ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Ανασκόπηση μεθόδων συμπλήρωσης ελλιπών δεδομένων

Διπλωματική Εργασία

Οικονόμου Κωνσταντίνος

Επιβλέπων: Παναπακίδης Ιωάννης

Σεπτέμβριος 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

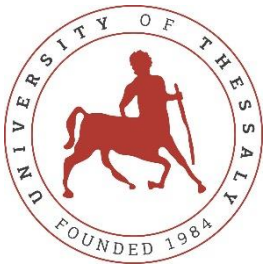
Ανασκόπηση μεθόδων συμπλήρωσης ελλιπών δεδομένων

Διπλωματική Εργασία

Οικονόμου Κωνσταντίνος

Επιβλέπων: Παναπακίδης Ιωάννης

Σεπτέμβριος 2021



UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Review of missing data completion methods

Diploma Thesis

Oikonomou Konstantinos

Supervisor: Panapakidis Ioannis

September 2021

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων	**Παναπακίδης Ιωάννης** Επίκουρος Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας
Μέλος	**Μπαργιώτας Δημήτριος** Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας
Μέλος	**Δασκαλοπούλου Ασπασία** Επίκουρος Καθηγήτρια, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

ΕΥΧΑΡΙΣΤΙΕΣ ή ΣΧΟΛΙΑ

Θέλω να ευχαριστήσω θερμά την οικογένεια μου, που είναι πάντα εκεί για εμένα, στις σπουδές μου, στα όνειρά μου, στη ζωή μου.

Θέλω επίσης να ευχαριστήσω τον επιβλέποντα καθηγητή μου Παναπακίδη Ιωάννη, για τη συνέπειά του και την άριστη συνεργασία μας, από το ξεκίνημα ως την περάτωση της διπλωματικής μου εργασίας.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Δηλώνω επίσης ότι τα αποτελέσματα της εργασίας δεν έχουν χρησιμοποιηθεί για την απόκτηση άλλου πτυχίου. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο Δηλών

Οικονόμου Κωνσταντίνος

Σεπτέμβριος 2021

ΠΕΡΙΛΗΨΗ

Το πρόβλημα των ελλιπών δεδομένων έχει αναγνωριστεί ως ένα σημαντικό ζήτημα, καθώς επηρεάζει την εγκυρότητα δεδομένων και ερευνών. Αφού ο σκοπός κάθε προσπάθειας συλλογής δεδομένων είναι η λήψη πληροφοριών, ένα σύνολο δεδομένων με ελλείψεις θα περιορίσει τα συμπεράσματα μας.

Τα ελλιπή δεδομένα συναντώνται παντού γύρω μας στον πραγματικό και τον ψηφιακό κόσμο, για παράδειγμα ως αποτέλεσμα αναπάντητων ερωτήσεων σε ερεύνες, κακής διαχείρισης αρχείου, λάθος λειτουργία ενός σένσορα ή μια συσκευής μετρήσεων. Η λάνθασμένη διαχείριση των ελλιπών δεδομένων συχνά μπορεί να οδηγήσει σε κακή ερμηνεία των αποτελεσμάτων και λάθος εικόνα σε κάποια μελέτη ανεξάρτητα από το πόσα μη ελλιπή δεδομένα υπάρχουν.

Όσο η συμπλήρωση ελλιπών δεδομένων αποδεικνύεται απαιτητικό πρόβλημα, και μεθοδολογικά αλλά και υπολογιστικά, ο σκοπός της εργασίας αυτής είναι η εξέταση της διαδικασίας συμπλήρωσης κάνοντας χρήση διαφόρων μεθόδων. Οι τεχνικές που εφαρμόζονται είναι τόσο παραδοσιακές μέθοδοι συμπλήρωσης, όπως χρήση μέσου όρου ή των k-κοντινότερων γειτλόνων, όσο και πιο προχωρημένες όπως μέθοδοι ομαδοποίησης και κατανομής συχνοτήτων.

Οι μέθοδοι αυτοί εφαρμόστηκαν σε ημερολογιακά δεδομένα ηλεκτρικού φορτίου για την περιοχή της Νέας Ελβετίας Θεσσαλονίκης για το έτος 2011 και σε ημερολογιακά δεδομένα για την ταχύτητα ανέμου στην πόλη του Βόλου για τα έτη 2018-2020.

Οι μέθοδοι εφαρμόζονται στα δεδομένα αυτά και έπειτα συγκρίνονται μεταξύ τους, όσον αφορά την ακρίβεια πρόβλεψης, το ποσοστό απόκλισης και τον χρόνο εκτέλεσης.

Η υλοποίηση του προγραμματιστικού μέρους της εργασίας έγινε σε Matlab.

ABSTRACT

Missing data problem has been recognized as a major issue, since it affects the validity of datasets and surveys. Since every collection of data aims to collect information, a dataset with missing values can restrict our conclusions and findings.

Missing data can be found anywhere around as, both in the virtual world as well as in real life, for example as a result of non-answered questions in a survey, bad record-keeping, functional faults of a sensor or a measurement device. Poor handling of missing data can lead to misinterpretation of results and bad image of findings of a survey, regardless of how many complete data exist.

As missing data imputations appears to be a demanding task, both methodologically and computationally, the purpose of this thesis is to review a number of missing data imputation methods. The methods that have been examined include simple traditional imputation methods, such as mean imputation or k-nearest neighbours, as well as more advanced methods, such as clustering methods and frequency distribution.

The imputation methods have been applied to the daily data of electrical load for Nea Elvetia in Thessaloniki, Greece, for the year 2011, as well as daily data of wind speed for the city of Volos, Greece, for the years 2018-2020.

The methods are applied to these datasets and are then compared based on prediction accuracy, deviation percentage and execution time.

The programming part of the thesis was implemented in Matlab.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο της εργασίας

Τα σύνολα δεδομένων συχνά χαρακτηρίζονται από ελλειπή δεδομένα, που οφείλονται σε διαφόρους παράγοντες, όπως αναπάντητες ερωτήσεις σε έρευνες, λανθασμένη λειτουργία ενός σένσορα ή μιας μηχανής μετρήσεων. Πολλές φορές που η πληροφορία χρειάζεται να παρθεί και να αναλυθεί real-time, δεν υπάρχει αρκετός χρόνος να διαπιστωθεί ο λόγος για τον οποίο λείπουν τα δεδομένα. Συνεπώς απαιτούνται τεχνικές που να είναι αποτελεσματικές να συμπληρώσουν τα ελλιπή δεδομένα, ανεξάρτητα του λόγου για τον οποίο λείπουν.

Σκοπός της εργασίας αυτής είναι να εξετάσει την αποτελεσμάτικοτητα μεθόδων συμπλήρωσης διαφόρων ειδών σε σύνολα δεδομένων που αφορούν ηλεκτρικό φορτίο και δεδομένα ταχύτητας αέρα και να συγκρίνει τις μεθόδους μεταξύ τους.

Χρησιμοποιούνται δυο σύνολα δεδομένων, με το ένα να περιέχει ημερήσια στοιχεία μετρήσεων ηλεκτρικού φορτίου σε per unit, για την περιοχή Νέα Ελβετία Θεσσαλονίκης του έτους 2011 και το δεύτερο με ημερήσιες μετρήσεις ταχύτητας ανέμου στα 10 μέτρα για την πόλη του Βόλου για τα έτη 2018-2020. Υποθέτοντας για καθένα από αυτά τα σύνολα πως ένα ποσοστό τιμών 10,20 ή 30% είναι ελλιπής, εφαρμόζονται οι μέθοδοι και συγκρίνονται τα δεδομένα με τα αρχικά πλήρη για να εξεταστεί η ακρίβεια και η απόκλιση των μεθόδων.

Δοκιμάζονται διαφόρων ειδών μέθοδοι στα δυο αυτά σύνολα δεδομένων, μέθοδοι με χρήση κλασσικής στατιστικής, μέθοδοι ομαδοποίησης δεδομένων και χρήση κατανομής συχνοτήτων.

1.2 Μηχανισμοί ελλιπών δεδομένων

Η γνώση των μηχανισμών των ελλιπών δεδομένων μπορεί να βοηθήσει στην εξήγηση του γιατί τα δεδομένα λείπουν αλλά και στην επιλογή της κατάλληλης μεθόδου συμπλήρωσης. Οι μηχανισμοί αυτοί μπορούν να εξηγήσουν αν η έλλειψη στοιχείων οφείλεται σε βαθύτερα αίτια ή οχι.

Διακρίνονται λοιπόν οι εξής κατηγορίες μηχανισμών [1]:

- Τυχαία έλλειψη (MAR) (Missing At Random)
- Εντελώς τυχαία έλλειψη (MCAR) (Missing Completely At Random)
- Μη τυχαία έλλειψη (MNAR) (Missing Not At Random)

Έστω Υ το ολοκληρωμένο σύνολο δεδομένων, τότε $\boldsymbol{Y} = \{\boldsymbol{Y_o}, \boldsymbol{Y_m}\}$

Με Y_o συμβολίζεται το σύνολο των τιμών που παρατηρούνται και Y_m το σύνολο των ελλιπών τιμών του Y.

Missing At Random (MAR) [2]:

$$\boldsymbol{P(M|Y_0, Y_m) = P(M|Y_0)}$$

όπου $\boldsymbol{M}$ ο δείκτης έλλειψης που ισούται με 1 αν Y παρατηρήσιμο και με 0 αν Y λείπει.

Δηλαδή είναι ο μηχανισμός στον οποίο τα στοιχεία δε λείπουν εντελώς τυχαία και η έλλειψη στοιχείου δεν εξαρτάται από την ίδια του την αξία, αν δεσμευτεί για μια άλλη μεταβλητή.

Missing Completely At Random (MAR) [2]:

$$\boldsymbol{P(M|Y_0, Y_m) = P(M)}$$

όπου ούτε οι τιμές του Y_m ούτε του Y_o μπορούν να βοηθήσουν στην εκτίμηση των ελλιπών τιμών και για αυτές τις περιπτώσεις προτείνεται μέθοδος διαγραφής των ελλιπών δεδομένων. Τα ελλιπή στοιχεία είναι ανεξάρτητα από κάθε άλλη μεταβλητή.

Missing Not At Random (MNAR) [2]:

$$\boldsymbol{P(M|Y_0, Y_m)}$$

όπου ο μηχανισμός ελλιπών δεδομένων σχετίζεται μετα ίδια τα ελλιπή δεδομένα.

1.3 Επισκόπηση βιβλιογραφίας

Ο χειρισμός των ελλιπών δεδομένων γίνεται κατά βάση με τρεις τρόπους: (α) απαλοιφή των ελλιπών τιμών, (β) συμπλήρωση με χρήση πιθανότερων τιμών και (γ) συμπλήρωση με χρήση μεθόδων και παραγόμενων τιμών.

Η απαλοιφή ελλιπών τιμών αποτελεί βολικό τρόπο αντιμετώπισης για μικρό σύνολο δεδομένων, ωστόσο μειώνεται η ακρίβεια και πιθανώς χάνεται βασική πληροφορία του συνόλου. [3]

Κάποιες απλές μέθοδοι συμπλήρωσης με χρήση παραγόμενων τιμών είναι η συμπλήρωση με χρήση μέσης τιμής, χρήση γραμμικής παρεμβολής ή χρήση επόμενων/προηγουμένων τιμών.

Η συμπλήρωση με χρήση του μέσου δεν εκμεταλλεύεται το πλεονέκτημα των χαρακτηριστικων του χρόνου ή των σχέσεων μεταξυ των μεταβλητών, πρόκειται για πολυ γρήγορη μέθοδο, ωστόσο μειώνει τη διασπορά του συνόλου δεδομένων. [3]

Η συμπλήρωση με χρήση γραμμικής παρεμβολής ειναι αποτελεσματική για χρονοσειρές αλλά όχι ιδιαίτερα για εποχικά δεδομένα, που έχουν αποκλίσεις μεταξύ τους. [3]

Η συμπλήρωση με προηγούμενες ή επόμενες τιμές είναι ευρέως χρησιμοποιημένη και αποτελεσματική για δεδομένα επαναλαμβανόμενα, όμως μπορεί να έχουν μεγάλη μεροληψία. [3]

Οι μέθοδοι που έχουν εφαρμοστεί πρόσφατα για την συμπλήρωση ελλιπών δεδομένων είναι συμπλήρωση με αλγόριθμο κ-κοντινότερων γειτόνων (KNN) [4,5], συμπλήρωση με χρήση του μέσου όρου [4,5], ασαφής ομαδοποίηση (fuzzy clustering) [6], μέθοδοι ομαδοποίησης σε σύνολο δεδομένων μειωμένων διαστάσεων [7, 8] και χρήση νευρώνων Perceptron [9].

Η μέθοδος των κ-κοντινότερων γειτόνων (KNN) εφαρμόστηκε σε ιατρικά δεδομένα (δεδομένα για ηπατίτιδα, διαβήτη και άλλα) [4], στα οποία εφαρμόστηκαν επίσης οι μέθοδοι του μέσου όρου και της απαλοιφής ελλιπών τιμών. Αποδείχτηκε πως η χρήση της μεθόδου κ-κοντινότερων γειτόνων έχει καλύτερη απόδοση συγκριτικά με τις άλλες μεθόδους, ειδικά όταν το ποσοστό ελλιπών τιμών αυξάνεται. Παρόλα αυτά έχει αρκετό υπολογιστικό κόστος.

Οι παραπάνω μέθοδοι εφαρμόστηκαν επίσης σε βάση δεδομένων μαθητών, όπου επίσης η ακρίβεια της KNN μεθόδου ήταν η καλύτερη. [5]

Η μέθοδος της ασαφούς ομαδοποίησης εφαρμόστηκε σε χρονοσειρές δεδομένων καρκινοπαθών ασθενών [6] και αποδείχτηκε να έχει καλύτερη απόδοση απο παρόμοιες μεθόδους.

Στην έρευνα [7] εφαρμόστηκε η μέθοδος ομαδοποίησης του συνόλου δεδομένων και η μείωση των διαστάσεών του. Τα ευρήματα της έρευνας ήταν πως η μέθοδος είναι αρκετά ικανή να συμπληρώσει αριθμητικές τιμές ιατρικών δεδομένων.

Παρόμοια μέθοδος ομαδοποίησης και μείωσης διαστάσεων πραγματοποιήθηκε και στην έρευνα [8], πάλι σε ιατρικά δεδομένα, μέθοδος που χρησιμοποιήθηκε αποτελεσματικά για συμπλήρωση αλλά και πρόληψη ασθενειών.

Στην έρευνα [9] εφαρμόστηκε τεχνική νευρωνικού δικτύου με νευρώνες Perceptron σε δεδομένα κίνησης οδικού δικτύου, που αποδείχτηκαν να έχουν καλύτερη απόδοση από δίκτυα βαθιάς μάθησης (Deep Learning Networks (DLNs))

1.4 Σκοπός της εργασίας

Καθώς οι τιμές ηλεκτρικού φορτίου και ταχύτητας ανέμου, εμφανίζουν μη γραμμικότητα και πολυπλοκότητα, οι αλγόριθμοι αναγνώρισης προτύπων μπορούν να προταθούν δικαίως ως μεθοδολογία για αυτών των ειδών τα δεδομένα. Οι αλγόριθμοι ομαδοποίησης μπορούν να ομαδοποιήσουν το σύνολο δεδομένων σε καλώς ταξινομημένες κλάσεις, καθώς και να μειώσουν τις διαστάσεις του αρχικού συνόλου δεδομένων. Η μέθοδος ομαδοποίησης έχει εφαρμοστεί με επιτυχία σε πολλούς αλγορίθμους αναγνώρισης προτύπων. [10]

Με βάση την παραπάνω επισκόπηση βιβλιογραφίας, είναι εμφανές πως το αντικείμενο της συμπλήρωσης ελλιπών δεδομένων χρειάζεται περαιτέρω έρευνα. Πολλές από τις ήδη εξεταζόμενες μεθόδους δεν έχουν εφαρμοστεί σε σύνολα δεδομένων μεγάλων διαστάσεων, ενώ αρκετές έχουν αποδειχθεί να έχουν υψηλό υπολογιστικό κόστος και αναζητώνται τρόποι βελτιστοποίησης. Η βιβλιογραφία για την αντιμετώπιση ελλιπών δεδομένων είναι εστιασμένη σε ιατρικά δεδομένα, και χρήση μεθόδων για πρόληψη ασθενειών, χωρίς να γίνεται λόγος για ανανεώσιμες πηγές ενέργειας ή ηλεκτρικό φορτίο.

Σκοπός της εργασίας αυτής είναι η πρόταση μεθόδων για συμπλήρωση ελλιπών δεδομένων ηλεκτρικού φορτίου και ταχύτητας ανέμου, έπειτα από υλοποίηση και εξέταση τους. Οι μέθοδοι εφαρμόζονται σε δυο σύνολα δεδομένων, με το ένα να περιέχει πραγματικά δεδομένα ημερήσιων μετρήσεων ηλεκτρικού φορτίου και το δεύτερο μετρήσεις ταχύτητας ανέμου.

Η εργασία επικεντρώνεται στην εφαρμογή των μεθόδων για συμπλήρωση ελλιπών τιμών ώστε να εκτιμώνται και να υπάρχει πλήρη εικόνα των δεδομένων για περαιτέρω εκτιμήσεις.

Η χρήση των μεθόδων ομαδοποίησης δεν έχει εξεταστεί στη βιβλιογραφία αρκετά για δεδομένα φορτίου και ανέμου, καθώς η χρήση τους είναι επικεντρωμένη σε ιατρικά δεδομένα. Στην παρούσα εργασία υλοποιούνται και εξετάζονται 5 τέτοιες μέθοδοι, καθώς επίσης και μέθοδοι με χρήση πιθανοτήτων μέσω κατανομής συχνοτήτων.

1.5 Δεδομένα και μεθοδολογία της εργασίας

Τα δεδομένα που χρησιμοποιούνται στην παρούσα εργασία, αφορούν μετρήσεις ηλεκτρικού φορτίου και μετρήσεις ταχύτητας ανέμου. Το πρώτο σύνολο δεδομένων, αποτελείται απο ημερήσιες μετρήσεις για κάθε ώρα, με τις τιμές του peakηλεκτρικού φορτίου σε per unit. Περιλαμβάνει δηλαδή 8760 μετρήσεις. Το δεύτερο σύνολο δεδομένων αποτελείται απο ημερήσιες μετρήσεις της ταχύτητας ανέμου στα 10 μέτρα, για 2 χρόνια, περιλαμβάνει δηλαδή 731 τιμές. Το σύνολο δεδομένων με τις ελλιπείς τιμές προκύπτει από τα παραπάνω με τυχαίες ελλιπείς τιμές σε ποσοστό 10-20-30 % των τιμών του συνόλου.

Η μεθοδολογία που ακολουθείται λοιπόν είναι η εξής:

- Φόρτωση του συνόλου δεδομένων στο προγραμματιστικό περιβάλλον του Matlab
- Τυχαία ανάθεση τιμών σε ελλιπείς σε ποσοστό 10-20-30 % του συνόλου τιμών
- Εφαρμογή της μεθόδου συμπλήρωσης ελλιπών τιμών, με σάρωση του συνόλου και εφαρμογή της αντίστοιχης μεθοδολογίας, σε περίπτωση εύρεσης ελλιπούς τιμής
- Αποθήκευση του συμπληρωμένου συνόλου δεδομένων
- Αξιολόγηση της απόδοσης και της ακρίβειας της μεθόδου με δείκτες και σύγκριση του αρχικού συνόλου με το συμπληρωμένο

1.6 Δομή της εργασίας

Στο Κεφάλαιο 2 γίνεται αρχικά μια εισαγωγή στις μεθόδους και τα σύνολα δεδομένων που χρησιμοποιούνται στης εργασία. Πραγματοποιείται μια γενική περιγραφή της βασικής μεθοδολογίας που ακολουθείται κατά βάση σε όλες τις μεθόδους που

εξετάστηκαν. Έπειτα γίνεται ταξινόμηση των μεθόδων σε 3 κατηγορίες: μέθοδοι με χρήση κλασσικής στατιστικής, μέθοδοι ομαδοποίησης δεδομένων και μέθοδοι κατανομής συχνοτήτων με χρήση πιθανοτήτων. Ακολουθεί αναλυτική περιγραφή μαθηματικά και μεθοδολογικά της κάθε διαδικασίας. Για κάθε μια μέθοδο, παρατίθενται οι αναγκαίες εξισώσεις στις οποίες και βασίστηκαν οι προγραμματιστικές υλοποιήσεις τους στο Matlab, ένα σενάριο ελλιπής τιμής και περιγραφή διαδικασίας συμπλήρωσης καθώς και διάγραμμα ροής με την παρουσίαση της κάθε μεθόδου.

Στο Κεφάλαιο 3 γίνεται παρουσίαση των αποτελεσμάτων της κάθε μεθόδου και σύγκρισή τους. Οι μέθοδοι εφαρμόζονται και στα δυο σύνολα δεδομένων που μελετώνται, στα σενάρια όπου οι ελλιπείς τιμές αποτελούν το 10,20 και 30% των τιμών του συνόλου.

Στο Κεφάλαιο 4 γίνεται η εξαγωγή των συμπερασμάτων και η παρουσίαση τους, με βάση τους πίνακες αποτελεσμάτων του Κεφαλαίου 3. Προτείνονται επίσης κάποιες ιδέες για μελλοντική έρευνα επάνω στο αντικείμενο.

ΚΕΦΑΛΑΙΟ 2

ΜΑΘΗΜΑΤΙΚΟ ΥΠΟΒΑΘΡΟ ΜΕΘΟΔΩΝ ΣΥΜΠΛΗΡΩΣΗΣ ΕΛΛΙΠΩΝ ΔΕΔΟΜΕΝΩΝ

2.1 Περιγραφή μεθοδολογίας

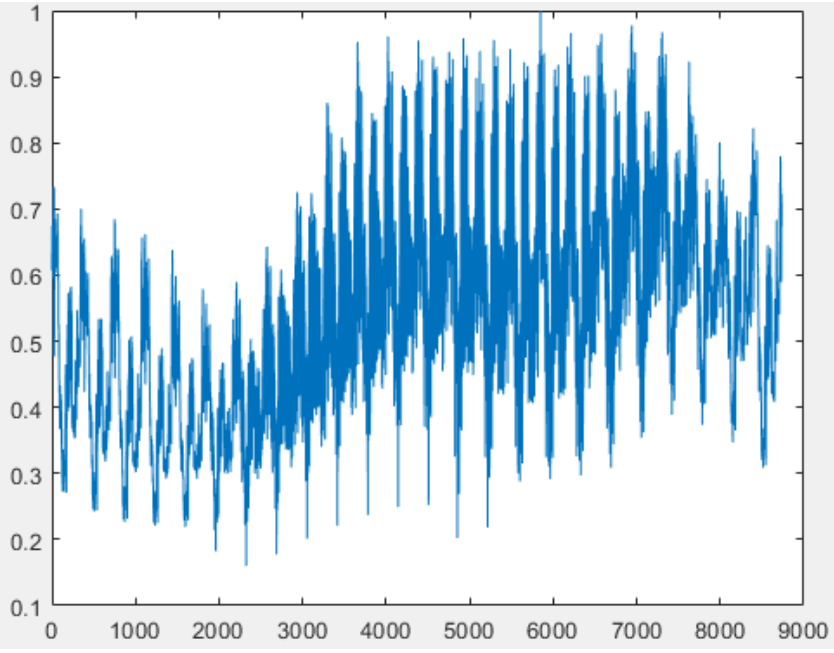
Στην παρούσα εργασία χρησιμοποιούνται δυο διαφορετικά σύνολα δεδομένων. Το πρώτο περιλαμβάνει μετρήσεις του ηλεκτρικού φορτίου στην περιοχή Νέα Ελβετία Θεσσαλονίκης για το έτος 2011, με μετρήσεις ανά ημέρα και ώρα, συνεπώς οι διαστάσεις του συνόλου δεδομένων είναι 365x24 = 8760 τιμές. Το δεύτερο σύνολο δεδομένων που χρησιμοποιείται περιλαμβάνει ημερήσιες μετρήσεις της ταχύτητας ανέμου για την περιοχή του Βόλου στα 10 μέτρα για την περίοδο δυο ετών 2018 έως 2020, συνεπώς οι διαστάσεις του συνόλου δεδομένων είναι 731x1 = 731 τιμές.

Εκφράζουμε συμβολικά αρχικά το κάθε σύνολο δεδομένων ως εξής:

Έστω

$$\boldsymbol{S_i = \{x_1, x_2, \dots, x_n\}, i \in [1, 24], n = 365}$$

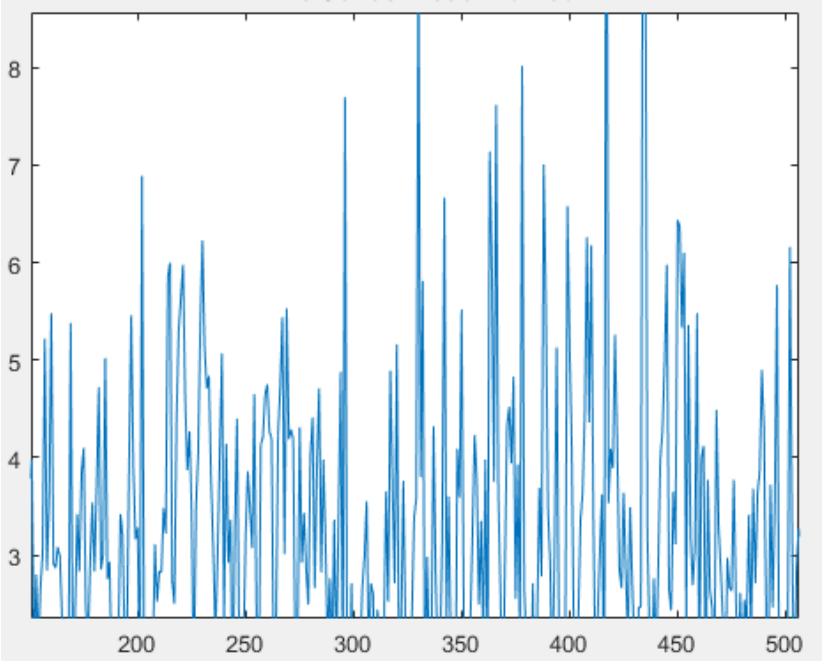
το πρώτο σύνολο δεδομένων όπου S_i η χρονοσειρά κάθε ώρας μεγέθους 365 τιμών για κάθε ημέρα, όπως παρουσιάζεται και στο Σχήμα 2.1



Σχήμα 2.1: Σχήμα χρονοσειράς φορτίου Ν. Ελβετίας σε μονάδες perunit (2011)

$$\boldsymbol{Z_i = \{x_1, x_2, \ldots, x_n\}, n = 731}$$

το δεύτερο σύνολο δεδομένων όπου Z_i η κάθε μέτρηση ταχύτητας ανέμου για κάθε μια ημέρα από τις 731 των δυο ετών, όπως παρουσιάζεται και στο Σχήμα 2.2.



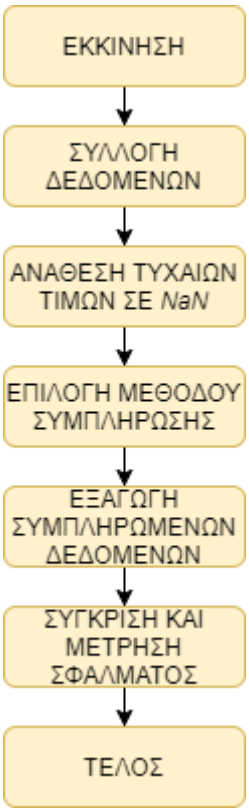
Σχήμα 2.2: Σχήμα χρονοσειράςταχύτητας ανέμου στα 10mγια την περιοχή του Βόλου (2018-2020)

Για την εξέταση των μεθόδων, θέτονται τυχαίες τιμές σε *NaN*, θεωρούνται δηλαδή ελλιπείς, με ποσοστό ελλιπών τιμών 10-20-30% των τιμών σε κάθε σύνολο δεδομένων.

Έπειτα με χρήση μεθόδων συμπλήρωσης ελλιπών δεδομένων στο προγραμματιστικό περιβάλλον του Matlab, γίνεται η εκτίμηση των *NaN* τιμών, κάνοντας χρήση των γνωστών τιμών. Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.3.

Στην παρούσα εργασία γίνεται χρήση τριών ειδών μεθόδων για τη συμπλήρωση των ελλιπών δεδομένων:

- Κλασσικές μέθοδοι στατιστικής
- Μέθοδοι ομαδοποίησης δεδομένων (clustering)
- Μέθοδοι με χρήση πιθανοτήτων



Σχήμα 2.3: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης συνοπτικά.

2.2 Κλασσικές μέθοδοι στατιστικής

2.2.1. Μέθοδος συμπλήρωσης με χρήση της μέσης τιμής (mean) στήλης

Στη μέθοδο αυτή κάθε μια ελλιπή τιμή συμπληρώνεται με τον μέσο όρο (mean value) των μη ελλιπών τιμών της στήλης. Γίνεται σάρωση του συνόλου δεδομένων και όταν βρεθεί ελλιπής τιμή, εξάγεται η μέση τιμή της στήλης με την οποία και γίνεται η συμπλήρωση.

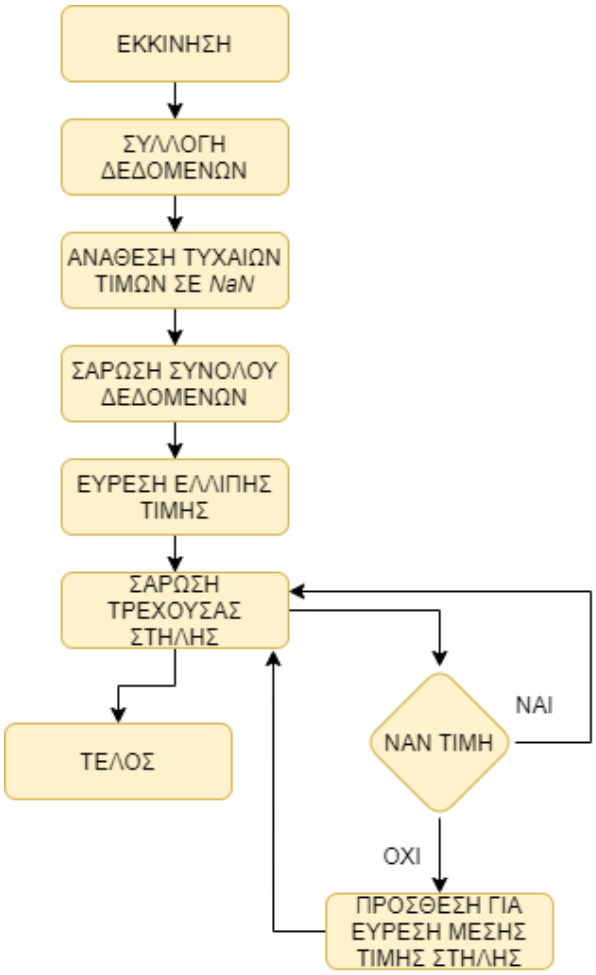
Έστω πως σαρώνοντας το σύνολο δεδομένων διαπιστώνεται πως λείπει η τιμή:

$$\boldsymbol{x(i,j)}$$

τότε συμπληρώνεται με την τιμή:

$$\boldsymbol{x(i,j) = \frac{1}{n}\sum_{k=1}^{n} x(k,j)\,, n = length}$$

αν η τιμή $\boldsymbol{x(k,j)}$ ειναι διάφορη του NaN, δεν είναι δηλαδή ελλιπής. Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.4.



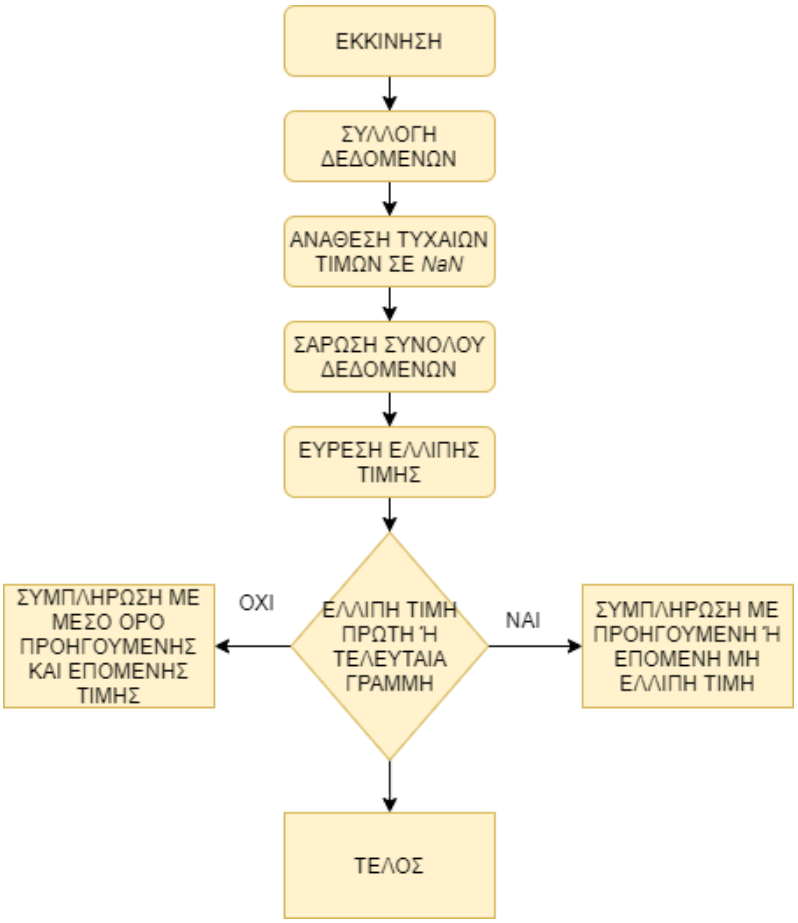
Σχήμα 2.4: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήση μέσης τιμής στήλης.

2.2.2. Μέθοδος συμπλήρωσης με χρήση της μέσης τιμής (mean) προηγούμενης και επόμενης τιμής

Στη μέθοδο αυτή οι ελλιπείς τιμές συμπληρώνονται με τον μέσο όρο της προηγούμενης και της επόμενης τιμής της στήλης. Γίνεται σάρωση του συνόλου δεδομένων και όταν βρεθεί ελλιπής τιμή συμπληρώνεται με τον μέσο όρο της προηγούμενης και της επόμενης μη ελλιπούς τιμής. Σε περίπτωση που η ελλιπής τιμής βρίσκεται στην πρώτη γραμμή του συνόλου δεδομένων ή την τελευταία, η συμπλήρωση γίνεται με την επόμενη και την τελευταία μη ελλιπή τιμή αντίστοιχα. Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.5.

Έστω πως σαρώνοντας το σύνολο δεδομένων διαπιστώνεται πως λείπει η τιμή $\boldsymbol{x(i,j)}$ τότε συμπληρώνεται με την παραγόμενη τιμή:

$$\boldsymbol{x(i,j) = \frac{1}{2}[x(i-1,j) + x(i+1,j)]}$$



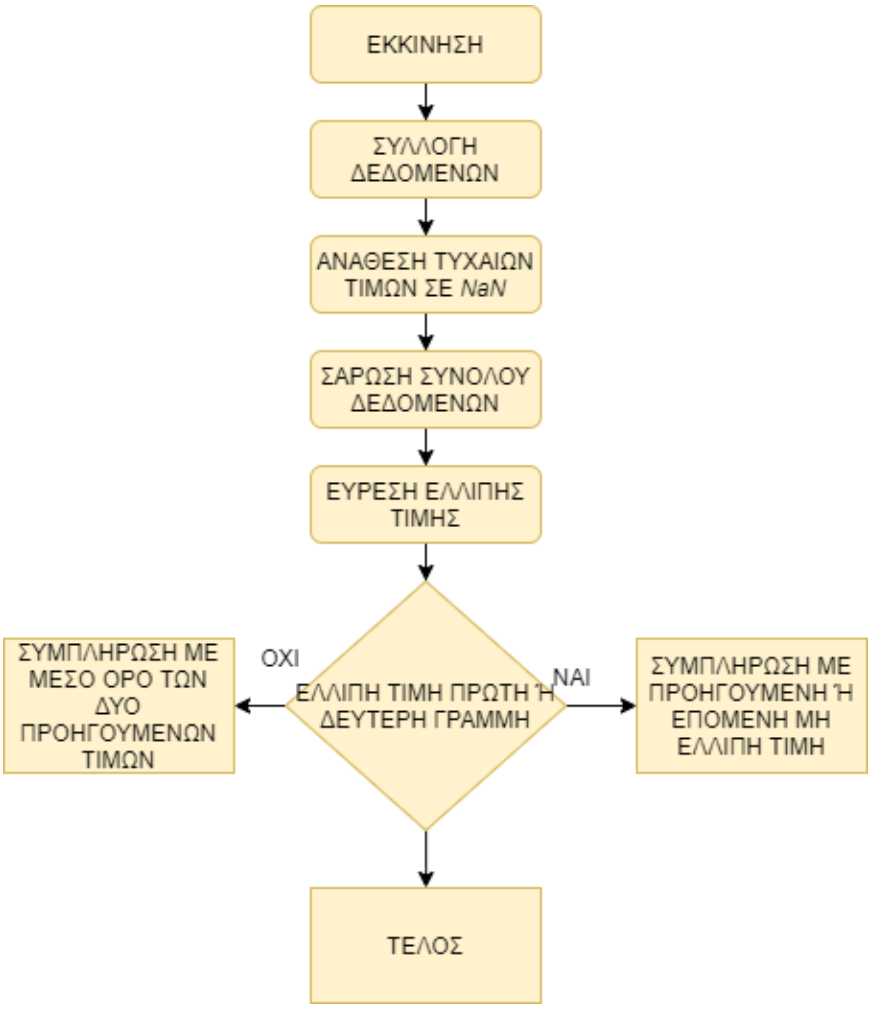
Σχήμα 2.5: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήση μέσης τιμής προηγούμενης και επόμενης τιμής.

2.2.3. Μέθοδος συμπλήρωσης με χρήση της μέσης τιμής (mean) των δυο προηγουμένων τιμών.

Στη μέθοδο αυτή η ελλιπής τιμή συμπληρώνεται με τον μέσο όρο δυο προηγουμένων μη ελλιπών τιμών. Σε περίπτωση που η ελλιπής τιμής βρίσκεται στη δεύτερη γραμμή του συνόλου δεδομένων η συμπλήρωση γίνεται με την προηγούμενη τιμή, ενώ αν βρίσκεται στην πρώτη γραμμή, η συμπλήρωση γίνεται με την πρώτη μη ελλιπή τιμή της στήλης.Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.6.

Έστω πως σαρώνοντας το σύνολο δεδομένων διαπιστώνεται πως λείπει η τιμή $\boldsymbol{x(i,j)}$, τότε αυτή συμπληρώνεται με την παραγόμενη τιμή:

$$\boldsymbol{x(i,j) = \frac{1}{2}[x(i-1,j) + x(i-2,j)]}$$



Σχήμα 2.6: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήση μέσης τιμής των δυο προηγούμενων τιμών.

2.2.4. Μέθοδος συμπλήρωσης με χρήση της ελάχιστης Ευκλείδειας απόστασης.

Στην μέθοδο αυτή, όταν βρίσκεται κατά την σάρωση των δεδομένων μια ελλιπής τιμή, βρίσκεται η Ευκλείδεια απόσταση της προηγούμενης τιμής με όλες τις υπόλοιπες τιμές της στήλης. Η τιμή στα δεδομένα που έχει την ελάχιστη απόσταση αποθηκεύεται και η ελλιπής τιμή συπληρώνεται με την επόμενη μη ελλιπή τιμή αυτής.

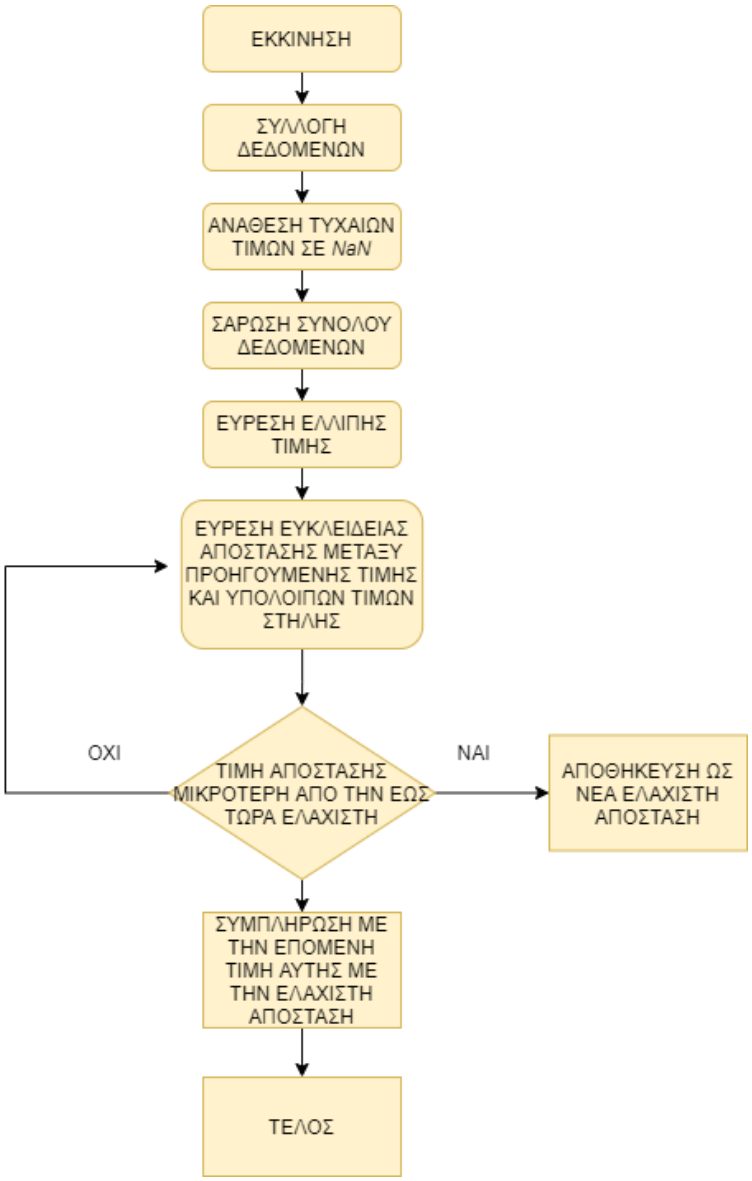
Οι τιμές των Ευκλείδειων αποστάσεων υπολογίζονται ως:

$$\boldsymbol{\delta_1} = \sum_{i=1}^{n} |\boldsymbol{x_i} - \boldsymbol{y_i}|$$

Έστω πως σαρώνοντας το σύνολο δεδομένων διαπιστώνεται πως λείπει η τιμή $\boldsymbol{x(i,j)}$, τότε αυτή συμπληρώνεται με την παραγόμενη τιμή:

$$\boldsymbol{x(i,j) = x(k+1,j)}$$

Η τιμή $x(k,j)$ είναι η τιμή με την ελάχιστη Ευκλείδεια απόσταση. Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.7.



Σχήμα 2.7: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήση ελάχιστης Ευκλείδειας απόστασης.

2.2.5. Μέθοδος συμπλήρωσης με χρήση επιλεγόμενης απόστασης προηγούμενης τιμής.

Στην μέθοδο αυτή, ο χρήστης επιλέγει κατά πόσες θέσεις πίσω θέλει να βρίσκεται η τιμή με την οποία θα συμπληρωθεί κάθε ελλιπής τιμή. Πρόκειται για μια πρακτική μέθοδο ειδικά όσον αφορά τα σύνολα δεδομένων που μελετώνται στην παρούσα εργασία, μιας και πρόκειται για δεδομένα ημερήσια. Έτσι μπορεί να επιλεχθεί για παράδειγμα η

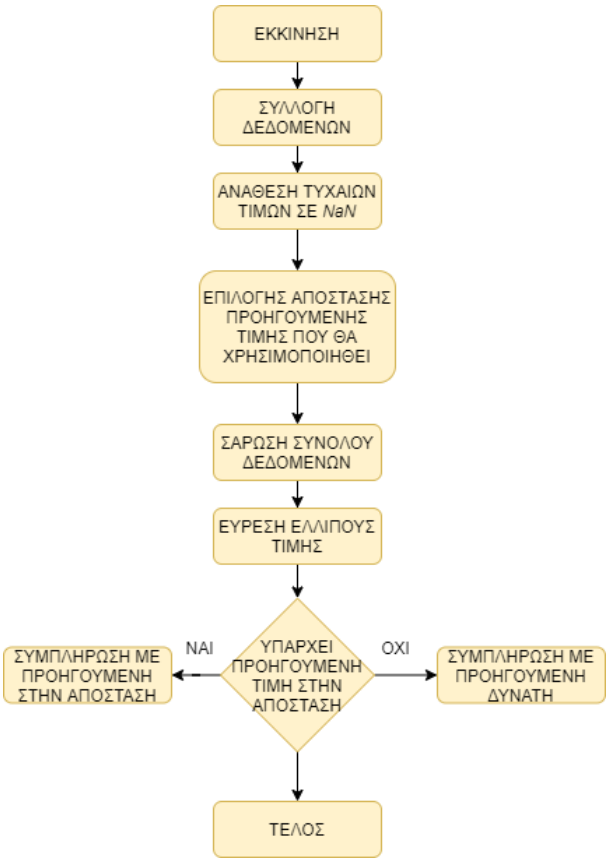
συμπλήρωση με την τιμή της ίδιας μέρας προηγούμενης εβδομάδας με επιλεγόμενη απόσταση 7. Αν δεν υπάρχουν οι ζητούμενες προηγούμενες θέσεις στην στήλη του συνόλου δεδομένων μας η ελλιπής τιμή συμπληρώνεται με την αμέσως προηγούμενη δυνατή μη ελλιπή τιμή. Υποκατηγορία της μεθόδου αυτής αποτελεί μια ευρέως διαδεδομένη και πρακτική μέθοδος, γνωστή ως Last Observation Carried Forward (LOCF) μέθοδος [11], με την οποία η ελλιπής τιμή συμπληρώνεται με την αμέσως προηγούμενη μη ελλιπή τιμή δεδομένων. Έτσι, η ελλιπής τιμή $x(i,j)$ με επιλεγόμενη απόσταση kγια προηγούμενη τιμή, συμπληρώνεται ως:

$$\boldsymbol{x(i,j) = x(i-k,j)}$$

Ενώ για $k = 1$ θεωρείται η περίπτωση της LOCF μεθόδου, δηλ. :

$$\boldsymbol{x(i,j) = x(i-1,j)}$$

Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.8.



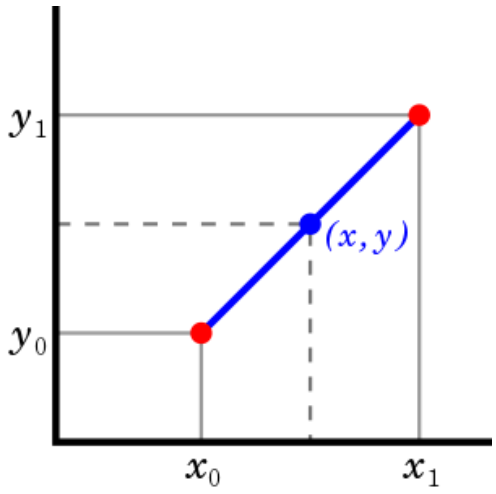
Σχήμα 2.8: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήση επιλεγόμενης απόστασης προηγούμενης τιμής.

2.2.6. Μέθοδος συμπλήρωσης με χρήση γραμμικης παρεμβολής με επιλεγόμενο window.

Στη μέθοδο αυτή, αφού γίνει σάρωση του συνόλου δεδομένων και βρεθούν οι ελλιπείς τιμές, κάθε ελλιπής τιμή συμπληρώνεται μεχρήση γραμμικής παρεμβολής λαμβάνοντας υπόψιν επόμενες και προηγούμενες τιμές με window τιμών 1,2 ή 3.

Η γραμμική παρεμβολή είναι μια διαδικασία που επιτρέπει να συναχθεί μια τιμή μεταξύ καλά καθορισμένων τιμών (Σχήμα 2.9), οι οποίες μπορούν να είναι σε έναν πίνακα ή σε ένα γραμμικό γράφημα (Σχήμα 2.10).

Αποτελεί μια μέθοδο που χρησιμοποιείται ευρέως για προβλήματα ελλιπών δεδομένων, καθώς δημιουργεί νέα σημεία μεταξύ του εύρους των επιμέρους υπαρχόντων δεδομένων.



Σχήμα 2.9: Γραφική απεικόνιση γραμμικής παρεμβολής στο διάστημα (x_0, x_1)[12]

Στηρίζεται στην εξής λογική:

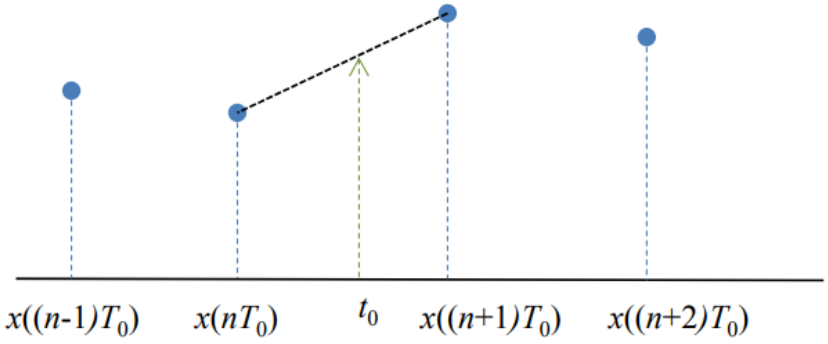
Ας υποθέσουμε ότι έχουμε γνωστές συντεταγμένες (x_0, y_0) και (x_1, y_1):

$$\frac{\boldsymbol{y - y_0}}{\boldsymbol{x - x_0}} = \frac{\boldsymbol{y_1 - y_0}}{\boldsymbol{x_1 - x_0}}$$

Επιλύοντας την εξίσωση για y, καταλήγουμε σε:

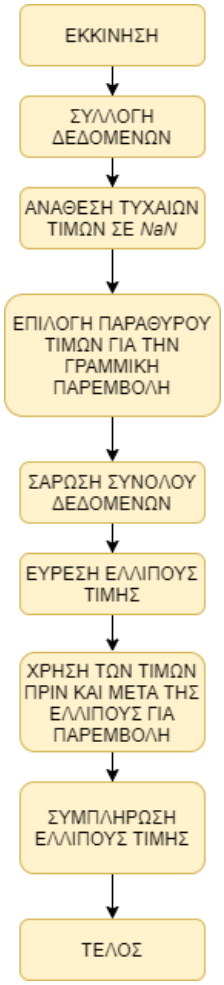
$$\boldsymbol{y = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0} = \frac{y_0(x_1 - x) + y_1(x - x_0)}{x_1 - x_0}}$$

που είναι η φόρμουλα για την γραμμική παρεμβολή στο διάστημα (x_0,x_1) [12].



Σχήμα 2.10: Γραφική απεικόνιση γραμμικής παρεμβολής διακριτής χρονοσειράς

Συνεπώς στην μέθοδο αυτή, αφού επιλέξω το παράθυρο τιμών που θα παρεμβάλω για να βγει η τιμή με την οποία θα συμπληρωθεί η ελλιπής τιμή, γίνεται η γραμμική παρεμβολή. Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.11.



Σχήμα 2.11: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήση γραμμικής παρεμβολής με επιλεγόμενο παράθυρο τιμών.

2.2.7. Μέθοδος συμπλήρωσης με χρήση μέσης τιμής k-κοντινότερων γειτόνων τιμών

Στη μέθοδο αυτή, αφού γίνει σάρωση του συνόλου δεδομένων και βρεθεί ελλιπής τιμή, έχοντας θέσει εξαρχής τον αριθμό του k, η ελλιπής τιμή συμπληρώνεται με τον σταθμισμένο μέσο όρο των γειτονικών kτιμών.

Ο KNN είναι ένας αλγόριθμος που χρησιμοποιείται συχνά για το ταίριασμα ενός σημείου με τους k κοντινότερους γείτονές του σε πολυδιάστατο χώρο. Η υπόθεση και γενικότερη ιδέα γύρω από τον αλγόριθμο είναι πως κάθε τιμή μπορεί να εκτιμηθεί με βάση τα κοντινότερα σημεία σε αυτή σε εξάρτηση με άλλες μεταβλητές.

Για την εύρεση και σύγκριση των γειτόνων χρησιμοποιείται ο τύπος της Ευκλείδειας απόστασης, που είναι και ο συνηθέστερος τρόπος εύρεσης ομοίων δεδομένων:

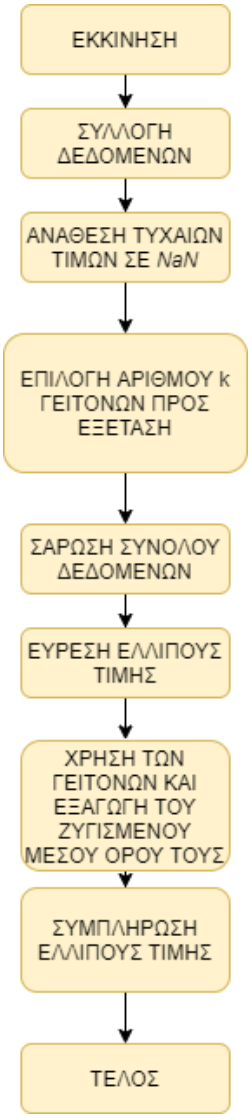
$$\boldsymbol{\delta = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}}$$

Ο αριθμός των γειτόνων προς εξέταση ζητείται στην αρχή της μεθόδου. Προτιμάται μικρός αριθμός για να μη ξεφεύγει η περιοχή εξέτασης στο σύνολο δεδομένων.

Σχήμα 2.12: Παραδείγματα 1ου-2ου-3ου κοντινότερου γείτονα για ένα testδεδομένο [13]

Στο παραπάνω παράδειγμα του Σχήματος 2.12, έχοντας δυο κλάσεις A και B, το τεστ δεδομένο που είναι συμβολισμένο με ερωτηματικό (?) πρόκειται να κατηγοριοποιηθεί. Δημιουργείται μια λίστα με τους k-κοντινότερους γείτονές του. Λαμβάνοντας υπόψιν το σύνολο δεδομένων ξεκινά υπολογίζοντας τις αποστάσεις μεταξύ του τεστ δεδομένου και όλων των υπολοίπων στο σύνολο δεδομένων. Έπειτα επιλέγει το σετ συντεταγμένων (x, y) που έχει την ελάχιστη απόσταση, όπου για 1-κοντινότερο γείτονα είναι το σύνολο

{A}, για 2-κοντινότερους γείτονες είναι το σύνολο {A,B} ενώ για $k = 3$ είναι το {A,A,B}. [13]. Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.13.



Σχήμα 2.13: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήση μέσης τιμής των k-κοντινότερων γειτόνων

2.3 Μέθοδοι ομαδοποίησης δεδομένων (clustering)

2.3.1. Εισαγωγή στην έννοια του clustering

Πολλές φορές σε περιπτώσεις συνεχούς μεταβλητής, η αναλυτική αναφορά όλων των τιμών δεν εξυπηρετεί για την παρουσίαση των δεδομένων, ούτε για την εξαγωγή συμπερασμάτων. Στην περίπτωση μας δεν εξυπηρετεί και στη συμπλήρωση ελλιπών τιμών με βάση τις γνωστές τιμές.

Για να επεξεργαστούμε καλύτερα στατιστικά δεδομένα ομαδοποιούμε τις τιμές σε διαστήματα (κλάσεις) μη επικαλυπτόμενα μεταξύ τους.

Μια κλάση είναι ένα κατάλληλα επιλεγμένο διάστημα του συνόλου τιμών μιας μεταβλητής.

Υπάρχουν δυο γενικότερα είδη ομαδοποίησης [14]:

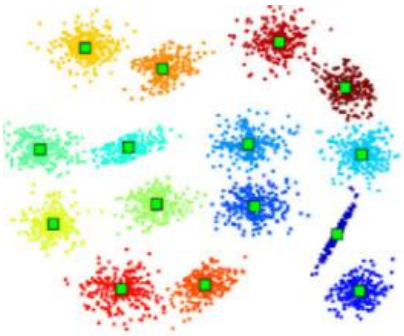
- Hard Clustering: στο οποίο κάθε τιμή ανήκει εξ ολοκλήρου σε μια κλάση ή οχι
- Soft Clustering: στην οποία αντί να τοποθετείται αυστηρά μια τιμή σε κάποια κλάση, τοποθετείται η πιθανότητα με την οποία πιθανώς να ανήκει σε κάποια κλάση

2.3.2. Γενική περιγραφή της μεθοδολογίας ομαδοποίησης

Η ομαδοποίηση των δεδομένων περιλαμβάνει τα παρακάτω βήματα:

Πρώτα γίνεται επιλογή του πλήθους των κλάσεων k, το οποίο ορίζεται αυθαίρετα, ανάλογα και το μέγεθος του δείγματος (Σχήμα 2.14).

Έπειτα προσδιορίζεται το πλάτος των κλάσεων και πραγματοποιείται η κατασκευή των κλάσεων.



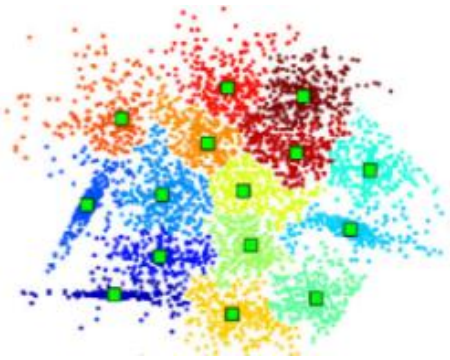
Σχήμα 2.14: Σχήμα αναπαράστασης δεδομένων πριν την εφαρμογή ομαδοποίησης [15]

Στην παρούσα εργασία θα ασχοληθούμε με ομαδοποίηση με βάση το κέντρο δεδομένων (centroid-based clustering) και ομαδοποίηση με βάση τη συνεκτικότητα δεδομένων (connectivity-based clustering).

Στην ομαδοποίηση με βάση τα κέντρα (centroid-based clustering)οι κλάσεις περιγράφονται από ένα διάνυσμα κέντρου, το οποίο μπορεί και να μην είναι μέρος του συνόλου δεδομένων. Όταν ο αριθμός των κλάσεων τεθεί σε k, η k-means ομαδοποίηση ορίζει ένα πρόβλημα βελτιστοποίησης: να βρεθεί ο αριθμός k κέντρων κλάσεων και να

γίνει ανάθεση των δεδομένων στο κοντινότερο κέντρο κλάσης, έτσι ώστε οι αποστάσεις από την κλάση να ελαχιστοποιηθούν (Σχήμα 2.15).

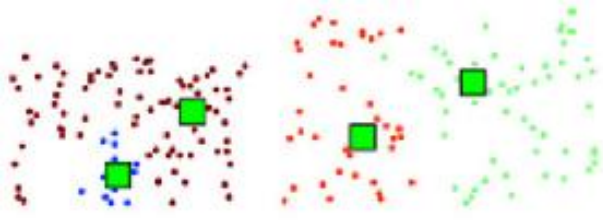
Παραλλαγές του k-means συνήθως περιλαμβάνουν βελτιστοποιήσεις, όπως επιλογή της καλύτερης από πολλές επαναλήψεις ή οριοθέτηση των κέντρων δεδομένων (k-medoids) ή ασαφή ανάθεση κλάσεων (fuzzy c-means).



Σχήμα 2.15: Σχήμα αναπαράστασης δεδομένων ως αποτέλεσμα εφαρμογής ομαδοποίησης[15]

Στην ομαδοποίηση με βάση τη συνεκτικότητα των δεδομένων (connectivity-based clustering), επίσης γνωστή ως hierarchical clustering, η βασική ιδέα είναι να ομαδοποιούνται δεδομένα με άλλα κοντινότερα δεδομένα και όχι σε μεγάλη απόσταση. Η ομαδοποίηση λοιπόν γίνεται με βάση την απόσταση. Η κλάση μπορεί να χαρακτηριστεί από την μέγιστη απόσταση που διανύεται για να συνδεθούν μέλη της κλάσης.

Οι μέθοδοι ομαδοποίησης διαφέρουν ως προς τον τρόπο υπολογισμού των αποστάσεων, όπου εκτός από κλασσική Ευκλείδεια απόσταση μπορεί να χρησιμοποιηθεί ζυγισμένος αριθμητικός μέσος όρος και άλλοι τρόποι μέτρησης.



Σχήμα 2.16: Σχήμα αναπαράστασης δεδομένων ως αποτέλεσμα εφαρμογής partitioning, ώστε να γίνει ομαδοποίηση των δεδομένων.[15]

2.3.3. Μέθοδος συμπλήρωσης με χρήση k-means αλγορίθμου

Η μέθοδος k-means στοχεύει στην ομαδοποίηση των δεδομένων σε κκλάσεις στις οποίες κάθε δεδομένο ανήκει στην κλάση με το κοντινότερο κέντρο (cluster centers).

Έστω ένα σύνολο δεδομένων $(x_1, x_2, \dots x_n)$ όπου το κάθε δεδομένο είναι ένα διάνυσμα μήκους d, ο k-means στοχεύει στο να ομαδοποιήσει τα δεδομένα σε k σύνολα $S = (S_1, S_2, \dots S_k)$ έτσι ώστε να ελαχιστοποιηθεί η απόσταση μεταξύ δεδομένων κλάσης και μέσων [16].

$$\boldsymbol{arg_s min} \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \sum_{\boldsymbol{x \in S_i}} \|\boldsymbol{x - \mu_i}\| = \boldsymbol{arg_s min} \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} |\boldsymbol{S_i}| \, \boldsymbol{Var S_i}$$

όπου μ_i το μέσο των σημείων του S_i. Αυτό είναι ανάλογο με την ελαχιστοποίηση των αποκλίσεων ανά ζεύγη των σημείων της ίδιας κλάσης.

$$\boldsymbol{arg_s min} \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \frac{\mathbf{1}}{\boldsymbol{2|S_i|}} \sum_{\boldsymbol{x,y \in S_i}} \|\boldsymbol{x - y}\|^2$$

Ο αλγόριθμος εναλλάσσεται μεταξύ δυο βημάτων [18]:

Βήμα ανάθεσης: Πραγματοποιείται η ανάθεση κάθε δεδομένου στην κλάση με τον κοντινότερο μέσο, αυτού με την ελάχιστη Ευκλείδεια απόσταση δ:

$$\boldsymbol{\delta} = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{x_i - y_i})^2}$$

$$\boldsymbol{S_i^{(t)}} = \{\boldsymbol{x_p}: \|\boldsymbol{x_p - m_i^{(t)}}\|^2 \leq \|\boldsymbol{x_p - m_j^{(t)}}\|^2 \forall \boldsymbol{j}, \mathbf{1} \leq \boldsymbol{j} \leq \boldsymbol{k}\}$$

όπου κάθε x_p γίνεται ανάθεση σε ακριβώς ένα S_i, ακόμη κι αν θα μπορούσε να γίνει ανάθεση σε δυο ή περισσότερα.

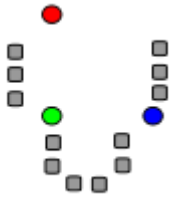
Βήμα ενημέρωσης: Γίνεται εκ νέου υπολογισμός των μέσων (centroids) για τα δεδομένα που τοποθετήθηκαν σε κάθε κλάση.

$$\boldsymbol{m_i^{(t+1)}} = \frac{\mathbf{1}}{\boldsymbol{|S_i^{(t)}|}} \sum_{\boldsymbol{x_j \in S_i^{(t)}}} \boldsymbol{x_j}$$

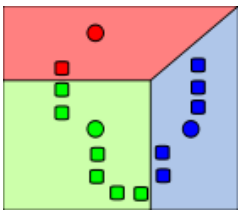
Μεθοδολογία τρεξίματος επαναληπτικού αλγόριθμου k-means:

- Καθορισμός του αριθμού των επιθυμητών κλάσεων
- Τυχαία ανάθεση κάθε δεδομένου σε κάποια κλάση
- Υπολογισμός των κέντρων των κλάσεων
- Ανάθεση εκ νέου κάθε δεδομένου στο κοντινότερο κέντρο
- Υπολογισμός εκ νέου των κέντρων των κλάσεων
- Επανάληψη των δυο προηγουμένων βημάτων μέχρι να μη χρειάζονται βελτιώσεις

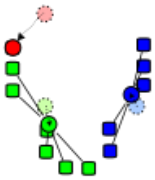
Αναπαράσταση τρεξίματος με σχήματα (Σχήματα 2.17 έως 2.20):



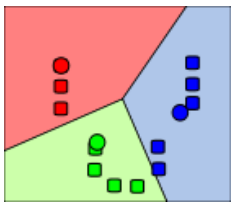
Σχήμα 2.17: Σχήμα αρχικού βήματος k-means, k αρχικοί μέσοι ανατίθενται τυχαία [17]



Σχήμα 2.18: Σχήμα επόμενου βήματος k-means, k κλάσεις σχηματίζονται συνδέοντας κάθε δεδομένο με το κοντινότερο κέντρο [17]



Σχήμα 2.19: Σχήμα τρίτου βήματος k-means, τα κέντρα των k κλάσεων γίνονται οι νέοι μέσοι [17]



Σχήμα 2.20: Σχήμα τελικό k-means, όπου τα δυο προηγούμενα βήματα έχουν επαναληφθεί ώστε να μην χρειάζεται άλλη βελτίωση [17]

Στην περίπτωση μας, για να γίνει η συμπλήρωση των ελλιπών τιμών πραγματοποιείται ομαδοποίηση των προηγούμενων τιμών της στήλης, όταν βρεθεί κάποια ελλιπής τιμή. Αυτό εξυπηρετεί στο να γίνει εντοπισμός προτύπων στα ημερήσια δεδομένα και με βάση τα πρότυπα αυτά εξετάζεται αν και οι αμέσως προηγούμενες τιμές της ελλιπούς τιμής ακολουθούν κάποιο πρότυπο.

Το παράθυρο τιμών για να δημιουργηθεί το πρότυπο πριν την ελλιπή τιμή ορίζεται αρχικά σε 3. Αν δε βρεθεί το πρότυπο μειώνεται κατά ένα και ξεκινά την αναζήτηση προτύπου μήκους 2. Ομοίως αν πάλι δε βρεθεί το μήκος παραθύρου μειώνεται σε 1.

Όταν βρεθεί το patternσε προηγούμενα στοιχεία της στήλης κρατείται η επόμενη τιμή μετά από κάθε εύρεση του προτύπου. Έχοντας όλες τις τιμές που ακολουθούν τις εμφανίσεις του αναζητούμενο προτύπου, συμπληρώνεται κάθε ελλιπήςτιμή με τον μέσο όρο των τιμών αυτών, των ακολούθων δηλαδή των εμφανίσεων του προτύπου.

Το ζητούμενο παράθυρο τιμών επιστρέφει στο μήκος 3, σε κάθε αναζήτηση ελλιπούς τιμής και μειώνεται ανάλογα σε κάθε βήμα, εάν είναι απαραίτητο.

Έστω πως βρίσκεται ελλιπής τιμή στη θέση $\boldsymbol{x(i,j)}$.

Πραγματοποιείται σάρωση του συνόλου δεδομένων από την αρχή της στήλης ως την προηγούμενη τιμή και ομαδοποίηση των δεδομένων $(\boldsymbol{x(1,j)}, \boldsymbol{x(2,j)}, \ldots \boldsymbol{x(i-1,j)})$ σε κλάσεις έστω 3 διαφορετικές.

Έχοντας τα επιμέρους 3 σύνολα δεδομένων $S = (S_1, S_2, S_3)$ και έχοντας ταξινομήσει κάθε μια από τις προηγούμενες τιμές σε ένα από αυτά τα σύνολα, αναζητείται το patternσυνόλων στα οποία ανήκουν οι τιμές $\mathbf{x(i-3,j)x(i-2,j)x(i-1,j)}$.

Εάν δε βρεθεί εμφάνιση του pattern, αναζητείται το patternσυνόλων των τιμών

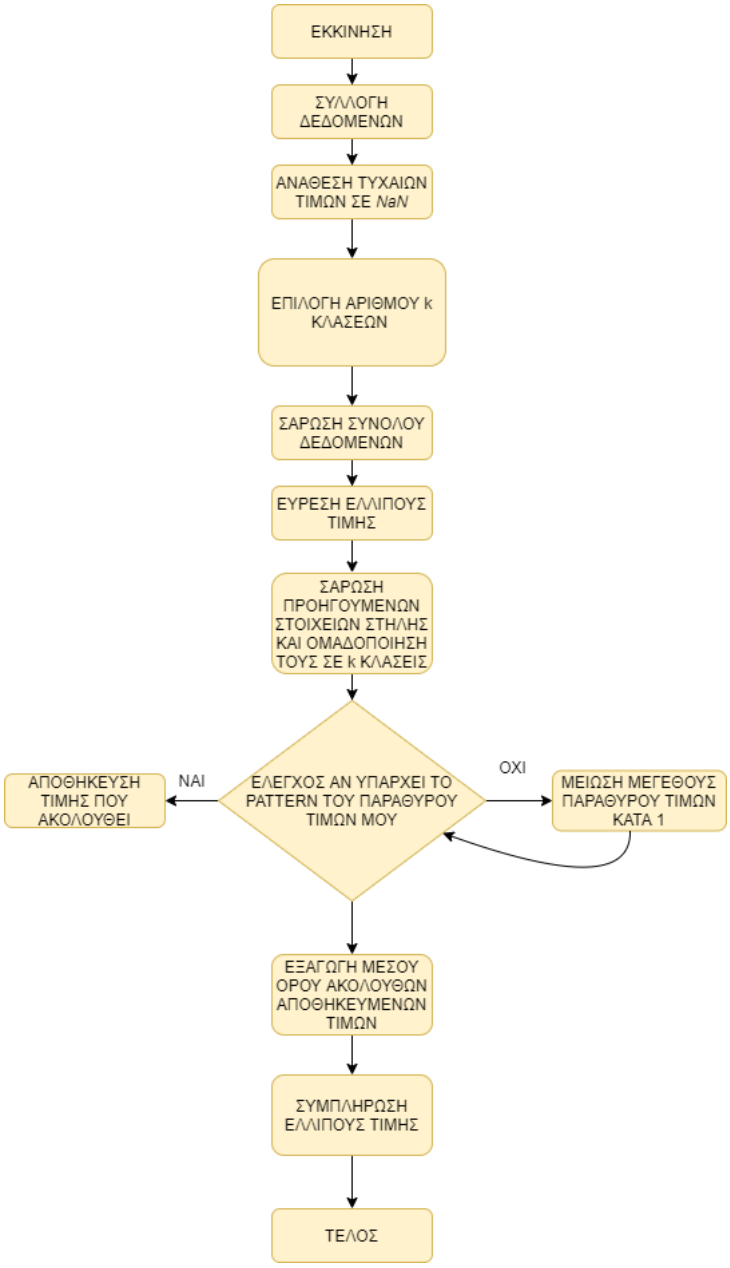
$\boldsymbol{x(i-2,j)x(i-1,j)}$.

Όταν βρεθούν οι εμφανίσεις των ζητούμενων pattern κρατείται σε άθροισμα η καθεμιά επόμενη τιμή των εμφανίσεων, και τέλος υπολογίζεται ο μέσος όρος τους.

Με τον μέσο όρο αυτόν συμπληρώνεται τελικά και η ελλιπής τιμή.

$$\mathbf{x(i,j) = \frac{1}{n}\sum_{k=1}^{n} x(a,j)}$$

όπου n ο αριθμός εμφανίσεων του pattern στη στήλη και $\boldsymbol{x}(\boldsymbol{a}, \boldsymbol{j})$ η ακόλουθη κάθε φορά τιμή. Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.21.



Σχήμα 2.21: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήση k-means clustering

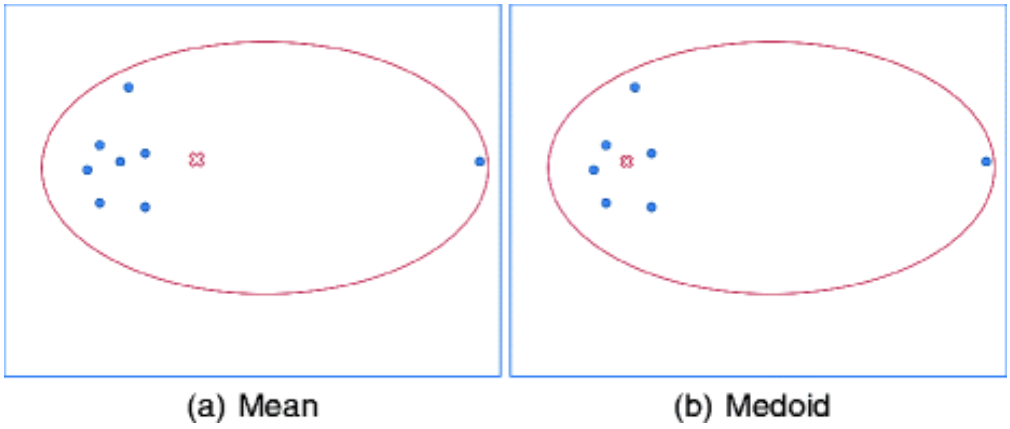
2.3.4. Μέθοδος συμπλήρωσης με χρήση k-medoids αλγορίθμου

Ο αλγόριθμος k-medoids αποτελεί μια μέθοδο ομαδοποίησης, παρόμοια του k-means clustering. Και οι δυο αλγόριθμοι στοχεύουν στο να διαιρέσουν το σύνολο δεδομένων σε ομάδες και προσπαθούν να ελαχιστοποιήσουν την απόσταση μεταξύ σημείων των κλάσεων και των κέντρων των συστάδων. [19]

Σε αντίθεση με τον k-means αλγόριθμο, ο k-medoids, επιλέγει υπαρκτά σημεία δεδομένων ως κέντρα και επομένως βοηθά στην ερμηνεία τους περισσότερο, σχετικά με τον k-means, όπου τα κέντρα των κλάσεων δεν είναι απαραίτητα κάποιο από τα δεδομένα (είναι η μέση τιμή μεταξύ των δεδομένων της κλάσης).

Ο k-medoids αποτελεί μια κλασσική τεχνική διάσπασης και ομαδοποίησης του συνόλου δεδομένων των ναντικειμένων σε k κλάσεις, όπου το k ορίζεται πριν αρχίσει η εκτέλεση της μεθόδου.

Ως medoid μιας κλάσης ορίζεται ως το αντικείμενο της κλάσης, του οποίου η μέση διαφορά με όλα τα αντικείμενα της κλάσης είναι η μικρότερη, και είναι συνεπώς το πιο κεντρικό σημείο της κλάσης. Η διαφορά μεταξύ μέσης τιμής και medoid κλάσης παρουσιάζεται στο Σχήμα 2.22.



Σχήμα 2.22: Σχήμα διαφοράς μεταξύ mean και medoid μιας κλάσης [20]

Μεθοδολογία k-medoids [22]:

- Αρχικά γίνεται ο ορισμός των k τυχαίων σημείων του συνόλου δεδομένων n ως αρχικά medoids.
- Πραγματοποιείται σύνδεση καθενός από τα δεδομένα με το κοντινότερο medoid χρησιμοποιώντας οποιαδήποτε μορφή απόστασης
- Για κάθε medoid m, για κάθε δεδομένο o, το οποίο δεν είναι medoid κάποιας κλάσης, γίνεται αλλαγή μεταξύ m και o, σύνδεση καθενός από τα δεδομένα στο κοντινότερο medoid και υπολογισμός του κόστους

- Αν το συνολικό κόστος είναι περισσότερο από ότι πριν, αναιρείται η αλλαγή μεταξύ m και o

Η διαφορά μεταξύ ενός medoid (C_i) και ενός αντικειμένου (P_i) υπολογίζεται από τη σχέση [22]:

$$\mathbf{E} = |\mathbf{P_i} - \mathbf{C_i}|$$

Ενώ το κόστος του k-medoids αλγορίθμου υπολογίζεται ως [22]:

$$\mathbf{c} = \sum_{\mathbf{c_i}} \sum_{\mathbf{P_i} \in \mathbf{C_i}} |\mathbf{P_i} - \mathbf{C_i}|$$

Η παραπάνω προσέγγιση αποτελεί μια από τις ποικίλες παραλλαγές του k-medoids, η οποία ξεκινά από το αρχικό σύνολο των medoids και ελέγχει κάθε φορά αν βελτιώνεται η συνολική απόσταση ορίζοντας ως medoids διαφορετικά δεδομένα και συγκρίνοντας σε κάθε βήμα.

Η μέθοδος αυτή είναι γνωστή και ως Partitioning Around Medoids (PAM) [21].

Στην περίπτωση μας,όμοια με την μέθοδο του k-means για να γίνει η συμπλήρωση των ελλιπών τιμών πραγματοποιείται ομαδοποίηση των προηγούμενων τιμών της στήλης, όταν βρεθεί κάποια ελλιπής τιμή, ώστε να εντοπιστούν τα πρότυπα (patterns) στα δεδομένα και με βάση αυτά να γίνει η συμπλήρωση.

Το παράθυρο τιμών για να δημιουργηθεί το πρότυπο πριν την ελλιπή τιμή ορίζεται αρχικά σε 3. Αν δε βρεθεί το πρότυπο μειώνεται κατά ένα και ξεκινά την αναζήτηση προτύπου μήκους 2. Ομοίως αν πάλι δε βρεθεί το μήκος παραθύρου μειώνεται σε 1.

Όταν βρεθεί το pattern σε προηγούμενα στοιχεία της στήλης κρατείται η επόμενη τιμή μετά από κάθε εύρεση του προτύπου. Έχοντας όλες τις τιμές που ακολουθούν τις εμφανίσεις του αναζητούμενο προτύπου, συμπληρώνεται κάθε ελλιπής τιμή με τον μέσο όρο των τιμών αυτών, των ακολούθων δηλαδή των εμφανίσεων του προτύπου.

Το ζητούμενο παράθυρο τιμών επιστρέφει στο μήκος 3, σε κάθε αναζήτηση ελλιπούς τιμής και μειώνεται ανάλογα σε κάθε βήμα, εάν είναι απαραίτητο.

Έστω πως βρίσκεται ελλιπής τιμή στη θέση $\boldsymbol{x(i,j)}$.

Πραγματοποιείται σάρωση του συνόλου δεδομένων από την αρχή της στήλης ως την προηγούμενη τιμή και ομαδοποίηση των δεδομένων $(\boldsymbol{x(1,j)}, \boldsymbol{x(2,j)}, \ldots \boldsymbol{x(i-1,j)})$σε κλάσεις έστω 3 διαφορετικές.

Έχοντας τα επιμέρους 3 σύνολα δεδομένων $S = (S_1, S_2, S_3)$και έχοντας ταξινομήσει κάθε μια από τις προηγούμενες τιμές σε ένα από αυτά τα σύνολα, αναζητείται το patternσυνόλων στα οποία ανήκουν οι τιμές $\mathbf{x(i-3,j)x(i-2,j)x(i-1,j)}$.

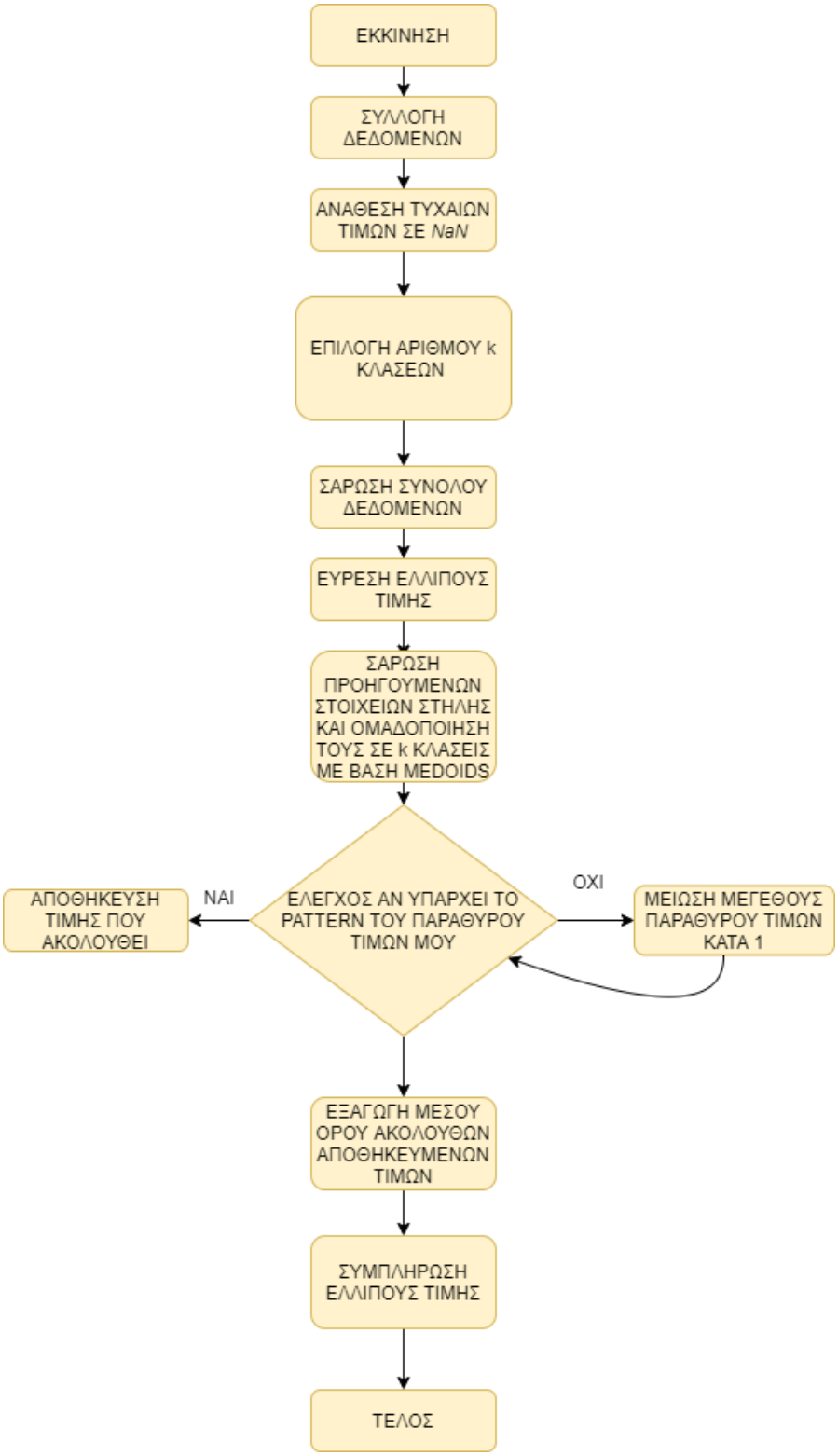
Εάν δε βρεθεί εμφάνιση του pattern, αναζητείται το patternσυνόλων των τιμών $\mathbf{x(i-2,j)x(i-1,j)}$.

Όταν βρεθούν οι εμφανίσεις των ζητούμενων patternκρατείται σε άθροισμα η καθεμιά επόμενη τιμή των εμφανίσεων, και τέλος υπολογίζεται ο μέσος όρος τους.

Με τον μέσο όρο αυτόν συμπληρώνεται τελικά και η ελλιπής τιμή.

$$\mathbf{x(i,j) = \frac{1}{n}\sum_{k=1}^{n} x(a,j)}$$

όπου n ο αριθμός εμφανίσεων του pattern στη στήλη και $x(a,j)$ η ακόλουθη κάθε φορά τιμή. Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.23.



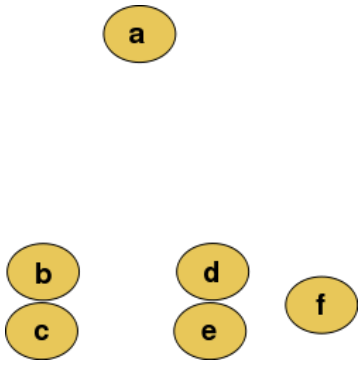
Σχήμα 2.23: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήση k-medoids clustering

2.3.5. Μέθοδος συμπλήρωσης με χρήση ιεραρχικής ομαδοποίησης (hierarchical clustering)

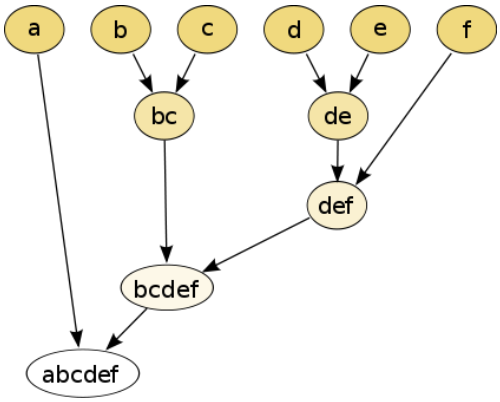
Η ιεραρχική ομαδοποίηση (hierarchical clustering), όπως λέει και το όνομά της, πρόκειται για αλγόριθμο που χτίζει μια ιεραρχία από κλάσεις. Ο αλγόριθμος ξεκινά αναθέτοντας καθένα από τα δεδομένα του συνόλου δεδομένων σε ένα ξεχωριστό ατομικό cluster (Σχήμα 2.24). Έπειτα, τα δυο κοντινότερα cluster ενώνονται σε ένα ίδιο. Στο τέλος, ο αλγόριθμος τερματίζει όταν έχει πια μείνει μόνο ένα cluster (Σχήμα 2.25).

Συγκεκριμένα, υπάρχουν δυο μηχανισμοί ιεραρχικής ομαδοποίησης [23]:

- Agglomerative: ο μηχανισμός που περιγράψαμε και παραπάνω, κατά τον οποίο κάθε δεδομένο του συνόλου ξεκινά από ένα ξεχωριστό μοναδικό cluster και η ιεραρχία πηγαίνει από κάτω προς τα πάνω μέχρι να μείνει μια μόνο κλάση
- Divisive: ο αντίστροφος μηχανισμός, κατά τον οποίο όλα τα δεδομένα ξεκινούν μαζί από μια κλάση και πραγματοποιούνται διαχωρισμοί των κλάσεων, καθώς η ιεραρχία πηγαίνει από πάνω προς τα κάτω

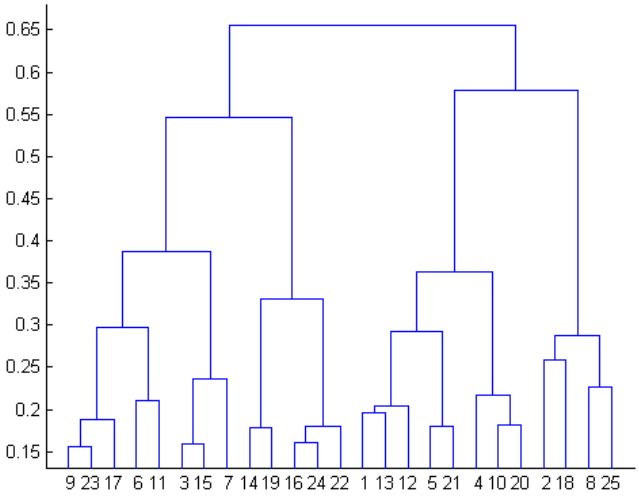


Σχήμα 2.24: Αρχικά δεδομένα πριν την ιεραρχική ομαδοποίηση [24]



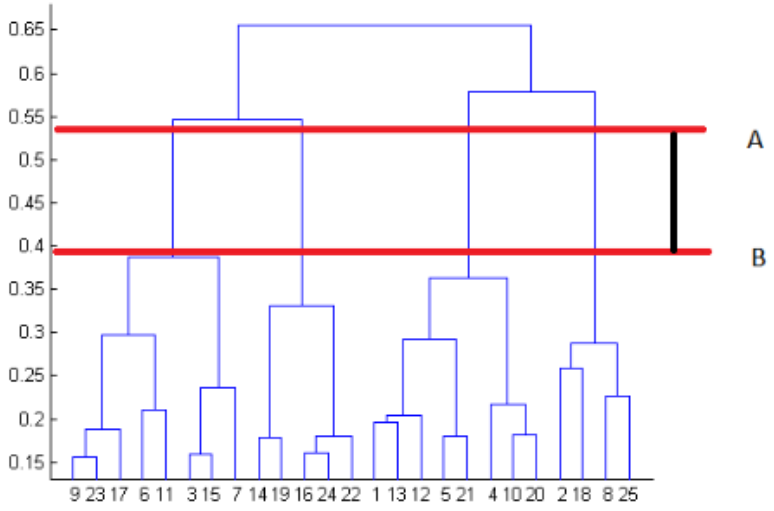
Σχήμα 2.25: Αναπαράσταση δεδομένων μετά την ιεραρχική ομαδοποίηση [24]

Τα αποτελέσματα της ιεραρχικής ομαδοποίησης μπορούν να παρουσιαστούν και σε διάγραμμα, όπως στα Σχήματα 2.26 και 2.27.



Σχήμα 2.26: Παράδειγμα αναπαράστασης δεδομένων κατά την ιεραρχική ομαδοποίηση (25 δεδομένα) [23]

Η απόφαση για τον αριθμό των κλάσεων μπορεί να αναχθεί από την παρατήρηση του δενδροδιαγράμματος. Για παράδειγμα η καλύτερη επιλογή αριθμού κλάσεων μπορεί να βρεθεί παίρνοντας οριζόντια γραμμή στο δενδροδιάγραμμα καλύπτοντας μέγιστη κάθετη απόσταση χωρίς να διασταυρωθεί με άλλη κλάση.



Σχήμα 2.27: Παράδειγμα εύρεσης ιδανικού αριθμού κλάσεων (αριθμός = 4 στο παράδειγμά μας) [23]

Ο αλγόριθμος του παραπάνω παραδείγματος υλοποιήθηκε με την προσέγγιση από κάτω προς τα πάνω.

Η απόφαση για την ένωση δυο κλάσεων παίρνεται με βάση μετρικές (metrics) αποστάσεων για να διαπιστωθεί πόσο κοντά βρίσκονται δυο κλάσεις [23]:

- Ευκλείδεια απόσταση:

$$\|\boldsymbol{a}-\boldsymbol{b}\|_2 = \sqrt{(\sum(\boldsymbol{a_i}-\boldsymbol{b_i}))}$$

- Τετραγωνική ευκλείδεια απόσταση:

$$\|\boldsymbol{a}-\boldsymbol{b}\|_2^2 = (\sum(\boldsymbol{a_i}-\boldsymbol{b_i})^2)$$

- Απόσταση Manhattan:

$$\|\boldsymbol{a}-\boldsymbol{b}\|_1 = \sum|\boldsymbol{a_i}-\boldsymbol{b_i}|$$

- Μέγιστη απόσταση:

$$\|\boldsymbol{a}-\boldsymbol{b}\|_\infty = \boldsymbol{max_i}|\boldsymbol{a_i}-\boldsymbol{b_i}|$$

- Απόσταση Mahalanobis:

$$\sqrt{(\boldsymbol{a}-\boldsymbol{b})^T\boldsymbol{S}^{-1}(\boldsymbol{a}-\boldsymbol{b})}$$

  όπου S ο πίνακας διακύμανσης (covariancematrix)

Το κριτήριο συνδέσμου (linkage) καθορίζει την απόσταση μεταξύ σετ δεδομένων σαν συνάρτηση αποστάσεων ανά ζεύγη. Κάποια συχνά χρησιμοποιημένα κριτήρια συνδέσμου για δυο σετ δεδομένων Α και Β είναι [25][26]:

- Unweighted average linkage clustering (UPGMA):

$$\frac{\mathbf{1}}{|\boldsymbol{A}||\boldsymbol{B}|}\sum_{\boldsymbol{a}\in\boldsymbol{A}}\sum_{\boldsymbol{b}\in\boldsymbol{B}}\boldsymbol{d}(\boldsymbol{a},\boldsymbol{b})$$

  όπου σε κάθε βήμα ομαδοποίησης, η ενημερωμένη απόσταση μεταξύ ενωμένων κλάσεων AUB και μια νέα κλάση X δίνεται από τη μέση τιμή των αποστάσεων $d_{A,x}$ και $d_{B,x}$

- Weighted average linkage clustering (WPGMA):

$$\boldsymbol{d}(\boldsymbol{i}\cup\boldsymbol{j},\boldsymbol{k}) = \frac{\boldsymbol{d}(\boldsymbol{i},\boldsymbol{k})+\boldsymbol{d}(\boldsymbol{j},\boldsymbol{k})}{\mathbf{2}}$$

  όπου σε κάθε βήμα οι κοντινότερες δυο κλάσεις i και j ενώνονται σε παραπάνω επίπεδο στην κλάση $i \cup j$και η απόσταση από μια άλλη κλάση k υπολογίζεται από την μέση τιμή των μέσων αποστάσεων μεταξύ μελών k και I κλάσεων

- Centroid distance (UPGMC):

$$\|\boldsymbol{c_s}-\boldsymbol{c_t}\|$$

όπου c_s και c_t τα κέντρα βάρους των κλάσεων s και t

- Maximum ή complete linkage clustering: μέγιστηαπόσταση

$$\boldsymbol{max}\{\boldsymbol{d}(\boldsymbol{a},\boldsymbol{b}):\boldsymbol{a}\in\boldsymbol{A},\boldsymbol{b}\in\boldsymbol{B}\}$$

- Minimum ή single linkage clustering: ελάχιστηαπόσταση

$$\boldsymbol{min}\{\boldsymbol{d}(\boldsymbol{a},\boldsymbol{b}):\boldsymbol{a}\in\boldsymbol{A},\boldsymbol{b}\in\boldsymbol{B}\}$$

- Ward's criterion: με το οποίο επιτυγχάνεται η ελαχιστοποίηση της συνολικής διακύμανσης (minimum variance) μεταξύ μιας κλάσης και χρήση της τετραγωνικής απόστασης μεταξύ κέντρων κλάσεων για αλλαγή διακύμανσης[19]

$$\boldsymbol{d_{ij}}=\boldsymbol{d}(\{\boldsymbol{X_i}\},\{\boldsymbol{X_j}\})=\|\boldsymbol{X_i}-\boldsymbol{X_j}\|^{\boldsymbol{2}}$$

Στην περίπτωση μας,για να γίνει η συμπλήρωση των ελλιπών τιμών πραγματοποιείται ομαδοποίηση των προηγούμενων τιμών της στήλης, όταν βρεθεί κάποια ελλιπής τιμή, ώστε να εντοπιστούν τα πρότυπα (patterns) στα δεδομένα και με βάση αυτά να γίνει η συμπλήρωση.

Το παράθυρο τιμών για να δημιουργηθεί το πρότυπο πριν την ελλιπή τιμή ορίζεται αρχικά σε 3 (ορίζεται ως Maxclust στο περιβάλλον του Matlab). Αν δε βρεθεί το πρότυπο μειώνεται κατά ένα και ξεκινά την αναζήτηση προτύπου μήκους 2. Ομοίως αν πάλι δε βρεθεί το μήκος παραθύρου μειώνεται σε 1.

Όταν βρεθεί το pattern σε προηγούμενα στοιχεία της στήλης κρατείται η επόμενη τιμή μετά από κάθε εύρεση του προτύπου. Έχοντας όλες τις τιμές που ακολουθούν τις εμφανίσεις του αναζητούμενο προτύπου, συμπληρώνεται κάθε ελλιπής τιμή με τον μέσο όρο των τιμών αυτών, των ακολούθων δηλαδή των εμφανίσεων του προτύπου.

Το ζητούμενο παράθυρο τιμών επιστρέφει στο μήκος 3, σε κάθε αναζήτηση ελλιπούς τιμής και μειώνεται ανάλογα σε κάθε βήμα, εάν είναι απαραίτητο.

Έστω πως βρίσκεται ελλιπής τιμή στη θέση $x(i,j)$.

Πραγματοποιείται σάρωση του συνόλου δεδομένων από την αρχή της στήλης ως την προηγούμενη τιμή και ομαδοποίηση των δεδομένων $(x(1,j),x(2,j),\dots x(i-1,j))$ σε κλάσεις έστω 3 διαφορετικές.

Έχοντας τα επιμέρους 3 σύνολα δεδομένων $S = (S_1, S_2, S_3)$ και έχοντας ταξινομήσει κάθε μια από τις προηγούμενες τιμές σε ένα από αυτά τα σύνολα, αναζητείται το pattern συνόλων στα οποία ανήκουν οι τιμές $x(i-3,j)x(i-2,j)x(i-1,j)$.

Εάν δε βρεθεί εμφάνιση του pattern, αναζητείται το patternσυνόλων των τιμών $x(i-2,j)x(i-1,j)$.

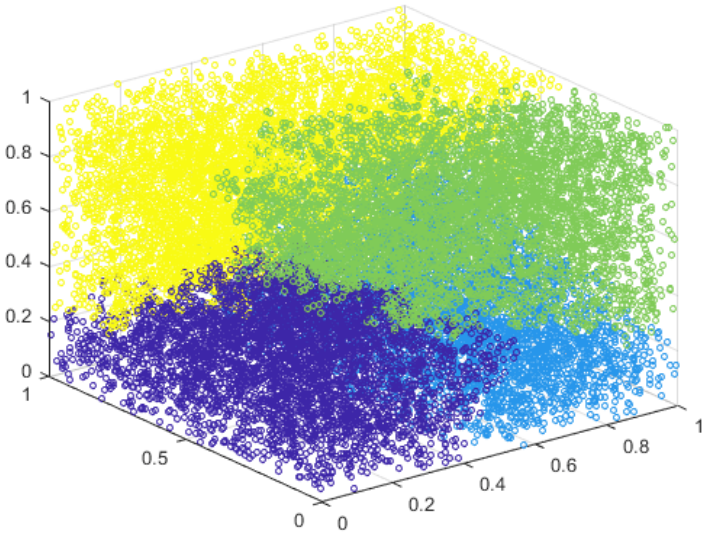
Όταν βρεθούν οι εμφανίσεις των ζητούμενων patternκρατείται σε άθροισμα η καθεμιά επόμενη τιμή των εμφανίσεων, και τέλος υπολογίζεται ο μέσος όρος τους.

Με τον μέσο όρο αυτόν συμπληρώνεται τελικά και η ελλιπής τιμή.

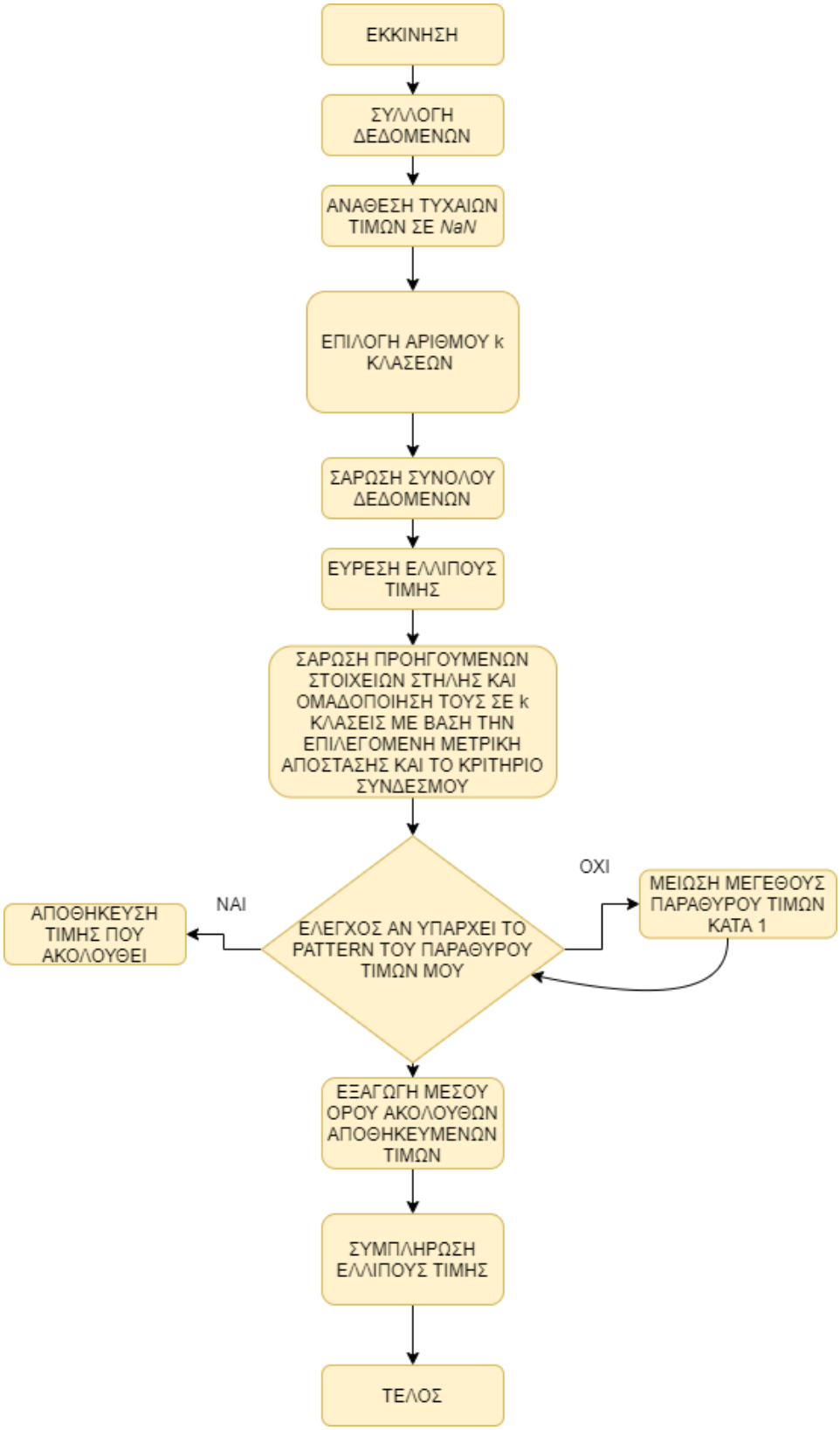
$$\mathbf{x(i,j) = \frac{1}{n}\sum_{k=1}^{n} x(a,j)}$$

όπου no αριθμός εμφανίσεων του patternστη στήλη και $x(a,j)$ η ακόλουθη κάθε φορά τιμή. Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.29.

Γίνεται εφαρμογή περισσότερων από ένα κριτηρίων συνδέσμου (linkage criteria) και γίνεται εξαγωγή αποτελεσμάτων και σύγκριση των αποτελεσμάτων. Το αποτέλεσμα μεθοδολογίας ομαδοποίησης στο γραφικό περιβάλλον του Matlab απεικονίζεται στο Σχήμα 2.28.



Σχήμα 2.28: Παράδειγμα clustering σε 4 κλάσεις μέσω του Matlab

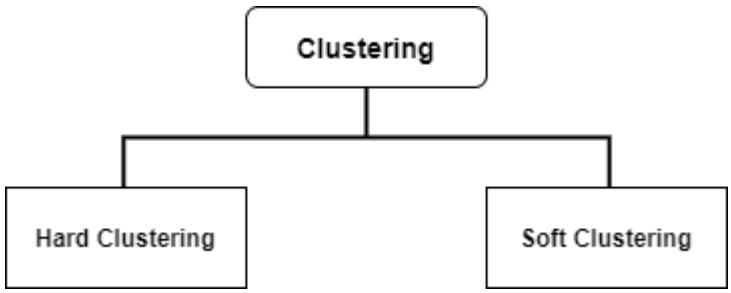


Σχήμα 2.29: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήση hierarchical clustering

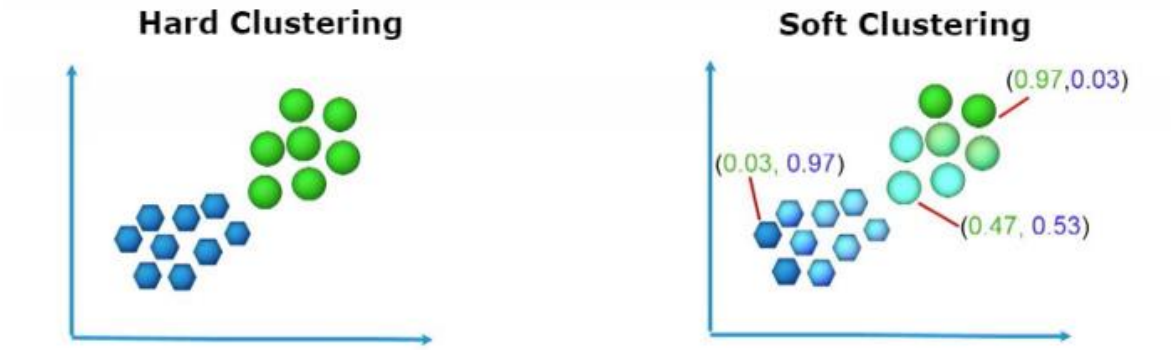
2.3.6. Μέθοδος συμπλήρωσης με χρήση ασαφούς ομαδοποίησης (fuzzy clustering)

Οι μέθοδοι ομαδοποίησης διακρίνονται σε δυο μεγάλες κατηγορίες, όπως φαίνεται στο Σχήμα 2.30. Στην μη ασαφή ομαδοποίηση (hard clustering), το σύνολο των δεδομένων διαιρείται σε διακριτές κλάσεις, όπου κάθε δεδομένο μπορεί να ανήκει μόνο σε μια ακριβώς κλάση.

Η ασαφής ομαδοποίηση (επίσης γνωστή ως soft clustering ή soft k-means) αποτελεί ένα είδος ομαδοποίησης στην οποία κάθε δεδομένο του συνόλου μπορεί να ανήκει σε περισσότερες από μια κλάσεις (Σχήμα 2.31).



Σχήμα 2.30: Υποσύνολα μεθόδου ομαδοποίησης [29]



Σχήμα 2.31: Σύγκριση hard clustering και soft clustering [29]

Ένας από τους πιο γνωστούς και ευρέως χρησιμοποιημένους αλγορίθμους της ασαφούς ομαδοποίησης είναι ο fuzzy c-means (FCM) αλγόριθμος ομαδοποίησης, που αποτελεί μια soft-clustering προσέγγιση, στην οποία κάθε δεδομένο του συνόλου ανατίθεται με πιθανότητα να ανήκει σε κάποια κλάση.

Γενική περιγραφή του αλγορίθμου FCM [30]:

- Επιλογή του αριθμού των clusters
- Ανάθεση τυχαίων συντελεστών με τους οποίους κάθε δεδομένο του συνόλου ανήκει σε κάποια κλάση
- Επανάληψη μέχρι ο αλγόριθμος να συγκλίνει (δηλαδή οι συντελεστές δεν αλλάξουν αρκετά από το ένα βήμα στο άλλο)
- Υπολογισμός του κέντρου βάρους (centroid) της κάθε κλάσης
- Υπολογισμός για κάθε δεδομένο του συνόλου των συντελεστών που συμβολίζουν κατά πόσο ανήκει σε κάποια κλάση

Κάθε δεδομένο x έχει ένα σετ συντελεστών που δείχνει τον βαθμό με τον οποίο ανήκει στο k-cluster $w_k(x)$.

Το κέντρο βάρους μιας κλάσης είναι ο μέσος όρος όλων των σημείων, ζυγισμένος από τον βαθμό με τον οποίο ανήκουν στην κλάση:

$$\boldsymbol{c_k} = \frac{\sum_x \boldsymbol{w_k}(\boldsymbol{x})^{\boldsymbol{m}} \boldsymbol{x}}{\sum_x \boldsymbol{w_k}(\boldsymbol{x})^{\boldsymbol{m}}}$$

όπου m η παράμετρος που δείχνει πόσο ασαφής θα είναι η κλάση. Όσο μεγαλύτερη η τιμή της, τόσο πιο ασαφής θα είναι η κλάση στο τέλος.

Ο αλγόριθμος FCM προσπαθεί να διαιρέσει μια τετελεσμένη χρονοσειρά nστοιχείων $X = \{x_1, \ldots, x_n\}$ σε μια συλλογή από c ασαφείς κλάσεις.

Ο αλγόριθμος επιστρέφει μια λίστα από c κέντρα κλάσεων $C = \{c_1, \ldots, c_c\}$ και έναν πίνακα διχοτόμησης $W = w_{i,j} \in [0,1], i = 1, \ldots, n, j = 1, \ldots, c$, όπου κάθε στοιχείο $w_{i,j}$ δείχνει τον βαθμό με τον οποίο το στοιχείο x_i ανήκει στην κλάση c_j

Ο FCM στοχεύει στην ελαχιστοποίηση της συνάρτησης:

$$\boldsymbol{arg_C min} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{c}} \boldsymbol{w_{ij}^m} \|\boldsymbol{x_i} - \boldsymbol{c_j}\|^{\boldsymbol{2}}$$

όπου

$$\boldsymbol{w_{ij}} = \frac{\mathbf{1}}{\sum_{\boldsymbol{k=1}}^{\boldsymbol{c}} \left(\frac{\|\boldsymbol{x_i}-\boldsymbol{c_j}\|}{\|\boldsymbol{x_i}-\boldsymbol{c_k}\|}\right)^{\frac{2}{m-1}}}$$

με τιμές w_{ij}(membership values) και συντελεστή ασάφειας (fuzzifier) $m \in R, m \geq 1$.

Στην περίπτωση μας,για την συμπλήρωση των ελλιπών δεδομένων, ακολουθείται όμοια διαδικασία όπως με την μέθοδο του k-means. Πραγματοποιείται ομαδοποίηση των προηγούμενων τιμών της στήλης, όταν βρεθεί κάποια ελλιπής τιμή, ώστε να εντοπιστούν τα πρότυπα (patterns) στα δεδομένα και με βάση αυτά να γίνει η συμπλήρωση.

Το παράθυρο τιμών για να δημιουργηθεί το πρότυπο πριν την ελλιπή τιμή ορίζεται αρχικά σε 3. Αν δε βρεθεί το πρότυπο μήκους 3 μειώνεται κατά ένα και ξεκινά την αναζήτηση προτύπου μήκους 2. Ομοίως αν πάλι δε βρεθεί το μήκος παραθύρου μειώνεται σε 1.

Όταν βρεθεί το pattern σε προηγούμενα στοιχεία της στήλης κρατείται η επόμενη τιμή μετά από κάθε εύρεση του προτύπου. Έχοντας όλες τις τιμές που ακολουθούν τις εμφανίσεις του αναζητούμενο προτύπου, συμπληρώνεται κάθε ελλιπής τιμή με τον μέσο όρο των τιμών αυτών, των ακολούθων δηλαδή των εμφανίσεων του προτύπου.

Το ζητούμενο παράθυρο τιμών επιστρέφει στο μήκος 3, σε κάθε αναζήτηση ελλιπούς τιμής και μειώνεται ανάλογα σε κάθε βήμα, εάν είναι απαραίτητο.

Έστω πως βρίσκεται ελλιπής τιμή στη θέση $x(i,j)$.

Πραγματοποιείται σάρωση του συνόλου δεδομένων από την αρχή της στήλης ως την προηγούμενη τιμή και ομαδοποίηση των δεδομένων $(x(1,j), x(2,j), \ldots x(i-1,j))$σε κλάσεις έστω 3 διαφορετικές.

Έχοντας τα επιμέρους 3 σύνολα δεδομένων $S = (S_1, S_2, S_3)$και έχοντας ταξινομήσει κάθε μια από τις προηγούμενες τιμές σε ένα από αυτά τα σύνολα, αναζητείται το patternσυνόλων στα οποία ανήκουν οι τιμές $x(i-3,j)x(i-2,j)x(i-1,j)$.

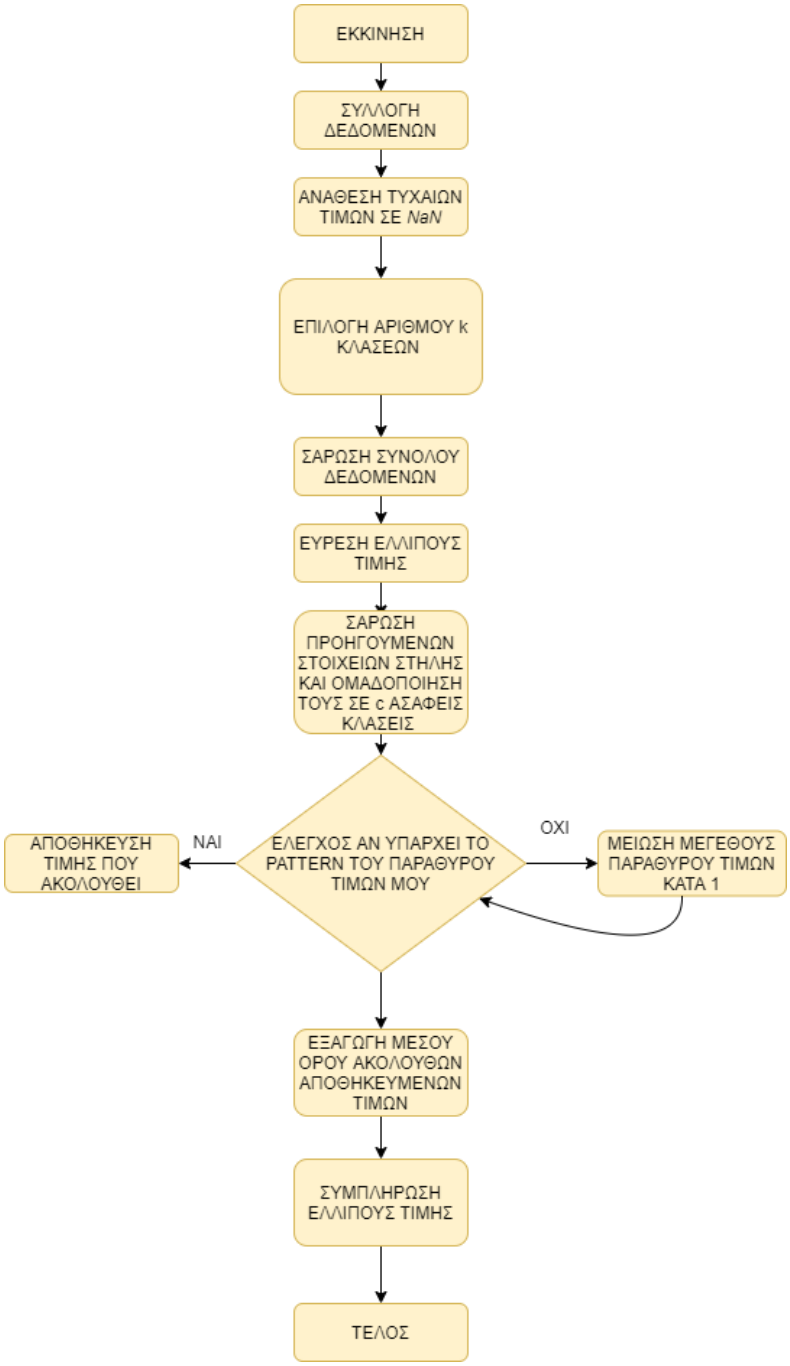
Εάν δε βρεθεί εμφάνιση του pattern, αναζητείται το pattern συνόλων των τιμών

$x(i-2,j)x(i-1,j)$.

Όταν βρεθούν οι εμφανίσεις των ζητούμενων pattern κρατείται σε άθροισμα η καθεμιά επόμενη τιμή των εμφανίσεων, και τέλος υπολογίζεται ο μέσος όρος τους.

Με τον μέσο όρο αυτόν συμπληρώνεται τελικά και η ελλιπής τιμή.

$\boldsymbol{x(i,j) = \frac{1}{n} \sum_{k=1}^{n} x(a,j)}$ όπου no αριθμός εμφανίσεων του pattern στη στήλη και $x(a,j)$ η ακόλουθη κάθε φορά τιμή.Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.32.



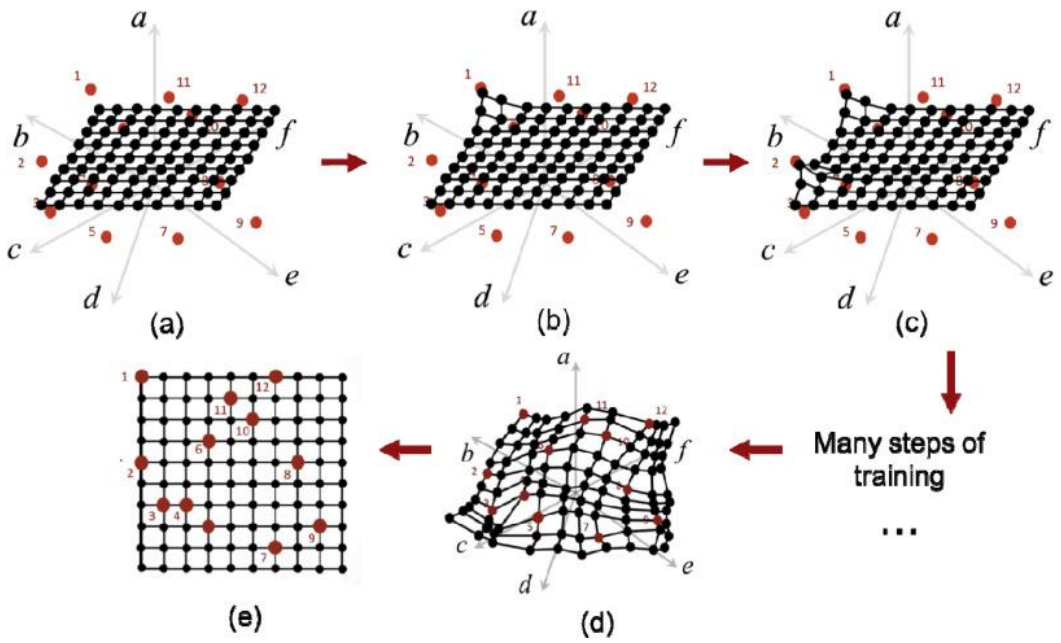
Σχήμα 2.32: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήση ασαφούς ομαδοποίησης(fuzzyclustering)

2.3.7. Μέθοδος συμπλήρωσης με χρήση ομαδοποίησης με αυτοοργανώμενο χάρτη (self-organizing map)

Ο αυτο-οργανώμενος χάρτης (self-organizing map) αποτελεί ένα είδος τεχνητού νευρωνικού δικτύου, όπου η εκμάθηση γίνεται χωρίς επίβλεψη (unsupervised learning).

Στην εκμάθηση με επίβλεψη υπάρχει ένα στοχευμένο output για κάθε πρότυπο input και το δίκτυο μαθαίνει να παράγει τα απαραίτητα outputs. Αντίθετα, στην εκμάθηση χωρίς επίβλεψη, το δίκτυο μαθαίνει να σχηματίζει τις δικές του ταξινομήσεις χωρίς περαιτέρω βοήθεια, αλλά με χρήση κοινών χαρακτηριστικών μεταξύ των δεδομένων.

Το νευρωνικό μοντέλο περιλαμβάνει κόμβους output, που είναι συνδεδεμένοι με το επίπεδο εισόδου και το είδος αυτό αναπαράστασης είναι γνωστό ως τοπογραφικός χάρτης (topographic map), όπως φαίνεται στο Σχήμα 2.33. [31]



Σχήμα 2.33: Σχήμα απεικόνισης τοπογραφικού χάρτη και βημάτων αυτοοργάνωσης [32]

Η εκμάθηση χωρίς επίβλεψη στηρίζεται στην ανταγωνιστική μάθηση, κατά την οποία οι νευρώνες ανταγωνίζονται μεταξύ τους για το ποιος θα ενεργοποιηθεί, μέχρι να ενεργοποιηθεί κάποιος νικητής νευρώνας (winning neuron ή BMU (Best Matching Unit)).

Αν το σύνολο δεδομένων εισόδου συμβολιστεί ως $\mathbf{x} = \{\mathbf{x_i}: \mathbf{i} = \mathbf{1}, \ldots, \mathbf{D}\}$ και τα βάρη των συνδέσεων μεταξύ δεδομένων εισόδου Ικαι νευρώνων jστον χώρο υπολογισμών ως $\boldsymbol{w_j} = \{\boldsymbol{w_{ji}}: \boldsymbol{j} = \mathbf{1}, \ldots, \boldsymbol{N}; \boldsymbol{i} = \mathbf{1}, \ldots, \boldsymbol{D}\}$

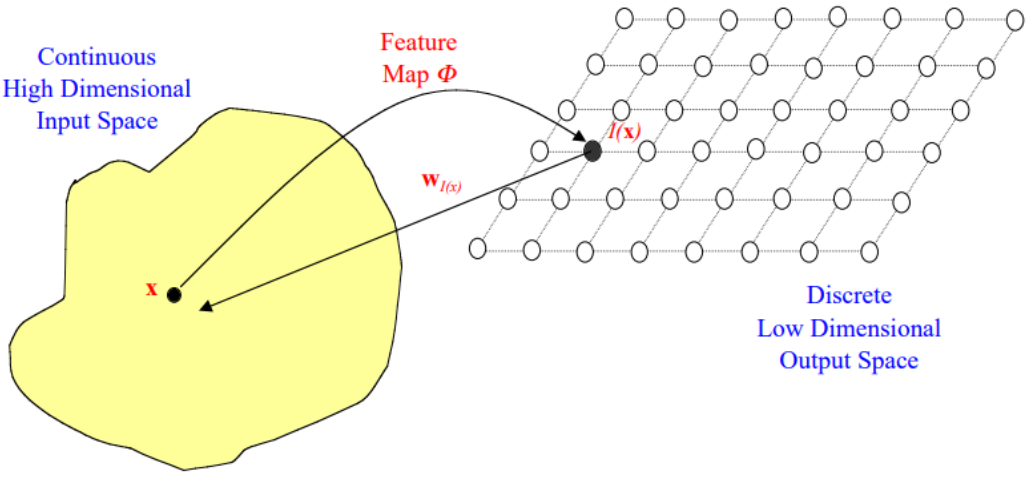
Τότε μπορεί να οριστεί ως συνάρτηση διαχωρισμού η τετραγωνική Ευκλείδεια απόσταση μεταξύ διανύσματος εισόδου xκαι διανύσματος βάρους w_j για κάθε νευρώνα j:

$$\boldsymbol{d_j(x)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{D}} (\boldsymbol{x_i} - \boldsymbol{w_{ji}})^2$$

δηλαδή ο νευρώνας του οποίου το διάνυσμα βάρους είναι πιο κοντά στο διάνυσμα εισόδου, είναι και ο νικητής νευρώνας.

Ως αποτέλεσμα οι νευρώνες αναγκάζονται να οργανωθούν από μόνοι τους, συνεπώς και η ονομασία αυτοοργανώμενου χάρτη (SOM). Στόχος της εκμάθησης του χάρτη είναι η απεικόνιση του πολυ-διάστατου χώρου σε χώρο 2-διαστάσεων, σε ένα τετράγωνο πλέγμα 2-D, όπως φαίνεται και στο Σχήμα 2.33.

Η οργάνωση των x σημείων του χώρου εισόδου σε σημεία του διδιάστατου χώρου εξόδου φαίνεται στα Σχήματα 2.34 και 2.35. Κάθε σημείο του χώρου εξόδου θα αντιστοιχηθεί με κάποιο του χώρου εισόδου.



Σχήμα 2.34: Σχήμα απεικόνισης οργάνωσης δεδομένων εισόδου στο 2-D πλέγμα [31]

Η διαδικασία εκμάθησης του αυτοοργανωμένου χάρτη περιγράφεται στα εξής βήματα [31]:

- *Αρχικοποίηση:* Τυχαία επιλογή τιμών για τα αρχικά διανύσματα βάρους w_j
- *Δειγματοληψία:* Επιλογή κάποιου δείγματος του διανύσματος εισόδου
- *Ταίριασμα:* Η ανταγωνιστική διαδικασία επιλογής του νικητή νευρώνα, μέσω της συνάρτησης διαχωρισμού

$$\boldsymbol{d_j(x) = \sum_{i=1}^{D} (x_i - w_{ji})^2}$$

και δημιουργία της γειτονιάς κόμβων γύρω του. Έπειτα λαμβάνεται υπ'όψιν η πλευρική αντίδραση των γειτόνων, ενός όρου της Νευρολογίας, που ορίζει πως οι κοντινότεροι γείτονες είναι πιο πιθανό να ενεργοποιηθούν από ότι μακρινοί κόμβοι. Με αυτή την αρχή ορίζουμε την πλευρική απόσταση μεταξύ νευρώνων i και j, ως

$$\boldsymbol{T_{j,I(x)} = exp(-S^2_{j,I(x)}/2\sigma^2)}$$

ως την τοπολογική γειτονιά, όπου $I(x)$ ο αριθμός του νικητή νευρώνα.

- *Ενημέρωση:* Ενημέρωση των βαρών του νικητή νευρώνα και της γειτονιάς του, εφαρμόζοντας της εξίσωση

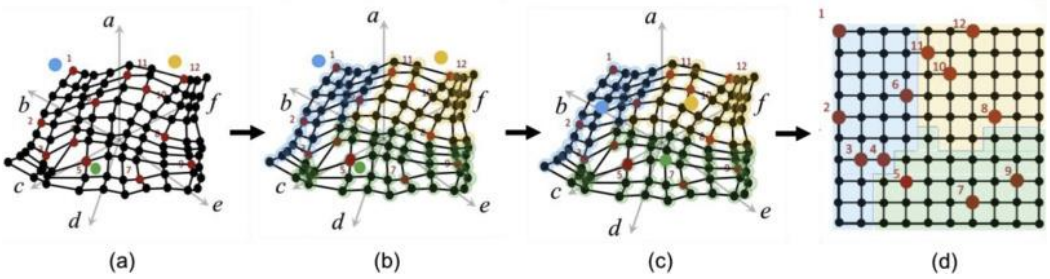
$$\boldsymbol{\Delta w_{ji} = \eta(t) T_{j,I(x)}(t)(x_i - w_{ji})}$$

στην οποία υπάρχει περίοδος χρόνου t εξαρτώμενος του ρυθμού εκμάθησης

$$\boldsymbol{\eta(t) = \eta_0 exp(-t/\tau_\eta)}$$

και οι ενημερώσεις πραγματοποιούνται πολλές περιόδους.

- *Επανάληψη:* Επανάληψη των τριών προηγούμενων βημάτων μέχρι η τοπογραφία του χάρτη να σταματήσει να αλλάζει.

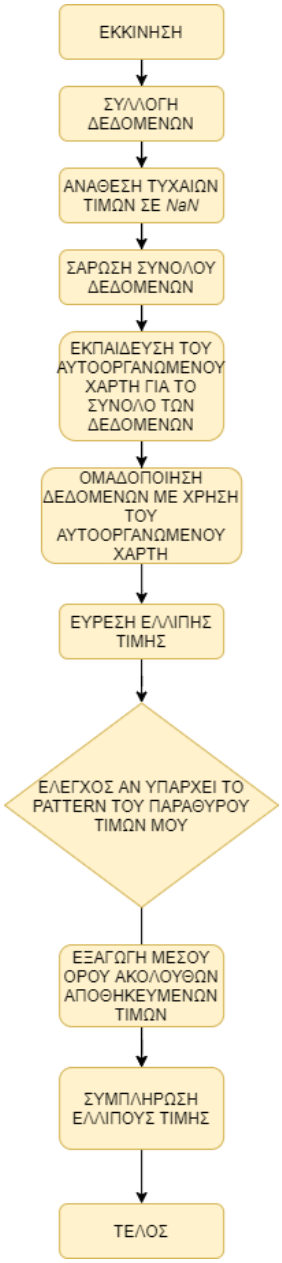


Σχήμα 2.35: Αποτέλεσμα ομαδοποίησης με αυτοοργανωμένο χάρτη σε 2-D πλέγμα [32]

Έχοντας ένα εκπαιδευμένο μοντέλο αυτοοργανωμένου χάρτη (Self-Organizing Map) μπορούμε να εκτιμήσουμε ελλιπείς τιμές του συνόλου δεδομένων.

Για κάθε ελλιπή τιμή υπολογίζεται BMU (Best Matching Unit). Οι ελλιπείς τιμές αγνοούνται κατά την επιλογή του BMU ενώ η τιμή με την οποία συμπληρώνεται είναι η αντίστοιχη BMU τιμή.

Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.36.



Σχήμα 2.36: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήση αυτοοργανωμένου χάρτη (Self-Organizing Map)

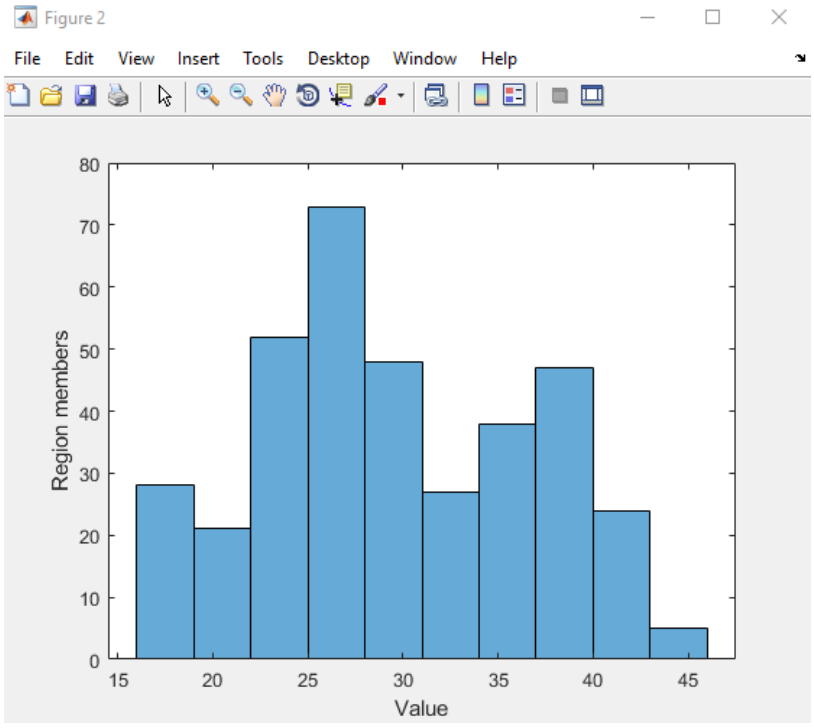
2.4 Μέθοδοι με χρήση πιθανοτήτων

2.4.1. Μέθοδος συμπλήρωσης με χρήσηκατανομής συχνοτήτων στήλης και μέγιστη πιθανότητα

Στην μέθοδο αυτή πραγματοποιείται διάσπαση της στήλης σε καθορισμένο αριθμό περιοχών (bins), όπου βρέθηκε ελλιπής τιμή. Με βάση τις περιοχές αυτές και έχοντας τα όρια των περιοχών αυτών γίνεται κατανομή συχνοτήτων των τιμών, βρίσκεται δηλαδή σε ποια περιοχή βρίσκονται οι περισσότερες τιμές. Η περιοχή δηλαδή με την μεγαλύτερη πυκνότητα τιμών, θα είναι και αυτή η οποία θα ληφθεί υπ'όψιν για την συμπλήρωση της ελλιπούς τιμής.

Η κατανομή συχνοτήτων είναι μια αναπαράσταση είτε σε μορφή γραφήματος είτε σε μορφή πίνακα των αριθμών των δεδομένων σε κάποιες περιοχές.

Στην περίπτωσή μας η κατανομή συχνοτήτων αναπαρίσταται γραφικά μέσω ιστογράμματος στο Matlab, όπως φαίνεται και στο Σχήμα 2.37.



Σχήμα 2.37: Σχήμα ιστογράμματος με την κατανομή τιμών σε στήλη που βρέθηκε ελλιπής τιμή. Στον κάθετο άξονα τα μέλη της κάθε περιοχής και στον οριζόντια τα όρια των περιοχών.

Αρχικά γίνεται σάρωση του συνόλου δεδομένων μέχρι να βρεθεί κάποια ελλιπής τιμή. Όταν βρεθεί γίνεται διάσπαση των δεδομένων της στήλης σε προκαθορισμένο αριθμό περιοχών (bins).

Έστω ελλιπής τιμή $x(i,j)$, γίνεται η διάσπαση των μη ελλιπών τιμών $x(:,j)$ σε περιοχές, από τις οποίες κρατούνται τα όρια. Στη συνέχεια υπολογίζεται ο πίνακας πιθανοτήτων γνωρίζοντας πόσες τιμές βρίσκονται σε κάθε περιοχή με το θεώρημα κλασσικής πιθανότητας:

$$\boldsymbol{P(x \in A) = \frac{\#A}{\#\Omega}}$$

δηλαδή η πιθανότητα να ανήκει σε μια περιοχή, ως ο αριθμός εμφανίσεων στην περιοχή προς το σύνολο των τιμών.

Δημιουργείται ο πίνακας πιθανοτήτων, όπως φαίνεται στο παράδειγμα του Σχήματος 2.38 και κρατείται η μέγιστη πιθανότητα περιοχής, στην οποία δηλαδή η τιμή είναι πιθανότερο να ανήκει.

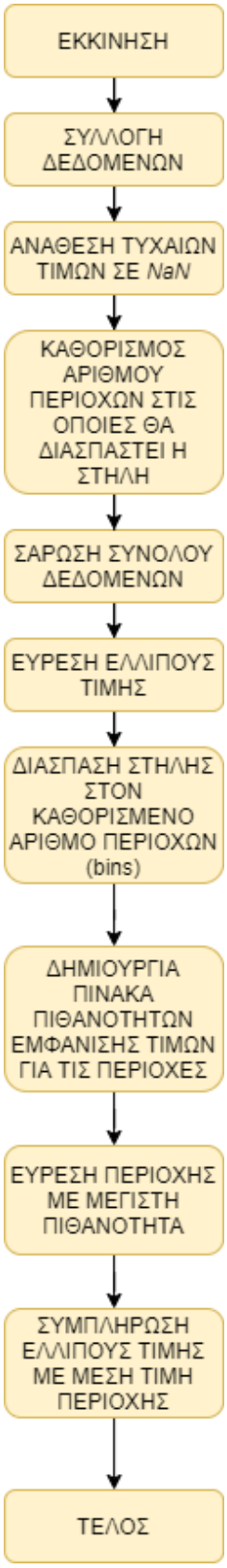
```
probability_array =

    0.0028    0.0110    0.0248    0.1212    0.1543    0.1543    0.2562    0.1653    0.0882    0.0220
```

Σχήμα 2.38: Παράδειγμα εξαγωγής πίνακα πιθανοτήτων για κάθε περιοχή στην οποία διαιρέθηκε το σύνολο δεδομένων.

Τέλος, ή ελλιπής τιμή συμπληρώνεται με την μέση τιμή της περιοχής με την μεγαλύτερη πιθανότητα εμφάνισης.

$$\boldsymbol{x(i,j) = \frac{1}{2}(k + l)}$$

όπου k και l το αριστερό και δεξί όριο της περιοχής με την μεγαλύτερη πιθανότητα αντίστοιχα.Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.39.



Σχήμα 2.39: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήσηκατανομής συχνοτήτων στήλης και μέγιστη πιθανότητα

2.4.2. Μέθοδος συμπλήρωσης με χρήση κατανομής συχνοτήτων στήλης και αθροιστική πιθανότητα

Μια παραλλαγή της παραπάνω μεθόδου, αποτελεί η κατανομή συχνοτήτων και συμπλήρωση με την αθροιστική πιθανότητα. Στην μέθοδο αυτή, όμοια με προηγουμένως γίνεται διάσπαση της στήλης στην οποία βρέθηκε ελλιπής τιμή σε καθορισμένο αριθμό περιοχών (bins).

Στη συνέχεια γίνεται κατανομή συχνοτήτων, πόσες τιμές δηλαδή της στήλης βρίσκονται σε κάθε περιοχή και υπολογίζεται ο πίνακας πιθανότητων που περιέχει τις πιθανότητας εύρεσης σε κάθε περιοχή.

Η συμπλήρωση της ελλιπούς τιμής γίνεται με βάση την αθροιστική συχνότητα, έχοντας τα μέσα των περιοχών και την πιθανότητα για κάθε περιοχή και προσθέτοντας τα γινόμενα τους.

$$\boldsymbol{F_i = \sum_{j=1}^{i} f_j}$$

όπου $x_j \leq x_i$για $j \leq i$ (αθροιστική συχνότητα)

$$P_i = \sum_{j=1}^{i} p_j$$

όπου $x_j \leq x_i$για $j \leq i$ (αθροιστική πιθανότητα)

Αρχικά γίνεται σάρωση του συνόλου δεδομένων μέχρι να βρεθεί κάποια ελλιπής τιμή. Όταν βρεθεί γίνεται διάσπαση των δεδομένων της στήλης σε προκαθορισμένο αριθμό περιοχών (bins).

Έστω ελλιπής τιμή $x(i,j)$, γίνεται η διάσπαση των μη ελλιπών τιμών $x(:,j)$ σε περιοχές, από τις οποίες κρατούνται τα όρια. Στη συνέχεια υπολογίζεται ο πίνακας πιθανοτήτων (Σχήμα 2.40) γνωρίζοντας πόσες τιμές βρίσκονται σε κάθε περιοχή με το θεώρημα κλασσικής πιθανότητας:

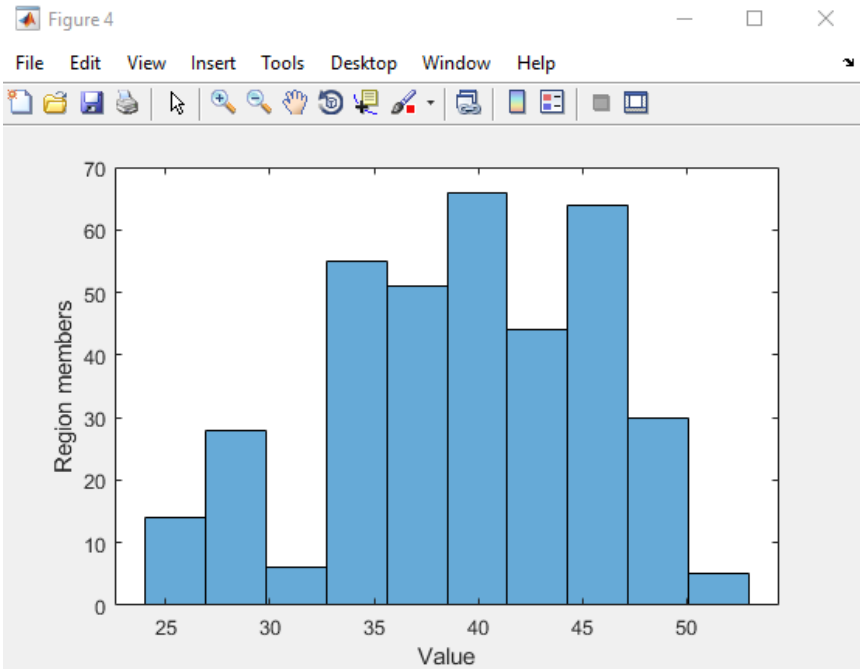
$$\boldsymbol{P(x \in A) = \frac{\#A}{\#\Omega}}$$

δηλαδή η πιθανότητα να ανήκει σε μια περιοχή, ως ο αριθμός εμφανίσεων στην περιοχή προς το σύνολο των τιμών.

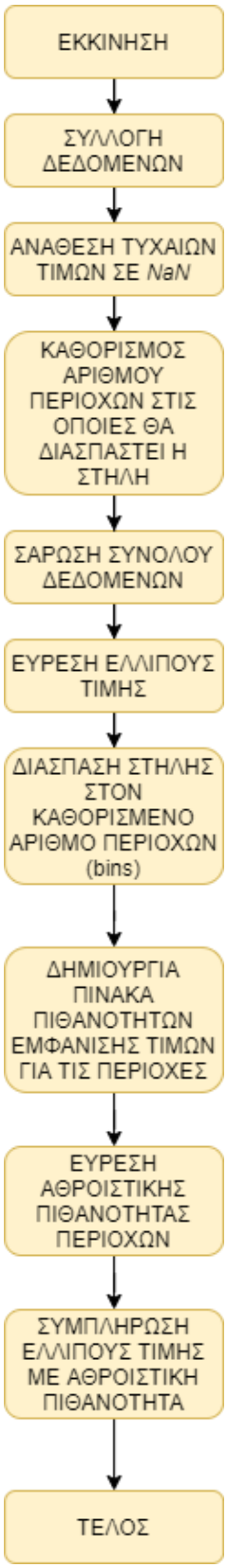
```
edges =

  Columns 1 through 10

         0    5.6000   11.2000   16.8000   22.4000   28.0000   33.6000   39.2000   44.8000   50.4000

  Column 11

   56.0000


probability_array =

    0.2617         0         0    0.0110    0.0992    0.1460    0.1543    0.1295    0.0909    0.1074
```

Σχήμα 2.40: Παράδειγμα εξαγωγής πίνακα πιθανοτήτων για κάθε περιοχή στην οποία διαιρέθηκε το σύνολο δεδομένων και των ορίων των περιοχών.

Πολλαπλασιάζοντας τις πιθανότητες κάθε περιοχής με το μέσο της περιοχής και αθροίζοντας τα επιμέρους αυτά γινόμενα, προκύπτει η αθροιστική πιθανότητα με την οποία συμπληρώνεται τελικά η ελλιπής τιμή. Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.42.



Σχήμα 2.41: Σχήμα ιστογράμματος με την κατανομή τιμών σε στήλη που βρέθηκε ελλιπής τιμή. Στον κάθετο άξονα τα μέλη της κάθε περιοχής και στον οριζόντια τα όρια των περιοχών.



Σχήμα 2.42: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήσηκατανομής συχνοτήτων στήλης και αθροιστική πιθανότητα.

2.4.3. Μέθοδος συμπλήρωσης με χρήση κατανομής συχνοτήτων παραθύρου τιμών και μέγιστη πιθανότητα

Στην μέθοδο αυτή επιλέγεται συγκεκριμένο παράθυρο τιμών, πριν και μετά την ελλιπή τιμή, το οποίο και θα διαιρεθεί σε καθορισμένο αριθμό περιοχών (bins). Με βάση αυτό το παράθυρο τιμών, θα γίνει και η κατανομή συχνοτήτων και ο υπολογισμών πιθανοτήτων εύρεσης σε κάθε περιοχή.

Η μέθοδος αυτή αποτελεί παραλλαγή της προηγούμενης μεθόδου συμπλήρωσης με χρήση στήλης και μέγιστη πιθανότητα. Είναι όμως πιο βολική, ειδικά στην περίπτωσή μας, όπου το σύνολο δεδομένων που μελετάται είναι ημερολογιακά δεδομένα.

Δίνεται λοιπόν η δυνατότητα να μελετηθούν στοιχεία περιορισμένου χρονικού και ημερολογιακού διαστήματος, όπως για παράδειγμα κατανομή σε περίοδο μιας εβδομάδας, ενός μήνα κ.ο.κ.

Αρχικά γίνεται σάρωση του συνόλου δεδομένων μέχρι να βρεθεί κάποια ελλιπής τιμή. Όταν βρεθεί γίνεται διάσπαση των δεδομένων της στήλης σε προκαθορισμένο αριθμό περιοχών (bins).

Έστω ελλιπής τιμή $x(i,j)$, γίνεται η διάσπαση των μη ελλιπών τιμών $x(i-window,j)$ και $x(i+window,j)$ σε περιοχές (bins), από τις οποίες κρατούνται τα όρια. Στη συνέχεια υπολογίζεται ο πίνακας πιθανοτήτων γνωρίζοντας πόσες τιμές βρίσκονται σε κάθε περιοχή με το θεώρημα κλασσικής πιθανότητας:

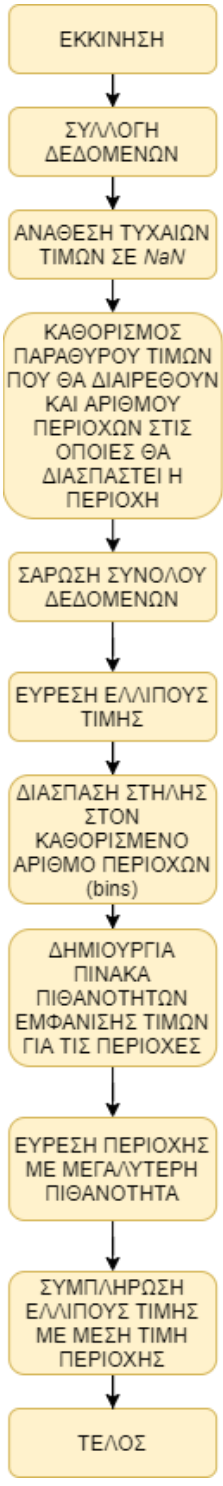
$$P(x \in A) = \frac{\#A}{\#\Omega}$$

δηλαδή η πιθανότητα να ανήκει σε μια περιοχή, ως ο αριθμός εμφανίσεων στην περιοχή προς το σύνολο των τιμών.

Δημιουργείται ο πίνακας πιθανοτήτων, όπως φαίνεται στο παράδειγμα του Σχήματος 2.43 και κρατείται η μέγιστη πιθανότητα περιοχής, με τη μέση τιμή της οποίας συμπληρώνεται η ελλιπής τιμή. Η μέθοδος περιγράφεται σχηματικά στο Σχήμα 2.44.

```
probability_array =

    0.0028    0.0110    0.0248    0.1212    0.1543    0.1543    0.2562    0.1653    0.0882    0.0220
```

Σχήμα 2.43: Παράδειγμα εξαγωγής πίνακα πιθανοτήτων για κάθε περιοχή στην οποία διαιρέθηκε το σύνολο δεδομένων.



Σχήμα 2.44: Διάγραμμα ροής μεθοδολογίας συμπλήρωσης με χρήσηκατανομής συχνοτήτων παραθύρου τιμών και μέγιστη πιθανότητα

ΚΕΦΑΛΑΙΟ3

ΠΑΡΟΥΣΙΑΣΗ ΚΑΙ ΑΝΑΛΥΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

3.1 Γενική περιγραφή μεθόδου λήψης και ανάλυσης αποτελεσμάτων

Στα πλαίσια της εργασίας οι μέθοδοι εφαρμόστηκαν σε δυο διαφορετικά σύνολα δεδομένων. Το ένα περιέχει ημερολογιακά δεδομένα μετρήσεων ηλεκτρικού φορτίου για την περιοχή Νέα Ελβετία Θεσσαλονίκης για το έτος 2011, με μέγεθος 24 επί 365 = 8760 τιμές. Το δεύτερο περιέχει ημερήσια δεδομένα για την ταχύτητα ανέμου στην περιοχή του Βόλου για τα έτη 2018-2020, με μέγεθος δηλαδή 365 και 366 = 731 τιμές.

Τα αρχεία αυτά περάστηκαν στο προγραμματιστικό περιβάλλον του Matlab με τη μορφή πινάκων. Έπειτα προγραμματίστηκε να τεθούν τυχαίες τιμές σε ελλιπείς (NaN στη μορφή του Matlab), τιμές με ποσοστό 10, 20 και 30% για καθένα από τα δυο αρχεία. Ο αριθμός των ελλιπών τιμών και οι θέσεις τους εμφανίζονται κατά την αρχή του τρεξίματος της κάθε μεθόδου.

Στη συνέχεια εφαρμόστηκαν στα σύνολα δεδομένων με τις ελλιπείς τιμές οι μέθοδοι, κατά τις οποίες σαρώνεται κάθε πίνακας και μόλις συναντηθεί κάποια ελλιπής τιμή, αυτή συμπληρώνεται ανάλογα τη μέθοδο.

Στο τέλος της κάθε μεθόδου τυπώνεται μήνυμα πως όλες οι ελλιπείς τιμές έχουν συμπληρωθεί και δεν έχει μείνει κάποια, και υπολογίζεται η ακρίβεια και η απόδοση της μεθόδου, καθώς και ο χρόνος εκτέλεσης βάσει δεικτών.

3.2 Δείκτες μετρήσεων αποτελεσμάτων

Η αξιολόγηση των μεθόδων όσον αφορά την ακρίβεια και την αποτελεσματικότητα εφαρμογής τους, γίνεται με χρήση ορισμένων δεικτών:

- Μέσο απόλυτο σφάλμα (MAE: Mean Absolute Error) :

  Το μέσο απόλυτο σφάλμα είναι μια μέτρηση λαθών μεταξύ παρατηρήσεων του αρχικού συνόλου δεδομένων και του συμπληρωμένου με χρήση της μεθόδου συμπλήρωσης.

  Είναι ο αριθμητικός μέσος του αθροίσματος των απόλυτων τιμών των λαθών για κάθε ένα δεδομένο του συνόλου.

$$\boldsymbol{MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}}$$

[33]

- Μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE: Mean Absolute Percentage Error) :

$$\mathbf{MAPE = \frac{100}{n} \sum_{t=1}^{n} |\frac{A_t - F_i}{A_i}|}$$

[34]

όπου A_t η πραγματική τιμή και F_i η προβλεπόμενη (συμπληρωμένη) τιμή.

- Μέσο σχετικό σφάλμα (MRE: Mean Relative Error) :

$$\mathbf{MRE = \frac{\sum_{i=1}^{n} y_i - x_i}{n} = \frac{\sum_{i=1}^{n} e_i}{n}}$$

- Μέσο σχετικό ποσοστιαίο σφάλμα (MRPE: Mean Relative Percentage Error) :

$$\boldsymbol{MRPE = \frac{100}{n} \sum_{t=1}^{n} \frac{A_i - F_i}{A_i}}$$

- Χρόνος εκτέλεσης συμπλήρωσης όλων των ελλιπών τιμών :

  Μέτρηση με την χρήση της εντολής tic του προγραμματιστικού περιβάλλοντος του Matlab. [35]

3.3 Παρουσίαση αποτελεσμάτων της κάθε μεθόδου

3.3.1. Μετρήσεις μεθόδων συμπλήρωσης με χρήση της μέσης τιμής

Στους Πίνακες 3.1 και 3.2 απεικονίζονται οι δείκτες μετρήσεων για τις μεθόδους συμπλήρωσης με χρήση μέσης τιμής, δηλαδή συμπλήρωση με μέση τιμή στήλης, μέση τιμή δυο προηγουμένων τιμών και μέση τιμή επόμενης και προηγούμενης τιμής.

Σύμφωνα με τα αποτελέσματα παρατηρείται μικρή αύξηση του απόλυτου ποσοστιαίου σφάλματος, όσο ανεβαίνει το ποσοστό ελλιπών τιμών από 10% σε 20% και 30%. Ο χρόνος εκτέλεσης και συμπλήρωσης όλων των ελλιπών τιμών παραμένει πολύ μικρός ανεξάρτητα του ποσοστού ελλιπών τιμών.

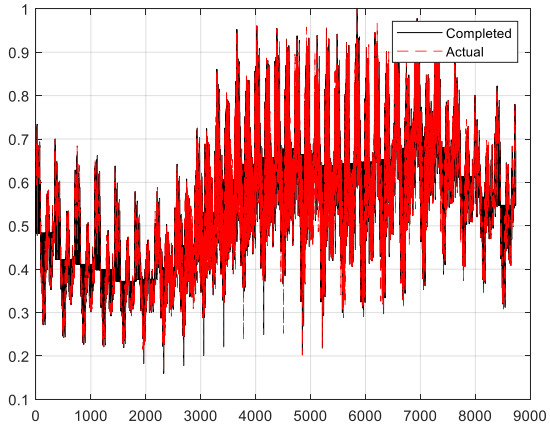
Στα Σχήματα 3.1 έως και 3.9 παρουσιάζονται γραφικά τα αποτελέσματα των μεθόδων με χρήση μέσης τιμής, με σύγκρισεις συνολικων και ημερήσιων χρονοσειρών.

Πίνακας 3.1: Τιμές δεικτών των μεθόδων συμπλήρωσης με χρήση μέσης τιμής, για το σύνολο δεδομένων ηλεκτρικού φορτίου.

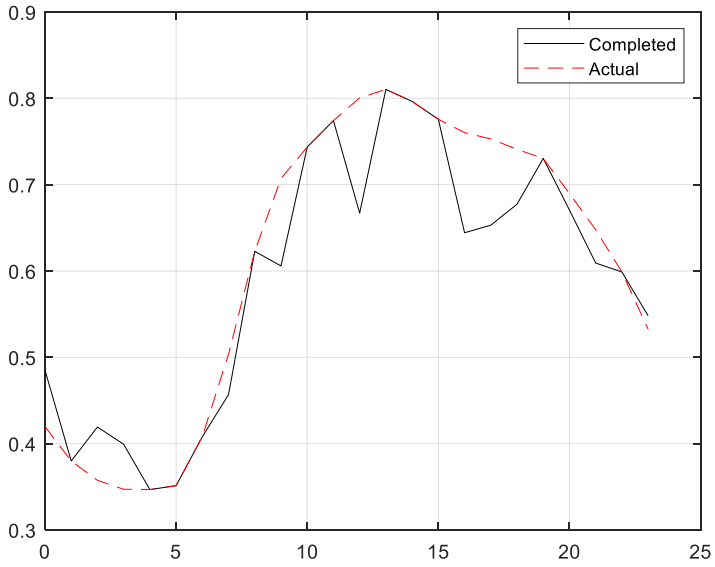
	Column Mean			Mean Two Previous			Mean Previous Next		
Missing Values (%)	10%	20%	30%	10%	20%	30%	10%	20%	30%
MAE	0.090	0.019	0.029	0.006	0.012	0.018	0.004	0.008	0.015
MAPE	1.879	3.753	5.606	1.048	2.079	3.303	0.824	1.615	2.690
MRE	-0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
MRPE	-0.630	-0.909	-1.483	-0.215	-0.140	-0.568	-0.117	-0.223	-0.366
Exec. Time	0.045	0.025	0.027	0.011	0.010	0.015	0.071	0.137	0.206

Πίνακας 3.2: Τιμές δεικτών των μεθόδων συμπλήρωσης με χρήση μέσης τιμής, για το σύνολο δεδομένων ταχύτητας ανέμου.

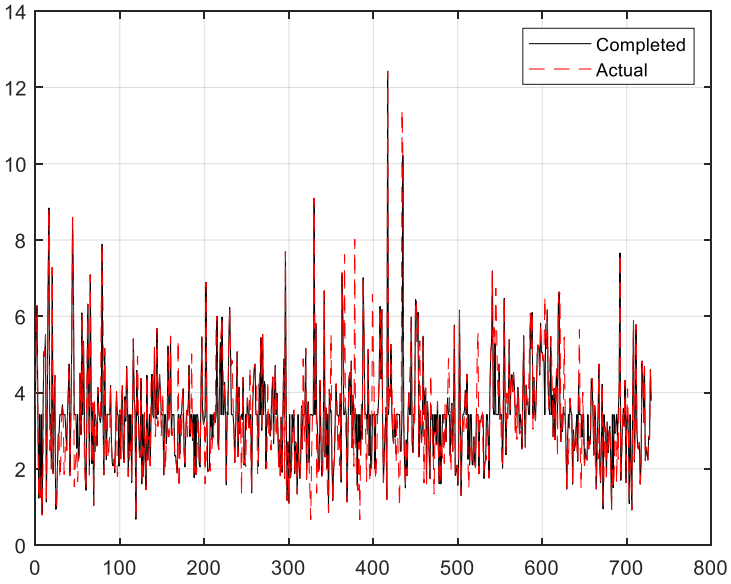
	Column Mean			Mean Two Previous			Mean Previous Next		
Missing Values (%)	10%	20%	30%	10%	20%	30%	10%	20%	30%
MAE	0.131	0.215	0.332	0.143	0.210	0.389	0.107	0.206	0.350
MAPE	5.501	8.230	13.819	4.742	7.567	14.841	3.408	7.703	13.038
MRE	-0.002	0.001	-0.048	0.007	0.033	-0.033	0.033	0.026	-0.027
MRPE	-3.233	-4.285	-8.339	-1.974	-1.948	-7.745	-0.131	-2.582	-6.246
Exec. Time	0.007	0.003	0.003	0.003	0.003	0.004	0.011	0.013	0.017



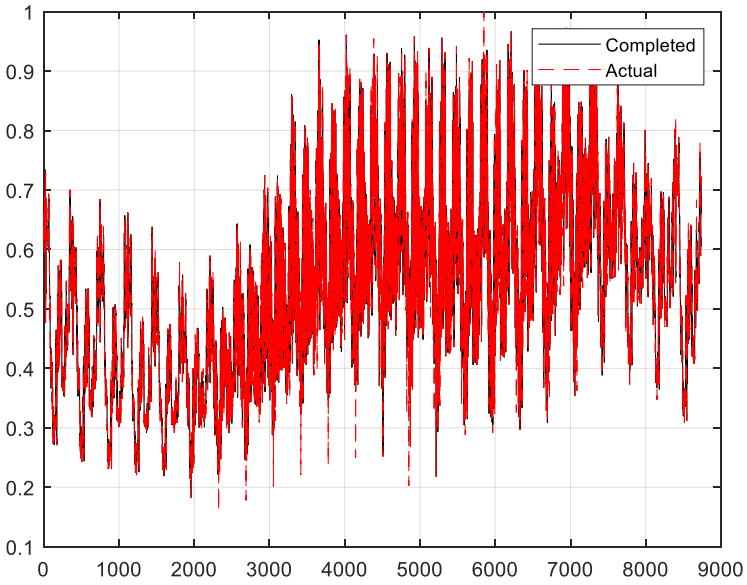
Σχήμα 3.1: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση μέσης τιμής στήλης.



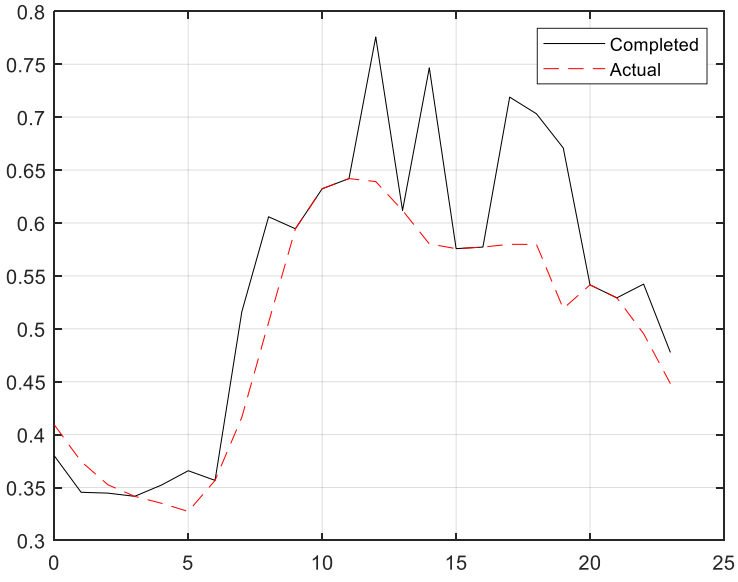
Σχήμα 3.2: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση μέσης τιμής στήλης.



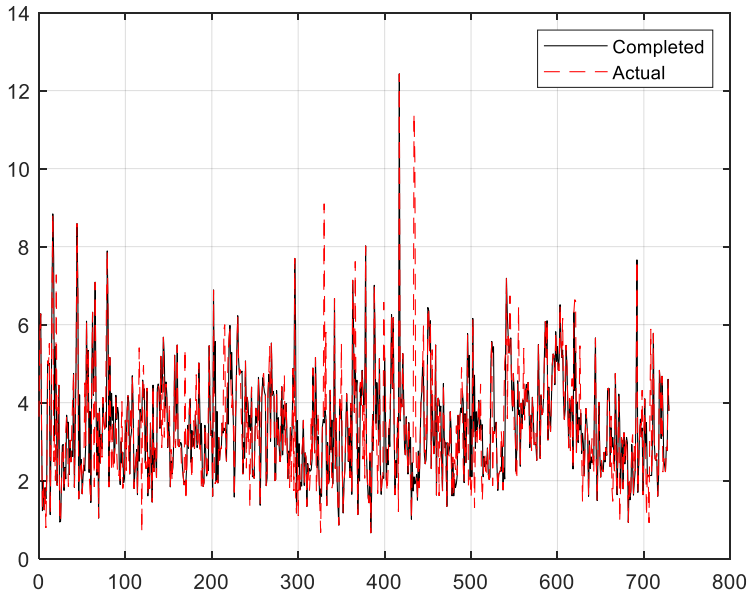
Σχήμα 3.3: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ταχύτητας ανέμου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση μέσης τιμής στήλης.



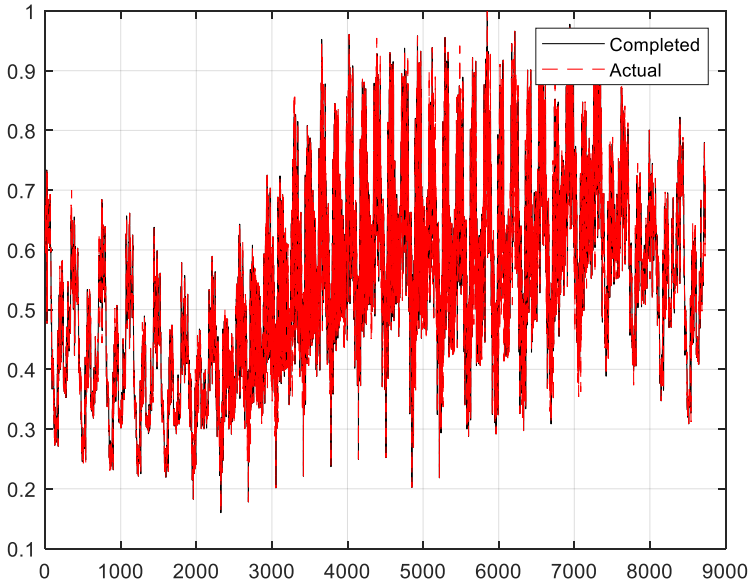
Σχήμα 3.4: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση μέσης τιμής των δυο προηγουμένων τιμών.



Σχήμα 3.5: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση μέσης τιμής των δυο προηγουμένων τιμών.

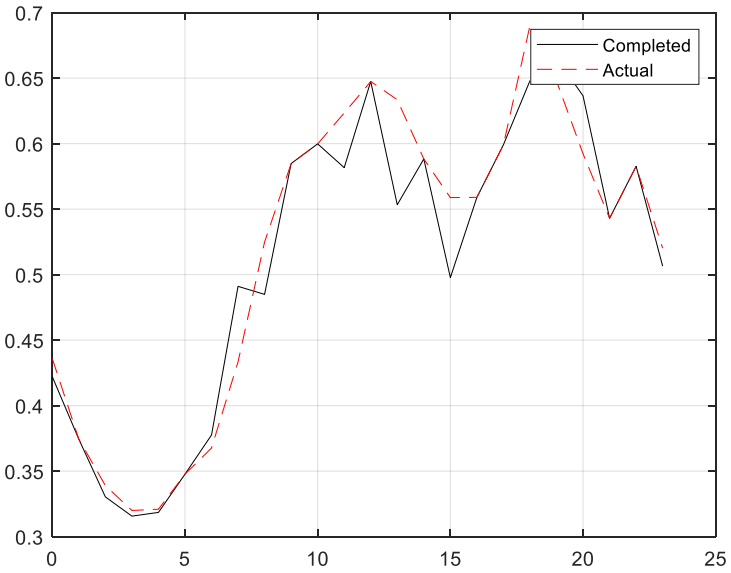


Σχήμα 3.6: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ταχύτητας ανέμου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση μέσης τιμής των δυο προηγουμένων τιμών.

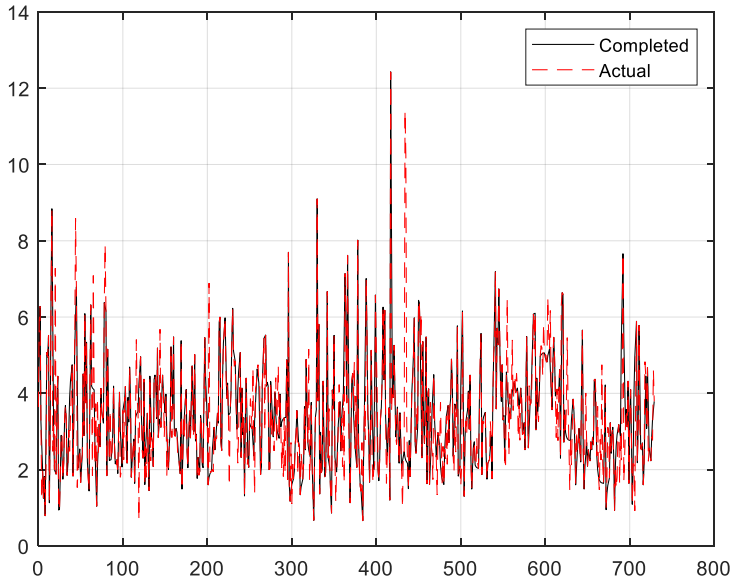


.

Σχήμα 3.7: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση μέσης τιμής προηγούμενης και επόμενης τιμής.



Σχήμα 3.8: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση μέσης τιμής προηγούμενης και επόμενης τιμής.



Σχήμα 3.9: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ταχύτητας ανέμου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση μέσης τιμής προηγούμενης και επόμενης τιμής.

3.3.2.Μετρήσεις μεθόδου συμπλήρωσης με χρήση της ελάχιστης Ευκλείδειας απόστασης

Στους Πίνακες 3.3 και 3.4 παρουσιάζονται οι μετρήσεις τις μεθόδου συμπλήρωσης με ελάχιστη Ευκλείδεια απόσταση για κάθε σύνολο δεδομένων.

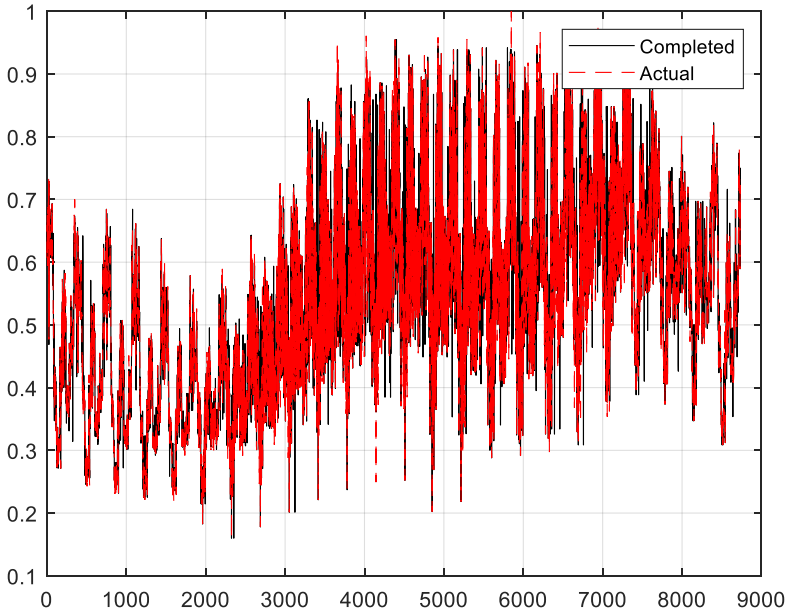
Πίνακας 3.3: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση ελάχιστης Ευκλείδειας απόστασης, για το σύνολο δεδομένων ηλεκτρικού φορτίου.

Missing Values (%)	10%	20%	30%
MAE	0.007	0.014	0.022
MAPE	1.199	2.616	4.043
MRE	0.001	-0.001	-0.001
MRPE	-0.021	-0.388	-0.581
Exec. Time	35.906	74.037	109.821

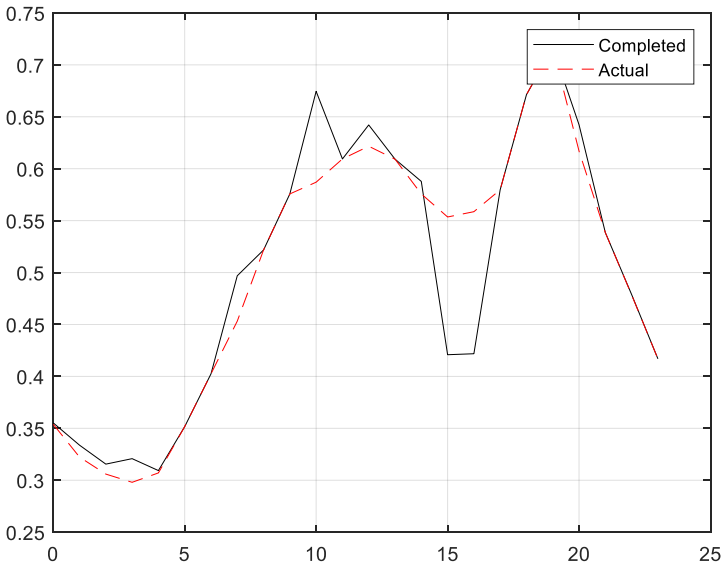
Πίνακας 3.4: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση ελάχιστης Ευκλείδειας απόστασης, για το σύνολο δεδομένων ταχύτητας ανέμου.

Missing Values (%)	10%	20%	30%
MAE	0.159	0.314	0.405
MAPE	5.166	9.273	14.051
MRE	-0.011	0.044	0.056
MRPE	-1.659	-1.734	-3.345
Exec. Time	4.569	9.082	14.532

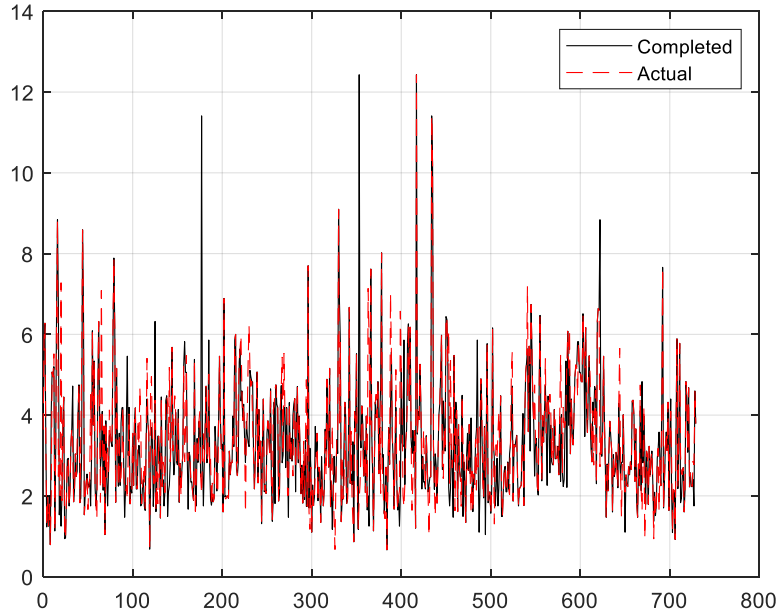
Παρατηρείται από τους Πίνακες 3.3 και 3.4 σημαντική αύξηση του σφάλματος μεταξύ του ποσοστού ελλιπών τιμών 10, 20 και 30% καθώς και σημαντική αύξηση του χρόνου εκτέλεσης. Στα Σχήματα 3.10 έως 3.12 παρουσιάζονται γραφικά τα αποτελέσματα, με συγκρίσεις χρονοσειρών.



Σχήμα 3.10: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με ελάχιστη Ευκλείδεια απόσταση.



Σχήμα 3.11: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με ελάχιστη Ευκλείδεια απόσταση.



Σχήμα 3.12: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ταχύτητας ανέμου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση ελάχιστης Ευκλείδειας απόστασης.

3.3.3. Μετρήσεις μεθόδου συμπλήρωσης με χρήση επιλεγόμενης προηγούμενης τιμής

Στους Πίνακες 3.5 και 3.6 παρουσιάζονται οι μετρήσεις μεθόδου με επιλεγόμενη προηγούμενη τιμή για κάθε ένα σύνολο δεδομένων.

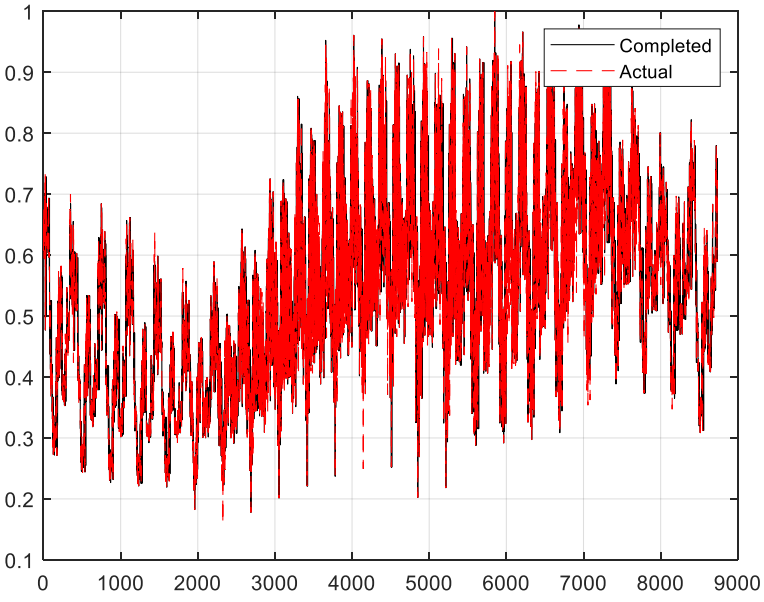
Πίνακας 3.5: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση επιλεγόμενης προηγούμενης τιμής, για το σύνολο δεδομένων ηλεκτρικού φορτίου.

Missing Values (%)	10%	20%	30%
MAE	0.008	0.016	0.024
MAPE	1.443	3.006	4.568
MRE	-0.001	-0.001	0.001
MRPE	-0.208	-0.479	-0.5682
Exec. Time	0.015	0.012	0.027

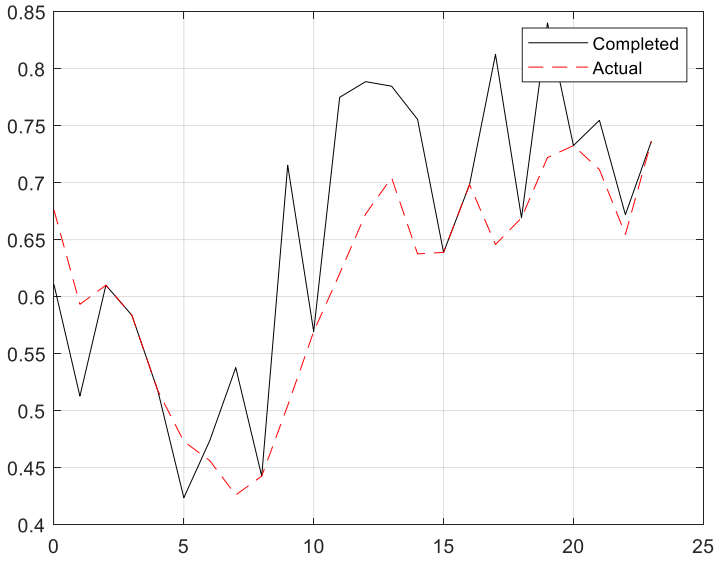
Πίνακας 3.6: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση επιλεγόμενης προηγούμενης τιμής, για το σύνολο δεδομένων ταχύτητας ανέμου.

Missing Values (%)	10%	20%	30%
MAE	0.155	0.309	0.453
MAPE	5.391	10.635	15.171
MRE	-0.005	0.0109	0.024
MRPE	-1.992	-3.629	-4.339
Exec. Time	0.004	0.004	0.004

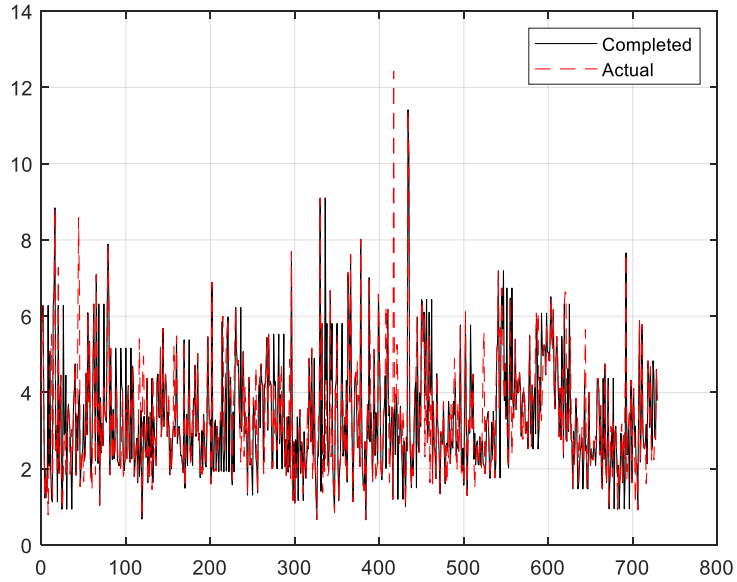
Τα αποτελέσματα των Πινάκων 3.5 και 3.6 προέκυψαν με επιλεγόμενη προηγούμενη τιμή σε απόσταση 7, για την τιμή δηλαδή της προηγούμενης εβδομάδας. Παρατηρείται ελάχιστος χρόνος εκτέλεσης, λόγω της άμεσης συμπλήρωσης χωρίς υπολογιστικό κόστος. Επίσης το μέγεθος του σφάλματος μειώνεται ελάχιστα στο πρώτο σύνολο δεδομένων όσο ανεβαίνει το ποσοστό ελλιπών τιμών, σε αντίθεση με το δεύτερο σύνολο δεδομένων που αυξάνεται σημαντικά, λόγω διαφοράς τιμών. Στα Σχήματα 3.13 έως 3.15 παρουσιάζονται γραφικά τα αποτελέσματα με συγκρίσεις χρονοσειρών.



Σχήμα 3.13: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με επιλεγόμενη προηγούμενη τιμή.



Σχήμα 3.14: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με επιλεγόμενη προηγούμενη τιμή.



Σχήμα 3.15: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ταχύτητας ανέμου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με επιλεγόμενη προηγούμενη τιμή.

3.3.4. Μετρήσεις μεθόδου συμπλήρωσης με χρήση γραμμικής παρεμβολής

Στους Πίνακες 3.7 και 3.8 παρουσιάζονται οι τιμές δεικτών της μεθόδου συμπλήρωσης με γραμμική παρεμβολή.

Πίνακας 3.7: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση γραμμικής παρεμβολής, για το σύνολο δεδομένων ηλεκτρικού φορτίου.

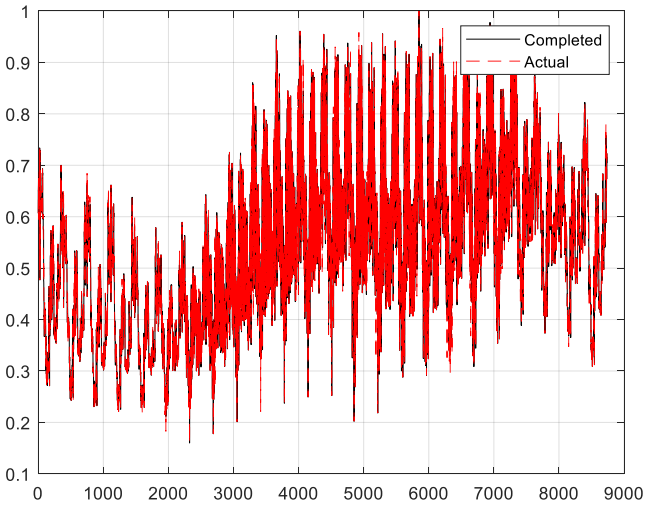
Missing Values (%)	10%	20%	30%
MAE	0.005	0.011	0.017
MAPE	0.924	1.897	3.019
MRE	-0.001	0.001	0.001
MRPE	-0.167	-0.220	-0.252
Exec. Time	1.012	1.089	1.252

Πίνακας 3.8: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση γραμμικής παρεμβολής, για το σύνολο δεδομένων ταχύτητας ανέμου.

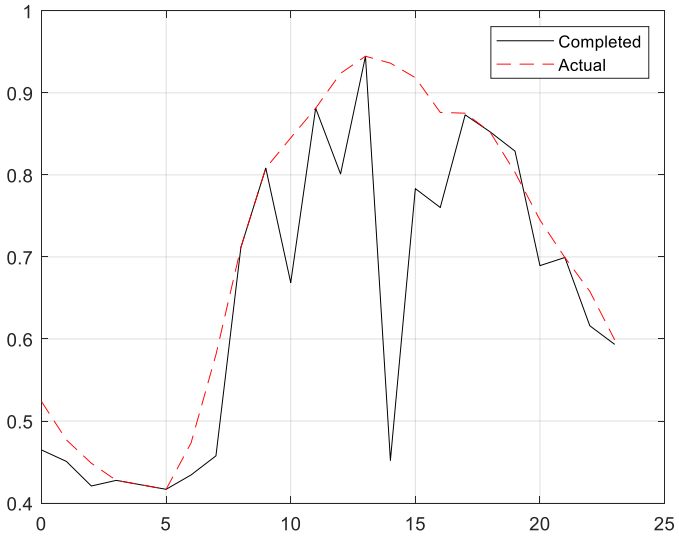
Missing Values (%)	10%	20%	30%
MAE	0.116	0.264	0.363
MAPE	5.183	9.387	13.364
MRE	-0.014	0.004	-0.012
MRPE	-3.127	-3.617	-5.993
Exec. Time	0.066	0.099	0.104

Παρατηρούμε μικρή αύξηση του χρόνου εκτέλεσης όσο αυξάνεται το ποσοστό ελλιπών τιμών, ωστόσο σημαντική αύξηση του λάθους ειδικά στο δεύτερο σύνολο δεδομένων.

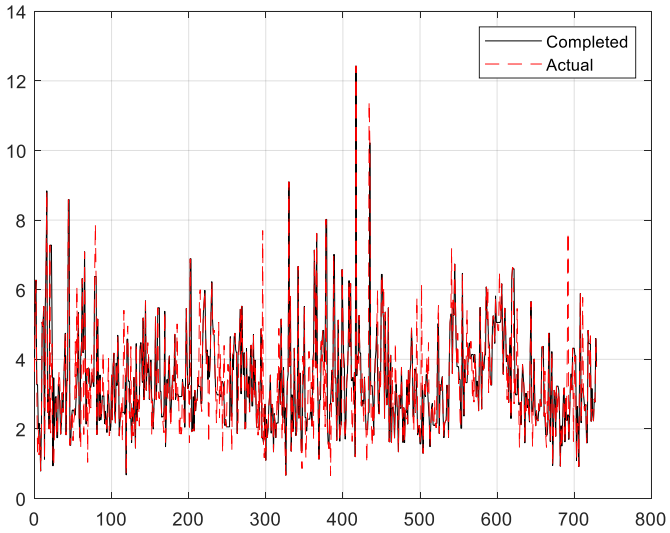
Τα αποτελέσματα παρουσιάζονται γραφικά στα Σχήματα 3.16 έως 3.18 με συγκρίσεις χρονοσειρών για καθέ ένα απο τα σύνολα δεδομένων.



Σχήμα 3.16: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με γραμμική παρεμβολή.



Σχήμα 3.17: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με γραμμική παρεμβολή.



Σχήμα 3.18: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ταχύτητας ανέμου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με γραμμική παρεμβολή.

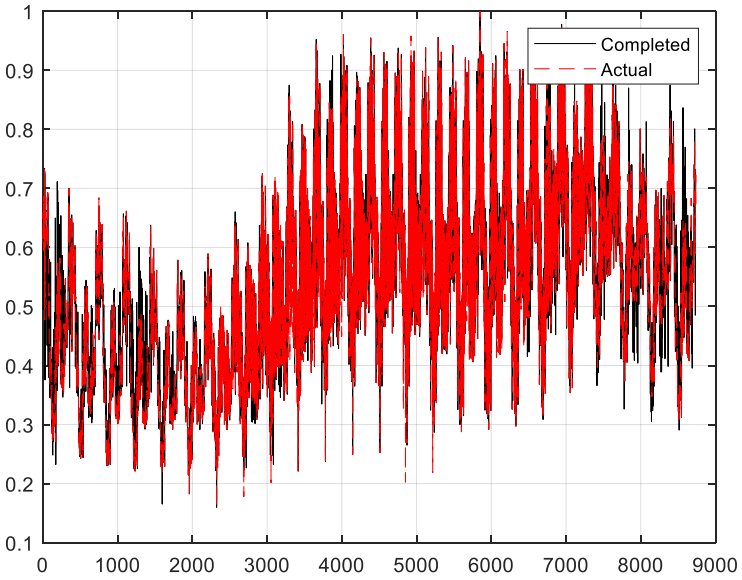
3.3.5. Μετρήσεις μεθόδου συμπλήρωσης με χρήση μέσης τιμής κ-κοντινότερων γειτόνων

Στον Πίνακα 3.9 παρουσιάζονται οι τιμές δεικτών για την μέθοδο συμπλήρωσης με χρήση μέσης τιμής κ-κοντινότερων γειτόνων.

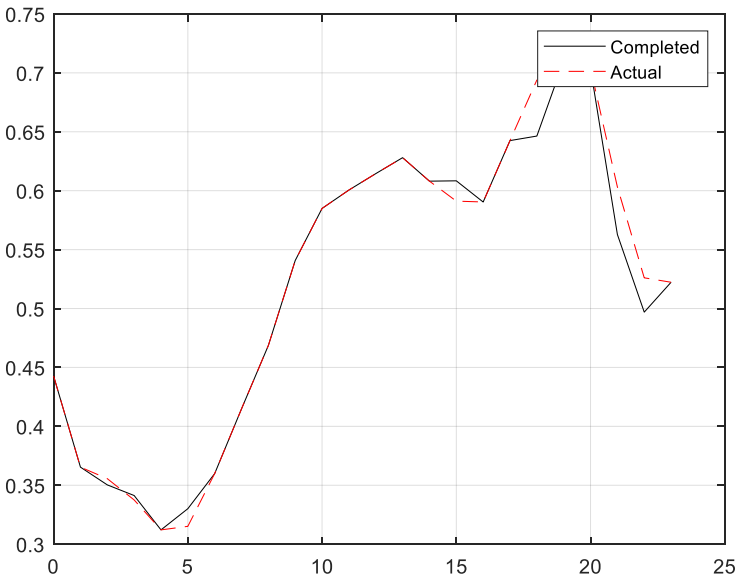
Πίνακας 3.9: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση μέσης τιμής κ-κοντινότερων γειτόνων, για το σύνολο δεδομένων ηλεκτρικού φορτίου.

Missing Values (%)	10%	20%	30%
MAE	0.002	0.009	0.014
MAPE	0.469	1.778	2.874
MRE	0.001	-0.001	-0.001
MRPE	0.020	-0.337	-0.444
Exec. Time	0.009	0.016	0.020

Παρατηρείται σχετικά μικρή αύξηση του σφάλματος με αύξηση του ποσοστού ελλιπών τιμών καθώς και αμελητέα αύξηση του χρόνου εκτέλεσης. Τα αποτελέσματα περιγράφονται γραφικά στα Σχήματα 3.19 και 3.20 με σύγκριση χρονοσειρών.



Σχήμα 3.19: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με μέση τιμή κ-κοντινότερων γειτόνων.



Σχήμα 3.20: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με μέση τιμή κ-κοντινότερων γειτόνων.

3.3.6. Μετρήσεις μεθόδου συμπλήρωσης με χρήση του αλγορίθμου k-means

Στους Πίνακες 3.10 και 3.11 παρουσιάζονται οι τιμές των δεικτών μετρήσων για την μέθοδο συμπλήρωσης με χρήση του αλγορίθμου k-means.

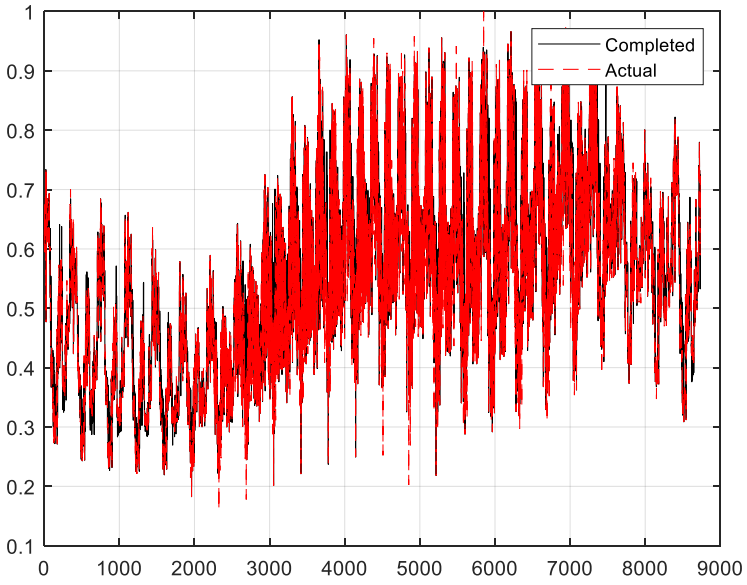
Πίνακας 3.10: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση k-meansαλγορίθμου, για το σύνολο δεδομένων ηλεκτρικού φορτίου.

Missing Values (%)	10%	20%	30%
MAE	0.006	0.013	0.021
MAPE	1.098	2.526	4.033
MRE	-0.001	-0.002	-0.005
MRPE	-0.276	-0.869	-1.545
Exec. Time	2.227	2.922	4.139

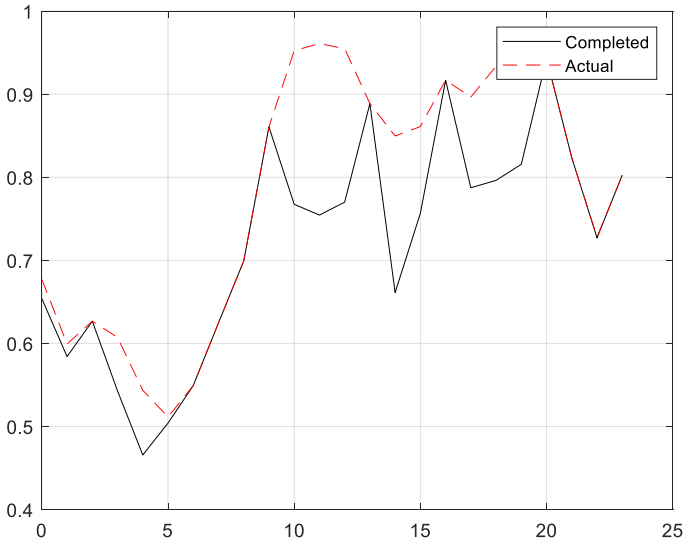
Πίνακας 3.11: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση k-meansαλγορίθμου, για το σύνολο δεδομένων ταχύτητας ανέμου.

Missing Values (%)	10%	20%	30%
MAE	0.108	0.261	0.356
MAPE	3.577	8.839	13.821
MRE	0.032	0.012	-0.005
MRPE	-0.899	-3.755	-7.154
Exec. Time	0.149	0.314	0.460

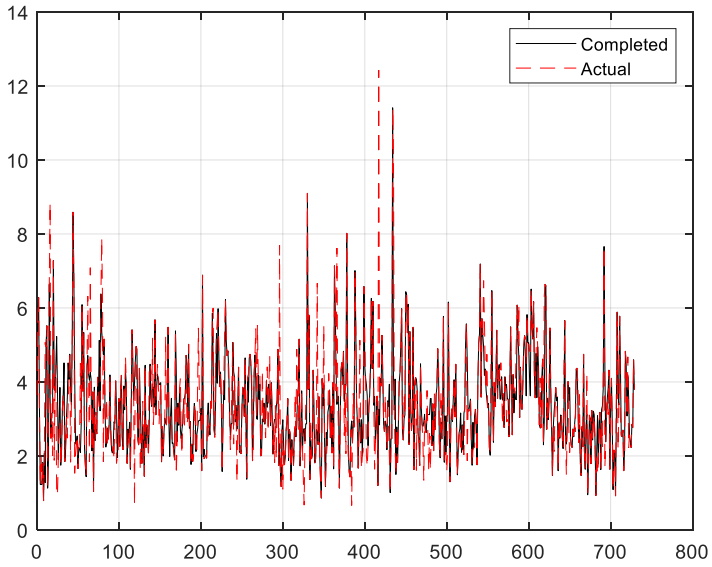
Παρατηρείται μικρή αύξηση του σφάλματος όσο αυξάνεται το ποσοστό ελλιπών τιμών στο πρώτο σύνολο δεδομένων, ενώ βλέπουμε σημαντική αύξηση στο δεύτερο σύνολο δεδομένων. Ο χρόνος εκτέλεσης παραμένει πολύ μικρός σε κάθε σενάριο εκτέλεσης.Τα αποτελέσματα παρουσιάζονται γραφικά στα Σχήματα 3.21 έως 3.23 με συγκρίσεις χρονοσειρών.



Σχήμα 3.21: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση k-meansαλγορίθμου.



Σχήμα 3.22: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση k-means αλγορίθμου.



Σχήμα 3.23: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ταχύτητας ανέμου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση k-means αλγορίθμου.

3.3.7. Μετρήσεις μεθόδου συμπλήρωσης με χρήση του αλγορίθμου k-medoids

Στους Πίνακες 3.12 και 3.13 παρουσιάζονται οι μετρήσεις δεικτών για την μέθοδο συμπλήρωσης με χρήση του αλγορίθμου k-medoids.

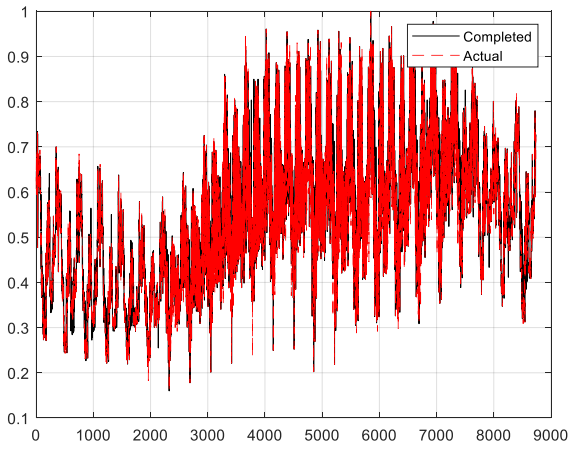
Πίνακας 3.12: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση k-medoids αλγορίθμου, για το σύνολο δεδομένων ηλεκτρικού φορτίου.

Missing Values (%)	10%	20%	30%
MAE	0.006	0.013	0.020
MAPE	1.113	2.400	3.823
MRE	-0.001	-0.002	-0.002
MRPE	-0.271	-0.672	-1.065
Exec. Time	1.453	3.282	13.351

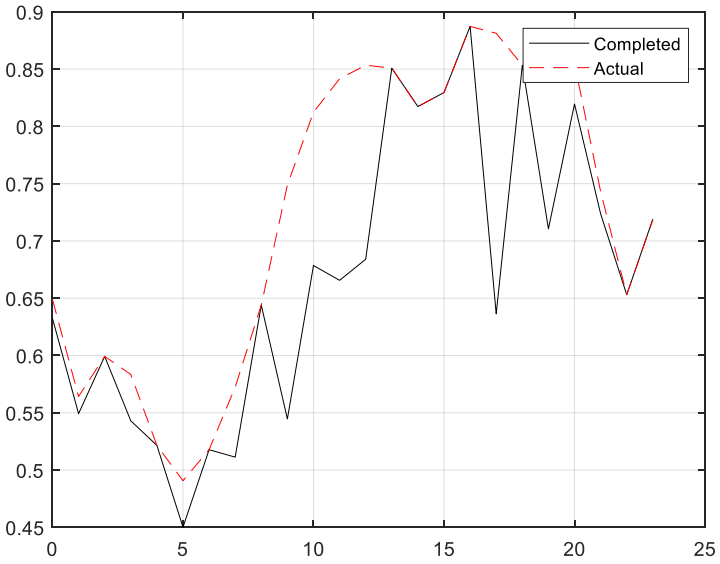
Πίνακας 3.13: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση k-medoids αλγορίθμου, για το σύνολο δεδομένων ταχύτητας ανέμου.

Missing Values (%)	10%	20%	30%
MAE	0.118	0.248	0.340
MAPE	4.420	9.464	13.670
MRE	-0.025	0.007	-0.025
MRPE	-2.486	-4.674	-7.485
Exec. Time	1.567	2.635	3.758

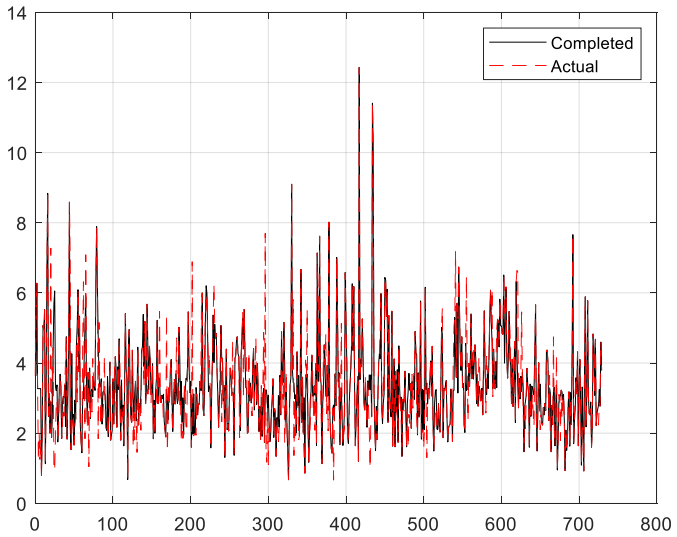
Παρατηρείται μικρή αύξηση του σφάλματος όσο αυξάνεται το ποσοστό ελλιπών τιμών στο πρώτο σύνολο δεδομένων και σημαντική αύξηση στο δεύτερο σύνολο δεδομένων. Ο χρόνος υπολογισμού αυξάνεται σημαντικά στο 30% ελλιπών τιμών. Τα αποτελέσματα παρουσιάζονται γραφικά στα Σχήματα 3.24 έως 3.26 με συγκρίσεις χρονοσειρών.



Σχήμα 3.24: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση k-medoids αλγορίθμου.



Σχήμα 3.25: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση k-medoids αλγορίθμου.



Σχήμα 3.26: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ταχύτητας ανέμου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση k-medoids αλγορίθμου.

3.3.8. Μετρήσεις μεθόδου συμπλήρωσης με χρήση ιεραρχικής ομαδοποίησης

Στους Πίνακες 3.14 και 3.15 παρουσιάζονται οι τιμές των δεικτών για την μέθοδο συμπλήρωσης με χρήση ιεραρχικής ομαδοποίησης.

Πίνακας 3.14: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση ιεραρχικής ομαδοποίησης, για το σύνολο δεδομένων ηλεκτρικού φορτίο.

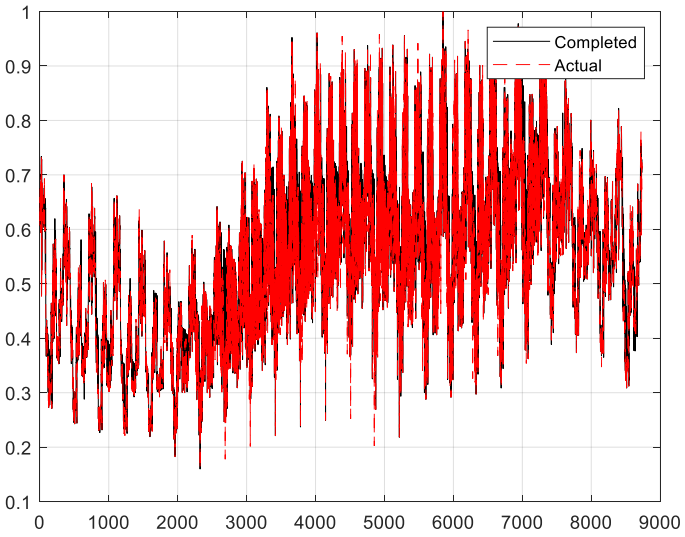
Missing Values (%)	10%	20%	30%
MAE	0.006	0.013	0.021
MAPE	1.159	2.525	3.990
MRE	-0.001	-0.001	-0.003
MRPE	-0.253	-0.723	-1.325
Exec. Time	0.996	1.512	1.921

Πίνακας 3.15: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση ιεραρχικής ομαδοποίησης, για το σύνολο δεδομένων ταχύτητας ανέμου.

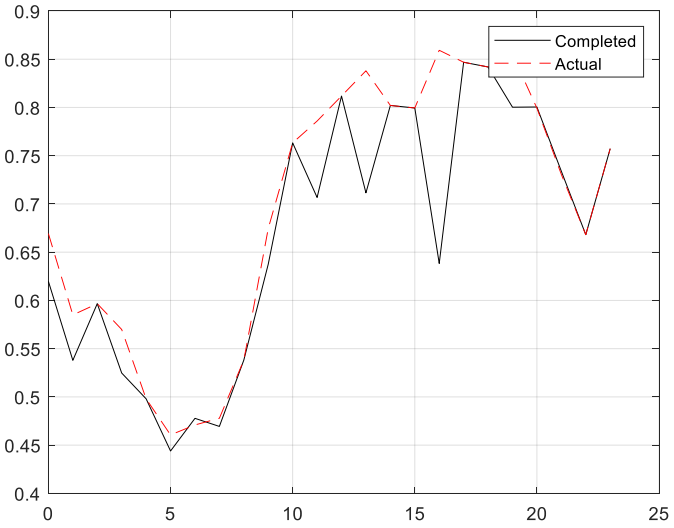
Missing Values (%)	10%	20%	30%
MAE	0.132	0.208	0.384
MAPE	4.767	7.859	13.552
MRE	0.025	-0.007	0.023
MRPE	-1.861	-3.870	-5.522
Exec. Time	0.110	0.270	0.352

Παρατηρείται μικρή αύξηση του σφάλματος με αύξηση του ποσοστού των ελλιπών τιμών στο πρώτο σύνολο δεδομένων, ενώ σημαντική αύξηση στο δεύτερο σύνολο δεδομένων. Ο χρόνος εκτέλεσης παραμένει πολύ μικρός σε κάθε σενάριο.

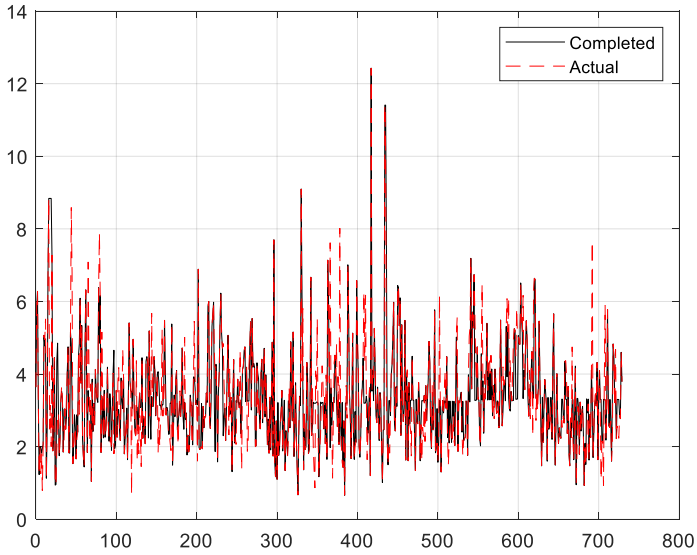
Τα αποτελέσματα παρουσιάζονται γραφικά στα Σχήματα 3.27 έως 3.29 με συγκρίσεις χρονοσειρών.



Σχήμα 3.27: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση ιεραρχικής ομαδοποίησης.



Σχήμα 3.28: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση ιεραρχικής ομαδοποίησης.



Σχήμα 3.29: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ταχύτητας ανέμου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση ιεραρχικής ομαδοποίησης.

3.3.9. Μετρήσεις μεθόδου συμπλήρωσης με χρήση ασαφούς ομαδοποίησης

Στους Πίνακες 3.16 και 3.17 παρουσίαζονται οι τιμές των δεικτών για την μέθοδο συμπλήρωσης με χρήση ασαφούς ομαδοποίησης.

Πίνακας 3.16: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση ασαφούς ομαδοποίησης, για το σύνολο δεδομένων ηλεκτρικού φορτίο.

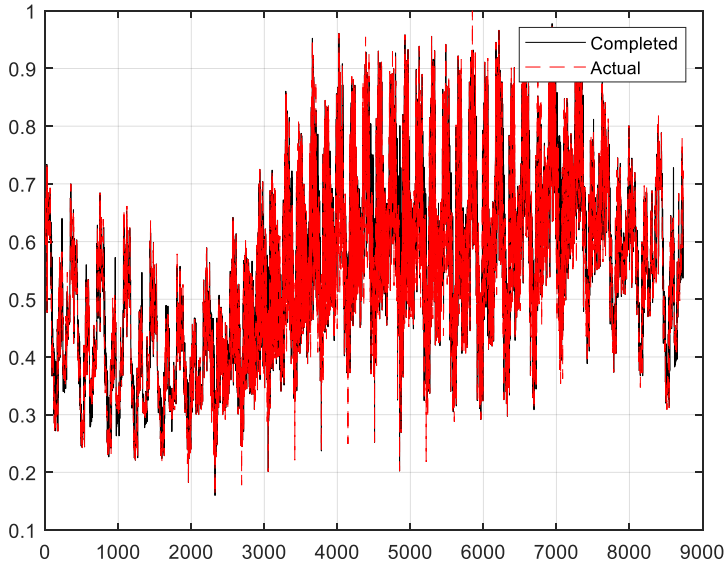
Missing Values (%)	10%	20%	30%
MAE	0.006	0.013	0.019
MAPE	1.166	2.482	3.835
MRE	-0.001	-0.002	-0.004
MRPE	-0.359	-0.685	-1.340
Exec. Time	2.055	3.708	5.104

Πίνακας 3.17: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση ασαφούς ομαδοποίησης, για το σύνολο δεδομένων ταχύτητας ανέμου.

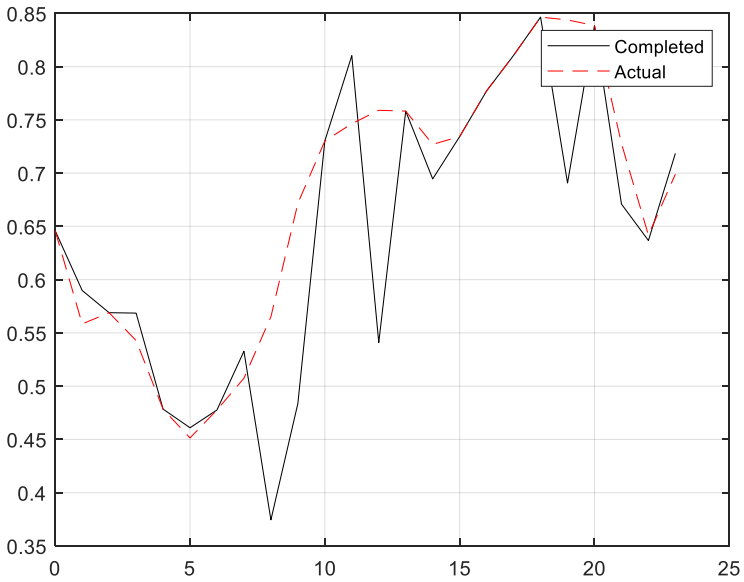
	10%	20%	30%
MAE	0.122	0.249	0.362
MAPE	4.861	9.639	11.913
MRE	0.004	-0.032	0.066
MRPE	-2.396	-5.391	-4.157
Exec. Time	0.633	0.992	1.360

Παρατηρείται μικρή αύξηση του λάθους, όσο αυξάνεται το ποσοστό ελλιπών τιμών στο πρώτο σύνολο δεδομένων, ενώ σημαντική αύξηση στο δεύτερο σύνολο δεδομένων. Ο χρόνος εκτέλεσης αυξάνεται εμφανώς αλλά όχι σημαντικά στο πρώτο σύνολο δεδομένων.

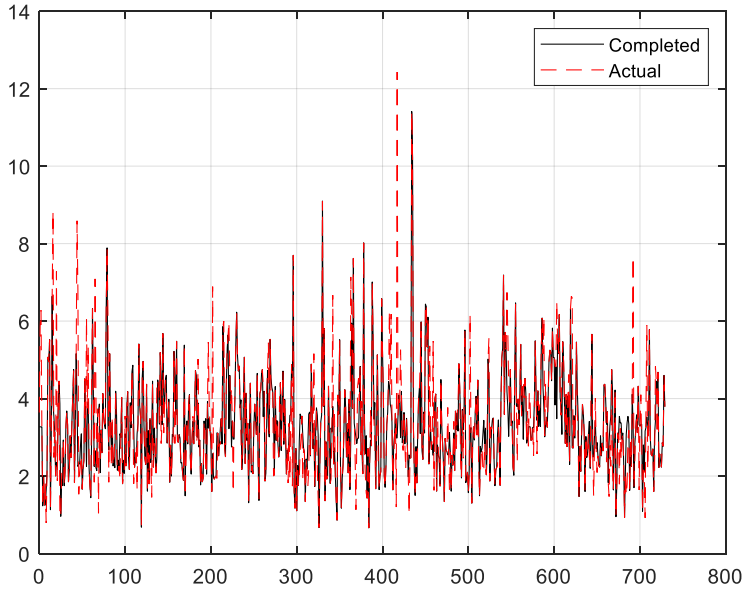
Τα αποτελέσματα παρουσιάζονται γραφικά στα Σχήματα 3.30 έως 3.32, με συγκρίσεις χρονοσειρών.



Σχήμα 3.30: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση ασαφούς ομαδοποίησης.



Σχήμα 3.31: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση ασαφούς ομαδοποίησης.



Σχήμα 3.32: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ταχύτητας ανέμου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση ασαφούς ομαδοποίησης.

3.3.10. Μετρήσεις μεθόδου συμπλήρωσης με χρήση αυτο-οργανώμενου χάρτη

Στους Πίνακες 3.18 και 3.19 παρουσιάζονται οι τιμές των δεικτών μέτρησης για την μέθοδο συμπλήρωσης με χρήση αυτο-οργανώμενου χάρτη.

Πίνακας 3.18: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση αυτο-οργανώμενου χάρτη, για το σύνολο δεδομένων ηλεκτρικού φορτίου.

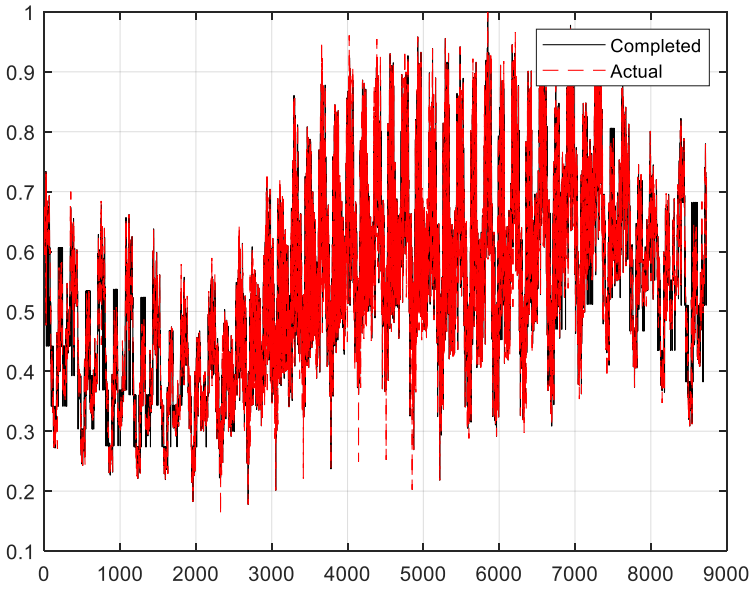
Missing Values (%)	10%	20%	30%
MAE	0.005	0.010	0.015
MAPE	0.973	1.920	2.894
MRE	-0.001	0.001	-0.001
MRPE	-0.168	-0.369	-0.579
Exec. Time	1.714	1.720	1.878

Πίνακας 3.19: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση αυτο-οργανώμενου χάρτη, για το σύνολο δεδομένων ταχύτητας ανέμου.

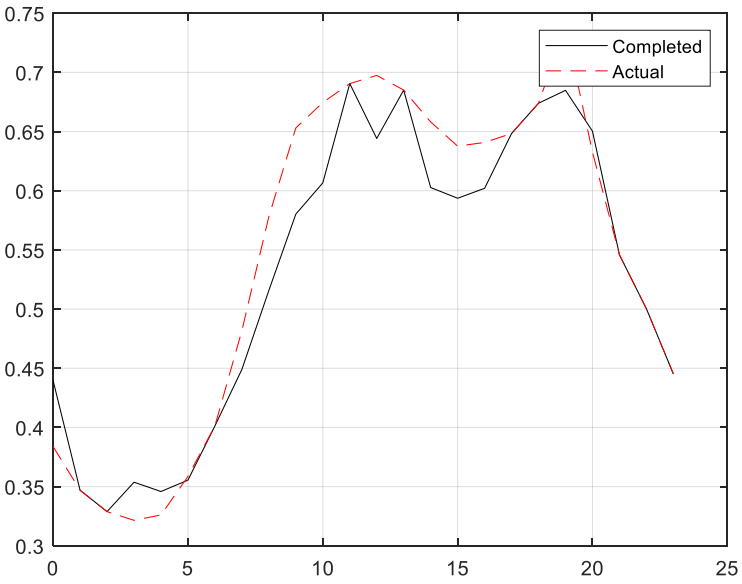
Missing Values (%)	10%	20%	30%
MAE	0.133	0.218	0.385
MAPE	5.159	7.858	13.251
MRE	-0.021	0.013	-0.036
MRPE	-2.680	-2.244	-5.638
Exec. Time	1.826	2.117	1.825

Ο χρόνος εκτέλεσης παραμένει πολύ μικρός σε κάθε σενάριο εκτέλεσης, ενώ παρατηρείται μικρή αύξηση του σφάλματος σε αύξηση ποσοστού ελλιπών τιμών στο πρώτο σύνολο δεδομένων και σημαντική αύξηση του στο δεύτερο.

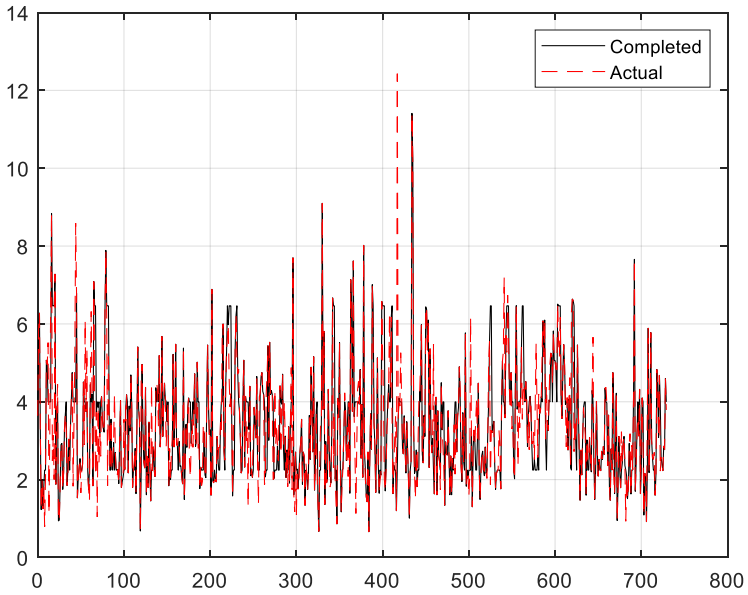
Τα αποτελέσματα παρουσιάζονται γραφικά στα Σχήματα 3.33 έως 3.35 με συγκρίσεις χρονοσειρών.



Σχήμα 3.33: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση αυτο-οργανώμενου χάρτη.



Σχήμα 3.34: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση αυτο-οργανώμενου χάρτη.



Σχήμα 3.35: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ταχύτητας ανέμου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με χρήση αυτο-οργανώμενου χάρτη.

3.3.11. Μετρήσεις μεθόδου συμπλήρωσης με χρήση κατανομής συχνοτήτων και μέγιστη πιθανότητα

Στους Πίνακες 3.20 και 3.21 παρουσιάζονται τα αποτελέσματα, με τις τιμές των δεικτών για την μέθοδο συμπλήρωσης με κατανομή συχνοτήτων και μέγιστη πιθανότητα.

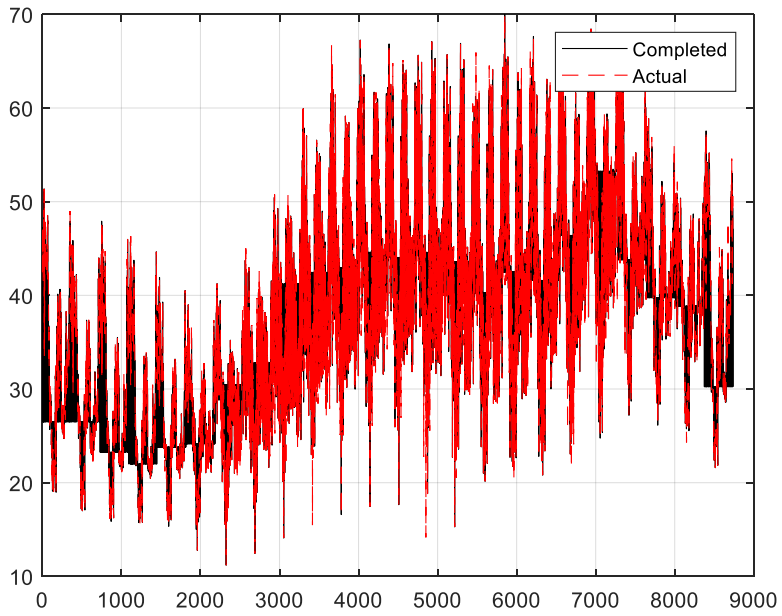
Πίνακας 3.20: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση κατανομής συχνοτήτων και μέγιστη πιθανότητα, για το σύνολο δεδομένων ηλεκτρικού φορτίου.

Missing Values (%)	10%	20%	30%
MAE	0.713	1.388	2.110
MAPE	1.828	3.706	5.485
MRE	0.226	0.404	0.677
MRPE	0.154	0.238	0.516
Exec. Time	8.060	15.752	24.477

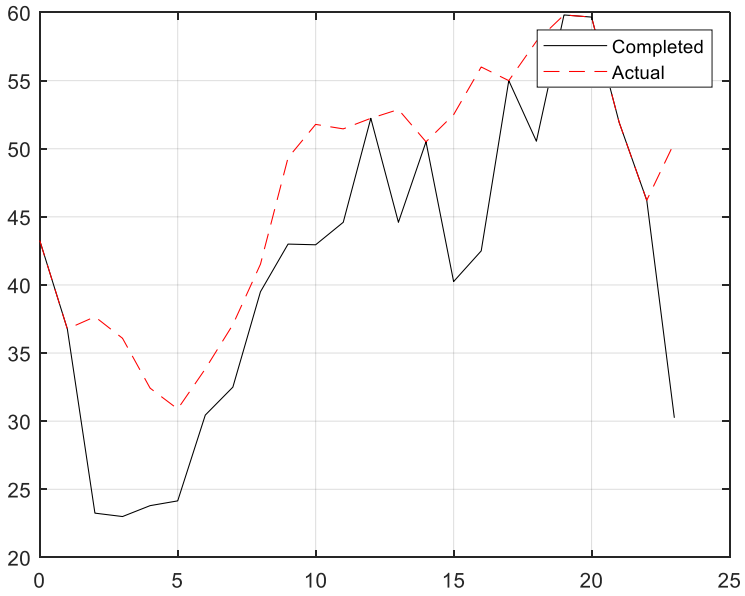
Πίνακας 3.21: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση κατανομής συχνοτήτων και μέγιστη πιθανότητα, για το σύνολο δεδομένων ταχύτητας ανέμου.

Missing Values (%)	10%	20%	30%
MAE	0.116	0.322	0.339
MAPE	3.392	8.220	11.094
MRE	0.044	0.304	0.064
MRPE	-0.551	6.696	-3.509
Exec. Time	0.384	0.893	1.209

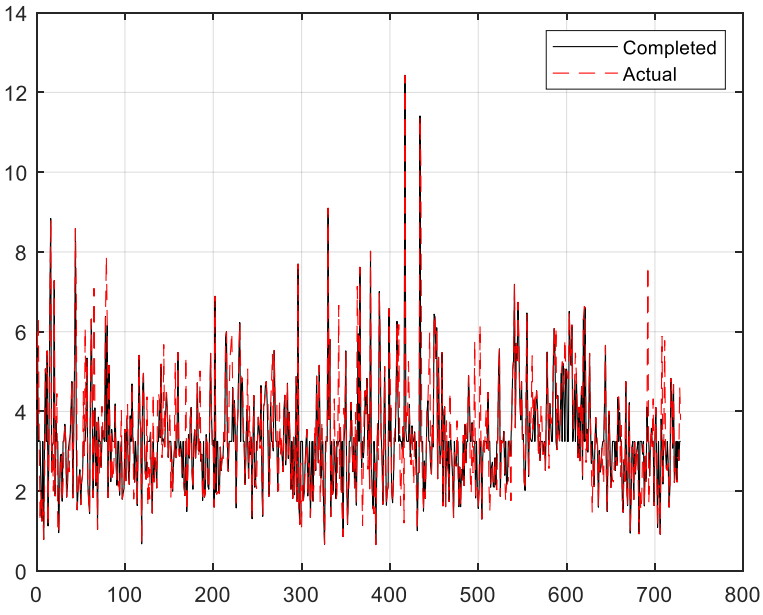
Παρατηρείται εμφανής αύξηση στο σφάλμα, με την αύξηση του ποσοστού ελλιπών τιμών, καθώς και μεγάλη αύξηση του χρόνου εκτέλεσης, ειδικά στο πρώτο σύνολο δεδομένων. Τα αποτελέσματα παρουσιάζονται γραφικά στα Σχήματα 3.36 έως 3.38 με συγκρίσεις χρονοσειρών.



Σχήμα 3.36: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με κατανομή συχνοτήτων και μέγιστη πιθανότητα.



Σχήμα 3.37: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με κατανομή συχνοτήτων και μέγιστη πιθανότητα.



Σχήμα 3.38: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ταχύτητας ανέμου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με κατανομή συχνοτήτων και μέγιστη πιθανότητα.

3.3.12. Μετρήσεις μεθόδου συμπλήρωσης με χρήση κατανομής συχνοτήτων και αθροιστική πιθανότητα

Στους Πίνακες 3.22 και 3.23 παρουσιάζονται τα αποτελέσματα με τις τιμές των δεικτών για την μέθοδο συμπλήρωσης με κατανομή συχνοτήτων και αθροιστική πιθανότητα.

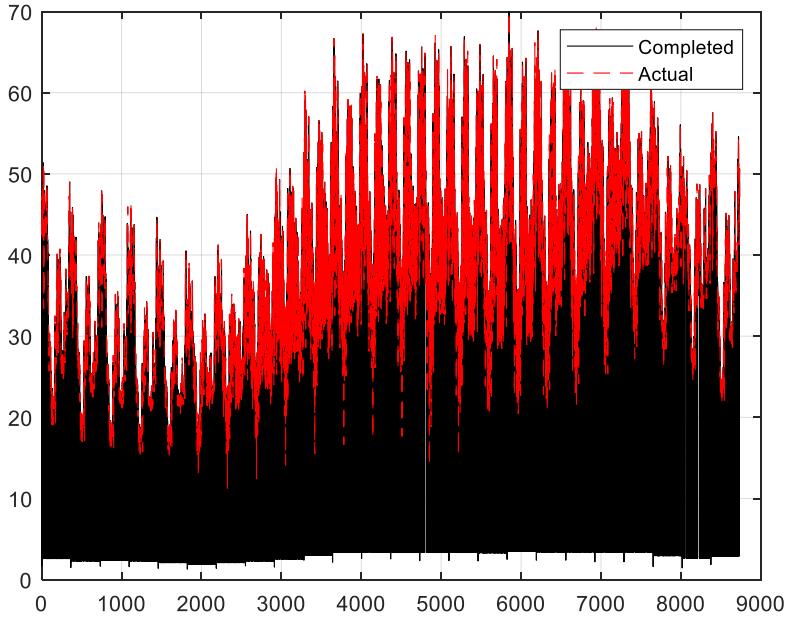
Πίνακας 3.22: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση κατανομής συχνοτήτων και αθροιστική πιθανότητα, για το σύνολο δεδομένων ηλεκτρικού φορτίου.

Missing Values (%)	10%	20%	30%
MAE	3.655	7.335	10.854
MAPE	9.223	18.462	27.675
MRE	3.655	7.335	10.854
MRPE	9.223	18.462	27.675
Exec. Time	1.997	4.401	6.564

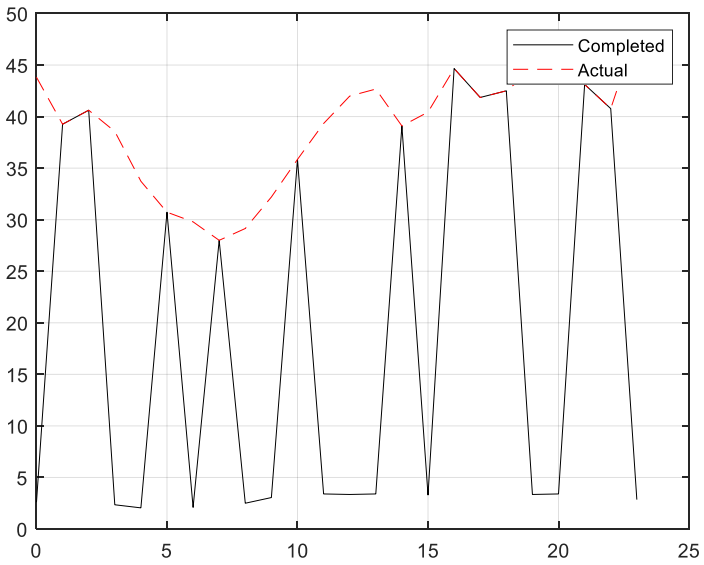
Πίνακας 3.23: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση κατανομής συχνοτήτων και αθροιστική πιθανότητα, για το σύνολο δεδομένων ταχύτητας ανέμου.

Missing Values (%)	10%	20%	30%
MAE	0.313	0.534	0.861
MAPE	7.830	15.477	23.379
MRE	0.313	0.534	0.861
MRPE	7.830	15.477	23.379
Exec. Time	0.212	0.480	0.832

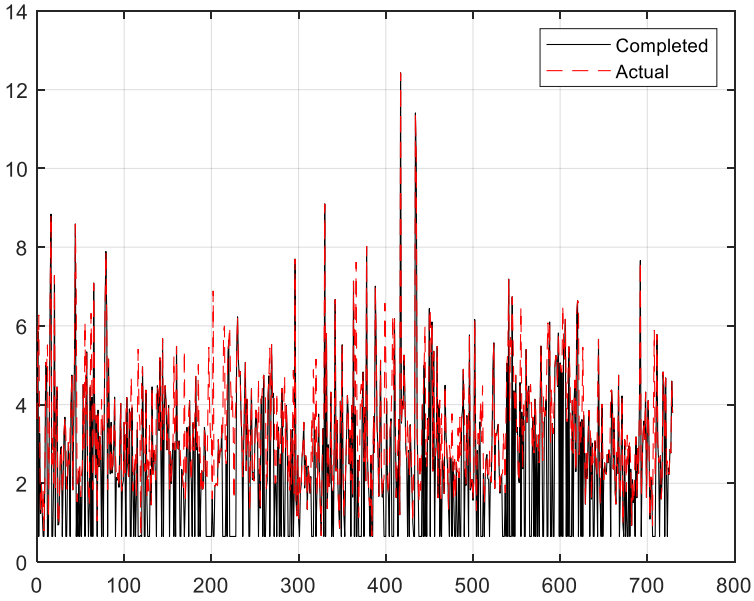
Παρατηρείται σημαντική αύξηση στο λάθος σε αύξηση του ποσοστού των ελλιπών τιμών και στα δυο σύνολα δεδομένων. Ο χρόνος εκτέλεσης αυξάνεται εμφανώς στο πρώτο σύνολο δεδομένων στην αύξηση του ποσοστού ελλιπών τιμών. Τα αποτελέσματα παρουσιάζονται γραφικά στα Σχήματα 3.39 έως 3.41 με συγκρίσεις χρονοσειρών.



Σχήμα 3.39: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με κατανομή συχνοτήτων και αθροιστική πιθανότητα.



Σχήμα 3.40: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με κατανομή συχνοτήτων και αθροιστική πιθανότητα.



Σχήμα 3.41: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ταχύτητας ανέμου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με κατανομή συχνοτήτων και αθροιστική πιθανότητα.

3.3.13. Μετρήσεις μεθόδου συμπλήρωσης με χρήση κατανομής συχνοτήτων παραθύρου τιμών και μέγιστη πιθανότητα

Στους Πίνακες 3.24 και 3.25 παρουσιάζονται τα αποτελέσματα με τις τιμές των δεικτών για τη μέθοδο συμπλήρωσης κατανομής συχνοτήτων παραθύρου τιμών και μέγιστη πιθανότητα.

Πίνακας 3.24: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση κατανομής συχνοτήτων παραθύρου τιμών και μέγιστη πιθανότητα, για το σύνολο δεδομένων ηλεκτρικού φορτίου.

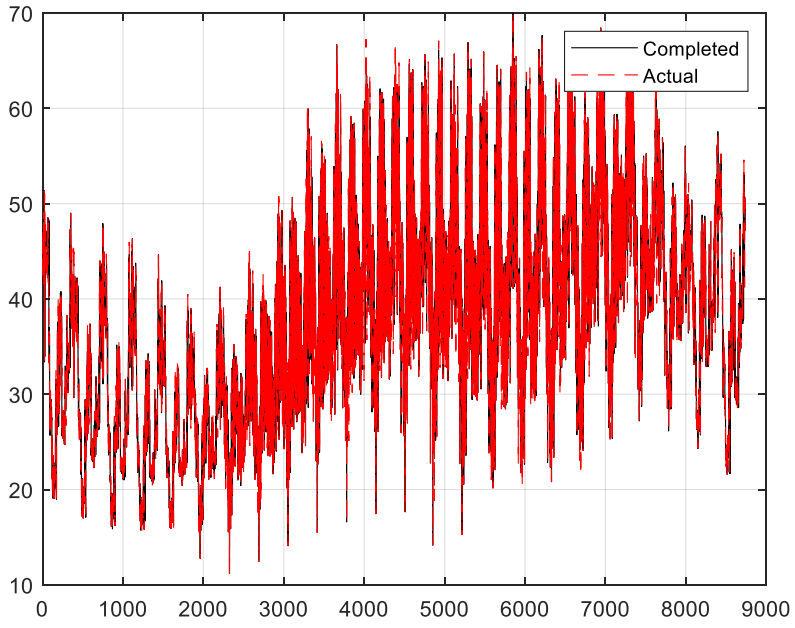
Missing Values (%)	10%	20%	30%
MAE	0.411	0.909	1.350
MAPE	1.093	2.382	3.531
MRE	0.017	0.019	0.059
MRPE	-0.122	-0.271	-0.304
Exec. Time	7.230	13.977	21.348

Πίνακας 3.25: Τιμές δεικτών μεθόδου συμπλήρωσης με χρήση κατανομής συχνοτήτων παραθύρου τιμών και μέγιστη πιθανότητα, για το σύνολο δεδομένων ταχύτητας ανέμου.

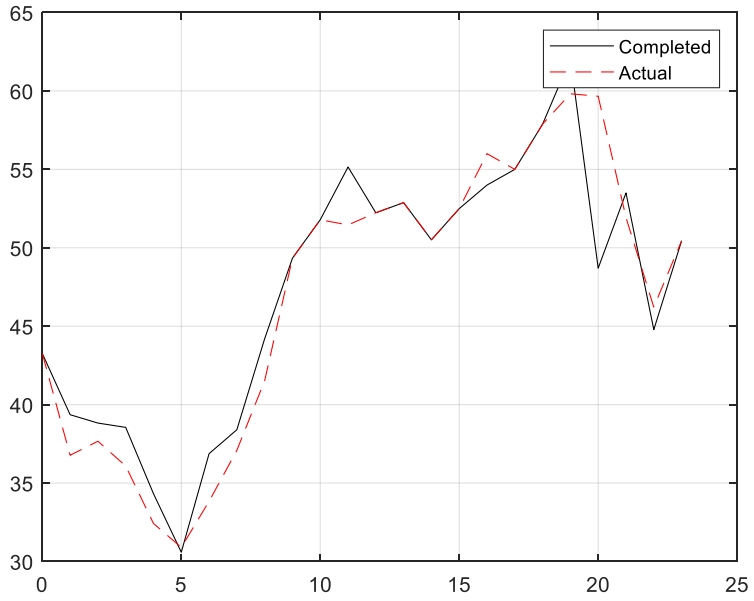
Missing Values (%)	10%	20%	30%
MAE	0.119	0.247	0.456
MAPE	3.483	8.134	12.140
MRE	0.027	0.077	0.275
MRPE	-0.217	-0.709	2.918
Exec. Time	0.164	0.237	0.423

Παρατηρείται μικρή αύξηση του σφάλματος με την αύξηση του ποσοστού ελλιπών τιμών στο πρώτο σύνολο δεδομένων, ωστόσο σημαντική αύξηση του χρόνου. Στο δεύτερο σύνολο δεδομένων παρατηρείται σημαντική αύξηση του σφάλματος, με τον χρόνο εκτέλεσης να παραμένει μικρός.

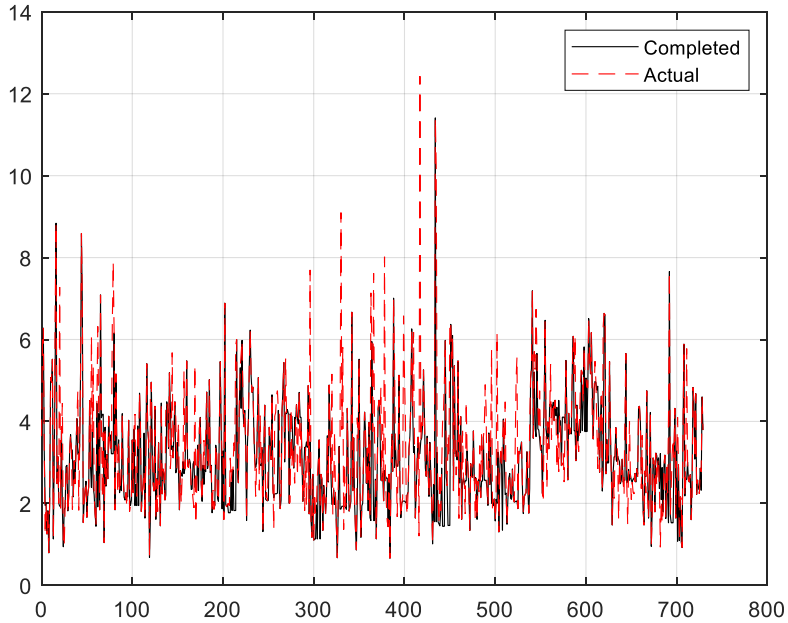
Τα αποτελέσματα παρουσιάζονται γραφικά στα Σχήματα 3.42 έως 3.44 με συγκρίσεις χρονοσειρών.



Σχήμα 3.42: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με κατανομή συχνοτήτων παραθύρου τιμών και μέγιστη πιθανότητα.



Σχήμα 3.43: Σύγκριση αρχικής και συμπληρωμένης ημερήσιας χρονοσειράς ηλεκτρικού φορτίου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με κατανομή συχνοτήτων παραθύρου τιμών και μέγιστη πιθανότητα.



Σχήμα 3.44: Σύγκριση αρχικής και συμπληρωμένης χρονοσειράς συνόλου δεδομένων ταχύτητας ανέμου για ποσοστό 30% ελλιπών τιμών με μέθοδο συμπλήρωσης με κατανομή συχνοτήτων παραθύρου τιμών και μέγιστη πιθανότητα.

3.4 Παρουσίασηαποτελεσμάτων συγκεντρωτικά

Πίνακας 3.26: Τιμές δεικτών μεθόδων συμπλήρωσης ιεραρχικοποιημένες κατά MAPE για 10% ελλιπείς τιμές στο σύνολο δεδομένων ηλεκτρικού φορτίου.

	MAPE	**MAE**	**MRE**	**MRPE**	**Exec. Time**
KNN	0.469	0.002	0.001	0.020	0.009
Mean Previous Next	0.824	0.004	0.001	-0.117	0.071
Linear Interpolation	0.924	0.005	-0.001	-0.167	1.012
SOM	0.973	0.005	-0.001	-0.168	1.714
Mean two previous	1.048	0.006	0.001	-0.215	0.011
Frequency distribution windowed	1.093	0.411	0.017	-0.122	7.230
k-means	1.098	0.006	-0.001	-0.276	2.227
k-medoids	1.113	0.006	-0.001	-0.253	0.996
Hierarchical clustering	1.159	0.006	-0.001	-0.253	0.996
Fuzzy clustering	1.166	0.006	-0.001	-0.359	2.055
Min. Euclidean distance	1.199	0.007	0.001	-0.021	35.906
Values back	1.443	0.008	-0.001	-0.208	0.015
Frequency distribution max prob.	1.828	0.713	0.226	0.154	8.060
Mean column	1.879	0.090	-0.001	-0.630	0.045
Frequency distribution sum prob.	9.223	3.665	3.655	9.223	1.997

Πίνακας 3.27: Τιμές δεικτών μεθόδων συμπλήρωσης ιεραρχικοποιημένες κατά MAPE για 20% ελλιπείς τιμές στο σύνολο δεδομένων ηλεκτρικού φορτίου.

	MAPE	**MAE**	**MRE**	**MRPE**	**Exec. Time**
Mean Previous Next	1.615	0.008	0.001	-0.223	0.137
KNN	1.778	0.009	-0.001	-0.337	0.016
Linear Interpolation	1.897	0.011	0.001	-0.220	1.089
SOM	1.92	0.010	0.001	-0.368	1.720
Mean two previous	2.074	0.012	0.001	-0.140	0.010
Frequency distribution max prob. windowed	2.382	0.909	0.019	-0.271	13.977
k-medoids	2.4	0.013	-0.002	-0.672	3.282
Fuzzy clustering	2.482	0.013	-0.002	-0.685	3.708
Hierarchical clustering	2.525	0.013	-0.001	-0.723	1.512
k-means	2.526	0.013	-0.002	-0.869	2.922
Min. Euclidean distance	2.616	0.014	-0.001	-0.388	74.037
Values back	3.006	0.016	-0.001	-0.479	0.012
Frequency distribution max prob.	3.706	1.388	0.404	0.238	15.752
Mean column	3.753	0.019	0.001	-0.909	0.025
Frequency distribution sum prob.	18.462	7.335	7.335	18.462	4.401

Πίνακας 3.28: Τιμές δεικτών μεθόδων συμπλήρωσης ιεραρχικοποιημένες κατά MAPE για 30% ελλιπείς τιμές στο σύνολο δεδομένων ηλεκτρικού φορτίου.

	MAPE	**MAE**	**MRE**	**MRPE**	**Exec. Time**
Mean Previous Next	2.69	0.015	0.001	-0.366	0.206
KNN	2.874	0.014	-0.001	-0.444	0.020
SOM	2.894	0.015	-0.001	-0.579	1.878
Linear Interpolation	3.019	0.017	0.001	-0.252	1.252
Mean two previous	3.303	0.018	0.001	-0.568	0.015
Frequency distribution max prob. windowed	3.531	1.350	0.059	-0.304	21.348
k-medoids	3.823	0.020	-0.002	-1.065	13.351
Fuzzy clustering	3.835	0.019	-0.004	-1.340	5.104
Hierarchical clustering	3.99	0.021	-0.003	-1.325	1.921
k-means	4.033	0.021	-0.005	-1.545	4.139
Min. Euclidean distance	4.043	0.022	-0.001	-0.581	109.821
Values back	4.568	0.024	0.001	-0.568	0.027
Frequency distribution max prob.	5.485	2.110	0.677	0.516	24.477
Mean column	5.606	0.029	0.001	-1.483	0.027
Frequency distribution sum prob.	27.675	10.854	10.854	27.675	6.564

Πίνακας 3.29: Τιμές δεικτών μεθόδων συμπλήρωσης ιεραρχικοποιημένες κατά MAPE για 10% ελλιπείς τιμές στο σύνολο δεδομένων ταχύτητας ανέμου.

	MAPE	**MAE**	**MRE**	**MRPE**	**Exec. Time**
Frequency distribution max prob.	3.392	0.116	0.044	-0.551	0.384
Mean Previous Next	3.408	0.107	0.033	-0.131	0.011
Frequency distribution max prob. windowed	3.483	0.119	0.027	-0.217	0.164
k-means	3.577	0.108	0.032	-0.889	0.149
k-medoids	4.420	0.118	-0.025	-2.486	1.567
Mean two previous	4.742	0.143	0.007	-1.974	0.003
Hierarchical clustering	4.767	0.132	0.025	-1.861	0.110
Fuzzy clustering	4.861	0.122	0.004	-2.396	0.633
SOM	5.159	0.133	-0.021	-2.680	1.826
Min. Euclidean distance	5.166	0.159	-0.011	-1.659	4.569
Linear Interpolation	5.183	0.116	-0.014	-3.127	0.066
Values back	5.391	0.155	-0.005	-1.992	0.004
Mean column	5.501	0.131	-0.002	-3.233	0.007
Frequency distribution sum prob.	7.830	0.313	0.313	7.830	0.212

Πίνακας 3.30: Τιμές δεικτών μεθόδων συμπλήρωσης ιεραρχικοποιημένες κατά MAPE για 20% ελλιπείς τιμές στο σύνολο δεδομένων ταχύτητας ανέμου.

	MAPE	**MAE**	**MRE**	**MRPE**	**Exec. Time**
Mean two previous	7.567	0.210	0.033	-1.948	0.003
Mean Previous Next	7.703	0.206	0.026	-2.582	0.013
SOM	7.858	0.218	0.013	-2.244	2.117
Hierarchical clustering	7.859	0.208	-0.007	-3.870	5.522
Frequency distribution max prob. windowed	8.134	0.247	0.077	-0.709	0.237
Frequency distribution max prob.	8.220	0.322	0.304	6.696	0.893
Mean column	8.230	0.215	0.001	-4.285	8.339
k-means	8.839	0.261	0.012	-3.755	0.314
Min. Euclidean distance	9.273	0.314	0.044	-1.734	3.345
Linear Interpolation	9.387	0.264	0.004	-3.617	0.099
k-medoids	9.464	0.248	0.007	-4.674	2.635
Fuzzy clustering	9.639	0.249	-0.032	-5.391	0.992
Values back	10.635	0.309	0.011	-3.629	0.004
Frequency distribution sum prob.	15.477	0.534	0.534	15.477	0.480

Πίνακας 3.31: Τιμές δεικτών μεθόδων συμπλήρωσης ιεραρχικοποιημένες κατά MAPE για 30% ελλιπείς τιμές στο σύνολο δεδομένων ταχύτητας ανέμου.

	MAPE	**MAE**	**MRE**	**MRPE**	**Exec. Time**
Frequency distribution max prob.	11.094	0.339	0.064	-3.509	1.209
Fuzzy clustering	11.913	0.362	0.066	-4.157	1.360
Frequency distribution max prob. windowed	12.140	0.456	0.275	2.918	0.423
Mean Previous Next	13.038	0.350	-0.027	-6.246	0.017
SOM	13.251	0.385	-0.036	-5.638	1.825
Linear Interpolation	13.364	0.363	-0.012	-5.993	0.104
Hierarchical clustering	13.552	0.384	0.023	-5.522	0.352
k-medoids	13.670	0.340	-0.025	-7.485	3.758
Mean column	13.819	0.332	-0.048	-8.339	0.003
k-means	13.821	0.356	-0.005	-7.154	0.460
Min. Euclidean distance	14.051	0.405	0.056	-3.345	14.532
Mean two previous	14.841	0.389	-0.033	-7.745	0.004
Values back	15.171	0.453	0.024	-4.339	0.004
Frequency distribution sum prob.	23.379	0.861	0.861	23.379	0.832

ΚΕΦΑΛΑΙΟ 4

ΣΥΜΠΕΡΑΣΜΑΤΑ

4.1 Σύνοψη και συμπεράσματα

Η συλλογή δεδομένων έχει ως απώτερο σκοπό την εξαγωγή συμπερασμάτων, συνεπώς οι ελλιπείς τιμές είναι πάντοτε ανεπιθύμητες. Η έλλειψη τιμών μπορεί να μειώσει τις πληροφορίες που μπορούμε να εξάγουμε, αλλά και να οδηγήσει σε λάθος συμπεράσματα που αφορούν το σύνολο δεδομένων που μελετάται. Η έλλειψη δεδομένων μπορεί να οφείλεται σε πολλούς παράγοντες, όπως λανθασμένη λειτουργία ενός σένσορα, σφάλμα του οργάνου μετρήσεων ή κακή αποθήκευση αρχείου δεδομένων. Η κακή διαχείριση των ελλιπών τιμών μπορεί να οδηγήσει σε λανθασμένη ερμηνεία και μοντελοποίηση των δεδομένων. Συνεπώς η προσοχή στρέφεται στην συμπλήρωση των ελλιπών τιμών με παραγόμενες τιμές, με βάση το σύνολο δεδομένων.

Σκοπός της εργασίας αυτής ήταν να εξετάσει και να συγκρίνει μεταξύ τους μεθόδους συμπλήρωσης ελλιπών δεδομένων με παραγόμενες τιμές, δείχνοντας αν και κατα πόσο πιο πολύπλοκοι υπολογιστικά μέθοδοι όπως clusteringκαι πιθανοτήτων αποδίδουν καλύτερα από τις κλασσικές μεθόδους συμπλήρωσης που χρησιμοποιούνται ευρέως. Οι μέθοδοι αυτοί εφαρμόστηκαν σε δεδομένα μετρήσεων ηλεκτρικού φορτίου και δεδομένα ταχύτητας ανέμου.

Η εξέταση των μεθόδων απαιτεί κατάλληλη επιλογή δεικτών ώστε να γίνει η εξαγωγή των συμπερασμάτων. Οι υλοποιήσεις των μεθόδων έγιναν στο προγραμματιστικό περιβάλλον και η σύγκρισή τους έγινε με χρήση δεικτών: μεσο απόλυτο σφάλμα, μέσο απόλυτο ποσοστιαίο σφάλμα, μέσο σχετικό σφάλμα και χρόνος εκτέλεσης.

Οι μέθοδοι που εξετάστηκαν ταξινομούνται σε τρεις κατηγορίες:

- Μέθοδοι κλασσικής στατιστικής: συμπλήρωση με μέση τιμή στήλης, συμπλήρωση με μέση τιμή επόμενης και προηγούμενης τιμής, συμπλήρωση με μέση τιμή των δυο προηγουμένων τιμών, συμπλήρωση με χρήση ελάχιστης Ευκλείδειας απόστασης, συμπλήρωση με επιλεγόμενη προηγούμενη τιμή, συμπλήρωση με γραμμική παρεμβολή (LI), συμπλήρωση με ζυγισμένο μέσο όρο κ-κοντινότερων γειτόνων (KNN)

- Μέθοδοι ομαδοποίησης δεδομένων (clustering): συμπλήρωση με χρήση του k-means αλγορίθμου, συμπλήρωση με χρήση του k-medoids αλγορίθμου, συμπλήρωση με ιεραρχική ομαδοποίηση (hierarchical clustering), συμπλήρωση με ασαφή ομαδοποίηση (fuzzy clustering), συμπλήρωση με ομαδοποίηση με αυτοοργανώμενο χάρτη (self-organizing map)
- Μέθοδοι με χρήση πιθανοτήτων: συμπλήρωση με κατανομή συχνοτήτων στήλης και μέγιστη πιθανότητα, συμπλήρωση με κατανομή συχνοτήτων στήλης και αθροιστική πιθανότητα, συμπλήρωση με κατανομή συχνοτήτων παραθύρου τιμών και μέγιστη πιθανότητα

Τα κυριότερα συμπεράσματα από την εξέταση των αλγορίθμων μπορούν να συνοψιστούν στα εξής:

- Δεν υπάρχει καθολικά ιδανικός αλγόριθμος για κάθε σύνολο δεδομένων.Οι μέθοδοι με την καλύτερη απόδοση διαφέρουν μεταξύ των συνόλων δεδομένων. Χαρακτηριστικό παράδειγμα αποτελεί η μέθοδος κατανομής συχνοτήτων και συμπλήρωση με μέγιστη πιθανότητα. Για το σύνολο δεδομένων ταχύτητας ανέμου, αποτελεί τη μέθοδο με την καλύτερη απόδοση (πίνακες 3.29, 3.31), ανεξάρτητα του ποσοστού ελλιπών τιμών. Αντίθετα στο σύνολο δεδομένων ηλεκτρικού φορτίου είναι απο τις μεθόδους με τη λιγότερη καλή απόδοση (πίνακες 3.26, 3.27, 3.28). Αντίθετα, η μέθοδος συμπλήρωσης με γραμμική παρεμβολή παρουσιάζει πολύ καλή απόδοση για τα δεδομένα ηλεκτρικού φορτίου, ενώ κακή για τα δεδομένα ταχύτητας ανέμου. Συνεπώς, κάθε σύνολο δεδομένων έχει τις ιδιαιτερότητες του και χρειάζεται ξεχωριστή εξέταση μεθόδων, καθώς δεν υπάρχει συνολικά ιδανική μέθοδος συμπλήρωσης.
- Για το σύνολο δεδομένων ηλεκτρικού φορτίου καλύτερες αποδόσεις σύμφωνα με τους συγκεντρωτικούς πίνακες (πίνακες 3.26, 3.27, 3.28) παρουσιάζουν οι μέθοδοι συμπλήρωσης κ-κοντινότερων γειτόνων, μέσης τιμής προηγούμενης και επόμενης τιμής, γραμμικής παρεμβολής και αυτοοργανώμενου χάρτη. Επιπλέον, καλή απόδοση παρουσιάζει η μέθοδος κατανομής συχνοτήτων παραθύρου τιμών, όμως με αυξημένο κόστος εκτέλεσης που φτάνει τα 20 δευτερόλεπτα (πίνακας 3.28).

- Για το σύνολο δεδομένων ηλεκτρικού φορτίου χειρότερες αποδόσεις σύμφωνα με τους συγκεντρωτικούς πίνακες (πίνακες 3.26, 3.27, 3.28) παρουσιάζουν οι μέθοδοι συμπλήρωσης με κατανομή συχνοτήτων στήλης και μέγιστη πιθανότητα, καθώς και η συμπλήρωση με μέση τιμή στήλης.
- Για το σύνολο δεδομένων ταχύτητας ανέμου καλύτερες αποδόσεις σύμφωνα με τους συγκεντρωτικούς πίνακες (πίνακες 3.29, 3.30, 3.31) παρουσιάζουν οι μέθοδοι κατανομής συχνοτήτων στήλης και παραθύρου τιμών με μέγιστη πιθανότητα, καθώς και η συμπλήρωση με μέση τιμή προηγούμενης και επόμενης τιμής.
- Για το σύνολο δεδομένων ταχύτητας ανέμου καλύτερες αποδόσεις σύμφωνα με τους συγκεντρωτικούς πίνακες (πίνακες 3.29, 3.30, 3.31) παρουσιάζουν οι μέθοδοι κατανομής συχνοτήτων και συμπλήρωση με αθροιστική πιθανότητα, καθώς και η συμπλήρωση με επιλεγόμενης απόστασης προηγούμενη τιμή, καθώς δεν επαναλαμβάνονται παρόμοιες τιμές σε διάστημα μιας εβδομάδας.
- Η μέθοδος συμπλήρωσης κατανομής συχνοτήτων στήλης και συμπλήρωση με αθροιστική πιθανότητα φαίνεται να έχει την χειρότερη απόδοση με το μεγαλύτερο απόλυτο ποσοστιαίο σφάλμα σε κάθε σενάριο εκτέλεσης και για τα δυο σύνολα δεδομένων. Συνεπώς αποτελεί μια μέθοδο που εξετάστηκε και δεν προτείνεται.
- Το μέγεθος του συνόλου δεδομένων παίζει σημαντικό ρόλο στην απόδοση των μεθόδων. Το ένα σύνολο δεδομένων που μελετήθηκε έχει μέγεθος 365x24 και το δεύτερο 731x1. Οι μέθοδοι εξετάστηκαν σε ποσοστό ελλιπών τιμών 10,20 και 30% των συνολικών τιμών. Μια μείωση απόδοσης και αύξηση του σφάλματος παρατηρείται σε όλες τις μεθόδους όσο αυξάνεται το ποσοστό των ελλιπών τιμών. Μεγαλύτερη αύξηση παρατηρείται στη μέθοδο συμπλήρωσης με κατανομή συχνοτήτων και αθροιστική πιθανότητα, όπου το σφάλμα αυξάνεται κατά ποσοστό που αγγίζει το 50% όσο αυξάνεται το ποσοστό ελλιπών τιμών και στα δυο σύνολα δεδομένων (πίνακες 3.22, 3.23).
- Οι μέθοδοι ομαδοποίησης όπως ομαδοποίηση με αυτοοργανώμενο χάρτη, k-means, k-medoids και ιεραρχική ομαδοποίηση δεν παρουσιάζουν σημαντική

αύξηση σφάλματος με την αύξηση του ποσοστού ελλιπών τιμών.Καταφέρνουν να επιτυγχάνουν καλή μέση απόδοση σε κάθε σενάριο εκτέλεσης για το σύνολο δεδομένων ηλεκτρικού φορτίου.

- Οι μέθοδοι κατανομής συχνοτήτων παρουσιάζουν καλή απόδοση σε γενική εικόνα και στα δυο σύνολα δεδομένων. Συγκεκριμένα η μέθοδος κατανομής συχνοτήτων παραθύρου τιμών έχει από τις καλύτερες αποδόσεις σε κάθε σενάριο εκτέλεσης για το σύνολο δεδομενων ηλεκτρικού φορτίου. Όσον αφορά το σύνολο δεδομένων ταχύτητας ανέμου, η μέθοδος κατανομής συχνοτήτων στήλης και συμπλήρωση με μέγιστη πιθανότητα έχει από τις καλύτερες αποδόσεις για κάθε σενάριο εκτέλεσης.
- Ο χρόνος εκτέλεσης φαίνεται να αυξάνεται σημαντικά σε όλες τις μεθόδους όσο αυξάνεται το ποσοστό ελλιπών τιμών. Σημαντικότερη αύξηση παρατηρείται στην μέθοδο με χρήση ελάχιστης Ευκλείδειας απόστασης για το σύνολο δεδομένων ηλεκτρικού φορτίου, όπου λόγω του υψηλού υπολογιστικού κόστους, ο χρόνος εκτέλεσης αγγίζει τα 100 δευτερόλεπτα (πίνακας 3.28). Επίσης, οι μέθοδοι κατανομής συχνοτήτων για το σύνολο δεδομένων ηλεκτρικού φορτίου αγγίζει τα 30 δευτερόλεπτα για ποσοστό ελλιπών τιμών 30% (πίνακας 3.28).
- Οι μέθοδοι συμπλήρωσης με μέση τιμή προηγούμενης και επόμενης τιμής, κ-κοντινότερων γειτόνων και αυτοοργανώμενου χάρτη επιτυγχάνουν συγκριτικά μικρό χρόνο εκτέλεσης και μικρό ποσοστιαίο σφάλμα.

Εν κατακλείδι, αποδείχτηκε πως ιδανικότερες μέθοδοι για το σύνολο δεδομένων ηλεκτρικού φορτίου, με μικρό μέσο σφάλμα καθώς και μικρό χρόνο εκτέλεσης αποτελούν οι μέθοδοι ομαδοποίησης, καθώς και κάποιες κλασσικές μέθοδοι στατιστικής, όπως γραμμική παρεμβολή και μέσος όρος προηγούμενης και επόμενης τιμής. Από την άλλη, για το σύνολο δεδομένων ταχύτητας ανέμου οι μέθοδοι κατανομής συχνοτήτων αποδείχτηκαν να έχουν την καλύτερη μέση απόδοση με μικρό ποσοστό σφάλματος και γρήγορους χρόνους εκτέλεσης.

Ωστόσο, όπως προαναφέρθηκε, τα αποτελέσματα της παρούσας εργασίας δεν μπορούν να γενικευτούν καθώς δεν υπάρχει καθολικά ιδανικός αλγόριθμος συμπλήρωσης ελλιπών δεδομένων, που θα απέδιδε καλά ως προς την ακρίβεια και τον χρόνο

εκτέλεσης για κάθε σύνολο δεδομένων. Κάθε σύνολο δεδομένων έχει τα δικά του ιδιαίτερα χαρακτηριστικά και προτείνεται να εξετάζεται ξεχωριστά.

4.2 Προτάσεις για μελλοντική έρευνα

Κάποιες ιδέες για μελλοντική έρευνα είναι οι εξής:

- Εξέταση αλγορίθμων μηχανικής μάθησης για την συμπλήρωση ελλιπών δεδομένων για δεδομένα ενέργειας. Για παράδειγμα, οι μέθοδοι naïve Bayes, SVM (Support Vector Machines) και νευρωνικά δίκτυα, αποτελούν δημοφιλείς μεθόδους συμπλήρωσης. Με την προσθήκη των μεθόδων αυτών, θα μπορούσε να γίνει περαιτέρω σύγκριση με τις κλασσικές μεθόδους συμπλήρωσης, τις μεθόδους ομαδοποίησης και πιθανοτήτων. Οι μέθοδοι που εξετάστηκαν στην παρούσα εργασία υλοποιήθηκαν στο προγραμματιστικό περιβάλλον του Matlab, ενώ οι μέθοδοι μηχανικής μάθησης είναι ήδη υλοποιημένες σε πακέτα της γλώσσας προγραμματισμού R.
- Μελλοντική εξέταση των μεθόδων θα μπορούσε να επικεντρωθεί στις σχέσεις μεταξύ των μεθόδων, καθώς και στα χαρακτηριστικά της καθεμιάς. Με τον τρόπο αυτό, θα μπορέσουν να εξαχθούν χρήσιμα συμπεράσματα ως προς το γιατί η κάθε μέθοδος αποδίδει καλύτερα συγκριτικά με τις υπόλοιπες, ανάλογα με το σύνολο δεδομένων που μελετάται.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Tshilidzi Marwala, *Computational intelligence for Missing Data Imputation, Estimation and Management, Knowledge Optimization Techniques*, 2009, pp. 4-6.

[2] J. A. Little, Donald B. Rubin, *Statistical Analysis with Missing Data*, 1987.

[3] Alvira Swalin, *How to Handle Missing Data*https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4Ημερομηνία πρόσβασης: 16-08-2021

[4] Edgar Acuna, Caroline Rodriguez, *The treatment of missing values and its effect in the classifier accuracy*, Research Gate, 2015.

[5] Minakshi, Dr. Rajan Vohra, Gimpy, *Missing Value Imputation in Multi Attribute Data Set*, IJCSIT, Vol. 5, 2014.

[6] Lúcia Maria Pina Moreira, *Fuzzy Clustering of Short Time Series with Missing Data for the Survival Prediction of Oncological Patients*, 2015.

[7] Yelipe UshaRani, Dr.P.Sammulal, *An Innovative Imputation and Classification Approach for Accurate Disease Prediction*, International Journal of Computer Science and Information Security (IJCSIS), Vol. 14 S1, 2016.

[8] Y.Usha Rani, P. Sammulal, *A Novel Approach for Imputation of Missing Attribute Values for Efficient Mining of Medical Datasets – Class Based Cluster Approach*, Rev. Téc. Ing. Univ. Zulia. Vol. 39, No 2, 2016.

[9] Teresa Pamuła, *Impact of Data Loss for Prediction of Traffic Flow on an Urban Road Using Neural Networks*, IEEE Transactions On Intelligent Transportation Systems, 2018.

[10] Steinley, D. K-*means clustering: A half-century synthesis*. Br. J. Math. Stat. Psychol. 2006, 59, 1–34.

[11] Buuren Stef van, *Flexible imputation of missing data*, Chapman &Hall, 2018, pp. 16-17.

[12] *Linear Interpolation*, Wikipedia https://en.wikipedia.org/wiki/Linear_interpolation Ημερομηνία πρόσβασης: 10-08-2021

[13] *Nearest Neighbor Classifiers,* http://trevorwhitney.com/data_mining/classification Ημερομηνία πρόσβασης: 10-08-2021

[14] Keyvan Golalipour, Ebrahim Akbari, Seyed Saeed Hamidi, Malrey Lee, Rasul Enayatifar, *From clustering to clustering ensemble selection: A review, Engineering Applications of Artificial Intelligence*, Volume 104, 2021, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2021.104388

[15] Zhou Liang, Pei Chen*, An automatic clustering algorithm based on the density-peak framework and Chameleon method,* Pattern Recognition Letters,Volume 150,2021,Pages 40-48,ISSN 0167-8655,https://doi.org/10.1016/j.patrec.2021.06.017

[16] Hartigan, J. A., & Wong, M. A., A k-means clustering algorithm, Applied Statistics, 1979, pp. 100-108.

[17] Di, Haibin & Shafiq, Muhammad &Alregib, Ghassan.*Multi-attribute k-means clustering for salt-boundary delineation from three-dimensional seismic data.* Geophysical Journal International, 2018

[18] MacKay, David."*Chapter 20. An Example Inference Task: Clustering*" (PDF). Information Theory, Inference and Learning Algorithms. Cambridge University Press, 2003, pp. 284–292

[19] Aruna Bhat, *K-medoids clustering using partitioning around medoids for performing face recognition*, Department of Electrical Engineering, IIT Delhi, Hauz Khas, New Delhi, 2014, pp. 3-5.

[20] Jin X., Han J., *K-Medoids Clustering*. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA., 2011 https://doi.org/10.1007/978-0-387-30164-8_426

[21] Kaufman, L. and Rousseeuw, P.J., *Clustering by means of Medoids, in Statistical Data Analysis Based on the L1 - Norm and Related Methods*,1987, edited by Y. Dodge, North-Holland,pp. 405– 416.

[22] *ML | K-Medoids clustering with solved example*, https://www.geeksforgeeks.org/ml-k-medoids-clustering-with-example/Ημερομηνία πρόσβασης: 10-08-2021

[23] Saurav Kaushik, *An Introduction to Clustering and different methods of clustering*, 2016.

[24] Tavish Srivastava*, Getting Clustering right,* Analytics Vidhya, 2013*, https://www.analyticsvidhya.com/blog/2013/11/getting-clustering-right/* Ημερομηνία πρόσβασης: 10-08-2021

[25] *The CLUSTER Procedure: Clustering Methods*. SAS/STAT 9.2 Users Guide. SAS Institute Ημερομηνία πρόσβασης: 10-08-2021

[26] *Linkage*, Mathworks, https*://www.mathworks.com/help/stats/linkage.html*
Ημερομηνία πρόσβασης: 10-08-2021

[27] Ward, J. H., Jr., *Hierarchical Grouping to Optimize an Objective Function*, Journal of the American Statistical Association, 58,1963, pp. 236–244.

[28] *Fuzzy Clustering*, Wolfram, https://reference.wolfram.com/legacy/applications/fuzzylogic/Manual/12.html Ημερομηνία πρόσβασης: 10-08-2021

[29] Satyam Kumar, *Fuzzy C-Means Clustering —Is it Better than K-Means Clustering?, https://towardsdatascience.com/fuzzy-c-means-clustering-is-it-better-than-k-means-clustering-448a0aba1ee7* Ημερομηνία πρόσβασης: 10-08-2021

[30] Qian Liu, Jianxin Liu, Min Li, Yang Zhou, *Approximation algorithms for fuzzy C-means problem based on seeding method*, Theoretical Computer Science,Volume 885,2021, Pages 146-158, ISSN 0304-3975 https://doi.org/10.1016/j.tcs.2021.06.035.

[31] John A. Bullinaria. *Self-organizing maps: Fundamentals*. 2004. http://www.cs.bham.ac.uk/~jxb/NN/l16.pdf Ημερομηνία πρόσβασης: 10-08-2021

[32] Jimin Qian, Nam Phuong Nguyen, Yutaka Oya, Gota Kikugawa, Tomonaga Okabe, Yue Huang, Fumio S. Ohuchi*, Introducing self-organized maps (SOM) as a visualization tool for materials research and education,*Results in Materials,Volume 4, 2019, 100020, ISSN 2590-048X, https://doi.org/10.1016/j.rinma.2019.100020 Ημερομηνία πρόσβασης: 10-08-2021

[33] Laura Frías-Paredes, Fermín Mallor, Martín Gastón-Romeo, Teresa León, *Assessing energy forecasting inaccuracy by simultaneously considering temporal and absolute errors,* Energy Conversion and Management, Volume 142,2017, Pages 533-546, ISSN 0196-8904, https://doi.org/10.1016/j.enconman.2017.03.056.

[34] Arnaud de Myttenaere, Boris Golden, Bénédicte Le Grand, Fabrice Ross*i, Mean Absolute Percentage Error for regression models,* Neurocomputing,Volume 192,2016, Pages 38-48, ISSN 0925 2312, https://doi.org/10.1016/j.neucom.2015.12.114.

[35] *Tic Matlab Documentation,* Mathworks, https://www.mathworks.com/help/matlab/ref/tic.html Ημερομηνία πρόσβασης: 12-08-2021