



Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Πολυτεχνική Σχολή

Τμήμα Μηχανολόγων Μηχανικών

Ενεργειακός Τομέας

Ερευνητική Ομάδα Περιβαλλοντικής Πληροφορικής

Διπλωματική Εργασία

Ανάλυση και μοντελοποίηση δεδομένων
κόμβων παρακολούθησης ποιότητας αέρα με
μεθόδους Υπολογιστικής Νοημοσύνης

Παναγιώτης Κεφαλάς

Επιβλέπων Καθηγητής

Κωνσταντίνος Καρατζάς

Θεσσαλονίκη, 2022

Περίληψη

Η αέρια ρύπανση αποτελεί τον σημαντικότερο περιβαλλοντικό παράγοντα επιβάρυνσης της υγείας στην Ευρώπη, με επιβλαβείς επιδράσεις στο καρδιαγγειακό και αναπνευστικό σύστημα, οδηγώντας έως και σε πρόωρη θνησιμότητα. Στην παρούσα εργασία πραγματοποιείται ανάλυση και παρουσιάζεται η τρέχουσα εικόνα ποιότητας αέρα στην πόλη της Θεσσαλονίκης, σύμφωνα με δεδομένα μετρήσεων που προέρχονται από κόμβους παρακολούθησης ποιότητας αέρα και μετεωρολογικό σταθμό, οι οποίοι βρίσκονται εγκατεστημένοι στην οροφή του κτιρίου Ε14 της Πολυτεχνικής Σχολής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης. Συγκρίνονται οι μετρήσεις των κόμβων ως προς την απόκλιση μεταξύ τους. Έμφαση δίνεται στη μελέτη του υδρόθειου (H_2S). Καταστρώνονται στατιστικά μοντέλα και μοντέλα μηχανικής μάθησης με στόχο την συμπλήρωση ελλείψεων στα δεδομένα μετρήσεων και την πρόβλεψη μελλοντικών επιπέδων αέριας ρύπανσης ενώ διερευνάται η απόδοσή τους. Τα αποτελέσματα έδειξαν ικανοποιητική ευστοχία των μοντέλων κατόπιν κατάλληλης παραμετροποίησης για την πλειοψηφία των υπό μελέτη ρύπων.

Λέξεις κλειδιά:

Ποιότητα αέρα, Αέρια ρύπανση, Αισθητήρες ποιότητας αέρα, Κόμβοι παρακολούθησης ποιότητας αέρα, Ανάλυση δεδομένων, Περιγραφική στατιστική, Μηχανική μάθηση, Υδρόθειο, Πρόγνωση

Abstract

Air pollution is considered the single most important environmental health risk factor in Europe, causing respiratory and cardiovascular disease, leading even to premature deaths. In this thesis analysis is being performed and results outlined concerning the current profile of air quality for the city of Thessaloniki, according to data measured by Air Quality and Weather Monitoring Nodes located on the roof of the building E14, Faculty of Engineering, of Aristotle University of Thessaloniki. The nodes are being compared in order to investigate correlation between measurements. Concentration levels of Hydrogen Sulfide (H_2S) are being investigated. Machine learning and statistical models are being utilized to fill missing data and forecast air pollutants future values. Their integrity is being validated. Results were accurate concerning most pollutants after proper model calibration.

Keywords

Air quality; Air pollution; Data analysis; Air quality sensors; Descriptive statistics; Machine learning; Hydrogen Sulfide; Forecasting

Περιεχόμενα

1. Εισαγωγή	12
1.1. Επιπτώσεις της αέριας ρύπανσης στο αστικό περιβάλλον	12
1.2. Εκτίμηση συγκέντρωσης αέριων ρύπων.....	12
1.3. Ερευνητικά ερωτήματα.....	14
2. Υλικά και Μέθοδοι	15
2.1. Υλικά.....	15
2.1.1. Αισθητήρες.....	16
2.1.2. Δεδομένα	17
2.1.2.1. Σύγκριση ΚΠΠΑ.....	18
2.1.3. Υλοποίηση μελέτης	18
2.2. Μέθοδοι.....	19
2.2.1. Ανάλυση Δεδομένων.....	19
2.2.1.1. Προεπεξεργασία – Συμπλήρωση ελλειπουσών τιμών	19
2.2.1.2. Αναδειγματοληψία	20
2.2.1.3. Τυπικά προφίλ	20
2.2.1.4. Διερεύνηση συσχέτισης	20
2.2.1.5. Περιγραφική Στατιστική.....	20
2.2.1.6. Αποσύνθεση χρονοσειράς	21
2.2.1.7. Κανονικοποίηση	21
2.2.1.8. Συνεισφορά ανέμου	22
2.2.1.9. Οπτικοποίηση – Συσταδοποίηση.....	22
2.2.2. Μοντελοποίηση δεδομένων πρόγνωσης.....	23
2.2.2.1. Παλινδρόμηση Ελαχίστων Τετραγώνων	24
2.2.2.2. SARIMAX.....	24
2.2.2.3. Random Forest Regressor	25
2.2.2.4. Διαδικασία επαλήθευσης μοντέλων	26
2.2.2.5. Αξιολόγηση πρόγνωσης	26
3. Αποτελέσματα και συζήτηση	28
3.1. Σύγκριση κόμβων	28
3.2. Ανάλυση δεδομένων.....	29
3.2.1. Προεπεξεργασία – Συμπλήρωση ελλειπουσών τιμών	29
3.2.2. Αναδειγματοληψία	32
3.2.3. Τυπικά προφίλ.....	33

3.2.4. Διερεύνηση συσχέτισης	34
3.2.5. Περιγραφική Στατιστική.....	35
3.2.6. Αποσύνθεση χρονοσειράς	37
3.2.7. Συνεισφορά ανέμου	39
3.2.8. Οπτικοποίηση – Συσταδοποίηση	42
3.2.8.1. Self-Organizing Maps (SOM)	42
3.2.8.2. K-Means	44
3.3. Μοντελοποίηση δεδομένων πρόγνωσης.....	47
3.3.1. Παλινδρόμηση ελαχίστων τετραγώνων – SARIMAX.....	48
3.3.2. Random Forest Regressor	53
3.3.3. Σύγκριση μοντέλων πρόγνωσης.....	56
4. Συμπεράσματα.....	57
5. Βιβλιογραφία	59
6. Παράρτημα.....	63
6.1. Διαγράμματα.....	64
6.1.1. Σύγκριση κόμβων	64
6.1.2. Συμπλήρωση ελλειπουσών τιμών.....	66
6.1.3. Αναδειγματοληψία	68
6.1.4. Τυπικά προφίλ.....	71
6.1.5. Διερεύνηση συσχέτισης	74
6.1.6. Περιγραφική Στατιστική.....	76
6.1.7. Αποσύνθεση χρονοσειράς	79
6.1.8. Συνεισφορά ανέμου	86
6.1.9. Παλινδρόμηση ελαχίστων τετραγώνων.....	89
6.1.10. SARIMAX.....	90
6.1.11. Random Forest	91

Κατάλογος Πινάκων

Πίνακας 1. Σύγκριση χαρακτηριστικών αναλυτικών οργάνων και αισθητήρων χαμηλού κόστους	13
Πίνακας 2. Κατασκευαστές και μετρούμενα μεγέθη ΚΠΠΑ	15
Πίνακας 3. Ενδεικτική μορφή δεδομένων ΚΠΠΑ Kunak	17
Πίνακας 4. Σφάλμα συμπλήρωσης ελλειπουσών τιμών θερμοκρασίας	31
Πίνακας 5. Σφάλμα συμπλήρωσης ελλειπουσών τιμών NO2	31
Πίνακας 6. Παράμετροι περιγραφικής στατιστικής (Aqmesh)	35
Πίνακας 7. Παράμετροι μοντέλου SARIMAX	49
Πίνακας 8. Συντελεστές παλινδρόμησης SARIMAX για πρόγνωση NO ₂ (ΚΠΠΑ Kunak)	49

Πίνακας 9. Παράμετροι αξιολόγησης SARIMAX κυλιόμενου δείγματος (Kunak)	51
Πίνακας 10. Παράμετροι αξιολόγησης SARIMAX αυξανόμενου δείγματος (Kunak)	52
Πίνακας 11. Αποτελέσματα 10-fold cross validation εκπαίδευσης μοντέλου Random Forest.....	53
Πίνακας 12. Παράμετροι αξιολόγησης πρόγνωσης μοντέλου Random Forest (Kunak) ...	53
Πίνακας 13. Παράμετροι αξιολόγησης μοντέλου RF κυλιόμενου δείγματος (Kunak)	54
Πίνακας 14. Παράμετροι αξιολόγησης RF αυξανόμενου δείγματος (Kunak)	55

Κατάλογος Εικόνων

Εικόνα 1. Σύγκριση στατιστικών μεγεθών για NO ₂ (ΚΠΠΑ Aqmesh, KASTOM, Kunak)....	28
Εικόνα 2. Συσχέτιση μετρήσεων NO2 μεταξύ κόμβων (ΚΠΠΑ Aqmesh, KASTOM, Kunak).	
Διαγράμματα συσχέτισης και ιστογράμματα (αριστερά), πίνακας συντελεστών συσχέτισης κατά Pearson (δεξιά).....	29
Εικόνα 3. Δοκιμαστικό τμήμα χρονοσειράς Θερμοκρασίας, προ συμπλήρωσης (ΚΠΠΑ Aqmesh)	29
Εικόνα 4. Δοκιμαστικό τμήμα χρονοσειράς Διοξειδίου του Αζώτου, προ συμπλήρωσης (ΚΠΠΑ Aqmesh)	30
Εικόνα 5. Δοκιμή συμπλήρωσης ελλειπουσών τιμών Θερμοκρασίας (ΚΠΠΑ Aqmesh)....	30
Εικόνα 6. Δοκιμή συμπλήρωσης ελλειπουσών τιμών NO2 (ΚΠΠΑ Aqmesh)	31
Εικόνα 7. Ωριαία (πάνω), ημερήσια (μέση) και εβδομαδιαία (κάτω) αναδειγματοληψία ρύπων (ΚΠΠΑ Kunak)	32
Εικόνα 8. Τυπικά εικοσιτετράωρα (πάνω) και εβδομαδιαία (κάτω) προφίλ (ΚΠΠΑ Kunak)	33
Εικόνα 9. Πίνακας συσχέτισης κατά Pearson, τιμών τυπικού εικοσιτετράρου (ΚΠΠΑ Kunak και KASTOM).....	34
Εικόνα 10. Ιστογράμματα ωριαίου ρυθμού δειγματοληψίας μετρούμενων μεγεθών (ΚΠΠΑ Aqmesh)	35
Εικόνα 11. Violin plot μέσου ημερήσιου ρυθμού δειγματοληψίας ρύπων (ΚΠΠΑ Kunak)	36
Εικόνα 12. Συνθετικά προσθετικού μοντέλου χρονοσειράς υδρόθειου (ΚΠΠΑ Kunak) ..	37
Εικόνα 13. Γρήγορος μετασχηματισμός Fourier χρονοσειράς υδρόθειου (ΚΠΠΑ Kunak) ..	38
Εικόνα 14. Ημερήσιες τάσεις εβδομαδιαίας περιοδικότητας ρύπων (Kunak).....	38
Εικόνα 15. Μέση τιμή συγκεντρώσεων ρύπων ανά διεύθυνση ανέμου (ΚΠΠΑ Kunak) ..	39
Εικόνα 16. Ροδογράμματα ρύπων ανά διεύθυνση και συχνότητα ανέμου (ΚΠΠΑ Kunak)	40
Εικόνα 17. Επίδραση μέσης κυρίαρχης διεύθυνσης ανέμου στις τιμές H ₂ S τυπικού εικοσιτετράρου (ΚΠΠΑ Kunak)	41
Εικόνα 18. Τιμές υδρόθειου υψηλότερες από το όριο πρόκλησης δυσσοσμίας.....	41
Εικόνα 19. Απεικόνιση χαρακτηριστικών SOM (ΚΠΠΑ Kunak)	42
Εικόνα 20. SOM συσταδοποίηση ημερήσιων προφίλ συγκεντρώσεων ρύπων (ΚΠΠΑ Kunak).....	43
Εικόνα 21. Τρισδιάστατη απεικόνιση δεδομένων μεθόδου PCA (ΚΠΠΑ Kunak)	44
Εικόνα 22. Elbow method για προσδιορισμό πλήθους συστάδων k (ΚΠΠΑ Kunak)	44
Εικόνα 23. Τρισδιάστατη απεικόνιση συσταδοποιημένων δεδομένων κατά K-Means (ΚΠΠΑ Kunak)	45

Εικόνα 24. Διαδιάστατη απεικόνιση συσταδοποιημένων δεδομένων κατά K-Means (ΚΠΠΑ Kunak)	46
Εικόνα 25. Ημερήσια προφίλ ρύπανσης και συχνότητα εμφάνισής τους κατά K-Means (ΚΠΠΑ Kunak)	46
Εικόνα 26. Διαχωρισμός της χρονοσειράς σε τμήμα εκπαίδευσης και τμήμα αξιολόγησης.	47
Εικόνα 27. Σύγκριση αποτελεσμάτων πρόγνωσης OLSR και πραγματικών δεδομένων (ΚΠΠΑ Kunak).	48
Εικόνα 28. Διαγράμματα αυτοσυσχέτισης αρχικής χρονοσειράς, διαφορισμένης κατά ένα βήμα και διαφορισμένης κατά 24 βήματα.	48
Εικόνα 29. Σύγκριση αποτελεσμάτων πρόγνωσης SARIMAX και πραγματικών δεδομένων (ΚΠΠΑ Kunak)	50
Εικόνα 30. Διακύμανση παραμέτρων αξιολόγησης SARIMAX κυλιόμενου δείγματος (ΚΠΠΑ Kunak). (. (Τιμές υπό κλίμακα για λόγους ευκρίνειας σύγκρισης).....	51
Εικόνα 31. Διακύμανση παραμέτρων αξιολόγησης SARIMAX αυξανόμενου δείγματος (ΚΠΠΑ Kunak). (. (Τιμές υπό κλίμακα για λόγους ευκρίνειας σύγκρισης).....	52
Εικόνα 32. Σύγκριση αποτελεσμάτων πρόγνωσης Random Forest και πραγματικών δεδομένων (ΚΠΠΑ Kunak)	53
Εικόνα 33. Βαρύτητα μεταβλητών στην διαμόρφωση του μοντέλου Random Forest (ΚΠΠΑ Kunak)	54
Εικόνα 34. Διακύμανση παραμέτρων αξιολόγησης RF κυλιόμενου δείγματος (ΚΠΠΑ Kunak) (. (Τιμές υπό κλίμακα για λόγους ευκρίνειας σύγκρισης).....	55
Εικόνα 35. Διακύμανση παραμέτρων αξιολόγησης RF αυξανόμενου δείγματος (ΚΠΠΑ Kunak) (. (Τιμές υπό κλίμακα για λόγους ευκρίνειας σύγκρισης).....	55
Εικόνα 36. Σύγκριση μοντέλων για κάθε ρύπο (τιμές υπό κλίμακα για λόγους ευκρίνειας)	56
Εικόνα 37. Τοποθεσία ΚΠΠΑ (OpenStreetMap contributors, 2017)	63
Εικόνα 38. Σύγκριση στατιστικών μεγεθών ημερήσιων μέσων NO2 (ΚΠΠΑ Aqmesh, Kunak).....	64
Εικόνα 39. Συσχέτιση μετρήσεων O3 μεταξύ κόμβων (ΚΠΠΑ Aqmesh, Kunak).	
Διαγράμματα συσχέτισης και ιστογράμματα (αριστερά), πίνακας συντελεστών συσχέτισης κατά Pearson (δεξιά).....	64
Εικόνα 40. Σύγκριση στατιστικών μεγεθών ημερήσιων μέσων PM2.5 (KASTOM, Kunak, Airly)	65
Εικόνα 41. Συσχέτιση μετρήσεων PM2.5 μεταξύ κόμβων (ΚΠΠΑ KASTOM, Kunak, Airly).	
Διαγράμματα συσχέτισης και ιστογράμματα (αριστερά), πίνακας συντελεστών συσχέτισης κατά Pearson (δεξιά).....	65
Εικόνα 42. Δοκιμή συμπλήρωσης ελλειπουσών τιμών υγρασίας (Kunak).....	66
Εικόνα 43. Δοκιμή συμπλήρωσης ελλειπουσών τιμών O3 (Kunak)	66
Εικόνα 44. Δοκιμή συμπλήρωσης ελλειπουσών τιμών H2S (ΚΠΠΑ Kunak)	66
Εικόνα 45. Δοκιμή συμπλήρωσης ελλειπουσών τιμών NO2 (Aqmesh).....	67
Εικόνα 46. Δοκιμή συμπλήρωσης ελλειπουσών τιμών O3 (ΚΠΠΑ Aqmesh)	67
Εικόνα 47. Ωριαία (πάνω), ημερήσια (μέση) και εβδομαδιαία (κάτω) αναδειγματοληψία ρύπων (ΚΠΠΑ Airly)	68
Εικόνα 48. Ωριαία (πάνω), ημερήσια (μέση) και εβδομαδιαία (κάτω) αναδειγματοληψία ρύπων (ΚΠΠΑ Aqmesh)	69
Εικόνα 49. Ωριαία (πάνω), ημερήσια (μέση) και εβδομαδιαία (κάτω) αναδειγματοληψία ρύπων (ΚΠΠΑ KASTOM)	70

Εικόνα 50. Τυπικά εικοσιτετράωρα (πάνω) και εβδομαδιαία (κάτω) προφίλ (ΚΠΠΑ Airly)	71
Εικόνα 51. Τυπικά εικοσιτετράωρα (πάνω) και εβδομαδιαία (κάτω) προφίλ (ΚΠΠΑ Aqmesh)	72
Εικόνα 52. Τυπικά εικοσιτετράωρα (πάνω) και εβδομαδιαία (κάτω) προφίλ (ΚΠΠΑ KASTOM)	73
Εικόνα 53. Συσχέτιση ωριαίων τιμών (ΚΠΠΑ Airly)	74
Εικόνα 54. Συσχέτιση ωριαίων τιμών (ΚΠΠΑ Aqmesh)	74
Εικόνα 55. Συσχέτιση ωριαίων τιμών (ΚΠΠΑ Kunak)	75
Εικόνα 56. Ιστογράμματα ωριαίου ρυθμού δειγματοληψίας μετρούμενων μεγεθών (ΚΠΠΑ Airly)	76
Εικόνα 57. Violin plot μέσου ημερήσιου ρυθμού δειγματοληψίας ρύπων (ΚΠΠΑ Airly)	76
Εικόνα 58. Ιστογράμματα ωριαίου ρυθμού δειγματοληψίας μετρούμενων μεγεθών (ΚΠΠΑ KASTOM)	77
Εικόνα 59. Violin plot μέσου ημερήσιου ρυθμού δειγματοληψίας ρύπων (ΚΠΠΑ KASTOM)	77
Εικόνα 60. Ιστογράμματα ωριαίου ρυθμού δειγματοληψίας μετρούμενων μεγεθών (ΚΠΠΑ Kunak)	78
Εικόνα 61. Violin plot μέσου ημερήσιου ρυθμού δειγματοληψίας ρύπων (ΚΠΠΑ Kunak)	78
Εικόνα 62. Συνθετικά προσθετικού μοντέλου χρονοσειράς PM10 (ΚΠΠΑ Airly)	79
Εικόνα 63. Συνθετικά προσθετικού μοντέλου χρονοσειράς PM2.5 (ΚΠΠΑ Airly)	79
Εικόνα 64. Συνθετικά προσθετικού μοντέλου χρονοσειράς PM1 (ΚΠΠΑ Airly)	80
Εικόνα 65. Ημερήσιες τάσεις εβδομαδιαίας περιοδικότητας ρύπων (ΚΠΠΑ Airly)	80
Εικόνα 66. Συνθετικά προσθετικού μοντέλου χρονοσειράς NO2 (ΚΠΠΑ Aqmesh)	81
Εικόνα 67. Συνθετικά προσθετικού μοντέλου χρονοσειράς O3 (ΚΠΠΑ Aqmesh)	81
Εικόνα 68. Ημερήσιες τάσεις εβδομαδιαίας περιοδικότητας ρύπων (ΚΠΠΑ Aqmesh)	82
Εικόνα 69. Συνθετικά προσθετικού μοντέλου χρονοσειράς NO2 (ΚΠΠΑ KASTOM)	82
Εικόνα 70. Συνθετικά προσθετικού μοντέλου χρονοσειράς PM2.5 (ΚΠΠΑ KASTOM)	83
Εικόνα 71. Συνθετικά προσθετικού μοντέλου χρονοσειράς CO (ΚΠΠΑ KASTOM)	83
Εικόνα 72. Συνθετικά προσθετικού μοντέλου χρονοσειράς NO2 (ΚΠΠΑ Kunak)	84
Εικόνα 73. Συνθετικά προσθετικού μοντέλου χρονοσειράς O3 (ΚΠΠΑ Kunak)	84
Εικόνα 74. Συνθετικά προσθετικού μοντέλου χρονοσειράς PM2.5 (ΚΠΠΑ Kunak)	85
Εικόνα 75. Μέση τιμή συγκεντρώσεων ρύπων ανά διεύθυνση ανέμου (ΚΠΠΑ Airly)	86
Εικόνα 76. Ροδογράμματα ρύπων ανά διεύθυνση και συχνότητα ανέμου (ΚΠΠΑ Airly)	86
Εικόνα 77. Μέση τιμή συγκεντρώσεων ρύπων ανά διεύθυνση ανέμου (ΚΠΠΑ Aqmesh)	87
Εικόνα 78. Ροδογράμματα ρύπων ανά διεύθυνση και συχνότητα ανέμου (ΚΠΠΑ Aqmesh)	87
Εικόνα 79. Μέση τιμή συγκεντρώσεων ρύπων ανά διεύθυνση ανέμου (ΚΠΠΑ KASTOM)	88
Εικόνα 80. Ροδογράμματα ρύπων ανά διεύθυνση και συχνότητα ανέμου (ΚΠΠΑ KASTOM)	88
Εικόνα 81. Σύγκριση αποτελεσμάτων πρόγνωσης OLSR και πραγματικών δεδομένων (ΚΠΠΑ Kunak και KASTOM)	89
Εικόνα 82. Σύγκριση αποτελεσμάτων πρόγνωσης μοντέλου SARIMAX και πραγματικών δεδομένων (ΚΠΠΑ Kunak και KASTOM)	90
Εικόνα 83. Σύγκριση αποτελεσμάτων πρόγνωσης μοντέλου Random Forest και πραγματικών δεδομένων (ΚΠΠΑ Kunak και KASTOM)	91

Συντομογραφίες - Συμβολισμοί

ΚΠΠΑ	Κόμβος Παρακολούθησης Ποιότητας Αέρα
ΑΧΚ	Αισθητήρας(ες) Χαμηλού Κόστους
ΠΑ	Ποιότητα Αέρα
ΝΟ ₂	Διοξείδιο του Αζώτου
O ₃	Όζον
PM	Particulate Matter (Αιωρούμενα Σωματίδια)
TSP	Συνολικά Αιωρούμενα Σωματίδια
H ₂ S	Υδρόθειο
CO	Μονοξείδιο του Άνθρακα
ppb	Parts per billion
hPa	hectopascals = mbar
RH%	Σχετική Υγρασία %
FRM	Federal Reference Methods
KNN	K-Nearest Neighbors
PCA	Principal Component Analysis
PC	Principal Component
SOM	Self-Organizing Map
ARIMA	Autoregressive Integrated Moving Average
SARIMA	Seasonal Autoregressive Integrated Moving Average
RF	Random Forest
LSTM	Long short-term memory
OLSR	Ordinary Least Squares Regression
TP	Tree Predictors
MAE	Mean Absolute Error
MSE	Mean Square Error
MAPE	Mean Absolute Percentage Error
R ²	R Squared

1. Εισαγωγή

1.1. Επιπτώσεις της αέριας ρύπανσης στο αστικό περιβάλλον

Η αέρια ρύπανση αποτελεί τον σημαντικότερο περιβαλλοντικό παράγοντα επιβάρυνσης της υγείας στην Ευρώπη, με επιβλαβείς επιδράσεις στο καρδιαγγειακό και αναπνευστικό σύστημα, οδηγώντας έως και σε πρόωρη θνησιμότητα. (ΕΕΑ, 2022)

Συγκεκριμένα, για το αστικό περιβάλλον, μελέτες έχουν υποδείξει την σύνδεση μεταξύ μεταβολών στον αριθμό ημερήσιων θανάτων και στα επίπεδα ρύπανσης προηγούμενων ημερών (Makri & Stilianakis, 2008). Οι κάτοικοι, ειδικά, αστικών κέντρων και περιοχών υψηλής κυκλοφοριακής συμφόρησης εκτίθενται σε υψηλά επίπεδα ρύπανσης με αντίστοιχες αρνητικές επιπτώσεις στην υγεία.

Κύριοι αέριοι ρύποι που παρατηρούνται στις αστικές περιοχές είναι:

- Αιωρούμενα σωματίδια (PM) – Θεωρούνται τα διάσπαρτα στην ατμόσφαιρα στερεά και υγρά σωματίδια. Οι σχετικές οδηγίες εστιάζουν στα σωματίδια μέσης αεροδυναμικής διαμέτρου μικρότερης των 10 μμ (PM10), όπως η σκόνη, η αιθάλη κ.α. Εισέρχονται στο κυκλοφορικό σύστημα μέσω της εισπνοής.
- Μονοξείδιο του Άνθρακα (CO) – Δηλητηριώδες για τον άνθρωπο αέριο, άοσμο και άχρωμο. Κύρια ανθρωπογενής πηγή οι εκπομπές μηχανών εσωτερικής καύσης.
- Νιτρικά Οξείδια (NO_x) – Προϊόντα καύσης, αποτελούν την πλειοψηφία των ανθρωπογενών ρύπων στην Ευρώπη, με κύρια πηγή τα οχήματα.
- Πτητικές οργανικές ενώσεις (VOC) – Ενώσεις που εξατμίζονται σε θερμοκρασίες δωματίου. Κύριες πηγές χημικές ουσίες καθαρισμού, μπογιές, καύσιμα κ.α.
- Όζον (O_3) – Επιβλαβές σε επίπεδο εδάφους, δεν εκπέμπεται απευθείας παρά σχηματίζεται ως προϊόν χημικών αντιδράσεων στις οποίες συμμετέχουν νιτρικά οξείδια και πτητικές οργανικές ενώσεις, υπό την επίδραση της ηλιακής ακτινοβολίας.
- Οξείδια του Θείου (SO_x) – Επιβλαβής ρύπος για τον άνθρωπο και το περιβάλλον. Κύρια ανθρωπογενής πηγή η βιομηχανία.

1.2. Εκτίμηση συγκεντρωσης αέριων ρύπων

Οι συγκεντρώσεις των ατμοσφαιρικών ρύπων εκτιμώνται με τη βοήθεια πολύπλοκου, συνήθως μη φορητού εξοπλισμό που υπακούει σε σχετικά πρότυπα όπως οι Ομοσπονδιακές Μέθοδοι Αναφοράς – FRM στις Η.Π.Α. ή σε ισοδύναμα πρότυπα. Τέτοια όργανα είναι τεχνολογίας υπέρυθρων ακτινών μετασχηματισμού Fourier, χρωματογράφου, φασματόμετρου κ.α. Ωστόσο, αυτά τα όργανα αναλυτικής μέτρησης προϋποθέτουν εμπειρία του χειριστή ενώ χαρακτηρίζονται από υψηλό κόστος αγοράς και συντήρησης. Άλλα χαρακτηριστικά τους είναι οι απαιτήσεις του χώρου εγκατάστασης, η αργή απόκριση και ο

όγκος των σχετικών συσκευών. Οι μετρήσεις τους αντιπροσωπεύουν την περιοχή εγκατάστασής τους, αλλά δεν δύνανται να αποδώσουν την ποιότητα αέρα στο μικροπεριβάλλον εντός του οποίου ζει και δραστηριοποιείται ο κάτοικος μιας πόλης. Συγκεκριμένα, η ατμοσφαιρική ρύπανση παρουσιάζει υπερτοπικές διακυμάνσεις, ιδίως κοντά στις πηγές εκπομπών. Εκτός από τα μοτίβα δραστηριότητας των ανθρώπων, αυτή η διακύμανση είναι ενδεχομένως το πλέον σημαντικό στοιχείο που καθορίζει την έκθεση σε σχετικά επίπεδα ρύπων. Η έκθεση αυτή εξαρτάται από τη χρονική δραστηριότητα και τη διαδρομή εντός του αστικού ιστού που ακολουθεί ένας κάτοικος, με συγκεκριμένες συμπεριφορές όπως ο τρόπος μετακίνησης και ο χρόνος παραμονής κοντά σε δρόμο ή σε πάρκο να παίζουν καθοριστικό ρόλο (Frederickson, et al., 2022).

Για τους λόγους αυτούς, πιο πρακτικές, συμπαγείς, ευέλικτες και χαμηλότερου κόστους συσκευές, τα ονομαζόμενα όργανα χαμηλού ή αποδεκτού κόστους, που αποτελούν Κόμβους Παρακολούθησης Ποιότητας Αέρα (ΚΠΠΑ). Πρόκειται για συσκευές κόστους έως 2500 Ευρώ σύμφωνα με το US EPA Air Sensor Toolbox, (Williams, et al., 2014) ή έως 5000 Ευρώ (Castell, et al., 2017), που αποτελούν μία εναλλακτική λύση, μια και μπορούν να αναπτυχθούν σε υψηλή πυκνότητα, και να αποδώσουν ακριβέστερα τη χωρική (αλλά και χρονική) διακύμανση των επιπέδων ρύπων στα οποία εκτίθεται ο κάτοικος μιας πόλης. Οι μέθοδοι μέτρησης στις περιπτώσεις αυτές ακολουθούν πρότυπα που έχουν αναπτυχθεί για την συγκεκριμένη κατηγορία οργάνων, όπως το IEEE 1451 (Kumar, et al., 2015). Σε κάθε περίπτωση, πρόκειται για ηλεκτρονικούς αισθητήρες οι οποίοι βελτιώθηκαν σημαντικά με την εξέλιξη της τεχνολογίας, οι οποίοι μετατρέπουν ένα φυσικό μέγεθος (ύπαρξη αέριου ρύπου) σε ένα μέγεθος (συνήθως ηλεκτρικό) του οποίου η τιμή συναρτάται με τη συγκέντρωση του πρώτου. Συνήθεις τύποι τέτοιων οργάνων είναι ηλεκτροχημικοί, υπέρυθρων ακτινών, καταλυτικών σφαιριδίων, φωτοϊονισμού, οπτικής ανίχνευσης και σταθερής κατάστασης (solid-state). Στον Πίνακας 1 συγκρίνονται τα χαρακτηριστικά των αναλυτικών οργάνων και των αισθητήρων χαμηλού κόστους (ΑΧΚ).

Πρόσφατα έχουν αναπτυχθεί αισθητήρες τύπου ημιαγωγού, ικανοί για τον εντοπισμό πολλαπλών αέριων ρύπων και πλεονεκτούν σε σχέση με τους καταλυτικούς και ηλεκτροχημικούς ως προς το χρόνο ζωής, το συμπαγές μέγεθος και το κόστος (Kularatna & Sudantha, 2008).

Πίνακας 1. Σύγκριση χαρακτηριστικών αναλυτικών οργάνων και αισθητήρων χαμηλού κόστους

Χαρακτηριστικό	Αναλυτικά ΟΡΓΑΝΑ	ΑΧΚ
Ανάλυση	Βέλτιστη	Συγκρίσιμη
Κόστος	Πολύ Υψηλό	Λογικό
Μέγεθος	Ογκώδες	Συμπαγές
Στιβαρότητα	Ευαίσθητα	Στιβαρά
Αντίδραση	Αργή	Γρήγορη
Μαζική Παραγωγή	Δύσκολη	Εφικτή

1.3. Ερευνητικά ερωτήματα

Στη βάση των προαναφερθέντων, η παρούσα μελέτη αποσκοπεί να δώσει απαντήσεις στα ακόλουθα ερευνητικά ερωτήματα:

- Ποια τα βασικά χαρακτηριστικά, τα προφίλ, οι ομοιότητες και οι διαφορές των επιπέδων συγκέντρωσης ρύπων που προκύπτουν από όργανα χαμηλού κόστους τα οποία λειτουργούν σε αστικό περιβάλλον όπως αυτό της Θεσσαλονίκης;

Για να απαντηθεί το ερώτημα αυτό χρησιμοποιήθηκαν τεχνικές στατιστικής και μηχανικής μάθησης και εφαρμόστηκαν μέθοδοι οπτικοποίησης σε χώρο αντιληπτών διαστάσεων (dimensionality reduction visualization) και συσταδοποίησης (clustering), με στόχο την αναγνώριση προτύπων. Πραγματοποιήθηκε σύγκριση μεταξύ διαθέσιμων μετρήσεων ΚΠΠΑ. Παρουσιάστηκαν σε διαγράμματα, για κάθε κόμβο, η εξέλιξη στον άξονα του χρόνου, η κατανομή και τα στατιστικά χαρακτηριστικά των μετρήσεων.

- Ποια η εικόνα ποιότητας αέρα που προκύπτει από ΚΠΠΑ σε περιοχή του κέντρου της Θεσσαλονίκης που δύναται να αντιπροσωπεύει την περιβαλλοντική φόρτιση σε σχέση με την ποιότητα αέρα στην Πανεπιστημιούπολη του ΑΠΘ;

Για να απαντηθεί το ερώτημα αυτό πραγματοποιήθηκε ανάλυση των διαθέσιμων δεδομένων μετρήσεων. Συγκεκριμένα, για το χρονικό διάστημα δειγματοληψίας, προσδιορίστηκε η εξέλιξη των επιπέδων του εκάστοτε ρύπου, τα προφίλ ποιότητας αέρα (ΠΑ) σε εβδομαδιαία και ημερήσια βάση, οι πίνακες συσχέτισης μεταξύ των μεγεθών, ενώ έγινε στατιστική περιγραφή και «αποσύνθεση χρονοσειράς» με στόχο τον εντοπισμό περιοδικότητας, τάσης και θορύβου και μελετήθηκε η επίδραση του ανέμου.

- Εστιάζοντας σε ρύπους που έχουν ενοχοποιηθεί για προβλήματα οσμών στο πολεοδομικό συγκρότημα, ποια εικόνα προκύπτει σε σχέση με τα επίπεδα, την ημερήσια διακύμανση και την εξάρτηση από μετεωρολογικές συνθήκες ως προς το υδρόθειο;

Εκτός των κύριων αέριων ρύπων που αναφέρθηκαν, ιδιαίτερη έμφαση δόθηκε στην μελέτη του υδρόθειου (H2S), το οποίο ενώ δεν προσεγγίζει επίπεδα τοξικότητας, προκαλεί επεισόδια δυσοσμίας με αντίστοιχες διαμαρτυρίες των κατοίκων της πόλης.

- Μπορεί η χρήση μεθόδων μηχανικής μάθησης να συνεισφέρει στο να συμπληρώνει στοιχεία μετρήσεων που λείπουν, ή να κάνει πρόγνωση μελλοντικών επιπέδων ρύπανσης, σε όργανο χαμηλού κόστους;

Επιστρατεύτηκαν μέθοδοι μηχανικής μάθησης προς συμπλήρωση των ελλειπουσών τιμών ώστε να αποκατασταθεί η συνέχεια της χρονοσειράς δεδομένων. Κατόπιν, λαμβάνοντας υπόψιν τα συμπεράσματα που προέκυψαν από την στατιστική ανάλυση καταστρώθηκαν μοντέλα, στατιστικά και τεχνητής νοημοσύνης, βραχυπρόθεσμης πρόβλεψης τιμών αέριων ρύπων και διερευνήθηκε η αποτελεσματικότητά τους.

Τελικά παρουσιάστηκαν τα συνολικά αποτελέσματα σε γραφήματα και πίνακες και εξήχθησαν τα σχετικά συμπεράσματα.

2. Υλικά και Μέθοδοι

2.1. Υλικά

Η εκτίμηση της συγκέντρωσης αέριων ρύπων και μετεωρολογικών μεγεθών πραγματοποιήθηκε από ΚΠΠΑ αισθητήρων χαμηλού κόστους και μετεωρολογικό σταθμό, οι οποίοι βρίσκονται εγκατεστημένοι στην οροφή του κτιρίου Ε14 της Πολυτεχνικής Σχολής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, όπως διακρίνεται στην Εικόνα 37 του Παραρτήματος. Ο κατασκευαστής και τα μετρούμενα μεγέθη κάθε κόμβου παρουσιάζονται στον Πίνακας 2.

Πίνακας 2. Κατασκευαστές και μετρούμενα μεγέθη ΚΠΠΑ

Κατασκευαστής	Μετρούμενοι ρύποι	Μετεωρολογικά μεγέθη
Kunak (Kunak Technologies S.L., 2022)	PM _{1-2.5-10} ($\mu\text{g}/\text{m}^3$) H ₂ S (ppb) NO ₂ (ppb) O ₃ (ppb)	Dew Point (°C) Relative Humidity (RH%) Pressure (hPa) Temperature (°C)
Aqmesh (AQMesh Environmental Instruments Ltd, 2022)	NO (ppb) NO ₂ (ppb) O ₃ (ppb)	Relative Humidity (RH%) Pressure (mbar) Temperature (°C)
KASTOM (KASTOM, 2022)	PM _{1-2.5-10} ($\mu\text{g}/\text{m}^3$) NO ₂ (ppb) O ₃ (ppb) CO ($\mu\text{g}/\text{m}^3$)	Relative Humidity (RH%) Pressure (mbar) Temperature (°C)
Airly (Airly Sp. Z o.o., 2022)	PM _{1-2.5-10} ($\mu\text{g}/\text{m}^3$)	Relative Humidity (RH%) Pressure (hPa) Temperature (°C)
Davis (Davis Instruments Corporation, 2022)		Ταχύτητα Ανέμου (m/s) Διεύθυνση Ανέμου (°)

2.1.1. Αισθητήρες

Οι υπό μελέτη ΚΠΠΑ βασίζονται κυρίως σε τεχνολογίες ηλεκτροχημικών και οπτικών αισθητήρων για την εκτίμηση της συγκέντρωσης αέριων ρύπων και αιωρούμενων σωματιδίων αντίστοιχα ενώ για τις μετεωρολογικές συνθήκες χρησιμοποιούνται αισθητήρες στερεάς κατάστασης (solid-state).

Ηλεκτροχημικοί αισθητήρες

Πρόκειται για ευρέως δοκιμασμένη τεχνολογία που περιλαμβάνει ηλεκτρόδια και ηλεκτρολύτη. Το αέριο προκαλεί ιονισμό και μεταφορά φορτίου μεταξύ των ηλεκτροδίων. Λαμβάνεται, ανάλογα με την τύπο του αισθητήρα, είτε η διαφορά δυναμικού (potentiometric) είτε η ροή ρεύματος (amperometric) που δημιουργείται και αφού ενισχυθεί το σήμα μεταφράζεται στην μέτρηση της συγκέντρωσης του αέριου ρύπου.

Οπτικοί αισθητήρες

Για τον υπολογισμό συγκεντρώσεων αιωρούμενων σωματιδίων χρησιμοποιούνται αισθητήρες οπτικής ανίχνευσης. Λειτουργούν με ακτίνες (laser ή infrared led) οι οποίες αντανακλώνται όταν προσπίπτουν σε σωματίδια και καταλήγουν σε ανιχνευτή (φωτοδιόδιο) όπου υπολογίζεται το πλήθος των σωματιδίων και εκτιμάται η συγκέντρωσή τους βάσει πυκνότητας.

Αισθητήρες στερεάς κατάστασης

Οι μετεωρολογικές παράμετροι εκτιμώνται μέσω αισθητήρων στερεάς κατάστασης. Το μέγεθος προς εκτίμηση προκαλεί αλλαγές στις ηλεκτρικές ιδιότητες του αισθητήρα και υπολογίζεται η αντίσταση/αγωγιμότητα του υλικού, η οποία μεταφράζεται στη μετρούμενη ποσότητα.

Η ταχύτητα και διεύθυνση του ανέμου μετρούνται με ανεμόμετρο κυπέλλων και ανεμοδείκτη αντίστοιχα. Οι περιστροφές του ανεμόμετρου καταγράφονται και λαμβάνεται η μέση τιμή της ταχύτητας ανέμου ανά 3 δευτερόλεπτα ενώ για την διεύθυνση λαμβάνεται το μέσο διάνυσμα για το ίδιο χρονικό διάστημα. (Davis Instruments Corporation, 2022)

2.1.2. Δεδομένα

Τα δεδομένα των μετρήσεων ελήφθησαν από τις αντίστοιχες διαδικτυακές βάσεις δεδομένων των κατασκευαστών, κάθε ΚΠΠΑ, για το χρονικό διάστημα 10 Απριλίου 2022 έως 5 Σεπτεμβρίου 2022. Ο ρυθμός δειγματοληψίας είναι ωριαίος και η μορφή των δεδομένων κατόπιν της αρχικής επεξεργασίας παρουσιάζεται ενδεικτικά στον Πίνακας 3.

Πίνακας 3. Ενδεικτική μορφή δεδομένων ΚΠΠΑ Kunak

Datetime	Dew Point (C)	H2S (ppb)	NO2 (ppb)	Relative Humidity (%)	O3 (ppb)	PM1 ($\mu\text{g}/\text{m}^3$)	PM2.5 ($\mu\text{g}/\text{m}^3$)	PM10 ($\mu\text{g}/\text{m}^3$)	Temp (C)
2022-04-10 00:00:00	9.74	16.6	60.03	42.0	0.0	6.61	8.56	12.61	17.6
2022-04-10 01:00:00	9.55	11.56	59.0	38.98	0.0	6.22	8.15	12.43	17.67
2022-04-10 02:00:00	9.39	13.33	59.05	37.07	0.0	6.26	8.22	12.04	17.49
2022-04-10 03:00:00	8.08	1.31	58.41	29.14	4.68	4.69	6.38	8.87	16.26
2022-04-10 04:00:00	7.46	0.33	58.28	29.25	2.04	3.49	4.99	7.16	15.64
2022-04-10 05:00:00	7.2	2.11	58.07	33.18	0.27	3.72	5.28	8.31	15.41
2022-04-10 06:00:00	7.16	0.84	56.61	34.14	0.0	4.14	5.75	8.57	15.76
2022-04-10 07:00:00	4.55	0.0	45.42	14.43	17.89	4.02	5.36	7.95	16.44
2022-04-10 08:00:00	3.91	0.0	40.1	12.32	22.9	3.04	4.21	6.6	17.65
2022-04-10 09:00:00	4.99	0.01	41.05	22.93	11.0	3.7	5.05	7.68	18.48

2.1.2.1. Σύγκριση ΚΠΠΑ

Εντοπίστηκαν οι κοινοί ρύποι μεταξύ των διαφορετικών ΚΠΠΑ και εφαρμόστηκαν οι μεθοδολογίες γραφικής απεικόνισης και στατιστικής περιγραφής και συσχέτισης που περιγράφονται στην ενότητα των μεθόδων (κεφ. 2). Κατασκευάστηκαν συγκριτικά διαγράμματα ανά ρύπο και εξάχθηκαν τα αποτελέσματα της σύγκρισης.

2.1.3. Υλοποίηση μελέτης

Το σύνολο της ανάλυσης δεδομένων, της μοντελοποίησης και της παρουσίασης (γραφήματα) υλοποιήθηκε σε προγραμματιστικό περιβάλλον JupyterLab (Kluyver, et al., 2016), ανοιχτού κώδικα Python (Python Software Foundation, 2022) με χρήση των βιβλιοθηκών:

- Pandas (McKinney, 2010)
- Numpy (Harris, Millman, van, & S.J., 2020)
- Matplotlib (Hunter, 2007)
- Seaborn (Waskom, 2021)
- Statsmodels (Seabold & Perktold, 2010)
- Windrose (Roubeyrie & Celles, 2018)
- Scipy (Virtanen, et al., 2020)
- Scikit-learn (Pedregosa, et al., 2011)
- MiniSom (Vettigli, 2018)
- PyCaret (Ali M. , 2020)

2.2. Μέθοδοι

2.2.1. Ανάλυση Δεδομένων

Η ανάλυση δεδομένων περιλαμβάνει παρατήρηση, εκκαθάριση, μετασχηματισμό, μορφοποίηση και μοντελοποίηση των δεδομένων με στόχο την εξαγωγή χρήσιμων συμπερασμάτων.

2.2.1.1. Προεπεξεργασία – Συμπλήρωση ελλειπουσών τιμών

Λόγω σφαλμάτων ή διακοπών κατά τη λειτουργία των αισθητήρων προκύπτουν ελλείψεις στο σύνολο των μετρήσεων. Το μέγιστο ποσοστό ελλειπουσών τιμών ανά μετρούμενο μέγεθος για το παρόν δείγμα υπολογίστηκε στο 3%.

Οι ελλείπουσες τιμές σε χρονοσειρές μπορούν γενικά να συμπληρωθούν με μεθόδους παρεμβολής (interpolation), καταλογισμού (imputation) και γραμμικού μετασχηματισμού. Οι τεχνικές παρεμβολής αξιοποιούνται σε μονομεταβλητά σύνολα δεδομένων ενώ οι τεχνικές καταλογισμού έχει παρατηρηθεί ότι συνήθως αποδίδουν καλύτερα σε δεδομένα πολλαπλών μεταβλητών (Samal, Panda, Babu, & Das, 2021).

Στο πλαίσιο της παρούσας εργασίας εξετάστηκαν τρεις μέθοδοι. Γραμμική παρεμβολή, παρεμβολή τμηματικών κυβικών πολυωνυμικών συναρτήσεων (Cubic Spline) και μία μέθοδος καταλογισμού μηχανικής μάθησης, αυτή του αλγορίθμου πλησιέστερης γειτνίασης (K-Nearest Neighbors).

Η μέθοδος παρεμβολής Cubic Spline αποτελεί υποσύνολο της παρεμβολής Spline. Αντί της προσέγγισης με χρήση ενός υψηλόβαθμου πολυωνύμου για όλες τις τιμές, υπολογίζονται επί μέρους, κυβικά πολυώνυμα, σε υποσύνολα των τιμών. Τα κυβικά πολυώνυμα ορίζονται κατά τρόπο ώστε τα προσκείμενα σημεία τιμών να συμφωνούν μεταξύ τους ως προς κλίση και καμπυλότητα στο σημείο σύνδεσης (Dyer & Dyer, 2001).

Ο αλγόριθμος KNN (Fix & Hodges, 1951) διαχειρίζεται μία σειρά K μεταβλητών ως συντεταγμένες σε ένα χώρο K διαστάσεων και λαμβάνει τις ελλείπουσες τιμές σύμφωνα με την πλησιέστερη γειτονική σειρά σε αυτόν τον χώρο, όταν είναι διαθέσιμες. Ταυτόχρονα δίδεται βαρύτητα στην ευκλείδεια απόσταση, ανάλογη του ποσοστού ελλείψεων στη συγκεκριμένη γειτονική σειρά. Κατ' αυτόν τον τρόπο σειρές με περισσότερες ελλείψεις συμμετέχουν λιγότερο στη διαδικασία ώστε να αντισταθμιστεί η χαμηλή τους αξιοπιστία (Junninen, Niska, & Tuppurainenc, 2004).

Πραγματοποιήθηκαν δοκιμές σε πολλαπλά μεγέθη ρύπων και μετεωρολογικών παραμέτρων. Επιλέχθηκαν πλήρη τμήματα χρονοσειρών και διαγράφηκαν τιμές τους ανά τακτά διαστήματα (testing sets). Το τεχνητό ποσοστό έλλειψης διατηρήθηκε όμοιο με αυτό των χρονοσειρών πραγματικών ελλείψεων. Κατόπιν συμπλήρωσης τιμών και σύγκρισης μεταξύ συμπληρωμένων και πραγματικών σετ, αξιολογήθηκε η επίδοση των μεθόδων με εκτιμητές σφαλμάτων όπως περιγράφεται στην [ενότητα 2.2.2.4](#).

2.2.1.2. Αναδειγματοληψία

Εκκινώντας με βάση τις ωριαίες τιμές του αρχικού δείγματος πραγματοποιήθηκε αναδειγματοληψία, σε ημερήσια και εβδομαδιαία βάση μέσων όρων, με στόχο την αντίστοιχη απεικόνιση των επιπέδων των μεγεθών στο συνολικό χρονικό διάστημα.

2.2.1.3. Τυπικά προφίλ

Ομαδοποιώντας ανά ώρα της ημέρας, ανά ημέρα της εβδομάδας και υπολογίζοντας τον μέσο όρο για κάθε ομάδα δεδομένων, δημιουργήθηκαν τα αντίστοιχα τυπικά εικοσιτετράωρα και εβδομαδιαία προφίλ ρύπων.

2.2.1.4. Διερεύνηση συσχέτισης

Λαμβάνοντας τα δεδομένα ωριαίας και ημερήσιας βάσης διερευνήθηκε η συσχέτιση μεταξύ των μετρούμενων μεγεθών, ρύπων και μετεωρολογικών, κατά την μέθοδο Pearson (Freedman, Pisani, & Purves, 2007). Στη συνέχεια παρήχθη ο αντίστοιχος πίνακας συσχέτισης.

Θεωρώντας δεδομένα ζευγαρωτών μεταβλητών $\{(x_1, y_1), \dots, (x_n, y_n)\}$ πλήθους n , ο συντελεστής συσχέτισης δείγματος Pearson ορίζεται:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

όπου:

- n : το πλήθος δείγματος
- x_i, y_i : τα σημεία του κάθε ζεύγους
- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$: ο δειγματικός μέσος του x , ομοίως και για \bar{y} .

2.2.1.5. Περιγραφική Στατιστική

Για κάθε μεταβλητή της χρονοσειράς υπολογίστηκαν τα στατιστικά μεγέθη:

- Πλήθος (count)
- Μέση τιμή (mean)
- Τυπική απόκλιση (std)
- Ελάχιστο (min)
- 25% -50% - 75%
- Μέγιστο (max)
- Κύρτωση (kurtosis)
- Λοξότητα (skew)

Στη συνέχεια απεικονίστηκαν σε θηκόγραμμα (και σε violin plot).

2.2.1.6. Αποσύνθεση χρονοσειράς

Η χρονοσειρά, εκτιμώντας προσθετικό μοντέλο, μπορεί να αναλυθεί ως εξής:

$$y_t = T_t + S_t + R_t \quad (2)$$

Όπου:

- y_t : Η παρατηρούμενη χρονοσειρά
- T_t : Η τάση (trend). Εκτιμάται με την μέθοδο των κινούμενων μέσων όρων.
- S_t : Η εποχικότητα (seasonality). Υπολογίζεται ως ο μέσος όρος όλων των ακολουθιών εποχικότητας αφού αφαιρεθεί η τάση από το αρχικό σήμα.
- R_t : Το υπόλοιπο ή αλλιώς θόρυβος (residual). Προκύπτει ως το υπόλοιπο της αφαίρεσης της εποχικότητας και της τάσης από το αρχικό σήμα.

(Dudek, 2015)

Η ύπαρξη και ο προσδιορισμός περιοδικότητας-εποχικότητας επαληθεύτηκε με μεθόδους Γρήγορου Μετασχηματισμού Fourier (FFT) (Brigham, 1988) και αυτοσυσχέτισης (Autocorrelation) (Flores, Engel, & Pinto, 2012).

2.2.1.7. Κανονικοποίηση

Αρκετοί αλγόριθμοι μηχανικής μάθησης αποδίδουν ακριβέστερα έπειτα από κανονικοποίηση των δεδομένων, συμπεριλαμβανομένων των αλγορίθμων που αξιοποιούν την Ευκλείδεια απόσταση, μεταξύ σημείων δεδομένων, όπως αυτοί που επιστρατεύονται στην παρούσα μελέτη. Ιδιαίτερα για δεδομένα, όπου οι μονάδες των μεταβλητών διαφέρουν μεταξύ τους και τα μεγέθη δεν είναι άμεσα συγκρίσιμα, η διαδικασία κρίνεται απαραίτητη πριν την μοντελοποίηση.

Η μέθοδος κανονικοποίησης που χρησιμοποιείται ακολουθεί την σχέση:

$$y = \frac{x - \min}{\max - \min} \quad (3)$$

Όπου:

- y : η κανονικοποιημένη τιμή
- x : η αρχική τιμή
- \min : η ελάχιστη παρατηρούμενη τιμή της μεταβλητής
- \max : μέγιστη παρατηρούμενη τιμή της μεταβλητής

Έτσι οι νέες τιμές που προκύπτουν για κάθε μεταβλητή βρίσκονται στο διάστημα μεταξύ 0 και 1. Η μέθοδος χρησιμοποιήθηκε ως προεπεξεργασία για την εφαρμογή των αλγορίθμων SOM και K-Means.

2.2.1.8. Συνεισφορά ανέμου

Προσαρμόστηκαν τα δεδομένα ταχύτητας και διεύθυνσης ανέμου στις χρονοσειρές μετρήσεων των ρύπων. Κατόπιν κατασκευάστηκαν ραβδογράμματα απεικόνισης των επιπέδων ρύπων ανά διεύθυνση του ανέμου και ροδογράμματα όπου συνδυάστηκαν τα παραπάνω με τη συχνότητα εμφάνισης κάθε διεύθυνσης. Διαχωρίστηκαν χρωματικά τα επίπεδα ρύπανσης, σύμφωνα με τα όρια που έχουν τεθεί από τη σχετική ευρωπαϊκή οδηγία. (DIRECTIVE 2008/50/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 21 May 2008 on ambient air quality and cleaner air for Europe)

2.2.1.9. Οπτικοποίηση – Συσταδοποίηση

Η οπτικοποίηση δεδομένων αποσκοπεί στη «μετάφραση» των δεδομένων σε μια μορφή εύκολα αντιληπτή από τον ανθρώπινο εγκέφαλο. Όταν πρόκειται για δεδομένα πολλών μεταβλητών, συχνά επιστρατεύονται τεχνικές μείωσης διαστάσεων σε δύο ή τρεις, ώστε αυτά να μπορούν να απεικονιστούν ευχερέστερα. Κατ' αυτόν τον τρόπο γίνεται, αρχικά, εφικτός ο εντοπισμός μοτίβων, τάσεων και ακραίων τιμών σε σύνθετα σετ δεδομένων.

Κατά την συσταδοποίηση τα δεδομένα διαχωρίζονται σε ομάδες (clusters) ώστε τα αντικείμενα που ανήκουν στην ίδια ομάδα να παρουσιάζουν μεγαλύτερη ομοιότητα σε σχέση με αυτά άλλης ομάδας.

Κατόπιν της επεξεργασίας και μοντελοποίησης επαναλαμβάνεται η διαδικασία της οπτικοποίησης ώστε να εξαχθούν τα σχετικά συμπεράσματα.

Self-Organizing Map (SOM)

Η τεχνική των αυτό-οργανούμενων απεικονίσεων SOM (Kohonen, 1990) αποτελεί μια μη επιβλεπόμενη μέθοδο μάθησης, που μπορεί να χρησιμοποιηθεί για την εύρεση συσχετίσεων, τη μείωση διαστάσεων και εν δυνάμει την κατηγοριοποίηση δεδομένων πολλαπλών μεταβλητών. Στόχος είναι η συμπίεση των πληροφοριών διατηρώντας παράλληλα τις τοπολογικές σχέσεις μεταξύ των δεδομένων. Αυτό επιτυγχάνεται συνδυάζοντας μη γραμμική προβολή μέσω μιας τοπολογίας τεχνητού νευρωνικού δικτύου, και συσταδοποίηση, παράγοντας έναν «χάρτη» όπου αναδεικνύεται η αλληλεξάρτηση μεταξύ των μετασχηματισμένων δεδομένων που δημιουργήθηκαν.

Ο αλγόριθμος SOM έχει εφαρμοστεί σε περιβαλλοντικές εφαρμογές, κυρίως για χωρικό εντοπισμό πηγών ρύπανσης, για μείωση διαστάσεων (Katsifarakis & Karatzas, 2017), για προσέγγιση συσταδοποίησης προφίλ πολλαπλών ρύπων (Pearce, et al., 2014) κ.α.

Στην συγκεκριμένη μελέτη χρησιμοποιήθηκε με σκοπό την μείωση διαστάσεων και τον εντοπισμό ημερών παρόμοιου προφίλ ρύπανσης.

K-Means

Ο πλέον καθιερωμένος στη συσταδοποίηση αλγόριθμος, K-Means (Lloyd & P., 1982), διαχωρίζει τα δεδομένα σε k πλήθος συστάδων (clusters). Συνήθως προηγείται η μέθοδος PCA όταν παρατηρούνται πολλαπλές μεταβλητές.

Αρχικά ορίζονται τυχαία k «μέσοι» ως κέντρα (centroids). Στη συνέχεια μέσω επαναληπτικής διαδικασίας σχηματίζονται συστάδες, κατά πλησιέστερη Ευκλείδεια απόσταση κάθε σημείου δεδομένων από τα κέντρα. Σε κάθε επανάληψη ανανεώνονται τα κέντρα, διατηρώντας το πλήθος k, μέσω εκ νέου υπολογισμού των μέσων για κάθε συστάδα. Η διαδικασία τερματίζεται όταν επιτευχθεί το κριτήριο σύγκλισης που έχει οριστεί (συνήθως η ευστάθεια στη θέση των κέντρων και η μη «ανάθεση» σημείων μεταξύ διαφορετικών συστάδων).

Η επιλογή του πλήθους k βασίζεται στο περιεχόμενο του προβλήματος καθώς και στο βέλτιστο αριθμό συστάδων ώστε να προσδίδουν την μέγιστη δυνατή πληροφορία συνυπολογίζοντας τη φθίνουσα απόδοση (diminishing returns). Το κριτήριο εκφράζεται μέσω του διαγράμματος Elbow Method (Marutho, Handaka, Wijaya, & Muljono, 2018).

Ανάλυση Κυρίων Συνιστωσών (PCA)

Κατά την μέθοδο των πρωτευουσών συνιστωσών (Principal Component Analysis (Pearson, 1901)), τα πολυμεταβλητά δεδομένα εκφράζονται μέσω επαναπροσδιορισμού των συντεταγμένων σε ένα νέο, ορθοκανονικό σύστημα, κατάλληλο για την εκάστοτε εφαρμογή. Το νέο σύστημα συντεταγμένων προκύπτει ως γραμμικός συνδυασμός των αρχικών πολλαπλών μεταβλητών, με στόχο τον εντοπισμό των κύριων συνιστωσών (principal components), εκείνων δηλαδή που εμπεριέχουν το μεγαλύτερο ποσοστό της αρχικής πληροφορίας (ποσοστό της συνολικής μεταβλητότητας των δεδομένων). Έτσι, είναι δυνατόν να αποδοθούν γραφικά τα κύρια συστατικά σε αντιληπτές διαστάσεις, διατηρώντας την μέγιστη δυνατή πληροφορία του αρχικού συστήματος.

Συγκεκριμένα η μέθοδος PCA χρησιμοποιήθηκε κατά το στάδιο επεξεργασίας των δεδομένων πριν την εφαρμογή των αλγορίθμων SOM και K-Means. Κατ' αυτόν τον τρόπο ο SOM δεν εκκινεί τυχαία αλλά βασίζεται στις κύριες συνιστώσες, ο K-Means αποδίδει με υψηλότερη ακρίβεια λόγω μείωσης θορύβου ενώ και οι δύο αλγόριθμοι εκτελούνται ταχύτερα λόγω μείωσης των διαστάσεων.

2.2.2. Μοντελοποίηση δεδομένων πρόγνωσης

Η μοντελοποίηση χρονοσειρών γενικότερα μπορεί να αντιμετωπιστεί με:

- Κλασικά / στατιστικά μοντέλα όπως: Κινούμενοι μέσοι όροι, εκθετική εξομάλυνση, ARIMA, SARIMA κ.α.
- Μοντέλα υπολογιστικής νοημοσύνης όπως: Γραμμική παλινδρόμηση, Random Forest, νευρωνικά δίκτυα όπως LSTM κ.α.

Στην μελέτη που πραγματοποιήθηκε αναπτύχθηκαν τρία μοντέλα:

1. Εκτίμησης μέσω Απλής Παλινδρόμησης Ελαχίστων Τετραγώνων (OLSR).
2. Παλινδρόμηση Ελαχίστων Τετραγώνων με διαχείριση σφαλμάτων μέσω εφαρμογής SARIMA (SARIMAX).
3. Μοντέλο μηχανικής μάθησης «τυχαίου δάσους» (Random Forest Regressor).

Επιλέχθηκε ένα τμήμα των δεδομένων για την κατάστρωση-εκπαίδευση των μοντέλων (training) ενώ το υπόλοιπο για την αξιολόγηση ευστοχίας της μοντελοποίησης (testing).

Οι τεχνικές μοντελοποίησης αφορούν πρόγνωση ωριαίων τιμών και δοκιμάστηκαν για διαφορετικές περιόδους κατά τη διάρκεια της δειγματοληψίας ώστε να διερευνηθεί η αποτελεσματικότητα τους υπό διαφορετικές συνθήκες. Επιλέχθηκε κυλιόμενο παράθυρο δείγματος τεσσάρων εβδομάδων με στόχο την πρόβλεψη του επόμενου εικοσιτετράωρου και καταγράφηκαν τα κριτήρια ευστοχίας. Κατόπιν, τέθηκαν ως στόχος πρόβλεψης τα τρία τελευταία εικοσιτετράωρα της δειγματοληψίας. Οι μέθοδοι υλοποιήθηκαν επαναληπτικά, αυξάνοντας σταδιακά το μέγεθος δείγματος κατά μία εβδομάδα σε κάθε επανάληψη, και διερευνήθηκε η μεταβολή της ευστοχίας πρόβλεψης.

2.2.2.1. Παλινδρόμηση Ελαχίστων Τετραγώνων

Η γενική μορφή της απλής μεθόδου παλινδρόμησης ελαχίστων τετραγώνων εφαρμόστηκε για ωριαία πρόγνωση ρύπων, αγνοώντας αρχικά συμπεριφορές τάσεων, εποχικότητας και αυτοσυσχέτισης. Όπως έχει αποδειχθεί (Akpan & Moffat, 2018) τα υπολειπόμενα στοιχεία της παλινδρόμησης (residuals) κατά την εφαρμογή σε χρονοσειρές, χαρακτηριζόμενες από τα παραπάνω, είναι συχνά αυτοσυσχετιζόμενα, οδηγώντας σε μειωμένη ευστοχία πρόγνωσης.

2.2.2.2. SARIMAX

Στην [ενότητα 2.2.1.6](#) διερευνήθηκε η περιοδικότητα των δεδομένων. Όταν παρατηρούνται κυκλικά φαινόμενα, μοντέλα τύπου SARIMA θεωρούνται πιο επαρκή κατά την βραχυπρόθεσμη κυρίως πρόγνωση (Gacheva-Ilieva, et al., 2013).

Ο όρος ARIMA (Box & Jenkins, 1976) λαμβάνει υπόψιν τα χαρακτηριστικά:

- AR: Autoregression (αυτοπαλινδρόμηση)
- I: Integrated (διαφόριση)
- MA: Moving Average (κινούμενοι μέσοι)

Το μοντέλο SARIMA αποτελεί εξέλιξη του ARIMA ώστε να συνυπολογίζει την επίδραση της εποχικότητας (S: seasonality).

Η γενική μορφή του μοντέλου SARIMA βασίζεται στις μη αρνητικές παραμέτρους:

- p , d , q : οι παράμετροι που καθορίζουν την αυτοπαλινδρόμηση (AR), την διαφόριση (I) και τον βαθμό κινούμενων μέσων (MA) αντίστοιχα.

- P, D, Q, m: οι παράμετροι που περιγράφουν τα παραπάνω μεγέθη αλλά για την εποχικότητα που χαρακτηρίζει την χρονοσειρά. Ο όρος m εκφράζει το μέγεθος της περιόδου, ανάλογα με το ρυθμό δειγματοληψίας. Π.χ. για ωριαίο ρυθμό εικοσιτετράωρης περιόδου όπως στα δεδομένα της μελέτης, m=24.

Οι παραπάνω παράμετροι εντοπίστηκαν κατά τη διαδικασία της αποσύνθεσης χρονοσειράς, προσδιορίστηκαν κατά τη διερεύνηση της αυτοσυσχέτισης και εποχικότητας ([ενότητα 2.2.1.6](#)) και κατόπιν υλοποιήθηκε η πρόγνωση.

Το X στην κατάληξη της ονομασίας SARIMAX εκφράζει την εφαρμογή του μοντέλου SARIMA στην διαχείριση των σφαλμάτων που προκύπτουν από την απλή παλινδρόμηση ελαχίστων τετραγώνων (OLSR).

2.2.2.3. Random Forest Regressor

Σε περιπτώσεις όπου τα δεδομένα παρουσιάζουν υψηλή μεταβλητότητα έχουν εφαρμοστεί πιο ισχυρά, μη παραμετροποιημένα και υβριδικά μοντέλα όπως νευρωνικά δίκτυα και μοντέλα τύπου δένδρων (Gocheva-Ilieva, et al., 2013).

Ο αλγόριθμος Random Forest (Breiman, 2001) αποτελεί συνδυασμό προβλέψεων τύπου δένδρων (Tree Predictors) έτσι ώστε κάθε δένδρο να βασίζεται σε ένα τυχαίο, ανεξάρτητου δείγματος, διάνυσμα, με την ίδια κατανομή για όλα τα δένδρα του δάσους.

Οι πιο σημαντικές παράμετροι είναι ο αριθμός δένδρων (ntree) και το πλήθος των μεταβλητών (mtry) ώστε:

$$Y = h(X) = \frac{1}{mtry} \sum_{i=1}^{ntree} h_i(X) \quad (4)$$

όπου:

- Y: το διάνυσμα εξόδου
- h: ο εκτιμητής
- X: Το διάνυσμα μεταβλητών εισόδου

(Ali & J., 2015)

Ο αλγόριθμος χωρίζει τυχαία τα δεδομένα σε ένα έναν αριθμό από ίσα υποσύνολα, κάθε ένα από τα οποία περιλαμβάνει ένα υποσύνολο από τις μεταβλητές του αρχικού συνόλου δεδομένων. Κάθε δένδρο εκπαιδεύεται βάσει του εκάστοτε υποσυνόλου. Το σύνολο των δένδρων «ψηφίζει» για την παραγωγή της τελικής πρόγνωσης. Με τον τρόπο αυτό προκύπτει ένας συνδυαστικός (ensemble based) αλγόριθμος, ενώ ταυτόχρονα αναδεικνύονται και οι πλέον σημαντικές (ως προς τη συμμετοχή τους στην τελική πρόγνωση) μεταβλητές των δεδομένων (Καρατζάς, 2012).

2.2.2.4. Διαδικασία επαλήθευσης μοντέλων

Κατά την εκπαίδευση μοντέλων μηχανικής μάθησης εφαρμόζεται η τεχνική αξιολόγησης διασταυρούμενης επαλήθευσης (Cross-Validation) με πιο διαδεδομένη την μέθοδο k-fold cross validation. Στη μέθοδο αυτή το σύνολο των δεδομένων διαιρείται σε k επιμέρους μικρότερα τμήματα. Έπειτα η εκπαίδευση του μοντέλου αξιοποιεί τα k-1 τμήματα και αξιολογείται με αυτό που απομένει, με τη διαδικασία να επαναλαμβάνεται k φορές, αφήνοντας κάθε φορά διαφορετικό τμήμα των δεδομένων έξω από τη φάση εκπαίδευσης, ώστε να χρησιμοποιηθεί για την επαλήθευση του μοντέλου.

Ειδικά όμως για πρόγνωση χρονοσειρών, η παραπάνω μέθοδος αποδίδει σφάλματα καθώς, όπως έχει αναφερθεί, τα στοιχεία χρονοσειρών είναι αυτοσυσχετιζόμενα, ενώ τεχνικές όπως η k-fold θεωρούν τα δείγματα ανεξάρτητα. Για τον λόγο αυτό χρησιμοποιείται η μέθοδος “Time Series Split” (Pedregosa, et al., 2011) κατά το cross validation, η οποία διατηρεί την σειρά των τμημάτων (folds) αυξάνοντας σταδιακά το δείγμα εκπαίδευσης.

2.2.2.5. Αξιολόγηση πρόγνωσης

Η αξιολόγηση των μοντέλων πρόγνωσης πραγματοποιήθηκε εκτιμώντας τα μέτρα:

- Μέσο απόλυτο σφάλμα (Mean Absolute Error)
- Μέσο τετραγωνικό σφάλμα (Mean Squared Error)
- Μέσο απόλυτο ποσοστιαίο σφάλμα (Mean Absolute Percentage Error)
- Συντελεστής προσδιορισμού (Coefficient of Determination – R²)

Μέσο απόλυτο σφάλμα (MAE)

Αποτελεί τον αριθμητικό μέσο του απόλυτου σφάλματος. Υπολογίζεται εύκολα, τα σφάλματα φέρουν την ίδια βαρύτητα λόγω απόλυτης τιμής και δεν παρουσιάζει ευαισθησία σε ακραίες τιμές. Χαμηλότερο MAE συνεπάγεται ευστοχότερη πρόβλεψη. (Padhma, 2021)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

Όπου

- y_i : Η πραγματική τιμή
- \hat{y}_i : Η προβλεπόμενη τιμή
- n: Το πλήθος δείγματος

Μέσο τετραγωνικό σφάλμα (MSE)

Διαφορά με το ΜΑΕ αποτελεί η ύψωση του σφάλματος στο τετράγωνο αντί της απόλυτης τιμής. Λόγω αυτού είναι πιο ευαίσθητος εκτιμητής σε ακραίες τιμές σφάλματος. Χαμηλότερο MSE συνεπάγεται ευστοχότερη πρόβλεψη. (Padhma, 2021)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6)$$

Μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE)

Η απόλυτη διαφορά μεταξύ πραγματικής και προβλεπόμενης τιμής διαιρείται με την πραγματική τιμή και μετατρέπεται σε ποσοστό. Αυξάνεται γραμμικά κατά την αύξηση του σφάλματος. Επειδή ως εκτιμητής είναι ανεξάρτητος της κλίμακας των δεδομένων χρησιμοποιήθηκε και για την σύγκριση απόδοσης προγνώσεων μεταξύ διαφορετικών μεγεθών. Χαμηλότερο MAPE συνεπάγεται ευστοχότερη πρόβλεψη. (Padhma, 2021)

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} * 100\% \quad (7)$$

Συντελεστής προσδιορισμού (R^2)

Ο συντελεστής R^2 υποδεικνύει το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής η οποία μοντελοποιείται επιτυχώς μέσω των ανεξάρτητων. Υπολογίζεται ως το άθροισμα τετραγώνων παλινδρόμησης προς το συνολικό άθροισμα τετραγώνων. Εκφράζεται ως ποσοστό, ανεξάρτητο της κλίμακας των δεδομένων και αξιοποιήθηκε και για σύγκριση προβλέψεων μεταξύ διαφορετικών μετρούμενων μεγεθών. Υψηλότερη τιμή R^2 συνεπάγεται ακριβέστερη προσέγγιση.

$$R^2 = \frac{SS_{regression}}{SS_{total}} \quad (8)$$

Όπου

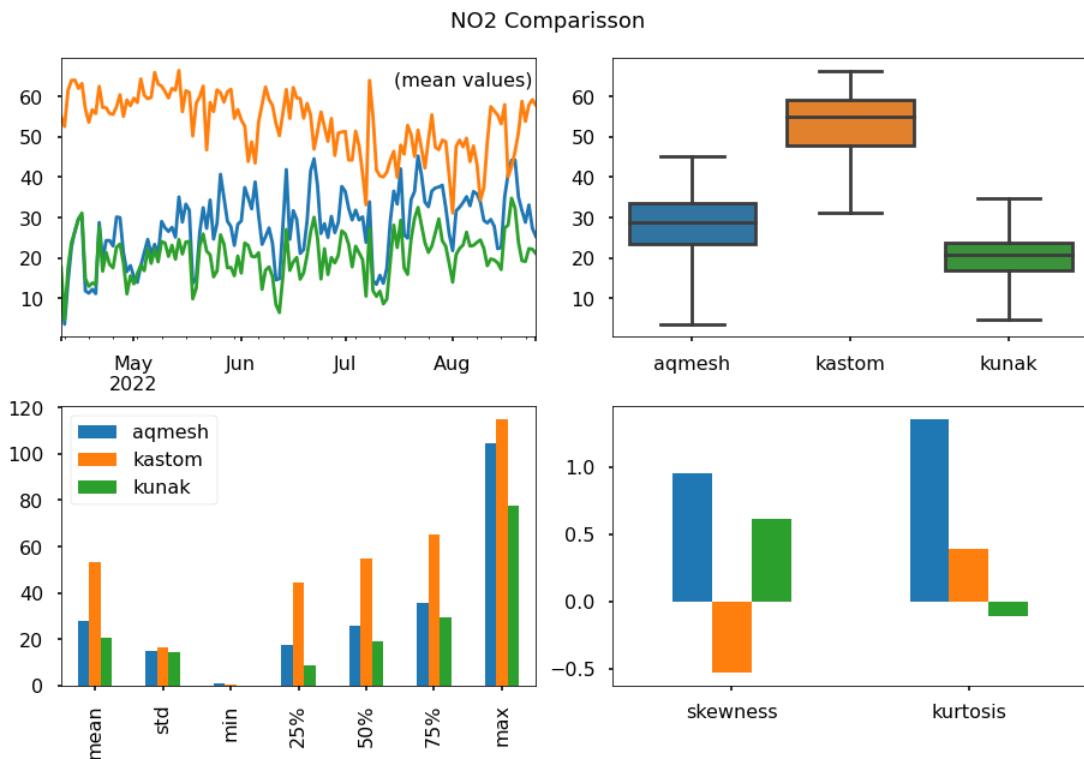
- $SS_{regression} = \left(\hat{y}_i - \frac{1}{n} \sum_{i=1}^n y_i \right)^2$
- $SS_{residual} = \sum_{i=1}^n \left(y_i - \frac{1}{n} \sum_{i=1}^n y_i \right)^2$
- $SS_{total} = SS_{regression} + SS_{residual}$

3. Αποτελέσματα και συζήτηση

Ακολουθεί παρουσίαση αποτελεσμάτων για κάθε ενότητα. Επιλέχθηκαν αντιπροσωπευτικά παραδείγματα εφαρμογής κάθε μεθόδου για συγκεκριμένο ΚΠΠΑ ή μετρούμενο μέγεθος. Πλήρη διαγράμματα και πίνακες για κάθε ΚΠΠΑ και ρύπο περιλαμβάνονται στο Παράρτημα.

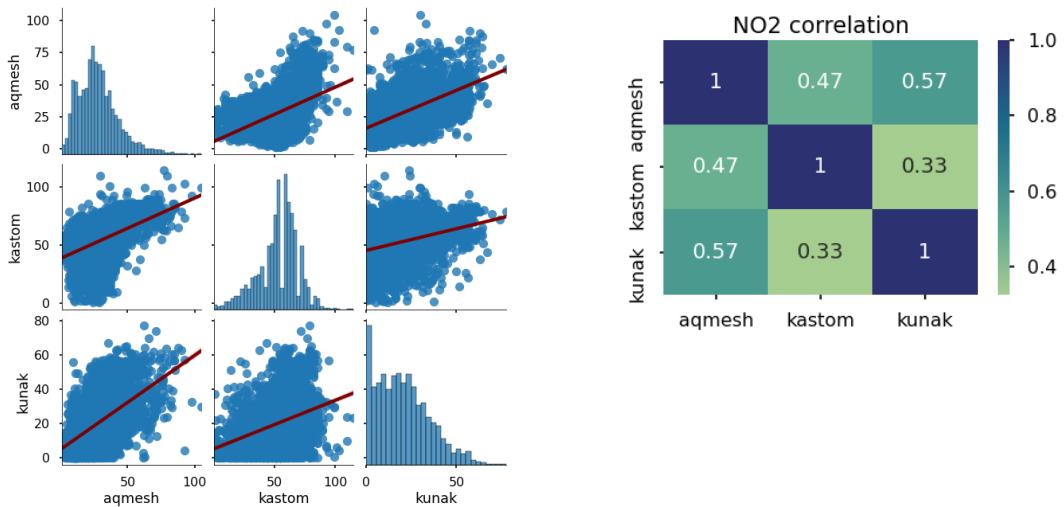
3.1. Σύγκριση κόμβων

Κατά την σύγκριση μεγεθών περιγραφικής στατιστικής για το διοξείδιο του αζώτου μεταξύ των ΚΠΠΑ Aqmesh, KASTOM και KUNAK (Εικόνα 1), παρατηρείται σημαντική απόκλιση μεταξύ KASTOM και των υπόλοιπων συσκευών. Οι τιμές του KASTOM είναι αρκετά υψηλότερες, ιδιαίτερα στο αρχικό τμήμα της χρονοσειράς. Η τυπική απόκλιση μεταξύ των τριών απέχει λιγότερο.



Εικόνα 1. Σύγκριση στατιστικών μεγεθών για NO₂ (ΚΠΠΑ Aqmesh, KASTOM, Kunak).

Παρατηρείται επίσης ότι η διαφορά μεταξύ των επιπέδων συγκέντρωσης του 20%, 50% και 75% είναι αρκετά όμοια ανεξάρτητα από τον ΚΠΠΑ. Η λοξότητα των Aqmesh και Kunak προκύπτει θετική ενώ του KASTOM αρνητική. Υψηλή κύρτωση παρατηρείται στις τιμές του ΚΠΠΑ Aqmesh, όπως φαίνεται και στα ιστογράμματα (Εικόνα 2)



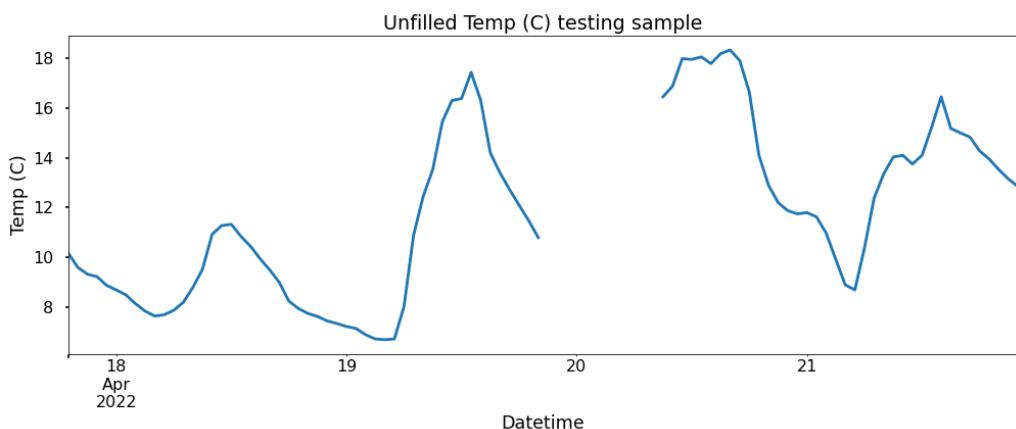
Εικόνα 2. Συσχέτιση μετρήσεων NO_2 μεταξύ κόμβων (ΚΠΠΑ Aqmesh, KASTOM, Kunak). Διαγράμματα συσχέτισης και ιστογράμματα (αριστερά), πίνακας συντελεστών συσχέτισης κατά Pearson (δεξιά).

Όπως φαίνεται στην Εικόνα 2, τα επίπεδα των συγκεντρώσεων διοξειδίου του αζώτου για τους διαφορετικούς κόμβους παρουσιάζουν μέτρια θετική συσχέτιση, ιδιαίτερα οι τιμές των κόμβων Aqmesh και Kunak.

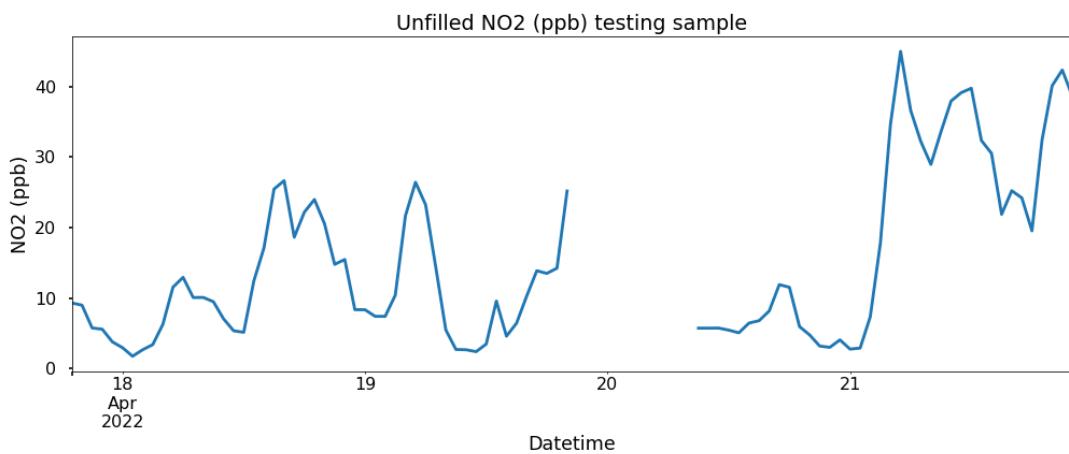
3.2. Ανάλυση δεδομένων

3.2.1. Προεπεξεργασία – Συμπλήρωση ελλειπουσών τιμών

Για κάθε κόμβο και μετρούμενο μέγεθος δοκιμάστηκαν τεχνικές συμπλήρωσης ελλειπουσών τιμών και εκτιμήθηκαν τα σχετικά σφάλματα. Για την τελική διαμόρφωση της χρονοσειράς επιλέχθηκε η μέθοδος με τον χαμηλότερο δείκτη σφάλματος για το εκάστοτε μέγεθος. Ακολουθεί παράδειγμα μεταξύ συμπλήρωσης τιμών NO_2 και Θερμοκρασίας του ΚΠΠΑ Aqmesh.

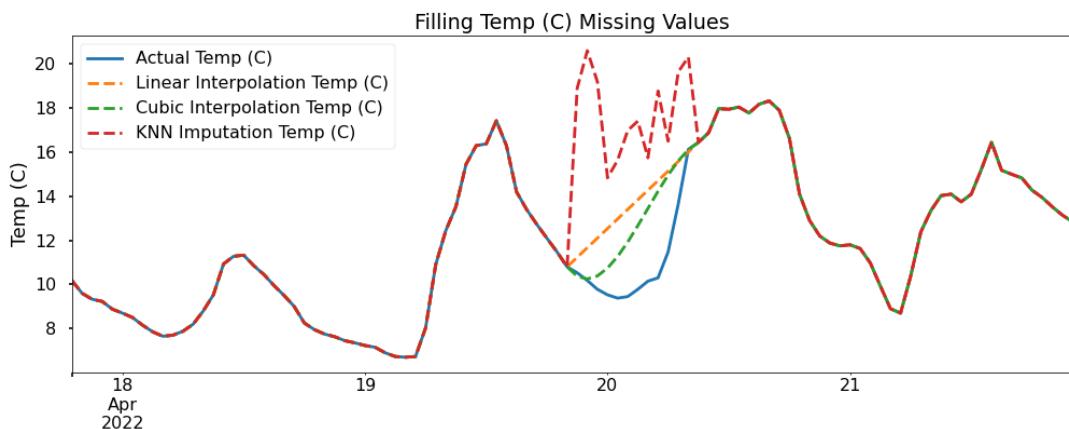


Εικόνα 3. Δοκιμαστικό τμήμα χρονοσειράς Θερμοκρασίας, προ συμπλήρωσης (ΚΠΠΑ Aqmesh)



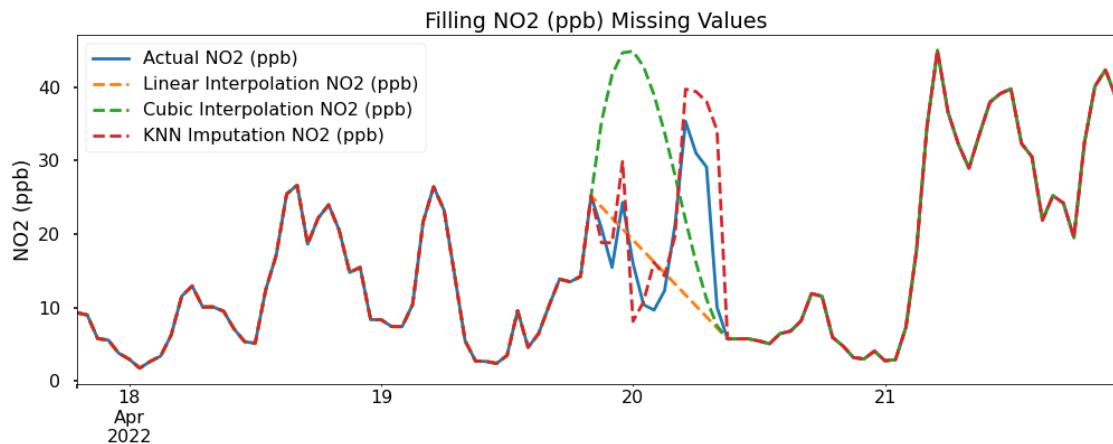
Εικόνα 4. Δοκιμαστικό τμήμα χρονοσειράς Διοξειδίου του Αζώτου, προ συμπλήρωσης (ΚΠΠΑ Aqmesh)

Στην Εικόνα 3 και Εικόνα 4 παρουσιάζεται δείγμα μετρήσεων NO2 και Θερμοκρασίας, αντίστοιχα, από το οποίο αφαιρέθηκαν ενδεικτικά τιμές ώστε να συμπληρωθούν στη συνέχεια με στόχο την αξιολόγηση των μεθόδων.



Εικόνα 5. Δοκιμή συμπλήρωσης ελλειπουσών τιμών Θερμοκρασίας (ΚΠΠΑ Aqmesh)

Στην Εικόνα 5 και Εικόνα 6 απεικονίζεται τμήμα της συμπληρωμένης δοκιμαστικής χρονοσειράς με τις μεθόδους Linear Interpolation, Cubic Interpolation και KNN Imputation, για Θερμοκρασία και Διοξείδιο του Αζώτου αντίστοιχα. Παρατηρείται η διαφορετική προσέγγιση της πραγματικής καμπύλης (Actual), κάθε μεθόδου και η αντίστοιχη ευστοχία όπως εκφράζεται και από τα σφάλματα που υπολογίστηκαν.



Εικόνα 6. Δοκιμή συμπλήρωσης ελλειπουσών τιμών NO2 (ΚΠΠΑ Aqmesh)

Όπως διακρίνεται στον Πίνακας 4 και Πίνακας 5, για τις τιμές θερμοκρασίας, χαμηλότερο συνολικά σφάλμα και υψηλότερο συντελεστή προσδιορισμού αποδίδει η μέθοδος Cubic Interpolation ενώ αντίστοιχα για το Διοξείδιο του Αζώτου η μέθοδος KNN Imputer.

Πίνακας 4. Σφάλμα συμπλήρωσης ελλειπουσών τιμών θερμοκρασίας

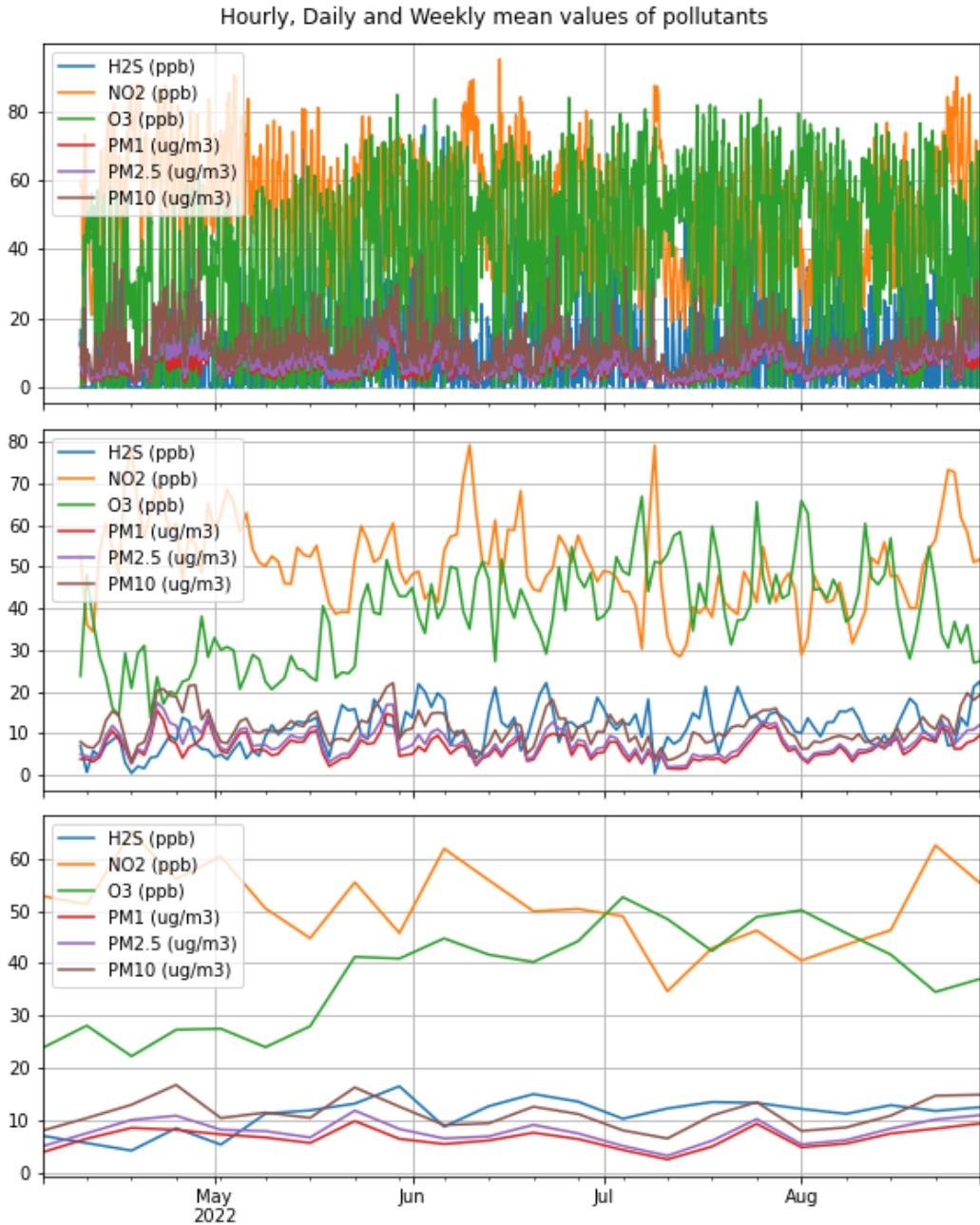
	MAE	MSE	R2
KNN	0.285	1.415	0.965
Linear	0.242	0.933	0.977
Cubic	0.164	0.487	0.988

Πίνακας 5. Σφάλμα συμπλήρωσης ελλειπουσών τιμών NO2

	MAE	MSE	R2
KNN	1.119	20.24	0.908
Linear	1.249	25.888	0.884
Cubic	2.195	89.325	0.693

3.2.2. Αναδειγματοληψία

Στα παρακάτω διαγράμματα παρουσιάζονται οι τιμές συγκεντρώσεων κάθε ρύπου, ανά ρυθμό δειγματοληψίας.

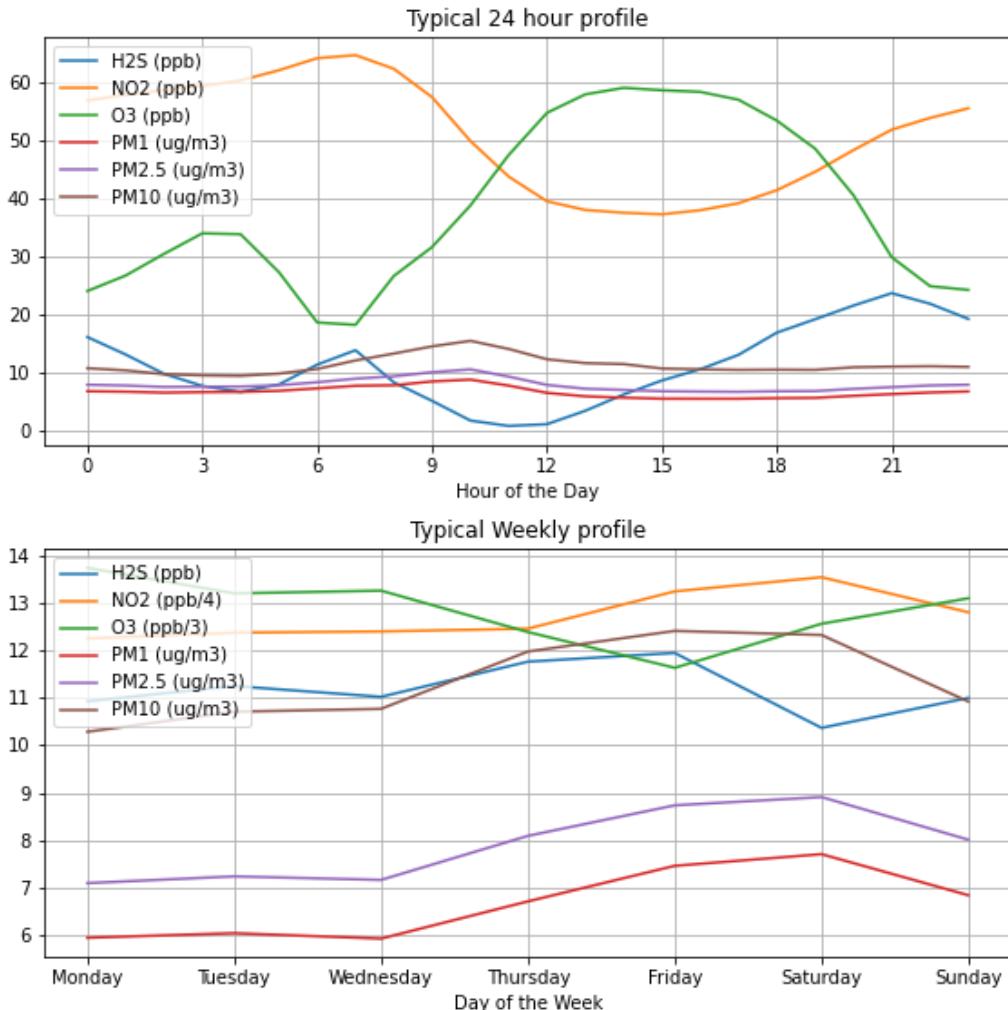


Εικόνα 7. Ωριαία (πάνω), ημερήσια (μέση) και εβδομαδιαία (κάτω) αναδειγματοληψία ρύπων (ΚΠΠΑ Kunak)

Όσο μειώνεται η συχνότητα δειγματοληψίας παρατηρούνται ευκολότερα οι τάσεις και άλλα χαρακτηριστικά των χρονοσειρών. Για παράδειγμα η ήπια αυξητική τάση του Όζοντος στο διάγραμμα εβδομαδιαίου ρυθμού, ως αποτέλεσμα της ηλιακής ακτινοβολίας των καλοκαιρινών μηνών και της μείωσης των εκπομπών ΝΟ λόγω μείωσης της κυκλοφορίας οχημάτων τον μήνα Αύγουστο.

3.2.3. Τυπικά προφίλ

Στα παρακάτω διαγράμματα παρουσιάζονται τα τυπικά προφίλ των συγκεντρώσεων των ρύπων σύμφωνα με τις μετρήσεις του ΚΠΠΑ Kunak.



Εικόνα 8. Τυπικά εικοσιτετράωρα (πάνω) και εβδομαδιαία (κάτω) προφίλ (ΚΠΠΑ Kunak)

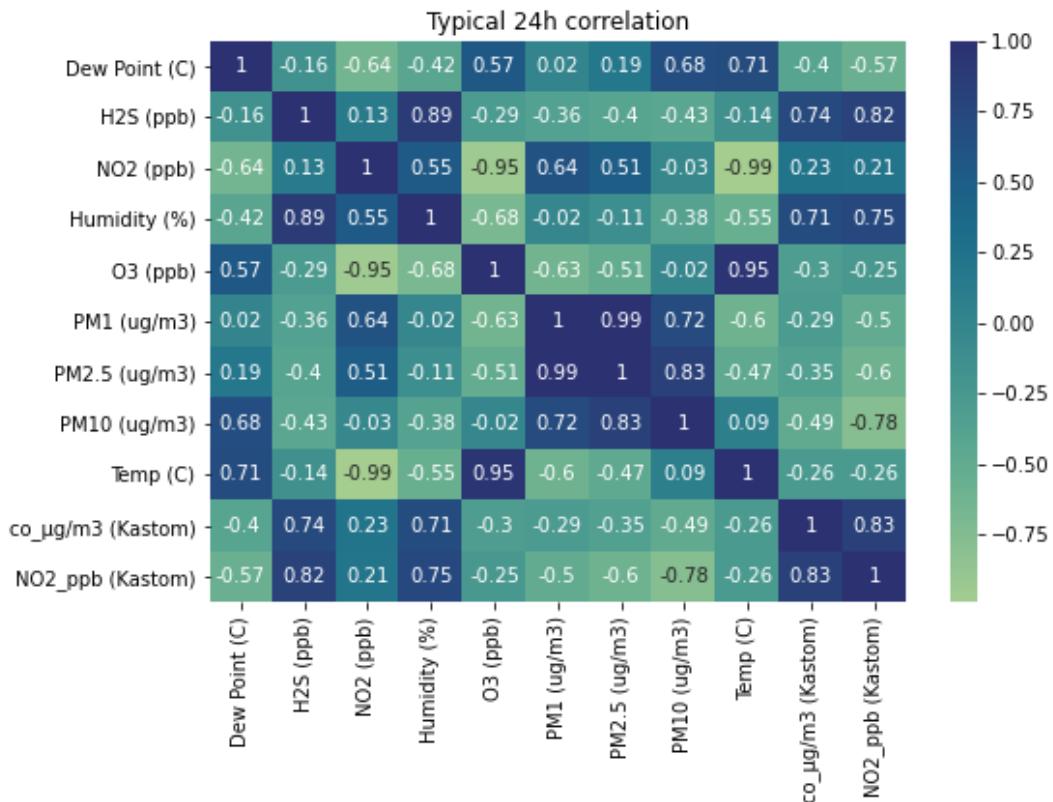
Στο εβδομαδιαίο διάγραμμα οι μονάδες του O₃ και NO₂ διαιρέθηκαν με 3 και 4 αντίστοιχα για λόγους ευκρίνειας.

- Τυπικό εικοσιτετράωρο: Το Υδρόθειο και το Διοξείδιο του Αζώτου παρουσιάζουν δύο κορυφές, τις πρωινές ώρες (7:00) και τις βραδινές (21:00). Το Όζον ακολουθεί αντιστρόφως ανάλογη πορεία κατά τη διάρκεια της ημέρας. Τα αιωρούμενα σωματίδια δείχνουν μια ήπια αύξηση στις 12:00.
- Τυπική εβδομάδα: Το Όζον παρουσιάζει σταθερότητα κατά την εξέλιξη της τυπικής εβδομάδας εκτός μίας ύφεσης που παρατηρείται τις Παρασκευές. Το Υδρόθειο κορυφώνεται τις Παρασκευές ενώ οι υπόλοιποι ρύποι ακολουθούν αυξητική τάση μέχρι τα Σάββατα και κατά τη διάρκεια του Σαββατοκύριακου μειώνονται.

3.2.4. Διερεύνηση συσχέτισης

Όπως προκύπτει από τον πίνακα συσχέτισης (Εικόνα 9), οι ρύποι εμφανίζουν θετική ή αρνητική συσχέτιση μεταξύ τους ή με τις μετεωρολογικές συνθήκες για ένα τυπικό εικοσιτετράωρο. Συγκεκριμένα:

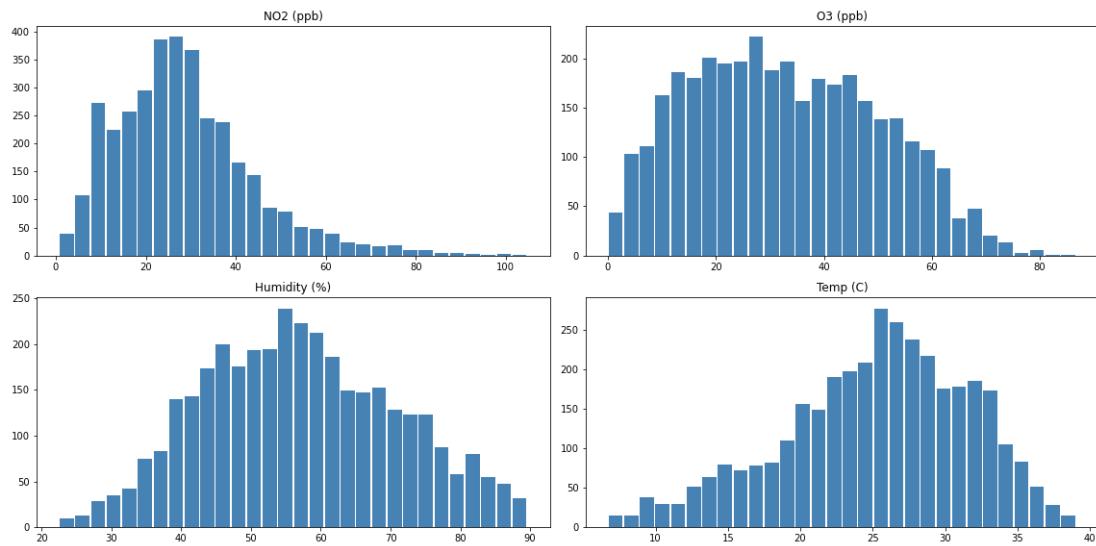
- οι τιμές του υδρόθειου (H_2S) παρουσιάζουν έντονη θετική συσχέτιση με αυτές της σχετικής υγρασίας, του μονοξειδίου του άνθρακα και του διοξειδίου του αζώτου, ενώ ήπια αρνητική με τις τιμές αιωρούμενων σωματίδιων.
- το διοξείδιο του αζώτου (NO_2) του ΚΠΠΑ Kunak, συσχετίζεται αρνητικά με το σημείο δρόσου και έντονα αρνητικά με το όζον και τη θερμοκρασία, ενώ ήπια θετικά με τα αιωρούμενα σωματίδια ($PM_{1,2}$) και την υγρασία.
- το διοξείδιο του αζώτου (NO_2) του ΚΠΠΑ Kastom, δείχνει θετική συσχέτιση με το υδρόθειο, την υγρασία και το μονοξείδιο του άνθρακα, ενώ αρνητική με το σημείο δρόσου και τα σωματίδια.
- το όζον αντίστοιχα φαίνεται να συσχετίζεται θετικά με το σημείο δρόσου και ειδικά με τη θερμοκρασία, ενώ αρνητικά με την υγρασία, τα σωματίδια και έντονα αρνητικά με το διοξείδιο του αζώτου.
- το μονοξείδιο του άνθρακα παρουσιάζει θετική συσχέτιση με το υδρόθειο, την υγρασία και το διοξείδιο του αζώτου του ΚΠΠΑ Kastom.



Εικόνα 9. Πίνακας συσχέτισης κατά Pearson, τιμών τυπικού εικοσιτετράωρου (ΚΠΠΑ Kunak και KASTOM)

3.2.5. Περιγραφική Στατιστική

Στην Εικόνα 10 απεικονίζεται σε ιστογράμματα η συχνότητα εμφάνισης τιμών των μετρούμενων μεγεθών και στον Πίνακας 6 οι παράμετροι περιγραφικής στατιστικής για τον ΚΠΠΑ Aqmesh.

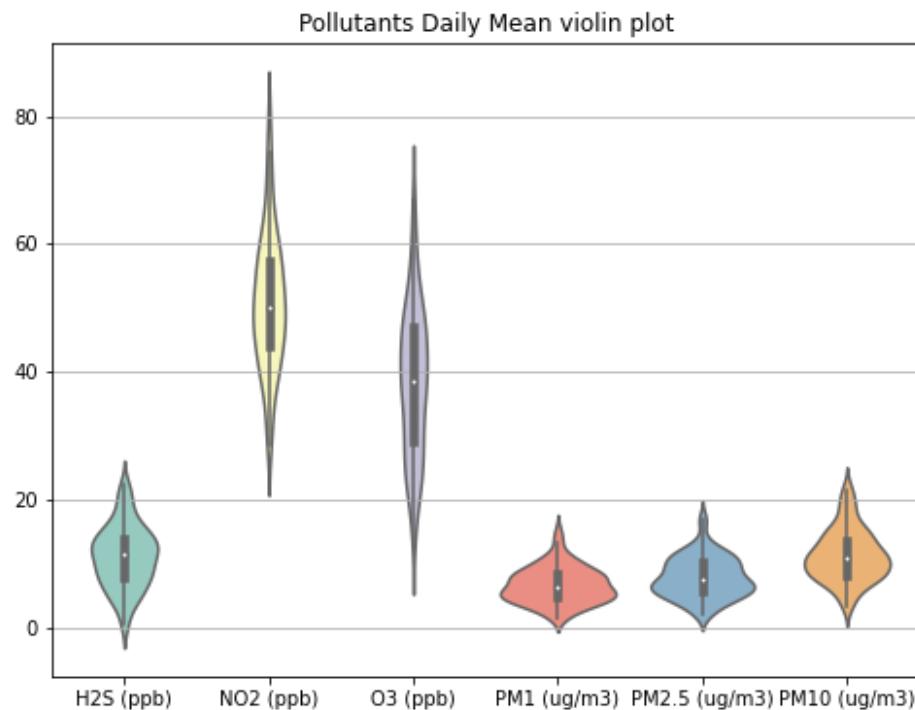


Εικόνα 10. Ιστογράμματα ωριαίου ρυθμού δειγματοληψίας μετρούμενων μεγεθών (ΚΠΠΑ Aqmesh)

Πίνακας 6. Παράμετροι περιγραφικής στατιστικής (Aqmesh)

	NO ₂ (ppb)	O ₃ (ppb)	Relative Humidity (%)	Temp (C)
count	3552.0	3552.0	3552.0	3552.0
mean	28.6	32.83	57.04	25.22
std	15.31	17.28	13.95	6.46
min	0.73	0.01	22.5	6.7
25%	17.9	18.8	46.43	21.2
50%	26.55	31.6	56.25	25.84
75%	36.48	46.04	67.18	29.96
max	104.77	86.68	89.58	39.05
kurtosis	1.57	-0.8	-0.6	-0.22
skew	1.01	0.23	0.14	-0.46

Το διάγραμμα της Εικόνα 11 αποκαλείται Violin Plot. Αποτελεί συνδυαστική μέθοδο θηκογράμματος και πυκνότητας πιθανότητας. Εκτός από επίπεδα τιμών, απεικονίζει επιπλέον τις κατανομές ποσοτικών δεδομένων για κατηγορικές μεταβλητές ώστε να μπορούν αυτές να συγκριθούν.

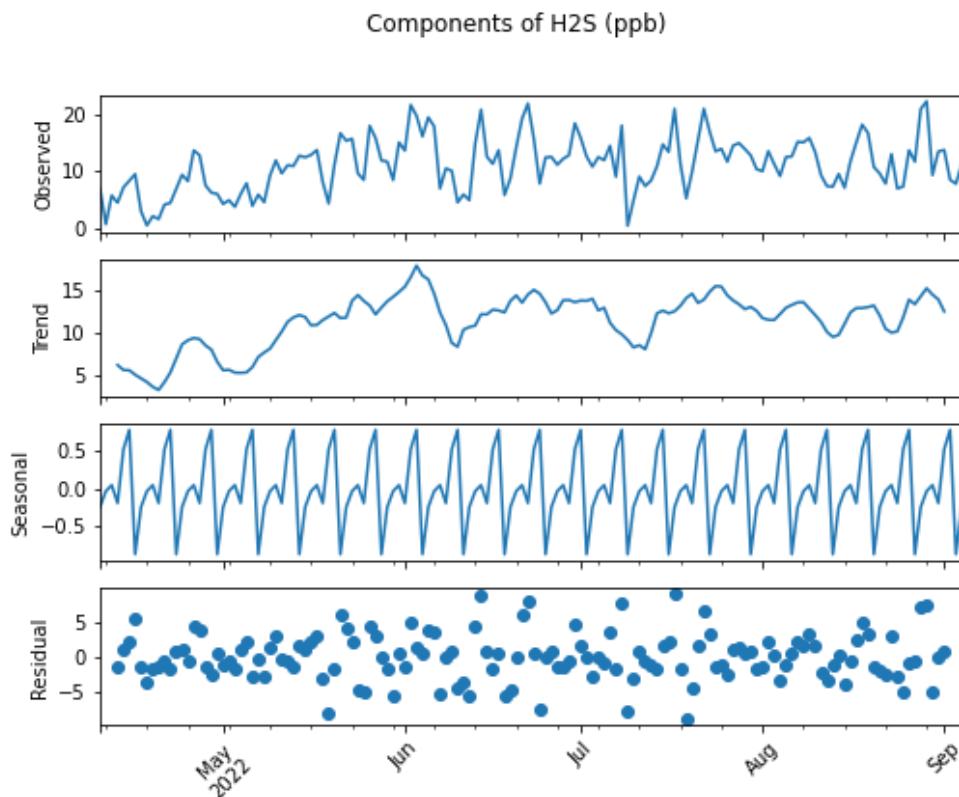


Εικόνα 11. Violin plot μέσου ημερήσιου ρυθμού δειγματοληψίας ρύπων (ΚΠΠΑ Kunak)

3.2.6. Αποσύνθεση χρονοσειράς

Στην Εικόνα 12 παρουσιάζονται τα συνθετικά στοιχεία της χρονοσειράς Υδρόθειου, ημερήσιας συχνότητας, των μετρήσεων του κόμβου Kunak. Στην 1^η σειρά καταγράφεται η πλήρης παρατηρούμενη χρονοσειρά, στη 2^η η τάση, στην 3^η η εποχικότητα-περιοδικότητα και στην 4^η ο θόρυβος που προκύπτει σύμφωνα με την ενότητα [2.2.1.6](#).

Παρατηρείται μια ήπια αύξουσα τάση κατά την πάροδο των μηνών, ενώ η εποχικότητα που προκύπτει υποδεικνύει περίοδο διάρκειας 7 ημερών.

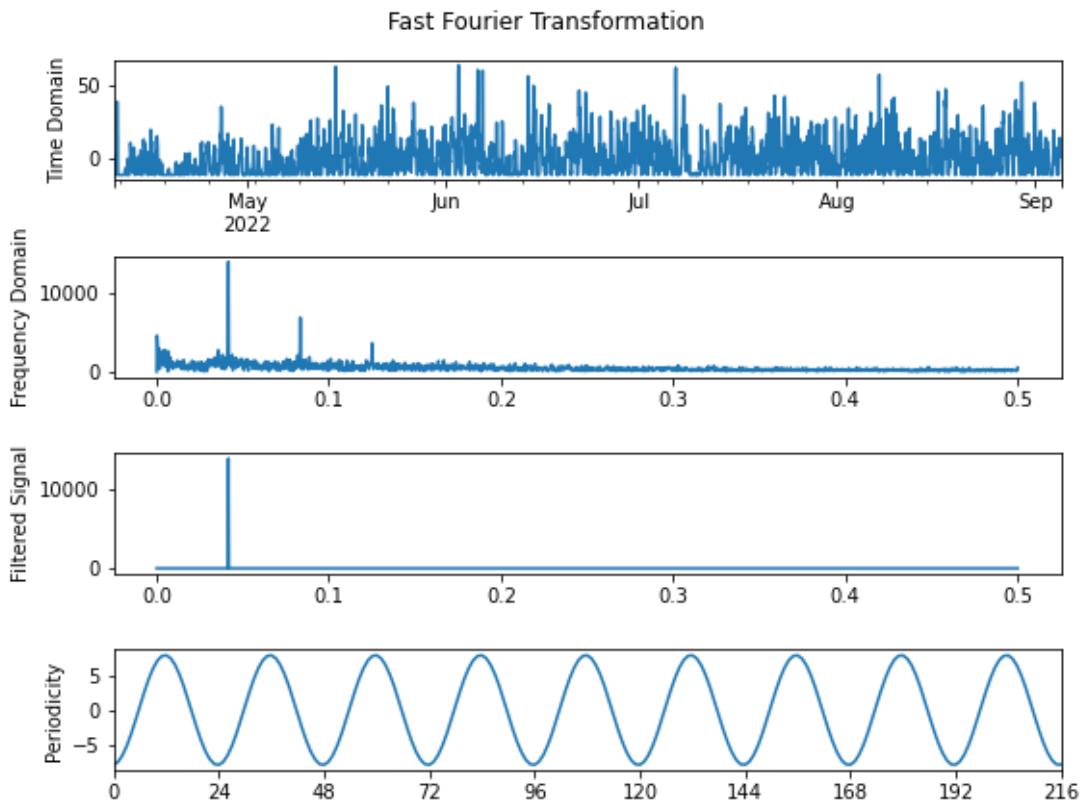


Εικόνα 12. Συνθετικά προσθετικού μοντέλου χρονοσειράς υδρόθειου (ΚΠΠΑ Kunak)

Περιοδικότητα

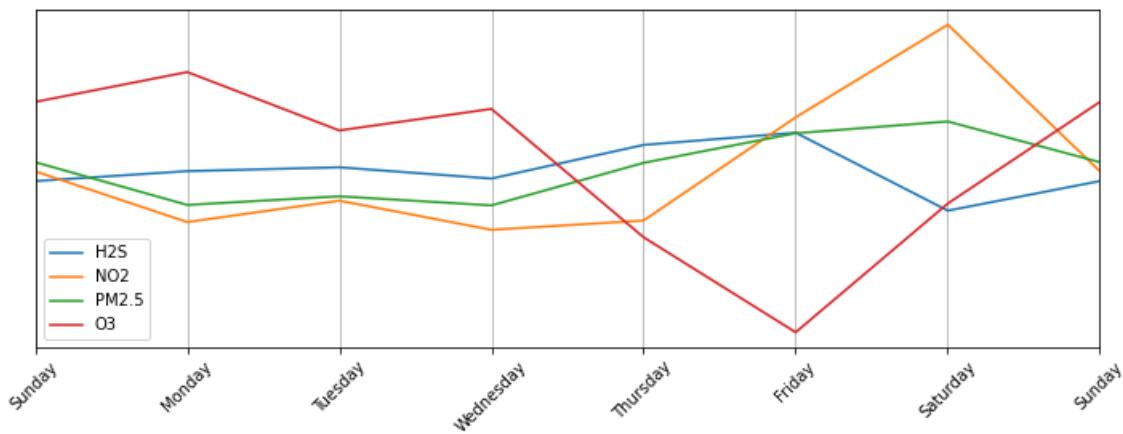
Η ύπαρξη περιοδικότητας επαληθεύτηκε με την μέθοδο Fast Fourier Transformation. Στην Εικόνα 13 διακρίνονται τα στάδια του μετασχηματισμού:

1. Διάγραμμα παρατηρούμενης χρονοσειράς, ωριαίας συχνότητας.
2. Μετασχηματισμός στο πεδίο συχνότητας κατόπιν κανονικοποίησης (ενότητα [2.2.1.7](#))
3. Απομόνωση της κύριας συχνότητας
4. Αντίστροφος μετασχηματισμός του φιλτραρισμένου σήματος στο πεδίο του χρόνου.



Εικόνα 13. Γρήγορος μετασχηματισμός Fourier χρονοσειράς υδρόθειου (ΚΠΠΑ Kunak)

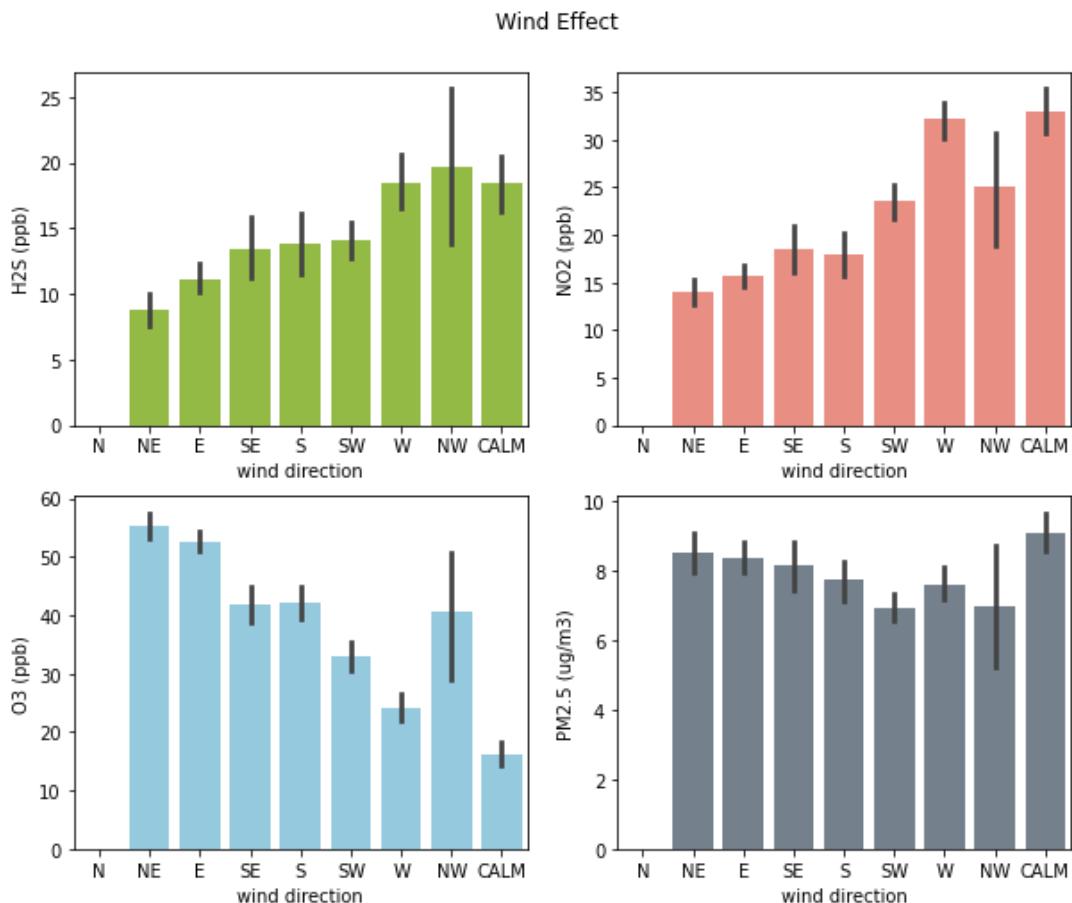
Η περιοδικότητα που χαρακτηρίζει το σύνολο των εξεταζόμενων ρύπων είναι εβδομαδιαία. Κατόπιν εστίασης σε μία περίοδο του προσθετικού μοντέλου, προκύπτουν οι ημερήσιες τάσεις, όπως διαμορφώνονται στο πέρας μίας εβδομάδας (Εικόνα 14). Τα αποτελέσματα επαληθεύουν αυτά της ενότητας [3.1.2](#), όπου προέκυψαν ακολουθώντας διαφορετική προσέγγιση.



Εικόνα 14. Ημερήσιες τάσεις εβδομαδιαίας περιοδικότητας ρύπων (Kunak)

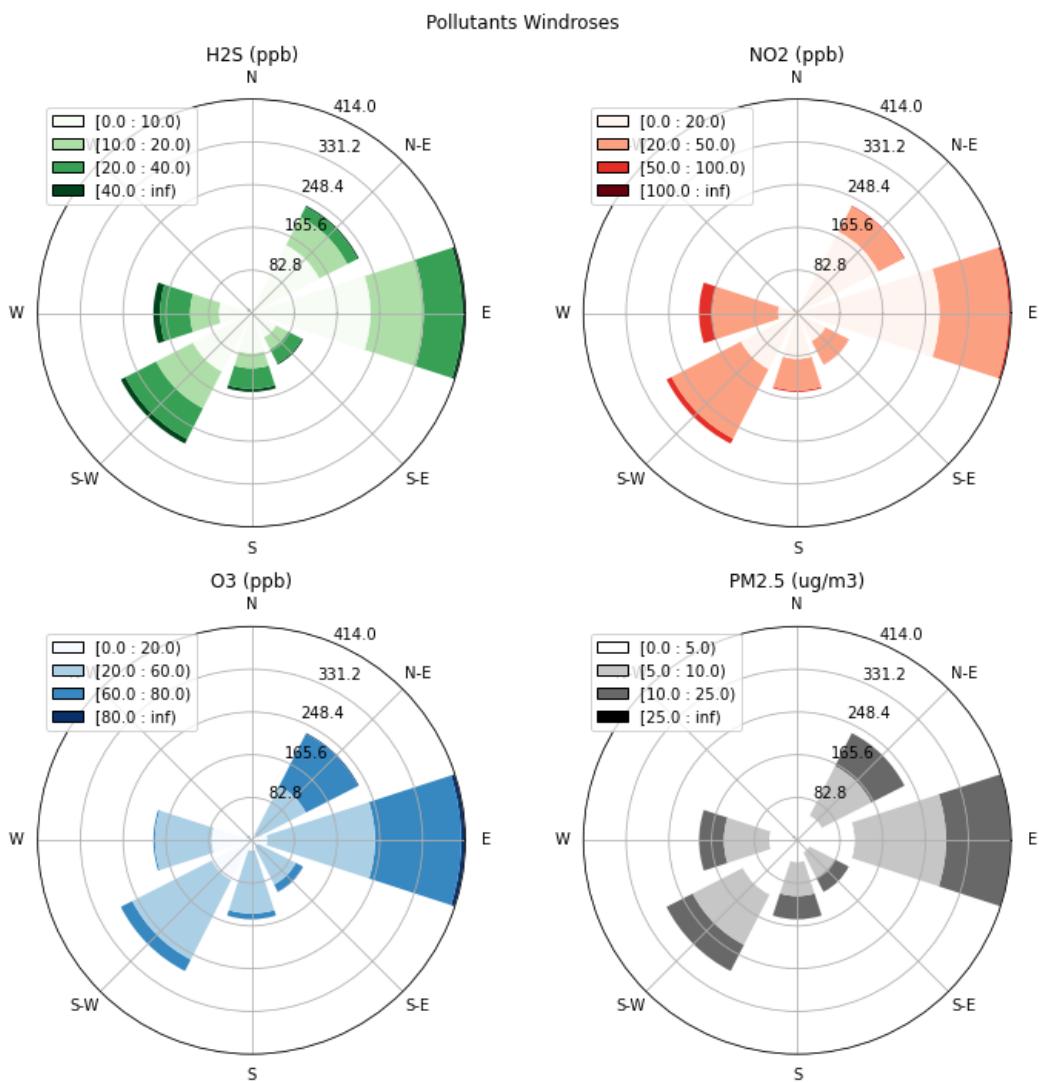
3.2.7. Συνεισφορά ανέμου

Η μέση τιμή συγκέντρωσης ρύπων του ΚΠΠΑ Kunak, ανά διεύθυνση του ανέμου, διακρίνεται στην Εικόνα 15, ενώ σε συνδυασμό με την συχνότητα εμφάνισης κάθε διεύθυνσης, εκτός των συνθηκών νηνεμίας (calm), σε ροδογράμματα του Εικόνα 16. Από τα ροδογράμματα προκύπτει ότι κυρίαρχες διευθύνσεις ανέμου για το διάστημα δειγματοληψίας είναι η ανατολική και νοτιοδυτική.



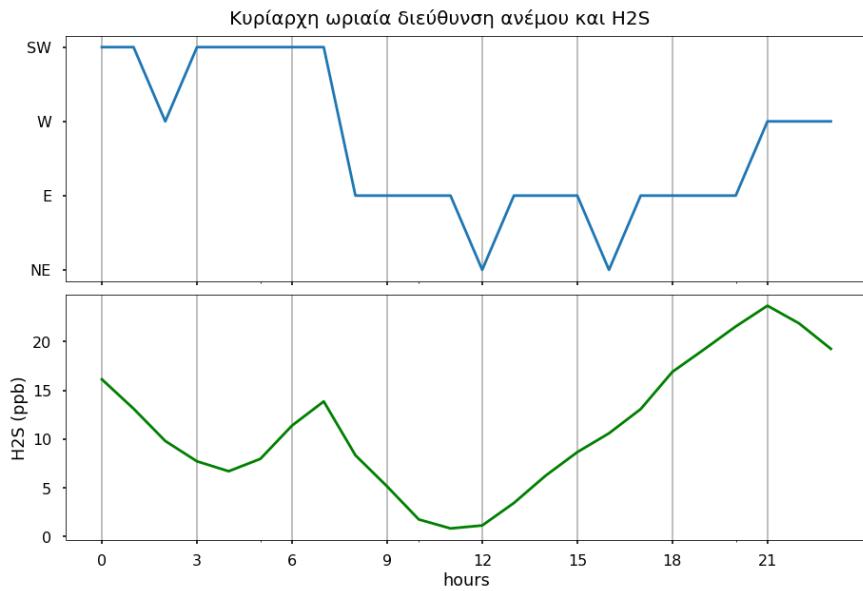
Εικόνα 15. Μέση τιμή συγκεντρώσεων ρύπων ανά διεύθυνση ανέμου (ΚΠΠΑ Kunak)

- Για το υδρόθειο, παρατηρούνται υψηλότερα επίπεδα όταν η διεύθυνση ανέμου είναι δυτική και βορειοδυτική (δηλαδή η συνήθης, σύμφωνα με τις μετρήσεις), ή ταχύτητας μικρότερης των 0.5 m/s. Να σημειωθεί ότι φαίνομενα δυσοσμίας προκαλούνται για συγκέντρωση υψηλότερη των 20 ppb H₂S (Kourtidis, Kelesis, & Petrakakis, 2008).
- Το διοξείδιο του αζώτου παρουσιάζει υψηλότερες τιμές για βόρειο άνεμο ή συνθήκες νηνεμίας.
- Το οζόνιο για βόρειο-βορειοανατολικές διευθύνσεις, αντίθετη συμπεριφορά με αυτή του NO₂.
- Η συμπεριφορά των αιρούμενων σωματιδίων δεν επηρεάζεται σημαντικά από τη διεύθυνση του ανέμου αν και παρατηρούνται υψηλότερες συγκεντρώσεις σε χαμηλές ταχύτητες ανέμου.



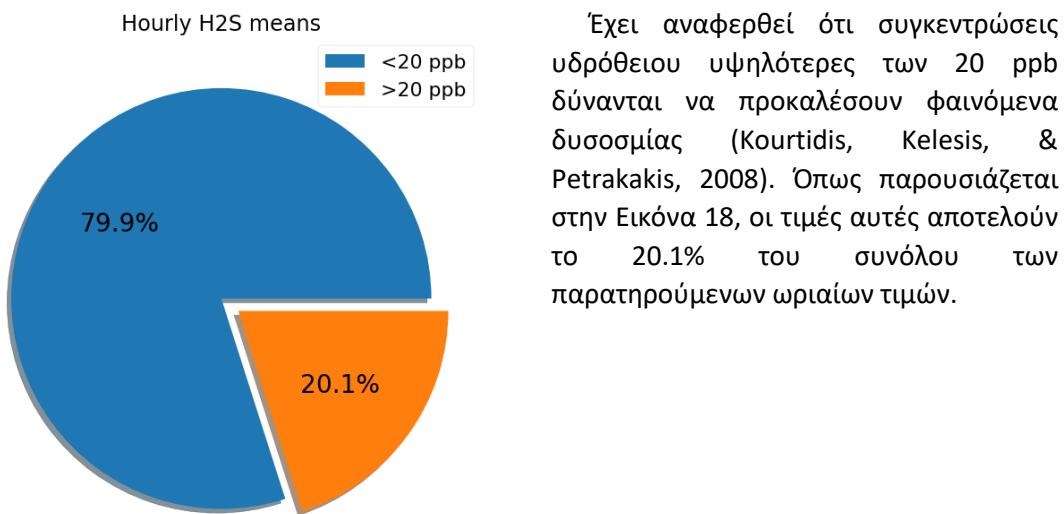
Εικόνα 16. Ροδογράμματα ρύπων ανά διεύθυνση και συχνότητα ανέμου (ΚΠΠΑ Kunak)

Η συχνότητα εμφάνισης Βόρειων και Βορειοδυτικών ανέμων κρίνεται αμελητέα ενώ δεν λαμβάνονται υπόψιν συγκεντρώσεις για ταχύτητα ανέμου < 0.5 m/s (νηνεμία).



Εικόνα 17. Επίδραση μέσης κυρίαρχης διεύθυνσης ανέμου στις τιμές H_2S τυπικού εικοσιτετράωρου (ΚΠΠΑ Kunak)

Σύμφωνα με την Εικόνα 17 οι κυρίαρχες διευθύνσεις ανέμου κατά τη διάρκεια τυπικού εικοσιτετράωρου είναι ΝΔ, Δ, Α, ΒΑ. Όπως παρατηρήθηκε νωρίτερα, το υδρόθειο παρουσιάζει υψηλότερες τιμές για νοτιοδυτικό και δυτικό άνεμο, αντίστοιχα και οι υψηλότερες μέσες τιμές εικοσιτετράωρου παρατηρούνται όταν επικρατούν οι ίδιες διευθύνσεις. Η κορύφωση στις 21:00 πιθανόν δείχνει συσσώρευση του ρύπου λόγω της ημερήσιας δραστηριότητας.

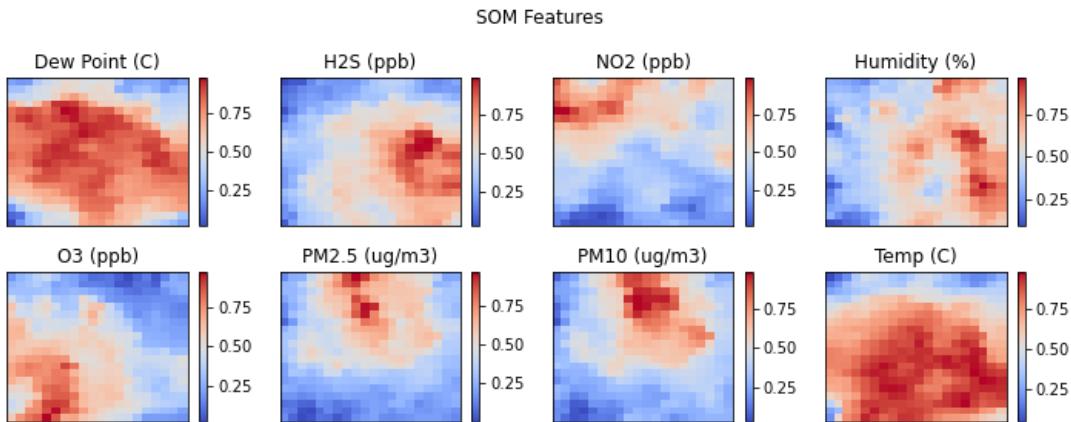


Εικόνα 18. Τιμές υδρόθειου υψηλότερες από το όριο πρόκλησης δυσοσμίας

3.2.8. Οπτικοποίηση – Συσταδοποίηση

3.2.8.1. Self-Organizing Maps (SOM)

Οργανώνοντας με την μέθοδο του των SOM κάθε μεταβλητή των δεδομένων, μπορούν αυτές να συγκριθούν οπτικά. Όπως παρατηρείται στην Εικόνα 19, το σημείο δρόσου με την θερμοκρασία και τα αιωρούμενα σωματίδια διαμέτρου 2.5 με αυτά των 10 μικρομέτρων, παρουσιάζουν παρόμοια τοπολογική εικόνα. Αυτό το στοιχείο ενισχύει την άποψη πως οι σχετικές παράμετροι σχετίζονται μεταξύ τους, κάτι το αναμενόμενο.

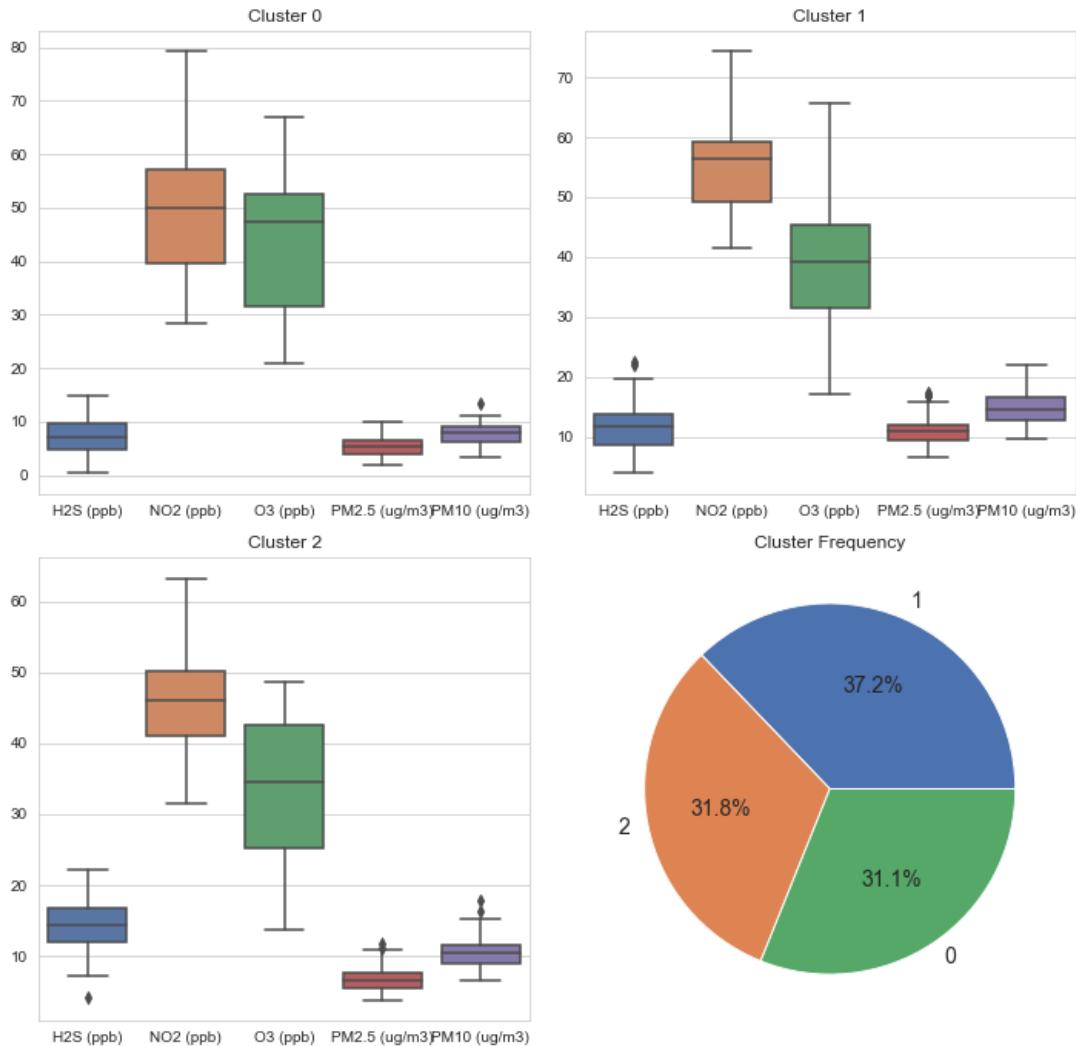


Εικόνα 19. Απεικόνιση χαρακτηριστικών SOM (ΚΠΠΑ Kunak)

Με εφαρμογή των SOM στο ενιαίο σετ δεδομένων, ημερήσιας συχνότητας, και εντοπισμό των συντεταγμένων των κυρίαρχων νευρώνων (winning neurons), ανατίθεται κάθε χρονικό βήμα στην αντίστοιχη ομάδα παραμέτρων. Έτσι δημιουργούνται συστάδες προκαθορισμένου πλήθους, τέτοιου ώστε να παρέχεται πληροφορία χρήσιμη για την εξήγηση της συμπεριφοράς των παραμέτρων.

Στην Εικόνα 20, απεικονίζονται οι τρεις συστάδες (clusters) όπως προέκυψαν έπειτα από εφαρμογή της μεθόδου στα δεδομένα του ΚΠΠΑ Kunak, και η συχνότητα εμφάνισης της κάθε μίας. Καθώς η συχνότητα ορίστηκε ως εικοσιτετράωρη, οι συστάδες απεικονίζουν εικοσιτετράωρα προφίλ ρύπανσης.

SOM Daily Profile Clustering



Εικόνα 20. SOM συσταδοποίηση ημερήσιων προφίλ συγκεντρώσεων ρύπων (ΚΠΠΑ Kunak)

Οι ομάδες είναι τρεις, με αρίθμηση από 0 έως 2.

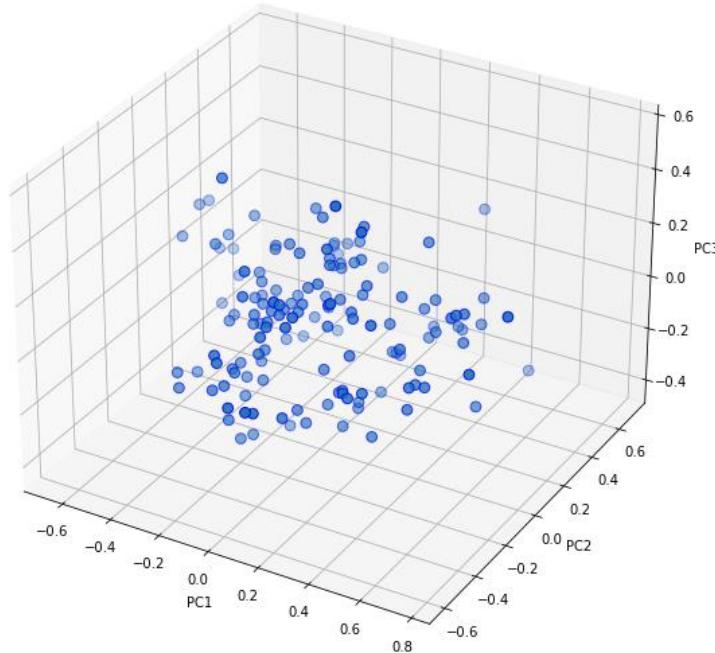
Κατά φθίνουσα συχνότητα παρατήρησης:

- Cluster 1: Προφίλ ημέρας όπου παρατηρούνται αισθητά υψηλότερες τιμές NO₂ από αυτές του Όζοντος. Τα αιωρούμενα σωματίδια βρίσκονται σε υψηλότερα σχετικά επίπεδα. Το υδρόθειο κάτω του ορίου των 20 ppb εκτός μιας ακραίας τιμής.
- Cluster 2: Παρατηρούνται οι μέγιστες τιμές υδρόθειου συγκριτικά με τα υπόλοιπα προφίλ, ξεπερνώντας τα 20 ppb.
- Cluster 0: Έντονα χαμηλότερα επίπεδα συγκέντρωσης υδρόθειου και συγκέντρωσης αιωρούμενων σωματιδίων. Παρατηρούνται οι υψηλότερες τιμές διοξειδίου του αζώτου.

3.2.8.2. K-Means

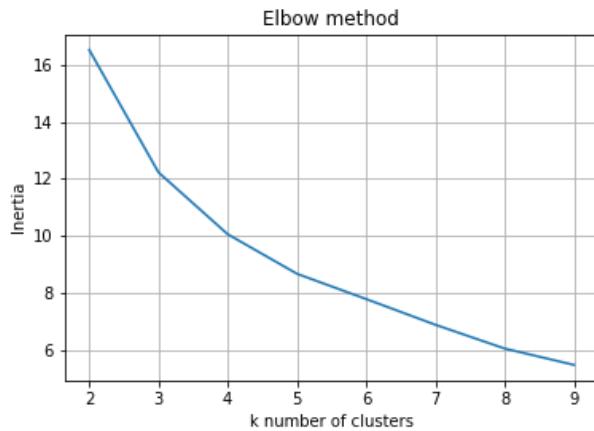
Εφαρμόζοντας την μέθοδο PCA που αναπτύχθηκε στην ενότητα 2.2.1.9 καθίσταται δυνατή η απεικόνιση του σετ δεδομένων πολλών μεταβλητών σε χώρο τριών διαστάσεων (Εικόνα 21). Κάθε κύρια συνιστώσα (PC) φέρει αντίστοιχα το 44.4%, 30.3% και 17.3%, δηλαδή το 92% της συνολικής πληροφορίας.

3D PCA visualization



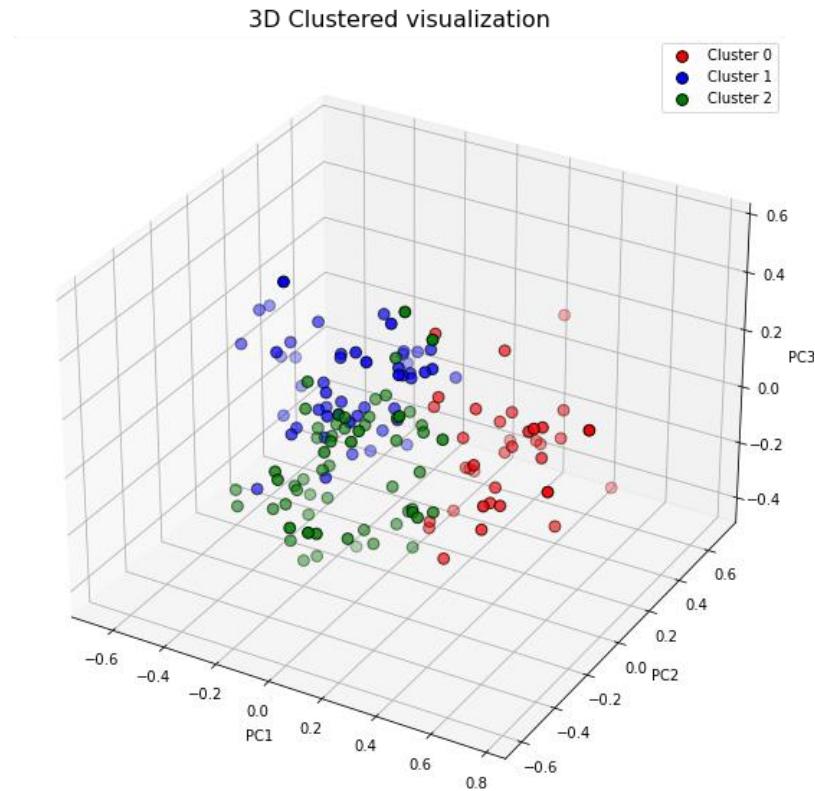
Εικόνα 21. Τρισδιάστατη απεικόνιση δεδομένων μεθόδου PCA (ΚΠΠΑ Kunak)

Όπως εξηγήθηκε, ο προσδιορισμός του πλήθους των συστάδων (clusters) βασίζεται στην βέλτιστα εξηγούμενη πληροφορία αλλά και στον τύπο του εκάστοτε προβλήματος.



Εικόνα 22. Elbow method για προσδιορισμό πλήθους συστάδων k (ΚΠΠΑ Kunak)

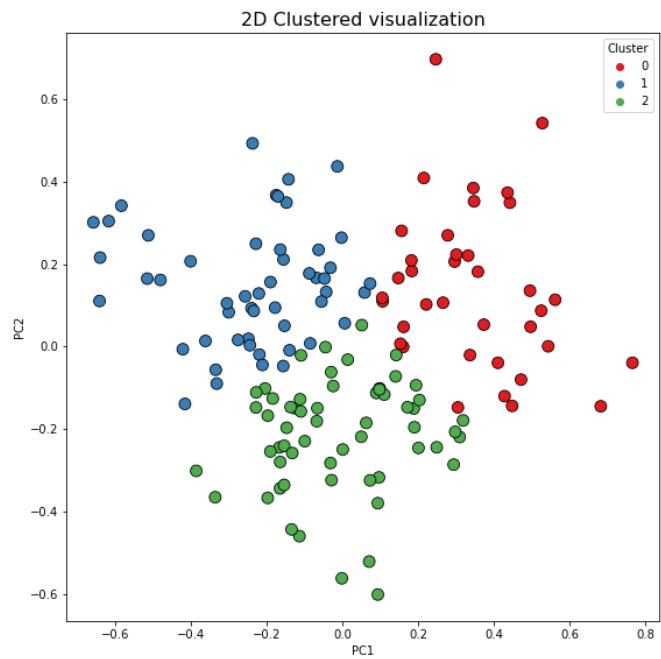
Αφού δοκιμάστηκαν πολλαπλά πλήθη k , η εξηγούμενη πληροφορία ή αλλιώς «αδράνεια» για κάθε k , απεικονίστηκε στην Εικόνα 22. Προφανές είναι ότι όσο αυξάνει ο αριθμός k μειώνεται η αδράνεια αλλά κάθε συστάδα παρέχει λιγότερη πληροφορία. Το σημείο καμπής (elbow) ($k = 3$) ενδείκνυται ως πλήθος συστάδων σύμφωνα με το Elbow method, χωρίς όμως να παρουσιάζει έντονη γωνία και χωρίς να λαμβάνεται υπόψιν η ανθρώπινη αντίληψη και η φύση του προβλήματος.



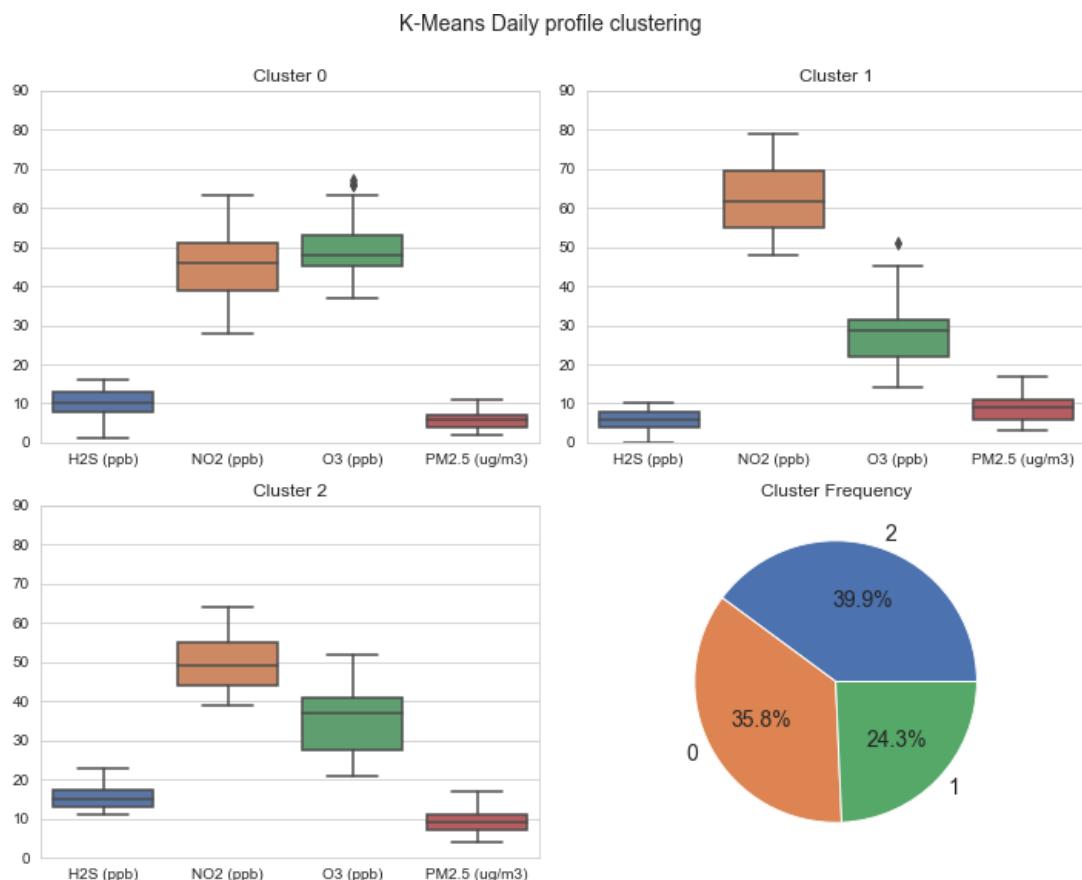
Εικόνα 23. Τρισδιάστατη απεικόνιση συσταδοποιημένων δεδομένων κατά K-Means (ΚΠΠΑ Kunak)

Η μέθοδος clustering K-Means όπως φαίνεται στην Εικόνα 23, ομαδοποιεί τα δεδομένα που παρουσιάζουν όμοια συμπεριφορά μεταξύ τους σε σχέση με τα υπόλοιπα. Οι συστάδες δεν προκύπτουν απόλυτα ανεξάρτητες μεταξύ τους λόγω της μορφής των δεδομένων.

Λαμβάνοντας μόνον τις δύο κύριες συνιστώσες, PC1, PC2 μπορεί να δημιουργηθεί δισδιάστατο διάγραμμα των συστάδων για λόγους ευκρίνειας.



Εικόνα 24. Δισδιάστατη απεικόνιση συσταδοποιημένων δεδομένων κατά K-Means (ΚΠΠΑ Kunak)



Εικόνα 25. Ημερήσια προφίλ ρύπανσης και συχνότητα εμφάνισής τους κατά K-Means (ΚΠΠΑ Kunak)

Κατόπιν ανάθεσης ετικετών ομάδας (labeling) στα στοιχεία της χρονοσειράς και αντιστροφής της κανονικοποίησης, εξήχθησαν τα ημερήσια προφίλ για τρία clusters (Εικόνα 25). Η συνολική εικόνα δείχνει να περιγράφει επαρκώς την συμπεριφορά των μεταβλητών όπως αυτή προσδιορίστηκε κατά την ανάλυση.

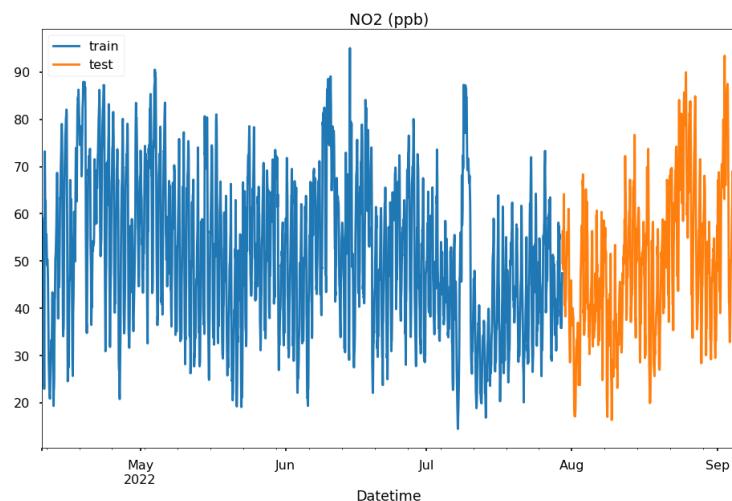
Κατά φθίνουσα συχνότητα εμφάνισης:

- Cluster 2: Παρατηρούνται οι υψηλότερες τιμές υδρόθειου συγκριτικά με τα υπόλοιπα προφίλ.
- Cluster 0: Η συγκέντρωση του όζοντος ξεπερνά αυτή του διοξειδίου του αζώτου. Υδρόθειο και αιωρούμενα σωματίδια σε χαμηλότερα επίπεδα.
- Cluster 1: Οι υψηλότερες παρατηρούμενες τιμές συγκέντρωσης NO_2 ενώ αντίστοιχα οι χαμηλότερες τιμές H_2S και O_3 .

Σημειώνεται ότι οι μέθοδοι K-Means και SOM αποφέρουν διαφορετικά αποτελέσματα για συσταδοποίηση.

3.3. Μοντελοποίηση δεδομένων πρόγνωσης

Αρχικά τα δεδομένα χρονοσειράς διαχωρίστηκαν σε υποσύνολα εκπαίδευσης (training – 75%) και επαλήθευσης (testing – 25%). (Εικόνα 26)

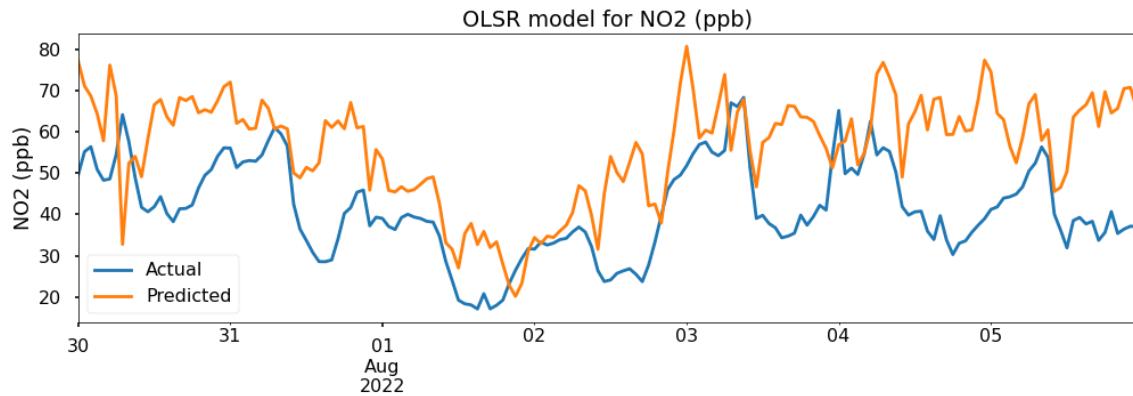


Εικόνα 26. Διαχωρισμός της χρονοσειράς σε τμήμα εκπαίδευσης και τμήμα αξιολόγησης.

Παρακάτω παρουσιάζεται η μοντελοποίηση πρόγνωσης ωριαίων τιμών διοξειδίου του αζώτου, για την τελευταία εβδομάδα της δειγματοληψίας (30 Αυγούστου έως 6 Σεπτεμβρίου 2022), με τις μεθόδους Παλινδρόμησης ελαχίστων τετραγώνων, μοντέλου SARIMAX και Random Forest Regressor. Έπειτα παρουσιάζονται τα αποτελέσματα και η αξιολόγηση επίδοσης των μοντέλων SARIMAX και RF για κυλιόμενο παράθυρο δείγματος και αυξανόμενο δείγμα.

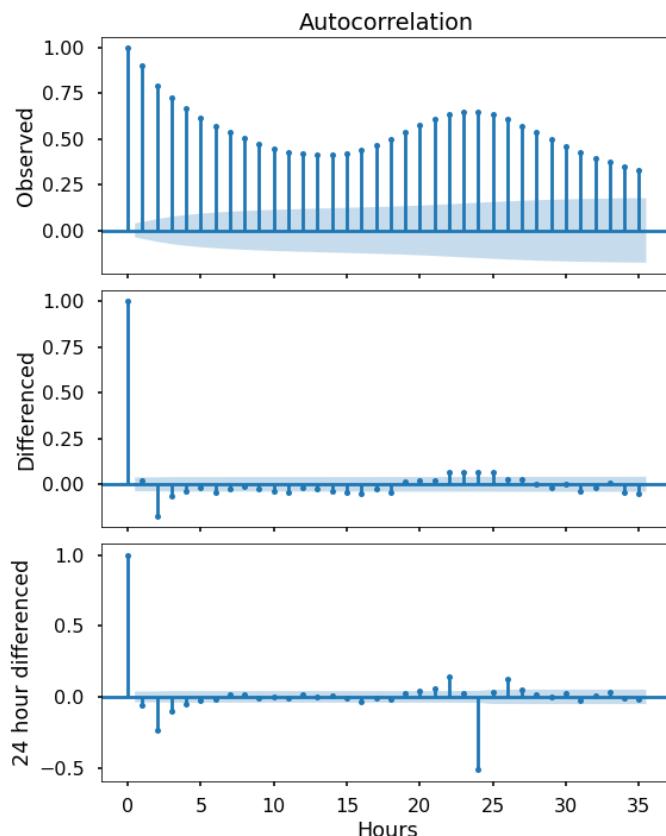
3.3.1. Παλινδρόμηση ελαχίστων τετραγώνων – SARIMAX

Κατόπιν υλοποίησης απλής παλινδρόμησης ελαχίστων τετραγώνων, για διάστημα μίας εβδομάδας, είναι εμφανές ότι η προσέγγιση δεν είναι ικανοποιητική (Εικόνα 27).



Εικόνα 27. Σύγκριση αποτελεσμάτων πρόγνωσης OLSR και πραγματικών δεδομένων (ΚΠΠΑ Kunak).

Εξετάστηκε εάν τα υπολειπόμενα στοιχεία της παλινδρόμησης (residual errors) είναι πράγματι αυτοσυσχετιζόμενα.



Εικόνα 28. Διαγράμματα αυτοσυσχέτισης αρχικής χρονοσειράς, διαφορισμένης κατά ένα βήμα και διαφορισμένης κατά 24 βήματα.

Στην πρώτη γραμμή της Εικόνα 28, διακρίνεται έντονη αυτοσυσχέτιση της αρχικής χρονοσειράς, για ένα βήμα υστέρησης (Lag 1) και στη συνέχεια για είκοσι τέσσερα (Lag 24). Διαφορίζοντας κατά 1 ώρα ($d=1$), η αυτοσυσχέτιση στο 1^o βήμα είναι οριακά μηδενική ($p=0$). Περαιτέρω διαφόριση κατά 24 βήματα για εξάλειψη της περιοδικής αυτοπαλινδρόμησης, όπως διακρίνεται στην 3^η σειρά του Εικόνα 28, δίνει πιο επίπεδη καμπύλη με έντονη αυτοσυσχέτιση στο 1^o εικοσιτετράωρο βήμα ($P=1$ για βήμα $m=24$ με $D=0$). Η ύπαρξη τάσης, όπως διαπιστώθηκε κατά την αποσύνθεση της χρονοσειράς, απαιτεί μοντελοποίηση κινούμενων μέσων όρων για το βασικό και για το εποχικό μοντέλο ($q=1$, $Q=1$).

Οι παρατηρήσεις αυτές οδηγούν στον προσδιορισμό των παραμέτρων του μοντέλου SARIMAX, όπως στην ενότητα [2.2.2.2](#). Δηλαδή:

Πίνακας 7. Παράμετροι μοντέλου SARIMAX

Παράμετρος	Περιγραφή	Τιμή
p	Αυτοπαλινδρόμηση	0
d	Διαφόριση	1
q	Κινούμενοι Μέσοι	1
P	Εποχική Αυτοπαλινδρόμηση	1
D	Εποχική Διαφόριση	0
Q	Εποχικοί Κινούμενοι Μέσοι	1
m	Περιοδικότητα	24

Στον Πίνακας 8 παρουσιάζονται οι συντελεστές της παλινδρόμησης, όπως προέκυψαν για την χρονοσειρά NO_2 . Στις πρώτες γραμμές καταγράφονται οι μεταβλητές εισόδου. Οι μεταβλητές των αιωρούμενων σωματιδίων δεν ήταν στατιστικά σημαντικές για το μοντέλο της παλινδρόμησης, με p -value > 0.05 και για αυτό παραλήφθηκαν. Η σημαντικότητα των παραμέτρων $(0, 1, 1)x(1, 0, 1, 24)$ εκφράζεται από το οριακά μηδενικό p -value και τους υψηλούς σχετικά συντελεστές παλινδρόμησης (coefficient).

Πίνακας 8. Συντελεστές παλινδρόμησης SARIMAX για πρόγνωση NO_2 (ΚΠΠΑ Kunak)

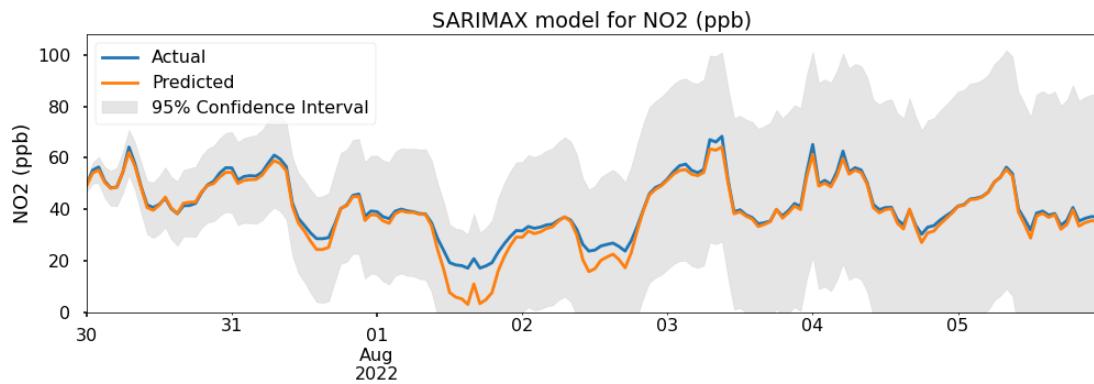
	Coefficient	Std error	Z	P> z
Dew Point (C)	2.7593	0.022	126.062	0.000
H2S (ppb)	0.0204	0.004	4.941	0.000
Relative Humidity (%)	-0.0267	0.006	-4.394	0.000
O3 (ppb)	-0.0255	0.004	-6.638	0.000
Temp (C)	-2.7624	0.020	-139.195	0.000
ma.L1	0.3084	0.012	26.151	0.000
ar.S.L24	0.9844	0.005	198.434	0.000
ma.S.L24	-0.9252	0.012	-76.012	0.000
sigma2	1.5893	0.025	63.990	0.000

Όπου:

ma.L1: Moving Average at Lag 1

ar.S.L24: Auto Regression Seasonal at Lag 24

ma.S.L24: Moving Average Seasonal at Lag 24



Εικόνα 29. Σύγκριση αποτελεσμάτων πρόγνωσης SARIMAX και πραγματικών δεδομένων (ΚΠΠΑ Kunak)

Έτσι, κατόπιν εφαρμογής του παραμετροποιημένου μοντέλου SARIMAX προκύπτει επιτυχής προσέγγιση της πραγματικής χρονοσειράς, εκτός σύντομων υποεκτιμήσεων, όπως φαίνεται στην Εικόνα 29. Το διάστημα εμπιστοσύνης 95% διευρύνεται όσο μακροπρόθεσμα δοκιμάζεται η πρόγνωση, κάτι αναμενόμενο λόγω της αδυναμίας διατήρησης του ίδιου επιπέδου ακρίβειας σε βάθος χρόνου πρόβλεψης.

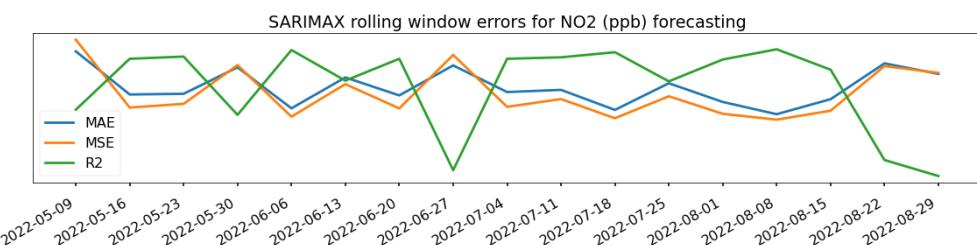
Κυλιόμενο παράθυρο δείγματος

Ορίστηκε κυλιόμενο παράθυρο δείγματος ενός μήνα, το οποίο μεταφέρεται με βήμα μίας εβδομάδας, από την αρχή έως το τέλος του συνολικού δείγματος. Στόχος είναι η πρόβλεψη του επόμενου εικοσιτετράωρου από το τέλος του δειγματικού παραθύρου. Αξιολογήθηκε η ευστοχία πρόγνωσης μέσω προσδιορισμού των σφαλμάτων (Πίνακας 9).

Πίνακας 9. Παράμετροι αξιολόγησης SARIMAX κυλιόμενου δείγματος (Kunak)

Προβλεπόμενο Εικοσιτετράωρο	MAE	MSE	MAPE	R2
2022-05-09	9.52	111.26	0.15	0.15
2022-05-16	3.6	18.0	0.08	0.85
2022-05-23	3.69	23.08	0.15	0.88
2022-05-30	7.32	76.27	0.13	0.08
2022-06-06	1.7	5.46	0.04	0.97
2022-06-13	5.94	50.22	0.09	0.55
2022-06-20	3.47	16.82	0.06	0.85
2022-06-27	7.59	90.29	0.16	-0.68
2022-07-04	3.93	19.0	0.09	0.85
2022-07-11	4.23	29.68	0.5	0.87
2022-07-18	1.48	3.34	0.04	0.94
2022-07-25	5.14	33.53	0.13	0.54
2022-08-01	2.57	9.42	0.06	0.84
2022-08-08	0.89	1.41	0.02	0.98
2022-08-15	2.95	13.65	0.08	0.7
2022-08-22	7.87	75.01	0.18	-0.54
2022-08-29	6.42	65.63	0.09	-0.76

Η «ευστοχία» του μοντέλου για διαφορετικά διαστήματα δεν διατηρείται σταθερή (Εικόνα 30). Η αστάθεια απόδοσης οφείλεται στην ανάγκη προσαρμογής των παραμέτρων του μοντέλου SARIMAX, έπειτα από διερεύνηση της συμπεριφοράς σφαλμάτων της παλινδρόμησης, από τον χρήστη. Στην παρούσα δοκιμή κυλιόμενου δειγματικού παραθύρου διατηρήθηκαν σταθερές οι παράμετροι « p, d, q, P, D, Q, m » ως αντιτροσωπευτικές του συνόλου της χρονοσειράς. Για κάποια χρονικά διαστήματα όμως, δεν είναι κατάλληλες, γεγονός που οδηγεί σε εσφαλμένη πρόγνωση. Θυσιάζοντας συνολική ακρίβεια, είναι εφικτό να προσαρμοστεί το μοντέλο ώστε να αποφευχθούν έντονες αστοχίες, εφόσον δεν δύναται να επέμβει ο χρήστης σε κάθε νέα δοκιμή πρόγνωσης.



Εικόνα 30. Διακύμανση παραμέτρων αξιολόγησης SARIMAX κυλιόμενου δείγματος (ΚΠΠΑ Kunak). (Τιμές υπό κλίμακα για λόγους ευκρίνειας σύγκρισης).

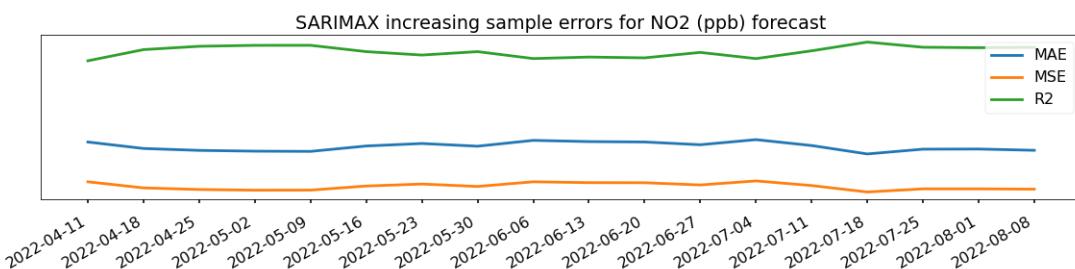
Αυξανόμενο δείγμα

Ορίστηκε ως αρχικό δείγμα η τελευταία εβδομάδα της δειγματοληψίας, το οποίο αυξήθηκε επαναληπτικά κατά μία εβδομάδα, με στόχο πρόγνωσης τα τρία τελευταία εικοσιτετράωρα μέσω μοντέλου SARIMAX.

Πίνακας 10. Παράμετροι αξιολόγησης SARIMAX αυξανόμενου δείγματος (Kunak)

Ημερομηνία έναρξης δείγματος	MAE	MSE	MAPE	R2
2022-08-08	3.38	15.69	0.07	0.82
2022-08-01	3.44	15.82	0.07	0.82
2022-07-25	3.44	15.82	0.07	0.82
2022-07-18	3.21	14.35	0.07	0.84
2022-07-11	3.61	17.37	0.08	0.8
2022-07-04	3.88	19.53	0.08	0.77
2022-06-27	3.65	17.64	0.08	0.8
2022-06-20	3.77	18.68	0.08	0.77
2022-06-13	3.79	18.72	0.08	0.77
2022-06-06	3.85	19.12	0.08	0.77
2022-05-30	3.58	16.94	0.07	0.8
2022-05-23	3.7	18.07	0.08	0.78
2022-05-16	3.59	17.15	0.08	0.8
2022-05-09	3.33	15.24	0.07	0.83
2022-05-02	3.34	15.19	0.07	0.83
2022-04-25	3.38	15.53	0.07	0.82
2022-04-18	3.47	16.27	0.07	0.81
2022-04-11	3.77	19.12	0.08	0.76

Παρατηρείται ότι η «ευστοχία» του μοντέλου παραμένει σχετικά σταθερή όσο αυξάνεται το μέγεθος δείγματος. Το χρονικό διάστημα πρόγνωσης των τριών τελευταίων εικοσιτετράωρων δεν μεταβάλλεται, όπως και οι παράμετροι του μοντέλου SARIMAX. Οι τελευταίες προσδιορίστηκαν από το σύνολο των δεδομένων και όχι από το εκάστοτε δείγμα καθώς η διαδικασία απαιτεί παρέμβαση του χρήστη.



Εικόνα 31. Διακύμανση παραμέτρων αξιολόγησης SARIMAX αυξανόμενου δείγματος (ΚΠΠΑ Kunak). (Τιμές υπό κλίμακα για λόγους ευκρίνειας σύγκρισης).

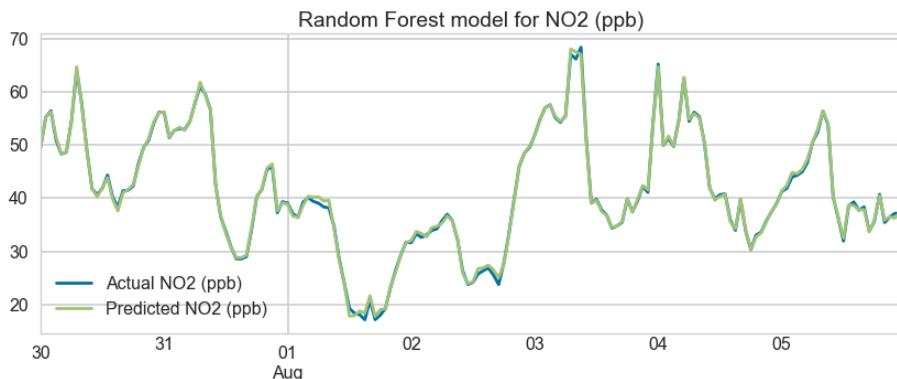
3.3.2. Random Forest Regressor

Καταστρώθηκε το μοντέλο μηχανικής μάθησης Random Forest με αξιολόγηση εκπαίδευσης 10-fold cross validation, προσέγγισης Time Series Split, όπως αναπτύχθηκε στην ενότητα 2.2.2.3. Τα αποτελέσματα παρουσιάζονται στον Πίνακας 11.

Πίνακας 11. Αποτελέσματα 10-fold cross validation εκπαίδευσης μοντέλου Random Forest

Fold	MAE	MSE	MAPE	R2
0	6.9440	75.5847	8.6939	0.7071
1	2.5124	15.2510	3.9052	0.9259
2	3.1432	22.1298	4.7042	0.9012
3	5.0731	43.1242	6.5669	0.8039
4	3.0256	17.5144	4.1850	0.9202
5	2.1392	12.3434	3.5133	0.9524
6	1.5352	5.8705	2.4229	0.9671
7	1.0778	3.5327	1.8795	0.9855
8	0.8495	2.1059	1.4512	0.9868
9	0.5873	0.7609	0.8723	0.9937
Mean	2.6887	19.8218	3.8195	0.9144
Std	1.9018	22.0842	2.2877	0.0874

Υλοποιήθηκε η πρόγνωση της τελευταίας εβδομάδας δειγματοληψίας και η σύγκριση με τα πραγματικά δεδομένα του ίδιου διαστήματος φαίνεται στην Εικόνα 32.



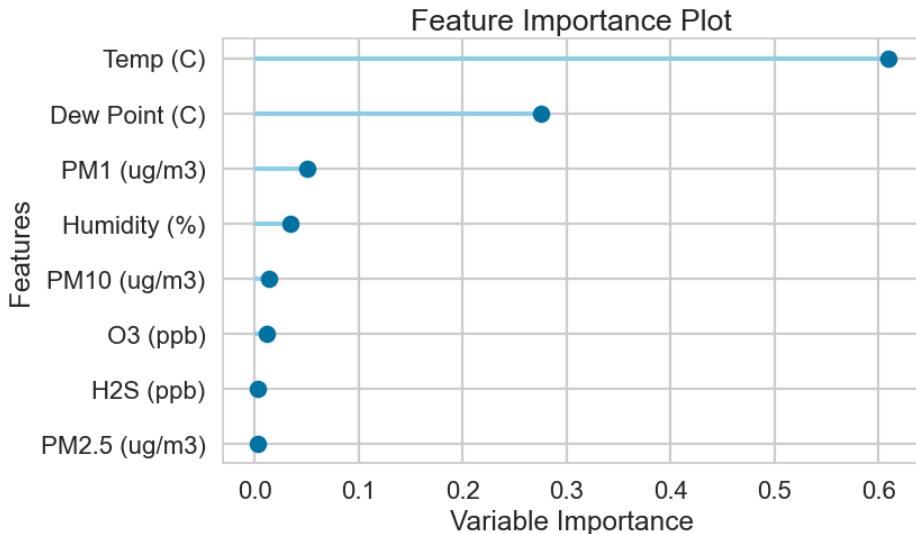
Εικόνα 32. Σύγκριση αποτελεσμάτων πρόγνωσης Random Forest και πραγματικών δεδομένων (ΚΠΠΑ Kunak)

Η πρόγνωση του μοντέλου κρίνεται ικανοποιητική όπως φαίνεται και από τις παραμέτρους αξιολόγησης του Πίνακας 12.

Πίνακας 12. Παράμετροι αξιολόγησης πρόγνωσης μοντέλου Random Forest (Kunak)

Random Forest	
MAE	0.337
MSE	0.211
MAPE	0.010
R2	0.998

Επισημαίνεται ότι δεν συμμετέχουν όλες οι μεταβλητές εισόδου με την ίδια βαρύτητα κατά την εκπαίδευση του μοντέλου μηχανικής μάθησης. Όπως παρατηρείται στην Εικόνα 33, οι τιμές θερμοκρασίας και σημείου δρόσου συμμετέχουν με αθροιστική σημαντικότητα 90%.



Εικόνα 33. Βαρύτητα μεταβλητών στην διαμόρφωση του μοντέλου Random Forest (ΚΠΠΑ Kunak)

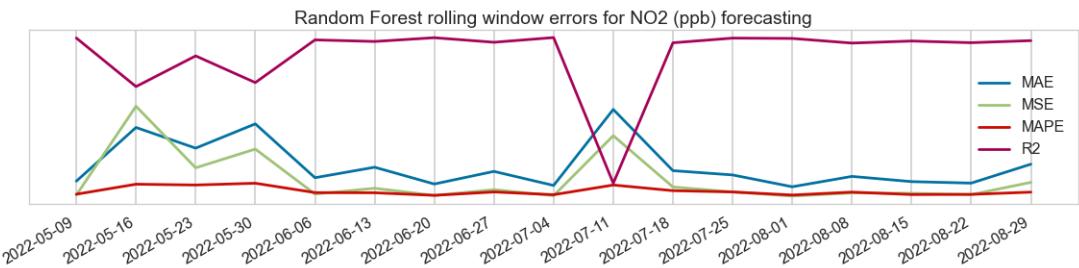
Κυλιόμενο δείγμα

Ορίστηκε κυλιόμενο δείγμα όπως περιγράφεται στην ενότητα 2.2.2.

Πίνακας 13. Παράμετροι αξιολόγησης μοντέλου RF κυλιόμενου δείγματος (Kunak)

Προβλεπόμενο Εικοσιτετράρο	MAE	MSE	MAPE	R2
2022-05-09	0.998721	1.400794	0.019	0.986109
2022-05-16	4.323429	56.373420	0.081	0.685418
2022-05-23	3.041946	18.246970	0.076	0.875362
2022-05-30	4.546058	29.849157	0.087	0.710242
2022-06-06	1.211850	2.027644	0.029	0.974636
2022-06-13	1.861409	5.579424	0.028	0.964700
2022-06-20	0.821608	1.147353	0.012	0.988038
2022-06-27	1.603054	4.720434	0.033	0.959798
2022-07-04	0.733079	1.028628	0.016	0.989014
2022-07-11	5.439854	38.152441	0.076	0.086714
2022-07-18	1.648967	6.367408	0.041	0.956533
2022-07-25	1.383383	3.316237	0.033	0.985530
2022-08-01	0.651012	0.731154	0.014	0.983474
2022-08-08	1.293163	2.720894	0.032	0.954827
2022-08-15	0.969754	2.465276	0.017	0.967483
2022-08-22	0.877617	1.673521	0.018	0.957145
2022-08-29	2.049621	9.242711	0.032	0.969947

Η ευστοχία πρόγνωσης του μοντέλου Random Forest, για NO_2 , διατηρείται σχετικά σταθερή για όλο το διάστημα δοκιμών, εκτός κάποιων ανωμαλιών όπως στις 11/7/2022. Δεν απαιτείται παρέμβαση του χρήστη καθώς το μοντέλο μηχανικής μάθησης εκπαιδεύεται εκ νέου όταν εφαρμόζεται σε διαφορετικά δεδομένα.



Εικόνα 34. Διακύμανση παραμέτρων αξιολόγησης RF κυλιόμενου δείγματος (ΚΠΠΑ Kunak). (Τιμές υπό κλίμακα για λόγους ευκρίνειας σύγκρισης).

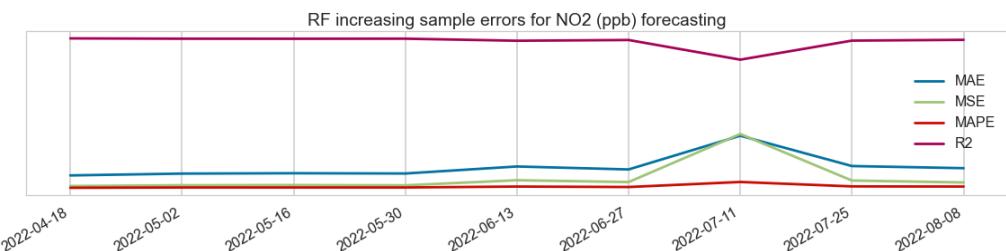
Αυξανόμενο δείγμα

Ορίστηκε ως αρχικό δείγμα η τελευταία εβδομάδα της δειγματοληψίας, το οποίο αυξήθηκε επαναληπτικά κατά μία εβδομάδα, με στόχο πρόγνωσης τα τρία τελευταία εικοσιτετράωρα μέσω μοντέλου Random Forest.

Πίνακας 14. Παράμετροι αξιολόγησης RF αυξανόμενου δείγματος (Kunak)

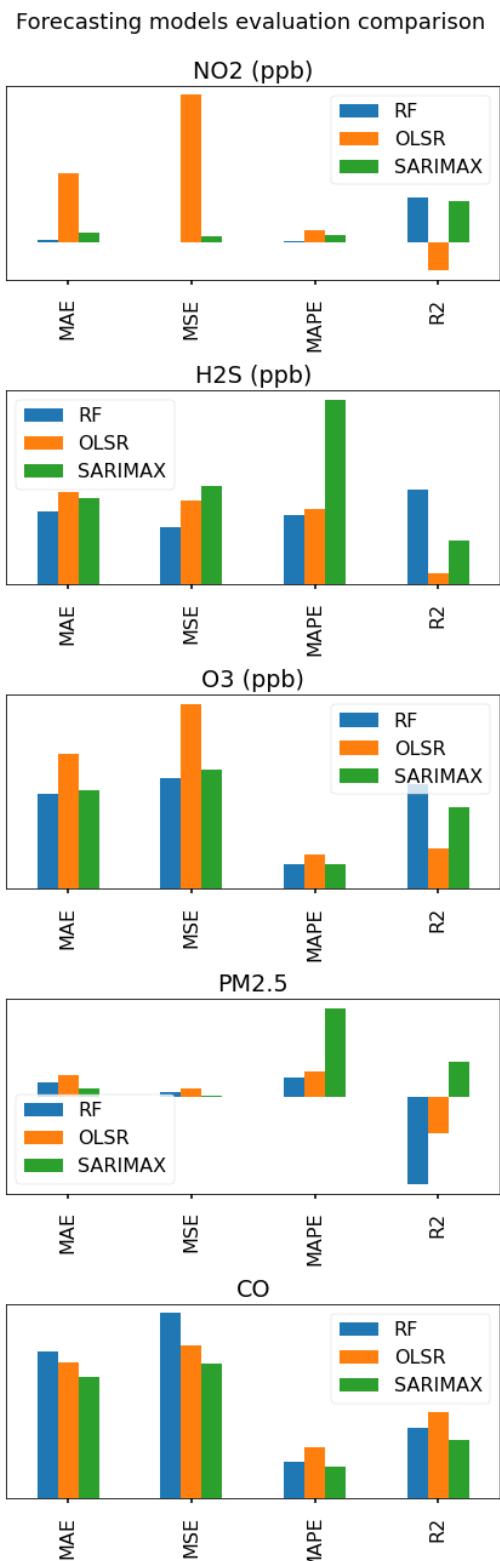
Ημερομηνία έναρξης δείγματος	MAE	MSE	MAPE	R2
2022-08-08	1.41	4.77	0.02	0.980
2022-07-25	1.55	6.03	0.02	0.975
2022-07-11	3.54	36.47	0.05	0.851
2022-06-27	1.33	5.05	0.02	0.979
2022-06-13	1.52	6.23	0.02	0.975
2022-05-30	1.06	2.92	0.02	0.988
2022-05-16	1.08	3.06	0.02	0.987
2022-05-02	1.06	2.97	0.02	0.988
2022-04-18	0.94	2.46	0.01	0.990

Κατά την αύξηση του δείγματος δεν παρατηρείται κάποια βελτίωση της απόδοσης.



Εικόνα 35. Διακύμανση παραμέτρων αξιολόγησης RF αυξανόμενου δείγματος (ΚΠΠΑ Kunak). (Τιμές υπό κλίμακα για λόγους ευκρίνειας σύγκρισης).

3.3.3. Σύγκριση μοντέλων πρόγνωσης



Εικόνα 36. Σύγκριση μοντέλων για κάθε ρύπο
(τιμές υπό κλίμακα για λόγους ευκρίνειας)

Τα μοντέλα πρόγνωσης OLSR, SARIMAX και RF δεν απέδωσαν εξίσου σε κάθε ρύπο (Εικόνα 36). Η απλή παλινδρόμηση ελαχίστων τετραγώνων ως η πιο βασική αλλά και γρήγορη μέθοδος εμφανίζει τα υψηλότερα σφάλματα και τον χαμηλότερο συντελεστή προσδιορισμού. Το μοντέλο Random Forest παρουσίασε υψηλότερη ευστοχία συνολικά, εκτός της πρόγνωσης για τις τιμές του CO. Συγκεκριμένα:

- NO₂: Έντονη διαφορά μεταξύ OLSR και SARIMAX, RF, με τα τελευταία να αποδίδουν εξίσου εύστοχα.
- H₂S: Επικρατέστερο το μοντέλο RF, κυρίως στον συντελεστή R².
- O₃: Ήπιες διαφορές μεταξύ των μοντέλων, πιο αποτελεσματικό το μοντέλο που βασίζεται σε RF.
- PM_{2.5}: Δυσκολία πρόγνωσης ακριβείας¹ της χαμηλής διακύμανσης των τιμών. Καλύτερο το μοντέλο SARIMAX, έπειτα από λεπτομερή προσαρμογή, πιθανό overfitting.
- CO: Παρόμοια απόδοση και των τριών μοντέλων. Μοναδική περίπτωση όπου το OLSR υπερτερεύει του RF.

¹ Κατά την διαμόρφωση των μοντέλων δεν ελήφθησαν υπόψιν οι μεταβλητές PM1, PM10. Η πρόγνωση βασίστηκε στα υπόλοιπα μεγέθη.

4. Συμπεράσματα

Η επιστράτευση οργάνων χαμηλού κόστους δύναται να προσδώσει επαρκές πληροφοριακό περιεχόμενο σχετικά με την ποιότητα αέρα αστικών περιοχών, κατόπιν μεθοδικής επεξεργασίας των δεδομένων και εξαγωγής αποτελεσμάτων. Με την ολοκλήρωση της μελέτης τα συμπεράσματα που προκύπτουν καταγράφονται ως εξής:

Όσον αφορά τη σύγκριση μεταξύ κόμβων, παρατηρήθηκαν αποκλίσεις μεταξύ επιμέρους συσκευών, κατά την εκτίμηση των επιπέδων συγκεντρώσεων αέριων ρύπων και σωματιδίων. Οι διαφορές εντοπίστηκαν κυρίως στην αριθμητική στάθμη των επιπέδων ενώ η διακύμανση και η τυπική απόκλιση παρουσίασαν όμοια συμπεριφορά. Καταγράφηκε σημαντική θετική συσχέτιση μεταξύ μετρήσεων διαφορετικών ΚΠΠΑ για συγκεντρώσεις Ο₃ και αιωρούμενων σωματιδίων και αντίστοιχα ήπια θετική συσχέτιση για συγκεντρώσεις ΝΟ₂. Οι αναφερθείσες αποκλίσεις ενδεχομένως οφείλονται στην διαφορετική τεχνολογική προσέγγιση των κατασκευαστών.

Κατά την συμπλήρωση ελλειπουσών τιμών, εφαρμόζοντας την κατάλληλη μέθοδο, στατιστική είτε μηχανικής μάθησης επετεύχθη αποδεκτής ακρίβειας συμπλήρωση ελλείψεων της χρονοσειράς. Ιδιαίτερα για χαμηλά ποσοστά ελλειπουσών τιμών στο σύνολο των δεδομένων, οι τεχνικές που δοκιμάστηκαν καθιστούν σχετικά ασφαλή την εξαγωγή συμπερασμάτων παρά τις οποιεσδήποτε παύσεις λειτουργίας των οργάνων.

Η διακύμανση των τιμών των συγκεντρώσεων ρύπων κατά τη διάρκεια τυπικού εικοσιτετράρου και τυπικής εβδομάδας καθώς και η περιοδικότητα που προσδιορίστηκε για κάθε μέγεθος, εκφράζουν πιθανή συσχέτιση με τον ρυθμό τυπικής αστικής ανθρώπινης δραστηριότητας, όπως η κυκλοφορία οχημάτων και η λειτουργία εγκαταστάσεων.

Τα παραπάνω συμπεραίνονται και από τα αποτελέσματα διερεύνησης συσχέτισης. Ρύποι όπως το υδρόθειο συσχετίζονται θετικά με το μονοξείδιο του άνθρακα και το διοξείδιο του αζώτου, υποδεικνύοντας πιθανότητα σύνδεσης με καύση υδρογονανθράκων. Μετεωρολογικές συνθήκες όπως η θερμοκρασία και το σημείο δρόσου έδειξαν να έχουν άμεση επίδραση στα επίπεδα συγκέντρωσης αέριων ρύπων και σωματιδίων ενώ ο άνεμος στη διασπορά και μεταφορά αυτών από την πηγή προς το σημείο μέτρησης.

Όσον αφορά το υδρόθειο, προσδιορίστηκαν οι συνθήκες για τις οποίες παρατηρήθηκαν τιμές υψηλότερες του επιπέδου αντιληπτής δυσοσμίας και το ποσοστό αυτών των τιμών σε σχέση με το σύνολο δείγματος, το οποίο κρίνεται μη αμελητέο.

Κατόπιν μείωσης των διαστάσεων και συσταδοποίησης των δεδομένων εξήχθησαν ημερήσια προφίλ ρύπανσης. Στα διαγράμματα που σχηματίστηκαν αποτυπώνονται διακριτά μοτίβα συγκεντρώσεων και η συχνότητα εμφάνισής τους στο δειγματοληπτικό διάστημα, παρέχοντας πληροφορία για την τυπική εικόνα της ποιότητας αέρα για την αστική περιοχή.

Με τα διαγράμματα της περιγραφικής στατιστικής σχηματίστηκε εικόνα του συνόλου των μετρήσεων και της κατανομής τους ενώ με τη διάσπαση των χρονοσειρών λήφθηκαν τα επιμέρους συνθετικά τους, προσδιορίζοντας έτσι την τάση και την περιοδικότητα των χρονοσειρών.

Η μελέτη των παραπάνω και ο προσδιορισμός στατιστικών παραμέτρων κατέστησαν δυνατή την κατάστρωση στατιστικού μοντέλου και μοντέλου μηχανικής μάθησης,

πρόγνωσης μελλοντικών τιμών. Όπως εκφράζεται από τους συντελεστές αξιολόγησης σφάλματος, η επίδοσή τους κρίθηκε συνολικά ικανοποιητική, με μεταβλητή ευστοχία για διαφορετικά μεγέθη και χρονικά παράθυρα πρόβλεψης.

Περεταίρω έρευνα στο ίδιο πλαίσιο θα μπορούσε να πραγματοποιηθεί, λαμβάνοντας εκτενέστερο, χρονικά, δείγμα. Εφαρμογή των παρόντων μεθόδων για σύγκριση μεταξύ συνθηκών διαφορετικής εποχής, όπως χειμώνας-καλοκαίρι. Επίσης πιθανή μελέτη επιπτώσεων σε σχέση με την ενεργειακή κρίση, τις διαφορετικές πηγές ενέργειας και την μεταβαλλόμενη ανθρώπινη δραστηριότητα ανάλογα με τον ρυθμό διαβίωσης. Επιπλέον, πιθανή εκμετάλλευση δεδομένων ΚΠΠΑ τοποθετημένων σε βιομηχανικές ζώνες, κομβικά σημεία κυκλοφορίας οχημάτων κ.α. Θα καθιστούσε δυνατή την διερεύνηση συσχέτισης των επιπέδων συγκεντρώσεων στο κέντρο της πόλης συγκριτικά με τις τυπικές πηγές ρύπανσης.

5. Βιβλιογραφία

- Airly Sp. Z o.o. (2022). ul.Mogilska 43, Palo Alto, CA 94304-1115, 31-545 Kraków, Poland.
- Akpan, E. A., & Moffat, I. U. (2018). Modeling the Autocorrelated Errors in Time Series Regression:. *Journal of Advances in Mathematics and Computer Science* 26(4), 1-15.
- Ali, L., & J., B. H. (2015). Random forests model for one day ahead load forecasting. *6th International Renewable Energy Congress, IREC 2015*.
- Ali, M. (2020, April). *PyCaret: An open source, low-code machine learning library in Python - version 1.0.0*. Retrieved from <https://www.pycaret.org>
- AQMash Environmental Instruments Ltd. (2022). United Kingdom.
- Box, G., & Jenkins, G. (1976). *Time series analysis, forecasting and control, revised edn*. San Francisco: Holden Day.
- Breiman, L. (2001). Random Forests. *Machine Learning* 45 (σσ. 5-32). Netherlands: 2001 Kluwer Academic Publishers.
- Brigham, E. O. (1988). *The fast Fourier transform and its applications*. USA: Prentice-Hall, Inc.
- Castell, N., Dauge, F. R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., . . . Bartonova, A. (2017). Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International* 99, 293-302.
- Davis Instruments Corporation. (2022). Hayward, 3465 Diablo Avenue, CA 94545, USA.
- DIRECTIVE 2008/50/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 21 May 2008 on ambient air quality and cleaner air for Europe. (χ.χ.).
- Dudek, G. (2015, August). STD: A Seasonal-Trend-Dispersion. *JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8*.
- Dyer, S., & Dyer, J. (2001). Cubic-spline interpolation. 1. *IEEE Instrumentation & Measurement Magazine* 4(1), 44-46.
- EEA. (2022, April 01). *Europe's air quality status 2022*. Ανάκτηση από European Environment Agency: <https://www.eea.europa.eu/publications/status-of-air-quality-in-Europe-2022/europees-air-quality-status-2022>
- EPA. (2022, June 24). *EPA scientists develop and evaluate Federal Reference & Equivalent Methods for measuring key air pollutants*. Ανάκτηση από United States Environmental Protection Agency: <https://www.epa.gov/air-research/epa-scientists-develop-and-evaluate-federal-reference-equivalent-methods-measuring-key>
- Fix, E., & Hodges, J. L. (1951). *Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4*. Randolph Field, Texas: USAF School of Aviation Medicine.
- Flores, J. H., Engel, P. M., & Pinto, R. C. (2012). Autocorrelation and partial autocorrelation functions to improve neural networks models on univariate time series forecasting.

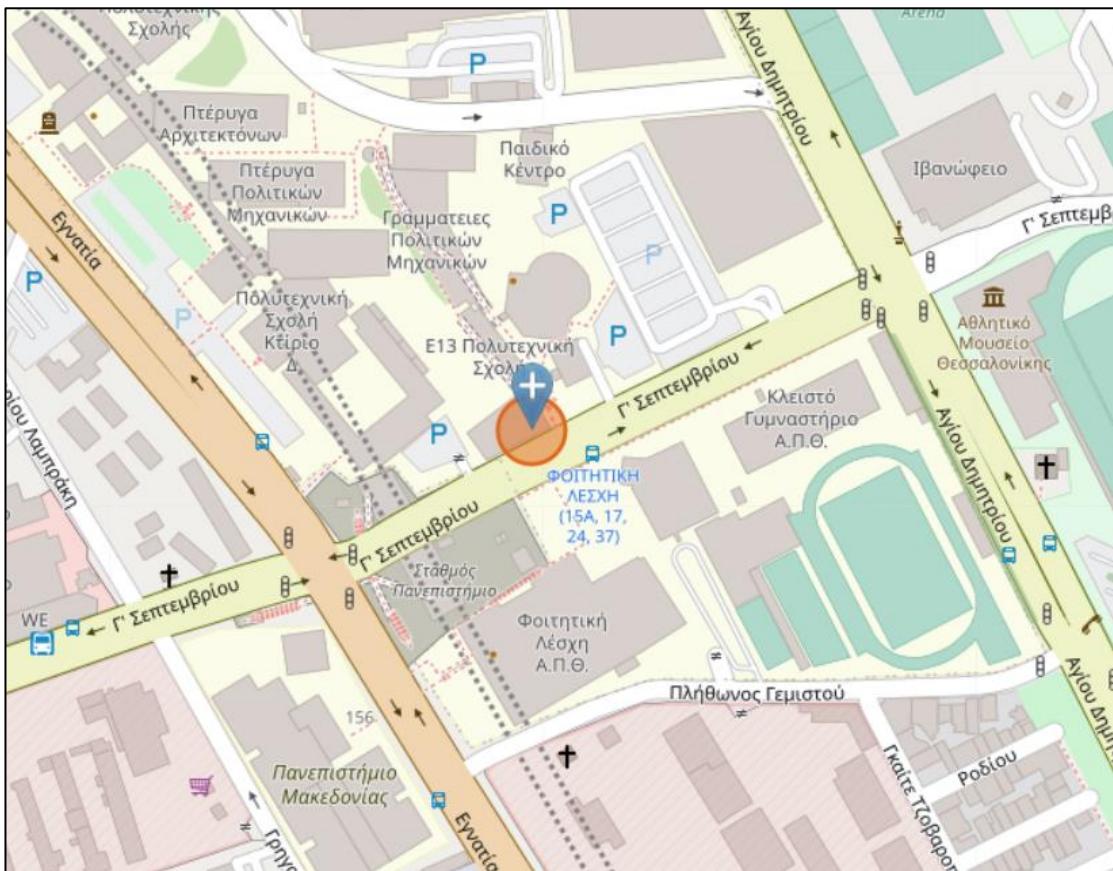
Neural Networks (IJCNN), The 2012 International Joint Conference on Incremental Gaussian Mixture Models.

- Frederickson, B., L., Sidaraviciute, R., Schmidt, A., J., . . . S., M. (2022). Are dense networks of low-cost nodes really useful for monitoring air pollution? A case study in Staffordshire. *Atmos. Chem. Phys.*, 22, 13949–13965.
- Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics (international student edition) 4th Edn.* New York: WW Norton & Company.
- Gocheva-Illieva, Ivanov, S. a., A., Voynikova, Desislava, Boyadzhiev, & Doychin. (2013). Time series analysis and forecasting for air pollution in small urban area: An SARIMA and factor analysis approach. *Stochastic Environmental Research and Risk Assessment* vol. 28, 1045-1060.
- Harris, C., Millman, K., v. d., & S.J. (2020). Array programming with NumPy. *Nature* 585, 357–362.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering Volume* 9, 90-95.
- Junninen, H., Niska, H., & Tuppurainenc, K. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment* 38, 2895–2907.
- Karatzas, K. (2022). *Air Quality (Micro) Sensors*. Thessaloniki: Aristotle University of Thessaloniki.
- KASTOM. (2022). KASTOM Innovative system for air quality monitoring & forecasting. Thessaloniki.
- Katsifarakis, N., & Karatzas, K. (2017). A New Feature Selection Methodology for Environmental Modelling Support: The Case of Thessaloniki Air Quality. *Hřebíček, J., Denzer, R., Schimak, G., Pitner, T. (eds) Environmental Software Systems. Computer Science for Environmental Protection. ISESS 2017. IFIP Advances in Information and Communication Technology, vol 507 (σσ. 61–70)*. Springer, Cham.
- Kluyver, T., Ragan-Kelley, B., Fern, Perez, o., Granger, B., Bussonnier, M., . . . Willing, C. (2016). Jupyter Notebooks -- a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (σσ. 87-90). F. Loizides and B. Schmidt - IOS Press. Ανάτηση από Positioning and Power in Academic Publishing: Players, Agents and Agendas.
- Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, vol. 78, no. 9, 1464-1480.
- Kourtidis, K., Kelesis, A., & Petrakakis, M. (2008). Hydrogen sulfide (H₂S) in urban ambient air. *Atmospheric Environment Volume* 42, Issue 32, σσ. 7476-7482.
- Kularatna, N., & Sudantha, B. H. (2008). An Environmental Air Pollution Monitoring System. *IEEE SENSORS JOURNAL, VOL. 8, NO. 4*, 415-422.
- Kumar, Anuj, Srivastava, Vishal, Singh, Kumar, M., . . . P., G. (2015). Current Status of the IEEE 1451 Standard-Based Sensor Applications. *IEEE Sensors Journal* 15(5), 2505-2513.

- Kunak Technologies S.L. (2022). P. E. La Muga 9, Planta 4, Oficina 1, 31160, Orcoyen - Navarra, Spain.
- Lloyd, & P., S. (1982, 2 28). Least squares quantization in PCM. *Information Theory, IEEE Transactions*, σσ. 129-137.
- Makri, A., & Stilianakis, N. I. (2008). Vulnerability to air pollution health effects. *International Journal of Hygiene and Environmental Health* 211(3-4), 326-336.
- Marutho, D., Handaka, S. H., Wijaya, E., & Muljono. (2018). The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. *International Seminar on Application for Technology of Information and Communication*, 553-538.
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. *PROC. OF THE 9th PYTHON IN SCIENCE CONF.*, 56 - 61.
- OpenStreetMap contributors. (2017). *Map data copyrighted OpenStreetMap*. Ανάκτηση από <https://www.openstreetmap.org>
- Padhma, M. (2021, October 28). *End-to-End Introduction to Evaluating Regression Models*. Ανάκτηση από Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/#h2_3
- Pearce, o. L., Waller, L. A., Chang, H. H., Klein, M., Sarnat, J. A., Sarnat, S. E., . . . Mulholland, J. A. (2014, July 03). Using self-organizing maps to develop ambient air quality classifications: a time series example. *Environ Health* 13, 56.
- Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559–572.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Perrot, M. (2011). *Scikit-learn: Machine Learning in Python*. Ανάκτηση από JMLR: <https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>
- Python Software Foundation. (2022). *Python Language Reference, version 3.8.8*. Ανάκτηση από <https://www.python.org/>
- Roubeyrie, L., & Celles, S. (2018). Windrose: A Python Matplotlib, Numpy library to manage wind and pollution data, draw windrose. *Journal of Open Source Software*, 3(29), 268.
- Samal, K. K., Panda, A. K., Babu, K. S., & Das, S. K. (2021). An improved pollution forecasting model with meteorological impact using multiple imputation and fine-tuning approach. *Sustainable Cities and Society, Volume 70*.
- Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.
- The pandas development team. (2020, Feb). *pandas-dev/pandas: Pandas 1.4.3*. Ανάκτηση από Zenodo: <https://doi.org/10.5281/zenodo.3509134>

- Vettigli, G. (2018). *Minisom: minimalistic and NumPy-based implementation of the Self Organizing Map*. Ανάκτηση από <https://github.com/JustGlowing/minisom>
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., . . . Mayorov, N. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17(3), 261-272.
- Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.
- Williams, R., Kilaru, V., Snyder, E., Kaufman, A., Dye, T., . . . Hafner, A. H. (2014). *Air Sensor Guidebook*. Washington, DC: U.S. Environmental Protection Agency, EPA/600/R-14/159 (NTIS PB2015-100610).
- Καρατζάς, Κ. (2012). *Σημειώσεις Περιβαλλοντικής Πληροφορικής*. Θεσσαλονίκη: Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.

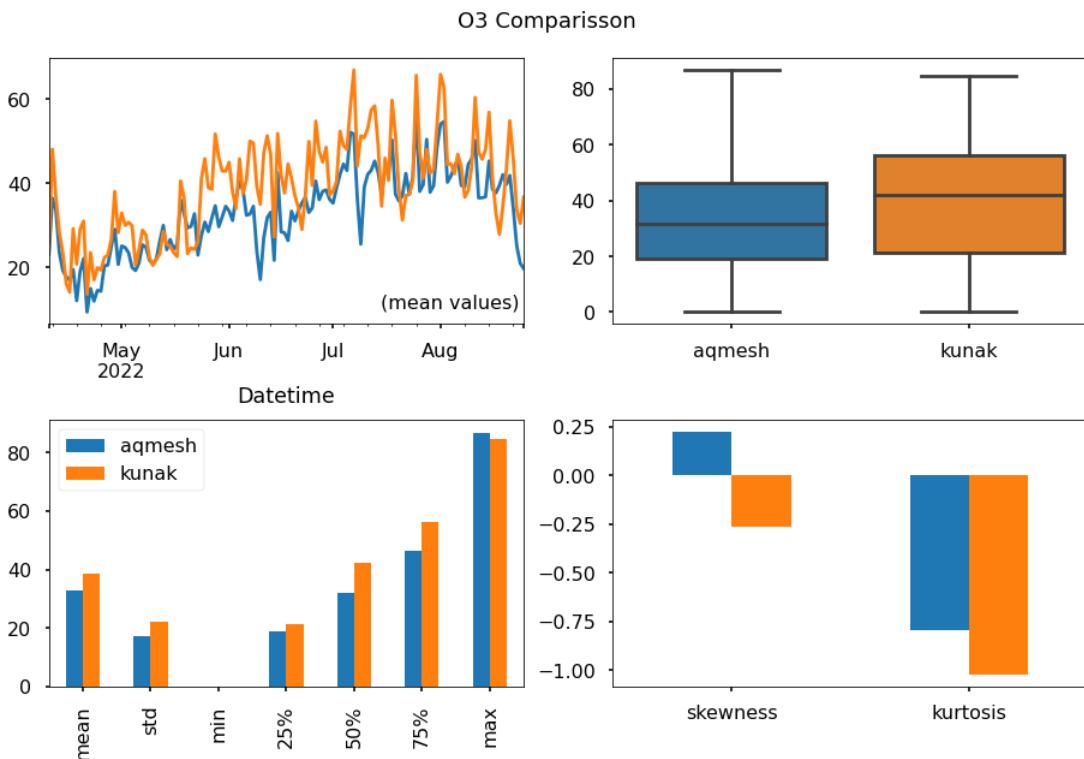
6. Παράρτημα



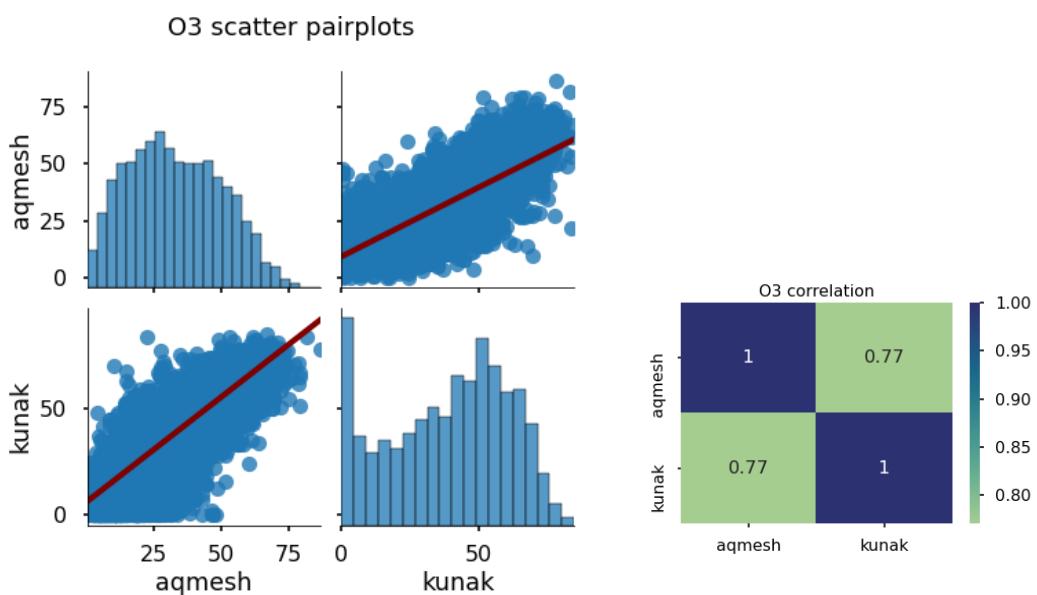
Εικόνα 37. Τοποθεσία ΚΠΠΑ (OpenStreetMap contributors, 2017)

6.1. Διαγράμματα

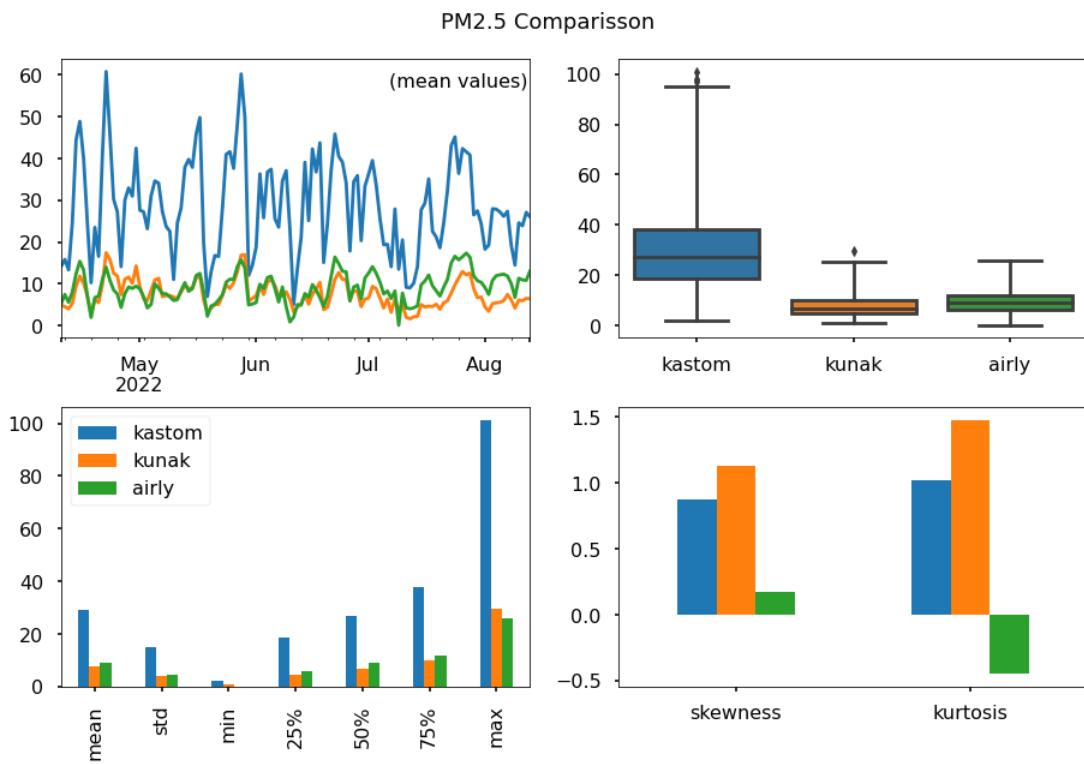
6.1.1. Σύγκριση κόμβων



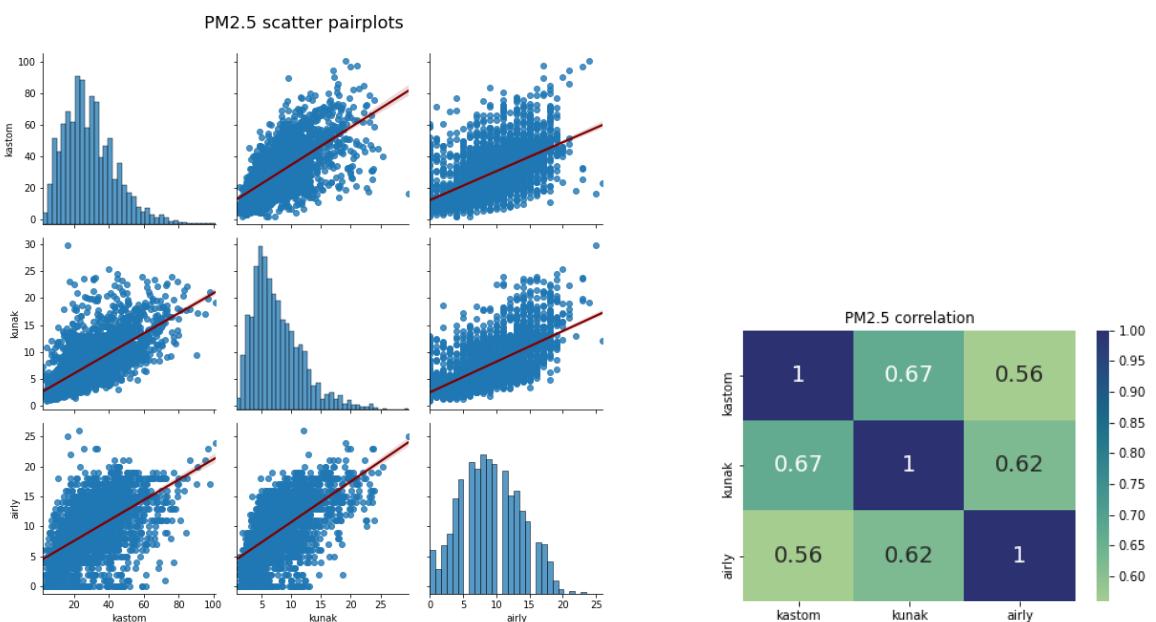
Εικόνα 38. Σύγκριση στατιστικών μεγεθών ημερήσιων μέσων NO₂ (ΚΠΠΑ Aqmesh, Kunak)



Εικόνα 39. Συσχέτιση μετρήσεων O₃ μεταξύ κόμβων (ΚΠΠΑ Aqmesh, Kunak). Διαγράμματα συσχέτισης και ιστογράμματα (αριστερά), πίνακας συντελεστών συσχέτισης κατά Pearson (δεξιά).



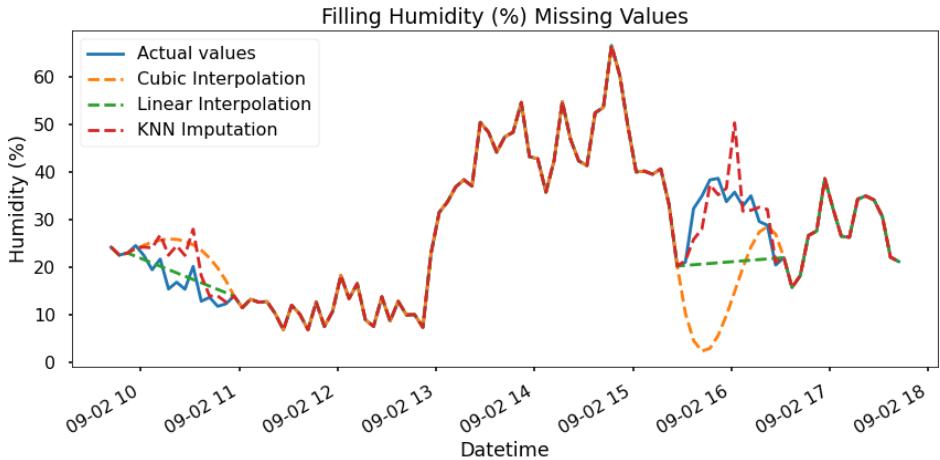
Εικόνα 40. Σύγκριση στατιστικών μεγεθών ημερήσιων μέσων PM2.5 (KASTOM, Kunak, Airly)



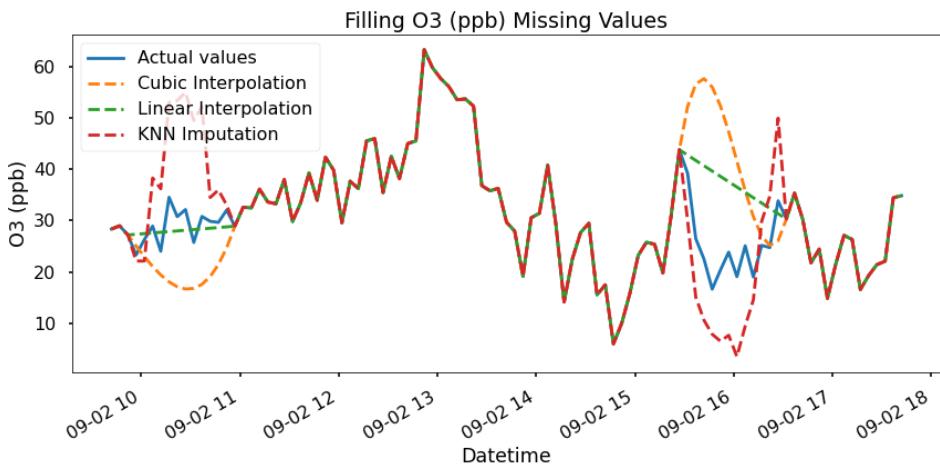
Εικόνα 41. Συσχέτιση μετρήσεων PM2.5 μεταξύ κόμβων (ΚΠΠΑ KASTOM, Kunak, Airly). Διαγράμματα συσχέτισης και ιστογράμματα (αριστερά), πίνακας συντελεστών συσχέτισης κατά Pearson (δεξιά).

6.1.2. Συμπλήρωση ελλειπουσών τιμών

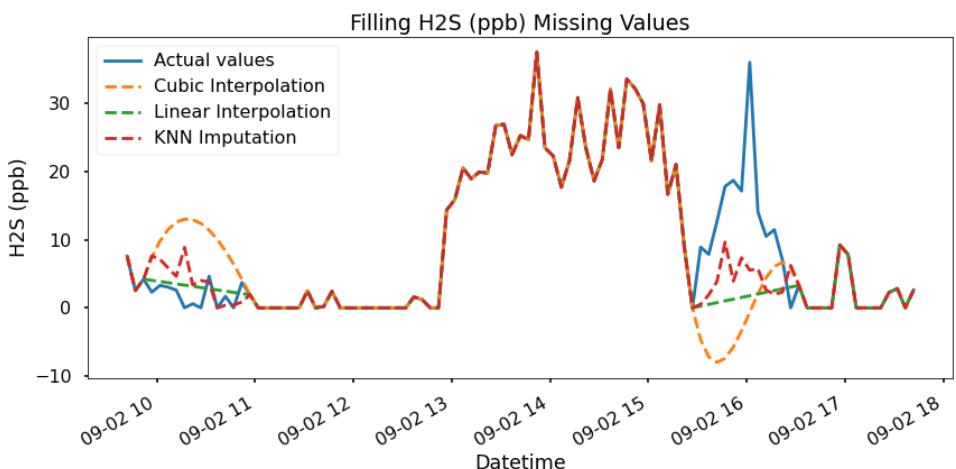
Kunak



Εικόνα 42. Δοκιμή συμπλήρωσης ελλειπουσών τιμών υγρασίας (Kunak)

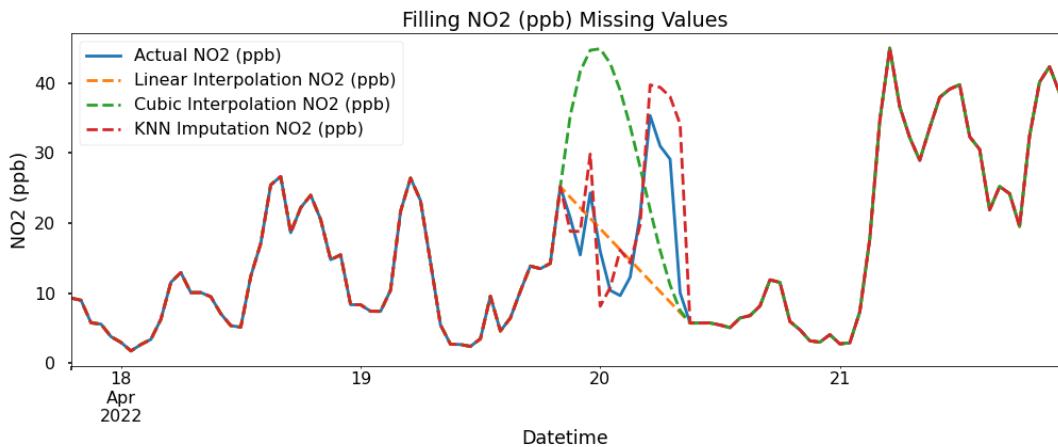


Εικόνα 43. Δοκιμή συμπλήρωσης ελλειπουσών τιμών O3 (Kunak)

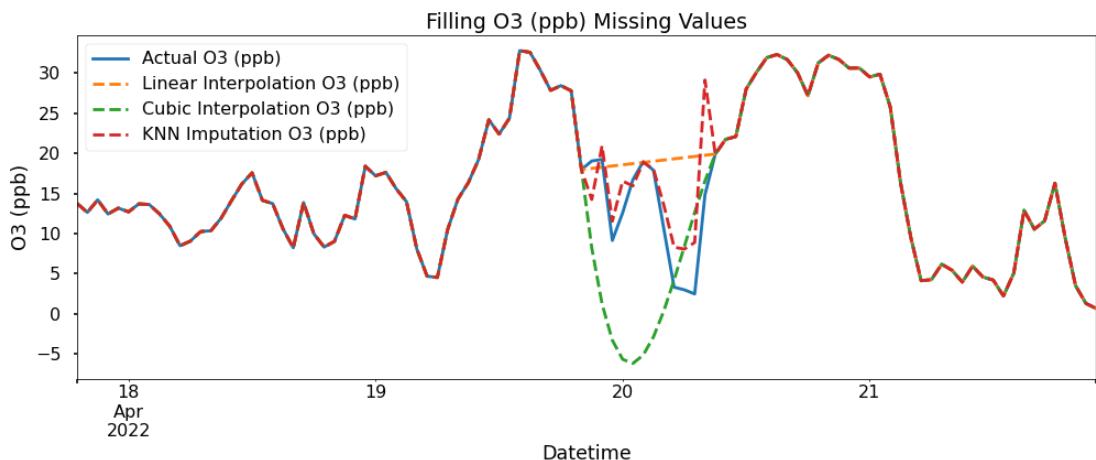


Εικόνα 44. Δοκιμή συμπλήρωσης ελλειπουσών τιμών H2S (ΚΠΠΑ Kunak)

Aqmesh



Εικόνα 45. Δοκιμή συμπλήρωσης ελλειπουσών τιμών NO2 (Aqmesh)



Εικόνα 46. Δοκιμή συμπλήρωσης ελλειπουσών τιμών O3 (ΚΠΠΑ Aqmesh)

Airly

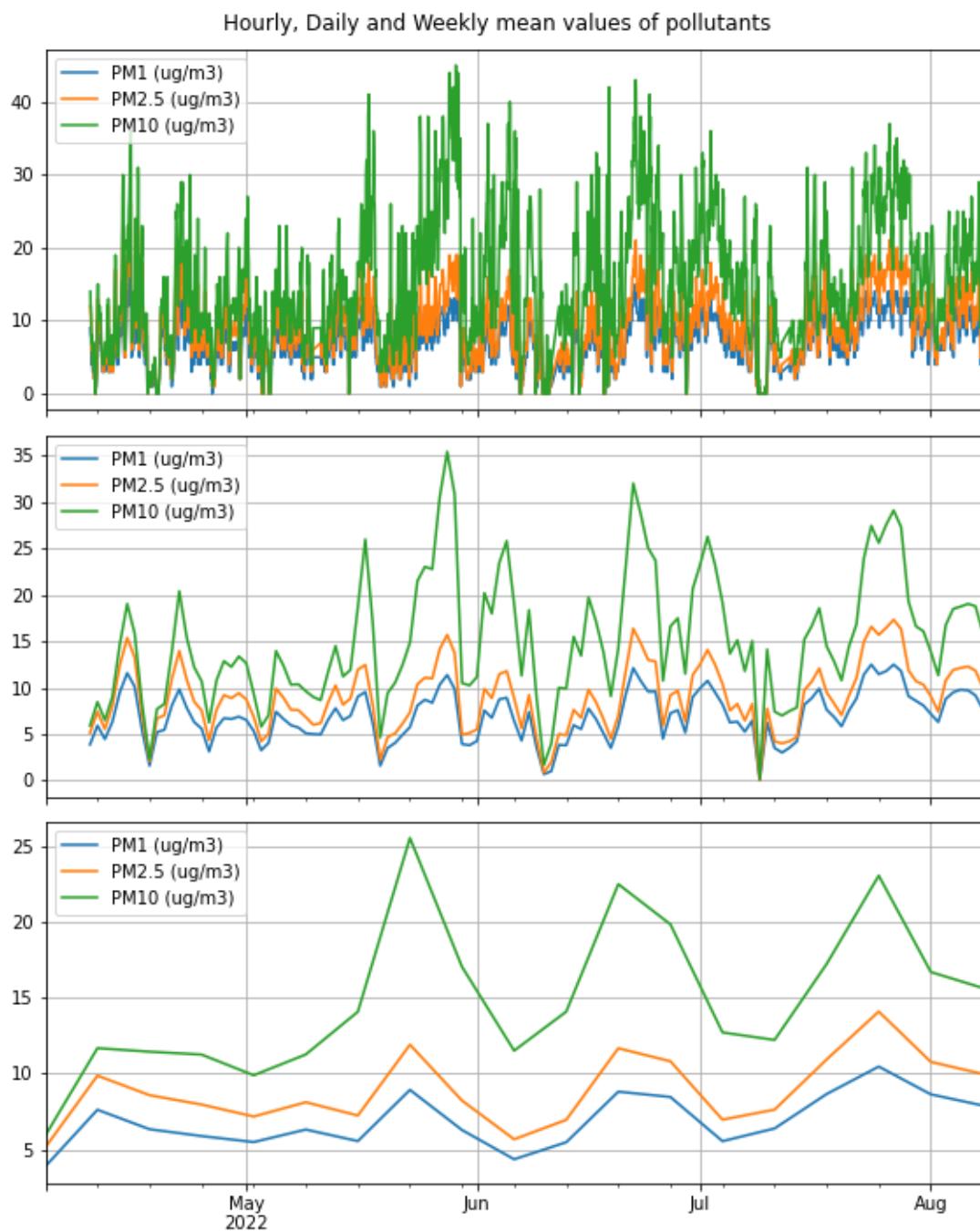
Οι ελλείψεις κρίθηκαν αμελητέες.

KASTOM

Οι ελλείψεις κρίθηκαν αμελητέες.

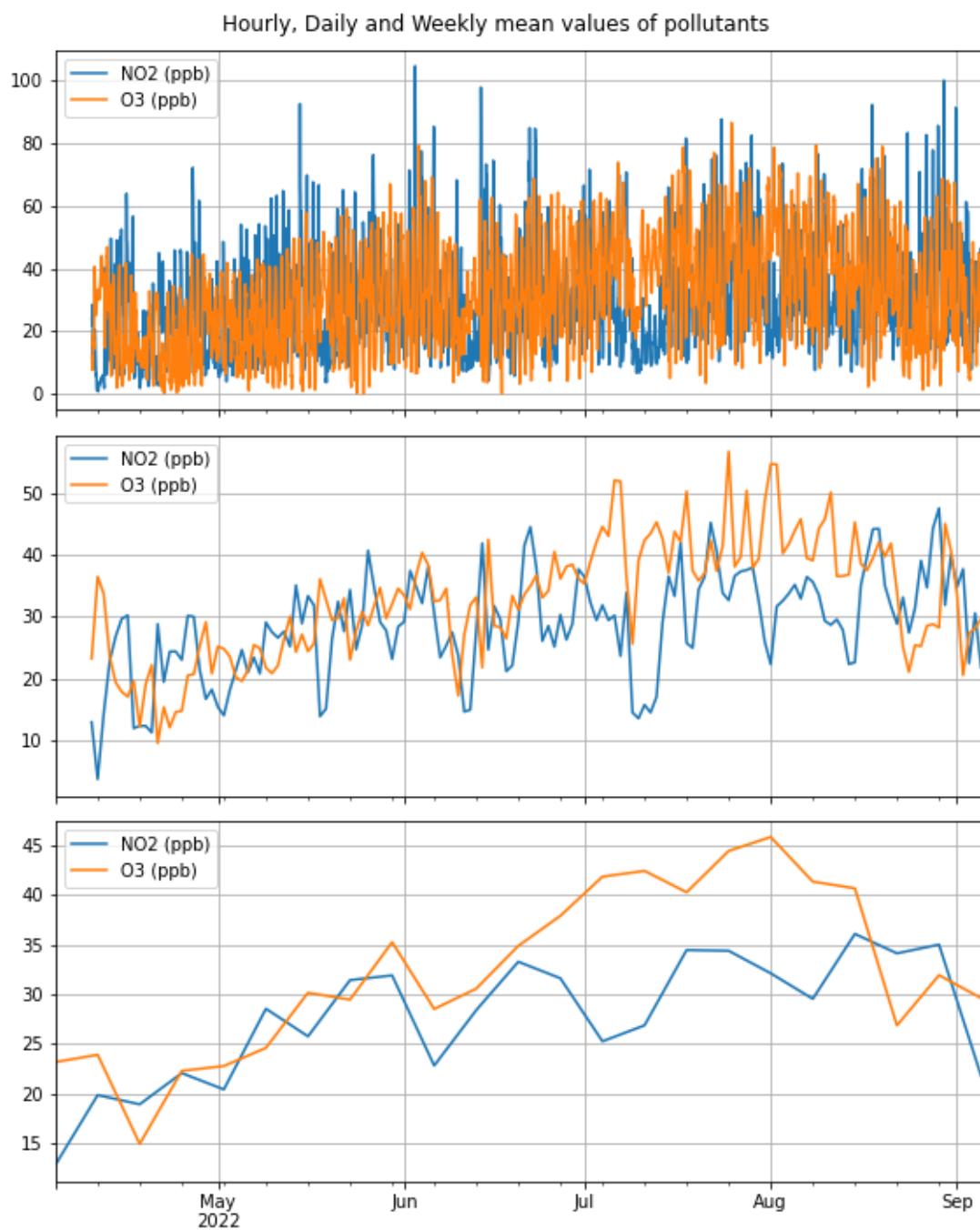
6.1.3. Αναδειγματοληψία

Airly



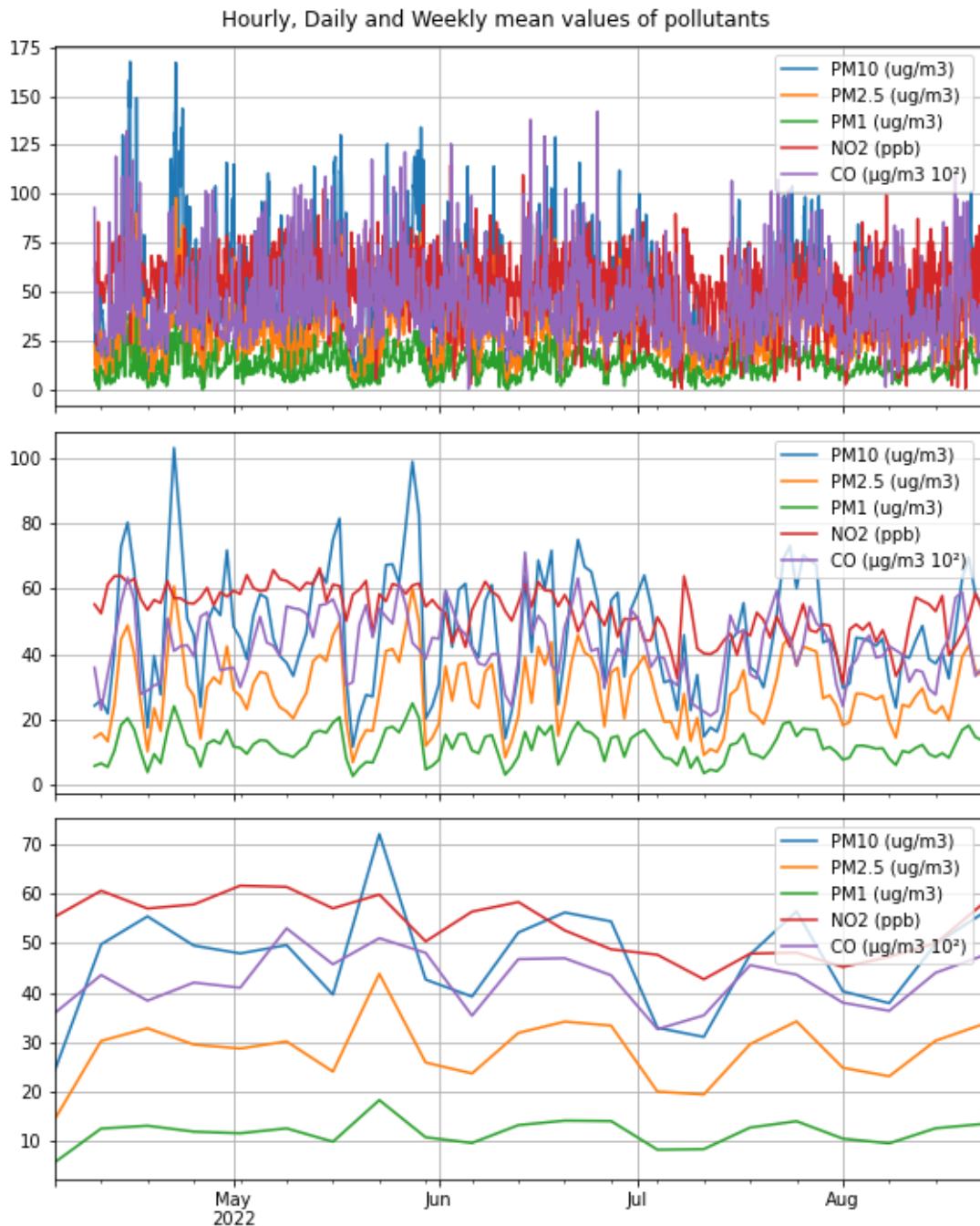
Εικόνα 47. Ωριαία (πάνω), ημερήσια (μέση) και εβδομαδιαία (κάτω) αναδειγματοληψία ρύπων (ΚΠΠΑ Airly)

Aqmesh



Εικόνα 48. Ωριαία (πάνω), ημερήσια (μέση) και εβδομαδιαία (κάτω) αναδειγματοληψία ρύπων (ΚΠΠΑ Aqmesh)

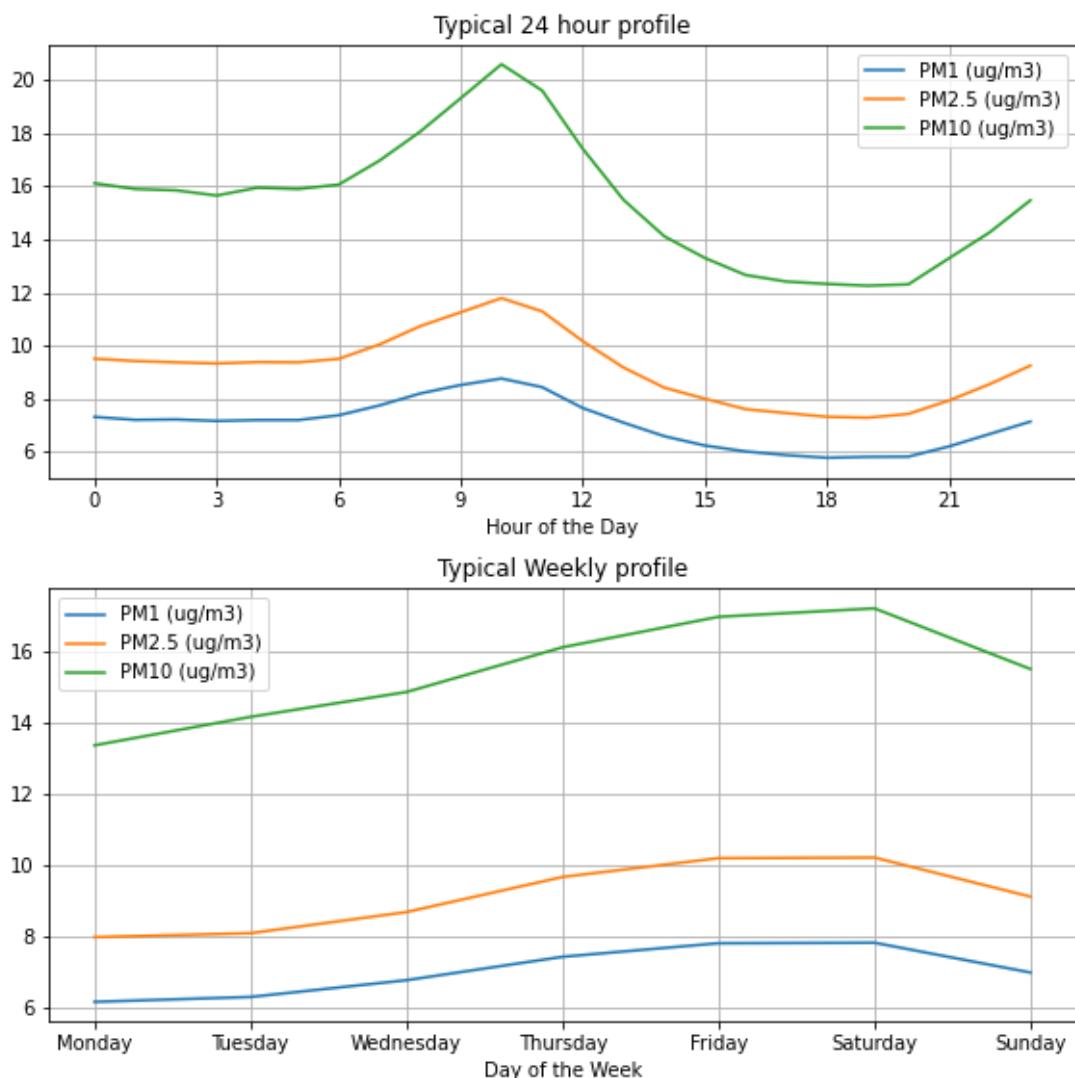
KASTOM



Εικόνα 49. Ωριαία (πάνω), ημερήσια (μέση) και εβδομαδιαία (κάτω) αναδειγματοληψία ρύπων (ΚΠΠΑ KASTOM)

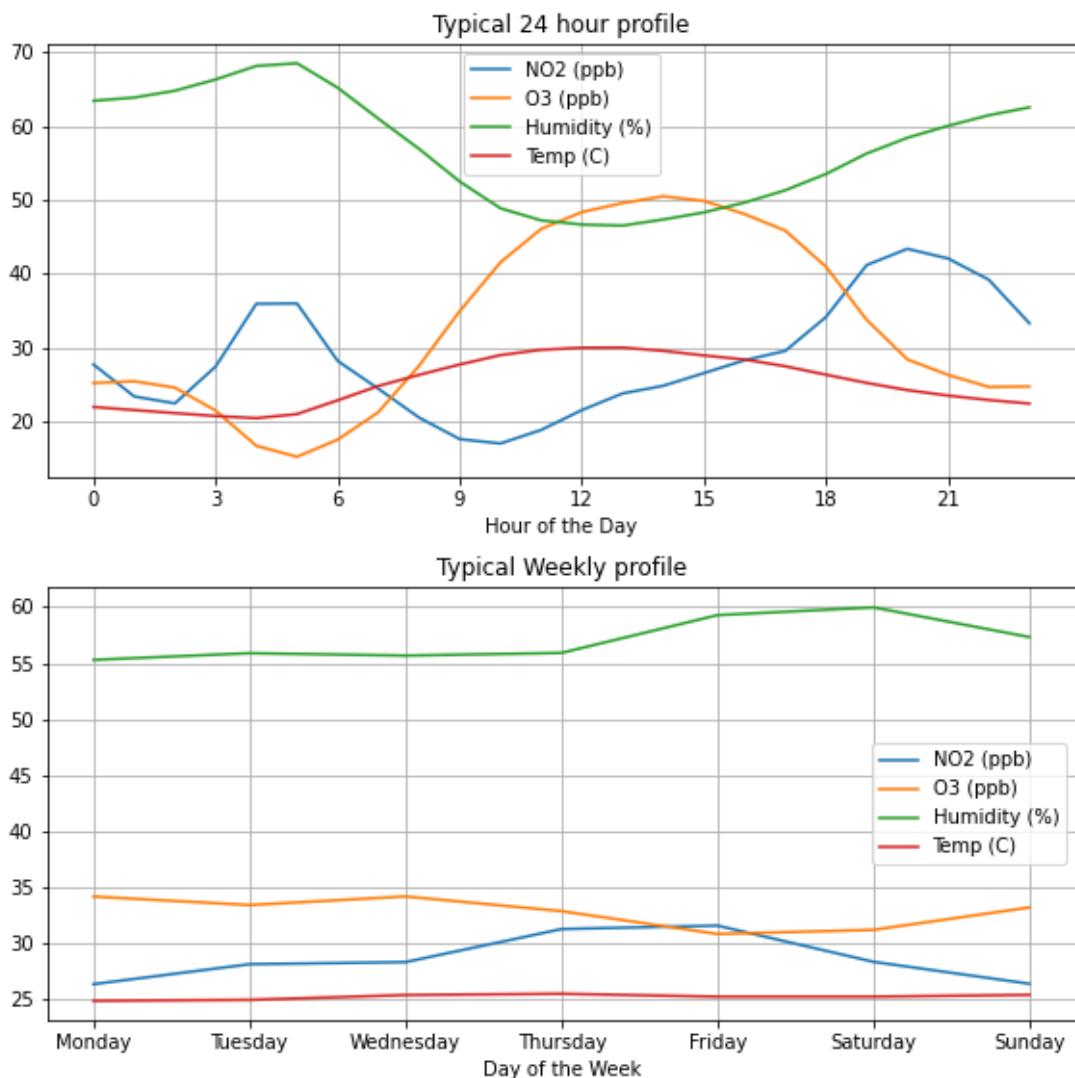
6.1.4. Τυπικά προφίλ

Airly

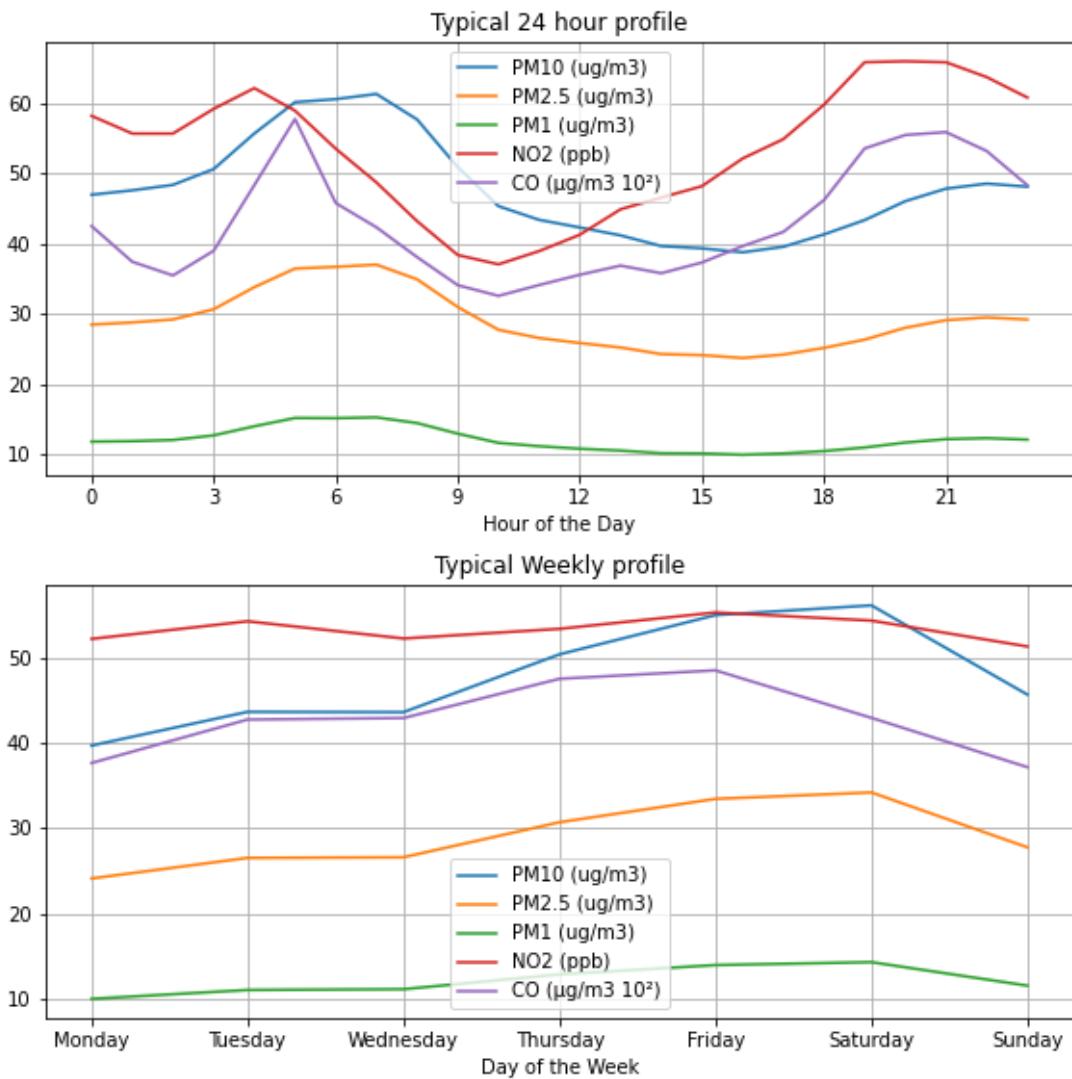


Εικόνα 50. Τυπικά εικοσιτετράωρα (πάνω) και εβδομαδιαία (κάτω) προφίλ (ΚΠΠΑ Airly)

Aqmesh



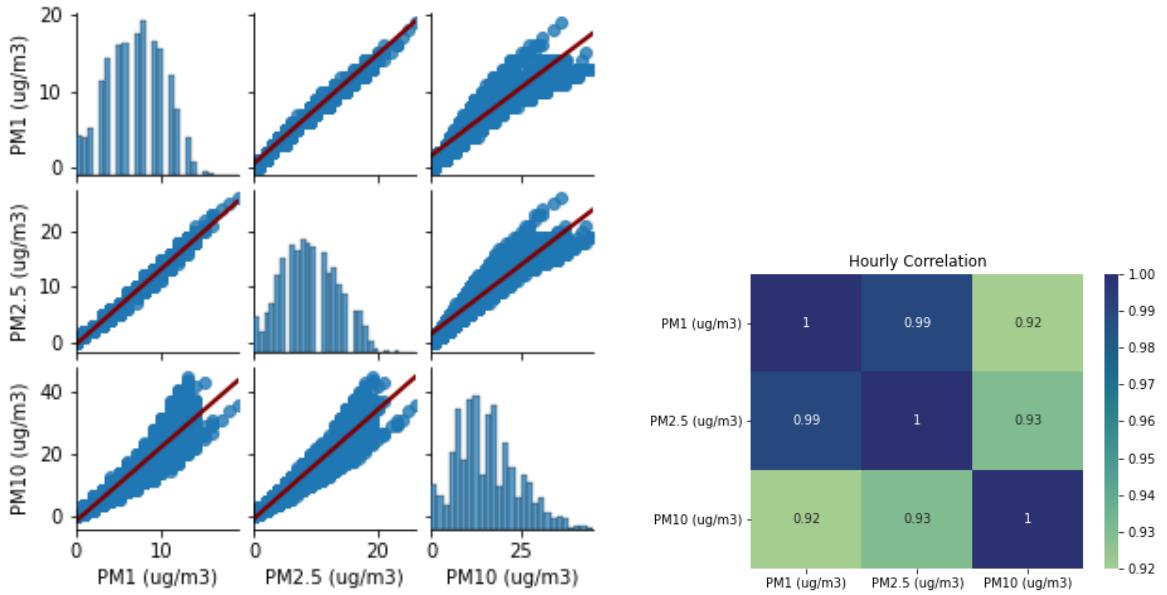
Εικόνα 51. Τυπικά εικοσιτετράωρα (πάνω) και εβδομαδιαία (κάτω) προφίλ (ΚΠΠΑ Aqmesh)



Εικόνα 52. Τυπικά εικοσιτετράωρα (πάνω) και εβδομαδιαία (κάτω) προφίλ (ΚΠΠΑ KASTOM)

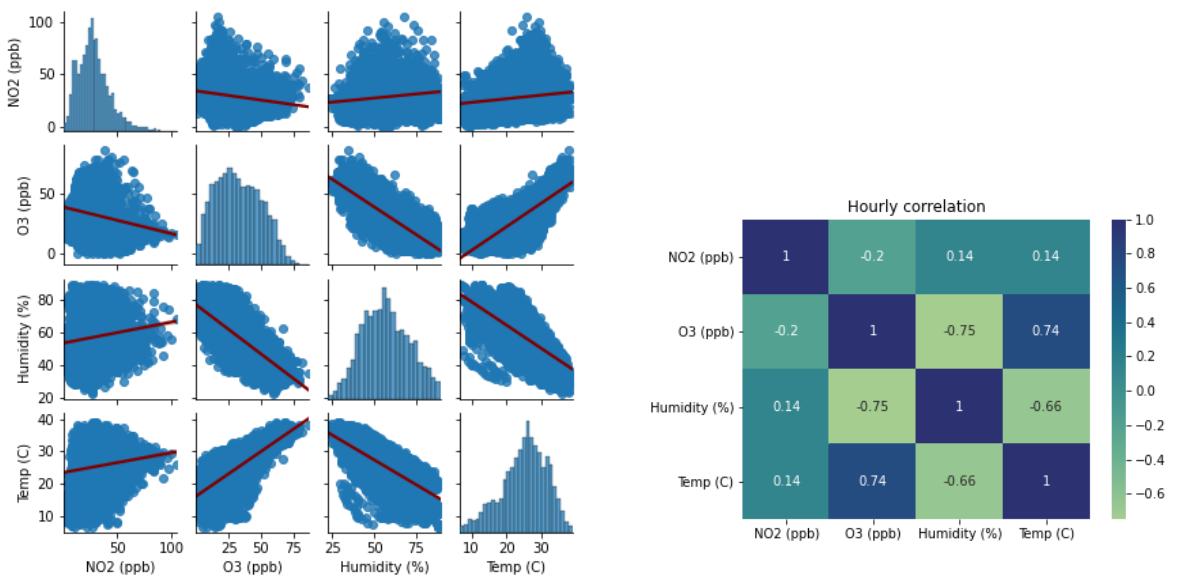
6.1.5. Διερεύνηση συσχέτισης

Airly



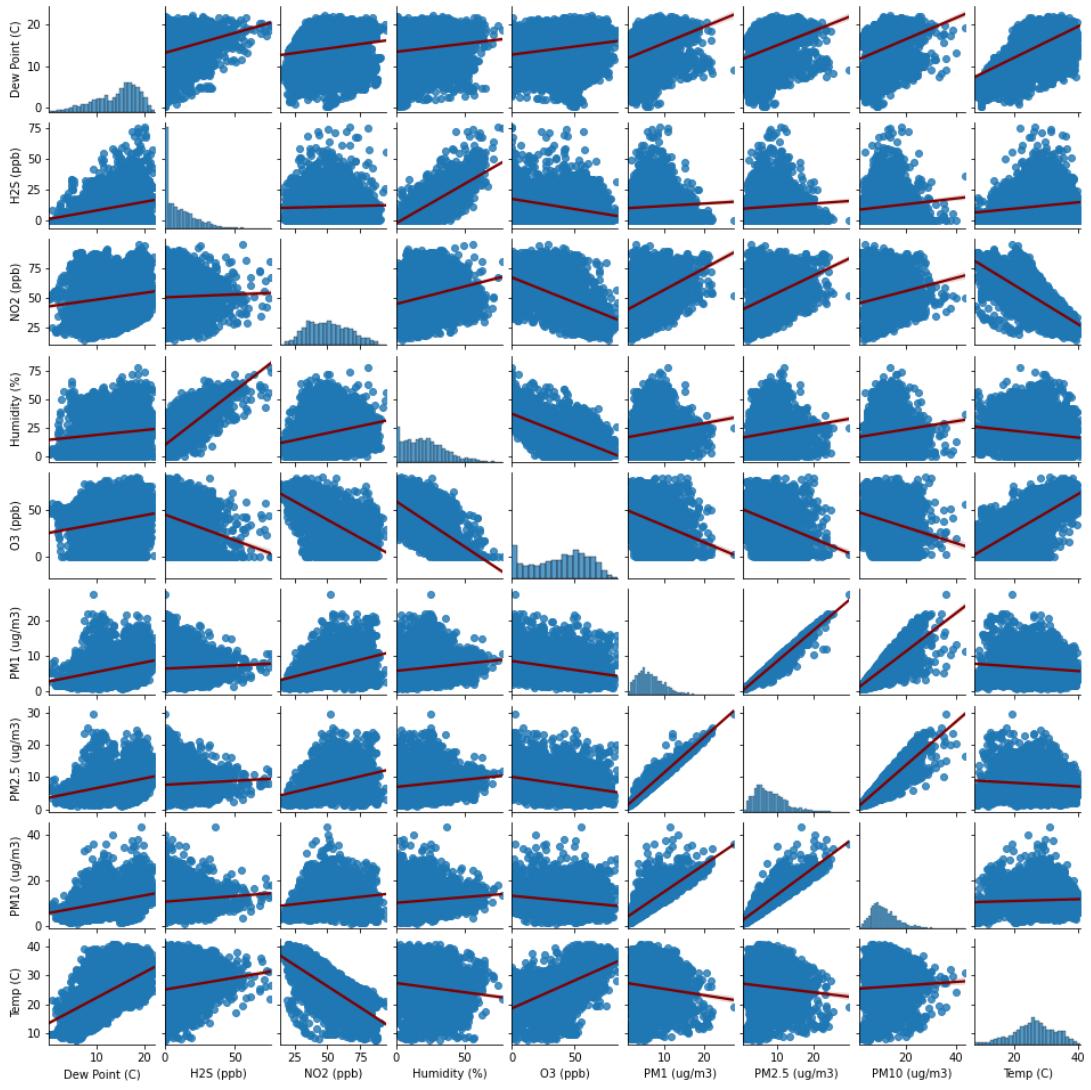
Εικόνα 53. Συσχέτιση ωριαίων τιμών (ΚΠΠΑ Airly)

Aqmesh



Εικόνα 54. Συσχέτιση ωριαίων τιμών (ΚΠΠΑ Aqmesh)

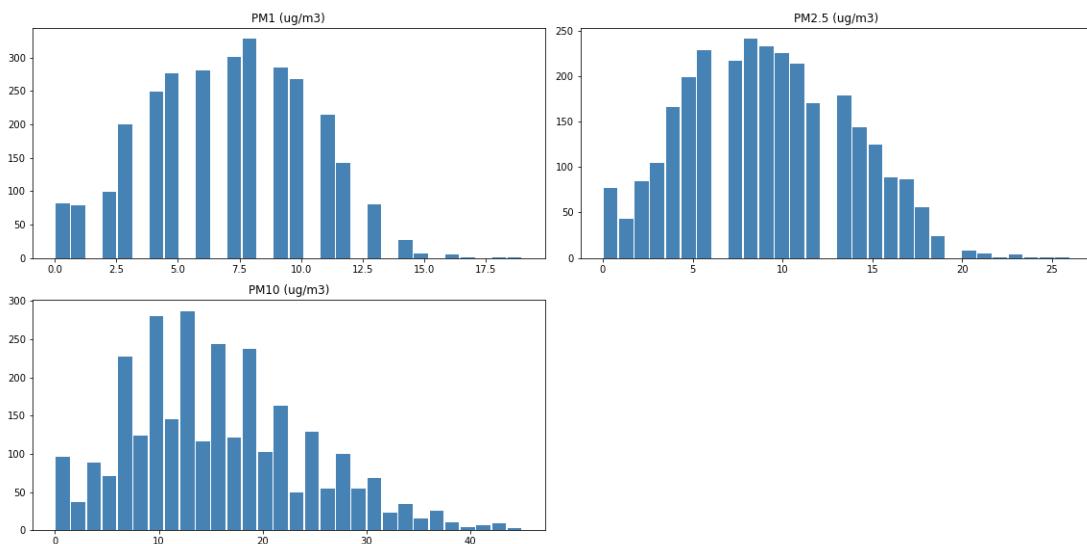
Kunak



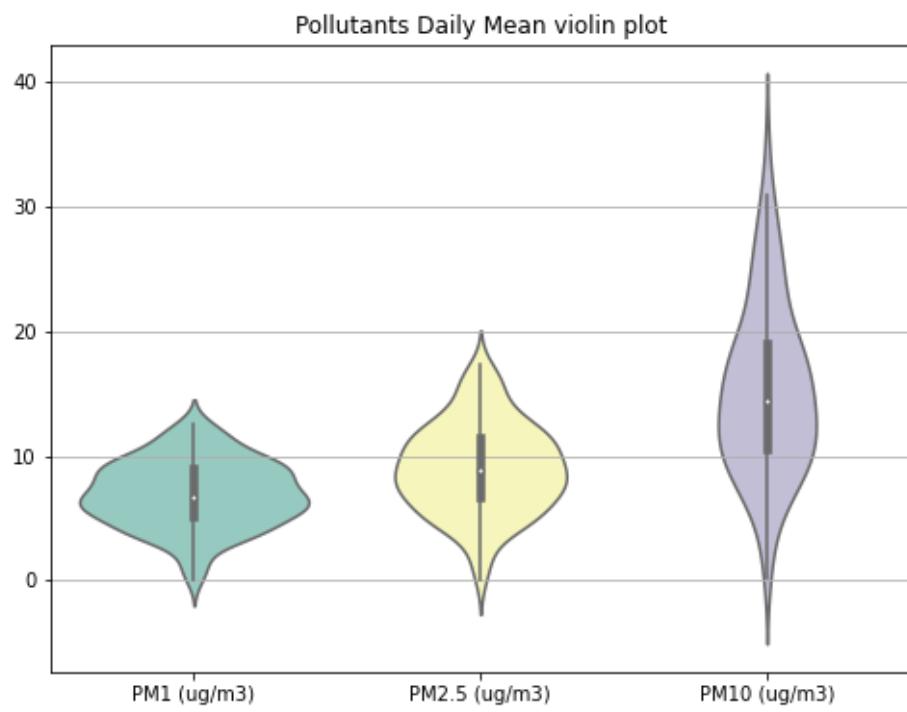
Εικόνα 55. Συσχέτιση ωριαίων τιμών (ΚΠΠΑ Kunak)

6.1.6. Περιγραφική Στατιστική

Airly

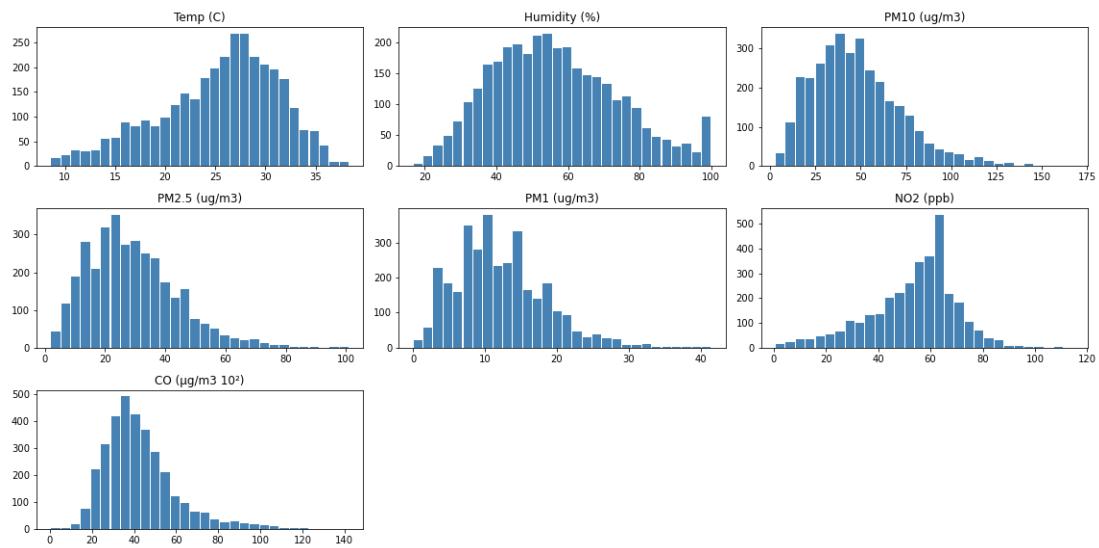


Εικόνα 56. Ιστογράμματα ωριαίου ρυθμού δειγματοληψίας μετρούμενων μεγεθών (ΚΠΠΑ Airly)

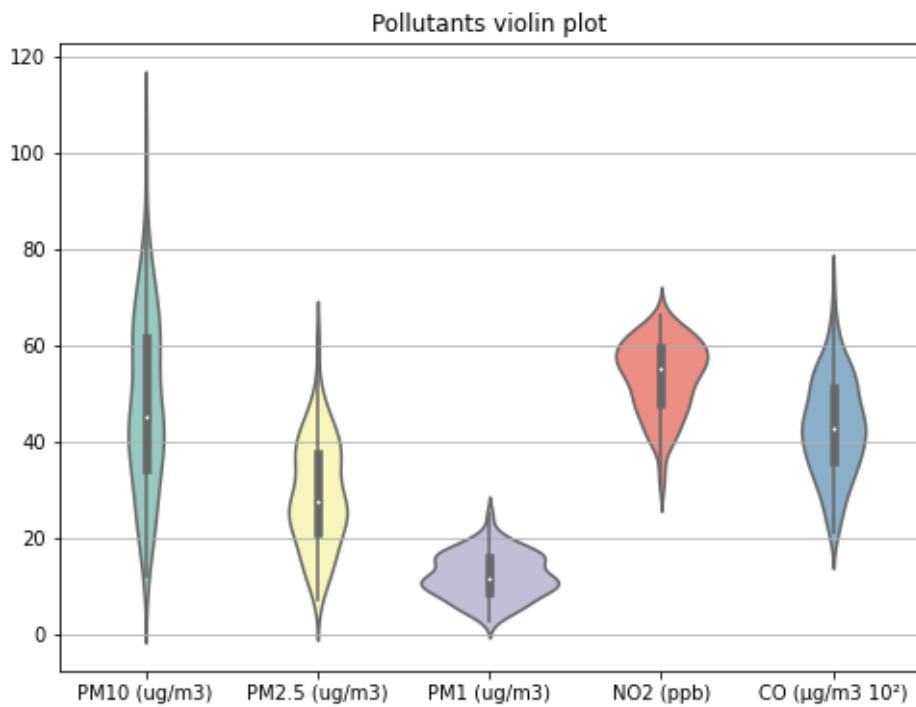


Εικόνα 57. Violin plot μέσου ημερήσιου ρυθμού δειγματοληψίας ρύπων (ΚΠΠΑ Airly)

KASTOM

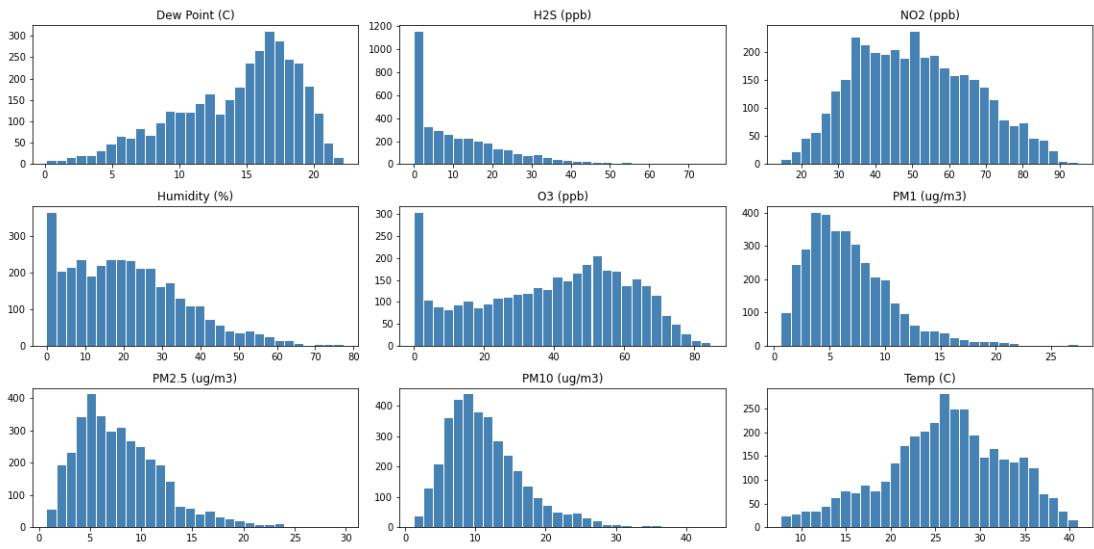


Εικόνα 58. Ιστογράμματα ωριαίου ρυθμού δειγματοληψίας μετρούμενων μεγεθών (ΚΠΠΑ KASTOM)

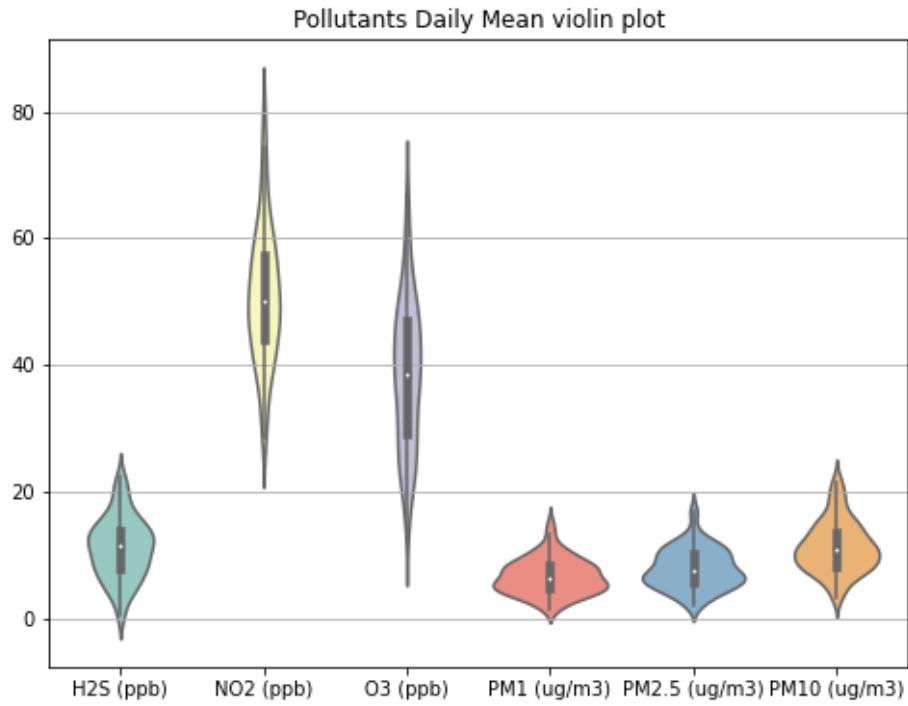


Εικόνα 59. Violin plot μέσου ημερήσιου ρυθμού δειγματοληψίας ρύπων (ΚΠΠΑ KASTOM)

Kunak



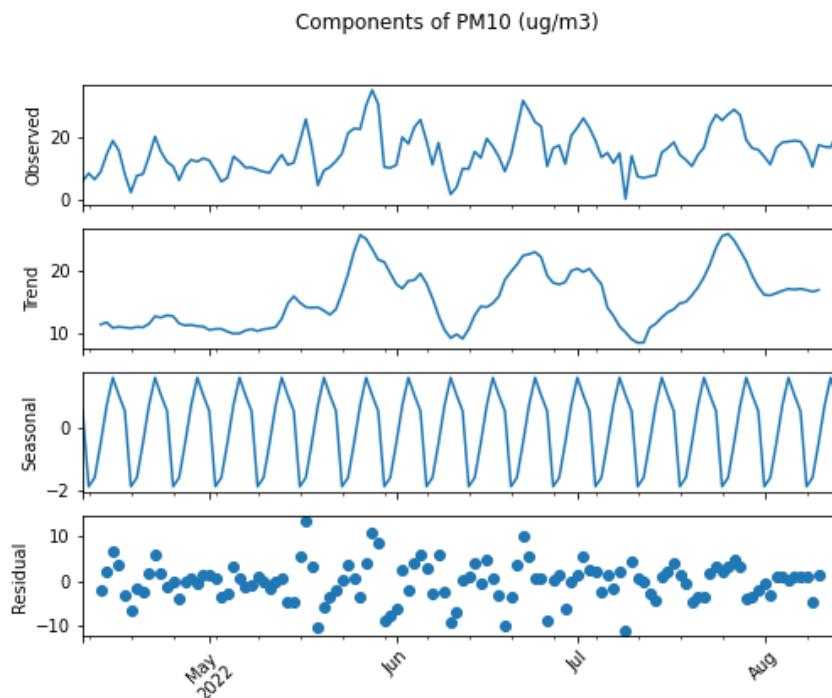
Εικόνα 60. Ιστογράμματα ωριαίου ρυθμού δειγματοληψίας μετρούμενων μεγεθών (ΚΠΠΑ Kunak)



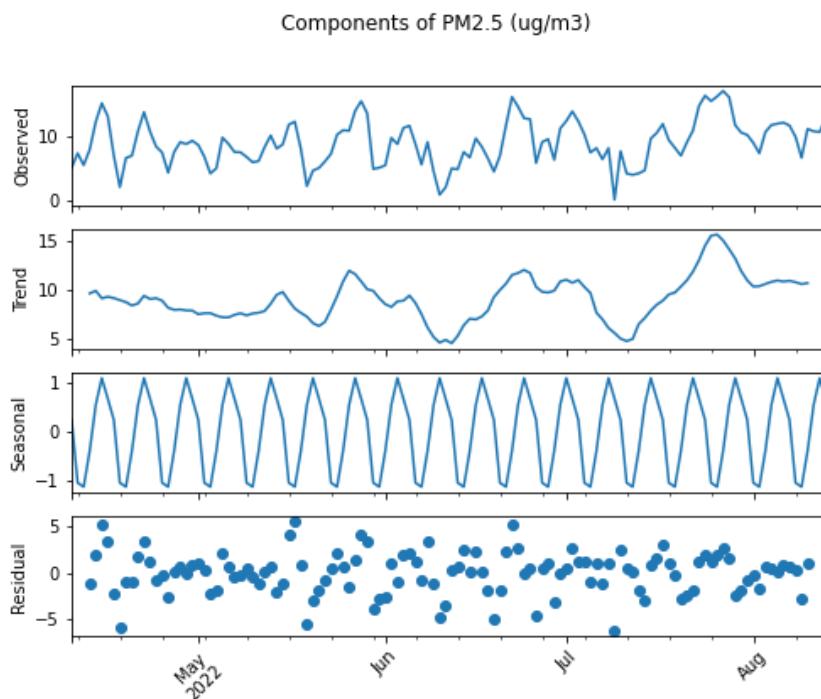
Εικόνα 61. Violin plot μέσου ημερήσιου ρυθμού δειγματοληψίας ρύπων (ΚΠΠΑ Kunak)

6.1.7. Αποσύνθεση χρονοσειράς

Airly

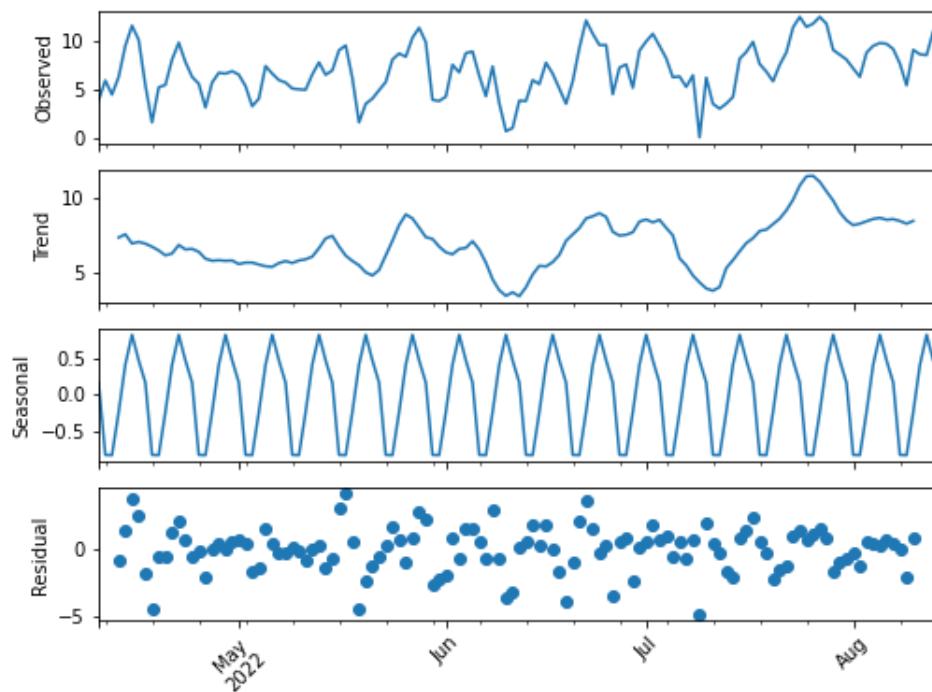


Εικόνα 62. Συνθετικά προσθετικού μοντέλου χρονοσειράς PM10 (ΚΠΠΑ Airly)

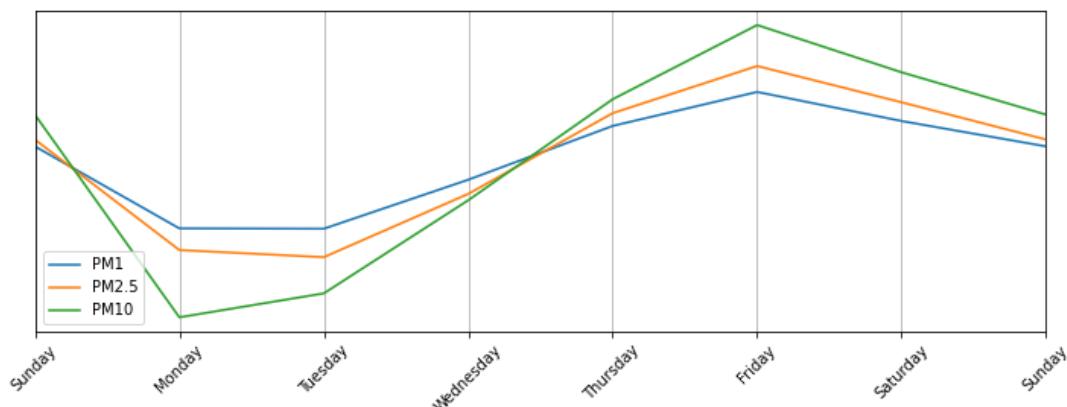


Εικόνα 63. Συνθετικά προσθετικού μοντέλου χρονοσειράς PM2.5 (ΚΠΠΑ Airly)

Components of PM1 ($\mu\text{g}/\text{m}^3$)



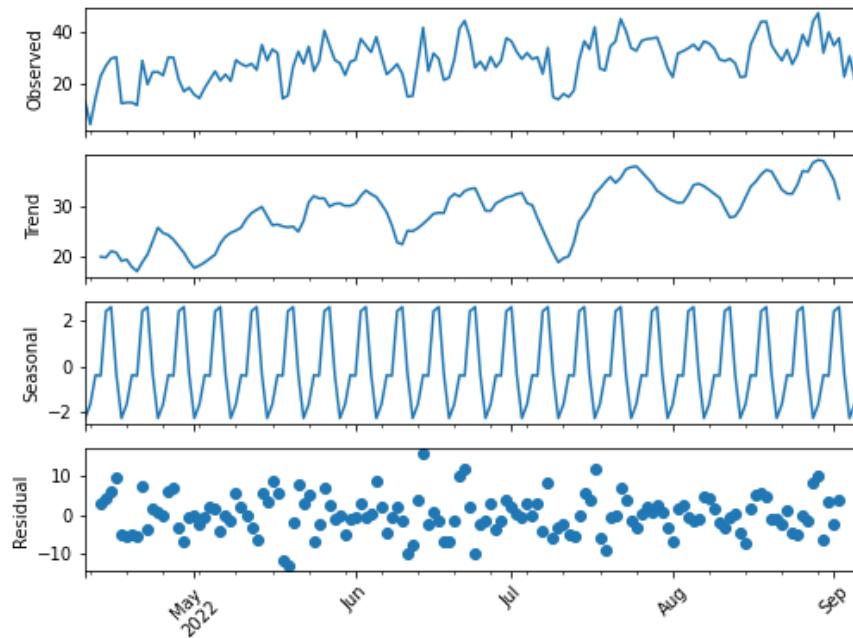
Εικόνα 64. Συνθετικά προσθετικού μοντέλου χρονοσειράς PM1 (ΚΠΠΑ Airly)



Εικόνα 65. Ημερήσιες τάσεις εβδομαδιαίας περιοδικότητας ρύπων (ΚΠΠΑ Airly)

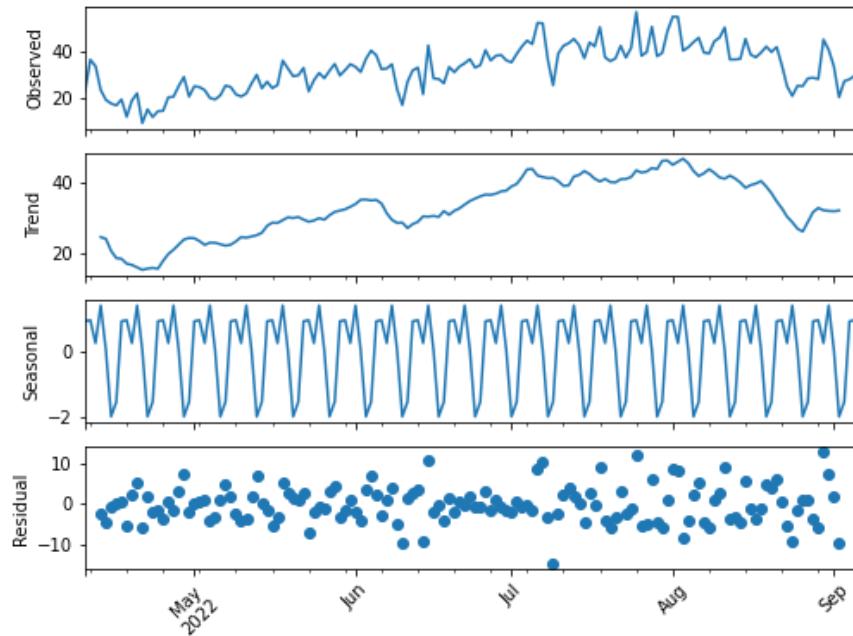
Aqmesh

Components of NO₂ (ppb)

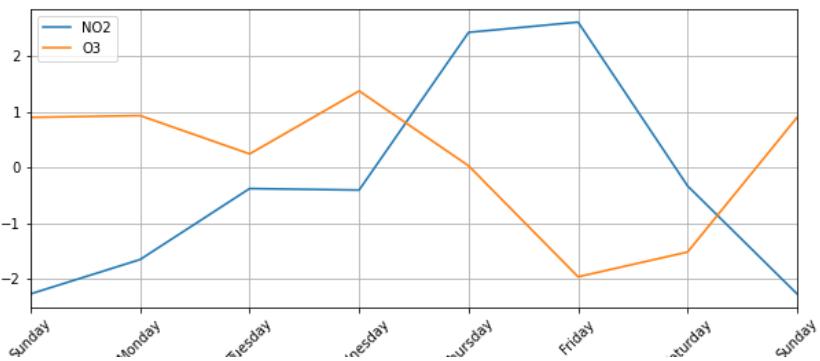


Εικόνα 66. Συνθετικά προσθετικού μοντέλου χρονοσειράς NO₂ (ΚΠΠΑ Aqmesh)

Components of O₃ (ppb)

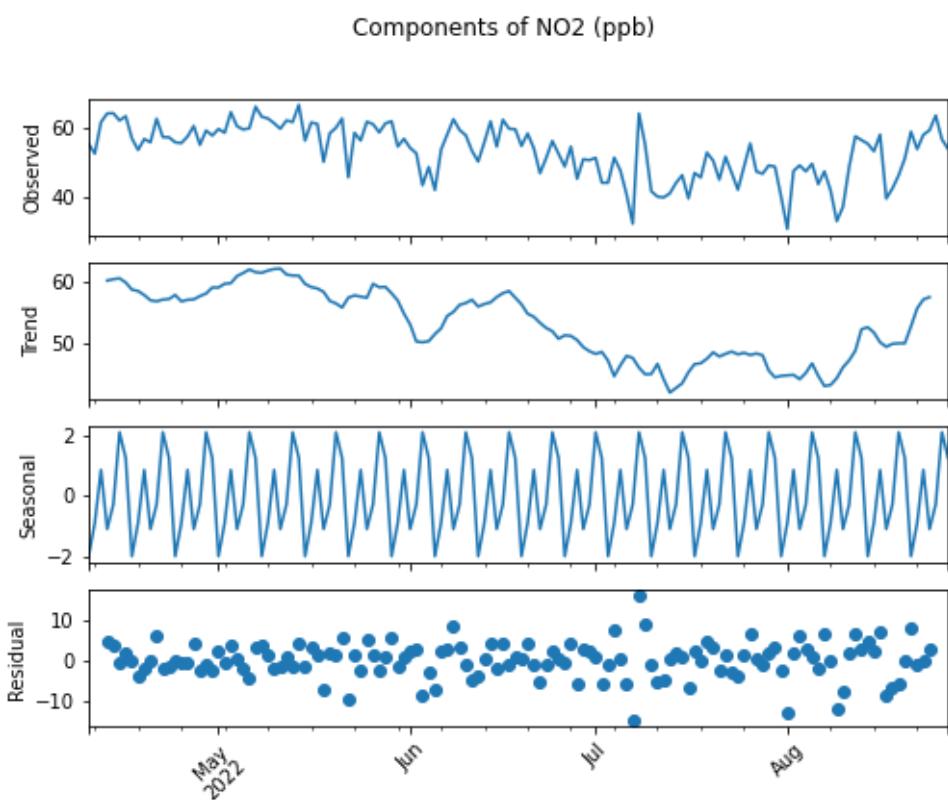


Εικόνα 67. Συνθετικά προσθετικού μοντέλου χρονοσειράς O₃ (ΚΠΠΑ Aqmesh)

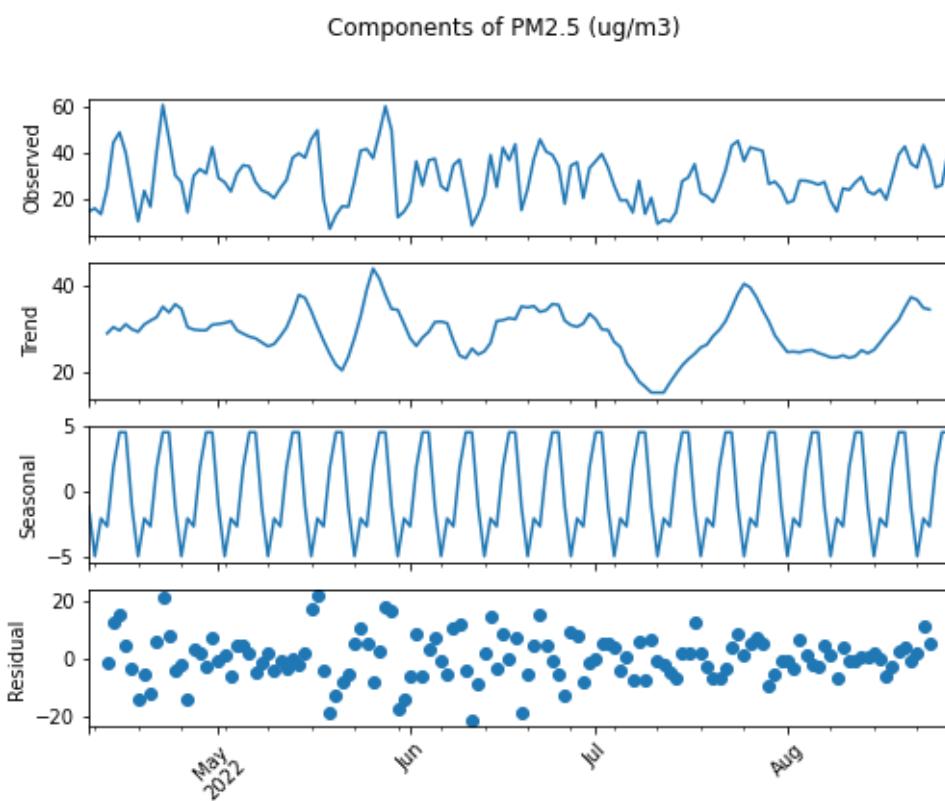


Εικόνα 68. Ημερήσιες τάσεις εβδομαδιαίας περιοδικότητας ρύπων (ΚΠΠΑ Aqmesh)

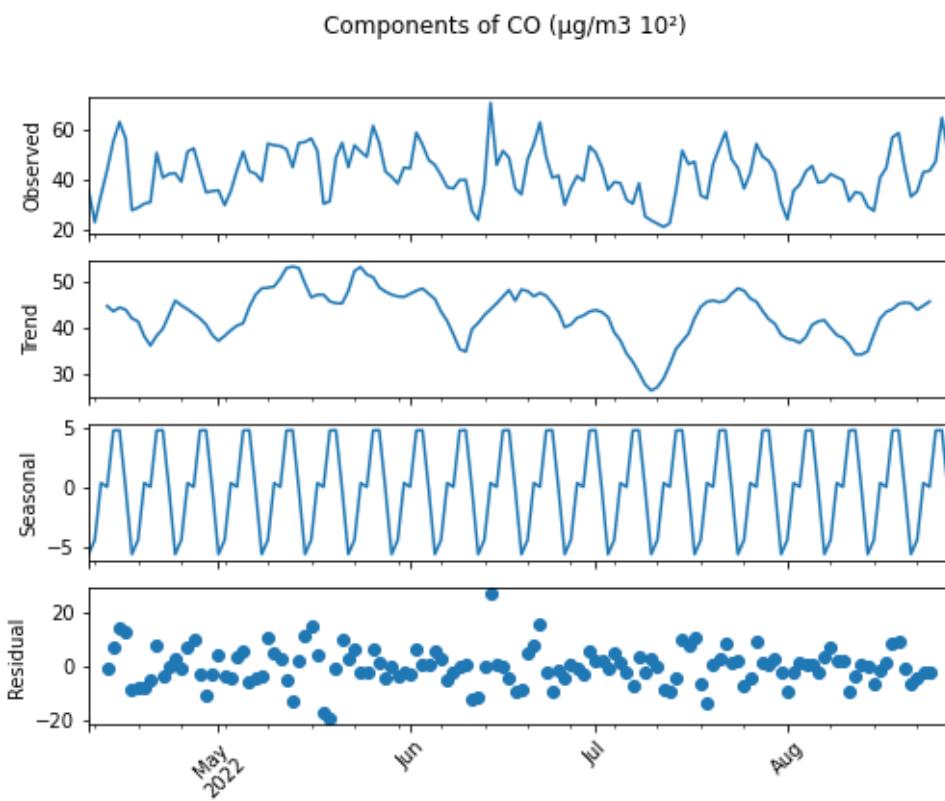
KASTOM



Εικόνα 69. Συνθετικά προσθετικού μοντέλου χρονοσειράς NO₂ (ΚΠΠΑ KASTOM)

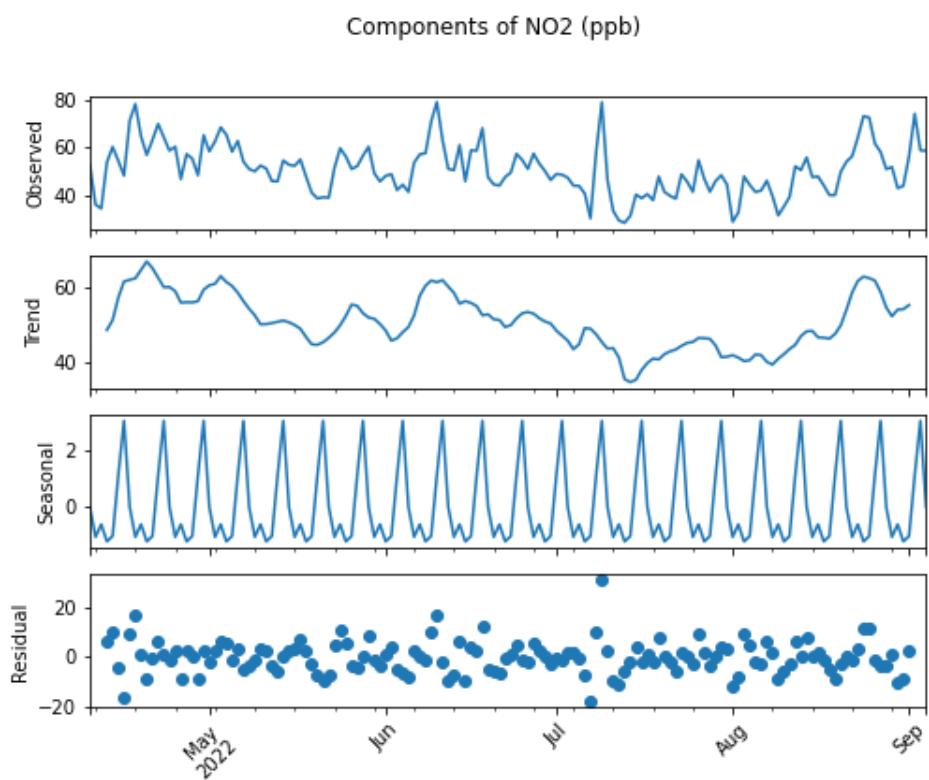


Εικόνα 70. Συνθετικά προσθετικού μοντέλου χρονοσειράς PM2.5 (ΚΠΠΑ KASTOM)

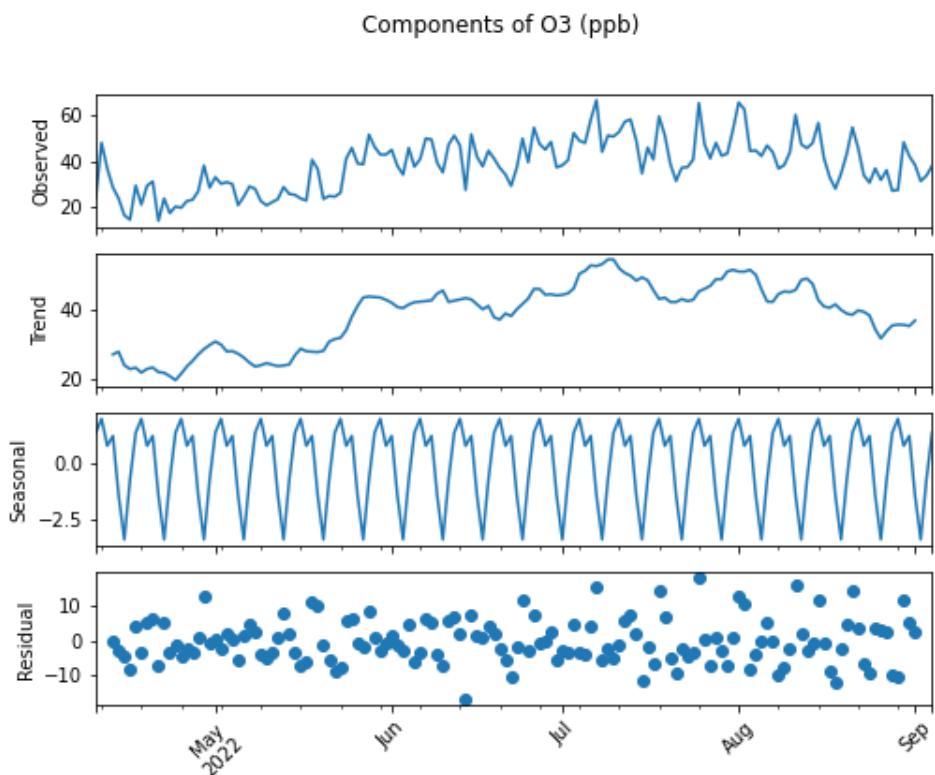


Εικόνα 71. Συνθετικά προσθετικού μοντέλου χρονοσειράς CO (ΚΠΠΑ KASTOM)

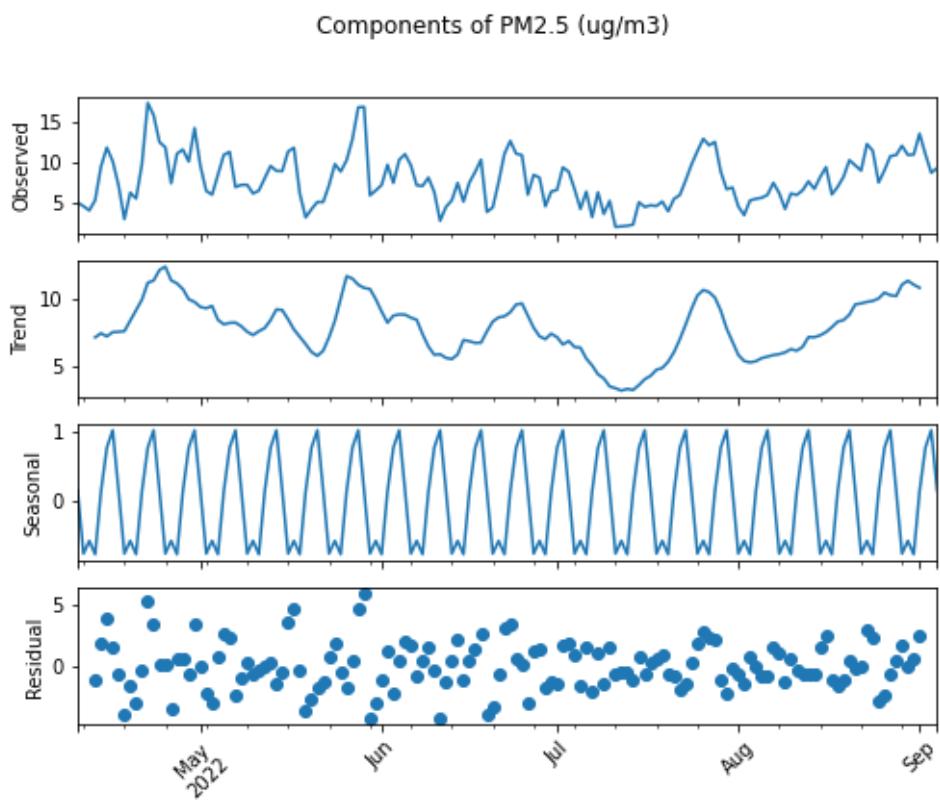
Kunak



Εικόνα 72. Συνθετικά προσθετικού μοντέλου χρονοσειράς NO₂ (ΚΠΠΑ Kunak)



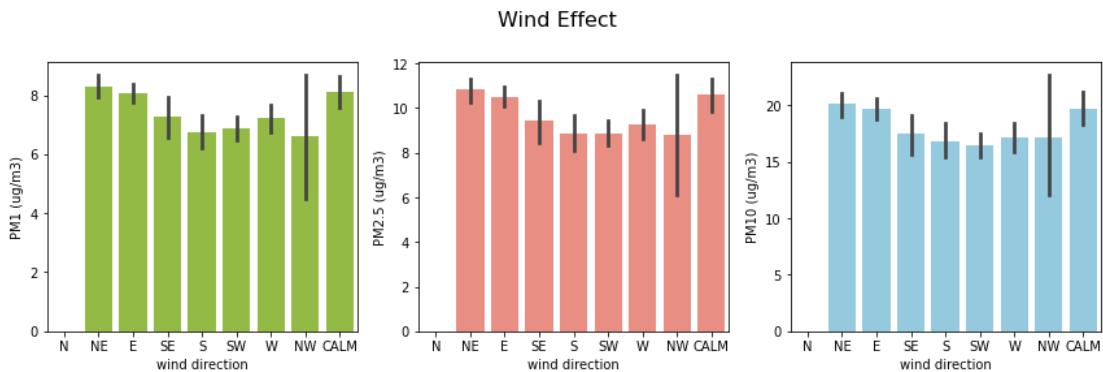
Εικόνα 73. Συνθετικά προσθετικού μοντέλου χρονοσειράς O₃ (ΚΠΠΑ Kunak)



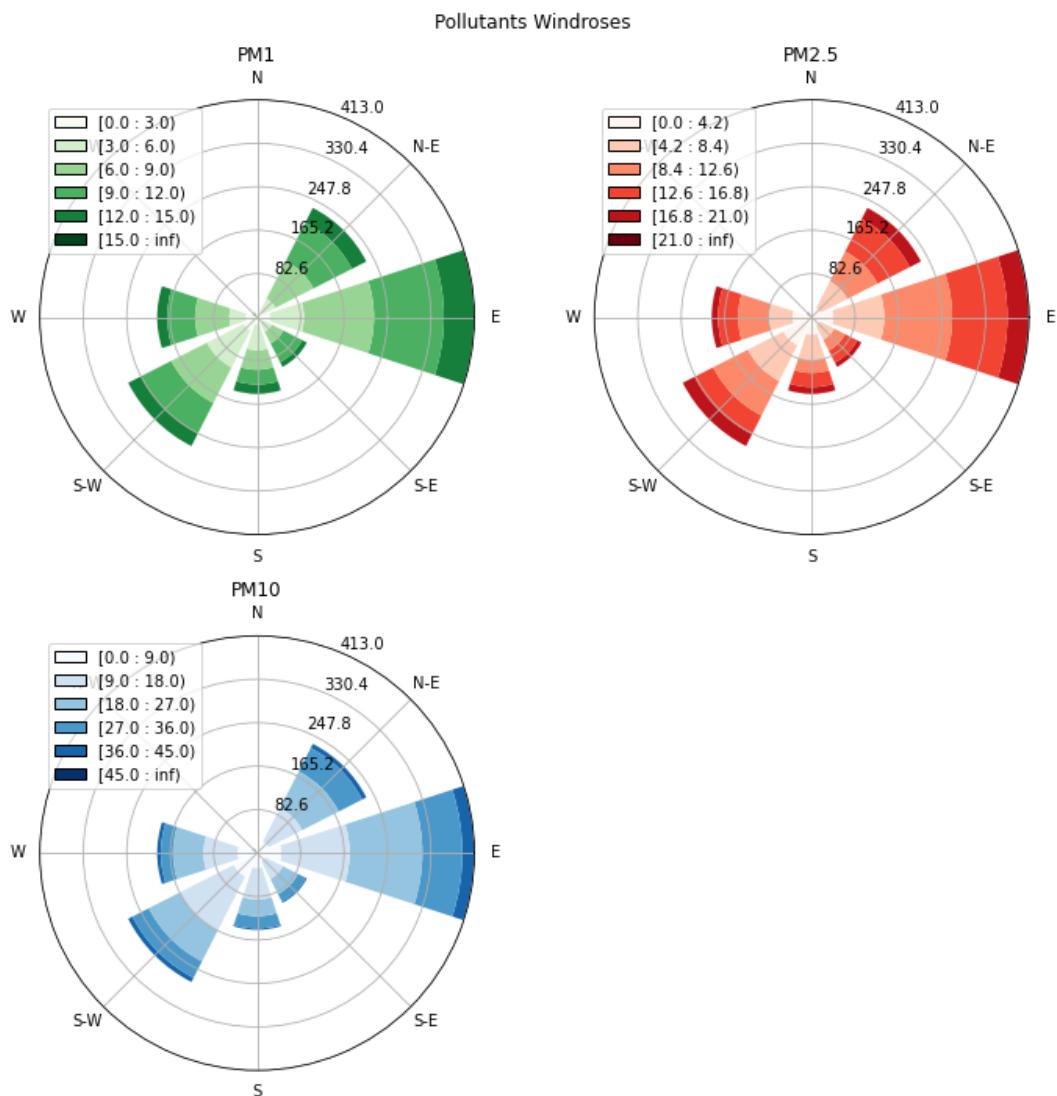
Εικόνα 74. Συνθετικά προσθετικού μοντέλου χρονοσειράς PM2.5 (ΚΠΠΑ Kunak)

6.1.8. Συνεισφορά ανέμου

Airly

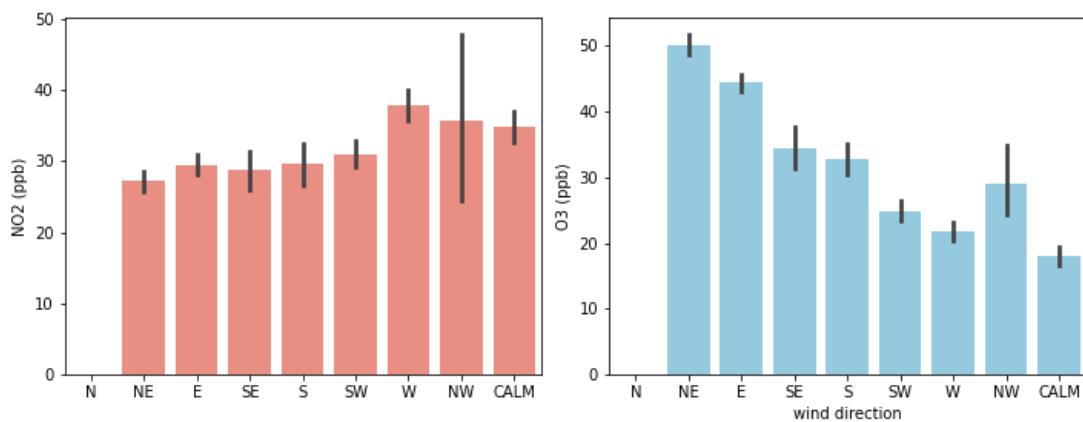


Εικόνα 75. Μέση τιμή συγκεντρώσεων ρύπων ανά διεύθυνση ανέμου (ΚΠΠΑ Airly)

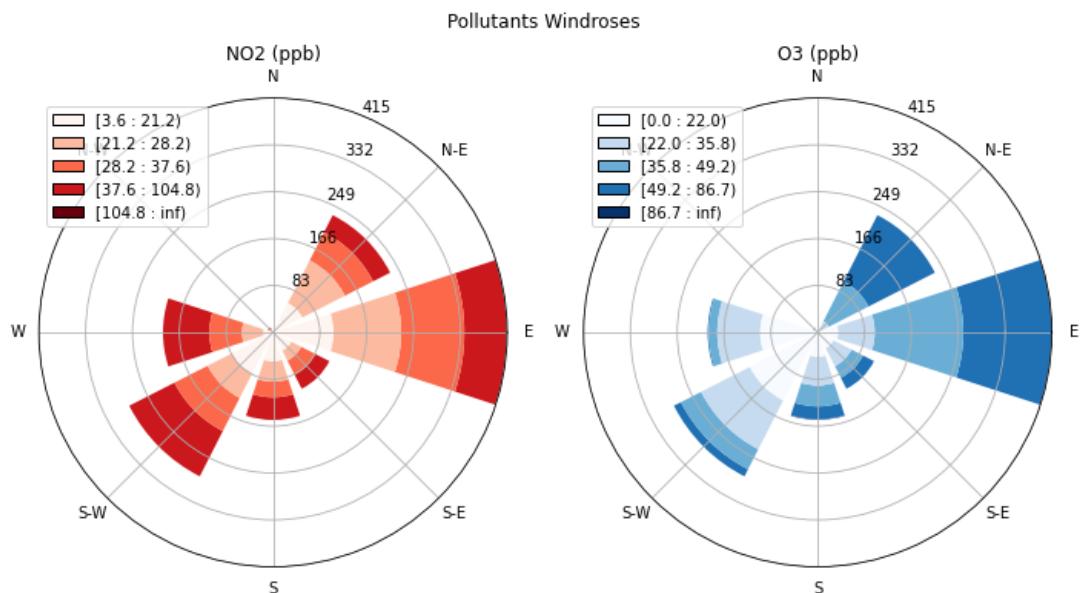


Εικόνα 76. Ροδογράμματα ρύπων ανά διεύθυνση και συχνότητα ανέμου (ΚΠΠΑ Airly)

Aqmesh

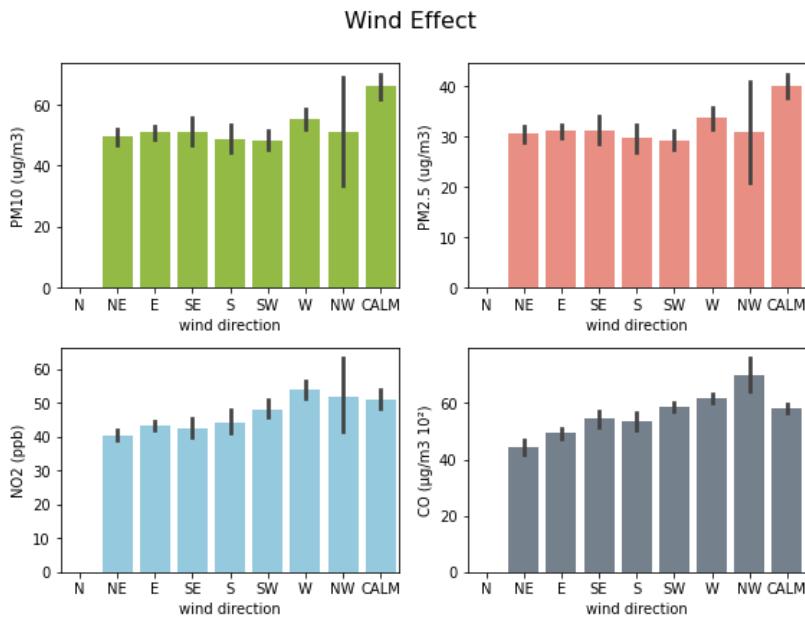


Εικόνα 77. Μέση τιμή συγκεντρώσεων ρύπων ανά διεύθυνση ανέμου (ΚΠΠΑ Aqmesh)

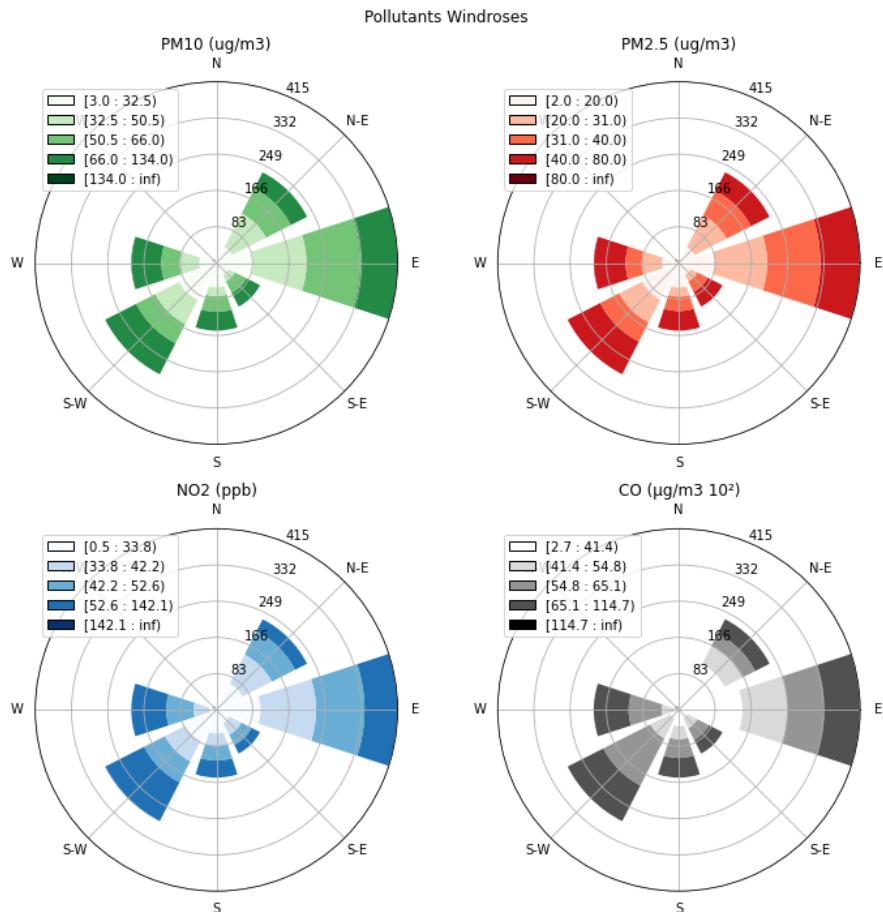


Εικόνα 78. Ροδογράμματα ρύπων ανά διεύθυνση και συχνότητα ανέμου (ΚΠΠΑ Aqmesh)

KASTOM

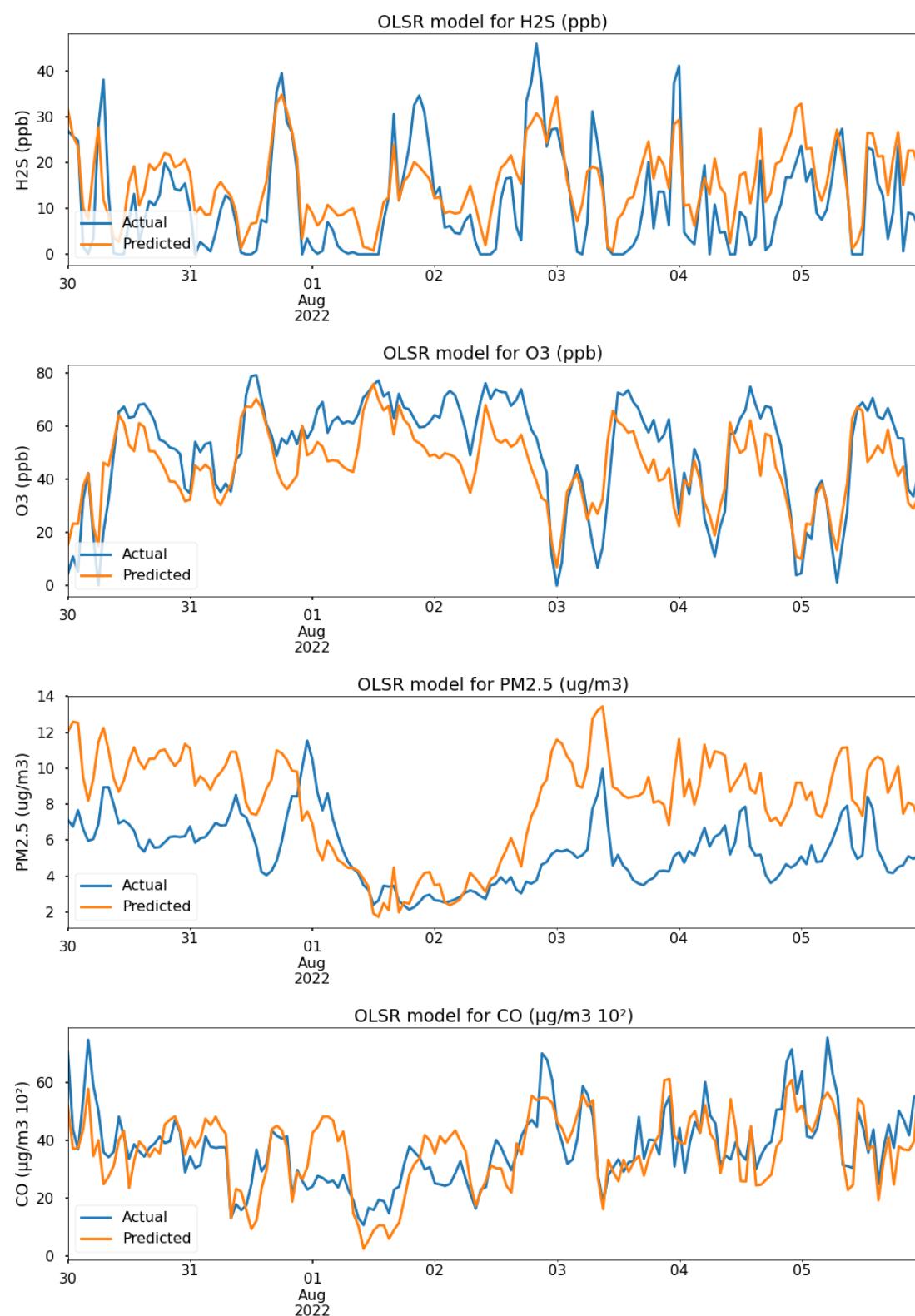


Εικόνα 79. Μέση τιμή συγκεντρώσεων ρύπων ανά διεύθυνση ανέμου (ΚΠΠΑ KASTOM)



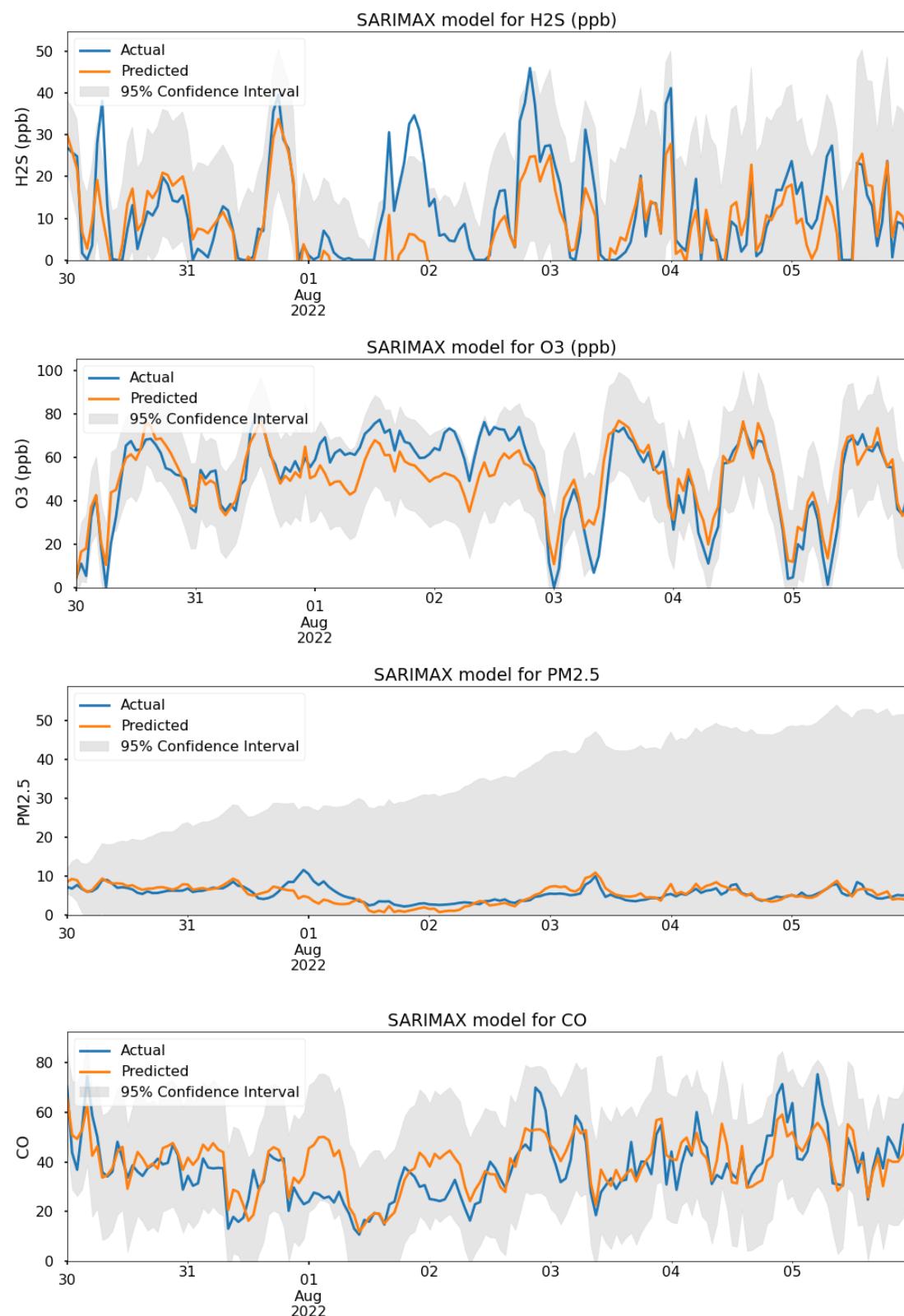
Εικόνα 80. Ροδογράμματα ρύπων ανά διεύθυνση και συχνότητα ανέμου (ΚΠΠΑ KASTOM)

6.1.9. Παλινδρόμηση ελαχίστων τετραγώνων



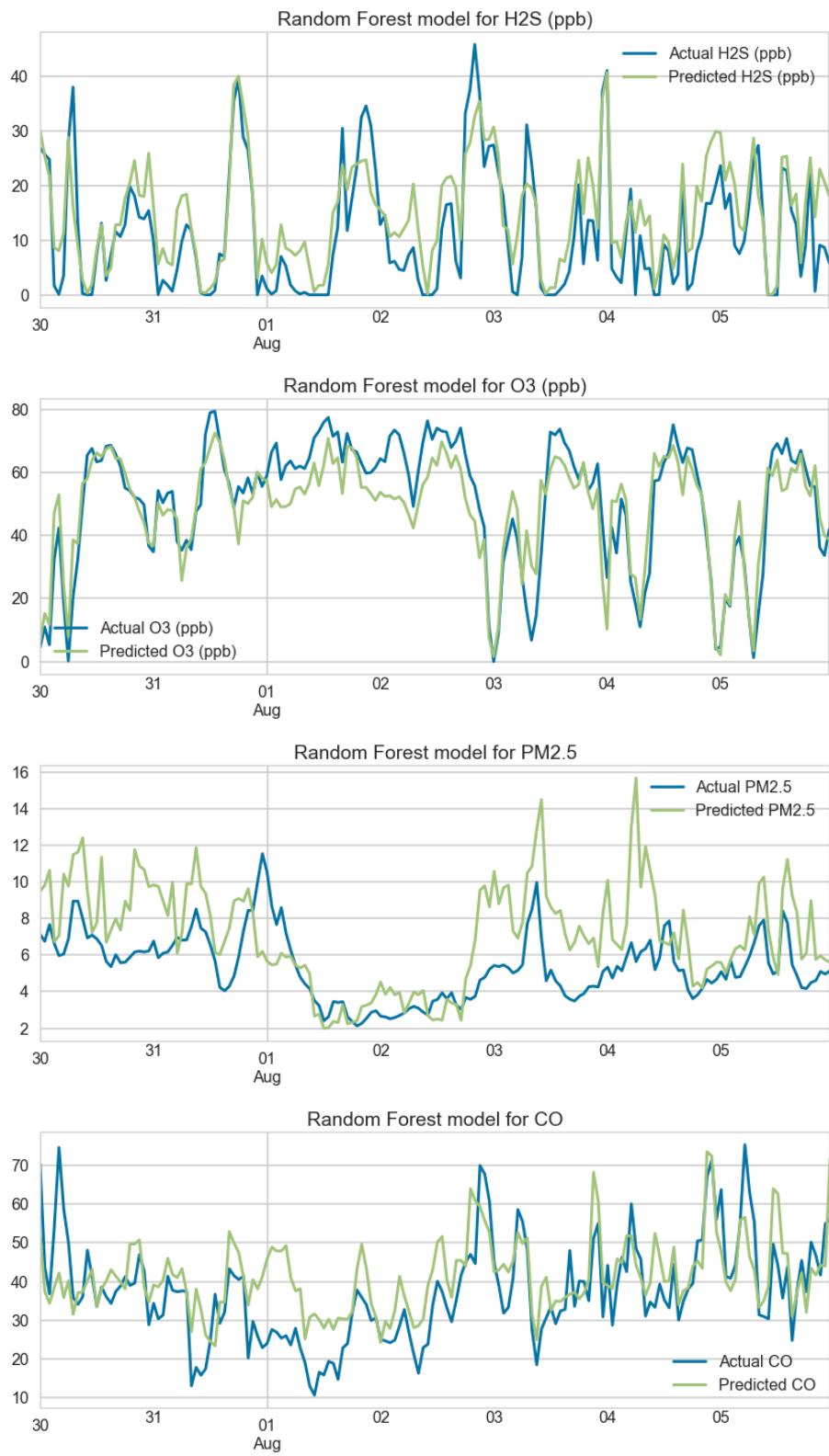
Εικόνα 81. Σύγκριση αποτελεσμάτων πρόγνωσης OLSR και πραγματικών δεδομένων (ΚΠΠΑ Kunak και KASTOM)

6.1.10. SARIMAX



Εικόνα 82. Σύγκριση αποτελεσμάτων πρόγνωσης μοντέλου SARIMAX και πραγματικών δεδομένων (ΚΠΠΑ Kunak και KASTOM)

6.1.11. Random Forest



Εικόνα 83. Σύγκριση αποτελεσμάτων πρόγνωσης μοντέλου Random Forest και πραγματικών δεδομένων (ΚΠΠΑ Kunak και KASTOM)