The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-M-6-2025
ISPRS, EARSeL & DGPF Joint Istanbul Workshop "Topographic Mapping from Space" dedicated to Dr. Karsten Jacobsen's 80th Birthday
29–31 January 2025, Istanbul, Türkiye

Estimation of Ground-Level NO_2 Concentrations Over Megacities Using Sentinel-5P and Machine Learning Models: A Case Study of Istanbul

Nur Yagmur Aydin [1], Ilyas Aydin [1]

[1] GTU, Engineering Faculty, Dept. of Geomatics Engineering 41400, Kocaeli, Türkiye – nyagmur@gtu.edu.tr, ilyasaydin@gtu.edu.tr

Keywords: Air Quality Regression, NO_2 Concentration, Sentinel-5P, GEE, Machine Learning.

Abstract

Air pollution is a serious issue in terms of public and environmental health. In this regard, it is important to determine its compliance with the standards by continuous monitoring. For this purpose, air quality ground monitoring stations have been established as part of monitoring systems. While these stations provide highly accurate data, they are point-based and costly. Satellite data, which provides a wide coverage area, enables local and global analysis while providing data with low spatial resolution. Integration of ground and satellite data using machine learning (ML) algorithms enables more accurate regional analysis. For this purpose, estimation analysis of the NO_2 parameter, which is the most measured parameter at ground monitoring stations and has a major impact on its formation by human activities, was conducted for the Istanbul megacity using freely accessible Sentinel-5P satellite data. The performance of three ML algorithms, namely multi-layer perceptron (MLP), support vector regression (SVR), and XGBoost regression (XGB), in estimating the ground level-NO_2 parameter was evaluated both quantitatively using RMSE and MAE accuracy metrics and qualitatively by visual analysis. The model was trained with data covering the years 2019-2022, validated with data for 2023, and tested with data for 2024. According to the results obtained, while the three models gave similar results with RMSE values of 19.59, 19.65, and 20.03 $\mu g/m^3$ and MAE values of 15.00, 14.34, and 15.90 $\mu g/m^3$ in the test data, SVR and MLP models provided higher accuracy in the seasonal assessment. In the visual assessment, the SVR model results provide a more accurate approach.

1. Introduction

Increasing population rates lead to higher consumption levels and accelerate raw material production. As populations concentrate in specific areas, metropolitan cities emerge, and communities within these limited regions contribute significantly to air pollution due to their production and consumption activities. Factors such as logistics, industrialization, housing, and heating all play a role in exacerbating air pollution. Identifying and managing the pollution generated by these factors are crucial for human health.

There are various pollutants such as nitrogen dioxide (NO_2), sulphur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), and particular matter ($PM_{2,5}$ and PM_{10}). Daily exposure of these pollutants emitted by anthropogenic activities such as transportation and industrial activities and natural processes such as wildfires and lightning causes health issues that may lead to death (Semlali, 2019). Among these pollutants, the concentration of NO_2 gas in the air is a commonly used indicator for assessing air quality and pollution levels. Combustion of fossil fuels and biomass burning are the main sources of NO_2 pollutant in megacities (Zhang et al., 2003). Accumulation of high concentrations of NO_2 around emission sources (Chi et al., 2022), results in long-term exposure to NO_2. This leads to respiratory diseases (Manisalidis et al., 2020), and NO_2 contributes to the creation of secondary pollutants like O_3 and aerosol nitrates, which can result in acid rain and decreased visibility (Seinfeld and Pandis, 2016). The World Health Organization (WHO) has stated that following the guidelines for NO_2 could prevent many deaths caused by air pollution (Song et al., 2023). To ensure health-safe living environments, this parameter should be continuously monitored, and any changes in its concentration should be identified.

Air quality in urban sites has been monitored with air quality monitoring stations which are among the main sources for collecting data for this purpose (Cedeno Jimenez and Brovelli, 2023). However, the need for periodic calibration of ground monitoring stations and the fact that they are affected by environmental changes constitute the disadvantages of these data sources. The cost of installation, maintenance, and calibration of ground stations negatively affects the interest in air quality analysis for local governments. These reasons have triggered research on alternative sources for monitoring air quality. In this process, satellite-based air quality analyses have gained importance, but the detection capabilities of these sources in the troposphere layer pose difficulties in observing the interaction with ground-based activities. Unlike point-based ground stations, satellite sources enable regional analysis due to their wide coverage and repeated data acquisition from the same area. Additionally, they provide information about regions where ground monitoring stations have not been established. However, since accuracy analysis cannot be performed in these regions, studies have been conducted to enhance data reliability by models integrating ground station measurements and satellite sources. Multiple studies have established models that integrate the Sentinel-5P TROPOMI dataset with ground monitoring stations using machine learning (ML) algorithms for various regions, including China (Chi et al., 2022), East Asia (Kang et al., 2021), and Europe (Shetty et al., 2024). Although studies have examined the distribution of NO_2 over Istanbul province (Kaplan et al., 2019; Makineci, 2022; Cavdaroglu and Arik, 2023), no study has been identified that specifically integrates satellite data with ground monitoring data using machine learning algorithms for Istanbul.

In this study, the potential of ML regression models developed with freely available Sentinel-5P satellite data and measurements of air quality ground monitoring stations was investigated for ground-level NO_2 estimation over Istanbul.

With this aim, the models were developed using data collected between the years 2018 and 2023 and were tested for 2024. The performance of three different ML algorithms was compared using accuracy metrics such as root mean square error (RMSE) and mean absolute error (MAE). Quantitative and qualitative assessments of the results obtained were also seasonally evaluated.

2. Study Area and Materials

Istanbul, a metropolitan city in Turkey with a population of approximately 15.6 million, was used as the study area. The air pollution in this city, characterized by frequent traffic congestion and industrial zones, is continuously monitored by 40 ground-based air quality monitoring stations. For this study, 32 stations that consistently measure NO_2 levels were included in the estimation process. The hourly measurements of ground monitoring stations were collected within the scope of the Air Quality Monitoring Project carried out under the administration of Istanbul Metropolitan Municipality and serve on the web portal (https://havakalitesi.ibb.gov.tr/).

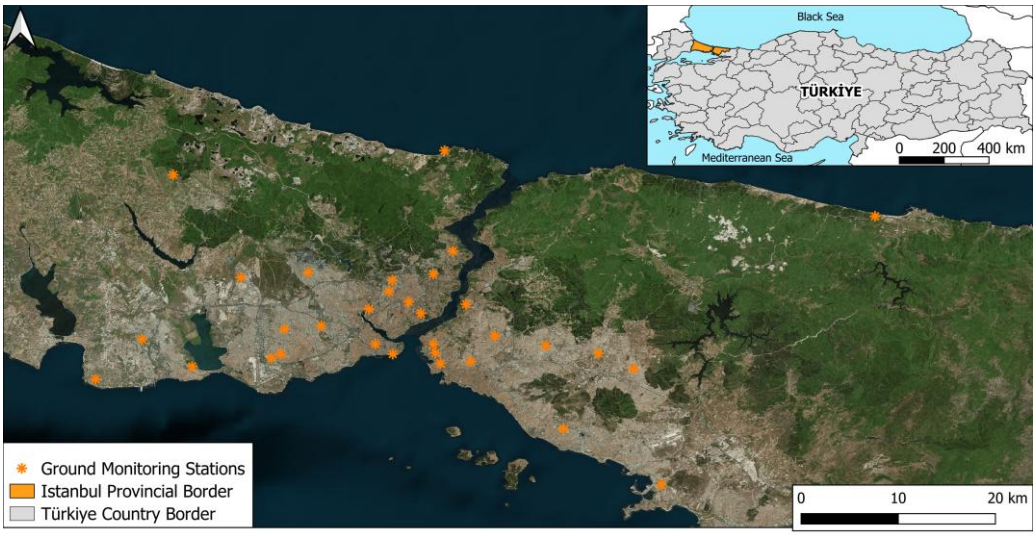

Figure 1. Study area region with the ground air quality monitoring stations on Google Earth image.

NO_2 concentration estimation was realized using not only ground monitoring stations but also the Sentinel-5P TROPOMI dataset. Sentinel-5P satellite carrying the TROPOMI instrument was developed in collaboration with the European Space Agency (ESA) and The Netherlands with the aim of performing atmospheric measurements with high spatio-temporal resolution. Sentinel-5P data has 5.0×3.5 km spatial resolution in the satellite flight direction and the perpendicular direction at nadir. In the study, ground monitoring stations were matched with the corresponding pixels of the Sentinel-5P datasets on the same date. The Sentinel-5P datasets were extracted by using the Google Earth Engine cloud computing platform which stores archive data and enables to collect and analyze satellite images and spatial datasets.

The Sentinel-5P satellite images and ground-based observations for the years 2018-2024 were collected. While the values corresponding to each station in the satellite pixel were taken as input data, the output data was determined by the hourly station data corresponding to the satellite passing hours. Additionally, station values lower than 1 µg/m³ and higher than 300 µg/m³ were removed to eliminate their misleading impact on the model. After this data preparation step, data was divided into 3 parts: training, validation, and test. The data collected from 2018 to 2022 serves as the training data, the data from 2023 is used for validation, and the data from 2024 is used for testing. The statistical specifications of the data are given in Table 1.

Data	Min (µg/m³)	Max (µg/m³)	Mean (µg/m³)	STD (µg/m³)
Training	1.05	234.10	34.14	27.25
Validation	1.05	253.00	31.28	23.34
Test	1.30	257.05	29.59	21.81

Table 1. The statistical properties of the ground-based data.

The training dataset varies from 1.05 to 234 µg/m³, while the validation dataset is between 1.05 to 253 µg/m³, and the test dataset is between 1.30 to 257 µg/m³. The standard deviations of datasets are 27.25, 23.34, and 21.81 µg/m³, respectively.

3. Methodology

In the study, the performance of ML models in the estimation of NO_2 was evaluated using Sentinel-5P TROPOMI and ground monitoring stations. The flowchart of the study is given in Fig. 2.

Firstly, the data preparation step was realized by integrating and filtering of satellite and ground station data. After this step, data was divided into three parts: training, validation, and test. Training and validation datasets were used for hyperparameters optimization and estimation of ML models. The hyperparameter tuning step of each ML model was processed with the *GridSearch-CV* algorithm.

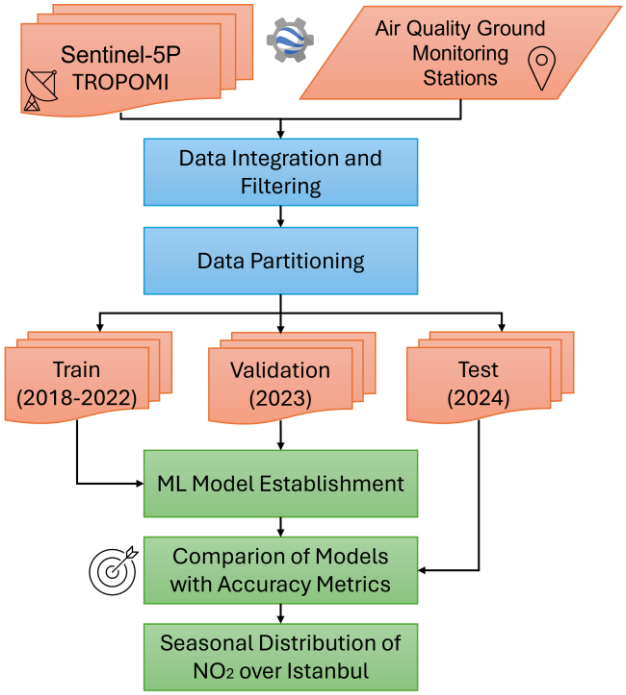

Figure 2. The workflow of the study.

Multiple models used for regression purposes produce different prediction results depending on their operating principles. In the NO_2 estimation process step, multiple ML regression models were conducted to obtain more accurate results. For this purpose, these models can be listed as Multi-Layer Perceptron (MLP) regression, Support Vector Regression (SVR), and eXtreme Gradient Boosting (XGB) regression, respectively.
MLP is a neural network model which consists of a number of neurons connected between layers by weight and bias (Atkinson and Tatnall, 1997). MLP approach is realized by performing non-parametric regression analysis and contains input layer, hidden layer/s and output layer. It is superior to single-layer perceptron due to its ability to learn linear and nonlinear relationships between input and output data and increase computational efficiency (Madhiarasan and Deepa, 2017).

SVR, developed by Smola and Schölkopf (2004), is an algorithm used for regression tasks and is successful in drawing highly effective conclusions from complex datasets with the help of support vectors. In nonlinear datasets, it can solve various regression problems by utilizing kernel structures like the Radial Basis Function (RBF). This model structure offers users the flexibility to define key parameters, including the regularization parameter (C) and the error sensitivity parameter (ε), facilitating performance optimization through hyperparameter tuning.

XGB algorithm proposed by Chen and Guestrin (2016) relies on a gradient boosting algorithm. XGB iteratively creates new trees that are used to fit residuals (differences of predictions and actual values) of the previous trees.

All aforementioned ML models have been widely used in various research areas; however, their performance and accuracy can vary based on the data structure. To determine the accuracy of the models, two accuracy metrics were used in the study. These are RMSE and MAE and their equations are given in Eq. 1 and 2.

$$\mathrm{RMSE} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-1}} \quad (1)$$

$$\mathrm{MAE} = \frac{\sum|y_i - \hat{y}_i|}{n} \quad (2)$$

where y_i is the actual measured value and $\hat{y}_i$ is the estimated value.

4. Results

Using the parameters determined through hyperparameter optimization, the models were trained using training and validation datasets and tested with test data that models had not seen before. To assess model accuracy, RMSE and MAE accuracy metrics were computed using the estimated values and measured values, and results are given in Table 2.

Model	Training (2018-2022)		Validation (2023)		Test (2024)	
[µg/m^3]	RMSE	MAE	RMSE	MAE	RMSE	MAE
MLP	26.83	21.03	21.53	16.07	19.59	15.00
SVR	25.34	17.70	21.38	15.23	19.65	14.34
XGB	24.55	18.42	22.05	17.11	20.03	15.90

Table 2. Accuracy assessment results of ML models.

The results obtained showed that the prediction values are within the acceptable value ranges and all three models performed successfully. In the training part, the model accuracies can be ranked from highest to lowest as follows: XGB – SVR – MLP. In the validation part, the order changed as SVR – MLP – XGB. However, in the test part, the model accuracies can be ranked as MLP – SVR – XGB. The RMSE values being lower than the standard deviation (21.81µg/m^3) indicate that models performed well. Since the concentration of the NO_2 parameter varies depending on the seasons, accuracy assessment was performed on a seasonal basis on the test data. The RMSE and MAE values obtained are given in Fig. 3.

According to Fig. 3, it is observed that the error values decreased in the seasonal evaluation in all ML models considering Table 1. In the RMSE values obtained in Fig. 3, the SVR model performed well in all seasons except winter. In the winter season, MLP performed relatively better than SVR. Considering MAE values, the SVR model performed well in all seasons. In the winter season, MLP has lower RMSE value than SVR, although SVR has lower MAE value than MLP. This indicates that the residuals between measured and estimated values obtained in the SVR model are higher in the winter season. The XGB model has the highest RMSE and MAE values compared with other models in all seasons. In general, the lowest error values were observed in the summer and spring seasons, while the highest error values were observed in the autumn and winter seasons, respectively. A possible reason is that NO_2 concentration is high density in winter and low density in summer (Shen et al., 2021).

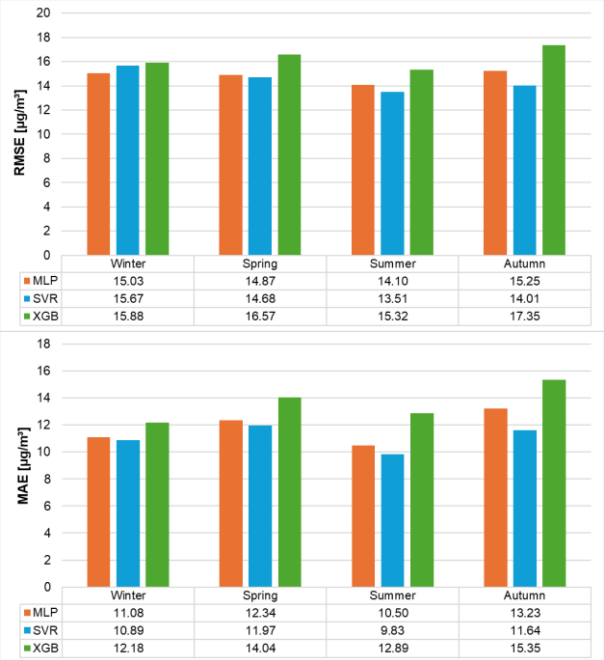

Figure 3. Accuracy assessment results on a seasonal basis.

The quantitative evaluation may not be sufficient to determine the most appropriate model. That is why, qualitative analysis was realized by visual interpretation of the results obtained. To do this, seasonal NO_2 estimation maps for Istanbul were produced for all models and are given in Fig. 4.

According to Figure 4, the spatial distribution of NO_2 in the MLP and SVR models was found to be better in all seasons compared to the XGB model. The result obtained in the XGB model showed the spatial distribution of NO_2 as clustered. In all model results, NO_2 distribution is high in winter, decreases in spring and shows the lowest distribution in summer. It increases again in autumn. Heating activities during the winter in urban areas can be shown as the main reason for this situation (Vîrghileanu et al., 2020; McDuffie et al., 2020; Morillas et al., 2024).

The other main source of NO_2 is road transport which significantly affects its distribution over the region (McDuffie et al., 2020). In light of the results obtained, it is observed that the NO_2 distribution is concentrated around the Bosphorus. It is noteworthy that the regions where the Bosphorus connects to the Sea of Marmara have the highest NO_2 concentration in all four seasons. These regions are areas with dense settlements and traffic density. When the model results are examined, although the models visually show similar harmony in seasonal transitions, the SVR model, which has the lowest RMSE and MAE values in the accuracy analysis, shows more realistic results. Although both models give successful results in general, in the MLP model results, while the NO_2 value range of the seasons has close values, the density difference between the seasons can be clearly seen in the SVR model.

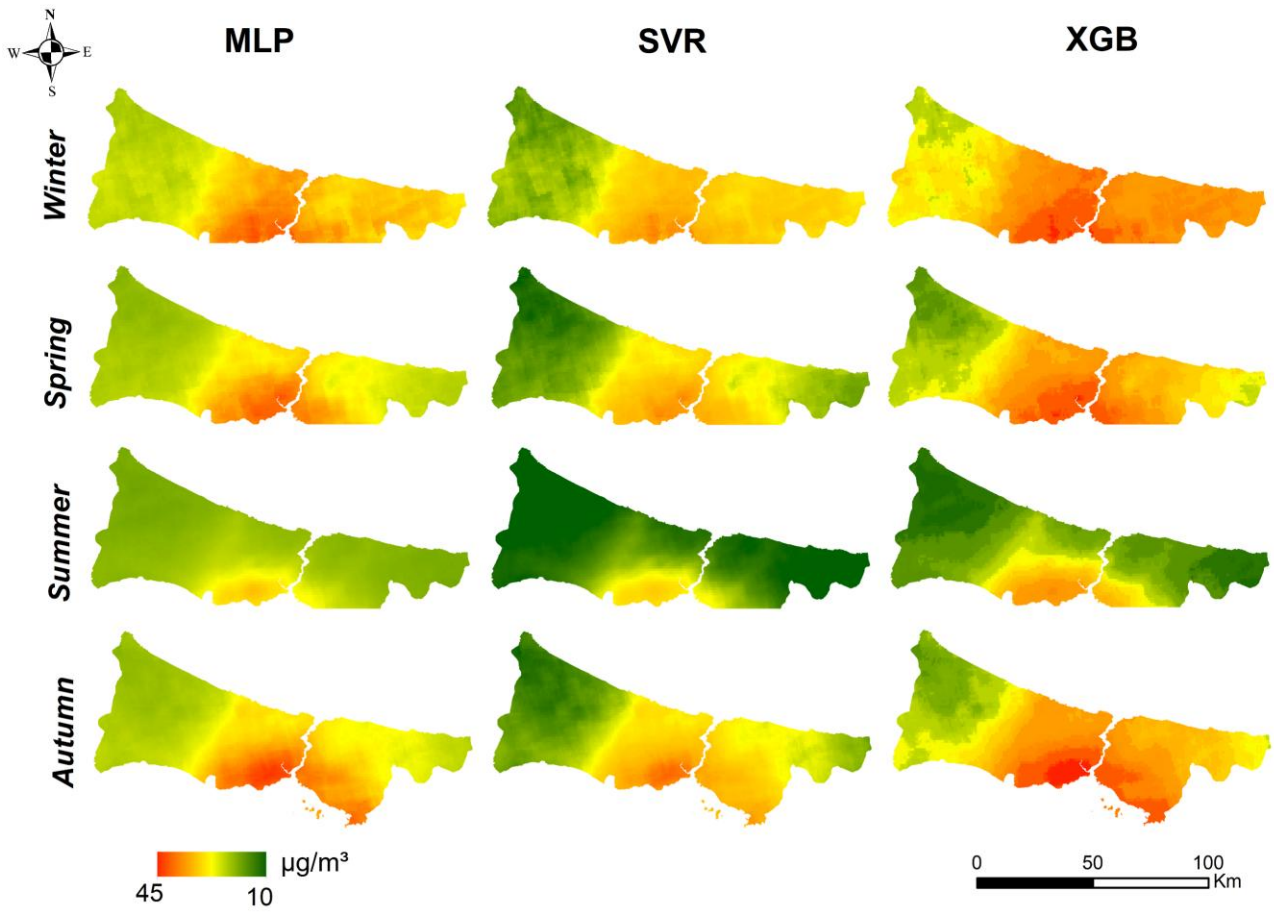

Figure 4. Seasonal distribution of NO_2 in Istanbul for 2024 according to the ML models.

5. Conclusions

This study demonstrates the applicability of ML models to estimate ground-level NO_2 over the Istanbul megacity. For this purpose, three ML algorithms, namely MLP, SVR, and XGB, were implemented to Sentinel-5P satellite data and ground air quality monitoring stations and compared both quantitatively and qualitatively. Additionally, the performance of the models according to seasons was also examined.

In the study, it is revealed that ML regression models yield successful results in estimating ground-level NO_2 using freely available satellite-based data sources, such as Sentinel-5P, which has significant potential for air quality monitoring. It is observed that higher accuracy performance was achieved with lower error values specific to the seasons. Seasonal changes were also successfully detected from the results obtained. In the overall evaluation, the SVR model showed relatively better results compared to the MLP model, while the XGB model could not visually demonstrate the expected performance.

The effect of environmental factors on model performance was not evaluated in the analysis. Therefore, future studies will aim to enhance model performance by enriching its feature space with environmental factors (e.g., land use/land cover, population), meteorological variables (e.g., air temperature), and topographic attributes (e.g., surface elevation). In addition, the accuracy of the models will be improved by developing season-specific models.

Acknowledgements

The authors express their gratitude to the Istanbul Metropolitan Municipality for providing the air quality ground monitoring station data used in this study.

References

Atkinson, P.M., Tatnall, A.R.L., 1997. Neural networks in remote sensing-Introduction. *International Journal of Remote Sensing*, 18(4), 699-709.

Cavdaroglu, G.Ç., Arik, A.O., 2023. Spatial-Temporary Analysis of Istanbul Air Pollution During the Pandemic using Google Earth Engine and Google Community Mobility Reports. *Veri Bilimi*, 6(1), 1-14.

Cedeno Jimenez, J.R., Brovelli, M.A., 2023. NO_2 Concentration Estimation at Urban Ground Level by Integrating Sentinel 5P Data and ERA5 Using Machine Learning: The Milan (Italy) Case Study. *Remote Sensing*, 15(22), 5400. doi.org/10.3390/rs15225400.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. doi.org/10.1145/2939672.2939785.

Chi, Y., Fan, M., Zhao, C., Yang, Y., Fan, H., Yang, X., Yang, J., Tao, J., 2022. Machine learning-based estimation of ground-level NO_2 concentrations over China. *Science of The Total Environment*, 807, 150721. doi.org/10.1016/j.scitotenv.2021.150721.

Kang, Y., Choi, H., Im, J., Park, S., Shin, M., Song, C.K., Kim, S., 2021. Estimation of surface-level NO_2 and O_3 concentrations using TROPOMI data and machine learning over East Asia. *Environmental Pollution*, 288, 117711. doi.org/10.1016/j.envpol.2021.117711.

Kaplan, G., Avdan, Z.Y., Avdan, U., 2019. Spaceborne nitrogen dioxide observations from the Sentinel-5P TROPOMI over Turkey. *Proceedings*, 18(1), 4. doi.org/10.3390/ECRS-3-06181.

Madhiarasan, M., Deepa, S.N., 2017. Comparative analysis on hidden neurons estimation in multi layer perceptron neural networks for wind speed forecasting. *Artificial Intelligence Review*, 48, 449-471. doi.org/10.1007/s10462-016-9506-6.

Makineci, H.B., 2022. Investigation of NO_2 and CO Emissions in Istanbul Province Central Districts using Remote Sensing and Terrestrial Station Data. *Turkish Journal of Remote Sensing*, 4(2), 62-74. (in Turkish) doi.org/10.51489/tuzal.1160333.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E., 2020. Environmental and health impacts of air pollution: a review. *Frontiers in Public Health*, 8, 14. doi.org/10.3389/fpubh.2020.00014.

McDuffie, E.E., Smith, S.J., O'Rourke, P., Tibrewal, K., Venkataraman, C., Marais, E.A., Zheng, B., Crippa, M., Brauer, M., Martin, R.V., 2020. A global anthropogenic emission inventory of atmospheric pollutants from sector-and fuel-specific sources (1970–2017): an application of the Community Emissions Data System (CEDS). *Earth System Science Data*, 12(4), 3413-3442. doi.org/10.5194/essd-12-3413-2020.

Morillas, C., Alvarez, S., Serio, C., Masiello, G., Martinez, S., 2024. TROPOMI NO_2 Sentinel-5P data in the Community of Madrid: A detailed consistency analysis with in situ surface observations. *Remote Sensing Applications: Society and Environment*, 33, 101083. doi.org/10.1016/j.rsase.2023.101083.

Seinfeld, J.H., Pandis, S.N., 2016. *Atmospheric chemistry and physics: from air pollution to climate change*. Wiley, New York.

Semlali, B.B., 2019. Towards remote sensing datasets collection and processing. *International Journal of Embedded and Real-Time Communication Systems (IJERTCS)*, 10(3), 49-67.

Shen, Y., Jiang, F., Feng, S., Zheng, Y., Cai, Z., Lyu, X., 2021. Impact of weather and emission changes on NO_2 concentrations in China during 2014–2019. *Environmental Pollution*, 269, 116163. doi.org/10.1016/j.envpol.2020.116163.

Shetty, S., Schneider, P., Stebel, K., Hamer, P.D., Kylling, A., Berntsen, T.K., 2024. Estimating surface NO_2 concentrations over Europe using Sentinel-5P TROPOMI observations and Machine Learning. *Remote Sensing of Environment*, 312, 114321. doi.org/10.1016/j.rse.2024.114321.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. *Statistics and Computing*, 14, 199-222. doi.org/10.1023/b:stco.0000035301.49549.88.

Song, J., Wang, Y., Zhang, Q., Qin, W., Pan, R., Yi, W., Xu, Z., Cheng, J., Su, H., 2023. Premature mortality attributable to NO2 exposure in cities and the role of built environment: A

global analysis. *Science of the Total Environment*, 866, 161395. doi.org/10.1016/j.scitotenv.2023.161395.

Vapnik, V., Golowich, S., Smola, A., 1997. Support vector method for function approximation, regression estimation and signal processing. *Advances in Neural Information Processing Systems*, 281-287.

Vîrghileanu, M., Săvulescu, I., Mihai, B.A., Nistor, C., Dobre, R., 2020. Nitrogen Dioxide (NO_2) Pollution monitoring with Sentinel-5P satellite imagery over Europe during the coronavirus pandemic outbreak. *Remote Sensing*, 12(21), 3575. doi.org/10.3390/rs12213575.

Zhang, R., Tie, X., Bond, D.W., 2003. Impacts of anthropogenic and natural NOx sources over the US on tropospheric chemistry. *Proceedings of the National Academy of Sciences*, 100(4), 1505-1509. doi.org/10.1073/pnas.252763799.