Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
Εθνικό Μετσόβιο Πολυτεχνείο

Διπλωματική Εργασία
Ανάλυση και Πρόβλεψη Χρονοσειρών
Time Series Analysis and Forecasting

Χαράλαμπος Σιημιλλάς 09118712
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
Τομέας Μαθηματικών

Κατεύθυνση: Στατιστική και Εφαρμοσμένη Ανάλυση
Επιβλέπων: Ιωάννης Κολέτσος, Αναπληρωτής Καθηγητής

Η τριμελής εξεταστική επιτροπή

Ι.Κολέτσος	Δ.Φουσκάκης	Π.Στεφανέας
Αναπλ.Καθηγητής	Καθηγητής	Αναπλ.Καθηγητής

Αθήνα, 2023

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή μου, κ. Ιωάννη Κολέτσο, για την υποστήριξή και την καθοδήγηση του κατά την διάρκεια της εκπόνησης αυτής της εργασίας. Επίσης ευχαριστώ ιδιαίτερα τα μέλη της εξεταστικής επιτροπής μου για τον χρόνο που αφιέρωασαν. Τέλος θέλω να ευχαριστήσω τους γονείς μου, για τη συνεχή τους υποστήριξη.

Περίληψη

Η παρούσα διπλωματική εργασία έχει ως σκοπό την μελέτη μεθόδων και εργαλείων που χρησιμοποιούνται για την παραγωγή προβλέψεων. Οι χρονοσειρές είναι η βάση αρκετών μεθόδων πρόβλεψης. Η ποσοτική και ποιοτική τους ανάλυση είναι το κλειδί για ικανά μοντέλα πρόβλεψης.

Στο εισαγωγικό κεφάλαιο γίνεται αναφορά στην επιχειρησιακή έρευνα και στη σημασία που απέκτησε ως κλάδος μέσα από μια ιστορική αναδρομή. Στην συνέχεια αναλύονται διάφορες μέθοδοι ανάλυσης και επεξεργασίας μιας χρονοσειράς, για την εξαγωγή των προτύπων που αυτή κρύβει. Ακολουθούν διάφορα μοντέλα πρόβλεψης και τεχνικές βελτίωσης της απόδοσης τους. Τέλος γίνεται η εφαρμογή των μεθόδων αυτών σε σύνολο ζήτησης ηλεκτρικής ενέργειας.

Λέξεις/φράσεις κλειδία: Ανάλυση χρονοσειρών, Μοντέλα Προβλέψεων, ARIMA, Wavelet Transform, STL

Abstract

This thesis aims to investigate the methodologies and instruments employed in the field of forecasting. Time series data forms the foundation of numerous forecasting techniques, and the thorough quantitative and qualitative analysis of these data is pivotal in developing proficient forecasting models.

The introductory chapter provides a historical overview, highlighting the significance of operations research as a discipline. It subsequently delves into diverse methodologies for analyzing and processing time series data, enabling the extraction of underlying patterns that lie within. Furthermore, the thesis explores a range of forecasting models and techniques aimed at enhancing their performance. Ultimately, these methodologies are applied to a dataset comprising electricity demand, demonstrating their practical application and effectiveness.

Kye words/phrases: Time Series, Forecasting, ARIMA, Wavelet Transform, STL

Περιεχόμενα

1 Εισαγωγή

1.1 Εισαγωγή στην Επιχειρησιακή έρευνα

1.1.1 Ιστορική Αναδρομή

Το 1940 θεωρείται ως το έτος γέννησης του επιστημονικού κλαδου της επιχειρησιακής έρευνας. Βέβαια οι ρίζες της επιχειρησιακής έρευνας μπορούν να εντοπισθούν πολλές δεκαετίες νωρίτερα, όταν έγιναν οι πρώτες προσπάθειες να χρησιμοποιηθεί μια επιστημονική προσέγγιση στη διαχείριση των οργανισμών. Πρωτοπόρος σε αυτές τις προσπαθειες ήταν ο Charles Babbage (1791- 1871) ο οποίος χαρακτηρίστηκε ως ο "πατέρας της Επιχειρησιακής έρευνας" καθώς η έρευνά του για το κόστος μεταφοράς και το κόστος ταξινόμησης της αλληλογραφίας συντέλεσε στη δημιουργία του γενικού αγγλικού "Ταχυδρομείου της πένας" το 1840. Το 1917, ο Δανός μηχανικός A.K.Erlang (1878-1929) εξέτασε προβλήματα που είχαν σχέση με το χρόνο απασχόλησης των τηλεφωνικών κέντρων. Υπήρξαν βέβαια και άλλες αξιόλογες επιστημονικές προσεγγίσεις γύρω από τη διοίκηση των επιχειρήσεων. Ωστόσο η έκρηξη του Β' Παγκοσμίου Πολέμου τον Σεπτέμβριο του 1939 συντάραξε την ανθρωπότητα και καταλυτικός παράγοντας για την έκβαση του ιστορικού αυτού γεγονότος, που καθόρισε το μέλλον των λαών, ήταν η εφαρμογή της "επιχειρησιακής έρευνας".

Κατά τον Β' Παγκόσμιο πόλεμο υπήρχε επείγουσα ανάγκη να κατανεμηθούν οι περιορισμένοι διαθέσιμοι πόροι, αρχικά στις διάφορες στρατιωτικές επιχειρήσεις, και στη συνέχεια στις δραστηριότητες κάθε επιχείρησης με αποτελεσματικό τρόπο. Ως εκ τούτου, η βρετανική και στη συνέχεια η αμερικανική στρατιωτική διοίκηση κάλεσε ένα μεγάλο αριθμό επιστημόνων να εφαρμόσουν μια **επιστημονική προσέγγιση** για την αντιμετώπιση αυτού όπως και άλλων στρατηγικών και τακτικών προβλημάτων. Στην πραγματικότητα, **τους ζητήθηκε να κάνουν έρευνα για τις (στρατιωτικές) επιχειρήσεις**. Αυτές οι ομάδες επιστημόνων ήταν οι πρώτες ομάδες Επιχειρησιακής Έρευνας. Τρεις από τις πιο χαρακτηριστικές εφαρμογές ήταν:

- (α) Η μελέτη για την καλύτερη αξιοποίηση (της νέας τότε) συσκευής ανίχνευσης και εντοπισμού εχθρικών αεροσκαφών (ραντάρ) στη Μεγάλη Βρετανία

- (β) Η μελέτη για τον προσδιορισμό του κατάλληλου βάθους εκρήξεων βομβών βυθού, που χρησιμοποιούσε η Αγγλική αεροπορία κατά των Γερμανικών υποβρυχίων

- (γ) Η μελέτη μιας ομάδας Επιχειρησιακών Ερευνών, του Υπουργείου Άμυνας των ΗΠΑ, για τον προσδιορισμό του βέλτιστου μεγέθους νηοπομπών (πλήθος φορτηγών πλοίων συνοδευόμενων από πολεμικά) μεταφοράς πολεμικού υλικού και στρατιωτών από τις ΗΠΑ στην Ευρώπη.

Μετά τον πόλεμο, η κορύφωση της βιομηχανικής επανάστασης δημιούργησε αρκετά προβλήματα λόγω της αυξανόμενης πολυπλοκότητας των οργανισμών και της εξειδίκευσης. Αυξήθηκαν οι ανάγκες παραγωγής και παροχής υπηρεσιών, υπήρχε μεγάλος ανταγωνισμός και το περιβάλλον της τεχνολογίας και της αγοράς άλλαζε συνεχώς. Συγκεκριμένα, οι εταιρείες εξαπλώνονταν σε εθνική και πολυεθνική κλίμακα καλύπτοντας όλο και περισσότερο γεωγραφικό χώρο, με αποτέλεσμα να αυξηθεί σημαντικά η κατανομή της εργασίας και ο επιμερισμός των δραστηριοτήτων μεταξύ των ατόμων που διοικούσαν. Η εισαγωγή νέων μορφών τεχνολογίας απαιτούσε εξειδικευμένους τεχνίτες, και αυτό με τη σειρά του εμφάνισε **το πρόβλημα του συντονισμού των διαφόρων τμημάτων της επιχείρησης**, για την επίτευξη κοινού στόχου ως προς το συμφέρον της

επιχείρησης. Τα προβλήματα ήταν πολλά και το ενδιαφέρον της βιομηχανίας για την επιχειρησιακή έρευνα δεν άργησε να έρθει, ύστερα και από τις προφανείς επιτυχείς εφαρμογές της στον στρατιωτικό τομέα. Έτσι η βιομηχανία παρακινημένη από την φανερή επιτυχία της επιχειρησιακής έρευνας στον στρατιωτικό τομέα , άρχισε να ενδιαφέρεται για τον νέο αυτό κλάδο. Στην Αμερική ιδρύθηκε το Operations Evaluation Group (OEG) σε συνεργασία με το MIT και υπογράφηκε σύμβαση με την Douglas Aircraft Company για το έργο RAND (Research and Development), το οποίο αφορούσε την επέκταση της χρήσης των ερευνητών επιχειρησιακής έρευνας για μεγάλο χρονικό διάστηκα μετά τη λήξη του πολέμου. Στην Μεγάλη Βρετανία το 1957 ιδρύεται η Διεθνής Συνομοσπονδία Εταιριών Επιχειρησιακών Ερευνών "IFORS". Στην Ελλάδα το 1963 , μια ομάδα πρωτοπόρων επιστημών ιδρύει την "Ελληνική Εταιρία Επιχειρησιακών Ερευνών" (Ε.Ε.Ε.Ε) ή όπως αποκαλείται διεθνώς "Hellenic Operational Research Society" (HELORS), μια επιστημονική, μη κερδοσκοπική εταιρία που είχε ως κύριο σκοπό να προαγάγει και να διαδώσει την Επιχειρησιακή Έρευνα στην Ελλάδα. Η "HELORS" αποτελεί μέλος της "IFORS" από τις αρχές του έτους 1968.

1.1.2 Τι είναι η Επιχειρησιακή Έρευνα;

Η Επιχειρησιακή Έρευνα είναι, ένα κύριο εργαλείο, το οποίο χρησιμοποιείται για την άσκηση της διοίκησης και ειδικότερα για τη λήψη αποφάσεων που αποσκοπεί στον καλύτερο δυνατό σχεδιασμό και λειτουργία ενός συστήματος, συνήθως υπό συνθήκες που απαιτούν την κατανομή σπάνιων πόρων. Με τον όρο σύστημα εννοούμε μια οργάνωση αλληλοεξαρτώμενων στοιχείων που συνεργάζονται για την επίτευξη του στόχου του συστήματος. Για παράδειγμα, η Ford Motor Company είναι ένα σύστημα του οποίου στόχος συνίσταται στη μεγιστοποίηση του κέρδους που μπορεί να αποκομιστεί από την παραγωγή ποιοτικών οχημάτων. Οι τεχνικές της επιχειρησιακής έρευνας υλοποιούνται συνήθως με την χρήση του Η/Υ. Το όνομα της προέρχεται από τη μετάφραση τον Αγγλικού όρου "Operational Research" ή όπως συνηθίζεται στην Αμερική "Operations Research" και σημαίνει "έρευνα στις επιχειρήσεις". Η απόδοση αυτή του όρου στα Ελληνικά μας λέει κάτι τόσο για την προσέγγιση όσο και για το πεδίο εφαρμογής του επιστημονικού αυτού τομέα, όμως θα μπορούσαμε να πούμε πως δεν είναι και η ακριβέστερη δυνατή. Η απόδοση "Λειτουργική Έρευνα" θα ήταν πιο επιτυχημένη, αφού αποδίδει πληρέστερα το περιεχόμενο του αντικειμένου, δηλαδή έρευνα πάνω στη λειτουργία ή στις λειτουργίες πολύπλοκων συστημάτων αποτελούμενων από ανθρώπους και μηχανές. Στον στρατό όμως, απ' όπου και ξεκίνησε η ανάπτυξη και η εφαρμογή της Επιχειρησιακής έρευνας, οι κυριότερες λειτουργίες ονομάζονται επιχειρήσεις. Επομένως, ίσως γι' αυτόν τον λόγο καθιερώθηκε αυτός ο όρος γενικότερα στην Ελλάδα. Επίσης, ο όρος "Επιχειρησιακής έρευνα" είναι πιο κατανοητός από τους ανθρώπους στον τομέα των επιχειρήσεων.

Τα τελευταία χρόνια χρησιμοποιούνται διεθνώς και οι εναλλακτικές ονομασίες Διοικητική Επιστήμη (Management Science), Επιστήμη Αποφάσεων (decision science) και Ανάλυση συστημάτων (System Analysis). Επίσης χρησιμοποιείται συχνά η συνδυαστική ονομασία Operations Research Management Science (OR/MS) για να υποδηλώσει ότι οι δύο όροι σημαίνουν περίπου το ίδιο. Τι σημαίνει όμως ο όρος Επιχειρησιακή Έρευνα; Ο καλύτερος τρόπος να απαντήσουμε αυτήν την ερώτηση, θα ήταν να δώσουμε έναν ορισμό. Κατά καιρούς έχουν προταθεί διάφοροι ορισμοί οι οποίοι δεν διαφέρουν ιδιαίτερα μεταξύ τους. Αναφέρουμε μερικούς από τους σημαντικότερους.

Ο R. Watson-Watt, ο οποίος μαζί με τον A.P.Rowe, φαίνεται ότι πρότεινε το όνομα " Operational Research" και ο οποίος ήταν ένας από τους πρωτεργάτες της εισαγωγής και ανάπτυξης αυτού του αντικειμένου στη Βρετανική αεροπορία, έδωσε τον εξής ορισμό:

> "Η Επιχειρησιακή Έρευνα αποσκοπεί στο να ερευνήσει ποσοτικά εαν ένας οργανισμός παίρνει από τη λειτουργία τον εξοπλισμού του, τη βέλτιστη δυνατή συνεισφορά σε σχέση με τον ολικό αντικειμενικό σκοπό του, ποιες αλλαγές σε εξοπλισμό και μεθόδους απαιτούνται για τη βελτίωση των αποτελεσμάτων με το μικρότερο δυνατό κόστος σε προσπάθεια και χρόνο και τέλος σε ποιο βαθμό οι μεταβολές στους επιμέρους αντικειμενικούς σκοπούς θα συνεισέφεραν στην πιο οικονομική και έγκαιρη εκτέλεση του ολικού στρατηγικού αντικειμενικού σκοπού"

από την εταιρία Επιχειρησιακής Έρευνας της Μεγάλης Βρετανίας (Operations Research Society) έχει προταθεί ο παρακάτω ορισμός:

> "Η Επιχειρησιακή Έρευνα είναι η εφαρμογή της σύγχρονης επιστήμης πάνω σε πολύπλοκα προβλήματα, τα οποία ανακύπτουν στη διεύθυνση και διοίκηση μεγάλων συστημάτων, αποτελούμενων από ανθρώπους, μηχανές, υλικά και κεφάλαια, στη βιομηχανία, τις Επιχειρήσεις, τις Κυβερνητικές Υπηρεσίες και την Άμυνα. Η χαρακτηριστική της μεθοδολογία συνίσταται στην ανάπτυξη επιστημονικού μοντέλου του υπό μελέτη συστήματος, που περιλαμβάνει μετρήσεις τυχαίων παραγόντων και με το οποίο προβλέπει και συγκρίνει τα αποτελέσματα εναλλακτικών αποφάσεων, στρατηγικών και ελέγχων. Ο σκοπός της είναι να βοηθήσει τη διοίκηση να καθορίσει την πολιτική και τις ενέργειές της επιστημονικά (κατά τον καλύτερο δυνατό τρόπο). Από τους Russell Lincoln Ackoff και Maurice W. Sasieni, όπως περιλαμβάνεται στο σύγγραμμα τους με τίτλο "Fundamentals of Operations Research": Επιχειρησιακή Έρευνα μπορεί να θεωρηθεί ότι είναι: η εφαρμογή επιστημονικών μεθόδων από μικτές ομάδες σε προβλήματα που αφορούν τον έλεγχο οργανωμένων συστημάτων (αποτελούμενων από ανθρώπους και μηχανές) κατά τρόπο, ώστε να παρέχουν λύσεις που εξυπηρετούν κατά τον καλύτερο δυνατό τρόπο τους σκοπούς του οργανισμού ως συνόλου".

Από την E.E.E.E. προτάθηκε ο ιδιαίτερα εύστοχος ορισμός:

> "Επιχειρησιακή Έρευνα είναι η επιστημονική προετοιμασία των αποφάσεων της διοικήσεως (με την επιστημονική ανάλυση των δεδομένων και τη δημιουργία μαθηματικών προτύπων)". Από τους Hans G. Daellenbach και John A. George το 1978: "Επιχειρησιακή έρευνα είναι η συστηματική εφαρμογή ποσοτικών μεθόδων, τεχνικών και εργαλείων στην ανάλυση προβλημάτων που εμπεριέχουν την λειτουργία συστημάτων".

1.1.3 Βασικά στάδια της Επιχειρησιακής Έρευνας

Όπως έχουμε δει οι μελέτες της Επιχειρησιακής Έρευνας αποσκοπούν στην εξεύρεση λύσεων για την λήψη αποφάσεων. Οι λύσεις αυτές είναι κατα κανόνα "βέλτιστες" ως προς τα δεδομένα του προβλήματος και ως προς κάποια μετρήσιμα κριτήρια (π.χ ελάχιστο κόστος, μέγιστη ευστάθεια, μέγιστο όρια αντοχής κ.τ.λ). Στον πραγματικό κόσμο όμως, συχνά τα δεδομένα του προβλήματος είναι ελλιπή και οι βέλτιστες αποφάσεις λαμβάνονται υπο συνθήκες αβεβαιότητας. Για τις περιπτώσεις αυτές χρησιμοποιούμε την μεθοδολογία των Πιθανοτήτων και της Στατιστικής, με τους αντίστοιχους δείκτες αξιοπιστίας. Η επιστημονική μεθοδολογία που ακολουθείται στην Επιχειρησιακή Έρευνα είναι σχεδόν πάντα η ίδια ανεξάρτητα από το πεδίο εφαρμογής ή το μοντέλο που θα χρησιμοποιηθεί. Τα στάδια μιας μελέτης επιχειρησικής έρευνας είναι τα εξής:

Στάδια μιας μελέτης επιχειρησιακής έρευνας:

Ανάλυση συστήματος

- Προσδιορίζουμε τη δομή και τον τρόπο λειτουργίας του συστήματος με σκοπό την κατανόηση του.
- Για να αποκτήσουμε σαφή αντίληψη του συστήματος, το αναλύουμε στα υποσυστήματα του, και για κάθε υποσύστημα προσδιορίζουμε στρατηγικές με τις οποίες μπορούμε να επηρεάσουμε το υπο μελέτη σύστημα.
- Αναγνωρίζουμε τις μεταβλητές-παραμέτρους του συστήματος, και τους διάφορους περιορισμούς που επιβάλει η δομή, η λειτουργία και το περιβάλλον του.

Διατύπωση στόχων

- Προσδιορίζουμε τους στόχους που θέλουμε να πετύχουμε. Ως προς αυτούς θα καθοριστεί η βέλτιστη λύση.
- Είναι καθοριστικό στάδιο καθώς η διατύπωση των στόχων θα καθορίσει και την επιτυχία και την αποτελεσματικότητα των λύσεων που θα προταθούν.
- Εαν υπάρχουν πολλαπλοί στόχοι πρέπει να ιεραρχηθούν. Εαν η επίτευξη ενός στόχου ακυρώνει κάποιον άλλο τότε χρειάζεται είτε να περιορίσουμε τους στόχους μας, είτε να διαφοροποιηθούν κάποιοι ώστε να επιτευχθεί στο τέλος η καλύτερη δυνατή λειτουργία του συστήματος.

Διατύπωση του μοντέλου (Μαθηματική μοντελοποίηση)

- Αναπαριστούμε το πραγματικό σύστημα όσο πιο απλά γίνεται μέσω ενός μαθηματικού μοντέλου.
- Το μοντέλο αποτελείται από μαθηματικές σχέσεις, που εκφράζουν τους στόχους του προβλήματος και τους περιορισμούς του περιβάλλοντος.

Επίλυση του μοντέλου

- Επίλυση του μοντέλου με χρήση διάφορων τεχνικών που στηρίζονται σε ανώτερα μαθηματικά (γραμμική άλγεβρα, βελτιστοποίηση, αριθμητική ανάλυση), την θεωρία πιθανοτήτων και την Στατιστική (ανάλυση παλινδρόμησης, παραγοντική ανάλυση, εκτίμηση παραμέτρων) καθώς και σε μεθόδους της επιχειρησιακής (γραμμικός προγραμματισμός, δένδρα αποφάσεων, θεωρία παιγνίων).
- Ο καθορισμός της/των μεθόδου/ων επίλυσης εξαρτάται από τον τύπο και την πολυπλοκότητα του μαθηματικού μοντέλου.

Ανάλυση ευαισθησίας(Sensitivity Analysis)

Αξιολόγηση της προτεινόμενης λύσης του μοντέλου μέσω του ελέγχου ευαισθησίας (μικρές μεταβολές στις τιμές των παραμέτρων επηρεάζει την λύση).

Υλοποίηση της λύσης

Η επιλογή της στρατηγικής που θα ακολουθήσουμε δεν είναι το τελευταίο στάδιο μιας μελέτης επιχειρησιακής έρευνας. Το τελευταίο και δυσκολότερο στάδιο, είναι αυτό της εφαρμογής της στρατηγικής που επιλέξαμε. Τα αποτελέσματα της στρατηγικής θα πρέπει να μετατραπούν σε λειτουργικές οδηγίες, παρουσιασμένες με κατανοητό τρόπο στα άτομα που θα διαχειριστούν το προτεινόμενο σύστημα, έτσι ώστε η βελτίωση που επιτεύχθηκε να υλοποιηθεί στο πραγματικό σύστημα .

1.2 Εισαγωγή στην Πρόβλεψη

Η επιτυχία κάθε επιχειρηματικού οργανισμού είναι άμεσα συνυφασμένη με μελλοντικά γεγονότα και αυτό καθιστά τις προβλέψεις ζωτικής σημασίας. Εαν υπήρχε ένας μαγικός τρόπος να γνωρίζαμε ότι είναι να συμβεί πριν αυτό συμβεί, τότε θα μπορούσε η κάθε επιχείρηση, και γενικότερα ο κάθε άνθρωπος, να παίρνει τις ιδανικότερες αποφάσεις σήμερα για ζητήματα που αφορούν το μέλλον. Για να μπορέσει ένας οργανισμός να θωρακιστεί έναντι των μελλοντικών εκβάσεων διάφορων παραγόντων, που επηρεάζουν την λειτουργία του, θα πρέπει σήμερα να **προβλέψει** όσο το δυνατό καλύτερα το "μέλλον" τους. Προφανώς και μια πρόβλεψη ενός "συστήματος" δεν μπορεί πάντοτε να συμφωνεί με την πραγματική έκβαση της τιμής του. Επομένως πάντοτε η πρόβλεψη συνοδεύεται από αβεβαιότητα, την οποία επιθυμούμε να την περιορίζουμε όσο το δυνατό περισσότερο, διαφορετικά η επιστημονικά διατυπωμένη πρόβλεψη μας δεν θα διαφέρει σε τίποτα από μια τυχαία μαντεψιά.

Προφανώς δεν μπορούμε να προβλέψουμε μελλοντικές τιμές/εκβάσεις όλων των "γεγονότων" με την ίδια ακρίβεια. Το πόσο καλά μπορούμε να προβλέψουμε την μελλοντική τιμή (predictability) μιας ποσότητας καθορίζεται κυρίως από τους εξής παράγοντες:

- Πόσο καλά μπορούμε να κατανοήσουμε και να ελέγξουμε τους παράγοντες με τους οποίους σχετίζεται.
- Πόσα **ποιοτικά** δεδομένα έχουμε στην διάθεση μας.
- Κατά πόσο οι προβλέψεις των ειδικών, μπορούν να επηρεάσουν το αντικείμενο του οποίου την μελλοντική τιμή προσπαθούμε να προβλέψουμε.

Για παράδειγμα η πρόβλεψη της ζήτησης ηλεκτρικής ενέργειας μπορεί να επιτευχθεί με μεγάλη ακρίβεια καθώς και οι 3 πιο πάνω συνθήκες ικανοποιούνται. Οι παράγοντες που επηρεάζουν (π.χ θερμοκρασία) την ζήτηση ελέγχονται εύκολα και επιπλέον υπάρχουν αρκετά ιστορικά δεδομένα για την ζήτηση. Επίσης μια πρόβλεψη της ζήτησης που θα προβληθεί στα μέσα μαζικής ενημέρωσης δεν θα επηρεάσει την επιθυμία του κάθε πολίτη για χρήση ηλεκτρικού ρεύματος. Αντίθετα όταν προσπαθούμε να προβλέψουμε τις τιμές ισοτιμίας συναλλάγματος μόνο μία από τις 3 συνθήκες ικανοποιείται. Ναι μέν υπάρχει αφθονία διαθέσιμων ιστορικών δεδομένων αλλά οι παράγοντες που επηρεάζουν την μελλοντική τους τιμή είναι περίπλοκοι. Επιπλέον εαν υπάρχουν ευρέως δημοσιευμένες προβλέψεις ότι η συναλλαγματική ισοτιμία θα αυξηθεί, οι αγοραστές θα προσαρμόσουν άμεσα την τιμή που είναι διατεθειμένοι να πληρώσουν και με αυτό τον τρόπο θα έχουμε αυτοεκπληρούμενες προβλέψεις. Κατά μία έννοια, οι συναλλαγματικές ισοτιμίες μετατρέπονται στις ίδιες τις προβλέψεις. Αυτό είναι ένα παράδειγμα της "υπόθεσης της αποτελεσματικής αγοράς" ("efficient market hypothesis").

Το περιβάλλον της πρόβλεψης σπανίως είναι αμετάβλητο:

Κάθε περιβάλλον [1] αλλάζει, και ένα καλό μοντέλο πρόβλεψης αντιλαμβάνεται τον τρόπο με τον οποίο τα πράγματα αλλάζουν. Για αυτό το λόγο οι "forecasters" δεν υποθέτουν ότι το περιβάλλον είναι αμετάβλητο, αλλά υποθέτουν ότι ο τρόπος με τον οποίο αυτό αλλάζει θα παραμένει σταθερός. Επομένως ένα μοντέλο πρόβλεψης πρέπει να αντιλαμβάνεται τόσο πού βρίσκεται το καθετι, όσο και προς τα πού κινείται. Αυτή ήταν και η άποψη του Abraham Lincoln ο οποίος τόνισε "Εαν μπορούσαμε να ορίσουμε πρώτα που βρισκόμαστε και

[1] με τον όρο περιβάλλον θα εννοούμε το γενικό περιβάλλον με το οποίο αλληλεπιδρά το αντικείμενο που προβλέπουμε

προς τα πού οδεύουμε, τότε θα κρίναμε καλύτερα το τι θα έπρεπε να κάνουμε στο μέλλον και πώς να το κάνουμε."

1.2.1 Μερικές εφαρμογές των μοντέλων προβλέψεων:

Πιο κάτω παρουσιάζονται ορισμένοι τομείς στους οποίους οι προβλέψεις χρησιμοποιούνται ευρέως.

Πρόβλεψη πωλήσεων (Sales Forecasting):

Κάθε εταιρεία που πωλεί αγαθά, πρέπει να προβλέπει όσο το δυνατό καλύτερα τη ζήτηση για τα αγαθά αυτά. Οι κατασκευαστές πρέπει να γνωρίζουν πόσο να παράγουν. Οι χονδρέμποροι και οι λιανοπωλητές πρέπει να γνωρίζουν πόσα πρέπει να αποθηκεύσουν. Η ουσιαστική υποεκτίμηση της ζήτησης κατα πάσα πιθανότητα θα οδηγήσει σε πολλές χαμένες πωλήσεις (κόστος ευκαιρίας-Opportunity Loss), δυσαρεστημένους πελάτες και ίσως να επιτρέψει στον ανταγωνισμό να αποκτήσει το πάνω χέρι στην αγορά. από την άλλη πλευρά, η σημαντική υπερεκτίμηση της ζήτησης είναι επίσης πολύ δαπανηρή λόγω του υψηλού κόστους αποθήκευσης, των αναγκαστικών μειώσεων στις τιμές και χαμένες ευκαιρίες για την εμπορία πιο κερδοφόρων αγαθών. Οι επιτυχημένοι διευθυντές μάρκετινγκ και παραγωγής κατανοούν πολύ καλά την σημασία της καλής πρόβλεψης των πωλήσεων.

Πρόβλεψη των αναγκών σε πρώτες ύλες και ανταλλακτικά:

Παρόλο που η αποτελεσματική πρόβλεψη των πωλήσεων αποτελεί κλειδί για σχεδόν κάθε εταιρεία, ορισμένοι οργανισμοί πρέπει να βασίζονται σε άλλους τύπους προβλέψεων. Ένα χαρακτηριστικό παράδειγμα αφορά τις προβλέψεις για τις ανάγκες σε πρώτες ύλες και ανταλλακτικά. Πολλές εταιρείες πρέπει να διατηρούν ένα απόθεμα ανταλλακτικών για να μπορούν να επισκευάζουν γρήγορα, είτε τον δικό τους εξοπλισμό είτε τα προϊόντα τους που πωλούνται ή μισθώνονται στους πελάτες. Για παράδειγμα οι συνέπειες για μια αεροπορική εταιρία να μήν έχει ένα ανταλλακτικό διαθέσιμο σε κάποιο συγκεκριμένο αεροδρόμιο, το οποίο απαιτείται για να συνεχίσει το αεροπλάνο να κάνει τα προγραμματισμένα δρομολόγια, είναι τουλάχιστο μία ακυρωση πτήσης.

Πρόβλεψη οικονομικών τάσεων (Forecasting economic trends):

Τι είναι η πρόβλεψη για το ποσοστό του πληθωρισμού; Το ποσοστό ανεργίας; Το εμπορικό ισοζύγιο; Τα στατιστικά μοντέλα για την πρόβλεψη των οικονομικών τάσεων ονομάζονται **οικονομετρικά μοντέλα**. Χρησιμοποιώντας ιστορικά δεδομένα για να κάνουν προβλέψεις για το μέλλον, αυτά τα οικονομετρικά μοντέλα συνήθως εξετάζουν ένα πολύ μεγάλο αριθμό παραγόντων που συμβάλλουν στην ανάπτυξη της οικονομίας. Ορισμένα μοντέλα περιλαμβάνουν εκατοντάδες μεταβλητές και εξισώσεις. Ωστόσο, εκτός από το μέγεθος και το πεδίο εφαρμογής τους, τα μοντέλα αυτά μοιάζουν με κάποιες από τις στατιστικές μεθόδους πρόβλεψης που χρησιμοποιούν οι επιχειρήσεις για την πρόβλεψη πωλήσεων. Επιπλέον αυτά τα οικονομετρικά μοντέλα μπορούν να έχουν μεγάλη επιρροή στον καθορισμό των κυβερνητικών πολιτικών.

Πρόβλεψη αναγκών σε προσωπικό (Staffing needs):

Μία από τις σημαντικότερες τάσεις στις σύγχρονες μεγάλες οικονομίες είναι η μετατόπιση της έμφασης από τη μεταποίηση στις υπηρεσίες. Όλο και περισσότερα από τα βιομηχανικά προϊόντα παράγονται εκτός της χώρας (όπου η εργασία είναι φθηνότερη) και στη συνέχεια εισάγονται. Ταυτόχρονα, ένας αυξανόμενος αριθμός επιχειρήσεων ειδικεύεται στην παροχή κάποιου είδους υπηρεσίας (π.χ,ταξίδια, τουρισμός, ψυχαγωγία, νομική βοήθεια, υπηρεσίες υγείας, χρηματοοικονομικές, εκπαιδευτικές, σχεδιασμός, συντήρηση κ.λπ.). Για μια τέ-

τοια επιχείρηση, η πρόβλεψη των "πωλήσεων" μετατρέπεται σε πρόβλεψη της ζήτησης υπηρεσιών, η οποία στη συνέχεια μεταφράζεται σε πρόβλεψη των αναγκών σε προσωπικό για την παροχή αυτών των υπηρεσιών.

1.2.2 Τα βασικά βήματα της διαδικασίας πρόβλεψης

Η διαδικασία της **πρόβλεψης** είναι αρκετά σύνθετη και αποτελείται από αρκετά στάδια. Τα γενικά βήματα της διαδικασίας πρόβλεψης έχουν ως εξής:

Προσδιορισμός του προβλήματος:
Συχνά είναι το δυσκολότερο κομμάτι της διαδικασίας πρόβλεψης. Ο ορθός και προσεκτικός ορισμός του προβλήματος απαιτεί πλήρη κατανόηση του τρόπου με τον οποίο η πρόβλεψη θα χρησιμοποιηθεί, ποιος είναι ο πελάτης και πώς αυτή επηρεάζει τον οργανισμό που την απαιτεί. Ο "προγνώστης" πρέπει να αφιερώσει αρκετό χρόνο για συζήτηση με όσους θα εμπλακούν στην διαδικασία πρόβλεψης (αυτούς που θα συλλέξουν τα δεδομένα και αυτούς που θα χρησιμοποιήσουν τα μοντέλα πρόβλεψης για μελλοντικούς σκοπούς).

Επιλογή του χρονικού ορίζοντα της πρόβλεψης:
Ένα πετυχημένο μοντελο πρόβλεψης είναι απαραίτητο να σχηματιστεί με βάση τον κατάλληλο χρονικό ορίζοντα. Δηλαδή πρέπει να καθοριστεί **πόσο παλιά μπορούν να είναι τα δεδομένα μας και μέχρι πόσο μπροστά στον χρόνο μπορεί να προβλέψει το μοντέλο.**

Συλλογή Πληροφορίας (Gathering Information):
Σε κάθε περίπτωση υπάρχουν δύο είδη πληροφορίας που απαιτούνται.

1. Στατιστικά δεδομένα
2. Η γνώση και η άποψη των εμπειρογνωμόνων (που θα συλλέξουν τα δεδομένα) και του πελάτη (που θα χρησιμοποιήσει την πρόβλεψη).

Συχνά είναι δύσκολο να συλλέξουμε επαρκή ιστορικά δεδομένα έτσι ώστε να είμαστε σε θέση να προσαρμόσουμε ένα καλό στατιστικό μοντέλο. Σε αυτή την περίπτωση οι "προβλέψεις με κρίση" μπορούν και πρέπει να χρησιμοποιηθούν είτε αυτόνομες είτε ως συμπληρωματικές. Επίσης σε κάποιες περιπτώσεις τα ιστορικά δεδομένα θα είναι λιγότερο χρήσιμα, καθώς το σύστημα που θα προβλέψουμε δεν διατηρεί το ίδιο μοντέλο εξέλιξης με το πέρασμα του χρόνου ή έχει αλλάξει αρκετά σε σχέση με αυτό του παρελθόντος. Σε αυτές τις περιπτώσεις τα πιο προσφατα ιστορικά δεδομένα θα είναι χρησιμότερα. Βέβαια χρειάζεται να είμαστε προσεκτικοί στο ποια δεδομένα θα αφαιρέσουμε, καθώς για να μπορεί το μοντέλο να προσαρμόζεται με ορθό τρόπο σε μελλοντικές αλλαγές, θα πρέπει να εκπαιδευτεί σε κατάλληλο σύνολο δεδομένων, στο οποίο να εμφανίζονται τα σωστά πρότυπα. Το πόσο πίσω στον χρόνο θα μπορούν να "πηγαίνουν" τα δεδομένα καθορίζεται από τις προϋποθέσεις του μοντέλου που επιλέξαμε.

Προκαταρκτική (διερευνητική) ανάλυση:
Το πρώτο βήμα της διαδικασίας, αφού συλλέξουμε δεδομένα, είναι η γραφική αναπαράσταση τούς. Από τις γραφικές αυτές, ελέγχουμε κατα πόσο υπάρχουν συστηματικά μοτίβα (patterns), τάση (Trend), σημάδια εποχικότητας (Seasonality). Επιπλέον εξετάζουμε αν υπάρχουν ενδείξεις επιχειρηματικών κύκλων (business cycles). Τέλος εντοπίζουμε τα δεδομένα (αν υπάρχουν) που αποκλίνουν πολύ από τα υπόλοιπα (Outliers), που

πρέπει να ερμηνευτούν από άτομα με την κατάλληλη γνώση;

Επιλογή και προσαρμογή μοντέλου πρόβλεψης που θα χρησιμοποιηθεί:
Το "καλύτερο" μοντέλο που θα χρησιμοποιηθεί εξαρτάται από τα διαθέσιμα ιστορικά δεδομένα, τη δύναμη των δεσμών μεταξύ της μεταβλητής πρόβλεψης και των επεξηγηματικών μεταβλητών, και το τρόπο με τον οποίο η πρόβλεψη θα χρησιμοποιηθεί. Πολύ συχνά χρησιμοποιούμε περισσότερα από ένα μοντέλο για την πρόβλεψη, συγκρίνοντας τα αποτελέσματα τους. Για τον σκοπό αυτό, απαιτείται η γνώση των διαφόρων μοντέλων πρόβλεψης, σε ποιες καταστάσεις αυτά είναι εφαρμόσιμα, πόσο αξιόπιστο είναι το καθένα από αυτά και τι είδους δεδομένα απαιτούνται για το καθένα. Σε αρκετές περιπτώσεις μπορεί να επιλεγούν περισσότερα από ένα μοντέλα, και η τελική πρόβλεψη να επιλεγεί με βάση κάποιο κατάλληλο κριτήριο.

Χρήση και αξιολόγηση του μοντέλου πρόβλεψης
(Using and evaluating a forecasting model):
Η απόδοση ενός μοντέλου μπορεί να υπολογιστεί επαρκώς όταν τα δεδομένα για την περίοδο πρόβλεψης είναι διαθέσιμα. Αρκετές μέθοδοι έχουν αναπτυχθεί για να βοηθήσουν στην αξιολόγηση της ακρίβειας των προβλέψεων. Στην πράξη όμως, για να μπορέσουμε να αξιολογήσουμε ένα μοντέλο ακολουθείται μια ιδιαίτερη τακτική. Το σύνολο των δεδομένων χωρίζεται σε δύο υποσύνολα, το σύνολο εκπαίδευσης (training set) και το σύνολο ελέγχου (test set). Το μοντέλο πρόβλεψης "εκπαιδεύετε" με βάση το σύνολο εκπαίδευσης και στην συνέχεια αξιολογείται με βάση την απόδοση του στο σύνολο ελέγχου. Έτσι με αυτό τον τρόπο μπορούμε να γνωρίζουμε πόσο καλά **αναμένουμε** να αποδώσει το μοντέλο, που θα επιλέξουμε, στα μελλοντικά άγνωστα δεδομένα.

Επιπλέον η πρόβλεψη που λαμβάνεται μέσω οποιουδήποτε μοντέλου δεν θα πρέπει να χρησιμοποιείται, ως έχει, στα τυφλά. Θα πρέπει να αξιολογείται από την άποψη του διαστήματος εμπιστοσύνης. Συνήθως όλα τα καλά μοντέλα πρόβλεψης διαθέτουν μεθόδους υπολογισμού της ανώτερης και της κατώτερης τιμής εντός των οποίων αναμένεται να παραμείνει η συγκεκριμένη πρόβλεψη, με προκαθορισμένο επίπεδο σημαντικότητας. Μπορεί επίσης να αξιολογηθεί από λογική άποψη, κατά πόσον η τιμή που λαμβάνεται είναι λογικά εφικτή; Μπορεί επίσης να αξιολογηθεί σε σχέση με κάποια σχετική μεταβλητή ή φαινόμενο. Έτσι, είναι δυνατόν, και μερικές φορές σκόπιμο να τροποποιηθεί η στατιστικώς προβλεπόμενη τιμή με βάση την αξιολόγηση.

1.2.3 Η Στατιστική φύση των προβλέψεων

Τα πράγματα που προσπαθούμε να προβλέψουμε είναι άγνωστα, επομένως μπορούμε να τα θεωρήσουμε ως τυχαίες μεταβλητές. Για παράδειγμα οι πωλήσεις του επόμενου μήνα μπορούν να είναι σε ένα εύρος πιθανών τιμών, και μέχρι την χρονική στιγμή της ωρίμανσης [2] η τιμή τους δεν είναι γνωστή . Στο παράδειγμα η χρονική στιγμη ωρίμανση είναι το τέλος της τελευταίας μέρας του μήνα της πρόβλεψης μας. Επομένως μέχρι να μάθουμε ακριβώς το ύψος των πωλήσεων του επόμενου μήνα, το ύψος των πωλήσεων είναι μια τυχαία ποσότητα.

Επειδή ο επόμενος μήνας είναι σχετικά κοντά, συνήθως μπορούμε να έχουμε μια αρκετά καλή διαίσθηση για το ύψος των μηνιαίων πωλήσεων. Από την άλλη πλευρά, εαν προβλέπουμε τις μηνιαίες πωλήσεις για τους επόμενους τρείς μήνες, οι πιθανές τιμές της τυχαίας μεταβλητής που εκφράζει τις πωλήσεις του τρίτου στην σειρά επόμενου μήνα, είναι πολύ περισσότερες και υπάρχει πολύ μεγαλύτερη διακύμανση μεταξύ των πιθανών τιμών που μπορεί να πάρει. Στις πλείστες περιπτώσεις προβλέψεων, η διακύμανση που σχετίζεται με το "αντικείμενο" πρόβλεψης μειώνεται όσο η χρονική στιγμή ωρίμανσης είναι πλησιέστερη στο παρόν. Με άλλα λόγια, όσο πιο μελλοντική πρόβλεψη κάνουμε τόσο περισσότερη αβεβαιότητα έχουμε για την ορθότητα της.

Όταν προβλέπουμε την εξέλιξη ενός "αντικειμένου" ουσιαστικά προβλέπουμε την μελλοντική του εξέλιξη και τη μελλοντική του τιμή τη χρονική στιγμή ωρίμανσης. Μπορούμε λοιπόν να φανταστούμε διάφορες πιθανές μελλοντικές εκβάσεις στο περιβάλλον της πρόβλεψης και να τις απεικονίσουμε σε ένα γράφημα. Τι μας δίνουν στην πραγματικότητα τα στατιστικά μοντέλα προβλέψεων; Αυτο που μας δίνουν είναι η αναμενόμενη τιμή της τυχαίας μεταβλητής που αντιπροσωπεύει το αντικείμενο πρόβλεψης. Συνήθως μια πρόβλεψη συνδυάζεται με ένα "διάστημα εμπιστοσύνης πρόβλεψης" το οποίο δίνει ένα εύρος τιμών που μπορεί να πάρει η τ.μ, με μια σχετικά (προκαθορισμένη) μεγάλη πιθανότητα. Όταν λοιπόν κάποιος θέλει να του προβλέψουμε κάτι για μια μελλοντική στιγμή, συνήθως του λέμε μια σημειακή εκτίμηση και ένα διάστημα εμπιστοσύνης στο οποίο με μεγάλη πιθανότητα θα ανήκει η πραγματική τιμή του αντικείμενου πρόβλεψης τη στιγμή της ωρίμανσης.

Η κατάλληλη μέθοδος πρόβλεψης εξαρτάται σε μεγάλο βαθμό από τα διαθέσιμα δεδομένα. Εαν δέν υπάρχουν διαθέσιμα ιστορικά δεδομένα τότε επιβάλλεται η αξιοποίηση ποιοτικών μεθόδων πρόβλεψης (qualitative forecasting) ενώ όταν υπάρχουν διαθέσιμα δεδομένα υιοθετούνται συνήθως ποσοτικές μέθοδοι (Quantitative Forecasting). Βέβαια πέραν της ύπαρξης ιστορικών δεδομένων θα πρέπει να είναι λογικό να θεωρήσουμε ότι κάποιες πτυχές του παρελθόντος θα συνεχίσουν να ισχύουν και στό μέλλον έτσι ώστε τα ιστορικά δεδομένα να συμβάλουν σε ένα καλό μοντέλο πρόβλεψης.

[2] Χρονική στιγμή ωρίμανσης πρόβλεψης: η χρονική στιγμή στην οποία προβλέπουμε την τιμή, της υπο μελέτης μεταβλητής

2 Δεδομένα Χρονοσειρών (Time series data)

Οι πλείστες ποσοτικές μέθοδοι, τις οποίες μελετάμε στην παρούσα διπλωματική εργασία, χρησιμοποιούν δεδομένα χρονοσειρών (time series data). Επομένως είναι εύλογο να εξηγήσουμε τι είναι τα δεδομένα χρονοσειρών και πώς πρέπει να αξιοποιούνται.

Ορισμός Χρονοσειράς δεδομένων:
Χρονοσειρά δεδομένων ονομάζεται μια ακολουθία από ιστορικές τιμές της υπό εξέτασης μεταβλητής, που λαμβάνονται με τη πάροδο του χρόνου και καταγράφονται σε τακτά χρονικά διαστήματα (είτε με ομοιόμορφο τρόπο είτε όχι). Με άλλα λόγια, πρόκειται για δεδομένα που συλλέγονται διαδοχικά, με κάθε παρατήρηση να συμβαίνει σε μια συγκεκριμένη χρονική στιγμή.

Οι χρονοσειρές χρησιμοποιούνται συνήθως σε τομείς όπως η οικονομία, τα χρηματοοικονομικά και η πρόγνωση του καιρού, καθώς μια χρονοσειρά μπορεί να παρέχει πληροφορίες για φαινόμενα όπως τάσεις, μοτίβα και εποχιακές διακυμάνσεις που αρκετά συχνά χαρακτηρίζουν τα ιστορικά δεδομένα.

Μαθηματικά, μια χρονοσειρά ορίζεται από τις τιμές $Y_1, Y_2, ...Y_n$ της μεταβλητής "στόχου"[3] Y κατά τις χρονικές στιγμές $t_1, t_2, ..., t_n$. Επομένως η μεταβλητή Y είναι μια συνάρτηση του χρόνου t , και συμβολιζεται με $Y = F(t)$.

Οι πιο απλές μέθοδοι πρόβλεψης χρονοσειρών (Time series forecasting methods) χρησιμοποιούν μόνο την πληροφορία που πηγάζει από την υπο μελέτη μεταβλητή και δέν επιχειρούν να εντοπίσουν τους παράγοντες που επηρεάζουν την συμπεριφορά της. Ως εκ'τούτου στηρίζονται στα μοτίβα (patterns) που εντοπίζουν στη χρονοσειρά και αγνοούν όλες τις υπόλοιπες δυνατές πληροφορίες που θα μπορούσαν να αξιοποιηθούν (π.χ μελετώντας τη σύνδεση της μεταβλητής στόχου με άλλες μεταβλητές).

Οι πιο διαδεδομένες μέθοδοι πρόβλεψης χρονοσειρών είναι τα μοντέλα αποσύνθεσης (decomposition models), μοντέλα εκθετικής εξομάλυνσης (exponential smoothing models) και τα μοντέλα Arima.

2.1 Γραφική αναπαρασταση δεδομένων χρονοσειράς

Γενικότερα στην ανάλυση δεδομένων πρώτο μέλημα μας είναι η γραφική αναπαράσταση των δεδομένων. Η οπτικοποίηση των δεδομένων, επιτρέπει να αναδυθούν τα ποιοτικά τους χαρακτηριστικά συμπεριλαμβανομένων των μοτίβων, των ασυνήθιστων παρατηρήσεων και των συσχετίσεων μεταξύ μεταβλητών. Τα χαρακτηριστικά που εντοπίζονται μέσω των γραφικών αναπαραστάσεων θα πρέπει να ενσωματώνονται όσο το δυνατό αποτελεσματικότερα στις μεθόδους προβλέψεων που θα χρησιμοποιηθούν. Ανάλογα με τη μορφή που έχουν τα δεδομένα μας, αποφασίζουμε ποιες γραφικές αναπαραστάσεις θα μας βοηθήσουν στην εξερεύνηση τους.

Η γραφική παράσταση της συνάρτησης $Y = F(t)$, παρουσιάζει την εξέλιξη της μεταβλητής Y στο χρόνο. Ένας ενδιαφέρον τρόπος να την αντιλαμβανόμαστε είναι σαν τη κίνηση ενός σημείου καθώς κυλάει ο χρόνος. Η γραφική αναπαράσταση της 'ιστορίας' της μεταβλητής Y αποκαλύπτει κάποιες μορφές χαρακτηριστικών κινήσεων

[3]Μεταβλητή στόχος: Έτσι θα αναφερόμαστε αρκετά συχνά στην μεταβλητή που μελετάμε ή για τη μεταβλητή όπου θέλουμε να κάνουμε πρόβλεψη

(characteristic movements), που εμφανίζονται συνηθέστερα σε κάποιο βαθμό. Η μετέπειτα ανάλυση και μελέτη αυτών των κινήσεων αποτελεί βασικό στάδιο για να επιτευχθεί καλή πρόβλεψη των μελλοντικών τιμών της χρονοσειράς. Ενδεικτικά πιο κάτω φαίνεται η εξέλιξη της μετοχής της Apple από τον Ιανουάριο του 2021 εώς και τις 18/03/2023.

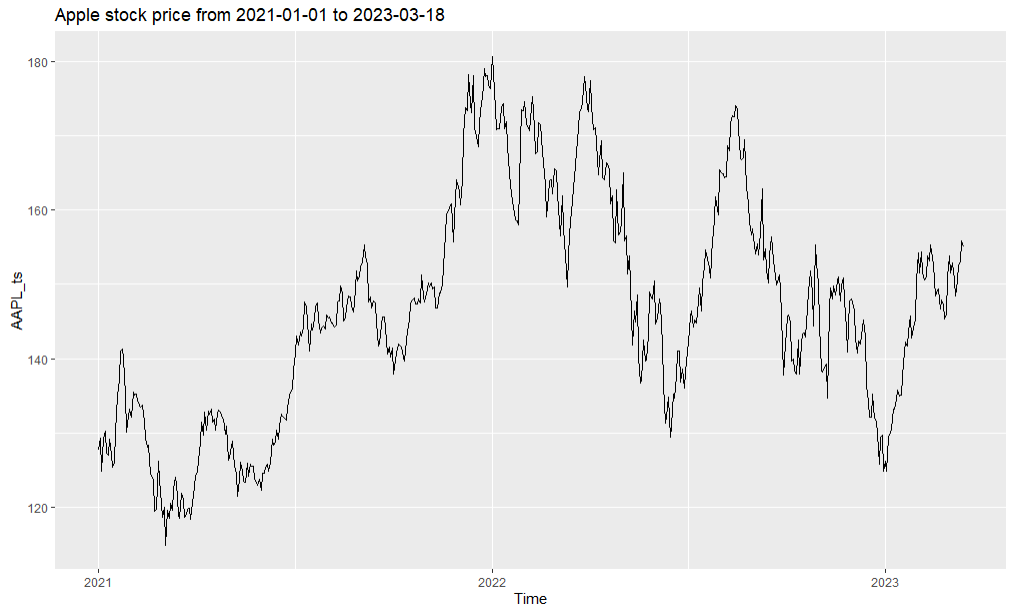


Figure 2.1.1: Η τιμή της μετοχή της εταιρίας Apple κατα τη περίοδο 01/01/2021 - 18/03/2023

2.2 Βασικά Χαρακτηριστικά χρονοσειρών

Όπως αναφέραμε και πιο πριν, μέσω της γραφικής αναπαράστασης της χρονοσειράς μπορούν να φανούν κάποια από τα ποιοτικά χαρακτηριστικά της, όπως η τάση, η εποχικότητα, η κυκλικότητα και η τυχαιότητα. Αυτά τα χαρακτηριστικά τα ονομάζουμε πρότυπα χρονοσειρών και τα αναλύουμε πιο κάτω.

Τάση (Trend):

Στην χρονοσειρά λέμε ότι υπάρχει τάση όταν υπάρχει μια μεγάλη περίοδος συνεχόμενων αυξήσεων ή μειώσεων στις τιμές της μεταβλητής που μελετάμε. Η αύξηση/μείωση δεν πρέπει να είναι γραμμική. Κάποιες φόρες δύναται να εκτιμηθεί από διάφορες οικογένειες καμπυλών, όπως μια ευθεία γραμμή, ένα πολυώνυμο ανώτερης τάξης ή μια εκθετική καμπύλη. Για να γίνει έλεγχος για το αν μια σειρά παρουσιάζει τάση θα πρέπει να υπάρχει επαρκής αριθμός παρατηρήσεων και να εκτιμηθεί σε ένα κατάλληλο χρονικό διάστημα. Μερικές φορές θα αναφερόμαστε σε μια τάση των δεδομένων ως την “μεταβολή της κατεύθυνσης”, όταν αυτή αλλάζει από αύξουσα σε φθίνουσα.

Εποχικότητα (Seasonality):

Η εποχικότητα σε μια χρονοσειρά ορίζεται το φαινόμενο κατα το οποίο εμφανίζονται περιοδικές διακυμάνσεις, μικρής χρονικής διάρκειας, οι οποίες οφείλονται σε εποχιακούς παράγοντες (seasonality Factors). Εποχιακός παράγοντας μπορεί να είναι η “μια μέρα της εβδομάδας”, “μια εποχή του χρόνου”, “τα Χριστούγεννα”. Για παράδειγμα οι μηνιαίες πωλήσεις ενός καταστήματος με παιδικά παιχνίδια κατά τη διάρκεια ενός έτους (που προφανώς μπορούν να αναπαρασταθούν ως μια χρονοσειρά) θα εμφανίσουν ιδιαίτερη αύξηση κατα του μήνες

Δεκεμβρίου και Ιανουαρίου λόγω των εορτών των Χριστουγέννων και της Πρωτοχρονιάς και αυτό το "ημερολογιακό γεγονός" αποτελεί τον εποχιακό παράγοντα που επηρεάζει τη διακύμανση των τιμών της χρονοσειράς.

Κυκλικότητα (Cyclic):

Ένας"κύκλος" εμφανίζεται σε μια χρονοσειρά όταν εμφανίζονται αυξομειώσεις χωρίς σταθερή συχνότητα. Αυτές οι διακυμάνσεις οφείλονται συνήθως σε οικονομικές και τεχνολογικές αλλαγές. Επιπλέον η κυκλικότητα σχετίζεται άμεσα με τους επιχειρηματικούς κύκλους (business cycles). Ο επιχειρηματικός κύκλος αναφέρεται στις περιόδους επέκτασης (ανάπτυξη της οικονομίας) που ακολουθούνται από περιόδους συρρίκνωσης (μείωση οικονομικών δραστηριοτήτων). Αρκετά συχνά η κυκλικότητα μπερδεύετε με την εποχικότητα οδηγώντας σε λανθασμένη διαχείριση της χρονοσειράς. Όταν η συχνότητα εμφάνισης των διακυμάνσεων είναι σταθερή και κατα το χρονικό διάστημα εμφάνισης τους υπάρχει κάποιος "εποχιακός-ημερολογιακός παράγοντας" που προκαλεί αυτές τις διακυμάνσεις τότε η χρονοσειρά εμφανίζει εποχικότητα. Αντίθετα όταν η συχνότητα των διακυμάνσεων δεν είναι σταθερή τότε εμφανίζεται η κυκλικότητα. Ακόμα μια σημαντική διαφορά των δύο παρατηρείται στην χρονική διάρκεια και την ισχύ των διακυμάνσεων η οποία στο φαινόμενο της κυκλικότητας είναι συνήθως μεγαλύτερη.

Μη κανονικές διακυμάνσεις-Τυχαιότητα:

Οι μη κανονικές διακυμάνσεις είναι κύριο χαρακτηριστικό των περισσότερων χρονοσειρών και αντιμετωπίζεται αρκετά δύσκολα. Συνήθως, αυτές οι διακυμάνσεις αντιπροσωπεύουν την επιρροή μιας στοχαστικής διαδικασίας στην εξέλιξη του υπό μελέτη μεγέθους ή κάποια ασυνέχεια που συνδέεται με κάποιο εξαίρετο γεγονός. Ακριβώς λόγω της στοχαστικής φύσης της εμφάνισης και της μεταβολής των παραμέτρων που προκαλούν αυτές τις διακυμάνσεις, εν γένει θεωρούνται ως εκείνες που απομένουν όταν η τάση, η κυκλικότητα και η εποχικότητα έχουν απορριφθεί για την επεξήγηση μιας διακύμανσης η της μορφής της χρονοσειράς. Τέτοιες διακυμάνσεις μπορεί να οφείλονται σε μια μεγάλη ποικιλία παραγόντων, όπως ξαφνικές καιρικές μεταβολές, απεργία ή κοινωνικές αναταράξεις. Η επίδραση αυτών των γεγονότων μπορεί να εξαλειφθεί με την εξομάλυνση των δεδομένων της χρονοσειράς.

Τα πιο πάνω "μοτίβα" που μπορεί να εμφανίσει μια χρονοσειρά αποτελούν καταλυτικούς παράγοντες στην επιλογή ενός ικανού μοντέλου πρόβλεψης δεδομένων χρονοσειρών. Ένα ικανό μοντέλο θα πρέπει να αντιλαμβάνεται και να εκμεταλλεύεται όλη τη πληροφορία που κρύβουν τα πιο πάνω μοτίβα.

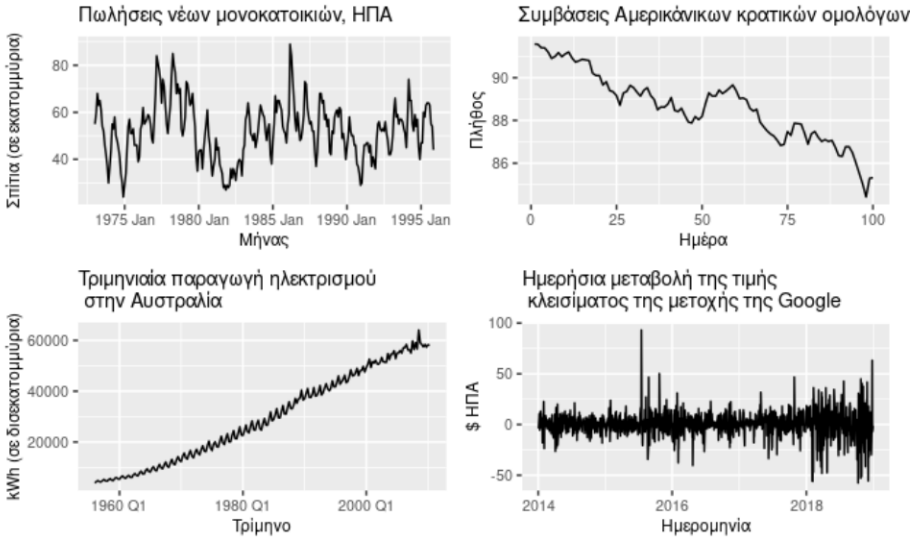


Figure 2.2.1: Τέσσερα παραδείγματα χρονοσειρών που παρουσιάζουν τα 4 μοτίβα που αναλύσαμε

1. Οι μηνιαίες πωλήσεις (πάνω αριστερά) παρουσιάζουν ισχυρή εποχικότητα σε κάθε χρόνο αφού οι πωλήσεις τείνουν να είναι μεγαλύτερες σε συγκεκριμένους μήνες κάθε χρόνο.

2. Οι συμβάσεις Αμερικάνικων κρατικών ομολόγων (πάνω δεξιά) δείχνουν αποτελέσματα από την αγορά της πόλης του Σικάγο για 100 συνεχόμενες ημέρες διαπραγμάτευσης το 1981. Εδώ δε βλέπουμε κάποια εποχικότητα, αλλά μια προφανή πτωτική τάση. Ενδεχομένως, αν είχαμε μια πολύ μακρύτερη χρονοσειρά στην διάθεση μας, θα βλέπαμε ότι αυτή η πτωτική τάση είναι στην πραγματικότητα μέρος ενός μεγαλύτερου κύκλου, αλλά όταν το εξετάζουμε μόνο σε 100 ημέρες, φαίνεται να είναι μια τάση.

3. Η τριμηνιαία παραγωγή ενέργειας στην Αυστραλία παρουσιάζει έντονη αυξητική τάση.

4. Τέλος στην κάτω δεξιά γραφική βλεπουμε ότι η ημερήσια αλλαγή στην τιμή με την οποία κλείνει η μετοχή της Google δέν παρουσιάζει ούτε τάση, ούτε εποχικότητα ούτε κυκλικότητα. Χαρακτηρίζεται από τυχαίες διακυμάνσεις και επομένως δύσκολα προβλέπονται.

2.3 Στασιμότητα χρονοσειράς

Στάσιμη Χρονοσειρά (Stationary Time series) : Μια χρονοσειρά καλείται στάσιμη εαν ικανοποιεί τις πιο κάτω προϋποθέσεις:

$$\mathbb{E}[y_t] = \mathbb{E}[y_{t-1}] = \mu$$
$$\mathbb{V}(y_t) = \sigma < \infty$$
$$Cov(y_t, y_{t-k}) = c_k < \infty, \forall t$$

Επομένως γενικά μια στάσιμη χρονοσειρά είναι εκείνη της οποίας οι στατιστικές ιδιότητες δεν εξαρτώνται από τη χρονική στιγμή κατά την οποία παρατηρείται η σειρά. Επομένως οι χρονοσειρές με τάση και εποχικότητα δεν είναι στάσιμες ενώ η χρονοσειρά λευκού θορύβου είναι στάσιμη καθώς το πότε θα την παρατηρήσουμε δεν έχει σημασία ως προς τα στατιστικά μεγέθη που θα δούμε. Μια χρονοσειρά με κυκλικότητα αλλά χωρίς εποχικότητα και τάση, είναι στάσιμη καθώς παρόλο που οι συμπεριφορές επαναλαμβάνονται σε κύκλους, οι κύκλοι δέν έχουν σταθερό μήκος, οπότε προτού παρατηρήσουμε τη σειρά δεν μπορούμε να είμαστε σίγουροι πού θα είναι οι κορυφές και τα κοιλώματα των κύκλων. Επομένως σε γενικές γραμμές, μια στάσιμη χρονοσειρά δεν θα έχει, μακροπρόθεσμα, προβλέψιμα πρότυπα. Έτσι το χρονοδιάγραμμα της στάσιμης χρονοσειράς αναμένουμε ότι θα είναι περίπου οριζόντιο (και ίσως μια κυκλική συμπεριφορά), με σταθερή διακύμανση. Αρκετές χρονοσειρές δεν ικανοποιούν και τις 3 πιο πάνω υποθέσεις γεγονός που τις καθιστά μή στάσιμες χρονοσειρές.

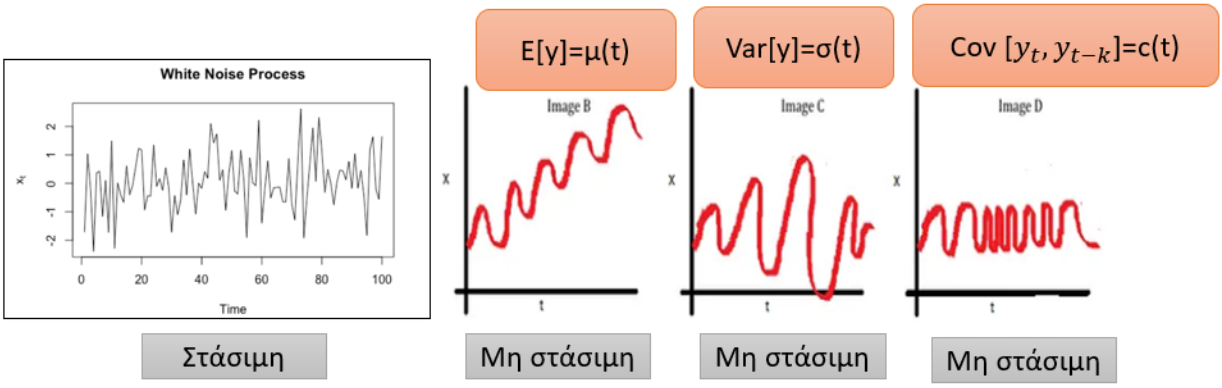


Figure 2.3.1: Παραδείγματα στάσιμων και μή στάσιμων χρονοσειρών

2.4 Γραφήματα Εποχικότητας (Seasonal plots)

Ένα Διάγραμμα εποχικότητας (Seasonal plot) είναι παρόμοιο με ένα χρονοδιάγραμμα (time plot) με τη διαφορά ότι τα δεδομένα της υπο μελέτης μεταβλητής απεικονίζονται ξεχωριστά για κάθε "χρονιά" στην οποία παρατηρήθηκαν. Πιο κάτω φαίνεται ένα ενδεικτικό παράδειγμα.

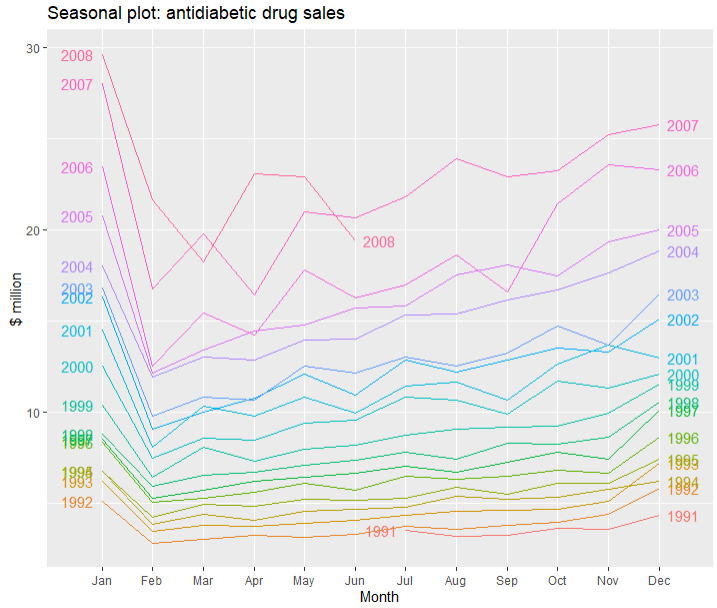


Figure 2.4.1: Αναπαράσταση των πωλήσεων φαρμάκων κατά του διαβήτη ξεχωριστά για κάθε "χρονιά"

Ένα εποχιακό γράφημα επιτρέπει στο υποβόσκων εποχιακό μοτίβο να μελετηθεί αναλυτικότερα και είναι ιδιαίτερα χρήσιμο στον εντοπισμό της εποχής και της "χρονιάς" κατα την οποία η συμπεριφορά των δεδομένων αλλάζει. Για παράδειγμα από το γράφημα (2.4.1), είναι σαφές ότι υπάρχει μια απότομη αύξηση των πωλήσεων τον Ιανουάριο κάθε έτους. Στην πραγματικότητα, πρόκειται πιθανότατα για πωλήσεις στα τέλη Δεκεμβρίου, καθώς οι πελάτες αγοράζουν και αποθηκεύουν τα φάρμακα πριν από το τέλος κάθε ημερολογιακού έτους, αλλά οι πωλήσεις αυτές δεν καταγράφονται από την κυβέρνηση παρά μόνο μία ή δύο εβδομάδες αργότερα. Το γράφημα δείχνει επίσης ότι υπήρξε και ένας ασυνήθιστα μικρός αριθμός πωλήσεων τον Μάρτιο του 2008 (οι περισσότερες από τις άλλες χρονιές δείχνουν αύξηση μεταξύ Φεβρουαρίου και Μαρτίου). Ο μικρός αριθμός πωλήσεων τον Ιούνιο του 2008 οφείλεται πιθανώς στην ελλιπή μέτρηση των πωλήσεων κατά τη στιγμή της συλλογής των δεδομένων.

Γραφήματα υποσειρών εποχικότητας (Seasonal subseries plots):

Ένα εναλλακτικό γράφημα που εστιάζει στην μελέτη των εποχιακών προτύπων είναι οι εποχιακές υποσειρές εποχικότητας, στις οποίες τα δεδομένα για κάθε εποχή προβάλλονται σε ξεχωριστό γράφημα. Είναι ιδιαίτερα χρήσιμα στον εντοπισμό αλλαγών εντός συγκεκριμένων εποχών.

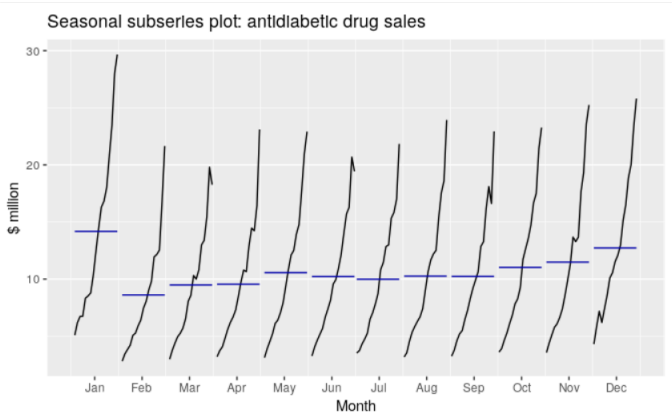


Figure 2.4.2: Γραφήματα υποσειρών εποχικότητας για τις μηνιαίες πωλήσεις φαρμάκων κατά του διαβήτη στην Αυστραλία

Οι οριζόντιες μπλε γραμμές αναπαριστούν τις μέσες τιμές για την αντίστοιχη εποχή. Είναι εξαιρετικά χρήσιμα γραφήματα για την μελέτη της συμπεριφοράς των δεδομένων στα πλαίσια μιας συγκεκριμένης εποχής με το πέρασμα του χρόνου.

2.5 Διαγράμματα Διασποράς

Αρκετά συχνά είναι χρήσιμο να μελετήσουμε τη σύνδεση της μεταβλητής που θα προβλέψουμε με άλλες μεταβλητές (επεξηγηματικές μεταβλητές). Στα αρχικά στάδια της μελέτης αυτής αναπαριστούμε τις τιμές της μεταβλητής-στόχου συναρτήσει των τιμών της επεξηγηματικής μεταβλητής. Αυτή η γραφική παράσταση μας επιτρέπει να οπτικοποιήσουμε την σχέση που συνδέει τις δύο μεταβλητές.

2.6 Γραμμική συσχέτιση μεταβλητών

Αρκετά συχνά υπολογίζουμε το δειγματικό συντελεστή **γραμμικής** συσχέτισης (κατά Pearson) δύο μεταβλητών ο οποίος εκφράζει τον "βαθμό" στον οποίο μπορούμε να εκτιμήσουμε γραμμικά την μία τ.μ όταν γνωρίζουμε την τιμή της άλλης. Η γραμμική συσχέτιση μεταξύ των μεταβλητών x και y δίνεται από:

$$r = \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum(x_t - \bar{x})^2}\sqrt{\sum(y_t - \bar{y})^2}}, \qquad r \in [-1, 1]$$

Αρνητική τιμή του r εκφράζει αρνητική γραμμική συσχέτιση, και όσο πιο κοντά στην τιμή -1 είναι τόσο ισχυρότερη είναι. Αντίθετα θετική τιμή εκφράζει θετική γραμμική συσχέτιση των 2 μεταβλητών. Επειδή ο δείκτης γραμμικής συσχέτισης ποσοτικοποιεί μόνο τη γραμμική συσχέτιση 2 μεταβλητών, αρκετές φορές μπορεί να οδηγήσει σε λανθασμένες εντυπώσεις για την συσχέτιση 2 μεταβλητών. Σε αρκετές περιπτώσεις ο δείκτης μπορεί να λάβει τιμές κοντά στο 1 ή -1 αλλά μια πολυωνυμική συσχέτιση να είναι πιό λογική απ'ότι η γραμμική μεταξύ των μεταβλητών. Για παράδειγμα στην πιο κάτω γραφική παράσταση βλέπουμε ότι ο συντελεστής συσχέτισης λαμβάνει την τιμή $r = 0.82$ αλλά το διάγραμμα διασποράς συνηγορεί ότι η υπόθεση πολυωνυμικής συσχέτισης των μεταβλητών είναι πιο λογική.

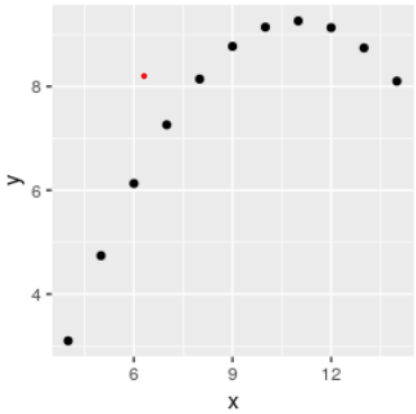


Figure 2.6.1: Διάγραμμα διασποράς 2 μεταβλητών με συντελεστή συσχέτισης $r = 0.82$

2.7 Διαγράμματα υστέρησης και αυτοσυχέτιση (Lag plots and Autocorrelation)

Τα διαγράμματα υστέρησης (lag plots) χρησιμοποιούνται όταν θέλουμε να μελετήσουμε τη σχέση μεταξύ της τιμής μιας μεταβλητής με την χρονικά-υστερημένη τιμή της (lagged value: η τιμή της ίδιας μεταβλητής με μια χρονική καθυστέρηση). Επομένως ένα lag plot είναι ένα διάγραμμα διασποράς των τιμών μιας μεταβλητής συναρτήσει των χρονικά-υστερημένων τιμών της. Στη πιο κάτω γραφική παρουσιάζουμε ένα ενδεικτικό παράδειγμα.

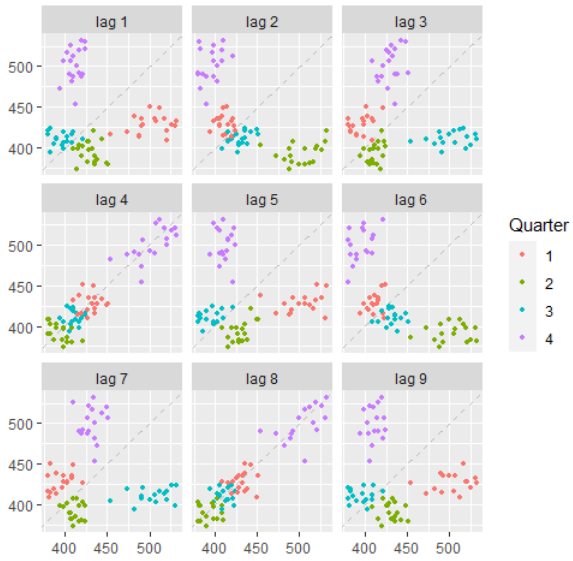


Figure 2.7.1: Lagged Scatterplots για τριμηνιαία παραγωγή μπύρας

- Κάθε ένα από τα 9 γραφήματα απεικονίζει την μεταβλητή y_t συναρτήσει της y_{t-k} για διαφορετικές τιμές του k (k=m για το "Lag plot m")

- Σε κάθε γράφημα οι παρατηρήσεις διαφορετικών "εποχών" (τρίμηνα) παρουσιάζονται με διαφορετικό χρώμα.

- Η συσχέτιση είναι ισχυρά θετική σε χρονική καθυστέρηση 4 περιόδων (lag plot 4) και 8 περιόδων (lag plot 8) αντανακλώντας την ισχυρή εποχικότητα, περιόδου 4, που υπάρχει στα δεδομένα.

- Η αρνητική συσχέτιση που παρατηρούμε στα lag plot 2 και lag plot 6 προκαλείται λόγω του ότι οι κορυφές του $4^{ου}$ τριμήνου αναπαρίστανται συναρτήσει των κοιλάδων στο $2^{ο}$ τρίμηνο.

Γενικά σε ένα διάγραμμα υστέρησης συνήθως προκύπτουν τα εξής σημαντικά μοτίβα:

- Διαγώνιος γραμμή: Όταν παρατηρούμε ότι τα δεδομένα τείνουν να απλωθούν σε διαγώνιο γραμμή τότε αυτό σημαίνει ότι τα δεδομένα είναι θετικά αυτοσυσχετιζόμενα (autocorrelated), δηλαδή οι υψηλές/χαμηλές τιμές (της y_{t-k}) τείνουν να ακολουθούνται από υψηλές/χαμηλές τιμές (της y_t).

- Καμπύλη γραμμή: Σε αυτή την περίπτωση οι υψηλές τιμές της μεταβλητής y_{t-k} τείνουν να ακολουθούνται από χαμηλές τιμές (της y_t) ενώ οι χαμηλές τιμές ακολουθούνται από υψηλές. Αυτο εκφράζει την αρνητική αυτοσυχέτιση των δεδομένων της μεταβλητής.

- Συστάδες σημείων (Cluster points): Όταν τα σημεία εμφανίζονται κατα συστάδες χωρίς να τα συνδέει κάποιος συστηματικό τρόπος τότε πολύ πιθανό να υπάρχουν ομάδες από κυκλικά επαναλαμβανόμενες τιμές που επανεμφανίζονται με τα πέρασμα του χρόνου.

Αυτοσυχέτιση (Autocorrelation): Ακριβώς όπως η συσχέτιση μετρά την έκταση μιας γραμμικής σχέσης μεταξύ δύο μεταβλητών, η αυτοσυχέτιση μετρά τη γραμμική σχέση μεταξύ υστερημένων τιμών μιας χρονοσειράς. Δηλαδή ποσοτικοποιεί τη **γραμμική** συσχέτιση μεταξύ χρονικά-υστερημένων τιμών (lagged values) μιας χρονοσειράς. Υπάρχουν αρκετοί συντελεστές αυτοσυχέτισης οι οποίοι αντιστοιχούν σε κάθε ένα τμήμα του διαγράμματος υστέρησης. Επειδή τόσο η μέση τιμή όσο και η διασπορά των μεταβλητών που μελετάμε, είναι συνήθως άγνωστοι, χρησιμοποιούμε τους δειγματικούς συντελεστές αυτοσυσχέτισης r_k. Για παράδειγμα, ο r_1 μετράει τη συσχέτιση μεταξύ y_t και y_{t-1}, ο r_2 μετράει τη συσχέτιση μεταξύ y_t και y_{t-2}, και ούτω καθεξής.

Η τιμή του r_k μπορεί να γραφτεί ως:

$$r_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}, \qquad \text{T: Το μήκος της χρονοσειράς}$$

Βέβαια, οι συντελεστές αυτοί δίνουν ασφαλή συμπεράσματα για στάσιμες χρονοσειρές, αφού για να προκύψει αξιοποιείται η ιδιότητα της ομοσκεδαστικότητας και της σταθερής μέσης τιμής. Οι συντελεστές αυτοσυχέτισης σχηματίζουν τη συνάρτηση αυτοσυχέτισης ACF.

Για την εύκολη μελέτη των συντελεστών r_k για διάφορες τιμές του k κατασκευάζεται το ακόλουθο γράφημα που είναι γνωστό ως ACF γράφημα (ή κορελόγραμμα-correlogram) το οποίο απεικονίζει τη συνάρτηση αυτοσυχέτισης.

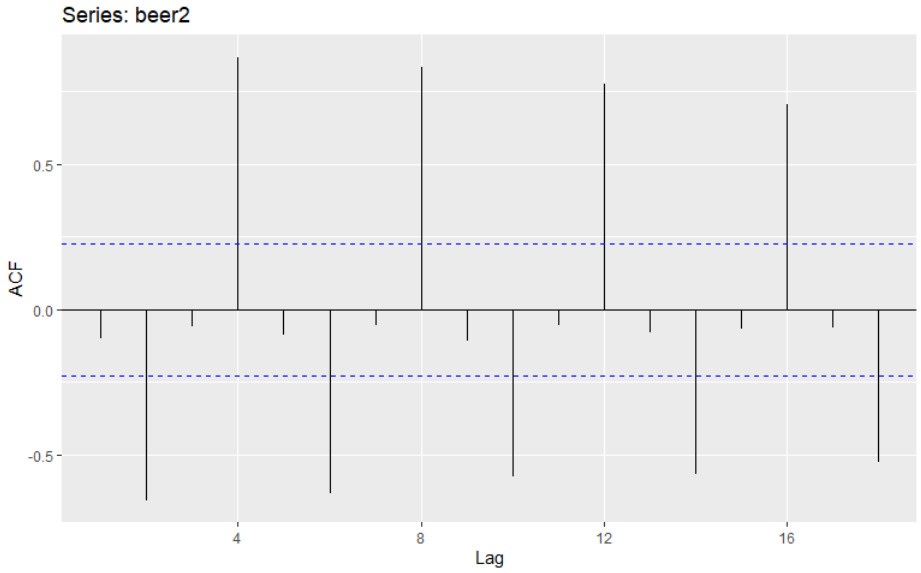


Figure 2.7.2: Συνάρτηση αυτοσυχέτισης για τη τριμηνιαία παραγωγή μπύρας

Επεξήγηση του γραφήματος ACF:

1. Οι διακεκομμένες μπλε γραμμές υποδεικνύουν κατα πόσο η γραμμική αυτοσυχέτιση είναι στατιστικά σημαντικά διαφορετική από το 0.

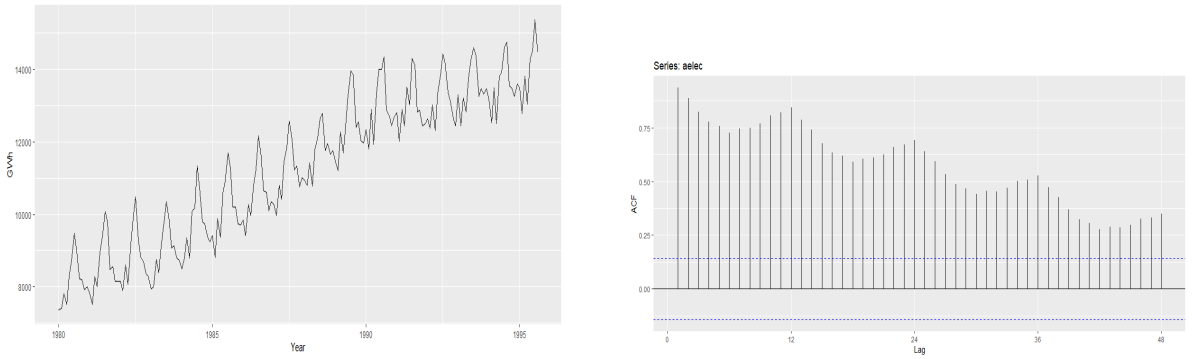
2. Παρατηρούμε ότι το r_4 έχει την υψηλότερη θετική τιμή από όλα τα υπόλοιπα. Αυτό οφείλεται στο εποχιακό μοτίβο στα δεδομένα όπου: οι κορυφές, κοιλάδες τείνουν να εμφανίζονται κάθε 4 τρίμηνα. Δηλαδή κάθε τιμή της μεταβλητής σχετίζεται μέσω γραμμικής σχέσης με την κατά 4-χρονικές στιγμές προηγούμενη της.

3. Ο δείκτης r_2 έχει την πιο αρνητική τιμή υποδεικνύοντας ότι οι κοιλάδες τείνουν να απέχουν 2 τετράμηνα από τις κορυφές.

Όταν μια χρονοσειρά χαρακτηρίζεται από τάση, οι αυτοσυχέτισης για μικρές χρονικές υστερήσεις τείνουν να είναι θετικές και μεγάλες καθώς οι παρατηρήσεις που απέχουν ελάχιστα χρονικά απέχουν και ελάχιστα σε τιμή. Επιπλέον η ACF μιας χρονοσειράς με τάση, συνήθως έχει θετικές τιμές οι οποίες εξασθενούν σε ένταση όσο αυξάνονται οι χρονικές υστερήσεις.

Όταν μια χρονοσειρά χαρακτηρίζεται από εποχικότητα τότε οι αυτοσυχετίσεις θα είναι μεγαλύτερες για τις εποχιακές υστερήσεις [4]. Στην περίπτωση όπου η χρονοσειρά χαρακτηρίζεται τόσο από τάση όσο και από εποχικότητα το ACF γράφημα παρουσιάζει ένα συνδυασμό των προαναφερθέντων φαινομένων.

[4]Εποχιακές υστερήσεις (seasonal lags): Εκφράζουν τις χρονικές-υστερήσεις που αντιστοιχούν σε εποχιακά μοτίβα, και είναι πολλαπλάσια της εποχιακής συχνότητας. π.χ αν μετράμε τις μηνιαίες πωλήσεις παιδικών παιχνιδιών οι εποχιακές χρονικές υστερήσεις θα αναφέρονται σε 12 χρονικές στιγμές όση και η εποχιακή περίοδος μηνιαίων πωλήσεων



(a) Χρονοδιάγραμμα χρονοσειράς με θετική τάση και εποχικότητα

(b) Συνάρτηση αυτοσυχέτισης χρονοσειράς με θετική τάση και εποχικότητα

Figure 2.7.3: Χρονοσειρά με θετική τάση και εποχικότητα

Παρατήρηση για την γραφική 2.7.3:
Η σταδιακή πτώση της ACF οφείλεται στην τάση ενώ οι διακυμάνσεις οφείλονται στην εποχικότητα.

Μια ιδιαίτερα σημαντική διαδικασία η οποία αξιοποιείται στην μοντελοποίηση αρκετών μοντέλων πρόβλεψης που θα αναφέρουμε παρακάτω είναι η **διαδικασία λευκού θορύβου (White noise)**. Μια χρονοσειρά καλείται λευκός θόρυβος αν δεν παρουσιάζει αυτοσυχετίσεις, δηλαδή κάθε παρατήρηση του είναι ανεξάρτητη από τις υπόλοιπες. Επίσης μια διαδικασία λευκού θορύβου χαρακτηρίζεται απο πεπερασμένη μέση τιμή και διασπορά. Το διακριτό ανάλογο στην διαδικασία λευκού θορύβου είναι η ακουλουθία τ.μ $\Upsilon_1, \Upsilon_2, ..., \Upsilon_n$ ανεξάρτητων με την ίδια κατανομή (iid random variables).

3 Ανάλυση Χρονοσειρών

Μια χρονοσειρά μπορεί να "κρύβει" αρκετά πρότυπα τα οποία για να γίνουν διακριτά χρειάζεται πρώτα να επεξεργαστούμε κατάλληλα τα δεδομένα της χρονοσειράς. Μια σημαντική διαδικασία είναι η ανάλυση της χρονοσειράς σε συνιστώσες, καθεμία απο τις οποίες να αναπαριστά μια υποκείμενη κατηγορία προτύπων.

Όταν αναλύουμε μια χρονοσειρά σε συνιστώσες, συνήθως συνδυάζουμε την τάση και την κυκλικότητα σε μια συνιστώσα τάσης-κυκλικότητας (συχνά για λόγους απλότητας καλείται απλά τάση). Έτσι, μπορούμε να θεωρήσουμε ότι μια χρονοσειρά αποτελείται από τρεις συνιστώσες: Μια συνιστώσα τάσης-κυκλικότητας, μια συνιστώσα εποχικότητας και μια συνιστώσα υπολοίπων (που περιγράφει οτιδήποτε άλλο σε μια χρονοσειρά). Για κάποιες χρονοσειρές (π.χ. για αυτές που παρατηρούνται τουλάχιστον ημερησίως), μπορεί να υπάρχουν περισσότερες από μια εποχιακές συνιστώσες, που αντιστοιχούν σε διαφορετικές εποχιακές περιόδους. Συνήθως μια τέτοια προσέγγιση βοηθά στην κατανόηση των χρονοσειρών, αλλά μπορεί, επίσης, να χρησιμοποιηθεί και για τη βελτίωση της ακρίβειας των προβλέψεων.

Αρκετές φορές, προτού γίνει η ανάλυση της χρονοσειράς σε συνιστώσες είναι απαραίτητο πρώτα να γίνει κάποιος μετασχηματισμός των δεδομένων της έτσι ώστε η ανάλυση της σε συνιστώσες, να γίνει πιο απλή και αποδοτική. Για αυτό το λόγο προτού αναφερθούμε στην ανάλυση συνιστωσών θα μιλήσουμε για μετασχηματισμούς και προσαρμογές.

3.1 Μετασχηματισμοί και προσαρμογές

Η προσαρμογή των ιστορικών δεδομένων μπορεί συχνά να οδηγήσει σε απλούστερες χρονοσειρές. Θα εξετάσουμε εδώ τέσσερα είδη προσαρμογών: τις ημερολογιακές προσαρμογές, τις πληθυσμιακές προσαρμογές, τις πληθωριστικές προσαρμογές και τους μαθηματικούς μετασχηματισμούς.

3.1.1 Ημερολογιακές προσαρμογές

Ορισμένες από τις διακυμάνσεις που εμφανίζονται σε εποχιακά δεδομένα ενδέχεται να οφείλονται σε απλά ημερολογιακά φαινόμενα. Σε τέτοιες περιπτώσεις, είναι συνήθως πολύ πιο εύκολο να αφαιρεθούν τέτοιες διακυμάνσεις πριν την οποιαδήποτε ανάλυση.
Για παράδειγμα, εαν μελετάτε τις συνολικές μηνιαίες πωλήσεις σε ένα κατάστημα λιανικής, θα παρατηρήσετε διακυμάνσεις σε αυτές, που οφείλονται στο διαφορετικό πλήθος των ημερών που πραγματοποιούνται οι συναλλαγές κάθε μήνα ή και των εποχιακών διακυμάνσεων κατά τη διάρκεια του έτους. Είναι εύκολο να αφαιρέσετε αυτήν τη διακύμανση υπολογίζοντας τις μέσες πωλήσεις ανά ημέρα συναλλαγής για κάθε μήνα και όχι τις συνολικές πωλήσεις του μήνα. Με αυτό τον τρόπο αφαιρούμε αποδοτικά τις ημερολογιακές διακυμάνσεις.

3.1.2 Πληθυσμιακές προσαρμογές

Τα δεδομένα που επηρεάζονται από πληθυσμιακές αλλαγές μπορούν να προσαρμοστούν ώστε να παρέχουν κατά κεφαλήν δεδομένα. Με άλλα λόγια προτιμήστε να λάβετε υπόψη τα δεδομένα ανά άτομο (ή ανά χίλια άτομα ή ανά εκατομμύριο άτομα) και όχι το άθροισμά τους.

Για παράδειγμα, εαν μελετάτε το πλήθος των νοσοκομειακών κλινών σε μια συγκεκριμένη περιοχή με την πάροδο του χρόνου, είναι πολύ πιο εύκολο να ερμηνευθούν τα αποτελέσματα εαν καταργήσετε τις επιπτώσεις των

πληθυσμιακών μεταβολών, λαμβάνοντας υπόψη το πλήθος των κρεβατιών ανά χίλια άτομα. Μπορείτε να δείτε, τότε, αν υπήρξαν πραγματικές αυξήσεις στον αριθμό των κλινών ή αν οι αυξήσεις οφείλονταν εξ ολοκλήρου σε πληθυσμιακές αυξήσεις. Είναι πιθανό να αυξηθεί ο συνολικός αριθμός κλινών, αλλά να μειωθεί ο αριθμός κλινών ανά χίλια άτομα.

3.1.3 Πληθωριστικές προσαρμογές

Τα δεδομένα που επηρεάζονται από την αξία του χρήματος θα ήταν προτιμότερο να προσαρμοστούν καλύτερα πριν από τη μοντελοποίηση τους. Για παράδειγμα, το μέσο κόστος μιας νεόδμητης κατοικίας θα έχει αυξηθεί τις τελευταίες δεκαετίες λόγω πληθωρισμού. Η σημερινή αξία μιας κατοικίας της τάξης των 200000$ δεν είναι η ίδια με αυτή των 200000$ για το ίδιο σπίτι πριν από είκοσι χρόνια. Για το λόγο αυτό, οι οικονομικές χρονοσειρές συνήθως προσαρμόζονται έτσι ώστε όλες οι τιμές να αναφέρονται στην τιμή του δολαρίου σε ένα συγκεκριμένο έτος.
Για να πραγματοποιηθούν αυτές τις προσαρμογές, πρέπει να χρησιμοποιηθεί ένας δείκτης τιμών. Οι δείκτες τιμών κατασκευάζονται συχνά από κυβερνητικές υπηρεσίες. Για καταναλωτικά αγαθά, ένας κοινός δείκτης τιμών είναι ο Δείκτης Τιμών Καταναλωτή (ή ΔΤΚ). Εαν z_t δηλώνει τον δείκτη τιμών και y_t δηλώνει την αρχική τιμή μιας κατοικίας το έτος t , τότε $x_t = \frac{y_t}{z_t} * z_{2000}$ δίνει την προσαρμοσμένη τιμή κατοικίας σε δολάρια του έτους 2000 .

3.1.4 Μαθηματικοί μετασχηματισμοί

Εαν τα δεδομένα παρουσιάζουν διακυμάνσεις που αυξάνονται ή μειώνονται ανάλογα με το επίπεδο της σειράς, τότε ένας μετασχηματισμός που συχνά είναι χρήσιμος είναι ο λογαριθμικός μετασχηματισμός. Αν συμβολίσουμε τις αρχικές παρατηρήσεις ως $y_1, ..., y_T$ και τις μετασχηματισμένες ως $w_1, ..., w_T$ τότε $w_t = log(y_t)$. Οι λογάριθμοι είναι χρήσιμοι επειδή είναι ερμηνεύσιμοι: οι μεταβολές σε μια λογαριθμική τιμή είναι σχετικές (ή ποσοστιαίες) μεταβολές της αρχικής κλίμακας.
Μια χρήσιμη οικογένεια μετασχηματισμών, που περιλαμβάνει τόσο λογάριθμους όσο και εκθετικούς μετασχηματισμούς, είναι η οικογένεια μετασχηματισμών Box-Cox (Box Cox, 1964), που εξαρτώνται από την παράμετρο λ και ορίζονται ως ακολούθως:

$$w_t = \begin{cases} \log(y_t) & \lambda=0 \\ sign(y_t)(|y_t|^\lambda - 1)/\lambda & \text{Διαφορετικά} \end{cases}$$

Μια καλή τιμή του λ είναι αυτή που κάνει το μέγεθος της εποχιακής διακύμανσης περίπου το ίδιο σε ολόκληρη τη σειρά, καθώς μετά απαιτείται ένα απλούστερο μοντέλο πρόβλεψης. Για να επιλέξουμε μια τιμή του λάμδα μπορούμε να χρησιμοποιήσουμε τη μέθοδο Guerrero (Guerrero, 1993).

Οι μαθηματικοί μετασχηματισμοί των δεδομένων είναι ιδιαίτερα χρήσιμοι όταν θέλουμε να διασφαλίσουμε ότι οι προβλέψεις παραμένουν εντός κάποιων προκαθορισμένων ορίων ή όταν θέλουμε να παραμένουν πάντα θετικές.

Μετασχηματισμός για αντιμετώπιση του περιορισμού: Θετικές προβλέψεις
Για να ενσωματώσουμε αυτό το περιορισμό στο μοντέλο πρόβλεψης εργαζόμαστε στην λογαριθμική κλίμακα (Box-Cox μετασχηματισμός για λ = 0). Είναι σημαντικό να αναφερθεί ότι με αυτό το μετασχηματισμό οι κατανομές των προβλέψεων περιορίζονται σε θετικές τιμές, γεγονός που τις κάνει να γίνονται ασύμμετρες με την πάροδο του χρόνου.

Μετασχηματισμός για αντιμετώπιση του περιορισμού: Προβλέψεις που περιορίζονται σε ένα διάστημα ($\hat{y}_{T+h|T} \in [a, b]$)
Σε αυτή τη περίπτωση εφαρμόζουμε ένα κλιμακωτό λογιστικό μετασχηματισμό (Scaled logit transform), ο οποίος αντιστοιχεί το διάστημα [a,b] σε ολόκληρη την πραγματική γραμμή. Ο μετασχηματισμός ορίζεται ως

$$y = log(\frac{x-a}{b-x})$$

όπου με y συμβολίζουμε τα μετασχηματισμένα δεδομένα και με x συμβολίζουμε τα δεδομένα στην αρχική τους κλίμακα. Όπως και στην προηγούμενη περίπτωση, έτσι και εδώ τα διαστήματα πρόβλεψης γίνονται ασύμμετρα και θα είναι μεταξύ του διαστήματος [a,b].

3.2 Συνιστώσες χρονοσειράς

Εαν υποθέσουμε ότι μια χρονοσειρά αναλύεται προσθετικά, τότε μπορούμε να γράψουμε:

$$y_t = S_t + T_t + R_t$$

όπου y_t είναι τα δεδομένα, S_t είναι η συνιστώσα της εποχικότητας, T_t είναι η συνιστώσα της τάσης-κυκλικότητας και R_t η συνιστώσα των υπολοίπων, κατα την χρονική στιγμή t.

Εναλλακτικά, εαν μια χρονοσειρά μπορεί να αναλυθεί πολλαπλασιαστικά τότε η ανάλυση εκφράζεται μέσω της σχέσης

$$y_t = S_t \times T_t \times R_t$$

Η προσθετική ανάλυση είναι καταλληλότερη εαν το μέγεθος των εποχιακών αυξομειώσεων ή της διακύμανσης γύρω από την τάση-κυκλικότητα, δεν μεταβάλλεται ανάλογα με το επίπεδο των χρονοσειρών. Όταν η διακύμανση στο μοτίβο της εποχικότητας, ή όταν η διακύμανση γύρω από την τάση-κυκλικότητα, φαίνεται να είναι ανάλογη με το επίπεδο των χρονοσειρών, τότε η πολλαπλασιαστική ανάλυση είναι καταλληλότερη. Η πολλαπλασιαστική ανάλυση συνηθίζεται σε οικονομικές χρονοσειρές. Μια εναλλακτική στη χρήση πολλαπλασιαστικής ανάλυσης είναι ο εκ των προτέρων μετασχηματισμός των δεδομένων, έως ότου η διακύμανση της σειράς να φαίνεται σταθερή με την πάροδο του χρόνου, και στη συνέχεια η εφαρμογή μιας προσθετικής ανάλυσης. Αυτό είναι χρήσιμο καθώς κάποιες μέθοδοι ή αλγόριθμοι που έχουν αναπτυχθεί δέν υποστηρίζουν την πολλαπλασιαστική ανάλυση.

Προσαρμογή εποχιακών δεδομένων
Εαν η συνιστώσα της εποχικότητας αφαιρεθεί απο τα αρχικά δεδομένα, οι τιμές που προκύπτουν είναι τα "εποχιακά προσαρμοζόμενα" δεδομένα. Για μια προσθετική ανάλυση, η προσαρμογή των εποχιακών δεδομένων προκύπτει από την σχέση $y_t - S_t$ και για την πολλαπλασιαστική ανάλυση δίνεται από την σχέση y_t/S_t .

Εαν η διακύμανση λόγω εποχικότητας δεν είναι πρωταρχικού ενδιαφέροντος, η εποχιακή προσαρμογή της σειράς μπορεί να φανεί χρήσιμη. Για παράδειγμα, τα μηνιαία δεδομένα ανεργίας συνήθως προσαρμόζονται εποχιακά προκειμένου να επισημανθεί περισσότερο η διακύμανση λόγω της υποκείμενης κατάστασης της οικονομίας παρά η εποχιακή διακύμανση.

3.2.1 Κινητοί μέσοι

Οι κινητοί μέσοι αποτελούν μια κλασσική μέθοδο ανάλυσης χρονοσειρών που δημιουργήθηκε τη δεκαετία του 1920 και χρησιμοποιήθηκε ευρέως μέχρι και τη δεκαετία του 1950. Ακόμα και σήμερα αποτελεί τη βάση πολλών μεθόδων ανάλυσης χρονοσειρών, επομένως είναι σημαντικό να κατανοήσουμε πώς λειτουργεί. Συχνά το πρώτο βήμα σε μια κλασική ανάλυση είναι η εκτίμηση της τάσης-κυκλικότητας μέσω της μεθόδου κινητών μέσων.

Εξομάλυνση με κινητούς μέσους

Ένας κινητός μέσος τάξης m μπορεί να γραφτεί ως:

$$\hat{T}_t = \frac{1}{m} \sum_{j=-k}^{k} y_{t+j}, \text{ όπου } m = 2k + 1$$

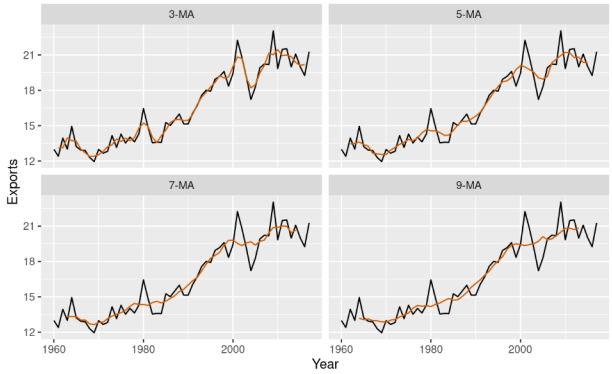
Με άλλα λόγια, η εκτίμηση της τάσης-κυκλικότητας τη χρονική στιγμή t επιτυγχάνεται με τις μέσες τιμές των χρονοσειρών εντός k χρονικών περιόδων του t. Οι παρατηρήσεις που βρίσκονται σε γειτονικές χρονικές στιγμές είναι, επίσης, πιθανό να έχουν παρόμοια τιμή. Επομένως, ο μέσος όρος εξαλείφει κάποια από την τυχαιότητα στα δεδομένα, αφήνοντας μια ομαλή συνιστώσα τάσης-κυκλικότητας. Αυτό το ονομάζουμε m-MA, που σημαίνει κινητός μέσος όρος (Moving Average - MA) τάξης m.

Year	Exports	5-MA
1960	12.99	
1961	12.40	
1962	13.94	13.46
1963	13.01	13.50
1964	14.94	13.61
1965	13.22	13.40
1966	12.93	13.25
1967	12.88	12.66
...	...	...
2010	19.84	21.21
2011	21.47	21.17
2012	21.52	20.78
2013	19.99	20.81
2014	21.08	20.37
2015	20.01	20.32
2016	19.25	
2017	21.27	

Figure 3.2.1: Ετήσιες εξαγωγές αγαθών και υπηρεσιών στην Ελλάδα:1960-2017

Στην τελευταία στήλη του πίνακα, φαίνεται ένας κινητός μέσος τάξης 5, που παρέχει μια εκτίμηση της τάσης-κυκλικότητας. Η πρώτη τιμή σε αυτήν τη στήλη είναι ο μέσος όρος των πέντε πρώτων παρατηρήσεων, 1960-1964, η δεύτερη τιμή στη στήλη 5-MA είναι ο μέσος όρος των τιμών για την περίοδο 1961-1965, και ούτω καθεξής. Κάθε τιμή στη στήλη 5-MA είναι ο μέσος όρος των παρατηρήσεων στο πενταετές παράθυρο με επίκεντρο το αντίστοιχο έτος. Δεν υπάρχουν τιμές ούτε για τα k πρώτα χρόνια ούτε για τα k τελευταία, γιατί μας λείπουν k

παρατηρήσεις και στις δύο περιπτώσεις. Αργότερα θα χρησιμοποιήσουμε πιο εξελιγμένες μεθόδους εκτίμησης της τάσης-κυκλικότητας που επιτρέπουν εκτιμήσεις στα ακραία σημεία. Πιο κάτω παρουσιάζουμε την εκτίμηση της συνιστώσας της τάσης μέσω της μεθόδου των κινητών μέσων για διάφορες τιμές του k.



Παρατηρήστε ότι η τάση-κυκλικότητα (με πορτοκαλί χρώμα) είναι ομαλότερη από τα αρχικά δεδομένα και αποτυπώνει την κύρια κίνηση των χρονοσειρών χωρίς όλες αυτές τις μικρές διακυμάνσεις. Η τάξη του κινητού μέσου όρου καθορίζει την ομαλότητα της εκτίμησης της τάσης-κυκλικότητας. Γενικά, μεγαλύτερη τάξη σημαίνει ομαλότερη καμπύλη.

3.2.2 Κεντρικός κινητός μέσος

Είναι δυνατόν να εφαρμοστεί ένας κινητός μέσος σε έναν κινητό μέσο. Ένας λόγος για να γίνει αυτό είναι να μετατρέψουμε έναν άρτιας τάξης κινητό μέσο σε συμμετρικό.

Όταν ένας 2-MA ακολουθεί έναν κινητό μέσο άρτιας τάξης (όπως το 4), ονομάζεται "κεντρικός κινητός μέσος τάξης 4". Αυτό συμβαίνει επειδή τα αποτελέσματα είναι πλέον συμμετρικά .

Εκτίμηση της τάσης-κυκλικότητας σε εποχιακά δεδομένα

Η πιο συνηθισμένη χρήση των κεντρικών κινητών μέσων είναι για την εκτίμηση της τάσης-κυκλικότητας σε εποχιακά δεδομένα. Θεωρείστε τον 2×4-MA:

$$\hat{T}_t = \frac{1}{8}y_{t-2} + \frac{1}{4}y_{t-1} + \frac{1}{4}y_t + \frac{1}{4}y_{t+1} + \frac{1}{8}y_{t+2}$$

Όταν εφαρμόζεται σε τριμηνιαία δεδομένα, κάθε τρίμηνο του έτους έχει ίση βαρύτητα, καθώς ο πρώτος και ο τελευταίος όρος αντιστοιχούν στο ίδιο τρίμηνο σε διαδοχικά έτη. Έτσι όλα τα τρίμηνα έχουν βαρύτητα ίση με 1/4. Κατά συνέπεια, η εποχιακή διακύμανση θα εξισορροπηθεί και οι προκύπτουσες τιμές $\hat{T}_t$ θα έχουν ελάχιστη ή καθόλου εποχιακή διακύμανση

Γενικά, ένας 2 × m-MA είναι ισοδύναμος με έναν σταθμισμένο μέσο τάξης m+1, όπου όλες οι παρατηρήσεις έχουν βάρος 1/m, εκτός από τους πρώτους και τους τελευταίους όρους που έχουν βάρη 1/(2m). Έτσι, εαν η εποχιακή περίοδος είναι άρτια και τάξης m, χρησιμοποιούμε έναν 2×m-MA για να εκτιμήσουμε την τάση-κυκλικότητα. Εαν η εποχιακή περίοδος είναι περιττή και τάξης m, χρησιμοποιούμε έναν m-MA για να εκτιμήσουμε την τάση-κυκλικότητα. Για παράδειγμα, για την εκτίμηση της τάσης-κυκλικότητας των μηνιαίων

δεδομένων μπορεί να χρησιμοποιηθεί ένας 2×12-MA και για την εκτίμηση της τάσης-κυκλικότητας των καθημερινών δεδομένων με μια εβδομαδιαία εποχικότητα ένας 7-MA.

Άλλες επιλογές για την τάξη του MA συνήθως οδηγούν σε αλλοίωση των εκτιμήσεων της τάσης-κυκλικότητας από την εποχικότητα των δεδομένων

3.2.3 Σταθμισμένοι κινητοί μέσοι

Οι συνδυασμοί κινητών μέσων έχουν ως αποτέλεσμα σταθμισμένους κινητούς μέσους. Για παράδειγμα, ο $2x4 - MA$ που συζητήθηκε παραπάνω είναι ισοδύναμος με το σταθμισμένο 5-MA με βάρη που δίνονται από το διάνυσμα $\left[\frac{1}{8}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}\right]$.

Σε γενικές γραμμές, ένας σταθμισμένος m-MA μπορεί να γραφτεί ως

$$\hat{T}_t = \sum_{j=-k}^{k} a_j y_{t+j} \quad \text{όπου} \quad k = (m-1)/2$$

Είναι σημαντικό τα βάρη να αθροίζονται στη μονάδα και να είναι συμμετρικά ($a_j = a_{-j}$). Ο απλός m-MA είναι μια ειδική περίπτωση όπου όλα τα βάρη ισούνται με $1/\mathrm{m}$. Ένα σημαντικό πλεονέκτημα των σταθμισμένων κινητών μέσων είναι ότι αποδίδουν μια ομαλότερη εκτίμηση της τάσης-κυκλικότητας. Αντί τα πλήρη βάρη να εισάγονται ή να αφαιρούνται από τους υπολογισμούς, αυξάνονται και στη συνέχεια μειώνονται αργά, με αποτέλεσμα μια ομαλότερη καμπύλη.

3.2.4 Κλασική Ανάλυση χρονοσειρών

Η κλασική μέθοδος ανάλυσης δημιουργήθηκε τη δεκαετία του 1920. Είναι μια σχετικά απλή διαδικασία και αποτελεί το σημείο εκκίνησης για τις περισσότερες άλλες μεθόδους ανάλυσης χρονοσειρών. Υπάρχουν δύο μορφές κλασσικής ανάλυσης: η προσθετική και η πολλαπλασιαστική ανάλυση. Οι 2 μέθοδοι κλασσικής ανάλυσης περιγράφονται ακολούθως σε μια χρονοσειρά με εποχιακή περίοδο m .

Προσθετική ανάλυση

Βήμα 1

Εαν το m είναι άρτιος αριθμός, η συνιστώσα της τάσης-κυκλικότητας $\hat{T}_t$ υπολογίζεται χρησιμοποιώντας έναν 2×m-MA ενώ αν το m είναι περιττός αριθμός τότε υπολογίζεται χρησιμοποιώντας έναν m-MA.

Βήμα 2

Υπολογίζεται η σειρά χωρίς την τάση: $y_t - \hat{T}_t$

Βήμα 3

Η συνιστώσα εποχικότητας κάθε εποχής υπολογίζεται ως η μέση τιμή των τιμών της χρονοσειρά χωρίς την τάση, για την συγκεκριμένη εποχή. Για παράδειγμα, με μηνιαία δεδομένα, η εποχιακή συνιστώσα για το Μάρτιο είναι ο μέσος όρος όλων των τιμών του Μαρτίου στα δεδομένα χωρίς την τάση. Η εποχιακή συνιστώσα λαμβάνεται ενώνοντας αυτές τις μηνιαίες τιμές και, στη συνέχεια, αναπαράγοντας την ακολουθία για κάθε έτος δεδομένων.

Βήμα 4
Η συνιστώσα των υπολοίπων υπολογίζεται αφαιρώντας τις εκτιμώμενες συνιστώσες εποχικότητας και τάσης-κυκλικότητας: $\hat{R}_t = y_t - \hat{T}_t - \hat{S}_t$

Πολλαπλασιαστική ανάλυση
Η κλασική πολλαπλασιαστική ανάλυση είναι παρόμοια με την προσθετική, με τη διαφορά ότι οι αφαιρέσεις αντικαθίστανται από διαιρέσεις.

Βασική υπόθεση της κλασικής ανάλυσης είναι ότι η εποχιακή συνιστώσα είναι σταθερή από χρόνο σε χρόνο. Αυτή η βασική υπόθεση τη καθιστά προβληματική στην πράξη. Για παράδειγμα τα μοτίβα ζήτησης ενέργειας έχουν αλλάξει αρκετά σε σχέση με τις περασμένες δεκαετίες. Πριν μερικές δεκαετίες ο κλιματισμός το καλοκαίρι δέν ήταν τόσο συνηθισμένος ενώ τώρα είναι, γεγονός που οδήγησε στην αλλαγή του εποχιακού μοτίβου με τη πάροδο των χρόνων. Ακόμα ένα αρνητικό στοιχείο της κλασσικής ανάλυσης είναι ότι η εκτίμηση της συνιστώσας της τάσης δέν είναι διαθέσιμη για τις πρώτες και τελευταίες m παρατηρήσεις. Επίσης η $\hat{T}_t$ τείνει να υπερομαλοποιεί πολύ γρήγορες αυξήσεις και μειώσεις στα δεδομένα. Τέλος, η κλασσική ανάλυση είναι κατάλληλη κυρίως για μηνιαία και τριμηνιαία δεδομένα.

Ανάλυση SEATS
"SEATS" σημαίνει Εξαγωγή Εποχικότητας σε Χρονοσειρές ARIMA (Seasonal Extraction in ARIMA Time Series). Η διαδικασία αυτή αναπτύχθηκε στην Τράπεζα της Ισπανίας και χρησιμοποιείται ευρέως από κυβερνητικές υπηρεσίες σε όλο τον κόσμο.
Μερικά από τα θετικά της μεθόδου αυτής είναι:

- Μπορεί να χειρίζεται δεδομένα χρονοσειρών με πολλαπλά εποχικά πρότυπα ή ακανόνιστες διακυμάνσεις.

- Ενσωματώνει προηγμένες στατιστικές τεχνικές, όπως η μοντελοποίηση ARIMA, για την εκτίμηση και την εξαγωγή των υποκείμενων συνιστωσών της χρονοσειράς. Αυτές οι τεχνικές τη καθιστούν πιο ανθεκτική σε ακραίες τιμές και θόρυβο στα δεδομένα, ενισχύοντας την ακρίβεια των αποτελεσμάτων της αποσύνθεσης.

- Διατηρεί τη προσθετική φύση της αρχικής χρονοσειράς. Αυτό σημαίνει ότι το άθροισμα των αποσυνδεδεμένων συνιστωσών (τάση, εποχιακή και υπόλοιπο) θα ισούται με την αρχική σειρά, εξασφαλίζοντας τη συνέπεια και την ερμηνευσιμότητα των αποτελεσμάτων της αποσύνθεσης.

Ενώ τα κύρια μειονεκτήματα/περιορισμοί αυτής της μεθόδου είναι:

- Απαιτεί τον προσδιορισμό διαφόρων παραμέτρων, όπως η τάξη του μοντέλου ARIMA και η συχνότητα των εποχικών προτύπων. Η ακρίβεια των αποτελεσμάτων της αποσύνθεσης μπορεί να είναι ευαίσθητη σε αυτές τις επιλογές παραμέτρων και η επιλογή των βέλτιστων τιμών μπορεί να απαιτεί κάποια δοκιμή και σφάλμα ή γνώση από ειδικούς.

- Περιλαμβάνει πολύπλοκους υπολογισμούς, συμπεριλαμβανομένων επαναληπτικών διαδικασιών μοντελοποίησης και εκτίμησης. Αυτό τη καθιστά υπολογιστικά απαιτητική, ιδίως για μεγάλα σύνολα δεδομένων ή σύνολα

δεδομένων χρονοσειρών υψηλής συχνότητας.

- Υποθέτει ότι η σειρά είναι στάσιμη (οι στατιστικές ιδιότητες των δεδομένων παραμένουν σταθερές με την πάροδο του χρόνου)

3.3 Ανάλυση STL

Η μέθοδος STL (Seasonality and Trend decomposition using Loess) αναπτύχθηκε το 1990 και αποτελεί μια διαδικασία "φιλτραρίσματος" μιας χρονοσειράς για την αποσύνθεση της σε 3 συνιστώσες: Τάση-Κυκλικότητα (Trend-Cyclic), Εποχικότητα (Seasonality) και Υπόλοιπο (Noise). Η ανάπτυξη της είχε σκοπό να επιλύσει τα μειονεκτήματα της κλασσικής ανάλυσης. Τα κυριότερα πλεονεκτήματα της έναντι της κλασσικής ανάλυσης είναι τα εξής:

- Αντιμετωπίζει κάθε είδους εποχικότητας και όχι μόνο μηνιαία και τριμηνιαία δεδομένα
- Η εποχιακή συνιστώσα επιτρέπεται να αλλάζει με την πάροδο του χρόνου και ο ρυθμός μεταβολής της μπορεί να ελεγχθεί από τον χρήστη
- Η ομαλότητα της τάσης-κυκλικότητας μπορεί επίσης να ελεγχθεί από το χρήστη .
- Δεν επηρεάζεται έντονα από ακραίες τιμές (outliers)

Ένα από τα σημαντικότερα μειονεκτήματα της έναντι των μεθόδων κλασσικής ανάλυσης και SEAT είναι ότι υλοποιείται μόνο για προσθετική ανάλυση χρονοσειράς. Βέβαια εαν υπάρχουν ισχυρές ενδείξεις ότι η πολλαπλασιαστική ανάλυση είναι αποτελεσματικότερη από την προσθετική για μια χρονοσειρά τότε ένας λογαριθμικός μετασχηματισμός των δεδομένων μπορεί να επιλύσει αυτό το μειονέκτημα. Πιο συγκεκριμένα αρχικά τα δεδομένα λογαριθμίζονται και στην συνέχεια εφαρμόζεται η STL ανάλυση. Ακόμα ένα μειονέκτημα της είναι ότι δεν μπορεί να διαχειριστεί αυτόματα την ημέρα συναλλαγών ή τις ημερολογιακές διακυμάνσεις.

Η STL είναι μια απλή μέθοδος η οποία συνίσταται από μια σειρά εφαρμογών του εξομαλυντή Loess (Loess smoother). Η απλότητα της μεταφράζεται σε μικρό υπολογιστικό κόστος γεγονός που τη καθιστά ικανή να αναλύσει γρήγορα μεγάλες χρονοσειρές και να διαχειριστεί περίπλοκες μορφές εποχικότητας και τάσης.

Εξομάλυνση τάσης (Trend Smoothing):
Είναι η διαδικασία κατά τη οποία αφαιρούνται οι μικρής-διάρκειας διακυμάνσεις (short-term fluctuations) από τη χρονοσειρά και εξάγεται η μεγάλης-διάρκειας τάση (long-term trend). Το αποτέλεσμα είναι μια ομαλή καμπύλη τάσης (smoothed trend line) η οποία παρουσιάζει τη γενική κατεύθυνση της χρονοσειράς.

Εξομάλυνση εποχικότητας (Seasonal Smoothing):
Η εποχιακή εξομάλυνση είναι η διαδικασία κατα τη οποία αφαιρείται η εποχιακή συνιστώσα από τη χρονοσειρά. Η εποχιακά προσαρμοσμένη χρονοσειρά μπορεί να χρησιμοποιηθεί για την ανάλυση της μή-εποχιακής συμπεριφοράς της χρονοσειράς.

3.3.1 Η συνάρτηση Loess (LOcally WEighted Scatter-plot Smoother)

Έστω $x_i, y_i, i = 1, ..., n$ τα δεδομένα, όπου x_i η ανεξάρτητη μεταβλητή και y_i η εξαρτημένη.
Η καμπύλη παλινδρόμησης Loess (Loess regression curve) $\hat{g}(x)$ είναι η εξομάλυνση (παλινδρόμηση) της y δοθέντος της x η οποία ορίζεται για κάθε x και όχι απλά στα x_i

Υπολογισμός της $\hat{g}(x)$

Ο υπολογισμός της καμπύλης Loess συνοψίζεται στον πιό κάτω αλγόριθμο.

1. Επιλογή ενός $q > 0$

2. Εντοπίζουμε τα q στο πλήθος x_i που είναι πλησιέστερα στο x και σε καθένα από αυτά τα x_i αντιστοιχούμε ένα βάρος γειτονίας (neighborhood weight) το οποίο καθορίζεται με βάση την απόσταση του από το x .

$$\lambda_q(x) = \text{ η απόσταση του q-οστού πιο μακρινού } x_i \text{ από το } x$$

3. Το "βάρος γειτονιάς" για κάθε x_i ορίζεται ως:

$$u_i = W\left(\frac{|x_i - x|}{\lambda_q(x)}\right)$$

όπου W η εξής συνάρτηση:

$$W(u) = \begin{cases} (1-u^3)^3 & 0 \le u < 1 \\ 0 & u \ge 1 \end{cases}$$

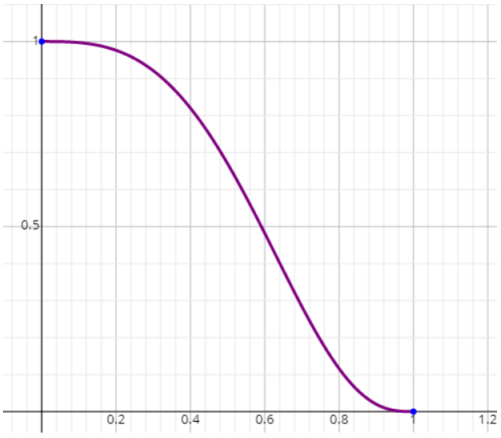


Figure 3.3.1: Η συνάρτηση W στο διάστημα [0,1]

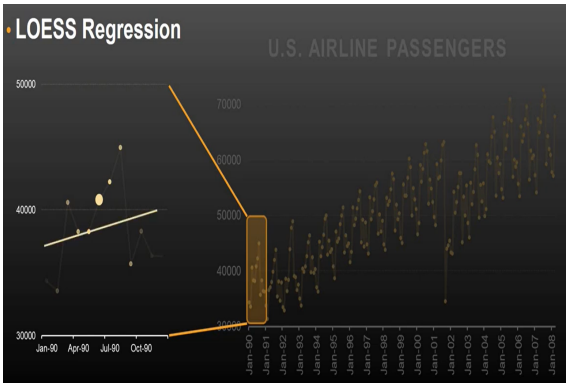
Με βάση λοιπόν τον ορισμό των βαρών γειτονιάς, εύκολα αντιλαμβανόμαστε ότι όσο πλησιέστερο είναι το x_i στο x, τόσο μεγαλύτερο βάρος γειτονίας του αντιστοιχείται. Τα βάρη λαμβάνουν όλο και μικρότερη τιμή όσο

το x_i απομακρύνεται από το x και είναι ίσο με μηδέν από το q-οστό μακρύτερο σημείο και έπειτα.

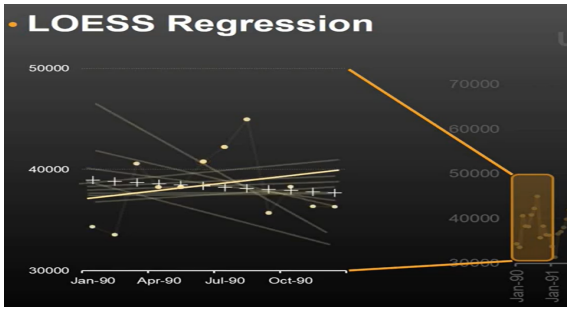
Στην συνέχεια προσαρμόζεται ένα πολυώνυμο βαθμού d όπου οι συντελεστές του προκύπτουν από την ελαχιστοποίηση του σταθμισμένου αθροίσματος των τετραγώνων των σφαλμάτων που ορίζεται από την σχέση

$$\sum_{i=1}^{q} u_i \varepsilon_i^2$$

Σε κάθε ε_i^2 αντιστοιχείται το βάρος u_i. Η τιμή του **τοπικά** προσαρμοσμένου πολυωνύμου στο σημείο x είναι ίση με $\hat{g}(x)$. Για κάθε σημείο προσαρμόζεται διαφορετικό πολυώνυμο και προσδιορίζεται για κάθε x η τιμή $\hat{g}(x)$. Στο τέλος όλες οι τιμές $\hat{g}(x)$ καθορίζουν την συνάρτηση Loess. Στο πιο κάτω σχήμα φαίνεται πώς σχηματίζεται η συνάρτηση Loess.



(a) Τοπικά προσαρμοσμένο πολυωνύμου βαθμού d=1 στο σημείο x.



(b) Τοπικά προσαρμοσμένα πολυώνυμα βαθμού d=1 σε όλα τα σημεία x του υποσυνόλου της χρονοσειράς. Η $\hat{g}(x)$ φαίνεται με σύμβολο "+"

Παρατηρήσεις

Όσο μεγαλύτερο είναι το q, δηλαδή όσες περισσότερες παρατηρήσεις καθορίζουν το πολυώνυμο που θα προσαρμοστεί στο σημείο x , τόσο ομαλότερη είναι η $\hat{g}(x)$. Όταν το $q \to \infty$ τότε τα $u_i(x)$ τείνουν στο 1 για κάθε i και για κάθε x, γεγονός που οδηγεί την $\hat{g}(x)$ να τείνει να γίνει ένα τυπικό πολυώνυμο ελαχίστων τετραγώνων βαθμού d.

Η λειτουργία της STL στηρίζεται σε 2 εμφωλευμένες αναδρομικές διαδικασίες . Ο εσωτερικός βρόγχος (inner loop), επαναλαμβάνεται $n_{(i)}$φορές, εμφωλευμένος σε ένα εξωτερικό βρόγχο (outer loop) ο οποίος επαναλαμβάνεται $n_{(0)}$ φορές.

Στον εσωτερικό βρόγχο, σε κάθε επανάληψη πραγματοποιείτε εξομάλυνση τάσης και εποχικότητας και στην συνέχεια οι συνιστώσες της τάσης και της εποχικότητας ενημερώνονται. Στον εξωτερικό βρόγχο η κάθε επανάληψη αποτελείται από τον εσωτερικό βρόγχο και τον υπολογισμό των βαρών στιβαρότητας (Robustness weights). Τα βάρη στιβαρότητας χρησιμοποιούνται στην επόμενη επανάληψη του εσωτερικού βρόγχου, **με σκοπό να μειώσουν την επιρροή της παροδικής και παρεκκλίνουσας συμπεριφοράς στις συνιστώσες της τάσης και της εποχικότητας.**. Η αρχική εφαρμογή του εσωτερικού βρόγχου πραγματοποιείτε με αρχικά βάρη στιβαρότητας ίσα με 1.

Υποσειρές κυκλικότητας

Είναι μια υποσειρά της χρονοσειράς όπου όλες οι τιμές της αναφέρονται στην ίδια χρονική στιγμή του εποχιακού κύκλου (Seasonal Cycle). Για παράδειγμα για τριμηνιαία δεδομένα η πρώτη υποσειρά κυκλικότητας είναι τα δεδομένα του $1^{ου}$ τριμήνου , η δεύτερη υποσειρά τα δεδομένα του $2^{ου}$ τριμήνου. Στη πιο κάτω χρονοσειρά στην οποία τα δεδομένα είναι τριμηνιαία η κάθε υποσειρά (4 στο πλήθος)χρωματίζονται με διαφορετικό χρώμα.

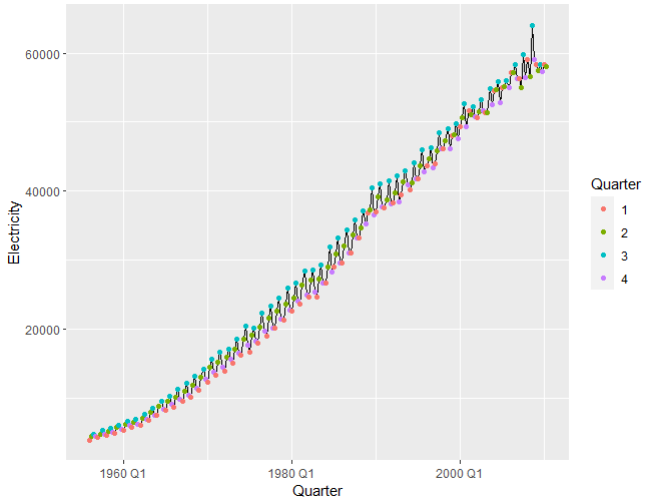


Figure 3.3.3: Γραφική αναπαράσταση τεσσάρων υποσειρών κυκλικότητας σε τριμηνιαία δεδομένα ζήτησης ηλεκτρικής ενέργειας

Το συνολικό πλήθος των υποσειρών-κυκλικότητας συμβολίζεται με n_p

3.3.2 Ο εσωτερικός βρόγχος (The inner loop)

Οι συνιστώσες της τάσης και της εποχικότητας μετά το πέρας της k-οστής επανάληψης του αλγορίθμου εκφράζονται μέσω των μεταβλητών $S_u^{(k)}$ και $T_u^{(k)}$ αντίστοιχα. Οι $S_u^{(k)}$ και $T_u^{(k)}$ ορίζονται για όλους τους χρόνους $(u = 1, ...N)$ ακόμα και για τους χρόνους όπου η Y_u είναι ελλιπής.

Υπολογισμός τών $S_u^{(k+1)}$ και $T_u^{(k+1)}$:

Βήμα 1: Detrending

Υπολογίζεται η σειρά χωρίς την επιρροή της τάσης (detrended series) $Y_u^{(k)} - T_u^{(k)}$. Εαν η Y_u είναι ελλιπής τότε η $Y_u^{(k)} - T_u^{(k)}$ είναι ελλιπής.

Βήμα 2: Εξομάλυνση υποσειρών κυκλικότητας (cycle subseries smoothing)

Κάθε υποσειρά κυκλικότητας της detrended series εξομαλύνεται μέσω της συνάρτησης Loess με παραμέτρους $q = n_{(s)}$ και $d = 1$. Οι εξομαλυμένες τιμές υπολογίζονται σε όλες τις χρονικές στιγμές ακόμα και στα σημεία όπου υπάρχουν ελλιπείς τιμές (missing values). Επιπλέον υπολογίζονται για μια χρονική στιγμή πριν την πρώτη καταγεγραμμένη παρατήρηση και μια χρονική στιγμή μετά τη τελευταία καταγεγραμμένη παρατήρηση.

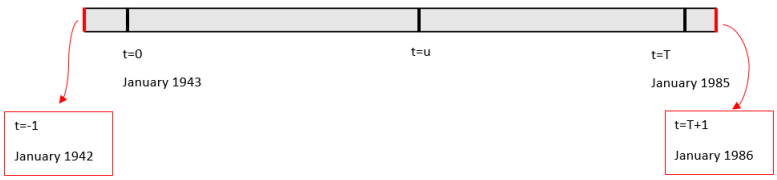


Figure 3.3.4: Παράδειγμα υπολογισμού των εξομαλυμένων τιμών της κυκλικής υποσειράς "Ιανουαρίου" για μηνιαία δεδομένα που συλλέχθηκαν από το 1943 εώς 1985

Η συλλογή όλων των εξομαλυμένων τιμών από όλες τις κυκλικές υποσειρές αποτελεί προσωρινή εποχιακή σειρά $C_u^{(k+1)}$ αποτελούμενη από $N + 2n_p$ τιμές για $u = -n_{(p)+1}, ..., (N + n_{(p)})$

Βήμα 3: Εφαρμογή βαθυπέρατου φιλτραρίσματος στις εξομαλυμένες υποσειρές κυκλικότητας

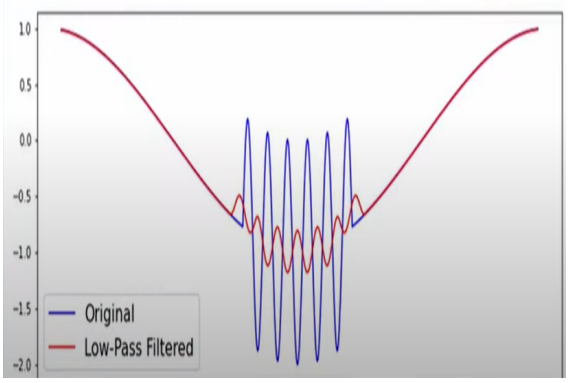
Στην χρονοσειρά $C_u^{(k+1)}$ που δημιουργήθηκε στο βήμα 2 εφαρμόζεται ένα βαθυπέρατο φιλτράρισμα (low pass filter) κατα σειρά:

1. 2 διαδοχικά φιλτραρίσματα κινητών μέσων (MA) μήκους $n_{(p)}$

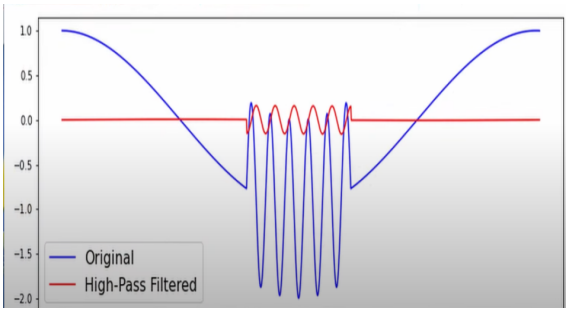
2. Φιλτράρισμα κινητών μέσων (MA) μήκους 3

3. Εξομάλυνση μέσω Loess συνάρτησης για d=1 και q= $n_{(l)}$

Με τον όρο Low-Pass filter εννοούμε την διαδικασία κατα την οποία η χρονοσειρά (σήμα) μετασχηματίζεται κατάλληλα έτσι ώστε να αφαιρεθούν οι θόρυβοι, όπου θόρυβοι θεωρούνται οι περιοχές στην οποία παρουσιάζονται υψηλές συχνότητες. Με το πέρας αυτής της διαδικασίας οι περιοχές χαμηλής συχνότητας μένουν ανεπηρέαστες ενώ οι περιοχές που εμφανίζουν υψηλές συχνότητες μετασχηματίζονται έτσι ώστε να μειωθεί η ένταση τους. Ένα High-Pass filter κάνει ακριβώς το αντίθετο. Στο πιό κάτω γράφημα βλέπουμε γραφικά αυτή την διαδικασία.



(a) Low-Pass filter



(b) High-Pass filter

Μετά το πέρας της πιο πάνω διαδικασίας προκύπτει η σειρά $L_u^{(k+1)}$ η οποία ορίζεται στα σημεία $u = 1, ...N$ (Το ότι η $C_u^{(k+_1)}$ ορίζεται για περισσότερα σημεία ήταν απαραίτητο έτσι ώστε να μπορεί να εφαρμοστεί η πιο πάνω διαδικασία και στο τέλος να έχουμε υπολογισμένη την $L_u^{(k+1)}$ στις χρονικές στιγμές της αρχικής χρονοσειράς)

Βήμα 4: Detrending of smoothed cycle subseries

$$S_u^{(k+1)} = C_u^{(k+1)} - L_u^{(k+1)}, \qquad u = 1, ...N$$

Οι τιμές της $L_u^{(k+1)}$ αφαιρούνται έτσι ώστε η εποχιακή συνιστώσα να μήν επηρεάζεται από χαμηλής εντάσεως δυνάμεις (Low-Frequency Power) [5] οι οποίες αναφέρονται σε μακροχρόνιες αλλαγές (Long-term changes) στα δεδομένα που προκύπτουν μέσα σε χρονικό πλαίσιο μεγαλύτερο από αυτό της εποχιακής συνιστώσας.

Βήμα 5: Deseaonalizing
Υπολογίζεται η χρονοσειρά χωρίς την εποχιακή συνιστώσα (deseasonalized series) $Y_u^{(k)} - S_u^{(k)}$. Εαν η Y_u είναι ελλιπής τότε η $Y_u^{(k)} - S_u^{(k)}$ είναι ελλιπής.

Βήμα 6: Trend Smoothing
Η deseasonalized series εξομαλύνεται μέσω της συνάρτησης Loess για q=$n_{(t)}$ & d=1 και προκύπτουν οι εξομαλυμένες τιμές $T_u^{(k+1)}$. Τα βήματα 2,3,4 έχουν ως σκοπό την εποχιακή εξομάλυνσης (seasonal Smoothing) ενώ το βήμα 5,6 έχει ως σκοπό την εξομάλυνση τάσης(trend smoothing).

[5]Frequency Power: Συχνά χρησιμοποιείται στην ανάλυση χρονοσειρών για να εκφράσει τον ρυθμό με τον οποία ένα σήμα ταλαντώνεται ή αυξομειώνεται

Αρχικές τιμές

Θέτουμε ως αρχική τιμή της $T_u^{(0)}$ να είναι η μηδενική η οποία λειτουργεί αρκετά καλά. Κανείς θα απορούσε γιατί να επιλέξουμε αυτή τη τιμή, αφού αυτό επιτρέπει κατα την πρώτη επανάληψη, η τάση να επηρεάσει την εποχιακή συνιστώσα διότι στο βήμα 1 δέν πραγματοποιείται η αφαίρεση της τάσης. Παρόλο που κατα την πρώτη επανάληψη οι διακυμάνσεις της τάσης απορροφούνται από τις εξομαλυμένες κυκλικές υποσειρές (δηλαδή από την $C_u^{(1)}$) στην συνέχεια μεγάλο μέρος αυτών αφαιρείται στο βήμα 4 (Detrending of smoothed cycle subseries) .

3.3.3 Ο εξωτερικός βρόγχος (The outer loop)

Θεωρούμε ότι έχουμε ήδη πραγματοποιήσει μια αρχική επανάληψη του εσωτερικού βρόγχου για να πάρουμε μια αρχική εκτίμηση των T_u, S_u για όλα τα u. Ορίζουμε τα υπόλοιπα ως εξής:

$$R_u = Y_u - T_u - S_u$$

Θα καθορίσουμε ενα βάρος στιβαρότητας για κάθε χρονική στιγμή (για όλα τα R_u). Μια παρατήρηση που αποτελεί ακραία τιμή (κάτι που προκαλεί μεγάλη τιμή στο $|R_u|$) θα έχει μηδενικό ή σχεδόν μηδενικό βάρος στιβαρότητας.

$$h = 6median(|R_u|)$$

$$p_u = B(|R_u|/h), \qquad \text{όπου} \qquad B(u) = \begin{cases} (1-u^2)^2 & 0 \le u < 1 \\ 0 & u > 1 \end{cases}$$

Μετά τον υπολογισμό των βαρών στιβαρότητας για κάθε R_u, επαναλαμβάνεται ο εσωτερικός βρόγχος με κάποιες διαφορές στα βήματα 2 και 6. Πιο συγκεκριμένα κατα τον υπολογισμό των βαρών γειτονιάς u_i, που αντιστοιχεί στην παρατήρηση που συμβαίνει την χρονική στιγμή t, για μια τιμή x, αυτό πολλαπλασιάζεται με το αντίστοιχο βάρος στιβαρότητας p_t

3.3.4 Επιπλέον εξομάλυνση της συνιστώσας εποχικότητας

Η εποχιακή συνιστώσα που υπολογίστηκε μέσω των 2 εμφωλευμένων βρόγχων δεν αποτελεί και την τελική συνιστώσα καθώς απαρτίζεται από τοπική τραχύτητα. Για αυτό τον λόγο πραγματοποιούμε εξομάλυνση χρησιμοποιώντας την συνάρτηση Loess και προκύπτει η τελική εποχιακή συνιστώσα η οποία είναι αρκετά τοπικά ομαλή.

3.4 Χαρακτηριστικά STL

Η ανάλυση χρονοσειρών μπορεί να χρησιμοποιηθεί για τη μέτρηση της ισχύος της τάσης και της εποχικότητας σε μια χρονοσειρά. Με την προσθετική ανάλυση η χρονοσειρά γράφεται, όπως αναφέρθηκε και πιο πάνω, ως

$$y_t = T_t + S_t + R_t$$

όπου T_t είναι η συνιστώσα εξομάλυνσης της τάσης, S_t η συνιστώσα εποχικότητας και R_t η συνιστώσα των υπολοίπων. Για δεδομένα με έντονη τάση, τα εποχιακά προσαρμοζόμενα δεδομένα θα πρέπει να έχουν πολύ μεγαλύτερη διακύμανση από τη συνιστώσα των υπολοίπων. Επομένως το κλάσμα $Var(R_t)/Var(T_t + R_t)$ θα πρέπει να είναι σχετικά μικρό. Όμως για δεδομένα με μικρή ή μηδενική τάση, οι δύο διακυμάνσεις θα πρέπει να είναι περίπου ίδιες. Έτσι ορίζουμε την ισχύ της τάσης ως:

$$F_t = max\left(0, 1 - \frac{Var(R_t)}{Var(T_t + R_t)}\right), \qquad F_t \in [0, 1]$$

Η ισχύς της εποχικότητας ορίζεται με παρόμοιο τρόπο, ως προς τα δεδομένα που έχει αφαιρεθεί η τάση κι όχι ως προς τα εποχιακά προσαρμοζόμενα δεδομένα.

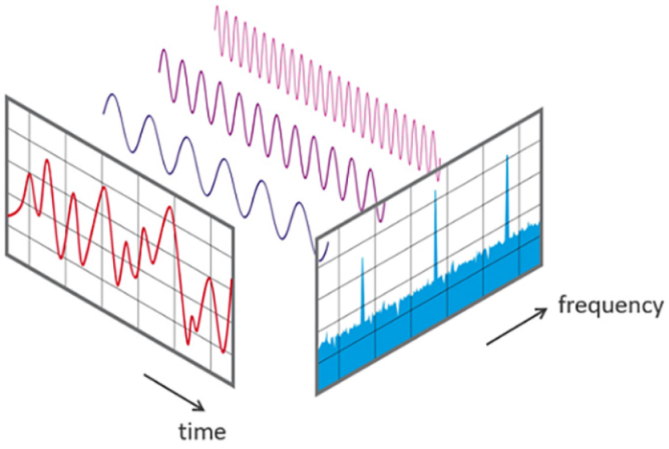
$$F_s = max\left(0, 1 - \frac{Var(R_t)}{Var(S_t + R_t)}\right), \qquad F_s \in [0, 1]$$

Μια σειρά με εποχιακή ισχύ F_s κοντά στο 0 δεν παρουσιάζει σχεδόν καμία εποχικότητα, ενώ μια σειρά με ισχυρή εποχικότητα θα έχει F_s κοντά στο 1, επειδή η $Var(R_t)$ θα είναι πολύ μικρότερη από τη $Var(S_t + R_t)$

Μπορούμε στη συνέχεια να χρησιμοποιήσουμε τα χαρακτηριστικά αυτά σε γραφικές παραστάσεις για να προσδιορίσουμε τον τύπο της σειράς με ισχυρή τάση και με τη μεγαλύτερη εποχικότητα. Αυτά τα μέτρα μπορούν να είναι χρήσιμα όταν, για παράδειγμα, έχουμε μια μεγάλη συλλογή χρονοσειρών και πρέπει να βρούμε τη σειρά με την μεγαλύτερη τάση ή με τη μεγαλύτερη εποχικότητα.

4 Wavelet Transformation, Time & Frequency decomposition

Ένας πολύ γνωστός τρόπος ανάλυσης μιας χρονοσειράς είναι μέσω της ανάλυσης Fourier η οποία αναλύει μια δοσμένη συνάρτηση (ένα σήμα ή χρονοσειρά) σε ένα άθροισμα από ημίτονα και συνημίτονα κατάλληλων συχνοτήτων. Μέσω της ανάλυσης Fourier μεταβαίνουμε από το πεδίο του χρόνου (time domain) στο πεδίο συχνοτήτων (Frequency Domain). Το πεδίο της συχνότητας θα μας πει τη σχετική συμβολή κάθε συχνότητας στη σύνθεση της συνάρτησης.



Το μειονέκτημα της μεθόδου Fourier είναι οτι δε μας δίνεται η οποιαδήποτε πληροφορία για το πότε ξεκινάει και πότε τελειώνει η κάθε συχνότητα, καθώς κατά την ανάλυση Fourier αρχικά η χρονοσειρά (σήμα) συρρικνώνεται στον χρόνο με αποτέλεσμα να χάνεται η πληροφορία αυτή. Για τον λόγο αυτό η ανάλυση Fourier δέν βοηθάει πολύ στην ανάλυση μιας χρονοσειράς ιδιαίτερα σε μή στάσιμες χρονοσειρές.

Γενικά για την ανάλυση μή στάσιμων χρονοσειρών (non-stationary) υπάρχουν 2 κύριες προσεγγίσεις.

- Μεθόδοι χρόνου (Time methods): Ανάλυση αυτοσυχέτισης (Autocorrelation Analysis), Μέθοδοι παλινδρόμησης (regression methods) κ.α

- Φασματοχρονικές μέθοδοι (Spectro-Temporal methods)

4.1 Μετασχηματισμός κυματιδίων (Wavelet transform)

Η μέθοδος μετασχηματισμού κυματιδίου (WT) ανήκει στην 2^{η} κατηγορία. Ένα από τα κύρια πλεονεκτήματα του WT είναι ότι μπορεί να αποσυνθέσει ένα σήμα απευθείας σύμφωνα με τη συχνότητα και να το αναπαραστήσει τόσο στο πεδίο της συχνότητα όσο και στο πεδίο του χρόνου. Επομένως τόσο η πληροφορια της συχνότητας όσο και του χρόνου διατηρούνται μετά από τον μετασχηματισμό. Είναι επομένως ένας πιο ισχυρός μετασχηματισμός για ανάλυση χρόνου-συχνότητας σε σχέση με τον μετασχηματισμό Fourier. Ο μετασχηματισμός κυματιδίων ως μαθηματική έννοια μπορεί να ερμηνευθεί ως η συνέλιξη (convolution) του σήματος με μια συνάρτηση κυματιδίου. Ο μετασχηματισμός κυματιδίων κυκλοφορεί σε δύο διαφορετικές εκδοχές: τον συνεχή και τον διακριτό μετασχηματισμό κυματιδίων.

Ορισμός Wavelet: Ένα wavelet είναι μια συνάρτηση $\psi \in \mathbb{L}^2(R)$ η οποία ικανοποιεί την εξής σχέση (admissibility condition):

$$C_\psi = \int_{-\infty}^{+\infty} \frac{|\hat{\Psi}(f)|^2}{|f|} df < \infty$$

όπου $\hat{\psi}(f) = \int_{-\infty}^{\infty} \psi(t) e^{-i(2\pi f)t} dt$ είναι ο μετασχηματισμός Fourier της ψ(t). Η C_ψ είναι η σταθερά επιτρεπτότητας (admissibility constant)

Η πιο πάνω ιδιότητα εξασφαλίζει την ύπαρξη αντίστροφης φόρμουλας για το συνεχή μετασχηματισμό wavelet. Από αυτή την ιδιότητα προκύπτει ότι $\hat{\psi}(\omega) \to 0$ καθώς $f \to 0$. Πράγματι αν $\hat{\psi}(f)$ είναι συνεχής, τότε $\hat{\psi}(0) = 0$ δηλαδή $\int_{-\infty}^{+\infty} \psi(t) = 0$.

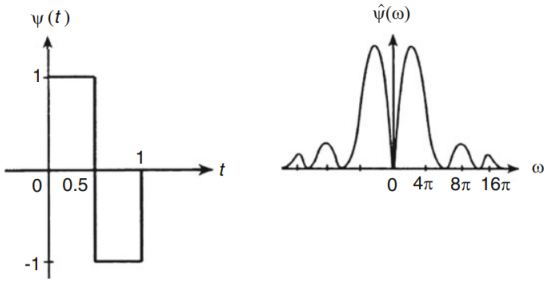
Ποιες οι προϋποθέσεις για να αποτελεί μια συνάρτηση ψ(t) κυματίδιο (Wavelet);
Με το όρο wavelet αναφερόμαστε σε μια οικογένεια συναρτήσεων οι οποίες αναπαριστούν σύντομες ταλαντώσεις περιορισμένες σε συγκεκριμένο χρονικό διάστημα (τοπικά κυματοειδής συνάρτηση). Είναι απλές συναρτήσεις του χρόνου. Για να ανήκει μια συνάρτηση στην οικογένεια των Wavelet θα πρέπει να ικανοποιεί τις εξής κύριες ιδιότητές:

- $\int_{-\infty}^{+\infty} \Psi(t) dt = 0$: Μηδενική μέση τιμή
- $\int_{-\infty}^{+\infty} |\Psi(t)^2| dt = 1$: Ενέργεια ίση με 1. Γενικότερα αρκεί να έχει πεπερασμένη ενέργεια (δηλαδή να ανήκει στον $L^2(\mathbb{R})$). Με κανονικοποίηση μπορούμε να κάνουμε την ψ να έχει ενέργεια ίση με 1. Αυτό είναι βολικό καθώς προκύπτει η έννοια της ορθογωνιότητας.

Παράδειγμα 1: Haar wavelet

$$\Psi(t) = \begin{cases} 1 & 0 \le t < \frac{1}{2} \\ -1 & \frac{1}{2} \le t < 1 \\ 0 & \text{Διαφορετικά} \end{cases}$$

Η ψ(t) και η $\hat{\psi}(t)$ αναπαρίστανται στο πίο κάτω σχήμα.



Στην συνέχεια δίνεται ένα σημαντικό θεώρημα που είναι χρήσιμο για την κατασκευή νέων, πιο σύνθετων κυματιδίων.

Θεώρημα:
Αν ψ είναι wavelet και φ είναι μια φραγμένη διαφορίσιμη συνάρτηση, τότε η συνέλιξη (convolution function) ψ*φ είναι wavelet. (Απόδειξη στο βιβλίο: [17])

Για να συλλάβει τις υψηλές και τις χαμηλές συχνότητες του σήματος, το κυματίδιο μετασχηματίζεται χρησιμοποιώντας μια βασική συνάρτηση (mother wavelet) που τεντώνεται (κλιμακώνεται) και μετατοπίζεται κατάλληλα.

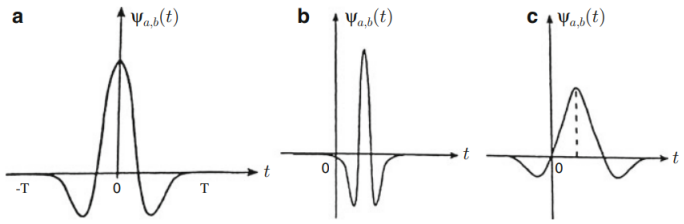
Κλιμάκωση των wavelets Η κλιμάκωση ενός κυματιδίου σημαίνει απλώς το τέντωμα (stretched) ή τη συμπίεση του (squeezed). Εισάγουμε ένα παράγοντα κλίμακας, s, έτσι ώστε $\psi_s(t) = \psi(t/s)$. Όσο μικρότερος είναι ο παράγοντας κλίμακας, τόσο πιο "συμπιεσμένο" είναι το κυματίδιο. Είναι φυσικό να σκεφτούμε μια αντιστοιχία μεταξύ των κλιμάκων κυματιδίων και της συχνότητας. Μια μικρή κλίμακα s δημιουργεί ένα συμπιεσμένο κυματίδιο. Με τη σειρά του, αυτό το συμπιεσμένο κυματίδιο καθιστά τις λεπτομέρειες να αλλάζουν γρήγορα. Κατά συνέπεια, μια μικρή κλίμακα s μπορεί να συλλάβει μια ταλάντωση υψηλής συχνότητας. Αντίθετα, υψηλή τιμή κλίμακας s μπορεί να συλλάβει κινήσεις χαμηλής συχνότητας.

Μετατόπιση των wavelets Η μετατόπιση ενός κυματιδίου σημαίνει απλώς τη μετακίνηση του σε διάφορες θέσεις. Η αριστερή μετατόπιση καλείται προώθηση ενώ η δεξιά μετατόπιση καλείται καθυστέρηση. Μαθηματικά, η καθυστέρηση μιας συνάρτησης ψ(t) κατά p παριστάνεται με $\psi(t-p)$.

Χρησιμοποιώντας αυτές τις δύο ιδιότητες, ο μετασχηματισμός κυματιδίων προσαρμόζεται έξυπνα για να συλλάβει χαρακτηριστικά σε ένα ευρύ φάσμα συχνοτήτων και συνεπώς έχει την ικανότητα να συλλαμβάνει γεγονότα που είναι τοπικά στο χρόνο. Έτσι ορίζουμε ένα wavelet με τον πιό κάτω τρόπο

$$\Psi_{s,p}(t) = \frac{1}{\sqrt{|s|}}\psi\left(\frac{t-p}{s}\right), s, p \in \mathbb{R}, s \neq 0$$

Η συνάρτηση Ψ(t) είναι το λεγόμενο μητρικό κυματίδιο (mother wavelet) και $\Psi_{s,p}(t)$ είναι "παιδί" του στην κλίμακα s και θέση p. Πιό κάτω παρουσιάζουμε ένα παράδειγμα πώς είναι το μητρικό wavelet και δύο παιδιά του που προκύπτουν για διαφορετικές τιμές των παραμέτρων s,p.



(a)Τυπικό παράδειγμα mother wavelet, (b)Συμπιεσμένο και δεξιά μετατοπισμένο wavelet $\psi_{s,p}, 0 < |s| << 1, p > 0$, (c) Επιμηκυμένο και δεξιά μετατοπισμένο wavelet $\psi_{s,p}, |s| >> 1, p > 0$

4.2 Συνεχής μετασχηματισμός κυματιδίων

Ο συνεχής μετασχηματισμός κυματιδίων (CWT) ορίζεται ως το ολοκλήρωμα, ορισμένο σε όλο το χρόνο του σήματος, πολλαπλασιασμένο με κλιμακοποιημένες και μετατοπισμένες (scaled and shifted) εκδόσεις του μητρικού κυματιδίου: $\psi(scale, position, time)$. Οι τιμές των παραγόντων κλιμάκωσης και μετάθεσης είναι συνεχείς, πράγμα που σημαίνει ότι μπορεί να υπάρχει άπειρος αριθμός κυματιδίων.

$$T(s,p) = \mathbb{W}_\psi[y](scale, position) = \int_{-\infty}^{+\infty} y_t \overline{\psi_{s,p}(Scale, Position, t)dt} =< y, \psi_{s,p} > = \frac{1}{\sqrt{s}} \int y_t \overline{\psi\left(\frac{t-p}{s}\right)} dt \tag{4.2.1}$$

Όπου $\overline{\psi_{s,p}(Scale, Position, t)}$ η συζυγής συνάρτηση της $\psi_{s,p}(Scale, Position, t)$. Με βάση λοιπόν το πιο πάνω ορισμό μπορούμε να καταλήξουμε στα εξής συμπεράσματα:

- Ο πυρήνας $\psi_{s,p}(t)$ στη πιο πάνω σχέση παίζει τον ίδιο ρόλο με το πυρήνα $e^{-i\omega t}$ στο μετασχηματισμό Fourier.

- Όπως και ο μετασχηματισμός Fourier, ο συνεχής μετασχηματισμός κυματιδίων $\mathbb{W}$ είναι γραμμικός. Ωστόσο, σε αντίθεση με το μετασχηματισμό Fourier, ο συνεχής μετασχηματισμός κυματιδίων δεν είναι μονοσήμαντα ορισμένος αλλά αποτελεί οποιοδήποτε μετασχηματισμό που λαμβάνεται για συνάρτηση Wavelet.

- Ως συνάρτηση του p (position), για συγκεκριμένη τιμή στη παράμετρο κλίμακας s, η $\mathbb{W}_\psi[y](s,p)$ αναπαριστά τη λεπτομερή πληροφορία που περιέχεται στο σήμα y(t) στην κλίμακα s.

- Ο μετασχηματισμός κυματιδίων έχει ονομαστεί "μαθηματικό μικροσκόπιο", όπου p είναι η θέση στη χρονοσειρά που "βλέπουμε" και s σχετίζεται με τη μεγέθυνση στη θέση p

Τα αποτελέσματα της CWT είναι πολλοί συντελεστές κυματιδίου T(s,p), οι οποίοι αποτελούν συνάρτηση της κλίμακας και της θέσης. Η κλίμακα και η θέση μπορούν να πάρουν οποιαδήποτε τιμή συμβατή με την περιοχή της χρονοσειράς y_t. Οι συντελεστές αυτοί $T(s,p)$ αναπαριστούν πόσο καλά το "παιδί" wavelet $\psi_{s,p}$ "συλλαμβάνει" το σήμα (goodness of fit). Τα χρονικά τμήματα όπου το κυματίδιο και το σήμα έχουν ίδιο πρόσημο έχουν ως αποτέλεσμα μια θετική συνεισφορά στο ολοκλήρωμα της εξίσωσης 4.2.1 (Περιοχές A,B στο σχήμα 4.2.1). Περιοχές όπου το σήμα και το κυματίδιο έχουν αντίθετο πρόσημο οδηγούν σε αρνητικές συνεισφορές στο ολοκλήρωμα - για παράδειγμα, οι περιοχές C, D και E στο σχήμα 4.2.1

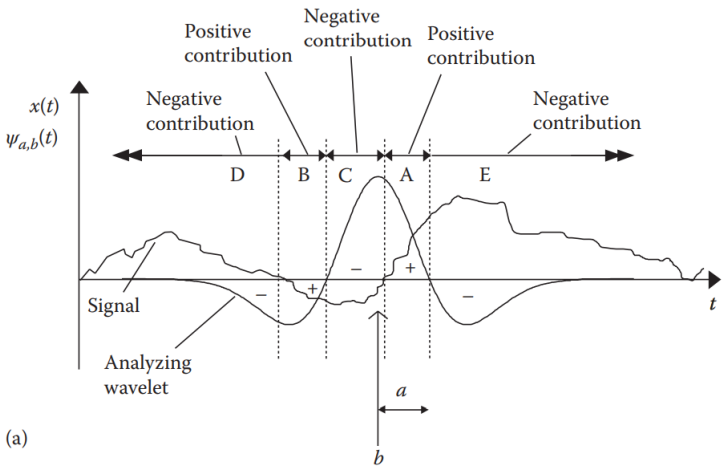


Figure 4.2.1: Ερμηνεία συντελεστών του WT

Όταν η συνάρτηση κυματιδίου είναι είτε πολύ συμπιεσμένη είτε πολύ τεντωμένη σε σύγκριση με τα χαρακτηριστικά του σήματος, ο μετασχηματισμός δίνει σχεδόν μηδενικές τιμές.

Η γραφική παράσταση των συντελεστών T(s,p) συναρτήσει των παραμέτρων s,p καλείται γράφημα μετασχηματισμού κυματιδίων (wavelet transform plot). Μπορεί να αναπαρασταθεί είτε μέσω γραφήματος επιφάνειας (γράφημα στις 3 διαστάσεις) είτε με contour plot. Πιό κάτω παρουσιάζουμε ένα τυπικό παράδειγμα.

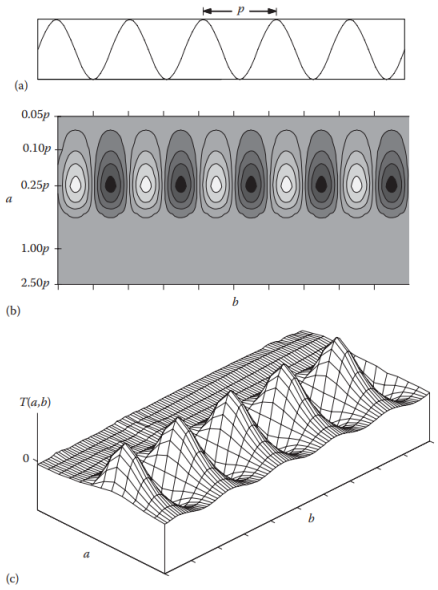


Figure 4.2.2: Γράφημα μετασχηματισμού κυματιδίων. Οι παράμετροι s,p συμβολίζονται με a,b αντίστοιχα.

Η ένταση ή το χρώμα κάθε σημείου στο διάγραμμα δείχνει το μέγεθος ή την ισχύ του συντελεστή κυματιδίου στη συγκεκριμένη θέση και κλίμακα. Συχνά σε εφαρμογές χρησιμοποιείται το γράφημα των τετραγώνων των συντελεστών, $T(s,p)^2 = E(s,p)$, που εκφράζουν την ενέργεια του κυματιδίου ως προς την υπο εξέταση χρονοσειρά στην κλίμακα s και μετατόπιση ίση με p. Το γράφημα αυτό ονομάζεται scalogram.

Ο αντίστροφος μετασχηματισμό ορίζεται ως:

$$y_t = \frac{1}{C_\psi} \int \int T(s,p)\psi \left(\frac{t-p}{s} \right) \frac{dsdp}{s^2}$$

όπου $\psi(t)$ είναι το βασικό κυματίδιο και s,p $\in \mathbb{R}$ είναι πραγματικές συνεχείς μεταβλητές. Αυτό επιτρέπει την ανάκτηση του αρχικού σήματος από το μετασχηματισμό κυματιδίου του με ολοκλήρωση σε όλες τις κλίμακες και θέσεις.

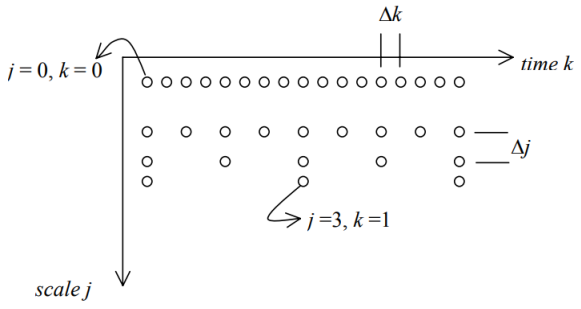
Πλεονεκτήματα της CWT:
Η CWT μπορεί να συλλάβει τόσο βραχυχρόνιες όσο και μακροχρόνιες αλλαγές στο πεδίο συχνοτήτων ενός σήματος και μπορεί να χειριστεί μη στάσιμα σήματα με μεταβαλλόμενες συνιστώσες συχνότητας. Η CWT μπορεί να παρέχει υψηλή ανάλυση τόσο στο πεδίο του χρόνου όσο και στο πεδίο της συχνότητας, επιτρέποντας την ακριβή ανίχνευση μεταβατικών γεγονότων.

Μειονεκτήματα της CWT:
Η CWT απαιτεί πολλούς υπολογιστικούς πόρους και μπορεί να είναι αργή για μεγάλα σύνολα δεδομένων. Επίσης είναι ευαίσθητη στην επιλογή της συνάρτησης wavelet και στον αριθμό των κλιμάκων που χρησιμοποιούνται στην ανάλυση.

4.3 Διακριτός μετασχηματισμός Wavelet

Όταν μιλάμε για τον Διακριτό Μετασχηματισμό κυματιδίων, η κύρια διαφορά είναι ότι ο DWT χρησιμοποιεί διακριτές τιμές για τον παράγοντα κλίμακας και μετατόπισης. Η DWT αναλύει ένα σήμα σε μια σειρά διακριτών συντελεστών κυματιδίων, οι οποίοι μπορούν να χρησιμοποιηθούν για την ανακατασκευή του αρχικού σήματος ή για την ανάλυση του συχνοτικού περιεχομένου του σήματος σε διαφορετικές κλίμακες.

Ένας τρόπος επιλογής των διακριτών τιμών κλίμακας και θέσης είναι με ένα δισδιάστατο πλέγμα, όπου το μητρικό κυματίδιο κλιμακώνεται (scaled) μέσω δυνάμεων του 2 (s=2^m) και επιμηκύνετε/συρρικνώνεται μέσω μιας σταθεράς $p = n2^m$ όπου n $\in 1, .., 2^{-m}N$ (N = το πλήθος των παρατηρήσεων). Το m τρέχει από το 0 στο J (όπου J το συνολικό πλήθος κλιμάκων που χρησιμοποιούνται). Στο πιό κάτω γράφημα αναπαριστάται το δυαδικό πλέγμα (όπου m=j , n=k).



Τα (ορθοκανονικά) δυαδικά διακριτά κυματίδια ορίζονται ως

$$\psi_{m,n}(t) = 2^{\frac{-m}{2}}\psi(2^{-m}t - n) = 2^{\frac{-m}{2}}\psi\left(\frac{t - n2^m}{2^m}\right)$$

Οι $\psi_{m,n}(t)$ επιλέγονται με τέτοιο τρόπο ώστε να σχηματίζουν μια ορθοκανονική βάση του $\mathbb{L}^2(\mathbb{R})$. Οι όροι $\{2^m\}_{0,..J}$ είναι μια ακολουθία κλιμάκων. Με αυτό το τρόπο τα κυματίδια κεντράρονται στο $n2^m$ με κλίμακα 2^m. Ο όρος $n2^m$ ονομάζεται παράμετρος μετατόπισης (shift parameter). Οι τιμές των m, n καθορίζουν το στήριγμα (το πεδίο στο οποίο η συνάρτηση είναι μή μηδενική) της συνάρτησης. Όταν το m γίνεται μεγαλύτερο, ο παράγοντας κλίμακας 2^m γίνεται μεγαλύτερος, και η συνάρτηση $\psi_{m,n}(t)$ μετατοπίζεται πιο δεξιά και γίνεται πιο διαδεδομένη (απλωμένη) και αντίστροφα όταν το m γίνεται μικρότερο. Ο όρος $2^{-m/2}$ διατηρεί τη νόρμα των συναρτήσεων βάσης ψ(t) στο 1 .

Οι συντελεστές της DWT προκύπτουν μέσω της σχέσης:

$$T_{m,n} = T(2^m, n2^m) = 2^{-m/2}\int_{-\infty}^{+\infty} y(t)\psi_{m,n}(t)dt =< y(t), \psi_{m,n}(t) >$$

Το αρχικό σήμα προκύπτει ως εξής:

$$y(t) = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} T_{m,n}\Psi_{m,n}(t) \tag{4.3.1}$$

4.4 Συνάρτηση κλιμάκωσης και αναπαράσταση πολλαπλής ανάλυσης (Scaling Function and Multiresolution Representation)

Τα ορθοκανονικά δυαδικά διακριτά κυματίδια συνδέονται με **τις συναρτήσεις κλιμάκωσης (scaling functions)** και τις **εξισώσεις διαστολής (dilation equations)** τους. Η συνάρτηση κλιμάκωσης σχετίζεται με την εξομάλυνση του σήματος (συνδέεται με ένα low-pass filter) και έχει την ίδια μορφή με το κυματίδιο, που δίνεται από τη σχέση

$$\varphi_{m,n}(t) = 2^{-m/2}\varphi(2^{-m}t - n)$$

και έχουν την εξής ιδιότητα $\int_{-\infty}^{\infty}\varphi_{0,0}(t)dt = 0$. Η συνάρτηση $\varphi_{0,0}(t) = \varphi(t)$ καλείται "father wavelet-father scaling function"

Η συνάρτηση κλιμάκωσης μπορεί να συνεληχθεί με το σήμα, να συλλάβει τις χαμηλές συχνότητες του σήματος, για να παραχθούν συντελεστές προσέγγισης ως εξής

$$S_{m,n} = \int_{-\infty}^{\infty} y(t)\varphi_{m,n}(t)dt$$

Οι συντελεστές προσέγγισης σε μια συγκεκριμένη κλίμακα m είναι συλλογικά γνωστοί ως η διακριτή προσέγγιση του σήματος σε αυτή την κλίμακα. Μια συνεχής προσέγγιση του σήματος **στην κλίμακα m** μπορεί να παραχθεί με την άθροιση μιας ακολουθίας συναρτήσεων κλιμάκωσης σε αυτή την κλίμακα.

$$y_m(t) = \sum_{n=-\infty}^{\infty} S_{m,n}\varphi_{m,n}(t)dt$$

Όπου $y_m(t)$ είναι μια ομαλή έκδοση της $y(t)$ περιορισμένη στην κλίμακα m. Όσο μικρότερο το m τόσο πιό πολύ η $y_m(t)$ τείνει να ταυτιστεί με την $y(t)$.

Επομένως ένας διαφορετικός τρόπος να αναπαραστήσουμε ένα σήμα y(t), είναι μέσω ενός συνδυασμένου αναπτύγματος σειράς, χρησιμοποιώντας τόσο τους συντελεστές προσέγγισης όσο και τους συντελεστές κυματιδίου (λεπτομέρειας) ως εξής

$$y(t) = \sum_{n=-\infty}^{\infty} S_{m_0,n}\varphi_{m_0,n}(t) + \sum_{m=-\infty}^{m_0}\sum_{n=-\infty}^{\infty} T_{m,n}\psi m, n(t) \tag{4.4.1}$$

Μπορούμε να δούμε από αυτή την εξίσωση ότι το αρχικό συνεχές σήμα εκφράζεται ως συνδυασμός μιας προσέγγισης του ίδιου του εαυτού του, σε αυθαίρετο δείκτη κλίμακας m_0, που προστίθεται σε μια διαδοχή λεπτομερειών του σήματος από τις κλίμακες m_0 έως το αρνητικό άπειρο. Η λεπτομέρεια του σήματος σε κλίμακα m δίνεται από την σχέση

$$d_m(t) = \sum_{n=-\infty}^{\infty} T_{m,n}\Psi_{m,n}(t)dt$$

από την εξίσωση (4.4.1) εύκολα προκύπτει η σχέση

$$y_{m-1}(t) = y_m(t) + d_m(t) \tag{4.4.2}$$

η οποία μας λέει ότι αν προσθέσουμε τη λεπτομέρεια του σήματος, σε μια αυθαίρετη κλίμακα (δείκτης m), στη προσέγγιση σε αυτή την κλίμακα, παίρνουμε την προσέγγιση του σήματος σε αυξημένη ανάλυση (δηλαδή σε μικρότερη κλίμακα, δείκτης m - 1). **Αυτό ονομάζεται αναπαράσταση πολλαπλής ανάλυσης**.

Στην συνέχεια δείχνουμε πώς συνδέεται η συνάρτηση κλιμάκωσης με την συνάρτηση κυματιδίου. Η εξίσωση κλιμάκωσης περιγράφει τη συνάρτηση κλιμάκωσης φ(t) ως προς συρρικνωμένες και μετατοπισμένες εκδοχές της ως εξής:

$$\varphi(t) = \sum_k c_k\varphi(2t-k)$$

Αυτή η εξίσωση μας λέει ότι μπορούμε να κατασκευάσουμε μια συνάρτηση κλιμάκωσης σε μια κλίμακα από έναν αριθμό συναρτήσεων κλιμάκωσης σε προηγούμενες κλίμακες. Οι c_k ονομάζονται συντελεστές κλιμάκωσης(scaling coefficient). Οι συναρτήσεις κλιμάκωσης συνδέονται με τις συναρτήσεις κυματιδίου μέσω της σχέσης

$$\psi(t) = \sum_k (-1)^k c_{N_k-1-k}\varphi(2t-k)$$

Όπου $N_k - 1$ είναι το πλήθος συντελεστών κλιμάκωσης που χρειάζονται για να αναπαραχθεί ένα κυματίδιο με συμπαγή στήριγμα (με αυτά ασχολούμαστε). Ορίζοντας $b_k = (-1)^k c_{N_k-1-k}$ προκύπτει ότι

$$\psi(t) = \sum_k^{N_k-1} b_k\varphi(2t-k)$$

Εκμεταλλευόμενοι τις πιο πάνω σχέσεις και μετά από πράξεις (που δίνονται αναλυτικότερα στο [17]) προκύπτουν οι σχέσεις (4.4.3) που αναπαριστούν τον αλγόριθμο αποσύνθεσης πολλαπλής ανάλυσης (Multiresolution

decomposition algorithm)

$$\begin{aligned} S_{m+1,n} &= \frac{1}{\sqrt{2}} \sum_k c_k S_{m,2n+k} = \frac{1}{\sqrt{2}} \sum_k c_{k-2n} S_{m,k} \\ T_{m+1,n} &= \frac{1}{\sqrt{2}} \sum_k b_k S_{m,2n+k} = \frac{1}{\sqrt{2}} \sum_k b_{k-2n} S_{m,k} \end{aligned} \tag{4.4.3}$$

Ως εκ τούτου, χρησιμοποιώντας αυτές τις εξισώσεις, μπορούμε να δημιουργήσουμε τους συντελεστές προσέγγισης/κυματιδίου στο δείκτη κλίμακας m+1 χρησιμοποιώντας τους συντελεστές κλιμάκωσης/κυματιδίου στην προηγούμενη κλίμακα.

Η επαναληπτική εφαρμογή των πιο πάνω εξισώσεων εκτελεί, αντίστοιχα, ένα υψιπέρατο (High-pass) και ένα βαθυπέρατο (Low-pass) φιλτράρισμα της εισόδου. Δηλαδή φιλτράρονται οι συντελεστές $S_{m,2n+k}$ για να προκύψουν οι έξοδοι $S_{m+1,n}$ και $T_{m+1,n}$. Τα διανύσματα που περιέχουν τις ακολουθίες $(1/\sqrt{2})c_k$ και $(1/\sqrt{2})b_k$ αντιπροσωπεύουν τα φίλτρα

- $(1/\sqrt{2})c_k$ είναι το βαθυπέρατο φίλτρο, που αφήνει να περάσουν οι χαμηλές συχνότητες του σήματος και συνεπώς μια εξομαλυμένη έκδοση του σήματος

- $(1/\sqrt{2})b_k$ είναι το υψιπέρατο φίλτρο, που αφήνει να περάσουν οι υψηλές συχνότητες που αντιστοιχούν στις λεπτομέρειες του σήματος.

Μπορούμε να πάμε προς την αντίθετη κατεύθυνση και να ανακατασκευάσουμε το $S_{m,n}$ από το $S_{m+1,n}$ και το $T_{m+1,n}$. Μετά από πράξεις που δίνονται στο ([17]) προκύπτει ο αλγόριθμος ανακατασκευής

$$S_{m-1,n} = \frac{1}{\sqrt{2}} \sum_k c_{n-2k} S_{m,k} + \frac{1}{\sqrt{2}} \sum_k b_{n-2k} T_{m,k}$$

Οι 2 πιό πάνω αλγόριθμοι συνθέτουν την μέθοδο Fast Wavelet Transform.

Μέχρι στιγμής, έχουμε εξετάσει τον διακριτό ορθοκανονικό μετασχηματισμό κυματιδίων ενός συνεχούς σήματος y(t), όπου αποδείχθηκε πώς η συνεχής συνάρτηση μπορεί να αναπαρασταθεί ως ένα σειριακό ανάπτυγμα συναρτήσεων κυματιδίων σε όλες τις κλίμακες και θέσεις (εξίσωση 4.3.1), ή ένα συνδυασμένο ανάπτυγμα σειράς που περιλαμβάνει τις συναρτήσεις κλίμακας και τις συναρτήσεις κυματιδίου (Εξίσωση 4.4.1). Τώρα θα θεωρούμε διακριτά σήματα εισόδου που καθορίζονται σε ακέραιες αποστάσεις, όπως ακριβώς είναι και οι χρονοσειρές δεδομένων που λαμβάνουμε. Αρχικά υποθέτουμε ότι το αντίστοιχο συνεχές σήμα $y(t)$ του διακριτού σήματος είναι γνωστό και μας χρησιμεύει για την ανάλυση.

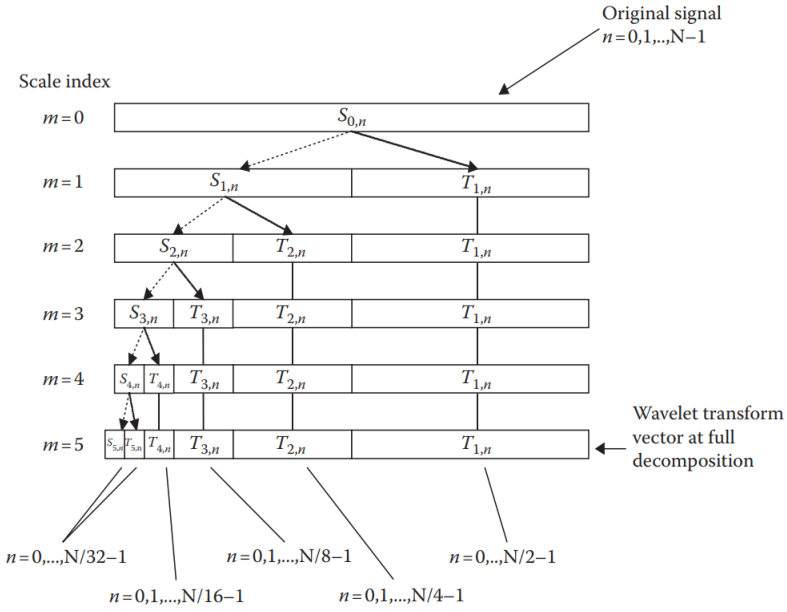
Οι συντελεστές προσέγγισης του σήματος στο δείκτη κλίμακας m=0 αντιπροσωπεύουν το πιο χονδροειδές επίπεδο προσέγγισης του αρχικού σήματος. Αυτοί οι συντελεστές συλλαμβάνουν τις συνιστώσες χαμηλής συχνότητας του σήματος και παρέχουν μια χονδρική προσέγγιση της συνολικής συμπεριφοράς του σήματος. Για να εντάξουμε το διακριτό σήμα στο πλαίσιο του αλγορίθμου πολλαπλής ανάλυσης (Multiresolution algorithm), το σήμα εισόδου στον αλγόριθμο πρέπει να είναι οι συντελεστές προσέγγισης του διακριτού σήματος σε δείκτη κλίμακας m=0. Αυτό σημαίνει ότι στο πρώτο βήμα της διαδικασίας αποσύνθεσης, το αρχικό σήμα αποσυντίθεται

σε συντελεστές προσέγγισης στην πιο χονδροειδή κλίμακα αξιοποιώντας την συνεχή μορφή του σήματος y(t).

$$S_{0,n} = \int_{-\infty}^{\infty} y(t)\varphi_{0,n}(t)$$

Όταν το συνεχές σήμα y(t) δέν είναι γνωστό αλλά γνωρίζουμε μόνο τις διακριτές τιμές του $y(t_i) = y_i, i = 0, ...N-1$ τότε θεωρούμε κατευθείαν ότι $S_{0,n} = y_n, n = 0, ..., N-1$

Αυτοί οι συντελεστές μπορούν στη συνέχεια να αποσυντεθούν περαιτέρω σε λεπτότερες κλίμακες για να καταγράψουν λεπτομερέστερες πληροφορίες σχετικά με το σήμα. Αυτό θα μας επιτρέψει να δημιουργήσουμε όλους τους επόμενους συντελεστές προσέγγισης και λεπτομέρειας, $S_{m,n}$ και $T_{m,n}$, σε δείκτες κλίμακας μεγαλύτερους από m=0 μέσω των σχέσεων (4.4.3). Αυτό μπορεί να γίνει για δείκτες κλίμακας $m > 0$, μέχρι μια μέγιστη κλίμακα που καθορίζεται από το μήκος του σήματος εισόδου. Επίσης στην πράξη, το διακριτό σήμα εισόδου $S_{0,n}$ έχει πεπερασμένο μήκος N, το οποίο είναι μια ακέραια δύναμη του 2: $N = 2^M$. Έτσι, το εύρος των κλιμάκων που μπορούμε να διερευνήσουμε είναι $0 < m < M$. Το διάνυσμα του μετασχηματισμού κυματιδίων μετά την πλήρη αποσύνθεση έχει τη μορφή $W^{(M)} = (S_M, T_M, T_{M-1}, ...T_m...T_2, T_1)$ όπου το T_m αντιπροσωπεύει το υπο-διάνυσμα που περιέχει τους συντελεστές $T_{m,n}$ σε δείκτη κλίμακας m, όπου το n κυμαίνεται από 0 έως $2^{M-m} - 1$. Μπορούμε να σταματήσουμε τη διαδικασία μετασχηματισμού πριν από την πλήρη αποσύνθεση. Αν το κάνουμε αυτό, ας πούμε σε ένα αυθαίρετο επίπεδο m_0, το διάνυσμα του μετασχηματισμού έχει τη μορφή $W^{(m_0)} = (S_{m_0}, T_{m_0}, T_{m_0-1}, ..., T_2, T_1)$. Σε αυτή τη περίπτωση, το διάνυσμα μετασχηματισμού δε περιέχει μια μόνο συνιστώσα προσέγγισης αλλά την ακολουθία των συνιστωσών προσέγγισης $S_{m_0,n}$. Πιό κάτω φαίνεται γραφικά ο υπολογισμός των συντελεστών προσέγγισης και λεπτομέρειας, $S_{m,n}$ και $T_{m,n}$.



Μπορούμε να προσεγγίσουμε λοιπόν το αρχικό σήμα μέσω της σχέσης:

$$y_0(t) = y_M(t) + \sum_{m=1}^{M} d_m(t) \tag{4.4.4}$$

όπου $d_m(t)$ είναι οι συνιστώσες λεπτομέρειας στην κλίμακα m ενώ $y_M(t)$ είναι η συνιστώσα προσέγγισης του σήματος στό μέγιστο επίπεδο αποσύνθεσης M (όπου πρέπει $N = 2^M$). Οι συνιστώσες αυτές υπολογίζονται αξιοποιώντας τους συντελεστές $S_{m,n}$ και $T_{m,n}$, που έχουν ήδη υπολογιστεί, από τις σχέσεις

$$\begin{aligned} d_m(t) &= \sum_{n=0}^{2^{M-m}-1} T_{m,n}\psi_{m,n}(t) \\ y_M(t) &= S_{M,n}\varphi_{M,n}(t) \end{aligned} \tag{4.4.5}$$

Multi-Resolution Analysis(MRA)

Η MRA, η οποία ονομάζεται και αλγόριθμος πυραμίδα, ορίζεται ως μια ιεραρχική αναπαράσταση της DWT. Βασίζεται στην αποσύνθεση του αυθεντικού σήματος(χρονοσειράς) σε m επίπεδα μέσω διαδοχικής μεταφοράς και κλιμάκωσης (translating and convolving) του μητρικού wavelet χρησιμοποιώντας χαμηλης εντάσεως και υψηλης εντάσεως φίλτρα. Αυτά τα φίλτρα διατηρούν (επιστρέφουν) τις συνιστώσες της "Λεπτομέρειας"(Detail(D)) και της "προσέγγισης"(approximation(A)). Το αρχικό σήμα μπορεί να ανασυνταχθεί με το άθροισμα όλων των συντελεστών Detail και approximation.

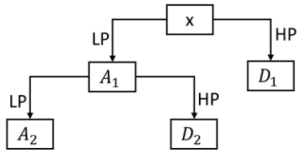


Figure 4.4.1: Multiresolution Ανάλυση . Το x είναι η αυθεντική χρονοσειρά, A_1, A_2 αναπαριστούν τις συνιστώσες προσέγγισης, D_1, D_2 αναπαριστούν του συντελεστές λεπτομέρειας. HP είναι το High-Pass φίλτρο και το LP είναι το Low-Pass φίλτρο.

Με βάση λοιπόν τα όσα αναφέρθηκαν παραπάνω αντιλαμβανόμαστε ότι στην πράξη, η DWT υλοποιείται πάντα ως τράπεζα φίλτρων (filter-bank) αξιοποιώντας τον Fast wavelet transform (και την αναπαράσταση πολλαπλής ανάλυσης). Αυτό σημαίνει ότι υλοποιείται ως καταιγισμός υψιπέρατων και βαθυπέρατων φίλτρων (high-pass and low-pass filters). Αυτό οφείλεται στο γεγονός ότι οι τράπεζες φίλτρων είναι ένας πολύ αποδοτικός τρόπος διαχωρισμού ενός σήματος σε διάφορες υπο-ζώνες συχνοτήτων. Για να εφαρμόσουμε την DWT σε ένα σήμα, ξεκινάμε με τη μικρότερη κλίμακα. Όπως είδαμε προηγουμένως, οι μικρές κλίμακες αντιστοιχούν σε υψηλές συχνότητες. Αυτό σημαίνει ότι αναλύουμε πρώτα τη συμπεριφορά των υψηλών συχνοτήτων. Στο δεύτερο στάδιο, η κλίμακα αυξάνεται με συντελεστή δύο (η συχνότητα μειώνεται με συντελεστή δύο) και αναλύουμε τη συμπεριφορά γύρω από το ήμισυ της μέγιστης συχνότητας. Στο τρίτο στάδιο, ο συντελεστής κλίμακας είναι τέσσερα και αναλύουμε συμπεριφορά συχνότητας γύρω από το ένα τέταρτο της μέγιστης συχνότητας. Και αυτό συνεχίζεται συνεχώς, μέχρι να φτάσουμε στο μέγιστο επίπεδο αποσύνθεσης.

Τι εννοούμε με το μέγιστο επίπεδο αποσύνθεσης; Για να το καταλάβουμε αυτό θα πρέπει επίσης να γνωρίζουμε ότι σε κάθε επόμενο στάδιο ο αριθμός των δειγμάτων του σήματος μειώνεται με συντελεστή δύο. Σε χαμηλότερες τιμές συχνότητας, θα χρειαστούν λιγότερα δείγματα για να ικανοποιήσετε τον ρυθμό Nyquist (Nyquist rate), οπότε δεν υπάρχει λόγος να διατηρούμε τον υψηλότερο αριθμό δειγμάτων στο σήμα (θα προκαλέσει μόνο την αύξηση του υπολογιστικού κόστους του μετασχηματισμού). Λόγω αυτής της υπο-

δειγματοληψίας, σε κάποιο στάδιο της διαδικασίας ο αριθμός των δειγμάτων στο σήμα μας θα γίνει μικρότερος από το μήκος του φίλτρου κυματιδίων και θα έχουμε φτάσει στο μέγιστο επίπεδο αποσύνθεσης. Για να δώσουμε ένα παράδειγμα, ας υποθέσουμε ότι έχουμε ένα σήμα με συχνότητες έως 1000 Hz. Στο πρώτο στάδιο χωρίζουμε το σήμα μας σε ένα τμήμα χαμηλών συχνοτήτων και ένα τμήμα υψηλών συχνοτήτων, δηλαδή 0-500 Hz και 500-1000 Hz. Στο δεύτερο στάδιο παίρνουμε το τμήμα χαμηλής συχνότητας και το χωρίζουμε πάλι σε δύο μέρη: 0-250 Hz και 250-500 Hz. Στο τρίτο στάδιο χωρίζουμε το τμήμα 0-250 Hz σε ένα τμήμα 0-125 Hz και ένα τμήμα 125-250 Hz. Αυτό συνεχίζεται μέχρι να φτάσουμε στο επίπεδο τελειοποίησης που χρειαζόμαστε ή μέχρι να εξαντλήσουμε τα δείγματα. Πιό κάτω βλέπουμε με γραφικό τρόπο πώς αυτή η διαδικασία υλοποιείται.

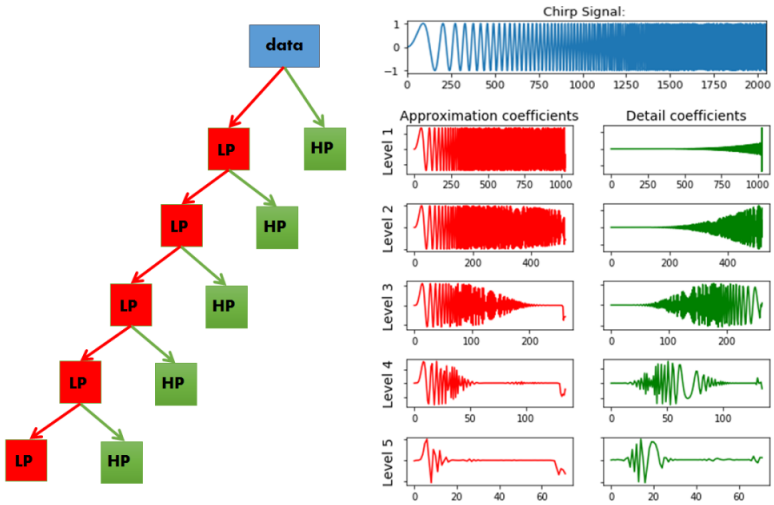


Figure 4.4.2: Οι συντελεστές προσέγγισης και λεπτομέρειας του κυματιδίου sym5 (επίπεδο 1 έως 5) που εφαρμόζεται σε ένα σήμα chirp, από το επίπεδο 1 έως 5. Αριστερά βλέπουμε μια σχηματική αναπαράσταση των υψιπέρατων και βαθυπέρατων φίλτρων που εφαρμόζονται στο σήμα σε κάθε επίπεδο.

4.5 Εφαρμογή Διακριτού μεασχηματισμού κυματιδίων με την μορφή αναπαράστασης πολλαπλής ανάλυσης

Ορίζουμε το κυματίδιο Harr που αποτελεί το πιο απλό παράδειγμα ορθογώνιου κυματιδίου. Η εξίσωση κλιμάκωσης του περιέχει μόνο 2 μή μηδενικούς συντελεστές κλιμάκωσης και δίνεται από την σχέση

$$\varphi(t) = \varphi(2t) + \varphi(2t-1)$$

επομένως οι συντελεστές κλιμάκωσης είναι οι $c_0 = c_1 = 1$. Η λύση της εξίσωσης κλιμάκωσης (συνάρτηση κλιμάκωσης-father wavelet) είναι ο απλός παλμός που δίνεται από την σχέση

$$\Phi(t) = \begin{cases} 1 & 0 \leq t < 1 \\ 0 & \text{Διαφορετικά} \end{cases}$$

Η σχέση που συνδέει αυτή την συνάρτηση κλιμάκωσης με το αντίστοιχο κυματίδιο είναι η

$$\psi(t) = \varphi(2t) - \varphi(2t-1)$$

Λύνοντας την εξισωση κυματιδίου προκύπτει το αντίστοιχο κυματίδιο Harr (μητρικό κυματίδιο-Mother wavelet) το οποίο δίνεται από την σχέση

$$\Psi(t) = \begin{cases} 1 & 0 \leq t < \frac{1}{2} \\ -1 & \frac{1}{2} \leq t < 1 \\ 0 & \Delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}$$

Επομένως οι συντελεστές b_k είναι οι $b_0 = 1, b_1 = -1$.
Πιό κάτω παρουσιάζονται γραφικά οι συναρτήσεις κλιμάκωσης και κυματιδίου καθώς και πώς αυτές προκύπτουν από τις αντίστοιχες εξισώσεις τους.

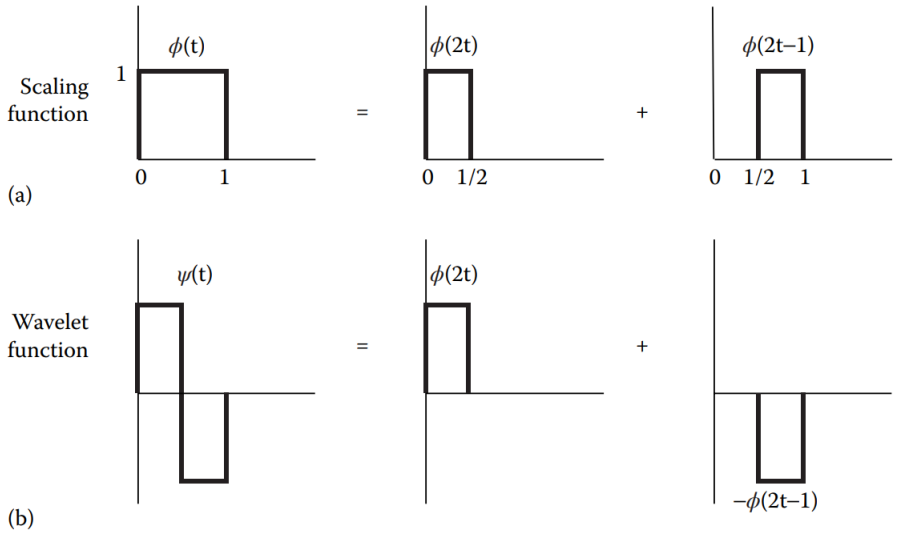


Figure 4.5.1: (a)Συνάρτηση κλιμάκωσης φ(t) . (b) συνάρτηση κυματιδίου ψ(t) από την οικογένεια κυματιδίων Harr.

Αντικαθιστώντας του συντελεστές c_0, c_1, b_0, b_1 στις εξισώσεις (4.3) λαμβάνουμε τις σχέσεις που δίνουν τους συντελσετές προσέγγισης και λεπτομέρειας στην ανώτερη κλίμακα οι οποίες είναι

$$\begin{aligned} S_{m_1,n} &= \frac{1}{\sqrt{2}}[S_{m,2n} + S_{m,2n+1}] \\ T_{m_1,n} &= \frac{1}{\sqrt{2}}[S_{m,2n} - S_{m,2n+1}] \end{aligned} \tag{4.5.1}$$

Αξιοποιώντας αυτές τις σχέσεις πραγματοποιούμε ανάλυση κυματιδίου Haar στο απλό σήμα (1,2,3,4). Επειδή το σήμα περιέχει μόνο 4 σημεία, μπορούν να πραγματοποιηθούν μόνο 2 επαναλήψεις ($2^{M=N}$) του αλγόριθμου αποσύνθεσης.

Αρχικά όπως αναφέραμε και πιο πάνω ορίζουμε

$$S_{0,n} = y_n \tag{4.5.2}$$

Επομένως ισχυεί ότι $(S_{0,1}, S_{0,2}, S_{0,3}, S_{0,4}) = (1, 2, 3, 4)$. Η πρώτη εφαρμογή του μετασχηματισμού δίνει

$$\begin{aligned} T_{1,0} &= \frac{1}{\sqrt{2}}[1-2] = \frac{-1}{\sqrt{2}} \quad , \quad T_{1,1} = \frac{1}{\sqrt{2}}[3-4] = \frac{-1}{\sqrt{2}} \\ S_{1,0} &= \frac{1}{\sqrt{2}}[1+2] = \frac{3}{\sqrt{2}} \quad , \quad S_{1,1} = \frac{1}{\sqrt{2}}[3+4] = \frac{7}{\sqrt{2}} \end{aligned} \tag{4.5.3}$$

Κατά την δεύτερη επανάληψη συμμετέχουν μόνο οι 2 συντελεστές προσέγγισης $S_{1,0}$ και $S_{1,1}$ και προκύπτουν οι συντελεστές

$$T_{2,0} = \frac{1}{\sqrt{2}}[3/\sqrt{2} - 7/\sqrt{2}] = -2 \quad , \quad S_{2,0} = \frac{1}{\sqrt{2}}[3/\sqrt{2} + 7/\sqrt{2}] = 5$$

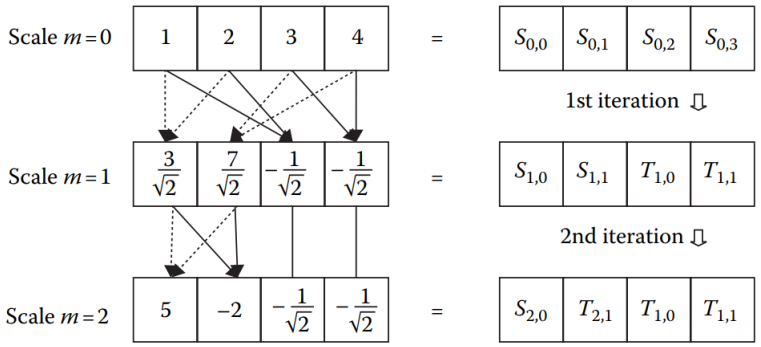


Figure 4.5.2: Γραφική αναπαράσταση διαδικασίας διακριτού μετασχηματισμού του σήματος (1,2,3,4) με το κυματίδιο Haar

Με τους συντελεστές αυτούς μπορούμε να κατασκευάσουμε τις συνιστώσες προσέγγισης και λεπτομέρειας αξιοποιώντας τις σχέσεις (4.4.5). Η συνιστώσα προσέγγισης ανώτερης κλίμακας (κλίμακας 2) δίνεται από την σχέση

$$y_2(t) = S_{2,0}\Phi_{2,0}(t) = 5.2^{-2/2}\varphi(2^{-2}t - 0) = \frac{5}{2}\mathbf{1}(\mathbf{t})_{[\mathbf{0},\mathbf{4}]} \tag{4.5.4}$$

Οι συνιστώσες λεπτομέρειας προκύπτουν από τις σχέσεις

$$\begin{aligned} d_2(t) &= T_{2,0}\psi_{2,0}(t) = (-2)2^{-\frac{2}{2}}\psi(2^{-2}t - 0) \\ &= -\left(\mathbf{1}(t)_{[0,2)} - \mathbf{1}(t)_{[2,4]}\right) \\ d_1(t) &= \sum_{n=0}^{2^{2-1}-1} T_{1,n}\psi_{1,n}(t) \\ &= T_{1,0}\psi_{1,0}(t) + T_{1,1}\psi_{1,1}(t) \\ &= \frac{-1}{\sqrt{2}}2^{-\frac{1}{2}}\psi(2^{-1}t - 0) + \frac{-1}{\sqrt{2}}2^{-\frac{1}{2}}\psi(2^{-1}t - 1) \\ &= -\frac{1}{2}\left(\mathbf{1}(t)_{[0,1)} - \mathbf{1}(t)_{[1,2]}\right) - \frac{1}{2}\left(\mathbf{1}(t)_{[2,3)} - \mathbf{1}(t)_{[3,4]}\right) \end{aligned}$$

Η προσέγγιση του αρχικού σήματος γίνεται μέσω της σχέσης (4.4.4). Επομένως το αρχικό σήμα προσεγγίζεται (εδώ προκύπτει ακριβής υπολογισμός) από την σχέση

$$y_0(t) = y_2(t) + d_2(t) + d_1(t) \tag{4.5.5}$$

Με το πιό κάτω σχήμα βλέπουμε γραφικά την ανακατασκευή του αρχικού σήματος.

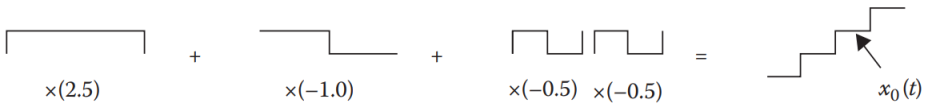


Figure 4.5.3: Γραφική ανπαράσταση προσέγγισης αρχικού σήματος από τις συνιστώσες λεπτομέρειας και προσέγγισης του μετασχηματισμού κυματιδίων

Πλεονεκτήματα της DWT:

Η DWT είναι υπολογιστικά αποδοτική και μπορεί να αναλύσει γρήγορα μεγάλα σύνολα δεδομένων. Η DWT μπορεί να χρησιμοποιηθεί για τον εντοπισμό μεταβατικών χαρακτηριστικών σε ένα σήμα και μπορεί να χειριστεί μη στάσιμα σήματα με μεταβαλλόμενες συνιστώσες συχνότητας. Η DWT μπορεί να παρέχει υψηλή ανάλυση στο πεδίο του χρόνου, επιτρέποντας την ακριβή ανίχνευση μεταβατικών γεγονότων.

Μειονεκτήματα της DWT:
Η DWT μπορεί να χάσει πληροφορίες σε υψηλότερες κλίμακες λόγω της υπο-δειγματοληψίας, γεγονός που μπορεί να επηρεάσει την ακρίβεια της ανάλυσης.

4.5.1 Wavelet Packet Decomposition(WPD)

Η WPD προτάθηκε από τους Coifman, Meyer, Wickenhauser και αποτελεί γενίκευση της Wavelet Decomposition η οποία προσφέρει υψηλότερης ποιότητας ανάλυση. Η DWT αναλύει σε κάθε επίπεδο μόνο την συνιστώσα προσέγγισης(A) ενώ η WPD αναλύει σε κάθε επίπεδο και τις 2 συνιστώσες όπως φαίνεται και στο πιό κάτω σχήμα.

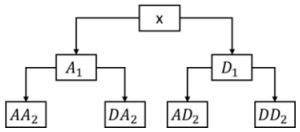


Figure 4.5.4: Αποσύνθεση του αρχικού σήματος x, 2 επιπέδων. A_1, D_1 είναι οι συνιστώσες προσέγγισης και λεπτομέρειας αντίστοιχα στο 1^o επίπεδο. AA_2 είναι η συνιστώσα προσέγγισης του A_1, DA_2 είναι η συνιστώσα λεπτομέρειας του A_1. AD_2 είναι η συνιστώσα προσέγγισης τής D_1, DD_2 είναι η συνιστώσα λεπτομέρειας τής D_1.

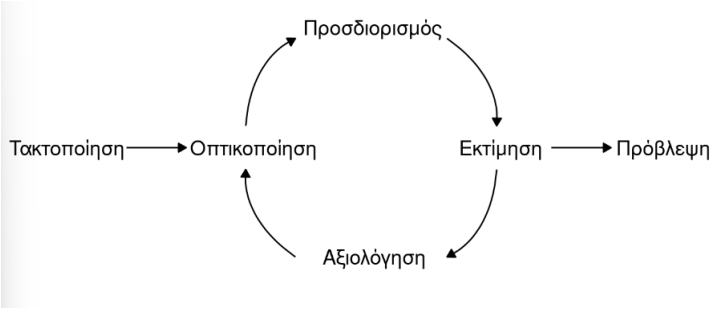
Πλεονεκτήματα της WPD:
Η WPD μπορεί να παρέχει πιο λεπτομερή ανάλυση σημάτων από την MRA, καθώς επιτρέπει την αποσύνθεση ενός σήματος σε μεγαλύτερο αριθμό ζωνών συχνοτήτων. Επιπλέον μπορεί να παρέχει μια πιο ευέλικτη αναπαράσταση του σήματος από το MRA, καθώς επιτρέπει την επιλογή συγκεκριμένων ζωνών συχνοτήτων για ανάλυση. Επίσης μπορεί να χειριστεί μη στάσιμα σήματα με μεταβαλλόμενες συνιστώσες συχνότητας.

Μειονεκτήματα της WPD:
Η WPD μπορεί να υποφέρει από υπερ-προσαρμογή, καθώς μπορεί να αποσυνθέσει το σήμα σε μεγάλο αριθμό πακέτων κυματιδίων, γεγονός που μπορεί να οδηγήσει σε ενίσχυση του θορύβου. Επίσης WPD απαιτεί την επιλογή κατάλληλων φίλτρων κυματιδίων, γεγονός που μπορεί να επηρεάσει την ακρίβεια της ανάλυσης. Τέλος η WPD μπορεί να είναι υπολογιστικά δαπανηρή για μεγάλα σύνολα δεδομένων.

5 Βασικά εργαλεία πρόβλεψης

5.1 Η ροή εργασιών κατά την διαδικασία της πρόβλεψης

Για την εξαγωγή ικανών σχημάτων πρόβλεψης δεν αρκεί απλά η επιλογή του μοντέλου, αλλα είναι μια σύνθετη διαδικασία και η συμβολή της κάθε διεργασίας είναι εξίσου σημαντική. Στο πιο κάτω σχήμα βλέπουμε σχηματικά μια τυπική διαδικασία παραγωγής προβλέψεων.



Στην συνέχεια περιγράφουμε συνοπτικά αυτές τις διαδικασίες:

Τακτοποίηση δεδομένων

Το πρώτο βήμα στην πρόβλεψη είναι η προετοιμασία των δεδομένων ώστε να αποκτήσουν τη σωστή μορφή. Αυτή η διαδικασία μπορεί να περιλαμβάνει την φόρτωση των δεδομένων, τον προσδιορισμό των τιμών που λείπουν, το φιλτράρισμα των χρονοσειρών καθώς και άλλες εργασίες προ-επεξεργασίας.

Γραφική απεικόνιση δεδομένων (οπτικοποίηση)

Η οπτικοιποίηση των δεδομένων είναι ένα πολύ σημαντικό βήμα. Αποτελεί την πρώτη επαφή με αυτά και την εξαγωγή των πρώτων και πολύ σημαντικών συμπερασμάτων. Η εξέταση των δεδομένων μας επιτρέπει να προσδιορίσουμε κοινά πρότυπα και στη συνέχεια, να καθορίσουμε ένα κατάλληλο μοντέλο.

Ορισμός μοντέλου (προσδιορισμός)

Λόγω της μεγάλης ανάγκης παραγωγής ποιοτικών προβλέψεων σχετιζόμενες με αρκετούς τομείς της κοινωνίας, έχουν αναπτυχθεί αρκετά μοντέλα πρόβλεψης και ακόμα και σήμερα η διαδικασία εύρεσης ικανών μοντέλων έχει έντονο ερευνητικό ενδιαφέρον. Ο καθορισμός του κατάλληλου μοντέλου για τα δεδομένα που έχουμε στη διάθεση μας είναι απαραίτητος για τη παραγωγή των κατάλληλων προβλέψεων.

Εκπαίδευση του μοντέλου (εκτίμηση)

Αφού καθοριστεί το κατάλληλο μοντέλο, πρέπει να το εκπαιδεύσουμε χρησιμοποιώντας κάποια από τα δεδομένα, τα οποία αποτελούν το σύνολο εκπαιδευσης (training set). Είναι σημαντικό να μην χρησιμοποιηθούν όλα τα διαθέσιμα δεδομένα για την εκπαίδευση έτσι ώστε τα εναπομείναντα δεδομένα να αξιοποιηθούν για την αξιολόγηση του μοντέλου προτού αυτό χρησιμοποιηθεί για πραγματικές προβλέψεις. Αυτό το σύνολο δεδομένων ονομάζεται σύνολο αξιολόγησης (Test/Valuation set).

Έλεγχος της απόδοσης του μοντέλου (αξιολόγηση)

Μόλις εκτιμηθεί ένα μοντέλο, είναι σημαντικό να ελεγχθεί η απόδοσή του πάνω στα δεδομένα αξιολόγησης. Η αξιολόγηση δεν γίνεται ποτέ στα δεδομένα στα οποία το μοντέλο εκπαιδεύτηκε επειδή το μοντέλο γνωρίζει ήδη τη κρυμμένη πληροφορία που αυτά έχουν και με βάση αυτά κάνει προβλέψεις. Υπάρχουν πολλά διαθέσιμα διαγνωστικά εργαλεία για τον έλεγχο της συμπεριφοράς του μοντέλου, καθώς και μέτρα για την ακρίβεια που

επιτρέπουν τη σύγκριση ενός μοντέλου με κάποιο άλλο. Αυτά θα συζητηθούν αργότερα (Δές Κεφάλαιο 5.6).

Δημιουργία προβλέψεων (πρόβλεψη)

Τελευταίο βήμα στην διαδικασία της παραγωγής προβλέψεων είναι η εφαρμογή του τελικού μοντέλου και η δημιουργία των προβλέψεων.

5.2 Τα βασικά μοντέλα πρόβλεψης

Οι πρώτες μέθοδοι που κατασκευάστηκαν για να κάνουν προβλέψεις είναι υπερβολικά απλές που κανείς μπορεί να αναρωτηθεί γιατί να αναφερθούν σε αυτή την εργασία. Ένας λόγος είναι επειδή οι ακόλουθες απλές μέθοδοι αποτελούν τη βάση για πολλές άλλες πιο σύνθετες μεθόδους όπως επίσης επειδή η απόδοση τους μπορεί να μας πει κατα πόσο ένα πιο περίπλοκο και υπολογιστικά δαπανηρό μοντέλο αξίζει να το χρησιμοποιούμε. Άλλωστε και στην ζωή δεν αποτελεί πάντα καλύτερη επιλογή το πιο περίπλοκο.

5.2.1 Average Method (Η μέθοδος της μέσης τιμής)

Σύμφωνα με αυτό το μοντέλο η πρόβλεψη **όλων** των μελλοντικών τιμών είναι ίση με τη μέση τιμή των ιστορικών δεδομένων:

$$\hat{y}_{T+h} = \bar{y} = \frac{\sum_{t=1}^{t=T} y_t}{T}$$

Στο πιο κάτω γράφημα βλέπουμε την εξέλιξη της τιμής κλεισίματος της μετοχής της εταιρίας Apple κατα την χρονική περίοδο 2/01/2019 - 26/04/2023 και τις προβλέψεις των μελλοντικών τιμών της (μπλε συμπαγής γραμμή)

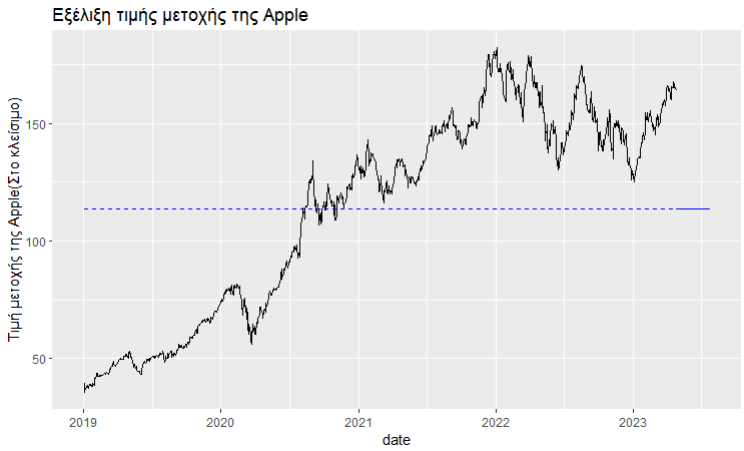


Figure 5.2.1: Προβλέψεις με τη μέθοδο της μέσης τιμής για τη τιμή κλεισίματος της μετοχής της εταιρίας Apple.

Από τη πιο πάνω γραφική εύκολα διαπιστώνουμε ότι οι προβλέψεις που κάναμε για τις 90 επόμενες τιμές κλεισίματος της μετοχής δεν είναι και τόσο καλές. Αυτό οφείλεται τόσο στο ότι το μοντέλο δεν είναι το κατάλληλο αλλά και στο γεγονός ότι τα δεδομένα που χρησιμοποιήσαμε για την εκπαίδευση του μοντέλου δεν ήταν τα καλύτερα δυνατά. Η μέση τιμή ελαττώθηκε αρκετά λόγω της επιρροής που άσκησαν πιο παλαιές τιμές (2019-2021) κατα τις οποίες η τιμή ήταν αρκετά πιο κάτω. Με άλλα λόγια το μοντέλο δέν εντόπισε την αυξητική τάση που παρατηρήθηκε κατα την χρονική περίοδο αυτή.

5.2.2 Naive Method

Πρόκειται για το πιο απλό μοντέλο πρόβλεψης που όσο παράδοξο και αν ακούγεται σε κάποιες περιπτώσεις μπορεί να έχει αρκετά καλά αποτελέσματα. Σύμφωνα με αυτό κάθε πρόβλεψη είναι ίση με τη τιμή της τελευταίας παρατήρησης. Η μέθοδος αυτή λειτουργεί αρκετά καλά σε μεταβλητές που συμπεριφέρονται όπως ένας τυχαίος περίπατος (Random Walk) και για βραχυπρόθεσμες προβλέψεις. Για αυτό τον λόγο ονομάζονται επίσης και προβλέψεις τυχαίου περιπάτου (random walk forecasts).

$$\hat{y}_{T+h} = y_T$$

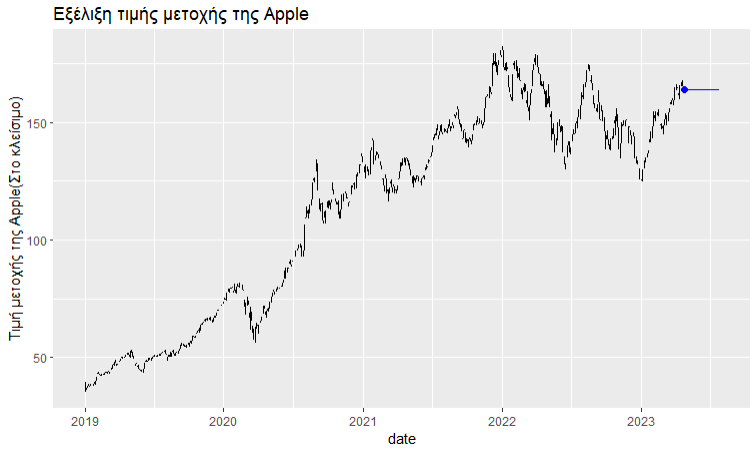


Figure 5.2.2: Προβλέψεις με τη μέθοδο naïve για τη τιμή κλεισίματος της μετοχής της εταιρίας Apple.

Παρατηρούμε ότι παρόλο που η μέθοδος αυτή είναι πολύ πιο απλή από την μέθοδο μέση τιμής, για δεδομένα μετοχών λειτουργεί πολύ καλύτερα ιδιαίτερα ότα θέλουμε μια πρόβλεψη μικρού χρονικού ορίζοντα (short-term forecast).

5.2.3 Seasonal Naive method(Εποχιακή naïve μέθοδος)

Η εποχιακή μέθοδος naïve αποτελεί μια παραλλαγή της μεθόδου Naive η οποία μπορεί να διαχειριστεί την ύπαρξη εποχικότητας σε μια χρονοσειρά. Σύμφωνα με αυτή η κάθε πρόβλεψη ορίζεται ως η τιμή της τελευταίας παρατήρησης από την ίδια εποχή.

$$\hat{Y}_{T+h} = y_{T+h-m(k+1)}, \qquad \text{όπου:}$$

- m: εποχιακή περίοδος (seasonal period).

- k: ακέραιο μέρος της ποσότητας $\frac{h-1}{m}$

Για παράδειγμα, με μηνιαία δεδομένα, η πρόβλεψη για όλες τις μελλοντικές τιμές Ιανουάριο είναι ίση με τη τελευταία τιμή που παρατηρήθηκε τον Ιανουάριο. Με τα τριμηνιαία δεδομένα, η πρόβλεψη όλων των μελλοντικών τιμών του τριμήνου Q2 είναι ίση με τη τελευταία παρατηρούμενη τιμή στο τρίμηνο Q2 (όπου Q2 σημαίνει το δεύτερο τρίμηνο). Παρόμοιοι κανόνες ισχύουν για άλλους μήνες και τρίμηνα αλλά και για άλλες εποχιακές

περιόδους. Πιο κάτω βλέπουμε τις προβλέψεις που προκύπτουν εφαρμόζοντας την εποχιακή naïve μέθοδο σε χρονοσειρά που εμφανίζει εποχικότητα.

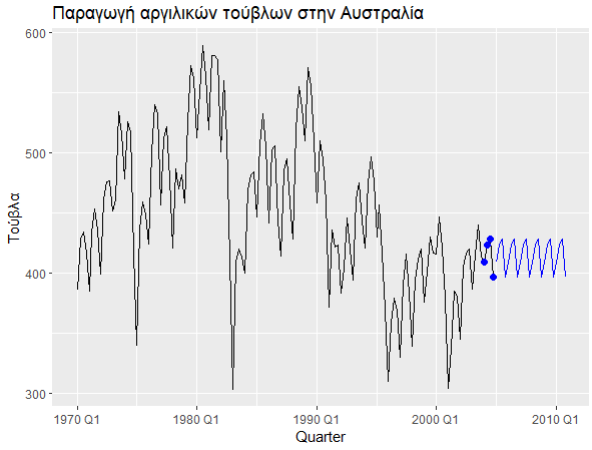


Figure 5.2.3: Προβλέψεις με την εποχιακή naïve μέθοδο της μελλοντικής παραγωγής αργιλικών τούβλων στην Αυστραλία.

5.2.4 Drift Method (Η μέθοδος περιπλάνησης)

Αποτελεί παραλλαγή της Naive μεθόδου κατα την οποία η πρόβλεψη είναι η τελευταία παρατήρηση με την προσθήκη ή την αφαίρεση ενός ποσού (το λεγόμενο drift) το οποίο ισούται με την μέση αλλαγή που εμφανίζεται στα ιστορικά δεδομένα. Δηλαδή η πρόβλεψη ορίζεται ως εξής:

$$\hat{y}_{T+h} = y_T + \frac{h}{T-1}\sum_{t=1}^{T}(y_t - y_{t-1}) = y_T + h(\frac{y_T - y_1}{T-1})$$

Επομένως με βάση την τελευταία σχέση εύκολα αντιλαμβανόμαστε ότι η πρόβλεψη αποτελεί το σημείο που ανήκει στην ευθεία που ενώνει τη πρώτη και την τελευταία παρατήρηση όταν αυτή προεκταθεί στο μέλλον.

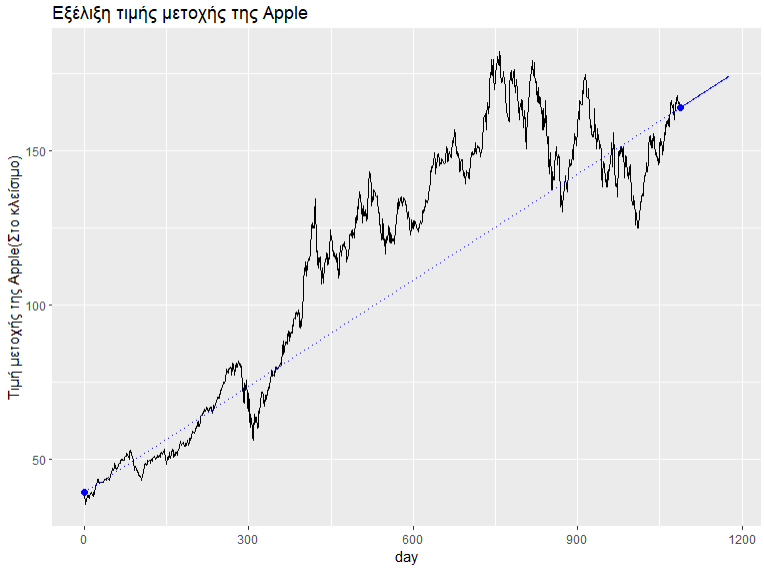


Figure 5.2.4: Προβλέψεις με τη μέθοδο περιπλάνησης για τη τιμή κλεισίματος της μετοχής της εταιρίας Apple.

5.3 Προσαρμοσμένες τιμές και υπόλοιπα

Προσαρμοσμένες τιμές

Ως προσαρμοσμένες τιμές (fitted values) ορίζουμε τη τιμή που αποδίδει το μοντέλο ως πρόβλεψη στα ιστορικά δεδομένα, αξιοποιώντας όλες τις προηγούμενες παρατηρήσεις. Τις συμβολίζουμε με $\hat{y}_{t|t-1}$ που δηλώνει την πρόβλεψη y_t με βάση τις παρατηρήσεις $y_1, \ldots, y_{t-1}$. Οι προσαρμοσμένες τιμές περιλαμβάνουν σχεδόν πάντα προβλέψεις ενός βήματος. Αξίζει να τονιστεί ότι οι προσαρμοσμένες τιμές δεν είναι αληθινές προβλέψεις αφού προκύπτουν από το μοντέλο που εκπαιδεύτηκε χρησιμοποιώντας όλες τις διαθέσιμες παρατηρήσεις των χρονοσειρών, συμπεριλαμβανομένων και των μελλοντικών παρατηρήσεων και της παρατήρησης ακριβώς την στιγμή που κάνουμε την πρόβλεψη.

Υπόλοιπα

Τα υπόλοιπα σε ένα μοντέλο χρονοσειρών είναι αυτό που απομένει μετά την προσαρμογή αυτού. Τα υπόλοιπα ισούνται με τη διαφορά μεταξύ των παρατηρήσεων και των αντίστοιχων προσαρμοσμένων τιμών:

$$e_t = y_t - \hat{y}_t$$

Όταν το μοντέλο που προσαρμόζεται χρησιμοποιεί μετασχηματισμούς των δεδομένων, είναι συχνά χρήσιμο να εξετάστούν τα υπόλοιπα σε μετασχηματισμένη κλίμακα. Τα υπόλοιπα αυτά ονομάζονται "καινοτόμα υπόλοιπα (innovation residuals)". Για παράδειγμα, ας υποθέσουμε ότι μοντελοποιήσαμε τους λογάριθμους των δεδομένων, $w_t = log(y_t)$. Τα καινοτόμα υπόλοιπα δίνονται από $w_t - \hat{w}_t$ ενώ τα κανονικά υπόλοιπα δίνονται από $y_t - \hat{y}_t$.

Τα υπόλοιπα είναι χρήσιμα για να ελέγξουμε εαν ένα μοντέλο έχει καταγράψει επαρκώς τις πληροφορίες που βρίσκονται στα δεδομένα. Μια καλή μέθοδος πρόβλεψης θα δώσει (καινοτόμα, εαν χρησιμοποιούμε μετασχηματισμό) υπόλοιπα με τις ακόλουθες ιδιότητες:

- Τα καινοτόμα υπόλοιπα να μην είναι συσχετισμένα. Εαν υπάρχουν συσχετισμοί μεταξύ τους, τότε μέρος της πληροφορίας που υπάρχει στα δεδομένα δέν αξιοποιήθηκε από το μοντέλο πρόβλεψης και παρέμεινε στα υπόλοιπα.

- Τα καινοτόμα υπόλοιπα να έχουν μηδενικό μέσο όρο. Σε αντίθετη περίπτωση οι προβλέψεις περιέχουν σφάλματα μεροληψίας.

Είναι πιθανό να υπάρχουν πολλές διαφορετικές μέθοδοι πρόβλεψης για το ίδιο σύνολο δεδομένων, οι οποίες να ικανοποιούν τις ιδιότητες αυτές. Ο έλεγχος αυτών των ιδιοτήτων είναι σημαντικός για να διαπιστωθεί εαν μια μέθοδος χρησιμοποιεί όλες τις διαθέσιμες πληροφορίες, αλλά δεν είναι καλός τρόπος για να επιλέξετε μια μέθοδο πρόβλεψης. Επίσης εαν η μέθοδος ικανοποιεί αυτές τις ιδιότητες δέν σημαίνει ότι δέν έχει περιθώρια βελτίωσης.
Εαν οι πιο πάνω ιδιότητες δεν ικανοποιούνται τότε η μέθοδος πρόβλεψης μπορεί να τροποποιηθεί για να δώσει καλύτερες προβλέψεις. Η προσαρμογή για σφάλματα μεροληψίας είναι εύκολη: εαν τα υπόλοιπα έχουν μέσο όρο m, τότε απλώς αφαιρέστε το m σε όλες τις προβλέψεις και επιλύεται το πρόβλημα των σφαλμάτων μεροληψίας. Η επίλυση του προβλήματος συσχέτισης είναι πιο δύσκολη.

Υπάρχουν και άλλες χρήσιμες ιδιότητες που ιδανικά θα θέλαμε (αλλά όχι απαραίτητες) να έχουν τα υπόλοιπα, και αυτές είναι:

- Ομοσκεδαστικότητα (homoscedasticity). Δηλαδή να έχουν σταθερή διακύμανση.

- Να ακολουθούν (τουλάχιστο προσεγγιστικά) την κανονική κατανομή.

Αυτές οι δύο ιδιότητες διευκολύνουν τον υπολογισμό των διαστημάτων πρόβλεψης. Αν δεν ισχύουν τότε είναι απαραίτητη μια εναλλακτική προσέγγιση για τον υπολογισμό των διαστημάτων πρόβλεψης.

5.4 Έλεγχος αυτοσυχέτισης Portmanteau

Όπως αναφέραμε στο (Κεφάλαιο 2.7), μέσω της εξέτασης του διαγράμματος ACF μπορούμε να ελέγξουμε για αυτοσυχετίσεις στα δεδομένα μας. Επομένως ο έλεγχος της αυτοσυχέτισης των υπολοίπων (residuals) μπορεί να γίνει μέσω του ACF γραφήματος. Όμως η εξέταση μέσω αυτό του γραφήματος εγείρει ένα σημαντικό πρόβλημα. Όταν εξετάζουμε το διάγραμμα της ACF για να δούμε αν κάθε σημείο βρίσκεται εντός των απαιτούμενων ορίων, πραγματοποιούμε εμμέσως πολλαπλούς ελέγχους υποθέσεων, καθένα με μικρή πιθανότητα να δώσει ένα ψευδώς θετικό αποτέλεσμα. Επομένως όταν πραγματοποιηθούν αρκετοί από αυτούς τους ελέγχους είναι πιθανό τουλάχιστον ένας να δώσει ένα ψευδώς θετικό αποτέλεσμα, και έτσι μπορούμε να συμπεράνουμε ότι τα υπόλοιπα έχουν κάποια εναπομένουσα αυτοσυχέτιση, ενώ στην πραγματικότητα αυτό δεν ισχύει.

Προκειμένου να ξεπεραστεί αυτό το πρόβλημα, έχουν προταθεί άλλοι έλεγχοι αυτοσυχέτισης. Οι έλεγχοι αυτοί στηρίζονται στην ιδέα **του να ελεγχθεί κατά πόσο οι πρώτες l αυτοσυχετίσεις διαφέρουν**

σημαντικά από αυτό που θα περίμενε κανείς από μια διαδικασία λευκού θορύβου (white noise). Ένας ελεγχος που εξετάζει μια ομάδα αυτοσυχετίσεων ονομάζεται **έλεγχος portmanteau** (στα γαλλικά σημαίνει βαλίτσα). Στην συνέχεια παρουσιάζουμε μερικούς γνωστούς ελέγχους Portmanteau.

Έλεγχος Box-Pierce

$$Q = T\sum_{k=1}^{l} r_k^2 \sim X_{l-m}^2,\ \text{m=πλήθος παραμέτρων προσαρμοσμένου μοντέλου}$$

Όπου l είναι η μέγιστη θεωρούμενη υστέρηση και T είναι ο συνολικός αριθμός των παρατηρήσεων.

Μεγάλες τιμές του Q υποδηλώνουν ότι οι αυτοσυχετίσεις **δεν** προέρχονται από μια σειρά λευκού θορύβου. Προτείνεται να χρησιμοποιούμε $l = 10$ για μη εποχιακά δεδομένα και $l = 2m$ για εποχιακά δεδομένα όπου m είναι η περίοδος της εποχικότητας. Ωστόσο, αυτός ο έλεγχος δεν είναι καλός όταν το l είναι μεγάλο, οπότε αν αυτές οι τιμές είναι μεγαλύτερες από $T/5$, τότε μια καλύετη τιμή για το l είναι η T/5.

Έλεγχος Ljung-Box

$$Q^* = T(T+2)\sum_{k=1}^{l}(T-k)^{-1}r_k^2$$

Και σε αυτή την περίπτωση, μεγάλες τιμές του Q^* υποδηλώνουν ότι οι αυτοσυχετίσεις δεν προέρχονται από μια σειρά λευκού θορύβου.

5.5 Διαστήματα πρόβλεψης από υπόλοιπα δειγματοθέτησης (bootstrapped residuals)

Όπως αναφέραμε πιο πάνω όταν η υπόθεση της κανονικής κατανομής των υπολοίπων είναι παράλογη τότε η κατασκευή διαστημάτων εμπιστοσύνης των προβλέψεων είναι μια δυσκολότερη διαδικασία. Ένας τρόπος κατασκευής είναι η χρήση **δειγματοθέτησης (bootstrapping), η οποία υποθέτει μόνο ότι τα υπόλοιπα είναι ασυσχέτιστα με σταθερή διακύμανση**.
Το σφάλμα της πρόβλεψης ενός βήματος ορίζεται ως

$$e_t = y_t - \hat{y}_{t|t-1}$$

Επομένως προκύπτει ότι $y_t = \hat{y}_{t|t-1} + e_t$. Έτσι μπορούμε να κατασκευάσουμε μια διαδικασία προσομοίωσης των μελλοντικών παρατηρήσεων μιας χρονοσειράς χρησιμοποιώντας την πιο κάτω σχέση.

$$y_{T+1} = \hat{y}_{T+1|T} + e_{T+1}$$

όπου $\hat{y}_{T+1}$ είναι η πρόβλεψη ενός βήματος και e_{T+1} είναι το μελλοντικό άγνωστο σφάλμα. Υποθέτοντας ότι τα σφάλματα των προβλέψεων θα είναι παρόμοια με τα προηγούμενα σφάλματα, μπορούμε να αντικαταστήσουμε το e_{T+1} με δειγματοληψία, **με επανάθεση**, από το πλήθος των σφαλμάτων που έχουμε ήδη υπολογίσει (δηλαδή, τα υπόλοιπα). Προσθέτοντας τη νέα προσομοιωμένη παρατήρηση στο σύνολο των δεδομένων μας, μπορούμε να επαναλάβουμε τη διαδικασία όσες φορές θέλουμε και ετσι προσομοιώνουμε ένα ολόκληρο σετ μελλοντικών τιμών για τη χρονοσειρά μας. Εαν επαναλάβουμε την ίδια διαδικασία k φορές τότε προσομοιώνουμε k (διαφορετικές)

εκβάσεις της χρονοσειράς στο μέλλον. Στο πιο κάτω γράφημα έχουμε δημιουργήσει 5 δειγματικές διαδρομές για τις επόμενες 30 ημέρες διαπραγμάτευσης της μετοχής της Apple .

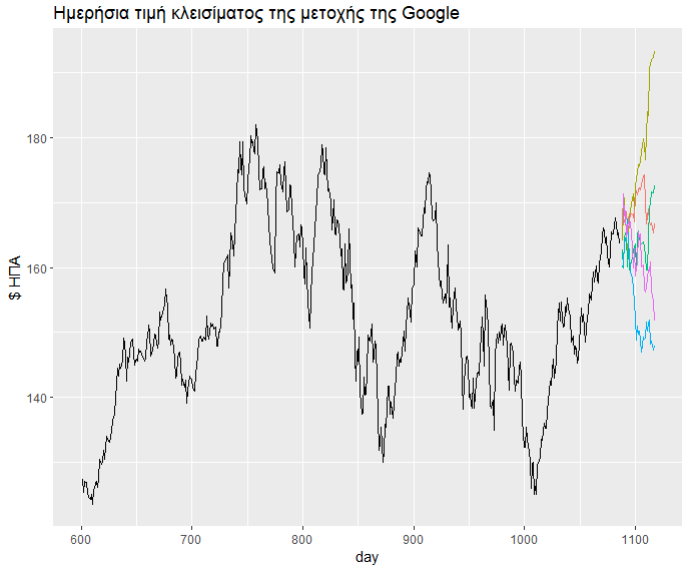


Figure 5.5.1: Πέντε προσομοιωμένες μελλοντικές δειγματικές διαδρομές της τιμής κλεισίματος της μετοχής της Apple με βάση ένα Drift μοντέλο με δειγματοθετημένα υπόλοιπα

Πώς όμως αυτές οι προσομοιώσεις μας βοηθάνε για την κατασκευή διαστημάτων πρόβλεψης; Προσομοιώνουμε πολλές μελλοντικές διαδρομές με την διαδικασία που περιγράψαμε και στην συνέχεια, υπολογίζουμε εκατοστημόρια των μελλοντικών δειγμάτων διαδρομών **για κάθε ορίζοντα πρόβλεψης** και έτσι κατασκευάζουμε διαστήματα πρόβλεψης για καθε χρονική στιγμή. Το αποτέλεσμα ονομάζεται δειγματοθετημένο (bootstrapped) διάστημα πρόβλεψης.

5.6 Αξιολόγηση ακρίβειας σημειακών προβλέψεων

Οι προβλέψεις θα πρέπει να αξιολογούνται με τη χρήση νέων δεδομένων (άγνωστων στο μοντέλο) προκειμένου να προσδιορίζεται η ακρίβειά τους. Αυτό οφείλεται στο γεγονός ότι το μέγεθος των υπολοίπων (η απόκλιση από τις πραγματικές τιμές) δεν παρέχει ακριβή ένδειξη του πόσο μεγάλα μπορεί να είναι τα πραγματικά σφάλματα πρόβλεψης καθώς αυτά προκύπτουν μέσω των δεδομένων στα οποία το μοντέλο εκπαιδεύτηκε (και επομένως τα έχει "μάθει").

Γενικά όταν κατασκευάζουμε μαθηματικά μοντέλα συνηθίζεται να διαχωρίζουμε τα συνολικά δεδομένα σε δύο σύνολα. Το σύνολο εκπαίδευσης (training set) και το σύνολο αξιολόγησης (test set). Το σύνολο εκπαίδευσης χρησιμοποιείται για την εκπαίδευση του μοντέλου, δηλαδή με βάση αυτά υπολογίζονται οι παράμετροι του μοντέλου (εαν έχει) και το σύνολο αξιολόγησης χρησιμοποιείται για να εξετάσουμε πόσο καλά αποδίδει το μοντέλο που έχουμε κατασκευάσει.

Το μέγεθος του συνόλου αξιολόγησης είναι συνήθως περίπου το 20% του συνολικού δείγματος, αν και αυτή η τιμή εξαρτάται από το μέγεθος των δεδομένων και τον ορίζοντα πρόβλεψης. Υπάρχουν εφαρμογές στις οποίες οι πιο πρόσφατες παρατηρήσεις είναι και οι πιο κρίσιμες επομένως σε αυτές τις περιπτώσεις καλύτερο είναι να συμπεριλάβουμε τις τελευταίες παρατηρήσεις στο μοντέλο και να αξιολογήσουμε το μοντέλο με άλλο τρόπο. Για παράδειγμα θα μπορούσαμε να εκπαιδεύσουμε το μοντέλο για το πρώτο 80% των δεδομένων μας, να το αξιολογήσουμε στο χρονικά επόμενο 20% και αν κριθεί καλό το μοντέλο τότε να το επανεκπαιδεύσουμε σε όλα τα δεδομένα. Το σύνολο αξιολόγησης πρέπει ιδανικά να είναι τουλάχιστον τόσο μεγάλο όσο ο μέγιστος απαιτούμενος ορίζοντας πρόβλεψης.

5.6.1 Σφάλματα πρόβλεψης

Ως σφάλμα μιας πρόβλεψης ορίζεται η διαφορά μεταξύ μιας παρατηρούμενης τιμής και της πρόβλεψης της. Εδώ, "Σφάλμα" δεν σημαίνει λάθος, σημαίνει το μέρος μιας παρατήρησης που δεν μπορεί να προβλεφθεί. Επομένως τα σφάλματα προβλέψεων δέν πρέπει να μπερδεύονται με τα υπόλοιπα (residuals). Τα υπόλοιπα υπολογίζονται στο σύνολο εκπαίδευσης ενώ τα σφάλματα πρόβλεψης υπολογίζονται στο σύνολο αξιολόγησης. Μια συνηθισμένη τακτική για να αξιολογούμε την ακρίβεια μοντέλων είναι η κατασκευή συναρτήσεων οι οποίες αθροίζουν τα σφάλματα με διαφορετικούς τρόπους. Πιο κάτω παρουσιάζουμε μερικά από τα πιό γνωστά μέτρα αξιολόγησης απόδοσης ενός μοντέλου.

- Μέσο Απόλυτο Σφάλμα: MAE $= mean(|e_t|)$
- Ρίζα του μέσου τετραγωνικού σφάλματος: RMSE $= \sqrt{mean(e_t^2)}$

Όταν θέλουμε να συγκρίνουμε την επίδοση των προβλέψεων μεταξύ διαφορετικών συνόλων δεδομένων χρειαζόμαστε μέτρα που δεν έχουν μονάδες μέτρησης. Τέτοια μέτρα βασίζονται στα

Ποσοστιαία σφάλματα: $p_t = 100\frac{e_t}{y_t}$.

Το πιο συχνά χρησιμοποιούμενο μέτρο είναι το **Μέσο Απόλυτο ποσοστιαίο σφάλμα**:

$$MAPE = mean(|p_t|)$$

Κλιμακούμενα σφάλματα

Ένας άλλος τρόπος σύγκρισης της ακρίβειας πρόβλεψης σε διαφορετικές σειρές με διαφορετικές μονάδες μέτρησης είναι μέσω των κλιμακούμενων σφαλμάτων που προτάθηκαν από τον Hyndman & Koehler (2006). Προτείνουν τη κλιμάκωση των σφαλμάτων με βάση το MAE εκπαίδευσης απο μια απλή μέθοδο πρόβλεψης. Για μη εποχιακές χρονοσειρές (εποχιακές χρονοσειρές), ένας χρήσιμος τρόπος ορισμού ενός κλιμακωτού σφάλματος είναι η χρήση naïve προβλέψεων:

$$q_j = \frac{e_j}{\frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|}$$

$$q_j = \frac{e_j}{\frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|}$$

Προφανώς το q_j είναι ανεξάρτητο από τη κλίμακα των δεδομένων καθώς αριθμητής και παρονομαστής περιλαμβάνουν τιμές στην κλίμακα των αρχικών δεδομένων. Όταν ένα κλιμακωτό σφάλμα είναι μικρότερο από 1 τότε προκύπτει από καλύτερη πρόβλεψη σε σχέση με τη naïve πρόβλεψη ενός βήματος που υπολογίζεται στα δεδομένα εκπαίδευσης. Αντίθετα όταν είναι μεγαλύτερο από 1 τότε η πρόβλεψη είναι χειρότερη από τη naïve πρόβλεψη ενός βήματος που υπολογίζεται στα δεδομένα εκπαίδευσης.

Το μέσο απόλυτο κλιμακωτό σφάλμα δίνεται από το τύπο:

$$MASE = mean(|q_j|)$$

ενώ η ρίζα του μέσου τετραγωνικού κλιμακωτού σφάλματος δίνεται από το τύπο

$$RMSSE = \sqrt{mean(q_j^2)}$$

όπου

$$q_j^2 = \frac{e_j^2}{\frac{1}{T-m}\sum_{t=m+1}^{T}(y_t - y_{t-m})^2}$$

και θέτουμε m = 1 για μη εποχιακά δεδομένα.

5.6.2 Αξιολόγηση ακρίβειας κατανεμημένων προβλέψεων

Τα μέτρα που αναφέραμε μέχρι τώρα αφορούν την ακρίβεια των σημειακών προβλέψεων. Για την αξιολόγηση των κατανεμημένων προβλέψεων απαιτείται η μελέτη άλλων μέτρων.

Μέτρο του Ποσοστημορίου (Quantile Score)

$$Q_{p,t} = \begin{cases} 2(1-p)(f_{p,t} - y_t) & y_t < f_{p,t} \\ 2p(y_t - f_{p,t}) & y_t \geq f_{p,t} \end{cases}$$

$$\text{Expect probability}(y_t \leq f_{p,t}) = p$$

Μια χαμηλή τιμή της $Q_{p,t}$ υποδεικνύει μια καλύτερη εκτίμηση του ποσοστημορίου.

Μέτρο Winkler (Winkler Score)

Είναι συχνά ενδιαφέρον να αξιολογήσουμε ένα διάστημα πρόβλεψης, παρά μερικά ποσοστημόρια, και το μέτρο του Winkler που προτείνεται από τον Winkler (1972) έχει σχεδιαστεί για αυτόν ακριβώς τον σκοπό. Εαν το διάστημα πρόβλεψης 100(1-α)% τη χρονική στιγμή t δίνεται από $[l_{a,t}, u_{a,t}]$ τότε το μέτρο Winkler ορίζεται ως το μήκος του διαστήματος συν έναν παράγοντα ποινής εαν η παρατήρηση είναι εκτός του διαστήματος:

$$\begin{cases} (u_{a,t} - l_{a,t}) + \frac{a}{2}(l_{a,t} - y_t) & y_t < l_{a,t} \\ (u_{a,t} - l_{a,t}) & l_{a,t} \leq y_t \leq u_{a,t} \\ (u_{a,t} - l_{a,t}) + \frac{a}{2}(y_t - y_{a,t}) & y_t > u_{a,t} \end{cases}$$

Για παρατηρήσεις που περιέχονται στο διάστημα, το μέτρο Winkler ισούται απλώς με το μήκος του διαστήματος. Ωστόσο, εαν η παρατήρηση βρίσκεται εκτός του διαστήματος τότε επιβάλλεται μια ποινή η οποία είναι ανάλογη με το πόσο μακριά βρίσκεται η παρατήρηση από τα όρια του διαστήματος. Όσο μικρότερο το μέτρο Winker τόσο καλύτερο το διάστημα πρόβλεψης. Βέβαια όταν γίνεται σύγκριση του ίδιου μοντέλου σε διαφορετικές χρονοσειρές με διαφορετική κλίμακα χρειάζεται ιδιαίτερη προσοχή, καθώς το πόσο μεγάλο θεωρείται το μέτρο θα πρέπει να αξιολογείται με βάση την κλίμακα των δεδομένων της χρονοσειράς.

Παρατηρήση:

Επειδή τα διαστήματα πρόβλεψης συνήθως κατασκευάζονται από τα ποσοστημόρια, ορίζοντας $l_{a,t} = f_{a/2,t}$ και $u_{a,t} = f_{1-a/2,t}$ εύκολα προκύπτει ότι το μέτρο Winker μπορεί να υπολογίζεται μέσω του Quantile Score:

$$W_{a,t} = (\frac{Q_{a/2,t} + Q_{1-a/2,t}}{a})$$

Δείκτης συνεχούς κατάταξης πιθανότητας (Continuous Ranked Probability Score) ή CRPS

Όταν ενδιαφερόμαστε να αξιολογήσουμε ολόκληρη την κατανομή των προβλέψεων παρά συγκεκριμένα ποσοστημόρια ή διαστήματα πρόβλεψης, χρησιμοποιούμε το CRPS το οποίο ορίζεται ως

$$CRPS(F, \hat{y}_{T+h}) = \int_{-\infty}^{\infty} (F(y_{T+h}) - \mathbf{1}_{\hat{y}_{T+h} \leq y_{T+h}})^2 dx$$

Όπου:

- y_{T+h} είναι η πραγματική τιμή την χρονική στιγμή T+h.
- $\hat{y}_{T+h}$ η πρόβλεψη για την χρονική στιγμή T+h
- $F(y_{T+h})$ είναι η προτεινόμενη από το μοντέλο αθροιστική κατανομή (cumulative distribution) για την παρατήρηση y_{T+h}

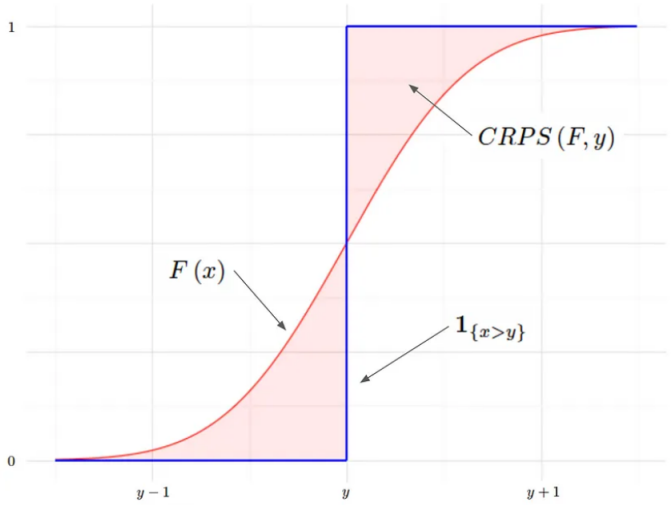


Figure 5.6.1: Γραφική αναπαράσταση του μέτρου CRPS

Συγκρίσεις χωρίς κλίμακα χρησιμοποιώντας δείκτες δεξιοτήτων

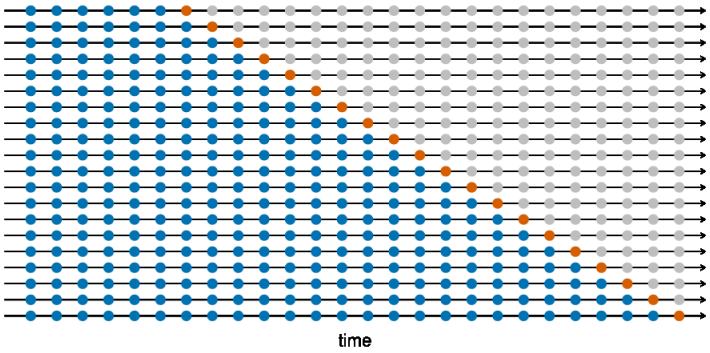
Πολύ σημαντικό προσόν ενός προγνώστη είναι να μπορεί να συγκρίνει την ακρίβεια κατανομής των διαφόρων μεθόδων σε διάφορες χρονοσειρές που έχουν **διαφορετικές κλίμακες**. Για το σκοπό αυτό, στις σημειακές προβλέψεις χρησιμοποιήσαμε κλιμακωτά σφάλματα. Μια άλλη προσέγγιση είναι να χρησιμοποιήσουμε δείκτες δεξιοτήτων (skill scores). Αυτά μπορούν να χρησιμοποιηθούν τόσο για ακρίβεια των σημειακών προβλέψεων όσο και για ακρίβεια των κατανεμημένων προβλέψεων. Με τους δείκτες δεξιοτήτων, συγκρίνουμε πόσο ακριβής είναι η πρόβλεψη σε σχέση με την πρόβλεψη μιας μεθόδου αναφοράς (μια απλή μέθοδος) ως προς κάποιο μέτρο απόδοσης, που επιλέγουμε εμείς. Ο δείκτης δεξιοτήτων CRPS ορίζεται ως εξής:

$$\frac{CRPS_{MA} - CRPS_M}{CRPS_{MA}}$$

Όπου $CRPS_{MA}$ το μέτρο CPRS για την μέθοδο αναφοράς ενώ $CRPS_M$ αναφέρεται στο μοντέλο που χρησιμοποιούμε και θέλουμε να το αξιολογήσουμε. Το ποσοστό που προκύπτει εκφράζει πόσο της εκατό καλύτερο (ως προς τον μέτρο CRPS) είναι το μοντέλο σε σχέση με το μοντέλο αναφοράς.

5.6.3 Διασταυρούμενη επικύρωση χρονοσειρών(Time series cross-validation)

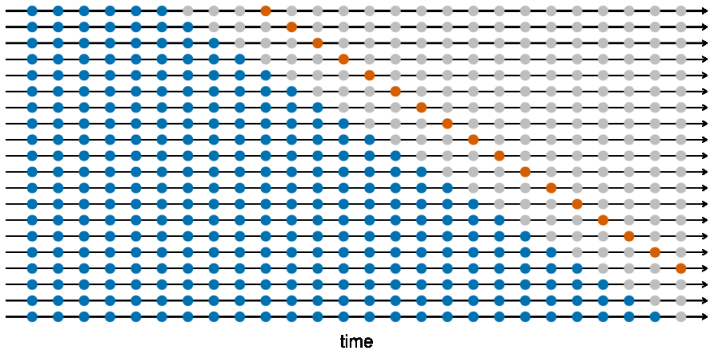
Για δεδομένα χρονοσειρών, τα σύνολα εκπαίδευσης και αξιολόγησης συνηθίζεται και είναι αποδοτικότερο να προκύπτουν από διασταυρούμενη επικύρωση της χρονοσειράς. Μια μέθοδος διασταυρούμενης επικύρωσης κατάλληλη για δεδομένα χρονοσειρών είναι η "Walk forward rolling/expanding window". Σε αυτήν τη διαδικασία, υπάρχει μια σειρά από σύνολα αξιολόγησης, καθένα από τα οποία αποτελείται από μία μόνο παρατήρηση. Το αντίστοιχο σύνολο εκπαίδευσης αποτελείται μόνο από παρατηρήσεις που πραγματοποιήθηκαν πριν από την παρατήρηση που σχηματίζει το σύνολο αξιολόγησης. Ο τρόπος σχηματισμού των συνόλων αξιολόγησης και εκπαίδευσης φαίνεται στο πιο κάτω γράφημα.



Όπως παρατηρούμε στο πιο πάνω γράφημα, εαν δεν έχουμε επαρκείς παρατηρήσεις στο σύνολο δεδομένων μας, τότε τα πρώτα (μονο-)σύνολα αξιολόγησης δεν αξιοποιούνται καθώς τα αντίστοιχα σύνολα εκπαίδευσης, λόγω του μικρού τους μεγέθους, δεν είναι ικανά να εκπαιδεύσουν επαρκώς το μοντέλο.

Η ακρίβεια πρόβλεψης υπολογίζεται εκτιμώντας το μέσο όρο στα σύνολα αξιολόγησης. Με αυτή τη διαδικασία σχεδόν κάθε στοιχείο από το σύνολο δεδομένων μας συμμετέχει στην αξιολόγηση του μοντέλου. Αυτή η διαδικασία είναι μερικές φορές γνωστή ως "αξιολόγηση σε κυλιόμενο σημείο έναρξης πρόβλεψης (evaluation on

a rolling forecasting origin)", επειδή η "έναρξη" στην οποία αρχίζει η πρόβλεψη προχωρά μπροστά στο χρόνο. Όταν σκοπεύουμε να κάνουμε προβλέψεις πολλών βημάτων, τότε η διαδικασία σχηματισμού των συνόλων αξιολόγησης διαφοροποιείται ελάχιστα. Έστω ότι ενδιαφερόμαστε για μοντέλα που παράγουν καλές προβλέψεις 4-βημάτων προς τα εμπρός. Το αντίστοιχο διάγραμμα φαίνεται παρακάτω



Υπάρχουν και άλλες μέθοδοι διασταυρούμενης επικύρωσης κατάλληλες για δεδομένα χρονοσειρών με τις πιό γνωστές να είναι οι

- Purged k-fold Cross-Validation

- Combinatorial purged Cross-Validation. Είναι αρκετά χρήσιμη για εφαρμογές της Μηχανικής μάθησης στα χρηματοοικονομικά.

6 Μοντέλα εκθετικής εξομάλυνσης

Οι μέθοδοι εκθετικής εξομάλυνσης παράγουν προβλέψεις που είναι σταθμισμένοι μέσοι προηγούμενων παρατηρήσεων, με τα βάρη να φθίνουν εκθετικά καθώς οι παρατηρήσεις γίνονται παλιότερες. Με αυτό το τρόπο, όσο πιο πρόσφατη είναι μία παρατήρηση τόσο περισσότερο συμβάλει στον καθορισμό της πρόβλεψης.

6.1 Απλή εκθετική εξομάλυνση (SES)

Η Απλή εκθετική εξομάλυνση είναι κατάλληλη για την πρόβλεψη δεδομένων χωρίς σαφή τάση ή εποχιακό μοτίβο. Οι προβλέψεις από το μοντέλο απλής εκθετικής παράγονται από τον πιο κάτω τύπο

$$\hat{y}_{T+1|T} = ay_T + a(1-a)y_{T-1} + a(1-a)^2 y_{T-2} + ... \tag{6.1.1}$$

όπου $0 \leq a \leq 1$ είναι η παράμετρος εξομάλυνσης.

Δηλαδή η πρόβλεψη ενός βήματος για τη χρονική T+1 είναι ένας σταθμισμένος μέσος όρος όλων των παρατηρήσεων της χρονοσειράς, με τις πιο πρόσφατες παρατηρήσεις να λαμβάνονται περισσότερο υπόψην απ'ότι οι παλαιότερες. Ο ρυθμός με τον οποίο μειώνονται τα βάρη ελέγχεται από την παράμετρο α. Στις ειδικές περιπτώσεις που

- α=1 τότε προκύπτει η μέθοδος naive

- α=1/N (όπου N το μήκος της χρονοσειράς) προκύπτει η μέθοδος του μέσου

Για οποιοδήποτε α μεταξύ 0 και 1, τα βάρη που συνδέονται με τις παρατηρήσεις μειώνονται εκθετικά καθώς πηγαίνουμε στο παρελθόν, έτσι εξηγείται και το όνομα "εκθετική εξομάλυνση".

Παράδειγμα: Πρόβλεψη πωλήσεων Ιουνίου

Έστώ ότι η παράμετρος εξομάλυνσης επιλέγεται ίση με 0.3 (a=0.3)

$$\hat{S}_6 = 0.3S_5 + 0.21S_4 + 0.147S_3 + 0.1029S_2 + ...$$

όπου οι συντελεστές προέκυψαν με τους υπολογισμούς:
$a(1-a)^0 = 0.3, a(1-a)^1 = 0.21, a(1-a)^2 = 0.147, a(1-a)^3 = 0.1029$

Όταν χρησιμοποιούμε εκθετική εξομάλυνση σε κάποια γλώσσα προγραμματισμού, η παράμετρος α συνήθως επιλέγεται με βάση ένα κριτήριο βελτιστοποίησης. Συνήθως επιλέγουμε το βέλτιστο α για το οποίο το SSE (άθροισμα τετραγώνων των υπολοίπων) γίνεται ελάχιστο. Σε κάποιες περιπτώσεις η παράμετρος α επιλέγεται με βάση την εμπειρία του προγνώστη.

Υπάρχουν ακόμα 2 μορφές με τις οποίες μπορούμε να αναπαραστήσουμε την πιο πάνω εξίσωση:

1. **Σταθμισμένη μορφή μέσου:**

$$\hat{y}_{T+1|T} = \sum_{j=0}^{T-1} a(1-a)^j y_{T-j} + (1-a)^T l_0 \tag{6.1.2}$$

Ο τελευταίος όρος γίνεται πολύ μικρός για μεγάλα T. Έτσι, η σταθμισμένη μορφή του μέσου οδηγεί στην ίδια εξίσωση πρόβλεψης (6.1.1)

2. **Συνιστώσα μορφή:**

Εξίσωση Πρόβλεψης: $\hat{y}_{t+h|t} = l_t$

Εξίσωση εξομάλυνσης: $l_t = ay_t + (1-a)l_{t-1}$

Όπου l_t είναι το επίπεδο (τιμή εξομάλυνσης) της σειράς τη χρονική στιγμή t. Για την εφαρμογή του αλγορίθμου της συνιστώσας μορφής, πέραν του καθορισμού της παραμέτρου εξομάλυνσης α, απαιτείται ο καθορισμός της αρχικής τιμής του επιπέδου, l_0. Συχνά ως αρχική τιμή επιλέγεται η πρώτη παρατήρηση y_0. Ωστόσο, ένας πιο αξιόπιστος και αντικειμενικός τρόπος καθορισμού της l_0 είναι η εκτίμησή της, μαζί με την παράμετρο εξομάλυνσης α, από τα παρατηρούμενα δεδομένα ελαχιστοποιώντας συνήθως το SSE. Σε αντίθεση με την περίπτωση παλινδρόμησης (όπου έχουμε τύπους που επιστρέφουν τις τιμές των συντελεστών παλινδρόμησης που ελαχιστοποιούν το SSE), αυτό συνεπάγεται ένα μη γραμμικό πρόβλημα ελαχιστοποίησης και πρέπει να χρησιμοποιήσουμε ένα εργαλείο βελτιστοποίησης για να το λύσουμε. Πιο συγκεκριμένα, χρησιμοποιούνται αλγόριθμοι βελτιστοποίησης για την αναζήτηση των βέλτιστων παραμέτρων που ελαχιστοποιούν το επιλεγμένο κριτήριο. Αυτοί οι αλγόριθμοι προσαρμόζουν επαναληπτικά τις τιμές των παραμέτρων έως ότου το κριτήριο ελαχιστοποιηθεί ή φθάσει σε ένα ικανοποιητικό επίπεδο.

Παρατήρηση:
Η απλή εκθετική εξομάλυνση έχει "επίπεδη" συνάρτηση πρόβλεψης. Δηλαδή, όλες οι προβλέψεις παίρνουν την ίδια τιμή, ίση με τη συνιστώσα του τελευταίου επιπέδου. Για αυτό ακριβώς τον λόγο οι προβλέψεις θα είναι κατάλληλες μόνο εαν οι χρονοσειρές δεν έχουν τάση ή συνιστώσα εποχικότητας. Στην πράξη η απλή εκθετική εξομάλυνση (και γενικότερα τα μοντέλα εκθετικής εξομάλυνσης) χρησιμοποιούνται για να παράγουν προβλέψεις ενός βήματος, και με κάθε νέα παρατήρηση προβλέπουν την επόμενη χρονική στιγμή.

Για παράδειγμα έστω ότι έχουμε τα ιστορικά δεδομένα πωλήσεων από την χρονική στιγμή T=5 εώς και την αρχική χρονική στιγμή t=0 και αυτά είναι: 35,38,39,42,45. Θέλουμε ένα μοντέλο που με το τέλος της κάθε εβδομάδας να προβλέπει τις προβλέψεις της επόμενης. Στον επόμενο πίνακα φαίνονται οι πραγματικές τιμές που προκύπτουν για την κάθε εβδομάδα (εβδομάδα 6 εώς 8) και οι αντίστοιχες προβλέψεις των τιμών τους.

Table 6.1.1: Παράδειγμα εφαρμογής μεθόδου απλής εκθετικής εξομάλυνσης με αρχική τιμή επιπέδου $l_0 = y_0$

Εβδομάδα	Πωλήσεις	Πρόβλεψη
6	39	$\hat{y}_{5+1\|5} = a35 + a(1-a)38 + a(1-a)^2 39 + a(1-a)^3 42 + a(1-a)^4 45 = l_5$
7	44	$\hat{y}_{6+1\|6} = a39 + (1-a)l_5 = l_6$
8	40	$\hat{y}_{7+1\|7} = a44 + (1-\alpha)l_6$

6.2 Μέθοδοι με τάση (Holt's Trend-Double exponential smoothing)

Ο Holt (1957) επέκτεινε την απλή εκθετική εξομάλυνση για να επιτρέψει την πρόβλεψη δεδομένων με τάση

$$\text{Εξίσωση Πρόβλεψης:} \quad \hat{y}_{t+h|t} = l_t + hb_t$$
$$\text{Εξίσωση Επιπέδου:} \quad l_t = ay_t + (1-a)(l_{t-1} + b_{t-1})$$
$$\text{Εξίσωση Τάσης:} \quad b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}$$

όπου η l_t δηλώνει μια εκτίμηση του επιπέδου της σειράς τη χρονική στιγμή t, η b_t υποδηλώνει μια εκτίμηση της τάσης (κλίση) της σειράς τη χρονική στιγμή t, η α είναι η παράμετρος εξομάλυνσης για το επίπεδο και β^* είναι η παράμετρος εξομάλυνσης για την τάση ($\alpha,\beta^* \in [0,1]$). Η πρόβλεψη κατά h βήματα μπροστά είναι ίση με το τελευταίο εκτιμώμενο επίπεδο συν h φορές την τελευταία εκτιμώμενη τιμή τάσης. Συνεπώς, οι προβλέψεις απότελούν μια γραμμική συνάρτηση του h .

Για την εφαρμογή του αλγορίθμου απαιτείται ο υπολογισμός των παραμέτρων $\alpha, \beta^*, l_0, b_0$, ο οποίος γίνεται ελαχιστοποιώντας το SSE για τα σφάλματα εκπαίδευσης ενός βήματος όπως ακριβώς στην μέθοδο απλής εκθετικής εξομάλυνσης. Αρκετές φορές οι αρχικές τιμές των συνιστωσών επιπέδου και τάσης l_0, b_0 υπολογίζονται από τις σχέσεις $l_0 = y_0, b_0 = y_0 - y_{-1}$

Table 6.2.1: Παράδειγμα εφαρμογής μεθόδου εκθετικής εξομάλυνσης με τάση με τις αρχικές τιμές συνιστωσών να υπολογίζονται από τις σχέσεις $l_0 = y_0, b_0 = y_0 - y_{-1}$. Με κόκκινο χρώμα είναι οι αρχικές εκτιμήσεις των συνιστωσών

Εβδομάδα	Πωλήσεις	$l_T(a=0.7)$	$b_T(\beta^*=0.6)$	y_{T+1}
-1	9.75	-	-	-
0	10.09	10.09	10.09-9.75=0.34	-
1	10.27	$a10.27 + (1-a)(10.09+0.34) = 10.32$	$\beta^*(10.32-10.09) + (1-\beta^*)0.34 = 0.272$	10.09+0.34=10.43
2	8.98	$a8.98 + (1-a)(10.32+0.272) = 9.46$	$\beta^*(9.46-10.32) + (1-\beta^*)0.272 = -0.407$	10.32+0.272=10.59
3	8.79	$a8.79 + (1-a)(9.46-0.407) = 8.87$	$\beta^*(8.87-9.46) + (1-\beta^*)(-0.407) = -0.517$	9.46-0.407=9.05
4	8.23	$a8.23 + (1-a)(8.87-0.517) = 8.27$	$\beta^*(8.27-8.87) + (1-\beta^*)(-0.517) = -0.567$	8.87-0.517=8.353

Ένα σημαντικό πρόβλημα που έχει η μέθοδος γραμμικής τάσης του Holt είναι ότι εμφανίζει μια επ' αόριστον σταθερή τάση (αύξουσα ή φθίνουσα) στο μέλλον γεγονός που οδηγεί σε υπερβολικές προβλέψεις, ειδικά για μεγαλύτερους ορίζοντες πρόβλεψης. Με αφορμή αυτή την παρατήρηση, ο Gardner McKenzie (1985) εισήγαγε παράμετρο που "αποσβένει" την τάση σε μια επίπεδη γραμμή κάποια στιγμή στο μέλλον. Η μέθοδος ονομάζεται
Damped Holt's method (Εκθετική εξομάλυνση με αποσβένουσα τάση) και ορίζεται ως:

$$\text{Εξίσωση Πρόβλεψης:} \quad \hat{y}_{t+h|t} = l_t + (\varphi + \varphi^2 + ... + \varphi^h)b_t$$
$$\text{Εξίσωση Επιπέδου:} \quad l_t = ay_t + (1-a)(l_{t-1} + b_{t-1})$$
$$\text{Εξίσωση Τάσης:} \quad b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}$$

Παρατηρούμε ότι η μέθοδος αυτή περιλαμβάνει επίσης μια παράμετρο απόσβεσης $\varphi \in (0,1)$ η οποία αποσβένει την τάση έτσι ώστε να πλησιάζει μια σταθερά (λόγω της σύγκλισης της γεωμετρικής σειράς $\varphi + \varphi^2 + ... + \varphi^h$) κάποια στιγμή στο μέλλον. Στην ειδική περίπτωση όπου φ=1 τότε η μέθοδος είναι ίδια με την γραμμική μέθοδο Holt. Αν παρατηρήσουμε καλύτερα την σχέση που δίνει τις προβλέψεις βλέπουμε

ότι αυτές συγκλίνουν στην $l_t + \frac{\varphi}{1-\varphi} b_T$ καθώς το $h \to \infty$ για οποιαδήποτε τιμή του $0 < \varphi < 1$. Επομένως οι βραχυπρόθεσμες προβλέψεις έχουν τάση ενώ οι μακροπρόθεσμες προβλέψεις δεν έχουν τάση. Στην πράξη καλό είναι το φ να διατηρείται πάντα πάνω από 0.8 αλλά όχι εντελώς κοντά στο 1 καθώς όπως αναφέραμε όταν φ=1 τότε προκύπτει ένα μοντέλο χωρίς απόσβεση.

Στην συνέχεια παρουσιάζουμε ένα παράδειγμα στο οποίο προσαρμόζουμε τις μεθόδους: Απλής εκθετικής εξομάλυνσης ("SES"), την μέθοδο του Holt (εκθετική εξομάλυνση με τάση) και την Damped Holt's method. Σε αυτό, τα δεδομένα είναι οι ετήσιες εξαγωγές της Αυστραλίας από το 1960 εώς και το 2017.

Χρησιμοποιείται διασταυρούμενη επικύρωση χρονοσειρών για να συγκριθεί η ακρίβεια πρόβλεψης ενός βήματος και στις τρεις μεθόδους. Σημειώνουμε ότι οι παράμετροι για κάθε μοντέλο επιλέχθηκαν να είναι οι βέλτιστες ως προς την ελαχιστοποίηση του SSE (άθροισμα τετραγωνικών σφαλμάτων), αξιοποιώντας τα υπόλοιπα που προκύπτουν $(y_t - \hat{y}_{t|t-1})$. Στον πιο κάτω πίνακα φαίνονται οι τιμές διάφορων μέτρων ικανότητας για τα μοντέλα με βάση τα οποία επιλέγεται το καταλληλότερο μοντέλο.

Table 6.2.2: Μέτρα ικανότητας για τα 3 μοντέλα εκθετικής Εξομάλυνσης

Μοντέλο	ME	RMSE	MAE	MAPE
Damped	0.5010322	1.618849	1.309446	7.439780
Holt	0.3904014	1.627631	1.320365	7.535499
SES	0.6484328	1.600030	1.280937	7.204111

Με βάση τα πιο πάνω μέτρα βλέπουμε ότι δεν υπάρχουν ουσιαστικές διαφορές στα 3 μοντέλα. Για την επιλογή λοιπόν του καταλληλότερου μοντέλου κατασκευάζεται η γραφική με τις προβλέψεις των εξαγωγών για τα επόμενα χρόνια και από τα 3 μοντέλα.

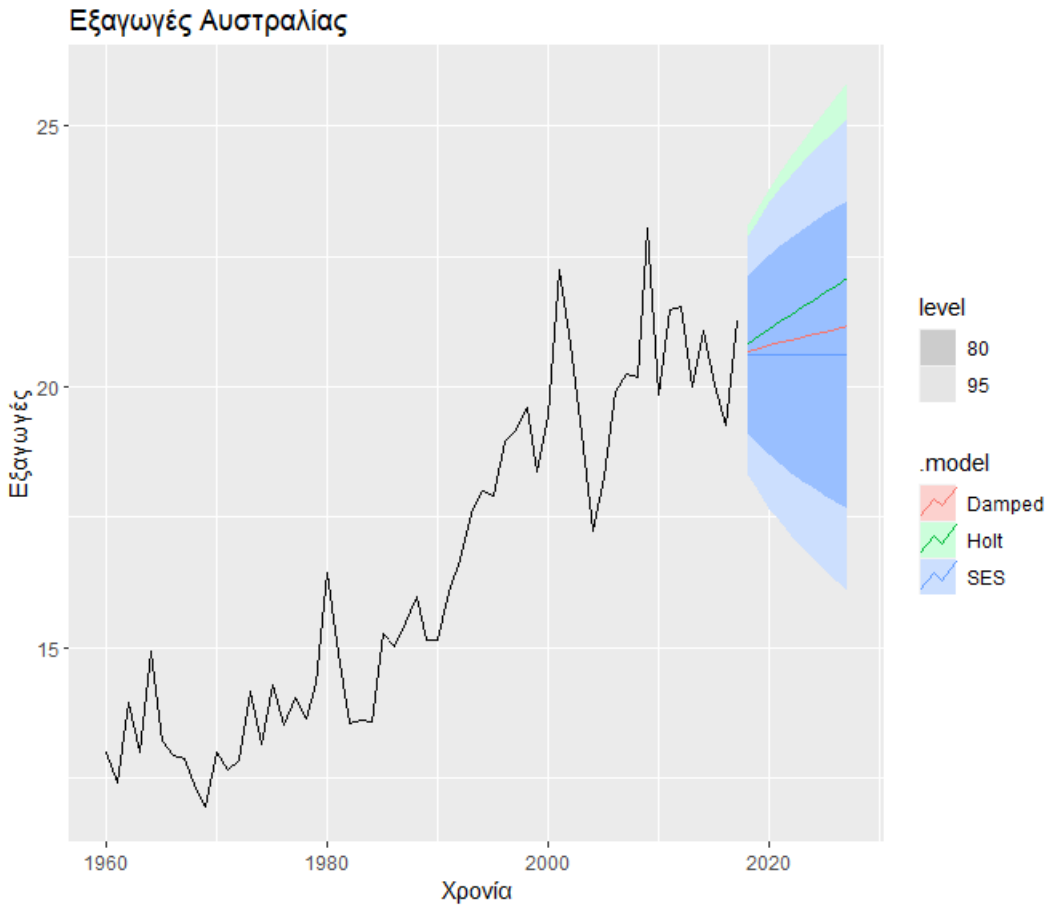


Figure 6.2.1: Οι προβλέψεις εξαγωγών για τα επόμενα 10 χρόνια με βάση και τα 3 μη εποχιακά μοντέλα εκθετικής εξομάλυνσης

Παρατηρούμε ότι η μέθοδος απλής εκθετικής εξομάλυνσης δεν καταφέρνει να αποδώσει στις προβλέψεις την αυξητική τάση των εξαγωγών ενώ η μέθοδος του Holt φαίνεται να υπερεκτιμά τις μελλοντικές προβλέψεις αφου όπως αναφέραμε και πιο πάνω εμφανίζει μια επ' αόριστον σταθερη αυξητική τάση. Η μέθοδος εκθετικής εξομάλυνσης με αποσβένουσα τάση φαίνεται να παράγει τις καλύτερες και πιο ασφαλείς προβλέψεις αφου λαμβάνει υπόψη την αυξητική τάση την οποία όμως δεν θεωρεί δεδομένη στο μέλλον (έτσι την αποσβένει) με αποτέλεσμα οι προβλέψεις για τις εξαγωγές να είναι μικρότερες από αυτές της μεθόδου του Holt.

6.3 Μέθοδοι με εποχικότητα

Η εποχιακή μέθοδος εκθετικής εξομάλυνσης των Holt-Winters μπορεί να αποτυπώσει και την εποχικότητα. Αποτελείται από την εξίσωση πρόβλεψης και τρεις εξισώσεις εξομάλυνσης: Μια για το επίπεδο l_t, μια για την τάση b_t και μια για την συνιστώσα της εποχικότητας s_t, με αντίστοιχες παραμέτρους εξομάλυνσης α, β^* και γ.

Υπάρχουν δύο παραλλαγές της μεθόδου, η προσθετική μέθοδος και η πολλαπλασιαστική. Ανάλογα με το ποια μέθοδο χρησιμοποιούμε, αλλάζει και η φύση (η ερμηνεία) της συνιστώσας της εποχικότητας.

Εποχιακή εκθετική εξομάλυνση με την προσθετική μέθοδο

- Η προσθετική μέθοδος προτιμάται όταν οι εποχιακές διακυμάνσεις είναι περίπου σταθερές σε όλη τη σειρά .

- Η συνιστώσα της εποχικότητας εκφράζεται σε απόλυτους όρους στην κλίμακα της παρατηρούμενης σειράς

- Στην εξίσωση επιπέδου η σειρά προσαρμόζεται εποχιακά αφαιρώντας την συνιστώσα της εποχικότητας

Εποχιακή εκθετική εξομάλυνση με την πολλαπλασιαστική μέθοδο

- Η πολλαπλασιαστική μέθοδος προτιμάται όταν οι εποχιακές διακυμάνσεις μεταβάλλονται ανάλογα με το επίπεδο της σειράς.
- Η συνιστώσα της εποχικότητας εκφράζεται σε σχετικούς όρους (ποσοστά).

- Η σειρά προσαρμόζεται εποχιακά διαιρώντας με την εποχιακή συνιστώσα.

Προσθετική μέθοδος των Holt-Winters
Η συνιστώσα μορφή της προσθετικής μεθόδου είναι:

$$\begin{aligned}\hat{y}_{t+h|t} &= l_t + hb_t + s_{t+h-m(k+1)}\\ l_t &= a(y_t - s_{t-m}) + (1-a)(l_{t-1} + b_{t-1})\\ b_t &= \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}\\ s_t &= \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} = \gamma^*(1-\alpha)(y_t - l_{t-1} - b_{t-1}) + [1-\gamma^*(1-\alpha)]s_{t-m}\end{aligned}$$

όπου k είναι το ακέραιο μέρος της (h-1)/m . Ο δείκτης χρόνου στην συνιστώσα της εποχικότητας στην εξίσωση πρόβλεψης διασφαλίζει ότι οι εκτιμήσεις των εποχιακών δεικτών που χρησιμοποιούνται για την πρόβλεψη προέρχονται από το τελευταίο έτος του δείγματος. Η εποχιακή εξίσωση δείχνει έναν σταθμισμένο μέσο μεταξύ του τρέχοντος εποχιακού δείκτη, $(y_t - l_{t-1} - b_{t-1})$ και του εποχιακού δείκτη της ίδιας περσινής εποχής (δηλαδή, πριν από m χρονικές περιόδους). Ισχύει ότι γ ∈[0,1-α], γ*∈[0,1]

Για να εφαρμοστεί ο αλγόριθμός της συνιστώσας μορφής της προσθετικής μεθόδου Holt-Winters απαιτείται ο υπολογισμός των παραμέτρων α, β^*, γ και των αρχικών τιμών l_0, b_0 και $s_k, k = 0, ..., m-1$ όπου m η εποχιακή περίοδος. Ο υπολογισμός των παραμέτρων αυτών γίνεται όπως και στις άλλες μεθόδους εκθετικής εξομάλυνσης μέσω μη γραμμικού προβλήματος βελτιστοποίησης. Συνήθως μια καλή επιλογή αρχικών τιμών των συνιστωσών δίνονται από τις πιο κάτω σχέσεις

$$
\begin{aligned}
s_k &= y_k - average(y_0, ..., y_{m-1}) \quad , \quad k = -m+1, ..., 0 \\
l_0 &= y_0 - s_{-m+1} \\
b_0 &= (y_0 - s_{-m+1}) - y_{-1} - s_{-1}
\end{aligned}
$$

Στον πιο κάτω πίνακα παρουσιάζεται μια εφαρμογή της μεθόδου Holt-Winters με προσθετική εποχικότητα για την πρόβλεψη του εγχώριου τουρισμού στην Αυστραλία.

Table 6.3.1: Με απαλό κόκκινο οι αρχικές τιμές που χρειάζονται για να μπορεί να εφαρμοστεί ο αλγόριθμος υπολογισμού των προβλέψεων ενός βήματος.
Θεωρείται ότι a=0.2620, β=0.1646, γ=0.0001 και οι αρχικές τιμές των συνιστωσών υπολογίστηκαν μέσω του μη γραμμικού προβλήματος βελτιστοποίησης

Τρίμηνο	Χρόνος(t)	y_t	$l_t(a = 0.2620)$	$b_t(\beta^* = 0.1646)$	s_t	$\hat{y}_{t\|t-1}$
1997 Q1	-3	-	-	-	1.5	-
1997 Q2	-2	-	-	-	-0.3	-
1997 Q3	-1	-	-	-	-0.7	-
1997 Q4	0	-	9.8	0.0	-0.5	-
1998 Q1	1	11.8	9.9=a(11.8-1.5)+(1-a)(9.8+0)	$0.0 \approx 0.016 = \beta^*(9.9 - 9.8) + (1 - \beta^*)0$	1.5=γ(11.8-9.8-0)+(1-γ)1.5	11.3=(9.8+0)+1.5
1998 Q2	2	9.3	9.9=a(9.3+0.3)+(1-a)(9.0+0)	$0.0 = \beta^*(9.9 - 9.9) + (1 - \beta^*)0$	-0.3=γ(9.3-9.9-0)+(1-γ)(-0.3)	9.6=(9.9+0.0)-0.3
1998 Q3	3	8.6	$9.7 \approx a(8.6 + 0.7) + (1 - a)(9.9 + 0)$	$0.0 \approx 0.032 = \beta^*(9.7 - 9.9) + (1 - \beta^*)0$	-0.7=γ(8.6-9.9-0)+(1-γ)(-0.7)	9.2=(9.9+0.0)-0.7
1998 Q4	4	9.3	9.8	0.0	-0.5	9.2

Πολλαπλασιαστική μέθοδος των Holt-Winters

Η συνιστώσα μορφή της πολλαπλασιαστικής μεθόδου είναι:

$$
\begin{aligned}
\hat{y}_{t+h|t} &= (l_t + hb_t)s_{t+h-m(k+1)} \\
l_t &= a(\frac{y_t}{s_{t-m}}) + (1-a)(l_{t-1} + b_{t-1}) \\
b_t &= \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1} \\
s_t &= \gamma(\frac{y_t}{l_{t-1} + b_{t-1}}) + (1-\gamma)s_{t-m}
\end{aligned}
$$

Για τις παραμέτρους εξομάλυνσης των συνιστωσών και στα 2 μοντέλα ισχύει ότι όσο μικρότερη η παράμετρος εξομάλυνσης της εποχικότητας γ τόσο λιγότερο αλλάζει η εποχικότητα με την πάροδο του χρόνου. Το ίδιο ισχύει και για την παράμετρο εξομάλυνσης της τάσης β^* . Στην πολλαπλασιαστική μέθοδος των Holt-Winters οι αρχικές τιμές των συνιστωσών μπορούν να προκύψουν από τις σχέσεις

$$
\begin{aligned}
s_k &= \frac{y_k}{average(y_0, ..., y_{m-1})} \quad , \quad k = -m+1, ..., 0 \\
l_0 &= \frac{y_0}{s_{-m+1}} \\
b_0 &= (\frac{y_0}{s_{-m+1}} - \frac{y_{-1}}{s_{-1}})
\end{aligned}
$$

6.3.1 Μέθοδος απόσβεσης των Holt-Winters

Μια μέθοδος που συχνά παρέχει ακριβείς και εύρωστες προβλέψεις για εποχιακά δεδομένα είναι η μέθοδος των Holt-Winters με αποσβένουσα τάση και πολλαπλασιαστική εποχικότητα:

$$\hat{y}_{t+h|t} = (l_t + (\varphi + \varphi^2 + ... + \varphi^h)b_t)s_{t+h-m(k+1)}$$
$$l_t = a(\frac{y_t}{s_{t-m}}) + (1-a)(l_{t-1} + \varphi b_{t-1})$$
$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)\varphi b_{t-1}$$
$$s_t = \gamma(\frac{y_t}{l_{t-1} + \varphi b_{t-1}}) + (1-\gamma)s_{t-m}$$

Θα εφαρμόσουμε τις εποχιακές μεθόδους εκθετικής εξομάλυνσης για ένα σύνολο δεδομένων σχετικά με τον τουρισμό στον Ηνωμένο Βασίλειο. Πιο κάτω βλέπουμε το χρονοδιάγραμμα των δεδομένων.

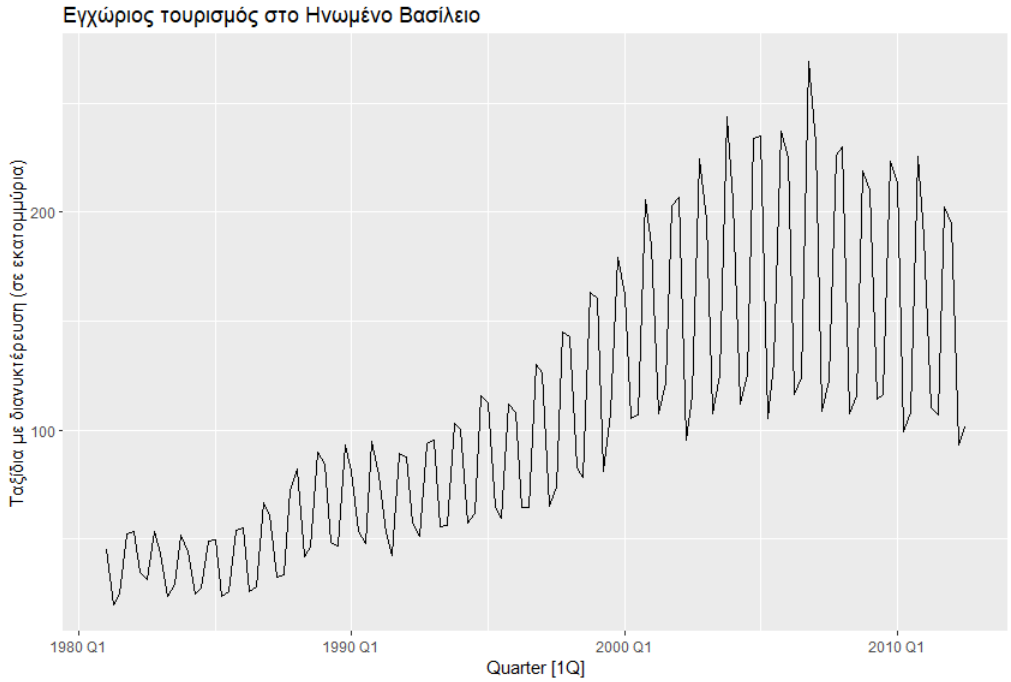


Figure 6.3.1: Τουρισμός στο Ηνωμένο Βασίλειο

Παρατηρούμε ότι οι εποχιακές διακυμάνσεις αλλάζουν με την πάροδο του χρόνου. Για αυτό το λόγο προτιμάμε την πολλαπλασιαστική μέθοδο των Holt-Winters έναντι της προσθετικής. Επίσης θα εφαρμόσουμε και τη μέθοδο απόσβεσης των Holt-Winters για να συγκρίνουμε τις προβλέψεις από τις 2 εποχιακές μεθόδους. Πιο κάτω παρουσιάζουμε γραφικά τις προβλέψεις από τις 2 μεθόδους.

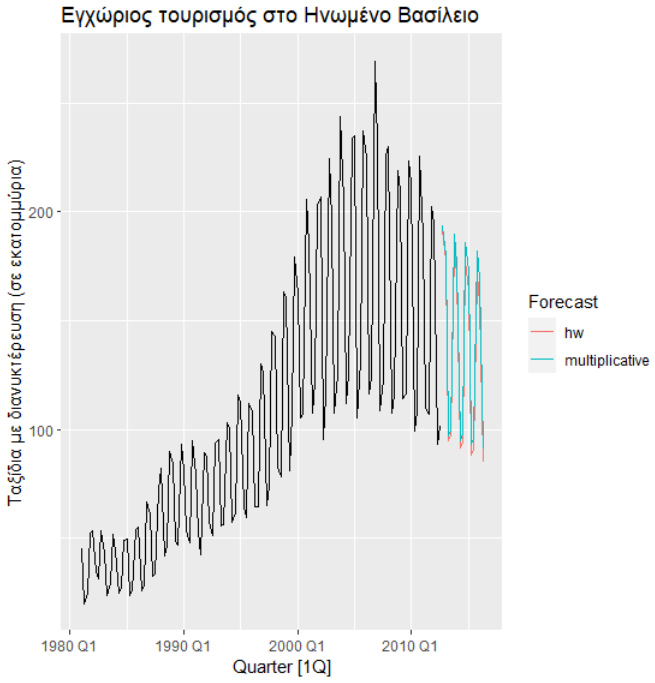


Figure 6.3.2: Προβλέψεις τουρισμού στο Ηνωμένο Βασίλειο για τα επόμενα 15 τρίμηνα με τις μεθόδους: Πολλαπλασιαστική μέθοδος των Holt-Winters & Μέθοδος απόσβεσης των Holt-Winters

Στον πιο κάτω πίνακα παρουσιάζουμε τις εκτιμήσεις των παραμέτρων (επιλέγονται οι καλύτερες με κριτήριο βελτιστοποίησης το SSE) μέσω της γλώσσας προγραμματισμού R.

Table 6.3.2: Παράμετροι των μοντέλων

.model	term	estimate
multiplicative	α	0.4600103
multiplicative	β^*	0.0231754
multiplicative	γ	0.0011777
multiplicative	l[0]	36.2850638
multiplicative	b[0]	0.8439505
multiplicative	s[0]	1.3489132
multiplicative	s[-1]	0.7075021
multiplicative	s[-2]	0.6815007
multiplicative	s[-3]	1.2620840
hw	alpha	0.4125391
hw	β^*	0.0426955
hw	γ	0.0011621
hw	φ	0.9800000
hw	l[0]	36.1192788
hw	b[0]	0.4697212
hw	s[0]	1.3500175
hw	s[-1]	0.7075763
hw	s[-2]	0.6819762
hw	s[-3]	1.2604300

Για την πολλαπλασιαστική μέθοδο των Holt-Winters οι παράμετροι εξομάλυνσης για τις συνιστώσες είναι α =0.46, $\beta^* = 0.023$ και γ =0.0012. Αυτό υπονοεί ότι ούτε η συνιστώσα της τάσης ούτε της εποχικότητας αλλάζει τόσο έντονα. Αυτό επιβεβαιώνεται και αν αναπαραστήσουμε τις προβλέψεις και από την προσθετική μέθοδο οι οποίες δεν θα έχουν μεγάλη διαφορά. Αυτό συμβαίνει καθώς στην εκθετική εξομάλυνση έχουν περισσότερη βαρύτητα οι πρόσφατες παρατηρήσεις με αποτέλεσμα το μοντέλο να μην επηρεάζεται και τόσο από τις μικρότερες εποχιακές διακυμάνσεις που παρατηρήθηκαν πριν το 2000. Αν αυτές οι διαφοροποιήσεις γίνονταν μεταξύ μικρού χρονικού διαστήματος (1-3 χρόνια) τότε η παράμετρο γ θα ήταν αρκετά μεγαλύτερη καθώς τότε πράγματι θα είχαμε έντονες διαφορές στις εποχιακές διακυμάνσεις. Βλέπουμε επίσης και τις αρχικές εκτιμήσεις. Για την εποχιακή συνιστώσα επειδή η εποχικότητα είναι ίση με m=4 απαιτούνται 4 αρχικές εκτιμήσεις.

7 Μοντέλα ARIMA (AutoRegressive Integrated Moving Average)

Τα μοντέλα ARIMA πραγματοποιούν προβλέψεις εξετάζοντας με διαφορετικό τρόπο τα δεδομένα σε σχέση με τα μοντέλα εκθετικής εξομάλυνσης που είδαμε στο προηγούμενο κεφάλαιο. Στοχεύουν στην **περιγραφή της αυτοσυχέτισης των δεδομένων** σε αντίθεση με τα μοντέλα εκθετικής εξομάλυνσης που βασίζονται στην περιγραφή της τάσης και της εποχικότητας των δεδομένων. Επειδή τα μοντέλα ARIMA λειτουργούν καλύτερα με στάσιμες χρονοσειρές, προτού τα ορίσουμε, θα παρουσιάσουμε την τεχνική διαφοροποίησης μιας χρονοσειράς η οποία συμβάλει στην μετατροπή μιας μη στάσιμης χρονοσειράς σε στάσιμη

7.1 Διαφοροποίηση (Differencing time series)

Αρχικά ορίζουμε τον τελεστή υστέρησης B.

Τελεστής υστέρησης: $By_t = y_{t-1}$

Με άλλα λόγια, ο B εφαρμόζεται στην y_t, και έχει ως αποτέλεσμα τη μετατόπιση των δεδομένων πίσω κατά μία περίοδο. Με εφαρμογή του k-φορές (B^k) προκαλεί μετατοπίσεις k-βημάτων.

Διαφοροποίηση (differencing):

Η διαφοροποίηση αποτελεί μια τεχνική που αρκετές φορές μπορεί να μετατρέψει μια μη-στάσιμη χρονοσειρά σε κάποια στάσιμη, από την οποία να μπορούμε να εξάγουμε τις ίδιες πληροφορίες. Η διαφοροποίηση μπορεί να συμβάλει στη σταθεροποίηση του μέσου όρου μιας χρονοσειράς αφαιρώντας τις μεταβολές στο επίπεδο μιας χρονοσειράς και κατά συνέπεια την εξάλειψη (ή τη μείωση) της τάσης και της εποχικότητας. Για την σταθεροποίηση της διακύμανσης, μετασχηματισμοί όπως οι λογαριθμικοί, συχνά αποδεικνύονται αποτελεσματικοί.

Αρκετές φορές, με την παρατήρηση απλά του χρονοδιαγράμματος δεν είναι εύκολο να διακρίνουμε αν μια χρονοσειρά είναι στάσιμη ή όχι. Το γράφημα αυτοσυχέτισης ACF συμβάλει σε αυτή την διαδικασία. Για στάσιμες χρονοσειρές, το ACF θα πέσει στο μηδέν σχετικά γρήγορα, ενώ το ACF μη στάσιμων δεδομένων θα μειώνεται αργά. Επίσης, για μη στάσιμα δεδομένα, η τιμή του r_1 είναι συχνά μεγάλη και θετική.

Η διαφοροποιημένη σειρά:
Αποτελεί τη μεταβολή μεταξύ διαδοχικών παρατηρήσεων στην αρχική σειρά και μπορεί να γραφτεί ως

$$\begin{aligned} y'_t &= y_t - y_{t-1}, \mathrm{t} \in [2,\mathrm{T}] \\ &= (1-B)y_t \end{aligned}$$

Εαν εφαρμόσουμε στην διαφοροποιημένη σειρά ξανά διαφοροποίηση τότε προκύπτουν οι διαφορές $2^{ης}$ τάξης οι οποίες ορίζονται ως

$$\begin{aligned} y_t'' &= y_t' - y_{t-1}' \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2} \\ &= (1-B)^2 y_t \qquad\qquad \text{t} \in [3,\text{T}] \end{aligned}$$

Όταν η διαφοροποιημένη σειρά (δηλαδή οι προσαυξήσεις σε διαδοχικά χρονικά σημεία) είναι λευκός θόρυβος, το μοντέλο για την αρχική σειρά μπορεί να γραφτεί ως ένα μοντέλο "τυχαίου περιπάτου"

$$y_t = y_{t-1} + \varepsilon_t, \text{ t} \in [2,\text{T}]$$

Όπου ε_t εκφράζει λευκό θόρυβο. Έτσι, το μοντέλο τυχαίου περιπάτου αποτελεί τη βάση για τις προβλέψεις της μεθόδου naïve. Τα μοντέλα τυχαίου περιπάτου χρησιμοποιούνται ευρέως για μη στάσιμα δεδομένα, ιδιαίτερα σε χρηματοοικονομικά και οικονομικά δεδομένα. Οι τυχαίοι περίπατοι χαρακτηρίζονται από

- Μεγάλες περιόδους εμφανών ανοδικών ή καθοδικών τάσεων
- Ξαφνικές και απρόβλεπτες μεταβολές στην κατεύθυνση.

Ένα στενά συνδεδεμένο μοντέλο που είναι χρήσιμο όταν ο μέσος όρος των μεταβολών μεταξύ διαδοχικών παρατηρήσεων είναι ίσο με $\text{c} \neq 0$ δίνεται πιό κάτω.

$$y_t - y_{t-1} = c + \varepsilon_t \implies y_t = c + y_{t-1} + \varepsilon_t$$

Αυτό είναι το μοντέλο πίσω από τη μέθοδο περιπλάνησης (drift method)

Εποχιακή διαφοροποίηση (διαφορές m-υστέρησης)
Μια εποχιακή διαφορά είναι η διαφορά μεταξύ μιας παρατήρησης και της προηγούμενης παρατήρησης από την ίδια εποχή και ορίζεται ως

$$\begin{aligned} y_t' &= y_t - y_{t-m} \\ &= (1-B^m) y_t \end{aligned}$$

όπου m=είναι ο αριθμός των εποχών. Εαν τα εποχιακά διαφοροποιημένα δεδομένα φαίνονται να συμπεριφέρονται ως λευκός θόρυβος, τότε ένα κατάλληλο μοντέλο για τα αρχικά δεδομένα είναι το

$$y_t = y_{t-m} + \varepsilon_t$$

Δηλαδή, αυτό το μοντέλο δίνει naïve εποχιακές προβλέψεις.

Αρκετά συχνά, για τη λήψη στάσιμων χρονοσειρών, είναι απαραίτητο να ληφθεί τόσο μια εποχιακή διαφορά όσο και μια πρώτη διαφορά (διαφορές με υστέρηση 1). Όταν εφαρμόζονται εποχιακές και πρώτες διαφορές, το αποτέλεσμα θα είναι το ίδιο ανεξάρτητα από τη σειρά εφαρμογής τους. Ωστόσο, εαν τα δεδομένα έχουν ισχυρό εποχιακό μοτίβο, καλό είναι να πραγματοποιούμε πρώτα τις εποχιακές διαφορές, επειδή η σειρά που προκύπτει μερικές φορές θα είναι στάσιμη και δεν θα υπάρχει ανάγκη για περαιτέρω διαφοροποίηση.

Ένα σημαντικό ζήτημα με τις διαφοροποιήσεις είναι ότι πρέπει να είναι ερμηνεύσιμες για να είναι πράγματι βοηθητικές. Οι διαφορές $1^{ης}$ τάξης είναι η μεταβολή μεταξύ μιας παρατήρησης και της επόμενης. Οι εποχιακές διαφορές είναι η μεταβολή από έτος σε έτος. Όλες οι άλλες διαφορές δέν ερμηνεύονται εύκολα.

Για να ελέγξουμε κατα πόσο μια χρονοσειρά χρειάζεται διαφοροποίηση έχει προταθεί ο στατιστικός έλεγχος μοναδιαίας ρίζας (unit root test). Υπάρχουν αρκετοί διαθέσιμοι έλεγχοι μοναδιαίας ρίζας, οι οποίοι βασίζονται σε διαφορετικές υποθέσεις και μπορεί να οδηγήσουν σε αντικρουόμενες απαντήσεις. Ένας πολύ γνωστός είναι ο έλεγχος Dickey-Fuller ο οποίος έχει ως μηδενική υπόθεση ότι ο συντελεστής φ_1 ενός μοντέλου AR(1) που προσαρμόζεται στα δεδομένα (το ορίζουμε στην συνέχεια) ανήκει στον μοναδιαίο κύκλο (είναι ίσος με 1) το οποίο συνεπάγεται μή στάσιμη χρονοσειρά. Μία επέκταση του στατιστικού ελέγχου αυτού είναι ο έλεγχος augmented Dickey–Fuller test ο οποίος χρησιμοποιείται για έλεγχο στασιμότητας σε μοντέλα AR(p). Στην εργασία των Kwiatkowski, Phillips, Schmidt και Shin ([12]) προτάθηκε ένα άλλος έλεγχος, με μηδενική υπόθεση ότι τα δεδομένα είναι στάσιμα. Το έκαναν αυτό επειδή για χρόνια παρατηρήθηκε ότι οι κλασσικοί έλεγχοι μοναδιαίας ρίζα (όπως τον Dickey-Fuller) αδυνατούσαν να απορρίψουν την μηδενική υπόθεση για αρκετές χρονοσειρές ενώ τα δεδομένα δεν έμοιαζαν να είναι μή στάσιμα. Αυτό συνέβαινε κυρίως επειδή περίπλοκες χρονοσειρές δέν είχαν επαρκή στοιχεία για να υπάρξουν **ισχυρές** ενδείξεις που να συνηγορούν στην απόρριψη της μηδενικής υπόθεσης. Σε αυτή την εργασία θα στηριχτούμε κυρίως στον έλεγχο Kwiatkowski-Phillips-Schmidt-Shin (KPSS) στον οποίο η μηδενική υπόθεση είναι ότι τα δεδομένα είναι στάσιμα .

Παρατήρηση:
Ο τελεστής υστέρησης είναι ιδιαίτερα χρήσιμος και καλό είναι να χρησιμοποιείται για την έκφραση των διαφορών καθώς μπορεί να αντιμετωπιστεί χρησιμοποιώντας συνήθεις αλγεβρικούς κανόνες. Συγκεκριμένα, οι όροι που περιλαμβάνουν το B μπορούν να πολλαπλασιαστούν μαζί. Για παράδειγμα, μια εποχιακή διαφορά ακολουθούμενη από μια διαφορά πρώτης τάξης μπορεί να γραφτεί ως

$$\begin{aligned}(1-B)(1-B^m)y_t &= (1-B-B^m+B^{m+1})y_t \\ &= y_t - y_{t-1} - y_{t-m} + y_{t-m-1}\end{aligned}$$

7.2 Αυτοπαλίνδρομα μοντέλα (Autoregressive models)

Σε ένα αυτοπαλίνδρομο μοντέλο, προβλέπουμε τη μεταβλητή ενδιαφέροντος χρησιμοποιώντας έναν γραμμικό συνδυασμό των προηγούμενων τιμών της μεταβλητής. Για αυτο το λόγο ονομάζεται αυτοπαλίνδρομο καθώς είναι μια παλινδρόμηση της μεταβλητής ενάντια στον εαυτό της.

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

όπου ε_t είναι λευκός θόρυβος. Αυτό το μοντέλο ονομάζεται AR(p) μοντέλο, δηλαδή ένα αυτοπαλίνδρομο μοντέλο τάξης p. Τόσο η τάξη του μοντέλου όσο και η επιλογή των παραμέτρων καθορίζουν το μοντέλο.

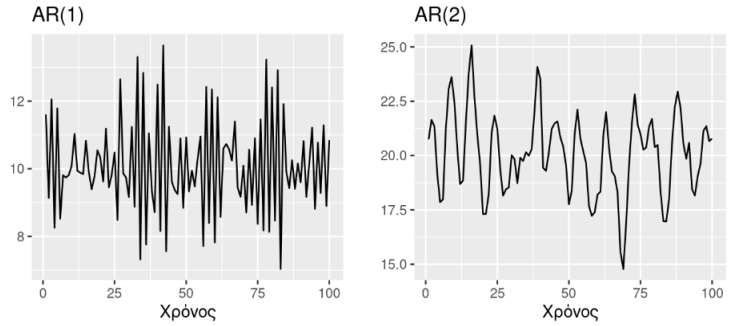


Figure 7.2.1: Δύο παραδείγματα δεδομένων από αυτοπαλίνδρομα μοντέλα με διαφορετικές παραμέτρους

Ορίζοντας το πολυώνυμο AR(p):

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 + ... - \varphi_p B^p$$

Τότε η πιο πάνω εξίσωση μπορεί να εκφραστεί μέσω της σχέσης:

$$\varphi(B) y_t = c + \varepsilon_t$$

Συνήθως περιορίζουμε τα αυτοπαλίνδρομα μοντέλα στα στάσιμα δεδομένα, οπότε απαιτούνται ορισμένοι περιορισμοί στις τιμές των παραμέτρων. Αποδεικνύεται ότι η διαδικασία AR(p) είναι στάσιμη όταν οι p στο πλήθος (μιγαδικές) ρίζες του πολυωνύμου φ(B) βρίσκονται όλες έξω από τον μοναδιαίο κύκλο .Ενδεικτικά παρουσιάζουμε τους περιορισμούς για τάξη 1 και 2.

- Για ένα μοντέλο $AR_{(1)}$: $-1 < \varphi_1 < 1$
- Για ένα μοντέλο $AR_{(2)}$: $-2 < \varphi_2 < 1\ , \quad \varphi_1 + \varphi_2 < 1\ , \quad \varphi_2 - \varphi_1 < 1$

7.3 Μοντέλα κινητών μέσων (Moving average model)

Ένα μοντέλο κινητών μέσων (moving average model) **χρησιμοποιεί τα προηγούμενα σφάλματα πρόβλεψης** για να κάνει παλινδρόμηση. Κάθε τιμή της y_t μπορεί να θεωρηθεί ως ο σταθμισμένος κινητός μέσος των τελευταίων σφαλμάτων πρόβλεψης. Δηλαδή

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}$$

Όπου και πάλι ε_t είναι λευκός θόρυβος. Ορίζοντας το πολυώνυμο:

$$\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q$$

Τότε η πιο πάνω εξίσωση μπορεί να εκφραστεί μέσω της σχέσης:

$$y_t = c + \theta(B) \varepsilon_t$$

Αυτό το μοντέλο ονομάζεται Μοντέλο MA(q) .

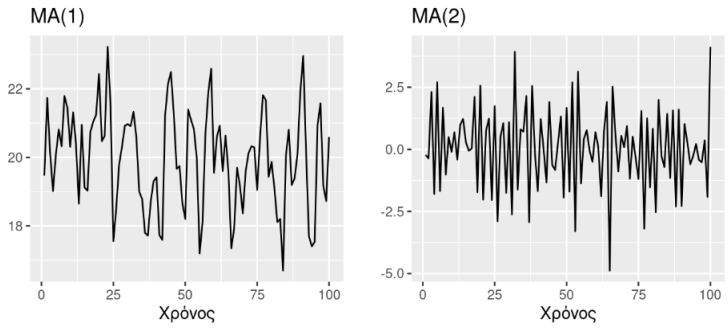


Figure 7.3.1: Δύο παραδείγματα δεδομένων από μοντέλα κινητών μέσων με διαφορετικές παραμέτρους

Τα μοντέλα κινητών μέσων είναι διαφορετική έννοια από τον **κινητό μέσο εξομάλυνσης** που συζητήθηκε στο κεφάλαιο (3.2.1). Ένα μοντέλο κινητών μέσων χρησιμοποιείται για την πρόβλεψη μελλοντικών τιμών, ενώ ο κινητός μέσος εξομάλυνσης χρησιμοποιείται για την **εκτίμηση του κύκλου τάσεων των προηγούμενων τιμών.**

Αξίζει να σημειωθεί ότι οποιοδήποτε στάσιμο AR(p) μπορεί να γραφεί ως $MA(\infty)$. Στην συνέχεια παρουσιάζεται πώς αυτό προκύπτει για ένα μοντέλο AR(1).

$$\begin{aligned} y_t &= \varphi_1 y_{t-1} + \varepsilon_t \\ &= \varphi_1(\varphi_1 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \varphi_1^2 y_{t-2} + \varphi_1 \varepsilon_{t-1} + \varepsilon_t \\ &= \dots \\ &= \varphi_1^m y_{t-m} + \varphi_1^{m-1} \varepsilon_{t-m+1} + \dots + \varepsilon_t \end{aligned}$$

Με την προϋπόθεση στασιμότητας του AR(p), για $m \to \infty$, προκύπτει ότι

$$y_t = \varepsilon_t + \varphi_1 \varepsilon_{t-1} + \varphi_1^2 \varepsilon_{t-2} + \varphi_1^3 \varepsilon_{t-3} + \dots$$

Για να ισχύει το αντίστροφο αποτέλεσμα, να μπορούμε να γράψουμε μία διαδικασία MA(q) ως μια AR(∞), πρέπει να επιβάλουμε ορισμένους περιορισμούς στις παραμέτρους του MA, και τότε το MA ονομάζεται αντιστρέψιμο. Αποδεικνύεται ότι για να είναι μια διαδικασία MA(q) αντιστρέψιμη θα πρέπει οι q στο πλήθος (μιγαδικές) ρίζες του πολυωνύμου ϑ(B) να είναι όλες έξω από τον μοναδιαίο κύκλο. Τα μοντέλα κινητών μέσων είναι εξ'ορισμού στάσιμα, αφού η διαδικασία λευκού θορύβου είναι στάσιμη.

7.4 Μη εποχιακά μοντέλα ARIMA

Ένα μη εποχιακό μοντέλο ARIMA προκύπτει συνδυάζοντας διαφοροποίηση με αυτοπαλινδρόμηση και ένα μοντέλο κινητού μέσου. Επομένως το μοντέλο εκφράζεται από την σχέση

$$y_t\prime = c + \varphi_1 y'_{t-1} + ... + \varphi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{7.4.1}$$

Όπου y'_t είναι η διαφοροποιημένη σειρά (μπορεί να έχει διαφοροποιηθεί περισσότερες από μία φορές). Ονομάζουμε αυτό το μοντέλο ως μοντέλο ARIMA(p,d,q), όπου

- p = τάξη του αυτοπαλινδρόμενου μέρους
- d = τάξη πρώτων διαφορών
- q = τάξη του μέρους του κινητού μέσου

Πολλά από τα μοντέλα που αναφέραμε εώς τώρα είναι ειδικές περιπτώσεις μοντέλων ARIMA

Λευκός θόρυβος	ARIMA(0,0,0) χωρίς σταθερά
Τυχαίος περίπατος	ARIMA(0,1,0) χωρίς σταθερά
Τυχαίος περίπατος με περιπλάνηση	ARIMA(0,1,0) με σταθερά
Αυτοπλαινδρόμηση	ARIMA(p,0,0)
Κινητός μέσος	ARIMA(0,0,q)

Figure 7.4.1: Ειδικές περιπτώσεις μοντέλων ARIMA

Η σχέση (7.4.1) που εκφράζει το μοντέλο ARIMA(p,d,q) μπορεί να γραφτεί (με αναδιάταξη όρων) χρησιμοποιώντας τον τελεστή υστέρησης ως

$$\underset{\uparrow\\AR(p)}{(1 - \varphi_1 B - ... - \varphi_p B^p)} \quad \underset{\uparrow\\d-differences}{(1-B)^d} \quad y_t = c + \underset{\uparrow\\MA(q)}{(1 + \theta 1 B + ... + \theta_q B^q)} \varepsilon_t$$

Η σταθερά c έχει σημαντική επίδραση στις μακροπρόθεσμες προβλέψεις που λαμβάνονται από αυτά τα μοντέλα.

- c=0 και d=0 , οι μακροπρόθεσμες προβλέψεις θα μηδενιστούν.
- c=0 και d=1 , οι μακροπρόθεσμες προβλέψεις θα τείνουν σε μια μη μηδενική σταθερά
- c=0 και d=2 , οι μακροπρόθεσμες προβλέψεις θα ακολουθούν μια ευθεία γραμμή.
- c≠0 και d=0 , οι μακροπρόθεσμες προβλέψεις θα τείνουν στο μέσο όρο των δεδομένων.
- c≠0 και d=1 , οι μακροπρόθεσμες προβλέψεις θα ακολουθούν μια ευθεία γραμμή.
- c≠0 και d=2 , οι μακροπρόθεσμες προβλέψεις θα ακολουθούν μια τετραγωνική τάση.

Η τιμή του d επηρεάζει, επίσης, και τα διαστήματα πρόβλεψης. Όσο μεγαλύτερη είναι η τιμή του d, τόσο πιο γρήγορα αυξάνεται το μέγεθος των διαστημάτων πρόβλεψης. Για d=0, η τυπική απόκλιση της

μακροπρόθεσμης πρόβλεψης θα τείνει στην τυπική απόκλιση των ιστορικών δεδομένων, επομένως όλα τα διαστήματα πρόβλεψης θα είναι ουσιαστικά τα ίδια.

Η τιμή του p είναι σημαντική αν τα δεδομένα εμφανίζουν κύκλους. Για να ληφθούν προβλέψεις με κυκλικότητα, είναι απαραίτητο να οριστεί p≥2, μαζί με ορισμένες πρόσθετες συνθήκες στις παραμέτρους. Για παράδειγμα σε ένα μοντέλο $AR(2)$, η κυκλική συμπεριφορά εμφανίζεται εαν $\varphi_1^2 + 4\varphi_2 < 0$

7.4.1 Καθορισμός παραμέτρων p,d,q στα μοντέλα ARIMA

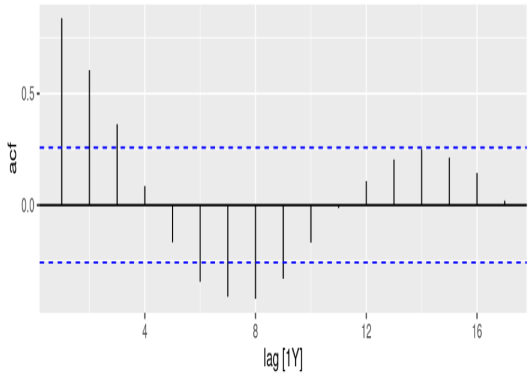
Παρατηρώντας απλά το χρονοδιάγραμμα δεν είναι εύκολη υπόθεση ο καθορισμός των παραμέτρων p,q,d. Το γράφημα της αυτοσυχέτισης ACF και το πολύ σχετικό διάγραμμα μερικής αυτοσυχέτισης PACF, αποτελούν δύο σημαντικά εργαλεία για να προσδιοριστούν οι κατάλληλες τιμές για τα p και q. Το PACF χρησιμοποιείται για να ξεπεράσει ένα βασικό πρόβλημα των γραφημάτων ACF. Ένα διάγραμμα ACF μας δείχνει τις αυτοσυχετίσεις μεταξύ y_t και y_{t-k} για τις διάφορες τιμές του k. Το ζήτημα που προκύπτει είναι ότι αν τα y_t και y_{t-1} συσχετίζονται, τότε θα πρέπει οι y_{t-1} και y_{t-2} να συσχετίζονται απλώς και μόνο επειδή συνδέονται και τα δύο με τη y_{t-1}, και όχι λόγω κάποιας νέας πληροφορίας που περιέχεται στο y_{t-2} που θα μπορούσε να χρησιμοποιηθεί στην πρόβλεψη της y_t.

Οι μερικές αυτοσυχετίσεις (partial autocorrelations) συμβάλουν στο να ξεπεραστεί αυτό το πρόβλημα. Αυτές μετρούν τη σχέση μεταξύ y_t και y_{t-k} αφού αφαιρεθεί η επίδραση των υστερήσεων 1,2,3,...,k-1. Κάθε μερική αυτοσυχέτιση μπορεί να εκτιμηθεί ως ο τελευταίος συντελεστής σε ένα αυτοπαλινδρόμενο μοντέλο. Συγκεκριμένα, ο k-οστός συντελεστής μερικής αυτοσυχέτισης (a_k), ισούται με την εκτίμηση του ϕ_k σε ένα μοντέλο AR(k). Στην πράξη, υπάρχουν πιο αποτελεσματικοί αλγόριθμοι για τον υπολογισμό του α_k.

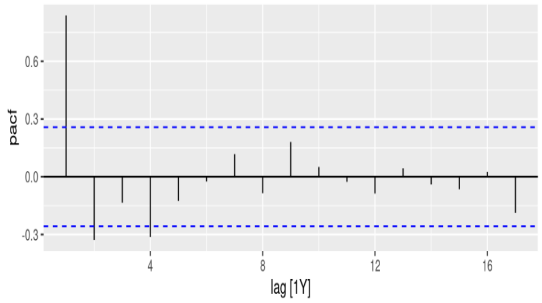
Τα δεδομένα ενδέχεται να ακολουθούν ένα μοντέλο ARIMA(p,d,0), εαν τα διαγράμματα των ACF και PACF **των διαφοροποιημένων δεδομένων** εμφανίζουν τα ακόλουθα πρότυπα:

- το ACF φθίνει εκθετικά ή ημιτονοειδώς.
- υπάρχει σημαντική κορυφή στην υστέρηση p στο PACF, αλλά καμία πέραν της υστέρησης p.

Πιο κάτω παρουσιάζουμε ένα ενδεικτικό παράδειγμα αυτής της συμπεριφοράς



(a) Υπάρχει ένα εξασθενημένο ημιτονοειδές πρότυπο στο ACF



(b) Τελευταία σημαντική κορυφή στην υστέρηση 4

Figure 7.4.2: ACF και PACF που θα περίμενε κανείς από ένα μοντέλο ARIMA(4,0,0)

Τα δεδομένα ενδέχεται να ακολουθούν ένα μοντέλο ARIMA(0,d,q), εαν τα διαγράμματα ACF και PACF των διαφοροποιημένων δεδομένων εμφανίζουν τα ακόλουθα πρότυπα:

- το PACF φθίνει εκθετικά ή ημιτονοειδώς.
- υπάρχει σημαντική κορυφή στην υστέρηση q στο ACF, αλλά καμία πέραν της υστέρησης q.

Το **Κριτήριο Πληροφορίας του Akaike** (Akaike's Information Criterion - AIC) είναι χρήσιμο για τον προσδιορισμό της τάξης ενός μοντέλου ARIMA. Το μέτρο AIC ορίζεται ως

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

όπου L είναι η πιθανοφάνεια των δεδομένων, $k = 1$ εαν $c \neq 0$ και $k = 0$ εαν $c = 0$. Ο τελευταίος όρος στην παρένθεση είναι το πλήθος των παραμέτρων στο μοντέλο (συμπεριλαμβανομένου του σ^2, της διακύμανσης των υπολοίπων). Προτιμάτε το μοντέλο με τη μικρότερη τιμή AIC.

Για τα μοντέλα ARIMA, **το διορθωμένο AIC**, το οποίο είναι χρήσιμο όταν το μέγεθος της χρονοσειράς δεν είναι πολύ πιο μεγάλο απο το πλήθος παραμέτρων του μοντέλο μπορεί να γραφτεί ως

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

Το AIC τείνει να ευνοεί περισσότερο τα μοντέλα με πολλές παραμέτρους όταν το μέγεθος της χρονοσειρά είναι μικρό καθώς οδηγείται στην υπερ-προσαρμογή που έχει σαν αποτέλεσμα η πιθανοφάνεια να λαμβάνει πολύ μεγάλη τιμή. Ουσιαστικά το μοντέλο κατορθώνει να καταγράψει πλήρως την συμπεριφορά των δεδομένων της χρονοσειράς αλλά αδυνατεί να γενικεύσει την καλή του απόδοση σε νέα, άγνωστα δεδομένα. Για αυτό τον λόγο είναι προτιμότερο να χρησιμοποιείται το AIC_c. Όταν το μέγεθος της χρονοσειράς είναι μεγάλο, οι δύο δείκτες τείνουν να ταυτιστούν.

Το **Bayesian κριτήριο πληροφορίας** (Bayesian Information Criterion) μπορεί να γραφτεί ως

$$BIC = AIC + [log(T) - 2](p + q + k + 1)$$

Η βασική του διαφορά απο το AIC είναι ότι η εισαγωγή επιπρόσθετων παραμέτρων αποθαρρύνεται σε μεγαλύτερο βαθμό από το AIC.

Αυτά τα κριτήρια πληροφορίας είναι βοηθητικά μόνο για την επιλογή των τιμών p και q και όχι για την παράμετρο d. Αυτό συμβαίνει επειδή η διαφοροποίηση αλλάζει τα δεδομένα στα οποία υπολογίζεται η πιθανοφάνεια, καθιστώντας τις τιμές του AIC μεταξύ των μοντέλων με διαφορετικές τάξης διαφοροποίησης μη συγκρίσιμες. Επομένως αρχικά προσδιορίζουμε την τιμή του d μέσω κάποιου άλλου κριτηρίου (π.χ επαναληπτική εφαρμογή του ελέγχου KPSS) και στην συνέχεια μπορούμε να χρησιμοποιήσουμε το AIC_c για να επιλέξουμε τις p και q. Η εκτίμηση των παραμέτρων $c, \phi_1, ..., \phi_p, \theta_1, ..., \theta_q$ γίνεται με βάση κάποιο κριτήριο βελτιστοποίησης. Συνήθως χρησιμοποιείται η εκτίμηση μέγιστης πιθανοφάνειας (maximum likelihood estimation). Αυτή η τεχνική βρίσκει τις τιμές των παραμέτρων που μεγιστοποιούν την πιθανότητα να πάρουμε τα δεδομένα που έχουμε παρατηρήσει.

7.5 Εποχιακά μοντέλα ARIMA

Τα μοντέλα ARIMA μπορούν, επίσης, να μοντελοποιήσουν ένα ευρύ φάσμα εποχιακών δεδομένων. Ένα εποχιακό μοντέλο ARIMA διαμορφώνεται περιλαμβάνοντας πρόσθετους εποχιακούς όρους στα μοντέλα ARIMA που έχουμε δει μέχρι στιγμής. Συμβολίζεται ως

$$ARIMA \quad \underset{\substack{\uparrow \\ \text{Μη εποχιακό} \\ \text{μέρος του μοντέλου}}}{(p,d,q)} \quad \underset{\substack{\uparrow \\ \text{Εποχιακό μέρος} \\ \text{του μοντέλου}}}{(P,D,Q)_m}$$

όπου m είναι ίσο με την εποχιακή περίοδο. Το εποχιακό μέρος προκύπτει σχεδόν όπως το μη εποχιακό μέρος αλλά περιλαμβάνει και υστερήσεις της εποχιακής περιόδου.

Για παράδειγμα, ένα μοντέλο $ARIMA(1,1,1)(1,1,1)_4$ (χωρίς σταθερά) προορίζεται για τριμηνιαία δεδομένα (m=4) και μπορεί να γραφτεί ως

$$(1 - \varphi_1 B)(1 - \Phi_1 B^4)(1 - B)(1 - B^4)y_t = (1 + \theta_1 B)(1 + \Theta_1 B^4)\varepsilon_t$$

7.5.1 Αξιοποίηση διαγραμμάτων Διαγράμματα ACF/PACF για τον καθορισμό των παραμέτρων

Το εποχιακό μέρος ενός μοντέλου AR ή MA θα **εμφανίζεται στις εποχιακές υστερήσεις του PACF και του ACF**. Για παράδειγμα, ένα μοντέλο $ARIMA(0,0,0)(0,0,1)_{12}$ θα παρουσιάζει:

- Μια κορυφή στην υστέρηση 12 στο ACF αλλά όχι άλλες σημαντικές κορυφές.
- Εκθετική μείωση στις εποχιακές υστερήσεις του PACF (δηλ. στις υστερήσεις 12, 24, 36, ...).

Ομοίως, ένα μοντέλο $ARIMA(0,0,0)(1,0,0)_{12}$ θα παρουσιάσει:

- Εκθετική μείωση στις εποχιακές υστερήσεις του ACF
- Μία μοναδική σημαντική κορυφή στην υστέρηση 12 στο PACF

Πιο κάτω παρουσιάζουμε ένα ενδεικτικό παράδειγμα από τα 2 γραφήματα και στην συνέχεια σχολιάζουμε πώς με βάση αυτά εντοπίζουμε καλές τιμές για τις παραμέτρους p,q του μοντέλου.

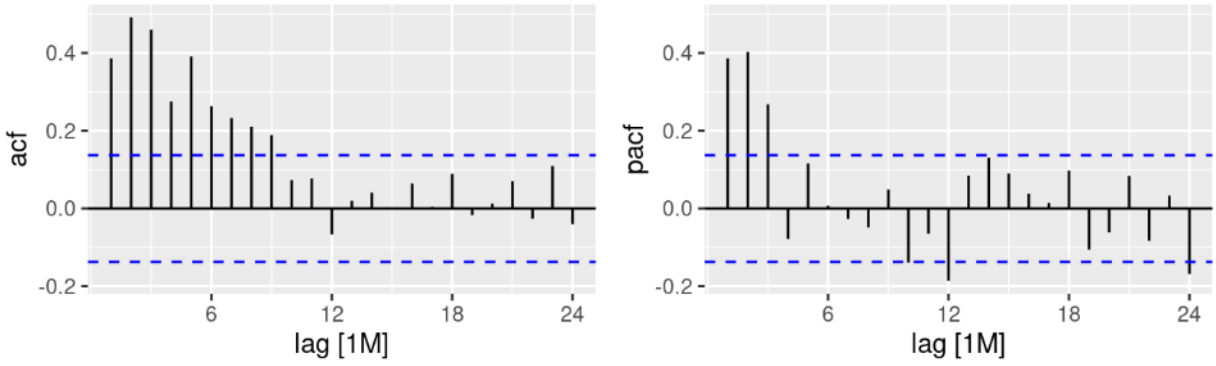


Figure 7.5.1: ACF και PACF για τον καθορισμό των παραμέτρων p,q ενός μοντέλου ARIMA

Παρατηρούμε ότι στο γράφημα PACF υπάρχουν κορυφές στις υστερήσεις 12 και 24 ενώ στο γράφημα ACF δέν υπάρχει καμία κορυφή σε εποχιακές υστερήσεις (υστερήσεις πολλαπλάσιες της εποχικότητας). Αυτό το πρότυπο υποδηλώνει ένα εποχιακό όρο AR(2). Επίσης στο γράφημα PACF παρατηρούμε 3 σημαντικές κορυφές σε μή εποχικές υστερήσεις, γεγονός που που υποδηλώνει έναν πιθανό μή εποχιακό όρο AR(3). Επομένως καταλήγουμε ότι ένα πιθανό μοντέλο για τα δεδομένα μας είναι ένα $ARIMA(3,0,0)(2,1,0)_{12}$.

7.6 Πρόβλεψη με μοντέλα ARIMA

Στα προηγούμενα υπο-κεφάλαια εξηγήθηκε η λειτουργία των μοντέλων ARIMA και πώς επιλέγονται οι παράμετροι τους έτσι ώστε να μπορούν να μοντελοποιούν τα διάφορα μοτίβα που κρύβει μια χρονοσειρά. Δεν εξηγήθηκε όμως πώς ακριβώς από τις εξισώσεις ARIMA(p,d,q)(P,D,Q) προκύπτουν οι εξισώσεις από τις οποίες παράγονται οι μελλοντικές προβλέψεις. Η διαδικασία με την οποία προκύπτει η εξίσωση πρόβλεψης δίνεται μέσω ενός παραδείγματος για να γίνει εύκολα κατανοητή. Παρουσιάζουμε την διαδικασία αυτή χρησιμοποιώντας ένα μοντέλο ARIMA(3,1,1) που μπορεί να γραφτεί ως εξής:

$$(1-\hat{\varphi}_1 B-\hat{\varphi}_2 B^2-\hat{\varphi}_3 B^3)(1-B)y_t=(1+\hat{\theta}_1 B)\varepsilon_t$$

Βήμα 1:
Επεξεργασία της εξίσωσης ARIMA έτσι ώστε το y_t να βρίσκεται στην αριστερή πλευρά και όλοι οι άλλοι όροι να είναι στα δεξιά.

$$[1-(1+\hat{\varphi}_1)B+(\hat{\varphi}_1-\hat{\varphi}_2)B^2+(\hat{\varphi}_2-\hat{\varphi}_3)B^3+\hat{\varphi}_3 B^4]y_t=(1+\hat{\theta}_1 B)\varepsilon_t$$

Στην συνέχεια εφαρμόζεται ο τελεστής υστέρησης B για να προκύψει

$$y_t-(1+\hat{\varphi}_1)y_{t-1}+(\hat{\varphi}_1-\hat{\varphi}_2)y_{t-2}+(\hat{\varphi}_2-\hat{\varphi}_3)y_{t-3}+\hat{\varphi}_3 y_{t-4}=\varepsilon_t+\hat{\theta}_1\varepsilon_{t-1}$$

Τέλος, μετακινούμε όλους τους όρους εκτός από τον y_t στη δεξιά πλευρά:

$$y_t=(1+\hat{\varphi}_1)y_{t-1}-(\hat{\varphi}_1-\hat{\varphi}_2)y_{t-2}-(\hat{\varphi}_2-\hat{\varphi}_3)y_{t-3}-\hat{\varphi}_3 y_{t-4}+\varepsilon_t+\hat{\theta}_1\varepsilon_{t-1}$$

Βήμα 2:
Στην τελευταία σχέση αντικαθιστούμε το t με T+1. Επίσης υποθέτοντας ότι έχουμε γνώση των τιμών έως και τη χρονική στιγμή T, όλες οι τιμές στη δεξιά πλευρά είναι γνωστές εκτός από την ε_{T+1}, την οποία αντικαθιστούμε με μηδέν. Την ε_T την αντικαθιστούμε με το τελευταίο παρατηρούμενο υπόλοιπο e_T

Με αυτό το τρόπο λαμβάνεται η πρόβλεψη ενός βήματος. Για να ληφθεί η πρόβλεψη δύο βημάτων αντικαθιστούμε το T+1 με T+2 και για την άγνωστη τιμή y_{T+1} αντικαθιστούμε την πρόβλεψη που λάβαμε ακριβώς πριν, δηλαδή τη τιμή $\hat{y}_{T+1}$. Τέλος ο άγνωστος όρος e_{T+1} θεωρείται ίσος με μηδέν. Με αυτή τη διαδικασία η πρόβλεψη δύο βημάτων προκύπτει από τη σχέση

$$\hat{y}_{T+2|T+1}=(1+\hat{\varphi}_1)\hat{y}_{T+1}-(\hat{\varphi}_1-\hat{\varphi}_2)y_T-(\hat{\varphi}_2-\hat{\varphi}_3)y_{T-1}-\hat{\varphi}_3 y_{T-2}$$

Για προβλέψεις μεγαλύτερου χρονικού ορίζοντα συνεχίζεται η διαδικασία που περιγράφηκε για την πρόβλεψη χρονικού ορίζοντα δύο βημάτων.

Οι εκτιμήσεις των συντελεστών των μοντέλων ARIMA(p,d,q), γίνονται συνήθως με την μέθοδο μέγιστης πιθανοφάνειας. Υπάρχουν 2 τρόποι για να γίνει αυτο. Ο πρώτος τρόπος, είναι μέσω της δεσμευμένης συνάρτησης πιθανοφάνειας (Conditional likelihood function), ο οποίος δεν καταλήγει σε κλειστό τύπο υπολογισμού των παραμέτρων αλλά σε ένα μή γραμμικό πρόβλημα ελαχιτοσποίησης, το οποίο συνήθως λύνεται με αριθμητικές μεθόδους (π.χ Newton Raphson). Ο δεύτερος τρόπος χρησιμοποιεί την μή δεσμευμένη συνάρτηση πιθανοφάνειας (Unconditional likelihood function) και στηρίζεται στην μέθοδο

της προς τα πίσω πρόβλεψης (back forecasting). Ο τρόπος με τον οποίο λειτουργούν αυτές οι 2 μέθοδοι υπολογισμού των συντελεστών του μοντέλου ARIMA(p,d,q) περιγράφονται αναλυτικά στο [6]

7.7 Δυναμική παλινδρόμηση-ARIMAX (ARIMA με παράγοντες πρόβλεψης)

Αρκετά συχνά για την επεξήγηση της συμπεριφοράς μιας χρονοσειράς αξιοποιούνται παράγοντες πρόβλεψης. Δηλαδή αξιοποιούνται άλλες μεταβλητές όπου η συμπεριφορά τους σχετίζεται άμεσα με τη συμπεριφορά της χρονοσειράς. Αρκετά συχνά αυτή η προσέγγιση έχει καλά αποτελέσματα αλλά αυτό γίνεται μόνο όταν οι μελλοντικές τιμές των παραγόντων πρόβλεψης μπορούν να εξαχθούν με αρκετά καλή ακρίβεια. Για παράδειγμα, η επίδραση της θερμοκρασίας και των διακοπών για πρόβλεψης ζήτηση ηλεκτρικής ενέργειας, η δραστηριότητα των ανταγωνιστών και η ευρύτερη οικονομία για την πρόβλεψη των πωλήσεων μιας επιχείρησης. Γενικότερα οι παράγοντες πρόβλεψης μπορούν να εξηγήσουν μερικές από τις ιστορικές διακυμάνσεις και αυτό μπορεί να οδηγήσει σε πιο ακριβείς προβλέψεις. Το αρνητικό αυτής της προσέγγισης στην πρόβλεψη χρονοσειρών είναι ότι αρκετές φορές οι μελλοντικές τιμές των παραγόντων πρόβλεψης πρέπει και αυτές να εκτιμηθούν με βάση κάποιο μοντέλο πρόβλεψης. Αυτή η διαδικασία εισάγει επιπλέον αβεβαιότητα στις προβλέψεις της χρονοσειράς, και αυτές όπως και τα αντίστοιχα διαστήματα εμπιστοσύνης τους, ευσταθούν μόνο υπο την προϋπόθεση ότι προέκυψαν οι προβλέψεις των παραγόντων πρόβλεψης που έχουν ήδη γίνει. Από την άλλη πλευρά για την επεξήγηση της συμπεριφοράς μιας χρονοσειράς αρκετές φορές η εκμετάλλευση απλά και μόνο των ιστορικών τιμών της είναι ικανοποιητική.

Πολύ συχνά για να μπορέσουμε να εξάγουμε όλα τα απαραίτητα πρότυπα και στοιχεία από μια χρονοσειρά χρειάζεται να συνδυάσουμε τόσο τα ιστορικά στοιχεία της όσο και παράγοντες πρόβλεψης (μέσω παλινδρόμησης). Αυτά τα μοντέλα ονομάζονται **μοντέλα δυναμικής παλινδρόμησης.** Για να επιτευχθεί αυτό επεκτείνουμε τα μοντέλα ARIMA έτσι ώστε να επιτρέψουμε άλλες πληροφορίες να συμπεριληφθούν στα μοντέλα.

Ένα μοντέλο γραμμικής παλινδρόμησης μπορεί να εκφραστεί μέσω της πιο κάτω σχέσης

$$y_t = \beta_0 + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \varepsilon_t$$

Όπου ε_t συνήθως θεωρείται ο μη συσχετισμένος όρος σφάλματος (δηλαδή, είναι λευκός θόρυβος). Ένα μοντέλο ARIMA(p,d,q) αναπαριστάται μέσω της σχέσης:

$$(1 - \varphi_1 B - ... - \varphi_p B^p)(1 - B)^d y_t = c + (1 + \theta 1 B + ... + \theta_q B^q)\varepsilon_t$$

Για την κατασκευή του δυναμικού μοντέλου παλινδρόμησης θα επιτρέψουμε στα σφάλματα από μια παλινδρόμηση να περιέχουν αυτοσυχέτιση, ονομάζοντας τον όρο του σφάλματος ως η_t αντί ε_t. Η σειρά σφάλματος η_t θεωρείται ότι ακολουθεί ένα μοντέλο ARIMA. Επομένως η γενική μορφή του μοντέλου είναι η εξής

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \eta_t \\ (1 - \varphi_1 B - ... - \varphi_p B^p)(1 - B)^d \eta_t &= c + (1 + \theta 1 B + ... + \theta_q B^q)\varepsilon_t \end{aligned}$$

Γενικότερα αυτό το μοντέλο πέραν της καλής του απόδοσης, δίνει λύση στο πρόβλημα των αυτοσυσχετιζόμενων σφαλμάτων που μπορεί να προκύψουν από ένα απλό μοντέλο παλινδρόμησης. Τα αυτοσυσχετιζόμενα

σφάλματα υποδηλώνουν ότι το μοντέλο δέν αξιοποιεί επαρκώς τις διαθέσιμες πληροφορίες.

Για να εκτιμήσουμε τις παραμέτρους του μοντέλου, πρέπει να ελαχιστοποιήσουμε το άθροισμα των τετραγώνων των τιμών ε_t. Αν ελαχιστοποιήσουμε το άθροισμα των τετραγώνων των τιμών η_t (κάτι που θα συνέβαινε αν εκτιμούσαμε το μοντέλο παλινδρόμησης αγνοώντας τους αυτοσυσχετισμούς στα σφάλματα), θα προκύψουν πολλά προβλήματα. Ένα πολύ σημαντικό στοιχείο που πρέπει να λάβουμε υπόψη μας κατά την **εκτίμηση μιας παλινδρόμησης με σφάλματα ARMA** είναι ότι όλες οι μεταβλητές του μοντέλου θα πρέπει πρώτα να είναι στάσιμες. Επομένως αρχικά ελέγχουμε κατα πόσο η y_t και οι παράγοντες πρόβλεψης $x_{1,t}, ..., x_{k,t}$ είναι στάσιμοι. Σε διαφορετική περίπτωση, οι εκτιμώμενοι συντελεστές δέν θα είναι συνεπής εκτιμήσεις (συγκλίνουν κατα πιθανότητα στην τυχαία μεταβλητή που εκτιμούν). Συχνά είναι επιθυμητό να διατηρηθεί η μορφή της σχέσης μεταξύ του y_t και των παραγόντων πρόβλεψης, και κατά συνέπεια είναι σύνηθες να διαφοροποιούνται όλες οι μεταβλητές εαν κάποια από αυτές χρειάζεται διαφοροποίηση.

Εαν όλες οι μεταβλητές στο μοντέλο είναι στάσιμες, τότε πρέπει να λάβουμε υπόψη μόνο τα σφάλματα ARMA για τα υπόλοιπα. Δηλαδή να ακολουθεί μοντέλο ARIMA(p,0,q), και άρα να μήν έχει διαφοροποιηθεί.

Τα μοντέλα ARIMA, SARIMA (seasonal ARIMA) και ARIMAX αποδίδουν πολύ καλά για προβλέψεις μικρού χρονικού ορίζοντα αλλά δεν είναι τόσο καλές για μακροχρόνιες προβλέψεις.

Στοχαστικές και αιτιοκρατικές τάσεις

Αρκετά συχνά ως παράγοντας πρόβλεψης χρησιμοποιείται ο όρος της γραμμικής τάσης. Μια γραμμική τάση μπορεί να μοντελοποιηθεί με 2 διαφορετικούς τρόπους οι οποίοι ονομάζονται αιτιοκρατική και στοχαστική τάση αντίστοιχα.
Η αιτιοκρατική και στοχαστική τάση προκύπτουν από τα πιό κάτω μοντέλα

Αιτιοκρατική τάση

$$y_t = \beta_0 + \beta_1 t + \eta_t$$
$$\eta_t \sim ARIMA(p, 0, q)$$

Στοχαστική τάση

$$y_t = \beta_0 + \beta_1 t + \eta_t$$
$$\eta_t \sim ARIMA(p, 1, q)$$

Εαν όμως στην περίπτωση της στοχαστικής τάσης διαφοροποιήσουμε και τα 2 μέλη της εξίσωσης προκύπτει ότι:

$$y_t = y_{t-1} + \beta_1 + \eta'_t$$

όπου η η'_t είναι μια διαδικασία ARMA.

Αυτή η εξίσωση θυμίζει μια διαδικασία τυχαίου περιπάτου με περιπλάνηση με τη διαφορά ότι ο όρος

του σφάλματος είναι μια διαδικασία ARMA και όχι λευκός θόρυβος. Μια από τις κύριες διαφοροποιήσεις των μοντελοποιήσεων της αιτιοκρατικής και στοχαστικής τάσης είναι ότι οι στοχαστικές τάσεις έχουν πολύ ευρύτερα διαστήματα πρόβλεψης, επειδή τα σφάλματα δεν είναι στάσιμα. Οι σημειακές προβλέψεις που θα προκύψουν και από τα 2 μοντέλα δεν διαφέρουν ιδιαίτερα.

Με τις αιτιοκρατικές τάσεις υπάρχει μια έμμεση υπόθεση που είναι αρκετά δεσμευτική ως προς την αξιοπιστία των προβλέψεων. Υποθέτουν ότι η **κλίση της τάσης δεν πρόκειται να αλλάξει με την πάροδο του χρόνου**. Από την άλλη πλευρά, με τις στοχαστικές τάσεις μπορεί να αλλάξει και η εκτιμώμενη αύξηση θεωρείται ότι θα είναι μόνο η μέση αύξηση κατά την ιστορική περίοδο. Κατά συνέπεια, είναι ασφαλέστερο να προβλέπουμε με στοχαστικές τάσεις, ειδικά για μεγαλύτερους ορίζοντες πρόβλεψης, καθώς τα διαστήματα πρόβλεψης επιτρέπουν μεγαλύτερη αβεβαιότητα σε μελλοντική αύξηση.

7.7.1 Δυναμική Αρμονική Παλινδρόμηση

Όταν υπάρχουν μεγάλες εποχιακές περίοδοι (π.χ m=52, m=365), μια δυναμική παλινδρόμηση με όρους Fourier είναι συχνά καλύτερη επιλογή. Το εποχιακό μοντέλο εκθετικής εξομάλυνσης (SETS) επιτρέπει εποχικότητα εώς m=24 (καλύπτει δηλαδή εώς ωριαία δεδομένα). Το εποχιακό πρότυπο διαμορφώνεται χρησιμοποιώντας όρους Fourier με τις βραχυπρόθεσμες δυναμικές των χρονοσειρών να αντιμετωπίζονται από ένα σφάλμα ARMA. Το μοντέλο ορίζεται από την πιο κάτω σχέση:

$$
\begin{aligned}
y_t &= \beta_0 + \sum_{k=1}^{K} [\alpha_k s_k(t) + \gamma_k c_k(t)] + \eta_t \\
s_k(t) &= \sin(\frac{2\pi k t}{m}) \\
c_k(t) &= \cos(\frac{2\pi k t}{m}) \\
\eta_t &\sim \text{ μη εποχιακή ARIMA διαδικασία}
\end{aligned}
$$

Όπου α_k , γ_k είναι συντελεστές παλινδρόμησης. Το K δέν μπορεί να είναι μεγαλύτερο από m/2. Συνήθως για την επιλογή του βέλτιστου K στηριζόμαστε στην τιμή του AICc. Γενικά αυτό το εποχιακό μοντέλο σε σχέση με τα όσα είδαμε έως τώρα (SARIMA,SETS) έχει τα εξής πλεονεκτήματα:

- Μπορεί να χρησιμοποιηθεί για οποιαδήποτε μήκος εποχικότητας
- Μπορεί να χειριστεί δεδομένα με περισσότερες από μία εποχικότητα (π.χ ημερήσια δεδομένα, εμφανίζουν και εβδομαδιαία και χρονιαία εποχικότητα) συμπεριλαμβάνοντας όρους Fourier διαφορετικών συχνοτήτων
- Η βραχυπρόθεσμη δυναμική αντιμετωπίζεται εύκολα με ένα απλό σφάλμα ARMA
- Η ομαλότητα του εποχιακού προτύπου μπορεί να ελεγχθεί από το K, το πλήθος των Fourier ζευγαριών sin και cos. Όσο μικρότερο το K τόσο πιο ομαλό το εποχιακό μοτίβο.

Το σημαντικό μειονέκτημα αυτής της μεθόδου σε σχέση με ένα μοντέλο SARIMA είναι ότι η εποχικότητα θεωρείται σταθερή με την πάροδο του χρόνου. Επομένως καλό είναι να αξιοποιείται σε συστήματα όπου

το εποχιακό μοτίβο δεν φαίνεται να αλλάζει με την πάροδο του χρόνου και για χρονικό ορίζοντα που δεν ξεπερνά μια πλήρη επανάληψη όλων των εποχών. Στην πράξη αυτό το μοντέλο είναι ιδιαίτερα χρήσιμο για εβδομαδιαία δεδομένα με εποχικότητα m=52. Στην πραγματικότητα το m δέν είναι ακριβώς 52 αλλά ίσο με 365/7 που δέν είναι ακέραιος αριθμός.

Πιο κάτω βλέπουμε ένα παράδειγμα του μοντέλου αυτού για εβδομαδιαία (m=52) δεδομένα για τις πωλήσεις βαρελιών βενζίνης στις Η.Π.Α.

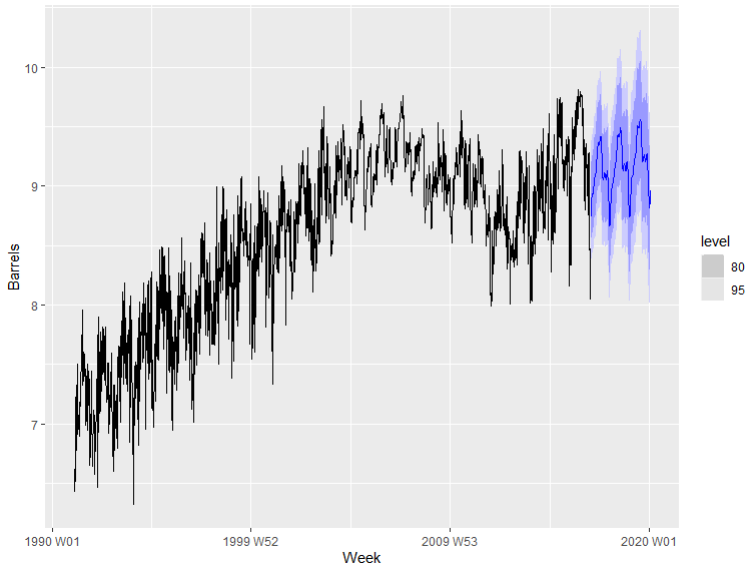


Figure 7.7.1: Δυναμική Αρμονική παλινδρόμηση για την πρόβλεψη μελλοντικών πωλήσεων βαρελιών βενζίνης στις Η.Π.Α

Στον πιο κάτω πίνακα βλέπουμε τους συντελεστές του δυναμικού αρμονικού μοντέλου παλινδρόμησης που προσαρμόσαμε στην γλώσσα προγραμματισμού R.

Table 7.7.1: Model coefficients for ARIMA with Fourier terms on US gasoline data

term	estimate
ma1	-0.8956
fourier(K = 6)C1_52	-0.1122
fourier(K = 6)S1_52	-0.2300
fourier(K = 6)C2_52	0.0419
fourier(K = 6)S2_52	0.0316
fourier(K = 6)C3_52	0.0832
fourier(K = 6)S3_52	0.0345
fourier(K = 6)C6_52	-0.0522
fourier(K = 6)S6_52	0.0002
intercept	0.0014

Με βάση λοιπόν τον πιο πάνω πίνακα μπορούμε να αντιληφθούμε ότι το μοντέλο που προσαρμόστηκε είναι το εξής:

$$y_t = 0.0014 - 0.2300\sin(\frac{2\pi t}{52}) - 0.1122cos(\frac{2\pi t}{2}) + ... + 0.0002sin(\frac{12\pi t}{52}) - 0.0522cos(\frac{12\pi t}{42}) + \varepsilon_t$$

$$(1-B)\eta_t = (1-0.8956B)\varepsilon_t \Leftrightarrow \eta_t - \eta_{t-1} = \varepsilon_t - 0.8956\varepsilon_{t-1} \quad (\eta_t \sim ARIMA(0,1,1))$$

7.7.2 Ειδικές περιπτώσεις παραγόντων πρόβλεψης

Κατά την μοντελοποίηση ενός μοντέλου πρόβλεψης πρέπει να λάβουμε υπόψη κάποια σημαντικά ζητήματα και να μοντελοποιηθούν με κατάλληλο τρόπο για να μπορέσουν να αντιμετωπιστούν μέσω του μοντέλου. Στην συνέχεια παρουσιάζονται κάποια ειδικά ζητήματα που πρέπει να λαμβάνονται υπόψη και πώς αυτά μοντελοποιούνται .

Παράγοντες πρόβλεψης με υστέρηση και κατανεμημένες υστερήσεις (Lagged predictors and Distributed lags)

Μερικές φορές οι επιπτώσεις ενός παράγοντα πρόβλεψης που περιλαμβάνεται σε ένα μοντέλο πρόβλεψης **δεν θα είναι άμεσες και απλές**. Για παράδειγμα μια διαφημιστική καμπάνια μιας εταιρίας που άρχισε τον Ιανουάριο, θα επηρεάσει τις πωλήσεις της τόσο τον Ιανουάριο όσο και των επόμενων 2-3 μηνών (η επίδραση κατανέμεται σε ένα χρονικό διάστημα 4 μηνών). Η επιρροή στις πωλήσεις στου επόμενους μήνες λογικά θα φθίνει με το πέρασμα του χρόνου και ίσως η επιρροή στις πωλήσεις να γίνει πιο εμφανής τον Φεβρουαρίου αντί το Ιανουάριο (μέχρι να αποδώσει καρπούς η διαφήμιση). Η μοντελοποίηση τέτοιων παραγόντων πρόβλεψης δίνεται από την σχέση

$$y_t = \beta_0 + \gamma_0 x_t + +\gamma_1 x_{t-1} + ... + \gamma_k x_{t-k} + \eta_t$$

Το k επιλέγεται με βάση κάποιο κριτήριο βελτιστοποίησης (π.χ AICc).

Τάση ως παράγοντας πρόβλεψης

Είναι σύνηθες στα δεδομένα χρονοσειρών να εκτιμάται η τάση τους. Μια γραμμική τάση μπορεί να μοντελοποιηθεί ως παράγοντας πρόβλεψης μέσω της σχέσης

$$y_t = \beta_1 t + \varepsilon_t$$

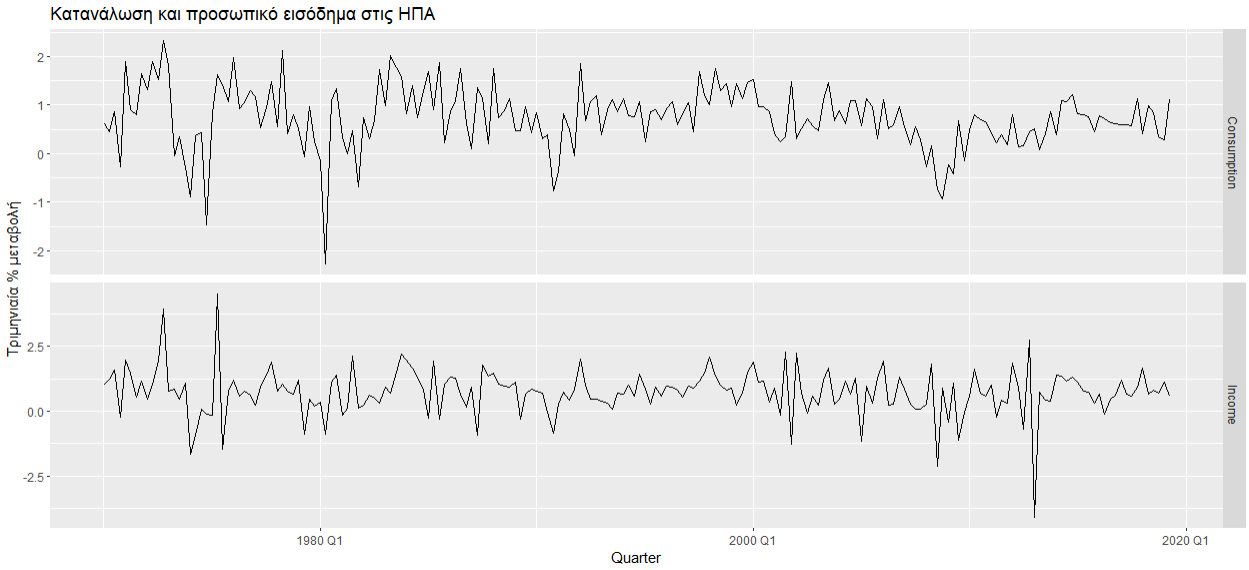
Μεταβλητές παρέμβασης (Intervention variables)

- **Μεταβλητές κορυφής (Spike):** Είναι ισοδύναμες με τις κατηγορικές μεταβλητές (dummy variables) για την αντιμετώπιση ακραίων τιμών (outliers).

- **Δείκτριες:** Δείκτριες μεταβλητές για την αντιμετώπιση επιδράσεων που είναι ενεργές σε συγκεκριμένο χρονικό διάστημα ή από κάποια χρονική στιγμή και έπειτα.

Εφαρμογή:

Θα κατασκευάσουμε ένα μοντέλο δυναμικής παλινδρόμησης για την πρόβλεψη των μέσων δαπανών ενός πολίτη ανάλογα με το εισόδημα του. Αρχικά βλέπουμε τα χρονοδιαγράμματα αυτών των μεταβλητών για να δούμε αν πράγματι φαίνεται να υπάρχει κάποια σχέση που να συνδέει τα έσοδα με τις δαπάνες.



Ποσοστιαίες μεταβολές στις τριμηνιαίες προσωπικές καταναλωτικές δαπάνες και προσωπικά διαθέσιμα έσοδα για τις ΗΠΑ, 1970 Q1 έως 2019 Q2

Φαίνεται να υπάρχει σύνδεση των 2 μεταβλητών. Είναι λογικό επίσης να υποθέσουμε ότι όταν ένα άτομο έχει μείωση στο εισόδημα του, ή όταν μείνει άνεργος, οι δαπάνες του δεν μειώνονται κατευθείαν αλλά μετά απο ένα χρονικό διάστημα. Επομένως θα χειριστούμε τη μεταβλητή του εισοδήματος ως παράγοντα πρόβλεψης με υστέρηση. Μετά απο μελέτη τόσο της σημαντικότητα των υστερήσεων της μεταβλητής εισοδήματος όσο και το κριτήριο AICc επιλέχθηκε να χρησιμοποιηθεί μόνο μία υστέρηση της μεταβλητής εισοδήματος. Τέλος οι δαπάνες φαίνεται να ακολουθούν ένα εποχιακό πρότυπο περιόδου 4 (μεγαλύτερες δαπάνες το καλοκαίρι και τον χειμώνα). Πιο κάτω βλέπουμε το διάγραμμα σκέδασης της μεταβλητής δαπανών συναρτήσει της μεταβλητής εισοδήματος αλλά και συναρτήσει της μεταβλητής που αθροίζει το εισόδημα με τη κατα μία χρονικά υστερημένη τιμή της.

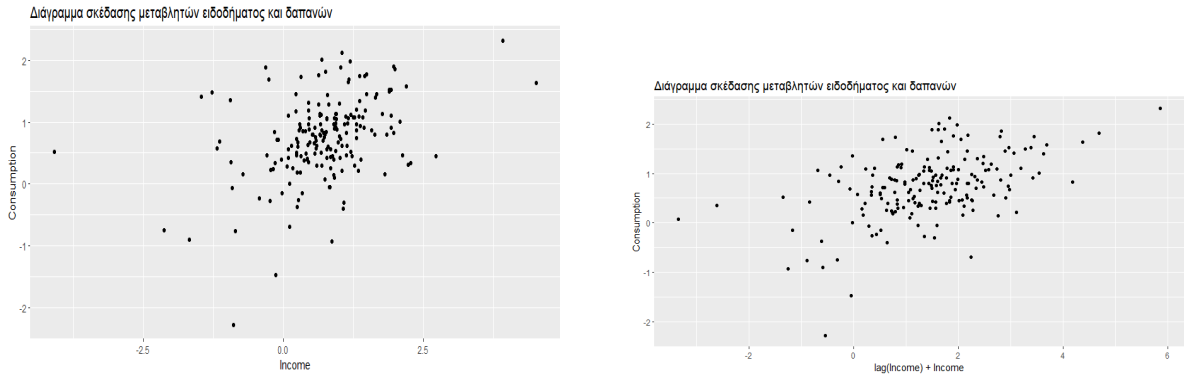


Figure 7.7.2: Διαγράμματα σκέδασης δαπανών συναρτήσει εισοδήμτος

Προτού προχωρήσουμε στην προσαρμογή οποιουδήποτε μοντέλου, ελέγχουμε κατα πόσο τα δεδομένα είναι στάσιμα. Απο το χρονοδιάγραμμα φαίνεται να είναι. Για να έχουμε όμως καλύτερη εικόνα εξετάζουμε την στασιμότητα της χρονοσειράς με τον στατιστικό έλεγχο KPSS. Η p-value του ελέγχου είναι μεγαλύτερη από 0.1, επομένως δέν υπάρχουν ισχυρές ενδείξεις για να απορριφθεί η υπόθεσης της στασιμότητας (μηδενική υπόθεση του ελέγχου). Με τον ίδιο στατιστικό έλεγχο, και η χρονοσειρά του εισοδήματος προκύπτει να είναι στάσιμη.

Έτσι, αφού χωρίσαμε το σύνολο των δεδομένων σε σύνολο εκπαιδευσης και σύνολο ελέγχου προσαρμόσαμε το πιο κάτω μοντέλο δυναμικής παλινδρόμησης.

$$y_t = 0.4795 + 0.2072I_t + 0.1590I_{t-1} + \eta_t$$
$$(1 - 0.0949B - 0.1745B^2 - 0.2193B^3)(1 + 0.0483B^4 + 0.1243B^8)\eta_t = \varepsilon_t$$
$$\varepsilon_t \sim NID(0, 0.2999)$$

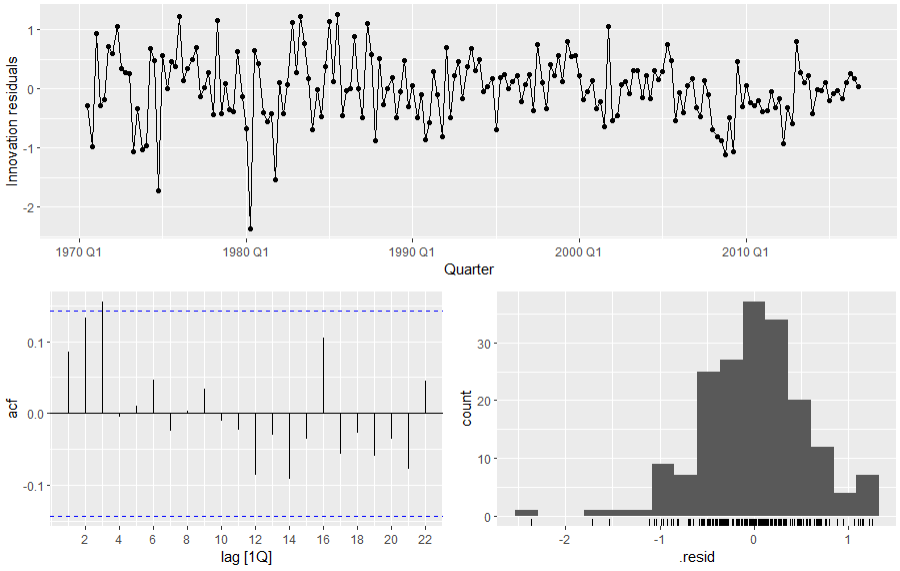
Όλοι οι συντελεστές του μοντέλου είναι στατιστικά σημαντικοί επομένως δέν αφαιρούμε κάποιο επεξηγηματικό παράγοντα.

```
Series: Consumption
Model: LM w/ ARIMA(3,0,0)(2,0,0)[4] errors

Coefficients:
      ar1     ar2     ar3     sar1     sar2     Income  lag(Income)  intercept
     0.0949  0.1745  0.2193  -0.0483  -0.1243  0.2072   0.1590        0.4795
s.e. 0.0745  0.0742  0.0748   0.0786   0.0752  0.0472    0.0467       0.0842


sigma^2 estimated as 0.2999:  log likelihood=-149.36
AIC=316.71   AICc=317.72   BIC=345.84
```

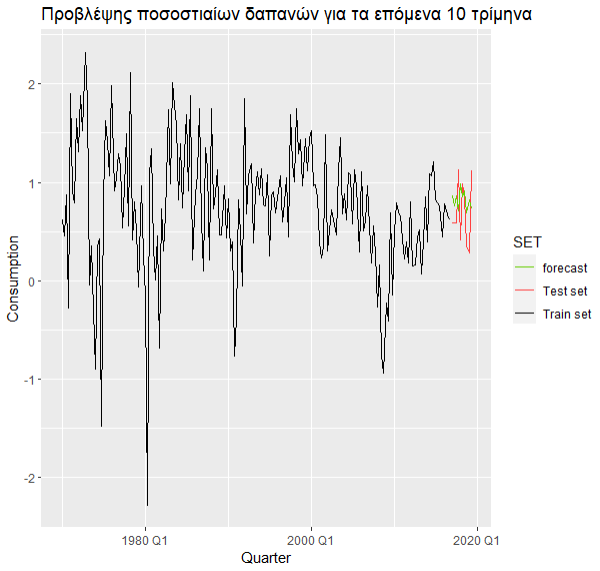
Στην συνέχεια παρουσιάζουμε τους διαγνωστικούς ελέγχους υπολοίπων (του μοντέλου ARIMA) που επιβεβαιώνουν ότι τα καινοτόμα υπόλοιπα δέν διαφέρουν σημαντικά απο τον λευκό θόρυβο:



Τα καινοτόμα υπόλοιπα (δηλαδή, τα εκτιμώμενα σφάλματα ARIMA) δεν διαφέρουν σημαντικά από το λευκό θόρυβο.

Επίσης πραγματοποιούμε τον στατιστικό έλεγχο αυτοσυχέτισης Portmanteau και πιο συγκεκριμένα τον έλεγχο Ljung-Box για τις 10 πρώτες αυτοσυχετίσεις. Ο έλεγχος αυτός δίνει p-value=0.982 επομένως τα υπόλοιπα φαίνεται ότι δεν ξεχωρίζουν από μια σειρά λευκού θορύβου. Η υπόθεση κανονικότητας δέν είναι παράλογη αλλά ούτε εύκολα αποδεκτή. Με βάση τον έλεγχο καλής προσαρμογής Shapiro-Wilk η κατανομή των υπολοίπων δέν φαίνεται να είναι κανονική. Αυτό μας κάνει να μήν μπορούμε να ερμηνεύσουμε τους συντελεστές που εκτιμήθηκαν, και τα διαστήματα εμπιστοσύνης θα πρέπει να προκύψουν μέσω δειγματοθετημένων υπολοίπων.

Στην συνέχεια πραγματοποιούμε προβλέψεις για το σύνολο ελέγχου.



Παρατηρούμε ότι η απόδοση του μοντέλου δέν είναι ικανοποιητική. Παρόλο λοιπόν που τα υπόλοιπα δέν διαφέρουν ουσιαστικά απο λευκό θόρυβο, πολλή πληροφορία της χρονοσειράς δέν καταγράφηκε απο αυτό το μοντέλο. Χρειάζονται σίγουρα και άλλες επεξηγηματικές μεταβλητές όπως για παράδειγμα η επιρροή διακοπών (μή στοχαστική ποσότητα), οι τιμές βασικών αγαθών και υπηρεσιών (στοχαστική ποσότητα). Βλέπουμε λοιπόν ότι η ικανοποίηση των διαγνωστικών ελέγχων για τα υπόλοιπα δέν είναι αρκετή για να καταλήξουμε σε ένα ικανό μοντέλο πρόβλεψης.

8 Πρόβλεψη Ιεραρχικών και ομαδοποιημένων χρονοσειρών

Αρκετές φορές έχουμε στην διάθεση μας μεγάλο όγκο δεδομένων με αποτέλεσμα αναλόγως κάποιων χαρακτηριστικών/ιδιοτήτων να μπορούμε να περιοριστούμε σε ένα υπόχωρο του συνολικού χώρου και να δημιουργούμε διάφορες υποσειρές. Για παράδειγμα τα δεδομένα πωλήσεων της Apple μπορούν να χωριστούν ανάλογα με την γεωγραφική τοποθεσία (κατά χώρα ή και κατά πολιτείες), κατά προϊόν κ.τ.λ.

Ο διαχωρισμός μιας χρονοσειράς σε υποσειρές με βάση κάποια χαρακτηριστικά μπορεί να προκύψει σε 2 μορφές οι οποίες είναι οι εξής:

Ιεραρχικές χρονοσειρές:
Προκύπτουν συχνά λόγω γεωγραφικών διαιρέσεων. Για παράδειγμα, οι συνολικές πωλήσεις της Apple μπορούν να κατηγοριοποιηθούν ανά χώρα, στη συνέχεια σε κάθε χώρα ανά πολιτεία, σε κάθε πολιτεία ανά περιοχή και ούτω καθεξής μέχρι το τελικό επίπεδο.

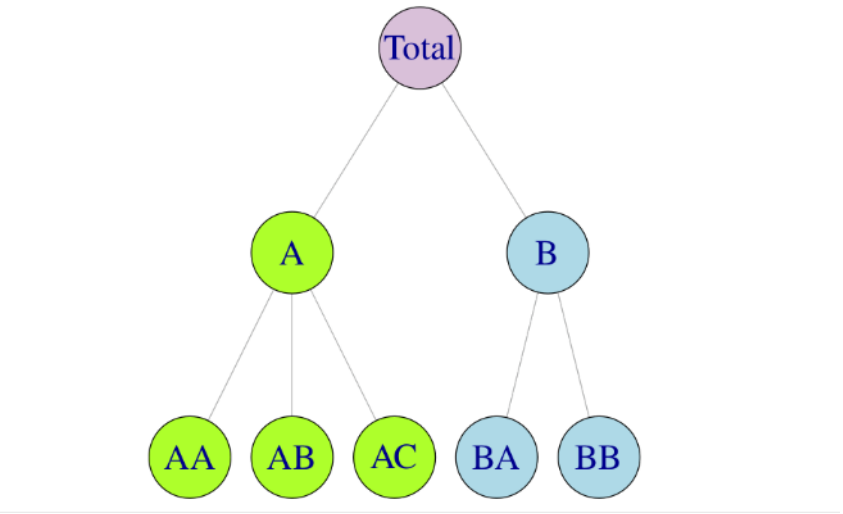


Figure 8.0.1: Ιεραρχικό διάγραμμα δύο επιπέδων

Είναι σημαντικό να αναφερθεί ότι για οποιαδήποτε χρονική στιγμή t, οι παρατηρήσεις στο κατώτατο επίπεδο της ιεραρχίας (σειρές παιδιά) αθροίζονται στις παρατηρήσεις της παραπάνω σειράς (σειρά πατέρας). Για παράδειγμα για το πιο πάνω σχήμα θα ισχύει

$$y_t = y_{AA,t} + y_{AB,t} + y_{AC,t} + y_{BA,t} + y_{BB,t},$$
$$y_{A,t} = y_{AA,t} + y_{AB,t} + y_{AC,t}, \qquad y_{B,t} = y_{BA,t} + y_{BB,t}$$

Ομαδοποιημένες χρονοσειρές:
Προκύπτουν όταν τα χαρακτηριστικά ενδιαφέροντος επικαλύπτονται και δεν είναι εμφωλευμένα. Για παράδειγμα, ένας ιδιοκτήτης οινοποιείου μπορεί να ενδιαφέρεται για πωλήσεις που βασίζονται σε χαρακτηριστικά, όπως το μέγεθος της φιάλης, το είδος (ξηρό, γλυκό, ημίγλυκο), το εύρος τιμών κλπ. Τέτοια χαρακτηριστικά δεν διαχωρίζονται φυσικά με μοναδικό ιεραρχικό τρόπο, καθώς τα χαρακτηριστικά αυτά δεν είναι εμφωλευμένα. Οι ομαδοποιημένες χρονοσειρές μπορεί μερικές φορές να θεωρηθούν ως ιεραρχικές χρονοσειρές που δεν επιβάλλουν μια μοναδική ιεραρχική δομή, με την έννοια ότι η σειρά με την οποία μπορεί να ομαδοποιηθεί η χρονοσειρά δεν είναι μοναδική.

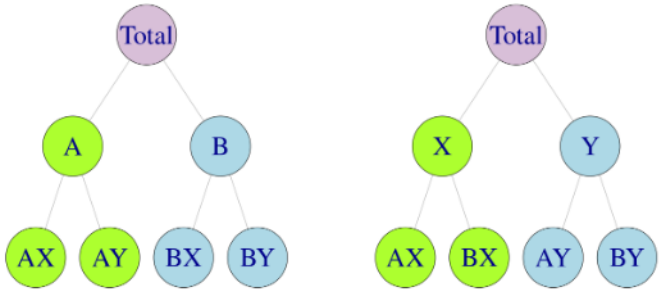


Figure 8.0.2: Εναλλακτικές αναπαραστάσεις ομαδοποιημένης δομής 2 επιπέδων

Οι δύο πιο πάνω δομές μπορούν συνδυαστούν και τότε προκύπτουν πιο περίπλοκες δομές. Συχνά κατα την διαδικασία πρόβλεψης θέλουμε να έχουμε προβλέψεις τόσο για τις αναλυτικές σειρές (δεν διαιρούνται άλλο) αλλά και για τις αθροιστικές (αυτές που εχουν υποδιαιρεθεί). Επομένως είναι φυσικό να θέλουμε να προσθέτονται οι προβλέψεις με τον ίδιο τρόπο όπως τα δεδομένα. Για παράδειγμα, το άθροισμα των προβλέψεων των περιφερειακών πωλήσεων θα πρέπει να είναι ίσο με το άθροισμα των προβλέψεων των πολιτειακών πωλήσεων, που με τη σειρά του θα πρέπει να είναι ίσο με την πρόβλεψη των εθνικών πωλήσεων.

Μια σημαντική απαίτηση όταν δημιουργούμε μια δομή χρονοσειρών είναι ότι οι προβλέψεις πρέπει **να είναι συνεκτικές σε ολόκληρη την δομή συγκέντρωσης**. Δηλαδή, απαιτούμε να αθροίζονται οι προβλέψεις με τρόπο που να είναι σύμφωνος με τη δομή συγκέντρωσης της ιεραρχίας ή της ομάδας που καθορίζει τη συλλογή χρονολογικών σειρών.

Όπως αναφέραμε και πιο πάνω η διαδικασία δημιουργίας ιεραρχικής ή ομαδοποιημένης δομής για μια χρονοσειρά μπορεί να αποδειχθεί αποτελεσματική για την δημιουργία καλών προβλέψεων και την αναγνώριση των κρυμμένων χαρακτηριστικών σε μια χρονοσειρά.

Παραδοσιακά, οι προβλέψεις ιεραρχικών ή ομαδοποιημένων χρονοσειρών περιελάμβαναν την **επιλογή ενός επιπέδου συγκέντρωσης** και τη δημιουργία προβλέψεων για αυτό το επίπεδο. Στην συνέχεια αυτές οι προβλέψεις **συγκεντρώνονται σε υψηλότερα επίπεδα είτε διαχωρίζονται για χαμηλότερα επίπεδα** , έτσι ώστε να προκύψουν οι **συνεκτικές προβλέψεις** για το υπόλοιπο

της δομής.

Οι κλασσικές μέθοδοι δημιουργίας συνεκτικών προβλέψεων είναι οι 2 ακόλουθες:

- Η προσέγγιση από κάτω προς τα πάνω (bottom-up)
- Προσεγγίσεις από πάνω προς τα κάτω (top-down)

Η προσέγγιση από κάτω προς τα πάνω

Με αυτή την προσέγγιση αρχικά δημιουργούνται προβλέψεις για κάθε σειρά στο κατώτερο επίπεδο ("Αρχικές προβλέψεις") και στην συνέχεια αυτές αθροίζονται για την παραγωγή προβλέψεων για όλες τις σειρές της δομής.

Προσεγγίσεις από πάνω προς τα κάτω

Η προσέγγιση αυτή προϋποθέτει αρχικά να γίνουν προβλέψεις για τη συνολική σειρά y_t και στην συνέχεια την κατανομή των προβλέψεων αυτών κάτω στην ιεραρχία (στις υπόλοιπες υπο-σειρές). Πώς όμως ακριβώς γίνετε η κατανομή μιας πρόβλεψης της αρχικής σειράς στις υποσειρές που προκύπτουν σε κάθε επίπεδο διαίρεσης (μέσω ενός διακριτικού χαρακτηριστικού);
Θεωρούμε ένα σύνολο αναλογιών $p_1, p_2, ...p_m$ που καθορίζουν τον τρόπο κατανομής των προβλέψεων των συνολικών σειρών για τη λήψη προβλέψεων για κάθε σειρά στο **κατώτερο επίπεδο της δομής**. Για παράδειγμα εαν χρησιμοποιήσουμε αυτή την προσέγγιση για προβλέψεις στην δομή του σχήματος 8.0.1. Οι προβλέψεις του κατώτερου επιπέδου προκύπτουν με τον ακόλουθο τρόπο.

$$\tilde{y}_{AA,t} = p_1\hat{y}_t \ , \ \tilde{y}_{AB,t} = p_2\hat{y}_t \ , \ \tilde{y}_{AC,t} = p_3\hat{y}_t \ , \ \tilde{y}_{BA,t} = p_4\hat{y}_t \ , \ \tilde{y}_{BB,t} = p_5\hat{y}_t$$

Μόλις δημιουργηθούν προβλέψεις κατώτατου επιπέδου h-βημάτων μπροστά, αυτές συναθροίζονται για να δημιουργήσουν συνεκτικές προβλέψεις για τις υπόλοιπες σειρές.

Το μοναδικό αναπάντητο ερώτημα είναι πώς προκύπτουν οι τιμές των αναλογιών p_j για $j = 1, 2..., m$. Υπάρχουν διαφορετικές μέθοδοι κατασκευής των αναλογιων με μερικές από αυτές να αναλύονται πιό κάτω

Μέσες ιστορικές αναλογίες

$$p_j = \frac{1}{T}\sum_{t=1}^{T}\frac{y_{j,t}}{y_t}, j = 1, ..., m$$

Αναλογίες των ιστορικών μέσων όρων

$$p_j = \frac{\sum_{t=1}^{T}\frac{y_{j,t}}{T}}{\sum_{t=1}^{T}\frac{y_t}{T}}, j = 1, ..., m$$

Αναλογίες προβλέψεων

Οι ιστορικές αναλογίες που είδαμε εώς τώρα, αγνοούν τον τρόπο με τον οποίο οι αναλογίες μπορούν να αλλάξουν με την πάροδο του χρόνου. Επιπλέον οι προσεγγίσεις από πάνω προς τα κάτω με βάση ιστορικές αναλογίες τείνουν να παράγουν λιγότερο ακριβείς προβλέψεις σε χαμηλότερα επίπεδα της ιεραρχίας σε σχέση με τις από κάτω προς τα πάνω προσεγγίσεις. Οι αναλογίες προβλέψεων αξιοποιούν πληροφορίες που προκύπτουν από προβλέψεις μελλοντικών τιμών για να δημιουργήσουν τις αναλογίες και με αυτό το τρόπο αντιμετωπίζουν το πρόβλημα που μόλις θίξαμε. Για την ακρίβεια πρώτα δημιουργούμε "αρχικές προβλέψεις" για όλες τις σειρές της δομής (δεν είναι συνεκτικές). Ο γενικός κανόνας λήψης των αναλογικών προβλέψεων είναι ο ακόλουθος:

$$p_j = \prod_{l=0}^{K-1} \frac{\hat{y}_{j,h}^{(l)}}{\hat{S}_{j,h}^{(l+1)}}, j = 1, ..., m$$

όπου $\hat{y}_{j,h}^{(l)}$ η αρχική πρόβλεψη της σειράς (h-βήματα μπροστά) που αντιστοιχεί στον κόμβο που είναι l επίπεδα πάνω από j. $\hat{S}_{j,h}^{(l+1)}$ είναι το άθροισμα των αρχικών προβλέψεων h-βήματα μπροστά **κάτω από τον κόμβο** που είναι l επίπεδα πάνω από τον κόμβο j και συνδέονται απευθείας με αυτόν τον κόμβο. Το K εκφράζει το πλήθος των επιπέδων στην ιεραρχία.

Ένα μειονέκτημα όλων των προσεγγίσεων από πάνω προς τα κάτω, είναι ότι δεν παράγουν αμερόληπτες συνεκτικές προβλέψεις (Hyndman, Ahmed, Athanasopoulos, Shang, 2011) ακόμη και αν οι αρχικές προβλέψεις είναι αμερόληπτες.

8.1 Συμβιβαστική πρόβλεψη (Forecast reconciliation)

Παρατηρήθηκε ότι οι πιο πάνω μέθοδοι δημιουργίας προβλέψεων είχαν αρκετό περιθώριο βελτίωσης. Για να γίνει αντιληπτή η δυνατότητα παραγωγής καλύτερων προβλέψεων μια πρώτη προσέγγιση ήταν η αναπαράσταση των συναθροίσεων μέσω σημειογραφίας πινάκων αντί αναλυτικών εξισώσεων. Για παράδειγμα η ιεραρχική δομή του σχήματος 8.0.1 μπορούμε να γράψουμε

$$\begin{bmatrix} y_t \\ y_{A,t} \\ y_{B,t} \\ y_{AA,t} \\ y_{AB,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BA,t} \\ y_{BB,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{AA,t} \\ y_{AB,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BA,t} \\ y_{BB,t} \end{bmatrix}$$

ή οποία σε συμπιεσμένη μορφή εκφράζεται ως

$$y_t = Sb_t$$

Για να έχουμε όμως ένα γενικότερο ορισμό που να αντιπροσωπεύουμε όλες τις μεθόδους πρόβλεψης για ιεραρχικές ή ομαδοποιημένες χρονοσειρές χρησιμοποιώντας ένα κοινό συμβολισμό ακολουθούμε την ακόλουθη προσέγγιση.

Υποθέτουμε ότι προβλέπουμε όλες τις σειρές και αυτές τις προβλέψεις τις ονομάζουμε **βασικές προβλέψεις** και συμβολίζονται με $\hat{y}_h$. Στην συνέχεια όλες οι προσεγγίσεις συνεκτικών προβλέψεων για ιεραρχικές ή ομαδοποιημένες δομές μπορούν να αναπαρασταθούν ως

$$\tilde{y}_h = SG\hat{y}_h$$

όπου G είναι ο πίνακας που απεικονίζει τις αρχικές προβλέψεις στο κατώτερο επίπεδο και ο αθροιστικός πίνακας S τις αθροίζει χρησιμοποιώντας την δομή συνάθροισης.

Ο πίνακας G καθορίζεται αναλόγως ποια μέθοδος χρησιμοποιείται. Για τις μεθόδους που αναλύσαμε πιο πάνω, "από κάτω προς τα πάνω " και "από πάνω προς τα κάτω" για την ιεραρχία του σχήματος 8.0.1 ο πίνακας G προκύπτει αντίστοιχα ως

$$G = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad , \quad G = \begin{bmatrix} p_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

8.2 Συμβιβαστική Πρόβλεψη-Η προσέγγιση βέλτιστου συμβιβασμού MinT (ή Minimum Trace)

Όπως είδαμε πιο πάνω η εφαρμογή του τελεστή SG σε οποιουδήποτε συνόλου βασικών προβλέψεων ($\hat{y}_h$) θα επιστρέψει ένα σύνολο συνεκτικών προβλέψεων. ($\tilde{y}_h$).

Το βασικό πρόβλημα των κλασσικών μεθόδων (για ιεραρχικές ή ομαδοποιημένες χρονοσειρές) είναι ότι αξιοποιούν περιορισμένες πληροφορίες από το σύνολο της δομής. Πιό συγκεκριμένα χρησιμοποιούν μόνο πληροφορίες από τις βασικές προβλέψεις στο επίπεδο συνάθροισης που είτε έχουν συγκεντρωθεί είτε διαχωριστεί για τη λήψη προβλέψεων σε όλα τα άλλα επίπεδα.

Με βάσης λοιπόν την σχέση $\tilde{y}_h = SG\hat{y}_h$ μπορούμε να σκεφτούμε διαφορετικές επιλογές για τον πίνακα G έτσι ώστε να λαμβάνουμε τις συνεκτικές προβλέψεις. Στην πραγματικότητα, μπορούμε να βρούμε τον βέλτιστο πίνακα G (ως προς κάποιο κατάλληλο κριτήριο) για να δώσουμε πιο ακριβείς συμβιβαστικές προβλέψεις. Μία προσέγγιση για βελτιστοποίηση είναι να βρεθεί ο πίνακας G που ελαχιστοποιεί την συνολική διακύμανση πρόβλεψης του συνόλου των συνεκτικών προβλέψεων. Με αυτή την προσέγγιση ασχολήθηκαν οι Wickramasuriya, Athanasopoulos, και Hyndman ([22]) και κατέληξαν στον βέλτιστο πίνακα G.

Ας υποθέσουμε ότι δημιουργούμε συνεκτικές προβλέψεις χρησιμοποιώντας την εξίσωση $\tilde{y}_h = SG\hat{y}_h$. Η πρώτη απαίτηση για τον πίνακα G ήταν να είναι τέτοιος ώστε να προκύπτουν αμερόληπτες προβλέψεις. Στην εργασία [10] απέδειξαν ότι εαν οι βασικές προβλέψεις $\hat{y}_h$ είναι αμερόληπτες, τότε οι συνεκτικές προβλέψεις $\tilde{y}_h$ θα είναι αμερόληπτες, με την προϋπόθεση ότι $SGS = S$.
Ο πίνακας διακύμανσης-συνδιακύμανσης των σφαλμάτων των συνεκτικών προβλέψεων h-βημάτων μπροστά δίνεται από

$$\mathbb{V}_h = Var[y_{T+h} - \tilde{y}_h] = SGW_hG'S'$$

Όπου $W_h = Var[y_{T+h} - \hat{y}_h]$ είναι ο πίνακας διακύμανσης-συνδιακύμανσης από τα αντίστοιχα σφάλματα της αρχικής πρόβλεψης. Επομένως αναζητούμε το βέλτιστο G που ελαχιστοποιεί τις διακυμάνσεις σφάλματος των συνεκτικών προβλέψεων. Επειδή οι διακυμάνσεις σφάλματος βρίσκονται στη διαγώνιο του πίνακα V_h, το άθροισμα όλων των διακυμάνσεων του σφάλματος δίνεται από το ίχνος του πίνακα V_h. Για αυτό τον λόγο η προσέγγιση που θα προκύψει με την χρήση αυτού του βέλτιστου πίνακα G ονομάζεται **προσέγγιση βέλτιστου συμβιβασμού MinT (ή Minimum Trace)**

Ο πίνακας G που ελαχιστοποιεί το ίχνος του V_h έτσι ώστε $SGS = S$, δίνεται από την σχέση

$$G = (S'W_h^{-1}S)^{-1}S'W_h^{-1}$$

Επομένως οι βέλτιστες συμβιβαστικές προβλέψεις δίνονται από

$$\tilde{y}_h = S(S'W_h^{-1}S)^{-1}S'W_h^{-1}\hat{y}_h \tag{8.2.1}$$

Το κύριο πρόβλημα σε αυτή την προσέγγιση είναι ο υπολογισμός του πίνακα W_h που συχνά δεν μπορεί να γίνει με αναλυτικό τρόπο. Επομένως χρησιμοποιούμε προσεγγίσεις αυτού του πίνακα. Μερικές από αυτές που λειτουργούν πολύ καλά στην πράξη είναι οι εξής

- $W_h = k_h\mathbf{I}$, όπου $k_h > 0$ σταθερά αναλογικότητας η οποία δεν χρειάζεται να καθοριστεί καθώς όταν αντικατασταθεί αυτός ο W_h στην σχέση που δίνει τις συνεκτικές προβλέψεις, διαγράφεται.

- $W_h = k_h diag(\hat{W}_1)$, $k_h > 0$, $\hat{W}_1 = 1/T\sum_{t=1}^{T} e_t e_t'$ και e_t είναι ένα n-διαστατο διάνυσμα με τα υπόλοιπα των μοντέλων που δημιουργούν τις βασικές προβλέψεις τοποθετημένα με την ίδια σειρά όπως και τα δεδομένα.

- $W_h = k_h\mathbf{\Lambda}, k_h > 0, \Lambda = diag(S\mathbf{1})$ όπου $\mathbf{1}$ είναι ένα μοναδιαίο διάνυσμα διάστασης m(ο αριθμός των σειρών κατώτατου επιπέδου). Προϋπόθεση αυτής της προσέγγισης είναι ότι κάθε σφάλμα στο κατώτατο επίπεδο βασικής πρόβλεψης έχει διακύμανση k_h και δεν σχετίζονται μεταξύ των κόμβων.

- $W_h = k_h W_1$. Εδώ υποθέτουμε μόνο ότι οι πίνακες συνδιακύμανσης σφάλματος είναι ανάλογοι μεταξύ τους και εκτιμούμε άμεσα τον πλήρη πίνακα συνδιακύμανσης ενός βήματος W_1

9 Το μοντέλο Prophet

Αυτό το μοντέλο προβλέψεων παρουσιάστηκε από την Facebook(2018) με αρχικό σκοπό την πρόβλεψη ημερήσιων δεδομένων με εβδομαδιαία και ετήσια εποχικότητα συμπεριλαμβάνοντας και την επίδραση των διακοπών, αλλά στην συνέχεια επεκτάθηκε για να αντιμετωπίζει και άλλου είδους εποχικότητες. Είναι ένα μοντέλο που λειτουργεί καλύτερα με χρονοσειρές που έχουν ισχυρή εποχικότητα και πολλές εποχές ιστορικών δεδομένων. Σχεδιάστηκε για να έχει διαισθητικούς και ερμηνεύσιμους παράγοντες που μπορούν να ρυθμιστούν χωρίς να χρειάζεται κάποιος να γνωρίζει εις βάθος την θεωρία του υποκείμενου μοντέλου. Αποτελεί ένα μη γραμμικό μοντέλο παλινδρόμησης, και εντάσσεται στα γενικευμένα προσθετικά μοντέλα-GAMs, με γενικό τύπο:

$$y_t = g(t) + s(t) + h(t) + \varepsilon_t$$

όπου

- g(t) είναι όρος της τάσης ("όρος ανάπτυξης")
- s(t) περιγράφει τα διάφορα εποχιακά μοτίβα
- h(t) καταγράφει την επίδραση διακοπών
- ε_t λευκός θόρυβος

Προτού εξηγήσουμε περαιτέρω τις συνιστώσες του μοντέλου αναφέρουμε ποιες αδυναμίες άλλων μοντέλων παρατήρησαν κατασκευαστές του μοντέλου αυτού, τις οποίες είχαν σκοπό να αντιμετωπίσουν.

- Οι αυτοματοποιημένες (οι παράμετροι επιλέγονται αυτόματα με βάση ένα κριτήριο βελτιστοποίησης) προβλέψεις με μοντέλα ARIMA είναι επιρρεπής σε μεγάλα σφάλματα στην εκτίμηση της τάσης όταν παρουσιάζεται αλλαγή συμπεριφοράς της τάσης κοντά στο σημείο όπου σταματάμε την εκπαιδευση του μοντέλου.

- Η μέθοδος εκθετικής εξομάλυνσης και εποχιακή naive μέθοδος μπορούν να αντιληφθούν εβδομαδιαία εποχικότητα αλλά αποτυγχάνουν να αντιληφθούν εποχικότητες μεγαλύτερης συχνότητας. Για παράδειγμα μπορούν να "υπερ-αντιδράσουν" σε μια πτώση της τιμής ενδιαφέροντος κοντά στο τέλος της χρονιάς επειδή δέν μοντελοποιούν επαρκώς ετήσια εποχικότητα, σύμφωνα με την οποία στο τέλος κάθε χρόνου η τιμή της μεταβλητής παρουσιάζει μια πτώση.

Η διατύπωση ενός μοντελου GAM έχει το πλεονέκτημα ότι

- Προσαρμόζεται πολύ γρήγορα χρησιμοποιώντας είτε αλγόριθμο Backfitting είτε L-BGFS(1995) ο οποίος είναι ένας αλγόριθμος βελτιστοποίησης που ανήκει στην οικογένεια των Quasi-Newton μεθόδων και προσεγγίζει τον BFGS αλγόριθμο χρησιμοποιώντας περιορισμένη ποσότητα υπολογιστικής μνήμης (πολύ γνωστός σε προβλήματα Μηχανικής μάθησης).

- Το μοντέλο προσαρμόζεται εύκολα σε κατάλληλες συνιστώσες ανάλογα με τις ανάγκες. Για παράδειγμα όταν στα ιστορικά δεδομένα εντοπιστεί μια νέα πηγή εποχικότητας οι συνιστώσες προσαρμόζονται κατάλληλα για να την συμπεριλάβουν στην μοντελοποίηση

Το μοντέλο αυτό διαμορφώνει το πρόβλημα της πρόβλεψης ως ένα πρόβλημα προσαρμογής καμπύλης το οποίο είναι εγγενώς διαφορετικό από τα μοντέλα χρονοσειρών που λαμβάνουν ρητά υπόψη τους την χρονική εξάρτηση στα δεδομένα. Τα πλεονεκτήματα αυτής της προσέγγισης είναι τα εξής

- Μπορούμε να προσαρμόσουμε εποχικότητες με πολλαπλές περιόδους.

- Σε αντίθεση με τα μοντέλα ARIMA οι μετρήσεις δέν χρειάζεται να είναι σε τακτά χρονικά διαστήματα και δέν χρειάζεται να παρεμβάλλουμε τις ελλειπούσες τιμές.

- Προκύπτουν εύκολα ερμηνεύσιμες παράμετροι (βασικό χαρακτηριστικό μοντέλων παλινδρόμησης)

9.1 Μοντελοποίηση της τάσης (όρου ανάπτυξης) g(t)

Οι κατασκευαστές του μοντέλου αντιμετώπισαν τον όρο της τάσης ως ένα όρο ανάπτυξης όπως ακριβώς κάνουμε στην διαδικασία μοντελοποίησης της ανάπτυξης ενός πληθυσμού, επηρεασμένοι από τον τρόπο που ερμήνευσαν την ανάπτυξη της εταιρίας Facebook. Για την πρόβλεψη της ανάπτυξης λοιπόν χρειάζεται να απαντήσουμε σε 2 ερωτήματα. Πώς έχει αυξηθεί ο πληθυσμός εώς τώρα και πώς αναμένεται να συνεχίσει να αναπτύσσεται.

Επομένως μια προσέγγιση κατα την μοντελοποίηση του όρου ανάπτυξης είναι να θεωρηθεί ότι είναι μη γραμμικός και ότι έχει ένα μέγιστο όριο ανάπτυξης, πέραν του οποίου δεν μπορεί να παρατηρηθεί επιπλέον ανάπτυξη.

9.1.1 Μη γραμμικός όρος ανάπτυξης, με φραγμένη ανάπτυξη

Ένα παράδειγμα αυτού του είδους ρυθμού ανάπτυξης αποτελεί ένα λογιστικό μοντέλο ανάπτυξης (logistic growth model) το οποίο ορίζεται ως

$$g(t) = \frac{C(t)}{1 + exp(-k(t)(t - \gamma(t)))}$$

όπου C(t) η μέγιστη τιμή που μπορεί να λάβει ο πληθυσμός ο οποίος μπορεί να αλλάζει με την πάροδο του χρόνου. Η παράμετρος k(t) εκφράζει τον δυναμικό ρυθμό ανάπτυξης και γ(t) είναι δυναμικός όρος αντιστάθμισης (offset parameter) ο οποίος ρυθμίζει και μετατοπίζει κατάλληλα ολόκληρη την καμπύλη παράλληλα στον y-άξονα με σκοπό η τελική μορφή της g(t) να είναι μια ομαλή συνάρτηση (χωρίς ασυνέχειες).

Για την ενσωμάτωση αλλαγών της τάσης στο μοντέλο ορίζουμε ρητά σημεία αλλαγής (change points) στα οποία η ανάπτυξη μπορεί να αλλάξει. Θεωρούμε ότι υπάρχουν συνολικά S στο πλήθος σημεία αλλαγής ανάπτυξης στις χρονικές στιγμές $s_j, j = 1, ..., S$. Με δ_j ορίζουμε την αλλαγή στον ρυθμό ανάπτυξης την χρονική στιγμή s_j και ορίζουμε $\delta = (\delta_1, \delta_2, ..., \delta_s)$.

Έτσι ο δυναμικός ρυθμός ανάπτυξης μπορεί να οριστεί ως

$$k(t) = k + \sum_{j=1}^{S} \mathbf{1}_{\{t>s_j\}}\delta_j = k + a(t)'\delta$$

Όπου $a(t)' \in \{0,1\}^S$ και $a_j(t) = \mathbf{1}_{\{t\geq s_j\}}$. Με k συμβολίζουμε τον αρχικό σταθερό ρυθμό ανάπτυξης. Για να μπορεί η g(t) να είναι ομαλή συνάρτηση οι συνιστωσες του s-διαστατου διανύσματος γ(t) ορίζονται ως

$$\gamma_j = (s_j - m - \sum_{l<j}\gamma_l)(1 - \frac{k + \sum_{l<j}\gamma_l}{k + \sum_{l\leq j}\gamma_l})$$

Συνήθως, όπως προτάθηκε και από τους σχεδιαστές του μοντέλου ισχύει ότι $\delta_j \sim Laplace(0,\tau)$. Η παράμετρος τ ελέγχει την ευελιξία του μοντέλου στην αλλαγή του ρυθμού ανάπτυξης. Όταν το τ τείνει στο 0 τότε το μοντέλο δεν εντοπίζει τα σημεία αλλαγών και τότε k και γ τείνουν να γίνουν σταθερές (τότε γ=m). Έτσι ο όρος ανάπτυξης ορίζεται τελικά ως

$$g(t) = \frac{C(t)}{1 + exp(-(k + a(t)'\delta)(t - (m + a(t)'\gamma)))}$$

Η επέκταση αυτού του μοντέλου σε άλλες οικογένειες καμπυλών είναι απλή.

9.1.2 Γραμμικός όρος ανάπτυξης

Για προβλήματα προβλέψεων που δεν παρουσιάζουν κορεσμό της ανάπτυξης, ένα κομμάτι με γραμμικό ρυθμό ανάπτυξης παρέχει ένα λιτό και συχνά χρήσιμο μοντέλο.

$$g(t) = (k + a(t)'\delta)t + m + a(t)'\gamma, \gamma_j = -s_j\delta_j$$

9.2 Μοντελοποίηση εποχικότητας

Χρονοσειρές με δεδομένα επιχειρήσεων πολύ συχνά παρουσιάζουν πολυ-περιοδική εποχικότητα. Για παράδειγμα η πενθήμερη εργασία μπορεί να προκαλέσει πρότυπα που επαναλαμβάνονται κάθε εβδομάδα ενώ γεγονότα που συνδέονται με ημέρες διακοπών/αργιών (π.χ Χριστούγεννα, κλείσιμο σχολείων) παράγουν πρότυπα που επαναλαμβάνονται κάθε χρόνο.
Για την πρόβλεψη αυτών των χρονοσειρών χρειάζεται να προσδιορίσουμε μοντέλα εποχικότητας που είναι περιοδικές συναρτήσεις του χρόνου t. Η μέθοδος Prophet βασίζεται στις σειρές Fourier που δίνουν ευέλικτο μοντέλο για περιοδικά φαινόμενα. Επομένως μπορούμε να προσδιορίσουμε τις αυθαίρετα ομαλές εποχιακές επιδράσεις με την εξής συνάρτηση.

$$s(t) = \sum_{n=1}^{N} a_n \cos(\frac{2\pi nt}{P}) + b_n \sin(\frac{2\pi nt}{P})$$

Η προσαρμογή της $s(t)$ απαιτεί την εκτίμηση των 2N παραμέτρων $\beta = [a_1, b_1, ..., a_N, b_N]$. Συχνά γίνεται η υπόθεση ότι $\beta \sim N(0, \sigma^2)$

9.3 Μοντελοποίηση Διακοπών και Ειδικών γεγονότων

Η επιρροή μιας "ειδικής" μέρας (π.χ αργία, γιορτή) σε μια χρονοσειρά είναι συχνά ίδια από χρόνο σε χρόνο. Βέβαια αρκετές γιορτές είναι κινητές με αποτέλεσμα τέτοιου είδους επιρροές να μήν μπορουν να μοντελοποιηθούν μέσω της συνιστώσας εποχικότητας. Επίσης συχνά εμφανίζονται ως στιγμιαία άλματα (shocks) ή κατανεμημένες γύρω από ένα μικρό χρονικό διάστημα (λίγο πριν και λίγο μετά την συγκεκριμένη μέρα). Για κάθε "ειδική" μέρα i, D_i είναι το σύνολο των ιστορικών και μελλοντικών ημερομηνιών για αυτή την γιορτή και k_i είναι η αντίστοιχη αλλαγή που επιφέρει η μέρα αυτή στην πρόβλεψη. Υποθέτουμε ότι υπάρχουν L στο σύνολο ειδικές μέρες.

Εαν ορίσουμε $z(t) = [\mathbf{1}_{\{t\in D_1\}}, ..., \mathbf{1}_{\{t\in D_L\}}]$ τότε ο όρος h(t) στο μοντέλο Prophet ορίζεται ως

$$h(t) = z(t)\mathbf{k} \quad , \quad k \sim Normal(0, v^2)$$

Ένα σημαντικό πλεονέκτημα του αποσυνθετικού μοντέλου είναι ότι μας επιτρέπει να εξετάσουμε κάθε συνιστώσα της πρόβλεψης ξεχωριστά. Αυτό παρέχει ένα χρήσιμο εργαλείο για τους αναλυτές ώστε να αποκτήσουν εικόνα του προβλήματος πρόβλεψης, εκτός από την απλή παραγωγή μιας πρόβλεψη.
Οι αναλυτές που κάνουν προβλέψεις έχουν συχνά εκτεταμένες γνώσεις σχετικά με την ποσότητα που που προβλέπουν, αλλά περιορισμένες στατιστικές γνώσεις. Στο μοντέλο Prophet υπάρχουν πολλά σημεία όπου οι αναλυτές μπορούν να τροποποιήσουν το μοντέλο για να εφαρμόσουν την τεχνογνωσία τους, χωρίς να απαιτείται κατανόηση των υποκείμενων στατιστικών στοιχείων.
Η παράμετρος τ είναι ένα απλό κουμπί που ρυθμίζεται κατάλληλα για να αυξήσει ή να μειώσει την ευελιξία της τάσης, και το σ είναι ένα κουμπί για την αύξηση ή τη μείωση της ισχύος της συνιστώσας εποχικότητας. Με οπτικοποίηση των αποτελεσμάτων των προβλέψεων του μοντέλου στα ιστορικά δεδομένα ο αναλυτής μπορεί να ρυθμίσει κατάλληλα αυτές τις παραμέτρους μέχρι να δεί ότι πράγματι το μοντέλο αντιλαμβάνεται αλλαγές στην τάση και την εποχικότητα των ιστορικών δεδομένων.

9.4 Εφαρμογή του μοντέλου Prophet

Χρησιμοποιήθηκαν ωριαία δεδομένα κατανάλωσης ενέργειας από την εταιρία PJM από το 2002 εώς και το 2018. Για την προσαρμογή του μοντέλου χρησιμοποιήσαμε την γλώσσα προγραμματισμού Python. Αρχικά διαχωρίσαμε το σύνολο των δεδομένων σε σύνολο εκπαίδευσης και σύνολο ελέγχου και στην συνέχεια παρατηρήσαμε το χρονοδιάγραμμα των δεδομένων

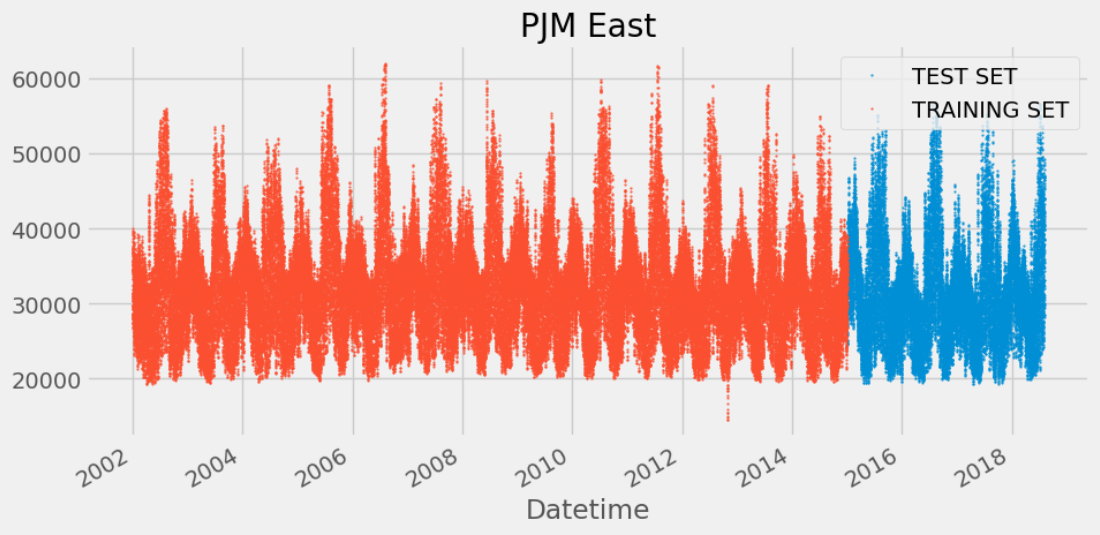
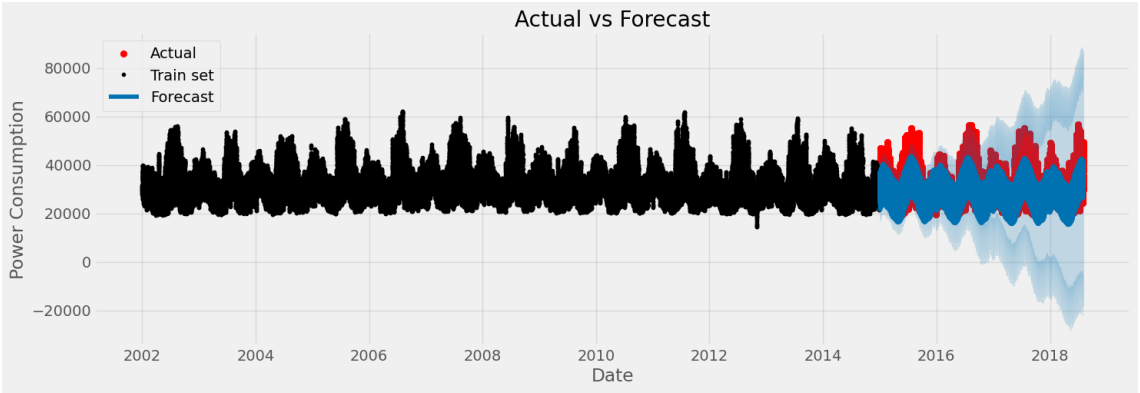


Figure 9.4.1: Ωριαία κατανάλωση ενέργειας για το διάστημα 2002-2018

Επειδή η κατανάλωση ενέργειας επηρεάζεται έντονα σε μέρες διακοπών και γενικότερα σε "ειδικές" μέρες, προσθέσαμε στο μοντέλο μας και την συνιστώσα των διακοπών, δίνοντας ως είσοδο στο μοντέλο τις ονομασίες τους καθώς και τις ημερομηνίες όπου αυτές συμβαίνουν. Οι προβλέψεις για τις τιμές της κατανάλωσης για το σύνολο ελέγχου όπως και οι πραγματικές τιμές φαίνονται στο πιό κάτω σχήμα.



Όπως αναφέραμε και πιο πάνω, ένα από τα πλεονεκτήματα του μοντέλου αυτού είναι ότι μπορεί να αντιμετωπίσει δεδομένα με πολυ-περιοδική εποχικότητα, η οποία είναι συχνά εμφανής σε ωριαία δεδομένα. Πιο κάτω παρουσιάζουμε τις προβλέψεις για κάθε συνιστώσα του μοντέλου ξεχωριστά.

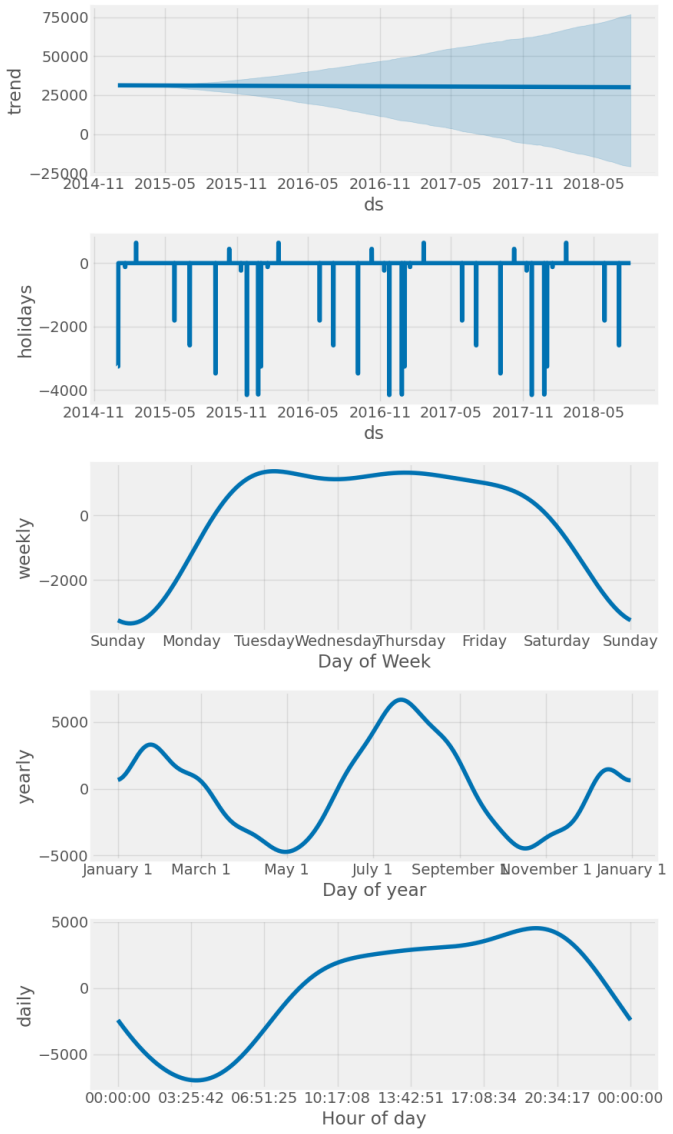


Figure 9.4.2: Οι προβλέψεις για κάθε συνιστώσα του μοντέλου Prophet

Το μοντέλο Prophet έχει το πλεονέκτημα ότι είναι πολύ πιο γρήγορο στην εκτίμηση από ότι τα μοντέλα που εξετάσαμε προηγουμένως και είναι εντελώς αυτοματοποιημένο. Ωστόσο, σπάνια δίνει καλύτερη ακρίβεια πρόβλεψης από τις εναλλακτικές προσεγγίσεις. Έγινε ιδιαίτερα γνωστό επειδή είναι πολυ εύκολο στην χρήση και παράγει καλές προβλέψεις με μικρό υπολογιστικό κόστος.

10 Μέθοδοι βελτίωσης προβλέψεων

Αρκετές φορές αφού επιλεχθεί ένα μοντέλο για πρόβλεψη, στην συνέχεια εξετάζουμε μεθόδους που μπορούν να βελτιώσουν τις προβλέψεις. Μια γνωστή μέθοδος είναι η "συνάθροιση δειγματοθετήσεων" (bootstrap aggregating). Μια άλλη διαδικασία που μπορεί να βελτιώσει τις προβλέψεις είναι ο συνδυασμός προβλέψεων από διαφορετικά μοντέλα, μέσω μιας συνάρτησης συνάθροισης των επιμέρους προβλέψεων. Στα ακόλουθα δύο υπο-κεφάλαια αναλύονται αυτές οι δύο μέθοδοι.

10.1 Δειγματοθέτηση και ενθυλάκωση (bootstrapping and bagging)

10.1.1 Δειγματοθέτηση Χρονοσειρών

Γενικά η διαδικασία δειγματοθέτησης (bootstrapping) στηρίζεται στην ιδέα δημιουργίας τυχαίων εκδόσεων των δεδομένων μας έτσι ώστε να αποφευχθεί η επανάληψη της πειραματικής διαδικασίας (συλλογή δεδομένων) η οποία αρκετές φορές είναι χρονοβόρα και κοστίζει. Είναι αρκετά χρήσιμη, ιδιαίτερα για τις 2 πιο κάτω περιπτώσεις

- Απόκτηση της κατανομής του σφάλματος (όταν στατιστικές υποθέσεις για την συμπεριφορά του δέν είναι λογικές. π.χ η υπόθεση κανονικότητας) για να μπορούμε να αξιολογήσουμε την βεβαιότητα της εκτίμησης.

- Για την παραγωγή της εκτίμησης " bagging" (bagging estimate) μέσω της συνάθροισης των αποτελεσμάτων πολλαπλών μοντέλων.

Η διαδικασία της δειγματοθέτησης στηρίζεται στην ακόλουθη διαδικασία.

(a) Έστω ένα αρχικό σύνολο δεδομένων $Z = Z_1, Z_2, ...Z_n$

(b) Εφαρμόζουμε τυχαία δειγματοληψία από το αρχικό σύνολο δεδομένων με επανάθεση μέχρι το δειγματοθετημένο σύνολο δεδομένων να έχει n στοιχεία.

(c) Επανάληψη του βήματος 2, B φορές. Τότε έχουμε B συνθετικά σύνολα που συμβολίζονται με $Z_b^*, b = 1, ..., B$. Αυτά ονομάζονται δειγματοθετημένα σύνολα δεδομένων (bootstrapped Dataset)

(d) Υπολογίζω την επιθυμητή (στατιστική) ποσότητα αξιοποιώντας με κατάλληλο τρόπο κάθε δειγματοθετημένο σύνολο δεδομένων. Η προκύπτουσα ποσότητα συμβολίζεται με $S(Z_b^*)$.

(e) Υπολογισμός bagging prediction η οποία ορίζεται από την σχέση $\sum_{b=1}^{B} S(Z_b^*)/B$. Επίσης η κατανομή της $S(Z_b^*)$, εαν κατασκευάσουμε το ιστόγραμμα των τιμών, θα μας δώσει την πιθανή κατανομή της S(Z)

Τα πιο πάνω βήματα αναπαριστώνται γραφικά με το επόμενο γράφημα.

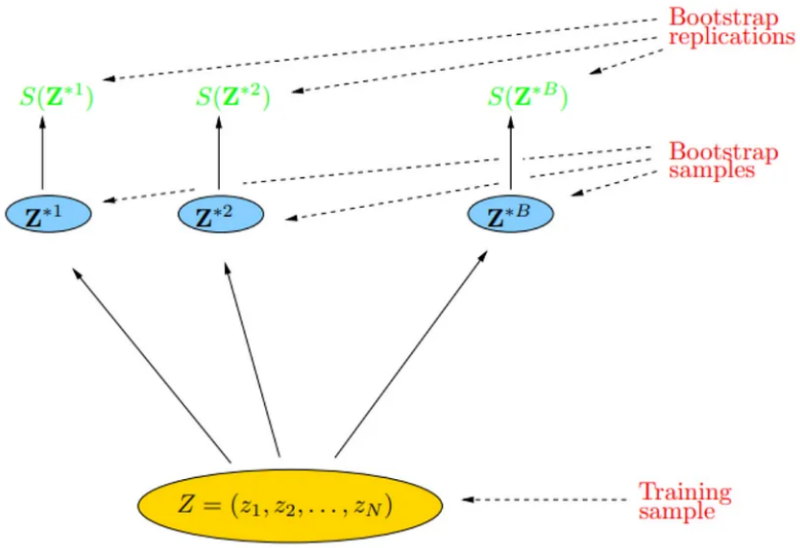


Figure 10.1.1: Κλασσική μέθοδος δειγματοθέτησης

Η εφαρμογή της μεθόδου δειγματοθέτησης σε δεδομένα χρονοσειρών χρειάζεται κάποιες μετατροπές για να μπορεί να λειτουργήσει αποτελεσματικά. Επειδή τα δεδομένα χρονοσειρών χαρακτηρίζονται από αυτοσυχέτιση εντός των θορύβων μεταξύ γειτονικών χρονικών στιγμών, μια τυχαία ανακατανομή των δεδομένων θα ανακατέψει αυτές τις αυτοσυχετίσεις. Για να αντιμετωπιστεί αυτό το πρόβλημα η δειγματοληψία γίνεται πάνω σε μπλοκ δεδομένων (ομάδα συνεχόμενων παρατηρήσεων).
Επομένως για την εφαρμογή της μεθόδου δειγματοθέτησης σε δεδομένα χρονοσειρών ακολουθούμε την ακόλουθη διαδικασία.

(a) Μετασχηματισμός της σειράς αν είναι απαραίτητο (π.χ Την μετατρέπουμε σε στάσιμη εαν δέν είναι).

(b) Ανάλυση της χρονοσειράς σε συνιστώσες Τάσης, Εποχικότητας, και υπόλοιπο χρησιμοποιώντας για παράδειγμα την μέθοδο STL.

(c) Συλλέγουμε την συνιστώσα υπόλοιπου του πιο πάνω βήματος και δημιουργούμε μπλοκ δεδομένων, όπου σε κάθε μπλοκ τοποθετούμε διαδοχικές τιμές υπολοίπων. Οι 3 πιό γνωστές μέθοδοι δημιουργίας των μπλοκ είναι οι moving block Boostrap (MBB), Circular block Boostrap (CBB), και Stationary Boostrap (SB). Εδώ θα χρησιμοποιήσουμε την MBB, σύμφωνα με την οποία καθορίζουμε αρχικά το μέγεθος των μπλοκ (μπλοκ ίσου μεγέθους) και στην συνέχεια μετακινούμε το κάθε μπλόκ κατα ένα βήμα κάθε φορά για να προκύψει το επόμενο μπλόκ όπως φαίνεται στο σχήμα (10.1.2). Εαν το μήκος της χρονοσειράς είναι n και το μήκος κάθε μπλόκ είναι l τότε δημιουργούνται με την MBB $n - l - 1$ μπλόκ.

(d) Το μήκος των μπλόκ καθορίζεται έτσι ώστε να αποτυπώνει την πιθανή αυτοσυχέτιση στα δεδομένα. π.χ σύμφωνα με την εργασία τους C.Bergmeir ([2]) για ετήσια και τριμηνιαία δεδομένα $l = 8$ ενώ για μηνιαία δεδομένα $l = 24$.

(e) Πραγματοποιούμε δειγματοληψία με επανάθεση $\left[\frac{n}{l}\right] + 2$ φορές ώστε η ακολουθία των μπλόκ να καλύπτει ολόκληρο το σύνολο δεδομένων.

(f) Προσθήκη της χρονοσειράς που προέκυψε στο βήμα (e) πίσω στην συνιστώσα της τάσης και εποχικότητας για να προκύψει μια προσομοιωμένη χρονοσειρά (simulated time series dataset).

(g) Επαναλαμβάνουμε τα βήματα (e) και (f) B φορές και έτσι προκύπτουν B συνθετικές χρονοσειρές οι οποίες λαμβάνουν υπόψη την αυτοσυχέτιση της χρονοσειράς.

Το πιό κάτω σχήμα αναπαριστά τα πιό πάνω βήματα γραφικά

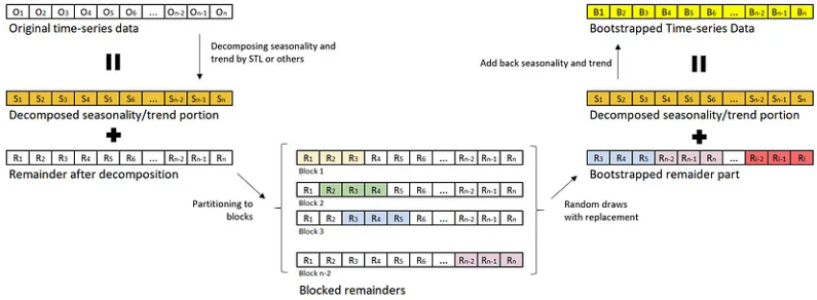


Figure 10.1.2: Διαδικασία δειγματοθέτησης σε δεδομένα χρονοσειρών για την αντιμετώπιση αυτοσυχετίσεων. Γίνεται χρήση τη μεθόδου MBB

10.1.2 Ενθυλακωμένες προβλέψεις (bagging predictions)

Όπως αναφέραμε και πιο πάνω οι δειγματοθετημένες χρονοσειρές μπορούν να χρησιμοποιηθούν για την παραγωγή της ενθυλακωμένης πρόβλεψης (bagging predictions). Για να γίνει αυτό παράγουμε προβλέψεις από καθεμιά από τις B δειγματοθετημένες χρονοσειρές που προέκυψαν μέσω της διαδικασίας που περιγράψαμε παραπάνω, και στην συνέχεια αξιοποιώντας μία συνάρτηση συνάθροισης προκύπτει η ενθυλακωμένη πρόβλεψη. Συνήθως η συνάρτηση συνάθροισης επιλέγεται ως ο μέσος όρος των τιμών των επιμέρους προβλέψεων. Παρόλο που κατα καιρούς προτάθηκαν άλλες πιό σύνθετες συναρτήσεις συνάθροισης, η απόδοσή τους δέν είναι ιδιαίτερα καλύτερη από αυτή της συνάρτησης μέσο όρου. Με αυτή την διαδικασία σε πολλές περιπτώσεις προκύπτουν καλύτερες προβλέψεις από ό'τι αν απλώς προβλέψουμε απευθείας τις αρχικές χρονοσειρές. Αυτό ονομάζεται "ενθυλάκωση (bagging)" που προκύπτει από τη φράση "bootstrap aggregating" (Συνάθροιση Δειγματοθετήσεων).

10.2 Συνδυαστική πρόβλεψη

Ένας απλός αλλά αποτελεσματικός τρόπος βελτίωσης ακρίβειας των προβλέψεων, είναι η πρόβλεψη με πολλές διαφορετικές μεθόδους στις ίδιες χρονοσειρές και ο συνδυασμός τους μέσω κάποιας συνάρτησης, με συνηθέστερη προσέγγιση τον υπολογισμό της μέσης τιμής των προβλέψεων που προκύπτουν. Ενώ υπήρξε σημαντική έρευνα σχετικά με τη χρήση σταθμισμένων μέσων ή κάποιας άλλης πιο σύνθετης προσέγγισης συνδυασμού, η χρήση ενός απλού μέσου όρου έχει αποδειχθεί ότι είναι δύσκολο να ξεπεραστεί.

10.3 Διαχείριση ειδικών χαρακτηριστικών χρονοσειρών για καλύτερες προβλέψεις

10.3.1 Εβδομαδιαία, ημερήσια, υπο-ημερήσια δεδομένα

Δεδομένα αυτής της συχνότητας αποτελούν πρόκληση ως προς την πρόβλεψη μελλοντικών τιμών. Το σημαντικό πρόβλημα όταν αντιμετωπίζουμε χρονοσειρά με εβδομαδιαία δεδομένα είναι ότι η εποχιακή περίοδος (πλήθος εβδομάδων σε ένα έτος) είναι μεγάλη και δεν είναι ακέραια, καθώς ο μέσος αριθμός εβδομάδων σε ένα έτος είναι 52.18 $((3\times\frac{365}{7}+\frac{366}{7})/4)$. Οι περισσότερες εποχιακές μέθοδοι που αναλύσαμε απαιτούν οι εποχικές περίοδοι να είναι ακέραιοι και δεν ανταποκρίνονται τόσο καλά σε μεγάλες εποχιακές περιόδους.

Η απλούστερη προσέγγιση για αντιμετώπιση αυτού του προβλήματος είναι η χρήση μεθόδου ανάλυσης χρονοσειράς σε συνιστώσες (π.χ STL) και εφαρμογή μη-εποχιακής μεθόδου στα εποχιακά προσαρμοσμένα δεδομένα. Μια εναλλακτική προσέγγιση είναι να χρησιμοποιήσουμε **μοντέλο δυναμικής αρμονικής Παλινδρόμησης**. Αυτή η προσέγγιση αποδίδει καλύτερα από την 1^{η} όταν υπάρχουν συμμεταβλητές που αποτελούν χρήσιμους παράγοντες πρόβλεψης καθώς αυτοί μπορούν να προστεθούν ως επιπρόσθετοι παλινδρομητές.

Τα ημερήσια και υπο-ημερήσια δεδομένα περιλαμβάνουν πολλαπλά εποχιακά μοτίβα και για αυτό το λόγο απαιτείται η χρήση μεθόδων που να χειρίζονται τόσο περίπλοκες εποχικότητες. Βέβαια όταν η χρονοσειρά έχει μικρό μέγεθος τότε πιθανό να μην φτάνει να εμφανίσει περισσότερες από μία εποχιακή περίοδο με αποτέλεσμα να μην περιπλέκεται το πρόβλημα πρόβλεψης. Όταν οι χρονοσειρές όμως είναι μεγάλες ώστε να εμφανιστούν εποχιακά πρότυπα πολλών περιόδων (και εποχιακές περίοδοι μεγάλης τάξης) τότε είναι απαραίτητο να υιοθετηθεί μια εκ των δύο προσεγγίσεων που αναφέραμε πιο πάνω ή η χρήση του μοντέλου Prophet. Η χρήση του μοντέλου Prophet έχει το θετικό ότι, μπορεί να ρυθμιστεί κατάλληλα για να αντιμετωπίσει κινητές γιορτές αλλά αυτό μπορεί να γίνει και στα μοντέλα ARIMA με χρήση εικονικών μεταβλητών.

11 Εφαρμογή - Ζήτηση ηλεκτρικής ενέργειας

11.1 Περιγραφή προβλήματος

Η ΔΕΗ θέλει να προβλέψει την ημίωρη ζήτηση ηλεκτρικής ενέργειας για τον επόμενο μήνα (Ιούλιος 2022). Για το σκοπό αυτό συλλέχθηκαν τα δεδομένα των προηγούμενων 12 μηνών. Μετά από ανάλυση των δεδομένων παρατηρήθηκε ότι τα δεδομένα της περσινής χρονιάς δεν είναι τόσο αντιπροσωπευτικά καθώς οι τιμές της ζήτησης επηρεάστηκαν κατα πολύ από την πανδημία (Covid-19) και η χρησιμοποίηση των προτύπων που αυτά προτείνουν θα επιφέρει λανθασμένα αποτελέσματα. Για αυτό το λόγο η μελέτη στηρίχθηκε στους 2 προηγούμενους μήνες όπως επίσης σε παράγοντες που επηρεάζουν την ημίωρη ζήτηση ηλεκτρικής ενέργειας. Η χρονοσειρά των δεδομένων, η οποία φαίνεται πιο κάτω, διαχωρίστηκε σε 2 σύνολα, το σύνολο εκπαίδευσης και σύνολο αξιολόγησης.

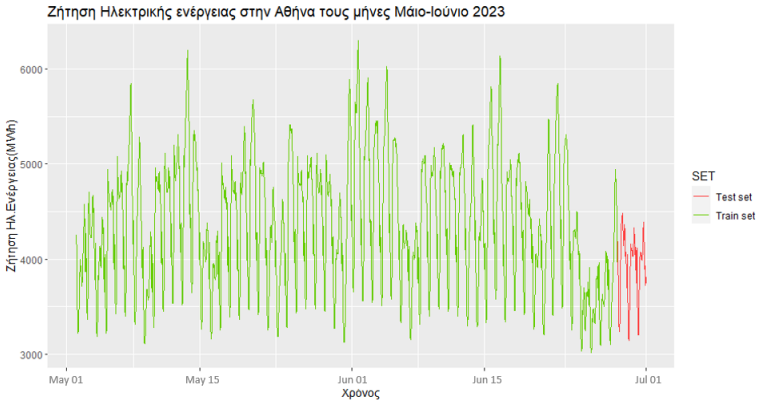
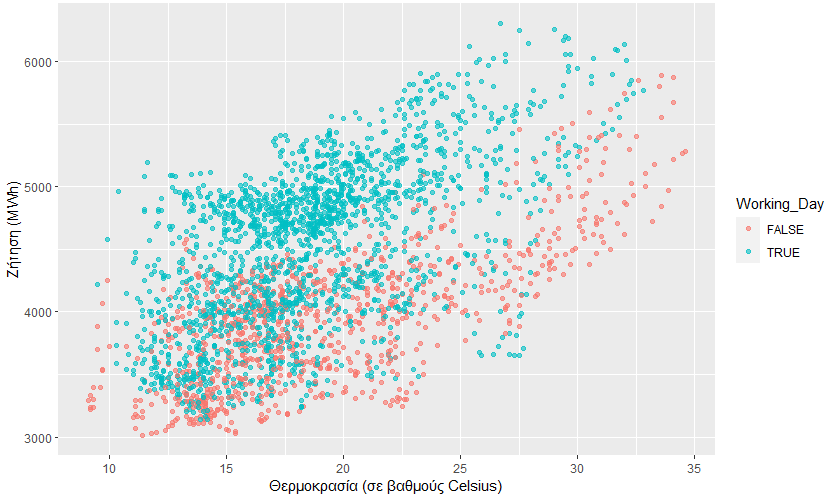


Figure 11.1.1: Ζήτηση ηλεκτρικής ενέργειας στην Αθήνα

11.1.1 Διαδικασία Μοντελοποίησης προβλήματος

Για την κατασκευή ενός αξιόπιστου μοντέλου πρόβλεψης κρίθηκε απαραίτητο να συμπεριληφθούν στην μοντελοποίηση και παράγοντες πρόβλεψης. Δύο σημαντικοί παράγοντες πρόβλεψης είναι η θερμοκρασία που επικρατεί όπως επίσης κατα πόσο η μέρα είναι καθημερινή ή αργία (συμπεριλαμβανομένου του Σαββατοκύριακου). Η εξάρτηση της ζήτησης από αυτούς τους 2 παράγοντες παρουσιάζεται στο πιο κάτω διάγραμμα σκέδασης.



Επιπλέον παρουσιάζουμε στο ίδιο γράφημα τη χρονοσειρά των δεδομένων εκπαίδευσης και των αντίστοιχων θερμοκρασιών έτσι ώστε να γίνει πιο εμφανές η μεταξύ τους συσχέτιση.

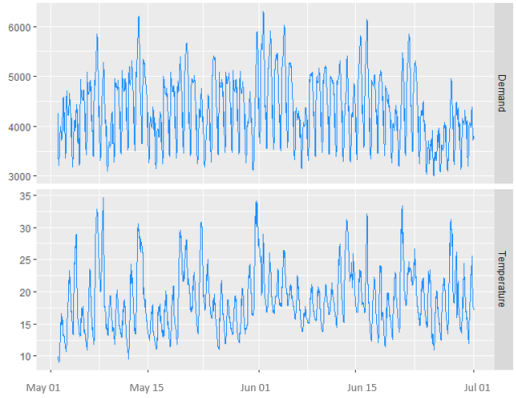


Figure 11.1.2: Διαγράμματα χρονοσειρών ζήτησης και θερμοκρασίας για το χρονικό διάστημα στο οποίο θα εκπαιδευτούν τα μοντέλα.

Από τις πιο πάνω γραφικές είναι εμφανές ότι οι δύο αυτοί παράγοντες πρόβλεψης μπορούν να επεξηγήσουν αρκετά καλά τη συμπεριφορά και τα πρότυπα της ζήτησης. Πιο συγκεκριμένα φαίνεται ότι σε υψηλότερες θερμοκρασίες η ζήτηση είναι αυξημένη ενώ σε αργίες παρατηρείται μείωση στην ζήτηση σε σχέση με τις εργάσιμες μέρες.

Βέβαια, η θερμοκρασία για μελλοντικές στιγμές δεν είναι ντετερμινιστική ποσότητα, και για να αξιοποιηθεί χρειάζεται να προβλεφθεί και αυτή. Το γεγονός αυτό προσθέτει αβεβαιότητα στις τελικές προβλέψεις. Για να μπορέσουμε λοιπόν να εκμεταλλευτούμε με ορθό τρόπο την συσχέτιση της ζήτησης με την θερμοκρασία, το μοντέλο που θα κατασκευάσουμε θα προβλέπει με χρονικό ορίζοντα τριών ημερών και θα ανανεώνεται καθημερινά για να προβλέπει τις επόμενες μέρες. Η γνώση της ζήτησης των επόμενων τριών ημερών είναι αρκετή για τον προγραμματισμό των διάφορων εργασιών που επηρεάζονται από την ζήτηση, επομένως ο περιορισμός του χρονικού ορίζοντα πρόβλεψης που θέσαμε δεν επηρεάζει τον τελικό στόχο μας. Βέβαια, για κάθε μέρα θα προκύπτουν 3 προβλέψεις, μια με χρονικό ορίζοντα 3, μια με χρονικό ορίζοντα 2 και

μία με χρονικό ορίζοντα 1. Προφανώς οι προβλέψεις με χρονικό ορίζοντα ένα και δύο θα είναι αρκετά πιο καλές όμως δεν θα απέχουν από την αρχική πρόβλεψη τριών βημάτων η οποία προσφέρει χρόνο για την κατάλληλη προετοιμασία ενώ οι άλλες δύο είναι χρήσιμες για τυχόν διορθώσεις και αντιμετώπισης ακραίων φαινομένων που μπορεί να προκύψουν.

Για την επιλογή του καταλληλότερου μοντέλου μελετήσαμε την απόδοση μερικών ικανών μοντέλων που μελετήθηκαν στα προηγούμενα κεφάλαια. Τα μοντέλα πού εξετάστηκαν είναι:

- SARIMA
- SARIMAX, με συντελεστές πρόβλεψης τη θερμοκρασία και τη δείκτρια μεταβλητή αργίας
- ARIMAX, με συντελεστές πρόβλεψης τη θερμοκρασία και τη δείκτρια μεταβλητή αργίας. Για τον εντοπισμό των εποχιακών προτύπων γίνεται χρήση όρων Fourier.
- Πρόβλεψη με ανάλυση STL (STLF).
- Πρόβλεψη με διακριτό μετασχηματισμό κυματιδίων (WTF).

Στην συνέχεια, σε αρκετά σημεία αντί να αναφρόμαστε σε κάθε χρονική στιγμή ως ημι-ώρα θα την καλούμε απλά ώρα.

11.2 Μοντέλο 1: SARIMA

Ξεκινήσαμε την μελέτη μας με την εξέταση της απόδοσης εποχιακών μοντέλων ARIMA τα οποία αναλύσαμε στο κεφάλαιο 7.5. Τα δεδομένα εκπαίδευσης είναι προφανές (από την γραφική 11.1.2) μη στάσιμα, επομένως θα πάρουμε πρώτα μια εποχιακή διαφορά. Τα εποχιακά διαφοροποιημένα δεδομένα μαζί με τα αντίστοιχα διαγράμματα ACF και PACF φαίνονται στο πιο κάτω σχήμα.

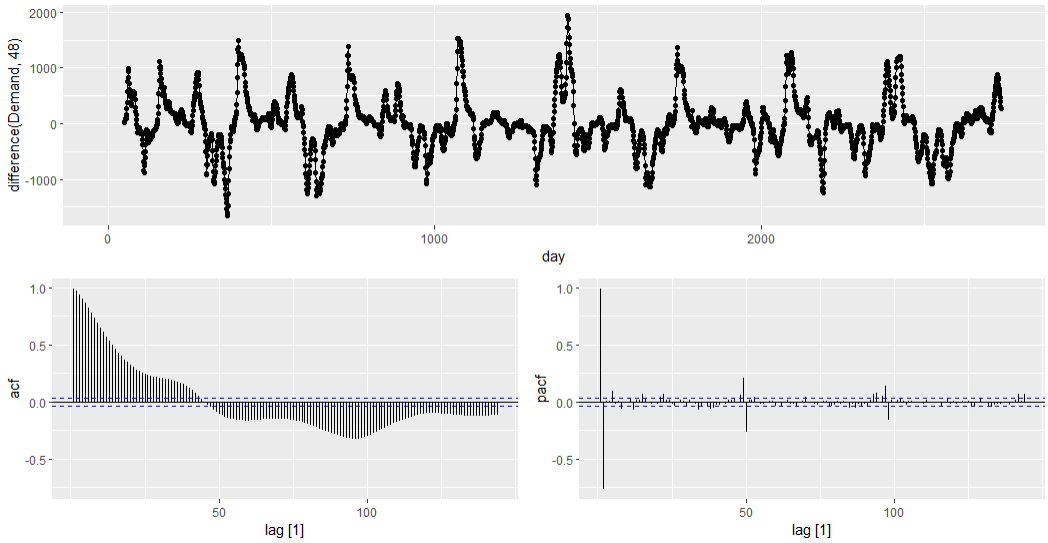


Figure 11.2.1: Εποχιακά διαφοροποιημένη ζήτηση ηλεκτρικής ενέργειας

Παρατηρώντας το πιο πάνω γράφημα φαίνεται να είναι και αυτή η χρονοσειρά μη στάσιμη, οπότε παίρνουμε ακόμα μια διαφορά πρώτης τάξης.

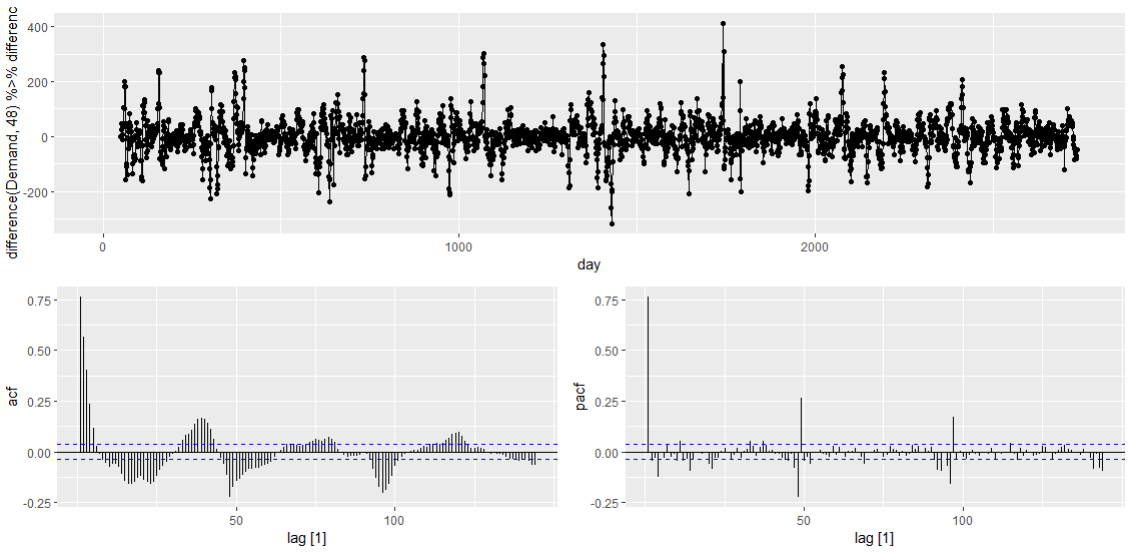
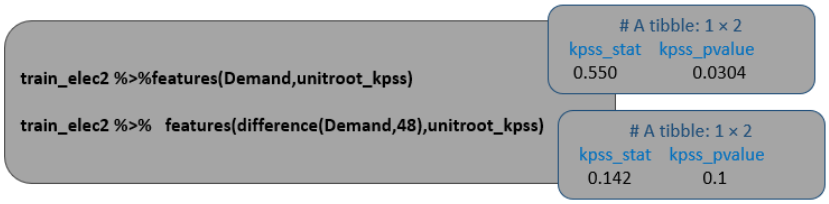


Figure 11.2.2: Διπλά διαφοροποιημένη ζήτηση ηλεκτρικής ενέργειας

Η χρονοσειρά πλέον είναι ξεκάθαρα στάσιμη. Τον γραφικό έλεγχο στασιμότητας μπορεί να συμπληρώσει ο στατιστικός έλεγχος KPSS. Σύμφωνα με αυτόν, η εποχιακή διαφοροποίηση είναι απαραίτητη αλλά δεν ισχύει το ίδιο για τη διαφοροποίηση πρώτης τάξης. Προτιμάμε να κρατήσουμε και τη διαφοροποίηση πρώτης τάξης καθώς με βάση τις γραφικές φαίνεται να είναι απαραίτητη.



Στόχος μας τώρα είναι να βρούμε ένα κατάλληλο μοντέλο ARIMA με βάση τα ACF και PACF που φαίνονται στο Σχήμα (11.2.2). Παρατηρούμε ότι οι 3 πρώτες υστερήσεις είναι πολύ σημαντικές στο ACF, όμως μόνο η πρώτη είναι σημαντική στο PACF. Επομένως αυτό υποδεικνύει ένα μη εποχιακό όρο MA(1). Επίσης η πολύ σημαντική πρώτη υστέρηση του διαγράμματος PACF, σε συνδυασμό με την ημιτονοειδές μείωση των τιμών στο ACF υποδεικνύουν τον μή εποχιακό όρο AR(1). Επιπλέον παρατηρούμε ότι οι εποχιακές υστερήσεις του PACF (48,96,..) φθίνουν εκθετικά, και στο διάγραμμα ACF η τελευταία έντονη εποχιακή υστέρηση είναι η δεύτερη. Επομένως προσθέτουμε ένα εποχιακό όρο MA(2). Με αντίστοιχο επιχείρημα μπορούμε να καταλήξουμε ότι μπορουμε να συμπεριλάβουμε και εποχιακό όρο AR(2) αλλά αυτό δέν το υιοθετούμε στο τέλος καθώς αυξάνεται κατα πολύ η υπολογιστική πολυπλοκότητα χωρίς ιδιαίτερη βελτίωση της απόδοσης στο τελικό μοντέλο. Επομένως το τελικό εποχιακό μοντέλο που επιλέγουμε είναι το $\mathbf{ARIMA}(\mathbf{1},\mathbf{1},\mathbf{1})(\mathbf{0},\mathbf{1},\mathbf{2})_{[\mathbf{48}]}$, το οποίο εκπαιδεύσαμε στο σύνολο εκπαιδευσης και προέκυψε τον μοντέλο:

Μοντέλο 1

$$\mathbf{ARIMA(1,1,1)(0,1,2)_{[48]}}$$

$$(1-\varphi_1 B)(1-B)(1-B^{48})y_t = (1+\theta_1 B)(1+\Theta_1 B^{48}+\theta_2 B^{96})\varepsilon_t$$
$$(1-0.7379B)(1-B)(1-B^{48})y_t = (1+0.1063B)(1-0.6797B^{48}-0.1813B^{96})$$

Όλοι οι συντελεστές του μοντέλου είναι στατιστικά σημαντικοί, για αυτό δεν αφαιρούμε κάποιον. Αυτό προκύπτει απο το γεγονός ότι το διάστημα εμπιστοσύνης για τον κάθε συντελεστή δεν περιέχει την μηδενική τιμη.

Αξίζει να σημειωθεί ότι μπορούσαμε να χρησιμοποιούσαμε την αυτόματη επιλογή των παραμέτρων μέσω της συνάρτησης ARIMA() του πακέτου fable στην R. Αυτή η συνάρτηση χρησιμοποιεί ένα αλγόριθμο επιλογής παραμέτρων, ο οποίος αρχικά προσδιορίζει το πλήθος εποχιακών διαφορών (D) που χρειάζονται, με την αξιοποίηση του δείκτη ισχύος εποχικότητας F_s (κεφάλαιο 3.4.). Πιο συγκεκριμένα εάν $F_s \geq 0.64$ τότε εφαρμόζεται εποχιακή διαφοροποίησης. Με επαναληπτική εφαρμογή του στατιστικού ελέγχου KPSS προσδιορίζει την παράμετρο d. Στην συνέχεια αρχίζοντας με ένα αρχικό συνδυασμό τιμών για τις παραμέτρους p,q,P,Q, μεταβαίνει σε άλλους πιθανούς συνδυασμούς που μειώνουν την τιμή του AICc μέχρι να εξαντλήσει τις πιθανές επιλογές, χωρίς να αυξάνει όμως ιδιαίτερα την πολυπλοκότητα του μοντέλου για να αποφευχθεί το φαινόμενο της υπερ-εκπαίδευσης. Με αυτό λοιπόν τον αλγόριθμο, θα προέκυπτε ένα μοντέλο $ARIMA(3,0,1)(2,1,0)_{48}$. Επειδή ο αλγόριθμος χρησιμοποιεί την επαναληπτική εφαρμογή του στατιστικού ελέγχου KPSS για τον καθορισμού των διαφορών πρώτης τάξης, δεν εφαρμόζει διαφορά πρώτης τάξης καθώς όπως αναφέρθηκε πιο πάνω ο στατιστικός έλεγχος KPSS προτείνει μόνο την εποχιακή διαφοροποίηση. Τόσο η τιμή του AIC όσο και η απόδοση στο σύνολο ελέγχου των 2 αυτών μοντέλων είναι σχεδόν ίδια, επομένως προτιμάμε το μοντέλο $ARIMA(1,1,1)(0,1,2)_{48}$ που είναι πιο απλό.

Προτού χρησιμοποιήσουμε αυτό το μοντέλο για προβλέψεις πραγματοποιούμε τους διαγνωστικούς ελέγχους για τα υπόλοιπα. Αρχικά εξετάζουμε κατα πόσο τα καινοτόμα υπόλοιπα είναι ασυσχέτιστα. Αυτό το εξετάζουμε τόσο γραφικά αλλά και με τον στατιστικό έλεγχο Ljung-Box.

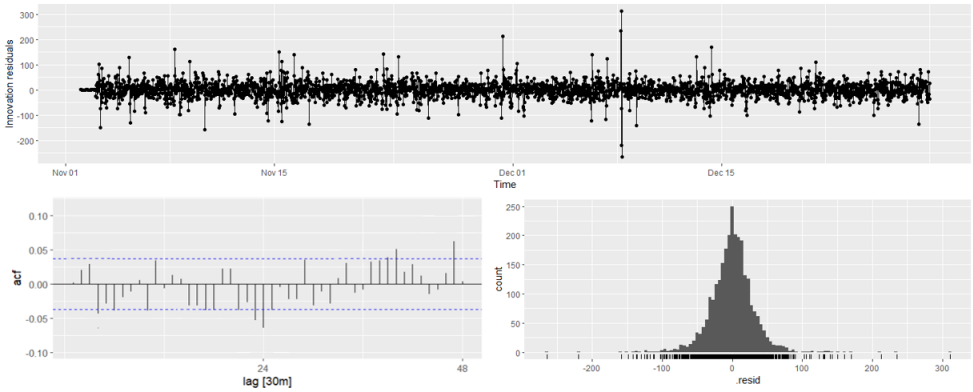


Figure 11.2.3: Καινοτόμα υπόλοιπα (δηλαδή, τα εκτιμώμενα σφάλματα ARIMA) του μοντέλου 1.

Παρατηρούμε ότι υπάρχουν κάποιες αυτοσυχετίσεις που είναι στατιστικά σημαντικές και κάποιες που οριακά δεν είναι. Οι στατιστικά σημαντικές αυτοσυχετίσεις σχετίζονται κυρίως με περίοδο 24 και 48 χρονικών στιγμών. Αυτό έχει φυσικό νόημα με βάση τα δεδομένα μας. Ο στατιστικός έλεγχος Ljung-Box δίνει p-value λίγο πιο κάτω από 0.05, και επομένως φαίνεται να υπάρχουν ενδείξεις για την απόρριψη της υπόθεσης μή αυτοσυχέτισης υπολοίπων, γεγονός που επιβεβαιώνει την εικόνα του ACF γραφήματος. Για να εξαλειφθούν αυτές οι αυτοσυχετίσεις απαιτείται ένα πιο περίπλοκο μοντέλο το οποίο δέν μελετάμε στην παρούσα εργασία. Πιο συγκεκριμένα μια καλή προσέγγιση είναι να αντιμετωπιστεί η κάθε μισή ώρα ως ξεχωριστή χρονοσειρά. Έτσι με αυτό τον τρόπο μπορούν να μοντελοποιηθούν περισσότερες πληροφορίες της χρονοσειράς με αποτέλεσμα τα υπόλοιπα του κάθε μοντέλου να είναι ασυσχέτιστα ([7]). Η μέση τιμή των υπολοίπων φαίνεται να είναι ίση με μηδέν και η υπόθεση κανονικότητας δεν φαίνεται παράλογη. Η συμπεριφορα των υπολοίπων είναι σχεδόν πανομοιότυπη σε όλα τα μοντέλα που θα ακολουθήσουν και την θεωρούμε ικανοποιητική καθώς, ο κύριος σκοπός της εφαρμογής είναι η ανάδειξη μεθόδων πρόβλεψης που χρησιμοποιούν τεχνικές αποσύνθεσης της χρονοσειράς όπως η STL και ο μετασχηματισμός κυματιδίων.

Στην συνέχεια παρουσιάζουμε, γραφικά, την απόδοση του μοντέλου αυτού καθώς και την απόδοση του (στο σύνολο αξιολόγησης) ως προς κάποια στατιστικά μέτρα απόδοσης σημειακών προβλέψεων που εισάγαμε στο κεφάλαιο 5.6.

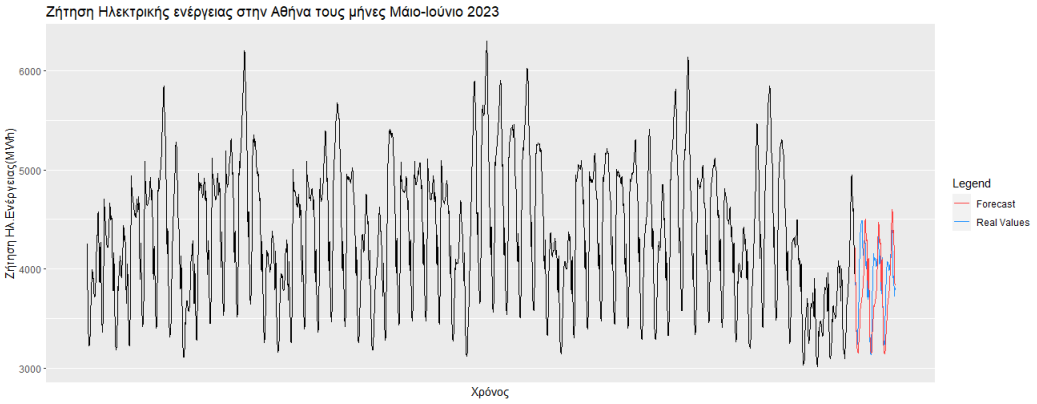


Figure 11.2.4: Προβλέψεις ημίωρης ζήτησης ηλεκτρικής ενέργειας για τις επόμενες 3 μέρες με το Μοντέλο 1

Χαρακτηριστικά απόδοσης Μοντέλου 1	
MAPE	6.59
RMSE	303
MAE	263

11.3 Μοντέλο 2: SARIMAX

Το μοντέλο 1 εκμεταλλεύεται τις πληροφορίες από προηγούμενες παρατηρήσεις και δεν αξιοποιεί καθόλου τις πληροφορίες που υπάρχουν σε παράγοντες πρόβλεψης όπως για παράδειγμα η θερμοκρασία και κατα πόσο είναι αργία ή όχι. Στόχος μας τώρα είναι η κατασκευή μοντέλου με καλύτερη απόδοση από το μοντέλο 1. Για αυτό το λόγο θα συμπεριλάβουμε και παράγοντες πρόβλεψης στο μοντέλο SARIMA, δηλαδή θα προσαρμόσουμε ένα μοντέλο SARIMAX. Για να το κάνουμε αυτό, αρχικά ορίζουμε τις μεταβλητές για τους παράγοντες πρόβλεψης.

$$T_t = \text{θερμοκρασία την χρονική στιγμή t}$$

$$\mathbf{1(WD)_t} = \begin{cases} 1 & \text{εαν η χρονική στιγμή t εμπίπτει σε μέρα αργίας} \\ 0 & \Delta\text{ιαφορετικά} \end{cases}$$

Επειδή μέρος της εποχικότητας και των προτύπων της χρονοσειράς θα επεξηγηθούν μέσω των παραγόντων πρόβλεψης, η επιλογή των παραμέτρων του εποχιακού ARIMA, στηρίχθηκε στην ελαχιστοποίηση του AIC (γίνεται με βάση το σύνολο εκπαίδευσης) σε συνδυασμό με την μεγιστοποίηση της απόδοσης του μοντέλου στο σύνολο αξιολόγησης (ελαχιστοποίηση κάποιου μέτρου σφάλματος). Πιο συγκεκριμένα οι παράμετροι επιλέχθηκαν με βάση την ελαχιστοποίηση του μέτρου MAPE.

Από αυτή την διαδικασία προέκυψε το εξής μοντέλο.

Μοντέλο 2

$$y_t = 3.2289T - 4.8425\mathbf{1(WD)_t} + \eta_t, \eta_t \sim \text{ARIMA}(1,0,3)(2,1,0)_{[48]}$$

$$(1 - 0.9787B)(1 + 0.4325B^{48} + 0.2677B^{96})(1 - B^{48})\eta_t = (1 + 0.8480B + 0.5504B^2 + 0.3275B^3)\varepsilon_t$$

Όλοι οι συντελεστές του πιο πάνω μοντέλου είναι στατιστικά σημαντικοί εκτός απο τον συντελεστής της επεξηγηματικής μεταβλητής $\mathbf{1(WD)_t}$. Βέβαια το διάστημα εμπιστοσύνης οριακά περιέχει το μηδέν μέσα επομένως επιλέγουμε να διατηρήσουμε το παράγοντα πρόβλεψης αυτό αφού όταν αφαιρεθεί η απόδοση στο σύνολο αξιολόγησης μειώνεται (ελάχιστα).

Η απόδοση του μοντέλου αυτού στο σύνολο αξιολόγησης συνοψίζεται στην πιο κάτω γραφική και πίνακα.

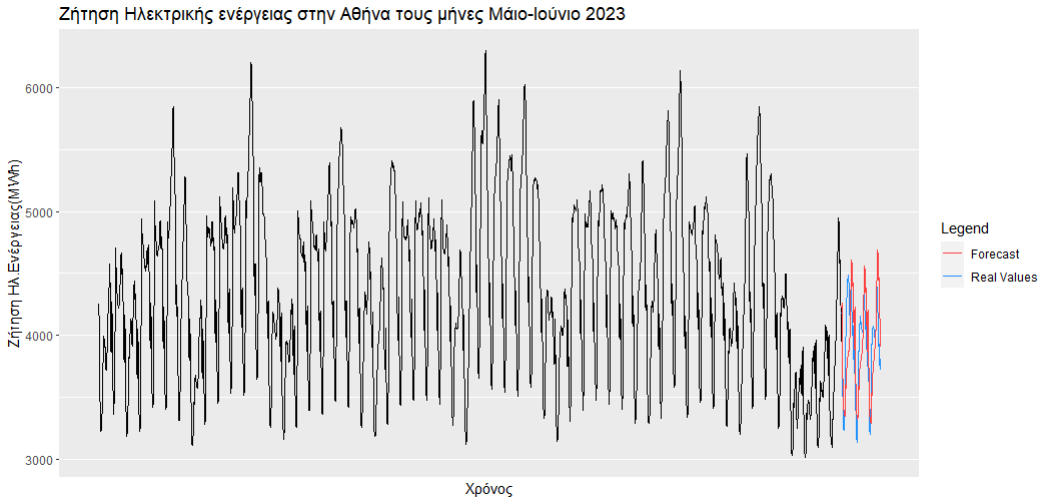


Figure 11.3.1: Προβλέψεις ημίωρης ζήτησης ηλεκτρικής ενέργειας για τις επόμενες 3 μέρες με το Μοντέλο 2

Χαρακτηριστικά απόδοσης Μοντέλου 1	
MAPE	6.19
RMSE	296
MAE	247

11.4 Μοντέλο 3: ARIMAX με χρήση όρων Fourier για την διαχείριση της εποχικότητας

Το αρνητικό του μοντέλου 2 είναι το γεγονός ότι, τα εποχιακά μοντέλα ARIMA δεν μπορούν να διαχειριστούν εύκολα εποχικότητες μεγάλης περιόδου και για αυτό τον λόγο το μοντέλο 2 είναι υπολογιστικά κοστοβόρο. Επιπλέον μπορεί να μοντελοποιηθεί επαρκώς μόνο η εποχιακή περίοδος m=48 και αγνοείτε έτσι η εβδομαδιαία εποχικότητα. Βέβαια η δείκτρια μεταβλητή $\mathbf{1}(WD)$ μπορεί να αποδώσει μέρος της εβδομαδιαίας εποχικότητας. Για αυτούς τους λόγους, προσαρμόστηκε ένα μοντέλο δυναμικής παλινδρόμησης με σφάλματα που ακολουθούν μη εποχιακό μοντέλο ARIMA και χρήση όρων Fourier εποχιακής περιόδου 48 και 336 για την αντιμετώπιση της ημερήσιας και εβδομαδιαίας εποχικότητας. Το μοντέλο που προσαρμόστηκε είναι το εξής

Μοντέλο 3

$$y_t = \beta_1 T + \beta_2 \mathbf{1}(\mathbf{WD}) + \sum_{k=1}^{K} [a1_k s_k(t) + \gamma 1_k c_k(t)]_{m=48} + \sum_{k=1}^{K} [a2_k s_k(t) + \gamma 2_k c_k(t)]_{m=336} + \eta_t$$

$\eta_t \sim$ ARIMA(p,d,q)-Μη εποχικό

Η απόδοση του μοντέλου αυτού στο σύνολο αξιολόγησης συνοψίζεται στην πιο κάτω γραφική και πίνακα.

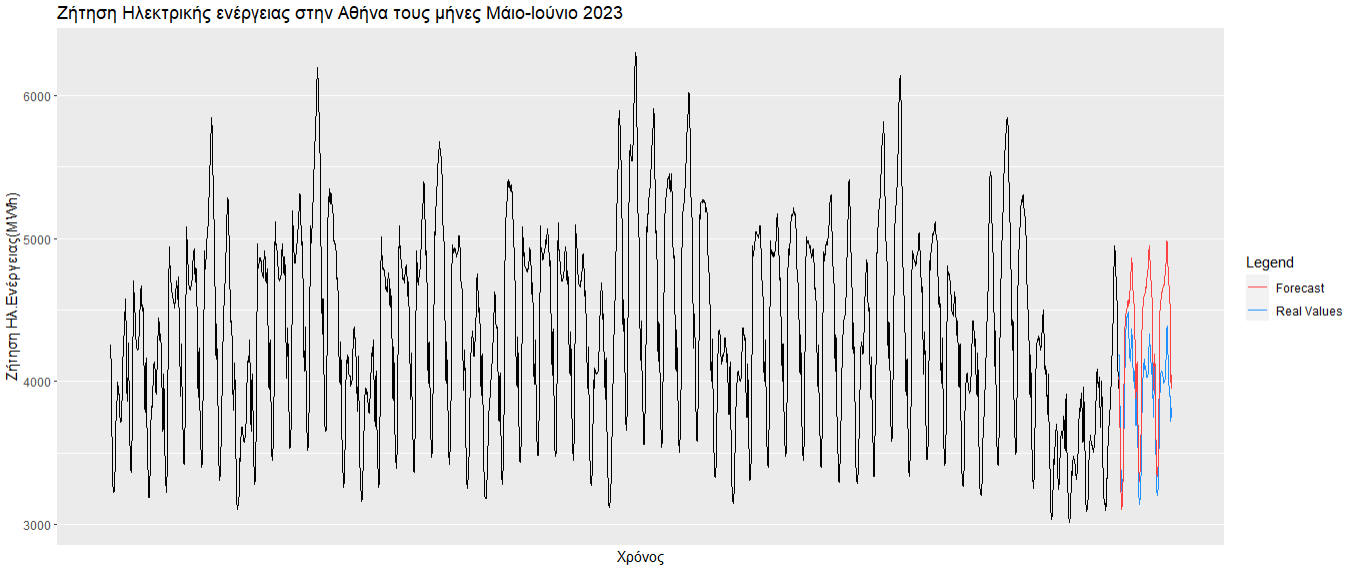


Figure 11.4.1: Προβλέψεις ημίωρης ζήτησης ηλεκτρικής ενέργειας για τις επόμενες 3 μέρες με το Μοντέλο 3

Χαρακτηριστικά απόδοσης Μοντέλου 1	
MAPE	9.88
RMSE	450
MAE	393

Παρατηρούμε ότι το μοντέλο αυτό δεν αποδίδει τόσο καλά όσο τα μοντέλα 1 και 2. Αυτό οφείλεται στο γεγονός ότι οι όροι Fourier δεν μπορούν να αποδώσουν τόσο καλά την εποχικότητα σε δεδομένα με έντονη πληροφορία. Στην συνέχεια θα δούμε ότι σε απλούστερες χρονοσειρές το μοντέλο 3 θα είναι ιδιαίτερα χρήσιμο τόσο για την υπολογιστική του απλότητα όσο και για την απόδοση του.

11.5 Μοντέλο 4: Πρόβλεψη με ανάλυση STL (STLF)

Με αυτό το μοντέλο θα στηριχτούμε στην ανάλυση μιας χρονοσειράς σε συνιστώσες. Αρχικά γίνεται ανάλυση της χρονοσειράς σε συνιστώσες με την μέθοδο STL και στην συνέχεια σε κάθε συνιστώσα εφαρμόζεται ένα κατάλληλο μοντέλο πρόβλεψης. Για την ανάλυση σε συνιστώσες χρησιμοποιήθηκε η προσθετική ανάλυση και η αποσυντεθειμένη χρονοσειρά μπορεί να γραφτεί ως.

$$y_t = \hat{S}_t + \hat{A}_t$$

Όπου $\hat{A}_t = \hat{T}_t + \hat{R}_t$ είναι η εποχιακά προσαρμοσμένη χρονοσειρά. Για την παραγωγή προβλέψεων από την αποσυντεθειμένη χρονοσειρά, κάνουμε προβλέψεις για την εποχιακή συνιστώσα και την εποχιακά προσαρμοσμένη, ξεχωριστά και στην συνέχεια οι προβλέψεις τους αθροίζονται για να δώσουν τις προβλέψεις της αρχική χρονοσειράς. Εαν οι συνθήκες του προβλήματος που μελετάμε το επιτρέπουν, για την εποχιακή συνιστώσα επιλέγεται μια απλή εποχιακή μέθοδος όπως η SNAIVE.

Το θετικό αυτής της διαδικασία είναι ότι η ανάλυση STL μπορεί να αντιμετωπίσει σύνθετες εποχικότητες σε αντίθεση με τις μεθόδους που αναφέρθηκαν πιο πάνω. Η ημίωρη ζήτηση ηλεκτρικής ενέργειας πέραν της 48-ωρης εποχιακής περιόδου, εμφανίζει και εβδομαδιαία εποχιακή περίοδο. Επομένως πραγματοποιώντας ανάλυση STL στην χρονοσειρά της ζήτησης ηλεκτρικής ενέργειας, την
διασπάμε σε 3 χρονοσειρές, την εποχιακά προσαρμοσμένη, την εποχιακή συνιστώσα περιόδου 48-ωρών και την εποχιακή συνιστώσα περιόδου 336-ωρών. Οι συνιστώσες αυτές φαίνονται στο πιο κάτω γράφημα.

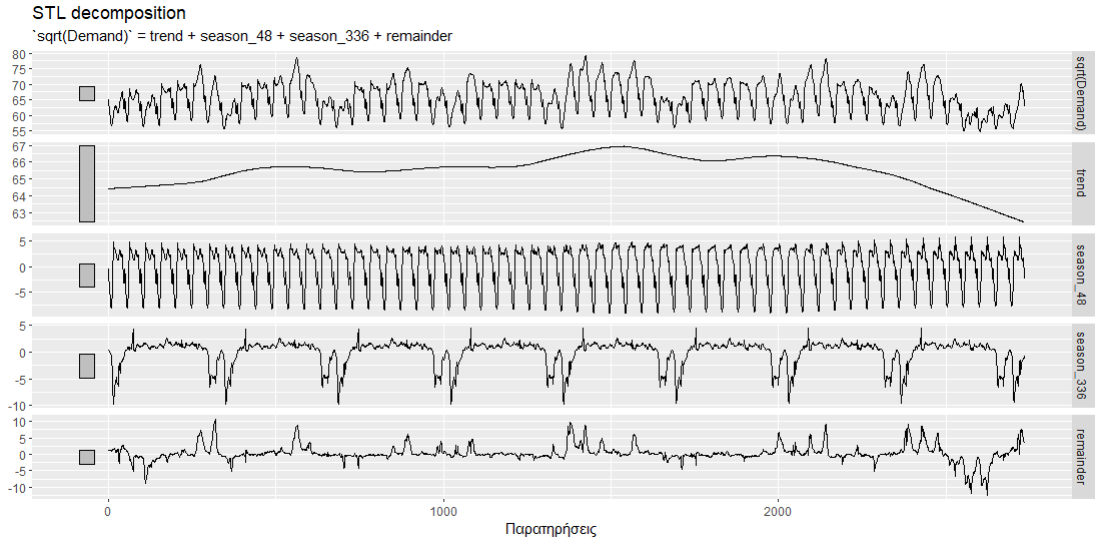
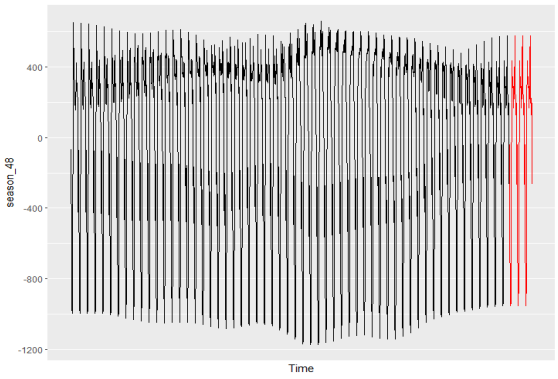


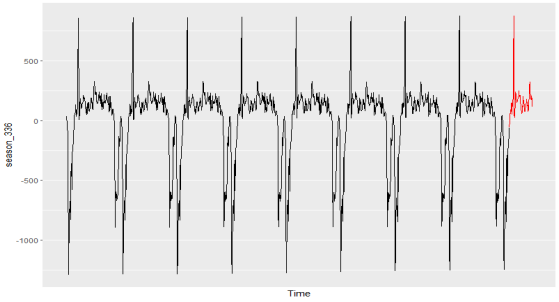
Figure 11.5.1: Συνιστώσες χρονοσειρές μετά την ανάλυση STL

Παρατηρούμε δύο ξεχωριστά εποχιακά πρότυπα που αποτυπώνονται στις συνιστώσες "$season_{48}$" και "$season_{336}$". Για τη συνιστώσα της εβδομαδιαίας εποχικότητας επιλέγεται ένα απλό SNAIVE μοντέλο πρόβλεψης με 336 υστερήσεις καθώς το μοτίβο δεν αλλάζει με την πάροδο του χρόνου και τυχόν επιρροές της θερμοκρασίας θα συμπεριληφθούν στις άλλες συνιστώσες. Για τη συνιστώσα της εποχικότητας "$Season''_{48}$ παρόλο που φαίνεται σταθερό το πρότυπο της, αντί για απλό SNAIVE προτιμάται δυναμική παλινδρόμηση με σφάλματα ARIMA και παράγοντες πρόβλεψης τους T και $\mathbf{1}(\mathbf{WD})$.

Πιο κάτω φαίνονται γραφικά οι προβλέψεις από τα δύο μοντέλα αυτά.



(a) Πρόβλεψη με μοντέλο Δυναμικής παλινδρόμησης και σφάλματα $ARIMA(1,0,5)(0,1,0)_{[48]}$ για την ημερήσια εποχικότητα



(b) Πρόβλεψη με το μοντέλο SNAIVE για την συνιστώσα εβδομαδιαίας εποχικότητας

Figure 11.5.2: Προβλέψεις μελλοντικών τιμών των σειρών εποχικότητας

Το αρνητικό που παρατηρούμε από αυτή την ανάλυση είναι ότι η συνιστώσα υπολοίπου δεν μοιάζει με λευκό θόρυβο, γεγονός που υποδεικνύει ότι υπάρχουν και άλλα πρότυπα τα οποία δεν μπόρεσαν οι συνιστώσες τις εποχικότητας και τάσης να συμπεριλάβουν. Για αυτό τον λόγο η όποια πρόβλεψη γίνει στην εποχιακά προσαρμοσμένη χρονοσειρά δεν θα στηρίζεται σε κάποιο ξεκάθαρο πρότυπο καθιστώντας την αναξιόπιστη. Απόδειξη σε αυτό είναι και η πολύ κακή προβλεπτική ικανότητα που είχε ένα μοντέλο δυναμικής παλινδρόμησης με σφάλματα ARIMA, αφού το μοντέλο δεν λαμβάνει καθόλου υπόψη τον παράγοντα της θερμοκρασίας διότι η ακαταστασία της εποχιακά προσαρμοσμένης χρονοσειράς δεν επέτρεπε να προσαρμοστεί ένα ικανό μοντέλο. Πιο κάτω παρουσιάζονται γραφικά οι προβλέψεις από το καλύτερο δυνατό μοντέλο που προσαρμόστηκε (εκθετική εξομάλυνση) στην εποχιακά προσαρμοσμένη χρονοσειρά για να γίνουν πιό εμφανές τα όσα σχολιάστηκαν

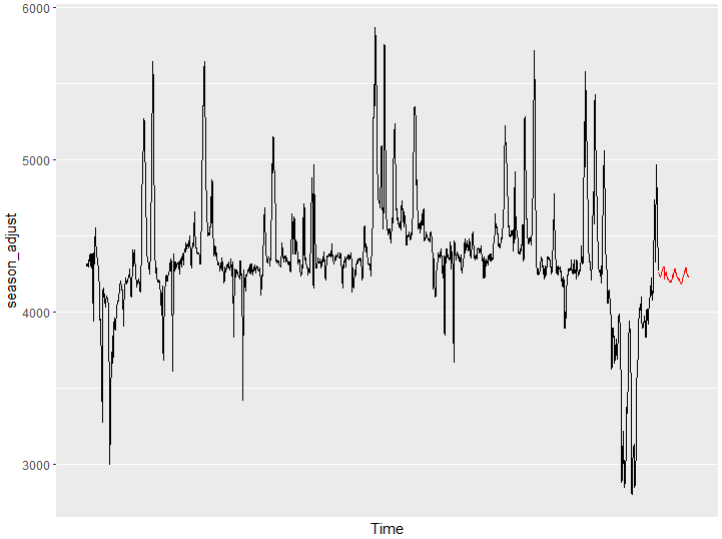


Figure 11.5.3: Πρόβλεψη των μελλοντικών τιμών της εποχιακά προσαρμοσμένης χρονοσειράς με μοντέλο εκθετικής εξομάλυνσης

Μια λύση σε αυτό το πρόβλημα είναι η αντικατάσταση της εποχιακά προσαρμοσμένης χρονοσειράς από τη συνιστώσα της τάσης. Για την κατασκευή ενός ικανού μοντέλου, το οποίο θα είναι ανθεκτικό σε απρόσμενες για την εποχή αλλαγές θερμοκρασίας (οι οποίες μπορούν να προβλεφθούν αρκετά καλά από ένα μετεωρολόγο), στο μοντέλο πρόβλεψης της τάσης καλό είναι να συμπεριληφθεί ο παράγοντας πρόβλεψης της θερμοκρασίας και η $\mathbf{1(WD)_t}$. Λαμβάνοντας προβλέψεις για την συνιστώσα της τάσης με μοντέλο δυναμικής παλινδρόμησης, και αθροίζοντας τις προβλέψεις αυτές με τις προβλέψεις που προέκυψαν από τα μοντέλα στις εποχιακές συνιστώσες προκύπτουν οι εξής προβλέψεις για την ζήτηση ηλεκτρικής ενέργειας των επόμενων 3 ημερών.

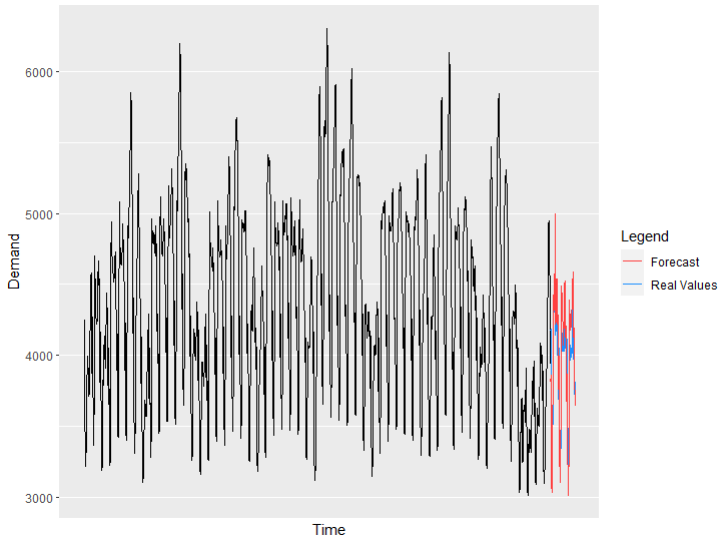


Figure 11.5.4: Πρόβλεψη των μελλοντικών τιμών της ζήτησης ηλεκτρικής ενέργειας με το μοντέλο 4, με την χρήση της συνιστώσας της τάσης στην θέση της εποχιακά προσαρμοσμένης χρονοσειράς

Χαρακτηριστικά απόδοσης Μοντέλου 4	
MAPE	4.768172
RMSE	230.5097
MAE	186.9912

Παρόλο που το μοντέλο αυτό έχει πολύ καλύτερη απόδοση από τα προηγούμενα τρία, το γεγονός ότι αγνοήσαμε τις πληροφορίες που υπήρχαν στην συνιστώσα υπολοίπου (και συνάμα στην εποχιακά προσαρμοσμένη χρονοσειρά) εγκυμονεί κινδύνους για την χρήση του μοντέλου αυτού σε μελλοντικές προβλέψεις, ιδιαίτερα εαν τα εποχιακά πρότυπα αυτά επηρεάζουν μελλοντικές τιμές (πέραν του συνόλου αξιολόγησης).

11.6 Μοντέλο 5: Πρόβλεψη με ανάλυση κυματιδίων(WTF)

Στην συνέχεια θα προσαρμόσουμε ένα πιο περίπλοκο μοντέλο το οποίο βασίζεται στον μετασχηματισμό κυματιδίων και είναι ικανό να επιλύσει προβλήματα που προέκυψαν κατά τη μοντελοποίηση του μοντέλου πρόβλεψης με την ανάλυση STL. Το μοντέλο αυτό αναλύεται στα εξής βήματα

Διακριτός μετασχηματισμός κυματιδίων

Αρχικά χρησιμοποιώντας το διακριτό μετασχηματισμό κυματιδίων (με τη μορφή της αναπαράστασης πολλαπλής ανάλυσης-MRA) αναλύουμε την χρονοσειρά σε συνιστώσες προσέγγισης και συνιστώσες λεπτομέρειας, οι οποίες καλούνται **συνιστώσες χρονοσειρές**. Οι συνιστώσες χρονοσειρές που προκύπτουν είναι αρκετά πιο ομαλές και στάσιμες σε σχέση με την αρχική χρονοσειρά. Επίσης τα συνολικά πρότυπα που κρύβει η αρχική χρονοσειρά μπορούν να γίνουν πιο εμφανή σε κάθε συνιστώσα χρονοσειρά, με καθεμιά να εκφράζει και άλλο πρότυπο **διαφορετικής συχνότητας**, καθώς όπως αναφέραμε στο κεφάλαιο 4, κάθε συνιστώσα παρουσιάζει διαφορετικές συχνότητες της χρονοσειράς. Η συνιστώσα προσέγγισης είναι αρκετά όμοια με την αρχική χρονοσειρά αλλά ομαλότερη (όσο πιο μεγάλη η τάξη του μετασχηματισμού, τόσο ομαλότερη είναι) ενώ οι συνιστώσες λεπτομέρειας αναπαριστούν κυρίως εποχιακά μοτίβα σε διαφορετικές συχνότητες.

Πιο κάτω βλέπουμε τις συνιστώσες που προκύπτουν μετά από ένα μετασχηματισμό με χρήση του κυματιδίου "Daubechies10", μαζί με την αρχική χρονοσειρά.

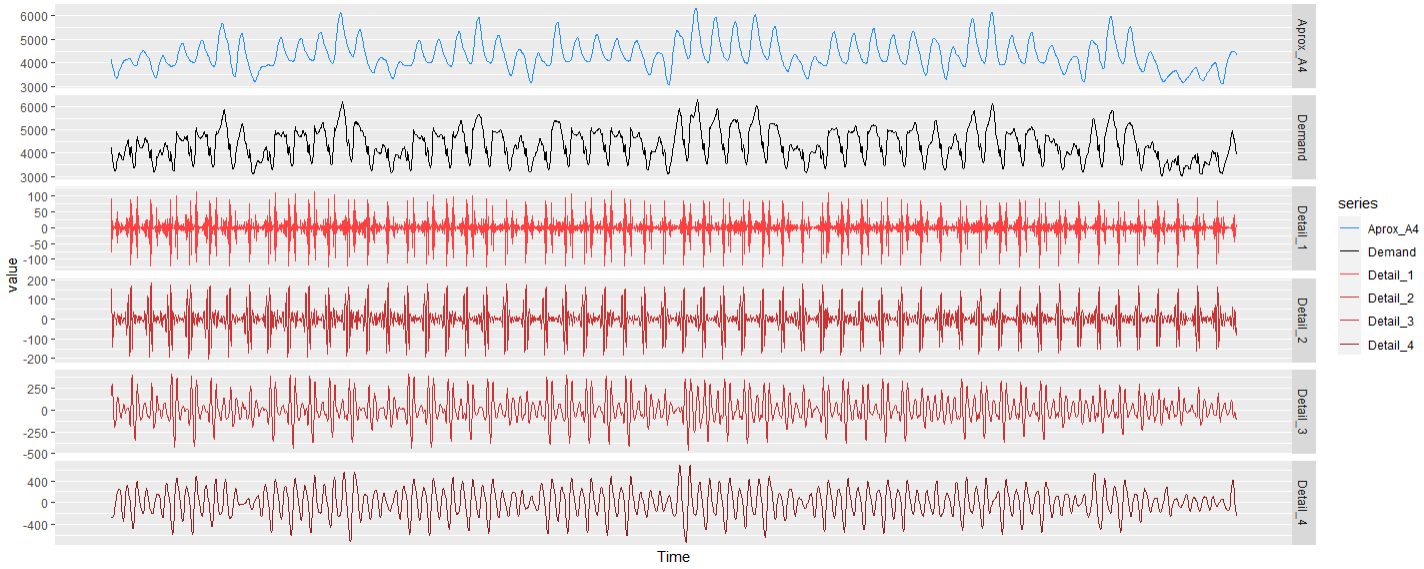


Figure 11.6.1: Συνιστώσες χρονοσειρές που προκύπτουν από τον διακριτό μετασχηματισμό κυματιδίων

Πράγματι, η συνιστώσα προσέγγισης (τάξης 4) μοιάζει αρκετά με την πραγματική χρονοσειρά αλλά έχει λιγότερο "θόρυβο". Στις χρονοσειρές λεπτομέρειας φαίνονται ξεκάθαρα εποχιακά πρότυπα τα οποία χαρακτηρίζονται από μεγαλύτερη συχνότητα στην πρώτη συνιστώσα λεπτομέρειας και μικρότερη όσο αυξάνεται η κλίμακα. Η διαδικασία πρόβλεψης σε καθεμία από αυτές τις συνιστώσες είναι ευκολότερη καθώς τα πρότυπα μπορούν να εντοπιστούν με μεγαλύτερη ευκολία. Σε κάθε συνιστώσα κυριαρχεί εποχιακό πρότυπο περιόδου 48 ωρών και εβδομαδιαίο εποχιακό πρότυπο (περιόδου 7×48). Είναι σημαντικό να αναφέρουμε ότι,

εαν χρησιμοποιούσαμε χρονοσειρά μεγαλύτερου μήκους (π.χ 2 χρόνια) τότε στις συνιστώσες λεπτομέρειας μεγάλης κλίμακας θα είχαν εμφανιστεί χρονιαία εποχιακά πρότυπα ενώ στις συνιστώσες μικρότερης κλίμακας θα είχαν εμφανιστεί τα ημίωρα πρότυπα.

Συνεχής μετασχηματισμός κυματιδίων για εντοπισμό συχνοτήτων με το πέρασμα του χρόνου

Στην συνέχεια πραγματοποιούμε ένα συνεχή μετασχηματισμό κυματιδίων (με κυματίδιο Morlet) έτσι ώστε να λάβουμε το scalogram, το οποίο είναι χρήσιμο εργαλείο για τον εντοπισμό των συχνοτήτων που παρατηρούνται με την πάροδο του χρόνου.

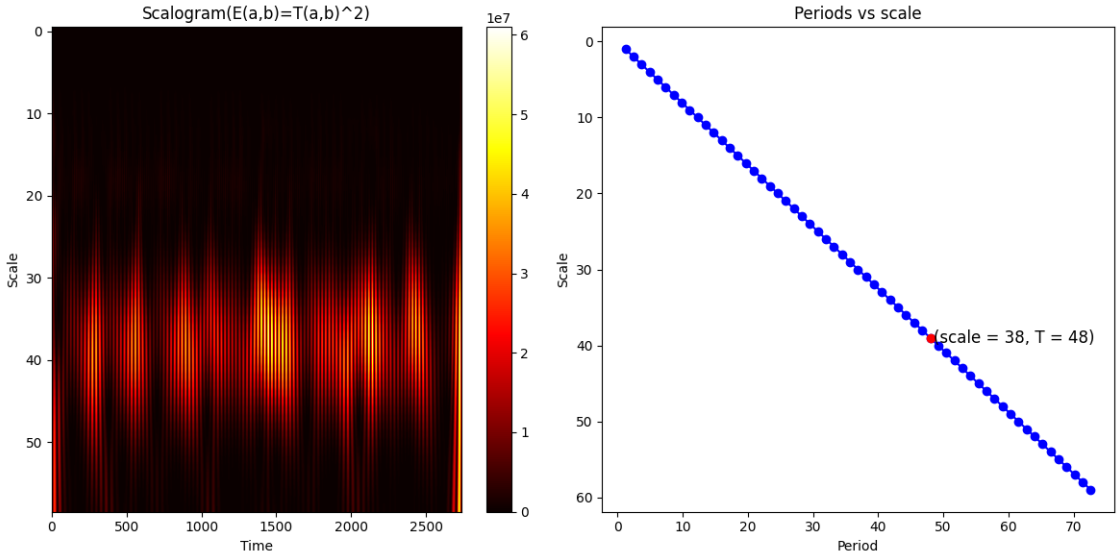


Figure 11.6.2: Scalogram συνεχή μετασχηματισμού κυματιδίων, με το Morlet κυματίδιο, για κλίμακες στο διάστημα [1,60]

Από το πιο πάνω γράφημα είναι εμφανές ότι καθ'όλη την διάρκεια του χρόνου εμφανίζονται οι συχνότητες που αντιστοιχούν σε κλίμακες μεταξύ 35 και 42. Μελετώντας καλύτερα (στην Python) τις συνιστώσες που αναπαριστώνται στο διάγραμμα αυτό, εντοπίζετε ότι κυρίαρχη κλίμακα είναι η 38 η οποία αντιστοιχεί σε περίοδο 48 ωρών. Η εβδομαδιαία εποχικότητα σε αυτό το γράφημα δεν είναι εμφανής καθώς αντιστοιχεί σε κλίμακα μεγαλύτερη από αυτές που αναπαριστά το πιο πάνω διάγραμμα. Πραγματοποιώντας τον συνεχή μετασχηματισμό κυματιδίων για μεγαλύτερο εύρος κλιμάκων προκύπτει το scalogram στο οποίο είναι εμφανής η εβδομαδιαία εποχικότητα.

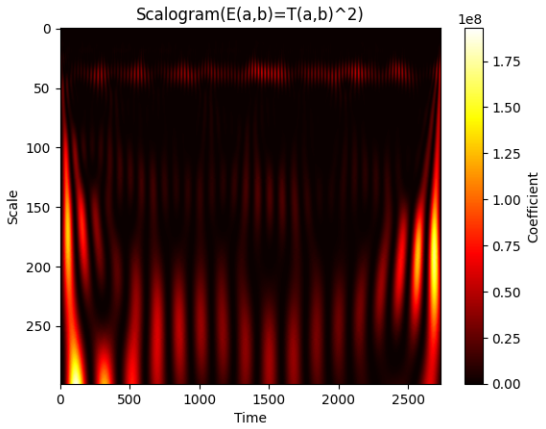


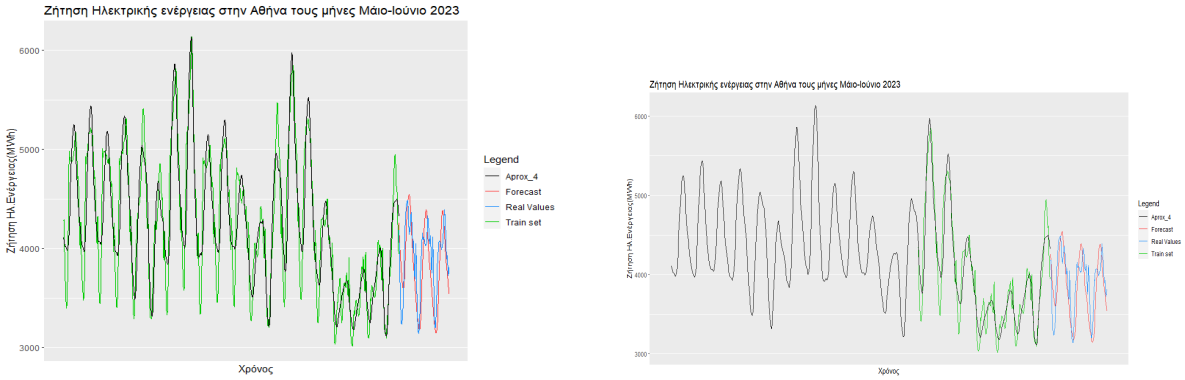
Figure 11.6.3: Scalogram συνεχή μετασχηματισμού κυματιδίων με το Morlet κυματίδιο για κλίμακες στο διάστημα [1,300]

Μελετώντας καλύτερα τις τιμές των συντελεστών που αναπαριστώνται, εντοπίστηκε ότι σε κλίμακα 272 παρατηρείται η συχνότητα $f = 0.00297619047$ η οποία αντιστοιχεί σε περίοδο 336 ωρών, δηλαδή αντιστοιχεί στην εβδομαδιαία εποχικότητα.

Προσαρμογή κατάλληλου μοντέλου για κάθε συνιστώσα χρονοσειρά

Επόμενο βήμα είναι η προσαρμογή ενός κατάλληλου μοντέλου σε κάθε συνιστώσα χρονοσειρά. Οι προβλέψεις αυτές, τοποθετούνται στο τέλος της αντίστοιχης συνιστώσας χρονοσειράς, για να προκύψουν οι **εκτεταμένες συνιστώσες χρονοσειρές**.

Για τη συνιστώσα προσέγγισης, ένα κατάλληλο μοντέλο είναι αυτό της δυναμικής παλινδρόμησης, με παράγοντες πρόβλεψης τις T και **1**(**WD**), με σφάλματα ARIMA που έχει την ικανότητα να αποτυπώσει τα μοτίβα της χρονοσειράς σε μελλοντικές προβλέψεις αλλά επίσης σέβεται τις αλλαγές που επιβάλει η αλλαγή της θερμοκρασίας, και προσαρμόζει τις προβλέψεις αναλόγως αν είναι εργάσιμη μέρα η όχι. Οι προβλέψεις για την συνιστώσα προσέγγιση φαίνονται πιο κάτω.



(a) Ολόκληρο το σύνολο εκπαίδευσης

(b) 7 τελευταίες μέρες συνόλου εκπαίδευσης

Figure 11.6.4: Προβλέψεις μελλοντικής εξέλιξης της συνιστώσας προσέγγισης τάξης 4, του μετασχηματισμού κυματιδίων

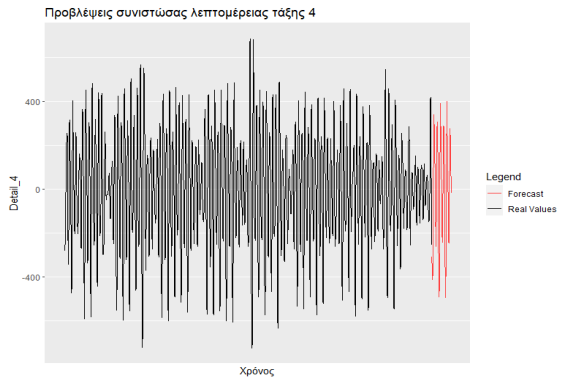
Για τις συνιστώσες λεπτομέρειας τάξης 4 και 3, απαιτείται ένα μοντέλο που να μπορεί να συλλαμβάνει τόσο ημερήσια όσο και εβδομαδιαία εποχικότητα. Ένα εποχιακό μοντέλο ARIMA, εκτός του ότι δεν μπορεί να αντιληφθεί ταυτόχρονα δύο εποχιακές περιόδου, αδυνατεί να λειτουργήσει αποτελεσματικά και γρήγορα όταν η περίοδος της εποχικότητας είναι μεγάλη, όπως συμβαίνει με τα δεδομένα ημίωρης ζήτησης ηλεκτρικής ενέργειας. Ένα μοντέλο ικανό να αντιμετωπίσει δεδομένα με σύνθετες και μεγάλες εποχιακές περιόδους, είναι η δυναμική αρμονική παλινδρόμηση με σφάλματα ARIMA (με χρήση όρων Fourier ως επεξηγηματικούς παράγοντες πρόβλεψης για την αντιμετώπιση εποχιακών προτύπων). Το αρνητικό αυτού του μοντέλου είναι ότι η εποχικότητα θεωρείται ότι είναι σταθερή, δηλαδή το εποχικό πρότυπο δεν επιτρέπεται να αλλάζει με την πάροδο του χρόνου. Για αυτό το λόγο στο μοντέλο συμπεριλαμβάνονται και οι παράγοντες πρόβλεψης T και $\mathbf{1}(\mathbf{WD})$. Επιπλέον το εποχιακό πρότυπο κάθε συνιστώσας λεπτομέρειας του μετασχηματισμού κυματιδίων είναι αρκετά σταθερό με την πάροδο του χρόνου, γεγονός που αποδεικνύει με ακόμα ένα τρόπο την αξία του μετασχηματισμού. Πιο συγκεκριμένα οι προβλέψεις προκύπτουν από την εξής σχέση.

Μοντέλο πρόβλεψης συνιστωσών λεπτομέρειας τάξης 3 και 4

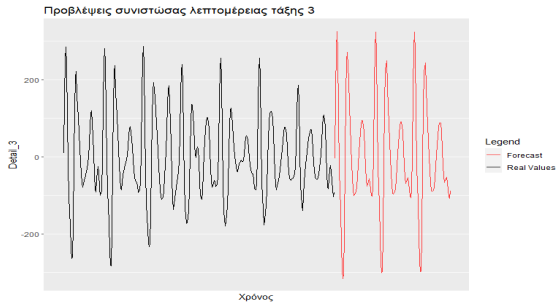
$$y_t = \beta_1 T + \beta_2 \mathbf{1}(\mathbf{WD}) + \sum_{k=1}^{K}[a_{1k}s_k(t) + \gamma_{1k}c_k(t)]_{m=48} + \sum_{k=1}^{K}[a_{2k}s_k(t) + \gamma_{2k}c_k(t)]_{m=336} + \eta_t$$

$\eta_t \sim$ ARIMA(p,d,q)-Μη εποχικό

Οι προβλέψεις την συνιστώσας λεπτομέρειας τάξης 4 και 3 παρουσιάζονται στις γραφικές (11.6.5 (a)) και (11.6.5(b)) αντίστοιχα.



(a) Προβλέψεις συνιστώσας λεπτομέρειας τάξης 4 (μαζί με ολόκληρο το σύνολο εκπαίδευσης)



(b) Προβλέψεις συνιστώσας λεπτομέρειας τάξης 3 (μαζί με τις 7 τελευταίες μέρες συνόλου εκπαίδευσης)

Figure 11.6.5: Προβλέψεις μελλοντικής εξέλιξης των συνιστωσών λεπτομέρειας τάξης 4 και 3 του μετασχηματισμού κυματιδίων

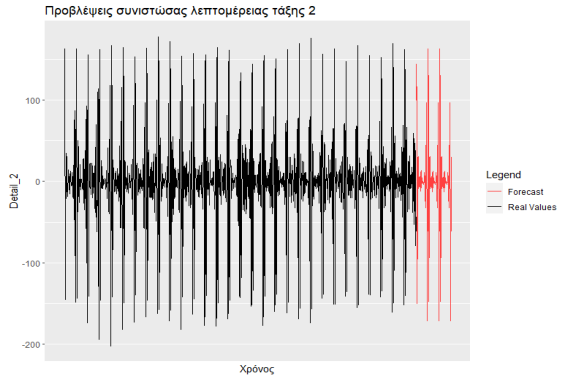
Όπως αναφέρθηκε και στο κεφάλαιο 4, οι συνιστώσες λεπτομέρειας μικρότερης τάξης του διακριτού μετασχηματισμού κυματιδίων, αντιστοιχούν σε μεγαλύτερες συχνότητες. Αυτό είναι εμφανές στις συνιστώσες λεπτομέρειας 1 και 2, όπου φαίνεται να παρουσιάζεται έντονα μόνο το ημερήσιο πρότυπο και όχι τόσο το εβδομαδιαίο. Για αυτό τον λόγο, για την πρόβλεψη μελλοντικών τιμών τους, χρησιμοποιείται δυναμική αρμονική παλινδρόμηση με όρους Fourier περιόδου μόνο 48. Επίσης λόγο της σταθερότητας των συνιστωσών αυτών η χρήση των παραγόντων πρόβλεψης T και $\mathbf{1}(\mathbf{WD})$ δεν είναι απαραίτητη. Αρκετές φορές οι συνιστώσες λεπτομέρειας μικρής τάξης μπορούν να αγνοηθούν εντελώς, γεγονός που συχνά βελτιώνει και την απόδοση του μοντέλου πρόβλεψης με μετασχηματισμό κυματιδίων. Επειδή όμως, το επιπεδο ανάλυσης που έγινε δεν είναι μεγάλο, θα χρησιμοποιηθούν για την παραγωγή των τελικών προβλέψεων. Το μοντέλο λοιπόν που προσαρμόστηκε για αυτές τις 2 συνιστώσες, εκφράζεται από τις σχέσεις

Μοντέλο πρόβλεψης συνιστωσών λεπτομέρειας τάξης 2 και 1

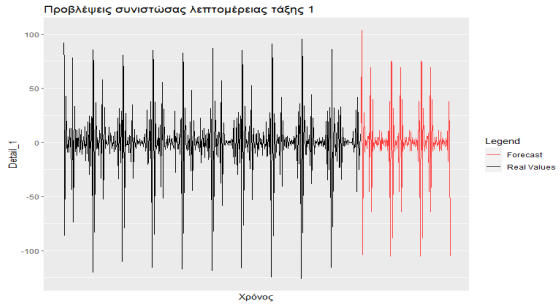
$$y_t = \sum_{k=1}^{K} [a_k s_k(t) + \gamma_k c_k(t)]_{m=48} + \eta_t$$

$$\eta_t \sim \text{ARIMA(p,d,q)-Μη εποχικό}$$

και οι γραφικές των προβλέψεων είναι οι εξής:



(a) Προβλέψεις μελλοντικής εξέλιξης της συνιστώσας λεπτομέρειας τάξης 2, του μετασχηματισμού κυματιδίων



(b) Προβλέψεις μελλοντικής εξέλιξης της συνιστώσας λεπτομέρειας τάξης 1, του μετασχηματισμού κυματιδίων

Figure 11.6.6: Προβλέψεις μελλοντικής εξέλιξης συνιστωσών λεπτομέρειας μικρότερης τάξης, του μετασχηματισμού κυματιδίων

Αντίστροφος διακριτός μετασχηματισμός κυματιδίων

Τελευταίο βήμα του μοντέλου, είναι ο αντίστροφος μετασχηματισμός κυματιδίων των εκτεταμένων συνιστωσών χρονοσειρών. Στον διακριτό μετασχηματισμό, με την μορφή της αναπαράσταση πολλαπλής ανάλυσης-MRA, αυτό μεταφράζεται σε άθροισμα των εκτεταμένων συνιστωσών χρονοσειρών.

Εαν συμβολίσουμε με $\hat{A4}_{T+h}$ την πρόβλεψη για την συνιστώσα προσέγγισης με χρονικό ορίζοντα h και $\hat{D1}_{T+h}, ..\hat{D4}_{T+h}$ τις προβλέψεις για τις συνιστώσες λεπτομέρειας, τότε η πρόβλεψη h-βημάτων της αρχικής χρονοσειράς προκύπτει από τη σχέση

$$\hat{y}_{T+h} = \hat{A4}_{T+h} + \hat{D1}_{T+h} + ... + \hat{D4}_{T+h}$$

Εφαρμόζοντας λοιπόν τον αντίστροφο μετασχηματισμό κυματιδίων, οι τελικές προβλέψεις για την ημίωρη ζήτηση ηλεκτρικής ενέργειας για τις επόμενες 3 ημέρες παρουσιάζονται στην πιο κάτω γραφική.

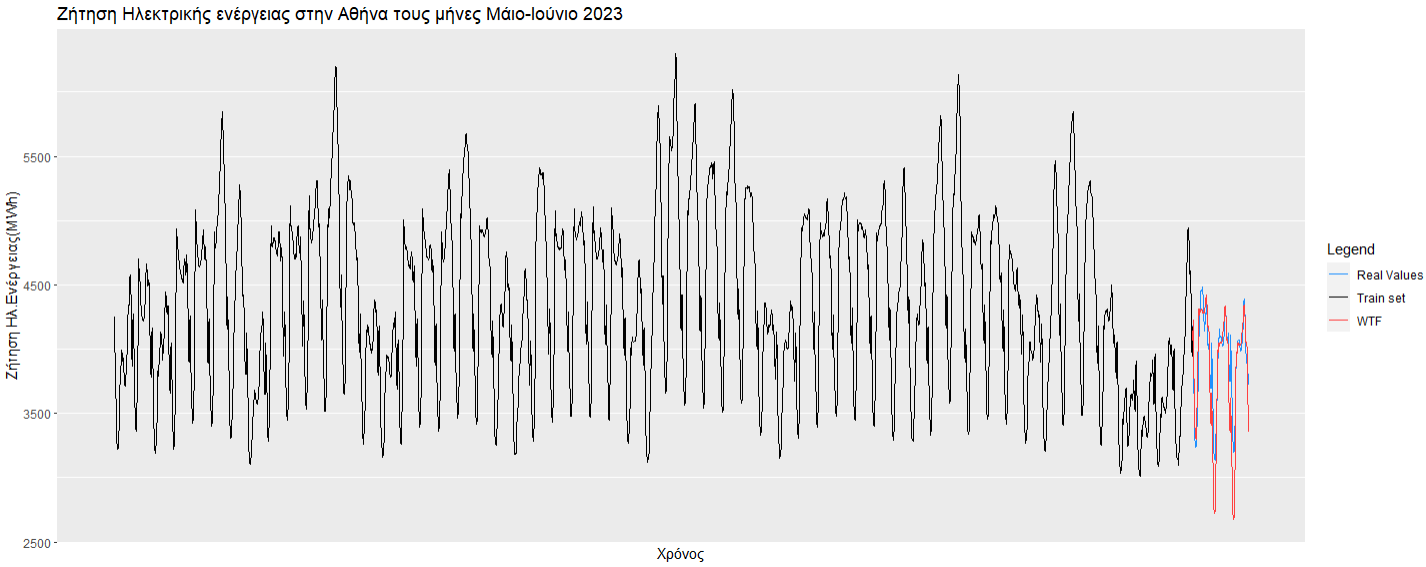


Figure 11.6.7: Προβλέψεις ημίωρης ζήτησης ηλεκτρικής ενέργειας για τις επόμενες 3 μέρες με το Μοντέλο 5

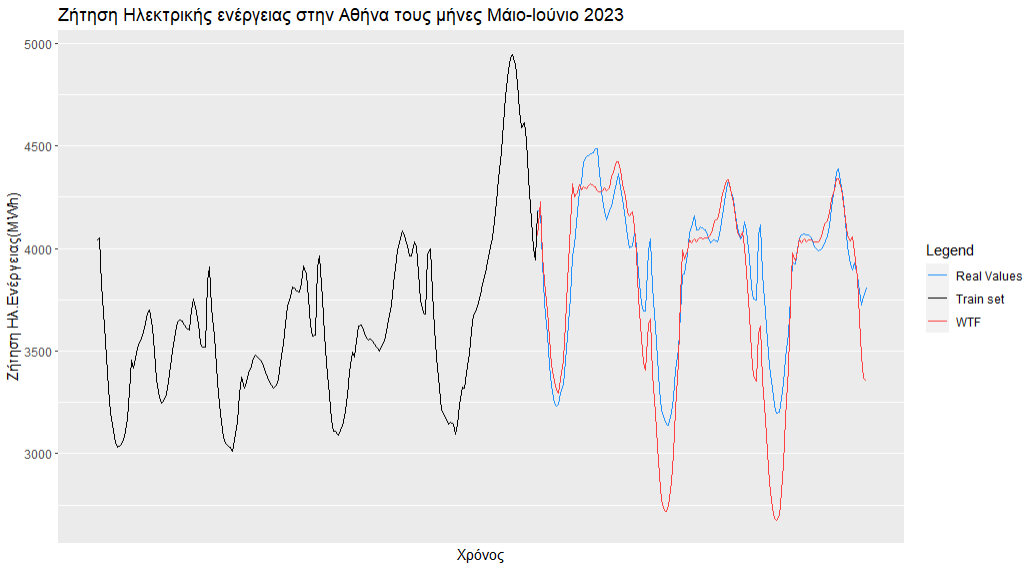


Figure 11.6.8: Προβλέψεις ημίωρης ζήτησης ηλεκτρικής ενέργειας για τις επόμενες 3 μέρες με το Μοντέλο 5, μαζί με τα δεδομένα εκπαίδευσης μόνο των τελευταίων 7 ημερών

Μια σημαντική παρατήρηση από τις πιο πάνω γραφικές είναι ότι εαν επιβάλαμε κάποια κατώτερη τιμή πρόβλεψης τότε η απόδοση του μοντέλου θα ήταν ακόμα καλύτερη. Είναι γνωστό τόσο από τα διαθέσιμα δεδομένα όσο και από τους αρμόδιους υπαλλήλους της ΔΕΗ ότι, η ημίωρη ζήτηση ηλεκτρικής ενέργειας δεν είναι κάτω από 3000 MWh, εκτός από ειδικές περιπτώσεις. Επομένως μπορεί να οριστεί ένα υπο-μοντέλο, το οποίο θα εφαρμόζεται όταν δεν εμφανίζεται κάποιο ακραίο γεγονός που θα αναγκάσει την ζήτηση κάποια στιγμή (κάποιο ημίωρο) της ημέρας να είναι κάτω από 3000, και θα έχει ακόμα καλύτερη απόδοση από το μοντέλο αυτό. Για να γίνει αυτό στηριζόμαστε σε μαθηματικούς μετασχηματισμούς των δεδομένων. Πιο συγκεκριμένα εφαρμόζουμε ένα κλιμακωτό λογιστικό μετασχηματισμό ο οποίος αντιστοιχεί το διάστημα [3000,7000] σε ολόκληρη την πραγματική γραμμή και ορίζεται από την σχέση

$$y = log(\frac{x - 3000}{7000 - x})$$

όπου με y συμβολίζουμε τα μετασχηματισμένα δεδομένα και με x συμβολίζουμε τα δεδομένα στην αρχική τους κλίμακα. Η απόδοση του μοντέλου 5 και του υπο-μοντέλου του, στο σύνολο αξιολόγησης, συνοψίζεται ως

Χαρακτηριστικά απόδοσης Μοντέλου 5(υπο-Μοντέλου 5)	
MAPE	4.443903 (3.825728)
RMSE	228.6182 (195.5997)
MAE	163.7542 (143.8264)

11.7 Μοντέλο 6: Πρόβλεψη με την ιεραρχική δομή του διακριτού μετασχηματισμού κυματιδίων (HSWTF)

Ένας εναλλακτικό τρόπος να προκύψουν οι προβλέψεις της αρχικής χρονοσειράς μέσω των προβλέψεων των συνιστωσών χρονοσειρών, του μετασχηματισμού κυματιδίου, είναι να εργαστούμε στο πλαίσιο των ιεραρχικών χρονοσειρών. Η ιεραρχική δομή που προκύπτει μετά από την ανάλυση κυματιδίων της χρονοσειράς δίνεται από το πιο κάτω διάγραμμα

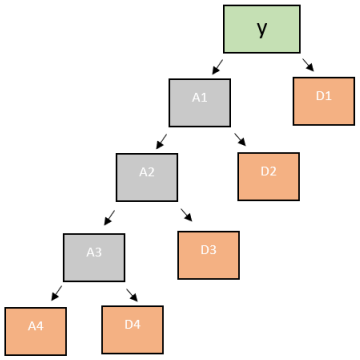


Figure 11.7.1: Ιεραρχική δομή αρχικής χρονοσειράς μετά την ανάλυση κυματιδίων. Με πορτοκαλί χρώμα εκφράζονται οι χρονοσειρές του κατώτερου επιπέδου. Με Ai εκφράζεται η προσέγγιση τω i-στού επιπέδου και Di είναι η λεπτομέρεια του i-οστού επιπέδου

Εαν θεωρήσουμε το διάνυσμα $X = [y, A_1, A_2, A_3, A_4, D_1, D_2, D_3, D_4]'$ που εκφράζει το σύνολο των χρονοσειρών και υποσειρών που ανήκουν στην πιο πάνω δομή, και το διάνυσμα $\mathbf{X_b} = [A_4, D_4, D_3, D_2, D_1]'$, που περιέχει τις υποσειρές βάσης της ιεραρχικής δομής, τότε η ιεραρχική δομή χρονοσειράς επιβάλει την πιο κάτω σχέση και περιορισμούς

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \mathbf{X_b} = S\mathbf{X_b}$$

Εαν πραγματοποιήθουν προβλέψεις για κάθε χρονοσειρά της δομής, δηλαδή να ληφθούν οι βασικές προβλέψεις $\hat{x} = [\hat{y}, \hat{A}_1, \hat{A}_2, \hat{A}_3, \hat{A}_4, \hat{D}_1, \hat{D}_2, \hat{D}_3, \hat{D}_4]'$, τότε όπως αναφέρθηκε στο υποκεφάλαιο 8.1 οι συνεκτικές

προβλέψεις προκύπτουν από την σχέση

$$\tilde{\mathbf{X}} = SG\hat{\mathbf{X}} \quad , \quad G = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Υπάρχουν αρκετοί πίνακες G για τους οποίους προκύπτουν οι συνεκτικές προβλέψεις. Στόχος είναι να βρεθεί ο πίνακας G για τον οποίο ελαχιτοποιείται η συνολική διακύμανση πρόβλεψης, του συνόλου των συνεκτικών προβλέψεων, και οι προβλέψεις να είναι αμερόληπτες. Δηλαδή το πρόβλημα ανάγεται σε πρόβλημα βελτιστοποίησης υπο περιορισμούς.

$$\begin{aligned} & min_{[G]}\mathbf{V}_h \quad , \quad \mathbf{V}_h = Var[X_{T+h} - \tilde{X}_h] \\ & \mathbf{E}[\tilde{X}_h] = X_h \end{aligned}$$

Όπως αναφέρθηκε στο κεφάλαιο 8 το πρόβλημα αυτό, εφόσον οι βασικές προβλέψεις είναι αμερόληπτες, ανάγεται στο εξής

$$\begin{aligned} & min_{[G]}\left(trace(\mathbf{V}_h)\right) \\ & SGS = S \end{aligned}$$

Η λύση του προβλήματος βελτιστοποίησης, δίνεται από την σχέση 8.2.1 που δίνει τις προβλέψεις με την προσέγγιση βέλτιστου συμβιβασμού MinT. Εαν επιλεχθεί ο πίνακας $k_h\mathbf{I}$ ως προσέγγιση του πίνακα W_h^{-1} τότε οι συμβιβαστικές προβλέψεις προκύπτουν από την σχέση

$$\tilde{X}_h = S(S'S)^{-1}S'\hat{x}_h \tag{11.7.1}$$

η οποία είναι η σχέση που δίνει την ευθεία παλινδρόμησης ελαχίστων τετραγώνων για πίνακα σχεδίασης $X = S$ και τιμές μεταβλητής πρόβλεψης $\hat{x}$.

Επομένως το μόνο που υπολείπεται για να λάβουμε τις συμβιβαστικές προβλέψεις, με την προσέγγιση βέλτιστου συμβιβασμού MinT, είναι οι προβλέψεις για κάθε χρονοσειρά της δομής που επιβάλει ο μετασχηματισμός κυματιδίων. Για τις συνιστώσες λεπτομέρειας και τη συνιστώσα προσέγγισης τάξης 4 έχουν ήδη παραχθεί προβλέψεις. Απομένει λοιπόν να παραχθούν προβλέψει και για τις συνιστώσες προσέγγισης τάξης τρία, δύο και ένα. Οι συνιστώσες αυτές δεν δίνονται από τον αλγόριθμο υπολογισμού του διακριτού μετασχηματισμού στην Python, αλλά προκύπτουν εύκολα μέσω της σχέσης 4.4.2

$$\begin{aligned} A_3 &= A_4 + D_4 \\ A_2 &= A_3 + D_3 \\ A_1 &= A_2 + D_2 \end{aligned} \tag{11.7.2}$$

Αφού προσαρμόστηκε ένα κατάλληλο μοντέλο δυναμικής παλινδρόμησης και για τις συνιστώσες προσέγγισης τάξης 1, 2 και 3, αξιοποιώντας την σχέση (11.7.1) προέκυψαν οι συμβιβαστικές προβλέψεις για όλες τις χρονοσειρές της δομής. Η πρώτη γραμμή του πίνακα $\tilde{y}_h$ περιέχει τις συμβιβαστικές προβλέψεις για την αρχική χρονοσειρά, δηλαδή τις προβλέψεις της ημίωρης ζήτησης ηλεκτρικής ενέργειας στην Αθήνα

για τις επόμενες 3 ημέρες. Πιο κάτω παρουσιάζονται στο ίδιο γράφημα οι προβλέψεις από το μοντέλο 5 και από το μοντέλο 6.

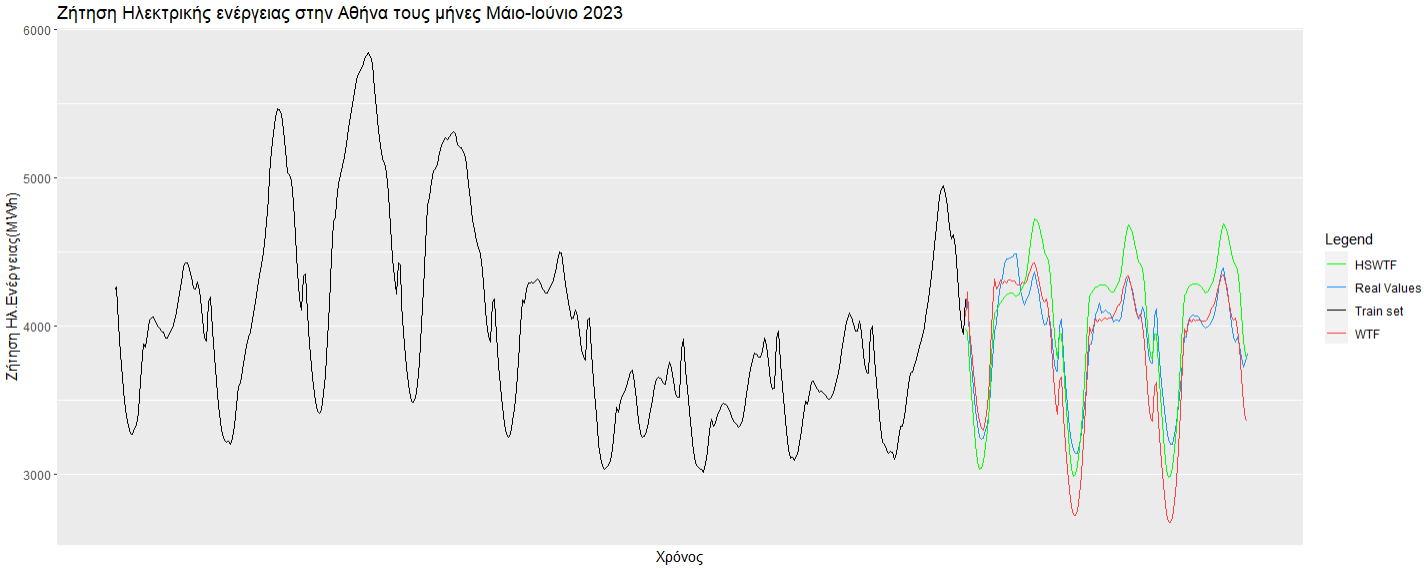


Figure 11.7.2: Προβλέψεις ημίωρης ζήτησης ηλεκτρικής ενέργειας για τις επόμενες 3 μέρες με το Μοντέλο 6 (HSWTF) και το μοντέλο 5(WTF)

Από την πιο πανω γραφική είναι εμφανές ότι σε κάποια σημεία το μοντέλο 6 πραγματοποιεί πιο καλές προβλέψεις ενώ σε άλλα σημεία υπερτερεί το μοντέλο 5. Μια συνήθης πρακτική για την παραγωγή καλύτερων προβλέψεων, όπως αναφέρθηκε στο υποκεφάλαιο 10.2, είναι ο συνδυασμός προβλέψεων από ικανά μοντέλα μέσω κάποιας συνάρτησης συνάθροισης. Παρατηρήθηκε ότι οι συνδυαστικές προβλέψεις από τα μοντέλα 5 και 6 μέσω της συνάρτησης του μέσου είναι καλύτερες από τις προβλέψεις των επιμέρους μοντέλων, όπως φαίνεται και στο πιο κάτω διάγραμμα και πίνακα μέτρων απόδοσης.

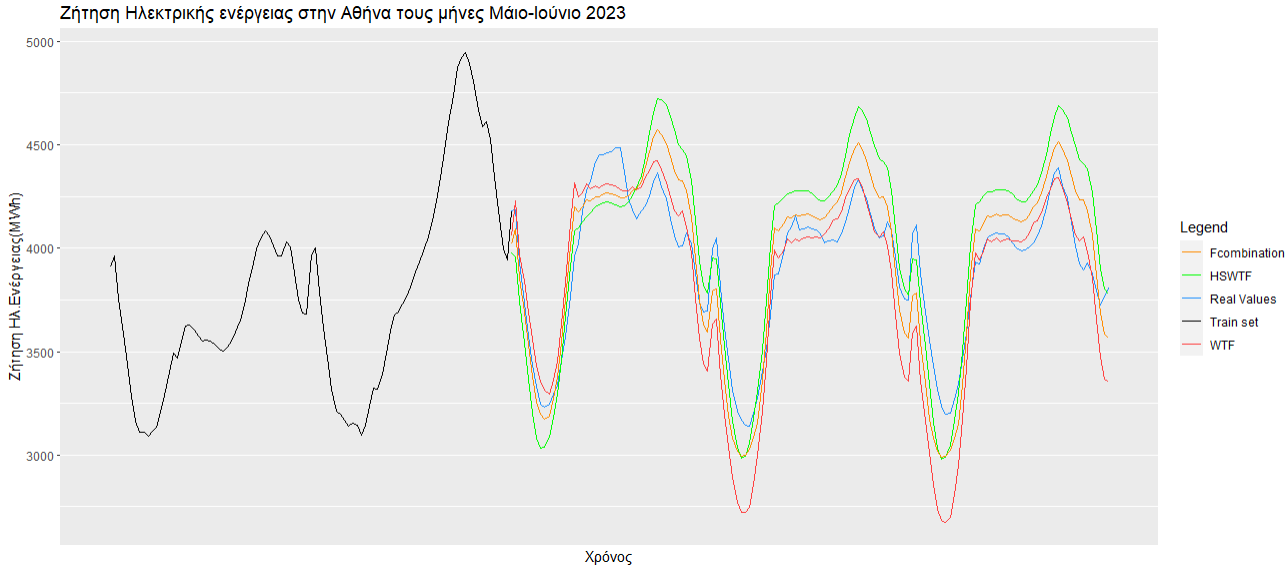


Figure 11.7.3: Συνδυαστικές προβλέψεις, από τα μοντέλα 5 και 6, της ημίωρης ζήτησης ηλεκτρικής ενέργειας για τις επόμενες 3 μέρες, στο ίδιο γράφημα με τις προβλέψεις από τα μοντέλα 5 και 6

	Μετρικές Απόδοσης		
	MAPE	**RMSE**	**MAE**
Μοντέλο 5	4.443903	228.6182	163.7542
Μοντέλο 6	5.711294	254.8709	226.0616
Συνδιασμένες προβλέψεις	4.082811	180.2903	159.1019

Table 11.7.1: Μετρικές απόδοσης μοντέλων που περιλαμβάνουν τον μετασχηματισμό κυματιδίων

Γενικό συμπέρασμα:

Με βάση την απόδοση των μοντέλων που σχολιάστηκαν, μπορούμε να καταλήξουμε στο συμπέρασμα ότι για δεδομένα χρονοσειρών με μεγάλες εποχιακές περιόδους, είναι προτιμότερο να χρησιμοποιείται κάποια μέθοδος αποσύνθεσης της ολικής χρονοσειράς σε απλούστερες με πιο ξεκάθαρη πληροφορία. Επίσης ο μετασχηματισμός κυματιδίων φάνηκε να είναι εξαιρετικά χρήσιμος για τέτοιου είδους δεδομένα τόσο για την παραγωγή προβλέψεων αλλά και για την καλύτερη κατανόηση των προτύπων που μια χρονοσειρά κρύβει. Γενικότερα δεν μπορεί να υπάρξει απάντηση στο πιο είναι το καλύτερο μοντέλο για προβλέψεις, αλλά το μόνο που μπορούμε να απαντήσουμε είναι ότι η κάθε χρονοσειρά, πρέπει να αναλύεται όσο το δυνατό περισσότερο έτσι ώστε να μπορούν να γίνει εμφανές ποιες τεχνικές και μέθοδοι μπορούν να χρησιμοποιηθούν για την επεξήγηση της.

Όπως αναφέρθηκε πιο πάνω, για δεδομένα ημίωρης ζήτησης ηλεκτρικής ενέργειας έχει προταθεί η προσέγγιση της προσαρμογής διαφορετικού μοντέλου για κάθε μισή ώρα της ημέρας. Δηλαδή να προσαρμοστούν 48 διαφορετικά μοντέλα. Με αυτό τον τρόπο καταγράφονται καλύτερα τα πρότυπα της χρονοσειράς και οι αυτοσυχετίσεις στα υπόλοιπα (του κάθε μοντέλου) είναι στατιστικά μή σημαντικές. Τέτοια μοντέλα μελετήθηκαν στην εργασία [7]. Ένα καλύτερο μοντέλο που είναι εξέλιξη του μοντέλου 5 και 6, και συνδυάζει την προσέγγιση που αναφέραμε, είναι η χρήση ενός διορθωτικού όρου για κάθε ημιώρα, που θα

αποτυπώνει πληροφορίες που λήφθηκαν απο το αντίστοιχο μοντέλο της κάθε ημιώρας.

Βιβλιογραφία

[1] Naadimuu G Bronson R. Επιχειρησιακή Έρευνα,schaum's outline of theory and problems of operations research second edition. 17.

[2] J. M Benitez C. Bergmeir, Rob J. Hyndman. Bagging exponential smoothing methods using stl decomposition and box-cox transformation. 2017.

[3] Tsan-Ming Choi, Yong Yu, and Kin-Fan Au. A hybrid sarima wavelet transform method for sales forecasting. *Decision Support Systems*, 51(1):130–140, 2011.

[4] Terence C.Mills. Applied time series analysis. 2019.

[5] Murat Kulahci Dougla C.Montgomery, Cheryl L. Jennings. Time series analysis and forecasting. 2015.

[6] Jean-Marie Dufour. Estimation of arma models by maximum likelihood. *McGill University*, First version:1981,Revised: February 1991, September 2000.

[7] Shu Fan and Rob J. Hyndman. Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems*, 27(1):134–141, 2012.

[8] FREDERICK S. HILLIER FREDERICK S. HILLIER. Introduction to operations research,seven edition. 17.

[9] Dr. Ravi Mahendra Gor. Introduction to operations research,seven edition,chapter 6 : Forecasting techniques.

[10] Rob J. Hyndman, Roman A. Ahmed, George Athanasopoulos, and Han Lin Shang. Optimal combination forecasts for hierarchical time series. *Computational Statistics Data Analysis*, 55(9):2579–2589, 2011.

[11] Rob J Hyndman and George Athanasopoulos. Forecasting: principles and practice. 2018.

[12] Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159–178, 1992.

[13] Imed Riadh Farah Manel Rhif, Ali Ben Abbes. Wavelet transform application for/in non-stationary time-series analysis: A review. 2019.

[14] Davis F.Hendry Michael P.Clements. A companion to economic forecasting. 2004.

[15] Anastasios Panagiotelis, George Athanasopoulos, Puwasala Gamakumara, and Rob J. Hyndman. Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting*, 37(1):343–359, 2021.

[16] Jean E. McRae Irma Terpenning Robert B. Cleveland, William S. Clevelands. A seasonal-trend decomposition procedure based on loess. 2014.

[17] Paul S.Addison. The illustrated wavelet transform handbook. 2017.

[18] Lena Sasal, Tanujit Chakraborty, and Abdenour Hadid. W-transformers: A wavelet-based transformer framework for univariate time series forecasting. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 671–676, 2022.

[19] Zhongfu Tan, Jinliang Zhang, Jianhui Wang, and Jun Xu. Day-ahead electricity price forecasting using wavelet transform combined with arima and garch models. *Applied Energy*, 87(11):3606–3610, 2010.

[20] Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

[21] Jerome Friedman Trevor Hastie, Robert Tibshirani. The elements of statistical learning. 2014.

[22] Shanika L. Wickramasuriya, George Athanasopoulos, and Rob J. Hyndman. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819, 2019.

[23] Peter C. Young, Diego J. Pedregal, and Wlodek Tych. Dynamic harmonic regression. *Journal of Forecasting*, 18(6):369–394, 1999.

[24] Keyi Zhang, Ramazan Gençay, and M. Ege Yazgan. Application of wavelet decomposition in time-series forecasting. *Economics Letters*, 158:41–46, 2017.

[25] Δ.Στογιάννης Ι.Κολέτσος. ΕΠΙΧΕΙΡΗΣΙΑΚΗ ΕΡΕΥΝΑ : Θεωρία, αλγόριθμοι εφαρμογές. 2021,Έκδοση 1.