Atmospheric Pollution Research 12 (2021) 101205

Contents lists available at ScienceDirect

Atmospheric Pollution Research

journal homepage: www.elsevier.com/locate/apr

External validation for statistical NO_2 modelling: A study case using a high-end mobile sensing instrument

Meng Lu [a,*], Ruoying Dai [b], Cjestmir de Boer [c], Oliver Schmitz [b], Ingeborg Kooter [c], Simona Cristescu [d], Derek Karssenberg [b]

[a] Department of Geography, University of Bayreuth Universitaetsstraße 30, 95447 Bayreuth, Germany
[b] Department of Physical Geography, Utrecht University, Princetonlaan 8a, 3584 CB, Utrecht, The Netherlands
[c] Netherlands Organization for Applied Research, TNO, Princetonlaan 6, 3584 CB, Utrecht, The Netherlands
[d] Institute for Molecules and Materials, Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands

ARTICLE INFO

Keywords:
Nitrogen Dioxide
Statistical modelling
Model validation
Hyperparameter optimisation
High-end mobile sensing instruments
Spatial prediction

ABSTRACT

Statistical learning models have been applied to study the spatial patterns of ambient Nitrogen Dioxide (NO_2), which is a highly dynamic, traffic-related air pollutant. Commonly, the validation process in most studies is based on bootstrapped split-sampling of training and test sets from fixed ground station measurements. As the ground stations distribute mostly sparsely over a region or country, this kind of cross-validation validation method does not consider how well models are capable of representing spatial variations in air pollution mostly occurring over distances shorter than the ground station sampling spacing. This may lead to inadequate hyperparameter optimisation and bias when comparing different statistical models. External mobile measurements are therefore needed for more reliable model evaluations as these provide detailed and spatially continuous information on air pollution patterns. However, most current designs of mobile NO_2 sensing instruments suffer from the trade-off between flexibility and measurement accuracy, as high-end sensors are commonly too heavy to be carried by a person or on a bike. In addition, sufficient repetitions over time are needed so that the measurements are representative to concentrations over a relatively long-term period. In this study, we installed a mobile air quality station onboard a cargo-bike to collect a dataset suitable for external validation. With the external validation dataset the model hyperparameter setting and statistical model comparison results alter. Our model comparison results also differ from previous studies relying only on ground stations for cross-validation.

1. Introduction

Chronic exposure to air pollution poses a threat to public health. The World Health Organisation estimated air pollution to contribute to 7 million deaths worldwide in 2016 (Organization et al., 2018). From a medical perspective, common air pollutants such as particulate matter and Nitrogen Dioxides (NO_2) damage the cardiovascular and respiratory systems (Anderson et al., 2012; Pascal, 2009). In the Netherlands, the NO_2 concentration limit set in 2017 was exceeded several times in areas with busy traffic (Atlasleefomegeving, 2020). Spatial prediction of NO_2 is needed for making scientific-based recommendations to reduce NO_2 emissions and meet the Goal 13 (topic: Atmosphere) of the Sustainability Development Goals 2030 (SDGs, 2017). Emissions from traffic can be direct and indirect. Atmospheric NO_2 is mostly traffic-related as an indirect secondary emission of an oxidation result of emitted NO, while the direct primary emission as NO_2 is minor (UK Department for Environment and Affairs, 2004; Atlasleefomegeving, 2020).

Ground monitoring stations of NO_2 can be routinely run or are project-oriented, which involves considerable investments (Hoek et al., 2008b). Recent studies show a rise in urban low-cost sensors in air quality studies (Spinelle et al., 2015; Schneider et al., 2017; Isiugo et al., 2018). Low-cost sensors for NO_2 mainly include electrochemical and metal oxide sensors, the former is based on chemical reactions between gases in the air and electrodes in the liquid in a sensor and the later is based on conductivity change of the sensing material. Low-cost ground sensors have been used to monitor air quality (Rai et al., 2017) and to understand spatiotemporal variations of air pollutant (Nagendra et al., 2019) and the relationships between the spatiotemporal variation of air pollutants and the urban environment (Miskell et al., 2018). They are used independently (Hasenfratz et al., 2015; Marjovi et al., 2015; Apte

* Corresponding author.
E-mail address: meng.lu@uni-bayreuth.de (M. Lu).

https://doi.org/10.1016/j.apr.2021.101205
Received 5 July 2021; Received in revised form 14 September 2021; Accepted 14 September 2021
Available online 18 September 2021

et al., 2017) or fused with ground monitoring stations (Gressent et al., 2020) for more accurate prediction (than using ground monitoring stations alone) of air pollution in space–time. Meanwhile, the deficits of low-cost sensors are obvious. Measurements from low-cost sensors are subject to sensor drift and interference effects. Sensor drift denotes the growing bias of sensor response due to the ageing electrochemical cell. Interference effects refer to the sensors' response to other pollutants, gases, temperature and relative humidity, leading to the measurements being presented lower (negative interference effects) or higher (positive interference effects) than the actual concentrations (van Zoest et al., 2019). For instance, an application of low-cost sensors in the Netherlands carried out in Amsterdam showed a significant signal drift over a period of two months. Calibration with temperature and relative humidity however improved the fit with ground stations (Mijling et al., 2018).

Methods for spatiotemporal mapping of NO_2 include dispersion models (Holmes and Morawska, 2006; Health Effects Institute, 2010), statistical models (Chen et al., 2019b), or hybrid models (Mölter et al., 2010b; Marshall et al., 2008; Beelen et al., 2010; Dijkema et al., 2010; Akita et al., 2014). Dispersion models simulate the emission, transformation, transportation and deposition of atmospheric particles, but require detailed emission inventory data and are computationally intensive; scaling a dispersion model towards larger areas can be at the cost of computational intractability. Statistical models aim at finding relationships between ground monitor station observations, satellite measurements, and ancillary NO_2 data, called geospatial predictors (Rivera et al., 2013; Park and Kwan, 2017; Kharol et al., 2015; Isiugo et al., 2018; Chen et al., 2019b; Lu et al., 2020a; Li et al., 2020; Chan et al., 2021). The geospatial predictors are variables that relate to the emission sources (e.g., transport network) and dispersion processes (e.g., meteorological data) of the pollutants (Briggs et al., 2000). The statistical approach has been used to predict high-resolution NO_2 at various spatial scales. Conventionally, linear regression methods have been used (Hoek et al., 2008a; Larkin et al., 2017) for mapping and analysing the pollutant sources. More recently, ensemble tree-based (Dou et al., 2021; Song et al., 2021) and neural networks-based (Li et al., 2020; Shams et al., 2021) machine learning methods have shown to be capable of further improving the prediction accuracy for the spatial and spatiotemporal prediction of NO_2. Environmental epidemiological studies have shown that mapping temporally resolved, long-term NO_2 climate, e.g. NO_2 of each hour of a day aggregated over multiple years (Lu et al., 2020b), to be preferable to account for human space–time activities in personal exposure assessment (Lu et al., 2019). However, validation thus far is mostly done against ground station measurements, which does not enable extensive validation of the spatial pattern of air pollution predictions. To validate spatiotemporal predictions regarding spatiotemporal patterns, external measurements providing more detailed spatiotemporal information is necessary.

Several studies compared the accuracy of high-resolution NO_2 mapping of various statistical models, including standard and regularised linear regression (Briggs et al., 2000), ensemble tree-based models (dos Santos et al., 2020), transformation-based (e.g. support vector machine), and artificial neural networks (Chen et al., 2019b; Kerckhoffs et al., 2019; Lu et al., 2020a). Kerckhoffs et al. (2019) modelling UFP (Ultra Fine Particles) and Chen et al. (2019b) modelling NO_2 over several European countries, concluded that the ensemble tree-based and linear regression models obtain similar prediction accuracy. However, neither of these two studies showed or evaluated the gridded prediction (i.e. the predicted continuous surface of NO_2). Also, in the global NO_2 modelling study of Lu et al. (2020a), it is found that various ensemble tree-based methods obtaining almost the same accuracy based on Cross-Validation on Ground Stations (CVGS) may give very different prediction patterns. There is thus a need to evaluate prediction models by comparing modelled and observed spatial patterns of NO_2 and preferably use observed patterns also in model building additional to CVGS.

Using external air pollution measurements that measures local spatial variation for validating air pollution models (both statistical and numerical) has been shown to be necessary for a reliable model evaluation (Khreis et al., 2018), however, remains a challenge due to limited air pollution measurements and the selection of an external validation dataset (Liu et al., 2018; Khreis et al., 2018). Most studies used external ground monitoring measurements for external validation. Liu et al. (2018) used ground station measurements to correct numerical models through the application of a correction factor to the ground station measurements. Mölter et al. (2010a) used air dispersion model output to LUR models and ground monitoring stations for validation. Ren et al. (2020) used monitoring stations that do not fulfil the 75% completeness criteria (i.e. less than 25% of data is missing) as external validations additional to other cross-validation strategies for a comparison between machine learning and LUR models. Similarly, Chen et al. (2019b) used external ground station measurements to evaluate different statistical air pollution models for comparison and concluded the strengthened model evaluation by the use of an external evaluation dataset. Khreis et al. (2018) evaluated the impact of using different ground measurements to validate statistical and numerical to the model evaluation results. The use of low-cost sensors in model validation are less frequent, likely due to the challenge in quality control and assurance. Lu et al. (2019b) used low-cost sensors (PurpleAir sensors) to evaluate a LUR (Land Use Regression) statistical $PM_{2.5}$ model. They compared block-level LUR predictions with PurpleAir sensor measurements and found higher values of the PurpleAir sensors and that the LUR predictions explain on average 20% of the variations from the PurpleAir sensor observations.

The objectives of our study are to (1) design a study to collect high quality measurements of air pollutants that are suitable for validating statistical models, and (2) use these measurements to evaluate and compare statistical NO_2 models built using the official national ground air quality monitoring network. We will answer the question if we can better optimise hyperparameters (parameters whose values are used to control the model training process and are not obtained from model training) in the models using these measurements and whether model comparison results differ from those based on CVGS.

We monitor NO_2 concentrations by installing a mobile air quality station onboard a cargo-bike to evaluate temporally resolved NO_2 models based on different statistical algorithms. Specifically, we focused on comparing three methods, Lasso (Tibshirani, 1994), Random Forest (RF, Breiman, 2001), and eXtreme Gradient Boosting (XGBoost, XGB, Chen et al., 2019a). These methods are selected as they are representative for the most recent spatial prediction techniques for NO_2 mapping, with Lasso and Random Forest compared in Chen et al. (2019b), Kerckhoffs et al. (2019) and Lu et al. (2020a). We compared the Lasso, RF, and XGB models of the corresponding hours with the cargo-bike measurements to understand the amount of spatial variability a statistical model could capture, and to compare different models and hyperparameter settings beyond CVGS accuracy of the models.

2. Materials and methods

2.1. Data

2.1.1. Cargo-bike and instruments

The cargo-bike (Fig. 1) was designed to carry reference apparatus in a compact package, to operate with optimal freedom of movement while providing reliable measurement data. It was equipped with two 12,8 V/100 Ah LiFePo4 batteries (Victron), a high-efficiency MultiPlus Compact inverter/charger (Victron) and a 150Wp solar panel. The cargo-bike and the instruments onboard weight 160 kg, has an e-bike support motor, and can operate 3–5 h continuously.

The air quality monitoring apparatus were: 42i NO_x monitor, 49i O3 monitor (Thermo Fisher Scientific), model 3321 APS (TSI), WXT536 climate sensor (Vaisala), MI70 + probe GM70 CO2 sensor (Vaisala),

GPS, a PC and several low-end air quality sensors. The housing was custom-made from sandwich panels and extruded aluminium profiles.

The cargo-bike covered a route shown in Fig. 2, in Nijmegen, the Netherlands, from July 16 to 19, 2019. The total route length is 29 km. The cargo-bike measured NO_2 every morning from 8 am to 11 am, the route was repeated daily. A two-point calibration was performed on the Thermo 42i NO_x monitor using synthetic zero gas (0 ppb) and calibration gasses of NO (410 ppb) and NO_2 (366 ppb).

We compared the cargo-bike measurements with ground station sensor measurements for validation. Along the cargo-bike route, there are two national ground stations established, one is the Nijmegen-Graafseweg station (called Graafseweg station, Latitude: 51.941372, Longitude 5.857777), which monitors mainly air pollution from traffic. The other is Nijmegen-Ruyterstraat station (called Ruyterstraat station, Latitude: 51.838221, Longitude: 5.856938), which monitors mainly background pollution. The stations are managed by RIVM. Over the period of the cargo-bike measurements considered, the mean NO_2 level is 23.25 $\mu g/m^3$ and 14.03 $\mu g/m^3$, respectively, for the Graafseweg and the Ruyterstraat stations. We acquired minutely ground station measurements and did the comparison using the cargo-bike and ground station measurements 2 min before and after when the cargo-bike was at the national ground stations (we cycled slower when we were close to the two ground stations), to account for differences in sampling duration and frequency between the cargo-bike and the ground stations. To facilitate the comparison, we also averaged the ground stations and cargo-bike measurements, respectively. Fig. 3 shows that the cargo-bike measurements matches better with the measurements from the Ruyterstraat station than the Graafseweg station. This could be explained by the much higher traffic intensity around the Graafseweg whereas the Ruyterstraat is in a neighbourhood away from the larger busy street. The cargo-bike is closer to the emission outlets of vehicles and has a higher sampling rates. These may explain that they show higher variations compared to the ground station measurements. A notably high value of a cargo-bike measurement on 17-07 may be caused by a heavy-emission vehicle passing by.

To use the cargo-bike measurements as external validations of statistical models, which are trained on ground station measurements in unit $\mu g/m^3$ (micrograms per cubic metre). The cargo-bike measurements are converted from ppb to $\mu g/m^3$ under the assumption of the ideal gas behaviour. Specifically, we firstly use the ideal gas law to determine the molar volume at the pressure and temperature that were measured during the bike-ride sampling:

$$MolarVolume\mathrm{NO2} = nRT/P$$

where n is 1 mol, R is the gas constant (0.082057366080960), T is temperature in Kelvin and P is pressure in atm. Then the NO_2 in $\mu g/m^3$ is calculated as:

$$\mathrm{NO2}[\mu g/m^3] = \mathrm{NO2}[\mathrm{ppb}] * (MolarWeight\mathrm{NO2}/MolarVolume\mathrm{NO2})$$

The cargo-bike measurements are aggregated in space–time to every minute and in each (25 m) grid cell of the prediction map. The measurements are also aggregated over the four days, to eliminate effects from random vehicles to represent general emission patterns.

2.1.2. Ground monitor stations used for statistical modelling

In the Netherlands (41,543 km^2), the national air quality ground monitoring network consists of 66 ground stations. These ground stations are managed by the National Institute for Public Health and the Environment (RIVM, National Institute for Public Health and the Environment, 2017). We further incorporated the ground monitor network from Germany (357.386 km^2, 376 stations) to better identify NO_2-predictors relationships using machine learning models. The ground monitoring station measurements are from the European Environment Agency (2021) for the Netherlands and the Umweltbundesamt (2021) for Germany. Three stations with inadequate measurements (i.e. missing values at certain hours) are neglected. The measurements are downloaded for the same days as the cargo-bike measurements and from 7:00 am–11:59 am.

Fig. 1. The cargo-bike and instruments that are used to sensor the NO_2 in our study.

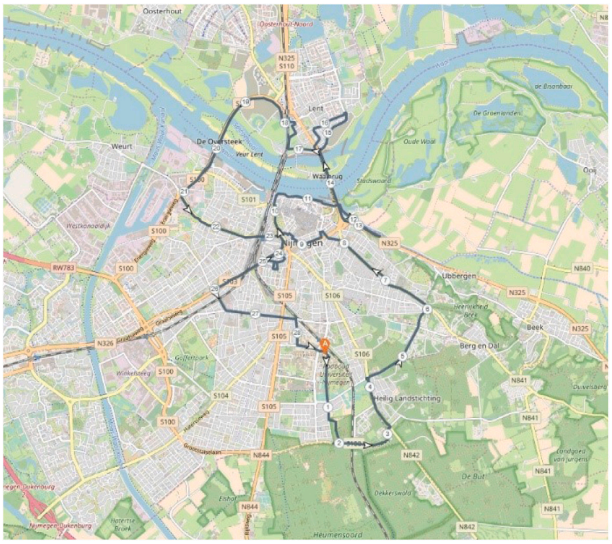

Fig. 2. Routes taken by the cargo-bike, the background map is from OpenStreetMap (OpenStreetMap contributors, 2019). The start and end of routes were at the same location, indicated by the point "A".

The ground station measurements used for prediction are aggregated in two ways, one is the mean of all hours and days measured by cargo-bike, called NEDL-avg dataset, and the other the mean of each hour of the days corresponding to the time of cargo-bike measurements, called NEDL-hr dataset. The NEDL-avg is used for XGB and RF hyperparameter optimisation and to identify the grid resolution for mapping. The NEDL-hr is used for building hourly statistical models.

2.1.3. Geospatial predictors

The geospatial predictors (Table 1) were calculated at 25 m resolution. They are either spatial attributes aggregated within a circular ring centred at each sensor or prediction location, called buffered predictors, or values of the spatial attribute at the observation or prediction location, called gridded variables. The buffered predictors include industry areas, roads, VIIRS (Visible Infrared Imaging Radiometer Suite) Nighttime Day/Night Band (DNB) radiance values (nightlight, NOAA, 2021) and population. Gridded variables include wind speed and temperature (Dee et al., 2011), elevation (Amante and Eakins, 2009), mean of the NO_2 column density from TROPOMI level

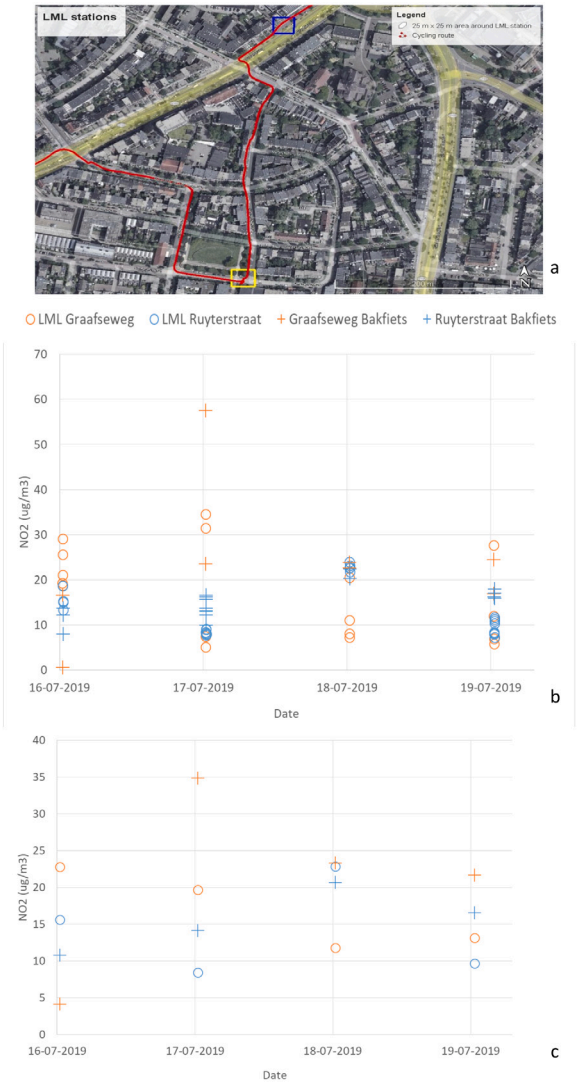

Fig. 3. Comparing the cargo-bike measurements with the Dutch national ground monitoring stations in the region. (a) Locations of the two ground stations, blue block is around the Graafseweg station and yellow the Ruyterstraat station. Bounding boxes (min latitude, max latitude, min longitude, max longitude) for the Graafseweg station are (51.84126, 51.84151, 5.85753, 5.85786) and for the Ruyterstraat station are (51.83810, 51.83830, 5.85701, 5.85719). The red line indicates the cargo-bike route. Within the two blocks the cargo-bike measurements are compared with the ground station measurements at the corresponding times. (b) The cargo-bike and ground station measurements while the cargo-bike was close to the ground stations. (c) Average values of the ground stations and the cargo-bike measurements.

3 product (Google Earth Engine, 2019) in July 2019 (3.5×7 km^2 resolution). The buffered predictors of road and industry are calculated from OpenStreetMap (OpenStreetMap contributors, 2019). For detailed descriptions of the sources of geospatial predictors and how they are calculated please refer to Lu et al. (2020a).

2.2. Research design

The methodological framework is designed as follows:

1. A set of statistical learning methods is applied to DENL-avg to identify hyperparameters for DENL-hr and to identify grid resolution for mapping based on variable importance.
2. The statistical learning methods and optimised hyperparameters are applied to DENL-hr to model hourly NO_2. These statistical learning models are evaluated using CVGS.
3. Hourly NO_2 is predicted for the study area with each of the statistical models using DENL-hr. XGB hyperparameters are adjusted while observing the prediction pattern.
4. The predictions from hourly models using the hyperparameters optimised in step (1) and for XGB additionally the parameters optimised in step (3) are compared to the cargo-bike measurements.

2.3. Random forest, extreme gradient boosting, Lasso

Lasso is a linear regression algorithm that uses an L1 norm to shrink variable coefficients to zero to reduce model variance. RF and XGB are ensemble tree-based methods, which mitigate two negative effects of a single tree model: instability and coarse separation. Large trees are subject to instability, while small trees are inaccurate for their piece-wise constant approximations. Bagging overcomes these two constraints by using small trees to add stability and avoid coarse approximation by averaging over small trees (Friedman, 2001). RF is based on Bagging, which grows trees independently while XGB is based on gradient boosting, which grows trees subsequently based on current model residuals. RF extends from bagging by choosing the variable to split from a subset a of variables. XGB is a scalable gradient boosting algorithm, which enables multiple penalisation paths to control model complexity to prevent model over-fitting, including regularisation on tree width and terminal node values, as well as dropping trees.

The variable importance is calculated for XGB and RF. We use the averaged ranking in 20 times bootstrapping for each method (Lu et al., 2020a). For the XGB the gain scores (Chen and Guestrin, 2016) are used and for the RF the permutation test is used to calculate the variable importance.

2.4. RF and XGB hyperparameter optimisation

The NLDE-avg dataset is used for hyperparameter optimisation, through grid search, with 5-fold CVGS. For XGB, the learning rate (eta), number of iterations (rounds), maximum tree depth (max.tree.depth) and gamma are tuned, each time 70% of data is drawn from the training set. The search grid for the number of iterations (rounds) was from 200 to 3000, with a step of 200; maximum tree depth (max-depth) from 3 to 6 with a step of 1, learning rate (eta) from 0.001 to 0.1 with a step of 0.05, the penalty term gamma (Chen et al., 2019a) from 1 to 5 with a step of 1. The 5-fold CVGS result indicates the optimal hyperparameters to be eta = 0.051, rounds = 200, max-depth = 3, gamma = 1. For RF, the minimum number of trees on the end nodes (min.node.size), and number of variables that are randomly drawn for each tree (mtry) are optimised. The optimal setting for RF is min.node.size equals 5 and mtry equals 12, the number of trees is set to 1000 for random forest, which is a safe choice as the high number of trees will not negatively affect model performance.

2.5. High setting of XGB hyperparameters

As the spatial pattern of XGB can vary greatly with different hyperparameter settings despite the CVGS accuracy remaining the same (Lu et al., 2020a), we observed spatial prediction patterns from multiple XGB hyperparameter settings. We tested increased learning rates of 0.002, 0.001, and 0.0005, and estimators (i.e. the number of trees) of 300, 3000, and 5000. We found altering the learning rate from 0.002 to 0.0005 only affect the prediction patterns subtly but setting the learning rate to 0.002 and 0.051 (original setting optimised by CVGS) makes a considerable difference in their predicted NO_2 patterns. The CVGS accuracy is optimised at 3000 trees with this new learning rate, which remains approximately the same compared to the original setting (Table 4). We also increased the L1 norm (lambda) from 2 to 10, with a step of 2, and gamma (Chen and Guestrin, 2016) to 5, to further control respectively extreme values at the terminal node and model complexity.

Table 1
Predictors used in this study. "_mon" indicates months, (mon = 1, …,12). "_buf" indicates the buffer radius for road density and industry areas. The buffered predictors with buffer radii of 25 m, 50 m, 100 m, 300 m, 500 m, 800 m, 1000 m, 3000 m, 5000 m are calculated. "_bufnl" indicates the buffer radius for the nightlight. The buffer radii of 450 m, 900 m, 3150 m, 4950 m, are calculated. The original resolution is provided for gridded (raster) variables and data types for vector variables.

Predictor	Variable name	Unit	Resolution/data type
Monthly wind speed at 10 m altitude.	Wind_speed_10m_mon	km/h	10 km
Monthly temperature at 2 m altitude.	temperature_2m_mon	Celsius	10 km
TROPOMI July 2019 mean vertical column density.	TROP0719; Tropomi	mol/cm^2	0.01 arc degrees
Population in 5 km grid	population_5000	Count	5 km
Population in 3 km grid	population_3000	Count	3 km
Population in 1 km grid	population_1000	Count	1 km
Nightlight	nightlight_bufnl	W cm^{-2} sr^{-1}	500 m
Total length of highway	road_1_buf	m	Polygon, lineString
Total length of primary roads	road_2_buf	m	Polygon, lineString
Total length of local roads	road_M345_buf	m	Polygon, lineString
Area of industry	I_1_buf	m^2	Polygon, lineString

With these settings the spatial prediction patterns of XGB, as well as their correlations with cargo-bike measurements change subtly. We will show the result of XGB with the maximum tree depth set to 5, learning rate 0.002, number of estimators 3000, lambda 2, and gamma 5. We refer to this hyperparameter setting of XGB as "high setting" (XGB hs); compared to the original setting, it searches the gradient much more slowly, correspondingly with more iterations, and uses additional penalties to control model over-fitting.

2.6. Accuracy assessment

The RMSE (Root Mean Squared Error, $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{predicted} - y_{observed})^2}$) provides a general insight into the variance and magnitude of the error. In addition, we calculated the MAE (Mean Absolute Error, $MAE = mean(|y_{predicted} - y_{observed}|)$) for the magnitude of the error and the IQR (Inter-Quartile Range, $IQR = Q_3 - Q_1$) for the variance of the error. To make the accuracy assessed at different times comparable, we calculated the R^2 (R-squared, $R^2 = 1 - mean((y_{predicted} - y_{observed})^2)/var(y_{observed})$), where var(.) indicates the variance function. We used 80% of the dataset for modelling and 20% for validation. A 20-time bootstrapped cross-validation was used. Thus, the accuracy metrics described above were calculated on validation datasets 20 times, and the median of these 20 results was used as the final accuracy measure.

Table 2
Ranking of the top 15 most important variables of XGBoost and Random Forest, using NLDE-avg. Please refer to Table 1 for the variable description.

Rank	XGBoost	Random Forest
1	population_3000	population_3000
2	road_class_3_3000	road_class_3_3000
3	road_class_2_25	road_class_2_25
4	radiation	population_5000
5	population_1000	road_class_2_50
6	population_5000	road_class_3_5000
7	road_class_2_100	road_class_2_100
8	road_class_3_5000	population_1000
9	road_class_2_50	nightlight_3150
10	nightlight_450	elevation
11	elevation	wind_speed_10m_9
12	road_class_1_5000	nightlight_450
13	TROP0719	TROP0719
14	road_class_3_100	nightlight_4950
15	temperature_2m_2	road_class_3_100

Table 3
20 times bootstrapped cross-validation results of the XGB, RF, and Lasso using NLDE-avg.

	RMSE	IQR	MAE	R-squared
XGB	7.8	6.9	5.4	0.65
RF	7.8	6.9	5.4	0.65
Lasso	8.3	8.5	6.0	0.61

3. Results

3.1. Models trained on NLDE-avg

Table 2 compares the variable importance obtained by the RF and XGB models trained on the NLDE-avg dataset. The top three ranked variables are the same, which consists of the emission-related variables primary road length within 25 m buffers. This indicates the maps can capture spatial variability at 25 m resolution. Therefore, the statistical modelling is based on 25 m resolution grids. Other emission-related variables include primary roads in 50 m and 100 m buffers and local roads in 100 m buffers. The variables are ranked similarly between RF and XGB. Over the study area, the Pearson's correlation coefficients between XGB and RF predictions is 0.97, between RF and Lasso 0.94, XGB and Lasso 0.90.

3.2. Hourly models

The results of XGB, XGB hs, RF, and Lasso models built respectively for 8–9, 9–10, and 10–11 am (referred to as 8 am, 9 am, and 10 am, respectively), averaged over 4 days, using the NLDE-hr dataset, and a comparison of variability with the cargo-bike measurements at corresponding hours are shown in Tables 4 and 6 and Figs. 5 and 7. In all the time slots, the XGB (i.e. with the hyperparameters tuned based on grid search using k-fold CV), XGB hs and RF obtained similar CVGS accuracy, and both outperformed Lasso. Fig. 4 shows the impact of learning rate and compared XGB and XGB hs prediction patterns. It could be observed that a too high learning rate leads to sporadic spatial patterns and the spatial prediction from XGB is noisier compared to XGB hs. The R^2 of the linear regression between model predictions and the cargo-bike measurements (Table 6) also indicated the XGB hs having a higher correlation with the cargo-bike measurements compared to the XGB at 8 am and other times similar. Therefore, for the rest of the comparisons between models and with the cargo-bike measurements, we used XGB hs instead of XGB.

The spatial predictions of the three models at the three time slots and the corresponding cargo-bike measurements are shown in Fig. 5. All the models predicted highest NO_2 along the primary road and show a decreasing trend away from the city centre to the suburban areas. The Lasso prediction shows the least spatial variation and the XGB the most. At 9 am, the highest NO_2 is measured near the river Waal (Fig. 2), this is captured by XGB and RF predictions but is completely missed by the Lasso prediction. At 10 am, the cargo-bike measurements are higher along the roads, and this is consistent across the model predictions.

The XGB hs, RF, and Lasso predictions along the cargo-bike track are compared to the cargo-bike measurements (Fig. 6). The mean (Table 5) at 8 am between model predictions and the cargo-bike measurements are the closest. The mean of model predictions at 9 and 10 am are higher than the cargo-bike measurements (less than 25%

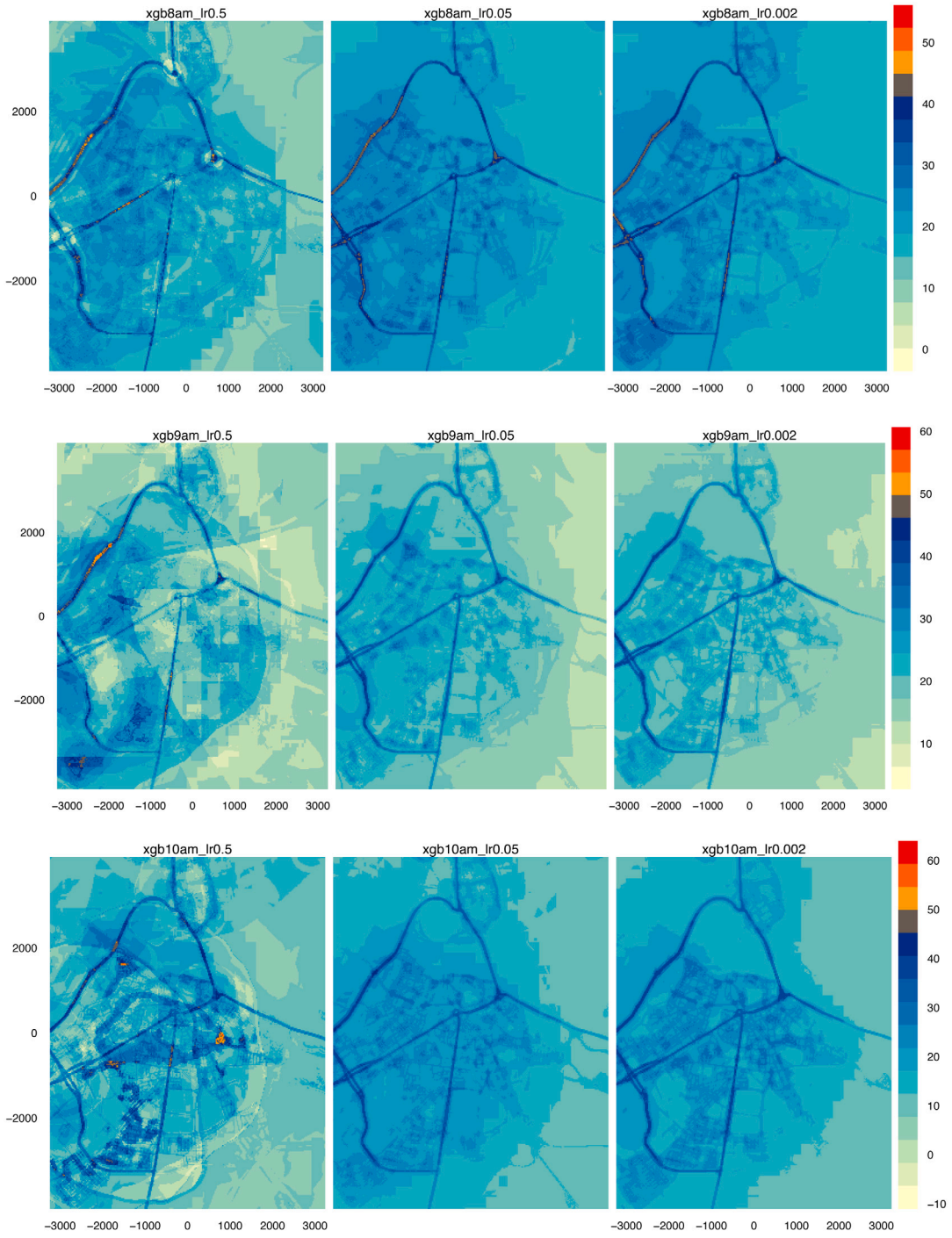

Fig. 4. Spatial predictions (μg/m^3) of XGB using different hyperparameter settings. From left to right: original hyperparameter setting with learning rate 0.5, original hyperparameter setting (learning rate 0.051), high-setting hyperparameters (learning rate 0.002).

higher). To facilitate visualising the spatial variations of the statistical model predictions compared to cargo-bike measurements, we regressed three model predictions along the cargo-bike track each against the cargo-bike measurements (Fig. 7). The model predictions along the cargo-bike track vary more similarly between each other compared to the variations in the cargo-bike measurements. The best match between the cargo-bike measurements and the three model predictions occurs at 10 am, all the models are capable of predicting the peaks between points 7500–9000 at 10 am, when the cargo-bike left the main road. This may be the reason that the R^2 (Table 6) between model predictions and cargo-bike measurements are the highest at 10 am. Compared with RF and XGB, the Lasso model prediction obtained the highest correlation with cargo-bike measurements at 10 am despite the lowest CVGS R^2. At 8 am, the cargo-bike is mostly in an area with less traffic. The R^2 of the XGB hs against cargo-bike measurements (0.27) is notably higher compared to that of Lasso (0.00) and RF predictions (0.1). At 9 am, the cargo-bike is mostly on the main road. The R^2 are similarly low in all model comparisons.

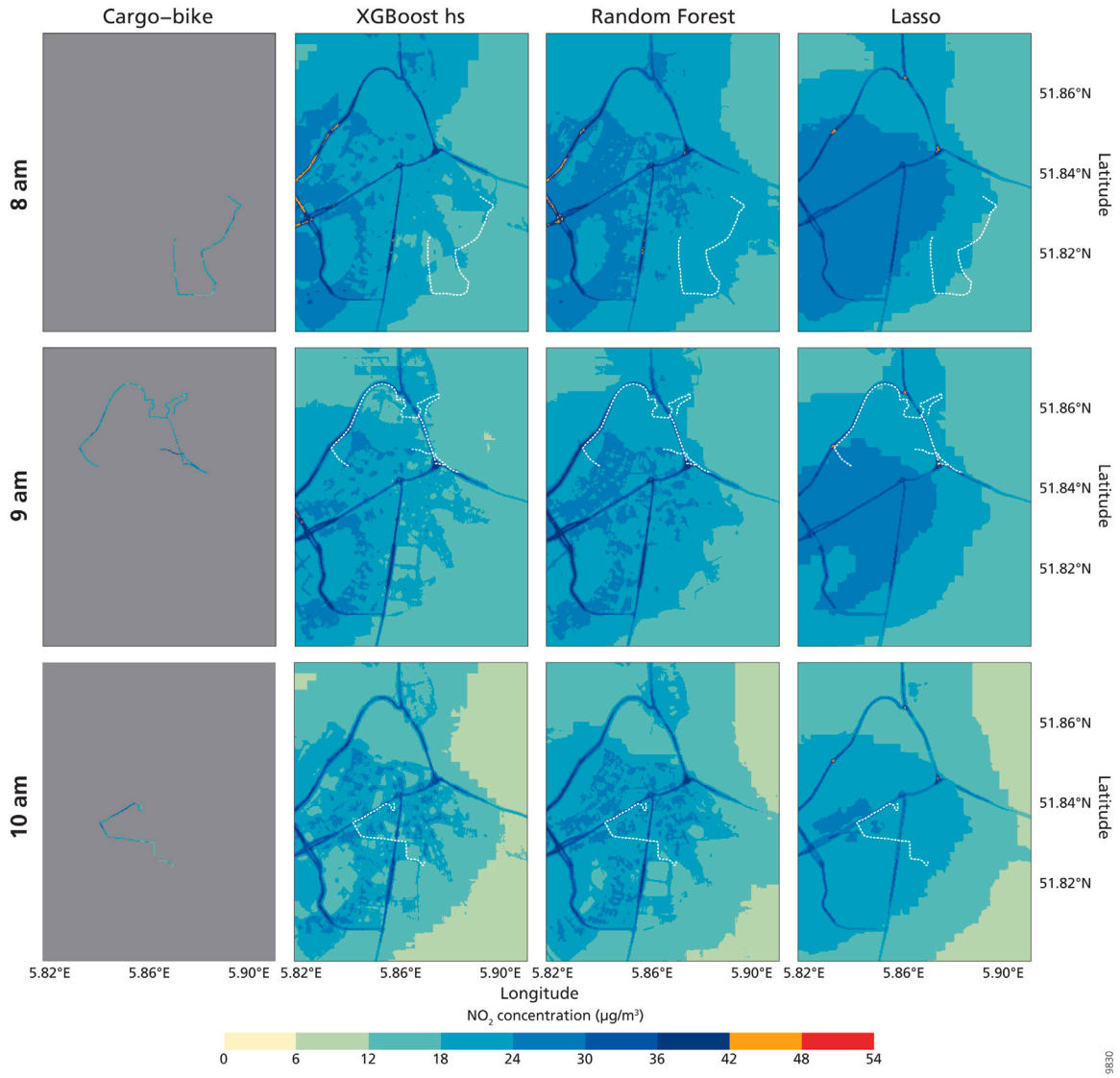

Fig. 5. Maps of cargo-bike measurements and predictions from XGB hs, RF, Lasso, hourly models. The dashed lines in the prediction maps indicate the cargo-bike routes.

Table 4
20 times bootstrapped cross-validation results of the XGBoost (original hyperparameter setting), XGBoost-hs (high-setting XGBoost), Random Forest, and Lasso using NLDE-hr. The units of RMSE, IQR, and MAE are µg/m^3.

Method	Time	RMSE	IQR	MAE	R-squared
XGBoost	8 am	7.8	8.1	5.6	0.70
	9 am	8.7	7.5	6.0	0.60
	10 am	8.7	7.3	5.8	0.57
XGBoost-hs	8 am	7.9	7.0	5.6	0.67
	9 am	8.6	7.1	5.7	0.63
	10 am	8.7	6.8	5.9	0.57
Random Forest	8 am	8.0	7.2	5.7	0.66
	9 am	8.4	7.1	5.8	0.63
	10 am	8.7	7.7	6.0	0.60
Lasso	8 am	9.0	10.0	6.6	0.58
	9 am	9.7	10.1	7.0	0.53
	10 am	9.6	9.1	6.6	0.51

Table 5
Mean NO_2 of cargo-bike, RF, Lasso, high-setting XGB (XGB hs) and cargo-bike measurements.

Time	Cargo-bike	RF	Lasso	XGB hs
8 am	21.25	21.9	22.0	20.0
9 am	22.5	28.6	25.6	27.2
10 am	20.0	25.3	24.4	23.9

Table 6
R^2 between each RF, Lasso, high setting of XGB (XGB hs) and cargo-bike measurements.

Time	XGB	RF	Lasso	XGB hs
8 am	0.1	0.09	0.00	0.27
9 am	0.1	0.16	0.09	0.09
10 am	0.51	0.48	0.55	0.48

4. Discussion

In this study, we used a cargo-bike with various air quality sensors and high-end instrumentation for gathering spatially more detailed ground validation data. The high-end monitors we used are equivalent to those used for official air quality monitoring. With high-end monitors, i.e. apparatus, we can measure as accurate as reasonably possible in (most) circumstances, not or minimally affected by temperature, humidity and other cross-reactivity (e.g., NO_2 sensors can be cross reactive to NO and O_3). As these monitors are usually quite large (19″ rack form factor or comparable), until now this equipment (as far as we

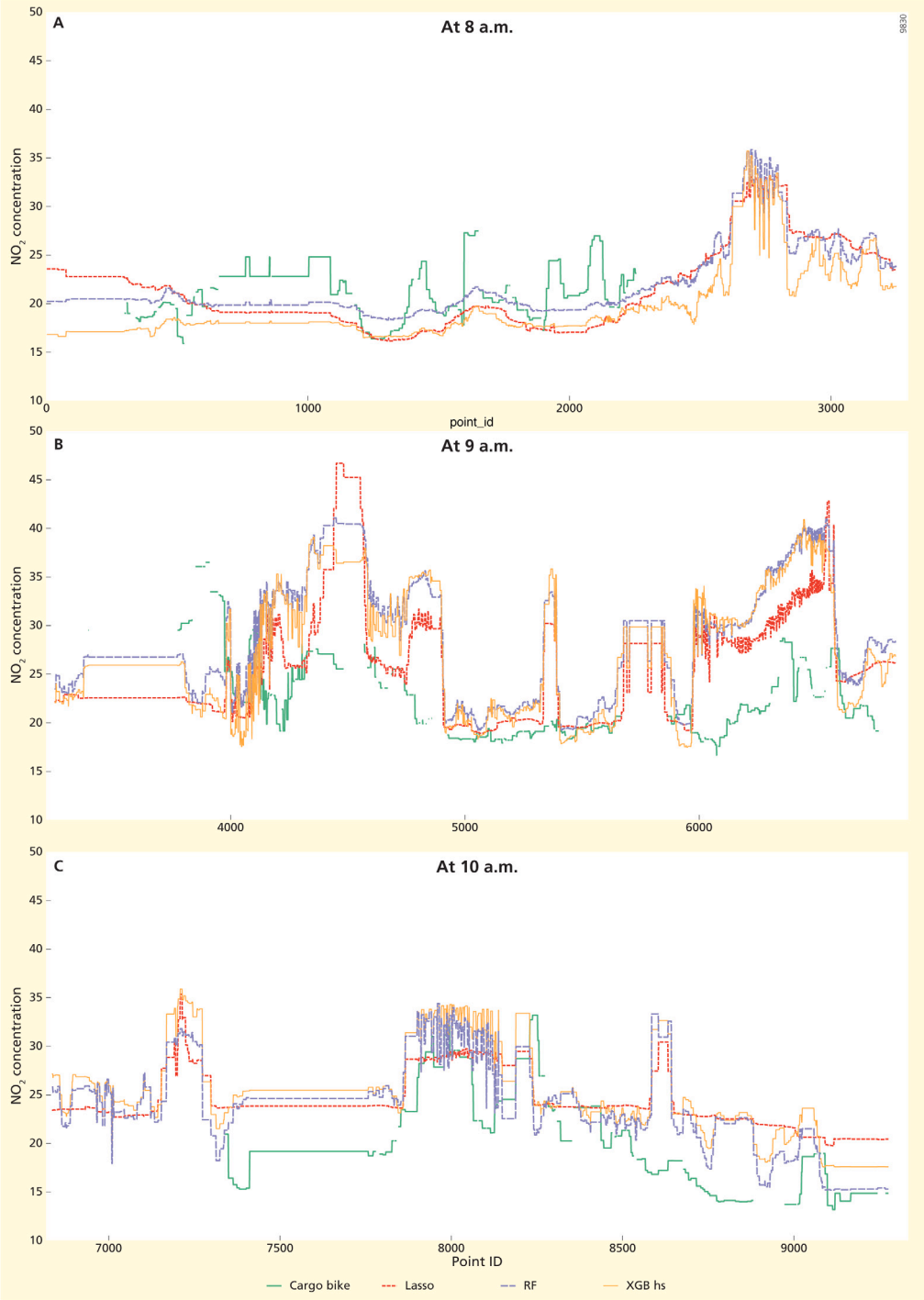

Fig. 6. Cargo-bike measurements and the XGB (high setting), RF, Lasso predictions and the cargo-bike measurements, visualised in the 1-D view. The point_id on the *x*-axis refers to the position on the track followed by the cargo bike.

know) has only been used stationary or on a car or van. With the cargo-bike we could reach considerably more areas in a city, still using this high-end equipment. High-end equipment onboard a cargo-bike gives the possibility to measure right at the place of interest such as densely populated areas (e.g., city centre) or where events occur.

With our unique cargo-bike dataset, this study provides a strong indication that the hyperparameter optimisation and model evaluation results based solely on CVGS may be misleading. In this study, the XGB and RF models with hyperparameters optimised based on k-fold CVGS obtained similar modelling CVGS accuracy. However, a comparison of spatial patterns in NO_2 predictions and cargo-bike NO_2 measurements as well as the validation of predictions against cargo-bike data indicate the RF model is more favourable. Using a lower learning rate for the XGB gives similar CVGS results compared to RF, but very different and seemingly more detailed spatial prediction patterns. The cargo-bike measurements provide a quantitative measure to understand which model gives the most realistic spatial predictions, and in our study this was the high-setting XGB model. The conventional CVGS-based model evaluation and hyperparameter optimisation lack this information and may lead to wrong conclusions regarding model comparison. We demonstrated that it is important to consider spatial prediction patterns when evaluating model predictions of NO_2. In addition, we

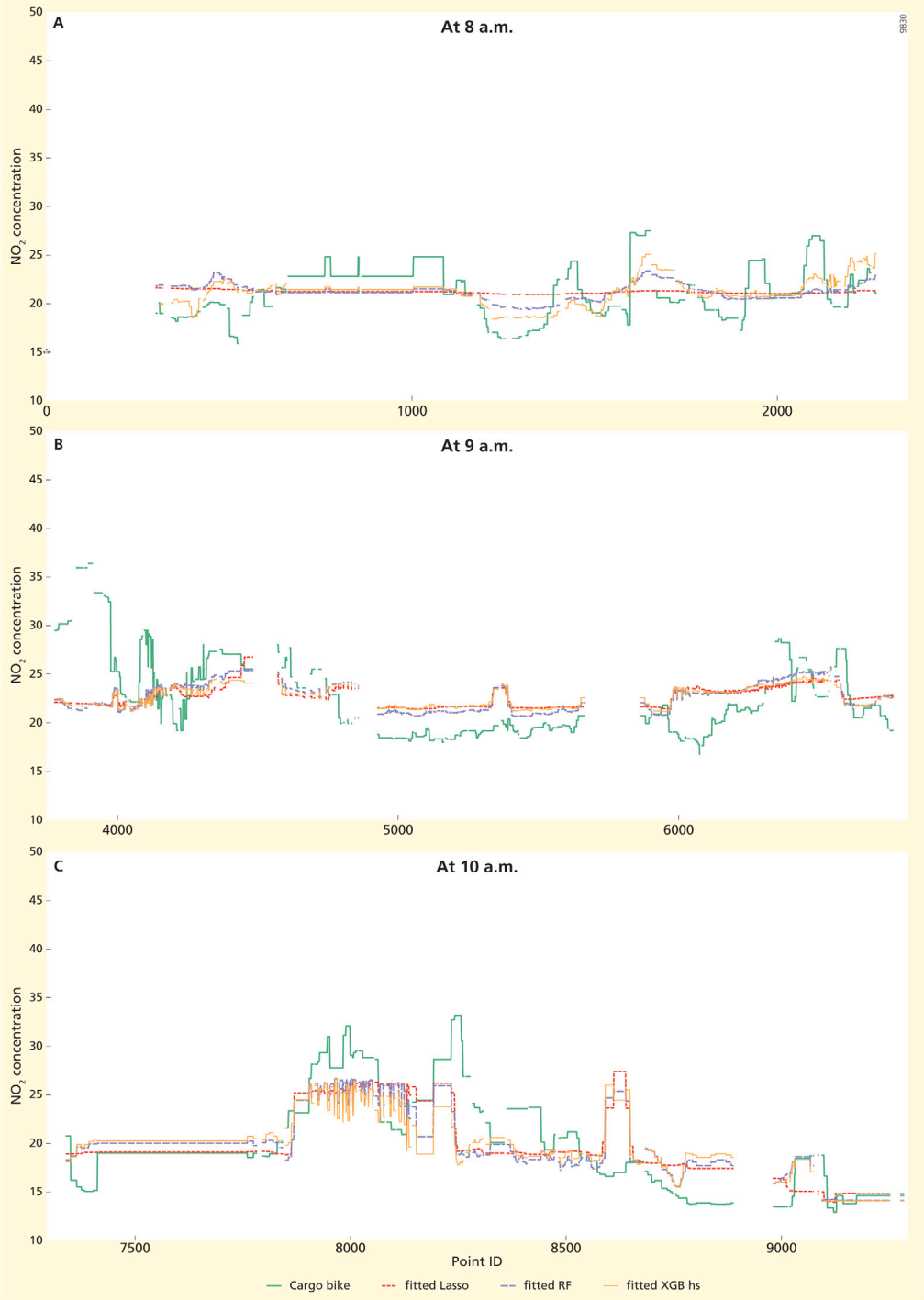

Fig. 7. Cargo-bike measurements and the fitted values of a linear regression between respectively the XGB (high setting), RF, Lasso predictions and the cargo-bike measurements, visualised in the 1-D view. The point_id on the *x*-axis refers to the position on the track followed by the cargo bike.

have shown that spatially-continuous measurements are a valuable source of information to improve the hyperparameter tuning and model evaluation. Importantly, external measurements provide us with an objective, quantitative measure to involve the spatial prediction pattern in the hyperparameter tuning and modelling processes. The route track taken at 8 am by the cargo-bike is mostly far away from major traffic roads (further than 500 m away from "primary roads" defined by OpenStreetMaps), and the route track taken at 9 am and 10 am are mostly in a traffic area. Among the prediction models, XGB hs obtained the highest R^2 with the cargo-bike measurements at 8 am, which shows the least variations compared to other times (Fig. 6). This may indicate that the XGB is less prone to over-fitting when its hyperparameters are properly set. The best match between all the models and the cargo-bike measurements is at 10 am. This indicates that the statistical models are capable of capturing the traffic emission related variations. The inconsistency between the R^2 of CVGS (Table 3, i.e. the lowest for Lasso and for 10 am) and when comparing with the cargo-bike measurements (Table 6 the highest for Lasso and 10 am) may be explained by that Lasso captured an "on and far-away from the main road" pattern with ground station measurements and this pattern is presented in this part of the trip. This pattern is not everywhere as evidenced by the low CVGS obtained by the Lasso. Future studies need to include traffic

volumes to better understand the amount of local variations that could be captured with statistical models as well as using the designed mobile sensing technique to sample in more areas to better understand the spatial heterogeneity in model predictions.

The hyperparameter that affects the prediction pattern of XGB the most is the learning rate. If the learning rate is too high, the model may miss the minimum of the objective function and not fully learn the patterns in the data, causing sporadic effects. When the learning rate is optimal, the sporadic effects diminish and will only change minimally when the learning rate is further reduced. In this study, further reducing the learning rate from the optimal learning rate identified by cross-validation led to a clearer pattern (Fig. 4). Moreover, other model over-fitting control strategies, such as increasing gamma, lambda, and reducing sub-sampling do not considerably alter the prediction patterns and the CVGS accuracy, which may indicate that the model is not subject to over-fitting.

An important difference between the validation on cargo-bike and CVGS is that the CVGS measurements include all scales of spatial variations (short range, medium range, large range), while the cargo-bike measurements include only local variations. This means testing with CVGS evaluates the model also on its capability to predict patterns of different scales, while testing on the cargo-bike measurements mainly evaluates the model on its capability to predict detailed local variations. A model trained on a dataset that includes all scales of variation (based on CVGS) is likely not the model that performs best when mapping the detailed local variations. To the other side, satellite imagery (e.g. Tropomi instrument measurements) can be used to evaluate how the model is scalable to regions where dense ground monitoring networks are not available.

5. Conclusion

In this study, we designed a novel high-end instrument to measure detailed NO_2 not only along the primary roads but also in the city centre and used the measurements to further evaluate and compare high-resolution statistical NO_2 models. The spatially dense measurements from the cargo-bike allow us to compare different statistical methods and hyperparameter settings accounting for their spatial patterns. This is complementary to the CVGS-based accuracy assessment. As the ground truths are scattered, ignoring the spatial prediction patterns in hyperparameter optimisation and model evaluation may lead to one-sided model evaluation and comparison. The accuracy of model predictions is spatially heterogeneous. We showed that while the CVGS accuracy stays the same, the XGB model predictions vary non-trivially with different hyperparameter settings, particularly the learning rate. With advanced mobile sensing techniques, this study provides an approach for a more in-depth look into NO_2 statistical modelling and model comparison and highlights the possible pitfall of exclusively depending on CVGS accuracy in model hyperparameter optimisation and comparison.

CRediT authorship contribution statement

Meng Lu: Conceptualization, Methodology, Data curation, Visualisation, Investigation, Writing. **Ruoying Dai:** Data curation. **Cjestmir de Boer:** Data curation, Writing – review & editing. **Oliver Schmitz:** Data curation. **Ingeborg Kooter:** Data curation. **Simona Cristescu:** Funding acquisition, Writing – review & editing. **Derek Karssenberg:** Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is funded by the Global Geo Health Data Centre (Utrecht University) and the Startimpulsprogramma Meten en Detecteren van Gezond Gedrag (Dutch Science Foundation). We are thankful to Evert Duistermaat, who cycled the cargo-bike and Sieger Henke, who contributed with practical, coordination and some scientific works. The authors appreciate Ton Markus for his advises and contributions on improving the figures. The authors are grateful to the editors and reviewers for their contributions.

References

Akita, Yasuyuki, Baldasano, Jose M., Beelen, Rob, Cirach, Marta, De Hoogh, Kees, Hoek, Gerard, Nieuwenhuijsen, Mark, Serre, Marc L., De Nazelle, Audrey, 2014. Large scale air pollution estimation method combining land use regression and chemical transport modeling in a geostatistical framework. Environ. Sci. Technol. 48 (8), 4452–4459.

Amante, Christopher, Eakins, Barry W., 2009. ETOPO1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. NOAA Technical Memorandum NESDIS NGDC-24.

Anderson, Jonathan O., Thundiyil, Josef G., Stolbach, Andrew, 2012. Clearing the air: a review of the effects of particulate matter air pollution on human health. J. Med. Toxicol. 8 (2), 166–175.

Apte, Joshua S., Messier, Kyle P., Gani, Shahzad, Brauer, Michael, Kirchstetter, Thomas W., Lunden, Melissa M., Marshall, Julian D., Portier, Christopher J., Vermeulen, Roel C.H., Hamburg, Steven P., 2017. High-resolution air pollution mapping with Google street view cars: exploiting big data. Environ. Sci. Technol. 51 (12), 6999–7008.

Atlasleefomegeving, 2020. Nitrogen dioxide. https://www.atlasleefomgeving.nl/meer-weten/lucht/stikstofdioxide.

Beelen, Rob, Voogt, Marita, Duyzer, Jan, Zandveld, Peter, Hoek, Gerard, 2010. Comparison of the performances of land use regression modelling and dispersion modelling in estimating small-scale variations in long-term air pollution concentrations in a Dutch urban area. Atmos. Environ. 44 (36), 4614–4621.

Breiman, Leo, 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Briggs, David J., de Hoogh, Cornelis, Gulliver, John, Wills, John, Elliott, Paul, Kingham, Simon, Smallbone, Kirsty, 2000. A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. Sci. Total Environ. 253 (1–3), 151–167.

Chan, Ka Lok, Khorsandi, Ehsan, Liu, Song, Baier, Frank, Valks, Pieter, 2021. Estimation of surface NO2 concentrations over Germany from TROPOMI satellite observations using a machine learning method. Remote Sens. 13 (5), 969.

Chen, Tianqi, Guestrin, Carlos, 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794.

Chen, Tianqi, He, Tong, Benesty, Michael, Khotilovich, Vadim, Tang, Yuan, Cho, Hyunsu, Chen, Kailong, Mitchell, Rory, Cano, Ignacio, Zhou, Tianyi, Li, Mu, Xie, Junyuan, Lin, Min, Geng, Yifeng, Li, Yutian, 2019a. Xgboost: Extreme gradient boosting. URL https://CRAN.R-project.org/package=xgboost. R package version 0.82.1.

Chen, Jie, de Hoogh, Kees, Gulliver, John, Hoffmann, Barbara, Hertel, Ole, Ketzel, Matthias, Bauwelinck, Mariska, van Donkelaar, Aaron, Hvidtfeldt, Ulla A., Katsouyanni, Klea, et al., 2019b. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ. Int. 130, 104934.

Dee, Dick P., Uppala, S.M., Simmons, A.J., Berrisford, Paul, Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, d.P., et al., 2011. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137 (656), 553–597.

Dijkema, Marieke B., Gehring, Ulrike, van Strien, Rob T., Van Der Zee, Saskia C., Fischer, Paul, Hoek, Gerard, Brunekreef, Bert, 2010. A comparison of different approaches to estimate small-scale spatial variation in outdoor NO_2 concentrations. Environ. Health Perspect. 119 (5), 670–675.

dos Santos, Rochelle Schneider, Vicedo-Cabrera, Ana, Sera, Francesco, Stafoggia, Massimo, de Hoogh, Kees, Kloog, Itai, Reis, Stefan, Vieno, Massimo, Gasparrini, Antonio, 2020. A satellite-based spatio-temporal machine learning model to reconstruct daily PM2. 5 concentrations across Great Britain. MedRxiv.

Dou, Xinyu, Liao, Cuijuan, Wang, Hengqi, Huang, Ying, Tu, Ying, Huang, Xiaomeng, Peng, Yiran, Zhu, Biqing, Tan, Jianguang, Deng, Zhu, Wu, Nana, Sun, Taochun, Ke, Piyu, Liu, Zhu, 2021. Estimates of daily ground-level NO2 concentrations in China based on Random Forest model integrated K-means. Adv. Appl. Energy (ISSN: 2666-7924) 2, 100017. http://dx.doi.org/10.1016/j.adapen.2021.100017, URL https://www.sciencedirect.com/science/article/pii/S266679242100010X.

European Environment Agency, 2021. Download service for E1a and E2a data. "https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm". Last Accessed: 20.02.2020.

Friedman, Jerome H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Statist. 1189–1232.

Google Earth Engine, 2019. Sentinel-5P NRTI NO2: Near real-time nitrogen dioxide. URL "https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_NRTI_L3_NO2#description". last assessed Sep 13, 2021.

Gressent, Alicia, Malherbe, Laure, Colette, Augustin, Rollin, Hugo, Scimia, Romain, 2020. Data fusion for air quality mapping using low-cost sensor observations: Feasibility and added-value. Environ. Int. (ISSN: 0160-4120) 143, 105965. http://dx.doi.org/10.1016/j.envint.2020.105965, URL https://www.sciencedirect.com/science/article/pii/S0160412020319206.

Hasenfratz, David, Saukh, Olga, Walser, Christoph, Hueglin, Christoph, Fierz, Martin, Arn, Tabita, Beutel, Jan, Thiele, Lothar, 2015. Deriving high-resolution urban air pollution maps using mobile sensor nodes. Pervasive Mob. Comput. 16, 268–285.

Health Effects Institute, 2010. Traffic-Related Air Pollution: A Critical Review of the Literature on Emissions, Exposure, and Health Effects. Number 17. Health Effects Institute.

Hoek, Gerard, Beelen, Rob, De Hoogh, Kees, Vienneau, Danielle, Gulliver, John, Fischer, Paul, Briggs, David, 2008b. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 42 (33), 7561–7578.

Hoek, Gerard, Beelen, Rob, de Hoogh, Kees, Vienneau, Danielle, Gulliver, John, Fischer, Paul, Briggs, David, 2008a. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 7561–7578.

Holmes, Nicholas S., Morawska, Lidia, 2006. A review of dispersion modelling and its application to the dispersion of particles: an overview of different dispersion models available. Atmos. Environ. 40 (30), 5902–5928.

Isiugo, Kelechi, Newman, Nicholas, Jandarov, Roman, Grinshpun, Sergey A., Reponen, Tiina, 2018. Assessing the accuracy of commercially available gas sensors for the measurement of ambient ozone and nitrogen dioxide. J. Occup. Environ. Hyg. 15 (11), 782–791.

Kerckhoffs, Jules, Hoek, Gerard, Portengen, Lützen, Brunekreef, Bert, Vermeulen, Roel C.H., 2019. Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. Environ. Sci. Technol. 53 (3), 1413–1421.

Kharol, S.K., Martin, R.V., Philip, S., Boys, B., Lamsal, L.N., Jerrett, M., Brauer, M., Crouse, D.L., McLinden, C., Burnett, R.T., 2015. Assessment of the magnitude and recent trends in satellite-derived ground-level nitrogen dioxide over North America. Atmos. Environ. 118, 236–245.

Khreis, Haneen, de Hoogh, Kees, Zietsman, Josias, Nieuwenhuijsen, Mark J., 2018. The impact of different validation datasets on air quality modeling performance. Transp. Res. Rec. 2672 (25), 57–66.

Larkin, Andrew, Geddes, Jeffrey A., Martin, Randall V., Xiao, Qingyang, Liu, Yang, Marshall, Julian D., Brauer, Michael, Hystad, Perry, 2017. Global land use regression model for nitrogen dioxide air pollution. Environ. Sci. Technol. 51 (12), 6957–6964.

Li, Tongwen, Wang, Yuan, Yuan, Qiangqiang, 2020. Remote sensing estimation of regional NO2 via space-time neural networks. Remote Sens. (ISSN: 2072-4292) 12 (16), URL https://www.mdpi.com/2072-4292/12/16/2514.

Liu, Fei, van der A, Ronald J., Eskes, Henk, Ding, Jieying, Mijling, Bas, 2018. Evaluation of modeling $NO2$ concentrations driven by satellite-derived and bottom-up emission inventories using in situ measurements over China. Atmos. Chem. Phys. 18 (6), 4171–4186.

Lu, Meng, Schmitz, Oliver, de Hoogh, Kees, Kai, Qin, Karssenberg, Derek, 2020a. Evaluation of different methods and data sources to optimise modelling of NO2 at a global scale. Environ. Int. (ISSN: 1873-6750) 142, 105856. http://dx.doi.org/10.1016/j.envint.2020.105856.

Lu, Meng, Schmitz, Oliver, Vaartjes, Ilonca, Karssenberg, Derek, 2019. Activity-based air pollution exposure assessment: differences between homemakers and cycling commuters. Health & place 60, 102233.

Lu, Meng, Soenario, Ivan, Helbich, Marco, Schmitz, Oliver, Hoek, Gerard, van der Molen, Michiel, Karssenberg, Derek, 2020b. Land use regression models revealing spatiotemporal co-variation in NO2, NO, and O3 in the Netherlands. Atmos. Environ. 223, 117238.

Lu, T., Wan, Y., Bechle, M., Presto, A., Hankey, S., et al., 2019b. External validation of national land use regression models for PM2. 5 using a low-cost sensor network. Environ. Epidemiol. 3, 251.

Marjovi, Ali, Arfire, Adrian, Martinoli, Alcherio, 2015. High resolution air pollution maps in urban environments using mobile sensor networks. In: 2015 International Conference on Distributed Computing in Sensor Systems. IEEE, pp. 11–20.

Marshall, Julian D., Nethery, Elizabeth, Brauer, Michael, 2008. Within-urban variability in ambient air pollution: comparison of estimation methods. Atmos. Environ. 42 (6), 1359–1369.

Mijling, Bas, Jiang, Qijun, De Jonge, Dave, Bocconi, Stefano, 2018. Field calibration of electrochemical NO 2 sensors in a citizen science context. Atmos. Meas. Tech. 11 (3), 1297–1312.

Miskell, Georgia, Salmond, Jennifer A., Williams, David E., 2018. Use of a handheld low-cost sensor to explore the effect of urban design features on local-scale spatial and temporal air quality variability. Sci. Total Environ. 619, 480–490.

Mölter, A., Lindley, S., de Vocht, F., Simpson, A., Agius, R., 2010a. Modelling air pollution for epidemiologic research — Part I: A novel approach combining land use regression and air dispersion. Sci. Total Environ. (ISSN: 0048-9697) 408 (23), 5862–5869. http://dx.doi.org/10.1016/j.scitotenv.2010.08.027, URL https://www.sciencedirect.com/science/article/pii/S0048969710008880. Special Section: Integrating Water and Agricultural Management Under Climate Change.

Mölter, A., Lindley, S., de Vocht, F., Simpson, A., Agius, R., 2010b. Modelling air pollution for epidemiologic research—Part I: A novel approach combining land use regression and air dispersion. Sci. Total Environ. 408 (23), 5862–5869.

Nagendra, S.M. Shiva, Yasa, Pavan Reddy, Narayana, M.V., Khadirnaikar, Seema, Rani, Pooja, 2019. Mobile monitoring of air pollution using low cost sensors to visualize spatio-temporal variation of pollutants at urban hotspots. Sustainable Cities Soc. 44, 520–535.

National Institute for Public Health and the Environment, 2017. Air quality monitoring network. https://www.luchtmeetnet.nl/. Accessed July 1, 2017.

NOAA, 2021. DMSP and VIIRS data download. "https://ngdc.noaa.gov/eog/download.html". Last Accessed: 11.03.2021.

OpenStreetMap contributors, 2019. Planet dump 7 Jan 2019 retrieved from https://planet.osm.org.

World Health Organization et al, 2018. Burden of Disease from the Joint Effects of Household and Ambient Air Pollution for 2016. Social and Environmental Determinants of Health Departmen, Geneva.

Park, Yoo Min, Kwan, Mei-Po, 2017. Individual exposure estimates may be erroneous when spatiotemporal variability of air pollution and human mobility are ignored. Health Place 43, 85–94.

Pascal, Laurence, 2009. Effets à court terme de la pollution atmosphérique sur la mortalité. Rev. Fr. Allergol. 49 (6), 466–476.

Rai, Aakash C., Kumar, Prashant, Pilla, Francesco, Skouloudis, Andreas N., Di Sabatino, Silvana, Ratti, Carlo, Yasar, Ansar, Rickerby, David, 2017. End-user perspective of low-cost sensors for outdoor air pollution monitoring. Sci. Total Environ. (ISSN: 0048-9697) 607–608, 691–705. http://dx.doi.org/10.1016/j.scitotenv.2017.06.266, URL https://www.sciencedirect.com/science/article/pii/S0048969717316935.

Ren, Xiang, Mi, Zhongyuan, Georgopoulos, Panos G., 2020. Comparison of machine learning and land use regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States. Environ. Int. (ISSN: 0160-4120) 142, 105827. http://dx.doi.org/10.1016/j.envint.2020.105827, URL https://www.sciencedirect.com/science/article/pii/S0160412020317827.

Rivera, Claudia, Stremme, Wolfgang, Grutter, Michel, 2013. Nitrogen dioxide DOAS measurements from ground and space: comparison of zenith scattered sunlight ground-based measurements and OMI data in central Mexico. Atmósfera 26 (3), 401–414.

Schneider, Philipp, Castell, Nuria, Vogt, Matthias, Dauge, Franck R., Lahoz, William A., Bartonova, Alena, 2017. Mapping urban air quality in near real-time using observations from low-cost sensors and model information. Environ. Int. 106, 234–247.

UN SDGs, 2017. Sustainable Development Knowledge Platform. United Nations. United Nations, Nd Web.

Shams, Seyedeh Reyhaneh, Jahani, Ali, Kalantary, Saba, Moeinaddini, Mazaher, Khorasani, Nematollah, 2021. Artificial intelligence accuracy assessment in NO 2 concentration forecasting of metropolises air. Sci. Rep. 11 (1), 1–9.

Song, Xin-Yi, Gao, Ya, Peng, Yubo, Huang, Sen, Liu, Chao, Peng, Zhong-Ren, 2021. A machine learning approach to modelling the spatial variations in the daily fine particulate matter (PM2. 5) and nitrogen dioxide (NO2) of Shanghai, China. Environ. Plan. B 48 (3), 467–483.

Spinelle, Laurent, Gerboles, Michel, Villani, Maria Gabriella, Aleixandre, Manuel, Bonavitacola, Fausto, 2015. Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide. Sensors Actuators B 215, 249–257.

Tibshirani, Robert, 1994. Regression Shrinkage and selection via the Lasso. J. R. Stat. Soc. B 58, 267–288.

Food UK Department for Environment and Rural Affairs, 2004. Ukno2. https://uk-air.defra.gov.uk/library/assets/documents/reports/aqeg/nd-contents-chapter1.pdf.

Umweltbundesamt, 2021. Important environmental indicators. "https://www.umweltbundesamt.de/en". Last Accessed: 20.02.2021.

van Zoest, Vera, Osei, Frank B., Stein, Alfred, Hoek, Gerard, 2019. Calibration of low-cost NO2 sensors in an urban air quality network. Atmos. Environ. 210, 66–75.