Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων

Βελτίωση ακρίβειας στατιστικών μεθόδων πρόβλεψης σε χρονοσειρές μικρού ιστορικού με χρήση τεχνικών συσταδοποίησης εποχιακών δεικτών συναφών δεδομένων

Διπλωματικη Εργασια

του

ΕΥΑΓΓΕΛΟΥ ΝΤΑΒΕΛΗ

Επιβλέπων: Βασίλειος Ασημακόπουλος
Καθηγητής Ε.Μ.Π.

Μονάδα Προβλεψεων και Στρατηγικης

Αθήνα, Ιούλιος 2017

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρικών Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων
Μονάδα Προβλέψεων και Στρατηγικής

Βελτίωση ακρίβειας στατιστικών μεθόδων πρόβλεψης σε χρονοσειρές μικρού ιστορικού με χρήση τεχνικών συσταδοποίησης εποχιακών δεικτών συναφών δεδομένων

Διπλωματικη Εργασια

του

ΕΥΑΓΓΕΛΟΣ ΝΤΑΒΕΛΗΣ

Επιβλέπων: Βασίλειος Ασημακόπουλος
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 20η Ιουλίου 2017.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Βασίλειος Ασημακόπουλος	Ιωάννης Ψαρράς	Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2017

(Υπογραφή)

.....

Ευαγγελος Νταβελης

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

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρικών Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων
Μονάδα Προβλέψεων και Στρατηγικής

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Βασίλειο Ασημακόπουλο για την ευκαιρία που μου έδωσε να εκπονήσω τη παρούσα διπλωματική και την υποστήριξή του σε όλη την πορεία της.

Επίσης, θα ήθελα να ευχαριστήσω τους καθηγητές κ. Ιωάννη Ψαρρά και κ. Δημήτριο Ασκούνη για την τιμή που μου έκαναν να συμμετάσχουν στην επιτροπή εξέτασης της διπλωματικής.

Ευχαριστώ ιδιαίτερα τον υποψήφιο διδάκτορα Ευάγγελο Σπηλιώτη για την καθοδήγηση, στήριξη και καθοριστική βοήθεια που μου παρείχε, όπως και τα υπόλοιπα μέλη της Μονάδας Προβλέψεων και Στρατηγικής.

Θερμές ευχαριστίες θα ήθελα να απευθύνω στον Δρ Χριστόφορο Αναγνωστόπουλο και την εταιρία Mentat Innovations για την καθοδήγησή τους στα πρώτα βήματα αυτής της εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου Γιώργο, Γρηγόρη, Κατερίνα και Μαρία.

Περίληψη

Αντικείμενο της διπλωματικής εργασίας είναι η ανάπτυξη μεθοδολογίας για τη βελτίωση της ακρίβειας στατιστικών μεθόδων πρόβλεψης σε χρονοσειρές που έχουν μικρό ιστορικό παρατηρήσεων μέσω τεχνικών συσταδοποίησης εποχιακών δεικτών από συναφείς χρονοσειρές.

Οι κλασικές μέθοδοι αποσύνθεσης απαιτούν ένα ελάχιστο πλήθος παρατηρήσεων για να μπορέσουν να εξάγουν το μοτίβο της εποχιακότητας μιας χρονοσειράς. Στη πράξη, όμως, συναντάμε συχνά χρονοσειρές που αποτελούνται από μικρό πλήθος τιμών, ενώ συγχρόνως περιγράφουν εποχιακά μεγέθη.

Παράλληλα, τα τελευταία χρόνια υπάρχει αφθονία στα δεδομένα που έχουμε στη διάθεσή μας. Η παρούσα εργασία βασίζεται στην υπόθεση ότι μπορούμε να χρησιμοποιήσουμε τη διαθέσιμη πληροφορία για να εξάγουμε αντιπροσωπευτικούς δείκτες εποχιακότητας που μπορούμε να χρησιμοποιήσουμε για να αναλύσουμε και να προεκτείνουμε χρονοσειρές που χαρακτηρίζονται από μικρό ιστορικό.

Για να το κάνουμε αυτό πρέπει αρχικά να συγκεντρώσουμε ένα πλήθος χρονοσειρών που περιγράφει παρόμοια φυσικά μεγέθη. Έπειτα, χρησιμοποιώντας τεχνικές συσταδοποίησης στους δείκτες εποχιακότητας αυτών που έχουν επαρκή δεδομένα για να εφαρμόσουμε τις κλασικές μεθόδους αποσύνθεσης, δημιουργούμε συστάδες παρόμοιας εποχιακής συμπεριφοράς. Ελέγχουμε, κατόπιν, αν οι μικρές χρονοσειρές μπορούν να υπαχθούν σε αυτές τις συστάδες και αν ναι, τις προβλέπουμε με δεδομένο ότι οι δείκτες εποχιακότητας τους είναι οι ίδιοι με τους μέσους δείκτες των συστάδων.

Για να ελέγξουμε την υπόθεση, εφαρμόσαμε την μεθοδολογία που περιγράφηκε σε ένα σύνολο χρονοσειρών ζήτησης φυσικού αερίου και λάβαμε θετικά αποτελέσματα. Συγκεκριμένα συγκρίναμε τη προτεινόμενη προσέγγιση με τη κλασική, που προεκτείνει τις μικρές χρονοσειρές βάσει των αρχικών τους δεδομένων και παρατηρήσαμε σημαντική βελτίωση της ακρίβειας.

Λέξεις Κλειδιά

Χρονοσειρές, Τεχνικές Προβλέψεων, Εποχιακότητα, Συσταδοποίηση, Μικρό ιστορικό, Φυσικό Αέριο.

Abstract

The purpose of this diploma thesis is to develop a methodology for improving the accuracy of statistical forecasting methods on timeseries with short history through the use of clustering techniques on the seasonal indices of other similar timeseries.

Classical decomposition methods require a minimum number of observations to be able to detect the seasonality pattern of a timeseries. In practice, however, we often encounter timeseries lacking enough data, while at the same time describing seasonal values.

Meanwhile, in recent years, there is an abundance of accessible data. This thesis draws upon the hypothesis that we can utilise the available infomation to extract representative seasonality indices that we can use in order to analyse and extend timeseries that are characterised by short history.

In order to achieve this, we initially have to gather a large number of timeseries describing similar values. Afterwards, we create clusters of similar seasonal behaviour by using clustering techniques on the seasonality indices of series with sufficient data. Then, we check if the shorter timeseries qualify to be a part of these clusters and if so, we predict their future values as they were characterised by the seasonal behaviour of the mean indices of the cluster members.

To test our hypothesis, we applied the described methodology to a set of natural gas demand timeseries and received positive results. In particular, we compared the proposed approach to the classical one, which forecasts short timeseries based on their original data, and we have measured a significant overall improvement in accuracy.

Keywords

Timeseries, Forecasting Techniques, Seasonality, Clustering, Short history, Natural Gas.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Στις κλασικές στατιστικές μεθόδους πρόβλεψης χρονοσειρών η συνηθισμένη διαδικασία ακολουθεί μια συγκεκριμένη διαδικασία βημάτων. Αρχικά, προετοιμάζουμε τη χρονοσειρά για ανάλυση. Έπειτα, την αναλύουμε στις τέσσερις της συνιστώσες: την τάση, την εποχιακότητα, την κυκλικότητα και την τυχαιότητα. Η πρόβλεψη γίνεται στην αποεποχικοποιημένη χρονοσειρά και μετά ενσωματώνεται σε αυτή το στοιχείο της εποχιακότητας.

Όμως, τα εργαλεία για αποσύνθεση της χρονοσειράς που έχουμε στη διάθεσή μας απαιτούν η χρονοσειρά να έχει τουλάχιστον ένα πλήθος παρατηρήσεων. Σε αντίθετη περίπτωση, η χρονοσειρά προβλέπεται σαν να μην χαρακτηρίζεται από εποχιακή συμπεριφορά. Το πρόβλημα προκύπτει λοιπόν στην αδυναμία να συνυπολογίσουμε την επιρροή της εποχιακότητας στην πρόβλεψη χρονοσειρών με μικρό ιστορικό.

Παράλληλα, ζούμε σε μια εποχή που την χαρακτηρίζει αφθονία δεδομένων. Έτσι συναντάμε όλο και περισσότερο χρονοσειρές που είναι μέρος ενός ευρύτερου συνόλου που περιγράφει παρόμοια μεγέθη.

1.1 Αντικείμενο της διπλωματικής

Στη παρούσα διπλωματική θα προσπαθήσουμε να χρησιμοποιήσουμε τη πληροφορία που μας δίνεται από ένα μεγάλο σύνολο δεδομένων για να προβλέψουμε με μεγαλύτερη ακρίβεια μικρές χρονοσειρές.

Θα χωρίσουμε, λοιπόν, τις χρονοσειρές σε μικρές και μεγάλες, δηλαδή με επαρκές ιστορικό για αποεποχικοποίηση. Στις μεγάλες χρονοσειρές θα χρησιμοποιήσουμε τις γνωστές μεθόδους αποσύνθεσης για να παραγάγουμε τους δείκτες εποχιακότητας για κάθε μία από αυτές. Στις μικρές, θα υπολογίσουμε ένα σύνολο δεικτών ψευδο-εποχιακότητας.

Χρησιμοποιώντας τεχνικές συσταδοποίησης στη πρώτη ομάδα των χρονοσειρών θα ελέγξουμε αν υπάρχουν πράγματι μοτίβα εποχιακότητας που χαρακτηρίζουν υποσύνολα των δεδομένων και θα τα εντοπίσουμε. Μετά θα εξετάσουμε αν οι μικρές χρονοσειρές δύνανται να καταταχθούν σε κάποια από τις συστάδες που υπολογίσαμε βάσει των δεικτών ψευδο-εποχιακότητας τους.

Στη συνέχεια, θα προεκτείνουμε τις μικρές χρονοσειρές που βρήκαμε να ανήκουν σε κάποια

συστάδα στο μέλλον με δύο τρόπους. Πρώτα, θα τις προβλέψουμε όπως γίνεται συνήθως. Έπειτα, θα τις αποεποχικοποιήσουμε βάσει της μέσης εποχιακότητας των μεγάλων χρονοσειρών που ανήκουν στην ίδια συστάδα με αυτές, θα τις προβλέψουμε και θα τις επαναεποχικοποιήσουμε.

Τελικά θα εκτιμήσουμε αν η ακρίβεια της προτεινόμενης μεθόδου είναι μεγαλύτερη από τη κλασική προσέγγιση.

1.1.1 Συνεισφορά

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. Μελετήθηκε ένα σύνολο χρονοσειρών που περιγράφει το επίπεδο φυσικού αερίου σε δεξαμενές
2. Έγινε ανάλυση τους και μετατράπηκαν σε μορφή για μεσοπρόθεσμη πρόβλεψη
3. Υπολογίστηκαν οι ομάδες συνάφειας των μοτίβων εποχιακότητας
4. Βάσει αυτών αποεποχικοποιήσαμε χρονοσειρές μικρού ιστορικού
5. Μετρήσαμε ότι η ακρίβεια πρόβλεψης βελτιώνεται με την προτεινόμενη μέθοδο αποεποχικοποίησης

1.2 Οργάνωση του τόμου

Ακολουθεί εκτενής περίληψη της διπλωματικής στο Κεφάλαιο 2. Στο Κεφάλαιο 3 θα περιγράψουμε τι είναι μια χρονοσειρά, τα δομικά της στοιχεία και θα δωθεί ιδιαίτερη προσοχή στην εποχιακή της συμπεριφορά, πώς μπορούμε να την απομονώσουμε ή να την ενσωματώσουμε στο μοντέλο της πρόβλεψης. Η μεθοδολογία που ακολουθείται για τη πρόβλεψη μιας χρονοσειράς, δηλαδή η προετοιμασία της, η προέκτασή της στο μέλλον και τέλος η αξιολόγησή ακρίβειας αναλύονται στο Κεφάλαιο 4. Στο Κεφάλαιο 5 παραθέτουμε αναλυτικά τα βήματα που ακολουθούμε για να αντιμετωπίσουμε το πρόβλημα των εποχιακών μικρών χρονοσειρών. Συγκεκριμένα, πώς εφαρμόζονται αυτά τα βήματα στο υπό εξέταση σύνολο δεδομένων φαίνεται στο Κεφάλαιο 6. Στο Κεφάλαιο 7 συνοψίζουμε τα αποτελέσματα και συζητάμε μελλοντικές επεκτάσεις. Τέλος, στο παράρτημα, ακολουθούν οι γλώσσες προγραμματισμού, οι βιβλιοθήκες και τα περιβάλλοντα ανάπτυξης που χρησιμοποιήθηκαν για την υλοποίηση του πειράματος της διπλωματικής και η παράθεση των λεπτομερών αποτελεσμάτων μέτρησης της ακρίβειας.

Κεφάλαιο 2

Εκτενής Περίληψη

Χρονοσειρά ονομάζουμε ένα σύνολο παρατηρήσεων που περιγράφουν την εξέλιξη της συμπεριφοράς ενός μεγέθους στο χρόνο. Θεωρούμε, μάλιστα, ότι αυτές οι παρατηρήσεις δεν είναι ανεξάρτητες μεταξύ τους χωρίς βέβαια αυτό να υπονοεί μια ντετερμινιστική σύνδεση μεταξύ αυτών.

Οι χρονοσειρές μπορούν να αναλυθούν σε τέσσερις χαρακτηριστικές υποσειρές. Την τάση, που περιγράφει πως μεταβάλλεται χρονικά το επίπεδο της χρονοσειράς. Τη κυκλικότητα που μετράει την κατά περιόδους μεταβολή της χρονοσειράς που οφείλεται σε εξωγενείς συνθήκες. Την εποχιακότητα που αποτελεί το μοτίβο διακύμανσης που παρουσιάζει η χρονοσειρά και επαναλαμβάνεται ανά σταθερές χρονικές περιόδους. Ο αστάθμητος παράγοντας της τύχης μιας χρονοσειράς καλείται τυχαιότητα.

Υπάρχει ένα σύνολο μεθόδων που μας επιτρέπει να αποσυνθέσουμε τη χρονοσειρά στα επιμέρους της στοιχεία. Εν γένει αντιμετωπίζουμε τη χρονοσειρά είτε ως άθροισμα των συνθετικών της μονάδων, είτε ως γινόμενο. Η πρώτη προσέγγιση είναι η προσθετική, ενώ η δεύτερη η πολλαπλασιαστική. Χρησιμοποιώντας το στατιστικό εργαλείο των κινητών μέσων όρων μπορούμε να εφαρμόσουμε τη κλασική μέθοδο αποσύνθεσης είτε με την προσθετική είτε με την πολλαπλασιαστική προσέγγιση. Συμπληρωματικά, για να εξασφαλίσουμε υψηλή ακρίβεια στον υπολογισμό των δεικτών εποχιακότητας μπορούμε να ενσωματώσουμε στη διαδικασία υπολογισμού της εποχιακότητας της χρονοσειράς μία μέθοδο συρρίκνωσης συντελεστών όπως είναι η μέθοδος James-Stein ή η μέθοδος Lemon-Krutchkoff.

Μία άλλη μέθοδος αποσύνθεσης είναι η μέθοδος STL. Αποτελεί ουσιαστικά μία διαδικασία φιλτραρίσματος της χρονοσειράς που μας επιτρέπει να την αποδομήσουμε σε τρία βασικά χαρακτηριστικά: την τάση, την εποχιακότητα και τα εναπομείνοντα στοιχεία. Η διαδικασία βασίζεται στη μέθοδο Loess.

Εκτός των παραπάνω μεθόδων που είναι ανεξάρτητες της διαδικασίας της πρόβλεψης, αν και αποτελεί συνήθως τον τελικό στόχο, υπάρχουν μοντέλα που ενσωματώνουν την εποχιακή ανάλυση. Ένα τέτοιο παράδειγμα είναι η μέθοδος Holt-Winters που στηρίχθηκε στη μέθοδο εκθετικής εξομάλυνσης για χρονοσειρές γραμμικής τάσης και είναι ικανή να διαχειριστεί και την εποχιακότητα. Επίσης, τα ολοκληρωμένα αυτοπαλινδρομικά μοντέλα κινητών μέσω όρων (ARIMA) δύνανται να μοντελοποιήσουν ένα μεγάλο εύρος εποχιακών δεδομένων.

Γενικά η διαδικασία της πρόβλεψης είναι μια πολυβηματική διαδικασία που αποσκοπεί στη χρήση της γνώσης των παρελθοντικών παρατηρήσεων μιας χρονοσειράς για να εκτιμήσει πως αυτή θα εξελιχθεί στο μέλλον.

Αρχικά, πρέπει η χρονοσειρά να έρθει σε κατάλληλη μορφή για να μπορέσουμε να εφαρμόσουμε τις στατιστικές μεθόδους που την προεκτείνουν στο μέλλον. Έτσι, το πρώτο βήμα είναι να αναπαραστήσουμε γραφικά τα δεδομένα έτσι ώστε να αποκτήσουμε μια εποπτεία στα ποιοτικά της χαρακτηριστικά. Βάσει αυτών και της γνώσης που έχουμε εν γένει για τα δεδομένα, καλούμαστε στη συνέχεια να διαχειριστούμε τις ιδιομορφίες της χρονοσειράς. Αυτές μπορεί να είναι μη ιδανική δειγματοληψία, κενές τιμές ή μηδενικές. Επίσης, πολλές φορές η χρονοσειρά χρήζει ημερολογιακών προσαρμογών αφότου οι παρατηρήσεις της μπορούν να επηρεάζονται από τις ημέρες ανθρώπινης εργασίας και συνεπώς από τα σαββατοκύριακα ή και τις αργίες.

Αφότου τα δεδομένα έχουν έρθει σε κατάλληλη μορφή και μπορούμε να τα αποσυνθέσουμε όπως περιγράψαμε προηγουμένως, χρειάζεται να τα προεκτείνουμε στο μέλλον. Μία πρώτη προσέγγιση είναι να θεωρήσουμε ότι κάθε μελλοντική στιγμή είναι ταυτόσημη με αυτή που προηγείται. Αυτή προσέγγιση περιγράφει τη μέθοδο Naive που συνήθως χρησιμοποιείται ως βάση σύγκρισης.

Η ανάλυση της παλινδρόμησης θεωρώντας τον χρόνο ανεξάρτητη μεταβλητή, αποτελεί μία άλλη μέθοδο προέκτασης της χρονοσειράς. Εν γένει αποσκοπεί στην εύρεση συσχετίσεων μεταξύ μιας εξαρτημένης μεταβλητής και μίας ή περισσότερων ανεξάρτητων μεταβλητών και γι΄ αυτό το λόγο εκτός από μοντέλο πρόβλεψης αυτή καθ΄ αυτή μπορεί να χρησιμοποιηθεί και ως υποβοήθημα για άλλες μεθόδους.

Τα μοντέλα εκθετικής εξομάλυνσης είναι απλά μοντέλα, εύκολα στη χρήση, με μικρές υπολογιστικές απαιτήσεις που έχουν την δυνατότητα να παράξουν ακριβείς προβλέψεις ακόμα και με σχετικά μικρό ιστορικό παρατηρήσεων. Τα συγκεκριμένα μοντέλα εξαρτώνται από τη μορφή της τάσης (Σταθερού επιπέδου, Γραμμικής, Εκθετικής ή Φθίνουσας τάσης) και από το πρότυπο εποχιακότητας (Χωρίς, Προσθετική, Πολλαπλασιαστική).

Η μέθοδος Θ βασίζεται στη μεταβολή των τοπικών καμπυλοτήτων μιας χρονοσειράς για να παράξει προβλέψεις. Η χρονοσειρά αναλύεται σε δύο ή περισσότερες γραμμές Theta και κάθε μία από αυτές προβλέπεται ξεχωριστά, είτε με το ίδιο είτε με διαφορετικό μοντέλο πρόβλεψης.

Αφότου έχουμε ολοκληρώσει τη διαδικασία της πρόβλεψης πρέπει να αξιολογήσουμε κατά πόσο το μοντέλο μας παρήγαγε ακριβείς προβλέψεις ή ποιο από τα μοντέλα που εφαρμόσαμε είναι καλύτερο. Για να το πετύχουμε αυτό χρησιμοποιούμε ένα σύνολο στατιστικών δεικτών αξιολόγησης της ακρίβειας. Αυτοί οι δείκτες μπορεί να εξαρτώνται από τη κλίμακα των δεδομένων ή να εκφράζονται από ποσοστιαία σφάλματα. Μπορούν να αντικατοπτρίζουν πραγματικά ή σχετικά μεγέθη, ή να είναι αποτέλεσμα κανονικοποίησης.

Στη παραπάνω διαδικασία, όμως, έχουμε πρόβλημα στη περίπτωση που η χρονοσειρά μας δεν διαθέτει αρκετά ιστορικά δεδομένα και συνεπώς δε μπορούμε να εξάγουμε το στοιχείο της εποχιακότητας. Η συνηθισμένη αντιμετώπιση είναι να θεωρήσουμε τη χρονοσειρά ως μη εποχιακή και να την προβλέψουμε ως τέτοια. Αλλά, έτσι ουσιαστικά δεν έχουμε απαλλάξει τη χρονοσειρά από τις εποχιακές της διακυμάνσεις και αυτές θα επηρεάσουν την ακρίβεια της

πρόβλεψής μας.

Ένας τρόπος να αντιμετωπίσουμε αυτό το ζήτημα είναι αν έχουμε ένα σύνολο χρονοσειρών που περιγράφουν συναφή μεγέθη. Έτσι, εκμαιεύουμε την πληροφορία για την εποχικότητα παρόμοιων χρονοσειρών και την χρησιμοποιούμε για να προβλέψουμε τις μικρότερες χρονοσειρές που φαίνεται να έχουν αντίστοιχη εποχιακή συμπεριφορά.

Για να το καταφέρουμε αυτό χρησιμοποιήθηκε η ακόλουθη μεθοδολογία. Αρχικά, φέρνουμε τα δεδομένα σε κατάλληλη μορφή για τις στατιστικές μεθόδους που θέλουμε να εφαρμόσουμε με τις συνηθισμένες τεχνικές διαχείρισης των ιδιομορφιών των χρονοσειρών και με χρήση της πληροφορίας που έχουμε για τα δεδομένα. Κατόπιν, χωρίζουμε τις χρονοσειρές σε δύο ομάδες: αυτές με επαρκή ιστορικά δεδομένα για να εξάγουμε την εποχιακή τους συμπεριφορά και τις υπόλοιπες.

Στη πρώτη κατηγορία, βρίσκουμε τους δείκτες εποχιακότητας με το κλασικό πολλαπλασιαστικό μοντέλο αποσύνθεσης. Στη συνέχεια, εφαρμόζουμε τον αλγόριθμο συσταδοποίησης DBSCAN στους υπολογισμένους δείκτες για να εντοπίσουμε αν υπάρχουν ομάδες με συνάφεια μεταξύ τους και να τις εντοπίσουμε. Για κάθε συστάδα υπολογίζουμε τους μέσους δείκτες εποχιακότητας.

Στη δεύτερη κατηγορία, υπολογίζουμε μία ψευδο-εποχιακότητα που θα χρησιμοποιήσουμε σαν κριτήριο συνάφειας για τις συστάδες που προέκυψαν προηγουμένως. Αν, λοιπόν, ένα διάνυσμα δεικτών έχει μικρότερη μέση τετραγωνική απόσταση από το μέσο όρο των εποχιακών δεικτών μίας συστάδας από την μέγιστη απόσταση που υπολογίζουμε από τις χρονοσειρές τις συστάδας και του μέσου όρου τους, τότε κατατάσσουμε την μικρή χρονοσειρά στην συστάδα.

Για τις μικρές χρονοσειρές που βρέθηκε να παρουσιάζουν κοντινή εποχιακή συμπεριφορά με κάποια συστάδα, χρησιμοποιούμε τους μέσους εποχιακούς δείκτες της συστάδας που ανήκουν ως την χαρακτηριστική τους εποχιακότητα. Κατόπιν, προβλέπουμε βάσει αυτής. Συγχρόνως, κάνουμε πρόβλεψη στα αρχικά δεδομένα και τελικώς συγκρίνουμε την ακρίβεια των δύο προσεγγίσεων σύμφωνα με το κανονικοποιημένο δείκτη μέσου απόλυτου σφάλματος.

Τα δεδομένα που έχουμε στη διάθεσή μας περιγράφουν τη στάθμη υγρού φυσικού αερίου σε ένα σύνολο δεξαμενών στη Γαλλία. Ο αρχικός σκοπός της ανάλυσης ήταν να εκτιμήσουμε πότε οι δεξαμενές αυτές θα αδειάσουν, έτσι ώστε να μπορεί η εταιρία διανομής να το αποτρέψει αλλά και να σχεδιάσει βέλτιστα τον ανεφοδιασμό τους.

Για τη προετοιμασία τους ακολουθήσαμε τις εξής ενέργειες. Μετατρέψαμε τις αρνητικές και μηδενικές τιμές σε κενές τιμές. Μετατρέψαμε τα δεδομένα σε ημερήσια κρατώντας, στις περιπτώσεις πολλαπλών παρατηρήσεων εντός μίας ημέρας, τη μικρότερη από αυτές. Μετά συμπληρώσαμε τις κενές τιμές με τη μέθοδο της γραμμικής παρεμβολής βάσει του χρόνου.

Λόγω του ότι θέλουμε να προβλέψουμε πότε θα αδειάσει η κάθε δεξαμενή χρειαζόμαστε την χρονοσειρά ζήτησης. Γι΄ αυτό εφαρμόζοντας πρώτες διαφορές λαμβάνουμε την ημερήσια ζήτηση. Αφαιρέσαμε τους ανεφοδιασμούς και τους διαχειριστήκαμε ως κενές τιμές. Τέλος, έχοντας ως στόχο την μεσοπρόθεσμη ζήτηση, μετατρέψαμε τα δεδομένα μας σε μηνιαία, υπολογίζοντας τον μέσο όρο της ημερήσιας ζήτησης κάθε μήνα.

Αφότου εφαρμόσαμε την αποεποχικοποίηση όπως περιγράφηκε προηγουμένως, εφαρμόσαμε τον αλγόριθμο DBSCAN που μας επέστρεψε μία συστάδα. Υπολογίσαμε τις μικρές χρο-

νοσειρές που ανήκουν σε αυτή και εφαρμόσαμε τις δύο προσεγγίσεις για τη παραγωγή προβλέψεων. Για κάθε μία προσέγγιση, χρησιμοποιήσαμε έξι μοντέλα πρόβλεψης

Βρήκαμε ότι πράγματι βελτιώθηκε σημαντικά η ακρίβεια των προβλέψεών μας. Όλα τα μοντέλα πρόβλεψης είχαν μεγαλύτερη ακρίβεια, ενώ στο μεγαλύτερο ποσοστό των χρονοσειρών έδειξαν καλύτερα αποτελέσματα με την προτεινόμενη προσέγγιση.

Κεφάλαιο 3

Χρονοσειρές με εποχιακή συμπεριφορά

3.1 Εισαγωγή

3.1.1 Γενικά για τις Χρονοσειρές

Ως χρονοσειρά μπορούμε να ορίσουμε ένα σύνολο τιμών-παρατηρήσεων που περιγράφουν την εξέλιξη της συμπεριφοράς ενός μεγέθους στο πεδίου του χρόνου. Τα μεγέθη που περιγράφονται μπορούν να είναι οποιασδήποτε φύσης αρκεί να μπορούν να ποσοτικοποιηθούν. Έτσι, συναντάμε χρονοσειρές που περιγράφουν φυσικά μεγέθη όπως είναι ο όγκος νερού βροχόπτωσης για μια περιοχή ανά ημέρα αλλά και οικονομικά όπως είναι οι πωλήσεις ενός προϊόντος συναρτήσει του χρόνου.

Θεωρείται ότι οι διαδοχικές τιμές μιας χρονοσειράς δεν είναι ανεξάρτητες μεταξύ τους. Ωστόσο, αυτό δεν συνεπάγεται ότι πρόκειται για μια ντετερμινιστική διαδικασία που μας επιτρέπει να καθορίσουμε επακριβώς τις μελλοντικές τιμές της από τις προηγούμενες. Αντίθετα, οι χρονοσειρές αποτελούν στοχαστικές διαδικασίες καθότι η εξέλιξη των τιμών τους ενέχουν τυχαιότητα, δεδομένου ότι περιγράφουν την εξέλιξη ενός μεγέθους στον πραγματικό κόσμο.

3.1.2 Συνθετικά στοιχεία μια χρονοσειράς

Σύμφωνα με τις παραδοσιακές μεθόδους ανάλυσης, μια χρονοσειρά αναλύεται σε τέσσερα δομικά χαρακτηριστικά: την τάση, τη κυκλικότητα, την εποχικότητα και την τυχαιότητα.

Η **τάση** μιας χρονοσειράς περιγράφει πως μακροπρόθεσμα μεταβάλλεται το μέσο επίπεδο των τιμών μιας χρονοσειράς. Η τάση μπορεί να είναι ανοδική, φθίνουσα ή μηδενική.

Η **κυκλικότητα** περιγράφει την κατά περιόδους μεταβολή της χρονοσειράς εξαιτίας ειδικών εξωγενών συνθηκών. Συναντάμε συχνά μεταβολές στον κυκλικό παράγοντα της σειράς που οφείλονται σε οικονομικές συνθήκες, ενώ χαρακτηριστικά παραδείγματα βρίσκουμε σε χρονοσειρές που περιγράφουν δείκτες παραγωγής, μετοχών, ΑΕΠ και οικονομικών μεγεθών εν γένει.

Η **εποχιακότητα** ορίζεται ως ένα μοτίβο διακύμανσης που παρουσιάζει η χρονοσειρά

και επαναλαμβάνεται ανά τακτές χρονικές περιόδους, συνήθως μικρότερες του έτους. Μερικά παραδείγματα χρονοσειρών που χαρακτηρίζονται από έντονο στοιχείο εποχιακότητας είναι χρονοσειρές θερμοκρασιών, κατανάλωσης πετρελαίου θέρμανσης και γενικά εποχιακών προϊόντων.

Η **τυχαιότητα** αποτελεί τον αστάθμητο παράγοντα της τύχης κατά την εξέλιξη μιας χρονοσειράς.

3.2 Προσθετικό και Πολλαπλασιαστικό μοντέλο αποσύνθεσης

Για να αναλύσουμε μια χρονοσειρά στα επιμέρους στοιχεία της μπορούμε να χρησιμοποιήσουμε απλές μαθηματικές σχέσεις. Αρχικά ορίζουμε την χρονοσειρά ως μία συνάρτηση των δομικών χαρακτηριστικών της:

$$Y_t = f(S_t, T_t, C_t, R_t)$$

Με: Y_t να είναι η χρονοσειρά
S_t η συνιστώσα της εποχιακότητας
T_t η συνιστώσα της τάσης
C_t η συνιστώσα της κυκλικότητας
R_t η συνιστώσα της τυξαιότητας.
Δύο απλές μορφές της συνάρτησης f είναι η προσθετική:

$$Y_t = S_t + T_t + C_t + R_t$$

και η πολλαπλασιαστική:

$$Y_t = S_t * T_t * C_t * R_t$$

Αξίζει να σημειώσουμε ότι τα δύο μοντέλα έχουν μία λογαριθμική σχέση μεταξύ τους. Δηλαδή, αν πάρουμε τον λογάριθμο της προσθετικής σχέσης προκύπτει η πολλαπλασιαστική.

3.2.1 Κλασική Μέθοδος Αποσύνθεσης με Κινητούς Μέσους Όρους

Η κλασική μέθοδος αποσύνθεσης αποτελεί μια εύκολη διαδικασία για να διασπάσουμε τη χρονοσειρά στα τέσσερα δομικά της στοιχεία. Συχνά, τη συναντάμε στη βιβλιογραφία με την επωνυμία $X-11$ και αποτελείται από πέντε βασικά βήματα. Μάλιστα, όπως θα δούμε μπορεί να χρησιμοποιηθεί τόσο δεδομένης πολλαπλασιαστικής σχέσης όσο και προσθετικής μεταξύ των συστατικών στοιχείων της χρονοσειράς.

Βήμα 1ο

Αρχικά, απομονώνουμε την τάση και τον κύκλο της χρονοσειράς. Αυτό το καταφέρνουμε με το να υπολογίσουμε τον κινητό μέσο όρο της χρονοσειράς στο μήκος της εποχιακότητας. Το

αποτέλεσμα δεν μεταφέρει την εποχιακή συμπεριφορά της αρχικής χρονοσειράς, ενώ παράλληλα εξαλείφεται σχεδόν και η τυχαιότητα, αφού οι τυχαίες διακυμάνσεις της χρονοσειράς χάνονται παίρνοντας τον μέσο όρο. Η σειρά που προκύπτει, λοιπόν, θεωρούμε ότι είναι η σειρά τάσης κύκλου και περιγράφεται από τη παρακάτω σχέση για το προσθετικό μοντέλο:

$$KMO(n) = T + C$$

και την αντίστοιχη, για το πολλαπλασιαστικό:

$$KMO(n) = T * C$$

όπου και στις δύο περιπτώσεις το $KMO(n)$ είναι ένας κινητός μέσος όρος μήκους n και τα T και C όπως τα ορίσαμε προηγουμένως. Πρέπει να σημειωθεί ότι εν γένει προτιμάται η χρήση του κεντρικού κινητού μέσου όρου μήκους αντίστοιχου της εποχιακότητας, όταν αυτή είναι άρτια σε μία χρονοσειρά, αντί του απλού κινητού μέσου όρου.

Βήμα 2ο

Είναι χρήσιμο σε αυτό το σημείο να παραγάγουμε τη χρονοσειρά των υπόλοιπων στοιχείων της αρχικής χρονοσειράς. Στο προσθετικό μοντέλο αυτό γίνεται με αφαίρεση της σειράς τάσης-κύκλου από τα αρχικά δεδομένα και με διαίρεση στο πολλαπλασιαστικό. Οπότε έχουμε αντίστοιχα τις παρακάτω σχέσεις:

$$S + R = Y - (T + C)$$

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

Βήμα 3ο

Σε αυτό το βήμα αφαιρούμε την τυχαιότητα από τους λόγους εποχιακότητας που προέκυψαν στο 2ο Βήμα. Βρίσκοντας, λοιπόν τη μέση τιμή των λόγων που αναφέρονται σε αντίστοιχες περιόδους του εποχιακού κύκλου, δημιουργούμε τους δείκτες εποχιακότητας της αρχικής μας χρονοσειράς. Πρέπει, το άθροισμα τον εν λόγω λόγων να ισούται με το μήκος του εποχιακού μας κύκλου. Λόγου χάρη, σε περίπτωση μηνιαίων δεδομένων αυτό αντιστοιχεί στο ένα έτος, δηλαδή 12. Αν αυτό δε συμβαίνει, θα χρειαστεί να κανονικοποιήσουμε τα δεδομένα.

Στη πράξη, πολλές φορές, χρειάζεται να διαχειριστούμε χρονοσειρές που παρουσιάζουν μεγάλη τυχαιότητα και ασυνήθιστες τιμές. Σε αυτές τις περιπτώσεις, μπορούμε να παραλείψουμε τη μέγιστη και την ελάχιστη τιμή στον υπολογισμό του μέσου όρου έτσι ώστε να επιτευχθεί μια σταθεροποίηση του αποτελέσματος.

Βήμα 4ο

Για να προσδιορίσουμε την αποεποχικοποιημένη σειρά εργαζόμαστε ως εξής:
Για το προσθετικό μοντέλο:

$$Y - S = T \times C \times S \times R - S$$

και την αντίστοιχη, για το πολλαπλασιαστικό:

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

Δηλαδή, στη πρώτη περίπτωση, αφαιρούμε τους δείκτες από τις αρχικές τιμές, ενώ στη δεύτερη τους διαιρούμε.

Βήμα 5ο

Για να αφαιρέσουμε την τυχαιότητα από το αποτέλεσμα του προηγούμενου βήματος χρησιμοποιούμε τον κινητό μέσο όρο μήκους 3 ή 6, ή διπλού κινητού μέσου όρου 3×3 σε αυτό. Η προκύπτουσα χρονοσειρά είναι μία ομαλή και ακριβής σειρά τάσης-κύκλου. Μπορούμε εύκολα να πάρουμε την σειρά της τυχαιότητας αντίστοιχα με προηγουμένως.

Για το προσθετικό μοντέλο:

$$T \times C \times R - KMO(3 \times 3) = T \times C \times R - T \times C = R$$

και την αντίστοιχη, για το πολλαπλασιαστικό:

$$\frac{T \times C \times R}{KMO(3 \times 3)} = \frac{T \times C \times R}{T \times C} = R$$

Βήμα 6ο

Έχουμε, λοιπόν, καταφέρει ως τώρα να λάβουμε μια χρονοσειρά που περιέχει τάση και κυκλικότητα. Υπάρχουν περιπτώσεις που έχει κάποιο νόημα να διαχωρίσουμε ατά τα στοιχεία. Τότε πρέπει να επιλέξουμε κατάλληλο μοντέλο τάσης και να το απαλείψουμε από τη χρονοσειρά $T \times C$. Έτσι, έστω ότι η χρονοσειρά μας παρουσιάζει γραμμική τάση, εφαρμόζουμε το μοντέλο της απλής γραμμικής παλινδρόμησης. Η προκύπτουσα ευθεία περιγράφει την τάση, ενώ αν την αφαιρέσουμε από την $T \times C$ λαμβάνουμε τον κύκλο. Προφανώς, το παραπάνω ισχύει για το προσθετικό μοντέλο ενώ στο πολλαπλασιαστικό μοντέλο παίρνουμε τον κύκλο με διαίρεση των σειρών.

3.2.2 Διαφορές των δύο μοντέλων και κατάλληλη επιλογή

Στο προσθετικό μοντέλο αποσύνθεσης οι μεταβολές που οφείλονται σε κάθε δομικό χαρακτηριστικό της χρονοσειράς εφαρμόζονται ανεξάρτητα για να τη συνθέσουν. Αντίθετα, στο πολλαπλασιαστικό μοντέλο τα στοιχεία της σειράς συσχετίζονται μεταξύ τους. Ως αποτέλεσμα, η εποχιακότητα μιας χρονοσειράς που παρουσιάζει έντονο το στοιχείο της τάσης θα έχει αντίστοιχα μεγαλύτερη ένταση σε περίπτωση που χρησιμοποιήσουμε το πολλαπλασιαστικό μοντέλο, καθώς έχουμε αύξηση της εποχιακής διακύμανσης όσο το επίπεδο της χρονοσειράς αλλάζει στον χρόνο. Χρησιμοποιώντας το προσθετικό μοντέλο έχουμε προσθήκη της ίδιας εποχιακής διακύμανσης ανεξάρτητα από το επίπεδο που θα μας οδηγήσει σε ομοιόμορφη εφαρμογή της εποχιακότητας.

Μπορούμε να χρησιμοποιήσουμε την εξής μέθοδο για να αποφασίσουμε ποια προσέγγιση στην αποσύνθεση ταιριάζει περισσότερο στη χρονοσειρά μας: Αφαιρούμε από την αρχική

χρονοσειρά το στοιχείο της εποχιακότητας με χρήση του πολλαπλασιαστικού μοντέλου και προσαρμόζουμε στη προκύπτουσα χρονοσειρά την ευθεία της απλής γραμμικής παλινδρόμησης. Στη περίπτωση που εντοπίσουμε στατιστική σημαντικότητα στη κλίση παραμέτρου της ευθείας του μοντέλου παλινδρόμησης, θεωρούμε ότι η επιλογή μας ήταν σωστή, ειδάλλως εφαρμόζουμε το προσθετικό μοντέλο.

3.2.3 Μέθοδοι συρρίκνωσης συντελεστών

Είναι μείζονος σημασίας να έχουμε υψηλή ακρίβεια στον υπολογισμό των δεικτών εποχιακότητας, έτσι ώστε να καταφέρουμε να παραγάγουμε και ακριβείς προβλέψεις. Το τελευταίο μπορεί να συμβεί μόνο αν εφαρμόσουμε το μοντέλο πρόβλεψής μας σε μία αποεποχικοποιημένη χρονοσειρά που είναι επαρκώς εξομαλυμένη. Για να το διασφαλίσουμε αυτό χρησιμοποιούμε μεθόδους συρρίκνωσης συντελεστών στους δείκτες εποχιακότητας της χρονοσειράς.

Μέθοδος Συρρίκνωσης James-Stein

Η βασική λειτουργία αυτής της μεθόδου περιγράφεται από την παρακάτω εξίσωση:

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

Το S δηλώνει τις τιμές των δεικτών εποχιακότητας και $S^J S_j$ είναι οι δείκτες της μεθόδου James-Stein. Για τον υπολογισμό του συντελεστή W χρησιμοποιούμε τον παρακάτω τύπο:

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

Όπου το μήκος του κύκλου εποχιακότητας δηλώνεται με $pos(periodsofseasonality)$. Επίσης τα:

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

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

είναι οι διαφορές λόγω δειγματοληπτικών σφαλμάτων. Το S_{jk} δηλώνει τον λόγο της εποχιακότητας για την εποχή j και κατά τον εποχιακό κύκλο k και το K_j δηλώνει το πλήθος των λόγων εποχιακότητας που έχουν υπολογιστή για την εν λόγω εποχή. Στη περίπτωση που βρούμε το A να είναι αρνητικό, το θέτουμε ίσο με το μηδέν.

Μέθοδος Συρρίκνωσης Lemon-Krutchkoff

Ο εποχιακός δείκτης Lemon-Krutchkoff βρίσκεται ως εξής:

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

Το $W_{j^*,j}$ υποδηλώνει τα βάρη των σχετικών πιθανοτήτων $L_{j^*,j}$ να παρατηρηθεί ο εκτιμητής S_{j^*}, δεδομένου S_j που είναι ο πραγματικός παράγοντας. Έχουμε:

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

Θεωρούμε ότι η πιθανότητα $L_{j^*,j}$ ακολουθεί κανονική κατανομή με διακύμανση ίση με $\sigma = \sqrt{V}$.

3.3 STL : Μία μέθοδος αποσύνθεσης Εποχιακότητας/Τάσης βασισμένη στη μέθοδο Loess

Η STL είναι μια διαδικασία φιλτραρίσματος που μας επιτρέπει να αποδομήσουμε μια χρονοσειρά σε τρία στοιχεία: την τάση, την εποχιακότητα και τα εναπομείναντα στοιχεία. Εκφραζόμενα σε μία μαθηματική σχέση που η χρονοσειρά συμβολίζεται με Ψ και τα χαρακτηριστικά της με T, S και R, αντίστοιχα.

$$Y = T + S + R$$

Οι σχεδιαστές της μεθόδου προσπάθησαν να ικανοποιήσουν τα παρακάτω κριτήρια:

1. Η μέθοδος STL να χαρακτηρίζεται από απλό σχεδιασμό και εύκολη χρήση.

2. Ύπαρξη ελαστικότητας κατά τον προσδιορισμό της ποσότητας διακύμανσης στα στοιχεία της τάσης και της εποχιακότητας.

3. Προσδιορισμός του αριθμού παρατηρήσεων ανά κύκλο του στοιχείου εποχιακότητας σε οποιονδήποτε ακέραιο μεγαλύτερο του 1.

4. Δυνατότητα αποσύνθεσης χρονοσειρών με κενές τιμές.

5. Ισχυρή τάση και εποχιακότητα που δεν αλλοιώνεται από τη μεταβατική, ανώμαλη συμπεριφορά που μπορεί να χαρακτηρίζει τα δεδομένα

6. Εύκολη υλοποίηση σε υπολογιστικό περιβάλλον και ταχύ υπολογισμό της, ακόμα και για μεγάλες χρονοσειρές.

3.3.1 Ο ορισμός της μεθόδου STL

Loess

Έστω ότι x_i και y_i, με $i = 1, \ldots, n$ είναι οι τιμές μίας ανεξάρτητης και μίας εξαρτημένης μεταβλητής, αντίστοιχα. Η καμπύλης παλινδρόμησης Loess, $g(x)$ είναι μία εξομάλυνση της y δεδομένης της x που μπορεί να υπολογιστεί για οποιαδήποτε τιμή της x στο πεδίο ορισμού της εξαρτημένης μεταβλητής.

Η $g(x)$ υπολογίζεται ως εξής. Επιλέγουμε έναν θετικό ακέραιο q, έστω $q \leq n$. Οι τιμές του q που είναι εγγύτερες στο x επιλέγονται και δίνεται στη κάθε μία ένα *βάρος γειτνίασης* βασισμένο στην απόσταση του από το x. Έστω $\lambda_q(x)$ η απόσταση του q-οστού πιο απομακρυσμένου x_i από το x. Έστω, τώρα, ότι W είναι η τρικυβική συνάρτηση βάρους:

$$W(u) = \begin{cases} (1-u^3)^3 & \text{αν } 0 \leq u < 1 \\ 0 & \text{αν } u \geq 1 \end{cases} \tag{3.1}$$

Το βάρος γειτνίασης για οποιοδήποτε x_i είναι:

$$v_i(x) = W(\frac{|x_i - x|}{\lambda_q(x)})$$

Συνεπώς όσο πιο κοντά στο x είμαστε τόσο μεγαλύτερο είναι το βάρος. Μάλιστα, αυτό μηδενίζεται για το χ-οστό σημείο σε απόσταση. Στο επόμενο βήμα προσαρμόζουμε μια πολυωνυμική καμπύλη βαθμού d στα δεδομένα με βάρος $v_i(x)$ στο σημείο (x_i, y_i). Η τιμή της τοπικά προσαρμοσμένης πολυωνυμικής καμπύλης στο ξ είναι η $g(x)$.

Τώρα σε περίπτωση που το $q > n$, η $\lambda_n(x)$ είναι η απόσταση από το ξ στο x_i που βρίσκεται πιο μακρυά από αυτό και ορίζουμε το $\lambda_q(x)$ ως:

$$\lambda_q(x) = \lambda_n(x)\frac{q}{n}$$

και αντίστοιχα ορίζονται τα βάρη γειτνίασης.

Για να χρησιμοποιήσουμε τη μέθοδο Loess πρέπει να επιλεχθούν κατάλληλα d και q. Η μέθοδος επιλογής τους ξεφεύγει από τα πλαίσια αυτής της διπλωματικής.

Αξίζει να αναφερθεί βέβαια, ότι όσο το q αυξάνεται η $g(x)$ γίνεται πιο ομαλή, όταν προσεγγίζει το άπειρο η $v_i(x)$ συγκλίνει στο ένα και η $g(x)$ προσεγγίζει το πολυώνυμο ελαχίστων τετραγώνων βαθμού d. Στη περίπτωση που η χρονοσειρά έχει ήπια καμπυλότητα είναι βάσιμο να πάρουμε τη διάσταση d ίση με ένα, ενώ όταν παρουσιάζει έντονη καμπυλότητα, με κορυφές και κοιλάδες, η διάσταση βαθμού δύο αποτελεί καλύτερη επιλογή.

Έστω τώρα ότι κάθε παρατήρηση $(x_i(x), y_i(x))$ έχει ένα βάρος ρ_i που εκφράζει την αξιοπιστία της παρατήρησης σε σχέση με τις υπόλοιπες. Ενσωματώνοντας αυτά τα βάρη στη διαδικασία εξομάλυνσης βελτιώνεται η στιβαρότητα(robustness) της μεθόδου STL.

Στοιχεία σχεδίασης: εσωτερικός και εξωτερικός βρόχος

Η μέθοδος STL εμπεριέχει δύο αναδρομικές διαδικασίες: έναν εσωτερικό βρόχο εμφωλευμένο μέσα σε έναν εξωτερικό. Σε κάθε πέρασμα του εσωτερικού βρόχου, τα στοιχεία εποχιακότητας και τάσης ανανεώνονται. Κάθε πέρασμα του εξωτερικού βρόχου εμπεριέχει την εκτέλεση του εσωτερικού κόμβου και ακολουθεί με τον υπολογισμό βαρών στιβαρότητας, τα οποία χρησιμοποιούνται στην επόμενη εκτέλεση του εσωτερικού κόμβου για να ελαφρύνουν την επιρροή όποιας μεταβατικής, ανώμαλης συμπεριφοράς στα στοιχεία τάσης και εποχιακότητας.

Ας θεωρήσουμε, τώρα, τις χρονοσειρές που συνθέτουν οι παρατηρήσεις κάθε περιόδου, ή κύκλου, της εποχιακότητας, έτσι ώστε να έχουμε τόσες στο πλήθος όσο το μήκος της

εποχιακότητας. Για παράδειγμα, ο κύκλος της εποχιακότητας είναι το έτος και έχουμε μηνιαία δεδομένα, παίρνουμε 12 τέτοιες χρονοσειρές, τη χρονοσειρά το Ιανουαρίου, του Μαρτίου και ούτω καθεξής. Ονομάζουμε αυτές τις χρονοσειρές *υποσειρές-κύκλου*.

Ο εσωτερικός βρόχος

Κάθε πέρασμα του εσωτερικού βρόχου περιλαμβάνει μία εποχιακή εξομάλυνση, που ανανεώνει το δομικό στοιχείο της εποχιακότητας, ακολουθούμενη από την εξομάλυνση της τάσης, που ανανεώνει το συστατικό της τάσης αντίστοιχα. Τα βήματα που περιγράφουν την διαδικασία είναι τα εξής:

1. **Βήμα 1:** Αφαίρεση τάσης
2. **Βήμα 2:** Εξομάλυνση υποσειράς-κύκλου
3. **Βήμα 3:** Βαθυπερατό φιλτράρισμα της εξομαλυσμένης υποσειράς-κύκλου
4. **Βήμα 4:** Αφαίρεση τάσης από την εξομαλυσμένη υποσειρά-κύκλου
5. **Βήμα 5:** Αποεποχικοποίηση
6. **Βήμα 6:** Εξομάλυνση τάσης

Έτσι, στα βήματα 2,3 και 4 συναντάμε το κομμάτι εξομάλυνσης εποχιακότητας, ενώ στο βήμα 6 το κομμάτι εξομάλυνσης της τάσης.

Ο εξωτερικός βρόχος

Έχοντας διατρέξει τον εσωτερικό βρόχο έχουμε λάβει προσεγγίσεις για στοιχεία τάσης και εποχιακότητας, τα T_v και S_v αντίστοιχα. Τότε το εναπομείναν στοιχείο βρίσκεται από τη σχέση:

$$R_v = Y_v - T_v - S_v$$

Ορίζουμε για κάθε παρατήρηση Y_v τα βάρη στιβαρότητας, που μας δείχνουν πόσο ακραία είναι η χρονοσειρά του εναπομείναντος στοιχείου, ως

$$\rho_v = B(\frac{|R_v|}{h})$$

όπου το h ορίζεται ακολούθως

$$h = 6 * median(|R_v)$$

και το B είναι η διτετραγωνική συνάρτηση βάρους:

$$B(u) = \begin{cases} (1-u^2)^2 & \text{αν } 0 \leq u < 1 \\ 0 & \text{αν } u > 1 \end{cases} \tag{3.2}$$

Σε αυτό το σημείο επαναλαμβάνεται ο εσωτερικός βρόχος, αλλά στα βήματα 2 και 6 χρησιμοποιείται το βάρος εποχιακότητας.

Μετα-εξομάλυνση της εποχιακότητας

Η παραπάνω εξομάλυνση δεν μας εγγυάται ότι η μετάβαση από ένα σημείο στο επόμενο στην υπολογισμένη εποχιακότητα θα είναι ομαλή. Υπάρχουν περιπτώσεις, όμως, που θέλουμε το στοιχείο εποχιακότητας της χρονοσειράς μας να χαρακτηρίζεται από ομαλές μεταβάσεις. Μία απλή αντιμετώπιση στο παραπάνω ζήτημα είναι, επιπρόσθετα, να εξομαλύνουμε το στοιχείο της εποχιακότητας με τη μέθοδο Loess. Οι εξομαλυμένες, πλέον, τιμές αποτελούν τη τελική μορφή της εποχιακότητας.

3.3.2 Σχόλια επί της μεθόδου

Η μέθοδος STL είναι μία πολύπλευρη και στιβαρή μέθοδος αποσύνθεσης. Μερικά πλεονεκτήματα που παρουσιάζει σε σχέση με άλλες μεθόδους αποσύνθεσης είναι:

- Μπορεί να διαχειριστεί οποιαδήποτε τύπο εποχιακότητας.
- Το στοιχείο εποχιακότητας δύναται να αλλάζει όσο προχωράει ο παράγοντας του χρόνου και ο ρυθμός μεταβολής μπορεί να καθοριστεί από το χρήστη.
- Η ομαλότητα της σειράς τάσης-κύκλου μπορεί να καθοριστεί επίσης από τον χρήστη.
- Παρουσιάζει στιβαρότητα σε ακραίες τιμές με αποτέλεσμα περιστασιακές ιδιαίτερες παρατηρήσεις δεν επηρεάζουν την εκτίμηση της τάσης-κύκλου και της εποχιακότητας.

Από την άλλη η μέθοδος δε μπορεί από μόνη της να διαχειριστεί μέρες διαπραγματεύσεων ή ημερολογιακές παραλλαγές.

Επίσης, αν και ο ορισμός που δώσαμε περιγράφει μία προσθετική αντιμετώπιση της αποσύνθεσης, η μέθοδος μπορεί εύκολα να τροποποιηθεί για πολλαπλασιαστικές χρονοσειρές.

3.4 Μέθοδοι Πρόβλεψης με ενσωματωμένη την αποεποχικοποίηση

Ενώ οι μέθοδοι που είδαμε ως τώρα είναι ανεξάρτητοι της διαδικασίας πρόβλεψης της χρονοσειράς, συνήθως τις χρησιμοποιούμε για να προετοιμάσουμε τη χρονοσειρά μας για το τελικό μας στόχο, τη πρόβλεψη. Αναλυτικά, διάφορες μέθοδοι πρόβλεψης χρονοσειρών θα μελετηθούν στο επόμενο κεφάλαιο, αλλά προς το παρόν θα εξετάσουμε δύο μεθόδους που έχουν ενσωματωμένη τη διαχείριση της εποχιακής συμπεριφοράς της χρονοσειράς.

3.4.1 Η εποχιακή μέθοδος Holt-Winters

Η μέθοδος Holt αποτελεί μία μέθοδο πρόβλεψης εκθετικής εξομάλυνσης που έχει χτιστεί ώστε να διαχειρίζεται χρονοσειρές που παρουσιάζουν γραμμική τάση. Στηριζόμενοι πάνω στην κλασική μέθοδο οι Holt και Winters την επέκτειναν έτσι ώστε να είναι ικανή να διαχειριστεί και την εποχιακότητα. Η νέα μορφή της μεθόδου αποτελείται συνολικά από τέσσερις εξισώσεις,

μία για την πρόβλεψη, μία για το επίπεδο της χρονοσειράς, μία για την τάση και μία για την εποχιακότητα.

Παρόμοια με τις μεθόδους αποσύνθεσης που είδαμε στις προηγούμενες ενότητες η εποχιακή μέθοδος Holt-Winters δύναται να χρησιμοποιηθεί τόσο με προσθετική προσέγγιση όσο και με πολλαπλασιαστική.

Προσθετικό μοντέλο Holt-Winters

Χρησιμοποιούμε την προσθετική μέθοδο όταν οι εποχιακές διακυμάνσεις παραμένουν επί το πλείστον σταθερές καθ' όλη τη χρονική έκταση της χρονοσειράς. Σε αυτή τη περίπτωση η η εποχιακότητα εκφράζεται κατά απόλυτη τιμή στη κλίμακα των παρατηρήσεων, και μπορούμε να λάβουμε την αποεποχικοποιημένη χρονοσειρά αφαιρώντας το στοιχείο της εποχιακότητας. Προσθέτοντας τις τιμές της εποχιακότητας για έναν κύκλο της το αποτέλεσμα πρέπει να είναι μηδέν.

Συμβολίζουμε το μήκος του κύκλου εποχιακότητας ως m και τον ορίζοντα πρόβλεψης ως h. Ο ορίζοντας πρόβλεψης, μία έννοια που θα αναλύσουμε λεπτομερέστερα στο επόμενο κεφάλαιο, μας δείχνει πόσο βήματα θα προχωρήσει στο χρόνο το μοντέλο πρόβλεψης μας σε σχέση με τα δεδομένα της αρχικής χρονοσειράς. Έτσι, έχουμε τη προσθετική μέθοδο Holt-Winters να περιγράφεται από τις ακόλουθες εξισώσεις:

$$y_{t+h|t} = l_t + hb_t + s_{t-m+h_m^+}$$
$$l_t = \alpha(y_t - s_{t-m} + (1-\alpha)(l_{t-1} + b_{t-1}$$
$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}$$
$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$$

Με h_m^+ να η θέση της παρατήρησης h στον κύκλο εποχιακότητας μ που ξεκινάει στο $h = 1$. Έτσι εξασφαλίζουμε ότι κατά την διαδικασία πρόβλεψης χρησιμοποιούμε δείκτες εποχιακότητας που προέρχονται από τον τελευταίο εποχιακό κύκλο των παρατηρήσεων. Επίσης τα α, β^* και γ είναι οι παράμετροι εξομάλυνσης του μοντέλου μας. Στην εξίσωση επιπέδου της χρονοσειράς μας συναντάμε ένα σταθμισμένο μέσο μεταξύ την αποεποχικοποιημένης παρατήρησής μας και της μη-εποχιακής πρόβλεψης για τη χρονική στιγμή t. Όπως θα δούμε παρακάτω η εξίσωση της τάσης είναι παρόμοια με αυτή της γραμμικής μεθόδου Holt και τελικά η εξίσωση της εποχιακότητας εμπεριέχει ένα σταθμισμένο μέσο μεταξύ της παρούσας χρονικής στιγμής και του δείκτη εποχιακότητας της ίδιας περιόδου του προηγούμενου κύκλου εποχιακότητας.

Κάνοντας πρόβλεψη ενός βήματος θεωρούμε το σφάλμα να είναι ίσο με:

$$e_t = y_t - (L_{t-1} + b_{t-1} + s_{t-m}) = y_t - y_{t|t-1}$$

και αντίστοιχα η μορφή της διόρθωσης σφάλματος των εξισώσεων εξομάλυνσης είναι:

$$l_t = l_{t-1} + b_{t-1} + \alpha e_t$$
$$b_t = b_{t-1} + \alpha\beta^* e_t$$
$$s_t = s_{t-m} + \gamma e_t$$

Πολλαπλασιαστικό μοντέλο Holt-Winters

Τη πολλαπλασιαστική μέθοδο τη χρησιμοποιούμε όταν οι διακυμάνσεις της εποχιακής συμπεριφοράς είναι ανάλογες του επιπέδου της χρονοσειράς. Εδώ, η εποχιακότητα εκφράζεται σε σχετική (ποσοστιαία) μορφή και μπορούμε να λάβουμε την αποεποχικοποιημένη χρονοσειρά διαιρώντας την αρχική μας σειρά με την εποχιακότητα. Το άθροισμα των τιμών ενός κύκλου εποχιακότητας της πολλαπλασιαστικής προσέγγισης αθροίζει στο μήκος του κύκλου. Για παράδειγμα, σε ετήσιες χρονοσειρές με μηνιαία δεδομένα το άθροισμα των τιμών για όλους τους μήνες ενός χρόνου πρέπει να είναι 12.

Το σύστημα εξισώσεων της πολλαπλασιαστικής μεθόδου είναι:

$$y_{t+h|t} = (l_t + hb_t)s_{t-m+h_m^+}$$

$$l_t = \alpha\frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1}$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}$$

$$s_t = \gamma\frac{y_t}{l_{t-1} - b_{t-1}} + (1-\gamma)s_{t-m}$$

και το αντίστοιχο διόρθωσης σφάλματος:

$$l_t = l_{t-1} + b_{t-1} + \alpha\frac{e_t}{s_{t-m}}$$

$$b_t = b_{t-1} + \alpha\beta^*\frac{e_t}{s_{t-m}}$$

$$s_t = s_{t-m} + \gamma\frac{e_t}{l_{t-1} + b_{t-1}}$$

με το σφάλμα να είναι:

$$e_t = y_t - (L_{t-1} + b_{t-1})s_{t-m}$$

3.4.2 Εποχιακά Μοντέλα ARIMA

Τα ολοκληρωμένα αυτοπαλινδρομικά μοντέλα κινητών μέσων όρν ARIMA (Auto Regressive Integrated Moving Average) είναι στοχαστικά μοντέλα ανάλυσης και πρόβλεψης της εξέλιξης μεγεθών. Σε αντίθεση με ντετερμενιστικά μοντέλα που θα δούμε στο επόμενο κεφάλαιο, τα μοντέλα ARIMA βασίζονται στον υπολογισμό της πιθανότητας που περιγράφει πως η τιμή ενός μεγέθους παίρνει τιμές σε ένα διάστημα. Οι Box και Jenking μελέτησαν εκτενώς αυτά τα μοντέλα και έτσι πολλές φορές τα συναντάμε στη βιβλιογραφία με τα ονόματα τους.

Ενσωματωμένο στα μοντέλα ARIMA είναι και το σφάλμα πρόβλεψης, δηλαδή ο τυχαίος παράγοντας,οι τιμές του μεγέθους που βρήκαμε στις προηγούμενες περιόδους και σχετικούς στοχαστικού παράγοντες. Μπορούμε να εκφράσουμε ένα μοντέλο ARIMA σαν γραμμικό συνδυασμό των παραπάνω παραγόντων, που ο βέλτιστος συνδυασμός μπορεί να παράξει τις καλύτερες προβλέψεις. Σε πραγματικά δεδομένα, δεν είναι εύκολο να τον εντοπίσουμε πάντα, αλλά μπορούμε να τον προσεγγίσουμε σε ικανοποιητικό βαθμό.

Βέβαια πρέπει να πληρούνται κάποιες προϋποθέσεις για να μπορεί να εφαρμοστεί ένα μοντέλου ARIMA.

- Η χρονοσειρά πρέπει να είναι διακριτή
- Η χρονοσειρά πρέπει να είναι στάσιμη
- Στόχος μας να είναι η βραχυπρόθεσμη πρόβλεψη

Μοντέλα ARIMA για εποχιακή πρόβλεψη

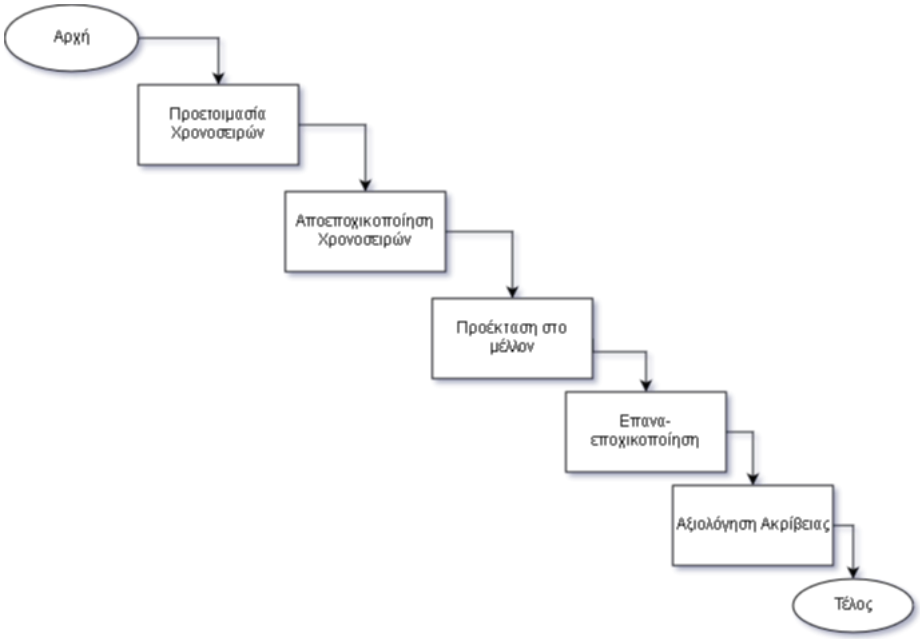
Τα μοντέλα ARIMA δύνανται να μοντελοποιήσουν ένα μεγάλο εύρος εποχιακών δεδομένων και αυτό επιτυγχάνεται ενσωματώνοντας στο μοντέλο επιπρόσθετους εποχιακούς όρους. Εκφράζουμε τα παραπάνω ως εξής:

$$ARIMA(p, d, q)(P, D, Q)_m$$

Τα μικρά γράμματα δηλώνουν τα κλασικά μη εποχικά κομμάτια του μοντέλου, ενώ τα κεφαλαία δηλώνουν τα εποχιακά κομμάτια. Όπως και πριν το m δηλώνει το μέγεθος του κύκλου εποχιακότητας. Τα εποχιακά κομμάτια του μοντέλου συμπεριλαμβάνουν όρους που είναι συναφείς με τους μη εποχιακούς, αλλά εμπεριέχουν βαςκσηιφτς του εποχιακού κύκλου. Οι εν λόγω όροι απλά πολλαπλασιάζονται με τους μη εποχιακούς.

Συναρτήσεις αυτοσυσχέτισης και μερικής αυτοσυσχέτισης: ACF/PACF

Τα μοντέλα ARIMA απαρτίζονται από ένα μοντέλο autoregression (AR) και ένα μοντέλο moving averages (MA). Το εποχιακό κομμάτι οποιουδήποτε από τα δύο αυτά μοντέλα θα γίνει εμφανές στις εποχιακές καθυστερήσεις των PACF και ACF. Έτσι, για να χρησιμοποιήσουμε τη σωστή εποχιακή σειρά για ένα μοντέλο, πρέπει να εστιάσουμε τη προσοχή μας στις εποχιακές καθυστερήσεις. Η μοντελοποίηση είναι παρόμοια με αυτή των μη εποχιακών δεδομένων, παρά το γεγονός ότι πρέπει να προσδιορίσουμε και τους εποχιακούς όρους που περιγράψαμε παραπάνω.



Σχήμα 4.1: Κλασική μεθοδολογία πρόβλεψης χρονοσειρών

Κεφάλαιο 4

Τεχνικές Προβλέψης Χρονοσειρών

4.1 Εισαγωγή

Μέσα από την εφαρμογή των κλασικών μεθόδων προβλέψεων αποσκοπούμε με χρήση των παρελθοντικών παρατηρήσεων μιας χρονοσειράς να εκτιμήσουμε τις μελλοντικές. Η πρόβλεψη αποτελεί μια πολυβηματική διαδικασία που θα αναπτυχθεί στο παρόν κεφάλαιο και περιγράφεται από το διάγραμμα ροής του Σχήματος 4.1.

4.2 Προετοιμασία Χρονοσειρών

4.2.1 Γραφική Αναπαράσταση Δεδομένων

Κατά τη διαδικασία της επεξεργασίας και πρόβλεψης μιας χρονοσειράς είναι σημαντικό ο ερευνητής να μπορεί να αποκτήσει μία διαίσθηση πάνω στα δεδομένα. Μέσω της αναπαράστασης τους σε μια γραφική απεικόνιση μπορεί εύκολα να διαπιστώσει ποια ποιοτικά χαρακτηριστικά της χρονοσειράς έχουν έντονο βάρος. Μπορεί έτσι να εντοπίσει ότι η χρονοσειράς χαρακτηρίζεται από ακραίες τιμές, έχει έντονη τάση ή μοτίβα που επαναλαμβάνονται και εποχιακότητα.

4.2.2 Διαχείριση Ιδιομορφίας Χρονοσειρών

Είναι αρκετά σύνηθες τα δεδομένα που προορίζονται για πρόβλεψη να παρουσιάζουν δυσμορφίες. Μη σταθερή συχνότητα μεταξύ δύο συνεχόμενων τιμών μιας χρονοσειράς τη καθιστά αταίριαστη για πολλές από τις κλασικές μεθόδους πρόβλεψης. Τα δεδομένα βρίσκονται σε συχνότητα διαφορετική από αυτή που θέλουμε να προβλέψουμε, λόγου χάρη ημερήσια, ενώ είναι επιθυμητό να παραχθούν μηνιαίες προβλέψεις. Επίσης, υπάρχει η περίπτωση ενώ εν γένει η σειρά αποτελείται από τιμές που ισαπέχουν μεταξύ τους στο πεδίο του χρόνου, παρουσιάζονται περιπτώσεις που δεν έχουν σημειωθεί κάποιες παρατηρήσεις. Αυτές τις ελλείψεις τις λέμε κενές τιμές. Μία άλλη ιδιομορφία που μπορεί να συναντήσουμε και καθιστά τη χρονοσειρά μη διαχειρίσιμη από αρκετές μεθόδους είναι η ύπαρξη μηδενικών τιμών.

Επαναδειγματοληψία

Στη περίπτωση που το χρονικό διάστημα μεταξύ δύο παρατηρήσεων δεν είναι σταθερό ή ακόμα και διαφορετικό από αυτό που θέλουμε να έχουμε στη πρόβλεψή μας, πρέπει να φέρουμε τα δεδομένα μας σε μια πιο κανονικοποιημένη μορφή. Μία μέθοδος που μπορούμε να ακολουθήσουμε είναι η μέθοδος της επαναδειγματοληψίας (resampling). Η χρονοσειρά, συχνά, περιγράφει φαινόμενα συνεχούς χρόνου μέσα από διακριτό χρόνο. Με την επαναδειγματοληψία χρησιμοποιούμε τη πληροφορία που έχουμε ήδη στη διάθεση μας για να δημιουργήσουμε μια νέα χρονοσειρά που της έχουμε ορίσει τους ακριβείς χρόνους των παρατηρήσεων. Η επιλογή αυτή γίνεται τόσο βάσει των τιμών που ήδη έχουμε αλλά και του στόχου μας για το αποτέλεσμα στη πρόβλεψη.

Έχοντας επιλέξει τη νέα συχνότητα των δεδομένων, μένει να επιλέξουμε με ποιόν τρόπο θα γίνει η μετατροπή. Μπορούμε να διακρίνουμε δύο βασικές περιπτώσεις ανάλογα με το αν μεταβαίνουμε από μεγαλύτερη σε μικρότερη συχνότητα. Την δειγματοληψία προς τα πάνω (upsampling) και τη δειγματοληψία προς τα κάτω (downsampling).

Για παράδειγμα, μπορούμε να έχουμε δεδομένα που έχουν συλλεχθεί σε ωριαία βάση, ενώ η πρόβλεψή μας θέλουμε να γίνει σε ημερήσιο επίπεδο. Τότε χρησιμοποιούμε downsampling, μετατρέποντας τα δεδομένα με μια κατάλληλη συνάρτηση, όπως το άθροισμα των ωριαίων τιμών, ο μέσος όρος τους, το μέγιστο ή το ελάχιστο. Αντίστοιχα αν τα δεδομένα μας θέλουμε να αλλάξουν από μικρότερη συχνότητα σε μεγαλύτερη, όπως από ημερήσια σε ωριαία, χρησιμοποιούμε upsampling, που συνδυάζεται συνήθως συμπληρώνοντας παρεμβολικά τις νέες

χρονικές στιγμές που προκύπτουν. Η παρεμβολή θα περιγραφεί στη διαχείριση κενών τιμών

Διαχείριση κενών τιμών

Αρχικά, εφόσον αυτό είναι δυνατό, προσπαθούμε να βρούμε τις τιμές που μας λείπουν από άλλες πηγές, πέραν της αρχικής που λάβαμε τα δεδομένα. Επίσης, αν οι σειρές που αναλύουμε είναι λίγες στο πλήθος και γνωρίζουμε τη φύση της, μπορούμε να ορίσουμε απευθείας τις κενές τιμές, στη περίπτωση που βάσει των παραπάνω μπορούμε να κάνουμε μία ασφαλή εκτίμηση.

Στη πράξη τα παραπάνω δεν είναι συχνά εφικτά, ιδίως στις περιπτώσεις που πρέπει να γίνει διαχείριση μεγάλου αριθμού χρονοσειρών. Έτσι διακρίνουμε δύο περιπτώσεις: χρονοσειρές με εποχιακή συμπεριφορά και μη.

Όταν η χρονοσειρά προς επεξεργασία παρουσιάζει σαφή εποχιακή συμπεριφορά, μπορούμε να εκτιμήσουμε τη τιμή βάσει των τιμών των αντίστοιχων περιόδων με την εν λόγω κενή τιμή. Ένας τρόπος είναι να πάρουμε τον μέσο όρο. Έτσι, αν σε μία ετήσια χρονοσειρά με μηνιαία δεδομένα, αν μας λείπει η τιμή από κάποιον Ιανουάριο, μπορούμε να τη συμπληρώσουμε ως τον μέσο όρο των άλλων.

Στη περίπτωση που δε μιλάμε για εποχιακή συμπεριφορά, χρησιμοποιούμε τη μέθοδο της παρεμβολής για να συμπληρώσουμε τις κενές τιμές μας. Χρησιμοποιώντας τις γειτονικές τιμές μιας ακολουθίας κενών τιμών, τις συμπληρώνουμε με χρήση κατάλληλης μεθόδου. Μια κλασική προσέγγιση είναι να συμπληρωθούν γραμμικά, ως σημεία της ευθείας που ορίζουν οι δύο οριακές τιμές. Ειδάλλως μπορούμε να χρησιμοποιήσουμε άλλες μεθόδους όπως να δώσουμε τη τιμή της πιο κοντινής από τις δύο τιμές, να τις συμπληρώσουμε ως σημεία ενός πολυωνύμου δευτέρου, τρίτου ή μεγαλύτερου βαθμού. Πιο σύνθετες μέθοδοι είναι αυτές των splines, krogh, βαρυκεντρίκή και πολυωνυμική κατά σημείο.

Διαχείριση μηδενικών τιμών

Διακρίνουμε δύο περιπτώσεις, αν η παρατήρηση είναι πράγματι μηδενική ή έχει πάρει αυτή τη τιμή λόγω λάθους στη συλλογή δεδομένων. Προφανώς, στη δεύτερη περίπτωση χειριζόμαστε αυτές τιμές σαν να ήταν κενές και ακολουθούμε τις διαδικασίες που περιγράφηκαν προηγουμένως.

Αλλιώς, αν οι μηδενικές τιμές είναι λίγες, δεν θα επηρεάσουν σε μεγάλο βαθμό τα μοντέλα πρόβλεψης μας. Σε αντίθετη περίπτωση, χρησιμοποιούμε ειδικά μοντέλα για τη διαχείριση διακοπτόμενης ζήτησης.

4.2.3 Ημερολογιακές προσαρμογές

Η χρονοσειρές είναι συχνό να περιγράφουν τιμές που σχετίζονται με ανθρώπινες ενέργειες και επηρεάζονται από τον τρόπο που είναι οργανωμένη η ανθρώπινη κοινωνία. Έτσι προετοιμάζοντας μια χρονοσειρά για πρόβλεψη πρέπει να εντοπίσουμε με πιο τρόπο συμβαίνει αυτή η επιρροή.

Σε ημερήσιες χρονοσειρές, αυτό γίνεται προσαρμόζοντας τα δεδομένα βάσει ημερολογιακών γεγονότων. Καθορίζουμε, λοιπόν, τις εργάσιμες μέρες συσχετισμένες με το μέγεθος που

περιγράφει η σειρά και τις αντίστοιχες αργίες. Έχοντας προσδιορίσει τα παραπάνω, υπολογίζουμε τις εργάσιμες μέρες για κάθε περίοδο των δεδομένων μας (N_t) και τον μέσο όρο τους για όλες τις περιόδους (N_{avg}).

Μπορούμε έτσι να εξομαλύνουμε τη τιμή της κάθε παρατήρησης μας ως εξής:

$$Y_t' = Y_t.\frac{N_{avg}}{N_t}$$

4.3 Προέκταση χρονοσειρών

4.3.1 Εισαγωγή

Αφότου έχουμε προετοιμάσει τη χρονοσειρά και την έχουμε αποσυνθέσει στα επιμέρους της στοιχεία, όπως περιγράφηκε στο προηγούμενο κεφάλαιο, προχωράμε στην εφαρμογή κάποιου μοντέλου πρόβλεψης στα αποεποχικοποιημένα δεδομένα. Στόχος στην εφαρμογή του μοντέλου είναι το αποτέλεσμα που θα προκύψει να είναι όσο πιο ακριβές δύναται, δηλαδή οι τιμές που θα παράξει το μοντέλο να είναι όσο το δυνατόν πιο κοντά στις πραγματικές που θα έχουμε στη διάθεση μας με τη πάροδο του χρόνου. Στο παρόν κεφάλαιο θα μελετήσουμε τις στατιστικές μεθόδους πρόβλεψης.

4.3.2 Naive: η αφελής μέθοδος

Η Naive είναι η πιο απλή μέθοδος στατιστικής πρόβλεψης. Θεωρεί ότι η τιμή που θα ακολουθήσει χρονικά ταυτίζεται με τη τιμή της παρούσας περιόδου, σύμφωνα με τη παρακάτω σχέση, όπου το F_t είναι η πρόβλεψη κατά τη χρονική στιγμή t και Y_t η πραγματική τιμή της σειράς:

$$F(t+1) = Y(t)$$

Είναι φυσικό ότι η Naive δεν παράγει ακριβείς προβλέψεις αλλά μπορούμε να τη χρησιμοποιήσουμε ως βάση σύγκρισης (benchmark) άλλων μεθόδων.

4.3.3 Μοντέλα Παλινδρόμησης

Η ανάλυση της παλινδρόμησης αποσκοπεί στην εύρεση συσχετίσεων μεταξύ μιας εξαρτημένης μεταβλητής και μίας ή περισσοτέρων ανεξάρτητων μεταβλητών. Έχει ευρεία χρήση στη διαδικασία των προβλέψεων, τόσο ως μοντέλο πρόβλεψης αλλά και ως υποβοήθημα σε άλλες μεθόδους, όπως θα δούμε παρακάτω. Κυρίως, όμως μας επιτρέπει να βγάλουμε συμπεράσματα για τη συσχέτιση της ανεξάρτητης μεταβλητής και των εξαρτημένων μεταβλητών.

Απλή Γραμμική Παλινδρόμηση

Η μέθοδος, που φέρει και το όνομα μέθοδος των ελαχίστων τετραγώνων, περιγράφει μία ευθεία με την ελάχιστη απόσταση ανά σημείο από τα πραγματικά δεδομένα. Για να τη λάβουμε πρέπει να ελαχιστοποιηθεί το άθροισμα σφαλμάτων στη δεύτερη εξίσωση που ακολουθεί. Επίσης, φαίνεται αναλυτικά πως προκύπτουν και οι συντελεστές:

$$\hat{Y_i} = \alpha + \beta X_i$$

$$\sum e_i^2 = \sum (Y_i - \hat{Y_i})^2$$

$$\beta = \frac{\frac{\sum X_i Y_i}{n} - \bar{X}\bar{Y}}{\frac{\sum X_i^2}{n} - \bar{X}^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\alpha = \bar{Y} - \beta \bar{X}$$

Για να χρησιμοποιήσουμε τη μέθοδο της Απλής Γραμμικής Παλινδρόμησης, πρέπει να υπάρχει εξάρτηση της ανεξάρτητης μεταβλητής από τη τιμή ή τη μεταβολή κάποιας άλλης. Για να ελέγξουμε αν αυτό συμβαίνει, χρησιμοποιούμε τον συντελεστή γραμμικής συσχέτισης r που αποτελεί έναν δείκτη του βαθμού που συσχετίζονται δύο μεταβλητές μεταξύ τους και για δύο μεταβλητές X, Y προκύπτει ως εξής:

$$Cov_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

$$Cov_{XX} = \frac{\sum (X_i - \bar{X})^2}{n} = Var_X = S_X^2$$

$$Cov_{YY} = \frac{\sum (Y_i - \bar{Y})^2}{n} = Var_Y = S_Y^2$$

$$r_{XY} = \frac{Cov_{XY}}{\sqrt{Cov_{YY} * Cov_{XX}}} = \frac{Cov_{XY}}{S_Y * S_X}$$

Και προκύπτει ότι ο συντελεστής r_{XY} λαμβάνει τιμές στο διάστημα από -1 έως 1.
Ο συντελεστής γραμμικής συσχέτισης επιδέχεται δύο ερμηνείες:

- Μας δείχνει την κατεύθυνση της σχέσης μεταξύ των δύο μεταβλητών, δηλαδή αν όταν οι τιμές της μίας αυξάνονται, οι τιμές τις άλλης μειώνονται οι αυξάνονται. Επίσης, μπορεί να μας δείξει ότι η μεταβολές της μίας είναι ανεξάρτητες από τις άλλης.

- Μας δείχνει τον βαθμό συσχέτισης και συνεπώς τη δυνατότητα της γραμμής παλινδρόμησης να εκφράσει τη σχέση μεταξύ των μεταβλητών. Όσο το r_{XY} πλησιάζει κατά απόλυτη τιμή τη μονάδα τόσο μικρότερη είναι η απόκλιση των πραγματικών τιμών της εξαρτημένης μεταβλητής από αυτές του μοντέλου.

Η μέθοδος των ελάχιστων τετραγώνων χρησιμοποιείται ως μοντέλο πρόβλεψης χρονοσειρών, απλά θέτοντας ως ανεξάρτητη μεταβλητή το χρόνο. Η γραμμή προκύπτει όπως περιγράψαμε προηγουμένως από τα ιστορικά δεδομένα, και προεκτείνοντας τη στο χρόνο λαμβάνουμε τις τιμές τις πρόβλεψης για τις χρονικές στιγμές που μας ενδιαφέρουν.

Πολλαπλή Παλινδρόμηση

Υπάρχουν περιπτώσεις που έχουμε πληροφορία για περισσότερες ανεξάρτητες μεταβλητές, τότε χρησιμοποιούμε τη μέθοδο της πολλαπλής παλινδρόμησης, η παίρνει την εξής μορφή:

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \cdots + b_k * X_k + e$$

Αντίστοιχα με πριν, πρέπει να ελαχιστοποιηθεί το τετραγωνικό σφάλμα. Για να βρούμε τις τιμές των συντελεστών που μας οδηγούν στο ποθητό αποτέλεσμα, προσδιορίζουμε τις τιμές των μερικών παραγώγων της συνάρτησης του σφάλματος για κάθε συντελεστή. Κατόπιν λύνουμε το γραμμικό σύστημα που προκύπτει αν θέσουμε τη κάθε μερική παράγωγο ίση με το μηδέν.

4.3.4 Μοντέλα Εκθετικής Εξομάλυνσης

Οι μέθοδοι εκθετικής εξομάλυνσης εμφανίστηκαν στο τέλος της δεκαετίας του 40′, από τον Brown με σκοπό την πρόβλεψη αποθεμάτων και συνεχίζονται να εξελίσσονται ως σήμερα, αποτελώντας τις πιο δημοφιλείς μεθόδους προβλέψεων. Ο λόγος είναι ότι πρόκειται για σχετικά απλά μοντέλα με ευκολία χρήσης, μικρές υπολογιστικές απαιτήσεις και την δυνατότητα να παράξουν ακριβείς προβλέψεις ακόμα και με σχετικά μικρό ιστορικό παρατηρήσεων. Έτσι, χρησιμοποιούνται συχνά για τη παράλληλη πρόβλεψη πολλών χρονοσειρών, συνήθως με μικρό ορίζοντα πρόβλεψης.

Σε όλα τα μοντέλα που ακολουθούν έχουμε μια εξομάλυνση των ιστορικών δεδομένων. Χρησιμοποιούνται συντελεστές βαρύτητας, που μειώνονται εκθετικά όσο πηγαίνουμε πίσω στο χρόνο, για να υπολογίσουμε τον μέσο όρο των παρατηρήσεων και να απαλλαχθούμε από τις τυχαίες διακυμάνσεις.

Προκύπτουν συνολικά 12 τύποι μοντέλων εξομάλυνσης ως συνδυασμός των παρακάτω:

- Πρότυπα τάσης:
  - Σταθερού Επιπέδου
  - Γραμμικής Τάσης
  - Εκθετικής Τάσης
  - Φθίνουσας Τάσης
- Πρότυπα εποχιακότητας:
  - Άνευ Εποχιακότητας
  - Προσθετικής Εποχιακότητας
  - Πολλαπλασιαστικής Εποχιακότητας

Ακολουθούν κάποια βασικά μοντέλα εκθετικής εξομάλυνσης.

Απλή εκθετική εξομάλυνση

Το μοντέλο της απλής εκθετικής εξομάλυνσης (Simple Exponential Smoothing) είναι ένα μοντέλο σταθερού επιπέδου που είναι ιδανικό για πρόβλεψη ενός βήματος. Επίσης, είναι κατάλληλο για χρονοσειρές που παρουσιάζουν υψηλό θόρυβο ή έντονο το στοιχείο της τυχαιότητας. Περιγράφεται από τις παρακάτω σχέσεις:

$$e_t = Y_t - F_t$$

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

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

Εκτός των F_t και Y_t που δηλώνουν τα ίδια μεγέθη όπως στη $Naive$, έχουμε το e που δηλώνει το σφάλμα, το S που δηλώνει το επίπεδο και το α, τον συντελεστή εξομάλυνσης της μεθόδου με δυνατές τιμές ανάμεσα στο 0 και στο 1. Βλέπουμε, λοιπόν, ότι πρέπει να χρησιμοποιήσουμε δύο παραμέτρους: τη πρώτη τιμή της πρόβλεψη F_1 και το παράγοντα εξομάλυνσης α. Αν επιλέξουμε μεγάλη τιμή για το α, η τελική τιμή της πρόβλεψης βασίζεται λιγότερο στις αρχικές τιμές και με $\alpha = 1$ η μέθοδος ταυτίζεται με τη Naive. Επίσης για χρονοσειρές πολύ μεγάλου μήκους η συνεισφορά της αρχικής πρόβλεψης στη τελική τιμή μειώνεται εκθετικά με το μήκος.

Υπάρχουν διάφοροι τρόποι να ορίσουμε το αρχικό επίπεδο μια χρονοσειράς, μερικοί εκ των οποίων είναι:

- Ο μέσος όρος όλων των τιμών της χρονοσειράς
- Ο μέσος όρος ν το πλήθος αρχικών τιμών της σειράς
- Η πρώτη τιμή
- Να υπολογίσουμε τη γραμμή της γραμμικής παλινδρόμησης και να θέσουμε το αρχικό επίπεδο ως το σταθερό επίπεδο αυτής της γραμμής

Για να εντοπίσουμε τον κατάλληλο συντελεστή εξομάλυνσης για βέλτιστη ακρίβεια πρέπει να λάβουμε υπόψιν μας τόσο το θόρυβο της σειράς όσο και τη σταθερότητα του μέσου όρου της. Σε περίπτωση έντονου θορύβου, μικρή τιμή του συντελεστή μας διασφαλίζει ότι το μοντέλο μας δε θα αντιδρά υπερβολικά στο θόρυβο. Πρόσθετα, στη περίπτωση που μεταβάλλεται ο μέσος όρος των τιμών της σειράς όσο προχωράμε στο πεδίο του χρόνου, ένα μεγάλος συντελεστής εξομάλυνσης επιτρέπει στο μοντέλο να ακολουθεί αυτή τη μεταβολή.

Εκθετική εξομάλυνση γραμμικής τάσης

Μια επέκταση της προηγούμενης μεθόδου είναι το μοντέλο εκθετικής εξομάλυνσης για γραμμικής τάση Holt Exponential Smoothing που εισήχθηκε από τον Holt το 1957. Η μέθοδος περιγράφεται από τις επόμενες εξισώσεις:

$$e_t = Y_t - F_t$$

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

$$T_t = T_{t-1} + \alpha * \beta * e_t$$

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

Όπου το νέο στοιχείο T_t δηλώνει την τάση του μοντέλου κατά τη χρονική στιγμή t. Εδώ το α αποτελεί τον συντελεστή εξομάλυνσης του επιπέδου ενώ το β το συντελεστή εξομάλυνσης της τάσης. Τόσο το αρχικό επίπεδο και η αρχική τάση πρέπει να προσδιοριστούν με προσοχή καθώς έχουν μεγάλη επιρροή στις τιμές του μοντέλου της πρόβλεψης. Για το πρώτο ακολουθούμε μία από τις μεθόδους υπολογισμού που περιγράφηκαν στην απλή εκθετική εξομάλυνση. Για την αρχική τάση μπορούμε να χρησιμοποιήσουμε μία από τις παρακάτω:

- Διαφορά των δύο πρώτων τιμών της χρονοσειράς
- Διαφορά της τιμής μια παρατήρησης της χρονοσειράς με τη πρώτη, και διαίρεση του αποτελέσματος με τον αριθμό των χρονικών βημάτων που μεσολαβούν μεταξύ τους
- Να υπολογίσουμε τη γραμμή της γραμμικής παλινδρόμησης και να θέσουμε την αρχική τάση ως τη κλίση αυτής της γραμμής

Μοντέλο μη γραμμική τάσης

Το μοντέλο μη γραμμικής τάση εισήχθηκε το 1985 από τους Gardner και McΚενζιε, αφότου είχε παρατηρηθεί ότι το μοντέλο γραμμικής τάσης πολλές φορές παρουσίαζε θετική προκατάληψη, ειδικά όταν εφαρμοζόταν με στόχο μεσοπρόθεσμες ή μακροπρόθεσμες προβλέψεις. Το μοντέλο περιγράφεται από τις εξής σχέσεις:

$$e_t = Y_t - F_t$$

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

$$T_t = T_{t-1} + \alpha * \beta * e_t$$

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

Παρατηρούμε ότι η αλλαγή σε σχέση με το προηγούμενο μοντέλο που περιγράψαμε εντοπίζεται στη τελευταία εξίσωση που, αντί να πολλαπλασιάζεται η τελευταία τάση που εντοπίσαμε με τη χρονική διαφορά σε περιόδους της τιμής προς πρόβλεψη από τη τελευταία τιμή των δεδομένων μας, έχουμε τη τελευταία τάση να πολλαπλασιάζεται με το άθροισμα των στοιχείων γεωμετρικής προόδου με λόγο ίσο με ϕ και μήκος m. Η παράμετρος ϕ, που ονομάζεται παράμετρος διόρθωσης της τάσης, καθορίζει το μοντέλο ως εξής:

- Για $\phi = 0$, το μοντέλο ταυτίζεται με αυτό της απλής εκθετικής εξομάλυνσης
- Για $0 < \phi < 1$, το μοντέλο χαρακτηρίζεται από φθίνουσα τάση
- Για $\phi = 1$, το μοντέλο ταυτίζεται με το μοντέλο γραμμικής τάσης

- Για $\phi > 1$, έχουμε εκθετική εξομάλυνση με εκθετική τάση

Για να προσδιορίσουμε το αρχικό επίπεδο και την αρχική τάση, χρησιμοποιούμε τις μεθόδους που περιγράφηκαν στις προηγούμενες υποενότητες.

4.3.5 Μέθοδος Theta

Η μέθοδος Θ είναι μια μονοδιάστατη μέθοδος πρόβλεψης που εισήχθηκε από τον Ασημακόπουλο το 1999. Κατά το μοντέλο, η χρονοσειρά αναλύεται σε δύο ή περισσότερες χρονοσειρές ή αλλιώς γραμμές Theta. Κάθε μία από τις προκύπτουσες σειρές προβλέπεται ξεχωριστά, είτε με το ίδιο είτε με διαφορετικό μοντέλο πρόβλεψης, και η τελική πρόβλεψη είναι ο συνδυασμός των επιμέρους αποτελεσμάτων.

Στην απλούστερη περίπτωση, το κλασσικό μοντέλο Θ, έχουμε δύο γραμμές Theta. Η πρώτη είναι μία ευθεία γραμμή που προκύπτει από την ευθεία γραμμικής παλινδρόμησης που μοντελοποιεί τα δεδομένα, η Theta Line (0). Η δεύτερη, Theta Line (2) προκύπτει ως εξής:

$$ThetaLine(2) = 2 * Y - LRL$$

Όπου Y η αρχική χρονοσειρά και LRL η ευθεία που προκύπτει από τη μέθοδο ελαχίστων τετραγώνων.

Η μέθοδος Θ βασίζεται στην μεταβολή των τοπικών καμπυλοτήτων μιας χρονοσειράς. Η μεταβολή αυτή λαμβάνει χώρα με τη χρήση της παραμέτρου ϑ που εφαρμόζεται στις δεύτερες διαφορές τις χρονοσειράς ως εξής:

$$Y_t^{\theta} = \theta * Y_t^{''}$$

$$Y_t^{''} = Y_t - 2 * Y_{t-1} + Y_{t-2}$$

Διακρίνουμε διαφορετικές συμπεριφορές των γραμμών Θ ανάλογα με τις τιμές τις παραμέτρου:

- Αν $\theta = 0$ η σειρά ταυτίζεται με την ευθεία γραμμικής παλινδρόμησης
- Αν $\theta = -1$ η σειρά που προκύπτει είναι η συμμετρική της αρχικής ως προς την ευθεία γραμμικής παλινδρόμησης
- Αν $\theta > 1$ έχουμε πιο έντονες καμπυλότητες, ανάλογες του βαθμού της παραμέτρου

Το τελικό μοντέλο πρόβλεψης προκύπτει ως γραμμικός συνδυασμός των γραμμών Θ, έτσι ώστε τα βάρη της κάθε σειράς να αθροίζουν στο 1. Στην κλασσική μέθοδο, όπως είδαμε στην εξίσωση υπολογισμού της Theta Line (2), έχουμε πάρει τα βάρη ίσα με 0.5 και για τις δύο γραμμές.

4.4 Αξιολόγηση Προβλέψεων

Αφότου έχουμε ολοκληρώσει τη διαδικασία πρόβλεψης πρέπει να αξιολογήσουμε κατά πόσο το μοντέλο μας παρήγαγε ακριβές προβλέψεις και στη περίπτωση που εφαρμόσαμε περισσότερα μοντέλα, πιο από αυτά είναι το πιο ακριβές. Για να το πετύχουμε αυτό χρησιμοποιούμε ένα σύνολο στατιστικών δεικτών αξιολόγησης της ακρίβειας. Αυτοί οι δείκτες, ή αλλιώς σφάλματα, μπορούν να χρησιμοποιηθούν για να μετρήσουν την απόδοση του υπό εφαρμογή μοντέλου στο σύνολο των δεδομένων του ιστορικού της χρονοσειράς, που είναι δηλαδή γνωστά στο μοντέλο μας (in-sample error, είναι στις τιμές του ορίζοντα πρόβλεψης (out-of-sample error). Αφότου ο σκοπός μας είναι να έχουμε ακριβή αποτύπωση της μελλοντικής εξέλιξης της χρονοσειράς, ενδιαφέρον για τη μέτρηση της απόδοσης του μοντέλου παρουσιάζει η δεύτερη κατηγορία.

Γενικά ορίζουμε ως σφάλμα της πρόβλεψης μιας περιόδου:

$$e_i = Y_i - F_i$$

Το απλό ποσοστιαίο σφάλμα ως:

$$p_i = \frac{100 * e_i}{Y_i}(\%)$$

Επίσης, έχουμε το απλό σχετικό σφάλμα, που είναι ο λόγος του σφάλματος της υπό εξέταση μεθόδου σε σχέση με κάποια άλλη μέθοδο. Συνήθως η μέθοδος σύγκρισης είναι η Naive, που περιγράφηκε πρωτύτερα:

$$r_i = \frac{e_i}{e_i^*}$$

Το απλό κανονικοποιημένο σφάλμα βρίσκεται από τη παρακάτω σχέση:

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

Στον Πίνακα 4.1 βλέπουμε τους πιο κοινούς δείκτες ακρίβειας. Το mean υποδηλώνει τον αριθμητικό μέσο όρο, το median τη διάμεσο και το gmean το γεωμετρικό μέσο. Επίσης το **1** λαμβάνει τη τιμή 1 στη περίπτωση που αυτό που εσωκλείει είναι αληθές, αλλιώς παίρνει την τιμή 0.

Στηριζόμενοι στη δουλειά των Hyndman και Koehler μπορούμε να χωρίσουμε τους δείκτες αυτούς στις εξής κατηγορίες:

Δείκτες που εξαρτώνται από τη κλίμακα των δεδομένων

Η κατηγορία αυτή αποτελείται από τους δείκτες MSE, RMSE, MAE, MdAE που συναρτώνται της απόλυτης τιμής των δεδομένων. Βοηθούν ιδιαίτερα όταν καλούμαστε να συγκρίνουμε την απόδοση διαφόρων μοντέλων πρόβλεψης επί της ίδιας χρονοσειράς. Σε περίπτωση που τους εφαρμόζουμε, όμως, σε ένα σύνολο χρονοσειρών με διαφορετική κλίμακα μπορούν να παράξουν αποπροσανατολιστικά αποτελέσματα. Το προτέρημα του δείκτη RMSE έναντι του

Συντομογραφία	Πλήρες Όνομα	Τύπος
ME	Mean Error	$mean(e_i)$
MSE	Mean Squared Error	$mean(e_i^2)$
RMSE	Rooted Mean Squared Error	$\sqrt{mean(e_i^2)}$
MAE	Mean Absolute Error	$mean(\|e_i\|)$
MdAE	Median Absolute Error	$median(\|e_i\|)$
MAPE	Mean Absolute Percentage Error	$mean(\|p_i\|)$
MdAPE	Median Absolute Percentage Error	$median(\|p_i\|)$
sMAPE	Symmetric Mean Absolute Percentage Error	$mean(\frac{200*\|Y_i-F_i\|}{Y_i+F_i)})$
sMdAPE	Symmetric Median Absolute Percentage Error	$median(\frac{200*\|Y_i-F_i\|}{Y_i+F_i)})$
MRAE	Mean Relative Absolute Error	$mean(\|r_i\|)$
MdRAE	Median Relative Absolute Error	$median(\|r_i\|)$
GMRAE	Geometric Mean Relative Absolute Error	$gmean(\|r_i\|)$
RelMAE	Relative Mean Absolute Error	MAE/MAE_b
RelRMSE	Relative Mean Squared Error	$RMSE/RMSE_b$
LMR	Log Mean Squared Error Ratio	$\log(RelMSE)$
PB	Percentage Better	$100 * mean(\mathbf{1}\{\|r_i\| < 1\})$
PB(MAE)	Percentage Better (MAE)	$100 * mean(\mathbf{1}\{MAE < MAE_b\})$
PB(MSE)	Percentage Better (MSE)	$100 * mean(\mathbf{1}\{MSE < MSE_b\})$
MAsE	Mean Absolute Scaled Error	$mean(\|q_i\|)$
MdAsE	Median Absolute Scaled Error	$median(\|q_i\|)$

Πίνακας 4.1: Δείκτες Ακρίβειας

απλού MSE είναι ότι μας δίνει μετρήσεις στην ίδια κλίμακα με αυτή της χρονοσειράς στην οποία εφαρμόζεται. Επίσης, οι δύο παραπάνω δείκτες είναι ιδιαίτερα ευαίσθητοι στις ακραίες τιμές μιας χρονοσειράς, σε σύγκριση με του MAE, MdAE λόγω του τετραγωνισμού του σφάλματος. Αυτό τους καθιστά ακατάλληλους για χρήση στην αξιολόγηση της ακρίβειας πρόβλεψης.

Δείκτες που βασίζονται σε ποσοστιαία σφάλματα

Σε αυτή την ομάδα, από τους δείκτες που είδαμε, ανήκουν οι MAPE, MdAPE, sMAPE και sMdAPE. Οι δείκτες MAPE και MdAPE υστερούν στο γεγονός ότι δεν δύνανται να λάβουν τιμή όταν οι πραγματικές παρατηρήσεις είναι μηδενικές και επιπρόσθετα παρουσιάζουν έντονη ασυμμετρία όταν οι παρατηρήσεις είναι κοντά στο μηδέν. Έτσι, ο δείκτες MAPE παρουσιάζει χαρακτηριστικά μεγάλες τιμές σε σχέση με τον δείκτη MdAPE. Μπορούμε, βέβαια, με τη χρήση λογαριθμικών μετασχηματισμών να προσδώσουμε μια σταθερότητα στους δείκτες. Επίσης, οι δύο αυτοί δείκτες δίνουν μεγαλύτερη βαρύτητα στα θετικά έναντι των αρνητικών

σφαλμάτων. Αντίθετα, οι δείκτες sMape και sMdAPE δεν παρουσιάζουν στον ίδιο βαθμό το πρόβλημα των μηδενικών τιμών. Αν και το όνομα τους υποδηλώνει συμμετρία, έχουν το μειονέκτημα ότι οι αισιόδοξες και οι απαισιόδοξες προβλέψεις δεν υπολογίζονται με το ίδιο βάρος.

Δείκτες σχετικών σφαλμάτων

Από τους δείκτες που έχουμε εξετάσει οι MRAE, MdRAE και GMRAE ανήκουν σε αυτή την κατηγορία. Το πρόβλημα με αυτή των ομάδα δεικτών είναι ότι στις περιπτώσεις που το σφάλμα αναφοράς λαμβάνει αρκετά μικρές τιμές ο λόγος r_i τείνει να έχει άπειρη διακύμανση.

Σχετικοί δείκτες

Το πλεονέκτημα των σχετικών δεικτών RelMAE, RelRMSE και PB είναι ότι αποφεύγουν το πρόβλημα που είδαμε στη προηγούμενη κατηγορία δεικτών με τις άπειρες τιμές. Επίσης, μας επιτρέπουν να αποφασίσουμε εύκολα πιο από τις μεθόδους που συγκρίνουμε είναι πιο ακριβής, ανάλογα με τον αν ο δείκτες έχει τιμή μεγαλύτερη της μονάδας. Βέβαια, καθότι απαιτούν αρκετές προβλέψεις δε μπορούμε να τους εφαρμόσουμε όταν έχουν πρόβλεψη με ορίζοντα ίσο με ένα.

Κανονικοποιημένοι δείκτες

Οι κανονικοποιημένοι δείκτες MAsE, MdAsE προσφέρουν τόσο ευκολία στην ερμηνεία αντίστοιχη των σχετικών δεικτών, ενώ συγχρόνως μπορούν να εφαρμοστούν για μοναδική περίοδο πρόβλεψης. Επιπρόσθετα, είναι απαλλαγμένες από το επίπεδο της κάθε χρονοσειράς και εφαρμόσιμες σε μαζική πρόβλεψη χρονοσειρών, αποφεύγοντας τις απροσδιοριστίες των ποσοστιαίων σφαλμάτων.

Κεφάλαιο 5

Προτεινόμενη Μεθοδολογία

5.1 Εισαγωγή

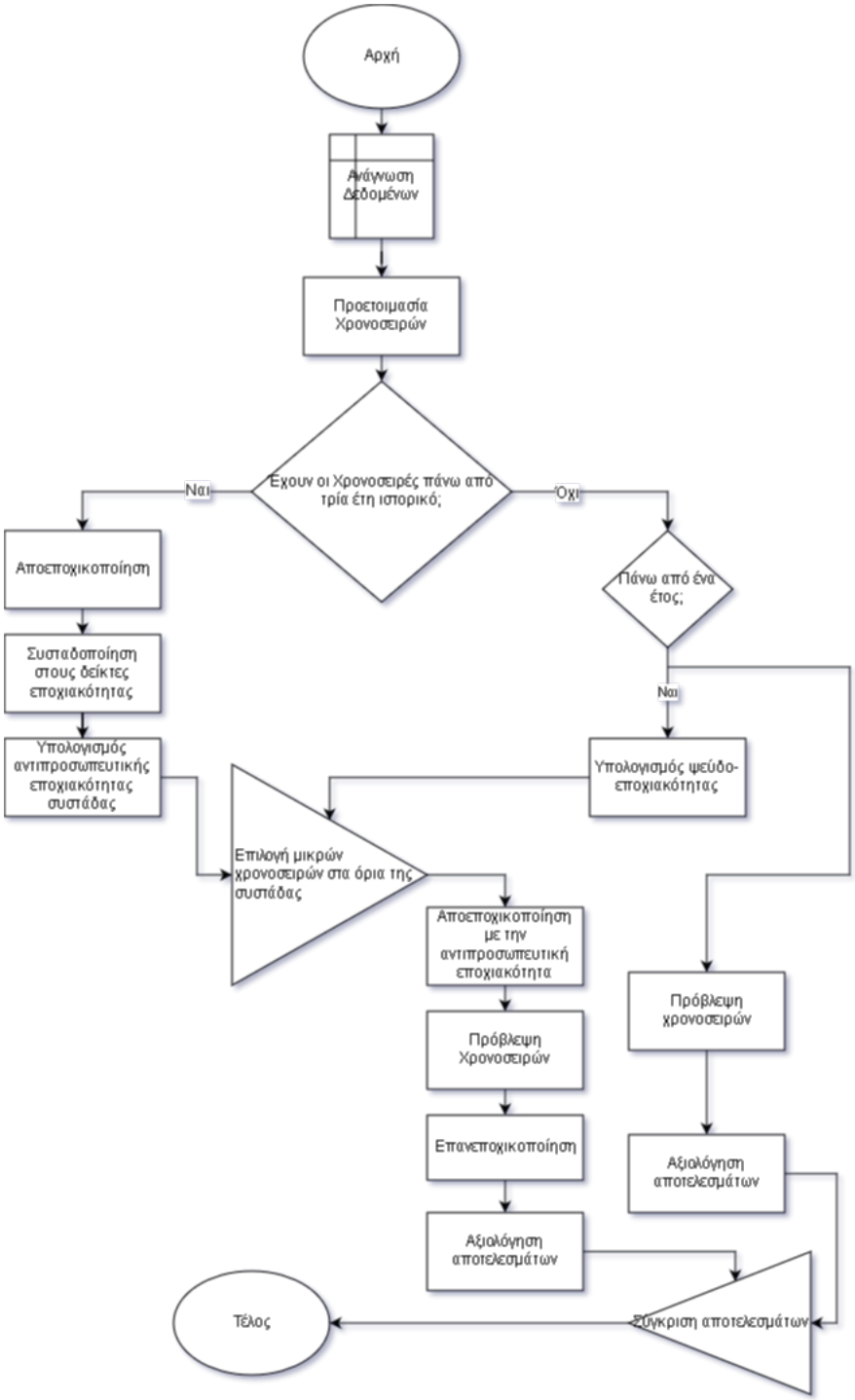
Στις προηγούμενες παραγράφους είδαμε τη συνηθισμένη διαδικασία πρόβλεψης μιας χρονοσειράς:

1. Προετοιμάζουμε τα δεδομένα και τα φέρνουμε στην επιθυμητή μορφή για ανάλυση
2. Αποσυνθέτουμε τη χρονοσειρά στα επιμέρους συνθετικά της στοιχεία
3. Προεκτείνουμε με τις μεθόδου πρόβλεψης την αποεποχικοποιημένη χρονοσειρά στο μέλλον
4. Ενσωματώνουμε την εποχιακότητα στο μοντέλο πρόβλεψης
5. Αξιολογούμε την ακρίβεια της πρόβλεψης

Μπορούμε όμως να παρατηρήσουμε ένα πρόβλημα στην παρακάτω διαδικασία. Στις περιπτώσεις που η χρονοσειρά μας δεν διαθέτει αρκετά ιστορικά δεδομένα, δεν είναι δυνατό για εμάς να εξάγουμε το στοιχείο της εποχιακότητας. Έτσι, διαχειριζόμαστε τη χρονοσειρά ως μη εποχιακή και εφαρμόζουμε τις μεθόδους πρόβλεψης στα αρχικά δεδομένα. Είναι προφανές, όμως, ότι μπορούμε να έχουμε στη διάθεση μας χρονοσειρές που είναι εποχιακές και που θα μπορούσαμε να τις προβλέψουμε καλύτερα στη περίπτωση που είχαμε γνώση για την εποχιακή τους συμπεριφορά.

Στη παρούσα διπλωματική προσπαθούμε να ξεπεράσουμε αυτό το πρόβλημα με χρήση της επιπλέον πληροφορίας που μπορούμε να λάβουμε από άλλες συναφείς χρονοσειρές με την υπό εξέταση χρονοσειρά. Έτσι, αν έχουμε στη διάθεσή μας δεδομένα που περιγράφουν παρόμοια μεγέθη, όπως λόγου χάρη πωλήσεις συναφών προϊόντων, υποθέτουμε ότι μπορούμε να εκμαιεύσουμε την ζητούμενη πληροφορία από το σύνολο των χρονοσειρών και να βελτιώσουμε την ακρίβεια της πρόβλεψης στις χρονοσειρές που στερούνται επαρκούς ιστορικού.

Σε αυτό το κεφάλαιο θα περιγράψουμε πώς εξετάστηκε αυτή η υπόθεση. Αρχικά, φέρνουμε τα δεδομένα μας σε κατάλληλη μορφή για την στατιστικές μεθόδους που θέλουμε να εφαρμόσουμε. Έπειτα, χωρίζουμε τις χρονοσειρές σε δύο ομάδες: εκείνες με επαρκές



Σχήμα 5.1: Διάγραμμα Ροής Μεθοδολογίας

ιστορικό για να εξάγουμε το στοιχείο της εποχιακότητας και τις υπόλοιπες. Στη πρώτη κατηγορία, εντοπίζουμε τους δείκτες εποχιακότητας κάθε χρονοσειρά και κατόπιν εξετάζουμε αν υπάρχει συνάφεια μεταξύ τους. Το αποτέλεσμα είναι να δημιουργηθούν συστάδες με παρόμοια εποχιακότητα και υπολογίζουμε τους αντιπροσωπευτικούς δείκτες εποχιακότητας της συστάδας. Στις χρονοσειρές που στερούνται επαρκής πληροφορίας, προσπαθούμε να βρούμε

μία ψεύδο-εποχιακότητα που θα χρησιμοποιήσουμε σαν κριτήριο συνάφειας με τις συστάδες που προέκυψαν προηγουμένως. Για τις μικρές χρονοσειρές που βρέθηκε να παρουσιάζουν παρόμοια εποχιακή συμπεριφορά με κάποια ομάδα, εφαρμόζουμε τα μοντέλα πρόβλεψης κάνοντας αποεποχικοποίηση με του λόγους εποχιακότητας της συστάδας. Κατόπιν κάνουμε πρόβλεψη στις ίδιες χρονοσειρές χωρίς αποεποχικοποίηση. Τελικώς, συγκρίνουμε τα αποτελέσματα των δύο προσεγγίσεων.

Μια αναπαράσταση της διαδικασία μπορούμε να δούμε στο διάγραμμα ροής του σχήματος 5.1.

5.2 Προετοιμασία και Αποσύνθεση Χρονοσειρών

Έλεγχος τιμών, resampling και διαχείριση κενών τιμών

Αρχικά, πρέπει να φέρουμε τα δεδομένα σε μία μορφή κατάλληλη για την ανάλυση και εφαρμογή μεθόδων που θέλουμε να κάνουμε. Για να το πετύχουμε αυτό εφαρμόζουμε τις τεχνικές επαναδειγματοληψίας και διαχείρισης κενών τιμών που αναπτύξαμε στο προηγούμενο κεφάλαιο, αφότου έχουμε επιλέξει από το σύνολο δεδομένων μας τη πληροφορία που μας είναι χρήσιμη.

Βέβαια, η παραπάνω διαδικασία δεν μπορεί να γίνει μόνο ως κλειστό κουτί (black box, δηλαδή να διαχειρίζεται πλήρως από το σύστημά μας. Πρέπει να αποκτήσουμε μία εποπτική διαίσθηση πάνω στα δεδομένα και στηριζόμενοι στη γνώση που έχουμε για τη φύση τους να διορθώσουμε τυχούσες ατέλειες. Μία συχνή τέτοια περίπτωση είναι οι αρνητικές τιμές σε ένα φυσικό μέγεθος που δεν δύναται να παίρνει τέτοιες.

Με παρόμοιο τρόπο πρέπει να σχεδιάσουμε τη διαδικασία του resampling και της διαχείρισης κενών τιμών. Ο λόγος είναι ότι καλούμαστε να επιλέξουμε τη κατάλληλη συνάρτηση ή προσέγγιση για να γίνουν τα παραπάνω και για να μην αλλοιωθεί η πληροφορία που φέρουν τα δεδομένα κατά αυτή τη διαδικασία, πρέπει να ορίσουμε ένα σύνολο φυσικών κανόνων που απορρέει από το προσδιορισμό της φύσης τους.

Τελικά, για να μπορέσουμε να αξιολογήσουμε τη προτεινόμενη μέθοδο πρέπει να αποκρύψουμε τις τελευταίες τιμές των χρονοσειρών μας. Το χρονικό διάστημα που αφαιρούμε από τα δεδομένα θα αποτελέσει και τον ορίζοντα πρόβλεψης των μοντέλων που θα χρησιμοποιήσουμε. Έτσι, το αποτέλεσμα των μεθόδων θα συγκριθεί με τις κρυμμένες τιμές και θα έχουμε τη δυνατότητα να αξιολογήσουμε την ακρίβεια της προσέγγισής μας.

5.2.1 Αποεποχικοποίηση

Οι μέθοδοι αποεποχικοποίησης που συναντήσαμε στο Κεφάλαιο 2, απαιτούν η αρχική μας χρονοσειρά να έχει επαρκές πλήθος δεδομένων. Κάνουμε μια διαμέριση, λοιπόν, των δεδομένων μας σε δύο ομάδες: αυτές που έχουν πάνω από τρία έτη παρατηρήσεων και αυτές που δεν έχουν.

Ικανοποιούνται τα κριτήρια για την αποσύνθεση στη πρώτη ομάδα χρονοσειρών και εφαρμόζουμε το πολλαπλασιαστικό μοντέλο της κλασσικής μεθόδου αποσύνθεσης. Το γεγονός

ότι σκοπεύουμε να χρησιμοποιούμε τους δείκτες εποχιακότητας που θα προκύψουν ως το σημείο αναφοράς για να εντοπίσουμε τις συστάδες που υπάρχουν στο σύνολο των δεδομένων μας, καθιστά την διασφάλιση της υψηλής ακρίβειας ακόμα πιο σημαντική κατά τον υπολογισμό τους. Έτσι, χρησιμοποιούμε τη μέθοδο συρρίκνωσης James-Stein κατά τη διαδικασία υπολογισμού των λόγων εποχιακότητας. Επίσης, είναι απαραίτητο όλες οι χρονοσειρές να έχουν τους δείκτες τους στην ίδια κλίμακα και συνεπώς κανονικοποιούνται έτσι ώστε το άθροισμά τους να είναι ίδιο με το μήκος της εποχιακότητας.

Πρέπει, όμως, να μπορέσουμε να αποκτήσουμε και μια εικόνα για την εποχιακότητα των χρονοσειρών που στερούνται επαρκές πλήθος παρατηρήσεων. Έτσι, για τις υπόλοιπες χρονοσειρές, θα χρησιμοποιήσουμε τους εξής τύπους για να λάβουμε το σύνολο των 'ψεύδο-εποχιακών' δεικτών:

$$CxSxR = \frac{Y}{LRL}$$

Θέτουμε την τάση της χρονοσειράς να είναι ίση με την ευθεία γραμμικής παλινδρόμησης επί των δεδομένων. Θεωρώντας ότι η χρονοσειρά μας είναι πολλαπλασιαστική, αφαιρούμε την τάση από τα αρχικά δεδομένα διαιρώντας την αρχική χρονοσειρά με την ευθεία ελαχίστων τετραγώνων.

$$PS_t = mean(CxSxR_t, CxSxR + t + l_{season}, \ldots, CxSxR_{t+i*l_{season}}),$$

$$t + i * l_{season} < L, i = 1, 2, \ldots$$

$$PS_{norm} = \frac{PS * l_{season}}{\sum PS_t}$$

Όπου, ΠΣ είναι η σειρά της ψευδο-εποχιακότητας, l_{season} το μήκος της εποχιακότητας και L το μήκος της χρονοσειράς. Συνεπώς, ορίζουμε την ψευδο-εποχιακότητα ως το μέσο όρων των τιμών ανά περίοδο της χρονοσειράς, αφότου της έχουμε αφαιρέσει την τάση. Κατόπιν, κανονικοποιούμε το προηγούμενο αποτέλεσμα έτσι ώστε το άθροισμα των δεικτών ψευδο-εποχιακότητας να αθροίζει στο μήκος του εποχιακού μοτίβου.

5.3 Συσταδοποίηση Δεικτών Εποχιακότητας

Έχοντας το σύνολο των εποχιακών δεικτών μπορούμε πλέον να εφαρμόσουμε ανάλυση συστάδων (cluster analysis) σε αυτό. Μας ενδιαφέρει το σύνολο των χρονοσειρών των οποίων η εποχιακότητα προέκυψε με τις κλασικές μεθόδους αποσύνθεσης.

Για να το καταφέρουμε αυτό, θα χρησιμοποιήσουμε μεθόδους συσταδοποίησης (clustering) από το πεδίο της Μηχανικής Μάθησης (Machine Learning). Δύο μέθοδοι που ελέγξαμε είναι οι μέθοδοι k-means και DBSCAN.

5.3.1 K-means

Ο αλγόριθμος K-means γνωστός και ως ο αλγόριθμος του Lloyd, παράγει οικογένειες δεδομένων που χαρακτηρίζονται από παρόμοια διακύμανση, προσπαθώντας να ελαχιστοποιήσει

την ενδοοικογενειακό τετραγωνική διαφορά. Το πλήθος των συστάδων που θα προκύψουν από την εφαρμογή του αλγορίθμου είναι μία παράμετρος που ορίζει ο χρήστης.

Κάθε συστάδα (cluster) χαρακτηρίζεται από το κέντρο της (centroid), που ουσιαστικά είναι ο μέσος όρος όλων των στοιχείων που την αποτελούν. Βάσει αυτού, υπολογίζεται το κριτήριο της τετραγωνικής διαφοράς μια συστάδας, που προσπαθεί ο αλγόριθμος να ελαχιστοποιήσει.

Η μέθοδος χαρακτηρίζεται από τρία βασικά βήματα, που μετά το πρώτο ο αλγόριθμος επαναλαμβάνει τα άλλα δύο:

1. Επιλογή αρχικών κεντρών για τις συστάδες
2. Κατανομή των στοιχείων (εδώ χρονοσειρών) σε συστάδες, ανάλογα με την απόσταση από το κέντρο τους
3. Υπολογισμός νέου κέντρου κάθε συστάδας, ως ο μέσος όρος των στοιχείων που την αποτελούν

Ο αλγόριθμος σταματάει, όταν δεν έχουμε μεγάλη διαφορά μεταξύ των κεντρών που υπολογίστηκαν ανάμεσα από δύο επαναλήψεις.

5.3.2 DBSCAN

Ο αλγόριθμος DBSCAN (Density-Based Spatial Clustering of Applications with Noise) αντιμετωπίζει τις συστάδες ως περιοχές που χαρακτηρίζονται από μεγάλη πυκνότητα και χωρίζονται μεταξύ τους από περιοχές που δεν έχουν αυτό το χαρακτηριστικό. Έτσι, σε αντίθεση με τον K-means μπορεί να παράξει συστάδες που δεν χρειάζεται να είναι σε κυρτές περιοχές του χώρου.

Στον αλγόριθμο DBSCAN δεν χρειάζεται να οριστεί από πριν το πλήθος των συστάδων, ενώ αντιμετωπίζει κάποια στοιχεία του συνόλου ως ακραίες τιμές που δεν κατατάσσει σε καμία συστάδα. Όμως, πρέπει να οριστεί τι εννοούμε πυκνή περιοχή και αυτό γίνεται με δύο παραμέτρους. Ένα βασικό στοιχείο της μεθόδου είναι τα βασικά δείγματα core samples, δηλαδή ένα δείγμα από το σύνολο των δεδομένων μας που έχει τουλάχιστον ν το πλήθος άλλα στοιχεία του συνόλου, το πολύ σε απόσταση ϵ από αυτό. Τα ν και ϵ είναι οι παράμετροι που ορίζουν πόσο πυκνό πρέπει να είναι ένα cluster.

5.3.3 Επιλογή και εφαρμογή μεθόδου

Ζητούμενο σε αυτό το βήμα της μεθοδολογίας είναι να ελέγξουμε αν πράγματι υπάρχει συνάφεια στην εποχιακή συμπεριφορά των χρονοσειρών που έχουμε στη διάθεσή μας και να κάνουμε συστάδες βάσει αυτού. Συνεπώς, η ανάλυση συστάδων γίνεται με χρήση του αλγορίθμου DBSCAN.

Εντοπίζουμε με πειραματισμό τις κατάλληλες παραμέτρους για τον αλγόριθμο. Κριτήριο μας είναι τα cluster που προκύπτουν να μην έχουν μικρό αριθμό, έτσι ώστε να μπορεί να παραχθεί αντιπροσωπευτική εποχιακότητα μέσα από αυτό.

Έχοντας παράξει τα clusters με εφαρμογή του αλγορίθμου, πρέπει να δούμε αν οι υπόλοιπες χρονοσειρές βάσει των ψευδο-εποχιακών δεικτών, θα μπορούσαν να ανήκουν σε αυτά. Ένας τρόπος για να το διαπιστώσουμε αυτό είναι να το ελέγξουμε βάσει του κριτηρίου των τετραγώνων του K-means. Έτσι, ελέγχουμε ποια είναι η μέγιστη τετραγωνική απόσταση μέσα σε κάθε συστάδα από το κέντρο της. Αν η απόσταση της ψευδο-εποχιακότητας μιας χρονοσειράς έχει μικρότερη απόσταση από το κέντρο της συστάδας από αυτή που υπολογίσαμε, τότε την κατατάσσουμε σε αυτή.

Τελικώς προκύπτει ένα υποσύνολο των χρονοσειρών που δεν είχαν αρκετές παρατηρήσεις να προσεγγίζουν την εποχιακή συμπεριφορά άλλων που είχαν. Έτσι μπορούμε να περιμένουμε ότι πράγματι είναι εποχιακές χρονοσειρές και χρησιμοποιώντας το κέντρο της συστάδας που την κατατάξαμε, να την αποεποχικοποιήσουμε. Μένει να δούμε αν προβλέποντας κατά αυτό τον τρόπο αποεποχικοποιημένη χρονοσειρά μπορούμε να έχουμε καλύτερα αποτελέσματα από το να κρίναμε αυτές τις χρονοσειρές μη εποχιακές και να τις προεκτείναμε ως τέτοιες.

5.4 Πρόβλεψη

Για την πρόβλεψη, τόσο των αποεποχικοποιημένων χρονοσειρών όσο και των αρχικών τους δεδομένων θα χρησιμοποιήσουμε τις μεθόδους που είδαμε στο Υποκεφάλαιο 3.3. Συγκεκριμένα θα χρησιμοποιήσουμε τις εξής μεθόδους:

- Naive
- Γραμμική Παλινδρόμηση (LRL)
- Απλή Εκθετική Εξομάλυνση (SES)
- Εκθετική Εξομάλυνση Γραμμικής Τάσης (Holt Exponential Smoothing)
- Εκθετική Εξομάλυνση Φθίνουσας Τάσης (Damped Exponential Smoothing)
- Κλασική Μέθοδο Θ (Theta Classic)

Για να βρεθούν οι βέλτιστες τιμές των παραμέτρων για κάθε μία από τις μεθόδου που εφαρμόζουμε εδώ, χρησιμοποιήθηκε το μέσο τετραγωνικό σφάλμα MSE. Συγκεκριμένα, ελέχθησαν όλες οι δυνατές τιμές των παραμέτρων με βήμα 0.01 και εκείνες που παρήγαγαν το μικρότερο μέσο τετραγωνικό εντός-του-δείγματος σφάλμα (out-of-sample error) χρησιμοποιήθηκαν για την επέκταση της χρονοσειράς στο μέλλον.

Οι παραπάνω μέθοδοι εφαρμόστηκαν δύο φορές. Μία αφότου αποεποχικοποιήσαμε τα δεδομένα και μία στην αρχική τους μορφή. Στη πρώτη περίπτωση ενσωματώθηκε στο μοντέλο της πρόβλεψης και η εποχιακότητα της συστάδας στην οποία άνηκε η εκάστοτε χρονοσειρά.

5.5 Αξιολόγηση Πρόβλεψης

Το σημαντικότερο βήμα της διαδικασίας, καθότι αξιολογώντας της κάθε προσέγγιση και συγκρίνοντας τα αποτελέσματα μπορούμε να αποφανθούμε αν πράγματι η λύση του προβλήμα-

τος για τις μικρές εποχιακές χρονοσειρές, που προτείνεται στη παρούσα διπλωματική, προσφέρει βελτίωση στην ακρίβεια.

Στο πρόβλημα αυτό εφαρμόζουμε τον δείκτη ακριβείας σε ένα μεγάλο πλήθος χρονοσειρών που μπορεί να χαρακτηρίζεται από διαφορετικά επίπεδα στις χρονοσειρές που εμπεριέχει. Γι΄ αυτό τον λόγο χρειαζόμαστε ένα σφάλμα που είναι στην ίδια κλίμακα για όλες τις χρονοσειρές και παράγει διαφορετικά αποτελέσματα ανάλογα με το μέγεθος των παρατηρήσεων.

Επίσης, η χρονοσειρές, λόγω και της εποχιακότητας που τις χαρακτηρίζει, θα έχουν τιμές σε κάποιες χρονικές περιόδους που είναι πιθανό να βρίσκονται πολύ κοντά στο μηδέν. Ο δείκτης που θα χρησιμοποιήσουμε, λοιπόν, δε πρέπει να είναι ευαίσθητος στις μικρές τιμές.

Ένας δείκτης που μπορούμε να ορίσουμε που ικανοποιεί τα προαναφερθέντα κριτήρια είναι ο κανονικοποιημένος δείκτης μέσο απόλυτου σφάλματος που δίνεται από τον ακόλουθο τύπο:

$$MAE_{norm} = mean(e_i)/mean(Y)$$

Βάσει, λοιπόν, αυτού του δείκτη βλέπουμε αν η μέθοδος μας παράγει καλύτερα αποτελέσματα ανά μέθοδο. Αυτό μπορούμε να το δούμε τόσο στο κατά μέσο όρο σφάλμα ανά μέθοδο όσο και στο ποσοστό των χρονοσειρών που παρουσίασαν πιο ακριβείς προβλέψεις με τη προτεινόμενη προσέγγιση.

Χρησιμοποιώντας τη πληροφορία που λάβαμε από το τελευταίο βήμα, μπορούμε να δούμε πώς θα μπορούσαμε να βελτιώσουμε τη μεθοδολογία εν συνόλω.

Κεφάλαιο 6

Πειραματική Εφαρμογή

6.1 Δεδομένα

Στη διάθεση μας έχουμε ένα σύνολο δεδομένων που περιγράφουν την στάθμη υγρού φυσικού αερίου σε ένα σύνολο δεξαμενών στη Γαλλία. Ο αρχικός σκοπός της ανάλυσης ήταν να μπορέσουμε να προβλέψουμε πότε αυτές οι δεξαμενές θα αδειάσουν λόγω της κατανάλωσης έτσι ώστε η εταιρεία να φροντίσει να ανεφοδιαστούν. Με αυτό τον τρόπο θα αποφευχθεί οποιαδήποτε πρόβλημα με τους καταναλωτές αλλά συγχρόνως θα επιτραπεί στην εταιρεία να σχεδιάσει κατά βέλτιστο τρόπο την διαδικασία αναπλήρωσης υγρού φυσικού αερίου στις δεξαμενές.

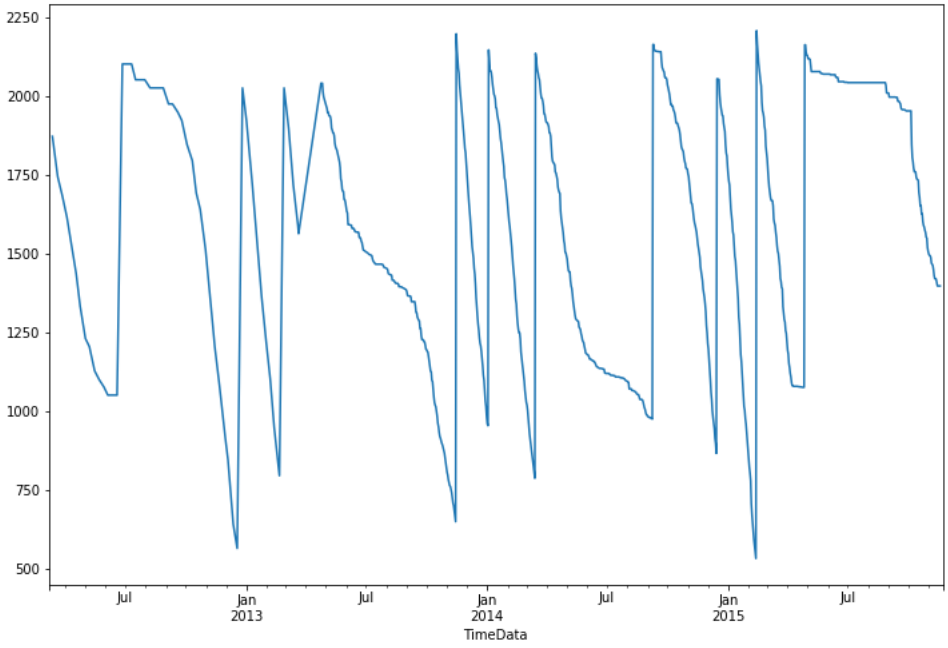
Τα δεδομένα ξεκινούν τον Αύγουστο του 2012 και συνεχίζουν ως και τον Νοέμβρη του 2015. Η συχνότητα μεταξύ των παρατηρήσεων κυμαίνεται από μερικές ώρες σε αρκετές μέρες μέχρι την επόμενη παρατήρηση. Περιγράφουν όγκο, άρα και η μονάδα που τα περιγράφει είναι αντίστοιχα μονάδα μέτρησης όγκου.

6.2 Προετοιμασία Χρονοσειρών

Παρατηρούμε ότι τα δεδομένα δεν χαρακτηρίζονται από σταθερή συχνότητα δειγματοληψίας, ούτε μεταξύ του αλλά ούτε κατά μήκος της χρονοσειράς που περιγράφει καθένα από αυτά. Χαρακτηριστικό παράδειγμα είναι χρονοσειρές που μπορεί για κάποιες ημέρες να έχουν πολλές παρατηρήσεις, ενώ σε άλλες να ακολουθεί μεγάλο διάστημα μέχρι την επόμενη μέρα που έχουμε δεδομένα.

Τα δεδομένα προέκυψαν από αισθητήρες εντός των δεξαμενών, που παρήγαγαν δεδομένα σε διαφορετικές στιγμές για κάθε δεξαμενή. Επίσης, υπήρχε περίπτωση σφάλματος στη μέτρηση, με τη στάθμη να μετράται μεγαλύτερη από κάποια προηγούμενη χρονική στιγμή, χωρίς να έχει συμβεί ανεφοδιασμός ανάμεσα.

Επίσης στα δεδομένα είναι εμφανείς οι χρονικές στιγμές που έγινε ανεφοδιασμός, καθώς βλέπουμε μια αλλαγή επιπέδου στη χρονοσειρά που τα περιγράφει, όπως φαίνεται στο Σχήμα 6.1. Στη συγκεκριμένη χρονοσειρά, παρατηρούμε ότι μεγάλο πλήθος από ανεφοδιασμούς στο χρονικό διάστημα για το οποίο έχουμε πληροφορία.



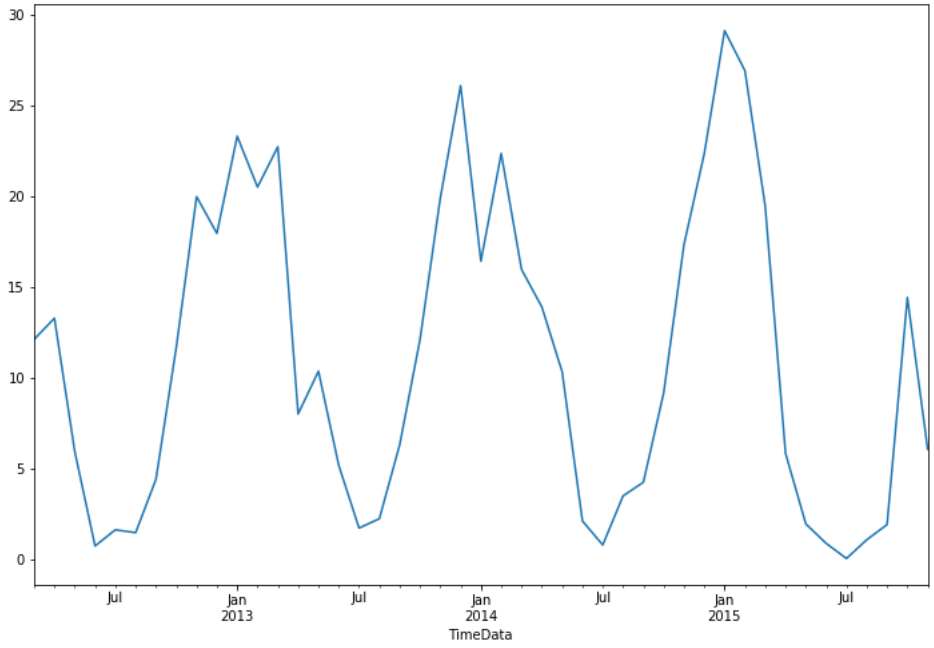
Σχήμα 6.1: Χρονοσειρά επιπέδου υγρού φυσικού αερίου σε δεξαμενή

Λόγω της φύσης της πληροφορίας που έχουμε, γνωρίζουμε ότι καμία στιγμή, σε ενεργές δεξαμενές δε μπορούμε να έχουμε μηδενικές τιμές και προφανώς δεν μπορεί μία δεξαμενή να έχει αρνητικό όγκο φυσικού αερίου.

Τα παραπάνω μας οδηγούν στο να κάνουμε τις εξής ενέργειες για τη προετοιμασία των χρονοσειρών:

- Μετατρέπουμε τυχούσες μη θετικές τιμές σε κενές τιμές
- Φέρνουμε τα δεδομένα σε επίπεδο ημέρας κρατώντας, στις περιπτώσεις πολλαπλών παρατηρήσεων, τη μικρότερη από αυτές
- Συμπληρώνουμε τις κενές τιμές με τη μέθοδο της γραμμικής παρεμβολής βάσει του χρόνου

Στόχος μας, όμως, εν τέλει είναι να προβλέψουμε πότε η δεξαμενή θα αδειάσει. Για να το εντοπίσουμε αυτό πρέπει να εστιάσουμε τη προσοχή μας στο ρυθμό κατανάλωσης της χρονοσειράς. Έτσι, χρησιμοποιούμε πρώτες διαφορές στα δεδομένα για να δούμε τι κατανάλωση είχαμε μέσα σε μία μέρα. Σε αυτό το σημείο, παρατηρούμε ότι τις ημέρες που είχαμε ρεφιλλς η χρονοσειρά μας παρουσιάζει ακραίες τιμές που δεν ανταποκρίνονται στον φυσιολογικό ρυθμό κατανάλωσης που χαρακτηρίζει τη δεξαμενή. Συνεπώς, θέτουμε εκείνη την ημέρα σας κενή τιμή και ξαναχρησιμοποιούμε τη μέθοδο της παρεμβολής για να τη διαχειριστούμε.



Σχήμα 6.2: Χρονοσειρά μέσης μηνιαίας κατανάλωσης

Επίσης, σκοπός μας είναι να γνωρίζουμε αρκετά πριν πότε μέλλεται μία δεξαμενή να αδειάσει. Βγάζει λοιπόν περισσότερο νόημα να χρησιμοποιήσουμε τα μοντέλα μας για να προβλέψουμε τη μηνιαία κατανάλωση. Έτσι, φέρνουμε τα δεδομένα σε επίπεδο μήνα, παίρνοντας τον μέσο όρο της ημερήσιας κατανάλωσης και λαμβάνουμε τη χρονοσειρά της μηνιαίας κατανάλωσης.

Στο Σχήμα 6.2 βλέπουμε τη νέα μορφή των δεδομένων για την ίδια δεξαμενή. Η ετήσια εποχιακή συμπεριφορά της χρονοσειράς είναι πλέον εμφανής.

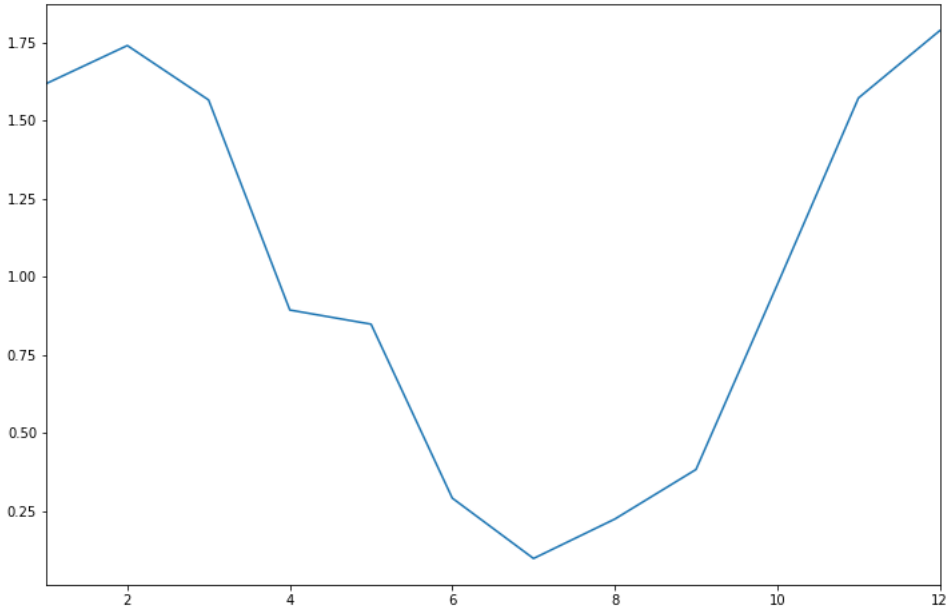
Τελικώς, για να μπορούμε αξιολογήσουμε τη μέθοδο πρόβλεψης, αποκρύβουμε τις 6 τελευταίες παρατηρήσεις για να τις χρησιμοποιήσουμε ως ορίζοντα πρόβλεψης

6.3 Αποεποχικοποίηση

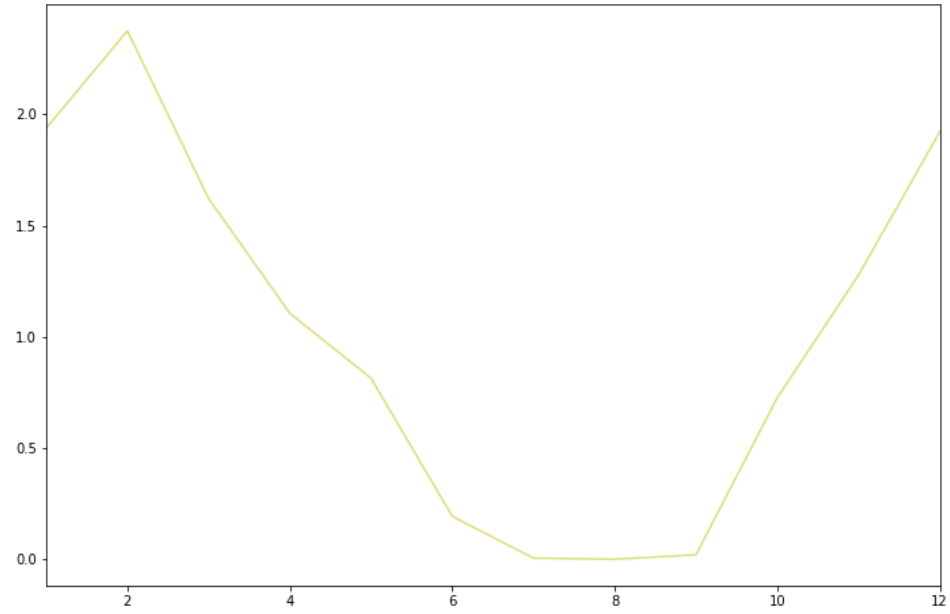
Όπως περιγράψαμε στο κεφάλαιο της μεθοδολογίας, χωρίζουμε τις χρονοσειρές σε δύο κατηγορίες: αυτές που έχουν πάνω από τρία έτη παρατηρήσεις και αυτές που έχουν λιγότερες.

Ακολουθώντας έτσι την μεθοδολογία της κλασσικής μεθόδου αποσύνθεσης με το πολλαπλασιαστικό μοντέλο και της συρρίκνωσης συντελεστών με τη μέθοδο James-Stein λαμβάνουμε τους εποχιακούς δείκτες για τις χρονοσειρές τις πρώτης ομάδας. Μπορούμε να δούμε την εποχιακότητα της χρονοσειράς των προηγούμενων σχημάτων στο Σχήμα 6.3.

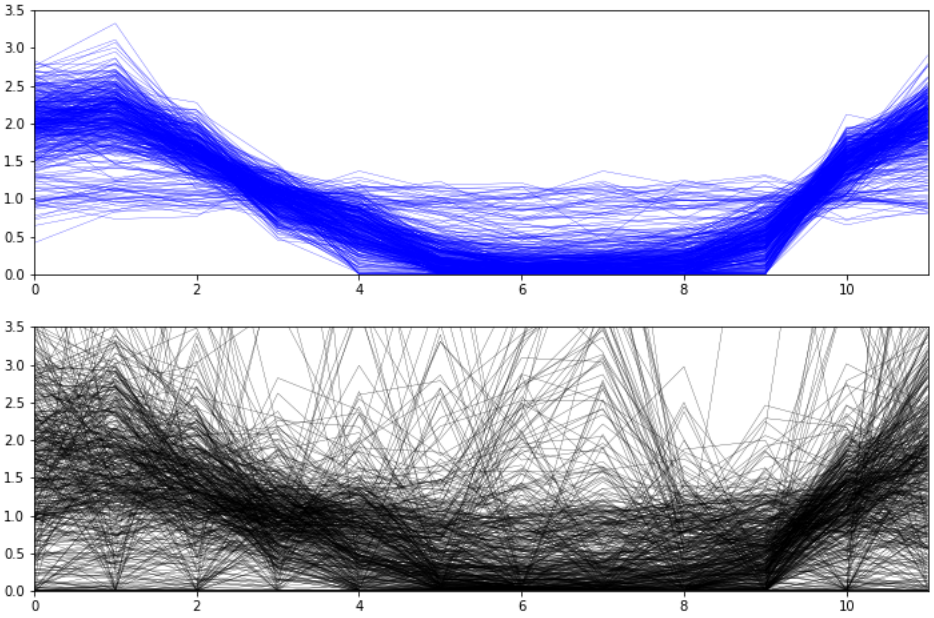
Αντίστοιχα στο Σχήμα 6.4 βλέπουμε το αποτέλεσμα του υπολογισμού της ψευδο-εποχιακότητας σε μία χρονοσειρά.



Σχήμα 6.3: Δείκτες εποχιακότητας



Σχήμα 6.4: Δείκτες ψευδο-εποχιακότητας



Σχήμα 6.5: Συσταδοποίηση δεικτών εποχιακότητας

Παρατηρούμε ότι τα αποτελέσματα των δύο κινούνται στην ίδια κλίμακα, και συνεπώς μπορούν να είναι συγκρίσιμα και να μας επιτρέψουν να φτιάξουμε τις συστάδες κατά τον τρόπο που περιγράψαμε στο προηγούμενο κεφάλαιο.

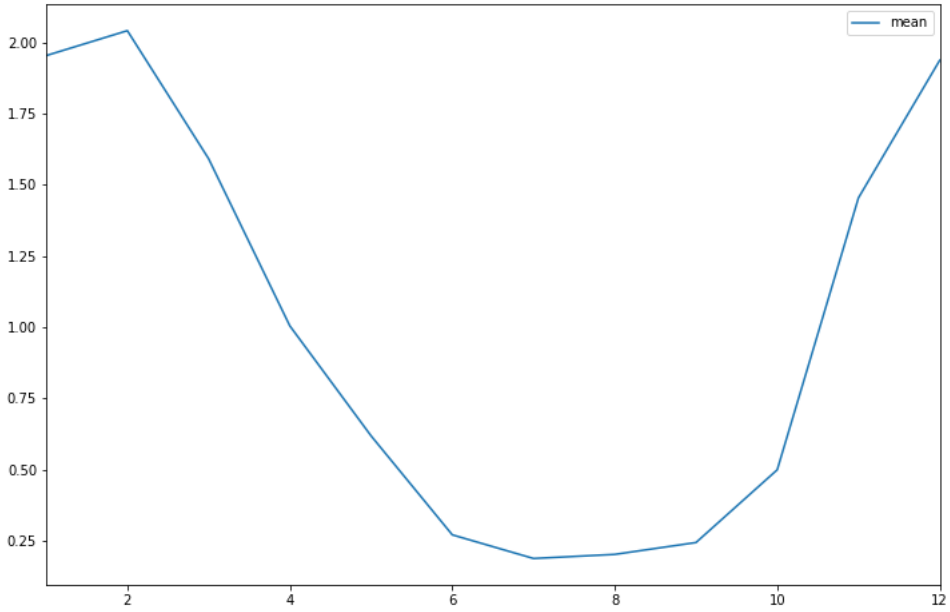
6.4 Συσταδοποίηση Δεικτών Εποχιακότητας

Χρησιμοποιώντας τον αλγόριθμο DBSCAN στους δείκτες εποχιακότητας των χρονοσειρών με επαρκείς παρατηρήσεις προέκυψε η διαμέριση που φαίνεται στο Σχήμα 6.5, με το τελευταίο σύνολο να μην μπορεί να δημιουργήσει συστάδες.

Βλέπουμε ότι ένα μεγάλο πλήθος χρονοσειρών δεν παρουσιάζει συνάφεια με άλλες χρονοσειρές. Δοκιμάζοντας διαφορετικές τιμές στις παραμέτρους του αλγορίθμου, προκύπτουν περισσότερες συστάδες με μεγαλύτερη πυκνότητα αλλά με μικρότερο πλήθος χρονοσειρών η κάθε μία. Αφότου, όμως σκοπεύουμε να χρησιμοποιήσουμε το αποτέλεσμα για μάθουμε την εποχιακή συμπεριφορά του συνόλου των δεδομένων μας πρέπει να αποφύγουμε την υπερπροσαρμογή (overfitting) σε συγκεκριμένα μοτίβα.

Το επόμενο βήμα είναι να υπολογίσουμε το κέντρο της συστάδας, το οποίο λάβαμε υπολογίζοντας τη μέση τιμή των δεικτών εποχιακότητας που ανήκουν σε αυτή. Το αποτέλεσμα φαίνεται στο Σχήμα 6.6.

Υπολογίζουμε, κατόπιν τη μέγιστη τετραγωνική διαφορά μεταξύ των δεικτών εποχιακότητας του cluster και του κέντρου του. Ελέγχουμε κάθε σύνολο δεικτών ψευδο-εποχιακότητας



Σχήμα 6.6: Δείκτες εποχιακότητας κέντρου συστάδας

αν έχει απόσταση μικρότερη της μέγιστης διαφοράς που υπολογίσαμε. Οι χρονοσειρές που πληρούν αυτό το κριτήριο φαίνονται στο Σχήμα 6.7 μαζί με το κέντρο της συστάδας.

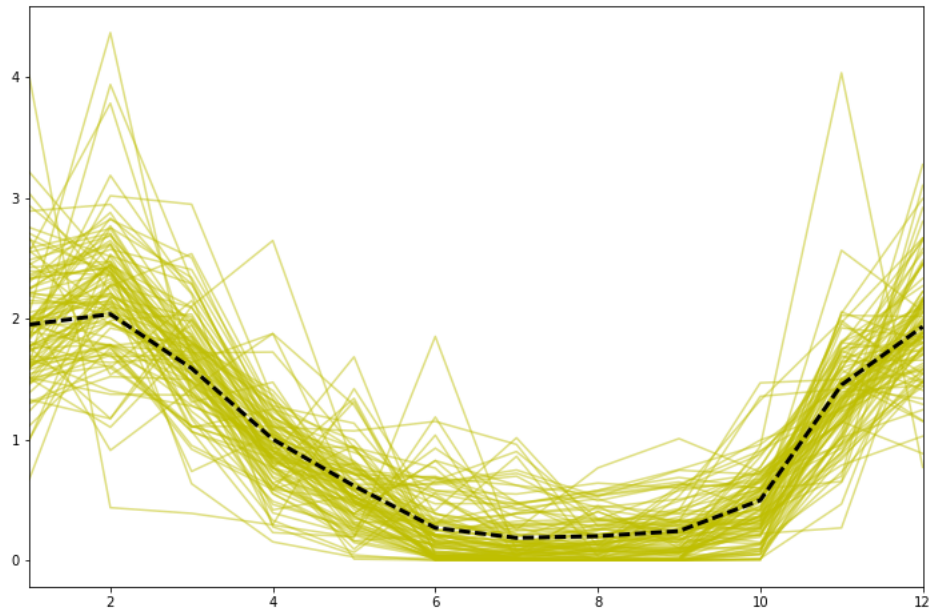
6.5 Πρόβλεψη

Έχοντας εντοπίσει τις χρονοσειρές που μας ενδιαφέρουν, τις προεκτείνουμε στο μέλλον με δύο τρόπους:

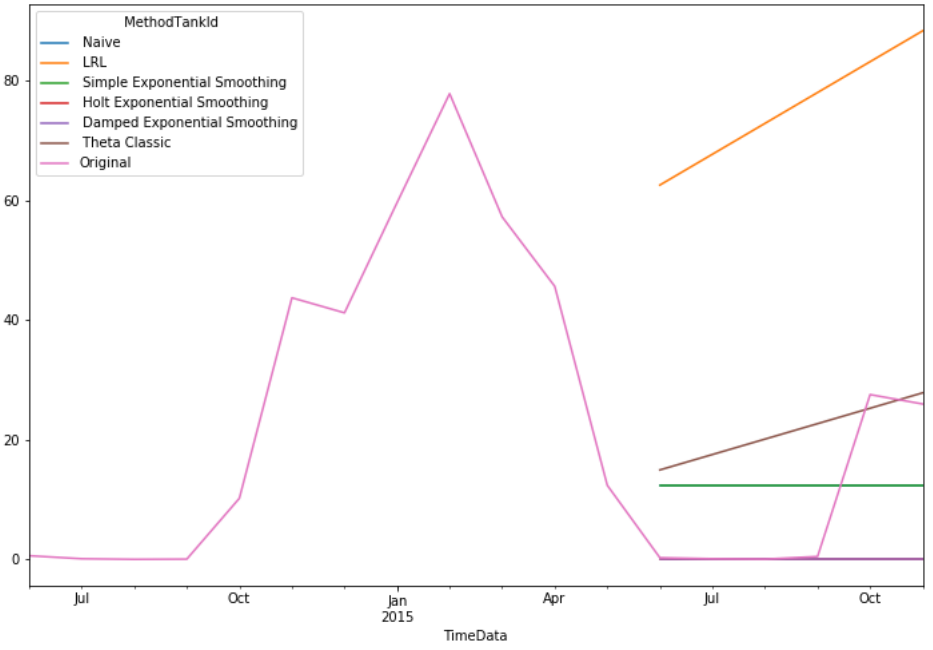
1. Πρόβλεψη της αρχικής χρονοσειράς
2. (α′) Αποεποχικοποίηση με τους κεντρικούς δείκτες
   (β′) Πρόβλεψη
   (γ′) Επαναεποχικοποίηση

Βλέπουμε τα αποτελέσματα της πρώτης μεθόδου στο Σχήμα 6.8 για μία από τις χρονοσειρές. Για την ίδια, τα αποτελέσματα της προτεινόμενης μεθόδου φαίνονται στα Σχήματα 6.9 και 6.10, για την αποεποχικοποιημένη και τελική πρόβλεψη αντίστοιχα.

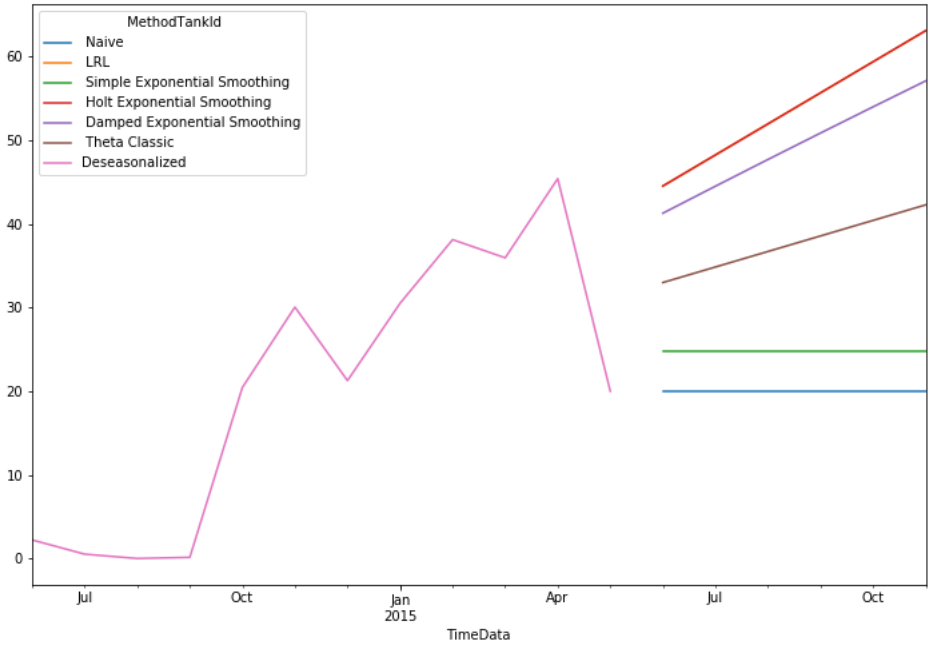
Πρέπει να σημειώσουμε ότι αφότου λάβαμε τα αποτελέσματα των προβλέψεων, οι τυχούσες αρνητικές τιμές μετατράπηκαν σε μηδενικές. Αυτό έγινε καθώς δε δύναται να έχουμε αρνητική κατανάλωση χωρίς εξωγενείς παράγοντες (π.χ. αναπλήρωση φυσικού αερίου), που μάλιστα τους έχουμε αφαιρέσει από την ανάλυση μας.



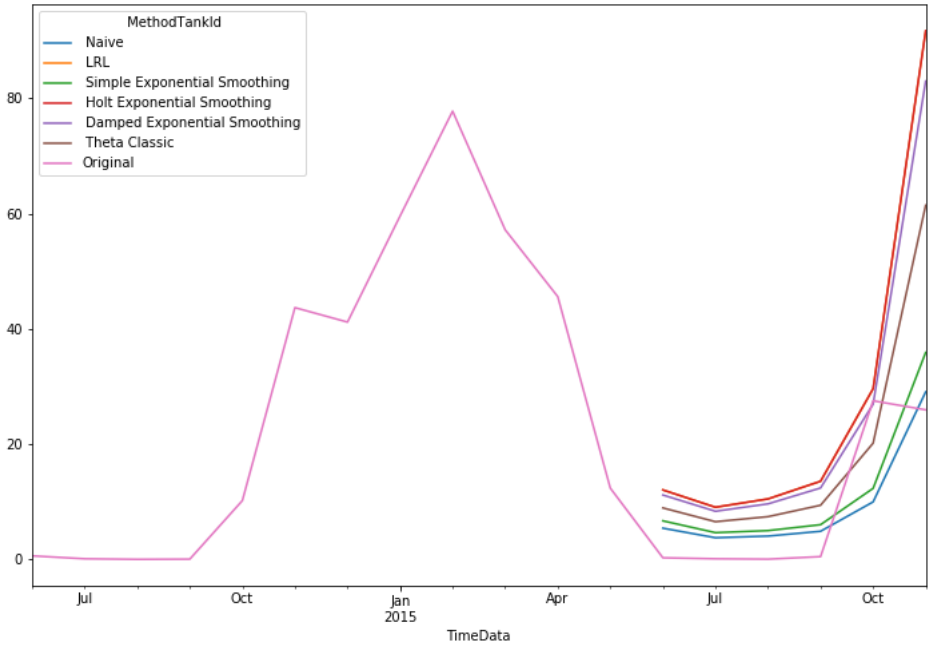
Σχήμα 6.7: Δείκτες ψευδο-εποχιακότητας συστάδας



Σχήμα 6.8: Πρόβλεψη κλασικής προσέγγισης



Σχήμα 6.9: Πρόβλεψη αποεποχικοποιημένης χρονοσειράς



Σχήμα 6.10: Τελική πρόβλεψη προτεινόμενης προσέγγισης

6.6 Αξιολόγηση Μεθόδου

Όπως περιγράψαμε στο τελευταίο βήμα του τελευταίου κεφαλαίου, μετράμε την ακρίβεια της μεθόδου με το κανονικοποιημένο μέσο απόλυτο σφάλμα. Μπορούμε να δούμε τα αποτελέσματα εν συνόλω για τη κάθε μέθοδο στον Πίνακα 6.1. Η πρώτη στήλη δείχνει τις μεθόδους πρόβλεψης, η δεύτερη τον δείκτη ακρίβειας για τη κλασική προσέγγιση πρόβλεψης χρονοσειρών με μικρό ιστορικό, η τρίτη τον δείκτη ακριβείας για τη προτεινόμενη προσέγγιση και τελικώς φαίνεται το ποσοστό των υπό εξέταση χρονοσειρών που είχαμε καλύτερα αποτελέσματα με τη νέα μέθοδο.

Μέθοδος Πρόβλεψης	**Κλασική**	**Προτεινόμενη**	**Ποσοστό (%)**
Naive	1.01	**0.81**	70.97
LRL	3.15	**0.99**	94.74
Simple Exponential Smoothing	1.31	**1.02**	74.74
Holt Exponential Smoothing	1.18	**0.89**	72.63
Damped Exponential Smoothing	1.05	**0.93**	73.03
Theta Classic	1.22	**0.96**	73.68

Πίνακας 6.1: Ακρίβεια Μεθόδων

Τα αποτελέσματα μας δείχνουν ότι για το κατά πολύ μεγαλύτερο μέρος των χρονοσειρών τις προεκτείναμε με μεγαλύτερη ακρίβεια χρησιμοποιώντας ανάλυση συστάδων για αποεποχικοποίηση. Όλες οι μέθοδοι πρόβλεψης είχαν καλύτερα αποτελέσματα σε σύγκριση με τη κλασική προσέγγιση.

Κεφάλαιο 7

Επίλογος

Σε αυτό το κεφάλαιο θα σκιαγραφήσουμε την πορεία ανάλυσης των δεδομένων που μας οδήγησαν εν τέλει στη διαδικασία της πρόβλεψης και επικύρωσης της υπόθεσής μας. Θα συζητήσουμε τα αποτελέσματα που προέκυψαν και τι περαιτέρω βήματα μπορούν να γίνουν για να ισχυροποιήσουν αυτά τα αποτελέσματα.

7.1 Σύνοψη και συμπεράσματα

Στις κλασική μεθοδολογία πρόβλεψης μιας χρονοσειράς, αφού την προετοιμάσουμε την αποσυνθέτουμε στα βασικά της χαρακτηριστικά: τη τάση, τον κύκλο, την τυχαιότητα και την εποχιακότητα. Αφότου έχουμε υπολογίσει τους εποχιακούς δείκτες, τους χρησιμοποιήσουμε για να απαλλάξουμε τα αρχικά δεδομένα από τις αλλαγές που υφίστανται λόγω της περιόδου που βρίσκονται μέσα στο μήκος του μοτίβου της εποχιακότητας. Έτσι προκύπτει μία πιο σταθερή πλέον χρονοσειρά, αφού της έχουμε αφαιρέσει τις προβλεπόμενες διακυμάνσεις. Αυτή τη χρονοσειρά, λοιπόν, χρησιμοποιούμε ως είσοδο στο μοντέλο της πρόβλεψής μας. Κατόπιν, λαμβάνουμε το αποτέλεσμα του μοντέλου και ενσωματώνουμε την εποχιακή συμπεριφορά για να προκύψει η τελική πρόβλεψη.

Το πρόβλημα εμφανίζεται όταν η υπό εξέταση χρονοσειρά δεν χαρακτηρίζεται από επαρκώς μεγάλο ιστορικό για να μας επιτρέψει να εξάγουμε το μοτίβο της εποχιακότητάς της. Σε αυτή τη περίπτωση, η συνηθισμένη προσέγγιση είναι να αντιμετωπίζουμε τη χρονοσειρά ως μη εποχιακή και να τη προβλέπουμε χωρίς να κάνουμε ανάλυση εποχιακότητας. Όμως, οι εποχιακές διακυμάνσεις, σε περίπτωση που η σειρά περιέγραφε πράγματι ένα εποχιακό μέγεθος, συνεχίζουν να χαρακτηρίζουν τα δεδομένα και τελικώς επηρεάζουν την ακρίβεια των μοντέλων.

Παράλληλα, όσο περνάνε τα χρόνια το πλήθος των δεδομένων που εν γένει έχουμε στη διάθεσή μας αυξάνεται εκθετικά. Αυτό το γεγονός ανοίγει δρόμους σε νέες προσεγγίσεις για να προσπαθήσουμε να αυξήσουμε την ακρίβεια της διαδικασίας της πρόβλεψης. Έτσι, θεωρώντας ότι έχουμε ένα μεγάλο πλήθος χρονοσειρών που περιγράφουν παρόμοιας φύσης δεδομένα, υποθέσαμε ότι μπορούμε να εκμαιεύσουμε πληροφορία για την εποχιακή συμπεριφορά των δεδομένων της συγκεκριμένης φύσης.

Δοκιμάσαμε την υπόθεσή μας σε ένα σύνολο δεδομένων χρονοσειρών από τα δεδομένα για

τον όγκο υγρού φυσικού αερίου σε δεξαμενές. Στόχος ήταν να μπορέσουμε να εκτιμήσουμε σε ποια χρονική στιγμή στο μέλλον κάθε δεξαμενή θα αδειάσει με σκοπό να το αποτρέψουμε, έτσι ώστε να μην υπάρξει πρόβλημα στους πελάτες της εταιρίας που παρέχει το φυσικό αέριο αλλά και να της δώσουμε τη γνώση που χρειάζεται για να σχεδιάσει βέλτιστα τον ανεφοδιασμό των δεξαμενών.

Πρώτο βήμα στην ανάλυση ήταν να καθαρίσουμε τα δεδομένα μας και να τα φέρουμε στη μορφή που θα μας βοηθήσει να πετύχουμε τον στόχο μας. Φροντίσαμε, λοιπόν, να τα φέρουμε όλα σε ημερήσια συχνότητα και να διαχειριστούμε τις κενές και μηδενικές τιμές. Κατόπιν, είδαμε ότι για να προβλέψουμε πότε θα καταναλωθεί όλο το υγρό φυσικό αέριο στις δεξαμενές μας ενδιαφέρει ο ρυθμός κατανάλωσης σε κάθε μία από αυτές. Έτσι, παίρνοντας τις διαφορές μεταξύ δύο διαδοχικών ημερών στον όγκο του υγρού, λάβαμε την ημερήσια κατανάλωση. Σκοπεύοντας να προβλέψουμε μεσοπρόθεσμα το πότε θα αδειάσει μία δεξαμενή, υπολογίσαμε τον μέσο όρο της ημερήσια κατανάλωσης για κάθε μήνα και προέκυψε η μηνιαία χρονοσειρά της μέσης κατανάλωσης.

Τελειώνοντας, λοιπόν, το πρώτο βήμα της ανάλυσης καταλήξαμε να έχουμε ένα σύνολο χρονοσειρών που περιγράφουν την μέση μηνιαία κατανάλωση ανά δεξαμενή. Όλες αυτές η χρονοσειρές μετράνε παρόμοια μεγέθη, ενώ ανάμεσά τους υπάρχει ένα ποσοστό που δεν έχουν επαρκή δεδομένα για να εξάγουμε την εποχιακή τους συμπεριφορά. Από την άλλη, η κατανάλωση φυσικού αερίου, που είναι συνδεδεμένη άμεσα με την θερμοκρασία αφότου μία κύρια χρήση του είναι η θέρμανση, περιμένουμε να χαρακτηρίζεται από εποχιακές διακυμάνσεις στο μήκος του χρόνου.

Έτσι, διαμερίσαμε τις χρονοσειρές σε αυτές που μπορούμε να χρησιμοποιήσουμε τις κλασικές μεθόδους αποσύνθεσης και σε αυτές που δεν μπορούμε. Στη πρώτη κατηγορία τις εφαρμόσαμε κανονικά και λάβαμε τους δώδεκα ετήσιους δείκτες εποχιακότητας για κάθε μία εξ΄ αυτών. Για τη δεύτερη κατηγορία αφαιρέσαμε από την χρονοσειρά τις τιμές της ευθείας γραμμικής παλινδρόμησης που αντιστοιχούσαν σε κάθε παρατήρηση, και από το αποτέλεσμα πήραμε τον μέσο όρο των τιμών για κάθε περίοδο. Προέκυψαν, συνεπώς, δώδεκα τιμές που ονομάσαμε δείκτες ψευδο-εποχιακότητας.

Επόμενο βήμα ήταν να ελέγξουμε αν πράγματι υπήρχαν κοινά μοτίβα στην εποχιακότητα των χρονοσειρών της πρώτης κατηγορία και να τα εντοπίσουμε. Για αυτό το λόγο χρησιμοποιήσαμε τον αλγόριθμο συσταδοποίησης DBSCAN. Βρήκαμε, ότι αν και πολλές χρονοσειρές δεν ακολουθούσαν συγκεκριμένο μοτίβο, υπήρχε ένα σύνολο εξ αυτών που παρουσίαζε συνάφεια στους εποχιακούς του δείκτες. Γι΄ αυτό το μοτίβο υπολογίσαμε την μέση εποχιακότητα.

Για να ελέγξουμε αν οι χρονοσειρές με μικρό ιστορικό ανήκουν στην συστάδα που προέκυψε, χρησιμοποιήσαμε το κριτήριο με το οποίο ο αλγόριθμος k-means κατατάσσει τα δεδομένα που πραγματεύεται σε συστάδες. Έτσι, για το σύνολο των χρονοσειρών με συνάφεια στους δείκτες εποχιακότητας, υπολογίσαμε τη μέγιστη μέση τετραγωνική διαφορά του κέντρου από τους δείκτες των υπόλοιπων του χρονοσειρών. Αν η μέση τετραγωνική διαφορά μεταξύ των δεικτών ψευδο-εποχιακότητας και του κέντρου ήταν μικρότερη από το αποτέλεσμα, θεωρήσαμε ότι μπορεί να αντιπροσωπευθεί η εποχιακότητα των χρονοσειρών τους από αυτή του κέντρου.

Έχοντας στη διάθεση μας ένα σύνολο χρονοσειρών που πληρούσε πλήρως τα κριτήρια που

θέσαμε για την υπόθεσή μας περάσαμε στη διαδικασία της πρόβλεψης. Αρχικά, προβλέψαμε τις χρονοσειρές κατά την κλασική προσέγγιση: τις προεκτείναμε στο μέλλον βάσει των αρχικών τους δεδομένων. Έπειτα, αποεποχικοποιήσαμε αυτές τις χρονοσειρές, με τους μέσους δείκτες εποχιακότητας της συστάδας, και εφαρμόσαμε τα μοντέλα πρόβλεψης κατόπιν. Στη δεύτερη περίπτωση, πολλαπλασιάσαμε πάλι τους δείκτες εποχιακότητας με το προϊόν των μοντέλων.

Σε ποσοστό μεγαλύτερο του 70% για κάθε μέθοδο πρόβλεψης είχαμε η προτεινόμενη μέθοδος να έχει μεγαλύτερη ακρίβεια, σημειώνοντας χαμηλότερη τιμή σφάλματος σε κάθε μία από τις μεθόδους που εξετάσαμε. Το αποτέλεσμα αυτό επαλήθευσε την υπόθεσή μας και έδειξε τη σημασία της διαχείρισης της εποχιακότητας των χρονοσειρών ακόμα και όταν αυτές έχουν μικρό ιστορικό.

Από τις μεθόδους που χρησιμοποιήσαμε μεγαλύτερη βελτίωση είδαμε στην ακρίβεια της μεθόδου της γραμμικής παλινδρόμησης. Παράλληλα, είχαμε το μεγαλύτερο ποσοστό των χρονοσειρών που πήγε καλύτερα η προτεινόμενη μέθοδος. Αυτό είναι λογικό, καθότι η μέθοδος υπολογίζει την τάση της χρονοσειράς και την προεκτείνει στο μέλλον. Δεδομένου ότι οι υπό εξέταση χρονοσειρές έχουν μικρό ιστορικό ενώ χαρακτηρίζονται από εποχιακή συμπεριφορά, η γραμμή ελαχίστων τετραγώνων επηρεάζεται από τις κορυφές τις χρονοσειράς και τις βλέπει σαν αύξηση της τάσης.

Τόσο η προτεινόμενη όσο και η κλασική προσέγγιση είχαν μεγαλύτερη ακρίβεια για τη μέθοδο Naive. Αυτό συμβαίνει διότι σε τόσο μικρό αριθμό δεδομένων η επιρροή της τυχαιότητας είναι πολύ μεγάλη κατά την πρόβλεψη.

7.2 Μελλοντικές επεκτάσεις

Στη παρούσα εργασία είδαμε ότι εκμεταλλευόμενοι την πληροφορία που μας δίνεται από έναν μεγάλο όγκο συναφών δεδομένων μπορούμε να εκτιμήσουμε την εποχιακή συμπεριφορά μικρότερων χρονοσειρών εξ αυτών.

Βέβαια, η μελέτη έγινε σε ένα συγκεκριμένο σύνολο χρονοσειρών ενεργειακών χρονοσειρών. Το πρώτο βήμα, λοιπόν, είναι να εξεταστούν και άλλα σύνολα δεδομένων. Ιδιαίτερο ενδιαφέρον, δε, θα έχει η μελέτη οικονομικών δεδομένων και η εφαρμογή στη διαχείριση αποθήκης.

Ένας άλλος παράγοντας που αξίζει να εξεταστεί είναι εναλλακτικές μέθοδοι αποσύνθεσης της χρονοσειράς. Αντί της κλασικής πολλαπλασιαστικής μεθόδου αποσύνθεσης, μπορεί να γίνει ανάλυση με το προσθετικό μοντέλο ή το μοντέλο STL, που είδαμε επίσης σε προηγούμενο κεφάλαιο. Επίσης, μπορεί να χρησιμοποιηθεί διαφορετική μέθοδος συρρίκνωσης συντελεστών για τους δείκτες εποχιακότητας, ενώ θα μπορούσε να εφαρμοστεί συρρίκνωση και στους δείκτες ψευδο-εποχιακότητας.

Όσον αναφορά την ανάλυση συστάδων, εκτός των DBSCAN και k-means θα μπορούσαν να χρησιμοποιηθούν και άλλοι αλγόριθμοι συσταδοποίησης. Επιπρόσθετα, στη περίπτωση που το σύνολο δεδομένων περιέχει και άλλες πληροφορίες που να δίνουν μεγαλύτερη διάσταση σε αυτά, όπως λόγου χάρη τοποθεσία ή είδος πελάτη, θα μπορούσε να χρησιμοποιηθεί αυτή σαν παράμετρος στην ανάλυση συστάδων.

Μεγάλο ενδιαφέρον θα παρουσίαζε επίσης, η σύγκριση της προτεινόμενης μεθόδου με άλλες προσεγγίσεις προέκτασης χρονοσειρών. Μπορεί να γίνει σύγκριση με μεθόδους πρόβλεψης με χρήση νευρωνικών δικτύων, μοντέλων ARIMA και με μπεϋζιανές τεχνικές προβλέψεων.

Βιβλιογραφία

[1] V. Assimakopoulos and K. Nikolopoulos. The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, 16(4):521–530, oct 2000.

[2] Derek W. Bunn and Angelos I. Vassilopoulos. Comparison of seasonal estimation methods in multi-item short-term forecasting. *International Journal of Forecasting*, 15(4):431 – 443, 1999.

[3] NumPy Developers. Numpy, 2017.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[5] David F. Findley, Brian C. Monsell, William R. Bell, Mark C. Otto, and Bor-Chung Chen. New capabilities and methods of the x-12-arima seasonal-adjustment program. *Journal of Business & Economic Statistics*, 16(2):127–152, 1998.

[6] Everette S. Gardner. Exponential smoothing: The state of the art. *Journal of Forecasting*, 4(1):1–28, 1985.

[7] Everette S. Gardner. Exponential smoothing: The state of the art—part ii. *International Journal of Forecasting*, 22(4):637 – 666, 2006.

[8] Jan G. De Gooijer and Rob J. Hyndman. 25 years of time series forecasting. *International Journal of Forecasting*, 22(3):443 – 473, 2006. Twenty five years of forecasting.

[9] Charles C. Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5 – 10, 2004.

[10] Rob Hyndman and Andrey V. Kostenko. Minimum sample size requirements for seasonal forecasting models. *Foresight: The International Journal of Applied Forecasting*, (6):12–15, 2007.

[11] Rob J Hyndman and George Athanasopoulos. Forecasting: Principles and practice, 2012.

[12] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679 – 688, 2006.

[13] W. James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379, Berkeley, Calif., 1961. University of California Press.

[14] Project Jupyter. Project jupyter, 2017.

[15] Marcin Kozak. "a dendrite method for cluster analysis" by caliński and harabasz: A classical work that is far too often incorrectly cited. *Communications in Statistics - Theory and Methods*, 41(12):2279–2280, 2012.

[16] Inc. Lambda Foundry and PyData Development Team. Pandas, 2017.

[17] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982.

[18] Don M. Miller and Dan Williams. Shrinkage estimators of time series seasonal factors and their effect on forecasting accuracy. *International Journal of Forecasting*, 19(4):669 – 684, 2003.

[19] K Nikolopoulos and V Assimakopoulos. *Generalizing the Theta model.* unknown, 2004.

[20] C. Carl Pegels. Exponential forecasting: Some new variations. *Management Science*, 15(5):311–315, 1969.

[21] Peter R. Winters. Forecasting sales by exponentially weighted moving averages. 6:324–342, 04 1960.

[22] scikit-learn developers. scikit-learn, 2017.

[23] Φώτιος Πετρόπουλος, Βασίλειος Ασημακόπουλος. *Επιχειρησιακές Προβλέψεις*. Συμμετρία, 2011.

[24] Forecasting & Strategy Unit. Σημειώσεις Μαθήματος Τεχνικών Προβλέψεων, 2017.

[25] Richard Withycombe. Forecasting with combined seasonal indices. *International Journal of Forecasting*, 5(4):547 – 552, 1989.

[26] Matplotlib Development Team ("MDT"). Matplotlib, 2017.

Παράρτημα Α΄

Τεχνικές λεπτομέρειες

Σε αυτό το κεφάλαιο θα περιγράψουμε τα εργαλεία προγραμματισμού και στατιστικής ανάλυσης που χρησιμοποιήθηκαν για την υλοποίηση της παρούσας διπλωματικής εργασίας. Χρησιμοποιήθηκαν δύο γλώσσες προγραμματισμού, με ένα μεγάλο πλήθος διαφορετικών βιβλιοθηκών και δύο προγραμματιστικά περιβάλλοντα.

Α΄.1 Γλώσσες προγραμματισμού

Για να υλοποιηθεί ο κώδικας επίλυσης του προβλήματος που άπτεται η εργασία χρησιμοποιήθηκαν δύο γλώσσες προγραμματισμού: η Python και η VB.NET.

Α΄.1.1 Python

Στη παρούσα εργασία χρησιμοποιήθηκε η γλώσσα Python καθότι προσφέρει μία πληθώρα εργαλείων για στατιστική ανάλυση και μηχανική μάθηση. Άλλωστε είναι από τις πιο διαδεδομένες γλώσσες γι΄ αυτό το σκοπό. Συγκεκριμένα χρησιμοποιήθηκαν οι βιβλιοθήκες που περιγράφονται ακολούθως, που είναι όλες ανοικτού κώδικα.

NumPy

Το πακέτο NumPy είναι βασικό πακέτο της γλώσσας Python που επιτρέπει προγραμματισμό για επιστημονική χρήση. Επιτρέπει τη διαχείριση δεδομένων μεγάλων διαστάσεων και προσφέρει μεγάλο πλήθος συναρτήσεων για την επεξεργασία τους.

Pandas

Το πακέτο Pandas είναι ιδανικό για την ανάλυση χρονοσειρών. Προσφέρει εύκολες στη χρήση δομές δεδομένων, και επιτρέπει την άμεση οπτικοποίηση και επεξεργασία τους.

Scikit-learn

Το Scikit-learn προσφέρει ένας πλήθος αλγορίθμων Μηχανικής Μάθησης. Συγχρόνως, παρέχει απλά και αποδοτικά εργαλεία για εξόρυξη και ανάλυση δεδομένων.

Matplotlib

Το Matplotlib είναι μία βιβλιοθήκη της Python που παράγει υψηλής ποιότητας γραφήματα. Βρίσκεται σε πλήρη αρμονία με τις προηγούμενες βιβλιοθήκες-πακέτα, ενώ δίνει στο χρήστη πολλές επιλογές για παραμετροποίηση.

Α΄.1.2 VB.NET

Η γλώσσα VB.NET είναι μία αντικειμενοστραφής γλώσσα προγραμματισμού που μας επέτρεψε να χτίσουμε μία βιβλιοθήκη που διαχειρίζεται πλήρως τις χρονοσειρές και τα μοντέλα πρόβλεψης ως αντικείμενα που απορρέει το ένα από το άλλο. Έτσι, η εφαρμογή των μοντέλων πρόβλεψης της παρούσας εργασίας έγινε αποκλειστικά μέσω της VB.NET.

Α΄.2 Πλατφόρμες και προγραμματιστικά εργαλεία

Χρησιμοποιήθηκαν δύο διαφορετικά περιβάλλοντα για να αναπτύξουμε τον κώδικα της διπλωματικής, ένα για το κομμάτι που αναπτύχθηκε σε Python και ένα για αυτό σε VB.NET.

Α΄.2.1 Jupyter Notebook

Το Jupyter Notebook αποτελεί μία πλατφόρμα ανοικτού κώδικα που τρέχει στο πρόγραμμα περιηγητή (browser) και επιτρέπει τη δημιουργία και διαμοιρασμό αρχείων που περιέχουν κώδικα, εξισώσεις, οπτικοποίηση δεδομένων και επεξηγηματικό κείμενο. Κάνει εύκολη τη επεξεργασία δεδομένων, την αριθμητική προσομοίωση, τη στατιστική μοντελοποίηση και την εφαρμογή μεθόδων μηχανικής μάθησης μεταξύ άλλων. Όλα αυτά, δίνει τη δυνατότητα να γίνουν με έναν διαδραστικό τρόπο.

Α΄.2.2 Visual Studio 2013

Η ανάπτυξη της βιβλιοθήκης προβλέψεων έγινε στη πλατφόρμα του Visual Studio 2013. Η πλατφόρμα αποτελεί ένα ολοκληρωμένο περιβάλλον ανάπτυξης (integrated development environment (IDE)) που προσφέρει εργαλεία για

- Σχεδιασμό
- Ανάπτυξη Κώδικα
- Χτίσιμο και Ανάπτυξη Εφαρμογών
- Αποσφαλμάτωση (Debugging)
- Συνεργασία Προγραμματιστών

Παράρτημα Β΄

Αναλυτικά Αποτελέσματα

Ακολουθούν τα αναλυτικά αποτελέσματα ακρίβειας των μεθόδων με τη προτεινόμενη μέθοδο στις αριστερές στήλες και την κλασική στις δεξιές:

	DES	HES	LRL	Naive	SES	Θ	DES	HES	LRL	Naive	SES	Θ
1	1.46	1.66	1.45	1.39	1.39	1.53	1.99	2.25	3.15	1.8	1.84	2.05
2	0.9	0.99	0.99	1.0	0.84	0.82	1.04	1.06	1.06	1.0	1.0	1.03
3	0.97	0.93	0.93	0.85	1.11	1.07	1.04	1.37	2.25	1.37	1.37	1.37
4	0.86	0.83	0.94	0.85	0.85	0.84	0.87	1.06	3.64	1.01	1.01	1.04
5	0.7	0.67	0.67	0.86	0.75	0.75	1.19	1.19	1.24	1.13	1.19	1.19
6	0.77	0.76	1.38	0.73	0.78	0.77	0.99	1.04	5.62	1.04	1.04	1.04
7	0.66	0.77	0.77	0.67	0.35	0.44	1.1	1.08	1.08	0.78	1.16	1.12
8	0.67	0.65	0.54	0.69	0.67	0.66	0.6	0.76	1.01	0.87	0.87	0.82
9	0.5	0.47	0.46	0.5	0.5	0.48	0.46	0.51	0.72	0.51	0.51	0.51
10	0.33	0.35	0.35	0.82	0.71	0.3	1.32	1.43	1.31	1.44	1.44	1.44
11	10.52	4.28	4.28	1.0	18.86	14.15	1.0	3.7	30.34	1.0	11.37	6.76
12	1.04	0.9	0.87	1.29	1.29	1.09	0.98	0.75	1.54	1.05	1.05	0.8
13	2.98	2.6	2.6	1.21	3.79	3.75	0.97	0.97	3.22	1.23	1.23	0.97
14	3.4	4.72	4.29	5.02	5.02	4.87	0.74	5.79	15.39	6.63	6.63	6.21
15	0.81	0.74	0.7	0.87	0.82	0.78	0.56	0.57	0.57	0.83	0.55	0.55
16	0.89	0.89	0.74	0.91	0.88	0.89	0.85	1.36	1.77	1.39	1.41	1.39
17	1.0	1.0	0.65	0.99	0.99	1.0	1.0	1.0	3.18	1.0	1.0	1.0
18	0.76	0.69	0.69	0.83	0.91	0.92	1.29	1.3	2.07	1.17	1.29	1.29
19	0.5	0.5	0.55	0.5	0.5	0.5	0.83	0.94	1.27	0.43	0.43	0.43
20	0.6	0.59	0.59	0.6	0.61	0.6	0.56	0.39	0.75	0.33	0.33	0.36
21	1.51	1.49	1.22	1.51	1.51	1.5	1.67	1.77	2.46	1.79	1.79	1.78
22	0.41	0.4	0.4	0.43	0.43	0.42	0.5	0.44	0.91	0.43	0.43	0.43
23	0.46	0.5	0.64	0.46	0.46	0.47	0.25	0.33	2.33	0.26	0.26	0.29
24	0.85	0.86	0.89	0.83	0.83	0.85	0.9	0.9	0.9	0.89	0.89	0.9
25	0.5	0.51	0.56	0.53	0.53	0.52	1.35	1.71	3.77	1.72	1.72	1.71
26	0.62	0.61	0.64	0.62	0.62	0.62	0.75	0.46	1.36	0.48	0.48	0.47

27	0.44	0.43	0.43	0.74	0.63	0.53	1.0	1.04	3.08	0.95	1.04	1.04
28	0.62	0.7	0.87	0.63	0.63	0.67	0.93	1.0	2.81	0.78	0.78	0.78
29	0.76	0.71	0.71	0.71	0.71	0.71	1.0	1.0	1.88	1.1	1.1	1.1
30	0.55	0.57	0.57	0.58	0.5	0.5	1.0	0.89	1.51	0.89	0.89	0.89
31	0.73	0.69	0.69	0.69	0.8	0.78	1.04	1.02	1.3	1.0	1.0	1.01
32	1.0	0.69	0.69	0.98	0.59	0.64	1.0	0.75	3.21	0.97	0.97	0.86
33	0.48	0.45	0.45	0.42	0.55	0.53	0.94	0.75	1.65	0.78	0.78	0.76
34	0.89	0.93	4.34	0.52	0.53	0.81	1.0	0.58	11.11	0.26	0.33	0.37
35	1.0	1.0	1.15	1.0	1.0	1.0	1.0	0.97	5.45	1.0	1.0	0.99
36	0.72	0.97	0.97	0.81	1.1	0.81	1.16	1.0	1.0	0.73	4.55	3.16
37	0.73	0.76	0.76	0.58	0.56	0.64	1.05	1.11	2.96	1.12	1.12	1.11
38	0.92	0.85	1.02	0.92	0.92	0.89	1.0	0.81	2.72	0.95	0.95	0.87
39	0.56	0.54	0.54	0.61	0.63	0.62	0.74	0.83	1.38	0.58	0.73	0.78
40	0.6	0.58	0.58	0.6	0.66	0.63	0.89	0.79	0.79	0.37	1.16	0.98
41	0.43	0.39	0.39	0.42	0.42	0.41	0.38	0.41	0.62	0.41	0.41	0.41
42	0.61	0.6	1.05	0.61	0.61	0.61	0.96	0.46	5.8	0.55	0.55	0.51
43	0.62	0.76	4.62	0.64	0.64	0.82	1.09	1.15	6.07	1.18	1.18	1.17
44	0.89	0.88	0.88	0.85	0.94	0.91	0.49	0.57	2.28	0.49	0.51	0.54
45	0.54	0.61	1.06	0.96	0.52	0.57	1.0	0.51	5.27	0.96	0.96	0.72
46	0.84	0.96	0.96	0.4	0.68	0.78	1.0	3.08	3.08	0.85	1.04	1.04
47	0.51	0.47	0.47	0.55	0.54	0.53	0.37	0.37	0.82	0.37	0.37	0.37
48	0.32	0.24	0.24	1.08	0.77	0.59	1.69	1.97	3.39	1.77	1.77	1.84
49	0.38	0.41	0.41	0.29	0.3	0.33	1.0	0.99	1.66	0.85	0.99	1.0
50	1.18	1.03	1.03	2.08	2.59	1.58	6.59	5.62	5.62	3.29	11.4	9.2
51	0.9	0.81	1.16	0.9	0.9	0.86	0.94	0.75	4.36	0.94	0.94	0.84
52	0.52	0.57	0.57	0.93	0.51	0.45	1.0	0.62	3.73	0.91	0.91	0.69
53	0.49	0.4	0.7	0.5	0.5	0.45	1.21	1.26	1.78	1.26	1.26	1.26
54	0.51	0.84	0.84	0.55	0.53	0.41	1.32	1.16	1.16	0.86	0.98	1.92
55	1.0	1.0	1.0	1.0	0.59	0.95	0.96	1.01	1.01	1.0	2.07	1.53
56	0.68	0.82	0.82	0.71	0.82	0.72	1.11	1.11	2.3	1.0	1.11	1.11
57	1.02	1.0	1.0	0.9	0.9	1.01	1.0	1.01	2.02	1.11	1.11	1.04
58	0.62	0.64	0.64	0.61	0.58	0.61	0.76	0.83	2.97	0.8	0.8	0.8
59	0.69	0.77	0.77	0.66	0.91	0.89	1.35	1.36	3.68	1.37	1.37	1.36
60	0.78	0.8	0.8	0.83	0.83	0.8	0.37	0.38	0.35	0.48	0.35	0.36
61	1.0	0.99	0.6	0.99	0.99	0.99	1.0	0.99	3.54	1.0	1.0	0.99
62	0.68	0.66	0.51	0.71	0.68	0.67	1.0	0.54	2.37	0.67	0.67	0.6
63	1.0	0.89	1.0	0.88	0.88	0.55	0.94	0.84	1.56	1.09	1.09	0.94
64	0.92	1.0	1.0	0.41	0.5	0.38	0.87	0.89	0.89	0.58	3.02	2.93
65	0.91	0.78	0.51	0.74	0.74	0.76	1.0	1.0	3.51	0.8	0.8	0.7
66	0.74	0.7	0.7	0.61	1.1	0.88	1.2	0.99	0.99	0.52	0.52	0.58
67	0.73	0.72	0.72	0.73	0.73	0.73	0.9	0.99	0.66	0.26	0.26	0.23

68	0.39	0.45	0.45	0.46	0.26	0.24	0.59	0.79	0.9	0.58	0.58	0.66
69	1.0	0.8	1.1	0.92	0.92	0.86	1.0	1.08	5.7	1.1	1.1	1.09
70	0.55	0.56	0.56	0.5	0.55	0.55	0.95	1.0	2.28	1.18	1.18	1.15
71	0.61	0.62	0.62	0.61	0.61	0.61	0.85	0.22	0.97	0.27	0.27	0.24
72	0.62	0.86	0.86	0.46	0.55	0.46	0.95	0.97	2.2	0.8	0.8	0.8
73	0.82	0.68	0.68	0.48	1.01	0.95	1.0	1.08	2.08	0.83	0.93	0.96
74	0.46	0.38	0.38	0.83	1.22	0.86	0.97	0.97	3.62	0.81	0.81	0.95
75	1.69	1.92	1.92	0.98	1.1	1.34	1.29	2.71	7.19	1.48	1.48	2.1
76	1.97	1.49	0.98	2.13	2.46	1.98	1.67	1.5	1.34	2.69	3.38	2.03
77	0.44	0.45	0.48	0.36	0.36	0.41	0.77	0.82	2.38	0.84	0.84	0.83
78	0.55	0.58	0.58	0.45	0.41	0.48	0.97	0.53	2.53	0.61	0.61	0.54
79	0.62	0.69	0.69	0.49	0.49	0.52	0.64	0.82	2.76	0.69	0.69	0.71
80	0.43	0.44	0.56	0.41	0.43	0.44	0.97	0.79	1.88	0.79	0.79	0.79
81	0.86	1.0	1.0	0.71	0.71	0.99	1.01	1.0	1.2	0.92	0.92	1.04
82	0.58	0.59	0.59	0.52	0.54	0.55	0.54	0.93	2.21	0.6	0.6	0.59
83	0.49	0.51	0.51	0.5	0.5	0.49	0.54	0.47	1.39	0.54	0.54	0.5
84	1.0	0.55	1.73	0.89	0.89	0.72	1.0	1.07	8.42	1.12	1.12	1.08
85	0.57	0.61	0.61	0.89	0.37	0.41	1.83	2.05	3.58	1.83	1.83	1.94
86	0.54	0.6	0.6	0.41	0.45	0.5	0.72	1.24	3.21	0.65	0.78	0.93
87	0.68	0.81	0.81	0.53	0.53	0.48	0.33	0.34	1.15	0.35	0.35	0.35
88	0.45	0.45	0.45	0.46	0.5	0.47	1.66	1.87	1.87	0.38	0.39	0.46
89	0.64	0.81	1.61	0.54	0.54	0.67	1.04	1.0	6.91	0.78	0.78	0.87
90	4.04	4.54	4.54	2.69	2.23	3.11	4.41	7.15	13.5	3.79	3.79	5.47
91	0.87	0.93	0.93	0.49	0.49	0.59	0.56	0.94	1.64	0.39	0.39	0.38
92	0.42	0.49	0.49	0.43	0.43	0.4	0.88	0.94	2.5	0.54	0.54	0.56
93	0.44	0.48	0.48	0.81	0.19	0.25	1.0	1.0	2.59	0.75	0.75	0.46
94	1.81	2.06	2.06	0.7	0.86	1.37	1.0	1.0	7.33	1.42	1.42	1.44
95	0.97	0.98	0.85	0.96	0.96	0.97	1.0	0.76	0.74	0.96	0.96	0.86

Πίνακας Β΄.1: Δείκτες Ακρίβειας για όλες τις χρονοσειρές

Γλωσσάριο

Ελληνικός όρος	Αγγλικός όρος
στιβαρότητα	robustness
κινητοί μέσοι όροι	moving averages
επαναδειγματοληψία	resampling
δειγματοληψία προς τα πάνω	upsampling
δειγματοληψία προς τα κάτω	downsampling
βάση σύγκρισης	benchmark
εκθετική εξομάλυνση	exponential smoothing
γραμμές Θ	theta lines
μηχανική μάθηση	machine learning
ανάλυση συστάδων	cluster analysis
συστάδα	cluster
συσταδοποίηση	clustering
υπερπροσαρμογή	overfitting
περιηγητής	browser