Understanding the true effects of the COVID-19 lockdown on air pollution by means of machine learning

Mario Lovrić[1, *], Kristina Pavlović[1], Matej Vuković[2], Stuart K. Grange[3,4], Michael Haberl[5], Roman Kern[1]

[1] Know-Center, Inffeldgasse 13/6, AT-8010 Graz

[2] Pro2Future, Inffeldgasse 25F, AT-8010 Graz

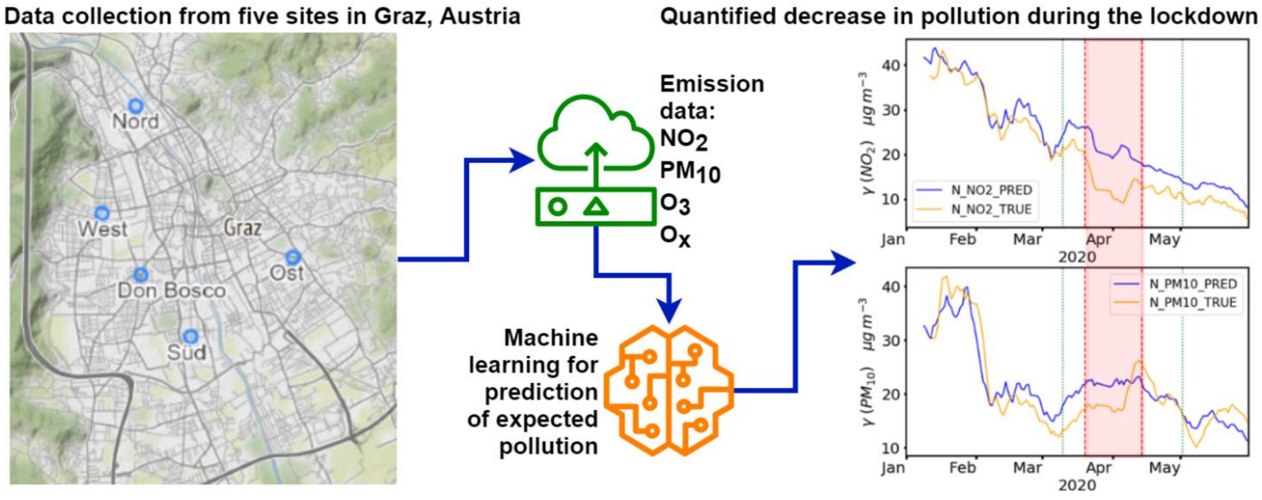
[3] Empa, Swiss Federal Laboratories for Materials Science and Technology, 8600 Dübendorf, Switzerland

[4] Wolfson Atmospheric Chemistry Laboratories, University of York, York, YO10 5DD, United Kingdom

[5] Institute of Highway Engineering and Transport Planning, Rechbauerstraße 12, AT-8010 Graz

*Corresponding author: mlovric@know-center.at

Keywords: PM_{10}, NO_2, Total oxidant, O_x, O_3, Random forest, Corona crisis



Graphical Abstract

Abstract

During March, 2020, most European countries implemented lockdowns to restrict the transmission of SARS-CoV-2, the virus which causes COVID-19 through their populations. These restrictions had positive impacts for air quality due to a dramatic reduction of economic activity and emissions. In this work, a machine learning approach was designed and implemented to analyze local air quality improvements during the COVID-19 lockdown in Graz, Austria. The machine learning approach was used as a robust alternative to simple, historical measurement comparisons for various individual pollutants. Concentrations of NO_2 (nitrogen dioxide), PM_{10}, O_3 (ozone) and O_x (total oxidant) were selected from five measurement sites in Graz and were set as target variables for random forest regression models to predict their expected values during the city's lockdown period. The true vs. expected difference is presented here as an indicator of true pollution during the lockdown. The machine learning models showed a high level of generalization for predicting the concentrations. Therefore, the approach was suitable for analyzing reductions in pollution concentrations. Results on the validation set showed very good performance for O_x and NO_2 when compared to PM_{10} and O_3. The analysis indicated that the city's average concentration reductions for the lockdown period were: -36.9 to -41.6%, and -6.6 to -14.2% for NO_2 and $PM_{10,}$ respectively. However, an increase of 11.6 to 33.8% for O_3 was estimated. The reduction in pollutant concentration, especially NO_2 can be explained by significant drops in traffic-flows during the lockdown period (-51.6 to -43.9%). The results presented give a real-world example of what pollutant concentration reductions can be achieved by reducing traffic-flows and other economic activities.

Introduction

The COVID-19 pandemic has caused disastrous health and socio-economic crises across the globe [1], [2]. Questions have been raised whether atmospheric pollution is a co-factor in disease development causing a higher lethality rate, especially in highly populated and polluted areas such as those in Italy [3], [4]. A study from China suggests there is a statistically confirmed relationship between air pollution by means of elevated concentrations of $PM_{2.5}$, PM_{10}, CO, NO_2 and O_3 and the COVID-19 infection rate [5]. An interplay of air quality and the pandemic seems obvious.

On the other side, lockdowns have caused significant changes in air quality [6]. A study on 44 Chinese cities [7] showed a decrease in main air pollutants from 5.93-24.67% during the lockdown while megacities such as Sao Paulo showed even higher concentration drops (40-70%) for some pollutants [8]. A study on $PM_{2.5}$ in capital cities showed concentration drops of 20-60% during the COVID-19 crisis [9]. It is suggested that the pollution drop was mainly driven by a reduction in traffic [10] and industrial activities [11]. Even if lockdowns hinder economic growth and might cause various negative effects in the long term, drops in pollution concentrations may act as another factor which slows disease transmission in tandem with limiting human contact. Lockdowns in Europe were instituted gradually by means of governmental interventions [12]. This massive intervention also poses a unique opportunity to study the change in various aspects of air quality, thus motivating our study.

We discuss and explore that for complete understanding of the true factors influencing pollutant concentrations, pure statistical tests or single-day comparisons might be inadequate since weather conditions, particle persistence and seasonality affect concentrations by linear and non-linear processes [13]. We investigate the effects of the gradually introduced lockdown on air quality in an urbanized area in Graz, Styria, Austria. Due to the high degree of traffic influence, we have included traffic data into our analysis. Furthermore, we have investigated in detail which of the pollutants' concentrations were more affected by the lockdown. We have employed machine learning to understand the true effects of the intervention measures and discern them from random and other factors [5], such as weather conditions. As such, the outcome of our study serves as a guide for future interventions and their expected associated change in the pollutants' concentration changes.

Materials and methods

Our study contains traditional exploratory statistical analysis, such as principal component analysis (PCA) to explore the key attributes. In addition, the main analysis is based on machine learning models build to capture the historical relationships between the attributes and compare the predictions after the interventions took place. In particular, we utilize historical data which matches the time frame of the lockdown for the preceding years. We further include traffic data and present the drop in mobility. First, we present the data sources and the data cleaning and preprocessing procedures.

Data

We collected environmental, pollution and weather data from publicly available sources provided by the Austrian government[1]. To have a realistic picture of air quality during the lockdown, we analyzed long term measurement data from January 2014 to May 2020 from five measurement sites in the Austria city of Graz (Süd (*eng. South*) - S, Nord (*eng. North*) - N, West (*eng. West*) - W, Don Bosco – D, Ost (*eng. East*) – O); Figure 1). The latter two sites are situated on arterial roads with high traffic volumes, especially during the morning and evening rush hours. The most polluted measurement site of Graz is Don Bosco that struggles to meet annual NO_2 and PM_{10} regulatory limits of the EU-Council directive 96/62/EC. This is primarily because of traffic related emissions, but also because of emissions from a nearby steel- and iron-mill [14]. Although Graz East is located at a heavily frequented commuter-arterial the situation is not that severe. Graz South is situated at a secondary road segment but also records higher pollutant concentrations due to an industrial complex nearby. Graz North and West are classed as urban background sites and are located near to minor roads with no specific emission contributors in immediate vicinity. A more detailed site description, photos of the sites and historical overview of the sites is given in Moser et al., 2019.

To understand the potential effect of traffic, traffic flow was accessed for the city of Graz. The traffic flow data were mainly measured with inductive loop detectors where the detectors measure the change in field when objects pass over them. Once a vehicle drives over a loop sensor, the loop field changes which allows for the detection of the presence of an object (a vehicle). The “Traffic control and street lighting unit of the city of Graz” monitors and records the data at one-minute time frequency and provided data from January 2017 to May 2020 for two sites, namely Don Bosco and Ost.

To determine the lockdown time frame we extracted data from a list of governmental decisions and intervention measures during the lockdown available from[2] and published in [12]. It consists of country codes, dates and measures countries took to control the pandemic.

[1] https://www.umwelt.steiermark.at/cms/ziel/2060750/DE/

[2] https://github.com/amel-github/covid19-interventionmeasures

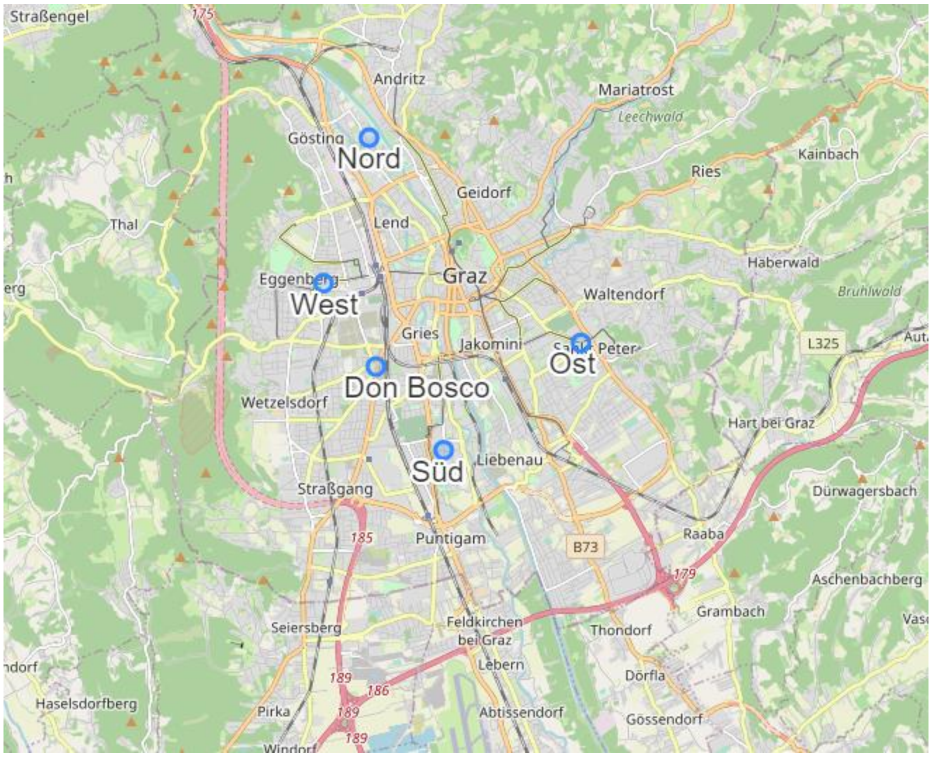


Figure 1. A city map of Graz indicating the five measurement sites: Süd – 47.041692° N, 15.433078° E; Nord – 47.09437° N, 15.415122° E; West - 47.069506° N, 15.403728° E; Don Bosco – 47.055617° N, 15.416539° E; Ost – 47.059530° N, 15.466634° E.

Data processing

Due to a very high correlation between $PM_{2.5}$ and PM_{10} (up to 97% for Graz South), we did not take $PM_{2.5}$ into account for this study. The variables in the accompanied dataset are abbreviated as follows: <site>_<pollutant>, *i.e.* S_NO2 would be NO_2 measured on the Graz South measurement site. The measurements analyzed were daily means.

Some data points were excluded from the analysis due to the presence of outliers. Observations between 1st and 3rd January each year were excluded due to high PM_{10} concentrations caused by New Year firework shows. Additionally, PM_{10} observations between 26th and 30th March, 2020 were excluded because of abnormally high values driven by a Saharan dust event (Federal Office: MeteoSwiss, 2020; Hansen, n.d.; also Supplementary Figure 1). Total oxidant (O_x; NO_2 + O_3) was calculated and included in the analysis as an additional pollutant [18]. O_x was included because it will indicate if the hypothesized changes in NO_2 and O_3 due to the lockdown measures were caused by a repartitioning of these two species which has consequences for air quality management.

Table 1. Site description for the air pollution measurement sites, data taken from [15] and http://app.luis.steiermark.at/luft2/suche.php. The number in brackets shows the amount of missing values in data, referring to 2324 values (full number of data entries).

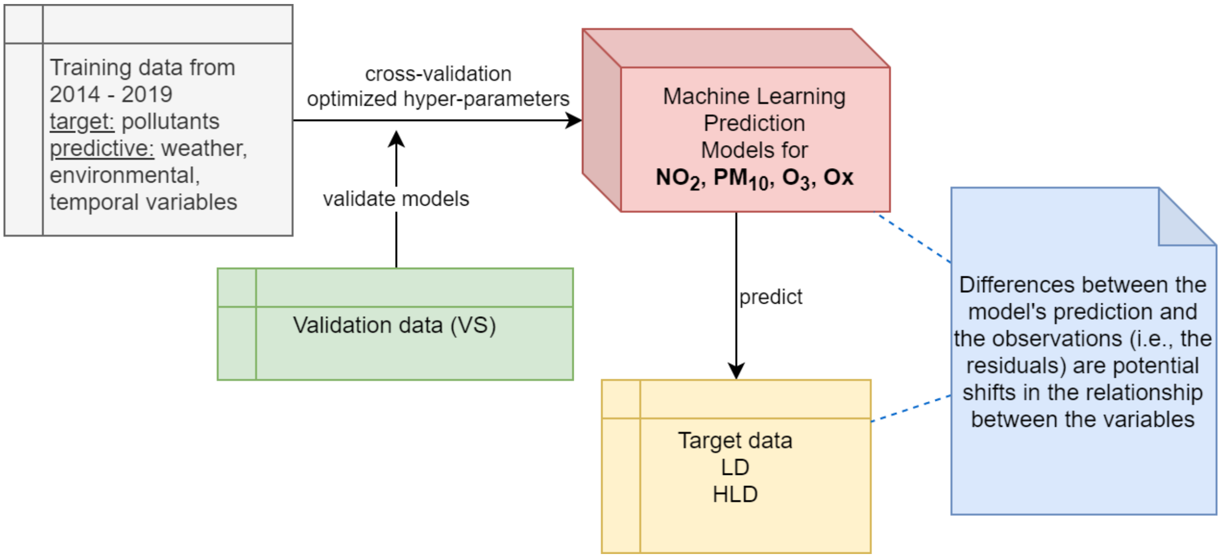
Measurement	**Don Bosco (D)**	**Nord (N)**	**Ost (O)**	**Süd (S)**	**West (W)**
O_3 [μg/m^3]		x (23)		x (39)	
PM_{10} [μg/m^3]	x (13)	x (14)	x (15)	x (31)	x (30)
NO_2 [μg/m^3]	x (3)	x (8)	x (10)	x (12)	x (30)
%RH	x (0)	x (12)	x (1244)	x (21)	x (8)
Air temperature [°C]	x (0)	x (12)	x (1239)	x (45)	x (0)
Precipitation [l/m^2]		x (12)			
Wind speed [m/s]		x (12)	x (1239)	x (4)	x (0)
Wind direction [Deg]		x (12)	x (1239)	x (4)	x (0)
Air pressure [mbar]		x (12)	x (1239)		

Inspired by Grange et al., 2018a; Šimić et al., 2020, we created binary encoded temporal variables for season, month, weekday, and day of year. The other predictive variables were relative humidity, air pressure, air temperature and precipitation, wind direction and wind speed. Variables with 1000+ missing values were removed, variables with rare missing values were imputed by backfilling. The number of missing value points is provided in Table 1. The processed data consists of 2324 days and 60 variables in total and is provided in table format within a persistent data repository [20]. The traffic data were aggregated to a day frequency and stored as a time series for the two sites (O, D). The processed traffic data ranges from January 2017 to May 2020.

Exploratory analysis and machine learning methods

The data was analyzed by means of explorative analysis and regression models. Data exploration was conducted by means of swarm plots (showing distribution of median concentrations in the HLD time frame over the years) and principal component analysis (PCA). For a comprehensive discussion of PCA, readers are advised to consult Abdi and Williams, 2010. We used PCA to inspect any cluster formation among the measurement sites and pollutants. Machine learning was employed to model the expected concentrations (as if no lockdown happened). For machine learning one needs a set of predictive variables and a target variable. The methodology follows closely the procedures described previously in [13]. The target variables in the models are the various pollutant concentrations. The predictive variables (X) are weather and environmental conditions as well as temporal variables accompanied by their respective lag-values (values from the previous two days). These predictive variables allow the machine learning model to capture seasonal behavior including activities like industrial production, and traffic (surrogate variables).

The machine learning algorithm used was random forest regression (RF) [22] which has been utilized in a number of previous air pollution models and air quality data analysis studies [13], [19]. The differences between past work and this analysis include: the exclusion of lag-values of the respective (predicted) pollutant and exclusion of other pollutant concentrations in the respective models. Data processing and model training was conducted with Python. The functions, scripts and libraries are presented in prior work [13], [23].


Training data from
2014 - 2019
target: pollutants
predictive: weather, environmental, temporal variables
cross-validation
optimized hyper-parameters
validate models
Validation data (VS)
Machine Learning Prediction Models for
NO_2, PM_{10}, O_3, Ox
predict
Target data
LD
HLD
Differences between the model's prediction and the observations (i.e., the residuals) are potential shifts in the relationship between the variables

Schema 1: Overview of the study methodology in order to detection changes in the relationships between the (dependent and independent) variables. Table 2 conveys more details on the methodology.

The proposed investigation of the lockdown pollution consists of training RF models for each of the pollutants' concentration (as the target) and observing the difference of predicted to the true values, a concept presented in [24] and depicted in Schema 1. That is, any increase in residuals can be mainly attributed to changes directly or indirectly associated with the lockdown. To achieve models with good generalization we used Bayesian optimization and feature selection by means of permutation importance [23]. The hyperparameter optimization of the RF models was conducted with a 10-fold cross-validation. The period of model training was between 3rd January 2014 and 31st December 2019. The data from 2020 was separated as an external validation set, VS (3rd January 2020 – 10th March 2020), a lockdown set, LD (10th March 2020 – 2nd May 2020) and a hard lockdown set, HLD (20th March 2020 – 14th April 2020) being the time frame of interest in this work. The trained models were then used to predict on the VS and the HLD. An overview of the dataset splits together with our expectation is given in Table 2.

Table 2. An overview of the dataset splits and their utilization in this study.

Name	From	To	Usage	Description	Expectation
Training	3rd January 2014	31st December 2019	ML model training	Training data for the ML models	Training data is sufficient to build a good predictive model
HLD 2014-2019	20th March 20xx	14th April 20xx	Comparison	Subset of the training data, for the relevant time frame in the years up to 2019	Long term comparison with HDL 2020, expect a pronounced drop in pollutant concentration
HLD 2019	20th March 2019	14th April 2019	Comparison	Subset of the training data, for the hard lockdown time frame the year prior to lockdown	Short term comparison with HDL 2020, expected drop in pollutant concentration (but closer match than long term)
VS	3rd January 2020	10th March 2020	Validation	Validation data set, prior to the lockdown	Good prediction quality, i.e., unchanged relationships between the variables w.r.t. to the training time frame
LD	10th March 2020	2nd May 2020	Prediction	Entire lockdown phase, including early measures and the initial easing phase	Loose fit between model and observations, due to exogenous factors, e.g., change in behavior
HLD	20th March 2020	14th April 2020	Prediction	Phase of enforced tight lockdown	Residuals expected to be contributed to lockdown measures

Concept validation and method comparison

The obtained machine learning results, that is, the predicted (expected) pollutant concentrations were compared to the measured (true) values. Our machine learning approach was then evaluated against historical changes in the data (short and long term). To understand the potential causes we evaluated the drop in pollutant concentration against the drop in traffic density. Furthermore, the concentration drops were aggregated to understand the city average concentration drops as an indicator for overall air pollution.

Results and Discussion

Explorative analysis

To gain a better understanding of the relationship between the pollutants we first conducted a principal component analysis (PCA) on the pollutants' concentration data from 2014 to 2020 (Figure 2). The PCA loadings plots show that the distinct measured pollutants group based on their chemical composition, *i.e.* the same pollutants group together independent of the measurement site. Despite PM_{10} have a wide range of sources, this pollutant was similar to NO_2, indicating that the PM_{10} concentrations in Graz were primarily sourced from traffic processes during the analysis period. One can observe that in the two groups (PM_{10} and NO_2), the northern site N has lower PC2 loadings. Furthermore, for NO_2 and PM_{10}, N is closer to W, which could be explained by both sites being less burdened by traffic emissions. On the contrary for the Don Bosco site (D), NO_2 and PM_{10} are close in the plot, which might be explained by a rather common source, being traffic since it is site with dense traffic. O_3 showed distinctly different patterns because O_3 is generated by secondary processes and not directly emitted. It can be inferred that the PC2 is more weighted by traffic.

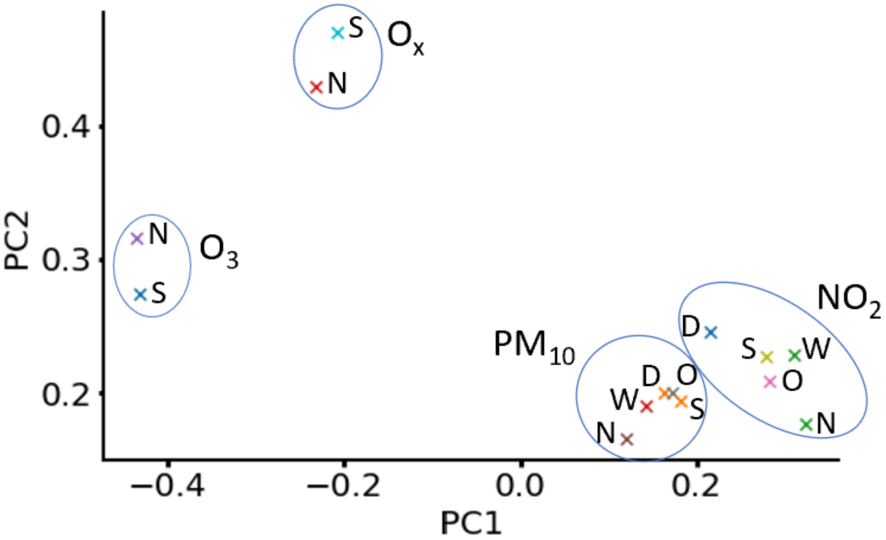


Figure 2. A PCA loadings plot calculated from pollutant concentrations (PM_{10}, NO_2, O_x, O_3) from five sites in (D, N, O, S, W) Graz, Austria, described in Figure 1. The pollutants form groups by means of chemical composition.

To understand the changes during the lockdown, we compared the HLD data in the time frame (20th March --14th April 2020) to the same time frames during 2014-2019. The swarm plots of the pollutants' concentrations are presented in Figure 3. It can be seen that in the given time frames, particularly for NO_2, the concentrations were lower in 2020 as compared to the years 2014 to 2019. PM_{10} and O_x do not show clear patterns, whereas O_3 appears to have higher average concentration compared to data from the 6 years before. Therefore, a general trend is present, traffic-sourced pollutants show concentration drops, while PM_{10} shows a drop at Don Bosco, which is a traffic-burdened site.

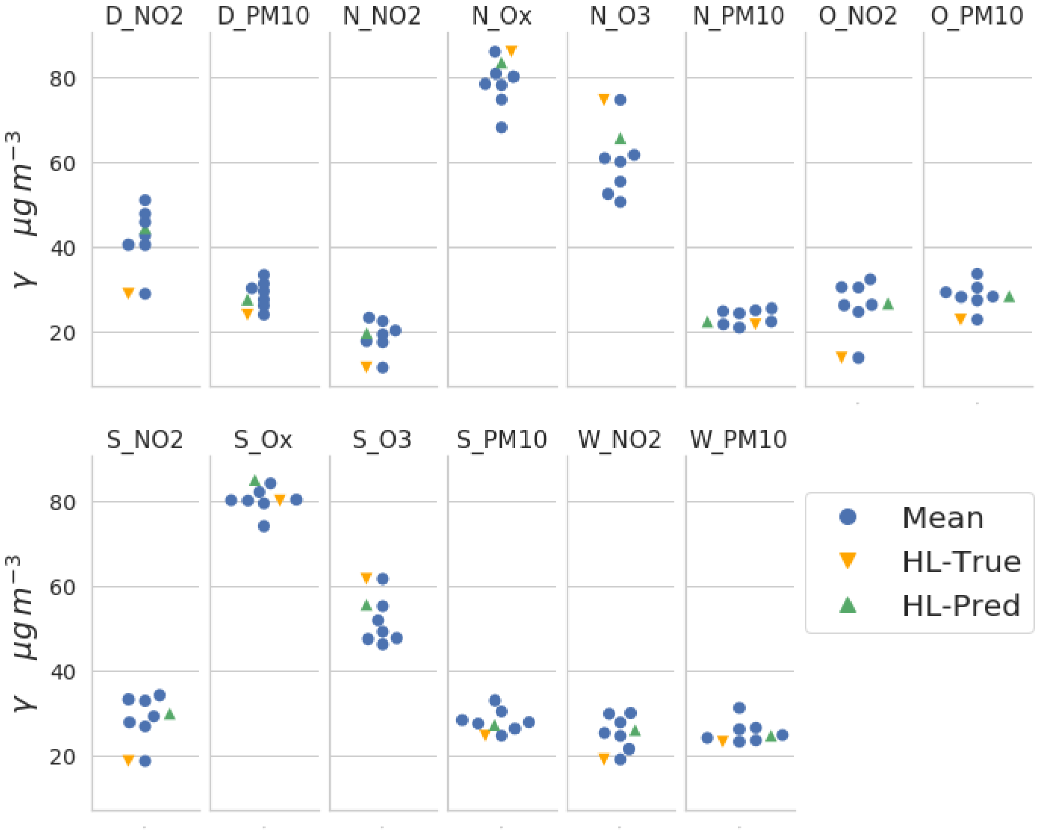


Figure 3. Swarm plots of pollutant concentrations at the five sites during the hard lockdown (HLD) time frame (20th March – 14th April) through years 2014 to 2020. The 2020 value is colored according to whether it is true or predicted by machine learning (see Machine Learning results section).

Machine learning results

For understanding the true change in pollution, we trained RF models with pollutants' concentrations as target variables. The optimal/best models (results in Supplementary Table 1) obtained were re-trained on the complete training data sets between 2014 and 2019 and subsequently fitted to data from 2020 (VS, HLD, LD). The results of the RF modelling by means of the coefficient of determination (R^2), the root mean square error (RMSE) and the normalized RMSE (NRMSE%) [13] are given in Table 3. Time series plots (as 7-day moving averages) for the predictions (in 2020) and their respective true values for the six pollutants are shown in Figure 4.

Table 3. R^2, RMSE and %NRMSE value for the external validation and lockdown sets from the best model.

	R2 VS	**%NRMSE VS**	**%NRMSE SL**	**%NRMSE HL**	**RMSE VS**	**RMSE SL**	**RMSE HL**	**True Mean**	**True Max**	**True Min**
W_PM10	0.66	29.1	28.3	30.88	8.11	6.23	7.11	23.11	69.47	5.58
W_NO2	0.81	13.47	38.59	49.27	4.74	7.91	9.32	25.9	62.28	3.11
S_PM10	0.71	25.08	28.76	33.33	8.52	6.65	8.16	26.46	78.23	6.93
S_O3	0.84	33.37	17.35	13.9	6.43	9.47	8.56	39.33	86.13	1.08
S_ Ox	-0.07	21.46	13.78	8.18	11.92	10.31	6.56	65.6	97.77	28.67
S_NO2	0.45	20.04	53.97	68.26	7.26	10.9	12.67	26.26	65.05	2.92
O_PM10	0.42	38.78	37.27	36.96	13.44	8.02	8.4	26.34	111.51	7.59
O_NO2	0.5	22.35	80.1	103.82	7.63	12.3	14.18	23.12	58.14	1.98
N_PM10	0.72	28.81	27.96	30.15	7.09	5.33	6.51	20.71	65.56	5.6
N_O3	0.87	24.6	14.01	13.9	6.76	9.38	10.38	48.67	93.71	3.64
N_ Ox	0.5	12.46	8.96	7.1	7.19	7.17	6.11	68.73	106.89	34.14
N_NO2	0.76	17.52	59.89	81.29	5.3	7.83	9.25	20.06	53.36	1.14
D_PM10	0.61	26.29	29.63	35.98	9.04	7.22	8.6	27.58	78.77	6.63
D_NO2	**0.55**	**15.32**	**46.64**	**58.28**	**7.43**	**14.48**	**16.76**	**38**	**81.8**	**8.02**

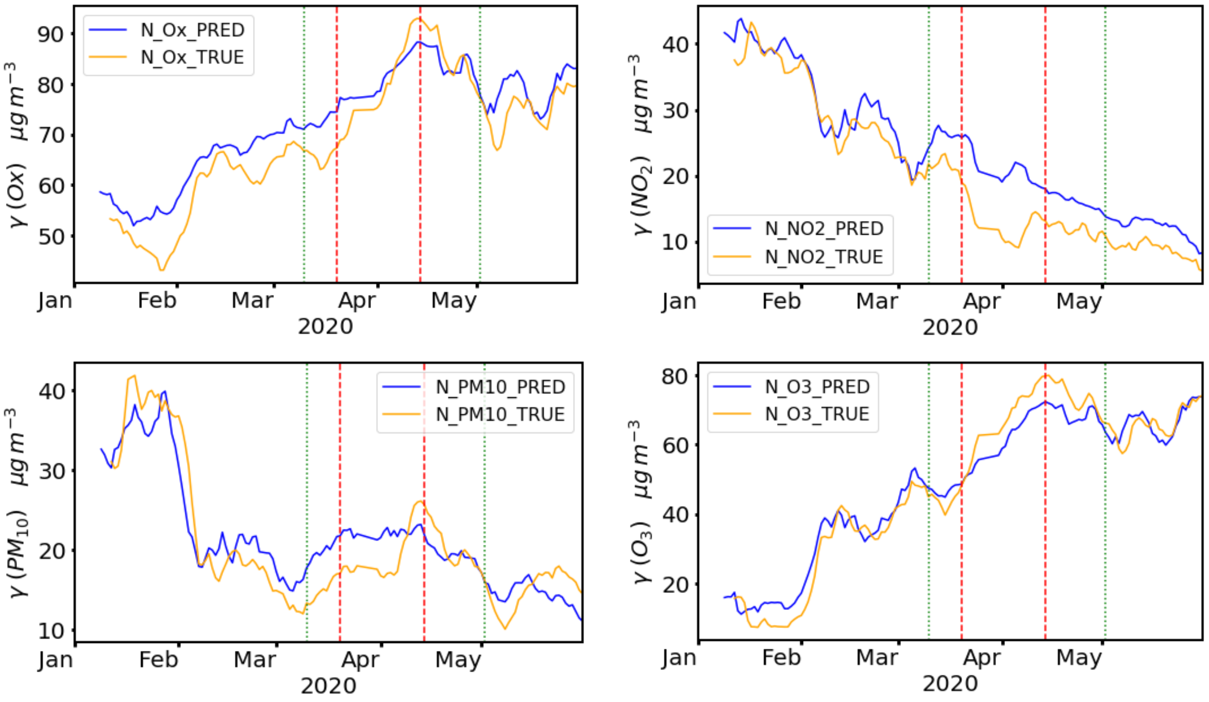


Figure 4. Time series plots for four pollutants' concentrations measured at Graz Nord (N). Orange line (measured) is compared to their predicted values (blue). The plots present a 7-day moving average (for better visibility) for the data in 2020. Prior to the green line is the validation set (3rd January 2020 – 10th March 2020). The green dashed lines show the LD time frame (10th March 2020 – 2nd May 2020) and red dashed lines show the HLD time frame (20th March 2020-03-20 – 14th April 2020). Top left is O_x, top right is NO_2, bottom left is PM_{10} and bottom right O_3. Additional plots are provided in Supplementary Figure 2.

Results from Table 3 and Figures 4 show that for the validation period (VS), the models' prediction quality by means of %NRMSE declines on average in following order $O_x < NO_2 < O_3 < PM_{10}$, while for the R^2 score the prediction declines as follows $O_x < NO_2 < O_3 < PM_{10}$. A number of additional processes such as long-range transport and secondary generation generally drives PM concentrations and these processes are not as relevant to the other gaseous pollutants analyzed here [25], [26]. This increased complication is likely the reason for decreased model performance for the PM_{10} pollutant.

NO_2 and O_3 have reasonable prediction performance (validation set), *i.e.*, O_3 has R^2 scores 0.84 (N) and 0.87 (S), while NO_2 has R^2 scores as high as 0.81 (W) and 0.76 (N) with lower scores for the traffic-loaded sites (S – 0.45, O – 0.50, D – 0.55). A similar pattern appears with PM_{10} where the lower R^2 scores are related to traffic-loaded sites (O – 0.42, D – 0.61). Better scores were achieved at less traffic-loaded sites (W – 0.66, N - 0.72) and S -0.71. These results overall suggest that the concept of using ML models can support understanding the true pollution.

For the PM_{10} and NO_2 pollutants, the models show concentration reductions (Figure 4 and Supplementary Figure 2) where the observed concentrations were lower than those which were predicted by the RF models. O_3 showed the opposite behavior where observed concentrations were higher than those predicted. The increases in O_3 concentrations can be explained by a reduction of the NO-O_3 titration cycle when NO_x emissions (and concentrations) were low during the lockdown period.

When comparing the ratios of %NMRSE VS to %NMRSE HLD (our proposed indicator for true pollution difference), one can see the largest ratios with NO_2 across the sites (2.77-3.98) meaning that they show the largest error in predicted (expected) vs true (measured) concentrations during the lockdown. The ratio is on average lower with O_3 (0.41 – 0.56). The inverse results are due to a concentration rise instead of a concentration drop. O_x shows a less clear pattern; underprediction for North and overprediction for South (Supplementary Figure 2). With PM_{10} the ML models suggest that the HLD and VS errors (0.95-1.36) do not deviate largely from each other, pointing to PM_{10} not being largely affected by the lockdown. Also, there is a period of unexpected high PM_{10} at the end of HLD which we attribute to lockdown fatigue. Regarding the sites, one can see the largest ratios across pollutants (%NMRSE VS to HLD) are at the East (O) and Don Bosco (D) sites which are more traffic-loaded than others.

Reduction in pollution/Method comparison

To support the contribution of machine learning in understanding the pollution we present here a method comparison. Pollution during HLD (median 2020) was compared to the respective medians of 2019 and 2014-2019 (*i.e.* historical comparison) as well as the median of the predicted HLD 2020. The comparison is presented in Figure 5.

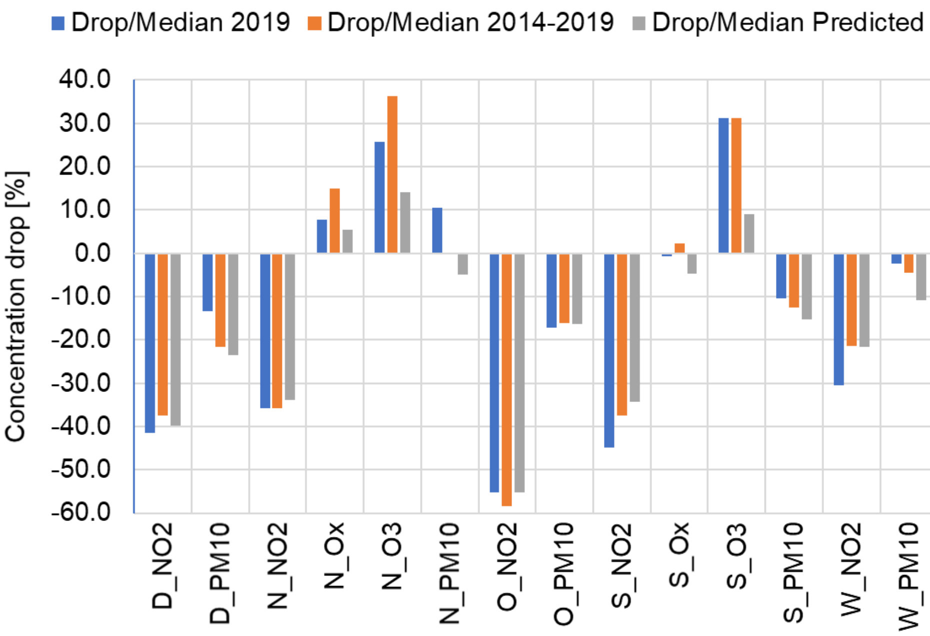


Figure 5. Calculated pollution reduction [%] from the median HLD values in the time frame (20th March - 14th April). The concentration drops i.e. their medians (HLD 2020 to 2019, *i.e.* short-term) are compared per pollutant to concentration drops (HLD 2020 to 2014-2019, *i.e.* long-term) and the concentration drops calculated from the predicted values (Data for the plot is provided in Supplementary Table 3).

The results show an overall good agreement for change in pollution of HLD 2020 referred to 2019 (short-term, in blue) and referred to 2014-2019 (long-term, in orange). Notable exceptions are: N_PM10 which shows an increase of 10.5% short-term and a 0.2% reduction long-term and W_NO2 showing a larger concentration drop in the short-term (30.6%) when compared to long-term (21.4%). One can also observe generally good agreement when comparing changes referred to 2019 and to the predictions of HLD 2020. Pollutants, where the results disagree (short-term vs predicted) the most are N_PM10, N_Ox, N_O3, S_O3, S_NO2 and W_NO2. For NO_2 and O_3, where good prediction results were achieved, one can assume that the true pollution change is lower than estimated by comparing it to values from previous years. The agreement between the methods validates the machine learning approach as an alternative for analyzing the pollution drop as no ground truth data are available during the lockdown.

For the pollutants where five measurement sites are available, the median concentration drops were averaged across all measurement sites (D, N, O, S, W) to a “city average concentration drop”. When comparing the city average concentration drops for 2020 to the short-term (vs HLD 2019), long-term (vs HLD 2014-2019) and prediction (*i.e.* predicted HLD 2020) one can see that for PM_{10} the ML models show a larger concentration drop whereas for NO_2 we see a smaller concentration drop with the ML models (Table 4).

Table 4. The median pollution reductions averaged across sites (D, N, O, S, W) in 2020 when comparing the HLD median to 2019, to HLD median 2014-2019 and to HLD predicted 2020 in percentage. (Data is provided in Supplementary Table 3)

	City average conc. drop vs 2019	City average conc. drop vs 2014-2019	City average conc. drop vs predicted
NO_2	**-41.6%**	-38.1%	-36.9%
PM_{10}	-6.6%	-11.0%	**-14.2%**

From the available traffic data, we calculated a reduction in median traffic during HLD 2020 against the same median time frame in 2019 and in 2017-2019 for the measurement sites D and O that are mainly traffic relevant. The traffic measured at the detector loops at these measurement sites showed a reduction of 45.6% at O and 51.6% respectively 43.9% at D, see Table 5. This reduction of traffic can be correlated with the massive reduction of NO_2 at measurement site D and O, see Figure 5, as traffic is especially one of the main sources of these compounds. In contrast the reduction of PM_{10} is not as pronounced. This demonstrates that traffic is just one part of the air quality problems concerning PM_{10}. Since also the industry was significantly curtailed in and around Graz, it was expected that PM_{10} should decrease more strongly. This indicates a wider range of influences on PM_{10} concentrations, and/or a high lag in the relationships involved in the PM_{10} related variables, *i.e.*, the duration of the intervention was too short to lead to a significant concentration drop in pollutant.

Table 5. Calculated traffic reduction in 2020 when comparing the HLD median to 2019 and to the 2017-2019 HLD median in percentage.

Year	Drop in traffic density (D)	Drop in traffic density (O)
Median 2019	-51.6%	-45.6%
Median 2017-2019	-43.9%	-45.6%

Conclusions

In this work we have examined air pollution concentrations during the COVID-19 lockdown for the city of Graz, Austria to gain better insights into the relative influences of the observed variables on a wide range of pollutants a historic event in human behavior. Besides using explorative methods, we employed random forest regression to differentiate between predicted (expected) values depending on environmental data (not affected by the lockdown) and the lockdown affected (true) pollution levels.

Our prediction models performed well for a series of pollutants indicating the selection of independent variables (predictors) is sufficient to explain changes in the observations as good generalization was observed for some of the pollutants. For PM_{10} and NO_2, the predicted values were found to be above the measured concentrations during the lockdown. O_3 was underpredicted during the lockdown time frame which is expected due to relationship with NO_x concentrations which were reduced during the lockdown because of much lower traffic volumes and therefore emissions. Our findings show that machine learning is a suitable tool to analyze pollution changes during events such as COVID-19 lockdowns. Although, the expected to true differences in pollutant concentration based on machine learning models showed similar results with regard to the historical comparisons, it was an important technique to employ because it enabled for far more robust comparisons with the observed time series.

Still, additional studies are needed with a wider scope (in terms of different geographical regions, additional possible influencing factors, as well as temporal analysis) to improve model generalization in order to obtain more better estimates of event-based air pollution reductions.

Declarations

Funding

The Know-Center is funded within the Austrian COMET Program – Competence Centers for Excellent Technologies – under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG. SKG is supported by the Swiss Federal Office for the Environment (FOEN) and by the Natural Environment Research Council (NERC) while holding associate status at the University of York.

Availability of data

The pollution data used for modelling is published [20]. There is no informed consent for making the traffic data accessible or published due to internal guidelines.

Competing interests

The authors declare that they have no conflict of interest.

Authors' contributions

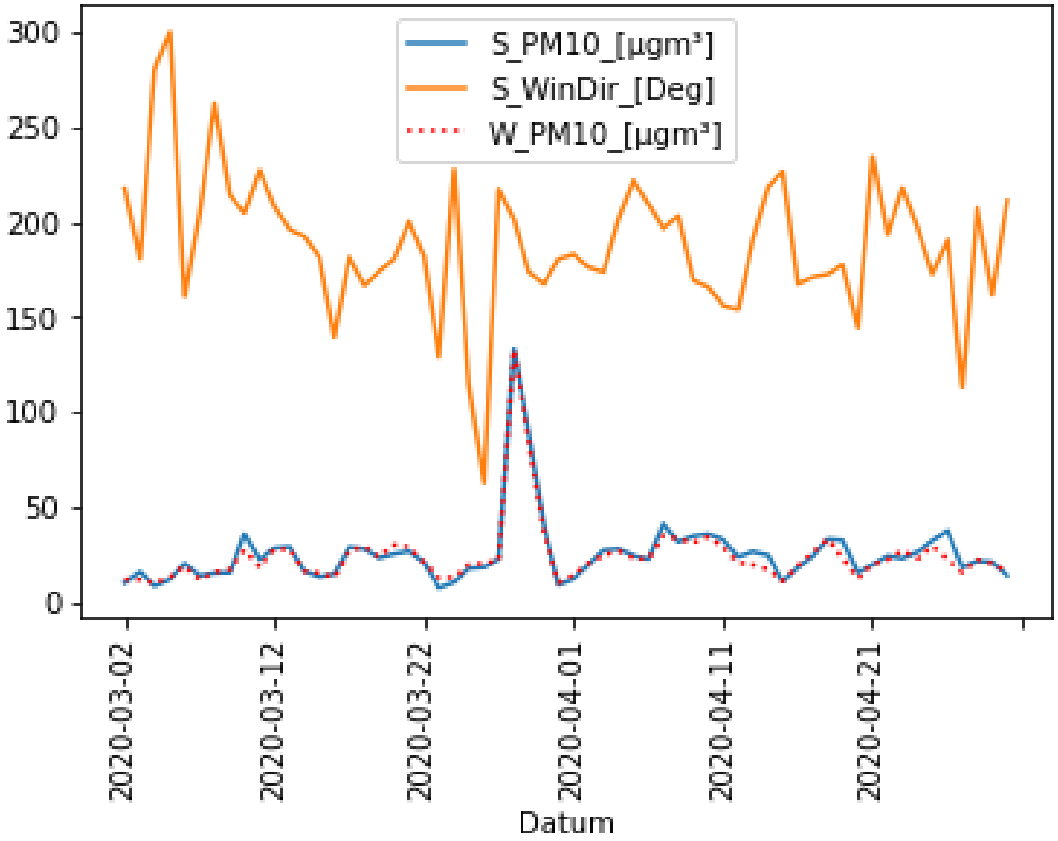
Mario Lovrić – Concept, Machine learning, Writing; Kristina Pavlović – Literature research, Visualization, Data collection and curation; Matej Vuković – Machine learning, Visualization, Writing; Stuart K. Grange – Writing, Interpretation, Revision and editing; Michael Haberl – Data collection and curation, Writing, Interpretation; Roman Kern – Concept, Writing, Supervision, Revision and editing

Acknowledgments

Not applicable.

Supplementary material

Supplementary Figure 1. A time series plot with an observation of a “Saharan dust” event. The plot shows two PM_{10} sites (Graz South and West) related to a change in wind direction observed at Graz South.

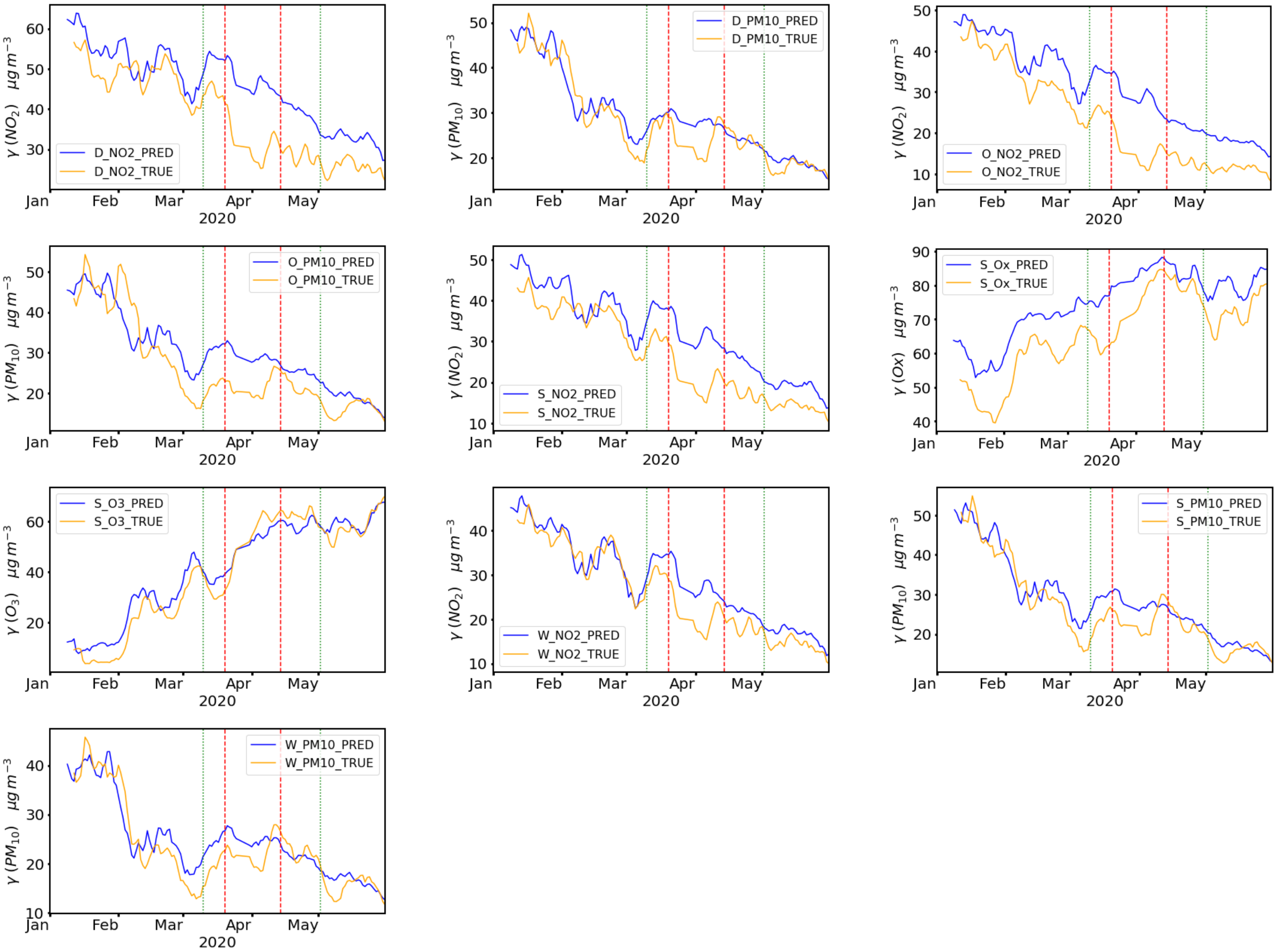


Supplementary Table 1. Hyperparameter optimization results from Random Forest training. The table shows the cross-validation error by means of CV-RMSE, whether the model used feature selection or not and how many features were included in the best model.

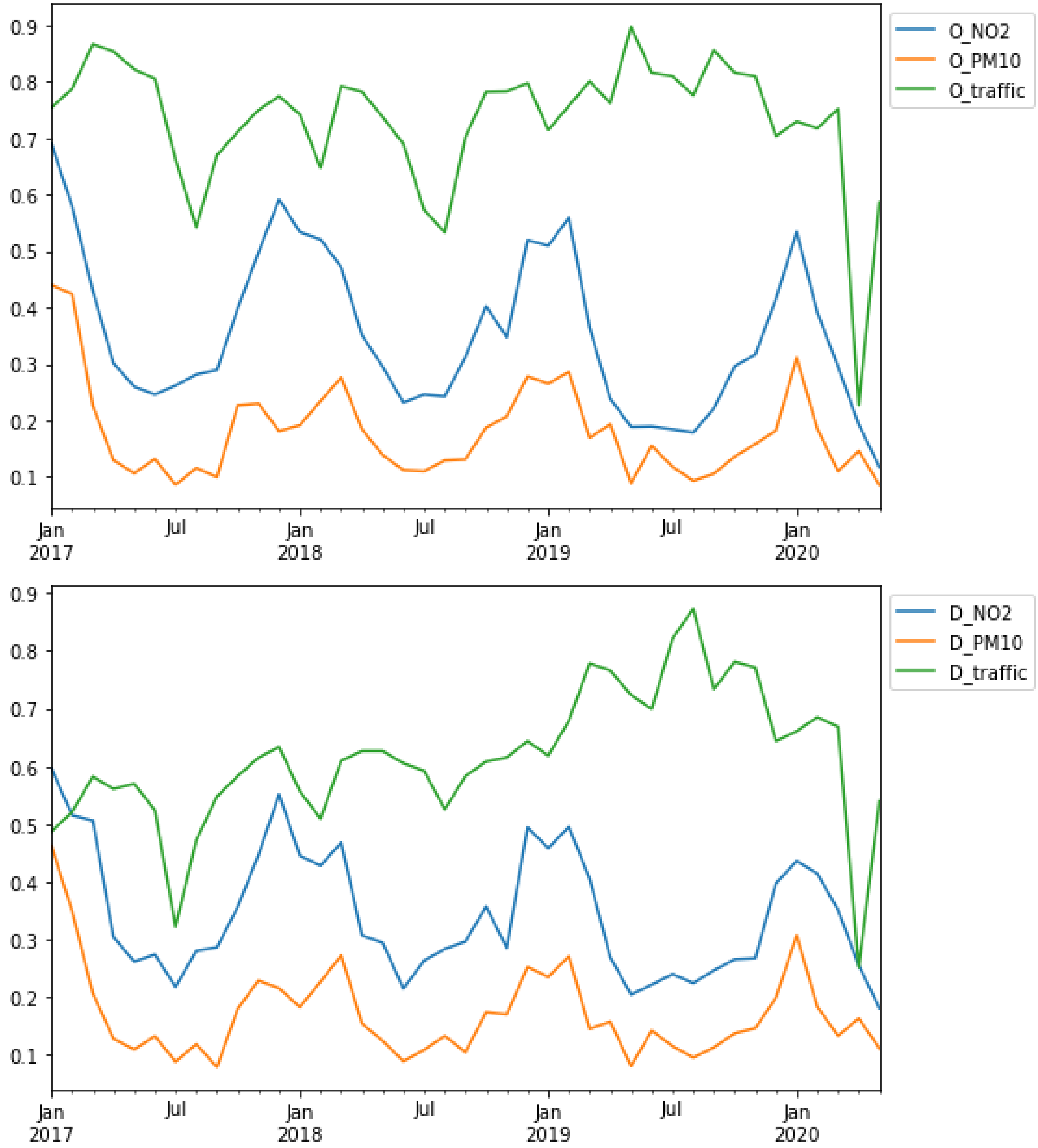
Pollutant	**CV-RMSE**	**Features**	**Feature selection**
D_NO2	7.82	66	True
D_PM10	10.24	66	False
N_NO2	5.55	66	False
N_O3	9.52	60	True
N_ Ox	8.92	66	False
N_PM10	8.98	66	False
O_NO2	6.29	64	True
O_PM10	10.29	66	False
S_NO2	7.00	60	True
S_O3	8.96	58	True
S_ Ox	9.25	66	False
S_PM10	10.44	60	True
W_NO2	6.36	66	False
W_PM10	9.63	66	True

Supplementary Table 2. Median values for each pollutant throughout years 2014 - 2020 during the hard lockdown timeframe (20th March – 14th April). The lowest value per pollutant is marked in bold with an asterisk

28.8	***16.8**	19.9	29.1	23.5	22.9	22.4	**W_PM10**
30.8	24.3	21.3	28.8	25.2	29.3	***20.3**	**W_NO2**
32.8	23.1	***22.4**	29.9	24.7	25.7	23	**S_PM10**
44.9	51.8	48.6	***46**	53.9	48.5	63.7	**S_O3**
81	85.4	74.8	***79**	79.2	82.7	82	**S_Ox**
34.7	27	26.3	31.6	25.8	34.6	***19**	**S_NO2**
32.1	***20**	23.9	30.5	29.6	28.3	23.4	**O_PM10**
31.2	21.8	26	29.3	31.4	26	***11.6**	**O_NO2**
23	***19.4**	19.8	23.4	25.5	19.7	21.8	**N_PM10**
56	62.8	52.7	***48.8**	60.4	60.7	76.3	**N_O3**
80.5	78.1	68.7	***74.3**	81.5	81.4	87.7	**N_Ox**
23.2	17.7	17.6	21.1	22.2	19.8	***12.7**	**N_NO2**
32.7	22.3	27.8	28.9	25.3	24.4	***21.1**	**D_PM10**
50.8	36.6	40.6	46.6	43.8	48.7	***28.5**	**D_NO2**
2014	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**year**



Supplementary Figure 2. Time series plots of exemplary pollutants' concentrations measured at all Graz sites which are not covered by Figure 4 in the Manuscript. Orange line (measured) compared to their predicted values (blue). The plots present a 7-day moving average for the data in 2020. Prior to the green line is the validation set (3rd January 2020 – 10th March 2020). The green dashed lines show the total lockdown time frame (10th March 2020 – 2nd May 2020) and red dashed lines show the hard lockdown time window (20th March 2020 – 14th April 2020).



Supplementary Figure 3. Time series plots vs traffic on the two traffic-loaded sites Graz, Don Bosco (top plot) and Graz, Ost (bottom plot). The values are monthly resampled to show a long-term pattern. All values were scaled to a 0-1 (MinMax) scale.

Supplementary Table 3. Calculated pollution reduction in % as a concentration drop in the median value based on the hard lockdown time frame (20th March – 14th April). The median of the HLD in 2020 is compared here to median of the same time frame in 2019, 2014-2019 and the predicted values in 2020.

	drop/median 2019	**drop/median 2014-2019**	**drop/median predicted**
D_NO2	-41.5%	-37.4%	-39.7%
D_PM10	-13.4%	-21.6%	-23.6%
N_NO2	-35.7%	-35.7%	-33.8%
N_Ox	7.7%	14.9%	5.5%
N_O3	25.7%	36.3%	14.1%
N_PM10	10.5%	-0.2%	-4.9%
O_NO2	-55.3%	-58.4%	-55.3%
O_PM10	-17.2%	-16.2%	-16.4%
S_NO2	-44.9%	-37.4%	-34.2%
S_Ox	-0.8%	2.3%	-4.7%
S_O3	31.3%	31.3%	9.1%
S_PM10	-10.4%	-12.6%	-15.2%
W_NO2	-30.6%	-21.4%	-21.7%
W_PM10	-2.3%	-4.6%	-10.9%

References

[1] W. Alabdulmonem, A. Shariq, and Z. Rasheed, "COVID-19: A global public health disaster.," *Int. J. Health Sci. (Qassim).*, vol. 14, no. 3, pp. 7–8, 2020, Accessed: Jun. 20, 2020. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/32536842.

[2] M. McKee and D. Stuckler, "If the world fails to protect the economy, COVID-19 will damage health not just now but also in the future," *Nature Medicine*, vol. 26, no. 5. Nature Research, pp. 640–642, May 01, 2020, doi: 10.1038/s41591-020-0863-y.

[3] E. Conticini, B. Frediani, and D. Caro, "Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy?," *Environ. Pollut.*, vol. 261, p. 114465, 2020, doi: 10.1016/j.envpol.2020.114465.

[4] D. Fattorini and F. Regoli, "Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy," *Environ. Pollut.*, vol. 264, p. 114732, 2020, doi: 10.1016/j.envpol.2020.114732.

[5] Y. Zhu, J. Xie, F. Huang, and L. Cao, "Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China," *Sci. Total Environ.*, vol. 727, no. December 2019, p. 138704, 2020, doi: 10.1016/j.scitotenv.2020.138704.

[6] F. Dutheil, J. S. Baker, and V. Navel, "COVID-19 as a factor influencing air pollution?," *Environ. Pollut.*, vol. 263, pp. 2019–2021, 2020, doi: 10.1016/j.envpol.2020.114466.

[7] R. Bao and A. Zhang, "Does lockdown reduce air pollution? Evidence from 44 cities in northern China," *Sci. Total Environ.*, vol. 731, no. 1954, p. 139052, 2020, doi: 10.1016/j.scitotenv.2020.139052.

[8] P. Krecl, A. C. Targino, G. Y. Oukawa, and R. P. Cassino Junior, "Drop in urban air pollution from COVID-19 pandemic: Policy implications for the megacity of São Paulo," *Environ. Pollut.*, vol. 265, pp. 19–21, 2020, doi: 10.1016/j.envpol.2020.114883.

[9] D. Rodríguez-Urrego and L. Rodríguez-Urrego, "Air quality during the COVID-19: PM2.5 analysis in the 50 most polluted capital cities in the world," *Environmental Pollution*, vol. 266. Elsevier Ltd, p. 115042, Nov. 01, 2020, doi: 10.1016/j.envpol.2020.115042.

[10] A. Kerimray *et al.*, "Assessing air quality changes in large cities during COVID-19 lockdowns: The impacts of traffic-free urban conditions in Almaty, Kazakhstan," *Sci. Total Environ.*, vol. 730, p. 139179, 2020, doi: 10.1016/j.scitotenv.2020.139179.

[11] L. Li *et al.*, "Air quality changes during the COVID-19 lockdown over the Yangtze River Delta Region: An insight into the impact of human activity pattern changes on air pollution variation," *Sci. Total Environ.*, vol. 732, 2020, doi: 10.1016/j.scitotenv.2020.139282.

[12] A. Desvars-Larrive *et al.*, "A structured open dataset of government interventions in response to COVID-19," Cold Spring Harbor Laboratory Press, May 2020. doi: 10.1101/2020.05.04.20090498.

[13] I. Šimić, M. Lovrić, R. Godec, M. Kröll, and I. Bešlić, "Applying machine learning methods to better understand, model and estimate mass concentrations of traffic-related pollutants at a typical street canyon," *Environ. Pollut.*, vol. 263, p. 114587, Apr. 2020, doi: 10.1016/j.envpol.2020.114587.

[14] M. Hinterhofer, "Anteil der verkehrsbedingten PM10 und PM2,5 Emissionen aus Abrieb und Wiederaufwirbelung an der Feinstaubbelastung in Österreich," Technische Universität Graz, 2014.

[15] F. Moser, U. Kleb, and H. Katz, “Statistische Analyse der Luftqualitätin Graz anhand von Feinstaub und Stickstoffdioxid,” Graz, 2019.

[16] Federal Office: MeteoSwiss, “Saharan dust events - MeteoSwiss,” 2020. https://www.meteoswiss.admin.ch/home/climate/the-climate-of-switzerland/specialties-of-the-swiss-climate/saharan-dust-events.html (accessed Jul. 31, 2020).

[17] K. Hansen, “More Dust Blows Out from North Africa,” *2020*. https://www.earthobservatory.nasa.gov/images/146407/more-dust-blows-out-from-north-africa (accessed Jul. 31, 2020).

[18] S. K. Grange and D. C. Carslaw, “Using meteorological normalisation to detect interventions in air quality time series,” *Sci. Total Environ.*, vol. 653, pp. 578–588, 2019, doi: 10.1016/j.scitotenv.2018.10.344.

[19] S. K. Grange, D. C. Carslaw, A. C. Lewis, E. Boleti, and C. Hueglin, “Random forest meteorological normalisation models for Swiss PM10 trend analysis,” *Atmos. Chem. Phys.*, vol. 18, no. 9, pp. 6223–6239, May 2018, doi: 10.5194/acp-18-6223-2018.

[20] M. Lovrić, K. Pavlović, R. Kern, S. K. Grange, M. Vuković, and M. Haberl, “Air pollution 01.2014 - 05.2020 (including COVID-19 lockdown) data from Graz, Austria,” Jul. 2020, doi: 10.5281/zenodo.3982670.

[21] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4. pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[22] L. Breiman, *Random Forests*, vol. 45, no. 1. 2001, pp. 5–32.

[23] M. Lovrić *et al.*, “Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: generalization, complexity or predictive ability?,” *chemrxiv*, Aug. 2020, doi: 10.26434/chemrxiv.12746948.

[24] M. Lovrić, L. Fadljević, R. Kern, T. Steck, J. Gerdenitsch, and E. Peche, “Prediction of anode lifetime in electro galvanizing lines by big data analysis,” 2020, Accessed: Apr. 21, 2020. [Online]. Available: https://www.researchgate.net/publication/340816111_PREDICTION_OF_ANODE_LIFE_TIME_IN_ELECTRO_GALVANIZING_LINES_BY_BIG_DATA_ANALYSIS.

[25] S. K. Grange, D. C. Carslaw, A. C. Lewis, E. Boleti, and C. Hueglin, “Random forest meteorological normalisation models for Swiss PM10 trend analysis,” *Atmos. Chem. Phys.*, vol. 18, no. 9, pp. 6223–6239, 2018, doi: 10.5194/acp-18-6223-2018.

[26] M. Viana *et al.*, “Source apportionment of particulate matter in Europe: A review of methods and results,” *Journal of Aerosol Science*, vol. 39, no. 10. Elsevier Ltd, pp. 827–849, Oct. 01, 2008, doi: 10.1016/j.jaerosci.2008.05.007.