ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

“ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ”

Σύγκριση προβλεπτικών προσεγγίσεων και αλγορίθμων για το μονοξείδιο του άνθρακα

Από

Σπυρίδων Παπαθανασίου

Υποβάλλεται

για την εκπλήρωση των προϋποθέσεων λήψης

Μεταπτυχιακού Διπλώματος

στην ειδίκευση «ΜΔΑ»

του ΠΜΣ “Πληροφοριακά Συστήματα & Υπηρεσίες”

στο

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Μάϊος 2026

Επιβλέπων/Επιβλέπουσα: Μιχαήλ Φιλιππάκης

Ακαδημαϊκή Θέση: Καθηγητής

Comparison of predictive approaches and algorithms for carbon monoxide

By
Spyridon Papathanasiou

Submitted
in partial fulfilment of the requirements for the degree of
Master of Science – Digital Systems and Services: Big Data and Analytics
at the
UNIVERSITY OF PIRAEUS
May 2026

Thesis Supervisor: Michael Filippakis
Title: Professor

Συγγραφέας / Author: Σπυρίδων Παπαθανασίου

ΣΕΛΙΔΑ ΕΓΚΥΡΟΤΗΤΑΣ

Ονοματεπώνυμο Φοιτητή: Σπυρίδων Παπαθανασίου

Τίτλος Μεταπτυχιακής Διπλωματικής Εργασίας: Σύγκριση προβλεπτικών προσεγγίσεων και αλγορίθμων για το μονοξείδιο του άνθρακα

Η παρούσα Μεταπτυχιακή Διπλωματική Εργασία υποβάλλεται ως μερική εκπλήρωση των απαιτήσεων του Προγράμματος Μεταπτυχιακών Σπουδών "Πληροφοριακά Συστήματα & Υπηρεσίες" του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς και εγκρίθηκε στις 21/5/2026 από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή

Επιβλέπων: Μιχαήλ Φιλιππάκης, Καθηγητής, Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς

Μέλος Εξεταστικής Επιτροπής: Μαρία Χαλκίδη, Καθηγήτρια, Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς

Μέλος Εξεταστικής Επιτροπής: Δημοσθένης Κυριαζής, Καθηγητής, Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΑΥΘΕΝΤΙΚΟΤΗΤΑΣ

Ο Σπυρίδων Παπαθανασίου, γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Σύγκριση προβλεπτικών προσεγγίσεων και αλγορίθμων για το μονοξείδιο του άνθρακα», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο 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.

Επιπλέον *δηλώνω υπεύθυνα ότι η συγκεκριμένη Μεταπτυχιακή Διπλωματική Εργασία έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει αξιολογηθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό.*

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου. Σε κάθε περίπτωση, αναληθούς ή ανακριβούς δηλώσεως, υπόκειμαι στις συνέπειες που προβλέπονται τις διατάξεις που προβλέπει η Ελληνική και Κοινοτική Νομοθεσία περί πνευματικής ιδιοκτησίας.

Ο ΔΗΛΩΝ

Ονοματεπώνυμο: Σπυρίδων Παπαθανασίου
Αριθμός Μητρώου: ΜΕ2427
Υπογραφή: ΣΠ

Αρχικά, θα ήθελα να ευχαριστήσω ιδιαίτερα τον κ. Φιλιππάκη για την υποστήριξη που μου προσέφερε καθ’ όλη την διάρκεια του μεταπτυχιακού αλλά και για την εμπιστοσύνη που μου έδειξε με το να αναλάβω την συγκεκριμένη διπλωματική εργασία.

Παράλληλα, θέλω να ευχαριστήσω από καρδίας την σύντροφο μου, τους φίλους μου και την οικογένειά μου που με στηρίζουν όλα αυτά τα χρόνια, σε κάθε μου εγχείρημα.

Πίνακας Περιεχομένων

Κατάλογος - Πίνακες

Κατάλογος – Διαγράμματα

Abstract

This MSc dissertation focuses on the application of forecasting algorithms to air quality data. Specifically, the study uses the dataset known as the “UCI Air Quality” dataset, available from the UCI repository, and focuses on forecasting carbon monoxide (CO). Air quality data can be considered difficult to predict, as they often exhibit unstable variability and extreme values over irregular intervals. In addition, the specific dataset contains a significant number of missing values, making it challenging to use, as appropriate preprocessing is required before applying forecasting algorithms. For this reason, particular emphasis is placed on the detection and handling of missing values, through the application and evaluation of different imputation techniques. Beyond this, the study presents four levels of comparisons regarding the forecasting algorithms applied: the comparison of statistical models (SARMA – SARIMA) with machine learning models (Random Forest), the comparison of “multi-step” and “single-step” forecasting approaches for these models, which leads to the comparison of “short-term” and “medium-term” forecasting horizons, as well as the comparison between “univariate” and “multivariate” modeling approaches. The forecasting horizon is defined as one week, where “single-step” approaches focus on predicting one hour ahead iteratively over a one-week period, while “multi-step” approaches focus on directly predicting values for the entire one-week horizon.

Regarding the results, model performance is primarily evaluated using the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). In general, it is confirmed that single-step approaches outperform multi-step approaches. More specifically, Random Forest models produce the lowest error metrics. The univariate Random Forest model, using a single-step approach, achieves the best performance, with an MAE of 0,34 and an RMSE of 0,53.

Keywords: Forecasting, Air Quality, Imputation, CO, Single step, Multi step, SARIMA, Random Forest, Missing Values

Περίληψη

Η παρούσα διπλωματική εργασία εστιάζει στην εφαρμογή αλγόριθμων πρόβλεψης σε δεδομένα ποιότητας αέρα. Συγκεκριμένα, η εργασία χρησιμοποιεί το σύνολο δεδομένων γνωστό ως «UCI Air Quality», διαθέσιμο από το αποθετήριο UCI, και εστιάζει στην προβλεπτική του μονοξείδιου του άνθρακα (CO). Τα δεδομένα ποιότητας αέρα μπορούν να χαρακτηριστούν ως δύσκολα προς πρόβλεψη καθώς συχνά παρουσιάζουν ασταθή διακύμανση και ακραίες τιμές σε ασταθή διαστήματα ενώ παράλληλα το συγκεκριμένο σύνολο δεδομένων παρουσιάζει σημαντικό αριθμό ελλιπών τιμών, πράγμα που το καθιστά δύσκολο ως προς την χρήση του, καθώς είναι αναγκαία η κατάλληλη προ - επεξεργασία ώστε να μπορούν να εφαρμοστούν οι διάφοροι προβλεπτικοί αλγόριθμοι. Για αυτόν τον λόγο, στην εργασία δόθηκε ιδιαίτερη έμφαση στην ανίχνευση και αντιμετώπιση των ελλιπών τιμών, μέσω εφαρμογής και αξιολόγησης διαφορετικών τρόπων αντιμετώπισης του προαναφερόμενου προβλήματος. Πέρα από αυτά, η εργασία παρουσιάζει 4 επίπεδα ερωτημάτων - συγκρίσεων σχετικά με τους προβλεπτικούς αλγόριθμους που εφαρμόστηκαν: Την σύγκριση στατιστικών μοντέλων (SARMA – SARIMA) με μοντέλα μηχανικής μάθησης (Random Forest), την σύγκριση προσεγγίσεων “multi-step” και “single-step” για τα προαναφερόμενα μοντέλα, που οδηγούν στην σύγκριση των δημιουργούμενων “βραχυπρόθεσμων” και “μεσοπρόθεσμων” οριζόντων πρόβλεψης, καθώς και την σύγκριση “μονομεταβλητών” και “πολυμεταβλητών” προσεγγίσεων των προαναφερόμενων αλγορίθμων. Ο ορίζοντας πρόβλεψης ορίστηκε ως 1 εβδομάδα, όπου οι προσεγγίσεις “singe-step” εστιάζουν στην πρόβλεψη μίας ώρας για διάρκεια μίας εβδομάδας ενώ οι προσεγγίσεις “multi-step” εστιάζουν στην πρόβλεψη τιμών μιας εβδομάδας.

Ως προς τα αποτελέσματα, αξιολογήθηκαν κυρίως με βάση τις μετρικές Mean Absolute Error (MAE) και Root Mean Square Error (RMSE). Γενικά, επιβεβαιώθηκε ότι οι προσεγγίσεις single – step υπερτερούν σε σχέση με τις multi – step προσεγγίσεις ενώ πιο συγκεκριμένα τα μοντέλα Random Forest παρήγαγαν τις χαμηλότερες μετρικές σφαλμάτων. Το μονοδιάστατο (univariate) μοντέλο Random Forest, με προσέγγιση single – step παρήγαγε τις χαμηλότερες μετρικές σφάλματος, καθώς το MAE βρέθηκε ως 0,34 και το RMSE ως 0,53.

Λέξεις κλειδιά: Προβλέψεις, Ποιότητα αέρα, Αντιμετώπιση ελλιπών τιμών, CO, Single step, Multi step, SARIMA, Random Forest

1 Εισαγωγή - Βιβλιογραφική ανασκόπηση

Η ραγδαία ανάπτυξη της σύγχρονης βιομηχανίας και των μεταφορών, σε συνδυασμό με την έντονη πληθυσμιακή αύξηση και την αστικοποίηση, έχει καταστήσει την ατμοσφαιρική ρύπανση ένα παγκόσμιο ζήτημα. Πολλοί ατμοσφαιρικοί ρύποι υποβαθμίζουν το περιβάλλον και συμβάλλουν σε σοβαρά περιβαλλοντικά φαινόμενα, όπως το φαινόμενο του θερμοκηπίου, η καταστροφή του όζοντος και το φωτοχημικό νέφος. Οι επιπτώσεις αυτές αυξάνουν σημαντικά τους κινδύνους για την ανθρώπινη υγεία παγκοσμίως, προκαλώντας ασθένειες όπως αναπνευστικές διαταραχές, χρόνιες και καρδιαγγειακές παθήσεις, ακόμη και καρκίνο [1]. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (WHO), η ατμοσφαιρική ρύπανση χαρακτηρίζεται εξαιρετικά επικίνδυνη καθώς συμβάλει στον πρόωρο θάνατο περίπου 7.000.000 ατόμων ανά έτος όπου τα 600.000 από αυτά τα άτομα είναι παιδιά [2]. Οι επιδράσεις της ατμοσφαιρικής ρύπανσης στην ανθρώπινη υγεία είναι πολυδιάστατες και μπορούν να επηρεάσουν σχεδόν όλα τα βασικά συστήματα του οργανισμού. Μεταξύ αυτών περιλαμβάνονται το αναπνευστικό και το ανοσοποιητικό σύστημα, το δέρμα και οι βλεννογόνοι, τα αισθητήρια όργανα, καθώς και το κεντρικό και περιφερικό νευρικό σύστημα και το καρδιαγγειακό [2].

Ειδικότερα, στο κατώτερο αναπνευστικό σύστημα η έκθεση σε ατμοσφαιρικούς ρύπους μπορεί να προκαλέσει τόσο βραχυπρόθεσμες όσο και μακροχρόνιες διαταραχές της πνευμονικής λειτουργίας, αύξηση της εμφάνισης αναπνευστικών συμπτωμάτων, μεγαλύτερη ευαισθησία των αεραγωγών σε αλλεργιογόνους παράγοντες και επιδείνωση λοιμώξεων του αναπνευστικού, όπως ρινίτιδα, ιγμορίτιδα, πνευμονία και νόσο των λεγεωνάριων. Καθοριστικό ρόλο στην πρόκληση αυτών των επιπτώσεων διαδραματίζουν κυρίως οι ρύποι που προέρχονται από διαδικασίες καύσης, όπως το διοξείδιο του θείου (SO_2), το διοξείδιο του αζώτου (NO_2), τα αιωρούμενα σωματίδια με αεροδυναμική διάμετρο μικρότερη από 10 μm (SPM), καθώς και το μονοξείδιο του άνθρακα (CO) [25].

Οι κυριότεροι ατμοσφαιρικοί ρύποι περιλαμβάνουν το SO_2, τα αιωρούμενα σωματίδια TSP - Total Suspended Particulate (σκόνη, PM10, PM2.5), τα NOx, το CO και το O_3, τα οποία αποτελούν μείγμα σωματιδίων και αερίων και αποτελούν ολοένα αυξανόμενο πρόβλημα. Τα αιωρούμενα σωματίδια (PM) δεν αποτελούν έναν ενιαίο τύπο σωματιδίων, αλλά ένα σύνολο μικροσκοπικών σωματιδίων που βρίσκονται στην ατμόσφαιρα, συμβάλλοντας στην υποβάθμιση της ποιότητας του αέρα και στην συχνή εμφάνιση αιθαλομίχλης (haze). Έρευνες έχουν δείξει ότι σωματίδια με διάμετρο έως 10 μm μπορούν να φτάσουν στο ανώτερο αναπνευστικό σύστημα, ενώ σωματίδια μικρότερα από 5 μm είναι δυνατόν να εισχωρήσουν βαθύτερα στους βρόγχους. Επιπλέον, σωματίδια μικρότερα από 1 μm μπορούν να διεισδύσουν μέχρι και στις κυψελίδες των πνευμόνων. Το διοξείδιο του θείου (SO_2) παράγεται κυρίως από ηφαιστειακές εκρήξεις αλλά και από βιομηχανικές δραστηριότητες [1].

Δεν υπάρχει σαφής κατηγοριοποίηση που να παρουσιάζει με ακρίβεια τον βαθμό επικινδυνότητας κάθε ρύπου. Τα μόνα δεδομένα που θα μπορούσαν να χρησιμοποιηθούν για την αξιολόγηση της επικινδυνότητας των ρύπων είναι εκείνα που σχετίζονται με τους πρόωρους θανάτους που αποδίδονται σε αυτούς. Παρ' όλα αυτά, λόγω της δυσκολίας στον ακριβή προσδιορισμό αυτής της σχέσης, τέτοιου είδους πληροφορίες είναι περιορισμένες [2].

Τα ορυκτά καύσιμα, όπως ο άνθρακας και το πετρέλαιο, περιέχουν θείο και κατά την καύση τους απελευθερώνεται SO_2 στην ατμόσφαιρα. Η παρουσία του μπορεί επίσης να συμβάλει στη δημιουργία όξινης βροχής, προκαλώντας ζημιές στο φυσικό και τεχνητό περιβάλλον, ενώ παράλληλα συνδέεται με αυξημένη συχνότητα αναπνευστικών παθήσεων [1]. Οι περισσότερες εκπομπές διοξειδίου του αζώτου (NO_2) προέρχονται από την καύση ορυκτών καυσίμων σε βιομηχανικές δραστηριότητες και οχήματα. Το NO_2 αποτελεί επίσης έναν από τους βασικούς παράγοντες που συμβάλλουν στη

δημιουργία όξινης βροχής και συμμετέχει σημαντικά στον σχηματισμό του όζοντος. Παράλληλα, επηρεάζει αρνητικά την ανθρώπινη υγεία και τα οικοσυστήματα, αυξάνοντας την εμφάνιση πνευμονικών ασθενειών [1]. Επίσης, έχουν παρατηρηθεί ασθενείς συσχετίσεις μεταξύ της βραχυχρόνιας έκθεσης σε NO_2, που προέρχεται από τη χρήση αερίου στο μαγείρεμα, και της εμφάνισης αναπνευστικών συμπτωμάτων καθώς και της μείωσης ορισμένων παραμέτρων της πνευμονικής λειτουργίας στα παιδιά, χωρίς όμως να παρατηρείται σταθερά το ίδιο φαινόμενο σε εκτεθειμένες γυναίκες. Όσον αφορά τη μακροχρόνια έκθεση, τα παιδιά σε αντίθεση με τους ενήλικες, παρουσιάζουν αυξημένα αναπνευστικά συμπτώματα, μειωμένη πνευμονική λειτουργία και μεγαλύτερη συχνότητα εμφάνισης χρόνιου βήχα, βρογχίτιδας και επιπεφυκίτιδας. Παρ' όλα αυτά, δεν έχει ακόμη αποδειχθεί σαφής αιτιώδης σχέση μεταξύ της έκθεσης σε NO_2 και των δυσμενών επιπτώσεων στην υγεία [25].

Το όζον (O_3), λόγω της ισχυρής οξειδωτικής του δράσης, θεωρείται επιβλαβές αέριο όταν βρίσκεται σε υψηλές συγκεντρώσεις κοντά στην επιφάνεια του εδάφους. Σε αυξημένα επίπεδα μπορεί να προκαλέσει βλάβες στο ανώτερο αναπνευστικό σύστημα, καθώς και ερεθισμούς στο δέρμα, τα μάτια και τη μύτη. Επιπλέον, οι υψηλές συγκεντρώσεις O_3 επηρεάζουν αρνητικά την παραγωγή γεωργικών καλλιεργειών, με τις απώλειες στην παραγωγή να αναμένεται να αυξηθούν στο μέλλον [1].

Το μονοξείδιο του άνθρακα (CO) είναι ένα άχρωμο και άοσμο αέριο που παράγεται κυρίως από την ατελή καύση καυσίμων που περιέχουν άνθρακα. Στα αστικά περιβάλλοντα, η οδική κυκλοφορία αποτελεί τη σημαντικότερη πηγή εκπομπών του. Οι συγκεντρώσεις CO στην ατμόσφαιρα επηρεάζονται σε μεγάλο βαθμό από την πυκνότητα των οχημάτων και συνήθως παρουσιάζουν υψηλότερες τιμές κατά τις ώρες αιχμής, δηλαδή το πρωί και το απόγευμα [25]. Το μονοξείδιο του άνθρακα (CO) συνδέεται στους πνεύμονες με την αιμοσφαιρίνη του αίματος, σχηματίζοντας καρβοξυαιμοσφαιρίνη (COHb), γεγονός που μειώνει την ικανότητα μεταφοράς οξυγόνου στον οργανισμό. Οι επιπτώσεις του CO στην υγεία περιλαμβάνουν υποξία, νευρολογικές διαταραχές και αλλαγές στη νευροσυμπεριφορά, καθώς και αύξηση της ημερήσιας θνησιμότητας και των εισαγωγών στα νοσοκομεία λόγω καρδιαγγειακών παθήσεων. Τα φαινόμενα αυτά έχουν παρατηρηθεί ακόμη και σε πολύ χαμηλές συγκεντρώσεις CO, γεγονός που υποδηλώνει ότι δεν υπάρχει σαφές κατώφλι συγκέντρωσης για την εμφάνισή τους. Ωστόσο, παραμένει ασαφές αν η σχέση μεταξύ της ημερήσιας θνησιμότητας και της έκθεσης σε CO είναι άμεση αιτιώδης ή αν το CO λειτουργεί ως δείκτης παρουσίας αιωρούμενων σωματιδίων (SPM – Suspended Particulate Matter). Επιπλέον, το CO που υπάρχει στην ατμόσφαιρα μπορεί να έχει ακόμη πιο σοβαρές επιπτώσεις στην υγεία από εκείνες που σχετίζονται αποκλειστικά με τον σχηματισμό της καρβοξυαιμοσφαιρίνης, ακόμη και σε χαμηλότερα επίπεδα από αυτά που προκαλούν αύξηση της COHb [25].

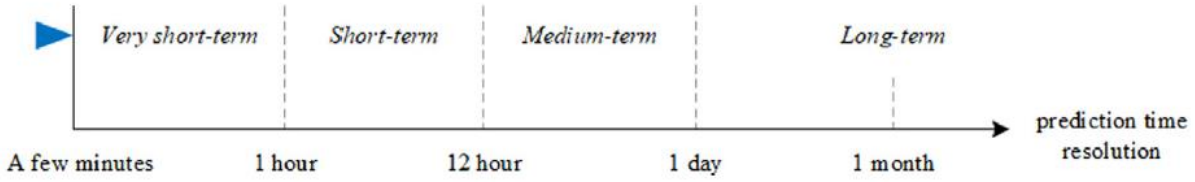
Σύμφωνα με στοιχεία του Ευρωπαϊκού Οργανισμού Περιβάλλοντος, για το έτος 2018 εκτιμάται ότι τα PM2.5, το NO_2 και το O_3 προκάλεσαν περίπου 379.000, 54.000 και 19.000 πρόωρους θανάτους αντίστοιχα στην Ευρωπαϊκή Ένωση [2]. Για άλλους ρύπους δεν υπάρχουν διαθέσιμα αντίστοιχα δεδομένα. Επιπλέον, σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, τα σωματίδια PM10 θεωρούνται επιβλαβή για την ανθρώπινη υγεία, αν και σε μικρότερο βαθμό σε σύγκριση με τα PM2.5 λόγω του μεγαλύτερου μεγέθους τους. Σχετικά με το CO, η Επιτροπή του Ηνωμένου Βασιλείου για τις Ιατρικές Επιπτώσεις των Ατμοσφαιρικών Ρύπων αναφέρει ότι τα τελευταία χρόνια παρατηρείται σημαντική μείωση των συγκεντρώσεών του στο εξωτερικό περιβάλλον, γεγονός που έχει συμβάλει και στη μείωση της επικινδυνότητάς του [2].

Η ρύπανση στους αστικούς δρόμους παρακολουθείται συστηματικά σε πολλές πόλεις και οι μετρήσεις δείχνουν ότι οι μέγιστες βραχυχρόνιες συγκεντρώσεις ρύπων όπως το CO, το NO_2 και τα αιωρούμενα σωματίδια μπορεί να υπερβαίνουν τα επιτρεπόμενα όρια ποιότητας αέρα κατά δύο έως και τέσσερις φορές, ανάλογα με την ένταση της κυκλοφορίας και τις συνθήκες διασποράς των ρύπων

στον δρόμο. Στην Ευρώπη, εκτιμάται ότι μεταξύ 9 και 18 εκατομμυρίων ανθρώπων εκτίθενται σε τόσο υψηλά επίπεδα ρύπανσης. Οι οδικές μεταφορές αποτελούν σημαντική πηγή δημιουργίας αιθαλομίχλης (smog), καθώς και μακροχρόνιων μέσων συγκεντρώσεων ουσιών όπως ο μόλυβδος, το βενζόλιο, τα αιωρούμενα σωματίδια και το βενζο[α]πυρένιο. Κατά μέσο όρο, η κυκλοφορία των οχημάτων ευθύνεται για περισσότερο από το ήμισυ των συγκεντρώσεων NO_2 και περίπου για το 40% των συγκεντρώσεων πτητικών οργανικών ενώσεων (VOC). Σε αρκετές πόλεις, η συμβολή της οδικής κυκλοφορίας στη ρύπανση από τις συγκεκριμένες ουσίες είναι ακόμη μεγαλύτερη [25].

Οι παραπάνω παρατηρήσεις συμβάλλουν στην κατανόηση της συσχέτισης μεταξύ της επικινδυνότητας των ρύπων και του αριθμού των επιστημονικών μελετών που επικεντρώνονται στην πρόβλεψή τους. Αξίζει να σημειωθεί ότι πολλές από τις σχετικές μελέτες εξετάζουν ταυτόχρονα περισσότερους από έναν ρύπους [2].

Βάσει όλων των προαναφερόμενων, γίνεται αντιληπτό ότι οι ατμοσφαιρικοί ρύποι μπορούν να επιφέρουν σημαντικές επιπτώσεις στην υγεία του πληθυσμού, επομένως είναι αναγκαία η μελέτη και κατ' επέκταση, η δημιουργία μοντέλων πρόβλεψης για τις ρυπογόνες ενώσεις για την πρόωρη αποτύπωση της κατάστασης και αντιμετώπιση των επιπτώσεων. Η πρόγνωση της ποιότητας του αέρα μπορεί να παρέχει χρήσιμα δεδομένα σχετικά με την περιβαλλοντική κατάσταση τόσο για την κοινωνία όσο και για τις κυβερνήσεις, ενώ παράλληλα επιτρέπει την έγκαιρη αποτύπωση των τάσεων της περιβαλλοντικής ρύπανσης. Υπάρχουν ορισμένοι βασικοί δείκτες για την κατά προσέγγιση ταξινόμηση των μοντέλων πρόγνωσης της ποιότητας του αέρα, όπως η χρονική κλίμακα των προβλέψεων, η μεθοδολογία πρόγνωσης και ο τύπος των δεδομένων εισόδου. Με βάση τη χρονική ανάλυση των δεδομένων, η πρόβλεψη της ποιότητας του αέρα μπορεί να διακριθεί σε πολύ βραχυπρόθεσμη, βραχυπρόθεσμη, μεσοπρόθεσμη και μακροπρόθεσμη, όπως φαίνεται από το διάγραμμα 1 [1, 26].



Διάγραμμα 1 Κατηγοριοποίηση χρονικού ορίζοντα. Πηγή: [1]

Κοινώς, δεν έχει νόημα να χρησιμοποιηθεί το μέγεθος του συνόλου δεδομένων ως ένδειξη του χρονικού ορίζοντα αλλά η χρονική κλίμακα που χρησιμοποιείται. Για παράδειγμα, 2 χρονοσειρές μπορούν να έχουν το ίδιο χρονικό μήκος αλλά να χρησιμοποιούν διαφορετική χρονική κλίμακα για τις παρατηρήσεις τους. Έστω μια χρονοσειρά με 96 παρατηρήσεις, εάν η χρονική κλίμακα είναι 15 λεπτά τότε το σετ δεδομένων ανταποκρίνεται σε μια ημέρα (1,440 λεπτά ή 24 ώρες) ενώ εάν η χρονική κλίμακα είναι 1 ώρα τότε ανταποκρίνεται σε 4 ημέρες δεδομένων (96 ώρες) [26].

Οι Liu et al. [1] αναφέρουν ότι οι περισσότερες μελέτες που διενεργούνται στα πλαίσια της πρόβλεψής ποιότητας του αέρα, χρησιμοποιούν ωριαία ή ημερήσια χρονική κλίμακα, τόσο στα δεδομένα που χρησιμοποιούνται για την εκπαίδευση των μοντέλων όσο και στον χρονικό ορίζοντα πρόβλεψης ,όπου τα δεδομένα προέρχονται από μελέτες πάνω του ενός έτους, που εμπεριέχουν στοιχεία για την μακροχρόνια τάση της κατάστασης [1].

Γενικά, ένας πιο σύντομος χρονικός ορίζοντας πρόβλεψης μπορεί να επιτύχει αποτελέσματα με υψηλότερη ακρίβεια, ενώ ένας μεγαλύτερος χρονικός ορίζοντας μπορεί να παρέχει μακροπρόθεσμες πληροφορίες, όπως την μακροχρόνια τάση του φαινόμενου. Η αξιοποίηση τέτοιων προβλέψεων

συμβάλλει στον σχεδιασμό και την εφαρμογή στρατηγικών μακροχρόνιου ελέγχου της ρύπανσης, ενώ οι προβλέψεις σε ωριαία βάση μπορούν να υποστηρίξουν την παρακολούθηση της ποιότητας του αέρα και τη βραχυπρόθεσμη διαχείρισή της με μεγαλύτερη ακρίβεια. Η ανάπτυξη μοντέλων ποιότητας αέρα αποτελεί μια σύνθετη διαδικασία συστημικής μηχανικής και ένα απαιτητικό πεδίο έρευνας στις περιβαλλοντικές επιστήμες, καθώς συμβάλλει στη μελέτη της σχέσης μεταξύ των αιτίων και των επιπτώσεων των ρύπων και υποστηρίζει την ανάπτυξη αποτελεσματικών λύσεων για τη μείωση της ρύπανσης στο μέλλον μέσω κατάλληλης ανάλυσης [1]. Παράλληλα, χρειάζεται να αναφερθεί ότι η επιλογή βραχυπρόθεσμου ή μακροπρόθεσμου χρονικού ορίζοντα πρόβλεψης εμπίπτει στο φαινόμενο ή στα κλάδο των δεδομένων υπό μελέτη.

Για την κατηγοριοποίηση της ποιότητας του αέρα χρησιμοποιείται ο Δείκτης Ποιότητας Αέρα (Air Quality Index). Ο Δείκτης Ποιότητας Αέρα (AQI) είναι μια τμηματικά γραμμική συνάρτηση που βασίζεται στις συγκεντρώσεις συγκεκριμένων ατμοσφαιρικών ρύπων, όπως το όζον (O_3), τα αιωρούμενα σωματίδια (PM2.5 και PM10), το μονοξείδιο του άνθρακα (CO), το διοξείδιο του θείου (SO_2) και το διοξείδιο του αζώτου (NO_2). Ωστόσο, δεν υπάρχει ένα ενιαίο παγκόσμιο πρότυπο για τον AQI, καθώς διαφορετικές χώρες και περιοχές χρησιμοποιούν δικούς τους δείκτες, οι οποίοι βασίζονται στα αντίστοιχα εθνικά πρότυπα ποιότητας αέρα [2].

Μάλιστα, οι περισσότερες μελέτες εστιάζουν στην πρόβλεψη του AQI, που αποτελεί τη μεταβλητή που προβλέπεται συχνότερα στις σχετικές μελέτες, ενώ από τις συγκεντρώσεις ρύπων, η πιο συχνά προβλεπόμενη είναι εκείνη των αιωρούμενων σωματιδίων PM2.5. Οι μεταβλητές που χρησιμοποιούνται συχνότερα ως προβλεπτικοί παράγοντες είναι τα χαρακτηριστικά των ρύπων (περίπου 50%), ακολουθούμενα από τα μετεωρολογικά χαρακτηριστικά (περίπου 35%), ενώ ο συνδυασμός και των δύο κατηγοριών μεταβλητών φαίνεται να προσφέρει την καλύτερη ακρίβεια στις προβλέψεις. Επιπλέον, ενδείξεις δείχνουν ότι οι αλγόριθμοι βαθιάς μάθησης είναι πιο αποτελεσματικοί σε σύγκριση με τις μεθόδους παλινδρόμησης. Παρά τη σημαντική αύξηση της χρήσης της βαθιάς μάθησης τα τελευταία χρόνια, η αναλογία των αλγορίθμων βαθιάς μάθησης σε σχέση με τους αλγορίθμους παλινδρόμησης παραμένει περίπου σταθερή με την πάροδο του χρόνου [2].

Ο AQI χρησιμοποιείται από διάφορους οργανισμούς και θεσμούς για την ενημέρωση του πληθυσμού σχετικά με τον βαθμό ατμοσφαιρικής ρύπανσης σε πόλεις ή γειτονιές, λαμβάνοντας υπόψη τις συγκεντρώσεις των βασικών ρύπων στην ατμόσφαιρα. Για τον λόγο αυτό, είναι λογικό ο AQI να αποτελεί τη συχνότερα μελετώμενη μεταβλητή σε έρευνες που αφορούν την πρόβλεψη της ποιότητας του αέρα. Αξίζει επίσης να σημειωθεί ότι διαφορετικές χώρες χρησιμοποιούν διαφορετικές κλίμακες AQI. Ενδεικτικά, μπορούν να αναφερθούν οι ακόλουθοι δείκτες [2]:

- Common Air Quality Index, που χρησιμοποιείται στην Ευρώπη από το 2006, περιλαμβάνει 5 κατηγορίες κινδύνου και κυμαίνεται από 0 (χαμηλός κίνδυνος) έως 100 (ανθυγιεινή ποιότητα αέρα).

- China Air Quality Index, ο οποίος περιλαμβάνει 6 κατηγορίες κινδύνου και κυμαίνεται από 0 (εξαιρετική ποιότητα αέρα) έως 500 (πολύ σοβαρή ρύπανση).

- US Air Quality Index, που κυμαίνεται επίσης από 0 έως 500 και χωρίζεται σε 6 κατηγορίες που αντιστοιχούν σε διαφορετικά επίπεδα ποιότητας αέρα.

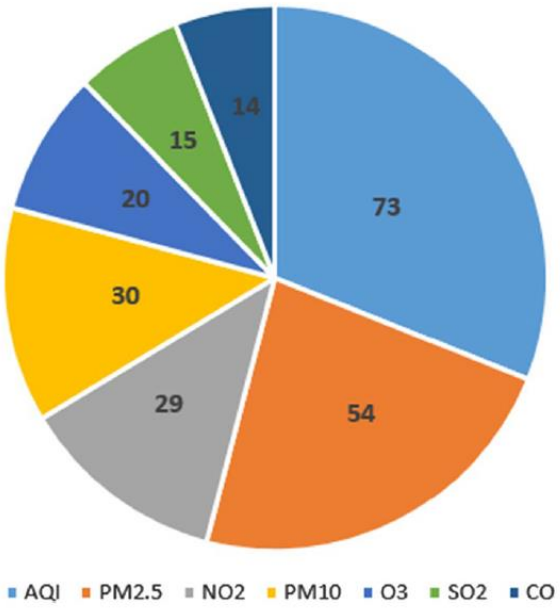
- Στο Ηνωμένο Βασίλειο, ο αντίστοιχος δείκτης κυμαίνεται από 0 έως 10 και κατατάσσει τις τιμές αυτές σε 4 επίπεδα ποιότητας αέρα.

Όσο αφορά τους αλγόριθμους που χρησιμοποιούνται για την παραγωγή προβλεπτικών μοντέλων, ουσιαστικά μπορούν να ομαδοποιηθούν σε τρείς διακριτές κατηγορίες: Στατιστικά ή γραμμικής παλινδρόμησης, μηχανικής μάθησης και βαθιάς μάθησης. Στην παρούσα εργασία θα χρησιμοποιηθούν αλγόριθμοι στατιστικής και μηχανικής μάθησης. Οι στατιστικές μέθοδοι αποτυπώνουν κυρίως τις στατιστικές συσχετίσεις μεταξύ διαφόρων παραγόντων και των ατμοσφαιρικών ρύπων σε χρονοσειρές, χρησιμοποιώντας ιστορικά δεδομένα για την πρόβλεψη της ποιότητας του αέρα, αντί να βασίζονται σε φυσικές, χημικές ή βιολογικές διεργασίες. Πρόκειται για μοντέλα που βασίζονται στα δεδομένα και στηρίζονται στη θεωρία της στατιστικής, της πιθανότητας και των στοχαστικών διεργασιών. Τα παραδοσιακά στατιστικά μοντέλα που χρησιμοποιούνται στην πρόγνωση της ατμοσφαιρικής ρύπανσης περιλαμβάνουν το αυτοπαλίνδρομο ολοκληρωμένο μοντέλο κινητού μέσου (ARIMA – Autoregressive Integrated Moving Average), το μοντέλο Grey (GM – Grey Model) και διάφορα μοντέλα παλινδρόμησης. Η ακρίβεια πρόβλεψης του μοντέλου GM εξαρτάται σε μεγάλο βαθμό από τα χαρακτηριστικά των δεδομένων και τις παραμέτρους του μοντέλου [1]. Τα μοντέλα παλινδρόμησης που χρησιμοποιούνται για την πρόβλεψη των συγκεντρώσεων ρύπων περιλαμβάνουν κυρίως τη βηματική παλινδρόμηση (stepwise regression), την παλινδρόμηση κύριων συνιστωσών (PCR – Principal Component Regression) και την πολλαπλή γραμμική παλινδρόμηση (MLR – Multiple Linear Regression) [1]. Τα στατιστικά μοντέλα βασίζονται στην περιγραφή των σχέσεων μεταξύ μεταβλητών μέσω πιθανοτήτων και στατιστικών μέσων τιμών και μπορούν γενικά να επιτύχουν ικανοποιητική ακρίβεια στην πρόβλεψη των μελλοντικών επιπέδων συγκέντρωσης ρύπων. Ωστόσο, εξακολουθούν να υπάρχουν περιθώρια βελτίωσης όσον αφορά την ακρίβεια των προβλέψεων, καθώς η συμπεριφορά των ατμοσφαιρικών ρύπων και τα ιδιαίτερα χαρακτηριστικά κάθε περιοχής μπορεί να είναι πολύπλοκα, ακανόνιστα και έντονα μη γραμμικά. Για τον λόγο αυτό, απαιτούνται πιο αποτελεσματικές προσεγγίσεις για την καλύτερη μοντελοποίηση και πρόβλεψη της ποιότητας του αέρα [1].

Όπως αναφέρθηκε προηγουμένως, αρκετές μελέτες χρησιμοποιούν κάποιον δείκτη AQI για την αξιολόγηση της ποιότητας του αέρα ωστόσο αυτό δεν μπορεί να εφαρμοστεί στο παρόν σετ δεδομένων «UCI Air Quality» για δύο βασικούς λόγους. Αρχικά, θα μπορούσαν να χρησιμοποιηθούν μόνο οι μετρήσεις που δεν προέρχονται από την πειραματική συσκευή των 5 χημικών αισθητήρων μεταλλικού οξειδίου [3], δηλαδή μόνο όσες ενώσεις έχουν μετρηθεί ως “ground truth (GT)”, αφού οι συγκεκριμένες μεταβλητές – μετρήσεις, είναι οι ακατέργαστες “έξοδοι” των αισθητήρων. Δηλαδή, δεν μπορούν να χρησιμοποιηθούν για τον δείκτη AQI ή οποιαδήποτε άλλη πρακτική χρήση, εκτός εάν μετατραπούν σε συγκεντρώσεις ρύπων μέσω Δηλαδή, για την χρήση των παραπάνω μετρικών είναι απαραίτητες εξειδικευμένες γνώσεις (Baseline subtraction - normalization, Drift correction, Temperature/humidity compensation), οι οποίες δεν βρίσκονται εντός πλαισίου της παρούσας διπλωματικής εργασίας. Επομένως οι ενώσεις που παρέχονται ως Ground Truth (GT) και χρησιμοποιούνται στους περισσότερους δείκτες AQI είναι το μονοξείδιο του άνθρακα (CO) και το διοξειδίου του αζώτου (NO_2). Είναι πιθανό τα παραπάνω να είναι οι βασικοί λόγοι βάσει των οποίων οι Kumar et al. [5] εστίασαν την έρευνα τους αποκλειστικά στις ενώσεις CO και NO_2. Μια άλλη προσέγγιση από τους Kaur και Sandha [28] είναι η κατασκευή ενός απλοποιημένου δείκτη τύπου AQI, συγκεκριμένα χρησιμοποιώντας τις ενώσεις CO(GT), NO2(GT) and C6H6 [28]. Σε αντίθεση, οι Liu et al. [4] δεν εστίασαν σε ενώσεις που περιέχονται συχνά στους δείκτες AQI, ούτε κατασκεύασαν κάποιον διαφορετικό δείκτη αλλά εστίασαν στο NOx. Ανεξάρτητα της μεθοδολογίας που χρησιμοποιήθηκε, όλες οι προαναφερόμενες πηγές χρησιμοποίησαν μεταβλητές που δεν προέκυψαν από τους χημικούς

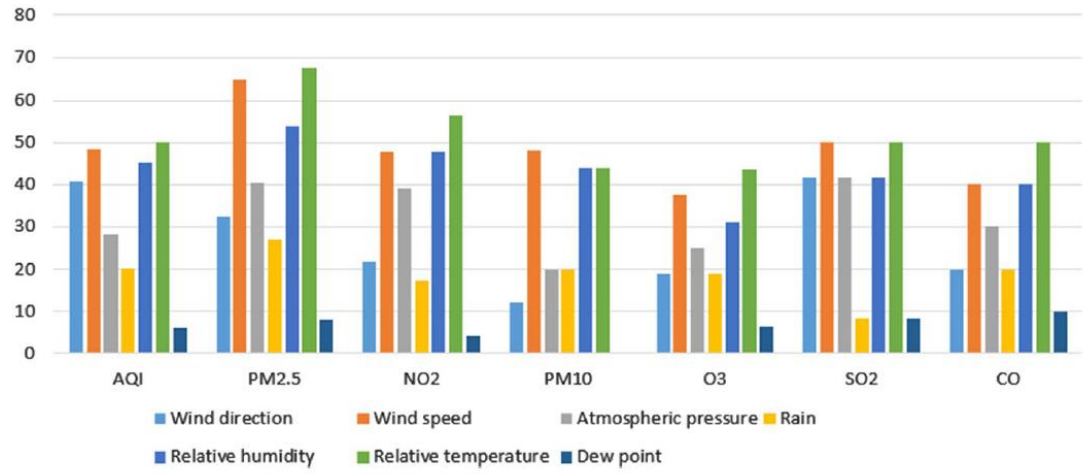
αισθητήρες. Η παρούσα διπλωματική εργασία εστιάζει αποκλειστικά στο μονοξείδιο του άνθρακα (CO) καθώς είναι μια από τις δύο ρυπογόνες ενώσεις που χρησιμοποιείται συχνά στους δείκτες AQI και παράλληλα παρέχεται από το σετ δεδομένων. Επιπλέον, είναι μια από τις λιγότερο μελετημένες ενώσεις στα πλαίσια της ατμοσφαιρικής ρύπανσης όπως φαίνεται από το διάγραμμα 2 [2].

Fig. 7 Publications by predicted pollutants



Διάγραμμα 2 Αριθμός δημοσιευμένων ερευνών ανά ρυπογόνο ένωση. Πηγή: [2]

Συγκεκριμένα στο διάγραμμα 2, οι Mendez, Merayo και Núñez παρουσιάζουν ένα σημαντικό πλήθος δημοσιεύσεων για την περίοδο 2011–2021, βάσει της προβλεπόμενης ρυπογόνου ένωσης. αναδεικνύουν ότι οι περισσότερες δημοσιεύσεις (73) σχετικά με την πρόβλεψη αέριων ρυπογόνων θέτουν ως τιμή στόχο τον δείκτη AQI. Πέρα από τον δείκτη AQI, οι λοιπές δημοσιεύσεις θέτουν ως τιμή στόχο τους παρακάτω ρύπους, κατά φθίνουσα σειρά: PM2.5 (54 δημοσιεύσεις), PM10 (30 δημοσιεύσεις), NO2 (29 δημοσιεύσεις) O3 (20 δημοσιεύσεις), SO2(15 δημοσιεύσεις) και CO(14 δημοσιεύσεις) [2]. Παράλληλα, οι ίδιοι παρουσίασαν και τις πιο συχνές μετεωρολογικές – καιρικές μεταβλητές που χρησιμοποιούνται για την πρόβλεψη αέριων ρυπογόνων [2] όπως φαίνεται στο παρακάτω διάγραμμα 3.



Διάγραμμα 3 Συχνά χρησιμοποιούμενες μετεωρολογικές μεταβλητές ανά ρυπογόνο ένωση. Πηγή: [2]

Το διάγραμμα 3 δείχνει ότι ανεξάρτητα ορισμένης μεταβλητής – στόχου, οι 3 πιο συχνές μετεωρολογικές μεταβλητές κατά φθίνουσα σειρά χρήσης, είναι η σχετική θερμοκρασία (relative temperature), ταχύτητα αέρα (wind speed) και σχετική υγρασία (relative temperature) [2]. Εστιάζοντας στο διοξείδιο του άνθρακα, είναι πολύ θετικό ότι το σετ δεδομένων του UCI, παρέχει δύο από τις τρείς πιο συχνά χρησιμοποιούμενες μετεωρολογικές μεταβλητές, την θερμοκρασία και την σχετική υγρασία.

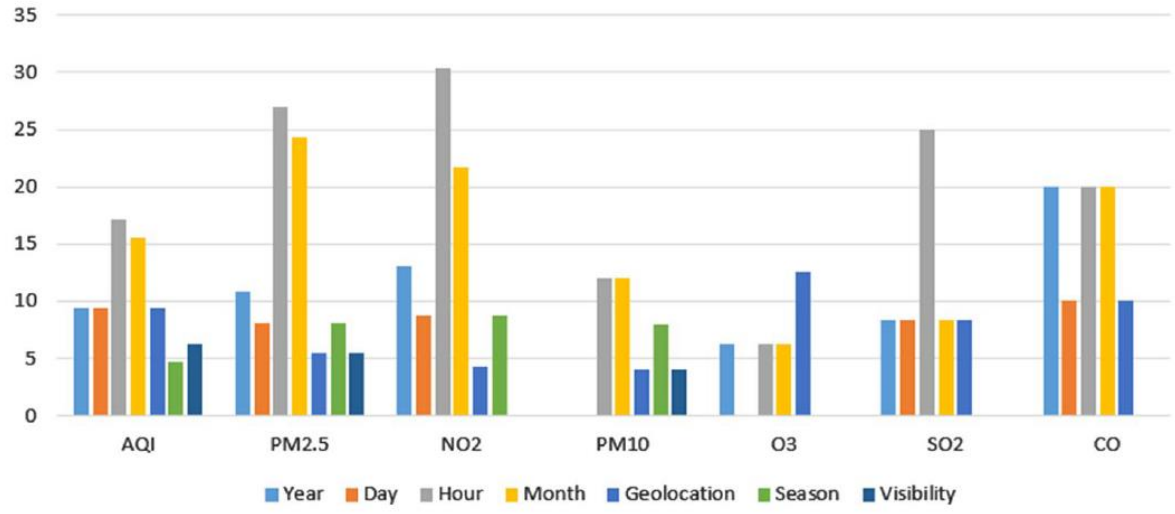


Fig. 10 Other variables by dependent variable (percentage)

Διάγραμμα 4 Συχνά χρησιμοποιούμενες λοιπές μεταβλητές ανά ρυπογόνο ένωση. Πηγή: [2]

Πέρα από τις μετεωρολογικές μεταβλητές, οι Mendez, Merayo και Núñez [2], συγκέντρωσαν και τις πιο συχνές μη – μετεωρολογικές μεταβλητές ανά προβλεπόμενη ρυπογόνο ένωση. Το διάγραμμα 4 παρουσιάζει την επίδραση πρόσθετων μεταβλητών στα προτεινόμενα μοντέλα. Οι μεταβλητές εξετάστηκαν αφορούσαν κυρίως την χρονική διάσταση, όπως η ώρα, η ημέρα, ο μήνας και το έτος. Επιπλέον, εξετάστηκε και η επίδραση της γεωγραφικής θέσης, δηλαδή των γεωγραφικών

συντεταγμένων και της ορατότητας. Γενικά, οι μεταβλητές που χρησιμοποιούνται συχνότερα είναι ο μήνας και η ώρα, καθώς εμφανίζονται περίπου στο 25% των μοντέλων. Οι προβλεπτικές αυτές μεταβλητές θεωρούνται ιδιαίτερα σημαντικές, καθώς λειτουργούν ως εξωτερικοί παράγοντες που επηρεάζουν τις προβλέψεις. Όσον αφορά την ώρα, για παράδειγμα, η νυχτερινή περίοδος συνδέεται συνήθως με μειωμένη κυκλοφορία και περιορισμένη βιομηχανική δραστηριότητα, γεγονός που μπορεί να οδηγήσει σε χαμηλότερες συγκεντρώσεις ρύπων. Αντίστοιχα, και ο μήνας παρουσιάζει παρόμοια επίδραση. Σε πολλές πόλεις, για παράδειγμα, η επαγγελματική δραστηριότητα κατά τη διάρκεια του καλοκαιριού είναι σημαντικά μειωμένη σε σχέση με το υπόλοιπο έτος, με αποτέλεσμα να αναμένεται και μείωση στα επίπεδα ατμοσφαιρικής ρύπανσης [2].

Πέρα από τα παραπάνω, υπάρχει σημαντικός αριθμός δημοσιευμένων ερευνών εντός της θεματικής που παρέχουν ιδιαίτερα χρήσιμες πληροφορίες σχετικά με τις μεθοδολογικές και αλγοριθμικές προσεγγίσεις που εμφανίζονται συχνά στην προβλεπτική της ποιότητας του αέρα, με μερικές από αυτές να παρουσιάζονται συνοπτικά παρακάτω.

Οι Kumar και Jain [79] χρησιμοποίησαν διάφορα μοντέλα ARIMA με σκοπό την πρόβλεψη ρυπογόνων χημικών ενώσεων όπως O_3, NO, NO_2, και CO. Χρησιμοποίησαν δεδομένα μετρήσεων εντός της πόλης Delhi της Ινδίας. Λόγω της φύσεως των δεδομένων, εφάρμοσαν διάφορους τύπους μετασχηματισμών (Πρώτες διαφορές, λογαριθμικούς μετασχηματισμούς, πρώτες διαφορές επί του λογαριθμικού μετασχηματισμού, μετασχηματισμούς τετραγωνικής ρίζας, πρώτες διαφορές επί των μετασχηματισμών τετραγωνικής ρίζας και single period return (SPR)). Για την αξιολόγηση των μεθόδων χρησιμοποιήθηκαν διαφορετικοί βραχυπρόθεσμοι ορίζοντες (10, 15, 20 βήματα) με προσέγγιση out of sample - one step ahead. Μέσω των αποτελεσμάτων τους γίνεται αντιληπτό ότι τα μοντέλα ARIMA μπορούν να παράγουν ικανοποιητικές προβλέψεις για τις συγκεντρώσεις των αέριων ρυπογόνων ενώσεων, ωστόσο ανάλογα τα χαρακτηριστικά της κάθε ένωσης, η προβλέψεις διαφέρουν ως προς την ακρίβεια τους [79].

Οι Liu et al. [80] παρουσίασαν συγκεντρωτικά τις μεθόδους που χρησιμοποιούνται για πρόβλεψη δεδομένων ποιότητας αέρα, με τις γενικές μεθόδους να διακρίνονται σε κλασσικά - στατιστικά μοντέλα, μοντέλα μηχανικής μάθησης και υβριδικά μοντέλα. Τα κλασσικά μοντέλα ανταποκρίνονται σε γραμμικά ή απλά δεδομένα, ενώ τα μοντέλα μηχανικής μάθησης ανταποκρίνονται καλύτερα σε μη – γραμμικά μοτίβα. Συνήθως, τα πραγματικά δεδομένα (real world data) παρουσιάζουν γραμμικά και μη – γραμμικά χαρακτηριστικά, επομένως τότε χρησιμοποιούνται τα υβριδικά μοντέλα. Τέλος, γίνεται αναφορά στην σημασία της προεπεξεργασίας, την κατασκευή των μοντέλων αλλά και η χρήση προσεγγίσεων παράλληλου υπολογισμού (parallel computing) για την πιο γρήγορη εκτέλεση των μεθόδων [80].

Οι Kaur et al. [81] επίσης παρουσιάζουν συγκεντρωτικά τις μεθόδους που χρησιμοποιούνται ενώ παράλληλα παρουσιάζουν το θεωρητικό υπόβαθρο για την μελέτη της ποιότητας του αέρα. Δίνεται μεγαλύτερη έμφαση στις μεθόδους μηχανικής μάθησης όπως δέντρα απόφασης, random forests και support vector machines, οι οποίες μπορούν ανταποκριθούν σε πολύπλοκα μοτίβα που παρουσιάζονται στα δεδομένα. Επίσης, γίνεται αναφορά στην σημασία της ποιότητας των δεδομένων, παρακολούθηση δεδομένων σε πραγματική χρόνο αλλά και στην μετάβαση προς συστημάτων που να μπορούν διαχειριστούν μεγάλο όγκο δεδομένων [81].

Οι Mirzadeh και Omranpour [66] δημιούργησαν ένα ενισχυμένο μοντέλο Random Forest με σκοπό την πρόβλεψη ρυπογόνων ενώσεων στο σύνολο δεδομένων Beijing PM2.5 και UCI Air Quality, που χρησιμοποιείται και στην παρούσα εργασία. Στην συνέχεια συγκρίνουν την απόδοση της μεθόδου σε σχέση με διάφορα άλλα μοντέλα (για παράδειγμα ARIMA, VARMA, SVR, RF, LSTM, GRU). Χρησιμοποιούν multi – step προσέγγιση, ωστόσο παρουσιάζουν τα σφάλματα για κάθε βήμα του

χρονικού ορίζοντα, δηλαδή από 1η ώρα μέχρι την 6η. Το προτεινόμενο μοντέλο, ακολουθούμενο από το κλασσικό μοντέλο RF, παρουσίαζαν τα χαμηλότερα σφάλματα MAE [66].

Οι Suradhaniwar et al. [26] εξετάζουν την απόδοση διαφορετικών αλγόριθμων (SARIMA, SVR, MLP, LSTM) σε μονοδιάστατα άγρο – μετεωρολογικά δεδομένα. Πέρα από τους διαφορετικούς αλγόριθμους, γίνεται σύγκριση μεταξύ one / single – step προσεγγίσεων (1 ώρα για 96 βήματα στον χρονικό ορίζοντα) και multi – step προσεγγίσεων (96 ώρες). Μέσω αυτής της σύγκρισης, αναδείχτηκε όσο μεγαλώνει ο χρονικός ορίζοντας πρόβλεψης παράλληλα αυξάνονται τα σφάλματα και ότι οι single – step προσεγγίσεις υπερτερούν ως προς την ακρίβεια των προβλέψεων σε σχέση με multi – step προσεγγίσεις. Αυτό οφείλεται στο γεγονός ότι οι multi – step προσεγγίσεις ενσωματώνουν τις προβλέψεις στα δεδομένα εκπαίδευσης, με αποτέλεσμα τα σφάλματα να συσσωρεύονται, μειώνοντας δραματικά την ακρίβεια για κάθε επόμενη πρόβλεψη [26].

Οι Athanasiadis et al. [27] εξέτασαν την χρήση αλγόριθμων ταξινόμησης (Naïve Bayes, Decision Trees, Nearest Neighbor και άλλα) και κλασσικών στατιστικών αλγόριθμων (ARIMA, LCA, PCA) με σκοπό την παραγωγή προβλέψεων από δεδομένα ποιότητας αέρα. Χρησιμοποίησαν δεδομένα από έναν μετεωρολογικό σταθμό εντός της Αθήνας (Μαρούσι). Παράλληλα, εφάρμοσαν διαφορετικούς χρονικούς ορίζοντες πρόβλεψης (8, 24, 48, 72 ώρες) ώστε να εξετάσουν την απόδοση των διαφορετικών αλγόριθμων και κατέληξαν ότι ο αλγόριθμος C.45, δηλαδή ένας αλγόριθμος decision tree, παρήγαγε τα καλύτερα αποτελέσματα [27].

Οι Kumar et al. [5] παρουσιάζουν ένα πλαίσιο βασισμένο στη μηχανική μάθηση για την πρόβλεψη και παρακολούθηση της ποιότητας του αέρα, συνδυάζοντας πολλαπλά μοντέλα παλινδρόμησης με μια προσέγγιση πρόβλεψης χρονοσειρών. Η μελέτη χρησιμοποιεί το σύνολο δεδομένων UCI Air Quality και εφαρμόζει διάφορα μοντέλα (Decision Tree, Random Forest, SVM, NN) για την πρόβλεψη των συγκεντρώσεων ρύπων, διαπιστώνοντας ότι τα μοντέλα Random Forest και Decision Tree επιτυγχάνουν τα χαμηλότερα σφάλματα πρόβλεψης, ανάλογα με τον ρύπο. Παράλληλα, συγκρίνουν τα προηγούμενα με ένα μοντέλο ARIMA για την πρόβλεψη μελλοντικών συγκεντρώσεων ρύπων, δείχνοντας ότι παρέχει αξιόπιστη απόδοση για βραχυπρόθεσμες προβλέψεις. Τα αποτελέσματα υπογραμμίζουν ότι, ενώ τα μοντέλα μηχανικής μάθησης είναι αποτελεσματικά στην αποτύπωση των σχέσεων μεταξύ των μεταβλητών, το ARIMA παραμένει χρήσιμο για την αποτύπωση των χρονικών προτύπων στις χρονοσειρές [5].

Οι Kaur και Sandha [28] προτείνουν ένα σύστημα βασισμένο στη μηχανική μάθηση για την πρόβλεψη της ποιότητας του αέρα, εκτιμώντας τόσο τις τιμές του Δείκτη Ποιότητας Αέρα (Air Quality Index – AQI) όσο και τις κατηγορίες ποιότητας αέρα (Good, Moderate, Unhealthy). Η μελέτη χρησιμοποιεί σύνολα δεδομένων UCI Air Quality καθώς και δεδομένα ατμοσφαιρικής ρύπανσης από το Δελχί, εφαρμόζοντας ένα μοντέλο Random Forest για παλινδρόμηση του AQI και ένα νευρωνικό δίκτυο βασισμένο στο TensorFlow για ταξινόμηση. Τα αποτελέσματα δείχνουν ότι το Random Forest παρέχει ισχυρή βασική απόδοση, ενώ το νευρωνικό δίκτυο βελτιώνει τη συνέπεια της ταξινόμησης μεταξύ διαφορετικών συνόλων δεδομένων, επιτυγχάνοντας ακρίβεια έως περίπου 76–78%. Η μελέτη αναδεικνύει βασικές προκλήσεις, όπως η ανισορροπία κλάσεων, τα προβλήματα ποιότητας δεδομένων και η γενίκευση των μοντέλων, και τις αντιμετωπίζει μέσω τεχνικών προεπεξεργασίας όπως η αντικατάσταση ελλιπών τιμών με την μέση τιμή μιας περιόδου και η μέθοδος SMOTE [28].

Οι Shat et al. [82] χρησιμοποιούν ένα μοντέλο FAR (Functional Auto-Regressive), το οποίο μπορεί να χαρακτηριστεί ως μια επέκταση του κλασσικού μοντέλου AR για τη βραχυπρόθεσμη πρόβλεψη ωριαίων συγκεντρώσεων όζοντος, και το συγκρίνουν με παραδοσιακά στατιστικά μοντέλα (ARIMA, VAR) και μεθόδους μηχανικής μάθησης (NNAR, Random Forest, SVM). Η μεθοδολογία βασίζεται στη διάσπαση της χρονοσειράς του όζοντος σε ντετερμινιστικά (εποχικά) και στοχαστικά συστατικά, τα

οποία μοντελοποιούνται ξεχωριστά πριν συνδυαστούν για την τελική πρόβλεψη. Χρησιμοποιώντας ωριαία δεδομένα όζοντος από το Λος Άντζελες (για την περίοδο 2013–2017), η μελέτη πραγματοποιεί προβλέψεις μιας ημέρας (one – step / day ahead) μπροστά και αξιολογεί την απόδοση των μεθόδων. Τα αποτελέσματα δείχνουν ότι το μοντέλο FAR υπερέχει σταθερά όλων των άλλων μεθόδων, ακολουθούμενο από τα VAR και ARIMA, ενώ τα μοντέλα μηχανικής μάθησης, όπως το Random Forest και το NNAR, παρουσιάζουν μέτρια απόδοση και το SVM τη χαμηλότερη [82].

Ο Kamińska [83] εξετάζει την χρήση του αλγόριθμου Random Forest για την βραχυπρόθεσμη πρόβλεψη αέριων ρυπογόνων ενώσεων (NO_2, NOx, PM2.5) με την χρήση επεξηγηματικών δεδομένων (μετεωρολογικά, μετακίνησης και χρονικά). Τα δεδομένα είναι σε ωριαία χρονική κλίμακα και αφορούν την πόλη Wroclaw της Πολωνίας. Ένα βασικό εύρημα της μελέτης είναι ότι διαφορετικοί ρύποι επηρεάζονται από διαφορετικούς παράγοντες. Για τα οξείδια του αζώτου (NO_2 και NOx), η κυκλοφορία οχημάτων αποτελεί τον σημαντικότερο παράγοντα πρόβλεψης, ενώ για τα PM2.5 μεγαλύτερη σημασία έχουν οι μετεωρολογικές συνθήκες, όπως η θερμοκρασία, η ταχύτητα και η διεύθυνση του ανέμου. Ένα επίσης σημαντικό εύρημα είναι ότι η εποχικότητα επηρεάζει σημαντικά την απόδοση των μοντέλων, με καλύτερα αποτελέσματα όταν τα μοντέλα εκπαιδεύονται σε συγκεκριμένα εποχικά υποσύνολα δεδομένων (θερινή περίοδος – χειμερινή περίοδος) αντί για ολόκληρο το σύνολο δεδομένων. Αυτό υποδηλώνει ότι οι σχέσεις μεταξύ των μεταβλητών μεταβάλλονται με την πάροδο του χρόνου και ότι η λήψη υπόψη της εποχικότητας μπορεί να βελτιώσει την ακρίβεια των προβλέψεων. Συνολικά, η μελέτη καταλήγει ότι το Random Forest αποτελεί ένα χρήσιμο εργαλείο για την πρόβλεψη της ποιότητας του αέρα, ιδιαίτερα όταν συνδυάζεται με κατάλληλες εξωγενείς μεταβλητές, αλλά η απόδοσή του εξαρτάται από τον τύπο του ρύπου, τα εποχικά πρότυπα και τα χαρακτηριστικά των δεδομένων [83].

Οι Kane et al. [65] συγκρίνουν την απόδοση μοντέλων ARIMA και Random Forest για την πρόβλεψη γρίπης των πτηνών (H5N1) στην Αίγυπτο, με δεδομένα για την περίοδο 2005 - 2011. Το μοντέλο ARIMA επιλέγεται μέσω αυτοματοποιημένης διαδικασίας βάσει κριτηρίων επιλογής μοντέλου και αποτυπώνει γραμμικές χρονικές εξαρτήσεις στα δεδομένα. Αντίθετα, το μοντέλο Random Forest χρησιμοποιεί πολλαπλά lags των μεταβλητών (προηγούμενα κρούσματα, θερμοκρασία, υγρασία) και δημιουργεί ένα σύνολο δέντρων απόφασης. Επίσης, παρατηρούν ότι το Random Forest είναι πιο ικανό να αντιληφθεί ακραίες αλλαγές (δηλαδή ένα σοβαρό φαινόμενο – outbreak) σε σχέση με το μοντέλο ARIMA. Παράλληλα το μοντέλο RF παρουσιάζει σημαντικά χαμηλότερα σφάλματα σε σχέση με το μοντέλο ARIMA. Συνολικά, οι συγγραφείς καταλήγουν ότι το Random Forest αποτελεί πιο αποτελεσματική προσέγγιση για την πρόβλεψη επιδημιών σε αυτό το πλαίσιο, ιδιαίτερα όταν τα δεδομένα είναι σύνθετα και μη γραμμικά. Παρ' όλα αυτά, επισημαίνεται ότι και τα δύο μοντέλα παρουσιάζουν περιορισμούς και ότι απαιτείται περαιτέρω έρευνα για τη βελτίωση της απόδοσης, ειδικά σε περιπτώσεις ακραίων τιμών και υψηλής δυναμικής των χρονοσειρών [65].

Οι Freeman et al. [60] παρουσιάζουν μια από τις πρώτες εφαρμογές LSTM – RNN σε δεδομένα ποιότητας αέρα, συγκεκριμένα παράγοντας προβλέψεις για την μέση τιμή O_3 ανά 8 ώρες. Η μελέτη χρησιμοποιεί ωριαία δεδομένα ποιότητας αέρα και μετεωρολογικών μεταβλητών από το Κουβέιτ και αναπτύσσει ένα μοντέλο πρόβλεψης που μπορεί να προβλέπει έως και 72 ώρες μπροστά. Ένα βασικό στοιχείο της μεθοδολογίας είναι η προσεκτική προεπεξεργασία των δεδομένων, συμπεριλαμβανομένης της συμπλήρωσης ελλείπων τιμών και της επιλογής χαρακτηριστικών μέσω δέντρων απόφασης, η οποία βελτίωσε την απόδοση του μοντέλου. Σχετικά με την αντιμετώπιση των ελλιπών τιμών, σε κενά διαστήματα άνω των 8 ωρών, χρησιμοποιείται το διάστημα ± μιας ημέρας, για την κάθε ώρα που είναι ελλιπής, διαιρώντας το αποτέλεσμα δια 2, ώστε να παραχθεί η μέση τιμή του επιλεγμένου παραθύρου. Το μοντέλο LSTM καταφέρνει να αποτυπώσει αποτελεσματικά χρονικές εξαρτήσεις και μη γραμμικές σχέσεις, υπερέχοντας σε σχέση με παραδοσιακές μεθόδους όπως

ARIMA και Feed Forward Neural Networks (FFNN). Το μοντέλο LSTM επιτυγχάνει χαμηλά σφάλματα πρόβλεψης (MAE < 2 για μεγαλύτερους ορίζοντες) και μπορεί επίσης να προβλέψει τη διάρκεια υπερβάσεων ρύπανσης, κάτι ιδιαίτερα σημαντικό για τη διαχείριση της ποιότητας του αέρα [60].

Οι Liu et al. [4] παρουσιάζουν και συγκρίνουν μοντέλα μηχανικής (SVR – Random Forest) για την πρόβλεψη του Δείκτη Ποιότητας Αέρα (AQI) και της συγκέντρωσης του NOx, χρησιμοποιώντας Support Vector Regression (SVR) και Random Forest Regression (RFR). Η μελέτη βασίζεται σε δύο σύνολα δεδομένων: δεδομένα ποιότητας αέρα από το Πεκίνο (Beijing Air Quality Dataset) και το σύνολο δεδομένων UCI Air Quality. Συγκεκριμένα, το SVR παρουσιάζει καλύτερη απόδοση στην πρόβλεψη του AQI, ενώ το Random Forest αποδίδει καλύτερα στην πρόβλεψη των συγκεντρώσεων NOx, γεγονός που υποδηλώνει ότι η καταλληλόλητα του μοντέλου εξαρτάται από τη μεταβλητή-στόχο [4].

2 Σύνολο Δεδομένων

Το σετ δεδομένων ανακτήθηκε από το UCI Repository [48] και αφορά τα δεδομένα από την δημοσίευση των De Vito et al. [3]. Το σετ δεδομένων αποτελείται από 9358 μετρήσεις μέσων τιμών ανά ώρα, οι οποίες παράχθηκαν από ένα σύνολο 5 χημικών αισθητήρων μεταλλικού οξειδίου (metal oxide chemical sensors) ενσωματωμένων σε μια συσκευή πολλαπλών αισθητήρων χημικών ουσιών για την εκτίμηση της ποιότητας του αέρα. Η συσκευή τοποθετήθηκε σε μια περιοχή με σημαντικά επίπεδα ρύπανσης σε επίπεδο δρόμου, σε μια ιταλική πόλη. Τα δεδομένα καταγράφηκαν από τον Μάρτιο του 2004 έως τον Φεβρουάριο του 2005 (σχεδόν ένα έτος) και αντιπροσωπεύουν τις μεγαλύτερου μεγέθους ελεύθερα διαθέσιμες καταγραφές μετρήσεων συσκευών χημικών αισθητήρων για την ανίχνευση και μέτρησης της ποιότητας αέρα που έχουν τοποθετηθεί στο πεδίο. Οι ωριαίες μέσες συγκεντρώσεις μονοξείδιου του άνθρακα (CO), μη-μεθανικών υδρογονανθράκων(NMHC), βενζολίου (C6H6), οξειδίων του αζώτου (NOx) και διοξειδίου του αζώτου (NO2) παρέχονται παράλληλα από έναν πιστοποιημένο αναλυτή αναφοράς που βρισκόταν στον ίδιο χώρο [48]. Οι μετρήσεις που προέρχονται από τον αναλυτή αναφοράς χαρακτηρίζονται ως Ground Truth (GT), δηλαδή για παράδειγμα η μεταβλητή του μονοξείδιου του άνθρακα εμφανίζεται ως “CO (GT)” ενώ οι μετρήσεις από την πειραματική συσκευή έχουν μορφή “ PT08.Sx ”, όπου για κάθε x λαμβάνει τιμές από 1 έως 5, με την καθεμιά να αναφέρεται σε έναν από τους 5 αισθητήρες της πειραματικής συσκευής [3,48].

Στον πίνακα 1, παρουσιάζονται αναλυτικά τα γνωρίσματα του σύνολού δεδομένων, μια σύντομη μετάφραση – επεξήγηση και η μονάδα μέτρησης του κάθε γνωρίσματος.

Πίνακας 1 Γνωρίσματα και επεξήγηση του συνόλου δεδομένων

Γνώρισμα	**Επεξήγηση - Μετάφραση**	**Μονάδα Μέτρησης**
Date	Ημερομηνία	Ημέρα/Μήνας/ Έτος
Time	Ώρα	Ώρα.Λεπτά.Δευτερόλεπτα
CO(GT)	Πραγματική μέση συγκέντρωση ανά ώρα για CO	mg/m^3
PT08.S1(CO)	Μέση τιμή ανά ώρα εξόδου χημικού αισθητήρα PT08.S1 - CO	επίπεδα τάσης/αντίστασης
NMHC(GT)	Πραγματική μέση συγκέντρωση ανά ώρα για NMHC	$μg/ m^3$

C6H6(GT)	Πραγματική μέση συγκέντρωση ανά ώρα για C6H6	μg/ m^3
PT08.S2(NMHC)	Μέση τιμή ανά ώρα εξόδου χημικού αισθητήρα PT08.S2 - NMHC	επίπεδα τάσης/αντίστασης
NOx(GT)	Πραγματική μέση συγκέντρωση ανά ώρα για NOx	Ppb – parts per billion
PT08.S3(NOx)	Μέση τιμή ανά ώρα εξόδου χημικού αισθητήρα PT08.S3 - NOx	επίπεδα τάσης/αντίστασης
NO2(GT)	Πραγματική μέση συγκέντρωση ανά ώρα για NO2	μg/ m^3
PT08.S4(NO2)	Μέση τιμή ανά ώρα εξόδου χημικού αισθητήρα PT08.S4 - NO2	επίπεδα τάσης/αντίστασης
PT08.S5(O3)	Μέση τιμή ανά ώρα εξόδου χημικού αισθητήρα PT08.S5 - O3	επίπεδα τάσης/αντίστασης
T	Θερμοκρασία	°C
RH	Σχετική υγρασία	%
AH	Απόλυτη υγρασία	g/ m^3

3 Μεθοδολογία

3.1 Αρχική προ-επεξεργασία δεδομένων

Αρχικά, τα δεδομένα μεταφορτώθηκαν σε έναν φάκελο στο Google Drive, καθώς ο κώδικας θα αναπτυσσόταν σε περιβάλλον Google Collab, με την χρήση ενός τυποποιημένου Virtual Machine. Είναι σημαντικό, για την ορθή φόρτωση των δεδομένων, μέσω της συνάρτησης “read_csv()” της βιβλιοθήκης pandas, να οριστούν ως παράμετροι ο χαρακτήρας διαχωρισμού στηλών (column separator), που στην προκειμένη περίπτωση ήταν το semi-colon (;) και παράλληλα να οριστεί ο χαρακτήρας των δεκαδικών, όπου στην προκειμένη περίπτωση ήταν το κόμμα (,). Το αποτέλεσμα της συνάρτησης “read_csv()”, δηλαδή το σετ δεδομένων ορίστηκε ως “df”. Πιθανώς λόγω κακού formatting, δημιουργούνται 2 κενά γνωρίσματα – στήλες, χωρίς ονομασία. Λόγω αυτού τους ανατίθεται αυτόματα η ονομασία “unnamed:16”και “unnamed:17”. Προφανώς, αφού δεν περιέχουν κάποια πληροφορία αφαιρούνται από το σετ δεδομένων.

Ακολουθεί η αρχική επισκόπηση των δεδομένων, χρησιμοποιώντας τα πέντε πρώτα στιγμιότυπα του σετ δεδομένων, όπως φαίνεται στον πίνακα 2.

Πίνακας 2 Επισκόπηση αρχικών 5 στιγμιότυπων του συνόλου δεδομένων

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
0	10/03/2004	18.00.00	2.6	1360.0	150.0	11.9	1046.0	166.0	1056.0	113.0	1692.0	1268.0	13.6	48.9	0.7578
1	10/03/2004	19.00.00	2.0	1292.0	112.0	9.4	955.0	103.0	1174.0	92.0	1559.0	972.0	13.3	47.7	0.7255
2	10/03/2004	20.00.00	2.2	1402.0	88.0	9.0	939.0	131.0	1140.0	114.0	1555.0	1074.0	11.9	54.0	0.7502
3	10/03/2004	21.00.00	2.2	1376.0	80.0	9.2	948.0	172.0	1092.0	122.0	1584.0	1203.0	11.0	60.0	0.7867
4	10/03/2004	22.00.00	1.6	1272.0	51.0	6.5	836.0	131.0	1205.0	116.0	1490.0	1110.0	11.2	59.6	0.7888

Όπως προαναφέρθηκε το σετ δεδομένων παρουσιάζει τα παραπάνω γνωρίσματα ενώ παράλληλα υπενθυμίζεται ότι τα γνωρίσματα που περιέχουν (GT) αναφέρονται σε Ground Truth, δηλαδή αξιόπιστες μετρήσεις από εξειδικευμένες συσκευές. Τα γνωρίσματα που ξεκινούν με το αλφαριθμητικό “PT0.8”, αναφέρονται σε μετρήσεις από την πειραματική συσκευή μετρήσεων που παρουσιάστηκε από τους De Vito et al. [3]. Πέρα από τα προαναφερόμενα γνωρίσματα που αναφέρονται στις ρυπογόνες ενώσεις ανάλογα την συσκευή μέτρησης, υπάρχει το γνώρισμα “Date” που αφορά την ημερομηνία λήψης του δείγματος σε μορφής «ημέρα/μήνας/έτος», το γνώρισμα “Time” που αφορά την χρονική στιγμή της λήψης του δείγματος με μορφή «Ώρα:Λεπτά:Δευτερόλεπτα», το γνώρισμα “T” που αφορά την θερμοκρασία την στιγμή της λήψης δείγματος σε βαθμούς Κελσίου (°C), το γνώρισμα “RH” που αφορά την σχετική υγρασία και τέλος το γνώρισμα “AH” που αφορά την απόλυτη υγρασία.

Ύστερα από την αρχική επισκόπηση, ελέγχονται οι διαστάσεις του σετ δεδομένων, μέσω του χαρακτηριστικού (attribute) “shape”.Το σετ δεδομένων παρουσιάζει 9471 δείγματα και 15 γνωρίσματα, ωστόσο αυτό διαφέρει από τον πραγματικό αριθμό δειγμάτων που παρέχεται από την πηγή δεδομένων που ορίζει ότι στο σετ δεδομένων υπάρχουν 9358 δείγματα [48]. Η αιτία του παραπάνω θα εξεταστεί στην συνέχεια.

Έπειτα γίνεται έλεγχος για τον τύπο δεδομένων κάθε γνωρίσματος μέσω του attribute “dtypes” που εφαρμόζεται στο σετ δεδομένων, όπως φαίνεται στον πίνακα 3.

Πίνακας 3 Τύπος δεδομένων ανά γνώρισμα

```
Date             object
Time             object
CO(GT)          float64
PT08.S1(CO)     float64
NMHC(GT)        float64
C6H6(GT)        float64
PT08.S2(NMHC)   float64
NOx(GT)         float64
PT08.S3(NOx)    float64
NO2(GT)         float64
PT08.S4(NO2)    float64
PT08.S5(O3)     float64
T               float64
RH              float64
AH              float64
dtype: object
```

Όλα τα γνωρίσματα, βρίσκονται στην κατάλληλη μορφή float64, εκτός των Date και Time που βρίσκονται σε μορφή object το οποίο είναι λογικό. Στην συνέχεια, έγινε έλεγχος για ελλιπείς τιμές

ανά γνώρισμα, όπου παρατηρήθηκε ότι υπήρχαν 114 ελλιπείς τιμές σε όλα τα γνωρίσματα, όπως φαίνεται στον πίνακα 4.

Πίνακας 4 Καταμέτρηση ελλιπών τιμών ανά γνώρισμα

```
Date             114
Time             114
CO(GT)           114
PT08.S1(CO)      114
NMHC(GT)         114
C6H6(GT)         114
PT08.S2(NMHC)    114
NOx(GT)          114
PT08.S3(NOx)     114
NO2(GT)          114
PT08.S4(NO2)     114
PT08.S5(O3)      114
T                114
RH               114
AH               114
dtype: int64
```

Εφόσον οι ελλιπείς τιμές παρατηρούνται σε όλα τα γνωρίσματα και για τους ίδιους δείκτες (indexes), προφανώς πρόκειται για κάποιο επιπλέον σφάλμα στην κωδικοποίηση του σετ δεδομένων και μπορούν να αφαιρεθούν. Ο πίνακας 5, δείχνει τους δείκτες - Indexes που παρουσιάζουν ελλιπείς τιμές για το γνώρισμα CO(GT), οι οποίοι ήταν οι εξής:

Πίνακας 5 Παρουσίαση και καταμέτρηση δεικτών με ελλιπής τιμές

```
Index([9357, 9358, 9359, 9360, 9361, 9362, 9363, 9364, 9365, 9366,
       ...
       9461, 9462, 9463, 9464, 9465, 9466, 9467, 9468, 9469, 9470],
      dtype='int64', length=114)
```

Δηλαδή, πράγματι υπάρχουν 114 συνεχόμενες ελλιπείς τιμές για το γνώρισμα CO(GT), από το index 9357 έως το 9470. Για να υπάρξει επιβεβαίωση ότι αυτό ισχύει και στα υπόλοιπα γνωρίσματα, θα γίνει έλεγχος για τα τελευταία 115 δείγματα του σετ δεδομένων μέσω της συνάρτησης “tail() , όπως φαίνεται να ισχύει στον πίνακα 6.

Πίνακας 6 Συνολική παρουσίαση και καταμέτρηση δεικτών με ελλιπής τιμές

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
9356	04/04/2005	14.00.00	2.2	1071.0	-200.0	11.9	1047.0	265.0	654.0	168.0	1129.0	816.0	28.5	13.1	0.5028
9357	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9358	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9359	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9360	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9466	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9467	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9468	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9469	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9470	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

115 rows × 15 columns

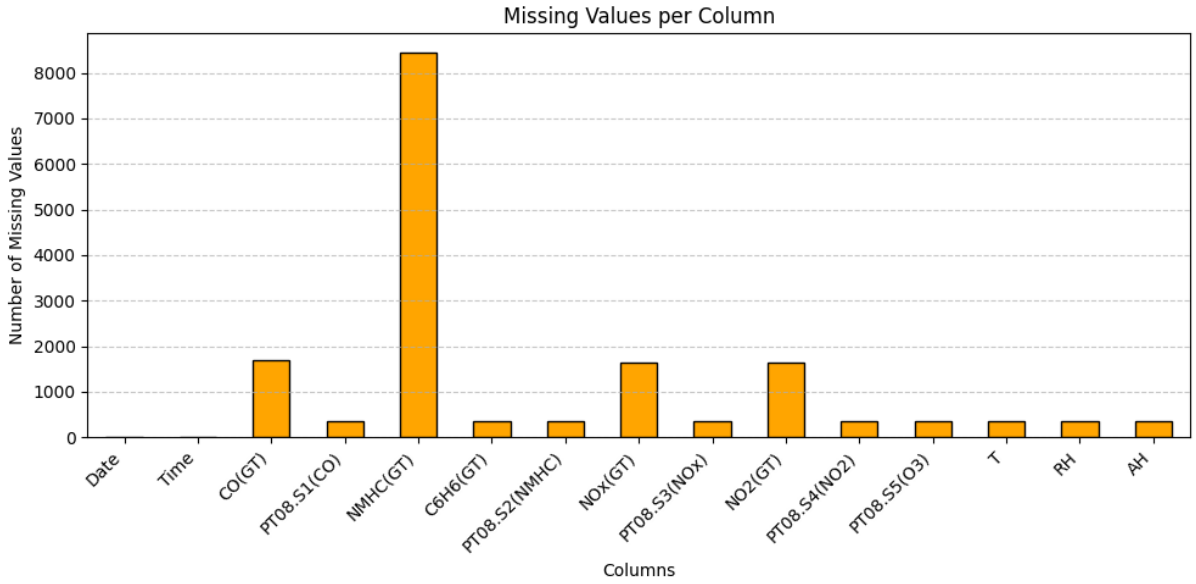
Άρα πράγματι πρόκειται για κάποιο σφάλμα κωδικοποίησης και οι τιμές μπορούν να αφαιρεθούν. Με την αφαίρεση των παραπάνω, το σετ δεδομένων έχει πλέον τον σωστό αριθμό δειγμάτων.

Οι συντάκτες του σετ δεδομένων αναφέρουν ότι έχουν κωδικοποιήσει τις πραγματικές ελλιπείς τιμές με την τιμή "-200" [3]. Επομένως, μέσω της εντολής "pandas.replace()", μπορεί να γίνει η αντικατάσταση των τιμών "-200" με NaN's. Έπειτα επαναλαμβάνεται η ανίχνευση ελλιπών τιμών ανά γνώρισμα, με τα αποτελέσματα να απεικονίζονται στον πίνακα 7.

Πίνακας 7 Καταμέτρηση ελλιπών τιμών ανά γνώρισμα ύστερα από επεξεργασία

```
Date              0
Time              0
CO(GT)         1683
PT08.S1(CO)     366
NMHC(GT)       8443
C6H6(GT)        366
PT08.S2(NMHC)   366
NOx(GT)        1639
PT08.S3(NOx)    366
NO2(GT)        1642
PT08.S4(NO2)    366
PT08.S5(O3)     366
T               366
RH              366
AH              366
dtype: int64
```

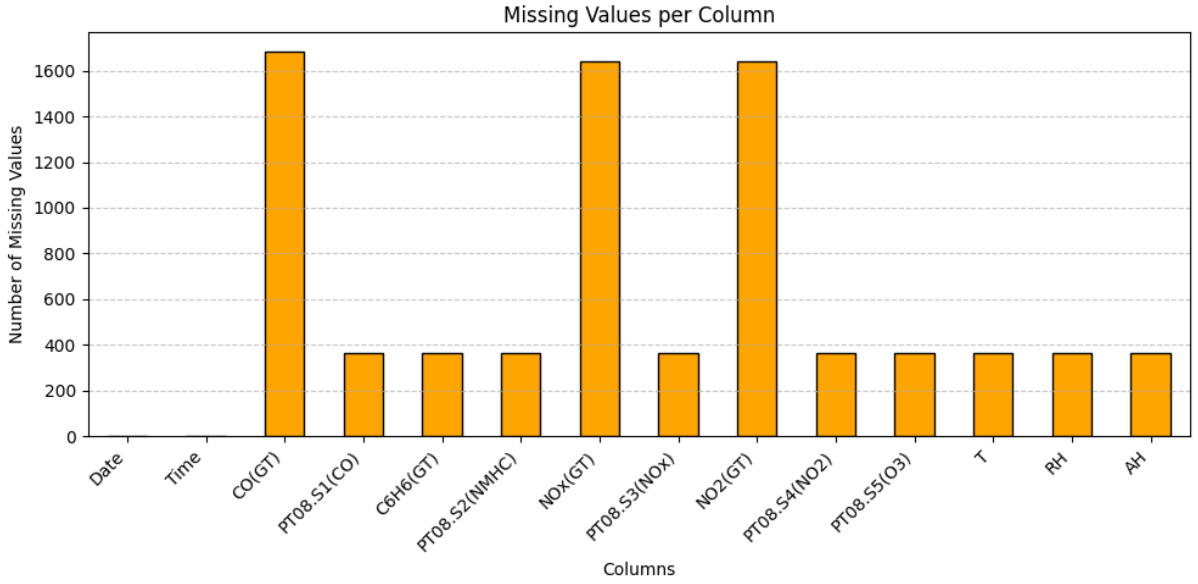
Μια καλή πρακτική είναι η οπτικοποίηση τέτοιων μεγεθών ώστε να γίνουν πιο εύκολα αντιληπτές οι διαφορές μεγεθών ανά γνώρισμα που μπορεί να υπάρχουν, με το διάγραμμα 5 να παρουσιάζει την εν λόγω οπτικοποίηση.



Διάγραμμα 5 Απεικόνιση ελλιπών τιμών ανά γνώρισμα

Το διάγραμμα 5 απεικονίζει τις ελλιπείς τιμές ανά γνώρισμα σε μορφή ραβδογράμματος (bar plot). Το γνώρισμα "NMHC(GT)", παρουσιάζει με σημαντική διαφορά τις περισσότερες ελλιπής τιμές, πράγμα διόλου παράξενο αφού οι Vito et al., αναφέρουν ότι ο συγκεκριμένος μετρητής

χρησιμοποιήθηκε μόνο για τις πρώτες 8 ημέρες και έπειτα έπαψε να χρησιμοποιείται [3]. Όπως αναφέρουν και οι Kauer et al., η ύπαρξη ελλιπών τιμών είναι ένα συχνό φαινόμενο σε δεδομένα μετρήσεων ποιότητας αέρα, που μπορεί να οφείλονται σε σφάλματα των συσκευών μετρήσεων όπως προβλήματα με τις μπαταρίες τους ή ακόμα και αδυναμίας επικοινωνίας με το διαδίκτυο για την μεταφορά των δεδομένων [81].



Διάγραμμα 6 Απεικόνιση ελλιπών τιμών ανά γνώρισμα ύστερα από αφαίρεση του NMHC(GT)

Βάσει αυτού, το γνώρισμα αφαιρέθηκε από το σετ δεδομένων καθώς δεν παρέχει αρκετές μετρήσεις ώστε να χρησιμοποιηθεί. Ακολουθούν τα γνωρίσματα “CO(GT)” με 1683 τιμές, “NO2” με 1642 τιμές, Nox με 1639 ενώ όλα τα λοιπά γνωρίσματα παρουσιάζουν 366, όπως φαίνεται στο διάγραμμα 6.

Όπως αναφέρουν οι Hua et al., η παρουσία ελλιπών τιμών είναι ένα πολύ συχνό φαινόμενο στα δεδομένα ποιότητας αέρα σε μορφή χρονοσειρών και οφείλονται σε διάφορους λόγους με τους πιο συνήθεις να είναι οι βλάβες στους αισθητήρες ή ακόμα και η προσωρινή παύση των συσκευών μέτρησης λόγω σφάλματος τροφοδοσίας ηλεκτρικής ενέργειας [6].

Καθώς τα περισσότερα γνωρίσματα παρουσιάζουν σημαντικό αριθμό ελλιπών τιμών, η αφαίρεση τους δεν είναι ορθή πρακτική αφού θα αναιρούταν σημαντικό μέρος χρήσιμων πληροφοριών από το σετ δεδομένων. Βάσει αυτού είναι αναγκαία η εύρεση μιας μεθοδολογίας που να μπορεί να διαχειριστεί το υψηλό πλήθος ελλιπών τιμών.

Αφού δεδομένα έχουν μορφή χρονοσειράς, η εν λόγω μεθοδολογία χρίζει ιδιαίτερης προσοχής καθώς είναι σημαντικό να μην αλλοιωθεί σημαντικά το «σήμα» της χρονοσειράς, και κατ’ επέκταση τα χαρακτηριστικά της δηλαδή η εποχικότητα, η τάση, η κυκλικότητα και τα κατάλοιπα της. Για παράδειγμα, οι Kumar, Chaudhri και Rajput, αναφέρουν ότι αντικατέστησαν τις ελλιπείς τιμές του σετ δεδομένων με την χρήση μέσων και ενδιάμεσων τιμών [5]. Παρόλο που αυτή η μέθοδος είναι ευρέως χρησιμοποιούμενη και εύκολη στην υλοποίηση της, με την μέση τιμή να χρησιμοποιείται για συνεχόμενα αριθμητικά δεδομένα ενώ η ενδιάμεση τιμή για κατηγορικές ή διατακτικές μεταβλητές, η χρήση της σε χρονοσειρές παρουσιάζει βασικά προβλήματα όπως υπό – εκτίμηση της διακύμανσης, παράληψη των συσχετίσεων μεταξύ των μεταβλητών με αποτέλεσμα μεροληπτικές (biased) εκτιμήσεις των συσχετίσεων και της συνδιακύνασης [7,6]. Οι Hua et al απέδειξαν μέσω πειραμάτων σε διαφορετικά σετ δεδομένων ότι η χρήση μέσης – ενδιάμεσης τιμής έχει λόγο να χρησιμοποιηθεί σε μικρά «κενά» δεδομένων, τα οποία χαρακτηρίζονται ως MCAR (Missing Completely At Random),

δηλαδή που δεν παρουσιάζουν κάποιο επαναληπτικό μοτίβο ή δομή ωστόσο παρήγαγε τα υψηλότερα σφάλματα σε σχέση με άλλες μεθόδους imputation όπως kNNI ή MICE [6].

Πριν από οποιαδήποτε εφαρμογή μεθόδου, είναι αναγκαία η κατασκευή του γνωρίσματος “Datetime”, που θα λειτουργήσει ως Index της χρονοσειράς. Για την κατασκευή του θα χρησιμοποιηθούν τα γνωρίσματα “Date” και “Time”, αφού λάβουν την κατάλληλη μορφή. Πιο συγκεκριμένα, η μορφή που οφείλει να έχει το γνώρισμα ορίζεται από την βιβλιοθήκη NumPy με τύπο “datetime64”, με την δομή να έχει ως εξής: «Έτος – Μήνας – Ημερομηνία Ώρα: Λεπτά: Δευτερόλεπτα» ή με την χρήση ενός παραδείγματος: «2004-03-10 18:00:00» [8].

Για να επιτευχθεί η παραπάνω μορφή, το γνώρισμα “Date”, θα μετατραπεί προσωρινά σε μορφή αλφαριθμητικού (string) και το σύμβολο “/” θα αντικατασταθεί με “-“ με την χρήση της μεθόδου replace(). Στο γνώρισμα “Time”, θα γίνει αντικατάσταση του συμβόλου “.” με το σύμβολο “:”. Ορίζεται ένα καινούργιο γνώρισμα ως “Datetime”, το οποίο εφόσον και τα 2 γνωρίσματα έχουν δομή αλφαριθμητικού μπορούν να “προστεθούν”, και το συνδυαστικό αλφαριθμητικό να χρησιμοποιηθεί στην συνάρτηση «to_datetime» από την βιβλιοθήκη pandas [9], με σκοπό την μετατροπή των δεδομένων σε ένα αντικείμενο datetime, δηλαδή με δομή datetime64 όπως προαναφέρθηκε. Ως παράμετρος εντός της συνάρτησης χρειάζεται να οριστεί το format, δηλαδή με ποια σειρά θα γίνει parse το αλφαριθμητικό που στην συγκεκριμένη περίπτωση είχε μορφή: '%d-%m-%Y %H:%M:%S' βάσει των γνωρισμάτων. Μετά την προαναφερόμενη διαδικασία, δημιουργήθηκαν μερικά καινούργια γνωρίσματα όπως το “Datetime_rep”, το οποίο είναι αντίγραφο του προαναφερόμενου γνωρίσματος “Datetime” και μέσω της βιβλιοθήκης Pandas και πιο συγκεκριμένα μέσω του accessor “dt” [10] που εφαρμόζεται σε ολόκληρα series και ξεχωριστά σε κάθε αντικείμενο datetime, δημιουργήθηκε το γνώρισμα “Day” μέσω της μεθόδου “day_name()”, το γνώρισμα “Month” μέσω της μεθόδου “month_name()” και το γνώρισμα “Hour” μέσω του γνωρίσματος “hour”. Τέλος, ορίζεται ως index το γνώρισμα “Datetime”, με τα αποτελέσματα των παραπάνω να φαίνεται στον πίνακα 8. Όλα τα παραπάνω γίνονται για περεταίρω ανάλυση των δεδομένων, όπως για παράδειγμα την εύρεση της ημέρας της εβδομάδος ή της ώρας που παρουσιάζει τα μεγαλύτερα μεγέθη συγκέντρωσης ρυπογόνων ενώσεων.

Πίνακας 8 Παρουσίαση του γνωρίσματος Datetime

Datetime	Date	Time	CO(GT)	PT08.S1(CO)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH	Datetime_rep	Day	Month	Hour
2004-03-10 18:00:00	10-03-2004	18:00:00	2.6	1360.0	11.9	1046.0	166.0	1056.0	113.0	1692.0	1268.0	13.6	48.9	0.7578	2004-03-10 18:00:00	Wednesday	March	18
2004-03-10 19:00:00	10-03-2004	19:00:00	2.0	1292.0	9.4	955.0	103.0	1174.0	92.0	1559.0	972.0	13.3	47.7	0.7255	2004-03-10 19:00:00	Wednesday	March	19
2004-03-10 20:00:00	10-03-2004	20:00:00	2.2	1402.0	9.0	939.0	131.0	1140.0	114.0	1555.0	1074.0	11.9	54.0	0.7502	2004-03-10 20:00:00	Wednesday	March	20
2004-03-10 21:00:00	10-03-2004	21:00:00	2.2	1376.0	9.2	948.0	172.0	1092.0	122.0	1584.0	1203.0	11.0	60.0	0.7867	2004-03-10 21:00:00	Wednesday	March	21
2004-03-10 22:00:00	10-03-2004	22:00:00	1.6	1272.0	6.5	836.0	131.0	1205.0	116.0	1490.0	1110.0	11.2	59.6	0.7888	2004-03-10 22:00:00	Wednesday	March	22

Ύστερα γίνεται έλεγχος για τον τύπο δεδομένων του κάθε γνωρίσματος όπως φαίνεται στον πίνακα 9, όπου παρατηρείται ότι πράγματι είναι στην κατάλληλη μορφή.

Πίνακας 9 Έλεγχος τύπου δεδομένων ανά γνώρισμα

```
Date                    object
Time                    object
CO(GT)                 float64
PT08.S1(CO)            float64
C6H6(GT)               float64
PT08.S2(NMHC)          float64
NOx(GT)                float64
PT08.S3(NOx)           float64
NO2(GT)                float64
PT08.S4(NO2)           float64
PT08.S5(O3)            float64
T                      float64
RH                     float64
AH                     float64
Datetime_rep    datetime64[ns]
Day                     object
Month                   object
Hour                     int32
dtype: object
['Date', 'Time', 'CO(GT)', 'PT08.S1(CO)', 'C6H6(GT)', 'PT08.S2(NMHC)', 'NOx(GT)', 'PT08.S3(NOx)', 'NO2(GT)', 'PT08.S4(NO2)', 'PT08.S5(O3)', 'T', 'RH', 'AH', 'Datetime_rep', 'Day', 'Month', 'Hour']
```

Επιπλέον, δημιουργείται ένα αντίγραφο του dataframe ως “df_original”, με σκοπό την μετέπειτα χρήση για σύγκριση γραφημάτων. Πλέον, αφού η χρονοσειρά είναι πλέον σε κατάλληλη μορφή, μπορεί να γίνει εντοπισμός της διάρκειας των κενών στα γνωρίσματα υπό μελέτη, CO(GT) και NO2, και μετέπειτα να επιλεχτεί η κατάλληλη μέθοδος για την συμπλήρωση των ελλιπών τιμών. Για την κατάλληλη επιλογή μεθοδολογίας αντιμετώπισης των ελλιπών τιμών, μια καλή πρακτική είναι η μελέτη των διαστημάτων με τις ελλείπεις τιμές [11, 12, 13, 14]. Τέτοιου είδους προβλήματα που αφορούν την ομαδοποίηση συνεχών διαστημάτων (είτε αφορούν ελλιπής είτε μη – ελλιπής τιμές) και την διαφοροποίηση τους από άλλα συνεχόμενα διαστήματα, στην βιβλιογραφία αναφέρονται ως “Islands and Gaps Problems” [11]. Για αυτό τον σκοπό ορίστηκε μια μεταβλητή “mask”, επιλέχθηκε το γνώρισμα υπό μελέτη (CO)GT και με την μέθοδο “isna()”, δημιουργείται ένα boolean series ή αλλιώς ένα μονοδιάστατο (1 X N) διάνυσμα όπου N ο αριθμός των δειγμάτων, όπου εάν ένα δείγμα παρουσιάζει ελλιπής τιμή χαρακτηρίζεται ως True και διαφορετικά ως False, για παράδειγμα έστω το παρακάτω διάνυσμα όπου η πρώτη σειρά αναφέρεται στα indexes και η δεύτερη σειρά στα Bools:

$$mask = \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ F & T & T & F & T & T \end{matrix}$$

Δηλαδή στο παράδειγμα υπάρχουν 2 διακριτά «κενά», το πρώτο κενό στους δείκτες 1,2 και το δεύτερο κενό στους δείκτες 4,5.

Έπειτα δημιουργείται μια άλλη μεταβλητή “gap_groups”, όπου εντός ορίζεται ένας Boolean έλεγχος μεταξύ της προαναφερόμενης μεταβλητής “mask” και της “mask” με την εφαρμογή της μεθόδου “shift()”, με την οποία το διάνυσμα μετακινείται κατά +1 index. Δηλαδή, η «νέα» μεταβλητή mask, βάσει του προαναφερόμενου παραδείγματος θα γινόταν:

$$mask.shift() = \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ NaN & F & T & T & F & T \end{matrix}$$

Επομένως, με την χρήση του not equal operator (!=), συγκρίνονται οι παραπάνω μεταβλητές και όπου στο καινούργιο διάνυσμα εάν οι Boolean τιμές είναι ίδιες (T/T ή F/F), τότε αντιστοιχεί False ενώ σε κάθε άλλη περίπτωση, δηλαδή αναντιστοιχίας, αντιστοιχεί True. Δηλαδή, με βάσει τα προαναφερόμενα παραδείγματα θα ισχύει:

$$\text{gap\_groups} = \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ T & T & F & T & T & F \end{matrix}$$

Τέλος, στην μεταβλητή “gap_groups” εφαρμόζεται η μέθοδος “cumsum()” της βιβλιοθήκης pandas. Υπενθυμίζεται ότι οι Boolean τιμές μπορούν να αντιστοιχηθούν με ακέραιους πραγματικούς αριθμούς, όπου False = 0 Και True = 1. Επομένως το διάνυσμα μετατρέπεται σε:

$$\text{gap\_groups}.cumsum() = \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 2 & 3 & 4 & 4 \end{matrix}$$

Βάσει του παραπάνω γίνεται αντιληπτό ότι πλέον τα «κενά διαστήματα» είναι τα indexes που έχουν την ίδια ακέραια τιμή, δηλαδή οι θέσεις 1,2 και 4,5 και λόγο αυτού, μπορούν να ομαδοποιηθούν. Εφόσον έχουν ομαδοποιηθεί, αφαιρώντας την ελάχιστη τιμή από την μέγιστη τιμή, αφού το index αφορά datetimes, μπορεί να υπολογιστεί η χρονική διάρκεια κάθε κενού. Για να επιτευχθεί το παραπάνω, δημιουργήθηκε μια μεταβλητή "missing_blocks", στην οποία αρχικά λαμβάνονται τα indexes του dataframe βάσει της προαναφερόμενης μεταβλητής "mask", Δηλαδή λαμβάνονται μόνα τα indexes που παρουσιάζουν True στην συνθήκη παρουσιάζοντας NaN τιμές. Έπειτα το διάνυσμα μετατρέπεται σε series μέσω της μεθόδου "to_series()" και με την χρήση της μεθόδου "group_by" στο " gap_groups[mask]" εκτελείται η ομαδοποίηση. Με βάση τα προαναφερόμενα παραδείγματα, ισχύει:

$$\mathrm{df.index} = \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ t0 & t1 & t2 & t3 & t4 & t5 \end{matrix}$$

Όπου t_x, το timestamp – datetime για την x ώρα, δηλαδή έστω $t_0 = 2004 - 03 - 10\ 18{:}00{:}00$ και όπως προαναφέρθηκε, έστω "mask":

$$mask = \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ F & T & T & F & T & T \end{matrix}$$

Επομένως, διατηρώντας τα true statements, ισχύει:

$$\mathrm{df.index[mask].to\_series()} = \begin{matrix} 1 & 2 & 4 & 5 \\ t1 & t2 & t4 & t5 \end{matrix}$$

και:

$$\mathrm{gap\_groups[mask]} = \begin{matrix} 1 & 2 & 4 & 5 \\ 2 & 2 & 4 & 4 \end{matrix}$$

Πλέον, το dataframe έχει ομαδοποιηθεί βάσει των «κενών», και θα μπορούσε να αναπαρασταθεί ως:

$$missing\_blocks = \begin{matrix} 2 & 4 \\ t1, t2 & t4, t5 \end{matrix}$$

όπου ως index χρησιμοποιείται το group των «κενών», από την εφαρμογή της μεθόδου "group_by" στο "gap_groups[mask]" και ως row τα αντιστοιχούμενα timestamps – datetimes. Τέλος, χρησιμοποιώντας την μέθοδο ".agg(['min', 'max'])", δημιουργούνται 2 επιπλέον γνωρίσματα στο dataframe, όπου min αντιστοιχεί στο πρώτο timestamp του «κενού» και max στο τελευταίο timestamp. Ο λόγος που μπορεί να εκτελεστεί η παραπάνω πράξη είναι επειδή, όπως προαναφέρθηκε, η μεταβλητή "Datetime" έχει δομή datetime64 και εσωτερικά κάθε στοιχείο "datetime64[ns]" αποθηκεύεται ως ακέραιος αριθμός 64-bit, σε nanoseconds ως προεπιλογή, από την ορισμένη εποχή που χρησιμοποιείται από το σύστημα UTC – Unix 1970 [15,16].

Επομένως, ορίζοντας το γνώρισμα "duration_hours" εντός του "missing_blocks", αφαιρώντας το "min" από το "max" και με την χρήση της μεθόδου "total_seconds()" της βιβλιοθήκης datetime, το μέγεθος μετατρέπεται σε δευτερόλεπτα ώστε αν διαιρεθεί με 3600, το μέγεθος των δευτερολέπτων ανά ώρα, το αποτέλεσμα μετατρέπεται σε ώρες. Καθώς στην Python η καταμέτρηση στοιχείων ξεκινά από το 0, γίνεται αύξηση κατά 1 για αποφυγή σύγχυσης, καθώς διαφορετικά τα κενά μίας ώρας θα εμφανίζονται ως μηδενικά. Έπειτα, εφαρμόζονται 2 φίλτρα, να διατηρηθούν οι τιμές μεγαλύτερες του 0 και να ταξινομηθούν κατά φθίνουσα σειρά βάσει του γνωρίσματος "duration_hours".

Πίνακας 10 Παρουσίαση των 10 μεγαλύτερων συνεχών διαστημάτων για το γνώρισμα CO

	min	max	duration_hours
CO(GT)			
220	2004-10-13 11:00:00	2004-10-20 15:00:00	173.0
50	2004-04-17 03:00:00	2004-04-23 04:00:00	146.0
214	2004-10-01 18:00:00	2004-10-07 15:00:00	142.0
168	2004-07-26 03:00:00	2004-07-31 04:00:00	122.0
292	2004-12-28 01:00:00	2005-01-01 00:00:00	96.0
182	2004-08-19 21:00:00	2004-08-23 08:00:00	84.0
172	2004-08-03 01:00:00	2004-08-06 04:00:00	76.0
188	2004-08-28 14:00:00	2004-08-31 13:00:00	72.0
128	2004-06-02 01:00:00	2004-06-04 15:00:00	63.0
196	2004-09-11 01:00:00	2004-09-13 09:00:00	57.0

187 discrete gaps have been detected

Ο πίνακας 10 δείχνει τα 10 μεγαλύτερα συνεχόμενα κενά τιμών, κατά φθίνουσα σειρά. Αρχικά παρατηρείται ότι υπάρχουν πάνω από 10 κενά μεγαλύτερα της μιας ημέρας, δηλαδή 24 ωρών. Το μεγαλύτερο διάστημα με συνεχόμενες ελλιπείς τιμές είναι 173 ώρες, δηλαδή περίπου 1 εβδομάδα. Η μέθοδος “describe()”, μπορεί να χρησιμοποιηθεί για να παράγει στατιστικά μεγέθη για το γνώρισμα υπό μελέτη “duration_hours”:

Πίνακας 11 Στατιστικά μέτρα συνεχών διαστημάτων ελλιπών τιμών

count	187.000000
mean	9.000000
std	25.660522
min	1.000000
25%	1.000000
50%	1.000000
75%	1.000000
max	173.000000

Name: duration_hours, dtype: float64

Δηλαδή μέσω του πίνακα 11, γίνεται αντιληπτό ότι συνολικά υπάρχουν 187 συνολικά διαστήματα ελλιπών τιμών. Βάσει των ποσοστών των τεταρτημόριων φαίνεται ότι τα περισσότερα κενά διαστήματα αφορούν ελλείψεις διάρκειας μίας ώρας. Ωστόσο, παρατηρείται ότι η τυπική απόκλιση είναι 25,6 ώρες και η μέγιστη διάρκεια συνεχών ελλιπών τιμών είναι 173 ώρες.

Εάν γίνει αλλαγή στο προαναφερόμενο φίλτρο ώστε να παρακαμφθούν οι ελλιπές τιμές διάρκειας μιας ώρας, προκύπτει ο πίνακας 12:

Πίνακας 12 Στατιστικά μέτρα συνεχών διαστημάτων ελλιπών τιμών μεγαλύτερα από 1 ώρα

```
count     42.000000
mean      36.619048
std       44.490587
min        2.000000
25%        4.000000
50%       19.000000
75%       54.000000
max      173.000000
Name: duration_hours, dtype: float64
```

Πλέον παρατηρούνται 42 διακριτά κενά και φαίνεται ότι υπάρχουν αρκετά διαστήματα με διαφορετικά μεγέθη ελλιπών τιμών, βάσει των τεταρτημόριων.

3.2 Μέθοδοι αντιμετώπισης ελλιπών τιμών

Η ανάλυση χρονοσειρών δεν μπορεί να ξεκινήσει ακόμη και όταν υπάρχει έστω και μια ελλείπουσα τιμή ανάμεσα στις παρατηρήσεις για το σύνολο δεδομένων. Αυτό συμβαίνει επειδή οι παραδοσιακές τεχνικές μοντελοποίησης χρονοσειρών, όπως το ARIMA και τα μοντέλα LSTM, έχουν σχεδιαστεί για την ανάλυση αυτοσυσχέτισης και προτύπων σε διαδοχικά δεδομένα χωρίς κενά. Επομένως, οι ελλείπουσες τιμές πρέπει να αντιμετωπιστούν και να συμπληρωθούν κατάλληλα πριν πραγματοποιηθεί οποιαδήποτε ανάλυσης [30].

Τα δεδομένα είναι σε κατάλληλη μορφή για την αντιμετώπιση των ελλιπών τιμών , καθώς έχει οριστεί το γνώρισμα “Datetime” ως δείκτης των δεδομένων ενώ παράλληλα έχουν εντοπιστεί τα μεγέθη των κενών. Βέβαια, χρειάζεται να σημειωθεί ότι όπως αναφέρουν οι Junger και Ponce de Leon [18], η εφαρμογή και αξιολόγηση μεθόδων «imputation» ή αντιμετώπισης ελλιπών τιμών είναι ένα ιδιαίτερα δύσκολο έργο, καθώς εξ’ ορισμού δεν υπάρχει το ολοκληρωμένο σύνολο δεδομένων και η προσομοίωση ενός συνόλου δεδομένων με χρόνο - χωρικές εξαρτήσεις δεν είναι μια απλή υπόθεση καθώς ακόμη και το «καλύτερο» μοντέλο μπορεί να μην αποτυπώνει επαρκώς την δυναμική της πραγματικής κατάστασης.

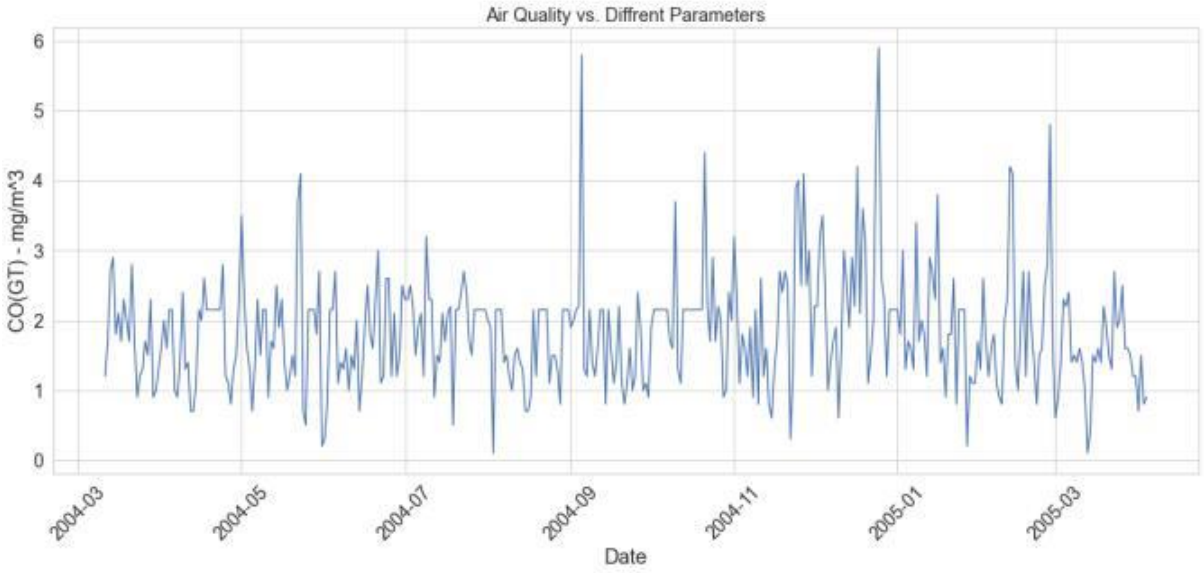
Ανάλογα με τον τρόπο και τους λόγους που οδηγούν στην απουσία παρατηρήσεων, τα ελλιπή δεδομένα ταξινομούνται σε δεδομένα που λείπουν εντελώς τυχαία (MCAR – Missing Completely At Random), δεδομένα που λείπουν τυχαία υπό προϋποθέσεις (MAR – Missing At Random) και δεδομένα που δεν λείπουν τυχαία (MNAR – Missing Not At Random). Στην περίπτωση του MCAR θεωρείται ότι οι ελλείπουσες παρατηρήσεις οφείλονται αποκλειστικά στην τύχη και δεν σχετίζονται ούτε με τις ίδιες τις τιμές των μεταβλητών ούτε με τις ήδη παρατηρούμενες τιμές. Η υπόθεση MAR υποδηλώνει ότι η απουσία τιμών δεν είναι απολύτως τυχαία, αλλά μπορεί να συνδέεται με τις άλλες μεταβλητές. Στις περιπτώσεις MCAR και MAR, η πιθανότητα εμφάνισης ελλιπών δεδομένων δεν εξαρτάται από τις ίδιες τις ελλείπουσες τιμές, γεγονός που καθιστά την αιτία της απουσίας τους «αγνοήσιμη» κατά τη στατιστική ανάλυση. Αντίθετα, στην περίπτωση MNAR θεωρείται ότι η έλλειψη δεδομένων σχετίζεται άμεσα με τα χαρακτηριστικά των ίδιων των τιμών που λείπουν [30].

Μια πιο απλοποιημένη μορφή των παραπάνω είναι ο χαρακτηρισμός των MCAR, ως η απουσία τιμών ανεξάρτητα από οποιαδήποτε άλλη τιμή, οι τιμές MAR όπου η απουσία τους εξαρτάται μόνο από τις παρατηρούμενες τιμές και οι τιμές MNAR, όπου η απουσία τους σχετίζεται τόσο με τις παρατηρούμενες όσο και με τις μη παρατηρούμενες τιμές [6].

Παράλληλα, οι Junnigen et al., επισημαίνουν ότι η απόδοση των μεθόδων συμπλήρωσης ελλιπών τιμών δεν εξαρτάται μόνο από το ποσοστό των ελλιπών δεδομένων, αλλά και από τα χαρακτηριστικά των προτύπων απουσίας τους. Επιπλέον, ο μηχανισμός εμφάνισης ελλιπών δεδομένων στην ποιότητα του αέρα θεωρείται συνήθως τυχαίος υπό προϋποθέσεις (MAR), με την έννοια ότι η πιθανότητα να λείπει μια τιμή δεν εξαρτάται από την ίδια την ελλείπουσα τιμή [31].

Υπάρχουν διάφοροι τρόποι να κατηγοριοποιηθούν οι διαφορετικές μέθοδοι αντιμετώπισης ελλιπών τιμών, όπως με βάσει αν το σετ δεδομένων είναι μονοδιάστατο ή πολυδιάστατο [18] ή μεταξύ κλασσικών - απλών μεθόδων, μεθόδων μηχανικής μάθησης και βαθιάς μάθησης [6]. Γενικά, οι κλασσικές μέθοδοι εφαρμόζονται σε μονοδιάστατα δεδομένα, δηλαδή σε “series” όπως για παράδειγμα το γνώρισμα CO(GT) που παρουσιάστηκε μέχρι τώρα, ενώ οι λοιπές προσεγγίσεις βασίζονται σε πολλαπλές διαστάσεις για την υλοποίηση τους. Σύμφωνα με τους Hua et al. [6] οι πιο συχνά χρησιμοποιούμενες απλές μέθοδοι είναι τρείς:

Διαγραφή, η οποία μπορεί να επηρεάσει σε πολύ μεγάλο βαθμό την ακεραιότητα των αποτελεσμάτων και πρακτικά δεν είναι εφικτή ως μέθοδος στις χρονοσειρές ιδιαίτερα όταν υπάρχουν μεγάλα διαστήματα κενών, παράλειψη – “ignoring” η οποία επίσης δεν εφαρμόζεται όταν υπάρχουν μεγάλα διαστήματα κενών και τέλος η αντικατάσταση των ελλιπών τιμών με ενδιάμεσης ή μέσης τιμής όπως προκύπτει. Υπενθυμίζεται ότι αυτή είναι η μέθοδος που ακολουθήθηκε από τους Kumar et al. [5] και είναι ευδιάκριτη ως προς τον εντοπισμό της από τις «σταθερές πεδιάδες» που δημιουργούνται μεταξύ των λοιπών διακυμάνσεων στις τιμές, όπως φαίνονται στο διάγραμμα 7.



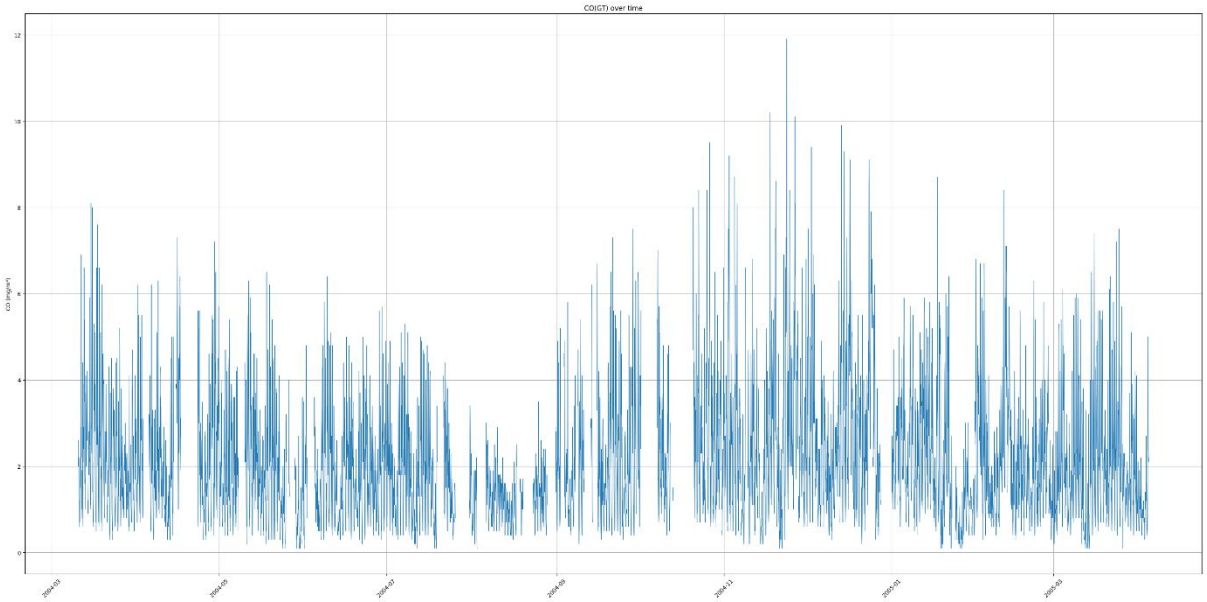
Διάγραμμα 7 Απεικόνιση CO ύστερα από εφαρμογή μέσης τιμής. Πηγή: [5]

Η χρήση της ενδιάμεσης τιμής είναι μια γρήγορη και πρακτική λύση ωστόσο οι Niako et al [30] ανέδειξαν ότι η συγκεκριμένη μέθοδος δεν μπορεί να ανταποκριθεί σε κενά διαστήματα μεγάλου μεγέθους αλλά ακόμα και σε μικρότερα διακριτά κενά, καθώς όσο αυξάνεται το ποσοστό των ελλιπών τιμών τόσο αυξάνονται και τα σφάλματα. Οι Junnigen et al. [31] αναφέρουν ότι η αντικατάσταση των ελλιπών τιμών με τη μέση ή την ενδιάμεση τιμή εφαρμόζεται συχνά και εξακολουθεί να προσφέρεται σε πολλά στατιστικά λογισμικά ως προεπιλογή για την αντιμετώπιση ελλιπών τιμών. Ωστόσο, η συγκεκριμένη μέθοδος μπορεί να αλλοιώσει σημαντικά την εσωτερική δομή των δεδομένων, προκαλώντας μεγάλες αποκλίσεις στους πίνακες συνδιακύμανσης και συσχέτισης και κατ’ επέκταση, να μειώσει την αποτελεσματικότητα των στατιστικών μοντέλων που θα εφαρμοστούν για πρόβλεψη [31].

Επιπλέον, άλλες μέθοδοι που χρησιμοποιούνται συχνά συμπεριλαμβάνουν την χρήση «two step Regression», ως βελτίωση της μεθόδου μέσου – ενδιάμεσης τιμής, “Last Observation Carried Forward (LOCF)” ή διαφορετικά “Forward Fill”, όπου για την αντικατάσταση των ελλιπών τιμών, χρησιμοποιούνται οι τιμές ενός προηγούμενου διαστήματος ή την αντίθετη μέθοδο “Back Fill”. Επίσης, όπως αναφέρθηκε προηγουμένως υπάρχουν μέθοδοι μηχανικής μάθησης για την εξάλειψη ελλιπών τιμών όπως Multivariate Imputation by Chained Equations (MICE) η οποία θεωρείται από τις βέλτιστες μεθόδους Imputation καθώς στην δημοσίευση των Hua et al. παρουσίαζε την καλύτερη ισορροπία μεταξύ απόδοσης μετρικών σφαλμάτων και χρόνου εκτέλεσης [6].

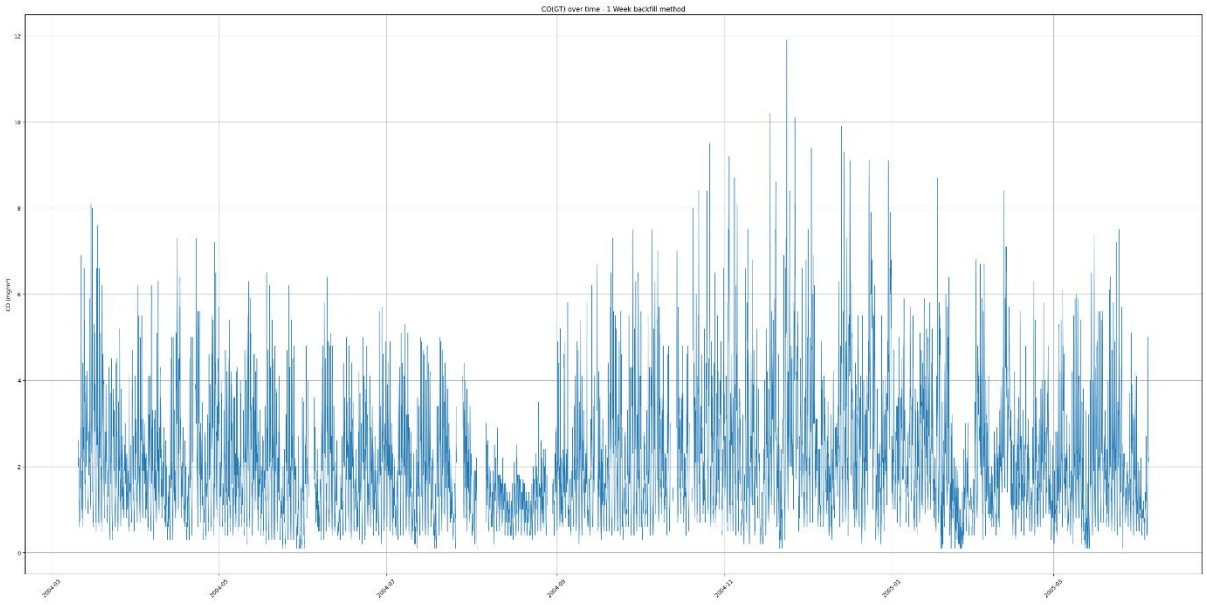
Η παρούσα εργασία εστιάζει στην εφαρμογή μεθόδων forward fill (LOCF) και back fill. Σύμφωνα με τους Niako et al. “παρόλο που πολλές μελέτες επισημαίνουν την ανάγκη προσοχής στη χρήση της μεθόδου LOCF για τη συμπλήρωση ελλιπών, λόγω περιορισμένης θεωρητικής τεκμηρίωσης και του κινδύνου εμφάνισης παραπλανητικών αποτελεσμάτων σε πολυετή αναλύσεις, τα δικά μας αποτελέσματα έδειξαν ότι το μοντέλο ARIMA που εφαρμόστηκε σε δεδομένα με συμπλήρωση μέσω LOCF, σε ποσοστό ελλιπών τιμών 25%”, που είναι παραπλήσιο σε αυτό που εμφανίζεται στο παρόν σετ δεδομένων (18%), “παρουσίασε την καλύτερη προβλεπτική απόδοση ως προς το RMSE και τη δεύτερη καλύτερη ως προς το MAPE” [30]. Επιπλέον, οι Junger και Ponce de Leon υποστηρίζουν ότι όσον αφορά univariate «απλές» μεθόδους, δηλαδή μεταξύ UM – Univariate Mean ,MD - Median και NN – Nearest Neighbor, το NN που αναφερόταν σε χρήση “last value carried forward” ή “next value carried backward”, είχε την καλύτερη απόδοση. Ωστόσο χρειάζεται να αναφερθεί ότι οι multivariate μέθοδοι παρήγαγαν πολύ καλύτερα αποτελέσματα [18].

Το διάγραμμα 8 δείχνει το γράφημα του γνωρίσματος CO(GT) πριν από οποιαδήποτε επεξεργασία σχετικά με την αντιμετώπιση των ελλιπών τιμών, στο οποίο υπάρχουν 1683 ελλιπής τιμές.



Διάγραμμα 8 Απεικόνιση CO πριν από επεξεργασία

Το διάγραμμα 9, δείχνει το γράφημα του γνωρίσματος CO(GT) αφού έχει εφαρμοστεί, όπου είναι εφικτό, αντικατάσταση με τις τιμές της προηγούμενης εβδομάδας ανά ώρα.



Διάγραμμα 9 Απεικόνιση CO ύστερα από χρήση forward fill

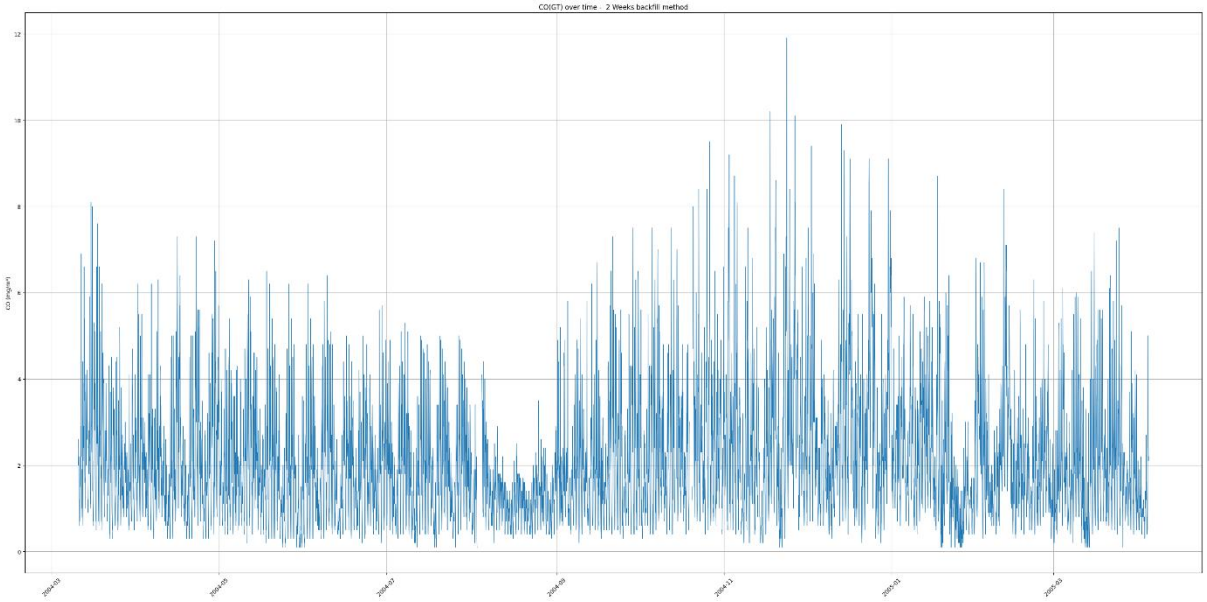
Σε σχέση με το αρχικό γράφημα, παρατηρείτε σημαντική βελτίωση καθώς πλέον εντοπίζονται 476 ελλιπείς τιμές, ωστόσο συνεχίζεται η ύπαρξη «κενών», πιθανώς επειδή η αντικατάσταση των τιμών δεν μπορεί να αντιμετωπίσει πλήρως κενά μεγαλύτερα ή ίσα της 1 εβδομάδας. Λόγω αυτού, έγινε δοκιμή διαφορετικών μεθοδολογιών για την εύρεση της βέλτιστης μεθόδου.

Πίνακας 13 Παρουσίαση των 10 μεγαλύτερων συνεχών διαστημάτων για το γνώρισμα CO ύστερα από χρήση forward fill

CO(GT)	min	max	duration_hours
152	2004-08-03 01:00:00	2004-08-06 04:00:00	76.0
146	2004-07-26 12:00:00	2004-07-28 17:00:00	54.0
126	2004-06-02 17:00:00	2004-06-04 14:00:00	46.0
158	2004-08-28 14:00:00	2004-08-30 08:00:00	43.0
184	2004-10-13 11:00:00	2004-10-14 15:00:00	29.0
188	2004-10-19 09:00:00	2004-10-20 08:00:00	24.0
182	2004-10-12 09:00:00	2004-10-13 08:00:00	24.0
44	2004-04-21 13:00:00	2004-04-22 08:00:00	20.0
166	2004-09-14 22:00:00	2004-09-15 16:00:00	19.0
178	2004-10-07 01:00:00	2004-10-07 11:00:00	11.0

Ο πίνακας 13 παρουσιάζει τα 10 μεγαλύτερα κενά διαστήματα για το γνώρισμα CO(GT), ύστερα από την εφαρμογή backfill 1 εβδομάδας. Πλέον, το μεγαλύτερο κενό είναι 76 ώρες, δηλαδή περίπου 3 μέρες.

3.2.1 1η Μέθοδος: 2 εβδομάδες forward fill - Time interpolation



Διάγραμμα 10 Απεικόνιση CO ύστερα από χρήση forward fill

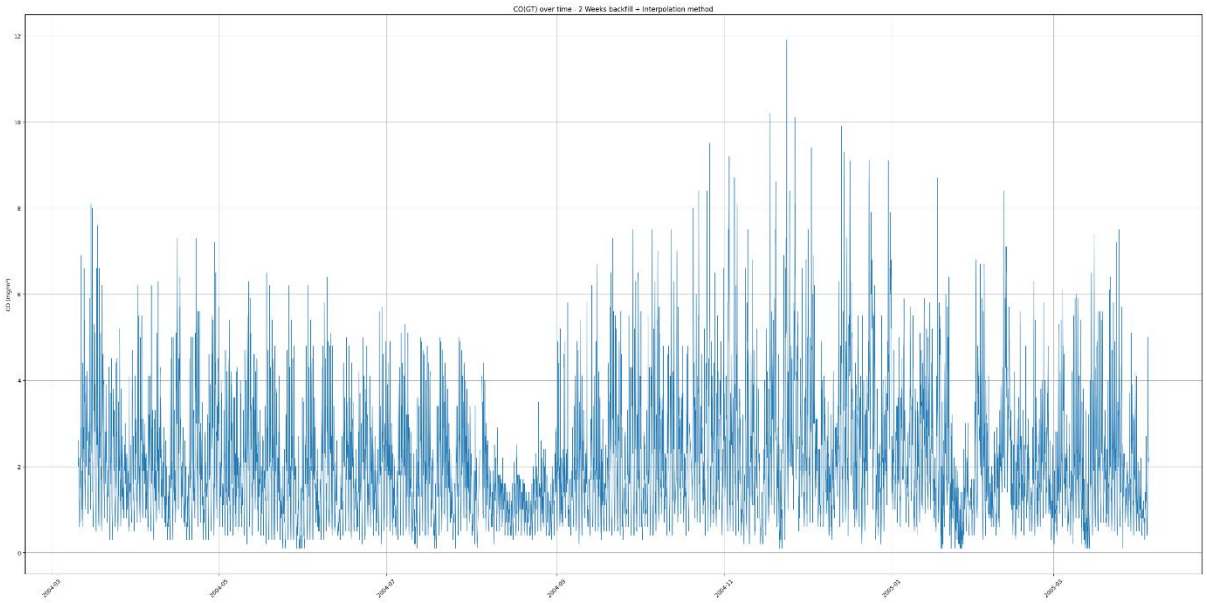
Αρχικά έγινα απόπειρα επικάλυψης των ελλιπών τιμών εφαρμόζοντας την μέθοδο 1 εβδομάδας forward fill, δηλαδή με επανάληψη της μεθόδου που προαναφέρθηκε. Με αυτόν τον τρόπο απομένουν 167 ελλιπής τιμές, επομένως είναι αναγκαία περεταίρω επεξεργασία.

Πίνακας 14 Παρουσίαση συνεχών διαστημάτων για το γνώρισμα CO ύστερα από διπλή χρήση forward fill

	min	max	duration_hours
CO(GT)			
126	2004-08-03 01:00:00	2004-08-04 17:00:00	41.0
148	2004-10-19 09:00:00	2004-10-20 08:00:00	24.0
144	2004-10-14 01:00:00	2004-10-14 11:00:00	11.0
150	2004-10-20 11:00:00	2004-10-20 15:00:00	5.0
128	2004-08-05 08:00:00	2004-08-05 09:00:00	2.0

5 discrete gaps have been detected

Είναι σημαντικό να αναφερθεί ότι προφανώς μπορεί να χρησιμοποιηθούν και μεγαλύτερα διαστήματα ως προς το interpolation, ωστόσο μεγαλύτερα διαστήματα δεν παρουσιάζουν ικανοποιητικά αποτελέσματα και έτσι ορίστηκε ως argument στην μέθοδο “interpolate()” “limit=24”. Ο πίνακας 14 δείχνει ότι με εξαίρεση το πρώτο και μεγαλύτερο διάστημα, το οποίο παρουσιάζει κενό 41 ώρες και το δεύτερο κενό που είναι οριακό, με 24 ώρες, τα υπόλοιπα είναι μικρότερα ενός 24-ώρου επομένως θεωρείται ότι η μέθοδος interpolation μπορεί να χρησιμοποιηθεί. Επίσης, υπενθυμίζεται ότι στην μέθοδο ανίχνευσης κενών έχει εφαρμοστεί φίλτρο ώστε να διατηρούνται μόνο τα κενά μεγαλύτερα της μιας ώρας. Δηλαδή πράγματι υπάρχουν 167 ελλιπής τιμές, όπου τα παραπάνω είναι τα συνεχόμενα κενά διαστήματα και οι υπόλοιπες 167 – 83 (που προκύπτει από το άθροισμα των συνεχών διαστημάτων του παραπάνω πίνακα) = 84 ώρες είναι κενά μιας ώρας.



Διάγραμμα 11 Απεικόνιση CO ύστερα από διπλή χρήση forward fill

Στο διάγραμμα 11 δείχνει το γράφημα του CO(GT) ύστερα από χρήση interpolation. Ωστόσο, παραμένουν 17 ελλιπής τιμές, οι οποίες εντοπίζονται για το διάστημα του πίνακα 15.

Πίνακας 15 Υπολειπόμενο διάστημα ελλιπών τιμών ύστερα από χρήση διπλού forward fill

	min	max	duration_hours
CO(GT)			
2	2004-08-04 01:00:00	2004-08-04 17:00:00	17.0

1 discrete gaps have been detected

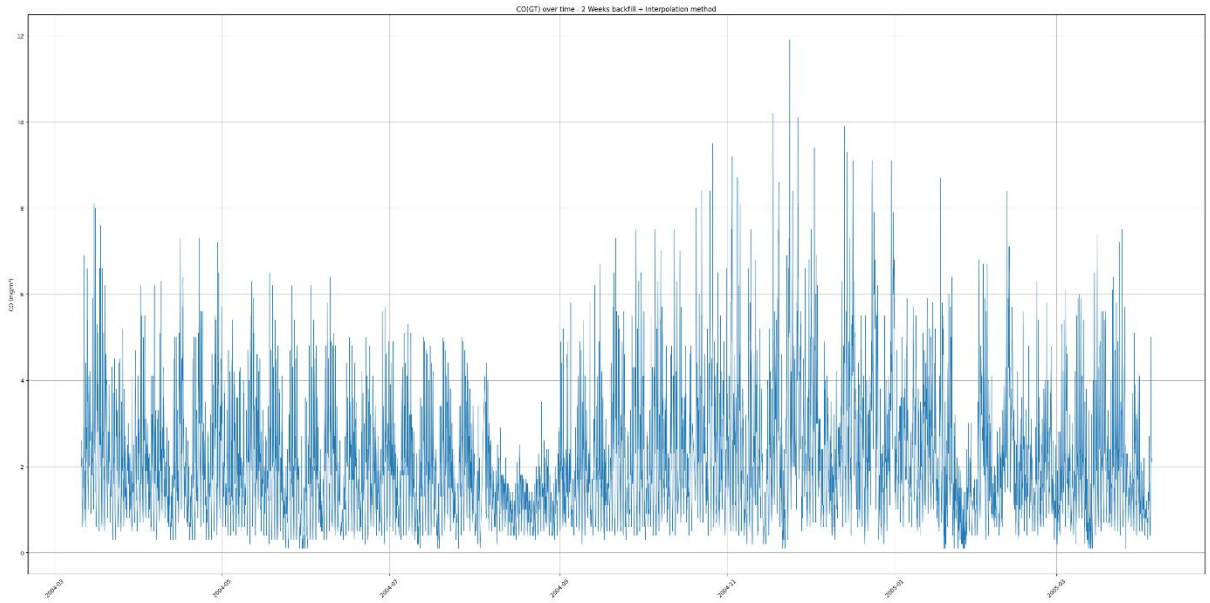
Δηλαδή, αφορά το αρχικό και μεγαλύτερο διάστημα που παρουσιάστηκε προηγουμένως και δεν ήταν εφικτό να επικαλυφθεί. Αυτό είναι απολύτως λογικό αφού το μεγαλύτερο κενό που παρατηρείται είναι 41 ώρες. Επομένως αφού χρησιμοποιηθεί interpolation 24 ωρών προκύπτει το κενό των 17 ωρών, αφού 41 – 24 = 17.

Εάν στην μέθοδο interpolation αλλάξει η τιμή της παραμέτρου “limit” σε 48 ώρες, η οποία είναι μεγαλύτερη από το μεγαλύτερο παρατηρούμενο κενό (40 ώρες), ως αποτέλεσμα δεν υπάρχουν πλέον ελλιπείς τιμές, όπως φαίνεται από τον πίνακα 16.

Πίνακας 16 Μη ύπαρξη ελλιπών τιμών

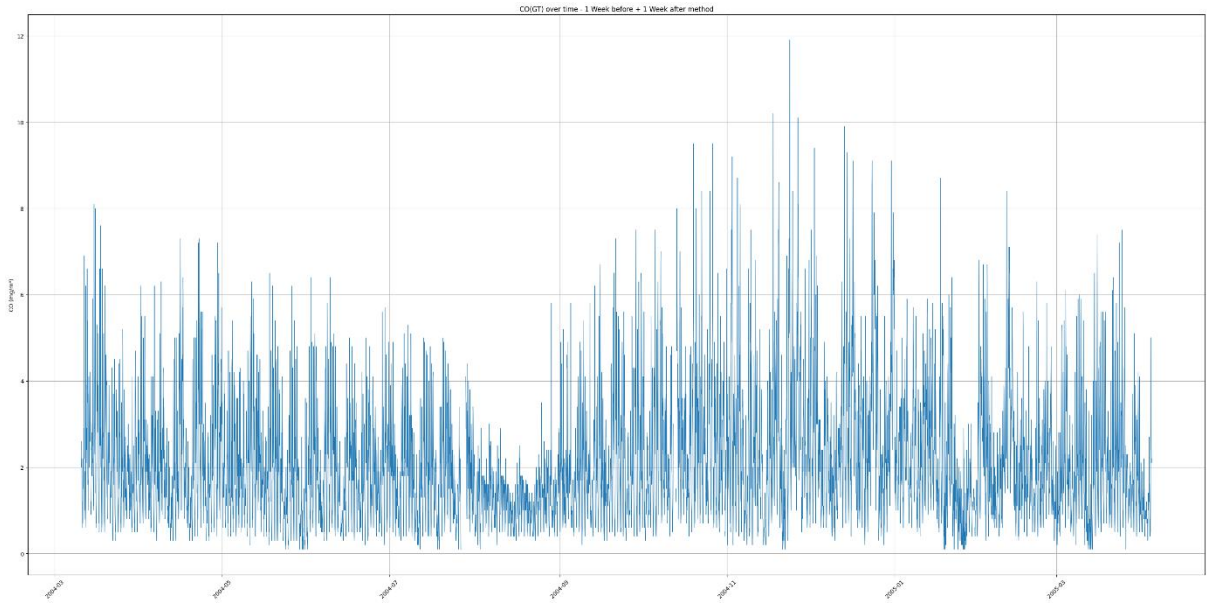
Empty DataFrame
Columns: [min, max, duration_hours]
Index: []

0 discrete gaps have been detected



Διάγραμμα 12 Απεικόνιση CO ύστερα από διπλή χρήση forward fill και time interpolation

3.2.2 2η Μέθοδος: 1 εβδομάδα forward fill - 1 εβδομάδα back fill - Time Interpolation



Διάγραμμα 13 Απεικόνιση CO με χρήση forward fill και back fill

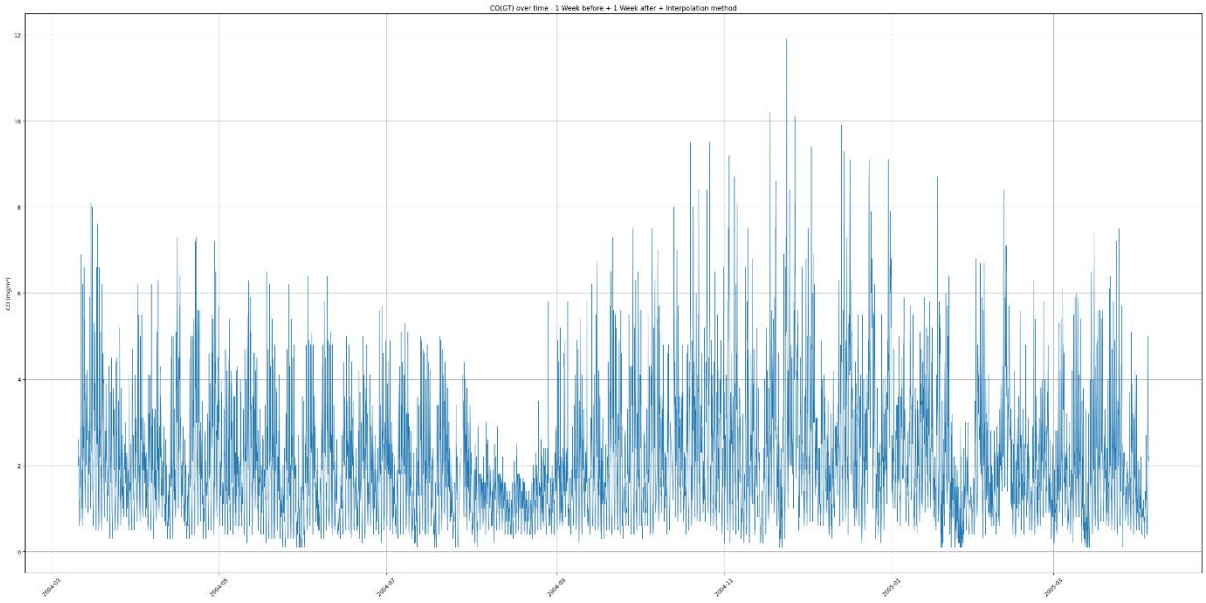
Το διάγραμμα 13, απεικονίζει το γράφημα του CO(GT) συνδυάζοντας την μεθοδολογία back- fill και forward – fill. Δηλαδή αρχικά αντιγράφονται, όπου είναι εφικτό, οι τιμές τις προηγούμενης εβδομάδας, όπως παρουσιάστηκαν στο διάγραμμα 10 και μετέπειτα αντιγράφονται οι τιμές της επόμενης εβδομάδας. Το σκεπτικό ήταν να μην δημιουργηθεί “bias” στα δεδομένα, χρησιμοποιώντας τιμές μόνο της προηγούμενης εβδομάδας, καθώς οι τιμές της επόμενης μπορεί να διαφέρουν σε σημαντικό βαθμό. Για παράδειγμα, οι καλοκαιρινοί μήνες παρουσιάζουν σημαντικά χαμηλότερα μεγέθη σε σχέση με τους λοιπούς μήνες και συνεπώς η χρήση μόνο της προηγούμενης εβδομάδας μπορεί να παρουσιάσει απότομες αλλαγές που πιθανώς να μην ήταν λογικές. Με την χρήση της

παραπάνω μεθόδου παρατηρούνται 159 ελλιπές τιμές, με τα κενά διαστήματα να παρουσιάζονται στην παρακάτω εικόνα.

Πίνακας 17 Παρουσίαση συνεχών διαστημάτων για το γνώρισμα CO ύστερα από χρήση forward fill και back fill

CO(GT)	min	max	duration_hours
110	2004-07-27 01:00:00	2004-07-28 17:00:00	41.0
132	2004-10-12 09:00:00	2004-10-13 08:00:00	24.0
128	2004-10-07 01:00:00	2004-10-07 11:00:00	11.0
134	2004-10-13 11:00:00	2004-10-13 15:00:00	5.0
112	2004-07-29 08:00:00	2004-07-29 09:00:00	2.0

Βάσει του πίνακα 17, τα κενά διαστήματα είναι τα ίδια με της 1ης μεθόδου, επομένως η διαφορές τους εντοπίζονται στις διακριτές ελλιπής τιμές. Όπως παρουσιάστηκε και στην 1η μέθοδο, εφαρμόστηκε time interpolation για τις λοιπές τιμές.



Διάγραμμα 14 Απεικόνιση CO με χρήση forward fill, back fill και time interpolation

Το διάγραμμα 14, παρουσιάζει τα αποτελέσματα ύστερα από την εφαρμογή time interpolation. Ακριβώς όπως και στην 1η μέθοδο, παρατηρούνται 17 ελλιπής τιμές, οι οποίες μπορούν να εξεταστούν για την παρουσία συνεχών και διακριτών διαστημάτων.

Πίνακας 18 Υπολειπόμενο διάστημα ελλιπών τιμών ύστερα από χρήση forward fill, back fill και time interpolation

CO(GT)	min	max	duration_hours
2	2004-07-28 01:00:00	2004-07-28 17:00:00	17.0

Πράγματι όπως φαίνεται στον πίνακα 18, παρατηρείται έναν εναπομείναντα κενό διάστημα καθώς και μια διακριτή ελλιπή τιμή, δηλαδή παραμένουν στο σύνολο 17 ελλιπής τιμές, δηλαδή όπως και στην 1η μέθοδο. Με παρόμοιο τρόπο, εάν οριστεί ως “limit” 48 ώρες, οι ελλιπείς τιμές δεν υφίστανται πλέον.

3.2.3 3η Μέθοδος: 1 εβδομάδα forward fill – εβδομαδιαία διάμεσος ανά ώρα

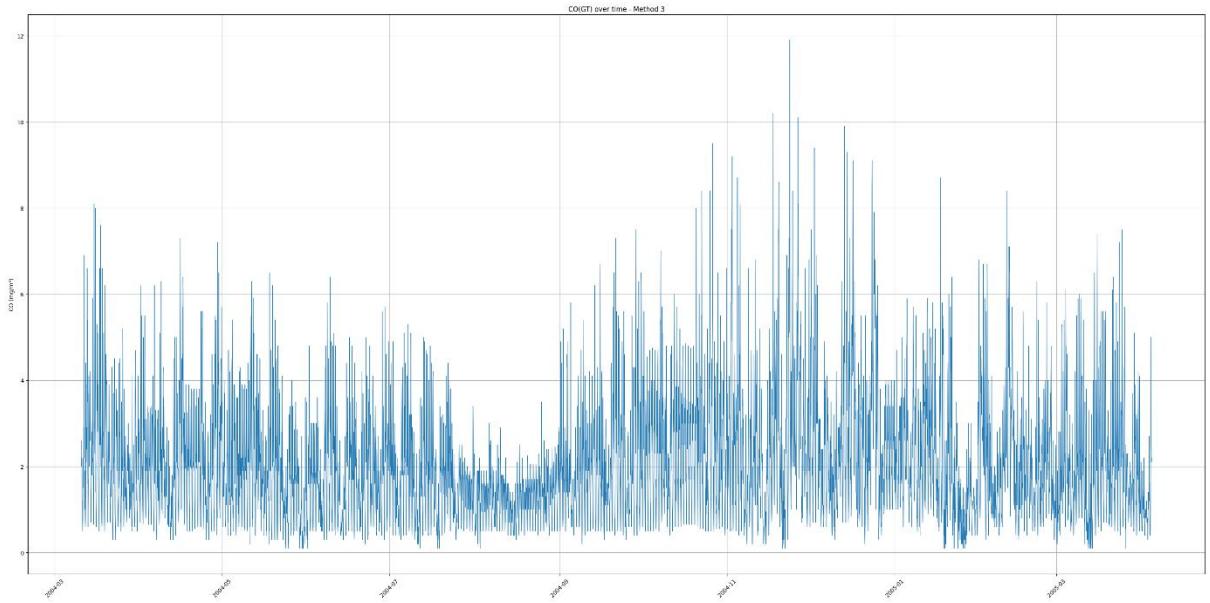
Στην συνέχεια έγινε απόπειρα μιας πιο σύνθετης μεθόδου. Αρχικά, όπως και στις προηγούμενες μεθοδολογίες, αντιγράφονται οι τιμές της προηγούμενης εβδομάδας. Στην συνέχεια, έναντι από την απλή αντιγραφή των τιμών της προηγούμενης και της επόμενης εβδομάδας, θεωρήθηκε ότι μια καλύτερη λύση θα ήταν η αντικατάσταση κάθε ελλιπής τιμής ανά ώρα Χ με την ενδιάμεση τιμή που προκύπτει από την τιμής της ίδιας ώρας προηγούμενης εβδομάδας και της ίδιας ώρας επόμενης εβδομάδας. Δηλαδή, για κάθε ελλιπή τιμή, ορίζεται ένα ± « συμμετρικό παράθυρο» για τις εβδομάδες που θα χρησιμοποιηθούν, επιλέγονται όλα τα δείγματα για την ίδια ώρα και ημέρα υπό μελέτη και υπολογίζεται η διάμεσος αντικαθιστώντας την τιμή. Σε περίπτωση που το «παράθυρο» δεν περιλαμβάνει ορθά δείγματα για την αντικατάσταση των τιμών, παραμένουν ως NaN. Θεωρείται ότι με αυτόν τον τρόπο διατηρούνται τα εποχιακά μοτίβα με μικρό βαθμό αλλοίωσης. Το παραπάνω προκύπτει από τις παρατηρήσεις των Freeman et al. [60], οι οποίοι εφάρμοσαν μια παρόμοια μεθοδολογία για συνεχόμενα κενά διαστήματα μεγαλύτερα των 8 ωρών [60].

Πιο συγκεκριμένα, ορίστηκε η συνάρτηση “fill_with_hourly_week_median”, με παραμέτρους “series”, όπου εισάγεται η υπό μελέτη χρονοσειρά – series και “window_weeks”, το «παράθυρο» για τις εβδομάδες. Η μεταβλητή “filled” δημιουργεί ένας αντίγραφο της παραμέτρου “series”. Στην συνέχεια εφαρμόζεται ένα “for loop” όπου διατηρούνται οι δείκτες των ελλιπών τιμών, με το καθένα να λαμβάνει επεξεργασία ξεχωριστά λόγο του προαναφερόμενου loop. Στην συνέχεια ορίζονται οι μεταβλητές “start” και “end” που θα λειτουργήσουν ως η αρχή και το τέλος του προαναφερόμενου παραθύρου, δηλαδή για κάθε τιμή – δείκτη, αφαιρείται το παράθυρο αφού μετατραπεί σε μορφή “Timedelta” ως αρχή και προστίθεται με τον ίδιο τρόπο για την εύρεση του τελικού δείκτη του παραθύρου. Επίσης, ορίζεται η μεταβλητή “hour” που θα λαμβάνει από τον δείκτη, που βρίσκεται σε μορφή “datetime64” την ώρα.

Πίνακας 19 Επεξήγηση μεταβλητών hour, end και hour

```
start = idx - pd.Timedelta(weeks=window_weeks)
end   = idx + pd.Timedelta(weeks=window_weeks)
hour  = idx.hour
```

Στην συνέχεια, εφαρμόζεται ένα φίλτρο “mask” ώστε οι τιμές να βρίσκονται εντός του παραθύρου και αφορούν την ίδια ώρα για το υπό μελέτη δείγμα. Δηλαδή, μεγαλύτερες ή ίσες της αρχής “start”, μικρότερες ή ίσες του τέλους “end” και “series.index.hour == hour”, ώστε το παράθυρο να διατηρεί μόνο τις τιμές της ίδιας ώρας – δείκτη υπό μελέτη. Σε περίπτωση που το παράθυρο περιέχει ελλιπής τιμές, αυτές οι τιμές αφαιρούνται, διότι διαφορετικά θα πρόκυπτε σφάλμα. Τέλος, εφόσον ισχύει η προηγούμενη συνθήκη, υπολογίζεται η ενδιάμεση τιμή του παραθύρου και επιστρέφεται ως το αποτέλεσμα της μεθόδου.



Διάγραμμα 15 Απεικόνιση CO με χρήση forward fill και εβδομαδιαίου διάμεσου ανά ώρα

Το διάγραμμα 15 παρουσιάζει το γράφημα του CO(GT) ύστερα από την εφαρμογή της μεθόδου με το παράθυρο να έχει οριστεί ως 1 εβδομάδα. Μέσω της εντολής df['CO(GT)'].isnull().sum()}, ανιχνεύονται 36 ελλιπής τιμές .Ύστερα, χρησιμοποιείτε πάλι η συνάρτηση detect gaps, ωστόσο θα είναι κενή αφού δεν υπάρχουν συνεχόμενα κενά διαστήματα, μόνο διακριτά κενά μιας ώρας σύμφωνα με τον πίνακα 20.

Πίνακας 20 Μη ύπαρξη ελλιπών τιμών ύστερα από την μέθοδο 3

```
Empty DataFrame
Columns: [min, max, duration_hours]
Index: []
```

Εάν αφαιρεθεί το φίλτρο >1 για την μέθοδο detect gaps θα είναι εφικτό να προσδιοριστούν τα μοναδιαία διαστήματα, όπως φαίνεται στον πίνακα 21 για τις πρώτες 12 τιμές.

Πίνακας 21 Μοναδιαία διαστήματα για την μέθοδο 3

CO(GT)	min	max	duration_hours
2	2004-05-08 04:00:00	2004-05-08 04:00:00	1.0
4	2004-05-09 04:00:00	2004-05-09 04:00:00	1.0
6	2004-05-10 04:00:00	2004-05-10 04:00:00	1.0
8	2004-05-11 04:00:00	2004-05-11 04:00:00	1.0
10	2004-05-12 04:00:00	2004-05-12 04:00:00	1.0
12	2004-05-13 04:00:00	2004-05-13 04:00:00	1.0
14	2004-05-14 04:00:00	2004-05-14 04:00:00	1.0
16	2004-05-15 04:00:00	2004-05-15 04:00:00	1.0
18	2004-05-16 04:00:00	2004-05-16 04:00:00	1.0
20	2004-05-17 04:00:00	2004-05-17 04:00:00	1.0
22	2004-05-18 04:00:00	2004-05-18 04:00:00	1.0
24	2004-05-19 04:00:00	2004-05-19 04:00:00	1.0
26	2004-05-20 04:00:00	2004-05-20 04:00:00	1.0
28	2004-05-21 04:00:00	2004-05-21 04:00:00	1.0

3.3 Αξιολόγηση μεθόδων αντιμετώπισης ελλιπών τιμών

Σύμφωνα με την βιβλιογραφία η επιλογή της κατάλληλης μεθόδου δεν μπορεί να γίνει αυθαίρετα αλλά μέσω της χρήσης “artificial missingness” [6, 31]. Για την αξιολόγηση των μεθόδων χρειάζεται να βρεθεί μια χρονική περίοδος που δεν θα παρουσιάζει ελλιπής τιμές. Στην συνέχεια θα εφαρμοστεί “artificial missingness”, δηλαδή θα δημιουργηθούν τεχνητά διαστήματα ελλιπών τιμών, παρόμοια με αυτά που παρουσιάζει το σετ δεδομένων αλλά υπό κλίμακα, δεδομένου ότι το επιλεγμένο χρονικό διάστημα προς αξιολόγηση θα είναι σημαντικά μικρότερο από το σύνολο. Έπειτα μπορούν να εφαρμοστούν οι διαφορετικές μέθοδοι για την αντιμετώπιση των ελλιπών τιμών και τα αποτελέσματα να συγκριθούν ως προς το αρχικό διάστημα χρησιμοποιώντας τα μέτρα Mean Absolute Error (MAE), Root Mean Square Error (RMSE) και R^2 [18, 60].

Σχετικά με την δημιουργία των τεχνητών κενών διαστημάτων χρειάζεται να οριστεί ο τρόπος που θα δημιουργηθούν αυτά τα «κενά», καθώς μπορούν είτε να είναι πραγματικά τυχαία κενά, δηλαδή διακριτά κενά ή να χρησιμοποιηθούν διαστήματα κενών. Οι Belachsen και Broday [17] δημιουργήσαν διάφορα N X L διαστήματα κενών, όπου ως N ορίστηκε ο αριθμός των διαστημάτων και L η χρονική διάρκεια του διαστήματος, για παράδειγμα 180 X 2 ώρες. Ωστόσο, θεωρήθηκε σκόπιμο αντί για τυχαία διαστήματα να χρησιμοποιηθούν τα πραγματικά κενά που παρατηρήθηκαν, ύστερα από την χρήση backfill 1 εβδομάδας, υπό κλίμακα αφού αυτά τα κενά παρατηρούνται στο σύνολο των δεδομένων, δηλαδή ενός χρόνου.

Αναλυτικότερα, πρώτα είναι αναγκαία η εύρεση ενός διαστήματος, έστω ενός μήνα, που να μην παρουσιάζει ελλιπής τιμές. Ως “s” ορίστηκε το γνώρισμα CO(GT) από το αντίγραφο του dataframe που αναφέρθηκε προηγουμένως (df_original), δηλαδή πριν από κάποια χρήση μεθόδου αντιμετώπισης ελλιπών τιμών.

Με παρόμοιο τρόπο όπως και στην μέθοδο για την εύρεση των κενών διαστημάτων, χρησιμοποιήθηκε ξανά το φίλτρο “mask”, όμως σε αντίθεση με την προηγούμενη φορά χρησιμοποιήθηκε η μέθοδος “notna()”, δηλαδή τιμή True θα λάβουν οι τιμές non – Nan ενώ False οι τιμές Nan.

Επιπλέον, η μεταβλητή “group_id” παρέχει την ίδια χρήση που παρουσιάστηκε για την μεταβλητή “gap_groups”, δηλαδή γίνεται αύξηση κατά +1 index και σύγκριση των μεταξύ True – False συνδυαστικά με την μέθοδο “cumsum()” για την εύρεση των διαστημάτων. Ύστερα ορίζεται η μεταβλητή - dataframe “blocks” εντός της οποίας γίνεται η ομαδοποίηση ανά “group_id” και η εύρεση της αρχής και τέλους του κάθε “group_id” βάσει των ελάχιστων και μέγιστων. Η βασική διαφορά των 2 μεθόδων είναι μεταξύ df_original.index.to_series() και df.index[mask], καθώς ο πρώτος όρος θα διατηρήσει όλα τα timestamps, δημιουργώντας ένα καινούργιο πίνακα, που θα έχει ως index τα timestamps Και ως γνώρισμα επίσης τα timestamps, ενώ ο δεύτερος όρος θα διατηρήσει μόνο τα indexes που έχουν λάβει τιμή True, δηλαδή όλες τις ελλιπής τιμές.

Στην συνέχεια ορίζεται η μεταβλητή “group_lengths”, όπου ελέγχεται η χρονική διάρκεια του κάθε διαστήματος, βάσει του “group_id” και “mask”, μέσω της μεθόδου “size()”. Πρακτικά εδώ γίνεται καταμέτρηση των timestamps ανά κάθε “group_id” αλλά αφού κάθε timestamp αντιστοιχεί σε μία ώρα, το μέγεθος του διαστήματος αντιστοιχεί στο μέγεθος του διαστήματος. Ωστόσο, όπως προαναφέρθηκε, επειδή διατηρούνται όλα τα διαστήματα, δηλαδή ελλιπή και μη – ελλιπή, είναι αναγκαία η διαφοροποίηση μεταξύ τους.

Για αυτόν τον λόγο δημιουργήθηκε η μεταβλητή “is_not_nan_block”. Όπως και στην προηγούμενη μέθοδο το series “s” ομαδοποιείται βάσει “group_id” όμως διαφοροποιείται καθώς εδώ χρησιμοποιείται η μέθοδος “all()”, δηλαδή ένας Boolean έλεγχος όπου αν όλες οι τιμές εντός ενός

“group_id” είναι μη ελλιπείς τιμές, το block χαρακτηρίζεται ως True, διαφορετικά χαρακτηρίζεται ως False. Πιο συγκεκριμένα, η μέθοδος “all” ελέγχει εάν όλα τα στοιχεία ενός επαναληπτικού στοιχείου (iterable) όπως λίστα, λεξικό και παρόμοια, είναι True, διαφορετικά επιστρέφει False [84]. Επειδή εφαρμόζεται στο φίλτρο “mask” υπό ομαδοποίηση “group_id”, που χρησιμοποιεί την μέθοδο “notna()”, ελέγχει εάν κάθε στοιχείο σε ένα “group_id” έχει τιμή “mask” == True, δηλαδή δεν είναι ελλιπής και αντίθετα. Προφανώς, εάν δεν διαταχθούν βάσει κάποιου άλλου γνωρίσματος, τότε τα διαστήματα θα εναλλάσσονται μεταξύ διαστημάτων ελλιπών και μη – ελλιπών τιμών. Ύστερα, ορίζεται ένα καινούργιο dataframe “results” για την απόδοση των παραπάνω με δείκτη το “group_id”. Τα 10 μεγαλύτερα διαστήματα σε φθίνουσα σειρά παρουσιάζονται στον πίνακα 22.

Πίνακας 22 Έλεγχος διαστημάτων ανά μήκος

CO(GT)	length	is_not_nan_block
220	173	False
50	146	False
214	142	False
168	122	False
292	96	False
153	95	True
182	84	False
172	76	False
273	73	True
188	72	False

Αρχικά, μπορεί να γίνει αντιστοίχιση των κενών διαστημάτων με αυτά του αρχικού σετ δεδομένων του πίνακα 10. Παρατηρείται ότι το μεγαλύτερο διάστημα συνεχών έγκυρων τιμών βρίσκεται στον δείκτη group id 153 με μήκος 95 ώρες ή διαφορετικά, σχεδόν 4 ημέρες. Παρόλο που το μεγαλύτερο αποδεκτό διάστημα έχει βρεθεί, θεωρείται καλή πρακτική να βρεθούν τα υπόλοιπα αποδεκτά διαστήματα. Για αυτό τον λόγο, ορίζεται στο dataframe “blocks” το γνώρισμα “all_valid”, το οποίο θα λάβει τις τιμές του “is_not_nan_block” και το γνώρισμα “duration” για τον υπολογισμό της διάρκειας του κάθε διαστήματος αφαιρώντας το μεγαλύτερο timestamp από το μικρότερο.

Για να διατηρηθούν μόνο τα αποδεκτά διαστήματα, δηλαδή αυτά που δεν είναι διαστήματα ελλιπών τιμών, ορίζεται το dataframe “valid_blocks”, επιλέγοντας τα δείγματα του “blocks” που τηρούν την συνθήκη “['all_valid']== True”. Τέλος, μπορεί να εφαρμοστεί στο καινούργιο dataframe η μέθοδος “sort_values()”, βάσει του γνωρίσματος “duration”, τα 10 μεγαλύτερα έγκυρα διαστήματα σε φθίνουσα σειρά παρουσιάζονται στον πίνακα 23.

Πίνακας 23 Έγκυρα συνεχή διαστήματα

CO(GT)	start	end	all_valid	duration
153	2004-07-08 05:00:00	2004-07-12 03:00:00	True	3 days 22:00:00
273	2004-11-28 03:00:00	2004-12-01 03:00:00	True	3 days 00:00:00
341	2005-02-18 05:00:00	2005-02-21 03:00:00	True	2 days 22:00:00
365	2005-03-20 05:00:00	2005-03-23 03:00:00	True	2 days 22:00:00
281	2004-12-10 05:00:00	2004-12-13 03:00:00	True	2 days 22:00:00
279	2004-12-07 05:00:00	2004-12-10 03:00:00	True	2 days 22:00:00
277	2004-12-04 05:00:00	2004-12-07 03:00:00	True	2 days 22:00:00
275	2004-12-01 05:00:00	2004-12-04 03:00:00	True	2 days 22:00:00
367	2005-03-23 05:00:00	2005-03-26 03:00:00	True	2 days 22:00:00
269	2004-11-23 05:00:00	2004-11-26 03:00:00	True	2 days 22:00:00

Επιβεβαιώνεται ότι το μεγαλύτερο συνεχόμενο διάστημα χωρίς ελλιπείς τιμές είναι 3 μέρες και 22 ώρες ενώ το δεύτερο μεγαλύτερο διάστημα είναι 3 μέρες. Τα υπάρχοντα διαστήματα είναι υπερβολικά μικρά σε διάρκεια ώστε να χρησιμοποιηθούν για την αξιολόγηση των μεθόδων, επομένως χρειάζεται να εφαρμοστεί μια διαφορετική μέθοδος εύρεσης κατάλληλου διαστήματος.

Βάσει του διαγράμματος 9, φαίνεται ότι υπάρχουν διαστήματα τουλάχιστον ενός μήνα που παρουσιάζουν εντός τους μικρότερα διαστήματα ελλιπών τιμών, τα οποία θα μπορούσαν να γίνουν κατάλληλα για χρήση με την εφαρμογή σύντομου time interpolation. Για την ανίχνευση αυτών των διαστημάτων το σκεπτικό ήταν να βρεθούν διαστήματα ενός μήνα τα οποία να είναι κατά X % αποδεκτά, δηλαδή έστω διάστημα ενός μήνα με 90% αποδεκτές τιμές και κατ' επακόλουθο 10% ελλιπής τιμές, άρα 24 ώρες x 31 ημέρες = 744 και 744 x 0.9 = 670 έγκυρες τιμές – ώρες.

Βάσει των παραπάνω, ορίστηκε η συνάρτηση "find_almost_clean_windows" με παραμέτρους "series" όπου θα εισάγεται η υπό μελέτη σειρά, "days" όπου θα εισάγεται το επιθυμητό διάστημα του παραθύρου σε ημέρες, "threshold" στο οποίο ορίζεται το κατώτατο όριο έγκυρων τιμών και "freq" όπου ορίζεται η συχνότητα της ζητούμενης σειράς.

Αρχικά, για την αποφυγή σφαλμάτων ορίζεται η μεταβλητή series "s" που λαμβάνει το εισαγόμενο argument "series", και εφαρμόζεται η μέθοδος "sort_index()" και η μέθοδος "asfreq()". Με αυτόν τον τρόπο γίνεται βέβαιο ότι η εισαγόμενη σειρά έχει την κατάλληλη μορφή, δηλαδή τα δείγματα είναι ανά μια ώρα και εάν υπάρχουν ελλιπείς τιμές λαμβάνουν τιμή NaN. Για να είναι αποδεκτό το argument "freq", ο χρήστης πρέπει να εισάγει κατάλληλη μορφή, δηλαδή: Ακέραιος αριθμός και μορφή μέτρησης χρόνου, όπως 1H για κάθε ώρα ή 1D για κάθε ημέρα.

Εάν ο χρήστης εισάγει απλά τον τύπο μέτρησης χρόνου, για παράδειγμα «H», υπονοώντας μια ώρα δημιουργείται σφάλμα. Για αυτό τον λόγο χρησιμοποιήθηκαν try – except – else statements. Εντός του try block ορίστηκε η μεταβλητή "freq_td" όπου εντός το αλφαριθμητικό argument που έχει εισαχθεί ως "freq" μετατρέπεται σε ένα pandas timedelta, δηλαδή το ζητούμενο χρονικό διάστημα σε κατάλληλη μορφή της μεθόδου "to_timedelta()". Σε περίπτωση που δημιουργηθεί σφάλμα Exception, μέσω χρήσης "H" η παραπλήσιου, εντός της μεθόδου "to_timedelta()" προστίθεται το αλφαριθμητικό "1" με το εισαγόμενο argument του χρήση, για τον υπολογισμό του timedelta.

Στην συνέχεια ορίζεται η μεταβλητή "window_size", εντός της οποίας το argument "days" μετατρέπεται σε timedelta, όπου για παράδειγμα 30 ημέρες μετατρέπονται σε 720 ώρες και διαιρούνται με την ορισμένη συχνότητα της σειράς "freq_td". Ως "valid" ορίζεται η σειρά "s_full" με την εφαρμογή της μεθόδου "notna()" και την μετατροπή των δειγμάτων ως True σε 1 και False σε 0. Στην συνέχεια ορίζεται "valid_count" το οποίο χρησιμοποιεί το προηγουμένως δημιουργημένο διάνυσμα "valid", σε συνδυασμό με την μέθοδο "rolling()". Η συγκεκριμένη μέθοδος δημιουργεί άλληλο-επικαλυπτόμενα παράθυρα ανάλογα με τα εισαγόμενα arguments, όπου "window_size" ορίζει την χρονική διάρκεια του παραθύρου, και επίσης ορίστηκε "min_periods**",** ως "window_size". Η παράμετρος αφορά τον ελάχιστο αριθμό παρατηρήσεων εντός του παραθύρου που έχουν κάποια τιμή ή με άλλον τρόπο, ο ελάχιστος αριθμός παρατηρήσεων που δεν είναι ελλιπής τιμές [85]. Διαφορετικά, το επιστρεφόμενο αποτέλεσμα είναι NaN (np.nan) [85]. Σε κάθε παράθυρο, εφαρμόζεται η μέθοδος "sum()", η οποία θα αθροίζει τις αληθές – έγκυρες τιμές, αφού όπως προαναφέρθηκε μέσω της μετατροπής σε ακέραιες τιμές, της μεθόδου notna() μπορεί να γίνει άθροισμα των έγκυρων τιμών.

Για παράδειγμα, έστω ότι το διάνυσμα "valid" είχε τις εξής τιμές με "window_size" = 3:

$$valid = \begin{matrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 1 & 0 & 1 & 1 \end{matrix}$$

Επομένως, με την χρήση rolling και sum, ισχύει ο παρακάτω πίνακας, με την πρώτη γραμμή να αντιστοιχεί στους δείκτες, η δεύτερη στα παράθυρα και η τρίτη στο άθροισμα των στοιχείων του κάθε παραθύρου:

$$validcount = \begin{matrix} 0 & 1 & 2 & 3 & 4 \\ - & - & [1,1,0] & [1,0,1] & [0,1,1] \\ NaN & NaN & 2 & 2 & 2 \end{matrix}$$

Για τον υπολογισμό του ποσοστού έγκυρων τιμών ορίζεται η μεταβλητή “valid_ratio” που διαιρεί τις ανιχνευμένες έγκυρες τιμές “valid_count” προς το αρχικό μέγεθος του παραθύρου “window_size”. Ορίζεται επίσης η μεταβλητή “ends”, η οποία θα διατηρήσει μόνο τους τελικούς δείκτες, δηλαδή τα timestamps που τελειώνει το παράθυρο, που έχουν ποσοστό έγκυρων εγγραφών (threshold) μεγαλύτερο ή ίσο από το ορισμένο. Έπειτα ορίζεται η λίστα “windows” και η μεταβλητή “delta”, όπου γίνεται αφαίρεση από το εισαγόμενο “days” argument με το εισαγόμενο “freq” argument. Αυτό γίνεται επειδή για να βρεθεί η αρχή του παραθύρου, θα πρέπει να γίνει αφαίρεση της τελικής ημερομηνίας και του timedelta των ζητούμενων ημερών. Όμως με αυτόν τον τρόπο, το παράθυρο θα περιέχει μια έξτρα ώρα. Αφαιρώντας το “freq” argument, το οποίο έστω ότι είναι “1H”, το κάθε παράθυρο θα είναι ακριβώς 720 ώρες – timestamps.

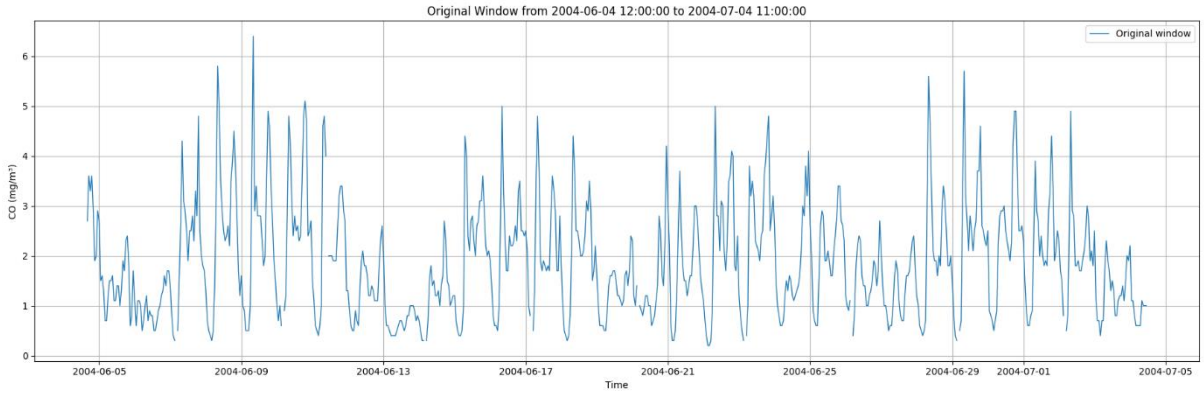
Εφαρμόζοντας την συνάρτηση “find_almost_clean_windows” στην σειρά “s” και arguments days = 30 και threshold = 0.98 (98%), δημιουργούνται τα αποτελέσματα του πίνακα 24.

Πίνακας 24 Πιθανά κατάλληλα (98% πληρότητα) διαστήματα

```
Found 1195 candidate 1-month windows
2004-06-04 12:00:00 to 2004-07-04 11:00:00  | valid_ratio = 0.981
2004-06-04 13:00:00 to 2004-07-04 12:00:00  | valid_ratio = 0.982
2004-06-04 14:00:00 to 2004-07-04 13:00:00  | valid_ratio = 0.983
2004-06-04 15:00:00 to 2004-07-04 14:00:00  | valid_ratio = 0.985
2004-06-04 16:00:00 to 2004-07-04 15:00:00  | valid_ratio = 0.986
2004-06-04 17:00:00 to 2004-07-04 16:00:00  | valid_ratio = 0.986
2004-06-04 18:00:00 to 2004-07-04 17:00:00  | valid_ratio = 0.986
2004-06-04 19:00:00 to 2004-07-04 18:00:00  | valid_ratio = 0.986
2004-06-04 20:00:00 to 2004-07-04 19:00:00  | valid_ratio = 0.986
2004-06-04 21:00:00 to 2004-07-04 20:00:00  | valid_ratio = 0.986
```

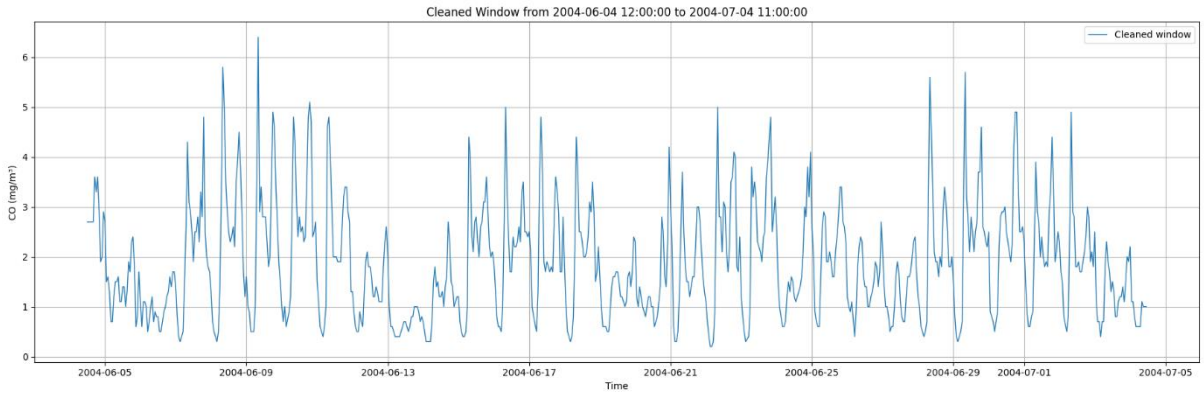
Υπάρχει μεγάλο πλήθος υποψήφιων διαστημάτων (1195) για χρονική διάρκεια ενός μήνα με ποσοστό έγκυρων τιμών 98%. Επιλέχτηκε το πρώτο δείγμα, παρόλο που οι λοιπές εγγραφές είχαν ελαφρώς μεγαλύτερα ποσοστά εγκυρότητας καθώς το ποσοστό ελλιπών τιμών είναι ικανοποιητικό.

Οι δείκτες – timestamps του παραθύρου αποθηκευτήκαν αντίστοιχα ως “start” και “end”. Βάσει αυτών και της μεθόδου “loc” βρέθηκαν στην σειρά “s” και αυτό αποθηκεύτηκε ως “win”. Παρατηρήθηκε ότι το συγκεκριμένο παράθυρο εμφανίζει 14 ελλιπής τιμές – ώρες. Το γράφημα του “win”, δηλαδή του επιλεγμένου διαστήματος παρουσιάζεται στο διάγραμμα 16.



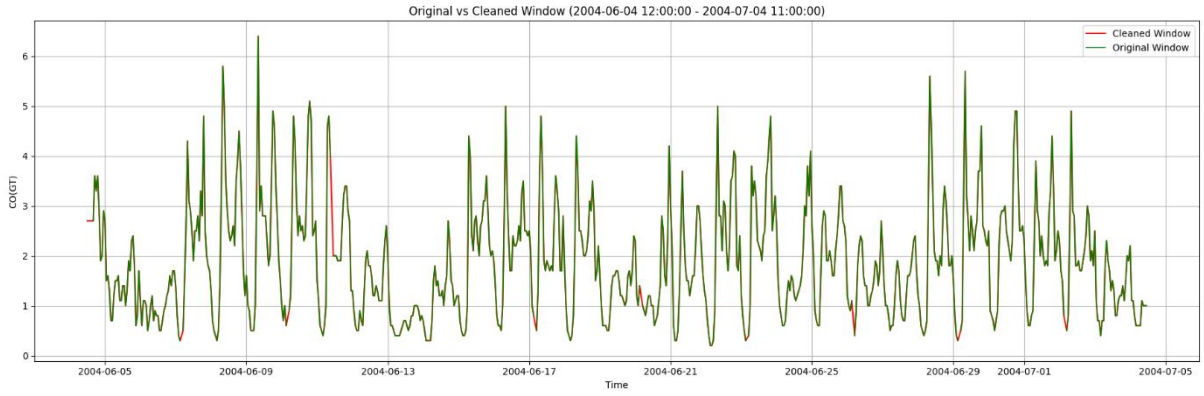
Διάγραμμα 16 Επιλεγμένο διάστημα ενός μήνα πριν από επεξεργασία

Η μέθοδος time interpolation είναι ιδανική για τα αντικατάσταση ελλιπών τιμών για μικρά ή διακριτά διαστήματα ελλιπών τιμών. Ύστερα από την εφαρμογή της μεθόδου, 4 τιμές έμειναν ως ελλιπής και για αυτό τον λόγο εφαρμόστηκε backfill με “limit” ίσο με 4 μέσω της μεθόδου “fillna()”, με το διάγραμμα ύστερα από την επεξεργασία να παρουσιάζεται παρακάτω.



Διάγραμμα 17 Επιλεγμένο διάστημα ενός μήνα ύστερα από επεξεργασία

Πλέον είναι εμφανές ότι δεν υπάρχουν ελλιπής τιμές στο επιλεγμένο διάστημα. Για τον εντοπισμό των διαφορών μεταξύ των διαγραμμάτων μια καλή πρακτική είναι η επικάλυψη (overlay) μεταξύ των 2 διαγραμμάτων με διαφορετικά χρώματα, όπως φαίνεται παρακάτω.



Διάγραμμα 18 Επικάλυψη πραγματικών και εκτιμώμενων τιμών

Πλέον το έγκυρο παράθυρο είναι έτοιμο για επεξεργασία καθώς δεν περιέχει ελλιπείς τιμές. Στην συνέχεια ακολουθεί η εφαρμογή του “artificial missingness”. Όπως αναφέρθηκε και προηγουμένως, καθώς τα κενά διαστήματα, διακριτά και μη, είναι αρκετά μεγάλα στην αρχική χρονοσειρά αποφασίστηκε να χρησιμοποιηθούν τα κενά διαστήματα ύστερα από χρήση backfill μιας εβδομάδας, όπως φαίνονται στον πίνακα 13, τα οποία αποθηκεύτηκαν σε μια λίστα “gap_lengths”. Επειδή τα κενά αφορούν διάστημα ενός χρόνου ενώ για την αξιολόγηση χρησιμοποιείτε διάστημα ενός μήνα, ορίστηκε η μεταβλητή “scale factor” η οποία είναι ίση με 1/12, δηλαδή 0.0833.

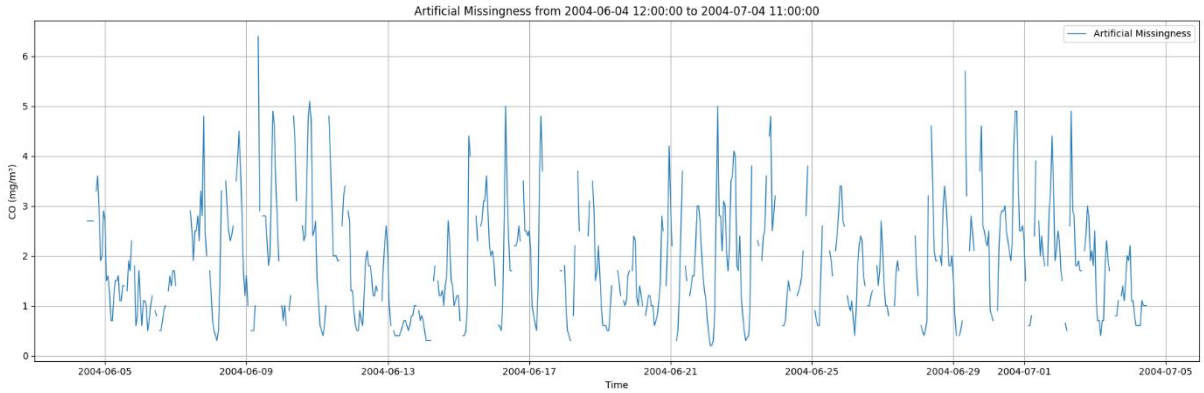
Για την εφαρμογή του scaling, εφαρμόζεται list comprehension, δηλαδή ακολουθεί την μορφή “newlist = [expression for item in iterable if condition == True]”, όπως φαίνεται παρακάτω.

Πίνακας 25 Εφαρμογή scaling

```
gap_lengths_scaled = [
    max(1, int(round(L * scale_factor))) for L in gap_lengths]
```

Αθροίζοντας τα στοιχεία της λίστα, προκύπτει ότι μπορούν να δημιουργηθούν 141 ελλιπής τιμές. Όπως προαναφέρθηκε το επιλεγμένο παράθυρο περιέχει 720 timestamps – ώρες, επομένως 141/720 = 0,195 ή 19,5% του παραθύρου θα είναι ελλιπής τιμές. Η αρχική χρονοσειρά, για το γνώρισμα CO(GT), όπως φαίνεται από τον πίνακα 7, παρουσιάζει 1683 ελλιπής τιμές και αποτελείται από 9356 τιμές, δηλαδή υπάρχουν 1683/9356= 0,179 ή 17,9% ελλιπής τιμές ως προς το σύνολο. Βάσει αυτού, θεωρείται ότι το παράθυρο μπορεί να προσεγγίσει σε ικανοποιητικό βαθμό τις συνθήκες του πραγματικού σετ δεδομένων για την σύγκριση των μεθόδων αντιμετώπισης ελλιπών τιμών.

Για την εφαρμογή των κενών ορίστηκε η συνάρτηση “apply_gaps_simple” με παραμέτρους “series” για την εισαγωγή της σειράς υπό μελέτη, “gap_lengths” για την εισαγωγή της λίστας με των κενών διαστημάτων και “seed” = 42 για μετέπειτα χρήση. Καθώς για την εφαρμογή των κενών θα χρησιμοποιηθεί η βιβλιοθήκη “random”, για να είναι σταθερή η εφαρμογή των κενών και κατ’ επακόλουθο, να μπορεί να γίνει ακριβής σύγκριση μεταξύ τους ορίστηκε rng = np.random.default_rng(seed). Ως “s” ορίζεται το αντίγραφο του εισαγόμενου series και “n” το μήκος του “s”. Για κάθε “L” αριθμό εντός της λίστας των κενών, μέσω χρήσης for loop, ορίζεται τυχαία η αρχή του κενού διαστήματος μεταξύ 0 έως n – L, ώστε να μην δημιουργηθούν σφάλματα “out of range”. Εδώ χρειάζεται να αναφερθεί ότι για να ισχύει το παραπάνω, πρέπει να ισχύει L ≤ n. Όλα τα παραπάνω ανατίθενται στην μεταβλητή “start”, επομένως με την χρήση της μεθόδου “iloc”, το διάστημα μπορεί να οριστεί ως “start” έως “start + L” και τελικά να οριστεί ως NaN. Σημαντικό είναι να αναφερθεί ότι με αυτόν τον τρόπο, μπορούν να δημιουργηθούν μεγαλύτερα κενά από αυτά που περιέχονται στην λίστα λόγω αλληλοεπικάλυψης.



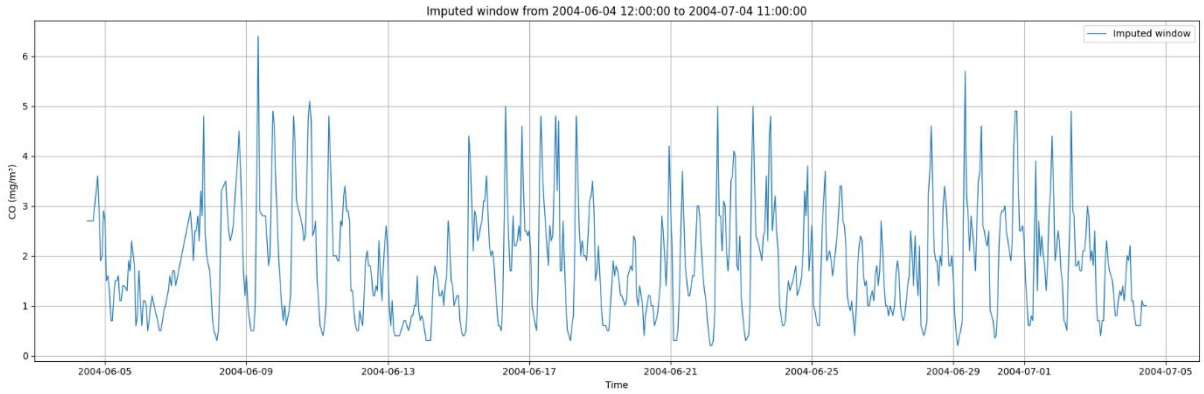
Διάγραμμα 19 Επιλεγμένο διάστημα με εφαρμογή τεχνητών κενών

Το διάγραμμα 19, απεικονίζει το επιλεγμένο παράθυρο ύστερα από την εφαρμογή “artificial missingness”. Τα κενά διαστήματα φαίνεται ότι εφαρμόστηκαν σχετικά ομοιόμορφα και επίσης παρατηρείται ότι λόγω της αλληλο – επικάλυψης τα περισσότερα διαστήματα είναι μεγαλύτερα της μίας ώρας.

3.3.1 1η Μέθοδος

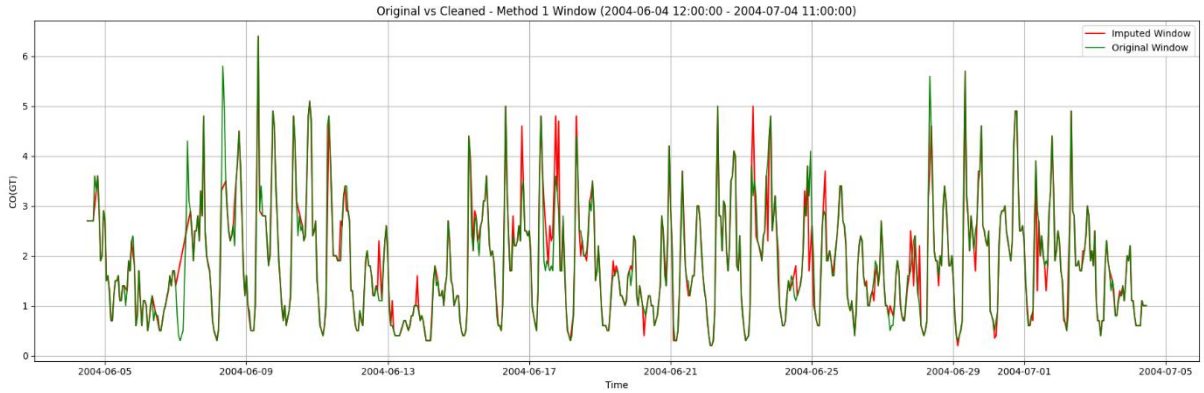
Αρχικά υπενθυμίζεται ότι ανεξάρτητα μεθόδου, αρχικά έχει εφαρμοστεί backfill μίας εβδομάδας, δηλαδή ακολουθείται μορφή «forward fill + method». Οι μέθοδοι βασίζονται στην βιβλιογραφία [6], [18], [17]. Σύμφωνα με τους Belachsen και Broday, [17], για την αξιολόγηση των εφαρμοζόμενων μεθόδων μπορούν να χρησιμοποιηθούν οι μετρικές MAE και R^2.

Η πρώτη μέθοδος αφορούσε την εφαρμογή backfill μίας εβδομάδας και χρήση time interpolation 48 ωρών. Βάσει αυτού ορίστηκε η συνάρτηση “impute_method_1_series” με παραμέτρους “series” για την εισαγωγή της χρονοσειράς, “lag_hours” για τον ορισμό των ωρών – timestamps για το backfill με προεπιλογή 24*7, “interp_limit” για τον ορισμό ωρών με την χρήση time interpolation με προεπιλογή 48 ώρες και “plot” καθώς η συνάρτηση μπορεί να παράγει το διάγραμμα της τροποποιημένης χρονοσειράς ωστόσο ορίζεται ως προεπιλογή False.



Διάγραμμα 20 Επιλεγμένο διάστημα με εφαρμογή 1ης μεθόδου

Το διάγραμμα 20, παρουσιάζει την χρονοσειρά με την χρήση της 1ης μεθόδου. Γενικά, φαίνεται ικανοποιητικό ως αποτέλεσμα ωστόσο είναι πολύ πιο χρήσιμο να γίνει σύγκριση με το αρχικό παράθυρο τιμών καθώς και χρήση μετρικών όπως R^2 και MAE.



Διάγραμμα 21 Σύγκριση 1ης μεθόδου με πραγματικές τιμές

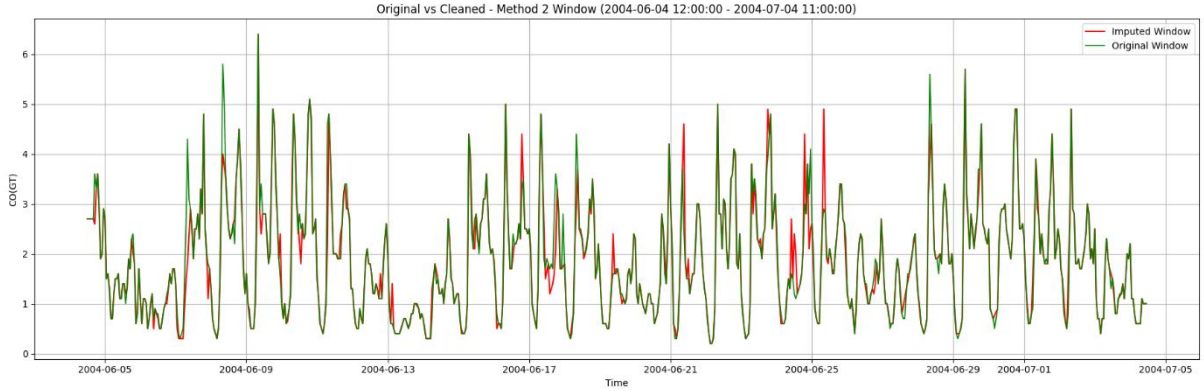
Το διάγραμμα 21, απεικονίζει την αρχική χρονοσειρά και την χρονοσειρά με την χρήση της μεθόδου 1. Μεταξύ 06-05 και 06-11 παρατηρείται απότομες αυξομειώσεις που απέχουν σημαντικά από το αρχικό μοτίβο των πιο ακραίων τιμών, οι οποίες πιθανώς οφείλονται στην χρήση του time interpolation, καθώς σε αυτό το σημείο δεν μπορεί να εφαρμοστεί backfill, καθώς δεν υπάρχουν προηγούμενα δεδομένα εβδομάδας που να μπορούν να χρησιμοποιηθούν. Σε αντίθεση, για το υπόλοιπο διάστημα φαίνεται ότι δημιουργούνται περισσότερες υψηλές τιμές από τις πραγματικές, δηλαδή περισσότερες «κορυφές». Το MAE βρέθηκε ως 0.54 mg/m³ και το R^2 βρέθηκε ως 0.57 ή 57%.

Πίνακας 26 Μετρικές 1ης Μεθόδου

Mean Absolute Error	0.5414634146341464
R - Squared	0.575946783636053

3.3.2 2η μέθοδος

Στην δεύτερη μέθοδο εφαρμόστηκε συνδυασμός forward fill και back fill, δηλαδή χρήση προηγούμενων και επόμενων τιμών σε συνδυασμό με time interpolation 48 ωρών. Το forward fill είχε ήδη εφαρμοστεί επομένως αρκεί να εφαρμοστεί μόνο το backfill και time interpolation. Πρακτικά, η μέθοδος είναι πανομοιότυπη με την προηγούμενη μέθοδο, με την βασική διαφορά να είναι η αλλαγή της προεπιλεγμένης τιμής για την παράμετρο “lag_hours” να ορίζεται σε (-24 x 7), για την λήψη των μετέπειτα τιμών, ανάλογα με το ορισμένο παράθυρο.



Διάγραμμα 22 Σύγκριση 2ης μεθόδου με πραγματικές τιμές

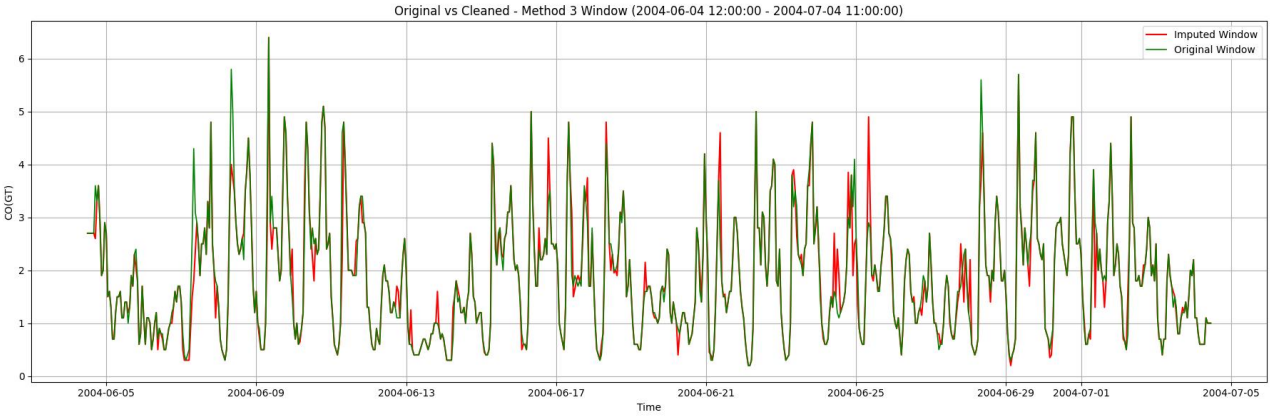
Το διάγραμμα 22, απεικονίζει την αρχική χρονοσειρά και την χρονοσειρά της μεθόδου 2. Φαίνεται ότι επιφέρει καλύτερα αποτελέσματα, καθώς δεν παρατηρούνται σταθερές σταδιακές αλλαγές που υποδηλώνουν χρήση time interpolation ή ιδιαίτερες αποκλίσεις από τις πραγματικές τιμές. Αυτό επιβεβαιώνεται και από τις μετρικές καθώς το MAE βρέθηκε ως 0.48 και το R^2 ως 62%, σημαντικά καλύτερα αποτελέσματα σε σχέση με την πρώτη μέθοδο.

Πίνακας 27 Μετρικές 2ης Μεθόδου

Mean Absolute Error	0.48722222222222217
R - Squared	0.625402902344409

3.3.3 3η Μέθοδος

Η τρίτη μέθοδος αφορά την χρήση συμμετρικών +- «παραθύρων» για την ίδια ημέρα και ώρα, χρησιμοποιώντας την μέση τιμή όπου είναι εφικτή. Σε περίπτωση που δεν μπορεί να υπολογιστεί η ενδιάμεση τιμή, επειδή κάποιο από τα παράθυρα περιέχει τιμή NaN, χρησιμοποιείται time interpolation. Ως παράμετροι έχουν οριστεί “series” για την εισαγωγή της χρονοσειράς, “window_weeks” για τον αριθμό των εβδομάδων που θα χρησιμοποιηθούν στο παράθυρο, “interp_limit” για το μέγιστο μέγεθος του interpolation και “use_fallback” που ορίζεται ως True, ώστε να ενεργοποιηθεί το time interpolation.



Διάγραμμα 23 Σύγκριση 3ης μεθόδου με πραγματικές τιμές

Φαίνεται ότι με αυτήν την μέθοδο δημιουργούνται υψηλότερες τιμές από τις πραγματικές, ύστερα από την στιγμή 06 – 13 - 2004. Το MAE βρέθηκε ως 0.49 και το R^2 ως 62%. Δηλαδή, το MAE παρουσιάζει μια μικρή αύξηση σε σχέση με την προηγούμενη μέθοδο (0.01) και μια πολύ μικρή μείωση (0.002) στο R^2.

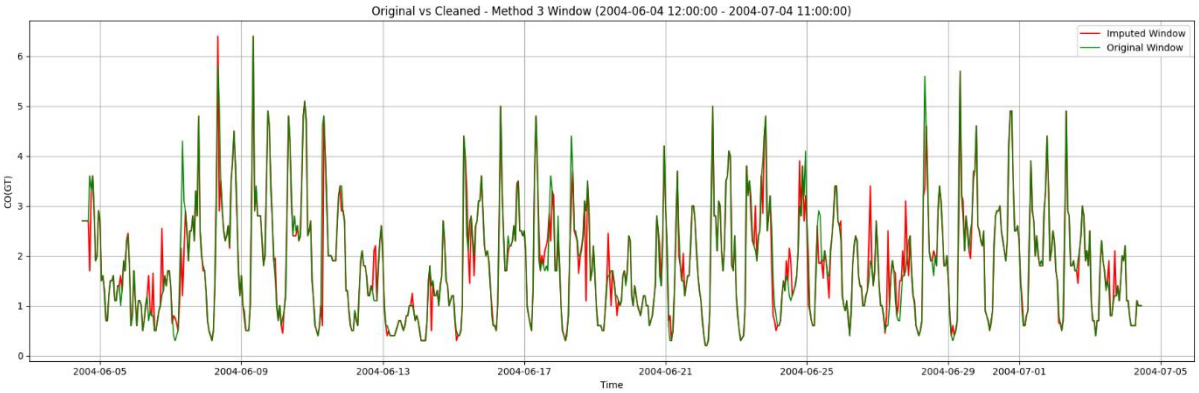
Παράλληλα, έγιναν δοκιμές για μεγαλύτερο παράθυρο ίσο με 2 εβδομάδων, όμως μεγαλύτερο παράθυρο δεν έχει νόημα να χρησιμοποιηθεί καθώς όπως προαναφέρθηκε, είναι συμμετρικό και ένας μήνας περιέχει 4 εβδομάδες επομένως δεν μπορεί να εφαρμοστεί στο συγκεκριμένο παράθυρο, ωστόσο παρατηρήθηκε σημαντική μείωση στις μετρικές απόδοσης.

Πίνακας 28 Μετρικές 3ης Μεθόδου

Mean Absolute Error	0.4994579945799458
R - Squared	0.6232349069893979

3.3.4 4η Μέθοδος

Η τέταρτη μέθοδος, παρόλο που δεν παρουσιάστηκε στο προηγούμενο κεφάλαιο σε αντίθεση με τις υπόλοιπες, αφορά μια παραλλαγή της προηγούμενης μεθόδου. Συγκεκριμένα, είναι η ίδια μέθοδος με την εξαίρεση ότι δεν γίνεται αντιστοίχιση σε ώρα και ημέρα αλλά μόνο σε ώρα. Επειδή δεν χρειάζεται να γίνει αντιστοίχιση σε ίδια ημέρα με αυτήν που παρουσιάζει ελλιπή τιμή, αντί για εβδομάδες ως παράθυρο χρησιμοποιούνται οι ημέρες. Δηλαδή, οι παράμετροι μένουν ίδιες όπως στην προηγούμενη μέθοδο με εξαίρεση το “window_weeks” που πλέον είναι “window_days” και κατ΄ επακόλουθο εντός του υπολογισμού timedelta χρησιμοποιείται argument “days” και όχι “weeks”.



Διάγραμμα 24 Σύγκριση 4ης μεθόδου με πραγματικές τιμές

Μέσω του διαγράμματος 24, παρατηρείται ότι η συγκεκριμένη μέθοδος δεν παρουσιάζει τις ακραίες τιμές σε σχέση με την προηγούμενη μέθοδο. Έγινε δοκιμή διαφορετικού μεγέθους ημερών για σύγκριση των μετρικών ως προς την μέθοδο 4, με τα αποτελέσματα να παρουσιάζονται στον πίνακα 29.

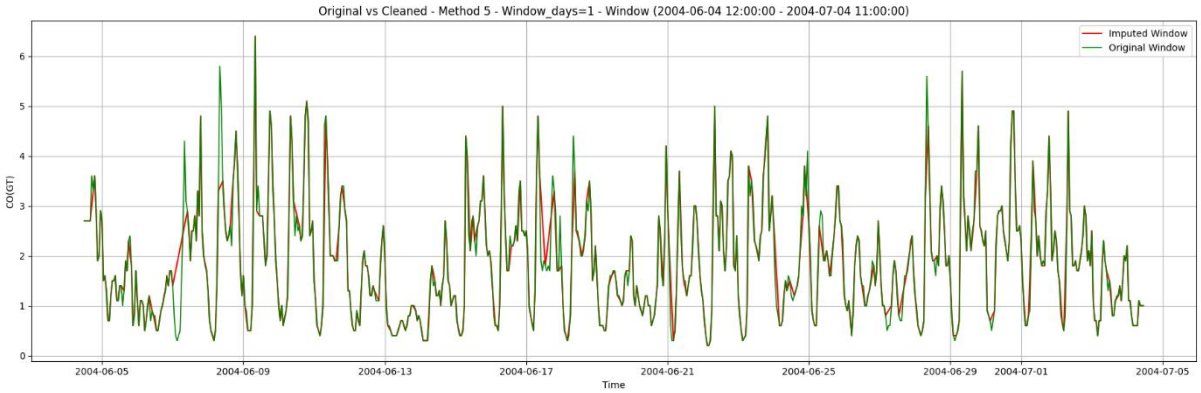
Πίνακας 29 Σύγκριση απόδοσης διαφορετικών παραμέτρων 4ης μεθόδου

Ημέρες – "window_days"	**MAE**	**R^2**
1	0.553968253968254	0.5079086228527024
2	0.6718253968253967	0.3378031394829243
3	0.7051587301587301	0.26608423576354123
7	0.6976190476190475	0.25059652553888534
14	0.6265873015873014	0.43623294065022555
21	0.6480158730158729	0.3763855469076949

Από τον πίνακα 29, φαίνονται τα αποτελέσματα των συγκριτικών μετρικών για διαφορετικό αριθμό ημερών, δηλαδή της μεταβλητής "window_days". Αρχικά, παρατηρείται ότι οι βέλτιστες μετρικές βρέθηκαν για τιμή παράθυρου 1. Για μεγαλύτερες τιμές παραθύρου, δηλαδή 2,3,7, παρατηρείται αύξηση του MAE και μείωση του R^2. Βάσει των 2 παραπάνω, προκύπτει ότι υπάρχει σημαντικό εβδομαδιαίο εποχιακό μοτίβο, αφού η χρήση μεγαλύτερου παραθύρου, χωρίς να γίνεται αντιστοίχιση στην ημέρα υπό μελέτη και κατ' επέκταση αγνοώντας το εβδομαδιαίο επιφέρει σημαντική μείωση στις μετρικές. Κοινώς, αυτό σημαίνει ότι δεν είναι όλες οι κοντινές χρονικές στιγμές εξίσου σημαντικές αλλά μόνο εκείνες που είναι κοντινές και εντός του εβδομαδιαίου μοτίβου. Τέλος, το παραπάνω επιβεβαιώνεται και από τα αποτελέσματα για την χρήση παραθύρου 2 εβδομάδων, όπου παρατηρείται μια μικρή βελτίωση στις μετρικές σε σχέση με τις λοιπές τιμές.

3.3.5 5η Μέθοδος

Ως τελευταία κρίθηκε σκόπιμο να παρουσιαστούν τα αποτελέσματα απλού time interpolation, όχι ως μέθοδο που μπορεί να εκτελεστεί στα πραγματικά δεδομένα, καθώς όπως έχει προαναφερθεί τα διαστήματα των κενών είναι πολύ μεγάλα για να μπορέσουν να καλυφθούν από την συγκεκριμένη μέθοδο αλλά για την σύγκριση των λοιπών μεθόδων με την συγκεκριμένη.



Διάγραμμα 25 Σύγκριση 5ης μεθόδου με πραγματικές τιμές

Το διάγραμμα 25, απεικονίζει το αρχικό γράφημα ως προς το διάγραμμα με την χρήση μεθόδου time interpolation για την εξάλειψη των κενών διαστημάτων και μεμονωμένων τιμών. Απευθείας παρατηρείται ότι η συγκεκριμένη μέθοδος παράγει πολύ θετικά αποτελέσματα, αφού υπάρχουν πολύ μικρές διαφοροποιήσεις μεταξύ των πραγματικών και των εκτιμώμενων τιμών. Το μόνο διαστήματα που δεν παρατηρούνται ικανοποιητικές τιμές αφορά μεταξύ 06-05 και 06-10, όπου παρατηρούνται αλλαγές με σταθερή κλίση, που πιθανώς οφείλονται στην ύπαρξη μεγάλου διαστήματος κενών τιμών.

Πίνακας 30 Μετρικές 5ης Μεθόδου

Mean Absolute Error	0.423130081300813
R - Squared	0.7160444890362961

Η παρατήρηση ότι η συγκεκριμένη μέθοδος φαίνεται να είναι η βέλτιστη όπως φαίνεται στο διάγραμμα 25, επιβεβαιώνεται και από τις μετρικές που παρουσιάζονται στον πίνακα 30, καθώς το R^2 βρέθηκε ως 72% και το MAE ως 0.42 mg/m³.

3.3.6 Σύγκριση μετρικών μεθόδων

Πίνακας 31 Συγκεντρωτικός πίνακας απόδοσης μεθόδων αντιμετώπισης ελλιπών τιμών

Μέθοδος	**MAE - mg/m³**	**R^2**
1 – Forward Fill + Time interpolation	0.5414634146341464	0.575946783636053
2 – Forward Fill + Back Fill + Time interpolation	0.48722222222222217	0.625402902344409
3 – Συμμετρική εβδομαδιαία μέση τιμή ανά ώρα	0.4994579945799458	0.6232349069893979
4 - Συμμετρικός ωριαία μέση τιμή	0.553968253968254	0.5079086228527024
5 – Time Interpolation	0.423130081300813	0.7160444890362961

Ο πίνακας 31, παρουσιάζει συγκεντρωτικά τις μετρικές για κάθε μέθοδο. Η βέλτιστη μέθοδος βρέθηκε ως καθαρό time interpolation με μετρικές MAE 0.42 mg/m³ και R^2 σχεδόν 72%. Ωστόσο, παρόλο που η συγκεκριμένη μέθοδος παρουσιάζει τις καλύτερες μετρικές, δεν μπορεί να χρησιμοποιηθεί όπως έχει στο σετ δεδομένων για τον απλό λόγο ότι τα κενά διαστήματα είναι πολύ μεγαλύτερα από αυτά που παρουσιάστηκαν στο παράδειγμα αξιολόγησης μεθόδων, όπως φαίνονται από το διάγραμμα 8. Βάσει αυτού, μια πιο ρεαλιστική μέθοδος με τις καλύτερες μετρικές για την εφαρμογή στο σετ δεδομένων θεωρείται η μέθοδος 2, η οποία συνδυάζει μεθόδους backfill, forward fill και time interpolation. Η δεύτερη μέθοδος έχει και την δεύτερη θέση ως προς τις υψηλότερες μετρικές αξιολόγησης ενώ παράλληλα θεωρείται ότι βάσει όλων των προαναφερόμενων ευρημάτων η μέθοδος αξιοποιεί το εβδομαδιαίο εποχιακό μοτίβο ενώ παράλληλα είναι εύκολα εξηγήσιμο και μπορεί να εφαρμοστεί χωρίς ιδιαίτερες δυσκολίες στο πραγματικό σετ δεδομένων.

Εδώ χρειάζεται να αναφερθεί ότι υπάρχουν δύο βασικοί λόγοι για τον βαθμό της προσοχής που δόθηκε για την εύρεση της βέλτιστης μεθόδου αντιμετώπισης ελλιπών τιμών και ανακατασκευής του συγκεκριμένου γνωρίσματος. Ο πρώτος λόγος αφορά το ίδιο το γνώρισμα, καθώς όπως προαναφέρθηκε είναι το βασικό γνώρισμα υπό μελέτη και επομένως χρίζει ιδιαίτερη προσοχή ώστε να μην αλλοιωθεί το σήμα του. Ο δεύτερος λόγος είναι ότι σύμφωνα με τις παρατηρήσεις των Niako et al. [30], όπου αναδεικνύουν ότι κάθε μέθοδος αντιμετώπισης ελλιπών τιμών, αυξάνει ή μειώνει την αυτοσυσχέτιση των δεδομένων και επομένως επηρεάζει την ποιότητα των αλγόριθμων πρόβλεψης [30].

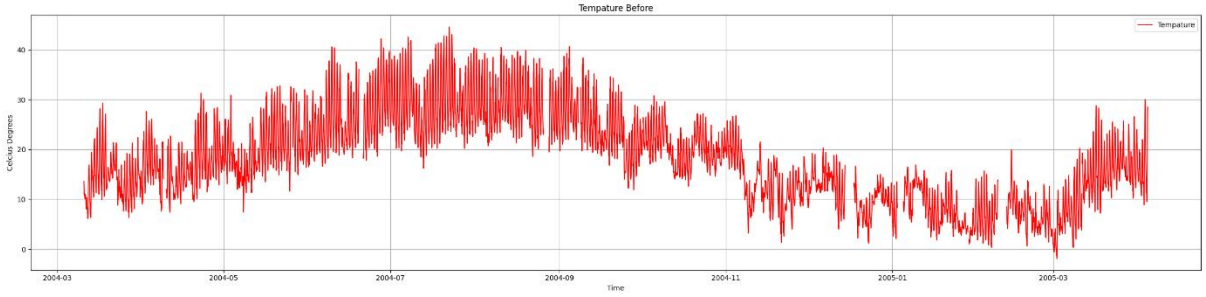
3.3.7 Εφαρμογή σε μετεωρολογικές μεταβλητές

Όπως φάνηκε και στον πίνακα 7, οι μετεωρολογικές μεταβλητές, δηλαδή θερμοκρασία ως “T”, σχετική υγρασία “RH” και απόλυτη υγρασία “AH”, παρουσιάζουν και αυτές ελλιπείς τιμές ωστόσο σε σημαντικά χαμηλότερο μέγεθος από το γνώρισμα “CO(GT)”. Συγκεκριμένα, φαίνεται ότι όλες οι μετεωρολογικές μεταβλητές και τα γνωρίσματα PT08.S5(O3), PT08.S4(NO2), PT08.S3(NOx), PT08.S2(NMHC), C6H6(GT) και PT08.S1(CO) παρουσιάζουν το ίδιο μέγεθος ελλιπών τιμών, δηλαδή 366 ώρες ή αλλιώς περίπου 15 ημέρες. Εστιάζοντας στα μετεωρολογικά γνωρίσματα, θεωρήθηκε σκόπιμο να γίνει έλεγχος στα κενά διαστήματα ώστε να διαπιστωθεί εάν μοιράζονται μεταξύ τους τα κενά διαστήματα. Για την επίτευξη του παραπάνω, χρησιμοποιήθηκε η δημιουργημένη συνάρτηση “detect_nan_gaps()” που παρουσιάστηκε προηγουμένως. Ορίστηκε μια μεταβλητή για κάθε γνώρισμα (“gaps_T”, “gaps_RH”, “gaps_AH”) για την εφαρμογή της συνάρτησης ενώ παράλληλα χρησιμοποιήθηκε η μέθοδος “len()” σε κάθε μια από τις παραπάνω μεταβλητές για την καταμέτρηση όλων των διαστημάτων.

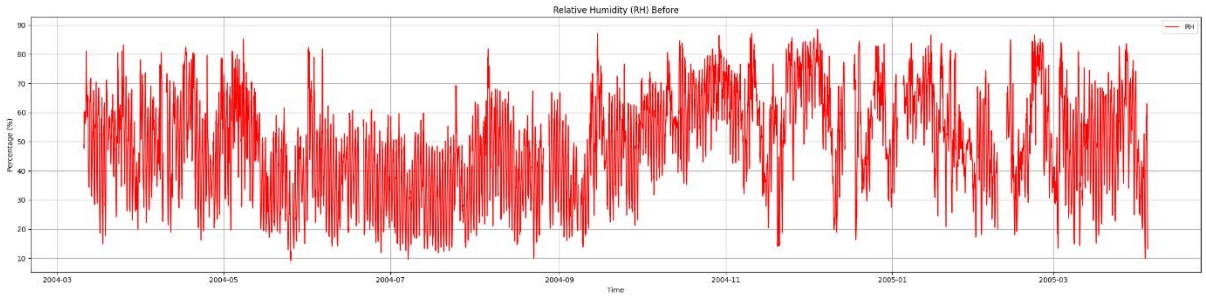
Πίνακας 32 Διαστήματα ελλιπών τιμών για θερμοκρασία, σχετική και απόλυτη υγρασία

```
                  min                 max  duration_hours
T
30 2005-02-08 17:00:00 2005-02-11 20:00:00            76.0
22 2004-12-14 17:00:00 2004-12-17 19:00:00            75.0
26 2005-01-02 21:00:00 2005-01-05 00:00:00            52.0
14 2004-08-26 06:00:00 2004-08-28 02:00:00            45.0
8  2004-06-19 14:00:00 2004-06-21 03:00:00            38.0
4  2004-04-08 23:00:00 2004-04-09 22:00:00            24.0
6  2004-05-25 19:00:00 2004-05-26 08:00:00            14.0
16 2004-09-07 23:00:00 2004-09-08 08:00:00            10.0
28 2005-01-28 17:00:00 2005-01-29 01:00:00             9.0
18 2004-09-08 10:00:00 2004-09-08 17:00:00             8.0

16 discrete gaps have been detected
                  min                 max  duration_hours
RH
30 2005-02-08 17:00:00 2005-02-11 20:00:00            76.0
22 2004-12-14 17:00:00 2004-12-17 19:00:00            75.0
26 2005-01-02 21:00:00 2005-01-05 00:00:00            52.0
14 2004-08-26 06:00:00 2004-08-28 02:00:00            45.0
8  2004-06-19 14:00:00 2004-06-21 03:00:00            38.0
4  2004-04-08 23:00:00 2004-04-09 22:00:00            24.0
6  2004-05-25 19:00:00 2004-05-26 08:00:00            14.0
16 2004-09-07 23:00:00 2004-09-08 08:00:00            10.0
28 2005-01-28 17:00:00 2005-01-29 01:00:00             9.0
18 2004-09-08 10:00:00 2004-09-08 17:00:00             8.0

16 discrete gaps have been detected
                  min                 max  duration_hours
AH
30 2005-02-08 17:00:00 2005-02-11 20:00:00            76.0
22 2004-12-14 17:00:00 2004-12-17 19:00:00            75.0
26 2005-01-02 21:00:00 2005-01-05 00:00:00            52.0
14 2004-08-26 06:00:00 2004-08-28 02:00:00            45.0
8  2004-06-19 14:00:00 2004-06-21 03:00:00            38.0
4  2004-04-08 23:00:00 2004-04-09 22:00:00            24.0
6  2004-05-25 19:00:00 2004-05-26 08:00:00            14.0
16 2004-09-07 23:00:00 2004-09-08 08:00:00            10.0
28 2005-01-28 17:00:00 2005-01-29 01:00:00             9.0
18 2004-09-08 10:00:00 2004-09-08 17:00:00             8.0

16 discrete gaps have been detected
```

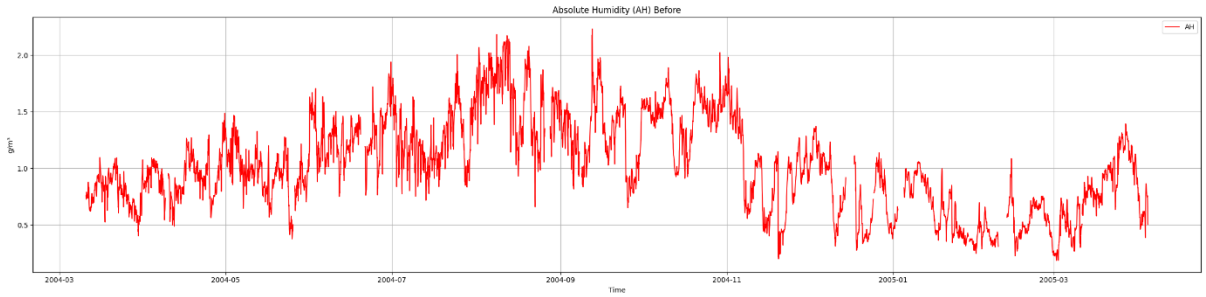
Από τον πίνακα 32, όπου παρουσιάζονται τα κενά διαστήματα για κάθε μετεωρολογική μεταβλητή καθώς και τα 10 μεγαλύτερα διαστήματα κατά φθίνουσα σειρά, διαπιστώνεται ότι πράγματι όλα τα διαστήματα είναι ακριβώς τα ίδια για όλα τα γνωρίσματα. Το μεγαλύτερο διάστημα είναι 76 ώρες ή περίπου 3 ημέρες. Στην συνέχεια δημιουργήθηκαν τα διαγράμματα 26,27,28, των γνωρισμάτων υπό μελέτη, με σκοπό την μελέτη των χαρακτηριστικών τους.



Διάγραμμα 26 Θερμοκρασία πριν από επεξεργασία

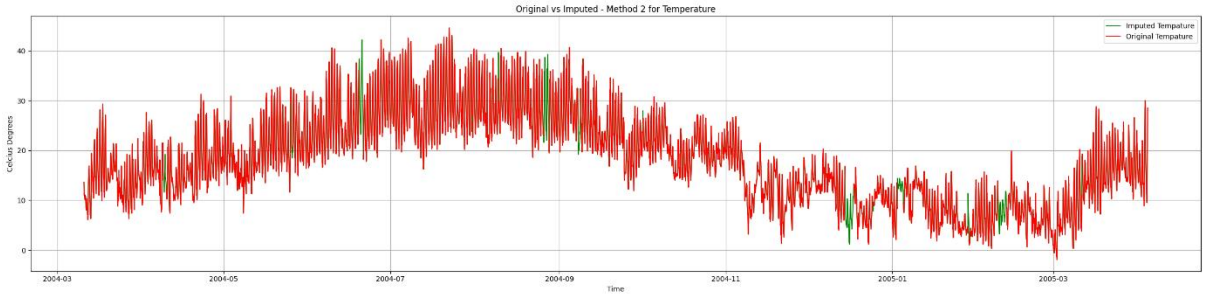


Διάγραμμα 27 Σχετική υγρασία πριν από επεξεργασία

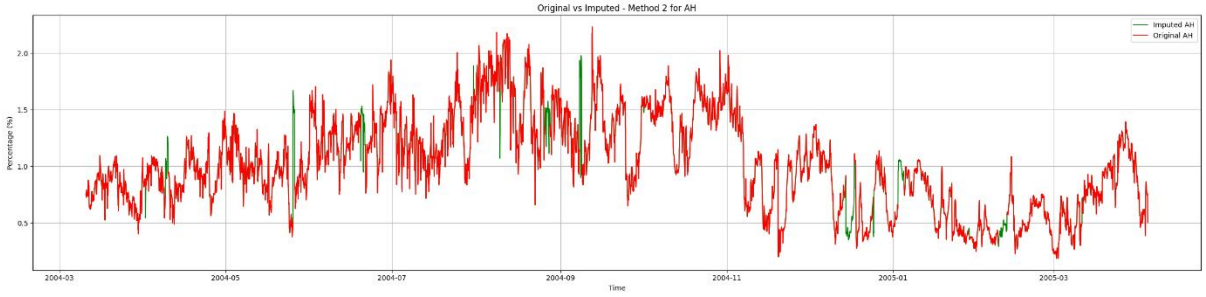


Διάγραμμα 28 Απόλυτη υγρασία πριν από επεξεργασία

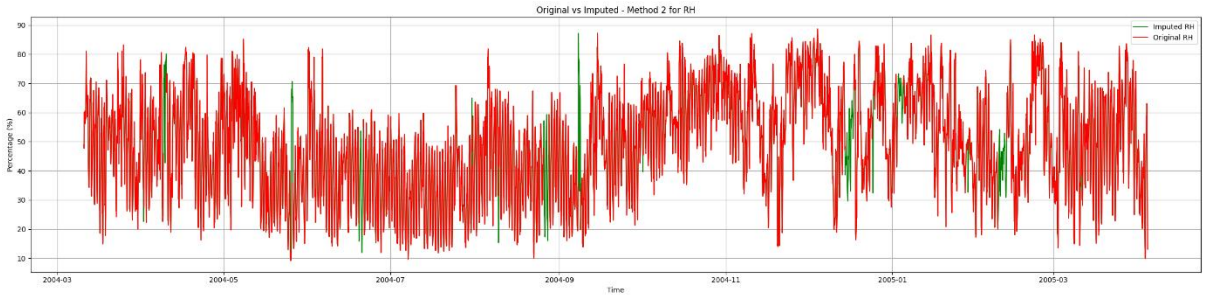
Εδώ είναι αναγκαίο να αναφερθεί μια παραδοχή σχετικά με την αντιμετώπιση των ελλιπών τιμών στα συγκεκριμένα γνωρίσματα. Λόγω του σχετικά μικρού μεγέθους των ελλιπών τιμών δεν θεωρήθηκε αναγκαία η αναλυτική σύγκριση διαφορετικών μεθόδων όπως έγινε για το «βασικό» γνώρισμα “CO(GT). Αντί αυτού έγινε σύγκριση μεταξύ 2 «απλοποιημένων» μεθόδων, τύπου 1 και 2. Δηλαδή, στην προκειμένη περίπτωση δεν εφαρμόζεται το αρχικό forward fill, καθώς στην μέθοδο 1 χρησιμοποιείται forward fill των προηγούμενων τιμών, διάρκειας μίας εβδομάδας και χρήση time interpolation σε περίπτωση που δεν επικαλυφθούν όλα τα κενά. Η μέθοδος 2 χρησιμοποιεί back fill, δηλαδή χρήση των επόμενων τιμών διάρκειας μιας εβδομάδας και χρήση time interpolation όπως στην η μέθοδο. Πρακτικά, χρησιμοποιείται η ίδια συνάρτηση που χρησιμοποιήθηκε στο προηγούμενο κεφάλαιο “impute_method_2()” με την βασική διαφορά μεταξύ τους να είναι η τιμή της παραμέτρου “lag_hours” ανάλογα το πρόσημο που χρησιμοποιείται και προφανώς η παράμετρος “column” για τον ορισμό του γνωρίσματος υπό μελέτη. Έτσι, επιλέχτηκε να εφαρμοστεί η μέθοδος που παρήγαγε τις καλύτερες μετρικές στο προηγούμενο κεφάλαιο, δηλαδή η μέθοδος 2, με τα τελικά διαγράμματα να παρατίθενται παρακάτω.



Διάγραμμα 29 Θερμοκρασία ύστερα από χρήση 2ης μεθόδου



Διάγραμμα 30 Απόλυτη υγρασία ύστερα από χρήση 2ης μεθόδου



Διάγραμμα 31 Σχετική υγρασία ύστερα από χρήση 2ης μεθόδου

Τέλος, έγινε ένας τελευταίος έλεγχος ελλιπών τιμών μέσω της συνάρτησης “detect_nan_gaps()” για τις προηγούμενα αναφερόμενες μεταβλητές, όπου εντοπίστηκε ότι πλέον δεν υπάρχει καμία ελλιπής τιμή στα γνωρίσματα που θα χρησιμοποιηθούν. Σε αυτό το στάδιο, μπορεί να οριστεί το τελικό σετ δεδομένων που θα χρησιμοποιηθεί με τις μεθόδους πρόβλεψης για την παραγωγή αποτελεσμάτων. Για αυτόν τον λόγο αφαιρέθηκαν όλα τα λοιπά γνωρίσματα που δεν πρόκειται να χρησιμοποιηθούν. Συγκεκριμένα δημιουργήθηκε η λίστα “dropped_variables” όπως εντός εισάχθηκαν τα αλφαριθμητικά «'PT08.S1(CO)', 'C6H6(GT)', 'PT08.S2(NMHC)', 'NOx(GT)', 'PT08.S3(NOx)', 'NO2(GT)', 'PT08.S4(NO2)', 'PT08.S5(O3)'». Επιπλέον, είναι σκόπιμη η μετατροπή των κατηγορικών μεταβλητών, δηλαδή των γνωρισμάτων “Date” και “Month”. Ωστόσο, στην συγκεκριμένη περίπτωση δεν είναι αναγκαία η χρήση κάποιας μέθοδού “encoding”, καθώς μέσω της βιβλιοθήκης “dt” και του γνωρίσματος “Datetime_rep”, υπάρχει πρόσβαση στα attributes “month” και “day” που ουσιαστικά μετατρέπουν τα γνωρίσματα σε αριθμητικά στοιχεία ή πιο συγκεκριμένα σε ποσοτικές μεταβλητές.

3.4 Κυκλική κωδικοποίηση χρονικών γνωρισμάτων

Στην συνέχεια, εφαρμόστηκε η μέθοδος “cyclical encoding” ή αλλιώς “κυκλικής κωδικοποίησης” [19, 31, 70, 71, 72] στα γνωρίσματα που παρουσιάζουν «κυκλικό» τύπο, όπως η ώρες, οι μέρες της βδομάδας ή οι μήνες. Το “cyclical encoding” εφαρμόζεται συχνά σε δεδομένα μορφής χρονοσειράς, καθώς σε αυτά τα δεδομένα παρατηρούνται με μεγαλύτερη συχνότητα γνωρίσματα «κυκλικής» φύσης – μορφής [60].

Ως γνωρίσματα «κυκλικής» φύσης, θεωρούνται αυτά όπου η σχέση των τιμών τους δεν είναι ορθό να γίνει αναπαράσταση ως γραμμική αλλά ως τριγωνομετρικός κύκλος. Πιο αναλυτικά, τα πιο συχνά εμφανιζόμενα κυκλικά γνωρίσματα είναι η ώρα (23:00 προς 0:00), οι μέρες της εβδομάδας (Κυριακή προς Δευτέρα), η κατεύθυνσή του ανέμου (359° προς 0°) ή οι μήνες (Δεκέμβριο προς Ιανουάριο). Το πρόβλημα που υφίσταται με όλα τα παραπάνω παραδείγματα είναι ότι δημιουργείται σημαντική διαφορά μεταξύ της αρχικής και τελικής τιμής ή διαφορετικά ελάχιστης και μέγιστης δυνατής τιμής, ενώ στην πραγματικότητα είναι παραπλήσιες τιμές. Εστιάζοντας στο παράδειγμα της ώρας, εάν εφαρμοστεί μια γραμμική κωδικοποίηση (encoding), ισχύει ο παρακάτω πίνακας 33 [19]:

Πίνακας 33 Παράδειγμα γραμμικής κωδικοποίησης ώρας

Ώρα	**Encoded Τιμή**
0:00	0
1:00	1
…	…
23:00	23

Η παραπάνω κωδικοποίηση - encoding, παρουσιάζει διάφορα προβλήματα όπως προαναφέρθηκε. Συγκεκριμένα, το πρώτο πρόβλημα είναι παρόλο που οι ώρες 23:00 και 0:00 απέχουν στην πραγματικότητα 1 ώρα, ένα μοντέλο θα αντιλαμβανόταν ότι απέχουν 23 ώρες. Επακόλουθο του παραπάνω είναι ότι δημιουργούνται λανθασμένες σχέσεις – συσχετίσεις μεταξύ των τιμών ή αγνοούνται άλλες σχέσεις , όπου για παράδειγμα εάν γίνει εστίαση στα μεσάνυχτα (0:00), η όγδοη ώρα (20:00) θα θεωρηθεί πιο σημαντική από την πρώτη πρωινή ώρα (1:00), λόγω απόστασης. Τέλος, το σήμα υφίσταται αλλοίωση στο μοτίβο του, για παράδειγμα έστω ένα σήμα που έχει ημερήσια εποχικότητα, όπως φαίνεται να ισχύει και στην περίπτωση του CO(GT). Με βάσει την κωδικοποίηση που έχει γίνει, το σήμα δεν θα ήταν ενιαίο πλέον αλλά θα διαχωριζόταν τεχνητά τα μεσάνυχτα, αφού 23:59 και 0:00 δεν θεωρούνται παραπλήσιες τιμές. Για την εφαρμογή του “cyclical encoding” σε ένα γνώρισμα, είναι απαραίτητη η χρήση δύο μετασχηματισμών: Ημίτονο (Sine) και Συνημίτονο (Cosine), όπως παρουσιάζονται παρακάτω στον πίνακα 34.

Πίνακας 34 Σχέσεις μετασχηματισμού ημίτονου και συνημίτονου

$$sine = \sin\left(\frac{2\pi * x}{T}\right)$$	(1)
$$cosine = \cos\left(\frac{2\pi * x}{T}\right)$$	(2)

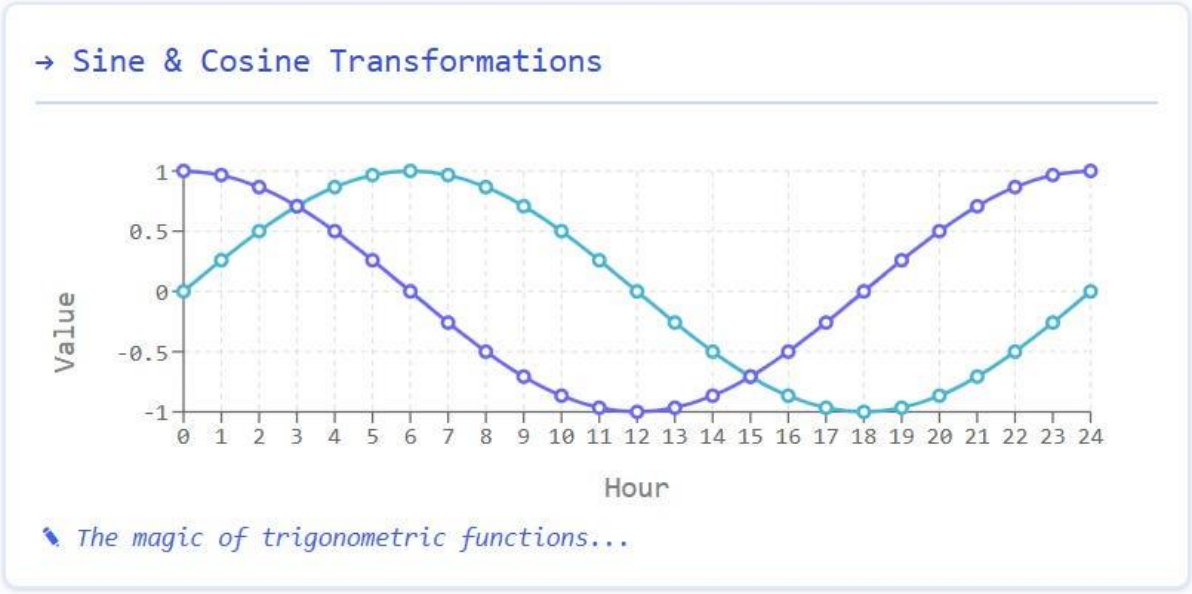
Όπου x, η μεταβλητή υπό μελέτη με σκοπό το cyclical encoding και T η χρονική μονάδα – συχνότητα της μεταβλητής υπό μελέτη.

Σε επίπεδο κώδικα, για κάθε μεταβλητή στην οποία θα εφαρμοστεί cyclical encoding, δηλαδή τα γνωρίσματα “Hour” – ώρα, “Weekday_num” - Ημέρα της εβδομάδας και “month_num” – μήνας, θα δημιουργηθούν δύο νέα γνωρίσματα. Για παράδειγμα, για το γνώρισμα “Hour”, ορίζονται τα νέα γνωρίσματα “hour_sin” και “hour_cos”, και εφαρμόζονται αντίστοιχα οι τύποι τους, όπως φαίνεται από τον πίνακα 34, με το T να ορίζεται ως 24, αφού αυτή είναι η μονάδα μέτρησης. Ο πίνακας 35 παρουσιάζει την υλοποίηση των προαναφερόμενων.

Πίνακας 35 Παράδειγμα υλοποίησης κυκλικής κωδικοποίησης

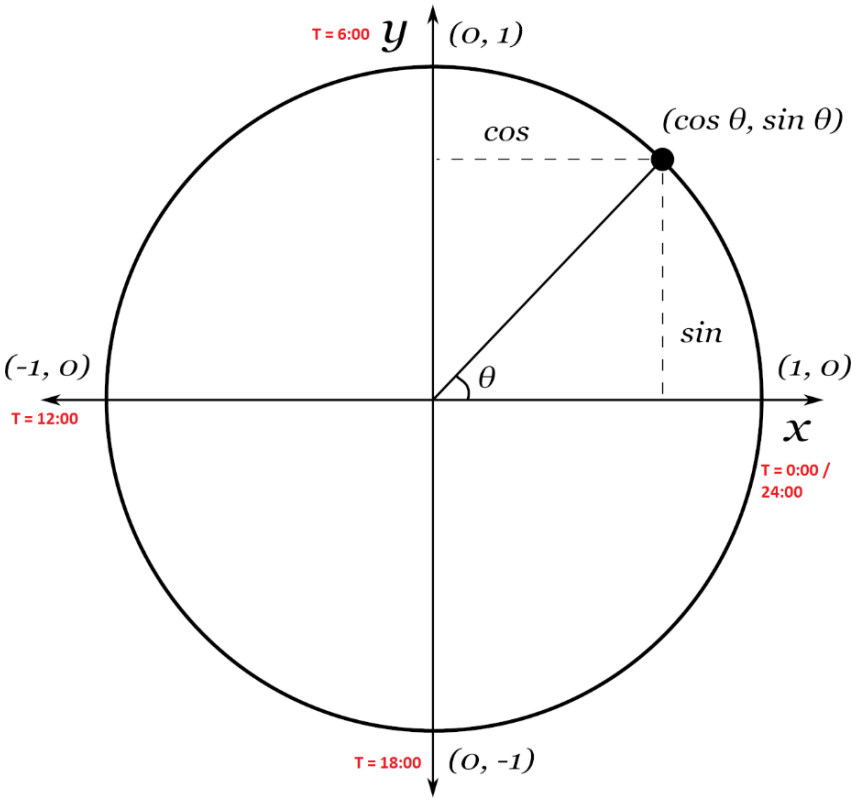
```
final_df["hour_sin"] = np.sin(2 * np.pi * final_df["Hour"] / 24)
final_df["hour_cos"] = np.cos(2 * np.pi * final_df["Hour"] / 24)
```

Το διάγραμμα 32 [19], απεικονίζει πως οι γραφικές παραστάσεις του ημίτονου και του συνημίτονου αποδίδουν την ώρα συνδυαστικά.



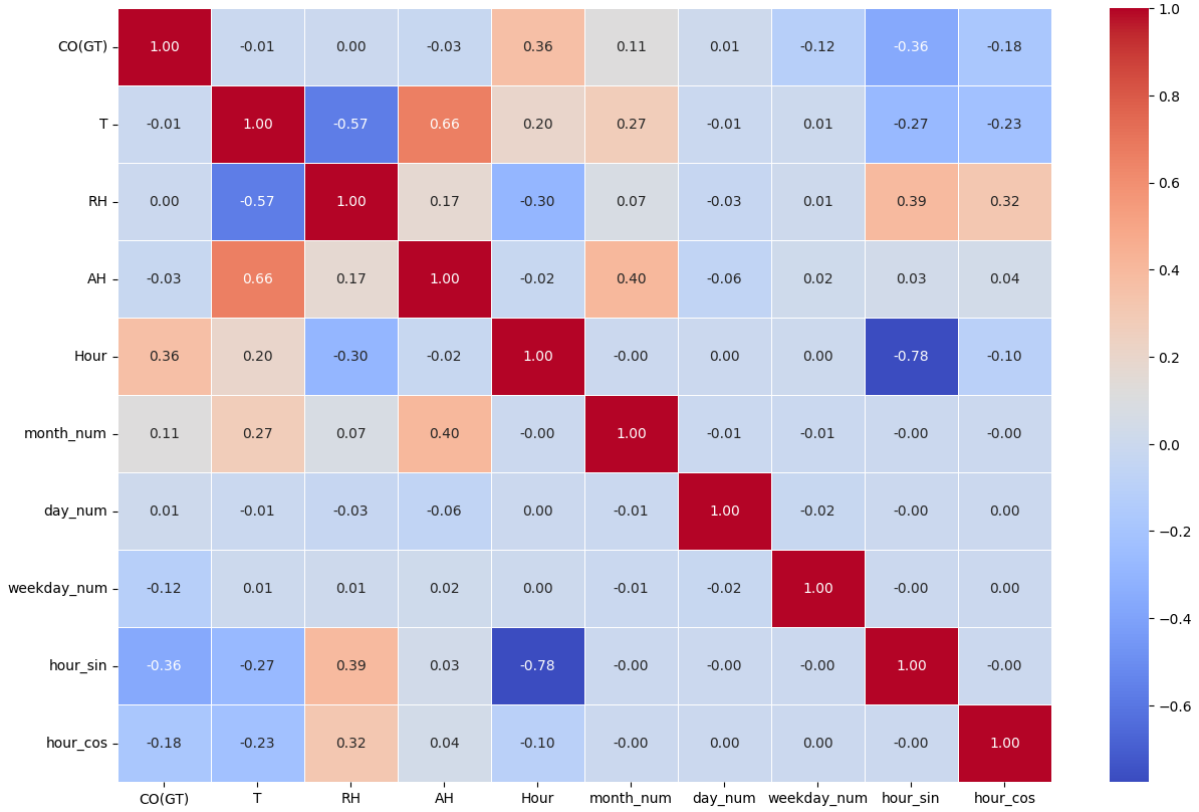
Διάγραμμα 32 Γραφικό παράδειγμα κυκλικής κωδικοποίησης. Πηγή: [19]

Στο διάγραμμα 32, για παράδειγμα μπορεί να παρατηρηθεί ότι για τις 12:00, η κωδικοποίηση θα είναι συνημίτονο -1 και ημίτονο 0. Πέρα από τις γραφικές παραστάσεις, θεωρείται χρήσιμο να αποδοθεί και ένα γράφημα του μοναδιαίου κύκλου σε συνδυασμό με τις κυκλικά κωδικοποιημένες ώρες, το οποίο παρουσιάζεται παρακάτω ως διάγραμμά 33.



Διάγραμμα 33 Απόδοση κωδικοποίησης στον μοναδιαίο κύκλο. Πηγή: [20,21]

Στο διάγραμμα 33, εντοπίζεται ότι τα τεταρτημόρια διαχωρίζονται ανά 6 ώρες και η κάθε ώρα έχει αντιστοιχηθεί σε έναν συνδυασμό τιμών ημιτόνου και συνημίτονου. Με βάση αυτήν την επεξεργασία, τα μεσάνυχτα (24:00) πλέον έχουν πράγματι παραπλήσια «θέση» με τις πρωινές ώρες όπως την μία το πρωί (1:00). Στην συνέχεια, θεωρήθηκε σκόπιμο να δημιουργηθεί ένα διάγραμμα συσχέτισης (correlation matrix), για τις ποσοτικές μεταβλητές και παρουσιάζεται στο διάγραμμα 34, παρακάτω.



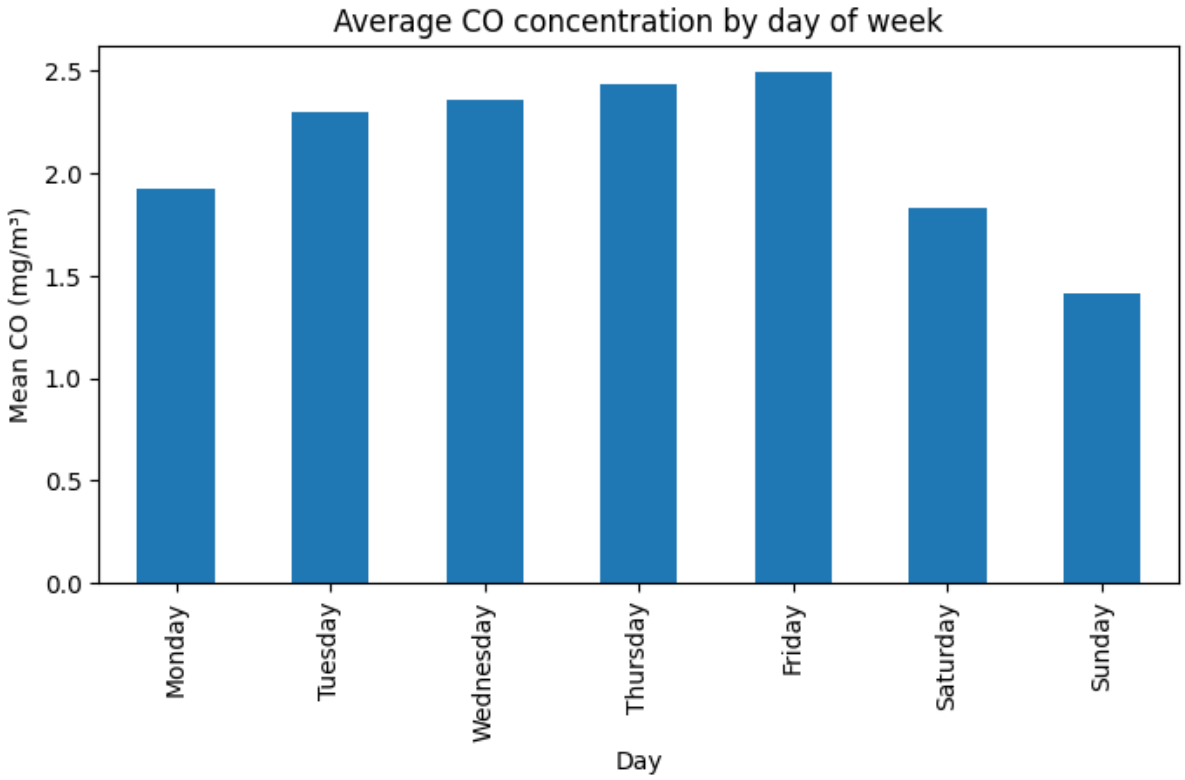
Διάγραμμα 34 Πίνακας συσχέτισης

Αρχικά, παρατηρείται ότι το γνώρισμα της ώρας παρουσιάζει θετική μέτρια - ασθενή σχέση (0,36) με το γνώρισμα υπό μελέτη CO(GT). Παράλληλα το κωδικοποιημένο γνώρισμα του μήνα εμφανίζει μια ασθενής θετική συσχέτιση (0,11). Τέλος, παρατηρείται μια αρνητική ασθενής συσχέτιση (-0,12) με την κωδικοποιημένη ημέρα της εβδομάδας.

Δεν έχει νόημα να χρησιμοποιηθούν οι κυκλικά κωδικοποιημένες μεταβλητές αφού εξ' ορισμού η σχέση τους δεν μπορεί να είναι γραμμική ενώ παράλληλα έχει νόημα να χρησιμοποιηθούν συνδυαστικά και όχι μεμονωμένα ως προς την τιμή στόχο. Για λόγους παραδείγματος, εισάχθηκαν στο διάγραμμα συσχέτισης.

Για τις κατηγορικές μεταβλητές, όπως την ημέρα της εβδομάδας ή τον μήνα, αφού η αριθμητική κωδικοποίηση ώστε να είναι σε κλίμακα θα όριζε μια γραμμική δομή που δεν υπάρχει στην πραγματικότητα .Επομένως θεωρήθηκε σκόπιμο να εξεταστεί η συσχέτισή τους με την τιμή στόχο CO(GT), μέσω ομαδοποιημένων γραφημάτων περιγραφικής στατιστικής.

Για την κατασκευή του διαγράμματος 35, πρώτα δημιουργήθηκε η λίστα "weekday_order" η οποία χρησιμοποιήθηκε για είναι εφικτό το διάγραμμα να παρουσιάζει την σωστή σειρά των ημερών και εντός εισάχθηκαν τα αλφαριθμητικά στοιχεία για κάθε ημέρα σε σειρά, δηλαδή από «Monday» έως «Sunday». Ύστερα ορίστηκε η μεταβλητή "day_group", εντός της οποίας εφαρμόστηκε η μέθοδος "groupby()" βάσει του γνωρίσματος "Day". Στην συνέχεια χρησιμοποιήθηκε η μέθοδος mean() για το γνώρισμα CO(GT) συνδυαστικά με την μέθοδο "reindex()" με argument την προαναφερόμενη λίστα "weekday_order".

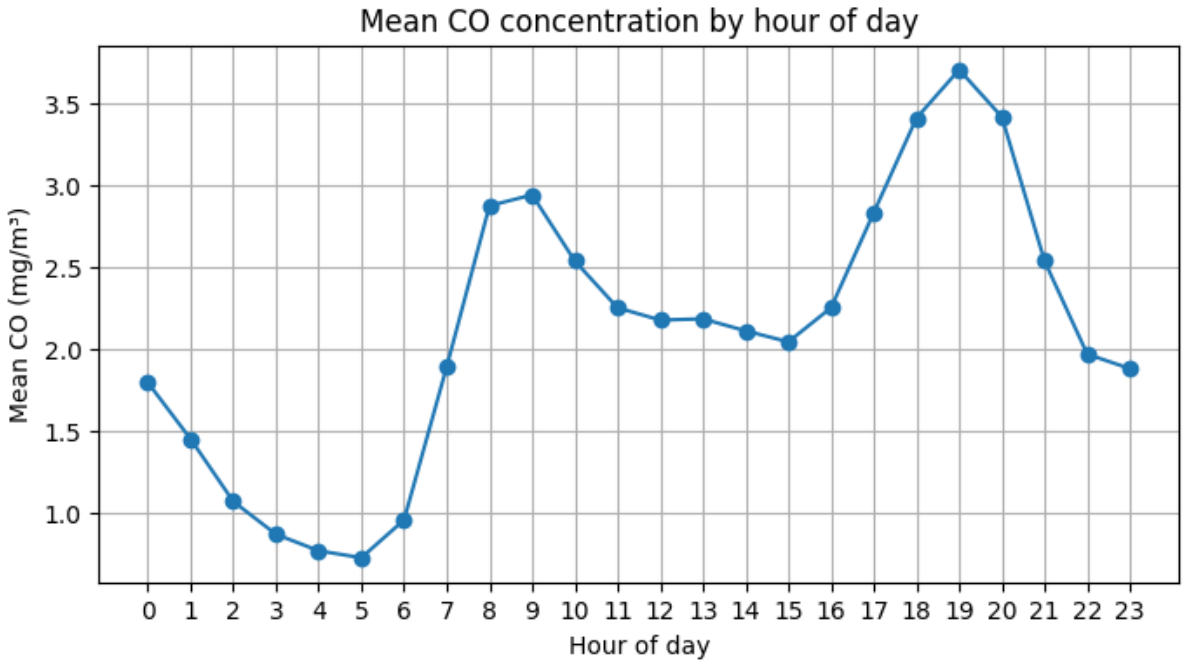


Διάγραμμα 35 Μέση συγκέντρωση CO ανά ημέρα της εβδομάδας

Το διάγραμμα 35, παρουσιάζει την μέση τιμή CO(GT) ομαδοποιημένη ανά ημέρα της εβδομάδας. Αρχικά, παρατηρείται μια μικρή σταδιακή αύξηση σε κάθε ημέρα από την Τρίτη με κορύφωση της τιμής την Παρασκευή. Στην συνέχεια παρατηρείται απότομη μείωση για το Σάββατο και η χαμηλότερη μέση τιμή παρατηρείται τις Κυριακές. Τέλος, η Δευτέρα παρουσιάζει αύξηση ωστόσο το μέγεθος της μέσης τιμής φαίνεται είναι παρόμοιο με αυτό του Σαββάτου και όχι της Τρίτης.

Όλα τα παραπάνω φαίνεται να συνάδουν με την βιβλιογραφία [23,24,25] ότι τα μεγέθη CO είναι άμεσα συνδεδεμένα με ανθρωπογενείς δραστηριότητες και συγκεκριμένα με την μετακίνηση των ανθρώπων μέσω των διαφόρων μεταφορικών τους μέσων που χρησιμοποιούν μηχανές εσωτερικής καύσης πετρελαίου ή βενζίνης. Συγκεκριμένα, θεωρείται ότι η υψηλότερη τιμή CO παρατηρείται την Παρασκευή επειδή ένας μεγαλύτερο αριθμός ατόμων βγαίνει για διασκέδαση σε συνδυασμό με το ήδη υπάρχον πλήθος των ατόμων που χρησιμοποιούν τα μέσα μεταφοράς τους κατά την διάρκεια της εβδομάδας. Αυτό μπορεί να συνεχίζεται και για το Σάββατο όμως φαίνεται ότι τις Κυριακές, οι περισσότεροι ξεκουράζονται επομένως δεν χρησιμοποιούν τα μεταφορικά τους μέσα. Η σημαντικά χαμηλότερη μέση τιμή της Δευτέρας μπορεί να οφείλεται είτε στο φαινόμενο “weekend lag” είτε στο γεγονός ότι αρκετά καταστήματα είναι ανοιχτά και τα Σάββατα, επομένως οι εργαζόμενοι μπορεί να ξεκουράζονται Κυριακές και Δευτέρες.

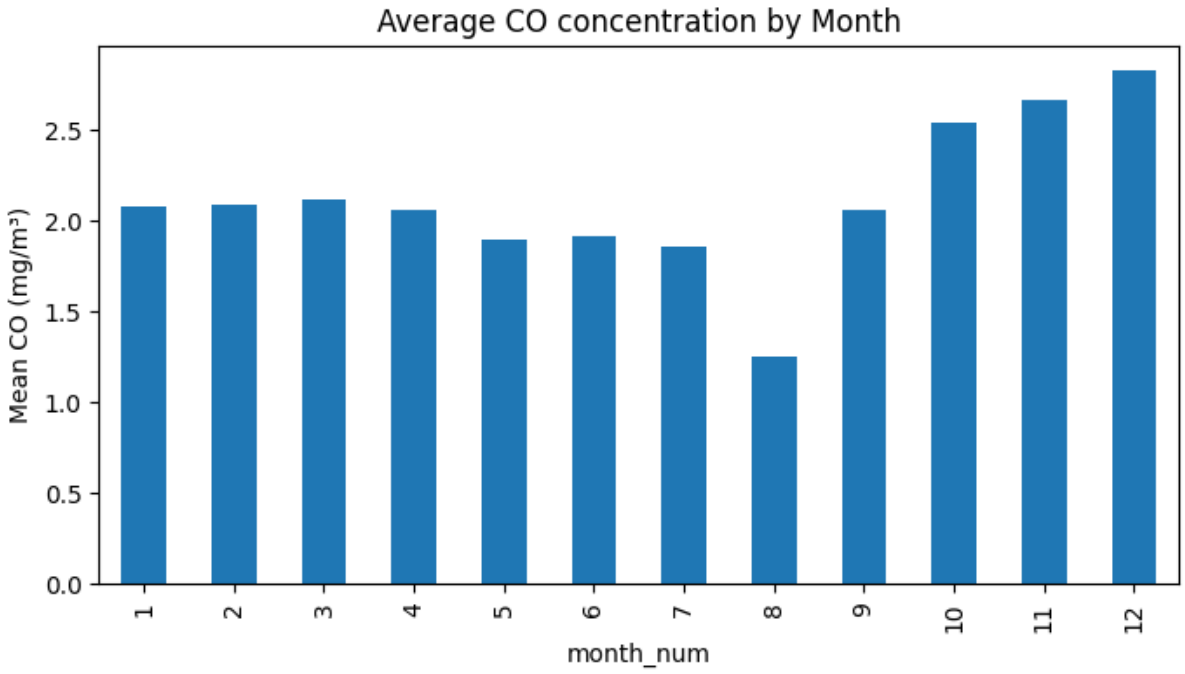
Ένας περαιτέρω τρόπος για να επιβεβαιωθούν πλήρως οι παραπάνω παρατηρήσεις είναι να δημιουργηθούν οι μέσες τιμές ανά ώρα για το σύνολο του σετ δεδομένων και να ελεγχθεί εάν υπάρχει ημερήσιος κύκλος με 2 κορυφές, τις πρωινές και απογευματινές ώρες.



Διάγραμμα 36 Μέση συγκέντρωση CO ανά ώρα της ημέρας

Το γράφημα 36, φαίνεται ότι επιβεβαιώνει τις παραπάνω υποθέσεις. Επιβεβαιώνεται το ημερήσιο μοτίβο, όπου από τις δώδεκα το βράδυ μέχρι τις πέντε το πρωί, όπου παρατηρείται και η ελάχιστη μέση τιμή. Από τις 5 μέχρι τις 9 το πρωί υπάρχει μια απότομη σταθερή αύξηση που φαίνεται να αντικατοπτρίζει την χρήση των μεταφορικών μέσων ώστε τα άτομα να μετακινηθούν στην εργασία τους. Από τις 9 μέχρι τις 3 φαίνεται μια σχεδόν σταθερή τάση, η οποία μπορεί να οφείλεται στο ότι τα περισσότερα βρίσκονται στην εργασία τους. Από τις 3 μέχρι 19 φαίνεται μια σταθερή αύξηση στην μέση τιμή CO, η οποία πιθανώς είναι η κίνηση των ατόμων είτε προς το σπίτι τους είτε προς άλλους προορισμούς.

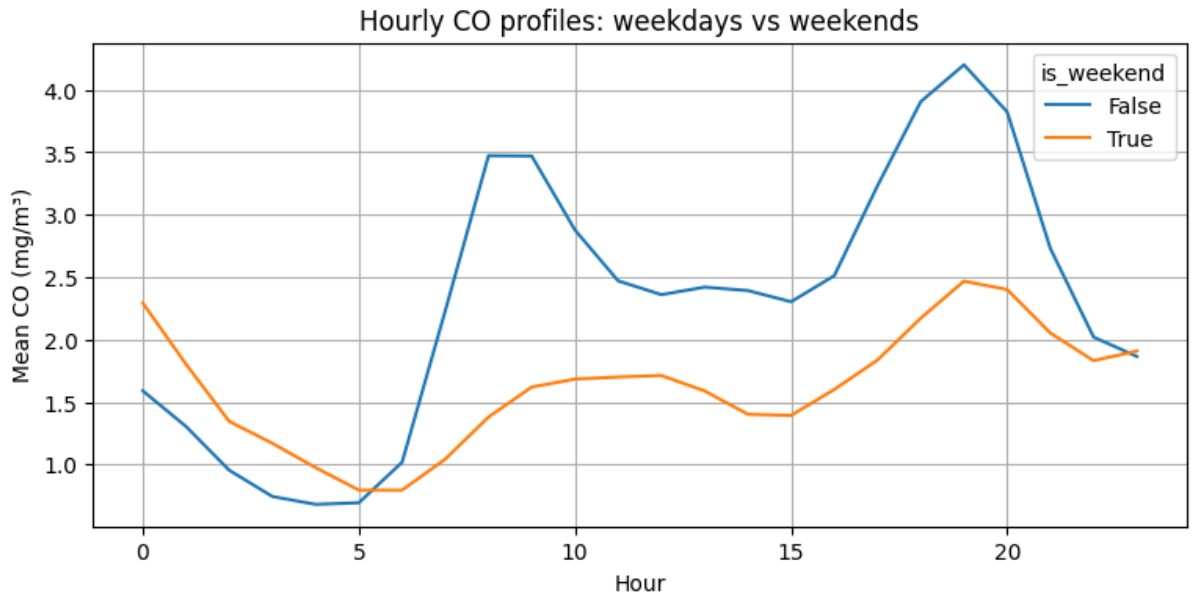
Είναι σημαντικό να παρατηρηθεί ότι ο παρατηρούμενος βραδινός «όγκος» είναι αρκετά μεγαλύτερος από τον παρατηρούμενο πρωινό όγκο. Αυτό πιθανώς να υποδηλώνει το συνδυαστικό φορτίο των ατόμων που επιστρέφουν από την εργασία τους συν των ατόμων που βγήκαν αποκλειστικά για άλλους σκοπούς, για παράδειγμα για διασκέδαση. Τέλος, παρατηρείται μια σχετικά απότομη μείωση των τιμών μέχρι τις 23, έτσι ώστε να ολοκληρώσουν τον κύκλο του μεγέθους των δεδομένων.



Διάγραμμα 37 Μέση συγκέντρωση CO ανά μήνα

Πέρα από το ημερήσιο μοτίβο, θεωρήθηκε ενδιαφέρον να ερευνηθεί ένα υπάρχει και κάποιο μηνιαίο μοτίβο. Συγκεκριμένα, η αρχική υπόθεση ήταν τους χειμερινούς μήνες θα υπάρχουν αυξημένες τιμές CO, καθώς περισσότερα άτομα θα προτιμήσουν να κινηθούν με τα οχήματα τους κυρίως λόγω πιο έντονων καιρικών συνθηκών όπως βροχή ή κρύο.

Από το γράφημα 37, φαίνεται να επιβεβαιώνεται η παραπάνω υπόθεση αφού ξεκινώντας από τον Σεπτέμβριο, τον 9^{ο} μήνα παρατηρείται μια σημαντική αύξηση της μέσης τιμής CO, με την υψηλότερη τιμή να παρατηρείται για τον Δεκέμβριο τον 12^{ο} μήνα δηλαδή. Ωστόσο, παρατηρείται μια απότομη μείωση από τον Δεκέμβριο προς τον Ιανουάριο ύστερα από την οποία η μέση τιμή παραμένει σχεδόν σταθερή για τον Ιανουάριο μέχρι τον Απρίλιο. Στην συνέχεια, παρατηρείται μια ακόμα μείωση από τον Απρίλιο στον Μάϊο με την τιμή πάλι να είναι σχεδόν σταθερή μέχρι τον Ιούλιο. Τέλος, τον Αύγουστο παρατηρείται η χαμηλότερη μέση τιμή CO.



Διάγραμμα 38 Σύγκριση μέσης τιμής CO μεταξύ εργάσιμων και Σαββατοκύριακων

Πέρα από τα προαναφερόμενα, θεωρήθηκε σημαντικό να ερευνηθεί μια ακόμα υπόθεση, εάν μπορεί να παρατηρηθεί σημαντική διαφορά στην μέση τιμή συγκέντρωσης CO, μεταξύ καθημερινών και σαββατοκύριακων. Πιο συγκεκριμένα, θεωρείται ότι επειδή περισσότερα άτομα δεν εργάζονται τα σαββατοκύριακα, άρα δεν χρησιμοποιούν τα μεταφορικά μέσα τους, η μέση τιμή για τα σαββατοκύριακα θα είναι χαμηλότερη από τις καθημερινές.

Πράγματι, αυτό επιβεβαιώνεται από το διάγραμμα 38, όπου παρουσιάζεται το γράφημα της μέσης τιμής ανά ώρα για τις καθημερινές, δηλαδή Δευτέρα με Παρασκευή, σε σχέση με Σάββατο και Κυριακή. Γενικά, παρατηρείται ότι οι καθημερινές έχουν σχεδόν την μισή τιμή για τις πρωινές μέχρι τις βραδινές ώρες, δηλαδή από τις 7:00 μέχρι τις 20:00. Πρακτικά, για τα σαββατοκύριακα διατηρείται η μορφή των 2 επαναλαμβανόμενων κορυφών που παρουσιάστηκε προηγουμένως στο διάγραμμα 36 αλλά με πιο ήπια μορφή. Αξιοσημείωτο είναι ότι παρατηρείται σύγκλιση των τιμών τα μεσάνυχτα καθώς και το γεγονός ότι τις πρωινές ώρες, συγκεκριμένα από τις 0:00 μέχρι τις 5:00, τα σαββατοκύριακα παρουσιάζουν μεγαλύτερες μέσες τιμές CO.

Πλέον, θεωρείται ότι έχει δημιουργηθεί μια καθαρή εικόνα για τα δεδομένα και είναι εφικτή η εφαρμογή διαφορετικών μεθόδων για την παραγωγή προβλεπτικών μοντέλων. Ωστόσο, πριν από οποιαδήποτε εφαρμογή μεθόδου, πρέπει να εφαρμοστεί διαχωρισμός δεδομένων (train – test split) για την αποφυγή Data Leak.

Οι Yang, Li και Jiang [32], αναφέρουν ότι σε προηγούμενες έρευνες, όταν χρησιμοποιήθηκαν κλασικές μέθοδοι αποσύνθεσης και τεχνικές βαθιάς μάθησης, ορισμένα χαρακτηριστικά του συνόλου δεδομένων που προοριζόταν για έλεγχο διέρρευσαν στο μοντέλο κατά τη διαδικασία εκπαίδευσης και αξιολόγησης, δηλαδή του test σετ. Αυτό είχε ως αποτέλεσμα το μοντέλο να εμφανίζει πολύ υψηλές μετρικές ακρίβειας που πρακτικά ήταν μια «ψευδαίσθηση» [32]. Η ύπαρξη data leakage οδηγεί σε μη αξιόπιστα αποτελέσματα κατά την εκπαίδευση και την αξιολόγηση των μοντέλων, ενώ καθιστά δύσκολη τη σωστή εκτίμηση της ικανότητάς τους να γενικεύουν. Στην πρόβλεψη χρονοσειρών, όπου αξιοποιούνται ιστορικά δεδομένα για την εκτίμηση μελλοντικών τιμών, είναι απαραίτητο τα μοντέλα να διαθέτουν ικανότητα γενίκευσης. Ωστόσο, όταν εμφανίζεται διαρροή πληροφορίας, δεδομένα του συνόλου ελέγχου - δοκιμής καθίστανται ουσιαστικά «γνωστά» στο

μοντέλο, με αποτέλεσμα να περιορίζεται η δυνατότητά του να ανταποκρίνεται σε πραγματικά "άγνωστα" δεδομένα. Επιπλέον, ο αριθμός των συνιστωσών επηρεάζεται σημαντικά από την κατανομή και το μήκος της χρονοσειράς. Οι επιπτώσεις της διαρροής πληροφορίας μπορούν να είναι ιδιαίτερα σοβαρές, καθώς το μοντέλο τείνει να προσαρμόζεται υπερβολικά σε συγκεκριμένα δείγματα και μοτίβα του συνόλου εκπαίδευσης αντί να μαθαίνει γενικευμένους κανόνες. Αυτό μπορεί να έχει ως αποτέλεσμα καλή απόδοση στο σύνολο ελέγχου – δοκιμής, αλλά χαμηλή αποτελεσματικότητα σε χρήση με πραγματικά δεδομένα. Σε περιπτώσεις όπου υπάρχει διαρροή πληροφορίας, η αξιολόγηση της απόδοσης στο σύνολο ελέγχου καθίσταται μη έγκυρη, επειδή το σύνολο αυτό δεν αντιπροσωπεύει επαρκώς άγνωστα δεδομένα. Κατά συνέπεια, μπορεί να προκύψει παραπλανητική εικόνα σχετικά με τις δυνατότητες και τους περιορισμούς του μοντέλου. Σήμερα, οι ερευνητές έχουν πλέον αναγνωρίσει το ζήτημα της εμπλοκής μελλοντικών, άγνωστων δεδομένων κατά τη διαδικασία Seasonal Decomposition [32].

3.5 Βασικά Στοιχεία Χρονοσειρών

Πριν την παρουσίαση πιο εξιδεικευμένων πληροφορίων και μεθόδων, εδώ θεωρείται ένα καλό σημείο για να δοθούν βασικές πληροφορίες σχετικά με την ανάλυση χρονοσειρών.

Έστω Y μια μεταβλητή που λαμβάνει τιμές Y_t, οι οποίες παρατηρούνται διαδοχικά σε τακτά χρονικά διαστήματα κατά τον χρόνο t, όπου t=1,2,3,4,…,n. Το σύνολο των παρατηρήσεων $Y_1, Y_2, Y_3 \dots, Y_n$ αποτελούν τα δεδομένα μιας χρονοσειράς. Αφού αυτές οι παρατηρήσεις είναι διαδοχικές, οι τιμές θα παρουσιάζουν ομοιότητες με αυτές που είναι χρονικά παραπλήσιές με αυτές. Το μέγεθος αυτής της ομοιότητας είναι γνωστό ως αυτοσυσχέτιση (autocorrelation). Χρονοσειρές που δεν παρουσιάζουν αυτοσυσχέτιση χαρακτηρίζονται ως λευκός θόρυβος (white noise) [30].

Όπως αναφέρουν οι Niako et al., η ύπαρξη εξάρτησης στα δεδομένα χρονοσειρών είναι ιδιαίτερα σημαντική για τη μοντελοποίηση της στοχαστικής συμπεριφοράς μιας διαδικασίας [30]. Τα δεδομένα χρονοσειρών μπορούν να εμφανίζουν διαφορετικά μοτίβα και συχνά είναι χρήσιμο να αποσυντίθενται σε επιμέρους συνιστώσες: την τάση-κύκλο, την εποχική συνιστώσα και το υπόλοιπο (ή τυχαίο) μέρος, καθεμία από τις οποίες αποτυπώνει έναν διαφορετικό τύπο συμπεριφοράς . Η τάση αντιπροσωπεύει μια μακροχρόνια αύξηση ή μείωση των τιμών και μπορεί να έχει διάφορες μορφές, όπως γραμμική ή μη γραμμική. Η εποχικότητα εμφανίζεται όταν οι τιμές επηρεάζονται από επαναλαμβανόμενους παράγοντες που εκδηλώνονται σε σταθερά και γνωστά χρονικά διαστήματα. Όταν παρατηρούνται αυξομειώσεις που δεν επαναλαμβάνονται περιοδικά, τότε γίνεται λόγος για κυκλική συμπεριφορά. Επειδή τα χρονικά διαστήματα μεταξύ των κύκλων δεν είναι σταθερά, η εκτίμηση της κυκλικής συνιστώσας είναι συνήθως δύσκολη. Τέλος, τα υπολείμματα περιλαμβάνουν ό,τι δεν εξηγείται από την τάση και την εποχικότητα, δηλαδή τις μη συστηματικές μεταβολές της χρονοσειράς [30].

3.6 Διαχωρισμός σετ εκπαίδευσης και δοκιμής - επικύρωσης

Είναι γεγονός ότι για δεδομένα που δεν είναι τύπου χρονοσειράς, δηλαδή τα δείγματα που απαρτίζουν το σύνολο δεδομένων είναι ανεξάρτητα μεταξύ τους και η σειρά που παρουσιάζονται δεν έχει καμιά σημασία για το τελικό ζητούμενο ανάλογα το πρόβλημα δηλαδή εάν είναι πρόβλημα επιβλεπόμενης ή μη επιβλεπόμενης μάθησης,

Στις περισσότερες εργασίες εποπτευόμενης μάθησης που περιλαμβάνουν ανεξάρτητες και ομοιόμορφα κατανεμημένες παρατηρήσεις, τα σύνολα δεδομένων συνήθως χωρίζονται σε υποσύνολα εκπαίδευσης και δοκιμής χρησιμοποιώντας τυχαία ανακατάταξη. Αυτή η στρατηγική εξασφαλίζει αντιπροσωπευτική δειγματοληψία και αμερόληπτη αξιολόγηση. Ωστόσο, στην πρόβλεψη χρονοσειρών, οι παρατηρήσεις είναι ταξινομημένες χρονικά και παρουσιάζουν

αυτοσυσχέτιση. Η τυχαία ανακατάταξη θα παραβίαζε τη χρονική αιτιότητα, θα μπορούσε να προκαλέσει διαρροή δεδομένων και θα παρήγαγε υπερβολικά αισιόδοξες μετρικές απόδοσης. Επομένως, τα δεδομένα χρονοσειρών πρέπει να χωρίζονται χρονολογικά, διατηρώντας τη φυσική σειρά των παρατηρήσεων. Συνήθως, στα προβλήματα εποπτευόμενης μάθησης εφαρμόζεται διαχωρισμός 70% / 30% ή 80% / 20%, τα μεν ποσοστά αφορούν την χρήση για εκπαίδευση και τα δε για δοκιμή – αξιολόγηση.

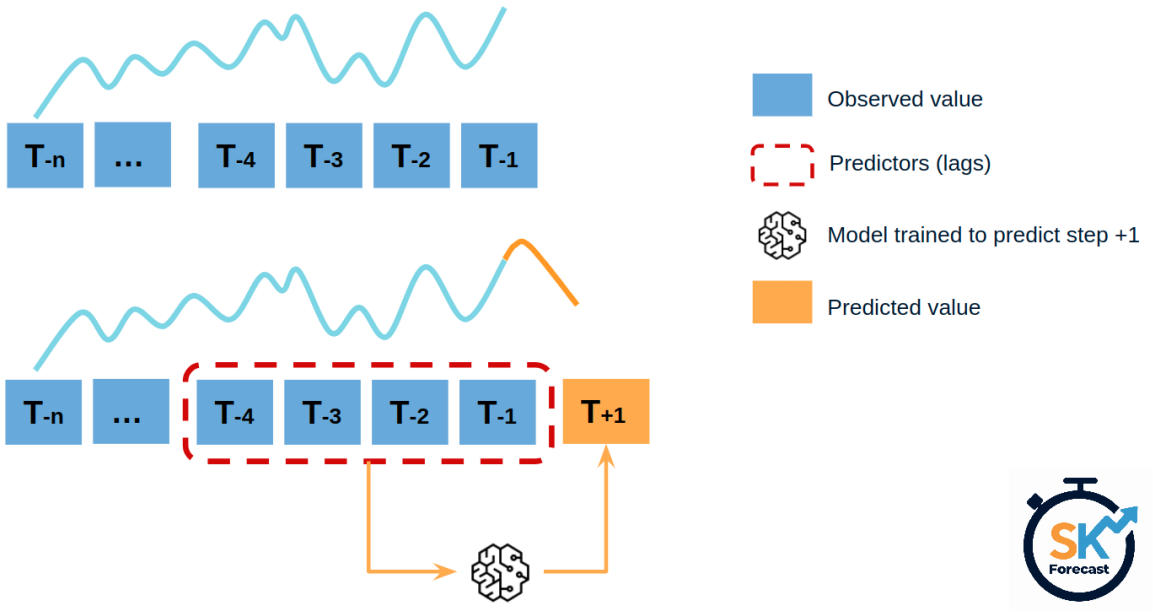
Ωστόσο, όπως αναφέρουν οι Suradhaniwar et al., η βασική παραδοχή αυτού του «κλασσικού» τρόπου διαχωρισμού είναι ότι η δειγματοληψία είναι αντιπροσωπευτική για το σύνολο των παρατηρήσεων. Δηλαδή, τυπικά αυτό σημαίνει ότι τα δείγματα που θα ληφθούν ως σετ εκπαίδευσης και δοκιμής, θα παρουσιάζουν παραπλήσιες - όμοιες καμπύλες κατανομής, μέσες τιμές και τιμές διακύμανσης μεταξύ τους. Εάν όλα τα προαναφερόμενα είναι σταθερά για μια χρονοσειρά, τότε χαρακτηρίζεται ως «στάσιμη». Έτσι, οι αλγόριθμοι πρόβλεψης χρονοσειρών βασίζονται στην υπόθεση ότι οι συσχετίσεις που παρατηρούνται ή μαθαίνονται κατά την εκπαίδευση παραμένουν έγκυρες και κατά την αξιολόγηση ή δοκιμή σε δεδομένα εκτός δείγματος. Ωστόσο, στην πράξη, τέτοιες παραδοχές συχνά οδηγούν σε σφάλματα, ιδιαίτερα σε έντονα στοχαστικές διεργασίες όπως η σχετική υγρασία. Για παράδειγμα, μια χρονοσειρά υγρασίας μπορεί να κυμαίνεται γύρω στο 30–40% για ένα χρονικό διάστημα και να αυξάνεται απότομα έως το 100% κατά τη διάρκεια βροχόπτωσης, δημιουργώντας ξαφνικές μεταβολές που δεν έχουν παρατηρηθεί προηγουμένως στα δεδομένα εκπαίδευσης. Τέτοιου είδους απότομες και χωρίς προηγούμενο μεταβολές στη διακύμανση των δεδομένων (ετεροσκεδαστικότητα για τα γραμμικά μοντέλα) θα οδηγήσουν σε υποεκτίμηση ή υπερεκτίμηση της απόδοσης των μοντέλων, καθώς ενσωματώνονται στη διαδικασία μοντελοποίησης και παραβιάζουν την προαναφερόμενη σταθερότητα ως προς την βασική παραδοχή του διαχωρισμού - δειγματοληψίας για τα σύνολα εκπαίδευσης και ελέγχου. Επιπλέον, οι μεταβολές αυτές επηρεάζουν αρνητικά την αξιολόγηση των μοντέλων, καθώς αυξάνουν τη μεταβλητότητα κατά την πρόβλεψη, οδηγώντας σε μεγαλύτερα σφάλματα. Αυτά σφάλματα αυτά είναι πιο έντονα στις βραχυπρόθεσμες προβλέψεις, ενώ σε μακροπρόθεσμες προβλέψεις όπως για παράδειγμα σε χρονική κλίμακα πρόβλεψης πάνω του ενός χρόνου, τείνουν να εξομαλύνονται, καθώς το μοντέλο είναι πιθανό να περιέχει αυτές τις ακραίες τιμές ή περιόδους, επομένως μειώνεται σημαντικά η επίδρασή τους στα προβλεπτικά μοντέλα [26].

Βάσει των προαναφερόμενων, ως βέλτιστη προσέγγιση θεωρήθηκε ο διαχωρισμός των δεδομένων βάσει σημείου. Δηλαδή, επιλέχτηκε ότι ως περίοδος test η τελευταία εβδομάδα δεδομένων, αφού θεωρείται αρκετή λόγω του βραχυπρόθεσμου ορίζοντα πρόβλεψής που προαναφέρθηκε, και ως περίοδος εκπαίδευσης οι λοιπές παρατηρήσεις. Λόγω του ότι θα χρησιμοποιηθούν univariate και multivariate μοντέλα, για παράδειγμα SARIMA και SARIMAX, πρακτικά δημιουργούνται 2 splits – διαχωρισμοί δεδομένων: Το “y_train, y_test” αφορά το βασικό γνώρισμα “CO(GT)” που θα χρησιμοποιηθεί για τα univariate μοντέλα. Το “x_train, x_test” αφορά τις εξωγενείς μεταβλητές που θα χρησιμοποιηθούν στα multivariate μοντέλα. Βάσει του διαγράμματος 34, επιλέχτηκαν οι χρονικές μεταβλητές, δηλαδή ώρα, ημέρα της εβδομάδας και μήνας με κυκλική κωδικοποίηση.

3.7 Προσεγγίσεις “Single-Step” / “Multi-Step” – “Recursive / Non-Recursive”

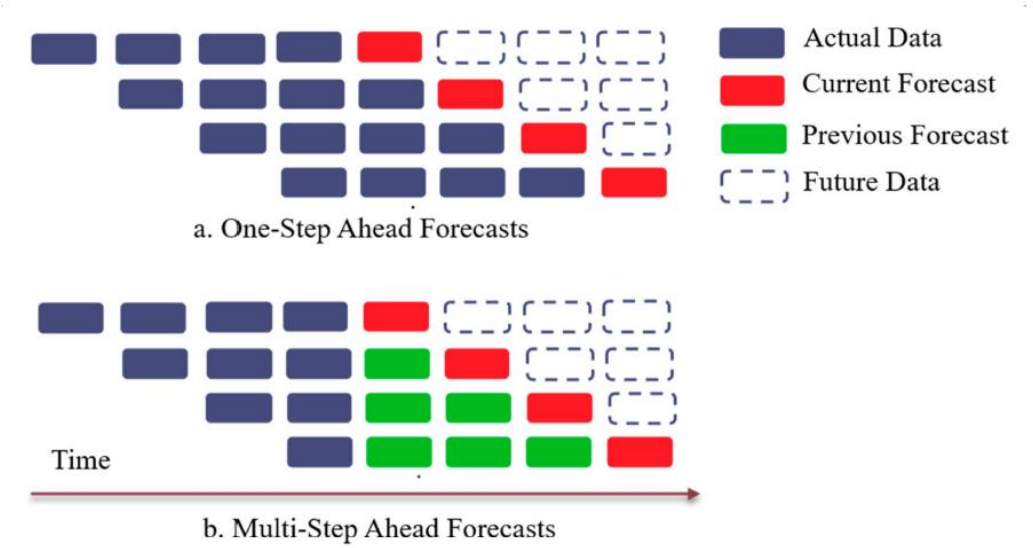
Όσον αφορά τον χρονικό ορίζοντα πρόβλεψης, υπενθυμίζεται ότι οι Liu et al., [1] αναφέρουν ότι οι βραχυπρόθεσμοι χρονικοί ορίζοντες πρόβλεψης παράγουν καλύτερες μετρικές απόδοσης ως προς

την πρόβλεψη ενώ οι Suradhaniwar et al., αναφέρουν ότι η βιβλιογραφία σχετικά με την χρήση προβλεπτικών αλγόριθμων για χρονοσειρές μετεωρολογικών δεδομένων, εστιάζει κυρίως σε “one-step ahead” προβλέψεις [26], δηλαδή πρόβλεψη όπου ο χρονικός ορίζοντας είναι ίσος με μία χρονική μονάδα [51]. Δηλαδή εάν τα δεδομένα είναι σε ωριαία κλίμακα αφορά πρόβλεψη μιας ώρας ενώ εάν είναι σε ημερήσια κλίμακα, αφορούν μια ημέρα. Τα παραπάνω δίνονται ως παράδειγμα στο διάγραμμα 39 παρακάτω.



Διάγραμμα 39 Παράδειγμα προσέγγισης one – step ahead. Πηγή: [51]

Σε αυτό το σημείο χρειάζεται να γίνει μια σημαντική παρατήρηση. Όταν γίνεται αναφορά σε μεθοδολογίες “one-step ahead”, χρειάζεται να προσδιοριστεί εάν είναι recursive ή non-recursive (αναδρομικά ή μη αναδρομικά). Εάν μια μέθοδος είναι αναδρομική, σημαίνει ότι οι προβλέψεις που δημιουργούνται εντάσσονται ως μέρος του συνόλου των δεδομένων ενώ εάν είναι μη – αναδρομικά, δεν εντάσσονται στο σύνολο δεδομένων. Για παράδειγμα μπορεί να χρησιμοποιηθεί το διάγραμμα 40.

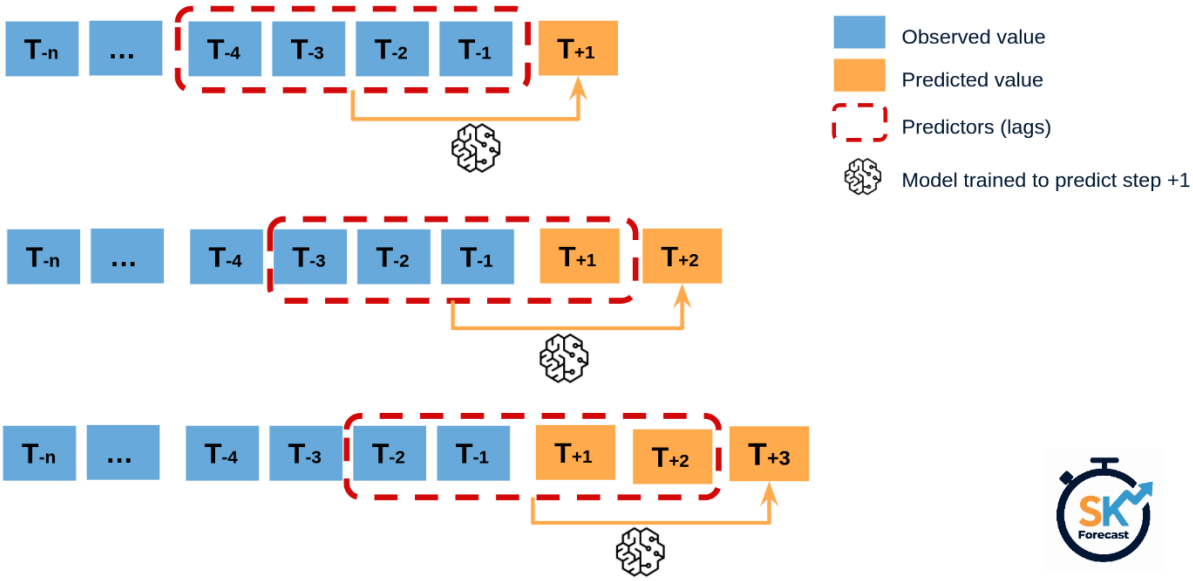


Διάγραμμα 40 Παράδειγμα σύγκρισης one-step ahead και multi-step προσεγγίσεων. Πηγή: [26]

Για "one-step ahead" προβλέψεις, η πρώτη σειρά που παρατηρείται περιλαμβάνει 4 δείγματα ως πραγματικές παρατηρήσεις με μπλε χρώμα, όπου έστω η τελευταία παρατήρηση να χαρακτηριστεί ως "n". Επομένως, η πρόβλεψη εντός ενός βήματος θα χαρακτηριστεί ως "n+1", με κόκκινο χρώμα. Στην επόμενη σειρά, εντοπίζεται η πρόβλεψη "n+2", όμως η προηγούμενη πρόβλεψη "n+1" δεν εισάγεται στο σετ δεδομένων αλλά αντικαθίσταται από την πραγματική παρατηρούμενη τιμή.

Σε αντίθεση, σε μια "Multi-step" προσέγγιση όπως φαίνεται στο διάγραμμα 41, η οποία είναι εξ' ορισμού αναδρομική [49] αλλά όχι απαραίτητα [50], ο χρονικός ορίζοντας πρόβλεψης θα είναι μεγαλύτερος του 1 και οι προβλέψεις εισάγονται στο σύνολο δεδομένων . Δηλαδή, η παρατήρηση "n+1" που παρατηρείται στην πρώτη «σειρά», εισάγεται στο σετ δεδομένων και η πρόβλεψη "n+2" γίνεται βάση της πρόβλεψης "n+1" και όχι της πραγματικής παρατηρούμενης τιμής, όπως φαίνεται στο διάγραμμα 40.

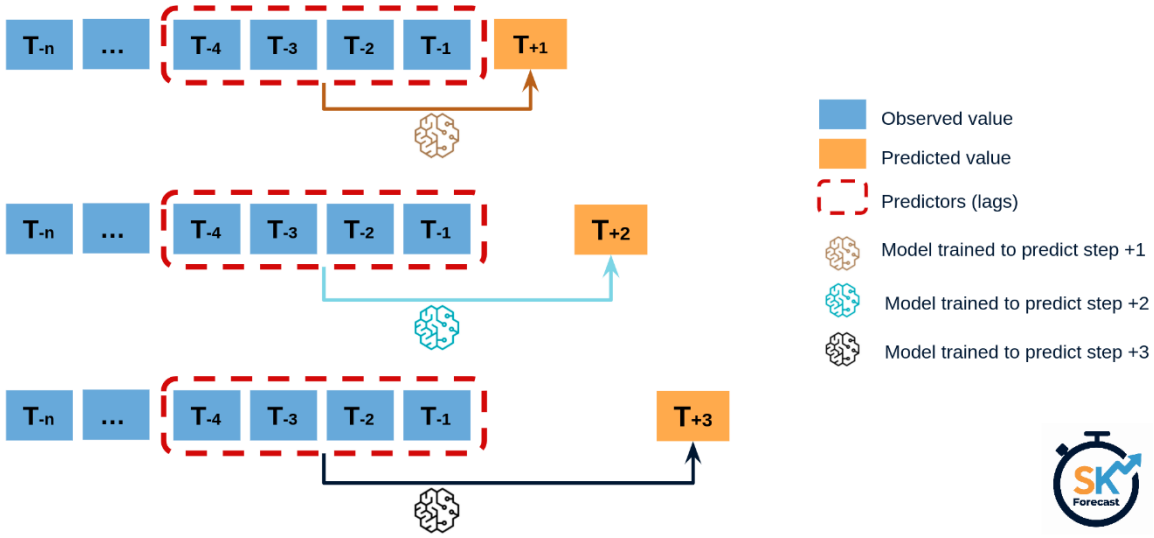
Μια πιο ορθή ορολογία θα ήταν ότι η πρόβλεψη n+1 χρησιμοποιείται ως "predictor – lag". Σχετικά με το ότι τα "Multi-step" μοντέλα είναι εξ΄ ορισμού αναδρομικά, αναφέρουν ότι πρακτικά αυτά τα μοντέλα είναι πολλαπλά "one-step" μοντέλα, δηλαδή αναδρομικά, που εκτελούνται μέχρι τον ορισμένο χρονικό ορίζοντα πρόβλεψης [26].



Διάγραμμα 41 Παράδειγμα προσέγγισης multi-step. Πηγή: [49]

Όσον αφορά ότι μια "multi-step" μέθοδος δεν είναι απαραίτητα αναδρομική, γίνεται αναφορά στην προσέγγιση γνωστή ως "Direct multi-step" [50] όπου δεν χρησιμοποιούνται οι προβλέψεις ως "predictor – lags", δηλαδή να εισάγονται στο σετ δεδομένων αλλά δημιουργούνται παράλληλα διαφορετικά μοντέλα για κάθε πρόβλεψη του ορίζοντα πρόβλεψης.

Για την επεξήγηση της συγκεκριμένης προσέγγισης μπορεί να χρησιμοποιηθεί το διάγραμμα 42. Εδώ, ο χρονικός ορίζοντας του παραδείγματος είναι ίσος με 3. Σε αντίθεση με την προηγούμενη προσέγγιση που ουσιαστικά μοιάζει με "sliding window", εδώ το παράθυρο μένει σταθερό και δημιουργούνται 3 διαφορετικά μοντέλα για κάθε n+1, 2 και 3, χωρίς να χρησιμοποιούνται οι προβλέψεις ως "predictor – lags".



Διάγραμμα 42 Παράδειγμα προσέγγισης direct multi-step. Πηγή:[50]

Ωστόσο τα παραπάνω φαίνεται να συγχέονται μεταξύ τους, καθώς για παράδειγμα οι Suradhaniwar et al., αναφέρουν ότι οι "single-step" προσεγγίσεις ενσωματώνουν τις μελλοντικές προβλέψεις ως "predictor – lags", δηλαδή είναι αναδρομικές προσεγγίσεις. Ανεξαρτήτως της πιθανής σύγχυσης, το

βασικό πλεονέκτημα των μη – αναδρομικών “single-step” προσεγγίσεων σε σχέση με αναδρομικές είναι ότι τα σφάλματα προβλέψεων δεν συσσωρεύονται, αφού δεν χρησιμοποιούνται ως “predictors”, επομένως η ακρίβεια πρόβλεψης δεν μειώνεται στο βάθος του χρονικού ορίζοντα πρόβλεψης [26]. Μάλιστα, οι ίδιοι χρησιμοποίησαν μια προσέγγιση την οποία χαρακτηρίζουν ως “walk – forward validation”, η οποία πρακτικά ήταν μια μη – αναδρομική “single-step”. Επίσης, οι ίδιοι αναφέρουν ότι η επιλογή μεταξύ μοντέλων single-step και multi-step εξαρτάται από τον επιθυμητό στόχο της κάθε έρευνας.

Πέρα από τα παραπάνω υπενθυμίζεται, οι προβλέψεις ποιότητας αέρα αξιολογούνται με την χρήση συγκεκριμένων χρονικών οριζόντων. Για παράδειγμα, οι Athanasiadis, Karatzas και Mitkas [27] αναδεικνύουν τις διαφορές μεταξύ διαφορετικών χρονικών οριζόντων ως προς τα αποτελέσματα των προβλέψεων, συγκεκριμένα μεταξύ οριζόντων 8, 24, 48 και 72 ωρών ενώ οι Mirzadeh και Omranpour [66] εστιάζουν σε ορίζοντές μεταξύ 1 και 6 ωρών.

3.8 Έλεγχος στασιμότητας και εποχικότητας

Πριν από την εφαρμογή μεθόδων μια καλή πρακτική είναι ο έλεγχος της χρονοσειράς ως προς την στασιμότητα της καθώς αρκετοί αλγόριθμοι έχουν ως παραδοχή λειτουργίας τους ότι η χρονοσειρά είναι στάσιμη. Ως στάσιμη, χαρακτηρίζεται μια χρονοσειρά, της οποίας οι στατιστικές ιδιότητες όπως ενδιάμεση τιμή ή διακύμανση, είναι σταθερές και δηλαδή δεν εξαρτούνται από την χρονική στιγμή που παρατηρούνται. Βάσει αυτού, εάν μια χρονοσειρά περιέχει τάση (trend) ή εποχικότητα (seasonality) δεν θεωρείται στάσιμη, καθώς είτε η τάση είτε η εποχικότητα θα επηρεάζει τις τιμές της χρονοσειράς σε διαφορετικές στιγμές. Ωστόσο, μια χρονοσειρά που θα χαρακτηριζόταν ως διεργασία “White Noise” ή αλλιώς “Λευκού Θορύβου” είναι στάσιμη [40]. Ως «White Noise», χαρακτηρίζονται οι χρονοσειρές που παρουσιάζουν σχεδόν μηδενική αυτοσυσχέτιση, χωρίς όμως να παρουσιάζουν πραγματική μηδενική αυτοσυσχέτιση λόγω τυχαίων διακυμάνσεων που προκύπτουν φυσικά. Συγκεκριμένα, το 95% των τιμών ACF σε μια διεργασία “White Noise” βρίσκονται εντός του διαστήματος $\pm 2/\sqrt{T}$, όπου T το μήκος της χρονοσειράς, δηλαδή ο συνολικός αριθμός παρατηρήσεων [53].

Ωστόσο, χρειάζεται να αναφερθεί ότι ορισμένες χρονοσειρές μπορεί να προκαλέσουν σύγχυση σχετικά με το παραπάνω, καθώς μπορεί να παρουσιάζουν κυκλική συμπεριφορά, χωρίς όμως να υπάρχει τάση ή εποχικότητα, χαρακτηρίζονται ως στάσιμες. Αυτό συμβαίνει επειδή η κυκλικότητα μπορεί να περιγράφει ως μια ακανόνιστη εποχικότητα, δηλαδή οι κύκλοι που δημιουργούνται είναι άγνωστο πότε θα εμφανιστούν και πόσο θα διαρκέσουν. Δηλαδή, βάσει του παραπάνω, μια στάσιμη χρονοσειρά είναι αυτή που δεν παρουσιάζει προβλέψιμα μοτίβα στην μακροχρόνια παρατήρηση της. Εάν δημιουργηθεί το γράφημα μιας στάσιμης χρονοσειράς θα πρέπει να είναι οριζόντιο, χωρίς ανοδικές - καθοδικές τάσεις ή εποχικότητα και ίσως παρουσιάζοντας κυκλική συμπεριφορά, με σταθερή διακύμανση στον χρόνο [40].

Όπως αναφέρουν οι Hyndman και Athanasopoulos, όταν η χρονική κλίμακα των δεδομένων είναι ημερήσια ή ωριαία όπως στο συγκεκριμένο σετ δεδομένων, η ανάλυση της χρονοσειράς είναι σημαντικά πιο απαιτητική καθώς συχνά υπάρχουν πολλά εποχικά μοτίβα [41]. Ένας σύνηθες τρόπος για τον έλεγχο στασιμότητας μιας χρονοσειράς, και κατ’ επέκταση ένα είναι αναγκαία η εφαρμογή πρώτων ή εποχιακών διαφορών, είναι η χρήση κάποιου “Unit Root Test” [40]. Τα 2 πιο συχνά τεστ τέτοιου τύπου είναι το ADF και το KPSS [54], όπου:

Για το ADF Test, η μηδενική υπόθεση (H0) αφορά την ύπαρξη unit root στην σειρά ενώ η ενναλακτική υπόθεση (H1) αφορά την μη – ύπαρξη μοναδιαίας ρίζας (unit root). Μάλιστα η ύπαρξη μοναδιαίας ρίζας μπορεί να χρησιμοποιηθεί ως εργαλείο ανίχνευσης ύπαρξης στοχαστικής τάσης [57].

Το KPSS test αντιστρέφει τις υποθέσεις, καθώς η μηδενική υπόθεση είναι ότι η σειρά είναι στάσιμη ως προς την τάση (trend stationary) ενώ η εναλλακτική υπόθεση είναι η μη - στασιμότητα της χρονοσειράς, δηλαδή η ύπαρξη μοναδιαίας ρίζας [54]. Μια καλή πρακτική είναι η συνδιαστική χρήση και των 2 τεστ, καθώς με αυτόν τον τρόπο μπορεί να επιβεβαιωθεί ότι μια χρονοσειρά είναι πραγματικά στάσιμη. Συγκεκριμένα, με την χρήση των 2 τεστ υπάρχουν 4 πιθανά αποτελέσματα [54]:

- 1η Περίπτωση: Τα 2 τεστ καταλήγουν ότι η χρονοσειρά δεν είναι στάσιμη, επομένως πράγματι δεν είναι στάσιμη.

- 2η Περίπτωση: Τα 2 τεστ καταλήγουν ότι η χρονοσειρά είναι στάσιμη, επομένως πράγματι η χρονοσειρά είναι στάσιμη.

- 3η Περίπτωση: Το KPSS τεστ καταλήγει σε στασιμότητα ενώ το ADF τεστ καταλήγει σε μη – στασιμότητα. Επομένως, η χρονοσειρά χαρακτηρίζεται ως στάσιμη γύρω από την τάση (trend stationary).
  Δηλαδή, η τάση πρέπει να αφαιρεθεί και μετέπειτα να ελεγχθεί ξανά ως προς την στασιμότητα.

- 4η Περίπτωση: Το KPSS τεστ καταλήγει σε μη - στασιμότητα ενώ το ADF τεστ καταλήγει σε στασιμότητα. Η χρονοσειρά είναι στάσιμη γύρω από τις διαφορές (difference stationary). Μετέπειτα ελέγχεται ξανά ως προς την στασιμότητα.

Όπου για τα παραπάνω ισχύει [57]:

- Trend Stationary - στάσιμη γύρω από την τάση: Η μέση τάση είναι ντετερμινιστική και μόλις εκτιμηθεί και αφαιρεθεί από την χρονοσειρά, τα κατάλοιπα χαρακτηρίζονται ως μια στάσιμη στοχαστική διαδικασία (stationary stochastic process).

- Difference Stationary - στάσιμη γύρω από τις διαφορές: Η μέση τάση είναι στοχαστική. Εάν εφαρμοστούν διαφορές στην χρονοσειρά, θα προκύψει μια στάσιμη στοχαστική διαδικασία.

Είναι σημαντική η διαφοροποίηση μεταξύ ντετερμινιστικών και στοχαστικών στοιχείων καθώς η ύπαρξη του καθενός επηρεάζει άμεσα και σε σημαντικό βαθμό την «συμπεριφορά» μιας διαδικασίας. Μια χρονοσειρά με ντετερμινιστική τάση πάντα θα επιστρέφει στην τάση της μακροχρόνια, δηλαδή αφού οι επιδράσεις ακραίων γεγονότων (shocks) σταματήσουν να επιδρούν. Σε αντίθεση, οι χρονοσειρές με στοχαστική τάση δεν επιστρέφουν ποτέ στην τάση πριν τα ακραία γεγονότα ή με άλλο τρόπο, οι αλλαγές που επιφέρουν τα shocks είναι μόνιμα [57].

3.8.1 Στοχαστικά και Ντετερμινιστικά μοτίβα

Πριν από την εφαρμογή κάποιου “seasonal – unit root test”, εδώ πρέπει να γίνει διαχωρισμός μεταξύ «στοχαστικών» και «ντετερμινιστικών» μοτίβων. Τα μοτίβα μπορούν αναφέρονται και στην τάση ωστόσο προς το παρόν θα γίνει εστίαση στην εποχικότητα [55, 46] :

Ως «Ντετερμινιστική» εποχικότητα, ορίζεται η εποχικότητα που παρουσιάζει σταθερό επαναλαμβανόμενο μοτίβο. Δηλαδή, οι κορυφές και οι κοιλότητες θα έχουν το ίδιο μέγεθος – ένταση ενώ παράλληλα θα εμφανίζονται σε κάποια σταθερά χρονικά πολλαπλάσια. Με άλλα λόγια, το χρονικό διάστημα μεταξύ των επαναλήψεων του εποχικού προτύπου είναι σταθερό [55]. Άρα, μια χρονοσειρά που περιέχει ντετερμινιστική εποχικότητα δεν μπορεί να είναι στάσιμη.

Η «Ντετερμινιστική» εποχικότητα μπορεί να αντιμετωπιστεί αποτελεσματικά με τη χρήση εποχικών ψευδομεταβλητών (dummy variables). Πρόκειται για κατηγορικές μεταβλητές που περιγράφουν την εποχική περίοδο, όπως για παράδειγμα ο μήνας που αντιστοιχεί σε κάθε χρονικό σημείο. Στη συνέχεια, η κατηγορική αυτή μεταβλητή μετατρέπεται σε ένα σύνολο δυαδικών μεταβλητών (indicator variables) μέσω της τεχνικής one-hot encoding. Εναλλακτικά, η εποχικότητα μπορεί να μοντελοποιηθεί με τη χρήση σειρών Fourier, οι οποίες βασίζονται σε ημιτονοειδείς και συνημιτονοειδείς συναρτήσεις με διαφορετικές περιόδους, επιτρέποντας την αποτύπωση περιοδικών μοτίβων στα δεδομένα [55, 46, 56].

Η «στοχαστική» εποχικότητα μπορεί να είναι στάσιμη και μη-στάσιμη. Η στοχαστική στάσιμη εποχικότητα μεταβάλλεται από περίοδο σε περίοδο (π.χ. από έτος σε έτος). Η έντασή της δεν είναι πλήρως προβλέψιμη, ωστόσο η περιοδικότητα παραμένει περίπου σταθερή. Αντίθετα, στην περίπτωση της ντετερμινιστικής εποχικότητας, η βέλτιστη πρόβλεψη για έναν συγκεκριμένο μήνα παραμένει ίδια ανεξάρτητα από το έτος. Στη στοχαστική στάσιμη εποχικότητα, όμως, η εκτίμηση για έναν μήνα επηρεάζεται από την τιμή του ίδιου μήνα στο προηγούμενο έτος [55].

Στην στοχαστική μη-στάσιμη εποχικότητα, τα εποχικά πρότυπα μεταβάλλονται σημαντικά κατά τη διάρκεια διαδοχικών εποχικών περιόδων. Οι μεταβολές αυτές μπορεί να οφείλονται στην ύπαρξη εποχικών μοναδιαίων ριζών, γεγονός που υποδηλώνει ότι η εποχικότητα είναι ολοκληρωμένη (integrated). Σε αυτό το είδος εποχικότητας δεν μεταβάλλεται μόνο η ένταση, αλλά και η περιοδικότητα με την πάροδο του χρόνου. Αυτό σημαίνει ότι οι κορυφές και τα χαμηλά σημεία δεν εμφανίζονται πλέον στις ίδιες χρονικές θέσεις. Παρόμοια φαινόμενα παρατηρούνται σε διάφορους τομείς, όπως σε χρονοσειρές κατανάλωσης ή δεδομένα βιομηχανικής παραγωγής. Η πρόβλεψη τέτοιων μεταβολών καθίσταται ιδιαίτερα δύσκολη όταν η χρονοσειρά χαρακτηρίζεται από ολοκληρωμένη εποχικότητα [55].

Συνοψίζοντας τα παραπάνω [56]:

- Εάν η χρονοσειρά δεν εμφανίζει εποχικότητα:
  - Και είναι στάσιμη, τότε προτείνεται η χρήση ARMA(p,q)
  - Εάν δεν είναι στάσιμη:
    - Υπάρχει ντετερμινιστική τάση
    - Υπάρχει στοχαστική τάση, χρειάζεται να παρθούν οι πρώτες διαφορές ή διαφορετικά να χρησιμοποιηθεί μοντέλο ARIMA(p,d,q)

- Εάν η χρονοσειρά εμφανίζει εποχικότητα:
  - Και είναι στάσιμη, τότε προτείνεται η χρήση SARMA(p,q)(P,Q)s
  - Εάν δεν είναι στάσιμη:
    - Υπάρχει ντετερμινιστική εποχικότητα και προτείνεται η χρήση dummy variables ή σειρών Fourier
    - Υπάρχει στοχαστική εποχικότητα, προτείνεται η χρήση SARIMA(p,d,q)(P,D,Q)s

Ωστόσο, είναι ιδιαίτερα αξιοσημείωτο ότι όλες οι προαναφερόμενες μορφές εποχικότητας δεν είναι αμοιβαία αποκλειστικές μεταξύ τους και όπως προαναφέρθηκε μια χρονοσειρά μπορεί να αποτελείται από διαφορετικά εποχικά μοτίβα [41]. Επομένως, είναι εφικτό μια χρονοσειρά να παρουσιάζει ντετερμινιστική και στοχαστική εποχικότητα [55].

3.8.2 ADF Test

Πίνακας 36 Αποτελέσματα ADF Test

```
  Στατιστικό ADF: -10.6896
  p-value: 0.0000
Reject the null hypothesis: series is trend stationary - no trend
------------------------------------------------------------
```

Στον πίνακα 36, παρουσιάζονται τα αποτελέσματα του ADF Test. Η τιμή ADF βρέθηκε ως -10.68 ενώ η p-value βρέθηκε ως 0, βάσει στρογγυλοποίησης. Εφόσον p-value είναι μικρότερη του επίπεδου σημαντικότητας 0.05, η αρχική υπόθεση απορρίπτεται, επομένως δεν υπάρχει unit root και η χρονοσειρά θεωρείται στάσιμη – δεν παρουσιάζει τάση.

3.8.3 KPSS Test

Πίνακας 37 Αποτελέσματα KPSS Test

```
KPSS Statistic: 1.2133915244508608
p-value: 0.01
Critical Values: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}
Lags used: 42
Reject the null hypothesis (series is not Trend-stationary)
------------------------------------------------------------
```

Στον πίνακα 37, παρουσιάζονται τα αποτελέσματα του KPSS Test. Η τιμή KPSS βρέθηκε ως 1.21 και η p-value ως 0.01. Εφόσον η p-value είναι μικρότερη του επίπεδου σημαντικότητας, η αρχική υπόθεση απορρίπτεται που σημαίνει ότι η χρονοσειρά δεν είναι στάσιμη.

3.8.4 Kruskal – Wallis Test

Πέρα από τα unit roots tests, το Kruskal Wallis test μπορεί να εφαρμοστεί στις χρονοσειρές για τον έλεγχο ύπαρξης εποχικότητας. Πρακτικά, η συγκεκριμένη μέθοδος ανιχνεύει εάν υπάρχουν σημαντικές διαφορές μεταξύ διαφορετικών ομάδων, όπου στην συγκεκριμένη περίπτωση η ομάδες αφορούν είτε ώρες, ημέρες ή την ανάλογη χρονική μονάδα που χρησιμοποιείτε στο εκάστοτε σετ δεδομένων [45]. Πέρα αυτού, η ύπαρξη εποχικότητας μπορεί να εντοπιστεί μέσω των διαγραμμάτων ACF, αφού οι αυτοσυσχετίσεις θα είναι μεγαλύτερες για τις εποχιακές υστερήσεις (σε πολλαπλάσια της εποχιακής συχνότητας) από ό,τι για τις άλλες υστερήσεις, πράγμα το οποίο θα παρουσιαστεί παρακάτω.

Πίνακας 38 Αποτελέσματα Kruskal - Wallis Test

```
Kruskal-Wallis for 'CO(GT)' by hour: p = 0.0000  + Εποχικότητα
```

Στον πίνακα 38, παρουσιάζονται τα αποτελέσματα του Kruskal – Wallis test άνα ώρα, αφού τα δεδομένα έχουν ωριαία χρονική κλίμακα. Παρατηρείται ότι η p-value είναι μικρότερη του 0.05. Αυτό πρακτικά σημαίνει ότι τουλάχιστον 1 ώρα της ημέρας έχει διαφορετική κατανομή από τις υπόλοιπες, επομένως η τιμή του γνωρίσματος CO(GT) αλλάζει σημαντικά ανά ώρα, πράγμα που επιβεβαιώνεται από το διάγραμμα των μέσων τιμών ανά ώρα του διαγράμματος 36. Αυτό μπορεί να χαρακτηριστεί ως ντετερμινιστική ημερήσια εποχικότητα.

Συνοψίζοντας, το ADF Test επιβεβαιώνει ότι δεν υπάρχει unit root, ενώ η απόρριψη του KPSS Test υποδηλώνει ότι η χρονοσειρά δεν είναι στάσιμη, πιθανώς λόγω εποχικότητας ή τάσης ή και των δύο. Στην βιβλιογραφία, σε περίπτωση που το αποτέλεσμα του ADF Test αναδεικνύει στασιμότητα ενώ το KPSS Test αναδεικνύει μη-στασιμότητα, τότε η σειρά θεωρείται "difference stationary", δηλαδή η μέση τάση είναι στοχαστική και προτείνεται η χρήση πρώτων ή εποχιακών διαφορών ή και συνδυασμός των δύο, χωρίς αυτή η πρόταση να είναι αναγκαία [54,57]. Παράλληλα, το Kruskal – Wallis test επιβεβαιώνει ότι υπάρχει ντετερμινιστική ημερήσια εποχικότητα. Βάσει όλων των παραπάνω, είναι πιθανό η χρονοσειρά να παρουσιάζει έναν συνδυασμό από στοχαστική και ντετερμινιστική εποχικότητα, ίσως ακόμα και τάσης. Για να απορριφθεί η περίπτωση της τάσης, αποφασίστηκε να εφαρμοστεί Seasonal Decompositon, ώστε να εξεταστούν μεμονωμένα η τάση και η εποχικότητα.

3.9 Seasonal Decomposition

Όπως προαναφέρθηκε στο προηγούμενο κεφάλαιο, οι Hyndman και Athanasopoulos επισημαίνουν ότι τα δεδομένα με ωριαία χρονική κλίμακα συνήθως παρουσιάζουν πολλαπλά εποχιακά μοτίβα, δηλαδή δεν παρουσιάζουν σταθερή εποχικότητα, με αποτέλεσμα μέθοδοι που βασίζονται στην ύπαρξη μιας σταθεράς εποχικότητας όπως ETS ή μοντέλα SARIMA να μην μπορούν να ανταπεξέλθουν και προτείνετε η χρήση STL Decomposition ή μοντέλων Prophet [41].

Για την υλοποίηση του STL decomposition χρησιμοποιήθηκε η βιβλιοθήκη statsmodels μέσω του module "statsmodels.tsa.seasonal" [35]. Στην συγκεκριμένη περίπτωση δεν χρησιμοποιήθηκε για να δημιουργεί κάποιο μοντέλο πρόβλεψης βάσει του decomposition [39] αλλά για να παρουσιαστούν τα διαφορετικά στοιχεία που αποτελούν μια χρονοσειρά, δηλαδή την τάση, εποχικότητα και τα κατάλοιπα. Το decomposition μπορεί να χρησιμοποιηθεί πρακτικά για την εύρεση της κατάλληλης επεξεργασίας και κατ΄ επέκταση του κατάλληλου μοντέλου πρόβλεψης. Όπως αναφέρουν οι [26] μια αλλαγή στην τάση, είτε είναι αύξηση είτε είναι μείωση, υποδηλώνει ότι η χρονοσειρά πιθανώς δεν είναι στάσιμη, επομένως χρειάζεται να εφαρμοστούν τεστ ελέγχου στασιμότητας και πιθανώς να εφαρμοστεί "differencing", δηλαδή να παρθούν οι πρώτες διαφορές της χρονοσειρας.

Γενικά, υπάρχουν διαφορετικές προσεγγίσεις για την εκτέλεση ενός decomposition όπως classic, STL, X11 ή SEATS [36]. Όσο αφορά την κλασσική προσέγγιση υπάρχουν δύο μορφές κλασικής αποσύνθεσης: η προσθετική αποσύνθεση (additive) και η πολλαπλασιαστική αποσύνθεση (multiplicative) [37], οι οποίες περιγράφονται από τις παρακάτω σχέσεις [44 , 42] :

Πίνακας 39 Προσθετικό και πολλαπλασιαστικό μοντέλο αποσύνθεσης

$y_t = S_t + T_t + R_t$	(3)
$y_t = S_t * T_t * R_t$	(4)

Όπου S_t η εποχικότητα, T_t η τάση και R_t τα κατάλοιπα για την χρονική στιγμή t. Η προσθετική αποσύνθεση (additive decomposition) είναι καταλληλότερη όταν το μέγεθος των εποχικών διακυμάνσεων, ή η απόκλιση γύρω από την τάση-κύκλο, παραμένει σταθερό ανεξάρτητα από το επίπεδο της χρονοσειράς. Αντίθετα, όταν οι εποχικές μεταβολές ή οι διακυμάνσεις γύρω από την τάση αυξάνονται ή μειώνονται ανάλογα με το επίπεδο των δεδομένων, τότε προτιμάται η πολλαπλασιαστική αποσύνθεση (multiplicative decomposition) [30, 44, 42].

Μια εναλλακτική στην χρήση του πολλαπλασιαστικού μοντέλου είναι ο μετασχηματισμός των δεδομένων με τέτοιο τρόπο όπου η διακύμανση της χρονοσειρά έτσι ώστε να είναι σταθερή, με αποτέλεσμα να μπορεί να χρησιμοποιηθεί το προσθετικό μοντέλο, αφού αυτή είναι η βασική συνθήκη χρήσης του προσθετικού μοντέλου. Ο πιο συχνός τρόπος εφαρμογής του παραπάνω είναι ο λογαριθμικός μετασχηματισμός (log transformation). Δηλαδή, το πολλαπλασιαστικό μοντέλο ισοδυναμεί με το αθροιστικό εφόσον η χρονοσειρά, και κατ΄ επέκταση τα στοιχεία που την αποτελούν, έχουν μετατραπεί σε λογάριθμο, άρα ισχύει το παρακάτω: [44 , 42]

Πίνακας 40 Πολλαπλασιαστικό μοντέλο και μετατροπή του σε προσθετικό

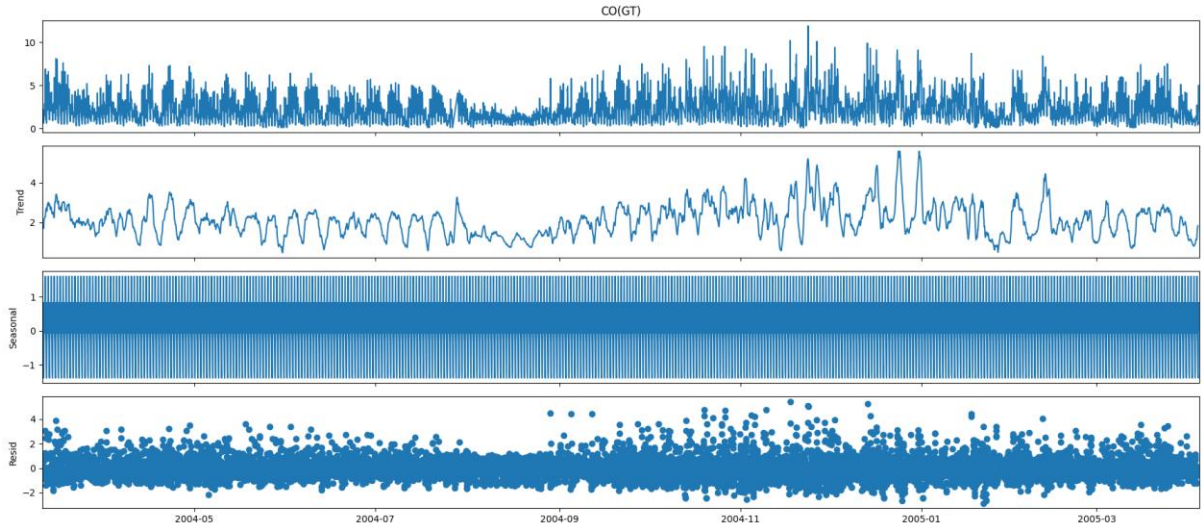
$y_t = S_t * T_t * R_t$ = $\log(y_t) = \log(S_t) + \log(T_t) + \log(R_t)$	(5)

Στην κλασσική αποσύνθεση, η βασική παραδοχή είναι ότι η εποχιακή «συνιστώσα» είναι σταθερή [37], δηλαδή ότι επαναλαμβάνεται σταθερά. Για αρκετές χρονοσειρές αυτό μπορεί να ισχύει, ωστόσο ειδικά σε χρονοσειρές με πολλές παρατηρήσεις (για παράδειγμα δεδομένα πολλών ετών) είναι πιθανό να μην ισχύει. Οι Hyndman και Athanasopoulos για παράδειγμα αναφέρουν ότι το μοτίβο της ζήτησης ηλεκτρικής ενέργειας έχει αλλάξει δραστικά λόγω της ευρείας χρήσης του κλιματισμού (air condition). Συγκεκριμένα, σε αρκετές τοποθεσίες η μέγιστη ζήτηση ηλεκτρικής ενέργειας εντοπιζόταν τον χειμώνα λόγω θέρμανσης ενώ πλέον το εποχικό μοτίβο έχει αλλάξει καθώς η μέγιστη ζήτηση παρατηρείται το καλοκαίρι λόγω κλιματισμού [37]. Βάσει των παραπάνω γίνεται αντιληπτό, ότι τα κλασσικά μοντέλα δεν μπορούν να διαχειριστούν αυτές τις εποχιακές μεταβολές. Ωστόσο, πέρα από τις μακροχρόνιες μεταβολές υπάρχουν και οι βραχυχρόνιες μεταβολές, οι οποίες μπορούν να προκαλέσουν εξίσου σημαντικές μεταβολές. Για παράδειγμα, η μηνιαία κίνηση αεροεπιβατών μπορεί να επηρεαστεί από μια απεργία του κλάδου ή την αύξηση τιμών των εισιτηρίων, με αποτέλεσμα την διαφοροποίηση της χρονοσειράς από τα κανονικά μεγέθη και μοτίβα της. Το κλασσικό μοντέλο αποσύνθεσης επίσης δεν μπορεί να διαχειριστεί αυτές τις βραχυπρόθεσμες μεταβολές [37].

Υπάρχουν άλλα μοντέλα αποσύνθεσης όπως STL, SEATS ή X-11. Ωστόσο, η μέθοδος STL παρουσιάζει αρκετά πλεονεκτήματα σε σχέση με τις υπόλοιπες μεθόδους. Αρχικά, το βασικό πλεονέκτημα της μεθόδου είναι ότι η εποχιακά συνιστώσα μπορεί να μεταβάλλεται στον χρόνο και αυτός ο ρυθμός αλλαγής μπορεί να προσαρμοστεί από τον χρήστη ανάλογα την περίπτωση υπό μελέτη. Παράλληλα, η μέθοδος STL μπορεί να διαχειριστεί όλους τους τύπους εποχικότητας – χρονικές κλίμακες, ενώ οι SEATS και X-11 μπορούν να εφαρμοστούν μόνο σε μηνιαία ή τετράμηνα δεδομένα. Η ομαλότητα της συνιστώσας τάσης-κύκλου (trend-cycle) μπορεί επίσης να ρυθμιστεί από τον χρήστη. Επιπλέον, η μέθοδος μπορεί να καταστεί ανθεκτική σε ακραίες τιμές (robust decomposition), ώστε μεμονωμένες ακραίες - ασυνήθιστες παρατηρήσεις να μην επηρεάζουν σημαντικά την εκτίμηση της τάσης και της εποχικής συνιστώσας, αλλά να αποτυπώνονται κυρίως στο υπόλοιπο (remainder component). Ωστόσο, η μέθοδος STL παρουσιάζει ορισμένους περιορισμούς. Συγκεκριμένα, δεν λαμβάνει αυτόματα υπόψη επιδράσεις όπως εργάσιμες ημέρες (trading days) ή ημερολογιακές μεταβολές και υποστηρίζει μόνο προσθετικές αποσυνθέσεις. Μια πολλαπλασιαστική αποσύνθεση μπορεί να προκύψει έμμεσα, εάν προηγηθεί λογαριθμικός μετασχηματισμός των δεδομένων και στη συνέχεια γίνει αντίστροφος μετασχηματισμός των συνιστωσών (back - transformation) [38].

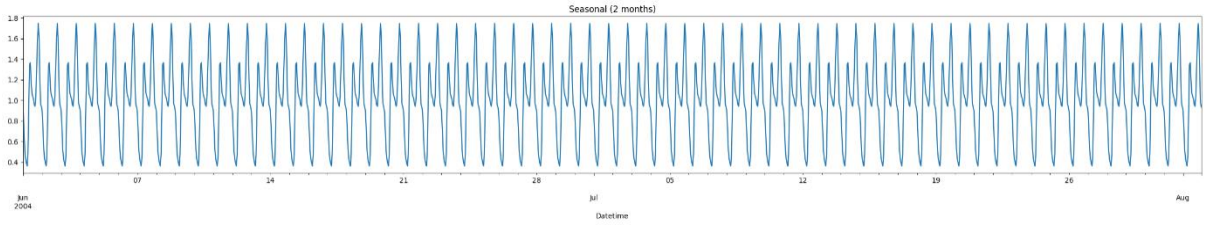
Συνοψίζοντας, η βασική διαφορά μεταξύ classical decomposition και STL decomposition, είναι ότι με την χρήση classical decomposition, η εποχικότητα θεωρείται σταθερή ενώ στην STL δεν είναι σταθερή και μπορεί να αλλάζει δραστικά από περίοδο σε περίοδο. Παράλληλα η STL δεν υποστηρίζει πολλαπλασιαστικές αποσυνθέσεις. Βάσει όλων των παραπάνω, αποφασίστηκε να εφαρμοστεί αρχικά

κλασσική προσθετική αποσύνθεση και στην συνέχεια STL αποσύνθεση. Για την υλοποίηση των παραπάνω μπορεί να χρησιμοποιηθεί η βιβλιοθήκη “statsmodels” μέσω των μεθόδων “seasonal_decompose” και “STL”.



Διάγραμμα 43 Κλασσική αποσύνθεση με την χρήση προσθετικού μοντέλου

Στο διάγραμμα 43, παρουσιάζεται η κλασσική προσθετική αποσύνθεση της χρονοσειράς. Αρχικά, δεν παρατηρείται μια μονότονη αύξηση ή μείωση της τάσης αλλά παρατηρείται μια σταθερή μείωση προς τους καλοκαιρινούς μήνες, με σχετικά σταθερή διακύμανση. Ύστερα, παρατηρείται μια σημαντική αύξηση τους χειμερινούς μήνες με σημαντικά υψηλότερη διακύμανση - μεταβλητότητα από τους καλοκαιρινούς μήνες. Από το παραπάνω μπορεί να γίνει αντιληπτό ότι η τάση μεταβάλλεται εποχικά και όχι τυχαία. Παράλληλα, η επιλογή ενός βραχυπρόθεσμου χρονικού ορίζοντα για ένα προβλεπτικό μοντέλου (ωρών μέχρι μερικών ημερών) θα παρήγαγε σημαντικά καλύτερα αποτελέσματα, αφού η τάση είναι πιο ομαλή βραχυπρόθεσμα. Στην συνέχεια, παρατηρείται η εποχικότητα της χρονοσειράς, η οποία δεν φαίνεται ξεκάθαρα και οριακά μοιάζει με μια πυκνή γραμμή. Όπως προαναφέρθηκε, η κλασσική αποσύνθεση θεωρεί την εποχικότητα σταθερή και λόγω του όγκου των παρατηρήσεων δεν εμφανίζεται κατάλληλα. Η εποχικότητα μπορεί να απομονωθεί για ένα μικρότερο χρονικό παράθυρο, ώστε να γίνει πιο ξεκάθαρη, πράγμα που παρουσιάζεται παρακάτω.

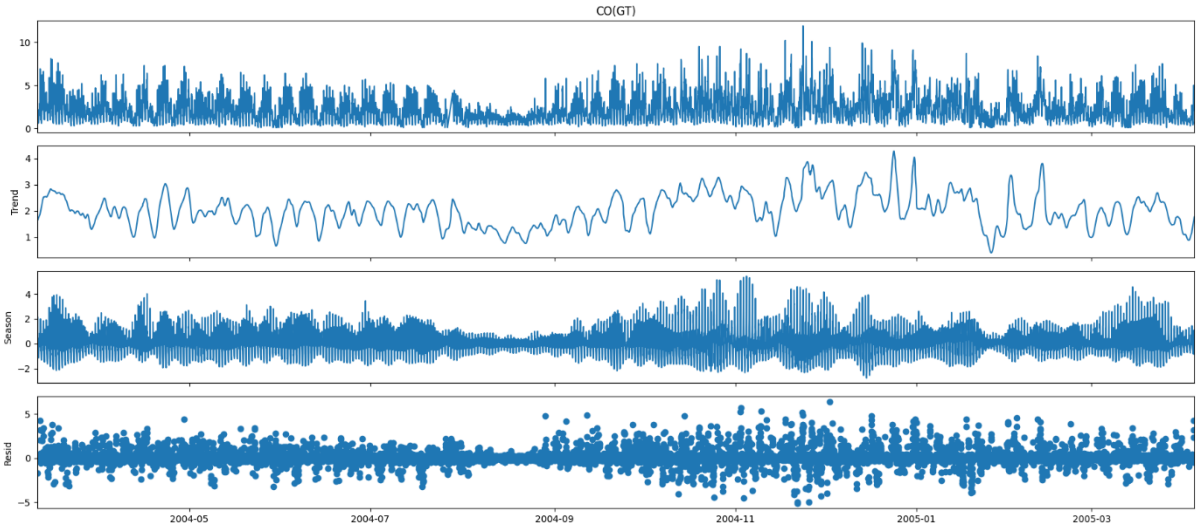


Διάγραμμα 44 Απεικόνιση εποχικότητας για 2 μήνες

Το διάγραμμα 44, παρουσιάζει το εποχιακό μοτίβο για την περίοδο 2004-06-01 έως 2004-08-01, δηλαδή τους καλοκαιρινούς μήνες Ιούνιο μέχρι Αύγουστο, που επιλέχτηκαν αυθαίρετα καθώς η εποχικότητα θα είναι ίδια για όλους τους μήνες, όπως προαναφέρθηκε. Με το διάγραμμα να είναι πλέον σε κατάλληλη μορφή, γίνεται αντιληπτό ότι η εποχικότητα αναφέρεται στο ημερήσιο διφασικό μοτίβο (diurnal) [60], με την χαμηλή κορυφή να ανταποκρίνεται στις πρωινές ώρες και την υψηλή

κορυφή στις βραδινές ώρες, όπως στο διάγραμμα 36. Ωστόσο, ακριβώς επειδή το εποχικό μοτίβο θεωρείται σταθερό οι μεταβολές που οφείλονται στα καιρικά – εποχικά φαινόμενα δεν μπορούν να ενσωματωθούν και πιθανώς να εισάγονται είτε στην τάση είτε στα κατάλοιπα.

Όσον αφορά τα κατάλοιπα, φαίνεται να είναι σχετικά συγκεντρωμένα γύρω από το 0, με μεγάλη διασπορά τους χειμερινούς μήνες και λιγότερη τους λοιπούς μήνες, ανάμεσα -2 και 2 που αναδεικνύει την ετεροσκεδαστικότητα των δεδομένων. Παράλληλα, παρατηρούνται και αρκετές ακραίες τιμές άνω του προαναφερόμενου περιθωρίου. Συνοψίζοντας τα παραπάνω, το προσθετικό κλασσικό μοντέλο αφήνει ένα σημαντικό μέρος της μεταβλητότητας της χρονοσειράς ανεξήγητη.



Διάγραμμα 45 STL αποσύνθεση

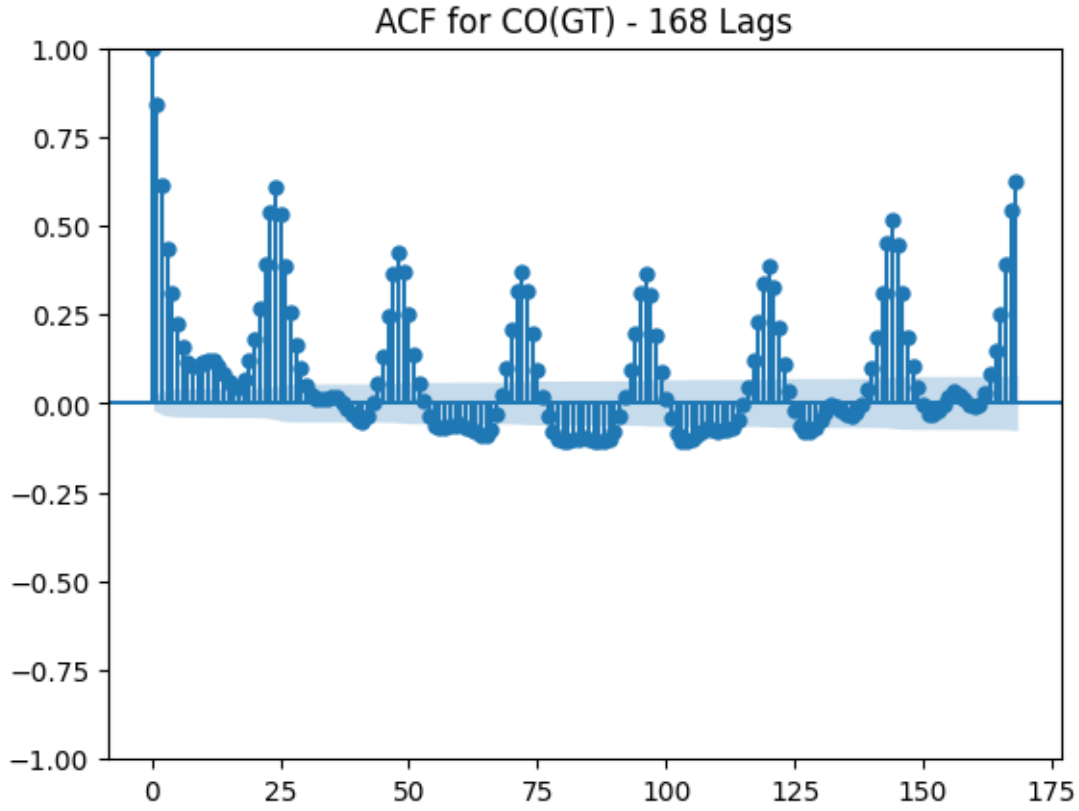
Το διάγραμμα 45, παρουσιάζει την STL αποσύνθεση. Αρχικά, σε αντίθεση με το διάγραμμα της κλασικής αποσύνθεσης παρατηρείται ότι η τάση είναι σημαντικά πιο ομαλή και φαίνεται να είναι σχετικά συγκεντρωμένη γύρω από το 2 ενώ παράλληλα συνεχίζουν να παρατηρούνται πιο ελαφρές αυξήσεις και μειώσεις για τους χειμερινούς και καλοκαιρινούς μήνες. Ωστόσο, η βασική διαφορά μεταξύ των 2 μεθόδων εντοπίζεται στην εποχικότητα. Οι μεταβολές της εποχικότητας φαίνεται να απορρόφησαν τις εποχικές μεταβολές που εντοπίζονταν στην τάση, αναδεικνύοντας ότι πράγματι υπάρχει σημαντική διαφοροποίηση στην εποχικότητα. Μάλιστα, επιβεβαιώνεται ότι οι χειμερινοί μήνες έχουν μεγαλύτερο εύρος μεταβολών ενώ οι καλοκαιρινοί έχουν σχετικά μηδαμινό εύρος. Τα κατάλοιπα επίσης παρουσιάζουν βελτίωση καθώς φαίνονται πιο συγκεντρωμένα γύρω από το μηδέν. Κατά την άνοιξη φαίνεται να υπάρχει μεγαλύτερο εύρος ωστόσο φαίνεται να είναι σχετικά σταθερό. Σε αντίθεση, οι χειμερινοί μήνες παρουσιάζουν σημαντικά μεγαλύτερο εύρος ενώ παράλληλα εντοπίζονται περισσότερες ακραίες τιμές και σημαντικά πιο ασταθές εύρος τιμών.

3.10 Διαγνωστικά διαγράμματα ACF - PACF

Η Συνάρτηση Αυτοσυσχέτισης (Autocorrelation Function - ACF) μετρά τον βαθμό συσχέτισης μιας χρονοσειράς με τις προηγούμενες τιμές της σε διαφορετικές χρονικές υστερήσεις (lags) και χρησιμοποιείται για την εύρεση των συνιστωσών για διαδικασίες MA (δηλαδή το q ή Q). Η Μερική Συνάρτηση Αυτοσυσχέτισης (Partial Autocorrelation Function - PACF) αξιολογεί τη συσχέτιση μεταξύ μιας χρονοσειράς και των lags της, αφαιρώντας την επίδραση των ενδιάμεσων χρονικών υστερήσεων και χρησιμοποιείται για την εύρεση των συνιστωσών για διαδικασίες AR(p ή P) [58]. Οι τιμές και των δύο συναρτήσεων μπορούν να λάβουν τιμές μεταξύ -1 και 1.

Τα διαγνωστικά διαγράμματα ACF – PACF, μπορούν να χρησιμοποιηθούν για τον έλεγχο στασιμότητας της χρονοσειράς αλλά και για την εύρεση των κατάλληλων μοντέλων και των παραμέτρων τους [30].

Ήδη έχει αναδειχθεί ότι τα δεδομένα παρουσιάζουν ημερήσιο μοτίβο (diurnal cycle), επομένως τα πιο σημαντικά μοτίβα βρίσκονται εντός 24 ωρών ή lags. Βάσει αυτού θεωρήθηκε ορθό να δημιουργηθούν διαγράμματα ACF – PACF με την χρήση μεγαλύτερων lags, ώστε να ερευνηθεί ένα υπάρχουν άλλα σημαντικά μοτίβα εκτός του ημερήσιου κύκλου. Συγκεκριμένα, χρησιμοποιήθηκαν τα lags 48 για την επιβεβαίωση του ημερήσιου μοτίβου και 168 για τον έλεγχο ύπαρξης εβδομαδιαίου μοτίβου, δηλαδή 2 ημέρων και μιας εβδομάδας, τα οποία παρατίθενται παρακάτω.

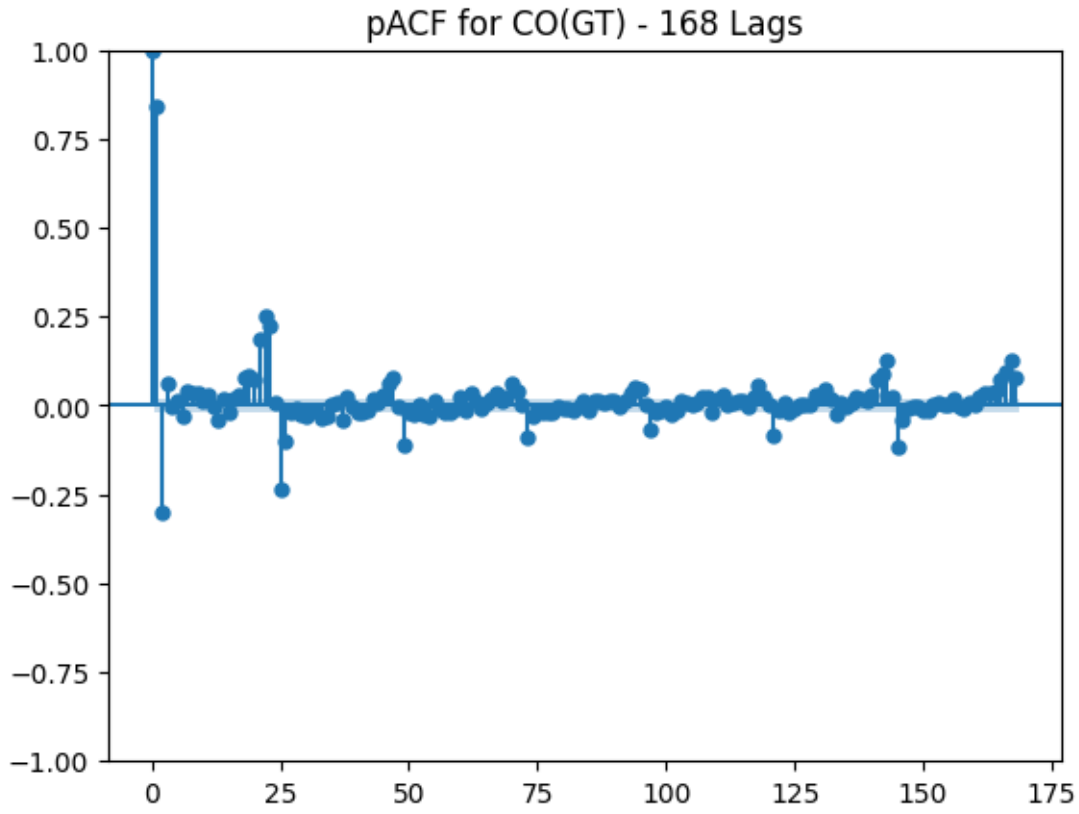


Διάγραμμα 46 Διάγραμμα ACF 168 Lags

Το διάγραμμα 46, παρουσιάζει το διάγραμμα ACF με την χρήση 168 lags. Παρατηρείται πολύ υψηλή συσχέτιση για lags 1,2,3 με μια σταδιακή μείωση. Ωστόσο, υπάρχει κυματοειδής επαναλαμβανόμενη μορφή, με spikes κάθε 24 ώρες , για παράδειγμα 24, 48, 72 lags, που επιβεβαιώνει την ύπαρξη του ημερήσιου μοτίβου ή αλλιώς «deterministic daily seasonality» [46], ωστόσο παρατηρείται ότι οι περισσότερες συσχετίσεις είναι μικρότερες ή ίσες του 50%. Η ύπαρξη του spike για lag 168 επιβεβαιώνει την ύπαρξη του εβδομαδιαίου μοτίβου, ωστόσο είναι χαμηλότερης σημασίας από αυτή του ημερήσιου μοτίβου.

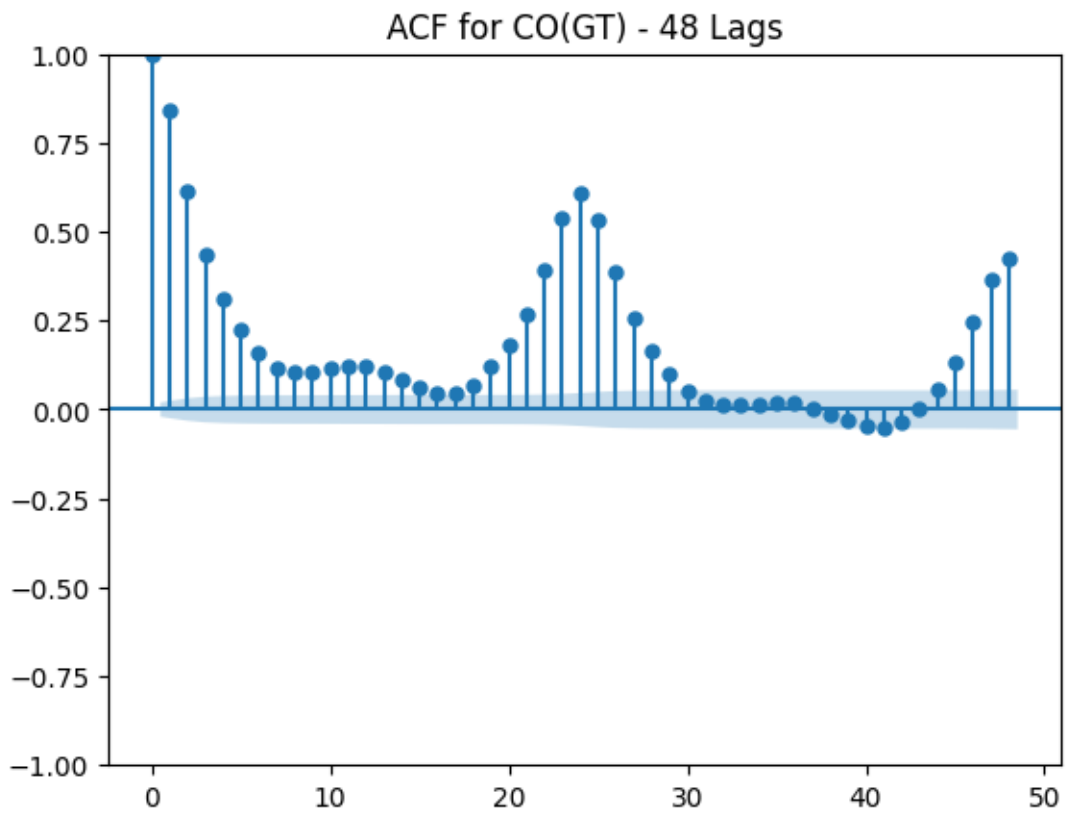
Παράλληλα, οι τιμές μεταξύ των κορυφών που παρατηρούνται μεταξύ κάθε 24 ώρες είναι είτε εντός του επιπέδου σημαντικότητας είτε είναι αρνητικές και χαμηλότερες του 25%, επομένως μπορούν να θεωρηθούν πρακτικά αμελητέες. Επίσης, παρατηρείται έλλειψη του «slow decay» σε συνδυασμό με την ύπαρξη μόνο θετικών συσχετίσεων και τα spikes να είναι κάτω του 50%.

Από τα παραπάνω επιβεβαιώνονται τα αποτελέσματα του ADF Test, δηλαδή ότι δεν υπάρχει «stochastic trend». Βάσει των παραπάνω, η παράμετρος q πιθανώς θα έχει τιμή 1 με μέγιστο 2.



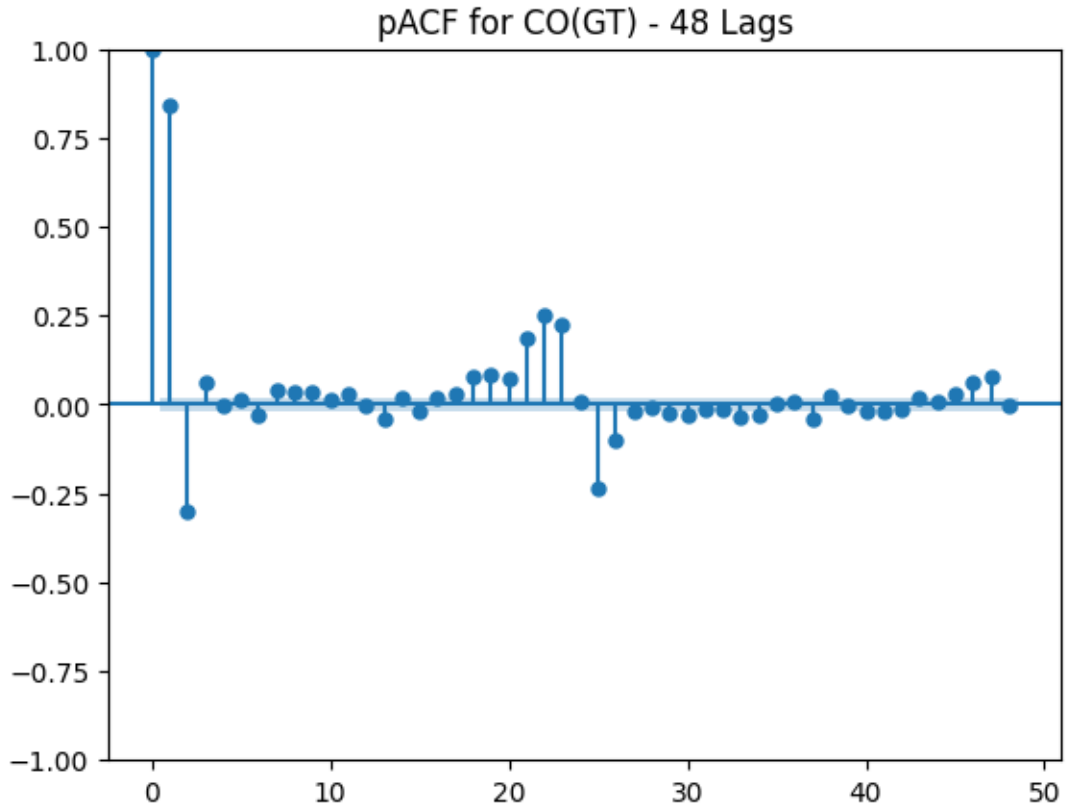
Διάγραμμα 47 Διάγραμμα PACF 168 lags

Στο διάγραμμα 47 παρουσιάζεται το διάγραμμα PACF με την χρήση 168 lags. Παρατηρείται πολύ υψηλή συσχέτιση για lags 1 και 2. Για lag 3 παρατηρείται αρνητική και χαμηλή αυτοσυσχέτιση. Για τα υπόλοιπα lags, οι τιμές κυμαίνονται γύρω από το 0, ωστόσο για κάθε 24 ώρες παρατηρούνται αρνητικά spikes χαμηλού μεγέθους, που φθίνουν για κάθε επανάληψη. Όλα τα προαναφερόμενα συνάδουν με τις παρατηρήσεις που εντοπίστηκαν για τα διαγράμματα ACF και παράλληλα αναδεικνύουν p = 1 ή μέγιστο 2.



Διάγραμμα 48 Διάγραμμα ACF 48 Lags

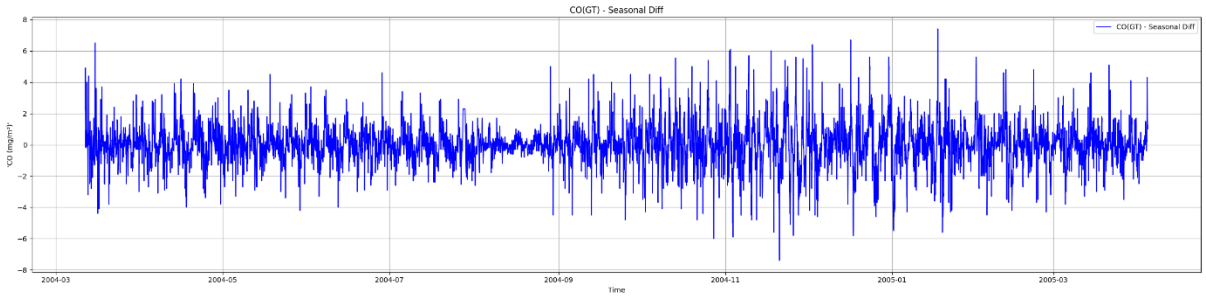
Το διάγραμμα 48, παρουσιάζει το διάγραμμα ACF με την χρήση 48 lags, και προσφέρεται κυρίως για ευκρίνεια σε συνδυασμό με την εικόνα Χ. Ισχύουν οι ίδιες παρατηρήσεις που έγιναν, δηλαδή τα πρώτα 2 lags παρουσιάζουν πολύ σημαντική συσχέτιση και ύστερα παρατηρείται μια φθίνουσα τάση για τα μεγέθη μέχρι περίπου το 20^{ο} lag και την μετέπειτα σημαντική κορυφή που παρατηρείται για το 24ο lag.



Διάγραμμα 49 Διάγραμμα PACF 48 Lags

Με παρόμοιο τρόπο, το διάγραμμα49 παρουσιάζει το διάγραμμα του PACF με την χρήση 48 lags και προσφέρεται συμπληρωματικά στο διάγραμμα 47. Ισχύουν όλες οι προαναφερόμενες παρατηρήσεις που έγινε στο διάγραμμα 47.

Όμως, όπως προαναφέρθηκε εάν τα αποτελέσματα του ADF τεστ δείχνουν στασιμότητα ενώ τα αποτελέσματα του KPSS δείχνουν μη – στασιμότητα, τότε προτείνετε η εφαρμογή διαφορών ώστε η χρονοσειρά να γίνει στάσιμη [54], με τα αποτελέσματα της εφαρμογής εποχιακών διαφορών να παρουσιάζονται παρακάτω.



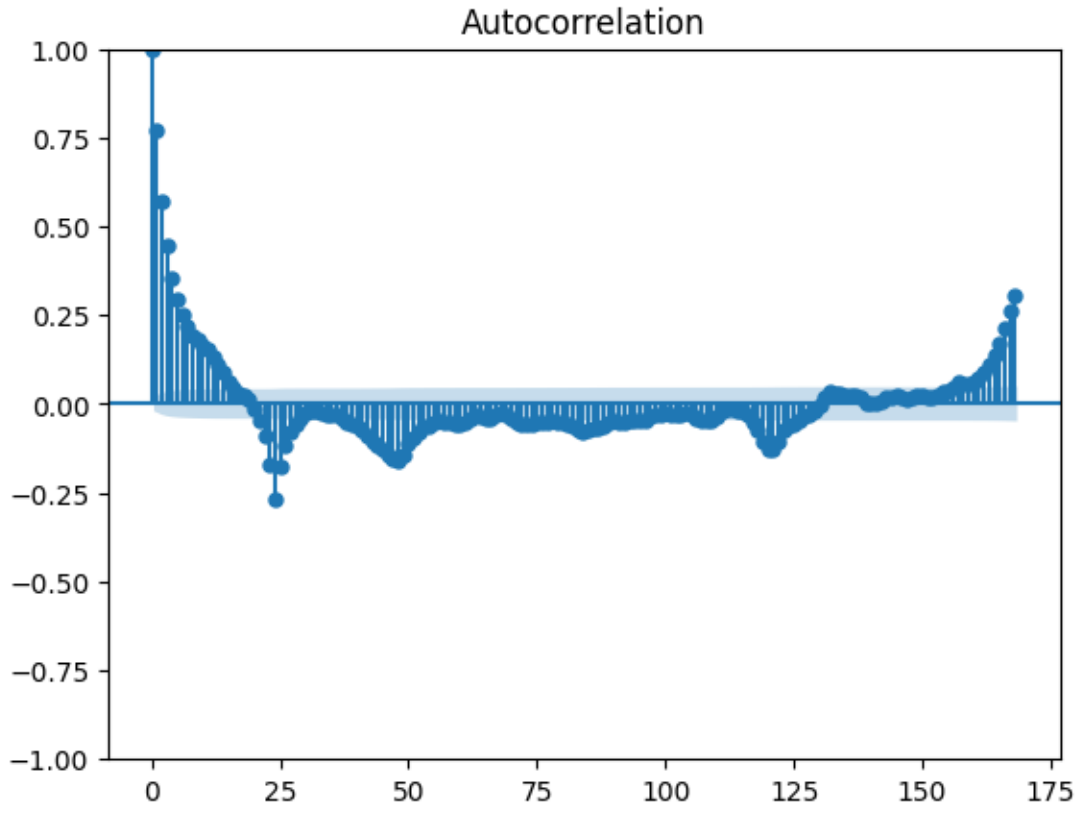
Διάγραμμα 50 Απεικόνιση χρονοσειράς ύστερα από χρήση εποχιακών διαφορών

Όπως φαίνεται στο διάγραμμα 50, πλέον οι τιμές της χρονοσειράς βρίσκονται γύρω από το 0. Η διακύμανση δεν φαίνεται να είναι πλήρως σταθερή, ωστόσο φαίνεται σχετικά πιο σταθερή σε σχέση με το γράφημα της χρονοσειράς χωρίς εποχικές διαφορές.

Πίνακας 41 Αποτελέσματα KPSS – ADF Test για χρονοσειρά εποχιακών διαφορών

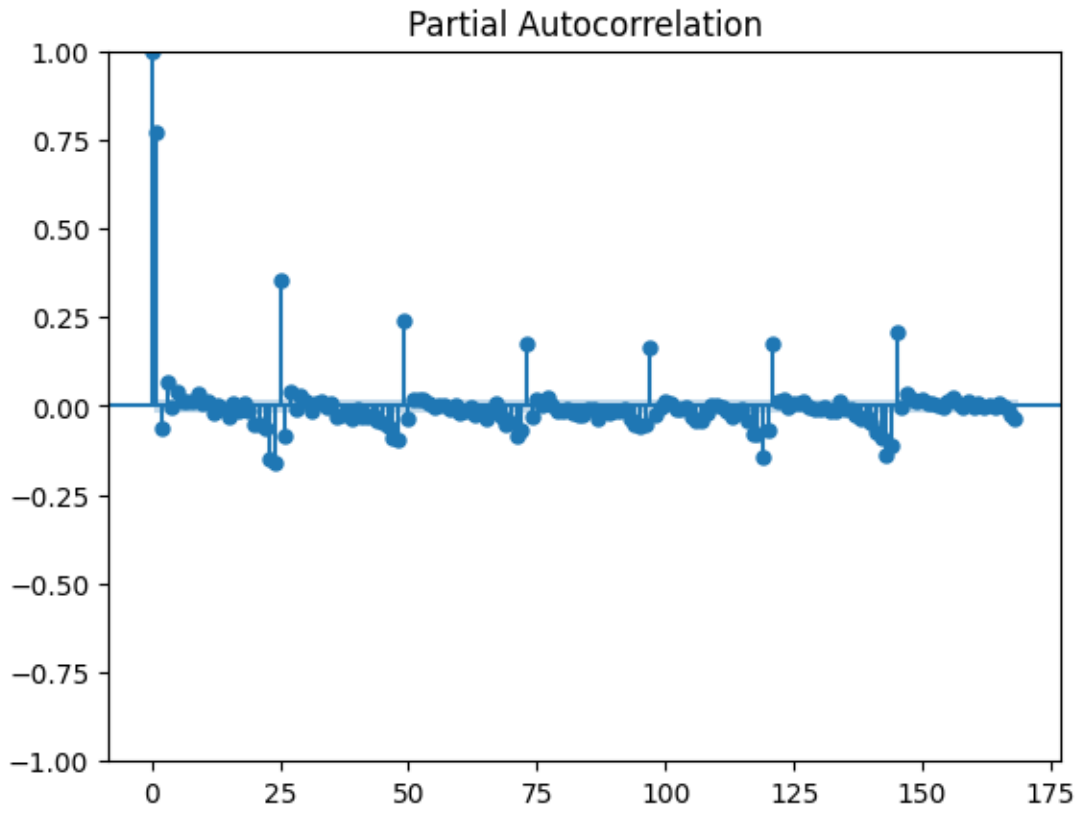
```
KPSS Statistic: 0.0033732216156595 59
p-value: 0.1
Critical Values: {'10%': 0.119, '5%': 0.146, '2.5%': 0.176, '1%': 0.216}
Lags used: 46
Fail to reject the null hypothesis (series is Trend-stationary)
----------------------------------------------------------
  Στατιστικό ADF: -16.0324
  p-value: 0.0000
  ✅ Στασιμή σειρά (χωρίς τάση)
----------------------------------------------------------
```

Από τον πίνακα 41, επιβεβαιώνεται ότι η χρονοσειρά είναι πλέον στάσιμη, καθώς το p-value του KPSS Test βρέθηκε ως 0,1 δηλαδή μεγαλύτερη του επίπεδου σημαντικότητας 0,05, άρα η αρχική υπόθεση δεν απορρίπτεται ενώ η p-value του ADF Test βρέθηκε ως 0, μικρότερη του 0,05 άρα η αρχική υπόθεση δεν απορρίπτεται. Επομένως, η χρονοσειρά είναι πράγματι στάσιμη.



Διάγραμμα 51 Διάγραμμα ACF για χρονοσειρά εποχιακών διαφορών 168 lags

Στο διάγραμμα 51, παρουσιάζεται το διάγραμμα ACF για την χρονοσειρά πρώτων εποχιακών διαφορών. Φαίνεται ότι έχει αφαιρεθεί το παλινδρομικό μοτίβο που παρατηρούνταν σε συνδυασμό με τις πολλαπλές κορυφές που παρουσιάζονταν κάθε 24 ώρες. Περίπου μέχρι την 20η υστέρηση – lag, παρατηρείται σταθερή μείωση και στο 24 παρατηρείται για αρνητική συσχέτιση περίπου -0,25. Παρόλο αυτό, θεωρείται πιθανός ο συντελεστής P = 1 ή 2.



Διάγραμμα 52 Διάγραμμα PACF για χρονοσειρά εποχιακών διαφορών 168 lags

Στο διάγραμμα 52 παρουσιάζεται το γράφημα PACF για την χρονοσειρά πρώτων εποχιακών διαφορών. Οι κορυφές που εντοπίζονταν κάθε 24 ώρες πλέον είναι θετικές και μεγαλύτερες σε μέγεθος. Επομένως, Q = 1 ή 2.

Όπως αναφέρουν οι Niako et al. [30], η τελική επιλογή ως προς την μέθοδο για την ανάλυση μιας χρονοσειράς εξαρτάται από τον ερευνητή και από την φύση των δεδομένων. Συνήθως για μονοδιάστατες χρονοσειρές (univariate time-series data) χρησιμοποιούνται στατιστικές μέθοδοι όπως exponential smoothing όπως Holt – Winters ή μοντέλα ARIMA. Όσο αφορά χρονοσειρές που παρουσιάζουν εποχικότητα, χρησιμοποιούνται εποχιακά μοντέλα ARIMA, γνωστά ως “Seasonal ARIMA” ή SARIMA, τα οποία παρουσιάζουν καλές μετρικές αποδόσεις σε προβλέψεις με βραχυπρόθεσμο ορίζοντα πρόβλεψης ενώ παράλληλα είναι πιο εύκολα σε κατανόηση από πιο πολύπλοκα μοντέλα όπως LSTM [30].

Βάσει όλων των παραπάνω, όσο αφορά στατιστικά μοντέλα, θεωρείται ότι η βέλτιστη μέθοδος είναι είτε SARIMA είτε SARIMAX. Θεωρείται ότι δεν έχει νόημα να εφαρμοστούν μέθοδοι ARIMA – ARIMAX, καθώς οι συγκεκριμένες μέθοδοι δεν είναι οι βέλτιστες για δεδομένα όπου η εποχικότητα επηρεάζει σε σημαντικό βαθμό την χρονοσειρά.

Παράλληλα, χρειάζεται να αναφερθεί ότι δεν είναι ξεκάθαρη η βέλτιστη μέθοδος προς το παρόν. Η χρονοσειρά παρουσιάζει σημαντική εποχικότητα κάθε 24 ώρες, με ένα από τα βασικά ερωτήματα να είναι εάν πρέπει ή όχι να εφαρμοστούν εποχιακές διαφορές. Πιο συγκεκριμένα, χρειάζεται να ερευνηθεί εάν η αφαίρεση της εποχικότητας, μέσω εποχιακών διαφορών, μπορεί να παράγει καλύτερα προβλεπτικά μοντέλα, δηλαδή εάν η εποχικότητα είναι θόρυβος ή δομή του σήματος. Για την διερεύνηση του παραπάνω, μπορεί να γίνει σύγκριση των μετρικών MAE / RMSE μεταξύ SARMA (1,0,1)(1,0,1), δηλαδή ενός μοντέλου που ενσωματώνει την εποχικότητα για την πρόβλεψη ή SARIMA

(1,0,1)(1,1,1), δηλαδή ενός μοντέλου που αφαιρεί την εποχικότητα μέσω εποχιακών διαφορών. Σκοπός δεν είναι απαραίτητα η εύρεση του βέλτιστου μοντέλου πρόβλεψης αλλά η σύγκριση των διαφορετικών προσεγγίσεων multi – step και single – step και για να είναι εφικτή η σύγκριση των αποτελεσμάτων, οι συντελεστές θα διατηρούνται σταθεροί.

4 Αποτελέσματα

4.1 Στατιστικά Μοντέλα: SARIMA – SARIMAX

Το μοντέλο ARIMA αποτελείται από τρία βασικά στοιχεία: το αυτοπαλίνδρομο μέρος (AR - (p)), τον βαθμό ολοκλήρωσης – διαφορών (I – (d)) και το μέρος κινητού μέσου (MA - (q)). Το αυτοπαλίνδρομο σκέλος AR(p) βασίζεται σε ένα μοντέλο παλινδρόμησης, στο οποίο οι προηγούμενες p τιμές της χρονοσειράς (yt−1 έως yt−p) χρησιμοποιούνται ως predictors για την πρόβλεψη της τρέχουσας τιμής yt [30]. Η τιμή d εκφράζει πόσες φορές πρέπει να διαφοροποιηθεί η χρονοσειρά ώστε να καταστεί στάσιμη. Το μέρος κινητών μέσων MA(q) χρησιμοποιεί τα q σφάλματα πρόβλεψης προηγούμενων χρονικών στιγμών σε ένα παλινδρομικό μοντέλο αντί των περασμένων τιμών του γνωρίσματος υπό μελέτη [30]. Για συντομία, η μέθοδος εμφανίζεται ως ARIMA (p,d,q). Ένα μοντέλο AR (p), αναπαρίσταται ως [30]:

Πίνακας 42 Εξίσωση μοντέλου AR

$Y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t$	(6)

Όπου Y_t μια στάσιμη (stationary) μεταβλητή, c μια σταθερά τιμή και έστω φ_i, όπου i = 1 έως p, οι συντελεστές αυτοσυσχέτισεις για τα αντίστοιχα lags 1 έως p και ε_t τα κατάλοιπα – μια σειρά λευκού θορύβου με μέση τιμή το μηδέν και διακύμανση σ^2. Επίσης, ένα μοντέλο MA (q) αναπαρίσταται ως [30, 64]:

Πίνακας 43 Εξίσωση μοντέλου MA

$Y_t = \mu + \theta_0 \varepsilon_t + \cdots + \theta_q \varepsilon_{t-q}$ *ή διαφορετικά:* $Y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$	(7)

Όπου Y_t μια στάσιμη μεταβλητή, $\mu = E(Y_t)$, δηλαδή c + ε_t, θ_j, όπου j = 1 έως q, c μια σταθερά και ε_t τα κατάλοιπα ή όπως προαναφέρθηκε, μια σειρά λευκού θορύβου με μέση τιμή το μηδέν και διακύμανση σ^2. Δηλαδή, το μοντέλο AR χρησιμοποιεί προηγούμενες τιμές του γνωρίσματος υπό μελέτη, σε μια παλινδρόμηση ενώ το μοντέλο MA χρησιμοποιεί τις τιμές σφάλματος των παρατηρήσεων ή αλλιώς τα κατάλοιπα σε μια παλινδρόμηση [64].

Επομένως, συνδυάζοντας τα παραπάνω, προκύπτει για μοντέλα ARMA (p,q) [30]:

Πίνακας 44 Εξίσωση μοντέλου ARMA

$Y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t + \theta_0 \varepsilon_t + \cdots + \theta_q \varepsilon_{t-q}$	(8)

Όπου ισχύει $\varphi_i \neq 0$, i = 1 εώς p, $\theta_j \neq 0$, j = 0 εως q και $\sigma^2 > 0$.

Όπως με τα περισσότερα μοντέλα, υπάρχουν ορισμένες παραδοχές - περιορισμοί που πρέπει να τηρούνται για την χρήση του αλγόριθμου ARIMA. Συγκεκριμένα, η χρονόσειρά που θα εφαρμοστεί ο αλγόριθμος πρέπει να είναι στάσιμη (stationary) και τα σφάλματα – κατάλοιπα να μην συσχετίζονται μεταξύ τους ενώ παράλληλα να παρουσιάζουν σταθερή διακύμανση, δηλαδή να παρουσιάζουν ομοσκεδαστικότητα. Δηλαδή όπως προαναφέρθηκε, μια στάσιμη χρονοσειρά ή αλλιώς μια διεργασία λευκού θορύβου είναι αυτή που η δομή πιθανών τιμών της δεν αλλάζει σημαντικά για το χρονικό διάστημα παρατήρησης, που παράλληλα παρουσιάζει σταθερή μέση τιμή [30].

Στις χρονοσειρές που παρουσιάζουν εποχικότητα εφαρμόζεται το εποχιακό μοντέλο ARIMA, γνωστό ως SARIMA (Seasonal Autoregressive Integrated Moving Average), το οποίο ενσωματώνει εποχιακούς και μη - εποχιακούς παράγοντες. Παρουσιάζει μορφή συντελεστών (p, d, q) × (P, D,Q)S, όπου οι συντελεστές (p, d, q) αφορούν το μη εποχιακό μέρος και οι συντελεστές (P, D,Q) το εποχιακό μέρος και όπου S η χρονική κλίμακα της χρονοσειράς [30].

Αρχικά, θεωρήθηκε χρήσιμο να εφαρμοστεί αυτοματοποιημένη μέθοδος εύρεσης βέλτιστων συντελεστών, τύπου grid-search ως προς τα μοντέλα SARIMA. Για την εκτέλεση του παραπάνω σκοπού χρησιμοποιήθηκε η βιβλιοθήκη “pmdarima” και συγκεκριμένα το Module “AutoArima”[47]. Σχετικά με τις παραμέτρους που χρησιμοποιήθηκαν, ως “y” εισάχθηκε το train split που αναφέρθηκε προηγουμένως. Η συνάρτηση προσφέρει την δυνατότητα εισαγωγής εξωγενών δεδομένων, δηλαδή πρακτικά την εφαρμογή ARIMAX ή SARIMAX, ωστόσο δεν χρησιμοποιήθηκε στην συγκεκριμένη περίπτωση καθώς θεωρήθηκε ότι δεν υπήρχαν διαθέσιμοι οι απαραίτητοι υπολογιστικοί πόροι, πράγμα το οποίο θα επιβεβαιωθεί και παρουσιαστεί αργότερα. Στην συνέχεια, βάσει του ADF Test και των διαγραμμάτων ACF – PACF, θεωρείται ότι δεν είναι αναγκαία η χρήση πρώτων ή εποχιακών διαφορών απευθείας, επομένως οι παράμετροι d (για τις πρώτες διαφορές) και D (για τις εποχιακές διαφορές) θα οριστούν αρχικά ως 0. Παράλληλα ορίζεται ως “m” ίσο με 24, δηλαδή ορίζεται ότι τα δεδομένα είναι ημερήσιας χρονικής κλίμακας και η παράμετρος “seasonal”, ορίζεται ως αληθής αφού σκοπός είναι η εύρεση του βέλτιστου SARIMA μοντέλου. Στην συνέχεια ορίζεται το πλαίσιο παραμέτρων (grid) των εποχιακών και μη εποχιακών παραμέτρων p,q,P,Q, όπου για όλα ορίζονται ως ελάχιστη τιμή 0 και ως μέγιστη 2 μέσω των παραμέτρων start_X, max_X, όπου X οι αντίστοιχες παράμετροι. Η παράμετρος “stationary” ορίζεται ως True, αφού το ADF Test απέδειξε ότι η χρονοσειρά είναι στάσιμη ενώ παράλληλα η παράμετρος “test” που αφορά την πραγματοποίηση κάποιου unit root test (όπως ADF ή KPSS) για τον έλεγχο στασιμότητας ορίζεται ως “None”. Τέλος, η παράμετρος “stepwise” ορίζεται ως αληθής, αφού ο συγκεκριμένος αλγόριθμος είναι σημαντικά πιο γρήγορος ως προς την εύρεση των βέλτιστων παραμέτρων σε σχέση για παράδειγμα με την χρήση τυχαίων παραμέτρων. Τα αποτελέσματα από την χρήση της προαναφερόμενης συνάρτησης παρουσιάζονται παρακάτω στον πίνακα 45, ωστόσο η συνάρτηση δεν μπορεί να ολοκληρωθεί λόγω έλλειψης RAM, με αποτέλεσμα την προσωρινή ανάκληση του VM υπό χρήση.

Πίνακας 45 Αποτελέσματα auto-pmdarima

Μοντέλο	AIC	Χρόνος
(0,0,0)(0,0,0)[24]	33515.330	0.28
(1,0,0)(1,0,0)[24]	19855.128	26.68
(0,0,1)(0,0,1)[24]	23876.117	16.83
(0,0,0)(0,0,0)[24]	44135.643	0.12
(1,0,0)(0,0,0)[24]	22138.617	0.52
(1,0,0)(2,0,0)[24]	19461.610	99.97

(1,0,0)(2,0,1)[24]	inf	244.80
(1,0,0)(1,0,1)[24]	inf	56.62
(0,0,0)(2,0,0)[24]	29129.569	58.67

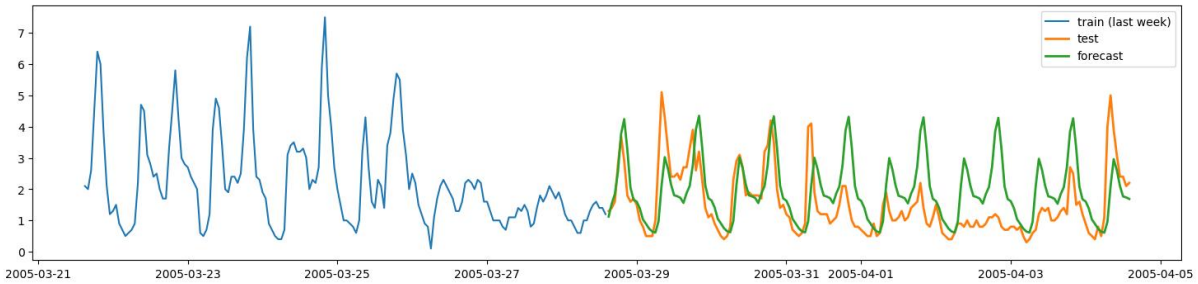
Επομένως, θα δημιουργηθούν ξεχωριστά συναρτήσεις για εφαρμογή μοντέλων SARIMA. Σχετικά με τους συντελεστές του αλγόριθμου, χρειάζεται να αναφερθεί ότι βάσει των διαγραμμάτων ACF και PACF του προηγούμενου κεφάλαιού δυνητικά μπορούν να χρησιμοποιηθούν μεγαλύτερου μεγέθους συντελεστές από αυτούς που αναφέρθηκαν. Ωστόσο, γενικά προτείνεται η χρήση χαμηλών συντελεστών, για λόγους απλότητας – χρόνου εκτέλεσης, και σε περίπτωση που τα αποτελέσματα δεν είναι ικανοποιητικά, μπορεί να γίνει σταδιακή αύξηση στους συντελεστές. Επίσης, όπως αναφέρουν και οι Mancini et al. [29], συχνά η χρήση υψηλών συντελεστών δεν επιφέρει αρκετές αλλαγές στα αποτελέσματα που να δικαιολογούν την πολυπλοκότητα του μοντέλου καθώς για παράδειγμα στην έρευνα τους τα μοντέλα APIMA(0,1,1) και ARIMA(14,1,14) παρουσίαζαν παρόμοια απόδοση [29].

Για την υλοποίηση των “multi step” αλλά και των “single step” προσεγγίσεων, μπορεί να χρησιμοποιηθεί η μέθοδος “SARIMAX” [73] της βιβλιοθήκης “statsmodels” [74]. Η διαδικασία εφαρμογής των μοντέλων είναι σχετικά απλή. Έστω η μεταβλητή “res = model.fit()”, που ανταποκρίνεται σε ένα αντικείμενο “SARIMAXResults” [75]. Έπειτα έστω “pred = res.get_forecast(…)”, δηλαδή εφαρμογή της μεθόδου “get_forecast” [76] στο προηγούμενο αντικείμενο, που επιστρέφει ένα αντικείμενο “PredictionResults” [77]. Από το αντικείμενο μπορεί να υπάρξει πρόσβαση στην ιδιότητα – χαρακτηριστικό (attribute) “predicted_mean” [77]. Καθώς κάθε βήμα στον προβλεπτικό ορίζοντα, αντιστοιχεί σε μια κατανομή πιθανοτήτων [68], επομένως αντιστοιχούν σε μια μέση τιμή και ένα τυπικό σφάλμα (μ ± 1.96σ για διάστημα εμπιστοσύνης 95%) [68], όπως για παράδειγμα ότι οι πιθανές τιμές βρίσκονται εντός του διαστήματος 2 ± 2 [79]. Επομένως, μέσω του “predicted_mean”, υπάρχει πρόσβαση στην κεντρική – μέση τιμή, όπου στο παράδειγμα αντιστοιχεί το 2. Άρα, μέσω του παραπάνω δημιουργείται ένα μονοδιάστατο διάνυσμα 1 x N, όπου N το μήκος των βημάτων για την πρόβλεψη, όπου κάθε τιμή αντιστοιχεί στην μέση τιμή. Μέσω αυτού του διανύσματος, μπορούν να γίνουν συγκρίσεις με τις πραγματικές – παρατηρούμενες τιμές ώστε να δημιουργηθούν οι μετρικές σφάλματος – αξιολόγησης MAE και RMSE.

Για την επίτευξη των προσεγγίσεων “single step” υπάρχει μια μικρή τροποποίηση της προαναφερόμενης μεθοδολογίας. Σκοπός είναι να μην υπάρχει συσσώρευση των σφαλμάτων μέσω της χρήσης προηγούμενων προβλέψεων ως τιμών εκπαίδευσης του μοντέλου και αντί αυτού να χρησιμοποιείται η πραγματική τιμή. Με την χρήση ενός “for loop” που θα εκτελείται για το σύνολο – μήκος το χρονικού ορίζοντα πρόβλεψης, κάθε επανάληψη θα αφορά ένα βήμα στον ορίζοντα – μια πρόβλεψη. Όπως και με την “multi step” μέθοδο, χρησιμοποιείται το χαρακτηριστικό “predicted_mean”, και εισάγεται στην λίστα “preds”, που αφορά τις προβλέψεις του μοντέλου. Στην συνέχεια, η πραγματική τιμή εισάγεται στην λίστα “res_upd”, ώστε για το επόμενο βήμα επανάληψης του “for_loop”, το μοντέλο να κάνει πρόβλεψη με βάση την τελευταία παρατηρούμενη τιμή και όχι την πρόβλεψη που παρήγαγε [69].

Για να συγκριθεί η απόδοση Multi – Step και Single – Step προσεγγίσεων, οι παράμετροι των μοντέλων θα παραμείνουν ίδιες. Ωστόσο, πρώτα χρειάζεται να ερευνηθεί εάν είναι καλύτερη η χρήση SARMA (1,0,1)(1,0,1) ή SARIMA (1,0,1)(1,1,1), καθώς δεν είναι ακόμα ξεκάθαρη η βέλτιστη προσέγγιση, όπως προαναφέρθηκε.

4.1.1 SARMA – Multi - Step



Διάγραμμα 53 Απόδοση Multistep SARMA

Το διάγραμμα 53 παρουσιάζει την εφαρμογή του μοντέλου Multi – Step (Recursive) SARIMA (1,0,1) x (1,0,1)24. Φαίνεται ότι το συγκεκριμένο μοντέλο εντοπίζει το βασικό ημερήσιο μοτίβο, δηλαδή την πρώτη κορυφή που παρατηρείται την πρωινή ώρα και την δεύτερη μεγαλύτερη κορυφή για τις βραδινές ώρες. Παράλληλα, δεν φαίνεται ότι μπορεί να αντιμετωπίσει οποιαδήποτε σημαντική διαφοροποίηση που μπορεί να προκύψει στο μοτίβο. Για παράδειγμα, μεταξύ 2005-04-02 και 2005-04-03, παρατηρείται ιδιαίτερα χαμηλή συγκέντρωση CO, με αποτέλεσμα η πρόβλεψη να απέχει σε σημαντικό βαθμό από τις πραγματικές παρατηρούμενες τιμές. Βάσει αυτών, φαίνεται ότι το μοντέλο εντοπίζει την μέση τιμή ανά ώρα και την επαναλαμβάνει.

Πίνακας 46 Μετρικές σφάλματος Multistep SARMA

Mean Absolute Error	Root Mean Squared Error
0.8187566751949479	1.0850912831285437

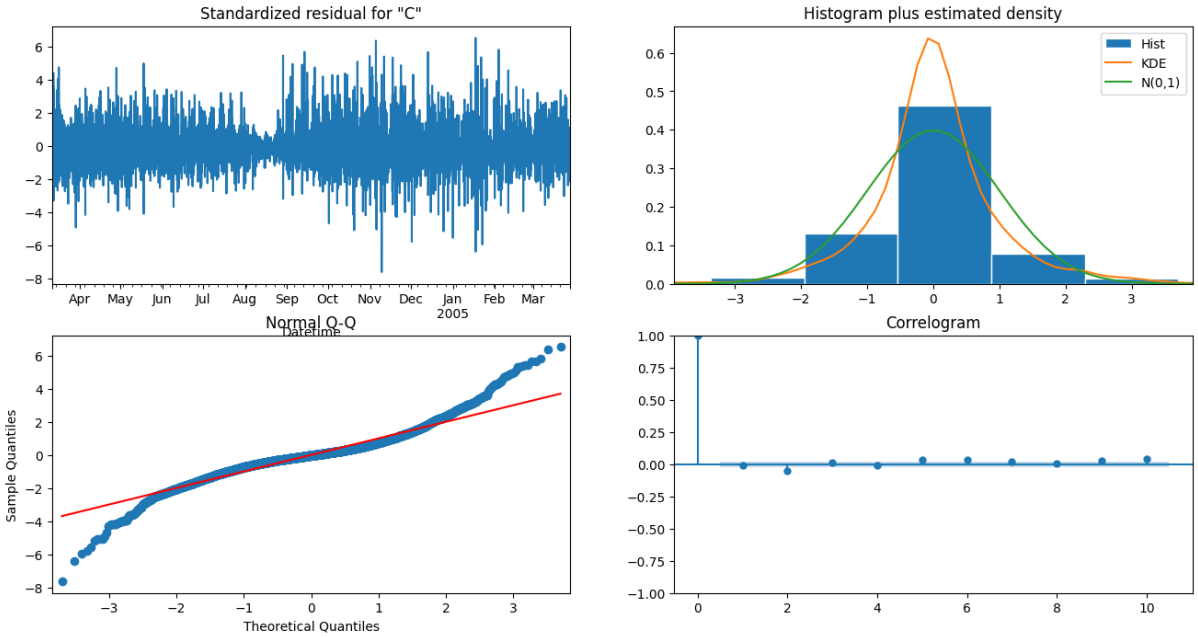
Παρόλο αυτά, το μοντέλο παρουσιάζει χαμηλά σφάλματα όπως φαίνεται στον πίνακα 46, με το MAE να βρέθηκε ως 0,81 και το RMSE ως 1,08.

Πίνακας 47 Λοιπές μετρικές Multistep SARMA

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                             CO(GT)   No. Observations:                 9189
Model:             SARIMAX(1, 0, 1)x(1, 0, 1, 24)   Log Likelihood               -8798.554
Date:                            Sun, 29 Mar 2026   AIC                          17607.107
Time:                                    16:04:00   BIC                          17642.722
Sample:                                03-10-2004   HQIC                         17619.215
                                     - 03-28-2005
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7737      0.006    128.825      0.000       0.762       0.786
ma.L1          0.1410      0.010     14.641      0.000       0.122       0.160
ar.S.L24       0.9962      0.001   1372.027      0.000       0.995       0.998
ma.S.L24      -0.9185      0.004   -249.881      0.000      -0.926      -0.911
sigma2         0.3976      0.003    120.561      0.000       0.391       0.404
===================================================================================
Ljung-Box (L1) (Q):                   0.64   Jarque-Bera (JB):              8881.35
Prob(Q):                              0.42   Prob(JB):                         0.00
Heteroskedasticity (H):               1.32   Skew:                             0.30
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.79
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
MAE: 0.8187566751949479
RMSE: 1.0850912831285437
```

Παράλληλα για την αξιολόγηση του μοντέλου, μπορούν να χρησιμοποιηθούν και άλλες μετρικές οι οποίες παρουσιάζονται στον πίνακα 47. Αρχικά το Alkaline Information Criterion (AIC) βρέθηκε ως 17607,107 ωστόσο δεν έχει νόημα προς σχολιασμό προς το παρόν καθώς θα χρησιμοποιηθεί για την σύγκριση με άλλα μοντέλα. Παράλληλα, υπάρχουν ορισμένες παρατηρήσεις που μπορούν να γίνουν σχετικά με τους συντελεστές του μοντέλου. Συγκεκριμένα, παρατηρείται ότι ο συντελεστής AR (p) λαμβάνει σχετικά υψηλή τιμή, περίπου 0,77 και ο εποχιακός συντελεστής AR (P) πολύ υψηλή τιμή 0,99. Επίσης, ο συντελεστής MA (q) παρουσιάζει χαμηλή τιμή 0,14 ενώ ο συντελεστής εποχιακού MA (Q) παρουσιάζει πολύ υψηλή αρνητική τιμή -0,91. Το Ljung – Box test βρέθηκε 0,42 δηλαδή μεγαλύτερο του 0,05 επομένως η αρχική υπόθεση, δηλαδή ότι τα κατάλοιπα ανεξάρτητα κατανεμημένα και παρομοιάζουν λευκό θόρυβο, δεν απορρίπτεται. Ωστόσο, το Jarque – Bera p, είναι κοντά στο 0, που σημαίνει ότι τα κατάλοιπα δεν ακολουθούν κανονική κατανομή.

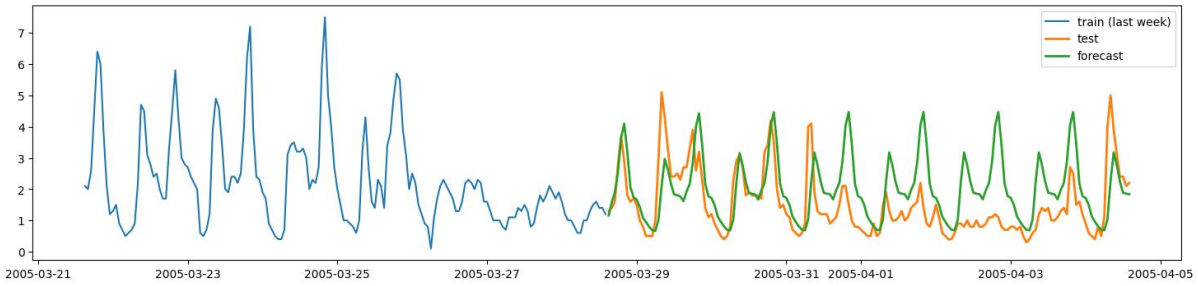
Για την καλύτερη αποτύπωση των αποτελεσμάτων σχετικά με τα κατάλοιπα, μπορεί να χρησιμοποιηθεί η μέθοδος “plot_diagnostics()”, εφόσον η μεταβλητή που θα εφαρμοστεί είναι παράγωγο της βιβλιοθήκης “statsmodels” [61] (για παράδειγμα έστω ένα αντικείμενο res = model.fit(), όπου το “model” έστω SARIMAX της βιβλιοθήκης “statsmodels”). Μέσω αυτής μπορούν να παραχθούν με ευκολία διαγράμματα σχετικά με τα κατάλοιπα, την κατανομή τους, την κανονικότητα τους και την αυτοσυσχέτιση τους.



Διάγραμμα 54 Διαγράμματα καταλοίπων για Multistep SARMA

Από το διάγραμμα 54, φαίνεται ότι πράγματι τα κατάλοιπα κυμαίνονται γύρω από το 0, μοιάζουν με λευκό θόρυβο και ότι η αυτοσυσχέτιση έχει αντιμετωπιστεί. Ωστόσο η διακύμανση δεν είναι σταθερή και παράλληλα υπάρχουν ακραίες τιμές. Από το διάγραμμα Q-Q και το ιστόγραμμα φαίνεται ότι η κατανομή τους είναι δεν είναι πλήρως κανονική αφού παρουσιάζουν ουρές (heavy tails), που επιβεβαιώνονται αφού η Κύρτωση βρέθηκε ως 7,79.

4.1.2 SARIMA – Multi - Step



Διάγραμμα 55 Απόδοση multistep SARIMA

Όπως προαναφέρθηκε, θεωρείται σκόπιμο να γίνει σύγκριση μεταξύ της υπάρχουσας χρονοσειράς και της χρονοσειράς εποχιακών διαφορών. Αυτό μπορεί να επιτευχθεί είτε μέσω της χρήσης του προηγούμενου μοντέλου στην χρονοσειρά εποχιακών διαφορών είτε στην αρχική χρονοσειρά με την χρήση μοντέλου που περιέχει συντελεστή εποχιακών διαφορών ίσο με 1 (D = 1), δηλαδή SARIMA (1,0,1)(1,1,1), με τα αποτελέσματα να παρουσιάζονται παρακάτω.

Πίνακας 48 Μετρικές σφάλματος Multistep SARIMA

Mean Absolute Error	Root Mean Squared Error
0.8752688895833016	1.1429205083403087

Ο πίνακας 48 παρουσιάζει τις μετρικές σφάλματος με το MAE να βρέθηκε περίπου 0,87 και RMSE 1,14. Συγκριτικά με το προηγούμενο μοντέλο παρατηρείται ότι τα σφάλματα αυξήθηκαν και επομένως οι προβλέψεις δεν είναι καλύτερες από το προηγούμενο μοντέλο.

Πίνακας 49 Λοιπές μετρικές Multistep SARIMA

SARIMAX Results

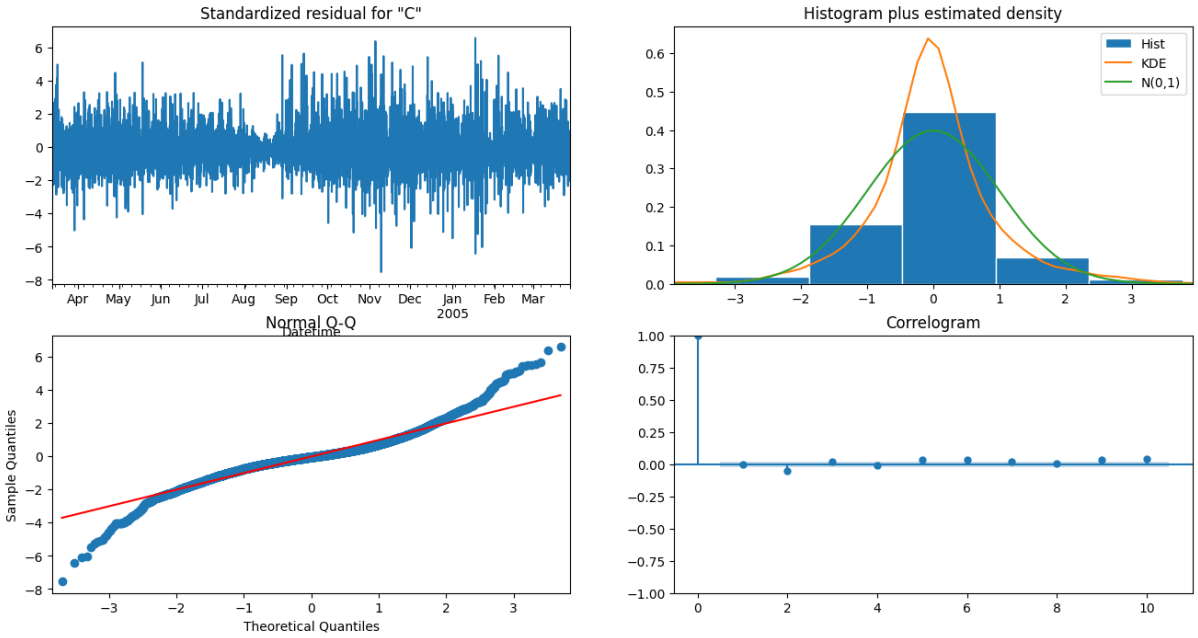
Dep. Variable:	CO(GT)	No. Observations:	9189
Model:	SARIMAX(1, 0, 1)x(1, 1, 1, 24)	Log Likelihood	-8740.102
Date:	Sun, 29 Mar 2026	AIC	17490.204
Time:	16:12:58	BIC	17525.805
Sample:	03-10-2004 - 03-28-2005	HQIC	17502.309
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.7671	0.006	125.547	0.000	0.755	0.779
ma.L1	0.1319	0.010	13.407	0.000	0.113	0.151
ar.S.L24	0.1094	0.008	13.120	0.000	0.093	0.126
ma.S.L24	-0.9380	0.003	-298.765	0.000	-0.944	-0.932
sigma2	0.3943	0.003	122.061	0.000	0.388	0.401

Ljung-Box (L1) (Q):	0.28	Jarque-Bera (JB):	8870.14
Prob(Q):	0.60	Prob(JB):	0.00
Heteroskedasticity (H):	1.34	Skew:	0.27
Prob(H) (two-sided):	0.00	Kurtosis:	7.80

Αρχικά, παρατηρείται ότι το AIC βελτιώθηκε (17490), δηλαδή μειώθηκε, σε σχέση με το προηγούμενο μοντέλο (17607). Επίσης, σημαντική μείωση παρατηρείται και στον συντελεστή του εποχιακού AR, που μειώθηκε στο 0,1094 από 0,9962, δηλαδή φαίνεται ότι πλέον η εφαρμογή εποχιακών διαφορών αντικατέστησε την δομή του εποχιακού συντελεστή. Ο εποχιακός συντελεστής MA συνεχίζει να παρουσιάζει σημαντικό αρνητικό μέγεθος (-0,9380). Το Ljung-Box τεστ συνεχίζει να είναι μεγαλύτερο του επιπέδου σημαντικότητας, επομένως τα κατάλοιπα είναι ανεξάρτητα και τυχαία, επομένως δεν παρουσιάζουν αυτοσυσχέτιση μεταξύ τους. Παράλληλα, το Jarque-Bera Test συνεχίζει να είναι 0 δηλαδή μικρότερο του επίπεδου σημαντικότητας, επομένως τα κατάλοιπα συνεχίζουν να μην ακολουθούν κανονική κατανομή.

Από όλα τα παραπάνω γίνεται αντιληπτό ότι τα κατάλοιπα δεν ακολουθούν κανονική κατανομή και παρουσιάζουν ουρές ανεξάρτητα του μοντέλου που χρησιμοποιείται. Δηλαδή, το πρόβλημα πηγάζει απευθείας από τα δεδομένα, για παράδειγμα στις ακραίες τιμές και την μη – σταθερή διακύμανση των τιμών. Πιθανώς, η χρήση κάποιου μετασχηματισμού (όπως για παράδειγμα ο λογαριθμικός μετασχηματισμός) να μείωνε σημαντικά την ύπαρξη ετεροσκεδαστικότητας και των ακραίων τιμών, χωρίς όμως αυτό να σημαίνει ότι η χρήση τέτοιων μετασχηματισμών είναι η βέλτιστη λύση.



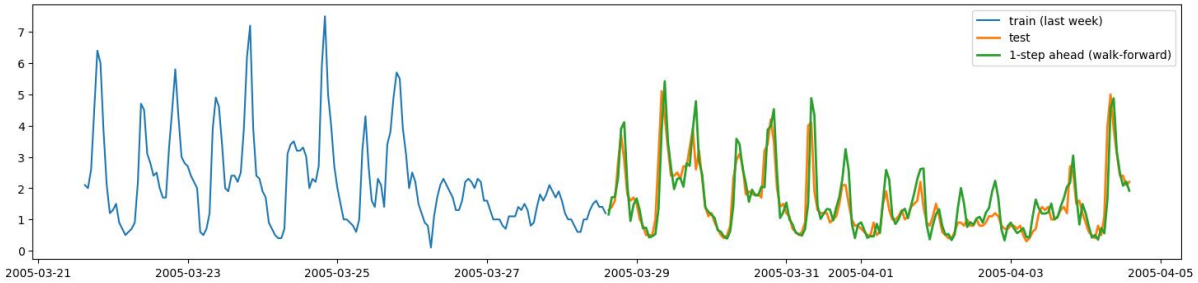
Διάγραμμα 56 Διαγράμματα καταλοίπων για Multistep SARIMA

Από το διάγραμμα 56, φαίνεται ότι το μοντέλο παρουσιάζει σχεδόν ίδια χαρακτηριστικά με το προηγούμενο μοντέλο, δηλαδή συνεχίζει να παρατηρείται ότι τα κατάλοιπα κυμαίνονται γύρω από το 0, χωρίς σταθερή διακύμανση, δεν παρατηρείται αυτοσυσχέτιση στα κατάλοιπα ενώ παράλληλα δεν ακολουθούν κανονική κατανομή και παρουσιάζουν ουρές (tails).

Εδώ γίνεται αντιληπτό, από την σύγκριση των 2 μοντέλων ότι χρειάζεται να γίνει μια επιλογή σχετικά με την εφαρμογή εποχιακών διαφορών. Δηλαδή, παρατηρήθηκε ότι το μοντέλο μη – εποχιακών διαφορών παρουσιάζει υψηλότερο AIC – BIC αλλά χαμηλότερα σφάλματα MAE – RMSE ενώ το μοντέλο εποχιακών παρουσιάζει αντίθετα μέτρα. Επομένως, ο κάθε ερευνητής πρέπει να αποφασίσει εάν στόχος του είναι η επιλογή του μοντέλου με την καλύτερη προσαρμογή στα δεδομένα (goodness of fit) ή την καλύτερη παραγωγή προβλέψεων.

Αποφασίστηκε να χρησιμοποιηθεί η χρονοσειρά εποχιακών διαφορών καθώς η αύξηση των σφαλμάτων μετρικών είναι σχετικά χαμηλά ενώ παράλληλα ο εποχιακός συντελεστής παρουσιάζει βελτίωση ως προς την μοντελοποίηση, κοινώς το μοντέλο μπορεί να χαρακτηριστεί ως πιο σταθερό. Επίσης, βάσει των περιπτώσεων ADF – KPSS test, όπως προαναφέρθηκε στο κεφάλαιο « Έλεγχος στασιμότητας και εποχικότητας», προτείνεται η εφαρμογή πρώτων – εποχιακών διαφορών ώστε η χρονοσειρά να είναι “πλήρως” στάσιμη.

4.1.3 SARIMA – Single - Step



Διάγραμμα 57 Απόδοση singlestep SARIMA

Το διάγραμμα 57, παρουσιάζει την εφαρμογή του μοντέλου Single – Step (Non - Recursive) SARIMA (1,0,1) x (1,1,1)24. Απευθείας παρατηρείται ότι το μοντέλο παρουσιάζει σημαντική βελτίωση σε σχέση με τις προηγούμενες μεθόδους, καθώς οι προβλέψεις φαίνεται να ακολουθούν τις παρατηρήσεις του δείγματος δοκιμής. Ωστόσο, φαίνεται ότι σε αρκετές κορυφές – υψηλές τιμές το μοντέλο παράγει υπερεκτιμήσεις σε σχέση με τις πραγματικά παρατηρούμενες τιμές.

Πίνακας 50 Μετρικές σφάλματος singlestep SARIMA

Mean Absolute Error	Root Mean Squared Error
0.3681206289921636	0.5762088094671861

Οι μετρικές σφάλματος επιβεβαιώνουν την παρατηρούμενη συμπεριφορά, καθώς είναι σημαντικά μικρότερες σε σχέση με τις προηγούμενες μεθόδους με το MAE να υπολογίστηκε ως 0,36 και το RMSE ως 0,57.

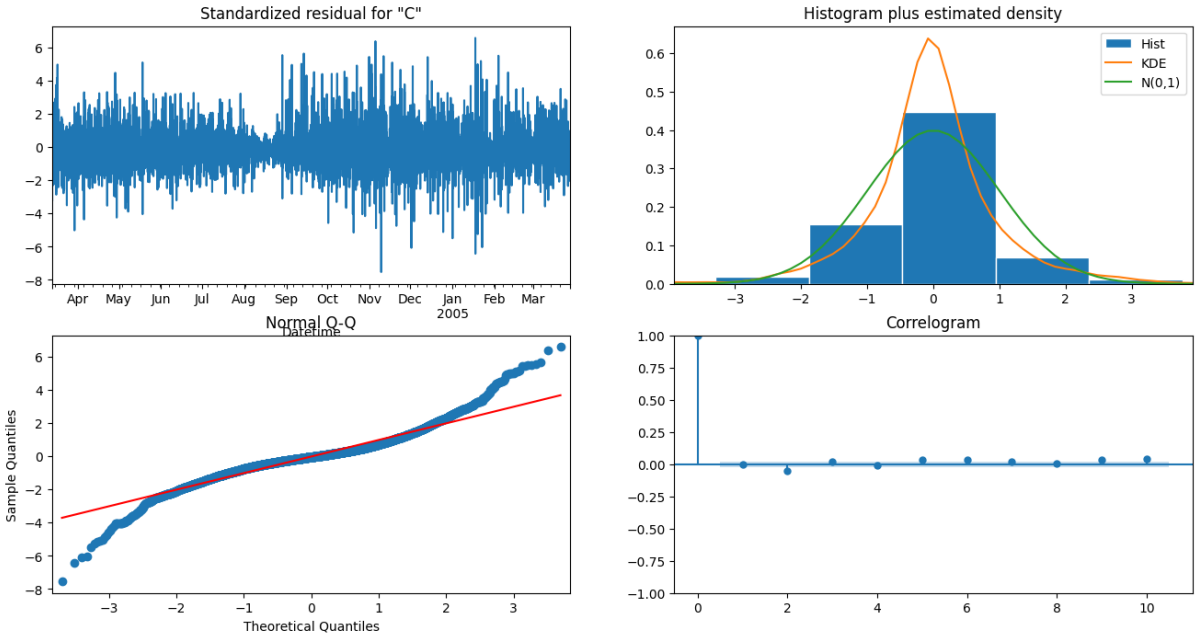
Πίνακας 51 Λοιπές μετρικές singlestep SARIMA

SARIMAX Results

Dep. Variable:			CO(GT)	No. Observations:		9189
Model:	SARIMAX(1, 0, 1)x(1, 1, 1, 24)			Log Likelihood		-8740.102
Date:			Sun, 26 Apr 2026	AIC		17490.204
Time:			09:40:15	BIC		17525.805
Sample:			03-10-2004	HQIC		17502.309
			- 03-28-2005			
Covariance Type:			opg			
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.7671	0.006	125.547	0.000	0.755	0.779
ma.L1	0.1319	0.010	13.407	0.000	0.113	0.151
ar.S.L24	0.1094	0.008	13.120	0.000	0.093	0.126
ma.S.L24	-0.9380	0.003	-298.765	0.000	-0.944	-0.932
sigma2	0.3943	0.003	122.061	0.000	0.388	0.401

Ljung-Box (L1) (Q):	0.28	Jarque-Bera (JB):	8870.14
Prob(Q):	0.60	Prob(JB):	0.00
Heteroskedasticity (H):	1.34	Skew:	0.27
Prob(H) (two-sided):	0.00	Kurtosis:	7.80

Ο πίνακας λοιπών μετρικών που παρουσιάζεται στο διάγραμμα 51, είναι πανομοιότυπος με την προηγούμενη μέθοδο παρόλο που τα αποτελέσματα των προβλέψεων είναι διαφορετικά μεταξύ τους. Επομένως, ισχύουν οι προαναφερόμενες παρατηρήσεις.

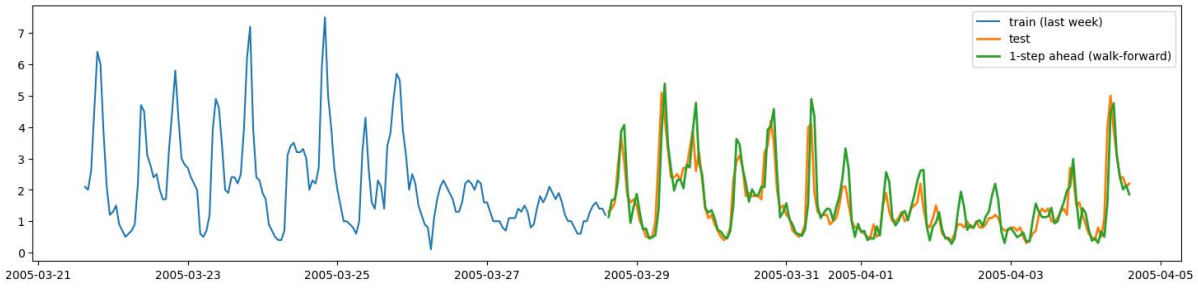


Διάγραμμα 58 Διαγράμματα καταλοίπων για singlestep SARIMA

Με παρόμοιο τρόπο, τα συγκεντρωτικά διαγράμματα καταλοίπων που παρουσιάζονται στο διάγραμμα 58, είναι επίσης πανομοιότυπα με αυτά της προηγούμενης μεθόδου.

Δηλαδή, εδώ γίνεται αντιληπτό ότι παρόλο που χρησιμοποιείται το ίδιο μοντέλο, με αποτέλεσμα να παράγονται τα ίδια κατάλοιπα (εκπαίδευσης), τα αποτελέσματα προβλέψεων – μετρικών σφαλμάτων είναι εντελώς διαφορετικά μεταξύ τους. Ο λόγος που προκύπτει το παραπάνω είναι ότι όπως προαναφέρθηκε το μοντέλο single – step δεν ενσωματώνει τις παρατηρήσεις που παράγει εντός των δεδομένων αλλά κάνει κάθε επόμενη πρόβλεψη βάσει των πραγματικών τιμών. Κοινώς, εδώ γίνεται αντιληπτό ότι η διαφορά μεταξύ των 2 είναι η μη συσσώρευση των σφαλμάτων, επομένως εάν στην προσέγγιση multi – step με κάποιον τρόπο αφαιρούνταν τα συσωρευτικά σφάλματα, τα αποτελέσματα προβλέψεων θα ήταν ίδια με την προσέγγιση multi – step αφού και τα 2 χρησιμοποιούν το ίδιο μοντέλο και παραμέτρους.

4.1.4 SARIMAX – Single – Step



Διάγραμμα 59 Απόδοση singlestep SARIMAX

Έπειτα, μπορεί να εφαρμοστεί η χρήση λοιπών μεταβλητών εντός του μοντέλου, δηλαδή το μοντέλο SARIMAX. Γενικά, βάσει του διαγράμματος 59, δεν φαίνεται η χρήση των επιπλέον μεταβλητών να επιφέρει ιδιαίτερες αλλαγές στην απόδοση του μοντέλου, καθώς είναι σχεδόν πανομοιότυπο με αυτό της προηγούμενης μεθόδου.

Πίνακας 52 Μετρικές σφάλματος singlestep SARIMAX

Mean Absolute Error	Root Mean Squared Error
0.36616583113619583	0.5749341444019184

Αυτό επιβεβαιώνεται και από τις μετρικές που παρουσιάζονται στον πίνακα 52. Σε σχέση με το προηγούμενο μοντέλο singlestep SARIMA το MAE βρέθηκε ως 0,36 παρουσιάζοντας μια βελτίωση 0,002 και το RMSE ως 0,57 που επίσης παρουσιάζει μια βελτίωση 0,002. Προφανώς, οι βελτιώσεις είναι τόσο χαμηλού μεγέθους που πρακτικά είναι αμελητέες.

Πίνακας 53 Λοιπές μετρικές singlestep SARIMAX

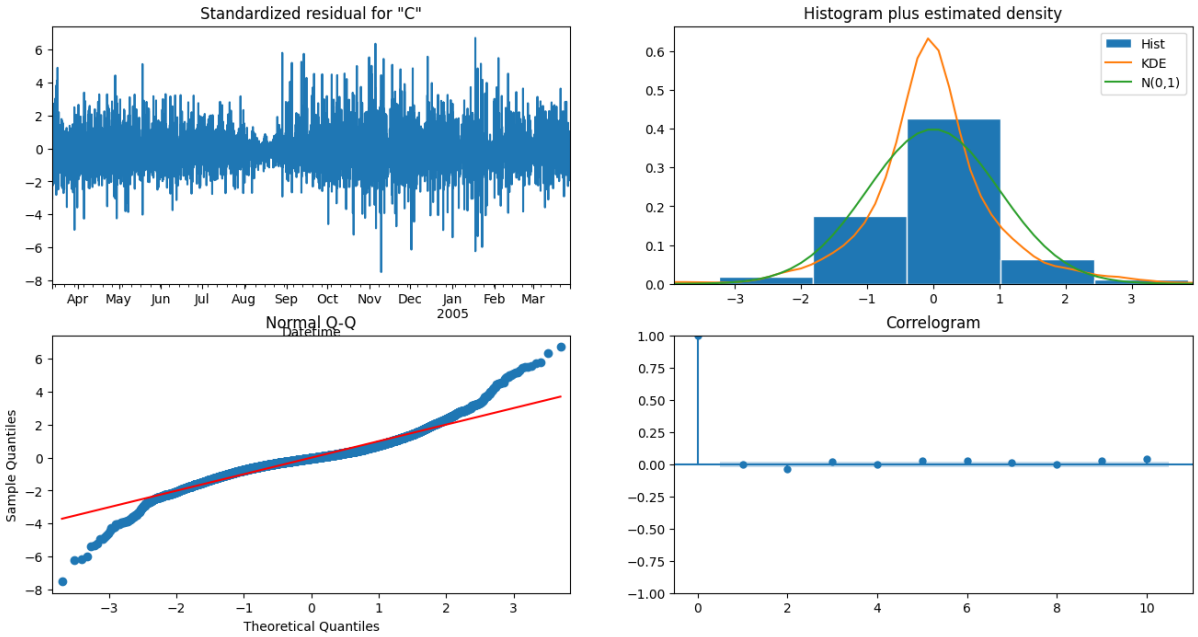
SARIMAX Results

Dep. Variable:	CO(GT)	No. Observations:	9189
Model:	SARIMAX(1, 0, 1)x(1, 1, 1, 24)	Log Likelihood	-8711.030
Date:	Sun, 26 Apr 2026	AIC	17444.060
Time:	09:58:02	BIC	17522.383
Sample:	03-10-2004 - 03-28-2005	HQIC	17470.691
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
hour_sin	1.71e-06	1.11e+04	1.54e-10	1.000	-2.17e+04	2.17e+04
hour_cos	-2.785e-06	9307.267	-2.99e-10	1.000	-1.82e+04	1.82e+04
dow_sin	0.2164	0.038	5.719	0.000	0.142	0.291
dow_cos	-0.1856	0.034	-5.423	0.000	-0.253	-0.119
month_sin	0.0021	0.150	0.014	0.989	-0.292	0.297
month_cos	0.2299	0.158	1.456	0.146	-0.080	0.540
ar.L1	0.7419	0.007	113.316	0.000	0.729	0.755
ma.L1	0.1460	0.010	14.595	0.000	0.126	0.166
ar.S.L24	0.1091	0.008	12.940	0.000	0.093	0.126
ma.S.L24	-1.0657	0.004	-291.481	0.000	-1.073	-1.059
sigma2	0.3449	0.004	91.695	0.000	0.338	0.352

Ljung-Box (L1) (Q):	0.22	Jarque-Bera (JB):	8790.50
Prob(Q):	0.64	Prob(JB):	0.00
Heteroskedasticity (H):	1.36	Skew:	0.31
Prob(H) (two-sided):	0.00	Kurtosis:	7.76

Οι λοιπές μετρικές παρουσιάζονται στον πίνακα 53. Αρχικά, το μοντέλου παρουσιάζει χαμηλότερο AIC σε σχέση με την προηγούμενη μέθοδο. Επίσης, παρατηρείται ότι τα σφάλματα των μεταβλητών “hour_sin” και “hour_cos” είναι πολύ υψηλά ενώ παράλληλα οι συντελεστές τους είναι σχετικά χαμηλοί. Πιθανώς, αυτό συμβαίνει επειδή το μοντέλο αντιλαμβάνεται την δομή του σήματος, αφού έχει οριστεί εποχικότητα ίση με 24, και τα στοιχεία της ώρας δεν προσφέρουν πληροφορία στο μοντέλο. Πέρα από αυτό, οι μεταβλητές “dow_sin” και “dow_cos”, φαίνεται να συνησφέρουν στην αύξηση της απόδοσης του μοντέλου. Οι λοιποί συντελεστές είναι σχεδόν ίδιοι με το προηγούμενο μοντέλο, που είναι λογικό.



Διάγραμμα 60 Διαγράμματα καταλοίπων για singlestep SARIMAX

Τέλος, τα διαγράμματα καταλοίπων του διαγράμματος 60 φαίνεται να μην διαφέρουν από αυτά της προηγούμενης μεθόδου.

4.2 Μοντέλα Μηχανικής Μάθησης: Random Forest Endo- Exo

Η χρήση του Random Forest Regressor εφαρμόζεται συχνά σε δεδομένα air quality όπως φαίνεται από την βιβλιογραφία [28], [5], [4], [65], [66] και έχει παράγει μοντέλα με χαμηλά σφάλματα. Παρόλο που χρησιμοποιείται ο ίδιος αλγόριθμος η χρήση του μπορεί να διαφέρει σημαντικά, όπου για παράδειγμα στο [4] χρησιμοποιήθηκε για την «ανακατασκευή» του Nox μέσω των λοιπών μεταβλητών του σετ δεδομένων.

Ο αλγόριθμος Random Forest ή αλλιώς Random Decision Forest, κατατάσσεται στην κατηγορία αλγόριθμων “ensemble learning” και χρησιμοποιείται για προβλήματα κατηγοριοποίησης, παλινδρόμησης αλλά μπορεί να χρησιμοποιηθεί με πολλούς διαφορετικούς τρόπους όπως για την ανακατασκευή γνωρισμάτων όπως προαναφέρθηκε [4]. Ο αλγόριθμος κατασκευάζει πολλαπλά δέντρα αποφάσεων με τυχαία επιλογή χαρακτηριστικών για κάθε κόμβο με διαφορετικά σύνολα εκπαίδευσης για το κάδε δέντρο. Στο τέλος εκτελείται ψηφοφορία (majority voting) σε περίπτωση προβλήματος κατηγοριοποίησης ή διαφορετικά επιλέγεται η μέση τιμή που προκύπτει από τα δέντρα για προβλήματα παλινδρόμησης – συνεχών τιμών [4].

Οι πιο συνήθεις παραμέτρους αφορούν τον αριθμό των δέντρων απόφασης που θα δημιουργηθούν καθώς και το μέγιστο βάθος που θα αναπτυχθούν ή διαφορετικά μετά από πόσους κόμβους (nodes) θα εφαρμοστεί η διαδικασία “pruning”. Ο αλγόριθμος μπορεί να περιγράφει από την εξίσωση 9:

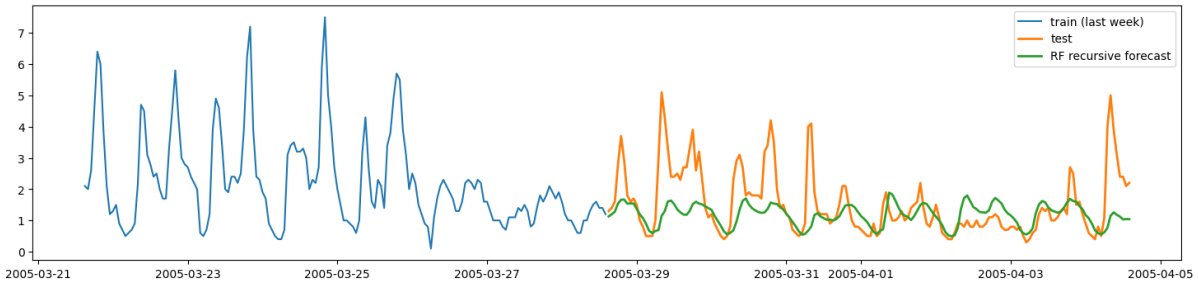
Πίνακας 54 Εξίσωση μοντέλου Random Forest

$$H(x) = \frac{1}{T}\sum_{i=1}^{T} h_i(x)$$	(9)

Όπου T ο αριθμός των παραγόμενων δένδρων παλινδρόμησης και $h_i(x)$ το αποτέλεσμα του I δέντρου παλινδρόμησης για το δείγμα x. Δηλαδή, η συνάρτηση $H(x)$, ή αλλιώς το αποτέλεσμα του Random Forest Regression, είναι η μέση τιμή του συνόλου των δέντρων απόφασης [4].

Για την εφαρμογή του αλγορίθμου Random Forest στα πλαίσια μελέτης χρονοσειρών, χρησιμοποιήθηκε το module ForecasterRecursive [59] σε συνδυασμό με την μέθοδο RandomForestRegressor της βιβλιοθήκης sklearn [78]. Το module επιτρέπει την δημιουργία lags μέσω της συνάρτησης, επομένως δεν είναι αναγκαία η προ – κατασκευή των lagged μεταβλητών. Επίσης, είναι σημαντικό να αναφερθεί ότι η παράμετρος "lags" μπορεί να δεχθεί είτε λίστες είτε ακέραιους αριθμούς. Ωστόσο, εάν εισαχθεί ακέραιος αριθμός X, αυτό συμπεριλαμβάνει όλους τους αριθμούς – lag predictors από το 1 έως το X. Σε περίπτωση που εισαχθεί λίστα, το παραπάνω δεν ισχύει και χρησιμοποιούνται μόνο τα lags εντός της εισαγόμενης λίστας [59]. Πέρα των παραπάνω, η υλοποίηση "single step" είναι πανομοιότυπη με αυτήν του προηγούμενου κεφάλαιού.

4.2.1 Univariate - Multistep Random Forest



Διάγραμμα 61 Απόδοση univariate multistep Random Forest

Το διάγραμμα 61, παρουσιάσει την εφαρμογή της μεθόδου Univariate Random Forest, δηλαδή με την χρήση του γνωρίσματος CO(GT), χωρίς εξωγενείς μεταβλητές. Επειδή η συγκεκριμένη μέθοδος είναι αναδρομική (Recursive), δεν χρειάζεται μεγάλο αριθμό υπολογιστικών πόρων, επομένως οι παράμετροι του Random Forest μπορούν να έχουν μεγαλύτερο μέγεθος και παρουσιάζονται αναλυτικά στον πίνακα 55.

Πίνακας 55 Παράμετροι univariate multistep Random Forest

Παράμετρος	**Τιμή**
n_estimators	300
max_depth	5
lags	1 – 24

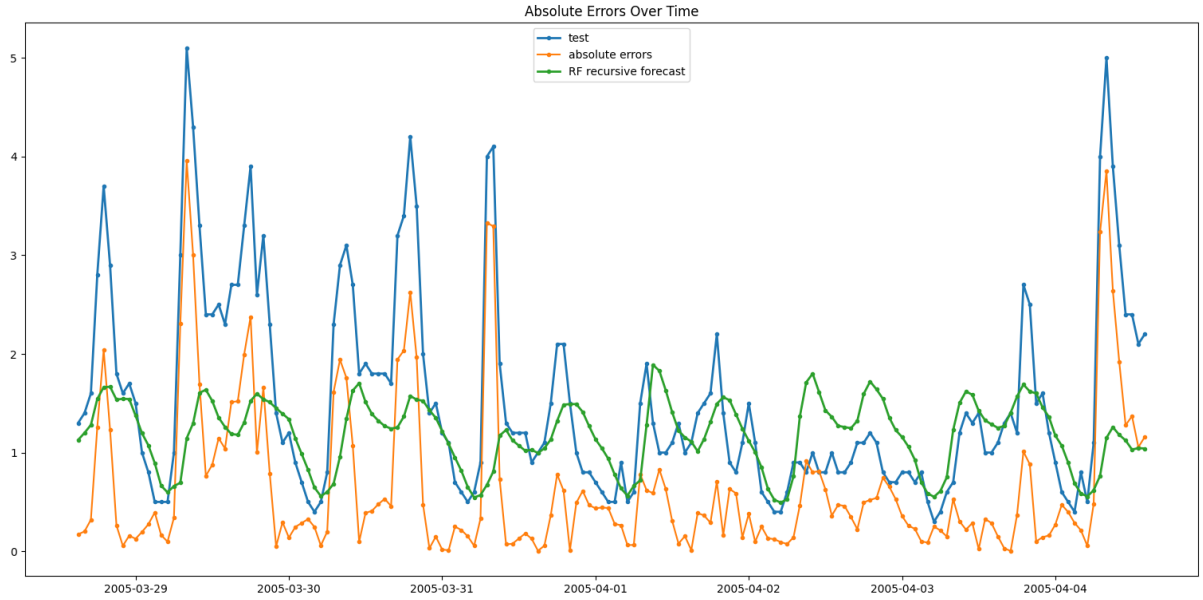
Το μοντέλο φαίνεται να παράγει μια μέση τιμή του σήματος ως πρόβλεψη, παρομοιάζοντας τα αποτελέσματα του διαγράμματος 53, δηλαδή Multi Step SARMA, χωρίς όμως να μπορεί να διαχειριστεί τις υψηλές τιμές. Ωστόσο, οι μετρικές σφαλμάτων είναι χαμηλότερες για την

συγκεκριμένη μέθοδο σε σχέση με την προαναφερόμενη μέθοδο, καθώς το MAE βρέθηκε ως 0.65 και RMSE 1.02, όπως φαίνονται στον πίνακα 56.

Πίνακας 56 Μετρικές σφάλματος univariate multistep Random Forest

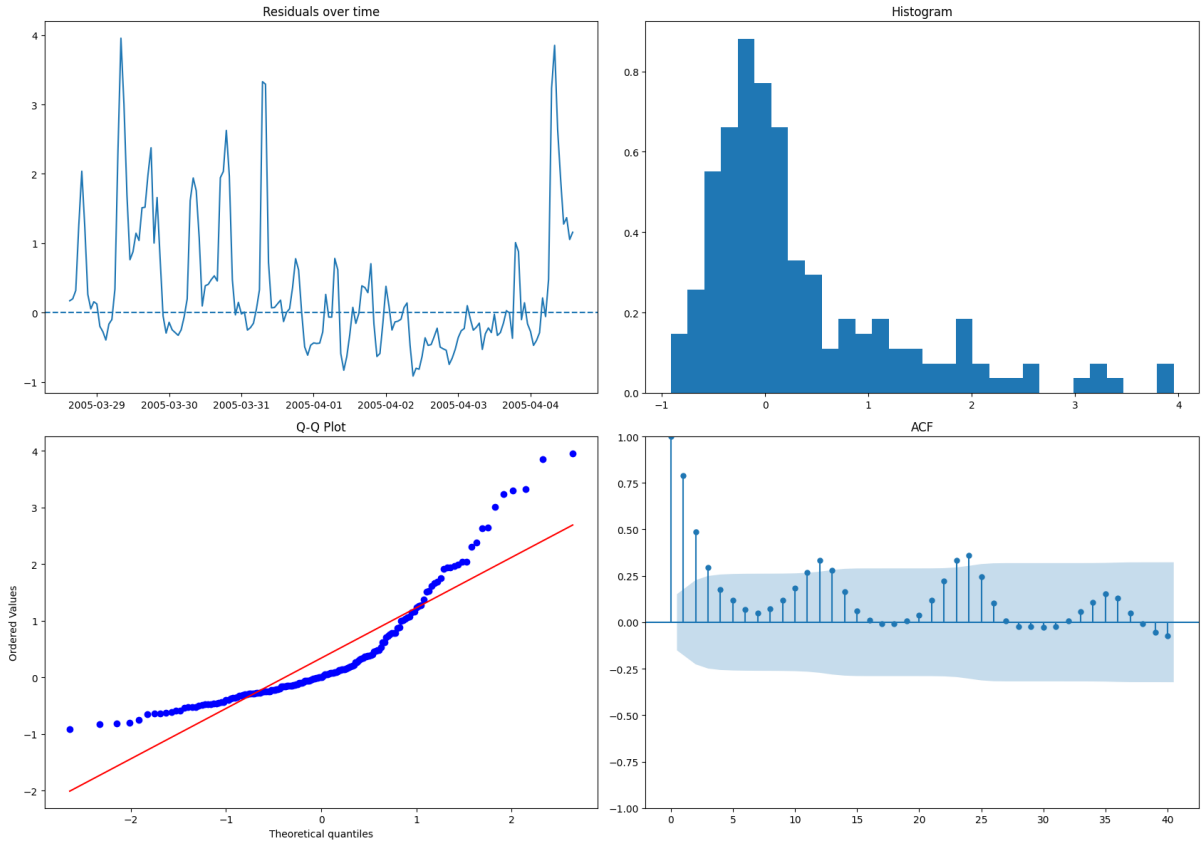
Mean Absolute Error	Root Mean Squared Error
0.6532502952569917	1.0221473282777487

Εκ πρώτης όψεως, οι μετρικές σφάλματος δεν φαίνεται να συμβαδίζουν με το προαναφερόμενο γράφημα, καθώς παρατηρούνται αρκετά σημεία πρόβλεψης που απέχουν σημαντικά από τα σημεία παρατηρήσεων. Για παράδειγμα, για το διάστημα 2005-04-04, παρατηρείται ότι ένα σημείο απέχει περίπου 4 μg/m^3 από την πραγματική τιμή του σετ δοκιμής. Πράγματι, μέσω της δημιουργίας ενός νέου γνωρίσματος “errors” εντός το οποίου έγινε η αφαίρεση των προβλέψεων από τις πραγματικές τιμές σε απόλυτη τιμή για τον υπολογισμό των απόλυτων σφαλμάτων, αποδείχθηκε ότι το μεγαλύτερο (απόλυτο) σφάλμα που παρατηρείται είχε τιμή 3.9543333333333335 μg/m^3. Βάσει αυτού δημιουργήθηκε το διάγραμμα 62, που παρουσιάζει συνδυαστικά τα απόλυτα σφάλματα, τιμές πρόβλεψης και πραγματικές τιμές σε μορφή γραφημάτων.



Διάγραμμα 62 Απόλυτα σφάλματα univariate multistep Random Forest

Από το διάγραμμα 62, επικυρώνονται τα αποτελέσματα των μετρικών σφαλμάτων ενώ παράλληλα μπορούν να γίνουν μερικές νέες παρατηρήσεις σχετικά με το μοντέλο. Πράγματι φαίνεται ότι τα περισσότερα σφάλματα συγκεντρώνονται γύρω από το 0,6 ενώ τα υψηλά σφάλματα οφείλονται στις ραγδαίες κορυφώσεις του CO που παρατηρήθηκαν. Μάλιστα είναι η αλλαγή των τιμών είναι τόσο απότομη όπου εντός διαστήματος μιας ώρας για τις 2005-04-04, η τιμή αυξήθηκε κατά 4 μg/m^3. Βάσει όλων των παραπάνω, θεωρείται ότι το μοντέλο μπορεί να προβλέψει το μέσο μοτίβο CO, ωστόσο δεν μπορεί να ανταπεξέλθει στις απότομες αλλαγές – κορυφές που μπορεί να προκύψουν, πράγμα το οποίο εξηγεί για ποιον λόγο το MAE είναι σχετικά χαμηλό ενώ το RMSE, το οποίο επηρεάζεται περισσότερο από ακραίες τιμές, είναι υψηλότερο.

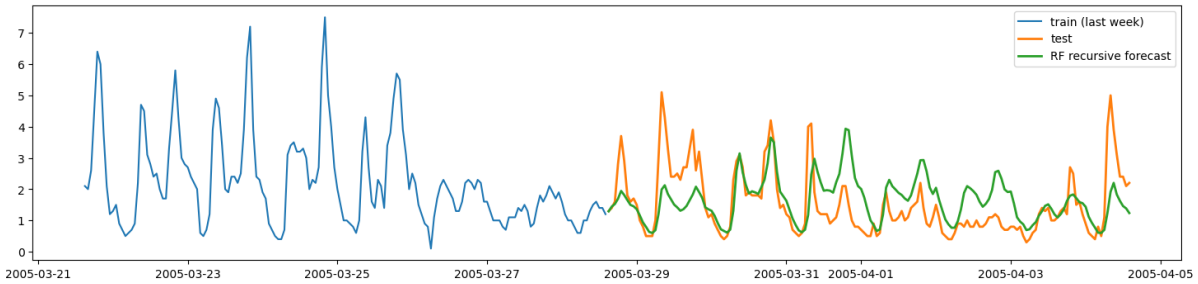


Διάγραμμα 63 Διαγράμματα καταλοίπων για univariate multistep Random Forest

Μπορούν επίσης να κατασκευαστούν και όλα τα διαγράμματα που παρουσιάστηκαν για την απόδοση των στατιστικών μεθόδων SARIMA. Ωστόσο, σε αντίθεση με τις στατιστικές μέθοδους όπου τα διαγράμματα μπορούν να κατασκευαστούν εύκολα χρησιμοποιώντας την μέθοδο “plot_diagnostics” [60], υπάρχουν διαφορετικές μέθοδοι που παρέχονται από την βιβλιοθήκη “skforecast” [62]. Με παρόμοιο τρόπο μπορούν να εφαρμοστούν σε αντικείμενα που δημιουργήθηκαν από μεθόδους της βιβλιοθήκης “skforecast”, όπως την μέθοδο “Forecaster Recursive” [59] που χρησιμοποιήθηκε για την υλοποίηση του αλγόριθμου Random Forest. Αντί αυτών δημιουργήθηκε η συνάρτηση “rf_residual_diagnostics ()”, για την παραγωγή των διαφόρων γραφημάτων.

Στο διάγραμμα 63, φαίνεται ότι τα κατάλοιπα δεν κυμαίνονται γύρω από το 0, και παρατηρείται μεγάλη διακύμανση των τιμών. Η κατανομή τους δεν είναι κανονική, φαίνεται να υπάρχει έντονη λοξότητα και θα χαρακτηριζόταν ως δεξιά ασύμμετρη (right-skewed - positively skewed) κατανομή, με μεγάλη ουρά. Αυτό επιβεβαιώνεται και από το διάγραμμα Q-Q καθώς τα κατάλοιπα δεν ακολουθούν την διαγώνιο και υπάρχουν αρκετές ακραίες τιμές. Τέλος, από το διάγραμμα ACF παρατηρούνται 6 κορυφές εκτός διαστήματος εμπιστοσύνης, ενώ παρουσιάζεται και ένα κυματοειδές μοτίβο, επιβεβαιώνοντας ότι τα κατάλοιπα δεν είναι λευκός θόρυβος και το μοντέλο δεν έχει ενσωματώσει πλήρως το μοτίβο - σήμα.

4.2.2 Multivariate – Multistep Random Forest



Διάγραμμα 64 Απόδοση multivariate multistep Random Forest

Στην συνέχεια, μπορεί να εφαρμοστεί ο αλγόριθμος Random Forest, χρησιμοποιώντας εξωγενείς μεταβλητές, δηλαδή χρησιμοποιώντας τα γνωρίσματα που βρίσκονται στο σετ "x_train", με το γράφημα της πρόβλεψης να παρουσιάζεται στο διάγραμμα 64. Εκ πρώτης όψεως φαίνεται να παρουσιάζει καλύτερα αποτελέσματα καθώς είναι πιο ευδιάκριτη η ημερήσια δομή ενώ παράλληλα δεν φαίνεται να επαναλαμβάνει το καθημερινό μέσο ημερήσιο μοτίβο αφού η πρόβλεψη αλλάζει μεγέθη για την διάρκεια του ορίζοντα πρόβλεψης. Οι παράμετροι του Random Forest δεν διαφοροποιούνται από το άλλο μοντέλο και παρουσιάζονται στον πίνακα 57.

Πίνακας 57 Παράμετροι multivariate multistep Random Forest

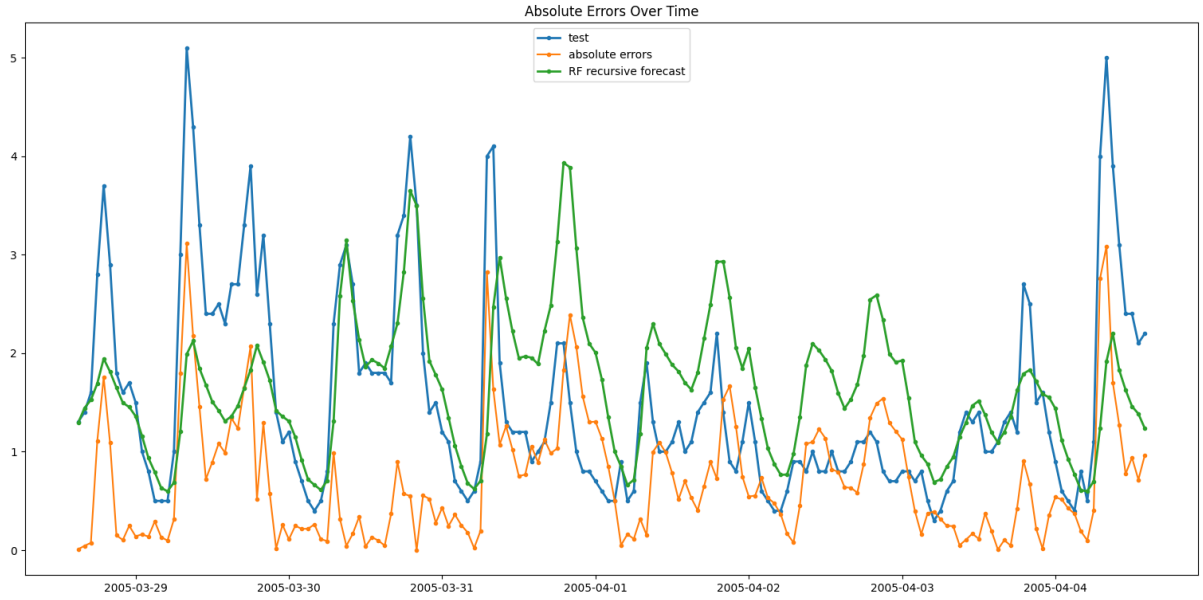
Παράμετρος	**Τιμή**
n_estimators	300
max_depth	5
lags	1 - 24

Οι παράμετροι έμειναν σταθεροί καθώς σκοπός δεν ήταν η σύγκριση διαφορετικών παραμέτρων αλλά η σύγκριση μεταξύ του ίδιου μοντέλου με την χρήση εξωγενών μεταβλητών.

Πίνακας 58 Μετρικές σφάλματος multivariate multistep Random Forest

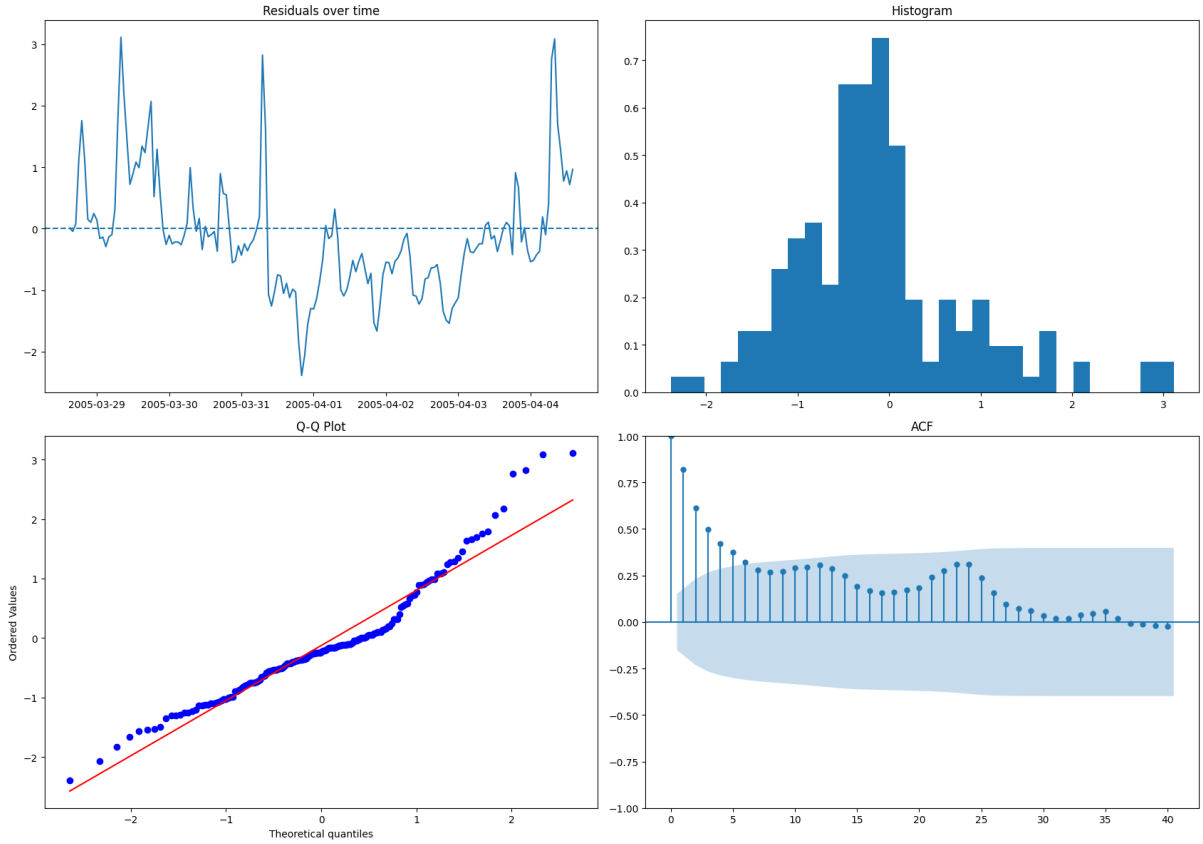
Mean Absolute Error	**Root Mean Squared Error**
0.7123943098072563	0.9521387851528437

Ωστόσο οι μετρικές σφάλματος αναδεικνύουν μια διαφορετική εικόνα, καθώς το MAE (0,71) αυξήθηκε ενώ το RMSE (0,95) μειώθηκε όπως φαίνεται στον πίνακα 58. Δηλαδή, το μοντέλο βελτιώθηκε ως προς τα υψηλά σφάλματα που οφείλονταν στις απότομες αλλαγές του CO, ωστόσο λόγω αυτού πιθανώς υπερεκτιμά τις υπόλοιπες πιο ήπιες τιμές, με αποτέλεσμα την αύξηση του MAE.



Διάγραμμα 65 Απόλυτα σφάλματα multivariate multistep Random Forest

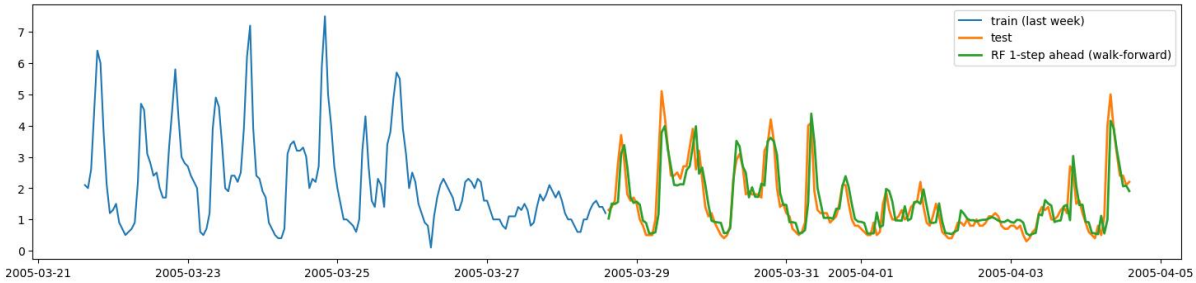
Μέσω του διαγράμματος 65, μπορεί να ερευνηθεί εάν πράγματι ισχύει η προαναφερόμενη υπόθεση. Πράγματι επιβεβαιώνονται καθώς το μοντέλο παρουσιάζει χαμηλότερα σφάλματα στις απότομες αλλαγές – κορυφές ωστόσο αυτό προκαλεί την σημαντική αύξηση των σφαλμάτων λόγω υπερεκτίμησης των λοιπών προβλέψεων.



Διάγραμμα 66 Διαγράμματα καταλοίπων για multivariate multistep Random Forest

Από το διάγραμμα 66, παρατηρείται μια σημαντική βελτίωση στα κατάλοιπα σε σχέση με το προηγούμενο μοντέλο. Τα κατάλοιπα φαίνεται να κυμαίνονται περισσότερο γύρω από το 0 ωστόσο συνεχίζουν να παρουσιάζουν ιδιαίτερα ασταθή διακύμανση. Η κατανομή φαίνεται να πλησιάζει την κανονική ωστόσο ακόμα παρατηρείται μικρότερη δεξιά ασσυμετρία και σημαντική ουρά θετικών τιμών. Αυτό αντικατοπτρίζεται και από το διάγραμμα Q-Q καθώς περισσότερα σημεία εντοπίζονται να ακολουθούν την διαγώνιο, με τις ακραίες τιμές να συνεχίζουν να υπάρχουν. Τέλος, συνεχίζεται να παρατηρείται αυτοσυσχέτιση, με 5 lags να παρατηρούνται εκτός της ζώνης εμπιστοσύνης.

4.2.3 Univariate – Single Step Random Forest



Διάγραμμα 67 Απόδοση univariate singlestep Random Forest

Τέλος, στο διάγραμμα 67, παρουσιάζονται οι προβλέψεις του univariate single – step Random Forest. Απευθείας γίνεται αντιληπτό ότι το συγκεκριμένο μοντέλο φαίνεται να παρουσιάζει τις καλύτερες προβλέψεις καθώς ακολουθούν τις παρατηρήσεις του δείγματος δοκιμής. Η συγκεκριμένη υλοποίηση είναι αργή ως προς την εκτέλεση, καθώς στα πλαίσιο ενός default Collab VM, χρειάζεται 9 με 10 λεπτά για την παραγωγή αποτελεσμάτων, με χρήση πιο ήπιων συντελεστών, όπως φαίνεται στον πίνακα 59 παρακάτω.

Πίνακας 59 Παράμετροι univariate singlestep Random Forest

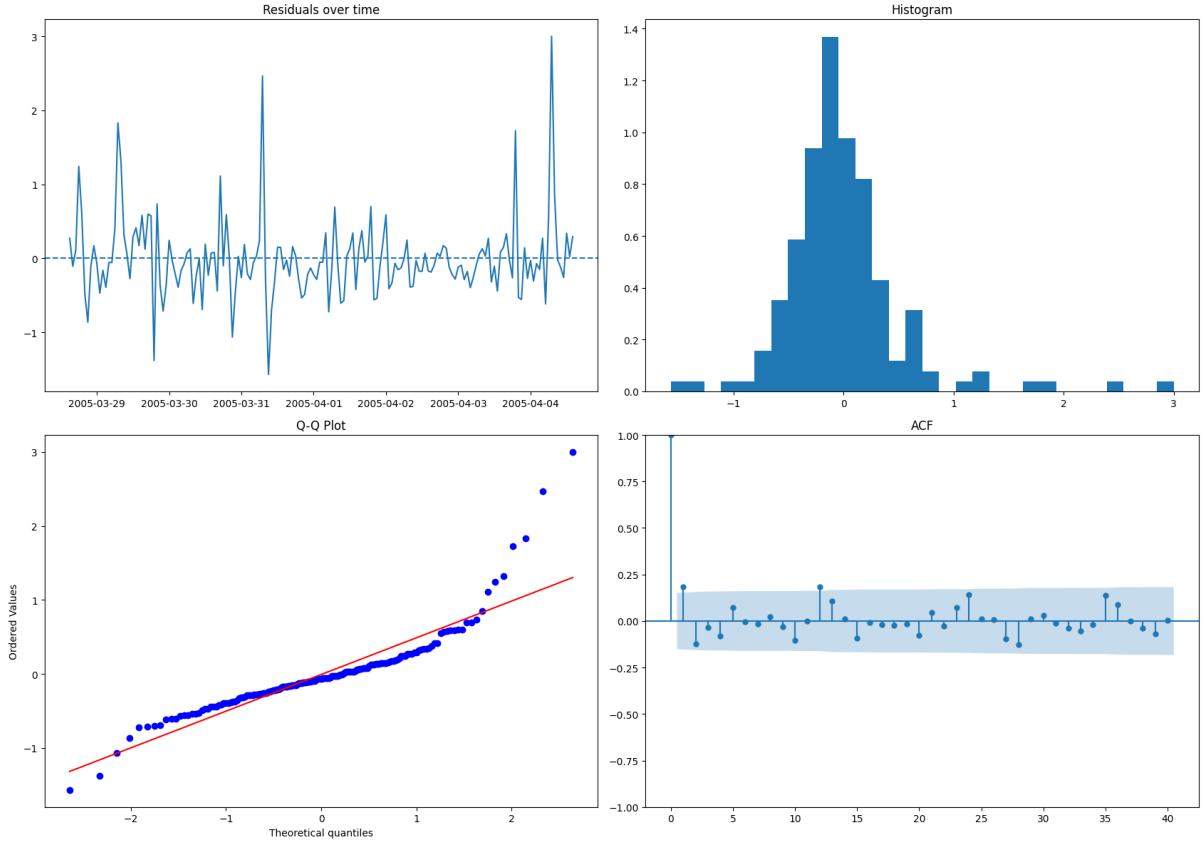
Παράμετρος	**Τιμή**
n_estimators	100
max_depth	5
lags	1 - 24

Αυτό είναι λογικό, καθώς η μέθοδος δεν είναι recursive, επομένως για κάθε βήμα που γίνεται πρόβλεψη παράγεται ένα νέο μοντέλο Random Forest. Όπως προαναφέρθηκε, ο χρονικός ορίζοντας ορίστηκε ως μια εβδομάδα και εφόσον η χρονική κλίμακα είναι ωριαία, αυτό σημαίνει ότι εκπαιδεύονται 168 μοντέλα, ένα για κάθε βήμα στον ορίζοντα πρόβλεψης. Όπως και με τις προηγούμενες μεθόδους, χρησιμοποιήθηκε η μέθοδος “Forecaster Recursive”, ωστόσο μέσω της εισαγωγής της σε ένα “for loop” μπορεί να επαναληφθεί για τα ορισμένα βήματα στον ορίζοντα πρόβλεψης. Όπως προαναφέρθηκε, οι προσεγγίσεις single-step δεν ενσωματώνουν τις προηγούμενες προβλέψεις στα δεδομένα εκπαίδευσης, επομένως δεν δημιουργείται συσωρευτικό σφάλμα για κάθε μετέπειτα πρόβλεψη. Επομένως, για να επιτευχθεί το παραπάνω, ύστερα από μια πρόβλεψη – επανάληψη του “for loop”, μπορεί εντός του να εισαχθεί η πραγματική παρατήρηση ώστε στην επόμενη επανάληψη, που θα δημιουργηθεί ένα καινούργιο μοντέλο Random Forest να χρησιμοποιήσει την παρατήρηση ως προηγούμενο βήμα και όχι την παραγόμενη παρατήρηση.

Πίνακας 60 Μετρικές σφάλματος univariate singlestep Random Forest

Mean Absolute Error	Root Mean Squared Error
0.3441877320701911	0.5381091694308399

Οι παρατηρήσεις σχετικά με την απόδοση της προσέγγισης επιβεβαιώνονται και από τις μετρικές σφαλμάτων που είναι σημαντικά χαμηλότερες σε σχέση με τις άλλες υλοποιήσεις multi-step Random Forest. Το MAE βρέθηκε ως 0,34 και το RMSE ως 0,53, δηλαδή είναι οι χαμηλότερες μετρικές σφάλματος που παρατηρούνται στις υλοποιήσεις Random Forest.



Διάγραμμα 68 Διαγράμματα καταλοίπων για univariate singlestep Random Forest

Τα διαγνωστικά γραφήματα των κατάλοιπων φαίνεται να επιβεβαιώνουν όλα τα προαναφερόμενα, όπως φαίνεται από το διάγραμμα 68. Τα κατάλοιπα πράγματι φαίνεται να κυμαίνονται γύρω από το 0 χωρίς σταθερή διακύμανση, λόγω της ύπαρξης ορισμένων ακραίων τιμών. Η κατανομή φαίνεται να είναι σχετικά κανονική ωστόσο η δεξιά ουρά συνεχίζεται να παρατηρείται ενώ επίσης και μέσω του διαγράμματος Q-Q φαίνεται ότι πλησιάζουν την διαγώνιο. Τέλος, μέσω του διαγράμματος παρατηρείται ότι τα περισσότερα lags βρίσκονται εντός ορίου εμπιστοσύνης με 2 lags να βρίσκονται οριακά εκτός, στα lags 1 και 12.

5 Συζήτηση - Συμπεράσματα

5.1 Σύγκριση αποτελεσμάτων

Βάσει όλων των πινάκων που παρουσιάστηκαν στα αποτελέσματα, παρακάτω παρουσιάζεται ο συγκεντρωτικός πίνακας των μετρικών MAE – RMSE ανά την μέθοδο.

Πίνακας 61 Συγκεντρωτικός πίνακας μετρικών σφαλμάτων

Μέθοδος	**MAE**	**RMSE**
SARMA - Multistep	0.8187566751949479	1.0850912831285437
SARIMA - Multistep	0.8752688895833016	1.1429205083403087
SARIMA - Singlestep	0.3681206289921636	0.5762088094671861
SARIMAX - Singlestep	0.36616583113619583	0.5749341444019184
RF – Univariate / Multistep	0.6532502952569917	1.0221473282777487
RF – Multivariate / Multistep	0.7123943098072563	0.9521387851528437
RF – Univariate / Singlestep	0.3441877320701911	0.5381091694308399

Ωστόσο, ίσως μια καλύτερη προσέγγιση είναι ο διαχωρισμός των μεθόδων βάσει της διάκρισης μεταξύ προσεγγίσεων single – step και multi – step, καθώς πρακτικά αφορούν διαφορετικούς χρονικούς ορίζοντες, παρόλο που θεωρητικά χρησιμοποιούν το ίδιο διάστημα πρόβλεψης, δηλαδή της 1 εβδομάδας. Όπως προαναφέρθηκε, οι single – step προσεγγίσεις χρησιμοποιούν τον ορίζοντα μίας ώρας που επαναλαμβάνεται για 1 εβδομάδα ενώ οι multi – step προσεγγίσεις αφορούν συνολικά μια εβδομάδα, λόγω της φύσης τους.

Πίνακας 62 Συγκεντρωτικός πίνακας μετρικών σφαλμάτων ανά singlestep προσεγγίσεις

Μέθοδος - Singlestep	**MAE**	**RMSE**
SARIMA - Singlestep	0.3681206289921636	0.5762088094671861
SARIMAX - Singlestep	0.36616583113619583	0.5749341444019184
RF – Univariate / Singlestep	0.3441877320701911	0.5381091694308399

Πρακτικά, όλες οι προσεγγίσεις single step παρουσιάζουν πολύ χαμηλές μετρικές σφαλμάτων, με αρκετά καλή απόδοση. Ως βέλτιστη, βρέθηκε η μέθοδος μονοδιάστατου Random Forest, καθώς παρουσιάζει το χαμηλότερο MAE (0.34) και το χαμηλότερο RMSE (0.53).

Πίνακας 63 Συγκεντρωτικός πίνακας μετρικών σφαλμάτων ανά multistep προσεγγίσεις

Μέθοδος - Multistep	**MAE**	**RMSE**
SARMA - Multistep	0.8187566751949479	1.0850912831285437
SARIMA - Multistep	0.8752688895833016	1.1429205083403087
RF – Univariate / Multistep	0.6532502952569917	1.0221473282777487
RF – Multivariate / Multistep	0.7123943098072563	0.9521387851528437

Σχετικά με τις προσεγγίσεις multistep, οι μέθοδοι μηχανικής μάθησης φαίνεται να υπερτερούν γενικά σε σχέση με τις στατιστικές μεθόδους. Ωστόσο, δεν είναι ξεκάθαρη η βέλτιστη μέθοδος, καθώς το

μονοδιάστατο μοντέλο παρουσιάζει το χαμηλότερο MAE ενώ το πολυδιάστατο παρουσιάζει το χαμηλότερο RMSE. Αυθαίρετα, θα επιλεχτεί ως βέλτιστο το μονοδιάστατο μοντέλο, δηλαδή αυτό με το χαμηλότερο MAE.

Συνοψίζοντας, στην εργασία παρουσιάζονται 4 επίπεδα συγκρίσεων. Αρχικά, γίνεται σύγκριση μεταξύ μοντέλα στατιστικής και μηχανικής μάθησης σχετικά με την προβλεπτική τους ικανότητα ως προς την χημική ένωση του διοξειδίου του άνθρακα (CO). Παράλληλα, εξετάζονται προσεγγίσεις multi – step και single – step, πράγμα που δημιουργεί ένα ακόμα επίπεδο σύγκρισης, αυτό του χρονικού ορίζοντα πρόβλεψης. Παρόλο που για όλες τις μεθόδους χρησιμοποιείται ο ίδιος χρονικός ορίζοντας της μιας εβδομάδας ή αλλιώς χρησιμοποιείται το ίδιο σετ ελέγχου μήκους 1 εβδομάδας ή 168 βήματα - ώρες, οι διαφορετικές προσεγγίσεις αφορούν διαφορετικούς χρονικούς ορίζοντες. Η προσέγγιση single – step θεωρείται βραχυπρόθεσμή ενώ η multi – step θεωρείται μεσοπρόθεσμη. Τέλος, εξετάζονται και μονοδιάστατα και πολυδιάστατα μοντέλα (univariate – multivariate), δηλαδή μοντέλα που χρησιμοποιούν μόνο τις προηγούμενες τιμές της μεταβλητής υπό μελέτη και μοντέλα που συνδυάζουν και εξωγενής – άλλες μεταβλητές πέρα από αυτήν υπό μελέτη.

Το σύνολο δεδομένων ήταν σχετικά δύσκολο ως προς την χρήση του για διάφορους λόγους. Γενικά, όπως προαναφέρθηκε, τα δεδομένα ποιότητας αέρα παρουσιάζουν ελλιπής τιμές και χαρακτηρίζονται από ετεροσκεδαστικότητα, καθώς είναι στοχαστικά από την φύση τους και συχνά παρουσιάζουν απότομες αλλαγές που δημιουργούν ακραίες τιμές. Μάλιστα, στην συγκεκριμένη περίπτωση υπήρχε σημαντικός αριθμός συνεχών διαστημάτων ελλιπών τιμών, με αποτέλεσμα η διαδικασία αντιμετώπισης ελλιπών τιμών να χρειάζεται ιδιαίτερη επεξεργασία και προσοχή.

Το πιο ενδιαφέρον αποτέλεσμα που προκύπτει, είναι ότι το ίδιο μοντέλο μπορεί να δημιουργήσει τα ίδια κατάλοιπα εκπαίδευσης αλλά λόγω της προσέγγισης single – step, μπορεί να παράγει προβλέψεις με σημαντικά χαμηλότερα σφάλματα.

Σχετικά με τα στατιστικά μοντέλα, φαίνεται ότι η χρήση μεταβλητών ή αλλιώς η εφαρμογή του μοντέλου SARIMAX, δεν επιφέρει ιδιαίτερες αλλαγές σε σχέση με ένα μοντέλο SARIMA. Πιθανώς, η χρήση του s = 24 που ορίζει την περιοδικότητα του μοντέλου ή διαφορετικά την χρονική κλίμακα των δεδομένων, σε συνδυασμό με τους συντελεστές AR και MA, είναι αρκετή ώστε το μοντέλο να ενσωματώσει την δομή των δεδομένων. Επομένως, η χρήση των κυκλικά κωδικοποιημένων μεταβλητών δεν επιφέρει ιδιαίτερες αλλαγές.

Σε αντίθεση, επειδή το μοντέλο μηχανικής μάθησης Random Forest δεν ενσωματώνει εκ φύσεως την χρονική εξάρτηση – δομή των μεταβλητών, και με την χρήση των κυκλικά κωδικοποιημένων μεταβλητών φαίνεται να υπάρχει βελτίωση στο μοντέλο. Ωστόσο, η χρήση τους τείνει να κάνει το μοντέλο να υπερεκτιμά σε σχέση με τις πραγματικές τιμές, αυξάνοντας τα σφάλματα του μοντέλου.

Επίσης, παρατηρείται ότι τα μοντέλα μηχανικής μάθησης παράγουν γενικά καλύτερα αποτελέσματα σε σχέση με τα στατιστικά μοντέλα, που συνάδει με την βιβλιογραφία όπως παρουσιάστηκε στο κεφάλαιο της βιβλιογραφικής ανασκόπησης.

5.2 Αδυναμίες της εργασίας

Προφανώς, εδώ χρειάζεται να αναφερθεί ότι δεν εκτελέστηκε πλήρης αναζήτηση των βέλτιστων συντελεστών για τα μοντέλα καθώς το παραπάνω δεν ήταν ο βασικός στόχος της παρούσας εργασίας. Επομένως, είναι πιθανό ότι με την χρήση μεγαλύτερων παραμέτρων τα στατιστικά μοντέλα να

πλησιάζουν τις μετρικές των μοντέλων μηχανικής μάθησης. Επίσης, θα μπορούσε να χρησιμοποιηθεί μια πολυδιάστατη – single – step μέθοδος RF, η οποία - πιθανώς να παρουσίαζε ακόμα καλύτερα αποτελέσματα.

Μια άλλη αδυναμία της εργασίας εντοπίζεται στον διαχωρισμό (train – test sets) των δεδομένων σε υποσύνολα. Η ετεροσκεδαστικότητα των δεδομένων μπορεί να οφείλεται σε διάφορους παράγοντες, όπως για παράδειγμα οι καλοκαιρινοί μήνες να παρουσιάζουν σταθερή και χαμηλή διακύμανση τιμών επειδή αρκετά άτομα λείπουν για διακοπές, επομένως δεν χρησιμοποιούνται μεταφορικά μέσα που επηρεάζουν άμεσα τα επίπεδα CO. Βάσει αυτού, όπως προτείνουν ο Kamińska [83], θα ήταν εφικτό να δημιουργηθούν διαφορετικά υποσύνολα βάσει κάποιος ομαδοποίησης, για παράδειγμα θερμών και ψυχρών μηνών.

5.3 Μελλοντικές Βελτιώσεις

Σχετικά με τις μελλοντικές βελτιώσεις, αρχικά προτείνεται η εφαρμογή μοντέλων βαθιάς μάθησης ώστε να είναι πιο πλήρης η σύγκριση των αλγοριθμικών προσεγγίσεων. Έπειτα, θα μπορούσε να γίνει εύρεση των βέλτιστων παραμέτρων στους αλγόριθμους που χρησιμοποιήθηκαν. Παράλληλα, θα μπορούσε να εφαρμοστεί λογαριθμικός μετασχηματισμός στο σύνολο δεδομένων όπως αναφέρουν οι Kumar και Jain, [79] για την εξάλειψη των ακραίων τιμών και της ετεροσκεδαστικότητας. Επίσης, πιθανώς η χρήση μοντέλων ARCH – GARCH, να παρήγαγε ακόμα καλύτερα αποτελέσματα στα πλαίσια των στατιστικών μεθόδων [80].

Άλλες βελτιώσεις θα μπορούσαν να εφαρμοστούν είναι η δημιουργία του “feature importance” για τα μοντέλα random forest, κυρίως για την επιβεβαίωση των σημαντικότερων lags την μεταβλητής υπό μελέτη, η δημιουργία διαγραμμάτων απόλυτων σφαλμάτων μεταξύ προσεγγίσεων single – step και Multi – step για την καλύτερη ανάδειξη των συσωρευτικών σφαλμάτων.

6 Βιβλιογραφία

[1] Liu, H., Yan, G., Duan, Z., & Chen, C. (2021). Intelligent modeling strategies for forecasting air quality time series: A review. *Applied Soft Computing*, 102, 106957, διαθέσιμο από: https://www.sciencedirect.com/science/article/abs/pii/S1568494620308954

[2] Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: A survey. *Artificial Intelligence Review, 56*, 10031–10066., διαθέσιμο από: https://link.springer.com/article/10.1007/s10462-023-10424-4

[3] De Vito, S., Massera, E., Piga, M., Martinotto, L., & Di Francia, G. (2008). On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical, 129*(2), 750–757, διαθέσιμο από: https://www.sciencedirect.com/science/article/abs/pii/S0925400507007691

[4] Liu, H., Li, Q., Yu, D., & Gu, Y. (2019). Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms. *Applied Sciences*, *9*(19), 4069, διαθέσιμο από: https://doi.org/10.3390/app9194069

[5] Kumar, D., Chaudhri, S. N., & Rajput, N. S. (2023). Air quality prediction and monitoring using machine learning-based forecasting approach. *Proceedings of the 2023 International Conference on IoT, Communication and Automation Technology (ICICAT)* (pp. 1–6), διαθέσιμο από: https://ieeexplore.ieee.org/document/10263594

[6] Hua, V., Nguyen, T., Dao, MS., Nguyen, HD., Nguyen, BT. (2024) The impact of data imputation on air quality prediction problem. PLOS ONE 19(9), διαθέσιμο από: https://journals.plos.org/plosone/article/citation?id=10.1371/journal.pone.0306303

[7] Abulkhair, A. (2023). Data Imputation Demystified, Kaggle, διαθέσιμο από: https://www.kaggle.com/code/ahmedabdulhamid/data-imputation-demystified-time-series-data#Mean/Median/Mode-Imputation

[8] Pandas development team, (xx), *User Guide: Time series / date functionality*, διαθέσιμο από: https://pandas.pydata.org/docs/user_guide/timeseries.html

[9] Pandas development team, (xx), *API reference: pandas.to_datetime*, διαθέσιμο από: https://pandas.pydata.org/docs/reference/api/pandas.to_datetime.html

[10] Pandas development team, (xx), *API reference: pandas.Series.dt*, διαθέσιμο από: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.dt.html

[11] Bedford, J. (2024) *Solving the Gaps and Islands Problem Using Python Pandas*, διαθέσιμο από: https://www.jbed.net/blog/2024-03-08-pandas-gaps-islands/

[12] Stack Overflow. (2021). *Pandas grouping by consecutive values*. διαθέσιμο από: https://stackoverflow.com/questions/69607480/pandas-grouping-by-consecutive-values

[13] Tao, C. (2020), *Pandas DataFrame Group by Consecutive Same Values.* Medium. Διαθέσιμο από: https://medium.com/data-science/pandas-dataframe-group-by-consecutive-same-values-128913875dba

[14] Data Science Stack Exchange. (2023). *Group rows partially [Python] [Pandas]*. Διαθέσιμο από: https://datascience.stackexchange.com/questions/119946/group-rows-partially-python-pandas

[15] Pandas development team, (xx), *User Guide: Timeseries - Epoch timestamps*, διαθέσιμο από: https://pandas.pydata.org/docs/user_guide/timeseries.html#epoch-timestamps

[16] NumPy development team. (xx). *API reference: Datetimes and timedeltas*. Διαθέσιμο από: https://numpy.org/doc/stable/reference/arrays.datetime.html

[17] Belachsen, I., Broday, D. M. (2022). Imputation of Missing $PM_{2.5}$ Observations in a Network of Air Quality Monitoring Stations by a New *k*NN Method. *Atmosphere*, *13*(11), 1934. Διαθέσιμο από: https://doi.org/10.3390/atmos13111934

[18]Junger, W. L., & Ponce de Leon, A. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment, 102*, 96–104. Διαθέσιμο από: https://www.sciencedirect.com/science/article/abs/pii/S1352231014009145?via%3Dihub

[19] Andrés, D. (2025). *Issue #89: Encoding cyclical features in time series*. MLPills (Substack). Διαθέσιμο από: https://mlpills.substack.com/p/issue-89-encoding-cyclical-features

[20] Sexton, J. (xx). *Unit circle calculator*. Inch Calculator, διαθέσιμο από: https://www.inchcalculator.com/unit-circle-calculator/

[21] Pelletier, H. (2024). *Cyclical encoding: An alternative to one-hot encoding for time series features*. Towards Data Science.https://towardsdatascience.com/cyclical-encoding-an-alternative-to-one-hot-encoding-for-time-series-features-4db46248ebba/

[22] Van Wyk, A. (2022). *Encoding Cyclical Features for Deep Learning*. Kaggle. Διαθέσιμο από: https://www.kaggle.com/code/avanwyk/encoding-cyclical-features-for-deep-learning

[23] Copernicus Atmosphere Monitoring Service. (2020). *Emissions changes due to lockdown measures during the first wave of the COVID-19 pandemic in Europe*. European Union Διαθέσιμο από: https://atmosphere.copernicus.eu/emissions-changes-due-lockdown-measures-during-first-wave-covid-19-pandemic-europe

[24] Green, F. (2025) *Air Quality Measurements Series: Carbon Monoxide (CO)*. Clarity. Διαθέσιμο από: https://www.clarity.io/blog/air-quality-measurements-series-carbon-monoxide-co

[25] Schwela, D. (2000). Air pollution and health in urban areas. *Reviews on Environmental Health*, 15(1–2), 13–42. Διαθέσιμο από: https://doi.org/10.1515/reveh.2000.15.1-2.13

[26] Suradhaniwar, S., Kar, S., Durbha, S. S., & Jagarlapudi, A. (2021). Time series forecasting of univariate agrometeorological data: A comparative performance evaluation via one-step and

multi-step ahead forecasting strategies. *Sensors*, 21(7), 2430. Διαθέσιμο από: https://doi.org/10.3390/s21072430

[27] Athanasiadis, I. N., Karatzas, K. D., & Mitkas, P. A. (2006). Classification techniques for air quality forecasting. In *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI) Workshop on Binding Environmental Sciences and Artificial Intelligence*. https://www.researchgate.net/publication/228778957_Classification_techniques_for_air_quality_forecasting

[28] Kaur, E., & Sandha, S. (2025). *Air Quality Prediction Using Machine Learning.* ResearchGate. Διαθέσιμο από: https://www.researchgate.net/publication/394490744_Air_Quality_Prediction_Using_Machine_Learning

[29] Mancini, S., Francavilla, A. B., Graziuso, G., & Guarnaccia, C. (2022). *An application of ARIMA modelling to air pollution concentrations during COVID pandemic in Italy. Journal of Physics: Conference Series, 2162*(1), 012009. Διαθέσιμο από: https://doi.org/10.1088/1742-6596/2162/1/012009

[30] Niako, N., Melgarejo, J. D., Maestre, G. E., & Vatcheva, K. P. (2024). *Effects of missing data imputation methods on univariate blood pressure time series data analysis and forecasting with ARIMA and LSTM. BMC Medical Research Methodology, 24,* Article 320. Διαθέσιμο από: https://doi.org/10.1186/s12874-024-02448-3

[31] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). *Methods for imputation of missing values in air quality data sets. Atmospheric Environment, 38*(18), 2895–2907. Διαθέσιμο από: https://doi.org/10.1016/j.atmosenv.2004.02.026

[32] Yang, X., Li, J., & Jiang, X. (2024). Research on information leakage in time series prediction based on empirical mode decomposition. Scientific Reports, 14, 28362. Διαθέσιμο από https://doi.org/10.1038/s41598-024-80018-9

[33] APX Machine Learning. (2026). *Selecting SARIMA Order (p, d, q)(P, D, Q)m* . Διαθέσιμο από: https://apxml.com/courses/time-series-analysis-forecasting/chapter-5-seasonal-arima-sarima/selecting-sarima-order

[34] Dabral, P. P., & Murry, M. Z. (2017). Modelling and forecasting of rainfall time series using SARIMA. *Environmental Processes, 4*(2), 399–419. https://doi.org/10.1007/s40710-017-0226-y

[35] Statsmodels development team. (xx). Seasonal-Trend decomposition using LOESS (STL). Διαθέσιμο από: https://www.statsmodels.org/devel/examples/notebooks/generated/stl_decomposition.html

[36] Hyndman, R.J., & Athanasopoulos, G. (2018). Time series decomposition in: *Forecasting: principles and practice, 2nd edition,* OTexts: Melbourne, Australia. Διαθέσιμο από:

https://otexts.com/fpp2/decomposition.html

[37] Hyndman, R.J., & Athanasopoulos, G. (2021). Classical decomposition *in Forecasting: principles and practice, 3rd edition*, OTexts: Melbourne, Australia. https://otexts.com/fpp3/classical-decomposition.html

[38] Hyndman, R.J., & Athanasopoulos, G. (2021). STL decomposition *in Forecasting: principles and practice, 3rd edition*, OTexts: Melbourne, Australia. https://otexts.com/fpp3/stl.html

[39] Hyndman, R.J., & Athanasopoulos, G. (2018). Forecasting with decomposition in: *Forecasting: principles and practice, 2nd edition,* OTexts: Melbourne, Australia. Διαθέσιμο από: https://otexts.com/fpp2/forecasting-decomposition.html

[40] Hyndman, R.J., & Athanasopoulos, G. (2021). Stationarity and differencing *in Forecasting: principles and practice, 3rd edition*, OTexts: Melbourne, Australia https://otexts.com/fpp3/stationarity.html

[41] Hyndman, R.J., & Athanasopoulos, G. (2021). Weekly, daily and sub-daily data *in Forecasting: principles and practice, 3rd edition*, OTexts: Melbourne, Australia https://otexts.com/fpp3/weekly.html

[42] APX Machine Learning. (2026). *Methods for Time Series Decomposition*. Διαθέσιμο από: https://apxml.com/courses/time-series-analysis-forecasting/chapter-2-decomposition-stationarity/decomposition-methods

[43] Dasari, N. (2025). *Time Series Forecasting Made Simple (Part 3.1): STL Decomposition*. Towards Data Science. Διαθέσιμο από: https://towardsdatascience.com/time-series-forecasting-made-simple-part-3-1-stl-decomposition-understanding-initial-trend-and-seasonality-prior-to-loess-smoothing/

[44] Hyndman, R.J., & Athanasopoulos, G. (2021). Time series components *in Forecasting: principles and practice, 3rd edition*, OTexts: Melbourne, Australia https://otexts.com/fpp3/components.html

[45] Lelwala, E., Seemasinghe, W., & Gunarathna, M. (2024). Nonparametric approach to detecting seasonality in time series: Application of the Kruskal–Wallis (KW) test on tourist arrivals to Sri Lanka. *South Asian Journal of Business Insights*, 4(1), 3–19. https://doi.org/10.4038/sajbi.v4i1.61

[46] Hansen, B. (2000). *ECON 390: Economic Forecasting – Lecture 8* [Lecture Slides]. University of Wisconsin–Madison. Διαθέσιμο από: https://users.ssc.wisc.edu/~behansen/390/2010/390Lecture8.pdf

[47] Pmdarima development team (xx). *API Reference: pmdarima.arima.auto_arima*. διαθέσιμο από: https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html#pmdarima.arima.auto_arima

[48] Vito, S. (2008). *Air Quality [Dataset].* UCI Machine Learning Repository. Διαθέσιμο από: https://archive.ics.uci.edu/dataset/360/air+quality

[49] Skforecast development team (xx). *User Guides: Recursive multi-step forecasting.* Διαθέσιμο από: https://skforecast.org/0.19.1/user_guides/autoregressive-forecaster

[50] Skforecast development team (xx). *User Guides: Direct multi-step forecaster.* Διαθέσιμο από: https://skforecast.org/0.19.1/user_guides/direct-multi-step-forecasting

[51] Skforecast development team (xx). *Intro to forecasting: Machine learning for forecasting – Single – step forecasting.* Διαθέσιμο από: https://skforecast.org/0.19.0/introduction-forecasting/introduction-forecasting.html#single-step-forecasting

[52] Hitchcock, D. (2025). STAT 520: Forecasting and Time Series – Lecture 9 [Lecture Slides]. University of South Carolina. Διαθέσιμο από: https://people.stat.sc.edu/hitchcock/stat520ch9slides.pdf

[53] Hyndman, R.J., & Athanasopoulos, G. (2018). White noise in: Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. Διαθέσιμο από: https://otexts.com/fpp2/wn.html

[54] Statsmodels development team. (xx). *Time series analysis: Stationarity and detrending (ADF/KPSS)*. Διαθέσιμο από: https://www.statsmodels.org/dev/examples/notebooks/generated/stationarity_detrending_adf_kpss.html

[55] Cerqueira, V. (2023). *3 Types of Seasonality and How to Detect Them*. Towards Data Science. Διαθέσιμο από: https://towardsdatascience.com/3-types-of-seasonality-and-how-to-detect-them-4e03f548d167/

[56] Stigler, M. (2009). *Time series econometrics – Seasonality / Lecture 8*. National Institute of Public Finance and Policy. New Delhi, India . Διαθέσιμο από https://matthieustigler.github.io/Lectures/Lect3Season.pdf

[57] Matlab development team. (2026). *Trend-Stationary vs. Difference-Stationary Processes*. Διαθέσιμο από: https://www.mathworks.com/help/econ/trend-stationary-vs-difference-stationary.html

[58] Hyndman, R.J., & Athanasopoulos, G. (2021). *Non-seasonal ARIMA models* in Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. Διαθέσιμο από: https://otexts.com/fpp3/non-seasonal-arima.html

[59] Skforecast development team (xx). *API Guide: ForecasterRecursive*. Διαθέσιμο από: https://skforecast.org/0.14.0/api/forecasterrecursive

[60] Freeman, B. S., Taylor, G., Gharabaghi, B., & Thé, J. (2018). Forecasting air quality time series using deep learning. *Journal of the Air & Waste Management Association*, 68(8), 866–886.

Διαθέσιμο από: https://doi.org/10.1080/10962247.2018.1459956

[61] Statsmodels development team. (xx). *User Guide: statsmodels.tsa.arima.model.ARIMAResults.plot_diagnostics*. Διαθέσιμο από: https://www.statsmodels.org/dev/generated/statsmodels.tsa.arima.model.ARIMAResults.plot_diagnostics.html

[62] Skforecast development team (xx). *API Reference: plot*. Διαθέσιμο από: https://skforecast.org/0.14.0/api/plot

[63] Silva, E. (2024). *Handling Gaps in Time Series*. Towards Data Science. Διαθέσιμο από: https://towardsdatascience.com/handling-gaps-in-time-series-dc47ae883990/#4f45

[64] https://otexts.com/fpp3/MA.html

[65] Kane, M. J., Price, N., Scotch, M., & Rabinowitz, P. (2014). Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1 outbreaks. *BMC Bioinformatics, 15*, 276. https://doi.org/10.1186/1471-2105-15-276

[66] Mirzadeh, H., & Omranpour, H. (2025). Extended random forest for multivariate air quality forecasting. *International Journal of Machine Learning and Cybernetics*, 16, 1175–1199. https://doi.org/10.1007/s13042-024-02329-7

[67] Hyndman, R.J., & Athanasopoulos, G. (2018). Seasonal ARIMA models in: Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. Διαθέσιμο από: https://otexts.com/fpp2/seasonal-arima.html

[68] Hyndman, R.J., & Athanasopoulos, G. (2018). Forecasting in: Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. Διαθέσιμο από: https://otexts.com/fpp2/arima-forecasting.html

[69] Brownlee, J. (2023). How to Create an ARIMA Model for Time Series Forecasting in Python. Machine Learning Mastery. Διαθέσιμο από: https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

[70] Bansal, A., Balaji, K., & Lalani, Z. (2025). Temporal encoding strategies for energy time series prediction (arXiv No. 2503.15456). arXiv. Διαθέσιμο από: https://arxiv.org/abs/2503.15456

[71] Mahajan, T., Singh, G., & Bruns, G. (2021). *An experimental assessment of treatments for cyclical data* [Conference session]. *2021 Computer Science Conference for CSU Undergraduates*. Διαθέσιμο από: https://scholarworks.calstate.edu/downloads/pv63g5147

[72] Li, W., & Law, K. (2024). Deep learning models for time series forecasting: A review. *IEEE Access*, *12*, https://www.researchgate.net/publication/381970036_Deep_Learning_Models_for_Time_Series_Forecasting_A_Review

[73] Statsmodels development team. (xx). *SARIMAX*. Διαθέσιμο από: https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html

[74] Seabold, S., & Perktold, J. (2010). *statsmodels: Econometric and statistical modeling with Python.* In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 92–96).

[75] Statsmodels development team. (xx). *SARIMAXRESULTS*. Διαθέσιμο από: https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAXResults.html

[76] Statsmodels development team. (xx). *SARIMAXRESULTS.get_forecast()*. Διαθέσιμο από: https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAXResults.get_forecast.html

[77] Statsmodels development team. (xx). *Prediction Results*. Διαθέσιμο από: https://www.statsmodels.org/dev/generated/statsmodels.tsa.base.prediction.PredictionResults.html

[78] Skforecast development team (xx). *API Reference: RandomForestRegressor*. Διαθέσιμο από: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[79] Kumar, U., & Jain, V. K. (2010). ARIMA forecasting of ambient air pollutants (O_3, NO, NO_2 and CO). *Stochastic Environmental Research and Risk Assessment, 24*(6), 751–760. https://doi.org/10.1007/s00477-009-0361-8

[80] Liu, Z., Zhu, Z., Gao, J., & Xu, C. (2021). Forecast methods for time series data: A survey. *IEEE Access, 9*, 91896–91912. https://doi.org/10.1109/ACCESS.2021.3091162

[81] Kaur, G., Gao, J., Chiao, S., Lu, S., & Xie, G. (2018). Air quality prediction: Big data and machine learning approaches. *International Journal of Environmental Science and Development, 9*(1), 8–16. https://doi.org/10.18178/ijesd.2018.9.1.1066

[82] Shah, I., Gul, N., Ali, S., & Houmani, H. (2024). Short-term hourly ozone concentration forecasting using functional data approach. *Econometrics, 12*(2), 12. https://doi.org/10.3390/econometrics12020012

[83] Kamińska, J. A. (2018). The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wrocław. *Journal of Environmental Management, 217*, 164–174. https://doi.org/10.1016/j.jenvman.2018.03.094

[84]Python Software Foundation. (n.d.). *Built-in functions: all()*. Διαθέσιμο από: https://docs.python.org/3/library/functions.html#all

[85] Pandas development team, (xx), *API Reference: pandas.DataFrame.rolling*, διαθέσιμο από: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rolling.html

[86] Αποθετήριο κώδικα, διαθέσιμο από: https://drive.google.com/drive/folders/1V6hUwZnMnwJMJXsjNkmgqs-OwHIikrh-?usp=sharing