International Journal of
Geo-Information

MDPI

Article

Estimating Daily NO_2 Ground Level Concentrations Using Sentinel-5P and Ground Sensor Meteorological Measurements

Jesus Rodrigo Cedeno Jimenez *, Angelly de Jesus Pugliese Viloria and Maria Antonia Brovelli

Department of Civil and Environmental Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
* Correspondence: jesusrodrigo.cedeno@polimi.it

Abstract: Environmental and health deterioration due to the increasing presence of air pollutants is a pressing topic for governments and organizations. Institutions such as the European Environment Agency have determined that more than 350,000 premature deaths can be attributed to atmospheric pollutants. The measurement of trace gas atmospheric concentrations is key for environmental agencies to fight against the decreased deterioration of air quality. NO_2, which is one of the most harmful pollutants, has the potential to cause diseases such as Chronic Obstructive Pulmonary Disease (COPD). Unfortunately, not all countries have local atmospheric pollutant monitoring networks to perform ground measurements (especially Low- and Middle-Income Countries). Although some alternatives, such as satellite technologies, provide a good approximation for tropospheric NO_2, these do not measure concentrations at the ground level. In this work, we aim to provide an alternative to ground sensor measurements. We used a combination of ground meteorological measurements with satellite Sentinel-5P observations to estimate ground NO_2. For this task, we used state-of-the-art Machine Learning models, linear regression models, and feature selection algorithms. From the results obtained, we found that a Multi-layer Perceptron Regressor and Kriging in combination with a Random Forest feature selection algorithm achieved the lowest RMSE (2.89 $\mu g/m^3$). This result, in comparison with the real data standard deviation and the models using only satellite data, represented an RMSE decrease of 55%. Future work will focus on replacing the use of meteorological ground sensors with only satellite-based data.

Keywords: atmospheric pollution; nitrogen dioxide; machine learning

check for updates

Citation: Cedeno Jimenez, J.R.; Pugliese Viloria, A.d.J.; Brovelli, M.A. Estimating Daily NO_2 Ground Level Concentrations Using Sentinel-5P and Ground Sensor Meteorological Measurements. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 107. https://doi.org/10.3390/ijgi12030107

Academic Editors: Godwin Yeboah and Wolfgang Kainz

Received: 15 December 2022
Revised: 20 February 2023
Accepted: 2 March 2023
Published: 4 March 2023

1. Introduction

Atmospheric pollution is a pressing topic for governments and organizations due to its negative impact on human health and on the environment. European regulatory authorities reported that atmospheric pollution was responsible for over 350,000 premature deaths in 2019 [1]. To encourage countries to make strict regulations on improving air quality, the United Nations (UN) Sustainable Development Goals (SDGs) address the reduction in atmospheric pollution. In particular, SDGs 3, 7 and 11 focus on clean energy production, people's well-being and making cities safer (healthier) [2]. However, according to Ref. [3], about 90% of the World's population lives in places where air quality does not comply with the UN pollution recommended limits. A key factor in improving this situation is the monitoring and reporting of air quality measurements [3].

Respiratory infections and diseases such as Chronic Obstructive Pulmonary Disease (COPD) are the fourth leading cause of death worldwide and can be linked to Nitrogen Dioxide (NO_2) [4]. Additionally, NO_2 contributes to the process of eutrophication and acidification of environmental components by acting as a precursor for ozone (O_3) and particulate matter (PM) [5]. Measurements in highly developed regions of the world, such as the USA or Europe have specified that the primary sources of NO_2 emissions are strictly related to human transportation, i.e., vehicles using combustion engines and stationary fuel combustion [6,7]. In fact, Ref. [8] reveals how the progressive reduction in population mobility during the COVID-19 period in the city of Milan (Italy) contributes to the reduction in NO_2 concentration. This occurred during the COVID-19 lockdown in 2020, when both transportation and supply chain emissions were significantly reduced. By the end of March, a reduction of 77% in private transportation, 66% in light and heavy-duty vehicles, 39% in combustion in manufacturing factories, and 20% in production processes led to a NO_2 reduction of one-third [8].

With the aim of controlling atmospheric pollutant levels in the air, some agencies, such as the European Environment Agency (EEA), established air quality directives. These directives establish thresholds that countries must not surpass on a yearly, daily and hourly basis [9]. To measure air quality, new technologies have been developed in the past decades. As an example, the Copernicus programme was developed by the European Commission and the European Space Agency (ESA). This programme utilises satellites and in-situ (non-satellite) data to measure air quality. These data provide European citizens and authorities with information about the condition of the atmosphere. The central appeal of the Copernicus programme is to use satellite technology to provide data about the Earth's atmosphere in regions that it would otherwise not be possible to obtain [10]. Among its satellite catalogue, Copernicus measures tropospheric NO_2 concentrations.

According to the European Union guidelines for the measurement of atmospheric NO_2, local authorities must provide air quality measurements through the use of certified ground stations [11]. These guidelines specify that stations must measure air pollutants at a height between 2 and 10 m from the ground. The main drawback of using ground stations as the source of measurements is twofold. The first is that ground sensors are only able to measure point data at the location where they are installed. The second is that the cost of installation, maintenance and operation requires both technical expertise and financial investment, which is not always available. This is a common case in low and middle-income regions such as Africa and Asia [3]. These constraints reduce the area of measurements, limiting the understanding of the physicochemical phenomena taking place. However, satellite technology has enabled scientists to overcome this limitation. Such is the case of the Copernicus satellite Sentinel-5P, a measurement tool equipped with state-of-the-art technology capable of capturing the presence of tropospheric trace gases. The main working unit of this satellite is the TROPOspheric Measurement Instrument (TROPOMI). TROPOMI is a sensor developed by the Netherlands Space Office and the ESA, which is able to measure NO_2, Carbon Monoxide (CO), aerosols and other atmospheric pollutants [12].

Even though the Sentinel-5P can be an initial approach to measuring atmospheric NO_2, this technology still has some disadvantages. The first is that this satellite provides measurements at the tropospheric level. This means that the satellite measures the total column of gas (i.e., the total concentration of a pollutant from the ground up to 10 km above ground level). Although the troposphere is the closest atmospheric layer to the Earth's surface, it does not represent the conditions that exist at the ground level (2 m to 10 m high). The second is regarding the spatial resolution, which is 3.5 × 5 km; this measurement is not adequate for the previously mentioned directives established by regional authorities. Another challenge when monitoring pollution using satellite technology is related to the measure itself. While satellites provide the number of NO_2 particles in an area (pixel), the ground stations provide a volume of the particles present in an air sample. Taking these characteristics into consideration, providing an accurate estimate of air quality a few meters from the ground can be challenging.

In this work, we focused on the creation of a model for estimating ground atmospheric NO_2 as an alternative to ground stations. According to Pinder et al., 2019, [13], air quality monitoring has been conducted for the past 50 years in high-income economies. The same study claims that this was made possible by spending millions of dollars to establish, operate, and maintain monitoring stations. As mentioned previously, this is not the situation in Low- and Middle-Income Countries (LMICs), where air quality monitoring is subpar or nonexistent. The situation is different for weather stations, which, according to the World Meteorological Organization (WMO), are monitored globally through a network of more than 8000+ stations [14]. Figure 1 shows the locations of these stations and whether they belong to an LMIC or a high-income country. Different from NO_2 sensors (Figure 2), meteorological ground sensors are present all over the world (including LMICs). This opens up the opportunity for LMICs to use a combination of Sentinel-5P NO_2 and meteorological indicators to perform an accurate estimation of the concentration volume of NO_2 at the ground level.

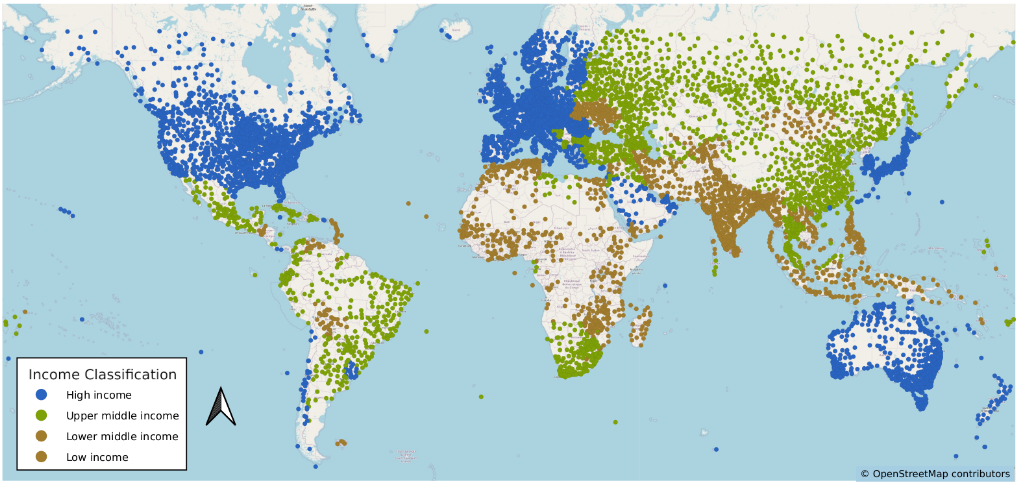


Figure 1. Weather stations registered by the WMO. The station's colour represents the countries' income classification (low, medium and high income) [14].

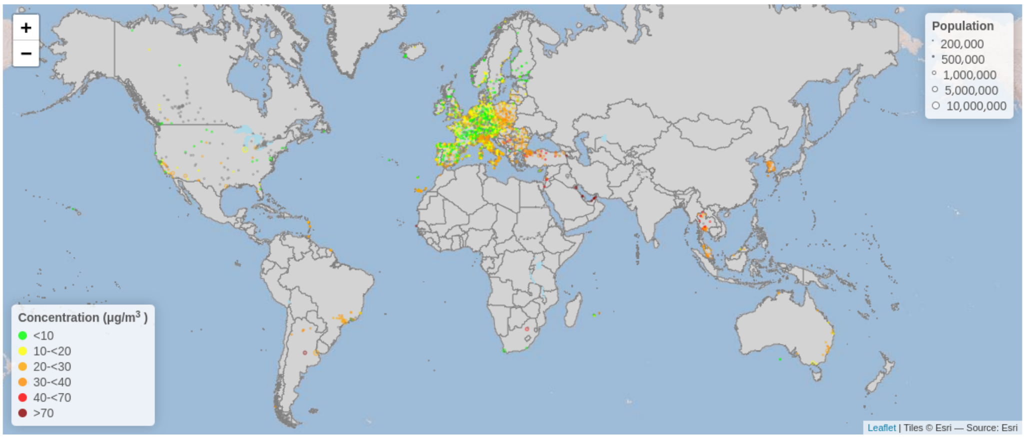


Figure 2. Global NO_2 stations registered by the World Health Organization (WHO). Global Air Quality Database App: app for exploring air quality in countries. WHO Global Air Quality Database (update 2018) edition. Version 1.0. Geneva, World Health Organization, 2018.

Recent studies have estimated ground level NO_2 concentrations based on satellite NO_2 observations by means of Machine Learning (ML) and linear models. Random Forest is among the most-used methods to model ground NO_2 [15–17]. Other ML methods used by authors include Decision Trees, Extremely Randomised Trees, Gradient Boost Decision Trees, Extreme Random Forest, XGboost, and Bayesian Maximum Entropy [15,18–20]. Linear regression and polynomial fitted regression are also implemented [21], as well as universal Kriging [22]. Several studies consider Sentinel-5P TROPOMI NO_2 measurements [4,15], which provide tropospheric NO_2 columns at the highest available resolution [23]; nevertheless, other studies consider the Aura Ozone Monitoring Instrument (OMI) NO_2 measurements as well [16,22,24], being both state-of-the-art satellite instruments [23]. The satellite NO_2 measurements are usually integrated with ground NO_2 concentrations derived from local air quality monitoring stations [24,25] or from authoritative datasets, e.g., the AirBase dataset from the EEA [26]. The auxiliary data (i.e., independent variables of the model) include mainly meteorological variables such as temperature, wind, and humidity, which are retrieved from both ground stations [19] and authoritative datasets such as the European Centre for Medium-Range Weather Forecasts (ECMWF) fifth-generation atmospheric reanalysis product (ERA5) [4,17], and ECMWF Modern-Era Retrospective Analysis Research and Application, version 2 (MERRA-2) [24]. Other auxiliary data include road and population density, terrain height, and land use/land cover [26]. As a case study for this work, we proposed a single daily estimate of ground NO_2 concentrations based on Sentinel-5P tropospheric NO_2 and ground meteorological data by means of a set of machine learning, linear regression, and feature selection models. The study is in line with the current state of the art, obtaining a better Root Mean Squared Error (RMSE) than similar studies.

The area of interest is focused in Lombardy, located in the North of Italy which is considered a pollution hotspot due to the physical features of this area (Figure 3). Specific topographic components such as the Alps' in the north and west of the region and the Apennines' presence in the south cause the wind to be entrapped in the Po Valley. Thermal inversions on a regular basis hinder the proper dispersion of air pollutants [27]. We concentrate on the air quality in the metropolitan city of Milan. This choice was made because Milan is a traffic-intensive and industrial area.

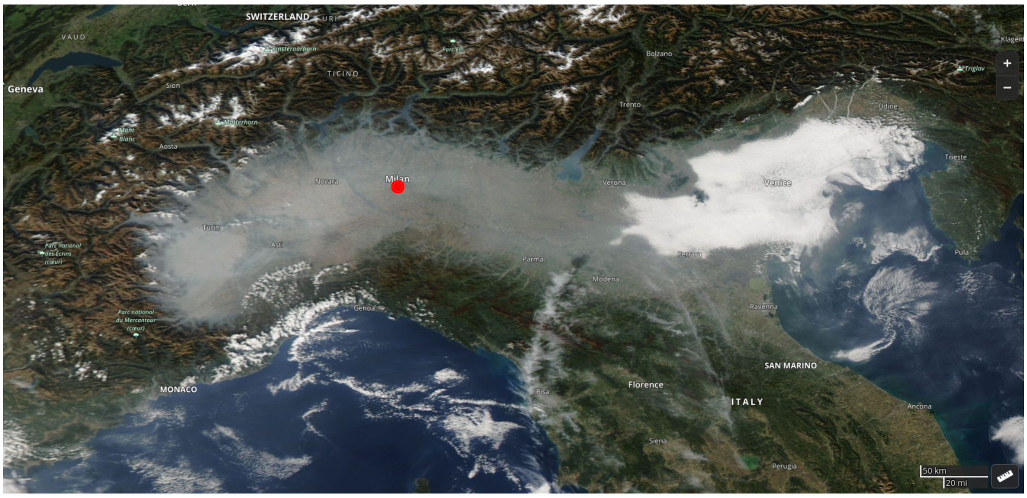


Figure 3. Satellite image (MODIS Aqua radiometer) of the Northern Italy Po Valley showing the aerosol layer entrapped in the area. The location of Milan is indicated with a red circle. Image extracted from https://worldview.earthdata.nasa.gov for the date 18 October 2017.

The project's methodology consists of three phases (data processing, training and testing). For the data processing, we created a Python pipeline that automatically processes, re-grids and registers the data into the data management system (Open Data Cube—ODC). For this work, we integrated ground sensors and satellite data into the ODC, a procedure established in other works [28]. The ODC is an indexing system that acts as an intermediary layer between the user and satellite data. It is commonly used to make analysis-ready data available and eliminate the processing steps for future users. For further details about the ODC, the official website can be reviewed (https://datacube-core.readthedocs.io/ (accessed on 12 September 2022)).

To implement the regression models, we used a selection of Machine Learning algorithms able to learn from real ground stations present in the Milan area. The input for these models was directly obtained from the ODC. The training procedure made use of a dataset of meteorological indicators from ARPA Lombardia (the Lombardy Regional Environment Protection Agency) and Sentinel-5P NO_2 measurements. The output of the models was an estimation of single daily NO_2 ground measurements at the time of passage of the satellite (from 12:00 h to 15:00 h UTC+1) [15,29].

The results indicate that the model's NO_2 estimation is accurate compared to the daily standard deviation and mean of the Metropolitan City of Milan from 12:00 h to 15:00 h UTC+1. The model's output has a lower Root Mean Square Error compared to using only satellite data for the estimation of NO_2 ground data.

The structure of the paper is as follows. In the first section, we describe the data used for the project, as well as the technologies involved. In the second section, we describe the Machine Learning algorithms implemented in this work. In the third section, we present the results. The final section contains a discussion of the results, conclusions and future work.

2. Materials and Methods

This section describes the data used to train and test the ground-air NO_2 estimation model. All the data used in this work focuses on the Metropolitan City of Milan (MCM) for the period of 1 January 2019 to 28 September 2022.

2.1. NO_2 Ground Sensors Data

The Lombardy Regional Environment Protection Agency (ARPA Lombardia) is the entity in charge of monitoring, among other things, the atmospheric pollutants in Lombardy. This is achieved by means of a network of 84 ground stations (Figure 4) compliant with the EEA measurement procedures [11]. The stations are capable of measuring NO_X (NO and NO_2), SO_2, CO, O_3, PM_{10}, $PM_{2.5}$ and benzene with an hourly average temporal resolution. The hourly average corresponds to the average of measurements during the 60 min before the indicated time. Even though the Lombardy region has a total of 84 ground stations for measuring NO_2, only 16 of these are located inside the Metropolitan City of Milan.

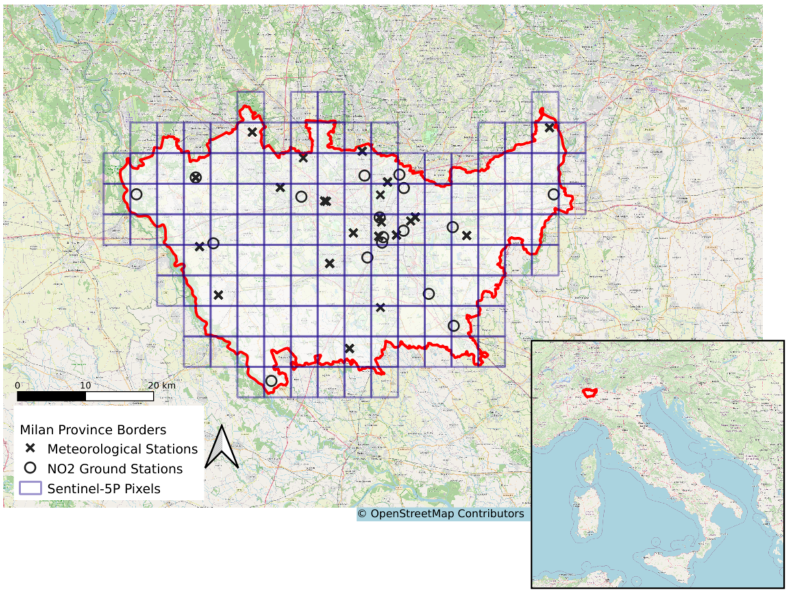


Figure 4. Location of NO_2 and meteorological ground sensors belonging to ARPA Lombardia alongside Sentinel-5P pixel grid [30].

The data are provided with an open license and can be downloaded through the web data portal (https://dati.lombardia.it, (accessed on 30 september 2022)). Measurements of the current year (2022) can also be accessed through an Application Programming Interface (API). Downloading through the API enables the user to access the data according to their needs by choosing only the needed pieces of information and not the dataset containing information outside the area of interest.

2.2. Meteorological Ground Sensors Data

Apart from trace gases, the ARPA Lombardia network also measures meteorological phenomena such as temperature, atmospheric pressure, humidity, precipitation, wind speed and others [31]. The meteorological network in the Metropolitan City of Milan is composed of 98 ground sensors at 25 different stations (Figure 4). Differently from NO_2 ARPA data, the meteorological sensors have a 10 min time resolution. Access to this dataset is achieved in the same way as described in Section 2.1. For this work, the meteorological parameters that we used were temperature, wind speed, wind direction, precipitation, global radiation, and relative humidity.

The original administrative boundaries and meteorological sensors of the MCM include the city San Colombano al Lambro city, which is outside the boundaries of those indicated in Figure 4. This portion of the MCM was excluded from our study area because no ground NO_2 stations are present here.

2.3. Satellite Data

Sentinel-5P was the chosen NO_2 satellite to train the model. Although other satellites, OMI from the National Aeronautics and Space Administration (NASA), provide NO_2 measurements, Sentinel-5P has a resolution of 5.5 × 3.5 km compared to 13 × 24 km of OMI [32]. Both Sentinel-5P and OMI have a daily temporal resolution. The daily satellite passage over the Metropolitan City of Milan is variable, but it is always in the time range between 11:00 and 14:00 UTM+0 (Figure 5).

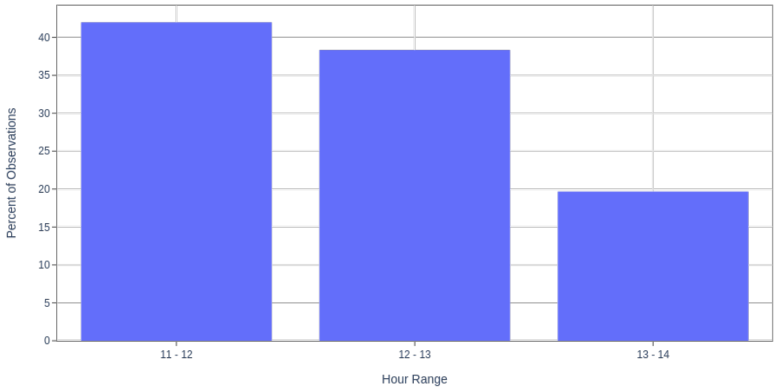


Figure 5. Percentage of observations at 1-h resolution for the Sentinel-5P passing over the MCM at UTM+0.

Access to the Sentinel data is free and open through the Open Access Hub (https://scihub.copernicus.eu/, (accessed on 30 september 2022)). For Sentinel-5P, batch data download is not available as for other Sentinel missions (e.g., Sentinel-2) in this portal. For this satellite, the website allows the download of one image at a time. The solution adopted in this work was the service provided by the Data and Information Access Services (DIAS). DIAS are private entities that deliver data processing through their platforms. Although the use of DIAS systems requires payment, they provide the original Sentinel observations through an API that allows batch mode download. For this work, we used the WEkEO DIAS (https://www.wekeo.eu/, (accessed on 30 september 2022)) API.

Once the satellite datasets were downloaded, we transformed them from Level 2 to Level 3. This consists of removing pixels containing disturbances (such as clouds), performing coordinates projection, as well as gridding and trimming to the area of interest. Level-2 transformation enables the user to choose the quality factor through the variable called 'pixel validity'. In this project, we used a pixel validity of 75 (from a range of 0 to 100), as suggested in the Sentinel-5P manuals [33]. The ESA developed the HARP tool, which we used for processing Sentinel-5P images. This is software designed for the processing of satellite images of Copernicus Sentinel missions [34].

To filter out satellite observations not belonging to the Metropolitan City of Milan, we used the official administrative boundaries published as a Shape File (https://www.istat.it/it/archivio/222527, (accessed on 30 september 2022)).

2.4. Data Pre-Processing

This work's three chosen data sources are delivered in different spatial and temporal resolutions. These spatiotemporal differences required a pre-processing step to make them coincide in an overlapping grid. On the one hand, satellite observations are regularly gridded with a resolution indicated in Table 1. On the other hand, ground NO_2 and meteorological ARPA station networks are both irregular but have different locations inside the MCM. To solve this situation, we considered the MCM as an average measurement for each of the measurements at the time period from 12:00 h to 15:00 h UTC+1. To validate this choice, we calculated the standard deviation and the mean for the ARPA NO_2 ground measurements inside the MCM for each day (from 12:00 h to 15:00 h UTC+1). As shown in Figure 6, the standard deviation is approximately 40% of the mean values. This indicated that we could consider the complete MCM city of Milan as a single average measurement of NO_2 for each day between 12:00 h and 15:00 h UTC+1.

Table 1. Data sources for spatial and temporal resolutions.

Dataset	Spatial Resolution (km × km)	Temporal Resolution
ARPA NO_2	Non-regular gridding	1-h
ARPA Meteorological	Non-regular gridding	10-min
Sentinel-5P	3.5 × 5	1 day

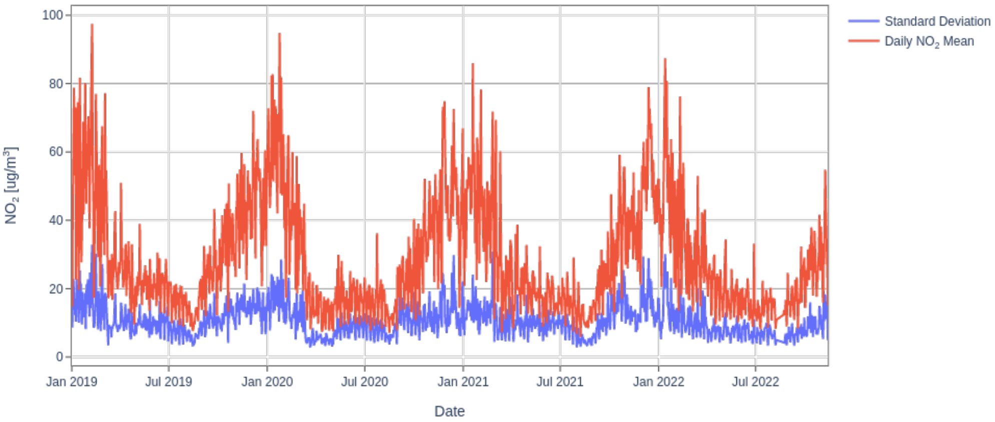


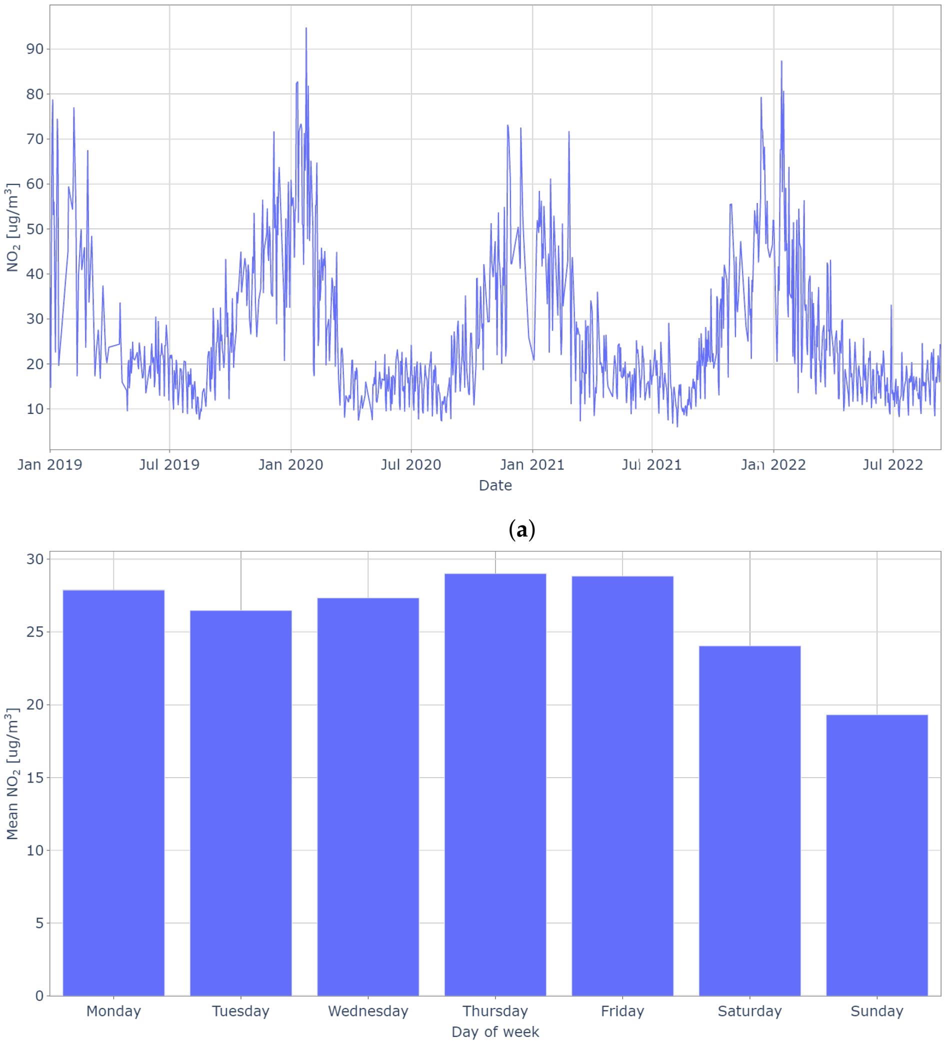
Figure 6. ARPA Lombardia NO_2 sensors' standard deviation vs. mean over time.

To describe better the physicochemical dynamics of the MCM area (e.g., the effect of wind in the time previous to the satellite passage), we considered meteorological data outside the time of passage of the satellite. Therefore, we created an average measurement of each meteorological indicator for the 21 h prior to the passage of the satellite. This was achieved by averaging daily measurements starting from 15:00 UTC+1 of the previous day until 12:00 UTC+1 on the day of the measurement. This 21 h period average was added to the data frame for every meteorological measurement as individual features.

Moreover, ARPA NO_2 data at the time of passage of the satellite presents a yearly seasonality (Figure 7a). The trend reveals that atmospheric pollution concentrations increase in winter and decrease in summer. In a similar fashion, the NO_2 weekly trend can be observed in Figure 7b, where NO_2 tends to decrease during the weekends. For this reason, we included both features, indicating the month and the day of the week. Given the difference in NO_2 concentrations between weekdays and weekends, we included a feature indicating this. All the variables considered for this study can be seen in Table 2.

Table 2. Variables considered to model ground NO_2 with the used temporal resolutions.

Used Temporal Resolution	Variable	Source
Daily	NO_2 Day of the week Month	Sentinel-5P - -
Daily (average from 12:00 h to 15:00 h)	NO_2 Temperature Wind speed Wind direction Precipitation Global radiation Relative humidity	ARPA atm. pollution ARPA meteo ARPA meteo ARPA meteo ARPA meteo ARPA meteo ARPA meteo
Daily (average from 15:00 h of previous day to 12:00 h current day)	NO_2 Temperature Wind speed Wind direction Precipitation Global radiation Relative humidity	ARPA atm. pollution ARPA meteo ARPA meteo ARPA meteo ARPA meteo ARPA meteo ARPA meteo

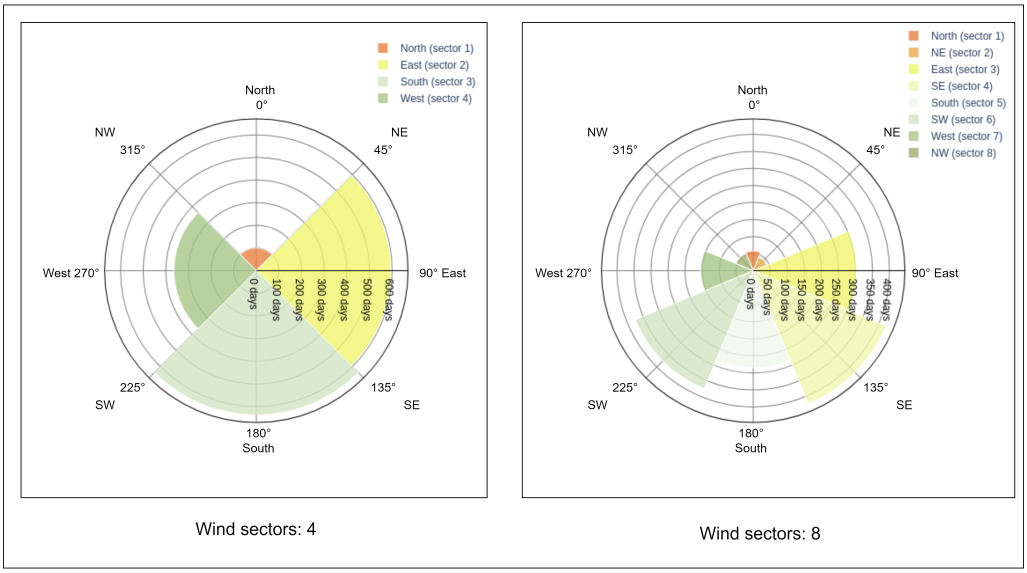


(**a**)

(**b**)

Figure 7. ARPA NO_2 ground sensor measurements for the Metropolitan City of Milan. (**a**) Average NO_2 from 12:00 h to 15:00 h UTC+1 for the complete study period. (**b**) Average NO_2 from 12:00 h to 15:00 h UTC+1 for each day of the week for the complete study period.

Wind direction data required further processing with respect to the rest of the meteorological data. Given that wind direction is originally delivered in degrees, this measure cannot be simply averaged for the satellite passage time. Doing so would lead to a wind direction result that is not true for that period. For this reason, we decomposed the wind direction into north and east vectors and multiplied them by the wind speed. When calculating the mean of this measure, we are considering the wind strength (velocity) to weigh the average over a period of time. This operation was performed both for the satellite passage time and for the 21 h period prior to the passage time. After obtaining this average, the measurement was transformed back to wind direction in degrees. For model training purposes, we classified wind direction into four and eight sectors (Figure 8) independently. The comparison of these two classifications will be further described in Section 2.7.2.



Wind sectors: 4

Wind sectors: 8

Figure 8. ARPA Lombardia wind direction data classified by sectors.

2.5. Methods

The task of estimating ground or surface level NO_2 concentrations has been of interest to several researchers to help with monitoring air quality. Many of the most recent approaches use Machine Learning (ML) or Deep Learning (DL) models. These techniques are used to integrate the meteorological and air quality data (both from satellite and ground stations). Furthermore, vehicle traffic data and spatially descriptive data, such as land cover, elevation, and distribution of pollutants sources are used to estimate NO_2 or a combined estimation of NO_2 and other air pollutants (e.g., $PM_{2.5}$, PM_{10}, O_3).

Recent ML approaches include Random Forest (RF) [15,16,35–37] as the method or one of the methods implemented in the studies. RF may also be used as a baseline to compare the results of more sophisticated methods such as DL models, e.g., total connection encoders and convolutional neural networks [38]. Other commonly used ML models, which are used individually or as part of ensemble models, include Support Vector Regression (SVR), Gradient Boosting (GB), and Extreme Gradient Boost (XGB) [17,35,39,40]. Additional studies utilised Linear Regression (LR) models such as Exponential Triple Smoothing (ETS) and Seasonal Autoregressive Integrated Moving Average (SARIMA) as the main methods to forecast NO_2 [41].

The studies that implement ML techniques to model NO_2 tend to use LR methods to compare with the ML results, e.g., comparing RF with LR [16], RF, SVR, and XGB with Multiple Linear Regression (MLR) [17].

Regarding NO_2 measurements, they can be retrieved from different sources such as Sentinel-5P TROPOMI NO_2 [15,17], Aura OMI NO_2 [35,37,38,40], Landsat 8 reflectance [42], and ground sensors [36,39,40]. Meanwhile, the meteorological information is usually derived from ground sensors belonging to regional or country monitoring networks [36,43]. Several studies also rely on other data relevant to model NO_2 and other pollutants, e.g., Digital Terrain Model (DTM) and power plants distribution [39], road density and population density [17], and land use [36,40].

Meteorological measurements and spatially descriptive data are the features used in the ML, DL or linear models. A subset of these features may be removed when they are not significant to model NO_2. The relevance of removing unnecessary features serves to avoid overfitting and to reduce processing time. The methods utilised for this purpose include recursive feature elimination [17], Spearman correlation coefficient calculation to examine the bivariate association between NO_2 and predictor variables [39], mutual information and maximum relevance-minimum redundancy [43].

Therefore, in this context, it is reasonable to test different ML approaches (Section 2.6) to find the one that has the lowest error with respect to the ground truth. This is performed with the aim of estimating daily NO_2 measurements at the city level (one average NO_2 value per day at the time of the passage of the Sentinel-5P).

2.6. Computational Regression Models

The ML models were implemented in Python due to the availability of scientific libraries. The Python scientific libraries used for this work were TensorFlow (https://www.tensorflow.org/, (accessed on 5 October 2022)) and SciKit-learn (https://scikit-learn.org/, (accessed on 5 october 2022)). SciKit-Learn is a library where Machine Learning models are implemented, and it can be used for training, fitting and testing the data. The trained SciKit-Learn models can then be exported and used for prediction purposes in other datasets containing the same features as the training one.

To achieve the estimation of ground NO_2 with Machine Learning, we used Non-Linear Multivariate Autoregressive Models. In addition, we treated the regression system as a series of discrete linear models by using splines and Kriging to evaluate the results with respect to ML models. Although pollution is a non-linear system, linear models are used to compare the performance of non-linear models [16]. Autoregressive models, such as the ones used for ML, have characteristics useful for this work, such as being theoretically grounded and generally not presenting overfitting [44]. From a practical point of view, these models are simple to train and configure.

Following Machine Learning best practices, 80% of the data were used for training and 20% were used for testing [44]. Normalisation was carried out to reduce the effect that using different units of measurement can have on the models. With this, all the variables had the same weight independently from their range of values. Normalisation was only performed using the training data to avoid introducing information from the testing set when training the model. Regarding the data splitting, we opted for not extracting data randomly but extracting the first 80% of the time period for training and using the last 20% for testing. This decision was taken on the one hand because this is the way in which the algorithm is planned to be used (i.e., daily consecutive estimations). On the other hand, it ensures that we introduce seasonality into the system by using all the data characteristics that were present in the training period (1 January 2019 to 6 March 2022). In the results section, we compare the results of doing this periodical data splitting with a randomised one.

For the purpose of this project, we decided to test 8 different models that are currently considered state-of-the-art. For the Machine Learning algorithms, we used Long Short-Term Memory (LSTM), Random Forest (RF), Epsilon-Support Vector Regression (SVR), Decision Tree Regression (DTR), Gradient Tree Boosting (GTB) and Multi-layer Perceptron Regressor (MLPR). Regarding the linear models, we used Kriging and B-Splines.

2.6.1. Long Short-Term Memory Algorithm

LSTM is a Neural Network (NN) supervised learning prediction architecture. This NN was used to take advantage of its capability to carry long-term dependencies to the future.

This characteristic is not present in other NN architectures [44]. Therefore, the LSTM was a good candidate for this dataset because data temporality is carried over almost 3 years.

2.6.2. Random Forest

RF is a machine-learning technique that makes predictions based on a group of regression trees (binary splits on predictor variables) [45]. As specified by Ref. [46], Regression Random Forests are created by developing trees based on a random vector. In comparison with other algorithms that only use bagging, RF prioritises random feature selection. We used this model because it is considered one of the state-of-the-art algorithms. Results have demonstrated a decrease in overall RMSE by using RF [46].

2.6.3. Support Vector Regression

SVR is commonly used for non-linear regression types. It is based on the same principle as Support Vector Machines (SVM). When analysing points in an n-dimensional space, the support vectors are those parallel to the hyperplane and closer to the points of interest. The purpose of this algorithm is to find those support vectors to create a model that best fits the data. Our interest in using this model is that in terms of predictability, the SVR has reportedly outperformed other linear, polynomial, and logistic regression models [47].

2.6.4. Decision Tree Regression

As described by [48], the DTR is based on a multistage or hierarchical decision scheme organised in a tree-like structure. It uses a collection of features together to accomplish regression in a single decision step. The tree is made up of a root node (the complete dataset), internal nodes, and terminal nodes. A binary choice is made at each node of the decision tree structure, which divides one class from the other classes. In a top-down procedure (such as the one used in this work), the processing is accomplished by travelling down the tree until it reaches the leaf node. The final solution is a result of splitting a decision into simpler ones [48]. Since this algorithm uses decision trees, we chose it as an alternative to compare its performance against RF.

2.6.5. Gradient Tree Boosting

GTB uses decision trees (Section 2.6.4) of a fixed size. With these trees, it creates a global model, which performs better than the single decision trees. We adopted this algorithm as a variation of DTR and RF because, as seen in the literature, it is used as an approach for regression with decision trees [17,35,39,40]. Given its performance, it is considered a state-of-the-art predictor with results at the level of Lasso and Random Forest [49].

2.6.6. Multi-Layer Perceptron Regressor

MLPR belongs to the class of Artificial Neural Networks (ANN). Each node of this ANN is a neuron that uses a non-linear activation function. Given that this model is able to discern between non-linearly separated data [50], it makes it a good candidate to use in this work. We used this model to analyse the data performance when distinguishing between non-linearly separated measurements.

2.6.7. B-Spline Regressor

B-splines is a linear regression tool. It is composed mainly of piecewise polynomials (usually cubic or quadratic) and has a smooth behaviour in comparison with higher-order polynomials. B-splines have continuous surfaces, reducing the problem of local polynomial approximations that usually present strong discontinuities. This is achieved by patching polynomial segments together to generate derivatives of lower order to join the intersections between each spline segment. We used this linear regression model to reduce the computational cost derived from Machine Learning algorithms [51].

2.6.8. Kriging Regressor

Kriging is a group of statistical methods for optimising spatial prediction. It has been applied to a wide range of fields and demonstrated a reduction in overall errors, such as RMSE [52]. As explained by Ref. [52], this regressor is a probabilistic predictor, and it assumes that the data are in the form of a statistical model. Since the prediction error is minimised and, on average, the predicted value and the true value agree, kriging predictors are referred to as optimal predictors. We used this tool not only because of its results in other areas (e.g., environmental sciences, mining and agriculture) but similar to B-Spline, it can help reduce the computational cost that is normally present in Machine Learning models [52].

2.7. Training and Testing of the Models

2.7.1. Training

As established in Section 2.6, originally, 80% of the data were used for training, and 20% were used for testing [44]. For the first phases of the project, we used only data from 1 January 2020 to 6 March 2022 to explore preliminary results. Later on, we decided to include data from 2019. To compare the results when introducing the 2019 datasets, the testing period (7 March 2022 to 28 September 2022) was maintained the same. This resulted in using approximately 85% for the training data and 15% for testing.

For training, we used as an input the Sentinel-5P NO_2 measurements and ARPA meteorological data for the period of 1 January 2019 to 6 March 2022 at the time of passage of the satellite and the 21-h period previous to this. As the target output, we provided the model with the ARPA NO_2 ground average measurement for each day from 12:00 h to 15:00 h UTC+1 for the MCM.

2.7.2. Testing

For the testing, we used the period from 7 March 2022 to 28 September 2022. This step consisted of using the trained model (Section 2.7.1) to estimate the NO_2 ground average measurement for each day from 12:00 h to 15:00 h UTC+1 for the MCM.

As mentioned in Section 2.4, testing was performed to determine if the models would have a lower error by classifying the wind direction into four or eight categories. This consisted of testing all the regression models either with a 4 or an 8-sector division.

A point to consider when analysing the testing data is that ARPA NO_2 and meteorological ground sensors were not available from 9 August 2022 to 23 August 2022. This was a specific situation for the MCM, and therefore, ground estimations do not consider this period of time.

3. Results

In this section, we present the results for each of the models using training and testing data for the period from 1 January 2019 to 28 September 2022. To evaluate the quality of the results and estimate the model error, we used the Root Mean Square Error (RMSE), as suggested in the literature [44].

Table 3 contains the RMSEs obtained using each of the models. The first column indicates whether the model belonged to the group of ML, linear models, the combination of more than one model or the feature selection algorithms. The second column contains the name of the model itself. The third and fourth columns show the RMSE for each of the listed models. The difference between these two is the method performed in the phase of data splitting (training and testing) specified in Section 2.7.1. The periodical sampling column involved using training data on the dates between 1 January 2019 and 6 March 2022. For the random sampling, we performed the training/testing splitting randomly but kept the 85/15% testing and training ratio.

Table 3. RMSE between ground NO_2 estimated values and ground truth from NO_2 ARPA ground stations.

Type of Model	Model	Periodical Sampling RMSE (μg/m^3)	Random Sampling RMSE (μg/m^3)
Machine Learning	LSTM	3.77	6.82
	Random Forest	3.70	7.13
	Support Vector Reg.	3.99	6.37
	Decision Tree Reg.	4.93	9.61
	Gradient Tree Boosting	3.53	6.80
	MLPR	3.23	6.42
Linear Models	Kriging	3.78	6.11
	B-Spline	4.19	6.63
Model Combination	Voting (MLPR + Kriging)	3.50	6.06
	Stacking (MLPR + Kriging)	3.55	5.98
Feature Selection	RF Feat. Sel. + Voting (MLPR + Kriging)	2.89	6.09
	CV Feat. Sel. + Voting (GTB + Kriging)	3.21	5.99
	Correlation > 0.6 + Voting (MLPR + Kriging)	4.15	6.87

From the results, we can observe a significant difference between the periodical and the random sampling, where periodical sampling has a lower RMSE. The difference can possibly be explained by the seasonality of the input data. On the one hand, we have the periodical sampling where we ensure that all the temporal, environmental and anthropological behaviours are used for training. On the other hand, random sampling does not necessarily take this into consideration. Random sampling extracts dates without taking into consideration if a particular season or period is included in the model training. An example of this is the COVID-19 Italian lockdown period.

The time span used for this work (1 January 2019 to 28 September 2022) comprises data from the COVID-19 pandemic, pre-pandemic and post-pandemic periods. The COVID-19 pandemic period represents the abnormal behaviour of people's transportation dynamics due to lockdown restrictions that were in place in Italy (9 March 2020 to 3 May 2020). Average NO_2 concentrations at the time of passage of the satellite declined for each day of the week during the lockdown period compared to the same period in 2021 (Figure 9) and 2019 [53]. Google Mobility data (https://www.google.com/covid19/mobility/, (accessed on 20 October 2022)) report the percentage change in people's movements during the COVID-19 lockdown period compared to a previous period where no lockdown was present [54]. By using this data, we can attribute that the decline in people's mobility was caused by the transportation restrictions imposed by local authorities.

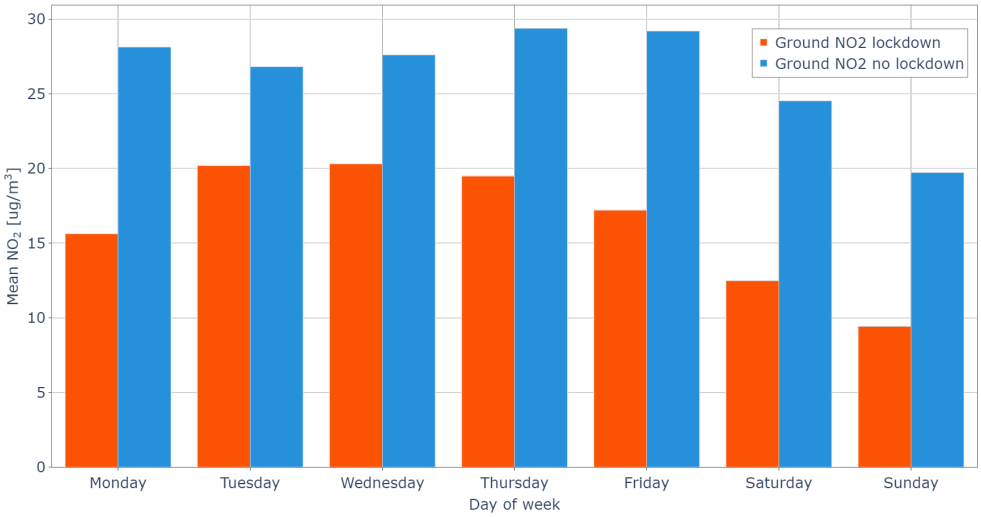


Figure 9. Mean ground NO_2 during 2020 Italy's COVID-19 lockdown vs. 2021 no-lockdown period.

The percentage decrease in Google mobility data (Figure 10) is calculated with respect to a baseline period before the pandemic outbreak. The baseline days are determined as the median value throughout the five-week period from 3 January to 6 February 2020. They indicate the normal value for that day of the week [54].

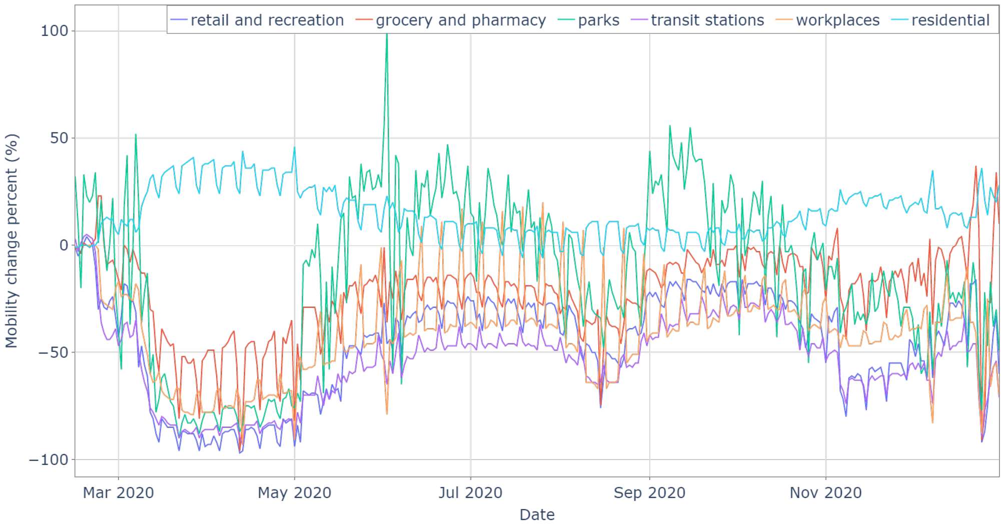


Figure 10. COVID-19 pandemic and post-pandemic Google mobility data for a variety of location categories [54].

In addition to the COVID-19 lockdown period, the wind direction (Figure 8) presents some features to highlight in terms of periodicity. When aggregating the wind direction into four groups, we observed that the most frequent one in the MCM was towards the south. Further division (into 8 sectors) reveals that the most frequent wind direction during the selected time period was the south-east direction and less frequent towards the north-east. We examined this difference during the testing phase by training and testing all the models first classifying them into four sectors and then into eight wind sectors. Results demonstrate that for the best models, we have an improvement from 3.00 $\mu g/m^3$ (4 wind sectors) to 2.89 $\mu g/m^3$ (8 wind sectors). Therefore, we can assume that the eight-sector wind division has a lower RMSE because the model is able to distinguish better the relationship between NO_2 and the wind direction. A clear example of this is the west and south-west directions, where the division into four sectors includes part of the west into the south-west. This distinction between west and south-west helps the model to assume that wind towards the west will result in a higher NO_2 concentration (Figure 11).

By making a more detailed analysis of the wind direction, we observed the presence of a relationship between wind direction and NO_2 concentrations. As seen in Figure 11, on average, there is a higher atmospheric pollution concentration when the wind direction is towards the north and the north-west. This behaviour is explained by the topographic characteristics illustrated in the introduction (Figure 3). Wind blowing from the south to the north and the north-west transports atmospheric pollutants, which due to the presence of the Alps and the Apennines, become entrapped in the Po Valley.

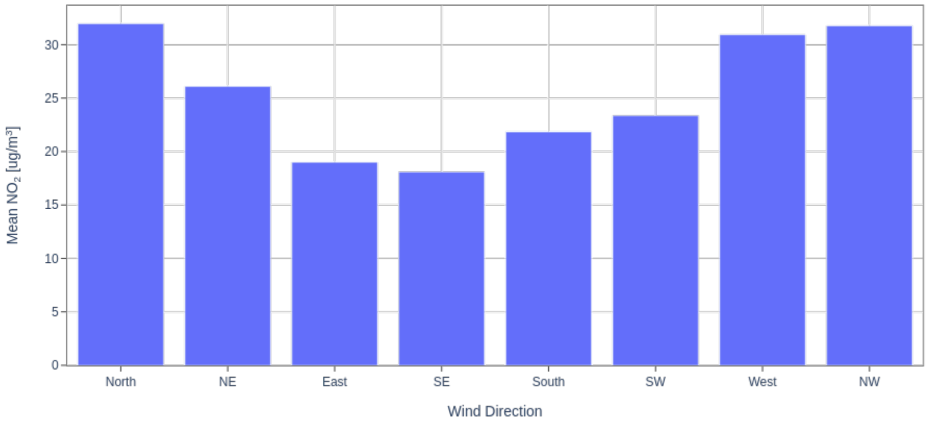


Figure 11. ARPA Lombardia NO_2 ground average measurements from 12:00 h to 15:00 h UTC+1 for each wind direction.

Given the points previously mentioned, we decided to use the periodical splitting method and eight sectors for the wind direction as features for our model. As observed in Table 3, the ML model that had the lowest RMSE was the MLPR. To combine characteristics from several models, we tested all the different pairs of model combinations by using Sklearn Voting and Stacking regressors. The result showed that the best model performance was obtained by using the voting mechanism to combine MLPR with Kriging.

Understanding if any of the used features could be removed was relevant because some of the features may not contribute towards the decrease in the RMSE. For this task, we used two Sklearn feature selection algorithms: the Random Forest and Cross Validation (CV). As an alternative, we calculated the Pearson correlation coefficient of all the inputs with respect to the real ground NO_2 measurement. We then used the features that had a Pearson correlation coefficient higher than 0.5 with respect to ground NO_2 to reduce the number of features. In Table 3, the RMSE results for these feature selection techniques demonstrate that by using the Random Forest feature selection algorithm in combination with the Voting Regressor, the RMSE decreased by approximately 18% with respect to the Voting (MLPR + Kriging) with no feature selection. The features selected by the best model (RF Feature Selection) and by correlation larger than 0.5 can be found in Table 4.

Table 4. Features selected by Random Forest algorithm and those with Pearson correlation coefficient larger than 0.5.

RF Selected Features	Features with Pearson Corr $>$ 0.5
Satellite NO_2	Satellite NO_2
Temperature at satellite passage time	Temperature at satellite passage
Wind speed at satellite passage time	Global radiation at satellite passage
Wind direction at satellite passage time	Relative humidity at satellite passage time
Global radiation at satellite passage time	Temperature before satellite passage
Temperature before satellite passage	Global radiation before satellite passage time
Wind speed before satellite passage	
Global radiation before satellite passage time	
Weekday/weekend classifier	
Day of the week	
The month of measurement	

Apart from the RMSE results of Table 3, Figure 12 plots the estimations obtained by the model with the lowest RMSE (voting algorithm combining MLPR and Kriging with RF feature selection). In this plot, we compare the estimated ground NO_2 and the ground truth for the testing period (7 March 2022 to 28 September 2022 from 12:00 h to 15:00 h UTC+1). We can observe that the trend and estimations are close to the ground truth. By calculating the residuals from the estimations and comparing them to three standard deviations of the ground truth, we found that the date of 24 March is the only one above this threshold. A relationship between this date and other factors (meteorological or high-intensity traffic) was explored, but none confirmed an explanation for the abnormality on this particular date.

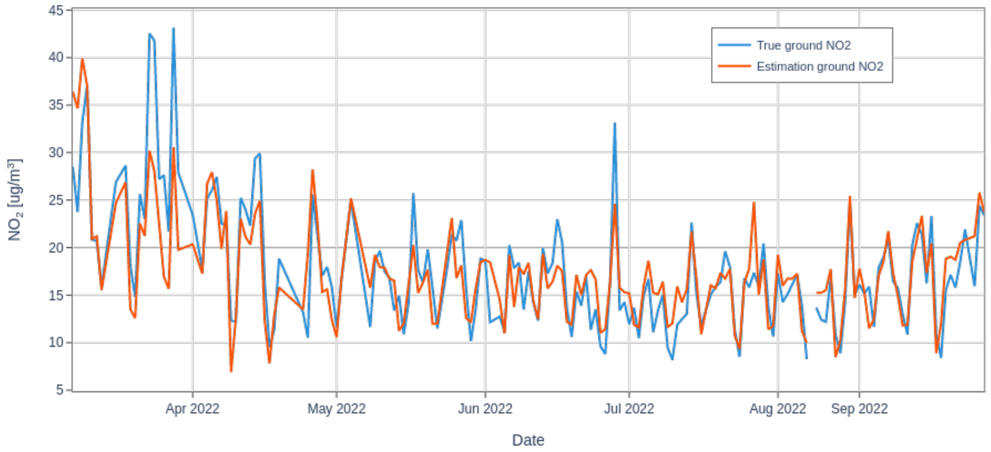


Figure 12. NO_2 ground truth vs. estimation from Voting Regression (MLPR + Kriging) and Random Forest feature selection algorithm.

Table 5 shows the statistical comparison between the ground truth and the best model (Voting Regression and RF feature selection with MLPR and Kriging) for the NO_2 ground estimations testing period (from 12:00 h to 15:00 h UTC+1). The best model's RMSE (2.89 μg/m^3) is significantly low because it is approximately lower by 50% compared to the ground truth's standard deviation. Additionally, it is 65% lower than the real minimum value observed during the testing period.

Table 5. Statistical comparison between Voting Regression (MLPR and Kriging) with RF feature selection NO_2 estimations and the ground truth NO_2 (from 12:00 h to 15:00 h UTC+1).

Data	**Measure**	**Value (μg/m^3)**
Ground Truth	Mean	17.69
	Standard Deviation	6.40
	Minimum	8.187
	Maximum	43.14
Model estimation	Mean	18.18
	Standard Deviation	5.97
	Minimum	7.41
	Maximum	37.493

As a comparison with our best result, we trained all the models using only Sentinel-5P data. The best model in this case had an RMSE of 6.18 μg/m^3 obtained with the SVR model. This demonstrates that the use of meteorological data in addition to satellite Sentinel-5P measurements decreases the estimation error of the NO_2 average (from 12:00 h to 15:00 h UTC+1) by approximately 45%.

4. Discussion and Conclusions

The main purpose of this study was to estimate ground NO_2 concentrations for the MCM from 12:00 h to 15:00 h UTC+1. We used ground meteorological data and Sentinel-5P NO_2 observations. This is relevant for regions where no NO_2 ground sensors are present, which in most cases, is a limitation for LMICs. The intention of using other networks that are globally available (i.e., meteorological indicators) apart from satellite data is to improve the estimation of NO_2 at the ground level.

Due to its topographical and anthropogenic characteristics, the study area for this work was the Metropolitan City of Milan. The period of time of the data was from 1

January 2019 to 28 September 2022 every day from 12:00 h to 15:00 h UTC+1. The estimation was accomplished by training linear and non-linear multivariate autoregressive models. The final output compares the different models based on the RMSE obtained from their estimations. In detail, the groups compared were ML models, linear models and a combination of ML and linear models. In order to decrease the number of features and reduce computational resources, we also used feature selection algorithms. The model with the lowest RMSE (2.89 $\mu g/m^3$) was the RF feature selection and a voting regression combining MLPR and Kriging. The number of features was reduced from 15 to 11 by using the RF feature selection algorithm.

In addition to the developed model, this study confirms the strong relationship that meteorological conditions have on NO_2 concentrations. This can be specially observed with features such as wind direction and ground temperature. Due to the NO_2 seasonal behaviour, the period of the year also has a strong influence on the model. Another factor which may be relevant to the NO_2 atmospheric ground concentrations is vehicle traffic. This is worth considering for future studies.

As indicated in Table 5, the estimations have an RMSE that is more than 50% lower than the ARPA NO_2 standard deviation for the daily average from 12:00 h to 15:00 h UTC+1. These results are significant compared to using only Sentinel-5P data for training the model, which resulted in a decrease of approximately 45% in the RMSE.

In future phases of this work, we will analyse the use of satellite data only (both meteorological and atmospheric). This will be achieved by using satellite meteorological open data (e.g., EUMETSAT from the EU or GOES from NASA) to ensure an air quality assessment independent of ground sensors. Additionally, we will explore the estimation of ground NO_2 at the resolution of 5.5 by 3.5 km as used by the Sentinel-5P.

Author Contributions: Conceptualization, Maria Antonia Brovelli and Jesus Rodrigo Cedeno Jimenez; methodology, Maria Antonia Brovelli and Jesus Rodrigo Cedeno Jimenez; software, Jesus Rodrigo Cedeno Jimenez and Angelly de Jesus Pugliese Viloria; validation, Maria Antonia Brovelli, Jesus Rodrigo Cedeno Jimenez, and Angelly de Jesus Pugliese Viloria; formal analysis, Maria Antonia Brovelli, Jesus Rodrigo Cedeno Jimenez, and Angelly de Jesus Pugliese Viloria; investigation, Jesus Rodrigo Cedeno Jimenez and Angelly de Jesus Pugliese Viloria; resources, Maria Antonia Brovelli; data curation, Jesus Rodrigo Cedeno Jimenez; writing—original draft preparation, Jesus Rodrigo Cedeno Jimenez; writing—review and editing, Maria Antonia Brovelli, Jesus Rodrigo Cedeno Jimenez, and Angelly de Jesus Pugliese Viloria; visualization, Jesus Rodrigo Cedeno Jimenez, Angelly de Jesus Pugliese Viloria; supervision, Maria Antonia Brovelli; project administration, Maria Antonia Brovelli; funding acquisition, Maria Antonia Brovelli. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially funded by the project Harmonia (identifier: H2020-LC-CLA-2018-2019-2020), an EU Horizon 2020 project dedicated to the development of a support system for improved resilience and sustainable urban areas to cope with climate change and urban areas.

Data Availability Statement: ARPA Lombardia NO_2 and meteorological ground measurements can be accessed and downloaded directly from their open portal (http://dati.lombardia.it, (accessed on 30 September 2022)); Sentinel-5P NO_2 tropospheric measurements can be downloaded from their Open Access Hub (https://s5phub.copernicus.eu/, (accessed on 30 September 2022)); Python processing and analysis scripts can be accessed directly from our GitHub repository (https://github.com/rodrigocedeno/NO2-GroundPollution-AI.git, (accessed on 11 November 2022)) by using the available Jupyter Notebooks.

Acknowledgments: Thanks to ARPA Lombardia for giving us support and providing the data through their open portal http://dati.lombardia.it, (accessed on 30 September 2022). Thanks to the Copernicus mission for providing Sentinel-5P open data NO_2 tropospheric measurements. Thanks to the WEKEO DIAS platform (https://www.wekeo.eu/, (accessed on 30 September 2022)), we were able to download Sentinel-5P data in batch mode. This helped to reduce data download time.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
API	Application Programming Interface
ARPA	Agenzia Regionale per la Protezione Ambientale
CO	Carbon Monoxide
COPD	Chronic Obstructive Pulmonary Disease
COVID	Corona Virus Disease
CV	Cross Validation
DIAS	Data and Information Access Services
DL	Deep Learning
DTM	Digital Terrain Model
DTR	Decision Tree Regression
EEA	European Environment Agency
ETS	Extreme Triple Smoothing
ESA	European Space Agency
EU	European Union
EUMETSAT	European Union Meteorological Satellites
GB	Gradient Boosting
GOES	Geostationary Operational Environment Satellite
GTB	Gradient Tree Boosting
LMICs	Low- and Middle-Income Countries
LR	Linear Regression
LSTM	Long Short-Term Memory
MCM	Metropolitan City of Milan
MDPI	Multidisciplinary Digital Publishing Institute
ML	Machine Learning
MLPR	Multi-Layer Perceptron Regressor
MLR	Multiple Linear Regression
MODIS	Moderate Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
NN	Neural Network
NO_2	Nitrogen Dioxide
NO	Nitrogen Monoxide
O_3	Ozone
ODC	Open Data Cube
OMI	Ozone Monitoring Instrument
PM	Particulate Matter
RF	Random Forest
RMSE	Root Mean Square Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SDGs	Sustainable Development Goals
SO_2	Sulphur Dioxide
SVM	Support Vector Machine
SVR	Support Vector Regression
TROPOMI	TROPOspheric Measurement Instrument
UN	United Nations
USA	United States of America
WHO	World Health Organization
WMO	World Meteorological Organization
XGB	Extreme Gradient Boost

References

1. European Environment Agency. *Health Impacts of Air Pollution in Europe, 2021*; EEA: Copenhagen, Denmark, 2021.
2. United Nations. The 17 Sustainable Development Goals. Available online: https://sdgs.un.org/goals (accessed on 25 September 2022).

3. Trushna, T.; Tiwari, R.R. Establishing the National Institute for Research in Environmental Health, India. *Bull. World Health Organ.* **2022**, *100*, 281–285.
4. Zhang, Z.; Wang, J.; Lu, W. Exposure to Nitrogen Dioxide and Chronic Obstructive Pulmonary Disease (COPD) in Adults: A Systematic Review and Meta-Analysis. *Environ. Sci. Pollut. Res.* **2018**, *25*, 15133–15145. [CrossRef] [PubMed]
5. Tyagi, S.; Chaudhary, M.; Ambedkar, A.K.; Sharma, K.; Gautam, Y.K.; Singh, B.P. Metal Oxide Nanomaterials based sensors for monitoring environmental NO_2 and its impact on plant ecosystem: A Review. *Sens. Diagn.* **2022**, *1*, 106–129. [CrossRef]
6. European Environment Agency. *Emissions from Road Traffic and Domestic Heating behind Breaches of EU Air Quality Standards across Europe*; European Environment Agency: Copenhagen, Denmark, 2022.
7. Pruitt, E. Review of the primary national ambient air quality standards for oxides of nitrogen. *Fed. Regist* **2018**, *83*, 17226–17278.
8. Piccoli, A.; Agresti, V.; Balzarini, A.; Bedogni, M.; Bonanno, R.; Collino, E.; Colzi, F.; Lacavalla, M.; Lanzani, G.; Pirovano, G.; et al. Modeling the Effect of COVID-19 Lockdown on Mobility and NO_2 Concentration in the Lombardy Region. *Atmosphere* **2020**, *11*, 1319. [CrossRef]
9. EEA. *Air Quality Standards*; European Environment Agency: Copenhagen, Denmark, 2021.
10. ESA. *Copernicus in Detail*; European Space Agency: Paris, France, 2019.
11. Reimann, S.; Wegener, R.; Claude, A.; Sauvage, S. *Updated Measurement Guideline for NOx and VOCs*; Actris: Dübendorf, Switzerland, 2018.
12. Kramer, H.J. Copernicus: Sentinel-5P (Precursor—Atmospheric Monitoring Mission). Publication Title: Copernicus: Sentinel-5P—Satellite Missions—eoPortal Directory. European Space Agency, Paris, France. 2012. Available online: https://www.eoportal.org/satellite-missions/copernicus-sentinel-5p#ground-segment (accessed on 25 August 2022).
13. Pinder, R.W.; Klopp, J.M.; Kleiman, G.; Hagler, G.S.W.; Awe, Y.; Terry, S. Opportunities and challenges for filling the air quality data gap in low- and middle-income countries. *Atmos. Environ.* **2019**, *215*, 116794. [CrossRef]
14. *WMO Weather Stations*; World Meteorological Organization (WMO): Geneva, Swizerland, 2021. Available online: http://www.wmo.int/datastat/wmodata_en.html (accessed on 7 May 2021).
15. Long, S.; Wei, X.; Zhang, F.; Zhang, R.; Xu, J.; Wu, K.; Li, Q.; Li, W. Estimating daily ground-level NO_2 concentrations over China based on TROPOMI observations and machine learning approach. *Atmos. Environ.* **2022**, *289*, 119310.
16. Araki, S.; Shima, M.; Yamamoto, K. Spatiotemporal land use random forest model for estimating metropolitan NO_2 exposure in Japan. *Sci. Total Environ.* **2018**, *634*, 1269–1277.
17. Kang, Y.; Choi, H.; Im, J.; Park, S.; Shin, M.; Song, C.K.; Kim, S. Estimation of surface-level NO_2 and O_3 concentrations using TROPOMI data and machine learning over East Asia. *Environ. Pollut.* **2021**, *288*, 117711. [CrossRef]
18. He, S.; Dong, H.; Zhang, Z.; Yuan, Y. An Ensemble Model-Based Estimation of Nitrogen Dioxide in a Southeastern Coastal Region of China. *Remote Sens.* **2022**, *14*, 2807.
19. Jiang, Q.; Christakos, G. Space-time mapping of ground-level PM2.5 and NO_2 concentrations in heavily polluted northern China during winter using the Bayesian maximum entropy technique with satellite data. *Air Qual. Atmos. Health* **2018**, *11*, 23–33.
20. Huang, C.; Sun, K.; Hu, J.; Xue, T.; Xu, H.; Wang, M. Estimating 2013–2019 NO_2 exposure with high spatiotemporal resolution in China using an ensemble model. *Environ. Pollut.* **2022**, *292*, 118285.
21. Naseer, E.; Basit, A.; Bhatti, M.K.; Siddique, M.A. *Machine Learning for Area-Wide Monitoring of Surface Level Concentration of NO_2 Using Remote Sensing Data*; Institute of Electrical and Electronics Engineers: Piscataway, NJ, USA, 2023; pp. 1–6. [CrossRef]
22. Chen, Z.Y.; Zhang, R.; Zhang, T.H.; Ou, C.Q.; Guo, Y. A kriging-calibrated machine learning method for estimating daily ground-level NO_2 in mainland China. *Sci. Total Environ.* **2019**, *690*, 556–564.
23. Sekiya, T.; Miyazaki, K.; Eskes, H.; Sudo, K.; Takigawa, M.; Kanaya, Y. A comparison of the impact of TROPOMI and OMI tropospheric NO_2 on global chemical data assimilation. *Atmos. Meas. Tech.* **2022**, *15*, 1703–1728.
24. Wu, Y.; Di, B.; Luo, Y.; Grieneisen, M.L.; Zeng, W.; Zhang, S.; Deng, X.; Tang, Y.; Shi, G.; Yang, F.; et al. A robust approach to deriving long-term daily surface NO_2 levels across China: Correction to substantial estimation bias in back-extrapolation. *Environ. Int.* **2021**, *154*, 106576.
25. Scheibenreif, L.; Mommert, M.; Borth, D. Toward Global Estimation of Ground-Level NO_2 Pollution With Deep Learning and Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14.
26. Kim, M.; Brunner, D.; Kuhlmann, G. Importance of satellite observations for high-resolution mapping of near-surface NO_2 by machine learning. *Remote Sens. Environ.* **2021**, *264*, 112573.
27. Diémoz, H.; Barnaba, F.; Magri, T.; Pession, G.; Dionisi, D.; Pittavino, S.; Tombolato, I.K.; Campanelli, M.; Ceca, L.S.D.; Hervo, M.; et al. Transport of po valley aerosol pollution to the northwestern Alps–Part 1: Phenomenology. *Atmos. Chem. Phys.* **2019**, *19*, 3065–3095. [CrossRef]
28. Cedeno Jimenez, J.R.; Oxoli, D.; Brovelli, M.A. Enabling Air Quality Monitoring with the Open Data Cube: Implementation for Sentinel-5P and Ground Sensor Observations. *Int. Arch. Photogramm. Remote Sensing And Spat. Inf. Sci.* **2021**, *XLVI-4/W2-2021*, 31–36. [CrossRef]
29. Li, M.; Wu, Y.; Bao, Y.; Liu, B.; Petropoulos, G.P. Near-Surface NO_2 Concentration Estimation by Random Forest Modeling and Sentinel-5P and Ancillary Data. *Remote Sens.* **2022**, *14*, 3612.
30. ARPA Lombardia. *Dati Sensori Aria: Open Data Regione Lombardia*; ARPA: Milan, Italy, 2021.
31. Lombardia, A. *Criteri di Rilevamento—Aria/Qualità Dell'Aria: ARPA Lombardia*; ARPA: Milan, Italy, 2021.
32. Earth Science Data Systems. Available online: https://www.earthdata.nasa.gov/sensors/omi (accessed on 15 October 2022).

33. van Geffen, J.H.G.M.; Eskes, H.J.; Boersma, K.F.; Veefkind, J.P. *TROPOMI ATBD of the Total and Tropospheric NO_2 Data Products*; Royal Netherlands Meteorological Institute, Ministry of Infrastructure and Water Management: Utrecht, The Netherlands, 2022; p. 88.
34. STDelftCorp. *HARP Manual*; Science and Technology: Delft, The Netherlands, 2021.
35. Di, Q.; Amini, H.; Shi, L.; Kloog, I.; Silvern, R.; Kelly, J.; Sabath, M.B.; Choirat, C.; Koutrakis, P.; Lyapustin, A.; et al. Assessing NO_2 concentration and model uncertainty with high spatiotemporal resolution across the contiguous united states using ensemble model averaging. *Environ. Sci. Technol.* **2020**, *54*, 1372–1384.
36. Gariazzo, C.; Carlino, G.; Silibello, C.; Renzi, M.; Finardi, S.; Pepe, N.; Radice, P.; Forastiere, F.; Michelozzi, P.; Viegi, G.; et al. A multi-city air pollution population exposure study: Combined use of chemical-transport and random-Forest models with dynamic population data. *Sci. Total Environ.* **2020**, *724*, 138102.
37. Kamińska, J.A. A random forest partition model for predicting NO_2 concentrations from traffic flow and meteorological conditions. *Sci. Total Environ.* **2019**, *651*, 475–483.
38. Liu, X.; Li, C.; Liu, D.; Grieneisen, M.L.; Yang, F.; Chen, C.; Zhan, Y. Hybrid deep learning models for mapping surface NO_2 across China: One complicated model, many simple models, or many complicated models? *Atmos. Res.* **2022**, *278*, 106339. [CrossRef]
39. Wong, P.Y.; Su, H.J.; Lee, H.Y.; Chen, Y.C.; Hsiao, Y.P.; Huang, J.W.; Teo, T.A.; Wu, C.D.; Spengler, J.D. Using land-use machine learning models to estimate daily NO_2 concentration variations in Taiwan. *J. Clean. Prod.* **2021**, *317*, 128411.
40. Chen, J.; Hoogh, K.D.; Gulliver, J.; Hoffmann, B.; Hertel, O.; Ketzel, M.; Bauwelinck, M.; Donkelaar, A.V.; Hvidtfeldt, U.A.; Katsouyanni, K.; et al. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. *Environ. Int.* **2019**, *130*, 104934.
41. Hong, W.Y.; Koh, D.; Mohtar, A.A.A.; Latif, M.T. Statistical Analysis and Predictive Modelling of Air Pollutants Using Advanced Machine Learning Approaches. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2020*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2020; ISBN 9781665419741. [CrossRef]
42. Yang, Q.; Yuan, Q.; Gao, M.; Li, T. A new perspective to satellite-based retrieval of ground-level air pollution: Simultaneous estimation of multiple pollutants based on physics-informed multi-task learning. *Sci. Total Environ.* **2023**, *857*, 159542. [CrossRef]
43. Iskandaryan, D.; Sabatino, S.D.; Ramos, F.; Trilles, S. Exploratory Analysis and Feature Selection for the Prediction of Nitrogen Dioxide. *AGILE GISci. Ser.* **2022**, *3*, 1–11.
44. Nielsen, A. *Practical Time Series Analysis: Prediction with Statistics and Machine Learning*, 1st ed.; O'Reilly Media: Sebastopol, CA, USA, 2019.
45. Speiser, J.L.; Miller, M.E.; Tooze, J.; Ip, E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst. Appl.* **2019**, *134*, 93–101. [CrossRef]
46. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
47. Parbat, D.; Chakraborty, M. A python based support vector regression model for prediction of COVID19 cases in India. *Chaos Solitons Fractals* **2020**, *138*, 109942. [CrossRef] [PubMed]
48. Xu, M.; Watanachaturaporn, P.; Varshney, P.K.; Arora, M.K. Decision tree regression for soft classification of remote sensing data. *Remote Sens. Environ.* **2005**, *97*, 322–336. [CrossRef]
49. Biau, G.; Cadre, B.; Rouvière, L. Accelerated gradient boosting. *Mach. Learn.* **2019**, *108*, 971–992. [CrossRef]
50. Mao, F.; Zhou, X.; Song, Y. Environmental and Human Data-Driven Model Based on Machine Learning for Prediction of Human Comfort. *IEEE Access* **2019**, *7*, 132909–132922.
51. Unser, M.; Aldroubi, A.; Eden, M. B-spline signal processing. I. Theory. *IEEE Trans. Signal Process.* **1993**, *41*, 821–833.
52. Krivoruchko, K. *Empirical Bayesian Kriging*; ESRI: Redlands, CA, USA, 2012; p. 5.
53. Oxoli, D.; Cedeno Jimenez, J.R.; Brovelli, M.A. Assessment of Sentinel-5P Performance for Ground-Level Air Quality Monitoring: Preparatory Experiments over the COVID-19 Lockdown Period. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *XLIV-3/W1-2020*, 111–116. [CrossRef]
54. Mathieu, E.; Ritchie, H.; Rodés-Guirao, L.; Appel, C.; Giattino, C.; Hasell, J.; Macdonald, B.; Dattani, S.; Beltekian, D.; Ortiz-Ospina, E.; et al. Coronavirus Pandemic (COVID-19). In *Our World in Data*; University of Oxford: Oxford, UK, 2020.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.