ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΗΣ ΙΣΧΥΟΣ

Μοντελοποίηση Φορτίων και Παραγωγής σε Δίκτυα Διανομής με χρήση Στατιστικής με σκοπό την Παραγωγή Τεχνητών Δεδομένων για Μελέτες Βέλτιστης Λειτουργίας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Νικόλαος Γκοβάτσος

Επίβλεψη : Νικόλαος Χατζηαργυρίου,
Ομότιμος Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΗΣ ΙΣΧΥΟΣ

Μοντελοποίηση Φορτίων και Παραγωγής σε Δίκτυα Διανομής με χρήση Στατιστικής με σκοπό την Παραγωγή Τεχνητών Δεδομένων για Μελέτες Βέλτιστης Λειτουργίας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Νικόλαος Γκοβάτσος

Επίβλεψη : Νικόλαος Χατζηαργυρίου,
Ομότιμος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 20η Μαρτίου 2024

.....	.....	.....
Νικόλαος Χατζηαργυρίου	Γεώργιος Κορρές	Παύλος Γεωργιλάκης
Ομότιμος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2024

.....

Γκοβάτσος Νικόλαος

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

ΠΕΡΙΛΗΨΗ

Ένα γενικό πρόβλημα, το οποίο σίγουρα συνδέεται και με την εμφάνιση νέων τεχνολογιών, είναι η ανεπάρκεια δεδομένων για την εκπαίδευση των αλγορίθμων μηχανικής μάθησης, με τη χρήση της οποίας καλούμαστε εμείς να κάνουμε προβλέψεις και αναλύσεις. Έτσι και στα δίκτυα διανομής για την εποπτεία, καθώς επίσης, την εύρυθμη λειτουργία των ηλεκτρικών δικτύων απαιτείται μεγάλος αριθμός δεδομένων. Ειδικά τα τελευταία χρόνια, με την ένταξη νέων φορτίων όπως τα ηλεκτρικά οχήματα και τις μονάδες αποθήκευσης ενέργειας δικτυακής κλίμακας, αναπόφευκτη ήταν η επέκταση των δικτύων, η ανανέωση και η αναβάθμιση τους. Ως αποτέλεσμα, σε πολλές των περιπτώσεων τα δεδομένα που έχουμε να διαχειριστούμε είναι ανεπαρκή ή ακόμα και αυτά τα ελάχιστα δεδομένα φορτίων του δικτύου που έχουμε στη διάθεση μας, παραμένουν στη διάθεση λίγων λόγω του κώδικα εμπιστευτικότητας.

Το πρόβλημα αυτό καλούμαστε να λύσουμε στην παρούσα διπλωματική εργασία. Χρησιμοποιώντας αλγορίθμους μηχανικής μάθησης και άλλες στατιστικές προσεγγίσεις, προσπαθούμε να πετύχουμε την απόκτηση περισσότερων δεδομένων μέσω της δημιουργίας συνθετικών. Στην παρούσα εργασία, η προτεινόμενη τεχνική υλοποιήθηκε σε περιβάλλον PyCharm με χρήση της γλώσσας προγραμματισμού ‘Python’ και εφαρμόστηκε για να δημιουργήσει συνθετικά δεδομένα ενός χρόνου, τόσο ενεργού, όσο και άεργου ισχύος για 3 διαφορετικούς ζυγούς. Βασικό μέλημα αποτελεί η διατήρηση της ίδιας συσχέτισης των δεδομένων φορτίου των διαφορετικών ζυγών μεταξύ τους, αρχικά και τελικά. Στο τέλος, τα δεδομένα ελέγχονται για την καταλληλόλητας τους, συγκρίνοντας κάποια στατιστικά τους μετρικά και μέσω μιας μελέτης ροών φορτίου.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

Μοντελοποίηση Φορτίων, Δίκτυα διανομής, Αύξηση δεδομένων, Χρονοσειρές, Γκαουσιανά μείγματα, Μελέτη ροών φορτίου

ABSTRACT

A general problem, which is certainly linked to the emergence of new technologies, is the inadequacy of data for the training of machine learning algorithms, with the use of which we are asked to make predictions and analyses. Thus, in the distribution networks, a large amount of data is required for the supervision, as well as the orderly operation of the electrical networks. Especially in recent years, with the inclusion of new loads such as electric vehicles and grid-scale energy storage units, it was inevitable to expand networks, renew and upgrade them. As a result, in many cases the data we have to manage is insufficient, or even the minimal network load data we have, remains available to a few because of the confidentiality code.

We are called to solve this problem in this thesis. Using machine learning algorithms and other statistical approaches, we strive to achieve more data acquisition through synthetic generation. In the present work, the proposed technique was implemented in a PyCharm environment using 'Python' programming language and applied to generate one-time synthetic data of both active and reactive power for 3 different buses. A key concern is to maintain the same correlation of the load data of the different buses with each other, initially and finally. Finally, the data is checked for suitability by comparing some statistical metrics and through a power flow analysis.

KEY WORDS

Load Modeling, Distributionnetworks, Data augmentation, Time series, Gaussian mixtures, Power flow analysis

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε κατά το ακαδημαϊκό έτος 2023–2024 υπό την επίβλεψη του κ. Νικόλαου Χατζηαργυρίου, Ομότιμο Καθηγητή της σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Ε.Μ.Π. στον οποίο οφείλω ιδιαίτερες ευχαριστίες για την ανάθεσή της, δίνοντάς μου την ευκαιρία να ασχοληθώ με ένα θέμα, το οποίο με το που το είδα, μου φάνηκε, ιδιαίτερα, ενδιαφέρον. Επίσης, θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα Παναγιώτη Πεδιαδίτη για την πολύτιμη βοήθεια και καθοδήγηση που μου παρείχε σε όλη τη διάρκεια εκπόνησης της εργασίας και που ήταν πάντα δίπλα μου για την επίλυση αποριών.

Ιδιαίτερα θα ήθελα να ευχαριστήσω την οικογένειά μου, την κοπέλα μου και τους φίλους μου, που ήταν δίπλα μου καθ’ όλη τη διάρκεια των σπουδών μου και δεν σταμάτησαν να πιστεύουν σε εμένα.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΙΝΑΚΑΣ ΕΙΚΟΝΩΝ

ΚΕΦΑΛΑΙΟ 1

1. Εισαγωγή

1.1. Βιβλιογραφική ανασκόπηση μεθόδων αύξησης δεδομένων (data augmentation) χρονοσειρών

Είναι γεγονός ότι, οι σύγχρονοι αλγόριθμοι μπορούν να πετύχουν με μεγάλη επιτυχία διεργασίες όπως η γέννηση καινούριων δεδομένων παρόμοιων με τα αρχικά (data generation), είτε, όπως, η αναγνώριση μοτίβων χρονοσειρών (pattern recognition) καθώς και άλλων πολλών διαδικασιών, εκτός του αντικειμένου της συγκεκριμένης διπλωματικής. Σημαντικός παράγοντας αυτής της επιτυχίας είναι το γεγονός ότι, τα προς επεξεργασία, δεδομένα είναι μεγάλα σε αριθμό και έτσι αυξάνεται η γενικότητα. Ωστόσο, στον τομέα της αναγνώρισης χρονοσειρών, πολλά σύνολα δεδομένων είναι συχνά πολύ μικρά. Μια μέθοδος αντιμετώπισης αυτού του προβλήματος είναι η αύξηση δεδομένων (data augmentation). Παρακάτω θα παραθέσουμε τις κύριες μεθόδους αύξησης δεδομένων χρονοσειρών, συμπεριλαμβανομένων μεθόδων που βασίζονται σε μετασχηματισμό, ανάμειξη προτύπων, μοντέλων παραγωγής και μεθόδων αποσύνθεσης.

1.2. Αύξηση δεδομένων με Τυχαίο Μετασχηματισμό

Οι τεχνικές αύξησης δεδομένων χρονοσειρών δανείζονται από τις τεχνικές αύξησης δεδομένων εικόνας, όπως η περικοπή, η αναστροφή και η προσθήκη θορύβου. Αυτές οι μέθοδοι αύξησης βασίζονται σε τυχαίους μετασχηματισμούς των αρχικών δεδομένων. Δηλαδή, η επαύξηση δεδομένων που βασίζεται σε τυχαίο μετασχηματισμό δημιουργεί ένα μοτίβο x′ χρησιμοποιώντας κάποια συνάρτηση μετασχηματισμού g(·), ή:

$$x' \leftarrow g(x)$$

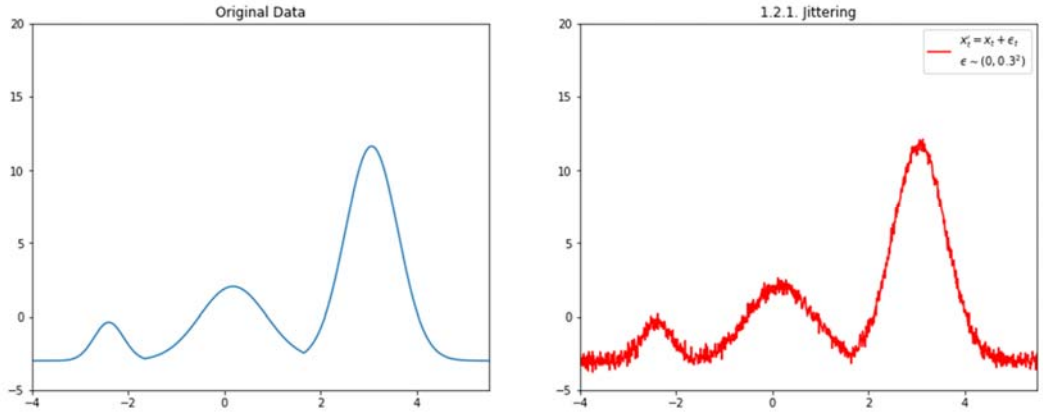
όπου x είναι μια ακολουθία αναφοράς x = x_1, …, x_t, …, x_T με T αριθμό χρονικών βημάτων από το σύνολο εκπαίδευσης.

1.2.1. Jittering

Μία από τις απλούστερες, αλλά αποτελεσματικές μεθόδους αύξησης δεδομένων που βασίζονται σε μετασχηματισμούς είναι το jittering, ή η πράξη της προσθήκης θορύβου στις χρονοσειρές. Το jittering μπορεί να οριστεί ως:

$$x' = x_1 + \epsilon_1, \dots, x_t + \epsilon_t, \dots, x_T + \epsilon_T$$

όπου ϵ είναι τυπικός Gaussian θόρυβος που προστίθεται σε κάθε χρονικό βήμα t και $\epsilon \sim N(0,\sigma^2)$. Η τυπική απόκλιση σ του προστιθέμενου θορύβου είναι μια υπερπαράμετρος που πρέπει να προκαθοριστεί. [1]



1 Αρχικά δεδομένα – Προσθήκη θορύβου [2]

1.2.2. Περιστροφή

Η περιστροφή ορίζεται ως:

$$x' = Rx_1, \dots, Rx_t, \dots, Rx_T$$

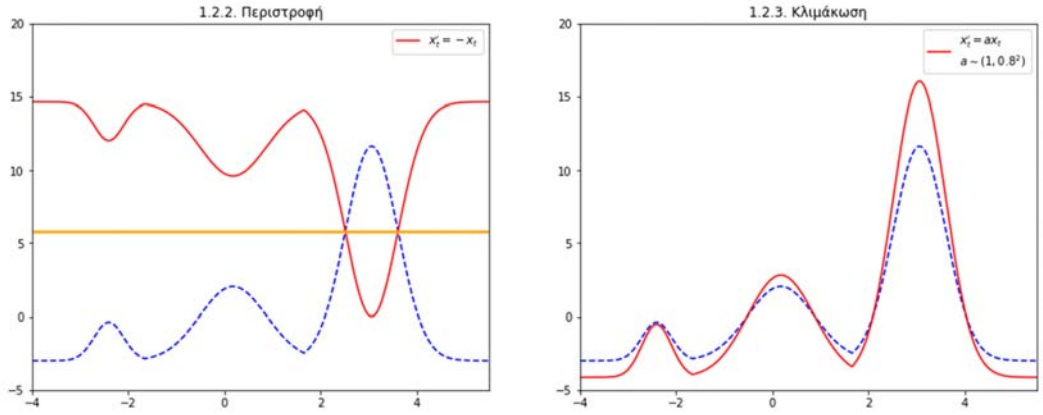
όπου το R είναι ένας πίνακας τυχαίας περιστροφής βάσει στοιχείων για γωνία $\theta \sim N(0,\sigma^2)$ για πολυμεταβλητές χρονικές σειρές και αναστροφή για μονομεταβλητές χρονικές σειρές. Ενώ η αύξηση δεδομένων περιστροφής μπορεί να δημιουργήσει εύλογα μοτίβα για την αναγνώριση εικόνας, μπορεί να μην είναι κατάλληλη για χρονοσειρές, καθώς η περιστροφή μιας χρονοσειράς μπορεί να αλλάξει την κλάση που σχετίζεται με το αρχικό δείγμα. [1]

1.2.3. Κλιμάκωση.

Η κλιμάκωση αλλάζει το συνολικό μέγεθος ή την ένταση μιας χρονοσειράς κατά μια τυχαία κλιμακωτή τιμή. Με την παράμετρο κλιμάκωσης α, η κλιμάκωση είναι ένας πολλαπλασιασμός του α σε ολόκληρη τη χρονοσειρά ή:

$$x' = \alpha x_1, \dots, \alpha x_t, \dots, \alpha x_T$$

Η παράμετρος κλιμάκωσης α μπορεί να προσδιοριστεί από μια κατανομή Gauss $\alpha \sim N(1,\sigma^2)$ με το α ως υπερπαράμετρο, ή μπορεί να είναι από μια τυχαία τιμή από ένα προκαθορισμένο σύνολο. Θα πρέπει να σημειωθεί ότι η «κλιμάκωση» όσον αφορά τις χρονοσειρές είναι διαφορετική από ό,τι στον τομέα της εικόνας. Για χρονοσειρές, αναφέρεται απλώς στην αύξηση του πλάτους των στοιχείων και όχι στη μεγέθυνση των χρονοσειρών. [1]



2 Περιστροφή – Κλιμάκωση αρχικών δεδομένων [2]

1.2.4. Παραμόρφωση Πλάτους

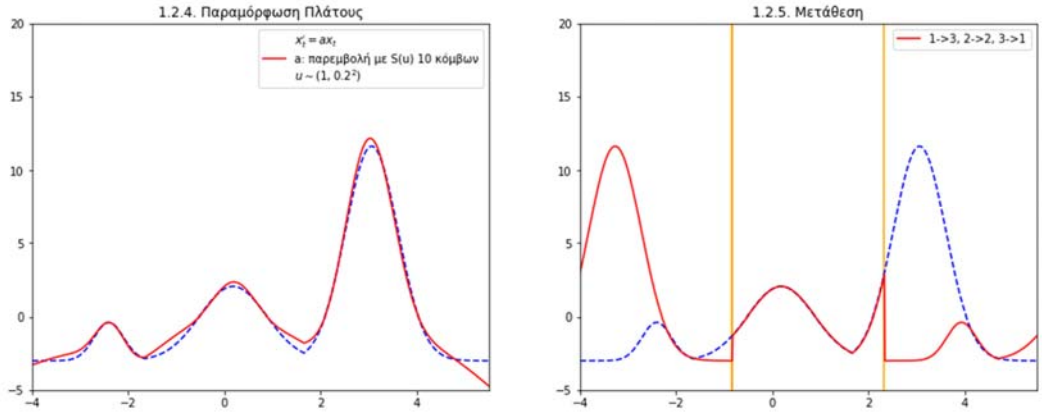
Η παραμόρφωση πλάτους είναι μια τεχνική αύξησης δεδομένων χρονοσειράς που παραμορφώνει το πλάτος ενός σήματος κατά μια εξομαλυνόμενη καμπύλη. Δηλαδή, η επαυξημένη χρονική σειρά x′ είναι η:

$$x' = \alpha_1 x_1, \dots, \alpha_t x_t, \dots, \alpha_T x_T$$

όπου α_1, …, α_t, …, α_T είναι μια ακολουθία που δημιουργείται με παρεμβολή ενός κυβικού σπιράλ S(u) με κόμβους u = u_1, …, u_i, …, u_I. Κάθε κόμβος u_i λαμβάνεται από μια κατανομή $u_i \sim N(1,\sigma^2)$ όπου ο αριθμός των κόμβων I και η τυπική απόκλιση σ είναι υπερπαράμετροι. Η ιδέα πίσω από την παραμόρφωση πλάτους είναι ότι μικρές διακυμάνσεις στα δεδομένα μπορούν να προστεθούν αυξάνοντας ή μειώνοντας τις τυχαίες περιοχές στη χρονοσειρά. Ωστόσο, τα μειονεκτήματα της παραμόρφωση πλάτους για την αύξηση δεδομένων είναι ότι εξακολουθεί να υποθέτει ότι ο τυχαίος μετασχηματισμός είναι ρεαλιστικός και εξαρτάται από δύο προκαθορισμένες υπερπαραμέτρους (ο αριθμός των κόμβων I και η τυπική απόκλιση του ύψους του κόμβου σ) αντί για μία όπως στις άλλες μεθόδ που βασίζονται σε μετασχηματισμό. [3]

1.2.5. Μετάθεση.

Η μετάθεση για την αύξηση δεδομένων προτάθηκε ως μέθοδος αναδιάταξης τμημάτων μιας χρονοσειράς προκειμένου να παραχθεί ένα νέο μοτίβο. Πρέπει να σημειωθεί ότι η μετάθεση δεν διατηρεί χρονικές εξαρτήσεις. Μπορεί να εκτελεστεί με δύο τρόπους, με τμήματα ίσου μεγέθους και με τμήματα μεταβλητού μεγέθους. Η χρήση μετάθεσης με τμήματα ίσου μεγέθους χωρίζει τη χρονοσειρά σε N αριθμό τμημάτων μήκους και τα μεταθέτει. Η χρήση τμημάτων μεταβλητού μεγέθους χρησιμοποιεί τμήματα τυχαίων μεγεθών.[3]



3 Παραμόρφωση πλάτους – Μετάθεση αρχικών δεδομένων [2]

1.2.6. Slicing.

Το Slicing είναι η αύξηση δεδομένων χρονοσειράς που ισοδυναμεί με την περικοπή για την αύξηση δεδομένων εικόνας. Η γενική ιδέα πίσω από τον τεμαχισμό είναι ότι τα δεδομένα επαυξάνονται με τον τεμαχισμό των χρονικών βημάτων από τα άκρα του μοτίβου ή:

$$x' = x_{\varphi}, \dots, x_t, \dots, x_{W+\varphi}$$

όπου W είναι το μέγεθος ενός παραθύρου και ϕ είναι ένας τυχαίος ακέραιος έτσι ώστε 1 ≤ ϕ ≤ T − W.[3]

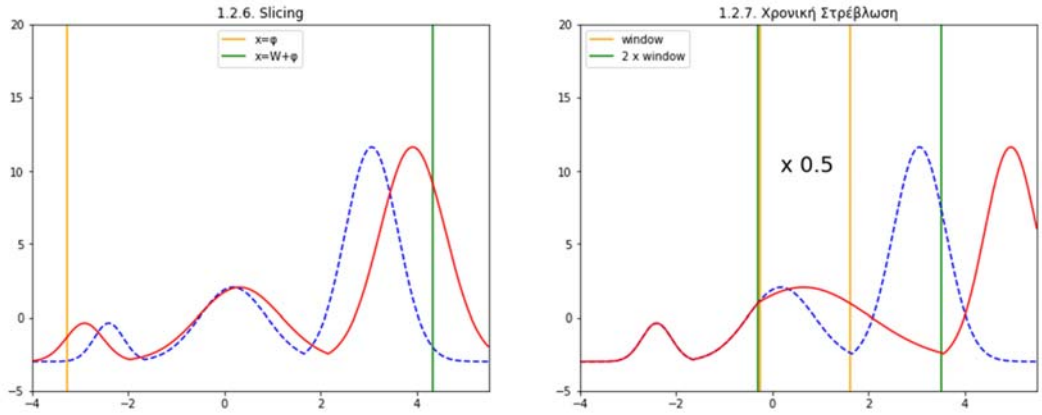
1.2.7. Χρονική Στρέβλωση.

(i) Η στρέβλωση του χρόνου είναι η πράξη της διατάραξης ενός μοτίβου στη χρονική διάσταση. Αυτό μπορεί να εκτελεστεί χρησιμοποιώντας μια ομαλή διαδρομή παραμόρφωσης ή μέσω ενός σταθερού τυχαίου παραθύρου. Όταν χρησιμοποιείτε στρέβλωση χρόνου με ομαλή διαδρομή παραμόρφωσης, η επαυξημένη χρονική σειρά γίνεται:

$$x' = x_{\tau(1)}, \dots, x_{\tau(t)}, \dots, x_{\tau(T)}$$

όπου τ(·) είναι μια συνάρτηση παραμόρφωσης που παραμορφώνει τα χρονικά βήματα με βάση μια ομαλή καμπύλη. Η ομαλή καμπύλη ορίζεται από ένα κυβικό σπιράλ S(u) με κόμβους u = u_1, …, u_i, …, u_I. Το ύψος των κόμβων u_i λαμβάνονται από 1. Με αυτόν τον τρόπο, τα χρονικά βήματα της σειράς έχουν ομαλή μετάβαση μεταξύ διατάσεων και συστολών. [3]

(ii) Εναλλακτικά, μια δημοφιλής μέθοδος στρέβλωσης χρόνου είναι η **παραμόρφωση παραθύρου**. Η παραμόρφωση παραθύρου παίρνει ένα τυχαίο παράθυρο της χρονοσειράς και το τεντώνει κατά 2 ή το συστέλλει κατά ½ . Ενώ οι πολλαπλασιαστές είναι σταθεροί στο ½ και 2, έχει παρατηρηθεί ότι μπορούν να τροποποιηθούν ή να βελτιστοποιηθούν και σε άλλες τιμές.[4]



4 Slicing – Χρονική στρέβλωση αρχικών δεδομένων [2]

1.3. Αύξηση δεδομένων με χρήση Γενετικών (ή Παραγωγής) Μοντέλων

Αντί να χρησιμοποιήσουμε τυχαίους μετασχηματισμούς, είναι δυνατό να δημιουργήσουμε νέες χρονοσειρές από κατανομές χαρακτηριστικών με χρήση γενετικών μοντέλων. Ταξινομούμε τα γενετικά μοντέλα σε δύο κατηγορίες, στα στατιστικά μοντέλα και στα μοντέλα που βασίζονται σε νευρωνικά δίκτυα.

1.3.1. Στατιστικά γενετικά μοντέλα

Υπάρχει μια μεγάλη ποικιλία στατιστικών, μαθηματικών ή στοχαστικών μοντέλων που χρησιμοποιούνται για τη δημιουργία και την αύξηση χρονοσειρών. Συνήθως, αυτές οι μέθοδοι αύξησης δημιουργούν ένα στατιστικό μοντέλο των δεδομένων και χρησιμοποιούνται συχνά στην πρόβλεψη. Για παράδειγμα, για την παραγωγή δεδομένων έχουν χρησιμοποιηθεί πολλοί τρόποι, σε πολλές εργασίες, όπως: Μείγματα γκαουσιανών δέντρων για την υπερδειγματοληψία μη ισορροπημένων κλάσεων και ταξινόμηση χρονοσειρών ή, ακόμα, μοντέλα μεικτής αυτοπαλίνδρομης λειτουργίας για την προσομοίωση χρονοσειρών. Υπάρχουν επίσης εργασίες που χρησιμοποιούν την τεχνική οπίσθιας δειγματοληψίας. [5]

1.3.2. Μοντέλα παραγωγής που βασίζονται σε νευρωνικά δίκτυα

Τα μοντέλα παραγωγής που βασίζονται σε νευρωνικά δίκτυα έχουν γίνει δημοφιλή τον τελευταίο καιρό. Ωστόσο, ενώ έχουν προταθεί πολλά δίκτυα παραγωγής, δεν χρησιμοποιήθηκαν όλα τα μοντέλα για την αύξηση δεδομένων. Η πιο βασική εφαρμογή των νευρωνικών δικτύων για τη δημιουργία χρονοσειρών είναι τα απευθείας δίκτυα αλληλουχίας σε ακολουθία όπως τα LSTM και τα χρονικά CNN. [6]

1.3.3. Δίκτυα κωδικοποιητή-αποκωδικοποιητή.

Τα δίκτυα κωδικοποιητή-αποκωδικοποιητή λαμβάνουν μια είσοδο υψηλών διαστάσεων ή δομική, την κωδικοποιούν σε ένα διάνυσμα χαμηλότερης διάστασης λανθάνον χώρο και στη συνέχεια την αποκωδικοποιούν πίσω σε μια έξοδο υψηλής διάστασης ή δομή. Οι μέθοδοι αύξησης δεδομένων δημιουργούν νέα μοτίβα αποκωδικοποιώντας διανύσματα που λαμβάνονται δειγματοληπτικά από τον λανθάνοντα χώρο. Σε ένα παράδειγμα, ένας αυτόματος κωδικοποιητής που βασίζεται σε LSTM χρησιμοποιήθηκε για τη δημιουργία δεδομένων για μια ταξινόμηση LSTM σχετικά με την αναγνώριση ανθρώπινης δράσης με βάση τον σκελετό. Ωστόσο, τα αποτελέσματα ήταν ανάμεικτα όταν συγκρίθηκαν τα αποτελέσματα της αύξησης δεδομένων χρησιμοποιώντας περιστροφή, κλιμάκωση και καμία αύξηση. [7]

1.3.4. Δίκτυα δημιουργίας αντιπάλων (GANs)

Τα GAN είναι μια κατηγορία παραγωγικών δικτύων που χρησιμοποιούν αντίθετη εκπαίδευση για τη βελτιστοποίηση από κοινού δύο νευρωνικών δικτύων, μιας γεννήτριας και ενός διαχωριστή. Παρόμοια με τα δίκτυα κωδικοποιητή-αποκωδικοποιητή, προκειμένου να δημιουργηθούν δείγματα, το GAN εκπαιδεύεται χρησιμοποιώντας το σύνολο δεδομένων εκπαίδευσης και στη συνέχεια γίνεται δειγματοληψία του διανύσματος z και χρησιμοποιείται με τη γεννήτρια για τη δημιουργία νέων χρονοσειρών. Έχουν προταθεί πολλές χρονοσειρές GAN. Ωστόσο, οι περισσότεροι στοχεύουν μόνο τη δημιουργία και όχι την αύξηση δεδομένων.

Τα υποκείμενα δίκτυα GAN για χρονικές σειρές μπορούν χονδρικά να χωριστούν σε τέσσερις αρχιτεκτονικές, GAN που βασίζονται σε πλήρως συνδεδεμένα δίκτυα ή MLP, επαναλαμβανόμενα GAN που χρησιμοποιούν RNN, GAN με προσωρινά CNN ή 1D CNN και GAN που δημιουργούν εικόνες βασισμένες σε φάσμα με 2D CNN. [8]

1.4. Αποσύνθεση χρονοσειρών

Αναφερόμενοι σε δεδομένα χρονοσειρών, η γέννηση δεδομένων μπορεί να χρησιμοποιηθεί για την αύξηση του μεγέθους του σετ εκπαίδευσης, τη βελτίωση της απόδοσης του μοντέλου και τη μείωση του κινδύνου υπερπροσαρμογής. Μια προσέγγιση για την αύξηση δεδομένων χρονοσειρών είναι η χρήση της αποσύνθεσης χρονοσειρών για τη δημιουργία νέων συνθετικών δεδομένων [9].

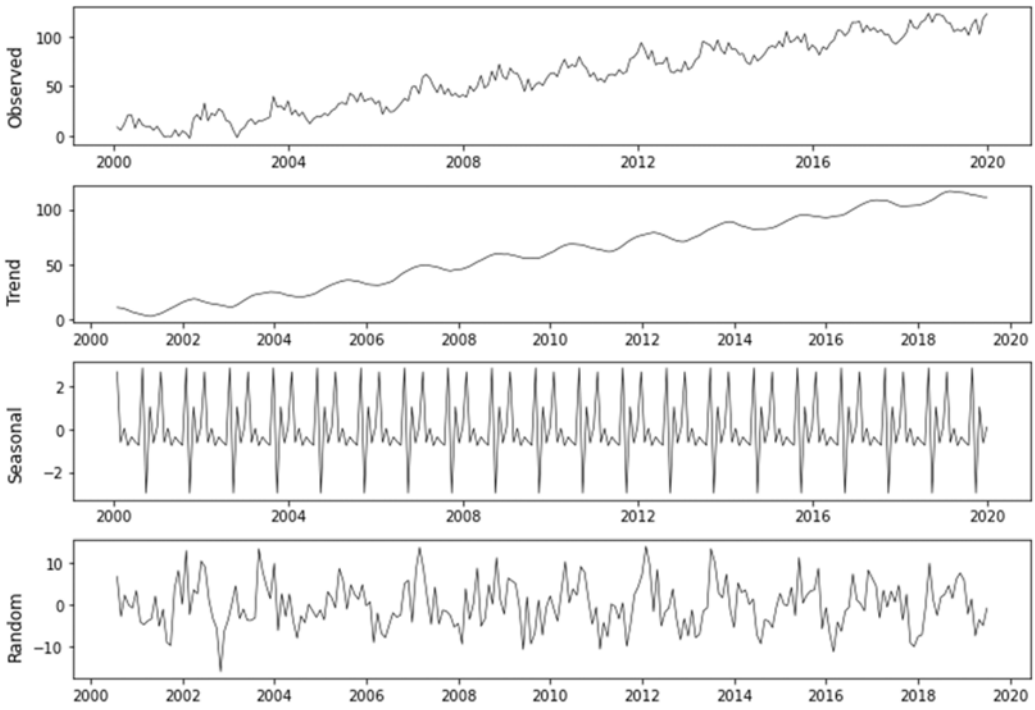
Η αποσύνθεση χρονοσειρών είναι μια στατιστική μέθοδος που χρησιμοποιείται για τον διαχωρισμό μιας χρονοσειράς στις τάσεις, τις εποχιακές και τις ακανόνιστες συνιστώσες της. Η συνιστώσα τάσης αντιπροσωπεύει το μακροπρόθεσμο μοτίβο στα δεδομένα, η εποχιακή συνιστώσα αντιπροσωπεύει τις περιοδικές διακυμάνσεις και η ακανόνιστη συνιστώσα αντιπροσωπεύει την τυχαία διακύμανση. Χρησιμοποιώντας την αποσύνθεση χρονοσειρών, μπορούμε να δημιουργήσουμε νέα δεδομένα χρονοσειρών συνδυάζοντας την τάση και τα εποχιακά στοιχεία με διαφορετικούς τρόπους.

Ένας τρόπος για τη δημιουργία νέων δεδομένων χρονοσειρών είναι η χρήση του bootstrapping. Το bootstrapping είναι μια στατιστική τεχνική που περιλαμβάνει την επαναδειγματοληψία των αρχικών δεδομένων με αντικατάσταση για τη δημιουργία νέων

συνόλων δεδομένων. Στο πλαίσιο των δεδομένων χρονοσειρών, το bootstrapping μπορεί να χρησιμοποιηθεί για τη δημιουργία νέων δεδομένων χρονοσειρών, δειγματίζοντας εκ νέου το ακανόνιστο στοιχείο της αρχικής χρονοσειράς και συνδυάζοντάς το με τα στοιχεία τάσης και τα εποχιακά στοιχεία.

Μια άλλη προσέγγιση για την αύξηση δεδομένων χρονοσειρών είναι η χρήση της αποσύνθεσης εποχικής τάσης χρησιμοποιώντας τη μέθοδο loess (STL). Η μέθοδος STL είναι μια μη παραμετρική μέθοδος για την αποσύνθεση χρονοσειρών που βασίζεται σε τοπικά σταθμισμένη παλινδρόμηση . Η μέθοδος STL είναι ιδιαίτερα χρήσιμη για δεδομένα χρονοσειρών που εμφανίζουν μη γραμμικές τάσεις ή πολύπλοκα εποχιακά μοτίβα. Χρησιμοποιώντας τη μέθοδο STL, μπορούμε να δημιουργήσουμε νέα δεδομένα χρονοσειρών συνδυάζοντας την τάση και τα εποχιακά στοιχεία με διαφορετικούς τρόπους[10].

Συνοπτικά, η αποσύνθεση χρονοσειρών μπορεί να χρησιμοποιηθεί ως ένα ισχυρό εργαλείο για την αύξηση και τη δημιουργία δεδομένων. Αποσυνθέτοντας μια χρονοσειρά στα επιμέρους στοιχεία της, μπορούμε να δημιουργήσουμε νέα συνθετικά δεδομένα επαναδειγματίζοντας το ακανόνιστο στοιχείο χρησιμοποιώντας bootstrapping ή χρησιμοποιώντας τη μέθοδο STL για να συνδυάσουμε την τάση και τα εποχιακά στοιχεία με διαφορετικούς τρόπους. Αυτές οι προσεγγίσεις μπορούν να βοηθήσουν στην αύξηση του μεγέθους του σετ εκπαίδευσης, στη βελτίωση της απόδοσης του μοντέλου και στη μείωση του κινδύνου υπερβολικής προσαρμογής, οδηγώντας σε πιο ακριβή και στιβαρά μοντέλα χρονοσειρών. Επιπλέον, η αποσύνθεση χρονοσειρών μπορεί να χρησιμοποιηθεί για τη δημιουργία συνθετικών δεδομένων για προβλήματα ταξινόμησης μη ισορροπημένων χρονοσειρών. Σε προβλήματα μη ισορροπημένης ταξινόμησης, ο αριθμός των δειγμάτων σε κάθε κατηγορία δεν είναι ίσος, οδηγώντας σε προκατειλημμένα μοντέλα που έχουν κακή απόδοση σε μειοψηφικές κατηγορίες. Χρησιμοποιώντας την αποσύνθεση χρονοσειρών, μπορούμε να δημιουργήσουμε νέα συνθετικά δεδομένα για την τάξη μειοψηφίας, επαναδειγματοληπτώντας το ακανόνιστο στοιχείο της χρονοσειράς και συνδυάζοντάς το με τα στοιχεία τάσης και εποχιακά.



5 Γράφημα αποσύνθεσης [2]

1.5. Προτεινόμενη επίλυση επαύξησης δεδομένων, για δεδομένα ενεργού και άεργου ισχύος

Αφού αναφέραμε πολλές απ' τις κατηγορίες στις οποίες εντάσσονται οι τεχνικές επαύξησης δεδομένων και αφού κάναμε και κάποιες δοκιμές στα δεδομένα μας με τη χρήση αυτών, αποφασίσαμε να απορρίψουμε την πρώτη κατηγορία, των τυχαίων μετασχηματιστών. Δεχτήκαμε να επιλύσουμε το πρόβλημα μας, με ένα στατιστικό γενετικό μοντέλο και, τέλος να κάνουμε μια προσπάθεια επίλυσης του προβλήματος, με αποσύνθεση χρονοσειρών. Η λογική πίσω απ' αυτή την επιλογή, προήλθε απ' το γεγονός ότι οι τυχαίοι μετασχηματισμοί χρησιμοποιούνται κυρίως, ως τεχνικές επαύξησης δεδομένων για εικόνες. Παρότι, τα αποτελέσματα τους μπορεί να φανούν ικανοποιητικά και στην επεξεργασία χρονοσειρών, η θέληση μας, δεν είναι να πάρουμε τα δεδομένα μας και να παραγάγουμε καινούρια, προσθέτοντας απλά μια συνιστώσα θορύβου, ή αναδιατάσσοντας τη χρονοσειρά μας. Θέληση μας, είναι να κατανοήσουμε πλήρως την συμπεριφορά των δεδομένων μας και έπειτα να παραγάγουμε, με πλήρη έλεγχο καινούρια.

Ξεκινήσαμε με την παραγωγή δεδομένων για ένα ζυγό. Στην προσπάθεια μας αυτή, επιλέξαμε *στατιστικά γενετικά μοντέλα*, και συγκεκριμένα, διαχωρισμό των μοτίβων και κατηγοριοποίηση τους με Γκαουσιανά Μείγματα και ύστερα γέννηση καινούριων δεδομένων με χρήση των κατανομών κάθε μοτίβου και τη λογική των αλυσίδων Markov. Απορρίψαμε την επίλυση με γενετικά μοντέλα άλλου τύπου, παρά την ευρεία χρήση τους

και τα εκπληκτικά τους αποτελέσματα, καθώς η απουσία πολλών δεδομένων, θα καθιστούσε την εκπαίδευση του αλγορίθμου δύσκολη και τα αποτελέσματα αυτού λανθασμένα.

Στη συνέχεια, επιλύσαμε το πρόβλημα μας και με τη χρήση της αποσύνθεσης χρονοσειρών. Αυτή η προσπάθεια μας, έγινε για να συγκρίνουμε τα αποτελέσματα, ενός πιο τεχνικού και γενικά χρησιμοποιούμενου, γενετικού αλγορίθμου, με έναν αλγόριθμο κατ' εξοχήν για χρονοσειρές και ανάλυσης αυτών.

Τέλος, επιλέξαμε έναν απ' τους δύο προαναφερθέντες τρόπους για την παραγωγή δεδομένων και για τους τρεις ζυγούς. Αφού παρήγαμε δεδομένα, επιλύσαμε μια μελέτη ροής φορτίου και γράψαμε τα συμπεράσματα μας τα οποία προήλθαν τόσο απ' αυτή όσο και από διάφορες στατιστικές μεθόδους σύγκρισης των παραγόμενων δεδομένων με τα αρχικά.

1.6. Δομή εργασίας

Το επιστημονικό περιεχόμενο που συντάχθηκε αποτελείται από τέσσερεις ενότητες. Το Κεφάλαιο 1 αποτελεί μια εισαγωγή της διπλωματικής εργασίας. Ο αναγνώστης μπορεί να σχηματίσει μία εικόνα για το τι θα ακολουθήσει, μία σύντομη περιγραφή του πειράματος που θα μελετηθεί και τη σκοπιά από την οποία προσεγγίστηκε το πρόβλημα.

Στο Κεφάλαιο 2 παρουσιάζονται οι απαραίτητες θεωρητικές γνώσεις για την πλήρη κατανόηση των αλγορίθμων που επιλέχθηκαν και τον τρόπο λειτουργίας τους. Επίσης, γίνεται εκτενής ανάλυση της μαθηματικής θεωρίας της λειτουργίας του κάθε αλγορίθμου παραγωγής δεδομένων.

Στο Κεφάλαιο 3 παρουσιάζεται πρώτα η θεωρία για τη διαχείριση δεδομένων η οποία θα γίνει στο πειραματικό μέρος και στη συνέχεια παρουσιάζονται κάποιες στατιστικές ιδιότητες οι οποίες χρησιμοποιούνται εκ των υστέρων για την εξαγωγή συμπερασμάτων μεταξύ των αρχικών και των παραγόμενων δεδομένων.

Το Κεφάλαιο 4 περιλαμβάνει το πειραματικό κομμάτι. Την παραγωγή, δηλαδή, δεδομένων για έναν ζυγό με διαφορετικούς τρόπους, στη συνέχεια την παραγωγή και για τους τρεις ζυγούς και, τέλος, την μελέτη ροής φορτίου που γίνεται και για τα δύο σύνολα δεδομένων· αρχικά και παραγόμενα.

ΚΕΦΑΛΑΙΟ 2

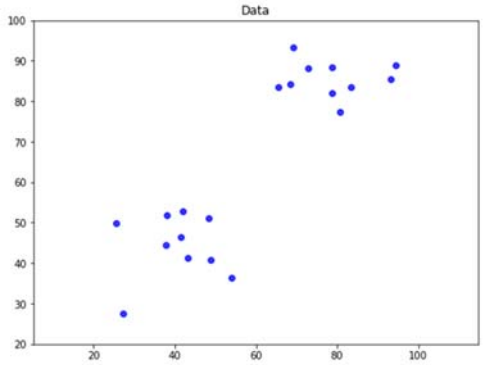
2. Απαραίτητη Θεωρία

2.1. Εισαγωγικά

Στον κόσμο της Μηχανικής Μάθησης, μπορούμε να διακρίνουμε δύο βασικούς τομείς: την *Supervised* και την *Unsupervised* μάθηση. Η κύρια διαφορά μεταξύ των δύο έγκειται στη φύση των δεδομένων καθώς και στις προσεγγίσεις που χρησιμοποιούνται για την αντιμετώπισή τους. Το *Clustering* είναι ένα πρόβλημα unsupervised μάθησης κατά το οποίο σκοπεύουμε να βρούμε ομάδες σημείων στο σύνολο δεδομένων μας που μοιράζονται ορισμένα κοινά χαρακτηριστικά.

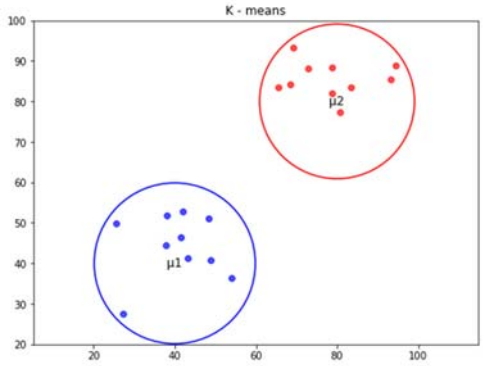
2.1. K-means

Ας υποθέσουμε ότι έχουμε ένα σύνολο δεδομένων που μοιάζει με αυτό:



6 Δεδομένα πριν την ομαδοποίηση μέσω K-means [2]

Στόχος μας είναι να βρίσκουμε σύνολα σημείων που φαίνονται κοντά. Σε αυτήν την περίπτωση, μπορούμε να αναγνωρίσουμε ξεκάθαρα δύο ομάδες σημείων που θα χρωματίσουμε μπλε και κόκκινο, αντίστοιχα:



7 K-means ομαδοποίηση δεδομένων [2]

Τα μ_1 και μ_2 είναι τα κεντροειδή κάθε συστάδας και είναι παράμετροι που προσδιορίζουν καθένα από αυτά. Ένας δημοφιλής αλγόριθμος ομαδοποίησης είναι γνωστός ως K-means. Ένα σημαντικό χαρακτηριστικό του K-means είναι ότι είναι μια αυστηρή μέθοδος ομαδοποίησης, που σημαίνει ότι θα συσχετίσει κάθε σημείο με ένα και μόνο ένα σύμπλεγμα. Ένας περιορισμός αυτής της προσέγγισης είναι ότι δεν υπάρχει μέτρο αβεβαιότητας ή πιθανότητα που να μας λέει πόσο ένα σημείο δεδομένων σχετίζεται με ένα συγκεκριμένο σύμπλεγμα [11]. Τι γίνεται λοιπόν με τη χρήση μιας μη αυστηρής ομαδοποίησης αντί για μια αυστηρή; Αυτό ακριβώς επιχειρούν να κάνουν τα Γκαουσιανά Μείγματα (Gaussian Mixture Models ή απλά τα GMM).

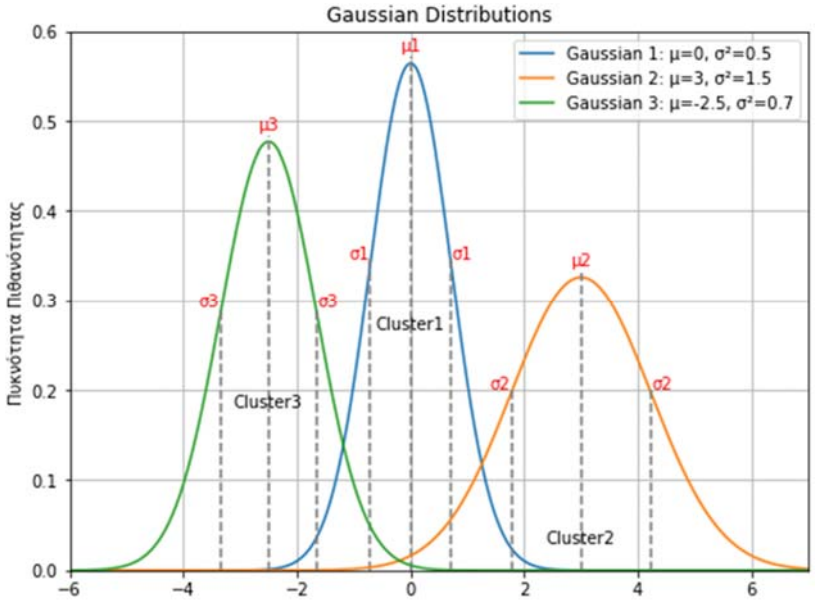
2.2. Γκαουσιανά Μείγματα

2.2.1. Ορισμοί – Αρχική θεωρία

Ένα Γκαουσιανό Μείγμα είναι μια συνάρτηση που αποτελείται από πολλές Gaussians, καθεμία από τις οποίες προσδιορίζεται με $k \in \{1,\ldots, K\}$, όπου K είναι ο αριθμός των συστάδων του συνόλου δεδομένων μας. Κάθε Gaussian k στο μείγμα αποτελείται από τις ακόλουθες παραμέτρους:

- Ένα μέσο όρο μ που ορίζει το κέντρο του.
- Μια συνδιακύμανση Σ που ορίζει το πλάτος του. Αυτό θα ήταν ισοδύναμο με τις διαστάσεις ενός ελλειψοειδούς σε ένα πολυμεταβλητό σενάριο.
- Μια πιθανότητα ανάμειξης π που ορίζει πόσο μεγάλη ή μικρή θα είναι η Gaussian συνάρτηση.

Ας απεικονίσουμε τώρα αυτές τις παραμέτρους γραφικά:



8 Συνιστώσες ενός Γκαουσιανού Μείγματος [2]

Εδώ, μπορούμε να δούμε ότι υπάρχουν τρεις Gaussian συναρτήσεις, επομένως K = 3. Κάθε Gaussian εξηγεί τα δεδομένα που περιέχονται σε καθένα από τα τρία διαθέσιμα συμπλέγματα. Οι συντελεστές ανάμειξης, π_k, είναι πιθανότητες και πρέπει να πληρούν αυτήν την προϋπόθεση:

$$\sum_{k=1}^{K} \pi_k = 1 \qquad (\mathbf{1})$$

Τώρα πώς προσδιορίζουμε τις βέλτιστες τιμές για αυτές τις παραμέτρους; Για να το πετύχουμε αυτό πρέπει να διασφαλίσουμε ότι κάθε Gaussian ταιριάζει στα σημεία δεδομένων που ανήκουν σε κάθε cluster. Αυτό ακριβώς κάνει η μέγιστη πιθανότητα.[12]

Γενικά, η συνάρτηση πυκνότητας Gauss δίνεται από:

$$N(\boldsymbol{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}\,|\Sigma|^{1/2}}\left(-\frac{1}{2}(\boldsymbol{x}-\mu)^T\Sigma^{-1}(\boldsymbol{x}-\mu)\right)$$

Όπου το x αντιπροσωπεύει τα σημεία δεδομένων μας, D είναι ο αριθμός των διαστάσεων κάθε σημείου δεδομένων. μ και Σ είναι ο μέσος όρος και η συνδιακύμανση, αντίστοιχα. Εάν έχουμε ένα σύνολο δεδομένων που αποτελείται από N = 1000 τρισδιάστατα σημεία (D = 3), τότε το x θα είναι ένας πίνακας 1000 × 3. Το μ θα είναι ένα διάνυσμα 1 × 3 και το Σ θα είναι ένας πίνακας 3 × 3. Για μεταγενέστερους σκοπούς, θα είναι επίσης χρήσιμο να πάρουμε τον φυσικό λογάριθμο αυτής της εξίσωσης, το οποίο δίνεται από: [12]

$$\ln\, N(\boldsymbol{x}|\mu,\Sigma) = -\frac{D}{2}\ln 2\pi - \frac{1}{2}\ln\Sigma\, -\frac{D}{2}(\boldsymbol{x}-\mu)^T\Sigma^{-1}(\boldsymbol{x}-\mu) \qquad (\mathbf{2})$$

Εάν διαφοροποιήσουμε αυτήν την εξίσωση ως προς τον μέσο όρο και τη συνδιακύμανση και στη συνέχεια την εξισώσουμε με το μηδέν, τότε θα μπορέσουμε να βρούμε τις βέλτιστες τιμές για αυτές τις παραμέτρους και οι λύσεις θα αντιστοιχούν στις Εκτιμήσεις Μέγιστης Πιθανότητας (MLE) για αυτήν τη ρύθμιση. Ωστόσο, επειδή έχουμε να κάνουμε όχι μόνο με μία, αλλά με πολλές Gaussians, τα πράγματα θα γίνουν λίγο περίπλοκα όταν έρθει η ώρα να βρούμε τις παραμέτρους για όλο το μείγμα. Από αυτή την άποψη, θα χρειαστεί να εισαγάγουμε ορισμένες πρόσθετες πτυχές που θα συζητήσουμε στην επόμενη ενότητα. [13]

Αρχικά, ας υποθέσουμε ότι θέλουμε να μάθουμε ποια είναι η πιθανότητα ένα σημείο x**n** να προέρχεται από μια Gaussian k. Μπορούμε να το εκφράσουμε ως εξής:

$$p(z_{nk} = 1 | \boldsymbol{x}_n)$$

Το οποίο διαβάζεται "δεδομένου ενός σημείου **x**, ποια είναι η πιθανότητα να προήλθε από μια Gaussian k;" Σε αυτήν την περίπτωση, το z είναι μια λανθάνουσα μεταβλητή που παίρνει μόνο δύο πιθανές τιμές. Είναι ένα όταν το x προέρχεται από τη Gaussian k, και μηδέν διαφορετικά. Γενικά, δεν μπορούμε να δούμε αυτή τη μεταβλητή z στην πραγματικότητα, αλλά η γνώση της πιθανότητας εμφάνισής της θα μας βοηθήσει να προσδιορίσουμε τις παραμέτρους του μείγματος Gauss, όπως θα συζητήσουμε αργότερα.

Ομοίως, μπορούμε να αναφέρουμε το εξής:

$$\pi_k = p(z_k = 1)$$

Που σημαίνει ότι η συνολική πιθανότητα παρατήρησης ενός σημείου που προέρχεται από τη Gaussian k είναι στην πραγματικότητα ισοδύναμη με τον συντελεστή ανάμειξης για αυτή τη Gaussian. Αυτό είναι λογικό, γιατί όσο μεγαλύτερη είναι η Gaussian, τόσο μεγαλύτερη θα περιμέναμε να είναι αυτή η πιθανότητα. Τώρα, έστω **z** το σύνολο όλων των πιθανών λανθανουσών μεταβλητών z, επομένως:

$$\boldsymbol{z} = \{z_1, \dots, z_K\}$$

Γνωρίζουμε εκ των προτέρων ότι κάθε z εμφανίζεται ανεξάρτητα από άλλα και ότι μπορούν να πάρουν την τιμή του ενός μόνο όταν το k είναι ίσο με το σύμπλεγμα από το οποίο προέρχεται το σημείο. Επομένως:

$$p(\boldsymbol{z}) = p(z_1 = 1)^{z_1} p(z_2 = 1)^{z_2} \dots p(z_K = 1)^{z_K} = \prod_{k=1}^{K} \pi_k^{z_K}$$

Τώρα, όμως, τι γίνεται με την εύρεση της πιθανότητας παρατήρησης των δεδομένων μας δεδομένου ότι προέρχονται από τη Gaussian k; Αποδεικνύεται ότι είναι στην πραγματικότητα η ίδια η Gaussian συνάρτηση. Ακολουθώντας την ίδια λογική που χρησιμοποιήσαμε για να ορίσουμε το p(**z**), μπορούμε ορίσουμε:

$$p(\boldsymbol{x}_n | \boldsymbol{z}) = \prod_{k=1}^{K} N(\boldsymbol{x}_n | \mu_k, \boldsymbol{\Sigma}_k)^{z_k}$$

Γιατί τα κάνουμε όλα αυτά; Ο αρχικός μας στόχος ήταν να προσδιορίσουμε ποια είναι η πιθανότητα του z με βάση την παρατήρησή μας x. Αποδεικνύεται ότι οι, εξισώσεις που

μόλις εξήγαμε, μαζί με τον κανόνα Bayes, θα μας βοηθήσουν να προσδιορίσουμε αυτήν την πιθανότητα. Από τον κανόνα του γινομένου των πιθανοτήτων, γνωρίζουμε ότι:

$$p(\boldsymbol{x}_n, \boldsymbol{z}) = p(\boldsymbol{x}_n|\boldsymbol{z})p(\boldsymbol{z})$$

Οι τελεστές στα δεξιά είναι αυτό που βρήκαμε πιο πάνω. Πώς όμως θα απαλλαγούμε από το z εδώ; Πρέπει απλώς να προσθέσουμε τους όρους στο z:

$$p(\boldsymbol{x}_n) = \sum_{k=1}^{K} p(\boldsymbol{x}_n|\boldsymbol{z})p(\boldsymbol{z}) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n|\mu_k, \Sigma_k)$$

Αυτή είναι η εξίσωση που ορίζει ένα Gaussian Mixture, και μπορούμε να δούμε ξεκάθαρα ότι εξαρτάται από όλες τις παραμέτρους που αναφέραμε προηγουμένως. Για να προσδιορίσουμε τις βέλτιστες τιμές για αυτές πρέπει να προσδιορίσουμε τη μέγιστη πιθανότητα του μοντέλου. Μπορούμε να βρούμε την πιθανότητα ως την κοινή πιθανότητα όλων των παρατηρήσεων **x**n, που ορίζεται από:

$$p(\mathbf{X}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n|\mu_k, \Sigma_k)$$

Όπως κάναμε για την αρχική συνάρτηση πυκνότητας Gauss, ας εφαρμόσουμε το φυσικό λογάριθμο σε κάθε πλευρά της εξίσωσης:

$$\ln\big(p(\mathbf{X})\big) = \sum_{n=1}^{N} ln \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n|\mu_k, \Sigma_k) \qquad (\mathbf{3})$$

Τώρα για να βρούμε τις βέλτιστες παραμέτρους για το μείγμα Gauss, το μόνο που έχουμε να κάνουμε είναι να διαφορίσουμε αυτήν την εξίσωση ως προς τις παραμέτρους. Όμως, υπάρχει ένας λογάριθμος που επηρεάζει τη δεύτερη άθροιση. Ο υπολογισμός της παραγώγου αυτής της έκφρασης και στη συνέχεια η επίλυση των παραμέτρων θα είναι πολύ δύσκολος.

Τι μπορούμε να κάνουμε; Πρέπει να χρησιμοποιήσουμε μια επαναληπτική μέθοδο για να εκτιμήσουμε τις παραμέτρους. Αλλά πρώτα, Πρέπει να βρούμε την πιθανότητα του z δεδομένου του x.

Από τον κανόνα του Bayes, ξέρουμε αυτό:

$$p(z_k = 1 \mid \mathbf{x}_n) = \frac{p(\mathbf{x}_n \mid z_k = 1)p(z_k = 1)}{\sum_{j=1}^{K} p\big(\mathbf{x}_n \mid z_j = 1\big)p\big(z_j = 1\big)}$$

Από τις προηγούμενες παράγωγες μας είδαμε ότι:

$$p(z_k = 1 \mid \mathbf{x}_n) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_n \mid \mu_j, \Sigma_j\right)} = \gamma(z_{nk}) \qquad (\mathbf{4})$$

Ας τα αντικαταστήσουμε τώρα στην προηγούμενη εξίσωση:

$$p(z_k = 1) = \pi_k, \qquad p(\mathbf{x}_n \mid z_k = 1) = \mathcal{N}(\mathbf{x}_n \mid \mu_k, \mathbf{\Sigma}_k)$$

Και αυτό είναι που ψάχναμε.[13]

2.2.2. Επιλογή αριθμού clusters

Δεδομένου ότι δεν γνωρίζουμε τη βασική αλήθεια των γεννητριών clusters, δηλαδή δεν γνωρίζουμε την αρχική διανομή που δημιούργησε τα δεδομένα, οι επιλογές μας σχετικά με την αξιολόγηση απόδοσης της διαδικασίας ομαδοποίησης είναι περιορισμένες και αρκετά θορυβώδεις.

Ωστόσο, θα διερευνήσουμε τρεις διαφορετικές τεχνικές για τυχαία δεδομένα που βρήκαμε για τη διατριβή αυτή.

2.2.3. Silhouette score

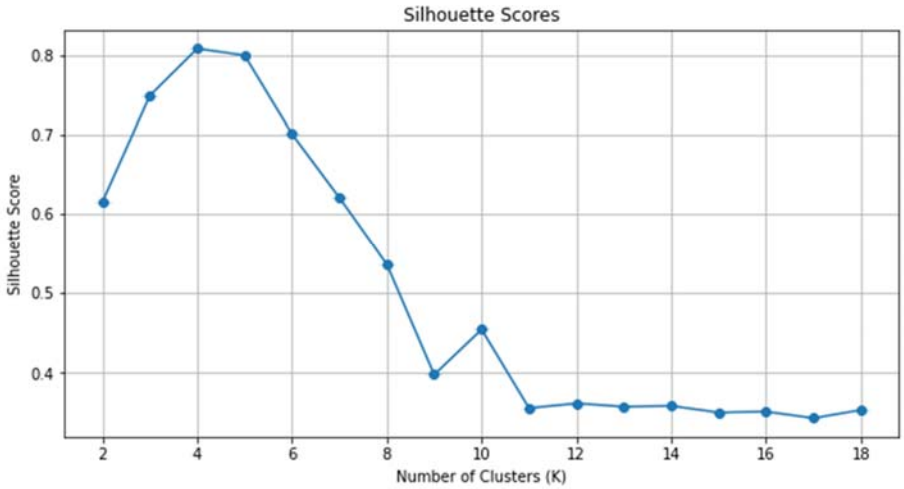
Αυτή η βαθμολογία, λαμβάνει υπόψη δύο μέτρα:

- Τη μέση απόσταση μεταξύ ενός δείγματος και όλων των άλλων σημείων στο ίδιο σύμπλεγμα.
- Τη μέση απόσταση μεταξύ ενός δείγματος και όλων των άλλων σημείων στο επόμενο κοντινότερο σύμπλεγμα.

δηλαδή ελέγχει πόσο συμπαγή είναι τα συμπλέγματα και καλά διαχωρισμένα. Όσο περισσότερο το σκορ είναι κοντά στο ένα, τόσο καλύτερη είναι η ομαδοποίηση. [14]

Εφόσον γνωρίζουμε ήδη ότι η διαδικασία προσαρμογής δεν είναι ντετερμινιστική, εκτελούμε είκοσι προσαρμογές για κάθε αριθμό συστάδων και, στη συνέχεια, λαμβάνουμε υπόψη τη μέση τιμή και την τυπική απόκλιση των καλύτερων πέντε εκτελέσεων.

Ας δούμε ότι περιγράψαμε μέσω ενός παραδείγματος υλοποίησης που φαίνεται στην παρακάτω εικόνα:

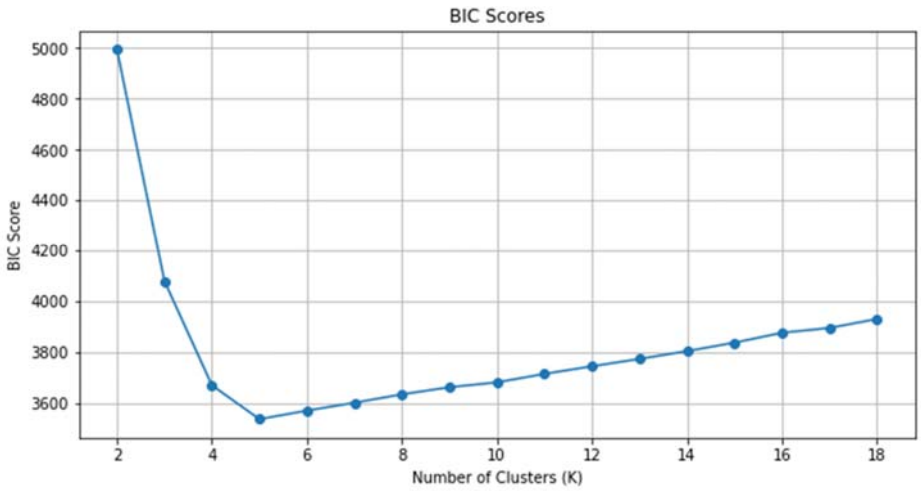


9 Silhouette Scores συναρτήσει του αριθμού των clusters για τον αλγόριθμο GMM [2]

Φαίνεται καθαρά ότι έχουμε την καλύτερη βαθμολογία με τέσσερα clusters. Πρέπει επίσης να λάβουμε υπόψη ότι και η διαμόρφωση των πέντε clusters είναι σχεδόν εξίσου καλή, αν λάβουμε υπόψη την τυπική απόκλιση (το «σφάλμα») και των δύο διαμορφώσεων. Άρα, αυτή η βαθμολογία δεν μας δίνει, πάντα, ξεκάθαρο αποτέλεσμα για τα δεδομένα μας.

2.2.4. Bayesian information criterion (BIC)

Αυτό το κριτήριο μας δίνει μια εκτίμηση για το πόσο καλό είναι το GMM όσον αφορά την πρόβλεψη των δεδομένων που έχουμε στην πραγματικότητα. Όσο χαμηλότερο είναι το BIC, τόσο καλύτερο είναι το μοντέλο για να προβλέψουμε πραγματικά τα δεδομένα που έχουμε, και κατ' επέκταση, την αληθινή, άγνωστη, κατανομή. Προκειμένου να αποφευχθεί η υπερβολική προσαρμογή, αυτή η τεχνική τιμωρεί μοντέλα με μεγάλο αριθμό clusters.[15]

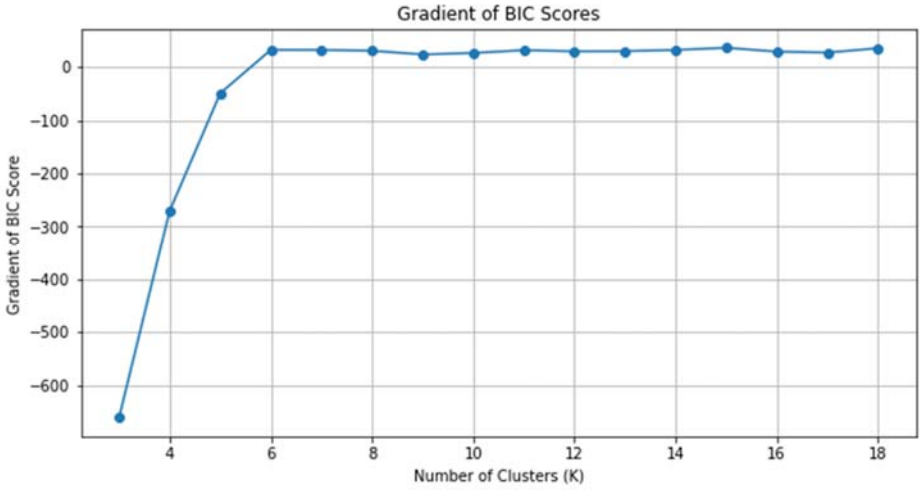


10 BIC Scores συναρτήσει του αριθμού των clusters για τον αλγόριθμο GMM [2]

Ακολουθώντας αυτό το κριτήριο, όσο μεγαλύτερος είναι ο αριθμός των clusters, τόσο καλύτερο θα πρέπει να είναι το μοντέλο. Πράγμα που σημαίνει ότι η ποινή που δίνουν τα κριτήρια BIC στα πολύπλοκα μοντέλα δεν μας σώζουν από την υπερπροσαρμογή.

Πριν απορρίψουμε αυτήν την τεχνική, μπορούμε να παρατηρήσουμε δύο πράγματα. Το πρώτο είναι ότι η καμπύλη είναι αρκετά ομαλή και μονότονη. Το δεύτερο είναι ότι η καμπύλη ακολουθεί διαφορετικές κλίσεις σε διαφορετικό τμήμα της. Ξεκινώντας από αυτές τις δύο παρατηρήσεις, ελέγχουμε πού είναι μεγάλη η κλίση της αλλαγής της καμπύλης BIC.

Τεχνικά, πρέπει να υπολογίσουμε την κλίση της καμπύλης βαθμολογιών BIC. Διαισθητικά, η έννοια της κλίσης είναι απλή: εάν δύο διαδοχικά σημεία έχουν την ίδια τιμή, η κλίση τους είναι μηδέν. Εάν έχουν διαφορετικές τιμές, η κλίση τους μπορεί να είναι είτε αρνητική, εάν το δεύτερο σημείο έχει χαμηλότερη τιμή ή θετική διαφορετικά. Το μέγεθος της κλίσης μας λέει πόσο διαφορετικές είναι οι δύο τιμές.



11 Κλίση της καμπύλης των BIC scores για επιλογή του αριθμού των clusters [2]

Όπως ήταν αναμενόμενο, όλες οι διαβαθμίσεις μέχρι πριν την τιμή για clusters ίση με έξι έχουν αρνητικές τιμές. Ακόμη, βλέπουμε πιο ξεκάθαρα ότι ξεκινώντας από τον αριθμό έξι, η διαβάθμιση γίνεται σχεδόν σταθερή, δηλαδή η αρχική συνάρτηση έχει μια πιο ήπια αύξηση, δηλαδή δεν υπάρχει μεγάλο κέρδος στην αύξηση του αριθμού των clusters. Εν ολίγοις, αυτή η τεχνική μας προτείνει να χρησιμοποιήσουμε πέντε clusters.

2.2.5. Εκτίμηση Παραμέτρων - Αλγόριθμος EM

Σε αυτό το σημείο θα ασχοληθούμε με το πως μπορούμε να καθορίσουμε τις τιμές των παραμέτρων ενός Γκαουσιανού Μείγματος, ώστε αυτό να περιγράφει κατά τον καλύτερο δυνατό τρόπο ένα σύνολο παρατηρήσεων. Συγκεκριμένα, έστω ότι έχουμε μια τυχαία μεταβλητή **X** για την οποία υποθέτουμε ότι ακολουθεί μια κατανομή Γκαουσιανού Μείγματος *K* συνιστωσών με άγνωστες παραμέτρους **Θ**. Για την ώρα, θα θεωρήσουμε ότι ο αριθμός, K, των πυρήνων στο μείγμα είναι γνωστός και σταθερός. Επίσης, θεωρούμε ότι έχουμε ένα σύνολο παρατηρήσεων S για την τυχαία μεταβλητή **X**. Το πρόβλημα που πρέπει

να αντιμετωπίσουμε αφορά την εκτίμηση, με βέλτιστο τρόπο, των τιμών των παραμέτρων **Θ**. [16]

Θα στηριχθούμε στην αρχή της μέγιστης πιθανοφάνειας. Όμως εδώ τα πράγματα είναι λίγο πιο περίπλοκα από ό,τι στην περίπτωση της απλής Γκαουσιανής, γιατί, ένα στιγμιότυπο της μεταβλητής παράγεται σε δύο στάδια. Το πρώτο στάδιο, η επιλογή δηλαδή του πυρήνα που αντιστοιχεί στην κάθε παρατήρηση, είναι «κρυφό» στην περίπτωσή μας. Με άλλα λόγια, έχουμε μεν τις παρατηρήσεις, αλλά δε μπορούμε να ξέρουμε από ποια Γκαουσιανή του μείγματος προήλθε η κάθε παρατήρηση. Το πρόβλημα αυτό είναι ένα κλασικό πρόβλημα ύπαρξης μη παρατηρήσιμων (unobserved) μεταβλητών. Ο κλασικός αλγόριθμος επίλυσής του είναι ο αλγόριθμος «μεγιστοποίησης της προσδοκίας» (Expectation-Maximization), ΕΜ. [16]

Ο αλγόριθμος ΕΜ είναι επαναληπτικός και περιλαμβάνει δύο βήματα. Συγκεκριμένα, έχουμε το βήμα Ε (Expectation) και το βήμα Μ (Maximization), για τα οποία ισχύουν τα παρακάτω:

1. Βήμα Ε: Υπολογισμός των προσδοκώμενων τιμών για κάθε κρυφή μεταβλητή, υποθέτοντας ότι οι τρέχουσες εκτιμήσεις των παραμέτρων ισχύουν.
2. Βήμα Μ: Υπολογισμός μιας νέας εκτίμησης μέγιστης πιθανοφάνειας για τις παραμέτρους, θεωρώντας ότι οι κρυφές μεταβλητές έχουν τιμές ίσες με τις προσδοκώμενες.

Στην προκειμένη περίπτωση θεωρούμε ότι κάθε παρατήρηση x_i είναι ένα *d*-διάστατο διάνυσμα [17]

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{id})^T$$

με το οποίο συνδέονται *K* κρυφές μεταβλητές, z_{i1}, z_{i2}, …, z_{iK}, οι οποίες καθορίζουν από ποια από τις *K* Γκαουσιανές του μείγματος παρήχθη η εν λόγω παρατήρηση. Οι δυνατές τιμές που μπορεί να πάρει κάθε μια από αυτές τις μεταβλητές είναι οι {0,1}. Το 1 (0) αντιστοιχεί στην περίπτωση κατά την οποία η παρατήρηση (δεν) προήλθε από τον αντίστοιχο πυρήνα. Επίσης, για κάθε i μόνο μια από τις K αντίστοιχες μεταβλητές φέρει την τιμή 1 και όλες οι άλλες ισούνται με 0, αφού μια συγκεκριμένη παρατήρηση μπορεί να παραχθεί από έναν και μόνο πυρήνα.

Το πρώτο βήμα του αλγορίθμου ΕΜ αφορά τον υπολογισμό των προσδοκώμενων τιμών $E(z_{ij})$ για τις προαναφερθείσες κρυφές μεταβλητές. Όπως έχουμε δει ήδη, η προσδοκώμενη τιμή για μια διακριτή τυχαία μεταβλητή Z δίνεται από τον παρακάτω τύπο:

$$E[z] = \sum_{\forall z} zP(Z = z)$$

Επομένως, για τις τυχαίες μεταβλητές που εξετάζουμε ισχύει ότι $E(z_{ij}) = 0P(z_{ij}=0) + 1P(z_{ij}=1)$ και άρα:

$$E(z_{ij}) = P(z_{ij}=1)$$

Όμως, η $P(z_{ij}=1)$, δηλαδή η πιθανότητα η κρυφή μεταβλητή z_{ij} να έχει την τιμή 1, προφανώς ισούται με την πιθανότητα ο j-οστός πυρήνας να έχει παράξει την i-οστή παρατήρηση. Αυτό μπορούμε να το διατυπώσουμε συμβολικά ως

$$P(z_{ij}=1) = P(j \mid \mathbf{x}_{ij})$$

Σύμφωνα με τον τύπο του Bayes, όπως αυτός δίνεται από την παρακάτω σχέση:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

προκύπτει ότι

$$P(j \mid \mathbf{x}_i) = \frac{P(j)p(\mathbf{x}_i \mid j)}{p(\mathbf{x}_i)}$$

Ενσωματώνοντας τώρα την υπόθεσή μας ότι οι παρατηρήσεις ακολουθούν κατανομή Γκαουσιανού Μείγματος, παίρνουμε την τελική μορφή:

$$P(j \mid \mathbf{x}_i) = \frac{\pi_j \phi(\mathbf{x}_i \mid \boldsymbol{\theta}_j)}{\sum_{k=1}^{K} \pi_k \phi(\mathbf{x}_i \mid \boldsymbol{O}_k)}$$

όπου η $\varphi(\mathbf{x} \mid \boldsymbol{\theta}_j)$ είναι j-οστή Γκαουσιανή συνάρτηση πυκνότητας πιθανότητας του μείγματος. Επομένως, από τα παραπάνω, η προσδοκώμενη τιμή για τις κρυφές μεταβλητές z_{ij} υπολογίζεται ως

$$\mathrm{E}[z_{ij}] = \frac{\pi_j \phi(\mathbf{x}_i \mid \boldsymbol{\theta}_j)}{\sum_{k=1}^{K} \pi_k \phi(\mathbf{x}_i \mid \boldsymbol{\theta}_k)}$$

Στο δεύτερο βήμα του αλγορίθμου ΕΜ χρησιμοποιούμε αυτές τις εκτιμήσεις για τις τιμές των κρυφών μεταβλητών, ώστε να υπολογίσουμε μια νέα εκτίμηση για τις παραμέτρους **Θ** του μοντέλου. Το συγκεκριμένο βήμα ανάγεται απευθείας στην απλή περίπτωση υπολογισμού των παραμέτρων μιας Γκαουσιανής κατανομής από ένα σύνολο παρατηρήσεων, αφού κατά το βήμα αυτό θεωρούμε ότι οι κρυφές μεταβλητές έχουν τιμές ίσες με τις προσδοκώμενες που υπολογίστηκαν στο προηγούμενο βήμα. Όμως, από τη στιγμή που ξέρουμε τις τιμές των κρυφών μεταβλητών, ουσιαστικά ξέρουμε ποιες παρατηρήσεις αντιστοιχούν σε ποιους πυρήνες και, επομένως, από τις παρατηρήσεις που αντιστοιχούν σε κάθε πυρήνα μπορούμε να εκτιμήσουμε τις βέλτιστες τιμές για τις παραμέτρους του. Συνεπώς, το αποτέλεσμα για τις νέες τιμές των παραμέτρων **Θ** προκύπτει με αρκετά διαισθητικό τρόπο, ως [17] :

$$\pi_j = \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}[z_{ij}]$$

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{N} \mathrm{E}[z_{ij}]\mathbf{x}_i}{\sum_{i=1}^{N} \mathrm{E}[z_{ij}]}$$

$$\boldsymbol{\Sigma}_j = \frac{\sum_{i=1}^{N} \mathrm{E}[z_{ij}]\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\sum_{i=1}^{N} \mathrm{E}[z_{ij}]} - \boldsymbol{\mu}_j\boldsymbol{\mu}_j^{\mathrm{T}}$$

Προφανώς, με βάση την ανωτέρω πιθανοτική ανάλυση και τον υπολογισμό των προσδοκώμενων τιμών για τις κρυφές μεταβλητές z_{ij} , κάθε παρατήρηση x_i δεν αντιστοιχίζεται σε μια μόνο Γκαουσιανή αλλά σε όλες, με διαφορετικό βαθμό στην καθεμιά όμως, και για αυτό το λόγο χρησιμοποιούνται σταθμισμένοι μέσοι όροι για τον υπολογισμό των παραμέτρων.

Πρακτικά μιλώντας πλέον, ο αλγόριθμος ΕΜ λειτουργεί ως εξής: Αρχικά θεωρούμε μια πρώτη εκτίμηση των παραμέτρων του μοντέλου. Για παράδειγμα, μπορούμε να θεωρήσουμε τα βάρη π_j των πυρήνων ίσα μεταξύ τους, κάθε μια από τις μέσες τιμές $\boldsymbol{\mu}_j$ να την εξισώσουμε με κάποια τυχαία επιλεγμένη παρατήρηση $\mathbf{x}_i$, και να θεωρήσουμε τους πίνακες συνδιακύμανσης $\boldsymbol{\Sigma}_j$ όλους ίσους με κάποιο κατάλληλο (θεωρώντας τη διασπορά του συνόλου των παρατηρήσεων) πολλαπλάσιο του αντίστοιχου μοναδιαίου πίνακα. Ένας πιο ενδεδειγμένος από θεωρητικής άποψης τρόπος αρχικοποίησης είναι να στηριχθούμε στον αλγόριθμο ομαδοποίησης k-means, και να χρησιμοποιήσουμε τα αποτελέσματα που μας δίνει αν τον εφαρμόσουμε στο σύνολο των παρατηρήσεων για Κ ομάδες (clusters), όσες και οι Γκαουσιανές του μοντέλου. Σε αυτήν την περίπτωση, ο αλγόριθμος k-means θα χωρίσει το σύνολο των παρατηρήσεων σε Κ υποσύνολα. Έπειτα, καθένα από αυτά τα υποσύνολα μπορεί να χρησιμοποιηθεί για την αρχικοποίηση μιας από τις Γκαουσιανές του μοντέλου. Συγκεκριμένα, για το j-οστό υποσύνολο μπορούμε να υπολογίσουμε το ποσοστό του συνόλου των παρατηρήσεων που του αντιστοιχούν, τη μέση τιμή και τη συνδιακύμανσή τους, και με αυτά να αρχικοποιήσουμε τις παραμέτρους π_j , $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$, αντίστοιχα, της j-οστής Γκαουσιανής κατανομής του μείγματος. [17]

Μετά από την αρχικοποίηση, θεωρώντας επομένως ότι έχουμε μια αρχική εκτίμηση για τις παραμέτρους $\boldsymbol{\Theta}$, έστω $\boldsymbol{\Theta}^0$, εκτελούμε το πρώτο βήμα του αλγορίθμου ΕΜ και παίρνουμε μια εκτίμηση για τις κρυφές μεταβλητές z_{ij} . Έπειτα, εκτελούμε το δεύτερο βήμα του ΕΜ και ανανεώνουμε την εκτίμησή μας για τις παραμέτρους $\boldsymbol{\Theta}$, από $\boldsymbol{\Theta}^0$ σε $\boldsymbol{\Theta}^1$. Κατόπιν, επιστρέφουμε πάλι στο πρώτο βήμα του ΕΜ και ούτω καθεξής. Συνεχίζοντας κατά αυτόν τρόπο και ακολουθώντας επαναληπτικά τα βήματα του αλγορίθμου ΕΜ, παίρνουμε διαδοχικές εκτιμήσεις, $\boldsymbol{\Theta}^0$, $\boldsymbol{\Theta}^1$,..., $\boldsymbol{\Theta}^t$,... για τις παραμέτρους του μοντέλου, και σταματάμε όταν ο αλγόριθμος συγκλίνει, δηλαδή όταν οι παράμετροι από επανάληψη σε επανάληψη πάψουν να αλλάζουν σημαντικά. Ένα άλλο κριτήριο τερματισμού – σύγκλισης προκύπτει από την παρατήρηση της τρέχουσας πιθανοφάνειας των παρατηρήσεων $\mathcal{L}$ ($\boldsymbol{\Theta}^t$), που ως μέτρο έχει την τάση να αυξάνεται από επανάληψη σε επανάληψη. Όταν η πιθανοφάνεια σταθεροποιηθεί μέσα στα όρια κάποιας προκαθορισμένης ακρίβειας, αυτό αποτελεί ένδειξη ότι ο αλγόριθμος έχει συγκλίνει.

Από την προηγηθείσα ανάλυση αποκαλύπτεται, λοιπόν, άλλο ένα πλεονέκτημα των Γκαουσιανών Μειγμάτων, το οποίο είναι η σχετικά εύκολη και μαθηματικά θεμελιωμένη εκπαίδευση του μοντέλου με δεδομένο ένα σύνολο παρατηρήσεων. Όμως, το πρόβλημα που πρέπει να επισημανθεί είναι το εξής: Ο αλγόριθμος ΕΜ, όπως αναφέρθηκε και παραπάνω, απαιτεί κάποια αρχικοποίηση, μια αρχική εκτίμηση των παραμέτρων του μοντέλου. Έπειτα, συγκλίνει σε ένα τοπικό μέγιστο της πιθανοφάνειας μέσω της επαναληπτικής εκτέλεσης των βημάτων που αναλύθηκαν προηγουμένως. Ο χώρος των παραμέτρων, όμως, είναι γενικά μεγάλης διάστασης και η συνάρτηση της πιθανοφάνειας

πολύπλοκη, με πολλά τοπικά ακρότατα. Αυτό έχει ως αποτέλεσμα η αρχικοποίηση να παίζει μεγάλο ρόλο ως προς το σημείο στο χώρο των παραμέτρων στο οποίο θα συγκλίνει ο αλγόριθμος. Με άλλα λόγια, ο κλασικός αλγόριθμος ΕΜ που περιγράψαμε σε αυτήν την ενότητα έχει μεγάλη εξάρτηση από την αρχικοποίηση. Για να αντιμετωπιστεί κάπως αυτό το πρόβλημα έχει προταθεί η εκτέλεση του αλγορίθμου περισσότερες της μίας φορές, με διαφορετική αρχικοποίηση την κάθε φορά, και η επιλογή του αποτελέσματος εκείνου που αντιστοιχεί σε μεγαλύτερη τελική πιθανοφάνεια των παρατηρήσεων. Όμως, αυτό, πέραν του ότι είναι πολύ απλοϊκό ως τρόπος αντιμετώπισης, οδηγεί και σε άσκοπη χρονική επιβάρυνση της εκπαίδευσης, αφού για το ίδιο σύνολο παρατηρήσεων, εκτός από το ένα μοντέλο που θα χρησιμοποιηθεί εν τέλει, εκπαιδεύονται και αρκετά άλλα που μένουν αχρησιμοποίητα.[17]

2.3. Αλυσίδες Markov

2.3.1. Εισαγωγικά

Μια αλυσίδα Markov είναι ένα μαθηματικό σύστημα που βιώνει μεταβάσεις από τη μια κατάσταση στην άλλη σύμφωνα με ορισμένους πιθανολογικούς κανόνες. Το καθοριστικό χαρακτηριστικό μιας αλυσίδας Markov είναι ότι ανεξάρτητα από το πώς έφτασε η διαδικασία στην παρούσα κατάστασή της, οι πιθανές μελλοντικές καταστάσεις είναι σταθερές. Με άλλα λόγια, η πιθανότητα μετάβασης σε οποιαδήποτε συγκεκριμένη κατάσταση εξαρτάται αποκλειστικά από την τρέχουσα κατάσταση και τον χρόνο που έχει παρέλθει. Ο χώρος κατάστασης, ή το σύνολο όλων των πιθανών καταστάσεων, μπορεί να είναι οτιδήποτε: γράμματα, αριθμοί, καιρικές συνθήκες, σκορ του μπέιζμπολ ή παραστάσεις μετοχών.[18]

Οι αλυσίδες Markov μπορούν να μοντελοποιηθούν από μηχανές πεπερασμένης κατάστασης και οι «τυχαίοι περίπατοι» παρέχουν ένα χαρακτηριστικό παράδειγμα της χρησιμότητάς τους στα μαθηματικά. Προκύπτουν ευρέως σε στατιστικά και θεωρητικά πλαίσια πληροφοριών και χρησιμοποιούνται ευρέως στα οικονομικά, τη θεωρία παιγνίων, τη θεωρία της ουράς (τηλεπικοινωνίες), τη γενετική και τα χρηματοοικονομικά. Ενώ είναι δυνατό να μελετηθούν αλυσίδες Markov με οποιοδήποτε μέγεθος χώρου καταστάσεων, η αρχική θεωρία και οι περισσότερες εφαρμογές επικεντρώνονται σε περιπτώσεις με πεπερασμένο (ή μετρήσιμο άπειρο) αριθμό καταστάσεων.[18]

2.3.2. Ιδιότητα Markov και αλυσίδα Markov

Μια ιδιότητα που κάνει τη μελέτη μιας τυχαίας διαδικασίας πολύ πιο εύκολη είναι η «ιδιότητα Markov». Σε απλή μετάφραση, η ιδιότητα Markov λέει, για μια τυχαία διαδικασία, ότι εάν γνωρίζουμε την τιμή που λαμβάνει η διαδικασία σε μια δεδομένη στιγμή, δεν θα λάβουμε πρόσθετες πληροφορίες σχετικά με τη μελλοντική συμπεριφορά της διαδικασίας συλλέγοντας περισσότερη γνώση για το παρελθόν. Με λίγο πιο μαθηματικούς όρους, για κάθε δεδομένη στιγμή, η υπό όρους κατανομή των μελλοντικών καταστάσεων της διεργασίας δεδομένης παρούσας και προηγούμενης κατάστασης εξαρτάται μόνο από την παρούσα κατάσταση και καθόλου από τις προηγούμενες καταστάσεις **(ιδιότητα χωρίς μνήμη)**. Μια τυχαία διαδικασία με την ιδιότητα Markov ονομάζεται **διαδικασία Markov**.[19]

Με βάση τον προηγούμενο ορισμό, μπορούμε τώρα να ορίσουμε τις «ομογενείς αλυσίδες Markov διακριτού χρόνου» (που θα δηλωθούν ως «αλυσίδες Markov» για απλότητα στη συνέχεια). Μια αλυσίδα Markov είναι μια διαδικασία Markov με διακριτό χρόνο και διακριτό χώρο κατάστασης. Δηλαδή, μια αλυσίδα Markov είναι μια διακριτή ακολουθία καταστάσεων, που η καθεμία προέρχεται από έναν διακριτό χώρο καταστάσεων (πεπερασμένος ή όχι) και ακολουθεί την ιδιότητα Markov.

Μαθηματικά, μπορούμε να υποδηλώσουμε μια αλυσίδα Markov όπως παρακάτω

$$X = (X_n)_{n\in\mathbb{N}} = (X_0, X_1, X_2, \dots)$$

όπου σε κάθε χρονική στιγμή η διεργασία παίρνει τις τιμές της σε ένα διακριτό σύνολο E έτσι ώστε

$$X_n \in E \quad \forall n \in \mathbb{N}$$

Έτσι, η ιδιότητα Markov υποδηλώνει ότι έχουμε

$$\mathbb{P}(X_{n+1} = s_{n+1} \mid X_n = s_n, X_{n-1} = s_{n-1}, X_{n-2} = s_{n-2}, \dots) = \mathbb{P}(X_{n+1} = s_{n+1} \mid X_n = s_n)$$

Παρατηρούμε για άλλη μια φορά ότι αυτός ο τελευταίος τύπος εκφράζει το γεγονός ότι για ένα δεδομένο ιστορικό (που βρίσκομαι τώρα και πού ήμουν πριν), η κατανομή πιθανοτήτων για την επόμενη κατάσταση (όπου πάω μετά) εξαρτάται μόνο από την τρέχουσα κατάσταση και όχι από την προηγούμενες.[19]

$$P(\text{ future} \mid \text{present, past}) = P(\text{ future} \mid \text{present, } \sout{\text{past}})$$

2.3.3. Πίνακες μετάβασης

Ένας πίνακας μετάβασης P_t για την αλυσίδα Markov {X} τη χρονική στιγμή t είναι ένας πίνακας που περιέχει πληροφορίες σχετικά με την πιθανότητα μετάβασης μεταξύ των καταστάσεων. Ειδικότερα, δεδομένης της ταξινόμησης των γραμμών και των στηλών ενός πίνακα από τον χώρο κατάστασης S, το (i, j)-στο στοιχείο του πίνακα P_t δίνεται από την :

$$(P_t)_{i,j} = \mathbb{P}(X_{t+1} = j \mid X_t = i)$$

Αυτό σημαίνει ότι κάθε γραμμή του πίνακα είναι ένα διάνυσμα πιθανότητας και το άθροισμα των εγγραφών του είναι 1.[19]

Οι πίνακες μετάβασης έχουν την ιδιότητα ότι το γινόμενο των επόμενων περιγράφει μια μετάβαση κατά μήκος του χρονικού διαστήματος που εκτείνεται από τους πίνακες μετάβασης. Δηλαδή, το $P_0 \circ P_1$ έχει στη θέση (i, j) πιθανότητα να είναι $X_2 = j$ δεδομένου ότι $X_0 = i$. Γενικά, το (i, j)-στο στοιχείο του $P_t \circ P_{t+1} \dots P_{t+k}$ είναι η πιθανότητα $\mathbb{P}(X_{t+k+1} = j \mid X_t = i)$

Ο πίνακας μετάβασης βήματος-*k* είναι ο

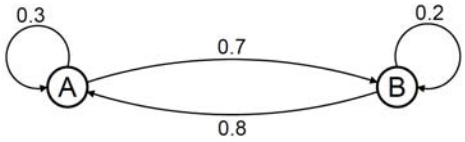
$$P_t^{(k)} = \begin{pmatrix} \mathbb{P}(X_{t+k} = 1 \mid X_t = 1) & \mathbb{P}(X_{t+k} = 2 \mid X_t = 1) & \dots & \mathbb{P}(X_{t+k} = n \mid X_t = 1) \\ \mathbb{P}(X_{t+k} = 1 \mid X_t = 2) & \mathbb{P}(X_{t+k} = 2 \mid X_t = 2) & \dots & \mathbb{P}(X_{t+k} = n \mid X_t = 2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{P}(X_{t+k} = 1 \mid X_t = n) & \mathbb{P}(x_{t+k} = 2 \mid X_t = n) & \dots & \mathbb{P}(X_{t+k} = n \mid X_t = n) \end{pmatrix}$$

και με βάση τα παραπάνω έχουμε:

$$P_t^{(k)} = P_t \cdot P_{t+1} \cdots P_{t+k-1}$$

2.3.4. Παράδειγμα

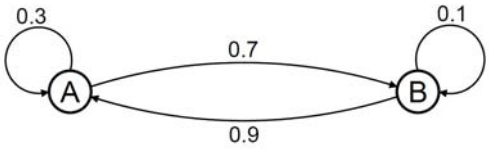
1) Μια (χρονικά ομοιογενής) αλυσίδα Markov χτισμένη στις καταστάσεις Α και Β απεικονίζεται στο παρακάτω διάγραμμα. Ποια είναι η πιθανότητα μια διεργασία που ξεκινά στο Α να βρίσκεται στο Β μετά από 2 κινήσεις;



Για να μετακινηθεί από το Α στο Β, η διαδικασία πρέπει είτε να παραμείνει στο Α κατά την πρώτη κίνηση και να μετακινηθεί στο Β στη δεύτερη κίνηση. ή να μετακινηθεί στο Β στην πρώτη κίνηση και μετά να μείνει στο Β στη δεύτερη κίνηση. Σύμφωνα με το διάγραμμα, η πιθανότητα είναι 0.3*0.7 + 0.7*0.3 = 0.35

Εναλλακτικά, η πιθανότητα η διαδικασία να είναι στο Α μετά από 2 κινήσεις είναι 0.3*0.3+ 0.7*0.8 = 0.65. Δεδομένου ότι υπάρχουν μόνο δύο καταστάσεις στην αλυσίδα, η διεργασία πρέπει να βρίσκεται στο Β αν δεν είναι στο Α, και επομένως η πιθανότητα η διεργασία να είναι στο Β μετά από 2 κινήσεις είναι 1 – 0.65 = 0.35

2) Για την ανεξάρτητη από το χρόνο αλυσίδα Markov που περιγράφεται στην παρακάτω εικόνα



Ο πίνακας μετάβασης είναι

$$P = \begin{pmatrix} 0.3 & 0.7 \\ 0.9 & 0.1 \end{pmatrix}$$

Από αυτό προκύπτει ότι ο πίνακας μετάβασης βήματος-2 είναι

$$P^2 = \begin{pmatrix} 0.3 & 0.7 \\ 0.9 & 0.1 \end{pmatrix} * \begin{pmatrix} 0.3 & 0.7 \\ 0.9 & 0.1 \end{pmatrix} = \begin{pmatrix} 0.72 & 0.28 \\ 0.36 & 0.64 \end{pmatrix}$$

2.4. Αποσύνθεση Χρονοσειρών

2.4.1. Είδη αποσύνθεσης

Τα δεδομένα χρονοσειρών μπορούν να παρουσιάσουν μια ποικιλία μοτίβων και συχνά είναι χρήσιμο να χωρίσουμε μια χρονοσειρά σε πολλά στοιχεία, καθένα από τα οποία αντιπροσωπεύει μια υποκείμενη κατηγορία προτύπων.

Κάθε χρονοσειρά αποτελείται από τρεις τουλάχιστον τύπους μοτίβων χρονοσειρών: την τάση, την εποχικότητα και τον κύκλο. Όταν αποσυνθέτουμε μια χρονική σειρά σε στοιχεία, συνήθως συνδυάζουμε την τάση και τον κύκλο σε ένα ενιαίο στοιχείο κύκλου-τάσης (μερικές φορές ονομάζεται τάση για απλότητα). Ως εκ τούτου, θεωρούμε ότι μια χρονοσειρά περιλαμβάνει τρία στοιχεία: μια συνιστώσα του κύκλου-τάσης, μια εποχιακή συνιστώσα και μια συνιστώσα υπόλοιπο (που περιέχει οτιδήποτε άλλο στη χρονοσειρά).

Παρακάτω θα εξετάζουμε μερικές κοινές μεθόδους για την εξαγωγή αυτών των στοιχείων από μια χρονολογική σειρά. Συχνά αυτό γίνεται για να βοηθήσει στη βελτίωση της κατανόησης των χρονοσειρών, αλλά μπορεί επίσης να χρησιμοποιηθεί για τη βελτίωση της ακρίβειας πρόβλεψης. [20]

Αν υποθέσουμε μια προσθετική αποσύνθεση, τότε μπορούμε να γράψουμε την χρονοσειρά ως:

$$y_t = S_t + T_t + R_t$$

όπου y_t είναι τα δεδομένα, S_t είναι η εποχιακή συνιστώσα, T_t είναι η συνιστώσα του κύκλου-τάσης και R_t είναι η υπόλοιπη συνιστώσα, όλα στην περίοδο t. Εναλλακτικά, μια πολλαπλασιαστική αποσύνθεση θα γραφόταν ως:

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

Η προσθετική αποσύνθεση είναι η καταλληλότερη εάν το μέγεθος των εποχιακών διακυμάνσεων ή η διακύμανση γύρω από τον κύκλο-τάσης δεν ποικίλλει ανάλογα με το επίπεδο της χρονοσειράς. Όταν η διακύμανση στο εποχιακό μοτίβο ή η διακύμανση γύρω από τον κύκλο-τάσης φαίνεται να είναι ανάλογη με το επίπεδο της χρονοσειράς, τότε είναι πιο κατάλληλη μια πολλαπλασιαστική αποσύνθεση. Οι πολλαπλασιαστικές αποσυνθέσεις είναι κοινές με τις οικονομικές χρονοσειρές.

Μια εναλλακτική λύση στη χρήση της πολλαπλασιαστικής αποσύνθεσης είναι πρώτα να μετασχηματιστούν τα δεδομένα μέχρις ότου η διακύμανση της σειράς φαίνεται να είναι σταθερή με την πάροδο του χρόνου και στη συνέχεια να χρησιμοποιήσουμε μια προσθετική αποσύνθεση. Όταν χρησιμοποιείται ένας μετασχηματισμός καταγραφής, αυτό ισοδυναμεί με τη χρήση πολλαπλασιαστικής αποσύνθεσης επειδή

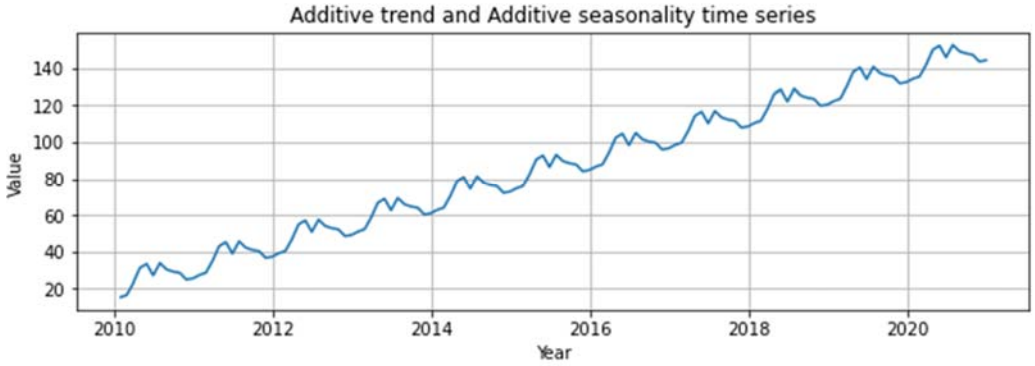
$y_t = S_t \times T_t \times R_t$ είναι ισοδύναμο με $\log y_t = \log S_t + \log T_t + \log R_t$ [20]

2.4.2. Είδη Χρονοσειρών

2.4.2.i. Προσθετική τάση και προσθετική εποχικότητα

Προσθετική τάση σημαίνει ότι η τάση είναι γραμμική (ευθεία γραμμή) και η προσθετική εποχικότητα σημαίνει ότι δεν υπάρχουν αλλαγές στα πλάτη ή τα ύψη των εποχιακών περιόδων με την πάροδο του χρόνου. [20]

Εδώ φαίνεται ένα παράδειγμα:



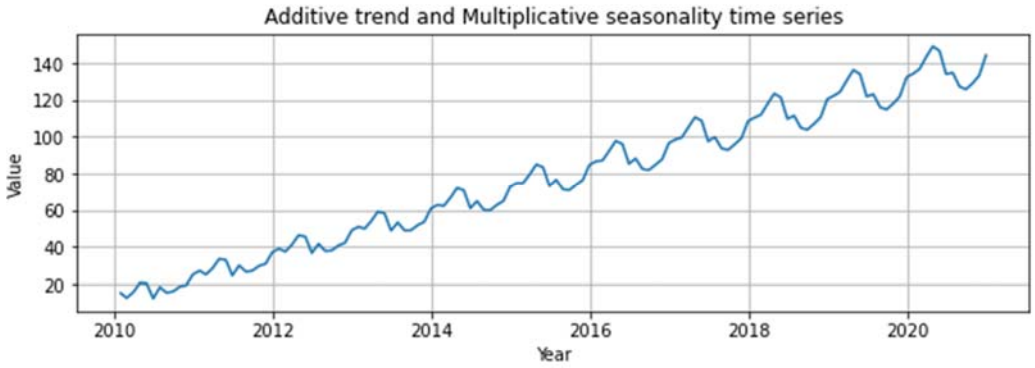
12 Χρονοσειρά με προσθετική τάση και προσθετική εποχικότητα [2]

Η χρονοσειρά δείχνει μια γραμμική τάση και εποχικότητα που δεν αλλάζει με την πάροδο του χρόνου. Δεν είναι η πιο τυπική χρονοσειρά, καθώς πιθανότατα το εύρος μιας εποχικής περιόδου θα αλλάξει με μια αυξητική τάση.

2.4.2.ii. Προσθετική τάση και πολλαπλασιαστική εποχικότητα

Η προσθετική τάση σημαίνει ότι η τάση είναι γραμμική (ευθεία γραμμή) και η πολλαπλασιαστική εποχικότητα σημαίνει ότι υπάρχουν αλλαγές στα πλάτη ή τα ύψη των εποχιακών περιόδων με την πάροδο του χρόνου. [20]

Εδώ φαίνεται ένα παράδειγμα:



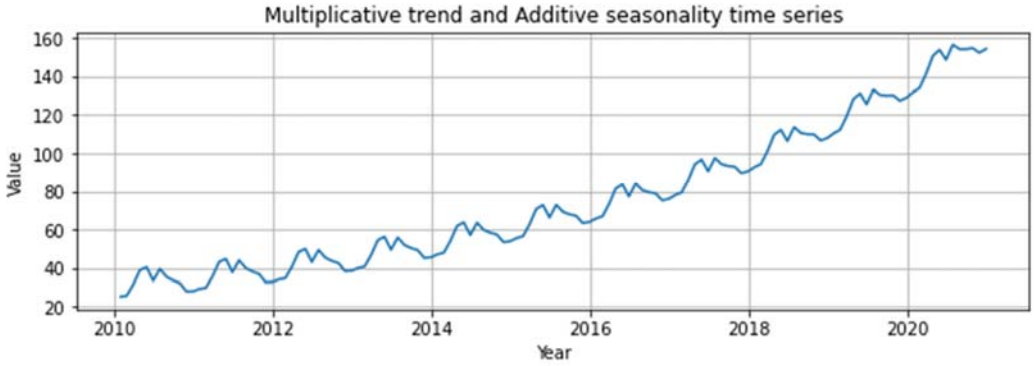
13 Χρονοσειρά με προσθετική τάση και πολλαπλασιαστική εποχικότητα [2]

Για άλλη μια φορά, η τάση είναι γραμμική, αλλά τα ύψη των εποχιακών περιόδων έχουν αυξηθεί με την πάροδο του χρόνου. Αυτή η συμπεριφορά είναι χαρακτηριστική για πολλές

χρονολογικές σειρές για προφανείς λόγους — περισσότερος όγκος (μια συνολική αύξηση στον άξονα y) εισάγει μεγαλύτερη μεταβλητότητα σε μια μόνο σεζόν.

2.4.2.iii. Πολλαπλασιαστική τάση και προσθετική εποχικότητα

Η πολλαπλασιαστική τάση σημαίνει ότι η τάση δεν είναι γραμμική (καμπύλη γραμμή) και η προσθετική εποχικότητα σημαίνει ότι δεν υπάρχουν αλλαγές στα πλάτη ή τα ύψη των εποχιακών περιόδων με την πάροδο του χρόνου.[20]

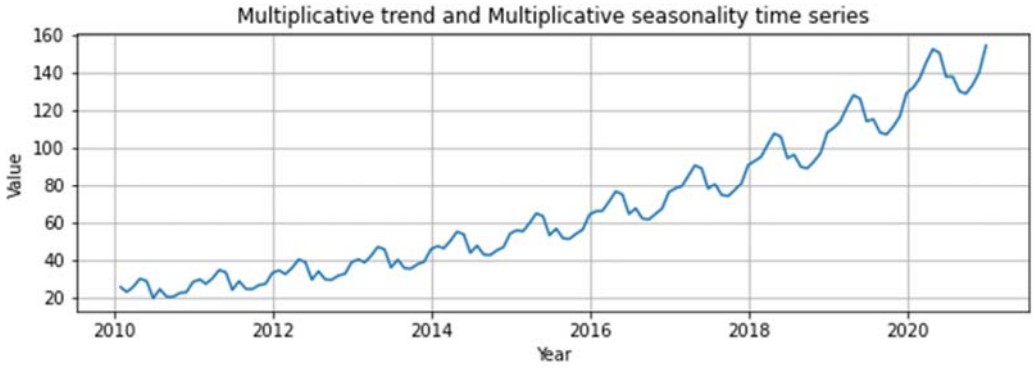


14 Χρονοσειρά με πολλαπλασιαστική τάση και προσθετική εποχικότητα [2]

Μπορούμε να δούμε πώς η τάση είναι ελαφρώς κυρτή. Δεν έχω δει πάρα πολλές χρονικές σειρές με αυτό το σχήμα επειδή οι εποχιακές περίοδοι τείνουν να διαφέρουν ως προς το πλάτος καθώς αυξάνεται η τιμή του άξονα y. Ωστόσο, εξακολουθεί να είναι ένα πιθανό σενάριο.

2.4.2.iv. Πολλαπλασιαστική τάση και πολλαπλασιαστική εποχικότητα

Η πολλαπλασιαστική τάση σημαίνει ότι η τάση δεν είναι γραμμική (καμπύλη γραμμή) και η πολλαπλασιαστική εποχικότητα σημαίνει ότι υπάρχουν αλλαγές στα πλάτη ή τα ύψη των εποχιακών περιόδων με την πάροδο του χρόνου. [20]



15 Χρονοσειρά με πολλαπλασιαστική τάση και πολλαπλασιαστική εποχικότητα [2]

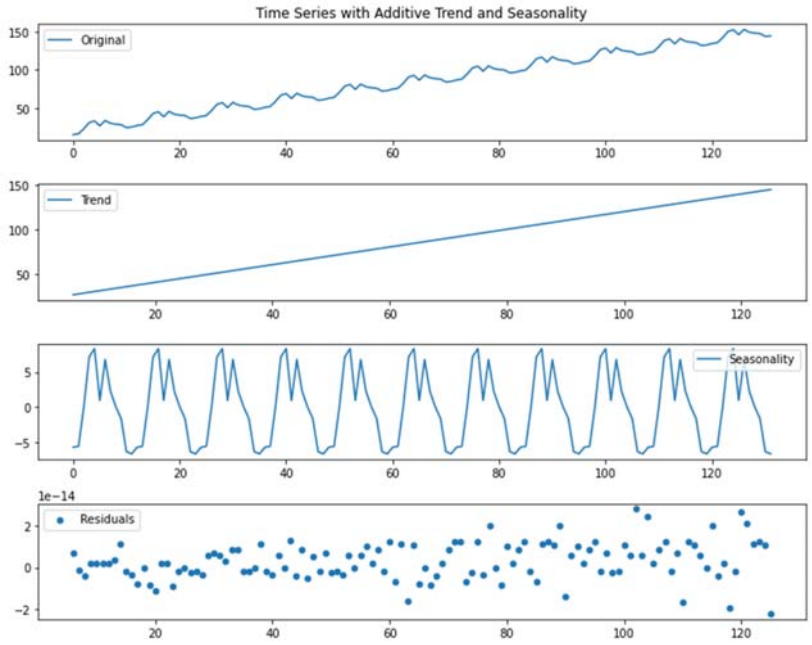
Αυτός είναι ο πιο ευρέως διαδεδομένος τύπος προτύπου στα δεδομένα χρονοσειρών. Παρατηρείται συχνά στα δεδομένα πωλήσεων, για παράδειγμα, όταν η ζήτηση για ένα

συγκεκριμένο προϊόν/υπηρεσία αυξάνεται με την πάροδο του χρόνου, αλλά οι περισσότερες από τις πωλήσεις γίνονται τους καλοκαιρινούς μήνες (σκεφτείτε τα αεροπορικά εισιτήρια).

Αφού παραθέσαμε τους διαφορετικούς τύπους μοτίβων που πρέπει να ξέρουμε πριν την αποσύνθεση μιας χρονοσειράς· στη συνέχεια, θα μάθουμε πώς να αναλύουμε πραγματικά μια χρονοσειρά σε τάσεις, εποχιακές και υπολειπόμενες συνιστώσες.

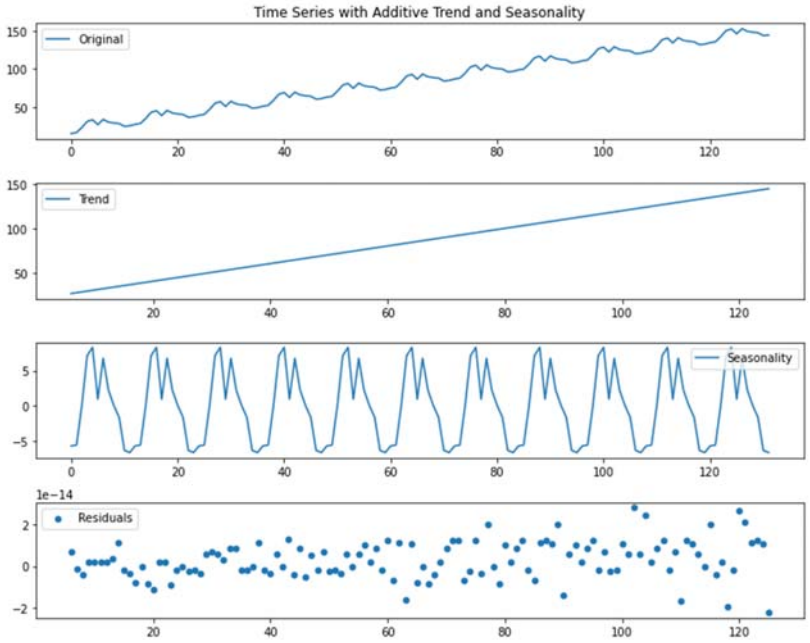
2.4.3. Παράδειγμα αποσύνθεσης

Σαν δεδομένα θα χρησιμοποιήσουμε μια χρονοσειρά με προσθετική τάση και εποχικότητα όπως φαίνεται στο πρώτο απ' τα παρακάτω τέσσερα σχήματα. Στη συνέχεια θα την αποσυνθέσουμε με βάση το προσθετικό μοντέλο αποσύνθεσης.



16 Αποσύνθεση χρονοσειράς με προσθετική τάση και εποχικότητα [2]

Στα υπόλοιπα σχήματα διακρίνονται κατά σειρά η τάση, η εποχικότητα και τα υπολείμματα. Στη συνέχεια, ας δούμε τι συμβαίνει εάν εφαρμόσουμε το πολλαπλασιαστικό μοντέλο αποσύνθεσης σε μια χρονοσειρά με πολλαπλασιαστική τάση και εποχικότητα:



17 Αποσύνθεση χρονοσειράς με πολλαπλασιαστική τάση και εποχικότητα [2]

Τα υπολείμματα είναι πλέον κεντραρισμένα γύρω στο 1 και έχουν πολύ μικρότερο εύρος και τυπική απόκλιση. Η διαφορά είναι εμφανής.

2.5. Bootstrap

2.5.1. Ανάλυση θεωρίας

Η βασική ιδέα που κρύβεται πίσω από το bootstrap είναι να εκτιμήσουμε τις ποσότητες που μας ενδιαφέρουν πραγματοποιώντας μια επαναδειγματοληψία με αντικατάσταση από το δείγμα που έχουμε στη διάθεσή μας. Όταν ασχολούμαστε με δεδομένα που συσχετίζονται χρονικά, η απλή διαδικασία επαναδειγματοληψίας αποτυγχάνει καθώς δεν είναι σε θέση να αναπαράγει τη δομή εξάρτησης. Για το λόγο αυτό, μελετώνται ορισμένες ad-hoc τεχνικές για την επίλυση της έλλειψης απόδοσης σε λειτουργίες bootstrap με χρονοσειρές. Τα πιο συνηθισμένα είναι γνωστά ως block bootstrap και residual bootstrap. Η προσέγγισή μας στο bootstrapping αποτελείται από έναν συνδυασμό των δύο αναφερόμενων μεθοδολογιών.[21]

Το Block bootstrap προσπαθεί να δημιουργήσει νέες σειρές, με την ίδια εξάρτηση των αρχικών δεδομένων, επαναδειγματοληπτώντας κομμάτια συνεχών παρατηρήσεων αντί για μεμονωμένες. Οι πιο συχνά χρησιμοποιούμενες μέθοδοι της οικογένειας των block εκκίνησης είναι το μη επικαλυπτόμενο block bootstrap (NBB), το κινούμενο block bootstrap (MBB), το κυκλικό block bootstrap (CBB) και το block bootstrap (SB).

Το Residual bootstrap είναι μια προσέγγιση που βασίζεται σε μοντέλα. Όπως υποδηλώνει το όνομα, το bootstrapping πραγματοποιείται στα υπολείμματα που λαμβάνονται ως αποτέλεσμα μιας λειτουργίας μοντελοποίησης στα πρωτογενή δεδομένα. Οι νέες χρονοσειρές δείχνουν την ίδια εξάρτηση των αρχικών δεδομένων, που ανιχνεύονται από το

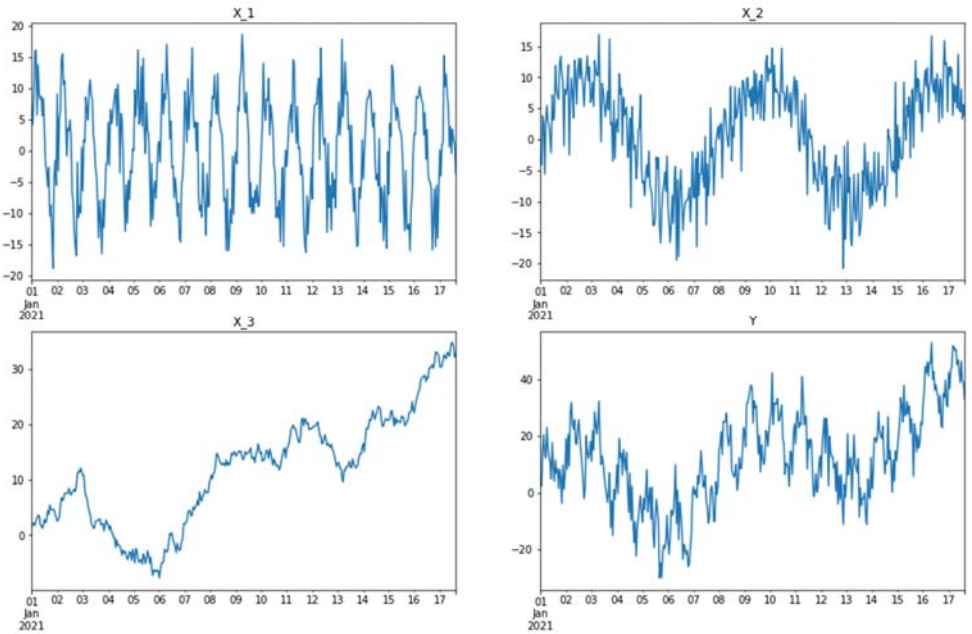
μοντέλο, συν ένα τυχαίο στοιχείο που λαμβάνεται με επαναδειγματοληψία από τα υπολείμματα.[21]

Στο πείραμα μας, συνδυάζουμε τη διαδικασία δειγματοληψίας block bootstrap με την εξάρτηση βάσης μοντέλου του υπολειπόμενου bootstrap. Πρώτον, έτσι οι χρονοσειρές εξομαλύνονται ανάλογα με την επιλεγμένη μέθοδο εξομάλυνσης. Δεύτερον, τα υπολείμματα που προκύπτουν από τη διαδικασία εξομάλυνσης επαναδειγματοληπτούνται με μια επιλεγμένη μέθοδο μπλοκ εκκίνησης. Τέλος, οι εξομαλυνόμενες γραμμές συν τα υπολειμματικά μπλοκ εκκίνησης αθροίζονται για να ληφθεί μια νέα χρονοσειρά. Η νέα σειρά που δημιούργησε αναλόγως αυτή τη λογική μοιράζεται την ίδια χρονική εξάρτηση. Το καλό της μεθοδολογίας εξαρτάται από την επιλεγμένη τεχνική εξομάλυνσης και την ένταση της ομαλοποίησης.

Ο συνδυασμός όλων των bootstrapped σειρών τείνει να σχηματίζει διαστήματα εμπιστοσύνης. Το bootstrapping χρονοσειρών είναι επίσης μια έγκυρη μέθοδος για τη δημιουργία διαστημάτων για την αξιολόγηση των ακραίων τιμών.

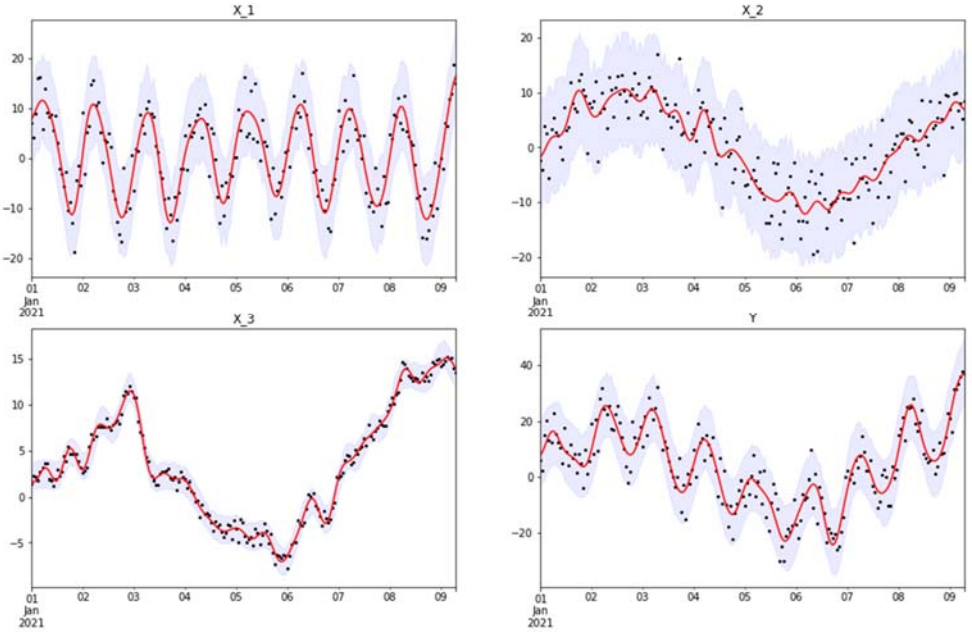
2.5.2. Πρόβλεψη με bootstrap

Για δεδομένα μας χρησιμοποιούμε τα τεχνητά δεδομένα Y τα οποία με βάση ότι έχουμε αναφέρει πιο πάνω στην παράγραφο με την αποσύνθεση, μπορούμε να διασπάσουμε σε συνιστώσες τάσης, εποχικότητας και τυχαιότητας. Εδώ, τα διασπάμε σε 2 συνιστώσες εποχικότητας (X_1 είναι ημερήσια, X_2 είναι ωριαία) και μία τάσης, X_3 η οποία περιέχει και την συνιστώσα της τυχαιότητας.



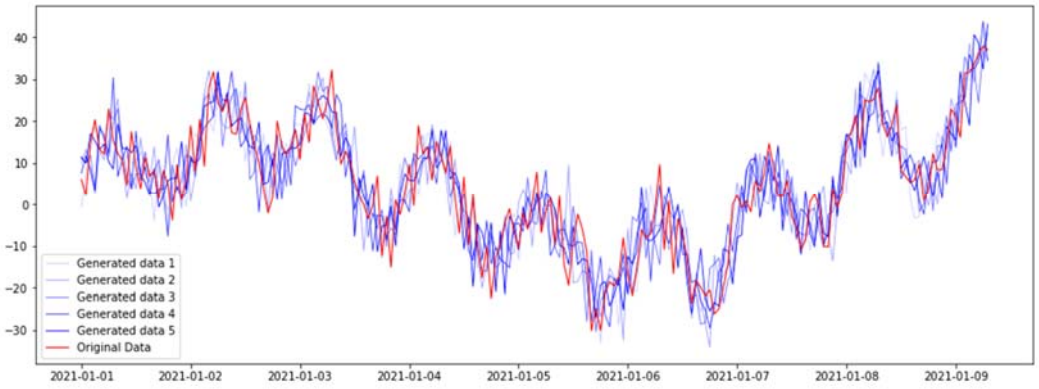
18 Οπτικοποίηση αρχικών δεδομένων και των συνιστωσών τους [2]

Έπειτα παρουσιάζονται οι χρονοσειρές αφού έχουν ομαλοποιηθεί με βάση τον αλγόριθμο του bootstrap και συγκεκριμένα για τις πρώτες εννιά απ’ τις δέκα εφτά ημέρες για λόγους οπτικοποίησης.



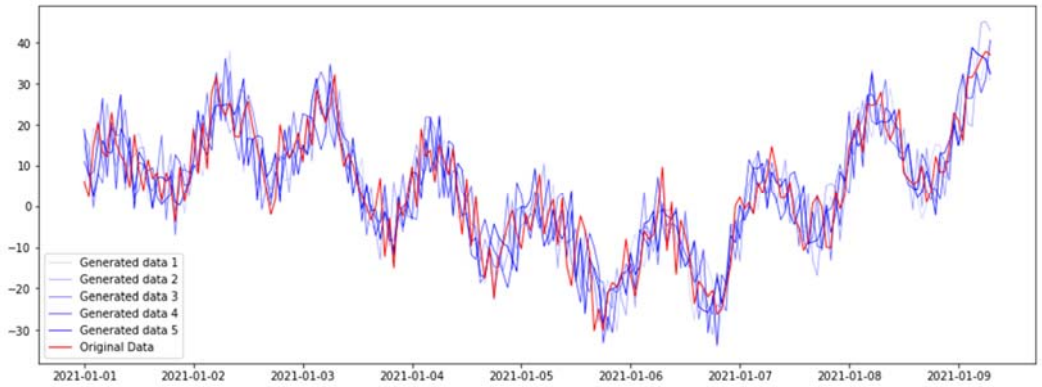
19 Bootstrap στα δεδομένα της εικόνας 18 [2]

Η κόκκινη χρονοσειρά στα παραπάνω διαγράμματα αποτελεί την ομαλοποιημένη καμπύλη μέσω του αλγορίθμου του bootstrap και το μπλε γέμισμα καταδεικνύει το εύρος των δειγμάτων που δημιουργήθηκαν. Στη συνέχεια θα παρουσιαστούν στο ίδιο διάγραμμα, πέντε παραγόμενες χρονοσειρές οι οποίες δημιουργήθηκαν σαν άθροισμα τριών παραγόμενων δειγμάτων κάθε φορά. Τα δύο προέκυψαν απ' τις δύο συνιστώσες εποχικότητας (X_1, X_2) και το τρίτο απ' τη συνιστώσα τάσης (X_3), η οποία περιέχει και την συνιστώσα της τυχαιότητας.



20 Παραγωγή δεδομένων μέσω bootstrap έπειτα από αποσύνθεση [2]

Ενώ τέλος θα παρουσιαστούν στο ίδιο διάγραμμα, πέντε παραγόμενες χρονοσειρές οι οποίες προέκυψαν σαν δείγματα απευθείας απ' τη χρονοσειρά των αρχικών δεδομένων.



21 Παραγωγή δεδομένων μέσω bootstrap απευθείας στα αρχικά δεδομένα [2]

Μπορούμε να παρατηρήσουμε ότι και στις δύο περιπτώσεις η παραγωγή δεδομένων μέσω bootstrap ακολουθεί την τάση και την εποχικότητα των αρχικών δεδομένων παρότι στη δεύτερη περίπτωση δεν έγινε αποσύνθεση σε πρώτο στάδιο. Συμπεραίνουμε ότι η επεξεργασία δεδομένων μέσω bootstrap μπορεί να δώσει καλά αποτελέσματα όταν υπάρχει έλλειψη δεδομένων προς εκπαίδευση του αλγόριθμου μας.

ΚΕΦΑΛΑΙΟ 3

3. Διαχείριση Δεδομένων

3.1. Προετοιμασία Δεδομένων - Διαχείριση κενών τιμών

Γιατί είναι σημαντικό να χειριζόμαστε δεδομένα που λείπουν;

Στα προβλήματα στον πραγματικό κόσμο έχουμε πολλά δεδομένα που λείπουν στις περισσότερες περιπτώσεις. Μπορεί να υπάρχουν διαφορετικοί λόγοι για τους οποίους λείπει κάθε τιμή. Μπορεί να υπάρχει απώλεια ή καταστροφή δεδομένων ή μπορεί να υπάρχουν και συγκεκριμένοι λόγοι. Τα δεδομένα που λείπουν θα μειώσουν την προγνωστική ισχύ του μοντέλου μας. Εάν εφαρμόσουμε αλγόριθμους με δεδομένα που λείπουν, τότε θα υπάρχει προκατάληψη στην εκτίμηση των παραμέτρων. Δεν μπορούμε να είμαστε σίγουροι για τα αποτελέσματά, εάν δεν χειριστούμε τα δεδομένα που λείπουν.[22]

Λόγοι πίσω από τις τιμές που λείπουν

Μερικοί από τους πιθανούς λόγους πίσω από την έλλειψη δεδομένων είναι:

Οι άνθρωποι δεν δίνουν πληροφορίες σχετικά με ορισμένες ερωτήσεις σε μια έρευνα συλλογής δεδομένων. Για παράδειγμα, μερικοί μπορεί να μην αισθάνονται άνετα να μοιράζονται πληροφορίες σχετικά με τον μισθό τους, το ποτό και τις συνήθειες καπνίσματος. Αυτά παραλείπονται σκόπιμα από τον πληθυσμό.

Οι ανακρίβειες κατά τη διαδικασία συλλογής δεδομένων συμβάλλουν επίσης στην έλλειψη δεδομένων. Για παράδειγμα, στη χειροκίνητη εισαγωγή δεδομένων, είναι δύσκολο να αποφευχθούν εντελώς ανθρώπινα λάθη

Ασυνέπειες στον εξοπλισμό που οδηγούν σε λανθασμένες μετρήσεις, οι οποίες με τη σειρά τους δεν μπορούν να χρησιμοποιηθούν, όπως δηλαδή και κάθε κενή τιμή που ενδέχεται να βρούμε στα δικά μας δεδομένα παρακάτω.[22]

Κάποιοι **τρόποι για την αντιμετώπιση του προβλήματος αυτού είναι**:

- Αρχικά γίνεται προσπάθεια εύρεσης της κενής τιμής από άλλες πηγές, αλλιώς γίνεται απευθείας ορισμός αυτής, αν υπάρχει ασφαλής κριτική εκτίμηση για το ύψος στο οποίο κυμάνθηκε.
- Η κενή τιμή ορίζεται ως το ημιάθροισμα (μέσος όρος) της προηγούμενης και της επόμενης παρατήρησης, όταν η χρονοσειρά χαρακτηρίζεται από στασιμότητα και δεν παρατηρείται εποχιακή συμπεριφορά.

- Αν η χρονοσειρά παρουσιάζει σαφή εποχιακή συμπεριφορά, τότε η κενή τιμή ορίζεται ως ο μέσος όρος των τιμών των αντίστοιχων περιόδων. Για παράδειγμα, αν τα δεδομένα αποτελούνται από μηνιαίες παρατηρήσεις και παρατηρηθεί κενή τιμή στο Μάρτιο κάποιου έτους, τότε η κενή αυτή τιμή ορίζεται ως ο μέσος όρος των λοιπών Μαρτίων[22]

3.2. Διαχείριση μηδενικών τιμών

Διακρίνουμε δύο επιλογές:

- Καμία αλλαγή στις μηδενικές τιμές συνήθως όταν είναι λίγες
- Διαχείριση μηδενικών τιμών – όπως τις κενές τιμές όταν είναι πολλές[22]

3.3. Εύρεση Ακραίων Τιμών

Οι ακραίες τιμές είναι παρατηρήσεις που διαφέρουν πολύ από τις περισσότερες παρατηρήσεις της χρονολογικής σειράς. Μπορεί να είναι λάθη ή μπορεί απλώς να είναι ασυνήθιστες τιμές. Σε αυτήν την περίπτωση, μπορεί να θέλουμε να αντικαταστήσουμε τις τιμές που λείπουν με μια εκτίμηση που είναι πιο συνεπής με την πλειονότητα των δεδομένων.

Η απλή αντικατάσταση των ακραίων στοιχείων χωρίς να σκεφτόμαστε γιατί έχουν συμβεί είναι μια επικίνδυνη πρακτική. Γιατί αυτές μπορεί να παρέχουν χρήσιμες πληροφορίες σχετικά με τη διαδικασία που παρήγαγε τα δεδομένα και οι οποίες θα πρέπει να λαμβάνονται υπόψη κατά την πρόβλεψη.

Ωστόσο, εάν είμαστε πρόθυμοι να υποθέσουμε ότι τα ακραία στοιχεία είναι πραγματικά σφάλματα ή ότι δεν θα συμβούν κατά την περίοδο πρόβλεψης, τότε η αντικατάστασή τους μπορεί να διευκολύνει την εργασία πρόβλεψης.

Υπάρχουν πολλοί αλγόριθμοι εντοπισμού ακραίων τιμών. Ωστόσο, εμείς θα αναλύσουμε και αυτόν τον οποίο χρησιμοποιήσαμε στο πειραματικό μέρος. Συγκεκριμένα, η διαδικασία αποσυνθέτει τη χρονοσειρά σε συνιστώσες τάσης, εποχιακής και υπολοίπων :

$$y_t = S_t + T_t + R_t$$

Το εποχιακό στοιχείο είναι προαιρετικό και μπορεί να περιέχει πολλά εποχιακά μοτίβα που αντιστοιχούν στις εποχιακές περιόδους στα δεδομένα. Η ιδέα είναι πρώτα να αφαιρέσουμε οποιαδήποτε εποχικότητα και τάση στα δεδομένα και στη συνέχεια να βρούμε ακραίες τιμές στην υπόλοιπη σειρά, R_t.

Για δεδομένα που παρατηρούνται συχνότερα από ετησίως, χρησιμοποιούμε μια ισχυρή προσέγγιση για την εκτίμηση των T_t και S_t εφαρμόζοντας πρώτα τη μέθοδο MSTL στα δεδομένα. Το MSTL θα εκτιμήσει επαναληπτικά την εποχική συνιστώσα.[23]

Στη συνέχεια, η ισχύς της εποχικότητας μετράται χρησιμοποιώντας τον τύπο:

$$F_s = 1 - \frac{\text{Var}\left(y_t - \hat{T}_t - \hat{S}_t\right)}{\text{Var}\left(y_t - \hat{T}_t\right)}$$

Εάν F_s>0,6, υπολογίζεται μια εποχικά προσαρμοσμένη σειρά:

$$y_t^* = y_t - \hat{S}_t$$

Εδώ χρησιμοποιείται ένα όριο εποχικής αντοχής, επειδή η εκτίμηση του $\hat{S}_t$ είναι πιθανό να είναι υπερβολικά προσαρμοσμένη και πολύ θορυβώδης εάν η υποκείμενη εποχικότητα είναι πολύ ασθενής (ή ανύπαρκτη), καλύπτοντας ενδεχομένως τυχόν ακραίες τιμές με την απορρόφησή τους στην εποχιακή συνιστώσα.

Εάν F_s≤0,6, ή εάν τα δεδομένα παρατηρούνται ετησίως ή λιγότερο συχνά, ορίζουμε απλώς $y_t^* = y_t$.

Στη συνέχεια, επανεκτιμούμε το στοιχείο τάσης από τις τιμές y_t^*. Για μη εποχιακές χρονοσειρές, όπως τα ετήσια δεδομένα, αυτό είναι απαραίτητο καθώς δεν έχουμε την εκτίμηση τάσης από την αποσύνθεση STL. Αλλά ακόμα κι αν έχουμε υπολογίσει μια αποσύνθεση STL, μπορεί να μην την έχουμε χρησιμοποιήσει εάν F_s≤0.6.

Η συνιστώσα τάσης T_t εκτιμάται με την εφαρμογή του ομαλοποιητή Friedman στα δεδομένα y_t^*. Αυτή η λειτουργία έχει δοκιμαστεί σε πολλά δεδομένα και τείνει να λειτουργεί καλά σε ένα ευρύ φάσμα προβλημάτων.

Αναζητούμε ακραίες τιμές στην εκτιμώμενη υπόλοιπη σειρά $\hat{R}_t = y_t^* - \hat{T}_t$

Εάν το Q1 υποδηλώνει το 25ο εκατοστημόριο και το Q3 το 75ο εκατοστημόριο των υπολοίπων τιμών, τότε το εύρος του διατεταρτημορίου ορίζεται ως IQR=Q3−Q1. Οι παρατηρήσεις επισημαίνονται ως ακραίες τιμές εάν είναι μικρότερες από Q1–3×IQR ή μεγαλύτερες από Q3+3×IQR. Αυτός είναι ο ορισμός που χρησιμοποίησε ο Tukey στην αρχική του πρόταση πλαισίου για τις «μακρινές» τιμές.[23]

Εάν οι υπόλοιπες τιμές κατανέμονται κανονικά, τότε η πιθανότητα μια παρατήρηση να αναγνωριστεί ως ακραία τιμή είναι περίπου 1 προς 427000.

3.3.1. Min-Max Κανονοποίηση

Η κανονικοποίηση Min-max είναι ένας από τους πιο συνηθισμένους τρόπους κανονικοποίησης δεδομένων. Για κάθε χαρακτηριστικό, η ελάχιστη τιμή αυτού του χαρακτηριστικού μετατρέπεται σε 0, η μέγιστη τιμή μετατρέπεται σε 1 και κάθε άλλη τιμή μετατρέπεται σε δεκαδικό μεταξύ 0 και 1.

Για παράδειγμα, εάν η ελάχιστη τιμή ενός χαρακτηριστικού ήταν 20 και η μέγιστη τιμή ήταν 40, τότε το 30 θα μετασχηματιζόταν σε περίπου 0,5 αφού βρίσκεται στα μισά του δρόμου μεταξύ 20 και 40. Ο τύπος είναι ο εξής:

X_new = (X - X_min)/(X_max - X_min)

Η κανονικοποίηση ελάχιστης μέγιστης τιμής έχει ένα αρκετά σημαντικό μειονέκτημα: δεν χειρίζεται πολύ καλά τις ακραίες τιμές. Για παράδειγμα, εάν έχουμε 99 τιμές μεταξύ του 0

και του 40 και μια τιμή είναι 100, τότε και όλες οι 99 τιμές θα μετατραπούν σε μια τιμή μεταξύ 0 και 0,4. [24]

3.3.2. Z-Score Normalization

Η κανονικοποίηση της Z-Score είναι μια στρατηγική ομαλοποίησης δεδομένων που αποφεύγει αυτό το ζήτημα με τις ακραίες τιμές. Ο τύπος για την κανονικοποίηση της Z-Score είναι ο παρακάτω:

$$X\_new = (X - \mu)/\sigma$$

Εδώ, μ είναι η μέση τιμή του χαρακτηριστικού και σ είναι η τυπική απόκλιση του χαρακτηριστικού. Εάν μια τιμή είναι ακριβώς ίση με τον μέσο όρο όλων των τιμών του χαρακτηριστικού, θα κανονικοποιηθεί στο 0. Εάν είναι κάτω από το μέσο όρο, θα είναι αρνητικός αριθμός και αν είναι πάνω από τον μέσο όρο θα είναι θετικός αριθμός. Το μέγεθος αυτών των αρνητικών και θετικών αριθμών καθορίζεται από την τυπική απόκλιση του αρχικού χαρακτηριστικού. Εάν τα μη κανονικοποιημένα δεδομένα είχαν μεγάλη τυπική απόκλιση, οι κανονικοποιημένες τιμές θα είναι πιο κοντά στο 0.

Με την κανονικοποίηση Z-Score τα σημεία βρίσκονται τώρα στην ίδια περίπου κλίμακα και για όλα τα χαρακτηριστικά.[24]

3.4. Συσχέτιση δεδομένων

Οι μεταβλητές σε ένα σύνολο δεδομένων μπορούν να σχετίζονται για πολλούς λόγους. Για παράδειγμα:

- Μια μεταβλητή θα μπορούσε να επιφέρει ή να εξαρτάται από τις τιμές μιας άλλης μεταβλητής.
- Μια μεταβλητή θα μπορούσε να συσχετιστεί ελαφρά με μια άλλη μεταβλητή.
- Δύο μεταβλητές θα μπορούσαν να εξαρτώνται από μια τρίτη άγνωστη μεταβλητή.

Μπορεί να είναι χρήσιμη στην ανάλυση δεδομένων και τη μοντελοποίηση, η καλύτερη κατανόηση των σχέσεων μεταξύ των μεταβλητών. Η στατιστική σχέση μεταξύ δύο μεταβλητών αναφέρεται ως συσχέτιση τους.

Μια συσχέτιση θα μπορούσε να είναι θετική, που σημαίνει ότι και οι δύο μεταβλητές κινούνται προς την ίδια κατεύθυνση ή αρνητική, που σημαίνει ότι όταν η τιμή μιας μεταβλητής αυξάνεται, οι τιμές των άλλων μεταβλητών μειώνονται. Η συσχέτιση μπορεί επίσης να είναι ουδέτερη ή μηδενική, που σημαίνει ότι οι μεταβλητές δεν σχετίζονται μεταξύ τους.[25]

- **Θετική συσχέτιση**: και οι δύο μεταβλητές αλλάζουν προς την ίδια κατεύθυνση.
- **Ουδέτερη συσχέτιση**: Καμία σχέση στην αλλαγή των μεταβλητών.
- **Αρνητική συσχέτιση**: οι μεταβλητές αλλάζουν σε αντίθετες κατευθύνσεις.

Η απόδοση ορισμένων αλγορίθμων μπορεί να επιδεινωθεί εάν δύο ή περισσότερες μεταβλητές συνδέονται στενά, πράγμα το οποίο ονομάζεται πολυσυγγραμμικότητα. Ένα

παράδειγμα είναι η γραμμική παλινδρόμηση, όπου μία από τις παραβατικές συσχετισμένες μεταβλητές θα πρέπει να αφαιρεθεί προκειμένου να βελτιωθεί η ικανότητα του μοντέλου.

Μπορεί επίσης να μας ενδιαφέρει η συσχέτιση μεταξύ των μεταβλητών εισόδου με τη μεταβλητή εξόδου, προκειμένου να παρέχουμε μια εικόνα για το ποιες μεταβλητές μπορεί να είναι ή να μην είναι σχετικές ως εισροές για την ανάπτυξη ενός μοντέλου.

Η δομή της σχέσης μπορεί να είναι γνωστή, π.χ. μπορεί να είναι γραμμικό ή μπορεί να μην έχουμε ιδέα αν υπάρχει σχέση μεταξύ δύο μεταβλητών ή ποια δομή μπορεί να έχει. Ανάλογα με το τι είναι γνωστό για τη σχέση και την κατανομή των μεταβλητών, μπορούν να υπολογιστούν διαφορετικές βαθμολογίες συσχέτισης.[25]

3.4.1. Συσχέτιση Pearson

Ο συντελεστής συσχέτισης Pearson, επίσης γνωστός ως, ο συντελεστής συσχέτισης προϊόντος-στιγμής ή στην καθομιλουμένη απλά ως ο συντελεστής συσχέτισης — είναι ένα μέτρο γραμμικής συσχέτισης μεταξύ δύο συνόλων δεδομένων. Είναι ο λόγος μεταξύ της συνδιακύμανσης δύο μεταβλητών και του γινόμενου των τυπικών αποκλίσεων τους. Επομένως είναι ουσιαστικά μια κανονικοποιημένη μέτρηση της συνδιακύμανσης, έτσι ώστε το αποτέλεσμα να έχει πάντα μια τιμή μεταξύ -1 και 1. Όπως και με την ίδια τη συνδιακύμανση, το μέτρο μπορεί να αντικατοπτρίζει μόνο μια γραμμική συσχέτιση μεταβλητών και αγνοεί πολλούς άλλους τύπους σχέσης ή συσχέτισης. [26]

3.4.1.i. Τύπος για έναν πληθυσμό

Ο συντελεστής συσχέτισης Pearson, όταν εφαρμόζεται σε έναν πληθυσμό, αντιπροσωπεύεται συνήθως από το ελληνικό γράμμα ρ και μπορεί να αναφέρεται ως συντελεστής συσχέτισης πληθυσμού ή συντελεστής συσχέτισης πληθυσμού Pearson. Δεδομένου ενός ζεύγους τυχαίων μεταβλητών (X,Y), ο τύπος για το ρ είναι:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

όπου: {cov} είναι η συνδιακύμανση

σ{X} είναι η τυπική απόκλιση του X

σ{Y} είναι η τυπική απόκλιση του Y

Ο τύπος για το *ρ* μπορεί να εκφραστεί με όρους μέσου όρου και προσδοκιών.

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)],$$

ο τύπος για *ρ* μπορεί επίσης να γραφτεί ως

$$\rho_{X,Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

3.4.1.ii. Συσχέτιση Spearman

Ο συντελεστής συσχέτισης του Spearman ή το ρ του Spearman, που ονομάστηκε από τον Charles Spearman και συχνά υποδηλώνεται και αυτό με το ελληνικό γράμμα ρ ή ως r_s είναι μη παραμετρικό μέτρο συσχέτισης κατάταξης (στατιστική εξάρτηση μεταξύ των

ταξινομήσεων δύο μεταβλητών). Αξιολογεί πόσο καλά μπορεί να περιγραφεί η σχέση μεταξύ δύο μεταβλητών χρησιμοποιώντας μια μονοτονική συνάρτηση.

Η συσχέτιση Spearman μεταξύ δύο μεταβλητών είναι ίση με τη συσχέτιση Pearson μεταξύ των τιμών κατάταξης αυτών των δύο μεταβλητών. ενώ η συσχέτιση του Pearson αξιολογεί τις γραμμικές σχέσεις, η συσχέτιση του Spearman αξιολογεί τις μονοτονικές σχέσεις (είτε γραμμικές είτε όχι). Εάν δεν υπάρχουν επαναλαμβανόμενες τιμές δεδομένων, εμφανίζεται μια τέλεια συσχέτιση Spearman με τιμή +1 ή −1 όταν κάθε μία από τις μεταβλητές είναι μια τέλεια μονότονη συνάρτηση της άλλης.

Διαισθητικά, η συσχέτιση Spearman μεταξύ δύο μεταβλητών θα είναι υψηλή όταν οι παρατηρήσεις έχουν παρόμοια (ή πανομοιότυπη για συσχέτιση 1) κατάταξη (δηλαδή ετικέτα σχετικής θέσης των παρατηρήσεων εντός της μεταβλητής: 1η, 2η, 3η, κ.λπ.) μεταξύ των δύο μεταβλητές και χαμηλή όταν οι παρατηρήσεις έχουν ανόμοια (ή εντελώς αντίθετο για συσχέτιση −1) κατάταξη μεταξύ των δύο μεταβλητών.

Ο συντελεστής Spearman είναι κατάλληλος τόσο για συνεχείς όσο και για διακριτές τακτικές μεταβλητές.

Ορισμός και υπολογισμός

Ο συντελεστής συσχέτισης Spearman ορίζεται ως ο συντελεστής συσχέτισης Pearson μεταξύ των μεταβλητών κατάταξης.[26]

Για ένα δείγμα μεγέθους n, οι n πρωτογενείς βαθμολογίες X_i,Y_i μετατρέπονται σε μεταβλητές κατάταξης $R(X_i)$, $R(Y_i)$, και το r_s υπολογίζεται όπως παρακάτω:

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\operatorname{cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)}}$$

Όπου:

- ρ υποδηλώνει τον συνήθη συντελεστή συσχέτισης Pearson, αλλά εφαρμόζεται στις μεταβλητές κατάταξης
- $\operatorname{cov}(\mathrm{R}(X), \mathrm{R}(Y))$ είναι η συνδιακύμανση των μεταβλητών κατάταξης,
- $\sigma_{\mathrm{R}(X)}$ και $\sigma_{\mathrm{R}(Y)}$ είναι οι τυπικές αποκλίσεις των μεταβλητών κατάταξης.

3.5. Συνδιακύμανση

Η συνδιακύμανση είναι ένα μέτρο της σχέσης μεταξύ δύο τυχαίων μεταβλητών. Η μέτρηση αξιολογεί πόσο – σε ποιο βαθμό – αλλάζουν οι μεταβλητές μαζί. Με άλλα λόγια, είναι ουσιαστικά ένα μέτρο της διακύμανσης μεταξύ δύο μεταβλητών. Ωστόσο, η μέτρηση δεν αξιολογεί την εξάρτηση μεταξύ των μεταβλητών.

Σε αντίθεση με τον συντελεστή συσχέτισης, η συνδιακύμανση μετράται σε μονάδες. Οι μονάδες υπολογίζονται πολλαπλασιάζοντας τις μονάδες των δύο μεταβλητών. Η διακύμανση μπορεί να λάβει οποιεσδήποτε θετικές ή αρνητικές τιμές. Οι τιμές ερμηνεύονται ως εξής [27]:

- Θετική συνδιακύμανση: Υποδεικνύει ότι δύο μεταβλητές τείνουν να κινούνται προς την ίδια κατεύθυνση.
- Αρνητική συνδιακύμανση: Αποκαλύπτει ότι δύο μεταβλητές τείνουν να κινούνται σε αντίστροφες κατευθύνσεις

3.5.1. Υπολογισμός Συνδιακύμανσης

Ο τύπος συνδιακύμανσης είναι παρόμοιος με τον τύπο συσχέτισης και ασχολείται με τον υπολογισμό των σημείων από τη μέση τιμή σε ένα σύνολο δεδομένων. Για παράδειγμα, η συνδιακύμανση μεταξύ δύο τυχαίων μεταβλητών X και Y μπορεί να υπολογιστεί χρησιμοποιώντας τον ακόλουθο τύπο (για πληθυσμό) [27]:

$$\mathrm{Cov}\,(\mathrm{X},\mathrm{Y}) = \frac{\sum\ (\mathrm{X_i} - \overline{\mathrm{X}})\left(\mathrm{Y_j} - \overline{\mathrm{Y}}\right)}{\mathrm{n}}$$

Για μια συνδιακύμανση δείγματος, ο τύπος προσαρμόζεται ελαφρώς:

$$\mathrm{Cov}\,(\mathrm{X},\mathrm{Y}) = \frac{\sum\ (\mathrm{X_i} - \overline{\mathrm{X}})\left(\mathrm{Y_j} - \overline{\mathrm{Y}}\right)}{\mathrm{n} - 1}$$

Όπου:

- X_i – οι τιμές της μεταβλητής X
- Y_j – οι τιμές της μεταβλητής Y
- $\bar{X}$ – ο μέσος όρος (μέσος όρος) της μεταβλητής X
- $\bar{Y}$ – ο μέσος όρος (μέσος όρος) της μεταβλητής Y
- n – ο αριθμός των σημείων δεδομένων

3.6. Συνδιακύμανση έναντι συσχέτισης

Η συνδιακύμανση και η συσχέτιση αξιολογούν κυρίως τη σχέση μεταξύ των μεταβλητών. Η πιο κοντινή αναλογία στη σχέση μεταξύ τους είναι η σχέση μεταξύ της διακύμανσης και της τυπικής απόκλισης.[27]

Η συνδιακύμανση μετρά τη συνολική διακύμανση δύο τυχαίων μεταβλητών από τις αναμενόμενες τιμές τους. Χρησιμοποιώντας τη συνδιακύμανση, μπορούμε μόνο να μετρήσουμε την κατεύθυνση της σχέσης (αν οι μεταβλητές τείνουν να κινούνται παράλληλα ή να δείχνουν μια αντίστροφη σχέση). Ωστόσο, δεν υποδηλώνει τη δύναμη της σχέσης, ούτε την εξάρτηση μεταξύ των μεταβλητών.

Από την άλλη πλευρά, η συσχέτιση μετρά την ισχύ της σχέσης μεταξύ των μεταβλητών. Η συσχέτιση είναι το κλιμακωτό μέτρο της συνδιακύμανσης. Είναι αδιάστατο. Με άλλα λόγια, ο συντελεστής συσχέτισης είναι πάντα μια καθαρή τιμή και δεν μετριέται σε καμία μονάδα.[27]

Η σχέση μεταξύ των δύο εννοιών μπορεί να εκφραστεί χρησιμοποιώντας τον παρακάτω τύπο:

$$\rho(X,Y) = \frac{\mathrm{Cov}\,(X,Y)}{\sigma_X \sigma_Y}$$

ΚΕΦΑΛΑΙΟ 4

4. Πειραματικό Μέρος

4.1. Το πρόβλημα

Αν και έγινε μια πρώτη επαφή τόσο στην «Περίληψη», όσο και στο «Κεφάλαιο 1» της διπλωματικής, αξίζει να αναλύσουμε λίγο καλύτερα το πρόβλημα μας και σε αυτήν την ενότητα. Κατ' ουσίαν, έχουμε δεδομένα για ένα ολόκληρο έτος τόσο ενεργού όσο και άεργου ισχύος, για τρεις ζυγούς στο σύνολο. Το ζητούμενο μας είναι να επεξεργαστούμε κατάλληλα, όπως θα καταδείξουμε παρακάτω, τα δεδομένα, ώστε να παράγουμε συνθετικά δεδομένα ενός ολόκληρου χρόνου για κάθε έναν από τους ζυγούς. Ύστερα, να τα ελέγξουμε και να δούμε, εάν αυτά θα μπορούσαν να είναι όμοια με τα αληθινά δεδομένα τα οποία μας δόθηκαν από τον μετρητή.

4.2. Τα δεδομένα

Όπως έχουμε ήδη αναφέρει, έχουμε στη διάθεση μας δεδομένα ενός ολόκληρου χρόνου, ενεργού και άεργου ισχύος για τρεις ζυγούς. Τα δεδομένα αυτά έχουν παρθεί μέσω δεκαπεντάλεπτης δειγματοληψίας. Ουσιαστικά, έχουμε 96 λήψεις καθημερινώς και 35.040 για κάθε ζυγό συνολικά. Τα δεδομένα μας δόθηκαν σε μορφή ενός excel και για την οπτικοποίηση-επεξεργασία τους χρησιμοποιήσαμε την γλώσσα προγραμματισμού python. Παρακάτω, παρουσιάζονται τα δεδομένα συνοπτικά, σε μορφή πίνακα:

Date	bus1act	bus1react	bus2act	bus2react	bus3act	bus3react
2018-01-01 00:00:00	41.76	2.82	61.44	6.06	20.16	3.66
2018-01-01 00:15:00	39.18	2.70	62.58	7.08	17.88	3.96
2018-01-01 00:30:00	38.10	2.58	64.44	6.42	18.42	3.96
2018-01-01 00:45:00	36.00	2.16	59.70	5.64	18.48	3.96
2018-01-01 01:00:00	34.56	1.92	53.76	4.80	15.42	3.12
...	...	...	...	...	...	...
2018-12-31 22:45:00	49.38	4.50	65.94	5.52	18.12	2.58
2018-12-31 23:00:00	51.78	4.86	64.14	6.00	18.06	2.76
2018-12-31 23:15:00	52.62	5.52	62.64	5.40	16.26	2.82
2018-12-31 23:30:00	52.62	4.92	56.64	4.74	15.18	2.04
2018-12-31 23:45:00	50.22	4.44	53.88	4.74	15.78	2.58

35040 rows × 6 columns

22 Συνοπτική παρουσίαση δεδομένων σε μορφή Dataframe

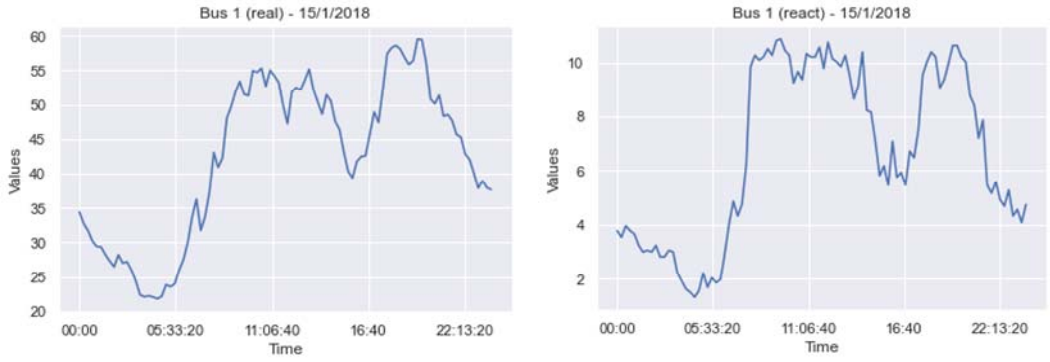
Ας δούμε ακόμη, κάποιες στατιστικές ιδιότητες για κάθε ζυγό:

	bus1act	bus1react	bus2act	bus2react	bus3act	bus3react
mean	27.68	6.36	33.19	5.96	25.14	5.41
std	12.01	3.28	18.72	1.89	13.27	2.76
min	8.76	0.00	10.08	1.86	8.04	0.00
25%	18.66	3.84	19.68	4.56	14.22	3.30
50%	25.02	5.76	27.36	5.76	21.96	4.92
75%	34.02	8.34	40.16	7.08	32.46	7.02
max	77.88	21.12	116.22	15.36	76.08	17.04

23 Βασικές στατιστικές ιδιότητες δεδομένων όλου του χρόνου

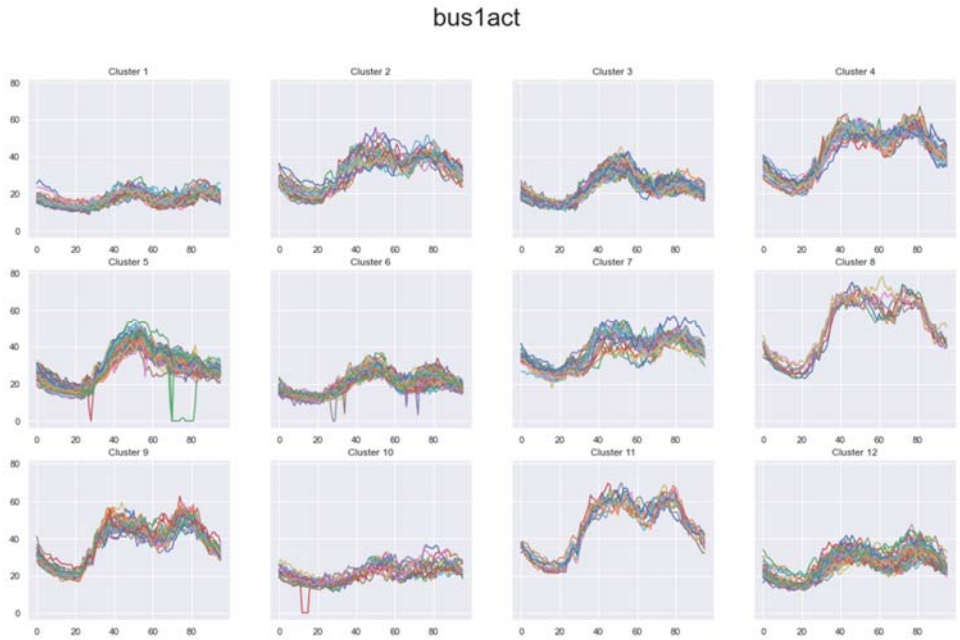
Απ' τους προηγούμενους δύο πίνακες, μπορούμε να καταλάβουμε λίγα πράγματα. Γίνεται εύκολα αντιληπτό ότι τα δεδομένα χωρίς κάποια οπτικοποίηση είναι δύσκολο να κατανοηθούν, να ερμηνευτούν και να εξάγουμε συμπεράσματα απ' αυτά. Ξεκινάμε οπτικοποιώντας, τα δεδομένα, του ζυγού Νο 1:

Για να γίνει περίπου, αντιληπτό, για τι δεδομένα μιλάμε παρακάτω θα απεικονίσουμε δεδομένα μιας τυχαίας ημέρας:

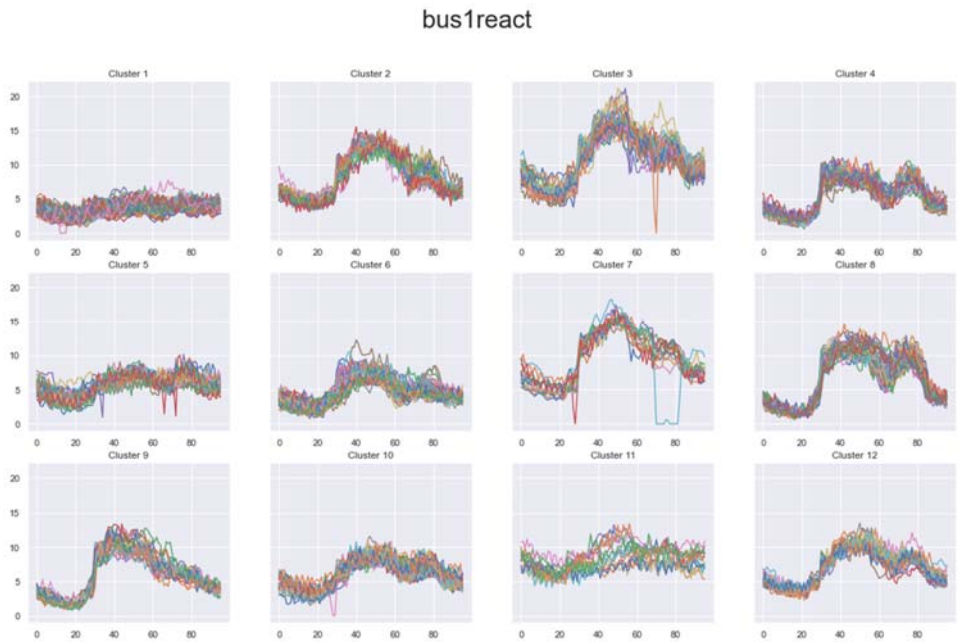


24 Οπτικοποίηση δεδομένων μίας ημέρας ενεργού και άεργου ισχύος για το ζυγό Νο 1

Και στη συνέχεια ολόκληρου του χρόνου για έναν ζυγό, τόσο ενεργού, όσο και άεργου ισχύος:



25 Οπτικοποίηση δεδομένων ενεργού ισχύος χωρισμένων σε clusters για όλο το χρόνο (Ζυγός 1)



26 Οπτικοποίηση δεδομένων άεργου ισχύος χωρισμένων σε clusters για όλο το χρόνο (Ζυγός 1)

Αυτό που βλέπουμε σε κάθε μια απ' τις δύο παραπάνω φωτογραφίες είναι, δώδεκα ομάδες/clusters χρονοσειρών. Κάθε, χρονοσειρά παριστάνει δεδομένα μίας ημέρας και έχει, όπως βλέπουμε διαφορετικό χρώμα σε σχέση με τις υπόλοιπες. Κάθε χρονοσειρά που

ανήκει στην ίδια ομάδα, έχει ουσιαστικά παρόμοιες στατιστικές ιδιότητες με κάθε άλλη που ανήκει στην ομάδα αυτή. Ο διαχωρισμός σε ομάδες έγινε με την βοήθεια ενός αλγορίθμου που βασίζεται στη θεωρία των γκαουσιανών μειγμάτων (βλ. ΚΕΦ 2, παρ 2). Κάθε cluster χαρακτηρίζεται από μια πιθανοφανή κανονική κατανομή, με μέση τιμή και διακύμανση που προσαρμόζονται αυτόματα κατά τη διαδικασία του clustering. Ο αριθμός των ομάδων/clusters δεν καθορίστηκε με τη βοήθεια κάποιου στατιστικού κριτηρίου, όπως αυτά που έχουμε αναφέρει στη θεωρία μας, αλλά έγινε κατά τυχαίο τρόπο, αφού αφορά μια πρώτη οπτικοποίηση των δεδομένων μας.

4.2.1. Απαλοιφή ακραίων τιμών

Κοιτώντας τις γραφικές παραστάσεις, παρατηρούνται κάποιες ασυνήθιστες τιμές σε σχέση με τις αντίστοιχες τιμές, των άλλων χρονοσειρών της ίδιας ομάδας (αναφερόμαστε, προφανώς, σε τιμές οι οποίες αφορούν την ίδια ώρα μέσα σε μια μέρα), είτε ακόμα παρατηρούνται και κάποιες μηδενικές τιμές. Αναφερόμενοι σε διάστημα, ενός χρόνου και 35.040 δεδομένων για κάθε ζυγό, είναι λογικό να συμβαίνουν και κάποια λάθη στις μετρήσεις μας. Αυτό μπορεί να δικαιολογηθεί είτε από κάποιο σφάλμα λειτουργίας στο ζυγό, είτε από σφάλμα στη μέτρηση. Σε κάθε περίπτωση, τα δεδομένα που θωρούμε ως ασυνήθιστα, θέλουμε να τα αντικαταστήσουμε με μια εκτίμηση που είναι πιο συνεπής με την πλειονότητα των δεδομένων. Για να βρούμε τα δεδομένα αυτά, ανατρέξαμε στα δεδομένα και φιλτράραμε όσα ικανοποιούν ανά cluster και ανά ώρα την παρακάτω συνθήκη:

$$\frac{X_t - \mu_\tau}{\sigma_\tau} > t, \qquad \tau o\, t\, (threshold)\, \tau o\, \theta \acute{\epsilon} \sigma a \mu \epsilon\, \acute{\iota} \sigma o\, \mu \epsilon\, 3$$

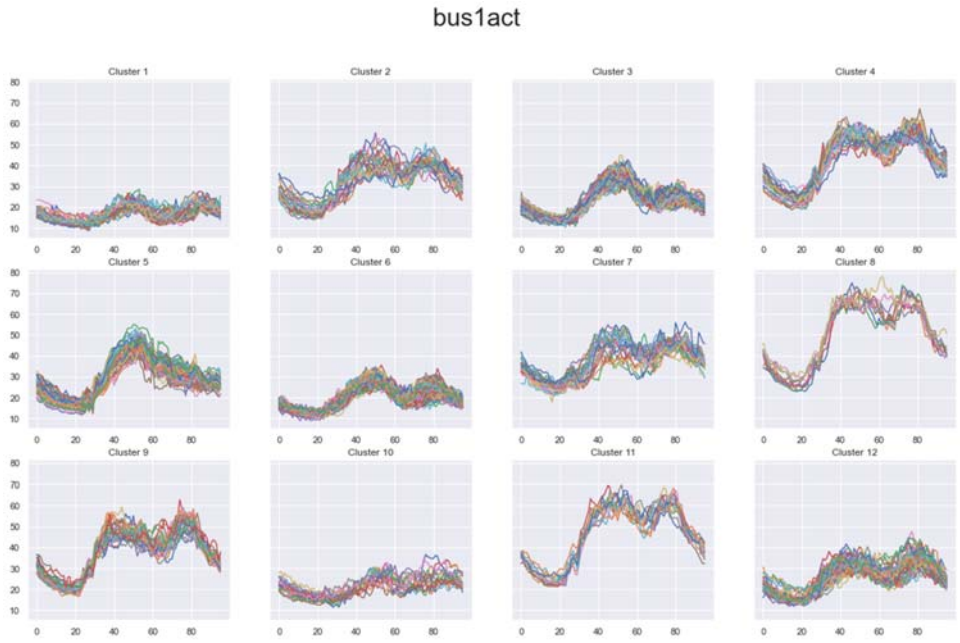
Όπου:

$\mathbf{X_t}$ είναι η παρατήρηση της τιμής που εξετάζεται.

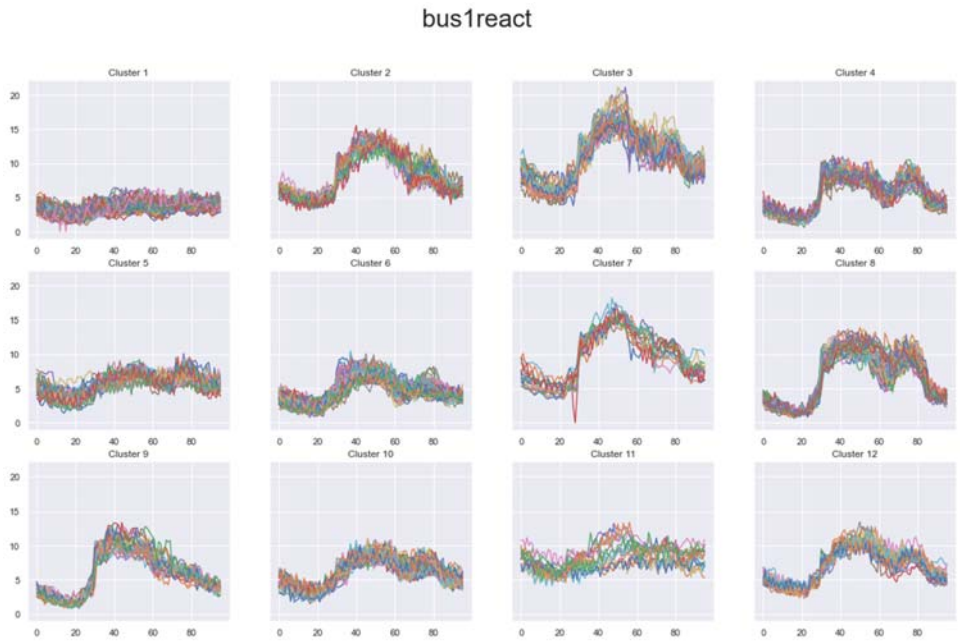
$\boldsymbol{\mu_\tau}$ είναι η μέση τιμή του συνόλου δεδομένων του συγκεκριμένου cluster εκείνη τη χρονική στιγμή.

$\boldsymbol{\sigma_\tau}$ είναι η τυπική απόκλιση του συνόλου δεδομένων του συγκεκριμένου cluster εκείνη τη χρονική στιγμή.

Αφού τα βρήκαμε, αντικαταστήσαμε καθένα απ' αυτά με το μ_τ, στο οποίο αναφερθήκαμε παραπάνω. Η οπτικοποίηση των δεδομένων απαλλαγμένων από ακραίες τιμές, διατηρώντας τις χρονοσειρές στα clusters τα οποία ήταν και πριν, μόνο για τον πρώτο ζυγό, φαίνεται παρακάτω:



27 Οπτικοποίηση δεδομένων ενεργού ισχύος εικ. 25 απαλλαγμένων από ακραίες τιμές



28 Οπτικοποίηση δεδομένων άεργου ισχύος εικ. 26 απαλλαγμένων από ακραίες τιμές

Παρατηρούμε ότι η εμφάνιση ακραίων τιμών έχει ελαχιστοποιηθεί. Εάν, θέλουμε και καλύτερη εξομάλυνση θα μπορούσαμε να μειώσουμε το threshold, αλλά κρίναμε ότι κατά αυτό τον τρόπο πολλά δεδομένα θα είχαν αλλαχθεί, γεγονός το οποίο θα επηρέαζε αρνητικά τις στατιστικές ιδιότητες των δεδομένων μας (πολλές αλλαγές σε σχέση με τα

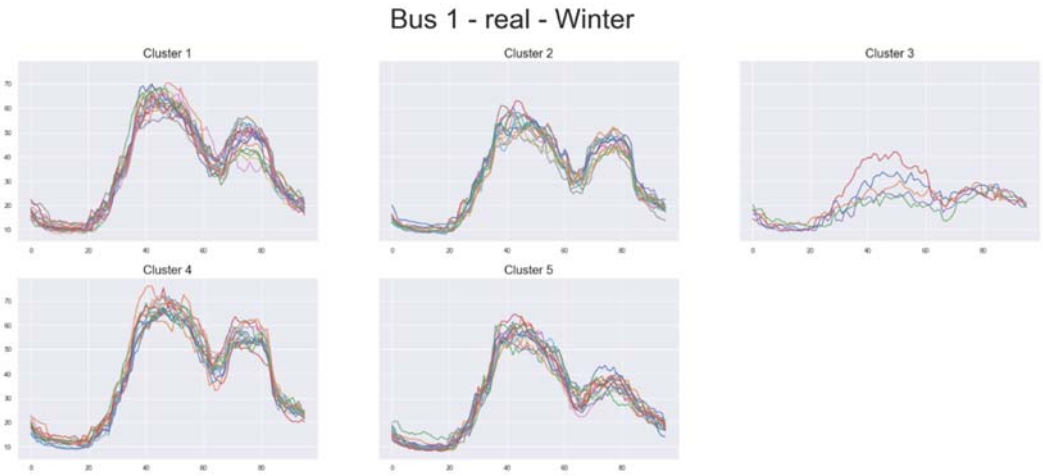
αρχικά).Μπορούμε έτσι να προχωρήσουμε με την εξομάλυνση των δεδομένων όλων των ζυγών μας. Αφού το κάνουμε, με τα δεδομένα μας, απαλλαγμένα από ακραίες τιμές, προχωράμε στη γέννηση καινούριων δεδομένων.

Αποφασίσαμε να χρησιμοποιήσουμε τρεις διαφορετικούς τρόπους γέννησης δεδομένων και ύστερα να συγκρίνουμε τα αποτελέσματα μας. Ξεκινάμε με την παραγωγή δεδομένων, μόνο του πρώτου ζυγού. Αφού, γνωρίσουμε τους τρεις τρόπους και συγκρίνουμε τα αποτελέσματα μας, θα δούμε στη συνέχεια ποιον απ' τους τρεις τρόπους θα χρησιμοποιήσουμε για να γεννήσουμε δεδομένα για όλους τους ζυγούς. Βασική προϋπόθεση είναι να διατηρούνται οι ιδιότητες τους συστήματος μας, καθώς οι τρεις ζυγοί είναι γειτονικοί και οι τιμές τους ενός επηρεάζουν τις τιμές του άλλου.

4.3. Πρώτος τρόπος επαύξησης δεδομένων

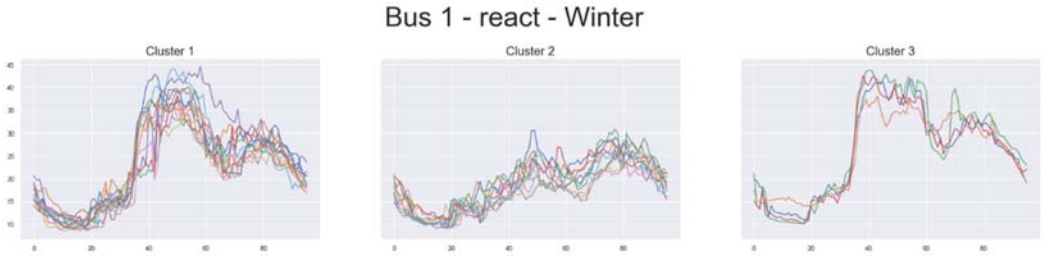
Ξεκινάμε και διαχειριζόμαστε τα δεδομένα ανά εποχή. Για κάθε εποχή, χωρίζουμε σε δεδομένα καθημερινών και δεδομένα από σαββατοκύριακα. Για κάθε μία υποκατηγορία, τρέχουμε τον αλγόριθμο που έχουμε φτιάξει και χωρίζουμε σε clusters. Η επιλογή του αριθμού των clusters δεν βασίζεται σε στατιστικά κριτήρια όπως τα BIC ή AIC, αλλά καθορίζεται πάλι βάσει της εμπειρίας και της παρατήρησης των δεδομένων μας. Συγκεκριμένα, για τις καθημερινές χρησιμοποιούμε πέντε ομάδες και για Σαββατοκύριακα χρησιμοποιούμε τρεις ομάδες.

Για να κατανοήσουμε καλύτερα αυτό που μόλις περιγράψαμε, ας δούμε παρακάτω τι εννοούμε. Πρώτα θα αναπαραστήσουμε τα ομαδοποιημένα, βάση των στατιστικών τους ιδιοτήτων, δεδομένα καθημερινών του χειμώνα:



29 Δεδομένα εργάσιμων ημερών ενεργού ισχύος χειμώνα για το ζυγό 1

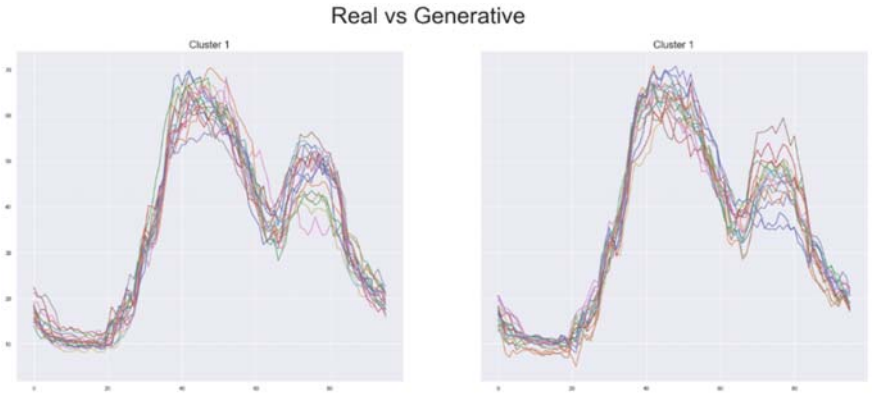
Στη συνέχεια θα παρουσιαστούν τα δεδομένα για τα Σαββατοκύριακα:



30 Δεδομένα μη εργάσιμων ημερών ενεργού ισχύος χειμώνα για το ζυγό 1

Μια πρώτη παρατήρηση κοιτάζοντας τις παραπάνω εικόνες είναι ότι οι χρονοσειρές εμφανίζουν διαφορά ως προς τα ύψη τους, γεγονός που επιβεβαιώνει την ορθότητα του να διαχειριστούμε ξεχωριστά εργάσιμες με μη εργάσιμες ημέρες. Έχοντας κατατάξει ουσιαστικά κάθε ημέρα σε ένα cluster, αυτό που κάνουμε για να παράγουμε καινούρια δεδομένα είναι να τρέξουμε σειριακά τον αλγόριθμο μας, ο οποίος σε πρώτη φάση θα αναγνωρίζει σε ποιο cluster ανήκει η κάθε μέρα της εποχής. Στη συνέχεια θα παράγει μέσω μιας συνάρτησης της βιβλιοθήκης scikit-learn, generated δεδομένα τα οποία θα ακολουθούν την κανονική κατανομή που αντιστοιχεί στο συγκεκριμένο cluster. Θα κάνουμε δηλαδή δειγματοληψία από την κανονική κατανομή αυτή. Για να είμαστε πιο συγκεκριμένοι, κάθε cluster στην GMM χαρακτηρίζεται από μια κανονική κατανομή, η οποία περιγράφεται από δύο βασικές παραμέτρους: τη μέση τιμή και τη διακύμανση. Βασικό γνώρισμα των generated μας δεδομένων, συνεπώς, είναι ότι θα έχουν ίδια μέση τιμή και διακύμανση με το cluster απ' το οποίο προέκυψαν

Ας δούμε μια οπτικοποίηση των αρχικών μας δεδομένων σε σύγκριση με λίγα generated για ένα συγκεκριμένο cluster για να καταδειχθεί η σχέση αρχικών με generated δεδομένων:



31 Σύγκριση αρχικών δεδομένων με τα generated για ένα cluster

Μπορούμε να παρατηρήσουμε την ομοιότητα που έχουν, αλλά παράλληλα έχουμε διατηρήσει και την τυχαιότητα που θέλαμε να έχουμε. Για να ολοκληρώσουμε την παραγωγή δεδομένων για ένα χρόνο, αρκεί να κάνουμε την ίδια διαδικασία και για τις 4 εποχές και να τα συνενώσουμε.

4.4. Δεύτερος τρόπος επαύξησης δεδομένων

Ο δεύτερος τρόπος έχει τη λογική του πρώτου μέχρι ένα σημείο. Πιο συγκεκριμένα, μεταχειριζόμαστε και εδώ τα δεδομένα μας ανά εποχή. Χωρίζουμε σε δεδομένα καθημερινών και δεδομένα από Σαββατοκύριακα. Για κάθε μία υποκατηγορία, τρέχουμε τον αλγόριθμο που έχουμε φτιάξει και χωρίζουμε σε clusters. Πάλι αυθαίρετα κάνουμε επιλογή για τον αριθμό των clusters. Συγκεκριμένα, για τις καθημερινές χρησιμοποιούμε πέντε ομάδες και για Σαββατοκύριακα χρησιμοποιούμε τρεις.

Ως εδώ η διαδικασία είναι η ίδια με τον πρώτο τρόπο. Τώρα, όμως αυτό που κάνουμε είναι να τρέξουμε σειριακά έναν αλγόριθμο, ο οποίος θα μας φτιάξει μια σειρά αριθμών, κάθε στοιχείο της οποίας θα υποδηλώνει σε ποια απ' τις πέντε ομάδες ανήκει κάθε καθημερινή. Το ίδιο προφανώς θα κάνουμε και για τα Σαββατοκύριακα. Για μεγαλύτερο εύρος επιλογής, αυξήσαμε τις ομάδες σε τρεις σε σχέση με τον προηγούμενο τρόπο. Ο λόγος που το κάναμε αυτό, είναι για να βρούμε για κάθε μία εκ των δύο σειρών, τον πίνακα μετάβασης σύμφωνα με τη θεωρία των Αλυσίδων Markov. Ουσιαστικά δεν κάνουμε τίποτα άλλο, πέρα απ' το να βρούμε, με δεδομένο ότι μια μέρα μας ανήκει σε ένα cluster, την πιθανότητα η επόμενη μέρα να ανήκει σε ένα άλλο cluster. Αυτό το κάνουμε θεωρώντας σαν αρχικές και σαν επόμενες καταστάσεις, όλα τα clusters μας. Για να καταλάβουμε οπτικά τι μπορεί να σημαίνει αυτό ας δούμε πως είναι ο πίνακας μετάβασης για καθημερινές (5 ομάδες δεδομένων) και Σαββατοκύριακα (3 ομάδες δεδομένων):

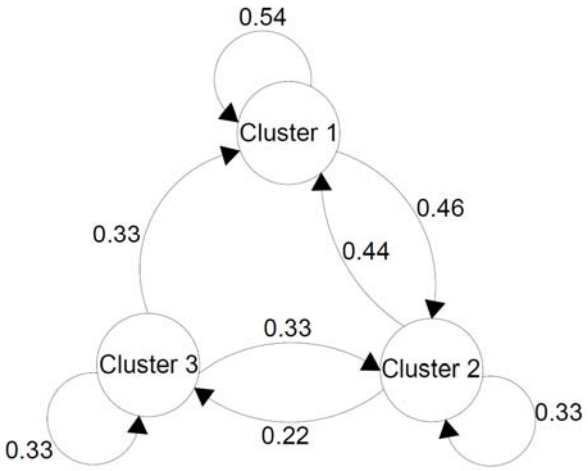
Next state / State	1	2	3	4	5
1	0.08	0.08	0.25	0.25	0.33
2	0.20	0.40	0.20	0.10	0.10
3	0.06	0.19	0.56	0.00	0.19
4	0.46	0.08	0.00	0.46	0.00
5	0.17	0.17	0.17	0.25	0.25

33 Πίνακας μετάβασης για εργάσιμες ημέρες

Next state / State	1	2	3
1	0.54	0.46	0.00
2	0.44	0.33	0.22
3	0.33	0.33	0.33

32 Πίνακας μετάβασης για μη εργάσιμες ημέρες

Και σχηματικά ας δούμε πως μοιάζει ο δεύτερος πίνακας σαν αλυσίδα Markov:



34 Αλυσίδες Markov για μη εργάσιμες ημέρες

Το επόμενο μας βήμα, εφόσον έχουμε τους πίνακες μετάβασης, είναι να εξομοιώσουμε κάθε εποχή σύμφωνα με τους πίνακες αυτούς. Αυτό, που θα κάνουμε αρχικά, είναι να θεωρήσουμε σαν πρώτη μέρα (αρχική κατάσταση), την αντίστοιχη πρώτη των αρχικών μας δεδομένων (για κάθε εποχή και για καθημερινές/σαββατοκύριακα ξεχωριστά αντίστοιχα). Στη συνέχεια, επιλέγεται η επόμενη κατάσταση μέσω ενός τυχαίου και αμερόληπτου ζαριού. Ο κώδικας λαμβάνει υπόψη τις πιθανότητες μετάβασης από την τρέχουσα κατάσταση προς όλες τις δυνατές επόμενες καταστάσεις και επιλέγει μια κατάσταση βάσει αυτών των πιθανοτήτων. Η επιλεγμένη κατάσταση γίνεται η νέα τρέχουσα κατάσταση και η διαδικασία επαναλαμβάνεται.

Παράλληλα με τη δημιουργία των παραπάνω λιστών, έχουμε ήδη παράξει δείγματα χρονοσειρών για κάθε ένα απ' τα cluster μας, τηρώντας τον κανόνα, τα δείγματα μας να ακολουθούν την κανονική κατανομή που αντιστοιχεί στο συγκεκριμένο cluster.

Έχουμε παράξει, για μια ολόκληρη εποχή, με τη βοήθεια των αλυσίδων Markov, μια ακολουθία από clusters και έχουμε για κάθε cluster, ένα μεγάλο αριθμό generated χρονοσειρών. Κάθε, στοιχείο της ακολουθίας, αντιπροσωπεύει μία ημέρα της εποχής. Για να φτιάξουμε μια generated ουσιαστικά εποχή, αυτό που πρέπει να κάνουμε είναι να επιλέξουμε τυχαία, με τη βοήθεια ενός τίμιου ζαριού μια χρονοσειρά για κάθε στοιχείο της λίστας από clusters. Κατά την διαδικασία, έχουμε προσέξει να μην υπάρχουν διπλότυπα δεδομένα, δηλαδή κάθε χρονοσειρά να μπορεί να επιλεχθεί μία φορά. Επιλέγοντας για κάθε μέρα, μια χρονοσειρά, φτιάχνουμε μια generated εποχή. Συνεχίζουμε την ίδια διαδικασία και για τις επόμενες εποχές και συνενώνουμε τα δεδομένα μας.

Για να καταλάβουμε καλύτερα τη διαδικασία, παρακάτω μπορούμε να δούμε τις 2 ακολουθίες που παρήχθησαν για το χειμώνα του ζυγού «1» με βάση τους δύο πίνακες κατάστασης που παρουσιάστηκαν παραπάνω.

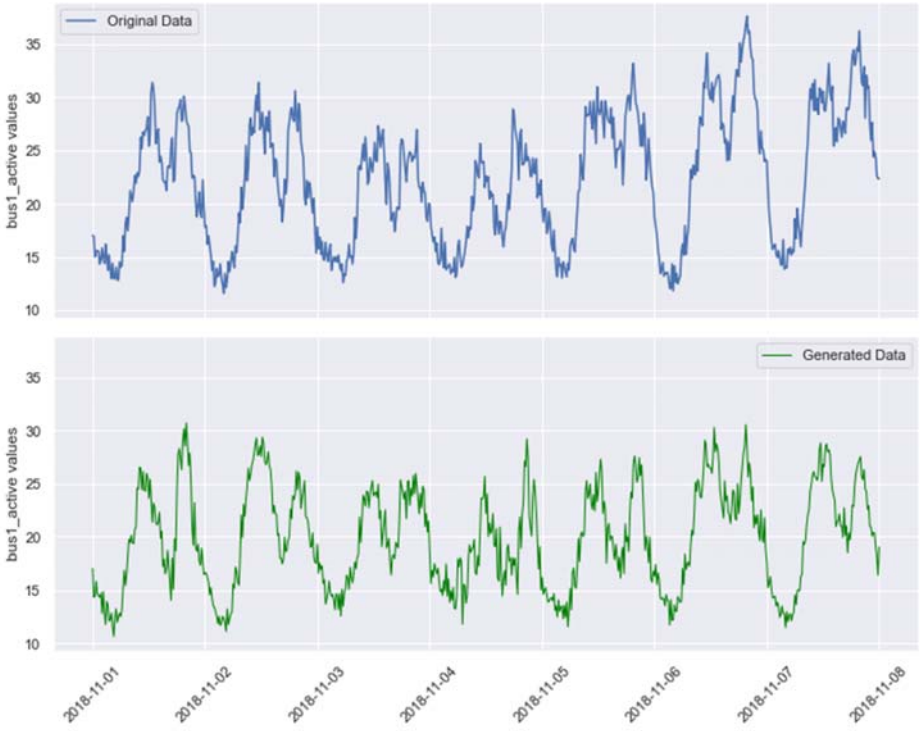
Για τις καθημερινές χωρισμένες σε 5 clusters, διαστήματος 2 εβδομάδων:

```
[2, 2, 1, 4, 2, 4, 4, 4, 4, 4, 2, 1, 5, 2, 3]
```

Για μη καθημερινές χωρισμένες σε 3 clusters, διαστήματος 2 εβδομάδων:

```
[2, 1, 2, 3, 3, 2]
```

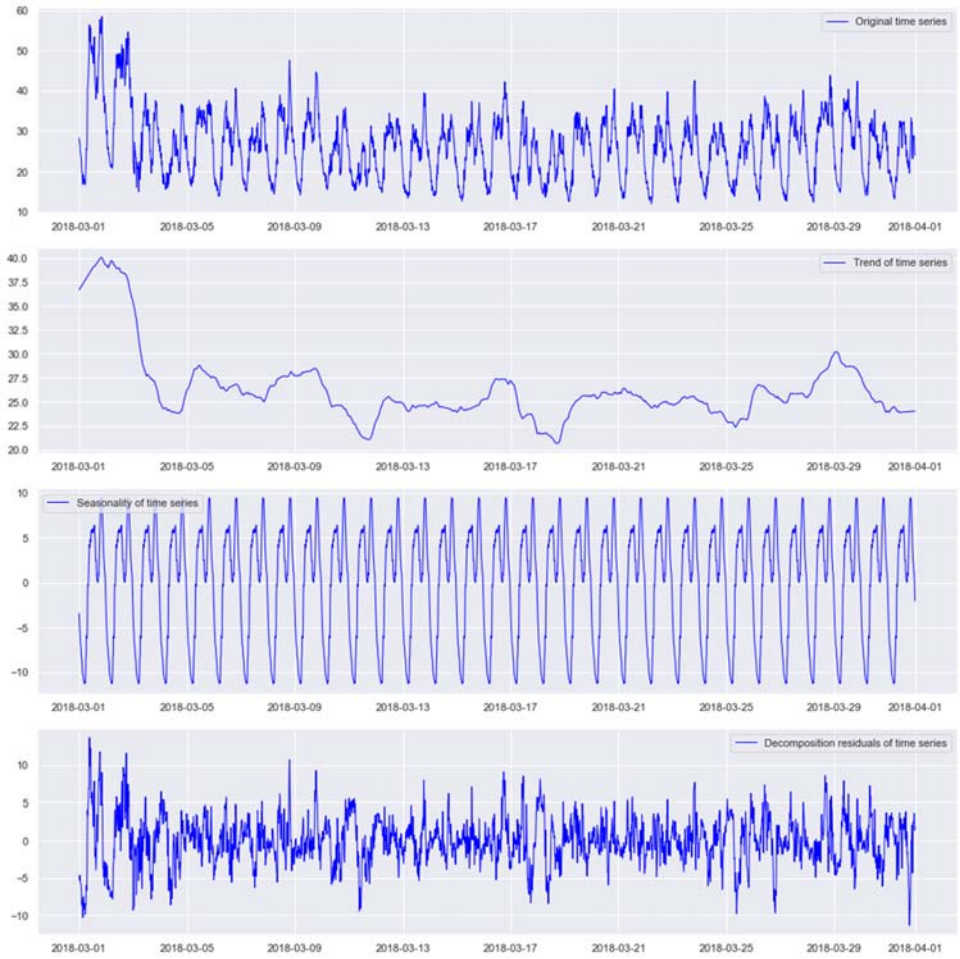
Με βάση την κάθε ακολουθία, επιλέγεται απ' τα δείγματα μας, μία χρονοσειρά η οποία να ανήκει στο cluster της κάθε υποομάδας (εργάσιμες και μη εργάσιμες ημέρες) που καταδεικνύεται κάθε φορά. Έτσι οπτικά για να καταλάβουμε τη διαφορά με τα πραγματικά δεδομένα, θα αναπαραστήσουμε την πρώτη εβδομάδα του Νοεμβρίου τόσο για τα αρχικά όσο και για τα generated δεδομένα, ενεργού ισχύος τους πρώτου ζυγού:



35 Σύγκριση αρχικών δεδομένων με generated σύμφωνα με τον δεύτερο τρόπο για μία εβδομάδα

4.5. Τρίτος τρόπος επαύξησης δεδομένων

Ο τρίτος τρόπος διαφέρει σημαντικά ως προς τη λογική επεξεργασίας των δεδομένων. Συγκεκριμένα, χειριζόμαστε τα δεδομένα ανά μήνα. Στα δεδομένα μήνα, κάνουμε αποσύνθεση χρονοσειρών σύμφωνα με τη θεωρία που περιγράψαμε στο Κεφάλαιο 2.4. Αυτό που κάνουμε ουσιαστικά είναι ότι αναλύουμε τις μηνιαίες χρονοσειρές μας σε τρεις συνιστώσες. Σε μια συνιστώσα εποχικότητας, μία τάσης και μία η οποία φανερώνει την τυχαιότητα ή αλλιώς το υπόλοιπο της χρονοσειράς. Ουσιαστικά το άθροισμα των τριών αυτών συνιστωσών μας φτιάχνει την αρχική μας χρονοσειρά. Παρακάτω, μπορούμε να δούμε οπτικά για τα δεδομένα ενεργού ισχύος του πρώτου ζυγού αυτό που, μόλις, περιγράψαμε.



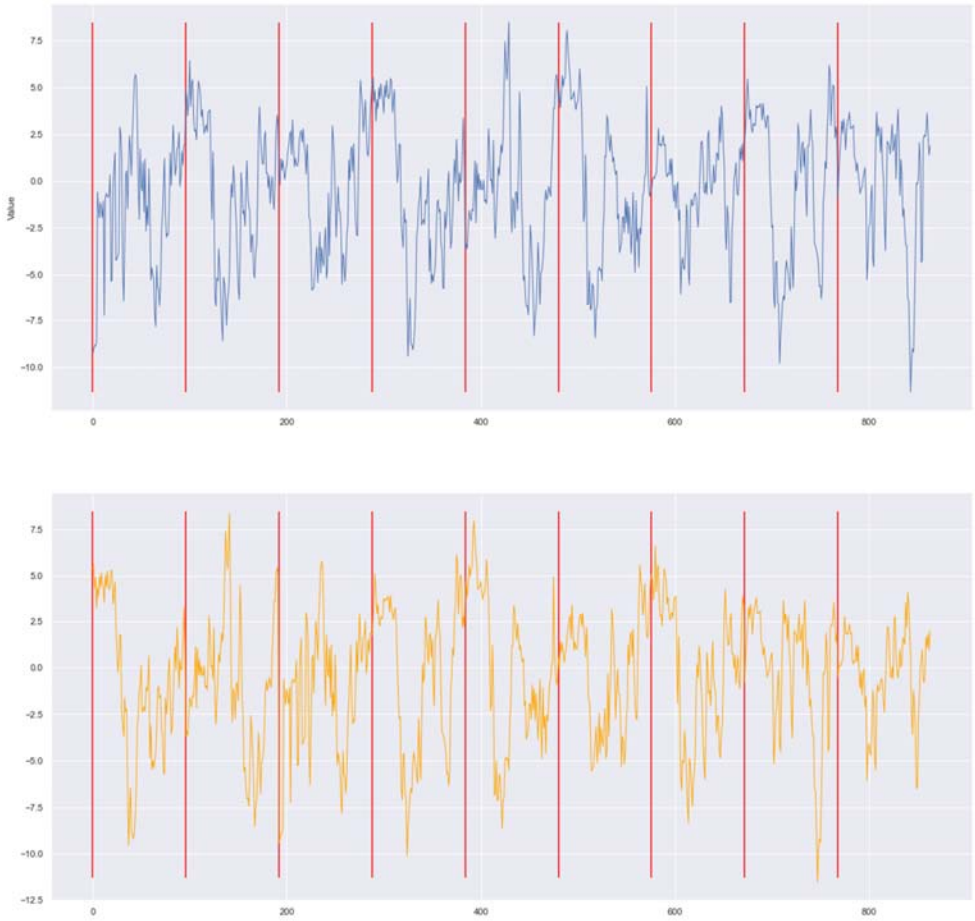
36 Αποσύνθεση δεδομένων ενός μήνα σε τάση, εποχικότητα και υπολείμματα

Μπορούμε να διαπιστώσουμε ότι η τάση είναι η συνιστώσα με το μεγαλύτερο πλάτος και ουσιαστικά διαμορφώνει την μορφή των δεδομένων. Απ' αυτή μπορούμε να καταλάβουμε εάν μια μέρα είναι εργάσιμη ή όχι ή να συγκρίνουμε μήνες μεταξύ τους ως προς τα πλάτη τους. Η εποχικότητα είναι μια συνιστώσα με σταθερό μοτίβο η οποία, αφού τα δεδομένα μας περιορίζονται σε ένα μήνα, έχει τη μορφή ημερήσιας εποχικότητας. Για κάθε μήνα είναι διαφορετική, όπως και οι υπόλοιπες συνιστώσες. Αυτή αν την απομονώσουμε δείχνει, πως τα δεδομένα μας έχουν την τάση να αυξομειώνονται μέσα σε μία μέρα (χαμηλά τις μη εργάσιμες ώρες και πιο υψηλά τις εργάσιμες). Τέλος, η συνιστώσα τυχαιότητας ο οποία αντιπροσωπεύει το απρόβλεπτο.

Ορμώμενοι απ' το τελευταίο, αποφασίσαμε για την παραγωγή καινούριων δεδομένων να κρατήσουμε σταθερές τις δύο πρώτες συνιστώσες και να πειράξουμε μόνο τη συνιστώσα της τυχαιότητας.

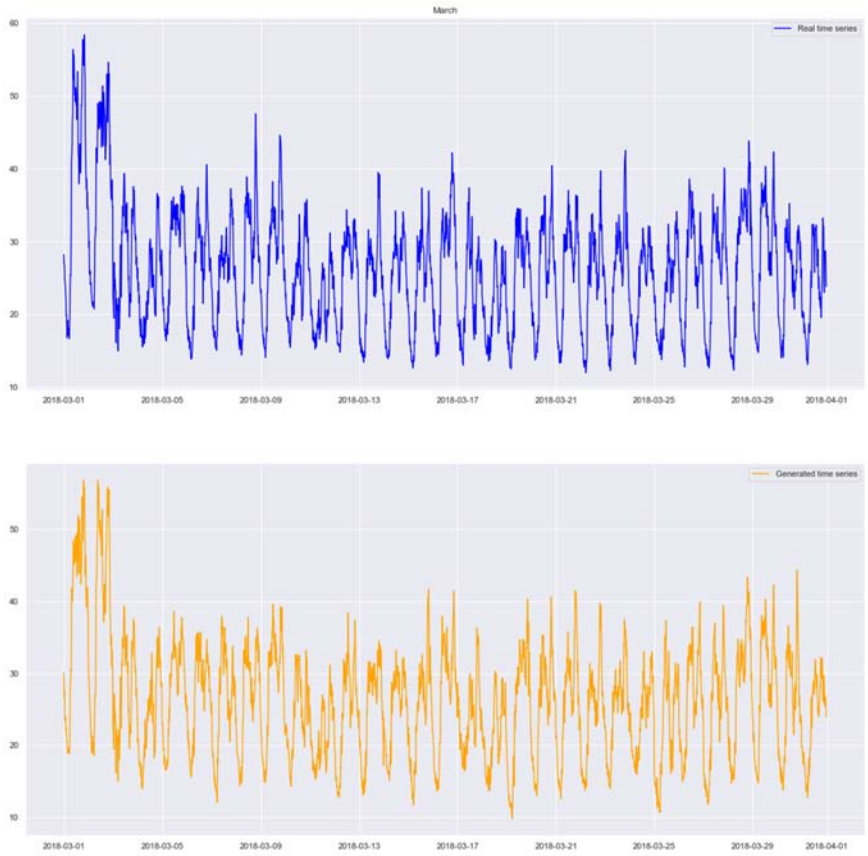
Αυτό που θα κάνουμε τώρα για να δημιουργήσουμε generated δεδομένα, είναι να πάρουμε τα residuals ανά μήνα και να τα χωρίσουμε σε δύο υποομάδες· σε residuals καθημερινών

και Σαββατοκύριακων. Μέσω της στατιστικής διαδικασίας του “bootstrap” θα ανακατανείμουμε στη συνέχεια τα residuals κάθε υποομάδας κατά τυχαίο τρόπο και θα τα προσθέσουμε στην τάση και την εποχικότητα των πραγματικών μας δεδομένων. Οπτικά θα δούμε παρακάτω τι εννοούμε ανακατανομή των residuals, για τα Σαββατοκύριακα ενός μήνα.



37 Ανακατανομή των residuals μέσω της διαδικασίας του ‘bootstrap’

Το μόνο που μένει για να δημιουργήσουμε generated δεδομένα είναι να προσθέσουμε τα ανακατανεμημένα residuals (πορτοκαλί χρονοσειρές) και για καθημερινές και για Σαββατοκύριακα στην τάση και την εποχικότητα όλου του μήνα στις θέσεις που πρέπει σειριακά. Μπορούμε να αναπαραστήσουμε παρακάτω τα πραγματικά δεδομένα ενός μήνα σε αντιπαράθεση με τα generated.



38 Σύγκριση αρχικών δεδομένων με generated σύμφωνα με τον τρίτο τρόπο για έναν μήνα

Και οι τρεις παραπάνω τρόποι, αφορούν παραγωγή generated δεδομένων για ένα μόνο ζυγό, χωρίς να λαμβάνουν υπόψιν τη θέση του στο δίκτυο και τη σχέση του με τους γειτονικούς ζυγούς του. Είναι δηλαδή προτεινόμενοι τρόποι παραγωγής δεδομένων για ανεξάρτητους ή μεμονωμένους ζυγούς. Ωστόσο, τα δεδομένα που μας δόθηκαν αφορούν γειτονικούς ζυγούς. Η συμπεριφορά της τάσης, της εποχικότητας και των ακραίων τιμών του ενός, επηρεάζεται άμεσα απ' τα αντίστοιχα χαρακτηριστικά του άλλου. Για το λόγο αυτό, τα καινούρια δεδομένα που θα παράγουμε στο επόμενο κεφάλαιο θα λαμβάνουν υπόψιν τα δεδομένα των γειτονικών, αυτού, ζυγών.

Ουσιαστικά, χρησιμοποιώντας τον τρόπο τον οποίο θεωρούμε ότι ήταν συνεπής και ως προς τα αποτελέσματα του, αλλά ήταν και σωστός ως προς την λογική του, θα παρουσιάσουμε τον προτεινόμενο τρόπο επίλυσης τους προβλήματος παραγωγής δεδομένων. Αυτός, θα λαμβάνει υπόψιν την συσχέτιση των δεδομένων και των τριών ζυγών ταυτόχρονα.

4.6. Προτεινόμενος τρόπος επίλυσης του προβλήματος

4.6.1. Εισαγωγή και διαχείριση των δεδομένων

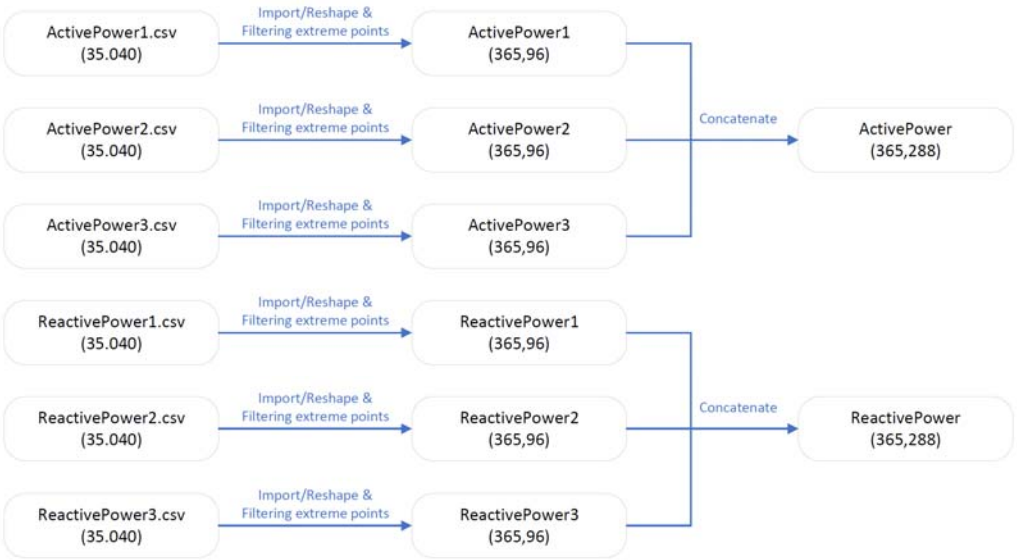
Ξεκινάμε, διαβάζοντας και αποθηκεύοντας τα δεδομένα μας ως Pandas Dataframes, ένα για κάθε αρχείο .csv που έχουμε στη διάθεση μας. Εφόσον τα δεδομένα μας, έχουν παρθεί μέσω δειγματοληψίας με συχνότητα 4 δείγματα/ώρα, δημιουργούμε έξι Dataframes, με μέγεθος (365*96). Έπειτα, ακολουθούμε τη διαδικασία που έχει περιγραφεί και στα προηγούμενα κεφάλαια και αφαιρούμε τυχόν μηδενικές τιμές ή ακραία σημεία.

Τα ονόματα των μεταβλητών-πινάκων, με τα φιλτραρισμένα δεδομένα μας μας φαίνονται παρακάτω:

active_power1, reactive_power1
active_power2, reactive_power2
active_power3, reactive_power3

Το επόμενο βήμα είναι να αποθηκεύσουμε, αφού συνενώσουμε τις παραπάνω μεταβλητές μας, σε δύο νέες μεταβλητές-πίνακες διάστασης (365, 288) ξεχωριστά, τα δεδομένα ενεργού και άεργου ισχύος. Αυτό σημαίνει ότι πλέον έχουμε τις εξής 2 μεταβλητές:

ActivePower
ReactivePower

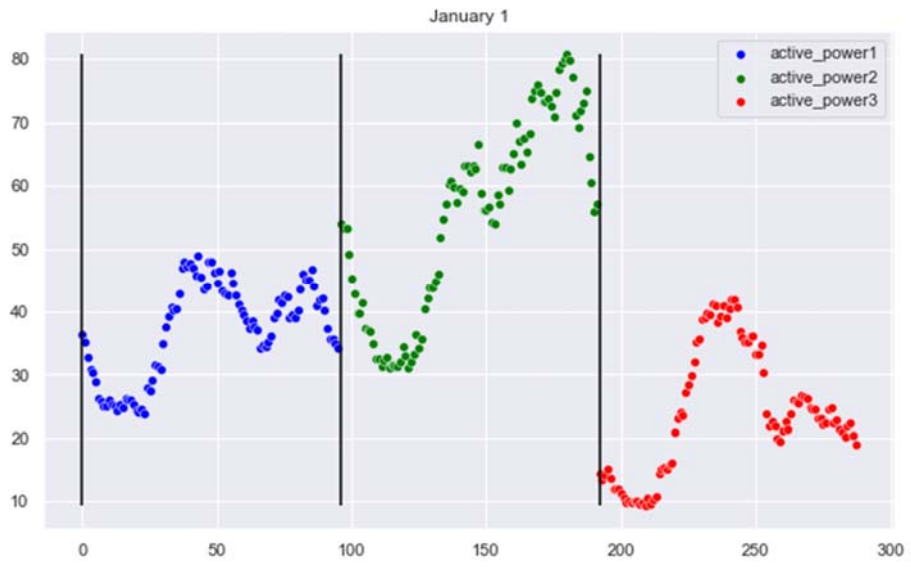


39 Διαγραμματική αναπαράσταση προεργασίας δεδομένων και δημιουργίας βασικών μεταβλητών

Ένας διαισθητικός τρόπος κατανόησης της δομής των δεδομένων, είναι αυτός της, παραπάνω, εικόνας. Με τον τρόπο που περιγράφεται, μας δίνεται η δυνατότητα να διαχειριστούμε τα δεδομένα και των τριών ζυγών σαν ένα διάνυσμα ανά ημέρα, διάστασης (1, 288). Η διαδικασία απ' το σημείο αυτό και μετά, δεν αλλάζει και πολύ με την παραγωγή

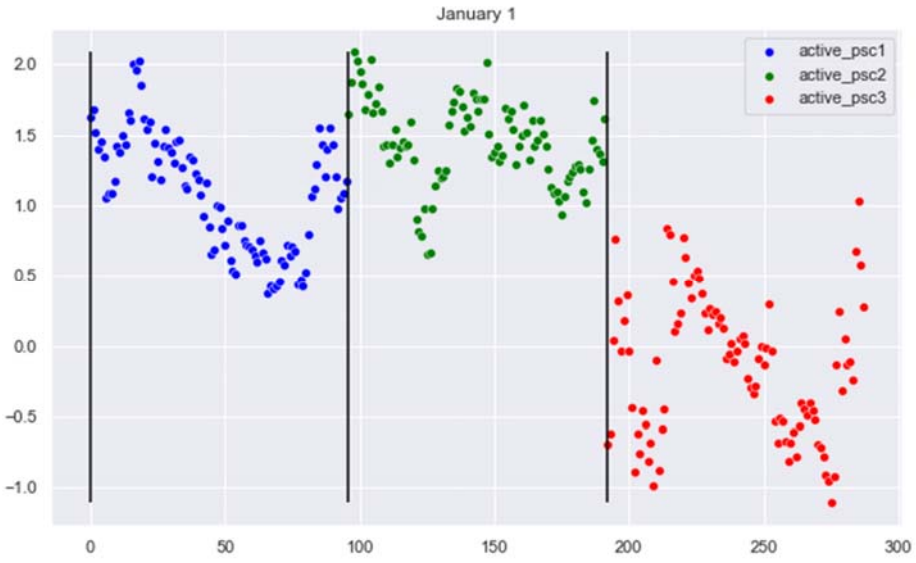
δεδομένων για έναν μόνο ζυγό. Απλώς τώρα, θα διαχειριζόμαστε ένα διάνυσμα μεγαλύτερο για κάθε ημέρα, το οποίο όταν το απομονώσουμε στις τρεις υποσυνιστώσες του, θα μας δίνει τα δεδομένα κάθε ζυγού ξεχωριστά.

Η ιδιαιτερότητα, ωστόσο, αυτού του συνενωμένου διανύσματος, είναι η ανομοιομορφία του ως προς τις τιμές του, όσο αφορά τον άξονα 'y', στα 3 σκέλη [1-96], [97-192], [193-288]. Προφανώς, ο αλγόριθμος μας δεν θα δώσει σωστά αποτελέσματα εάν δεν κάνουμε κάτι γι' αυτό. Έτσι, θα μετασχηματίσουμε με z-score κανονικοποίηση, πρώτα τα δεδομένα μας, σύμφωνα με τη θεωρία που περιγράψαμε στο 'κεφάλαιο 3' (βλ. 3.3.2). Τα δεδομένα μας θα μεταβληθούν, ώστε οι τιμές τους να διατηρούνται κοντά στο 0 και η τυπική τους απόκλιση κοντά στο 1. Θα αποφύγουμε δηλαδή, την παραγωγή ακραίων τιμών (στο βαθμό που αυτό είναι δυνατό) και θα βοηθήσουμε τον αλγόριθμο μας να επεξεργαστεί καλύτερα το διάνυσμα μας (1, 288) χωρίς τα διαφορετικά στατιστικά δεδομένα κάθε ζυγού (μέση τιμή, τυπική απόκλιση, ελάχιστο-μέγιστο) να επηρεάσουν την παραγωγή δεδομένων. Για να αντιληφθούμε τι εννοούμε, παρακάτω θα οπτικοποιήσουμε ένα ολικό διάνυσμα μιας τυχαίας ημέρας, πριν την κανονικοποίηση του.



40 Οπτικοποίηση δεδομένων ολικού διανύσματος και των 3 ζυγών για μία ημέρα

Μετά την κανονικοποίηση του, το ίδιο διάνυσμα μοιάζει όπως παρακάτω:



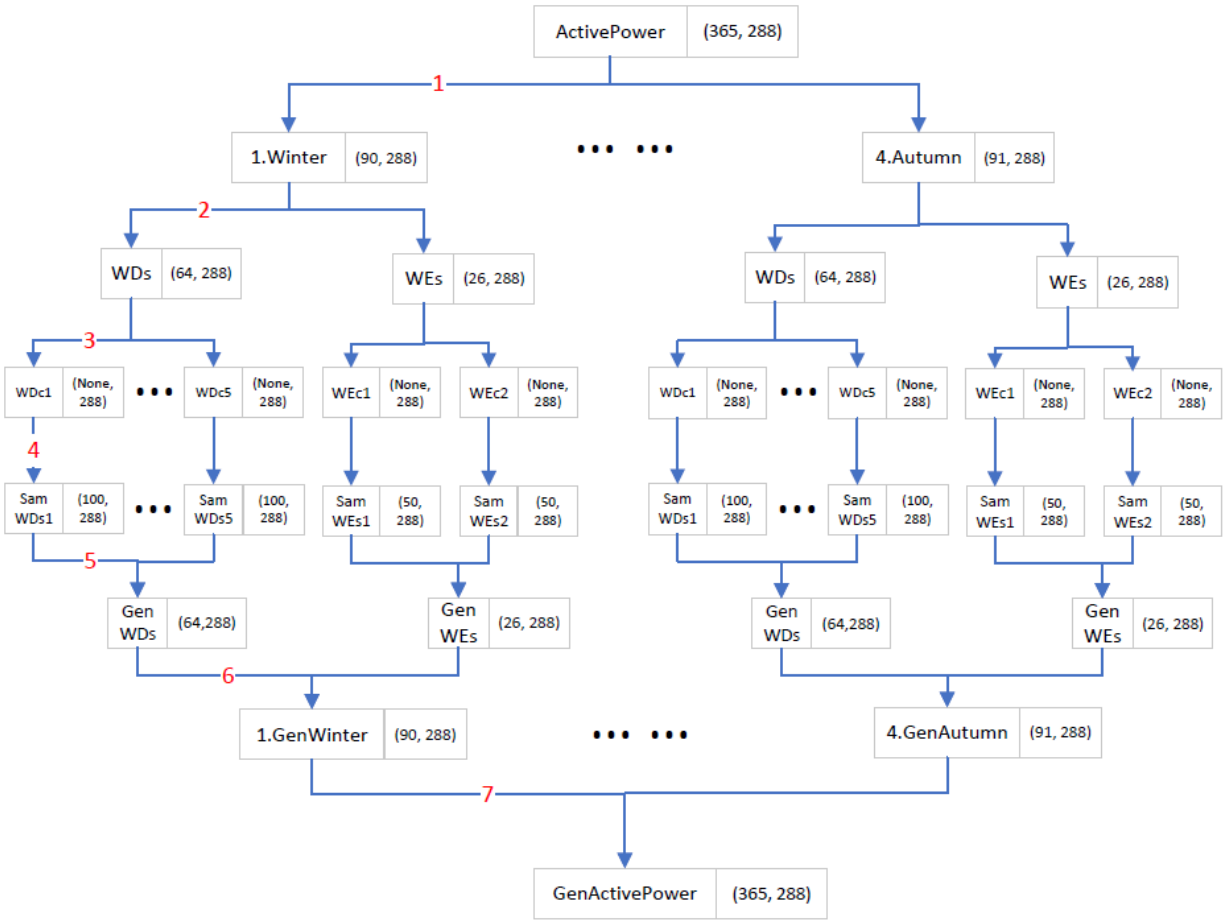
41 Οπτικοποίηση κανονικοποιημένων δεδομένων ολικού διανύσματος και των 3 ζυγών για μία ημέρα

4.6.2. Παραγωγή δεδομένων ενεργού ισχύος ενός χρόνου για τρεις ζυγούς

Απ' τις εικόνες, γίνεται εύκολα αντιληπτό, ότι εξηγήθηκε μέχρι τώρα και ακόμη, φαίνεται η μορφή που θα έχουν οι δύο μας βασικές μεταβλητές (ActivePower, ReactivePower) πριν και αφού μετασχηματιστούν. Βέβαια, έχουμε απεικονίσει μόνο μία ημέρα, δηλαδή ένα απ' τα 365 διανύσματα μεγέθους (1,288). Αφού, ολοκληρωθεί ο μετασχηματισμός ολόκληρης της μεταβλητής-πίνακα, διάστασης (365, 288), θα χωρίσουμε τα δεδομένα μας ανά εποχή και ύστερα θα τα χωρίσουμε σε δεδομένα εργάσιμων ημερών και δεδομένα από Σαββατοκύριακα. Ουσιαστικά, θα εκτελέσουμε τον 2^{ο} τρόπο, όπως περιεγράφηκε σε προηγούμενε κεφάλαιο (βλ 4.4). Εν συντομία, για να δημιουργήσουμε την μεταβλητή GenActivePower (βήμα 7^{ο} της εικ. 42) :

- Χωρίζουμε σε δεδομένα καθημερινών και δεδομένα από Σαββατοκύριακα (2^{ο} βήμα) και διενεργούμε GMM clustering με πέντε clusters για καθημερινές και δύο για Σαββατοκύριακα.
- Ουσιαστικά, το πόσες χρονοσειρές θα αναχθούν σε κάθε cluster, είναι κάτι που το καθορίζει ο αλγόριθμος μας. Έτσι, η πρώτη διάσταση των καινούριων μας μεταβλητών (βήμα 3^{ο}) είναι άγνωστη και συμβολίζεται με *None.*

- Αφού έχουμε τις χρονοσειρές μας χωρισμένες σε ομάδες, θα κάνουμε δειγματοληψία (4ο βήμα), δηλαδή, γένεση καινούριων χρονοσειρών (100 για κάθε cluster με δεδομένα καθημερινών και 50 για κάθε cluster με δεδομένα από Σαββατοκύριακα). Κοινό χαρακτηριστικό, των δειγμάτων μας, είναι ότι θα διατηρούν τις στατιστικές ιδιότητες (μέση τιμή, διακύμανση) της ομάδας.



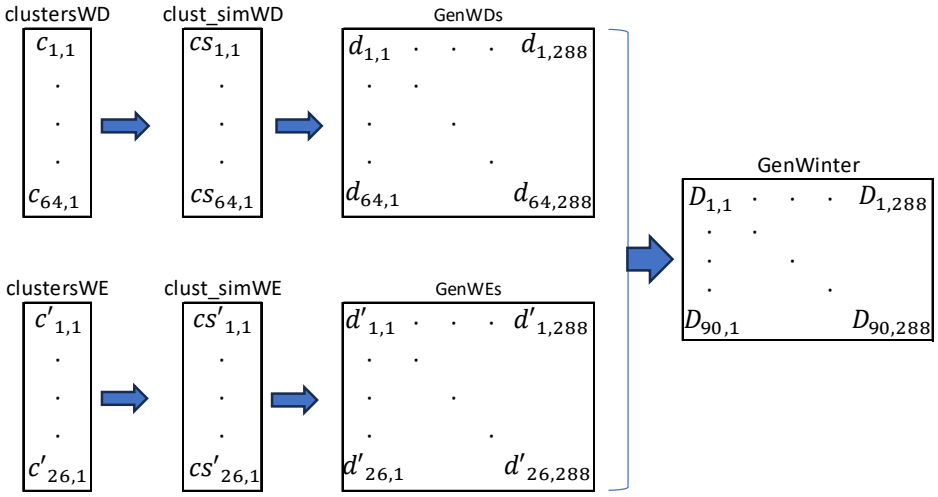
42 Διαγραμματική αναπαράσταση δημιουργίας GenActivePower

- Έχοντας δημιουργήσει αρκετά δείγματα για κάθε ομάδα, πρέπει να γίνει η επιλογή κάποιων απ' αυτά, τα οποία θα γίνουν συστατικά για τη δημιουργία μια ολόκληρης εποχής (βήμα 5ο). Η επιλογή θα γίνει με τον τρόπο τον οποίο θα περιγραφεί παρακάτω, μεσολαβεί του βήματος '4' αλλά παραλείπεται οπτικά στην εικόνα 41: α) Χαρτογραφούμε τα πραγματικά δεδομένα, ολόκληρης της εποχής τα οποία είναι αποθηκευμένα στην μεταβλητή του βήματος '1', με διάκριση την ομάδα η οποία τα χαρακτηρίζει στο βήμα '4'. Αποθηκεύουμε την τιμή της ομάδας για κάθε ημέρα· στην μεταβλητή 'clustersWD' εάν είναι εργάσιμη και στη μεταβλητή 'clustersWE' εάν μη εργάσιμη (βλ. εικ. 43). β) Έχοντας την χαρτογράφηση για εργάσιμες και μη εργάσιμες ημέρες, μπορούμε να βρούμε τους δύο πίνακες μετάβασης Markov ξεχωριστά για κάθε μεταβλητή μας. Οι πίνακες αυτοί αφορούν την πιθανοτική

μετάβαση από μια κατάσταση (δηλαδή ένα cluster) σε κάποια άλλη. Οι καταστάσεις μας είναι 5 για εργάσιμες και 2 για μη εργάσιμες ημέρες (προφανώς, όσα και τα clusters τα οποία επιλέξαμε στο βήμα '3' γ) Μπορούμε τώρα να εξομοιώσουμε σειριακά, μια ολόκληρη εποχή, λαμβάνοντας υπόψιν τους πίνακες μετάβασης που βρήκαμε προηγουμένως. Θα δημιουργήσουμε δύο καινούριες μεταβλητές μήκους όσου και του μήκους των 'clustersWD', 'clustersWE' αντίστοιχα, με ονόματα 'clust_simWD' και 'clust_simWE' (βλ. εικ. 43). Η εξομοίωση θα γίνει με τη χρήση ενός τίμιου ζαριού και θα τηρούνται οι πιθανότητες μετάβασης από ομάδα σε ομάδα δ) Από τις μεταβλητές διάστασης (64, 1) και (26, 1), όπου 64 είναι οι εργάσιμες μέρες του χειμώνα για το έτος το οποίο συλλέχθηκαν τα δεδομένα και 26 οι μη εργάσιμες, αντίστοιχα, πρέπει να καταλήξουμε σε δύο μεταβλητές διάστασης (64, 288) και (26, 288), τις 'GenWDs' και 'GenWEs'. Το 288 είναι το μήκος του διανύσματος για δεδομένα μίας ημέρας και των τριών ζυγών μαζί. Το πρόβλημα θα λυθεί με την μονοσήμαντη επιλογή για κάθε μία εκ των τιμών ομάδας ($c_{i,j}$, $cs'_{i,j}$), ενός διανύσματος (1,288) απ' την δειγματοληψία που έγινε στο βήμα '5'.

- Το μόνο που μένει τώρα για τη δημιουργία μιας ολόκληρης εποχής είναι να συνενώσουμε τις μεταβλητές 'GenWDs' και 'GenWEs', προσέχοντας να τηρήσουμε την ακολουθία εργάσιμων και μη εργάσιμων ημερών των πραγματικών μας δεδομένων.

Μπορούμε να κατανοήσουμε καλύτερα αυτό που περιγράψαμε στα προηγούμενα δύο βήματα, εάν κοιτάξουμε την παρακάτω εικόνα:

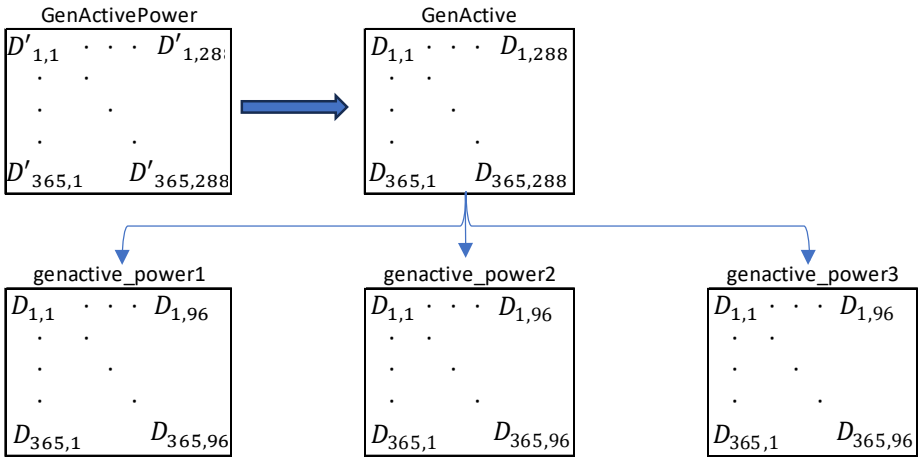


43 Διαγραμματική αναπαράσταση βημάτων 5-7 της εικόνας 42

Όπου: i) $c_{i,j}$, $cs_{i,j} \in$ {"WDc1", "WDc2", "WDc3", "WDc4", "WDc5"}
ii) $c'_{i,j}$, $cs'_{i,j} \in$ {"WEc1", "WEc2"}
iii) $d_{i,j}$, $d'_{i,j}$, $D_{i,j} \in (-\infty, +\infty)$

- Η παραπάνω διαδικασία πρέπει να συνεχιστεί και για τις επόμενες εποχές. Όταν ολοκληρωθεί, συνενώνοντας τα παραγόμενα δεδομένα και των τεσσάρων εποχών, καταλήγουμε στη μεταβλητή GenActicePower (βήμα 7ο)

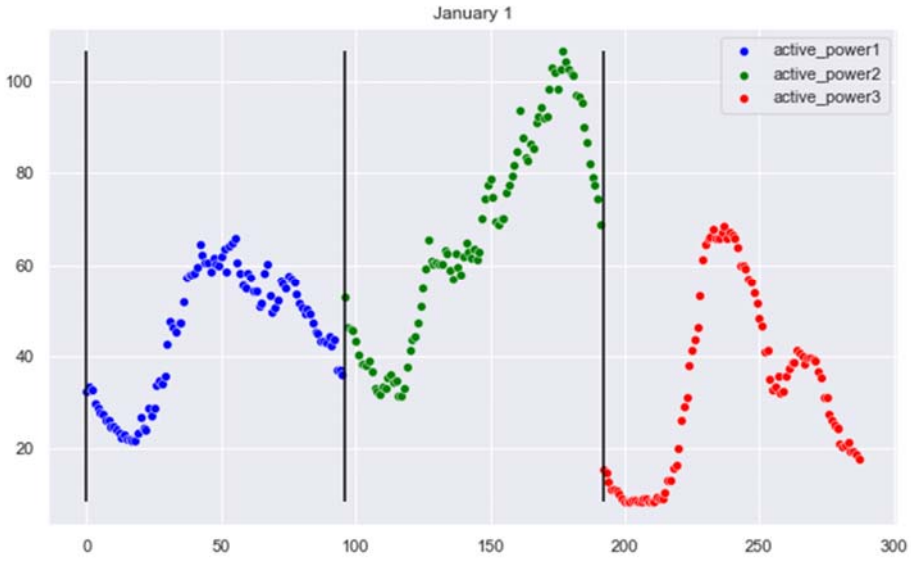
Έχοντας δημιουργήσει την μεταβλητή GenActicePower, έχουμε παράξει τεχνητά δεδομένα και για τους τρεις ζυγούς μας. Τα έχουμε όμως σε κανονικοποιημένη μορφή και συνενωμένα. Όπως έχει ήδη φανεί στην 'Εικόνα 41', τα δεδομένα αυτά, δεν έχουν πραγματική υπόσταση. Είναι όπως είχαμε εξηγήσει τιμές κοντά στο 0 και η τυπική τους απόκλιση κοντά στο 1. Πρέπει συνεπώς να ακολουθήσουμε πρώτα αντίστροφο μετασχηματισμό και έπειτα να διαχωρίσουμε την μεταβλητή μας στις τρεις επιμέρους συνιστώσες της, τις μεταβλητές genactive_power1, genactive_power2, genactive_power3 όπως θα φανεί παρακάτω:



44 Διαγραμματική αναπαράσταση διαχωρισμού ολικού διανύσματος σε ένα για κάθε ζυγό

Όπου $D'_{i,j} \in (-\infty, +\infty)$ και $D_{i,j} \in [0, +\infty)$

Τα τελικά παραγόμενα δεδομένα της «1ης Ιανουαρίου» (ημέρας για την οποία δείξαμε, προηγουμένως, τα αρχικά και τα κανονικοποιημένα δεδομένα), απαλλαγμένα απ' τον μετασχηματισμό τον οποίο τα είχαμε υποβάλει, είναι τα παρακάτω:



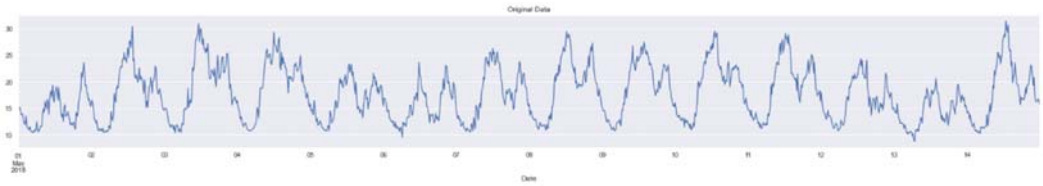
45 Generated δεδομένα του ολικού διανύσματος για την 1η Ιανουαρίου

4.6.3. Παραγωγή δεδομένων άεργου ισχύος ενός χρόνου για τρεις ζυγούς

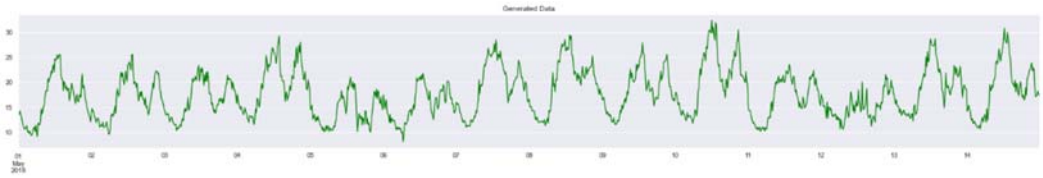
Προφανώς η διαδικασία είναι η ίδια με την παραγωγή δεδομένων ενεργού ισχύος. Το μόνο που αλλάζει, είναι η εισαγωγή στον αλγόριθμο μας της μεταβλητής ReactivePower έναντι της ActivePower. Δεν ακολουθούμε διαφορετική αντιμετώπιση για τα δεδομένα άεργου ισχύος καθώς επιθυμούμε τα δεδομένα μας να αποκτήσουν την ίδια τυχαιότητα με αυτή που θα αποκτήσουν τα δεδομένα ενεργού ισχύος. Το πρόβλημα ολοκληρώνεται με την παραγωγή των μεταβλητών genreactive_power1, genreactive_power2, genreactive_power3

4.7. Συγκριτικά αποτελέσματα

Έχοντας πλέον δημιουργήσει τα δικά μας τεχνητά δεδομένα, αξίζει να δούμε αρχικά πως αυτά φαίνονται οπτικά σε σύγκριση με τα αρχικά δεδομένα. Αναπαριστούμε δύο τυχαίες εβδομάδες του χρόνου:



46 Αρχικά δεδομένα για δύο τυχαίες εβδομάδες



47 Generated δεδομένα για τις δύο ίδιες εβδομάδες

Παρατηρούμε ότι διατηρούνται τα ολικά και τοπικά μέγιστα-ελάχιστα αλλά και η τάση των δεδομένων μας ως ένα βαθμό. Φαίνεται, ωστόσο, ότι μιλάμε για εντελώς διαφορετικά δεδομένα παρότι έχουμε αναπαραστήσει μόνο δύο εβδομάδες, διάστημα το οποίο αποτελεί στιγμιότυπο σε σχέση με ολόκληρο το χρόνο. Οπτικά, οι διαφορές μπορούν να φανούν εν μέρη αλλά σε καμία περίπτωση δε μπορεί να εξαχθεί πλήρες συμπέρασμα. Για το λόγο αυτό θα υπολογίσουμε κάποια στατιστικά μετρικά τα οποία θα μας βοηθήσουν να εξάγουμε χρήσιμα συμπεράσματα.

Bus 1

	Original Data	Generated Data	Error
Maximum Active Power(MW)	77,88	74,76	-4,17%
Maximum Reactive Power (MVAr)	21,12	19,42	-8,75%
Average Active Power (MW)	27,67	27,72	0,17%
Average Reactive Power (MVAr)	6,36	6,35	-0,07%
Active Energy Consumption (MWh)	242.399,33	242.808,81	0,17%
Reactive Energy Consumption (MVAr)	55.703,91	55.662,95	-0,07%

Bus 2

	Original Data	Generated Data	Error
Maximum Active Power(MW)	116,22	124,23	6,45%
Maximum Reactive Power (MVAr)	15,36	14,99	-2,47%
Average Active Power (MW)	33,18	34,46	3,73%
Average Reactive Power (MVAr)	5,96	6,06	1,70%
Active Energy Consumption (MWh)	290.646,91	301.912,21	3,73%
Reactive Energy Consumption (MVAr)	52.191,46	53.093,42	1,70%

Bus 3

	Original Data	Generated Data	Error
Maximum Active Power(MW)	76,08	74,37	-2,30%
Maximum Reactive Power (MVAr)	17,04	16,95	-0,53%
Average Active Power (MW)	25,14	24,96	-0,71%
Average Reactive Power (MVAr)	5,41	5,45	0,69%
Active Energy Consumption (MWh)	220.184,15	218.636,95	-0,71%
Reactive Energy Consumption (MVAr)	47.421,96	47.753,80	0,69%

48 Συγκριτικά αποτελέσματα στατιστικών μετρικών για αρχικά και Generated δεδομένα

Όπως μπορούμε να δούμε το πρώτο μετρικό και ιδιαιτέρως σημαντικό για δεδομένα φορτίου, είναι η μέγιστη τιμή. Η μέγιστη ισχύς είναι πολύ σημαντική στην ανάλυση του δικτύου σε θέματα που αφορούν την προστασία του δικτύου, την βελτιστοποίηση του, την εξισορρόπηση του φορτίου ανά ζυγό, την αποτροπή αποτυχιών και την τιμολόγηση. Στο αρχικό σύνολο δεδομένων ο ‘ζυγός 1’ έχει μέγιστη ενεργή ισχύ 77,88 MW, ο ‘ζυγός 2’ 116,22 MW και ο ‘ζυγός 3’ 76,08 MW. Στο παραγόμενο σύνολο δεδομένων αυτές οι τιμές είναι 74,76 MW, 124,23 MW και 74,37 MW αντίστοιχα. Η δεύτερη μέτρηση που επιλέξαμε είναι η μέση ισχύς. Εν γένει, η μέση τιμή ισχύος για ένα ζυγό, είναι αυτή που μας βοηθάει στο σχεδιασμό, τη διαχείριση του, μας βοηθάει να κάνουμε προβλέψεις και να αναγνωρίσουμε περιθώρια βελτίωσης του δικτύου. Η μέση ισχύς και κυρίως η σύγκριση των δύο μέσων τιμών (αρχικά/παραγόμενα δεδομένα), είναι αυτή που θα μας βοηθήσει να καταλάβουμε τη συνολική εικόνα όσο αφορά τα καινούρια δεδομένα μας. Ως προς αυτό, θα παρατηρήσουμε έναν δείκτη, που προστέθηκε στην τρίτη στήλη των παραπάνω πινάκων και είναι το ποσοστιαίο σφάλμα. Θα παρατηρήσουμε ότι, αναφερόμενοι, στη μέγιστη τιμή ισχύος, το απόλυτο σφάλμα κυμαίνεται από τιμές της τάξεως του 0,53% έως το 8,75% (και για ενεργό και για άεργο ισχύ). Μιλώντας για τη μέση ισχύ, όμως, το απόλυτο σφάλμα παίρνει τιμές από 0,16% έως 3,71%. Αυτό σημαίνει ότι τα παραγόμενα δεδομένα μας έχουν άμεση συσχέτιση με τα αρχικά (το είδαμε και οπτικά πιο πριν). Η ελάχιστη τιμή σαν μετρικό δεν λήφθηκε από επιλογή υπόψιν και δεν καταμετρήθηκε. Αυτό συνέβη, όπως έχει εξηγηθεί και σε προηγούμενο κεφάλαιο διότι τα αρχικά δεδομένα μας είχαν ελάχιστη τιμή ίση με το μηδέν, αυτή διορθώθηκε και τα παραγόμενα δεδομένα δημιουργήθηκαν με αφετηρία δεδομένα χωρίς μηδενικές ή ακραίες τιμές. Άλλωστε, εάν ο χρήστης θέλει

ανάλογα με την προβλεπόμενη εφαρμογή, τέτοια σημεία μπορούν εύκολα να εισαχθούν τυχαία μετά την παραγωγή των δεδομένων. Η τελική μέτρηση είναι η κατανάλωση ενέργειας καθ' όλη τη διάρκεια του έτους. Μπορεί εύκολα να γίνει κατανοητό ότι δείχνει τα ίδια αποτελέσματα με τη μέση ισχύ, αφού η μία μπορεί να προκύψει από την άλλη πολλαπλασιάζοντας με 8.760. Ωστόσο, παρουσιάζεται εδώ για ευκολότερη σύγκριση των αποτελεσμάτων.

4.8. Μελέτη ροών φορτίου

Η ανάλυση ροών φορτίου συνεπάγεται τον υπολογισμό των αγνώστων τάσεων των ζυγών και των αγνώστων ροών ισχύος ενός συστήματος ηλεκτρικής ενέργειας για μία δεδομένη επιλογή ισχύων παραγωγής, τάσεων γεννητριών και φορτίων. Οι μελέτες ροών φορτίου είναι πολύ χρήσιμες για διάφορους λόγους. Μερικοί από αυτούς αναφέρονται παρακάτω:

1. Είναι απαραίτητες για την εκλογή της πλέον οικονομικής λειτουργίας των γεννητριών του συστήματος. Επειδή στην περίοδο της ημέρας τα φορτία μεταβάλλονται συνεχώς, απαιτείται συχνά ο υπολογισμός εκ νέου της παραγόμενης ισχύος κάθε γεννήτριας.

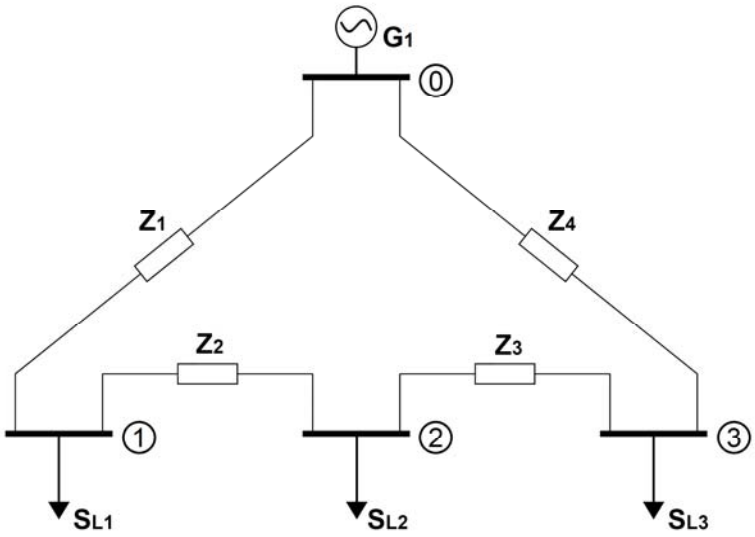
2. Είναι απαραίτητες στη διατήρηση τάσεων και ροών εντός προκαθορισμένων ορίων λειτουργίας.

3. Είναι απαραίτητες στη μελέτη ενδεχομένων διαταραχών.

4. Είναι απαραίτητες σε μελέτες επέκτασης του συστήματος παραγωγής και μεταφοράς ηλεκτρικής ενέργειας.

Από τα παραπάνω γίνεται φανερό ότι οι μελέτες ροών φορτίου είναι θεμελιώδους σημασίας στη μελέτη συστημάτων ηλεκτρικής ενέργειας. Η ανάλυση ροών φορτίου προϋποθέτει τη συμμετρική κατάσταση του συστήματος και για το σκοπό αυτό επαρκεί το μονογραμμικό διάγραμμα του συστήματος ηλεκτρικής ενέργειας. Ένα τυπικό και ίσως το πιο απλό μονογραμμικό διάγραμμα του συστήματος που μελετάμε στο πείραμα μας, μπορεί να απεικονιστεί όπως στην 'Εικόνα 49' (να σημειωθεί ότι το διάγραμμα που απεικονίζεται στην εικόνα, έχει επιλεχθεί αυθαίρετα και καμία σχέση έχει με το πραγματικό δίκτυο απ' το οποίο και πάρθηκαν τα δεδομένα μας):

Όσο αφορά το σχήμα, ο ζυγός '0', είναι ο ζυγός αναφοράς και για τη ρύθμιση της τάσης του συστήματος έχουμε προσθέσει μια γεννήτρια πάνω σε αυτόν. Οι ζυγοί '1', '2' και '3' είναι ζυγοί φορτίου (μιγαδικού). Σαν φορτία θα χρησιμοποιήσουμε τα δεδομένα που έχουμε στο πείραμα μας καθώς $S_{Li} = P_i + jQ_i$. Με βάση όσα αναφέρθηκαν και παραπάνω οι άγνωστοι μας αφού δημιουργήσουμε τις εξισώσεις για τη μελέτη ροής φορτίου, είναι οι τρεις διανυσματικές τάσεις των ζυγών φορτίου και οι ροές ισχύος των γραμμών μεταφοράς. Εμείς με τα δεδομένα που έχουμε στο πείραμα μας μπορούμε να δημιουργήσουμε τόσα προβλήματα μελέτης ροών φορτίου, όσα και τα δεδομένα ισχύος τα οποία διαθέτουμε ανά ζυγό. Μπορούμε, δηλαδή, να λύσουμε 365*96 σετ εξισώσεων, τα οποία ένα προς ένα μας δείχνουν ένα στιγμιότυπο της κατάστασης όσο αφορά τις τάσεις, τις γωνίες και τις ροές ισχύος. Δεν έχει κάποια αξία, η επίλυση μόνο ενός σετ εξισώσεων το οποίο θα προκύψει από ένα σετ φορτίων (φορτία μιας χρονικής στιγμής). Για το λόγο αυτό θα λύσουμε όλες τις



49 Τυπικό μονογραμμικό διάγραμμα ενός ΣΗΕ σαν του πειράματος μας

εξισώσεις που θα προκύψουν απ' τα φορτία που έχουμε στη διάθεση μας τόσο για τα αρχικά δεδομένα όσο και για τα παραγόμενα για να δούμε την μεταβολή μέσα στο χρόνο τόσο των τάσεων, όσο και των ροών ισχύος και να δούμε τι συσχέτιση έχουν τα δύο διαφορετικά σύνολα δεδομένων. Είναι σημαντικό το ότι επιλέξαμε τα δύο διαφορετικά σύνολα φορτίων να τα εξομοιώσουμε μέσα απ' το ίδιο σύστημα ζυγών, καθώς έτσι θα μπορούμε να δούμε και εάν εν τέλει τα φορτία που παρήχθησαν θα έχουν ίδιο αντίκτυπο στο δίκτυο, όποιο και εάν είναι αυτό.

Για τον υπολογισμό των ροών ισχύος στο ηλεκτρικό δίκτυο, επιλέχθηκε ο αλγόριθμος Newton-Raphson. Αυτή η επιλογή έγινε μεταξύ άλλων αλγορίθμων όπως οι μέθοδοι Gauss-Seidel και Fast Decoupled. Ο Newton-Raphson είναι ένας αρκετά αξιόπιστος και ακριβής αλγόριθμος και μας επαρκεί για τον υπολογισμό των τάσεων και των ροών ισχύος του δικτύου μας. Ουσιαστικά, ο αλγόριθμος Newton-Raphson είναι ένας επαναληπτικός αλγόριθμος ο οποίος λειτουργεί επιδιώκοντας τη σύγκλιση των τάσεων και των ροών ισχύος μεταξύ δύο επαναλήψεων. Κριτήριο για να σταματήσει ο αλγόριθμος, είναι οι διαφορές μεταξύ των διαδοχικών επαναλήψεων για τις τάσεις και τις ροές ισχύος να είναι κάτω από ένα προκαθορισμένο κατώφλι σφάλματος. Αυτό το κριτήριο σύγκλισης ορίζει την επιθυμητή ακρίβεια των υπολογισμών και στο πείραμα μας τέθηκε ίσο με '1e-8' για πολύ μεγάλη ακρίβεια. Επιπλέον, μπορεί να οριστεί και ένα μέγιστο πλήθος επαναλήψεων, ώστε ο αλγόριθμος να σταματήσει ακόμα και εάν δεν υπάρχει σύγκλιση, βάση του κατωφλίου. Στον αλγόριθμο μας, ωστόσο, ο αριθμός αυτός τέθηκε ίσος με '100' ώστε να υπερισχύει η ακρίβεια, έναντι της ταχύτητας.

Ας δούμε παρακάτω τα αποτελέσματα που έδωσε ο αλγόριθμος μας συνοπτικά για όλο το χρόνο:

Bus1

	Orig_mean	Gen_mean	Orig_max	Gen_max	Orig_min	Gen_min
Voltage Magnitude (p.u.)	0.99	0.99	1.00	1.00	0.97	0.98
Voltage Angle (p.u.)	-0.57	-0.58	-0.18	-0.13	-1.72	-1.72

Bus2

	Orig_mean	Gen_mean	Orig_max	Gen_max	Orig_min	Gen_min
Voltage Magnitude (p.u.)	0.99	0.99	1.00	1.00	0.96	0.97
Voltage Angle (p.u.)	-0.79	-0.81	-0.24	-0.16	-2.53	-2.55

Bus3

	Orig_mean	Gen_mean	Orig_max	Gen_max	Orig_min	Gen_min
Voltage Magnitude (p.u.)	0.99	0.99	1.00	1.00	0.97	0.98
Voltage Angle (p.u.)	-0.56	-0.57	-0.18	-0.13	-1.68	-1.66

50 Αποτελέσματα μελέτης ροής ισχύος για του ζυγούς

Line 1

	Orig_mean	Gen_mean	Orig_max	Gen_max	Orig_min	Gen_min
P_From (MW)	16.01	16.51	58.89	56.09	3.80	3.34
Q_From (MVAR)	2.84	2.82	8.92	8.50	-0.11	-0.17
P_To (MW)	-15.96	-16.45	-3.79	-3.33	-58.29	-55.56
Q_To (MVAR)	-2.75	-2.73	0.25	0.23	-8.04	-7.88
PLoss (MW)	0.06	0.06	0.60	0.53	0.00	0.00
QLoss (MVAR)	0.09	0.10	0.97	0.86	0.00	0.00

Line 2

	Orig_mean	Gen_mean	Orig_max	Gen_max	Orig_min	Gen_min
P_From (MW)	-17.22	-17.51	-1.29	-2.05	-61.48	-61.32
Q_From (MVAR)	-3.21	-3.17	-0.25	-0.24	-7.82	-7.93
P_To (MW)	17.29	17.58	62.14	61.97	1.29	2.05
Q_To (MVAR)	3.32	3.29	8.76	8.55	0.37	0.37
PLoss (MW)	0.07	0.07	0.65	0.65	0.00	0.00
QLoss (MVAR)	0.11	0.12	1.06	1.05	0.00	0.00

Line 3

	Orig_mean	Gen_mean	Orig_max	Gen_max	Orig_min	Gen_min
P_From (MW)	-43.68	-44.33	-14.39	-13.62	-123.70	-120.69
Q_From (MVAR)	-9.20	-9.15	-2.32	-1.87	-23.39	-23.72
P_To (MW)	44.08	44.73	126.34	123.17	14.43	13.66
Q_To (MVAR)	9.83	9.80	24.82	24.70	2.72	1.99
PLoss (MW)	0.39	0.40	2.64	2.48	0.04	0.03
QLoss (MVAR)	0.64	0.65	4.29	4.02	0.06	0.05

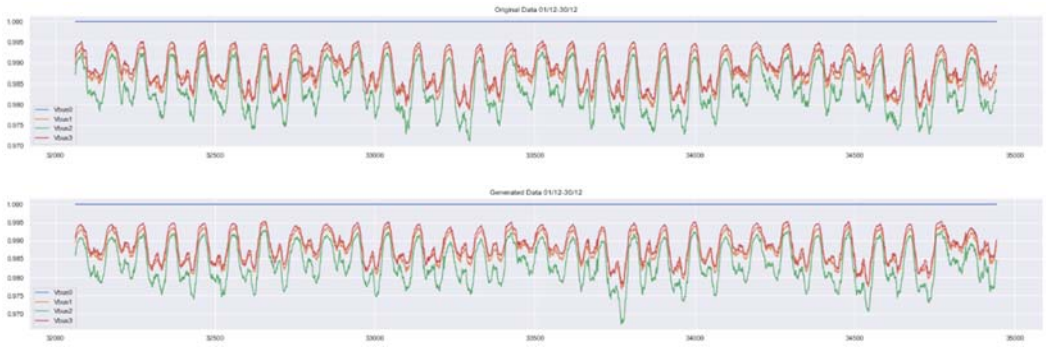
Line 4

	Orig_mean	Gen_mean	Orig_max	Gen_max	Orig_min	Gen_min
P_From (MW)	-42.43	-43.28	-14.42	-13.42	-120.92	-116.42
Q_From (MVAR)	-8.74	-8.71	-2.55	-2.00	-20.72	-22.88
P_To (MW)	42.80	43.67	123.44	118.73	14.46	13.45
Q_To (MVAR)	9.34	9.33	24.03	24.16	2.81	2.31
PLoss (MW)	0.37	0.38	2.52	2.30	0.04	0.03
QLoss (MVAR)	0.60	0.62	4.10	3.74	0.06	0.05

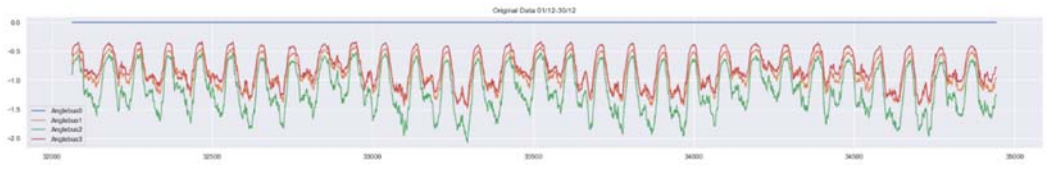
51 Αποτελέσματα μελέτης ροής ισχύος για τις γραμμές μεταφοράς

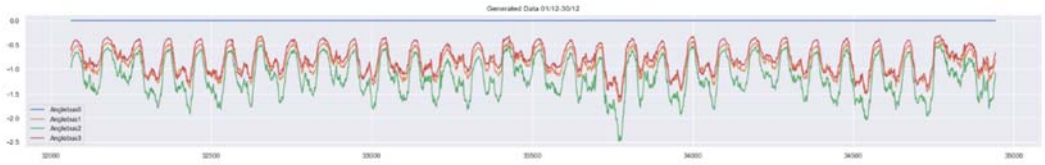
Κοιτώντας τα δεδομένα μας παρατηρούμε μια σχετική συσχέτιση ως προς τις τιμές όσο αφορά τη διακύμανση των τάσεων και των γωνιών για τους ζυγούς και ακόμη της έκχυσης ισχύος και των απωλειών όσο αφορά τις γραμμές μεταφοράς. Συγκεκριμένα η μέση τάση είναι ίση και για αρχικά και για τελικά δεδομένα, η μέση γωνία παρουσιάζει μέγιστη διαφορά στον ζυγό 2, της τάξεως του 2,4%. Όσο αφορά τις μέγιστες διαφορές στις γραμμές μεταφοράς, αυτές παρουσιάζονται στην πρώτη γραμμή και συγκεκριμένα η διαφορά της ενεργού ισχύος που εισέρχεται στη γραμμή από τον ζυγό αναφοράς (slack) περιέχει σφάλμα 3%, η ενεργός ισχύς που εξέρχεται στο 'ζυγό 1' παρουσιάζει σφάλμα 2,9%, ενώ τέλος συνολική απώλεια επαγωγικής ισχύος στη γραμμή παρουσιάζει σφάλμα 10%.

Τα παραπάνω πινακάκια και οι όποιες επεξηγήσεις έγιναν, δεν μας αφήνουν να αντιληφθούμε εξ ολοκλήρου την συνάφεια μεταξύ των δεδομένων που πήραμε απ' τις δύο διαφορετικές μελέτες ροής φορτίου που διεξήχθησαν, παρά τα μικρά σφάλματα. Για το λόγο αυτό, παρακάτω, θα αναπαρασταθούν τα διαγράμματα των τάσεων, των γωνιών και της ενεργού ισχύος που εισέρχεται σε κάθε γραμμή για ένα μήνα για να μπορέσει να γίνει η οπτική σύγκριση.

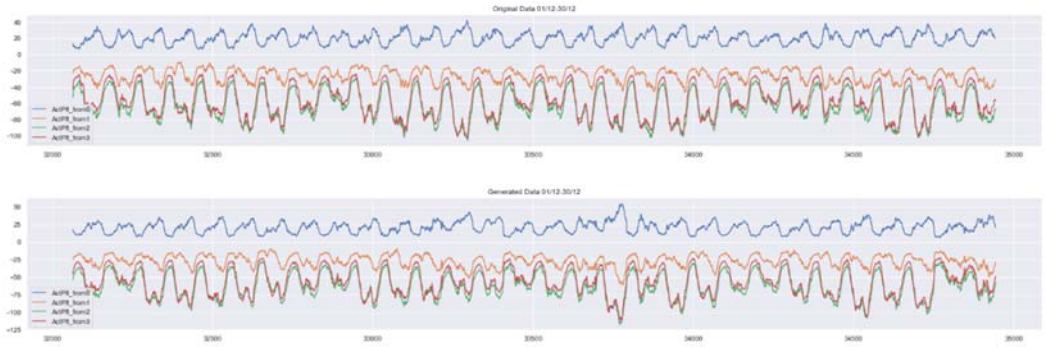


52 Οπτικοποίηση τάσεων των 4 ζυγών με αρχικά-τεχνητά φορτία για ένα μήνα





53 Οπτικοποίηση γωνιών των τάσεων των 4 ζυγών με αρχικά-τεχνητά φορτία για ένα μήνα



54 Οπτικοποίηση ενεργού ισχύος των 4 ζυγών με αρχικά-τεχνητά φορτία για ένα μήνα

Παρατηρούμε ότι παρά την ανεξαρτησία κάθε διαγράμματος των παραγόμενων δεδομένων και την εμφανή μοναδικότητά τους σε σχέση με τα αντίστοιχα αρχικά, υπάρχει μια συσχέτιση όσο αφορά τα ολικά μέγιστα και ελάχιστα. Ειδικότερα, εάν εστιάσουμε σε δεδομένα ημερών, είναι εμφανές ότι έχουμε παρόμοια συμπεριφορά σε ώρες εργάσιμες, με αντίστοιχη άνοδο και παρόμοια συμπεριφορά σε αρχή και τέλος της ημέρας όπου οι ώρες είναι μη εργάσιμες.

Συμπερασματικά, η μελέτη ροής φορτίου πραγματοποιήθηκε σε δύο στάδια. Στο πρώτο στάδιο, εκτελέστηκε η ροή φορτίου σε ένα δίκτυο που αποτελείται από τους τρεις ζυγούς φορτίου με φορτία τα αρχικά μας δεδομένα και έναν τέταρτο ζυγό με γεννήτρια. Στο δεύτερο στάδιο, αντικαταστάθηκαν οι τρεις αρχικοί ζυγοί με τρεις νέους, αυτούς για τους οποίους παρήγαμε δεδομένα στο ‘κεφ. 4.6’. Οι νέοι ζυγοί παρουσίασαν παρόμοια μέση τιμή και ίδια μοτίβα στα αποτελέσματά τους όπως και οι αρχικοί ζυγοί. Τε γεγονός αυτό φανερώνει σταθερότητα συστήματος και ομοιότητα ως προς τους ζυγούς. Με βάση αυτό, μπορούμε να πούμε ότι οι ζυγοί, παρόλο που είναι διαφορετικοί στο σύνολο των δεδομένων, παρουσιάζουν παρόμοια συμπεριφορά στο δίκτυο.

ΚΕΦΑΛΑΙΟ 5

5. Συμπεράσματα και μελλοντικοί στόχοι

5.1. Μεθοδολογία και αποτελέσματα

Στην παρούσα διπλωματική εργασία, αναλύθηκε η χρήση σύγχρονων μεθόδων στην επεξεργασία και ανάλυση δεδομένων ηλεκτρικής ενέργειας. Ζητούμενο ήταν η δημιουργία νέων δεδομένων όταν μόνο περιορισμένος αριθμός δεδομένων εκπαίδευσης είναι διαθέσιμος. Η εφαρμογή προηγμένων τεχνικών, όπως η χρήση Gaussian Mixture Models και Markov Chains και η αποσύνθεση χρονοσειρών, αποδείχθηκε ικανή να δημιουργήσει δεδομένα που αποδίδουν με ακρίβεια τις χρονικές και εποχιακές μεταβολές της ζήτησης ενέργειας. Η προσέγγιση αυτή αποδείχθηκε ιδιαίτερα χρήσιμη στην παραγωγή συνθετικών δεδομένων (όπως φάνηκε στο 'Κεφάλαιο 4') ικανών να μιμούνται τις πραγματικές συνθήκες, παρέχοντας μια σταθερή βάση για μελλοντικές αναλύσεις και εφαρμογές στον τομέα των έξυπνων δικτύων.

5.2. Μελλοντικοί Στόχοι

Όσον αφορά τη μελλοντική έρευνα, είναι σημαντικό να εξεταστεί η δυνατότητα ενσωμάτωσης περισσότερων δεδομένων από ποικίλες πηγές, σε πιο σύνθετα δίκτυα και μεγαλύτερο όγκο δεδομένων. Μια σημαντική πρόκληση μπορεί να είναι η ενσωμάτωση περισσότερων παραμέτρων, όπως η ενσωμάτωση μετεωρολογικών δεδομένων και η διερεύνηση της σχέσης τους με τα φορτία των ζυγών του δικτύου. Στόχος, συμπερασματικά, μπορεί είναι η εξέταση της δυναμικής αλληλεπίδρασης μεταξύ διαφορετικών παραμέτρων και η ανάπτυξη μοντέλων που θα μπορούν να προβλέψουν τις ανάγκες των έξυπνων δικτύων με μεγαλύτερη ακρίβεια και αποδοτικότητα.

[1] T. T. Um *et al.*, “Data augmentation of wearable sensor data for Parkinson’s disease monitoring using convolutional neural networks,” in *ICMI 2017 - Proceedings of the 19th ACM International Conference on Multimodal Interaction*, Association for Computing Machinery, Inc, Nov. 2017, pp. 216–220. doi: 10.1145/3136755.3136817.

[2] Nikolaos Gkovatsos, “Modeling Loads and Production in Distribution Networks using Statistics to Generate Artificial Data for Optimal Operation Studies”, Accessed: Jan. 16, 2024. [Online]. Available: https://github.com/nicksgov/Thesis

[3] B. K. Iwana and S. Uchida, “Time Series Data Augmentation for Neural Networks by Time Warping with a Discriminative Teacher,” Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.08780

[4] A. Le Guennec, S. Malinowski, and R. Tavenard, “Data Augmentation for Time Series Classification using Convolutional Neural Networks,” 2016. [Online]. Available: https://shs.hal.science/halshs-01357973

[5] H. Cao, V. Y. F. Tan, and J. Z. F. Pang, “A parsimonious mixture of gaussian trees model for oversampling in imbalanced and multimodal time-series classification,” *IEEE Trans Neural Netw Learn Syst*, vol. 25, no. 12, pp. 2226–2239, Dec. 2014, doi: 10.1109/TNNLS.2014.2308321.

[6] Y. Hou, Y. Liu, W. Che, and T. Liu, “Sequence-to-Sequence Data Augmentation for Dialogue Language Understanding.” [Online]. Available: https://github.com/AtmaHou/Seq2SeqDataAugmentationForLU.

[7] J. Tu, H. Liu, F. Meng, M. Liu, and R. Ding, “SPATIAL-TEMPORAL DATA AUGMENTATION BASED ON LSTM AUTOENCODER NETWORK FOR SKELETON-BASED HUMAN ACTION RECOGNITION.” [Online]. Available: https://github.com/Damilytutu/LSTMAE

[8] I. J. Goodfellow *et al.*, “Generative Adversarial Nets.” [Online]. Available: http://www.github.com/goodfeli/adversarial

[9] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. A. Muller, “Deep learning for time series classification: a review,” *Data Min Knowl Discov*, vol. 33, no. 4, pp. 917–963, Jul. 2019, doi: 10.1007/s10618-019-00619-1.

[10] RB Cleveland, WS Cleveland, JE McRae, and I Terpenning, “STL: A seasonal-trend decomposition procedure,” 1990.

[11] P. S. Bradley, K. P. Bennett, and A. Demiriz, “Constrained K-Means Clustering,” 2000.

[12] E. Patel and D. S. Kushwaha, “Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model,” in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 158–167. doi: 10.1016/j.procs.2020.04.017.

[13] W. Jannah and D. R. S. Saputro, “Parameter estimation of Gaussian mixture models (GMM) with expectation maximization (EM) algorithm,” *AIP Conf Proc*, vol. 2566, no. 1, p. 040002, Nov. 2022, doi: 10.1063/5.0117119.

[14] K. R. Shahapure and C. Nicholas, “Cluster Quality Analysis Using Silhouette Score,” in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 2020, pp. 747–748. doi: 10.1109/DSAA49011.2020.00096.

[15] M. Nishida and T. Kawahara, “Unsupervised speaker indexing using speaker model selection based on Bayesian information criterion,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03).*, 2003, pp. I–I. doi: 10.1109/ICASSP.2003.1198744.

[16] Hidenori Watanabe, Shogo Muramatsu, and Hisakazu Kikuchi, *Interval Calculation of EM Algorithm for GMM Parameter Estimation*. IEEE, 2010.

[17] D. Wu, “Communicated by Tomi Kinnunen Parameter Estimation for α-GMM Based on Maximum Likelihood Criterion.”

[18] A. C. C. Coolen, “Markov Chains Compact Lecture Notes and Exercises,” *Department of Mathematics, King’s College London*, pp. 2–40, 2009.

[19] R. M. Blumenthal, “AN EXTENDED MARKOV PROPERTY,” 1957.

[20] E. C. Nwogu *et al.*, “Choice between Mixed and Multiplicative Models in Time Series Decomposition,” *Int J Stat Appl*, vol. 2019, no. 5, pp. 153–159, 2019, doi: 10.5923/j.statistics.20190905.04.

[21] J. P. Kreiss and S. N. Lahiri, “Bootstrap Methods for Time Series,” in *Handbook of Statistics*, vol. 30, Elsevier B.V., 2012, pp. 3–26. doi: 10.1016/B978-0-444-53858-1.00001-6.

[22] Asimakopoulos V and Spiliotis E, “Forecasting Techniques (ECE - NTUA) - Preparation & Time Series Analysis - Lecture 1.” Accessed: Jan. 21, 2024. [Online]. Available: https://www.fsu.gr/el/component/jdownloads/finish/6/1272

[23] A. T. Williams, R. E. Sperl, and S. M. Chung, “Anomaly Detection in Multi-Seasonal Time Series Data,” *IEEE Access*, vol. 11, pp. 106456–106464, 2023, doi: 10.1109/ACCESS.2023.3317791.

[24] M. Mazziotta and A. Pareto, “Normalization methods for spatio-temporal analysis of environmental performance: Revisiting the Min–Max method,” *Environmetrics*, vol. 33, no. 5, Aug. 2022, doi: 10.1002/env.2730.

[25] S. Senthilnathan, “Usefulness of Correlation Analysis,” *SSRN Electronic Journal*, Jul. 2019, doi: 10.2139/ssrn.3416918.

[26] J. Benesty, J. Chen, Y. Huang, and I. Cohen, “Pearson correlation coefficient,” in *Springer Topics in Signal Processing*, vol. 2, Springer Science and Business Media B.V., 2009, pp. 1–4. doi: 10.1007/978-3-642-00296-0_5.

[27] K. G. Jbreskog, “STRUCTURAL ANALYSIS OF COVARIANCE AND CORRELATION MATRICES,” 1978.