αξιολογήσουμε αποδοτικά πληθώρα μεθόδων χωρίς να εφαρμόσουμε μία πολύπλοκη πολυκριτήρια ανάλυση.

Το σύνολο των τιμών πρόβλεψης αποτελείται από δύο μέρη. Οι πρώτες n παρατηρήσεις αυτού, για τις οποίες υπάρχουν διαθέσιμα πάντα και τα αντίστοιχα πραγματικά δεδομένα, καλείται διάστημα προσαρμογής του μοντέλου πρόβλεψης (*in-sample* ή *fit period*), ενώ οι υπόλοιπες αποτελούν τις μελλοντικές προβλέψεις (*out-of-sample* ή *test period*) και έχουν μήκος ίσο με τον ορίζοντα πρόβλεψης h (*forecasting horizon* ή *lead time*). Βάσει του *in-sample* σφάλματος γίνεται η επιλογή του μοντέλου πρόβλεψης και η παραμετροποίησή (εκπαίδευση) του, ενώ μέσω του *out-of-sample* σφάλματος η αξιολόγηση του μοντέλου ως προς την αποτελεσματικότητά του.

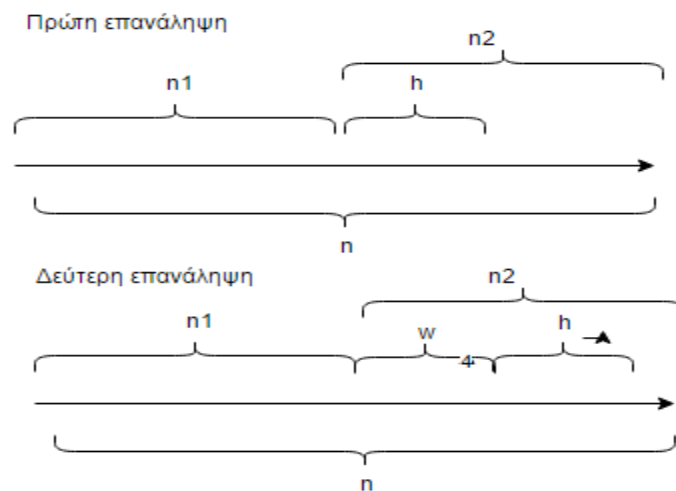
Όπως γίνεται κατανοητό, για την *out-of-sample* αξιολόγηση δεν υπάρχουν πάντα διαθέσιμα τα πραγματικά δεδομένα, καθώς συνήθως χρησιμοποιείται όλο το διαθέσιμο πλήθος παρατηρήσεων για την εκπαίδευση του μοντέλου και την παραγωγή προβλέψεων. Έτσι η αξιολόγηση του μοντέλου μπορεί να γίνει εκ των υστέρων όταν και γίνουν γνωστές οι μελλοντικές πραγματικές τιμές του προβλεπόμενου μεγέθους. Παρόλα αυτά συχνά οι αναλυτές «κρύβουν» παρατηρήσεις προκειμένου η εν λόγω ανάλυση να πραγματοποιείται πάντα και να εξάγονται χρήσιμα συμπεράσματα. Η εν λόγω διαδικασία είναι ιδιαίτερα σημαντική καθώς έχει διαπιστωθεί ότι το *in-sample* σφάλμα τείνει να είναι αρκετά μικρότερο του *out-of-sample*, το οποίο είναι άλλωστε και το σφάλμα που πρακτικά μας ενδιαφέρει (πραγματική επίδοση μοντέλου). Αυτό είναι και ένα από τα βασικά συμπεράσματα των τριών διαγωνισμών Makridakis-Competition (*M-Competition*) αλλά και αρκετών άλλων σχετικών μελετών (*Makridakis et al., 1982 και Makridakis et al., 1993 και Makridakis et al., 2000*).

Σε αυτήν την κατεύθυνση, και κυρίως κατά τη διαδικασία επιλογής κατάλληλου μοντέλου πρόβλεψης, έχουν αναπτυχθεί αρκετές μέθοδοι αξιολόγησης μοντέλων βάσει του *out-of-sample* σφάλματός τους. Κλασικό παράδειγμα αποτελούν τα *out-of-sample tests* κατά τα οποία η αρχική χρονοσειρά χωρίζεται σε δύο μέρη μήκους n_1 και n_2 , όπου $n_1+n_2=n$. Το πρώτο μέρος χρησιμοποιείται για την εκπαίδευση του μοντέλου και το δεύτερο για την αξιολόγησή του συγκρίνοντας τις παραγόμενες προβλέψεις με τις πραγματικές «κρυμμένες» παρατηρήσεις της χρονοσειράς.

Στην πιο απλή περίπτωση της μεθόδου θέτουμε $n_2=h$, δηλαδή αξιολογούμε το μοντέλο μία φορά για τον ορίζοντα πρόβλεψης που έχει επιλεγεί. Αυτή η μέθοδος καλείται *fixed-origin evaluation* και το μοντέλο εκπαιδεύεται χρησιμοποιώντας $n-h$ τιμές (*Tashman, 2000*). Επιλέγοντας ωστόσο μικρότερο μήκος n_1 η εν λόγω διαδικασία αξιολόγησης μπορεί να πραγματοποιηθεί πλέον για τον ίδιο ορίζοντα πρόβλεψης περισσότερες από μία φορές, ολισθαίνοντας σταδιακά το σημείο έναρξης των προβλέψεων (*forecasting origin*). Η διαδικασία αυτή, γνωστή και ως *rolling-origin evaluation*, μπορεί να φανεί ιδιαίτερα ευεργετική καθώς όσες περισσότερες αξιολογήσεις πραγματοποιούνται, τόσο πιο αντικειμενικά αναμένεται να είναι τα αντίστοιχα αποτελέσματα, εξαλείφοντας τον παράγοντα της τυχαιότητας (π.χ. ύπαρξη κάποιου ειδικού γεγονότος εντός του

διαστήματος αξιολόγησης). Για αυτό και θα πρέπει να προτιμάται όταν δεν υπάρχουν περιορισμοί που σχετίζονται με το υπολογιστικό κόστος. Διαφοροποιήσεις στην ακρίβεια του μοντέλου καθώς αυτό πραγματοποιεί τη διαδικασία rolling-origin μπορεί να οφείλονται είτε απλά στην προσθήκη δεδομένων (*updating*), είτε στον επανυπολογισμό των παραμέτρων του καθώς νέα δεδομένα γίνονται διαθέσιμα (*recalibrating*).

Φυσικά, παράμετροι όπως το μήκος n_1 , n_2 και h , καθώς και το διάστημα ολίσθησης του σημείου εκκίνησης πρόβλεψης (*window*) αποτελούν σημαντικές μεταβλητές οι οποίες καθορίζουν κατά πολύ τα τελικά συμπεράσματα της μεθόδου και θα πρέπει να ορίζονται ανάλογα με τις ανάγκες του εκάστοτε προβλήματος που ο αναλυτής καλείται να ικανοποιήσει. Όταν το σημείο εκκίνησης ανανεώνεται κατά μία περίοδο ανά επανάληψη η μέθοδος καλείται *one-step-ahead rolling origin*, ενώ όταν ανανεώνεται με βήμα w , *w-step ahead rolling origin*, αντίστοιχα. Στη δεύτερη περίπτωση, και ειδικά όταν θέτουμε $w=h$, υπάρχει το πλεονέκτημα ότι οι παρατηρήσεις που γίνονται διαθέσιμες δεν ξαναχρησιμοποιούνται στην αξιολόγηση, ελαχιστοποιώντας έτσι τον παράγοντα της τυχειότητας στα αποτελέσματα. Ωστόσο έχουν το μειονέκτημα ότι, ειδικά για μεγάλους ορίζοντες πρόβλεψης ή μικρού μήκους χρονοσειρές, το αρχικό δείγμα προβλέψεων μειώνεται σημαντικά, ενώ ο τελικός αριθμός επαναλήψεων είναι σαφώς μικρότερος οδηγώντας σε αυξημένη διακύμανση μεταξύ των επιμέρους αποτελεσμάτων.



Εικόνα 2.15: Παρουσίαση μεθόδου rolling origin με ανανέωση του fit period σε κάθε επανάληψη κατά w

Στη βιβλιογραφία που επεκτείνεται εκτός των τεχνικών προβλέψεων παρόμοια τεστ είναι τα *leave-k-out cross-validation* (k παρατηρήσεις εξαιρούνται σε κάθε επανάληψη) και *k-fold cross-validation* (όπου k τυχαία δείγματα λαμβάνονται από τις χρονοσειρές και ένα εξαιρείται για την τελική αξιολόγηση). Φυσικά, στα μοντέλα των χρονοσειρών όπου οι παρατηρήσεις πρέπει να βρίσκονται σε χρονολογική σειρά, η δεύτερη μέθοδος είναι μη εφαρμόσιμη.

2.5.2 Στατιστικοί δείκτες αξιολόγησης

Προχωρώντας στην αξιολόγηση των προβλέψεων αυτήν καθαυτή, ως σφάλμα πρόβλεψης στο σημείο i ορίζεται η αλγεβρική διαφορά μεταξύ της πραγματικής τιμής της χρονοσειράς και της πρόβλεψης

$$e_i = Y_i - F_i$$

Περίοδος	Δεδομένα	Πρόβλεψη	Διάστημα
1	Y_1	F_1	<i>In-sample</i>
2	Y_2	F_2	
3	Y_3	F_3	
⋮	⋮	⋮	
n-2	Y_{n-2}	F_{n-2}	
n-1	Y_{n-1}	F_{n-1}	
n	Y_n	F_n	
n+1		F_{n+1}	<i>Out-of-sample</i>
⋮		⋮	
n+h		F_{n+h}	

Πίνακας 2.6: Παρατηρήσεις και προβλέψεις χρονοσειράς προς στατιστική ανάλυση απόδοσης προβλέψεων.

Όπως αναφέρθηκε νωρίτερα, προτού γίνουν διαθέσιμες οι πραγματικές τιμές Y_{n+1} έως Y_{n+h} , η τιμή του σφάλματος μπορεί να υπολογιστεί μόνο για τις in-sample τιμές. Όταν, όμως, έχει παρέλθει κατάλληλο χρονικό διάστημα και τα δεδομένα έχουν γίνει διαθέσιμα, μπορούμε να υπολογίσουμε τόσο το σφάλμα του μοντέλου πρόβλεψης, όσο και το πραγματικό σφάλμα. Όλοι οι στατιστικοί δείκτες σφάλματος που θα αναφερθούν είναι εκφρασμένοι έτσι ώστε να αναφέρονται στο σφάλμα του μοντέλου πρόβλεψης, αλλά με κατάλληλες αλλαγές στις περιόδους που συμμετέχουν για τον υπολογισμό αυτών, μπορούν να αναφερθούν στην ακρίβεια των προβλέψεων πέραν της περιόδου n , δεδομένης της ύπαρξης νέων πραγματικών δεδομένων.

Μέσο σφάλμα (Mean Error): Υπολογίζεται από τον απλό προσημασμένο μέσο όρο των σφαλμάτων και εκφράζει ένα μέτρο συστηματικότητας του σφάλματος. Όσο η τιμή αυτού είναι κοντά στο μηδέν, τόσο τα σφάλματα είναι τυχαία και όχι συστηματικά. Αν ο δείκτης παίρνει θετικές τιμές, δηλώνει απαισιοδοξία στις προβλέψεις, μιας και οι προβλέψεις ήταν κατά μέσο όρο μικρότερες των πραγματικών τιμών. Από την άλλη, αρνητικές τιμές του δείκτη δηλώνουν αισιοδοξία. Συχνά ο δείκτης αναφέρεται και ως bias και υπολογίζεται από τον παρακάτω τύπο.

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

Μέσο απόλυτο σφάλμα (Mean Absolute Error): Αποτελεί ένα μέτρο της ακρίβειας της πρόβλεψης έναντι των πραγματικών τιμών διατηρώντας τις μονάδες μέτρησης της αρχικής χρονοσειράς. Εκφράζει τη μέση αστοχία της πρόβλεψης, χωρίς να δίνεται σημασία στην κατεύθυνση της πρόβλεψης. Όσο μεγαλύτερη είναι η τιμή του δείκτη, τόσο

μικρότερη προκύπτει η ακρίβεια της μεθόδου που εφαρμόστηκε. Αρνητικό του εν λόγω δείκτη είναι ότι εξαρτάται από την κλίμακα της χρονοσειράς και συνεπώς δεν μπορεί να χρησιμοποιηθεί για να εξαγεί συμπεράσματα σε σεντ πολλών χρονοσειρών.

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

Μέσο τετραγωνικό σφάλμα (Mean Squared Error): Όπως και το απόλυτο σφάλμα, αποτελεί μέτρο της ακρίβειας πρόβλεψης, το οποίο όμως δίνει πολύ μεγαλύτερο βάρος στα μεγάλα σφάλματα. Αυτό γίνεται δεδομένου ότι τα σφάλματα τετραγωνίζονται και συνεπώς σφάλματα με μικρές τιμές λαμβάνουν μικρότερο βάρος κατά τον υπολογισμό του. Χρησιμοποιείται ευρέως για τον υπολογισμό των βέλτιστων παραμέτρων εξομάλυνσης και την προσαρμογή μοντέλων πρόβλεψης.

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

Ρίζα μέσου τετραγωνικού σφάλματος (Root Mean Squared Error): Υπολογίζεται άμεσα από την τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος. Έχει τις ίδιες ιδιότητες με αυτό, αλλά είναι εκφρασμένο στις μονάδες της αρχικής χρονοσειράς προκειμένου να γίνεται πιο κατανοητό.

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

Μέσο απόλυτο ποσοστιαίο σφάλμα (Mean Absolute Percentage Error): Συχνά είναι απαραίτητος ο υπολογισμός των σφαλμάτων πρόβλεψης σε ποσοστιαία μορφή. Αυτό είναι ιδιαίτερα χρήσιμο όταν θέλουμε για παράδειγμα να συγκρίνουμε την ακρίβεια μιας μεθόδου πρόβλεψης που έχει εφαρμοσθεί σε περισσότερες από μία χρονοσειρές διαφορετικού επιπέδου μέσης τιμής ή όταν οι πραγματικές τιμές είναι ιδιαίτερα υψηλές.

Το μέσο απόλυτο ποσοστιαίο σφάλμα είναι εκφρασμένο επί τοις εκατό και λαμβάνει τιμές μεγαλύτερες ή ίσες του μηδενός, με τις μικρότερες τιμές να υποδηλώνουν και καλύτερη απόδοση της μεθόδου πρόβλεψης. Στην πραγματικότητα αποτελεί δηλαδή μία ποσοστιαία έκφραση του απόλυτου σφάλματος, με το μειονέκτημα βέβαια ότι οδηγεί σε απροσδιοριστίες για μηδενικές τιμές και σε ακραίες τιμές για τιμές μικρότερες της μονάδας.

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

Συμμετρικό μέσο απόλυτο ποσοστιαίο σφάλμα (Symmetric Mean Absolute Percentage Error): Αποτελεί μια παραλλαγή του μέσου απόλυτου ποσοστιαίου σφάλματος και υπολογίζεται ως εξής:

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

Σε αντίθεση με το μέσο απόλυτο σφάλμα, το απόλυτο του σφάλματος δε διαιρείται απλώς με την πραγματική τιμή, αλλά με το ημιάθροισμα της πραγματικής τιμής και της πρόβλεψης. Με την αλλαγή αυτή, ο δείκτης αυτός αποκτά και πάνω όριο και πλέον μπορεί να πάρει τιμές στο διάστημα [0%,200%]. Έτσι λύνονται τα προβλήματα απροσδιοριστίας και ακραίων τιμών που αναφέρθηκαν στην περίπτωση του απόλυτου ποσοστιαίου σφάλματος.

Ένα πρόβλημα βέβαια με το συγκεκριμένο δείκτη είναι ότι οι αισιόδοξες και οι απαισιόδοξες προβλέψεις δεν μεταχειρίζονται αντίστοιχα. Έτσι, μοντέλα με την ίδια απόλυτη ακρίβεια αλλά διαφορετική προκατάληψη θα έχουν διαφορετικές τιμές σφάλματος, με τις αισιόδοξες προβλέψεις να υπερτερούν. Ο εν λόγω δείκτης είναι από τους πιο διαδεδομένους στις τεχνικές προβλέψεων και λόγω των πλεονεκτημάτων του χρησιμοποιείται ιδιαίτερα σε διαγωνισμό προβλέψεων.

Σχετικά σφάλματα (Relative Measures): Συχνά ζητούμενο δεν είναι η αξιολόγηση της απόλυτης ακρίβειας πρόβλεψης, αλλά η εξακρίβωση του κατά πόσο μία μέθοδος είναι πιο αποτελεσματική από μία άλλη. Τέτοια περίπτωση μπορεί να είναι η μελέτη αντικατάστασης μίας μεθόδου πρόβλεψης σε μία επιχείρηση με μία καινούργια. Έτσι, λόγω των προβλημάτων και των περιορισμών που συνεπάγονται τα σφάλματα που παρουσιάστηκαν νωρίτερα, ένας ασφαλής τρόπος μέτρησης της απόδοσης μίας μεθόδου είναι ο υπολογισμός του σχετικού της σφάλματος ως προς μία μέθοδο αναφοράς (*benchmark*).

Για λόγους απλότητας, και όταν δεν υπάρχει προφανής απάντηση, ως μέθοδο αναφοράς τίθεται η απλοϊκή μέθοδος ή μία άλλη η οποία χαρακτηρίζεται από σχετική απλότητα. Σκοπός είναι να εξακριβωθεί αν η συγκρινόμενη μέθοδος, που είναι και πιο πολύπλοκη, οδηγεί σε βελτίωση της προβλεπτικής ικανότητας. Αν δηλαδή αξίζει να θυσιάσουμε υπολογιστικό κόστος προκειμένου να γίνουμε πιο ακριβείς. Το σχετικό σφάλμα ορίζεται με τη χρήση ενός μέτρου ακρίβειας (*error metric EM*) ως εξής:

$$RelEM = \frac{EM_i}{EM_{benchmark}}$$

Έτσι, σε περίπτωση που το ζητούμενο είναι η ακρίβεια, ως δείκτη μπορούμε να χρησιμοποιήσουμε το μέσο απόλυτο σφάλμα, ενώ αν το ζητούμενο είναι η αμεροληψία, το μέσο σφάλμα. Το σχετικό σφάλμα με τιμές μικρότερες της μονάδας φανερώνει ότι η συγκρινόμενη μέθοδος υπερτερεί της μεθόδου αναφοράς, ενώ με τιμές μεγαλύτερες της μονάδας το αντίθετο. Έτσι, μία τιμή π.χ. 0.8 φανερώνει ότι

η συγκρινόμενη μέθοδος είναι κατά μέσο όρο κατά 20% καλύτερη από αυτήν της αναφοράς.

Μέσο απόλυτο κανονικοποιημένο σφάλμα (Mean Absolute Scaled Error): Οι *Hyndman & Koehler (2006)*, προκειμένου να αντιμετωπίσουν τις περιπτώσεις απροσδιοριστίας των δεικτών *MAPE* και *sMAPE* αλλά και να δώσουν την ίδια βαρύτητα στα μικρά και τα μεγάλα σφάλματα, σε αντίθεση με τους δείκτες *MSE* και *RMSE*, πρότειναν τον ακόλουθο στατιστικό δείκτη για τη μέτρηση της ακρίβειας πρόβλεψης:

$$MASE \frac{\frac{1}{n} \sum_{i=1}^n |Y_i - F_i|}{\frac{1}{n-1} \sum_{i=2}^n |Y_i - Y_{i-1}|}$$

Πρόκειται στην ουσία για υποπερίπτωση του σχετικού σφάλματος κατά την οποία κανονικοποιούμε με τη μέση τιμή των διαφορών πρώτου βαθμού της χρονοσειράς, με την πρόβλεψη δηλαδή της απλοϊκής μεθόδου.

Σε αυτό το σημείο αναφέρεται ότι, αν και θεωρητικά δεν μπορεί να τεκμηριωθεί, το κανονικοποιημένο σφάλμα έχει ως κάτω όριο το 0.5. Το εν λόγω συμπέρασμα έχει αποδειχτεί εμπειρικά μέσα από μελέτες που πραγματοποιήθηκαν σε μεγάλα δείγματα χρονοσειρών διαφόρων τύπων και χαρακτηριστικών και οι οποίες έδειξαν ότι υπό φυσιολογικές συνθήκες μία μέθοδος δεν μπορεί πρακτικά να είναι καλύτερη πάνω από 2 φορές από την απλοϊκή. Αντίθετα, αν είναι αποτελεσματική, η τιμή του σφάλματος αναμένεται να είναι κοντά στο 0.7 και σπάνια να υπερβαίνει το 0.9. Αυτό συμβαίνει λόγω της τυχαιότητας των χρονοσειρών η οποία έχει ως αποτέλεσμα να μην μία μέθοδος να υπερτερεί μίας άλλης, αλλά ποτέ να μην πετυχαίνει τέτοια ακρίβεια που να εκμηδενίζει την απόδοση της αναφοράς. Έτσι, αν μία αντικειμενικά αποτελεσματικότερη μέθοδος πετύχει μεμονωμένα ακρίβεια μικρότερη από 0.5 ή μεγαλύτερη του 0.9, το εν λόγω αποτέλεσμα οφείλεται πιθανότατα σε τύχη.

Theil's U-Statistic: Ο *Theil (1966)* πρότεινε ένα δείκτη σφάλματος, ο οποίος συνδυάζει χαρακτηριστικά σχετικών σφαλμάτων, δίνοντας παράλληλα μεγαλύτερο βάρος στα μεγαλύτερα σφάλματα, καθώς μεγάλα σφάλματα για μια επιχείρηση συνεπάγονται και ισοδύναμα μεγάλη ζημιά. Ο στατιστικός αυτός δείκτης επιτρέπει τη συγκριτική συσχέτιση της ακρίβειας των προβλέψεων με τις προβλέψεις της μεθόδου Naive και ταυτόχρονα τετραγωνίζει τα σφάλματα, ούτως ώστε να δίνεται μεγαλύτερο βάρος στις μεγάλες αποκλίσεις. Η μαθηματική διατύπωση του δείκτη έχει ως εξής:

$$U = \sqrt{\frac{\sum_{i=2}^n \left(\frac{|Y_i - F_i|}{|Y_{i-1}|}\right)^2}{\sum_{i=2}^n \left(\frac{|Y_i - Y_{i-1}|}{|Y_{i-1}|}\right)^2}}$$

• Αν $U=1$, τότε η μέθοδος Naive είναι εξίσου ακριβής με την μέθοδο πρόβλεψης που εφαρμόστηκε.

Κεφάλαιο 2: Ανάλυση χρονοσειρών και απόδοσης μοντέλων πρόβλεψης

- Αν $U < 1$, τότε η μέθοδος πρόβλεψης που εφαρμόστηκε έχει καλύτερη απόδοση από πλευράς ακρίβειας σε σχέση με τη μέθοδο Naive (όσο μικρότερη τιμή, τόσο καλύτερη απόδοση).
- Αν $U > 1$, τότε η μέθοδος πρόβλεψης που εφαρμόστηκε έχει χειρότερη απόδοση από πλευράς ακρίβειας σε σχέση με τη μέθοδο Naive, οπότε δεν υπάρχει λόγος να εφαρμοσθεί (όσο μεγαλύτερη τιμή, τόσο χειρότερη απόδοση).

Percentage Better: Μια συσχετιστική προσέγγιση, που επιτρέπει την σύγκριση διαφορετικών μεθόδων που έχουν εφαρμοσθεί σε μεγάλο πλήθος χρονοσειρών διαφορετικών επιπέδων, είναι η χρήση του ποσοστού για το οποίο μια μέθοδος βάσει ενός δείκτη EM είναι ακριβέστερη έναντι μιας μεθόδου αναφοράς για ένα συγκεκριμένο δείκτη σφάλματος. Ο δείκτης ορίζεται ως εξής:

$$\text{Percentage Better} = \text{mean}(\{EM_i < EM_{\text{benchmark}}\}) * 100 (\%)$$

Το μειονέκτημα του παραπάνω δείκτη είναι ότι δε δίνει καμία αίσθηση του επιπέδου βελτίωσης. Έτσι, αν μία μέθοδος υπερτερεί της αναφοράς οριακά στο μεγαλύτερο ποσοστό των περιπτώσεων, αλλά υστερεί σημαντικά στο υπόλοιπο, ο δείκτης θα ισχυριστεί ότι η μέθοδος είναι σαφώς καλύτερη.

Κεφάλαιο 2: Ανάλυση χρονοσειρών και απόδοσης μοντέλων πρόβλεψης

Το πρόβλημα με τις προβλέψεις είναι ότι είναι σωστές πολλές φορές για να τις αγνοούμε και λάθος επίσης πολλές για να βασιζόμαστε πάνω τους

Patrick Young

Κεφάλαιο 3: Μοντέλα πρόβλεψης

3.1 Γενικά για τα μοντέλα πρόβλεψης

Μία από τις πλέον χρήσιμες εφαρμογές της ανάλυσης χρονοσειρών είναι ο εντοπισμός της σχέσης που συνδέει μία ακολουθία παρατηρήσεων ενός μεγέθους και η χρήση αυτής για την εκτίμηση της εξέλιξης του στο μέλλον. Η διαδικασία αυτή μπορεί να τυποποιηθεί επιτυχώς μέσω μαθηματικών μοντέλων, γνωστά και ως στατιστικά μοντέλα πρόβλεψης. Στόχος των μοντέλων είναι να παρέχουν όσο το δυνατόν πιο εύστοχες προβλέψεις, να ελαχιστοποιούν δηλαδή την απόκλιση των εκτιμήσεών τους από τις πραγματικές τιμές του μεγέθους που προσπαθούν να περιγράψουν στο μέλλον.

Από το 1980 και μετά, ο τομέας των προβλέψεων έχει γνωρίσει σημαντική ανάπτυξη, βρίσκοντας ευρεία εφαρμογή τόσο σε ακαδημαϊκό, όσο και σε επιχειρησιακό περιβάλλον. Αρκετά είναι τα μοντέλα πρόβλεψης που έχουν προταθεί και τα οποία διαφέρουν μεταξύ τους τόσο ως προς τη φύση τους όσο και ως προς τα πεδία στα οποία μπορούν να βρουν εφαρμογή. Το βασικό κριτήριο καταλληλότητας συνεχίζει να είναι φυσικά η ακρίβεια πρόβλεψης, η οποία και εξαρτάται από το βαθμό στον οποίο το εκάστοτε μοντέλο πρόβλεψης είναι ικανό να περιγράψει τη χρονοσειρά δεδομένου των ποιοτικών της χαρακτηριστικών, του μήκους της, της συχνότητάς της και άλλων σχετικών παραγόντων.

Στο παρόν κεφάλαιο θα πραγματοποιηθεί μία εισαγωγή στα μοντέλα πρόβλεψης και θα παρουσιαστούν βασικά στατιστικά μοντέλα που χρησιμοποιούνται για την πρόβλεψη μεγεθών σε μεγάλο εύρος εφαρμογών. Τέλος, θα αναλυθεί η μέθοδος θ και θα προταθεί μία νέα γενικευμένη της μορφή η οποία επιτρέπει την αποτελεσματικότερη μοντελοποίηση των μοτίβων των χρονοσειρών.

3.2 Κατηγορίες μοντέλων πρόβλεψης

Τα μοντέλα μπορούν να αναλυθούν σε δύο βασικές κατηγορίες ανάλογα με τον τρόπο που εξάγονται οι προβλέψεις, στα **ποσοτικά**, ή αλλιώς στατιστικά μοντέλα, και στα **ποιοτικά** (Πετρόπουλος & Ασημακόπουλος 2011). Τα πρώτα διαχωρίζονται στο *μοντέλο χρονοσειρών* και το *αιτιοκρατικό μοντέλο*, ενώ τα δεύτερα στην *κριτική πρόβλεψη* και την *πρόβλεψη στόχου*.

Η στατιστική πρόβλεψη αναφέρεται στην εφαρμογή ενός μαθηματικού μοντέλου πάνω σε μια σειρά δεδομένων προκειμένου να παραχθεί με συστηματικό τρόπο η πρόβλεψη της πορείας της ακολουθίας. Για να καταστεί αυτό εφικτό τα ποσοτικά μοντέλα προσπαθούν συνήθως να αποσαφηνίσουν τη σχέση που συνδέει τις παρατηρήσεις μεταξύ τους (μοντέλο χρονοσειρών) ή με αυτές άλλων συσχετισμένων μεγεθών (αιτιοκρατικό μοντέλο), την οποία και αξιοποιούν στη συνέχεια για την προέκτασή τους. Όπως γίνεται κατανοητό, σε αυτήν την περίπτωση οι προβλέψεις παράγονται χωρίς την εμπλοκή του ανθρώπου

και είναι τόσο ακριβείς όσο επιτυχημένα κατάφερε το μοντέλο να εντοπίσει τις εκάστοτε σχέσεις.

Από την άλλη πλευρά, η κριτική πρόβλεψη βασίζεται στη γνώση, την κρίση και την πείρα ενός ατόμου ή μιας ομάδας ατόμων χωρίς να απαιτούνται απαραίτητα ιστορικά δεδομένα. Αν και δεν οδηγεί σε αυτοματοποιημένη παραγωγή προβλέψεων και εμπεριέχει σε σημαντικό βαθμό την προκατάληψη των ανθρώπων σε κρίση, πλεονεκτεί έναντι των στατιστικών μεθόδων ως προς το ότι λαμβάνει υπόψη ειδικά γεγονότα, αντισταθμίζει τις ανεπάρκειες και εξισορροπεί τις ανομοιογένειες των καταγεγραμμένων δεδομένων. Ιδανικά, και όταν υπάρχει διαθέσιμη εξωτερική πληροφορία σε αδόμητη μορφή, η στατιστική πρόβλεψη συνδυάζεται με την κριτική για την επίτευξη καλύτερων αποτελεσμάτων. Μερικές από τις βασικότερες μεθόδους της είναι η απλή κρίση, η μέθοδος Delphi και οι δομημένες αναλογίες.

Τέλος, η πρόβλεψη στόχου είναι ουσιαστικά η επιθυμητή πορεία της χρονοσειράς που θέλει η επιχείρηση να έχει και όχι τόσο η πραγματική πρόβλεψή της. Με ανάλυση των ιστορικών δεδομένων του μεγέθους που μελετάται υπολογίζεται ο ρυθμός ανάπτυξής του για το παρελθόν διάστημα. Στη συνέχεια, η εφαρμογή των στατιστικών προβλέψεων δίνει μια εκτίμηση της πορείας της χρονοσειράς και πλέον μένει ο χειριστής των δεδομένων να τροποποιήσει το μελλοντικό ρυθμό ανάπτυξης, σύμφωνα με την επιθυμητή μελλοντική πορεία της επιχείρησης. Η τελική πρόβλεψη είναι συνεπώς εδώ η πρόβλεψη του διευθυντή ανάπτυξης (χειριστή) και περιέχει αρκετή αισιοδοξία και μεροληψία, καθώς καθορίζει τους στόχους της επιχείρησης.

Στην παρούσα εργασία θα εξεταστούν αποκλειστικά τα στατιστικά μοντέλα και συγκεκριμένα τα μοντέλα χρονοσειρών καθώς στόχος της είναι η ανάπτυξη μίας αυτοματοποιημένης και γενικευμένης μεθοδολογίας πρόβλεψης για μεγάλο πλήθος χρονοσειρών ποικίλων χαρακτηριστικών. Έτσι, η χρήση ποιοτικών μοντέλων κρίνεται αδύνατη ενώ η εφαρμογή αιτιοκρατικών μοντέλων άστοχη δεδομένου ότι αυτά απαιτούν τη συσχέτιση κάθε χρονοσειράς με άλλα σχετικά μεγέθη, διαδικασία αδύνατη να αυτοματοποιηθεί με αποδοτικό τρόπο.

3.3 Μοντέλο χρονοσειρών

Το μοντέλο χρονοσειρών είναι το πλέον γνωστό μοντέλο στατιστικής πρόβλεψης, η εφαρμογή του οποίου προϋποθέτει αποκλειστικά την ύπαρξη επαρκών ιστορικών δεδομένων σε προηγούμενες, σταθερές χρονικές περιόδους (διακριτή χρονοσειρά). Βασιζόμενο στις παρελθοντικές τιμές της υπό εξέταση τυχάιας μεταβλητής ανιχνεύει ένα πρότυπο συμπεριφοράς το οποίο και αξιοποιεί για να παράγει προβλέψεις προεκτείνοντάς το κατάλληλα στο μέλλον. Έτσι, η χρήση του μοντέλου χρονοσειρών παράγει ικανοποιητικά αποτελέσματα στις περιπτώσεις που η εξέλιξη της χρονοσειράς γίνει με βάση τις τωρινές συνθήκες, δηλαδή δεν μεταβληθεί το πρότυπο συμπεριφοράς που έχει εντοπιστεί. Σημαντικό μειονέκτημα του μοντέλου χρονοσειρών είναι ωστόσο η αδυναμία συσχέτισης του προβλεπόμενου μεγέθους με εξωτερικούς παράγοντες που πιθανότατα το επηρεάζουν σημαντικά, όπως π.χ. η προώθηση ενός προϊόντος ή η είσοδος ενός ανταγωνιστή στην αγορά.

Στις μεθόδους χρονοσειρών συγκαταλέγονται η απλοϊκή μέθοδος (Naïve), η μέθοδος αποσύνθεσης (Decomposition Method), η μέθοδος εκθετικής εξομάλυνσης (Exponential Smoothing), τα αυτοπαλινδρομούμενα μοντέλα κινητού μέσου όρου *ARMA* (Auto Regressive Moving Average) και η μέθοδος θ .

3.3.1 Απλοϊκή μέθοδος (Naïve)

Πρόκειται για την απλούστερη μέθοδο πρόβλεψης που μπορούμε να ορίσουμε κατά την οποία υποθέτουμε ότι το μέλλον θα ταυτίζεται πάντα με το παρόν. Έτσι, οι προβλέψεις της απλοϊκής μεθόδου, γνωστής και ως αφελής μεθόδου, δίνονται από τη σχέση:

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

Προφανώς, επειδή η εν λόγω μέθοδος δεν λαμβάνει υπόψη της ούτε πιθανές διακυμάνσεις στο επίπεδο της χρονοσειράς αλλά ούτε και τάση, η ακρίβειά της αναμένεται να είναι ιδιαίτερα χαμηλή, ειδικά για μακροπρόθεσμες προβλέψεις. Έτσι, χρησιμοποιείται συνήθως ως σημείο αναφοράς ακρίβειας για άλλες μεθόδους (*benchmark*) προκειμένου να γνωρίζουμε πόσες φορές καλύτερα αποδίδει η εκάστοτε μέθοδος από την πλέον απλοϊκή (βλ. σχετικά σφάλματα). Γνωστή είναι και η Naïve 2, ή αλλιώς εποχιακή απλοϊκή μέθοδος, κατά την οποία η Naïve εφαρμόζεται επί των αποεποχικοποιημένων δεδομένων ούτως ώστε να ληφθεί υπόψη η εποχιακότητα.

Άλλη χρησιμότητα της Naïve είναι ο εντοπισμός «μη προβλέψιμων» χρονοσειρών, μεγεθών δηλαδή για τα οποία η απλοϊκή μέθοδος αποδίδει καλύτερα από οποιοδήποτε άλλο μοντέλο πρόβλεψης. Τέτοιες χρονοσειρές, όπως γίνεται κατανοητό, δεν χαρακτηρίζονται από κάποιο μοτίβο συμπεριφοράς και διαθέτουν σημαντική τυχαιότητα. Έτσι είναι μάταιη και πρακτικά αδύνατη η επέκτασή τους.

3.3.2 Μέθοδος αποσύνθεσης

Όπως παρουσιάστηκε στην ενότητα 2.2, κάθε χρονοσειρά μπορεί να αποσυντεθεί στα επιμέρους ποιοτικά της χαρακτηριστικά, δηλαδή την τάση, την εποχιακότητα, τον κύκλο και την τυχαιότητα. Έτσι, αν κανείς προεκτείνει τα εν λόγω χαρακτηριστικά στο μέλλον δύναται να παράγει συνδυάζοντάς τα εκ νέου προβλέψεις για την υπό εξέταση χρονοσειρά. Η διαδικασία που ακολουθείται είναι η εξής:

- *Μοντέλο αποσύνθεσης*: Θεωρούμε ένα μοντέλο αποσύνθεσης για τον υπολογισμό των χαρακτηριστικών της χρονοσειράς. Αυτό μπορεί να είναι είτε πολλαπλασιαστικής είτε προσθετικής μορφής και φανερώνει τον τρόπο με τον οποίο συνδυάζονται τα επιμέρους χαρακτηριστικά προκειμένου να προκύψουν τα αρχικά δεδομένα. Στην περίπτωση της κλασικής πολλαπλασιαστικής μεθόδου αποσύνθεσης θεωρούμε δηλαδή:

$$Y = T * C * S * R$$

, όπου T , C , S και R είναι αντίστοιχα τα χαρακτηριστικά της τάσης, του κύκλου, της εποχιακότητας και της τυχαιότητας.

- *Προέκταση τάσης*: Η τάση προεκτείνεται εύκολα στο μέλλον αξιοποιώντας την εξίσωση που χρησιμοποιήθηκε εξ αρχής για την εκτίμησή της. Στην περίπτωση

π.χ. γραμμικής τάσης υπολογίζεται κατά την αποσύνθεση της χρονοσειράς η ευθεία της απλής γραμμικής παλινδρόμησης. Γνωρίζοντας τους συντελεστές αυτής της ευθείας μπορούμε να εκτιμήσουμε στη συνέχεια την τιμή της τάσης για οποιοδήποτε ορίζοντα πρόβλεψης.

- *Προέκταση κύκλου*: Ο κύκλος είναι ένα μέγεθος το οποίο δεν μπορούμε να το επεκτείνουμε με προφανή τρόπο καθώς φανερώνει μακροχρόνιες διακυμάνσεις στο επίπεδο της χρονοσειράς που εξαρτώνται κυρίως από εξωτερικούς παράγοντες. Ο ασφαλέστερος τρόπος για την προέκτασή του λοιπόν είναι να θεωρήσουμε ότι η τελευταία γνωστή τιμή του θα παραμείνει σταθερή στον ζητούμενο ορίζοντα πρόβλεψης.
- *Προέκταση εποχιακότητας*: Στην πραγματικότητα η εποχιακότητα δεν προεκτείνεται. Απλώς εποχικοποιούνται οι προβλέψεις που παράγονται χρησιμοποιώντας τους δείκτες εποχιακότητας που εκτιμήθηκαν κατά την αποσύνθεση για κάθε μία περίοδο στη συχνότητα της χρονοσειράς.
- *Προέκταση τυχαιότητας*: Η τυχαιότητα είναι ένα στοχαστικό μέγεθος το οποίο δεν προβλέπεται καθώς εκφράζει επί της ουσίας τον θόρυβο που εμπεριέχεται στα δεδομένα, ή καλύτερα την μη ερμηνευθείσα διακύμανση της χρονοσειράς. Έτσι, κατά την προέκταση της χρονοσειράς θεωρείται πως έχει μηδενική επίδραση. Εδώ θεωρούμε δηλαδή ότι $R_{t+n}=1$ για κάθε ορίζοντα πρόβλεψης n .

3.3.3 Μοντέλα εκθετικής εξομάλυνσης

Η εκθετική εξομάλυνση (*Hyndman et al., 2008*), ως μέθοδος πρόβλεψης παράγει τα αποτελέσματά της έπειτα από εξομάλυνση των δεδομένων για την απομόνωση των πραγματικών προτύπων της χρονοσειράς από τις τυχαιές διακυμάνσεις εντός αυτής. Η βασική αρχή της μεθόδου είναι ότι όσο πιο πρόσφατα είναι τα δεδομένα τόσο σημαντικότερη είναι και η πληροφορία που εμπεριέχεται σε αυτά. Έτσι, ανάλογα βέβαια με τη συστηματικότητα των προτύπων συμπεριφοράς που εντοπίζονται, τα μοντέλα εκθετικής εξομάλυνσης δίνουν μεγάλη βαρύτητα στα πρόσφατα δεδομένα, η οποία και φθίνει εκθετικά καθώς κινούμαστε προς το παρελθόν. Για παράδειγμα, η απλή εκθετική εξομάλυνση χρησιμοποιεί έναν σταθμισμένο μέσο όρο για την εκτίμηση του επιπέδου της χρονοσειράς τα βάρη του οποίου φθίνουν ανάλογα με το πόσο απέχει η εκάστοτε παρατήρηση από το σημείο για το οποίο ζητείται εκτίμηση.

Η απλότητα τους και η ικανοποιητική ακρίβεια που παρέχουν σε σχέση με άλλα, αρκετά πιο πολύπλοκα μοντέλα, καθιστούν τα μοντέλα εκθετικής εξομάλυνσης ιδιαίτερα δημοφιλή παρά το γεγονός ότι πρωτοεμφανίστηκαν στα μέσα της δεκαετίας του '60 (*Brown 1959 και Gardner 1985*). Το συγκριτικό τους πλεονέκτημά οφείλεται κυρίως στο γεγονός ότι αποσυνθέτουν τη χρονοσειρά στις βασικές της συνιστώσες, καθεμία εκ των οποίων εξομαλύνεται και επαναπροσδιορίζεται καθώς περνά ο χρόνος, ωθώντας έτσι το μοντέλο να μην επηρεάζεται εύκολα από ακραίες τιμές και μεγάλες διακυμάνσεις λόγω π.χ. ειδικών γεγονότων. Γενικά τα μοντέλα εκθετικής εξομάλυνσης αποδίδουν καλύτερα σε περιπτώσεις στάσιμων χρονοσειρών και χρονοσειρών με μικρό και σταθερό ρυθμό

ανάπτυξης. Τα πλέον γνωστά μοντέλα εκθετικής εξομάλυνσης είναι η απλή εκθετική εξομάλυνση, η εκθετική εξομάλυνση γραμμικής τάσης και η εκθετική εξομάλυνση μη γραμμικής ή φθίνουσας τάσης.

Πιο συγκεκριμένα, στη γενική τους περίπτωση τα μοντέλα εκθετικής εξομάλυνσης αποσυνθέτουν τη χρονοσειρά σε τέσσερις συνιστώσες (*components*): το επίπεδο, την τάση, την εποχιακότητα και το υπολειπόμενο σφάλμα (*Hyndman et al., 2002*). Το επίπεδο καθορίζει τη τιμή γύρω από την οποία κινείται η χρονοσειρά. Η τάση μοντελοποιεί την βραχυπρόθεσμη κατεύθυνση της χρονοσειράς, η εποχιακότητα τις συστηματικές διακυμάνσεις της χρονοσειράς σε προκαθορισμένες περιόδους εντός της συχνότητας που έχει ορισθεί και το υπολειπόμενο σφάλμα τα χαρακτηριστικά που οφείλονται σε τυχαιότητα και τα οποία δεν μπορούν συνεπώς να προβλεφθούν με κάποιον τρόπο.

Όπως φαίνεται στον *Πίνακα 3.1*, ανάλογα με τον τύπο της κάθε συνιστώσας η μέθοδος εκθετικής εξομάλυνσης επεκτείνεται σε συνολικά 30 μοντέλα. Το σφάλμα της μεθόδου μπορεί είναι είτε προσθετικό (*Additive*) είτε πολλαπλασιαστικό (*Multiplicative*), η τάση απούσα (*None*), προσθετική, πολλαπλασιαστική, φθίνουσα προσθετική (*Additive damped*) και φθίνουσα πολλαπλασιαστική (*Multiplicative damped*), ενώ η εποχιακότητα απούσα, πολλαπλασιαστική ή προσθετική. Από το σύνολο των μοντέλων σημειώνεται ότι τα 11 είναι εξορισμού αρκετά ασταθή (*Hyndman & Athanasopoulos, 2016, κεφάλαιο 7.7*) καθώς συνδυάζουν προσθετικά με πολλαπλασιαστικά μοτίβα και συνεπώς δύναται να υποπέσουν σε απροσδιοριστίες. Έτσι, είθισται να χρησιμοποιούνται τα υπόλοιπα 19 ευσταθή μοντέλα για την παραγωγή αξιόπιστων προβλέψεων. Σημειώνεται για παράδειγμα ότι τα μοντέλα πολλαπλασιαστικού σφάλματος δεν ορίζονται όταν τα δεδομένα περιέχουν αρνητικές ή μηδενικές τιμές ενώ τα μοντέλα προσθετικού σφάλματος με πολλαπλασιαστική τάση ή εποχιακότητα δεν ορίζονται όταν τα δεδομένα περιέχουν μηδενικές τιμές.

Πίνακας 3.1 Μοντέλα εκθετικής εξομάλυνσης ανάλογα με τον τύπο σφάλματος, εποχιακότητας και τάσης.

Συνιστώσα τάσης	Συνιστώσα εποχιακότητας		
	N (None)	A (Additive)	M (Multiplicative)
N (None)	N, N	N, A	N, M
A (Additive)	A, N	A, A	A, M
A _d (Additive damped)	A _d , N	A _d , A	A _d , M
M (Multiplicative)	M, N	M, A	M, M
M _d (Multiplicative damped)	M _d , N	M _d , A	M _d , M

Κάθε μοντέλο εκθετικής εξομάλυνσης μπορεί να γραφτεί χρησιμοποιώντας τις ακόλουθες εξισώσεις:

$$Y_t = h(x_{t-1}) + k(x_{t-1})\varepsilon_t$$

$$x_t = f(x_{t-1}) + g(x_{t-1})\varepsilon_t$$

, όπου ε_t είναι ο λευκός θόρυβος. Αν ορίσουμε επιπλέον τα ακόλουθα μεγέθη

$$x_t = (l_t, b_t, s_t, s_{t-1}, \dots, s_{t-(m-1)})$$

$$e_t = k(x_{t-1})\varepsilon_t$$

$$\mu_t = h(x_{t-1})$$

$$Y_t = \mu_t + e_t$$

, τότε για το προσθετικό μοντέλο προκύπτει ότι $Y_t = \mu_t + \varepsilon_t$, ενώ με $\mu_t = F_{(t-1)+1}$ συμβολίζεται η πρόβλεψη του μοντέλου για ορίζοντα ένα τη στιγμή $t-1$. Σε αυτήν την περίπτωση $k(x_{t-1}) = 1$. Αντίθετα, το πολλαπλασιαστικό μοντέλο γράφεται ως $Y_t = \mu_t(1 + \varepsilon_t)$ με $k(x_{t-1}) = \mu_t$ και $\varepsilon_t = e_t / \mu_t = (Y_t - \mu_t) / \mu_t$. Η μόνη διαφορά δηλαδή μεταξύ μοντέλων προσθετικού και πολλαπλασιαστικού σφάλματος είναι ότι αντικαθιστούν τον παράγοντα ε_t με $\mu_t \varepsilon_t$. Οι αναλυτικοί τύποι υπολογισμού ανά μοντέλο πρόβλεψης δίνονται στον Πίνακα 3.2.

Πίνακας 3.2: Υπολογισμός μοντέλων εκθετικής εξομάλυνσης προσθετικού σφάλματος. Για τα μοντέλα πολλαπλασιαστικού σφάλματος αντικαθιστούμε τον παράγοντα ε_t με $\mu_t \varepsilon_t$.

Συνιστώσα τάσης	Συνιστώσα εποχιακότητας		
	N (none)	A (additive)	M (multiplicative)
N (none)	$\mu_t = l_{t-1}$ $l_t = l_{t-1} + \alpha \varepsilon_t$	$\mu_t = l_{t-1} + s_{t-m}$ $l_t = l_{t-1} + \alpha \varepsilon_t$ $s_t = s_{t-m} + \gamma \varepsilon_t$	$\mu_t = l_{t-1} s_{t-m}$ $l_t = l_{t-1} + \alpha \varepsilon_t / s_{t-m}$ $s_t = s_{t-m} + \gamma \varepsilon_t / l_{t-1}$
A (additive)	$\mu_t = l_{t-1} + b_{t-1}$ $l_t = l_{t-1} + b_{t-1} + \alpha \varepsilon_t$ $b_t = b_{t-1} + \beta \varepsilon_t$	$\mu_t = l_{t-1} + b_{t-1} + s_{t-m}$ $l_t = l_{t-1} + b_{t-1} + \alpha \varepsilon_t$ $b_t = b_{t-1} + \beta \varepsilon_t$ $s_t = s_{t-m} + \gamma \varepsilon_t$	$\mu_t = (l_{t-1} + b_{t-1}) s_{t-m}$ $l_t = l_{t-1} + b_{t-1} + \alpha \varepsilon_t / s_{t-m}$ $b_t = b_{t-1} + \beta \varepsilon_t / s_{t-m}$ $s_t = s_{t-m} + \gamma \varepsilon_t / (l_{t-1} + b_{t-1})$
M (multiplicative)	$\mu_t = l_{t-1} b_{t-1}$ $l_t = l_{t-1} b_{t-1} + \alpha \varepsilon_t$ $b_t = b_{t-1} + \beta \varepsilon_t / l_{t-1}$	$\mu_t = l_{t-1} b_{t-1} + s_{t-m}$ $l_t = l_{t-1} b_{t-1} + \alpha \varepsilon_t$ $b_t = b_{t-1} + \beta \varepsilon_t / l_{t-1}$ $s_t = s_{t-m} + \gamma \varepsilon_t$	$\mu_t = (l_{t-1} b_{t-1}) s_{t-m}$ $l_t = l_{t-1} b_{t-1} + \alpha \varepsilon_t / s_{t-m}$ $b_t = b_{t-1} + \beta \varepsilon_t / (s_{t-m} l_{t-1})$ $s_t = s_{t-m} + \gamma \varepsilon_t / (l_{t-1} b_{t-1})$
Ad (Additive damped)	$\mu_t = l_{t-1} + \phi b_{t-1}$ $l_t = l_{t-1} + \phi b_{t-1} + \alpha \varepsilon_t$ $b_t = \phi b_{t-1} + \beta \varepsilon_t$	$\mu_t = l_{t-1} + \phi b_{t-1} + s_{t-m}$ $l_t = l_{t-1} + \phi b_{t-1} + \alpha \varepsilon_t$ $b_t = \phi b_{t-1} + \beta \varepsilon_t$ $s_t = s_{t-m} + \gamma \varepsilon_t$	$\mu_t = (l_{t-1} + \phi b_{t-1}) s_{t-m}$ $l_t = l_{t-1} + \phi b_{t-1} + \alpha \varepsilon_t / s_{t-m}$ $b_t = \phi b_{t-1} + \beta \varepsilon_t / s_{t-m}$ $s_t = s_{t-m} + \gamma \varepsilon_t / (l_{t-1} + \phi b_{t-1})$
Md (Multiplicative damped)	$\mu_t = l_{t-1} b_{t-1}^\phi$ $l_t = l_{t-1} b_{t-1}^\phi + \alpha \varepsilon_t$ $b_t = b_{t-1}^\phi + \beta \varepsilon_t / l_{t-1}$	$\mu_t = l_{t-1} b_{t-1}^\phi + s_{t-m}$ $l_t = l_{t-1} b_{t-1}^\phi + \alpha \varepsilon_t$ $b_t = b_{t-1}^\phi + \beta \varepsilon_t / l_{t-1}$ $s_t = s_{t-m} + \gamma \varepsilon_t$	$\mu_t = (l_{t-1} b_{t-1}^\phi) s_{t-m}$ $l_t = l_{t-1} b_{t-1}^\phi + \alpha \varepsilon_t / s_{t-m}$ $b_t = b_{t-1}^\phi + \beta \varepsilon_t / (s_{t-m} l_{t-1})$ $s_t = s_{t-m} + \gamma \varepsilon_t / (l_{t-1} b_{t-1}^\phi)$

Όπως γίνεται κατανοητό, προκειμένου να υπολογιστεί το εκάστοτε μοντέλο εκθετικής εξομάλυνσης πρέπει αρχικά να ορισθούν οι συντελεστές εξομάλυνσης $\theta = (\alpha, \beta, \gamma, \phi)$ και τα αρχικά επίπεδα $X_0 = (l_0, b_0, s_0, s_{-1}, s_{-m+1})$. Για να γίνει κάτι τέτοιο μπορεί κανείς να

μεγιστοποιήσει το κριτήριο τη πιθανοφάνειας (*maximum likelihood*) ή εναλλακτικά να ελαχιστοποιήσει την τιμή

$$L^*(\theta, X_0) = n \log \left(\sum_{t=1}^n \frac{e_t^2}{k(x_{t-1})^2} \right) + 2 \sum_{t=1}^n \log |k(x_{t-1})|$$

η οποία ισούται με δύο φορές την αρνητική τιμή του λογαρίθμου της πιθανοφάνειας. Βασικό πλεονέκτημα της εν λόγω τεχνικής είναι ότι βελτιστοποιεί ταυτόχρονα όλες τις παραμέτρους του μοντέλου χωρίς να χρειάζεται να ορισθούν εξαρχής με αυθαίρετο τρόπο τα αρχικά επίπεδα για τον υπολογισμό των συντελεστών εξομάλυνσης, μία πρακτική που εφαρμοζόταν ευρέως παλαιότερα με χρήση εμπειρικών κανόνων. Πλεονεκτεί επίσης έναντι της απλής γραμμικής αναζήτησης (*grid search*) κατά την οποία αναζητείται από διακριτές τιμές η πλέον κατάλληλη τιμή παραμέτρου, καθώς σε αυτή την περίπτωση όλες οι μεταβλητές είναι συνεχείς.

Σημειώνεται για τις τιμές των παραμέτρων ότι, αν και συνήθως λαμβάνουν τιμές στο διάστημα (0,1), πρακτικά ισχύουν για λόγους ευστάθειας οι ακόλουθοι περιορισμοί:

$$0.1 \leq \alpha \leq 0.9, 0.1 \leq \beta \leq 0.9, 0.1 \leq \gamma \leq 0.9 \text{ και } \beta \leq \varphi \leq 1$$

Άλλες εναλλακτικές για την παραμετροποίηση των μοντέλων είναι η ελαχιστοποίηση κάποιου σφάλματος προσαρμογής του μοντέλου (π.χ. του MSE ή του MAPE) ή η ελαχιστοποίηση της διακύμανσης των υπολειπόμενων σφαλμάτων. Ωστόσο, βιβλιογραφικά, η μεγιστοποίηση της πιθανοφάνειας προτιμάται καθώς μπορεί να συνδυαστεί στη συνέχεια με κριτήρια πολυπλοκότητας (*information criteria*) όπως το *AIC* (*Akaike information criterion*) και το *BIC* (*Bayesian information criterion*) και να οδηγήσει στην επιλογή του απλούστερου, πλην ακριβούς μοντέλου πρόβλεψης βάσει προσαρμογής του στα ιστορικά δεδομένα (βλ. ενότητα 3.4).

Αξίζει να σημειωθεί ότι λόγω της απλότητας και της αποτελεσματικότητάς τους, οι μέθοδοι εκθετικής εξομάλυνσης βρίσκονται υλοποιημένες σε αρκετά δημοφιλή στατιστικά πακέτα, χαρακτηριστικότερα από τα οποία είναι αυτά της *R*. (*Hyndman, 2014 και Hyndman & Khandakar, 2008*).

Στη συνέχεια παρουσιάζονται αναλυτικά τα τρία βασικότερα και πλέον εμφανιζόμενα στη βιβλιογραφία μοντέλα εκθετικής εξομάλυνσης.

3.3.3.1 Απλή εκθετική εξομάλυνση σταθερού επιπέδου (Simple Exponential Smoothing)

Η απλή εκθετική εξομάλυνση SES, όπως καλείται το μοντέλο σταθερού επιπέδου, έχει ως βασική της υπόθεση την απουσία τάσης από τα δεδομένα, δηλαδή την ύπαρξη στασιμότητας. Δεδομένου μάλιστα ότι δεν υποθέτει εποχιακή συμπεριφορά, εκφράζεται επί της ουσίας μέσω του μοντέλου *ANV* της μεθόδου εκθετικής εξομάλυνσης. Έτσι, η εκάστοτε πρόβλεψη του μοντέλου προκύπτει προεκτείνοντας με μια οριζόντια ευθεία γραμμή το τρέχον επίπεδο της χρονοσειράς. Αυτό δεν είναι άλλο από το σταθμισμένο μέσο όρο των παρατηρήσεων της χρονοσειράς, τα βάρη των οποίων εξαρτώνται από την παράμετρο εξομάλυνσης.

Για την παραγωγή πρόβλεψης χρειάζεται πρώτα απ' όλα να οριστεί το αρχικό επίπεδο της χρονοσειράς l_0 , το οποίο όπως παρουσιάστηκε νωρίτερα μπορεί να γίνει με αρκετές εναλλακτικές. Κάθε νέα τιμή του προβλεπόμενου μεγέθους ισούται με το αμέσως προηγούμενο επίπεδο που υπολογίζεται. Ο υπολογισμός του επιπέδου χρονοσειράς γίνεται σε κάθε χρονική περίοδο και εξομαλύνεται από το σφάλμα της διαδικασίας προσαρμογής του μοντέλου ε_t . Η διαδικασία αυτή επαναλαμβάνεται για όλα τα καταγεγραμμένα δεδομένα της χρονοσειράς και οι προβλέψεις του μοντέλου ισούνται με την τελευταία υπολογισμένη τιμή του επιπέδου της χρονοσειράς. Οι εξισώσεις που περιγράφουν το συγκεκριμένο μοντέλο πρόβλεψης είναι οι εξής:

$$l_t = l_{t-1} + a \cdot \varepsilon_t$$

$$F_t = \mu_t = l_{t-1}$$

Σημειώνεται ότι σε αυτήν την περίπτωση ο βέλτιστος συντελεστής εξομάλυνσης καθορίζεται από δύο αλληλεξαρτώμενους παράγοντες. Ο πρώτος είναι το ποσοστό θορύβου της χρονοσειράς, η τιμή του οποίου συνήθως επηρεάζει αντιστρόφως ανάλογα την τιμή του συντελεστή εξομάλυνσης ούτως ώστε το μοντέλο να μην αντιδρά υπερβολικά σε ακραίες, μη αντιπροσωπευτικές τιμές. Ο δεύτερος παράγοντας είναι η στασιμότητα, δηλαδή η σταθερότητα του μέσου επιπέδου της χρονοσειράς. Όσο ασθενέστερη είναι η στασιμότητα της χρονοσειράς, τόσο μεγαλύτερος πρέπει να είναι ο συντελεστής εξομάλυνσης ούτως ώστε να μπορεί το μοντέλο να παρακολουθεί τις μεταβολές που παρουσιάζονται στα δεδομένα. Έτσι, για τιμή παραμέτρου $a=1$ η πρόβλεψη είναι ίση με την τελευταία τιμή της χρονοσειράς και το μοντέλο ταυτίζεται ουσιαστικά με την απλοϊκή μέθοδο πρόβλεψης. Στην περίπτωση πάλι που $a=0$, οι προβλεπόμενες τιμές ισούνται όλες με το αρχικό επίπεδο της χρονοσειράς.

3.3.3.2 Εκθετική Εξομάλυνση Γραμμικής Τάσης (Holt Exponential Smoothing)

Το μοντέλο της εξομάλυνσης γραμμικής τάσης μοιάζει πολύ με αυτό της SES με τη διαφορά ότι λαμβάνει επιπροσθέτως υπόψη του τη τάση της χρονοσειράς. Πρόκειται δηλαδή για προέκταση του μοντέλου σταθερού επιπέδου και θεωρητικά μπορεί να προσδώσει αρκετά καλύτερα αποτελέσματα από αυτό, δεδομένου ότι ταυτόχρονα με την προσαρμογή του επιπέδου της χρονοσειράς υπάρχει και προσαρμογή στη τάση αυτής. Θα μπορούσε κανείς να πει επίσης ότι η Holt πλησιάζει τη φιλοσοφία της απλής γραμμικής παλινδρόμησης, ωστόσο λόγω του συστήματος επαναπροσδιορισμού που αυτή διαθέτει δίνει σταδιακά όλο και μεγαλύτερη έμφαση στις πρόσφατες παρατηρήσεις σε σύγκριση με τις αρχικές. Οι προβλέψεις της συγκεκριμένης μεθόδου γίνονται με τις παρακάτω εξισώσεις:

$$l_t = l_{t-1} + b_{t-1} + a\varepsilon_t$$

$$b_t = b_{t-1} + b\varepsilon_t$$

$$F_t = \mu_t = l_{t-1} + b_{t-1}$$

Όπως φαίνεται, η τελική πρόβλεψη του μοντέλου είναι το άθροισμα της τρέχουσας τάσης και του τρέχοντος επιπέδου της χρονοσειράς και αντιστοιχεί στο μοντέλο AAN της μεθόδου εκθετικής εξομάλυνσης.

3.3.3.3 Εκθετική Εξομάλυνση Μη Γραμμικής Τάσης (Damped)

Το μοντέλο εκθετικής εξομάλυνσης μη γραμμικής τάσης, το οποίο και αντιστοιχεί στο μοντέλο AA_dN της μεθόδου εκθετικής εξομάλυνσης, αποτελεί προέκταση του μοντέλου Holt προκειμένου να μπορέσουν να περιγράψουν χρονοσειρές που παρουσιάζουν μη γραμμικές διακυμάνσεις επιπέδου, δηλαδή μεταβαλλόμενη τάση. Αυτό γίνεται μέσω της εισαγωγής μίας επιπρόσθετης παραμέτρου φ , γνωστή και ως παραμέτρου διόρθωσης τάσης. Η αύξηση του εύρους των χρονοσειρών που μπορούμε να επεξεργαστούμε αποτελεσματικά με μία τέτοια μη γραμμική μέθοδο γίνεται εύκολα αντιληπτή, ενώ έχει διαπιστωθεί και εμπειρικά ότι η Damped παράγει πολύ πιο ακριβή αποτελέσματα σε σχέση με την SES και την Holt, κυρίως όταν δεν είναι εύκολος ο εντοπισμός του μοτίβου τάσης που ακολουθεί μία χρονοσειρά. Άλλο πλεονέκτημα της μεθόδου Damped είναι ότι, σε αντίθεση με τις υπόλοιπες δύο μεθόδους εκθετικής εξομάλυνσης, αυτή μπορεί να δώσει πολύ καλύτερες προβλέψεις και σε μακροπρόθεσμο επίπεδο καθώς αναθεωρεί συστηματικά τόσο το επίπεδο όσο και τον ρυθμό ανάπτυξης της χρονοσειράς. Οι εξισώσεις στις οποίες βασίζεται το μοντέλο είναι οι παρακάτω:

$$l_t = l_{t-1} + \varphi b_{t-1} + a\varepsilon_t$$

$$b_t = \varphi b_{t-1} + b\varepsilon_t$$

$$F_t = \mu_t = l_{t-1} + \varphi b_{t-1}$$

Όπως φαίνεται οι εξισώσεις της Damped είναι πανομοιότυπες με αυτές της Holt και γι' αυτό δεν θα γίνει ξανά κάποια επεξήγηση παρά μόνο στη μοναδική τους διαφοροποίηση που την προσφέρει η παράμετρος φ . Έτσι, ο συντελεστής φ προκαλεί πλέον μη γραμμικό υπολογισμό της πρόβλεψης F_t , σε αντίθεση με πριν όπου η τάση αυξανόταν γραμμικά σε κάθε περίοδο. Ανάλογα τη τιμή του φ μπορούμε να προσεγγίσουμε διαφορετικού μοτίβου χρονοσειρές, αυξανόμενης τάσης, σταθερού επιπέδου και φθίνουσας τάσης. Συγκεκριμένα, για $\varphi < 0$ προκύπτει φθίνουσα τάση και το μέγεθος κατά το οποίο αυξάνει η τιμή των προβλέψεων μικραίνει κάθε χρονική περίοδο, ενώ για $\varphi > 0$ προκύπτει εκθετική τάση και το μέγεθος κατά το οποίο αυξάνει η τιμή των προβλέψεων μεγαλώνει κάθε φορά. Τέλος, για $\varphi = 0$ έχουμε εφαρμογή της μεθόδου SES και για $\varphi = 1$ εφαρμογή του μοντέλου γραμμικής τάσης.

3.3.4 Αυτοπαλινδρομούμενα μοντέλα κινητού μέσου όρου (ARMA)

Τα αυτόπαλινδρομούμενα μοντέλα κινητών μέσων όρων $ARMA$ (Auto Regressive Moving Average) ανήκουν στα στοχαστικά στατιστικά μοντέλα πρόβλεψης. Η εφαρμογή τους βασίζεται αποκλειστικά στον υπολογισμό της πιθανότητας με την οποία η τιμή του μεγέθους που καλούμαστε να προβλέψουμε βρίσκεται μία δεδομένη στιγμή εντός κάποιου διαστήματος, ενώ προ-απαιτούμενα αποτελούν η ύπαρξη μίας διακριτής και στάσιμης χρονοσειράς, καθώς και η παραγωγή βραχυπρόθεσμων προβλέψεων. Τα μοντέλα $ARMA$

μελετήθηκαν εκτεταμένα από τους *Box και Jenkins (1970)* και συχνά συναντώνται στη βιβλιογραφία με την ομώνυμη ονομασία.

Στην γενική τους μορφή τα μοντέλα *ARMA* περιέχουν τον τυχαίο παράγοντα, τιμές του μεγέθους που εμφανίστηκαν σε προηγούμενες περιόδους και σχετικούς στοχαστικούς παράγοντες. Κάθε μοντέλο μπορεί να εκφραστεί ως γραμμικός συνδυασμός των παραπάνω παραγόντων και στόχος μας είναι να ανακαλύψουμε εκείνον που παράγει τις καλύτερες προβλέψεις (*Hyndman & Athanasopoulos, 2016 κεφάλαιο 8.5*). Έτσι, με την έκφραση *ARMA(p,q)* συμβολίζουμε ένα μοντέλο το οποίο για να προβλέψει τη στιγμή t λαμβάνει υπόψη του τις p προηγούμενες τιμές της χρονοσειράς (*παράγοντας AR*) και τις διορθώνει εξετάζοντας τα σφάλματα που παρήγαγε κατά τις q προηγούμενες προβλέψεις του (*παράγοντας MA*). Στην περίπτωση που η χρονοσειρά που μελετάται δεν είναι στάσιμη, τα μοντέλα *ARMA* επεκτείνονται μέσω της διαφορίσης (βλ. *ενότητα 2.4.1*) στα ολοκληρωμένα (*Integrated*) αυτόπαλινδρομούμενα μοντέλα κινητών μέσω των *ARIMA(p,d,q)* τα οποία και αποτελούν την πλέον γενικευμένη τους έκφραση. Σε αυτή την περίπτωση ο παράγοντας d φανερώνει το βαθμό διαφορίσης της χρονοσειράς και οι παράγοντες p και q αφορούν πλέον τη διαφορισμένη χρονοσειρά. Μάλιστα, με τη χρήση εποχιακής διαφορίσης, την απομάκρυνση δηλαδή των συστηματικών εποχιακών διακυμάνσεων, τα μοντέλα *ARIMA* δύναται να περιγράψουν πλέον αποτελεσματικά και εποχιακές χρονοσειρές

Η επιλογή του καταλληλότερου μοντέλου *ARIMA* δεν είναι πάντα προφανής. Συχνά περισσότερα από ένα μοντέλα μπορούν να οδηγούν σε παραπλήσιες προβλέψεις, αφήνοντας την τελική επιλογή στην κρίση μας. Επίσης, ενδέχεται κάποιο μοντέλο να προσαρμόζεται καλύτερα από ένα άλλο σε μία χρονοσειρά αλλά η πολυπλοκότητά του να είναι σημαντικά μεγαλύτερη και για αυτό το λόγο να απορριφτεί. Θέλοντας ωστόσο να αυτοματοποιήσουμε με κάποιον τρόπο τη διαδικασία εύρεσης του βέλτιστου μοντέλου *ARIMA*, ακολουθούμε την παρακάτω διαδικασία που περιλαμβάνει τρία στάδια: την *αναγνώριση*, την *εκτίμηση* και την *διάγνωση*.

1. Στο στάδιο της *αναγνώρισης* επιλέγονται ένα ή περισσότερα μοντέλα *ARIMA* τα οποία θεωρούμε βάση κάποιων στοιχείων ότι μπορούν να περιγράψουν ικανοποιητικά την χρονοσειρά. Αναγνωρίζουμε δηλαδή τα υποψήφια μοντέλα καθορίζοντας τους παράγοντες p , d και q . Στην πιο απλουστευμένη της μορφή η αναγνώριση μπορεί να πραγματοποιηθεί παρατηρώντας την εξέλιξη των συντελεστών αυτοσυσχέτισης *ACF* και μερικής αυτοσυσχέτισης *PACF* αυτής σε διαγράμματα. Συνοπτικά αναφέρουμε τα εξής:
 - ✓ Για ένα στάσιμο μοντέλο *AR(p)*:
 - Οι τιμές των συντελεστών *ACF* φθίνουν προς το μηδέν ακολουθώντας εκθετική ή ημιτονοειδή πορεία
 - Οι τιμές των συντελεστών *PACF* μηδενίζονται απότομα μετά από p περιόδους υστέρησης
 - ✓ Για ένα στάσιμο μοντέλο *MA(q)*

- Οι τιμές των συντελεστών ACF μηδενίζονται απότομα μετά από q περιόδους υστέρησης
 - Οι τιμές των συντελεστών $PACF$ φθίνουν προς το μηδέν ακολουθώντας εκθετική ή ημιτονοειδή πορεία
 - ✓ Για ένα στάσιμο μοντέλο $ARIMA(p,q)$
 - Οι τιμές των συντελεστών ACF φθίνουν προς το μηδέν μετά από $q-p$ περιόδους υστέρησης
 - Οι τιμές των συντελεστών $PACF$ φθίνουν προς το μηδέν μετά από $p-q$ περιόδους υστέρησης
2. Στο στάδιο της *εκτίμησης* υπολογίζουμε για κάθε ένα από τα υποψήφια μοντέλα τις παραμέτρους τους και τα υλοποιούμε. Αυτό μπορεί να γίνει με αρκετούς τρόπους, ο πιο διαδεδομένος εκ των οποίων είναι, όπως και στην περίπτωση των μοντέλων εκθετικής εξομάλυνσης, η μεγιστοποίηση της προσδοκώμενης πιθανοφάνειας (*likelihood*).

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

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

- ,όπου L η προσδοκώμενη πιθανοφάνεια ταύτισης του μοντέλου με τα αρχικά δεδομένα, $F(X_t)$ η προβλεπόμενη από το μοντέλο τιμή τη περίοδο t , n ο αριθμός παρατηρήσεων, e_t το σφάλμα πρόβλεψης, σ^2 η διακύμανση των σφαλμάτων του μοντέλου και RSS (*Round Sum of Squares*) το άθροισμα των τετραγωνικών σφαλμάτων του μοντέλου.
3. Στο στάδιο του *διαγνωστικού ελέγχου* εφαρμόζονται στατιστικά τεστ και άλλοι ποιοτικοί έλεγχοι προκειμένου να εξακριβωθεί ο βαθμός στον οποίο τα μοντέλα που αναγνωρίστηκαν και εκτιμήθηκαν είναι κατάλληλα. Ο διαγνωστικός έλεγχος προσαρμογής του μοντέλου πάνω στα αρχικά δεδομένα γίνεται μελετώντας την κατανομή των αντίστοιχων σφαλμάτων πρόβλεψης. Αν το μοντέλο είναι αρκετά περιγραφικό, τότε τα σφάλματα που παράγει θα *πρέπει να οφείλονται αποκλειστικά στην τυχαιότητα* και συνεπώς να μην συσχετίζονται μεταξύ τους με κάποιον τρόπο, ή αλλιώς:

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

, όπου p_k η συσχέτιση των σφαλμάτων για υστέρηση (διάστημα μεταξύ δύο παρατηρήσεων) k . Προφανώς τα σφάλματα ενός μοντέλου $ARIMA$ ποτέ δεν είναι τελείως ασυσχέτιστα, όσο καλά και αν αυτό περιγράφει τη χρονοσειρά. Αναμένεται λοιπόν για κάποιες υστερήσεις να υπάρξουν μη μηδενικούς δείκτες συσχέτισης. Για να ελέγξουμε αν αυτοί είναι σημαντικά διάφοροι του μηδενός, υπολογίζονται οι t -τιμές του τυπικού σφάλματός τους $S(r_k(e))$.

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

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

Για να μην είναι σημαντική μία συσχέτιση (π.χ. να ανήκει στο 5% της κανονικής κατανομής πιθανότητας) η τιμή του t-δείκτη δεν πρέπει θεωρητικά να ξεπερνά την τιμή 2. Στην πράξη, για υστέρηση 1, 2 και 3 πρέπει να είναι μικρότερη του 1.25 και για μεγαλύτερη υστέρηση μικρότερη του 1.6.

Άλλα κριτήρια που μπορούν να μας βοηθήσουν στην επιλογή του καταλληλότερου μοντέλου *ARIMA* είναι τα κριτήρια *AIC* (*Akaike, 1973*) και *BIC* (*Schwarz, 1978*). Πρόκειται για κριτήρια τα οποία αξιολογούν κατά πόσο αποδίδει το μοντέλο που εξετάζεται στη χρονοσειρά συναρτήσει της πολυπλοκότητάς του. Φανερώνουν έτσι κατά πόσο αξίζει το μοντέλο που θα επιλεγεί να γίνει περισσότερο πολύπλοκο προκειμένου να αυξηθεί η ακρίβειά του. Η ποσοτικοποίηση της εν λόγω υπόθεσης είναι εξαιρετικά κρίσιμη καθώς όσο αυξάνεται η πολυπλοκότητα ενός μοντέλου τόσο βελτιώνεται η προσαρμογή του (*fitting*), δηλαδή μειώνεται η προκατάληψή του. Ωστόσο, η αύξηση της πολυπλοκότητας οδηγεί ταυτόχρονα σε συστηματική αύξηση της διακύμανσης των σφαλμάτων του και συνεπώς σε έλλειψη ακρίβειας πρόβλεψης (*over-fitting*). Στόχος μας είναι η επίτευξη όσο το δυνατόν μικρότερης προκατάληψης και υψηλότερης ακρίβειας (βλ. ενότητα 2.5.1). Μειονέκτημα των εν λόγω κριτηρίων αποτελεί ωστόσο το γεγονός, ότι επειδή δεν έχουν ως βάση τους κάποια συγκεκριμένη υπόθεση ακρίβειας (π.χ. επίτευξη μηδενικού σφάλματος), δεν μας πληροφορούν άμεσα για το αν ένα μοντέλο είναι κατάλληλο, παρά μόνο για το ποιο μοντέλο είναι το καλύτερο από αυτά που εξετάστηκαν. Έτσι, η επιλογή του βέλτιστου μοντέλου γίνεται συγκρίνοντας την τιμή των κριτηρίων για όλα τα υποψήφια μοντέλα. Ο υπολογισμός τους για χρονοσειρά μήκους n δίνεται παρακάτω.

$$AIC = -2\log L + 2(p + q + k + 1)$$

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

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

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

Το *BIC* λειτουργεί σαν το *AIC*, δίνοντας όμως μεγάλο βάρος στην πολυπλοκότητα του μοντέλου για να αποφευχθούν φαινόμενα υπερ-προσαρμογής.

$$BIC = AIC + \log(n)(p + q + k + 1)$$

3.3.4.1 Αυτόπαλινδρομούμενα μοντέλα - AR(p)

Τα αυτόπαλινδρομούμενα μοντέλα θεωρούν την ύπαρξη μίας γραμμικής σχέσης ανάμεσα στην παρατήρηση της χρονοσειράς που εξετάζεται και στις προηγούμενες αυτής. Ένα τέτοιο μοντέλο p τάξης αναπαρίσταται αλγεβρικά ως εξής:

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

,όπου φ_i οι συντελεστές αυτοσυσχέτισης του μοντέλου AR για υστέρηση i και c μία σταθερά για την οποία αποδεικνύεται ότι ισχύει $c = \mu(1 - \varphi_1 - \varphi_2 - \dots - \varphi_p)$. Χρησιμοποιώντας τον τελεστή ολίσθησης το μοντέλο μπορεί να γραφτεί και ως:

$$(1 - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_p B^p) \bar{y}_t = e_t$$

, όπου $\bar{y}_t = y_t - \mu$.

3.3.4.2 Μοντέλα Κινητού Μέσου Όρου - MA(q)

Τα μοντέλα κινητού μέσου όρου θεωρούν γραμμικές σχέσεις ανάμεσα στην παρατήρηση της χρονοσειράς που εξετάζεται και στα σφάλματα που εμφάνισε το μοντέλο σε προηγούμενες περιόδους. Ένα τέτοιο μοντέλο γράφεται αλγεβρικά ως εξής:

$$y_t = c - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \dots - \theta_q e_{t-q} + e_t$$

,όπου θ_i οι συντελεστές μερικής αυτοσυσχέτισης του μοντέλου MA για υστέρηση i και $c = \mu$. Χρησιμοποιώντας τον τελεστή ολίσθησης, ένα μοντέλο MA μπορεί να γραφτεί και ως:

$$\bar{y}_t = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q) e_t$$

3.3.4.3 Μοντέλα ARIMA (p,d,q)

Τα μοντέλα AR και MA συνδυάζονται εισάγοντας στην εξίσωση τους ένα μοντέλο διαφορίσης για τη διασφάλιση της στασιμότητας. Το συνολικό μοντέλο αναπαρίσταται με τη χρήση του τελεστή ολίσθησης B ως εξής:

$$(1 - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_p B^p)(1 - B)^n(1 - B^m)^N y_t = c + (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q) e_t$$

, όπου n η τάξη διαφορίσης, N η τάξη εποχιακής διαφορίσης και m η συχνότητα της χρονοσειράς. Αποδεικνύεται ότι για $n=N=0$ η σταθερά c ισούται με $\mu(1 - \varphi_1 - \varphi_2 - \dots - \varphi_p)$, ενώ σε οποιαδήποτε άλλη περίπτωση η σταθερά ισούται με μηδέν.

Πρακτικά η χρονοσειρά που προκύπτει από τη διαφορίση δεν είναι ποτέ απολύτως στάσιμη γύρω από το μηδέν. Έτσι, συνηθίζεται να προστίθεται σταθερά ακόμα και σε χρονοσειρές που έχουν υποστεί διαφορίση. Συνοπτικά αναφέρουμε:

- Η μη διαφορίση ($d=0$) είναι δείγμα ύπαρξης σταθερότητας στην αρχική χρονοσειρά. Ωστόσο αυτή η σταθερότητα ενδέχεται εν τέλει να είναι ανεπαρκής. Η εισαγωγή λοιπόν μίας σταθεράς c για τον καλύτερο προσδιορισμό του επιπέδου της συνίσταται.

- Η διαφορίση πρώτης τάξης ($d=1$) συνίσταται κατά την ύπαρξη σταθερής τάσης στην αρχική χρονοσειρά. Αυτή θεωρητικά θα εξαλειφτεί μετά τη διαφορίση, ωστόσο ενδέχεται να έχει υποπέσει στην προηγούμενη περίπτωση. Έτσι, η εισαγωγή σταθεράς μπορεί να γίνει έχοντας συμβουλευτεί πρώτα κάποιο από τα κριτήρια πολυπλοκότητας.
- Η διαφορίση δεύτερης τάξης ($d=2$) συνεπάγεται ύπαρξη χρονικά μεταβαλλόμενης τάσης στην αρχική χρονοσειρά. Έτσι, η εισαγωγή σταθεράς σε αυτήν την περίπτωση θεωρείται άστοχη επιλογή.
- Η διαφορίση πρακτικά λαμβάνει τάξη μικρότερη ή ίση του δύο και η εποχιακή μικρότερη ή ίση του ένα.

Παρακάτω δίνονται κάποιοι περιορισμοί που πρέπει να πληρούνται για τα πιο διαδεδομένα μοντέλα *ARIMA*. Για τα υπόλοιπα μοντέλα (που είναι και περισσότερο σύνθετα) οι αντίστοιχοι περιορισμοί γίνονται ιδιαίτερα σύνθετοι και η παρουσίασή τους κρίνεται άσκοπη.

- AR(1): $-1 < \varphi_1 < 1$
- AR(2): $-1 < \varphi_2 < 1$ και $\varphi_1 + \varphi_2 < 1$ και $\varphi_2 - \varphi_1 < 1$
- MA(1): $-1 < \theta_1 < 1$
- MA(2): $-1 < \theta_2 < 1$ και $\theta_1 + \theta_2 > -1$ και $\theta_1 - \theta_2 < 1$
- Για μοντέλα AR(p) με $p > 2$ πρέπει τουλάχιστον να ισχύει $\varphi_1 + \varphi_2 + \varphi_3 + \dots + \varphi_p < 1$

Επίσης, δεδομένων των τιμών αυτοσυσχέτισης μίας χρονοσειράς, για τους συντελεστές των μοντέλων *ARIMA* ισχύει:

Πίνακας 3.3: Εκτίμηση παραμέτρων μοντέλων *ARIMA* μέσω της χρήσης των συντελεστών αυτοσυσχέτισης.

ARMA(p, q)	ρ_1	ρ_2
AR(1)	φ_1	-
MA(1)	$-\theta_1$	-
AR(2)	$\frac{\varphi_1}{1 + \theta_1^2}$	$\frac{\varphi_1^2}{1 - \varphi_2} + \varphi_2$
MA(2)	$\frac{-\theta_1(1 - \theta_2)}{1 + \theta_1^2 + \theta_2^2}$	$\frac{-\theta_2}{1 + \theta_1^2 + \theta_2^2}$
ARMA(1, 1)	$\frac{(1 - \varphi_1\theta_1)(\varphi_1 - \theta_1)}{1 + \theta_1^2 - 2\varphi_1\theta_1}$	$\rho_1\varphi_1$

Γενικά αποδεικνύεται ότι για αμιγώς *MA(q)* διαδικασίες ισχύει για τη συνάρτηση αυτοσυσχέτισης:

$$\rho_k = \frac{-\theta_k + \theta_{k+1}\theta_1 + \dots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \dots + \theta_q^2} \quad \text{για } k=1, 2, \dots, q \text{ και } \rho_k = 0 \text{ για } k > q.$$

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

$$\rho_k = \rho_1^k \text{ για } k=1,2,\dots,p$$

3.3.4.4 Πρόβλεψη με μοντέλα ARIMA

Όταν θέλουμε να υπολογίσουμε μέσω ενός μοντέλου ARIMA την τιμή της χρονοσειράς y την περίοδο t , τότε απαιτείται γνώση των τιμών $y_{t-1}, y_{t-2}, \dots, y_{t-p}$ ή/και των τιμών $\theta_{t-1}, \theta_{t-2}, \dots, \theta_{t-q}$. Για την πρόβλεψη λοιπόν της τιμής \widehat{y}_{t+T} , απαιτείται αντίστοιχα γνώση των τιμών $y_{t+T-1}, y_{t+T-2}, \dots, y_{t+T-n}$ ή/και των τιμών $\theta_{t+T-1}, \theta_{t+T-2}, \dots, \theta_{t+T-n}$, όπου T ο ορίζοντας πρόβλεψης. Για ένα μοντέλο π.χ. ARMA(1,1), αν θελήσουμε να προβλέψουμε 3 περιόδους μπροστά απαιτείται γνώση των τιμών y_t και θ_t για τη πρώτη πρόβλεψη, των τιμών y_{t+1} και θ_{t+1} για τη δεύτερη και των τιμών y_{t+2} και θ_{t+2} για τη τρίτη.

Εδώ βλέπουμε ότι εμφανίζεται αμέσως ένα πρόβλημα: Οι τιμές y_t και θ_t που πρέπει να χρησιμοποιηθούν για τη πρώτη πρόβλεψη είναι διαθέσιμες. Δεν ισχύει όμως το ίδιο και για τις y_{t+1} και θ_{t+1} που απαιτούνται στην πρόβλεψη του δεύτερου διαστήματος. Για να γίνει λοιπόν πρόβλεψη θα πρέπει να θεωρήσουμε ως y_{t+1} την νωρίτερα εκτιμημένη τιμή της χρονοσειράς \widehat{y}_{t+1} από το μοντέλο και μηδενικό σφάλμα. Αντίστοιχα για την τρίτη πρόβλεψη θεωρούμαι $y_{t+2} = \widehat{y}_{t+2}$ και μηδενικό σφάλμα. Μακροχρόνια το μοντέλο ARIMA εκφυλίζεται δηλαδή σε μοντέλο AR εξαρτώμενο μόνο από τις προβλέψεις που έχει κάνει το ίδιο και όχι από τις τιμές των δεδομένων. Αυτός είναι και ο λόγος που χρησιμοποιείται κυρίως για βραχυπρόθεσμες προβλέψεις.

Αξίζει να σημειωθεί ότι λόγω της αποτελεσματικότητάς τους σε βραχυπρόθεσμες προβλέψεις, οι μέθοδοι εκθετικής εξομάλυνσης βρίσκονται υλοποιημένες σε αρκετά δημοφιλή στατιστικά πακέτα, χαρακτηριστικότερα από τα οποία είναι αυτά της R. (Hyndman & Khandakar, 2008).

3.4 Η μέθοδος θ (Theta method)

Σε αυτό το σημείο θα παρουσιαστεί αναλυτικά η μέθοδος θ ούτως ώστε να γίνει αντιληπτή η ευρύτερη φιλοσοφία της και να αναδειχτούν τα πλεονεκτήματα και τα μειονεκτήματά της. Μέσω μίας εκτενούς βιβλιογραφικής ανασκόπησης που θα πραγματοποιηθεί θα εξακριβωθούν πιθανές προεκτάσεις της που θα οδηγήσουν στη γενίκευσή της και στην αποδοτικότερη χρήση της μέσω της καλύτερης προσαρμογής της σε χρονοσειρές ποικίλων χαρακτηριστικών.

3.4.1 Συνοπτική παρουσίαση της μεθόδου και βιβλιογραφική επισκόπηση

Η μέθοδος θ (Assimakopoulos & Nikolopoulos, 2000) είναι μία μονοδιάστατη μέθοδος πρόβλεψης η οποία βασίζεται στη μεταβολή των τοπικών διακυμάνσεων της χρονοσειράς μέσω της εφαρμογής της παραμέτρου θ στις πρώτες διαφορές της. Ο μετασχηματισμός οδηγεί στη δημιουργία νέων γραμμών, γνωστές ως γραμμές θ , οι οποίες διατηρούν τη μέση τιμή και την κλίση των αρχικών δεδομένων αλλά όχι και την καμπυλότητά τους. Όσο πιο μικρή γίνεται η τιμή της παραμέτρου θ , τόσο μειώνονται οι τοπικές διακυμάνσεις της χρονοσειράς, και το αντίστροφο. Έτσι, χρησιμοποιώντας μικρές τιμές της παραμέτρου $0 < \theta < 1$ μπορούμε να εντοπίσουμε στις αντίστοιχες γραμμές θ μακροπρόθεσμα χαρακτηριστικά της αρχικής χρονοσειράς όπως είναι η τάση (Assimakopoulos, 1995), και εκτιμώντας γραμμές με τιμή παραμέτρου $\theta > 1$ βραχυπρόθεσμα χαρακτηριστικά όπως

είναι το τρέχον επίπεδο. Η παράμετρος της μεθόδου μπορεί να θεωρηθεί λοιπόν στη πράξη ως ένας μετασχηματισμός καμπυλότητας η επίδραση του οποίου εξαρτάται από την απόσταση της εκάστοτε παρατήρησης από αυτήν της ευθείας της απλής γραμμικής παλινδρόμησης την ίδια χρονική περίοδο η οποία και υπολογίζεται για $\theta=0$. Έτσι, εκμεταλλευόμενοι τις ιδιότητες της κάθε γραμμής θ , καθίσταται εφικτή η αποσύνθεση της χρονοσειράς σε δύο ή περισσότερες γραμμές κάθε μία εκ των οποίων τονίζει διαφορετικά χαρακτηριστικά της αρχικής χρονοσειράς, η προέκτασή τους με κατάλληλο τρόπο και ο συνδυασμός των επιμέρους προβλέψεων για την αποτελεσματικότερη πρόβλεψη του ζητούμενου μεγέθους.

Στην κλασική της μορφή, η μέθοδος θ αποτελείται από δύο γραμμές με τιμές παραμέτρου μηδέν και δύο υπολογισμένες επί των αποεποχικοποιημένων δεδομένων. Η γραμμή μηδέν, η οποία ταυτίζεται με αυτήν της απλής γραμμικής παλινδρόμησης, έχει μηδενική καμπυλότητα ενώ η γραμμή δύο διπλάσια καμπυλότητα της αρχικής. Έτσι, η πρώτη γραμμή προβλέπεται επεκτείνοντας την ευθεία της απλής γραμμικής παλινδρόμησης και η δύο με χρήση της μεθόδου SES. Οι προβλέψεις συνδυάζονται με ίσα βάρη και στη συνέχεια εποχικοποιούνται χρησιμοποιώντας κατάλληλους δείκτες. Σημειώνεται ότι, βάσει όσων αναφέρθηκαν νωρίτερα, σε περίπτωση που πραγματοποιούνται μακροχρόνιες προβλέψεις μπορεί κανείς να αυξήσει το βάρος της γραμμής μηδέν για να λάβει σε μεγαλύτερο βαθμό υπόψη την μακροχρόνια τάση της χρονοσειράς. Το αντίστροφο μπορεί να γίνει στην περίπτωση παραγωγής βραχυπρόθεσμων προβλέψεων.

Η κλασική μέθοδος θ έγινε ιδιαίτερα γνωστή με τη συμμετοχή της στο διαγωνισμό *M3* (*Makridakis & Hibon, 2000*) καθώς αναδείχτηκε ως το πλέον ακριβές μοντέλο πρόβλεψης με εξαιρετικά αποτελέσματα στην περίπτωση των μηνιαίων και μικροοικονομικών δεδομένων. Αξιοσημείωτο ήταν επίσης το γεγονός ότι η μέθοδος κατάφερε να ξεπεράσει άλλες σαφώς πιο πολύπλοκες μεθόδους καθώς και στατιστικά πακέτα πρόβλεψης όπως είναι το *ForecastPro* και το *ForecastX*. Σήμερα, 20 χρόνια από την εμφάνισή της, συνεχίζει να αποτελεί ένα από τα πλέον αξιοσημείωτα benchmarks που καλείται η κάθε νέα μέθοδος που προτείνεται να ξεπεράσει.

Προκειμένου να ερευνηθεί η αποτελεσματικότητα της μεθόδου θ σε δεδομένα και εφαρμογές πέραν του διαγωνισμού *M3*, η ακρίβεια της μεθόδου μελετήθηκε σε πλήθος εργασιών. Οι *Nikolopoulos et al. (2012)* αξιολόγησαν την απόδοση της μεθόδου σε ένα μεγάλο σετ δεδομένων που δεν αφορούν ζήτηση προϊόντων προβλέποντας την εξέλιξη του δείκτη S&P500, ενώ αργότερα αποδείχτηκε ότι η χρήση της μεθόδου μπορεί να οδηγήσει σε σημαντικά οφέλη όταν χρησιμοποιείται για την οργάνωση και τη διανομή προϊόντων σε επιχειρήσεις (*Nikolopoulos et al., 2012*). Οι *Pagourtzi et al. (2007)* αξιολόγησαν τη μέθοδο για την πρόβλεψη τιμών ακινήτων στο Ηνωμένο Βασίλειο, οι *Petropoulos et al. (2007)* έλεγξαν την ακρίβεια της μεθόδου στα δεδομένα του διαγωνισμού T και οι *Petropoulos et al. (2008)* πρότειναν κάποιες ιδέες για τη χρήση της μεθόδου στην περίπτωση χρονοσειρών διακοπτόμενης ζήτησης (*Intermittent Demand forecasting*). Πιο πρόσφατα, οι *Thomakos & Nikolopoulos (2015)* πρότειναν μία μονο-παραγοντική (*univariate*) μορφή της μεθόδου για την πρόβλεψη χρονοσειρών πολλαπλών

μεταβλητών (*multivariate*) και αξιολόγησαν την απόδοσή της σε μακροοικονομικά δεδομένα και χρηματοοικονομικές χρονοσειρές.

Τα αποτελέσματα σε όλες τις παραπάνω μελέτες έδειξαν ότι η μέθοδος μπορεί να οδηγήσει σε υποσχόμενα αποτελέσματα ανεξαρτήτως εφαρμογής και τύπου δεδομένων και συνεπώς απέδειξαν τη χρησιμότητά της για την πρόβλεψη πληθώρας χρονοσειρών σε επιχειρήσεις και οργανισμούς. Για να γίνει βέβαια κάτι τέτοιο αποδοτικά απαιτείται η ενσωμάτωση της μεθόδου σε κατάλληλο στατιστικό πακέτο πρόβλεψης (*Forecasting Support System - FSS*). Έτσι, οι *Nikolopoulos & Assimakopoulos (2003)* ανέπτυξαν ένα σύστημα το οποίο μπορεί με αυτοματοποιημένους κανόνες και χρήση κριτικών αναπροσαρμογών επί των αποτελεσμάτων της μεθόδου θ να υποστηρίξουν αποφάσεις ενώ οι *Tavanidou et al. (2003)* ενσωμάτωσαν τη μέθοδο σε ένα web-based FSS.

Η έρευνα γύρω από το μοντέλο πρόβλεψης προχώρησε περαιτέρω στην κατεύθυνση βελτιστοποίησης της τιμής της παραμέτρου θ ανά χρονοσειρά και της γενίκευσης συνεπώς της χρήσης του σε μορφές πέραν της κλασικής. Έτσι, οι *Constantinidou et al. (2012)* πρότειναν τη χρήση ενός νευρωνικού δικτύου για την επιλογή των βέλτιστων βαρών δύο γραμμών θ , μία προσέγγιση που είχε προταθεί και νωρίτερα μέσω της χρήσης ευριστικών κανόνων (*Bougioukos et al. 2007 και Bougioukos et al. 2008*). Οι *Thomakos & Nikolopoulos (2014)* πρότειναν αργότερα μία προσέγγιση για τον προσδιορισμό της βέλτιστης τιμής της παραμέτρου θ για πρόβλεψη με μία και μόνο γραμμή και επιπλέον εισήγαγαν μία φόρμουλα για τον υπολογισμό των βέλτιστων βαρών των γραμμών θ για χρήση δύο γραμμών. Οι *Fiorucci et al. (2016)* παρουσίασαν μία προσέγγιση για τη βέλτιστη παραμετροποίηση της μεθόδου μέσω μίας ενιαίας και γενικευμένης μορφής στην περίπτωση χρήσης δύο γραμμών θ όπου η μία εξ αυτών ισούται με μηδέν. Τέλος, οι *Petropoulos & Nikolopoulos (2013)* μελέτησαν τη χρήση της μεθόδου με περισσότερες των δύο γραμμών προκειμένου να αυξήσουν την ακρίβειά της με αξιολογικά αποτελέσματα.

Όπως αναδεικνύεται μέσω της βιβλιογραφικής επισκόπησης που πραγματοποιήθηκε, παρόλο που η μέθοδος θ αποτελεί ένα γενικευμένο και ευέλικτο πλαίσιο αποσύνθεσης χρονοσειρών, η έρευνα περιορίζεται μέχρι στιγμής στον βέλτιστο προσδιορισμό των βαρών των δύο γραμμών θ που ορίζονται στην κλασική της προσέγγιση. Αυτό μπορεί να αποδειχτεί πως, αν επιθυμούμε οι δύο γραμμές να συνθέτουν την αρχική χρονοσειρά και η μία εξ αυτών είναι η μηδενική, ισοδυναμεί με τον εντοπισμό της βέλτιστης τιμής της παραμέτρου θ για τη δεύτερη. Έτσι, το πρόβλημα της βελτιστοποίησης και γενίκευσης της μεθόδου περιορίζεται πρακτικά στην επιλογή της γραμμής που καλείται να περιγράψει τα βραχυπρόθεσμα χαρακτηριστικά της χρονοσειράς, ή εναλλακτικά στον καθορισμό του βαθμού στον οποίο η μακροπρόθεσμη τάση συμμετέχει στην πρόβλεψη. Όπως αναφέρουν οι *Hyndman και Billah (2003)*, η μέθοδος θ μπορεί να αντικατασταθεί από το μοντέλο της απλής εκθετικής εξομάλυνσης (*SES*) πραγματοποιώντας μία μετατόπιση στις προβλέψεις αυτού η οποία σχετίζεται με την κλίση της γραμμής μηδέν και την παράμετρο θ . Αν και ο εν λόγω ισχυρισμός δεν είναι απολύτως αληθής (*Nikolopoulos & Assimakopoulos, 2005*), αποδεικνύει πως το μοντέλο είναι ικανό να περιγράψει αποτελεσματικά μόνο γραμμικές τάσεις. Ένα ερώτημα που προκύπτει λοιπόν είναι αν και με ποιον τρόπο η μέθοδος θ θα μπορούσε να τροποποιηθεί προκειμένου να περιγράψει

επιπρόσθετα μη γραμμικά μοτίβα τάσης (π.χ. εκθετικό) και συνεπώς να γενικευτεί περαιτέρω.

3.4.2 Η μέθοδος θ και κάποιες ενδιαφέρουσες διαπιστώσεις

Όπως αναφέρθηκε νωρίτερα, η μέθοδος θ βασίζεται στην αποσύνθεση των χρονοσειρών σε επιμέρους γραμμές με ίδια ροπή πρώτης και δεύτερης τάξης με αυτής των αρχικών δεδομένων αλλά διαφορετική καμπυλότητα. Δεδομένης μίας παραμέτρου $\theta \in R$ το εν λόγω αποτέλεσμα επιτυγχάνεται μέσω του μετασχηματισμού θ , της εφαρμογής δηλαδή της παραμέτρου στις δεύτερες διαφορές της χρονοσειράς για τη δημιουργία της γραμμής Y^θ ως εξής:

$$Y_t^\theta = \theta Y_{00}, \text{ με } Y_{00} = Y_t - 2Y_{t-1} + Y_{t-2}$$

όπου με Y_t συμβολίζεται η παρατήρηση της αρχικής χρονοσειράς τη στιγμή t . Αν η παράμετρος θ είναι μεγαλύτερη της μονάδας, η νέα γραμμή που προκύπτει διαθέτει αυξημένη καμπυλότητα ενώ για τιμές παραμέτρου $0 < \theta < 1$ οι καμπυλότητες συρρικνώνονται. Για $\theta = 1$ η γραμμή που προκύπτει ταυτίζεται με τα αρχικά δεδομένα και για $\theta = 0$ παράγεται η ευθεία της απλής γραμμικής παλινδρόμησης για την οποία και ισχύει:

$$Y^0 = a + bt$$

$$b = \frac{\sum_{i=1}^n (t - \bar{t})(Y_t - \bar{Y})}{\sum_{i=1}^n (t - \bar{t})^2}$$

$$a = \bar{Y} - b\bar{t}$$

Σημειώνεται επίσης πως για αρνητική τιμή της παραμέτρου πραγματοποιείται αντιστροφή των αρχικών δεδομένων με άξονα συμμετρίας τη γραμμή μηδέν. Η χρησιμότητα βέβαια τέτοιων γραμμών είναι περιορισμένη και συνεπώς δεν μελετώνται στα πλαίσια της παρούσας διατριβής.

Ο υπολογισμός των γραμμών θ απλοποιείται αρκετά χρησιμοποιώντας την εναλλακτική έκφραση της μεθόδου που πρότεινε ο Νικολόπουλος και οι συνεργάτες του το 2008 κατά την οποία κάθε γραμμή αποτελεί γραμμικό άθροισμα των αρχικών δεδομένων και της ευθείας της απλής γραμμικής παλινδρόμησης.

$$Y_t^\theta = \theta Y_t + (1 - \theta)(a + bt)$$

Βάσει της παραπάνω εξίσωσης γίνεται κατανοητό πως παρατηρήσεις οι οποίες απέχουν σημαντικά από τη γραμμή Y^0 θα λάβουν σημαντική μεταβολή στην τιμή τους, και το αντίστροφο. Έτσι, για 'λείες' χρονοσειρές ο μετασχηματισμός αναμένεται να έχει αμελητέα επίδραση, ενώ για χρονοσειρές με υψηλή διακύμανση να προκαλέσει σημαντικές τροποποιήσεις.

Μετά τον υπολογισμό των γραμμών θ που θα χρησιμοποιηθούν για την πρόβλεψη της αρχικής χρονοσειράς, σειρά έχει η προέκτασή τους με χρήση κατάλληλων μεθόδων. Μία, δύο ή και περισσότερες γραμμές θ μπορούν να χρησιμοποιηθούν με ποικίλες τιμές

παραμέτρου ανάλογα με το ποια χαρακτηριστικά της αρχικής χρονοσειράς επιθυμούμε να αναδείξουμε. Οι γραμμές με τιμή θ κοντά στο μηδέν αναδεικνύουν τα μακροπρόθεσμα χαρακτηριστικά της χρονοσειράς (π.χ. τη τάση) και συνεπώς ένα κατάλληλο μοντέλο πρόβλεψης θα ήταν η Holt. Αντίθετα, οι γραμμές με τιμή θ μεγαλύτερες της μονάδας αναδεικνύουν τα βραχυπρόθεσμα χαρακτηριστικά της χρονοσειράς (π.χ. επίπεδο) και συνεπώς ένα κατάλληλο μοντέλο πρόβλεψης θα ήταν η SES. Στις υπόλοιπες περιπτώσεις θα μπορούσε να χρησιμοποιηθεί ένα μοντέλο σαν την Damped η οποία είναι σε θέση να λαμβάνει από μόνη της υπόψη και τα δύο χαρακτηριστικά με αποδοτικό τρόπο.

Προκειμένου να παραχθούν προβλέψεις για τα αρχικά δεδομένα, οι επιμέρους προβλέψεις των γραμμών πρέπει να συνδυαστούν με κατάλληλα βάρη. Σημειώνεται ότι προκειμένου ο συνδυασμός των γραμμών θ να οδηγεί στην ανακατασκευή της αρχικής χρονοσειράς, το οποίο φυσικά είναι και το επιθυμητό, θα πρέπει να ισχύουν συγκεκριμένοι περιορισμοί. Για την περίπτωση π.χ. δύο γραμμών θ , οι *Fiorucci et al. (2016)* αποδεικνύουν ότι το βάρος της εκάστοτε γραμμής δίνεται από το ακόλουθο σετ εξισώσεων:

$$w_{\theta_1} = \frac{\theta_2 - 1}{\theta_2 - \theta_1}$$

$$w_{\theta_2} = 1 - w_{\theta_1}$$

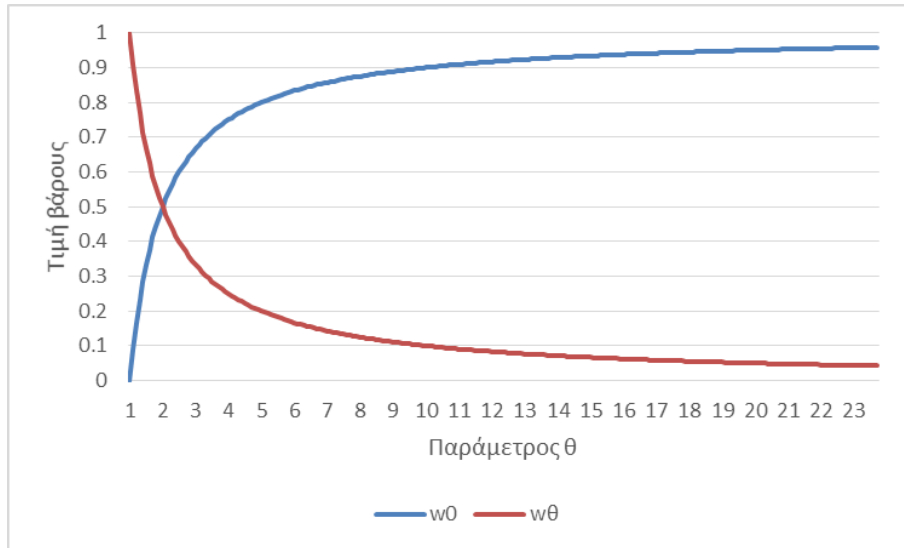
υπό τον περιορισμό $\theta_1 \leq 1$ και $\theta_2 \geq 1$

όπου θ_i and w_{θ_i} είναι η παράμετρος και το βάρος της γραμμής i που επιλέχθηκε. Στην περίπτωση μάλιστα που η μία εκ των γραμμών είναι η μηδενική, η παραπάνω σχέση απλοποιείται ως εξής:

$$w_0 = \frac{\theta - 1}{\theta}$$

$$w_\theta = \frac{1}{\theta}$$

υπό τον περιορισμό $\theta \geq 1$



Εικόνα 3.1: Μεταβολή της τιμής των βαρών συνδυασμού δύο γραμμών θ όταν η μία εξ αυτών είναι η μηδενική.

Όπως φαίνεται στην *Εικόνα 3.1* με την αύξηση της τιμής της παραμέτρου θ αυξάνεται σημαντικά η συμμετοχή της μη μηδενικής γραμμής στην πρόβλεψη, ενώ για $\theta=2$ (κλασική μέθοδος θ) τα βάρη των δύο γραμμών ταυτίζονται. Προκύπτει δηλαδή ότι η κατάλληλη αυξομείωση των βαρών δύο γραμμών Y^0 και Y^θ για την παρακολούθηση των μακροπρόθεσμων και βραχυπρόθεσμων χαρακτηριστικών μίας χρονοσειράς ισοδυναμεί στην πραγματικότητα με την επιλογή μίας εναλλακτικής γραμμής θ .

Σημειώνεται ότι ο συνδυασμός των γραμμών Y^0 και Y^θ είναι εκείνος που χρησιμοποιείται ευρύτερα στην πράξη, κυρίως λόγω της απλότητάς του και των εμπειρικών αποτελεσμάτων τα οποία αναφέρουν ότι προκειμένου οι προβλέψεις που παράγονται να είναι στιβαρές θα πρέπει η μία εκ των γραμμών θ που χρησιμοποιείται να είναι η μηδενική. Φυσικά, όπως αναφέρθηκε νωρίτερα, η μέθοδος θ παρέχει ένα γενικευμένο πλαίσιο αποσύνθεσης χρονοσειρών σε επιμέρους γραμμές και μπορεί να χρησιμοποιείται ελεύθερα για την παραγωγή προβλέψεων με χρήση περισσότερων των δύο γραμμών.

Σύμφωνα με την έκφραση των βαρών των γραμμών θ που παρουσιάστηκε προκύπτει λοιπόν ότι στην περίπτωση συνδυασμού δύο γραμμών η μέθοδος μπορεί να εκφραστεί από την ακόλουθη απλουστευμένη εξίσωση:

$$Y_t = \frac{\theta-1}{\theta} Y_t^0 + \frac{1}{\theta} Y_t^\theta = \frac{\theta-1}{\theta} (a + bt) + \frac{1}{\theta} Y_t^\theta, \theta \geq 1$$

Η κλασική μέθοδος θ , όπως αυτή έλαβε μέρος και στο διαγωνισμό *M3*, ακολουθεί μεθοδολογικά τα πέντε παρακάτω βήματα:

- *Αποεποχικοποίηση χρονοσειράς*: Τα αρχικά δεδομένα αποεποχικοποιούνται με χρήση της κλασικής πολλαπλασιαστικής μεθόδου αποσύνθεσης. Προκειμένου να

γίνει ωστόσο αυτό εκτελείται πρώτα ένας έλεγχος εποχιακότητας με διάστημα εμπιστοσύνης 90%. Σε περίπτωση που η χρονοσειρά δεν κριθεί εποχιακή το εν λόγω στάδιο παραλείπεται.

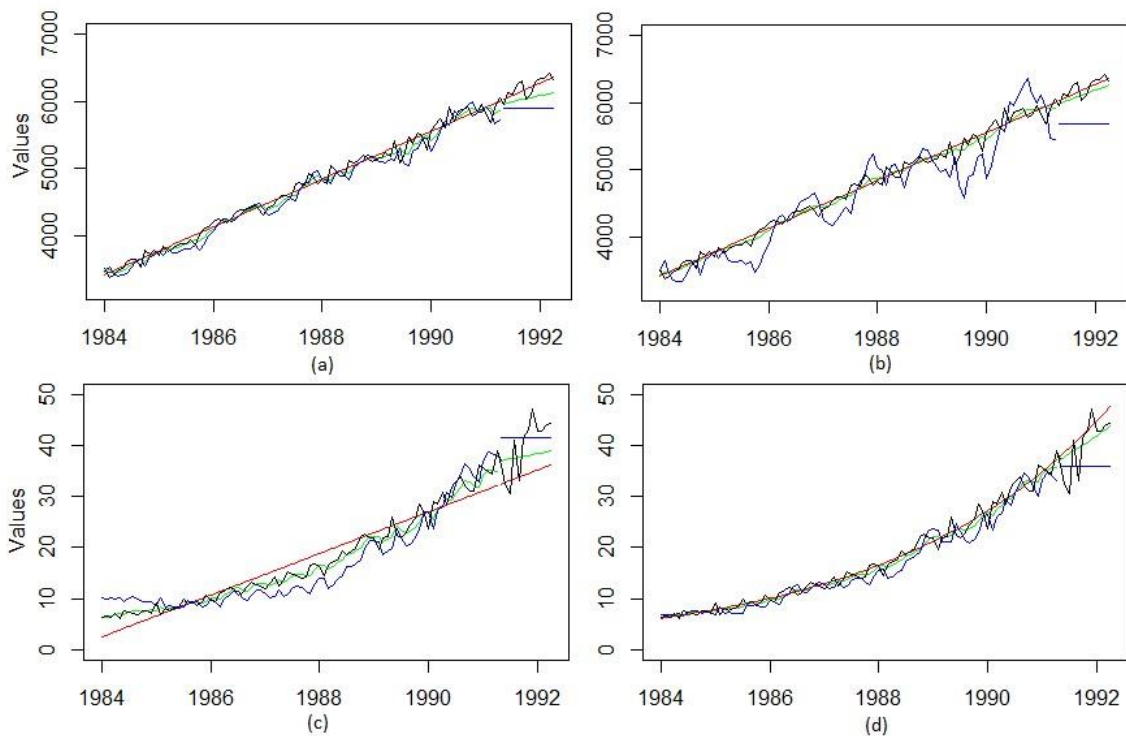
- *Εκτίμηση γραμμών θ* : Αξιοποιώντας τα αποεποχικοποιημένα δεδομένα υπολογίζονται οι γραμμές Y^0 και Y^2 .
- *Πρόβλεψη γραμμών θ* : Οι δύο γραμμές θ προβλέπονται με χρήση κατάλληλων μεθόδων. Συγκεκριμένα η γραμμή μηδέν επεκτείνεται με επέκταση της ευθείας της απλής γραμμικής παλινδρόμησης, ενώ η γραμμή δύο με χρήση της απλής εκθετικής εξομάλυνσης.
- *Συνδυασμός προβλέψεων*: Οι προβλέψεις των δύο γραμμών συνδυάζονται χρησιμοποιώντας ίσα βάρη.
- *Εποχικοποίηση προβλέψεων*: Αξιοποιώντας κατάλληλα τους δείκτες εποχιακότητας που υπολογίστηκαν στο πρώτο βήμα της μεθοδολογίας οι προβλέψεις εποχικοποιούνται. Σε περίπτωση που η χρονοσειρά δεν είχε κριθεί εποχιακή το εν λόγω στάδιο παραλείπεται.

Όπως σημειώθηκε και κατά την παρουσίαση της μεθόδου, παρόλο που η κλασική έκφρασή της έχει αποδειχτεί ιδιαίτερα ακριβής και ικανή να παράγει στιβαρές προβλέψεις, διαθέτει δύο σημαντικά μειονεκτήματα: (1) τη χρήση της γραμμής Y^2 ανεξαρτήτως της χρονοσειράς που καλούμαστε να προβλέψουμε και (2) τη θεώρηση ότι η μακροχρόνια τάση όλων των χρονοσειρών είναι γραμμικής μορφής.

Σε ότι αφορά το πρώτο μειονέκτημα της μεθόδου η βιβλιογραφία αποδεικνύει πως έχουν πραγματοποιηθεί σημαντικές προσπάθειες προκειμένου να επιλέγεται η βέλτιστη γραμμή θ και μάλιστα μέσω αρκετών διαφορετικών προσεγγίσεων. Συνεπώς το εν λόγω μειονέκτημα δεν κρίνεται σκόπιμο να ερευνηθεί περαιτέρω καθώς δεν εμπεριέχει κάποια καινοτομία. Το δεύτερο μειονέκτημα εμφανίζει ωστόσο ιδιαίτερο ενδιαφέρον καθώς αφενός δεν έχει αποτελέσει προηγουμένως αντικείμενο μελέτης και αφετέρου καθορίζει σημαντικά την απόδοση της μεθόδου σε μέσο- και μακροπρόθεσμες προβλέψεις.

Οι *Hyndman και Billah (2003)* προσπάθησαν να αποδείξουν ότι η κλασική μέθοδος θ μπορεί να αντικατασταθεί από το μοντέλο της απλής εκθετικής εξομάλυνσης πραγματοποιώντας μία μετατόπιση στις προβλέψεις αυτού ίση με το μισό της κλίση της γραμμής μηδέν. Η εν λόγω προσέγγιση ονομάστηκε '*SES with drift*', ή αλλιώς *ses-d*, και μάλιστα στην ίδια εργασία πραγματοποιήθηκε μία προσπάθεια μεταβολής του βαθμού κατά τον οποίο μετατοπίζεται η πρόβλεψη της *SES* ούτως ώστε να επιτευχθεί καλύτερη ακρίβεια. Τα αποτελέσματα ήταν αρκετά κοντά σε αυτά της κλασικής μεθόδου θ και σημαντική διαφοροποίηση μεταξύ των δύο προσεγγίσεων αποτέλεσε ο υπολογισμός του επιπέδου της ευθείας της απλής γραμμικής παλινδρόμησης επί της γραμμής Y^2 και όχι επί των αρχικών δεδομένων. Ανεξαρτήτως πάντως των αποτελεσμάτων, εκείνο που αδιαμφισβήτητα αποδεικνύει η εν λόγω μελέτη είναι η γραμμική φύση των προβλέψεων της μεθόδου θ η οποία και γεννά την απορία κατά πόσο αν δεν ίσχυε, αν δηλαδή υπήρχε η δυνατότητα προέκτασης της χρονοσειράς και μη γραμμικό τρόπο, η μέθοδος θα μπορούσε να αποδώσει ακόμα καλύτερα.

Προκειμένου ο παραπάνω ισχυρισμός να γίνει καλύτερα αντιληπτός, στην *Εικόνα 3.2* παρουσιάζονται δύο χαρακτηριστικά παραδείγματα χρονοσειρών με διαφορετικά μοτίβα τάσης, ένα γραμμικό και ένα εκθετικό. Χρησιμοποιώντας την κλασική μέθοδο θ , η χρονοσειρά γραμμικής τάσης προβλέπεται ικανοποιητικά (α), ενώ η ακρίβεια της μεθόδου μπορεί να βελτιωθεί ακόμα καλύτερα για χρήση μεγαλύτερης τιμής παραμέτρου θ , δηλαδή αύξηση της συμμετοχής της γραμμής Y^0 στην πρόβλεψη (β). Ωστόσο, το ίδιο δεν ισχύει για την περίπτωση της χρονοσειράς εκθετικής τάσης όπου και η τελική πρόβλεψη απέχει σημαντικά από την πραγματικότητα, ιδιαίτερα στους μεγαλύτερους ορίζοντες πρόβλεψης (γ). Αν κανείς αντικαθιστούσε λοιπόν στην εξίσωση του μοντέλου της ευθείας της απλής γραμμικής παλινδρόμησης με αυτήν μίας εκθετικής καμπύλης (δ) θα μπορούσε να οδηγηθεί σε σαφώς πιο ακριβείς προβλέψεις.



Εικόνα 3.2: Απόδοση της μεθόδου θ για μία χρονοσειρά γραμμικής και μη γραμμικής τάσης.

Τα διαγράμματα (α) και (γ) παρουσιάζουν την πρόβλεψη της κλασικής μεθόδου θ για χρονοσειρές γραμμικής και εκθετικής τάσης, το διάγραμμα (β) την απόδοση της μεθόδου για αντικατάσταση της γραμμής Y^2 με αυτήν της Y^7 , και τέλος το διάγραμμα (δ) την απόδοση της μεθόδου για αντικατάσταση της γραμμής μηδέν με μία εκθετική καμπύλη. Με κόκκινο χρώμα παρουσιάζεται η γραμμή Y^0 , με μπλε η Y^2 και με πράσινο ο συνδυασμός τους.

3.4.3 Επέκταση της μεθόδου θ για χρονοσειρές μη γραμμικής τάσης

Όπως παρουσιάστηκε στην προηγούμενη παράγραφο, η μέθοδος θ είναι ικανή να περιγράψει ικανοποιητικά χρονοσειρές με γραμμικά μοτίβα τάσης και μάλιστα να μεταβάλει σημαντικά την κλίση τους μέσω της παραμέτρου θ . Έτσι, υψηλές τιμές της

παραμέτρου μπορούν να επιλεγούν για χρονοσειρές με εμφανή γραμμική τάση και χαμηλές τιμές για στάσιμες χρονοσειρές ή χρονοσειρές με έντονο το χαρακτηριστικό του θορύβου. Στην περίπτωση πάντως που η χρονοσειρά εμφανίζει μη γραμμικό μοτίβο τάσης, τα αποτελέσματα της μεθόδου αναμένεται να μην είναι και τα πλέον ακριβή, ειδικά στην περίπτωση παραγωγής μεσο- και μακροπρόθεσμων προβλέψεων.

Σύμφωνα με τον απλουστευμένο της τύπο, η μέθοδος θ με χρήση δύο γραμμών Y^0 και Y^2 μπορεί να εκφραστεί ανεξαρτήτως του τύπου της γραμμής μηδέν ως εξής:

$$Y_t = \frac{\theta-1}{\theta} Y_t^0 + \frac{1}{\theta} Y_t^\theta, \theta \geq 1$$

Έτσι, στην περίπτωση πρόβλεψης της γραμμής θ με το μοντέλο της απλής εκθετικής εξομάλυνσης, το οποίο και δεν λαμβάνει υπόψη του την ύπαρξης τάσης στα δεδομένα, γίνεται εμφανές πως οποιαδήποτε μορφή τάσης λαμβάνεται υπόψη αποκλειστικά μέσω της επέκτασης της γραμμής μηδέν. Αντικαθιστώντας συνεπώς κανείς το μοντέλο της απλής γραμμικής παλινδρόμησης με κάποια άλλη μορφή καμπύλης και επεκτείνοντάς την κατάλληλα στο μέλλον, μπορεί αυτομάτως να μετατοπίσει το τρέχον επίπεδο της χρονοσειράς (προσδιοριζόμενο μέσω της SES στη γραμμή θ) με μη γραμμικό τρόπο.

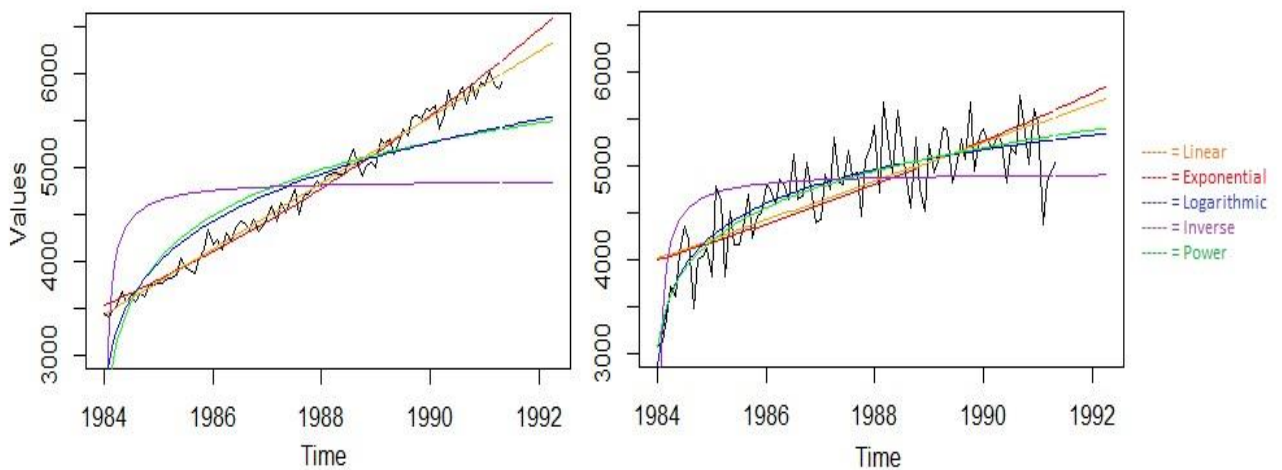
Στα πλαίσια της παρούσας διατριβής, και για λόγους απλότητας, εξετάζονται τέσσερις μη γραμμικοί τύποι τάσης: η εκθετική (*exponential*), η λογαριθμική (*logarithmic*), η αντίστροφη (*inverse*) και η δύναμη (*power*). Φυσικά οποιαδήποτε άλλη μορφή καμπύλης, όπως π.χ. ένα πολυώνυμο n βαθμού (σημειώνεται ότι και η ευθεία απλής γραμμικής παλινδρόμησης είναι πολυώνυμο πρώτου βαθμού), μπορεί να χρησιμοποιηθεί ελεύθερα ως Y^0 . Οι συγκεκριμένες επιλογές έγιναν δεδομένου ότι οι τέσσερις αυτές μορφές καμπύλης μπορούν να περιγράψουν ικανοποιητικά οποιαδήποτε μορφή μακροχρόνιας τάσης ενώ προκειμένου να εκτιμηθούν απαιτείται η εκτίμηση δύο και μόνο παραμέτρων ανά περίπτωση, όσων δηλαδή και στην περίπτωση της ευθείας. Έτσι, η μέθοδος θ γενικεύεται και γίνεται σαφώς πιο ευέλικτη χωρίς να αυξάνεται η συνολική της πολυπλοκότητα και η αβεβαιότητα κατά την παραμετροποίησή της.

Οι εξισώσεις για την εκτίμηση της γραμμής Y_t^0 για κάθε μία από τις προαναφερθείσες καμπύλες δίνονται παρακάτω.

- Ευθεία απλής γραμμικής παλινδρόμησης (*linear regression*): $Y_t^0 = b + at$
- Εκθετική καμπύλη (*exponential curve*): $Y_t^0 = be^{at}$, ή εναλλακτικά $\log(Y_t^0) = \log(b) + at$
- Λογαριθμική καμπύλη (*logarithmic curve*): $Y_t^0 = b + a \log(t)$
- Αντίστροφη καμπύλη (*inverse curve*): $Y_t^0 = b + a/t$
- Καμπύλη δύναμης (*power curve*): $Y_t^0 = bt^a$, ή εναλλακτικά $\log(Y_t^0) = \log(b) + a \log(t)$

όπου, όπως και για την περίπτωση της απλής γραμμικής παλινδρόμησης, οι παράμετροι a και b μπορούν να εκτιμηθούν εύκολα με χρήση της μεθόδου ελαχίστων τετραγώνων. Αυτός είναι και ο λόγος για τον οποίο μέσω λογαρίθμων όλες οι καμπύλες δίνονται σε γραμμική μορφή.

Στην *Εικόνα 3.3* οι προτεινόμενες καμπύλες χρησιμοποιούνται για την επέκταση της μακροχρόνιας τάσης δύο χρονοσειρών με διαφορετικό μοτίβο: γραμμικό και λογαριθμικό. Όπως φαίνεται, η νέα προσέγγιση μπορεί να οδηγήσει στη δεύτερη περίπτωση σε σημαντικές βελτιώσεις καθώς η κλασική γραμμική μορφή της μεθόδου οδηγεί σε μη ρεαλιστικές προβλέψεις με αισιόδοξη διάθεση. Αντίθετα, στην πρώτη περίπτωση η κλασική μέθοδος θ φαίνεται να είναι η βέλτιστη επιλογή.



Εικόνα 3.3: Χρήση εναλλακτικών καμπύλων ως Y_t^0 στην κλασική μέθοδο θ .

Εδώ θα ήταν καλό να σημειωθεί ότι παρόλο που το εν λόγω φαινόμενο μπορεί να περιοριστεί με χρήση κατάλληλης γραμμής θ , π.χ. με χρήση μικρής τιμής παραμέτρου για πρόβλεψη χρονοσειρών με λογαριθμική τάση, η νέα προσέγγιση οδηγεί σε σαφώς μεγαλύτερη ευελιξία και αυξάνει τις επιλογές μας στη διαδικασία της πρόβλεψης. Πιο συγκεκριμένα αναδεικνύει το γενικότερο πλαίσιο αποσύνθεσης της μεθόδου σε επιμέρους γραμμές θ και την ικανότητά της να περιγράψει ικανοποιητικά οποιαδήποτε χρονοσειρά ανεξαρτήτως χαρακτηριστικών. Άλλωστε, το πρόβλημα της βελτιστοποίησης της παραμέτρου θ έχει αποτελέσει αντικείμενο μελέτης σε προηγούμενες εργασίες και συνεχίζει να αποτελεί ένα ισχυρό εργαλείο στην κατεύθυνση παραγωγής προβλέψεων υψηλής ακρίβειας. Σημαντικό βέβαια βήμα προκειμένου να εκμεταλλευθεί κανείς πλήρως τα πλεονεκτήματα της προτεινόμενης προσέγγισης είναι η χρήση μίας μεθοδολογίας ικανής να αναγνωρίσει το μοτίβο τάσης για την επιλογή και χρήση της καταλληλότερης γραμμής μηδέν ανά περίπτωση.

Προκειμένου να αναδειχτεί η σημασία της αναθεώρησης του μακροπρόθεσμου μοτίβου τάσης των χρονοσειρών στη βελτίωση της προβλεπτικής ακρίβειας της μεθόδου θ , στον *Πίνακα 3.4* παρουσιάζεται η επίδοση των πέντε εναλλακτικών προσεγγίσεων μοτίβων

τάσης που προτάθηκαν νωρίτερα στα δεδομένα των διαγωνισμών *M1* και *M3* (4004 χρονοσειρές). Η επίδοση των μεθόδων αξιολογείται μέσω του δείκτη σφάλματος *sMAPE*, ενώ τα δεδομένα έχουν αποεποχικοποιηθεί όπου αυτό κρίνεται απαραίτητο μέσω της κλασικής μεθόδου αποσύνθεσης.

Όπως παρατηρείται, πέραν των ετήσιων χρονοσειρών του διαγωνισμού *M1* και των άλλων δεδομένων του διαγωνισμού *M3*, το εκθετικό μοτίβο τάσης φαίνεται να κυριαρχεί έναντι των υπολοίπων οδηγώντας σε συρρίκνωση του σφάλματος της κλασικής προσέγγισης της τάξεως του 2.66%. Ενθαρρυντικό είναι επίσης το γεγονός ότι στην περίπτωση των μηνιαίων δεδομένων του *M1*, πέραν του εκθετικού μοτίβου, και όλες οι υπόλοιπες προσεγγίσεις φαίνεται να υπερτερούν έναντι της κλασικής. Αυτό επιβεβαιώνει τον αρχικό ισχυρισμό πως η κάθε χρονοσειρά ενδέχεται να περιγράφεται αποδοτικά από διαφορετικό μοτίβο τάσης, το οποίο και πρέπει να αξιοποιείται κατά τη διαδικασία της πρόβλεψης.

*Πίνακας 3.4: Προβλεπτική ακρίβεια μεθόδου θ (sMAPE) για κλασική αποεποχικοποίηση των δεδομένων των διαγωνισμών *M1* και *M3* και χρήση εναλλακτικών μοτίβων τάσης.*

Δεδομένα	Γραμμικό	Εκθετικό	Λογαριθμικό	Αντίστροφο	Δύναμη
M3 Ετήσια	16.73	16.44	17.20	17.73	16.86
M3 Τριμηνιαία	9.29	9.28	9.60	9.81	9.56
M3 Μηνιαία	13.87	13.68	14.00	14.25	14.05
M3 Άλλα	4.92	5.31	5.87	6.27	5.96
M1 Ετήσια	20.19	20.40	22.47	23.23	21.73
M1 Τριμηνιαία	16.30	16.04	17.00	17.53	16.90
M1 Μηνιαία	16.52	14.73	15.14	14.95	14.76
Σύνολο	13.89	13.52	14.04	14.31	13.90

Ο παραπάνω ισχυρισμός περί επιλεκτικότητας και καταλληλότητας μοτίβων τάσης ανά χρονοσειρά ενισχύεται ελέγχοντας επιπλέον το ποσοστό στο οποίο στο σύνολο των 4004 χρονοσειρών κάθε μία από τις επιμέρους προσεγγίσεις αποτελεί τη βέλτιστη λύση. Τα αποτελέσματα δείχνουν πως η κλασική μέθοδος πρωτεύει μόλις στο 18.61% των χρονοσειρών, ενώ η εκθετική στο 42.96%. Τα ποσοστά στις περιπτώσεις του λογαρίθμου, του αντιστρόφου και της δύναμης φτάνουν τα 4.75%, 28.37% και 5.32% αντίστοιχα. Συγκρίνοντας τα εν λόγω ποσοστά με τα αντίστοιχα σφάλματα πρόβλεψης συμπεραίνει κανείς εύκολα λοιπόν αφενός πως όλα τα μοντέλα μπορούν να αξιοποιηθούν για τη μοντελοποίηση συγκεκριμένων μοτίβων τάσης, και αφετέρου πως κάποια από αυτά οδηγούν σε αρκετά απόλυτες λύσεις κάνοντάς τα ιδιαίτερα ακριβή σε χρονοσειρές αντίστοιχου μοτίβου και σημαντικά ανακριβή στην αντίθετη περίπτωση. Αυτός είναι και ο λόγος που ενώ π.χ. το αντίστροφο μοντέλο υπερτερεί σε πρωτιές έναντι του γραμμικού, εν τέλει το δεύτερο λόγω της καθολικής του χρήσης πετυχαίνει καλύτερη μέση ακρίβεια.

3.4.4 Αθροιστικό και πολλαπλασιαστικό μοντέλο θ

Μία επιπλέον παρατήρηση που μπορεί να γίνει για τις ιδιότητες του κλασικού μοντέλου θ πέραν της γραμμικότητάς του, είναι πως επί της ουσίας αυτό αποσυνθέτει τη χρονοσειρά στις συνιστώσες επιπέδου και τάσης, τις προεκτείνει ξεχωριστά και τις συνδυάζει προσθετικά με ανεξάρτητο τρόπο. Ως εκ τούτου προκύπτει αμέσως το συμπέρασμα πως η κλασική μέθοδος θ δεν μπορεί να περιγράψει ικανοποιητικά χρονοσειρές των οποίων το επίπεδο μεταβάλλεται συγκριτικά με την τάση της χρονοσειράς. Τέτοια ιδιότητα συναντά για παράδειγμα κανείς στην κλασική πολλαπλασιαστική μέθοδο αποσύνθεσης και στα πολλαπλασιαστικά μοντέλα εκθετικής εξομάλυνσης όπου η συνιστώσα π.χ. της εποχιακότητας αυξομειώνεται ανάλογα με το επίπεδο της χρονοσειράς.

Ο αθροιστικός χαρακτήρας του μοντέλου είναι μία σημαντική ιδιότητα της κλασικής μεθόδου θ καθώς περιορίζει επί της ουσίας την ευελιξία της, αφήνοντας περιθώριο σε άλλα μοντέλα, όπως π.χ. αυτά της εκθετικής εξομάλυνσης, να υπερτερούν για συγκεκριμένους τύπους χρονοσειρών. Εκτός αυτού, ένα πολλαπλασιαστικό μοντέλο έχει επιπλέον την ιδιότητα να διατηρεί τις προβλέψεις του θετικές ανεξαρτήτως δεδομένων εισόδου, τα οποία οφείλουν βέβαια σε αυτήν την περίπτωση να είναι και αυτά με τη σειρά τους θετικά. Έτσι, το εν λόγω μοντέλο θα μπορούσε να αξιοποιηθεί χωρίς επιπλέον ελέγχους για την άμεση πρόβλεψη μεγεθών για τα οποία δεν ορίζονται αρνητικές προβλέψεις, όπως π.χ. η τιμή μιας μετοχής ή οι πωλήσεις ενός προϊόντος.

Προκειμένου λοιπόν η μέθοδος θ να γίνει περισσότερο ευέλικτη και να μπορεί να προβλέπει αποτελεσματικά και χρονοσειρές για τις οποίες οι συνιστώσες τάσης κι επιπέδου δεν είναι ανεξάρτητες μεταξύ τους αλλά αλληλοεπιδρούν, προτείνεται η έκφραση της μεθόδου πέραν του προσθετικού, με πολλαπλασιαστικό τρόπο. Για να γίνει αυτό, οι σχέσεις που παρουσιάστηκαν νωρίτερα για τον υπολογισμό των γραμμών θ τροποποιούνται ως εξής:

$$\text{Προσθετική γραμμή } \theta: Y_t^\theta = \theta Y_t + (1 - \theta) Y_t^0$$

$$\text{Πολλαπλασιαστική γραμμή } \theta: Y_t^\theta = \frac{Y_t^\theta}{Y_t^{0\theta-1}}$$

Η πάνω σχέση αναφέρεται στο κλασικό προσθετικό μοντέλο θ και η κάτω σχέση στο πολλαπλασιαστικό. Όπως φαίνεται, και στις δύο περιπτώσεις μπορεί να χρησιμοποιηθεί κανονικά οποιαδήποτε γραμμή μηδέν για την εκτίμηση της μακροπρόθεσμης τάσης της μεθόδου και τον υπολογισμό αντίστοιχα της δεύτερης γραμμής που είναι υπεύθυνη για τον προσδιορισμό του επιπέδου. Στη δεύτερη περίπτωση ωστόσο, αναμένεται να είναι μεγαλύτερη η επίδραση της γραμμής μηδέν στον υπολογισμό της γραμμής θ , επηρεάζοντας με μη γραμμικό τρόπο τις τιμές της, ανάλογα πάντα με την απόσταση της γραμμής μηδέν από τα αρχικά δεδομένα. Σημειώνεται βέβαια πως, όπως και στην περίπτωση της κλασικής μεθόδου αποσύνθεσης, ενδέχεται για αρκετές χρονοσειρές και οι δύο προσεγγίσεις να οδηγούν σε παραπλήσια αποτελέσματα. Σημαντικές διαφορές αναμένεται να εμφανίζονται κυρίως σε χρονοσειρές με έντονες διακυμάνσεις.

Αξιοποιώντας λοιπόν κανείς τις αντίστοιχες γραμμές θ , μπορεί να προεκτείνει μια χρονοσειρά με τη μέθοδο θ ως εξής:

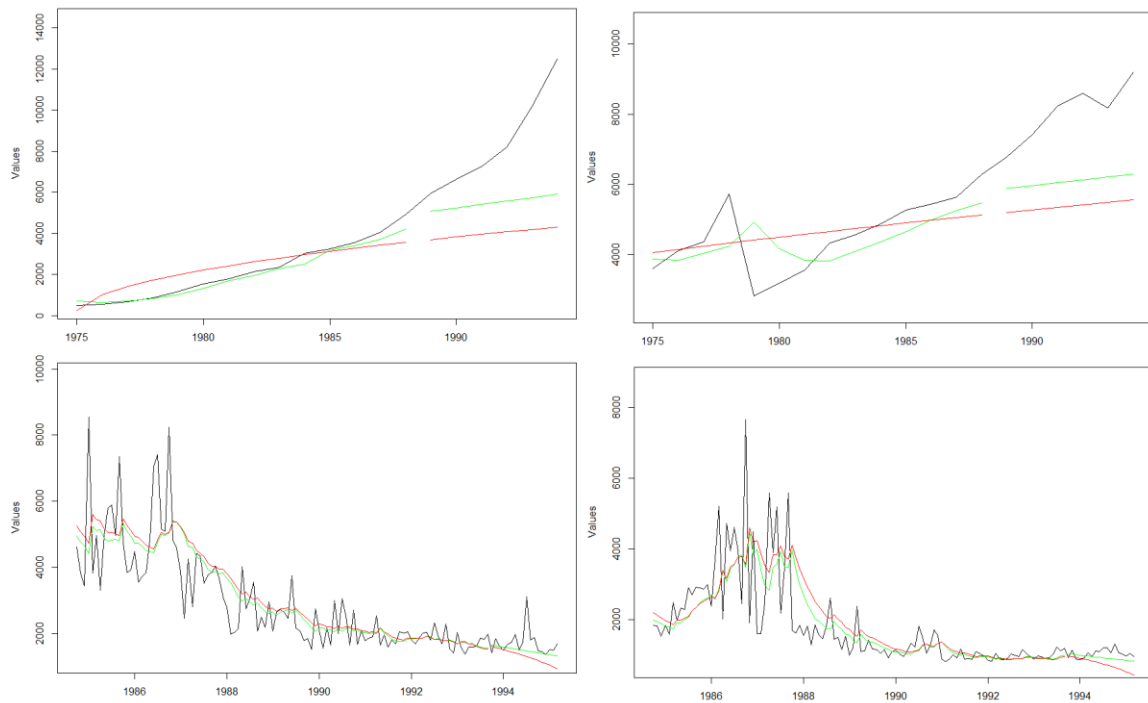
$$\text{Προσθετικό μοντέλο: } Y_t = \frac{\theta-1}{\theta} Y_t^0 + \frac{1}{\theta} Y_t^\theta$$

$$\text{Πολλαπλασιαστικό μοντέλο: } Y_t = (Y_t^\theta Y_t^0)^{\frac{1}{\theta}}$$

Όπως και στην κλασική μέθοδο θ , η οποία εδώ εκφράζεται από την πρώτη σχέση, η γραμμή μηδέν θα προεκταθεί και στις δύο περιπτώσεις ντετερμινιστικά βάσει της εξίσωσής της, ενώ η γραμμή θ μέσω SES. Η ειδοποιός διαφορά μεταξύ των δύο μοντέλων εντοπίζεται λοιπόν αφενός στην εκτίμηση διαφορετικού τρέχοντος επιπέδου λόγω του υπολογισμού των γραμμών θ με διαφορετικό τρόπο, και αφετέρου στο συνδυασμό των προβλέψεων επιπέδου και τάσης.

Σε αυτό το σημείο σημειώνεται πως ένας περιορισμός που ενδέχεται να προκύπτει για το πολλαπλασιαστικό μοντέλο είναι πως για μεγάλες τιμές της παραμέτρου θ η εκτίμηση των αντίστοιχων γραμμών μπορεί να καταστεί υπολογιστικά αδύνατη, ειδικά σε περιπτώσεις όπου οι τιμές της χρονοσειράς είναι μεγάλες. Σε αυτές τις περιπτώσεις μπορεί να χρησιμοποιήσει κανείς κανονικοποιημένα δεδομένα, ή να επιλέξει σχετικά μικρές τιμές θ .

Στην [Εικόνα 3.4](#) παρουσιάζονται ενδεικτικά κάποια παραδείγματα εφαρμογής των δύο μοντέλων σε τυχαίες ετήσιες (πάνω) και μηνιαίες (κάτω) χρονοσειρές. Όπως φαίνεται, ανάλογα με τη διακύμανση των δεδομένων και του τρόπου με τον οποίο αλληλεπιδρά το επίπεδο της χρονοσειράς με την τάση, τα δύο μοντέλα παράγουν ανά περιπτώσεις αρκετά διαφοροποιημένες προβλέψεις.



Εικόνα 3.4: Χρήση γραμμικού προσθετικού (πράσινο) και πολλαπλασιαστικού (κόκκινο) μοντέλου θ για την προέκταση δύο ετήσιων (πάνω) και μηνιαίων (κάτω) χρονοσειρών.

Σε αυτό το σημείο αξίζει να σημειωθεί πως, δεδομένου ότι η συνιστώσα της εποχιακότητας – αν βέβαια υπάρχει – δύναται να αφαιρεθεί είτε με προσθετικό, είτε με πολλαπλασιαστικό τρόπο, η μέθοδος θ μπορεί να μετατραπεί ουσιαστικά έπειτα από την προηγούμενη επέκτασή της σε μία ολοκληρωμένη οικογένεια μοντέλων πρόβλεψης, καθένα εκ των οποίων έχει διαφορετικές ιδιότητες. Συνολικά προκύπτουν δηλαδή 6 διαφορετικά μοντέλα πρόβλεψης ανάλογα με το είδος της εποχιακότητας που υποθέτουμε πως διαθέτει η χρονοσειρά (απουσία, προσθετική ή πολλαπλασιαστική) και του τρόπου με τον οποίο αλληλοεπιδρούν οι συνιστώσες τάσης και επιπέδου (προσθετικά ή πολλαπλασιαστικά). Μάλιστα, υποθέτοντας επιπλέον 5 διαφορετικά μοτίβα τάσης, το σύνολο των μοντέλων αυξάνεται στα 30, ενισχύοντας σημαντικά τις επιλογές μας και την ικανότητά μας να αποσυνθέσουμε και να προεκτείνουμε αποτελεσματικά οποιαδήποτε τύπο χρονοσειράς.

Η επέκταση της μεθόδου θ σύμφωνα με τα παραπάνω σε 30 (ή και περισσότερα) μοντέλα είναι ιδιαίτερα σημαντική καθώς με αυτό τον τρόπο άλλες γενικότερες οικογένειες μοντέλων πρόβλεψης, όπως είναι αυτή της εκθετικής εξομάλυνσης, παύουν να εμφανίζουν οποιοδήποτε θεωρητικό πλεονέκτημα απέναντι στη θ , η οποία διαθέτει πλέον ισοδύναμα μοντέλα αντίστοιχων ιδιοτήτων. Για να γίνει αυτό πιο κατανοητό στη συνέχεια συγκρίνεται επιγραμματικά η γενικευμένη μέθοδος θ με αυτή των μοντέλων εκθετικής εξομάλυνσης, επισημαίνοντας ταυτόχρονα ομοιότητες και διαφορές μεταξύ των δύο οικογενειών μοντέλων πρόβλεψης.

- Και οι δύο οικογένειες μοντέλων μπορούν να προβλέψουν υποθέτοντας είτε έλλειψη εποχιακότητας είτε προσθετική ή πολλαπλασιαστική εποχιακότητα.
- Στην περίπτωση της εκθετικής εξομάλυνσης η συνιστώσα της εποχιακότητας αλληλοεπιδρά διαρκώς με τις υπόλοιπες, επαναπροσδιορίζοντας συνεχώς την τιμή της. Αντίθετα, στην περίπτωση του μοντέλου θ εφαρμόζεται ντετερμινιστικά και παραμένει σταθερή σε κάθε περίοδο.
- Η τάση στα μοντέλα εκθετικής εξομάλυνσης επαναπροσδιορίζεται σε κάθε περίοδο βάσει σφάλματος μοντέλου σε μία προσπάθεια προσαρμογής τους στα πιο πρόσφατα δεδομένα. Αντίθετα η τάση στο μοντέλο θ υπολογίζεται εξ αρχής και εφαρμόζεται ντετερμινιστικά. Έτσι, αντί να υιοθετείται η βραχυπρόθεσμη τάση της χρονοσειράς για την προέκτασή της, υιοθετείται η μακροπρόθεσμη που αναμένεται να είναι και η πλέον στιβαρή.
- Και οι δύο οικογένειες μοντέλων μπορούν να συνδυάσουν τις συνιστώσες τάσης και επιπέδου με πολλαπλασιαστικό ή προσθετικό τρόπο. Η διαφορά έγκειται στο γεγονός πως στην περίπτωση της εκθετικής εξομάλυνσης αυτά αλληλοεπιδρούν σε κάθε περίοδο και το ένα επηρεάζει τη τιμή του άλλου βάσει του σφάλματος που παράγει ο συνδυασμός τους. Αντίθετα, στην περίπτωση του θ οι συνιστώσες εκτιμώνται και προεκτείνονται ξεχωριστά και στη συνέχεια συνδυάζονται για να προσδώσουν το τελικό αποτέλεσμα.

- Το επίπεδο και η τάση υπολογίζεται στην περίπτωση του θ επί μετασχηματισμένων δεδομένων (μεγαλύτερης ή μικρότερης καμπυλότητας αντίστοιχα) ενώ στην περίπτωση της εξομάλυνσης επί των αρχικών δεδομένων. Έτσι, ακόμα και στην απλούστερη μορφή των δύο μοντέλων, οι προβλέψεις τους θα διαφέρουν σημαντικά.
- Τα μοντέλα εκθετικής εξομάλυνσης μπορούν να υποθέσουν είτε γραμμική είτε μεταβαλλόμενη τάση. Χαρακτηριστικά παραδείγματα είναι π.χ. η Holt και η Damped. Για να γίνει το ίδιο στην περίπτωση του θ θα πρέπει να υποθεθεί γραμμική τάση ή εκθετική/λογαριθμική τάση ανάλογα με το πρόσημό της (αύξουσα ή φθίνουσα) και να χρησιμοποιηθεί αντίστοιχο μοντέλο.

Συμπερασματικά οι δύο οικογένειες μοντέλων, αν και έχουν παρόμοιες ιδιότητες, διακρίνονται σε μεγάλο βαθμό από αντίθετες φιλοσοφίες. Το μοντέλο θ αντιμετωπίζει ντετερμινιστικά τα δεδομένα θεωρώντας ότι η μέση κατάσταση που χαρακτηρίζει τη χρονοσειρά είναι και αυτή που θα οδηγήσει σε αντιπροσωπευτικότερες προβλέψεις. Αντίθετα, η εκθετική εξομάλυνση βασίζεται στον διαρκή επαναπροσδιορισμό και ανατροφοδότηση του μοντέλου για την παραγωγή προβλέψεων που αντιπροσωπεύουν όσο το δυνατό καλύτερα τις πιο πρόσφατες παρατηρήσεις της χρονοσειράς.

Σημειώνεται πως στα πλαίσια της παρούσας διατριβής θα εξεταστεί χωρίς βλάβη της γενικότητας μόνο η περίπτωση της πολλαπλασιαστικής εποχιακότητας -αν υπάρχει – καθώς, αφενός όπως αναφέρθηκε σε προηγούμενη ενότητα οι διαφορές μεταξύ των δύο προσεγγίσεων είναι αμελητέες, και αφετέρου η συρρίκνωση των εποχιακών συντελεστών που θα αξιοποιηθεί στη μεθοδολογία που θα κατασκευαστεί έχει οριστεί μόνο για την περίπτωση της κλασικής μεθόδου αποσύνθεσης.

Ενδεικτικά, και προκειμένου να αναδειχτεί η συμβολή της πολλαπλασιαστικής έκφρασης της μεθόδου θ στη γενίκευσή της, την αύξηση της ευελιξίας της και τη βελτίωση της προσαρμογής της σε χρονοσειρές ποικίλων τύπων, στον *Πίνακα 3.5* παρουσιάζεται η ακρίβεια 10 διαφορετικών μοντέλων θ : Οι χρονοσειρές αποεποχικοποιούνται βάσει της κλασικής πολλαπλασιαστικής μεθόδου αποσύνθεσης και στη συνέχεια χρησιμοποιείται το προσθετικό το πολλαπλασιαστικό μοντέλο θ για την προέκτασή τους. Μάλιστα, ως γραμμές μηδέν χρησιμοποιούνται οι πέντε καμπύλες που προτάθηκαν νωρίτερα, οδηγώντας συνολικά σε $2*5=10$ μοντέλα γραμμικής ή μη τάσης και προσθετικής ή πολλαπλασιαστικής αλληλεπίδρασής της με το επίπεδο της χρονοσειράς.

Πίνακας 3.5: Προβλεπτική ακρίβεια των προσθετικών και πολλαπλασιαστικών μοντέλων θ (sMAPE) για κλασική πολλαπλασιαστική αποεποχικοποίηση των δεδομένων των διαγωνισμών M1 και M3 και χρήση εναλλακτικών μοτίβων τάσης.

Μοντέλο	M3 Ετήσια	M3 Τριμηνιαία	M3 Μηνιαία	M3 Άλλα	M1 Ετήσια	M1 Τριμηνιαία	M1 Μηνιαία	Σύνολο
Γραμμικό-Προσθετικό	16.73	9.29	13.87	4.92	20.19	16.30	16.52	13.89
Εκθετικό-Προσθετικό	16.44	9.28	13.68	5.31	20.40	16.04	14.73	13.52
Λογαριθμικό-Προσθετικό	17.20	9.60	14.00	5.87	22.47	17.00	15.14	14.04
Αντίστροφο-Προσθετικό	17.73	9.81	14.25	6.27	23.23	17.53	14.95	14.31

Δύναμη-Προσθετικό	16.86	9.56	14.05	5.96	21.73	16.90	14.76	13.90
Γραμμικό-Πολ/στικό	16.88	9.21	14.20	4.98	20.11	16.38	16.54	14.03
Εκθετικό-Πολ/στικό	15.93	9.12	13.96	5.36	18.56	16.12	14.73	13.43
Λογαριθμικό-Πολ/στικό	17.24	9.56	14.29	5.96	22.21	17.08	14.91	14.11
Αντίστροφο-Πολ/στικό	18.15	9.92	14.60	6.27	22.90	18.22	15.04	14.55
Δύναμη-Πολ/στικό	16.64	9.44	14.38	6.04	21.12	16.87	14.82	13.95

Τα αποτελέσματα είναι ιδιαίτερα ενθαρρυντικά καθώς, όπως παρατηρείται, το πολλαπλασιαστικό εκθετικό μοντέλο θ είναι το πλέον ακριβές στο σύνολο των 4004 χρονοσειρών, αφήνοντας πίσω του το μέχρι πρότινος καλύτερο αθροιστικό εκθετικό μοντέλο. Η επίδοσή του είναι εντυπωσιακή ιδιαίτερα για την περίπτωση των ετήσιων δεδομένων όπου υπερέχει των υπολοίπων κατά περισσότερο από μία ποσοστιαία μονάδα. Βελτιωμένη είναι αντίστοιχα η επίδοση και άλλων μοντέλων μη γραμμικής τάσης, όπως π.χ. του λογαρίθμου και τη δύναμης για τα τριμηνιαία δεδομένα.

Στη γενική περίπτωση τα προσθετικά μοντέλα φαίνεται να υπερτερούν έναντι των πολλαπλασιαστικών, με τα τελευταία να αποδίδουν ωστόσο ιδιαιτέρως καλά σε χρονοσειρές μικρότερης συχνότητας, όπως π.χ. οι ετήσιες και οι τριμηνιαίες. Αυτό ίσως να σχετίζεται με την ιδιότητα των πολλαπλασιαστικών μοντέλων να υπερτονίζουν τον θόρυβο στις χρονοσειρές, γεγονός που στην περίπτωση των ετησίων δεδομένων οδηγεί στην έμφαση χρήσιμων ιδιαιτεροτήτων των χρονοσειρών, και στην περίπτωση των μηνιαίων, όπου είναι και εντονότερος, στον αποπροσανατολισμό του μοντέλου.

Επίσης, αν και στα επιμέρους σετ δεδομένων εντοπίζονται σημαντικές διαφορές μεταξύ των δύο μοντέλων για το ίδιο μοτίβο τάσης, στο σύνολό τους τα περισσότερα εμφανίζουν παραπλήσια συμπεριφορά, με εξαίρεση κυρίως το γραμμικό μοντέλο. Έτσι, συμπεραίνει κανείς πως η κατανομή των δεδομένων ανά συχνότητα, και πολύ πιθανό ανά ένταση ποιοτικού χαρακτηριστικού (τάση, εποχιακότητα κ.ο.κ.), επηρεάζει σημαντικά τα αποτελέσματα και αναδεικνύει τα οφέλη που μπορεί να προσδώσει η γενικευμένη έκφραση της μεθόδου θ μεμονωμένα σε επίπεδο χρονοσειράς.

*Έχω δει το μέλλον και είναι όπως το παρόν.
Διαφέρει μόνο ως προς το ότι είναι μακρύτερο.*

Dan Quisenberry

Κεφάλαιο 4: Εξομάλυνση χρονοσειρών

Οι χρονοσειρές, ή χρονολογικές σειρές, αποτελούν ένα σύνολο διαδοχικών παρατηρήσεων της τιμής ενός μεγέθους. Δεδομένου ότι οι επιμέρους παρατηρήσεις είναι συνεχείς και διατεταγμένες στον άξονα του χρόνου, γίνεται αμέσως αντιληπτό ότι παρέχουν σημαντική πληροφορία σχετικά με τα χαρακτηριστικά του εκάστοτε μεγέθους και τον τρόπο με τον οποίο αυτό εξελίσσεται. Δυστυχώς βέβαια η αποτελεσματική άντληση της εν λόγω πληροφορίας, η οποία και αποτελεί προαπαιτούμενο για την ικανοποιητική πρόβλεψη του μεγέθους, δεν είναι πάντοτε εύκολη, κυρίως λόγω της ύπαρξης θορύβου στα δεδομένα που καλούμαστε να προβλέψουμε.

Ως θόρυβος, ή τυχαιότητα, ορίζεται η μη ερμηνευθείσα διακύμανση μιας χρονοσειράς η οποία μπορεί να οφείλεται είτε στην επίδραση εξωτερικών μεταβλητών που δεν λαμβάνουμε υπόψη κατά την ανάλυσή της, είτε στη γενικότερη αβεβαιότητα που χαρακτηρίζει την εξέλιξη του κάθε μεγέθους. Σε κάθε περίπτωση ο θόρυβος μειώνει ανάλογα με την έκτασή του την ακρίβεια των μοντέλων πρόβλεψης αλλοιώνοντας τα πραγματικά μοτίβα της χρονοσειράς, το σήμα δηλαδή που καλούνται να εντοπίσουν και να προεκτείνουν. Χαρακτηριστικό παράδειγμα αποτελεί η επίδραση που έχουν οι ακραίες τιμές και άλλες μορφές *outlier* στην επίδραση των μοντέλων πρόβλεψης: Όταν το μοντέλο συναντήσει μία ασυνήθιστη τιμή θα προσπαθήσει να αναπροσαρμοστεί ούτως ώστε να μειώσει το σφάλμα προσαρμογής του. Καθώς γίνεται αυτό το μοντέλο υποθέτει στην πραγματικότητα ότι η εν λόγω ασυνήθιστη τιμή δεν αποτελεί *outlier* αλλά ένα νέο μοτίβο της χρονοσειράς το οποίο και οφείλει να εκτιμήσει. Έτσι, ακόμα και με το πέρασ της εν λόγω τιμής, το μοντέλο θα συνεχίσει να επηρεάζεται από αυτή μέχρις ότου μεσολαβήσει ένα διάστημα ικανό για να αντιληφθεί ότι πρέπει να το αγνοήσει. Στην περίπτωση θορύβου περιορισμένης έντασης το εν λόγω φαινόμενο είναι αμελητέο και η παραμετροποίηση του μοντέλου γίνεται με σχετική ασφάλεια. Ωστόσο, στην περίπτωση θορύβου υψηλής έντασης η παραμετροποίηση του μοντέλου εμπεριέχει σημαντική αβεβαιότητα οδηγώντας πιθανώς σε αστοχίες.

Για την ανάπτυξη λοιπόν μίας γενικευμένης μεθοδολογίας ικανής να παράγει ακριβείς προβλέψεις ανεξαρτήτως των χαρακτηριστικών της χρονοσειράς, και συνεπώς και του επιπέδου θορύβου σε αυτές, απαιτείται η αντίστοιχη προετοιμασία της ούτως ώστε να αποκαλυφθούν τα πραγματικά της μοτίβα και να μειωθεί η αβεβαιότητα κατά την παραμετροποίηση του μοντέλου με το οποίο θα παραχθούν οι προβλέψεις. Σε αυτό το κεφάλαιο θα παρουσιαστούν κάποιες μέθοδοι που μπορούν να χρησιμοποιηθούν για την επίτευξη του παραπάνω σκοπού, θα αναφερθούν οι λόγοι για τους οποίους η τεχνική της εξομάλυνση φαντάζει η πλέον αποδοτική λύση στην περίπτωση πρόβλεψης με τη μέθοδο θ και θα προταθεί μία νέα εκδοχή της που μπορεί να αξιοποιηθεί για τη βέλτιστη προετοιμασία των χρονοσειρών και την παραγωγή στιβαρών προβλέψεων.

4.1 Προετοιμασία χρονοσειρών: Αποκάλυψη μοτίβων και μείωση της αβεβαιότητας κατά την παραμετροποίηση των μοντέλων πρόβλεψης

Η προετοιμασία χρονοσειρών περιλαμβάνει την κατάλληλη επεξεργασία των δεδομένων πριν την εφαρμογή κάποιου μοντέλου πρόβλεψης με σκοπό τη γενικότερη βελτίωση της προβλεπτικής τους ακρίβειας. Φυσικά, το ποιες τεχνικές θα χρησιμοποιηθούν για την προετοιμασία μίας χρονοσειράς εξαρτάται σημαντικά τόσο από τα χαρακτηριστικά τη ίδιας της χρονοσειράς, όσο και από το μοντέλο πρόβλεψης.

Χαρακτηριστικό παράδειγμα αποτελεί η διαδικασία της *αποεποχικοποίησης* η οποία, όταν το μοντέλο πρόβλεψης που χρησιμοποιείται δεν είναι εποχιακό, είναι απαραίτητη προκειμένου να μην θεωρηθεί η εποχιακότητα ως μία επιπλέον μη συστηματική διακύμανση και να ληφθεί αποτελεσματικά υπόψη κατά την παραγωγή των προβλέψεων. Η εν λόγω διαδικασία δεν θα μας απασχολήσει εδώ καθώς έχει παρουσιαστεί ήδη αναλυτικά και έχουν αναφερθεί εκτενώς τα οφέλη της για μη εποχιακά μοντέλα, όπως είναι π.χ. η μέθοδος *θ*, η *holt* και η *damped*.

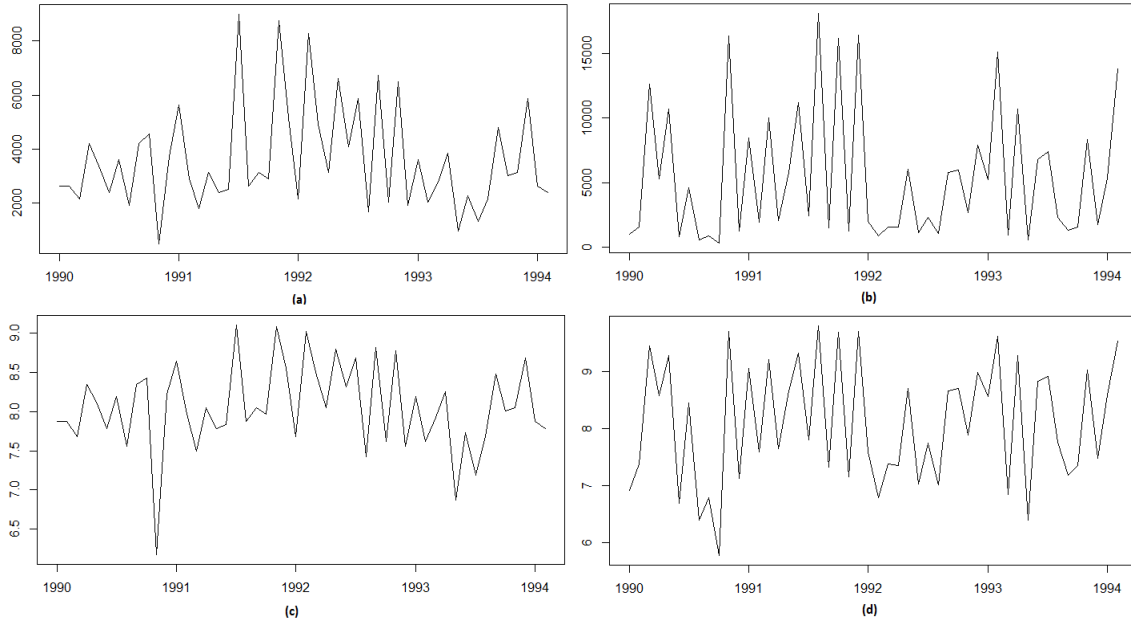
Άλλες διαδικασίες που μπορούν να ενταχθούν στις βασικές τεχνικές προετοιμασίας χρονοσειρών είναι ο *έλεγχος και η εξασφάλιση στασιμότητας και κανονικότητας* στα δεδομένα. Η πρώτη διαδικασία, η οποία όπως είδαμε στο [Κεφάλαιο 2](#) εφαρμόζεται μέσω αντίστοιχων στατιστικών ελέγχων και τη διαδικασία της διαφόρισης, είναι απαραίτητη μόνο στην περίπτωση μοντέλων που δεν λαμβάνουν υπόψη τους την τάση των δεδομένων, όπως είναι π.χ. τα μοντέλα *ARMA* και η μέθοδος της απλής εκθετικής εξομάλυνσης. Η κανονικότητα των δεδομένων από την άλλη μεριά μπορεί να φανεί ιδιαίτερα χρήσιμη για κάθε μοντέλο πρόβλεψης καθώς, επί της ουσίας, βοηθάει στον περιορισμό της διακύμανσης της χρονοσειράς και συνεπώς στην ανάδειξη των χαρακτηριστικών της, όπως π.χ. του τρέχοντος επιπέδου και της τάσης.

Αρκετές είναι οι μελέτες που έχουν πραγματοποιηθεί προκειμένου να εξακριβωθούν τα πιθανά οφέλη των μετασχηματισμών στην προβλεπτική ικανότητα των μοντέλων πρόβλεψης. Οι *Nelson & Granger (1979)* υποστήριξαν με μελέτη που έκαναν σε 21 χρονοσειρές με χρήση μοντέλων *ARIMA* πως ο μετασχηματισμός *Box-Cox* δεν οδηγεί απαραίτητα σε προβλέψεις βελτιωμένης ακρίβειας ενώ οι *Thyer et al. (2002)* ισχυρίστηκαν ότι ο εν λόγω μετασχηματισμός μπορεί υπό συγκεκριμένες συνθήκες να καταστεί ιδιαίτερα ωφέλιμος. Πιο πρόσφατα ο *Beaumont (2014)* πραγματοποίησε μία σειρά διαφόρων τύπων μετασχηματισμών στα δεδομένα του διαγωνισμού *M3* και απέδειξε πως στην περίπτωση των μηνιαίων δεδομένων οι προβλέψεις που προέρχονταν από μετασχηματισμένες χρονοσειρές ήταν σημαντικά καλύτερες από αυτές των αρχικών χρονοσειρών ενώ για συγκεκριμένες κατηγορίες μετασχηματισμών το ίδιο συμπέρασμα εξαγόταν και για την περίπτωση των ετήσιων και τριμηνιαίων δεδομένων. Οι *Proietti & Lütkepohl (2013)* πρότειναν μία νέα μέθοδο για τον καθορισμό του βέλτιστου μετασχηματισμού *Box-Cox* και μελέτησαν την αποτελεσματικότητά του σε μεγάλο σετ από μακροοικονομικά δεδομένα και χρονοσειρές βιομηχανικής παραγωγής. Τα αποτελέσματα έδειξαν ότι στο 1/5 των περιπτώσεων ο μετασχηματισμός οδηγούσε για μικρούς ορίζοντες σε σαφώς καλύτερες προβλέψεις. Καθώς ο ορίζοντας πρόβλεψης αυξανόταν τα οφέλη του μετασχηματισμού έφθιναν ενώ στις περισσότερες περιπτώσεις ο προτεινόμενος

μετασχηματισμός ήταν αυτός της απλής λογαρίθμησης. Τέλος, ο *McKenzie (1999)* ερεύνησε την επίδραση του μετασχηματισμού *Box-Cox* σε δεδομένα χρηματιστηρίου και κατέληξε στο συμπέρασμα πως μπορούν να υπάρξουν οφέλη, ωστόσο αυτά εξαρτώνται σε μεγάλο βαθμό από το σφάλμα που χρησιμοποιείται για την εκτίμησή τους.

Φυσικά, οι παρατηρήσεις μίας χρονοσειράς ενδέχεται να ακολουθούν ήδη την κανονική κατανομή οπότε και ο έλεγχος κανονικότητας είναι απαραίτητος πριν την εφαρμογή ενός μετασχηματισμού προκειμένου να μην πραγματοποιούνται περιττές διαδικασίες και να μην οδηγούμαστε σε υπερβολές. Βάσει των αποτελεσμάτων του ελέγχου μπορεί στη συνέχεια να εφαρμοστεί ο κατάλληλος μετασχηματισμός (π.χ. *Box-Cox* ή *Johnson*) για την εξασφάλιση κανονικότητας και τον περιορισμό της συνολικής διακύμανσης στην χρονοσειρά. Βασικό μειονέκτημα αυτής της κατά τα άλλα πολύτιμης διαδικασίας είναι, όπως αναφέρθηκε και νωρίτερα, αφενός ότι τα αποτελέσματα που προκύπτουν δεν είναι πάντοτε τα απολύτως επιθυμητά, και αφετέρου ότι συχνά προκύπτουν προβλήματα απροσδιοριστίας. Για παράδειγμα, η εφαρμογή του μετασχηματισμού του λογαρίθμου, ο οποίος έχει εμπειρικά αποδειχτεί ότι οδηγεί σε εξίσου ικανοποιητικά αποτελέσματα σε σύγκριση με άλλους πολύ πιο σύνθετους μετασχηματισμούς, δεν ορίζεται στην περίπτωση χρονοσειρών με αρνητικές και μηδενικές τιμές, ενώ επιπρόσθετα στην περίπτωση χρονοσειρών με ακραία μικρές τιμές σε σύγκριση με αυτές του υπολοίπου δείγματος, μπορεί να υπερτονίσει *outlier* αντί να τα εξαλείψει. Ανάλογα δηλαδή τη φύση του μετασχηματισμού που εφαρμόζεται υπάρχει περίπτωση στην προσπάθειά μας να περιορίσουμε τη συνολική διακύμανση της χρονοσειράς να δημιουργήσουμε τοπικά νέα *outliers*. Αυτό συμβαίνει καθώς οι μετασχηματισμοί λόγω του μη γραμμικού τους χαρακτήρα συμπεριφέρονται ανομοιόμορφα απέναντι σε παρατηρήσεις μικρής και μεγάλης τιμής. Επιπλέον, θα πρέπει να έχουμε υπόψη ότι για ορισμένες ιδιαίτερα ανομοιόμορφες χρονοσειρές ενδέχεται να μην υπάρχει μετασχηματισμός που να οδηγεί αν όχι σε απόλυτη, τουλάχιστον σε ικανοποιητική κανονικότητα.

Ένα παράδειγμα των παραπάνω ισχυρισμών παρουσιάζεται στην *Εικόνα 4.1* όπου ο μετασχηματισμός του λογαρίθμου εφαρμόζεται σε δύο χρονοσειρές, μία με μικρές και μεγάλες ακραίες τιμές (a) και μία μόνο με μεγάλες (b). Όπως φαίνεται, στην πρώτη περίπτωση ο λογάριθμος οδηγεί σε μείωση της συνολικής διακύμανσης αλλά εισάγει ένα νέο *outlier* (c). Αντίθετα, στη δεύτερη περίπτωση ο μετασχηματισμός οδηγεί σε ικανοποιητική κανονικότητα χωρίς να παράγει νέες ανομοιομορφίες (d).



Εικόνα 4.1: Εφαρμογή του μετασχηματισμού λογαρίθμου για περιορισμό της διακύμανσης σε δύο χρονοσειρές με ακραίες τιμές θετικής και αρνητικής κατεύθυνσης.

Συμπερασματικά, η αίσθηση που υπάρχει εμπειρικά γύρω από την εφαρμογή μετασχηματισμών για την προετοιμασία των χρονοσειρών πριν την παραγωγή προβλέψεων είναι ότι, αν και αυτοί δεν εγγυούνται βελτιωμένη προβλεπτική ακρίβεια, η πιθανότητα να υπάρξουν έστω και οριακά οφέλη ή τουλάχιστον να μην παραχθούν σημαντικά χειρότερες προβλέψεις σε σχέση με αυτές των αρχικών δεδομένων, είναι σχετικά μεγάλη. Επίσης, τα προσδοκόμενα οφέλη φαίνεται να μην αυξάνονται με τη χρήση πολύπλοκων μετασχηματισμών, γεγονός που καταστεί την εφαρμογή απλούστερων μετασχηματισμών, όπως π.χ. των Box-Cox ή ακόμα και της απλής λογαρίθμησης, επαρκή λύση. Οι μετασχηματισμοί αποτελούν συνεπώς ένα ισχυρό εργαλείο για τη βελτίωση της απόδοσης μίας προβλεπτικής διαδικασίας, το μέγεθος της οποίας εξαρτάται ωστόσο σημαντικά από τις ιδιαιτερότητες των δεδομένων. Αυτό αποδεικνύεται άλλωστε και από τα αντικρουόμενα αποτελέσματα μεταξύ των διαφόρων μελετών στη βιβλιογραφία.

Μία λύση για να επιτύχει κανείς βελτιωμένη ακρίβεια πρόβλεψης χωρίς την αβεβαιότητά και τους περιορισμούς των μετασχηματισμών, είναι να εντοπίσει και να απομακρύνει από τη χρονοσειρά πιθανά *outlier*. Το πλεονέκτημα της εν λόγω μεθόδου έγκειται στο γεγονός ότι η απομάκρυνση αποκλειστικά των ακραίων τιμών των χρονοσειρών εγγυάται συστηματική βελτίωση της ακρίβειας πρόβλεψης καθώς με αυτόν τον τρόπο δεν αλλοιώνεται κανένα από τα μοτίβα της χρονοσειράς ενώ περιορίζεται η διακύμανσή της. Σημαντικό μειονέκτημα βέβαια της εν λόγω προσέγγισης είναι ότι προκειμένου να εντοπιστούν τα *outlier* πρέπει να εφαρμοστούν επιπλέον διαγνωστικοί έλεγχοι για τους οποίους πρέπει να οριστούν παράμετροι όπως το τι θεωρείται 'ακραία' τιμή, με τη χρήση ποιου μοντέλου θα εντοπιστεί ένα *outlier* και το πόσο πρέπει να μεταβληθεί μία παρατήρηση προκειμένου να θεωρηθεί 'φυσιολογική'. Άλλο σημαντικό μειονέκτημα της

τεχνικής απαλοιφής *outlier* είναι ο σημαντικός χρόνος που απαιτείται προκειμένου να ολοκληρωθεί η όλη διαδικασία, παράγοντας που την καθιστά ακατάλληλη για την πρόβλεψη μεγάλου πλήθους χρονοσειρών.

Η απαλοιφή του θορύβου μπορεί να επιτευχθεί εναλλακτικά μέσω μίας διαδικασίας εξομάλυνσης. Η εξομάλυνση των χρονοσειρών βασίζεται στην λογική ότι, αν και κάποιες μεμονωμένες παρατηρήσεις ενδέχεται να αποτελούν *outlier* παραπλανώντας το μοντέλο πρόβλεψης σχετικά με τα μοτίβα της χρονοσειράς, στο σύνολο n διαδοχικών παρατηρήσεων τα μοτίβα θα είναι σαφώς πιο ασφαλές να εκτιμηθούν καθώς η πιθανότητα να χαρακτηρίζονται όλες από αυξημένη τυχαιότητα είναι σημαντικά μικρότερη. Έτσι, αν η τιμή κάθε παρατήρησης μεταβληθεί κατάλληλα βάσει των n γειτονικών της παρατηρήσεων, θα καταλήξουμε σε μία νέα σαφώς πιο ομαλή και εύκολο να προεκταθεί χρονοσειρά οι παρατηρήσεις τις οποίας θα είναι αρκετά πιο αντιπροσωπευτικές σε σύγκριση με τις αρχικές. Βασικό πλεονέκτημα της εν λόγω τεχνικής είναι ότι περιορίζει σημαντικά τη διακύμανση των χρονοσειρών εφαρμόζοντας απλούς μετασχηματισμούς στα δεδομένα και μάλιστα χωρίς να χρειάζεται η παραμετροποίηση κάποιου επιπρόσθετου μοντέλου ή η εφαρμογή κάποιου διαγνωστικού ελέγχου. Έτσι, η προβλεπτική ικανότητα των μοντέλων πρόβλεψης βελτιώνεται χωρίς να εισάγεται επιπλέον αβεβαιότητα στη διαδικασία της πρόβλεψης. Επιπλέον, δεδομένης της απλότητας των τεχνικών εξομάλυνσης, η πολυπλοκότητα και το υπολογιστικό κόστος της διαδικασίας πρόβλεψης δεν αυξάνονται σημαντικά καθιστώντας τη χρήση τους ιδανική επιλογή για την πρόβλεψη μεγάλου πλήθους χρονοσειρών ποικίλων χαρακτηριστικών.

Σε αυτό το σημείο αναφέρουμε ότι δύο άλλες τεχνικές που εντοπίζονται στη βιβλιογραφία για τον έμμεσο περιορισμό την επίδραση της λανθασμένης παραμετροποίησης των μοντέλων πρόβλεψης λόγω της επίδρασης του θορύβου και άλλων σχετικών παραγόντων είναι η χρήση της μεθόδου συνάθροισης δεδομένων σε εναλλακτικές συχνότητες (*temporal aggregation*) καθώς και ο συνδυασμός προβλέψεων (*forecast combination*). Στην πρώτη περίπτωση τα δεδομένα μίας υψηλής συχνότητας (π.χ. μηνιαία) αθροίζονται σε μία χαμηλότερης (π.χ. ετήσια) και πραγματοποιείται πρόβλεψη επί αυτών με την πεποίθηση ότι αφού στη νέα συχνότητα η διακύμανση θα είναι μικρότερη, η παραμετροποίηση του αντίστοιχου μοντέλου θα γίνει με μεγαλύτερη επιτυχία (*Silvestrini & Veredas, 2008*). Στη δεύτερη περίπτωση αποδεχόμαστε ότι κάποια μεμονωμένα μοντέλα θα επηρεαστούν σημαντικά από ανομοιομορφίες στη χρονοσειρά, ωστόσο θεωρούμε ότι αυτό δεν είναι δυνατόν να συμβεί για κάθε πιθανό μοντέλο πρόβλεψης (*Hibon & Evgeniou 2005*). Έτσι, αν συνδυάσουμε κατάλληλα τις προβλέψεις πολλών διαφορετικών μοντέλων η τελική μέση πρόβλεψη αναμένεται να έχει επηρεαστεί λιγότερο από τα μειονεκτήματα των επιμέρους μεθόδων. Μάλιστα, συχνά πραγματοποιείται συνδυασμός των παραπάνω τεχνικών (συνδυασμός προβλέψεων από όλες τις δυνατές συχνότητες) με αντίστοιχα πλεονεκτήματα, γνωστός και ως *multiple temporal aggregation* (*Kourentzes et al. 2014*).

Οι εν λόγω τεχνικές δεν κρίνονται ωστόσο ιδιαίτερα ωφέλιμες για την παρούσα μεθοδολογία καθώς αφενός το μοντέλο πρόβλεψης που θα χρησιμοποιηθεί θα είναι ένα, οπότε και ο συνδυασμός προβλέψεων είναι αδύνατος, και αφετέρου η πρόβλεψη σε μία

εναλλακτική συχνότητα απαιτεί την αναγνώριση του βέλτιστου επιπέδου συνάθροισης, το οποίο σε καμία περίπτωση δεν είναι εύκολο να καθοριστεί. Από την άλλη μεριά η πρόβλεψη σε πολλαπλές συχνότητες είναι μάλλον ανούσια όταν αυτό πραγματοποιείται μέσω της μεθόδου θ καθώς η γραμμή μηδέν που προκύπτει θα έχει την ίδια κλίση σε κάθε επίπεδο συνάθροισης και συνεπώς η μόνο πληροφορία που προστίθεται μέσω αυτής της τεχνικής θα προέρχεται από τη γραμμή δύο. Δεδομένης λοιπόν της αυξανόμενης πολυπλοκότητας της εν λόγω τεχνικής, η χρήση της για μεγάλο πλήθος χρονοσειρών είναι μάλλον απαγορευτική. Άλλωστε, τα οφέλη της πολλαπλής συνάθροισης οφείλονται σε μεγάλο βαθμό στην καλύτερη εκτίμηση των μακροχρόνιων χαρακτηριστικών της χρονοσειράς, κυρίως δηλαδή της τάσης της, η οποία και εκτιμάται ικανοποιητικά στην περίπτωση της μεθόδου θ μέσω της γραμμής μηδέν.

Έτσι, στις παραγράφους που ακολουθούν παρουσιάζονται κάποιες βασικές μέθοδοι εξομάλυνσης χρονοσειρών με σκοπό να εντοπιστούν τα πλεονεκτήματα και τα μειονεκτήματά τους. Στόχος είναι βάσει των παραπάνω συμπερασμάτων να προταθεί μία αποδοτική μέθοδος για την αποτελεσματική προετοιμασία των χρονοσειρών πριν την εφαρμογή της μεθόδου πρόβλεψης θ .

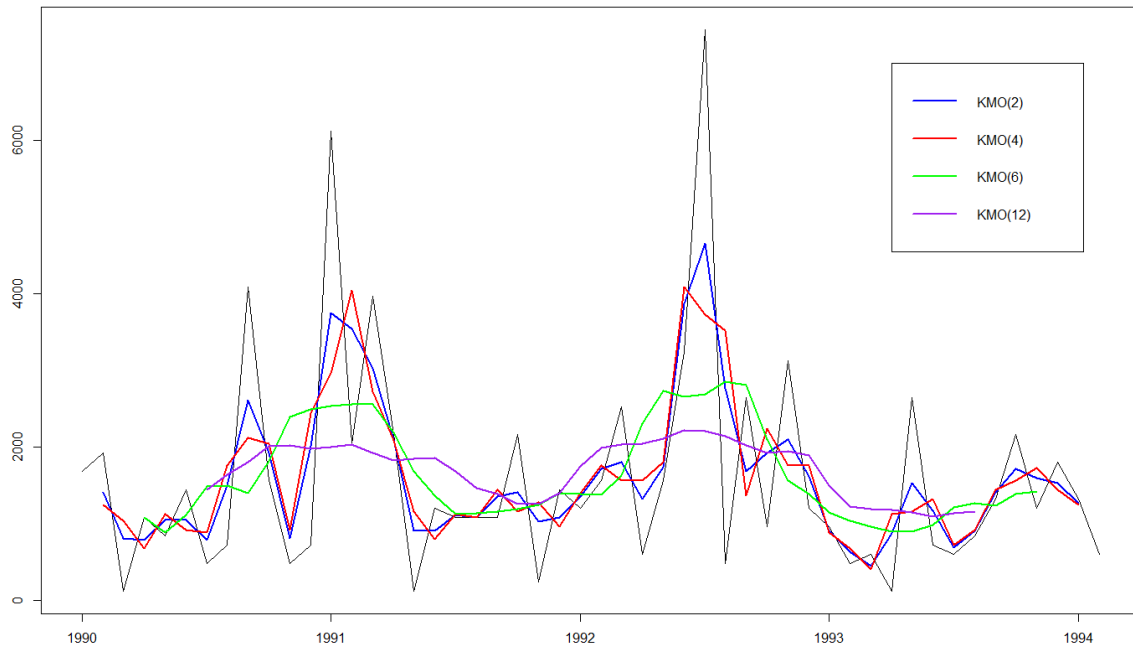
4.2 Εξομάλυνση χρονοσειρών με χρήση κινητών μέσων όρων

Οι κινητοί μέσοι όροι (*KMO*) είναι η απλούστερη μέθοδος εξομάλυνσης χρονοσειρών και η εφαρμογή τους οδηγεί όπως παρουσιάστηκε και στην περιγραφή των μεθόδων αποεποχικοποίησης στην απομόνωση της τάσης και του κύκλου της χρονοσειράς, στην απομάκρυνση δηλαδή των εποχιακών και τυχαίων διακυμάνσεων. Φυσικά, επειδή σε αυτό το σημείο θεωρούμε ότι αντιμετωπίζουμε αποεποχικοποιημένες χρονοσειρές, το στοιχείο της εποχιακότητας μπορούμε να θεωρήσουμε ότι έχει εξαλειφθεί αποτελεσματικά οπότε και η χρήση τους στοχεύει αποκλειστικά στον περιορισμό του θορύβου. Έτσι, ενώ στην περίπτωση των εποχιακών χρονοσειρών η τάξη του *KMO* που θα επιλεγεί εξαρτάται από τη συχνότητά τους, εδώ μπορεί να χρησιμοποιηθεί θεωρητικά *KMO* οποιασδήποτε τάξης.

Πιο συγκεκριμένα, όταν αναφερόμαστε σε *KMO* n τάξης, εννοούμε ότι n διαδοχικές παρατηρήσεις της χρονοσειράς συναθροίζονται προκειμένου ο μέσος τους όρος να αντικαταστήσει την παρατήρηση που βρίσκεται στο κέντρο του εν λόγω διαστήματος. Στην περίπτωση π.χ. ενός *KMO*(3) η παρατήρηση που βρίσκεται στη θέση t θα αντικατασταθεί από το μέσο όρο των παρατηρήσεων στις θέσεις $t-1$, t και $t+1$. Όσο αυξάνεται η τάξη του *KMO* αυξάνεται και το επίπεδο εξομάλυνσης καθώς αξιοποιείται πληροφορία από μεγαλύτερο χρονικό διάστημα και συνεπώς οδηγούμαστε σε σαφώς πιο λείες χρονοσειρές. Ένα χαρακτηριστικό παράδειγμα φαίνεται στην *Εικόνα 4.2*.

Βασικό πλεονέκτημα της μεθόδου αποτελεί η απλότητά της και συνεπώς το μικρό υπολογιστικό της κόστος, ενώ σημαντικό μειονέκτημά της είναι η αδυναμία εξομάλυνσης των άκρων της χρονοσειράς, η δημιουργία δηλαδή κενών τιμών στην αρχή και στο τέλος της λόγω έλλειψης δεδομένων. Μάλιστα, το παραπάνω φαινόμενο εντείνεται με την αύξηση της τάξης του *KMO* καθώς απαιτείται μεγαλύτερο πλήθος παρατηρήσεων για την εξομάλυνση κάθε παρατήρησης.

Στη συνέχεια παρουσιάζονται τα τέσσερα βασικότερα είδη ΚΜΟ καθώς και μία μέθοδο για τη συμπλήρωση των κενών τιμών που δημιουργούν, γνωστή και ως *backcasting*.



Εικόνα 4.2: Εφαρμογή ΚΜΟ διαφόρων τάξεων σε χρονοσειρά υψηλής τυχαιότητας.

Απλός κινητός μέσος όρος: Πρόκειται για την απλούστερη μορφή ΚΜΟ η οποία και συναθροίζει n γειτονικές παρατηρήσεις για την αντικατάσταση της παρατήρησης που βρίσκεται στο μέσο αυτών με τον μέσο τους όρο. Ο εν λόγω μετασχηματισμός μπορεί λοιπόν να εφαρμοστεί για όλες τις παρατηρήσεις της χρονοσειράς τις στιγμές t πλην των $n \bmod 2$ αρχικών και τελικών παρατηρήσεών της. Οι συγκεκριμένες τιμές λόγω έλλειψης δεδομένων θα παραμείνουν αναγκαστικά κενές ενώ όσο αυξάνεται η τάξη του ΚΜΟ τόσο θα μεγαλώνει και το πλήθος τους. Μάλιστα, για τον μέγιστο δυνατό βαθμό ΚΜΟ η τιμή εξομάλυνσης θα ταυτίζεται με τον μέσο όρο όλης της χρονοσειράς αφήνοντας όλες τις παρατηρήσεις Y της πλην την ενδιάμεση κενές. Μαθηματικά ο απλός ΚΜΟ, ο οποίος ορίζεται μόνο για περιττές τάξεις n , υπολογίζεται ως εξής:

$$KMO(n)_t = \frac{1}{n} \sum_{i=-(n \bmod 2)}^{n \bmod 2} Y_{t+i}$$

Διπλός κινητός μέσος όρος: Ο διπλός ΚΜΟ($n \times m$) αναφέρεται στη διπλή εφαρμογή του απλού ΚΜΟ με ίσα ή άνισα μήκη, στην εφαρμογή δηλαδή απλού ΚΜΟ(n) σε χρονοσειρά που έχει εξομαλυνθεί ήδη μέσω απλού ΚΜΟ(m). Η εν λόγω προσέγγιση οδηγεί σε πιο ομαλές χρονοσειρές σε σχέση με τον απλό ΚΜΟ αντίστοιχης τάξης, ωστόσο έχει το μειονέκτημα ότι παράγει περισσότερες κενές τιμές.

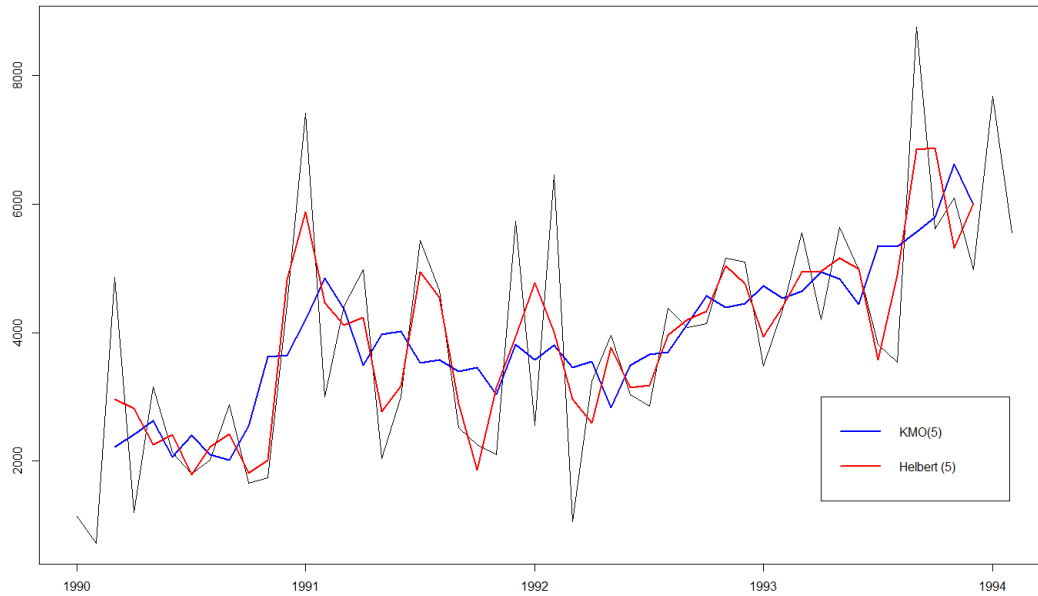
Κεντρικός κινητός μέσος όρος: Ο κεντρικός KMO αποτελεί εκδοχή του απλού KMO προκειμένου αυτός να βρίσκει εφαρμογή σε περιπτώσεις όπου η επιθυμητή τάξη είναι άρτιος αριθμός οπότε και υπάρχει θέμα καθορισμού της ενδιάμεσης παρατήρησης. Αν π.χ. επιθυμούμε τον υπολογισμό ενός KMO τάξης 4 τότε η θεωρητικά ενδιάμεση παρατήρηση των τεσσάρων παρατηρήσεων είναι η 2.5 η οποία φυσικά δεν αντιστοιχεί σε κάποια πραγματική τιμή της χρονοσειράς. Προκειμένου λοιπόν να γίνει η επιθυμητή αντιστοίχιση εκτιμάται τη χρονική περίοδο t ο $KMO(n)$ για τα σημεία $t-0.5$ και $t+0.5$ και ο μέσος όρος αυτών αποτελεί την εξομάλυνση της τρέχουσας παρατήρησης. Στην πραγματικότητα ο κεντρικός KMO αποτελεί δηλαδή υποπερίπτωση του διπλού KMO τάξης $(2\chi n)$.

Σταθμισμένος κινητός μέσος όρος: Πρόκειται για επέκταση του απλού KMO η οποία έχει ως σκοπό να οδηγήσει μία πιο εξομαλυμένο αποτέλεσμα. Η διαφοροποίησή του έγκειται στο γεγονός ότι κάθε μία από τις παρατηρήσεις που χρησιμοποιείται στη συνάθροιση για τον υπολογισμό του μέσου όρου διαθέτει διαφορετικό βάρος ανάλογα με την απόστασή της από την ενδιάμεση παρατήρηση. Μαθηματικά ο σταθμισμένος KMO υπολογίζεται ως εξής:

$$KMO(n)_t = \frac{1}{n} \sum_{i=-(n \bmod 2)}^{n \bmod 2} a_i Y_{t+i}$$

$$\sum_{i=-(n \bmod 2)}^{n \bmod 2} a_i = 1$$

Με βάση τα παραπάνω δύναται λοιπόν να προσδώσουμε μεγαλύτερο βάρος στις παρατηρήσεις που γειτνιάζουν περισσότερο με την προς εξομάλυνση παρατήρηση, οι τιμές των οποίων και οφείλουν να είναι συμμετρικές ως προς αυτή, δηλαδή $a_i = a_{-i}$. Σημειώνεται επίσης ότι δεν είναι αναγκαστικό τα βάρη να βρίσκονται μεταξύ 0 και 1 καθώς και ότι λόγω του ορισμού του μετασχηματισμού δεν απαιτείται στην πράξη να επιλέξουμε τάξη KMO αλλά βάρη για κάθε γειτονική παρατήρηση. Για παράδειγμα, αν θέλουμε να λαμβάνουμε υπόψη μόνο τις κοντινές παρατηρήσεις μπορούμε να μειώσουμε αισθητά το βάρος των απομακρυσμένων παρατηρήσεων χωρίς να περιορίσουμε την τάξη του KMO .



Εικόνα 4.3: Εφαρμογή απλού και σταθμισμένου $KMO(5)$ με βάρη Henderson.

Στη βιβλιογραφία προτείνονται αρκετές προσεγγίσεις για τον καθορισμό των βέλτιστων βαρών και τάξης σταθμισμένων KMO , χαρακτηριστικότερες εκ των οποίων είναι αυτές του *Spencer (1904)* και του *Henderson* για την απομόνωση του εποχιακού παράγοντα (*Ladiray & Quenneville, 2001*). Ένα παράδειγμα παρουσιάζεται στη *Εικόνα 4.5*. Επίσης, στην περίπτωση που τα βάρη υπολογιστούν μέσω κάποιας συνάρτησης *kernel* και σταθμιστούν κατάλληλα για να αθροίζονται στη μονάδα, ο σταθμισμένος KMO ταυτίζεται με την εξομάλυνση τύπου *kernel* η οποία και περιγράφεται στην επόμενη παράγραφο. Τέλος, λόγω της γενικευμένης έκφρασής του, ο σταθμισμένος KMO μπορεί να χρησιμοποιηθεί για την υλοποίηση όλων των εκδοχών KMO που παρουσιάστηκαν νωρίτερα με χρήση κατάλληλων βαρών. Στον πίνακα που ακολουθεί παρουσιάζονται κάποια τέτοια χαρακτηριστικά παραδείγματα.

Πίνακας 4.1: Τιμές βαρών σταθμισμένου κινητού μέσου όρου για κλασικές εκδοχές KMO .

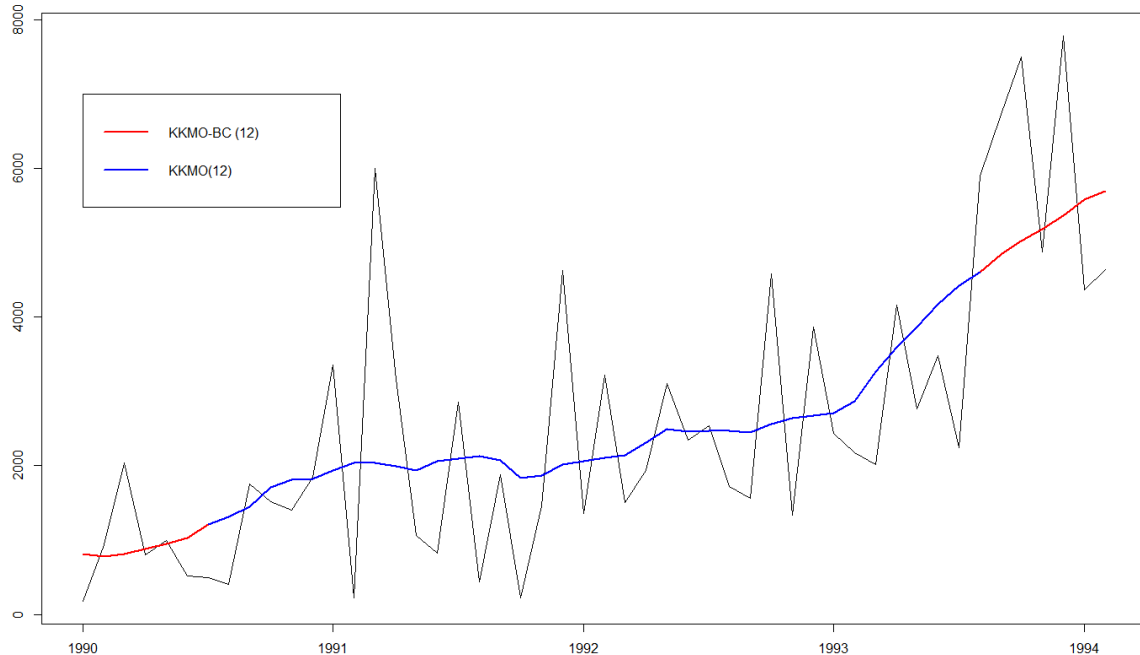
<i>KMO</i>	<i>a</i>₀	<i>a</i>₁	<i>a</i>₂	<i>a</i>₃	<i>a</i>₄	<i>a</i>₅	<i>a</i>₆	<i>a</i>₇	<i>a</i>₈	<i>a</i>₉	<i>a</i>₁₀	<i>a</i>₁₁
3	.333	.333										
5	.200	.200	.200									
2 x 12	.083	.083	.083	.083	.083	.083	.042					
3 x 3	.333	.222	.111									
3 x 5	.200	.200	.133	.067								
Hend. 5	.558	.294	-.073									
Hend. 9	.330	.267	.119	-.010	-.041							
Hend. 13	.240	.214	.147	.066	.000	-.028	-.019					
Hend. 23	.148	.138	.122	.097	.068	.039	.013	-.005	-.015	-.016	-.011	-.004
Spencer 15	.231	.209	.144	.066	.009	-.016	-.019	-.009				
Spencer 21	.171	.163	.134	.037	.051	.017	-.006	-.014	-.014	-.009	-.003	

Τεχνική Backcasting: Πρόκειται για την τεχνική που εφαρμόζεται προκειμένου η χρονοσειρά που παράγεται μετά την εξομάλυνση με ΚΜΟ να μην διαθέτει κενές τιμές. Όπως αναφέρθηκε νωρίτερα, ο απλός ΚΜΟ εφαρμόζεται σε όλες τις παρατηρήσεις της χρονοσειρά πλην των $n \bmod 2$ αρχικών και τελικών της τιμών καθώς για αυτές δεν επαρκούν τα διαθέσιμα δεδομένα. Κατά αντιστοιχία και στις υπόλοιπες περιπτώσεις ΚΜΟ υπάρχει ένα δεδομένο πλήθος κενών τιμών που θα προκύψει μετά την εφαρμογή τους. Αυτό αποτελεί σοβαρό πρόβλημα της εν λόγω τεχνικής εξομάλυνσης ως μέθοδο προετοιμασίας χρονοσειρών για την παραγωγή προβλέψεων καθώς μεταφέρει το σημείο έναρξης προβλέψεων πίσω στο παρελθόν και στερεί παρατηρήσεις από το μοντέλο πρόβλεψης τις οποίες και θα μπορούσε να αξιοποιήσει για την καλύτερη προσαρμογή του. Έτσι, η αναμενόμενη ακρίβεια πρόβλεψης μειώνεται σημαντικά.

Μία λύση για την επίλυση του εν λόγω προβλήματος είναι η κατάλληλη προέκταση των δύο άκρων της αρχικής χρονοσειράς με χρήση κάποιας μεθόδου πρόβλεψης. Η προέκταση του τέλους της χρονοσειράς επιτυγχάνεται με την εφαρμογή ενός μοντέλου πρόβλεψης σε αυτήν (*forecasting*), ενώ η προέκταση της αρχής της χρονοσειράς με αντιστροφή της χρονολογικής σειράς των παρατηρήσεων και εφαρμογή του ίδιου μοντέλου πρόβλεψης στη χρονοσειρά που προκύπτει (*backcasting*).

Προκειμένου βέβαια η τεχνική να κριθεί αποτελεσματική θα πρέπει αφενός ο ορίζοντας πρόβλεψης που θα χρησιμοποιηθεί να είναι τέτοιος ούτως ώστε η εξομάλυνση να μην δημιουργεί κενές τιμές για καμία από τις τρέχουσες παρατηρήσεις, και αφετέρου η μέθοδος πρόβλεψης να είναι τέτοια ούτως ώστε οι προεκτάσεις να είναι αντιπροσωπευτικές. Για παράδειγμα, η χρήση του μοντέλου της απλής εκθετικής εξομάλυνσης κρίνεται άστοχη επιλογή για την προέκταση μίας χρονοσειράς με εμφανή τάση. Η επιλογή της καταλληλότερης μεθόδου προέκτασης στα πλαίσια της τεχνικής *backcasting* εμπίπτει φυσικά στο γενικότερο πρόβλημα επιλογής μεθόδων πρόβλεψης και θα πρέπει να αντιμετωπίζεται αναλόγως. Στην πράξη βέβαια, προκειμένου να μην επιβαρύνεται με πρόσθετο υπολογιστικό κόστος η όλη διαδικασία πρόβλεψης, προτείνεται η χρήση όσο το δυνατόν πιο απλών μεθόδων. Άλλωστε, δεδομένου ότι η εξομάλυνση βασίζεται πάντοτε πέραν των προβλέψεων που παράγονται και στα πραγματικά δεδομένα, η επίδραση της μεθόδου πρόβλεψης γίνεται αισθητή κυρίως στις τελευταίες παρατηρήσεις της χρονοσειράς όπου και ο μέσος όρος εκτιμάται κατά πλειοψηφία από τις προβλέψεις, και μάλιστα για μικρές τάξεις ΚΜΟ.

Ένα παράδειγμα εφαρμογής της τεχνικής *backcasting* φαίνεται στην [Εικόνα 4.4](#) όπου ο ΚΚΜΟ(12) εφαρμόζεται για την εξομάλυνση μίας χρονοσειράς. Στην περίπτωση μη εφαρμογής της τεχνικής *backcasting* η εξομάλυνση θα οδηγούσε στην εισαγωγή έξι κενών τιμών στην αρχή και στο τέλος της χρονοσειράς δυσχεραίνοντας τη διαδικασία της πρόβλεψης. Όπως φαίνεται, προεκτείνοντας κατάλληλα τα αρχικά δεδομένα με χρήση του μοντέλου *damped* το εν λόγω πρόβλημα έχει εξαλειφθεί.



Εικόνα 4.4: Εφαρμογή της τεχνικής backcasting για τη συμπλήρωση των κενών τιμών κατά την εφαρμογή ενός $KKMO(12)$.

Επίδραση Κινητών Μέσων Όρων στην ακρίβεια πρόβλεψης: Προκειμένου να ερευνηθεί η επίδραση των KMO στην προβλεπτική ακρίβεια των μοντέλων πρόβλεψης, κρίνεται σκόπιμη η μελέτη των παραγόμενων σφαλμάτων για ποικίλα επίπεδα συνάθροισης και χρήση διαφορετικών μεθόδων πρόβλεψης κατά την εφαρμογή της τεχνικής *backcasting*. Στον [Πίνακα 4.2](#) παρουσιάζεται λοιπόν η απόδοση βάσει *sMAPE* πέντε κλασικών μοντέλων πρόβλεψης για εφαρμογή $KMO(3)$, $KMO(5)$, $KMO(12)$, $KMO(3 \times 3)$ και $KMO(3 \times 5)$ στα δεδομένα του διαγωνισμού $M3$ (3003 χρονοσειρές). Για την εφαρμογή της τεχνικής *backcasting* χρησιμοποιείται αρχικά το μοντέλο πρόβλεψης *Damped*, ενώ στη συνέχεια στον [Πίνακα 4.3](#) ελέγχεται η απόδοση και άλλων μοντέλων στην περίπτωση της εφαρμογής απλού $KMO(12)$.

Όπως φαίνεται, όσο αυξάνεται το επίπεδο συνάθροισης, τόσο επωφελούνται οι μέθοδοι πρόβλεψης για δεδομένα υψηλής συχνότητας (μηνιαία), ενώ για δεδομένα χαμηλότερων συχνοτήτων (τριμηνιαία) αρκεί ένας KMO μικρότερης τάξης για τη βελτίωση της απόδοσής τους. Μάλιστα, για τα μη περιοδικά δεδομένα (ετήσια και λοιπά) φαίνεται πως η επιλογή να μην εξομαλυνθεί η χρονοσειρά είναι στη γενικότερη περίπτωση η βέλτιστη λύση. Το εν λόγω φαινόμενο οφείλεται προφανώς στην αύξηση της τυχαιότητας στα δεδομένα υψηλότερων συχνοτήτων η οποία και περιορίζεται σημαντικά όταν αθροιστούν κατάλληλα οι γειτονικές παρατηρήσεις μεταξύ τους. Αντίθετα, η υπερβολική εξομάλυνση χρονοσειρών που χαρακτηρίζονται ήδη από ομαλότητα οδηγεί σε αλλοίωση του σήματος και αποβολή χρήσιμης πληροφορίας, κάτι που φυσικά έχει αρνητικές επιπτώσεις στην προβλεπτική ικανότητα των μεθόδων.

Σύμφωνα με τα αποτελέσματα η συχνότητα της χρονοσειράς θα μπορούσε λοιπόν να αποτελέσει σημαντικό δείγμα της τάξης συνάθροισης που απαιτείται για τη βελτίωση της προβλεπτικής ακρίβειας των μεθόδων, δηλαδή 12 για μηνιαίες, 3 για τριμηνιαίες και απουσία εξομάλυνσης για ετήσιες. Φυσικά η εφαρμογή του παραπάνω λογικού κανόνα δεν εγκλείατε σε καμία περίπτωση το βέλτιστο δυνατό αποτέλεσμα για κάθε μοντέλο πρόβλεψης καθώς μεγιστοποιεί στο σύνολο του διαγωνισμού την ακρίβεια μόνο των μοντέλων *SES* και *Naïve*. Μπορεί ωστόσο να θεωρηθεί μία σχετικά αποδοτική προσέγγιση δεδομένου ότι στις υπόλοιπες περιπτώσεις μοντέλων πρόβλεψης η ακρίβεια που επιτυγχάνεται είναι παραπλήσια της βέλτιστης.

Αξίζει να σημειωθεί ότι για επιλογή υψηλού επιπέδου συνάθροισης, τα αποτελέσματα της *SES* ταυτίζονται με αυτά της *Naïve*, γεγονός που σημαίνει ότι η παράμετρος εξομάλυνσης της πρώτης ταυτίζεται με το μηδέν. Αυτό είναι αρκετά λογικό αν σκεφτεί κανείς ότι στην περίπτωση πλήρως εξομαλυμένων χρονοσειρών (π.χ. *KMO(12)* ή *KMO(3x3)*) δεν υφίστανται συστηματικές αλλαγές επιπέδου και συνεπώς το τρέχον επίπεδο της χρονοσειράς προσδιορίζεται βέλτιστα από την τελευταία διαθέσιμη παρατήρηση.

Πίνακας 4.2: Απόδοση κλασικών μεθόδων πρόβλεψης χρονοσειρών βάσει *sMAPE* στα δεδομένα του διαγωνισμού *M3* για χρήση διαφόρων μορφών *KMO* – εφαρμογή *backcasting* μέσω *damped*.

Μέθοδος πρόβλεψης	ΚΜΟ(1) – Απουσία εξομάλυνσης				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naïve	17.88	9.97	16.71	6.30	14.66
SES	17.77	9.79	14.26	6.28	13.43
Holt	17.18	9.74	13.95	4.29	13.02
Damped	19.15	11.23	15.82	4.67	14.73
Theta	16.73	9.29	13.87	4.92	12.81

Μέθοδος πρόβλεψης	ΚΜΟ(3)				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naïve	18.17	9.55	14.80	6.53	13.72
SES	18.17	9.55	14.81	6.53	13.73
Holt	17.96	10.45	15.15	4.67	13.96
Damped	19.59	12.45	17.44	5.05	15.93
Theta	17.21	9.10	14.45	5.18	13.16

Μέθοδος πρόβλεψης	ΚΜΟ(5)				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naïve	18.48	9.58	14.49	6.67	13.66
SES	18.48	9.58	14.48	6.67	13.65
Holt	17.69	10.07	14.83	4.56	13.65
Damped	19.26	11.54	17.46	4.92	15.63
Theta	18.48	9.58	14.49	6.67	13.66

ΚΜΟ(12)

<i>Μέθοδος πρόβλεψης</i>	<i>Ετήσια</i>	<i>Τριμηνιαία</i>	<i>Μηνιαία</i>	<i>Λοιπά</i>	<i>Σύνολο</i>
Naïve	19.16	9.79	13.94	6.75	13.60
SES	19.16	9.79	13.94	6.75	13.60
Holt	17.85	9.70	14.12	4.49	13.25
Damped	18.15	9.95	15.62	4.56	14.10
Theta	17.97	9.32	13.76	5.40	13.06

<i>Μέθοδος πρόβλεψης</i>	<i>ΚΜΟ(3x3)</i>				
	<i>Ετήσια</i>	<i>Τριμηνιαία</i>	<i>Μηνιαία</i>	<i>Λοιπά</i>	<i>Σύνολο</i>
Naïve	18.23	9.55	14.67	6.57	13.68
SES	18.23	9.55	14.67	6.57	13.68
Holt	17.60	10.77	15.83	4.41	14.27
Damped	19.38	12.64	20.06	5.09	17.18
Theta	17.25	9.10	14.33	5.23	13.11

<i>Μέθοδος πρόβλεψης</i>	<i>ΚΜΟ(3x5)</i>				
	<i>Ετήσια</i>	<i>Τριμηνιαία</i>	<i>Μηνιαία</i>	<i>Λοιπά</i>	<i>Σύνολο</i>
Naïve	18.48	9.58	14.40	6.68	13.61
SES	18.48	9.58	14.40	6.69	13.61
Holt	17.46	10.38	15.26	4.51	13.88
Damped	18.46	11.32	19.07	4.84	16.16
Theta	17.49	9.13	14.08	5.34	13.06

<i>Μέθοδος πρόβλεψης</i>	<i>ΚΜΟ τάξης ίσης με τη συχνότητα των δεδομένων</i>				
	<i>Ετήσια</i>	<i>Τριμηνιαία</i>	<i>Μηνιαία</i>	<i>Λοιπά</i>	<i>Σύνολο</i>
Naïve	17.88	9.75	13.99	6.30	13.31
SES	17.77	9.75	13.99	6.28	13.29
Holt	17.18	9.87	13.90	4.29	13.03
Damped	19.15	11.11	15.15	4.67	14.38
Theta	16.73	9.31	13.88	4.92	12.83

Ένα άλλο συμπέρασμα που εξάγεται από τα αποτελέσματα είναι πως η χρήση διπλών *ΚΜΟ* δεν προσδίδει κάτι περαιτέρω στην προβλεπτική ακρίβεια των μεθόδων πρόβλεψης σε σχέση με τους απλούς *ΚΜΟ*, γεγονός που μπορεί να οφείλεται τόσο στη δημιουργία επιπλέον κενών τιμών, όσο και στην υπερβολική εξομάλυνση των χρονοσειρών βάσει της συχνότητάς τους: Κοινώς ο θόρυβος εξαλείφεται με παρόμοιο τρόπο από αρκετά πιο απλοϊκές μεθόδους και η χρήση διπλών *ΚΜΟ* είναι πρακτικά άσκοπη.

Όσον αφορά τη μέθοδο προέκτασης των χρονοσειρών για χρήση της τεχνικής *backcasting*, οι μέθοδοι *Holt*, *Damped* και *Theta*, φαίνεται να είναι οι πλέον αποδοτικές, πιθανότατα λόγω της ικανότητάς τους να αποτυπώνουν αποτελεσματικά την τάση των δεδομένων. Μάλιστα, οι διαφοροποιήσεις στην απόδοση είναι αρκετά σημαντικές, φαινόμενο που αποδεικνύει πόση καίρια είναι η επίδραση της τεχνικής *backcasting* και

Κεφάλαιο 4: Εξομάλυνση χρονοσειρών

της προσοχής που θα πρέπει να δίνεται κατά την εφαρμογή της. Ενδεικτικά αναφέρεται για το παράδειγμα που παρουσιάστηκε πως η διακύμανση του σφάλματος για χρήση της ίδιας μεθόδου πρόβλεψης και διαφορετικών για την εφαρμογή του backcasting κυμαίνεται από 1.6% έως 2.3%. Έτσι, η επιλογή του βέλτιστου μοντέλου πρόβλεψης κατά το backcasting καταλήγει να είναι εξίσου σημαντική με την επιλογή του βαθμού εξομάλυνσης.

Πίνακας 4.3: Απόδοση κλασικών μεθόδων πρόβλεψης χρονοσειρών βάσει sMAPE στα δεδομένα του διαγωνισμού M3 για εφαρμογή ΚΜΟ(3) και χρήση διαφορετικών μοντέλων πρόβλεψης για εφαρμογή της τεχνικής backcasting.

Μέθοδος πρόβλεψης	Naive				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naive	18.68	9.75	15.70	6.73	14.32
SES	18.68	9.75	15.66	6.73	14.30
Holt	17.29	10.46	16.29	5.04	14.38
Damped	18.82	12.25	18.56	5.09	16.25
Theta	17.47	9.26	15.30	5.33	13.67

Μέθοδος πρόβλεψης	SES				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naive	18.64	9.74	14.96	6.73	13.96
SES	18.65	9.74	14.95	6.73	13.96
Holt	17.17	10.29	15.25	5.03	13.82
Damped	18.43	12.04	17.26	5.09	15.49
Theta	17.44	9.25	14.54	5.33	13.30

Μέθοδος πρόβλεψης	Holt				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naive	18.12	9.50	14.78	6.48	13.69
SES	18.12	9.50	14.77	6.48	13.68
Holt	18.23	10.56	15.21	4.72	14.08
Damped	19.99	12.67	17.55	5.16	16.13
Theta	17.19	9.08	14.51	5.16	13.18

Μέθοδος πρόβλεψης	Damped				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naive	18.17	9.55	14.80	6.53	13.72
SES	18.17	9.55	14.81	6.53	13.73
Holt	17.96	10.45	15.15	4.67	13.96
Damped	19.59	12.45	17.44	5.05	15.93
Theta	17.21	9.10	14.45	5.18	13.16

Μέθοδος πρόβλεψης	Theta				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο

Naïve	18.44	9.69	14.83	6.62	13.84
SES	18.44	9.69	14.83	6.62	13.84
Holt	17.02	10.05	15.14	4.78	13.66
Damped	18.30	11.87	17.10	4.95	15.34
Theta	17.4	9.25	14.48	5.26	13.26

Η επίπτωση των σταθμισμένων *KMO* θα μελετηθεί αναλυτικά στην επόμενη παράγραφο αφού μπορούν να εκφραστούν πολύ πιο ολοκληρωμένα μέσω συναρτήσεων *kernel*.

4.3 Εξομάλυνση χρονοσειρών με χρήση Kernel

Η εξομάλυνση μέσω *kernel* (*Wand & Jones, 1995*) είναι μία τεχνική η οποία δημιουργήθηκε για τη μελέτη πολύπλοκων συναρτήσεων, όπως είναι αυτές της πυκνότητας πιθανότητας και της παλινδρόμησης μεταξύ μεταβλητών. Το πλεονέκτημα της εν λόγω προσέγγισης είναι ότι αποτελεί μη παραμετρική μέθοδο και συνεπώς μπορεί να αναδείξει άμεσα πιθανά μοτίβα που κρύβονται στα δεδομένα και τα οποία δεν είναι εύκολο να εντοπιστούν από παραμετρικά μοντέλα. Κατ' επέκταση λοιπόν η εξομάλυνση μέσω *kernel* μπορεί να χρησιμοποιηθεί για την ανάδειξη χαρακτηριστικών σε χρονοσειρές που έχουν ατονήσει λόγω της ύπαρξης θορύβου.

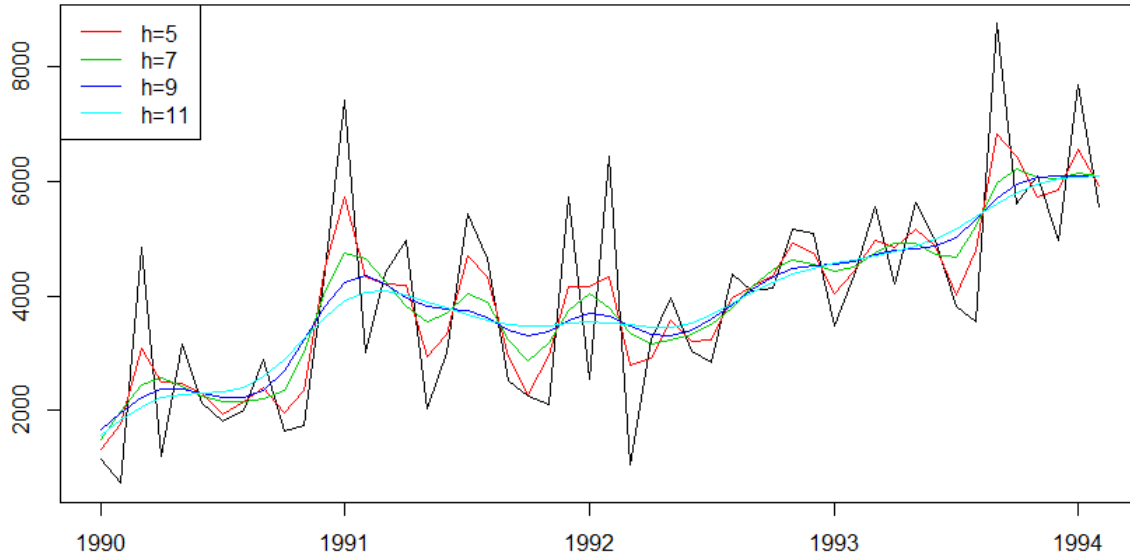
Δεδομένης λοιπόν μίας χρονοσειράς Y_1, Y_2, \dots, Y_n με συνεχή συνάρτηση f να περιγράφει την κατανομή της, η εκτιμήτρια πυκνότητας πιθανότητας *kernel* υπολογίζεται ως:

$$\hat{f}(x, h) = \frac{1}{nh} \sum_{i=1}^n K \frac{x - X_i}{h}$$

όπου K είναι η συνάρτηση *kernel* (πυρήνας) που αξιοποιείται για την εκτίμηση της κατανομής και h το εύρος της (*bandwidth*). Η συνάρτηση K μπορεί να είναι οποιαδήποτε συνάρτηση ωστόσο συνήθως επιλέγεται έτσι ώστε η καμπύλη της να είναι ομοιόμορφη και συμμετρική γύρω από το μηδέν. Το κέντρο της συνάρτησης τοποθετείται κάθε φορά στην τρέχουσα παρατήρηση και η επίδραση των γειτονικών παρατηρήσεων, το πλήθος των οποίων και καθορίζεται μέσω του εύρους της, υπολογίζεται βάσει της εξίσωσης που την περιγράφει. Έτσι, αν και για την εξομάλυνση κάθε σημείου της χρονοσειράς μπορεί να λαμβάνονται υπόψη αρκετές γειτονικές παρατηρήσεις, δύναται να προσδώσουμε μεγαλύτερη έμφαση σε αυτές που επιθυμούμε.

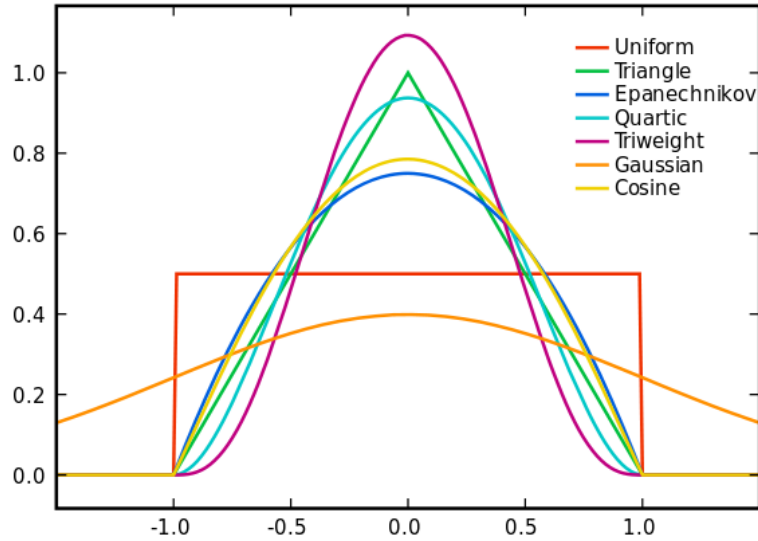
Από τα παραπάνω προκύπτει λοιπόν πως η εξομάλυνση μέσω *Kernel* θυμίζει σημαντικά την εξομάλυνση με χρήση σταθμισμένου *KMO* όπου η συνάρτηση K αξιοποιείται για τον καθορισμό των βαρών των γειτονικών παρατηρήσεων. Κατ' αντιστοιχία λοιπόν με το σταθμισμένο *KMO* όσο αυξάνει το εύρος του *kernel* τόσο βελτιώνεται το αποτέλεσμα της εξομάλυνσης, χωρίς βέβαια αυτό να σημαίνει ότι η εξομάλυνση που οδηγεί σε καλύτερη προβλεπτική ακρίβεια είναι και η πλέον εκτενής. Βιβλιογραφικά η επιλογή του εύρους θεωρείται ιδιαίτερα σημαντική διαδικασία (σημαντικότερη από τον καθορισμό της συνάρτησης K) και θα πρέπει να γίνεται με τέτοιο τρόπο ούτως ώστε να εξασφαλίζεται αφενός επαρκής εξομάλυνση και αφετέρου καλή προσαρμογή στα αρχικά δεδομένα. Ένα

χαρακτηριστικό παράδειγμα παρουσιάζεται στην *Εικόνα 4.5* όπου ο ίδιος πυρήνας χρησιμοποιείται για την εξομάλυνση μιας χρονοσειράς με διαφορετικό κάθε φορά εύρος (5, 7, 9, και 11). Όπως φαίνεται, η χρήση μικρού εύρους οδηγεί σε ανεπαρκή εξομάλυνση (*undersmoothing*) που είναι ανίκανη να απομακρύνει αποτελεσματικά το θόρυβο από τα δεδομένα, ενώ αντίστοιχα η χρήση μεγάλου εύρους σε υπερβολική εξομάλυνση (*oversmoothing*) που έχει ως αποτέλεσμα να χαθεί αρκετή χρήσιμη πληροφορία.



Εικόνα 4.5: Εξομάλυνση μηνιαίας χρονοσειράς μέσω kernel με χρήση κοινού πυρήνα αλλά διαφορετικού εύρους (5, 7, 9 και 11).

Συνάρτηση kernel: Όσον αφορά τη συνάρτηση *kernel*, όπως αναφέρθηκε και νωρίτερα αυτή μπορεί θεωρητικά να υλοποιηθεί από οποιαδήποτε καμπύλη, ωστόσο στην πράξη χρησιμοποιούμε ομοιόμορφες με κέντρο το μηδέν. Η εν λόγω επιλογή έχει ως σκοπό οι παρατηρήσεις συμμετρικά της τρέχουσας να συμμετέχουν με το ίδιο βάρος στον καθορισμό της νέας της τιμής και ταυτόχρονα αυτές που βρίσκονται πιο κοντά της να έχουν μεγαλύτερη επίδραση. Χαρακτηριστικά παραδείγματα κλασικών συναρτήσεων δίνονται στην *Εικόνα 4.6*. Π.χ. η ομοιόμορφη κατανομή (*Uniform*) προσδίδει σε όλες τις γειτονικές παρατηρήσεις το ίδιο βάρος (απλός *KMO*) ενώ η κανονική (*Gauss*) δίνει μεγαλύτερο βάρος στις κοντινότερες το οποίο μάλιστα και φθίνει σημαντικά καθώς απομακρυνόμαστε από την τρέχουσα παρατήρηση.



Εικόνα 4.6: Καμπύλες βασικών συναρτήσεων kernel.

Οι εν λόγω καμπύλες υλοποιούνται μέσω των παρακάτω συναρτήσεων, όπου $\mathbf{1}$ ισούται με μηδέν όταν $|u| \leq 1$ και με μηδέν σε αντίθετη περίπτωση. Η επίδρασή τους παρουσιάζεται ενδεικτικά για εξομάλυνση μίας μηνιαίας χρονοσειρά με εύρος 11 στην [Εικόνα 4.7](#).

$$K(u) = \frac{1}{2} \mathbf{1}\{|u| \leq 1\}, \textit{Uniform}$$

$$K(u) = (1 - |u|) \mathbf{1}\{|u| \leq 1\}, \textit{Triangular}$$

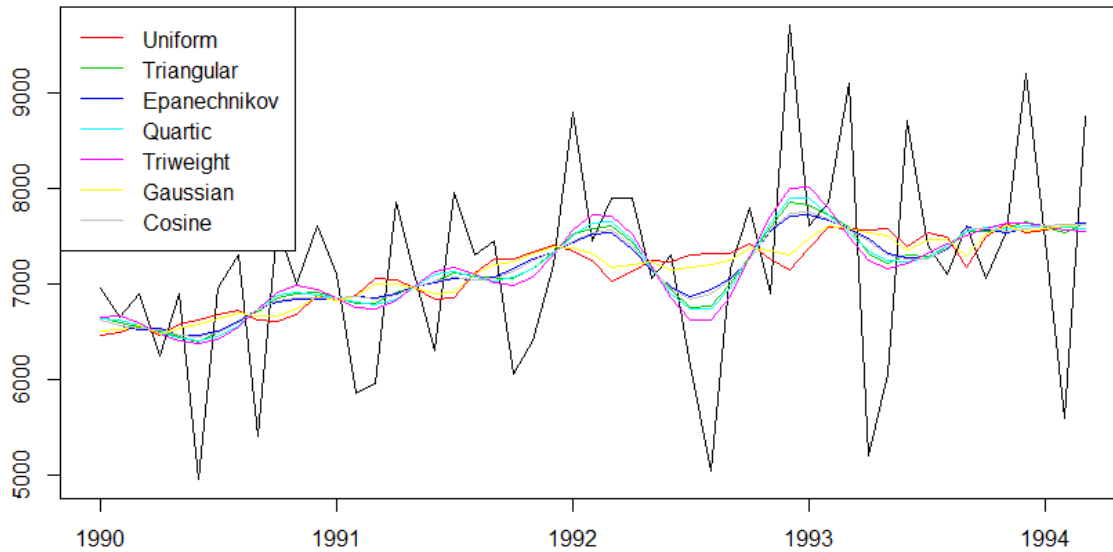
$$K(u) = \frac{3}{4} (1 - u^2) \mathbf{1}\{|u| \leq 1\}, \textit{Epanechnikov}$$

$$K(u) = \frac{15}{16} (1 - u^2)^2 \mathbf{1}\{|u| \leq 1\}, \textit{Quartic (biweight)}$$

$$K(u) = \frac{35}{32} (1 - u^2)^3 \mathbf{1}\{|u| \leq 1\}, \textit{Triweight}$$

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \mathbf{1}\{|u| \leq 1\}, \textit{Gaussian}$$

$$K(u) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}u\right) \mathbf{1}\{|u| \leq 1\}, \textit{Cosine}$$



Εικόνα 4.7: Εφαρμογή εξομάλυνσης μέσω kernel εύρους 11 σε μηνιαία χρονοσειρά για χρήση διαφορετικών πυρήνων.

Θεωρητικά, η καμπύλη *Epanechnikov* (1969) είναι εκείνη για την οποία παράγονται τα βέλτιστα βάρη συμμετοχής για κάθε παρατήρηση, οπότε και η απόδοση όλων των υπόλοιπων καμπυλών υπολογίζεται έχοντας ως άξονα αναφοράς την απόδοσή της. Η απόδοση αυτή, η οποία και μετράτε μέσω του *MISE* (*mean integrated squared error*) ή του *AMISE* (*asymptotic MISE*) δίνεται από τον παρακάτω τύπο.

$$\left(\int u^2 K(u) du \right)^{\frac{1}{2}} \int K(u)^2 du$$

ή αλλιώς

$$Efficiency = \frac{\sqrt{\frac{13}{55}}}{\left(\int u^2 K(u) du \right)^{\frac{1}{2}} \int K(u)^2 du} * 100\%$$

οπότε και για τις καμπύλες που αναφέρθηκαν προκύπτουν θεωρητικά οι αποδόσεις του Πίνακα 4.4.

Πίνακας 4.4: Απόδοση συναρτήσεων kernel ως προς τη θεωρητικά βέλτιστη Epanechnikov

Συνάρτηση kernel	Απόδοση (%)
Uniform	92.9
Triangular	98.6
Epanechnikov	100.0
Quartic	99.4
Triweight	98.7
Gaussian	95.1
Cosine	99.9

Εύρος kernel: Δεδομένης της εξέχουσας σημασίας του εύρους στην αποτελεσματικότητα της εξομάλυνσης, η επιλογή της βέλτιστης καμπύλης *kernel* θεωρείται συχνά δευτερευούσης σημασίας. Έτσι, μπορεί κανείς να καθορίζει εξ αρχής την εν λόγω συνάρτηση και δεδομένης αυτής να προσπαθήσει στη συνέχεια να βελτιστοποιήσει την απόδοση της εξομάλυνσης μεταβάλλοντας κατάλληλα το εύρος. Κριτήριο απόδοσης συνεχίζει να είναι η ελαχιστοποίηση του *MISE/AMISE*, ενώ σε περιπτώσεις χρονοσειρών με σημαντικές ανομοιομορφίες προτείνεται πριν την ελαχιστοποίηση του κριτηρίου η εφαρμογή κάποιου μετασχηματισμού προκειμένου οι παρατηρήσεις της να ακολουθούν όσο το δυνατόν την κανονική κατανομή. Σε αυτό το σημείο αναφέρεται πως υπάρχουν αρκετοί αλγόριθμοι για τον καθορισμό του βέλτιστου εύρους, οι σημαντικότεροι εκ των οποίων είναι οι *Unbiased (ή least squares) cross-validation*, *Biased cross-validation (Scott, 1992)*, *Direct plug-in rule* και *Solve-the-equation rule (Sheather & Jones, 1991)* με τους δύο τελευταίους να προτιμώνται στη βιβλιογραφία.

Λόγω της φύσης της, όπως και στην περίπτωση των *KMO*, η εξομάλυνση μέσω *kernel* επηρεάζεται σημαντικά από την έλλειψη δεδομένων στην αρχή και στο τέλος της χρονοσειράς η οποία και καθιστά αδύνατη την εφαρμογή της στα εκεί σημεία. Προκειμένου να λυθεί το εν λόγω πρόβλημα μπορεί κανείς να καταφύγει στην τεχνική *backcasting* ή να εφαρμόσει τη συνάρτηση *kernel* μόνο επί των διαθέσιμων παρατηρήσεων. Φυσικά, στη δεύτερη περίπτωση η τιμή της εξομαλυμένης χρονοσειράς στα δύο της άκρα θα έχει καθοριστεί αποκλειστικά από τις προηγούμενες τιμές της, μειώνοντας έτσι την αποτελεσματικότητα της μεθόδου.

Επίδραση εξομάλυνσης τύπου *kernel* στην ακρίβεια πρόβλεψης: Προκειμένου να ερευνηθεί η επίδραση της εξομάλυνσης μέσω *kernel* στην προβλεπτική ακρίβεια των μοντέλων πρόβλεψης, κρίνεται σκόπιμη η μελέτη των παραγόμενων σφαλμάτων για ποικίλα επίπεδα συνάθροισης, χρήση διαφορετικών μεθόδων πρόβλεψης για εφαρμογή της τεχνικής *backcasting* καθώς και εφαρμογή πληθώρας πυρήνων. Στον *Πίνακα 4.5* παρουσιάζεται λοιπόν αρχικά η απόδοση πέντε κλασικών μοντέλων πρόβλεψης για την εφαρμογή εξομάλυνσης εύρους ίσο με εννέα για χρήση επτά διαφόρων συναρτήσεων *kernel*. Για την εφαρμογή της τεχνικής *backcasting* χρησιμοποιείται αρχικά η μέθοδος πρόβλεψης θ , ενώ στη συνέχεια στον *Πίνακα 4.6* ελέγχεται η απόδοση και άλλων μοντέλων. Τέλος, διατηρώντας σταθερό ως μοντέλο πρόβλεψης στην τεχνική *backcasting* το μοντέλο u και αξιοποιώντας τη συνάρτηση *'Triweight'* ως πυρήνα, εξετάζεται στον *Πίνακα 4.7* η επίδραση του εύρους εξομάλυνσης. Τα δεδομένα που χρησιμοποιούνται είναι αυτά του διαγωνισμού *M3*, ενώ η αξιολόγηση γίνεται με χρήση του δείκτη *sMAPE*.

Πίνακας 4.5: Απόδοση κλασικών μεθόδων πρόβλεψης χρονοσειρών βάσει sMAPE στα εξομαλυμένα δεδομένα του διαγωνισμού M3 μέσω kernel - Εύρος 9, backcasting μέσω Theta και χρήση διαφόρων πυρήνων.

Μέθοδος πρόβλεψης	Uniform				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naïve	20.10	10.13	14.11	7.06	13.99
SES	20.10	10.13	14.11	7.06	13.99
Holt	17.71	9.50	14.20	4.90	13.23

Κεφάλαιο 4: Εξομάλυνση χρονοσειρών

Damped	18.20	9.75	15.92	4.96	14.22
Theta	20.10	10.13	14.11	7.06	13.99

<i>Μέθοδος πρόβλεψης</i>	Triangular				
	<i>Ετήσια</i>	<i>Τριμηνιαία</i>	<i>Μηνιαία</i>	<i>Λοιπά</i>	<i>Σύνολο</i>
Naïve	19.02	9.80	14.49	6.83	13.84
SES	19.02	9.80	14.48	6.83	13.84
Holt	16.47	9.90	15.42	4.49	13.62
Damped	17.22	10.74	19.87	4.71	16.12
Theta	17.87	9.37	14.18	5.47	13.26

<i>Μέθοδος πρόβλεψης</i>	Epanechnikov				
	<i>Ετήσια</i>	<i>Τριμηνιαία</i>	<i>Μηνιαία</i>	<i>Λοιπά</i>	<i>Σύνολο</i>
Naïve	19.25	9.85	14.39	6.91	13.86
SES	19.26	9.85	14.39	6.91	13.86
Holt	16.60	9.62	15.06	4.65	13.42
Damped	17.19	10.27	18.64	4.83	15.42
Theta	18.09	9.43	14.11	5.55	13.29

<i>Μέθοδος πρόβλεψης</i>	Quartic				
	<i>Ετήσια</i>	<i>Τριμηνιαία</i>	<i>Μηνιαία</i>	<i>Λοιπά</i>	<i>Σύνολο</i>
Naïve	19.00	9.79	14.49	6.82	13.83
SES	19.00	9.79	14.49	6.82	13.83
Holt	16.53	9.86	15.25	4.51	13.54
Damped	17.49	10.67	19.32	4.75	15.91
Theta	17.86	9.36	14.18	5.46	13.25

<i>Μέθοδος πρόβλεψης</i>	Triweight				
	<i>Ετήσια</i>	<i>Τριμηνιαία</i>	<i>Μηνιαία</i>	<i>Λοιπά</i>	<i>Σύνολο</i>
Naïve	18.81	9.76	14.58	6.76	13.82
SES	18.81	9.76	14.58	6.76	13.82
Holt	16.56	10.02	15.47	4.47	13.70
Damped	17.69	11.15	19.85	4.87	16.33
Theta	17.69	9.32	14.26	5.40	13.24

<i>Μέθοδος πρόβλεψης</i>	Gaussian				
	<i>Ετήσια</i>	<i>Τριμηνιαία</i>	<i>Μηνιαία</i>	<i>Λοιπά</i>	<i>Σύνολο</i>
Naïve	19.83	10.04	14.17	7.01	13.93
SES	19.83	10.04	14.17	7.01	13.93
Holt	17.18	9.53	14.59	4.81	13.30
Damped	17.62	9.77	16.74	4.91	14.49
Theta	18.64	9.62	13.96	5.66	13.39

<i>Μέθοδος πρόβλεψης</i>	Cosine				
	<i>Ετήσια</i>	<i>Τριμηνιαία</i>	<i>Μηνιαία</i>	<i>Λοιπά</i>	<i>Σύνολο</i>

Naïve	19.20	9.84	14.41	6.89	13.85
SES	19.20	9.84	14.41	6.89	13.85
Holt	16.60	9.68	15.12	4.62	13.46
Damped	17.23	10.33	18.80	4.81	15.52
Theta	18.04	9.41	14.12	5.53	13.28

Ένα πρώτο συμπέρασμα που εξάγεται σχετικά με την εξομάλυνση μέσω *kernel*, είναι ότι ανάλογα με τον πυρήνα που θα επιλεγεί ενδέχεται να υπάρξουν σημαντικές διαφοροποιήσεις στα παραγόμενα σφάλματα πρόβλεψης. Αυτές είναι αρκετά πιο έντονες ανά μέθοδο στην περίπτωση ελέγχου τους σε επίπεδο συχνότητας, ενώ περιορίζονται όταν μελετηθούν στο σύνολο του διαγωνισμού. Σε γενικές γραμμές οι καμπύλες *Uniform* και *Gauss*, οι οποίες προσδίδουν αρκετά ομοιόμορφα βάρη στις συμμετέχουσες παρατηρήσεις, φαίνεται να παράγουν παρεμφερή αποτελέσματα μεταξύ τους τα οποία και είναι σημαντικά διαφοροποιημένα σε σχέση με τις υπόλοιπες καμπύλες. Αυτό συμβαίνει καθώς οι τελευταίες προσδίδουν όπως φαίνεται και από τη μορφή τους μεγαλύτερο βάρος στις πλέον πρόσφατες παρατηρήσεις. Έτσι, τα μοντέλα *Naïve*, *SES* και *Theta* προκύπτει πως είναι προτιμότερο να εφαρμόζονται μέσω ανομοιόμορφης εξομάλυνσης ενώ αντίθετα η ομοιόμορφη εξομάλυνση, δηλαδή η χρήση απλών *KMO*, είναι προτιμότερη επιλογή για την περίπτωση των μοντέλων *Holt* και *Damped*. Σημειώνεται ενδεικτικά πως η εφαρμογή ενός πυρήνα *kernel* μπορεί να μεταβάλει την ακρίβεια πρόβλεψης ενός μοντέλου έναντι της χρήσης ενός απλού *KMO* από 0.5-15%, με τα μοντέλα των *Naïve*, *SES* και *Holt* να είναι τα πιο ευσταθή (μέση μεταβολή 1.5%) και τα μοντέλα των *Theta* και *Damped* να είναι αρκετά πιο ασταθή (μέση μεταβολή 5% και 9.9% αντίστοιχα).

Ένα άλλο συμπέρασμα που παραμένει αναλλοίωτο σε σχέση με τους απλούς *KMO*, είναι η σημασία του μοντέλου πρόβλεψης που χρησιμοποιείται για την εφαρμογή της τεχνικής *backcasting*. Χαρακτηριστικές είναι οι περιπτώσεις των μοντέλων *Damped* και *SES* όπου η μεταβολή της ακρίβειας μπορεί να φτάσει ανάλογα με την προσέγγιση που ακολουθείται τις 4 και 1 μονάδα αντίστοιχα. Μάλιστα, εύκολα εξάγεται το συμπέρασμα ότι αν οι διαφοροποιήσεις είναι σημαντικές για την περίπτωση της προέκτασης της χρονοσειράς στον ίδιο ορίζοντα με χρήση διαφορετικών μοντέλων, οι διαφοροποιήσεις θα είναι ακόμα μεγαλύτερες όταν μεταβάλλεται ταυτόχρονα και το εύρος της εξομάλυνσης. Ενδεικτικά αναφέρεται πως στο παράδειγμα που εξετάστηκε η διακύμανση μεταξύ των σφαλμάτων πρόβλεψης για χρήση της ίδιας μεθόδου πρόβλεψης αλλά διαφορετικών κατά την εφαρμογή του *backcasting* μπορεί να οδηγήσει σε διακυμάνσεις στην ακρίβεια από 2% έως 9%. Έτσι, η επιλογή του βέλτιστου μοντέλου πρόβλεψης κατά το *backcasting* καταλήγει να είναι εξίσου σημαντική με την επιλογή του βαθμού εξομάλυνσης και του πυρήνα του *kernel*.

Πίνακας 4.6: Απόδοση κλασικών μεθόδων πρόβλεψης χρονοσειρών βάσει sMAPE στα εξομαλυμένα δεδομένα του διαγωνισμού M3 μέσω kernel - πυρήνας Triweight, εύρος 9 και backcasting μέσω κλασικών μεθόδων πρόβλεψης.

Μέθοδος πρόβλεψης	Naive				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naive	19.16	9.83	15.48	6.94	14.35
SES	19.16	9.83	15.48	6.94	14.35
Holt	17.16	10.75	20.36	4.87	16.35
Damped	18.53	12.19	27.75	5.16	20.54
Theta	17.81	9.33	15.13	5.53	13.69

Μέθοδος πρόβλεψης	SES				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naive	19.16	9.83	14.73	6.94	14.00
SES	19.16	9.83	14.73	6.94	14.00
Holt	16.81	10.29	15.95	4.85	14.07
Damped	18.05	11.48	20.06	5.12	16.60
Theta	17.80	9.32	14.32	5.53	13.30

Μέθοδος πρόβλεψης	Holt				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naive	18.21	9.49	14.52	6.54	13.58
SES	18.21	9.49	14.52	6.54	13.58
Holt	18.36	10.86	15.97	4.49	14.53
Damped	19.53	12.40	20.73	5.07	17.47
Theta	17.28	9.07	14.30	5.23	13.10

Μέθοδος πρόβλεψης	Damped				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naive	18.32	9.56	14.55	6.61	13.65
SES	18.32	9.56	14.55	6.62	13.65
Holt	17.60	10.60	15.65	4.44	14.15
Damped	18.95	11.95	20.15	4.91	16.95
Theta	17.33	9.10	14.22	5.27	13.08

Μέθοδος πρόβλεψης	Theta				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naive	18.81	9.76	14.58	6.76	13.82
SES	18.81	9.76	14.58	6.76	13.82
Holt	16.56	10.02	15.47	4.47	13.70
Damped	17.69	11.15	19.85	4.87	16.33
Theta	17.69	9.32	14.26	5.40	13.24

Τέλος, όπως και στην περίπτωση των *KMO*, το εύρος της εξομάλυνσης φαίνεται να επηρεάζει και αυτό σημαντικά την προβλεπτική ακρίβεια των μοντέλων, όπου η βέλτιστη τιμή φαίνεται να σχετίζεται και πάλι με τη συχνότητα των δεδομένων. Συνοπτικά θα μπορούσε κανείς να επιλέξει δηλαδή να εξομαλύνει τα δεδομένα του, αν φυσικά υπάρχει διαθέσιμος επαρκής αριθμός παρατηρήσεων, με έναν πυρήνα *kernel* μήκους παρεμφερούς της συχνότητάς τους. Σημειώνεται ωστόσο πως για άλλη μια φορά η εξομάλυνση αποτελεί ικανή αλλά όχι και αναγκαία συνθήκη για την βελτίωση της συνολικής απόδοσης των μοντέλων καθώς και εντοπίζονται συστηματικές βελτιώσεις για τις περιπτώσεις των μοντέλων *Naïve*, *SES* και *Damped*, δεν ισχύει όμως το ίδιο και για τις υπόλοιπες μεθόδους.

Πίνακας 4.7: Απόδοση κλασικών μεθόδων πρόβλεψης χρονοσειρών βάσει *sMAPE* στα εξομαλυμένα δεδομένα του διαγωνισμού *M3* μέσω *kernel* - πυρήνας *Triweight*, *backcasting* μέσω *Theta* και χρήση εύρους 5, 7, 9 και 11.

Μέθοδος πρόβλεψης	Εύρος 5				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naïve	18.17	9.69	15.17	6.51	13.93
SES	18.17	9.69	15.16	6.51	13.93
Holt	16.85	10.85	15.57	5.09	14.05
Damped	18.98	12.91	17.64	4.92	16.00
Theta	17.14	9.22	14.80	5.14	13.34

Μέθοδος πρόβλεψης	Εύρος 7				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naïve	18.50	9.70	14.78	6.65	13.83
SES	18.50	9.70	14.78	6.65	13.83
Holt	16.67	10.33	15.82	4.58	13.97
Damped	18.29	12.22	18.86	5.20	16.27
Theta	17.42	9.26	14.44	5.28	13.24

Μέθοδος πρόβλεψης	Εύρος 9				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naïve	18.81	9.76	14.58	6.76	13.82
SES	18.81	9.76	14.58	6.76	13.82
Holt	16.56	10.02	15.47	4.47	13.70
Damped	17.69	11.15	19.85	4.87	16.33
Theta	17.69	9.32	14.26	5.40	13.24

Μέθοδος πρόβλεψης	Εύρος 11				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naïve	19.11	9.82	14.43	6.85	13.84
SES	19.11	9.82	14.43	6.85	13.84
Holt	16.55	9.72	15.23	4.53	13.51
Damped	17.32	10.45	18.75	4.73	15.54
Theta	17.96	9.39	14.14	5.50	13.26

4.4 Εξομάλυνση χρονοσειρών μέσω συρρίκνωσης τοπικών καμπυλοτήτων

Όπως παρουσιάστηκε νωρίτερα, οι βασικότερες από τις τεχνικές εξομάλυνσης που συναντώνται στη βιβλιογραφία βασίζονται στην ιδέα της σταθμισμένης συνάθροισης γειτονικών παρατηρήσεων με βάρη που προκύπτουν ανάλογα με την έμφαση που θέλουμε να προσδώσουμε στις πλέον κοντινές. Η εν λόγω συνάθροιση έχει ως αποτέλεσμα τον περιορισμό του παράγοντα της τυχαιότητας και συνεπώς την παραγωγή μίας νέας χρονοσειράς με μικρότερη διακύμανση και σταθερότερο επίπεδο. Ο βαθμός εξομάλυνσης και η αποτελεσματικότητα της τεχνικής εξαρτάται από εκεί και πέρα από μία πληθώρα παραγόντων, όπως π.χ. το επίπεδο (*πλήθος παρατηρήσεων που συναθροίζονται*) και τη μέθοδο (*γραμμική ή μη γραμμική*) συνάθροισης.

Το αποτέλεσμα που προκύπτει από τις παραπάνω μεθόδους δίνει λοιπόν πράγματι την εικόνα μίας αποτελεσματικά εξομαλυμένης χρονοσειράς. Διαθέτουν ωστόσο κάποια τρωτά σημεία που οφείλει κανείς να επισημάνει.

Στην γενική τους μορφή, προκειμένου οι τεχνικές εξομάλυνσης να λειάνουν μία χρονοσειρά στο σημείο n , χρειάζεται να συναθροίσουν με κάποια μέθοδο τα σημεία από $n-k$ έως $n+k$, όπου k το επίπεδο συνάθροισης. Όπως γίνεται κατανοητό, αυτό έχει ως αποτέλεσμα η αρχική χρονοσειρά να μην μπορεί να εξομαλυνθεί στα πρώτα και στα τελευταία k σημεία της καθώς τα διαθέσιμα δεδομένα δεν επαρκούν. Το εν λόγω πρόβλημα αντιμετωπίζεται συνήθως με τη χρήση τεχνικών όπως αυτή του *back-casting*, δηλαδή την κατάλληλη προέκταση των δύο άκρων της αρχικής χρονοσειράς με χρήση μοντέλων πρόβλεψης ούτως ώστε να υπάρξουν επαρκή δεδομένα και να καταστεί δυνατός ο μετέπειτα υπολογισμός της εξομαλυμένης χρονοσειράς.

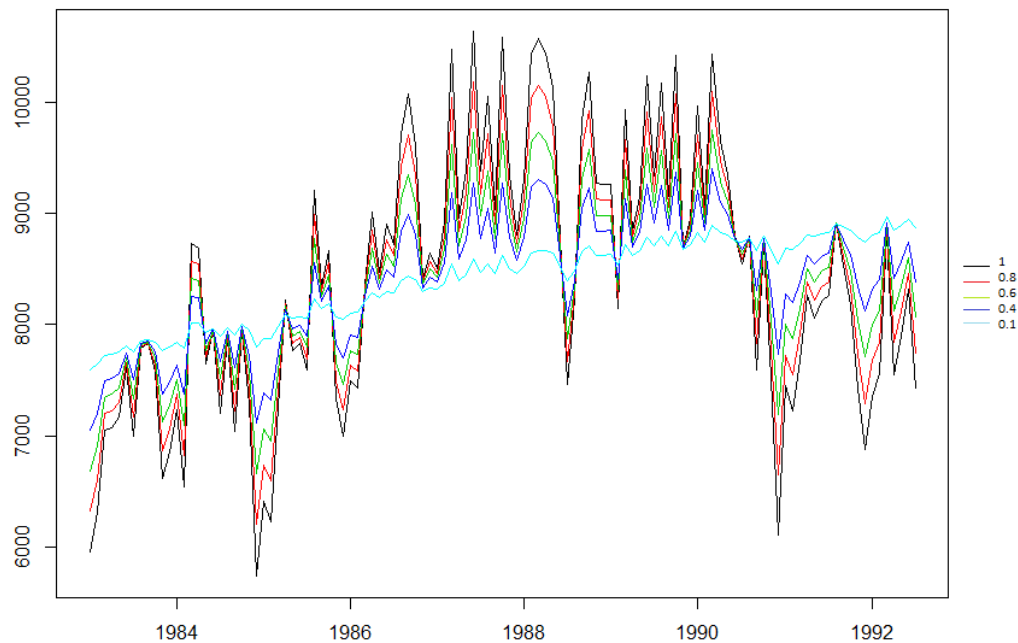
Το εν λόγω πρόβλημα μπορεί να θεωρηθεί αμελητέο για μικρές τιμές επιπέδων k , ωστόσο μπορεί να γίνει καθοριστικό όταν οι τεχνικές εξομάλυνσης εφαρμόζονται επί χρονοσειρών μεγάλης διακύμανσης, όπου και για να υπάρξει ικανοποιητικό αποτέλεσμα απαιτείται συνήθως υψηλό επίπεδο συνάθροισης. Σε τέτοιες περιπτώσεις θα χρειαστεί η αρχική χρονοσειρά να επεκταθεί για μεγάλο χρονικό ορίζοντα οπότε και το αποτέλεσμα που θα προκύψει θα βασίζεται σε σαφώς πιο αβέβαιες προβλέψεις και συνεπώς ενδέχεται να οδηγήσει σε παραπλανητικά αποτελέσματα και σε αλλοίωση των μοτίβων της χρονοσειράς, όπως είναι η τάση και το επίπεδο. Επίσης ερωτήματα όπως το ποια μέθοδος πρόβλεψης θα επιλεγεί για την προέκταση της χρονοσειράς σε αυτά τα σημεία αυξάνουν την πολυπλοκότητα του εν λόγω εγχειρήματος.

Προκειμένου να καταφέρει κανείς λοιπόν να λειάνει αποτελεσματικά μία χρονοσειρά καλείτε να απαντήσει στο ερώτημα κατά πόσο είναι διατεθειμένος να εισάγει αβεβαιότητα στη εξομαλυμένη χρονοσειρά χρησιμοποιώντας μεγάλες προεκτάσεις ούτως ώστε να οδηγηθεί σε μία πιο λεία, και κατά συνέπεια, πιο εύκολα επεκτάσιμη χρονοσειρά. Το εν λόγω ερώτημα απαντάται συνήθως μέσα από διάφορα στατιστικά τεστ ωστόσο, ακόμα και έτσι, και δεδομένου ότι η χρονοσειρά που θα προκύψει θα αποτελέσει τη βάση για την παραγωγή προβλέψεων, κανείς δεν μπορεί να εγγυηθεί για την αποτελεσματικότητα των παραπάνω πρακτικών.

Βάσει των παραπάνω γίνεται κατανοητό ότι ιδιαίτερο ενδιαφέρον θα είχε η δημιουργία μίας νέας τεχνικής εξομάλυνσης η οποία θα ήταν ικανή να παράγει παρεμφερή αποτελέσματα με αυτά των υψηλών επιπέδων συνάθροισης συμβατικών προσεγγίσεων χωρίς να απαιτούνται αντίστοιχες προεκτάσεις και παραμετροποιήσεις.

4.4.1 Ο κλασικός μετασχηματισμός θ

Αρκετά κοντά σε αυτή την κατεύθυνση κινείται ως φιλοσοφία ο μετασχηματισμός θ (Assimakopoulos & Nikolopoulos, 2000) ο οποίος δημιουργήθηκε αρχικά με σκοπό να εντοπίζονται αποτελεσματικά μακροπρόθεσμες τάσεις. Όπως αναφέρθηκε και στην περιγραφή του ομώνυμου μοντέλου πρόβλεψης, ο εν λόγω μετασχηματισμός συρρικνώνει επί της ουσίας τις τοπικές καμπυλότητες της χρονοσειράς για τιμές παραμέτρου στο διάστημα $(0,1)$, παράγοντας μία νέα σαφώς πιο λεία χρονοσειρά με ίδια κλίση και επίπεδο με την αρχική. Με αυτόν τον τρόπο απαλείφονται οι τοπικές διακυμάνσεις της χρονοσειράς και αναδεικνύονται τα χαμηλότερης συχνότητας χαρακτηριστικά της. Πέραν από τη διατήρηση των ροπών πρώτου και δεύτερου βαθμού, ο μετασχηματισμός θ έχει επιπλέον το πλεονέκτημα ότι βασίζεται αποκλειστικά στις πρώτες διαφορές της χρονοσειράς και συνεπώς μπορεί να οδηγήσει σε λείες χρονοσειρές αξιοποιώντας αποκλειστικά τα αρχικά δεδομένα και όχι προεκτάσεις της χρονοσειράς (*backcasting*).



Εικόνα 4.8: Ο μετασχηματισμός θ για τιμές παραμέτρου 1, 0.8, 0.6, 0.4 και 0.1.

Το μειονέκτημα της εν λόγω προσέγγισης έγκειται ωστόσο στο γεγονός ότι επειδή ακριβώς εφαρμόζεται γραμμικά επί των αρχικών δεδομένων, για να οδηγηθεί κανείς σε επαρκώς λείες χρονοσειρές χρειάζεται να χρησιμοποιήσει αρκετά μικρές τιμές θ , αλλοιώνοντας έτσι στη γενική περίπτωση σημαντικά την αρχικά διαθέσιμη πληροφορία. Όπως φαίνεται και στην *Εικόνα 4.8*, καθώς μειώνεται η τιμή της παραμέτρου και η εξομαλυμένη χρονοσειρά

προσεγγίζει την ευθεία της απλής γραμμικής παλινδρόμησης, απαλείφεται όλο και περισσότερο η τυχαιότητα των δεδομένων. Την ίδια στιγμή βέβαια εξαφανίζονται οι ευρύτερες τοπικές καμπυλότητες της χρονοσειράς με αποτέλεσμα η χρονοσειρά που προκύπτει να προσεγγίζει την απόλυτη ευθεία, κάτι που προφανώς προβλεπτικά μπορεί να αποδειχτεί άστοχο. Μάλιστα, για τιμή παραμέτρου θ ίση με 0, η εξομαλυμένη χρονοσειρά ταυτίζεται με αυτήν της απλής γραμμικής παλινδρόμησης επηρεάζοντας σημαντικά το σημείο έναρξης των προβλέψεων.

Συμπερασματικά, όταν χρησιμοποιείται ο μετασχηματισμός θ με τιμές παραμέτρου (0,1), συρρικνώνονται οι τοπικές καμπυλότητες της χρονοσειράς και ταυτόχρονα τα επιμέρους σημεία της προσεγγίζουν ανάλογα με την αρχική τους απόκλιση από την ευθεία της απλής γραμμικής παλινδρόμησης την εν λόγω ευθεία. Έτσι, κάθε επαρκώς εξομαλυμένη χρονοσειρά είναι αναγκασμένη να ακολουθεί ένα καθαρά γραμμικό μοτίβο τάσης, κάτι το οποίο δεν είναι πάντοτε προβλεπτικά επιθυμητό. Σε αυτήν την περίπτωση δεν απαιτείται ωστόσο πέραν της τιμής της παραμέτρου θ να καθοριστεί ούτε το επίπεδο συνάθροισης ούτε το βάρος της κάθε παρατήρησης, ενώ η εξομάλυνση πραγματοποιείται άμεσα χωρίς να δημιουργηθούν κενές τιμές.

4.4.2 Ο μη γραμμικός μετασχηματισμός θ

Μία λύση για τη διατήρηση της λογικής του μετασχηματισμού θ αλλά και τον ταυτόχρονο περιορισμό της εξάλειψης των ευρύτερα τοπικών καμπυλοτήτων της χρονοσειράς, είναι η χρήση του μη γραμμικού μετασχηματισμού θ (Assimakopoulos, 1995). Ο εν λόγω μετασχηματισμός στηρίζεται στην μείωση της καμπυλότητας που υπολογίζεται μεταξύ τριών γειτονικών παρατηρήσεων και πραγματοποιείται με τέτοιο τρόπο ούτως ώστε η ενδιάμεση/τρέχουσα παρατήρηση να προσεγγίσει τη νοητή ευθεία που ορίζουν οι δύο γειτονικές της. Η καμπυλότητα στο σημείο i υπολογίζεται βάσει του τύπου που ακολουθεί και είναι εμφανές ότι αν βρίσκεται πάνω στην εν λόγω ευθεία θα ισούται με μηδέν, ενώ όσο αποκλίνει από αυτήν θα αυξάνεται.

$$\delta_i = X_{i+1} - 2X_i + X_{i-1} \text{ ή}$$

$$\delta_i(\%) = 300 \frac{X_{i+1} - 2X_i + X_{i-1}}{X_{i+1} + X_i + X_{i-1}} * 100\%$$

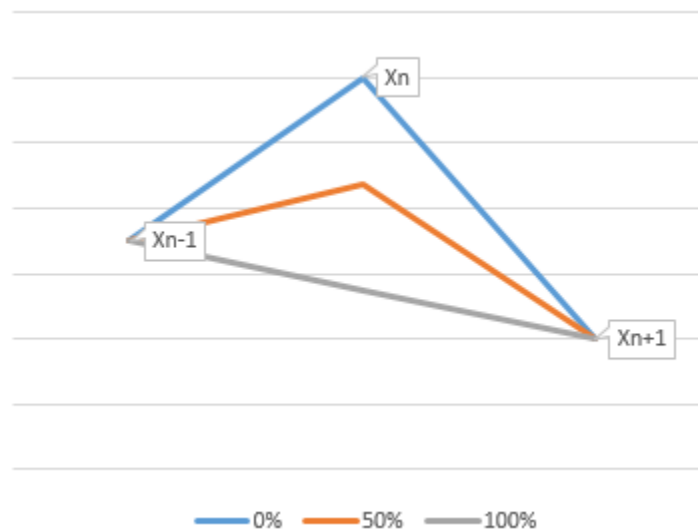
Αν λοιπόν X_i είναι η τρέχουσα παρατήρηση και c είναι μία παράμετρος εξομάλυνσης η οποία ορίζει την ποσοστιαία μεταβολή κατά την οποία επιθυμούμε αυτή να προσεγγίσει την ευθεία που ορίζουν οι δύο γειτονικές της, η νέα της τιμή θα ισούται μετά την εφαρμογή του μετασχηματισμού με

$$X'_i = X_i * (1 - c) + c * \left(\frac{X_{i+1} + X_{i-1}}{2} \right)$$

Ενδεικτικά η εφαρμογή του μετασχηματισμού σε μία τριάδα σημείων παρουσιάζεται για παράμετρο εξομάλυνσης 0, 0.5 και 1 στην *Εικόνα 4.9*.

Σε αντίθεση με τον κλασικό μετασχηματισμό θ , ο εν λόγω μετασχηματισμός δεν εφαρμόζεται ταυτόχρονα σε όλα τα σημεία της χρονοσειράς καθώς είναι προφανές ότι η

μεταβολή ενός σημείου της χρονοσειράς θα επηρεάσει ακαριαία την τοπική καμπυλότητα των δύο γειτονικών του. Έτσι, ο μετασχηματισμός εφαρμόζεται σειριακά από σημείο σε σημείο και με κριτήριο προτεραιότητας το μέγεθος της ισχύουσας καμπυλότητας. Πρακτικά, κατά την εφαρμογή του μετασχηματισμού πραγματοποιείτε αρχικά εξάλειψη των ακραίων τιμών της χρονοσειράς, στη συνέχεια περιορισμός των τοπικών διακυμάνσεων και τέλος απόλυτη ευθυγράμμιση των τιμών της ως προς την πρώτη και την τελευταία παρατήρησή της, τα οποία εξ ορισμού είναι και τα μόνα σημεία που παραμένουν σταθερά κατά την εξομάλυνση (δεν μπορεί να οριστεί τοπική διακύμανση για το πρώτο και το τελευταίο σημείο της χρονοσειράς λόγω έλλειψης γειτονικών σημείων).



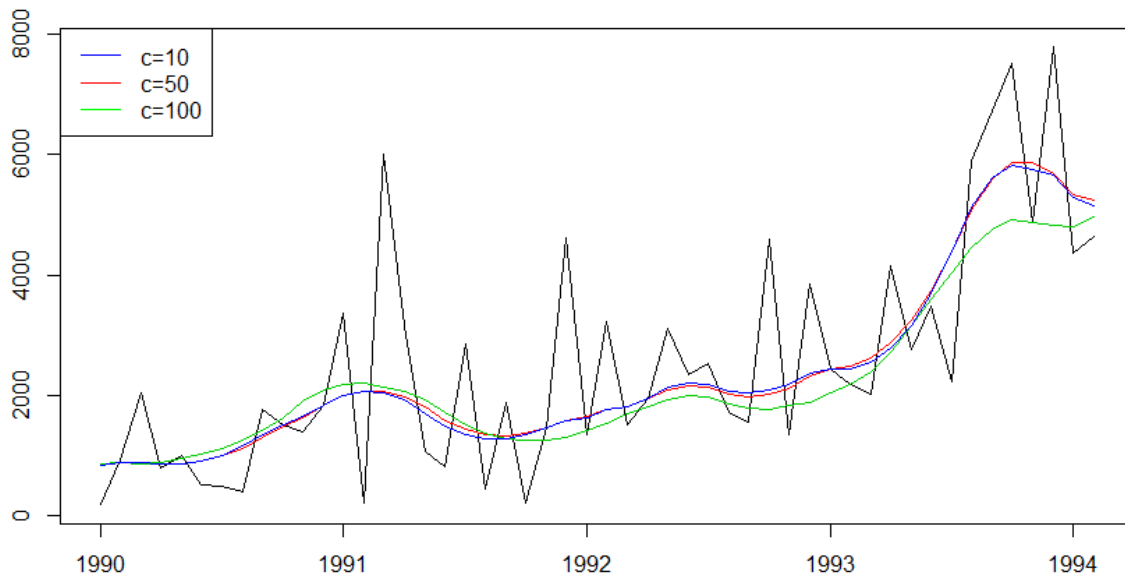
Εικόνα 4.9: Εξομάλυνση τριάδας παρατηρήσεων με χρήση του μη γραμμικού μετασχηματισμού θ για παράμετρο εξομάλυνσης c ίση με 0 (αρχικά δεδομένα), 0.5 και 1.

Πλεονεκτήματα της εν λόγω τεχνικής είναι αφενός η μη γραμμική φύση του μετασχηματισμού (μεταβολή καμπυλότητας ανάλογα με το αρχικό μέγεθός της), η επιλεκτικότητά της (μετασχηματισμός των παρατηρήσεων με υψηλή τοπική διακύμανση) και η δυνατότητα διατήρησης των ευρύτερων καμπυλοτήτων της χρονοσειράς λόγω του συνδυασμού των δύο προηγούμενων χαρακτηριστικών του μετασχηματισμού.

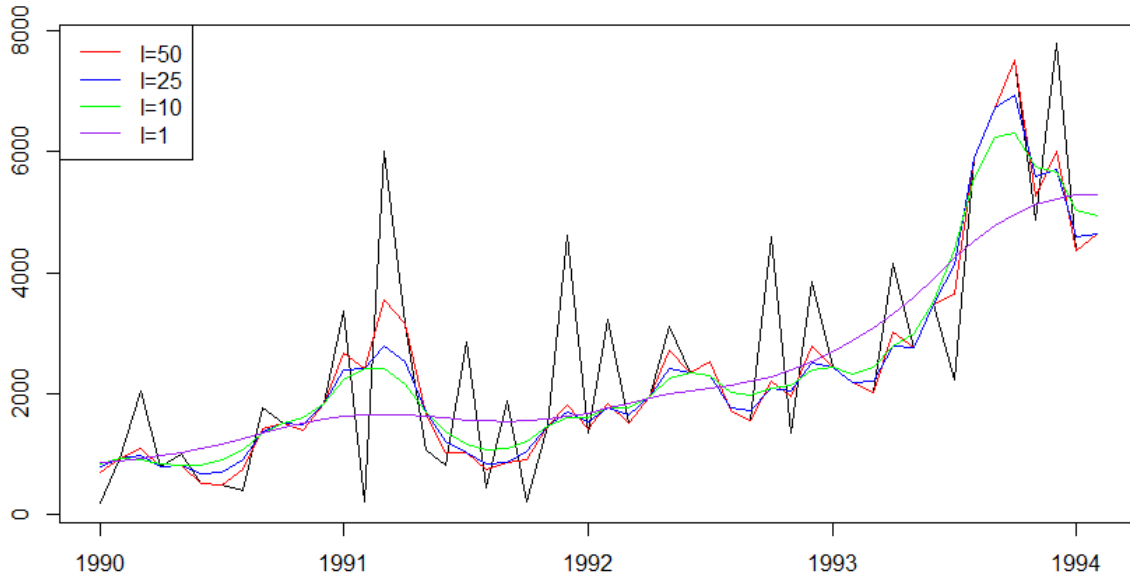
Δεδομένου βέβαια ότι στην πράξη πάντα θα υπάρχουν αποκλίσεις μεταξύ της εξομαλυμένης χρονοσειράς και της ευθείας που ορίζουν τα δύο ακριανά σημεία της, ένα ερώτημα που τίθεται είναι το πότε τερματίζει η εν λόγω διαδικασία, ούτως ώστε αφενός ο χρόνος εκτέλεσης της διαδικασίας να βρίσκεται εντός ανεκτών ορίων και αφετέρου η χρονοσειρά που θα προκύψει να μην είναι μια απόλυτη ευθεία. Για ποια τιμή καμπυλότητας δ μία χρονοσειρά θεωρείται δηλαδή επαρκώς ομαλή. Άλλα ερωτήματα που προκύπτουν σχετίζονται φυσικά με τον προσδιορισμό της παραμέτρου εξομάλυνσης c , αλλά και το κατά πόσο είναι ασφαλές να εξομαλύνει κανείς μία χρονοσειρά βάσει των δύο

ακριανών της σημείων τα οποία μπορούν κάλλιστα να διέπονται από αυξημένη τυχαιότητα.

Μετά από σειρά πειραμάτων που πραγματοποιήθηκαν αποδείχτηκε εμπειρικά ότι για να προκύψει εξομαλυμένη ευθεία με προβλεπτική ανωτερότητα σε σχέση με τα αρχικά δεδομένα, η όλη διαδικασία εξομάλυνσης μπορεί να τερματιστεί όταν όλες οι τοπικές καμπυλότητες γίνουν μικρότερες από 1% με παράμετρο εξομάλυνσης ίση με 0.1. Αυτό συμβαίνει καθώς απότομες μεταβολές (π.χ. ολίσθηση κατά 75% ή 100%) ελλοχεύουν τον κίνδυνο να εξαλειφθούν ακαριαία οι ευρύτερες καμπυλότητες της χρονοσειράς, ενώ ασθενέστερες συνθήκες τερματισμού οδηγούν σε ανομοιόμορφη επεξεργασία των πρωτογενών δεδομένων υψηλής διακύμανσης και κατά συνέπεια στον καθορισμό λανθάνοντος επιπέδου και τάσης. Εποπτικά οι παραπάνω ισχυρισμοί γίνονται αντιληπτοί στην *Εικόνα 4.10* για εφαρμογή της τεχνικής σε μηνιαία χρονοσειρά με παράμετρο εξομάλυνσης ίση με 0.1, 0.5 και 1 μέχρις ότου η καμπυλότητα όλων των σημείων γίνει μικρότερη του 5%, και στην *Εικόνα 4.11* για εφαρμογή της τεχνικής στην ίδια χρονοσειρά με παράμετρο εξομάλυνσης ίση με 0.1 μέχρις ότου η καμπυλότητα όλων των σημείων γίνει μικρότερη του 50%, 25%, 10% και 1% αντίστοιχα.



Εικόνα 4.10: Εξομάλυνση μηνιαίας χρονοσειράς με χρήση του μη γραμμικού μετασχηματισμού θ για παράμετρο εξομάλυνσης c ίση με 0.1, 0.5 και 1 και συνθήκη τερματισμού την επίτευξη διακύμανσης μικρότερης του 5%.



Εικόνα 4.11: Εξομάλυνση μηνιαίας χρονοσειράς με χρήση του μη γραμμικού μετασχηματισμού θ για παράμετρο εξομάλυνσης c ίση με 0.1 και συνθήκη τερματισμού την επίτευξη διακύμανσης μικρότερης του 50%, 25%, 10% και 1%.

Τα πειράματα έδειξαν επίσης ότι η εξομάλυνση βάσει των ακριανών σημείων της χρονοσειράς οδηγεί σε ομαλές χρονοσειρές με μειωμένη προβλεπτική αξία και ότι μέθοδοι που λαμβάνουν υπόψη το τρέχον επίπεδο της χρονοσειράς (όπως π.χ. η μέθοδος θ και η *damped*) θα πρέπει να χρησιμοποιούνται προκειμένου να τα αντικαθιστούν προεκτείνοντάς τες κατά μία παρατήρηση αντίστοιχα (*backcasting*). Μάλιστα, η εν λόγω παράμετρος, δηλαδή η επιλογή της μεθόδου που θα χρησιμοποιηθεί για να καθοριστούν τα σταθερά σημεία της χρονοσειράς κατά τη διαδικασία της εξομάλυνσης, αναδείχτηκε όπως και στην περίπτωση των *KMO* και των *kernel*, ως η πλέον καθοριστική. Σημαντικό πλεονέκτημα έναντι των κλασικών μεθόδων αποτελεί βέβαια αναμφισβήτητα το γεγονός ότι σε αυτήν την περίπτωση η εφαρμογή της τεχνικής *backcasting* απαιτεί πάντα την προέκταση της χρονοσειράς κατά μία και μόνο περίοδο ανεξαρτήτως του βαθμού εξομάλυνσης που επιλέγεται.

4.4.5 Μετασχηματισμός εξισορρόπησης τοπικών διακυμάνσεων

Όπως παρουσιάστηκε στην προηγούμενη παράγραφο η μη γραμμική μορφή του μετασχηματισμού θ παρουσιάζει τρία βασικά πλεονεκτήματα έναντι των υπολοίπων τεχνικών εξομάλυνσης χρονοσειρών που παρουσιάστηκαν και τα οποία συνοψίζονται ως εξής:

- Διατηρούνται οι τοπικές καμπυλότητες της χρονοσειράς ακόμα και για πολύ υψηλά επίπεδα εξομάλυνσης
- Οι διακυμάνσεις της χρονοσειράς περιορίζονται επιλεκτικά ανάλογα με την εκάστοτε τοπική καμπυλότητα που παρατηρείται

- Η τεχνική *backcasting* εφαρμόζεται πάντα για ορίζοντα πρόβλεψης ίσο με μία χρονική περίοδο ανεξάρτητα από το βαθμό εξομάλυνσης
- Η τεχνική βασίζεται σε ένα μετασχηματισμό μη γραμμικής φύσης
- Η εξομάλυνση δεν οδηγεί απαραίτητως σε γραμμικά μοτίβα τάσης

Στα πλαίσια ωστόσο μίας προσπάθειας γενίκευσης και βελτίωσης του μη γραμμικού μετασχηματισμού θ πραγματοποιήθηκε μία σειρά τροποποιήσεων ούτως ώστε αυτός να απαλλαχθεί όσο το δυνατό από τις αυθαίρετες παραμετροποιήσεις που αναφέρθηκαν νωρίτερα, δηλαδή τον καθορισμό του επιπέδου εξομάλυνσης και της συνθήκης τερματισμού. Αυτό έγινε επεκτείνοντας αφενός τον υπολογισμό της τοπικής διακύμανσης σε τετράδες σημείων και θέτοντας ως επιπλέον απαίτηση οι επιμέρους διακυμάνσεις εντός αυτών, δηλαδή οι διακυμάνσεις που ορίζονται για τις δύο τριάδες σημείων εντός της υπό εξέταση τετράδας, να εξισώνονται. Στόχος είναι η παραγωγή μίας λείας χρονοσειράς η οποία αν και τοπικά δεν θα παρουσιάζει διακυμάνσεις, στο σύνολό της θα διατηρεί τα βασικά μοτίβα της αρχικής χρονοσειράς.

4.4.5.1 Η μέθοδος εξομάλυνσης

Κατά αντιστοιχία με την εφαρμογή του μη γραμμικού μετασχηματισμού θ σε τριάδες σημείων, η συνολική διακύμανση μίας τετράδας σημείων μπορεί να οριστεί ως

$$C_i = \delta_i - \delta_{i-1} =$$

$$X_{i+1} - 2X_i + X_{i-1} - X_i + 2X_{i-1} - X_{i-2} =$$

$$X_{i+1} - 3X_i + 3X_{i-1} - X_{i-2}$$

Αν $\delta_i * \delta_{i-1} < 0$ και $|\delta_{i-1}| > |\delta_i|$, δηλαδή οι επιμέρους διακυμάνσεις είναι ετερόσημες και η πρώτη μεγαλύτερη από τη δεύτερη κατά απόλυτη τιμή, τότε προκειμένου αυτές να εξισωθούν θα πρέπει

$$X_{i+1} - 2X_i + X_{i-1}' = -X_i + 2X_{i-1}' - X_{i-2} \rightarrow$$

$$X_{i-1}' = X_{i+1} - X_i + X_{i-2}$$

Δηλαδή απαιτείται το δεύτερο κατά σειρά σημείο να ολισθήσει προς το τρίτο κατά απόσταση ίση με

$$|X_{i-1} - X_{i+1} + X_i - X_{i-2}|$$

Αντίστοιχα, αν $|\delta_{i-1}| < |\delta_i|$ θα έχουμε μετακίνηση του τρίτου κατά σειρά σημείου προς το δεύτερο κατά ίση απόσταση ως εξής

$$X_i' = X_{i+1} - X_{i-1} + X_{i-2}$$

Αν τώρα $\delta_i * \delta_{i-1} > 0$, δηλαδή οι επιμέρους διακυμάνσεις είναι ομόσημες, τότε προκειμένου να εξισωθούν οι τοπικές διακυμάνσεις θα πρέπει αφενός

$$X_{i+1} - 2X_i' + X_{i-1}' = X_i' - 2X_{i-1}' + X_{i-2} \rightarrow$$

$$X_i' - X_{i-1}' = \frac{1}{3}(X_{i+1} - X_{i-2})$$

και αφετέρου προκειμένου να διατηρηθεί το μοτίβο της χρονοσειράς στο υπό εξέταση σημείο

$$X_{i+1} - 2X_i' + X_{i-1}' + X_i' - 2X_{i-1}' + X_{i-2} = X_{i+1} - 2X_i + X_{i-1} + X_i - 2X_{i-1} + X_{i-2} \rightarrow$$

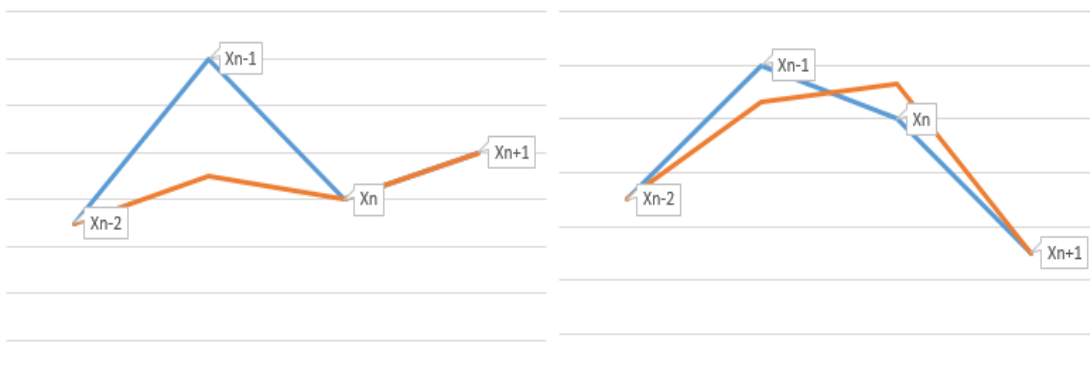
$$X_{i-1}' + X_i' = X_{i-1} + X_i$$

Λύνοντας το σύστημα των δύο εξισώσεων λαμβάνουμε τις νέες τιμές των δύο σημείων τα οποία και θα ολισθήσουν αντιδιαμετρικά και αναλογικά προς την ευθεία που ορίζουν τα δύο ακριανά σημεία της τετράδας ως εξής:

$$X_i' = \frac{1}{2}(X_{i-1} + X_i) + \frac{1}{6}(X_{i+1} - X_{i-2})$$

$$X_{i-1}' = \frac{1}{2}(X_{i-1} + X_i) - \frac{1}{6}(X_{i+1} - X_{i-2})$$

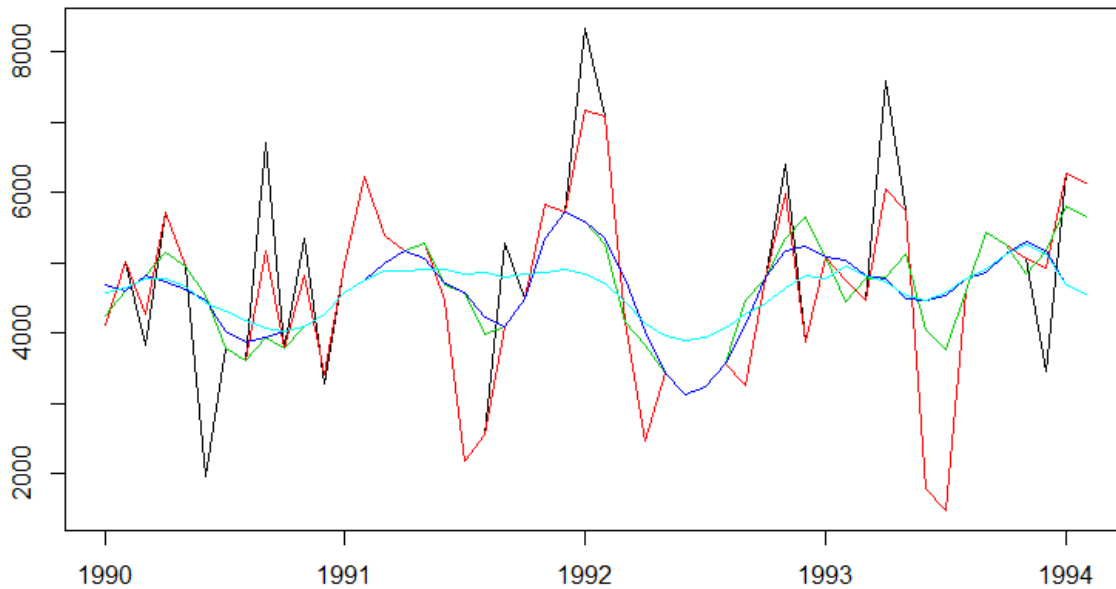
Και σε αυτήν την περίπτωση ο αλγόριθμος εξομάλυνσης εφαρμόζεται σειριακά με προτεραιότητα που ορίζεται μέσω της διακύμανσης των εκάστοτε τετράδων της χρονοσειράς ταξινομημένων σε φθίνουσα σειρά, ενώ για να ολοκληρωθεί η όλη διαδικασία θα πρέπει οι διακυμάνσεις που ορίζονται στο εσωτερικό όλων των τετράδων να είναι διάφορες μεταξύ τους και μεγαλύτερες του μηδενός.



Εικόνα 4.12: Εξομάλυνση μίας σειράς τεσσάρων σημείων σε περίπτωση ετερόσημων (αριστερά) και ομόσημων (δεξιά) τοπικών διακυμάνσεων.

Σημειώνεται επίσης ότι όπως και στην περίπτωση του μη γραμμικού μετασχηματισμού θ , η εν λόγω προσέγγιση δεν οδηγεί απαραίτητα σε απόλυτη ευθεία αλλά σε καμπύλες χρονοσειρές μηδενικών τοπικών διακυμάνσεων. Ενδεικτικά η εφαρμογή του μετασχηματισμού σε μία τετράδα σημείων παρουσιάζεται στην *Εικόνα 4.12*, ενώ η επιλεκτικότητα της τεχνικής αναδεικνύεται στην *Εικόνα 4.13* όπου παρουσιάζεται το

αποτέλεσμα του αλγορίθμου σε μηνιαία χρονοσειρά υψηλής διακύμανσης μετά από 10, 50, 100 και 200 επαναλήψεις.

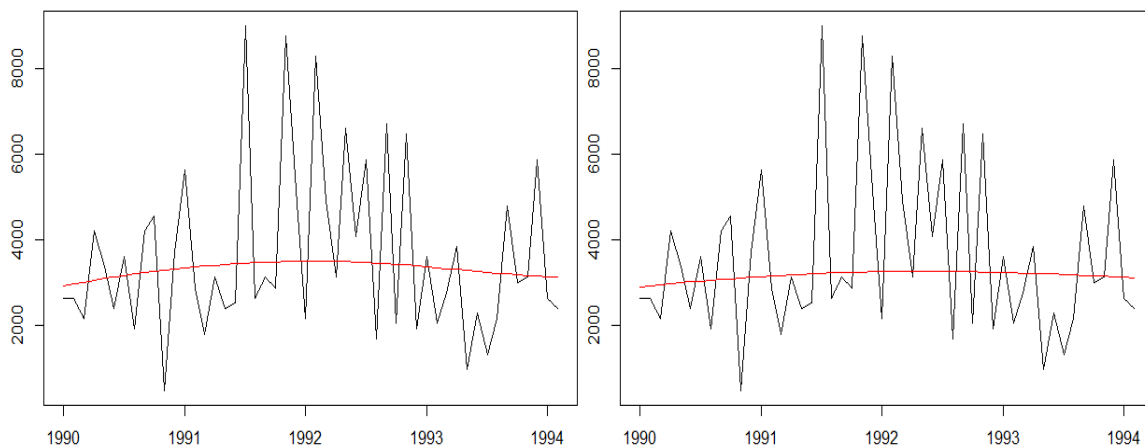


Εικόνα 4.13: Αποτέλεσμα εξομάλυνσης μηνιαίας χρονοσειράς υψηλής διακύμανσης (μαύρο) με χρήση της τεχνικής εξισορρόπησης τοπικών διακυμάνσεων μετά από 10 (κόκκινο), 50 (μπλε), 100 (πράσινο) και 200 (γαλάζιο) επαναλήψεις.

Φυσικά και εδώ το υπολογιστικό κόστος αποτελεί μελανό σημείο για τη διαδικασία καθώς, όπως γίνεται αντιληπτό, από ένα σημείο εξομάλυνσης και μετά οι μεταβολές μεταξύ των σημείων είναι τόσο μικρές που σε συνδυασμό με την έλλειψη απόλυτης υπολογιστικής ακρίβειας καθιστούν τη συνθήκη τερματισμού πρακτικά αδύνατη. Προκειμένου λοιπόν ο αλγόριθμος να τερματίζει σε ικανοποιητικούς χρόνους μπορεί κανείς να αντικαταστήσει στις ισότητες των σχέσεων το απόλυτο μηδέν με κάποια άλλη μικρή τιμή που παράγει ισάξιο αποτέλεσμα. Για να γίνει αυτό με επιτυχία και να μην εμπλακεί η όλη διαδικασία με παραμετροποιήσεις, προτείνεται η διαίρεση των επιμέρους διακυμάνσεων με τη μέση τιμή της χρονοσειράς και η χρήση ως εναλλακτικής συνθήκης τερματισμού η απαίτηση η διαφορά μεταξύ των διακυμάνσεων να είναι μικρότερη από 0.01%, δηλαδή πρακτικά μηδενικά.

Μετά από πειράματα που πραγματοποιήθηκαν αποδείχτηκε πως η εν λόγω συνθήκη είναι ιδιαίτερα αποτελεσματική επιλογή καθώς παράγει σχεδόν πανομοιότητα αποτελέσματα με αυτά αυστηρότερων συνθηκών (όπως π.χ. της 0.001%) που προσεγγίζουν το απόλυτο μηδέν, σε ασύγκριτα όμως μικρότερο χρόνο. Ο εν λόγω ισχυρισμός δεν αναφέρεται μόνο στο οπτικό αποτέλεσμα που προκύπτει αλλά και στην προβλεπτική ικανότητα κλασικών μεθόδων πρόβλεψης όταν εφαρμόζονται επί των εξομαλυμένων χρονοσειρών. Χαρακτηριστικά αποτελέσματα παρουσιάζονται στην [Εικόνα 4.14](#) και στον [Πίνακα 4.8](#), όπου οι μέθοδοι *Naive*, *SES*, *Holt*, *Damped* και θ εφαρμόζονται στην πρώτη περίπτωση

στα εξομαλυμένα δεδομένα του διαγωνισμού *M3* με συνθήκη τερματισμού 0.01% και 0.001% αντίστοιχα. Για την εφαρμογή της τεχνικής *backcasting* χρησιμοποιείται το μοντέλο *Damped*.



Εικόνα 4.14: Εξομάλυνση μέσω της τεχνικής εξισορρόπησης τοπικών διακυμάνσεων με συνθήκη τερματισμού 0.01% (αριστερά) και 0.001% σε μηνιαία χρονοσειρά του διαγωνισμού *M3*.

Όπως φαίνεται, οι εξομαλυμένες χρονοσειρές ταυτίζονται στα δύο τους άκρα για συνθήκη τερματισμού 0.01% και 0.001%, ενώ η όποια διαφοροποίηση εντοπίζεται στα ενδιάμεσα σημεία της χρονοσειράς τα οποία στη δεύτερη περίπτωση εμφανίζουν κάπως πιο περιορισμένη καμπυλότητα. Στην πραγματικότητα η διαφορά μεταξύ της θεωρητικά απόλυτης εξίσωσης των καμπυλοτήτων και αυτής που επιτυγχάνεται πρακτικά έχει να κάνει με το γεγονός ότι στην πρώτη περίπτωση το αποτέλεσμα που προκύπτει είναι η ευθεία γραμμή που ορίζεται από τα δύο άκρα της χρονοσειράς μετά την εφαρμογή της τεχνικής *backcasting*, ενώ στη δεύτερη είναι μία καμπύλη ανεπαίσθητης κυρτότητας. Αυτό έχει ως αποτέλεσμα οι δύο περιπτώσεις εξομάλυνσης να είναι ισοδύναμες προβλεπτικά αφού τόσο η κλίση όσο και το επίπεδό τους ταυτίζονται.

Πίνακας 4.8: Απόδοση κλασικών μεθόδων πρόβλεψης χρονοσειρών βάσει *sMAPE* στα δεδομένα του διαγωνισμού *M3* για εξομάλυνση με χρήση της τεχνικής εξισορρόπησης τοπικών διακυμάνσεων με συνθήκη τερματισμού 0.1%, 0.01% και 0.001% αντίστοιχα.

Μέθοδος πρόβλεψης	Συνθήκη 0.1%				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naïve	17.98	9.77	13.97	6.40	13.34
SES	17.98	9.77	13.97	6.40	13.34
Holt	17.02	10.89	16.03	4.65	14.29
Damped	16.81	10.61	15.36	4.52	13.85
Theta	16.70	9.23	13.76	5.05	12.74

Μέθοδος πρόβλεψης	Συνθήκη 0.01%				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naïve	17.95	9.90	14.03	6.41	13.39
SES	17.95	9.90	14.03	6.41	13.39
Holt	16.64	9.81	14.79	4.39	13.33
Damped	16.55	9.70	14.5	4.33	13.14
Theta	16.55	9.26	13.72	5.01	12.70

Μέθοδος πρόβλεψης	Συνθήκη 0.001%				
	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Naïve	17.94	9.96	14.06	6.42	13.42
SES	17.94	9.96	14.06	6.42	13.42
Holt	16.63	9.45	14.35	4.44	13.03
Damped	16.55	9.39	14.16	4.37	12.90
Theta	16.52	9.30	13.69	5.01	12.69

Στον Πίνακα 4.9 παρουσιάζονται συγκριτικά τα χαρακτηριστικά και οι απαιτήσεις της προτεινόμενης τεχνικής εξομάλυνσης έναντι των κλασικών προσεγγίσεων που παρουσιάστηκαν νωρίτερα προκειμένου να συνοψιστούν τα πλεονεκτήματά της.

Έχοντας απαλλαγεί πρακτικά από τις παραμέτρους που καθορίζουν τον τερματισμό της διαδικασίας εξομάλυνσης και τον βαθμό εξομάλυνσης, η μόνη παράμετρος που διατηρείται στην εν λόγω προσέγγιση είναι ο καθορισμός της μεθόδου που θα εφαρμοστεί για την προέκταση της χρονοσειράς στα δύο της άκρα και την εξομάλυνσή της βάσει αυτών. Το γεγονός ότι η προέκταση που απαιτείται είναι πάντα μοναδιαίου μήκους αποτελεί βέβαια από μόνο του πλεονέκτημα της προσέγγισης έναντι των συμβατικών μεθόδων εξομάλυνσης, για τις οποίες και το μήκος είναι ανάλογο του βαθμού εξομάλυνσης. Ωστόσο, επειδή όπως αναφέρθηκε και για την περίπτωση του μη γραμμικού μετασχηματισμού θ η εν λόγω παράμετρος είναι καθοριστικής σημασίας, σε αυτό το σημείο παρουσιάζεται μία μεθοδολογία για τον καλύτερο καθορισμό των δύο σημείων ενδιαφέροντος. Σημειώνεται βέβαια πως η μέθοδος καθορισμού των σημείων είναι ανεξάρτητη της διαδικασίας εξομάλυνσης και μπορεί να χρησιμοποιηθεί κάλλιστα οποιαδήποτε εναλλακτική προσέγγιση για την πραγματοποίηση της ίδιας διαδικασίας.

Πίνακας 4.9: Σύγκριση μεθόδων εξομάλυνσης ως προς τα χαρακτηριστικά και τις απαιτήσεις τους

Μέθοδος εξομάλυνσης	Παραμετροποίηση	Επέκταση χρονοσειράς	Τύπος εξομάλυνσης	Διατήρηση τοπικών καμπυλοτήτων	Προβλεπτική αξία εξομαλυμένης χρονοσειράς
Κινητοί μέσοι όροι	<ul style="list-style-type: none"> Επίπεδο εξομάλυνσης k Μέθοδος προέκτασης 	Ανάλογα το επίπεδο εξομάλυνσης k	Γραμμικός Μη γραμμικός	Ναι	Ικανοποιητική
Εξομάλυνση μέσω kernel	<ul style="list-style-type: none"> Επίπεδο εξομάλυνσης k 	Ανάλογα το επίπεδο	Μη γραμμικός	Ναι	Καλή

	<ul style="list-style-type: none"> • Πυρήνας <i>kernel</i> • Μέθοδος προέκτασης 	εξομάλυνσης k			
Γραμμικός μετασχηματισμός θ	Τιμή παραμέτρου θ	-	Γραμμικός	Όχι	Ικανοποιητική
Μη γραμμικός μετασχηματισμός θ	<ul style="list-style-type: none"> • Συνθήκη τερματισμού • Βαθμός εξομάλυνσης • Μέθοδος προέκτασης 	Κατά ένα βήμα στα δύο άκρα της χρονοσειράς	Μη γραμμικός	Ναι	Πολύ καλή
Μετασχηματισμός εξισορρόπησης τοπικών διακυμάνσεων	Μέθοδος προέκτασης	Κατά ένα βήμα στα δύο άκρα της χρονοσειράς	Μη γραμμικός	Ναι	Πολύ καλή

4.4.5.2 Μεθοδολογία βελτίωσης της απόδοσης της τεχνικής *backcasting*

Όπως έχει αναφερθεί εκτενώς στην παρούσα διατριβή, ο καθορισμός του βέλτιστου μοντέλου πρόβλεψης για μία χρονοσειρά είναι μία ιδιαίτερα απαιτητική διαδικασία, κυρίως λόγω της αβεβαιότητας που χαρακτηρίζει τις προβλέψεις αυτές καθαυτές. Το γεγονός ότι ένα μοντέλο πρόβλεψης εφαρμόζει καλά επί των ιστορικών δεδομένων έχει αποδειχτεί επανειλημμένα ότι είναι ανεξάρτητο της προβλεπτικής τους ακρίβειας και συνεπώς πολλές *in-sample* και *out-of-sample* τεχνικές αξιολόγησης που συναντώνται στη βιβλιογραφία (βλ. [Παράγραφο 2.5.1](#)) μπορούν να αποδειχτούν αναποτελεσματικές παρά την εκτενή ανάλυση που πραγματοποιούν επί της απόδοσης των επιμέρους μοντέλων. Το εν λόγω φαινόμενο διογκώνεται μάλιστα όταν η χρονοσειρά χαρακτηρίζεται από έντονη τυχαιότητα ή τα μοτίβα της αλλάζουν σημαντικά με την πάροδο του χρόνου. Έτσι, η αβεβαιότητα που εμπεριέχει η όλη διαδικασία της επιλογής βέλτιστης μεθόδου πρόβλεψης γεννά το ερώτημα κατά πόσο αξίζει κανείς να επενδύσει σε μία τέτοια για την αποτελεσματικότερη εφαρμογή της τεχνικής *backcasting*, δεδομένου του υψηλού υπολογιστικού κόστους που κάτι τέτοιο συνεπάγεται και του ότι αυτό θα προστεθεί μάλιστα στο κόστος της εξομάλυνσης και της μετέπειτα πρόβλεψης των χρονοσειρών. Έτσι, ο στόχος που τίθεται για άλλη μια φορά σε αυτό το σημείο είναι η βελτίωση της απόδοσης της μεθόδου πρόβλεψης που θα επιλεγεί να χρησιμοποιηθεί εντός της τεχνικής *backcasting* και όχι η επιλογή της αυτή καθαυτή μεταξύ άλλων μεθόδων.

Στην εισαγωγή του παρόντος κεφαλαίου παρουσιάστηκαν αρκετές τεχνικές πέραν της εξομάλυνσης για τη βελτίωση της προβλεπτικής ακρίβειας των μοντέλων πρόβλεψης χρονοσειρών, η απλούστερη εκ των οποίων είναι η χρήση μετασχηματισμών. Παρά τα όποια μειονεκτήματά τους, τα οφέλη των μετασχηματισμών σε προβλεπτική ακρίβεια είναι αρκετά πιθανά και εμπειρικά οι ερευνητές αποδέχονται τον ισχυρισμό ότι ένας επιτυχημένος μετασχηματισμός δεδομένων αν δεν βελτιώσει την ακρίβεια των προβλέψεων, αν μη τι άλλο δεν θα την περιορίσει σημαντικά. Αυτό το πολύ βασικό συμπέρασμα, σε συνδυασμό πάντα με την απλότητα της εφαρμογής τους, αποτελεί πολύ ισχυρό κίνητρο για την υιοθέτησή τους εντός μίας μεθοδολογίας για τη βελτίωση της απόδοσης της τεχνικής *backcasting*. Έτσι, βάσει όσων παρουσιάστηκαν στην [Παράγραφο 2.4.2](#) προτείνεται σε πρώτη φάση ο μετασχηματισμός των δεδομένων πριν την προέκτασή τους μέσω της συνάρτησης *Box-Cox* και η επιλογή της παραμέτρου λ μέσω της

ελαχιστοποίησης της συνάρτησης πιθανοφάνειας. Προκειμένου μάλιστα να περιοριστεί το εύρος των επιλογών εντός ενός λογικού πλαισίου προτείνεται η χρήση τιμών λ στο διάστημα $[-1,1]$. Αναφέρεται ότι ο μετασχηματισμός του *Johnson* απορρίφθηκε ως εναλλακτική, αφενός λόγω της πολυπλοκότητάς του και αφετέρου λόγω της απροσδιοριστίας στην οποία συχνά εμπίπτει.

Όσον αφορά τώρα το ποιο μοντέλο πρόβλεψης θα πρέπει να εφαρμοστεί για την προέκταση της μετασχηματισμένης χρονοσειράς η απάντηση δεν είναι προφανής. Η χρήση των γενικευμένων εκφράσεων του μοντέλου θ που παρουσιάστηκαν στην *Παράγραφο 3.4.3* δεν αποτελούν ικανοποιητικές λύσεις καθώς ο μετασχηματισμός δυνάμει συχνά αλλοιώνει τα μοτίβα τάσης και τα μετατρέπει σε γραμμικά καθιστώντας τις ανούσιες. Επιπλέον, η προέκταση της χρονοσειράς με μία μέθοδο όπως η *Naïve* και η *SES* είναι ανεπαρκής καθώς αυτές δεν λαμβάνουν υπόψη τους την τάση των δεδομένων, η οποία και εδώ είναι εξέχουσας σημασίας από τη στιγμή που η εξομάλυνση θα πραγματοποιηθεί βάσει των παραγόμενων σημείων. Από την άλλη μεριά η χρήση ενός μοντέλου όπως η *holt* ή η κλασική θ είναι επίσης ριψοκίνδυνη επιλογή καθώς η πρώτη επεκτείνει τη χρονοσειρά βάσει της πιο πρόσφατης τάσης της, που ενδέχεται να είναι υπερεκτιμημένη, ενώ η δεύτερη με το μισό της μακροχρόνιας τάσης της, που ενδέχεται να είναι υποεκτιμημένη. Η *damped* φαντάζει ως η πλέον λογική επιλογή, ωστόσο, όπως παρουσιάστηκε διεξοδικά στις περιπτώσεις των *KMO* και *kernel*, δεν είναι πάντοτε και η βέλτιστη. Στην κατεύθυνση λοιπόν του περιορισμού της αβεβαιότητας της επιλογής και παραμετροποίησης των μοντέλων πρόβλεψης προτείνεται ο συνδυασμός τριών κλασικών μεθόδων πρόβλεψης με ίσα βάρη: της *SES*, της *holt* και της *damped*. Ο συνδυασμός μοντέλων πρόβλεψης ανήκει και αυτός στις τεχνικές που παρουσιάστηκαν για τη βελτίωση της προβλεπτικής ακρίβειας των χρονοσειρών ενώ ο συγκεκριμένος είχε εξαιρετική επίδοση στα δεδομένα του διαγωνισμού *M3* και θεωρείται ιδιαίτερως στιβαρή επιλογή καθώς οι τρεις μέθοδοι διαχειρίζονται τα χαρακτηριστικά των χρονοσειρών με αρκετά διαφορετικό τρόπο και συνεπώς παράγουν διαφοροποιημένες προβλέψεις. Έτσι, ο συνδυασμός τους έχει σημαντική πιθανότητα να παράγει αντιπροσωπευτικές προβλέψεις.

Σημειώνεται ότι ακριβώς επειδή οι χρονοσειρές που καλούμαστε να προεκτείνουμε είναι απαλλαγμένες από εποχιακότητα (αποεποχικοποίηση πριν την πρόβλεψη) και θόρυβο (μετασχηματισμός πριν την εξομάλυνση), δεν απαιτείται η χρήση κάποιου πιο σύνθετου μοντέλου πρόβλεψης καθώς αυτό θα αυξήσει την πολυπλοκότητα της όλης διαδικασίας χωρίς να οδηγήσει αναγκαστικά σε καλύτερα αποτελέσματα. Άλλωστε έχει αποδειχτεί εμπειρικά ότι πολύπλοκες μέθοδοι συχνά αποτυγχάνουν να ξεπεράσουν προβλεπτικά αρκετά πιο απλοϊκές μεθόδους, ακόμα και όταν δεν έχει προηγηθεί επεξεργασία επί των δεδομένων των χρονοσειρών. Χαρακτηριστικό παράδειγμα αποτελούν τα συμπεράσματα των διαγωνισμών *M1* και *M3*. Επιπλέον, έχοντας τα παραπάνω ως δεδομένα, γίνεται κατανοητό ότι οι υπόλοιπες τεχνικές που αναφέρθηκαν για τη βελτίωση της προβλεπτικής ακρίβειας των μοντέλων πρόβλεψης χρονοσειρών, όπως είναι η απαλοιφή *outlier* και η χρήση τεχνικών συνάθροισης στο επίπεδο του χρόνου, μπορούν να θεωρηθούν περιττές καθώς στην πρώτη περίπτωση οι μετασχηματισμοί έχουν περιορίσει ήδη πιθανές ακραίες

τιμές ενώ στη δεύτερη περίπτωση η τεχνική έχει ελάχιστα περιθώρια να αναδείξει χαρακτηριστικά που είναι εμφανή σε εναλλακτικές συχνότητες.

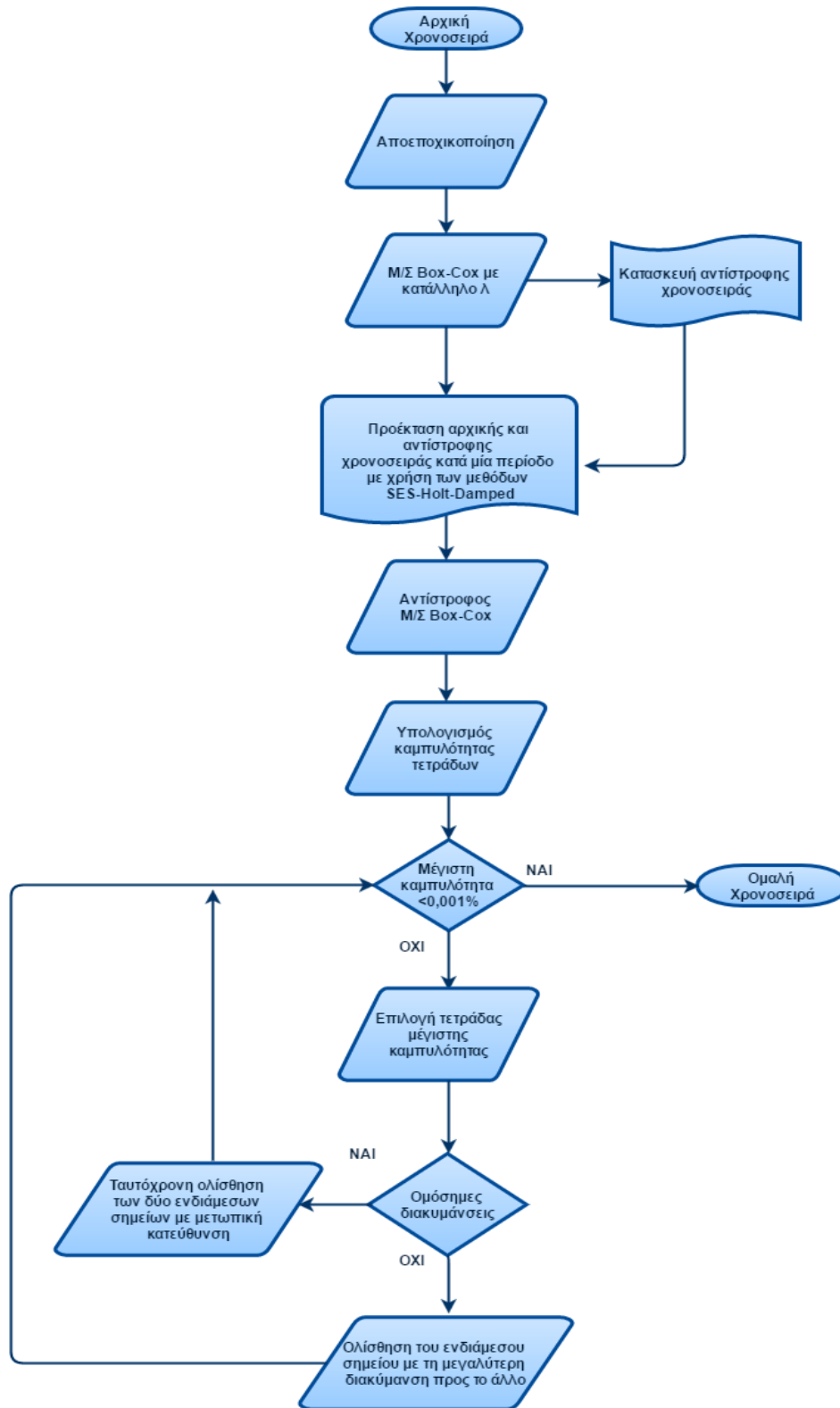
Η ανωτερότητα της προτεινόμενης μεθοδολογίας στην κατεύθυνση της βελτίωσης της ακρίβειας των μοντέλων πρόβλεψης στις εξομαλυμένες χρονοσειρές εξάγεται μέσα από τα αποτελέσματα του Πίνακα 4.10 όπου και ελέγχεται η ακρίβεια που επιτυγχάνεται για την κλασική μέθοδο θ στην περίπτωση μετασχηματισμού ή μη των δεδομένων πριν την εφαρμογή της τεχνικής *backcasting* και εφαρμογή αυτής για πληθώρα μοντέλων πρόβλεψης. Πιο συγκεκριμένα η αρχική χρονοσειρά προεκτείνεται στα δύο της άκρα μέσω των μεθόδων *Damped*, *Ses-Holt-Damped (S-H-D)* και *Theta*, έχοντας ή μη μετασχηματίσει (T) κατάλληλα τα αποεποχικοποιημένα δεδομένα με χρήση του Box-Cox.

Πίνακα 4.10: Απόδοση κλασικής μεθόδου θ βάσει sMAPE στα δεδομένα του διαγωνισμού M3 για εξομάλυνση μέσω της εξισορρόπησης τοπικών διακυμάνσεων. Για την εφαρμογή της τεχνικής *backcasting* χρησιμοποιείται πληθώρα μοντέλων πρόβλεψης ενώ εξετάζεται και η εφαρμογή μετασχηματισμού Box-Cox.

Μέθοδος πρόβλεψης	Ετήσια	Τριμηνιαία	Μηνιαία	Λοιπά	Σύνολο
Damped	16.55	9.26	13.72	5.01	12.70
S-H-D	16.65	9.29	13.68	5.14	12.72
Theta	16.93	9.59	13.78	5.24	12.91
Damped (T)	16.59	9.27	13.63	4.96	12.66
S-H-D (T)	16.75	9.29	13.60	5.12	12.70
Theta (T)	16.72	9.54	13.66	5.25	12.79

Όπως παρατηρείται, η εφαρμογή του μετασχηματισμού Box-Cox πριν την προέκταση των δύο άκρων της αποεποχικοποιημένης χρονοσειράς οδηγεί συνολικά σε ελαφρώς βελτιωμένα αποτελέσματα ανεξαρτήτως της μεθόδου πρόβλεψης που χρησιμοποιείται. Επίσης, σε αντίθεση με τις προηγούμενες περιπτώσεις μεθόδων εξομάλυνσης, οι διαφοροποιήσεις στην ακρίβεια πρόβλεψης είναι πλέον αρκετά πιο περιορισμένες μειώνοντας έτσι σημαντικά την επίδραση που έχει η μέθοδος που χρησιμοποιείται στο *backcasting* στην συνολική απόδοση της προβλεπτικής διαδικασίας. Αυτός είναι και ο λόγος που μία πιο γενικευμένη προσέγγιση, όπως είναι ο συνδυασμός των μεθόδων *Ses-Holt-Damped*, προτιμάται εν τέλει στη συγκεκριμένη περίπτωση από την *Damped*, παρόλο που η δεύτερη είναι ανεπαίσθητα πιο ακριβής στο συγκεκριμένο σετ δεδομένων.

Βάσει των παραπάνω καθορίζεται λοιπόν με σαφήνεια μία ολοκληρωμένη μεθοδολογία για την εφαρμογή της προτεινόμενης τεχνικής εξομάλυνσης μέσω της εξισορρόπησης τοπικών διακυμάνσεων η οποία περιλαμβάνει αφενός την εξομάλυνση της χρονοσειράς αυτή καθαυτή και αφετέρου τη βελτίωση της απόδοσης της τεχνικής *backcasting* προκειμένου να προκύψει το βέλτιστο δυνατό αποτέλεσμα. Η μεθοδολογία παρουσιάζεται σε μορφή διαγράμματος ροής στην *Εικόνα 4.15*.



Εικόνα 4.15: Διάγραμμα ροής για την ολοκληρωμένη εφαρμογή της προτεινόμενης τεχνικής εξομάλυνσης μέσω της εξισορρόπησης τοπικών διακυμάνσεων.

Είναι σημαντικό για τους επενδυτές να καταλάβουν τι ξέρουν και τι όχι. Μάθε λοιπόν πως δεν μπορείς να ξέρεις τι θα συμβεί στο μέλλον και ότι κάθε επένδυση που βασίζεται στην ακρίβεια μίας πρόβλεψης για το πού θα βρίσκεται η αγορά τον επόμενο χρόνο είναι καταδικασμένη.

Barry Ritholtz

Κεφάλαιο 5: Παραγωγή τεχνητών χρονοσειρών μέσω προσομοίωσης

Προκειμένου να επιβεβαιωθούν τα συμπεράσματα της παρούσας διατριβής, αλλά και να εμπλουτιστούν πιθανώς, πρωταρχικό μέλημα αποτελεί η εφαρμογή της μεθοδολογίας που θα προταθεί σε ένα μεγάλο και ποικιλόμορφο σετ δεδομένων. Για την απόκτηση ενός τέτοιου σετ μπορεί κανείς να στραφεί σε χρονοσειρές που χρησιμοποιούνται ευρέως στην επιστημονική κοινότητα για την αξιολόγηση τεχνικών προβλέψεων, στη συλλογή νέων από εξωτερικούς φορείς ή στην παραγωγή τεχνητών μέσω διαδικασιών προσομοίωσης (*simulation*) και αναδειγματοληψίας (*bootstrapping*).

Ένα από τα πλέον χρησιμοποιούμενα σετ δεδομένων για την αξιολόγηση μεθόδων και τεχνικών πρόβλεψης, αποτελούν οι χρονοσειρές των διαγωνισμών *M* (*Makridakis Competition*), και ιδίως αυτές του *M3*. Οι χρονοσειρές των εν λόγω διαγωνισμών έχουν χρησιμοποιηθεί μάλιστα σε τόσες πολλές μελέτες στο χώρο των προβλέψεων που πλέον αποτελούν το βασικότερο σημείο αναφοράς προκειμένου να εξακριβωθεί η επίδοση οποιασδήποτε νέας μεθόδου. Ωστόσο, τα δεδομένα των διαγωνισμών *M* έχουν δεχτεί κατά καιρούς σφοδρή κριτική σχετικά με την καταλληλότητά τους, κυρίως λόγω του μικρού τους όγκου και της ανομοιόμορφης κατανομής των χρονοσειρών στα επιμέρους τους χαρακτηριστικά, παράγοντες που μπορούν να οδηγήσουν πιθανώς στην υπερεκτίμηση ή υποτίμηση συγκεκριμένων μεθόδων. Αντίστοιχα προβλήματα μπορούν βέβαια να προκύψουν κάλλιστα και για οποιοδήποτε άλλο σετ πραγματικών δεδομένων για το οποίο δεν διαθέτουμε επαρκείς πληροφορίες σχετικά με τη φύση και τα χαρακτηριστικά των χρονοσειρών που εμπεριέχει.

Έχοντας ως βάση τα παραπάνω, στο παρόν κεφάλαιο παρουσιάζεται λοιπόν η μεθοδολογία που ακολουθήθηκε προκειμένου να παραχθούν τεχνητές χρονοσειρές ποικίλων χαρακτηριστικών και μάλιστα ομοιόμορφα κατανομημένων ως προς αυτά. Οι χρονοσειρές παράχθηκαν μέσω μίας διαδικασίας προσομοίωσης και πιο συγκεκριμένα με τη βοήθεια μίας γεννήτριας δεδομένων. Στις παραγράφους που ακολουθούν παρουσιάζονται αρχικά σετ δεδομένων από διαγωνισμούς προβλέψεων που αξιοποιούνται σήμερα για την αξιολόγηση της επίδοσης των μοντέλων πρόβλεψης, αναλύεται η ανάγκη για παραγωγή τεχνητών χρονοσειρών και τέλος περιγράφεται η μεθοδολογία κατασκευής και επεξεργασίας τους.

5.1 Διαγωνισμοί προβλέψεων

Οι διαγωνισμοί προβλέψεων αποτελούν οργανωμένες προσπάθειες που έχουν γίνει κατά καιρούς προκειμένου να εξαχθούν εμπειρικά συμπεράσματα σχετικά με την επίδοση των μεθόδων πρόβλεψης. Βασίζονται στη συλλογή πληθώρας χρονοσειρών διαφορετικής ή παραπλήσιας φύσης - τόσο ως προς την κατηγορία (π.χ. μακροοικονομικά δεδομένα) όσο και τα χαρακτηριστικά τους (π.χ. εποχιακές χρονοσειρές) – και την εφαρμογή γνωστών και νέων μεθόδων σε αυτές για την παραγωγή προβλέψεων σε συγκεκριμένα διαστήματα για τα οποία οι πραγματικές τιμές δεν είναι εξαρχής διαθέσιμες. Οι μέθοδοι αξιολογούνται

στη συνέχεια με χρήση προκαθορισμένων δεικτών σφαλμάτων ως προς την ακρίβεια, την προκατάληψη και την ευστάθειά τους ή και ως προς άλλα μεγέθη που μπορεί να σχετίζονται με την επίδοσή τους. Η επίδοση κάθε μεθόδου κρίνεται συγκρίνοντάς την είτε με αυτή των υπολοίπων μεθόδων που συμμετέχουν στο διαγωνισμό, είτε με αυτήν απλοϊκών μεθόδων πρόβλεψης (π.χ. της *Naïve*) που αποτελούν συνήθως σημεία αναφοράς λόγω της απλότητας και της μικρότερης θεωρητικά αναμενόμενης ακριβείας τους.

Αν και πολλοί στατιστικολόγοι συνεχίζουν να αμφισβητούν τα αποτελέσματα των διαγωνισμών λόγω της μη θεωρητικής τεκμηρίωσής τους, εν τούτοις στην επιστημονική κοινότητα των προβλέψεων έχουν γίνει πλήρως αποδεκτά, με τα δεδομένα τους να αποτελούν σημείο αναφοράς και μελέτης για όλους τους ερευνητές. Εξέχοντα ρόλο στην καθιέρωση και στην αποδοχή των διαγωνισμών αποτέλεσαν τα σχεδόν πανομοιότυπα συμπεράσματα που εξήχθησαν μέσω των διαγωνισμών *M*, τα οποία και συνεχίζουν να επαληθεύονται εμπειρικά μέσω άλλων πιο σύγχρονων μελετών. Οι διαγωνισμοί οι οποίοι έχουν λάβει χώρα μέχρι στιγμής είναι συνοπτικά οι ακόλουθοι:

- M Competition (*Makridakis et al., 1982*)
- M2 Competition (*Makridakis et al., 1993*)
- M3 Competition (*Makridakis et al., 2000*)
- T- Competition (*Makridakis et al., 1993*)
- Διαγωνισμοί Προβλέψεων Neutral Networks NN3 (2006-2007)
- Διαγωνισμοί Προβλέψεων Neutral Networks NN4 (2009)
- Διαγωνισμοί Προβλέψεων Neutral Networks NN5 (2008)

5.1.1 Διαγωνισμοί M

Οι Reid, Newbold και Granger ήταν οι πρώτοι οι οποίοι τη δεκαετία του 1970 θέλησαν να εκτιμήσουν εμπειρικά την ακρίβεια των μεθόδων πρόβλεψης μέσω της εφαρμογής απλών μεθόδων σε σετ δεδομένων. Λόγω βέβαια του υψηλού υπολογιστικού κόστους για την εποχή, ο αριθμός των χρονοσειρών που χρησιμοποιήθηκε στις μελέτες τους ήταν ιδιαίτερα περιορισμένος. Έτσι, οι *Makridakis και Hibon (1979)* έκαναν μία πρώτη προσπάθεια να εκτιμήσουν την ακρίβεια των μεθόδων χρησιμοποιώντας 111 χρονοσειρές μακροοικονομικών δεδομένων. Το βασικότερο συμπέρασμα της μελέτης τους ήταν ότι οι πιο περίπλοκες μέθοδοι δεν παρήγαγαν αναγκαστικά και ακριβέστερες προβλέψεις. Όπως ήταν αναμενόμενο η εν λόγω έρευνα βρήκε αρκετούς πολέμιους στο χώρο της στατιστικής οι οποίοι, δεδομένης της μη θεωρητικής τεκμηρίωσης των αποτελεσμάτων, συνέχιζαν να υποστηρίζουν πως τα πιο σύνθετα μοντέλα που κατασκεύαζαν, όπως π.χ. τα *ARIMA*, ήταν και ανώτερα προβλεπτικά.

5.1.1.1 Διαγωνισμός M1

Προκειμένου να στηρίξουν τα αποτελέσματα της ερευνάς τους, οι *Makridakis και Hibon* προχώρησαν το 1982 στη δημιουργία του διαγωνισμού *M1* (*Makridakis et al., 1982*). Αυτή τη φορά οι χρονοσειρές, οι οποίες αναφέρονταν σε διάφορες κατηγορίες δεδομένων (μικροοικονομικά-MICRO, βιομηχανία-INDUSTRY, μακροοικονομικά-MACRO και δημογραφικά-DEMOG), έφταναν σε αριθμό τις 1001 ενώ θα εφαρμόζονταν σε αυτές 15

διαφορετικά μοντέλα πρόβλεψης και 9 παραλλαγές τους. Η υλοποίηση των μεθόδων και η εξαγωγή των αποτελεσμάτων έγινε από τους δύο ερευνητές και τα συμπεράσματα ήταν τα ίδια με της προηγούμενης μελέτης τους. Σημειώνουμε ότι για μεθόδους με μεγάλο υπολογιστικό κόστος το σετ δεδομένων του διαγωνισμού μειωνόταν στις 111 χρονοσειρές.

Πίνακας 5.1: Πλήθος χρονοσειρών διαγωνισμού M1 ανά συχνότητα και κατηγορία.

Κατηγορία Χρονοσειρών						
Συχνότητα	MICRO	INDUSTRY	MACRO	DEMOG	Σύνολο	
Ετήσια	57	35	59	30	181	
Τριμηνιαία	42	18	104	39	203	
Μηνιαία	203	183	156	75	617	
Σύνολο	302	236	319	144	1001	

Συγκεκριμένα έδειξαν πως (α) στη γενική περίπτωση οι πιο απλές μέθοδοι εμφανίζουν καλύτερη απόδοση προβλεπτικά, ή τουλάχιστον οι πιο πολύπλοκες μέθοδοι δεν μπορούν να εγγυηθούν και καλύτερες προβλέψεις, (β) με την αύξηση του ορίζοντα πρόβλεψης αυξάνονται τα σφάλματα πρόβλεψης, (γ) ανάλογα το κριτήριο απόδοσης που χρησιμοποιείται ενδέχεται να αλλάζει και η κατάταξη μεταξύ των μεθόδων που συμμετέχουν στο διαγωνισμό, και (γ) ο συνδυασμός απλών μεθόδων αποτελεί συνήθως καλύτερη επιλογή από την πρόβλεψη με τις επιμέρους μεθόδους ξεχωριστά.

Τα δεδομένα, καθώς και περισσότερες πληροφορίες για το διαγωνισμό είναι διαθέσιμες στην ιστοσελίδα του *International Institute of Forecasters* (<http://forecasters.org/resources/time-series-data/m-competition>).

5.1.1.2 Διαγωνισμός M2

Ο διαγωνισμός M2 (Makridakis et al., 1993) αποτέλεσε μία ακόμα πιο οργανωμένη προσπάθεια των Makridakis και Hibon για την ενίσχυση των αποτελεσμάτων του διαγωνισμού M1 ενώ παράλληλα στόχευε στον εντοπισμό των παραγόντων που επηρεάζουν την απόδοση των μεθόδων πρόβλεψης και κυρίως την αξιολόγηση της ανθρώπινη κρίσης. Ο διαγωνισμός κράτησε περίπου 2 χρόνια και το ζητούμενο ήταν η παραγωγή προβλέψεων 15 μηνών μπροστά για 6 μακροοικονομικές χρονοσειρές τεσσάρων εταιρειών, όσων προβλέψεων παράγουν δηλαδή συνήθως και οι ίδιες οι εταιρείες για τον μεσοπρόθεσμο προγραμματισμό τους.

Χαρακτηριστικό του διαγωνισμού ήταν ότι, πέρα από τις ίδιες τις χρονοσειρές, οι ερευνητές είχαν τη δυνατότητα να επικοινωνούν με μέλη των συνεργαζόμενων εταιρειών, να συλλέγουν εξωτερικά στοιχεία που θεωρούσαν σημαντικά και να ενσωματώνουν οποιαδήποτε πληροφόρηση στην παραγωγή προβλέψεων. Τέτοιες πληροφορίες μπορεί να ήταν γενικά στοιχεία περί της οικονομίας των ΗΠΑ ή του κλάδου στον οποίο δραστηριοποιούταν η εκάστοτε εταιρεία ειδικότερα, επεξηγήσεις των δεδομένων και των μεταβλητών που επηρεάζουν τα προβλεπόμενα μεγέθη κ.ο.κ. Καθοριστικής σημασίας για την ποιότητα της έρευνας ήταν η υλοποίηση των μεθόδων και της όλης διαδικασίας

πρόβλεψης από διαφορετικούς εξειδικευμένους ερευνητές, πέραν των δύο συγγραφέων, γεγονός που ενίσχυσε την αντικειμενικότητα και την αξιοπιστία των αποτελεσμάτων.

Τα αποτελέσματα του διαγωνισμού ήταν πανομοιότυπα με αυτά του *M1*, με τις πιο απλές μεθόδους να σημειώνουν καλύτερη απόδοση συγκριτικά με τις πιο σύνθετες και τον συνδυασμό αυτών να αποδεικνύεται καλύτερη επιλογή από την εφαρμογή των επιμέρους μεθόδων μεμονωμένα. Τέλος, η ακρίβεια των διαφόρων μεθόδων πρόβλεψης φάνηκε να επηρεάζεται αρνητικά από το μέγεθος του ορίζοντα πρόβλεψης ενώ, ανάλογα με το κριτήριο αξιολόγησης που χρησιμοποιούσαν, άλλαζε η κατάταξη των μεθόδων που συμμετείχαν στο διαγωνισμό. Τα αποτελέσματα του διαγωνισμού παρουσιάστηκαν αναλυτικά στο *International Journal of Forecasting* και στο *International Symposium of Forecasting* και έγιναν διαθέσιμα μαζί με σχετικές πληροφορίες για την πιστή αναπαραγωγή τους σε πάνω από 600 ερευνητές.

Τα δεδομένα, καθώς και περισσότερες πληροφορίες για το διαγωνισμό είναι διαθέσιμες στην ιστοσελίδα του *International Institute of Forecasters* (<http://forecasters.org/resources/time-series-data/m2-competition>).

5.1.1.3 Διαγωνισμός *M3*

Ο διαγωνισμός *M3* (*Makridakis et al., 2000*) αποτελεί το μεγαλύτερο διαγωνισμό προβλέψεων που έχει διεξαχθεί μέχρι και σήμερα με τα αποτελέσματα και τα δεδομένα του να αποτελούν αντικείμενο μελέτης για εκατοντάδες ερευνητές και μελέτες. Οργανώθηκε με χορηγό το περιοδικό *International Journal of Forecasting* και συμμετείχαν 25 ομάδες, εκ των οποίων οι 19 ήταν ακαδημαϊκές και οι 6 εξειδικευμένα πακέτα λογισμικού επιχειρηματικών προβλέψεων.

Πίνακας 5.2: Πλήθος χρονοσειρών διαγωνισμού *M3* ανά συχνότητα και κατηγορία.

Κατηγορία Χρονοσειρών							
Συχνότητα	MICRO	INDUSTRY	MACRO	FINANCE	DEMOG	OTHER	Σύνολο
Ετήσια	146	102	83	58	245	11	645
Τριμηνιαία	204	83	336	76	57	0	756
Μηνιαία	474	334	312	145	111	52	1428
Ετήσια	4	0	0	29	0	141	174
Σύνολο	828	519	731	308	413	204	3003

Βασική ιδιαιτερότητα του διαγωνισμού ήταν αφενός το σαφώς μεγαλύτερο πλήθος χρονοσειρών (3003) διαφόρων κατηγοριών (μικροοικονομικά-MICRO, βιομηχανία-INDUSTRY, μακροοικονομικά-MACRO, χρηματοοικονομικά-FINANCE, δημογραφικά-DEMOG και λοιπά-OTHER) και αφετέρου η συμμετοχή μη στατιστικών μεθόδων πρόβλεψης, όπως τα νευρωνικά δίκτυα (*neural networks*) και τα έξυπνα συστήματα (*expert systems*). Συνοπτικά, από τις στατιστικές μεθόδους που συμμετείχαν αναφέρουμε τις μεθόδους εκθετικής εξομάλυνσης και τα μοντέλα *ARIMA*, από τα νευρωνικά δίκτυα τις τεχνικές *RBF* (*rule based forecasting*) και από τα έξυπνα συστήματα τα *ForecastPro*, *SmartForecasts*, *Autocast* και *Autobox*. Έτσι, τα αποτελέσματα των προηγούμενων

διαγωνισμών M θα δοκιμάζονταν σε μεγαλύτερο όγκο δεδομένων και για περισσότερες και πιο εξειδικευμένες μεθόδους.

Πίνακας 5.3: Αποτελέσματα διαγωνισμού $M3$ για το σύνολο των δεδομένων ($sMAPE$) ανά ορίζοντα πρόβλεψης.

Μέθοδος/ Ορίζοντας	1	2	3	4	5	6	8	12	15	18	1 - 18
Naïve2	10.5	11.3	13.6	15.1	15.1	15.9	14.5	16.0	19.3	20.7	15.5
Single	9.5	10.6	12.7	14.1	14.3	15.0	13.3	14.5	18.3	19.4	14.3
Holt	9.0	10.4	12.8	14.5	15.1	15.8	13.9	14.8	18.8	20.2	14.6
Dampen	8.8	10.0	12.0	13.5	13.7	14.3	12.5	13.9	17.5	18.9	13.6
Winter	9.1	10.5	12.9	14.6	15.1	15.9	14.0	14.6	18.9	20.2	14.7
Comb S-H-D	8.9	10.0	12.0	13.5	13.7	14.2	12.4	13.6	17.3	18.3	13.5
B-J automatic	9.2	10.4	12.2	13.9	14.0	14.8	13.0	14.1	17.8	19.3	14.0
Autobox1	9.8	11.1	13.1	15.1	16.0	16.8	14.2	15.4	19.1	20.4	15.2
Autobox2	9.5	10.4	12.2	13.8	13.8	14.9	13.2	15.2	18.2	19.9	14.4
Autobox3	9.7	11.2	12.9	14.6	15.8	16.5	14.4	16.1	19.2	21.2	15.3
Robust-Trend	10.5	11.2	13.2	14.7	15.0	15.9	15.1	17.5	22.2	24.3	16.3
ARARMA	9.7	10.9	12.6	14.2	14.6	15.6	13.9	15.2	18.5	20.3	14.7
Automat ANN	9.0	10.4	11.8	13.8	13.8	15.5	13.4	14.6	17.3	19.6	14.1
Flores/Pearce1	9.2	10.5	12.6	14.5	14.8	15.3	13.8	14.4	19.1	20.8	14.7
Flores/Pearce2	10.0	11.0	12.8	14.1	14.1	14.7	12.9	14.4	18.2	19.9	14.3
PP-autocast	9.1	10.0	12.1	13.5	13.8	14.7	13.1	14.3	17.7	19.6	14.0
ForecastPro	8.6	9.6	11.4	12.9	13.3	14.3	12.6	13.2	16.4	18.3	13.2
SmartFcs	9.2	10.3	12.0	13.5	14.0	15.1	13.0	14.9	18.0	19.4	14.1
Theta-sm	9.8	11.3	12.6	13.6	14.3	15.0	12.7	14.0	16.2	18.3	13.9
Theta	8.4	9.6	11.3	12.5	13.2	14.0	12.0	13.2	16.2	18.2	13.0
RBF	9.9	10.5	12.4	13.4	13.2	14.2	12.8	14.1	17.3	17.8	13.8
ForecastX	8.7	9.8	11.6	13.1	13.2	13.9	12.6	13.9	17.8	18.7	13.5
AAM1	9.8	10.6	11.2	12.6	13.0	13.5	14.1	14.9	18.0	20.4	14.6
AAM2	10.0	10.7	11.3	12.9	13.2	13.7	14.3	15.1	18.4	20.7	14.9

Οι χρονοσειρές του διαγωνισμού χωρίστηκαν σε ετήσιες, τριμηνιαίες, μηνιαίες και λοιπές χρονοσειρές. Για τις εκάστοτε συχνότητες ο ορίζοντας πρόβλεψης τέθηκε ίσος με 6, 8, 18 και 8 περιόδους, ενώ ο ελάχιστος αριθμός παρατηρήσεων ήταν 14, 16, 48 και 60 αντίστοιχα. Ο ικανοποιητικός αριθμός παρατηρήσεων των χρονοσειρών, σε συνδυασμό με τις αυστηρά θετικές τιμές τους σε όλο τους το εύρος, συνέβαλε στη σωστή εκπαίδευση όλων των μοντέλων πρόβλεψης που θα συμμετείχαν και στην αποφυγή οποιουδήποτε προβλήματος στις αξιολογήσεις τους μέσω των δεικτών $sMAPE$, $MAPE$ και άλλων σχετικών κριτηρίων. Σε αυτό το σημείο αναφέρετε ότι οι παρατηρήσεις στις περιόδους παραγωγής προβλέψεων ήταν άγνωστες στους συμμετέχοντες προκειμένου να εξασφαλιστεί η αντικειμενικότητα των αποτελεσμάτων.

Κεφάλαιο 5: Παραγωγή τεχνητών χρονοσειρών μέσω προσομοίωσης

Για ακόμα μία φορά τα συμπεράσματα των δύο προηγούμενων διαγωνισμών και ακόμα παλαιότερων μελετών επιβεβαιώθηκαν: Οι πιο απλές μέθοδοι πρόβλεψης είχαν την τάση να επιτυγχάνουν την ίδια ή και καλύτερη ακρίβεια σε σχέση με τις πιο πολύπλοκες μεθόδους, η αύξηση του ορίζοντα πρόβλεψης μείωνε σημαντικά την ακρίβεια των μοντέλων και η χρήση διαφορετικών δεικτών σφάλματος οδηγούσε σε σχετικές διαφοροποιήσεις στα συμπεράσματα.

Πίνακας 5.4: Οι μέθοδοι που δίνουν τα καλύτερα αποτελέσματα στο διαγωνισμό M3 για το σύνολο των δεδομένων ανά κατηγορία.

Κριτήριο Απόδοσης	Micro (828)	Industry (519)	Macro (731)	Finance (308)	Demographic (413)	Other (204)
sMAPE	Theta	Forecast X/	RBF/ARA RMA	AAM1/	ForecastX	Comb S-H-D
	Forecast Pro	Forecast Pro	Theta/	AAM2	Dampen	ARARMA
			Robust-Trend		ForecastPro/RBF	Forecast Pro
					SmartFcs	Comb S-H-D
Μέση κατάταξη	Theta	Forecast Pro	Robust-Trend	AAM1/	Robust-Trend	Theta
		Theta/		AAM2	ForecastX	Autobox2 /
		ForecastX				ARARMA
		Comb S-H-D				Forecast Pro
MAPE	Theta	Forecast X	Robust-Trend	Autobox3	RBF	Theta
	Forecast Pro	Theta	ARARMA	ForecastPro	Robust-Trend	Autobox2
MRAE	Theta	Theta	Robust-Trend	Robust-Trend	RBF	ARARMA
		RBF/	ARARMA	ARARMA		Theta
		Comb S-H-D	RBF	Theta		Autobox2
				AAM1/AAM2		Comb S-H-D

Η μονάδα προβλέψεων και στρατηγικής του Ε.Μ.Π. συμμετείχε στο διαγωνισμό με το μοντέλο αποσύνθεσης θ , το οποίο και σημείωσε όπως φαίνεται και στους αντίστοιχους πίνακες εξαιρετική απόδοση, ιδίως στις μηνιαίες χρονοσειρές και στα μικροοικονομικά δεδομένα. Τα αποτελέσματα είναι αρκετά εντυπωσιακά δεδομένου του ότι πρόκειται για μία μονοδιάστατη μέθοδο πρόβλεψης και του ότι κατάφερε να επικρατήσει έναντι σαφώς πιο εξειδικευμένων και πολυδιάστατων μεθόδων και πακέτων.

Τα δεδομένα, καθώς και περισσότερες πληροφορίες για το διαγωνισμό είναι διαθέσιμες στην ιστοσελίδα του *International Institute of Forecasters* (<http://forecasters.org/resources/time-series-data/m3-competition>).

5.1.2 Διαγωνισμοί Προβλέψεων Neutral Networks

Πρόκειται για μία οικογένεια τριών διαγωνισμών (*NN3*, *NN4* και *NN5*) οι οποίοι πραγματοποιήθηκαν το 2006-2007, 2009 και 2008 αντίστοιχα. Υπεύθυνος για την οργάνωση των διαγωνισμών ήταν μαζί με άλλους ακαδημαϊκούς ο S.Crone από το Lancaster University της Βρετανίας και κύριος στόχος ήταν η ανάδειξη μεθόδων πρόβλεψης που δεν βασίζονταν σε αμιγώς μαθηματικές διαδικασίες, όπως γινόταν στην πλειοψηφία των περιπτώσεων στους διαγωνισμούς *M*. Σε αυτόν μπορούσαν να συμμετάσχουν αναδρομικά νευρωνικά δίκτυα (*recurrent neural networks*), εκτιμητές ασαφούς λογικής (*fuzzy logic*), γενετικοί αλγόριθμοι (*genetic algorithms*), δέντρα παλινδρόμησης και απόφασης (*classification & regression*) και φυσικά υβριδικά μοντέλα (*hybrid models*).

Όλα τα δεδομένα του διαγωνισμού αναφέρονταν στον κλάδο των μεταφορών (κίνηση σε αυτοκινητόδρομους, κίνηση αυτοκινήτων σε σήραγγες, κυκλοφορία σε αυτόματα συστήματα πληρωμών στις εθνικές οδούς, κυκλοφορία ανθρώπων στα συστήματα του μετρό, αεροπορικά εισιτήρια σε εγχώρια αεροσκάφη, εισαγωγές ναυτιλίας, διέλευση συνόρων, ροές αγωγών και τις σιδηροδρομικές μεταφορές) και διέθεταν διαφορετικό πλήθος παρατηρήσεων και ημερομηνία εκκίνησης. Αναφέρονταν επίσης σε διαφορετικές συχνότητες (ετήσια, τριμηνιαία, μηνιαία, εβδομαδιαία, ημερήσια και ωριαία δεδομένα) και έφταναν το σύνολό τους τις 11 ανά κατηγορία (11 χρονοσειρές * 3 διαγωνισμοί * 6 συχνότητες = 198). Οι οριζόντες πρόβλεψης ανά συχνότητα έφταναν τις 6, 8, 12, 26, 14 και 48 παρατηρήσεις αντίστοιχα.

Πλεονέκτημα του διαγωνισμού ήταν η εξερεύνηση νέων μεθόδων σε μεγαλύτερο εύρος συχνοτήτων και κατηγορίες δεδομένων. Ωστόσο μεγάλο μειονέκτημά του ήταν το σαφώς μικρότερο πλήθος χρονοσειρών σε σύγκριση με τους διαγωνισμούς *M*, το οποίο μπορεί βέβαια να δικαιολογηθεί λόγω του υψηλού υπολογιστικού κόστους που απαιτούν τα μοντέλα μηχανικής μάθησης.

Τα δεδομένα, καθώς και περισσότερες πληροφορίες για το διαγωνισμό είναι διαθέσιμες στην ιστοσελίδα του διαγωνισμού (<http://www.neural-forecasting-competition.com/index.htm>).

5.1.3 Διαγωνισμός T

Ο εν λόγω διαγωνισμός οργανώθηκε το 2006 από το πανεπιστήμιο *INSEAD* με υπεύθυνη την M. Hibon (*Hibon et. al. 2006*) και είχε ως σκοπό τη συνέχιση της έρευνας των αποτελεσμάτων των διαγωνισμών *M* που προηγήθηκαν. Ο διαγωνισμός, πέρα των κλασικών κατηγοριών χρονοσειρών του *M3*, διέθετε και δεδομένα από τον κλάδο των μεταφορών που επεκτείνονταν μάλιστα σε υψηλότερες συχνότητες πέραν των συνηθισμένων (εισαγωγή ημερήσιων, ωριαίων και εβδομαδιαίων δεδομένων). Το σύνολο των χρονοσειρών έφτανε τις 368, με τον οριζόντα πρόβλεψης να εκτείνεται ανάλογα τη συχνότητα στις 6, 8, 12, 26, 14 και 48 περιόδους για ετήσια, τριμηνιαία, μηνιαία, εβδομαδιαία, ημερήσια και ωριαία δεδομένα αντίστοιχα.

Πίνακας 5.5: Πλήθος χρονοσειρών διαγωνισμού T ανά συχνότητα και κατηγορία.

Συχνότητα	Πλήθος	Μήκος
Ετήσια	37	30
Τριμηνιαία	42	102
Μηνιαία	161	122
Εβδομαδιαία	42	466
Ημερήσια	40	568
Ωριαία	46	1204
Σύνολο	368	

Πέραν από τις ανάλογες προβλέψεις, ο διαγωνισμός ζητούσε από τους συμμετέχοντες την κατάθεση των αντίστοιχων διαστημάτων εμπιστοσύνης σημαντικότητας 95%. Επίσης, για λόγους διευκόλυνσης, οι οργανωτές είχαν παραχωρήσει τους λόγους εποχιακότητας όλων των χρονοσειρών. Κάθε διαγωνιζόμενος μπορούσε να καταθέσει προβλέψεις για μεμονωμένα σετ δεδομένων ή προαιρετικά και για το σύνολο των χρονοσειρών. Η μέθοδος θ συμμετείχε στο διαγωνισμό στο κομμάτι των μηνιαίων χρονοσειρών, εκεί δηλαδή όπου στο διαγωνισμό $M3$ είχε σημειώσει και τη βέλτιστη απόδοση. Τα αποτελέσματα του διαγωνισμού δεν έχουν γίνει διαθέσιμα ακόμα δυστυχώς.

5.1.4 Προβληματισμός πάνω στα δεδομένα των διαγωνισμών προβλέψεων

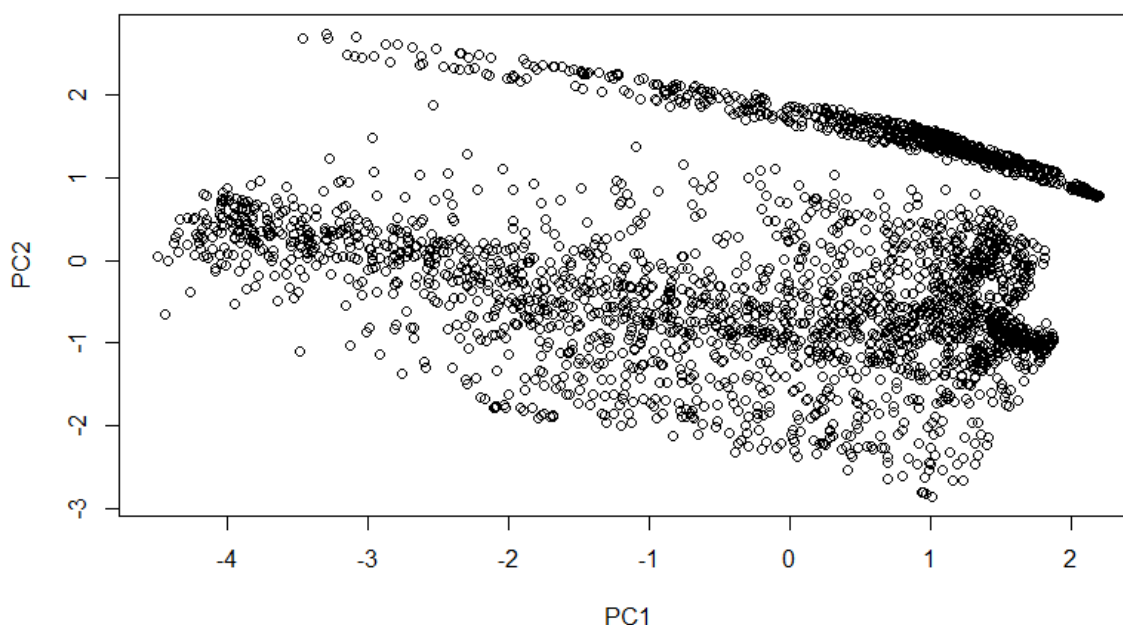
Οι διαγωνισμοί προβλέψεων οδήγησαν σε ιδιαίτερα χρήσιμα εμπειρικά συμπεράσματα σχετικά με την επίδοση των μοντέλων πρόβλεψης τα οποία γίνονται σήμερα ευρέως αποδεκτά από την επιστημονική κοινότητα και αποτελούν σημείο αναφοράς. Μάλιστα, προκειμένου τα εν λόγω συμπεράσματα να είναι όσον το δυνατό πιο αντιπροσωπευτικά, η απόδοση των μεθόδων δεν μελετήθηκε μόνο συνολικά αλλά και ανά κατηγορία, συχνότητα και χαρακτηριστικό χρονοσειράς. Ενδεικτικά παραδείγματα αποτελούν τα εξής:

- Η επίδοση μίας μεθόδου πρόβλεψης ενδέχεται να διαφοροποιείται σημαντικά ανάλογα με το δείκτη σφάλματος που χρησιμοποιείται.
- Η επίδοση μίας μεθόδου πρόβλεψης ενδέχεται να διαφοροποιείται σημαντικά ανάλογα με τη συχνότητα ή την κατηγορία της χρονοσειράς που καλείται να προβλέψει.
- Όσο αυξάνεται ο ορίζοντας πρόβλεψης, τόσο μειώνεται η επίδοση των μοντέλων πρόβλεψης.
- Απλές και πολύπλοκες μέθοδοι αναμένεται να έχουν παρεμφερή επίδοση όταν οι προβλέψεις βρίσκονται κοντά στις πιο πρόσφατες τιμές της χρονοσειράς. Αντίθετα, όταν αυξάνει ο ορίζοντας οι πιο πολύπλοκες μέθοδοι αναμένεται να αποδίδουν καλύτερα.
- Στην περίπτωση εποχιακών χρονοσειρών απαιτείται αποεποχικοποίηση των δεδομένων ή χρήση ενός εποχιακού μοντέλου πρόβλεψης. Επιπλέον, η επιλεκτική αποεποχικοποίηση ανάλογα με τη συστηματικότητα των περιοδικών διακυμάνσεων είναι προτιμότερη από την α -priori επεξεργασία τους.

- Χρονοσειρές με έντονη τάση θα πρέπει να προβλέπονται από μοντέλα τα οποία τη λαμβάνουν επαρκώς υπόψη. Η επίδοση τέτοιων μοντέλων αναμένεται μάλιστα να βελτιώνεται σημαντικά σε σύγκριση με άλλα που την αγνοούν στην περίπτωση των ετήσιων χρονοσειρών και των μεγαλύτερων οριζόντων πρόβλεψης, εκεί δηλαδή όπου και η συνιστώσα της τάσης υπερέρχει έναντι αυτής του θορύβου και της εποχιακότητας.
- Η αύξηση της διακύμανσης των δεδομένων αυξάνει το σφάλμα της πρόβλεψης. Χρονοσειρές με ιδιαίτερα υψηλή διακύμανση είναι προτιμότερο να προβλέπονται από απλές μεθόδους, ενώ χρονοσειρές με χαμηλά επίπεδα θορύβου από πιο σύνθετες μεθόδους.
- Η αύξηση του επιπέδου της τάσης και της εποχιακότητας μπορούν να οδηγήσουν σε μείωση της επίδοσης των μοντέλων πρόβλεψης.

Ωστόσο, παρά τη συστηματικότητα των συμπερασμάτων των διαγωνισμών, αρκετοί είναι εκείνοι που ισχυρίζονται ότι αυτά βασίζονται σε αποτελέσματα που χαρακτηρίζονται από αρκετή τυχαιότητα καθώς συνδέονται άμεσα με τα δεδομένα που χρησιμοποιήθηκαν στην εκάστοτε περίπτωση (*Ord, 2001*). Θεωρούν δηλαδή πως αφενός το πλήθος των χρονοσειρών που χρησιμοποιήθηκαν στους διαγωνισμούς δεν επαρκεί για την εξαγωγή ασφαλών συμπερασμάτων γενικής εμβέλειας και αφετέρου ότι το δείγμα των χρονοσειρών δεν είναι σε καμία περίπτωση αντιπροσωπευτικό ως προς τα χαρακτηριστικά και τις ιδιότητες των χρονοσειρών που περιλαμβάνει. Μάλιστα, σε σχόλια που έχουν γίνει για το διαγωνισμό *M3*, αναφέρθηκε μεταξύ άλλων πως, ακόμα και για τις κατηγορίες δεδομένων που εξετάστηκαν, ενδέχεται οι χρονοσειρές που επιλέχθηκαν να μην είναι επαρκώς αντιπροσωπευτικές (*Lawrence, 2001*).

Οι παραπάνω ισχυρισμοί δεν μπορούν να θεωρηθούν αβάσιμοι καθώς οι χρονοσειρές των διαγωνισμών πράγματι διαμοιράζονται ανομοιόμορφα ανά κατηγορία, συχνότητα και χαρακτηριστικό αποδίδοντας έτσι πιθανώς πλεονέκτημα σε συγκεκριμένες μεθόδους ικανές να προβλέψουν ικανοποιητικά την πλειοψηφία τους. Αυτό είναι και το βασικό συμπέρασμα στο οποίο κατέληξαν πρόσφατα οι *Hyndman et al. (2016)* οι οποίοι πραγματοποίησαν μία διεξοδική ανάλυσή στα δεδομένα του διαγωνισμού *M3* και απέδειξαν πως αυτός περιλαμβάνει χρονοσειρές με πολύ συγκεκριμένα χαρακτηριστικά και μάλιστα ανομοιόμορφα κατανεμημένα στο δείγμα (π.χ. συχνότητα, τάση, τυχαιότητα, εποχιακότητα, στασιμότητα και αυτοσυσχέτιση). Μάλιστα, ποσοτικοποιώντας τα εν λόγω χαρακτηριστικά ως προς την έντασή τους και ομαδοποιώντας τα κατάλληλα σε τεχνητές μεταβλητές (components) με χρήση της τεχνικής *Principal Components Analysis (Hotelling, 1933)* κατάφεραν να τοποθετήσουν κάθε χρονοσειρά του διαγωνισμού στο δισδιάστατο χώρο δείχνοντας έτσι πως μεγάλο μέρος αυτού παραμένει κενός (*Kang et al., 2017*). Το παραπάνω αποτέλεσμα παρουσιάζεται στην *Εικόνα 5.1* όπου τα πρώτα δύο components της ανάλυσης, τα οποία και εξηγούν το 67.67% της συνολικής διακύμανσης, χρησιμοποιούνται για την απεικόνιση των χρονοσειρών του διαγωνισμού. Όπως φαίνεται, στο δισδιάστατο χώρο που σχηματίζεται οι χρονοσειρές του *M3* συγκεντρώνονται σε συγκεκριμένες θέσεις και κάποιες άλλες παραμένουν παντελώς άδειες ή αραιοκατοικημένες.



Εικόνα 5.1: Δυσδιάστατη παρουσίαση των δεδομένων του διαγωνισμού M3 βάσει χαρακτηριστικών τους μέσω της μεθοδολογίας των Kang et al., 2017.

Πιο συγκεκριμένα, ανατρέχοντας στη μελέτη των [Kang et al. \(2017\)](#) συμπεραίνουμε βάσει της παραπάνω απεικόνισης πως το μεγαλύτερο πλήθος των χρονοσειρών του M3 χαρακτηρίζεται από χαμηλή τυχαιότητα (μειώνεται κατά τον άξονα PC1), αυξημένη τάση και αυτοσυσχέτιση (αυξάνεται κατά τον άξονα PC1) και σημαντική εποχιακότητα (μειώνεται κατά τον άξονα PC2), όταν βέβαια αναφερόμαστε σε εποχιακές χρονοσειρές.

Φυσικά οι διαγωνισμοί δεν παύει να έχουν αξία και να δίνουν μία αίσθηση της πραγματικότητας καθώς αναφέρονται σε πραγματικά δεδομένα και σε χρονοσειρές επιχειρήσεων που είναι περισσότερο πιθανό να κληθούμε να προεκτινουμε. Ωστόσο το ερώτημα σχετικά με το αν τα συμπεράσματα των διαγωνισμών θα διατηρούνταν αν π.χ. αυξανόταν περαιτέρω το πλήθος των εποχιακών χρονοσειρών ή αν μειωνόταν ο αριθμός των χρονοσειρών με σημαντική τάση, παραμένει αναπάντητο.

Σε αυτήν την κατεύθυνση οι [Koning et al. \(2005\)](#) πραγματοποίησαν μία μελέτη στα αποτελέσματα του διαγωνισμού M3 προκειμένου να διαπιστώσουν αν τα κομβικά συμπεράσματά του ήταν στατιστικά σημαντικά. Τα αποτελέσματα έδειξαν ότι πράγματι κάποιες μέθοδοι αποδίδουν σημαντικά καλύτερα από κάποιες άλλες παράγοντας αρκετά διαφοροποιημένες προβλέψεις. Επίσης φάνηκε πως δεν υπήρχε κάποια σχέση μεταξύ της πολυπλοκότητας της μεθόδου πρόβλεψης και της ακρίβειας που επιτυγχανόταν ενώ υπήρχε συσχέτιση των αποτελεσμάτων για τους διάφορους δείκτες σφαλμάτων που χρησιμοποιήθηκαν. Τέλος, η αίσθηση πως ο συνδυασμός μεθόδων πρόβλεψης αποδίδει καλύτερα από τις επιμέρους μεθόδους δεν επιβεβαιώθηκε. Αργότερα, οι [Gorr και Schneider \(2013\)](#) πραγματοποιώντας μελέτη στα μικροοικονομικά μηνιαία δεδομένα του

διαγωνισμού κατέληξαν στο συμπέρασμα ότι οι πολύπλοκες μέθοδοι είναι ακριβέστερες καθώς και ότι ο συνδυασμός τέτοιων είναι πιθανότερο να οδηγήσει σε βελτιωμένη προβλεπτική ακρίβεια. Όπως γίνεται κατανοητό, τα παραπάνω αντικρουόμενα αποτελέσματα ενίσχυσαν τους προβληματισμούς γύρω από τους διαγωνισμούς προβλέψεων καθώς κάποια από τα συμπεράσματα στα οποία αυτοί είχαν οδηγήσει βρέθηκαν να μην επιβεβαιώνονται για θεωρητικά μεγάλο πλήθος χρονοσειρών ή για συγκεκριμένα υποσύνολα χρονοσειρών.

Προκειμένου λοιπόν να τεκμηριωθούν τα συμπεράσματα της παρούσας διατριβής αλλά και να εμπλουτιστούν πιθανώς εξετάζοντας τη συμπεριφορά των υπό μελέτη μεθόδων σε μεγάλο και ποικιλόμορφο πλήθος χρονοσειρών, κρίνεται σκόπιμη η αξιοποίηση πέραν των δεδομένων των διαγωνισμών προβλέψεων και πληθώρας άλλων που καλύπτουν όλο ή το σημαντικότερο εύρος τύπων χρονοσειρών. Για να γίνει κάτι τέτοιο δεδομένης της έλλειψης πραγματικών χρονοσειρών, δημιουργείται η ανάγκη ανάπτυξης μίας γεννήτριας τεχνητών χρονοσειρών. Ο τρόπος με τον οποίο θα γίνει αυτό περιγράφεται στην παράγραφο που ακολουθεί. Σημειώνεται επίσης πως δεν εξετάστηκε η τεκμηρίωση των αποτελεσμάτων μέσω τεχνικών αναδειγματοληψίας καθώς αυτές ναι μεν βοηθούν στην καλύτερη αξιολόγηση των τεχνικών πρόβλεψης περιορίζοντας την επίδραση του παράγοντα της τυχαιότητας, ωστόσο δεν απαντούν στο ερώτημα αν η εκάστοτε μέθοδος πρόβλεψης συμπεριφέρεται καλύτερα σε χρονοσειρές συγκεκριμένων χαρακτηριστικών και ιδιοτήτων.

5.2 Γεννήτρια παραγωγής τεχνητών χρονοσειρών

Προκειμένου να παραχθούν τεχνητές χρονοσειρές με ποικίλα χαρακτηριστικά διαφορετικής έντασης, προαπαιτούμενο αποτελεί ο καθορισμός και η ποσοτικοποίηση αυτών των χαρακτηριστικών. Όπως αναφέρθηκε νωρίτερα, στην μελέτη των *Hyndman et al. (2016)* τα χαρακτηριστικά χρονοσειρών που μελετήθηκαν ήταν η εποχιακότητα, η τάση, η τυχαιότητα, η αυτοσυσχέτιση των παρατηρήσεων και η κανονικότητά τους. Άλλη ενδιαφέρουσα σχετική μελέτη είναι αυτή των *Petropoulos et al. (2014)* οι οποίοι θεώρησαν πως βάσει της κλασικής μεθόδου αποσύνθεσης κάθε χρονοσειρά μπορεί να αναλυθεί στα τέσσερα βασικά ποιοτικά της χαρακτηριστικά, δηλαδή την εποχιακότητα, την τάση, τον κύκλο και την τυχαιότητα, οπότε και κατά αντιστοιχία μπορεί κανείς να τα χρησιμοποιήσει για να παράγει νέες. Επιπλέον παράμετρος ήταν το μήκος της χρονοσειράς, δηλαδή το πλήθος των παρατηρήσεων.

Η τελευταία προσέγγιση φαντάζει αρκετά πιο λογική, αφενός λόγω του ότι βρίσκεται σαφώς πιο κοντά στην ανθρώπινη αίσθηση και αφετέρου του ότι τα εν λόγω χαρακτηριστικά είναι απολύτως ανεξάρτητα μεταξύ τους. Για παράδειγμα, η αυτοσυσχέτιση των παρατηρήσεων μπορεί να αποτελεί σε μη στάσιμες χρονοσειρές ένδειξη τάσης ή εποχιακότητας ενώ η κανονικότητα των παρατηρήσεων, αν και συνδέεται από μόνη της άμεσα με την αποτελεσματικότητα των μοντέλων πρόβλεψης, μπορεί να αποτελεί ένα ακόμα μέτρο ποσοτικοποίησης της έντασης της τυχαιότητας στα δεδομένα. Συμπερασματικά, αν οι χρονοσειρές που μελετώνται δεν είναι ή δε γίνουν στάσιμες στην πορεία, είναι μάλλον επίφοβο να τις αξιολογούμε ως προς κάποιο χαρακτηριστικό πέραν των τεσσάρων που αναφέρθηκαν, ειδικά από τη στιγμή που η διαδικασία

μετασχηματισμού χρονοσειρών σε στάσιμες βασίζεται σε διαγνωστικούς ελέγχους με υψηλά περιθώρια σφάλματος. Έτσι, η γεννήτρια που θα αναπτυχθεί στην παρούσα διατριβή κρίνεται σκόπιμο να στηρίζεται αποκλειστικά στη λογική της κλασικής μεθόδου αποσύνθεσης.

Κάποιες βασικές διαφοροποιήσεις που εξετάζονται ωστόσο σε σχέση με την μελέτη των *Petropoulos et al. (2014)* είναι η εξαίρεση του χαρακτηριστικού της κυκλικότητας και ο διαχωρισμός του τύπου της τάσης που εμπεριέχεται στα δεδομένα. Επιπλέον επιλέγεται όλες οι χρονοσειρές να έχουν ίσο μήκος προκειμένου η όποια διακύμανση στα σφάλματα των μεθόδων που θα εξεταστούν να οφείλεται αποκλειστικά στα δεδομένα των χρονοσειρών και όχι στις παρατηρήσεις που είναι διαθέσιμες για την εκπαίδευσή τους. Η εξαίρεση του παράγοντα της κυκλικότητας έχει να κάνει με το γεγονός ότι το εν λόγω χαρακτηριστικό γίνεται αντιληπτό μόνο σε περιπτώσεις χρονοσειρών εκτενούς ιστορικότητας (20 ή 30 χρόνια). Έτσι, οποιαδήποτε κυκλικότητα υπολογιστεί σε χρονοσειρές μικρότερου μήκους είναι μάλλον καταχρηστική και αντιπροσωπεύει μη ερμηνευθείσες διακυμάνσεις. Όσον αφορά τώρα τον τύπο της τάσης, όπως παρουσιάστηκε διεξοδικά στην *Παράγραφο 3.4.3*, το μοτίβο ανάπτυξης της χρονοσειράς μπορεί να επηρεάσει σημαντικά την επίδοση ενός μοντέλου πρόβλεψης και για αυτόν το σκοπό κρίνεται επιθυμητή η χρήση μοντέλων ικανών να μοντελοποιήσουν διαφορετικούς τύπους της, όπως π.χ. γραμμική, εκθετική, λογαριθμική κ.ο.κ. Ο εν λόγω διαχωρισμός είναι λοιπόν επιτακτικός στα πλαίσια της παρούσας διατριβής, όπου και εισάγεται μία γενικευμένη μορφή της κλασικής μεθόδου θ για την αποδοτικότερη προέκταση οποιουδήποτε μοτίβου χρονοσειράς.

5.2.1 Ανάλυση χαρακτηριστικών χρονοσειράς

Όπως αναφέρθηκε νωρίτερα, προκειμένου να παραχθούν τεχνικά χρονοσειρές, θα αξιοποιηθεί η εξίσωση της κλασικής πολλαπλασιαστικής μεθόδου αποσύνθεσης χωρίς το στοιχείο του κύκλου, δηλαδή η ακόλουθη φόρμουλα:

$$Y_i = T_i * S_i * R_i$$

Η λογική της γεννήτρια βασίζεται όπως φαίνεται στην υπόθεση ότι δεδομένης μίας τάσης, εποχιακότητας και τυχαιότητας συγκεκριμένης έντασης, μπορεί κανείς να παράξει μία νέα χρονοσειρά ως το γινόμενο τους. Φυσικά, ένα καίριο ερώτημα που προκύπτει σε αυτό το σημείο είναι πώς παράγονται οι εν λόγω συνιστώσες καθώς και πώς κατηγοριοποιείται ποιοτικά η έντασή τους.

Σε πρώτη φάση, απαιτείται λοιπόν αφενός η εισαγωγή κάποιων ειδικών δεικτών προκειμένου να ποσοτικοποιηθούν τα εν λόγω χαρακτηριστικά, και αφετέρου η χρήση ενός σετ δεδομένων ως βάση ποιοτικού διαχωρισμού της έντασής τους σε κλίμακες. Μία προφανής επιλογή βάσης είναι τα μηνιαία δεδομένα του διαγωνισμού *M3* καθώς διαθέτουν αρκετά ικανοποιητικό πλήθος και ποικιλία ιδιοτήτων για μία τέτοια διεργασία. Έτσι, αξιοποιώντας τις εκεί χρονοσειρές θα εκτιμηθεί για πραγματικές επιχειρησιακές εφαρμογές ποια π.χ. τάση θεωρείται υψηλής, μεσαίας ή χαμηλής έντασης, πόσο μεταβάλλονται οι δείκτες εποχιακότητας μίας χρονοσειράς εντός του έτους και ποιο είναι το εύρος τυχαιότητας που αναμένεται.

Για την απομόνωσή των παραπάνω χαρακτηριστικών εφαρμόστηκε η κλασική πολλαπλασιαστική μέθοδος αποσύνθεσης όπως παρουσιάστηκε στην *Παράγραφο 2.2* και στη συνέχεια η ένταση κάθε συνιστώσας ποσοτικοποιήθηκε αξιοποιώντας τους παρακάτω δείκτες:

$$\text{Ένταση εποχιακότητας} = \left(\frac{2}{11} \sum_{i=2}^{12} \left| \frac{S_i - S_{i-1}}{S_i + S_{i-1}} \right| \right) * 100\%$$

$$\text{Ένταση τυχαιότητας} = \left(\frac{2}{n-1} \sum_{i=2}^n \left| \frac{R_i - R_{i-1}}{R_i + R_{i-1}} \right| \right) * 100\%$$

$$\text{Ένταση τάσης} = |\text{Κλίση καμπύλης τάσης}|$$

Σημειώνεται ότι προκειμένου οι τιμές των δεικτών να είναι ανεξάρτητες της κλίμακας της χρονοσειράς και να εξαχθούν αντιπροσωπευτικά συμπεράσματα σχετικά με την ένταση της κάθε συνιστώσας στην εκάστοτε χρονοσειρά, τα αρχικά δεδομένα κανονικοποιούνται διαιρώντας τις τιμές τους με την ελάχιστη τιμή της χρονοσειράς. Επίσης, για να υπολογιστεί η ένταση της τάσης ανά μοτίβο ανάπτυξης, στο *6^ο Βήμα* της μεθοδολογίας της κλασικής αποσύνθεσης η γραμμική τάση αντικαθίσταται με εναλλακτικές καμπύλες ως εξής:

- Γραμμική τάση (**LIN**): $T_t = b + at$
- Εκθετική τάση (**EXP**): $T_t = be^{at}$
- Λογαριθμική τάση (**LOG**): $T_t = b + a \log(t)$
- Αντίστροφη τάση (**INV**): $T_t = b + a/t$
- Τάση σε μορφή δύναμης (**POW**): $T_t = bt^a$

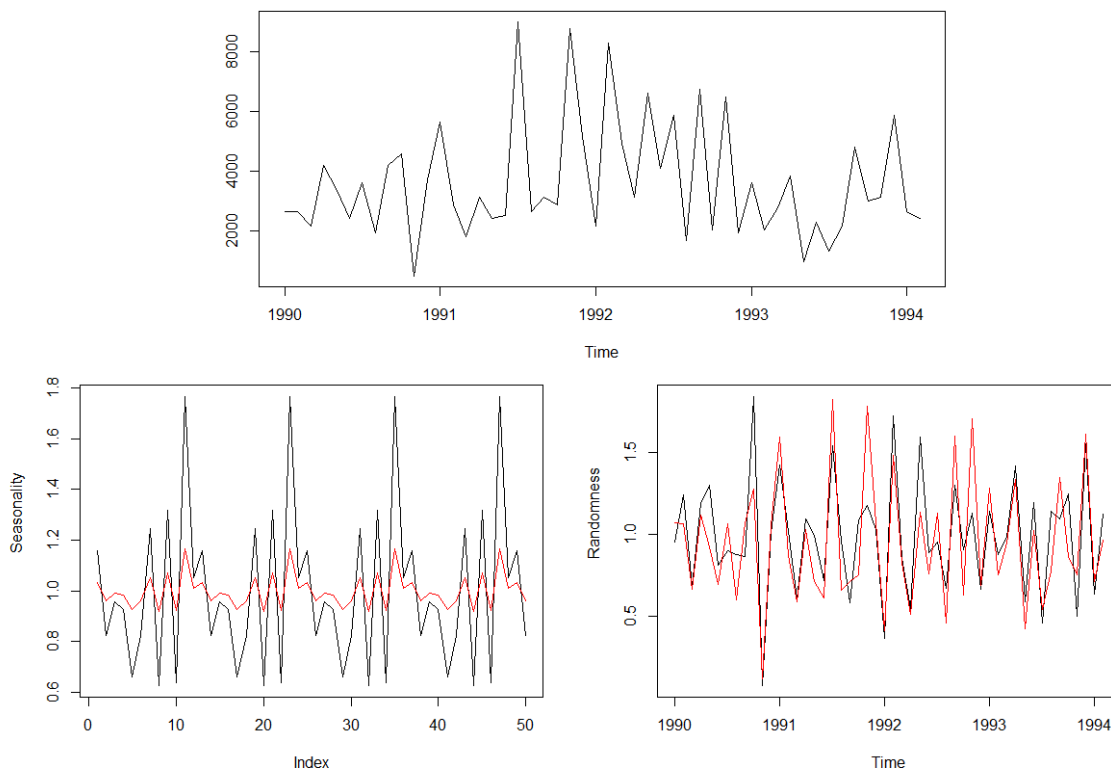
Όπως γίνεται κατανοητό, επειδή όλες οι χρονοσειρές έχουν την ίδια κλίμακα, η παράμετρος b δεν έχει καμία αξία για την αξιολόγηση της συνιστώσας της τάσης και η έντασή της μπορεί να εκτιμηθεί άμεσα ανά μοτίβο συγκρίνοντας απευθείας τις απόλυτες τιμές της αντίστοιχης κλίσης a . Τονίζεται βέβαια σε αυτό το σημείο πως μέσω αυτής της προσέγγισης ενώ μπορούμε να αποφανθούμε π.χ. μεταξύ δύο χρονοσειρών γραμμικής τάσης ποια παρουσιάζει την ταχύτερη ανάπτυξη, το ίδιο δεν είναι εφικτό για δύο χρονοσειρές διαφορετικού μοτίβου ανάπτυξης καθώς οι αντίστοιχοι συντελεστές a προκύπτουν από εντελώς διαφορετικές εξισώσεις. Έτσι μπορούμε να μιλάμε μόνο για ένταση τάσης ανά μοτίβο ανάπτυξης.

Τέλος, επισημαίνεται πως προκειμένου οι δείκτες εποχιακότητας να είναι όσον το δυνατό πιο αντιπροσωπευτικοί, για την εκτίμησή τους εφαρμόζεται επιπλέον η μέθοδος συρρίκνωσης συντελεστών *Miller-Williams*. Αυτό συμβαίνει καθώς, αν εφαρμόσει κανείς καταχρηστικά την κλασική αποσύνθεση σε μη εποχιακές χρονοσειρές χωρίς τη χρήση αντίστοιχου τεστ εποχιακότητας, ένα μεγάλο μέρος του θορύβου θα λογιστεί ως εποχιακότητα οδηγώντας στο εσφαλμένο συμπέρασμα ότι η εν λόγω χρονοσειρά είναι κατά ένα σημαντικό παράγοντα εποχιακή. Αυτό δεν συμβαίνει στην περίπτωση της

Κεφάλαιο 5: Παραγωγή τεχνητών χρονοσειρών μέσω προσομοίωσης

συρρίκνωσής καθώς ελέγχεται η σημαντικότητα των συντελεστών και στη συνέχεια αυτοί συρρικνώνονται αναλόγως.

Ένα χαρακτηριστικό παράδειγμα παρουσιάζεται στην *Εικόνα 5.2* όπου η κλασική μέθοδος αποσύνθεσης εφαρμόζεται σε εμφανώς μη εποχιακή χρονοσειρά αποδίδοντας δείκτες μεγάλης διακύμανσής, μία βασική ένδειξη εποχιακότητας. Για το ίδιο παράδειγμα παρουσιάζεται το αποτέλεσμα της αποεποχικοποίησης με τη μέθοδο *Miller-Williams* όπου οι νέοι δείκτες βρίσκονται όλοι αρκετά κοντά στη μονάδα. Οι νέοι συντελεστές είναι λοιπόν σαφώς πιο αντιπροσωπευτικοί και δεν χαρακτηρίζουν εσφαλμένα τη χρονοσειρά ως προς τα χαρακτηριστικά της. Στο κάτω δεξιά διάγραμμα φαίνεται επιπλέον η νέα συνιστώσα τάσης που υπολογίστηκε, η οποία στη δεύτερη περίπτωση λόγω της έλλειψης σημαντικής εποχιακότητας είναι αρκετά πιο έντονη, όπως και είναι το επιθυμητό. Ενδεικτικά αναφέρεται πως στην περίπτωση της κλασικής αποσύνθεσης οι δείκτες εποχιακότητας και τυχαιότητας θα λάμβαναν σύμφωνα με τις σχέσεις που παρουσιάστηκαν νωρίτερα τιμές 45.3% και 53.0%, ενώ στην περίπτωση της συρρίκνωσης 10.2% και 58.8% αντίστοιχα.



Εικόνα 5.2: Εφαρμογή διαδικασίας συρρίκνωσης συντελεστών εποχιακότητας Miller-Williams (κόκκινο) για την αποτελεσματικότερη αποτύπωση των χαρακτηριστικών της χρονοσειράς σε σχέση με την κλασική προσέγγιση (μαύρο).

Από εκεί και πέρα ο ποιοτικός διαχωρισμός της έντασης των τριών χαρακτηριστικών μπορεί να πραγματοποιηθεί εύκολα βάσει της κατανομής της τιμής τους στο σύνολο των 1428 μηνιαίων χρονοσειρών του διαγωνισμού. Στα πλαίσια της παρούσας εργασίας οι χρονοσειρές διαχωρίζονται ποιοτικά σε τρεις κατηγορίες προκειμένου αφενός η κατηγοριοποίηση να βρίσκεται σε αναλογία με την κοινή αίσθηση και αφετέρου οι χρονοσειρές που θα δημιουργηθούν να διαφέρουν σημαντικά μεταξύ τους ανά κατηγορία, όχι όμως ιδιαίτερα εντός αυτής. Πιο συγκεκριμένα η κατηγοριοποίηση πραγματοποιείται ελέγχοντας την κατανομή του εκάστοτε δείκτη έντασης ως εξής:

- 0-33% του δείγματος: Χρονοσειρές χαμηλής έντασης (*Low intensity*)
- 33-66% του δείγματος: Χρονοσειρές μεσαίας έντασης (*Medium intensity*)
- 66-100% του δείγματος : Χρονοσειρές υψηλής έντασης (*High intensity*)

Βάσει αυτού του διαχωρισμού μπορούμε να κατηγοριοποιήσουμε εύκολα οποιαδήποτε χρονοσειρά ανάλογα με το επίπεδο έντασης των συνιστωσών της χρησιμοποιώντας τον συμβολισμό $Trend(Z,Z,Z)$, όπου ο όρος $Trend$ αντιστοιχεί στον τύπο τάσης της χρονοσειράς και τα ορίσματα Z στην ένταση των συνιστωσών Τάσης, Εποχιακότητας και Τυχαιότητας αντίστοιχα. Έτσι, ο συμβολισμός $LOG(M,L,H)$ συμβολίζει μία χρονοσειρά λογαριθμικού μοτίβου τάσης μεσαίας έντασης, χαμηλής έντασης εποχιακότητας και υψηλού επιπέδου τυχαιότητας.

Σημειώνεται πως προκειμένου να μην υπάρξουν λογικά σφάλματα και ασυνέχειες στα εν λόγω διαστήματα ως ελάχιστη τιμή θεωρείται πάντα το μηδέν ανεξαρτήτως τιμής δείγματος, ενώ ως μέγιστη αυτή του 95% του δείγματος.

Αναφέρεται επίσης ότι με τον όρο «33% του δείγματος» δεν αναφερόμαστε στην τιμή που λαμβάνει ο δείκτης για το πρώτο 33% της κατανομής των χρονοσειρών του διαγωνισμού ταξινομημένο κατά αύξουσα σειρά βάσει της τιμής της αντίστοιχης συνιστώσας, αλλά στο 1/3 της απόλυτης διαφοράς μέγιστης-ελάχιστης τιμής του δείκτη στο δείγμα υποθέτοντας πως αυτή ακολουθεί μία κατανομή Poisson. Έτσι, για την περίπτωση π.χ. της εποχιακότητας, η οποία λαμβάνει μεν τιμές από 0% έως 68% αλλά φυσιολογικές μπορούν να θεωρηθούν μόνο όσες είναι μικρότερες του 18%, τα αντίστοιχα όρια είναι 0-5.97%, 5.98-11.95% και 11.96-18.1%. Οι εν λόγω τιμές παρουσιάζονται αναλυτικά ανά συνιστώσα στον [Πίνακα 5.6](#).

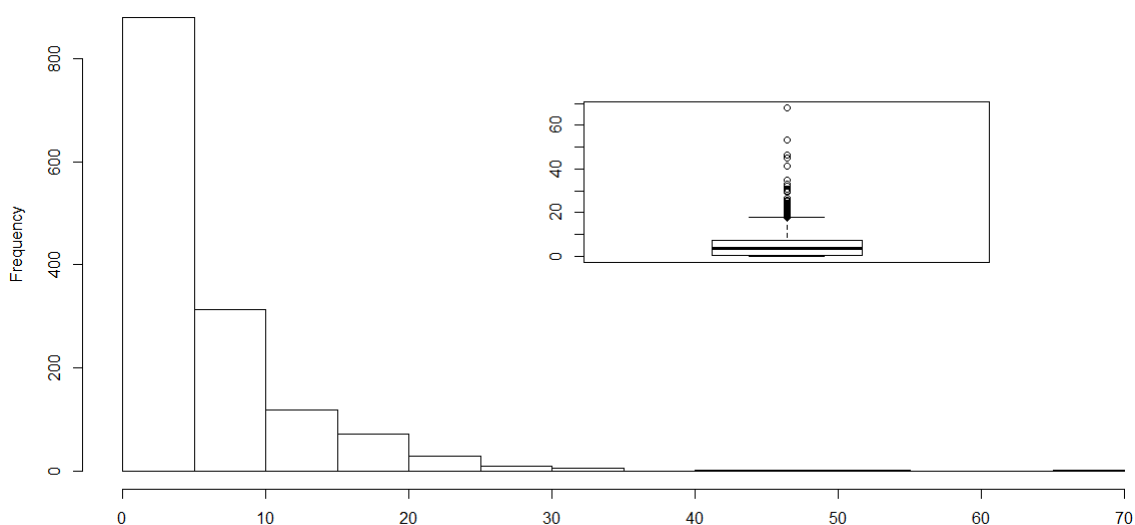
Ο λόγος για τον οποίο έγιναν οι παραπάνω επιλογές είναι προκειμένου τα όρια που θα τεθούν για κάθε κατηγορία να μην επηρεαστούν από το πλήθος των χρονοσειρών της βάσης που αντιστοιχίζονται σε κάθε τιμή δείκτη καθώς η ανομοιόμορφη κατανομή των δεδομένων στα διάφορα χαρακτηριστικά χρονοσειρών αποτελεί όπως αναφέρθηκε νωρίτερα από τα βασικότερα μειονεκτήματα των διαγωνισμών προβλέψεων. Από την άλλη μεριά ήταν ταυτόχρονα επιθυμητό τα όρια των δεικτών να απαλλαχθούν από την επίδραση ακραίων περιπτώσεων ούτως ώστε να βρίσκονται όσο το δυνατό εντός λογικών πλαισίων. Το φαινόμενο της ανομοιομορφίας, όπως και τα εύρη των τιμών που λαμβάνουν οι συνιστώσες που μελετώνται, οπτικοποιούνται στα ιστογράμματα των [Εικόνων 5.3-5.9](#). Χαρακτηριστικό είναι το γεγονός ότι η συντριπτική πλειοψηφία των χρονοσειρών του διαγωνισμού διαθέτει χαρακτηριστικά χαμηλής έντασης.

Κεφάλαιο 5: Παραγωγή τεχνητών χρονοσειρών μέσω προσομίωσης

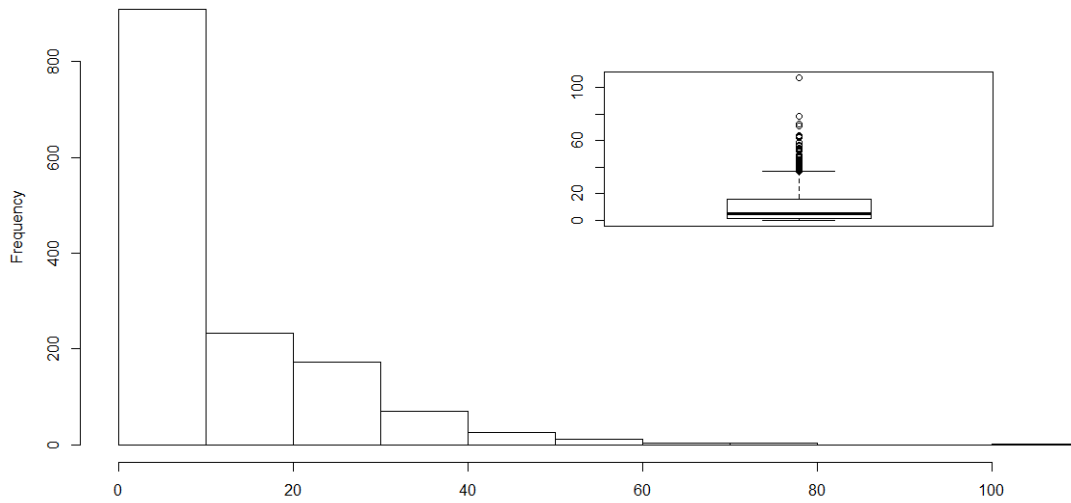
Πίνακα 5.6: Όρια δεικτών για ποιοτικό διαχωρισμό των χρονοσειρών βάσει της έντασης των δομικών τους χαρακτηριστικών.

Χαρακτηριστικό	Χαμηλής έντασης	Μεσαίας έντασης	Υψηλής ένταση
Εποχιακότητα	6.034	12.068	18.102
Τυχαιότητα	11.368	22.737	34.105
Γραμμική τάση	0.025	0.049	0.074
Εκθετική τάση	0.006	0.011	0.017
Λογαριθμική τάση	0.634	1.268	1.902
Αντίστροφη τάση	2.599	5.197	7.796
Τάση σε μορφή δύναμης	0.129	0.258	0.388

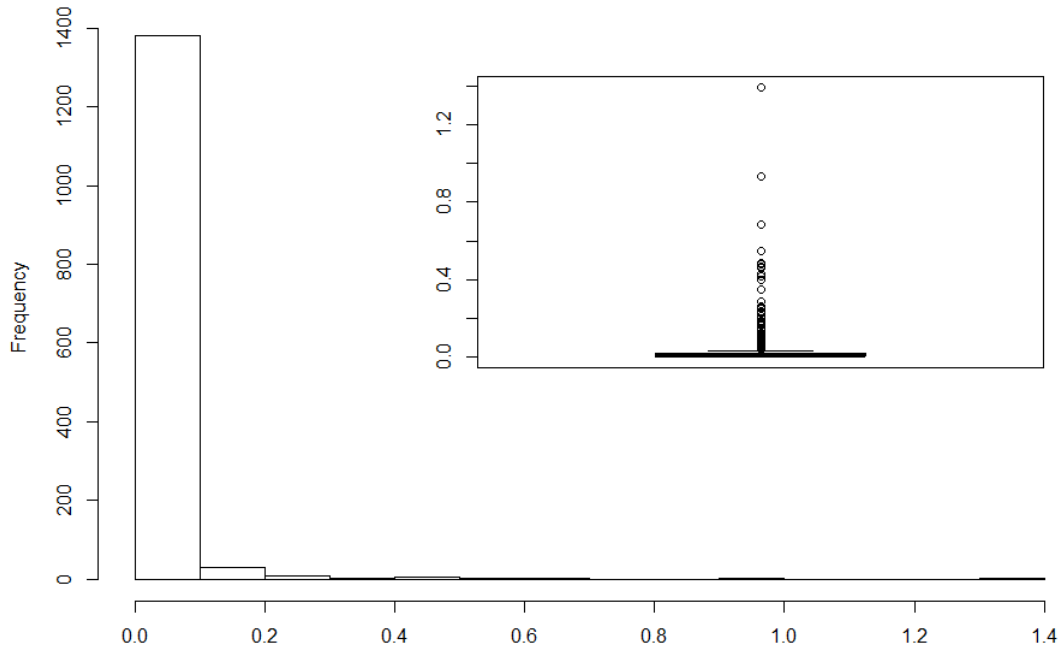
Για λόγους πληρότητας στα δεξιά καθενός από τα ιστογράμματα παρουσιάζεται και το *boxplot* της αντίστοιχης κατανομής. Στην περίπτωση π.χ. της εποχιακότητας, όπως φαίνεται, ο μεγαλύτερος όγκος των χρονοσειρών φτάνει σε ένταση έως και 30%. Από εκεί και πέρα οποιαδήποτε εποχιακότητα μεγαλύτερης έντασης λογίζεται ως ανωμαλία.



Εικόνα 5.3: Κατανομή των χρονοσειρών του διαγωνισμού M3 βάσει έντασης εποχιακότητας.

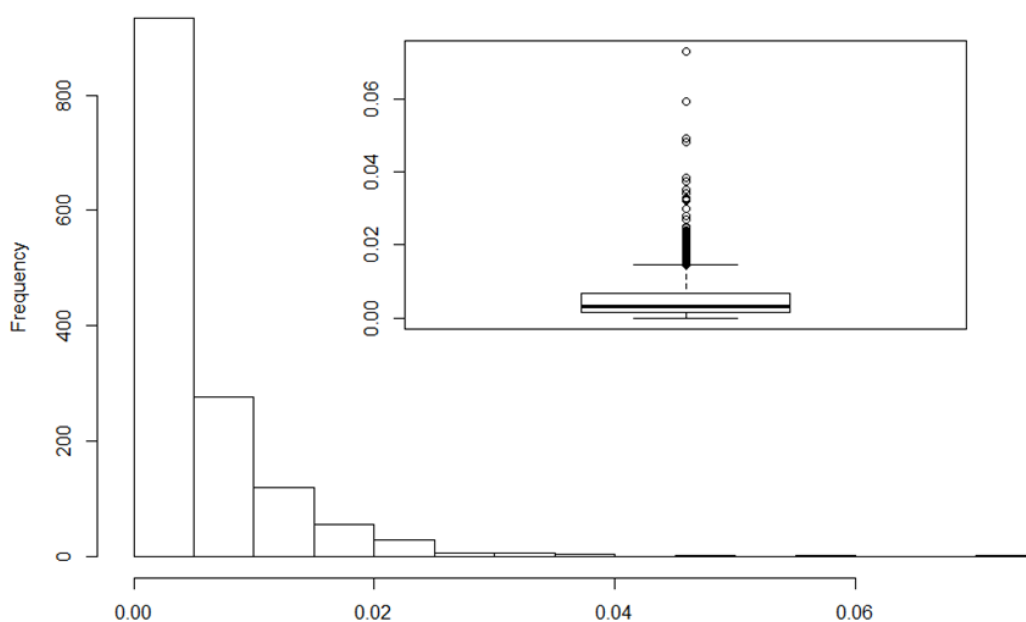


Εικόνα 5.4: Κατανομή των χρονοσειρών του διαγωνισμού M3 βάσει έντασης τυχαιότητας.

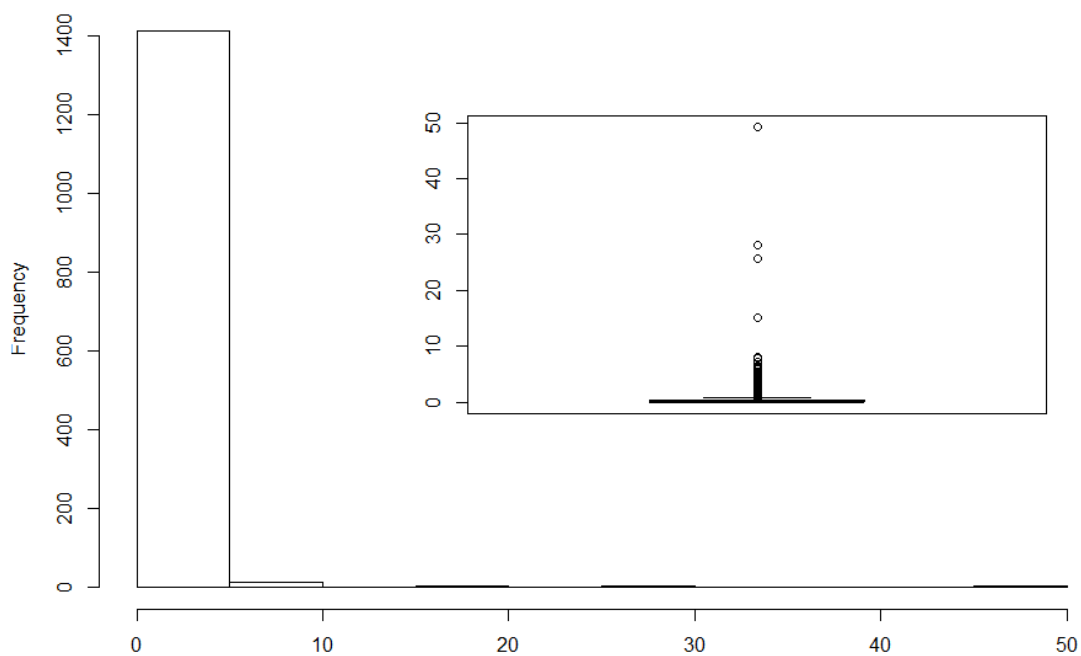


Εικόνα 5.5: Κατανομή κλίσης χρονοσειρών γραμμικού μοτίβου στα μηνιαία δεδομένα του διαγωνισμού M3.

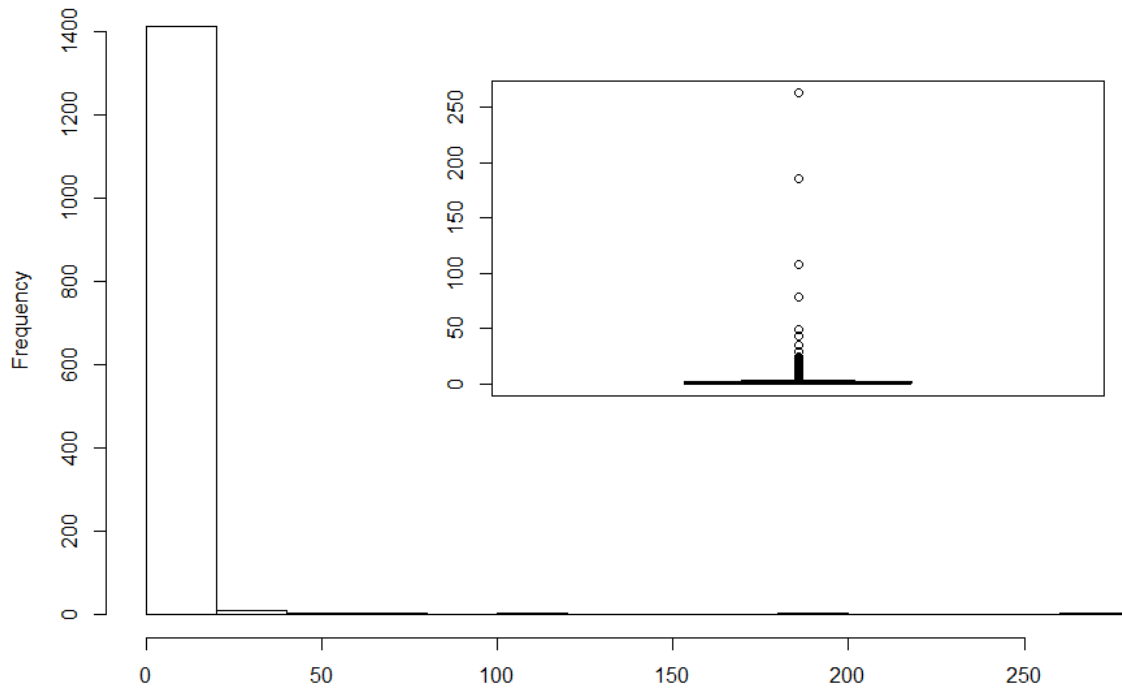
Κεφάλαιο 5: Παραγωγή τεχνητών χρονοσειρών μέσω προσομοίωσης



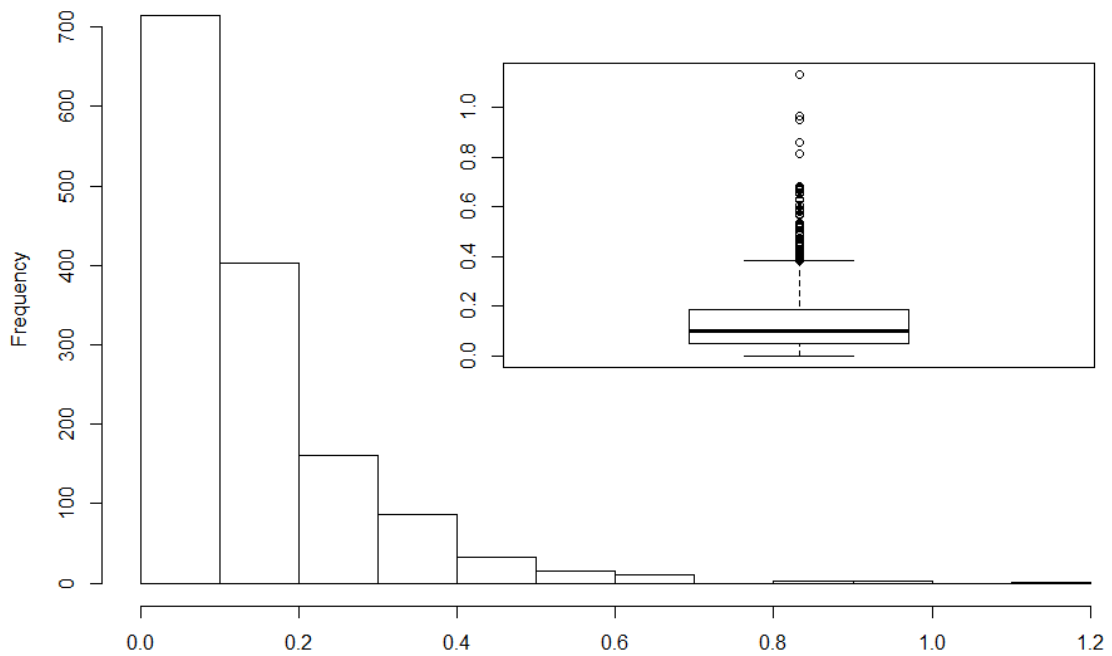
Εικόνα 5.6: Κατανομή κλίσης χρονοσειρών εκθετικού μοτίβου στα μηνιαία δεδομένα του διαγωνισμού M3.



Εικόνα 5.7: Κατανομή κλίσης χρονοσειρών λογαριθμικού μοτίβου στα μηνιαία δεδομένα του διαγωνισμού M3.



Εικόνα 5.8: Κατανομή κλίσης χρονοσειρών αντίστροφου μοτίβου στα μηνιαία δεδομένα του διαγωνισμού M3.



Εικόνα 5.9: Κατανομή κλίσης χρονοσειρών μοτίβου δύναμης στα μηνιαία δεδομένα του διαγωνισμού M3.

5.2.2 Μεθοδολογία παραγωγής τεχνητών χρονοσειρών

Έχοντας ορίσει πλέον τα όρια έντασης για κάθε κατηγορίας χαρακτηριστικού, σειρά έχει η παραγωγή των χρονοσειρών. Αυτές αποφασίστηκε να έχουν μήκος 100 παρατηρήσεων καθώς έτσι εξασφαλίζονται επαρκές τιμές για την εκτίμηση του μοντέλου πρόβλεψης χωρίς να αυξάνεται δραματικά το υπολογιστικό κόστος της διαδικασίας πρόβλεψης. Η παραγωγή των τεχνητών χρονοσειρών πραγματοποιείται σε επτά βήματα που παρουσιάζονται παρακάτω. Μάλιστα, για να γίνει ευκολότερα αντιληπτή η διαδικασία παραγωγής χρονοσειρών, δίνεται επιπλέον ως παράδειγμα η εφαρμογή της μεθοδολογίας για κατασκευή μίας χρονοσειράς εκθετικής τάσης υψηλής έντασης, μεσαίας εποχιακότητας και χαμηλού θορύβου, ή αλλιώς $EXP(H,M,L)$.

Βήμα 1^ο:

Επιλέγεται το μοτίβο τάσης και η κατηγορία έντασης της κλίσης. *Π.χ. Στο παράδειγμα εκθετική τάση υψηλής έντασης.*

Βήμα 2^ο:

Παράγεται ένας τυχαίος αριθμός TL (*Level of Trend*) ο οποίος ακολουθεί κανονική κατανομή με μέση τιμή το κέντρο της αντίστοιχης κατηγορίας και διακύμανση ίση με το 30% αυτής. Με αυτόν τον τρόπο καθορίζεται επιτυχώς με τυχαίο τρόπο η ένταση της τάσης της χρονοσειράς η οποία όπως αναφέρθηκε πρέπει να διαφέρει σημαντικά μεταξύ των κατηγοριών αλλά να όχι ιδιαίτερα εντός αυτών. *Στο παράδειγμα θα επιλεγθεί δηλαδή ένας αριθμός από την κατανομή με μέση τιμή το $(0.011+0.017)/2=0.014$ και διακύμανση $0.014/3=0.005$. Έστω $TL=0.017$.*

Βήμα 3^ο:

Ανάλογα με το είδος της τάσης που επιλέχθηκε νωρίτερα παράγεται η αντίστοιχη συνιστώσα τάσης T_t ως εξής:

Γραμμική τάση: $T_t = TL * t + 1$

Εκθετική τάση: $T_t = e^{TL * t}$

Λογαριθμική τάση: $T_t = 1 + TL * \log(t)$

Αντίστροφη τάση: $T_t = 1 + TL/t$

Τάση σε μορφή δύναμης: $T_t = t^{TL}$

*Στο παράδειγμα παράγονται δηλαδή οι τιμές $e^{0.017*1}$, $e^{0.017*2}$, ..., $e^{0.017*100}$*

Βήμα 4^ο:

Εφαρμόζεται ξανά το Βήμα 2 της μεθοδολογίας για την περίπτωση των συνιστωσών εποχιακότητας SL (*Level of Seasonality*) και τυχειότητας RL (*Level of Randomness*). *Στο παράδειγμα θα επιλεγθεί λοιπόν για την περίπτωση της εποχιακότητας μεσαίας*

έντασης ένας αριθμός από την κατανομή με μέση τιμή το $(6.034+12.068)/2=9.051\%$ και διακύμανση $9.051/3=3.017\%$. Έστω $SL=9.70\%$. Αντίστοιχα, για την περίπτωση της τυχαιότητας χαμηλής έντασης ένας αριθμός από την κατανομή με μέση τιμή το $(0+11.368)/2=5.684\%$ και διακύμανση $5.684/3=1.895\%$. Έστω $RL=5.050\%$.

Βήμα 5°:

Παράγονται τυχαία 100 αριθμοί που ακολουθούν την κανονική κατανομή με μέση τιμή ίση με το επίπεδο τυχαιότητας RL και διακύμανση ίση με το ένα τρίτο αυτής. Οι αριθμοί αυτοί προστίθενται ή αφαιρούνται στη συνέχεια τυχαία από τη μονάδα παράγοντας έτσι δείκτες τυχαιότητας R_i με ομοιόμορφη διακύμανση γύρω από την μονάδα εύρους ίσου με RL .

Βήμα 6°:

Ακολουθείται ξανά η διαδικασία του *Βήματος 5* για την περίπτωση των δεικτών εποχιακότητας, με τη διαφορά βέβαια ότι σε αυτή την περίπτωση οι δείκτες που θα παραχθούν θα είναι 12 και θα επαναλαμβάνονται κατά μήκος της χρονοσειράς. Σημειώνεται ότι οι δείκτες εποχιακότητας S_i που υπολογίζονται θα πρέπει να κανονικοποιηθούν προκειμένου να έχουν άθροισμα ίσο με 12 (τη συχνότητα των δεδομένων) και στη συνέχεια να αντιστοιχηθούν στις 100 παρατηρήσεις της χρονοσειράς που κατασκευάζεται ούτως ώστε όλες να διαθέτουν την κατάλληλη τιμή δείκτη εποχιακότητας.

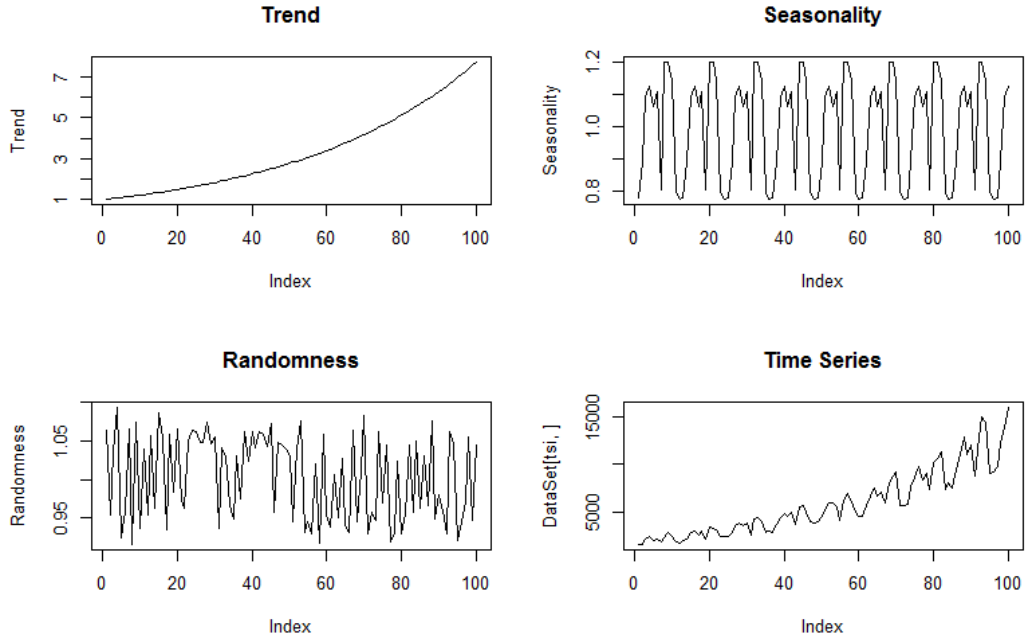
Βήμα 7°:

Οι επιμέρους συνιστώσες T_i , S_i και R_i πολλαπλασιάζονται μεταξύ τους παράγοντας την τελική χρονοσειρά με τα επιθυμητά χαρακτηριστικά. Βέβαια, επειδή το αρχικό της επίπεδο θα ισούται εξορισμού με τη μονάδα, κρίνεται σκόπιμη η θέσπιση ενός νέου επιπέδου ούτως ώστε να μην υπάρχουν ζητήματα που σχετίζονται με αρνητικές τιμές. Για να γίνει αυτό αρκεί ο πολλαπλασιασμός της χρονοσειράς που παράγεται με μία τυχαία θετική τιμή. Προκειμένου να υπάρχει αναλογία με τα δεδομένα του διαγωνισμού $M3$ το εν λόγω επίπεδο L ανήκει στο διάστημα $[1,000-9,000]$. Έτσι, η αρχική εξίσωση μπορεί να γενικευτεί στην ακόλουθη:

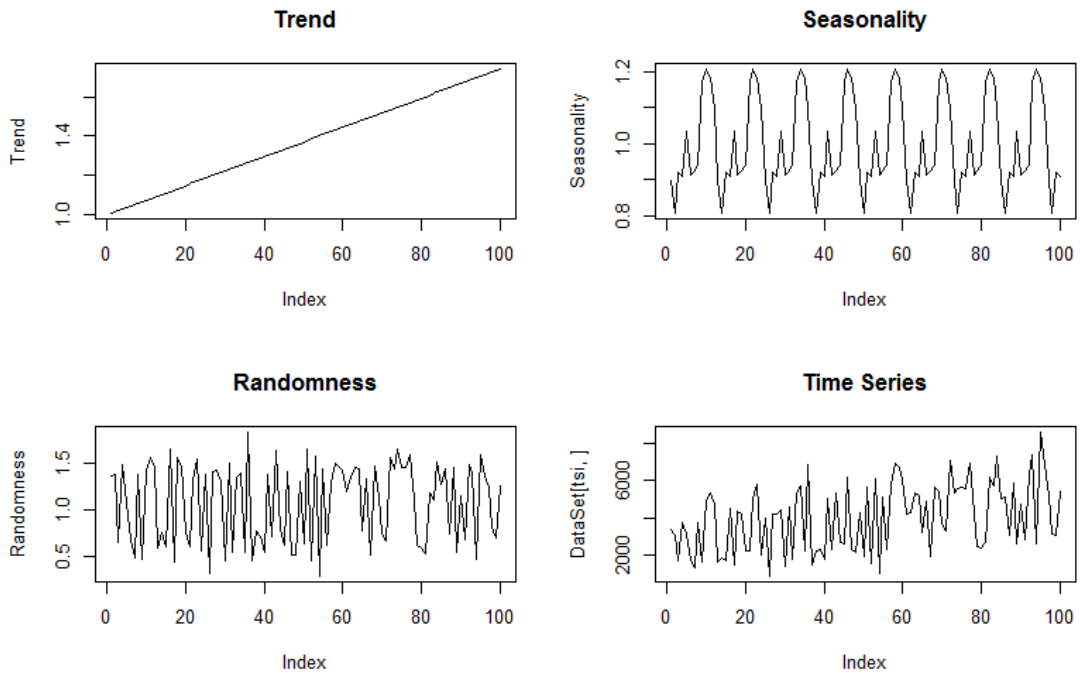
$$Y_i = L * T_i * S_i * R_i$$

Στη συνέχεια παρουσιάζονται στις *Εικόνες 5.10-5.14* κάποια παραδείγματα παραγωγής τεχνητών χρονοσειρών βάσει της προτεινόμενης μεθοδολογίας. Για λόγους πληρότητας, πέραν της τελικής χρονοσειράς απεικονίζονται σε κάθε περίπτωση και οι επιμέρους συνιστώσες της.

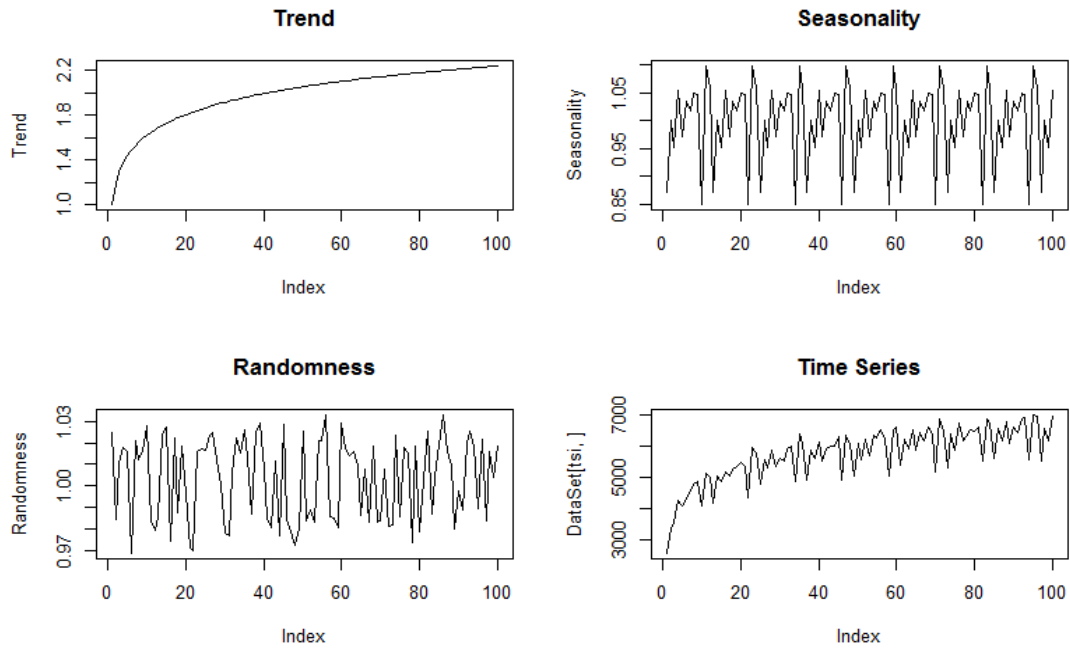
Κεφάλαιο 5: Παραγωγή τεχνητών χρονοσειρών μέσω προσομοίωσης



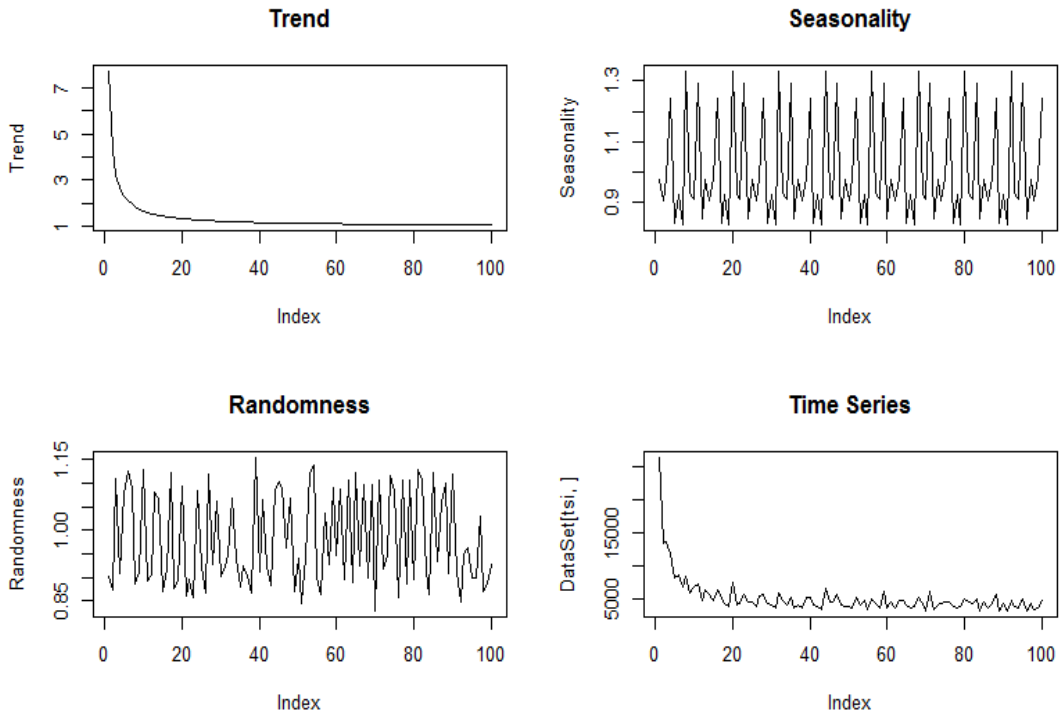
Εικόνα 5.10: Παραγωγή μίας τυχαίας χρονοσειράς EXP(H,M,L) μέσω της προτεινόμενης γεννήτριας.



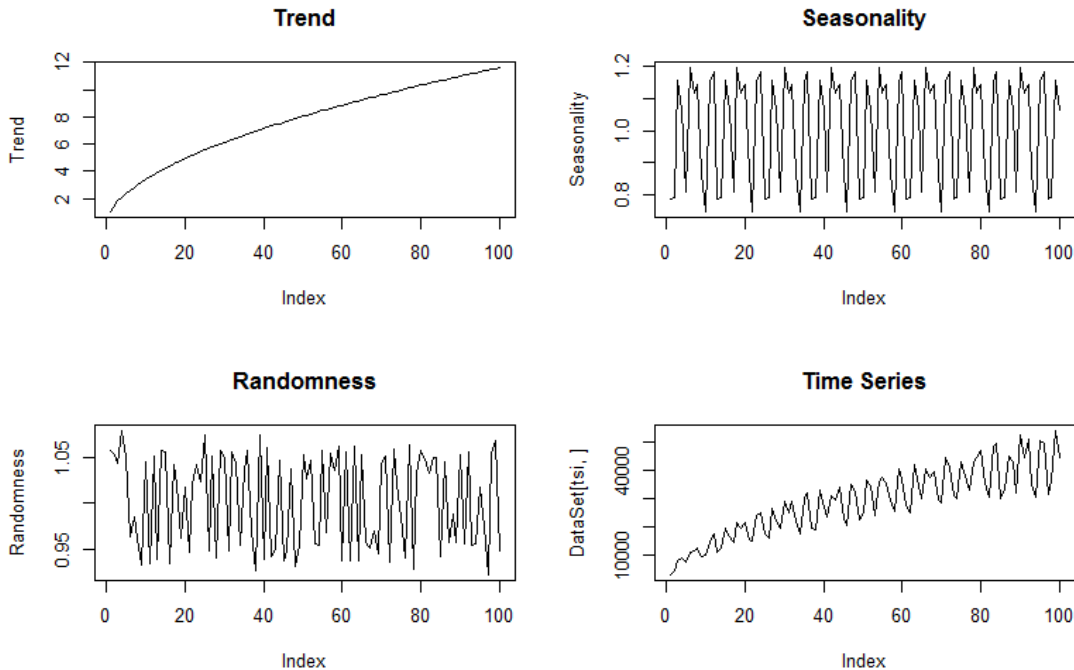
Εικόνα 5.11: Παραγωγή μίας τυχαίας χρονοσειράς LIN(L,M,H) μέσω της προτεινόμενης γεννήτριας.



Εικόνα 5.12: Παραγωγή μίας τυχαίας χρονοσειράς LOG(L,M,L) μέσω της προτεινόμενης γεννήτριας.



Εικόνα 5.13: Παραγωγή μίας τυχαίας χρονοσειράς INV(H,M,M) μέσω της προτεινόμενης γεννήτριας.



Εικόνα 5.14: Παραγωγή μίας τυχαίας χρονοσειράς POW(M,L,H) μέσω της προτεινόμενης γεννήτριας.

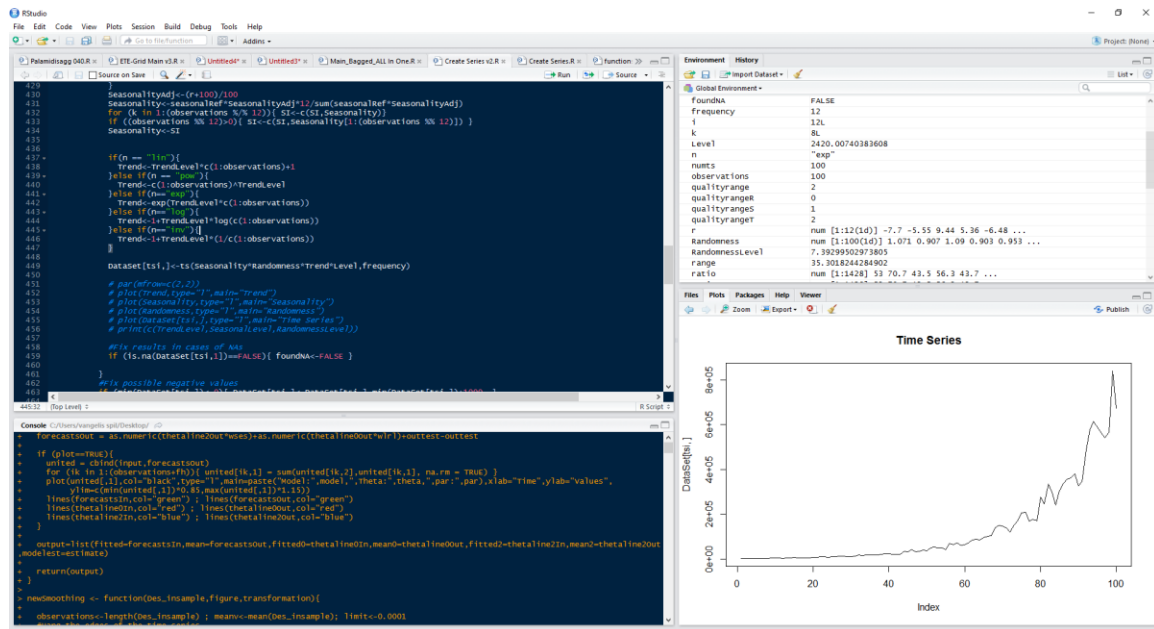
Βασικό πλεονέκτημα της γεννήτριας που αναπτύχθηκε είναι ότι μπορεί να αποκτήσει κανείς άμεσα και με πλήρως κατανοητό τρόπο χρονοσειράς ποικίλων χαρακτηριστικών, μήκους και πλήθους της επιλογής του. Επίσης, δεδομένου ότι οι κατηγορίες είναι σχετικά ευρείς και η παραγωγή των χρονοσειρών βασίζεται σε τυχαίους αριθμούς, οι παραγόμενες χρονοσειρές αναμένεται να διαφοροποιούνται επαρκώς ακόμα και εντός της κάθε κατηγορίας οδηγώντας έτσι σε αντιπροσωπευτικά αποτελέσματα.

Επισημαίνεται επίσης πως, σε περίπτωση που δεν θέλει κανείς να παράγει k χρονοσειρές από μία συγκεκριμένη υποκατηγορία, αλλά k εντελώς τυχαίες χρονοσειρές, μπορεί κάλλιστα να το πραγματοποιήσει γεννώντας $3 \cdot k$ σετ τυχαίων αριθμών κατανεμημένων εντός των ορίων των χαρακτηριστικών που παρουσιάστηκαν νωρίτερα και συνδυάζοντάς τα κατάλληλα (απευθείας εφαρμογή βημάτων 5 έως 7).

5.2.3 Υλοποίηση γεννήτριας και παραγωγή χρονοσειρών

Όπως παρουσιάστηκε στην προηγούμενη παράγραφο, η γεννήτρια που προτείνεται βασίζεται σε μία σύνθετη μεθοδολογία η οποία απαιτεί τη στατιστική ανάλυση ενός σετ δεδομένων που θα αποτελέσει τη βάση για τον καθορισμό των ορίων της κάθε κατηγορίας χρονοσειρών ανάλογα με την ένταση των χαρακτηριστικών τους, τη χρήση τεχνικών αποσύνθεσης, την παραγωγή τυχαίων αριθμών, τη γέννηση δεικτών που ακολουθούν συγκεκριμένες κατανομές και άλλες πολύπλοκες διεργασίες. Προκειμένου λοιπόν η διαδικασία που απαιτείται για τη γέννηση χρονοσειρών να υλοποιείται με έναν αυτόματο και συστηματικό τρόπο, απαιτείται αδιαμφισβήτητα η ανάπτυξή της σε προγραμματιστικό περιβάλλον.

Το περιβάλλον που χρησιμοποιήθηκε για αυτόν το σκοπό ήταν αυτό της R (R Development Core Team, 2008), μίας γλώσσας η οποία έχει γίνει ιδιαίτερα γνωστή τα τελευταία χρόνια στον τομέα της στατιστικής ανάλυσης κυρίως λόγω της open-source λογικής της η οποία επιτρέπει αφενός τη δωρεάν χρήση της και αφετέρου την ελεύθερη κυκλοφορία και επαναχρησιμοποίηση πακέτων και συναρτήσεων από τα μέλη της κοινότητάς της. Αυτό έχει ως αποτέλεσμα πολλές από τις διεργασίες που επιθυμεί να πραγματοποιήσει ο αναλυτής να είναι ήδη διαθέσιμες μειώνοντας έτσι σημαντικά τον χρόνο και την προσπάθεια για την ολοκλήρωση οποιασδήποτε ανάλυσης. Μάλιστα, πολλές από αυτές τις συναρτήσεις γίνονται διαθέσιμες από ακαδημαϊκούς αυξάνοντας έτσι την αξιοπιστία των αποτελεσμάτων. Το γραφικό περιβάλλον του προγράμματος, με την ονομασία *Rstudio*, παρουσιάζεται ενδεικτικά στην *Εικόνα 5.15*.



Εικόνα 5.15: Rstudio - Το γραφικό περιβάλλον της R

Για την υλοποίηση της γεννήτριας χρονοσειρών δεν αξιοποιήθηκε ωστόσο κάποιο από τα έτοιμα στατιστικά πακέτα της R πέραν του βασικού της “stats” καθώς κρίθηκε σκόπιμο για λόγους σαφήνειας και ακριβείας οι μέθοδοι αποσύνθεσης και ανάλυσης των χρονοσειρών να αναπτυχθούν από την αρχή. Στη συνέχεια παρατίθενται κάποια βασικά μέρη του κώδικα που αναπτύχθηκε για την υλοποίηση της γεννήτριας μαζί με συνοπτικά σχόλια για την κατανόησή του.

#Καθορισμός του επιπέδου L της χρονοσειράς. Αντιστοιχεί σε μία τιμή που επιλέγεται τυχαία μέσα από μία κανονική κατανομή που ορίζεται βάσει του συνόλου των επιπέδων του διαγωνισμού (dfElements\$Median)

`Level<-rnorm(1,dfElements$Median,(dfElements$Median*1/3))`

Κεφάλαιο 5: Παραγωγή τεχνητών χρονοσειρών μέσω προσομοίωσης

#Καθορισμός του επιπέδου εποχιακότητας (*SeasonalLevel*), τυχαιότητας (*RandomnessLevel*) και τάσης (*TrendLevel*) της χρονοσειράς βάσει του τύπου της τάσης που επιλέχθηκε (*type*) και της αντίστοιχης κατηγορίας έντασης (*qualityrangeS*, *qualityrangeR* και *qualityrangeT*) οι οποίες μπορούν να λάβουν τιμές *L*, *M*, και *H*.

```
SeasonalLevel<-ReturnRandomNumber(dfElements$Seasonality,qualityrangeS)

RandomnessLevel<-
ReturnRandomNumber(dfElements$Randomness,qualityrangeR)

if(type == "lin"){
  TrendLevel<-
ReturnRandomNumber(dfElements$CoefficientLinear,qualityrangeT)
} else if(type == "pow"){
  TrendLevel<-
ReturnRandomNumber(dfElements$CoefficientPower,qualityrangeT)
} else if(type == "exp"){
  TrendLevel<-
ReturnRandomNumber(dfElements$CoefficientExponential,qualityrangeT)
} else if(type == "log"){
  TrendLevel<-
ReturnRandomNumber(dfElements$CoefficientLogarithmic,qualityrangeT)
} else if(type == "inv"){
  TrendLevel<-
ReturnRandomNumber(dfElements$CoefficientInvese,qualityrangeT)
}
}
```

#Δημιουργία των δεικτών τυχαιότητας (*Randomness*)

```
r = array(0,observations)
for (i in 1:observations){
  if (sample(0:1, 1)==0){
    r[i] = + rnorm(1, RandomnessLevel, (RandomnessLevel*1/3))
  } else {
    r[i] = - rnorm(1, RandomnessLevel, (RandomnessLevel*1/3))
  }
}
Randomness = (r+100)/100
```

#Δημιουργία των δεικτών εποχιακότητας (*Seasonality*). Όπως φαίνεται οι παραγόμενοι δείκτες κανονικοποιούνται ώστε να έχουν άθροισμα 12.

```
#Create seasonality
Sl<-c()
r = array(0,12)#randomness
for (i in 1:12){
  if (sample(0:1, 1)==0){
    r[i] = + rnorm(1, SeasonalLevel, (SeasonalLevel*1/3))
  } else {
```

```

    r[i] = - rnorm(1, SeasonalLevel, (SeasonalLevel*1/3))
  }
}
SeasonalityAdj<-(r+100)/100
Seasonality<-seasonalRef*12/sum(seasonalRef)
for (k in 1:(observations %% 12)){ SI<-c(SI,Seasonality)}
if ((observations %% 12)>0){ SI<-c(SI,Seasonality[1:(observations %% 12)]) }
Seasonality<-SI

```

#Δημιουργία της συνιστώσας τάσης (Trend) ανάλογα με τον τύπο που επιλέχθηκε.

```

if(n == "lin"){
  Trend<-TrendLevel*c(1:observations)+1
}else if(n == "pow"){
  Trend<-c(1:observations)^TrendLevel
}else if(n=="exp"){
  Trend<-exp(TrendLevel*c(1:observations))
}else if(n=="log"){
  Trend<-1+TrendLevel*log(c(1:observations))
}else if(n=="inv"){
  Trend<-1+TrendLevel*(1/c(1:observations))
}

```

#Δημιουργία της τελικής χρονοσειράς (TimeSeries) ως γινόμενο των επιμέρους συνιστωσών της.

```

TimeSeries<-ts(Seasonality*Randomness*Trend*Level,frequency)

```

Σημειώνεται ότι το στοιχείο *dfElements* αποτελεί τη βάση αναφοράς και εμπεριέχει για τις 1428 μηνιαίες χρονοσειρές του *M3* τις τιμές των αντίστοιχων δεικτών. Επίσης η συνάρτηση *ReturnRandomNumber* είναι υπεύθυνη για τον καθορισμό των επιπέδων έντασης (*randnumberratio*) βάσει της κατηγορίας της χρονοσειράς (*qualityrange*), όπως φαίνεται παρακάτω.

```

ReturnRandomNumber<- function(ratio,qualityrange){

```

```

  #Ένταξη του ορίσματος ratio εντός μιας κανονικής κατανομής

```

```

  range<-quantile(range,0.95)

```

```

  if(qualityrange==2){
    rc<-((2*range/3)+range)/2
  }else if(qualityrange==1){
    rc<-((2*range/3)+(range/3))/2
  }else{
    rc<-range/6
  }

```

```

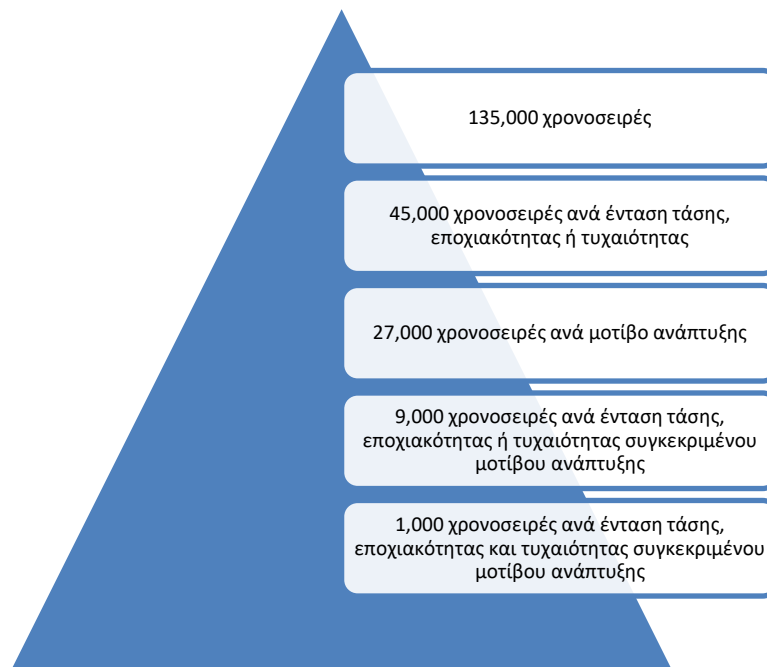
  randnumberratio=rnorm(1,rc,(rc*1/3))

```

```
} return(randnuberratio)
```

Η γέννηση των τεχνητών χρονοσειρών πραγματοποιείται όπως αναφέρθηκε νωρίτερα θεωρώντας ότι αυτές αποτελούνται από 100 παρατηρήσεις ούτως ώστε να υπάρχει επαρκές δείγμα παρατηρήσεων για την επεξεργασία τους και την παραμετροποίηση των μοντέλων πρόβλεψης χωρίς το σύστημα να επιβαρύνεται με σημαντικό υπολογιστικό κόστος. Για λόγους πληρότητας και ομοιομορφίας επιλέχθηκε επίσης να παράγεται ισάριθμο πλήθος χρονοσειρών για κάθε πιθανό συνδυασμό χαρακτηριστικών τους. Έτσι, δεδομένου ότι υπάρχουν συνολικά πέντε μοτίβα ανάπτυξης και τρία επίπεδα έντασης τάσης, εποχιακότητας και τυχαιότητας, για παραγωγή n χρονοσειρών ανά υποκατηγορία χρονοσειράς προκύπτουν συνολικά $5*3*3*3*n=135*n$ χρονοσειρές.

Αν και στην πράξη βρέθηκε ότι οι διαφοροποιήσεις που προκύπτουν στα συμπεράσματα όσον αφορά την επίδοση των μεθόδων πρόβλεψης σε κάθε πιθανή υποκατηγορία χρονοσειρών είναι αμελητέες για παραγωγή περισσότερων από $n=100$ χρονοσειρές, για λόγους εγκυρότητας η παράμετρος τέθηκε ίση με 1,000. Έτσι η γεννήτρια παρήγαγε συνολικά 135,000 χρονοσειρές οι οποίες κατανέμονται όπως φαίνεται στην πυραμίδα της *Εικόνας 5.16*.



Εικόνα 5.16: Πλήθος παραγόμενων χρονοσειρών από τη γεννήτρια ανά κατηγορία χρονοσειράς

5.3 Υποστήριξη της πληρότητας των παραγόμενων χρονοσειρών

Όπως παρουσιάστηκε στην προηγούμενη παράγραφο, βασικός σκοπός της γεννήτριας είναι να παράγει ένα σύνολο από χρονοσειρές οι οποίες θα αποτελούν αντιπροσωπευτικό δείγμα μελέτης, με την έννοια ότι θα περιλαμβάνουν ικανοποιητικό πλήθος από οποιοδήποτε τύπο χρονοσειράς, ή αλλιώς συνδυασμό τύπου τάσης, έντασης τάσης, εποχιακότητας και τυχαιότητας.

Με τον τρόπο με τον οποίο ορίστηκαν, μετρήθηκαν και αναπαραστάθηκαν τα παραπάνω χαρακτηριστικά, αναμένει κανείς πράγματι τα αποτελέσματα της γεννήτριας που αναπτύχθηκε να είναι ιδιαίτερος αντιπροσωπευτικά. Ωστόσο, για την εξακρίβωση της παραπάνω διαίσθησης απαιτείται προφανώς επαλήθευση με χρήση κατάλληλων μεθόδων. Σε αυτήν την κατεύθυνση, ένας εύκολος και πρακτικός τρόπος επαλήθευσης είναι η απεικόνιση της έντασης της τιμής των επιμέρους χαρακτηριστικών στον χώρο. Δεδομένου ωστόσο ότι οι μεταβλητές απεικόνισης είναι αρκετές (ένταση εποχιακότητας, τυχαιότητας και τάσης ανά είδος) το παραπάνω εγχείρημα είναι αρκετά πολύπλοκο.

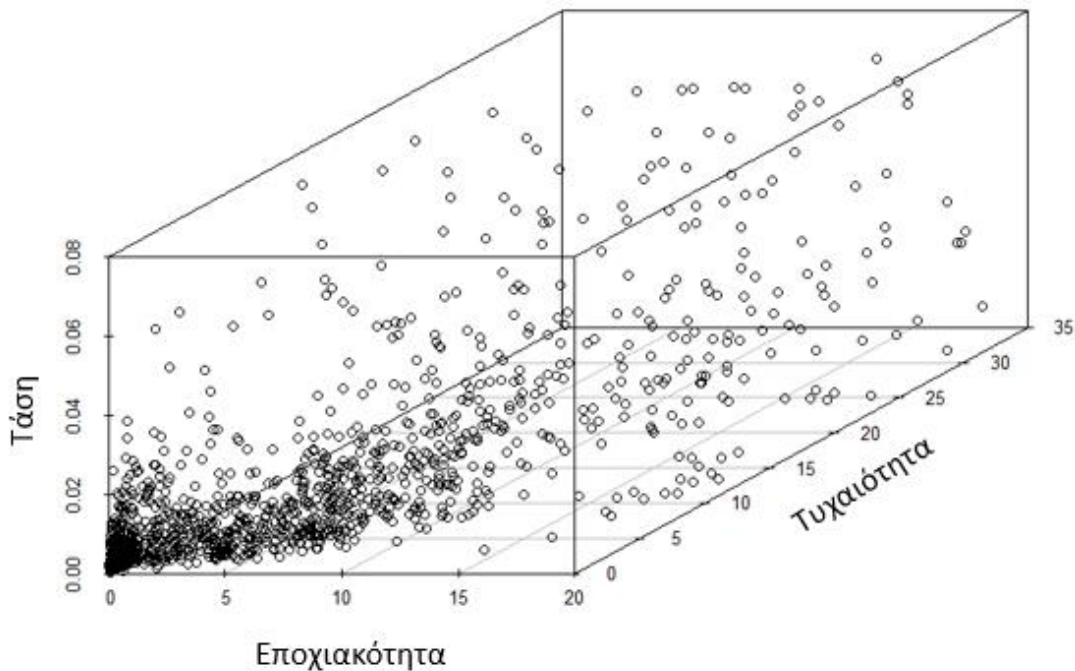
Προκειμένου λοιπόν να μειωθεί το πλήθος των εξεταζόμενων μεταβλητών, ανατρέχουμε σε πρώτη φάση στον πίνακα συσχέτισής τους, ο οποίος και φαίνεται παρακάτω. Αν δύο μεταβλητές βρεθεί πως είναι ιδιαίτερα συσχετισμένες, τότε η μία εξ αυτών θα μπορεί να εξαιρεθεί της μελέτης χωρίς να επηρεάζει ουσιαστικά τα αποτελέσματα. Όπως παρατηρείται, η εποχιακότητα φαίνεται να σχετίζεται ασθενώς με την τυχαιότητα, όχι ωστόσο σε σημαντικό βαθμό, ενώ και οι δύο παραπάνω μεταβλητές είναι πρακτικά ασυσχέτιστες με όλες τις μεταβλητές τάσης. Επίσης, η τάση γραμμικής μορφής συσχετίζεται σημαντικά με αυτή της λογαριθμικής και της δύναμης, ενώ η εκθετική με αυτή της δύναμης.

Συμπερασματικά, μπορεί κανείς λοιπόν να απεικονίσει στον τρισδιάστατο χώρο τη φύση των χρονοσειρών χρησιμοποιώντας ως μεταβλητές την τάση, την εποχιακότητα και την τυχαιότητα, όπου ως τάση λογίζεται είτε το εκθετικό είτε το γραμμικό μοτίβο τάσης.

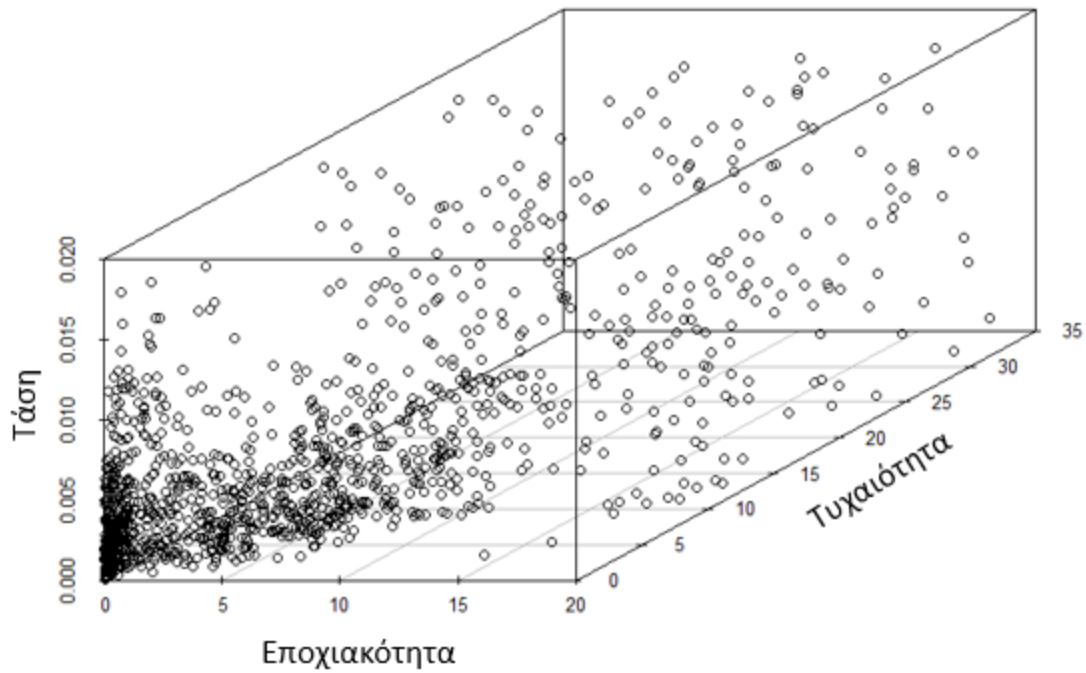
Πίνακας 5.7: Πίνακας συσχετίσεων έντασης των ποιοτικών χαρακτηριστικών των μηνιαίων δεδομένων του M3.

	Επο/τητα	Τυ/τητα	Τάση (linear)	Τάση (exp)	Τάση (log)	Τάση (inverse)	Τάση (power)
Επο/τητα	1.00	0.55	0.20	0.13	0.17	0.16	0.14
Τυ/τητα	0.55	1.00	0.41	0.36	0.33	0.29	0.28
Τάση (linear)	0.20	0.41	1.00	0.58	0.92	0.84	0.56
Τάση (exp)	0.13	0.36	0.58	1.00	0.39	0.28	0.87
Τάση (log)	0.17	0.33	0.92	0.39	1.00	0.97	0.49
Τάση (inverse)	0.16	0.29	0.84	0.28	0.97	1.00	0.39
Τάση (power)	0.14	0.28	0.56	0.87	0.49	0.39	1.00

Αρχικά απεικονίζεται λοιπόν ο χώρος στον οποίο εκτείνονται τα μηνιαία δεδομένα του διαγωνισμού *M3*, οι οποίες και αποτελούν τη βάση για την παραγωγή των τεχνητών χρονοσειρών και αντιπροσωπευτικότερο δείγμα για όλον τον διαγωνισμό. Επισημαίνεται πως προκειμένου τα αποτελέσματα να είναι πιο ευπαρουσίαστα, οι άξονες του κάθε χαρακτηριστικού έχουν λάβει ως άνω όριο αυτό του *Πίνακα 5.6*. Τα αποτελέσματα παρουσιάζονται για λόγους πληρότητας για το γραμμικό και για το εκθετικό μοτίβο τάσης ξεχωριστά στην *Εικόνα 5.17* και *5.18* αντίστοιχα, αν και τα συμπεράσματα είναι ποιοτικά τα ίδια. Όπως φαίνεται, ο χώρος στον οποίο εκτείνονται οι χρονοσειρές είναι στο μεγαλύτερό του μέρος κενός, καθώς η πλειοψηφία των χρονοσειρών έχει χαρακτηριστικά χαμηλής έντασης και μόνο ένα μικρό σύνολο χρονοσειρών διαφοροποιείται σημαντικά. Αυτό επιβεβαιώνει ακράδαντα τους ισχυρισμούς που θέλουν τα δεδομένα των διαγωνισμών πρόβλεψης μη πλήρη και ακατάλληλα πιθανώς για εξαγωγή καθολικών συμπερασμάτων για κάθε δυνατό τύπο χρονοσειράς.



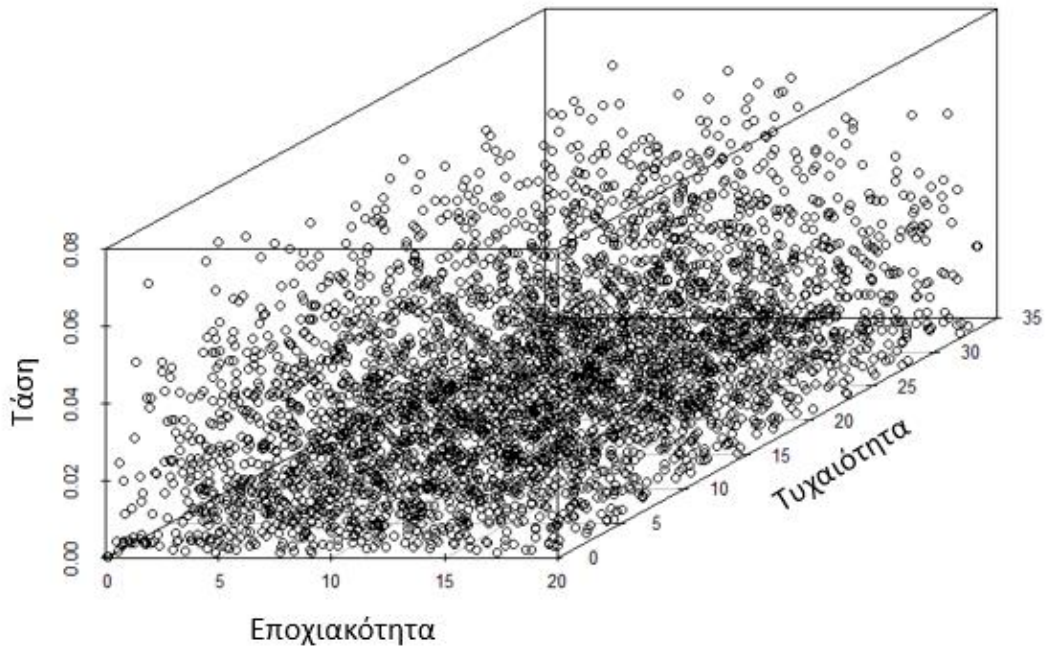
Εικόνα 5.17: Απεικόνιση ποιοτικών χαρακτηριστικών των μηνιαίων χρονοσειρών του *M3* στο χώρο - Γραμμική τάση.



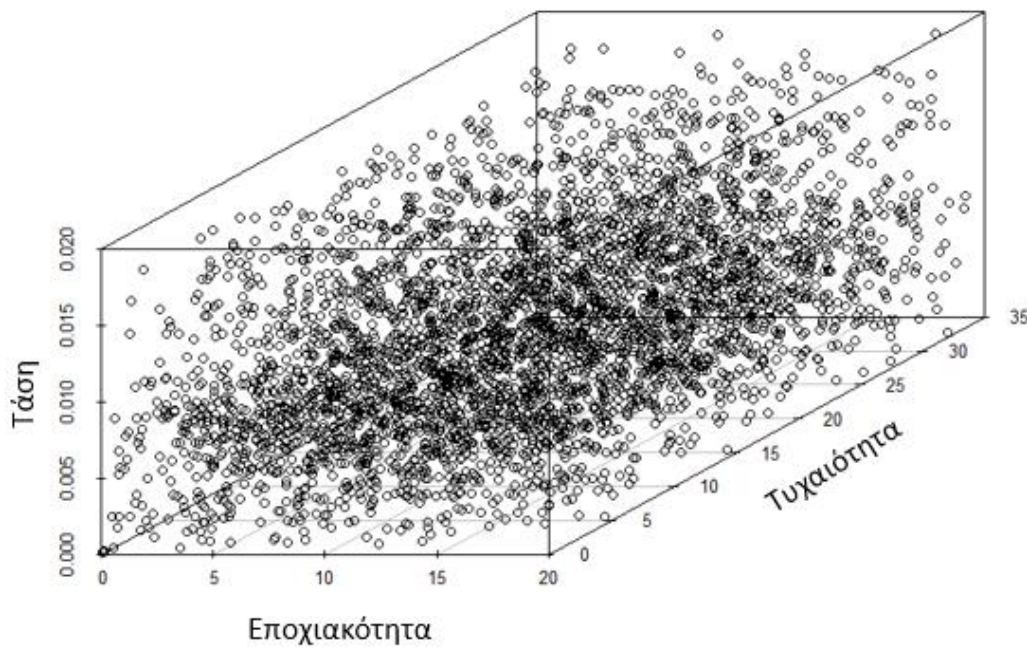
Εικόνα 5.18: Απεικόνιση ποιοτικών χαρακτηριστικών των μηνιαίων χρονοσειρών του M3 στο χώρο –Εκθετική τάση.

Εφαρμόζοντας στη συνέχεια την ίδια διαδικασία για το σύνολο των τεχνητών χρονοσειρών, όπως παρατηρείται, ο ίδιος ακριβώς χώρος συμπληρώνεται σαφώς πιο ολοκληρωμένα και ομοιόμορφα, με τις χρονοσειρές να διαθέτουν πλέον χαρακτηριστικά έντασης μεγάλου εύρους. Τα αποτελέσματα παρουσιάζονται για το γραμμικό και για το εκθετικό μοτίβο τάσης ξεχωριστά στην [Εικόνα 5.19](#) και [5.20](#) αντίστοιχα, χωρίς ωστόσο ποιοτικές διαφοροποιήσεις.

Κεφάλαιο 5: Παραγωγή τεχνητών χρονοσειρών μέσω προσομοίωσης



Εικόνα 5.19: Απεικόνιση ποιοτικών χαρακτηριστικών χρονοσειρών γεννήτριας στον χώρο - Γραμμική τάση.



Εικόνα 5.20: Απεικόνιση ποιοτικών χαρακτηριστικών χρονοσειρών γεννήτριας στον χώρο - Εκθετική τάση.

Ένας επιπρόσθετος έλεγχος που σχετίζεται με την αποτελεσματικότητα της γεννήτριας παραγωγής τεχνητών δεδομένων, είναι η απεικόνιση των χρονοσειρών στο χώρο που ορίστηκε από τους *Kang et al. (2017)*, οι οποίοι και όπως αναφέρθηκε νωρίτερα υποθέτουν πως χαρακτηριστικά των χρονοσειρών είναι πέραν της εποχιακότητας, της τάσης και της τυχαιότητας, η αυτοσυσχέτιση των παρατηρήσεων και η στασιμότητα, ή αλλιώς η σταθερότητα της διακύμανσης. Επίσης, καθώς μελέτησαν όλο τα δεδομένα του διαγωνισμού *M3*, έθεσαν ως επιπλέον μεταβλητή ανάλυσης αυτήν της συχνότητας. Φυσικά, δεδομένου ότι η παρούσα γεννήτρια δεν μοντελοποιεί ξεχωριστά τις επιπλέον διαστάσεις, περιμένει κανείς ο χώρος που ορίζουν αυτές οι μεταβλητές να μην γεμίζει με την ίδια αποδοτικότητα που παρουσιάστηκε νωρίτερα. Παραμένει ωστόσο αναμφίβολα ένας ενδιαφέρον έλεγχος της αποτελεσματικότητάς της.

Πιο συγκεκριμένα, οι συγγραφείς πρότειναν τη μέτρηση των μεγεθών που αναφέρθηκαν ως εξής:

1. Η χρονοσειρά Y_t αποσυντίθεται στις συνιστώσες τάσης (T), εποχιακότητας (S) και τυχαιότητας (R) με χρήση της προσθετικής αποσύνθεσης *stl* (ή μέσω *loess* σε περίπτωση χρονοσειράς ετήσιας συχνότητας).
2. Η συχνότητάς της κάθε χρονοσειράς είναι 12 σε περίπτωση μηνιαίων δεδομένων, 4 σε περίπτωση τριμηνιαίων και 1 σε περίπτωση ετήσιων ή άλλων δεδομένων.
3. Η ένταση τυχαιότητας ισοδυναμεί με την εντροπία της χρονοσειράς (*spectral entropy*) η οποία και υπολογίζεται ως $-\int_{-\pi}^{\pi} f_y(\omega) \log f_y(\omega) d\omega$, όπου $f_y(\omega)$ είναι η πυκνότητα εντροπίας (*spectral density*).
4. Η ένταση εποχιακότητας υπολογίζεται ως $1 - \frac{\text{var}(R_t)}{\text{var}(Y_t - T_t)}$ δηλαδή ως λόγος της διακύμανσης του τυχαίου παράγοντα προς τη διακύμανση του εποχιακού παράγοντα.
5. Η ένταση τάσης υπολογίζεται ως $1 - \frac{\text{var}(R_t)}{\text{var}(Y_t - S_t)}$ δηλαδή ως λόγος της διακύμανσης του τυχαίου παράγοντα προς τη διακύμανση του παράγοντα τάσης.
6. Η ένταση αυτοσυσχέτισης A ταυτίζεται με τον συντελεστή αυτοσυσχέτισης των παρατηρήσεων πρώτου βαθμού *AR(1)*.
7. Υπολογίζεται ο βέλτιστος συντελεστής λ μετασχηματισμού *Box-Cox* στο διάστημα $[0, 1]$ μέσω μεγιστοποίησης της πιθανοφάνειας και αυτός χρησιμοποιείται ως μέτρο στασιμότητας B .
8. Οι έξι μεταβλητές ομαδοποιούνται με χρήση της τεχνικής *Principal Component Analysis* (*PCA*) μέσω τις οποίας και εξαγονται ισάριθμες τεχνητές μεταβλητές που περιγράφουν τις επιμέρους χρονοσειρές χρησιμοποιώντας μία γραμμική σχέση μεταξύ των επιμέρους μεταβλητών βάσει της συσχέτισής τους.

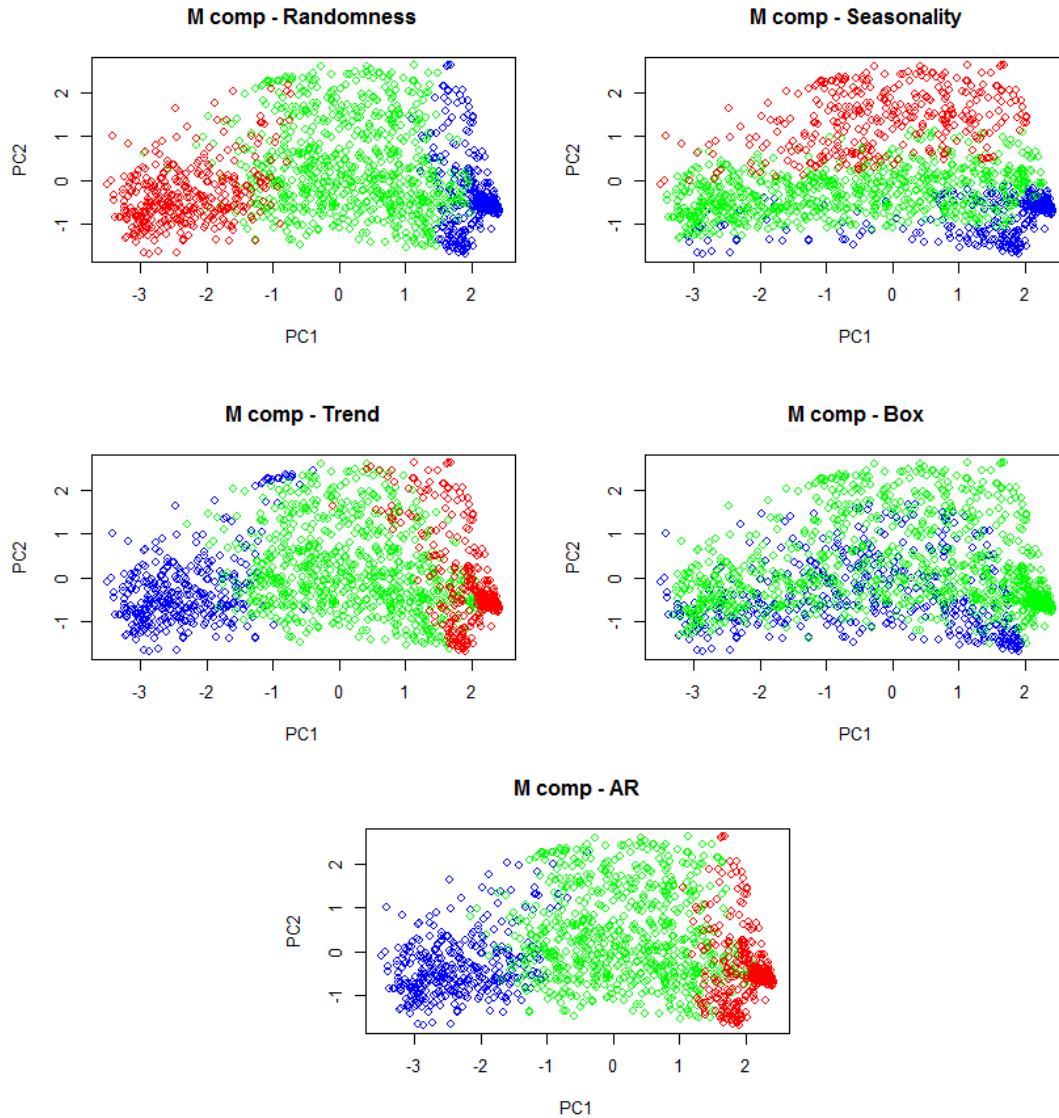
Θέλοντας να εφαρμόσουμε την παραπάνω ανάλυση μόνο στα μηνιαία δεδομένα του διαγωνισμού, εξαιρείται αρχικά ως παράμετρος η συχνότητα, η οποία και πλέον είναι ίδια για όλες τις χρονοσειρές του δείγματος, και εφαρμόζονται στη συνέχεια κανονικά τα υπόλοιπα βήματα της διαδικασίας. Οι πέντε τεχνητές μεταβλητές που παράγονται μέσω της *PCA* προκύπτουν ως γραμμικό άθροισμα των αρχικών χρησιμοποιώντας τους συντελεστές του [Πίνακα 5.8](#).

Κεφάλαιο 5: Παραγωγή τεχνητών χρονοσειρών μέσω προσομίωσης

Πίνακας 5.8: Συντελεστές δόμησης τεχνητών μεταβλητών PCA για την απεικόνιση των χρονοσειρών του M3 στο χώρο.

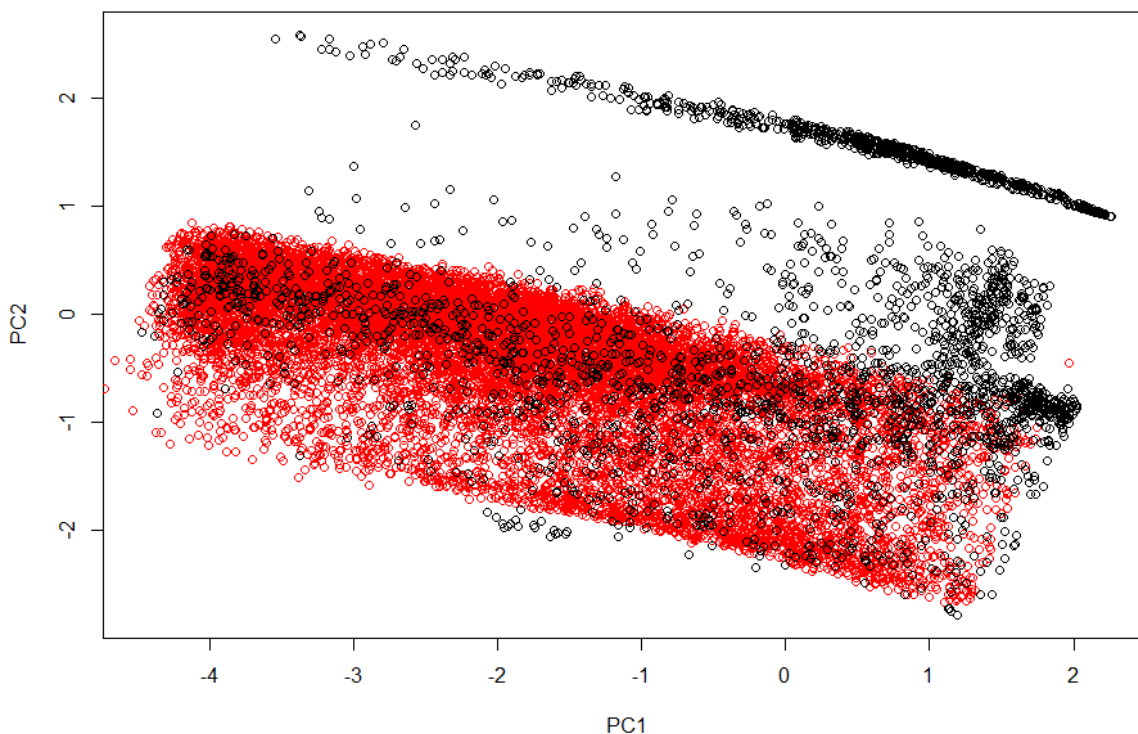
Μεταβλητή	PC1	PC2	PC3	PC4	PC5
R	-0.56	0.06	-0.05	-0.61	0.55
T	0.55	0.10	0.20	-0.76	-0.28
S	-0.15	0.91	0.35	0.13	-0.08
A	0.56	0.08	0.17	0.17	0.79
B	0.20	0.39	-0.90	-0.05	0.02

Μάλιστα, οι πρώτες δύο τεχνητές μεταβλητές επεξηγούν το 77.64% της συνολικής διακύμανσης στο δείγμα, και συνεπώς μπορούν να χρησιμοποιηθούν με ασφάλεια για την απεικόνιση των χρονοσειρών σε δισδιάστατο χώρο. Το αποτέλεσμα παρουσιάζεται στην *Εικόνα 5.21* όπου επιπρόσθετα για κάθε μία από τις αρχικές μεταβλητές απεικονίζεται η έντασή της στον χώρο (με κόκκινο οι χρονοσειρές Υψηλής ένταση χαρακτηριστικού, με πράσινο οι Μεσαίας και με μπλε οι Χαμηλής). Βασικό συμπέρασμα είναι ότι η πλειοψηφία των μηνιαίων χρονοσειρών του διαγωνισμού χαρακτηρίζονται από χαμηλή ένταση εποχιακότητας και τυχαιότητας και υψηλή ένταση τάσης. Επίσης οι περισσότερες χρονοσειρές δεν είναι στάσιμες, ενώ οι μεταβλητές τάσης-θορύβου και τάσης-αυτοσυσχέτισης φαίνονται να συσχετίζονται σημαντικά.



Εικόνα 5.21: Απεικόνιση της έντασης των ποιοτικών χαρακτηριστικών των μηνιαίων χρονοσειρών του M3 στον χώρο με χρήση PCA.

Χρησιμοποιώντας κανείς τις ίδιες μεταβλητές και αναλύοντας κατάλληλα τα δεδομένα των τεχνητών χρονοσειρών που παρήχθησαν νωρίτερα, δύναται να τις προβάλει στον ίδιο ακριβώς χώρο και να συγκρίνει το βαθμό στον οποίο αυτές τον συμπληρώνουν. Το αποτέλεσμα που προκύπτει φαίνεται στην *Εικόνα 5.22* και όπως παρατηρείται η γεννήτρια έχει καταφέρει να «γεμίσει» το χώρο σε αρκετά ικανοποιητικό βαθμό παράγοντας ομοιόμορφα χρονοσειρές ποικίλων χαρακτηριστικών. Σημειώνεται πως δεδομένης της μη ενσωμάτωσης των παραγόντων B και A στη διαδικασία παραγωγής χρονοσειρών από τη γεννήτρια, ο εν λόγω χώρος εξ αρχής δεν αναμενόταν να συμπληρωθεί τέλεια από τις χρονοσειρές που αυτή παράγει και η εμφάνιση χρονοσειρών σε όλο σχεδόν το εύρος του χώρου που ορίστηκε είναι συνεπώς ιδιαίτερα ενθαρρυντική.



Εικόνα 5.22: Απεικόνιση των χρονοσειρών του M3 (μαύρο) και των μηνιαίων τεχνητών της γεννήτριας (κόκκινο) στο χώρο που ορίστηκε μέσω της PCA και των μεταβλητών των Kang et al., 2017.

Πιο συγκεκριμένα, όπως παρατηρείται στο διάγραμμα, η γεννήτρια έχει καταφέρει να παράξει αρκετές θορυβώδεις χρονοσειρές που στο διαγωνισμό απουσίαζαν, ενώ το πλήθος των χρονοσειρών φαίνεται να διαμοιράζεται σχεδόν με ομοιόμορφο τρόπο στο χώρο ως προς τα υπόλοιπα χαρακτηριστικά (τάσης και εποχιακότητας). Αξίζει να σημειωθεί πως, τόσο βάσει δεδομένων διαγωνισμού όσο και γεννήτριας, συνεχίζουν να υπάρχουν κάποιες αραιοκατοικημένες περιοχές στο χώρο, όπως παραδείγματος χάριν θορυβώδεις χρονοσειρές με έντονη εποχιακότητα (κάτω αριστερά). Αυτό είναι απολύτως λογικό δεδομένου ότι, αν μια χρονοσειρά διακρίνεται από τυχαιότητα, είναι αδύνατο να εμφανίζει συστηματικά εποχιακά μοτίβα. Έτσι, συμπεραίνουμε συνολικά πως η γεννήτρια έχει καταφέρει να συμπληρώσει το χώρο όπως ακριβώς επιθυμούσαμε, και όπου βέβαια αυτό είναι φυσικώς εφικτό.

Βάσει των παραπάνω μπορούν να εξαχθούν συνοπτικά τα παρακάτω βασικά συμπεράσματα:

- Οι χρονοσειρές των διαγωνισμών προβλέψεων κατανέμονται ανομοιόμορφα ως προς τα χαρακτηριστικά τους.
- Τα βασικότερα χαρακτηριστικά των χρονοσειρών είναι η συχνότητα, η τάση, η εποχιακότητα και η τυχαιότητα, ενώ η στασιμότητα και η αυτοσυσχέτιση μπορεί και αυτή να προσδώσει πιθανώς κάποια επιπλέον πληροφορία.

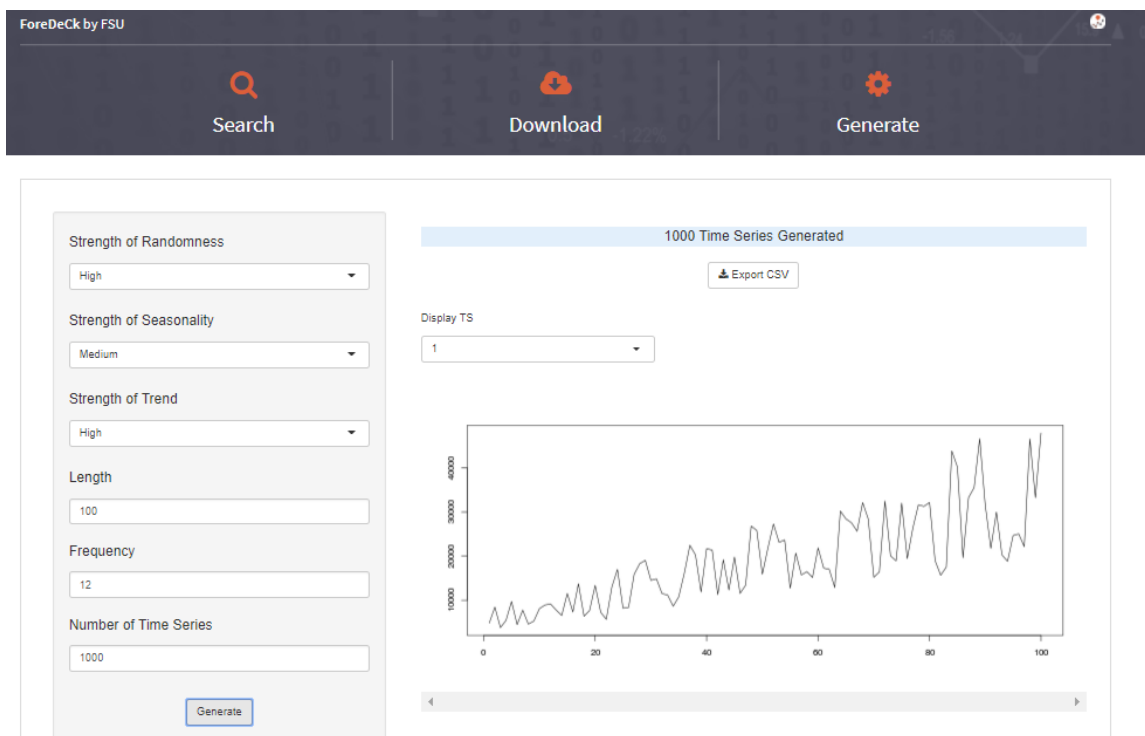
- Το είδος της τάσης (γραμμικό ή μη γραμμικό) φαίνεται να αποτελεί μία επιπλέον παράμετρο που χαρακτηρίζει τη χρονοσειρά.
- Χρησιμοποιώντας κανείς κατάλληλους μηχανισμούς μπορεί να γεννήσει τεχνητές χρονοσειρές χαρακτηριστικών που επιθυμεί και συνεπώς να δημιουργήσει αξιόλογα σετ δεδομένων για τον αποτελεσματικότερο έλεγχο της απόδοσης των μοντέλων πρόβλεψης.
- Η αποτελεσματικότητα μιας γεννήτριας να παράγει τεχνητά δεδομένα που συμπληρώνουν ικανοποιητικά τον χώρο εξαρτάται σημαντικά από το ποιες μεταβλητές τον δομούν και το ποιες λαμβάνονται υπόψη κατά την παραγωγή τους.
- Η γεννήτρια που κατασκευάστηκε αποδεικνύεται αποτελεσματική τόσο για τον καθορισμό ενός αντιπροσωπευτικού χώρου μελέτης, όσο και για τη συμπλήρωση άλλων που έχουν οριστεί μέσω διαφορετικών δειγμάτων και μεταβλητών.

Σε αυτό το σημείο αναφέρεται πως η παρούσα γεννήτρια είναι διαθέσιμη για επιστημονικούς σκοπούς ελεύθερα και στο διαδίκτυο στη διεύθυνση <http://fsudataset.com/>.

Για την αξιοποίησή της δεν απαιτείται εγγραφή σε κάποια υπηρεσία και ο χρήστης πρέπει απλά να προσδιορίσει όπως παρουσιάστηκε νωρίτερα το επίπεδο της τυχαιότητας, της εποχιακότητας και της τάσης (High, Medium ή Low), καθώς και τη συχνότητα των τεχνητών δεδομένων. Στη συνέχεια καθορίζεται το πλήθος των χρονοσειρών που επιθυμείται να παραχθούν και το μήκος τους.

Πατώντας το πλήκτρο 'Generate', όπως φαίνεται στην *Εικόνα 5.23*, οι ζητούμενες χρονοσειρές παράγονται σχεδόν ακαριαία και παρουσιάζονται ταυτόχρονα ανά αύξοντα αριθμό σε αντίστοιχο διάγραμμα στα δεξιά της εφαρμογής προκειμένου ο χρήστης να επιβεβαιώσει αν τα δεδομένα που παρήχθησαν είναι και τα επιθυμητά. Τέλος, πατώντας την ένδειξη 'Export CSV', ο χρήστης μπορεί να κατεβάσει τα δεδομένα σε μορφή .csv για περαιτέρω χρήση και να τα αξιοποιήσει όπως εκείνος επιθυμεί.

Κεφάλαιο 5: Παραγωγή τεχνητών χρονοσειρών μέσω προσομοίωσης



Εικόνα 5.23: Η διαδικτυακή εφαρμογή που αναπτύχθηκε για την παροχή των τεχνητών δεδομένων στην ευρύτερη επιστημονική κοινότητα.

Υπάρχουν τομείς όπου η εξειδίκευση είναι αδύνατη. Η αγορά μετοχών και η μακροπρόθεσμη πρόβλεψη για τον στρατηγικό σχεδιασμό είναι κάποια καλά παραδείγματα. Έχει αποδειχτεί ότι οι ειδικοί δεν είναι πολύ καλύτεροι από μία μαιμού που πετά ένα κέρμα.

Daniel Kahneman

Κεφάλαιο 6: Πειραματική διάταξη - Ανάλυση και εξέταση των αποτελεσμάτων

Παρά την όποια πρόοδο έχει σημειωθεί στον κλάδο των προβλέψεων, η έρευνα που αφορά την προέκταση των χρονοσειρών είναι αρκετά κατακερματισμένη. Αν και υπάρχουν δηλαδή εργαλεία που μπορούν να οδηγήσουν σε βελτίωση της ικανότητάς μας να προβλέψουμε με στατιστικό τρόπο, αυτά δεν αξιοποιούνται κατάλληλα ή δεν συνδυάζονται αποδοτικά μεταξύ τους για την επίτευξη του βέλτιστου δυνατού αποτελέσματος. Η βελτίωση της προβλεπτικής ακρίβειας των μεθόδων γίνεται συνήθως μέσω της χρήσης επιπλέον πληροφορίας ή την κριτική επεξεργασία των στατιστικών προβλέψεων από ειδικούς. Αυτό έχει ως αποτέλεσμα η αναμενόμενη ακρίβεια των μεθόδων να περιορίζεται και η όποια βελτίωσή της να αφορά, εξαιτίας της μη αυτοματοποίησης της όλης διαδικασίας, αποκλειστικά μεγέθη που ενδιαφέρουν πρωτίστως τις επιχειρήσεις. Τι γίνεται όμως στην περίπτωση που είναι επιθυμητή η ακριβής πρόβλεψη μεγάλου όγκου δεδομένων ποικίλων χαρακτηριστικών και ιδιαιτεροτήτων;

Όπως αναφέρθηκε στην παρούσα διατριβή η παραγωγή προβλέψεων μέσω μιας διαδικασίας υψηλής απόδοσης απαιτεί τη χρήση αποκλειστικά στατιστικά μεθόδων και μάλιστα των καλύτερων εξ αυτών. Προκειμένου μία μέθοδος ωστόσο να θεωρηθεί επιτυχημένη θα πρέπει αναμφισβήτητα να καταφέρνει να ξεπεράσει προβλεπτικά όλες τις κλασικές προσεγγίσεις. Η βιβλιογραφία αποδεικνύει πως κάτι τέτοιο σπάνια επιτυγχάνεται καθώς απλά μοντέλα, όπως είναι αυτά της θ (2000) και της *damped* (1968), συνεχίζουν να αποτελούν μέχρι και σήμερα κάποιες από τις αποδοτικότερες επιλογές για την προέκταση των χρονοσειρών. Σκοπός εδώ είναι λοιπόν να δημιουργηθεί ένα ολοκληρωμένο μεθοδολογικό πλαίσιο το οποίο θα αξιοποιεί και θα συνδυάζει τις πιο επιτυχημένες πρακτικές για να επιτευχθούν οι παραπάνω στόχοι.

Νωρίτερα προτάθηκε στο [Κεφάλαιο 3](#) μία νέα μέθοδος για την προέκταση χρονοσειρών η οποία αξιοποιεί τις ιδιότητες της κλασικής μεθόδου θ αλλά επεκτείνει την εφαρμογή της σε χρονοσειρές με μη γραμμικά μοτίβα τάσης. Επίσης παρουσιάστηκε η πολλαπλασιαστική έκφραση της μεθόδου που επιτρέπει την αλληλεπίδραση των συνιστωσών επιπέδου και τάσης, όπως αυτά υπολογίζονται από τις επιμέρους γραμμές θ . Αποτέλεσμα των παραπάνω επεκτάσεων ήταν η αύξηση της ευελιξίας και την προσαρμοστικότητας της μεθόδου και η βελτίωση της ικανότητάς της να περιγράφει αποδοτικά περισσότερους τύπους χρονοσειρών. Έτσι, επιλέγοντας κανείς το βέλτιστο μοντέλο θ από αυτά που προκύπτουν θα μπορούσε να επιτύχει σημαντικά καλύτερη ακρίβεια έναντι άλλων κλασικών μεθόδων πρόβλεψης.

Στο [Κεφάλαιο 4](#) εισάχθηκε επίσης μία νέα τεχνική μη γραμμικής εξομάλυνσης χρονοσειρών. Στόχος η αποτελεσματική απομάκρυνση του θορύβου από το σήμα της χρονοσειράς ούτως ώστε να βελτιωθεί η απόδοση των μοντέλων που καλούνται να την προεκτείνουν. Στο [Κεφάλαιο 2](#) παρουσιάστηκαν τέλος αρκετές τεχνικές αποσύνθεσης για την αποεποχικοποίηση των χρονοσειρών και αναφέρθηκαν τα οφέλη των μεθόδων

συρρίκνωσης δεικτών εποχιακότητας. Έτσι, στο παρόν κεφάλαιο θα μελετηθεί κατά πόσο ο συνδυασμός των παραπάνω τεχνικών προβλέψεων μπορεί να συμβάλει στην κατεύθυνση της επίτευξης υψηλής ακρίβειας με συστηματικό τρόπο.

Επιπλέον, προκειμένου τα συμπεράσματα της εργασίας να είναι όσο το δυνατό πιο αντιπροσωπευτικά, αξιοποιούνται δύο διαφορετικά σετ δεδομένων: Τα δεδομένα των διαγωνισμών *M* και τεχνητές χρονοσειρές που παρήχθησαν από τη γεννήτρια που παρουσιάστηκε στο *Κεφάλαιο 5*. Η επίδοση της προτεινόμενης μεθοδολογίας συγκρίνεται με αυτή των κλασικών μοντέλων τα οποία και οφείλει να ξεπεράσει σε όλους τους τομείς.

6.1 Παρουσίαση της προτεινόμενης μεθοδολογίας

Η μεθοδολογία που προτείνεται περιλαμβάνει πέντε βασικά βήματα τα οποία και οπτικοποιούνται στο διάγραμμα ροής της *Εικόνας 6.1*.

Βήμα 1^ο: Αποεποχικοποίηση δεδομένων

Όπως παρουσιάστηκε εκτενώς στο *Κεφάλαιο 2*, η εποχιακότητα μπορεί να αυξήσει σημαντικά τη διακύμανση της χρονοσειράς μειώνοντας έτσι την προβλεπτική ικανότητα οποιουδήποτε μη εποχιακού μοντέλου πρόβλεψης κληθεί να την προεκτείνει. Ωστόσο, επειδή η εν λόγω διακύμανση είναι συστηματική και περιοδική δύναται να απομακρυνθεί αποτελεσματικά αυξάνοντας έτσι το αναμενόμενο επίπεδο ακρίβειας. Η αποεποχικοποίηση μπορεί να γίνει με διάφορες μεθόδους, βασικότερη διαφοροποίηση μεταξύ των οποίων αποτελεί ο προσθετικός ή πολλαπλασιαστικός χαρακτήρας του μοντέλου που χρησιμοποιείται για την αποσύνθεση της χρονοσειράς.

Πιο συγκεκριμένα, οι μέθοδοι πολλαπλασιαστικού χαρακτήρα θεωρούν ότι καθώς αυξάνει η τιμή ενός μεγέθους μεγαλώνουν σταδιακά και οι εποχικές διαφορές μεταξύ των παρατηρήσεων. Έτσι, χρονοσειρές με σημαντική τάση θα εμφανίζουν σε κάθε νέα περίοδο μεγαλύτερες εποχιακές διαφοροποιήσεις. Από την άλλη μεριά οι μέθοδοι προσθετικού χαρακτήρα θεωρούν ότι η εποχιακότητα είναι ανεξάρτητη της τάσης στα δεδομένα και συνεπώς οι απόλυτες διαφορές μεταξύ των παρατηρήσεων θα πρέπει να παραμένουν ίδιες σε κάθε περίοδο.

Στην θεωρία υπάρχουν κάποιες τεχνικές για να αναγνωρίσει κανείς αν μία χρονοσειρά χαρακτηρίζεται από τάση πολλαπλασιαστικής ή προσθετικής μορφής. Στην πράξη βέβαια κάτι τέτοιο είναι μάλλον δύσκολο ενώ οι διαφορές μεταξύ των δύο προσεγγίσεων δεν είναι ιδιαίτερα σημαντικές, ειδικά σε περιπτώσεις χρονοσειρών μικρού μήκους ή ασθενούς ρυθμού ανάπτυξης. Έτσι, στην παρούσα μεθοδολογία προτείνεται η χρήση της κλασικής πολλαπλασιαστικής μεθόδου αποσύνθεσης η οποία και αποτελεί βιβλιογραφικά την πλέον διαδεδομένη προσέγγιση για την αποεποχικοποίηση χρονοσειρών. Βασική προσθήκη αποτελεί ωστόσο η συρρίκνωση των συντελεστών εποχιακότητας χρησιμοποιώντας την τεχνική των *Miller-Williams* προκειμένου οι δείκτες εποχιακότητας που θα προκύψουν να είναι όσο το δυνατόν πιο αντιπροσωπευτικοί.

Σε αυτό το σημείο υπενθυμίζεται ότι υπάρχουν και άλλες τεχνικές για τη συρρίκνωση των συντελεστών, όπως είναι αυτές των *Lemon-Krutchkoff* και *James-Stein*. Η εν λόγω τεχνική προτιμάται λόγω της επιλεκτικότητάς της καθώς ανάλογα με την κυρτότητα και την

σημαντικότητα των εποχιακών δεικτών επιλέγεται διαφορετική μορφή συρρίκνωσης, όπως παρουσιάζεται στην *Παράγραφο 2.2.2*.

Σημειώνεται επίσης ότι η αποεποχικοποίηση εφαρμόζεται επιλεκτικά επί των χρονοσειρών ανάλογα με τη συστηματικότητα των διακυμάνσεων που παρουσιάζονται ανά περίοδο. Ο έλεγχος εποχιακότητας πραγματοποιείται μέσω του τεστ που παρουσιάστηκε στην *Παράγραφο 2.3.3.4* ενώ το διάστημα εμπιστοσύνης τέθηκε ίσο με 90% ($t_{crit}=1.645$). Η επιλογή του διαστήματος εμπιστοσύνης είναι τέτοιο ώστε να εξασφαλίζεται μεν η ύπαρξη επαρκών ενδείξεων για τον χαρακτηρισμό μιας χρονοσειράς ως εποχιακή, χωρίς ωστόσο ο έλεγχος να είναι εξαντλητικός. Χαρακτηριστικό παράδειγμα που αποδεικνύει την καταλληλότητά του εν λόγω διαστήματος αποτελεί ο διαγωνισμός *M3* στον οποίο οι μέθοδοι που συμμετείχαν με υψηλότερο διάστημα εμπιστοσύνης για την αποεποχικοποίηση των δεδομένων σημείωσαν χαμηλότερα επίπεδα ακρίβειας.

Συνοψίζοντας, σε αυτό το βήμα ελέγχεται η εποχιακή συμπεριφορά των δεδομένων και σε περίπτωση που εντοπιστεί εποχιακότητα αυτά αποεποχικοποιούνται χρησιμοποιώντας την πολλαπλασιαστική μέθοδο εξομάλυνσης κινητών μέσων όρων με συρρικνωμένους συντελεστές εποχιακότητας κατά *Miller-Williams*.

Βήμα 2^ο: Εξομάλυνση δεδομένων

Στο *Κεφάλαιο 4* έγινε εκτενής περιγραφή μεθόδων και τεχνικών που μπορούν να αξιοποιηθούν για την προετοιμασία της χρονοσειράς πριν την πρόβλεψη. Οι περισσότερες εξ αυτών στοχεύουν στον περιορισμό των διακυμάνσεων και την εξασφάλιση κανονικότητας ή στασιμότητας. Παρουσιάστηκαν επίσης οι λόγοι για τους οποίους η εξομάλυνση των χρονοσειρών θεωρήθηκε ως η πλέον κατάλληλη πρακτική σε σύγκριση με άλλες προσεγγίσεις όπως είναι η χρήση μετασχηματισμών, ενώ έγινε σύγκριση μεταξύ διαφόρων μεθόδων εξομάλυνσης. Η ανάλυση των πλεονεκτημάτων κάθε μεθόδου οδήγησε στην ανάπτυξη μία νέας τεχνικής εξομάλυνσης, του μετασχηματισμού εξισορρόπησης τοπικών διακυμάνσεων. Ο εν λόγω μετασχηματισμός φάνηκε να αποδίδει σαφώς καλύτερα σε σχέση με τις υπόλοιπες μεθόδους στα δεδομένα του διαγωνισμού *M3* ενώ παρουσιάζει επιπλέον το πλεονέκτημα πως μπορεί να εφαρμοστεί καθολικά σε δεδομένα κάθε συχνότητας χωρίς να απαιτείται κάποια επιπλέον παραμετροποίηση. Έτσι, η υιοθέτηση της μεθόδου φαντάζει ιδιαίτερα ωφέλιμη και πολλά υποσχόμενη.

Ταυτόχρονα παρουσιάστηκε μία μεθοδολογία προκειμένου να βελτιωθεί η τεχνική *backcasting*, το τέχνασμα που χρησιμοποιείται δηλαδή εντός κάθε τεχνικής εξομάλυνσης προκειμένου να μην παράγονται κενές τιμές στην αρχή και στο τέλος της χρονοσειράς. Αυτό έγινε καθώς η εν λόγω τεχνική αποδείχτηκε πως επηρεάζει σημαντικά την τελική επίδοση του μοντέλου πρόβλεψης και συνεπώς πρέπει να εφαρμόζεται με προσοχή.

Συνοψίζοντας, σε αυτό το βήμα τα αποεποχικοποιημένα δεδομένα εξομαλύνονται μέσω του μετασχηματισμού εξισορρόπησης τοπικών διακυμάνσεων χρησιμοποιώντας ταυτόχρονα μία βελτιωμένη έκδοση της κλασικής τεχνικής *backcasting*.

Βήμα 3^ο: Προέκταση χρονοσειράς

Στο *Κεφάλαιο 3* παρουσιάστηκαν αναλυτικά διάφορα μοντέλα πρόβλεψης χρονοσειρών τα κυριότερα εκ των οποίων ήταν τα μοντέλα εκθετικής εξομάλυνσης, τα μοντέλα *ARIMA* και η μέθοδος θ . Αναφέρθηκαν τα πλεονεκτήματα της τελευταίας και παρουσιάστηκε εμπειρικά η ανωτερότητά της, κυρίως λόγω της ικανότητάς της να αναγνωρίζει την μακροχρόνια τάση των δεδομένων και να τη συνδυάζει κατάλληλα με το τρέχον επίπεδο. Έτσι, προτάθηκε μία βελτιωμένη της έκδοση η οποία, σε αντίθεση με την κλασική προσέγγιση, είναι ικανή να μοντελοποιεί μη γραμμικά μοτίβα τάσης καθώς και να συνδυάζει τις συνιστώσες επιπέδου και τάσης είτε προσθετικά, είτε πολλαπλασιαστικά.

Δεδομένου ότι το μοτίβο τάσης, αλλά και ο μηχανισμός με τον οποίο αλληλοεπιδρούν τα δεδομένα δεν είναι εξαρχής γνωστά, εδώ καλείται κανείς να επιλέξει αφενός ποιο θα είναι το μοτίβο ανάπτυξης στο μοντέλο θ που θα χρησιμοποιηθεί και αφετέρου η έκφραση του. Βάσει της επιλογής, η οποία θα γίνει σε επόμενο βήμα, θα ορισθεί η κατάλληλη γραμμή μηδέν της μεθόδου και θα υπολογιστούν οι αντίστοιχες γραμμές θ .

Σημειώνεται πως για λόγους λιτότητας στην παρούσα διατριβή δεν εξετάζεται η αξιοποίηση πολλαπλών γραμμών θ , ούτε η χρήση εναλλακτικών τιμών της παραμέτρου. Άλλωστε οι παραπάνω προσεγγίσεις έχουν αποτελέσει αντικείμενο μελέτης σε άλλες σχετικές εργασίες και συνεπώς δεν έχει κάποια ιδιαίτερη αξία η περαιτέρω έρευνά τους. Ταυτόχρονα τα αποτελέσματα της εργασίας είναι άμεσα συγκρίσιμα με αυτά της κλασικής προσέγγισης διατηρώντας την πολυπλοκότητα της όλης διαδικασίας πρόβλεψης σε σταθερά χαμηλά επίπεδα.

Συνοπτικά, στα εξομαλυμένα δεδομένα εφαρμόζεται η γραμμική ή η μη γραμμική προσέγγιση της μεθόδου θ θεωρώντας δύο γραμμές, την TL^0 και την TL^2 . Όσον αφορά τη γραμμή μηδέν εξετάζονται πέντε διαφορετικές μορφές της: Γραμμική, εκθετική, λογαριθμική, αντίστροφη και δύναμη. Η γραμμή TL^0 προβλέπεται ντετερμινιστικά επεκτείνοντάς την κατάλληλα για τον επιθυμητό ορίζοντα βάσει της εξίσωσής της. Από την άλλη μεριά η γραμμή TL^2 προεκτείνεται μέσω της μεθόδου *Naïve*. Ο συνδυασμός των γραμμών μηδέν και δύο μπορεί να γίνει όπως αναφέρθηκε με πολλαπλασιαστικό ή με προσθετικό τρόπο.

Σημειώνεται πως η εφαρμογή της μεθόδου *Naïve* είναι ισοδύναμη εντός της προτεινόμενης μεθοδολογίας με την χρήση του μοντέλου *SES* που ορίζει η κλασική προσέγγιση καθώς όπως αναφέρθηκε νωρίτερα στην αντίστοιχη παράγραφο, για επαρκώς εξομαλυμένες χρονοσειρές η παράμετρος εξομάλυνσης a της μεθόδου *SES* ισούται με 1 και συνεπώς αυτή εκφυλίζεται σε *Naïve*. Αυτό αποτελεί ένα ακόμα πλεονέκτημα της μεθοδολογίας καθώς η πολυπλοκότητα της μεθόδου πρόβλεψης μειώνεται περιορίζοντας τις παραμετροποιήσεις που απαιτούνται για την υλοποίησή της κατά δύο: Την παράμετρο εξομάλυνσης a και το αρχικό επίπεδο I_0 . Ταυτόχρονα περιορίζεται ανάλογα και η όποια αβεβαιότητα εισέρχεται κατά την παραμετροποίηση του μοντέλου πρόβλεψης, φαινόμενο το οποίο όπως έχει αναφερθεί είναι υψίστης σημασίας και το οποίο συχνά προσπαθούμε να περιορίσουμε με χρήση τεχνικών όπως είναι ο συνδυασμός μεθόδων πρόβλεψης και η χρονική συνάθροιση δεδομένων.

Βήμα 4^ο: Εποχικοποίηση δεδομένων

Αξιοποιώντας κατάλληλα τους συρρικνωμένους δείκτες εποχιακότητας που υπολογίστηκαν στο Βήμα 1, οι προβλέψεις που παρήχθησαν εποχικοποιούνται ούτως ώστε να ακολουθήσουν την εποχιακή συμπεριφορά των δεδομένων.

Βήμα 5^ο: Επιλογή βέλτιστου μοντέλου πρόβλεψης θ

Η εφαρμογή της μεθοδολογίας μέχρι και το προηγούμενο βήμα είναι ανεξάρτητη του πλήθους, της φύσης και των χαρακτηριστικών των χρονοσειρών προς πρόβλεψη ενώ μοναδικές παράμετροι σε αυτήν αποτελεί το μοτίβο ανάπτυξης που θεωρούμε για τη χρονοσειρά και η σχέση που συνδέει τις συνιστώσες επιπέδου και τάσης, δηλαδή το μοντέλο πρόβλεψης θ . Έτσι, σε περίπτωση που υπάρχει ένδειξη για το ποια είναι η βέλτιστη γραμμή TL^0 και σχέση διασύνδεσης των συνιστωσών των δεδομένων, μπορεί κανείς να την αξιοποιήσει προκειμένου να επιλέξει το βέλτιστο μοντέλο πρόβλεψης και να επεκτείνει αποδοτικότερα τη χρονοσειρά ενδιαφέροντάς.

Η επιλογή του βέλτιστου μοντέλου μπορεί να γίνει με χρήση κατάλληλου διαγνωστικού ελέγχου, την αξιολόγηση δηλαδή της προβλεπτικής ικανότητας των διαθέσιμων μοντέλων και την επιλογή του καλύτερου βάσει ενός κριτηρίου. Έτσι, ενώ αρχικά θα υπολογιστούν 10 μεμονωμένα μοντέλα θ (προσθετικό και πολλαπλασιαστικό μοντέλο για 5 μοτίβα τάσης), το επικρατέστερο εξ αυτών θα εντοπιστεί και θα αξιοποιηθεί για την τελική προέκταση της χρονοσειράς. Αυτό θα είναι και το τελικό αποτέλεσμα της προτεινόμενης μεθοδολογίας που θα αξιολογηθεί ανάλογα.

Όσον αφορά το διαγνωστικό έλεγχο, όπως αναφέρθηκε σε προηγούμενη ενότητα αυτός μπορεί να εφαρμοστεί αξιολογώντας είτε την *in-sample* είτε την *out-of-sample* απόδοση των μοντέλων πρόβλεψης με αντίστοιχα πλεονεκτήματα και μειονεκτήματα. Στην πρώτη περίπτωση, που είναι και η ταχύτερη δυνατή, η απόδοση του μοντέλου στο μέλλον εκτιμάται μέσω της απόδοσής του στο παρελθόν, του σφάλματος δηλαδή προσαρμογής. Αντίθετα, στη δεύτερη περίπτωση, η οποία είναι και χρονικά αρκετά πιο κοστοβόρα, η αξιολόγηση γίνεται μέσω της απόδοσης του μοντέλου σε ιστορικά δεδομένα που όμως δεν χρησιμοποιήθηκαν για την εκπαίδευσή του. Όπως γίνεται αντιληπτό, η ακρίβεια της πρώτης προσέγγισης μειονεκτεί θεωρητικά έναντι της δεύτερης καθώς υποθέτει πως το μέλλον θα είναι πανομοιότυπο με το παρελθόν, κάτι που φυσικά δεν ισχύει πάντοτε. Έτσι, οι επιλογές της δεύτερης τεχνικής που δοκιμάζει την ευστάθεια των μοντέλων πρόβλεψης σε άγνωστο δείγμα παρατηρήσεων αναμένεται να είναι πιο αντιπροσωπευτικές και να οδηγήσουν σε βελτιωμένη προβλεπτική ακρίβεια.

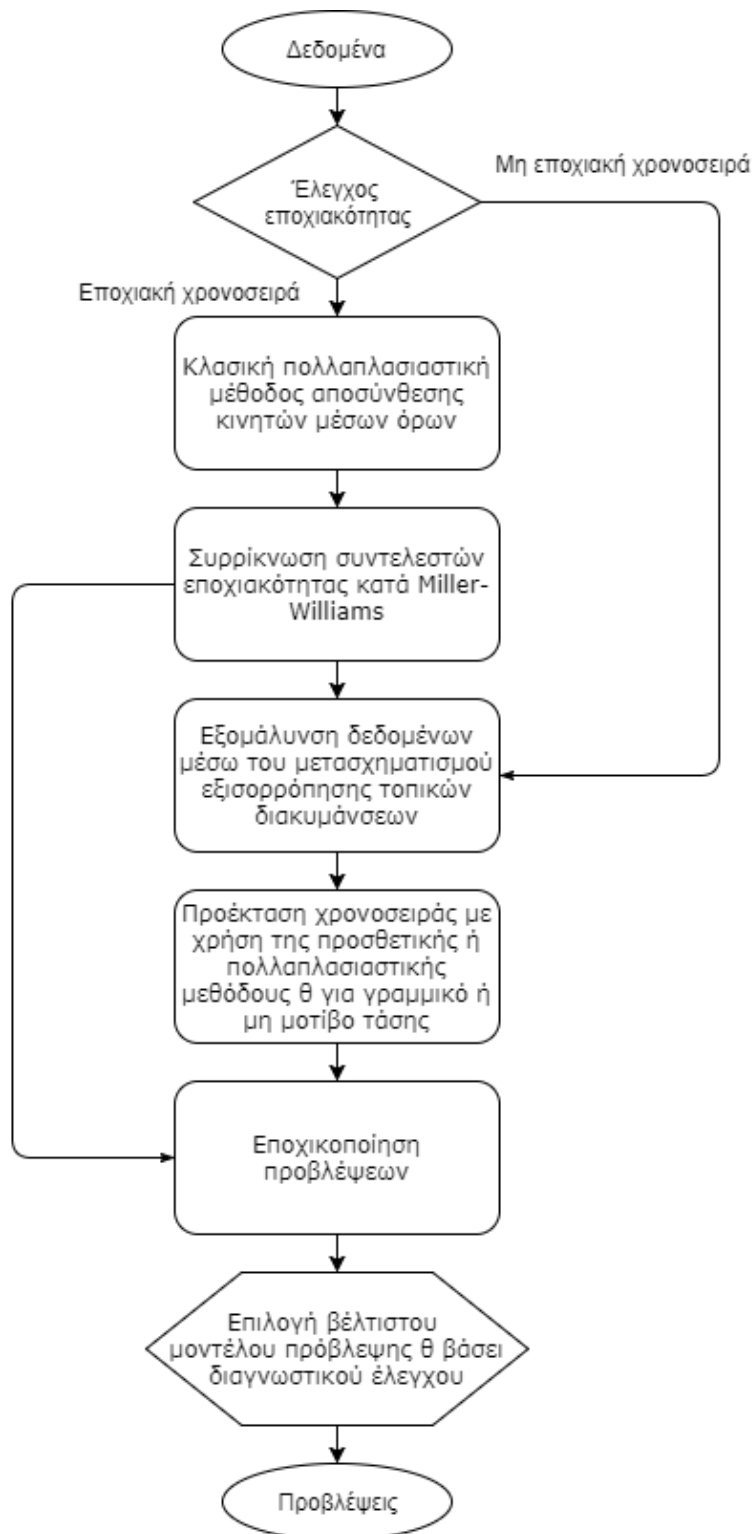
Δεδομένου ωστόσο ότι η παρούσα μεθοδολογία στοχεύει μεταξύ άλλων στην αποδοτική πρόβλεψη πληθώρας χρονοσειρών, καθώς και του ότι η προτεινόμενη διαδικασία πρόβλεψης έχει επιβαρυνθεί ήδη υπολογιστικά μέσω της τεχνικής εξομάλυνσης, η χρήση μίας εξαντλητικής τεχνικής επιλογής βέλτιστου μοντέλου πρόβλεψης κρίνεται μάλλον άσκοπη. Άλλωστε, λόγω της αβεβαιότητας που χαρακτηρίζει τα δεδομένα, η χρήση *out-of-sample* τεχνικών σε καμία περίπτωση δεν εγγυάται τον εντοπισμό της βέλτιστης δυνατής λύσης, ενώ παράλληλα έχει το μειονέκτημα πως απαιτεί για την εκτέλεσή της

μεγάλο πλήθος δεδομένων για την παραμετροποίηση και της αξιολόγηση των μεθόδων, κάτι που δεν είναι πάντοτε εφικτό.

Έτσι, αποφασίστηκε για την επιλογή του βέλτιστου μοντέλου πρόβλεψης θ να υπολογίζεται το μέσο *in-sample* τετραγωνικό σφάλμα (*MSE*) για όλα τα υποψήφια μοντέλα και να επιλέγεται εκείνο που το ελαχιστοποιεί. Η εν λόγω απλουστευμένη προσέγγιση χρησιμοποιείται και από αρκετούς άλλους δημοφιλείς αλγορίθμους αυτόματης παραγωγής προβλέψεων, όπως π.χ. το *ETS* (*Hyndman et al., 2002*) για την αυτόματη προέκταση χρονοσειρών με χρήση μεθόδων εκθετικής εξομάλυνσης, με ικανοποιητικά αποτελέσματα. Με αυτόν τον τρόπο η επιλογή γίνεται με μεγάλη ταχύτητα και υπό φυσιολογικές συνθήκες συνάδουν επαρκώς με την πραγματικότητα.

Σε αυτό το σημείο αναφέρεται πως η επιλογή του κριτηρίου *MSE* έγινε κυρίως λόγω της ευρείας χρήσης του στη βιβλιογραφία σε εφαρμογές βελτιστοποίησης μοντέλων πρόβλεψης (παραμετροποίηση και επιλογή μεθόδου) καθώς και του γεγονότος πως άλλα σχετικά σφάλματα, όπως π.χ. το *MAE*, οδηγούσαν σε παρεμφερή αποτελέσματα. Επίσης, αναφέρεται πως η χρήση *information criteria* ως εναλλακτικών για την αξιολόγηση μεταξύ άλλων και της πολυπλοκότητας της μεθόδου πρόβλεψης, όπως π.χ. το *AIC* και το *BIC*, δεν έχουν κανένα νόημα στην εν λόγω περίπτωση καθώς όλα τα μοντέλα που εξετάζονται είναι της ίδιας ακριβώς πολυπλοκότητας. Έτσι, η χρήση του *MSE* φαντάζει μονόδρομος και *information criteria* προτείνεται να χρησιμοποιηθούν εναλλακτικά μόνο σε περίπτωση που κάποιος επιλέξει να χρησιμοποιήσει γραμμές μηδέν πολυπλοκότητας μεγαλύτερης του δύο, όπως π.χ. ένα πολυώνυμο.

Τέλος, σημειώνεται πως στην παρούσα μεθοδολογία η παράμετρος θ παραμένει για λόγους απλότητας σταθερή και ίση με δύο. Σε περίπτωση που υπάρχει βέβαια γνώση σχετικά με το βαθμό που η μακροπρόθεσμη τάση πρέπει να συμμετέχει στην παραγωγή της τελικής πρόβλεψης, αυτή μπορεί να αξιοποιηθεί χωρίς πρόβλημα στην προτεινόμενη μέθοδο αυξομειώνοντας αντίστοιχα την παράμετρο στο επιλεχθέν μοντέλο πρόβλεψης.



Εικόνα 6.1: Διάγραμμα ροής για την ολοκληρωμένη εφαρμογή της προτεινόμενης μεθοδολογίας πρόβλεψης χρονοσειρών.

6.2 Πειραματική διάταξη

Σε αυτό το σημείο θα παρουσιαστεί η πειραματική διάταξη που θα χρησιμοποιηθεί για την αξιολόγηση της επίδοσης της προτεινόμενης μεθοδολογίας παραγωγής προβλέψεων. Αυτή περιλαμβάνει δύο διαφορετικά σετ δεδομένων: Ένα με πραγματικές χρονοσειρές και ένα με τεχνητές. Το πρώτο σετ δεν είναι άλλο από τα δεδομένα των διαγωνισμών *M1* και *M3* (4004 χρονοσειρές), ενώ το δεύτερο είναι το σετ των 135.000 χρονοσειρών που κατασκευάστηκε μέσω της γεννήτριας που παρουσιάστηκε στο [Κεφάλαιο 5](#). Στόχος είναι η εξακρίβωση της αποτελεσματικότητας της μεθοδολογίας σε πλήθος δεδομένων ποικίλων χαρακτηριστικών. Σε αυτό το σημείο σημειώνεται μάλιστα πως, αν και τα πραγματικά δεδομένα παρουσιάζουν ανά περιπτώσεις μεταβολές που καθορίζονται από εξωτερικές μεταβλητές και οι οποίες είναι αδύνατο να αναπαραχθούν πιστά με τεχνητό τρόπο, στο σύνολο ενός μεγάλου πλήθους χρονοσειρών τέτοια φαινόμενα μπορούμε να θεωρήσουμε ότι εξασθενούν σημαντικά. Έτσι, η αξιολόγηση της προτεινόμενης μεθοδολογίας στο σύνολο των τεχνητών χρονοσειρών θα προσδώσει μία ιδιαίτερα αντιπροσωπευτική αξιολόγηση για πραγματικές της εφαρμογές.

Για λόγους κατανόησης τα αποτελέσματα καθενός από τα σετ δεδομένων θα παρουσιαστούν ξεχωριστά και στο τέλος θα συγκριθούν μεταξύ τους προκειμένου να εξακριβωθεί η συνέπεια των συμπερασμάτων και να σχολιαστούν τυχόν διαφοροποιήσεις. Οι παράμετροι που θα μελετηθούν για τα δύο σετ χρονοσειρών παρατίθενται στη συνέχεια.

Μέθοδοι πρόβλεψης:

Συνολικά μελετάται η επίδοση 10 διαφορετικών μοντέλων πρόβλεψης τα οποία δεν είναι άλλα από τη γραμμική και μη εκδοχή της μεθόδου θ για τα μοτίβα τάσης που παρουσιάστηκαν νωρίτερα (5 συνολικά), εκφρασμένα είτε σε προσθετική είτε σε πολλαπλασιαστική μορφή. Σημειώνεται ότι τα εν λόγω μοντέλα εφαρμόζονται για δύο γραμμές θ με τιμές 0 και 2 και προεκτείνονται όπως ορίστηκε σε προηγούμενη παράγραφο.

1. Γραμμική τάση (**LIN**)
 2. Εκθετική τάση (**EXP**)
 3. Λογαριθμική τάση (**LOG**)
 4. Αντίστροφη τάση (**INV**)
 5. Τάση σε μορφή δύναμης (**POW**)
-
- A. Προσθετική έκφραση (**A**)
 - B. Πολλαπλασιαστική έκφραση (**M**)

Μέθοδοι αναφοράς:

Προκειμένου να εξακριβωθεί η επίδοση της μεθοδολογίας για κάθε μία από τις μεθόδους πρόβλεψης απαιτείται η σύγκριση της απόδοσής τους σε σχέση με κάποιες μεθόδους αναφοράς. Αυτό γίνεται αφενός επειδή η απόλυτη τιμή ενός μέτρου σφάλματος δεν λείει

κάτι από μόνη της και αφετέρου γιατί προκειμένου να κριθεί η προτεινόμενη μεθοδολογία ανώτερη από τις κλασικές προσεγγίσεις οφείλει να τις ξεπερνάει πάντα.

Έτσι, ως μέθοδοι αναφοράς τίθενται εδώ τα ίδια τα 10 μοντέλα θ που προκύπτουν όταν εφαρμοστούν απευθείας στα δεδομένα χωρίς κάποια περεταίρω επεξεργασία. Άλλοι κλασικοί μέθοδοι πρόβλεψης, όπως π.χ. μοντέλα εκθετικής εξομάλυνσης και ARIMA, δεν θεωρούνται κατάλληλες για αναφορά καθώς ήδη η απλούστερη προσέγγιση της μεθόδου θ τις ξεπερνά.

Φυσικά, η βασική μέθοδος αναφοράς θα είναι σε κάθε περίπτωση η κλασική μέθοδος θ καθώς, ανεξαρτήτως του οφέλους που προκύπτει από την προετοιμασία της χρονοσειράς, η οποία και αξίζει να αξιολογηθεί ξεχωριστά, οι νέες εκφράσεις της μεθόδου αποτελούν από μόνες τους βελτιώσεις και πρέπει να μελετηθούν αντίστοιχα.

Αξιολόγηση της επίδοσης των μεθόδων:

Η επίδοση κάθε μεθόδου αποφασίστηκε να αξιολογηθεί χρησιμοποιώντας ως κριτήριο το συμμετρικό απόλυτο ποσοστιαίο σφάλμα **sMAPE** καθώς το βασικότερο ζητούμενο στις προβλέψεις είναι η εξασφάλιση ικανοποιητικής ακρίβειας. Ο συγκεκριμένος δείκτης αποτελεί λόγω των ιδιοτήτων του ένα από τα καταλληλότερα μέτρα ακρίβειας καθώς διαθέτει το πλεονέκτημα της συμμετρικότητας όταν πρόκειται για ακραίες ή μηδενικές τιμές. Επίσης, η ποσοστιαία έκφρασή του καθιστά δυνατή την ταυτόχρονη αξιολόγηση της μεθόδου πρόβλεψης κατά μήκος πολλών χρονοσειρών διαφορετικής κλίμακας.

Σε αυτό το σημείο αναφέρουμε πως εξετάστηκε η χρήση και άλλων δεικτών σφάλματος ωστόσο απορρίφθηκαν καθώς αφενός οι μη ποσοστιαίοι δείκτες σφαλμάτων αδυνατούν να μας ενημερώσουν για την επίδοση μιας μεθόδου σε πλήθος διαφορετικών χρονοσειρών, και αφετέρου η χρήση σχετικών σφαλμάτων είναι ανούσια δεδομένου του ότι έχουν ήδη τεθεί μέθοδοι αναφοράς. Επιπλέον το **sMAPE**, το οποίο αποτέλεσε το βασικότερο κριτήριο επίδοσης στους διαγωνισμούς M , έχει επικρατήσει να χρησιμοποιείται ως το βασικό κριτήριο απόδοσης σε διαγωνισμούς προβλέψεων και σχετικές μελέτες. Τέλος, η χρήση ενός επιπλέον μέτρου για την ταυτόχρονη αξιολόγηση της μεθόδου από άποψη προκατάληψης θα καθιστούσε την ανάλυση των αποτελεσμάτων ιδιαίτερα πολύπλοκη χωρίς κάποιο ιδιαίτερο πρόσθετο όφελος, ειδικά από τη στιγμή που το εν λόγω χαρακτηριστικό θεωρείται δευτερευούσης σημασίας

Ορίζοντας πρόβλεψης:

Για λόγους συνάφειας των αποτελεσμάτων της παρούσας εργασίας με αυτά των διαγωνισμών M αποφασίστηκε ο ορίζοντας πρόβλεψης να είναι ανάλογος της συχνότητας των δεδομένων. Έτσι, αυτός ορίστηκε ίσος με $h=18$, 8 και 6 περιόδους για μηνιαία, τριμηνιαία και μη εποχιακά δεδομένα αντίστοιχα. Για κάθε χρονοσειρά λοιπόν μήκους n τα δεδομένα χωρίζονται σε δύο τμήματα: το δείγμα για την εκπαίδευση του μοντέλου πρόβλεψης (*in-sample*) και το δείγμα για την αξιολόγηση της επίδοσής του (*out-sample*). Το *in-sample* διαθέτει μήκος $n-h$ και το *out-sample* μήκος h στο οποίο και υπολογίζεται η τιμή του **sMAPE** για την εκάστοτε χρονοσειρά.

Κατηγορίες δεδομένων:

Δεδομένου ότι στην παρούσα διατριβή προτείνεται μία ολοκληρωμένη μεθοδολογία για την αποδοτική πρόβλεψη χρονοσειρών πληθώρας χαρακτηριστικών και ιδιοτήτων, κρίνεται απαραίτητο να ελεγχθεί η επίδοσή της όχι μόνο κατά μήκος όλων των χρονοσειρών αλλά και ανά κατηγορία δεδομένων. Η κατηγοριοποίηση μπορεί να γίνει βάσει:

- **Συχνότητας**, π.χ. ετήσιες ή μηνιαίες χρονοσειρές
- **Είδους εφαρμογής**, π.χ. δημογραφικά ή βιομηχανικά δεδομένα
- **Έντασης χαρακτηριστικού**, π.χ. εποχιακότητας ή τυχαιότητας
- **Είδους τάσης**, π.χ. γραμμικής ή εκθετικής ανάπτυξης

Η κατηγοριοποίηση βάσει συχνότητας και εφαρμογής θα εφαρμοστεί λόγω της φύσης των δεδομένων μόνο για την περίπτωση των δεδομένων του διαγωνισμού *M* ενώ η μελέτη ανά ένταση χαρακτηριστικού θα γίνει και για την περίπτωση των τεχνητών δεδομένων για τα οποία και υπάρχει διαθέσιμη η εν λόγω πληροφορία. Τέλος, η αξιολόγηση ανά είδος τάσης είναι εφικτή μόνο για την περίπτωση των τεχνητών δεδομένων καθώς για αυτά του διαγωνισμού δεν την γνωρίζουμε εξ αρχής και συνεπώς δεν κρίνεται αξιόπιστη μία τέτοια κατηγοριοποίηση όποια μεθοδολογία και αν αξιοποιηθεί για την πραγματοποίησή της.

Στη συνέχεια παρουσιάζονται αναλυτικά τα πρωτογενή δεδομένα και τα αποτελέσματα της προσέγγισης για κάθε σετ χρονοσειρών ξεχωριστά.

6.2.1 Αποτελέσματα μεθοδολογίας για τα τεχνητά δεδομένα της γεννήτριας

Όπως παρουσιάστηκε στο αντίστοιχο κεφάλαιο, η γεννήτρια που αναπτύχθηκε χρησιμοποιήθηκε προκειμένου να δημιουργήσει 135.000 τεχνητές χρονοσειρές ποικίλων εντάσεων χαρακτηριστικών. Στο εν λόγω κεφάλαιο παρουσιάζονται τα εν λόγω δεδομένα όπως και τα αποτελέσματα για πρόβλεψή τους μέσω της κλασικής και της προτεινόμενης προσέγγισης.

6.2.1.1 Κατηγοριοποίηση τεχνητών δεδομένων

Συχνότητα:

Δεδομένου ότι η γεννήτρια κατασκευάστηκε ούτως ώστε να παράγει μηνιαία χρονοσειρές, η συχνότητα στο παρόν δείγμα δεδομένων είναι ενιαία. Σημειώνεται ότι η παραπάνω υπόθεση εφαρμόζεται χωρίς βλάβη της γενικότητας καθώς, βάσει των αποτελεσμάτων των διαγωνισμών *M* και άλλων σχετικών μελετών, η επίδραση της συχνότητας στο σφάλμα είναι αμελητέα. Αντίθετα, σημασία έχει αναμφισβήτητα η έντασή της τάσης, του θορύβου και της εποχιακότητας, η οποία και λαμβάνεται κανονικά υπόψη κατά την παραγωγή των τεχνητών χρονοσειρών.

Είδος εφαρμογής:

Καθώς πρόκειται για τεχνητά δεδομένα, είναι προφανές πως δεν υπάρχει αντιστοιχία μεταξύ χρονοσειρών και είδους εφαρμογής.

Ένταση χαρακτηριστικού:

Όπως παρουσιάστηκε στο *Κεφάλαιο 5*, τα δεδομένα δύναται να χωριστούν σε υποκατηγορίες ανάλογα με την ένταση του κάθε χαρακτηριστικού τους: Τάσης, τυχαιότητας και εποχιακότητας. Λόγου του τρόπου με τον οποίο παρήχθησαν τα δεδομένα, το πλήθος των χρονοσειρών είναι ισοκαταναμημένο ως εξής:

Πίνακας 6.1: Κατηγοριοποίηση τεχνητών δεδομένων γεννήτριας βάσει έντασης χαρακτηριστικών.

Συχνότητα	Χαμηλό	Μεσαίο	Υψηλό
Εποχιακότητα	45,000	45,000	45,000
Τυχαιότητα	45,000	45,000	45,000
Γραμμική τάση	9,000	9,000	9,000
Εκθετική τάση	9,000	9,000	9,000
Λογαριθμική τάση	9,000	9,000	9,000
Αντίστροφη τάση	9,000	9,000	9,000
Τάση σε μορφή δύναμης	9,000	9,000	9,000

6.2.1.2 Αξιολόγηση μεθοδολογίας βάσει τεχνητών δεδομένων

Έχοντας προβεί στην κατηγοριοποίηση των δεδομένων βάσει έντασης χαρακτηριστικών, καθίσταται εφικτή η αξιολόγηση της προτεινόμενης μεθοδολογίας κατά μήκος διαφορετικών αξόνων. Ως μέτρο σύγκρισης χρησιμοποιείται όπως αναφέρθηκε η επίδοση των 10 μοντέλων θ επί των αποεποχικοποιημένων δεδομένων, η οποία και αντιπροσωπεύει την απλούστερη δυνατή προσέγγιση για την εφαρμογή τους. Σημειώνεται βέβαια ότι η θεώρηση μη γραμμικών μοτίβων και η πολλαπλασιαστική έκφραση της μεθόδου αποτελεί από μόνη της μια σημαντική διαφοροποίηση σε σχέση με την κλασική προσέγγιση της μεθόδου θ , η επίδοση της οποίας θεωρείται για το λόγο αυτό και το βασικότερο μέτρο αναφοράς.

Σε πρώτη φάση, προκειμένου να εξακριβωθεί η αποτελεσματικότητα της προτεινόμενης μεθοδολογίας πρόβλεψης, παρουσιάζεται η μέση ακρίβεια τεσσάρων διαφορετικών προσεγγίσεων στο σύνολο των τεχνητών χρονοσειρών. Σκοπός είναι να μελετηθεί το κέρδος που επιφέρει προβλεπτικά κάθε ένα από τα βασικά βήματα της μεθοδολογίας: Αποεποχικοποίηση, εξομάλυνση και προέκταση μέσω μη γραμμικών μοτίβων τάσης.

Κλασική προσέγγιση: Πρόβλεψη με χρήση των διαφόρων εκδοχών της μεθόδου θ . Η αποεποχικοποίηση εφαρμόζεται μέσω της κλασικής μεθόδου αποσύνθεσης.

Κλασική προσέγγιση με συρρίκνωση εποχιακών συντελεστών: Πρόβλεψη με χρήση των διαφόρων εκδοχών της μεθόδου θ . Η αποεποχικοποίηση εφαρμόζεται έχοντας προβεί σε συρρίκνωση συντελεστών κατά *Miller-Williams*.

Κλασική προσέγγιση με εξομάλυνση δεδομένων: Πρόβλεψη με χρήση των διαφόρων εκδοχών της μεθόδου θ . Η αποεποχικοποίηση εφαρμόζεται μέσω της κλασικής μεθόδου αποσύνθεσης ενώ η αποεποχικοποιημένη χρονοσειρά επιδέχεται εξομάλυνση.

Προτεινόμενη προσέγγιση: Πρόβλεψη με χρήση των διαφόρων εκδοχών της μεθόδου θ . Η αποεποχικοποίηση εφαρμόζεται έχοντας προβεί σε συρρίκνωση συντελεστών κατά *Miller-Williams* ενώ η αποεποχικοποιημένη χρονοσειρά επιδέχεται εξομάλυνση.

Τα αποτελέσματα ανά προσέγγιση φαίνονται στον *Πίνακα 6.2* και είναι υπολογισμένα για το δείκτη ακρίβειας *sMAPE*. Με την ένδειξη *Selective* σημειώνεται η απόδοση της προτεινόμενης μεθοδολογίας μετά και την επιλογή του καλύτερου μοντέλου πρόβλεψης θ βάσει σφάλματος προσαρμογής. Αντίστοιχα, με τις ενδείξεις *A* και *M* σημειώνεται η έκφραση του μοντέλου (προσθετική ή πολλαπλασιαστική) και ολογράφως το μοτίβο τάσης.

Πίνακας 6.2: Επίδοση προτεινόμενης μεθοδολογίας σε σχέση με άλλες απλούστερες προσεγγίσεις στα τεχνητά δεδομένα της γεννήτριας - Αναλυτικά αποτελέσματα.

Μοντέλο	Κλασική προσέγγιση	Κλασική προσέγγιση με συρρίκνωση εποχιακών συντελεστών	Κλασική προσέγγιση με εξομάλυνση δεδομένων	Προτεινόμενη προσέγγιση
Linear - A	19.26	19.23	19.35	19.32
Exponential- A	18.97	18.94	18.81	18.78
Logarithmic- A	19.44	19.40	19.24	19.21
Inverse- A	19.46	19.43	19.34	19.31
Power- A	19.31	19.28	19.11	19.08
Linear- M	19.53	19.50	19.78	19.75
Exponential- M	19.06	19.02	18.85	18.82
Logarithmic- M	19.43	19.40	19.27	19.24
Inverse- M	19.44	19.41	19.39	19.35
Power- M	19.20	19.17	19.08	19.05
Selective	18.72	18.68	18.63	18.52

Όπως παρατηρείται, πέραν του γραμμικού μοτίβου τάσης, για όλα τα μη γραμμικά μοντέλα θ η προτεινόμενη προσέγγιση οδηγεί συνολικά σε σημαντικά καλύτερες προβλέψεις. Στη γραμμική προσέγγιση η συρρίκνωση συντελεστών είναι η βέλτιστη λύση, με την εξομάλυνση να μεγεθύνει ελαφρώς το σφάλμα. Οι παράγοντες που διαμορφώνουν την παραπάνω εικόνα θα μελετηθούν εξονυχιστικά στη συνέχεια.

Το βασικό που πρέπει να παρατηρήσει κανείς εδώ σε πρώτη φάση είναι πως η προτεινόμενη λύση είναι και η βέλτιστη δυνατή, μειώνοντας το σφάλμα σε σχέση με την κλασική προσέγγιση κατά 3.8%. Επίσης, η συρρίκνωση των εποχιακών συντελεστών και η εξομάλυνση οδηγούν σε σταδιακή βελτίωση του μέσου σφάλματος σε όλες τις περιπτώσεις, τόσο ξεχωριστά όσο και συνδυαστικά.

Ενδιαφέρον παρουσιάζει ακόμα το γεγονός πως, πέραν του γραμμικού μοντέλου, οι διαφορές μεταξύ προσθετικής και πολλαπλασιαστικής έκφρασης είναι αρκετά περιορισμένες, με την προσθετική να είναι ελαφρώς καλύτερη. Αυτό ενισχύεται μάλιστα μετά την εφαρμογή της προτεινόμενης μεθοδολογίας για την καλύτερη προετοιμασία των αρχικών δεδομένων. Έτσι, για λόγους οικονομίας και κατανόησης από εδώ και πέρα τα μοντέλα θα μελετώνται συνολικά ανά μοτίβο παρουσιάζοντας τα μέσα αποτελέσματα των δύο εκφράσεων. Ο προηγούμενος πίνακας απλοποιείται λοιπόν στον *Πίνακα 6.3*.

Αξίζει να σημειωθεί πάντως πως, ενώ τα πολλαπλασιαστικά μοντέλα εμφανίζουν μεγαλύτερο μέσο σφάλμα από τα προσθετικά, η επιλογή του καλύτερου εξ αυτών γίνεται αποτελεσματικά μειώνοντας το τελικό σφάλμα της μεθοδολογίας. Έτσι αποδεικνύεται πως μεμονωμένα κάθε μοντέλο θ μπορεί να προσδώσει επιπλέον αξία προβλεπτικά στην προτεινόμενη προσέγγιση και πως η εισαγωγή των νέων εκφράσεων της μεθόδου συνεισφέρει καθοριστικά στην ευελιξία της μεθόδου και στη συνολική της απόδοση.

Πίνακας 6.3: Επίδοση προτεινόμενης μεθοδολογίας σε σχέση με άλλες απλούστερες προσεγγίσεις στα τεχνητά δεδομένα της γεννήτριας – Συνοπτικά αποτελέσματα.

Μοντέλο	Κλασική προσέγγιση	Κλασική προσέγγιση με συρρίκνωση με εποχιακών συντελεστών	Κλασική προσέγγιση με εξομάλυνση δεδομένων	Προτεινόμενη προσέγγιση
Linear	19.40	19.37	19.57	19.54
Exponential	19.02	18.98	18.83	18.80
Logarithmic	19.44	19.40	19.26	19.23
Inverse	19.45	19.42	19.37	19.33
Power	19.26	19.23	19.10	19.07
Selective	18.72	18.68	18.63	18.52

Πιο αναλυτικά, η γενικευμένη μέθοδος θ με εκθετικό μοτίβο τάσης φαίνεται να είναι η πιο ακριβής προσέγγιση από τα επιμέρους μοντέλα, μειώνοντας το μέσο σφάλμα πρόβλεψης στην κλασική προσέγγιση κατά 2.0% και κατά 3.1% στην προτεινόμενη. Ακολουθεί η τάση δύναμης, η λογαριθμική, η αντίστροφη και η γραμμική για την περίπτωση της προτεινόμενης προσέγγισης, και η δύναμη, η γραμμική, η λογαριθμική και η αντίστροφη για την κλασική. Φυσικά, η επιλογή του βέλτιστου μοντέλου είναι σταθερά η καλύτερη δυνατή προσέγγιση με οφέλη που κυμαίνονται από 2.8% έως 3.8%, ανάλογα με την προσέγγιση.

Χαρακτηριστικό είναι επίσης το γεγονός ότι ανά προσέγγιση τα παραπάνω ποιοτικά συμπεράσματα διατηρούνται με ελάχιστες και μικρές διαφοροποιήσεις. Έτσι συμπεραίνουμε πως συνολικά, ανεξαρτήτως της προσέγγισης που ακολουθείται, η επιλογή ενός αντιπροσωπευτικότερου μοτίβου τάσης, όπως είναι αυτό της εκθετικής τάσης, και μοντέλου, όπως είναι αυτό του προσθετικού, μπορεί να βελτιώσει σημαντικά την ακρίβεια των προβλέψεων.

Χρησιμοποιώντας ως μέτρο ακρίβειας την επίδοση της κάθε μεθόδου στην κλασική προσέγγιση, παρατηρούμε επίσης ότι η συρρίκνωση των εποχιακών συντελεστών έχει ως

Κεφάλαιο 6: Πειραματική διάταξη - Ανάλυση και εξέταση των αποτελεσμάτων

αποτέλεσμα τη μέση βελτίωση της ακρίβειας κατά 0.16%. Χαρακτηριστικό είναι επίσης το γεγονός ότι η μέθοδος βελτιώνει την μέση ακρίβεια σε όλα τα μοτίβα τάσης ανεξαιρέτως.

Η βελτίωση είναι σημαντικά μεγαλύτερη για την περίπτωση της εξομάλυνσης η οποία κυμαίνεται σε σχέση με την κλασική προσέγγιση και ανάλογα με το μοτίβο τάσης από 0.4% έως 1.0%. Η βελτίωση από την εξομάλυνση γίνεται μάλιστα αντιληπτή για όλα τα μη γραμμικά μοτίβα τάσης ανεξαιρέτως, με το γραμμικό να είναι το μόνο που εμφανίζει αρνητική επίπτωση.

Στην περίπτωση της προτεινόμενης μεθοδολογίας τα οφέλη είναι ακόμα μεγαλύτερα καθώς η μέση βελτίωση σε σχέση με την κλασική προσέγγιση ανάλογα με το μοτίβο τάσης που επιλέγεται κυμαίνεται από 0.6% έως 1.1%. Το ποσοστό αυτό είναι ίσο με το άθροισμα των επιμέρους οφελών από τις τεχνικές συρρίκνωσης εποχιακών συντελεστών και εξομάλυνσης δεδομένων, κάτι το οποίο αποδεικνύει ότι οι δύο τεχνικές είναι ανεξάρτητες μεταξύ τους και μπορούν και συνδυάζονται προσθετικά οδηγώντας σε σαφώς πιο βελτιωμένα αποτελέσματα.

Χρησιμοποιώντας ως μέτρο αναφοράς την κλασική προσέγγιση της μεθόδου θ , συμπεραίνουμε ότι για χρήση εκθετικού μοτίβου τάσης η ακρίβεια πρόβλεψης μπορεί να βελτιωθεί κατά 2.1% στην περίπτωση συρρίκνωσης εποχιακών συντελεστών, κατά 2.9% στην περίπτωση εξομάλυνσης των δεδομένων και κατά 3.1% για συνδυασμό των δύο τεχνικών. Η προτεινόμενη μεθοδολογία οδηγεί ακόμα σε βελτίωση της τάξης του 0.9% για λογαριθμικό μοτίβο τάσης, 0.3% για αντίστροφο και 1.7% για τάση σε μορφή δύναμης. Αυτό είναι αρκετά ενθαρρυντικό δεδομένου ότι για μη εφαρμογή π.χ. της εξομάλυνσης τα αντίστοιχα μη γραμμικά μοντέλα υστερούν έναντι της κλασικής μεθόδου. Η εξομάλυνση συντελεί δηλαδή καθοριστικά στη βελτίωση της προβλεπτικής ακρίβειας των μεθόδων πρόβλεψης καθιστώντας αποδοτικές εν τέλει ακόμα και μεθόδους που θεωρούνταν αρχικά ανακριβείς.

Εξετάζοντας την επίδοση των επιμέρους μεθόδων ως προς την ένταση της εποχιακότητας, μπορεί κανείς να παρατηρήσει πως καθώς αυξάνεται το επίπεδο της αυξάνεται και η απόδοση της προτεινόμενης προσέγγισης. Το κέρδος κυμαίνεται ανάλογα το μοντέλο από 0.6% έως 1.5% για την περίπτωση της υψηλής έντασης με μέση τιμή το 1%. Αυτό πέφτει στο 0.7% για τη μεσαία ένταση και μηδενίζει για την χαμηλή, όπως και αναμένεται για μη εποχιακές χρονοσειρές. Η επιλεκτική προσέγγιση εμφανίζει μέσο κέρδος σε όλες τις περιπτώσεις περί του 1%. Σημειώνεται επίσης πως στην περίπτωση της υψηλής έντασης εποχιακότητας η προτεινόμενη προσέγγιση βελτιώνει την απόδοση και του γραμμικού μοντέλου, πράγμα που σημαίνει ότι η συρρίκνωση των εποχιακών συντελεστών υπερνικά την επίδραση της εξομάλυνσης, η οποία και όπως είδαμε νωρίτερα μειώνει συνολικά την απόδοση του συγκεκριμένου μοντέλου.

Συνοψίζοντας, η προτεινόμενη μεθοδολογία φαίνεται να βελτιώνει στη γενική περίπτωση την ακρίβεια της πρόβλεψης, ειδικά όταν πρόκειται για χρονοσειρές με έντονη εποχιακότητα. Οι βελτιώσεις αυτές εξαρτώνται σημαντικά από το μοντέλο τάσης που επιλέγεται και αυξάνονται για εποχιακότητα μεσαίας και υψηλής έντασης.

Πίνακας 6.4: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση εποχιακότητας.

Εποχιακότητα	Linear	Exponential	Logarithmic	Inverse	Power	Selective
Κλασική προσέγγιση						
Χαμηλό	17.18	16.70	17.24	17.27	17.04	16.49
Μεσαίο	19.42	19.03	19.46	19.50	19.30	18.78
Υψηλό	20.62	20.30	20.64	20.64	20.47	19.92
Προτεινόμενη προσέγγιση						
Χαμηλό	17.79	16.65	17.13	17.17	16.88	16.32
Μεσαίο	19.57	18.83	19.24	19.36	19.10	18.59
Υψηλό	20.49	19.99	20.39	20.52	20.26	19.71

Όσον αφορά το επίπεδο της τυχαιότητας, όπως παρατηρείται η ακρίβεια πρόβλεψης βελτιώνεται στη γενική περίπτωση μέσω της προτεινόμενης μεθοδολογίας για όλα τα μοντέλα τάσης, ακόμα και για χρονοσειρές με χαμηλά επίπεδα θορύβου. Όπως είναι αναμενόμενο η βελτίωση αυξάνει καθώς αυξάνει το επίπεδο του θορύβου με τα ποσοστά να φτάνουν το 1.0%, 1.4% και 0.5% για την περίπτωση χαμηλού, μεσαίου και υψηλού επιπέδου θορύβου. Το περιορισμένο ποσοστό κέρδους για χρονοσειρές υψηλού θορύβου μπορεί να εξηγηθεί από το γεγονός πως όταν η τυχαιότητα είναι μεγάλη, η αναγνώριση του σήματος καθίσταται δευτερευούσης σημασίας για τον προσδιορισμό του μέλλοντος, ο οποίος και είναι πρακτικά αδύνατο να πραγματοποιηθεί αποδοτικά ακόμα και με αναλυτικές μεθόδους.

Αξίζει να σημειωθεί πως το γραμμικό μοντέλο είναι το μόνο που εμφανίζει χειρότερη απόδοση κατά την εφαρμογή της προτεινόμενης μεθοδολογίας και μάλιστα μόνο στην περίπτωση της χαμηλής έντασης θορύβου. Έτσι, συμπεραίνουμε πως η αφαίρεση του θορύβου είναι καλύτερο να γίνεται μόνο σε περιπτώσεις όπου αυτή είναι σχετικά αυξημένη καθώς σε αντίθετη περίπτωση μπορεί να οδηγήσει σε αλλοίωση του σήματος και αφαίρεση χρήσιμης πληροφορίας. Αυτό ισχύει ιδιαίτερα στο γραμμικό μοντέλο το οποίο λόγω της ακαμψίας του δεν μπορεί να συμμορφωθεί το ίδιο εύκολα σε χρονοσειρές πολύπλοκων μακροχρόνιων μοτίβων.

Στην περίπτωση του επιλεκτικού μοντέλου, τα κέρδη είναι σημαντικά σε κάθε περίπτωση και συγκεκριμένα ίσα με 3.8%, 1.0% και 0.5% για την περίπτωση χαμηλού, μεσαίου και υψηλού επιπέδου θορύβου. Συμπεραίνουμε λοιπόν πως ο διαγνωστικός έλεγχος αποδίδει ανεξαρτήτως θορύβου και αποφεύγει κακοτοπιές οδηγώντας σε σταθερά βελτιωμένες προβλέψεις.

Συμπερασματικά, το μοντέλο τάσης φαίνεται να επηρεάζει καθοριστικά την ακρίβεια που επιτυγχάνεται μέσω της προτεινόμενης μεθοδολογίας, η οποία ωστόσο βελτιώνει σταθερά την μέση απόδοσή των μοντέλων σε κάθε περίπτωση. Όταν ο θόρυβος είναι λίγος τα οφέλη από την εξομάλυνση είναι περιορισμένα έως και αρνητικά, ενώ όταν αυξάνεται σημαντικότερα.

Κεφάλαιο 6: Πειραματική διάταξη - Ανάλυση και εξέταση των αποτελεσμάτων

Πίνακας 6.5: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση τυχαιότητας.

Τυχαιότητα	Linear	Exponential	Logarithmic	Inverse	Power	Selective
Κλασική προσέγγιση						
Χαμηλό	8.72	8.09	9.00	9.14	8.73	7.85
Μεσαίο	18.19	17.83	18.19	18.15	18.02	17.51
Υψηλό	29.57	29.40	29.44	29.40	29.34	29.06
Προτεινόμενη προσέγγιση						
Χαμηλό	9.41	7.98	8.75	8.92	8.40	7.58
Μεσαίο	18.08	17.49	17.85	17.94	17.74	17.33
Υψηλό	29.45	29.20	29.39	29.44	29.34	28.92

Η αξιολόγηση της απόδοσης της μεθόδου ανά είδος και ένταση τάσης οδηγεί και αυτή με τη σειρά της σε παρεμφερή συμπεράσματα. Η επιλεκτική μέθοδος είναι η πλέον αποδοτική, οδηγώντας ανά ένταση γραμμικής τάσης σε κέρδη από 0.5% έως 1.2%. Ωστόσο, σε αυτήν την περίπτωση τα μέσα κέρδη της προτεινόμενης προσέγγισης περιορίζονται καθώς αυξάνει η ένταση της τάσης, προφανώς λόγω της ανικανότητας των μεμονωμένων μοντέλων να ακολουθήσουν πιστά το σήμα οποιασδήποτε χρονοσειράς σε ακραίες καταστάσεις. Για την περίπτωση της εκθετικής τάσης τα κέρδη διαμορφώνονται αντίστοιχα ανά επίπεδο έντασης από 0.7% έως 1.4%, για τη λογαριθμική από 0.6% έως 1.8%, για την αντίστροφη από 0.2% έως 0.8% και για τη τάση δύναμης από 0.1% έως 1.4%.

Πίνακας 6.6: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση γραμμικής τάσης.

Γραμμική τάση	Linear	Exponential	Logarithmic	Inverse	Power	Selective
Κλασική προσέγγιση						
Χαμηλό	18.68	18.30	18.31	18.12	18.23	17.82
Μεσαίο	19.46	19.10	19.90	20.16	19.68	18.92
Υψηλό	22.22	21.84	23.27	23.71	22.75	22.05
Προτεινόμενη προσέγγιση						
Χαμηλό	18.94	17.96	18.04	18.00	17.91	17.61
Μεσαίο	19.28	18.96	19.78	20.07	19.59	18.73
Υψηλό	22.43	22.00	23.17	23.53	22.90	21.93

Πίνακας 6.7: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση εκθετικής τάσης.

Εκθετική τάση	Linear	Exponential	Logarithmic	Inverse	Power	Selective
<i>Κλασική προσέγγιση</i>						
Χαμηλό	18.86	18.78	18.60	18.42	18.58	18.34
Μεσαίο	19.51	19.10	19.13	19.07	18.99	18.60
Υψηλό	20.08	19.26	21.52	22.04	20.98	19.61
<i>Προτεινόμενη προσέγγιση</i>						
Χαμηλό	18.97	18.53	18.45	18.40	18.40	18.21
Μεσαίο	19.68	18.72	18.85	18.90	18.69	18.34
Υψηλό	20.23	19.43	21.36	21.85	21.02	19.43

Πίνακα 6.8: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση λογαριθμικής τάσης.

Λογαριθμική τάση	Linear	Exponential	Logarithmic	Inverse	Power	Selective
<i>Κλασική προσέγγιση</i>						
Χαμηλό	17.53	17.38	17.49	17.46	17.46	17.12
Μεσαίο	20.63	19.88	20.43	20.34	20.18	19.36
Υψηλό	21.76	21.42	22.33	22.62	21.93	21.36
<i>Προτεινόμενη προσέγγιση</i>						
Χαμηλό	17.79	17.26	17.41	17.43	17.33	17.01
Μεσαίο	20.71	19.41	20.02	20.12	19.77	19.02
Υψηλό	21.70	21.36	22.15	22.43	21.92	21.17

Πίνακα 6.9: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση αντίστροφης τάσης.

Αντίστροφη τάση	Linear	Exponential	Logarithmic	Inverse	Power	Selective
<i>Κλασική προσέγγιση</i>						
Χαμηλό	17.03	16.93	17.25	17.34	17.23	16.90
Μεσαίο	19.06	18.66	19.41	19.58	19.24	18.44
Υψηλό	22.03	21.38	21.67	21.49	21.33	20.76
<i>Προτεινόμενη προσέγγιση</i>						
Χαμηλό	17.13	16.99	17.27	17.37	17.23	16.87
Μεσαίο	19.30	18.53	19.30	19.49	19.09	18.25
Υψηλό	22.15	20.84	21.16	21.20	20.90	20.39

Κεφάλαιο 6: Πειραματική διάταξη - Ανάλυση και εξέταση των αποτελεσμάτων

Πίνακα 6.10: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση τάσης σε μορφή δύναμης.

Τάση δύναμης	Linear	Exponential	Logarithmic	Inverse	Power	Selective
Κλασική προσέγγιση						
Χαμηλό	18.00	18.01	17.99	18.00	17.99	17.91
Μεσαίο	19.46	19.30	19.24	19.08	19.20	18.77
Υψηλό	19.94	19.27	20.15	20.27	19.82	19.02
Προτεινόμενη προσέγγιση						
Χαμηλό	18.08	18.03	18.03	18.03	18.02	17.89
Μεσαίο	19.53	18.93	19.00	19.01	18.91	18.55
Υψηλό	20.15	19.05	19.86	20.06	19.59	18.76

Αξίζει επίσης να σημειωθεί ότι, ενώ η απόδοση όλων των μοντέλων μειώνεται με την αύξηση της έντασης τάσης ανεξαρτήτως μοτίβου, και ενώ η επιλεκτική προσέγγιση είναι η καλύτερη για το σύνολο των χρονοσειρών, μεμονωμένα υπάρχουν μοντέλα τα οποία υπερτερούν έναντι άλλων για συγκεκριμένα είδη τάσης. Αυτό συμβαίνει καθώς ανάλογα με το μοτίβο μακροχρόνιας τάσης που επικρατεί, διαφορετικές γραμμές μηδέν πρέπει να χρησιμοποιούνται για την αποτελεσματική προέκτασή του. Αυτός είναι άλλωστε και ο βασικός λόγος για τον οποίο αποφασίστηκε εξαρχής στην παρούσα διατριβή να επεκταθεί η κλασική γραμμική μέθοδος θ για μη γραμμικά μοτίβα τάσης.

Η απόδοση των διαφόρων μοντέλων ανά είδος τάσης παρουσιάζεται αναλυτικά στον Πίνακα 6.11 για την κλασική και την προτεινόμενη προσέγγιση και όπως παρατηρείται υπάρχουν αρκετές διαφοροποιήσεις σε σχέση με τους προηγούμενους πίνακες. Το εκθετικό μοντέλο αποδίδει καλύτερα σε χρονοσειρές γραμμικού και εκθετικού μοτίβου, το μοντέλο δύναμης σε χρονοσειρές λογαριθμικού και γραμμικού μοτίβου, το γραμμικό μοντέλο σε χρονοσειρές μοτίβου δύναμης και το αντίστροφο σε χρονοσειρές αντίστοιχου μοτίβου. Μάλιστα, ο παραπάνω διαχωρισμός ισχύει ανεξαρτήτως της επεξεργασίας που εφαρμόζεται μέσω της προτεινόμενης προσέγγισης.

Τα παραπάνω συμπεράσματα είναι αρκετά λογικά αν σκεφτεί κανείς τον τρόπο με τον με τον οποίο λειτουργεί το κλασικό γραμμικό μοντέλο θ . Δεδομένου ότι η μακροχρόνια τάση ενσωματώνεται κατά το ήμισυ στις προβλέψεις λόγω της γραμμής 2, γίνεται αντιληπτό πως αν η τάση συνεχιστεί αυτούσια στο μέλλον - πράγμα που εδώ συμβαίνει λόγω του τρόπου με τον οποίο κατασκευάζονται τα τεχνητά δεδομένα - οι προβλέψεις που θα προκύψουν θα είναι θεωρητικά αισιόδοξες στην περίπτωση φθίνουσας τάσης και απαισιόδοξες στην περίπτωση αύξουσας τάσης. Έτσι, αν κανείς θελήσει να περιορίσει το εν λόγω φαινόμενο χρησιμοποιώντας μία εναλλακτική γραμμή μηδέν (και όχι καλύτερη παράμετρο θ), θα μπορούσε να αξιοποιήσει την εκθετική που μεταβάλλεται μη γραμμικά με το χρόνο και συνεπώς μπορεί να περιορίσει τέτοια φαινόμενα. Αντίθετα, στην περίπτωση της εκθετικής, της λογαριθμικής και αντίστροφης τάσης, όπου το μοτίβο είναι συγκεκριμένο και λόγω της καμπυλότητάς του αποδοτικά επεκτάσιμο μέσω αντίστοιχων γραμμών, το εκθετικό, το λογαριθμικό και το αντίστροφο μοντέλο είναι τα βέλτιστα δυνατά.

Τέλος, για την περίπτωση της δύναμης δεν μπορούν να υπάρξουν ασφαλή συμπεράσματα καθώς τα κοίλα της μακροπρόθεσμης τάσης μπορούν να διαφέρουν ανάλογα με την περίπτωση, οπότε και το εκθετικό ή το λογαριθμικό μοντέλο θα είναι τα βέλτιστα δυνατά. Ίσως αυτός να είναι και ο λόγος που σε αυτό το μοτίβο το γραμμικό μοντέλο, δηλαδή η μέση οδός, είναι και η βέλτιστη λύση.

Πίνακα 6.11: Επίδοση προτεινόμενης προσέγγισης στα τεχνητά δεδομένα της γεννήτριας ανά ένταση τάσης σε μορφή δύναμης.

Είδος Τάσης	Linear	Exponential	Logarithmic	Inverse	Power	Selective
<i>Κλασική προσέγγιση</i>						
Γραμμικό	18.93	18.52	20.01	20.44	19.59	18.91
Εκθετικό	19.44	18.78	20.38	20.67	20.19	18.86
Λογαριθμικό	18.14	18.45	18.03	18.20	18.01	18.05
Αντίστροφο	22.51	21.21	20.73	19.75	20.52	19.73
Δύναμη	17.95	18.11	18.02	18.19	17.95	18.03
<i>Προτεινόμενη προσέγγιση</i>						
Γραμμικό	19.02	18.57	19.81	20.18	19.56	18.68
Εκθετικό	19.61	19.04	20.44	20.77	20.28	18.87
Λογαριθμικό	18.09	18.38	18.11	18.24	18.08	18.01
Αντίστροφο	23.00	19.99	19.71	19.28	19.40	19.08
Δύναμη	17.94	18.02	18.07	18.19	18.02	17.96

Συνοπτικά η μελέτη των αποτελεσμάτων στο συγκεκριμένο δείγμα τεχνητών δεδομένων μας οδηγεί στα παρακάτω βασικά συμπεράσματα:

- ✓ Η συρρίκνωση των εποχιακών δεικτών βελτιώνει συστηματικά τις προβλέψεις των εποχιακών χρονοσειρών.
- ✓ Όσο αυξάνεται η εποχιακότητα, τόσο αυξάνει και το κέρδος της προτεινόμενης προσέγγισης.
- ✓ Στη γενική περίπτωση η εξομάλυνση των χρονοσειρών οδηγεί σε βελτιωμένη προβλεπτική ακρίβεια. Εξαιρέσεις ενδέχεται να υπάρξουν σε χρονοσειρές που είναι ήδη επαρκώς ομαλές.
- ✓ Από ένα επίπεδο θορύβου και μετά η προτεινόμενη προσέγγιση αδυνατεί να προσδώσει σημαντικά κέρδη καθώς η τυχαιότητα υπερέχει του σήματος.
- ✓ Όσο αυξάνεται η τάση της χρονοσειράς τόσο μειώνονται τα προσδοκώμενα οφέλη της μεθόδου.
- ✓ Για λογικά επίπεδα τάσης και τυχαιότητας η προτεινόμενη προσέγγιση οδηγεί σε σημαντικά καλύτερες προβλέψεις.
- ✓ Τα οφέλη συρρίκνωσης εποχιακών δεικτών και εξομάλυνσης είναι ανεξάρτητα μεταξύ τους.
- ✓ Η επιλογή ενός αντιπροσωπευτικού μοτίβου τάσης, όπως το εκθετικό, μπορεί να βελτιώσει σημαντικά την ακρίβεια των προβλέψεων ανεξαρτήτως συχνότητας χρονοσειράς και είδους εφαρμογής.

Κεφάλαιο 6: Πειραματική διάταξη - Ανάλυση και εξέταση των αποτελεσμάτων

- ✓ Ανάλογα με το μοτίβο τάσης της χρονοσειράς διαφορετικά μοντέλα αναμένεται να οδηγήσουν σε καλύτερες προβλέψεις: Εκθετικό για γραμμική τάση, Εκθετικό για εκθετική τάση, Λογαριθμικό για Λογαριθμική Τάση, Αντίστροφο για αντίστροφη τάση και Γραμμικό για τάση σε μορφή δύναμης.
- ✓ Τα προσθετικά μοντέλα είναι γενικά πιο ακριβή από τα πολλαπλασιαστικά. Ωστόσο οι διαφορές είναι σχετικά μικρές και σε μεμονωμένες περιπτώσεις και οι δύο εκφράσεις μπορούν να προσδώσουν οφέλη.
- ✓ Η επιλογή του βέλτιστου μοντέλου θ οδηγεί συνολικά σε καλύτερες προβλέψεις. Μεμονωμένα φυσικά ενδέχεται να μην επιλέγεται πάντα η βέλτιστη λύση.
- ✓ Με την εφαρμογή της προτεινόμενης μεθοδολογίας μπορεί κανείς να επιλέξει με σχετική ασφάλεια στη γενική περίπτωση το βέλτιστο μοντέλο θ για κάθε χρονοσειρά και να αποκομίσει σημαντικά οφέλη. Ωστόσο, αν η ένταση της τυχαιότητας είναι ιδιαίτερα χαμηλή, τότε ενδέχεται να πρέπει να παρακαμφθεί το βήμα της εξομάλυνσης.

6.2.2 Αποτελέσματα μεθοδολογίας για τα δεδομένα των διαγωνισμών M

Οι διαγωνισμοί $M1$ και $M3$ αποτελούνται από 4004 χρονοσειρές και τα δεδομένα τους κατηγοριοποιούνται βάσει συχνότητας και εφαρμογής. Ο διαγωνισμός $M1$ διαθέτει συνολικά 1001 χρονοσειρές ενώ ο $M3$ 3003 και μπορούν να κατηγοριοποιηθούν σύμφωνα με την παρακάτω ανάλυση.

6.2.2.1 Κατηγοριοποίηση δεδομένων στους διαγωνισμούς M

Συχνότητα:

Τα δεδομένα χωρίζονται σε τέσσερις κατηγορίες βάσει συχνότητας: Ετήσια, τριμηνιαία, μηνιαία και άλλα δεδομένα. Η κατηγορία «άλλα» υπάρχει μόνο στην περίπτωση του διαγωνισμού $M3$. Στον Πίνακα 6.12 παρουσιάζεται η κατανομή των χρονοσειρών ανά συχνότητα. Όπως παρατηρείται, οι μηνιαίες χρονοσειρές αποτελούν από μόνες τους το 50% του όγκου των δεδομένων των διαγωνισμών οπότε και η μέθοδος που θα τις προβλέψει καλύτερα αναμένεται να λάβει σημαντικό προβάδισμα στην τελική κατάταξη βάσει του συνόλου των χρονοσειρών. Αυτός είναι και ο λόγος που κρίνεται απαραίτητη η μελέτη των μεθόδων ανά συχνότητα δεδομένων.

Πίνακας 6.12: Κατηγοριοποίηση δεδομένων διαγωνισμού M βάσει συχνότητας.

Συχνότητα	M1	M3	Σύνολο	Σύνολο (%)
Ετήσια	181	645	826	20.6
Τριμηνιαία	203	756	959	24.0
Μηνιαία	617	1428	2045	51.1
Άλλα	0	174	174	4.3
Σύνολο	1001	3003	4004	100.0

Είδος εφαρμογής:

Τα δεδομένα χωρίζονται σε τέσσερις κατηγορίες βάσει εφαρμογής: Μικροοικονομικά - MICRO, Βιομηχανία - INDUSTRY, Μακροοικονομικά - MACRO, Χρηματοοικονομικά - FINANCE, Δημογραφικά - DEMOG και λοιπά - OTHER). Η κατηγορία «Άλλα» και «Χρηματοοικονομικά» υπάρχει μόνο στην περίπτωση του διαγωνισμού Μ3. Στον Πίνακα 6.13 παρουσιάζεται η κατανομή των χρονοσειρών ανά είδος εφαρμογής. Όπως παρατηρείται, τα Μικροοικονομικά και Μακροοικονομικά δεδομένα αποτελούν πάνω από το 50% του όγκου των δεδομένων των διαγωνισμών οπότε και η μέθοδος που θα τις προβλέψει καλύτερα αναμένεται να λάβει σημαντικό προβάδισμα στην τελική κατάταξη βάσει του συνόλου των χρονοσειρών. Αυτός είναι και ο λόγος που κρίνεται απαραίτητη η μελέτη των μεθόδων ανά είδος εφαρμογής.

Πίνακας 6.13: Κατηγοριοποίηση δεδομένων διαγωνισμού Μ βάσει είδους εφαρμογής.

Είδος εφαρμογής	M1	M3	Σύνολο	Σύνολο (%)
MICRO	302	828	1130	28.2
INDUSTRY	236	519	755	18.9
MACRO	319	731	1050	26.2
FINANCE	0	308	308	7.7
DEMOG	144	413	557	13.9
OTHER	0	204	204	5.1
Σύνολο	1001	3003	4004	100.0

Ένταση χαρακτηριστικού:

Όπως παρουσιάστηκε στο Κεφάλαιο 5, τα δεδομένα δύναται να χωριστούν σε υποκατηγορίες ανάλογα με την ένταση του κάθε χαρακτηριστικού τους: Τάσης, συχνότητας και εποχιακότητας. Στους Πίνακες 6.14 και 6.15 παρουσιάζεται λοιπόν η κατανομή των χρονοσειρών ανά υποκατηγορία για επίπεδο έντασης «Υψηλό», «Μεσαίο» και «Χαμηλό».

Πίνακας 6.14: Κατηγοριοποίηση δεδομένων διαγωνισμού Μ3 βάσει έντασης χαρακτηριστικών.

Συχνότητα	Χαμηλό	Μεσαίο	Υψηλό
Εποχιακότητα	2374	373	256
Τυχαιότητα	2364	367	272
Γραμμική τάση	2045	373	585
Εκθετική τάση	1301	595	1107
Λογαριθμική τάση	2449	292	262
Αντίστροφη τάση	2526	250	227
Τάση σε μορφή δύναμης	1506	805	692

Πίνακας 6.15: Κατηγοριοποίηση δεδομένων διαγωνισμού M1 βάσει έντασης χαρακτηριστικών.

Συχνότητα	Χαμηλό	Μεσαίο	Υψηλό
Εποχιακότητα	628	173	200
Τυχειότητα	809	140	52
Γραμμική τάση	651	122	228
Εκθετική τάση	395	214	392
Λογαριθμική τάση	798	81	122
Αντίστροφη τάση	830	67	104
Τάση σε μορφή δύναμης	482	286	233

Όπως παρατηρείται, και στους δύο διαγωνισμούς η συντριπτική πλειοψηφία των χρονοσειρών (περίπου 80%) διαθέτει χαρακτηριστικά χαμηλής έντασης, ενώ λίγες είναι εκείνες που χαρακτηρίζονται από υψηλή διακύμανση, εποχιακότητα ή τάση. Ακόμα μεγαλύτερο ενδιαφέρον παρουσιάζει το γεγονός ότι ελάχιστες χρονοσειρές ακολουθούν αμιγώς το μοντέλο ανάπτυξης του λογαρίθμου ή του αντιστρόφου, κάτι που αποδεικνύει τη σημασία διερεύνησης του βαθμού που ένα τέτοιο μοτίβο επηρεάζει την απόδοση των μοντέλων πρόβλεψης. Έτσι, κρίνεται απαραίτητη η μελέτη των μεθόδων ανά ένταση χαρακτηριστικού.

6.2.2.2 Αξιολόγηση μεθοδολογίας βάσει δεδομένων διαγωνισμών M

Έχοντας προβεί στην κατηγοριοποίηση των δεδομένων των διαγωνισμών M βάσει συχνότητας, είδους εφαρμογής και έντασης χαρακτηριστικών, καθίσταται εφικτή η αξιολόγηση της προτεινόμενης μεθοδολογίας σε πληθώρα διαφορετικών αξόνων. Ως μέτρο σύγκρισης χρησιμοποιείται και σε αυτήν την περίπτωση η επίδοση των 5 μη γραμμικών μορφών της μεθόδου θ επί των αποεποχικοποιημένων δεδομένων, η οποία και αντιπροσωπεύει την απλούστερη δυνατή προσέγγιση για την εφαρμογή τους.

Σημειώνεται βέβαια ότι η θεώρηση μη γραμμικών μοτίβων καθώς και αθροιστική ή πολλαπλασιαστικής σχέσης μεταξύ επιπέδου και τάσης, αποτελούν από μόνες τους σημαντικές διαφοροποιήσεις σε σχέση με την κλασική προσέγγιση της μεθόδου θ . Έτσι, η επίδοση της τελευταίας θα θεωρηθεί και το βασικότερο μέτρο αναφοράς.

Σε πρώτη φάση, προκειμένου να εξακριβωθεί λοιπόν η αποτελεσματικότητα της προτεινόμενης μεθοδολογίας πρόβλεψης, παρουσιάζεται ανά συχνότητα δεδομένων η ακρίβεια τεσσάρων διαφορετικών προσεγγίσεων. Σκοπός είναι να μελετηθεί το κέρδος που επιφέρει προβλεπτικά κάθε ένα από τα βασικά βήματα της μεθοδολογίας: Αποεποχικοποίηση, εξομάλυνση και προέκταση μέσω γενικευμένης μεθόδου θ .

Κλασική προσέγγιση: Πρόβλεψη με χρήση των διαφόρων εκδοχών της μεθόδου θ . Η αποεποχικοποίηση εφαρμόζεται μέσω της κλασικής μεθόδου αποσύνθεσης.

Κλασική προσέγγιση με συρρίκνωση εποχιακών συντελεστών: Πρόβλεψη με χρήση των διαφόρων εκδοχών της μεθόδου θ . Η αποεποχικοποίηση εφαρμόζεται έχοντας προβεί σε συρρίκνωση συντελεστών κατά *Miller-Williams*.

Κλασική προσέγγιση με εξομάλυνση δεδομένων: Πρόβλεψη με χρήση των διαφόρων εκδοχών της μεθόδου θ . Η αποεποχικοποίηση εφαρμόζεται μέσω της κλασικής μεθόδου αποσύνθεσης ενώ η αποεποχικοποιημένη χρονοσειρά επιδέχεται εξομάλυνση.

Προτεινόμενη προσέγγιση: Πρόβλεψη με χρήση των διαφόρων εκδοχών της μεθόδου θ . Η αποεποχικοποίηση εφαρμόζεται έχοντας προβεί σε συρρίκνωση συντελεστών κατά *Miller-Williams* ενώ η αποεποχικοποιημένη χρονοσειρά επιδέχεται εξομάλυνση.

Τα αποτελέσματα ανά προσέγγιση φαίνονται στον *Πίνακα 6.16* και είναι υπολογισμένα για το δείκτη ακρίβειας *sMAPE*. Με την ένδειξη *Selective* σημειώνεται η απόδοση της προτεινόμενης μεθοδολογίας μετά και την επιλογή του καλύτερου μοντέλου πρόβλεψης θ βάσει σφάλματος προσαρμογής. Αντίστοιχα, με τις ενδείξεις *A* και *M* σημειώνεται η έκφραση του μοντέλου (προσθετική ή πολλαπλασιαστική) και ολογράφως το μοτίβο τάσης.

Πίνακας 6.16: Επίδοση προτεινόμενης μεθοδολογίας σε σχέση με άλλες απλούστερες προσεγγίσεις στα δεδομένα των διαγωνισμών M – Αναλυτικά αποτελέσματα.

Μοντέλο	Κλασική προσέγγιση	Κλασική προσέγγιση με συρρίκνωση εποχιακών συντελεστών	Κλασική προσέγγιση με εξομάλυνση δεδομένων	Προτεινόμενη προσέγγιση
Linear - A	13.89	13.86	13.64	13.62
Exponential- A	13.52	13.49	13.25	13.24
Logarithmic- A	14.04	14.02	13.89	13.88
Inverse- A	14.31	14.29	14.17	14.16
Power- A	13.90	13.89	13.80	13.79
Linear- M	14.03	14.00	13.69	13.69
Exponential- M	13.43	13.40	13.27	13.26
Logarithmic- M	14.11	14.09	13.93	13.91
Inverse- M	14.55	14.48	14.24	14.23
Power- M	13.95	13.93	13.79	13.78
Selective	13.47	13.43	13.39	13.37

Όπως παρατηρείται, για όλα τα μοντέλα θ που κατασκευάστηκαν η προτεινόμενη προσέγγιση οδηγεί συνολικά σε σημαντικά καλύτερες προβλέψεις, κατά μέσο όρο της τάξεως του 1.6%. Μάλιστα, το κλασικό μοντέλο θ παρουσιάζει ακόμα μεγαλύτερα οφέλη που αγγίζουν το 2.2%, με το εκθετικό να ακολουθεί με παρόμοιο ποσοστό βελτίωσης. Επίσης, είναι ενθαρρυντικό το γεγονός πως κάθε τεχνική προετοιμασίας δεδομένων οδηγεί μεμονωμένα σε όλες τις περιπτώσεις σε αξιοσημείωτες βελτιώσεις, αποδεικνύοντας με αυτό τον τρόπο την αποτελεσματικότητά τους και την ορθότητα των επιλογών που έγιναν κατά τη δόμηση της προτεινόμενης συνδυαστικής μεθοδολογίας. Ενδεικτικά αναφέρεται πως το μέσο κέρδος που προκύπτει σε ακρίβεια από τη συρρίκνωση των εποχιακών συντελεστών αγγίζει το 0.2%, ενώ από την εξομάλυνση το 1.38%.

Κεφάλαιο 6: Πειραματική διάταξη - Ανάλυση και εξέταση των αποτελεσμάτων

Κάτι που πρέπει να παρατηρήσει κανείς εδώ είναι πως η προτεινόμενη επιλεκτική λύση, σε αντίθεση με το σετ των τεχνητών δεδομένων, δεν είναι και η βέλτιστη δυνατή ανά προσέγγιση καθώς όπως φαίνεται τα δύο εκθετικά μοντέλα ξεπερνούν το επιλεγόμενο μοντέλο από 0.03% έως 0.14%. Αυτό οφείλεται πιθανότατα στο γεγονός ότι στην περίπτωση των τεχνητών δεδομένων το μοτίβο τάσης διατηρείται εξ ορισμού κατά μήκος της χρονοσειράς, ακόμα και στην περίπτωση χρονοσειρών υψηλής τυχαιότητας. Έτσι, αν εντοπιστεί σωστά το μοτίβο τάσης στα ιστορικά δεδομένα, η πρόβλεψη για το μέλλον θα είναι λογικά και η βέλτιστη δυνατή. Αντίθετα, στην περίπτωση των πραγματικών δεδομένων, όπου και το μοτίβο τάσης ενδέχεται να αλλάξει αναπάντεχα στο μέλλον, δεν ισχύει η προηγούμενη παραδοχή οδηγώντας σε αρκετές περιπτώσεις σε εσφαλμένες επιλογές μοντέλου. Μάλιστα, δεν είναι καθόλου τυχαίο το γεγονός πως τα εκθετικά μοντέλα ανταποκρίνονται καλύτερα στις πραγματικές χρονοσειρές αφού η κλίση τους εξυπηρετεί μεταβολές στην τάση (damped), προσομοιώνοντας έτσι καλύτερα δεδομένα με έντονες διακυμάνσεις και αναπάντεχες μεταβολές επιπέδου.

Ενδιαφέρον παρουσιάζει ακόμα το γεγονός πως, πέραν του γραμμικού μοντέλου στην κλασική προσέγγιση, οι διαφορές μεταξύ προσθετικής και πολλαπλασιαστικής έκφρασης είναι αρκετά περιορισμένες, με την προσθετική να είναι ελαφρώς καλύτερη. Αυτό ενισχύεται μάλιστα μετά την εφαρμογή της προτεινόμενης μεθοδολογίας για την καλύτερη προετοιμασία των αρχικών δεδομένων. Έτσι, για λόγους οικονομίας και κατανόησης από εδώ και πέρα τα μοντέλα θα μελετώνται συνολικά ανά μοτίβο παρουσιάζοντας τα μέσα αποτελέσματα των δύο εκφράσεων. Ο προηγούμενος πίνακας απλοποιείται λοιπόν στον *Πίνακα 6.17*.

Αξίζει να σημειωθεί πάντως πως, ενώ τα πολλαπλασιαστικά μοντέλα εμφανίζουν μεγαλύτερο μέσο σφάλμα από τα προσθετικά, η επιλογή του καλύτερου εξ αυτών γίνεται αρκετά αποτελεσματικά μειώνοντας το τελικό σφάλμα της μεθοδολογίας. Έτσι αποδεικνύεται πως μεμονωμένα κάθε μοντέλο θ μπορεί να προσδώσει επιπλέον αξία προβλεπτικά στην προτεινόμενη προσέγγιση και πως η εισαγωγή των νέων εκφράσεων της μεθόδου συνεισφέρει καθοριστικά στην ευελιξία της μεθόδου και στη συνολική της απόδοση.

Πίνακας 6.17: Επίδοση προτεινόμενης μεθοδολογίας σε σχέση με άλλες απλούστερες προσεγγίσεις στα δεδομένα των διαγωνισμών – Συνοπτικά αποτελέσματα.

Μοντέλο	Κλασική προσέγγιση	Κλασική προσέγγιση με συρρίκνωση εποχιακών συντελεστών	Κλασική προσέγγιση με εξομάλυνση δεδομένων	Προτεινόμενη προσέγγιση
Linear	13.96	13.93	13.66	13.66
Exponential	13.47	13.45	13.26	13.25
Logarithmic	14.07	14.05	13.91	13.90
Inverse	14.43	14.38	14.21	14.19
Power	13.93	13.91	13.79	13.78
Selective	13.47	13.43	13.39	13.37

Πιο αναλυτικά, η γενικευμένη μέθοδος θ με εκθετικό μοτίβο τάσης φαίνεται να είναι η πιο ακριβής προσέγγιση από τα επιμέρους μοντέλα, μειώνοντας το μέσο σφάλμα πρόβλεψης στην κλασική προσέγγιση κατά 3.5% και κατά 5.1% στην προτεινόμενη. Ακολουθεί η γραμμική, η τάση δύναμης, η λογαριθμική και η αντίστροφη για την περίπτωση της προτεινόμενης προσέγγισης, και η δύναμη, η αντίστροφη, η γραμμική και η λογαριθμική και για την κλασική. Η επιλογή του βέλτιστου μοντέλου είναι η καλύτερη δυνατή λύση για την κλασική προσέγγιση και την προσέγγιση συρρίκνωσης συντελεστών, και η δεύτερη καλύτερη για τις υπόλοιπες με οφέλη που κυμαίνονται από 3.5% έως 4.2%, ανάλογα με την προσέγγιση.

Χαρακτηριστικό είναι επίσης το γεγονός ότι ανά προσέγγιση τα παραπάνω ποιοτικά συμπεράσματα διατηρούνται με ελάχιστες και μικρές διαφοροποιήσεις. Έτσι συμπεραίνουμε πως συνολικά, ανεξαρτήτως της προσέγγισης που ακολουθείται, η επιλογή ενός αντιπροσωπευτικότερου μοτίβου τάσης, όπως είναι αυτό της εκθετικής τάσης, και μοντέλου, όπως είναι αυτό του προσθετικού, μπορεί να βελτιώσει σημαντικά την ακρίβεια των προβλέψεων. Αναλυτικότερα, στο 29% των περιπτώσεων επιλέγεται γραμμικό μοντέλο, στο 34% εκθετικό, στο 14% δύναμη και αντίστροφο ενώ στο 10% λογαριθμικό. Επίσης, στο 66% των περιπτώσεων επιλέγεται προσθετικό μοντέλο, ενώ στο 44% πολλαπλασιαστικό, με το πλέον δημοφιλές και στις δύο περιπτώσεις να είναι το εκθετικό.

Χρησιμοποιώντας ως μέτρο ακρίβειας την επίδοση της κάθε μεθόδου στην κλασική προσέγγιση, παρατηρούμε επίσης ότι η συρρίκνωση των εποχιακών συντελεστών έχει ως αποτέλεσμα τη μέση βελτίωση της ακρίβειας κατά 0.2%. Χαρακτηριστικό είναι επίσης το γεγονός ότι η μέθοδος βελτιώνει την μέση ακρίβεια σε όλα τα μοτίβα τάσης ανεξαιρέτως.

Η βελτίωση είναι σημαντικά μεγαλύτερη για την περίπτωση της εξομάλυνσης η οποία κυμαίνεται σε σχέση με την κλασική προσέγγιση και ανάλογα με το μοτίβο τάσης από 0.6% έως 2.1%. Η βελτίωση από την εξομάλυνση γίνεται μάλιστα αντιληπτή για όλα τα μοτίβα τάσης ανεξαιρέτως, με τη μικρότερη βελτίωση να εμφανίζεται στην επιλεκτική προσέγγιση. Η τελευταία παρατήρηση οφείλεται πιθανότατα στο γεγονός ότι, μετά την εξομάλυνση της χρονοσειράς, το μοτίβο των δεδομένων μεταβάλλεται σημαντικά. Έτσι, είναι πολύ πιθανό η μακροπρόθεσμη τάση να αλλοιώνεται σε αρκετές περιπτώσεις και να επιλέγονται αντίστοιχα μη αντιπροσωπευτικά μοντέλα θ .

Για την περίπτωση της προτεινόμενης μεθοδολογίας, τα οφέλη είναι ακόμα μεγαλύτερα καθώς η μέση βελτίωση σε σχέση με την κλασική προσέγγιση ανάλογα με το μοτίβο τάσης που επιλέγεται κυμαίνεται από 1.1% έως 2.1%. Το ποσοστό αυτό είναι ίσο με το άθροισμα των επιμέρους οφελών από τις τεχνικές συρρίκνωσης εποχιακών συντελεστών και εξομάλυνσης δεδομένων, κάτι το οποίο αποδεικνύει ότι οι δύο τεχνικές είναι ανεξάρτητες μεταξύ τους και μπορούν και συνδυάζονται προσθετικά οδηγώντας σε σαφώς πιο βελτιωμένα αποτελέσματα.

Χρησιμοποιώντας ως μέτρο αναφοράς την κλασική προσέγγιση της μεθόδου θ , συμπεραίνουμε ότι για χρήση εκθετικού μοτίβου τάσης η ακρίβεια πρόβλεψης μπορεί να βελτιωθεί κατά 3.7% στην περίπτωση συρρίκνωσης εποχιακών συντελεστών, κατά 5.0%

Κεφάλαιο 6: Πειραματική διάταξη - Ανάλυση και εξέταση των αποτελεσμάτων

στην περίπτωση εξομάλυνσης των δεδομένων και κατά 5.1% για συνδυασμό των δύο τεχνικών. Η προτεινόμενη μεθοδολογία οδηγεί ακόμα σε βελτίωση της τάξης του 0.4% για λογαριθμικό μοτίβο τάσης και 1.3% για τάση σε μορφή δύναμης. Αυτό είναι αρκετά ενθαρρυντικό δεδομένου ότι για μη εφαρμογή π.χ. της εξομάλυνσης τα αντίστοιχα μη γραμμικά μοντέλα υστερούν έναντι της κλασικής μεθόδου. Η εξομάλυνση συντελεί δηλαδή καθοριστικά στη βελτίωση της προβλεπτικής ακρίβειας των μεθόδων πρόβλεψης καθιστώντας αποδοτικές εν τέλει ακόμα και μεθόδους που θεωρούνταν αρχικά ανακριβείς.

Εξετάζοντας την επίδοση των επιμέρους μεθόδων ως προς την ένταση της εποχιακότητας, μπορεί κανείς να παρατηρήσει πως, καθώς αυξάνεται το επίπεδο της αυξάνεται και η απόδοση της προτεινόμενης προσέγγισης. Το κέρδος κυμαίνεται ανάλογα το μοντέλο από 3.5% έως 4.9% για την περίπτωση της μεσαίας έντασης με μέση τιμή το 4.3%. Αυτό πέφτει στο 1.8% για τη μεγάλη ένταση και στο 0.8% για τη χαμηλή, όπως και αναμένεται για μη εποχιακές χρονοσειρές και δεδομένα για τα οποία χαρακτηρίζονται από σημαντικές διακυμάνσεις. Η επιλεκτική προσέγγιση εμφανίζει μέσο κέρδος σε όλες τις περιπτώσεις περί του 1.4%.

Πίνακας 6.18: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά ένταση εποχιακότητας.

Εποχιακότητα	Linear	Exponential	Logarithmic	Inverse	Power	Selective
Κλασική προσέγγιση						
Χαμηλό	11.98	11.46	12.17	12.56	11.96	11.44
Μεσαίο	16.38	16.38	16.56	16.71	16.61	16.37
Υψηλό	24.15	23.24	23.64	23.99	23.64	23.40
Προτεινόμενη προσέγγιση						
Χαμηλό	11.71	11.29	12.15	12.51	11.99	11.37
Μεσαίο	15.79	15.77	15.82	15.91	15.80	15.80
Υψηλό	23.92	23.16	23.10	23.24	23.12	23.57

Συνοψίζοντας, η προτεινόμενη μεθοδολογία φαίνεται να βελτιώνει στη γενική περίπτωση την ακρίβεια της πρόβλεψης, ειδικά όταν πρόκειται για χρονοσειρές με έντονη εποχιακότητα. Οι βελτιώσεις αυτές εξαρτώνται σημαντικά από το μοντέλο τάσης που επιλέγεται και αυξάνονται για εποχιακότητα μεσαίας έντασης. Αντίθετα, όταν αυτή αυξάνεται υπερβολικά, οπότε και οι διακυμάνσεις αλλοιώνουν σημαντικά τα μοτίβα της χρονοσειράς, ή περιορίζεται ιδιαίτερα, οπότε και η συνιστώσα της εποχιακότητας είναι πλέον αμελητέα, τότε τα οφέλη της μεθόδου συρρικνώνονται.

Όσον αφορά το επίπεδο της τυχαιότητας, όπως παρατηρείται η ακρίβεια πρόβλεψης βελτιώνεται στη γενική περίπτωση μέσω της προτεινόμενης μεθοδολογίας για όλα τα μοντέλα τάσης, ακόμα και για χρονοσειρές με χαμηλά επίπεδα θορύβου. Όπως είναι αναμενόμενο η βελτίωση αυξάνει καθώς αυξάνει το επίπεδο του θορύβου με τα ποσοστά να φτάνουν το 0.4%, 1.6% και 4.6% για την περίπτωση χαμηλού, μεσαίου και υψηλού επιπέδου θορύβου.

Στην περίπτωση του επιλεκτικού μοντέλου, τα κέρδη είναι σημαντικά σε κάθε περίπτωση και συγκεκριμένα ίσα με 1.1%, 0.4% και 0.9% για την περίπτωση χαμηλού, μεσαίου και υψηλού επιπέδου θορύβου. Συμπεραίνουμε λοιπόν πως ο διαγνωστικός έλεγχος αποδίδει ανεξαρτήτως θορύβου και αποφεύγει κακοτοπιές οδηγώντας σε σταθερά βελτιωμένες προβλέψεις. Ωστόσο, όπως φαίνεται από τη φθίνουσα πορεία του κέρδους, όσο περισσότερος θόρυβος εμπεριέχεται σε μία χρονοσειρά, τόσο δυσκολότερο είναι να εντοπίσει κανείς το μοντέλο που την περιγράφει αποδοτικότερα στο μέλλον.

Συμπερασματικά, το μοντέλο τάσης φαίνεται να επηρεάζει καθοριστικά την ακρίβεια που επιτυγχάνεται μέσω της προτεινόμενης μεθοδολογίας, η οποία ωστόσο βελτιώνει σταθερά την μέση απόδοση των μοντέλων σε κάθε περίπτωση. Όταν ο θόρυβος είναι λίγος τα οφέλη από την εξομάλυνση είναι περιορισμένα, ενώ όταν αυξάνεται σημαντικότερα, με τον αλγόριθμο επιλογής να δυσκολεύεται ωστόσο να εντοπίσει το βέλτιστο μοντέλο πρόβλεψης σε περιπτώσεις αυξημένης τυχαιότητας.

Πίνακας 6.19: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά ένταση τυχαιότητας.

Τυχαιότητα	Linear	Exponential	Logarithmic	Inverse	Power	Selective
Κλασική προσέγγιση						
Χαμηλό	9.91	9.56	10.37	10.77	10.21	9.66
Μεσαίο	24.65	23.17	22.78	22.80	22.55	23.14
Υψηλό	36.95	36.59	36.70	37.21	36.81	35.64
Προτεινόμενη προσέγγιση						
Χαμηλό	9.79	9.50	10.37	10.72	10.24	9.55
Μεσαίο	23.49	22.81	22.53	22.67	22.48	23.24
Υψηλό	36.08	35.06	34.86	34.98	34.88	35.33

Η αξιολόγηση της απόδοσης της μεθόδου ανά είδος και ένταση τάσης οδηγεί και αυτή με τη σειρά της σε παρεμφερή συμπεράσματα. Η επιλεκτική μέθοδος είναι η δεύτερη πιο αποδοτική επιλογή, οδηγώντας ανά ένταση γραμμικής τάσης σε μέσα κέρδη της τάξεως του 1.4%. Ωστόσο, σε αυτήν την περίπτωση τα μέσα κέρδη της προτεινόμενης προσέγγισης περιορίζονται καθώς αυξάνει πολύ η ένταση της τάσης, προφανώς λόγω της ανικανότητας των μεμονωμένων μοντέλων να ακολουθήσουν πιστά το σήμα οποιασδήποτε χρονοσειράς σε ακραίες καταστάσεις. Για την περίπτωση της εκθετικής τάσης τα κέρδη διαμορφώνονται στα επιμέρους μοντέλα αντίστοιχα ανά επίπεδο έντασης από 1.4% έως 1.9%, για τη λογαριθμική από 1.1% έως 3.2%, για την αντίστροφη από 1.2% έως 2.1% και για τη τάση δύναμης από 1.4% έως 2.0%.

Κεφάλαιο 6: Πειραματική διάταξη - Ανάλυση και εξέταση των αποτελεσμάτων

Πίνακας 6.20: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά ένταση γραμμικής τάσης.

Γραμμική τάση	Linear	Exponential	Logarithmic	Inverse	Power	Selective
<i>Κλασική προσέγγιση</i>						
Χαμηλό	10.03	9.98	10.32	10.46	10.31	10.03
Μεσαίο	15.44	15.20	16.00	16.44	15.97	15.30
Υψηλό	26.11	24.00	25.34	26.37	24.67	23.78
<i>Προτεινόμενη προσέγγιση</i>						
Χαμηλό	9.90	9.85	10.17	10.31	10.15	9.91
Μεσαίο	14.95	14.65	15.54	15.93	15.46	14.69
Υψηλό	25.32	23.68	25.25	26.01	24.79	24.03

Πίνακας 6.21: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά ένταση εκθετικής τάσης.

Εκθετική τάση	Linear	Exponential	Logarithmic	Inverse	Power	Selective
<i>Κλασική προσέγγιση</i>						
Χαμηλό	10.26	10.24	10.45	10.55	10.45	10.26
Μεσαίο	13.04	13.00	13.38	13.58	13.39	12.94
Υψηλό	18.65	17.38	18.55	19.28	18.15	17.40
<i>Προτεινόμενη προσέγγιση</i>						
Χαμηλό	10.10	10.07	10.24	10.32	10.23	10.16
Μεσαίο	12.89	12.79	13.22	13.42	13.19	12.80
Υψηλό	18.09	17.10	18.40	19.00	18.12	17.31

Πίνακας 6.22: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά ένταση λογαριθμικής τάσης.

Λογαριθμική τάση	Linear	Exponential	Logarithmic	Inverse	Power	Selective
<i>Κλασική προσέγγιση</i>						
Χαμηλό	10.88	10.76	11.31	11.52	11.27	10.87
Μεσαίο	20.22	20.04	21.10	22.02	20.76	19.61
Υψηλό	33.93	30.02	30.66	31.62	29.74	29.53
<i>Προτεινόμενη προσέγγιση</i>						
Χαμηλό	10.76	10.63	11.18	11.40	11.14	10.69
Μεσαίο	19.70	19.32	20.50	21.12	20.17	19.22
Υψηλό	32.23	29.54	30.46	31.13	29.92	30.36

Πίνακα 6.23: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά ένταση αντίστροφης τάσης.

Αντίστροφη τάση	Linear	Exponential	Logarithmic	Inverse	Power	Selective
Κλασική προσέγγιση						
Χαμηλό	10.98	10.85	11.44	11.70	11.38	10.92
Μεσαίο	21.81	21.98	22.45	23.83	22.28	21.76
Υψηλό	36.66	31.88	32.74	33.12	31.74	31.41
Προτεινόμενη προσέγγιση						
Χαμηλό	10.86	10.71	11.30	11.54	11.24	10.76
Μεσαίο	21.46	21.24	22.22	22.89	21.95	21.30
Υψηλό	34.57	31.34	32.19	32.73	31.67	32.25

Πίνακα 6.24: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά ένταση τάσης σε μορφή δύναμης.

Τάση δύναμης	Linear	Exponential	Logarithmic	Inverse	Power	Selective
Κλασική προσέγγιση						
Χαμηλό	10.61	10.63	10.86	10.95	10.86	10.63
Μεσαίο	12.71	12.55	13.28	13.63	13.25	12.68
Υψηλό	22.63	20.67	21.92	22.85	21.31	20.51
Προτεινόμενη προσέγγιση						
Χαμηλό	10.45	10.43	10.62	10.71	10.61	10.48
Μεσαίο	12.59	12.40	13.17	13.50	13.12	12.47
Υψηλό	21.81	20.32	21.79	22.50	21.37	20.63

Αξίζει επίσης να σημειωθεί ότι, ενώ η απόδοση όλων των μοντέλων μειώνεται με την αύξηση της έντασης τάσης ανεξαρτήτως μοτίβου, και ενώ η εκθετική προσέγγιση είναι η καλύτερη για το σύνολο των χρονοσειρών, μεμονωμένα υπάρχουν μοντέλα τα οποία υπερτερούν έναντι άλλων για συγκεκριμένα είδη τάσης. Αυτό συμβαίνει καθώς ανάλογα με το μοτίβο μακροχρόνιας τάσης που επικρατεί, διαφορετικές γραμμές μηδέν πρέπει να χρησιμοποιούνται για την αποτελεσματική προέκτασή του. Αυτός είναι άλλωστε και ο βασικός λόγος για τον οποίο αποφασίστηκε εξαρχής στην παρούσα διατριβή να επεκταθεί η κλασική γραμμική μέθοδος θ για μη γραμμικά μοτίβα τάσης.

Για την καλύτερη κατανόηση των αποτελεσμάτων, στη συνέχεια παρουσιάζεται αναλυτικά η επίδοση του κάθε μοντέλου πρόβλεψης θ ανά συχνότητα δεδομένων και διαγωνισμό (Πίνακας 6.25) αλλά και ανά συχνότητα συνολικά (Πίνακας 6.25).

Κεφάλαιο 6: Πειραματική διάταξη - Ανάλυση και εξέταση των αποτελεσμάτων

Πίνακας 6.25: Επίδοση προτεινόμενης μεθοδολογίας σε σχέση με άλλες απλούστερες προσεγγίσεις στα δεδομένα των διαγωνισμών *M* ανά διαγωνισμό και συχνότητα.

Δεδομένα	Linear (A)	Exponential (A)	Logarithmic (A)	Inverse (A)	Power (A)	Linear (M)	Exponential (M)	Logarithmic (M)	Inverse (M)	Power (M)	Selective
Κλασική προσέγγιση											
Ετήσια M3	16.73	16.44	17.20	17.73	16.86	16.88	15.93	17.24	18.15	16.64	15.88
Τριμηνιαία M3	9.29	9.28	9.60	9.81	9.56	9.21	9.12	9.56	9.92	9.44	9.23
Μηνιαία M3	13.87	13.68	14.00	14.25	14.05	14.20	13.96	14.29	14.60	14.38	13.88
Άλλα M3	4.92	5.31	5.87	6.27	5.96	4.98	5.36	5.96	6.27	6.04	5.09
Ετήσια M1	20.19	20.40	22.47	23.23	21.73	20.11	18.56	22.21	22.90	21.12	19.32
Τριμηνιαία M1	16.30	16.04	17.00	17.53	16.90	16.38	16.12	17.08	18.22	16.87	16.37
Μηνιαία M1	16.52	14.73	15.14	14.95	14.76	16.54	14.73	14.91	15.04	14.82	14.93
Σύνολο	13.89	13.52	14.04	14.31	13.90	14.03	13.43	14.11	14.55	13.95	13.47
Προτεινόμενη προσέγγιση											
Ετήσια M3	16.61	16.08	17.49	18.05	17.17	16.67	16.00	17.56	18.19	17.13	15.99
Τριμηνιαία M3	9.06	8.95	9.42	9.65	9.37	9.11	8.94	9.40	9.77	9.35	9.03
Μηνιαία M3	13.56	13.41	13.64	13.79	13.63	13.69	13.39	13.64	13.80	13.63	13.50
Άλλα M3	5.11	5.52	6.11	6.51	6.20	5.13	5.56	6.20	6.49	6.27	5.35
Ετήσια M1	20.35	19.10	22.40	23.31	21.84	20.44	19.69	22.68	23.47	21.72	18.65
Τριμηνιαία M1	16.12	15.87	16.70	16.96	16.61	16.14	15.87	16.76	16.99	16.56	16.02
Μηνιαία M1	15.84	14.76	14.88	15.05	14.87	15.81	14.82	14.93	15.07	14.91	15.46
Σύνολο	13.62	13.24	13.88	14.16	13.79	13.69	13.26	13.91	14.23	13.78	13.37

Πίνακας 6.26: Επίδοση προτεινόμενης μεθοδολογίας σε σχέση με άλλες απλούστερες προσεγγίσεις στα δεδομένα των διαγωνισμών M ανά συχνότητα.

Δεδομένα	Linear (A)	Exponential (A)	Logarithmic (A)	Inverse (A)	Power (A)	Linear (M)	Exponential (M)	Logarithmic (M)	Inverse (M)	Power (M)	Selective
Κλασική προσέγγιση											
Ετήσια	15.30	15.22	16.18	16.73	15.84	15.39	14.57	16.18	16.94	15.61	14.63
Τριμηνιαία	10.77	10.71	11.17	11.44	11.11	10.73	10.60	11.15	11.68	11.01	10.74
Μηνιαία	14.67	14.00	14.34	14.46	14.26	14.91	14.19	14.48	14.73	14.51	14.20
Σύνολο	13.89	13.52	14.04	14.31	13.90	14.03	13.43	14.11	14.55	13.95	13.47
Προτεινόμενη προσέγγιση											
Ετήσια	15.29	14.79	16.40	16.99	16.11	15.34	14.85	16.51	17.11	16.07	14.62
Τριμηνιαία	10.55	10.41	10.96	11.20	10.90	10.60	10.41	10.96	11.30	10.88	10.51
Μηνιαία	14.25	13.82	14.01	14.17	14.00	14.33	13.82	14.03	14.18	14.02	14.09
Σύνολο	13.62	13.24	13.88	14.16	13.79	13.69	13.26	13.91	14.23	13.78	13.37

Πίνακας 6.27: Επίδοση προτεινόμενης προσέγγισης στα δεδομένα των διαγωνισμών M ανά είδος εφαρμογής.

Δεδομένα	Linear (A)	Exponential (A)	Logarithmic (A)	Inverse (A)	Power (A)	Linear (M)	Exponential (M)	Logarithmic (M)	Inverse (M)	Power (M)	Selective
Κλασική προσέγγιση											
DEMOGRAPHIC	14.66	11.99	12.77	12.70	12.28	14.59	11.84	12.54	13.26	12.21	12.11
FINANCE	15.45	15.72	15.45	15.58	15.38	15.34	15.38	15.33	15.45	15.24	15.43
INDUSTRIAL	13.38	13.39	13.59	13.75	13.48	13.42	13.12	13.54	13.87	13.35	13.40
MACRO	8.34	7.88	9.07	9.42	8.95	8.35	7.97	9.06	9.51	8.95	8.26
MICRO	19.79	20.08	20.26	20.66	20.19	20.26	20.01	20.64	21.13	20.48	19.64
OTHER	7.24	7.48	8.17	8.74	8.35	7.43	7.52	8.36	8.81	8.49	7.15
Προτεινόμενη προσέγγιση											
DEMOGRAPHIC	13.16	11.57	12.46	12.86	12.39	13.10	11.77	12.53	12.90	12.44	12.43
FINANCE	15.19	15.09	15.30	15.44	15.24	15.20	14.96	15.29	15.48	15.22	15.12
INDUSTRIAL	13.28	13.17	13.43	13.60	13.33	13.40	13.09	13.43	13.69	13.32	13.24
MACRO	8.50	7.99	9.15	9.48	9.03	8.53	8.05	9.14	9.50	8.99	8.12
MICRO	19.56	19.51	19.87	20.10	19.74	19.68	19.50	19.94	20.24	19.72	19.38
OTHER	7.34	7.63	8.44	9.00	8.58	7.40	7.66	8.56	8.99	8.66	7.45

Όπως παρατηρείται, το εκθετικό αθροιστικό μοντέλο θ φαίνεται να είναι η πιο ακριβής προσέγγιση, μειώνοντας το μέσο σφάλμα πρόβλεψης για το σύνολο των δεδομένων στην κλασική προσέγγιση κατά 2.7% και στην προτεινόμενη κατά 4.7%. Η μέθοδος είναι σημαντικά καλύτερη από την κλασική προσέγγιση σε όλα τα σετ συχνοτήτων πέραν των ετησίων του διαγωνισμού $M1$ και των άλλων του διαγωνισμού $M3$, με το ποσοστό βελτίωσης να φτάνει έως και το 10.8%. Αντίθετα τα υπόλοιπα μοτίβα τάσης πετυχαίνουν καλύτερη ακρίβεια μόνο στα μηνιαία δεδομένα των διαγωνισμών και οδηγούν σε χειρότερες προβλέψεις στα υπόλοιπα σετ, με αρκετές βέβαια εξαιρέσεις μετά και την εφαρμογή την προτεινόμενης προσέγγισης.

Ιδιαίτερα καλές επιδόσεις παρουσιάζει και η πολλαπλασιαστική έκφραση του εκθετικού μοντέλου θ , γεγονός που σημαίνει πως το μοτίβο τάσης που επιλέγεται έχει μεγαλύτερη επίδραση προβλεπτικά από την έκφραση του μοντέλου, χωρίς βέβαια αυτό να σημαίνει πως η τελευταία δεν κάνει καμία διαφορά. Αυτό επιβεβαιώνεται από το γεγονός ότι για το ίδιο μοτίβο τάσης τα σφάλματα των δύο εκφράσεων είναι παρόμοια, όχι όμως και απαραίτητα. Πιο συγκεκριμένα, το σφάλμα αυξάνει για την πολλαπλασιαστική έκφραση στην περίπτωση της γραμμικής τάσης και της τάσης δύναμης και μειώνεται για την εκθετική, τη λογαριθμική και την αντίστροφη. Τα σφάλματα βέβαια γίνονται σχεδόν αμελητέα μετά την εφαρμογή της εξομάλυνσης καθώς αυτή περιορίζει τις διακυμάνσεις της χρονοσειράς και συνεπώς το πολλαπλασιαστικό μοντέλο παύει να παράγει σημαντικά διαφορετικές προβλέψεις από το προσθετικό.

Πολύ σημαντικό είναι επίσης το φαινόμενο πως, μετά την εφαρμογή της προτεινόμενης λύσης, και ανεξαρτήτως της συχνότητας και του σετ δεδομένων που εξετάζεται, η πλειοψηφία των μοντέλων καταφέρνει να ξεπεράσει τις επιδόσεις της κλασικής μεθόδου θ , συχνά μάλιστα με μεγάλες διαφορές. Έτσι, ενώ στην αρχική προσέγγιση το εκθετικό μοντέλο είναι το μόνο που υπερτερεί του γραμμικού, μετά την προτεινόμενη προσέγγιση, όλα πέραν των αντιστρόφων οδηγούν σε κέρδη.

Όσον αφορά τώρα το επιλεκτικό μοντέλο πρόβλεψης, όπως σημειώθηκε και νωρίτερα αυτό φαίνεται να δυσκολεύεται να επιλέξει σωστά για την περίπτωση των χρονοσειρών υψηλών συχνοτήτων και κυρίως για τα μηνιαία δεδομένα του $M3$ και τα τριμηνιαία του $M1$. Αυτό οφείλεται προφανώς στη τυχαιότητα που χαρακτηρίζει τα εν λόγω δεδομένα καθώς και σε φαινόμενα απότομη μεταβολής του μοτίβου τάσης που παρατηρείται στα ιστορικά δεδομένα στο μέλλον. Παρόλα αυτά το επιλεκτικό μοντέλο αποδίδει ιδιαίτερα ικανοποιητικά στα ετήσια δεδομένα, όπου και πρωτεύει, και στο σύνολο του διαγωνισμού όπου αποτελεί τη δεύτερη πιο συμφέρουσα λύση. Έτσι, δεδομένου ότι με κάποιον τρόπο πρέπει εξ ορισμού να αποφασίσει κανείς πιο μοντέλο θα επιλέξει για την προέκταση μιας χρονοσειράς, καθώς και του ότι σε θορυβώδεις χρονοσειρές η όποια επιλεκτική μέθοδος αναμένεται να οδηγήσει σε μη βέλτιστα αποτελέσματα, η προτεινόμενη προσέγγιση φαντάζει ιδιαίτερα αποδοτική. Συμβουλευμένοι μάλιστα τα αποτελέσματα της γεννήτριας, όπου και η εν λόγω μέθοδος άριστευσε, η υιοθέτηση της επιλεκτικής μεθόδου φαντάζει μία αξιόπιστη λύση.

Συναθροίζοντας τώρα τα αποτελέσματα ανά συχνότητα δεδομένων, συμπεραίνουμε ότι στη γενική περίπτωση η προτεινόμενη μεθοδολογία βελτιώνει την ακρίβεια του εκάστοτε μοντέλου τάσης κατά 2.0% για τριμηνιαίες χρονοσειρές και κατά 2.7% για μηνιαίες χρονοσειρές, ενώ τη χειροτερεύει κατά 0.9% για ετήσιες χρονοσειρές ή για δεδομένα χωρίς συχνότητα. Αυτό βέβαια δεν ισχύει για τα μοντέλα πρόβλεψης γραμμικής και εκθετικής τάσης όπου η βελτίωση στην πρώτη περίπτωση φτάνει το 0.1% για ετήσια δεδομένα, το 2% για τριμηνιαία και το 2.9% για μηνιαία, ενώ στη δεύτερη περίπτωση το 2.8% για ετήσια και τριμηνιαία δεδομένα και το 1.3% για μηνιαία. Επίσης, στην περίπτωση των εποχιακών δεδομένων όλα τα μοντέλα οδηγούν σε βελτιώσεις συγκριτικά με την κλασική του έκδοση, ανεξαρτήτως έκφρασής τους. Μάλιστα, το επιλεκτικό μοντέλο είναι το μόνο που μαζί με το γραμμικό οδηγεί σε βελτιώσεις ανεξαρτήτως συχνότητας με τις ετήσιες να αποτελούν το δυνατότερο σημείο της προσέγγισης. Το τελευταίο δικαιολογείται πιθανότατα από τη φύση των μη εποχιακών δεδομένων τα οποία χαρακτηρίζονται από χαμηλότερα επίπεδα τυχαιότητας και ως εκ τούτου είναι πολύ πιο εύκολο να εντοπίσει κανείς σε αυτά το μοτίβο μακροπρόθεσμης τάσης τους.

Τα αποτελέσματα μας οδηγούν επίσης στο συμπέρασμα ότι η τεχνική της εξομάλυνσης αποδίδει καλύτερα όταν η συχνότητα είναι τέτοια που να επιτρέπει τη διείσδυση θορύβου στα δεδομένα, οπότε και η απαλοιφή του έχει πράγματι νόημα, ενώ το μοντέλο τάσης που επιλέγεται είναι καθοριστικό για το επίπεδο ακρίβειας που θα επιτευχθεί. Σε όλες τις περιπτώσεις πάντως, η προτεινόμενη μεθοδολογία βελτιώνει την ακρίβεια πρόβλεψης κατά μήκος όλων των συχνοτήτων όταν χρησιμοποιείται το γραμμικό ή το εκθετικό μοντέλο τάσης, με το δεύτερο να είναι η αποδοτικότερη λύση.

Εξετάζοντας τέλος τα αποτελέσματα ανά τύπο δεδομένων ξεχωριστά συμπεραίνουμε ότι η προτεινόμενη μεθοδολογία οδηγεί σε βελτιωμένες προβλέψεις σε όλες τις περιπτώσεις εκτός των δημογραφικών και των άλλων δεδομένων. Το μέσο όφελος για τα μικροοικονομικά δεδομένα είναι 0.76%, για τα βιομηχανικά 0.9%, για τα οικονομικά 1.3% και για τα δημογραφικά 2.1%. Τα εν λόγω ποσοστά μεταβάλλονται ωστόσο σημαντικά ανάλογα με το μοντέλο τάσης που χρησιμοποιείται. Έτσι, η ακρίβεια πρόβλεψης βελτιώνεται π.χ. στη περίπτωση των δημογραφικών δεδομένων κατά 10.2% για χρήση γραμμικού μοτίβου τάσης, ποσοστό το οποίο πέφτει στο 3.5% για εκθετικό μοτίβο και στο -1.3% για αντίστροφη τάση.

Συνολικά, το εκθετικό μοτίβο τάσης φαίνεται να είναι και πάλι η αποδοτικότερη επιλογή, αν και η εξομάλυνση θα πρέπει να μην εφαρμόζεται για την περίπτωση των μακροοικονομικών δεδομένων. Επιπλέον, η κλασική μέθοδος θ φαίνεται να αποδίδει καλύτερα στα άλλα δεδομένα. Γενικά αποδεικνύεται πως για συγκεκριμένους τύπους δεδομένων κάποια μοντέλα να αποδίδουν πάντως ελαφρώς καλύτερα από άλλα, όπως π.χ. η τάση δύναμης με πολλαπλασιαστική έκφραση για τα οικονομικά δεδομένα, πράγμα ιδιαίτερα ενδιαφέρον από τη στιγμή που το εν λόγω μοντέλο δεν υπερέχει έναντι των υπολοίπων σε άλλες κατηγορίες του διαγωνισμού. Ίσως λοιπόν να υπάρχουν κατηγορίες δεδομένων οι οποίες να οδηγούν από μόνες τους στην υιοθέτηση συγκεκριμένων μοντέλων πρόβλεψης. Οι διαφορές πάντως σε κάθε περίπτωση δεν είναι δραματικές και

Κεφάλαιο 6: Πειραματική διάταξη - Ανάλυση και εξέταση των αποτελεσμάτων

συνεπώς οι επιλογές μπορούν να γενικευτούν στο μοντέλο εκθετικής τάσης ή το επιλεκτικό.

Συνοπτικά λοιπόν το μοντέλο τάσης επηρεάζει σημαντικά την ακρίβεια πρόβλεψης ενώ η προτεινόμενη μεθοδολογία βελτιώνει στις περισσότερες περιπτώσεις τύπου δεδομένων την απόδοσή τους. Σε κάθε περίπτωση πάντως η προτεινόμενη μεθοδολογία με εκθετικό μοντέλο τάσης εξασφαλίζει καλύτερη ακρίβεια από την κλασική προσέγγιση σχεδόν για κάθε τύπο δεδομένων, με τον επιλεκτικό αλγόριθμο να ακολουθεί.

Συνοπτικά η μελέτη των αποτελεσμάτων στο συγκεκριμένο δείγμα δεδομένων μας οδηγεί στα παρακάτω βασικά συμπεράσματα:

- ✓ Η συρρίκνωση των εποχιακών δεικτών βελτιώνει συστηματικά τις προβλέψεις των εποχιακών χρονοσειρών.
- ✓ Η προτεινόμενη μεθοδολογία βελτιώνει την ακρίβεια της πρόβλεψης καθώς αυξάνεται η ένταση εποχιακότητας, όχι όμως απεριόριστα. Για μεγάλες εποχιακές διακυμάνσεις είναι πιθανό τα αναμενόμενα κέρδη να μειωθούν.
- ✓ Στη γενική περίπτωση η εξομάλυνση των χρονοσειρών οδηγεί σε βελτιωμένη προβλεπτική ακρίβεια.
- ✓ Η προτεινόμενη μεθοδολογία βελτιώνει την ακρίβεια της πρόβλεψης καθώς αυξάνεται η ένταση τυχαιότητας.
- ✓ Η προτεινόμενη μεθοδολογία βελτιώνει την ακρίβεια της πρόβλεψης καθώς αυξάνεται η ένταση τάσης, όχι όμως απεριόριστα. Για μεγάλες τάσεις είναι πιθανό τα αναμενόμενα κέρδη να μειωθούν.
- ✓ Τα οφέλη συρρίκνωσης εποχιακών δεικτών και εξομάλυνσης είναι ανεξάρτητα μεταξύ τους.
- ✓ Η επιλογή ενός αντιπροσωπευτικού μοτίβου τάσης, όπως π.χ. το εκθετικό, μπορεί να βελτιώσει σημαντικά την ακρίβεια των προβλέψεων ανεξαρτήτως συχνότητας χρονοσειράς και είδους εφαρμογής.
- ✓ Τα οφέλη από τον εντοπισμό του καταλληλότερου μοτίβου τάσης είναι σημαντικότερα από τον εντοπισμό της καταλληλότερης έκφρασης της μεθόδου.
- ✓ Η επιλογή του καταλληλότερου μοντέλου θ γίνεται ευκολότερα για ομαλές χρονοσειρές και ειδικότερα για χρονοσειρές περιορισμένης τυχαιότητας και εποχιακότητας. Ωστόσο γίνεται με μεγαλύτερη επιτυχία για χρονοσειρές έντονης τάσης, όπου και το μοτίβο γίνεται αρκετά πιο εμφανές.
- ✓ Σε αντίθεση με τα τεχνητά δεδομένα, όπου το μοτίβο τάσης διατηρείται στο μέλλον, στα πραγματικά δεδομένα η επιλογή του βέλτιστου μοντέλου θ είναι αρκετά πιο δύσκολη καθώς εκεί ενδέχεται να αλλάξει ο τρόπος εξέλιξης του μεγέθους.
- ✓ Με την εφαρμογή της προτεινόμενης μεθοδολογίας το επιλεκτικό μοντέλο τάσης μπορεί στη γενική περίπτωση να εφαρμόζεται σε κάθε χρονοσειρά με σχετική ασφάλεια και πιθανότητες για σημαντικά οφέλη, ανεξαρτήτως της έντασης των χαρακτηριστικών της και του τύπου της.
- ✓ Για συγκεκριμένες κατηγορίες δεδομένων και συχνότητες ενδέχεται κάποια μοντέλα θ να αποδίδουν καλύτερα έναντι άλλων καθώς και να μεταβάλλεται η επίδοση του αλγορίθμου επιλογής. Για παράδειγμα, το εκθετικό μοντέλο είναι

ακριβέστερο για την περίπτωση των δημογραφικών δεδομένων και τα μοντέλο δύναμης για των οικονομικών. Επίσης για τα ετήσια δεδομένα, που χαρακτηρίζονται από ομαλότητα, το επιλεκτικό μοντέλο αποδίδει σαφώς καλύτερα από τα μηνιαία, όπου και η τυχαιότητα αυξάνει σημαντικά.

Όταν βλέπω μία πρόβλεψη δύο δεκαδικών ψηφίων αναρωτιέμαι αν οφείλεται στην αδυναμία κατανόησης των περιορισμών στα δεδομένα ή στη ψευδαίσθηση της ακρίβειας

Barry Ritholtz

Κεφάλαιο 7: Συμπεράσματα και προεκτάσεις

7.1 Ανασκόπηση

Στα πλαίσια της παρούσας διατριβής αναπτύχθηκε μία μεθοδολογία με στόχο την αποδοτική παραγωγή προβλέψεων ανεξαρτήτως πλήθους και τύπου δεδομένων. Η μελέτη μπορεί να αξιοποιηθεί για την αυτοματοποιημένη πρόβλεψη χρονοσειρών σε επιχειρήσεις και οργανισμούς που, αφενός αναζητούν υψηλή ακρίβεια στις προβλέψεις τους, και αφετέρου επιθυμούν τον περιορισμό του κόστους της όλης διαδικασίας πρόβλεψης, συμπεριλαμβανομένου του χρόνου εκτέλεσης και της ενασχόλησης του ανθρώπινου δυναμικού.

Προκειμένου να επιτευχθεί ο εν λόγω στόχος εξετάστηκε αρχικά η αξιοποίηση τεχνικών προβλέψεων της βιβλιογραφίας οι οποίες έχουν κατά καιρούς αποδείξει πως συμβάλουν στην εν λόγω κατεύθυνση. Ο εντοπισμός, η επέκταση και η οργάνωση τέτοιων τεχνικών εντός ενός ενιαίου μεθοδολογικού πλαισίου είναι καθοριστικής σημασίας για την κατασκευή της εν λόγω μεθοδολογίας καθώς, αν και συχνά προτείνονται αποδοτικά μοντέλα πρόβλεψης και μέθοδοι επεξεργασίας δεδομένων για τη βελτίωση της προβλεπτικής τους ικανότητας, σπάνια αυτές συνδυάζονται για την κατασκευή μίας ολοκληρωμένης λύσης. Αυτό αποδεικνύεται πρώτον από το γεγονός πως η βιβλιογραφία είναι αρκετά κατακερματισμένη σε ότι αφορά το συνδυασμό τεχνικών προβλέψεων και δεύτερον από το φαινόμενο πως πολλές νέες τεχνικές που προτείνονται αδυνατούν να ξεπεράσουν άλλες παλαιότερες και συχνά απλούστερες, χαρακτηριστικό παράδειγμα των οποίων αποτελούν τα μοντέλα εκθετικής εξομάλυνσης και η μέθοδος θ .

Η μεθοδολογία που προτείνεται κατασκευάστηκε λοιπόν βάσει της λογικής που σκιαγραφήθηκε νωρίτερα με σκοπό να ξεπεράσει προβλεπτικά οποιαδήποτε μεμονωμένη μέθοδο πρόβλεψης, τόσο σε ταχύτητα όσο και σε ακρίβεια. Να είναι δηλαδή όπως αναφέρει χαρακτηριστικά ο Steve Jobs “thinner lighter and faster”. Η επίτευξη του παραπάνω σκοπού προσεγγίστηκε μέσω των ακόλουθων βημάτων:

- Την αναζήτηση βέλτιστων τεχνικών προβλέψεων στη βιβλιογραφία
- Την υιοθέτηση εκείνων των μεθόδων προετοιμασίας χρονοσειρών που αποδεδειγμένα βελτιώνουν την προβλεπτική ικανότητα των στατιστικών μοντέλων πρόβλεψης
- Την ανάπτυξη νέων μεθόδων επεξεργασίας δεδομένων για την περαιτέρω βελτίωση της απόδοσης των μοντέλων πρόβλεψης
- Τον εντοπισμό του αποδοτικότερου βιβλιογραφικά μοντέλου πρόβλεψης
- Την επέκταση και γενίκευση του εν λόγω μοντέλου για την καθολική του εφαρμογή σε πλήθος χρονοσειρών ανεξαρτήτως τύπου.
- Την οργάνωση των παραπάνω τεχνικών εντός ενός ενιαίου μεθοδολογικού πλαισίου για την αυτοματοποίηση της διαδικασίας πρόβλεψης.

Βάσει των παραπάνω η προτεινόμενη μεθοδολογία αποτελείται ουσιαστικά από τέσσερα βήματα:

- i. Την αποεποχικοποίηση των δεδομένων για την εξάλειψη των εποχιακών διακυμάνσεων
- ii. Την εξομάλυνση της χρονοσειράς για τον περιορισμό της τυχαιότητας
- iii. Την προέκταση της χρονοσειράς μέσω ενός κατάλληλου μοντέλου πρόβλεψης
- iv. Την επαναεποχικοποίηση των προβλέψεων για την ενσωμάτωση του εποχιακού παράγοντα σε αυτές

Πιο συγκεκριμένα, για την αποεποχικοποίηση των δεδομένων επιλέχθηκε να εφαρμοστεί η κλασική πολλαπλασιαστική μέθοδος αποσύνθεσης κινητών μέσων όρων. Ωστόσο, προκειμένου να περιοριστεί η αβεβαιότητα κατά την εκτίμηση της εποχιακής συνιστώσας χρησιμοποιήθηκε επιπλέον η μέθοδος συρρίκνωσης συντελεστών *Miller-Williams*. Έτσι, αφού εφαρμοστεί ένας έλεγχος για την εξακρίβωση της εποχιακής συμπεριφοράς της χρονοσειράς, υπολογίζονται οι εποχιακοί δείκτες και εν συνεχεία αυτοί συρρικνώνονται ούτως ώστε να γίνουν όσο το δυνατό πιο αντιπροσωπευτικοί.

Η εξομάλυνση της χρονοσειράς πραγματοποιείται μέσω μίας νέας μεθόδου που αναπτύχθηκε για αυτό το σκοπό στα πλαίσια της διατριβής, και η οποία αναφέρεται ως «μετασχηματισμός εξισορρόπησης τοπικών διακυμάνσεων». Η εν λόγω μέθοδος εμπνεύστηκε από τον μετασχηματισμό θ για την ανάδειξη της μακροπρόθεσμης τάσης των χρονοσειρών και περιλαμβάνει μηχανισμούς ούτως ώστε να διατηρεί επιπλέον τις τοπικές καμπυλότητες των δεδομένων. Έτσι, η εξομαλυμένη χρονοσειρά είναι απαλλαγμένη από τυχαίες διακυμάνσεις, επιτρέποντας στο μοντέλο πρόβλεψης αφενός να εντοπίσει αποτελεσματικότερα το επίπεδό της και αφετέρου να προσαρμοστεί καλύτερα στο μοτίβο ανάπτυξής της, είτε αυτό είναι γραμμικό είτε όχι.

Ως μοντέλο πρόβλεψης επιλέχθηκε η μέθοδος θ κυρίως λόγω της απλότητάς της και της ικανότητάς της να οδηγεί σε ακριβείς προβλέψεις, ακόμα και συγκριτικά με άλλες αρκετά πιο πολύπλοκες μεθόδους. Η μέθοδος θ , η οποία αναδείχτηκε νικήτρια στον μεγαλύτερο διαγωνισμό προβλέψεων που έλαβε ποτέ χώρα, τον *M3*, βασίζεται στην αποσύνθεση της χρονοσειράς σε δύο ή περισσότερες γραμμές ίδιας μέσης τιμής και κλίσης, αλλά διαφορετικής καμπυλότητας. Έτσι, ανάλογα με τις γραμμές που επιλέγονται αναδεικνύονται αντίστοιχα τα βραχυπρόθεσμα ή τα μακροπρόθεσμα χαρακτηριστικά των δεδομένων, ο βέλτιστος εντοπισμός και συνδυασμός των οποίων μπορεί να οδηγήσει σε ιδιαίτερα ακριβείς προβλέψεις. Για τη γενίκευση ωστόσο της μεθόδου και την εξασφάλιση καλύτερης προσαρμογής σε κάθε τύπο δεδομένων, στην παρούσα διατριβή προτάθηκε η επέκτασή της σε δύο κατευθύνσεις:

- Η μέθοδος θ είναι στη βάση της ένα προσθετικό μοντέλο πρόβλεψης, με την έννοια ότι οι επιμέρους γραμμές θ (στην κλασική προσέγγιση 0 και 2) υπολογίζονται με γραμμικό τρόπο και συνδυάζονται προσθετικά. Έτσι, οι συνιστώσες τάσης και επιπέδου, οι οποίες περιγράφονται αντίστοιχα από γραμμές χαμηλής και υψηλής τιμής παραμέτρου θ , είναι ανεξάρτητες μεταξύ τους και δεν επηρεάζει η μία την άλλη κατά την εξέλιξη του μεγέθους. Αυτό φυσικά δεν ισχύει για όλα τα μεγέθη

καθώς μπορεί κάλλιστα το επίπεδο να αλληλεπιδρά με τη τάση των δεδομένων και να μεταβάλλεται ανάλογα με αυτό. Έτσι, μία πρώτη προέκταση της μεθόδου που προτείνεται είναι η εισαγωγή του πολλαπλασιαστικού μοντέλου θ .

- Στην κλασική της έκφραση η μέθοδος θ αποτελείται από δύο γραμμές, την 0 και τη 2, η πρώτη εκ των οποίων ταυτίζεται με την ευθεία απλής γραμμικής εξομάλυνσης και η δεύτερη με τα δεδομένα, διπλάσιας όμως καμπυλότητας. Η ανεξάρτητη προέκταση των δύο γραμμών θ έχει ως αποτέλεσμα την εκτίμηση της μακροπρόθεσμης τάσης της χρονοσειράς μέσω της γραμμής μηδέν, και του επιπέδου μέσω της γραμμής 2. Ο συνδυασμός τους οδηγεί σε μία πρόβλεψη με σημείο εκκίνησης την πρόβλεψη της γραμμής 2 και κλίση το ήμισυ της 0. Όπως γίνεται αντιληπτό, βάσει της παραπάνω ανάλυσης η μέθοδος θ είναι ικανή να προβλέπει αποκλειστικά γραμμικά μοτίβα τάσης, κάτι που προφανώς είναι αρκετά περιοριστικό. Έτσι, μία δεύτερη προέκταση της μεθόδου που προτείνεται είναι η εισαγωγή μη γραμμικών μοντέλων θ , μέσω της αντικατάστασης της κλασικής γραμμής μηδέν με άλλες μη γραμμικές καμπύλες όπως η εκθετική, η λογαριθμική κ.ο.κ..

Σε αυτό το σημείο αναφέρεται πως οι παραπάνω προεκτάσεις γεννούν επί της ουσίας πολλά νέα μοντέλα θ διαφορετικών ιδιοτήτων, υποθέτοντας αφενός διαφορετική διασύνδεση στις συνιστώσες της χρονοσειράς (αθροιστική ή πολλαπλασιαστική) και αφετέρου ποικίλα μοτίβα τάσης (γραμμικά ή μη). Έτσι, μέσω του σφάλματος προσαρμογής των μοντέλων που προκύπτουν στα ιστορικά δεδομένα δύναται να επιλεγεί το θεωρητικά καταλληλότερο και να αξιοποιηθεί για την αποτελεσματικότερη προέκταση της χρονοσειράς. Οι προβλέψεις που παράγονται εποχικοποιούνται μέσω των εποχιακών δεικτών που υπολογίστηκαν νωρίτερα, παράγοντας έτσι τις τελικές προβλέψεις της μεθόδου.

Η προτεινόμενη μεθοδολογία εφαρμόστηκε σε πλήθος πραγματικών και τεχνητών δεδομένων προκειμένου αρχικά να αξιολογηθεί η απόδοσή της σε πραγματικές συνθήκες και επιπλέον να εξεταστεί η συμπεριφορά της για προκαθορισμένες παραμέτρους έντασης, τάσης και τυχαιότητας. Ως πραγματικά δεδομένα επιλέχθηκαν οι χρονοσειρές των διαγωνισμών $M1$ και $M3$ (4004) ενώ ως τεχνητές 135.000 χρονοσειρές που παράχθηκαν μέσω κατάλληλης γεννήτριας που κατασκευάστηκε για πέντε διαφορετικά μοτίβα τάσης (γραμμικό, εκθετικό, λογαριθμικό, αντίστροφο και δύναμη) και τρία επίπεδα έντασης (υψηλό, μεσαίο και χαμηλό).

Τα αποτελέσματα είναι ιδιαίτερα ενθαρρυντικά και αποδεικνύουν πως στη γενική περίπτωση η προτεινόμενη μεθοδολογία οδηγεί σε καλύτερα αποτελέσματα έναντι άλλων κλασικών προσεγγίσεων, με τα οφέλη να οφείλονται τόσο στη βελτιωμένη προετοιμασία των χρονοσειρών, όσο και στη γενικευμένη μέθοδο θ . Στη συνέχεια παρουσιάζονται λοιπόν κάποια από τα βασικότερα συμπεράσματα της διατριβής και πιθανές προεκτάσεις της για την μελλοντική εξέλιξή της.

7.1 Συμπεράσματα διατριβής

Ένα πρώτο συμπέρασμα που προκύπτει από τα αποτελέσματα της διατριβής είναι πως η συρρίκνωση των εποχιακών δεικτών βελτιώνει συστηματικά τις προβλέψεις των εποχιακών χρονοσειρών, με τα οφέλη να αυξάνονται μάλιστα για υψηλά επίπεδα εποχιακότητας. Αυτό είναι πολύ σημαντικό καθώς η αποεποχικοποίηση είναι μία διαδικασία που εφαρμόζεται καθολικά στη βιβλιογραφία για την πρόβλεψη μέσω μη εποχιακών μοντέλων και απ' ό,τι φαίνεται εσφαλμένα δεν επιλέγεται πάντοτε η προτεινόμενη τεχνική για τη βελτιστοποίησή της. Ο παραπάνω ισχυρισμός ενισχύεται μάλιστα από το γεγονός πως η συρρίκνωση των εποχιακών δεικτών μπορεί να εφαρμοστεί με ελάχιστο υπολογιστικό κόστος. Ένας λόγος που μπορεί να αμελείται μέχρι στιγμής η εν λόγω διαδικασία είναι πιθανώς τα σχετικά περιορισμένα κέρδη της σε ακρίβεια, τα οποία ωστόσο δεν είναι και αμελητέα.

Όσον αφορά την τεχνική εξομάλυνσης, στη γενική περίπτωση φαίνεται να οδηγεί σε βελτιωμένη προβλεπτική ακρίβεια και μάλιστα ανεξαρτήτως του τύπου της χρονοσειράς. Το παραπάνω φαινόμενο μπορεί να ερμηνευθεί αρχικά από τον περιορισμό του θορύβου στο σήμα της χρονοσειράς που επιτρέπει στο μοντέλο πρόβλεψης να παραμετροποιηθεί με μεγαλύτερη ακρίβεια αποφεύγοντας αστοχίες που οφείλονται σε πιθανές ακραίες τιμές και απότομες μεταβολές επιπέδου. Ένας άλλος λόγος που η εν λόγω τεχνική οδηγεί σε βελτίωση της ακρίβειας είναι και η ανάδειξη του μοτίβου τάσης της χρονοσειράς. Δεδομένου ότι η γενικευμένη μέθοδος θ μπορεί να περιγράψει τόσο γραμμικά όσο και μη γραμμικά μοτίβα τάσης, γίνεται κατανοητό πως η παραπάνω ιδιότητα είναι ιδιαίτερα ωφέλιμη τόσο για την επιλογή του βέλτιστου μοντέλου θ για την προέκτασή του μοτίβου, όσο και για την καλύτερη εκτίμηση των παραμέτρων της εξίσωσης που το περιγράφει. Έτσι, η τεχνική εξομάλυνσης οδηγεί συνδυαστικά σε προβλέψεις οι οποίες είναι μία βελτιωμένη εκτίμηση του τρέχοντος επιπέδου της χρονοσειράς - απαλλαγμένο από την τυχαιότητα - προσαυξημένο κατά την κλίση του μοτίβου ανάπτυξης - τονισμένο από την εξομάλυνση.

Ένα ακόμα πλεονέκτημα της τεχνικής εξομάλυνσης είναι η απλοποίηση της μεθόδου θ αυτής καθαυτής καθώς, μετά την ομαλοποίηση της χρονοσειράς η απλή εκθετική εξομάλυνση που χρησιμοποιείται για την προέκταση της γραμμής 2 λαμβάνει παράμετρο εξομάλυνσης ίση με 1, δηλαδή ταυτίζεται με την απλοϊκή μέθοδο. Αυτό είναι αρκετά σημαντικό καθώς αφενός έτσι μειώνεται ο χρόνος εκτέλεσης της διαδικασία πρόβλεψης και αφετέρου απαλλάσσεται η προέκταση του επιπέδου από την αβεβαιότητα που αναφέρεται στην παραμετροποίηση του μοντέλου που το περιγράφει. Έτσι, η μέθοδος θ μετασχηματίζεται σε μία μέθοδο αποσύνθεσης το επίπεδο των προβλέψεων της οποίας καθορίζεται μη παραμετρικά βάσει της διαδικασίας εξομάλυνσης και η τάση τους βάσει του μοτίβου ανάπτυξης που εντοπίστηκε.

Σε αυτό το σημείο αναφέρεται βέβαια πως η τεχνικής εξομάλυνσης δεν είναι σε καμία περίπτωση πανάκεια καθώς ενδέχεται αν δεν τερματιστεί έγκαιρα να αλλοιώσει το σήμα της χρονοσειράς και να οδηγήσει σε μειωμένη ακρίβεια. Έτσι, σε περιπτώσεις χρονοσειρών που είναι επαρκώς ομαλές ενδέχεται να πρέπει να αποφευχθεί να εφαρμοστεί η εν λόγω τεχνική. Βέβαια, δεδομένου ότι σε ομαλές χρονοσειρές το σφάλμα

πρόβλεψης είναι ήδη περιορισμένο, το εν λόγω φαινόμενο δεν είναι αποθαρρυντικό για τη μεθοδολογία και μπορεί να θεωρηθεί δευτερευούσης σημασίας.

Ενθαρρυντικό είναι επίσης το γεγονός πως, σύμφωνα με τα αποτελέσματα, τα οφέλη συρρίκνωσης εποχιακών δεικτών και εξομάλυνσης φαίνεται να είναι ανεξάρτητα μεταξύ τους. Αυτό σημαίνει πως οι εν λόγω τεχνικές μπορούν να απομονωθούν όποτε αυτό κρίνεται απαραίτητο (π.χ. σε περιπτώσεις ομαλών χρονοσειρών) και να αξιοποιηθούν κατάλληλα και σε άλλες αντίστοιχες μεθοδολογίες. Για παράδειγμα, η συρρίκνωση εποχιακών δεικτών μπορεί να συνδυαστεί κάλλιστα με κλασικές μεθόδους πρόβλεψης όπως η εκθετική εξομάλυνση για την βελτίωση της προβλεπτικής τους ακρίβειας. Σημειώνεται ωστόσο πως στην περίπτωση της τεχνικής εξομάλυνσης αυτό θα πρέπει να γίνεται με κριτική σκέψη καθώς, όπως αναφέρθηκε, αν και η εν λόγω τεχνική βοηθά στον καλύτερο προσδιορισμό του επιπέδου της χρονοσειράς, τονίζει ταυτόχρονα την τάση της. Έτσι, αν στην εξομαλυμένη χρονοσειρά εφαρμοστεί ένα μοντέλο πρόβλεψης που στηρίζεται αποκλειστικά στα πρόσφατα δεδομένα της χρονοσειράς για την προέκτασή της (βλ. *Holt*) και όχι στην μακροπρόθεσμη συμπεριφορά της (βλ. θ) ενδέχεται να παραχθούν υπερβολικά προκατειλημμένες προβλέψεις μειωμένης ακρίβειας. Αυτό δεν ισχύει βέβαια στην περίπτωση μοντέλων που δεν περιγράφουν τη τάση της χρονοσειράς, όπως π.χ. η απλή μέθοδος εξομάλυνσης σταθερού επιπέδου, όπου και αναμένονται σημαντικές βελτιώσεις λόγω του καλύτερου προσδιορισμού του επιπέδου.

Όσον αφορά την επιλογή του βέλτιστου μοντέλου πρόβλεψης θ για την προέκταση χρονοσειρών, ένα βασικό συμπέρασμα που προκύπτει είναι πως αυτή μπορεί να επιτευχθεί με υψηλά επίπεδα επιτυχίας μόνο σε περιπτώσεις όπου τα μοτίβα της χρονοσειράς διατηρούνται αυτούσια στο μέλλον. Χαρακτηριστικό παράδειγμα αποτελούν οι εξαιρετικές επιδόσεις της επιλεκτικής μεθόδου στα τεχνητά δεδομένα και η πιο περιορισμένη αποτελεσματικότητά της στα δεδομένα των διαγωνισμών *M*. Αυτό συμβαίνει καθώς στην πρώτη περίπτωση γίνεται εμμέσως μέσω του ορισμού της γεννήτριας η υπόθεση πως κάθε χρονοσειρά που παράγεται θα διατηρεί το μοτίβο της τάσης της στο μέλλον και μάλιστα οι συνιστώσες αυτής θα συνδέονται μεταξύ τους πολλαπλασιαστικά. Αντίθετα, στα πραγματικά δεδομένα το μοτίβο εξέλιξης της χρονοσειράς μπορεί να είναι οποιοδήποτε και επιπλέον να αλλάζει αναπάντεχα, οδηγώντας έτσι ένα προβλεπτικά ανώτερο μοντέλο στο παρελθόν σε μη ακριβείς προβλέψεις στο μέλλον.

Φυσικά η τυχαιότητα επηρεάζει και αυτή με τη μεριά της σημαντικά την ικανότητα της προτεινόμενης μεθόδου να επιλέγει με επιτυχία το ακριβέστερο μοντέλο πρόβλεψης. Όσο αυξάνεται η τυχαιότητα τόσο αλλοιώνεται το σήμα της χρονοσειράς και το μοτίβο εξέλιξής του. Έτσι, οποιαδήποτε επιλογή και αν γίνει, ακόμα και μετά την εφαρμογή της εξομάλυνσης, έχει πιο περιορισμένες πιθανότητες να επιτύχει. Ωστόσο, σε οποιαδήποτε άλλη περίπτωση, και ειδικά σε περιπτώσεις χρονοσειρών με αυξημένη ένταση στη συνιστώσα της τάσης, η επιλεκτική μέθοδος αποδίδει βέλτιστα οδηγώντας σε σημαντικά καλύτερες προβλέψεις από τα μεμονωμένα μοντέλα. Χαρακτηριστικά αναφέρεται η περίπτωση των ετήσιων δεδομένων των διαγωνισμών τα οποία χαρακτηρίζονται από ομαλότητα και αυξημένη ένταση τάσης, όπου η επιλεκτική μέθοδος άριστευσε με διαφορά σε σύγκριση με τα μεμονωμένα μοντέλα.

Ένα άλλο χρήσιμο συμπέρασμα που εξάγεται είναι πως στη γενική περίπτωση τα προσθετικά μοντέλα θ είναι αρκετά πιο ακριβή από τα πολλαπλασιαστικά. Ωστόσο οι διαφορές τους είναι σχετικά μικρές και μειώνονται ακόμα περισσότερο για τα μοντέλα μη γραμμικής τάσης. Ενθαρρυντικό είναι επίσης το γεγονός πως μετά την επεξεργασία των δεδομένων, οι διαφορές γίνονται ακόμα μικρότερες, μειώνοντας έτσι την αβεβαιότητα κατά την επιλογή του βέλτιστου μοντέλου πρόβλεψης. Αυτό εξηγείται σε μεγάλο βαθμό από την επίδραση της τεχνικής εξομάλυνσης: Όταν η χρονοσειρά εξομαλύνεται τα δεδομένα γίνονται ομαλά και αρκετά γραμμικά. Έτσι, η απουσία σημαντικών διακυμάνσεων στη χρονοσειρά έχει ως αποτέλεσμα να μην παρατηρούνται έντονες αλληλεπιδράσεις μεταξύ επιπέδου και τάσης και το πολλαπλασιαστικό μοντέλο να προσεγγίζει αρκετά το γραμμικό. Σε κάθε περίπτωση πάντως μεμονωμένα και οι δύο εκφράσεις της μεθόδου μπορούν να προσδώσουν οφέλη και για αυτό το λόγο η επιλογή του βέλτιστου μοντέλου θ βάσει κατάλληλων διαγνωστικών ελέγχων οδηγεί συνολικά σε καλύτερες προβλέψεις.

Βάσει των παραπάνω παρατηρήσεων ακόμα μεγαλύτερη επίδραση φαίνεται να έχει λοιπόν στην ακρίβεια πρόβλεψης η σωστή επιλογή μοτίβου τάσης το οποίο και παραμένει σε σημαντικό βαθμό αναλλοίωτο μετά την εξομάλυνση της χρονοσειράς. Μάλιστα τα πλεονεκτήματα αναγνώρισης του βέλτιστου μοτίβου διατηρούνται ανεξαρτήτων συχνότητας, τύπου δεδομένων και είδους εφαρμογής. Έτσι, ανάλογα με το μοτίβο τάσης της χρονοσειράς, διαφορετικά μοντέλα μπορούν να αξιοποιηθούν και να οδηγήσουν σε βελτιωμένες προβλέψεις. Βάσει των αποτελεσμάτων στα δεδομένα της γεννήτριας, για τα οποία και το μοτίβο τάσης είναι ελεγχόμενο, αναφέρονται ενδεικτικά οι παρακάτω εμπειρικοί κανόνες επιλογής βέλτιστου μοντέλου βάσει μοτίβου που εντοπίστηκε: Εκθετικό για γραμμική τάση, Εκθετικό για εκθετική τάση, Λογαριθμικό για Λογαριθμική Τάση, Αντίστροφο για αντίστροφη τάση και Γραμμικό για τάση σε μορφή δύναμης.

Οι παραπάνω εμπειρικοί κανόνες μπορούν να εξηγηθούν και θεωρητικά βάσει της παραμέτρου $\theta=2$ που χρησιμοποιείται στα πειράματα. Για χρήση δύο γραμμών $\theta 0$ και 2 , η τάση των προβλέψεων θα είναι ως γνωστός το ήμισυ της υπολογισμένης. Έτσι, αν η χρονοσειρά έχει ανοδική γραμμική τάση και δεν προκύψουν μεταβολές στο μοτίβο ανάπτυξής της, τότε οι προβλέψεις που θα παραχθούν θα είναι απαισιόδοξες. Αντίστοιχα το ίδιο θα ισχύει για την περίπτωση καθοδικής γραμμικής τάσης όπου και οι προβλέψεις θα είναι αισιόδοξες. Για να λυθεί το εν λόγω φαινόμενο μπορεί κανείς είτε να αντικαταστήσει τη γραμμή μηδέν με εκθετική και λογαριθμική αντίστοιχα, ή να αυξήσει τη τιμή της παραμέτρου θ . Η δεύτερη περίπτωση δεν εξετάζεται στην παρούσα διατριβή και σε κάθε περίπτωση η πρώτη είναι περισσότερο αποδοτική λύση καθώς, ακόμα και αν υπολογιστεί βέλτιστα η παράμετρος, αυτό δεν θα οδηγήσει σε καμία περίπτωση σε τέλειες προβλέψεις λόγω της μη γραμμικότητας των δεδομένων. Έτσι, καταλαβαίνει κανείς πως, ανεξάρτητα του πραγματικού μοτίβου της χρονοσειράς, υπάρχουν επιπλέον λόγοι για τους οποίους κάποιο μοντέλο θ ενδέχεται να υπερτερεί έναντι των υπολοίπων και αναδεικνύεται η σημασία της γενίκευσης του μοντέλου θ για γραμμικά και μη γραμμικά μοτίβα τάσης.

Ως προς τα χαρακτηριστικά των χρονοσειρών και το πώς αυτά επηρεάζουν την ακρίβεια της προτεινόμενης μεθοδολογίας, αναφέρεται αρχικά πως καθώς αυξάνεται ο θόρυβος

αυξάνονται και τα οφέλη της μεθοδολογίας, κυρίως λόγω της επίδρασης της τεχνικής εξομάλυνσης. Στην κλασική προσέγγιση τα μοντέλα τείνουν να αποπροσανατολίζονται από τη τυχαιότητα στα δεδομένα και να παράγουν μη ρεαλιστικές προβλέψεις. Αντίθετα, με την προτεινόμενη προσέγγιση, μέσω της οποίας εντοπίζεται ένα στιβαρό επίπεδο για τη χρονοσειρά, εξασφαλίζονται αρκετά πιο αντιπροσωπευτικές προβλέψεις. Σημειώνεται ωστόσο πως από ένα επίπεδο θορύβου και μετά η προτεινόμενη προσέγγιση ενδέχεται να μην μπορεί να προσδώσει σημαντικά κέρδη καθώς σε αυτήν τη περίπτωση η τυχαιότητα υπερέχει του σήματος και συνεπώς ανεξαρτήτως ορθής επεξεργασίας ή μη οποιαδήποτε μέθοδος αδυνατεί να προσδιορίσει το μέλλον αποδοτικά.

Παρόμοια εικόνα σχηματίζεται και για την περίπτωση της εποχιακότητας όπου, αν και η τεχνική συρρίκνωσης δεικτών συμβάλει στην αποτελεσματικότερη πρόβλεψη των χρονοσειρών, από ένα επίπεδο έντασης με μετά αυτή αδυνατεί να πετύχει κάποια σημαντική διαφορά. Αυτό συμβαίνει καθώς σε έντονα εποχιακές χρονοσειρές, στις οποίες απουσιάζει προφανώς η τυχαιότητα, η συνιστώσα της εποχιακότητας είναι κυρίαρχη και μάλιστα ικανοποιητικά προβλέψιμη από οποιαδήποτε εναλλακτική τεχνική. Έτσι, η συρρίκνωση και ο επαναπροσδιορισμός των δεικτών λόγω αβεβαιότητάς δεν προσδίδει κάτι στη μεθοδολογία και οδηγεί σε παρεμφερείς προβλέψεις. Αντίθετα, για λογικά επίπεδα τάσης και τυχαιότητας, η προτεινόμενη προσέγγιση οδηγεί σε βελτιωμένες προβλέψεις μέσω του καλύτερου προσδιορισμού των συνιστωσών τάσης, επιπέδου και εποχιακότητας.

Τέλος, για την περίπτωση της τάσης, όπως αναφέρθηκε και νωρίτερα αυτή εξυπηρετεί όσο αυξάνεται τον εντοπισμό του καλύτερου μοντέλου θ και συνεπώς εκεί αποδίδει καλύτερα και η προτεινόμενη μεθοδολογία. Για χαμηλά επίπεδα έντασης τάσης η εν λόγω επιλογή αναμένεται λοιπόν να μην είναι η βέλτιστη δυνατή, οδηγώντας σε περιορισμένα κέρδη, ενώ για μεσαία και υψηλά αναμένεται να προσδώσει υψηλά κέρδη.

Συνοπτικά, με την εφαρμογή της προτεινόμενης μεθοδολογίας μπορεί κανείς να επιλέξει με σχετική ασφάλεια το βέλτιστο μοντέλο θ για κάθε χρονοσειρά και να αποκομίσει σημαντικά οφέλη σε ακρίβεια και υπολογιστικό κόστος προσδιορίζοντας αποδοτικότερα τις συνιστώσες εποχιακότητας, τάσης και επιπέδου. Αν η ένταση της τυχαιότητας είναι ιδιαίτερα χαμηλή, τότε ενδέχεται να πρέπει να παρακαμφθεί το βήμα της εξομάλυνσης για την αποφυγή τυχόν αλλοιώσεων, ενώ αν η χρονοσειρά είναι στάσιμη με σημαντικές διακυμάνσεις, τότε ενδέχεται η επιλογή του μοντέλου να μην είναι η βέλτιστη. Σε οποιαδήποτε άλλη περίπτωση η μεθοδολογία αναμένεται να αποδώσει καλύτερα από τις κλασικές μεθόδους πρόβλεψης και ιδιαίτερα σε αυτές σημαντικής τάσης, εποχιακότητας και τυχαιότητας.

7.2 Προεκτάσεις διατριβής

Αν και η προτεινόμενη μεθοδολογία αποδεικνύεται πως μπορεί να βελτιώσει σημαντικά την επίδοση της διαδικασίας πρόβλεψης σε σχέση με άλλες κλασικές προσεγγίσεις, υπάρχουν αρκετές προεκτάσεις της προκειμένου αυτή να καταστεί ακόμα πιο αποδοτική και να αποδώσει πραγματικό όφελος για επιχειρήσεις και οργανισμούς.

Μία πρώτη επέκταση της παρούσας μεθοδολογίας θα ήταν λοιπόν η ενσωμάτωση στην σε αυτήν επιπλέον τεχνικών προετοιμασίας δεδομένων προκειμένου η γενικευμένη μέθοδος θ να είναι σε θέση να οδηγήσει σε ακόμα καλύτερες προβλέψεις. Όπως φάνηκε από τα αποτελέσματα της εργασίας, τα οφέλη που προέρχονται από την προ-επεξεργασία των δεδομένων είναι ιδιαίτερα σημαντικά και ως εκ τούτου αξίζει να μελετηθεί περαιτέρω. Έτσι, η εισαγωγή π.χ. μίας τεχνικής για την αναγνώριση και την αφαίρεση ειδικών γεγονότων που αλλοιώνουν τα μοτίβα της χρονοσειράς και αποπροσανατολίζουν το μοντέλο πρόβλεψης, θα ήταν μία πρώτη χρήσιμη προέκταση. Μάλιστα, δεδομένου ότι η προτεινόμενη εξομάλυνση αφαιρεί ακραίες τιμές (additive outliers) αλλά δυσκολεύεται εκ φύσεως να απομακρύνει άλλα είδη ειδικών γεγονότων, όπως π.χ. μεταβολές επιπέδου (level shift), η παραπάνω διαδικασία θα μπορούσε να επικεντρωθεί σε συγκεκριμένες κατηγορίες outlier, περιορίζοντας έτσι ταυτόχρονα το επιπλέον υπολογιστικό κόστος που υπεισέρχεται από αυτήν.

Άλλη προέκταση της παρούσας διατριβής θα ήταν ο αναλυτικότερος έλεγχος του τρόπου με τον οποίο αλληλοεπιδρά η προτεινόμενη προσέγγιση επεξεργασίας δεδομένων με άλλα μοντέλα πρόβλεψης χρονοσειρών εκτός του θ . Για παράδειγμα, στο κεφάλαιο που παρουσιάστηκε η τεχνική εξομάλυνσης φάνηκε πως τα μοντέλα εκθετικής εξομάλισης με τάση, όπως η *holt*, δεν συνεργάζονται αποδοτικά με την προτεινόμενη τεχνική οδηγώντας στη γενική περίπτωση σε χειρότερες προβλέψεις. Αυτό οφείλεται όπως αναφέρθηκε στο γεγονός ότι η εξομάλυνση υπερτονίζει την πρόσφατη τάση στα δεδομένα, την οποία και τα μοντέλα εκθετικής εξομάλυνσης την εφαρμόζουν αυτούσια στην πρόβλεψη. Έτσι, σε αντίθεση με την μέθοδο θ που στηρίζεται στην μακροχρόνια και στιβαρή τάση της χρονοσειράς, αυτά οδηγούνται σε προκατειλημμένες προβλέψεις υψηλού σφάλματος. Ως εκ τούτου, θα είχε ιδιαίτερο ενδιαφέρον να μελετηθεί ποιοι μέθοδοι μπορούν να αξιοποιηθούν εναλλακτικά για τη βελτίωση της απόδοσής τους, ή αντίστοιχα ποιες παραμετροποιήσεις πρέπει να γίνουν στην προτεινόμενη τεχνική εξομάλυνσης για την παραγωγή ακριβέστερων προβλέψεων. Για παράδειγμα, οι μετασχηματισμοί δεδομένων έχει αποδειχτεί στη βιβλιογραφία πως συμβάλουν στη βελτίωση της προβλεπτικής ικανότητας των εν λόγω μοντέλων, σε περιορισμένο όμως βαθμό. Ίσως η μελέτη άλλων τεχνικών πέραν των μετασχηματισμών δεδομένων να κρύβουν σημαντικότερα οφέλη.

Άλλη προέκταση της εργασίας θα μπορούσε να αφορά τη βελτίωση ή την αντικατάσταση του αλγορίθμου επιλογής βέλτιστου μοντέλου θ με άλλη αποδοτικότερη προσέγγιση. Για παράδειγμα, θα μπορούσε να εξεταστεί η πιθανότητα του συνδυασμού των προβλέψεων που προέρχονται από τα επικρατέστερα μοντέλα θ , ή της εξαίρεσης των χειρότερων εξ αυτών. Έτσι, τα οφέλη του συνδυασμού προβλέψεων θα έβρισκαν άμεσα εφαρμογή στην παρούσα μεθοδολογία και θα οδηγούσαν πιθανώς σε καλύτερες προβλέψεις. Τέτοιες

πρακτικές ερευνήθηκαν σε κάποιο βαθμό ωστόσο τα αποτελέσματα ήταν διφορούμενα και χρήζουν περισσότερης μελέτης.

Μεγάλο ενδιαφέρον θα εμφάνιζε επίσης η εφαρμογή και η αξιολόγηση της προτεινόμενης προσέγγισης σε πραγματικό περιβάλλον. Μπορεί η μεθοδολογία που αναπτύχθηκε να εφαρμόστηκε με επιτυχία σε μεγάλο πλήθος χρονοσειρών πληθώρας χαρακτηριστικών, επιβεβαιώνοντας την προβλεπτική της ανωτερότητα, ωστόσο η χρήση της για παραγωγή προβλέψεων στο πραγματικό περιβάλλον μιας επιχείρησης θα μπορούσε να οδηγήσει σε επιπλέον χρήσιμα συμπεράσματα. Για παράδειγμα, ίσως να υπάρχουν τύποι δεδομένων που δεν εξετάστηκαν και για τους οποίους η μεθοδολογία δεν αποδίδει ιδιαίτερα καλά. Επίσης, είναι πιθανό βάσει χρόνου εκτέλεσης και ακρίβειας που επιθυμείται από την επιχείρηση να μην κρίνεται σε όλες της περιπτώσεις απαραίτητη η χρήση της μεθοδολογίας έναντι άλλων πιο απλοϊκών προσεγγίσεων. Ο εντοπισμός τέτοιων περιπτώσεων και η εύρεση πρακτικά αποδοτικών λύσεων είναι νευραλγική για την ουσιαστική προσφορά αντίστοιχων μεθόδων στις επιχειρήσεις.

Άλλη χρήσιμη επέκταση της παρούσας διατριβής είναι η ενσωμάτωσή της εντός ενός ολοκληρωμένου συστήματος υποστήριξης προβλέψεων προκειμένου αυτή να γίνει άμεσα και εύκολα διαθέσιμη σε επιχειρήσεις και οργανισμού. Η εν λόγω επέκταση έχει ήδη ξεκινήσει υπό την ονομασία *OMEN* και αυτή τη στιγμή βρίσκεται στο στάδιο ελέγχου ποιότητας λειτουργίας και προσθήκης επιπλέον δυνατοτήτων. Το *OMEN* είναι ένα διαδικτυακό εργαλείο παραγωγής προβλέψεων το οποίο κάνει χρήση των βιβλιοθηκών της γλώσσας προγραμματισμού R - στην οποία και αναπτύχθηκαν όλες οι μέθοδοι και οι τεχνικές προβλέψεων που παρουσιάζονται στην εργασία – για την ανάλυση και την προέκταση των χρονοσειρών, ενώ επιπλέον χρησιμοποιεί διεπαφές ούτως ώστε οι χρήστες να μπορούν να αξιοποιούν τις τεχνικές που παρέχονται μέσω εύχρηστων οθονών και εργαλείων. Η εφαρμογή είναι διαθέσιμη στη διεύθυνση <http://omen-project.eu/> και για λόγους πληρότητας, πέραν της παρούσας μεθοδολογίας, ενσωματώνει και μεγάλο πλήθος άλλων κλασικών και προηγμένων τεχνικών και μοντέλων προβλέψεων.

Η επέκταση της παρούσας μεθοδολογίας για χρονοσειρές διακοπτόμενης ζήτησης θα είχε και αυτή με τη σειρά της αρκετό ενδιαφέρον, κυρίως λόγω του ότι όταν κανείς αναφέρεται σε μεγάλο όγκο δεδομένων σε επιχειρήσεις και οργανισμού, έρχεται αντιμέτωπος συνήθως με την παραγωγή προβλέψεων σε επίπεδο καταστήματος και κωδικού προϊόντων. Σε αυτές τις περιπτώσεις τα δεδομένα χαρακτηρίζονται από διακοπές, και συνεπώς χρήζουν επιπλέον επεξεργασίας προκειμένου κοινά μοντέλα πρόβλεψης, όπως η μέθοδος θ , να μπορούν να τα διαχειριστούν αποδοτικά.

Τέλος, δεδομένου ότι η μέθοδος θ , αν απλοϊκή στη βάση της, αποδεικνύεται μία ευέλικτη και ιδιαίτερος ακριβής τεχνική προέκτασης χρονοσειρών, γίνεται κατανοητό πως η περαιτέρω έρευνα γύρω από της ιδιότητές της έχει ακόμα πολλά να προσφέρει. Για παράδειγμα, θα είχε ενδιαφέρον να ενσωματωθεί στη μεθοδολογία μία τεχνική για την αυτόματη επιλογή της βέλτιστης παραμέτρου θ , η οποία και όπως παρουσιάστηκε μπορεί να επηρεάσει σημαντικά την προβλεπτική ικανότητα των μοντέλων θ που προκύπτουν. Επιπλέον, θα μπορούσε να εξεταστεί η χρήση πολλαπλών γραμμών θ , η εισαγωγή νέων

γραμμών μηδέν, ο συνδυασμός πολλών διαφορετικών μοτίβων τάσης ταυτοχρόνως και η επιλογή διαφορετικών μοντέλων πρόβλεψης πέραν της SES για την προέκταση καθεμίας εκ των γραμμών θ που προκύπτουν.

Βιβλιογραφία

Πετρόπουλος, Φ., Ασημακόπουλος, Β. (2011). Επιχειρησιακές προβλέψεις, Συμμετρία, Αθήνα

Πετρόπουλος, Φ. (2012). Παράγοντες που επηρεάζουν την ακρίβεια πρόβλεψης χρονοσειρών: Προσομοίωση και ανάλυση. Διδακτορική διατριβή, Αθήνα, Απρίλιος 2012

Akaike, H., Petrov, B.N., Csáki, F. (1973). Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, p. 267-281

Assimakopoulos, V. (1995). A successive filtering technique for identifying long- term trends. *Journal of Forecasting*, John Wiley 14 (1), 35 – 43

Assimakopoulos, V., Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting* 16 (4), 521 – 530

Beaumont, A. N. (2014). Data transforms with exponential smoothing methods of forecasting. *International Journal of Forecasting*, 30 (4), 918-927

Bougioukos N., Nikolopoulos K., Assimakopoulos V. (2007). Combining the Theta Model with Artificial Intelligence. 27th International Symposium on Forecasting ISF 2007, June 23-27 2007, New York, USA

Bougioukos N., Tsiafa E., Nikolopoulos K. and Assimakopoulos V. (2008). ThetaAI: blending the theta model with Artificial Intelligence heuristics. 28th International Symposium on Forecasting ISF 2008, June 22-25 2008, Nice, France

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, 26, 211–252

Box, G., Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day

Brown, R.G. (1959). *Statistical forecasting for inventory control*, McGraw-Hill, New York

Cleveland, W. S., Dunn, D. M., Terpenning I. J. (1978). SABL: A Resistant Seasonal Adjustment Procedure with Graphical Methods for Interpretation and Diagnosis. Chapter in National Bureau of Economic Research book *Seasonal Analysis of Economic Time Series*, 201 - 241

Cleveland, R. B., Cleveland, W. S., McRae, J.E., Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, 6, 3–73

Constantinidou C., Nikolopoulos K., Bougioukos N., Tsiafa E., Petropoulos F. and Assimakopoulos V. (2012) "A neural network approach for the Theta model", in: Ming M.

(ed.), Information Engineering, Lecture Notes in Information Technology, Vol. 25, pp. 116-120

Engle, R. F., Yoo, B. S. (1987). Forecasting and testing in co-integrated systems. *Journal of Econometrics*, 35 (1), 143-159

Epanechnikov, V. A. (1969). Non-Parametric Estimation of a Multivariate Probability Density. *Theory of Probability and Its Applications*, 14 (1), 153–158

Farrell P. J., Stewart, K. R. (2006). Comprehensive study of tests for normality and symmetry: Extending the Spiegelhalter test. *Journal of Statistical Computation and Simulation*, 76(9), 803-816

Fiorucci, J. A., Pellegrini, T. R., Louzada, F., Petropoulos, F., Koehler, A. B. (2016). Models for optimising the theta method and their relationship to state space models. *International Journal of Forecasting*, 32 (4), 1151–1161

Gardner, E.S. (1985). Exponential smoothing: the state of the art. *Journal of Forecasting*, 4, 1–28

Gorr, W. L., Schneider, M. J. (2013). Large-change forecast accuracy: Reanalysis of M3-Competition data using receiver operating characteristic analysis. *International Journal of Forecasting*, 29 (2), 274–281

Guerrero, V.M. (1993). Time-series analysis supported by power transformations. *Journal of Forecasting*, 12, 37–48

Hibon, M., Evgeniou, T. (2005). To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting* 21, 15 – 24.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441 & 498–520.

Hyndman, R., Koehler, A., Snyder, R., Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 18 (3), 439–454

Hyndman R. J., Billah B. (2003). Unmasking the Theta method. *International Journal of Forecasting*, 19 (2), 287-290

Hyndman, R. J., Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679-688

Hyndman, R., Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 26 (3), 1–22

Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D. (2008). *Forecasting with exponential smoothing: the state space approach*, Springer. URL: <http://www.exponentialsMOOTHING.net>

Hyndman, R.J. (2014). Forecast: Forecasting functions for time series and linear models. R package version 5.6. URL: <http://CRAN.R-project.org/package=forecast>

Hyndman, R. (2015). forecast: Forecasting functions for time series and linear models. R package version 5.8. <http://github.com/robjhyndman/forecast>

Hyndman, R., Athanasopoulos, G. (2016). Forecasting: principles and practice. URL: <http://otexts.com/fpp/>

Hyndman, R. J., Kang, Y., and Smith-Miles, K. (2016). Exploring Time series collections used for forecast evaluation. 36th International Symposium on Forecasting (ISF 2016), June 19 – 22, 2016, Santander, Spain

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer New York Heidelberg Dordrecht London

Johnson, N. L. (1949). Systems of Frequency Curves Generated by Methods of Translation. URL: <http://www.jstor.org/stable/2332539>

Johnson, V. E. (2013). Revised standards for statistical evidence. Proceedings of the national academy of sciences of the United States of America, 110 (48), 19313-19317

Kang, Y., Hyndman, R. J., Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. International Journal of Forecasting 33 (2), 345-358.

Kendall, M., Stuart, A. (1983). The Advanced Theory of Statistics, volume 3. Griffin, 410–414

Keskin S. (2006). Comparison of Several Univariate Normality Tests Regarding Type I Error Rate and Power of the Test in Simulation based Small Samples. Journal of Applied Science Research, 2(5), 296-300

Koning, A. J. Franses, P. H., Hibon, M., Stekler, H. O. (2005). The M3 competition: Statistical tests of the results. International Journal of Forecasting, 21 (3), 397-409

Kourentzes, N., Petropoulos, F., Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. International Journal of Forecasting, 30 (2), 291 - 302

Kuiper, J. (1979). A Survey and Comparative Analysis of Various Methods of Seasonal Adjustment. Chapter in National Bureau of Economic Research book Seasonal Analysis of Economic Time Series, 57 - 96

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., Shin, Y. (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root. Journal of Econometrics, 54, 159–178

Ladiray, D., Quenneville, B. (2001). Seasonal adjustment with the X-11 method, Lecture notes in statistics, Vol. 158, Springer-Verlag.

Lawrence, M. (2001). Commentaries on the M3-Competition. Why another study? *International Journal of Forecasting* 17, 574–575.

Ledolter, J. (1989). The effect of additive outliers on the forecasts from ARIMA models. *International Journal of Forecasting*, 5 (2), 231 - 240

Makridakis, S., Hibon, M., Moser, C. (1979). Accuracy of Forecasting: An Empirical Investigation. *Journal of the Royal Statistical Society. Series A (General)*, 142 (2), 97-145

Makridakis, S., Wheelwright, S. C., Mc Gee, V. E. (1983). *Forecasting Methods and applications (Second Edition)*. New York: Wiley

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., Winkler, R. (1982). The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition. *Journal of Forecasting*, 1 (2), 111-153

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., Simmons, L.F. (1993). The M-2 Competition: a real-time judgmentally based forecasting study. *International Journal of Forecasting*, 9 (1), 5-23

Makridakis, S., Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting* 16 (4), 451 – 476

McKenzie, S. K. (1984). Concurrent Seasonal Adjustment With Census X-11. *Journal of Business & Economic Statistics*, 2 (3), 235-249

McKenzie, M. D. (1999). Power transformation and forecasting the magnitude of exchange rate changes, *International Journal of Forecasting*, Volume 15, 1 (1), 49-55

Mendes, M., Pala, A. (2003). Type I Error Rate and Power of Three Normality Tests. *Information Technology Journal*, 2, 135-139

Miller D.M., Williams D. (2003). Shrinkage estimators of time series seasonal factors and their effect on forecasting accuracy. *International Journal of Forecasting* 19 (3), 669 – 684

Nelson, H. L. Jr., Granger C. W. J. (1979). Experience with using the Box–Cox transformation when forecasting economic time series. *Journal of Econometrics*, 10, 57–69

Nelson, C. R., Plosser, C. R. (1982). Trends and random walks in macroeconomic time series. *Journal of Monetary Economics*, 10 (2), 139-162

Nikolopoulos, K., Assimakopoulos, V. (2003). Theta intelligent forecasting information system. *Industrial Management and Data Systems* 103 (9), 711 – 726

- Nikolopoulos K. and Assimakopoulos V. (2005) Theta model: decomposition approach or just SES with drift?. 25th International Symposium on Forecasting ISF 2005, June 12-15 2005, San Antonio, USA
- Nikolopoulos, K., Assimakopoulos, V., Bougioukos, N., Litsa, A., Petropoulos, F. (2012). The Theta model: An essential forecasting tool for Supply Chain planning. *Lecture Notes in Electrical Engineering* 123 (1), 431 – 437
- Nikolopoulos, K., Thomakos, D., Petropoulos, F., Litsa, A., Assimakopoulos, V. (2012). Forecasting S&P500 with the Theta Model. *International Journal of Financial Economics and Econometrics*, Accepted 4 (1), 73 – 780
- Ord, K. (2001). Commentaries on the M3-Competition. An introduction, some comments and a scorecard. *International Journal of Forecasting* 17, 537–541.
- Pagourtzi, E., Assimakopoulos, V., Litsa, A. (2007). Theta Model forecasts quarterly & monthly dwelling prices in UK. *Briefings in Real Estate Finance* 17 (1), 1 – 30
- Petropoulos F., Nikolopoulos K., Assimakopoulos V., Pagourtzi E., Bougioukos N. (2007). Optimizing the Theta model for the T-Competition data. 22nd European Conference on Operational Research EURO XXII, July 8-11 2007, Prague, Czech Republic
- Petropoulos, F., Nikolopoulos, K., Assimakopoulos, V. (2008). Theta-Croston and Croston-Theta: investigating new ideas for Intermittent Demand forecasting. 28th International Symposium on Forecasting ISF 2008, June 22-25 2008, Nice, France
- Petropoulos, F., Nikolopoulos, K. (2013). Optimizing the theta model for monthly time series. *Proceedings of the 5th International Conference of Agents and Artificial Intelligence*
- Petropoulos, F., Makridakis, S., Assimakopoulos, V., Nikolopoulos, K. (2014). 'Horses for Courses' in demand forecasting, *European Journal of Operational Research*, 237 (1), 152-163
- Proietti, T., Lütkepohl, H. (2013). Does the Box–Cox transformation help in forecasting
- R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. <http://www.R-project.org>
- Razali, N.M., Wah, Y.B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2 (1), 21-33
- Said, S. E. and Dickey, D. A. (1984). Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika*, 71, 599–607
- Shapiro S. S., Wilk, M. B. (1965). An Analysis of Variance Test for Normality. *Biometrika*, 52 (3/4), 591-611

Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6 (2), 461–464

Scott, D. W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.

Sheather, S. J., Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society series B*, 53, 683–690

Silvestrini, A., Veredas, D. (2008). Temporal aggregation of univariate and multivariate time series models: a survey. *Journal of Economic Surveys*, 22 (3), 458 - 497

Spencer, J. (1904). On the graduation of the rates of sickness and mortality. *Journal of the Institute of Actuaries*, 38, 334–343

Syntetos, A., Katsikopoulos, K. (2016). Bias-Variance Trade-offs in Demand Forecasting. *Foresight: The International Journal of Applied Forecasting*, Issue 40

Tashman, L.J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16 (4), 437–450

Tavanidou, E., Nikolopoulos, K., Metaxiotis, K., Assimakopoulos, V., 2003. eTIFIS: An innovative eForecasting Web application. *International Journal of Software Engineering and Knowledge Engineering* 13 (2), 215 – 236

Thiel, H. (1966). *Applied Economic Forecasting*. Chicago: Rand McNally

Thomakos, D., Nikolopoulos, K. (2014). Fathoming the theta method for a unit root process. *Oxford Journal of Management Mathematics* 25 (1), 105 – 124

Thomakos, D., Nikolopoulos, K. (2015). Forecasting multivariate time series with the theta method

Thyer, M., Kuczera, G., Wang, Q. J. (2002). Quantifying parameter uncertainty in stochastic models using the Box–Cox transformation. *Journal of Hydrology*, 265, 246–257

Venables, W. N., Ripley, B. D. (2002). *Modern Applied Statistics with S*, Fourth edition, Springer

Wald, A. (1936). Berechnung und Ausschaltung von Saisonschwankungen. *Beitrage zur Konjunkturforschung*, volume 9, Vienna, Austria

Wand, M. P., Jones, M. C. (1995). *Kernel Smoothing Monographs on Statistics and Applied Probability*. Chapman & Hall

Zaycoff, R. (1936). *Über die Zerlegung statistischer Zeitreihen in drei Komponenten*. Publication No. 4. Sophia: State University of Sophia, Statistical Institute for Economic Research

Π1. Σε επιστημονικά περιοδικά

1. **Spiliotis, E.**, Assimakopoulos, V., Nikolopoulos, K. "Thinner, lighter, faster..." and the way forward for time series extrapolation methods: superimposing smoothing, a variation of the Theta method and shrinkage of seasonal factors, *International Journal of Production Economics -accepted (Impact Factor: 3.49)*
2. Sarris, D., **Spiliotis, E.**, Assimakopoulos, V. Exploiting resampling techniques for model selection in forecasting: An empirical evaluation using out-of-sample tests, *Operational Research: An International Journal - in print (Impact Factor: 1.07)*
3. Marinakis, V., Doukas, H., **Spiliotis, E.**, Papastamatiou, I., 2017. Decision Support for Intelligent Energy Management in Buildings Using the Thermal Comfort Model, *International Journal of Computational Intelligence Systems*, 10 (1), 882–893 (*Impact Factor: 1.57*)
4. Papastamatiou, I., Doukas, H., **Spiliotis, E.**, Psarras, J., 2016. How “OPTIMUS” is a city in terms of energy optimization? e-SCEAF: A web based decision support tool for local authorities, *Information Fusion*, 29, 149-161 (*Impact Factor: 5.67*)
5. **Spiliotis E.**, Raptis A., Legaki Z.N., Assimakopoulos V., 2014. Forecasting electrical consumption of commercial buildings using energy performance indicators. *International Journal of Decision Support Systems*, 1(2), 164–182

Π2. Σε πρακτικά επιστημονικών συνεδρίων

1. **Spiliotis, E.**, Marinakis, V., Doukas, H., Psarras, J., 2016. Integrating a Decision Support System with smart grid infrastructures and ICT solutions towards energy cost reduction. 7th International Conference on Information, Intelligence, Systems and Applications (IISA 2016), 1-6
2. Doukas, H., Marinakis, V., **Spiliotis, E.**, Psarras, J., 2016. OPTIMUS Decision Support Tools: Transforming Multidisciplinary Data to Energy Management Action Plans. 7th International Conference on Information, Intelligence, Systems and Applications (IISA 2016), 1-5
3. **Spiliotis, E.**, Anastasopoulos, G., Dede, P., Marinakis, V. Doukas, H., 2015. A framework for integrating user experience in action plan evaluation through social media: Transforming user generated content into knowledge to optimise energy use in buildings. 6th International Conference on Information, Intelligence, Systems and Applications (IISA 2015), 1-6
4. Androulaki, S., **Spiliotis, E.**, Doukas, H., Papastamatiou, I., Psarras, J., 2014. Proposing a Smart City Energy Assessment Framework: Linking local vision with data sets. 5th International Conference on Information, Intelligence, Systems and Applications (IISA 2014), 50-56

Π3. Σε βιβλία

1. Androulaki, S., Doukas, H., **Spiliotis, E.**, Papastamatiou, I., Psarras, J., 2016. A framework to assess the behavior and performance of a city towards energy optimization. Chapter in book: Intelligent Computing Systems, Edition: 2016, Series Volume 627, Chapter: 9, Publisher: Springer-Verlag Berlin Heidelberg, Editors: George A. Tsihrintzis, Maria Virvou, Lakhmi C. Jain, DOI: 10.1007/978-3-662-49179-9

Π4. Σε κλαδικά περιοδικά

1. **Spiliotis, E.**, Raptis, A., Assimakopoulos V., 2016. Off-the-Shelf vs. Customized Forecasting Support Systems. Foresight: The International Journal of Applied Forecasting, Fall 2016 (Issue 43), 42-48

Π5. Σε διεθνή επιστημονικά συνέδρια

1. **Spiliotis, E.**, Patikos, A., Assimakopoulos V., Kouloumos, A., 2017. Data as a service: Providing new datasets to the forecasting community for time series analysis. 37th International Symposium on Forecasting (ISF 2017), June 25 – 28, 2017, Cairns, Australia
2. Skepetari, I., **Spiliotis, E.**, Raptis, A., Assimakopoulos, V., 2017. Assisting judgmental adjustments in hierarchical forecasts through interactive forecasting support systems. 37th International Symposium on Forecasting (ISF 2017), June 25 – 28, 2017, Cairns, Australia
3. **Spiliotis, E.**, Assimakopoulos, V., Nikolopoulos, K., 2016. "Thinner, lighter, faster..." and the way forward for time series extrapolation methods. 19th International Symposium on Inventories, August 22-26, 2016, Hilton Budapest Hotel, Budapest, Hungary
4. **Spiliotis, E.**, Assimakopoulos V., 2016. Transforming the theta model into a flexible decomposition method by considering nonlinear trends. 36th International Symposium on Forecasting (ISF 2016), June 19 – 22, 2016, Santander, Spain
5. Skepetari, I., **Spiliotis, E.**, Raptis, A., Assimakopoulos V., 2016. OMEN: Promoting forecasting support systems. 36th International Symposium on Forecasting (ISF 2016), June 19 – 22, 2016, Santander, Spain
6. **Spiliotis, E.**, Marinakis, V., Tsapelas, G., Skondrogianni, C., Doukas, H., 2016. Reducing Energy and CO2 Emissions at the City Premises: A Decision Support System for Local Authorities. 5th International Symposium and 27th National Conference on Operational Research (HELORS 2016), June 9-11, 2016, Aigaleo, Athens
7. **Spiliotis, E.**, Raptis, A., Skepetari, E., Assimakopoulos, V., 2015. Exploiting business intelligence of water companies. ForWarD: an online water demand forecasting tool. 35th International Symposium on Forecasting (ISF 2015), June 21 – 24, 2015, Riverside, California, USA
8. **Spiliotis, E.**, Petropoulos, F., Kourentzes, N., Assimakopoulos, V., 2014. Examining the effect of temporal aggregation on forecasting accuracy for

- hierarchical energy consumption time series. 34th International Symposium on Forecasting (ISF 2014), June 29 – July 2, 2014, Rotterdam, The Netherlands
9. Paterakis, S., **Spiliotis, E.**, Assimakopoulos, V., 2014. Forecasting Electrical Demand in Commercial Buildings through Energy Performance Indicators using Time Series Methods. 20th Conference of the International Federation of Operational Research Societies EURO (IFORS), July 13-18, 2014, Barcelona, Spain
 10. **Spiliotis, E.**, Raptis, A., Legaki, N. Z., Assimakopoulos, V., 2013. A method for predicting electrical consumption in energy-intensive buildings through energy performance indicators. 2nd International Symposium and 24th National Conference on Operational Research (HELORS 2013), 26 - 28 September 2013, Athens, Greece
 11. Gkolemis, N., Legaki, N.Z., Assimakopoulos, V., **Spiliotis, E.**, 2013. Modeling electrical energy operations of buildings to forecast and achieve energy efficiency. 26th European Conference on Operational Research EURO (INFORMS) MMXIII, July 1-4, 2013, Rome, Italy
 12. **Spiliotis, E.**, Raptis, A., Legaki, N.Z., Assimakopoulos, V., 2013. Forecasting energy consumption of buildings using performance indicators. 26th European Conference on Operational Research EURO (INFORMS) MMXIII, July 1-4, 2013, Rome, Italy

Π6. Σε φοιτητικά συνέδρια:

1. Σίμος, Ν., **Σπηλιώτης, Ε.**, Ασημακόπουλος Β., 2015. Ανάπτυξη Μεθοδολογίας για τη Μείωση των Ουρών Αναμονής στα Υποκαταστήματα Τραπέζης. 4ο Φοιτητικό Συνέδριο Ελληνικής Εταιρίας Επιχειρησιακών Ερευνών - "Επιχειρησιακή Έρευνα - Ευκαιρίες και Προκλήσεις Μέσα στην Κρίση", 17-18 Δεκ. 2015
2. Φακίνος, Γ., **Σπηλιώτης, Ε.**, Ασημακόπουλος Β., 2015. Πρόβλεψη ηλεκτρικής κατανάλωσης σε σύμπλεγμα εμπορικών κτιρίων με χρήση μεθόδων χρονικής συνάθροισης. 4ο Φοιτητικό Συνέδριο Ελληνικής Εταιρίας Επιχειρησιακών Ερευνών - "Επιχειρησιακή Έρευνα - Ευκαιρίες και Προκλήσεις Μέσα στην Κρίση", 17-18 Δεκ. 2015