A close look at using national ground stations for the statistical modeling of NO_2

Foeke Boersma and Meng Lu

Department of Geography, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany

Correspondence: Foeke Boersma (foekeboersma@hotmail.com)

Abstract.

Air pollution leads to various health and societal issues. It is therefore essential to model and predict air pollution over space. Many statistical models use geospatial data related to air pollution sources. However, an important aspect often disregarded is the spatial heterogeneity in which characteristics, phenomena, or relationships between geographically distributed variables may change over space. This study aims to evaluate and compare various spatial and non-spatial statistical and machine learning methods, with attention given to different spatial groups. Traffic- and population-related variables define the spatial groups. Models classify as local or global. Local models are based on the Amsterdam area and perform predictions on the same area. Global models are based on observations throughout Germany and The Netherlands while predictions apply to several smaller areas of interest in Germany and The Netherlands. We found that prediction accuracy differs substantially in different spatial groups. Predictions for places near roads with high populations show poor prediction accuracy, while prediction accuracy increases in low population density areas for both local and global models. Prediction accuracy is further increased in places far from roads for global models. This division into spatial groups also shows that global non-linear methods can have higher prediction accuracy than global linear methods. The spatial prediction patterns show that non-linear methods generally are less prone to overfitting than linear methods. Additionally, clusters of predicted air pollution differ between models within cities, and various predictors exhibit the highest influence on NO_2 concentrations across different cities. Lastly, applying the same methods to the local dataset yields poor metrics, especially for the non-linear methods.

1 Introduction

Quantifying air quality forms the base for environmental management and understanding the negative effects of air pollution. Air pollutants have been modelled at different spatial scales, up to globally. The models can be classified into statistical models, chemical transportation models and air dispersion models. The chemical transportation models are developed for large-scale air pollution modeling. The air dispersion models require detailed and spatially revolved emission data to model small-scale spatial variations in air pollutants sufficiently (Beelen et al., 2013). Statistical modeling is becoming more popular in high-resolution mapping at different spatial scales, due to an increment in available predictors (i.e. GIS variables) and computational capability. Land Use Regression (LUR) is the most well-known statistical modeling approach for air pollution. It is based on a linear regression model to capture the spatial variability of traffic-related air pollution in urban areas. Most LUR models are based on measurements from ground monitoring stations (Hoek et al., 2008; Wang et al., 2020). Geostatistical methods such as kriging could further capture the spatial correlations between the observations. However, several studies favor the simple character of the LUR models and conclude that they perform better than or equivalent to geostatistical methods (Hoek et al., 2008; Marshall et al., 2008; Beelen et al., 2013). These counter-intuitive results may be because the geostatistical models are not optimally specified.

Although the linear models have the advantage of being highly interpretable and can be extrapolated, they may not capture the complex air emission, dispersion, and deposition process (Wang et al., 2020). Data-driven, nonparametric models (most commonly known as machine learning methods in air pollution mapping), such as tree-based algorithms have been applied to air pollution mapping in recent years, as these methods may better capture the non-linear relationships between pollutants and predictors (Weichenthal et al., 2016; Reid et al., 2015; Lu et al., 2020). Brokamp et al. (2017) compare the Land Use Random Forest models (LURF) with the LUR models for elemental components of PM 2.5 in the urban area of Cincinnati, Ohio, and find that the LURF shows a lower prediction error variance for each elemental model with cross-validation. Kerckhoffs et al. (2019) report that machine learning algorithms such as bagging and random forest, explain more variability in ultra fine particle concentrations than multiple linear regression and regularized regression techniques. Ameer et al. (2019) advocate that the random forest regression is the best technique, compared to decision tree regression, Multi-Layer Perceptron regression, and gradient boosting regression, for pollution prediction for data sets of varying size, location, and characteristics. Ren et al. (2020) conclude that non-linear machine learning methods achieve higher accuracy than the linear LUR, thereby stressing that a careful design of hyperparameter tuning and flexible data splitting and validations is important to acquire stable and reliable results. Chen et al. (2019) compare 16 algorithms to predict the annual average fine particle (PM2.5) and nitrogen dioxide (NO_2) concentrations across Europe, and also conclude with a favor of the ensemble tree-based methods. However, this difference is more prevalent for the PM2.5 pollutant while NO_2 model predictions show a similar R^2. At the same time, a high correlation is reported between the predicted values of the various models used in the study. Furthermore, since they measure two pollutants, the most influential predictors on the variation of each pollutant differ substantially. To expand further, satellite observations and dispersion model estimates are among the most influential predictors for PM2.5 concentrations, whereas the variation in NO_2 is primarily attributable to traffic-related variables. The major contribution of road traffic to NO_2 concentrations is supported by Wong et al. (2021). Additionally, they mention that the rise of NO_2 might be related to a specific type of diesel particulate produced by heavy vehicles such as buses. In contrast, nitrogen is produced by, particularly long-range transport, gasoline-fueled passenger cars.

Over the past few years, there has been a notable rise in the utilization of statistical modeling for air pollution mapping, leading to the emergence of numerous local and global air pollution maps. These maps are now being increasingly utilized in urban and health studies. However, evaluating air pollution models and maps remains a challenge. One reason is the lack of air pollution measurements. A second reason may be attributable to a different focus on spatial heterogeneity in air pollution. For instance, He et al. (2022) acknowledge spatial heterogeneity in measurement stations as they show that the probability density function of concentrations (NO, NO_2, PM10, PM2.5) of different spatial categories (urban traffic; suburban/rural traffic; urban industrial; suburban/rural industrial; urban background; suburban background; rural background) show different patterns.

However, the research does not focus on modeling thus potential differences in prediction accuracy for every concentration type and spatial category. Another reason why evaluating air pollution models and maps remains challenging is that most of the current statistical modeling approaches only assess the overall accuracy but not the accuracy over space (Hoek et al., 2008; Chen et al., 2019). Hoek et al. (2008) state that a LUR model typically explains 60-70% of the variation in NO_2. However, the explained variation may be very low in areas near traffic. Chen et al. (2019) argue that most of the previous air pollution exposure assessment studies make no distinction in the characteristics of the monitoring sites when performing cross-validation, potentially leading to misrepresenting model results. Therefore, they opt to "evaluate models using pollution data collected from monitoring sites which represent the application locations" (Chen et al., 2019, p.3). Shaddick et al. (2020) also argue that the uncertainty in air pollutant measurements is only discussed in limited studies. A consequence of the inadequate evaluation is that the non-extrapolating property of most non-parametric machine learning methods is commonly ignored. Areas to be predicted could differ considerably in their societal and environmental properties compared to the training data, yielding highly biased predictions, which are not evaluated in multiple studies (Shaddick et al., 2020).

Given the increasing number of modeling and prediction techniques, and the presence of misrepresented prediction maps due to heterogeneity issues, this study aims to understand *to what extent can statistical models be used in predicting NO_2 concentrations given high-quality, high-temporal resolution ground station measurements: how does the performance of statistical models differ and how does it differ spatially?* The study area is in the Netherlands and Germany. Two datasets are used. One is the official national ground station measurements of the two countries (OpenAQ, 2017; EEA, 2021), and the other is the ground station measurements of the Amsterdam area (Gemeente Amsterdam, 2022). The aims are to compare modeling accuracy and prediction patterns 1) using the national ground stations and a set of more densely distributed local ground stations; 2) in areas close and far away to traffic and with different population densities, and 3) to using different statistical models considering, and understand the added value of modeling spatial correlation.

2 Methodology

2.1 Data

We refer to the dataset that consists of all the ground station measurements in Germany and the Netherlands as the "global dataset", and the dataset that consists of ground station measurements in the Amsterdam area as a "local dataset". The global- and local datasets contain the annual mean of NO_2 concentrations, measured in μg/m3, for 2017 (OpenAQ, 2017; EEA, 2021). Figure 1 shows the distribution of NO_2 concentrations at the global and local measurement stations.

Supplementary, figure 1a shows the spatial distribution of NO_2 measurement stations. Urban areas generally have a higher density of measurement stations. The urban area of Amsterdam is used to evaluate the global model and examine the differences between the global- and local models. Additionally, the direct, less densely populated areas around Amsterdam are included to examine the effect of the urban area on the predicted NO_2 concentration levels per local model.

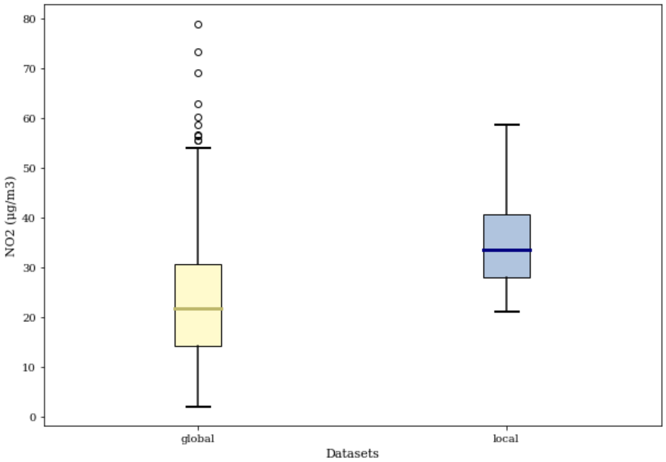


Figure 1. NO_2 distributions of global- (yellow) and local (blue) model

To examine whether the prediction quality differs between areas with different spatial characteristics (e.g. high vs. low road density), observations of the global and local datasets are split into three spatial groups, based on several population and traffic-oriented variables. Data for the year 2015 from the Global Human Settlement layer population grid is used for the population variable (JRC, 2015); The information on road length in meters is derived from OpenStreetMap (2019). Descriptive statistics of the variables that determine the spatial groups are visible in table 1.

Table 1. Descriptive statistics for each relevant variable in the determination of spatial groups for the local- and global datasets

Variable		Mean	Min	25%	75%	Max
Road class 1 100m (total length of highway (m))	Local data	2154.787	0	0	3001.109	12950.676
	Global data	12.295	0	0	0	982.912
Road class 2 100m (total length of primary roads (m))	Local data	4018.626	0	2367.599	5348.419	9596.102
	Global data	68.943	0	0	0	735.144
Road class 3 100m (total length of local roads (m))	Local data	25838.098	6483.437	18085.396	33039.556	50712.625
	Global data	272.059	0	29.281	406.097	1088.154
Population 1000m	Local data	111157.013	20097.258	106347.117	128723.570	137546.047
	Global data	6154.486	0	2204.520	9036.756	20300.887

The three spatial groups are defined as follows:

1. Urban: observations that belong to this group are within 100 meters of either road class 1 (highway) or 2 (primary roads) and the population 1000 (population density within 1000 meters of every measurement station) values are in the highest 25%; or the road class 3 (local roads) values are in the highest 25% and the population 1000 values are in the highest 25%.

2. Low population: observations belonging to this group are within 100 meters of either road class 1 or 2 and the population 1000 values are in the lowest 75%; or the road class 3 values are in the highest 25% and the population 1000 values are in the lowest 75%.

3. Far from roads: observations belonging to this group are further than 100 meters of either road class 1 or 2; or the road class 3 values are in the lowest 75%.

By applying this division [1], 85 observations classify as "urban", 138 as "low population" and 259 as "far from road", together compromising the 482 observations of the global dataset. Since the local dataset contains fewer samples and is characterized by higher population areas, the threshold is adjusted from 0.75 to 0.5 (i.e. "urban" is now related to the 50% highest values rather than the 75% highest values). For the local dataset, 56 observations are attributable to the spatial group "urban", 46 to "low population", and 30 to "far from road". Supplementary, figure 2 and 3 display the spatial distribution of observations through spatial groups, for the global- and local datasets, thereby including information on the spatial groups. Supplementary figure 4 (global dataset) and 5 (local dataset) show the measured NO_2 per station.

Spatial predictors

A set of variables with related data is already derived from Lu et al. (2020), including industrial areas from OpenStreetMaps, road length from OpenStreetMaps, population from GHS-POP R2019A population grid, and Earth night light from VIIRS in various buffers, wind speed and temperature at 2 m altitude from ERA-LAND 5 climate re-analysis model, elevation from 30 m Radar global product, Tropomi level 3 NO_2 of 2018 and global radiation. An overview of the variables derived from Lu et al. (2020) can be found in supplementary, table 1. We used the precipitation from weather stations (National Centers for Environmental Information, 2017) and conducted spatial interpolation using ordinary kriging to cover the NO_2 measurement stations. Kriging parameters can be found in supplementary, parameters. The precipitation consists of average monthly precipitation data, measured in millimeters. The “World Settlement Layer 2015” determines the building density. This layer is publicly available on figshare (Marconcini et al., 2020). Taking building density as an explanatory variable, multiple studies use several measurement scales, consisting of buffers that vary in size. Beelen et al. (2013) quantify several buffers representing building density whereby parameters are 100m, 300m, 500m, 1000m, and 5000m. Another study performed by Kheirbek et al. (2014) measures the correlation between air pollution and building density via 15 circular buffers, ranging from 50m to 1000m. Lu et al. (2020) also use buffers for one explanatory factor to encourage comprehensiveness within the methodology. Varying buffer sizes are implemented in our study too. As several measuring stations are close to each other, especially in urban areas, the maximum buffer size is set to 1000m. Complementary, 100m and 500m circular buffers are included. The Normalized Difference Vegetation Index (NDVI) values are obtained through NASA and are related to 2017 (NASA, 2017). The Dutch dataset for traffic volume is obtained via Nationaal Dataportaal Wegverkeer (NDW) (Rijkswaterstaat, 2017) whereas the German dataset for traffic volume is obtained via Bundesanstalt für Strassenwesen (BAST) (Bundesanstalt für Strassenwesen, 2017). Both the NDW- and BAST-datasets are generated via automatic counting stations. The traffic volume is expressed in

[1] The related code is visible in supplementary, Code 1

average hourly traffic, measured over 2017, and ranges from 25m, 50m, 100m, 400m, and 800m. The formula for calculating average hourly traffic can be found in supplementary, equations.

2.2 Modeling NO_2 globally and locally

2.2.1 Ensemble trees

The global models can be classified into two types of statistical learning methods. The first group composes ensemble tree-based approaches consisting of random forest, Light Gradient Boosting (LightGBM), and Extreme Gradient Boosting (XGboost). To potentially improve the accuracy of the non-linear models, parameters are tuned. With this, reducing overfitting, limiting complexity, improving computational efficiency, and encouraging the model's robustness are important aspects. For the random forest model, the number of estimators is set to 1000; the min_samples_split equals 10; the min_samples_leaf equals 5; the maximum features used per tree is set to 4; the maximum depth is 10; bootstrapping is allowed and the random state is included for reproducibility purposes. For both the LightGBM and XGboost models, the number of estimators is set to 50,000; the reg_alpha equals 2; the reg_lambda equals 0; the max_depth equals 5; the learning rate is 0.0005 and the random state is included to ensure reproducibility. Additionally, the gamma of the XGboost model is set to 5. Further details can be found in supplementary, parameters. The equations for the ensemble trees can be found in supplementary, equations.

2.2.2 Multiple linear regression

The key variables highlighted by the random forest model are chosen as predictors in Multiple Linear Regression (MLS). MLS, a statistical method employing multiple explanatory variables to forecast the response variable, operates within a linear framework, where the relationship between predictors and response follows the form Y = Xß. However, linear regression techniques can be characterized by complexity and overfitting. In this context, Least Absolute Shrinkage and Selection Operator (LASSO) and RIDGE regression emerge as broader forms of linear regression models, incorporating regularization terms, unlike "pure" linear regression lacking such regularization. The RIDGE regression puts constraints on the model coefficients whereby the penalty term, lambda λ, regularizes coefficients with large values, consequently reducing the complexity and multi-collinearity. The LASSO regression is different because the penalty equals the sum of the absolute values of the coefficients (Ren et al., 2020). Again the lambda serves as a penalization factor, however, this time stressing the magnitude of the predictors' coefficients, rather than the square of the coefficients. Consequently, coefficients can be equal to 0 which leads to feature selection. For modeling, the alpha for the LASSO and RIDGE models are tuned to 0.1, leading to the lowest MAE, RMSE, and highest R^2 out of options ranging from 0.1 to 1 with a step of 0.1. The parameters and equations for the linear regression, error term, RIDGE regression, and LASSO regression can be found in supplementary, parameters and supplementary, equations, respectively.

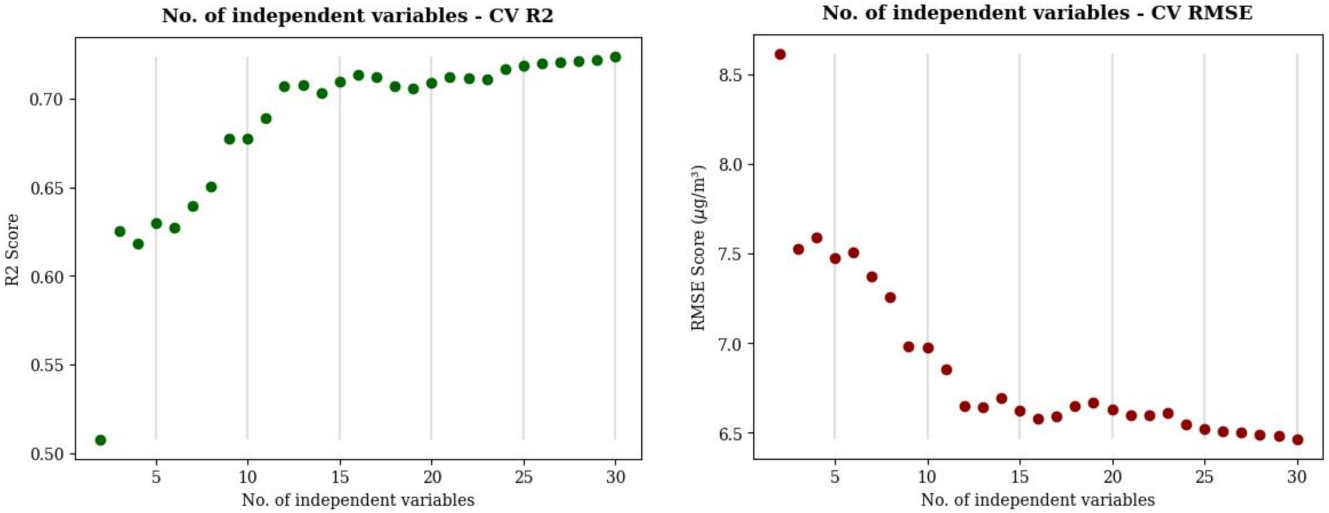
2.2.3 Mixed-effects model and kriging

To use the spatial information, we add mixed-effects modeling and kriging methods. With the mixed-effects model, fixed and random effects are included. Fixed effects consist of the most influential predictors while random effects account for potential spatial trends in the data. The spatial character of the observation, i.e. whether an observation is situated in an urban area, low-populated area, or far from road area, accounts for the random effect in the model. In contrast, the linear model composes all the fixed effects while neglecting the possibility of observation clustering. Additionally, two kriging methods are used for local modeling: ordinary and universal kriging. The kriging parameters are fit using the R package automap (Hiemstra et al., 2008). Details are found in supplementary parameters and equations. For the universal kriging and linear modeling method, a model is being created that does not account for the spatial groups, and a model is being created that does take the spatial groups into account. When accounting for spatial groups, a model is created for each spatial group from which NO_2 will be derived. Eventually, eleven models are used and compared, five based on the global dataset; six derived from the local dataset. The relevant equations can be found in supplementary, equations.

2.3 Feature selection

We first select features for global models based on the Shapley value (Shapley, 1953). The variable selection removes irrelevant or colinear predictors that would otherwise generate unstable estimates (Araki et al., 2018). Feature selection is based on examining the Shapley values of each feature (i.e. predictor). The Shapley value for a feature value j is determined by the contribution ϕ_j of feature j to the prediction, in this case, NO_2 concentration levels, compared to the average prediction of the dataset (Shapley, 1953). The contribution of a feature is calculated by examining the difference between the response variable that is obtained when the feature is present in comparison to the response variable that is obtained when the feature is absent (i.e. marginal contribution) (Algaba et al., 2019; Shapley, 1953). The formula for calculating the Shapley value can be found in supplementary, equations.

The random forest algorithm is used to determine the influence of each feature on the response variable (i.e. NO_2 concentrations). The idea of Shapley feature selection is embodied in an out-of-sample performance 10-fold cross-validation. The ranking of variables of each iteration is based on the Shapley value. After performing the iterations, the median of the out-of-sample performances 10-fold cross-validation is used to identify each feature's importance to the variance in NO_2. The relative positions of each predictor for the median-based approach can be found in supplementary, figure 6. The Shapley ranking of one fold is visible in supplementary, figure 7. To determine the preferred number of predictor variables, a random forest algorithm is applied to each number of most influential features, based on the average median ranking, ranging from the two most influential to the thirty most influential features. Thereafter, the RMSE and R^2 are used to determine the optimal number of predictors used for modeling. The number of predictor variables and the evaluation scores (R^2, RMSE) are visible in figure 2a, and figure 2b. A remarkable prediction accuracy improvement is obtained when considering at least twelve predictors. However, the improvement is marginal considering more than twelve covariates as the curve flattens.



(a) Number of features - R^2 score

(b) No. features - RMSE score

Figure 2. Out-of-sample performances ten-fold cross-validation: no. features - model performance (global)

Due to the poor performances by the random forest model over all the local station measurements (supplementary, figure 8a, 8b, and 8c), and per spatial group (supplementary, table 2), the random forest algorithm is not applicable to identify the (number of) variables for the local models. Rather, the best subset regression is used for variable selection for the local models. This approach consists of testing all possible combinations of predictor variables, thereafter selecting the model based on some statistical criteria (Kassambara, 2018). The maximum number of predictors considered is equal to 30. The relevant statistical criteria are the adjusted R^2, Mallows CP, and Bayesian Information Criteria (BIC) scores. Additional strategies may be necessary as the preferred number of features may differ per statistical criterion. Kassambara (2018) points to a rigorous approach whereby k-fold cross-validation is applied to select a model based on the prediction error computed on the new test data. Using the relationship between the number of features and prediction error, eventually nine features are identified that will be used for local modeling.

2.4 Model comparison

The global models are compared between the random forest model, the LightGBM model, and the XGboost model, representing tree-based methods, and the LASSO and RIDGE-models, representing linear models; the local models are compared between linear models, mixed effect models, and kriging models; For every model, the spatial groups are compared in terms of R^2, RMSE, and MAE, as these metrics are a common and useful approach in several studies (Rybarczyk and Zalakeviciute, 2018; Ameer et al., 2019; Chang et al., 2020). Furthermore, the prediction patterns of the local- and global mappings are examined. With this, the mobile NO_2 map is used as a benchmark to evaluate every model's accuracy, as the area of interest is similar (Kerckhoffs et al., 2019). Table 2 shows the global and local models, relevant selected predictors in sequential order of importance, and relevant evaluation methods. Predictions based on the global models are made for different areas with different demographic characteristics, including two 700,000+ inhabitant cities Amsterdam and Hamburg, Utrecht, which represents a middle-sized city (350,000+ inhabitants), and Bayreuth, which represents a city of a relatively low population (70,000+ inhabitants). Predictions derived from the local model apply to the Amsterdam area only. Table 3 shows complementary information for the models in terms of model complexity and the potential presence of a spatial component.

Table 2. Global and local models defined by selected predictors, models evaluated, and evaluation method.

Model	Selected predictors	Models evaluated	Evaluation
Global model	population_3000 road_class_3_3000 trafbuf25 population_1000 nightlight_450 nightlight_3150 trafbuf50 road_class_3_300 bldden100 ndvi road_class_2_25 trop_mean_filt_2019	random forest XGboost LightGBM LASSO RIDGE	cross validation over the entire area cross validation over different land types comparing with Kerckhoffs et al. (2019)
Local model	population 1000 nightlight_450 nightlight_4950 population_3000 road_class_1_5000 road_class_2_1000 road_class_2_5000 road_class_3_100 road_class_3_300	linear model linear model separating for spatial groups mixed-effects model ordinary kriging universal kriging universal kriging separating for spatial groups	cross validation over the entire area cross validation over different land types comparing with Kerckhoffs et al. (2019)

Table 3. Global and local models defined by model complexity and consideration of the spatial component.

Model	Model complexity	Consideration of the spatial component
Linear regression	No regularization	Using rules to separate different land types and fitting the model separately
LASSO	L2 regularization	Not explicitly
RIDGE	L1 regularization	Not explicitly
Mixed effect	No regularization	Using rules to separate different land types and consider it as a random variable.
Kriging	No regularization	Distance (second-order stationarity) Using rules to separate different land types and fitting the model separately
Random forest	Controlled by hyperparameters: number of trees, minimum number of samples for splitting, minimum number of samples per leaf, maximum features per tree, maximum depth, bootstrapping	Not explicitly
XGBoost	Controlled by hyperparameters: number of estimators, alpha, lambda, learning rate, maximum depth	Not explicitly
LightGBM	Controlled by hyperparameters: number of estimators, alpha, lambda, learning rate, maximum depth, gamma	Not explicitly

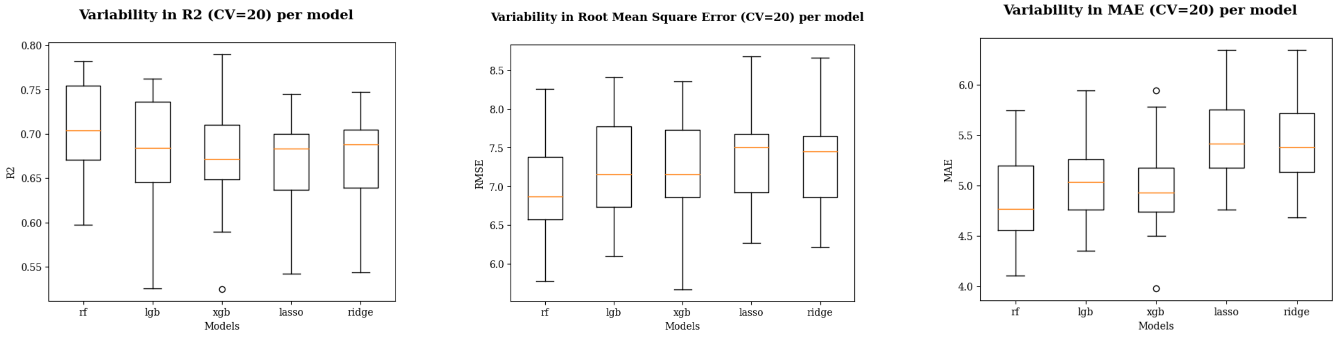
3 Results

3.1 Models

3.1.1 Global

Evaluating the different linear- and non-linear models is done by performing out-of-sample performances 20-fold cross-validation, thereby examining the variance and median statistics per model in terms of R^2, MAE and RMSE (figure 3a, figure 3b, and figure 3c). We select an out-of-sample performance 20-fold cross-validation to encourage stable estimates.

With out-of-sample performances 20-fold cross-validation, the linear models (i.e. LASSO and RIDGE) score similarly to the non-linear models, especially the R^2 and RMSE. Generally, the random forest model performs best out of the considered global models as the median R^2, median RMSE, and median MAE are best for this model. The robustness of the random forest model is relatively high too, as the standard deviation is lowest in terms of R^2 and RSME (figure 3a and figure 3b).



(a) R^2

(b) RMSE

(c) MAE

Figure 3. The out-of-sample performances 20-fold cross-validation: performance per model (a) R^2, (b) RMSE, (c) MAE (global). The upper- and lower quartiles indicate variability; RF = random forest, LGB = LightGBM, XGB = XGboost

Accounting for spatial characteristics

The comparison is not only made between linear and non-linear models, and global- and local models, but also between spatial groups, to account for the potential spatial heterogeneity of observations. Meyer and Pebesma (2021, 2022) argue that the increasing popularity of global maps, due to their ability to fit linear and non-linear and complex relationships, are subject to misinformation. A global model that is trained can make accurate predictions as long as global predictors are available, however at places where predictions are far beyond the data, poor predictions by a trained random forest model are not uncommon, given that predictor values of the predicted area do not resemble the training data. Therefore, the potential difference in global model prediction accuracy between different spatial groups is examined. Examining the descriptive statistics per spatial group already exposes interesting differences in terms of NO_2 concentration levels (table 4).

Table 4. Descriptive NO_2 statistics for each spatial group, measured in μg/m3

Group	Count	Mean	Sd.	Min	25%	50%	75%	Max
Urban	85	38.865	13.065	15.768	28.172	38.076	47.923	78.882
Low population	138	27.601	9.769	7.872	19.876	26.876	34.407	56.706
Far from road	259	16.653	8.341	2.122	10.331	15.892	22.518	48.887

Table 5 describes the performances, in terms of R^2, RMSE, and MAE, for each spatial group, per model. For each model and each performance criterion, the observations far from roads, perform considerably better than observations close to roads, for both urban and low-population areas. The non-linear models outperform the linear models when the data is trained on observations far from roads and observations close to roads but are characterized in low-population areas. For urban areas, the performances between the linear and non-linear methods are less distinguishable which might be explained by the relatively low number of observations. Due to the relatively limited number of observations and heterogeneous character of data in the urban class, ensemble tree-based methods have poor learning conditions, possibly having fewer abilities to learn and discover patterns.

			URB			LP			FFR		
Models			R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Non linear	RF	*Mean*	0.271	10.994	8.964	0.387	7.285	5.361	0.712	4.189	3.007
		SD.	0.099	1.298	0.950	0.185	1.323	0.762	0.102	0.983	0.550
	LightGBM	*Mean*	0.175	11.631	9.477	0.367	7.381	5.468	0.725	4.075	2.872
		SD.	0.145	0.226	0.955	0.226	1.513	0.739	0.120	1.040	0.537
	XGboost	*Mean*	0.228	11.230	9.147	0.426	7.060	5.228	0.737	3.991	2.774
		SD.	0.150	1.014	0.807	0.183	1.340	0.687	0.116	1.096	0.530
Linear	RIDGE	*Mean*	0.328	10.491	8.617	0.348	7.517	5.703	0.696	4.358	3.211
		SD.	0.127	1.080	0.860	0.167	1.139	0.606	0.103	1.133	0.564
	LASSO	*Mean*	0.265	10.936	9.017	0.282	7.859	6.047	0.613	4.912	3.749
		SD.	0.177	1.159	1.040	0.201	1.105	0.672	0.119	1.153	0.678

Table 5. Model performance per spatial group (CV = 20). RMSE and MAE are represented in NO_2 (μg/m3) values.

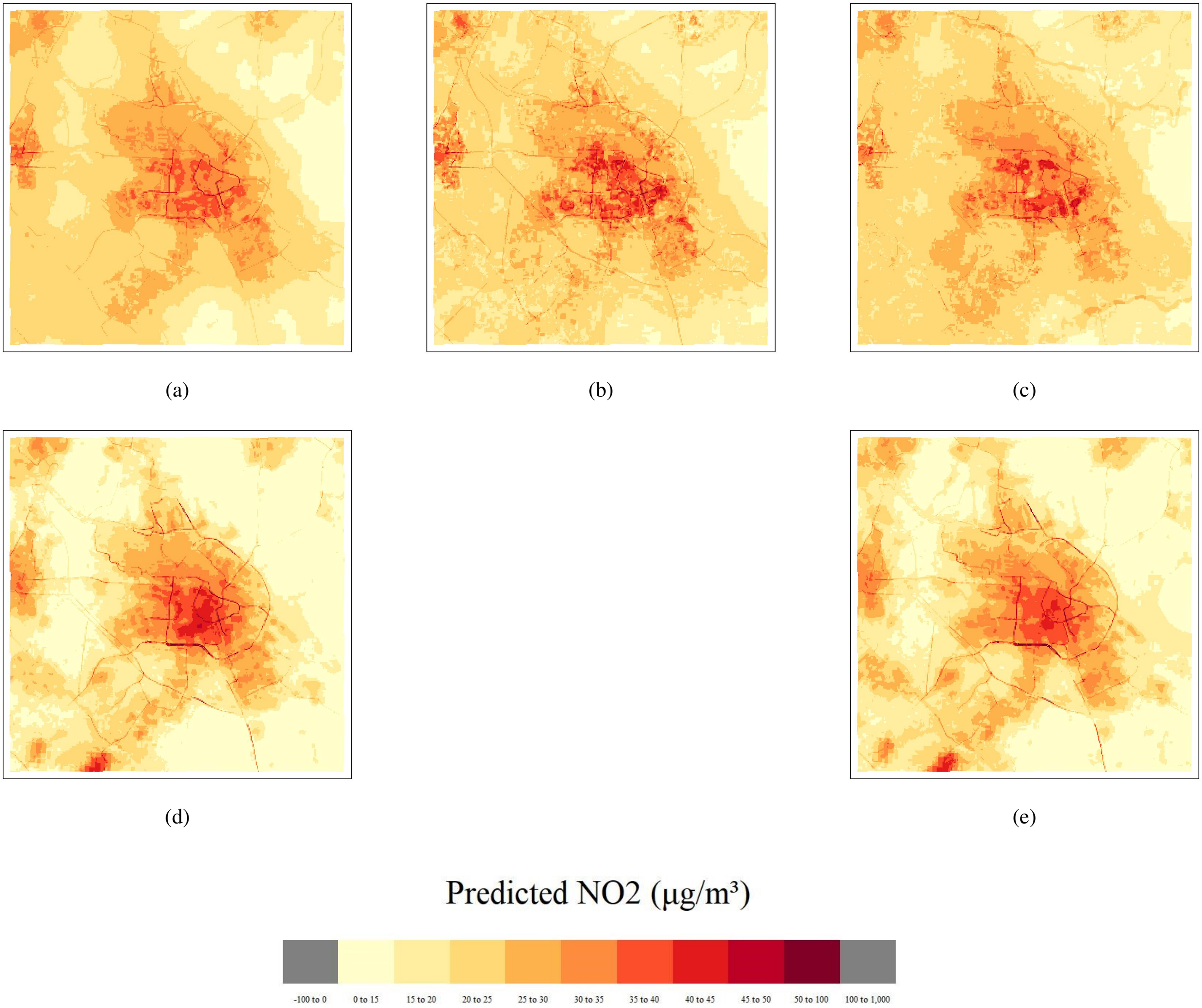
URB: Urban (near road, high population), LP: Low Population (near road, low population), FFR: Far From Road.

Spatial prediction patterns

Figure 4 shows the NO_2 spatial predictions for the Amsterdam area per model, a-c relating to the non-linear spatial predictions, d and e relating to the linear techniques. Generally, the linear spatial predictions seem more prone to overfitting as a higher share of extreme values characterizes these prediction maps compared to non-linear techniques (i.e. below 15μg/m3 or above 50μg/m3). Interestingly, linear techniques identify a high NO_2 hot spot in the southwestern part of the study area that is not identified by the non-linear techniques. Generally, major roads (highways, national roads) are identifiable and other urban areas (e.g. Haarlem) are characterized by high NO_2 concentration levels (supplementary, figure 9).

Figures 5 a-f show the spatial patterns of the predicted NO_2 concentrations for Hamburg (a and b), Utrecht (c and d) and Bayreuth (e and f) for the random forest and RIDGE models. The other models (LightGBM, XGboost, LASSO) for Hamburg, Utrecht, and Bayreuth (both zoomed in and out) can be found in supplementary sections figure 10a-c, figure 11a-c, figure 12a-c, figure 13a-e respectively. Comparing maps of these cities, there are noticeable differences in prediction patterns. A most important finding is that the highest air pollution seems to be situated around major roads in Hamburg while the urban center accounts for the highest air pollution concentrations in Utrecht. The high correlation between major roads and high air pollution could be reasonably explained considering that Hamburg is the 69th of the most congested cities in the world (Tomtom, 2021). Interestingly, the highest NO_2 concentration levels among highways differ spatially between the random forest and RIDGE models, for example, the highways in the southeastern and western part of the Hamburg area contain high NO_2 levels for the RIDGE models while a nuanced identification is related to the random forest prediction for the same area.

In the random forest prediction map for Hamburg, air pollution among roads in the center and northern part of the city is more pronounced compared to the RIDGE model equivalence. Additionally, the magnitude of high air pollution related to major roads is considerably higher for Hamburg, compared to Utrecht and Bayreuth. Still, the relationship between the presence of roads and heavier air pollution concentration is identifiable for both Utrecht and Bayreuth, especially with the RIDGE model predictions. For Utrecht, the urban center is more pronounced in terms of high NO_2 concentration levels, compared to Hamburg and Bayreuth. Moreover, the RIDGE model applied on Utrecht identifies more clusters (i.e. scattering) of NO_2 values in the periphery. In comparison, the predicted NO_2 values are more scattered in the urban center for the random forest when compared to the RIDGE model. Again, this difference in prediction patterns between a linear and non-linear model is apparent for the Amsterdam area. Bayreuth is characterized by moderate air pollution and very low ($<$15 μg/m3) pollution in rural areas surrounding the city - some clusters exceeding the 15 μg/m3 benchmark are noticeable that correspond to other villages in the area, hinting to the influence of population or building density on air pollution (see also supplementary, figure 13a-e). Supplementary, figure 14 shows the distribution of predicted NO2 per global model for each location.


(a)

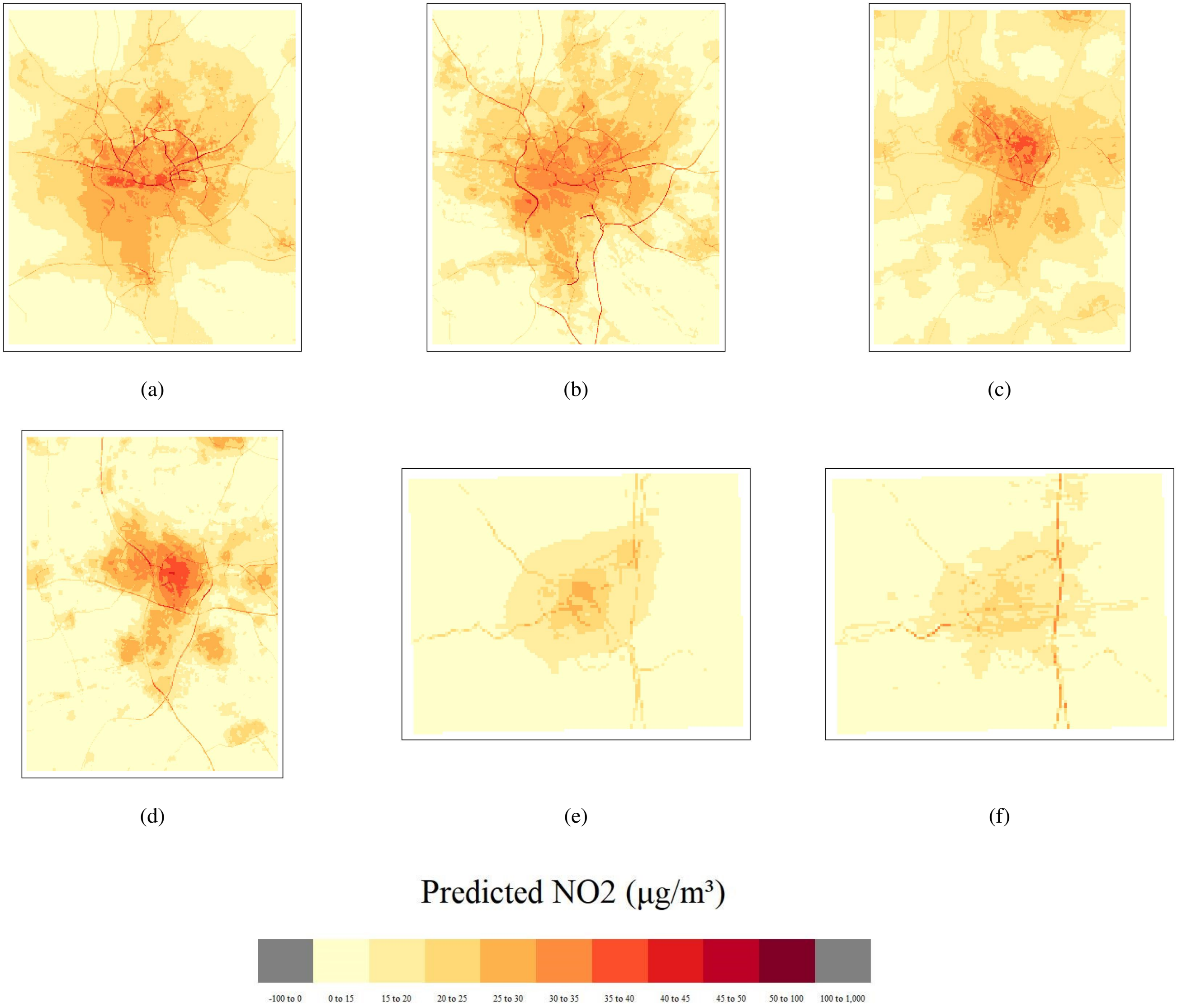
(b)

(c)

(d)

(e)

Figure 4. Spatial patterns of predicted NO_2 (100m), measured in μg/m3, per model for Amsterdam - non linear techniques (top): (a) = random forest, (b) = LightGBM, (c) = XGboost; linear techniques (bottom): (d) = LASSO, (e) = RIDGE. Extent = 30km x 30km



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5. Spatial patterns of predicted NO_2 (100m), measured in μg/m3, per model for Hamburg (extent = 30km x 30km), Utrecht (extent = 25km x 25km) and Bayreuth (extent = 10km x 10km) - top: left = random forest (Hamburg), right = RIDGE (Hamburg), right = random forest (Utrecht); bottom: left = RIDGE (Utrecht), middle = random forest (Bayreuth), right = RIDGE (Bayreuth)

3.1.2 Local

The performances of the local models are composed of the R^2, RMSE, and MAE. Table 6 shows the model performances for the linear model, the mixed-effects model, the ordinary kriging model, and the universal kriging model, whereby a leave-one-out cross-validation is applied. The ordinary kriging model shows the poorest performance, which can be explained by it's spatial prediction patterns (figure 6) and parameters (e.g. a limited range in which other observations are considered). Using auxiliary variables results in a prediction accuracy improvement as the metrics of the universal kriging model are considerably better than the metrics of the ordinal kriging model. However, accounting for spatial autocorrelation does not automatically result in a higher accuracy since the linear model performs better than the universal kriging method. At the same time, accounting for random effects yields a higher R^2, a lower RMSE, and a lower MAE.

Table 6. Model performance (CV = leave-one-out)

	R^2	RMSE (μg/m3)	MAE (μg/m3)
ordinary kriging	0.072	8.542	7.052
linear model	0.307	7.412	5.955
mixed-effects model	0.326	7.315	5.808
UK (model + kriged residuals)	0.277	7.749	6.097

Table 7 shows the model results per spatial group, again based on leave-one-out cross-validation. Similar to the results of the global model, the results for the local models indicate that models trained on "urban" observations perform poor - however, the proximity to the road does not necessarily influence the model performance since the R^2 of the low population class is higher than the R^2 of the far from road class. In contrast to the models trained on the global dataset, which perform best in far from road areas, the models trained on the local dataset perform best in areas with low populations and proximity to roads. An explanation may be that observations in "far from road" areas for the local dataset are more similar to observations in the urban and low-population areas when compared to the global dataset, as the predictor characteristics are more uniform in a relatively small area (i.e. local dataset) compared to a relatively large area (i.e. global dataset).

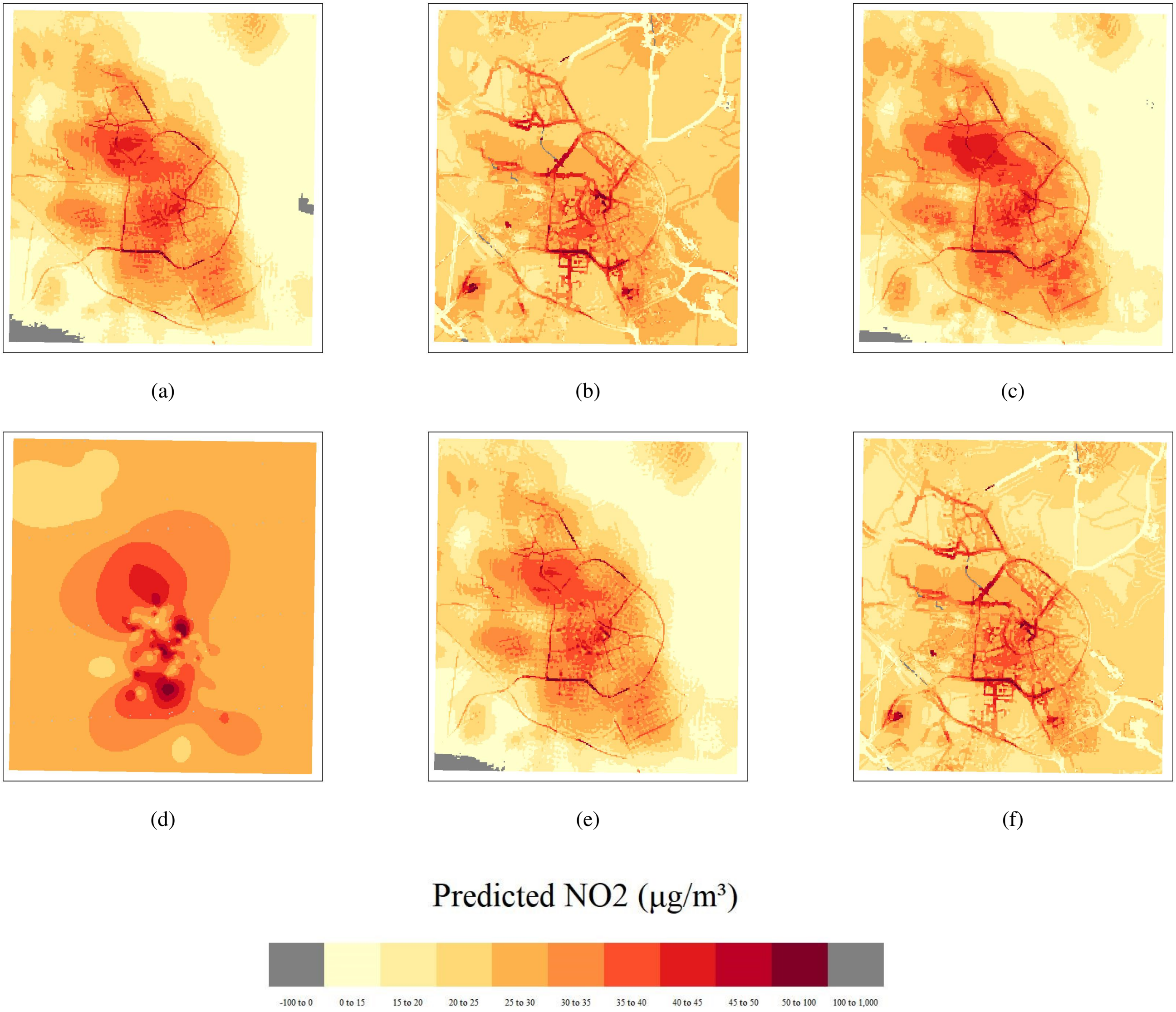
Table 7. Model performance per spatial group (CV = leave-one-out-cross-validation). RMSE and MAE in μg/m3

	URB			LP			FFR		
Models	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
ordinary kriging	0.072	8.257	6.772	0.223	8.558	6.575	0.072	9.029	8.303
linear model	0.140	7.890	6.360	0.509	6.8	5.301	0.147	7.390	6.202
mixed-effects model	0.141	7.874	6.316	0.524	6.505	5.298	0.115	7.404	5.644
UK (model + kriged residuals)	0.161	8.068	6.27	0.487	6.938	5.174	0.037	7.19	8.299

URB = Urban (near road, high population), LP = Low Population (near road, low population), FFR = Far From Road- Mixed-effects model results are identical

Spatial prediction patterns

Figure 6 shows the predicted NO_2 patterns based on the local dataset. The prediction map of the linear model (a) is fairly similar to the prediction maps of the mixed-effects (c) model and universal kriging (e) model - the models identify a high NO_2 concentration cluster at the northwestern part of Amsterdam. Further examination reveals that this cluster is likely highly influenced by the predictor road class 2 5000 (i.e. primary roads within 5000m), as this predictor shows a similar cluster at the same location (supplementary, figure 15, figure 16a-i). Two models that account for the spatial groups first, before the modeling process, show comparable patterns whereby the influence of roads is visible - be it via the predictors themselves, or the spatial groups (see also supplementary, figure 17). The relative low NO_2 values along the roads in the outer Amsterdam area can be attributed to the spatial grouping divisions. To extend, the presence of predictor values with high standard deviations can impact the NO_2 values for that specific spatial group, potentially leading to overestimation or underestimation in certain parts of the prediction area. The patterns that are along the roads belong to the spatial group "low population" whereby observations within this group are in the vicinity of roads (<100m). Comparing this spatial group, to the spatial group "far from road", the data distribution for every predictor in low population is substantially different than the data distribution for every predictor in the group "far from road", leading to different learning patterns which explain the relative high prediction values along the roads (supplementary, figure 18a-i). At some places, negative predicted values are apparent albeit few. This is likely a cause of the training dataset having different feature characteristics than the testing dataset. Comparing the local prediction patterns to the global prediction patterns, a cluster of high air pollution in the northwestern part of Amsterdam is visible in some local models that is not visible in the global models (as discussed, these models refer to the general linear-, mixed-effects-, and universal kriging models). A possible explanation for why the cluster is identified in the local dataset, as opposed to the global dataset, could be the difference in spatial distribution of NO_2 values between the local- and global datasets, resulting in different learning patterns between the local- and global models (figure 1). Supplementary, figure 19 shows the distribution of predicted NO2 per local model.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 6. Spatial patterns of predicted NO$_2$ (μg/m3) at 100m resolution based on local dataset - top: left = linear model, middle = linear model separating for spatial groups, right = mixed-effects model; bottom: left = ordinary kriging, middle = universal kriging, right = universal kriging separating for spatial groups

Model comparison

Figure 7 shows the correlation in predicted NO_2 values for the local- and global models, as well as the mobile NO_2 map of Kerckhoffs et al. (2019) which is used as a benchmark. Some outliers are present in the local linear model that accounts for spatial groups and the universal kriging model that accounts for spatial groups, resulting in a relatively low correlation with the open NO_2 dataset. Therefore, outliers are removed to enhance the correlation between the mentioned models on the one hand and the NO_2 dataset on the other. There are a plethora of outlier detection methods, however multiple techniques such as chi-squared test, Dixon test, Grubbs test, and 1.5 times IQR, label values that would be within the IQR of other models, as outliers. Therefore, a manual threshold is chosen based on other models' data distributions. Since the maximum value of the ten models (excluding the two where the outlier detection is applied first) is 85, the outlier detection threshold will be tuned to 85 as the upper bound. For the lower bound, negative values are considered as outliers. Afterward, the outlier filtering is applied to all models, potentially filtering out negative prediction values for other models. The new correlation matrix is visible in supplementary, figure 20. Furthermore, global models are highly correlated; interestingly the LASSO model has fewer prediction pattern similarities with the other global models. The ordinary kriging model has low similarity with other models which may not be surprising, given the model settings (e.g. a low searching radius) and prediction pattern. Comparing the models to the mobile NO_2 map, the local models generally show more similarity than global models. This may be because the local dataset and the mobile NO_2 map are projected on the Amsterdam area.

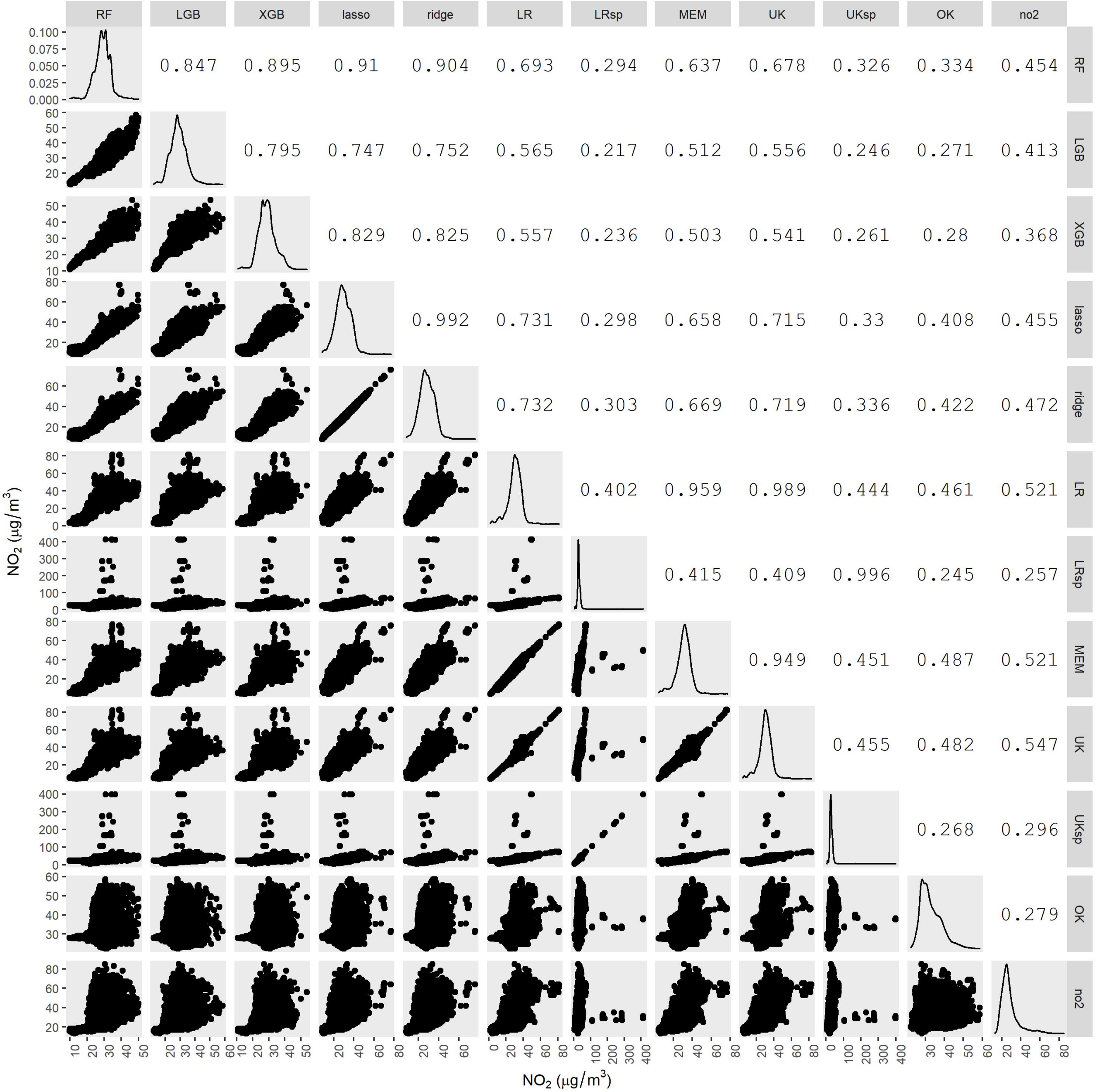


Figure 7. Comparing model predictions whereby the numbers equal the Pearson correlation coefficient. RF = random forest, LGB = LightGBM, XGB = XGboost, LR = linear regression, LRsp = linear regression accounting for spatial groups, MEM = mixed-effects model, UK = universal kriging, UKsp = universal kriging accounting for spatial groups, OK = ordinary kriging, no2 = mobile NO_2 map.

4 Discussion

Precise modeling and estimation of NO_2 concentration levels are essential for understanding air pollution comprehensively, thereby deducing political implications that can encourage a healthy society. Several important findings of this research, related to this, are discussed below.

Relationship between predictors and other pollutants

For both the global and local datasets, traffic and population density variables are selected among the most influential predictors, a similar finding to the statements of Beelen et al. (2013) as they argue that including such variables encourages prediction accuracy. Additionally, the high influence of traffic on NO_2 concentrations match the findings of Lu et al. (2020) and Chen et al. (2019). Chen et al. (2019) moreover find that the most important feature(s) differ per pollutant measured. As this research only predicts NO_2 concentrations, it will be interesting to see to what extent choosing a different pollutant results in different models and prediction patterns. For instance, major roads and highways are identifiable in Amsterdam, Bayreuth, Hamburg, and Utrecht, hinting at the high influence of traffic on NO_2 concentrations. As other features may be chosen as the highest influencers on pollutants other than NO_2, prediction maps are subject to change too, potentially leading to substantially different prediction maps for CO, O_3, PM2.5, PM10, and SO_2. Examining to what degree modeling and prediction patterns differ for different pollutants is therefore important to better understand the relationship between space and health, especially since the impact of air pollution on health may differ per pollutant (He et al., 2022).

Addition of temporal analysis

The temporal analysis in this research is limited and can be extended. For instance, Lu et al. (2020) account for day and night pollution. Complementary, a distinction between week- and weekend days can be executed, and certain events, such as COVID, could influence the concentration levels of pollutants. Accounting for those temporal aspects potentially influences the feature importance, modeling, prediction maps, and prediction quality per spatial group. For instance, Wong et al. (2021) state that the atmospheric environment is substantially improving during the COVID-19 pandemic - a period in which urban mobility is virtually zero. The importance of time on air pollution predictors such as traffic and air pollution itself is demonstrated by Kendrick et al. (2015) too, as they argue that NO_2 has seasonal and diurnal variation as a function of traffic volumes alongside a major arterial, and Minoura and Ito (2010) by stating that traffic signals are related to roadside air quality in Tokyo, Japan. Therefore, future research should implement temporal models, for instance, ARIMA and its relatives (Wong et al., 2021), or apply spatio-temporal models (van Zoest et al. (2020)) to examine the effect of time or space-time on feature influence and model accuracy, not only between global and local datasets but also between spatial groups (e.g. urban, low population, far from the road) and administrative regions such as cities. Simultaneously, the absence of the temporal dimension poses challenges in interpretation, uncertainty assessment, and spatial prediction. Still, joint spatio-temporal modeling greatly complicated the modeling and we believe it is more illustrative and reprehensible to first study at a lower dimensional (Wikle et al., 1998, 2019).

Accounting for spatial groups

Considering the global dataset, the differences between the linear and non-linear techniques are barely noticeable. Although the random forest model generally performs best (highest R^2, lowest MAE), the R^2 of the RIDGE is higher than the LightGBM and XGboost models. However, when accounting for the spatial groups - urban, low population, and far from roads - the differences in model performance between linear and non-linear techniques become more distinguishable, whereby the latter techniques perform better. This is especially true for observations far from roads, where data generally is more homogeneous, while the model performances between linear and non-linear techniques in urban areas are less pronounced, albeit both techniques perform poorly here. So, in the first instance, this study does not acknowledge better prediction performances by the non-linear techniques, as suggested by Weichenthal et al. (2016), Reid et al. (2015), Chen et al. (2019), and Lu et al. (2020), however when accounting for spatial characteristics, non-linear techniques perform considerably better than linear techniques, especially in more homogeneous areas which support the arguments made by Meyer and Pebesma (2021) that non-linear predictions are of higher quality in areas that have similar environmental variables as the training data. Although there are different cross-validation methods available, whereby some research claims that spatial cross-validation better captures the important phenomenon of autocorrelation, we used random bootstrap cross-validation instead of spatial cross-validation. Wadoux et al. (2021) argue that standard cross-validation (i.e. ignoring autocorrelation) results in smaller bias than spatial cross-validation. Moreover, they state that spatial cross-validation methods should not be used for map assessment as they have no theoretical underpinning. In contrast, standard cross-validation is applicable and is sufficient in clustered data scenarios (Wadoux et al., 2021; Lu et al., 2023).

Urban areas' more heterogeneous data nature converges performance between linear and non-linear techniques. Thereby, both techniques perform poorly which, when not controlling for spatial characteristics of the observations, may have gone unnoticed. The poor prediction accuracy in urban areas is worrisome given that the impact of air pollution can depend on the surrounding environment, i.e. people who live in the vicinity of traffic-heavy roads (which are often more present in, or around urban areas) and/or industries facing higher exposure to air pollution (He et al., 2022). Though spatial grouping greatly improves the predicting reliability, it can present counter-intuitive patterns. For instance, in some areas, the predicted NO_2 concentration levels are lower along roads than the concentration levels of the rural surroundings. Moreover, Patelli et al. (2023) identify three main categories in which random forests can be linked to spatial data, being pre-, in-, and post-processing. While random forest performance is linked with spatial groups in our study (which can arguably be linked to a form of post-processing), there is potential in better integrating spatial data in ensemble tree-based models such as random forests, to increase predictive performance potentially (Patelli et al., 2023).

Spatially varying on feature importance

While the feature importance is equal between cities, the influence of predictors on NO_2 concentrations differs between the case cities. For instance, building density and population are more prevalent contributors to air pollution in Utrecht, compared to Hamburg, while traffic has a higher influence on high NO_2 concentrations in Hamburg, compared to Utrecht. Therefore, it is suggested further to examine the effect of space on predictor influence to stress the importance of not only global and national but also local policies in managing air pollution. Additionally, global models are applied to different cities with the same predictors. As the case cities unravel that high NO_2 can be attributed to different predictors per city, applying models with different features may yield better prediction results. An important condition is that every city has enough observations to avoid unreliable predictions.

Model quality

The limited number of observations in the local dataset poses a problem in making accurate predictions, especially for non-linear techniques. Consequently, six different models are chosen for the local dataset. Interestingly, accounting for spatial autocorrelation does not equal model improvement while accounting for spatial groups does. Outliers are omitted after the model predictions to deal with unreliable predictions. Transforming the original data could avoid data out of range (e.g. $<$ 0 mg/m^3). In our study, such transformation, e.g., a log transformation, is not applied, however, the application could have yielded a better data distribution of predictors, potentially leading to prediction improvements. Complementary, airborne pollutant concentrations are often positively skewed (Maranzano et al., 2020). To adjust for positive skewness, transformations can be applied but also cause prediction changes which currently are not revised in our research. Simultaneously, Lu et al. (2023) examine several techniques such as transformations, likelihood functions, and loss functions to address the issue of non-Gaussian distributions. Thereby, they observed that using a transformation, likelihood function, and loss function that matches with the more-likely distribution (i.e. Gamma) does not improve the modeling results but worsens the prediction errors and the uncertainty quantification (Lu et al., 2023).

Moreover, while the LASSO and RIDGE models seem useful with the global dataset, the predictions are unsatisfactory with the local dataset. Given that the "road class 2 5000" has relatively high values and high influence on the NO_2 concentration levels in local modeling, using a LASSO and/or RIDGE model would have been useful. Also, important model features may be analyzed in more detail to get a deeper understanding of air pollution. In the models of this study, traffic is a prevalent feature, however, no distinction is made between traffic type (e.g. cars, buses, trucks), car type (e.g. electric, diesel), and distance traveled while such aspects can be influential to air pollution, as suggested by Wong et al. (2021). Performing a deeper analysis of features may unravel interesting findings that would otherwise have been unnoticed, for instance, distinguishing between vehicle types may show that relatively many trucks are on specific roads (e.g. going to or from the port of Hamburg) which might explain certain clusters of high NO_2 concentration levels at certain places.

The relationship between air pollution and socioeconomic status

As the relationship with socioeconomic status is not evaluated in this study, future research should examine the relationship between linear and non-linear predictions with socioeconomic status so that not only the quantification of exposure but also the assessment of inequality in health. For instance, Jiao et al. (2018) argue that the relationship between socioeconomic factors and the health environment is understudied. Additionally, Bai et al. (2019) mention that very few studies have analyzed the influences of air pollution from the perspective of meteorological factors, pollution sources, and socioeconomic status. Next to the missing link with socioeconomic status, the temporal aspect is important when examining air pollution and society, and human exposure to air pollution (Lu et al., 2020).

5 Conclusions

In this study, we compare various statistical models for predicting NO_2 concentrations at different scales (local vs. global) for their spatial and overall prediction accuracy, as well as NO_2 maps of cities with distinctive characteristics such as population, and the importance of features. One of the key findings of this study is that the model performance varies little with models of different levels of complexity, but spatially due to spatial heterogeneities in traffic and urban features. Non-linear techniques predict better in areas far from roads and in areas near roads but with low population density, compared to linear models. Additionally, global model prediction accuracy is considerably higher in areas far from roads than in areas near roads. Furthermore, population density is associated with global model prediction accuracy, whereby low-populated areas yield higher prediction accuracy than high-populated areas. In contrast, methods preferred in global modeling appear to be unfavorable in local modeling. Local models show that low-populated areas near roads yield higher prediction accuracy than high-populated areas near roads or areas far from roads. The relatively few NO_2 observations used in the local models could explain why non-linear models perform poorly. The prediction accuracy in spatial groups differs from the global models. We also found that modeling the spatial autocorrelation does not improve the local modeling accuracy, but accounting for spatial groups does. Lastly, non-linear prediction patterns are less prone to overfitting compared to linear methods, and different modeling techniques lead to different NO_2 clusters in the prediction map. These results suggest that only looking at the prediction accuracy is insufficient for evaluating the statistical models in air pollution prediction.

Code and data availability

Available via: https://github.com/FoekeBoersma/A-close-look-at-using-national-ground-stations-for-the-statistical-mapping-of-NO2 and https://doi.org/10.5281/zenodo.8397133

Datasets larger than 100MB are included and can be accessed in another repository via: https://doi.org/10.5281/zenodo.7948161

Author contributions.

Conceptualization, F.B. and M.L.; methodology, F.B. and M.L.; validation, F.B.; formal analysis, F.B.; investigation, F.B.; resources, F.B. and M.L.; data curation, F.B.; writing—original draft preparation, F.B. and M.L.; writing—review and editing, F.B. and M.L.; visualization, F.B.; supervision, M.L.; project administration, F.B. and M.L.; funding acquisition, F.B. and M.L. All authors have read and agreed to the published version of the manuscript.

Competing interests.

The authors declare that they have no conflict of interest.

References

Algaba, E., Fragnelli, V., and Sánchez-Soriano, J.: Handbook of the Shapley value, CRC Press, 2019.

Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., and Asghar, M. N.: Comparative analysis of machine learning techniques for predicting air quality in smart cities, IEEE Access, 7, 128 325–128 338, 2019.

Araki, S., Shima, M., and Yamamoto, K.: Spatiotemporal land use random forest model for estimating metropolitan NO2 exposure in Japan, Science of The Total Environment, 634, 1269–1277, 2018.

Bai, L., Jiang, L., Yang, D.-y., and Liu, Y.-b.: Quantifying the spatial heterogeneity influences of natural and socioeconomic factors and their interactions on air pollution using the geographical detector method: A case study of the Yangtze River Economic Belt, China, Journal of Cleaner Production, 232, 692–704, 2019.

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M.-Y., Künzli, N., Schikowski, T., Marcon, A., et al.: Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe–The ESCAPE project, Atmospheric Environment, 72, 10–23, 2013.

Brokamp, C., Jandarov, R., Rao, M., LeMasters, G., and Ryan, P.: Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches, Atmospheric Environment, 151, 1–11, 2017.

Bundesanstalt für Strassenwesen: Automatische Zähltellen 2017, https://www.bast.de/DE/Verkehrstechnik/Fachthemen/v2-verkehrszaehlung/Daten/2017_1/Jawe2017.html?nn=1819490, 2017.

Chang, Y.-S., Chiao, H.-T., Abimannan, S., Huang, Y.-P., Tsai, Y.-T., and Lin, K.-M.: An LSTM-based aggregated model for air pollution forecasting, Atmospheric Pollution Research, 11, 1451–1463, 2020.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U. A., Katsouyanni, K., et al.: A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide, Environment international, 130, 104 934, 2019.

EEA: Explore Air Pollution Data, https://www.eea.europa.eu/themes/air/explore-air-pollution-data, 2021.

Gemeente Amsterdam: Luchtkwaliteit-NO$_2$-metingen, https://maps.amsterdam.nl/no2/?LANG=nl, 2022.

He, H., Schäfer, B., and Beck, C.: Spatial heterogeneity of air pollution statistics in Europe, Scientific Reports, 12, 12 215, 2022.

Hiemstra, P., Pebesma, E., Twenh"ofel, C., and Heuvelink, G.: Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network, Computers Geosciences, dOI: http://dx.doi.org/10.1016/j.cageo.2008.10.011, 2008.

Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., and Briggs, D.: A review of land-use regression models to assess spatial variation of outdoor air pollution, Atmospheric environment, 42, 7561–7578, 2008.

Jiao, K., Xu, M., and Liu, M.: Health status and air pollution related socioeconomic concerns in urban China, International Journal for Equity in Health, 17, 1–11, 2018.

JRC: GHS population grid, derived from GPW4, multitemporal (1975, 1990, 2000, 2015)., European Commission, Joint Research Centre (JRC); Columbia University, Center for International Earth Science Information Network., 2015.

Kassambara, A.: Machine learning essentials: Practical guide in R, Sthda, 2018.

Kendrick, C. M., Koonce, P., and George, L. A.: Diurnal and seasonal variations of NO, NO2 and PM2. 5 mass as a function of traffic volumes alongside an urban arterial, Atmospheric Environment, 122, 133–141, 2015.

Kerckhoffs, J., Hoek, G., Portengen, L., Brunekreef, B., and Vermeulen, R. C.: Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces, Environmental science & technology, 53, 1413–1421, 2019.

Kheirbek, I., Ito, K., Neitzel, R., Kim, J., Johnson, S., Ross, Z., Eisl, H., and Matte, T.: Spatial variation in environmental noise and air pollution in New York City, Journal of Urban Health, 91, 415–431, 2014.

Lu, M., Schmitz, O., de Hoogh, K., Kai, Q., and Karssenberg, D.: Evaluation of different methods and data sources to optimise modelling of NO2 at a global scale, Environment international, 142, 105 856, 2020.

Lu, M., Cavieres, J., and Moraga, P.: A Comparison of Spatial and Nonspatial Methods in Statistical Modeling of NO 2: Prediction Accuracy, Uncertainty Quantification, and Model Interpretation, Geographical Analysis, 55, 703–727, 2023.

Maranzano, P., Fassò, A., Pelagatti, M., and Mudelsee, M.: Statistical modeling of the early-stage impact of a new traffic policy in Milan, Italy, International journal of environmental research and public health, 17, 1088, 2020.

Marconcini, M., Metz-Marconcini, A., Üreyen, S., Palacios-Lopez, D., Hanke, W., Bachofer, F., Zeidler, J., Esch, T., Gorelick, N., Kakarla, A., et al.: Outlining where humans live, the World Settlement Footprint 2015, Scientific Data, 7, 1–14, https://springernature.figshare.com/articles/dataset/World_Settlement_Footprint_WSF_2015/10048412?backTo=/collections/Outlining_where_humans_live_-_The_World_Settlement_Footprint_2015/4712852, 2020.

Marshall, J. D., Nethery, E., and Brauer, M.: Within-urban variability in ambient air pollution: comparison of estimation methods, Atmospheric Environment, 42, 1359–1369, 2008.

Meyer, H. and Pebesma, E.: Predicting into unknown space? Estimating the area of applicability of spatial prediction models, Methods in Ecology and Evolution, 12, 1620–1633, 2021.

Meyer, H. and Pebesma, E.: Machine learning-based global maps of ecological variables and the challenge of assessing them, Nature Communications, 13, 1–4, 2022.

Minoura, H. and Ito, A.: Observation of the primary NO2 and NO oxidation near the trunk road in Tokyo, Atmospheric environment, 44, 23–29, 2010.

NASA: Measuring Vegetation Enhanced Vegetation Index (EVI), https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring_vegetation_4.php, 2017.

National Centers for Environmental Information: Global Summary of the Month (GSOM), Version 1, https://www.ncei.noaa.gov/access/search/data-search/global-summary-of-the-month?startDate=2017-01-01T00:00:00&endDate=2017-12-31T23:59:59&bbox=55.441,2.959,47.100,15.557&dataTypes=PRCP, 2017.

OpenAQ: Fighting air inequality through open data, 2017.

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

Patelli, L., Cameletti, M., Golini, N., and Ignaccolo, R.: A path in regression Random Forest looking for spatial dependence: a taxonomy and a systematic review, arXiv preprint arXiv:2303.04693, 2023.

Reid, C. E., Jerrett, M., Petersen, M. L., Pfister, G. G., Morefield, P. E., Tager, I. B., Raffuse, S. M., and Balmes, J. R.: Spatiotemporal prediction of fine particulate matter during the 2008 Northern California wildfires using machine learning, Environmental science & technology, 49, 3887–3896, 2015.

Ren, X., Mi, Z., and Georgopoulos, P. G.: 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, Environment international, 142, 105 827, 2020.

Rijkswaterstaat: Intensiteit Wegvakken, https://data.overheid.nl/dataset/28311-intensiteit-wegvakken--inweva--2017, 2017.

Rybarczyk, Y. and Zalakeviciute, R.: Machine learning approaches for outdoor air quality modelling: A systematic review, Applied Sciences, 8, 2570, 2018.

Shaddick, G., Salter, J. M., Peuch, V.-H., Ruggeri, G., Thomas, M. L., Mudu, P., Tarasova, O., Baklanov, A., and Gumy, S.: Global Air quality: an inter-disciplinary approach to exposure assessment for burden of disease analyses, Atmosphere, 12, 48, 2020.

Shapley, L. S.: Stochastic games, Proceedings of the national academy of sciences, 39, 1095–1100, 1953.

Tomtom: Tomtom Traffic Index - Ranking 2021, https://www.tomtom.com/en_gb/traffic-index/ranking/, 2021.

van Zoest, V., Osei, F. B., Hoek, G., and Stein, A.: Spatio-temporal regression kriging for modelling urban NO2 concentrations, International journal of geographical information science, 34, 851–865, 2020.

Wadoux, A. M.-C., Heuvelink, G. B., De Bruin, S., and Brus, D. J.: Spatial cross-validation is not the right way to evaluate map accuracy, Ecological Modelling, 457, 109 692, 2021.

Wang, A., Xu, J., Tu, R., Saleh, M., and Hatzopoulou, M.: Potential of machine learning for prediction of traffic related air pollution, Transportation Research Part D: Transport and Environment, 88, 102 599, 2020.

Weichenthal, S., Van Ryswyk, K., Goldstein, A., Bagg, S., Shekkarizfard, M., and Hatzopoulou, M.: A land use regression model for ambient ultrafine particles in Montreal, Canada: A comparison of linear regression and a machine learning approach, Environmental research, 146, 65–72, 2016.

Wikle, C. K., Berliner, L. M., and Cressie, N.: Hierarchical Bayesian space-time models, Environmental and ecological statistics, 5, 117–154, 1998.

Wikle, C. K., Zammit-Mangion, A., and Cressie, N.: Spatio-temporal statistics with R, CRC Press, 2019.

Wong, M. S., Zhu, R., Kwok, C. Y. T., Kwan, M.-P., Santi, P., Liu, C. H., Qin, K., Lee, K. H., Heo, J., Li, H., et al.: Association between NO2 concentrations and spatial configuration: a study of the impacts of COVID-19 lockdowns in 54 US cities, Environmental Research Letters, 16, 054 064, 2021.