ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΡΟΒΛΕΨΗ ΤΗΣ ΑΤΜΟΣΦΑΙΡΙΚΗΣ ΡΥΠΑΝΣΗΣ ΜΕ ΧΡΗΣΗ ΧΡΟΝΟΣΕΙΡΩΝ ΚΑΙ ΜΕΘΟΔΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Διπλωματική Εργασία

Μπίτη Πολυξένη

Επιβλέπουσα: Τουσίδου Ελένη

Ιούλιος 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΡΟΒΛΕΨΗ ΤΗΣ ΑΤΜΟΣΦΑΙΡΙΚΗΣ ΡΥΠΑΝΣΗΣ ΜΕ ΧΡΗΣΗ ΧΡΟΝΟΣΕΙΡΩΝ ΚΑΙ ΜΕΘΟΔΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Διπλωματική Εργασία

Μπίτη Πολυξένη

Επιβλέπουσα: Τουσίδου Ελένη

Ιούλιος 2022

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

AIR POLLUTION PREDICTION USING TIME SERIES AND MACHINE LEARNING METHODS

Diploma Thesis

Biti Polyxeni

Supervisor: Tousidou Eleni

July 2022

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπουσα **Τουσίδου Ελένη**
Ε.ΔΙ.Π, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Βασιλακόπουλος Μιχαήλ**
Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Τσαλαπάτα Χαρίκλεια**
Ε.ΔΙ.Π, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ευχαριστίες

Η εργασία ολοκληρώθηκε με τη βοήθεια της επιβλέπουσας κυρίας Ελένης Τουσίδου, την οποία θα ήθελα να ευχαριστήσω θερμά για την καθοδήγηση και τις πολύτιμες συμβουλές της, καθώς και την άριστη επικοινωνία μας. Θα ήθελα επίσης να ευχαριστήσω τα μέλη της επιτροπής, κύριο Βασιλακόπουλο και κυρία Τσαλαπάτα για τη συμμετοχή τους. Τέλος, οφείλω το μεγαλύτερο ευχαριστώ στην οικογένεια μου και τους φίλους μου για τη στήριξη τους καθόλη τη διάρκεια των σπουδών μου και για τις ευκαιρίες που είχα χάρη σε αυτούς.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

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

Ο/Η Δηλών/ούσα

Μπίτη Πολυξένη

Διπλωματική Εργασία

ΠΡΟΒΛΕΨΗ ΤΗΣ ΑΤΜΟΣΦΑΙΡΙΚΗΣ ΡΥΠΑΝΣΗΣ ΜΕ ΧΡΗΣΗ ΧΡΟΝΟΣΕΙΡΩΝ ΚΑΙ ΜΕΘΟΔΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Μπίτη Πολυξένη

Περίληψη

Η αστικοποίηση και η βιομηχανοποίηση που πραγματοποιείται ολοένα και περισσότερο με την πάροδο των χρόνων εντείνουν το πρόβλημα της ατμοσφαιρικής ρύπανσης. Τα τελευταία χρόνια έχει δοθεί μεγαλύτερη βαρύτητα στο πρόβλημα της ρύπανσης, καθώς πλέον οι συνέπειές της γίνονται περισσότερο αντιληπτές καθημερινά. Η υγεία των ανθρώπων, ιδιαίτερα όσων κατοικούν γύρω απο μεγάλα αστικά κέντρα, είναι αποδεδειγμένα πολύ επιβαρυμένη εξαιτίας των συνθηκών ρύπανσης που αναγκάζονται να βιώνουν.

Λόγω των παραπάνω, η μελέτη και η πρόβλεψη της ατμοσφαιρικής ρύπανσης καθίσταται επείγουσα και απαραίτητη για τη βελτίωση της κατάστασης μακροπρόθεσμα. Η παρούσα εργασία, μελετά πραγματικά δεδομένα, τα οποία έχουν συλλεχθεί σε βάθος χρόνων, για διάφορες πόλεις. Στόχος είναι μέσω πειραματισμών με μεθόδους πρόβλεψης χρονοσειρών και μηχανικής μάθησης να γίνει έγκυρη πρόβλεψη για μελλοντικές τιμές της ατμοσφαιρικής ρύπανσης. Μελετώνται και προβλέπονται οι τιμές για πέντε ρυπαντικά στοιχεία, το όζον (O_3), το διοξείδιο του αζώτου (NO_2), το διοδείδιο του θείου (SO_2), τα σωματίδια $PM_{2.5}$ και τα σωματίδια PM_{10}. Η εργασία επίσης μελετά την επίδραση που έχουν οι καιρικές συνθήκες στα επίπεδα της ατμοσφαιρικής ρύπανσης, συγκρίνοντας την εγκυρότητα των προβλέψεων κατά την συμπερίληψη ή όχι στη μελέτη των μετεωρολογικών δεδομένων.

Λέξεις-κλειδιά:

Ατμοσφαιρική Ρύπανση, Μηχανική Μάθηση, Χρονοσειρές, Πρόβλεψη

Diploma Thesis

AIR POLLUTION PREDICTION

USING TIME SERIES

AND MACHINE LEARNING METHODS

Biti Polyxeni

Abstract

Urbanization and industrialization, which are appearing more intensively over the years, intensify the problem of air pollution. In recent years, more attention has been paid to the problem of air pollution, as its consequences are now becoming more visible every day. People's health, especially those living around large urban centers, is proven to be very burdensome due to the pollution conditions they are forced to experience.

Due to the above, the study and forecast of air pollution becomes urgent and necessary to improve the situation in the long run. This thesis studies real data, which have been collected over time, for a number of cities. The goal is to achieve the most valid prediction for future values of air pollution through experimentation with time series prediction methods and machine learning. The values of five pollutants are studied and predicted: ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), $PM_{2.5}$ particles and PM_{10} particles. The paper also studies the effect of weather conditions on air pollution levels, by comparing the validity of forecasts when meteorological data were included or not in the study.

Keywords:

Air Pollution, Machine Learning, Time Series, Prediction

Πίνακας περιεχομένων

Κατάλογος σχημάτων

Κατάλογος πινάκων

Συντομογραφίες

ADF	Augmented Dickey Fuller
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
AQI	Air Quality Index
ARIMA	AutoRegressive Integrated Moving Average
CNN	Convolutional Neural Network
LSTM	Long-Short Term Memory
MAE	Mean Absolute Error
MSE	Root Mean Squared Error
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
PM	Particle Matter
NO	Nitrogen Oxides
SO	Sulphur Oxides
O_3	Ozone

Κεφάλαιο 1

Εισαγωγή

Ο καθαρός αέρας είναι απαραίτητο αγαθό για όλα τα όντα του πλανήτη. Σήμερα όμως η ατμόσφαιρα είναι πιο επιβαρυμένη από ποτέ, λόγω των πολυάριθμων εργοστασιακών μονάδων, λόγω των οχημάτων που έχουν γίνει αναγκαία σε όλους σχεδόν τους εργαζόμενους αλλά και λόγω της καύσης υλικών για θέρμανση, και αυτοί είναι μόνο ορισμένοι από τους βασικούς παράγοντες που εντείνουν τη ρύπανση του αέρα. Είναι πλέον αποδεδειγμένο πως η ρύπανση του αέρα καταπονεί την υγεία των ζωντανών οργανισμών είτε βραχυπρόθεσμα είτε μακροπρόθεσμα. Άσθμα, πνευμονία, καρκίνος των πνευμόνων, αυξημένος κίνδυνος ανακοπής της καρδιάς και διάφορες δερματολογικές παθήσεις είναι μόνο μερικά παραδείγματα ασθενειών που φαίνεται να είναι στενά συνυφασμένες με την εκτεταμένη ρύπανση και φυσικά οδηγούν σε πρόωρους θανάτους.

Όλα τα προαναφερθέντα αποτέλεσαν τους λόγους για τους οποίους τα τελευταία χρόνια δίνεται ιδιαίτερη σημασία στη μελέτη της ατμοσφαιρικής ρύπανσης με στόχο την εύρεση μεθόδων και λειτουργιών που θα την αναχαιτίσουν και θα επισημάνουν τη σημασία του προβλήματος σε κάθε άκρη του πλανήτη. Η Ευρωπαϊκή Ένωση έχει θεσπίσει εδώ και αρκετά χρόνια κανόνες και νόμους, οι οποίοι αποσκοπούν στη μείωση της ατμοσφαιρικής ρύπανσης στην Ευρώπη. Στόχος είναι να μειωθούν οι πρόωροι θάνατοι από ασθένειες που σχετίζονται με την ατμοσφαιρική ρύπανση κατά 40% σε σχέση με τα επίπεδα της προηγούμενης δεκαετίας, αλλά και ο περιορισμός των δασικών εκτάσεων που υφίστανται βλάβες λόγω των ρύπων του αέρα.

Ένα απαραίτητο βήμα για την μείωση της ατμοσφαιρικής ρύπανσης είναι η πρόβλεψη των τιμών των ρύπων του αέρα για δεδομένο χρονικό διάστημα. Η έγκυρη πρόβλεψη είναι

αναγκαία για τη λήψη αποφάσεων και τη δημιουργία ενός πλάνου για την αναχαίτισή της, αλλά και για τον έλεγχο της αποτελεσματικότητας των πλάνων που ήδη βρίσκονται σε λειτουργία. Εκτός από την απόφαση για τις δράσεις που θα βελτιώσουν τα επίπεδα ρύπανσης, η πρόβλεψη μπορεί να ενημερώσει για τις περιόδους που η ποιότητα του αέρα είναι ασφαλέστερη για τη διενέργεια εξωτερικών δραστηριοτήτων.

Ωστόσο, η ποιοτική πρόβλεψη των τιμών της ατμοσφαιρικής ρύπανσης είναι αρκετά απαιτητική, καθώς βασίζεται εξ ολοκλήρου στην ύπαρξη αισθητήρων που παρέχουν συνεχείς και ακριβείς μετρήσεις σε πολλαπλά σημεία ανά τον πλανήτη. Επιπροσθέτως, τα επίπεδα ατμοσφαιρικής ρύπανσης εξαρτώνται από πολλούς εξωτερικούς παράγοντες, όπως για παράδειγμα οι καιρικές συνθήκες, που πρέπει να ληφθούν υπόψιν κατά τη διενέργεια της μελέτης.

1.1 Αντικείμενο της διπλωματικής

Σκοπός της συγκεκριμένης διπλωματικής είναι η πρόβλεψη της τιμής πέντε ρύπων του αέρα, του όζοντος (O_3), του διοξειδίου του αζώτου (NO_2), του διοξειδίου του θείου (SO_2), των σωματιδίων $PM_{2.5}$ και PM_{10}. Τα δεδομένα θα χρησιμοποιηθούν ως χρονοσειρές και οι προβλέψεις μελλοντικών τιμών θα γίνουν με τη μέθοδο ARIMA και Facebook Prophet για χρονοσειρές, αλλά και με τη χρήση νευρωνικών δικτύων όπως CNN και LSTM. Τα μοντέλα των νευρωνικών δικτύων έχουν δημιουργηθεί με δυο εκδοχές, μονομεταβλητά και πολυμεταβλητά. Στα μονομεταβλητά μοντέλα η πρόβλεψη των μελλοντικών τιμών γίνεται αποκλειστικά βάσει των παλαιότερων τιμών, ενώ στα πολυμεταβλητά μοντέλα οι προβλέψεις, εκτός των παλαιότερων τιμών, γίνονται και βάσει των καιρικών συνθηκών, πράγμα που φαίνεται να αυξάνει σημαντικά την ακρίβεια των προβλέψεων.

1.2 Οργάνωση του τόμου

Η παρούσα εργασία αποτελείται συνολικά απο έξι κεφάλαια. Το πρώτο κεφάλαιο αποτελεί την εισαγωγή και επεξήγηση των στόχων της διπλωματικής εργασίας. Έπειτα, στο δεύτερο κεφάλαιο, αναφέρονται και επεξηγούνται σύντομα εργασίες που έχουν ασχοληθεί με παρόμοια θεματολογία. Στο τρίτο κεφάλαιο γίνεται μια εισαγωγή στις έννοιες και τους όρους που απαιτούνται για την κατανόηση της μελέτης της παρούσας διπλωματικής. Το τέταρτο κεφάλαιο αποτελείται απο τη διαδικασία συλλογής, ανάλυσης και επεξεργασίας των δεδομένων

που χρησιμοποιήθηκαν για τα πειράματα της εργασίας, όπως επίσης και απο τα προγραμματιστικά εργαλεία με τα οποία αναπτύχθηκε η εργασία. Ακολουθεί το πέμπτο κεφάλαιο, στο οποίο γίνεται λεπτομερής επεξήγηση των μοντέλων προβλέψεων και της διαδικασίας που οδήγησε στην τελική τους μορφή, όπως επίσης και των αποτελεσμάτων που αυτά πέτυχαν. Τέλος, στο έκτο κεφάλαιο γίνεται η σύνοψη των μοντέλων και προκύπτουν τα τελικά συμπεράσματα.

Κεφάλαιο 2

Συναφείς Εργασίες

2.1 Εισαγωγή

Στο κεφάλαιο αυτό παρατίθενται προηγούμενες μελέτες που παρουσιάζουν τεχνικές που έχουν χρησιμοποιηθεί για την πρόβλεψη της ατμοσφαιρικής ρύπανσης, και έχουν δοκιμαστεί στην παρούσα διπλωματική.

2.2 Πρόβλεψη με μεθόδους χρονοσειρών

Το μοντέλο Prophet δημιουργήθηκε απο την Facebook με σκοπό να προβλέπει τις τιμές χρονοσειρών αυτόματα, είτε βραχυπρόθεσμα είτε μακροπρόθεσμα. Οι Justin Shen, Davesh Valagolam και Serena McCalla [1], σε έρευνα που δημοσίευσαν το 2020, χρησιμοποίησαν το μοντέλο Prophet, ώστε να προβλέψουν τα επίπεδα ατμοσφαιρικής ρύπανσης στη Σεούλ σε διάρκεια ενός έτους.

Εκτός απο το μοντέλο Prophet, ιδιαίτερα διαδεδομένο μοντέλο πρόβλεψης χρονοσειρών είναι και το ARIMA. Το μοντέλο αυτό έχει χρησιμοποιηθεί σε πολυάριθμες δοκιμές πρόβλεψης της ατμοσφαιρικής ρύπανσης, μια εκ των οποίων και η μελέτη του Ziyuan Ye [2], απο το Southern University of Science and Technology στην πόλη Σεντζέν της Κίνας. Στη μελέτη αυτή, υλοποιούνται και συγκρίνονται οι επιδόσεις των μοντέλων ARIMA και Prophet κατά την πρόβλεψη της ατμοσφαιρικής ρύπανσης της πόλης Σεντζέν, ενώ επίσης επιχειρείται να δημιουργηθεί ένα συνδυαστικό μοντέλο για μεγαλύτερη ακρίβεια.

Οι C. Guarnaccia, J. G. Ceron Breton, R. M. Ceron Breton, C. Tepedino, J. Quartieri, N. Mastorakis [3], διαθέτοντας ωριαία δεδομένα της συγκέντρωσης CO στην ατμόσφαιρα της

πόλης Μοντερέι του Μεξικό, χρησιμοποίησαν το μοντέλο ARIMA, ώστε να προβλέψουν τις τιμές της συγκέντρωσης για ένα 24ωρο.

2.3 Πρόβλεψη με μεθόδους μηχανικής μάθησης

Στη μελέτη [4] των Dewen Seng, Qiyan Zhang, Xuefeng Zhang, Guangsen Chen και Xiyuan Chen , γίνεται χρήση ενός LSTM νευρωνικού δικτύου με πολλαπλές εισόδους-εξόδους, εκπαιδευμένο με δεδομένα ατμοσφαιρικής ρύπανσης, τα οποία έχουν συλλεγεί από τέσσερις μετεωρολογικούς σταθμούς στο Πεκίνο. Το μοντέλο προέβλεψε τις τιμές πέντε ρυπογόνων στοιχείων της ατμόσφαιρας, $PM_{2.5}$, CO, NO_2, O_3 και SO_2, χρησιμοποιώντας τους δείκτες RMSE, MAE και R2 για την μέτρηση της ακρίβειας. Το μοντέλο δημιουργήθηκε ώστε να προβλέπει τα επίπεδα ρύπανσης μέσα στις επόμενες N ώρες και τα αποτελέσματα του συγκρίθηκαν με αυτά άλλων αλγορίθμων, όπως για παράδειγμα Linear Regression, Random Forest, SVM και ARMA και φαίνεται πως πέτυχε με διαφορά τη μεγαλύτερη ακρίβεια.

Επιπλέον, μια πρόσφατη δημοσίευση των Aysenur Gilik, Arif Selcuk Ogrenci και Atilla Ozmen [5], επιχειρεί να συνδυάσει τα μοντέλα CNN και LSTM με αποτέλεσμα ένα υβριδικό νευρωνικό δίκτυο για την πρόβλεψη της ατμοσφαιρική ρύπανσης ανά ώρα. Η ιδέα βασίζεται στο ότι τα CNN εξάγουν τις σχέσεις που υπάρχουν μεταξύ των χαρακτηριστικών του συνόλου δεδομένων που εξαρτώνται χωρικά ενώ τα LSTM αυτών που εξαρτώνται χρονικά. Όπως και στην προηγούμενη έρευνα, χρησιμοποιούνται δεδομένα για την ατμοσφαιρική ρύπανση τα οποία έχουν συλλεγεί από αισθητήρες σε πολλαπλές τοποθεσίες της ίδιας πόλης. Οι ερευνητές δημιούργησαν τρία μοντέλα. Το πρώτο είχε ως είσοδο τα ιστορικά δεδομένα ενός μόνο ρύπου και βάση αυτών προβλεπόταν η μελλοντική του τιμή. Το δεύτερο είχε πολλαπλές εισόδους που απαρτίζονταν από τις καιρικές συνθήκες και τις συγκεντρώσεις των ρύπων και μια μόνο έξοδο, τη μελλοντική τιμή ενός ρύπου. Το τρίτο μοντέλο ήταν παρόμοιο με το δεύτερο με τη διαφορά ότι είχε πολλαπλές εξόδους, μία για κάθε ρύπο. Το συμπέρασμα των δοκιμών ήταν πως τα μοντέλα με τις πολλαπλές εισόδους είχαν μεγαλύτερη ακρίβεια προβλέψεων.

Στη μεταπτυχιακή του έρευνα, ο Γεώργιος Καραμπέλας [6] προέβλεψε τις τιμές πέντε ατμοσφαιρικών ρύπων, $PM_{2.5}$, PM_{10}, NO_2, O_3 και SO_2 για την Αθήνα, την Νάπολη, το Γκντανσκ και το Λονδίνο, δημιουργώντας ένα συνδυαστικό πολυμεταβλητό νευρωνικό δίκτυο με πολλαπλές εξόδους, αποτελούμενο απο ένα μονοδιάστατο συνελικτικό και ένα LSTM δίκτυο διπλής κατεύθυνσης.

Κεφάλαιο 3

Θεωρητικό Υπόβαθρο

3.1 Εισαγωγή

Στο παρόν κεφάλαιο ορίζονται οι βασικές θεωρητικές έννοιες της εργασίας και αναλύονται τα εργαλεία και οι τεχνικές που χρησιμοποιήθηκαν για τη διεξαγωγή προβλέψεων.

3.2 Δείκτης ποιότητας του αέρα

Ο δείκτης ποιότητας του αέρα (AQI) είναι μια μονάδα μέτρησης της καθαρότητας της ατμόσφαιρας και περιγράφει πόσο επικίνδυνη είναι η έκθεση σε αυτήν. Ο δείκτης κινείται απο το 0 έως το 100. Όταν η τιμή του δείκτη κινείται πάνω απο το 80, η έκθεση στην ατμόσφαιρα καθίσταται ιδιαίτερα επιβλαβής για τον ανθρώπινο οργανισμό, αντίθετα, όταν η τιμή του δείκτη κυμαίνεται κάτω απο το 10, η ποιότητα του αέρα είναι ιδανική.

Ο δείκτης AQI περικλείει τις συγκεντρώσεις πέντε ρυπογόνων ουσιών: $PM_{2.5}$, PM_{10}, NO_2, O_3 και SO_2. Τα σωματίδια PM είναι μικροσκοπικά σωματίδια (στερεά ή υγρά) που υπάρχουν στον αέρα και πολλές φορές δεν μπορούμε να τα δούμε με γυμνό μάτι παρά μόνο με μικροσκόπιο. Ο όρος PM_{10} αναφέρεται σε σωματίδια που έχουν διάμετρο 10 μικρομέτρων ή μικρότερη ενώ ο όρος $PM_{2.5}$ αναφέρεται σε σωματίδια που έχουν διάμετρο 2.5 μικρομέτρων ή μικρότερη. Τα μεγάλα σωματίδια πάνω από 10mμ όπως η σκόνη δεν είναι τόσο επικίνδυνα για την υγεία μας καθώς προσκολλώνται στην αναπνευστική οδό και συνήθως δεν φτάνουν στους πνεύμονες. Τα μικροσκοπικά σωματίδια όμως κάτω από 2.5mμ είναι πολύ πιο επικίνδυνα. Φτάνουν στους πνεύμονες και από εκεί απευθείας στην κυκλοφορία του αίματος. Σύμφωνα με τους ειδικούς, η εισπνοή μεγάλης ποσότητας τέτοιων σωματιδίων

μπορεί να προκαλέσει εγκεφαλικό επεισόδιο, καρδιοπάθειες, καρκίνο του πνεύμονα και άλλες ασθένειες. Το διοξείδιο του θείου (SO_2), παράγεται κατά την καύση στερεών και υγρών καυσίμων. Μεγάλες επίσης ποσότητες διοξειδίου του θείου ελευθερώνονται στον αέρα κατά τις εκρήξεις των ηφαιστείων. Το όζον (O_3) χρησιμεύει για να προστατεύει τη γη από τις υπεριώδεις ακτίνες του ηλίου. Το επίπεδο μόλυνσης στην ατμόσφαιρα, σήμερα, είναι πάρα πολύ υψηλό και το όζον αντιδρά περισσότερο με τους ρύπους που βρίσκονται στην ατμόσφαιρα παρά με τις υπεριώδεις ακτίνες του ηλίου. Έτσι μειώνεται η ικανότητά του να μας προστατεύει. Το διοξείδιο του αζώτου (NO_2), παράγεται κατά τη λειτουργία των βενζινοκινητήρων. Με την επίδραση της ηλιακής ακτινοβολίας, από τα οξείδια του αζώτου παράγεται και όζον, το οποίο είναι ερεθιστικό αέριο.

	Good	Fair	Moderate	Poor	Very poor	Extremely poor
$PM_{2.5}$	0-10	10-20	20-25	25-50	50-75	75-800
PM_{10}	0-20	20-40	40-50	50-100	100-150	150-1200
NO_2	0-40	40-90	90-120	120-230	230-340	340-1000
O_3	0-50	50-100	100-130	130-240	240-380	380-800
SO_2	0-100	100-200	200-350	350-500	500-750	750-1250

Σχήμα 3.1: Air Quality Index.

3.3 Χρονοσειρές

Με τον όρο χρονοσειρά, αναφερόμαστε σε μια ακολουθία παρατηρήσεων με σταθερό βήμα δειγματοληψίας. Όπως σε κάθε άλλη ανάλυση δεδομένων, έτσι και στην ανάλυση των χρονοσειρών, θεωρείται πως τα δεδομένα επηρεάζονται απο έναν λευκό θόρυβο, ο οποίος μεταβάλλει τις τιμές της χρονοσειράς εντελώς τυχαία. Συγκεκριμένα, είναι απαραίτητο η χρονοσειρά που αναλύεται να είναι στάσιμη. Στάσιμη ονομάζεται η χρονοσειρά της οποίας η μέση τιμή και η διακύμανση είναι σταθερή και τα χαρακτηριστικά της δεν εξαρτώνται απο τη χρονική στιγμή την οποία αυτή εξετάζεται.

Οι περισσότερες χρονοσειρές μπορούν να περιγραφούν μέσω δυο βασικών τους στοιχείων, την εποχικότητα και την τάση. Ο όρος εποχικότητα αναφέρεται σε συγκεκριμένα μο-

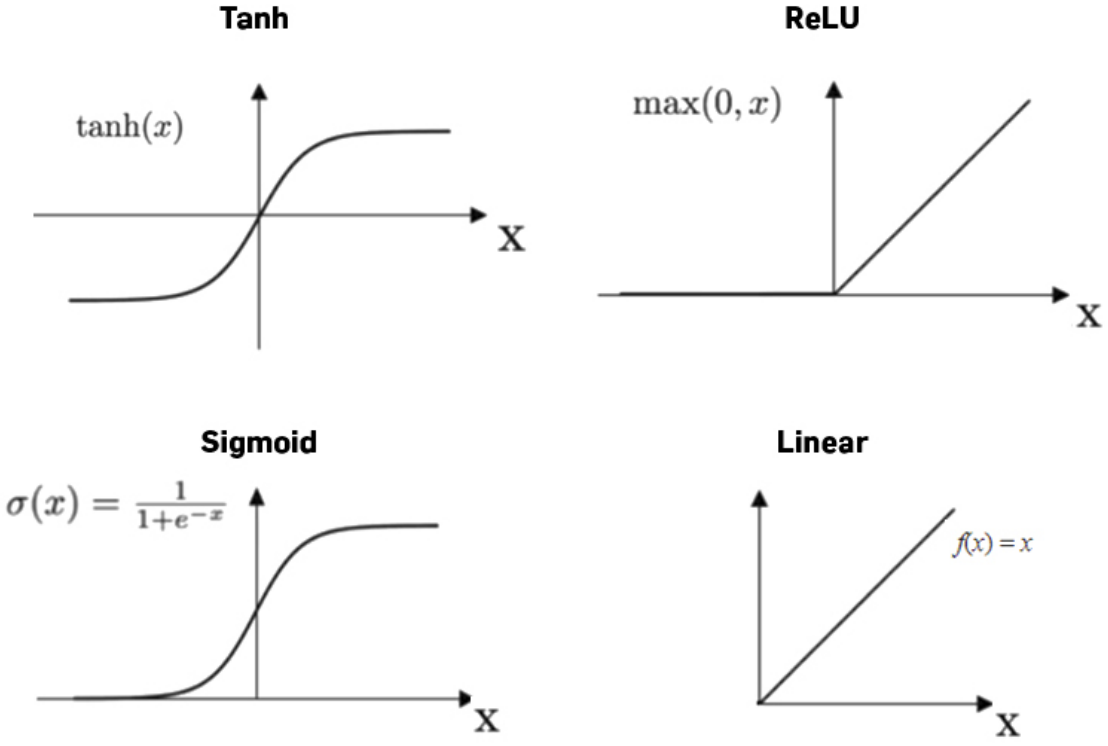
τίβα, τα οποία επαναλαμβάνονται συστηματικά στη διάρκεια του χρόνου. Ο όρος τάση αντιπροσωπεύει την αύξουσα ή φθίνουσα κλίση στη χρονοσειρά.

3.4 Τεχνητά νευρωνικά δίκτυα

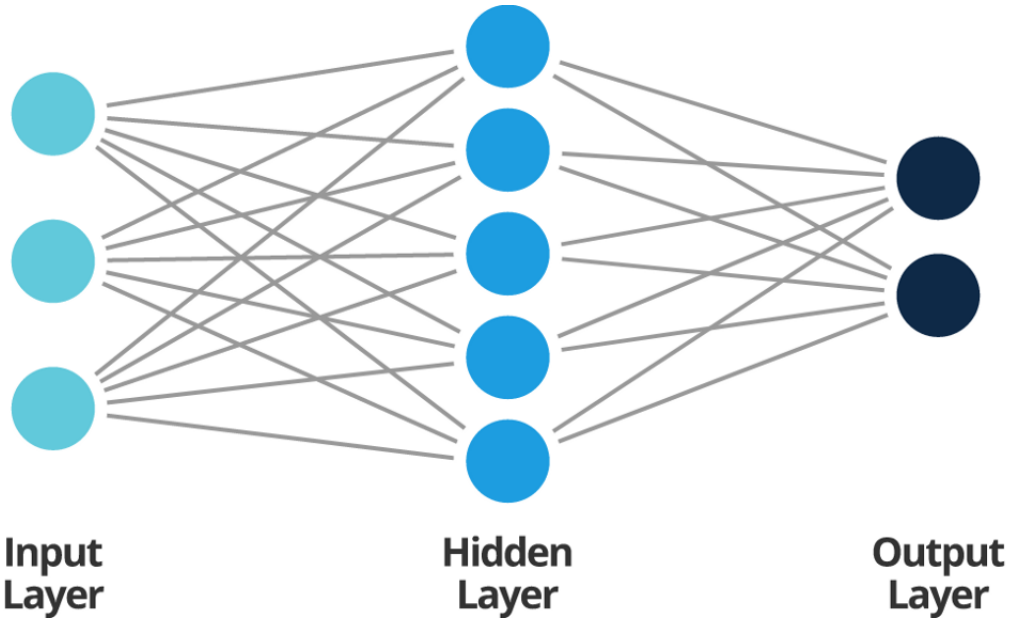
Τα τεχνητά νευρωνικά δίκτυα (ANN) είναι μαθηματικά μοντέλα τα οποία έχουν δημιουργηθεί βασισμένα στον τρόπο που λειτουργούν οι νευρώνες του ανθρώπινου εγκεφάλου. Τα ANN στοχεύουν στην αναγνώριση προτύπων σε ένα σύνολο δεδομένων. Τα ANN αποτελούνται απο τρία βασικά είδη νευρώνων, τα οποία εμπλουτίζονται ανάλογα με το συγκεκριμένο μοντέλο

- οι νευρώνες εισόδου
- οι υπολογιστικοί ή κρυμμένοι νευρώνες
- και οι νευρώνες εξόδου

Κάθε νευρώνας εκτελεί δυο λειτουργίες, τη συλλογή των εισόδων και την παραγωγή της εξόδου. Κάθε είσοδος πολλαπλασιάζεται με ένα βάρος και έπειτα προωθείται σε μια συνάρτηση ενεργοποίησης ώστε να μετασχηματιστεί σε μια έξοδο.



Σχήμα 3.2: Συναρτήσεις ενεργοποίησης.



Σχήμα 3.3: Αρχιτεκτονική τεχνητού νευρωνικού δικτύου.

Κάθε μια απο τις εισόδους συνδέεται με τους κρυμμένους νευρώνες, πολλαπλασιάζεται με ένα βάρος (weight) σε κάθε σύνδεση και προστίθεται ένας σταθερός όρος (bias), ο οποίος είναι ανεξάρτητος απο κάθε είσοδο και επηρεάζει την ακρίβεια της εξόδου.

Η πρόβλεψη μιας τιμής με τη χρήση ANN προϋποθέτει τη θέσπιση μιας συνάρτησης κόστους την οποία το νευρωνικό δίκτυο θα προσπαθεί να ελαχιστοποιήσει για να πετύχει το βέλτιστο αποτέλεσμα. Η συνάρτηση κόστους, ποσοτικοποιεί τη διαφορά μεταξύ της πραγματικής και της προβλεφθείσας τιμής που παράγεται απο το μοντέλο.

Βασικό χαρακτηριστικό των ANN είναι και το ποσοστό εκμάθησης, το οποίο ελέγχει σε τι βαθμό προσαρμόζονται τα βάρη του δικτύου με στόχο να ελαχιστοποιηθεί η συνάρτηση κόστους. Όσο πιο μικρή είναι η τιμή του ποσοστού εκμάθησης τόσο πιο αργά το μοντέλο κινείται προς το ολικό ελάχιστο της συνάρτησης κόστους. Σε περίπτωση επιλογής ενός πολύ μικρού ποσοστού εκμάθησης, η διαδικασία της εκπαίδευσης του μοντέλου γίνεται ιδιαίτερα χρονοβόρα και σε μερικές περιπτώσεις γίνεται ατέρμονη. Αντίθετα, η επιλογή ενός μεγάλου ποσοστού εκμάθησης πιθανόν να έχει ως αποτέλεσμα το μοντέλο να είναι ασταθές και να μην εντοπίσει ποτέ το ελάχιστο της συνάρτησης κόστους, επομένως να μην είναι βέλτιστο.

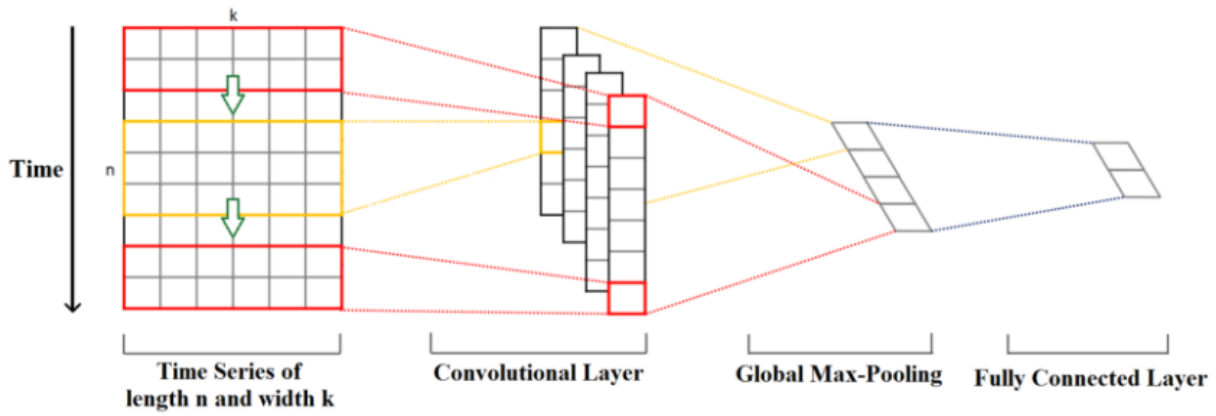
Μια πολύ σημαντική παράμετρος των νευρωνικών δικτύων είναι ο βελτιστοποιητής. Ο βελτιστοποιητής είναι αλγόριθμος που ορίζει ποσοτικά τις αλλαγές που θα πρέπει να συμβούν στο ποσοστό εκμάθησης και στα βάρη σε κάθε επανάληψη του μοντέλου.

3.5 Συνελικτικά νευρωνικά δίκτυα

Τα συνελικτικά νευρωνικά δίκτυα (CNN), ειδικεύονται στην επεξεργασία δεδομένων που η τοπολογία τους μοιάζει με πλέγμα, όπως, για παράδειγμα, μια εικόνα. Ωστόσο, είναι ιδιαίτερα πρακτικά και στην ανάλυση χρονοσειρών. Συγκεκριμένα, αντί να εξάγονται χωρικές πληροφορίες, τα CNN μιας διάστασης εξάγουν πληροφορίες κατά μήκος της διάστασης του χρόνου. Ένα CNN συνήθως αποτελείται απο τρία είδη επιπέδων. Το επίπεδο της συνέλιξης (convolutional layer), το επίπεδο υποδειγματοληψίας (pooling layer) και το πλήρως συνδεδεμένο επίπεδο (fully connected layer).

Το επίπεδο συνέλιξης είναι ο θεμέλιος λίθος των CNN και αναλαμβάνει τον μεγαλύτερο υπολογιστικό φόρτο του δικτύου. Αυτό το επίπεδο πολλαπλασιάζει δυο πίνακες ένας εκ των οποίων είναι τα φίλτρα (kernels) του δικτύου και ο άλλος είναι το κομμάτι των δεδομένων που γίνεται είσοδος στο δίκτυο. Η είσοδος έχει διαστάσεις n x k, όπου n ο αριθμός των χρονικών βημάτων που χρησιμοποιούνται για την πρόβλεψη, ενώ k είναι ο αριθμός των χαρακτηριστικών/μεταβλητών που λαμβάνονται υπόψιν. Ο πίνακας των φίλτρων έχει πάντα το πλάτος της εισόδου ενώ το μήκος μπορεί να διαφέρει, με αυτόν τον τρόπο, τα φίλτρα κινούνται προς μια κατεύθυνση απο την αρχή της χρονοσειράς ως το τέλος της. Τα στοιχεία του πίνακα των φίλτρων πολλαπλασιάζονται με τα αντίστοιχα στοιχεία της χρονοσειράς που καλύπτουν. Έπειτα, τα αποτελέσματα προστίθενται μεταξύ τους και εφαρμόζεται μια συνάρτηση ενεργοποίησης (σχήμα 3.2). Η τιμή που προκύπτει μετατρέπεται σε ενα νέο φιλτραρισμένο στοιχείο μιας μονομεταβλητής χρονοσειράς και η διαδικασία επαναλαμβάνεται έως ότου δημιουργηθούν τόσα νεα φιλτραρισμένα στοιχεία όσα και τα συνελικτικά φίλτρα.

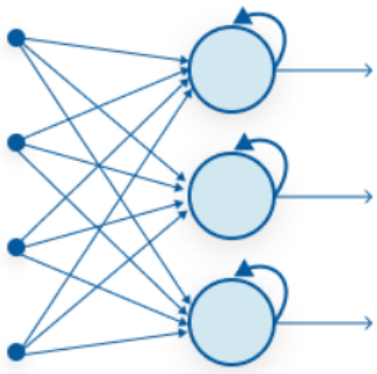
Ακολουθεί το επίπεδο υποδειγματοληψίας, ή αλλιώς max-pooling layer. Η φιλτραρισμένη χρονοσειρά που δημιουργήθηκε στο προηγούμενο επίπεδο χωρίζεται σε διανύσματα και ξεχωρίζεται η μεγαλύτερη τιμή απο καθένα απο αυτά. Ένα νέο διάνυσμα δημιουργείται απο τις μέγιστες τιμές και προωθείται σαν είσοδος στο πλήρως συνδεδεμένο επίπεδο.



Σχήμα 3.4: Αρχιτεκτονική συνελικτικού νευρωνικού δικτύου.

3.6 Ανατροφοδοτούμενα νευρωνικά δίκτυα

Τα ανατροφοδοτούμενα νευρωνικά δίκτυα (RNN) αποτελούν ενα είδος δικτύου που διατηρεί μνήμη με τις πρόσφατες πληροφορίες που έχει συλλέξει κατά τη διάρκεια της εκπαίδευσης του. Στα RNN, η έξοδος ενός προηγούμενου βήματος αποτελεί την είσοδο για το τρέχον βήμα της διαδικασίας εκπαίδευσης.



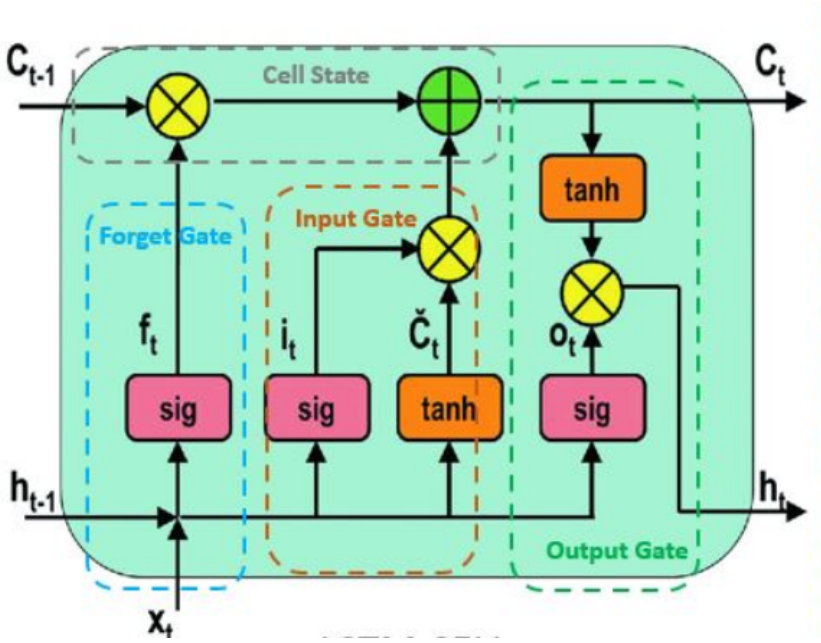
Σχήμα 3.5: Επαναλαμβανόμενη σύνδεση σε RNN.

Πιο συγκεκριμένα, τα κρυμμένα επίπεδα στα RNN παράγουν μια έξοδο την οποία ανατροφοδοτούν στο δίκτυο ως είσοδο για την επόμενη επανάληψη. Το δίκτυο ανανεώνει τα βάρη και για την τρέχουσα αλλά και για την προηγούμενη είσοδο. Έπειτα, πολλαπλασιάζει τις εισόδους με το αντίστοιχο βάρος και προωθεί το αποτέλεσμα ως είσοδο στο επόμενο επίπεδο, το οποίο εφαρμόζοντας μια συνάρτηση ενεργοποίησης (σχήμα 3.2) τροποποιεί τα

δεδομένα και παράγει τις προβλέψεις. Καθώς το δίκτυο παράγει προβλέψεις, παράλληλα υπολογίζει και το σφάλμα της πρόβλεψης, την απόκλιση απο την πραγματική τιμη. Το σφάλμα αυτό ελαχιστοποιείται με την κατάλληλη ανανέωση των βαρών μετά απο κάθε επανάληψη και η ανανέωση αυτή διαδίδεται μέσω της οπισθοδιάδοσης (backpropagation) σε όλο το δίκτυο.

3.6.1 Δίκτυα μακράς-βραχύχρονης μνήμης

Τα δίκτυα μακράς-βραχύχρονης μνήμης (LSTM) αποτελούν μια επέκταση των RNN δικτύων, τα οποία έχουν τη δυνατότητα να ανακαλύπτουν πρότυπα απο μεγαλύτερες ακολουθίες δεδομένων. Αποτελούνται απο κελιά μνήμης, πύλες εισόδου-εξόδου και πύλες επιλεκτικής συγκράτησης. Σε κάθε χρονικό βήμα, τα κελιά μνήμης συνδυάζουν γνώση απο την τρέχουσα είσοδο, τη βραχυπρόθεσμη και τη μακροπρόθεσμη μνήμη του δικτύου. Κάθε κόμβος ενός LSTM αποτελείται απο τέσσερα επίπεδα που αλληλεπιδρούν μεταξύ τους, όπως φαίνεται στο σχήμα 3.6.



Σχήμα 3.6: Εσωτερική δομή ενός LSTM κόμβου.

Η μαύρη γραμμή, η οποία ξεκινά απο τα αριστερά προς τα δεξιά αποτελεί την τρέχουσα κατάσταση του κόμβου. Η πρώτη πύλη είναι η πύλη επιλεκτικής συγκράτησης, η οποία μέσω μιας σιγμοειδούς συνάρτησης αποφασίζει εαν η τιμή εξόδου της προηγούμενης κατάστασης θα αγνοηθεί. Η έξοδος απο την πύλη επιλεκτικής συγκράτησης (f_t) πολλαπλασιάζεται με την προηγούμενη κατάσταση του κόμβου (C_{t-1}) και εαν το αποτέλεσμα είναι μηδέν, τότε η τιμή

θα αγνοηθεί.

Έπειτα, ακολουθεί η πύλη εισόδου κατά την οποία η τρέχουσα είσοδος και η προηγούμενη έξοδος περνούν μέσα απο μια δεύτερη σιγμοειδή συνάρτηση ώστε η έξοδος της (i_t), στη συνέχεια, να πολλαπλασιαστεί με την έξοδο ($\hat{C}_t$) απο μια εφαπτομενική συνάρτηση με την ίδια είσοδο. Το αποτέλεσμα του πολλαπλασιασμού αυτού προστίθεται με το γινόμενο $f_t \cdot C_{t-1}$ και συνολικά αποτελούν τη νέα κατάσταση του κόμβου.

Τέλος, ακολουθεί η πύλη εξόδου, κατά την οποία οι τιμές της τρέχουσας εισόδου και της προηγούμενης εξόδου ύστερα απο μια σιγμοειδή συνάρτηση πολλαπλασιάζονται με την έξοδο μιας εφαπτομενικής συνάρτησης απο την οποία περνά η νέα κατάσταση του κόμβου, το αποτέλεσμα απο την πράξη αυτή είναι η πληροφορία που θα περαστεί στο κρυφό επίπεδο του επόμενου βήματος της εκπαιδευσης του δικτύου.

3.7 ARIMA

Το μοντέλο ARIMA είναι ένα στοχαστικό μαθηματικό μοντέλο το οποίο μας βοηθάει να μελετήσουμε και να προβλέψουμε την εξέλιξη ενός μεγέθους. Αποτελεί την πιο διαδεδομένη μέθοδο πρόβλεψης χρονοσειρών, βάσει των τιμών του μεγέθους που έχουν εμφανιστεί σε προηγούμενες χρονικές περιόδους. Για να εφαρμοστεί το μοντέλο σε κάποια χρονοσειρά θα πρέπει να προσδιοριστούν τρεις συγκεκριμένοι παράγοντες.

Οι παράγοντες του μοντέλου ARIMA είναι οι p ή αλλιώς τάξη AR, q ή αλλιώς τάξη MA, d ή αλλιώς τάξη διαφόρισης. Το p αντιστοιχίζεται στο πλήθος των προηγούμενων τιμών που θα χρησιμοποιηθούν για την πρόβλεψη, το q αντιστοιχίζεται στο πλήθος των περιόδων των οποίων το σφάλμα απο την πρόβλεψη θα ληφθεί υπόψιν για τη νέα πρόβλεψη και τέλος, το d είναι το πλήθος των διαφορίσεων που θα πρέπει να υποστεί η χρονοσειρά προκειμένου να μετατραπεί σε στάσιμη.

3.7.1 Augmented Dickey-Fuller Test

Το Augmented Dickey-Fuller Test (ADFT) είναι ένα στατιστικό τεστ, το οποίο ελέγχει εάν μια χρονοσειρά διαθέτει μοναδιαία ρίζα, οπότε και η χρονοσειρά δεν θεωρείται στάσιμη. Το τεστ ελέγχει εάν η λύση της παρακάτω εξίσωσης ως προς α είναι η μονάδα.

$$y_t = c + \alpha y_{t-1} + \beta t + \phi \Delta y_{t-1} + error,$$

όπου y_t είναι η τιμή της χρονοσειράς τη χρονική στιγμή t, y_{t-1} είναι η πρώτη καθυστέρηση της χρονοσειράς και $\phi \Delta Y_{t-1}$ είναι η διαφορά της χρονοσειράς τις χρονικές στιγμές t και t-1. Συγκεκριμένα, ελέγχονται δυο υποθέσεις

$$H_0 : \alpha = 1$$

$$H_1 : \alpha \neq 1$$

Στην περίπτωση που ικανοποιείται η πρώτη υπόθεση, η χρονοσειρά δεν είναι στάσιμη και χρειάζεται να γίνει διαφόριση της ενώ στην περίπτωση που ικανοποιείται η δεύτερη υπόθεση η χρονοσειρά είναι ήδη στάσιμη.

3.8 Facebook Prophet

Το 2017, η Facebook δημιούργησε το μοντέλο Prophet για τις γλώσσες R και Python. Είναι ένα ιδιαίτερα πετυχημένο μοντέλο πρόβλεψης για τις χρονοσειρές που εμφανίζουν κάποια εποχικότητα. Η πρόβλεψη μιας μελλοντικής τιμής της χρονοσειράς μέσω αυτού του μοντέλου, γίνεται βάση ενός αθροιστικού τύπου ο οποίος έχει ως εξής:

$$y_t = g_t + s_t + h_t + e_t$$

Οι όροι του παραπάνω αθροίσματος δηλώνουν τα εξής, g_t είναι η τάση που έχει η χρονοσειρά η οποία μπορεί να είναι ανοδική, καθοδική ή εναλλασσόμενη, ο όρος s_t ονομάζεται συνάρτηση εποχικότητας και πρόκειται για μια σειρά Fourier ως προς το χρόνο, h_t είναι μια συνάρτηση η οποία βοηθάει το μοντέλο να προσαρμόζει τις προβλέψεις όταν αυτές αναφέρονται σε εορτασικές μέρες ή ξεχωριστά γεγονότα εντός του χρόνου και τέλος e_t είναι ο όρος σφάλματος.

3.9 Μονομεταβλητή και πολυμεταβλητή ανάλυση

Ο όρος μονομεταβλητή ανάλυση (Univariate Analysis) σημαίνει πως η πρόβλεψη τιμών μιας χρονοσειράς, γίνεται βάση των τιμών της χρονοσειράς σε προηγούμενα χρονικά βήματα, χωρίς να λαμβάνονται υπόψιν οι τιμές των άλλων χαρακτηριστικών/μεταβλητών. Για παράδειγμα, για την πρόβλεψη μελλοντικών τιμών της συγκέντρωσης O_3 στην ατμόσφαιρα, θα χρησιμοποιηθούν μόνο παρελθοντικές τιμές της συγκέντρωσης.

Η πολυμεταβλητή ανάλυση (Multivariate Analysis) απο την άλλη, λαμβάνει υπόψιν πάνω άπο δυο χαρακτηριστικά/μεταβλητές για τη διεξαγωγή προβλέψεων που αφορούν μελλοντικές τιμές. Τα πολυμεταβλητά μοντέλα αποδεικνύονται πιο ακριβή διότι αξιοποιούν τις σχέσεις εξάρτησης της μεταβλητής για την οποία γίνονται οι προβλέψεις με τις υπόλοιπες μεταβλητές, κάτι που τα μονομεταβλητά μοντέλα αγνοούν με αποτέλεσμα να χάνεται πληροφορία.

3.10 Μετρικές αξιολόγησης μοντέλων

Κατά την εξέταση της απόδοσης οποιουδήποτε μοντέλου, πρέπει να αξιολογούνται οι τιμές πρόβλεψης που παράγει. Η αξιολόγηση γίνεται με τον υπολογισμό των κατάλληλων μετρικών σφάλματος. Η μέτρηση σφάλματος είναι ένας τρόπος για να ποσοτικοποιηθεί η απόδοση ενός μοντέλου και να συγκριθεί αντικειμενικά με την απόδοση διαφορετικών μοντέλων. Στην παρούσα διπλωματική, για την αξιολόγηση των μοντέλων έχουν χρησιμοποιηθεί τρεις μετρικές αξιολόγησης. Η πρώτη μετρική είναι το μέσο τετραγωνικό σφάλμα (MSE), η οποία ακολουθεί τον μαθηματικό τύπο

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}\Big(y' - y\Big)^2,$$

όπου y' η προβλεφθείσα και y η πραγματική τιμή, όσο πιο κοντά στο μηδέν είναι η τιμή της συγκεκριμένης μετρικής, τόσο πιο ακριβές θεωρείται το μοντέλο. Η επόμενη μετρική είναι η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE), η οποία ακολουθεί τον μαθηματικό τύπο

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(y' - y\Big)^2},$$

όπου y' η προβλεφθείσα και y η πραγματική τιμή. Το σφάλμα αυτό είναι πάντα θετικό και οι μικρότερες τιμές σημαίνουν πιο ποιοτικό μοντέλο. Τέλος, η μετρική του μέσου απόλυτου σφάλματος (MAE) με μαθηματικό τύπο

$$MAE = \frac{1}{n}\Sigma_{i=1}^{n}\Big(|y' - y|\Big),$$

όπου y' η προβλεφθείσα και y η πραγματική τιμή. Όσο μικρότερη είναι η τιμή αυτής της μετρικής τόσο ακριβέστερο είναι το μοντέλο, ωστόσο, το MAE δεν υποδεικνύει το σχετικό μέγεθος του σφάλματος και καθίσταται δύσκολο να διαφοροποιηθούν τα μεγάλα από τα μικρά σφάλματα, γι' αυτόν το λόγο χρησιμοποιείται συνδυαστικά με τις προηγούμενες δυο μετρικές αξιολόγησης.

Κεφάλαιο 4

Δεδομένα και εργαλεία

4.1 Python

Όλες οι μελέτες, οι στατιστικές αναλύσεις και τα μοντέλα δημιουργήθηκαν μέσω της γλώσσας προγραμματισμού *Python*. Η *Python* αποτελεί μια απο τις πιο διαδεδομένες και εύχρηστες γλώσσες προγραμματισμού στην επιστήμη και την ανάλυση δεδομένων, προσφέροντας πολυάριθμες βιβλιοθήκες με αυτόματες μεθόδους. Συγκεκριμένα, για την ανάλυση των δεδομένων χρησιμοποιήθηκαν οι βιβλιοθήκες *sklearn*, *statsmodels* και *pandas*, ενώ για την οπτικοποίηση τους μέσω γραφικών παραστάσεων η βιβλιοθήκη *matplotlib*.

4.2 Jupyter Notebook

Το *Jupyter Notebook* αποτελεί μια web εφαρμογή ανοιχτού κώδικα, η οποία μπορεί να χρησιμοποιηθεί για τη δημιουργία αρχείων κώδικα. Χρησιμοποιείται ιδιαίτερα συχνά κατά την ανάλυση δεδομένων καθώς δίνει τη δυνατότητα στους χρήστες να χωρίζουν τον κώδικα σε ανεξάρτητα κομμάτια και να οπτικοποιούν τα αποτελέσματα διαδραστικά.

4.3 Συλλογή δεδομένων

Όλα τα δεδομένα αντλήθηκαν απο την πλατφόρμα Air Quality Open Data Platform, στην οποία βρίσκονται συγκεντρωμένα τα δεδομένα για την ποιότητα του αέρα απο το 2015 έως και το 2022, σε ημερήσια βάση, για περίπου 380 πόλεις ανά τον κόσμο. Τα δεδομένα για κάθε πόλη, βασίζονται στον μέσο όρο απο τις μετρήσεις πολλαπλών σταθμών. Παρέχον-

ται τιμές για την ελάχιστη, τη μέγιστη, τη διάμεση τιμή καθώς και την τυπική απόκλιση για κάθε ένα απο τα είδη ατμοσφαιρικών ρύπων (NO_2, SO_2, $O3$, PM_{10}, $PM_{2.5}$) αλλά και μετεωρολογικών δεδομένων (θερμοκρασία, υγρασία, ταχύτητα αέρα, κατεύθυνση αέρα). Όλες οι τιμές έχουν προσαρμοστεί στην κλίμακα του US EPA. Εκτός απο τις προαναφερθείσες τιμές, δίνεται και η πληροφορία για το πλήθος των δειγμάτων που χρησιμοποιούνται για τον υπολογισμό της διάμεσης και της τυπικής απόκλισης.

Ωστόσο, τα δεδομένα που αφορούν τις καιρικές συνθήκες ήταν αρκετά ελλιπή, συνεπώς ήταν απαραίτητο να συλλεχθούν απο άλλη πηγή. Τα καιρικά δεδομένα συγκεντρώθηκαν απο την πλατφόρμα Visual Crossing και ανταποκρίνονται στα ίδια χρονικά διαστήματα με τα δεδομένα της ατμοσφαιρικής ρύπανσης, για κάθε πόλη που μελετήθηκε.

Για την πραγματοποίηση των πειραμάτων της εργασίας έχουν επιλεγεί τέσσερις μεγάλες πόλεις, δυο Ευρωπαϊκές, η Αθήνα και το Παρίσι, και δυο Ασιατικές, το Πεκίνο και το Δελχί. Επιλέχθηκαν πόλεις οι οποίες δεν είχαν ιδιαίτερα γεωγραφικά και κλιματικά κοινά χαρακτηριστικά μεταξύ τους, με σκοπό τα συμπεράσματα των μοντέλων να είναι γενικευμένα. Τα δεδομένα που αφορούν την Αθήνα εκτείνονται στο διάστημα απο 07-11-2019 έως 18-03-2022, ενώ τα δεδομένα για τις υπόλοιπες πόλεις εκτείνονται σε πολύ μεγαλύτερο διάστημα, συγκεκριμένα απο 30-12-2014 έως 19-03-2022.

	no2	o3	pm10	pm25	so2	temp	dew	humidity	windspeed	winddir
2019-11-07	17.4	27.7	48.0	54.0	1.1	21.1	16.1	74.20	20.0	160.1
2019-11-08	13.3	27.7	31.0	54.0	1.6	20.0	14.5	71.00	18.7	117.1
2019-11-09	13.8	26.9	17.0	41.0	1.6	17.2	11.9	72.24	15.4	136.7
2019-11-10	8.3	30.1	12.0	33.0	1.6	18.2	14.5	79.00	18.2	213.3
2019-11-11	11.5	22.8	18.0	36.0	1.6	17.1	13.1	77.69	6.9	174.5
...	...	...	...	...	...	...	...	...	...	...
2022-03-15	11.9	31.3	24.0	61.0	4.6	7.7	-0.9	56.10	20.1	229.4
2022-03-16	13.8	31.7	30.0	74.0	5.1	8.3	3.1	70.50	20.0	172.8
2022-03-17	11.0	33.8	32.0	70.0	3.6	11.0	5.2	68.90	25.7	133.7
2022-03-18	7.4	36.2	22.0	53.0	2.1	7.2	-1.4	54.90	32.0	81.1
2022-03-19	6.9	36.6	18.0	46.0	1.1	5.5	-5.2	46.60	33.0	53.0

Σχήμα 4.1: Τα δεδομένα πριν την κανονικοποίηση.

	no2	o3	pm10	pm25	so2	temp	dew	humidity	windspeed	winddir
2014-12-30	0.393531	0.100872	0.054435	0.168860	0.298901	0.363248	0.313274	0.263279	0.453718	0.431864
2014-12-31	0.067385	0.255293	0.037298	0.046053	0.081319	0.279915	0.166372	0.130485	1.000000	0.997627
2015-01-01	0.319407	0.120797	0.059476	0.311404	0.320879	0.254274	0.176991	0.173210	0.180577	0.622373
2015-01-02	0.405660	0.069738	0.067540	0.309211	0.356044	0.303419	0.244248	0.218245	0.180577	0.414915
2015-01-03	0.603774	0.029888	0.118952	0.462719	0.868132	0.254274	0.313274	0.420323	0.016692	0.320000
...	...	...	...	...	...	...	...	...	...	...
2022-03-15	0.129380	0.287671	0.058468	0.243421	0.010989	0.510684	0.539823	0.498845	0.235205	0.279661
2022-03-16	0.117251	0.247821	0.070565	0.250000	0.010989	0.529915	0.571681	0.517321	0.235205	0.323051
2022-03-17	0.091644	0.247821	0.030242	0.092105	0.021978	0.350427	0.460177	0.640878	0.125948	0.394915
2022-03-18	0.146900	0.120797	0.045363	0.254386	0.021978	0.303419	0.518584	0.939954	0.071320	0.397966
2022-03-19	0.074124	0.292653	0.014113	0.129386	0.021978	0.356838	0.423009	0.518476	0.180577	0.231525

Σχήμα 4.2: Τα δεδομένα μετά την κανονικοποίηση.

4.4 Προεπεξεργασία δεδομένων

Αρχικά, απο το σύνολο δεδομένων που αφορούν την ατμοσφαιρική ρύπανση, δημιουργείται μια νέα στήλη για κάθε ατμοσφαιρικό ρύπο που θα προβλεφθεί, με χρήση της συνάρτησης *pivot_table*. Η συνάρτηση χρησιμοποιεί ένα μοναδικό γνώρισμα για κάθε γραμμή συνδυάζοντας συγκεκριμένες στήλες που καθορίζονται απο την παράμετρο *index*, προκειμένου να διασπάσει τις στήλες που καθορίζονται απο την παράμετρο *columns* σε πλήθος στηλών ίσο με το πλήθος των μοναδικών τιμών κάθε μίας. Συγκεκριμένα, χρησιμοποιούνται οι στήλες 'Date', 'Country' και 'City' ως μοναδικά αναγνωριστικά κάθε γραμμής προκειμένου να διασπαστεί η στήλη 'Specie', η οποία διαθέτει όλους τους ρύπους που μελετώνται, και κάθε ρύπος να βρεθεί σε ξεχωριστή στήλη. Ακολουθεί η αφαίρεση των στηλών που δεν αφορούν τους ατμοσφαιρικούς ρύπους που θα προβλεφθούν μέσω της συνάρτησης *drop*. Οι τιμές των στηλών που έχουν απομείνει μετατρέπονται σε αριθμητικές και ελέγχονται για τυχόν κενές τιμές, οι οποίες αντικαθίστανται με την αμέσως προηγούμενη έγκυρη τιμή του συγκεκριμένου ρύπου, μέσω της συνάρτησης *bfill*. Τέλος, αφαιρέθηκαν οι διπλότυπες γραμμές.

Έπειτα, το σύνολο δεδομένων που περιείχε τα καιρικά στοιχεία για κάθε πόλη, ελέγχθηκε για κενές τιμές, οι οποίες αντικαταστάθηκαν με τη συνάρτηση *bfill*, όπως και στα δεδομένα της ατμοσφαιρικής ρύπανσης. Στη συνέχεια, τα καιρικά δεδομένα συμπτύχθηκαν με τα δεδομένα των ατμοσφαιρικών ρύπων ώστε να αντιμετωπιστούν ως ενιαίο σύνολο. Τα δεδομένα του ενιαίου, πλέον, συνόλου κανονικοποιήθηκαν προκειμένου όλες οι στήλες να

βρίσκονται στην ίδια αριθμητική κλίμακα, όπως φαίνεται στους πίνακες 4.1 και 4.2, ώστε να είναι συγκρίσιμες οι επιδόσεις των μοντέλων για την πρόβλεψη καθενός απο τους ρύπους. Τέλος, καθώς πρόκειται για δεδομένα χρονοσειρών, μελετήθηκαν ως προς την εποχικότητα και την τάση τους ανα ημέρα (σχήμα 4.5), εβδομάδα (σχήμα 4.6), μήνα (σχήμα 4.7) και χρόνο (σχήμα 4.8), με χρήση της συνάρτησης *seasonal_decompose* και της επαναδειγματοληπτικής συνάρτησης *resample*. Η συνάρτηση *seasonal_decompose* διασπά κάθε χρονοσειρά που λαμβάνει ως παράμετρο σε ένα άθροισμα τριών συνιστωσών, της τάσης, της εποχικότητας και των τυχαίων υπολειμμάτων.

Για κάθε μια πόλη που επιλέχθηκε για μελέτη, τα δεδομένα χωρίστηκαν σε δυο υποσύνολα, αυτό που θα χρησιμοποιηθεί για την εκπαίδευση των μοντέλων, το οποίο αντιστοιχεί στο 80% του συνόλου, και αυτό που θα χρησιμοποιηθεί για τον έλεγχο της απόδοσης τους, δηλαδή το υπόλοιπο 20%. Εκτός απο τις ημερήσιες προβλέψεις, οι οποίες δεν παρουσίαζαν κάποια εποχικότητα, έγιναν και προβλέψεις που αφορούν τα εβδομαδιαία και μηνιαία δεδομένα, τα οποία εμφάνισαν ισχυρή εποχικότητα για όλες τις μελετηθείσες πόλεις. Τα διαστήματα εκπαίδευσης και ελέγχου για όλες τις συχνότητες κάθε πόλης, συγκεντρώνονται στους πίνακες που ακολουθούν.

Πόλη	Διάστημα Εκπαίδευσης	Διάστημα Ελέγχου
Αθήνα	07-11-2019 έως 07-09-2021	08-09-2021 έως 19-03-2022
Παρίσι	29-12-2014 έως 10-02-2021	11-02-2021 έως 18-03-2022
Πεκίνο	30-12-2014 έως 07-02-2021	08-02-2021 έως 19-03-2022
Δελχί	29-12-2014 έως 10-02-2021	11-02-2021 έως 19-03-2022

Πίνακας 4.1: Χωρισμός δεδομένων σε εκπαίδευσης και ελέγχου για ημερήσιες προβλέψεις

Πόλη	Διάστημα Εκπαίδευσης	Διάστημα Ελέγχου
Αθήνα	10-11-2019 έως 26-09-2021	03-10-2021 έως 20-03-2022
Παρίσι	04-01-2015 έως 04-10-2020	11-10-2020 έως 20-03-2022
Πεκίνο	04-01-2015 έως 04-10-2020	11-10-2020 έως 20-03-2022
Δελχί	04-01-2015 έως 04-10-2020	11-10-2020 έως 20-03-2022

Πίνακας 4.2: Χωρισμός δεδομένων σε εκπαίδευσης και ελέγχου για εβδομαδιαίες προβλέψεις

Πόλη	Διάστημα Εκπαίδευσης	Διάστημα Ελέγχου
Αθήνα	30-11-2019 έως 30-09-2021	31-10-2021 έως 31-03-2022
Παρίσι	31-12-2014 έως 30-09-2020	31-10-2020 έως 31-03-2022
Πεκίνο	31-12-2014 έως 30-09-2020	31-10-2020 έως 31-03-2022
Δελχί	31-12-2014 έως 30-09-2020	31-10-2020 έως 31-03-2022

Πίνακας 4.3: Χωρισμός δεδομένων σε εκπαίδευσης και ελέγχου για μηνιαίες προβλέψεις

4.5 Οπτικοποίηση των δεδομένων

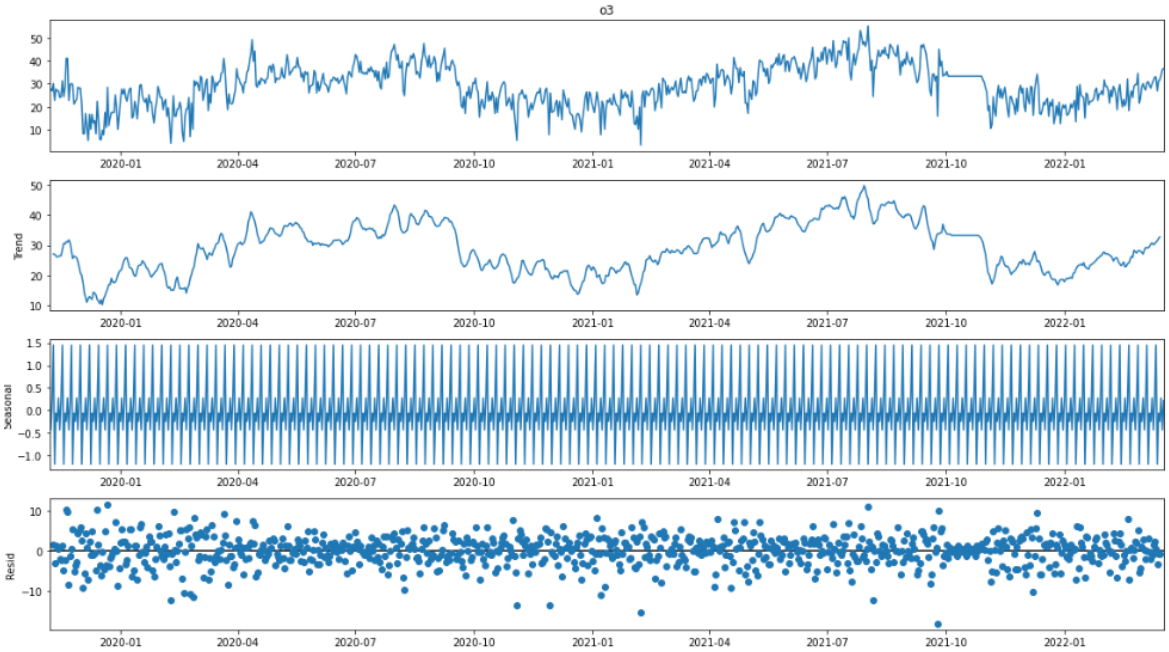
Για κάθε πόλη γίνεται, με τη χρήση της συνάρτησης *resample*, εβδομαδιαία (σχήμα 4.3), μηνιαία (σχήμα 4.4) και ετήσια επαναδειγματοληψία προκειμένου να εντοπιστούν μοτίβα που επαναλαμβάνονται κατά χρονικά διαστήματα. Η συνάρτηση *resample*, ομαδοποιεί τα διαθέσιμα ημερήσια δεδομένα κατά εβδομαδιαία και μηνιαία συχνότητα χρησιμοποιώντας το μέσο όρο κάθε διαστήματος. Προκειμένου να κατανοηθούν βαθύτερα τα δεδομένα χρειάστηκε να γίνει η αναπαράστασή τους γραφικά μέσω της συνάρτησης *seasonal_decompose*, η λειτουργία της οποίας επεξηγήθηκε στην προηγούμενη υποενότητα και αποτυπώνεται στα σχήματα 4.5-4.8.

	no2	o3	pm10	pm25	so2	temp	dew	humidity	windspeed	winddir
2019-11-10	0.420074	0.477842	0.377049	0.271429	0.029762	0.551036	0.803371	0.690846	0.290685	0.414695
2019-11-17	0.397770	0.442885	0.360656	0.523810	0.045351	0.436179	0.752274	0.858292	0.113796	0.447874
2019-11-24	0.236856	0.534544	0.288056	0.529252	0.028345	0.419273	0.745853	0.863483	0.182013	0.422389
2019-12-01	0.370685	0.435178	0.201405	0.438095	0.039683	0.455199	0.636704	0.609312	0.258489	0.536693
2019-12-08	0.423792	0.178915	0.484778	0.438095	0.090703	0.336855	0.522204	0.657618	0.255736	0.526997
...	...	...	...	...	...	...	...	...	...	...
2022-02-20	0.550186	0.376548	0.386417	0.434014	0.175737	0.305579	0.490102	0.683971	0.246559	0.582281
2022-02-27	0.322358	0.481971	0.292740	0.323810	0.153061	0.327980	0.484216	0.612641	0.446620	0.595846
2022-03-06	0.297398	0.458574	0.264637	0.337415	0.102041	0.256128	0.401819	0.625520	0.330682	0.498577
2022-03-13	0.224642	0.529590	0.177986	0.277551	0.124717	0.149197	0.290530	0.653061	0.390639	0.594823
2022-03-20	0.322181	0.567437	0.338798	0.411111	0.178571	0.208087	0.269039	0.500000	0.454675	0.372250

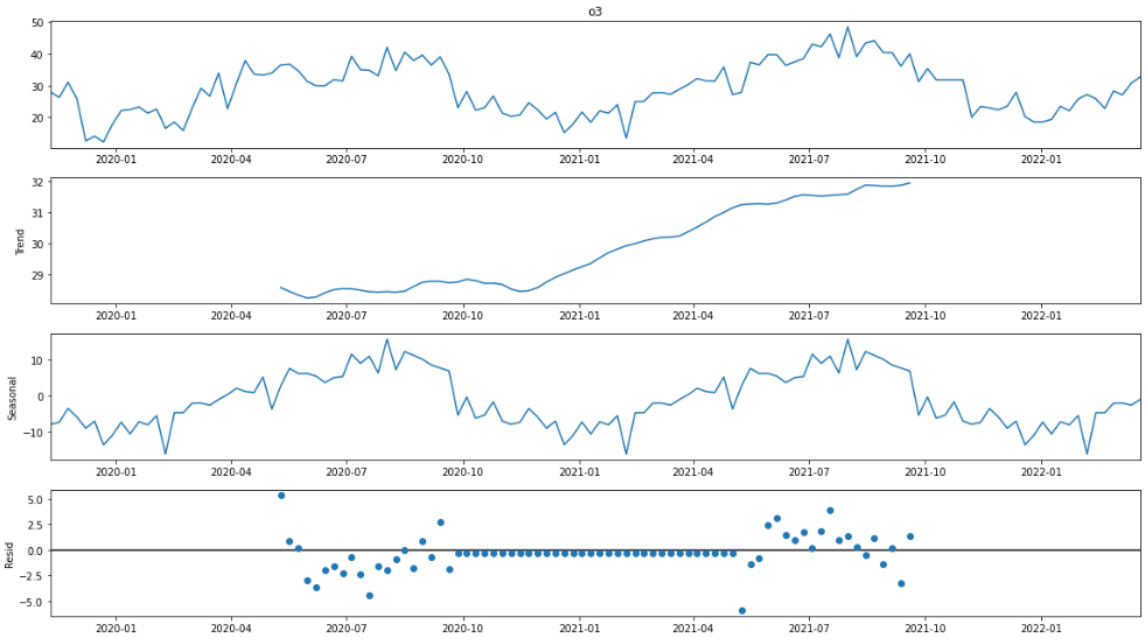
Σχήμα 4.3: Εβδομαδιαία επαναδειγματοληψία.

	no2	o3	pm10	pm25	so2	temp	dew	humidity	windspeed	winddir
2019-11-30	0.348358	0.471741	0.299863	0.461905	0.038029	0.459689	0.739076	0.775324	0.194593	0.458606
2019-12-31	0.398489	0.217540	0.352723	0.363441	0.139529	0.326971	0.523620	0.692148	0.277820	0.629856
2020-01-31	0.371507	0.376655	0.253305	0.335791	0.324117	0.252815	0.364262	0.564136	0.375561	0.634184
2020-02-29	0.372260	0.292007	0.231204	0.290969	0.336617	0.301265	0.443368	0.605385	0.338330	0.591757
2020-03-31	0.372227	0.484741	0.227393	0.269432	0.248336	0.356747	0.500060	0.594962	0.370657	0.477122
2020-04-30	0.258984	0.594990	0.224590	0.295873	0.125661	0.413807	0.480275	0.465039	0.412277	0.456559
2020-05-31	0.425351	0.602399	0.332099	0.296774	0.202509	0.590380	0.631147	0.390944	0.347448	0.459557
2020-06-30	0.235812	0.539949	0.230055	0.248889	0.255291	0.705030	0.760924	0.393084	0.415346	0.511748
2020-07-31	0.292121	0.638386	0.297726	0.331797	0.261137	0.812082	0.786758	0.268355	0.514126	0.498222
2020-08-31	0.303394	0.679533	0.269170	0.329339	0.291859	0.807501	0.802706	0.296094	0.408925	0.510957
2020-09-30	0.329740	0.555877	0.323497	0.306667	0.374603	0.740138	0.796130	0.372589	0.462455	0.565868

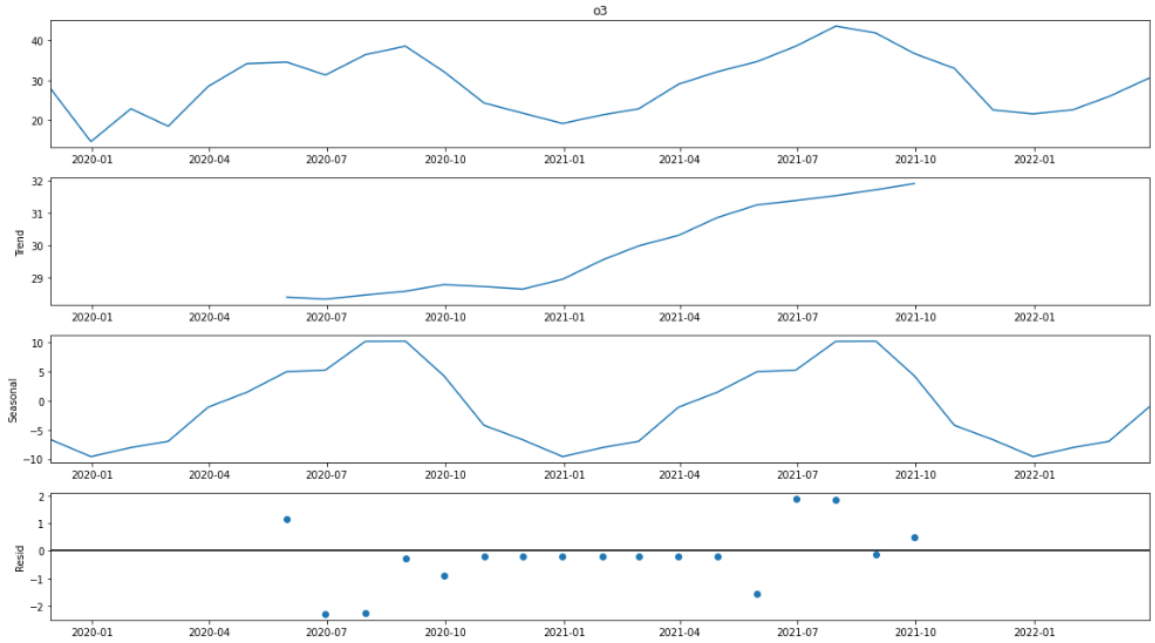
Σχήμα 4.4: Μηνιαία επαναδειγματοληψία.



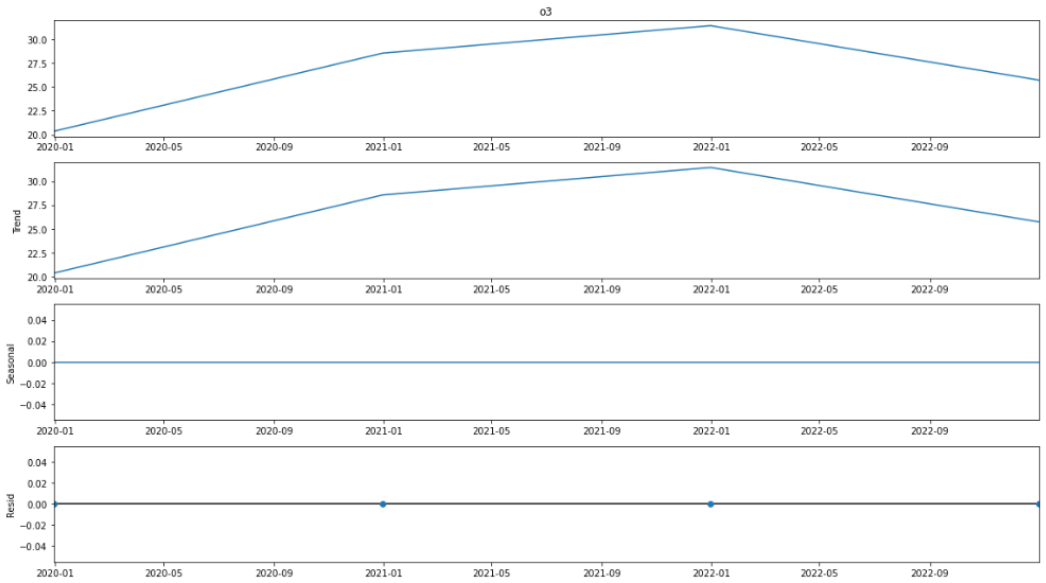
Σχήμα 4.5: Ανάλυση εποχικότητας και τάσης ημερήσιων δεδομένων O_3 για την Αθήνα.



Σχήμα 4.6: Ανάλυση εποχικότητας και τάσης εβδομαδιαίων δεδομένων O_3 για την Αθήνα.



Σχήμα 4.7: Ανάλυση εποχικότητας και τάσης μηνιαίων δεδομένων O_3 για την Αθήνα.



Σχήμα 4.8: Ανάλυση εποχικότητας και τάσης ετήσιων δεδομένων O_3 για την Αθήνα.

Καθένα απο τα σχήματα 4.5-4.8 αποτελείται απο τέσσερα γραφήματα. Το πρώτο γράφημα αφορά την αναπαράσταση των γνήσιων δεδομένων της χρονοσειράς, στην προκείμενη ενός απο τους πέντε ατμοσφαιρικούς ρύπους, το δεύτερο γράφημα (Trend) απεικονίζει την τάση των δεδομένων, το τρίτο γράφημα (Seasonal) αποτυπώνει την εποχικότητα των δεδομένων, και τέλος, στο τέταρτο γράφημα (Resid) απεικονίζεται το υπόλοιπο των γνήσιων δεδομένων όταν αφαιρεθούν απο αυτά η τάση και η εποχικότητα.

Κατά την ανάλυση των χρονοσειρών για κάθε πόλη, παρατηρήθηκε πως όλες οι χρονοσειρές, εκτός απο αυτές των συγκεντρώσεων NO_2 και SO_2 της Αθήνας, εμφάνιζαν συγκεκριμένη κατεύθυνση τάσης, όπως φαίνεται στον πίνακα 4.4. Επίσης, παρατηρείται πως για όλες τις χρονοσειρές κάθε πόλης τα εβδομαδιαία και μηνιαία δεδομένα εμφανίζουν ισχυρά μοτίβα εποχικότητας όπως φαίνεται και στα σχήματα 4.6, 4.7, τα οποία αναφέρονται στα δεδομένα της Αθήνας.

Πόλη	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
Αθήνα	↑	-	-	↑	↑
Παρίσι	↑	↓	↓	↓	↓
Πεκίνο	↑	↓	↓	↓	↓
Δελχί	↑	↓	↓	↓	↓

Πίνακας 4.4: Τάση χρονοσειρών.

Κεφάλαιο 5

Σχεδιασμός και ανάπτυξη μοντέλων πρόβλεψης

5.1 Εισαγωγή

Σε αυτό το κεφάλαιο, αναλύεται ο τρόπος με τον οποίο σχεδιάστηκαν τα μοντέλα πρόβλεψης, πώς καθορίστηκαν οι παράμετροί τους, καθώς και ο τρόπος με τον οποίο έγινε η επεξεργασία των δεδομένων ειδικά για κάθε μοντέλο. Οι μετρικές που χρησιμοποιήθηκαν για την μέτρηση της ακρίβειας είναι κοινές για όλα τα μοντέλα και αυτές είναι η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE), το μέσο τετραγωνικό σφάλμα (MSE) και το μέσο απόλυτο σφάλμα (MAE).

5.2 ARIMA

5.2.1 auto-ARIMA

Το μοντέλο auto-ARIMA αποτελεί αυτοματοποιημένη μέθοδο της στατιστικής βιβλιοθήκης *pmdarima* της *Python*, η οποία λαμβάνει ως είσοδο μια μονομεταβλητή χρονοσειρά και υπολογίζει, ύστερα απο δοκιμές, τον πιο αποδοτικό συνδυασμό των παραμέτρων AR, MA και τάξης διαφόρισης. Η εκπαίδευση του μοντέλου πραγματοποιείται καλώντας τη συνάρτηση *auto_arima* και έπειτα με τη συνάρτηση *predict* γίνονται προβλέψεις για το διάστημα που ορίζεται απο την παράμετρο της συνάρτησης αυτής.

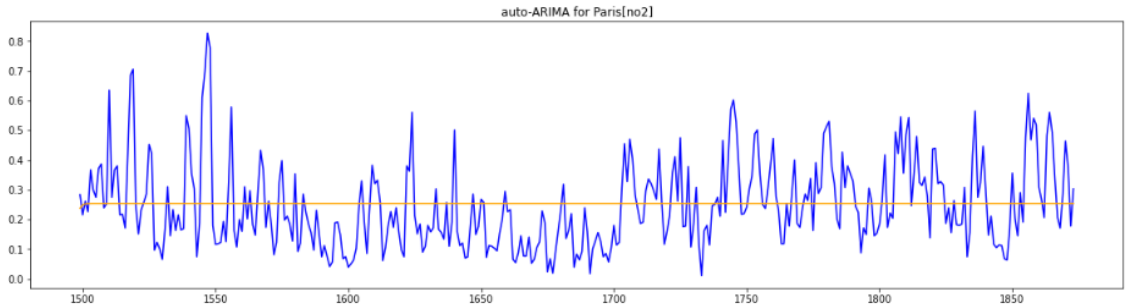
Δε χρειάστηκε κάποια ιδιαίτερη επεξεργασία των δεδομένων για το μοντέλο καθώς πρό-

κειται για ένα μονομεταβλητό μοντέλο που χρειάζεται μόνο τις ιστορικές τιμές της χρονοσειράς που μελετάται.

Οι προβλέψεις αφορούν την ημερήσια συχνότητα και τα σύνολα εκπαίδευσης και ελέγχου δεδομένων για κάθε πόλη ακολουθούν τον πίνακα 4.1.

```
Performing stepwise search to minimize aic
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=-1950.103, Time=1.36 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=-1568.577, Time=0.23 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=-1609.430, Time=0.18 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=-1683.998, Time=0.27 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=-1570.574, Time=0.06 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=-1953.571, Time=1.57 sec
 ARIMA(0,1,2)(0,0,0)[0] intercept   : AIC=-1930.190, Time=0.72 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=-1934.934, Time=0.96 sec
 ARIMA(1,1,3)(0,0,0)[0] intercept   : AIC=-1947.065, Time=1.12 sec
 ARIMA(0,1,3)(0,0,0)[0] intercept   : AIC=-1949.180, Time=0.91 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=-1950.456, Time=1.07 sec
 ARIMA(2,1,3)(0,0,0)[0] intercept   : AIC=-1950.241, Time=1.45 sec
 ARIMA(1,1,2)(0,0,0)[0]             : AIC=-1955.272, Time=0.39 sec
 ARIMA(0,1,2)(0,0,0)[0]             : AIC=-1932.019, Time=0.27 sec
 ARIMA(1,1,1)(0,0,0)[0]             : AIC=-1936.532, Time=0.30 sec
 ARIMA(2,1,2)(0,0,0)[0]             : AIC=-1954.040, Time=0.52 sec
 ARIMA(1,1,3)(0,0,0)[0]             : AIC=-1954.133, Time=0.62 sec
 ARIMA(0,1,1)(0,0,0)[0]             : AIC=-1685.981, Time=0.13 sec
 ARIMA(0,1,3)(0,0,0)[0]             : AIC=-1950.931, Time=0.43 sec
 ARIMA(2,1,1)(0,0,0)[0]             : AIC=-1952.170, Time=0.30 sec
 ARIMA(2,1,3)(0,0,0)[0]             : AIC=-1951.929, Time=0.75 sec

Best model:  ARIMA(1,1,2)(0,0,0)[0]
Total fit time: 13.628 seconds
auto-ARIMA MSE for Paris[no2]:0.020943482592867782
auto-ARIMA RMSE for Paris[no2]:0.14471863250068315
auto-ARIMA MAE for Paris[no2]:0.11544716015786874
```



Σχήμα 5.1: Αποτελέσματα πρόβλεψης (πορτοκαλί γραμμή) auto-ARIMA για την ημερήσια συγκέντρωση NO_2 στο Παρίσι.

Στο σχήμα 5.1 φαίνονται τα μοντέλα που εξετάστηκαν ώστε τελικά να επιλεχθεί αυτό με τον μικρότερο δείκτη AIC, ο οποίος προσεγγίζει το ποσοστό πληροφορίας που χάνεται στο εκάστοτε μοντέλο· όσο πιο μικρός είναι ο δείκτης τόσο λιγότερη πληροφορία χάνεται και το μοντέλο θεωρείται πιο ποιοτικό.

Ομοίως, επαναλήφθηκε η διαδικασία για τη συγκέντρωση των πέντε ατμοσφαιρικών ρύπων και για τις τέσσερις πόλεις που μελετήθηκαν, όπου παρατηρήθηκε παρόμοια συμπεριφορά των μοντέλων.

Οι προβλέψεις μέσω του μοντέλου auto-ARIMA δεν ήταν ιδιαίτερα ικανοποιητικές. Αυτό πιθανότατα οφείλεται στην αυτοματοποίηση της διαδικασίας, γεγονός που αφενός την απλουστεύει, αφετέρου όμως αφαιρεί τη δυνατότητα προσαρμογής των παραμέτρων στις

ανάγκες της εκάστοτε χρονοσειράς.

5.2.2 Manual ARIMA

Το πρόβλημα της χαμηλής ακρίβειας λόγω αυτοματοποίησης του μοντέλου auto-ARIMA λύθηκε με την δημιουργία ενός ξεχωριστού μοντέλου ARIMA του οποίου οι παράμετροι υπολογίζονται ύστερα απο ανάλυση της εκάστοτε χρονοσειράς.

Αρχικά, κάθε μια χρονοσειρά που αντιστοιχεί σε έναν ατμοσφαιρικό ρύπο εξετάζεται ως προς τη στασιμότητα με τη βοήθεια της συνάρτησης *adfuller*, η οποία εφαρμόζει το Augmented Dickey Fuller τεστ για στασιμότητα επαναληπτικά για κάθε μια χρονοσειρά των ρύπων. Για να επιβεβαιωθεί το αποτέλεσμα του τεστ, χρησιμοποιείται η συνάρτηση *ndiffs* της βιβλιοθήκης *pmdarima*, η οποία αυτόματα υπολογίζει και επιστρέφει πόσες φορές θα πρέπει να διαφοριστεί η χρονοσειρά προκειμένου να μετατραπεί σε στάσιμη. Τα αποτελέσματα του ADF τεστ για κάθε πόλη είχαν ως εξής:

```
Check if Athens_Data is stationary :
p-value for no2 : 0.000000
p-value for o3 : 0.185124
p-value for pm10 : 0.000000
p-value for pm25 : 0.000000
p-value for so2 : 0.000001
```

Σχήμα 5.2: Τεστ στασιμότητας των χρονοσειρών της Αθήνας.

```
Check if Paris_Data is stationary :
p-value for no2 : 0.000088
p-value for o3 : 0.000176
p-value for pm10 : 0.000000
p-value for pm25 : 0.000000
p-value for so2 : 0.000694
```

Σχήμα 5.3: Τεστ στασιμότητας των χρονοσειρών του Παρισιού.

```
Check if Delhi_Data is stationary :
p-value for no2 : 0.000491
p-value for o3 : 0.003891
p-value for pm10 : 0.000009
p-value for pm25 : 0.000003
p-value for so2 : 0.000114
```

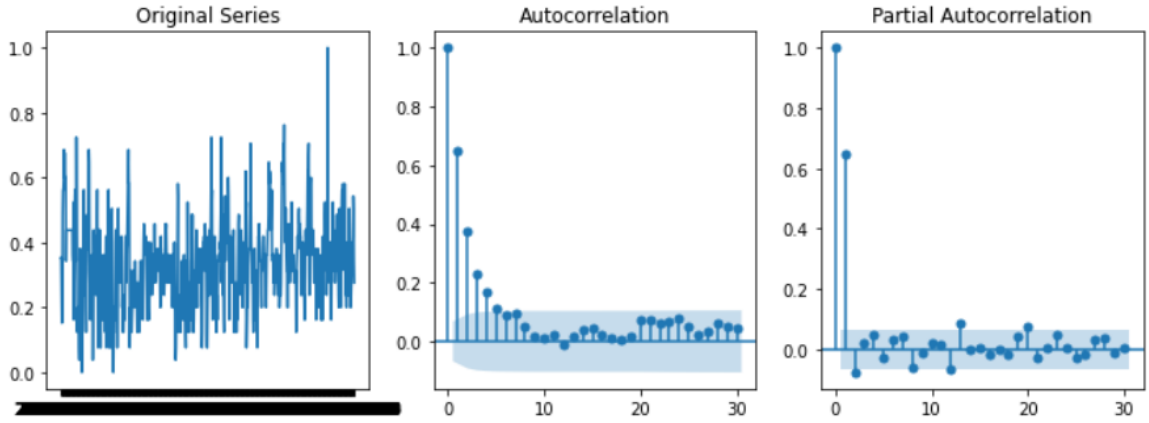
Σχήμα 5.4: Τεστ στασιμότητας των χρονοσειρών του Δελχί.

```
Check if Beijing_Data is stationary :
p-value for no2 : 0.000000
p-value for o3 : 0.000024
p-value for pm10 : 0.000000
p-value for pm25 : 0.000007
p-value for so2 : 0.000008
```

Σχήμα 5.5: Τεστ στασιμότητας των χρονοσειρών του Πεκίνου.

Οι χρονοσειρές που αφορούν τους ρύπους είναι στάσιμες όταν το p value που αναγράφεται στις εικόνες είναι μικρότερο απο 0.05. Παρατηρώντας τα παραπάνω αποτελέσματα, η μόνη χρονοσειρά που δεν είναι στάσιμη και απαιτείται να μετατραπεί, είναι η χρονοσειρά της ημερήσιας συγκέντρωσης O_3 στην Αθήνα, με p value ίσο με 0.18. Χρησιμοποιώντας τη συνάρτηση *ndiffs* φαίνεται οτι πρέπει να διαφοριστεί μια μόνο φορά ώστε να μετατραπεί σε στάσιμη.

Αφού έχει καθοριστεί η τάξη διαφόρισης, παρουσιάζονται τα διαγράμματα αυτοσυσχέτισης και μερικής αυτοσυσχέτισης απο τα οποία προκύπτουν οι παράμετροι MA και AR αντίστοιχα.



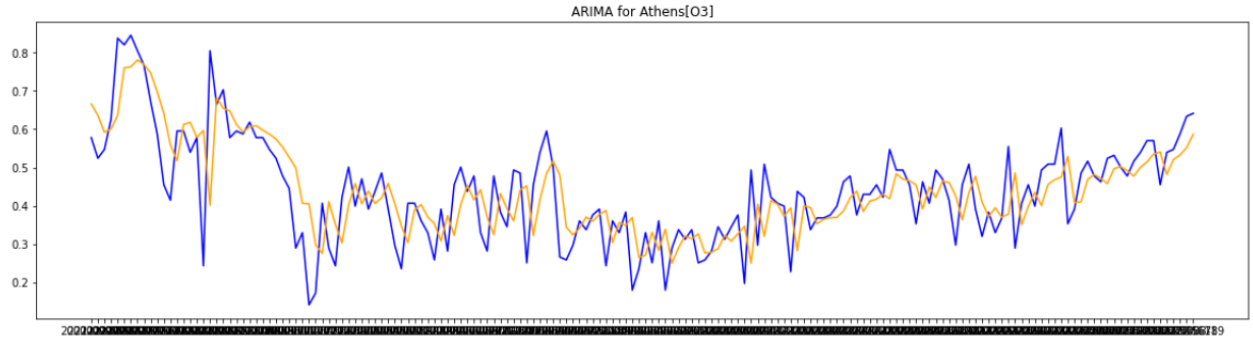
Σχήμα 5.6: Διαγράμματα αυτοσυχέτισης (Autocorrelation) και μερικής αυτοσυσχέτισης (Partial Autocorrelation).

Η αυτοσυσχέτιση και η μερική αυτοσυσχέτιση είναι μετρικές συσχέτισης μεταξύ των τρεχουσών και των προηγούμενων τιμών των χρονοσειρών και υποδεικνύουν ποιές τιμές προηγούμενων βημάτων είναι πιο χρήσιμες για την πρόβλεψη μελλοντικών τιμών. Με την αναπαράσταση των διαγραμμάτων, μέσω των συναρτήσεων *plot_acf* και *plot_pacf* οι οποίες λαμβάνουν ως παράμετρο την εκάστοτε χρονοσειρά, μπορεί να προσδιοριστεί η τάξη των παραμέτρων AR και MA σε ένα μοντέλο ARIMA. Πιο συγκεκριμένα, για τον υπολογισμό

των AR και MA παραμέτρων, προσμετρώνται τα πρώτα σημεία των διαγραμμάτων μερικής αυτοσυχέτισης και αυτοσυσχέτισης, αντίστοιχα, που βρίσκονται εκτός της μπλε σκιασμένης περιοχής, όπως φαίνεται στο σχήμα 5.6.

Αφού έχουν προσδιοριστεί οι παράμετροι μέσω της διαδικασίας που περιγράφηκε παραπάνω, το μοντέλο είναι έτοιμο να εξάγει τις προβλέψεις των νέων τιμών. Αρχικά πραγματοποιείται η εκπαίδευση του μοντέλου χρησιμοποιώντας τη συνάρτηση *ARIMA* της βιβλιοθήκης *statsmodels* και το σύνολο δεδομένων εκπαίδευσης που έχει οριστεί για κάθε πόλη, το οποίο ορίζεται ως παράμετρος στη συνάρτηση. Μέσα απο μία επαναληπτική διαδικασία, χρησιμοποιώντας τη συνάρτηση *forecast*, εξάγεται μια πρόβλεψη για ένα χρονικό βήμα κάθε φορά, η οποία στο επόμενο επαναληπτικό βήμα προστίθεται στο τέλος του συνόλου εκπαίδευσης, προκειμένου αυτό να επεκταθεί. Ακολουθεί εκ νέου εκπαίδευση του μοντέλου με το εμπλουτισμένο σύνολο. Η διαδικασία επαναλαμβάνεται τόσες φορές όσες και οι μελλοντικές τιμές που χρειάζεται να προβλεφθούν. Στο σχήμα 5.7 φαίνονται οι ημερήσιες προβλέψεις του manual ARIMA μοντέλου για τη συγκέντρωση O_3 στην Αθήνα, με παραμέτρους AR, διαφόρισης και MA ίσες με 2, 1, 2 αντίστοιχα.

```
ARIMA MSE for Athens[O3]:0.0095586081555493
ARIMA RMSE for Athens[O3]:0.09776813466334161
ARIMA MAE for Athens[O3]:0.07366203768836073
```

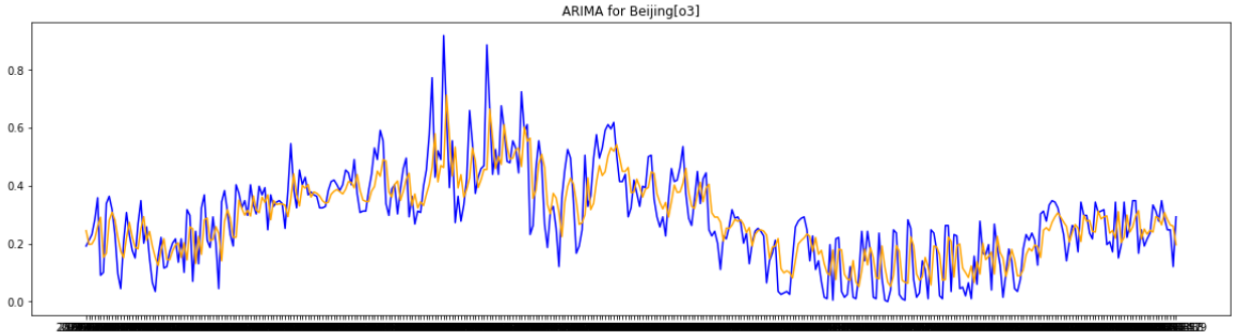


Σχήμα 5.7: Πρόβλεψη (πορτοκαλί γραμμή) manual ARIMA της ημερήσιας συγκέντρωσης O_3 στην Αθήνα.

Ωστόσο, για το Παρίσι, το Δελχί και το Πεκίνο ακολουθήθηκε μια διαφορετική διαδικασία υπολογισμού των παραμέτρων, καθώς λόγω του αρκετά μεγαλύτερου όγκου δεδομένων οι παράμετροι με την παραπάνω διαδικασία κατέληγαν να είναι ιδιαίτερα μεγάλοι αριθμοί, γεγονός που καθιστούσε το μοντέλο μη ποιοτικό και όχι τόσο ακριβές. Για να λυθεί το πρόβλημα αυτό, για τις πόλεις με πολλά δεδομένα, οι παράμετροι υπολογίστηκαν πειραματικά (Grid Search). Δοκιμάστηκαν επαναληπτικά, με τη διαδικασία που περιγράφηκε, μοντέλα

με διαφορετικούς συνδυασμούς των παραμέτρων AR, MA και τάξης διαφόρισης, με εύρος τιμών των παραμέτρων (0,10), (0,10) και (0,2) αντίστοιχα. Κατά τη διάρκεια εφαρμογής του εκάστοτε μοντέλου υπολογίζεται ο δείκτης AIC για καθένα συνδυασμό των παραμέτρων. Αφού υπολογιστεί ο δείκτης για όλους τους πιθανούς συνδυασμούς, επιλέγεται ως βέλτιστος ο συνδυασμός παραμέτρων που σημείωσε το μικρότερο δείκτη AIC. Έπειτα, εφαρμόζεται το μοντέλο ARIMA με τις βέλτιστες παραμέτρους ώστε να διεξάγει προβλέψεις. Η διαδικασία εκπαίδευσης του μοντέλου και της πρόβλεψης νέων τιμών, αφού βρεθούν οι βέλτιστες παράμετροι, είναι η ίδια που επεξηγήθηκε παραπάνω.

```
best AIC is:  -2956.1262147057364
ARIMA parameters:
 p:  5
 d:  0
 q:  5
ARIMA MSE for Beijing o3 0.009742265687334772
ARIMA RMSE for Beijing o3 0.09870291630612933
ARIMA MAE for Beijing o3 0.0764524667462647
```



Σχήμα 5.8: Grid Search και πρόβλεψη (πορτοκαλί γραμμή) manual ARIMA της ημερήσιας συγκέντρωσης O_3 στο Πεκίνο.

Συγκρίνοντας τα αποτελέσματα στους πίνακες 5.1-5.4 των ημερήσιων προβλέψεων των μοντέλων auto-ARIMA και manual ARIMA, παρατηρείται πως το manual μοντέλο πέτυχε συντριπτικά καλύτερα αποτελέσματα, με εξάιρεση τον δείκτη συγκέντρωσης SO_2 στο Παρίσι, όπου το αυτόματο μοντέλο ήταν πιο αποτελεσματικό.

5.2.3 Σημασία του όγκου δεδομένων για την ακρίβεια των προβλέψεων

Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, τα δεδομένα που αφορούν την Αθήνα περιορίζονται σε πολύ μικρότερο χρονικό διάστημα συγκριτικά με τις υπόλοιπες πόλεις που μελετώνται. Για τις πόλεις αυτές, οι ημερήσιες προβλέψεις που έγιναν επαναλήφθηκαν, αυτή τη φορά με πλήθος ημερήσιων δεδομένων ίσο με αυτό της Αθήνας, δηλαδή δεδομένα που αντιστοιχούσαν στις τελευταίες 839 ημέρες αντί για τις συνολικά 1896 διαθέσιμες.

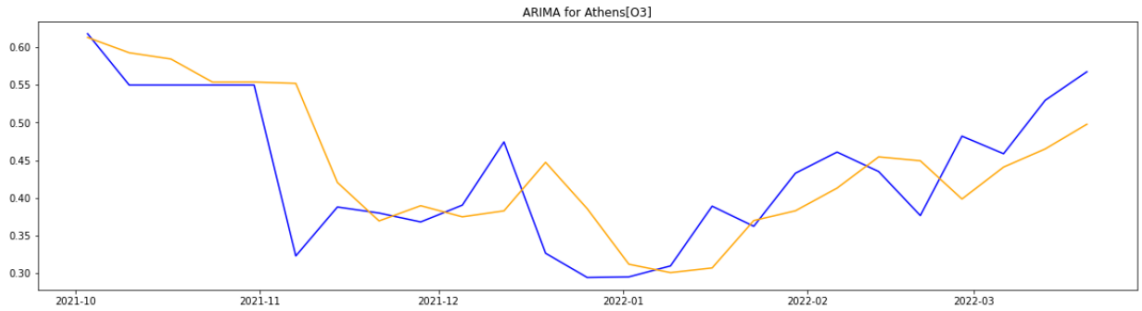
Συγκρίνοντας τα αποτελέσματα του συγκεκριμένου πειράματος για κάθε πόλη, τα οποία βρίσκονται καταγεγραμμένα στους πίνακες 5.1-5.4, με τις προβλέψεις όπου συμπεριλήφθη-

καν όλα τα δεδομένα, παρατηρούμε πως τα σφάλματα όλων των κατηγοριών είναι μεγαλύτερα για κάθε ατμοσφαιρικό ρύπο και για κάθε πόλη. Συνεπώς, όταν χρησιμοποιούνται περισσότερα δεδομένα κατα την εκπαίδευση του μοντέλου οι προβλέψεις είναι πιο έγκυρες και πιο κοντά στις πραγματικές τιμές.

5.2.4 Εβδομαδιαίες και μηνιαίες προβλέψεις

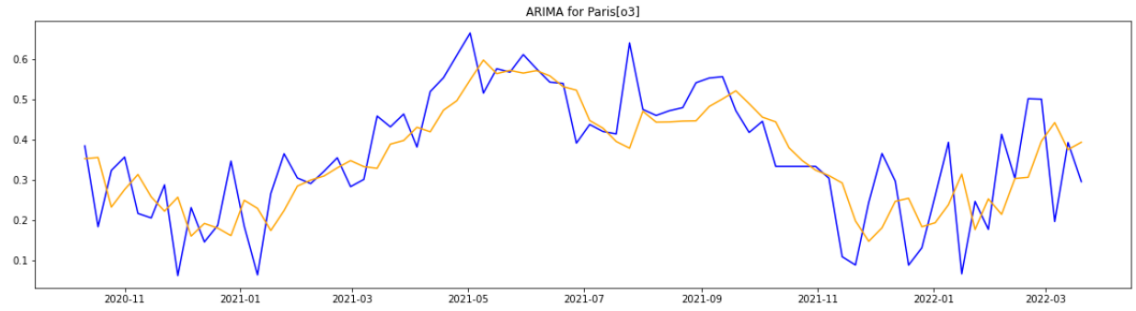
Λόγω της εποχικότητας που εμφάνισαν τα δεδομένα σε εβδομαδιαία και μηνιαία συχνότητα, έγιναν προβλέψεις που αφορούσαν αυτά τα διαστήματα για να ελεγχθεί η σημασία της εποχικότητας στην εγκυρότητα των προβλέψεων. Συγκεκριμένα, με χρήση της συνάρτησης *resample* της *Python*, τα ημερήσια δεδομένα ομαδοποιούνται κατά εβδομάδα και μήνα υπολογίζοντας το μέσο όρο των δεδομένων για κάθε διάστημα. Έπειτα, ακολουθείται η διαδικασία που περιγράφηκε παραπάνω για την εύρεση των κατάλληλων παραμέτρων για κάθε χρονοσειρά. Οι εβδομαδιαίες και μηνιαίες χρονοσειρές που δημιουργούνται, τροφοδοτούνται στο μοντέλο για να γίνουν προβλέψεις, όπως φαίνεται στα σχήματα 5.10-5.15. Το μοντέλο, λόγω μικρού όγκου μηνιαίων δεδομένων για την πόλη της Αθήνας δεν μπορούσε να λειτουργήσει, επομένως οι μηνιαίες προβλέψεις (σχήμα 5.15) έχουν πραγματοποιηθεί μόνο για τις υπόλοιπες πόλεις.

```
ARIMA MSE for Athens[O3]:0.004906441564469912
ARIMA RMSE for Athens[O3]:0.07004599606308637
ARIMA MAE for Athens[O3]:0.04975594278233996
```



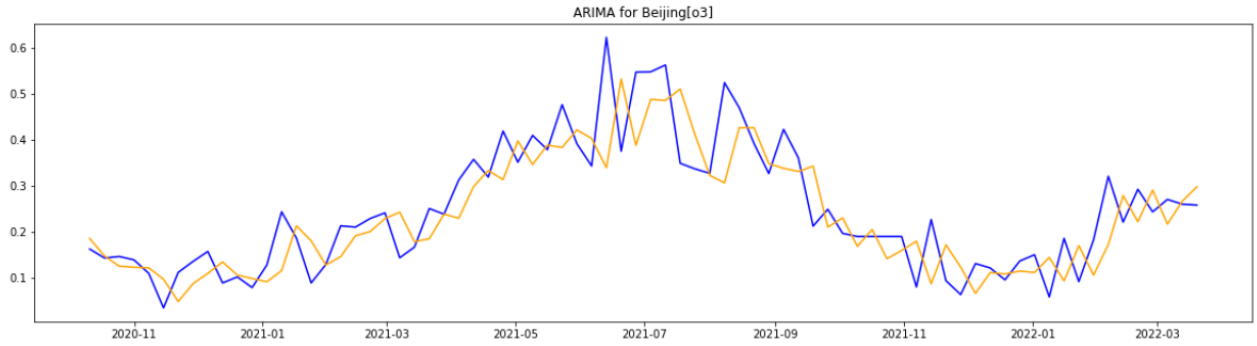
Σχήμα 5.9: Εβδομαδιαίες προβλέψεις (πορτοκαλί γραμμή) της συγκέντρωσης O_3 στην Αθήνα.

```
best AIC is:  -750.6232986690353
ARIMA parameters:
 p:  5
 d:  0
 q:  5
ARIMA MSE for Paris o3 0.010429700524903076
ARIMA RMSE for Paris o3 0.10212590525867116
ARIMA MAE for Paris o3 0.0783273479973195
```



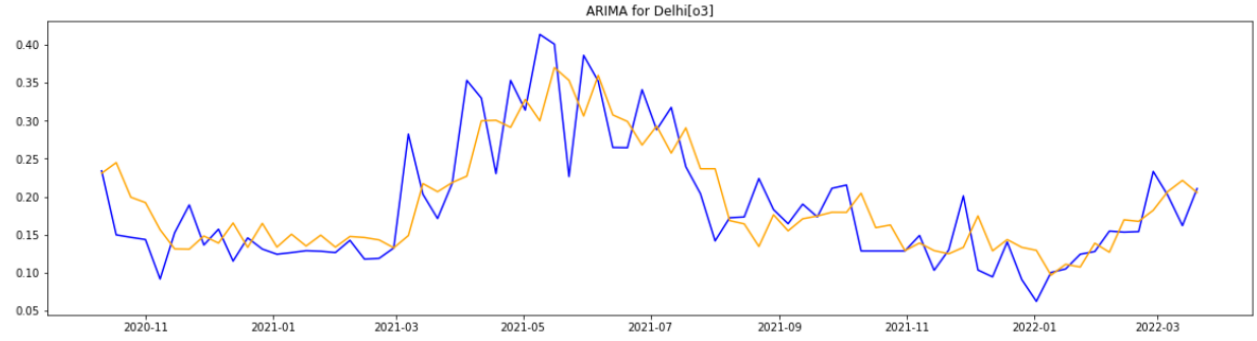
Σχήμα 5.10: Εβδομαδιαίες προβλέψεις (πορτοκαλί γραμμή) της συγκέντρωσης O_3 στο Παρίσι.

```
best AIC is:  -835.5483014690551
ARIMA parameters:
 p:  6
 d:  0
 q:  5
ARIMA MSE for Beijing o3 0.006161953833904571
ARIMA RMSE for Beijing o3 0.07849811356908248
ARIMA MAE for Beijing o3 0.05952523658428046
```



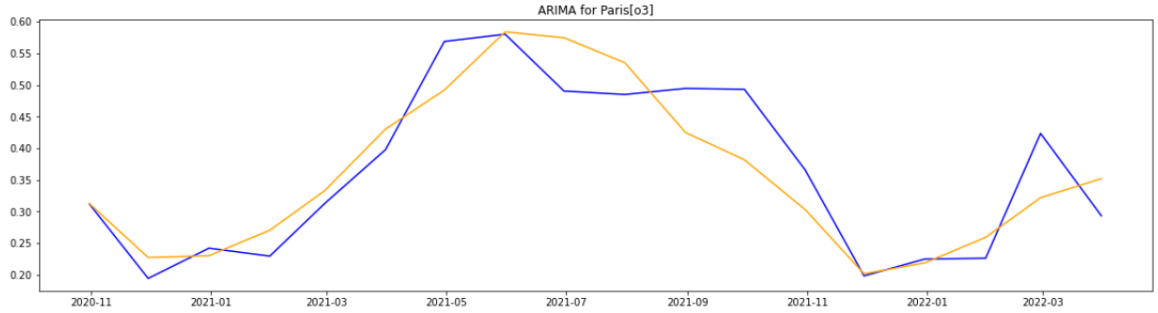
Σχήμα 5.11: Εβδομαδιαίες προβλέψεις (πορτοκαλί γραμμή) της συγκέντρωσης O_3 στο Πεκίνο.

```
best AIC is:  -1113.8421795767754
ARIMA parameters:
 p:  5
 d:  0
 q:  7
ARIMA MSE for Delhi o3 0.002371492078909634
ARIMA RMSE for Delhi o3 0.04869796791355502
ARIMA MAE for Delhi o3 0.03591339949667314
```



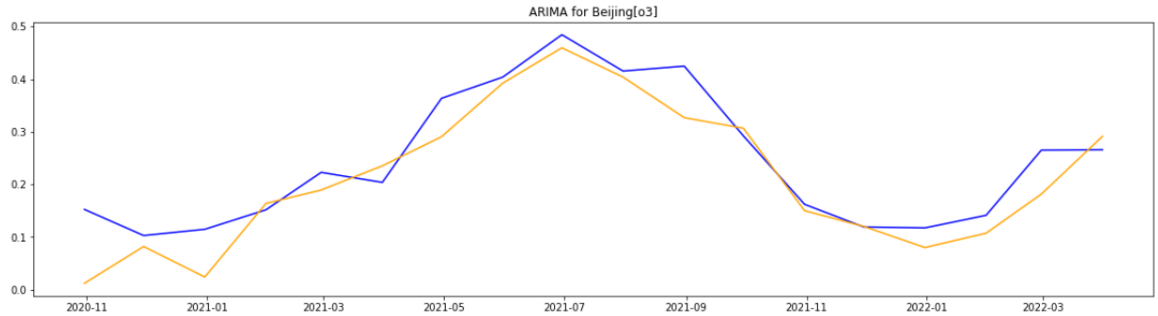
Σχήμα 5.12: Εβδομαδιαίες προβλέψεις (πορτοκαλί γραμμή) της συγκέντρωσης O_3 στο Δελχί.

```
best AIC is:  -136.77126168250862
ARIMA parameters:
 p:  5
 d: 0
 q:  5
ARIMA MSE for Paris o3 0.0031134062272632082
ARIMA RMSE for Paris o3 0.055797905222895315
ARIMA MAE for Paris o3 0.04449068803852748
```



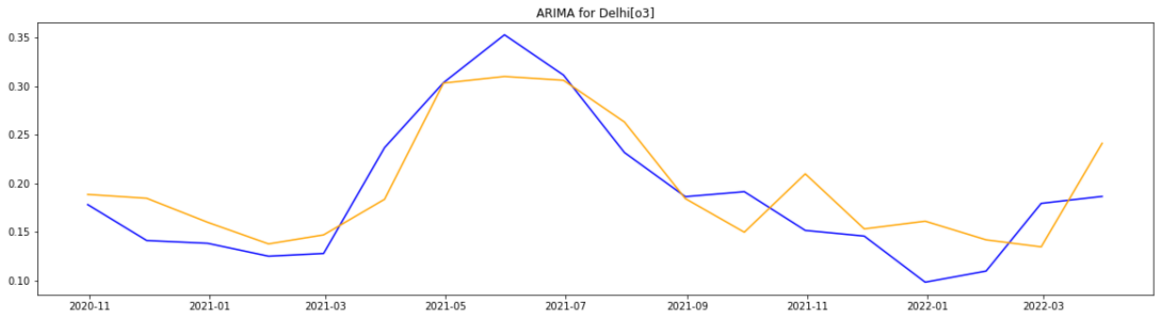
Σχήμα 5.13: Μηνιαίες προβλέψεις (πορτοκαλί γραμμή) της συγκέντρωσης O_3 στο Παρίσι.

```
best AIC is:  -188.36107172529935
ARIMA parameters:
 p:  10
 d: 0
 q:  8
ARIMA MSE for Beijing o3 0.003166843742153108
ARIMA RMSE for Beijing o3 0.056274716722104504
ARIMA MAE for Beijing o3 0.0419927098448906
```



Σχήμα 5.14: Μηνιαίες προβλέψεις (πορτοκαλί γραμμή) της συγκέντρωσης O_3 στο Πεκίνο.

```
best AIC is:  -286.10527279291716
ARIMA parameters:
 p:  7
 d: 0
 q:  7
ARIMA MSE for Delhi o3 0.001322531034873597
ARIMA RMSE for Delhi o3 0.036366619788943774
ARIMA MAE for Delhi o3 0.030230026268832606
```



Σχήμα 5.15: Μηνιαίες προβλέψεις (πορτοκαλί γραμμή) της συγκέντρωσης O_3 στο Δελχί.

5.2.5 Αποτελέσματα ARIMA

Οι εβδομαδιαίες και μηνιαίες προβλέψεις του manual ARIMA μοντέλου ήταν αρκετά ικανοποιητικές. Όπως διακρίνεται απο τα σχήματα 5.10-5.15 οι προβλέψεις κατάφεραν να πλησιάσουν σημαντικά τις πραγματικές τιμές των συγκεντρώσεων.

Στους πίνακες που ακολουθούν καταγράφονται συνολικά τα αποτελέσματα όλων των πειραμάτων που διεξήχθησαν. Τα μοντέλα ίδιων συχνοτήτων βρίσκονται στο ίδιο χρωματικό πλαίσιο. Οι χρωματιστές τιμές αντιστοιχούν στο μικρότερο RMSE σφάλμα όλων των μοντέλων της ίδια συχνότητας ανά ρύπο. Οι υπογραμμισμένες τιμές αποτελούν τη χρονοσειρά που σημείωσε το μικρότερο RMSE σφάλμα για την εκάστοτε μέθοδο.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
Auto-ARIMA (daily)	MSE	0.0259	0.0126	0.0002	0.0038	0.0129
	RMSE	0.1610	0.1124	0.0158	0.0618	0.1137
	MAE	0.1258	0.0989	0.0136	0.0306	0.0946
Manual ARIMA (daily)	MSE	0.0097	0.0072	0.00013	0.0034	0.008
	RMSE	0.0987	0.0849	0.0116	0.0583	0.0890
	MAE	0.0764	0.0619	0.0084	0.0222	0.070
Manual ARIMA less data(daily)	MSE	0.0177	0.0114	0.00014	0.0049	0.0105
	RMSE	0.1381	0.1071	0.0133	0.0721	0.1026
	MAE	0.0809	0.0854	0.0095	0.0385	0.0803
Manual ARIMA (weekly)	MSE	0.0061	0.0039	0.00006	0.00062	0.00475
	RMSE	0.0784	0.0626	0.008	0.0249	0.0689
	MAE	0.059	0.0478	0.0061	0.0155	0.0552
Manual ARIMA (monthly)	MSE	0.0031	0.0016	0.00011	0.0004	0.0015
	RMSE	0.0562	0.0407	0.0107	0.0201	0.0392
	MAE	0.0419	0.0329	0.0089	0.0156	0.0326

Πίνακας 5.1: Αποτελέσματα μοντέλων ARIMA για το Πεκίνο.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.0892	0.0329	0.0340	0.0193	0.0157
Auto-ARIMA (daily)	RMSE	0.2987	0.1814	0.1844	0.1391	0.1254
	MAE	0.2739	0.1448	0.1084	0.1009	0.0950
	MSE	0.0095	0.02457	0.01571	0.01390	0.01154
Manual ARIMA(daily)	RMSE	0.0977	0.1567	0.1253	0.1179	0.1074
	MAE	0.0736	0.1228	0.0823	0.083	0.0782
	MSE	0.0049	0.0142	0.0172	0.0054	0.0045
Manual ARIMA (weekly)	RMSE	0.070	0.1193	0.1311	0.0738	0.06762
	MAE	0.0497	0.0914	0.0709	0.0561	0.05457

Πίνακας 5.2: Αποτελέσματα μοντέλων ARIMA για την Αθήνα.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.0503	0.02094	0.00267	0.0211	0.0211
Auto-ARIMA (daily)	RMSE	0.2243	0.1447	0.0516	0.1455	0.1454
	MAE	0.1921	0.1154	0.039	0.1010	0.1202
	MSE	0.0152	0.0127	0.1044	0.0115	0.0096
Manual ARIMA (daily)	RMSE	0.1232	0.1128	0.3232	0.1076	0.098
	MAE	0.095	0.087	0.0473	0.08013	0.07423
	MSE	0.0199	0.0128	0.321	0.0126	0.0128
Manual ARIMA less data(daily)	RMSE	0.142	0.1135	0.572	0.1124	0.1134
	MAE	0.1003	0.0916	0.337	0.0856	0.0848
	MSE	0.0104	0.0073	0.00062	0.00956	0.00979
Manual ARIMA (weekly)	RMSE	0.1021	0.0854	0.02497	0.09781	0.0989
	MAE	0.0783	0.0677	0.0189	0.07243	0.07568
	MSE	0.0031	0.00932	0.00043	0.00764	0.0078
Manual ARIMA (monthly)	RMSE	0.0557	0.0965	0.0209	0.0874	0.08834
	MAE	0.0444	0.0804	0.0169	0.0757	0.07928

Πίνακας 5.3: Αποτελέσματα μοντέλων ARIMA για το Παρίσι.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.0157	0.0228	0.0114	0.0071	0.0267
Auto-ARIMA (daily)	RMSE	0.1253	0.1510	0.1068	0.0843	0.16365
	MAE	0.090	0.1312	0.0937	0.0697	0.14251
	MSE	0.0034	0.002	0.0013	0.0016	0.0033
Manual ARIMA (daily)	RMSE	0.0586	0.0044	0.0373	0.0404	0.0577
	MAE	0.0426	0.0332	0.0275	0.0261	0.039
	MSE	0.0042	0.0025	0.0085	0.0028	0.0053
Manual ARIMA less data(daily)	RMSE	0.0672	0.0504	0.092	0.0532	0.0733
	MAE	0.0537	0.0386	0.220	0.0343	0.0492
	MSE	0.00237	0.00367	0.0014	0.00297	0.0058
Manual ARIMA (weekly)	RMSE	0.0486	0.0606	0.03795	0.0545	0.07625
	MAE	0.0359	0.0456	0.0301	0.0389	0.0563
	MSE	0.0013	0.0025	0.00365	0.00174	0.0037
Manual ARIMA (monthly)	RMSE	0.0363	0.050	0.0604	0.0418	0.0615
	MAE	0.0302	0.0431	0.0493	0.0349	0.0491

Πίνακας 5.4: Αποτελέσματα μοντέλων ARIMA για το Δελχί.

Συγκρίνοντας τα αποτελέσματα των ημερήσιων προβλέψεων με αυτά των εβδομαδιαίων και μηνιαίων, διαπιστώνουμε πως η εβδομαδιαία και η μηνιαία συχνότητα προβλέψεων πετυχαίνουν πολύ καλύτερη επίδοση για κάθε πόλη. Αυτό οφείλεται στην εποχικότητα των δεδομένων και στη δυνατότητα του μοντέλου να την αντιλαμβάνεται και να τη χρησιμοποιεί.

Εκτός απο τη σύγκριση των επιδόσεων των εβδομαδιαίων και μηνιαίων προβλέψεων με τις αυτές των ημερήσιων, συγκρίνονται και οι επιδόσεις των ίδιων μοντέλων μεταξύ των πόλεων. Προκύπτει πως οι επιδόσεις των μοντέλων που εφαρμόστηκαν στα δεδομένα του Παρισιού, του Πεκίνου και του Δελχί είναι κατά πλειοψηφία καλύτερες απο τις επιδόσεις των μοντέλων τα οποία εφαρμόστηκαν στα δεδομένα της Αθήνας. Η παρατήρηση αυτή επιβεβαιώνεται απο τα σχήματα 5.10-5.12, στα οποία είναι φανερό οτι η γραμμή των προβλέψεων της συγκέντρωσης O_3 στις υπόλοιπες πόλεις πλησιάζει πιο κοντά στα πραγματικά δεδομένα. Η μειωμένη επίδοση των μοντέλων στα δεδομένα της Αθήνας οφείλεται, πιθανότατα, στον

αρκετά μικρότερο διαθέσιμο όγκο δεδομενων.

Σε κάθε πόλη, εκτός απο το Δελχί, υπάρχει ένας ρύπος του οποίου η πρόβλεψη επιφέρει το μικρότερο σφάλμα στην πλειοψηφία των μοντέλων που δοκιμάστηκαν. Για το Πεκίνο, η πρόβλεψη της συγκέντρωσης SO_2 είχε τα μικρότερα σφάλματα και στα πέντε μοντέλα, ενώ η συγκέντρωση του ίδιου ρύπου στο Παρίσι πέτυχε τα μικρότερα σφάλματα σε τρία απο τα πέντε μοντέλα, στις ημερήσιες προβλέψεις με το μοντέλο auto-ARIMA και στις εβδομαδιαίες και μηνιαίες με το manual ARIMA. Για την Αθήνα, η πρόβλεψη της συγκέντρωσης $PM_{2.5}$ είχε τα μικρότερα σφάλματα σε δυο απο τα τρία μοντέλα, στις ημερήσιες και εβδομαδιαίες προβλέψεις με το μοντέλο manual ARIMA.

5.3 Facebook Prophet

Η μέθοδος Prophet απαιτεί τα δεδομένα εισόδου να έχουν μια συγκεκριμένη μορφή προκειμένου να γίνουν οι προβλέψεις. Το σύνολο δεδομένων πρέπει να αποτελείται απο δυο στήλες, μια με τις τιμές της χρονοσειράς, η οποία δεν απαιτείται να έχει συγκεκριμένο όνομα, και μια η οποία αναγράφει την ημερομηνία που πραγματοποιήθηκε η μέτρηση της τιμής, η οποία θα πρέπει να ονομάζεται 'ds'. Όπως και στα μοντέλα ARIMA, οι εισαχθείσες χρονοσειρές θα πρέπει να είναι στάσιμες ώστε να λειτουργήσει αποδοτικά το μοντέλο, ωστόσο δεν υπάρχουν παράμετροι που πρέπει να προσδιοριστούν για το μοντέλο. Ακολουθείται και πάλι η διαδικασία του ADF τεστ, η οποία επεξηγήθηκε στην προηγούμενη ενότητα.

Η εκπαίδευση του μοντέλου γίνεται δημιουργώντας ένα μοντέλο μέσω της συνάρτησης *Prophet* της βιβλιοθήκης *fbprophet* χωρίς παραμέτρους και έπειτα καλείται η συνάρτηση *fit* με το σύνολο εκπαίδευσης ως παράμετρο. Στη συνέχεια, καλείται η συνάρτηση *forecast* με παράμετρο τη στήλη 'ds' του συνόλου δεδομένων ελέγχου, δηλαδή τη στήλη στην οποία βρίσκονται οι ημερομηνίες για τις οποίες θα προβλεφθούν οι συγκεντρώσεις των ρύπων.

Η έξοδος του μοντέλου αποτελεί ένα σύνολο δεδομένων που απαρτίζεται απο τέσσερις στήλες, τη στήλη 'ds',δηλαδή τις ημερομηνίες στις οποίες αντιστοιχούν οι προβλέψεις, τη στήλη 'yhat', δηλαδή τις τιμές των προβλέψεων και τέλος, τις στήλες 'yhat_lower' και 'yhat_upper', το κάτω και το άνω όριο, αντίστοιχα, του διαστήματος εμπιστοσύνης για την πρόβλεψη. Το διάστημα εμπιστοσύνης ορίζεται ως ένα διάστημα τιμών μέσα στο οποίο υπάρχει μεγάλο ποσοστό πιθανότητας να βρίσκεται η πραγματική τιμή της πρόβλεψης. Η πιθανότητα αυτή μπορεί να καθοριστεί χειροκίνητα για το μοντέλο Prophet, στις παρακάτω δοκιμές

ορίστηκε στο 80%.

5.3.1 Σημασία του όγκου δεδομένων για την ακρίβεια των προβλέψεων

Όπως και στο μοντέλο ARIMA, ελέγχθηκε εαν ο μεγαλύτερος όγκος δεδομένων εκπαίδευσης θα επιφέρει καλύτερα αποτελέσματα. Όμοια με την προηγούμενη ενότητα, το μοντέλο Prophet δοκιμάστηκε για την πρόβλεψη των ατμοσφαιρικών ρύπων με με πλήθος ημερήσιων δεδομένων ίσο με αυτό της Αθήνας. Τα δεδομένα αντιστοιχούσαν στις τελευταίες 839 αντί για τις δυνολικά 1896 διαθέσιμες ημέρες.

Συγκρίνοντας τους πίνακες 5.5-5.8 με τα αποτελέσματα των προβλέψεων με χρήση όλων των δεδομένων και ενός μέρους αυτών, για κάθε πόλη, είναι ξεκάθαρο πως οι προβλέψεις που βασίζονται στην εκπαίδευση με μικρότερο όγκο δεδομένων είναι λιγότερο ακριβείς με σημαντική διαφορά.

5.3.2 Αποτελέσματα Facebook Prophet

Όμοια με το μοντέλο ARIMA, έγιναν προβλέψεις που αφορούσαν διαστήματα εβδομάδων και μηνών, τα οποία εμφάνιζαν και ισχυρή εποχικότητα. Η διαδικασία για να δημιουργηθούν τα εβδομαδιαία και μηνιαία δεδομένα είναι η ίδια που περιγράφηκε στην προηγούμενη ενότητα. Τα αποτελέσματα όλων των πειραμάτων που διεξήχθησαν βρίσκονται στους πίνακες που ακολουθούν. Τα μοντέλα ίδιων συχνοτήτων βρίσκονται στο ίδιο χρωματικό πλαίσιο. Οι χρωματιστές τιμές αντιστοιχούν στο μικρότερο RMSE σφάλμα όλων των μοντέλων της ίδια συχνότητας ανά ρύπο. Οι υπογραμμισμένες τιμές αποτελούν τη χρονοσειρά που σημείωσε το μικρότερο RMSE σφάλμα για την εκάστοτε μέθοδο.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.0118	0.0090	0.0010	0.0030	0.0118
Prophet (daily)	RMSE	0.1086	0.0953	0.0320	0.0555	0.1089
	MAE	0.0862	0.0706	0.0271	0.0218	0.0881
	MSE	0.2280	0.0430	0.0001	0.0036	0.0210
Prophet less data (daily)	RMSE	0.4775	0.2075	0.0125	0.0603	0.1449
	MAE	0.4602	0.1701	0.0082	0.0522	0.1062
	MSE	0.0124	0.0069	0.00099	0.00096	0.0072
Prophet (weekly)	RMSE	0.1115	0.0832	0.0315	0.0310	0.0852
	MAE	0.0949	0.0695	0.0248	0.0202	0.0625
	MSE	0.0056	0.0018	0.0018	0.0003	0.0020
Prophet (monthly)	RMSE	0.0753	0.0430	0.0430	0.0178	0.0457
	MAE	0.0607	0.0319	0.0351	0.0097	0.0343

Πίνακας 5.5: Αποτελέσματα προβλέψεων FB Prophet για το Πεκίνο.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.1065	0.0552	0.0439	0.0296	0.0159
Prophet (daily)	RMSE	0.0113	0.2350	0.2096	0.1722	0.1261
	MAE	0.0765	0.1898	0.1716	0.1443	0.0963
	MSE	0.0673	0.0210	0.0218	0.0098	0.0039
Prophet (weekly)	RMSE	0.2594	0.1451	0.1478	0.0990	0.0631
	MAE	0.2400	0.1234	0.1051	0.0854	0.0515
	MSE	0.0631	0.0148	0.0155	0.0076	0.0020
Prophet (monthly)	RMSE	0.2513	0.1217	0.1244	0.0874	0.0450
	MAE	0.2361	0.1103	0.1129	0.0823	0.0382

Πίνακας 5.6: Αποτελέσματα προβλέψεων FB Prophet για την Αθήνα.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.0197	0.0245	0.0043	0.0194	0.0184
Prophet (daily)	RMSE	0.1404	0.1566	0.0661	0.1393	0.1356
	MAE	0.1105	0.1148	0.0558	0.1022	0.1049
	MSE	0.1902	0.0245	0.0032	0.0209	0.0364
Prophet less data (daily)	RMSE	0.4361	0.0607	0.0573	0.1445	0.1908
	MAE	0.3968	0.2123	0.0250	0.1060	0.1383
	MSE	0.0210	0.0076	0.0005	0.0083	0.0095
Prophet (weekly)	RMSE	0.1451	0.0871	0.0233	0.0912	0.0979
	MAE	0.1136	0.0658	0.0191	0.0695	0.0741
	MSE	0.0082	0.0053	0.0003	0.0050	0.0054
Prophet (monthly)	RMSE	0.0908	0.0730	0.0198	0.0711	0.0739
	MAE	0.0767	0.0598	0.0160	0.0592	0.0646

Πίνακας 5.7: Αποτελέσματα προβλέψεων FB Prophet για το Παρίσι.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.0057	0.0063	0.0049	0.0030	0.0065
Prophet (daily)	RMSE	0.0760	0.0797	0.0703	0.0555	0.0809
	MAE	0.0611	0.0650	0.0562	0.0431	0.0617
	MSE	0.0094	0.1009	0.0871	0.0325	0.1476
Prophet less data (daily)	RMSE	0.0973	0.3177	0.2951	0.1803	0.3841
	MAE	0.0855	0.2954	0.2702	0.1573	0.3525
	MSE	0.0041	0.0271	0.0044	0.0178	0.0434
Prophet (weekly)	RMSE	0.0643	0.1646	0.0670	0.1334	0.2083
	MAE	0.0523	0.1440	0.0554	0.1169	0.1829
	MSE	0.0010	0.0062	0.0055	0.0024	0.0060
Prophet (monthly)	RMSE	0.0320	0.0793	0.0747	0.0497	0.0780
	MAE	0.0264	0.0692	0.0598	0.0404	0.0603

Πίνακας 5.8: Αποτελέσματα προβλέψεων FB Prophet για το Δελχί.

Όπως παρατηρήσαμε και στα προηγούμενα πειράματα, με τη μέθοδο ARIMA, οι εβδομαδιαίες και μηνιαίες προβλέψεις είχαν καλύτερες επιδόσεις, γεγονός που οφείλεται στην ικανότητα του μοντέλου να αντιλαμβάνεται και να προβλέπει τις νέες τιμές βάσει της υπάρχουσας εποχικότητας που εντοπίζεται στα δεδομένα. Συγκεκριμένα, οι μηνιαίες προβλέψεις είχαν σε γενικές γραμμές καλύτερες επιδόσεις απο τις εβδομαδιαίες και ημερήσιες. Αυτό μπορεί να επεξηγηθεί βλέποντας το σχήμα 4.7, στο οποίο παρατηρούμε ότι το μοτίβο εποχικότητας στη γραφική παράσταση των μηνιαίων δεδομένων είναι αρκετά πιο ξεκάθαρο από ότι στα εβδομαδιαία.Η παρατήρηση αυτή ισχύει για όλες τις μελετηθείσες πόλεις.

Επιπλέον, τα μοντέλα που εφαρμόστηκαν στα δεδομένα της Αθήνας είχαν χειρότερες επιδόσεις συγκριτικά με τα αντίστοιχα μοντέλα τα οποία εφαρμόστηκαν στις υπόλοιπες πόλεις με μεγαλύτερο πλήθος διαθέσιμων δεδομένων.

Σε κάθε πόλη υπάρχει ένας ρύπος του οποίου η πρόβλεψη επιφέρει το μικρότερο σφάλμα στην πλειοψηφία των μοντέλων που δοκιμάστηκαν. Για το Πεκίνο η πρόβλεψη της συγκέντρωσης SO_2 έιχε τα μικρότερα σφάλματα στις ημερήσιες προβλέψεις ενώ η συγκέντρωση PM_{10} στις εβδομαδιαίες και μηνιαίες. Για την Αθήνα η πρόβλεψη της συγκέντρωσης $PM_{2.5}$ είχε τα μικρότερα σφάλματα σε δυο απο τα τρία μοντέλα, το εβδομαδιαίο και μηνιαίο. Η πρόβλεψη της συγκέντρωσης SO_2 στο Παρίσι πέτυχε τα μικρότερα σφάλματα και στα τέσσερα μοντέλα. Τέλος, η συγκέντρωση O_3 στο Δελχί απέφερε το μικρότερο σφάλμα σε τρία απο τα τέσσερα συνολικά μοντέλα, στην ημερήσια πρόβλεψη με τη χρήση μικρότερου πλήθους δεδομένων εκπαίδευσης και στις εβδομαδιαίες και μηνιαίες προβλέψεις.

5.4 Νευρωνικά δίκτυα

Παρά τα ικανοποιητικά αποτελέσματα που πέτυχαν οι μέθοδοι πρόβλεψης χρονοσειρών, έγινε μια προσπάθεια βελτιστοποίησης της ακρίβειας μέσω των νευρωνικών δικτύων. Τα νευρωνικά δίκτυα έχουν αποδεδειγμένα σημειώσει μεγάλη επιτυχία στην πρόβλεψη χρονοσειρών, δεδομένης της ικανότητάς τους να μαθαίνουν αυτόματα τις χρονικές εξαρτήσεις που υπάρχουν. Πιο συγκεκριμένα, επιλέχθηκαν δυο είδη δικτύων για την πρόβλεψη, τα δίκτυα μακράς-βραχύχρονης μνήμης (LSTM) και τα συνελικτικά δίκτυα (CNN).

Τα δίκτυα δοκιμάστηκαν με πολλούς συνδυασμούς επιπέδων και παραμέτρων και τελικά επιλέχθηκαν οι συνδυασμοί οι οποίοι απέφεραν το μικρότερο MSE σφάλμα.

Τα μοντέλα των νευρωνικών δικτύων έχουν χρησιμοποιηθεί για να προβλέπουν τις μελλοντικές τιμές κάθε χρονοσειράς με δυο τρόπους, είτε χρησιμοποιώντας μόνο τις τιμές της σε παρελθοντικές χρονικές στιγμές (μονομεταβλητά), είτε χρησιμοποιώντας τις προηγούμενες τιμές σε συνδυασμό με στήλες που αφορούν καιρικές συνθήκες (πολυμεταβλητά). Τα αντίστοιχα μονομεταβλητά και πολυμεταβλητά μοντέλα που δημιουργήθηκαν είχαν την ίδια δομή επιπέδων, προκειμένου να είναι συγκρίσιμα μεταξύ τους. Ωστόσο, στα μονομεταβλητά μοντέλα, χρησιμοποιήθηκε η συνάρτηση ενεργοποίησης ReLu, η οποία λόγω χαρακτηριστικών όπως η σταθερή τιμή της κλίσης της, οδηγεί σε αρκετά γρήγορη εκπαίδευση του μοντέλου. Η επιλογή αυτή, οφείλεται στο γεγονός πως η συνάρτηση ReLu μηδενίζει τις αρνητικές εισόδους. Συνεπώς, στα πολυμεταβλητά μοντέλα, τα οποία χρησιμοποιούν και τις καιρικές συνθήκες όπου μερικές απο αυτές, όπως η θερμοκρασία, μπορούν να λάβουν και αρνητικές τιμές, θεωρήθηκε σκόπιμο να χρησιμοποιηθεί η γραμμική συνάρτηση ενεργοποίησης. Όταν πρόκειται για μονομεταβλητό μοντέλο, η δεύτερη διάσταση της εισόδου, είναι πάντα μονάδα, ενώ όταν πρόκειται για πολυμεταβλητό είναι η τιμή 6 (μια στήλη για τις τιμές του ρύπου και πέντε για τις καιρικές συνθήκες). Για τις εβδομαδιαίες και μηνιαίες προβλέψεις, η πρώτη διάσταση της εισόδου εξαρτάται απο τα διαστήματα στα οποία κάθε χώρα εμφανίζει μοτίβο εποχικότητας, όπως φαίνεται στον πίνακα 5.9. Αντίθετα, για τις ημερήσιες προβλέψεις, η πρώτη διάσταση της εισόδου είναι η τιμή 7 για όλες τις χρονοσειρες. Η τιμή επιλέχθηκε αυθαίρετα καθώς δεν υπήρχαν επαναλαμβανόμενα μοτίβα.

5.4.1 Προετοιμασία των δεδομένων

Η ανάπτυξη μονομεταβλητών και πολυμεταβλητών μοντέλων απαιτούσε διαφορετική προεπεξεργασία των δεδομένων για το καθένα. Πιο συγκεκριμένα, στα μονομεταβλητά μοντέλα απομονώθηκαν οι στήλες που αφορούν τους ρύπους του αέρα. Οι στήλες αυτές αποτελούν τις χρονοσειρές όπου θα γίνει η πρόβλεψη των μελλοντικών τιμών και εισάγονται επαναληπτικά η μία μετά την άλλη στα μοντέλα. Για τα πολυμεταβλητά μοντέλα, οι στήλες που χρησιμοποιούνται είναι, εκτός απο τους ρύπους, η θερμοκρασία, η υγρασία, η ταχύτητα και η κατεύθυνση του αέρα και το σημείο δρόσου για κάθε μέρα, εβδομάδα ή μήνα που γίνεται η δειγματοληψία. Για κάθε πολυμεταβλητό πείραμα δημιουργούνται πέντε ξεχωριστά σύνολα δεδομένων, τα οποία αποτελούνται απο έξι στήλες. Η πρώτη στήλη αφορά τον εκάστοτε ρύπο που θα προβλεφθεί και οι υπόλοιπες πέντε στήλες αφορούν τις καιρικές συνθήκες.

Τέλος, για να γίνουν προβλέψεις μέσω των νευρωνικών δικτύων δημιουργήθηκε συνάρτηση η οποία χωρίζει κάθε σύνολο δεδομένων σε πολλά δείγματα τα οποία αποτελούνται απο έναν συγκεκριμένο αριθμό χρονικών βημάτων. Πιο συγκεκριμένα, ομαδοποιούνται, κατά υποσύνολα, τόσες τιμές απο τις στήλες όσες και τα χρονικά βήματα. Τα νέα υποσύνολα αντιστοιχίζονται με την τιμή του ρύπου μετά το τελευταίο χρονικό βήμα. Επομένως κάθε μια τιμή της συγκέντρωσης του εκάστοτε ρύπου είναι συνυφασμένη με τόσες προηγούμενες τιμές της όσα και τα χρονικά βήματα που έχουν επιλεγεί. Τα βήματα αυτά χρησιμοποιούνται ως ιστορικό για να γίνει πρόβλεψη της νέας τιμής. Στα μοντέλα ημερήσιων προβλέψεων, λόγω έλλειψης εποχικότητας, έχει οριστεί να χρησιμοποιούνται οι προηγούμενες επτά ημέρες για τη νέα πρόβλεψη, για όλες τις πόλεις. Στα μοντέλα των εβδομαδιαίων και μηνιαίων προβλέψεων, ωστόσο, κάθε πόλη εμφάνιζε εποχικότητα της ρύπανσης σε διαφορετικά χρονικά διαστήματα. Προκειμένου να γίνει σωστή εκμετάλλευση της ύπαρξης εποχικότητας, για κάθε μια πόλη ορίστηκαν διαφορετικά πλήθη χρονικών βημάτων βάσει του πίνακα 5.9.

Προβλέψεις	Αθήνα	Παρίσι	Δελχί	Πεκίνο
Ημερήσιες	7	7	7	7
Εβδομαδιαίες	51	43	42	46
Μηνιαίες	12	12	12	11

Πίνακας 5.9: Χρονικά βήματα για κάθε συχνότητα.

5.4.2 CNN

Το πρώτο επίπεδο αυτού του μοντέλου αποτελείται απο ένα μονοδιάστατο συνελικτικό στρώμα (convolutional layer) το οποίο δημιουργείται μέσω της συνάρτησης *Conv1D* της βιβλιοθήκης *keras.layers.convolutional*. Το επίπεδο δρα πάνω σε μια μονοδιάστατη ακολουθία, στην προκείμενη τη χρονοσειρά που αφορά έναν ατμοσφαιρικό ρύπο. Έχει χρησιμοποιηθεί ένας πίνακας 32 φίλτρων και η είσοδος έχει διαστάσεις της μορφης n x k, που διαφέρουν ανάλογα με το αν πρόκειται για ημερήσιες, εβδομαδιαίες ή μηνιαίες προβλέψεις.

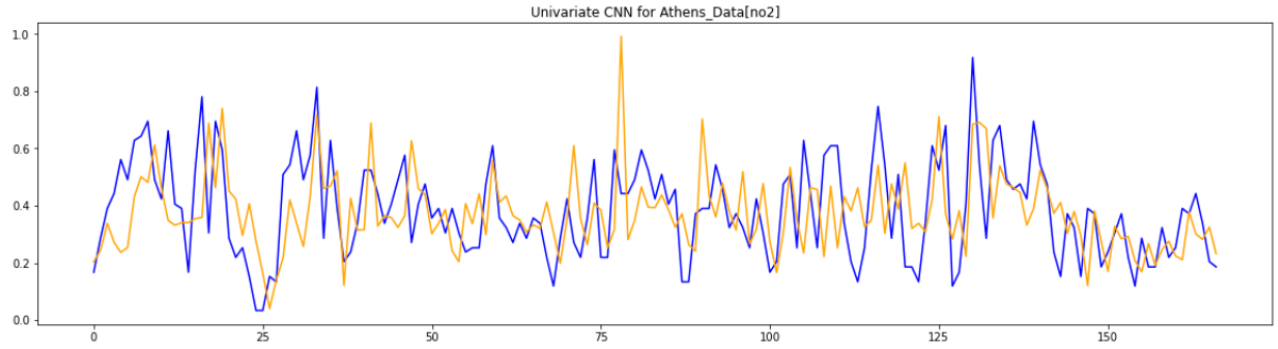
Πιο συγκεκριμένα, η πρώτη διάσταση αναφέρεται στα βήματα του παρελθόντος που λαμβάνονται υπόψιν για τη νέα πρόβλεψη, ενώ η δεύτερη διάσταση αναφέρεται στον αριθμό των στηλών/χαρακτηριστικών που χρησιμοποιούνται για την πρόβλεψη. Έπειτα ακολουθεί το επίπεδο συγκέντρωσης ή σμίκρυνσης (pooling layer), το οποίο τοποθετείται μέσω της συνάρτησης *MaxPooling1D* της βιβλιοθήκης *keras.layers.convolutional*, και αναλαμβάνει να αποστάξει το αποτέλεσμα του συνελικτικού στρώματος ώστε να κρατήσει μόνο τις πιο σημαντικές εισόδους. Στη συνέχεια, τα δεδομένα περνούν σε ένα flatten επίπεδο (flatten layer), το οποίο τοποθετείται μέσω της συνάρτησης *Flatten* της βιβλιοθήκης *keras.layers*, για τη μείωση των χαρακτηριστικών και τη μετατροπή των δεδομένων σε ένα μονοδιάστατο διάνυσμα. Έπεται ένα πυκνό πλήρως συνδεδεμένο στρώμα (dense layer) 50 νευρώνων, το οποίο δημιουργείται απο τη συνάρτηση *Dense* της βιβλιοθήκης *keras.layers*, που ερμηνεύει τα χαρακτηριστικά που εξάγονται από το συνελικτικό τμήμα του μοντέλου. Πριν το επίπεδο εξόδου της τελικής πρόβλεψης, βρίσκεται και ένα στρώμα εγκατάληψης (dropout layer) το οποίο ορίζει τυχαία τις εισόδους ίσες με 0, με πιθανότητα 0.2 σε κάθε βήμα της εκπαίδευσης. Στο στρώμα αυτό αντιστοιχεί η συνάρτηση *Dropout* της βιβλιοθήκης *keras.layers*) και που βοηθά στην αποφυγή υπερβολικής προσαρμογής του μοντέλου στα δεδομένα εκπαίδευσης (overfitting). Οι είσοδοι που δεν έχουν μηδενιστεί, κλιμακώνονται κατά 1/(1 - ρυθμό dropout) έτσι ώστε το άθροισμα όλων των εισόδων να παραμένει αμετάβλητο. Στο τέλος βρίσκεται το στρώμα εξόδου, το οποίο αποτελείται απο έναν νευρώνα που εξάγει ένα μοναδικό αριθμητικό αποτέλεσμα, την πρόβλεψη. Στο μονομεταβλητό μοντέλο όλες οι συναρτήσεις ενεργοποίησης έχουν οριστεί να είναι ReLu, ενώ στο πολυμεταβλητό γραμμικές.

Ως συνάρτηση κόστους στο μοντέλο ορίστηκε το σφάλμα MSE. Το μοντέλο προσπαθεί να ελαχιστοποιήσει αυτό το σφάλμα όλο και περισσότερο σε κάθε βήμα της εκπαίδευσης. Ο βελτιστοποιητής που χρησιμοποιήθηκε είναι ο Adam, ένας αλγόριθμος βελτιστοποίησης που μπορεί να χρησιμοποιηθεί αντί της κλασικής διαδικασίας στοχαστικής κλίσης κατάβασης για την επαναληπτική ενημέρωση των βαρών του δικτύου με βάση τα δεδομένα εκπαίδευσης. Ο βελτιστοποιητής καθώς και η συνάρτηση κόστους του δικτύου ορίζονται ως παράμετροι κατά την κλήση της συνάρτησης *compile* του μοντέλου.

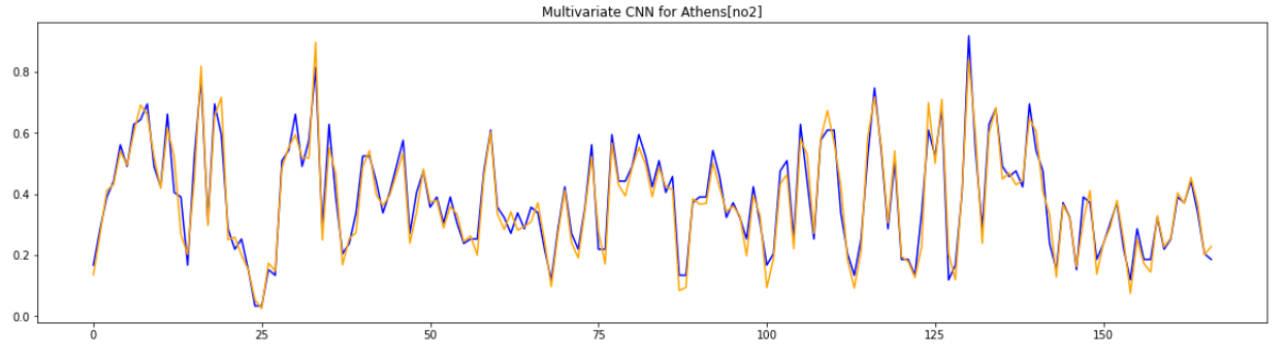
Ακολουθούν τα γραφήματα των ημερήσιων (σχήμα 5.16), εβδομαδιαίων (σχήμα 5.17) και μηνιαίων (σχήμα 5.18) προβλέψεων για τη συγκέντρωση NO_2 στην Αθήνα. Όπως διαπιστώνεται και απο τα σχήματα, οι προβλέψεις των πολυμεταβλητών μοντέλων βρίσκονται αρκετά πιο κοντά στις πραγματικές τιμές. Ωστόσο, οι προβλέψεις εβδομαδιαίας και μηνιαίας

συχνότητας, ιδιαίτερα οι τελευταίες, φαίνεται να έχουν μια σημαντική απόκλιση απο τις πραγματικές τιμές και στα δυο μοντέλα. Αυτό οφείλεται στον μικρό όγκο δεδομένων που διατίθεται για την Αθήνα, ο οποίος μετά την εβδομαδιαία και μηνιαία επαναδειγματοληψία λιγοστεύει ακόμα περισσότερο, με αποτέλεσμα το μοντέλο να μην εκπαιδεύεται επαρκώς.

```
Univariate CNN MSE for Athens[no2]:0.03062419
Univariate CNN RMSE for Athens[no2]:0.17499768668587173
Univariate CNN MAE for Athens[no2]:0.14188974
```

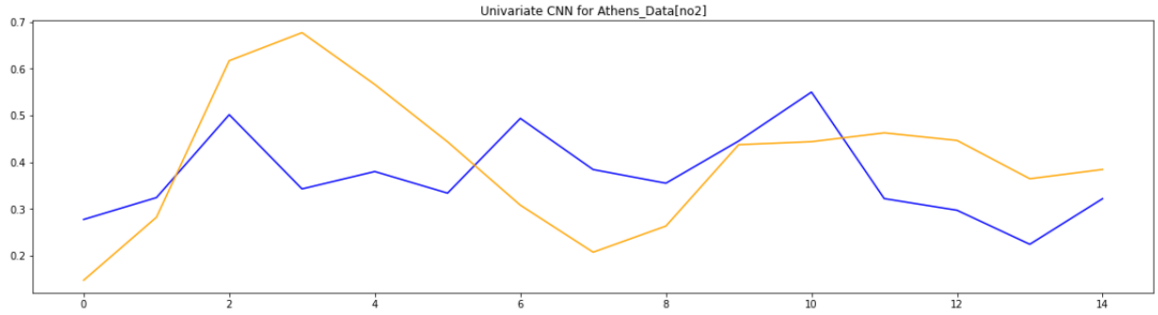



```
Multivariate CNN MSE for Athens[no2]:0.0017991749
Multivariate CNN RMSE for Athens[no2]:0.04241668225703542
Multivariate CNN MAE for Athens[no2]:0.033982728
```

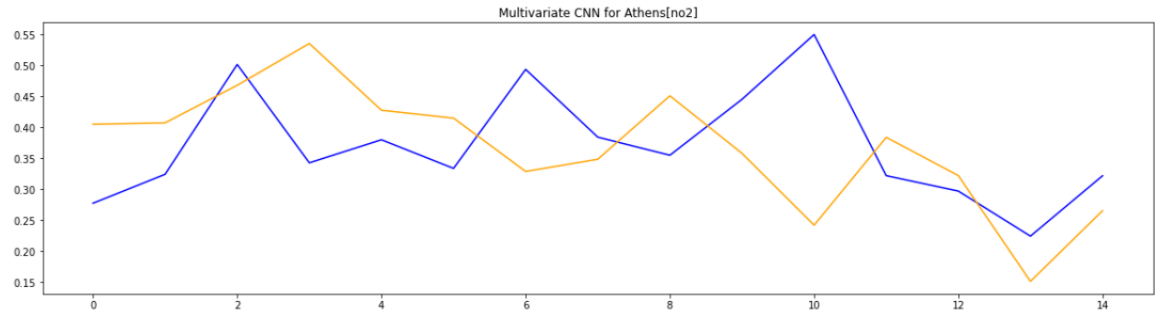


Σχήμα 5.16: Ημερήσιες προβλέψεις (πορτοκαλί γραμμή) της συγκέντρωσης NO_2 στην Αθήνα με μονομεταβλητό και πολυμεταβλητό μοντέλο CNN.

```
Univariate CNN MSE for Athens[no2]:0.022765093
Univariate CNN RMSE for Athens[no2]:0.15088105431666435
Univariate CNN MAE for Athens[no2]:0.1318882
```

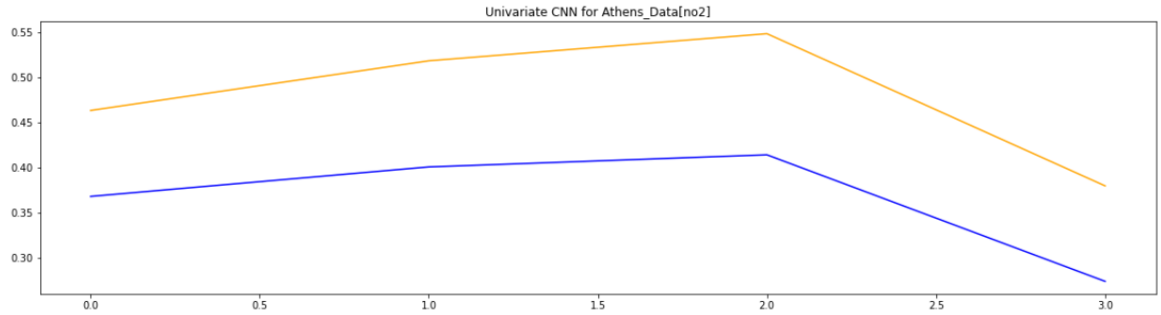



```
Multivariate CNN MSE for Athens[no2]:0.014882045
Multivariate CNN RMSE for Athens[no2]:0.121991986895374
Multivariate CNN MAE for Athens[no2]:0.09823395
```

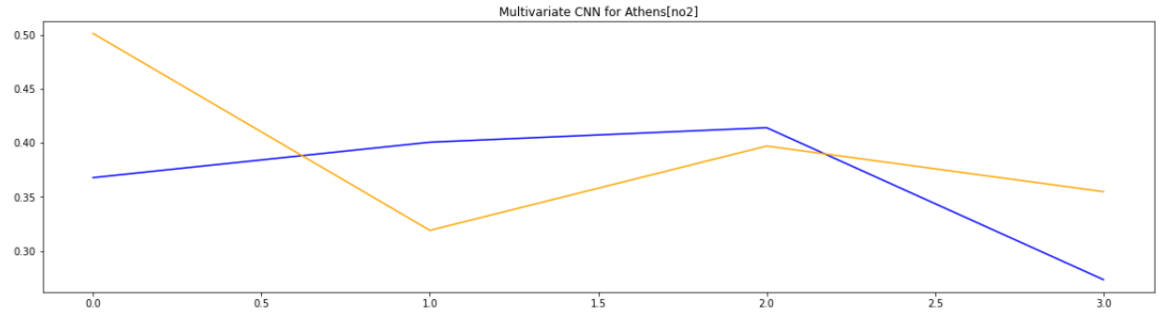


Σχήμα 5.17: Εβδομαδιαίες προβλέψεις (πορτοκαλί γραμμή) της συγκέντρωσης NO_2 στην Αθήνα με μονομεταβλητό και πολυμεταβλητό μοντέλο CNN.

```
Univariate CNN MSE for Athens[no2]:0.013096973
Univariate CNN RMSE for Athens[no2]:0.11444200845834611
Univariate CNN MAE for Athens[no2]:0.11351662
```




```
Multivariate CNN MSE for Athens[no2]:0.007835526
Multivariate CNN RMSE for Athens[no2]:0.08851850664880646
Multivariate CNN MAE for Athens[no2]:0.07831118
```



Σχήμα 5.18: Μηνιαίες προβλέψεις (πορτοκαλί γραμμή) της συγκέντρωσης NO_2 στην Αθήνα με μονομεταβλητό και πολυμεταβλητό μοντέλο CNN.

5.4.3 Αποτελέσματα CNN

Παρακάτω συγκεντρώνονται τα αποτελέσματα που πέτυχαν το μονομεταβλητό και πολυμεταβλητό μοντέλο CNN στις τρεις διαφορετικές συχνότητες για τις οποίες έγιναν οι προβλέψεις, για κάθε πόλη ξεχωριστά. Τα μοντέλα ίδιων συχνοτήτων βρίσκονται στο ίδιο χρωματικό πλαίσιο. Οι χρωματιστές τιμές αντιστοιχούν στο μικρότερο RMSE σφάλμα όλων των μοντέλων της ίδια συχνότητας ανά ρύπο. Οι υπογραμμισμένες τιμές αποτελούν τη χρονοσειρά που σημείωσε το μικρότερο RMSE σφάλμα για την εκάστοτε μέθοδο.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.0120	0.03062	0.02395	0.02012	0.01429
Univariate CNN (daily)	RMSE	0.10982	0.1749	0.15476	0.14186	0.11955
	MAE	0.0845	0.14188	0.0999	0.1024	0.0930
	MSE	0.00132	0.00179	0.00187	0.00127	0.00161
Multivariate CNN (daily)	RMSE	0.0363	0.0424	0.0432	0.0357	0.0401
	MAE	0.02807	0.03398	0.0299	0.0274	0.02985
	MSE	0.0064	0.0227	0.0447	0.0092	0.0080
Univariate CNN (weekly)	RMSE	0.0801	0.1508	0.2115	0.0962	0.0895
	MAE	0.0677	0.1318	0.1333	0.0822	0.0668
	MSE	0.0115	0.0148	0.0390	0.01563	0.0081
Multivariate CNN (weekly)	RMSE	0.1073	0.1219	0.1976	0.1250	0.0901
	MAE	0.0874	0.0982	0.1282	0.1031	0.0689
	MSE	0.0054	0.0130	0.0204	0.00606	0.0045
Univariate CNN (monthly)	RMSE	0.0735	0.1144	0.1431	0.0778	0.0671
	MAE	0.05745	0.1135	0.1037	0.0606	0.0464
	MSE	0.0288	0.00783	0.02108	0.00595	0.00266
Multivariate CNN (monthly)	RMSE	0.1699	0.0885	0.1451	0.0077	0.0516
	MAE	0.1657	0.0783	0.099	0.0612	0.0459

Πίνακας 5.10: Αποτελέσματα προβλέψεων CNN μοντέλων για την Αθήνα.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.0203	0.0170	0.00261	0.01341	0.0109
Univariate CNN (daily)	RMSE	0.1427	0.1304	0.0511	0.1158	0.1044
	MAE	0.1102	0.1022	0.0336	0.0882	0.0814
	MSE	0.00108	0.00083	0.00024	0.00095	0.00054
Multivariate CNN (daily)	RMSE	0.0329	0.0289	0.0157	0.0309	0.0233
	MAE	0.0252	0.0232	0.0126	0.0240	0.0182
	MSE	0.0133	0.0116	0.00084	0.0113	0.0143
Univariate CNN (weekly)	RMSE	0.1155	0.1081	0.0290	0.1063	0.1199
	MAE	0.0887	0.0890	0.0234	0.0855	0.0978
	MSE	0.0068	0.0050	0.0011	0.0036	0.0033
Multivariate CNN (weekly)	RMSE	0.0828	0.0713	0.0344	0.0606	0.0579
	MAE	0.0676	0.0594	0.0290	0.0487	0.0457
	MSE	0.0101	0.00663	0.00031	0.0061	0.0072
Univariate CNN (monthly)	RMSE	0.1008	0.0814	0.0177	0.0785	0.0850
	MAE	0.0738	0.0657	0.0140	0.0660	0.0713
	MSE	0.00247	0.0030	0.00048	0.00343	0.0021
Multivariate CNN (monthly)	RMSE	0.0497	0.0548	0.0220	0.0585	0.0466
	MAE	0.0402	0.0383	0.0173	0.0479	0.0387

Πίνακας 5.11: Αποτελέσματα προβλέψεων CNN μοντέλων για το Παρίσι.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.0139	0.0100	0.00017	0.00311	0.01076
Univariate CNN (daily)	RMSE	0.1179	0.100	0.01331	0.05577	0.1037
	MAE	0.0920	0.0719	0.0105	0.0207	0.0144
	MSE	0.00067	0.00053	0.00008	0.00022	0.00095
Multivariate CNN (daily)	RMSE	0.0259	0.0232	0.0091	0.0150	0.0308
	MAE	0.0202	0.0175	0.0073	0.0094	0.0254
	MSE	0.0103	0.00588	0.00011	0.00065	0.00646
Univariate CNN (weekly)	RMSE	0.1018	0.0766	0.0108	0.0256	0.0803
	MAE	0.0760	0.05977	0.0080	0.0164	0.0662
	MSE	0.0190	0.0129	0.0021	0.0019	0.0209
Multivariate CNN (weekly)	RMSE	0.1379	0.1138	0.0461	0.0436	0.1447
	MAE	0.1117	0.0919	0.0389	0.0342	0.1202
	MSE	0.00444	0.00205	0.00003	0.00072	0.00675
Univariate CNN (monthly)	RMSE	0.0667	0.0453	0.0059	0.0269	0.0821
	MAE	0.0533	0.0380	0.0051	0.0203	0.0690
	MSE	0.0026	0.00127	0.00038	0.00068	0.00145
Multivariate CNN (monthly)	RMSE	0.0510	0.0357	0.0196	0.0261	0.0381
	MAE	0.0392	0.0311	0.0146	0.0214	0.0305

Πίνακας 5.12: Αποτελέσματα προβλέψεων CNN μοντέλων για το Πεκίνο.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.00482	0.00243	0.00198	0.00237	0.00445
Univariate CNN (daily)	RMSE	0.0694	0.0493	0.04460	0.0487	0.0667
	MAE	0.0521	0.0383	0.0316	0.0288	0.0459
	MSE	0.00039	0.00013	0.00012	0.00009	0.00021
Multivariate CNN (daily)	RMSE	0.0199	0.0117	0.0112	0.0098	0.0145
	MAE	0.0146	0.0093	0.0090	0.0070	0.0115
	MSE	0.0029	0.0046	0.0060	0.0022	0.0062
Univariate CNN (weekly)	RMSE	0.0541	0.0680	0.0077	0.0478	0.0788
	MAE	0.0399	0.0547	0.0647	0.0347	0.0579
	MSE	0.0081	0.0077	0.0079	0.0083	0.0092
Multivariate CNN (weekly)	RMSE	0.0904	0.0881	0.08897	0.09138	0.0960
	MAE	0.0746	0.0704	0.0692	0.0762	0.079
	MSE	0.0009	0.0104	0.0128	0.0016	0.0022
Univariate CNN (monthly)	RMSE	0.0301	0.1023	0.1134	0.0401	0.0472
	MAE	0.0237	0.0711	0.0800	0.0316	0.0372
	MSE	0.0022	0.0074	0.0036	0.0033	0.0068
Multivariate CNN (monthly)	RMSE	0.0474	0.0687	0.0606	0.0577	0.0825
	MAE	0.0345	0.0582	0.0483	0.0515	0.0619

Πίνακας 5.13: Αποτελέσματα προβλέψεων CNN μοντέλων για το Δελχί.

Απο την παρατήρηση των παραπάνω πινάκων και τη σύγκριση των επιδόσεων των διαφορετικών μοντέλων για κάθε πόλη, προκύπτει ότι στην πλειοψηφία τους τα πολυμεταβλητά CNN μοντέλα πέτυχαν μικρότερα σφάλματα από τα αντίστοιχα μονομεταβλητά, με εξαίρεση τις εβδομαδιαίες προβλέψεις για τις συγκεντρώσεις ρύπων στο Πεκίνο και τις εβδομαδιαίες και μηνιαίες προβλέψεις για το Δελχί. Ακόμα, για όλες τις πόλεις, φαίνεται πως οι προβλέψεις εβδομαδιαίας και μηνιαίας συχνότητας με το μονομεταβλητό μοντέλο είχαν μικρότερα σφάλματα συγκριτικά με τις ημερήσιες προβλέψεις. Αντίθετα, το πολυμεταβλητό μοντέλο λειτούργησε καλύτερα στη διεξαγωγή ημερήσιων προβλέψεων.

Επιπλέον, τα μοντέλα που εφαρμόστηκαν στα δεδομένα της Αθήνας είχαν χειρότερες

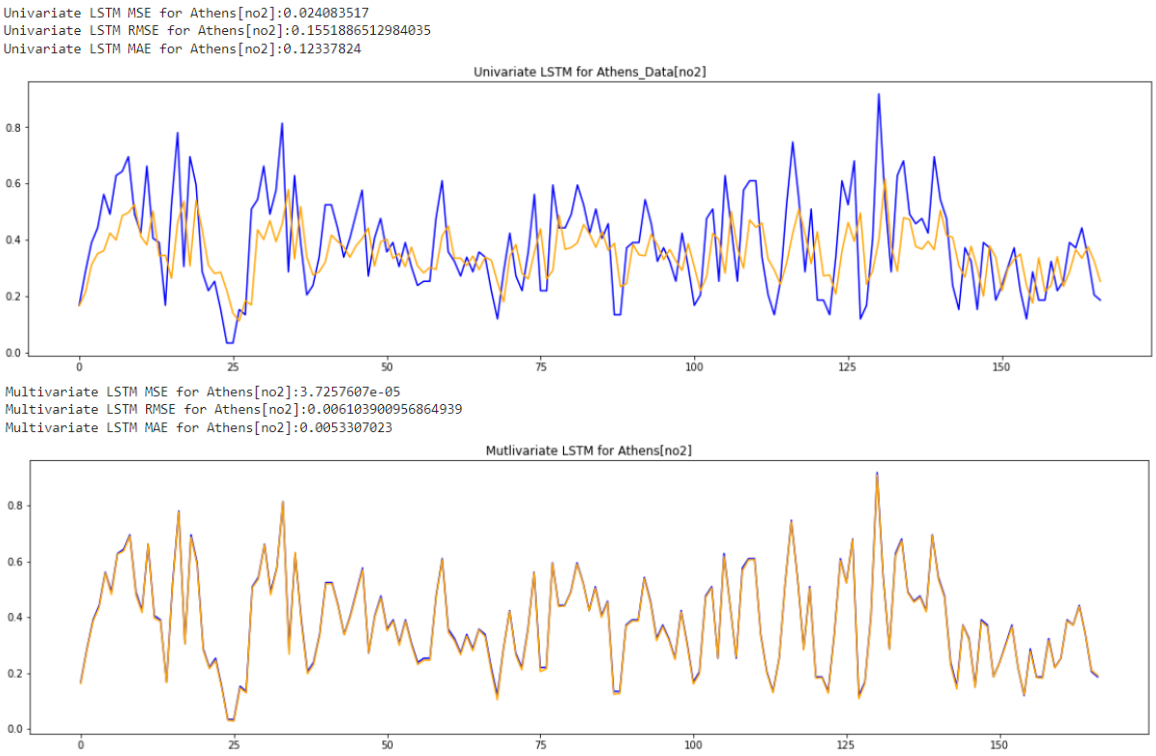
επιδόσεις συγκριτικά με τα αντίστοιχα μοντέλα τα οποία εφαρμόστηκαν στις υπόλοιπες πόλεις με μεγαλύτερο πλήθος διαθέσιμων δεδομένων.

Αναφορικά με τους ρύπους που προβλέφθηκαν, ξεχώρισε η πρόβλεψη της συγκέντρωσης SO_2 στο Παρίσι, η οποία είχε τα μικρότερα σφάλματα σε όλα τα μοντέλα και τις συχνότητες, αλλά και στο Πεκίνο, όπου σημείωσε τα μικρότερα σφάλματα σε όλες τις προβλέψεις εκτός των εβδομαδιαίων με το πολυμεταβλητό μοντέλο.

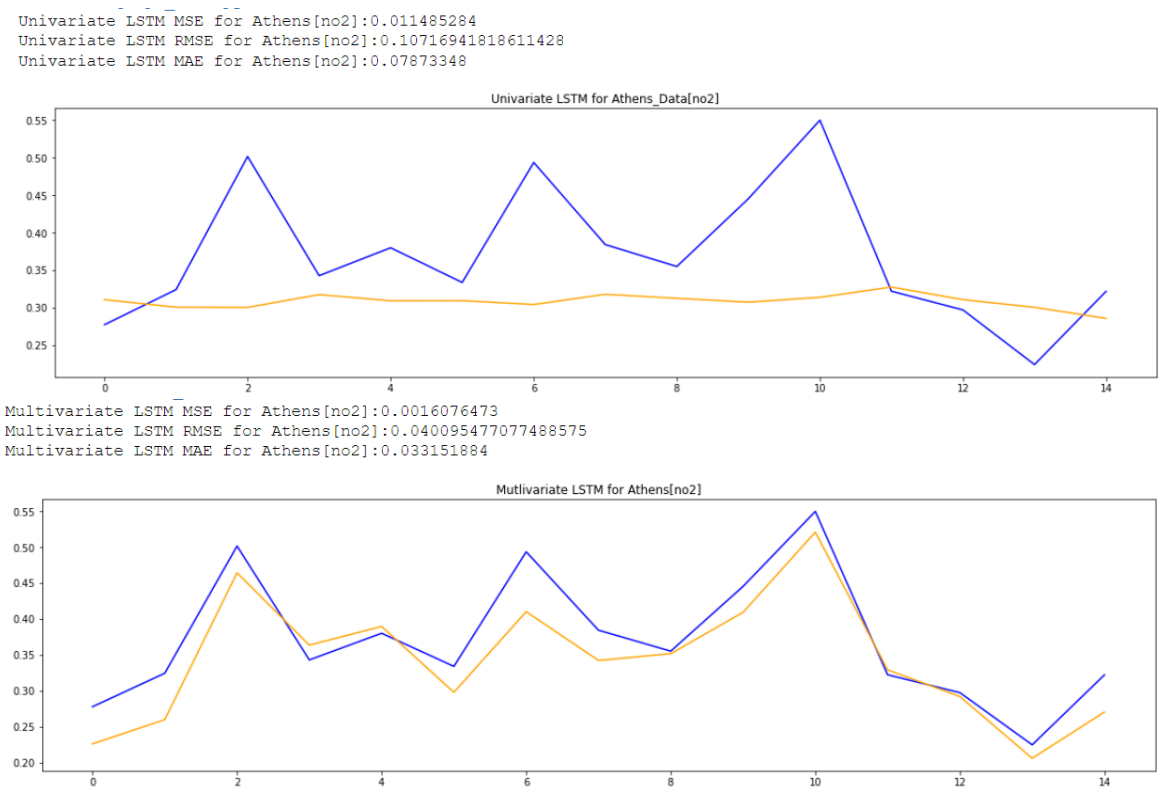
5.4.4 LSTM

Το πρώτο επίπεδο αυτού του μοντέλου αποτελεί ένα στρώμα 50 LSTM νευρώνων, τοποθετημένο μέσω της συνάρτησης *LSTM* της βιβλιοθήκης *keras.layers*, όπου οι διαστάσεις εισόδου ακολουθούν το πρότυπο που αναπτύχθηκε στα συνελικτικά δίκτυα. Έπειτα, τοποθετείται ένα πλήρως διασυνδεδεμένο πυκνό στρώμα 30 νευρώνων, μέσω της συνάρτησης *Dense* της βιβλιοθήκης *keras.layers*, για την ερμηνεία των εξαγόμενων χαρακτηριστικών απο τα δεδομένα. Στη συνέχεια, ακολουθεί ένα επίπεδο εγκατάληψης, το οποίο δημιουργείται μέσω της συνάρτησης *Dropout* της βιβλιοθήκης *keras.layers*, ώστε το μοντέλο να μην δημιουργεί εξαρτήσεις με καμία είσοδο. Στο επίπεδο αυτό οι είσοδοι μηδενίζονται τυχαία με πιθανότητα ίση με την παράμετρο της συνάρτησης, στην περίπτωση μας 0.2. Τέλος ακολουθεί το στρώμα εξόδου με έναν νευρώνα απο τον οποίο θα εξαχθεί η πρόβλεψη της νέας τιμής. Σε όλα τα επίπεδα, εκτός απο το επίπεδο εγκατάληψης, χρησιμοποιούνται συναρτήσεις ενεργοποίησης. Στη δομή του μονομεταβλητού μοντέλου όλες οι συναρτήσεις ενεργοποίησης έχουν οριστεί να είναι ReLu, ενώ στο πολυμεταβλητό γραμμικές.

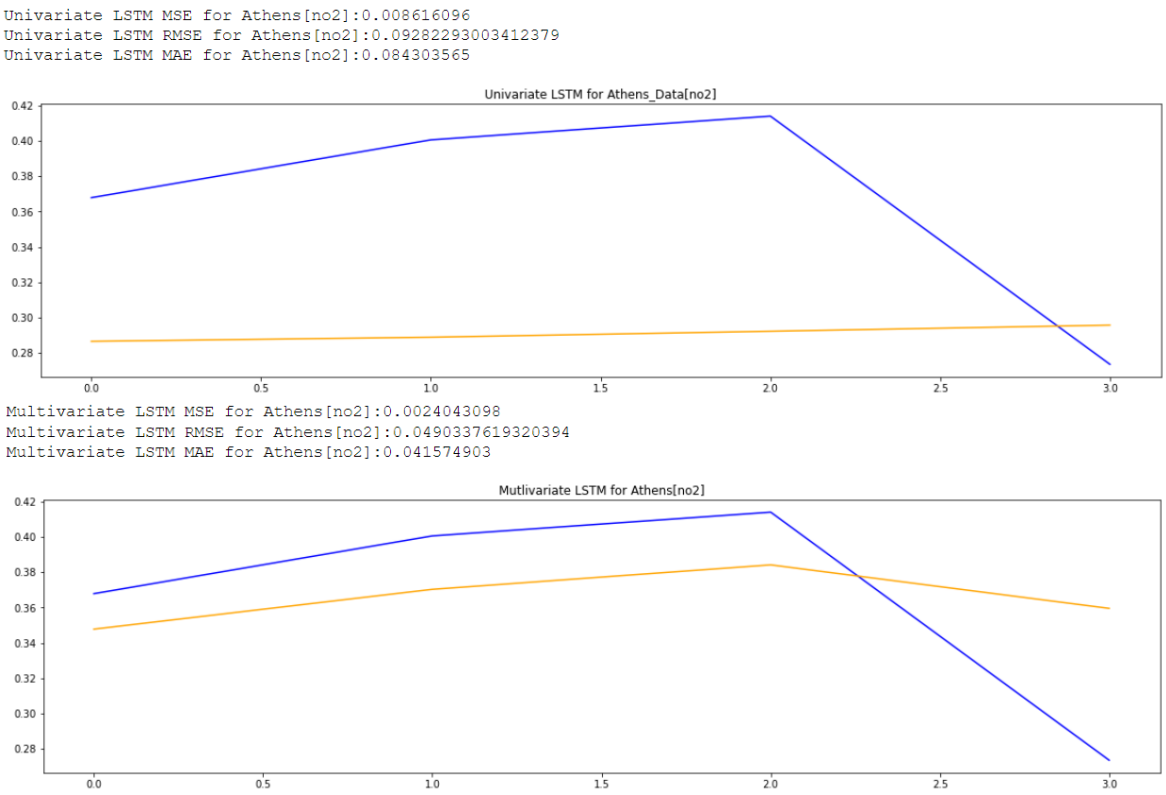
Η συνάρτηση κόστους που προσπαθεί να ελαχιστοποίησει το μοντέλο ώστε να βελτιώσει τις προβλέψεις είναι η MSE. Στο μοντέλο έχει χρησιμοποιηθεί ως βελτιστοποιητής ο RMSProp με ποσοστό εκμάθησης 0.0001, ο οποίος προσπαθεί να επιλύσει το πρόβλημα ότι οι κλίσεις της συνάρτησης κόστους μπορεί να ποικίλλουν πολύ σε μεγέθη. Ο βελτιστοποιητής καθώς και η συνάρτηση κόστους του δικτύου ορίζονται ως παράμετροι κατά την κλήση της συνάρτησης *compile* του μοντέλου.



Σχήμα 5.19: Ημερήσιες προβλέψεις (πορτοκαλί γραμμή) της συγκέντρωσης NO_2 στην Αθήνα με μονομεταβλητό και πολυμεταβλητό μοντέλο LSTM.



Σχήμα 5.20: Εβδομαδιαίες προβλέψεις (πορτοκαλί γραμμή) της συγκέντρωσης NO_2 στην Αθήνα με μονομεταβλητό και πολυμεταβλητό μοντέλο LSTM.



Σχήμα 5.21: Μηνιαίες προβλέψεις (πορτοκαλί γραμμή) της συγκέντρωσης NO_2 στην Αθήνα με μονομεταβλητό και πολυμεταβλητό μοντέλο LSTM.

Παρατηρώντας τα σχήματα 5.19-5.21, βλέπουμε πως οι προβλέψεις του πολυμεταβλητού LSTM μοντέλου φτάνουν αρκετά πιο κοντά στις πραγματικές τιμές και στις τρεις χρονικές συχνότητες, συγκριτικά με το μονομεταβλητό μοντέλο.

5.4.5 Αποτελέσματα LSTM

Ακολουθούν τα αποτελέσματα των μοντέλων για όλες τις χρονικές συχνότητες, καταγεγραμμένα σε έναν πίνακα για κάθε πόλη ξεχωριστά. Τα μοντέλα ίδιων συχνοτήτων βρίσκονται στο ίδιο χρωματικό πλαίσιο. Οι χρωματιστές τιμές αντιστοιχούν στο μικρότερο RMSE σφάλμα όλων των μοντέλων της ίδια συχνότητας ανά ρύπο. Οι υπογραμμισμένες τιμές αποτελούν τη χρονοσειρά του ρύπου που σημείωσε το μικρότερο RMSE σφάλμα για την εκάστοτε μέθοδο.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.0091	0.0240	0.0197	0.0132	0.0112
Univariate LSTM (daily)	RMSE	0.0955	0.1551	0.1404	0.1151	0.1062
	MAE	0.0717	0.1233	0.0881	0.0816	0.0778
	MSE	0.00003	0.00003	0.00004	0.00003	0.00001
Multivariate LSTM (daily)	RMSE	0.00558	0.00610	0.00663	0.00588	0.00432
	MAE	0.0043	0.0053	0.0044	0.0051	0.0035
	MSE	0.0038	0.0114	0.0318	0.0056	0.0042
Univariate LSTM (weekly)	RMSE	0.0621	0.1071	0.1784	0.0750	0.0653
	MAE	0.0535	0.0787	0.1158	0.0591	0.0558
	MSE	0.0018	0.0016	0.0171	0.0006	0.0007
Multivariate LSTM (weekly)	RMSE	0.0428	0.0400	0.1310	0.0245	0.0265
	MAE	0.0357	0.0331	0.0700	0.0181	0.0225
	MSE	0.0136	0.0086	0.0213	0.0036	0.00064
Univariate LSTM (monthly)	RMSE	0.1167	0.0928	0.1461	0.0602	0.0254
	MAE	0.1057	0.0843	0.1144	0.0543	0.0229
	MSE	0.0015	0.0024	0.0224	0.0009	0.0006
Multivariate LSTM (monthly)	RMSE	0.0396	0.0490	0.1499	0.0301	0.0258
	MAE	0.0340	0.0415	0.0961	0.0243	0.0222

Πίνακας 5.14: Αποτελέσματα προβλέψεων LSTM μοντέλων για την Αθήνα.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.0148	0.0132	0.0022	0.0117	0.0098
Univariate LSTM (daily)	RMSE	0.1218	0.1153	0.0477	0.1084	0.0992
	MAE	0.0941	0.0912	0.0326	0.0808	0.0769
	MSE	0.00002	0.00007	$3 * 10^{-6}$	$7 * 10^{-6}$	$8 * 10^{-6}$
Multivariate LSTM (daily)	RMSE	0.0052	0.0084	0.0018	0.0027	0.0028
	MAE	0.0041	0.0073	0.0014	0.0019	0.0025
	MSE	0.0097	0.0070	0.0005	0.0095	0.0101
Univariate LSTM (weekly)	RMSE	0.0989	0.0841	0.0229	0.0976	0.1007
	MAE	0.0764	0.0671	0.0180	0.0780	0.0790
	MSE	$8 * 10^{-5}$	0.00024	$8 * 10^{-5}$	$3 * 10^{-5}$	$8 * 10^{-5}$
Multivariate LSTM (weekly)	RMSE	0.0092	0.0156	0.0093	0.0058	0.0090
	MAE	0.0081	0.0149	0.0066	0.0043	0.0079
	MSE	0.0105	0.0056	0.0003	0.0050	0.0056
Univariate LSTM (monthly)	RMSE	0.1029	0.0749	0.0197	0.0713	0.0749
	MAE	0.0818	0.0609	0.0179	0.0599	0.0664
	MSE	0.0007	0.0010	0.0002	0.0014	0.0008
Multivariate LSTM (monthly)	RMSE	0.0267	0.0325	0.0166	0.0382	0.0297
	MAE	0.0200	0.0280	0.0154	0.0321	0.0236

Πίνακας 5.15: Αποτελέσματα προβλέψεων LSTM μοντέλων για το Παρίσι.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.0100	0.0076	0.0001	0.0030	0.0092
Univariate LSTM (daily)	RMSE	0.1002	0.0873	0.0114	0.0554	0.0961
	MAE	0.0797	0.0663	0.0080	0.0205	0.0780
	MSE	0.0001	0.00003	0.00002	0.00002	$6 * 10^{-6}$
Multivariate LSTM (daily)	RMSE	0.0123	0.0060	0.0052	0.00521	0.0025
	MAE	0.01193	0.00554	0.00518	0.00496	0.00203
	MSE	0.00508	0.00411	0.00009	0.00056	0.00586
Univariate LSTM (weekly)	RMSE	0.0713	0.0641	0.0098	0.0237	0.0765
	MAE	0.0541	0.0517	0.0087	0.0146	0.0654
	MSE	0.00010	0.00031	0.00017	0.00012	0.00002
Multivariate LSTM (weekly)	RMSE	0.0101	0.0178	0.01323	0.0113	0.0049
	MAE	0.00897	0.01712	0.01180	0.0085	0.00405
	MSE	0.00378	0.0030	0.00013	0.00044	0.03052
Univariate LSTM (monthly)	RMSE	0.0615	0.0547	0.0115	0.0210	0.1747
	MAE	0.0530	0.0451	0.0106	0.0158	0.1656
	MSE	0.00037	0.00045	0.00018	0.00027	0.00137
Multivariate LSTM (monthly)	RMSE	0.0192	0.0213	0.0135	0.0166	0.0371
	MAE	0.0156	0.0176	0.0115	0.0122	0.0327

Πίνακας 5.16: Αποτελέσματα προβλέψεων LSTM μοντέλων για το Πεκίνο.

Model	Error	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
	MSE	0.0034	0.0020	0.0014	0.0016	0.0032
Univariate LSTM (daily)	RMSE	0.0585	0.0452	0.0374	0.0400	0.0569
	MAE	0.0429	0.0343	0.0263	0.0257	0.0392
	MSE	$5 * 10^{-6}$	$3 * 10^{-6}$	$3 * 10^{-6}$	$1 * 10^{-6}$	0.000048
Multivariate LSTM (daily)	RMSE	0.00243	0.00191	0.00190	0.00036	0.00696
	MAE	0.00205	0.00153	0.00155	0.00329	0.00677
	MSE	0.00217	0.00297	0.00183	0.00205	0.00402
Univariate LSTM (weekly)	RMSE	0.0465	0.0545	0.0428	0.0453	0.0634
	MAE	0.0329	0.0406	0.0347	0.0334	0.0464
	MSE	0.00015	0.00006	0.00006	0.00008	0.00031
Multivariate LSTM (weekly)	RMSE	0.0125	0.00078	0.0079	0.0090	0.0176
	MAE	0.0107	0.0063	0.0060	0.0083	0.0157
	MSE	0.00064	0.00792	0.00599	0.00296	0.005715
Univariate LSTM (monthly)	RMSE	0.0253	0.0890	0.0774	0.0544	0.0755
	MAE	0.0202	0.0769	0.0696	0.0481	0.0648
	MSE	0.0008	0.0010	0.0019	0.0005	0.0011
Multivariate LSTM (monthly)	RMSE	0.0290	0.0317	0.0436	0.0234	0.0343
	MAE	0.0226	0.0285	0.0377	0.0214	0.0293

Πίνακας 5.17: Αποτελέσματα προβλέψεων LSTM μοντέλων για το Δελχί.

Απο τη καταγραφή και την παρατήρηση των αποτελεσμάτων για τα LSTM δίκτυα προκύπτουν όμοια αποτελέσματα με τα CNN δίκτυα. Τα πολυμεταβλητά μοντέλα λειτούργησαν αποδοτικότερα από ότι τα αντίστοιχα μονομεταβλητά για κάθε πόλη. Η σύγκριση των μονομεταβλητών μοντέλων διαφορετικής συχνότητας μεταξύ τους δείχνει ότι είναι ακριβέστερα στις εβδομαδιαίες και μηνιαίες προβλέψεις, ενώ τα πολυμεταβλητά μοντέλα στις ημερήσιες.

Οι επιδόσεις των μοντέλων που εφαρμόστηκαν στις πόλεις με μεγάλο όγκο δεδομένων ήταν περισσότερο ικανοποιητικές από τα αντίστοιχα μοντέλα τα οποία εφαρμόστηκαν στην Αθήνα.

Η συγκέντρωση που κατάφερε να προβλεφθεί με ιδιαίτερη επιτυχία ήταν η αυτή του SO_2 στο Παρίσι και το Πεκίνο, όπου και απέδωσε τα μικρότερα σφάλματα σε πέντε και σε

τέσσερα, αντίστοιχα, απο τα έξι μοντέλα και προβλέφθηκε με μεγάλη ακρίβεια για όλες τις συχνότητες.

5.5 Συγκριτικά αποτελέσματα των μοντέλων

Τα συνελικτικά νευρωνικά δίκτυα είναι φτιαγμένα ωστε να αναγνωρίζουν μοτίβα και πρότυπα σε ιδιαίτερα μεγάλες ακολουθίες δεδομένων. Με μικρούς όγκους μπορεί να οδηγηθούν σε υπερπροσαρμογή στα δεδομένα εκπαίδευσης, άρα και σε μεγαλύτερα σφάλματα κατά τη διαδικασία ελέγχου. Επιπλέον τα στατιστικά μοντέλα είναι ειδικά σχεδιασμένα ώστε να εντοπίζουν την εποχικότητα και την τάση στα δεδομένα των χρονοσειρών. Λόγω των προηγούμενων παρατηρήσεων, τα στατιστικά μοντέλα πέτυχαν καλύτερα αποτελέσματα στις προβλέψεις των εβδομαδιαίων και μηνιαίων δεδομένων, τα οποία εμφάνισαν εποχικότητα, συγκριτικά με τα CNN δίκτυα. Ωστόσο, τα LSTM δίκτυα πέτυχαν, κατά πλειοψηφία, καλύτερες επιδόσεις σε όλες τις συχνότητες των προβλέψεων, γεγονός που οφείλεται στη μνήμη την οποία διαθέτουν και χρησιμοποιούν προκειμένου να εξάγουν τις προβλέψεις τους.

Ακολουθούν συγκεντρωμένα τα αποτελέσματα των καλύτερων μεθόδων πρόβλεψης ανά πόλη και ανά ατμοσφαιρικό ρύπο.

Time Series	Daily	Weekly	Monthly
O_3	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM
NO_2	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM
SO_2	Multivariate LSTM	Multivariate LSTM	FB Prophet
PM_{10}	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM
$PM_{2.5}$	Multivariate LSTM	Multivariate LSTM	Univariate LSTM

Πίνακας 5.18: Επιτυχέστερα μοντέλα ανά συχνότητα, για την πρόβλεψη των ατμοσφαιρικών ρύπων στην Αθήνα.

Time Series	Daily	Weekly	Monthly
O_3	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM
NO_2	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM
SO_2	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM
PM_{10}	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM
$PM_{2.5}$	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM

Πίνακας 5.19: Επιτυχέστερα μοντέλα ανά συχνότητα, για την πρόβλεψη των ατμοσφαιρικών ρύπων στο Παρίσι.

Time Series	Daily	Weekly	Monthly
O_3	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM
NO_2	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM
SO_2	Multivariate LSTM	Manual ARIMA	Manual ARIMA
PM_{10}	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM
$PM_{2.5}$	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM

Πίνακας 5.20: Επιτυχέστερα μοντέλα ανά συχνότητα, για την πρόβλεψη των ατμοσφαιρικών ρύπων στο Πεκίνο.

Time Series	Daily	Weekly	Monthly
O_3	Multivariate LSTM	Multivariate LSTM	Univariate LSTM
NO_2	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM
SO_2	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM
PM_{10}	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM
$PM_{2.5}$	Multivariate LSTM	Multivariate LSTM	Multivariate LSTM

Πίνακας 5.21: Επιτυχέστερα μοντέλα ανά συχνότητα, για την πρόβλεψη των ατμοσφαιρικών ρύπων στο Δελχί.

Κεφάλαιο 6

Επίλογος

Έπειτα απο τη δημιουργία και τη δοκιμή όλων των μοντέλων, διεξήχθησαν ορισμένα συμπεράσματα που αφορούν την ακρίβεια καθενός απο αυτά, καθώς και την επίδραση που είχαν οι μετατροπές στα δεδομένα.

6.1 Σύνοψη των μοντέλων

6.1.1 Σύνοψη των μοντέλων ARIMA

Αναπτύχθηκαν δυο εκδοχές για το στατιστικό μοντέλο ARIMA για τις ημερήσιες προβλέψεις, μια αυτόματη (auto-ARIMA), την οποία παρέχει η βιβλιοθήκη *pmdarima* της *Python*, και μια η οποία δημιουργήθηκε χειροκίνητα (manual ARIMA). Όπως διαπιστώθηκε απο την καταγραφή των αποτελεσμάτων σε πίνακες, η ακρίβεια του μοντέλου auto-ARIMA δεν είναι ιδιαίτερα ικανοποιητική, αυτό συμβαίνει διότι το αυτόματο μοντέλο δεν προσαρμόζεται επαρκώς στις απαιτήσεις κάθε χρονοσειράς, καθώς επίσης υπάρχει και σημαντικός περιορισμός στο άνω όριο των παραμέτρων AR, MA και διαφόρισης. Αντίθετα, το manual ARIMA είχε ιδιαίτερα βελτιωμένα αποτελέσματα καθώς προηγήθηκε επεξεργασία των δεδομένων, με σκοπό να προσδιοριστούν οι κατάλληλες παράμετροι για κάθε χρονοσειρά.

Έπειτα, η διαδικασία ημερήσιων προβλέψεων για το Παρίσι, το Πεκίνο και το Δελχί επαναλήφθηκε με μικρότερο αριθμό δεδομένων, συγκεκριμένα ίδιο με το πλήθος ημερήσιων δεδομένων της Αθήνας, ο οποίος ήταν αρκετά μικρότερος. Το πείραμα επιβεβαίωσε πως ο όγκος των δεδομένων εκπαίδευσης διαδραματίζει σημαντικό ρόλο στις επιδόσεις των μοντέλων.

Τέλος, οι προβλέψεις που διεξήχθησαν στα επαναδειγματοληφθέντα δεδομένα με το μο-

ντέλο manual ARIMA, δηλαδή τα δεδομένα εβδομαδιαίας και μηνιαίας συχνότητας, φάνηκε να είναι περισσότερο ακριβείς απο τις ημερήσιες προβλέψεις για κάθε εξεταζόμενη πόλη. Το γεγονός αυτό, επιβεβαιώνει οτι για το στατιστικό μοντέλο ARIMA, η ύπαρξη εποχικότητας έχει σημασία για την εγκυρότητα του καθώς την αναγνωρίζει και μπορεί να την ενσωματώσει στις προβλεψεις, ώστε να πετύχει μεγαλύτερη ακρίβεια.

6.1.2 Σύνοψη του μοντέλου Prophet

Το μοντέλο Prophet αποτελεί επίσης ένα στατιστικό μοντέλο πρόβλεψης χρονοσειρών. Όμοια με το μοντέλο ARIMA, έγιναν προσπάθειες πρόβλεψης δεδομένων ημερήσιας, εβδομαδιαίας και μηνιαίας συχνότητας καθώς επίσης και ημερήσιες προβλέψεις με τη χρήση μικρότερου όγκου δεδομένων.

Και με αυτό το μοντέλο επιβεβαιώθηκε το γεγονός πως ο μεγαλύτερος όγκος δεδομένων εκπαίδευσης συμβάλλει στην εγκυρότητα του μοντέλου, διότι τα αποτελέσματα για τις ημερήσιες προβλέψεις με όλα τα διαθέσιμα δεδομένα των πόλεων είχαν σημαντικά καλύτερα αποτελέσματα.

Οι προβλέψεις για διαφορετικές χρονικές συχνότητες, οι οποίες εμφάνιζαν εποχικότητα, φαίνεται πως και σε αυτό το μοντέλο λειτούργησαν σημαντικά καλύτερα απο ότι οι ημερήσιες, οι οποίες δεν παρουσίαζαν μοτίβα εποχικότητας.

6.1.3 Σύνοψη των νευρωνικών δικτύων

Τα μοντέλα CNN και LSTM εξετάστηκαν στη διεξαγωγή ημερήσιων, εβδομαδιαίων και μηνιαίων προβλέψεων, μονομεταβλητά και πολυμεταβλητά. Οι ημερήσιες προβλέψεις δημιουργούνται χρησιμοποιώντας το ιστορικό των επτά προηγούμενων ημερών για κάθε ρύπο, ενώ το ιστορικό που χρησιμοποιούν οι προβλέψεις για την εβδομαδιαία και μηνιαία συχνότητα εξαρτάται απο το διάστημα εποχικότητας που εμφανίζει κάθε πόλη για κάθε συχνότητα (5.9).

Τα πολυμεταβλητά μοντέλα πέτυχαν συντριπτικά καλύτερα αποτελέσματα απο τα αντίστοιχα μονομεταβλητά στις ημερήσιες προβλέψεις. Αυτό συμβαίνει διότι η επίδραση των καιρικών και ατμοσφαιρικών συνθηκών διαδραματίζει πολύ σημαντικό ρόλο στο ποσοστό συγκέντρωσης ρύπων στην ατμόσφαιρα. Τα μονομεταβλητά μοντέλα δεν εκμεταλλεύονται την εξάρτηση αυτή, καθώς εξετάζουν μόνο τιμές της συγκέντρωσης των ρύπων σε προηγούμενα χρονικά βήματα. Τα πολυμεταβλητά μοντέλα πέτυχαν κατά πλειοψηφία καλύτερη επίδοση απο τα μονομεταβλητά και στις εβδομαδιαίες και μηνιαίες προβλέψεις, με εξαίρεση

τα μοντέλα CNN για το Δελχί πιθανότατα λόγω του ταραχώδους κλίματος, ωστόσο, όχι στο βαθμό που παρατηρήθηκε στα ημερήσια δεδομένα.

Εκτός απο τη σύγκριση που έγινε μεταξύ των προβλέψεων ίδιας συχνότητας με δυο διαφορετικά μοντέλα (μονομεταβλητό και πολυμεταβλητό), έγινε σύγκριση και μεταξύ προβλέψεων διαφορετικής συχνότητας με το ίδιο μοντέλο. Παρατηρήθηκε το γεγονός οτι τα πολυμεταβλητά μοντέλα έιχαν εξαιρετικά καλύτερες επιδόσεις στις ημερήσιες προβλέψεις, συγκριτικά με τις εβδομαδιαίες και μηνιαίες, ενώ τα μονομεταβλητά είχαν καλύτερη επίδοση στις τελευταίες. Το γεγονός αυτό, οφείλεται στο οτι κατά τη διάρκεια μιας εβδομάδας ή ενός μήνα, οι καιρικές συνθήκες μπορεί να είναι άστατες και να έχουν σημαντικές αποκλίσεις μεταξύ τους. Καθώς λοιπόν η επαναδειγματοληψία πραγματοποιείται συγκεντρώνοντας το μέσο όρο των τιμών κάθε ημέρας, ώστε να χωριστεί το σύνολο δεδομένων σε εβδομάδες και μήνες, δημιουργούνται άστατα δεδομένα που παραπλανούν το μοντέλο και οι προβλέψεις γίνονται λιγότερο έγκυρες.

Επιπλέον, παρατηρείται πως τα μοντέλα που προέβλεπαν τιμές για το Παρίσι, το Πεκίνο και το Δελχί, των οποίων η εκπαίδευση έγινε με αρκετά μεγαλύτερο αριθμό δεδομένων, απέφερε στην πλειοψηφία των ρύπων αρκετά καλύτερα αποτελέσματα απο αυτά της Αθήνας. Συνεπώς, επιβεβαιώνεται η σημασία του πλήθους δεδομένων εκπαίδευσης για το ποσοστό ακρίβειας των προβλέψεων των νευρωνικών δικτύων.

Συγκρίνοντας τα αποτελέσματα των δυο διαφορετικών νευρωνικών δικτύων, CNN και LSTM, το τελευταίο είχε εξέχουσα επίδοση, γεγονός που οφείλεται στην μακροπρόθεσμη μνήμη που διαθέτει, όπως επεξηγήθηκε στο τρίτο κεφάλαιο, η οποία επιτρέπει την εκμάθηση ακόμη περισσότερων παραμέτρων. Αυτό το καθιστά ιδιαίτερα ισχυρό για να κάνει προβλέψεις, ιδίως όταν στα δεδομένα υπάρχει μια μακροπρόθεσμη τάση, η οποία παρουσιάζεται στις περισσότερες χρονοσειρές που εξετάζονται στην εργασία αυτή.

6.1.4 Γενική σύγκριση των μοντέλων

Όπως φαίνεται απο την καταγραφή των αποτελεσμάτων όλων των μοντέλων στο πέμπτο κεφάλαιο, παρατηρείται το γεγονός οτι τα νευρωνικά δίκτυα πέτυχαν σημαντικά καλύτερες επιδόσεις απο τις στατιστικές μεθόδους πρόβλεψης χρονοσειρών, με την επίδοση του πολυμεταβλητού LSTM να ξεχωρίζει για τις ημερήσιες προβλεψεις, αλλά και για τις εβδομαδιαίες και μηνιαίες προβλέψεις, όπως διακρίνεται στους πίνακες 5.18, 5.19, 5.20 και 5.21. Η ύπαρξη εποχικότητας στα εβδομαδιαία και μηνιαία δεδομένα επηρεάζει ανοδικά την επίδοση των στατιστικών μεθόδων και των μονομεταβλητών νευρωνικών δικτύων. Επίσης,

ο όγκος των δεδομένων εκπαίδευσης ήταν σημαντικό χαρακτηριστικό για όλα τα μοντέλα, καθώς η εκπαίδευση με μεγαλύτερο πλήθος δεδομένων πετυχαίνει μεγαλύτερη ακρίβεια προβλέψεων και στα στατιστικά μοντέλα αλλά και στα νευρωνικά δίκτυα. Επιπλέον, η ύπαρξη συγκεκριμένης κατεύθυνσης στην τάση των δεδομένων φαίνεται να παίζει σημαντικό ρόλο για μοντέλα που αναπτύχθηκαν. Παρατηρώντας τους πίνακες 5.2, 5.6, 5.10 και 5.14 βλέπουμε πως οι προβλέψεις των χρονοσειρών NO_2 και SO_2 της Αθήνας, των οποίων η τάση δεν παρουσιάζει συγκεκριμένη κατεύθυνση, όπως φαίνεται στον πίνακα 4.4, είναι λιγότερο ακριβείς συγκριτικά με τις προβλέψεις των υπόλοιπων ρύπων.

6.2 Οδηγίες εκτέλεσης των πειραμάτων

Ο κώδικας που αναπτύχθηκε προκειμένου να αναλυθούν τα δεδομένα και να δημιουργηθούν τα μοντέλα πρόβλεψης, έχει αναρτηθεί στην πλατφόρμα GitHub (pbiti/Air-Pollution-Prediction_diploma-thesis). Δίνονται επίσης και τα επεξεργασμένα σύνολα δεδομένων κάθε πόλης, τα οποία περιέχουν συμπτυγμένα τα ατμοσφαιρικά και καιρικά δεδομένα. Τέλος, υπάρχουν κινούμενα αρχεία εικόνας τα οποία παρουσιάζουν την εξέλιξη των πέντε ρύπων κατά το διάστημα καταγραφής των δεδομένων για κάθε πόλη.

6.3 Μελλοντικές επεκτάσεις

Τα πειράματα που διεξήχθησαν απέφεραν αρκετά ικανοποιητικά αποτελέσματα και οι τιμές των ατμοσφαιρικών ρύπων κατάφεραν να προβλεφθούν με ελάχιστα σφάλματα. Ωστόσο, μελλοντικά θα μπορούσαν να γίνουν ορισμένες επεκτάσεις στη μελέτη, όπως αναγράφονται παρακάτω:

1. Δημιουργία υβριδικών μοντέλων που συνδυάζουν τα νευρωνικά δίκτυα και τις στατιστικές μεθόδους, προκειμένου να αξιοποιηθούν τα διαφορετικά οφέλη καθενός.
2. Μελέτη και πρόβλεψη των ατμοσφαιρικών ρύπων για πόλεις που ανήκουν σε ίδιες κλιματικές ζώνες, για παράδειγμα μεσογειακή ζώνη.
3. Επανάληψη των πειραμάτων για διαφορετικά χρονικά βήματα σε κάθε συχνότητα και σύγκριση των επιδόσεων.

Βιβλιογραφία

[1] Serena McCalla Justin Shen, Davesh Valagolam. Prophet forecasting model: a machine learning approach to predict the concentration of air pollutants (PM2.5, PM10, O3, NO2, SO2, CO) in seoul, south korea. 2020.

[2] Ziyuan Ye. Air pollutants prediction in shenzhen based on ARIMA and prophet method. In *2019 International Conference on Building Energy Conservation, Thermal Safety and Environmental Pollution Control (ICBTE 2019)*, Anhui, China, Nov. 2019.

[3] Claudio Guarnaccia, Julia Griselda Ceron Breton, Rosa Maria Ceron Breton, Carmine Tepedino, Joseph Quartieri, , and Nikos E. Mastorakis. Arima models application to air pollution data in monterrey, mexico. In *AIP Conference Proceedings*, Cambridge, UK, Feb. 2018. AIP Conference Proceedings.

[4] Dewen Seng, Qiyan Zhang, Xuefeng Zhang, Guangsen Chen, and Xiyuan Chen. Spatiotemporal prediction of air quality based on lstm neural network. In Changlong Wang Giulio Lorenzini, editor, *Cleaner Energy and Greener Environment: Problems, Solutions and Applications*, volume 60 of *Alexandria Engineering Journal*, pages 2021–2032. Science Direct, 2021. doi:10.1016/j.aej.2020.12.009.

[5] Arif Selcuk Ogrenci Aysenur Gilik and Atilla Ozmen. Air quality prediction using CNN and LSTM - based hybrid deep learning architecture. 2022.

[6] Georgios Karampelas. Analysis and prediction of air pollution in multiple countries using biLSTM-conv1D neural networks. Master's thesis, University of Piraeus - Department of Informatics, Nov. 2021.

[7] A. Bekkar, B. Hssina, and S. Douzi. Air-pollution prediction in smart city, deep learning approach. Dec. 2021.

[8] Ayşenur Özen. Seasonality analysis and forecast in time series. Feb. 2021.

[9] Alexandre Zajic. Introduction to AIC — akaike information criterion. Dec. 2019.

[10] Selva Prabhakaran. Arima model – complete guide to time series forecasting in python.

Aug. 2021.

[11] Air quality index (aqi) basics. https://www.airnow.gov/aqi/aqi-basics/.

[12] Ατμοσφαιρική Ρύπανση. https://www.eea.europa.eu/el/themes/air/intro.

[13] Η ρύπανση του αέρα. http://ebooks.edu.gr/ebooks/v/html/8547/2206/Chimeia_B-Gymnasiou_html-empl/index3_4.html.

[14] Time series analysis with facebook prophet: How it works and how to use it. https://towardsdatascience.com/time-series-analysis-with-facebook-prophet-how-it-works-and-how-to-use-it-f15ecf2c0e3a.

[15] Ν. Ζάχος Αλέξανδρος. Το νευρωνικό δίκτυο LSTM ως μοντέλο βροχής απορροής. Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών, Εργαστήριο Υπολογιστικών Συστημάτων, Mar. 2021.

[16] Andreas Jacobsen Lepperod. Air quality prediction with machine learning. Master's thesis, Norwegian University of Science and Technology, Faculty of Information Echnology and Electrical Engineering, Department of Computer Science, Jun. 2019.