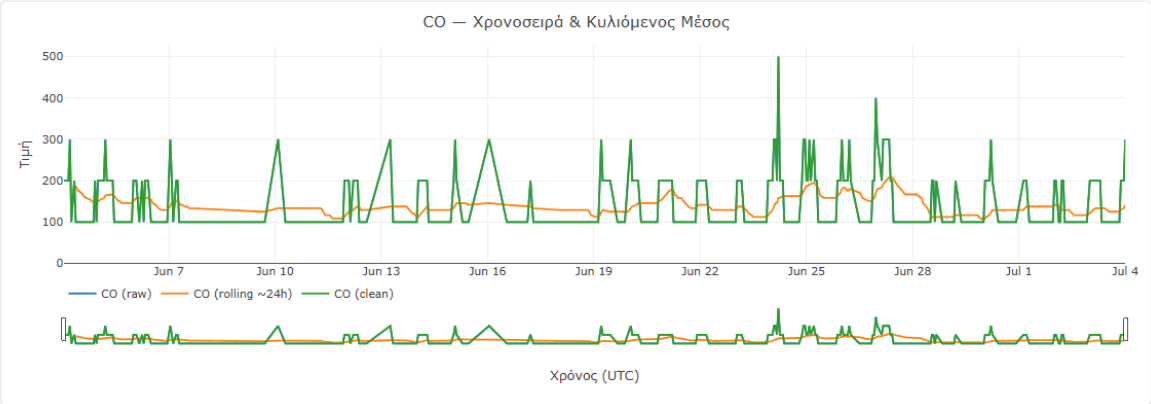
ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Σύστημα παρακολούθησης και ανάλυσης δεδομένων αισθητήρων ρύπων»



Φοιτητής

Σταφυλίδης Χρήστος

Α.Μ. 516126

Επιβλέπων

Κυριάκος Τσιακμάκης

Επίκουρος Καθηγητής

Σεπ 2025

Σύστημα παρακολούθησης και ανάλυσης δεδομένων αισθητήρων ρύπων

Κωδικός: 25175

Φοιτητής: Χρήστος Σταφυλίδης

Εισηγητής: Κυριάκος Τσιακμάκης

Ημερομηνία ανάληψης Π.Ε. 30-03-2025

Ημερομηνία περάτωσης Π.Ε. 08-09-2025

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως πτυχιακή εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η έγκριση της πτυχιακής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

Περίληψη

Η εργασία παρουσιάζει ένα Σύστημα παρακολούθησης και ανάλυσης δεδομένων αισθητήρων ρύπων που μετατρέπει ωμές μετρήσεις σε αξιόπιστη, διαδραστική μορφή για την ποιότητα αέρα. Εισάγει ωριαία δεδομένα (CO, NO, NO_2, SO_2, O_3, PM_{10}), τα ενοποιεί χρονικά σε UTC και εφαρμόζει ποιοτικό έλεγχο με ανθεκτικές μεθόδους (αρνητικές/αδύνατες τιμές, global IQR, Hampel). Παρέχεται εκδοχή με flags και διατήρηση των αρχικών. Πάνω σε αυτήν υλοποιούνται αναλύσεις χρονοσειρών (κυλιόμενοι/ημερήσιοι μέσοι, προφίλ 24ώρου, συσχέτιση, cross-correlation με υστέρηση) και ανίχνευση ανωμαλιών. Ενσωματώνονται προβλέψεις επόμενης ώρας με απλά, στιβαρά μοντέλα Naïve και rolling backtest. Οι συγκεντρώσεις χαρτογραφούνται παραμετρικά σε AQI και υπερβάσεις (ημερήσιος PM_{10}, 8ωρος O_3, ωριαίος NO_2). Η πλατφόρμα υλοποιείται σε Python/Flask με Pandas-NumPy και Plotly και παράγει αναφορά με επεξηγητικά κείμενα, πίνακες και διαδραστικά γραφήματα.

«Monitoring and Analytics System for Air Pollution Sensor Data»

Abstract

The thesis presents an Monitoring and Analytics System for Air Pollution Sensor Data transforms raw sensor readings into reliable, interactive insights on air quality. It ingests hourly data (CO, NO, NO_2, SO_2, O_3, PM10), harmonizes timestamps in UTC, and applies robust quality control (handling negative/invalid values, global IQR screening, and Hampel filtering). A “clean” version with flags is produced while preserving the originals. On this basis, the system performs time-series analyses (rolling/daily means, 24-hour profiles, correlation, and lagged cross-correlation) and anomaly detection. It incorporates next-hour forecasting using simple, robust models (Naïve) evaluated via rolling backtests. Concentrations are mapped parametrically to AQI sub-indices and exceedances (daily PM10, 8-hour O_3, hourly NO_2). The platform is implemented in Python/Flask with Pandas–NumPy and Plotly, generating a consolidated report with explanatory text, tables, and interactive charts.

Ευχαριστίες

Θέλω να ευχαριστήσω τους γονείς μου και τους καθηγητές μου που με βοήθησαν για την ολοκλήρωση της εργασίας.

Περιεχόμενα

Κατάλογος Σχημάτων

Κεφάλαιο 1ο: Εισαγωγή

1.1 Εισαγωγή

Η ποιότητα του αέρα είναι ζήτημα δημόσιας υγείας και καθημερινότητας. Ο αέρας που αναπνέουμε επηρεάζεται από μεταφορές, θέρμανση, βιομηχανία και φυσικά φαινόμενα, με αποτέλεσμα να περιέχει αιωρούμενα σωματίδια και αέρια που δεν είναι πάντα ορατά αλλά έχουν πραγματική επίδραση στην ευεξία και τη λειτουργία των πόλεων. Η παρακολούθηση της ατμοσφαιρικής ρύπανσης δεν είναι πολυτέλεια αλλά είναι βασική γνώση για τον πολίτη, τον φορέα και τον ερευνητή.

Οι βασικοί ρύποι που εξετάζονται διεθνώς είναι τα αιωρούμενα σωματίδια PM10, τα οξείδια του αζώτου (NO και κυρίως NO_2), το όζον (O_3), το διοξείδιο του θείου (SO_2) και το μονοξείδιο του άνθρακα (CO). Τα PM10 διεισδύουν στο αναπνευστικό σύστημα και σχετίζονται με αναπνευστικά και καρδιαγγειακά προβλήματα. Το NO_2 και το O_3 ερεθίζουν τους αεραγωγούς και μειώνουν τη λειτουργία των πνευμόνων, ενώ το SO_2 προκαλεί βρογχόσπασμό σε ευαίσθητες ομάδες. Το CO δεσμεύει την ικανότητα του αίματος να μεταφέρει οξυγόνο. Για αυτόν τον λόγο υπάρχουν όρια και δείκτες ποιότητας αέρα, σχεδιασμένοι ώστε να προστατεύουν την υγεία με τρόπο κατανοητό και εφαρμόσιμο.

Η μέτρηση των ρύπων είναι απαραίτητη γιατί χωρίς αξιόπιστα δεδομένα δεν μπορούμε να αναγνωρίσουμε πότε και πού ξεπερνιούνται τα όρια, ούτε να αξιολογήσουμε το αποτέλεσμα παρεμβάσεων. Τα δεδομένα δίνουν τη βάση για αποφάσεις, από τον σχεδιασμό συγκοινωνιών μέχρι την ενημέρωση των πολιτών. Η ζωντανή, σε πραγματικό χρόνο, παρακολούθηση προσθέτει μια άλλη διάσταση, επιτρέπει έγκαιρη προειδοποίηση όταν ξεκινά ένα επεισόδιο ρύπανσης, διευκολύνει λειτουργικές αποφάσεις (π.χ. αναβολή υπαίθριων δραστηριοτήτων) και βοηθά στον ποιοτικό έλεγχο του ίδιου του εξοπλισμού, αφού απότομες ασυνέχειες στις τιμές μπορεί να δηλώνουν πρόβλημα αισθητήρα ή τοπικό συμβάν.

Πέρα από τη συλλογή ένα ουσιαστικό βήμα είναι και η ανάλυση. Οι απλές χρονοσειρές, αν μείνουν χωρίς ανάλυση συχνά δίνουν αποσπασματική εικόνα. Η ανάλυση οργανώνει και καθαρίζει τα δεδομένα, εντοπίζει ακραίες τιμές και κενά με ανθεκτικές μεθόδους και κρατά ιχνηλασιμότητα. Επιπλέον αποκαλύπτει μοτίβα και τάσεις μέσω κυλιόμενων μέσων, ημερήσιων συνόψεων και προφίλ 24ώρου. Διαχωρίζει την εποχικότητα από την απλή τάση με αποσύνθεση τύπου STL και εξετάζει σχέσεις μεταξύ ρύπων μέσω συσχετίσεων και cross-correlation. Η ανίχνευση ανωμαλιών και αλλαγών καθεστώτος δείχνει στιγμές που αποκλίνουν από το «κανονικό» ενώ οι βασικές προβλέψεις επόμενης ώρας μας δίνουν καλή πληροφορία. Τέλος η χαρτογράφηση των τιμών σε δείκτες ποιότητας αέρα και υπερβάσεις μεταφράζει τεχνικά μεγέθη σε πληροφορία άμεσα αξιοποιήσιμη.

Η παρούσα εργασία υλοποιεί ένα Ηλεκτρονικό Σύστημα Παρακολούθησης Ρύπων που ενοποιεί όλη τη ροή, από το αρχείο μετρήσεων μέχρι τη διαδραστική αναφορά. Το κίνητρο ήταν διπλό. Πρώτα η ανάγκη για ένα ενιαίο, επαναλήψιμο εργαλείο που να διαχειρίζεται δεδομένα με σφάλματα, κενά και δεύτερον η ανάγκη για αποτέλεσμα χρήσιμο σε διαφορετικούς ρόλους, από φοιτητές και αναλυτές έως φορείς και πολίτες. Το σύστημα δέχεται τα δεδομένα, τα κανονικοποιεί χρονικά, εφαρμόζει ελέγχους ποιότητας με ανθεκτικά κριτήρια, εντοπίζει μοτίβα και σχέσεις, υπολογίζει δείκτες και υπερβάσεις, ανιχνεύει ανωμαλίες και παράγει βασικές προβλέψεις της επόμενης ώρας. Όλα αυτά παρουσιάζονται σε μια οργανωμένη, διαδραστική σελίδα με σαφή κείμενα μεθοδολογίας πριν από κάθε ενότητα, ώστε το αποτέλεσμα να είναι ταυτόχρονα τεχνικά ορθό και κατανοητό.

Η συνεισφορά της εργασίας είναι ότι προσφέρει έναν καλό και όμορφο τρόπο καθαρισμού και ανάλυσης, με αυστηρή διαχείριση του χρόνου και προσοχή στη στατιστική ορθότητα και ταυτόχρονα μια εμπειρία χρήστη που επιτρέπει σε μη ειδικούς να αποκτούν εικόνα χωρίς να γράφουν κώδικα. Η αρχιτεκτονική είναι αρθρωτή και επεκτάσιμη ώστε μελλοντικά να ενσωματωθούν πιο προηγμένα μοντέλα πρόβλεψης, επίσημες κλίμακες AQI ανά πλαίσιο και μετεωρολογικές παράμετροι.

Κεφάλαιο 2ο: Αισθητήρες και OpenData

2.1 Αισθητήρες

Στο σύστημά μας η μέτρηση ξεκινά από τον κόμβο αισθητήρων. Κάθε κόμβος είναι μια ολοκληρωμένη μονάδα που συλλέγει ενδείξεις για βασικούς ατμοσφαιρικούς ρύπους, τις μετατρέπει σε ψηφιακά δεδομένα με χρονοσήμανση και τις αποστέλλει για αποθήκευση και ανάλυση. Ο σχεδιασμός δίνει προτεραιότητα στη μετρολογική σταθερότητα, στη χαμηλή κατανάλωση και στη δυνατότητα συντήρησης στο πεδίο γιατί αυτό καθορίζει τελικά την αξιοπιστία όλων των επόμενων βημάτων.

Οι δικοί μας κόμβοι (από τα έτοιμα δεδομένα της πηγής μας [1]) στοχεύουν στους ρύπους που εμφανίζονται συστηματικά στο αστικό περιβάλλον: μονοξείδιο του άνθρακα (CO), οξείδια του αζώτου (NO και NO_2), διοξείδιο του θείου (SO_2), όζον (O_3) και αιωρούμενα σωματίδια PM_{10}. Για τα αέρια χρησιμοποιούνε ηλεκτροχημικές κυψέλες με κατάλληλο αναλογικό μέτωπο (conditioning, θερμοκρασιακή αντιστάθμιση και ψηφιοποίηση), επειδή προσφέρουν καλή ευαισθησία σε χαμηλές συγκεντρώσεις με μικρό ενεργειακό αποτύπωμα. Για τα σωματίδια βασίζονται σε οπτική σκέδαση φωτός, όπου η ροή αέρα διέρχεται από θάλαμο μέτρησης και η ένταση του σκεδαζόμενου φωτός μετασχηματίζεται σε συγκέντρωση PM_{10} μέσω εργοστασιακής καμπύλης και διορθώσεων υγρασίας.

Η αρχή λειτουργίας κάθε καναλιού έχει συγκεκριμένα σημεία προσοχής. Στο CO παρακολουθούνε την αργή μετατόπιση μηδενός με τον χρόνο και τη θερμοκρασία και χρησιμοποιούνε zero-offset παρακολούθηση και ήπια εξομάλυνση για να αποφεύγονται ψευδοαιχμές. Στο NO και στο NO_2 λαμβάνουνε υπόψη τη διασταυρούμενη ευαισθησία μεταξύ τους και με το O_3 γι' αυτό οι κόμβοι φέρουν και ξεχωριστό αισθητήριο O_3 και εφαρμόζουνε διορθωτικούς όρους που μειώνουν τη συνεισφορά ξένων αερίων στο σήμα. Στο SO_2 η σταθερότητα επηρεάζεται από υγρασία και θερμοκρασία, επομένως η θερμοκρασιακή αντιστάθμιση και η παρακολούθηση του περιβάλλοντος είναι απαραίτητες. Στο O_3 η φωτοχημεία του περιβάλλοντος οδηγεί σε έντονες ημερήσιες διακυμάνσεις, οπότε οι ηλεκτροχημικές ενδείξεις ελέγχονται για γραμμικότητα στο εύρος ενδιαφέροντος και συνδυάζονται με φίλτρα εξομάλυνσης που δεν αλλοιώνουν τις κορυφές του μεσημεριού. Στα PM_{10} το οπτικό σύστημα επηρεάζεται από μεγέθη γι' αυτό εφαρμόζουνε κανόνα διόρθωσης υγρασίας για να μειώνεται η σφάλμα-λόγω-νερού και προγραμματίζουνε περιοδικό καθαρισμό/έλεγχο της εισαγωγής ώστε να διατηρείται σταθερή ροή.

Κάθε πρωτογενής μέτρηση αποκτά χρονοσήμανση UTC από συγχρονισμένο ρολόι του κόμβου και οι τιμές αθροίζονται σε ωριαία προϊόντα με σαφή κανόνα (μέσος όρος ή άλλο ορισμένο σταθερά σε όλο το δίκτυο). Η χρήση UTC εξαλείφει σφάλματα θερινής ώρας και επιτρέπει δίκαιες συγκρίσεις μεταξύ κόμβων. Τα ωριαία προϊόντα που φτάνουν στην πλατφόρμα είναι ταξινομημένα αυστηρά σε αύξουσα σειρά, ώστε downstream αναλύσεις όπως συσχετίσεις, προφίλ 24ώρου και να πατούν σε συνεπή χρονοάξονα.

Οι ωριαίες μετρήσεις για CO, NO, NO_2, SO_2, O_3 και PM_{10} μάς δίνουν το πραγματικό φορτίο ρύπανσης στον χώρο και στον χρόνο. Χωρίς αυτές, δεν μπορούμε να πούμε αν μια περιοχή ξεπερνά τα όρια, αν ένα επεισόδιο ξεκινά ή τελειώνει, ούτε να εξηγήσουμε τις διαφορές από γειτονιά σε γειτονιά.

Οι τιμές τροφοδοτούν άμεσα την προστασία της υγείας. Μετρημένες συγκεντρώσεις μετατρέπονται σε δείκτες ποιότητας αέρα και σε προειδοποιήσεις, ώστε σχολεία, αθλητικές δραστηριότητες ή υπαίθριες εργασίες να προσαρμόζονται εγκαίρως. Σε βιομηχανικά ή κυκλοφοριακά περιβάλλοντα, οι μετρήσεις δείχνουν ζώνες και ώρες υψηλού κινδύνου και επιτρέπουν στοχευμένα μέτρα αντί για οριζόντιες απαγορεύσεις.

Οι τιμές είναι επίσης εργαλείο ελέγχου πολιτικών. Αν αλλάξει κάτι στην πόλη (π.χ. νέες κυκλοφοριακές ρυθμίσεις, μετατροπές καυσίμων, νέες πηγές εκπομπών), μόνο μέσα από τις χρονοσειρές μπορούμε να δούμε αν υπήρξε ουσιαστική βελτίωση ή επιδείνωση. Η ανάλυση των τάσεων, των ημερήσιων προφίλ και των συσχετίσεων απομονώνει το «σήμα» από τον θόρυβο και τεκμηριώνει αν μια παρέμβαση λειτούργησε.

Για εμάς που θέλουμε να αναλύσουμε τις τιμές βλέπουμε και την μετρολογική τους αξία. Δείχνουν drift, ασυνέχειες και ανωμαλίες του ίδιου του εξοπλισμού, απότομα «καρφιά», μηδενισμοί ή αδικαιολόγητες αποκλίσεις μεταξύ γειτονικών κόμβων ενεργοποιούν συντήρηση, επαναβαθμονόμηση ή έλεγχο τοποθέτησης.

Επίσης, με ιστορικά δεδομένα εκπαιδεύουμε απλές προβλέψεις της επόμενης ώρας και εκτιμούμε την πιθανότητα υπέρβασης ορίων ώστε οι αποφάσεις να είναι προληπτικές και όχι αντιδραστικές. Με λίγα λόγια, οι μετρήσεις των αισθητήρων μας δεν είναι απλοί αριθμοί, είναι η πληροφορία για έγκαιρη ενημέρωση.

2.2 Ανοιχτά Δεδομένα

Τα ανοιχτά δεδομένα είναι πληροφορίες που μπορεί να δει και να χρησιμοποιήσει ο καθένας, χωρίς εμπόδια [2-3]. Είναι σαν μια δημόσια βιβλιοθήκη στο διαδίκτυο: μπαίνεις, βρίσκουμε τα στοιχεία που χρειάζομαστε και τα αξιοποιουμε για μάθημα, εργασία, έρευνα. Όταν λέμε «ανοιχτά», εννοούμε ότι τα δεδομένα είναι διαθέσιμα δωρεάν, σε μορφή που κατεβαίνει εύκολα (π.χ. CSV/JSON), και με άδεια που επιτρέπει επανάχρηση αρκεί να αναφέρουμε την πηγή.

Στο θέμα της ποιότητας αέρα, τα open data έχουν μεγάλη αξία. Δίνουν σε όλους τη δυνατότητα να δουν πώς αλλάζουν οι ρύποι με την ώρα και την ημέρα, να ελέγξουν αν υπάρχουν υπερβάσεις ορίων και να συγκρίνουν περιοχές. Ένας δήμος μπορεί να τα χρησιμοποιήσει για να αποφασίσει μέτρα που μειώνουν τη ρύπανση. Ένας μαθητής μπορεί να τα κατεβάσει για να φτιάξει γράφημα ή να κάνει μια μικρή έρευνα. Ένας γονιός μπορεί να δει πότε είναι καλύτερο να γίνουν υπαίθριες δραστηριότητες. Το σημαντικό είναι ότι η πληροφορία δεν κρύβεται: είναι προσβάσιμη σε όλους, με τον ίδιο τρόπο.

Για να είναι χρήσιμα, τα ανοιχτά δεδομένα πρέπει να είναι καθαρά και τακτοποιημένα. Αυτό σημαίνει σωστές μονάδες (π.χ. μg/m³), σωστές ημερομηνίες και ώρες (ιδανικά σε κοινό ρολόι, όπως UTC), και ξεκάθαρες στήλες με ονόματα που καταλαβαίνονται. Χρειάζονται επίσης λίγες βασικές πληροφορίες γύρω τους, όπως από πού προέρχονται, πώς μετρήθηκαν και πότε ενημερώθηκαν τελευταία φορά. Με αυτά, οποιοσδήποτε μπορεί να ξανακάνει την ίδια ανάλυση και να πάρει το ίδιο αποτέλεσμα.

Υπάρχουν άδειες που λένε ακριβώς τι επιτρέπεται και συνήθως μπορείς να χρησιμοποιήσεις τα δεδομένα ελεύθερα, αλλά πρέπει να γράψεις την πηγή. Υπάρχουν επίσης θέματα ιδιωτικότητας και ασφάλειας: όταν δημοσιεύεις δεδομένα, φροντίζεις να μην αποκαλύπτεις ευαίσθητες πληροφορίες (π.χ. ακριβή διεύθυνση ενός ιδιωτικού σταθμού, αν αυτό δεν είναι σωστό να φαίνεται). Στόχος είναι να ωφεληθεί η κοινωνία, χωρίς να εκτεθεί κανείς.

Στο δικό μας σύστημα παρακολούθησης ρύπων, τα ανοιχτά δεδομένα παίζουν διπλό ρόλο. Από τη μία, μπορούμε να αντλούμε στοιχεία από δημόσιες πηγές για να εμπλουτίσουμε τις αναλύσεις μας (π.χ. σύγκριση με ευρύτερη εικόνα πόλης ή περιοχής). Από την άλλη, μπορούμε να δημοσιεύουμε τα αποτελέσματα που παράγουμε—όπως ημερήσιους μέσους, δείκτη ποιότητας αέρα (AQI) ή καταλόγους υπερβάσεων—με τρόπο απλό και επαναχρησιμοποιήσιμο. Έτσι, αυτό που μετράμε και αναλύουμε δεν μένει σε μια οθόνη γίνεται κοινό αγαθό, χρήσιμο σε σχολεία, σε ερευνητές και σε φορείς.

Κεφάλαιο 3ο: Εργαλεία υλοποίησης της πλατφόρμας

Η καρδιά της πλατφόρμας είναι η Python [4]. Την επιλέξαμε γιατί συνδυάζει εργαλεία ανάλυσης δεδομένων, γρήγορη ανάπτυξη web υπηρεσιών και ένα τεράστιο οικοσύστημα βιβλιοθηκών. Το αποτέλεσμα είναι ένας ενιαίος κώδικας, καθαρίζει και αναλύει χρονοσειρές ρύπων, δημιουργεί διαδραστικά γραφήματα και τα δίνει μέσα από ένα ελαφρύ Flask backend — όλα σε μία γλώσσα [5].

Η εφαρμογή τρέχει σε Python 3.x με απομονωμένο εικονικό περιβάλλον, ώστε οι εξαρτήσεις να είναι εύκολες. Η δομή του κώδικα ακολουθεί διαχωρισμό: ο φάκελος services/ φιλοξενεί την «εγκέφαλο» της ανάλυσης (modules όπως analysis.py, anomalies.py, forecast.py, aqi.py, plotly_helpers.py, narrative.py), ενώ το Flask app αναλαμβάνει μόνο ροές HTTP, upload αρχείων και απόδοση HTML μέσω Jinja2. Αυτή η αρχιτεκτονική κρατά το web-κομμάτι λεπτό και την ανάλυση ανεξάρτητη και επαναχρησιμοποιήσιμη.

Για τον πυρήνα της επεξεργασίας χρησιμοποιούμε Pandas και NumPy [6-7]. Οι μετρήσεις φορτώνονται σε DataFrame, τυποποιούνται οι στήλες (παράμετρος, τιμή, μονάδα, κόμβος, dt) και εφαρμόζονται ροές καθαρισμού που είναι πλήρως vectorized για απόδοση. Η βιβλιοθήκη pandas μάς επιτρέπει γρήγορο resampling σε ωριαία βάση, κυλιόμενα στατιστικά, ομαδοποιήσεις και συγχωνεύσεις σε κοινό χρονοάξονα. Το NumPy στηρίζει τις αριθμητικές πράξεις χαμηλού επιπέδου με σταθερή απόδοση πάνω σε μεγάλα arrays, ώστε ακόμη και μήνες ωριαίων μετρήσεων να επεξεργάζονται άνετα στη μνήμη.

Η χρονοσήμανση είναι κρίσιμη για ποιότητα αέρα και όλες οι σειρές μετατρέπονται σε datetime64[UTC] και ταξινομούνται αυστηρά σε αύξουσα σειρά. Οι υπολογισμοί (resampling, rolling, συσχετίσεις) εκτελούνται σε UTC για να αποφεύγονται ασυνέπειες θερινής ώρας, ενώ η παρουσίαση μπορεί να γίνει σε τοπική ζώνη με σαφή ένδειξη. Το αποτέλεσμα είναι ένας ενιαίος, συνεπής χρονικός άξονας σε όλα τα modules.

Στο τμήμα QC (Ποιοτικός Έλεγχος, Quality Control) εφαρμόζουμε ελέγχους για αρνητικές/αδύνατες τιμές, global outliers [8] με IQR[9], και εντοπισμό spikes με Hampel (rolling median και MAD) [10-11]. Οι τιμές δεν «σβήνονται»· σημαίνονται με flags και δημιουργείται «καθαρή» στήλη ανά ρύπο για downstream χρήση. Για πολύ μικρά κενά χρησιμοποιούμε συντηρητική παρεμβολή στον χρόνο, μόνο ώστε να μην «σπάνε» τα γραφήματα και οι κυλιόμενοι υπολογισμοί. Όλες οι αποφάσεις είναι αναπαραγώγιμες, καθώς οι παράμετροι συγκεντρώνονται σε μία υπηρεσία και εφαρμόζονται με τον ίδιο τρόπο σε κάθε dataset.

Οι βασικές αναλύσεις υλοποιούνται επίσης στην Python: ωριαίες χρονοσειρές με κυλιόμενο μέσο όρο για τάση, ημερήσιοι μέσοι με κανόνες πληρότητας, προφίλ 24ώρου (0–23) για τυπικό ωριαίο μοτίβο, συσχέτιση Pearson [12] σε κοινά timestamps και cross-correlation με lag για χρονική υστέρηση μεταξύ ρύπων [13]. Όπου υπάρχει διαθέσιμη βιβλιοθήκη statsmodels, αξιοποιούμε STL decomposition [14] (Trend/Seasonal/Residual, period=24 για ωριαία δεδομένα) με robust επιλογές. Όλα αυτά παράγονται

ως «sections» με έτοιμες περιγραφές μεθόδου, ώστε το αποτέλεσμα να είναι και τεχνικά σωστό και εκπαιδευτικό.

Για επιχειρησιακή χρησιμότητα ενσωματώσαμε προβλέψεις 1-βήμα μπροστά με απλές αλλά στιβαρές μεθόδους: Naïve, Seasonal Naïve (24h), Moving Average, Simple Exponential Smoothing και Drift. Η αξιολόγηση γίνεται με rolling backtest one-step-ahead και μετρικές MAE, RMSE, sMAPE, MASE [15-20]. Η Python επιτρέπει να κρατήσουμε όλη τη λογική forecasting σε ένα module (services/forecast.py) και να την καλούμε για κάθε ρύπο με κοινό interface.

Ο υπολογισμός AQI [21] στην Python γίνεται παραμετρικά: οι τιμές των ρύπων μετασχηματίζονται σε υπο-δείκτες μέσω breakpoints και γραμμικής παρεμβολής, ενώ ο συνολικός δείκτης είναι το μέγιστο ανά χρονική στιγμή. Οι υπερβάσεις υπολογίζονται σε κατάλληλα χρονικά παράγωγα (ημερήσιος PM10, 8-ώρος O_3, ωριαίος NO_2) με κανόνες πληρότητας. Ο κώδικας είναι ανεξάρτητος από συγκεκριμένο κανονιστικό πλαίσιο και δέχεται πίνακες breakpoints/thresholds, ώστε να προσαρμόζεται εύκολα.

Τα γραφήματα δημιουργούνται ως Plotly [22] JSON dictionaries στην Python και αποδίδονται στον browser χωρίς εξωτερικούς servers. Η επιλογή αυτή προσφέρει διαδραστικότητα (zoom, pan, range slider, ενοποιημένα tooltips) και συνέπεια στην όψη. Εφαρμόζουμε connectgaps και ήπια παρεμβολή μόνο για σκοπούς απεικόνισης, ώστε οι γραμμές να είναι συνεχείς όπου αυτό έχει νόημα, χωρίς να αλλοιώνονται τα πρωτογενή δεδομένα.

Η Python αναλαμβάνει και το web επίπεδο και με το Flask ορίζουμε τα endpoints για upload αρχείων και δημιουργούμε εξατομικευμένα reports ανά συνεδρία. Η Jinja2 αποδίδει τα templates, λαμβάνει από την Python τα sections (τίτλος, περιγραφή, narrtive, πίνακες, figures) και τα εμφανίζει με Bootstrap. Η λογική παραμένει στο backend: το frontend δεν «υπολογίζει», απλώς παρουσιάζει με διαδραστικό τρόπο ό,τι παράγει η ανάλυση.

Οι υπολογισμοί είναι vectorized, αποφεύγουμε loops σε Python όπου γίνεται και χρησιμοποιούμε resampling/rolling/groupby σε Pandas [23]. Η ταξινόμηση ASC και ο ενιαίος χειρισμός UTC προσφέρουν σταθερότητα σε όλες τις ροές. Η αναπαραγωγιμότητα εξασφαλίζεται με κλειδωμένες εκδόσεις εξαρτήσεων και καθαρή ιεραρχία modules· το ίδιο αρχείο εισόδου θα παράγει την ίδια αναφορά. Το σύστημα είναι self-contained, άρα μπορεί να τρέξει και χωρίς πρόσβαση στο διαδίκτυο, κάτι σημαντικό για ιδιωτικότητα και για εργαστήρια.

Ο κώδικας συνοδεύεται από τυπικές πρακτικές ποιότητας: σαφή type hints για καλύτερη τεκμηρίωση, προσεγμένο logging στην ανάλυση και στο Flask για εντοπισμό σφαλμάτων, και διακριτό χειρισμό εξαιρέσεων ώστε μια αποτυχημένη ενότητα να μην «ρίχνει» ολόκληρο το report αλλά να εμφανίζει ενημερωτικό μήνυμα στον χρήστη. Η διάσπαση σε modules επιτρέπει στοχευμένες βελτιώσεις (π.χ. μόνο στο aqi.py όταν αλλάζει ένα πλαίσιο κανόνων), χωρίς να επηρεάζεται ο υπόλοιπος κώδικας.

Η Python μας δίνει το ενιαίο υπόβαθρο για όλα: ανάγνωση και καθαρισμό μετρήσεων, στατιστική ανάλυση χρονοσειρών, παραγωγή δεικτών και υπερβάσεων, βασικές προβλέψεις και διαδραστική απεικόνιση, όλα ενσωματωμένα σε ένα απλό Flask app. Αυτή η επιλογή διασφαλίζει ότι η πλατφόρμα είναι ταυτόχρονα γρήγορη στην ανάπτυξη, πιστή στα δεδομένα και εύκολη στη συντήρηση — ακριβώς αυτό που χρειάζεται ένα ηλεκτρονικό σύστημα παρακολούθησης ρύπων που θέλει να παραμείνει χρήσιμο και να εξελίσσεται.

Κεφάλαιο 4ο: Μέθοδοι Ανάλυσης Δεδομένων

4.1 Εντοπισμός Global Outliers με IQR (Robust)

Ο δείκτης IQR (Interquartile Range) μετρά το «πάχος» του κεντρικού πυρήνα των τιμών, αγνοώντας τα άκρα. Αν οι περισσότερες τιμές μιας μεταβλητής πέφτουν σε ένα «λογικό» εύρος, ενώ λίγες τιμές είναι πολύ χαμηλές ή πολύ υψηλές, αυτές οι λίγες είναι υποψήφιες για outliers. Με το IQR:

- παίρνουμε τα τεταρτημόρια Q1 (25%) και Q3 (75%),
- ορίζουμε IQR = Q3 − Q1,
- φτιάχνουμε «φράχτες» κάτω και πάνω από την κανονική περιοχή,
- ό,τι πέφτει έξω από αυτούς τους φράχτες θεωρείται global outlier.

Λέγεται «global» γιατί βλέπει όλο το δείγμα σαν σακούλα τιμών, χωρίς να λαμβάνει υπόψη τη χρονική σειρά. Αυτό είναι πολύ χρήσιμο ως πρώτο φίλτρο για χοντρές αστοχίες.

- Q1: 25ο εκατοστημόριο (το 25% των τιμών είναι ≤ Q1)
- Q3: 75ο εκατοστημόριο (το 75% των τιμών είναι ≤ Q3)
- IQR: IQR = Q3 − Q1
- Κάτω όριο (lower fence): lower = Q1 − k * IQR
- Πάνω όριο (upper fence): upper = Q3 + k * IQR
- Κανόνας: αν value < lower ή value > upper → outlier = True

Η παράμετρος k ελέγχει την αυστηρότητα. Στα κλασικά boxplots συχνά k = 1.5. Σε περιβαλλοντικές/αισθητηριακές χρονοσειρές με φυσικά επεισόδια προτιμούμε συχνά k = 3 για λιγότερα ψευδώς θετικά.

Υπάρχουν διαφορετικοί «ορισμοί»/παραλλαγές (π.χ. Tukey, linear interpolation). Τα ακριβή Q1/Q3 μπορεί να διαφέρουν λίγο ανά λογισμικό, αλλά ο κανόνας με IQR λειτουργεί παρόμοια στην πράξη.

1ο Παράδειγμα με 5 τιμές

Ας πάρουμε 5 ωριαίες μετρήσεις μιας μεταβλητής (π.χ. CO σε μg/m3):

Τιμές: 12, 13, 14, 15, 120

Βήμα 1: Υπολογισμός Q1 και Q3

Με την «γραμμική» μέθοδο (όπως χρησιμοποιεί το pandas συνήθως):

Q1 = 13, Q3 = 15

Βήμα 2: IQR

IQR = Q3 − Q1 = 15 − 13 = 2

Βήμα 3: Φράχτες για k = 1.5 (κλασικός)

lower = 13 − 1.5*2 = 10*

upper = 15 + 1.52 = 18

Βήμα 4: Έλεγχος outliers

Οι τιμές κάτω από 10 ή πάνω από 18 είναι outliers.

Η τιμή 120 > 18, άρα είναι outlier.

Με k = 3:

lower = 13 − 3*2 = 7*

upper = 15 + 32 = 21

Πάλι 120 > 21, άρα outlier.

Παρατηρούμε ότι μια «πραγματικά» ακραία τιμή (120) θα χαρακτηριστεί outlier και με k=1.5 και με k=3. Όσο αυξάνουμε το k, τόσο λιγότερες τιμές θα θεωρούνται outliers.

Άρα

- Ο IQR βασίζεται στα Q1, Q3, όχι στη μέση τιμή/τυπική απόκλιση, άρα **δεν επηρεάζεται τόσο** από την ίδια την ακραία τιμή.
- Σε μικρά δείγματα τα Q1/Q3 μπορεί να φανούν «σκαλοπάτια», αλλά ο κανόνας εξακολουθεί να λειτουργεί.
- Το k ρυθμίζει την αυστηρότητα. Με k=3 είμαστε πιο «επιεικείς» και προστατευόμαστε από υπερ-σήμανση σε σειρές με φυσικά μεγάλα εύρη.

Ο IQR χρησιμοποιεί τη διάμεσο και τα τεταρτημόρια, όχι τον μέσο και την τυπική απόκλιση. Έτσι:

- λίγες ακραίες τιμές δεν τραβάνε τα Q1/Q3 όσο θα τραβούσαν τον μέσο/σ.τ.α.,
- ο κανόνας είναι αξιόπιστος ακόμη κι όταν η κατανομή δεν είναι κανονική (π.χ. ασύμμετρη ή με βαριές ουρές),
- είναι απλός και γρήγορος να υπολογιστεί.

Σύγκριση με z-score (mean ± N*std): ο z-score υποθέτει περίπου κανονικότητα και είναι ευαίσθητος στα ίδια τα outliers (ανεβάζουν το std). Ο IQR είναι **πιο ανθεκτικός** σε κατανομές.

Global vs Local

- Global IQR: αγνοεί τον χρόνο· βλέπει το σύνολο. Πλεονέκτημα: απλότητα, σταθερότητα, γρήγορη αναγνώριση «τελείως εκτός κλίμακας» τιμών. Μειονέκτημα: δεν «πιάνει» τοπικά spikes που είναι ακραία μόνο μέσα στο μικρό τους περιβάλλον.
- Local (π.χ. Hampel): βλέπει μικρά παράθυρα στον χρόνο, ιδανικό για στιγμιαίες αιχμές.

Στην πράξη, τα συνδυάζουμε:

1. Global IQR για γενικά ανεπίσημα άκρα.
2. Hampel για τοπικά spikes.

Έτσι καλύπτουμε και τις δύο περιπτώσεις.

Επιλογή του k (πόσο αυστηροί θέλουμε να είμαστε)

- k = 1.5: «κλασικό» για boxplots. Αρκετά αυστηρό σε πολλά σενάρια.
- k = 2.0: ενδιάμεσο.
- k = 3.0: πιο ανεκτικό, συνήθως κατάλληλο για αισθητήρια με πραγματικά επεισόδια (π.χ. αιφνίδιες αυξήσεις PM10 ή O3).

Όταν εφαρμόζουμε global IQR σε ρύπους όπως CO, NO2, O3, PM10, SO2 προσέχουμε:

1. Εποχικότητα και κύκλοι ημέρας: επειδή ο IQR είναι global, δεν ξεχωρίζει «ημέρα vs νύχτα» ή «χειμώνα vs καλοκαίρι». Αν οι τιμές αλλάζουν εποχικά, ο IQR μπορεί να γίνει πολύ «φαρδύς» ή πολύ «στενός».
2. Μικρά δείγματα: με πολύ λίγες τιμές τα Q1/Q3 δεν είναι τόσο σταθερά.
3. Πολλά ίδια ή σχεδόν ίδια: αν IQR ≈ 0 (δηλαδή Q1 ≈ Q3), πρακτικά «δεν υπάρχει εύρος». Σε αυτήν την περίπτωση ο κανόνας καταρρέει.
4. Μείγματα δεδομένων: αν έχουμε δύο τελείως διαφορετικές καταστάσεις (π.χ. «κανονικές μέρες» και «ημέρες σκόνης»), ο global IQR θα τείνει να καλύπτει και τις δύο. Κάνουμε outliers μέσα σε καθεμιά κατάσταση, δηλαδή IQR ανά κατηγορία.

2ο Παράδειγμα με 5 τιμές

Τιμές: 70, 72, 75, 78, 110

Q1 = 72 (περίπου), Q3 = 78 (περίπου)

IQR = 6

Για k = 1.5: lower = 72 − 9 = 63, upper = 78 + 9 = 87 → 110 είναι outlier

Για k = 3: lower = 72 − 18 = 54, upper = 78 + 18 = 96 → 110 είναι outlier

Το 110 είναι λιγότερο ακραίο από το 120 του πρώτου παραδείγματος, αλλά ακόμη έξω από το «φυσιολογικό» εύρος που περιγράφει το κεντρικό 50% των τιμών.

Δεν αρκεί μόνο το IQR

- Τοπικά spikes που συμβαίνουν για 1–2 δείγματα σε ένα κατά τα άλλα «ήσυχο» κομμάτι σειράς. Θέλουμε και τοπικό φίλτρο (π.χ. Hampel).

- Δεδομένα με ισχυρή εποχικότητα ή αλλαγές καθεστώτος (regime shifts). Εκεί πάμε σε IQR ανά «κατάσταση» (π.χ. εποχή, ώρα).

Ο IQR είναι ένα απλό, γρήγορο και ανθεκτικό εργαλείο για να εντοπίσουμε global outliers που δεν ταιριάζουν με τον κεντρικό πυρήνα των τιμών. Δουλεύει καλά ως πρώτο φίλτρο, ειδικά όταν δεν θέλουμε να υποθέσουμε κανονικότητα ή όταν λίγα ακραία σημεία θα αλλοίωναν τον μέσο και την τυπική απόκλιση. Για δεδομένα αισθητήρων, συνιστάται να ξεκινάμε με k = 3 και να συνδυάζουμε το IQR με ένα τοπικό φίλτρο όπως το Hampel, ώστε να καλύπτούμε και τις στιγμιαίες αιχμές.

4.2 Ανίχνευση Τοπικών Spikes με Hampel (rolling median + MAD)

Το Hampel είναι ένα «τοπικό ραντάρ» για στιγμιαίες αιχμές (spikes). Αντί να κοιτά όλο το δείγμα (global), κοιτάζει κάθε σημείο μέσα σε ένα μικρό, συρόμενο παράθυρο γύρω του και το συγκρίνει με την τοπική διάμεσο και τη MAD (Median Absolute Deviation).

Επειδή βασίζεται σε διάμεσο και απόλυτες αποκλίσεις (όχι σε μέσο και τυπική απόκλιση), είναι robust: λίγες ακραίες τιμές μέσα στο παράθυρο δεν το «στραβώνουν».

Με απλά λόγια:

Αν μια τιμή απέχει πολύ από το τυπικό επίπεδο των γειτόνων της (όπως αυτό αποτυπώνεται από τη διάμεσο), τη σημαίνουμε ως spike.

Ορισμοί και κανόνας (plain text)

Δουλεύουμε με παράθυρο μήκους w (μονοί αριθμοί συνήθως, π.χ. w = 13) το οποίο κινείται κατά μήκος της χρονοσειράς. Για το σημείο x(t):

1. Υπολογίζουμε rolling median στο παράθυρο: med = median(τιμών στο παράθυρο).
2. Υπολογίζουμε απόλυτη απόκλιση από τη median: d = |x(t) − med|.
3. Υπολογίζουμε MAD στο ίδιο παράθυρο: MAD = median(|x(i) − med|) για όλα τα i του παραθύρου.
4. Προσεγγίζουμε την «τοπική τυπική απόκλιση» ως: sigma ≈ 1.4826 * MAD.
5. Κανόνας Hampel: αν d > n_sigma * sigma, τότε το x(t) θεωρείται spike (hampel = True). Συνήθεις τιμές: n_sigma = 3.5 έως 6.0.

Ο συντελεστής 1.4826 είναι το κλασικό scale factor που κάνει τη MAD συγκρίσιμη με τυπική απόκλιση όταν η κατανομή είναι περίπου κανονική. Το ωραίο με το Hampel είναι ότι δεν χρειάζεται να είναι κανονική για να δουλέψει καλά.

1ο Παράδειγμα με 5 διαδοχικές τιμές (μία μεταβλητή)

Ας υποθέσουμε ωριαίες μετρήσεις (π.χ. CO σε μg/m3). Παίρνουμε 5 συνεχόμενες τιμές και θα ελέγξουμε ΜΟΝΟ το κεντρικό σημείο με παράθυρο w = 5 (ώστε να «χωράει» αριστερά και δεξιά συμμετρικά).

Χρόνος (UTC) και τιμές (raw):

t0: 150

t1: 158

t2: 510 ← ύποπτη αιχμή

t3: 162

t4: 155

Βήμα 1: rolling median στο παράθυρο [150, 158, 510, 162, 155]

Διατεταγμένες: [150, 155, 158, 162, 510]

median = 158

Βήμα 2: απόλυτες αποκλίσεις από τη median

|150 − 158| = 8

|158 − 158| = 0

|510 − 158| = 352

|162 − 158| = 4

|155 − 158| = 3

Βήμα 3: MAD = median(8, 0, 352, 4, 3) = 4 (οι αποκλίσεις σε σειρά: [0, 3, 4, 8, 352])

Βήμα 4: sigma ≈ 1.4826 * MAD = 1.4826 * 4 = 5.9304

Βήμα 5: Κανόνας με n_sigma = 5

Κατώφλι = n_sigma * sigma = 5 * 5.9304 ≈ 29.652

Απόκλιση του κεντρικού σημείου (t2): d = |510 − 158| = 352

Αποτέλεσμα: 352 > 29.652 ⇒ **spike**. Το t2 σημαίνεται ως hampel = True.

Αν αλλάξω n_sigma = 3.5, το κατώφλι πέφτει ≈ 20.756, και πάλι 352 > 20.756 ⇒ spike.

Για πολύ ήπιες αιχμές, η επιλογή n_sigma κάνει διαφορά.

Πώς χειριζόμαστε τα spikes στην πράξη

Το Hampel απαντά «ανιχνεύθηκε spike; ναι/όχι». Μετά:

- Συνήθως δεν διαγράφουμε τιμές. Τις σημαίνουμε (flag) και τις αντικαθιστούμε προσωρινά με κενό (NaN) για να μην επηρεάσουν υπολογισμούς/γραφικά.
- Αν το κενό είναι μικρό (π.χ. 1–2 διαδοχικά σημεία), κάνουμε γραμμική παρεμβολή στον χρόνο.
- Πάντα σημειώνουμε τι παρεμβλήθηκε (imputed = True). Τα raw διατηρούνται στην άκρη για αναθεώρηση.

Επιλογή παραμέτρων (w και n_sigma)

Παράθυρο w (μήκος):

- Ωριαία δεδομένα: συχνά w ≈ 13 (περίπου μισή μέρα).
- Αν τα γεγονότα που θες να κρατήσεις είναι βραχύβια, κράτα w μικρό για να μην «ισοπεδώνεται» η median.
- Για μη συστηματικές διακυμάνσεις (π.χ. κύκλο 24ώρου), βάζουμε w αρκετά μεγαλύτερο από τον τυπικό χρόνο μεταβολής που σε «ενοχλεί».
- Προτιμάμε μονούς αριθμούς ώστε η rolling median να έχει καθαρό κέντρο.

n_sigma (κατώφλι):

- 5.0 = συντηρητικό (λίγα false positives, ίσως χάσει ήπιες αιχμές).
- 4.0…3.5 = πιο ευαίσθητο (πιάνει περισσότερα spikes, αλλά αυξάνει false positives).
- Ξεκινάμε με 5.0 και, αν «περνάνε» ανεπιθύμητες αιχμές, κατεβάζουμε σταδιακά.

Σχέση με IQR και γιατί τα συνδυάζουμε

- **IQR (global)**: Βλέπει όλο το δείγμα σαν σακούλα τιμών· καλό για εξωφρενικές τιμές σε γενικό επίπεδο. Δεν έχει αίσθηση «γειτονιάς στον χρόνο».
- **Hampel (local)**: Βλέπει το **τοπικό** πλαίσιο. Αν μια τιμή είναι οκ γενικά αλλά «βγάζει μάτι» σε ένα μικρό τμήμα της σειράς, το Hampel θα τη βρει.

  Ο συνδυασμός δίνει καθαρή σειρά χωρίς να κρύβει πραγματικές, σταδιακές μεταβολές.

Τεχνικές λεπτομέρειες/παγίδες

- **Όρια παραθύρου**: Στην αρχή/τέλος της σειράς το παράθυρο δεν «χωράει». Συνήθως χρησιμοποιούμε min_periods (π.χ. τουλάχιστον w/2 σημεία) ή αγνοούμε τις άκρες.
- **MAD = 0**: Αν όλες οι τιμές στο παράθυρο είναι ίδιες ή σχεδόν ίδιες, τότε MAD ≈ 0, άρα sigma ≈ 0 και ο κανόνας θα «χαλάσει».

- **Ανομοιόμορφη δειγματοληψία**: Το Hampel μετρά σε πλήθος δειγμάτων, όχι σε πραγματικό χρόνο. Αν τα βήματα χρόνου είναι άνισα, προτιμάμε παράθυρο σε χρόνια μονάδα (π.χ. 12 ώρες) με rolling που βασίζεται στον χρόνο (σε κώδικα).
- **Πολυ-αιχμές**: Αν υπάρχουν δύο spikes μέσα στο ίδιο παράθυρο, η median/MAD μπορεί να τα «αντέξει», αλλά ίσως χρειαστεί δεύτερο πέρασμα ή ελαφρώς μικρότερο w.
- **Ασυμμετρίες**: Σε έντονα ασύμμετρες τοπικές κατανομές, ο κανόνας παραμένει robust, αλλά η ερμηνεία θέλει προσοχή.

2ο Μικρό δεύτερο παράδειγμα (ήπια αιχμή)

Τιμές σε 5ωρο: 120, 118, 145, 119, 121 (w = 5, κέντρο το 145)

Διατεταγμένες: [118, 119, 120, 121, 145]

rolling median = 120

Απόλυτες αποκλίσεις: [|120−120|=0, |118−120|=2, |145−120|=25, |119−120|=1, |121−120|=1]

MAD = median(0, 1, 1, 2, 25) = 1

sigma ≈ 1.4826 * 1 = 1.4826

Με n_sigma = 5: κατώφλι = 7.413

d για το 145: 25 > 7.413 ⇒ spike

Με n_sigma = 6: κατώφλι = 8.896

25 > 8.896 ⇒ πάλι spike

Το Hampel είναι ευαίσθητο σε μικρά παράθυρα — εδώ μια μέτρια αιχμή θεωρείται spike, και λογικά, γιατί ξεφεύγει καθαρά από τους γείτονες.

Το Hampel είναι ιδανικό για βραχύβια, τοπικά spikes σε χρονοσειρές αισθητήρων, επειδή:

- χρησιμοποιεί διάμεσο και MAD (ανθεκτικά στα outliers),
- είναι τοπικό — σέβεται το περιβάλλον κάθε σημείου,
- έχει δύο «κουμπιά» ρύθμισης (w και n_sigma) για να ταιριάζει σε διαφορετικά σενάρια.

Σε συνδυασμό με το IQR (global) και μια συντηρητική παρεμβολή μικρών κενών, προσφέρει καθαρές σειρές χωρίς να θυσιάζουμε την πραγματική πληροφορία των μετρήσεων.

4.3 Παρεμβολή μικρών κενών (time-based interpolation)

Γιατί εμφανίζονται κενά και πότε «επιτρέπεται» να τα γεμίζουμε

Σε πραγματικά δεδομένα αισθητήρων υπάρχουν αναπόφευκτα κενά: απώλειες επικοινωνίας, διακοπές ρεύματος, επανεκκινήσεις συσκευών, στιγμιαία σφάλματα. Η παρεμβολή μικρών κενών είναι μια προσεκτική πρακτική με δύο στόχους:

1. Να διατηρήσουμε τη συνέχεια της χρονοσειράς για υπολογισμούς που τη χρειάζονται (κινητοί μέσοι, ημερήσιοι μέσοι, συσχετίσεις).
2. Να μην αλλοιώσουμε τη φυσική μεταβλητότητα: γεμίζουμε μόνο μικρά κενά, και πάντα τα σημειώνουμε (imputed) για πλήρη διαφάνεια.

Βασική αρχή: παρεμβάλλουμε λίγο και συντηρητικά. Μεγάλα κενά μένουν κενά.

Τι σημαίνει «μικρό κενό» (πολιτική limit)

Ορίζουμε ένα όριο "limit" πέρα από το οποίο **δεν** παρεμβάλλουμε. Το όριο μπορεί να εκφραστεί:

- ως **πλήθος διαδοχικών δειγμάτων** (π.χ. έως 2 συνεχόμενα NaN σε ωριαία σειρά), ή/και
- ως **μέγιστη διάρκεια** (π.χ. έως 3 ώρες πραγματικού χρόνου).

Παράδειγμα πολιτικής:

"Παρεμβάλλουμε έως 2 συνεχόμενα κενά ή έως 3 ώρες, όποιο έρθει πρώτο, και μόνο όταν το κενό έχει έγκυρες τιμές πριν και μετά (anchors)."

Μέθοδος: γραμμική παρεμβολή στον χρόνο

Όταν υπάρχει έγκυρη τιμή **πριν** (t0, y0) και **μετά** (t1, y1) από το κενό, η γραμμική παρεμβολή δίνει την τιμή στη στιγμή t ως:

$y(t) = y0 + (y1 − y0) * ((t − t0) / (t1 − t0))$

Σημεία-κλειδιά:

- Αν το δείγμα είναι **τακτικό** (π.χ. ακριβώς ανά ώρα), τότε (t − t0) / (t1 − t0) ισούται με το **κλασματικό βήμα** μέσα στο πλήθος των ωρών.
- Αν το δείγμα είναι **ανομοιόμορφο**, χρησιμοποιούμε **πραγματικούς χρόνους** (διαφορά σε λεπτά/δευτερόλεπτα).
- Δεν παρεμβάλλουμε στην **αρχή** ή στο **τέλος** της σειράς όταν λείπει μία από τις δύο άγκυρες (δεν υπάρχει y0 ή y1).
- Κάθε παρεμβληθείσα τιμή σημειώνεται με **imputed = True**.

Διδακτικό Παράδειγμα Α: 5 ωριαίες τιμές με 2 συνεχόμενα κενά

Μεταβλητή: CO (μg/m3), ωριαίο βήμα.

Χρόνος (UTC) — Τιμή (raw)

2025-06-10 00:00 — 160

2025-06-10 01:00 — 155

2025-06-10 02:00 — NaN ← κενό

2025-06-10 03:00 — NaN ← κενό

2025-06-10 04:00 — 150

Έχουμε δύο διαδοχικά κενά. Το όριο είναι limit = 2, άρα **επιτρέπεται** παρεμβολή, επειδή υπάρχουν έγκυρες πριν και μετά (155 στις 01:00, 150 στις 04:00).

Βήμα ανά ώρα = (150 − 155) / (4 − 1) = −5/3 ≈ −1.67

2025-06-10 02:00: 155 − 1.67 = 153.33

2025-06-10 03:00: 153.33 − 1.67 = 151.67

Καθαρή σειρά: 160, 155, 153.33, 151.67, 150

Σημείωση: = True για 02:00 και 03:00.

Διδακτικό Παράδειγμα Β: κενό μεγαλύτερο από το όριο

Ίδια λογική, αλλά με 3 συνεχόμενα κενά:

Χρόνος — Τιμή (raw)

00:00 — 160

01:00 — 155

02:00 — NaN

03:00 — NaN

04:00 — NaN

05:00 — 150

limit = 2, αλλά έχουμε 3 NaN στη σειρά. **Δεν** παρεμβάλλουμε — παραμένουν NaN.

Η τεχνητή γεφύρωση τριών ωρών αυξάνει τον κίνδυνο παραποίησης τάσεων.

Διδακτικό Παράδειγμα Γ: ανομοιόμορφη δειγματοληψία

Μεταβλητή: O3 (μg/m3), ακανόνιστα timestamps.

Χρόνος — Τιμή (raw)

10:00 — 80

10:30 — NaN ← κενό

11:15 — 92

Θέλουμε την τιμή στις 10:30 (στο μέσο 10:00 → 11:15 δεν είναι ακριβώς μέσος χρόνος).

Διάρκεια t0 → t1 = 75 λεπτά

Διάρκεια t0 → t = 30 λεπτά

Λόγος = 30/75 = 0.4

y(10:30) = 80 + (92 − 80) * 0.4 = 80 + 12 * 0.4 = 84.8

Παρατήρηση: σε ανομοιόμορφα βήματα, η γραμμική παρεμβολή με χρόνο είναι πιο ακριβής από το «απλό μέσο».

Πότε να προτιμήσω άλλη μέθοδο

Η γραμμική είναι το «πρώτο εργαλείο». Άλλες επιλογές:

- Forward/Backward Fill (κρατά την τελευταία/επόμενη τιμή): χρήσιμο για ελάχιστα κενά με πολύ ήσυχες σειρές. Μειώνει τεχνητά τη μεταβλητότητα.
- Κινητός μέσος (moving average): όχι πραγματική παρεμβολή σημείου, αλλά εξομάλυνση — δεν γεμίζει NaN.
- Spline/LOESS: πιο ομαλές καμπύλες, αλλά μπορούν να δημιουργήσουν τεχνητές κορυφές/κοιλάδες. Προσοχή σε επεισοδιακές σειρές (π.χ. PM10).
- Στοχαστικά φίλτρα (π.χ. Kalman): ισχυρά σε δυναμικά συστήματα, αλλά απαιτούν μοντελοποίηση και είναι πιο «βαριά».

Για αισθητήρες ποιότητας αέρα, ο κανόνας «λίγο, γραμμικά, μόνο όταν έχει έγκυρες» κρατά την ισορροπία μεταξύ συνέχειας και ορθότητας.

Επιπτώσεις στη στατιστική

- Μέσος: μικρές παρεμβολές έχουν μικρή επίδραση, αλλά αν συμπληρώσεις μαζικά κενά, ο μέσος θα «τραβήξει».
- Διακύμανση/συντελεστής μεταβλητότητας: η παρεμβολή μειώνει τη διακύμανση σε σχέση με τα «κενά» spikes — γι’ αυτό δεν παρεμβάλλουμε μεγάλα κενά.
- Συσχετίσεις μεταξύ παραμέτρων: αν παρεμβάλλεις σε μία παράμετρο αλλά όχι στην άλλη, η συσχέτιση μπορεί να αλλάξει. Φρόντισε να κρατάς flags και να γνωρίζεις πού έγινε imputation.

Επιλογή παραμέτρων και πολιτική

- limit (σε πλήθος): τυπικά 1–2 δείγματα για ωριαία δεδομένα.
- limit (σε χρόνο): π.χ. έως 2–3 ώρες, ειδικά αν το δείγμα δεν είναι αυστηρά ωριαίο.
- έγκυρες: παρεμβολή μόνο όταν υπάρχει έγκυρη τιμή πριν και μετά.
- ζώνη ώρας: δουλεύουμε σε UTC για συνέπεια, ιδίως αν συνδυάζουμε αισθητήρες.
- ανά παράμετρο: σε ρύπους με απότομες φυσικές μεταβολές (π.χ. NOx) ίσως χρειάζεται πιο μικρό limit από ό,τι σε πιο «αργές» σειρές (π.χ. CO).

Συνηθισμένες παγίδες

- Παρεμβολή σε μεγάλα κενά «για να φαίνονται ωραία τα γραφήματα». Το αποτέλεσμα είναι παραπλανητικό.
- Αγνόηση ανομοιόμορφου ρυθμού: χρήση «βήματος 1» ενώ τα timestamps έχουν διαφορά 30 ή 90 λεπτών. Πάντα χρησιμοποίησε πραγματικούς χρόνους.
- Παρεμβολή σε κορυφές επεισοδίων: η γραμμική θα "κόψει" το peak. Αν το επεισόδιο είναι σημαντικό, προτίμησε να αφήσεις το κενό.
- Μη καταγραφή imputed: χάνεται η ιχνηλασιμότητα. Πάντα κράτα flags.

Η παρεμβολή μικρών κενών είναι ένα στοχευμένο εργαλείο για να κρατήσουμε τις χρονοσειρές αξιοποιήσιμες χωρίς να εισάγουμε τεχνητή πληροφορία. Με μια συντηρητική πολιτική limit, γραμμική παρεμβολή στον χρόνο και σαφή σηματοδότηση (imputed), πετυχαίνουμε πρακτική συνέχεια για αναλύσεις (π.χ. κυλιόμενοι/ημερήσιοι μέσοι, συσχετίσεις) ενώ παραμένουμε ειλικρινείς για το πού λείπουν δεδομένα.

4.4 Ποιοτικός Έλεγχος (QC) & Καθαρισμός Χρονοσειρών Αισθητήρων

Τα «ακατέργαστα» δεδομένα αισθητήρων είναι πολύτιμα, αλλά σπάνια είναι καθαρά. Θόρυβος μέτρησης, στιγμιαίες διακοπές τροφοδοσίας/επικοινωνίας, επανεκκινήσεις συσκευών, ή πραγματικά αλλά βραχύβια επεισόδια (spikes) μπορούν να αλλοιώσουν τα στατιστικά και τα γραφήματα. Ο Ποιοτικός Έλεγχος (Quality Control, QC) στοχεύει να:

- εντοπίσει **φανερά λανθασμένες τιμές** (π.χ. αρνητικές συγκεντρώσεις),
- εντοπίσει **ακραίες αποκλίσεις** (outliers) που είναι ασύμβατες με τη γενική κατανομή,
- εντοπίσει **spikes** (στιγμιαίες αιχμές) που δεν αντιπροσωπεύουν σταθερό φαινόμενο,
- **μην καταστρέψει** τα πρωτογενή δεδομένα: να σημαδέψει (flag) τι διορθώθηκε,
- **ανασυνθέσει** μικρά κενά με ασφαλή παρεμβολή, ώστε να συνεχίσει η ανάλυση.

Στόχος δεν είναι να «ωραιοποιήσουμε» τα δεδομένα, αλλά να αφαιρέσουμε τα εμφανώς λανθασμένα στοιχεία και να μειώσουμε τον τυχαίο θόρυβο που θολώνει τα συμπεράσματα.

Η μέθοδος που εφαρμόζουμε — βήμα προς βήμα

Έλεγχος αρνητικών τιμών

Για συγκεντρώσεις ρύπων (π.χ. σε μg/m3), αρνητικές τιμές δεν έχουν φυσικό νόημα. Κάθε τιμή < 0 σημειώνεται ως negative = True και δεν χρησιμοποιείται ως έχει.

Global outliers με IQR (robust)

Υπολογίζουμε τα τεταρτημόρια και το IQR στη συνολική σειρά:

- Q1 = 25ο εκατοστημόριο
- Q3 = 75ο εκατοστημόριο
- IQR = Q3 - Q1
- Κατώτερο όριο (lower) = Q1 − k*IQR
- Ανώτερο όριο (upper) = Q3 + k*IQR

Με k = 3 (πιο «χαλαρό» από το κλασικό 1.5) μειώνουμε τα ψευδώς θετικά σε σειρές με μεγάλη φυσική διακύμανση. Ό,τι είναι < lower ή > upper σημειώνεται ως iqr = True.

Τοπικά spikes με φίλτρο Hampel (rolling median + MAD)

Το Hampel εντοπίζει στιγμιαίες αιχμές σε κυλιόμενο παράθυρο μήκους w (π.χ. 13 χρονικά βήματα):

- rolling median = διάμεσος των τιμών μέσα στο παράθυρο,
- MAD = median(|x − rolling median|) μέσα στο ίδιο παράθυρο,
- Προσέγγιση τυπικής απόκλισης: sigma ≈ 1.4826 * MAD,
- Κανόνας σήμανσης: αν |x − rolling median| > n_sigma * sigma, τότε hampel = True (π.χ. n_sigma = 5).

Το Hampel «βλέπει» τοπικά spikes που ο IQR συχνά χάνει.

Χειρισμός flags & παρεμβολή μικρών κενών

- Όλα τα σημεία με negative ή/και iqr ή/και hampel γίνονται προσωρινά NaN.
- Αν τα κενά είναι μικρά (π.χ. έως 2 διαδοχικά δείγματα) και υπάρχουν τιμές πριν/μετά, γίνεται γραμμική παρεμβολή στον χρόνο.
- Τα σημεία που γεμίζονται σημειώνονται imputed = True. Τα πρωτογενή (raw) δεδομένα παραμένουν διαθέσιμα.

Σημείωση: σε μεγάλα κενά δεν παρεμβάλλουμε (αποφεύγουμε τεχνητές τάσεις).

Μικρό παράδειγμα με 5 τιμές (μία μεταβλητή)

Παράδειγμα 5 ωριαίων μετρήσεων CO (μονάδες: μg/m3):

Χρόνος (UTC)	Raw τιμή
2025-06-10 00:00	160
2025-06-10 01:00	155
2025-06-10 02:00	-2
2025-06-10 03:00	520
2025-06-10 04:00	150

Έλεγχος αρνητικών

Η τιμή -2 είναι αδύνατη φυσικά → negative = True.

IQR (global) με k = 3

Τα raw ταξινομημένα: [-2, 150, 155, 160, 520]

- Q2 (διάμεσος) = 155
- Q1 = διάμεσος των [-2, 150] = 74
- Q3 = διάμεσος των [160, 520] = 340
- IQR = 340 − 74 = 266
- lower = 74 − 3*266 = −724
- upper = 340 + 3*266 = 1138

Καμία τιμή δεν υπερβαίνει αυτά τα (ευρύχωρα) όρια → iqr = False για όλες.

Hampel με παράθυρο w = 3 και n_sigma = 5

Ελέγχουμε το 520 μέσα στο παράθυρο [160, 520, 150]:

- rolling median = median(160, 520, 150) = 160
- |x − rolling median|: 0, 360, 10
- MAD = median(0, 360, 10) = 10
- sigma ≈ 1.4826 * 10 = 14.826
- Κατώφλι = n_sigma * sigma = 5 * 14.826 ≈ 74.13
- Απόκλιση 520 από τη median: 360 > 74.13 ⇒ hampel = True.

Παρεμβολή μικρών κενών

Μετατρέπουμε τα flagged σε NaN: [160, 155, NaN, NaN, 150].

Υποθέτουμε ότι παρεμβάλλουμε μέχρι 2 διαδοχικά κενά. Έχουμε 2 (στις 02:00, 03:00) και γνωστές τιμές πριν/μετά (155 και 150), άρα επιτρέπεται γραμμική παρεμβολή.

Γραμμική παρεμβολή από 155 (01:00) προς 150 (04:00):

- Βήμα ανά ώρα = (150 − 155) / (4 − 1) = −5/3 ≈ −1.67
- 02:00 → 155 − 1.67 = 153.33

- 03:00 → 153.33 − 1.67 = 151.67

Καθαρή σειρά (clean): [160, 155, 153.33, 151.67, 150]

Σημειώνουμε True για τις δύο παρεμβληθείσες τιμές.

Τι συμπεραίνουμε από τη μέθοδο

1. Αρνητικές τιμές: απορρίπτονται στη φυσική ερμηνεία. Η επισήμανσή τους προστατεύει όλα τα downstream στατιστικά.
2. IQR (global): λειτουργεί σαν «φράκτης» ενάντια σε ακραία outliers με βάση τη συνολική κατανομή. Με k = 3 αποφεύγουμε να κόβουμε νόμιμες υψηλές τιμές σε ρύπους με μεγάλα εύρη. Αν θέλουμε αυστηρότερο έλεγχο, μικραίνουμε το k.
3. Hampel: εντοπίζει τοπικά spikes που ο IQR δεν πιάνει. Είναι robust (διάμεσος + MAD) και προσαρμόζεται στο τοπικό πλαίσιο.
4. Παρεμβολή μικρών κενών: επιτρέπει συνεχή ανάλυση χωρίς να εισάγονται τεχνητές τάσεις, εφόσον:
    - τα κενά είναι μικρά (π.χ. ≤ 2–3 διαδοχικά δείγματα),
    - υπάρχουν τιμές εκατέρωθεν,
    - καταγράφουμε τι παρεμβλήθηκε (imputed = True).

Συνδυαστικά, τα παραπάνω μειώνουν θόρυβο και λάθη χωρίς να αλλοιώνεται η στατιστική ουσία της σειράς.

Τι πληροφορία παίρνουμε ανά παράμετρο (ενδεικτικά)

- CO: συνήθως αντικατοπτρίζει καύσεις/κυκλοφορία. Spikes μπορεί να συνδέονται με διερχόμενα οχήματα ή κοντινές πηγές. Η καθαρή σειρά αναδεικνύει την ημερήσια «αναπνοή» της πόλης.
- NO / NO2: ευαίσθητα στην κυκλοφορία και στον καιρό. Τοπικά spikes είναι συχνά πραγματικά· το Hampel βοηθά να ξεχωρίζουμε στιγμιαίο από επίμονο επεισόδιο.
- O3: εμφανίζει μεσημβρινούς μέγιστους λόγω φωτοχημείας. Spikes εκτός μοτίβου μπορεί να είναι τεχνικά. Με IQR (k = 3) αποφεύγουμε αδικαιολόγητο «κόψιμο» φυσικών κορυφών.
- PM10: αιχμές από σκόνη, κυκλοφορία, έργα, άνεμο. Ο συνδυασμός IQR + Hampel είναι κρίσιμος, γιατί ένα spike μπορεί να είναι πραγματική εκπομπή.
- SO2: σχετίζεται με καύσιμα θείου/βιομηχανία. Σπάνια αλλά έντονα επεισόδια χρειάζονται QC για να ξεχωρίσουμε μετρήσεις από artifacts.

Κοινή αρχή: δεν εξαφανίζουμε τις τιμές αλλά τις σημειώνουμε, τις αντικαθιστούμε προσωρινά για να τρέξουν οι δείκτες/γραφήματα και κρατάμε πάντα την πρόσβαση στα raw.

Ο QC που εφαρμόσαμε είναι robust, διαφανής και αναστρέψιμος: σημαδεύει ό,τι διορθώνεται, αφήνει τα raw ανέπαφα και παράγει «καθαρή» εκδοχή κατάλληλη για στατιστική/οπτική ανάλυση. Με λίγες ρυθμίσεις (IQR – Hampel – Interpolation) βελτιώνουμε σημαντικά την αξιοπιστία των συμπερασμάτων χωρίς να χάνουμε την επαφή με την πραγματικότητα των μετρήσεων.

4.5 Ανάλυση Χρονοσειράς

Στόχος και γενικές αρχές

Η ανάλυση χρονοσειρών για δεδομένα αισθητήρων αποσκοπεί στο να ξεχωρίσουμε:

- την υποκείμενη τάση από τον βραχυπρόθεσμο θόρυβο,
- τις ημερήσιες/εβδομαδιαίες δομές από τα τυχαία σκαμπανεβάσματα,
- τις σχέσεις μεταξύ παραμέτρων (π.χ. O3 vs NO2).
-

Βασικές αρχές:

- Χρησιμοποιούμε πάντα χρονική ταξινόμηση (ASC) στα timestamps.
- Δουλεύουμε, όπου γίνεται, στην ίδια χρονική ζώνη (συνήθως UTC) και με σταθερό βήμα δειγματοληψίας (ή ρητή χρήση πραγματικού χρόνου).
- Χειριζόμαστε τα κενά συντηρητικά (βλ. κεφάλαιο για παρεμβολή μικρών κενών).
- Κρατάμε διαφάνεια: τι είναι raw, τι είναι clean, ποια σημεία έχουν imputed.

Παρακάτω, τέσσερις βασικές μέθοδοι:

(Α) Χρονοσειρά με Κυλιόμενο Μέσο (rolling mean)

(Β) Ημερήσιοι Μέσοι (daily means)

(Γ) Προφίλ 24ώρου (diurnal profile)

(Δ) Συσχέτιση Παραμέτρων (correlation)

4.5.1 Χρονοσειρά με Κυλιόμενο Μέσο (Rolling Mean ~24h)

Ο κυλιόμενος μέσος είναι ένας απλός, ισχυρός τρόπος να εξομαλύνουμε την ωριαία (ή γενικά βήμα-χρόνου) ακολουθία, ώστε να αναδειχθεί η τάση χωρίς να «καίγονται» οι λεπτομέρειες.

Η ιδέα είναι:

σε κάθε χρονική στιγμή παίρνουμε τον μέσο όρο των γειτονικών τιμών μέσα σε ένα παράθυρο μήκους w.

Τύπος (plain text):

rolling_mean(t) = μέσος όρος τιμών στο διάστημα [t − w/2, t + w/2]

Σε διακριτά δείγματα με σταθερό βήμα, w είναι πλήθος δειγμάτων: π.χ. w = 24 για περίπου 24 ώρες σε ωριαία δεδομένα.

Επιλογή παραθύρου:

- Μικρό w → κρατά περισσότερη λεπτομέρεια (λιγότερη εξομάλυνση).
- Μεγάλο w → πιο ομαλή καμπύλη.

Μικρό παράδειγμα 5 τιμών (ωριαίες)

Ώρες (UTC) και τιμές (π.χ. PM10 σε μg/m3):

t0: 100

t1: 120

t2: 80

t3: 60

t4: 90

Με w = 3 (κυλιόμενο παράθυρο 3 σημείων, κεντραρισμένο):

- rolling στο t1 = (100 + 120 + 80) / 3 = 100.00
- rolling στο t2 = (120 + 80 + 60) / 3 = 86.67
- rolling στο t3 = (80 + 60 + 90) / 3 = 76.67

Στα άκρα (t0, t4) δεν υπολογίζουμε κεντραρισμένο rolling ή χρησιμοποιούμε μικρότερο παράθυρο με προσοχή.

Τι συμπεραίνουμε

- Τα απότομα «καρφιά» της raw χρονοσειράς «μαλακώνουν» και φαίνεται η γενικότερη τάση.
- Μεγάλες αποκλίσεις μεταξύ raw και rolling σημαίνουν έντονη βραχυπρόθεσμη μεταβλητότητα ή spikes.
- Πολύ μεγάλο παράθυρο μπορεί να μην εχει καλά αποτελέσματα

4.5.2 Ημερήσιοι Μέσοι (Daily Means)

Ο ημερήσιος μέσος συμπυκνώνει όλα τα σημεία μιας ημέρας σε ένα νούμερο: είναι ιδανικός για σύγκριση ημερών/εβδομάδων, υπολογισμό τάσεων σε μεγαλύτερες κλίμακες και αποφεύγει τον ωριαίο θόρυβο.

Τύπος (plain text):

daily_mean(ημέρα D) = (άθροισμα όλων των τιμών της D) / (πλήθος έγκυρων τιμών της D)

Μικρό παράδειγμα 5 τιμών σε 2 ημέρες

Ημέρα 1 (D1):

08:00 → 20

12:00 → 40

18:00 → 30

Ημέρα 2 (D2):

08:00 → 25

12:00 → 35

Υπολογισμός:

- daily_mean(D1) = (20 + 40 + 30) / 3 = 30.0
- daily_mean(D2) = (25 + 35) / 2 = 30.0

Το συμπέρασμα εδώ είναι ότι, παρά διαφορετική ωριαία κατανομή, οι δύο ημέρες είχαν ίδιο ημερήσιο μέσο (30).

- Βλέπουμε περιόδους «υψηλών/χαμηλών» ημερών χωρίς να μας αποσπά ο ωριαίος θόρυβος.
- Χρήσιμο για εποχικότητα, εβδομαδιαία μοτίβα, αξιολόγηση πολιτικών/παρεμβάσεων.

4.5.3 Προφίλ 24ώρου (Diurnal Profile)

Το προφίλ 24ώρου είναι ο μέσος όρος ανά ώρα της ημέρας (0…23) συγκεντρώνοντας δεδομένα από πολλές ημέρες.

Δείχνει «Ποια είναι η τυπική εικόνα μέσα στη μέρα;» Π.χ. NOx με πρωινή/βραδινή αιχμή, O3 μεσημβρινή άνοδο, PM10 αργά το βράδυ σε ορισμένα περιβάλλοντα.

Τύπος (plain text):

diurnal_mean(ώρα h) = (άθροισμα τιμών που καταγράφηκαν στην ώρα h σε όλες τις ημέρες) / (πλήθος τέτοιων τιμών)

Μικρό παράδειγμα 5 τιμών σε 2 ημέρες

Ημέρα 1:

08:00 → 20

12:00 → 40

18:00 → 30

Ημέρα 2:

08:00 → 25

12:00 → 35

(στην 18:00 δεν έχουμε τιμή, δεν πειράζει)

Υπολογισμός:

- diurnal_mean(08:00) = (20 + 25) / 2 = 22.5
- diurnal_mean(12:00) = (40 + 35) / 2 = 37.5
- diurnal_mean(18:00) = 30 (μόνο μία διαθέσιμη τιμή)

Βλέπουμε συστηματικές ωριαίες κορυφές/κοιλάδες.

4.5.4 Συσχέτιση Παραμέτρων (Pearson Correlation)

Η συσχέτιση Pearson μετρά πόσο γραμμικά σχετίζονται δύο μεταβλητές: τιμές κοντά σε +1 σημαίνουν ισχυρή θετική σχέση, κοντά σε −1 ισχυρή αρνητική, κοντά στο 0 αδύναμη ή καθόλου γραμμική σχέση.

Τύπος (plain text):

r = cov(X, Y) / (std(X) * std(Y))

όπου

cov(X, Y) = [Σ((x_i − mean(X)) * (y_i − mean(Y)))] / (n − 1)

std = τυπική απόκλιση δείγματος = sqrt([Σ((x_i − mean)^2)] / (n − 1))

Ο υπολογισμός γίνεται μόνο στα κοινά timestamps όπου υπάρχουν και X και Y (drop NaN ζεύγη). Ελάχιστο n: τουλάχιστον 3 ζεύγη για να βγει κάποιο χρήσιμο νούμερο.

Μικρό παράδειγμα 5 ζευγών (NO2 vs O3)

Υποθετικές τιμές (ίδια χρονικά σημεία):

NO2: 60, 55, 50, 45, 40

O3 : 20, 30, 40, 50, 60

Μέσοι:
mean(NO2) = (60 + 55 + 50 + 45 + 40) / 5 = 50

mean(O3) = (20 + 30 + 40 + 50 + 60) / 5 = 40

Αποκλίσεις από τον μέσο:

NO2 dev: +10, +5, 0, −5, −10

O3 dev: −20, −10, 0, +10, +20

Πολλαπλασιασμοί αποκλίσεων και άθροισμα:

(+10)(−20) + (+5)(−10) + (0)(0) + (−5)(+10) + (−10)(+20) = −200 − 50 + 0 − 50 − 200 = −500

cov(NO2, O3) = −500 / (5 − 1) = −125

Τυπικές αποκλίσεις:

std(NO2) = sqrt((100 + 25 + 0 + 25 + 100) / 4) = sqrt(250 / 4) = sqrt(62.5) ≈ 7.9057
std(O3) = sqrt((400 + 100 + 0 + 100 + 400) / 4) = sqrt(1000 / 4) = sqrt(250) ≈ 15.8114

r = −125 / (7.9057 * 15.8114) ≈ −125 / 125 ≈ −1.00

Άρα έχουμε ισχυρή αρνητική γραμμική συσχέτιση (ενδεικτικό για ζεύγη όπως O3 vs NO2 σε αρκετά περιβάλλοντα).

Γενικά:

Η συσχέτιση δεν σημαίνει αιτιότητα.

4.6 Cross-correlation με υστέρηση (lag)

Ο «σταυρο-συσχετισμός» (cross-correlation) δείχνει πόσο δυνατά και προς ποια κατεύθυνση σχετίζεται μια μεταβλητή X με μια άλλη Y όταν μετακινούμε (υστερούμε ή προπορεύουμε) τη μία σε σχέση με την άλλη.

Αν το Y «ακολουθεί» το X με κάποια καθυστέρηση, η συσχέτιση θα κορυφώνει σε θετικό lag (ή αρνητικό, ανάλογα με τον ορισμό).

Ορισμοί (plain text)

- lag L (σε ώρες): συγκρίνουμε X(t) με Y(t + L).

L > 0: το Y είναι «μετά» από το X (X προηγείται).

L < 0: το Y είναι «πριν» από το X (Y προηγείται).

- Για κάθε L υπολογίζουμε Pearson r μόνο στα κοινά χρονικά σημεία.
- Διαλέγουμε ένα εύρος lags, π.χ. από −48 έως +48 ώρες.

Μικρό παράδειγμα (5 ζεύγη, απλό)

Έστω X και Y στο ίδιο βήμα χρόνου:

X: 10, 20, 30, 40, 50

Y: 8, 16, 24, 32, 40

Για lag L = 0: συσχέτιση r ≈ +1 (τέλεια γραμμική σχέση Y ≈ 0.8 * X).

Για lag L = +1 (Y «μετά»): συγκρίνουμε X(1..4) με Y(2..5) → πάλι ισχυρή θετική συσχέτιση.

Για lag L = −1: συγκρίνουμε X(2..5) με Y(1..4) → επίσης ισχυρή θετική.

Το μέγιστο r θα εμφανιστεί σε L όπου οι σειρές «κουμπώνουν» καλύτερα. Σε πραγματικά δεδομένα, το μέγιστο μπορεί να είναι σε +3 ώρες, π.χ. όταν ένας ρύπος ακολουθεί έναν άλλο με αυτήν τη χρονοκαθυστέρηση.

Το lag στο οποίο η συσχέτιση μεγιστοποιείται υποδεικνύει πιθανό «χρόνο απόκρισης» ή «μετατόπιση» μεταξύ δύο μεγεθών (π.χ. NO2 → O3).

Αν ο μέγιστος r είναι αρνητικός, έχουμε αντισυσχέτιση (όταν το ένα ανεβαίνει, το άλλο πέφτει) με συγκεκριμένο lag.

Γενικά συσχέτιση ≠ αιτιότητα. Τα lags βοηθούν στο «storytelling», αλλά δεν αποδεικνύουν αιτιώδεις σχέσεις.

4.7 Αποεποχικοποίηση με STL (Seasonal-Trend decomposition using Loess)

Η STL «σπάει» μια χρονοσειρά σε τρία μέρη:

1. **Trend**: αργές, μακροχρόνιες μεταβολές.
2. **Seasonal**: επαναλαμβανόμενο μοτίβο συγκεκριμένης περιόδου (π.χ. 24 ώρες).
3. **Remainder (Residual)**: ό,τι μένει (θόρυβος, επεισόδια, μη εξηγημένες μεταβολές).

Η ιδέα είναι να διαβάζουμε καθαρά την εποχικότητα/κύκλο (π.χ. ημερήσιο) και την υποκείμενη τάση, χωρίς να μας μπερδεύει ο θόρυβος.

Απαιτήσεις (plain text)

- Σταθερό βήμα χρόνου (ή επαναδειγμάτιση).
- Επιλογή περιόδου: σε ωριαία δεδομένα, period = 24 για ημερήσιο μοτίβο.

Μικρό παράδειγμα (5 σημεία, απλοποιημένο)

Έστω ωριαία τιμή R(t) με «baseline» 20, μια μικρή ημερήσια κυμάτωση και θόρυβο:

t0: 20, t1: 22, t2: 21, t3: 19, t4: 20

Η STL (με period αρκετά μεγαλύτερο από 5, εδώ απλώς για παράδειγμα) θα προσπαθήσει να βρει:

- Seasonal ≈ μικρές επαναλαμβανόμενες διαφορές ανά ώρα ημέρας,
- Trend ≈ ομαλή «γραμμή» που περνά μέσα από τις τιμές,
- Remainder ≈ ό,τι περισσεύει.

Σε πραγματικά δεδομένα με αρκετό μήκος, το **Seasonal** θα δείξει π.χ. μέγιστο O3 μεσημέρι και ελάχιστο νύχτα, το **Trend** π.χ. γενική άνοδο λόγω καλοκαιριού, και το **Remainder** θα κρατήσει τα επεισόδια ή τον θόρυβο.

Γενικά

- Το Seasonal αποκαλύπτει **τυπικό μοτίβο** (ημέρας, εβδομάδας, κ.λπ.).
- Το Trend δείχνει **μακρο-τάσεις** (π.χ. σταδιακή αύξηση PM10 λόγω εποχικότητας).
- Το Remainder είναι το «πεδίο έρευνας» για **ανωμαλίες/επεισόδια**.

4.8 Βασικές Προβλέψεις Επόμενης Τιμής (Forecasting Basics)

Θέλουμε να προβλέψουμε την επόμενη ωριαία τιμή (ή και μερικά βήματα μπροστά) μιας παραμέτρου (π.χ. NO2, O3, PM10) χρησιμοποιώντας μόνο το πρόσφατο ιστορικό της.
Η πρόβλεψη είναι χρήσιμη για: έγκαιρη προειδοποίηση επεισοδίων, βελτίωση λειτουργίας φίλτρων/συναγερμών, και δημιουργία αναμενόμενης «γραμμής βάσης» για έλεγχο ανωμαλιών.

Θα δούμε απλές, στιβαρές μέθοδοι που στήνονται γρήγορα και λειτουργούν ως baseline πριν περάσουμε σε πιο σύνθετα μοντέλα.

Ορίζουμε το πλαίσιο: 1-βήμα vs h-βήμα μπροστά

- 1-βήμα μπροστά (one-step-ahead): προβλέπουμε μόνο το επόμενο δείγμα (π.χ. την επόμενη ώρα).
- h-βήμα μπροστά (multi-step): προβλέπουμε 2, 3, … ώρες μπροστά. Συνήθως είτε:
    - αναδρομικά (recursive): χρησιμοποιούμε την πρόβλεψη του t+1 ως είσοδο για t+2 κ.ο.κ., ή
    - άμεσα (direct): εκπαιδεύουμε χωριστά για κάθε ορίζοντα (π.χ. ξεχωριστή φόρμουλα για +2 ώρες).

Στο παρόν κεφάλαιο εστιάζουμε στο 1-βήμα (εύκολο, πρακτικό baseline).

Μέθοδοι πρόβλεψης (απλές και χρήσιμες)

4.8.1 Persistence / Naïve

Η βασική ιδέα είναι «αυτό που είδα τώρα θα το δω και στο επόμενο βήμα».

- Πρόβλεψη: forecast(t+1) = value(t).
- Πλεονέκτημα: μηδενική πολυπλοκότητα, πολλές φορές εκπληκτικά ισχυρό baseline σε βραχυπρόθεσμους ορίζοντες.
- Μειονέκτημα: αγνοεί τάση και εποχικότητα.

4.8.2 Seasonal Naïve

Σε ωριαία δεδομένα, ο κύκλος 24 ωρών είναι συχνός.

Πρόβλεψη: forecast(t+1) = value(t+1−24).

- Πλεονέκτημα: αξιοποιεί το ημερήσιο μοτίβο.
- Μειονέκτημα: απαιτεί αρκετό ιστορικό και σταθερή εποχικότητα.

4.8.3 Κινητός Μέσος (Moving Average)

Πρόβλεψη: μέσος των τελευταίων w τιμών (π.χ. w=3 ή w=6).

Μαλακώνει θόρυβο, συχνά αποδίδει καλύτερα από Naïve.

Επιλογή w: μικρό w → πιο «ζωντανές» προβλέψεις, μεγάλο w → πιο αργές αλλά σταθερές.

4.8.4 Εκθετική Εξομάλυνση (Simple Exponential Smoothing, SES)

Πρόβλεψη βασισμένη σε εκθετικά φθίνοντα βάρη: οι πιο πρόσφατες τιμές μετρούν περισσότερο.

Ενημέρωση επιπέδου: L_t = α * x_t + (1 − α) * L_{t−1}, 0<α≤1

Πρόβλεψη 1 βήμα μπροστά: forecast(t+1) = L_t

Επιλογή α: 0.2–0.4 συνήθως λειτουργεί καλά σαν αρχή.

4.8.5 Μέθοδος «Drift» (γραμμική τάση από άκρο σε άκρο)

Υποθέτουμε γραμμική τάση από την πρώτη ως την πιο πρόσφατη τιμή.

Κλίση: slope = (x_last − x_first) / (n − 1)

Πρόβλεψη 1 βήμα: forecast = x_last + slope

Μίνι παράδειγμα με 5 ωριαίες τιμές για όλα

Έστω πέντε τιμές PM10:

t0=100, t1=120, t2=80, t3=60, t4=90. Θέλουμε πρόβλεψη για t5.

- **Naïve**: forecast = 90.
- **Seasonal Naïve** (για demo με περίοδο 3): forecast = τιμή στο t2 = 80.
- **Moving Average (w=3)**: forecast = mean(t2, t3, t4) = (80 + 60 + 90) / 3 = 76.67.
- **SES (α=0.3)**:

  L0=100,
  L1=0.3*120* + *0.7*100 = 106,

  L2=0.3*80* + *0.7*106 = 98.2,

  L3=0.3*60* + *0.7*98.2 = 86.74,

  L4=0.3*90* + *0.7*86.74 = 87.718 → forecast = 87.72.

- **Drift**: slope = (90 − 100) / 4 = −2.5 → forecast = 90 − 2.5 = 87.5.

Σύγκριση: Naïve=90, Seasonal(3)=80, MA(3)=76.67, SES≈87.72, Drift=87.5.

Δεν υπάρχει «μία σωστή» πρόβλεψη — η καλύτερη εξαρτάται από τη φύση της σειράς και αξιολογείται με **backtesting**.

4.8.6 Αξιολόγηση: Backtesting & Μετρικές

Για να κρίνουμε ποια μέθοδος «πιάνει» καλύτερα σε δικά μας δεδομένα:

4.8.6.1 Rolling/Expanding backtest (one-step-ahead)

- Χωρίζουμε σε train και test (π.χ. οι τελευταίες 2–4 εβδομάδες ως test).
- Προχωράμε βήμα-βήμα στο test: σε κάθε ώρα προβλέπουμε με ό,τι έχουμε δει ως τότε (χωρίς να «κοιτάμε» το μέλλον).
- Μαζεύουμε τα σφάλματα ε_t = y_t − ŷ_t.

4.8.6.2 Μετρικές

- **MAE** (Mean Absolute Error): μέσο |ε_t|.
- **RMSE** (Root MSE): sqrt(μέσο ε_t^2). Τιμωρεί περισσότερο τα μεγάλα λάθη.

- **MAPE**: μέσο |ε_t / y_t| σε %. Προσοχή κοντά στο 0 (ασταθής).
- **sMAPE**: 2*|ε|/(|y|+|ŷ|) σε %. Πιο σταθερό όταν υπάρχουν μηδενικά.
- **MASE**: MAE του μοντέλου / MAE του Naïve. Τιμές <1 σημαίνουν «κέρδισες» το Naïve.

4.8.6.3 Διαγνωστικός έλεγχος υπολοίπων

Θέλουμε τα υπόλοιπα (residuals) με μέση τιμή ~0 και να μην έχουν αυτοσυσχέτιση.

Αν υπάρχει δομή στα υπόλοιπα, σημαίνει ότι «μένει πληροφορία» να εξηγήσουμε (π.χ. εποχικότητα ή τάση που η μέθοδός μας δεν πιάνει).

4.8.6.4 Πότε επιλέγω τι

- **Naïve**: εξαιρετικό baseline σε βραχυπρόθεσμες προβλέψεις και πολύ θόρυβο.
- **Seasonal Naïve (24h)**: όταν υπάρχει σαφές ημερήσιο μοτίβο.
- **Moving Average**: όταν θες λίγο «σβήσιμο» του θορύβου.
- **SES**: όταν υπάρχει ήπια τάση/επίπεδο που αλλάζει.
- **Drift**: όταν η σειρά έχει καθαρή γραμμική τάση στο πρόσφατο παράθυρο.

Ό,τι κερδίζει στα backtests γίνεται το baseline παραγωγής.

Επέκταση για πολλαπλά βήματα μπροστά (h-step)

- Recursive: για +k ώρες, «ταΐζεις» συνεχώς την επόμενη πρόβλεψη πίσω στο μοντέλο. Απλό, αλλά τα λάθη συσσωρεύονται.
- Direct: εκπαιδεύεις ξεχωριστό μοντέλο για κάθε ορίζοντα (π.χ. +1h, +2h, …). Καλύτερο σε μεγαλύτερους ορίζοντες, αλλά πιο «βαρύ».
- Στα απλά μοντέλα (Naïve/Seasonal/MA/SES) συνήθως χρησιμοποιούμε recursive.

4.9 Ανίχνευση Ανωμαλιών & Αλλαγών Καθεστώτος (Anomaly & Changepoints)

Τι θεωρούμε «ανωμαλία»

- Σημειακή ανωμαλία (point anomaly): μεμονωμένη τιμή που «πετάγεται» μακριά από τις υπόλοιπες (π.χ. 85 όταν γύρω υπάρχουν 20–25).
- Συμφραζόμενη ανωμαλία (contextual): ασυνήθιστη τιμή σε συγκεκριμένο πλαίσιο ώρας/μέρας/εποχής (π.χ. πολύ υψηλό O_3 τα μεσάνυχτα).
- Συλλογική ανωμαλία (collective): «κομμάτι» χρονοσειράς με διαφορετική συμπεριφορά (π.χ. ξαφνική μόνιμη άνοδος επιπέδου).

Πρέπει να τις εντοπίζουμε έγκαιρα, χωρίς να γεμίζουμε ψευδείς συναγερμούς.

QC vs Anomaly Detection

- QC (Quality Control): καθαρίζει προφανή λάθη/artefacts (αρνητικά, spikes κ.λπ.) για να μπορεί να τρέξει σωστά η ανάλυση.
- Anomaly detection: επιχειρησιακό επίπεδο. Αν κάτι είναι «ασυνήθιστο» δεδομένου του προφίλ της σειράς, ενημερώνουμε/σηκώνουμε σήμα.

Στην πράξη, το QC τρέχει πρώτα, το anomaly detection πάνω στη «νοικοκυρεμένη» σειρά.

4.9.1 Μέθοδοι

4.9.1.1 Z-score (ή robust z-score)

Μετράμε πόσα «τυπικά σφάλματα» απέχει μια τιμή από το κέντρο.

- Κλασικό z: (x − mean) / std.
- **Robust z:** (x − median) / (1.4826 * MAD). Ανθεκτικό σε outliers.
- **Rolling z:** υπολογισμός σε μικρό παράθυρο χρόνου για να «πιάνονται» τοπικές ανωμαλίες. Κανόνας: αν |z| > όριο (π.χ. 3.5…5.0) ⇒ ανωμαλία.

4.9.1.2 Hampel (rolling median + MAD)

Το έχουμε ήδη χρησιμοποιήσει: «ραντάρ» spikes σε κυλιόμενο παράθυρο. Αν |x − rolling_median| > n_sigma * 1.4826 * rolling_MAD ⇒ spike.

4.9.1.3 IQR (global)

Βλέπει όλο το δείγμα: outlier αν x < Q1 − k*IQR ή x > Q3 + k*IQR (συχνά k = 3 για αισθητήρες).

4.9.1.4 Αλλαγές καθεστώτος (changepoints)

Ψάχνουμε ξαφνικές μετατοπίσεις σε μέση τιμή ή/και διακύμανση.

- **CUSUM (cumulative sum):** συσσωρεύει αποκλίσεις από «στόχο» (π.χ. ιστορικό μέσο).

  Αν περάσει κατώφλι h ⇒ αλλαγή.

  Παράμετροι:
  k (drift): πόση απόκλιση θεωρούμε «μικρή».

  h (threshold): πόσο άθροισμα χρειάζεται για συναγερμό.

- **Δίδυμα παράθυρα (rolling means/variances):** συγκρίνουμε μέσο/διακύμανση μεταξύ δυο γειτονικών παραθύρων (π.χ. 24h vs επόμενες 24h). Μεγάλη διαφορά ⇒ πιθανό changepoint.
- **Offline segmenting (π.χ. PELT/"ruptures"):** βρίσκει βέλτιστη τμηματοποίηση σε «κομμάτια» με σταθερό μέσο/variance. Πολύ χρήσιμο για αναδρομική ανάλυση.

4.9.1.5 Μικρά παραδείγματα

Point anomaly με robust z

Δεδομένα (μιας ώρας): 20, 22, 21, **85**, 23

Median = 22, αποκλίσεις |x−22|: 2, 0, 1, 63, 1 → MAD = median(0,1,1,2,63) = 1

Robust z στο 85: (85 − 22) / (1.4826*1) ≈ 42.5 ⇒ ανωμαλία.

Changepoint (shift στη μέση)

Τιμές: 30, 31, 29, **45**, 46

Οι τρεις πρώτες έχουν μέσο ≈ 30, οι δύο τελευταίες ≈ 45. Η μετατόπιση στη μέση είναι προφανής. Ένα απλό τεστ «κυλιόμενων μέσων» ή CUSUM θα τη σημαίνει.

Ρύθμιση ορίων

Όρια (thresholds): αυστηρά όρια → λίγοι ψευδώς θετικοί αλλά μπορεί να χάσουμε πραγματικές ανωμαλίες. Χαλαρά όρια → πολλά alerts.

4.10 Δείκτες Ποιότητας Αέρα (AQI) & Υπερβάσεις

Είναι ένας ενοποιημένος δείκτης που μετατρέπει τις συγκεντρώσεις ρύπων σε «υπο-δείκτες» (per-pollutant indices) μέσω breakpoints (διαστήματα τιμών). Για κάθε ρύπο βρίσκουμε το διάστημα [C_low, C_high] στο οποίο πέφτει η συγκέντρωση C και κάνουμε γραμμική παρεμβολή στο αντίστοιχο διάστημα δείκτη [I_low, I_high]. Ο συνολικός AQI στη χρονική στιγμή είναι συνήθως το μέγιστο των υπο-δεικτών (ο «χειρότερος» ρύπος).

Δίνει μια κοινή κλίμακα για διαφορετικούς ρύπους ώστε να χαρακτηρίζεται η ποιότητα αέρα με απλούς όρους (π.χ. “Μέτρια”, “Κακή”), να εντοπίζονται μέρες/ώρες κινδύνου και να γίνεται στοχευμένη ενημέρωση.

Υπερβάσεις

Είναι παραβιάσεις ρυθμιστικών ορίων (π.χ. ημερήσιος μέσος PM10 > όριο, 8-ώρος μέσος O3 > όριο). Συγκρίνουμε με το κατάλληλο threshold και μετράμε πλήθος υπερβάσεων ανά ημέρα/μήνα/έτος.

Κεφάλαιο 5ο: Παρουσίαση και Ανάλυση Αποτελεσμάτων

Φτιάξαμε μια ολοκληρωμένη, μοντέρνα ιστοσελίδα/πλατφόρμα ανάλυσης δεδομένων αισθητήρων που σχεδιάστηκε για να παίρνει «ακατέργαστα» αρχεία μετρήσεων και να τα μετατρέπει σε καθαρή, τεκμηριωμένη γνώση. Η πλατφόρμα υλοποιήθηκε με Python 3 και Flask στο backend, Jinja2 για templating, Pandas/NumPy για επεξεργασία δεδομένων, Plotly για διαδραστικά γραφήματα και Bootstrap 5 για την εμφάνιση. Όπου χρειάζεται εποχική αποσύνθεση, αξιοποιούμε προαιρετικά το statsmodels (STL). Ο κώδικας είναι οργανωμένος σε σαφή, επεκτάσιμα modules, ώστε κάθε επιμέρους μέθοδος να αναβαθμίζεται χωρίς να πειράζουμε τη βασική εφαρμογή.

Στο επίκεντρο έχουμε ένα Flask app που δέχεται αρχεία τύπου CSV/Excel, τα διαβάζει σε Pandas DataFrame και εφαρμόζει μια συνεπή αλυσίδα επεξεργασίας (pipeline): ενοποίηση/κανονικοποίηση χρόνου (UTC), ταξινόμηση ASC σε όλα τα timestamps, προαιρετικό resampling σε ωριαία βάση, και έξυπνη ενοποίηση των στηλών ώστε να σχηματίζονται κοινοί χρονικοί άξονες ανά ρύπο. Στο UI, Jinja2 συνθέτει δυναμικά τις ενότητες της αναφοράς (sections), ενώ τα γραφήματα υλοποιούνται ως Plotly JSON ώστε να είναι ελαφριά και πλήρως διαδραστικά (zoom, pan, tooltips, range slider). Το Bootstrap προσφέρει καθαρή, responsive επιφάνεια εργασίας, με προσεγμένη τυπογραφία και στοίχιση πινάκων.

Η λογική έχει σπάσει σε ανεξάρτητα services για καθαρό διαχωρισμό ευθυνών:

- analysis.py: ο κορμός που «μαζεύει» όλες τις ενότητες και ορίζει τη ροή.
- plotly_helpers.py: δημιουργοί γραφημάτων (timeseries, ημερήσιοι μέσοι, προφίλ 24ώρου, heatmaps, bar plots lags, κ.ά.).
- narrative.py: κείμενα/εξηγήσεις που προηγούνται κάθε ενότητας (μεθοδολογία και συμπεράσματα).
- anomalies.py: ανίχνευση ανωμαλιών και αλλαγών καθεστώτος.
- forecast.py: βασικές προβλέψεις επόμενης τιμής με backtesting.
- aqi.py: υπολογισμός υπο-δεικτών AQI, συνολικού AQI και υπερβάσεων (παραμετρικά). Η δομή αυτή μάς επιτρέπει να προσθέτουμε/αλλάζουμε μεθόδους χωρίς να αγγίζουμε τον σκελετό του app· ενημερώνουμε απλώς το αντίστοιχο service.

5.1 Η πλατφόρμα — από το αρχείο ως την αναφορά

1. **Φόρτωση & προεπεξεργασία**

   Ο χρήστης ανεβάζει το αρχείο του· η πλατφόρμα αναγνωρίζει αυτόματα τις στήλες χρόνου/ρύπων, μετατρέπει τα timestamps σε UTC, τα ταξινομεί αυστηρά σε αύξουσα σειρά, φροντίζει για ομοιόμορφο ωριαίο ρυθμό όπου χρειάζεται και δημιουργεί ενοποιημένο πίνακα με πεδία όπως co, no, pm10, no2, so2, o3, dt. Όλα τα downstream modules στηρίζονται σε αυτήν την συνεπή, «μαζεμένη» βάση.

2. **Έλεγχος ποιότητας (QC) & καθαρισμός**

   Πριν από την ανάλυση, εφαρμόζονται κανόνες robust καθαρισμού:

   - **Αρνητικές/αδύνατες τιμές** → σημαίνονται ως μη έγκυρες.

- **Global IQR (robust)** → εντοπίζονται ακραίες global τιμές με βάση τα τεταρτημόρια και καθορισμένο k*IQR.
- **Hampel (rolling median + MAD)** → εντοπισμός τοπικών spikes μέσα σε συρόμενο παράθυρο.
  Οι τιμές δεν διαγράφονται· σημαίνονται (flags). Για πολύ μικρά κενά εφαρμόζεται συντηρητική παρεμβολή στον χρόνο (γραμμική, με όριο διαδοχικών κενών) αποκλειστικά για να συνεχίσουν οι υπολογισμοί και να διαβάζονται σωστά τα διαγράμματα. Ένα αναλυτικό QC summary εμφανίζει πόσα σημεία επηρεάστηκαν από κάθε κανόνα, μαζί με στατιστικές πριν/μετά.

3. **Ανάλυση χρονοσειρών**

Η πλατφόρμα παράγει πλήρες σετ βασικών αναλύσεων, με κείμενο-εισαγωγή πριν από κάθε γράφημα:

- **Ωριαίες χρονοσειρές (raw) & κυλιόμενος μέσος (~24h)** για ανάδειξη τάσης, με «ενωμένες» γραμμές (connect-gaps) ώστε τα μικρά κενά να μην τεμαχίζουν την εικόνα.
- **Ημερήσιοι μέσοι** με κανόνα πληρότητας (π.χ. ≥18/24 ώρες), για καθαρή σύγκριση ημερών.
- **Προφίλ 24ώρου (diurnal 0–23)** από συγκέντρωση πολλών ημερών, για να φανεί το τυπικό ωριαίο μοτίβο (π.χ. πρωινές αιχμές NOx, μεσημβρινό O_3).
- **Συσχέτιση (Pearson)** με heatmap μόνο στα κοινά timestamps, για να ξεχωρίζουν θετικές/αρνητικές σχέσεις μεταξύ ρύπων.
- **Cross-correlation με υστέρηση (lags)** σε εύρος π.χ. −48..+48 ώρες, για να εντοπίζεται πιθανός χρόνος απόκρισης ενός ρύπου ως προς έναν άλλο (κορύφωση του r στο κατάλληλο lag).
- **Αποεποχικοποίηση με STL (Seasonal–Trend–Remainder)**, period=24 για ωριαία δεδομένα (όπου είναι διαθέσιμο το statsmodels): εμφανίζονται ξεχωριστά τα components Trend/Seasonal/Residual και ενδεικτικές συνεισφορές διακύμανσης.

4. **Ανωμαλίες & αλλαγές καθεστώτος**

Η πλατφόρμα περιλαμβάνει ενότητα «Ανωμαλίες & Αλλαγές Καθεστώτος» με δύο στιβαρά εργαλεία:

- **Robust z (rolling median + MAD)**: σημαίνει spikes όταν |z| υπερβαίνει ρυθμιζόμενο όριο (π.χ. 5.0).
- **Δίδυμα παράθυρα (twin-window)**: συγκρίνει μέσος/διακύμανση μεταξύ δύο διαδοχικών παραθύρων (π.χ. 24h/24h) και δίνει score αλλαγής καθεστώτος πάνω από κατώφλι (π.χ. 1.5×std).
  Τα γεγονότα εμφανίζονται πάνω στη γραμμή (markers για spikes, διακεκομμένες κάθετες για μετατοπίσεις), και συνοψίζονται σε πίνακα alerts με timestamp, τύπο και score.

5. **Προβλέψεις επόμενης ώρας (Forecasts)**

   Για κάθε διαθέσιμη παράμετρο εκτελείται rolling one-step backtest (χωρίς να «κοιτάμε» το μέλλον) πάνω σε απλές αλλά αξιόπιστες μεθόδους:

   - **Naïve** (επόμενο=τελευταίο),
   - **Seasonal Naïve 24h**,
   - **Moving Average (w)**,
   - **Simple Exponential Smoothing (SES, α)**,
   - **Drift** (γραμμική τάση).

     Υπολογίζονται **MAE, RMSE, sMAPE, MASE** και επιλέγεται αυτόματα η «καλύτερη» μέθοδος (μικρότερο MASE) για την προβολή. Στο γράφημα φαίνονται οι τελευταίες 72 ώρες (actual), η 1-βήμα πρόβλεψη της καλύτερης μεθόδου και ένας marker με την **πρόβλεψη της επόμενης ώρας**. Πίνακας μετρικών παρουσιάζει και τις τιμές πρόβλεψης +1h για όλες τις μεθόδους.

6. **AQI & Υπερβάσεις (παραμετρικά)**

   Η πλατφόρμα υπολογίζει υπο-δείκτες AQI ανά ρύπο με breakpoints (γραμμική παρεμβολή ανά διάστημα) και εμφανίζει τον συνολικό AQI ως μέγιστο υπο-δείκτη ανά χρονική στιγμή, μαζί με τον top contributor (ποιος ρύπος «κυριαρχεί»). Παράλληλα, υποστηρίζει υπερβάσεις σε παράγωγα όπως ημερήσιος μέσος PM10, 8-ώρος O_3, ωριαίο NO_2, με κανόνες πληρότητας. Τα breakpoints/όρια είναι παραμετρικά (π.χ. configuration ανά πλαίσιο: EU/WHO/US), ώστε να μπορούν να «κουμπώσουν» τα επίσημα κανονιστικά σχήματα χωρίς αλλαγή κώδικα. Για επίδειξη χρησιμοποιούνται ασφαλείς προεπιλογές (DEMO), με σαφή επισήμανση ότι πρέπει να αντικατασταθούν από τα επίσημα.

Κάθε ενότητα ξεκινά με σύντομη, διδακτική περιγραφή της μεθόδου, ακολουθούν γραφήματα και πίνακες σε προσεγμένη στοίχιση (numeric alignment, border/hover classes). Τα γραφήματα έχουν range slider, ενιαίο tooltip («hovermode: x unified»), ενωμένες γραμμές ακόμη και όταν υπάρχουν μικρά κενά (connect-gaps/ήπια οπτική παρεμβολή έως 2 ώρες), και σταθερή χρονολογική σειρά (ASC) σε όλες τις προβολές.

Η πλατφόρμα δεν αποθηκεύει μόνιμα τα δεδομένα του χρήστη· τα επεξεργάζεται στη μνήμη και τα εμφανίζει στην ίδια συνεδρία, με στόχο την ιδιωτικότητα και τη συμμόρφωση σε workflows όπου τα δεδομένα δεν πρέπει να «μένουν» στον server. Η υλοποίηση είναι self-contained και μπορεί να τρέξει τοπικά ή σε εσωτερικό περιβάλλον χωρίς πρόσβαση στο διαδίκτυο.

Με ένα μόνο ανέβασμα αρχείου, ο χρήστης παίρνει:

- αυτόματο **καθαρισμό** και **τεκμηριωμένο QC**,
- **πλήρη εικόνα χρονοσειράς** με τάση/ημερήσιους/ωριαία μοτίβα,
- **σχέσεις** μεταξύ ρύπων (συσχέτιση/lag),

- **ανωμαλίες** και **μετατοπίσεις καθεστώτος** με οπτικά και πίνακες,
- **προβλέψεις επόμενης ώρας** με ποσοτική αξιολόγηση λαθών,
- **AQI** και **υπερβάσεις** βάσει παραμετρικών ορίων,
- και όλα αυτά με **εξηγητικό κείμενο** πριν από κάθε ενότητα, ώστε η αναφορά να είναι χρήσιμη και σε μη ειδικούς.

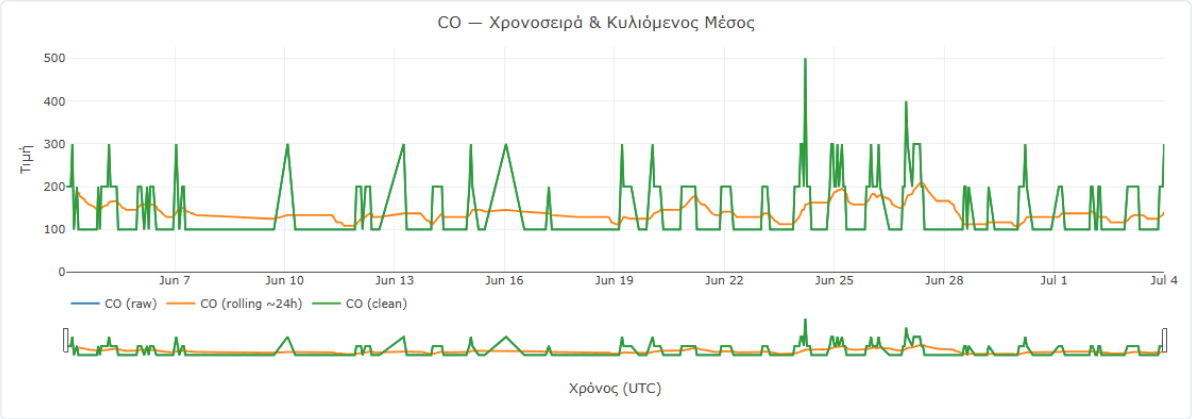
5.2 QC – Ποιοτικός Έλεγχος & Καθαρισμός

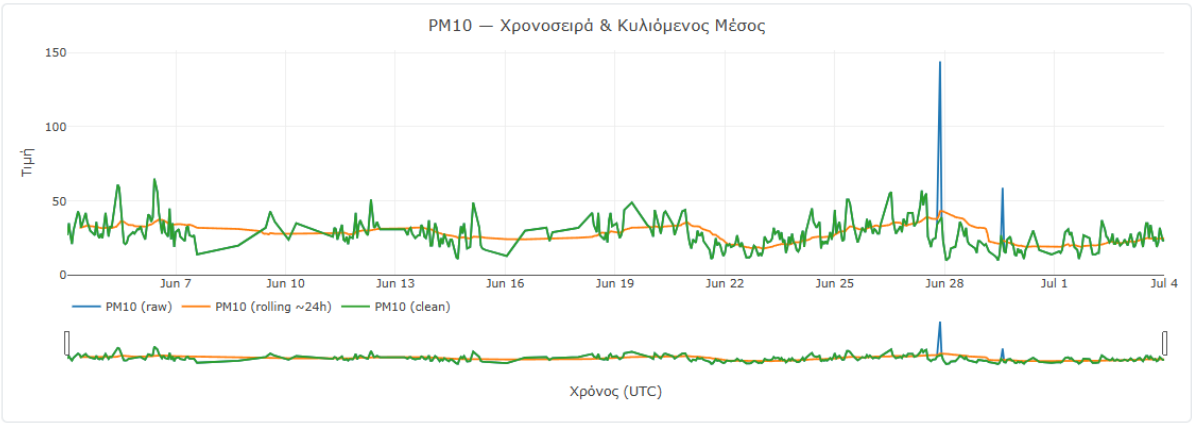
Τι κάνουμε: Εντοπίζουμε αρνητικές τιμές, global outliers με IQR (k=3) και spikes με Hampel (rolling median + MAD). Σημειώνουμε (flags) τα ύποπτα σημεία, τα αντικαθιστούμε προσωρινά με κενά (NaN), και εφαρμόζουμε χρονο-παρεμβολή για μικρά κενά (≤2).

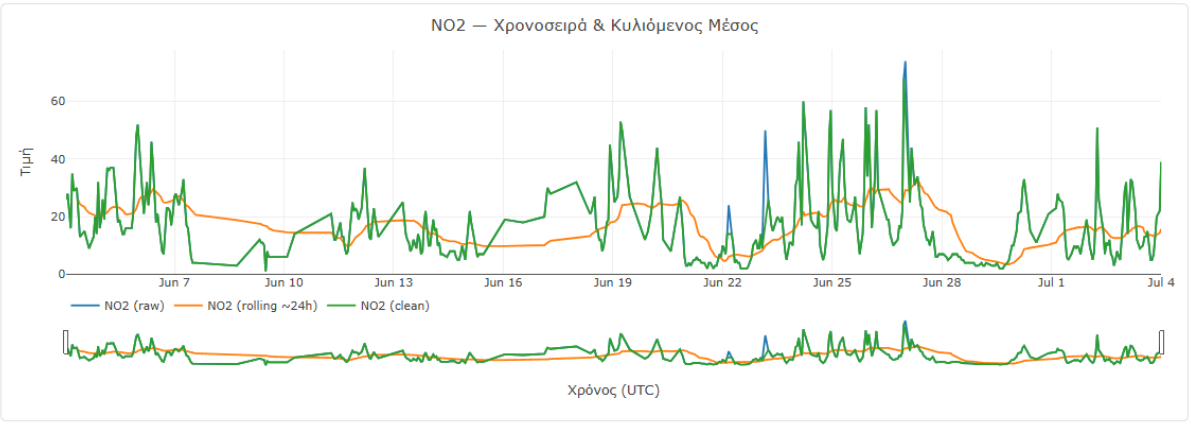
Τι κερδίζουμε: καθαρότερη σειρά για ανάλυση χωρίς να χάνονται τα πρωτογενή δεδομένα.

parameter	count_raw	missing_raw	n_negative	n_iqr	n_hampel	n_anyflag	n_imputed	mean_raw	mean_clean
co	421	0	0	0	0	0	0	140.143	140.143
no2	415	6	0	1	3	3	9	16.882	16.849
o3	415	6	0	0	20	20	22	59.518	59.909
pm10	410	11	0	2	2	3	13	28.168	27.875
so2	415	6	0	2	3	3	7	3.398	3.033

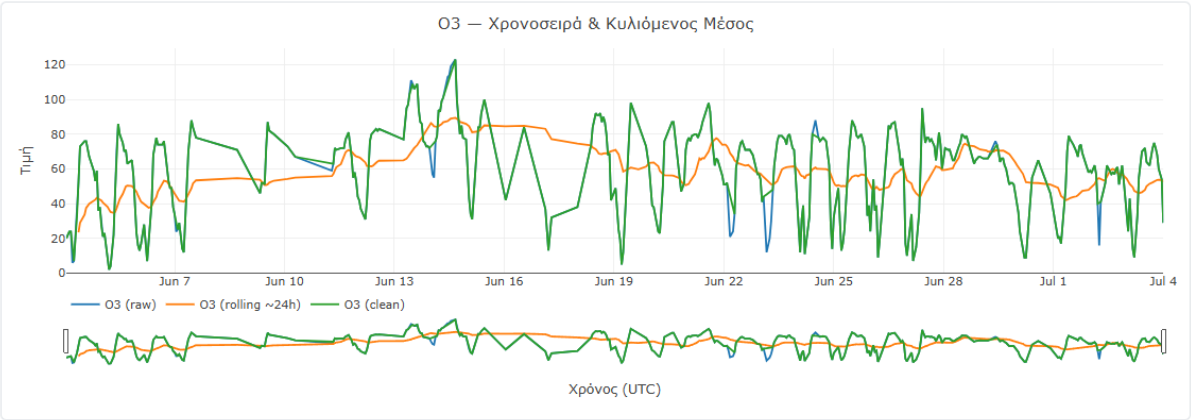
Εικόνα 5.1: Σύνοψη QC ανά παράμετρο




PM10 — Χρονοσειρά & Κυλιόμενος Μέσος
150
100
50
0
Τιμή
Jun 7
Jun 10
Jun 13
Jun 16
Jun 19
Jun 22
Jun 25
Jun 28
Jul 1
Jul 4
PM10 (raw)
PM10 (rolling ~24h)
PM10 (clean)
Χρόνος (UTC)


NO2 — Χρονοσειρά & Κυλιόμενος Μέσος
60
40
20
0
Τιμή
Jun 7
Jun 10
Jun 13
Jun 16
Jun 19
Jun 22
Jun 25
Jun 28
Jul 1
Jul 4
NO2 (raw)
NO2 (rolling ~24h)
NO2 (clean)
Χρόνος (UTC)


SO2 — Χρονοσειρά & Κυλιόμενος Μέσος
100
50
0
Τιμή
Jun 7
Jun 10
Jun 13
Jun 16
Jun 19
Jun 22
Jun 25
Jun 28
Jul 1
Jul 4
SO2 (raw)
SO2 (rolling ~24h)
SO2 (clean)
Χρόνος (UTC)



Εικόνα 5.2: Χρονοσειρές και Κυλιόμενοι Μέσοι για όλους τους αισθητήρες

Μέθοδος: Εντοπίζουμε αρνητικές τιμές, global outliers με IQR (k=3) και στιγμιαία spikes με Hampel (rolling median + MAD). Τα ύποπτα σημεία σημειώνονται (flags) και αντικαθίστανται προσωρινά με κενές τιμές, οι οποίες συμπληρώνονται με χρονο-παρεμβολή για μικρά κενά (≤2 συνεχόμενα δείγματα). Στόχος είναι καθαρότερη σειρά χωρίς να χάνονται τα πρωτογενή δεδομένα.

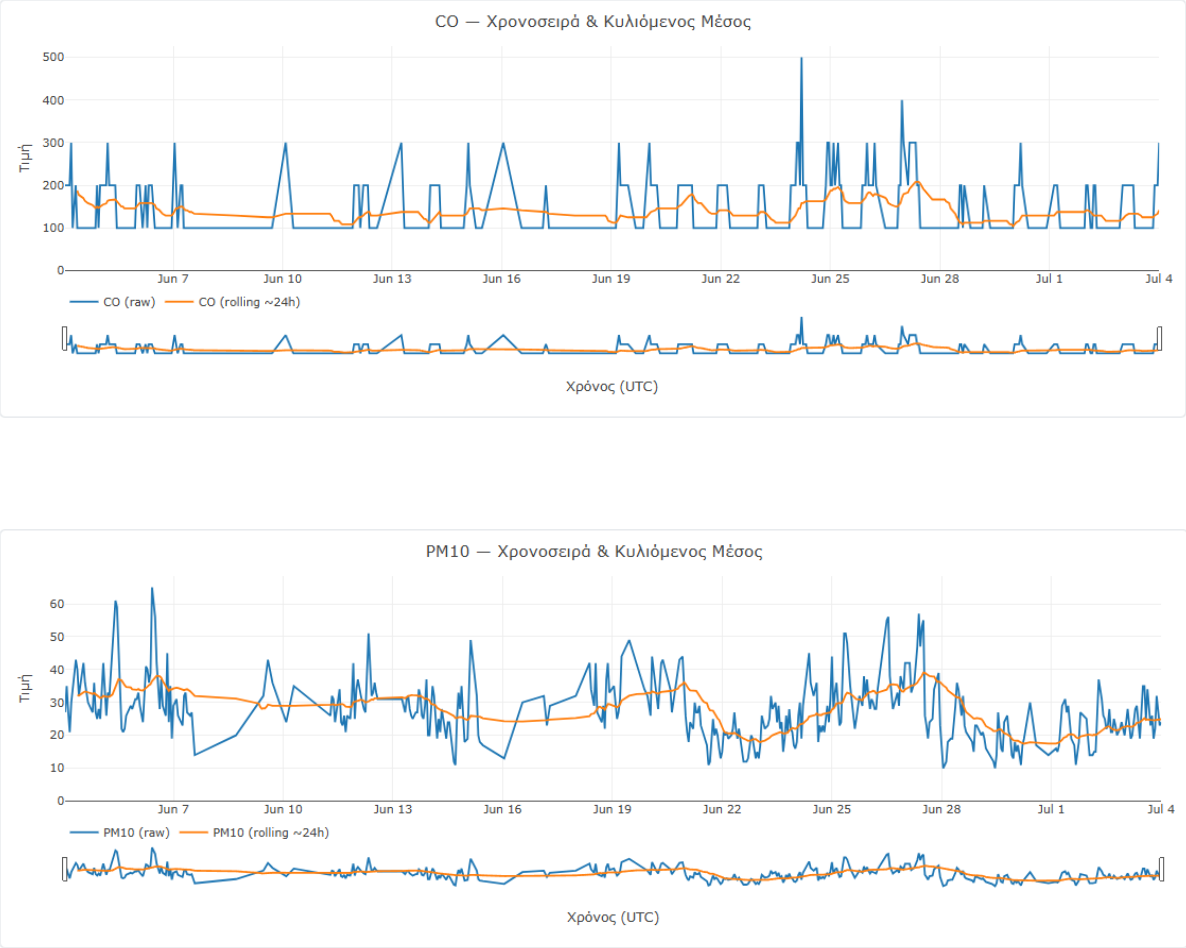
Τι συμπεραίνουμε: Το QC απομακρύνει ακραίες/λανθασμένες τιμές, μειώνει τον θόρυβο και επιτρέπει ασφαλέστερη στατιστική και οπτική ανάλυση. Οι σημαίες (flags) διατηρούνται ώστε να γνωρίζουμε πού έγιναν διορθώσεις.

Ανά παράμετρο:

- **CO (μονοξείδιο του άνθρακα):** δείγματα=421, flags=0 (0.0%), imputed=0 (0.0%), μέση (raw)=140.14, μέση (clean)=140.14.
- **NO_2 (διοξείδιο του αζώτου):** δείγματα=415, flags=3 (0.7%), imputed=9 (2.2%), μέση (raw)=16.88, μέση (clean)=16.85.
- **O_3 (όζον):** δείγματα=415, flags=20 (4.8%), imputed=22 (5.3%), μέση (raw)=59.52, μέση (clean)=59.91.
- **PM_{10} (αιωρούμενα σωματίδια):** δείγματα=410, flags=3 (0.7%), imputed=13 (3.2%), μέση (raw)=28.17, μέση (clean)=27.87.
- **SO_2 (διοξείδιο του θείου):** δείγματα=415, flags=3 (0.7%), imputed=7 (1.7%), μέση (raw)=3.40, μέση (clean)=3.03.

5.3 Χρονοσειρές

Εμφάνιση ωριαίων τιμών και εξομαλυμένης καμπύλης (κυλιόμενος μέσος ~24h) για ανάδειξη υποκείμενης τάσης και εντοπισμό επεισοδίων.



Εικόνα 5.3: Ανάδειξη υποκείμενης τάσης και εντοπισμό επεισοδίων.

Μέθοδος: Προβάλλουμε την ωριαία εξέλιξη των τιμών και μια εξομαλυμένη καμπύλη (κυλιόμενος μέσος ~24h) ώστε να διαχωρίζουμε την υποκείμενη τάση από στιγμιαίο θόρυβο.

Τι συμπεραίνουμε: Τα απότομα «καρφιά» υποδεικνύουν επεισόδια, ενώ η εξομαλυμένη καμπύλη δείχνει γενικές αυξομειώσεις και μεταβατικές περιόδους.

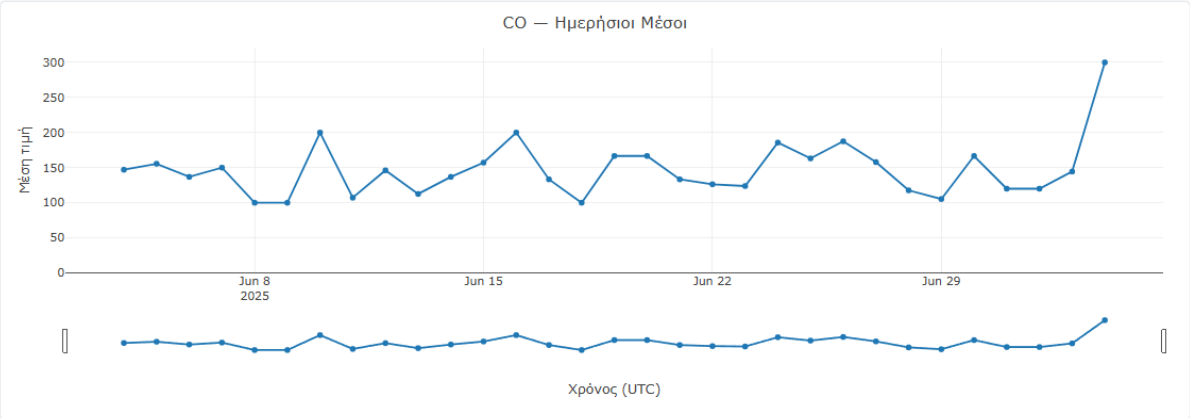
Χρονική κάλυψη δεδομένων: 2025-06-04 01:00:00+00:00 έως 2025-07-04 00:00:00+00:00 (διάρκεια: 29 days 23:00:00).

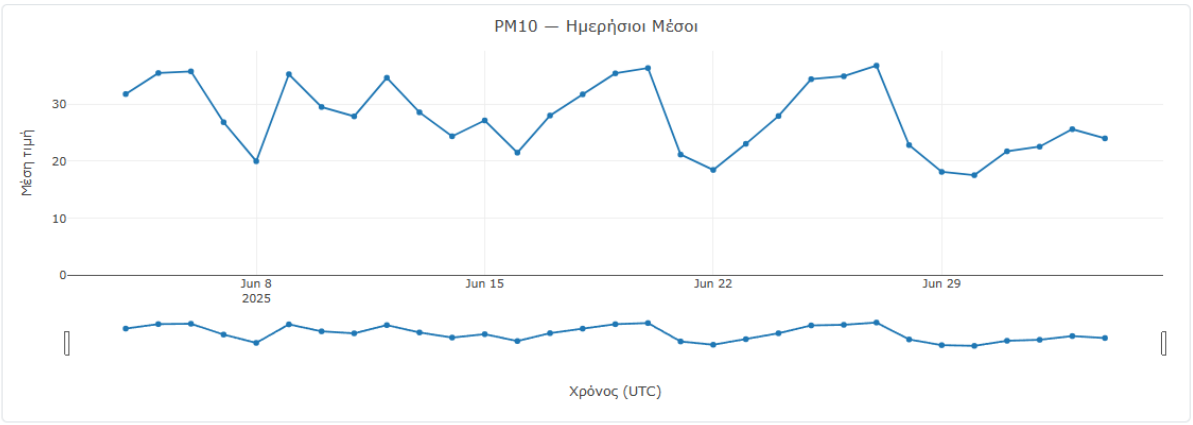
Ανά παράμετρο:

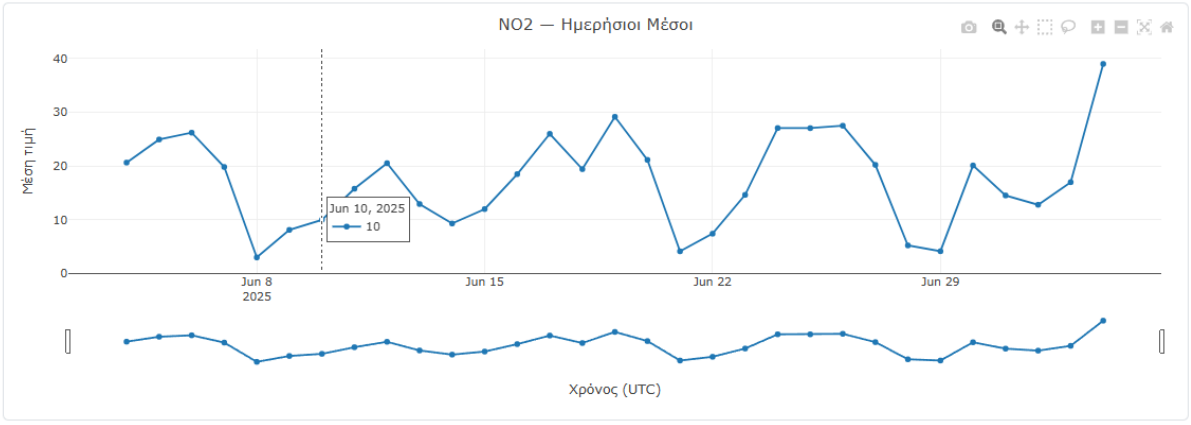
- **CO (μονοξείδιο του άνθρακα):** διάμεσος=100.00, P95=300.00, μέγιστο=500.00 (στιγμή: 2025-06-24 05:00:00+00:00); ενδεικτική τάση σε ημερήσια κλίμακα: αύξηση.
- **PM_{10} (αιωρούμενα σωματίδια):** διάμεσος=27.00, P95=45.00, μέγιστο=65.00 (στιγμή: 2025-06-06 10:00:00+00:00); ενδεικτική τάση σε ημερήσια κλίμακα: μείωση.
- **NO_2 (διοξείδιο του αζώτου):** διάμεσος=14.00, P95=42.00, μέγιστο=68.00 (στιγμή: 2025-06-26 23:00:00+00:00); ενδεικτική τάση σε ημερήσια κλίμακα: αύξηση.
- **SO_2 (διοξείδιο του θείου):** διάμεσος=3.00, P95=5.00, μέγιστο=6.00 (στιγμή: 2025-06-13 08:00:00+00:00); ενδεικτική τάση σε ημερήσια κλίμακα: αύξηση.
- **O_3 (όζον):** διάμεσος=65.00, P95=92.20, μέγιστο=123.00 (στιγμή: 2025-06-14 16:00:00+00:00); ενδεικτική τάση σε ημερήσια κλίμακα: μείωση.

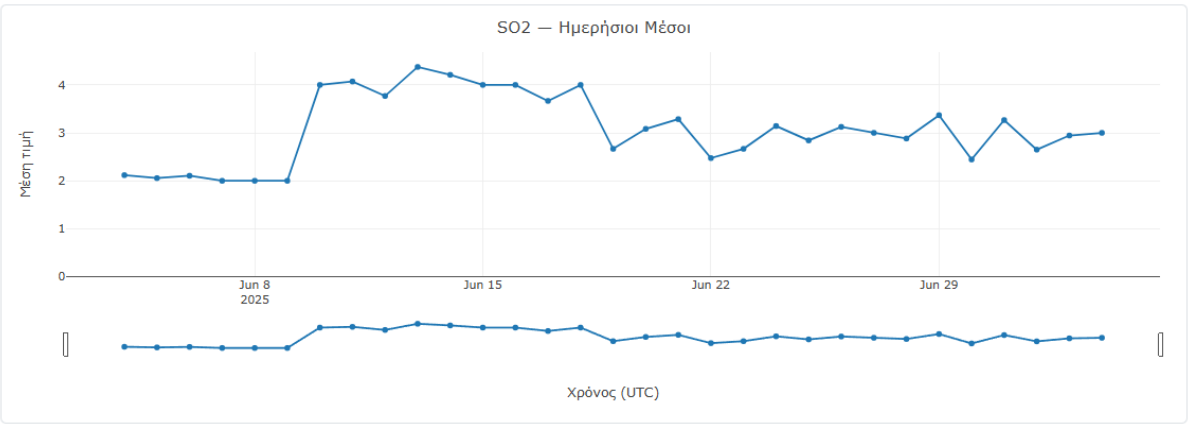
5.4 Ημερήσιοι Μέσοι

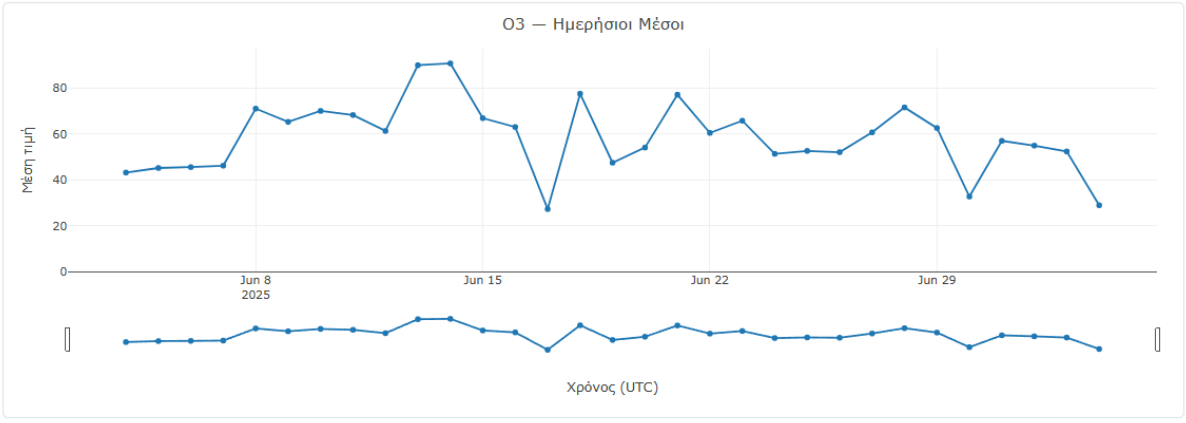
Συμπύκνωση ωριαίων σε μέση τιμή ανά ημέρα για καθαρή σύγκριση ημερών/εβδομάδων και ανάδειξη περιόδων με αυξημένες ή μειωμένες συγκεντρώσεις.




PM10 — Ημερήσιοι Μέσοι
Μέση τιμή
30
20
10
0
Jun 8
2025
Jun 15
Jun 22
Jun 29
Χρόνος (UTC)


NO2 — Ημερήσιοι Μέσοι
Μέση τιμή
40
30
20
10
0
Jun 10, 2025
10
Jun 8
2025
Jun 15
Jun 22
Jun 29
Χρόνος (UTC)


SO2 — Ημερήσιοι Μέσοι
Μέση τιμή
4
3
2
1
0
Jun 8
2025
Jun 15
Jun 22
Jun 29
Χρόνος (UTC)


O3 — Ημερήσιοι Μέσοι
Μέση τιμή
80
60
40
20
0
Jun 8
2025
Jun 15
Jun 22
Jun 29
Χρόνος (UTC)

Εικόνα 5.4: Ημερήσιοι Μέσοι για όλους τους αισθητήρες

Μέθοδος: Υπολογίζουμε μέση τιμή ανά ημέρα (ημερήσιοι μέσοι) για να αναδειχθούν καθαρά οι διαφορές μεταξύ ημερών χωρίς τον ωριαίο θόρυβο.

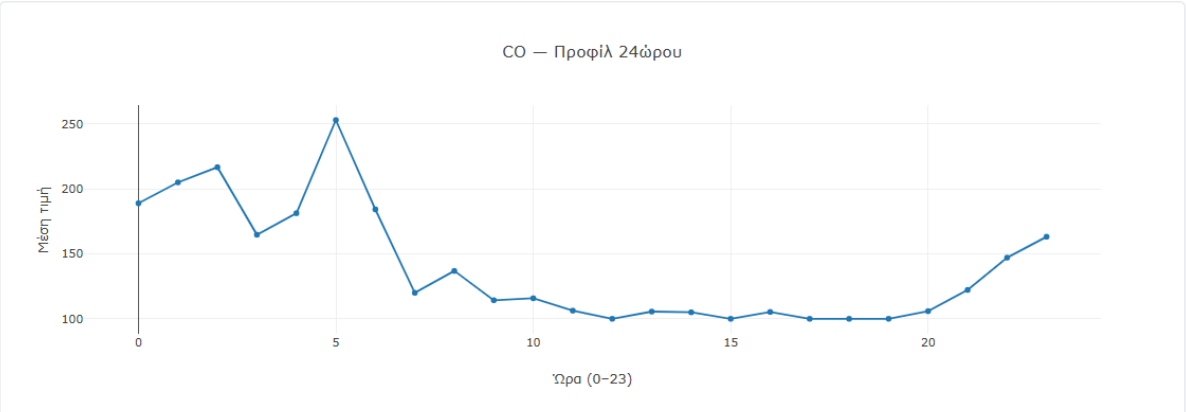
Τι συμπεραίνουμε: Η ενότητα βοηθά στον εντοπισμό περιόδων με αυξημένες ή μειωμένες συγκεντρώσεις και στη σύγκριση ημερών/εβδομάδων.

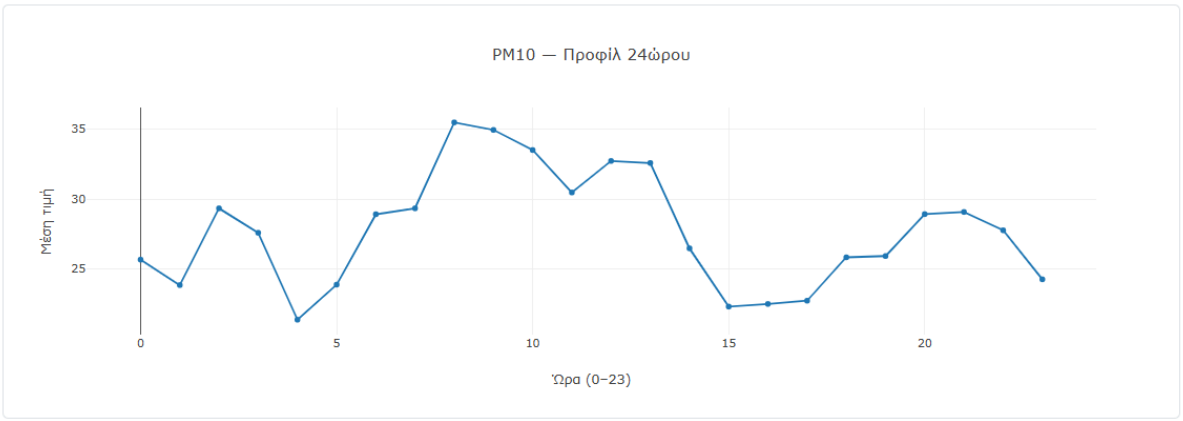
Ανά παράμετρο:

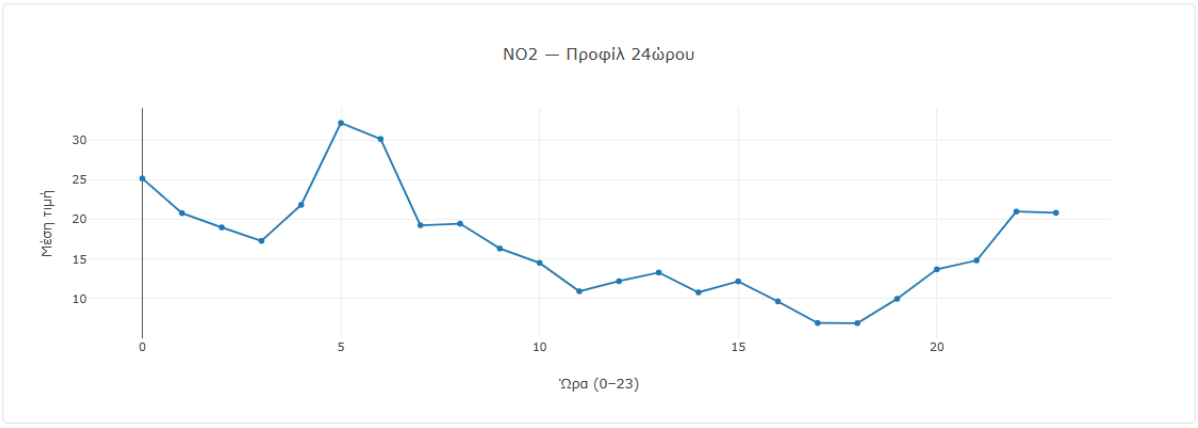
- **CO (μονοξείδιο του άνθρακα):** υψηλότερος ημερήσιος μέσος την 2025-07-04 (τιμή: 300.00); χαμηλότερος την 2025-06-08 (τιμή: 100.00).
- **PM_{10} (αιωρούμενα σωματίδια):** υψηλότερος ημερήσιος μέσος την 2025-06-27 (τιμή: 36.74); χαμηλότερος την 2025-06-30 (τιμή: 17.56).
- **NO_2 (διοξείδιο του αζώτου):** υψηλότερος ημερήσιος μέσος την 2025-07-04 (τιμή: 39.00); χαμηλότερος την 2025-06-08 (τιμή: 3.00).
- **SO_2 (διοξείδιο του θείου):** υψηλότερος ημερήσιος μέσος την 2025-06-13 (τιμή: 4.38); χαμηλότερος την 2025-06-07 (τιμή: 2.00).
- **O_3 (όζον):** υψηλότερος ημερήσιος μέσος την 2025-06-14 (τιμή: 90.66); χαμηλότερος την 2025-06-17 (τιμή: 27.33).

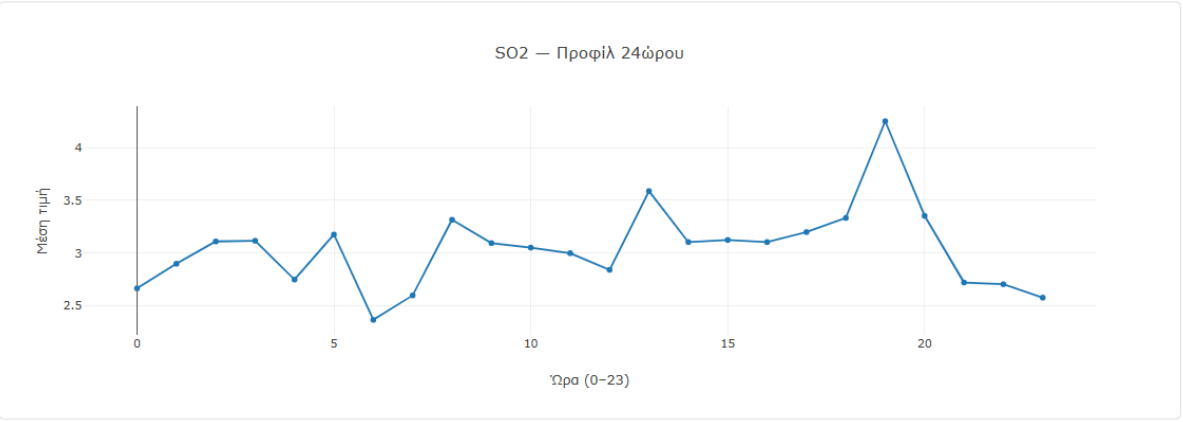
5.5 Προφίλ 24ώρου

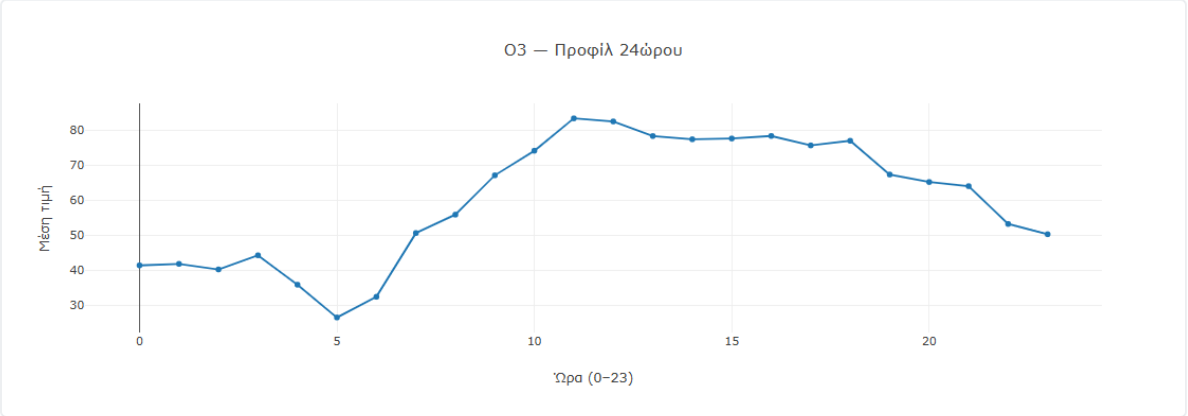
Μέσος όρος ανά ώρα (0–23) για να φανεί το τυπικό ωριαίο μοτίβο (π.χ. πρωινές/βραδινές αιχμές NOx, μεσημβρινό O_3).




PM10 — Προφίλ 24ώρου
Μέση τιμή
35
30
25
0
5
10
15
20
Ώρα (0–23)


NO2 — Προφίλ 24ώρου
Μέση τιμή
30
25
20
15
10
0
5
10
15
20
Ώρα (0–23)


SO2 — Προφίλ 24ώρου
Μέση τιμή
4
3.5
3
2.5
0
5
10
15
20
Ώρα (0–23)


O3 — Προφίλ 24ώρου
Μέση τιμή
80
70
60
50
40
30
0
5
10
15
20
Ώρα (0–23)

Εικόνα 5.5: Προφίλ 24ώρου για όλους τους αισθητήρες

Μέθοδος: Υπολογίζουμε τον μέσο όρο ανά ώρα της ημέρας (0–23) σε όλο το σύνολο για να φανεί το τυπικό ημερήσιο μοτίβο.

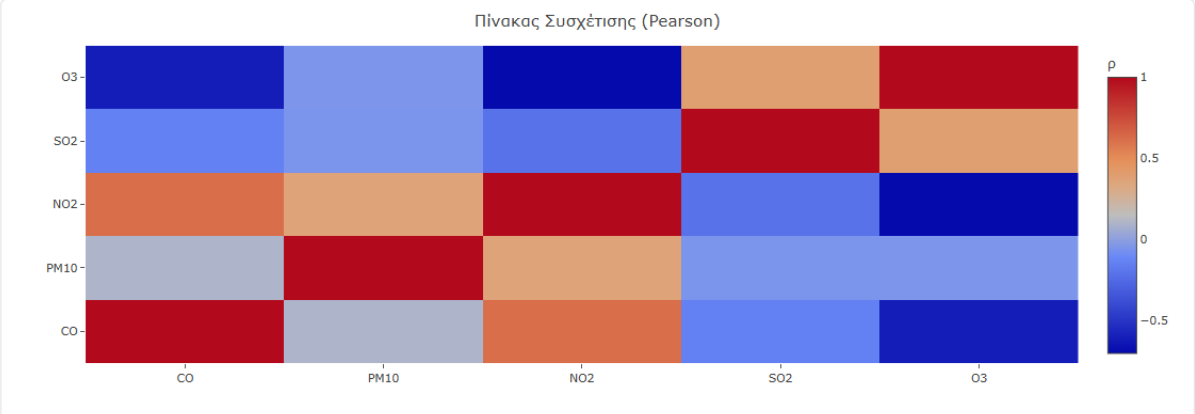
Τι συμπεραίνουμε: Παρατηρούμε ώρες με συστηματικά υψηλές/χαμηλές τιμές που συνδέονται με πηγές (π.χ. κυκλοφορία) ή φωτοχημεία (π.χ. O_3 μεσημέρι).

Ανά παράμετρο:

- **CO (μονοξείδιο του άνθρακα):** μέγιστο γύρω στις 5:00, ελάχιστο γύρω στις 12:00, εύρος (amplitude) ≈ 152.94.
- **PM_{10} (αιωρούμενα σωματίδια):** μέγιστο γύρω στις 8:00, ελάχιστο γύρω στις 4:00, εύρος (amplitude) ≈ 14.10.
- **NO_2 (διοξείδιο του αζώτου):** μέγιστο γύρω στις 5:00, ελάχιστο γύρω στις 18:00, εύρος (amplitude) ≈ 25.26.
- **SO_2 (διοξείδιο του θείου):** μέγιστο γύρω στις 19:00, ελάχιστο γύρω στις 6:00, εύρος (amplitude) ≈ 1.88.
- **O_3 (όζον):** μέγιστο γύρω στις 11:00, ελάχιστο γύρω στις 5:00, εύρος (amplitude) ≈ 56.68.

5.6 Συσχέτιση Παραμέτρων

Συσχέτιση: Συντελεστής Pearson μεταξύ παραμέτρων στα κοινά timestamps. Υψηλές θετικές τιμές υποδεικνύουν πιθανές κοινές πηγές/συνθήκες, αρνητικές αντιστροφή.



Εικόνα 5.6: Πίνακας Συσχέτισης

Μέθοδος: Υπολογίζουμε συσχετίσεις Pearson μεταξύ παραμέτρων (−1 έως +1).

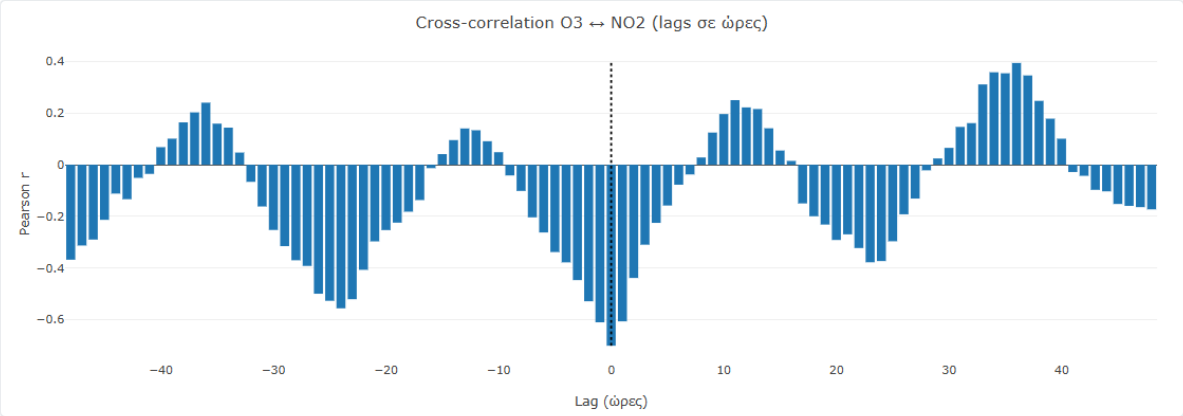
Τι συμπεραίνουμε: Υψηλές θετικές τιμές υποδεικνύουν κοινές πηγές ή συνθήκες, ενώ αρνητικές πιθανή αντιστροφή (π.χ. O_3 vs NO_2). Η συσχέτιση δεν συνεπάγεται αιτιότητα.

Ισχυρά ζεύγη:

- Θετική: CO (μονοξείδιο του άνθρακα) ↔ NO_2 (διοξείδιο του αζώτου) (ρ=0.619)
- Αρνητική: NO_2 (διοξείδιο του αζώτου) ↔ O_3 (όζον) (ρ=-0.700)
- Αρνητική: CO (μονοξείδιο του άνθρακα) ↔ O_3 (όζον) (ρ=-0.610)

5.7 Cross-correlation με υστέρηση (lag)

Μέθοδος: Pearson r για σειρά lags μεταξύ δύο μεταβλητών.



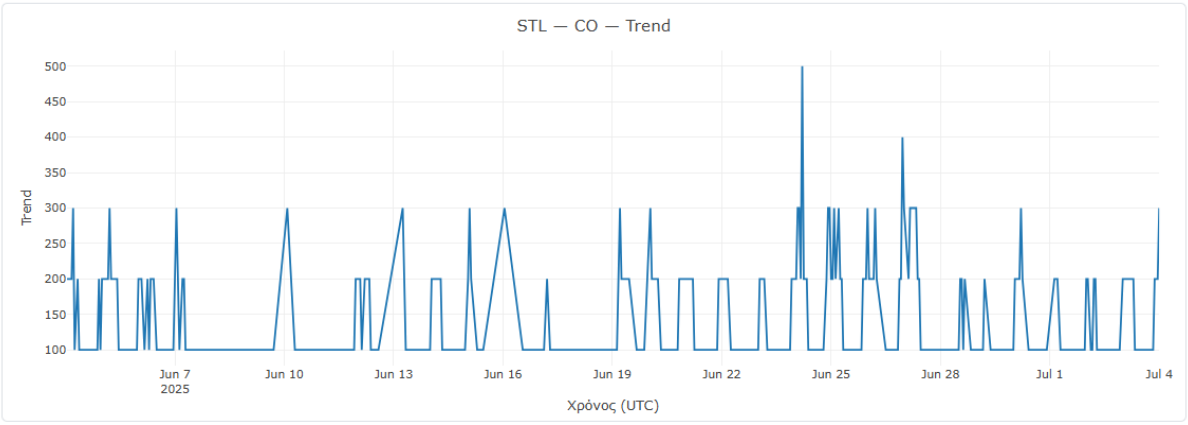
Εικόνα 5.7: Cross-correlation O3 με NO2

Μέθοδος: Υπολογίζουμε Pearson r για σειρά lags (π.χ. −48..+48 ώρες) μεταξύ του ζεύγους O3 ↔ NO2. Θετικό lag σημαίνει ότι η δεύτερη μεταβλητή ακολουθεί την πρώτη.

Τι συμπεραίνουμε: Το lag στο οποίο το |r| μεγιστοποιείται υποδεικνύει πιθανό χρόνο απόκρισης. Στα δικά σας δεδομένα, μέγιστο |r| στο lag 0 ώρες με r=-0.700.

5.8 Αποεποχικοποίηση (STL)

Μέθοδος: Διάσπαση της σειράς σε Trend, Seasonal και Remainder (period=24).



Εικόνα 5.8: Trend σειρά για το CO

Μέθοδος: Η STL διασπά τη σειρά σε Trend, Seasonal και Remainder με περίοδο 24 (π.χ. ημερήσιο μοτίβο για ωριαία δεδομένα).

Τι συμπεραίνουμε: Το Seasonal δείχνει το τυπικό μοτίβο, το Trend την υποκείμενη τάση, ενώ το Remainder κρατά θόρυβο/επεισόδια.

5.9 Προβλέψεις (1-βήμα μπροστά)

Μέθοδος: Απλές αλλά στιβαρές προβλέψεις επόμενης ώρας (Naïve, Seasonal 24h, Moving Average, Simple Exponential Smoothing, Drift). Χρησιμοποιείται rolling one-step backtest με μετρικές MAE, RMSE, sMAPE και MASE. Η επόμενη ώρα (+1h) προβάλλεται ως marker.

Μετρικές & Επόμενη Πρόβλεψη — CO

method	MAE	RMSE	sMAPE%	MASE	Next(+1h)
naive	22.105	51.978	11.42	1.000	300.000
seasonal24	29.362	60.142	15.90	1.328	200.000
ma	35.339	58.001	21.44	1.599	180.000
ses	38.613	57.132	24.71	1.747	196.555
drift	26.822	57.838	14.87	1.213	300.139

Μετρικές & Επόμενη Πρόβλεψη — PM_{10}

method	MAE	RMSE	sMAPE%	MASE	Next(+1h)
naive	5.220	6.956	19.19	1.000	24.000
seasonal24	7.842	10.389	28.75	1.502	24.000
ma	6.279	8.175	23.25	1.203	24.000
ses	5.827	7.478	21.56	1.116	24.944
drift	5.711	7.492	21.10	1.094	23.996

Μετρικές & Επόμενη Πρόβλεψη — NO_2

method	MAE	RMSE	sMAPE%	MASE	Next(+1h)
naive	4.845	8.185	27.85	1.000	39.000
seasonal24	9.017	12.971	57.80	1.861	32.000
ma	7.065	10.425	43.24	1.458	20.200
ses	6.537	9.385	40.05	1.349	22.707
drift	5.474	8.632	32.57	1.130	39.018

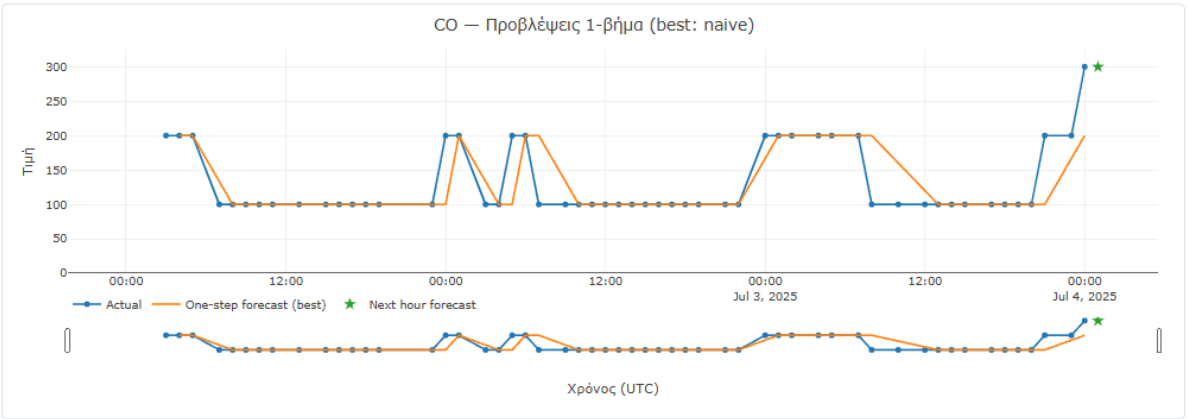
Μετρικές & Επόμενη Πρόβλεψη — SO_2

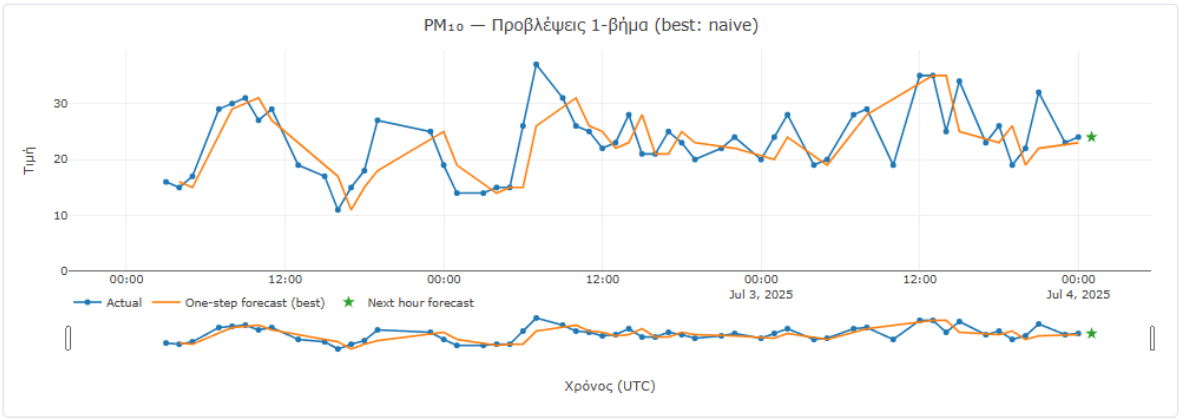
method	MAE	RMSE	sMAPE%	MASE	Next(+1h)
naive	0.541	0.847	16.77	1.000	3.000
seasonal24	0.643	0.987	20.50	1.189	3.000
ma	0.524	0.714	16.93	0.969	2.800
ses	0.536	0.733	17.28	0.991	2.732
drift	0.560	0.864	17.39	1.036	3.001

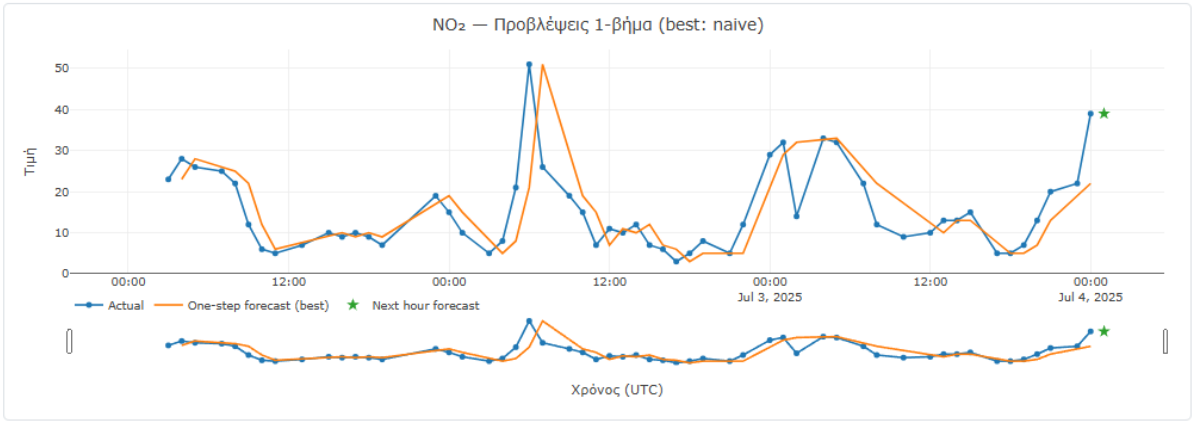
Μετρικές & Επόμενη Πρόβλεψη — O_3

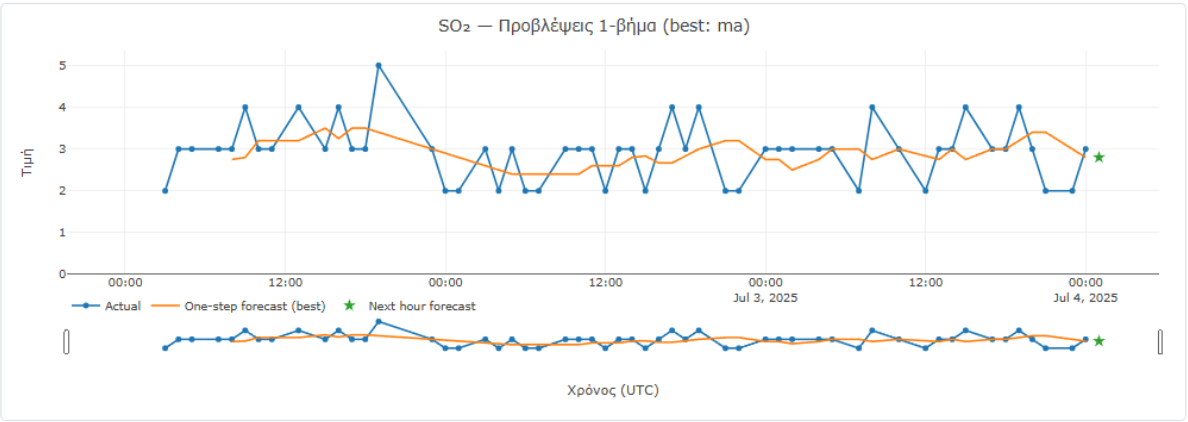
method	MAE	RMSE	sMAPE%	MASE	Next(+1h)
naive	6.469	9.859	15.57	1.000	29.000
seasonal24	13.810	18.738	29.88	2.135	24.000
ma	13.426	18.077	27.95	2.075	56.600
ses	13.858	18.167	27.96	2.142	52.291
drift	8.809	13.928	20.37	1.362	29.013

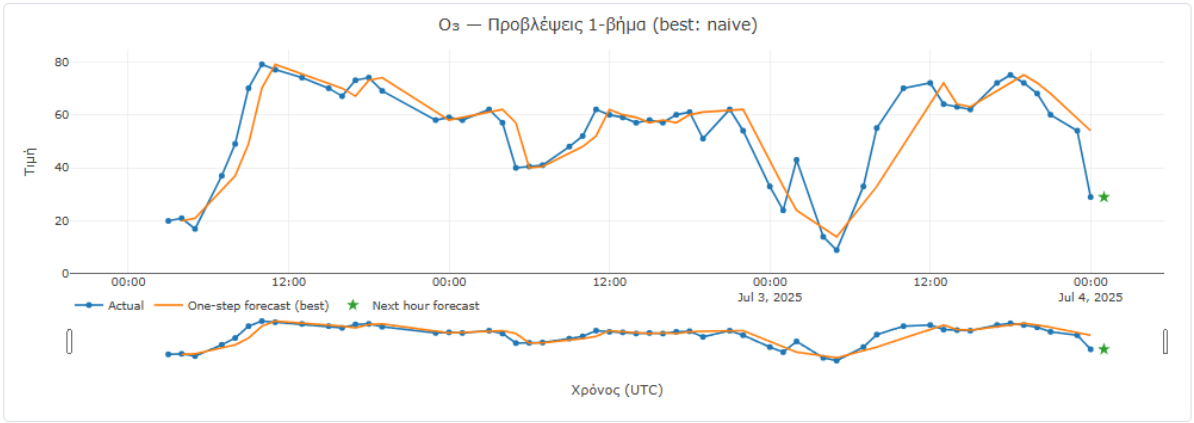
Εικόνα 5.9: Μετρικές και επόμενη πρόβλεψη για όλους τους αισθητήρες











Εικόνα 5.10: Προβλέψεις 1-βήμα για όλους τους αισθητήρες

CO: Αξιολογήθηκαν πέντε απλές μέθοδοι 1-βήμα (Naïve, Seasonal 24h, Moving Average, SES, Drift) με rolling backtest (>=48 ώρες ιστορικού). Καλύτερη κατά MASE αναδείχθηκε η *naive*. Στο γράφημα φαίνονται οι τελευταίες 72 ώρες (Actual) και η αντίστοιχη 1-βήμα πρόβλεψη της καλύτερης μεθόδου καθώς και ένας marker με την πρόβλεψη της επόμενης ώρας.

PM_{10}: Αξιολογήθηκαν πέντε απλές μέθοδοι 1-βήμα (Naïve, Seasonal 24h, Moving Average, SES, Drift) με rolling backtest (>=48 ώρες ιστορικού). Καλύτερη κατά MASE αναδείχθηκε η *naive*. Στο γράφημα φαίνονται οι τελευταίες 72 ώρες (Actual) και η αντίστοιχη 1-βήμα πρόβλεψη της καλύτερης μεθόδου καθώς και ένας marker με την πρόβλεψη της επόμενης ώρας.

NO_2: Αξιολογήθηκαν πέντε απλές μέθοδοι 1-βήμα (Naïve, Seasonal 24h, Moving Average, SES, Drift) με rolling backtest (>=48 ώρες ιστορικού). Καλύτερη κατά MASE αναδείχθηκε η *naive*. Στο γράφημα φαίνονται οι τελευταίες 72 ώρες (Actual) και η αντίστοιχη 1-βήμα πρόβλεψη της καλύτερης μεθόδου καθώς και ένας marker με την πρόβλεψη της επόμενης ώρας.

SO_2: Αξιολογήθηκαν πέντε απλές μέθοδοι 1-βήμα (Naïve, Seasonal 24h, Moving Average, SES, Drift) με rolling backtest (>=48 ώρες ιστορικού). Καλύτερη κατά MASE αναδείχθηκε η *ma*. Στο γράφημα φαίνονται οι τελευταίες 72 ώρες (Actual) και η αντίστοιχη 1-βήμα πρόβλεψη της καλύτερης μεθόδου καθώς και ένας marker με την πρόβλεψη της επόμενης ώρας.

O_3: Αξιολογήθηκαν πέντε απλές μέθοδοι 1-βήμα (Naïve, Seasonal 24h, Moving Average, SES, Drift) με rolling backtest (>=48 ώρες ιστορικού). Καλύτερη κατά MASE αναδείχθηκε η *naive*. Στο γράφημα φαίνονται οι τελευταίες 72 ώρες (Actual) και η αντίστοιχη 1-βήμα πρόβλεψη της καλύτερης μεθόδου καθώς και ένας marker με την πρόβλεψη της επόμενης ώρας.

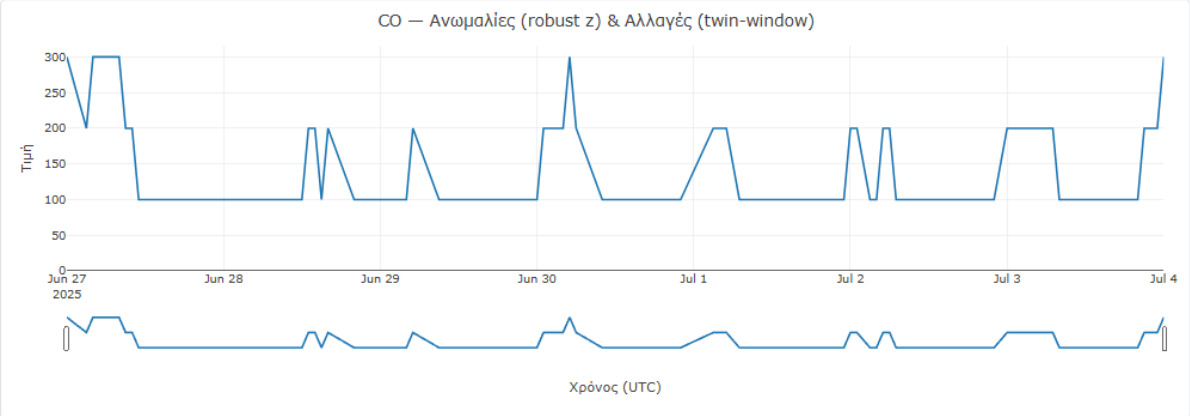
5.10 Ανωμαλίες & Αλλαγές Καθεστώτος

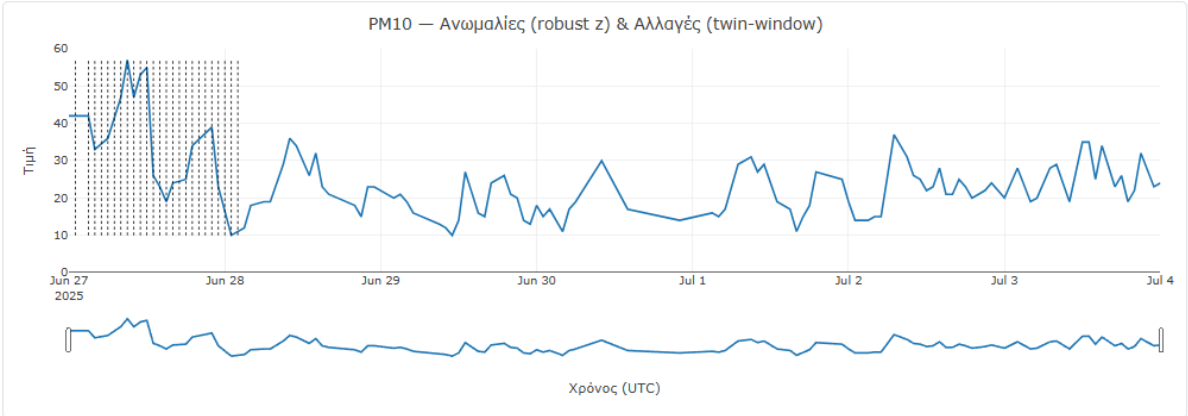
Robust z (rolling median+MAD) για στιγμιαίες αιχμές και δίδυμα παράθυρα 24h για μετατοπίσεις επιπέδου. Τα γεγονότα σημειώνονται στον χρόνο και εμφανίζονται σε πίνακα με score.

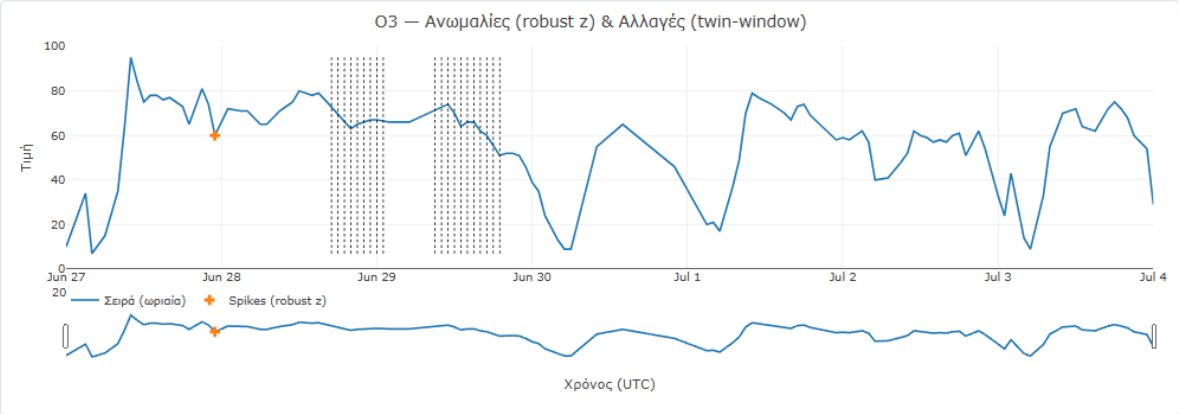
Alerts — CO

dt	type	score	value
2025-06-19 01:00:00+00:00	changepoint	2.94	100.0
2025-06-18 23:00:00+00:00	changepoint	2.75	NaN
2025-06-19 00:00:00+00:00	changepoint	2.75	NaN
2025-06-18 22:00:00+00:00	changepoint	2.62	100.0
2025-06-19 02:00:00+00:00	changepoint	2.61	NaN
2025-06-19 03:00:00+00:00	changepoint	2.61	100.0
2025-06-18 21:00:00+00:00	changepoint	2.33	100.0
2025-06-18 19:00:00+00:00	changepoint	2.24	100.0
2025-06-18 18:00:00+00:00	changepoint	2.16	100.0
2025-06-18 20:00:00+00:00	changepoint	2.11	100.0
2025-06-18 17:00:00+00:00	changepoint	2.08	100.0
2025-06-19 04:00:00+00:00	changepoint	1.84	200.0
2025-06-23 03:00:00+00:00	changepoint	1.56	NaN
2025-06-26 23:00:00+00:00	spike	5.40	400.0

Εικόνα 5.11: Alerts για το CO







Εικόνα 5.12: Ανωμαλίες και αλλαγές για όλους τους αισθητήρες

CO: Spikes με robust z (όριο |z|>5.0), αλλαγές καθεστώτος με δίδυμα παράθυρα 24h και κατώφλι 1.5×std. Τα όρια είναι ήπια προεπιλογή και μπορούν να ρυθμιστούν.

PM10: Spikes με robust z (όριο |z|>5.0), αλλαγές καθεστώτος με δίδυμα παράθυρα 24h και κατώφλι 1.5×std. Τα όρια είναι ήπια προεπιλογή και μπορούν να ρυθμιστούν.

NO2: Spikes με robust z (όριο |z|>5.0), αλλαγές καθεστώτος με δίδυμα παράθυρα 24h και κατώφλι 1.5×std. Τα όρια είναι ήπια προεπιλογή και μπορούν να ρυθμιστούν.

SO2: Spikes με robust z (όριο |z|>5.0), αλλαγές καθεστώτος με δίδυμα παράθυρα 24h και κατώφλι 1.5×std. Τα όρια είναι ήπια προεπιλογή και μπορούν να ρυθμιστούν.

O3: Spikes με robust z (όριο |z|>5.0), αλλαγές καθεστώτος με δίδυμα παράθυρα 24h και κατώφλι 1.5×std. Τα όρια είναι ήπια προεπιλογή και μπορούν να ρυθμιστούν.

5.11 AQI & Υπερβάσεις

Παραμετρικός υπολογισμός υπο-δεικτών με breakpoints και συνολικού AQI ως μέγιστο. Υπερβάσεις σε ημερήσιο PM10, 8ωρο O3, ωριαίο NO2.

Υπο-δείκτες (τελευταίες 48 εγγραφές)

	PM10_daily_index	O3_8h_index	NO2_1h_index	AQI	TopContributor
dt					
2025-07-02 01:00:00+00:00	NaN	NaN	5.0	5.00	NO2_1h
2025-07-02 02:00:00+00:00	NaN	NaN	NaN	NaN	None
2025-07-02 03:00:00+00:00	NaN	NaN	2.5	2.50	NO2_1h
2025-07-02 04:00:00+00:00	NaN	NaN	4.0	4.00	NO2_1h
2025-07-02 05:00:00+00:00	NaN	27.83	10.5	27.83	O3_8h
2025-07-02 06:00:00+00:00	NaN	26.75	25.5	26.75	O3_8h
2025-07-02 07:00:00+00:00	NaN	25.54	13.0	25.54	O3_8h
2025-07-02 08:00:00+00:00	NaN	24.88	NaN	24.88	O3_8h
2025-07-02 09:00:00+00:00	NaN	24.04	9.5	24.04	O3_8h
2025-07-02 10:00:00+00:00	NaN	24.32	7.5	24.32	O3_8h
2025-07-02 11:00:00+00:00	NaN	24.32	3.5	24.32	O3_8h
2025-07-02 12:00:00+00:00	NaN	24.54	5.5	24.54	O3_8h
2025-07-02 13:00:00+00:00	NaN	25.89	5.0	25.89	O3_8h
2025-07-02 14:00:00+00:00	NaN	27.07	6.0	27.07	O3_8h
2025-07-02 15:00:00+00:00	NaN	28.29	3.5	28.29	O3_8h
2025-07-02 16:00:00+00:00	NaN	28.31	3.0	28.31	O3_8h
2025-07-02 17:00:00+00:00	NaN	29.06	1.5	29.06	O3_8h

Εικόνα 5.13: Υποδείκτες για το AQI

Κατανομή ωρών ανά κλάση AQI

Class	Hours
Good	476
Moderate	5

Υπερβάσεις PM10 (ημερήσιος μέσος > 50.0)

Καμία.

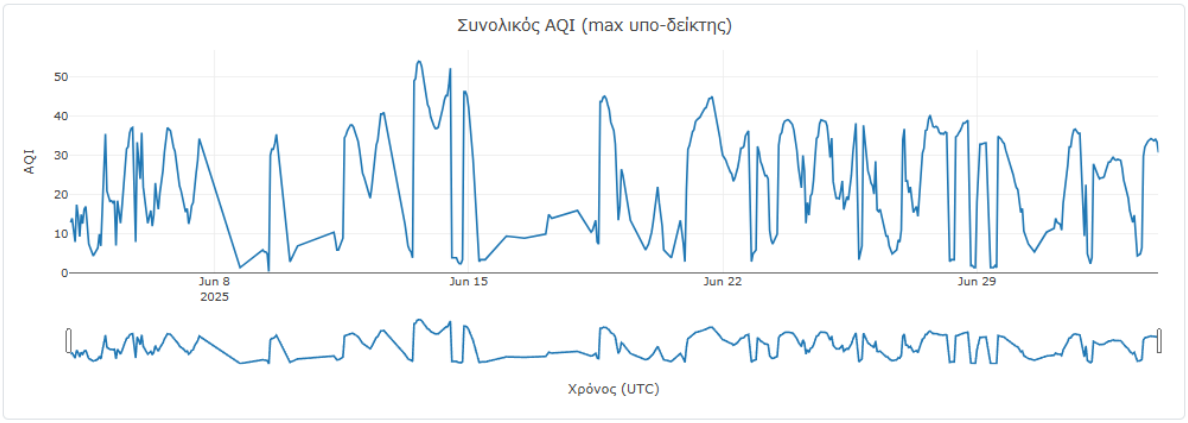
Υπερβάσεις O3 (8ωρος μέσος > 120.0)

Καμία.

Υπερβάσεις NO2 (ωριαίο > 200.0)

Καμία.

Εικόνα 5.14: Κατανομή ωρών ανά κλάση AQI και υπερβάσεις



Εικόνα 5.15: Συνολικός AQI

Ο συνολικός AQI απεικονίζει την χειρότερη (μεγαλύτερη) συμβολή ανά χρονική στιγμή. Ο top contributor βοηθά στην αιτιολόγηση. Οι υπερβάσεις εντοπίζονται στα κατάλληλα χρονικά παράγωγα.

Κεφάλαιο 6ο: Συμπεράσματα και προτάσεις βελτίωσης

Η εργασία παρουσίασε μια λειτουργική πλατφόρμα ανάλυσης δεδομένων αισθητήρων ατμοσφαιρικής ρύπανσης, με στόχο (α) την αξιόπιστη προεπεξεργασία και ενοποίηση δεδομένων, (β) την παραγωγή τυπικών αλλά και προχωρημένων αναλύσεων χρονοσειρών, (γ) την ανίχνευση ανωμαλιών και αλλαγών καθεστώτος, (δ) τη βασική πρόβλεψη επόμενης τιμής, και (ε) τον υπολογισμό παραμετρικών δεικτών AQI/υπερβάσεων.

Η αρχιτεκτονική (Flask + Pandas/NumPy + Plotly + Bootstrap, προαιρετικά statsmodels/STL) είναι modular, ώστε κάθε μέθοδος να αναβαθμίζεται ανεξάρτητα, και όλα τα γραφήματα/πίνακες να προκύπτουν αυτόματα από την ίδια «καθαρή» πηγή δεδομένων..

Η ακολουθία QC (έλεγχος αρνητικών/μη ρεαλιστικών τιμών, Global IQR, Hampel σε rolling παράθυρο) αποδείχθηκε επαρκής και ανθεκτική για δεδομένα αισθητήρων με σποραδικά spikes και ακανόνιστες κατανομές. Η παρεμβολή μικρών κενών (time-based, γραμμική, με σαφές limit σε διαδοχικά κενά) βελτιώνει τη συνέχεια της χρονοσειράς για υπολογισμούς/οπτικοποίηση, χωρίς να αλλοιώνει σημαντικά τη μεταβλητότητα. Όλες οι παρεμβάσεις σημαίνονται με flags για ιχνηλασιμότητα. Η κανονικοποίηση χρόνου σε UTC και η αυστηρή αύξουσα ταξινόμηση (ASC) σε όλα τα στάδια είναι κρίσιμες: αποφεύγονται σφάλματα λόγω μετατροπών ζώνης ώρας ή ανακατεμένων timestamps, ειδικά σε cross-parameter συσχετίσεις και resampling. Ο κυλιόμενος μέσος (~24h) λειτουργεί ως απλό αλλά αποτελεσματικό φίλτρο τάσης, επιτρέποντας οπτικό διαχωρισμό θορύβου/επεισοδίων από το υπόβαθρο. Οι ημερήσιοι μέσοι με κανόνα πληρότητας (π.χ. ≥18/24 ωρών) προσφέρουν συγκρίσιμη και σταθερή σύνοψη ημερών, χρήσιμη για επιχειρησιακές αναφορές. Το προφίλ 24ώρου (diurnal) αποτυπώνει τυπικές ωριαίες δομές (π.χ. αιχμές NOx πρωί/βράδυ, ανόδους O_3 μεσημέρι) και αποτελεί αξιόπιστη «υπογραφή» συστήματος. Η συσχέτιση Pearson στα κοινά timestamps αναδεικνύει σχέσεις και εν δυνάμει κοινούς μηχανισμούς (θετική) ή ανταγωνιστικές διεργασίες (αρνητική), ενώ η cross-correlation με lag δίνει ένδειξη πιθανής χρονικής καθυστέρησης μεταξύ ρύπων (π.χ. «ο Α ακολουθεί τον Β κατά k ώρες»). Η STL αποσύνθεση (Trend/Seasonal/Residual, period=24) διευκολύνει τον αναλυτικό διαχωρισμό ημερήσιας εποχικότητας και μακρο-τάσης από τα υπολειμματικά επεισόδια. Η robust παραλλαγή είναι κατάλληλη για αισθητήρες με outliers, με προϋπόθεση επαρκές μήκος σειράς.

Επίσης, ο robust z-score (rolling median + MAD) εντοπίζει αξιόπιστα σημειακές ανωμαλίες (spikes) χωρίς να «τιμωρεί» τη φυσική ασυμμετρία της κατανομής. Ο δείκτης δίδυμων παραθύρων (διαφορά διαδοχικών rolling μέσων/διακυμάνσεων) εντοπίζει μετατοπίσεις επιπέδου (changepoints) με χαμηλό υπολογιστικό κόστος και ερμηνεύσιμο κατώφλι. Η οπτική σήμανση (markers/κατακόρυφες γραμμές) και ο συνοδευτικός πίνακας alerts καθιστούν τον έλεγχο γεγονότων άμεσο και αναπαραγώγιμο.

Οσον αφόρα την πρόβλεψη οι baseline μέθοδοι (Naïve, Seasonal-24h, Moving Average, SES, Drift) με rolling one-step backtest και μετρικές MAE/RMSE/MASE παρέχουν τίμια, ερμηνεύσιμη απόδοση. Η αυτόματη επιλογή «καλύτερης» μεθόδου μέσω MASE είναι πρακτική: η αξιολόγηση σε πραγματικές συνθήκες δείχνει ότι οι απλές μέθοδοι στέκονται ικανοποιητικά σε βραχυπρόθεσμους ορίζοντες, ειδικά όταν η σειρά εμφανίζει ισχυρή ημερήσια εποχικότητα (Seasonal-Naïve) ή ομαλή τάση (SES/Drift).Η προβολή της επόμενης ώρας μαζί με το ιστορικό one-step forecast βελτιώνει την επιχειρησιακή χρησιμότητα (γρήγορη εκτίμηση κινδύνου).

Οσον αφορά τα AQI και υπερβάσεις η παραμετρική υλοποίηση υπο-δεικτών AQI (breakpoints + γραμμική παρεμβολή) επιτρέπει την εύκολη προσαρμογή σε διαφορετικά κανονιστικά πλαίσια. Ο συνολικός AQI ως μέγιστος υπο-δείκτης ανά timestamp, μαζί με τον top contributor, απαντά στο «πόσο καλή/κακή είναι η ποιότητα» και «ποιος ρύπος ευθύνεται περισσότερο». Η λογική υπερβάσεων (PM10 ημερήσιος, O_3 8ωρος, NO_2 ωριαίος) με κανόνες πληρότητας παράγει αναφορές συμβατές με επιχειρησιακούς ελέγχους.

Η επόμενη φάση μπορεί να επικεντρωθεί στην ενίσχυση της προγνωστικής ικανότητας και της επιχειρησιακής χρησιμότητας της πλατφόρμας. Πρώτα, στο κομμάτι του forecasting, προτείνεται η ενσωμάτωση μοντέλων Holt–Winters με 24ωρη εποχικότητα για βραχυπρόθεσμες προβλέψεις που σέβονται τον ημερήσιο κύκλο καθώς και SARIMA/SARIMAX ώστε να αξιοποιούνται εξωγενείς μεταβλητές (π.χ. μετεωρολογικά).

Στο κομμάτι αναφορών, είναι χρήσιμη η παραγωγή εξαγωγών PDF/Excel ανά ενότητα με ενσωματωμένα γραφήματα και συνοπτικά συμπεράσματα, ώστε να υποστηρίζονται τυπικές διαδικασίες τεκμηρίωσης ή αρχειοθέτησης. Η δημιουργία προτύπων (templates) για «γρήγορες» επιχειρησιακές αναφορές (π.χ. γεγονότα ανωμαλιών τελευταίας εβδομάδας, υπερβάσεις ανά ρύπο/μήνα) θα αυτοματοποιήσει την επικοινωνία ευρημάτων.

Για διαλειτουργικότητα και αυτοματοποίηση, προτείνεται η υλοποίηση connectors προς APIs όπως OpenAQ και μετεωρολογικές υπηρεσίες, με δυνατότητα προγραμματισμένων λήψεων και εκτελέσεων (cron jobs). Έτσι, η πλατφόρμα θα μπορεί να τραβά περιοδικά νέα δεδομένα, να εκτελεί QC/αναλύσεις/προβλέψεις και να δημοσιεύει ενημερωμένες αναφορές χωρίς χειροκίνητη παρέμβαση. Σε συνδυασμό με έναν απλό μηχανισμό ειδοποιήσεων (π.χ. email για υπερβάσεις), το σύστημα θα μετασχηματιστεί από εργαλείο διερεύνησης σε εργαλείο συνεχούς παρακολούθησης και υποστήριξης αποφάσεων.

Κατά τη σύνταξη της παρούσας εργασίας έγινε χρήση του εργαλείου chatgpt για υποστήριξη στη διατύπωση κειμένων, στη διόρθωση εκφράσεων και στη βελτίωση της σαφήνειας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] https://explore.openaq.org/locations/2162787

[2] https://data.europa.eu/el/dataeuropa-academy/what-open-data

[3] https://project.opendatamonitor.eu/

[4] https://www.python.org/

[5] https://flask.palletsprojects.com/en/stable/

[6] https://pandas.pydata.org/

[7] https://numpy.org/

[8] https://www.geeksforgeeks.org/machine-learning/understanding-global-outliers/

[9] https://en.wikipedia.org/wiki/Interquartile_range

[10] https://www.mathworks.com/help/signal/ref/hampel.html

[11] https://blogs.sas.com/content/iml/2021/06/01/hampel-filter-robust-outliers.html

[12] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[13] https://en.wikipedia.org/wiki/Cross-correlation

[14] https://www.geeksforgeeks.org/data-analysis/seasonal-decomposition-of-time-series-by-loess-stl/

[15] https://otexts.com/fpp2/simple-methods.html

[16] https://en.wikipedia.org/wiki/Moving_average

[17] https://otexts.com/fpp2/ses.html

[18] https://en.wikipedia.org/wiki/Concept_drift

[19] https://skforecast.org/0.14.0/faq/parameters-search-backtesting-vs-one-step-ahead.html

[20] https://www.pmorgan.com.au/tutorials/mae%2C-mape%2C-mase-and-the-scaled-rmse/

[21] https://en.wikipedia.org/wiki/Air_quality_index

[22] https://plotly.com/

[23] https://www.w3schools.com/python/pandas/default.asp

ΠΑΡΑΡΤΗΜΑ Α

```
# services/analysis.py

from __future__ import annotations
from .forecast import run_forecast_sections
from .anomalies import run_anomalies_sections
from .aqi import run_aqi_sections


from typing import List, Dict
import numpy as np
import pandas as pd

try:
    from statsmodels.tsa.seasonal import STL  # pip install statsmodels
    HAS_STL = True
except Exception:
    STL = None
    HAS_STL = False

# Βοηθητικές συναρτήσεις για σχήματα (Plotly) και επεξηγηματικά κείμενα (narratives)
from .plotly_helpers import (
    line_raw_and_rolling,
    daily_mean,
    diurnal_profile,
    correlation_heatmap,
    xcorr_bar,
    stl_component_figure,
)
from .narrative import (
    timeseries_narrative,
    daily_means_narrative,
    diurnal_narrative,
    correlation_narrative,
    xcorr_narrative,
    stl_narrative,
)

# Παράμετροι/ρύποι που υποστηρίζουμε (σειρά εμφάνισης)
POLLUTANTS = ["co", "no", "pm10", "no2", "so2", "o3"]

# Σύντομες περιγραφές που εμφανίζονται πάνω από κάθε ενότητα (η μεθοδολογία)
DESC_TIMESERIES = (
    "<p><strong>Χρονοσειρές:</strong> Εμφάνιση ωριαίων τιμών και εξομαλυμένης καμπύλης "
    "(κυλιόμενος μέσος ~24h) για ανάδειξη υποκείμενης τάσης και εντοπισμό επεισοδίων.</p>"
```

```
)

DESC_DAILY = (
    "<p><strong>Ημερήσιοι Μέσοι:</strong> Συμπύκνωση ωριαίων σε μέση τιμή ανά ημέρα για καθαρή "
    "σύγκριση ημερών/εβδομάδων και ανάδειξη περιόδων με αυξημένες ή μειωμένες συγκεντρώσεις.</p>"
)

DESC_DIURNAL = (
    "<p><strong>Προφίλ 24ώρου:</strong> Μέσος όρος ανά ώρα (0–23) για να φανεί το τυπικό ωριαίο "
    "μοτίβο (π.χ. πρωινές/βραδινές αιχμές NOx, μεσημβρινό O₃).</p>"
)

DESC_CORR = (
    "<p><strong>Συσχέτιση:</strong> Συντελεστής Pearson μεταξύ παραμέτρων στα κοινά timestamps. "
    "Υψηλές θετικές τιμές υποδεικνύουν πιθανές κοινές πηγές/συνθήκες, αρνητικές αντιστροφή.</p>"
)


def run_analysis_sections(df_clean: pd.DataFrame) -> List[Dict]:
    """
    Χτίζει όλες τις ενότητες ανάλυσης για το report:

      1) Χρονοσειρές (raw + rolling)
      2) Ημερήσιοι μέσοι
      3) Προφίλ 24ώρου
      4) Συσχέτιση (Pearson)
      5) Cross-correlation με υστέρηση (lag)
      6) Αποεποχικοποίηση (STL, period=24h) — προαιρετική (αν υπάρχει statsmodels)

    Σημειώσεις:
      • Προτιμά *_clean αν υπάρχει, αλλιώς raw.
      • Όλα τα plots δουλεύουν σε ASC dt.
      • Για xcorr/STL γίνεται resample σε ωριαίο ρυθμό.
    """
    sections: List[Dict] = []

    if df_clean is None or df_clean.empty or "dt" not in df_clean.columns:
        return sections

    # --- Ταξινόμηση dt μία φορά (ASC) ---
    df_plot = df_clean.sort_values("dt").reset_index(drop=True)

    # --- Επιλογή διαθέσιμων σειρών (προτίμηση *_clean) ---
    available: List[str] = []
    base_cols: List[str] = []  # ονόματα χωρίς _clean
    for p in POLLUTANTS:
        if f"{p}_clean" in df_plot.columns:
            available.append(f"{p}_clean")
```

```
            base_cols.append(p)
        elif p in df_plot.columns:
            available.append(p)
            base_cols.append(p)

    if not available:
        return sections

    # ====================================================
    # 1) ΧΡΟΝΟΣΕΙΡΕΣ (raw + rolling)
    # ====================================================
    figs_ts: List[Dict] = []
    for src, base in zip(available, base_cols):
        series_df = pd.DataFrame({"dt": df_plot["dt"], base: df_plot[src]})
        figs_ts.append(line_raw_and_rolling(series_df, base))
    sections.append({
        "title": "Χρονοσειρές",
        "description": DESC_TIMESERIES,
        "tables": [],
        "figures": figs_ts,
        "narrative": timeseries_narrative(
            pd.DataFrame({"dt": df_plot["dt"], **{b: df_plot[s] for s, b in zip(available,
base_cols)}}),
            base_cols
        ),
    })

    # ====================================================
    # 2) ΗΜΕΡΗΣΙΟΙ ΜΕΣΟΙ
    # ====================================================
    figs_daily: List[Dict] = []
    for src, base in zip(available, base_cols):
        figs_daily.append(daily_mean(pd.DataFrame({"dt": df_plot["dt"], base: df_plot[src]}), base))
    sections.append({
        "title": "Ημερήσιοι Μέσοι",
        "description": DESC_DAILY,
        "tables": [],
        "figures": figs_daily,
        "narrative": daily_means_narrative(
            pd.DataFrame({"dt": df_plot["dt"], **{b: df_plot[s] for s, b in zip(available,
base_cols)}}),
            base_cols
        ),
    })

    # ====================================================
    # 3) ΠΡΟΦΙΛ 24ΩΡΟΥ
    # ====================================================
```

```
    figs_diurnal: List[Dict] = []
    for src, base in zip(available, base_cols):
        figs_diurnal.append(diurnal_profile(pd.DataFrame({"dt": df_plot["dt"], base: df_plot[src]}),
base))
    sections.append({
        "title": "Προφίλ 24ώρου",
        "description": DESC_DIURNAL,
        "tables": [],
        "figures": figs_diurnal,
        "narrative": diurnal_narrative(
            pd.DataFrame({"dt": df_plot["dt"], **{b: df_plot[s] for s, b in zip(available,
base_cols)}}),
            base_cols
        ),
    })

    # ======================================================
    # 4) ΣΥΣΧΕΤΙΣΗ (PEARSON)
    # ======================================================
    corr_df = pd.DataFrame({"dt": df_plot["dt"]})
    # κράτησε μία φορά κάθε base όνομα (σε περίπτωση διπλοεγγραφών)
    uniq: List[str] = []
    for b in base_cols:
        if b not in uniq:
            uniq.append(b)
    for src, name in zip(available, base_cols):
        if name in uniq:
            corr_df[name] = df_plot[src]
    fig_corr = correlation_heatmap(corr_df, uniq)
    sections.append({
        "title": "Συσχέτιση Παραμέτρων",
        "description": DESC_CORR,
        "tables": [],
        "figures": [fig_corr],
        "narrative": correlation_narrative(corr_df, uniq),
    })

    # ======================================================
    # 5) CROSS-CORRELATION ΜΕ ΥΣΤΕΡΗΣΗ (LAG)
    #    Επιλέγουμε κατά προτεραιότητα O3 ↔ NO2,
    #    αλλιώς το πρώτο διαθέσιμο ζεύγος.
    # ======================================================
    pair = None
    if "o3" in uniq and "no2" in uniq:
        pair = ("o3", "no2")
    elif len(uniq) >= 2:
        pair = (uniq[0], uniq[1])
```

```
    if pair is not None:
        # μικρό df με τα δύο πεδία και resample σε ωριαίο
        df_pair = pd.DataFrame({"dt": df_plot["dt"]})
        name_to_series = {b: df_plot[s] for s, b in zip(available, base_cols)}
        if pair[0] in name_to_series and pair[1] in name_to_series:
            df_pair[pair[0]] = name_to_series[pair[0]]
            df_pair[pair[1]] = name_to_series[pair[1]]
            g = df_pair.set_index("dt").resample("H").mean()

            a = g[pair[0]]
            b = g[pair[1]]

            max_lag = 48  # ώρες
            lags = list(range(-max_lag, max_lag + 1))
            rvals: List[float] = []
            for L in lags:
                if L < 0:
                    # αρνητικό lag: το δεύτερο προηγείται (shift το πρώτο προς το μέλλον)
                    r = a.shift(-L).corr(b)
                else:
                    # θετικό lag: το δεύτερο ακολουθεί (shift το δεύτερο προς το μέλλον)
                    r = a.corr(b.shift(L))
                rvals.append(0.0 if pd.isna(r) else float(r))

            # καλύτερο lag με βάση |r|
            best_lag = 0
            best_r = 0.0
            if any(np.isfinite(rv) for rv in rvals):
                abs_arr = np.array([abs(rv) if np.isfinite(rv) else -np.inf for rv in rvals])
                idx = int(abs_arr.argmax())
                best_lag = lags[idx]
                best_r = rvals[idx]

            xcorr_fig = xcorr_bar(lags, rvals, title=f"Cross-correlation {pair[0].upper()} ↔
{pair[1].upper()} (lags σε ώρες)")
            sections.append({
                "title": "Cross-correlation με υστέρηση (lag)",
                "description": "<strong>Μέθοδος:</strong> Pearson r για σειρά lags μεταξύ δύο
μεταβλητών.",
                "tables": [],
                "figures": [xcorr_fig],
                "narrative": xcorr_narrative(
                    f"{pair[0].upper()} ↔ {pair[1].upper()}",
                    best_lag=best_lag,
                    best_r=best_r,
                    window_desc="−48..+48 ώρες",
                ),
            })
```

```
    # ====================================================
    # 6) STL DECOMPOSITION (period = 24 ώρες) — αν υπάρχει statsmodels
    # ====================================================
    if HAS_STL:
        period = 24  # ημερήσιος κύκλος για ωριαία δεδομένα
        stl_figs: List[Dict] = []
        stl_narr_parts: List[str] = []

        dt_index = pd.to_datetime(df_plot["dt"])
        # map base name -> series (προτεραιότητα *_clean)
        base_map: Dict[str, pd.Series] = {}
        for src, name in zip(available, base_cols):
            if name not in base_map:
                base_map[name] = df_plot[src]

        for name, series in base_map.items():
            # Resample ωριαίο + συντηρητική παρεμβολή για STL
            g = pd.DataFrame({"dt": dt_index, name: series}).set_index("dt").resample("H").mean()
            s = g[name].interpolate(limit=2)  # συμπλήρωση μικρών κενών μόνο για σταθερότητα STL
            # απαιτείται επαρκές μήκος (τουλ. ~3 periods) για νόημα
            if s.dropna().shape[0] < period * 3:
                continue

            # STL (robust=True για ανθεκτικότητα σε outliers)
            stl = STL(s, period=period, robust=True)
            res = stl.fit()

            # Γραφήματα components
            stl_figs.append(stl_component_figure(s.index, s.values, "Trend", f"{name.upper()} —
Trend"))
            stl_figs.append(stl_component_figure(s.index, res.seasonal.values, "Seasonal",
f"{name.upper()} — Seasonal (period=24h)"))
            stl_figs.append(stl_component_figure(s.index, res.resid.values, "Remainder",
f"{name.upper()} — Remainder"))

            # Ενδεικτική συνεισφορά διακύμανσης
            try:
                var_total = float(np.nanvar(s.values))
                shares = None
                if var_total > 0:
                    shares = {
                        "seasonal": 100.0 * float(np.nanvar(res.seasonal.values)) / var_total,
                        "trend":    100.0 * float(np.nanvar(res.trend.values)) / var_total,
                        "remainder":100.0 * float(np.nanvar(res.resid.values)) / var_total,
                    }
            except Exception:
                shares = None
```

```
            stl_narr_parts.append(stl_narrative(name.upper(), period=period, var_shares=shares))

        if stl_figs:
            sections.append({
                "title": "Αποεποχικοποίηση (STL)",
                "description": "<strong>Μέθοδος:</strong> Διάσπαση της σειράς σε Trend, Seasonal και
Remainder (period=24).",
                "tables": [],
                "figures": stl_figs,
                "narrative": "".join(stl_narr_parts),
            })
    else:
        # Προαιρετική ενημέρωση ώστε το report να είναι «αυτοεξηγούμενο»
        sections.append({
            "title": "Αποεποχικοποίηση (STL)",
            "description": "<strong>Μέθοδος:</strong> STL decomposition",
            "tables": [],
            "figures": [],
            "narrative": (
                "<p>Η ενότητα STL παραλείφθηκε επειδή το πακέτο <code>statsmodels</code> "
                "δεν είναι εγκατεστημένο. Εγκατάσταση: "
                "<code>python -m pip install --upgrade numpy scipy statsmodels==0.14.2</code>.</p>"
            ),
        })

    # ----------------------------------------------------
    # 7) ΠΡΟΒΛΕΨΕΙΣ (1-βήμα)
    # ----------------------------------------------------
    try:
        forecast_sections = run_forecast_sections(df_plot)
        if forecast_sections:
            sections.extend(forecast_sections)
    except Exception as ex:
        # Προαιρετικά: εμφάνισε ενημερωτικό narrative αντί για crash
        sections.append({
            "title": "Προβλέψεις (1-βήμα μπροστά)",
            "description": "<strong>Μέθοδος:</strong> Απλές προβλέψεις επόμενης ώρας.",
            "tables": [],
            "figures": [],
            "narrative": f"<p>Η ενότητα προβλέψεων παραλείφθηκε λόγω σφάλματος: {str(ex)}</p>",
        })
# 8) Ανωμαλίες & Αλλαγές Καθεστώτος
    try:
        anom_secs = run_anomalies_sections(df_plot)
        if anom_secs:
            sections.extend(anom_secs)
    except Exception as ex:
        sections.append({
```

```
            "title": "Ανωμαλίες & Αλλαγές Καθεστώτος",
            "description": "<strong>Μέθοδος:</strong> Robust z + twin-window",
            "tables": [], "figures": [],
            "narrative": f"<p>Παράλειψη λόγω σφάλματος: {str(ex)}</p>"
        })

    # 9) AQI & Υπερβάσεις (DEMO)
    try:
        aqi_secs = run_aqi_sections(df_plot)
        if aqi_secs:
            sections.extend(aqi_secs)
    except Exception as ex:
        sections.append({
            "title": "AQI & Υπερβάσεις",
            "description": "<strong>Μέθοδος:</strong> Παραμετρικός AQI",
            "tables": [], "figures": [],
            "narrative": f"<p>Παράλειψη λόγω σφάλματος: {str(ex)}</p>"
        })

    return sections
```