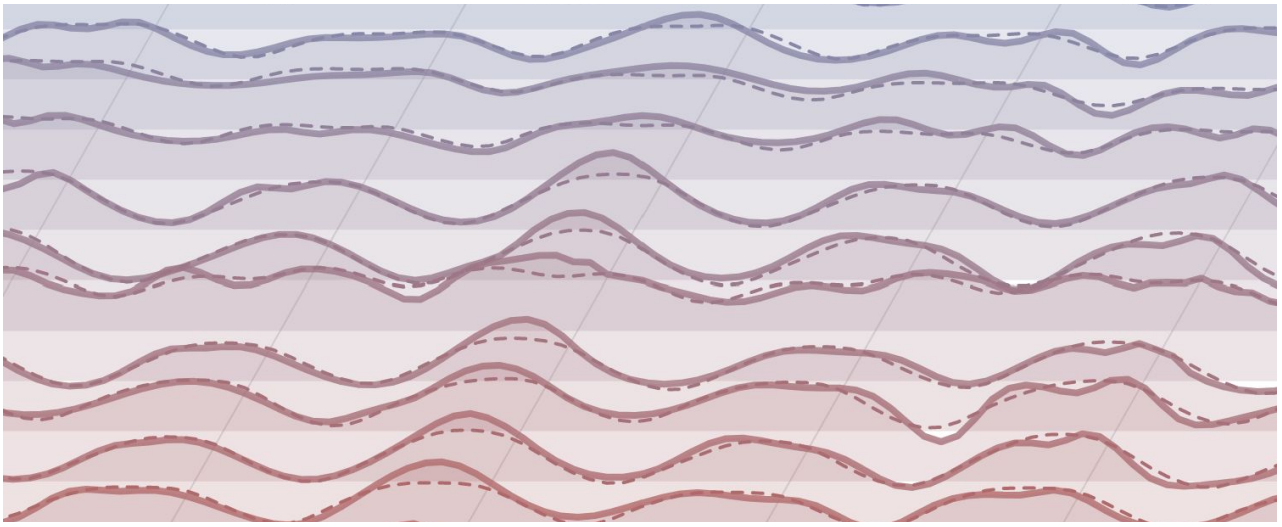
Geodan
President Kennedylaan 1
1079 MB Amsterdam (NL)
Tel. +31 (0)20 - 5711 311
Fax +31 (0)20 - 5711 333
E-mail: info@geodan.com
Web: www.geodan.com

Internship Report

Investigating air quality in Amsterdam during the 2020 COVID-19 lockdowns



Name student
MA Hamelberg
Registration number
910824302050
Mobile number student
+31652073899
Period of internship
2020-06-01 - 2020-09-30
Final date report
2020-09-30
Supervisor Geodan
T van Tilburg
Supervisor MGI
JGPW Clevers

Date 2020-09-30
Classification public
Status final
Version 1.0

Abstract

In this internship research report for *Geodan* we investigate air quality in and close to Amsterdam. In particular nitrogen dioxide (NO_2) is studied as this pollutant impacts our climate and health and its excess is generally produced by traffic according to earlier research. Amsterdam was subject to lockdowns during 2020 in response to the novel coronavirus disease (COVID-19). The economic and societal changes resulting from these lockdowns provide a unique opportunity to understand our air quality and how it reacts to external factors. Initially, four datasets are specified, which consist of ground / satellite based NO_2, road traffic, and weather data. The ground based NO_2 data is the response variable and the other datasets the external factors functioning as explanatory variables. The datasets are preprocessed, which includes temporal smoothing of the data and the conversion of data types to improve interoperability. Hereafter, their linear relationships are assessed. The external factors have a weak short term correlation to the ground based NO_2 data, with the exception of wind speed and satellite based NO_2 data. To further examine the relationship between the datasets, a prediction is performed using a machine learning model. The resulting variable importance indicates that wind speed was the most important variable in predicting ground based NO_2, followed by the wind direction and the road traffic intensity. The prediction accuracies are relatively high with an average Pearson's correlation coefficient of 0.824 and a coefficient of determination of 0.683. The prediction accuracy improves slightly when leaving out data during 2020 as well as with the inclusion of the previously filtered satellite data. In the final trend analysis, a harmonic model highlights deviations from the seasonal trend in the road traffic and NO_2 data during the COVID-19 lockdowns. NO_2 values dropped to ~65% in March 2020 and ~73% during the second quarter of 2020 compared to the mean NO_2 of the previous years. This coincided with similar drops in road traffic intensity. Lastly, when excluding data from 2020, the road traffic intensity and NO_2 seems to be negatively correlated. This relationship became positive when the 2020 data was included, where potentially both variables independently react to other factors, suggesting a neglectable impact of road traffic on NO_2 concentrations in and close to Amsterdam. Further research should investigate this phenomenon in greater detail by examining residual effects and subpopulations within the data.

List of abbreviations

Term	Abbreviation
Advanced Micro Devices	AMD
Application Programming Interface	API
Coefficient of Determination	r^2
Confidence Interval	CI
Earth Observatory	EO
European Environment Agency	EEA
Global Land Data Assimilation System	GLDAS
Goddard Earth Sciences Data and Information Services Center	GES DISC
Google Earth Engine	GEE
Koninklijk Nederlands Meteorologisch Instituut (Royal Netherlands Meteorological Institute)	KNMI
Landelijk Meetnet Luchtkwaliteit (National Air Quality Measuring Network)	LML
Limited Liability Company	LLC
Mean Squared Error	MSE
Nationale Databank Wegverkeersgegevens (National Data Warehouse for Traffic Information)	NDW
Nitrogen dioxide	NO_2
Novel Coronavirus Disease discovered late 2019	COVID-19
Particulate Matter (fijnstof)	PM
Pearson's Correlation Coefficient	r
Random Forest Regressor	RFR
Rijksinstituut voor Volksgezondheid en Milieu (National Institute for Public Health and the Environment)	RIVM
Root Mean Squared Error	RMSE
Sentinel-5 Precursor	S5P
Severe Acute Respiratory Syndrome Coronavirus 2	SARS-CoV-2
Sulfur dioxide	SO_2
Tropospheric Monitoring Instrument	TROPOMI
World Health Organization	WHO

Table of contents

1 Introduction

1.1 Internship organization background

Geodan was established more than 30 years ago by members of the *Vrije Universiteit Amsterdam*. Nowadays it is one of the largest geographic information technology companies in Europe. *Geodan* collects, combines, visualizes, and analyzes data such as proprietary, open source, contemporary, or historical data. The private company creates new spatial insights by connecting different data sources in innovative ways with a focus on usability and smart design. *Geodan* is academically oriented, and has a large research department constantly innovating. Their transparent approach to research, big data capabilities, and contribution to the public domain aligns well with the topic of this internship research.

1.2 Context and justification

Air pollution is a real problem in the era of climate change (Ramanathan and Feng 2009; Seinfeld and Pandis 1998; Silva et al. 2013) and is detrimental to our health (Brunekreef and Holgate 2002; Kampa and Castanas 2008). Also the Netherlands experiences moderate to bad air quality (EEA 2019). Different pollutants concentrate in the air, including particulate matter (PM) and nitrogen dioxide (NO_2). They have an adverse effect on our health (WHO 2013), where the first mainly originates from agriculture and the latter from road traffic (Temam et al. 2017). The air pollutants are measured by the *National Air Quality Measuring Network* (*Landelijk Meetnet Luchtkwaliteit*, LML) using ground sensors scattered over the Netherlands (Luchtmeetnet.nl 2020; RIVM 2020a), and more recently by the *Sentinel-5 Precursor* (S5P) satellite equipped with the *Tropospheric Monitoring Instrument* (TROPOMI) (ESA 2020; Griffin et al. 2019; Veefkind et al. 2012). These sources provide a quantifiable insight into our air quality. Namely the LML can be used for decentralized decision making as stated by the *Environmental Act* (*Omgevingswet) 2021* (Rijkswaterstaat 2020) equipped with official interpolation models (Manders et al. 2017; Wesseling et al. 2019) that conform to European guidelines of informing the public (The European Parliament and The Council Of The European Union 2008). Further investigating this data and its relationship to other factors, especially on a local scale, could provide decision makers with the right tools for future actions. This is also relevant to the greater public, which must be provided with accurate and consistent information. And more recently, an extraordinary anomaly in our modern history saw human behavior change dramatically with many unanswered questions about its impact on our local air quality.

1.2.1 Significance

This anomaly concerns the dramatic change in our society and economy in the year 2020 in response to the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) leading to the corresponding coronavirus disease (COVID-19). Major changes in our behavior were caused by national and local lockdowns enforced by governing bodies. Did this affect our air quality and to what degree was this our influence? Since this is a recent occurrence, a great opportunity arises to investigate this relatively unanswered question. It also provides an insight into the relationship and influence of different factors potentially affecting air pollution by comparing patterns and trends before and during this unique period.

1.2.2 Previous research

Air quality is a frequently discussed topic. Using Google Scholar we get 78,600 results with the keyword “air pollution” in the title and 48,500 results with “air quality”. Combining this with the keyword “COVID” gives us 143 results as of the writing of this report, and 10 results using “coronavirus”. A large part of the search results look at health and transmission problems associated with COVID-19 and air quality. Accounting for this only a handful of research papers and reports assess the effects of COVID-19 measures on air quality. Satellite measurements saw reduced air pollutant concentrations in China as early as February 2020 (EO 2020). This was followed globally by a 60% and 31% reduction in NO_2 and PM respectively measured by satellites and ground sensors in 34 countries during COVID-19 lockdowns (Venter et al. 2020). In the Netherlands the *National Institute for Public Health and the Environment* (RIVM) is more nuanced based on ground sensor measurements

(RIVM 2020b), but they acknowledge the effects of a decrease in traffic intensity on air quality. The *Royal Netherlands Meteorological Institute* (KNMI) did provide a more conclusive answer suggesting a 20% to 38% NO_2 drop in Western Europe by assessing satellite measurements (Bauwens et al. 2020).

1.2.3 Research contribution

Previous research exclusively looked at a European scale and had limited data points in the Netherlands. Moreover, earlier research does suggest that air pollution has a relationship with factors such as weather (Battista and de Lieto Vollaro 2017; Hargreaves et al. 2005; Weiner and Matthews 2003) and road traffic (Bohemen and Janssen van de Laak 2003; Comert et al. 2020; Jiang et al. 2019) where recent drops also coincided with stagnating vehicle transportation (Venter et al. 2020). However, information about COVID-19 related changes in the Netherlands remains limited. An assessment should be made that utilizes local air quality, road traffic and weather data, potentially with the help of remote sensing data, to investigate their relationships, importance, and trends. This is especially relevant given the rare abnormalities 2020 provides.

1.3 Research objective

To assess air quality, we need to look at the datasets used to measure this variable. The LML ground measurements are a good candidate as it provides hourly measurements for numerous years. It is also used as the official air quality indicator by the Dutch government and the European Union. The more recent S5P satellite can provide additional information, but is more limited in historic data and less established in terms of usage. Local weather (KNMI 2020) and road traffic data (NDW 2020) are spatiotemporally similar to the air quality data and readily available in the *Geodan* servers. This concerns real time hourly weather data from stations by the KNMI and minutely traffic measurements using road sensors by the *National Data Warehouse for Traffic Information* (*Nationale Databank Wegverkeersgegevens*, NDW). They are ideal candidates to be used as influential factors to assess their relationship and importance to air quality. The road traffic data is especially relevant as this may provide a measurable link between air quality and its potential human influence. The relationship is initially linearly assessed. However, to understand the nonlinear effects these factors have on air quality, a more advanced method should assess their level of importance. Namely by using machine learning, aiming to predict air quality using the mentioned factors. Assessing the effects by these factors on the predicted outcomes can tell us something about their importance within a complex web of interconnectivity. Doing this in the context of the COVID-19 lockdowns adds a unique comparison. Lastly, an analysis on the air quality trends could verify the local decrease of air pollution. This decrease is compared to weather and road traffic to support their established relationships and importance. This is initially performed on data in and around Amsterdam as many data points are contained within this area. The majority of atmospheric NO_2 finds its source by road traffic and is the single largest environmental health risk in Europe (Cattaneo 2019). Therefore, the focus will be on this component.

1.3.1 Research questions

The above mentioned aim of the research can be summarized to the following research questions:

1) How does ground (LML) / satellite (S5P) based air quality, road traffic (NDW) and weather (KNMI) data compare to each other?
    a) What are the specifications of each dataset?
    b) How should the data be prepared?
    c) Do the datasets linearly relate to each other?
2) Can S5P, NDW and KNMI data predict hourly LML air quality data using machine learning?
    a) What is the prediction accuracy of the machine learning model?
    b) Which training variables are important to maintain this accuracy?
    c) Does the prediction accuracy and variable importance change after excluding training data during the COVID-19 lockdown and what can this tell us?
    d) Can S5P data improve the prediction accuracy?
3) To what degree do values in all datasets deviate from the normal trend during the COVID-19 lockdown?

2 Methodology

2.1 Scientific procedures

The research in this report applies an empirical approach as the required data is readily available in a raw format instead of as scientifically verified data. Preprocessing steps are implemented to reduce inconsistencies, heterogeneity, and outliers in the data. This makes the process somewhat heuristic as no statistical inquiry is performed to the quality of the data and because the research is conducted with limited time. We assume that the preprocessed NO_2, weather, and road traffic data is homogeneous throughout space and time. They may not reflect the entirety of the research area due to their locality and inaccuracies may be present, but it is the closest available data to address the research questions and the focus lies on their relationships and influences, eliminating the need for absolute accuracies. Initially for research question one, a basic specification table is constructed for the datasets. The datasets are systematically preprocessed for optimal compatibility. This is followed by a linear regression analysis between each variable within the datasets. This provides an initial indication of their relationship, and how that may translate into their function and importance to predict air quality. The predictions showcase the strength of each variable using machine learning. In this case by the usage of a robust and general purpose random forest regressor (RFR). The advantage of this architecture is that it allows an easy deconstruction of the variable branching leading to the predictions in contrast to for example a deep neural network "black box" approach. However, for a more general and unbiased assessment of the variable importance, a permutation technique is implemented. The importance of each variable is addressed including and excluding data within COVID-19 lockdowns to answer the second research question. The same comparison is made where S5P data is included and excluded. The accuracies of the RFR are assessed using cross validation to avoid a temporal bias. The third research question provides a more descriptive overview of the general patterns and trends within each variable. For this a seasonal trend analysis is implemented that uses linear harmonics to fit a curve to normalize for seasonality. Deviations from this curve and historical differences are quantified to assess any disruption in the seasonal trend. A detailed approach to answer the research questions is presented in the paragraphs below.

2.2 Materials

2.2.1 Hardware

A laptop is provided by *Geodan* with the necessary accessories. This mainly serves as a portal to a powerful computational server housing the hardware necessary to conduct this research. The server stores and processes the data, and provides coding environments with customizable dependencies. The specifications consist of a 32 processing core *AMD* threadripper with 100 gigabyte random access memory and fast solid state drives that enables both high throughput and high performance computing.

2.2.2 Software

On the servers of *Geodan*, coding environments and databases are accessible through terminals and other applications. Coding is mainly performed in a *Python Anaconda* environment (Anaconda Inc 2020), written in *Jupyter* notebooks housed in *JupyterLab* (Project Jupyter 2020). External data is piped to databases on the server to be accessible directly from the notebooks. The main database management system is *Clickhouse* (Yandex LLC 2020). Additionally, application programming interfaces (APIs) to databases and tools are used for research purposes. Namely, the *Google Earth Engine* (GEE) *Python* API for remote sensing data (Google 2020). Important *Python* libraries for machine learning and statistics are *scikit-learn* (Pedregosa et al. 2011) and *SciPy* (SciPy developers 2020), where *NumPy* (NumPy 2020) and *pandas* (pandas 2020) are used for data preparation and structuring, and lastly *Bokeh* (Bokeh contributors 2020) and *Folium* (Story 2013) are responsible for the visualizations.

2.3 Methods

2.3.1 Research question 1 - Data comparison

Dataset specifications

It starts with a literature research of the LML and S5P air quality datasets and their specifications (e.g. availability, sensor type, capturing method, product level, spatiotemporal resolution, coverage, bands). This is expanded by the KNMI weather station data and NDW road traffic data in the Netherlands. A similar spatiotemporal resolution is important for further assessment. The outcome of this procedure is a table with the datasets and their specifications based on metainformation and technical properties (see table 1 to 3). Relevant variables within these datasets can be seen in figure 1.

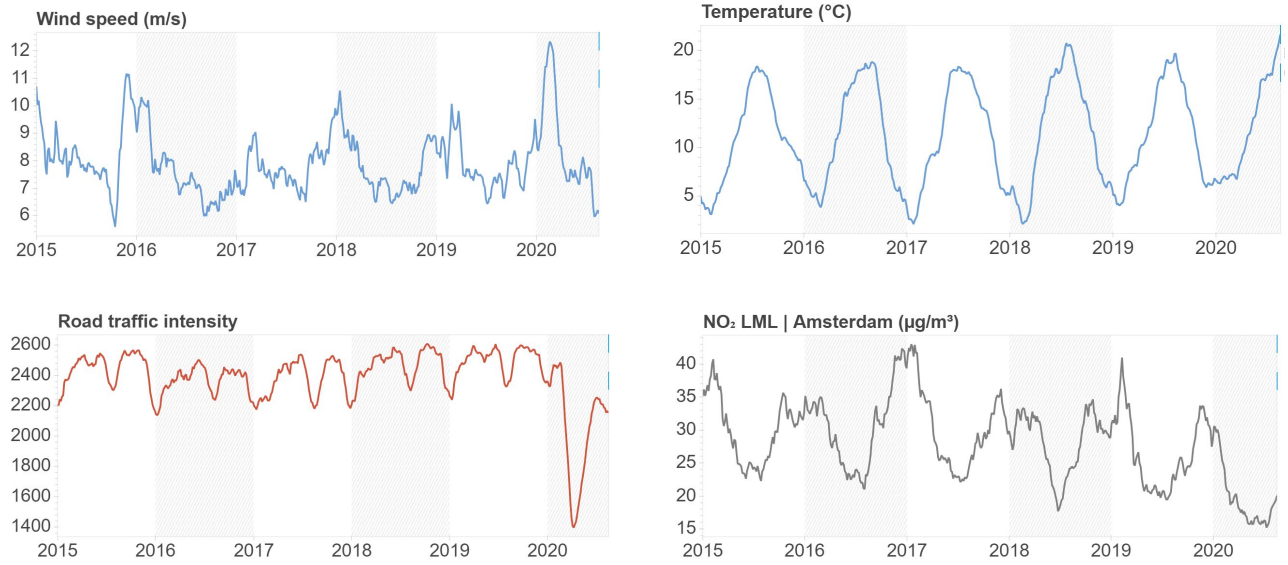


Figure 1 Smoothed time series (rolling mean window size of 40 days) of important variables within the selected datasets

Preparation

The datasets will be stored, preprocessed, and spatiotemporally matched. They are loaded into the *Clickhouse* database and accessed from the *Jupyter* notebook. The data points are selected based on spatial proximity and aggregated into similar time intervals. In this case it will be in and around Amsterdam with a date range of 1 January 2015 to 18 August 2020. The date range was chosen as this encompassed the available data with the exception of the S5P dataset. The LML NO_2 sensors are selected by the sensors within a 12 km radius (i.e. 18 sensors) from to the inner city center of Amsterdam. The farthest sensor is located near the A2 close to Breukelen. The KNMI data is taken by an automated weather station located at Schiphol. The available weather components are wind direction (°), wind speed (m/s), temperature (°C), horizontal view (m), air pressure (0.1 hPa), moisture (%), and dew point temperature (°C). The NDW data will be the mean of 10 road traffic sensors on major roads in and around Amsterdam. Here the road traffic speed (avg km/h over one minute) and intensity (sum vehicles/h over one minute) are taken. The selected sensors can be seen in figure 2. The S5P data will be the average value of a 10 km buffer radius around the center point of Amsterdam. It will encompass several cells of the S5P data. This is one variable indicated as NO_2 S5P | Amsterdam. The 18 LML air quality sensors will be individually assessed, but also by their total mean. The individually assessed sensors are indicated by the pollutant name, sensor type, and the corresponding location, e.g. NO_2 LML | Amsterdam-Oude Schans. Their total mean is indicated as NO_2 LML | Amsterdam.

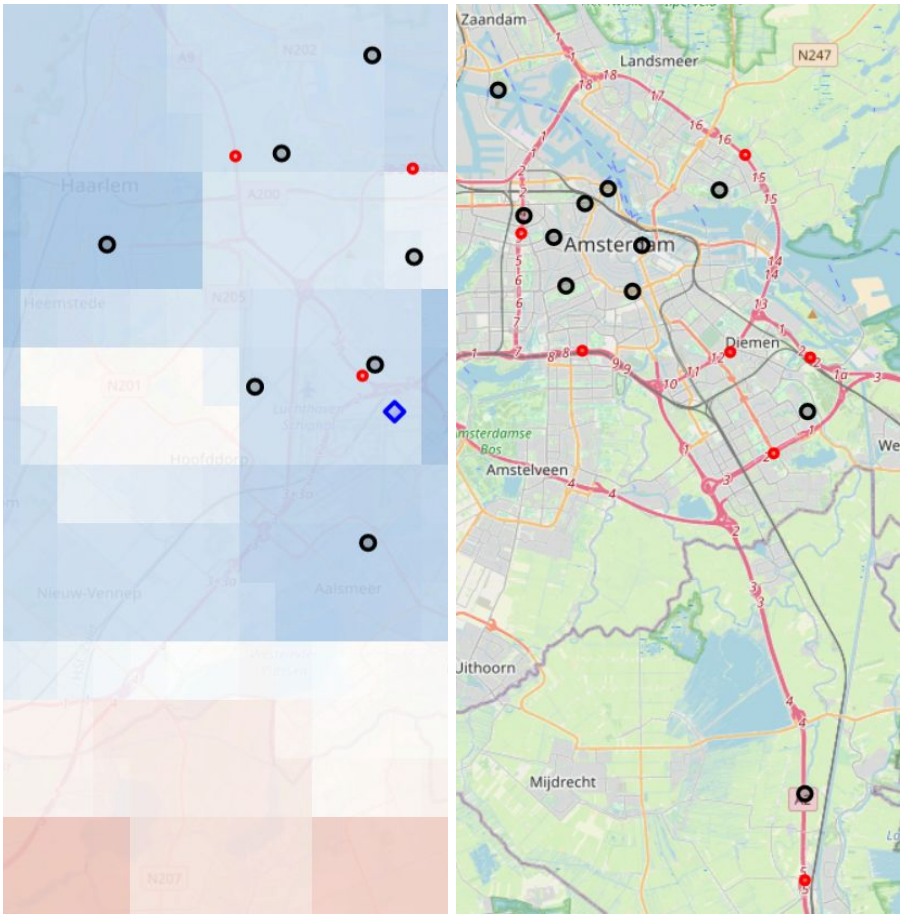


Figure 2 Map of the selected data points: air quality sensors (black circles, 18 in total), road traffic sensors (red circles, 10 in total), and a weather station (blue diamond). A single raster image of the S5P is overlaid on the left side of the map.

After selecting the data points and variables, preprocessing steps are implemented that optimize the relationships between the datasets. All variables in the datasets are aggregated to hourly intervals if there is more than one data point in an hour. Outliers in the data are filtered by the 0.001st and .999th percentile. Then the datasets are smoothed using a centered rolling mean with a window size of 6 hours (see figure 5 and 6). The weather variables receive extra smoothing as to reduce their volatility (see figure 7). Temporal information such as the hour of day, day of the year and total hours are included but will not be smoothed or filtered. The LML air quality values are exponential by nature thus log transformed to form a normal distribution (see figure 13). This allows for an improved linear relationship to the other normally distributed values and also improves the prediction performance. Cyclical variables have an extra preprocessing step where they are reduced to a machine interpretable format. In the case of the human readable clock of 24 hours a conversion is made to the trigonometric functions sine and cosine. This splits the value in two parts, which is less readable for humans but gains an increased interpretability by machines. The same applies for the degrees in which the meteorological wind direction is indicated, of which the final conversion can be seen in figure 3 and 8. Converting this data vastly improves their relationship to other variables and the prediction performance.

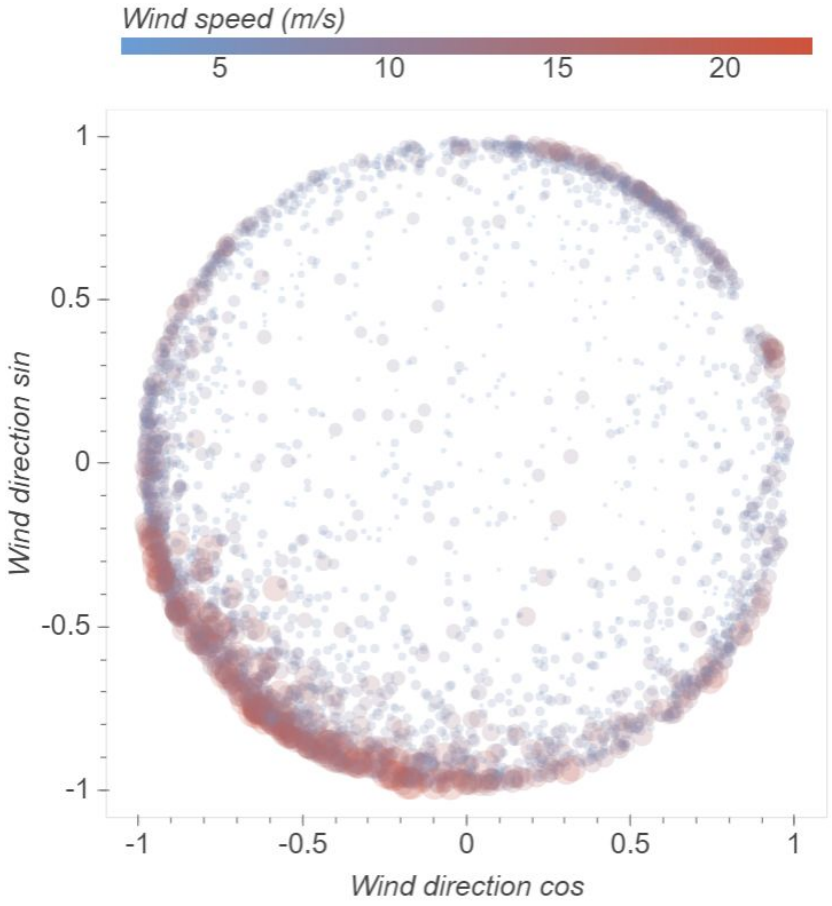


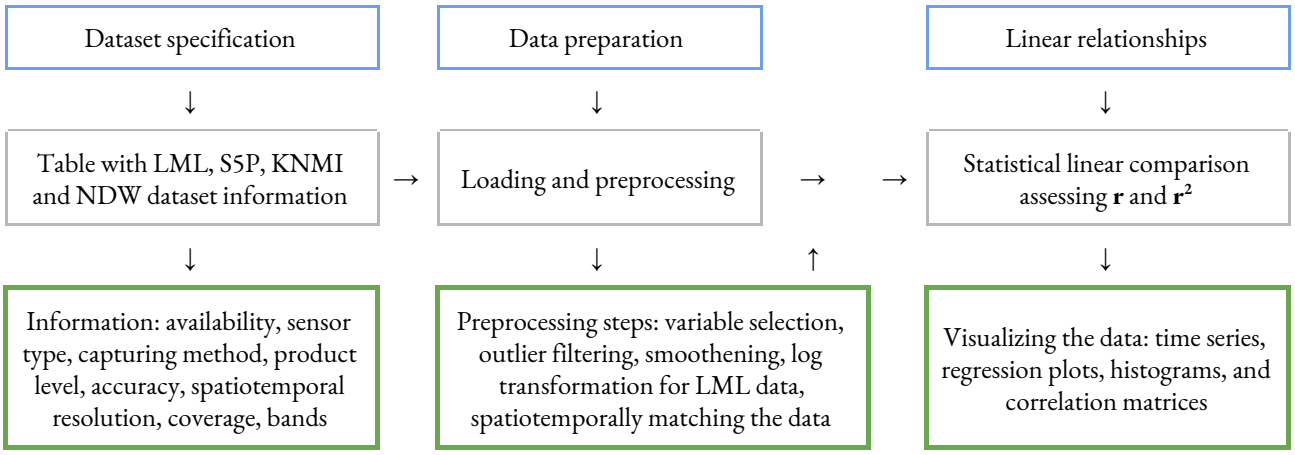
Figure 3 Converted cyclical wind direction (°) to sine and cosine. Note that the points are not on a perfect circle, this is the results of the data smoothing.

The S5P data has temporally sparse data, and is thus interpolated with a polynomial fit to the second degree. This data is not log transformed as this already has a normal distribution. S5P data is measured over the whole tropospheric column, averaging the air pollutant vertically and horizontally. This is electromagnetically derived from certain wavelengths remotely measured by the *TROPOMI*. In comparison, the LML chemically measures NO_2 in $\mu g/m^3$. The S5P air quality data values are expressed as mol/m^2 over the whole column. A direct comparison cannot be made between the LML and S5P data as they are fundamentally different (Ialongo et al. 2020), thus remaining separate variables. Nevertheless, their relationships are assessed and the S5P will be investigated into its effectiveness in predicting air quality at the surface. For an easier reference, the S5P values are linearly fitted to the LML values. Finally, all variables are integrated as columns in a *pandas* dataframe indexed by the date and hour of the day.

Linear relationships

After preparing the data, a comparison is made by assessing their linear relationships. As an overview, a correlation matrix (see figure 9 and 10) is generated indicating the Pearson's correlation coefficient (**r** or **PCC**) and the coefficient of determination ($\mathbf{r^2}$). The relations between the NO_2 to the other variables are visualized (see figure 11 and 12). The full time series are converted to linear regression plots with marginal distributions linked between the mean of the LML sensors and the other variables (see figure 13). This provides a more intuitive understanding of their relationships.

Flowchart - Research question 1



2.3.2 Research question 2 - Predicting air quality

Prediction accuracy

The preprocessed data is fed into the RFR. See figure 4 of a schematic overview of one decision tree within a RFR of a limited depth. The *Python* library *scikit-learn* has an established method to execute this model (scikit-learn developers 2020b). The hyper parameterizations are the default values. Except for the maximum depth, which is changed to 15 and the random state is set to 1. The *scikit-learn* library also provides a cross validation method (scikit-learn developers 2020a) which is used to assess the prediction accuracy of the response variables. It uses 10 k-folds. The statistics to compare the observed and predicted values from the cross validation analysis are the **r**, **r²**, mean squared error (**MSE**) and root mean squared error (**RMSE**). The accuracies are listed in a table (see table 4) and various sections of the results are presented in a time series graph (see figure 16) and a regression plot with linked histograms to visualize the predictions coinciding with the observations (see figure 17).

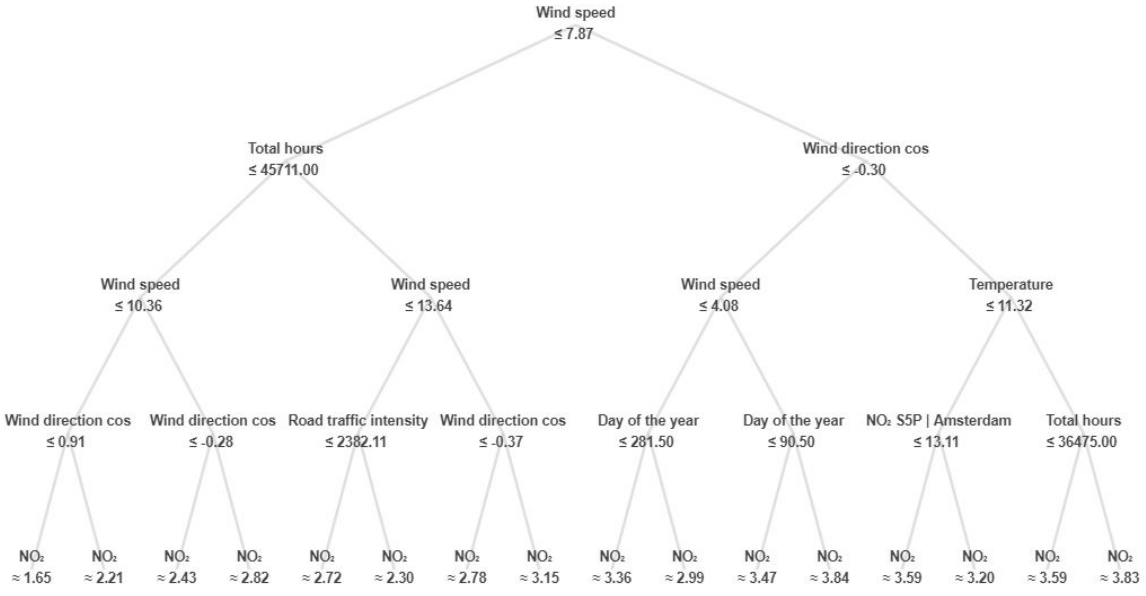


Figure 4 Schematic decision tree with a maximum depth of 5 layers.

Variable importance

The variable importance (i.e. feature importance) is assessed using a permutation method. This reigns superior to the classic build-in variable importance method that is tailored to a random forest and its hierarchical tree structure (Breiman 2001). The permutation importance, as found in the *scikit-learn* library, is a more generalized and less biased method that during training of a machine learning model randomizes one explanatory variable, runs the model, and assesses its contributed error (scikit-learn developers 2020c). If the error remains low, the explanatory variable reduces in importance. The variable importances are visualized using horizontal bars (see figure 14 and 15) and compared to the corresponding linear relationships of the previous research question. The variables with a low importance are removed, whereafter the prediction process is repeated. A 5% threshold of accuracy loss is held. The percentage differences of the prediction accuracies and variable importance are expressed in tables (see table 5 and 6).

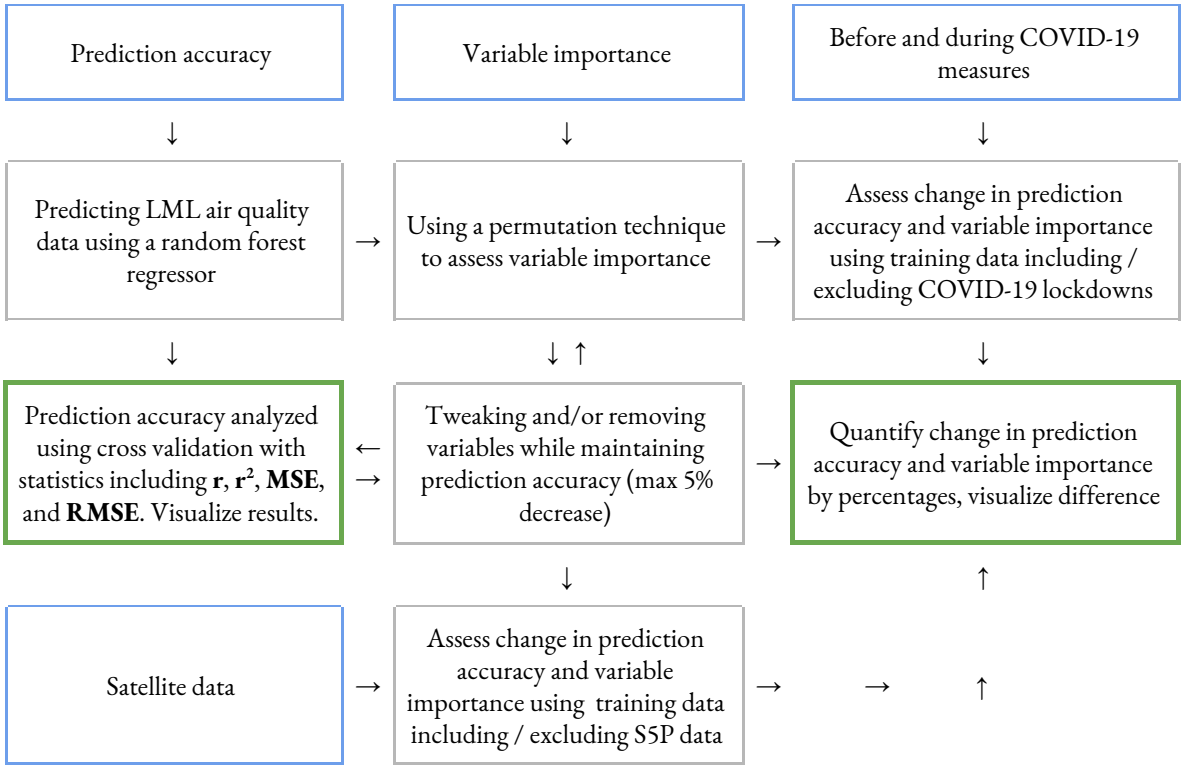
Before and during COVID-19 measures

Initially, the RFR is trained with data from 1 January 2015 to 18 August 2020. This includes a time period of national lockdowns in the Netherlands with the most restrictive ones enforced on 9 March 2020. As seen in figure 1, a major drop in road traffic is observed during this period. This data is quite different from the historical trend, which may influence the predictions. The RFR model is retrained excluding data from 2020 and its prediction accuracy (see table 7 and 8) and variable importance is compared to the previous runs (see table 9 and figure 18) where data within 2020 was included.

Satellite data

The S5P data used for training the prediction model lacks historical data before mid 2018. The absence of more historical data does not provide a fair comparison between the accuracy including and excluding the S5P variable as the empty data disturbs the machine learning model. Therefore, an extra comparison is made by retraining the prediction model with data from mid 2018 to 18 August 2020 where in one run S5P is included and the other excluded (see table 10 and 11 and figure 19).

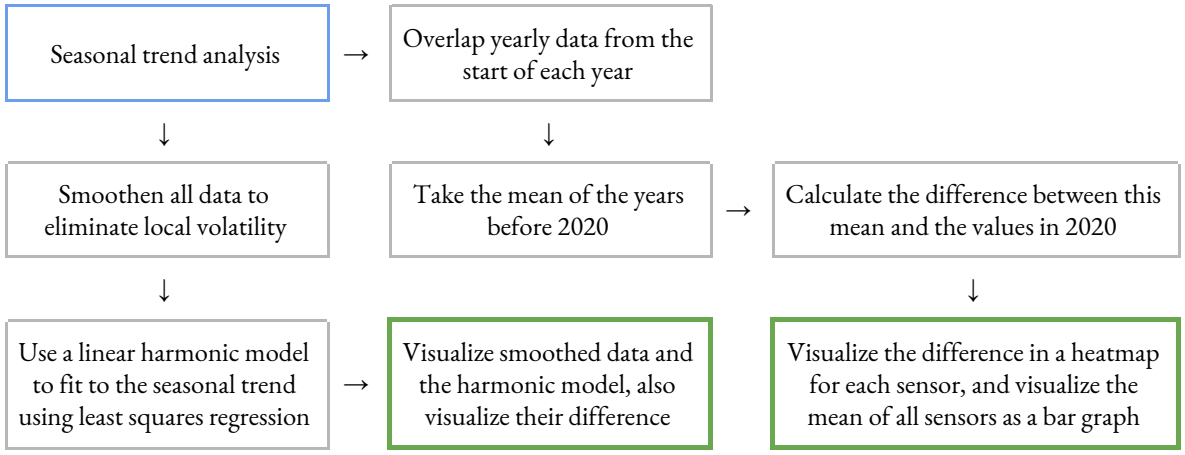
Flowchart - Research question 2



2.3.3 Research question 3 - Trends during COVID-19

Seasonality can be observed in the data as seen in figure 1. A linear harmonic model is therefore implemented to map this seasonality using equations 4.1 - 4.2 found in a book concerning time series analysis by Shumway and Stoffer (2017). Four waves are merged with biyearly, yearly, half yearly, and quarter yearly frequencies. These frequencies were found to align to the multi seasonal patterns in the data as seen in figure 1. The error and amplitude for each corresponding frequency wave is determined by the fit to each variable within the datasets. The four combined waves allow the mapping of multi seasonality in the data. Before fitting the convoluted-harmonic-linear-least-squares-regression-model to the data, a smoothing is applied using a centered rolling mean of 24 hours. This eliminates more volatility. For visualization purposes, a larger smoothing is performed with arbitrarily wide windows. The harmonic model, an extra linear equation and the seasonal smoothed data are visualized (see figure 20 and 21). The harmonic model is subtracted from the smoothed data to accentuate the difference (see figure 22). The trends are further analyzed using another method by taking a temporal section each year and overlaying them (see figure 23). This concerns the years 2015 to 2020. From this overlay, the mean value of each day of the years before 2020 is calculated. Hereafter, the ratio is taken by dividing the values in the year 2020 by the mean values of the years 2015-2019. This standardizes the deviations over time, accounting for seasonality and synchronized deviations. The ratio is expressed as a percentage in a table (see table 12) and visualized as a heatmap (see figure 24). The percentage deviations can be seen per variable within the data. This is also done for the other years (see figure 25). Finally, the dimensionality is reduced by averaging the variables to their mean, overlaying them in a bar chart (see figure 26) to get a better understanding of their relationship and comparative trends. Their long term relationships (see figure 27 and 28) are further investigated with a larger rolling mean window size including and excluding 2020.

Flowchart - Research question 3



2.4 Limitations and subsequent adaptations

One limitation is the lack of literature concerning the recent NO_2 changes due to the COVID-19 lockdowns and how this relates to other factors. This is momentarily still uncharted territory, especially on a local scale in the Netherlands. As seen in figure 1, the normal behavior of mainly the road traffic intensity is completely disrupted in the year 2020. A regular data comparison, prediction, and trend analysis using data during this time could skew expected outcomes. To adapt to this limitation, certain steps have to be repeated using data including and excluding data during 2020. This is performed in research question 2c. The same data isolation is implemented for the trend analysis in research question 3, where the harmonic model will be fitted to the data before 2020 and long term relationships are assessed including and excluding 2020 data. Even though this abnormality is a limitation, it also serves as an advantage to stress test the prediction model on overfitting. And more importantly, this unseen situation which could previously only be assessed hypothetically is now empirically available. It serves as a real world 'laboratory'. Subsequent limitations unrelated to COVID-19 lockdowns are empty values within the datasets

that could disrupt mainly the prediction accuracy, this is accounted for by removing all data rows that contain at least one empty cell ensuring that the model trains with complete data. Other limitations such as the limited S5P range are already accounted for in research question 2d and the preprocessing steps in research question 1b account for most data inconsistencies.

3 Results

3.1 Summary of the methodology

The relevant datasets will be listed in a specification table. Their data is loaded into a *Jupyter* notebook from the *Clickhouse* and GEE databases and further investigated using a *Python Anaconda* environment with scientific libraries. Preprocessing steps are data outlier filtering by the 0.001th and 0.999th percentile, smoothing by a rolling mean of 6 hours, averaging all geographic locations to one variable, as well as log transformation for NO_2 data, cyclical conversion of the wind direction and temporal variables, and polynomial interpolation for the S5P variable. The linear relationships between all variables are assessed using the **r** and $\mathbf{r^2}$ and presented in a correlation matrix. The correlations between the variables and the NO_2 LML variable are isolated as this relationship will be compared to the prediction results. The predictions are performed using an RFR, where the accuracies are assessed using an unbiased cross validator with 10 k-folds. The training (i.e. explanatory) variables within the NDW and KNMI datasets aim to predict NO_2 values of the LML dataset using data over the course of 1 January 2015 to 18 August 2020. The variable importance leading to the predictions are also assessed using an unbiased permutation technique. The same process is repeated for training data excluding data within the year 2020 to assess the potential impacts of the COVID-19 lockdowns on the variable importance and the prediction accuracy of the RFR. Another rerun with a variable importance and prediction accuracy assessment is performed where S5P data is included and excluded on training data that encompasses the range of the available S5P data from 1 Juli 2018. After the predictions a trend analysis is performed on relevant variables within the datasets. Mainly the NO_2 seasonal trends are modelled using a linear harmonic model which is compared to the actual measurements. Their difference is taken to accentuate potential deviations. This is also done for the road traffic intensity and weather variables as a comparison. Another method overlaps the values of similar temporal sections of each year. The years 2015 to 2019 are averaged, where after the rational difference between this average and the values found in 2020 are expressed as percentages. Finally, the mean of each year and month is taken of the NO_2, road traffic intensity and wind speed for a holistic comparison, and the long term linear correlations are assessed.

3.2 Findings

The findings provide the results in tabular and graphic forms to answer the research questions. The full *Python* scripts corresponding to the research questions can be found in this link: https://github.com/MarnixHamelberg/geodan.

3.2.1 Research question 1 - Data comparison

Dataset specifications

Four datasets are collected and stored. Relevant metadata is listed in tables 1 to 3.

Table 1 Selected datasets for research

Abbr	Name	Description	Application
LML	Landelijk Meetnet Luchtkwaliteit (LML)	Official and verified ground air quality measurement stations scattered over the Netherlands.	Surface air quality
S5P	Sentinel-5 Precursor (S5P)	Sentinel-5 Precursor is a satellite launched on 13 October 2017 by the European Space Agency to monitor air pollution. The onboard sensor is frequently referred to as Tropomi (TROPOspheric Monitoring Instrument).	Atmospheric air quality / weather (cloud fraction)
KNMI	Weather stations of The Royal Netherlands Meteorological Institute (KNMI)	The Royal Netherlands Meteorological Institute (KNMI) is the Dutch national weather service. Primary tasks of KNMI are weather forecasting and monitoring of weather, climate, air quality and seismic activity. KNMI is also the national research and information centre for meteorology, climate, air quality, and seismology.	Surface weather properties
NDW	Road sensor data of the National Data Warehouse for Traffic Information (NDW)	NDW has 19 public authorities working together on collecting, storing and distributing traffic data. This data is used to provide traffic information, to ensure effective traffic management, and to conduct accurate traffic analyses.	Road traffic properties

Table 2 Information dataset properties

Abbr	Producer/distributor	Source	Data type	Access	Spatial extent	Spatial resolution	Temporal extent	Temporal resolution	Last update
LML	RIVM/Luchtmeetnet/GGD Amsterdam/DCMR/ODRA/OMWB/RUDZL	https://www.luchtmeetnet.nl/	Vector	Open source	National	100 sensors at various locations, but more dense at major cities	2014-01-01 to present / Varies per sensor	Hourly	Present
S5P	European Union (EU) / European Space Agency (ESA) / Copernicus Programme	https://sentinel.esa.int/web/sentinel/missions/sentinel-5p/data-products	Raster	Open Source	Global	0.01 arc degrees (~7 km × ~3.5 km)	2018-07-10 to present	1-3 days	Present
KNMI	KNMI	http://projects.knmi.nl/klimatologie/uurgegevens	Vector	Open source	National	34 stations on land with a maximum distance of 30 km each	1951-01-01 to present	Hourly	Present
NDW	NDW consisting of the national government (Rijkswaterstaat), and provincial, city and municipal authorities	https://www.ndw.nu/pagina/nl/4/databank/	Vector	Open source	National	20,000 sensors on 4,300 km provincial, 3,400 km governmental and 3,000 km municipal roads	2009-06-01 to present / Varies per sensor	Minutely	Present

Table 3 Further information dataset properties

Abbr	Capturing method	Sensor type	Product level	Components / bands
LML	Ground sensor / station	Various automated chemical measurement methods		Mainly: Particulate Matter (PM10 and PM2.5), Ozone (O3) and Nitrogen Dioxide (NO_2).
S5P	Remotely (satellite)	Multispectral Instrument (MSI): TROPOMI (TROPOspheric Monitoring Instrument). Ultraviolet and visible (270–500 nm), near-infrared (675–775 nm) and shortwave infrared (2305–2385 nm) spectral bands.	Level 3	Cloud fraction (CF), Aerosol Index (AI), Carbon Monoxide (CO), Formaldehyde (CH2O), Nitrogen Dioxide (NO_2), Ozone (O3), sulphur dioxide (SO2), methane (CH4).
KNMI	Ground sensor / station	Various automated meteorological measurement methods		Mainly: Wind direction (DD), Wind speed (FF), Wind gusts (FX), Temperature (T), Dew point temperature (TD), Hourly total rain (RH), Air pressure (P), Horizontal view (VV), Cloudiness (N), Relative moisture (U)
NDW	Ground sensor / station	Automated subsurface induction loops		Traffic intensity (amount vehicles), traffic speed, estimated travel time, vehicle category

Preparation

After storing the data, each variable is loaded and preprocessed following the steps described in the methodology. The final preprocessed data compared to the raw data are visualized in figures 5 to 7 for a selection of variables. A window size of 6 hours of the rolling mean was empirically found to provide the best relationship and prediction results. Both smaller and larger windows sizes reduced accuracy.

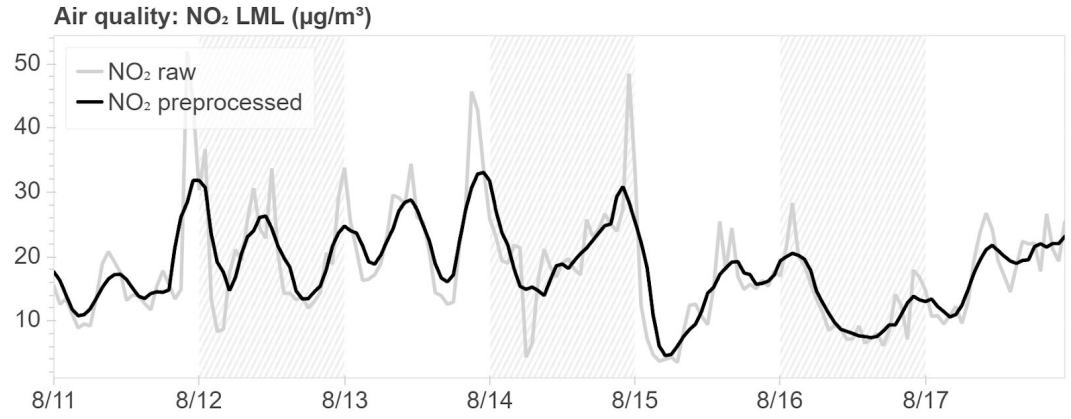


Figure 5 Time series of raw versus preprocessed air quality measurements during a section of 2020

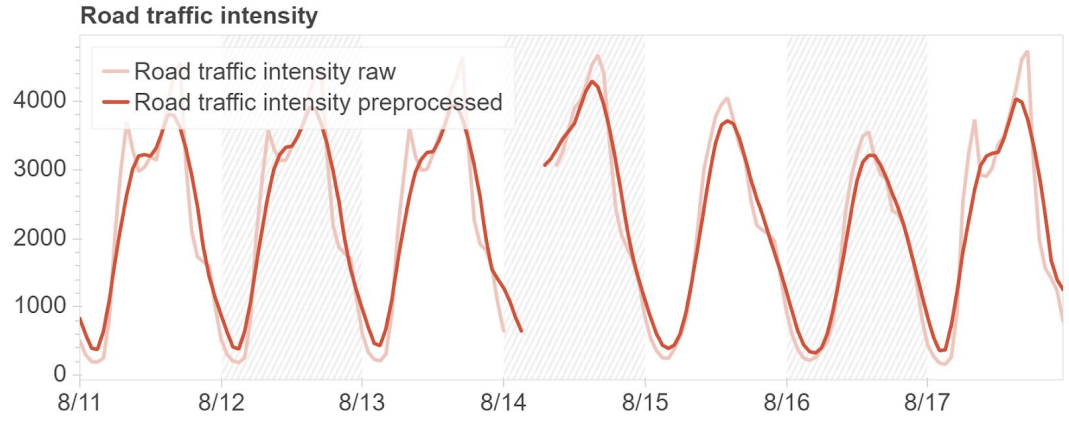


Figure 6 Time series of raw (with one extra step of averaging the minutely measurements to hourly data points) versus preprocessed road traffic intensity measurements during a section of 2020

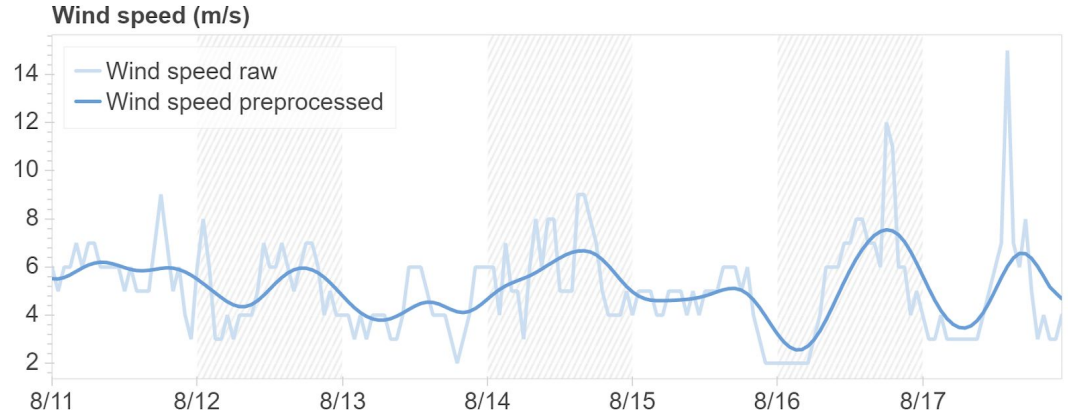


Figure 7 Time series of raw versus preprocessed wind speed measurements during a section of 2020

Figure 8 shows the conversion of the wind direction in degrees to sine and cosine. Now the jump between 360° and 1° is not drastic and represents the geometric distance. Especially noticeable is the sudden jump at the end of 15 August 2020. This sudden jump is accounted for by the trigonometric functions where the cyclical distance is accurately represented.

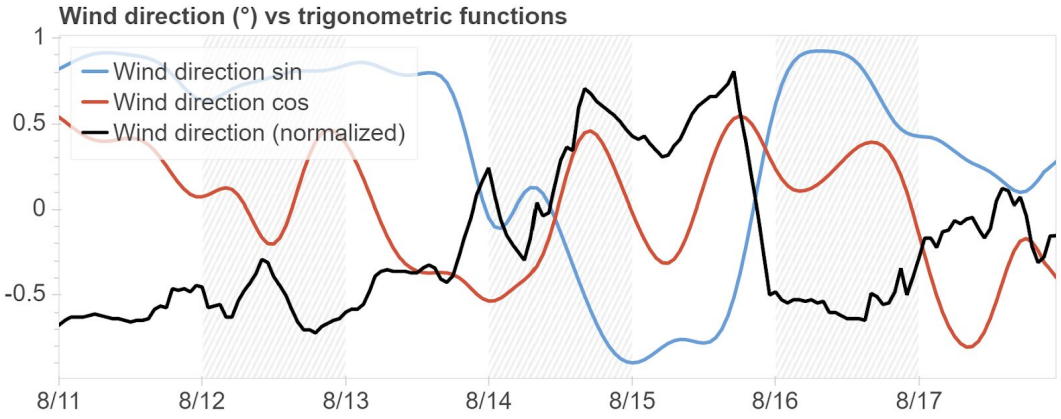


Figure 8 The cyclical wind direction variable compared to the sine and cosine conversions

Linear relationships

The relationships between all preprocessed variables within the datasets can be seen in the correlation matrix visualized as a heatmap in figure 9. This matrix shows the **r** and the correlation matrix in figure 10 shows the corresponding $\mathbf{r}^2$.

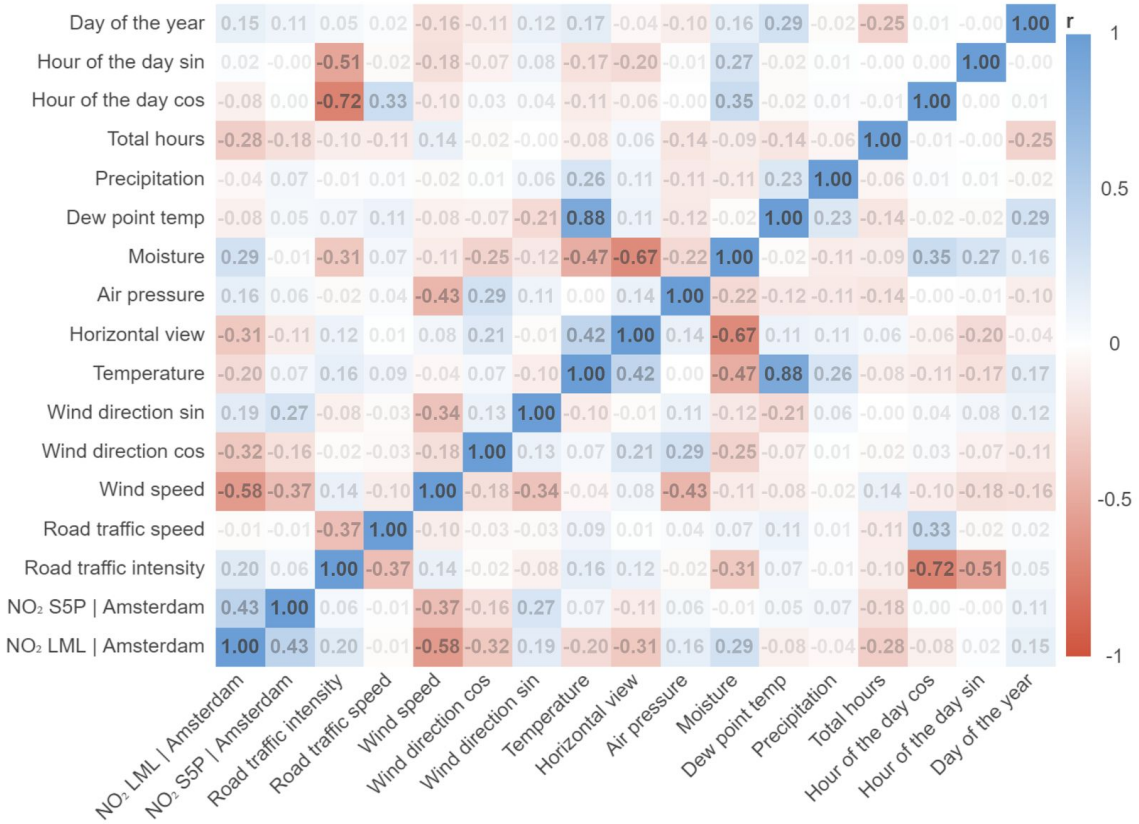


*Figure 9 Correlation matrix a as a heatmap showing the **r** of all variables in the datasets*

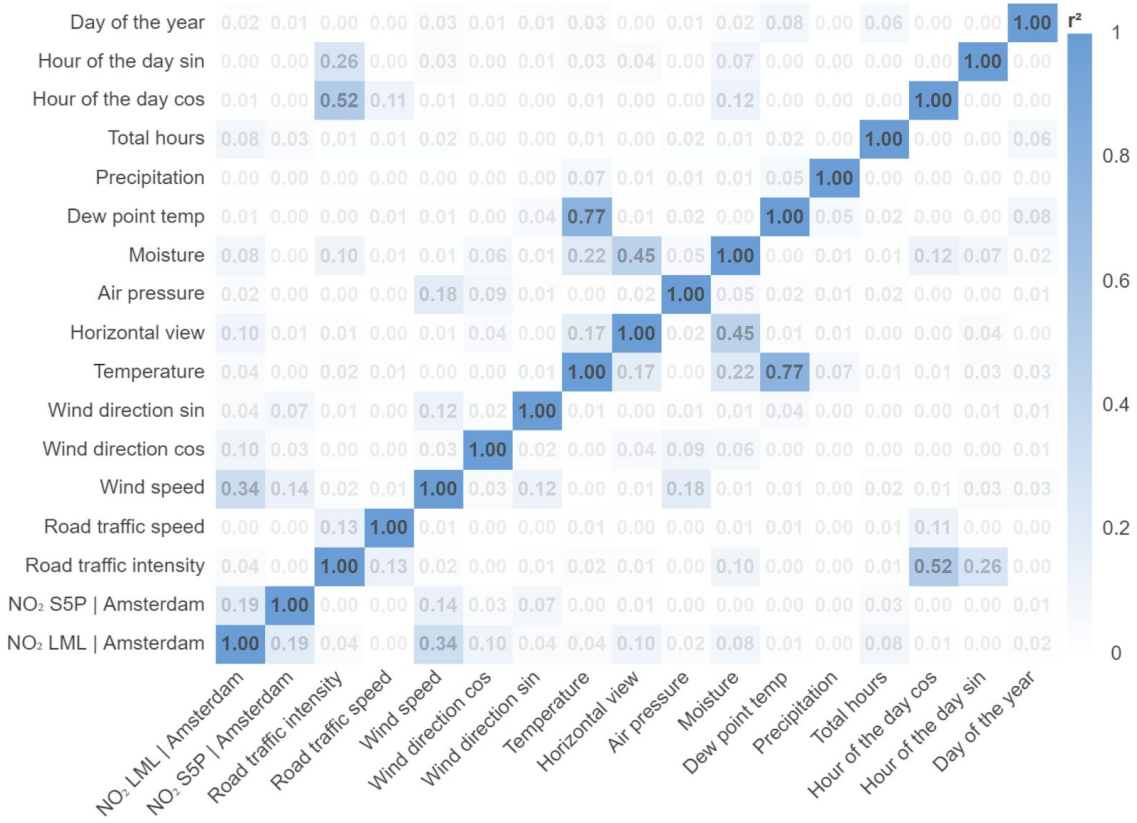


Figure 10 Correlation matrix a as a heatmap showing the r^2 *of all variables in the datasets*

Figure 11 and 12 visualize the first column of the correlation matrix, i.e. NO_2 LML | Amsterdam.

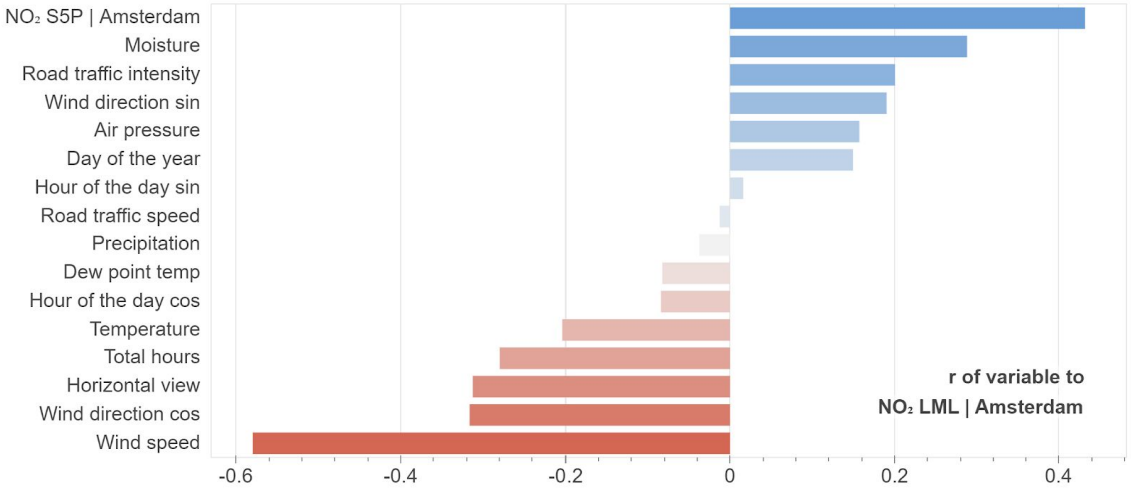


Figure 11 Bar chart of the r *of* NO_2 *LML | Amsterdam compared to other variables*

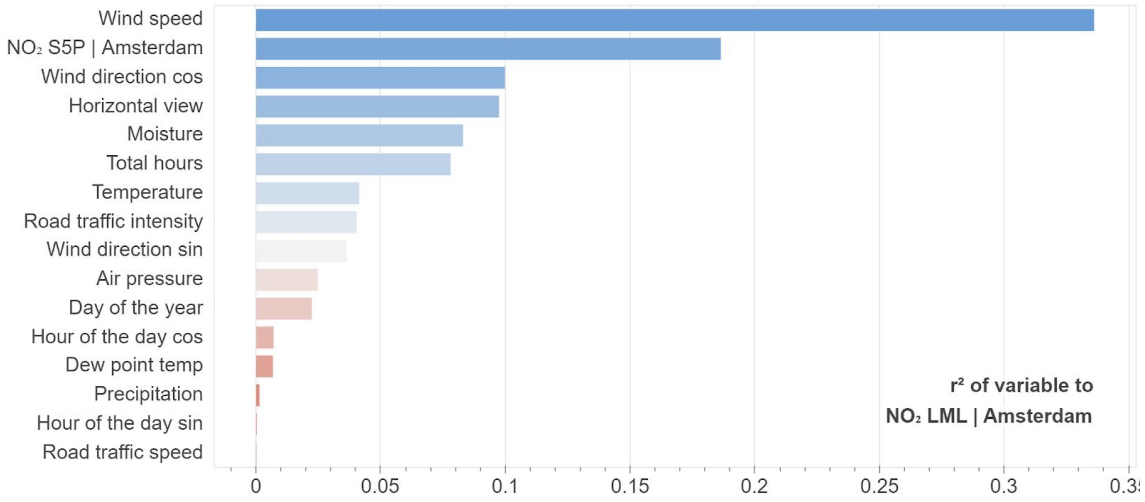
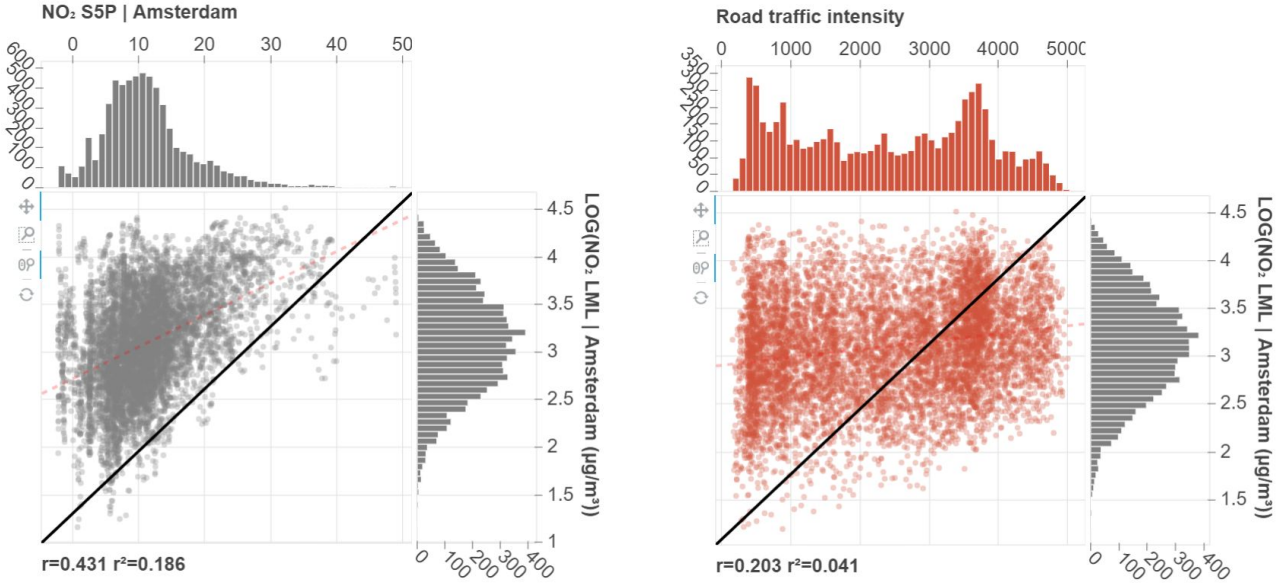


Figure 12 Bar chart of the r^2 of NO_2 LML | Amsterdam compared to other variables

The above figures are further visualized in figure 13 for some variables as a regression plot with linked histograms to visualize their relationship.



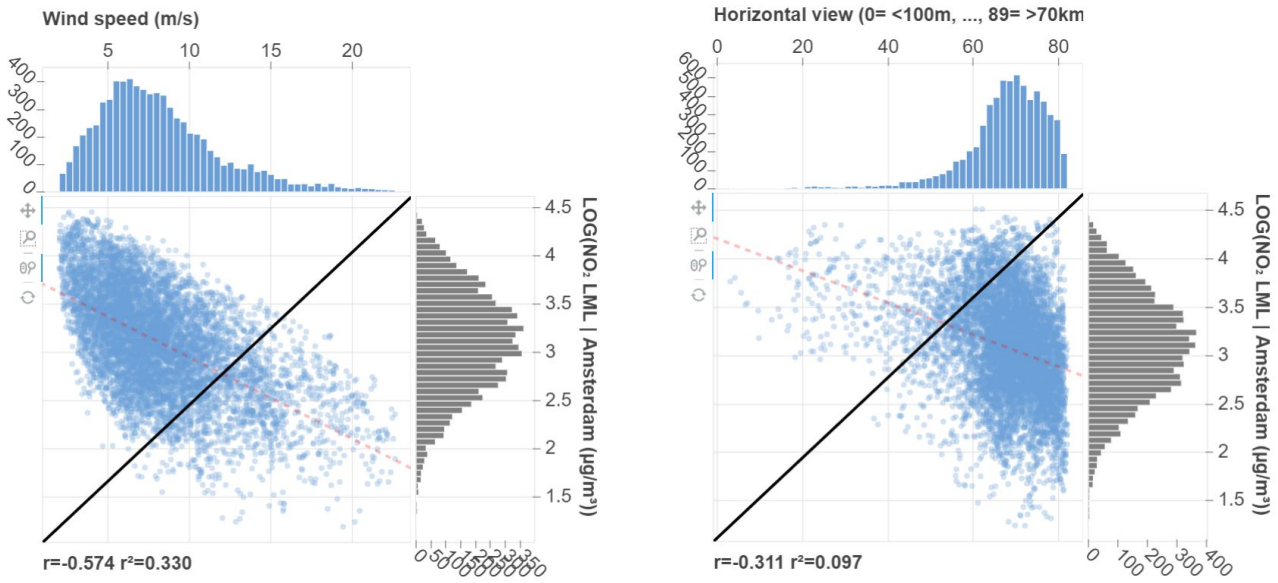


Figure 13 Regression plots with linked histograms of the relationship between the log transformed NO_2 and other variables.

3.2.2 Research question 2 - Predicting air quality

Prediction accuracy

After preprocessing the data, predictions are made using the RFR model. The accuracy results are per air quality sensor (ordered by distance to the first station) and can be seen in table 4. The total mean is calculated in the final row.

Table 4 Air quality prediction accuracy using all dataset variables

Location	r	-/+ 95% CI	r^2	-/+ 95% CI	MSE	-/+ 95% CI	RMSE	-/+ 95% CI
NO_2 LML | Amsterdam	0.854	0.107	0.732	0.172	0.063	0.050	0.248	0.084
NO_2 LML | Amsterdam-Oude Schans	0.777	0.168	0.610	0.241	0.112	0.106	0.327	0.142
NO_2 LML | Amsterdam-Stadhouderskade	0.766	0.170	0.593	0.242	0.085	0.076	0.285	0.112
NO_2 LML | Amsterdam-Van Diemenstraat	0.854	0.126	0.733	0.197	0.080	0.082	0.277	0.118
NO_2 LML | Amsterdam-Haarlemmerweg	0.828	0.092	0.688	0.150	0.079	0.056	0.277	0.097
NO_2 LML | Amsterdam-Jan van Galenstraat	0.844	0.112	0.716	0.174	0.072	0.062	0.264	0.095
NO_2 LML | Amsterdam-Vondelpark	0.772	0.165	0.603	0.237	0.117	0.091	0.337	0.117
NO_2 LML | Amsterdam-Nieuwendammerdijk	0.794	0.117	0.634	0.179	0.168	0.151	0.401	0.170
NO_2 LML | Amsterdam-Einsteinweg	0.891	0.062	0.794	0.109	0.069	0.057	0.258	0.094
NO_2 LML | Amsterdam-Ookmeer	0.815	0.104	0.667	0.164	0.195	0.152	0.434	0.164
NO_2 LML | Zaanstad-Hemkade	0.842	0.133	0.713	0.204	0.174	0.312	0.393	0.275
NO_2 LML | Amsterdam -Kantershof	0.805	0.087	0.650	0.139	0.129	0.037	0.359	0.051
NO_2 LML | Badhoevedorp-Sloterweg	0.835	0.092	0.700	0.149	0.113	0.075	0.332	0.104
NO_2 LML | Zaanstad-Hoogtij	0.868	0.060	0.755	0.101	0.140	0.049	0.373	0.062
NO_2 LML | Spaarnwoude-Machineweg	0.843	0.049	0.712	0.082	0.160	0.055	0.399	0.067
NO_2 LML | Hoofddorp-Hoofdweg	0.803	0.126	0.648	0.192	0.193	0.087	0.436	0.095
NO_2 LML | Oude Meer-Aalsmeerderdijk	0.806	0.100	0.652	0.155	0.124	0.047	0.350	0.062
NO_2 LML | Haarlem-Schipholweg	0.853	0.069	0.728	0.114	0.088	0.036	0.296	0.058
Total mean	**0.825**	**0.108**	**0.685**	**0.167**	**0.120**	**0.088**	**0.336**	**0.109**

Variable importance

The above predictions are made using all explanatory variables, of which the absolute values of their linear relationships and the permutation variable importance to the response variable can be seen in figure 14.

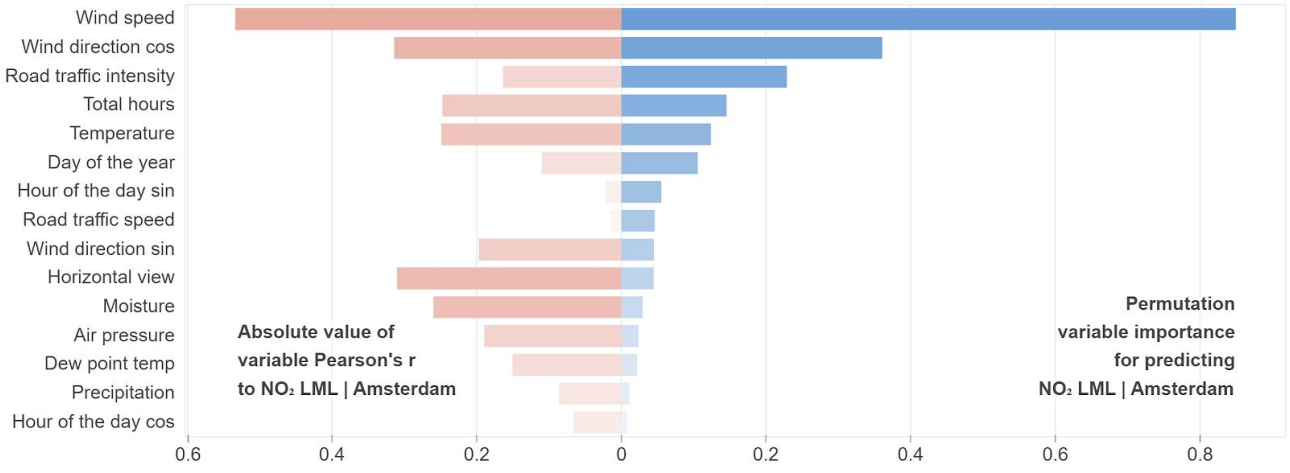


*Figure 14 The abs(**r**) and variable (i.e. feature) importance of the predictions*

The variables with a low importance are removed from the prediction model, while maintaining at least 95% of the original accuracies from the total mean. The new variable importance is displayed in figure 15 and new accuracies are presented in table 5 where table 6 shows the percentage difference from the initial accuracies.

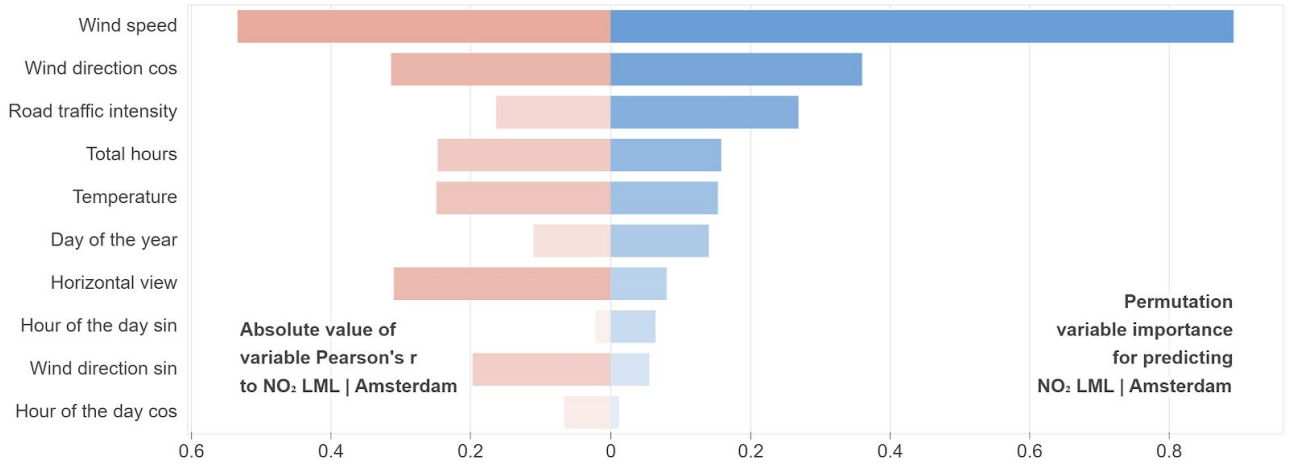


Figure 15 The variable importance of the predictions excluding unimportant variables

Table 5 NO_2 prediction accuracy using the selected dataset variables as seen in figure 15

Location	r	-/+ 95% CI	r^2	-/+ 95% CI	MSE	-/+ 95% CI	RMSE	-/+ 95% CI
NO_2 LML | Amsterdam	0.852	0.093	0.728	0.151	0.064	0.044	0.251	0.076
NO_2 LML | Amsterdam-Oude Schans	0.779	0.165	0.614	0.237	0.111	0.104	0.325	0.140
NO_2 LML | Amsterdam-Stadhouderskade	0.765	0.156	0.592	0.224	0.085	0.072	0.286	0.107
NO_2 LML | Amsterdam-Van Diemenstraat	0.851	0.119	0.728	0.186	0.082	0.079	0.280	0.114
NO_2 LML | Amsterdam-Haarlemmerweg	0.814	0.118	0.667	0.187	0.084	0.064	0.285	0.107
NO_2 LML | Amsterdam-Jan van Galenstraat	0.844	0.092	0.715	0.145	0.072	0.053	0.265	0.084
NO_2 LML | Amsterdam-Vondelpark	0.771	0.148	0.600	0.218	0.118	0.083	0.338	0.108
NO_2 LML | Amsterdam-Nieuwendammerdijk	0.805	0.102	0.651	0.158	0.157	0.121	0.390	0.142
NO_2 LML | Amsterdam-Einsteinweg	0.889	0.053	0.792	0.093	0.069	0.050	0.259	0.084
NO_2 LML | Amsterdam-Ookmeer	0.817	0.105	0.670	0.165	0.194	0.155	0.432	0.167
NO_2 LML | Zaanstad-Hemkade	0.841	0.134	0.711	0.205	0.175	0.312	0.395	0.275
NO_2 LML | Amsterdam -Kantershof	0.799	0.087	0.641	0.137	0.133	0.037	0.363	0.051
NO_2 LML | Badhoevedorp-Sloterweg	0.836	0.081	0.701	0.133	0.113	0.070	0.332	0.098
NO_2 LML | Zaanstad-Hoogtij	0.866	0.065	0.750	0.109	0.143	0.057	0.376	0.071
NO_2 LML | Spaarnwoude-Machineweg	0.841	0.048	0.707	0.079	0.163	0.058	0.402	0.069
NO_2 LML | Hoofddorp-Hoofdweg	0.803	0.126	0.648	0.192	0.193	0.088	0.436	0.096
NO_2 LML | Oude Meer-Aalsmeerderdijk	0.805	0.080	0.650	0.126	0.125	0.035	0.352	0.047
NO_2 LML | Haarlem-Schipholweg	0.853	0.058	0.729	0.097	0.088	0.031	0.296	0.050
Total mean	**0.824**	**0.102**	**0.683**	**0.158**	**0.120**	**0.084**	**0.337**	**0.105**

Table 6 Air quality prediction accuracy difference when predicting with all variables versus selected variables

Location	%Δ r	%Δ -/+ 95% CI	%Δ r^2	%Δ -/+ 95% CI	%Δ MSE	%Δ -/+ 95% CI	%Δ RMSE	%Δ -/+ 95% CI
NO_2 LML | Amsterdam	99.74%	86.72%	99.39%	87.56%	101.64%	89.01%	101.10%	90.44%
NO_2 LML | Amsterdam-Oude Schans	100.35%	98.23%	100.66%	98.41%	98.96%	98.75%	99.50%	98.89%
NO_2 LML | Amsterdam-Stadhouderskade	99.96%	91.95%	99.74%	92.50%	100.35%	94.85%	100.36%	95.33%
NO_2 LML | Amsterdam-Van Diemenstraat	99.75%	94.55%	99.45%	94.64%	101.46%	96.44%	100.91%	96.63%
NO_2 LML | Amsterdam-Haarlemmerweg	98.38%	128.37%	97.00%	124.99%	106.57%	113.90%	103.01%	110.48%
NO_2 LML | Amsterdam-Jan van Galenstraat	99.98%	82.21%	99.82%	83.58%	100.16%	86.64%	100.44%	88.52%
NO_2 LML | Amsterdam-Vondelpark	99.88%	90.06%	99.55%	91.69%	100.55%	91.34%	100.50%	92.53%
NO_2 LML | Amsterdam-Nieuwendammerdijk	101.43%	87.87%	102.74%	88.53%	93.75%	79.84%	97.38%	83.48%
NO_2 LML | Amsterdam-Einsteinweg	99.87%	84.84%	99.70%	85.22%	100.56%	87.56%	100.65%	88.62%
NO_2 LML | Amsterdam-Ookmeer	100.23%	100.66%	100.46%	100.56%	99.46%	101.73%	99.66%	101.72%
NO_2 LML | Zaanstad-Hemkade	99.83%	101.02%	99.67%	100.86%	100.64%	100.26%	100.36%	99.97%
NO_2 LML | Amsterdam -Kantershof	99.33%	99.43%	98.66%	98.60%	102.69%	102.29%	101.34%	100.19%
NO_2 LML | Badhoevedorp-Sloterweg	100.11%	88.01%	100.15%	89.18%	99.69%	93.12%	99.98%	94.05%
NO_2 LML | Zaanstad-Hoogtij	99.70%	108.16%	99.41%	107.55%	101.99%	115.97%	100.89%	114.66%
NO_2 LML | Spaarnwoude-Machineweg	99.71%	97.64%	99.42%	97.11%	101.62%	103.84%	100.80%	101.66%
NO_2 LML | Hoofddorp-Hoofdweg	99.99%	100.17%	99.98%	100.23%	100.00%	101.40%	99.99%	101.13%
NO_2 LML | Oude Meer-Aalsmeerderdijk	99.93%	80.08%	99.72%	81.29%	100.77%	75.18%	100.55%	76.30%
NO_2 LML | Haarlem-Schipholweg	100.05%	84.98%	100.06%	85.80%	99.84%	84.87%	100.05%	86.06%
Total mean	**99.89%**	**94.46%**	**99.73%**	**94.79%**	**100.31%**	**95.80%**	**100.33%**	**95.83%**

Figure 16 visualizes some sections of the predictions, displaying them in relation to the observed values. For some sections the prediction accuracy is poor, where others have high prediction accuracies as displayed in figure 17. Interesting to note is that the prediction accuracies are higher in the sections before the COVID-19 lockdowns, where the training and testing data are more similar. Keep in mind that for the true prediction performance the tables above should be used as these statistics were derived from the unbiased cross validator.

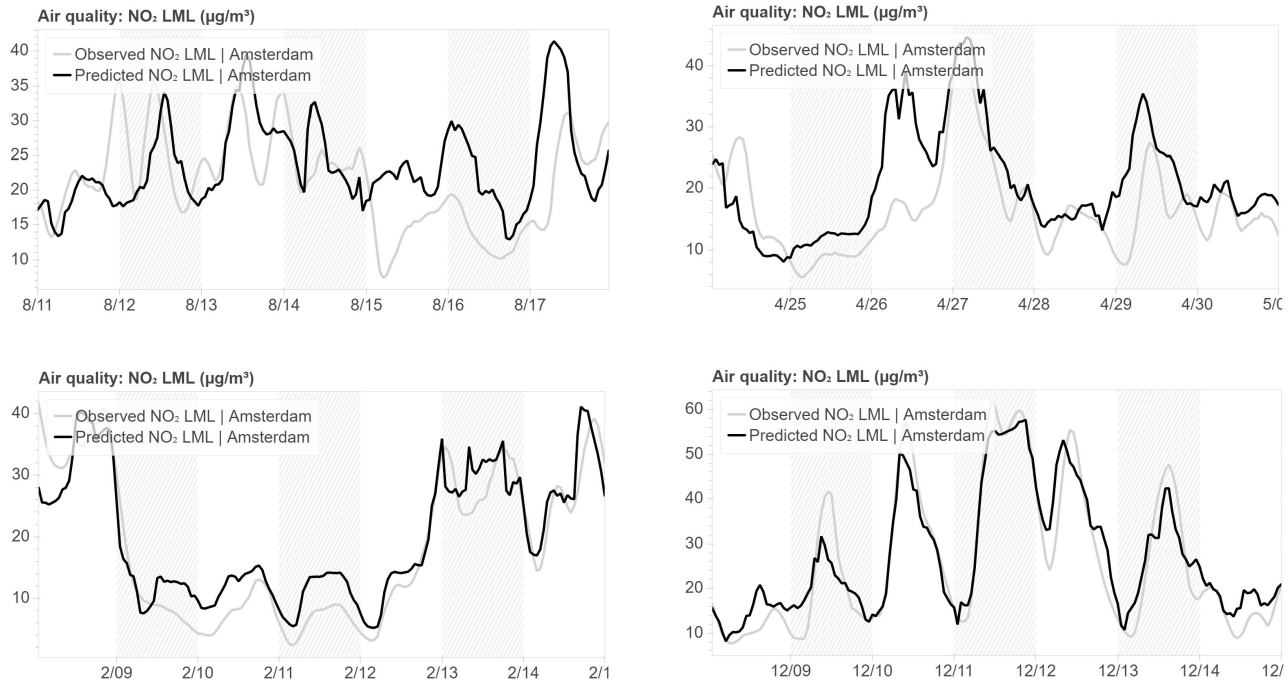
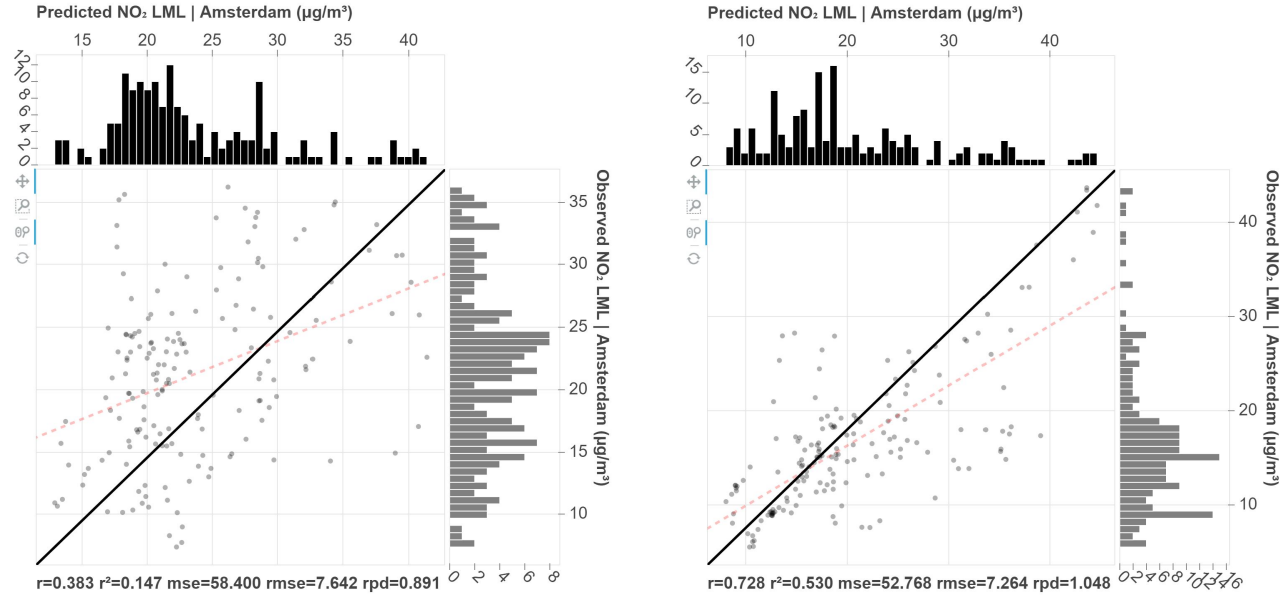


Figure 16 Four time series of different week long sections of NO_2 predictions vs observed values.



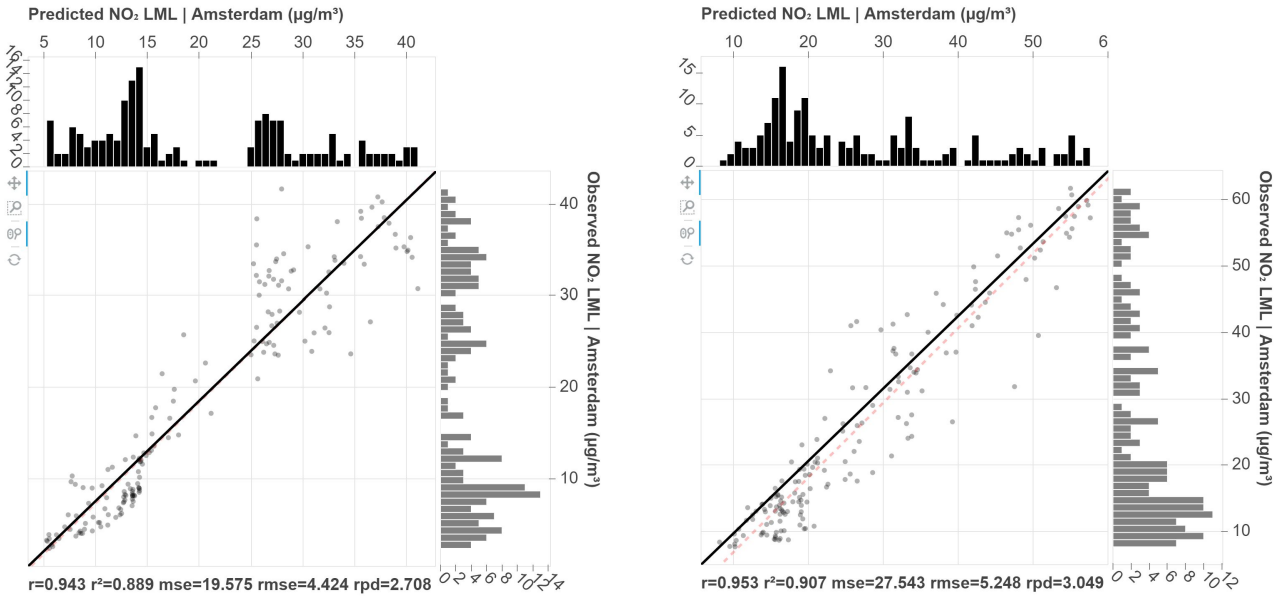


Figure 17 Four regression plots and linked histograms corresponding to the NO_2 values found in the time series in figure 16.

Before and during COVID-19 measures

The above mentioned prediction process is repeated using training data exclusively before the year 2020. The new accuracies can be found in table 7 and the percentage difference from the previous prediction accuracies are shown in table 8. The new variable importance graph is shown in figure 18 and the difference in variable importance indicated as percentages in table 9.

Table 7 Air quality prediction accuracy using the above selected dataset variables and testing/training data before March 2020

Location	r	-/+ 95% CI	r^2	-/+ 95% CI	MSE	-/+ 95% CI	RMSE	-/+ 95% CI
NO_2 LML | Amsterdam	0.877	0.031	0.769	0.053	0.056	0.014	0.235	0.030
NO_2 LML | Amsterdam-Oude Schans	0.818	0.072	0.670	0.115	0.094	0.047	0.305	0.071
NO_2 LML | Amsterdam-Stadhouderskade	0.813	0.032	0.661	0.052	0.069	0.022	0.261	0.041
NO_2 LML | Amsterdam-Van Diemenstraat	0.872	0.021	0.761	0.037	0.070	0.017	0.264	0.033
NO_2 LML | Amsterdam-Haarlemmerweg	0.832	0.075	0.694	0.125	0.075	0.051	0.270	0.088
NO_2 LML | Amsterdam-Jan van Galenstraat	0.860	0.033	0.740	0.056	0.064	0.012	0.252	0.023
NO_2 LML | Amsterdam-Vondelpark	0.806	0.085	0.652	0.135	0.102	0.028	0.319	0.043
NO_2 LML | Amsterdam-Nieuwendammerdijk	0.824	0.079	0.681	0.124	0.154	0.127	0.386	0.144
NO_2 LML | Amsterdam-Einsteinweg	0.891	0.062	0.794	0.106	0.065	0.035	0.254	0.064
NO_2 LML | Amsterdam-Ookmeer	0.845	0.058	0.715	0.097	0.172	0.126	0.409	0.138
NO_2 LML | Zaanstad-Hemkade	0.842	0.144	0.713	0.218	0.185	0.349	0.404	0.297
NO_2 LML | Amsterdam -Kantershof	0.816	0.070	0.666	0.110	0.132	0.047	0.363	0.063
NO_2 LML | Badhoevedorp-Sloterweg	0.857	0.066	0.735	0.111	0.104	0.064	0.319	0.088
NO_2 LML | Zaanstad-Hoogtij	0.879	0.026	0.773	0.046	0.137	0.040	0.369	0.054
NO_2 LML | Spaarnwoude-Machineweg	0.849	0.054	0.721	0.090	0.162	0.061	0.400	0.075
NO_2 LML | Hoofddorp-Hoofdweg	0.832	0.036	0.692	0.060	0.178	0.037	0.421	0.043
NO_2 LML | Oude Meer-Aalsmeerderdijk	0.826	0.053	0.683	0.087	0.119	0.025	0.344	0.037
NO_2 LML | Haarlem-Schipholweg	0.867	0.035	0.752	0.061	0.082	0.018	0.287	0.032
Total mean	**0.845**	**0.057**	**0.715**	**0.093**	**0.112**	**0.062**	**0.326**	**0.076**

Table 8 Air quality prediction accuracy difference when predicting NO_2 using testing/training data before and during 2020 (1 January 2015 to 18 August 2020) versus testing/training data before 2020 (1 January 2015 to 31 December 2019)

Location	%Δ r	%Δ -/+ 95% CI	%Δ r^2	%Δ -/+ 95% CI	%Δ MSE	%Δ -/+ 95% CI	%Δ RMSE	%Δ -/+ 95% CI
NO_2 LML | Amsterdam	102.93%	32.96%	105.67%	35.00%	86.36%	32.54%	93.81%	39.39%
NO_2 LML | Amsterdam-Oude Schans	104.95%	43.47%	109.14%	48.42%	85.15%	44.92%	93.77%	50.42%
NO_2 LML | Amsterdam-Stadhouderskade	106.21%	20.59%	111.70%	23.22%	80.97%	30.31%	91.26%	38.09%
NO_2 LML | Amsterdam-Van Diemenstraat	102.42%	17.78%	104.40%	19.81%	85.81%	22.09%	94.34%	28.98%
NO_2 LML | Amsterdam-Haarlemmerweg	102.18%	63.87%	104.07%	67.08%	88.77%	80.45%	94.59%	82.67%
NO_2 LML | Amsterdam-Jan van Galenstraat	101.85%	35.84%	103.47%	38.50%	88.33%	21.90%	95.07%	27.58%
NO_2 LML | Amsterdam-Vondelpark	104.57%	57.30%	108.66%	62.19%	86.92%	34.35%	94.20%	40.14%
NO_2 LML | Amsterdam-Nieuwendammerdijk	102.38%	76.91%	104.64%	78.05%	98.06%	104.90%	98.93%	101.72%
NO_2 LML | Amsterdam-Einsteinweg	100.15%	116.44%	100.33%	114.33%	94.72%	70.48%	97.82%	76.06%
NO_2 LML | Amsterdam-Ookmeer	103.45%	55.28%	106.70%	59.00%	88.97%	81.16%	94.73%	82.73%
NO_2 LML | Zaanstad-Hemkade	100.12%	107.22%	100.33%	106.09%	105.86%	111.56%	102.22%	108.27%
NO_2 LML | Amsterdam -Kantershof	102.02%	80.15%	103.96%	80.06%	99.77%	125.01%	99.76%	123.45%
NO_2 LML | Badhoevedorp-Sloterweg	102.47%	81.19%	104.91%	83.35%	91.84%	91.41%	95.95%	90.29%
NO_2 LML | Zaanstad-Hoogtij	101.56%	40.58%	103.03%	42.20%	95.78%	70.59%	98.05%	75.58%
NO_2 LML | Spaarnwoude-Machineweg	100.95%	113.02%	101.94%	114.21%	99.24%	105.79%	99.55%	109.21%
NO_2 LML | Hoofddorp-Hoofdweg	103.60%	28.64%	106.72%	31.31%	92.12%	41.88%	96.42%	45.43%
NO_2 LML | Oude Meer-Aalsmeerderdijk	102.52%	66.46%	104.95%	68.54%	95.11%	71.76%	97.60%	78.76%
NO_2 LML | Haarlem-Schipholweg	101.65%	60.52%	103.24%	62.79%	93.41%	59.20%	96.85%	62.95%
Total mean	**102.51%**	**56.35%**	**104.70%**	**59.19%**	**93.18%**	**74.04%**	**96.61%**	**72.40%**

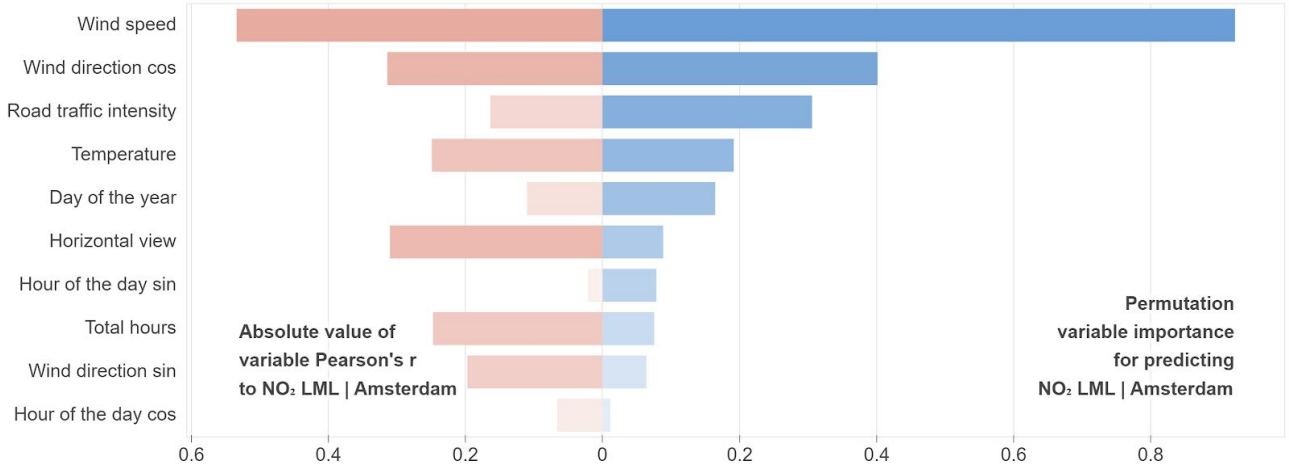


Figure 18 The variable importance of the predictions excluding COVID-19 lockdown measures testing/training data

Table 9 Difference in variable importance

Explanatory variable	Importance NO_2 LML | Amsterdam | incl COVID-19 data	Importance NO_2 LML | Amsterdam | excl COVID-19 data	Difference
Total hours	0.158	0.075	47.39%
Hour of the day cos	0.012	0.012	93.65%
Wind speed	0.892	0.923	103.45%
Horizontal view	0.080	0.088	109.94%
Wind direction cos	0.360	0.401	111.38%
Road traffic intensity	0.269	0.305	113.65%
Wind direction sin	0.055	0.064	116.38%
Day of the year	0.140	0.164	116.95%
Hour of the day sin	0.064	0.078	122.06%
Temperature	0.153	0.191	124.64%

Satellite data

The mean values of the prediction accuracies using training data only after 1 Juli 2018 and including and excluding the S5P variable can be seen in table 10. The variable importance is visualized in figure 19 and the difference indicated in table 11.

Table 10 Total mean of the prediction accuracies using data after 1 Juli 2018 including and excluding S5P data as an explanatory training variable.

Location	r	-/+ 95% CI	r^2	-/+ 95% CI	MSE	-/+ 95% CI	RMSE	-/+ 95% CI
NO_2 LML | Amsterdam excl S5P	0.780	0.173	0.616	0.270	0.086	0.050	0.291	0.081
NO_2 LML | Amsterdam incl S5P	0.788	0.147	0.626	0.227	0.086	0.044	0.291	0.073
Percentage difference NO_2 LML | Amsterdam	**100.98%**	**84.70%**	**101.60%**	**84.26%**	**99.83%**	**88.02%**	**100.08%**	**90.86%**
Total mean excl S5P	0.761	0.184	0.575	0.299	0.163	0.150	0.385	0.140
Total mean incl S5P	0.760	0.188	0.577	0.277	0.160	0.136	0.384	0.134
Percentage difference total mean	**99.90%**	**101.99%**	**100.34%**	**92.82%**	**98.73%**	**90.67%**	**99.82%**	**95.18%**

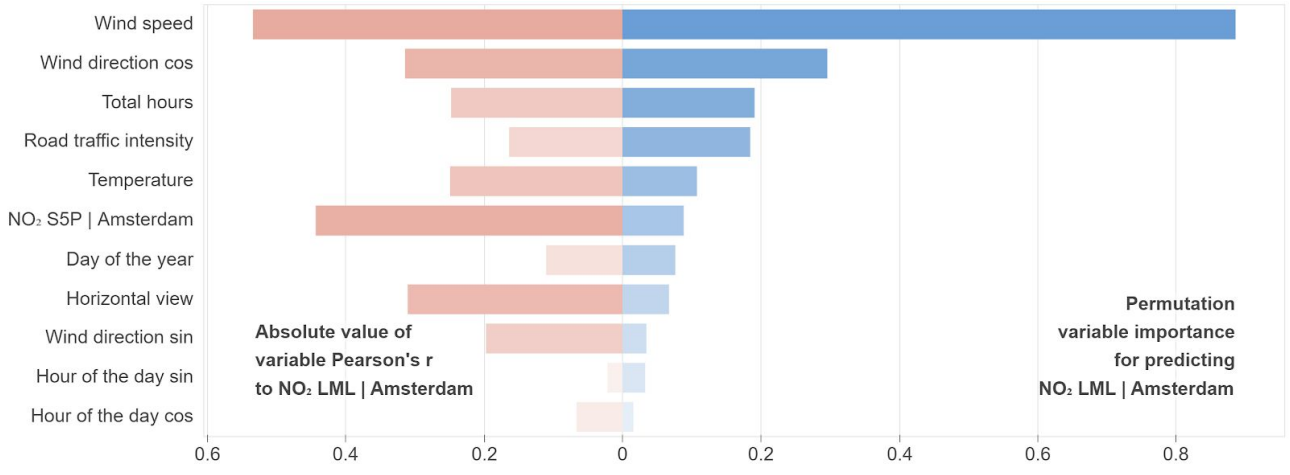


Figure 19 Variable importance of the prediction model using S5P data as an explanatory variable.

Table 11 Difference in variable importance excluding and including S5P data.

Explanatory variable	Importance NO_2 LML | Amsterdam | excl S5P data	Importance NO_2 LML | Amsterdam | incl S5P data	Difference
Wind direction sin	0.049	0.035	70.24%
Horizontal view	0.083	0.067	80.34%
Day of the year	0.093	0.076	81.40%
Hour of the day sin	0.039	0.033	84.33%
Total hours	0.216	0.190	88.29%
Road traffic intensity	0.206	0.184	89.51%
Wind direction cos	0.324	0.296	91.23%
Wind speed	0.960	0.886	92.26%
Temperature	0.111	0.107	96.90%
Hour of the day cos	0.016	0.016	100.62%

3.2.3 Research question 3 - Trends during COVID-19

The trend analysis consists of three methods displaying similar findings. The first method shows the data chronologically. Figure 20 displays relevant variables and their rolling mean values and the corresponding harmonic model.

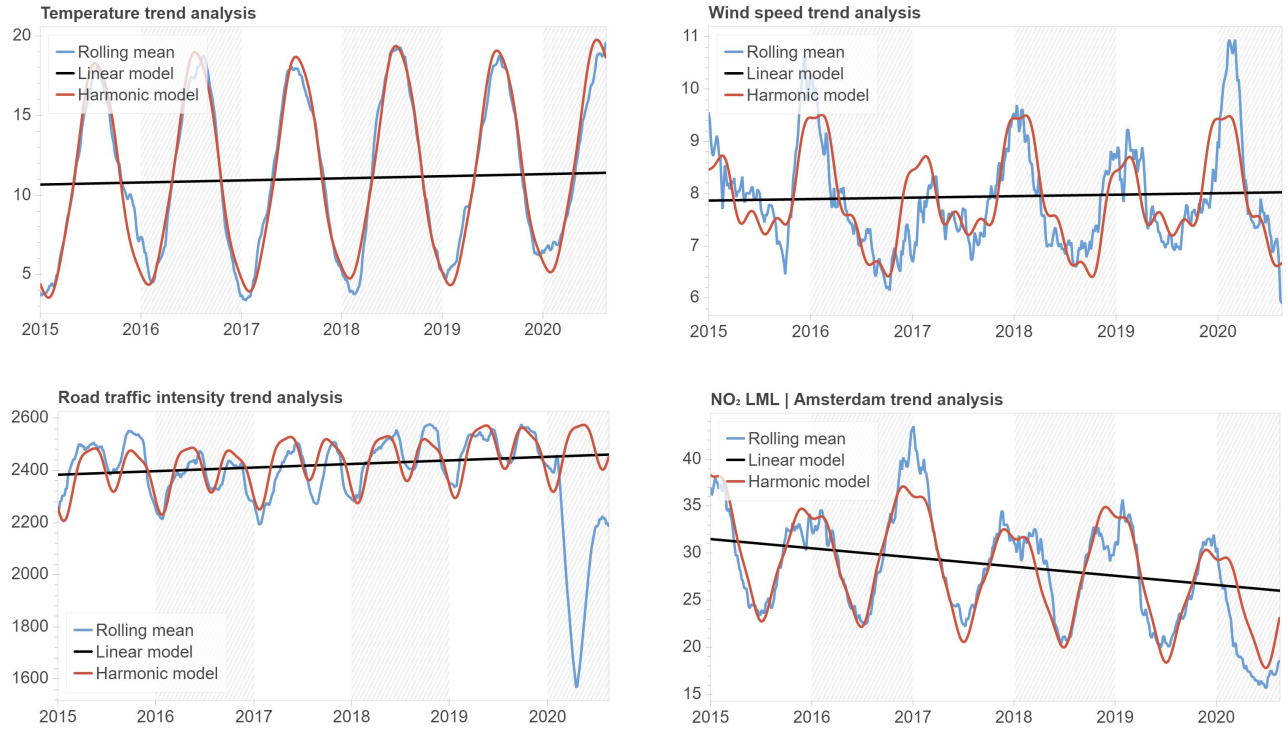


Figure 20 Time series showing the harmonic model and the rolling mean of the corresponding variable

Figure 21 shows the harmonic models (dashed lines) and the corresponding variables (solid lines) as a ridge line plot. Figure 22 displays the difference between the rolling mean and the harmonic model.

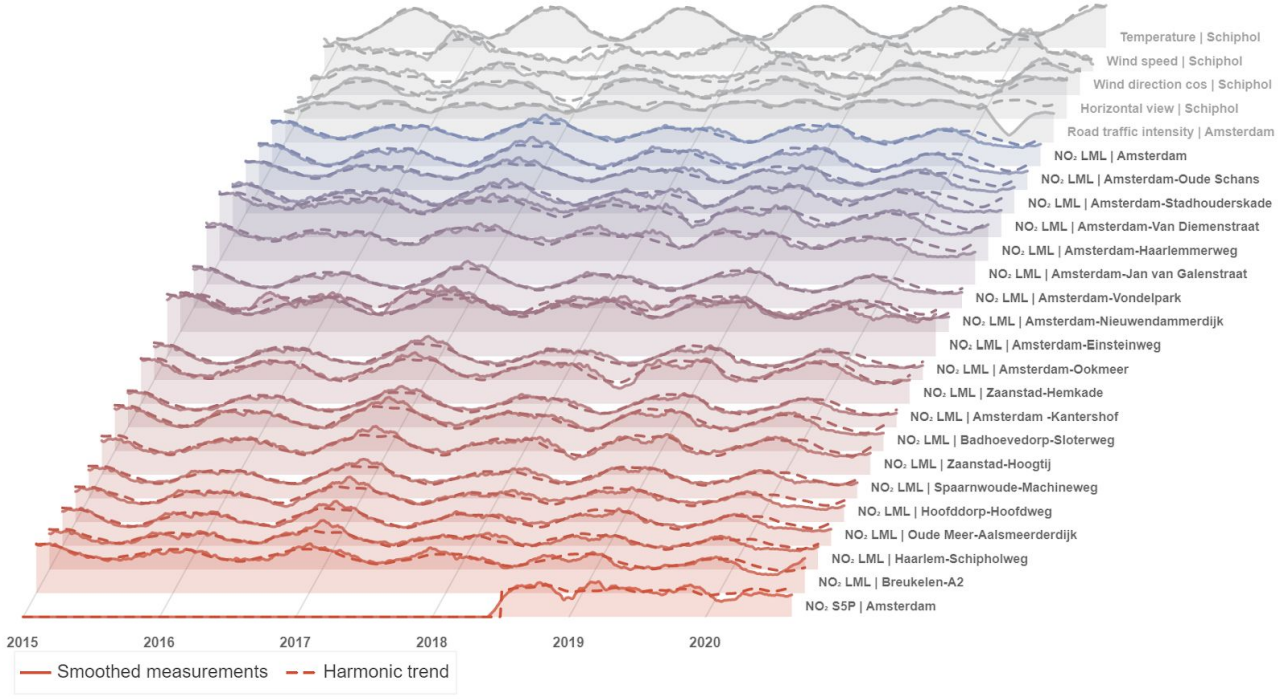


Figure 21 The harmonic model and smoothed measurements of relevant explanatory variables and all response variables

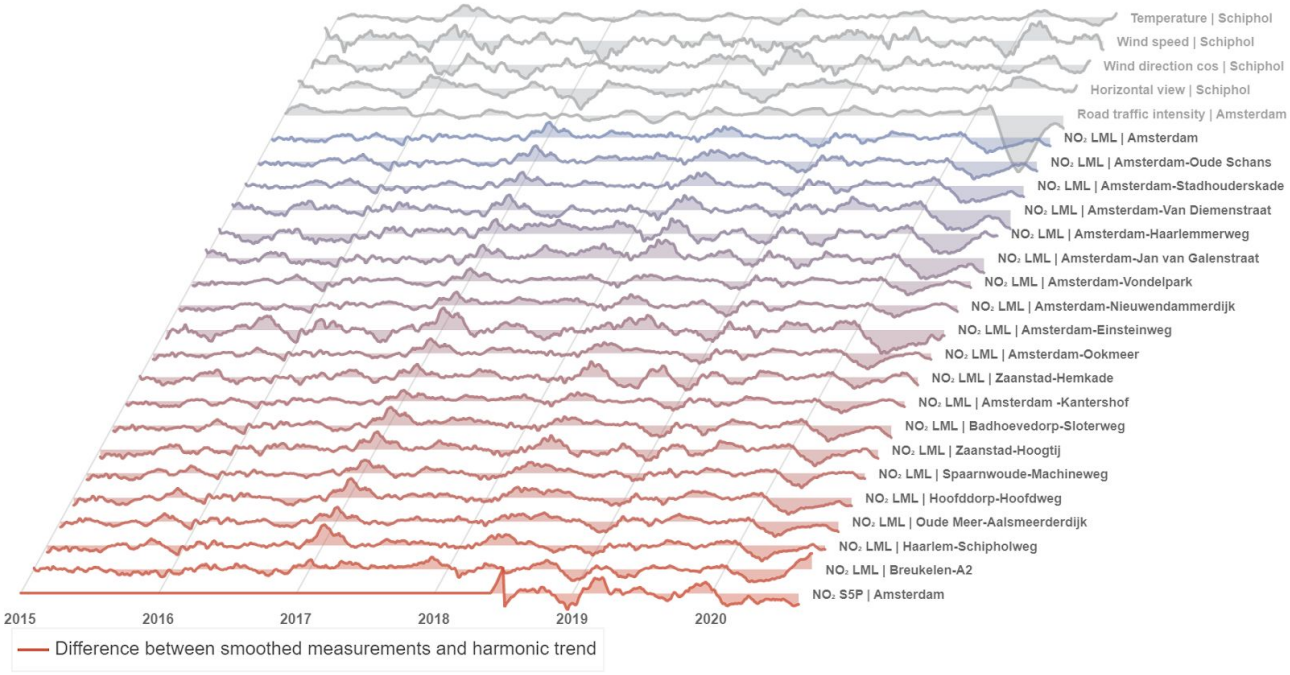


Figure 22 The difference between the smoothed measurements and harmonic model as seen in figure 21

The second method overlaps each year to compare their rational differences. See figure 23 for the mean NO_2 of the years 2015-2019 compared to the NO_2 of 2020. The percentage differences for all NO_2 locations in 2020 can be found in table 12, this includes the wind gusts and road traffic intensity as a comparison. Note that the S5P data is left out, as insufficient historical data is available.

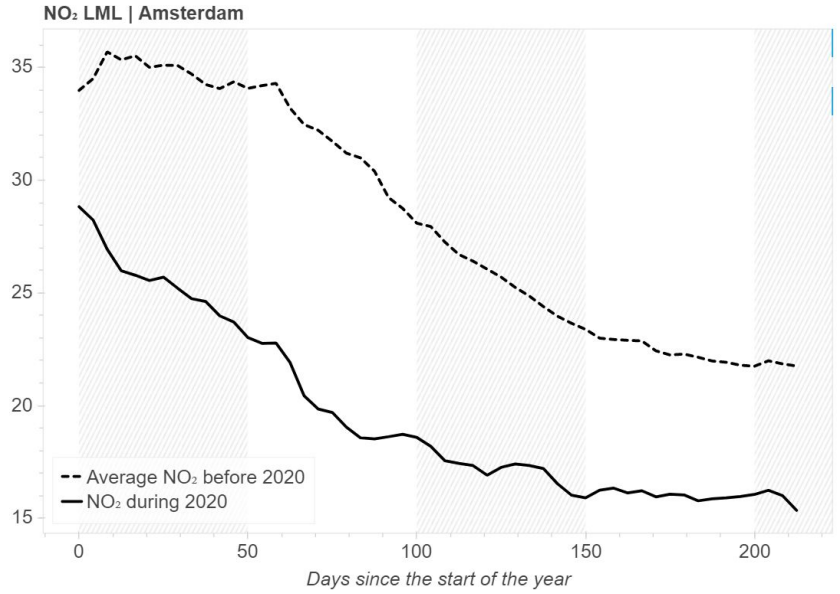


Figure 23 Time series of NO_2 before and during 2020 comparison

Table 12 Percentage difference of the values in 2020 compared to the mean value of the years 2015-2019.

Location	January	February	March	April	May	June	Juli	Total mean
NO_2 LML | Amsterdam-Haarlemmerweg	79.34%	65.22%	59.39%	53.11%	54.96%	68.60%	62.28%	**63.27%**
NO_2 LML | Hoofddorp-Hoofdweg	81.10%	47.12%	59.66%	69.86%	58.85%	72.80%	63.01%	**64.63%**
NO_2 LML | Amsterdam-Oude Schans	86.05%	60.49%	59.56%	64.48%	62.91%	74.69%	65.21%	**67.63%**
NO_2 LML | Amsterdam-Jan van Galenstraat	82.51%	62.69%	64.01%	63.59%	63.75%	73.09%	65.54%	**67.88%**
NO_2 LML | Amsterdam-Stadhouderskade	91.99%	65.42%	64.95%	60.28%	66.32%	66.49%	65.81%	**68.75%**
NO_2 LML | Badhoevedorp-Sloterweg	91.27%	56.13%	63.34%	67.13%	64.08%	68.98%	71.16%	**68.87%**
NO_2 LML | Amsterdam-Ookmeer	84.72%	46.27%	55.10%	75.07%	70.46%	84.37%	69.52%	**69.36%**
NO_2 LML | Breukelen-A2	83.66%	65.25%	60.10%	56.22%	60.80%	66.42%	94.88%	**69.62%**
NO_2 LML | Amsterdam	88.12%	59.91%	64.87%	68.43%	67.27%	77.10%	67.70%	**70.49%**
NO_2 LML | Amsterdam-Vondelpark	86.03%	56.44%	61.68%	68.74%	71.88%	74.14%	76.41%	**70.76%**
NO_2 LML | Haarlem-Schipholweg	82.77%	51.45%	64.84%	78.78%	67.71%	82.96%	66.99%	**70.78%**
NO_2 LML | Oude Meer-Aalsmeerderdijk	86.85%	53.17%	62.03%	80.09%	71.92%	79.31%	67.57%	**71.56%**
NO_2 LML | Amsterdam -Kantershof	87.19%	57.90%	64.28%	68.37%	68.53%	83.43%	71.66%	**71.62%**
NO_2 LML | Amsterdam-Einsteinweg	84.61%	53.97%	69.48%	78.46%	72.61%	80.75%	62.26%	**71.73%**
NO_2 LML | Amsterdam-Van Diemenstraat	93.06%	67.45%	67.39%	65.00%	68.48%	83.92%	70.14%	**73.64%**
NO_2 LML | Amsterdam-Nieuwendammerdijk	89.81%	64.24%	68.46%	68.15%	73.18%	84.19%	81.38%	**75.63%**
NO_2 LML | Spaarnwoude-Machineweg	91.15%	45.50%	70.13%	96.13%	78.45%	105.05%	64.75%	**78.74%**
NO_2 LML | Zaanstad-Hoogtij	106.00%	62.87%	74.88%	79.53%	86.24%	90.02%	80.02%	**82.79%**
NO_2 LML | Zaanstad-Hemkade	104.93%	77.84%	80.36%	76.59%	83.37%	87.62%	91.94%	**86.09%**
Total mean (NO_2)	**88.48%**	**58.91%**	**64.97%**	**70.42%**	**69.04%**	**79.16%**	**71.48%**	**71.78%**
Road traffic intensity | Amsterdam	*109.65%*	*104.26%*	*78.32%*	*59.91%*	*73.08%*	*84.06%*	*93.44%*	**86.10%**
Wind speed | Schiphol	*99.10%*	*158.98%*	*116.60%*	*96.21%*	*104.46%*	*96.76%*	*111.72%*	**111.97%**

Table 12 is visualized in figure 24. The visualization is expanded to the other years, as seen in figure 25.

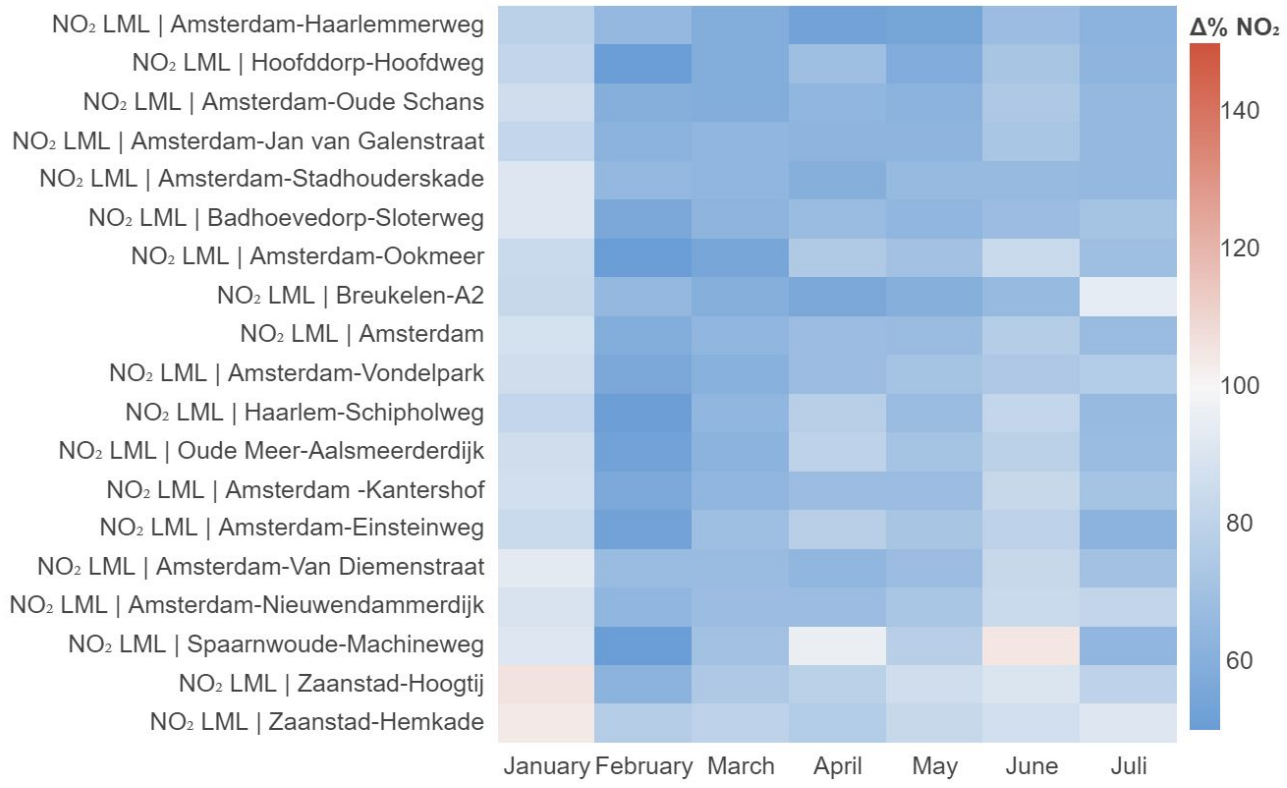


Figure 24 Heatmap of the percentage difference in NO_2 in 2020 compared to the mean value of the years 2015-2019

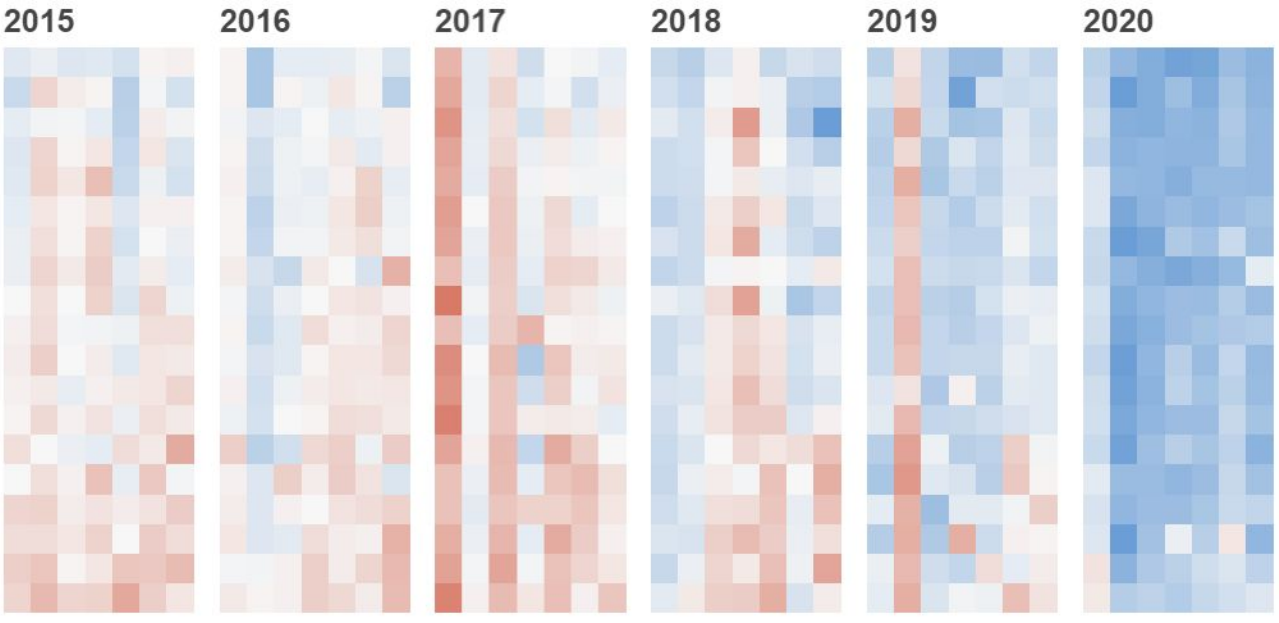


Figure 25 Heatmap of the percentage difference in NO_2 in each year compared to the mean value of the years 2015-2019

In the third method all LML air quality locations are averaged per month of each year and compared to the mean of the wind speed and road traffic intensity variables as visualized in figure 26. The correlation is recalculated on a larger moving average, once with data including 2020 as seen in figure 27 and once with data excluding 2020 as seen in figure 28.

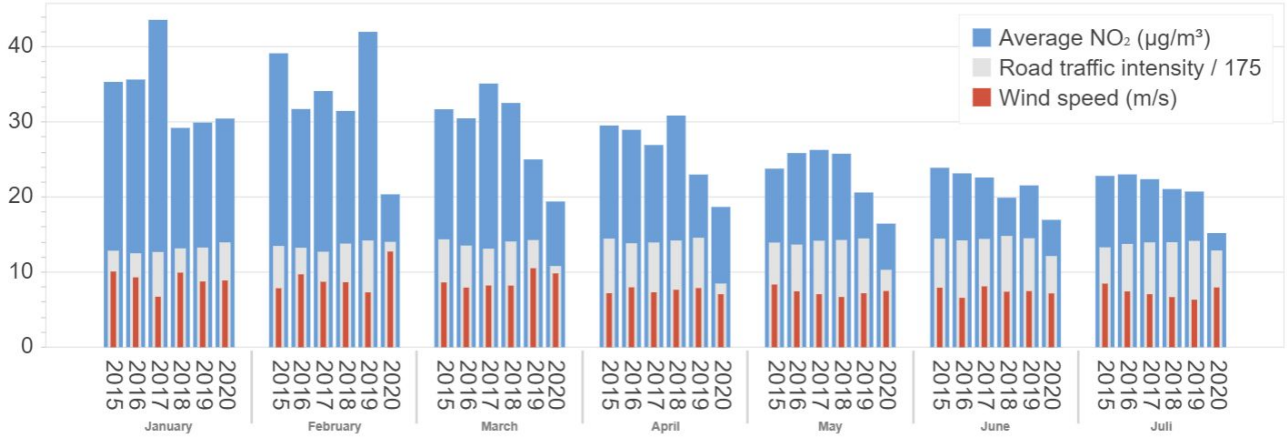


Figure 26 The mean values per month of each year of relevant variables

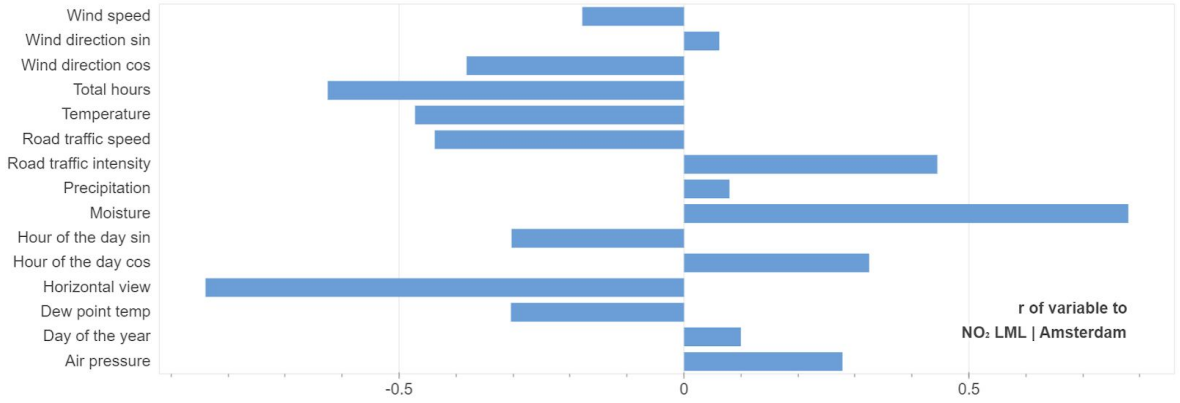


Figure 27 Long term linear correlations including 2020 data with smoothed values (window size = 160 days)

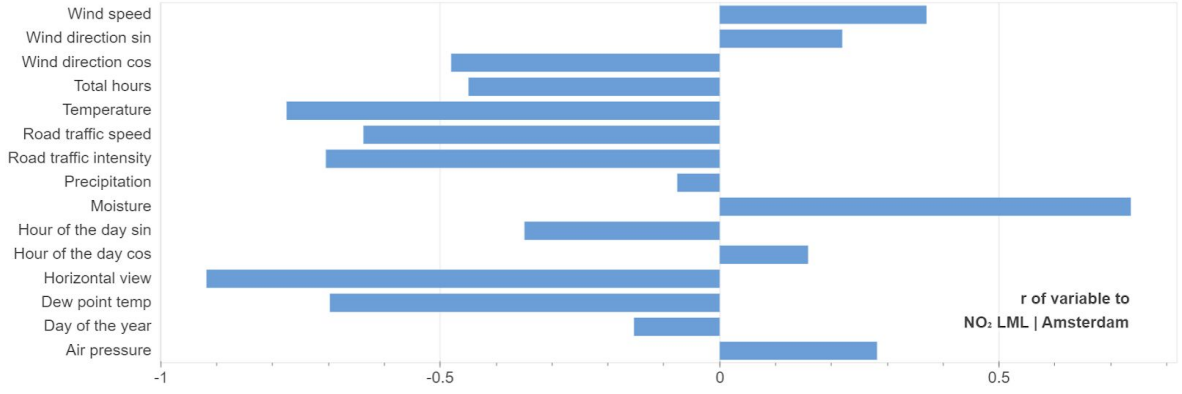


Figure 28 Long term linear correlations excluding 2020 data with smoothed values (window size = 160 days)

4 Discussion

4.1 Summary of the results

Four datasets are used, the NO_2 LML air quality, NO_2 S5P air quality, KNMI weather, and NDW road traffic data. All datasets require preprocessing steps to optimize their relationships. This includes the smoothening of the data and converting cyclical data to mathematically interpretable formats. The relationship of the NO_2 LML variables to the other variables are the most relevant, as the NO_2 LML is the response variable that will be predicted. There are weak positive and negative linear relationships between the NO_2 LML and the explanatory variables, staying within $\mathbf{-0.58 < r < 0.43}$ and $\mathbf{r^2 < 0.34}$. The wind speed variable has the largest negative relationship and the NO_2 S5P variable has the largest positive relationship. The road traffic intensity does not show a strong linear relationship on hourly time intervals. However, this changes when predicting NO_2 LML. Here road traffic ranks number three in variable importance to predict NO_2. It shatters in comparison to wind speed, which has by far the most predictive power. Variables that have a very low importance are filtered before the final predictions. These predictions have an accuracy of $\mathbf{r = 0.824}$ (±0.102 with a 95% confidence interval (CI)), $\mathbf{r^2 = 0.683}$ (±0.158 with a 95% CI), **MSE = 0.120** (±0.084 with a 95% CI), and **RMSE = 0.337** (±0.105 with a 95% CI). Another prediction was made excluding data during 2020, whereafter the $\mathbf{r^2}$ improved to **~105%** and the road traffic intensity importance rose to **~114%**. When using training data that covers the temporal extent of the S5P data, an increase in $\mathbf{r^2}$ of **~2%** and a narrower CI can be found when including the previously filtered NO_2 S5P data. Furthermore, the seasonal trend analysis shows a decrease in the mean NO_2 concentrations in the LML and S5P dataset during the COVID-19 lockdowns in 2020. The average NO_2 LML dropped to **~65%** in March 2020 and to **~73%** during the second quarter of 2020 compared to the mean NO_2 of the previous years. This coincides with a drop to ~78% in road traffic intensity on three major roads during March 2020, with a mean of ~72% during the second quarter. The wind speed measured near Schiphol may have played a role in the reduced NO_2 concentration due to their negative correlation as it increased to ~117% in March 2020. However, it stabilized to ~99% in the second quarter of the same year. Lastly, when excluding data from 2020, the road traffic intensity and NO_2 seem to have a slight negative relationship which was also seen in the harmonic model. This relationship became positive when the 2020 data was included, where potentially the NO_2 concentrations and road traffic intensity independently react to other factors.

4.2 Analysis of the results

4.2.1 Observations

The dataset specifications indicate that the LML, KNMI, NDW vector datasets have temporal properties with similar density, whereas the S5P raster dataset has a sparser temporal resolution and a denser coverage due to its gridded format (see table 1 to 3). The vector datasets have at least one measurement per hour. The spatial resolution varies as their geographic distribution is different, where the LML is positioned near roads, open fields, and in between buildings, the NDW data points on major roads in and around Amsterdam and a single KNMI data point in an open field at Schiphol. Also, the amount of measurement locations per area differ by large margins. However, each location is within a distance of 20 km in and around the city centre of Amsterdam. The assumption is that the spatial correlation is relatively high with this proximity and the averaging of geographic locations. A linear relationship between the averages of these data points can be observed. In order to increase their similarity, necessary preprocessing steps were implemented. The smoothening of the data by taking their rolling mean increased the similarity between the dataset and removed unnecessary volatility in the data, but at a cost of temporal detail (see figure 5 to 7). Conversion of cyclical data (see figure 3 and 8) and log transformation (see figure 13) were important steps as well that improved variable relationships and importances, and prediction accuracies without fundamentally changing the data. The polynomial interpolation in the temporally sparse S5P data is necessary to account for empty values between the daily measurements, and proves to have the highest positive correlation to the LML data.

The majority of the relationships of the NO_2 S5P, NDW and KNMI variables to the NO_2 LML variable have small to moderate correlations with low explanatory percentages (see figure 9 to 13). This may suggest a weak short term individual linear dependency of each variable to the NO_2 LML variable. However, this does not tell the whole story, as there may be nonlinear and interdependent relationships between the normally distributed variables. This was further investigated by predicting the NO_2 LML data where the explanatory variables are analyzed. Even though the individual relationships are weak, a decent prediction accuracy was established (see table 4). This prediction accuracy stayed similar when removing variables with low importance (e.g. precipitation, moisture, road traffic speed), which could be safely discarded (see figure 14 and 15 and table 5 and 6). The high prediction accuracy compared to the weak correlations can be explained by the interconnectivity between the datasets, where two individual variables with different properties at the same time interval explain a certain NO_2 concentration. For example, wind speed plays an important role, but only using the wind speed as a predictor may be difficult, because the wind faces different directions. Combining the wind speed with the wind direction generates a nonlinear multidimensional predictor, thus causing a higher prediction accuracy. Another advantage of using a nonlinear approach is that the distributions of the explanatory and response variables can be expressed as clusters, which are quite distinguishable as indicated in the regression plots with linked histograms (see figure 13).

The importance of each variable indicates that the wind speed far outweighs the other explanatory variables, which coincides with the relatively large negative correlation. Still, the other variables hold a certain degree of importance, and contribute to the final prediction accuracy. The road traffic intensity is ranked third in terms of importance, suggesting that this variable can explain NO_2 concentrations to a certain degree (see figure 15). Another prediction was performed excluding training data during the COVID-19 lockdowns in 2020 to further investigate the impact of the variables, namely the road traffic intensity which had a large deviation from the seasonal trend during 2020. The prediction accuracy improved slightly with narrower CIs after retraining the model without the 2020 data (see table 7 and 8). The relatively small prediction improvement indicates that the RFR model was not overfitted and handled the 2020 abnormalities accordingly. It should be stressed that parts of these abnormalities were also present in the training data. The importance of the road traffic intensity improved slightly when predicting without 2020 data (see figure 18 and table 9). This means that the road traffic and NO_2 concentrations have a stronger link before the drastic change in 2020. The variable that did change by a large percentage was the total hours. This temporal variable dropped in importance when excluding data during the COVID-19 lockdown. The drop in temporal importance may be explained by the abnormal patterns found in the data during the COVID-19 lockdowns, thus relying less on the historical trends partially explained by the total hours variable which may reflect the temporal seasonality of the data.

Another model training was performed to assess the importance of the NO_2 S5P satellite data. The measured NO_2 from the S5P was initially removed as it had incomplete data over the temporal range found in the other datasets. When retraining the prediction model within the temporal range of the S5P data, once excluding and once including the NO_2 S5P variable, a small increase in prediction accuracy was observed when this variable was included (see table 10). The NO_2 S5P has a larger spatial coverage, but a lower temporal resolution compared to the NO_2 LML data. Nevertheless, when interpolating and smoothening the data, a similarity can be observed that contributes to an improved NO_2 LML prediction. Directly interchanging these NO_2 dataset should be performed with caution, as their measurement techniques and values differ.

The seasonal trend analysis did see the NO_2 S5P and NO_2 LML values drop during the 2020 COVID-19 lockdowns (see figure 20 to 25). Simultaneously, wind speeds picked up early 2020, suggesting this may be the cause of the lower NO_2 concentrations due to their earlier established negative correlation and a similar inverse pattern in 2017 (see figure 20, 22 and 25). The exceptionally low NO_2 concentration during February 2020 was probably caused by the unusually high wind speeds during this month. However, the following months all the way up to July still see low NO_2 concentrations whilst the wind speed stabilizes. This could mean that the wind speed was not the dominant factor in the improved air quality. Conversely, the road traffic intensity did drop following a similar pattern to the deviation found in the NO_2 data (see table 12). Does this mean that the reduced NO_2 concentrations in 2020 are caused by the lower road traffic intensity? Looking at the years before 2020, the positive relationship between the road traffic intensity and NO_2 concentrations seems to dissipate and even become slightly

negatively correlated (see figure 27 and 28) where the road traffic intensity has an upward trend and the NO_2 LML a downward trend (see figure 20).

4.2.2 Explanations & reflections

The research describes the results in the form of literature research, relationships, predictions, variable importances and seasonal trends concerning relevant datasets. Each dataset needed preprocessing to meet the goal of this research to improve their compatibility. The preprocessing steps altered the datasets from their initial raw states, which may have impacted the results. However, most preprocessing steps adhered to common methodologies, such as log transformation which is a simple and effective step of normally distributing the data and cyclical data conversion which proved to drastically improve the results without fundamentally altering the data. Smoothening the data had the largest impact on its structure and was somewhat arbitrary in window size. Depending on the question being addressed, different window sizes were used to apply the smoothening. Initially, this size was relatively small and by empirical testing optimal relationships and prediction performances were noticed when the data was categorized on an hourly basis. Using this narrow window, the correlations between the datasets corresponded with detailed temporal changes in the data. Extreme volatility which may house valuable information was filtered out by this smoothening, but the general patterns remained visible.

The four weather elements that related the most to NO_2 are the wind speed, wind direction, horizontal view and moisture. This makes physical sense. In the case of moisture and the horizontal view an association between relative humidity and the rapid sulfur dioxide (SO_2) oxidation to particulate sulfate can be found in hazy situations, where NO_2 is one of the main oxidizers (Li, Hoffmann, and Colussi 2018; Wang et al. 2020). This causes a reduced visibility and coincides with a high atmospheric water content and presence of particulates. That can also explain their high negative correlation and the reduced importance of moisture for predicting NO_2 as it does not add any extra explanatory power due to its dependency to the horizontal view. The wind direction and wind speed have the highest negative correlation to NO_2, because these factors are dominant in the meteorological processes related to air pollution due to their transportational and directional properties, as well as their influence of relative humidity, particulates, and overcast (Weiner and Matthews 2003). Interestingly enough, the temperature variable which had a low correlation on an hourly base to NO_2 had a relatively high variable importance in the prediction. This may be due to its consistent seasonal nature as seen in the trend analysis, but also to a stronger nonlinear relationship and higher long term correlation, therefore providing a solid unaltered base for predicting NO_2 even when other variables fluctuate irregularly. In the end, combining these weather factors led to the relatively high prediction accuracies as seen in their variable importance rankings (see figure 15). However, for an analysis that is able to isolate the importance and influence of the other variables in relation to NO_2 concentrations a bit more, one may want to eliminate the weather contribution all together and analyze the residuals.

The relatively high correlation between the LML and S5P NO_2 values does not come as a surprise, as it aims to measure the same component and previous correlations have been found (Ialongo et al. 2020). But as stated before these variables should be kept separate as they measure respectively on a local scale and remotely over an entire tropospheric column. In terms of preprocessing, the polynomial fit between the sparse data points of the S5P is relatively arbitrary and could be optimized by other more advanced interpolation techniques. The S5P data may play a more important role when the goal is to predict NO_2 over larger time intervals, which align to the daily measurements of the S5P satellite mission.

The road traffic data adds a human factor to the equation, which still leaves room for interpretation as its hourly correlation to NO_2 is weak and its prediction importance moderate. This correlation may vary per measurement location, and should be further assessed for sub populations (see figure 24). The relationship of the averaged variables of each dataset becomes more apparent over longer time intervals, where conversely a negative correlation before the year 2020 is found. Other studies have found a positive correlation between road traffic and NO_2 concentrations (Bohemen and Janssen van de Laak 2003; Comert et al. 2020; Jiang et al. 2019), which does not apply to Amsterdam in the 5 years before 2020. This may have been due to the insignificant impact of vehicles or due to a reduction in vehicles with fuel producing toxic emissions such as Diesel (Olabi,

Maizak, and Wilberforce 2020) partially enforced by regulations (Amsterdam 2020). On the other hand, the negative correlation is quite small and may not hold any significance. Even though a slight negative correlation is historically normal, a sudden positive correlation and a comparative trend in 2020 between NO_2 and the road traffic intensity was observed, aligning with the recent study of Venter et al. (2020). This synchronicity may partially be explained by other anthropogenic factors, such as agricultural, economic, and industrial changes, to which these variables independently respond.

4.3 Significance of the results

Many of the results can be explained by physical properties where similar observations and relationships were established in earlier research. The weather data aligns with the expected patterns and trends, the S5P satellite data was similar to the ground measurements, and the road traffic intensity coincided with the drop in NO_2 during 2020. The addition of data during the COVID-19 lockdowns in 2020, the comparison to historical seasonal data, and the local spatial scale was a significant part of this research, as these spatiotemporal areas are relatively unexplored. For example the S5P satellite data resulted in an improved prediction accuracy and according to several analysis methods, a possibility exists that NO_2 and road traffic behave independently in and around Amsterdam. The methods within this research can be applied to the entire Netherlands where the datasets are present. This research focussed on NO_2, but can be easily applied to other air pollutants available in the LML data. The results derived from the methodology and data of this research can support local decision making, and provides us with a better understanding of the human influence on our air quality.

4.4 Further research

The notion that the road traffic has a weak negative correlation to NO_2 is an interesting phenomenon that should be further investigated. The sudden shift to a synchronized trend in 2020 is therefore even more relevant to understand, especially concerning the causality of the road traffic intensity to NO_2. The current methodology averaged a limited amount of NDW road traffic sensors. They should be individually assessed and queried by close proximity to NO_2 sensors for more detailed and local results. Weather variables can also be accounted for in future research where the residuals may play an important role in understanding the nuanced effects of road traffic on NO_2 before and during COVID-19 lockdowns. Moreover, the NO_2 sensors are only spatially correlated by proximity and do not take terrain roughness into account. These factors are partially accounted for by previous research, but should be applied to the spatiotemporal scale and range similar to this research. Furthermore, the S5P data was underrepresented in this research due to its shorter temporal range. The methods of this research could be expanded by including the S5P data in the dominant prediction model and trend analysis as new measurements become available.

Lastly, the results only scratch the surface in terms of analysis depth. They could be expanded by new and interesting questions or the addition of other datasets, such as more remote sensing data (e.g. the *GLDAS Noah Land Surface Model* (GES DISC 2020; Rodell et al. 2004)) and citizen science projects (e.g. *openSenseMap* (senseBox 2020)). In terms of preprocessing the window sizes of the rolling means could be altered to different temporal intervals that may prioritize long term predictions that conform to the lower temporal resolution of the S5P data. The relationships could be assessed using multiple nonlinear regression analysis methods and the basic RFR machine learning model could be replaced by more advanced deep learning models. The cross validation technique used to assess the prediction model performance is effective, but currently uses an arbitrary amount of k-folds and data partitioning was randomized using stratified sampling over the entire training and testing datasets. Future research could focus this sampling more to the desired temporal ranges of interest, such as the 2020 deviations, and exclude training data within this range to stress test the prediction model on overfitting. The trend analysis techniques may be expanded by more advanced seasonal modelling methods, and the historical comparisons could be further expanded by normalizing the value ranges based on the minima and maxima per year to eliminate the multi year trends.

5 Conclusions

The NO_2 concentrations in Amsterdam saw a steady decline in the last 5 years with drastically lower values in the year 2020. This drop coincided with the economic and societal changes in response to the COVID-19 lockdowns. A similar drop was observed in satellite imagery. Other factors, such as the road traffic intensity, had a steady incline over these past years, but also dropped during 2020. This contrasted the weather patterns which remained unaffected. In the short term these factors (i.e. explanatory variables) have a weak linear correlation to NO_2 ground measurements. The wind speed had the strongest negative correlation and the satellite imagery the strongest positive correlation. In between these variables the linear correlation was weak. Nevertheless, the predictions resulting from these factors to explain ground based NO_2 data were relatively accurate, even more so when the satellite data was included. Excluding data during 2020 did not affect the prediction accuracy by a large percentage, suggesting that the explanatory variables may have a small impact on major fluctuations in the response variable. The importance of each explanatory variable did see that the road traffic intensity had some importance in predicting NO_2 ranking third, topped by the wind direction and wind speed. This importance did not change when excluding 2020 data. We may conclude that short term hourly weather, road traffic, and satellite measurements can explain the ground based NO_2 concentrations in and around Amsterdam. It is nevertheless the question if these explanatory variables hold any causality, namely the road traffic intensity. This was further investigated by a trend analysis on longer time scales. Here a shift in linear correlations were observed, with a negative correlation between the road traffic intensity and NO_2 concentrations when 2020 data was excluded. Including the 2020 data did see a positive correlation and similar trend deviations between these variables as compared to the regular seasonality in the previous years. Even though they may be similar during 2020, the earlier observed weak short term correlation, the relatively small unaltered variable importance, and the negative long term correlation, may suggest that road traffic has a neglectable impact on the average NO_2 concentrations in and around Amsterdam. As mentioned before, they may react independently to other human factors providing an indirect reflection of how our economy and society operates. This statement should be further investigated, especially on a more local and individual scale looking at subpopulations in the data and accounting for residual effects. The current preprocessing steps which involved variable selection, smoothening, log transformation and cyclical data conversion did contribute to an improved prediction accuracy. These steps could be further expanded by other methods and arbitrariness could be eliminated. The analysis should be expanded by more datasets and the coverage should include other parts of the Netherlands. The proposed methodology in this research allows for this scaling, where other pollutants can be analyzed as well.

6 References

Amsterdam. 2020. "Low Emission Zone." *English Site*. Retrieved September 17, 2020 (https://www.amsterdam.nl/en/traffic-transport/low-emission-zone/).

Anaconda Inc. 2020. "Anaconda | The World's Most Popular Data Science Platform." *Anaconda*. Retrieved September 2, 2020 (https://www.anaconda.com/).

Battista, Gabriele, and Roberto de Lieto Vollaro. 2017. "Correlation between Air Pollution and Weather Data in Urban Areas: Assessment of the City of Rome (Italy) as Spatially and Temporally Independent Regarding Pollutants." *Atmospheric Environment* 165:240–47.

Bauwens, M., S. Compernolle, T. Stavrakou, J. -F. Müller, J. Gent, H. Eskes, P. F. Levelt, R. A, J. P. Veefkind, J. Vlietinck, H. Yu, and C. Zehner. 2020. "Impact of Coronavirus Outbreak on NO_2 Pollution Assessed Using TROPOMI and OMI Observations." *Geophysical Research Letters* 47(11).

Bohemen, H. D. Van, and W. H. Janssen van de Laak. 2003. "The Influence of Road Infrastructure and Traffic on Soil, Water, and Air Quality." *Environmental Management* 31(1):0050–0068.

Bokeh contributors. 2020. "Bokeh." Retrieved September 2, 2020 (https://bokeh.org/).

Breiman, L. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Brunekreef, Bert, and Stephen T. Holgate. 2002. "Air Pollution and Health." *The Lancet* 360(9341):1233–42.

Cattaneo, Bruno. 2019. "Air Quality: Traffic Measures Could Effectively Reduce NO_2 Concentrations by 40% in Europe's Cities." *EU Science Hub - European Commission*. Retrieved September 18, 2020 (https://ec.europa.eu/jrc/en/news/air-quality-traffic-measures-could-effectively-reduce-NO_2-concentrations-40-europe-s-cities).

Comert, Gurcan, Samuel Darko, Nathan Huynh, Bright Elijah, and Quentin Eloise. 2020. "Evaluating the Impact of Traffic Volume on Air Quality in South Carolina." *International Journal of Transportation Science and Technology* 9(1):29–41.

EEA. 2019. *Air Quality in Europe 2019 — European Environment Agency*. *Publication*. ISSN 1977-8449.

EO. 2020. "Airborne Nitrogen Dioxide Plummets Over China." Retrieved August 31, 2020 (https://earthobservatory.nasa.gov/images/146362/airborne-nitrogen-dioxide-plummets-over-china).

ESA. 2020. "Sentinel-5P - Missions - Sentinel Online." Retrieved September 2, 2020 (https://sentinel.esa.int/web/sentinel/missions/sentinel-5p).

GES DISC. 2020. "GES DISC Dataset: GLDAS Noah Land Surface Model L4 3 Hourly 0.25 x 0.25 Degree V2.1 (GLDAS_NOAH025_3H 2.1)." Retrieved September 21, 2020 (https://disc.gsfc.nasa.gov/datasets/GLDAS_NOAH025_3H_2.1/summary).

Google. 2020. "Google Earth Engine." Retrieved September 2, 2020 (https://earthengine.google.com).

Griffin, Debora, Xiaoyi Zhao, Chris A. McLinden, Folkert Boersma, Adam Bourassa, Enrico Dammers, Doug Degenstein, Henk Eskes, Lukas Fehr, Vitali Fioletov, Katherine Hayden, Shailesh K. Kharol, Shao Meng Li, Paul Makar, Randall V. Martin, Cristian Mihele, Richard L. Mittermeier, Nickolay Krotkov, Maarten Sneep, Lok N. Lamsal, Mark ter Linden, Jos van Geffen, Pepijn Veefkind, and Mengistu Wolde. 2019. "High-Resolution Mapping of Nitrogen Dioxide With TROPOMI: First Results and Validation Over the Canadian Oil Sands." *Geophysical Research Letters* 46(2):1049–60.

Hargreaves, P. R., J. U. Smith, S. Young, and K. W. T. Goulding. 2005. "Development of an Empirical Model to Predict Nitrogen Dioxide Concentrations from Weather Variables for Sites across the UK." *Atmospheric Environment* 39(3):409–17.

Ialongo, Iolanda, Henrik Virta, Henk Eskes, Jari Hovila, and John Douros. 2020. "Comparison of TROPOMI/Sentinel-5 Precursor NO_2 Observations with Ground-Based Measurements in Helsinki." *Atmospheric Measurement Techniques* 13(1):205–18.

Jiang, Lili, Ziheng Sun, Qingwen Qi, and An Zhang. 2019. "Spatial Correlation between Traffic and Air Pollution in Beijing." *The Professional Geographer* 71(4):654–67.

Kampa, Marilena, and Elias Castanas. 2008. "Human Health Effects of Air Pollution." *Environmental Pollution* 151(2):362–67.

KNMI. 2020. "KNMI - Uurgegevens van Het Weer in Nederland - Download." Retrieved September 1, 2020 (http://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi).

Li, Lijie, Michael R. Hoffmann, and Agustín J. Colussi. 2018. "Role of Nitrogen Dioxide in the Production of Sulfate during Chinese Haze-Aerosol Episodes." *Environmental Science & Technology* 52(5):2686–93.

Luchtmeetnet.nl. 2020. "Luchtmeetnet.nl." *Luchtmeetnet.nl*. Retrieved May 5, 2020 (https://www.luchtmeetnet.nl/).

Manders, Astrid M. M., Peter J. H. Builtjes, Lyana Curier, Hugo A. C. Denier van der Gon, Carlijn Hendriks, Sander Jonkers, Richard Kranenburg, Jeroen J. P. Kuenen, Arjo J. Segers, Renske M. A. Timmermans, Antoon J. H. Visschedijk, Roy J. Wichink Kruit, W. Addo J. van Pul, Ferd J. Sauter, Eric van der Swaluw, Daan Swart, John Douros, Henk Eskes, Erik van Meijgaard, Bert van Ulft, Peter van Velthoven, Sabine Banzhaf, Andrea C. Mues, Rainer Stern, Guangliang Fu, Sha Lu, Arnold Heemink, Nils van Velzen, and Martijn Schaap. 2017. "Curriculum Vitae of the LOTOS–EUROS (v2.0) Chemistry Transport Model."

NDW. 2020. "Nationale Databank Wegverkeersgegevens." Retrieved May 5, 2020 (https://www.ndw.nu/).

NumPy. 2020. "NumPy." Retrieved September 2, 2020 (https://numpy.org/about/).

Olabi, A. G., David Maizak, and Tabbi Wilberforce. 2020. "Review of the Regulations and Techniques to Eliminate Toxic Emissions from Diesel Engine Cars." *Science of The Total Environment* 748:141249.

pandas. 2020. "Pandas - Python Data Analysis Library." Retrieved September 2, 2020 (https://pandas.pydata.org/about/).

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12(85):2825–30.

Project Jupyter. 2020. "Project Jupyter." Retrieved September 2, 2020 (https://www.jupyter.org).

Ramanathan, V., and Y. Feng. 2009. "Air Pollution, Greenhouse Gases and Climate Change: Global and Regional Perspectives." *Atmospheric Environment* 43(1):37–50.

Rijkswaterstaat. 2020. "Aan de slag met de Omgevingswet." *Aan de slag met de Omgevingswet*. Retrieved May 5, 2020 (https://aandeslagmetdeomgevingswet.nl/).

RIVM. 2020a. "Landelijk Meetnet Luchtkwaliteit | RIVM." Retrieved May 5, 2020 (https://www.rivm.nl/landelijk-meetnet-luchtkwaliteit).

RIVM. 2020b. "Nog Geen Duidelijk Verband Zichtbaar Tussen Verminderde Mobiliteit En Concentraties Fijnstof En Stikstofdioxide in Nederland | RIVM." Retrieved August 31, 2020 (https://www.rivm.nl//lucht/nog-geen-duidelijk-verband-zichtbaar-tussen-verminderde-mobiliteit-en-concentraties-fijnstof-en).

Rodell, M., P. R. Houser, U. Jambor, J. Gottschalck, K. Mitchell, C. J. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, J. K. Entin, J. P. Walker, D. Lohmann, and D. Toll. 2004. "The Global Land Data Assimilation System." *Bulletin of the American Meteorological Society* 85(3):381–94.

scikit-learn developers. 2020a. "3.1. Cross-Validation: Evaluating Estimator Performance — Scikit-Learn 0.23.2 Documentation." Retrieved September 3, 2020 (https://scikit-learn.org/stable/modules/cross_validation.html).

scikit-learn developers. 2020b. "3.2.4.3.2. Sklearn.Ensemble.RandomForestRegressor — Scikit-Learn 0.23.2 Documentation." Retrieved September 2, 2020 (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html).

scikit-learn developers. 2020c. "4.2. Permutation Feature Importance — Scikit-Learn 0.23.2 Documentation." Retrieved September 3, 2020 (https://scikit-learn.org/stable/modules/permutation_importance.html#permutation-importance).

SciPy developers. 2020. "Scientific Computing Tools for Python — SciPy.Org." Retrieved September 2, 2020 (https://www.scipy.org/about.html#).

Seinfeld, John H., and Spyros N. Pandis. 1998. "From Air Pollution to Climate Change." *Atmospheric Chemistry and Physics*

1326.

senseBox. 2020. "OpenSenseMap." Retrieved September 21, 2020 (https://opensensemap.org/).

Shumway, Robert H., and David S. Stoffer. 2017. *Time Series Analysis and Its Applications: With R Examples*. Springer.

Silva, Raquel A., J. Jason West, Yuqiang Zhang, Susan C. Anenberg, Jean-François Lamarque, Drew T. Shindell, William J. Collins, Stig Dalsoren, Greg Faluvegi, Gerd Folberth, Larry W. Horowitz, Tatsuya Nagashima, Vaishali Naik, Steven Rumbold, Ragnhild Skeie, Kengo Sudo, Toshihiko Takemura, Daniel Bergmann, Philip Cameron-Smith, Irene Cionni, Ruth M. Doherty, Veronika Eyring, Beatrice Josse, I. A. MacKenzie, David Plummer, Mattia Righi, David S. Stevenson, Sarah Strode, Sophie Szopa, and Guang Zeng. 2013. "Global Premature Mortality Due to Anthropogenic Outdoor Air Pollution and the Contribution of Past Climate Change." *Environmental Research Letters* 8(3):034005.

Story, R. 2013. "Folium — Folium 0.11.0 Documentation." Retrieved September 2, 2020 (https://python-visualization.github.io/folium/).

Temam, Sofia, Emilie Burte, Martin Adam, Josep M. Antó, Xavier Basagaña, Jean Bousquet, Anne-Elie Carsin, Bruna Galobardes, Dirk Keidel, Nino Künzli, Nicole Le Moual, Margaux Sanchez, Jordi Sunyer, Roberto Bono, Bert Brunekreef, Joachim Heinrich, Kees de Hoogh, Debbie Jarvis, Alessandro Marcon, Lars Modig, Rachel Nadif, Mark Nieuwenhuijsen, Isabelle Pin, Valérie Siroux, Morgane Stempfelet, Ming-Yi Tsai, Nicole Probst-Hensch, and Bénédicte Jacquemin. 2017. "Socioeconomic Position and Outdoor Nitrogen Dioxide (NO_2) Exposure in Western Europe: A Multi-City Analysis." *Environment International* 101:117–24.

The European Parliament and The Council Of The European Union. 2008. "Richtlijn 2008/50/EG van het Europees Parlement en de Raad van 20 mei 2008 betreffende de luchtkwaliteit en schonere lucht voor Europa." *Official Journal of the European Union* OJ L(152):1–44.

Veefkind, J. P., E. a. A. Aben, K. McMullan, H. Forster, J. de Vries, G. Otter, J. Claas, H. J. Eskes, J. F. de Haan, Q. Kleipool, M. van Weele, O. Hasekamp, R. Hoogeveen, J. Landgraf, R. Snel, P. J. J. Tol, P. Ingmann, R. Voors, B. Kruizinga, R. Vink, and H. Visser. 2012. "TROPOMI on the ESA Sentinel-5 Precursor: A GMES Mission for Global Observations of the Atmospheric Composition for Climate, Air Quality and Ozone Layer Applications." *Remote Sensing of Environment* 120(SI):70–83.

Venter, Zander S., Kristin Aunan, Sourangsu Chowdhury, and Jos Lelieveld. 2020. "COVID-19 Lockdowns Cause Global Air Pollution Declines." *Proceedings of the National Academy of Sciences* 117(32):18984–90.

Wang, Junfeng, Jingyi Li, Jianhuai Ye, Jian Zhao, Yangzhou Wu, Jianlin Hu, Dantong Liu, Dongyang Nie, Fuzhen Shen, Xiangpeng Huang, Dan Dan Huang, Dongsheng Ji, Xu Sun, Weiqi Xu, Jianping Guo, Shaojie Song, Yiming Qin, Pengfei Liu, Jay R. Turner, Hyun Chul Lee, Sungwoo Hwang, Hong Liao, Scot T. Martin, Qi Zhang, Mindong Chen, Yele Sun, Xinlei Ge, and Daniel J. Jacob. 2020. "Fast Sulfate Formation from Oxidation of SO 2 by NO 2 and HONO Observed in Beijing Haze." *Nature Communications* 11(1):2844.

Weiner, Ruth F., and Robin A. Matthews, eds. 2003. "Chapter 18 - Meteorology and Air Pollution." Pp. 351–74 in *Environmental Engineering (Fourth Edition)*. Burlington: Butterworth-Heinemann.

Wesseling, Joost, Henri de Ruiter, Christa Blokhuis, Derko Drukker, Ernie Weijers, Hester Volten, Jan Vonk, Lou Gast, Marita Voogt, Peter Zandveld, Sjoerd van Ratingen, and Erik Tielemans. 2019. "Development and Implementation of a Platform for Public Information on Air Quality, Sensor Measurements, and Citizen Science." *Atmosphere* 10(8):445.

WHO. 2013. *Review of Evidence on Health Aspects of Air Pollution – REVIHAAP Project: Final Technical Report*. 07-0307/2011/604850/SUB/C3. World Health Organization.

Yandex LLC. 2020. "ClickHouse DBMS." Retrieved September 3, 2020 (https://clickhouse.tech/).