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. Modeling and predicting air pollution over space have important implications in health studies, urban planning, and policy-making. Many statistical models have been developed to understand the relationships between geospatial data and air pollution sources. An important aspect often neglected is spatial heterogeneity; however, the relationships between geographically distributed variables and air pollutants commonly vary over space. This study aims to evaluate and compare various spatial and non-spatial statistical modeling (including machine learning) methods within different spatial groups. The spatial groups are defined by traffic- and population-related variables. Models are classified into local and global models. Local models use air pollution measurements from the Amsterdam area. Global models use ground station observations 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. The prediction accuracy is further increased in places far from roads for global models. Modeling of air pollution in different spatial groups shows that non-linear methods can have higher prediction accuracy than linear methods. The spatial prediction patterns of global models show that non-linear methods generally are less sensitive to extreme values compared to linear methods. Additionally, clusters of predicted air pollution differ between models within cities despite similar prediction accuracy. Also, the influence of predictors on NO_2 concentrations varies across different cities. Using the local dataset of our study, explicitly accounting for spatial autocorrelation in the universal and ordinary kriging models does not improve accuracy; however, analyzing prediction performance across spatial groups provides valuable insights. Comparing local and global prediction patterns reveals that local models capture regional clusters of high air pollution which are not detected by global models. These findings highlight that solely relying on overall prediction accuracy can be insufficient and potentially misleading, underscoring the importance of considering spatial variability and model performance within different spatial groups.

1 Introduction

Modeling and estimating NO_2 concentration levels is essential for a comprehensive understanding of air pollution, which plays a critical role in urban planning and policy-making to foster a healthy society. Air pollutants have been modeled across various spatial scales, from local to global. These models can be broadly classified into three categories: statistical models,

chemical transport models, and air dispersion models. Chemical transport models are typically used for large-scale air pollution modeling, while air dispersion models require detailed, spatially resolved emission data to capture small-scale variations in pollutants (Beelen et al., 2013).

In recent years, statistical modeling has gained popularity for high-resolution mapping at different spatial scales, driven by the increase in available predictors (e.g., GIS variables) and advancements in computational capabilities. Land Use Regression (LUR) is the most well-known statistical approach for air pollution modeling, using linear regression to capture the spatial variability of traffic-related air pollution in urban areas. Most LUR models rely on data from ground monitoring stations (Hoek et al., 2008; Wang et al., 2020). Geostatistical methods like kriging can further account for spatial correlations between observations. However, several studies have favored the simplicity of LUR, often concluding that it performs as well as or better than geostatistical methods (Hoek et al., 2008; Marshall et al., 2008; Beelen et al., 2013). Notably, these conclusions are typically based on prediction accuracy alone, without considering the models' ability to quantify uncertainty, provide scientific interpretations, or integrate known mechanisms (Lu et al., 2023). Specifically, many studies neglect optimal estimation of the covariance function and the specification of priors in geostatistical modeling.

While linear models are advantageous for their interpretability and ability to extrapolate, they may fall short in capturing the complex processes of air emission, dispersion, and deposition (Wang et al., 2020). As a result, data-driven, non-parametric models—commonly referred to as machine learning methods in air pollution mapping—have become increasingly popular. These models, such as tree-based algorithms, are better suited for capturing the non-linear relationships between pollutants and predictors (Weichenthal et al., 2016; Reid et al., 2015; Lu et al., 2020). For instance, Brokamp et al. (2017) compared Land Use Random Forest (LURF) models with LUR models for elemental components of $PM_{2.5}$ in Cincinnati, Ohio, and found that LURF models demonstrated lower prediction error variance across all elemental models when cross-validated. Similarly, Kerckhoffs et al. (2019) reported that machine learning algorithms, such as bagging and random forest, explained more variability in ultra-fine particle concentrations than multiple linear regression and regularized regression techniques. Ameer et al. (2019) advocated for random forest regression as the best technique for pollution prediction across varying datasets, locations, and characteristics, outperforming decision tree regression, multi-layer perceptron regression, and gradient boosting regression. Ren et al. (2020) also concluded that non-linear machine learning methods achieve higher accuracy than linear LUR, emphasizing the importance of careful hyperparameter tuning and robust data splitting and validation to ensure stable, reliable results. Chen et al. (2019) compared 16 algorithms for predicting annual average fine particle ($PM_{2.5}$) and nitrogen dioxide (NO_2) concentrations across Europe. They found that ensemble tree-based methods were particularly effective for $PM_{2.5}$, while NO_2 models showed similar R^2 values across different methods. Importantly, they reported a high correlation between the predicted values of various models, noting that the most influential predictors differed substantially between pollutants. For example, satellite observations and dispersion model estimates were key predictors for $PM_{2.5}$ concentrations, while NO_2 variability was primarily driven by traffic-related variables. The significant contribution of road traffic to NO_2 levels is further supported by Wong et al. (2021), who found that nitrogen emissions are particularly influenced by long-range transport from gasoline-fueled passenger cars.

In recent years, the use of statistical modeling for air pollution mapping has surged, resulting in numerous local and global pollution maps that are increasingly applied in urban and health studies. However, evaluating these models and maps remains challenging. One challenge is the scarcity of air pollution measurements. Another is the varying focus on spatial heterogeneity in air pollution. For example, He et al. (2022) acknowledge spatial heterogeneity in measurement stations by demonstrating that the probability density functions of concentrations (NO, NO_2, PM_{10}, $PM_{2.5}$) vary across different spatial categories (e.g., urban traffic, suburban/rural traffic, urban industrial, suburban/rural industrial, urban background, suburban background, rural background). However, their study does not model potential differences in prediction accuracy across these categories. A third challenge is that most current statistical approaches assess only overall accuracy, neglecting spatial variation (Hoek et al., 2008; Chen et al., 2019). Hoek et al. (2008) reported that LUR models typically explain 60-70% of the variation in NO_2, but this explained variation may be significantly lower near traffic. Chen et al. (2019) argued that many air pollution exposure studies fail to account for the characteristics of monitoring sites when performing cross-validation, potentially misrepresenting model results. They suggest evaluating models using pollution data from monitoring sites that reflect the application locations (Chen et al., 2019).

Finally, a consistent and coherent method for quantifying uncertainty in air pollution mapping is lacking. Shaddick et al. (2020) pointed out that uncertainty in air pollutant measurements is rarely discussed. This inadequate evaluation can lead to overlooked biases, especially since non-parametric machine learning methods often lack extrapolation capabilities. When predicted areas differ significantly in societal and environmental characteristics from training data, highly biased predictions may result, which are not adequately evaluated in many studies (Shaddick et al., 2020).

Given the growing number of modeling and prediction techniques and the potential for misrepresented prediction maps due to heterogeneity issues, this study aims to investigate: *To what extent can statistical models predict NO_2 concentrations using high-quality, high-temporal-resolution ground station measurements? How do the performance of these models and their spatial accuracy vary?* The study focuses on the Netherlands and Germany, using two datasets: the official national ground station measurements from both countries (referred to as the global dataset) (OpenAQ, 2017; EEA, 2021), and the more densely distributed ground station measurements from the Amsterdam area (referred to as the local dataset) (Gemeente Amsterdam, 2022). The global dataset includes 482 measurement stations covering 398,000 km^2 with a point density of 0.0012 points per km^2, while the local dataset includes 132 stations covering 196 km^2 with a point density of 0.591 points per km^2. The study aims to compare and understand model behaviors and prediction patterns across 1) the two datasets, 2) different spatial groups classified by proximity to traffic and population density, and 3) various statistical models, to evaluate the added value of non-linear machine learning models and geostatistical approaches.

2 Methodology

2.1 Data

The global and local datasets include the annual mean NO_2 concentrations (measured in $\mu g/m^3$) for the year 2017 (OpenAQ, 2017; EEA, 2021). Figure 1 presents the distribution of NO_2 concentrations at the global and local measurement stations. The

terms "global" and "local" are chosen to reflect the relative scale of the datasets with "global" representing a broader, cross-national dataset and "local" focusing specifically on Amsterdam. While the "global" dataset includes only two neighboring countries, this terminology emphasizes its wider scope compared to the local dataset. The global dataset comprises ground station measurements from Germany and the Netherlands, while the local dataset includes data from ground measurement stations specifically in the Amsterdam area.

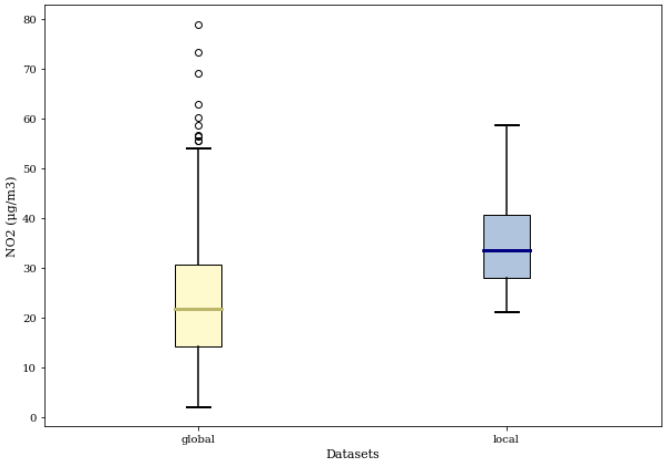


Figure 1. Distribution of NO_2 concentrations in the global (yellow) and local (blue) datasets.

The spatial distribution of NO_2 measurement stations is provided in the supplementary materials (Figure 1a, 1b). Urban areas generally have a higher density of measurement stations. This study focuses on the differences between global and local models, particularly in Amsterdam, while also considering the city's less densely populated areas to examine the urban impact on predicted NO_2 concentrations in the local models.

To evaluate whether prediction quality varies across areas with different spatial characteristics (e.g., high vs. low road density), the global and local datasets are divided into three spatial groups based on population density and traffic-oriented variables. Population data for 2015 from the Global Human Settlement Layer is used (JRC, 2015), and road length information is sourced from OpenStreetMap (2019). Descriptive statistics for the variables used to define spatial groups are presented in Table 1.

The three spatial groups are defined as follows:

1. **Urban:** Areas within 100 meters of road class 1 (highways) and 2 (primary roads) and with population density in the highest 25%; or areas where both road class 3 (local roads) values and population density are in the highest 25%.

2. **Suburban:** Areas within 100 meters of road class 1 and 2 with population density in the lowest 75%; or areas where road class 3 values are in the highest 25% and population density in the lowest 75%.

3. **Rural:** Areas further than 100 meters from road class 1 and 2; or areas where road class 3 values are in the lowest 75%.

Table 1. Descriptive statistics of variables determining spatial groups for the local and global datasets.

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

This classification resulted in 85 observations being labeled as "urban", 138 as "suburban", and 259 as "rural", totaling 482 observations in the global dataset. Given the smaller sample size of the local dataset, the threshold for defining "urban" was adjusted from the 75th percentile to the 50th percentile having a converging effect on the different group sizes. Moreover, the increase in samples classified as "urban" is encouraged as a result of the relatively high heterogeneity in this group. The local dataset consists of 56 observations that are classified as "urban," 46 as "suburban," and 30 as "rural."

Although this adjustment introduces some inconsistency between the global and local definitions of "urban," it addresses the challenge of unequal distribution of instances across groups in the local dataset, which could introduce bias into the statistical learning models. The threshold adjustment represents an initial step to mitigate such effects by ensuring a more balanced representation of spatial characteristics within the local model. Supplementary Figures 2 and 3 show the spatial distribution of observations between these groups for both datasets, while Supplementary Figures 4 and 5 show the measured NO_2 concentrations per station.

Spatial predictors

We utilized a set of variables derived from Lu et al. (2020), including data on industrial areas, road lengths, population density, Earth night lights, wind speed, temperature, elevation, Tropomi level 3 NO_2, and global radiation. A complete list of these variables is available in the supplementary material (Table 1). Precipitation data was sourced from weather stations (National Centers for Environmental Information, 2017) and interpolated using ordinary kriging to cover the NO_2 measurement stations. Kriging parameters are detailed in the supplementary material.

Building density was obtained from the "World Settlement Layer 2015" dataset available on Figshare (Marconcini et al., 2020). In line with previous studies (Beelen et al., 2013; Kheirbek et al., 2014), we considered various buffer sizes (100m, 500m, 1000m) around measurement stations to account for spatial proximity effects, especially in densely populated urban areas. NDVI values were obtained from NASA (NASA, 2017).

Traffic volume data was sourced from the "Nationaal Dataportaal Wegverkeer" (NDW) in the Netherlands (Rijkswaterstaat, 2017) and "Bundesanstalt für Strassenwesen" (BAST) in Germany (Bundesanstalt für Strassenwesen, 2017). This data, generated by automatic counting stations, is expressed as average hourly traffic over 2017, with buffer sizes of 25m, 50m, 100m, 400m, and 800m. The formula for calculating average hourly traffic is provided in the supplementary material.

2.2 Modeling NO_2 globally and locally

2.2.1 Ensemble trees

The global models use two types of statistical learning methods. The first group consists of ensemble tree-based approaches, including random forest and Extreme Gradient Boosting (XGBoost). Hyperparameters are tuned based on cross-validation error. For the random forest model, the number of estimators is set to 1000, with a minimum samples split of 10, minimum samples per leaf of 5, maximum features per tree of 4, and a maximum depth of 10. The XGBoost model uses 50,000 estimators, with a reg_alpha of 2, reg_lambda of 0, max_depth of 5, and a learning rate of 0.0005. Additionally, the gamma for the XGBoost model is set to 5. Further details can be found in the supplementary material, section parameters. Additionally, the Light Gradient Boosting (LightGBM) model was tested but did not yield significantly different results compared to XGBoost. The results of LightGBM analyses are shown in the supplementary material (Figure 6a-c, 7).

2.2.2 Multiple Linear Regression

Key variables identified by the random forest model are used as predictors in Multiple Linear Regression (MLR). Regularization techniques such as Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge regression are employed to prevent overfitting. LASSO differs from Ridge in that it uses the sum of the absolute values of the coefficients as a penalty, allowing some coefficients to be exactly zero, thus enabling feature selection (Ren et al., 2020). The alpha for both LASSO and Ridge models is tuned to 0.1, optimizing for the lowest Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and highest R^2 among options ranging from 0.1 to 1 in increments of 0.1. Detailed parameters and mathematical formulations for linear regression, error terms, Ridge regression, and LASSO regression are given in the supplementary material (section Parameters and section Equations).

2.2.3 Mixed-Effects Model and Kriging

The performance of random forest, XGBoost (Supplementary Table 2), LASSO and Ridge (Supplementary Table 3) are unsatisfactory for the local dataset. Spatial modeling approaches including mixed-effects modeling and kriging are applied.

Mixed-effects models could capture hierarchical or grouped structures. In our study, the fixed effects correspond to the most influential predictors, such as population density, road length, and other traffic-related variables, which are assumed to have consistent effects across the entire study region. The random effects capture the spatial trends specific to different geographic regions, such as urban, suburban, and rural areas. For instance, local topography, vegetation, or specific traffic patterns in a

region can create unique spatial trends. These spatial groups represent the variation in NO_2 concentrations due to local environmental factors and are modeled as random effects, allowing the model to account for spatial autocorrelation within regions.

This modeling approach is suitable because the spatial distribution of pollutants such as NO_2 is not random and tends to show clusters or gradients in traffic, land use, and population density. By modeling the spatial context as random effects, we capture these spatial dependencies and potentially improve the accuracy of predictions in different areas (Mullen and Birkeland, 2008; Lee et al., 2020).

Kriging is the Gaussian process developed in Geoscience. The residuals of a linear regression model are considered as random variables and a covariance function is modeled. It estimates NO_2 concentrations based on the spatial relationships of the observations as NO_2 is a relatively smooth spatial process.

In this study, ordinary and universal kriging are applied. Ordinary kriging assumes that the mean of the variable being predicted is constant but unknown over the entire study area. It focuses on modeling spatially correlated random variation without assuming any global trend. Universal kriging assumes that the variable being predicted has a deterministic trend (e.g. linear or polynomial). The `automap` package in R (Hiemstra et al., 2008) is used to initialize the covariance parameters and to perform the kriging interpolation. Two separate models are created, one that incorporates the spatial groups (urban, suburban, rural) and one that does not. These models help to compare the impact of spatial context on the accuracy of the predictions (Idir et al., 2021; Khan et al., 2023).

In total, ten models are fit and compared: four using the global dataset and six using the local dataset, enabling a comprehensive evaluation of model performance across different geographical scales. The equations for kriging and the linear model are provided in the supplementary materials, under the "Parameters" and "Equations" sections.

2.3 Feature selection

Feature selection for global models is initially based on Shapley values (Shapley, 1953). While the Variance Inflation Factor (VIF) is effective for detecting multicollinearity, it does not consider feature importance or interactions. Shapley values are preferred for their comprehensive evaluation, which aligns with our goal of enhancing model performance and interpretability. VIF results are available in the supplementary materials (Tables 4 and 5). Feature selection aims to remove irrelevant or highly correlated predictors that could generate unstable estimates (Araki et al., 2018).

Shapley values are calculated for each feature (i.e. predictor) based on its contribution ϕ_j to the prediction of NO_2 concentration levels, compared to the average prediction across the dataset (Shapley, 1953). The contribution of a feature is determined by comparing the difference in the response variable when the feature is present versus when it is absent (i.e., marginal contribution) (Algaba et al., 2019; Shapley, 1953). The formula for calculating Shapley values can be found in the supplementary materials.

In this study, feature selection is guided by the out-of-sample performance in a 10-fold repeated random sampling validation, where Shapley values are calculated in each iteration of the random forest models. Predictors are ranked based on the median Shapley value across all iterations. The relative positions of each predictor using the median-based approach are illustrated in

Supplementary Figures 8 and 9, with the Shapley ranking of a single fold shown in Figure 2. The most influential predictors for the global models include nightlight intensity (450m and 3150m buffers), population density (1000m and 3000m buffers), road class (class 2 within 25m and class 3 within 300m and 3000m buffers), trop mean filter 2018, building density (100m buffer), NDVI, and traffic buffers (25m and 50m buffers). Descriptive statistics of the most influential predictors for the global models are in table 2.

A random forest algorithm is applied iteratively to determine the optimal number of predictors, starting with the two most influential predictors and extending to the thirty most influential features. The RMSE and R^2 metrics are used to evaluate the optimal number of predictors. The number of predictor variables and their corresponding evaluation scores (R^2, RMSE) are shown in Figures 3a and 3b. In particular, prediction accuracy improves significantly when considering at least twelve predictors, although the improvement is marginal beyond this number.

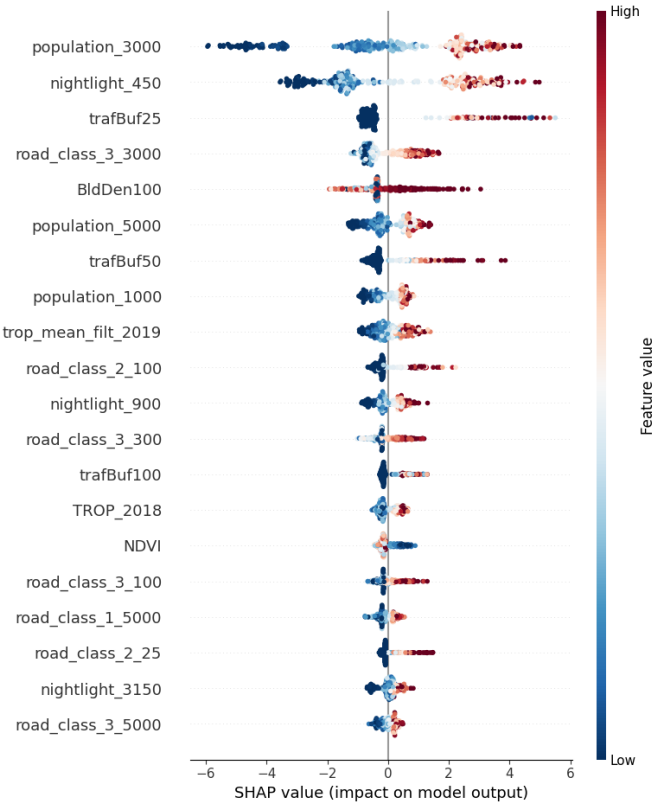


Figure 2. Shapley ranking of a single fold using the global dataset.

Table 2. Descriptive statistics for global predictors. bldden100 = Built area 100m buffer, ndvi = Normalized Difference Vegetation Index, nightlight_3150 = Nightlight 3150m buffer, nightlight_450 = Nightlight 450m buffer, population_1000 = Population in 1km grid, population_3000 = Population in 3km grid, road_class_2_25 = Total length of primary roads 25m buffer, road_class_3_300 = Total length of local roads 300m buffer, road_class_3_3000 = Total length of local roads 3000m buffer, trafbuf25 = Traffic count 25m buffer, trafbuf50 = Traffic count 50m buffer, trop_mean_filt_2018 = TROPOMI 2018 mean vertical column density

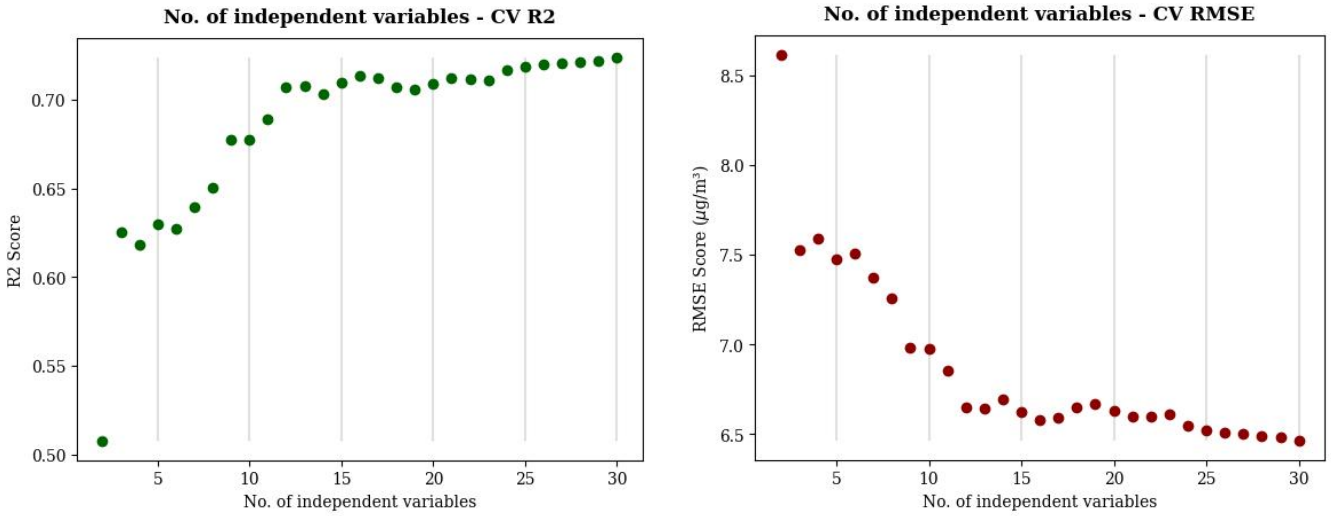
Variable	Unit	25th	50th	75th	Mean	Median	Max	Min
bldden100	%	0.4	0.88	0.99	0.68	0.88	1	0
ndvi	(x10,000)	2285.75	3153.5	4199.25	3331.37	3153.5	7775	747
nightlight_3150	$\mathrm{Wcm^{-2}sr^{-1}}$	2.9	8.2	16.51	11.04	8.2	101	0
nightlight_450	$\mathrm{Wcm^{-2}sr^{-1}}$	4.62	14.01	22.4	15.34	14.01	84.32	0
population_1000	count	2204.52	5945.54	9036.76	6154.49	5945.54	20300.89	0
population_3000	count	11452.7	33821.94	61824.05	41489.44	33821.94	165271.38	0
road_class_2_25	m	0	0	0	14.57	0	164.93	0
road_class_3_300	m	930.85	2447.69	3403.39	2314.03	2447.69	7239.33	0
road_class_3_3000	m	70823.23	134524.3	193091.82	136692.07	134524.3	444277.31	0
trafbuf25	count	0	0	0	128.7	0	5112.96	0
trafbuf50	count	0	0	0	146.89	0	5112.96	0
trop_mean_filt_2018	$\mathrm{mol/cm^2}$	0	0	0	0	0	0	0

* The values for *trop mean filt 2018* are very small, on the order of 10^{-5}.

Due to the random forest model's poor performance across all local station measurements (Supplementary Figures 10a-c) and per spatial group (Supplementary Table 2), the random forest algorithm is deemed unsuitable for identifying the number of variables for the local models. Instead, best subset regression is used for variable selection in local models. This approach tests all possible combinations of predictor variables (Kassambara, 2018), with a maximum of 30 predictors considered. The statistical criteria include adjusted R^2, Mallows CP, and Bayesian Information Criteria (BIC) scores. As a result, nine features are identified for the local models. The most influential predictors for the local models include nightlight intensity (450m and 4950m buffer), population density (3000m buffer), road class (class 1 within 5000m; class 2 within 1000m and 5000m buffers; road class 3 within 100m and 300m buffers), and traffic buffer (50m buffer). Descriptive statistics of the most influential predictors for the local models are in table 3.

2.4 Model comparison

In global modeling, comparisons are made among tree-based models—random forest and XGBoost—and linear models—LASSO and Ridge. For local modeling, we compare linear models, mixed-effect models, and kriging models. Each model is evaluated based on R^2, RMSE, and MAE, which are standard metrics in the field (Rybarczyk and Zalakeviciute, 2018; Ameer et al., 2019; Chang et al., 2020). For model evaluation, leave-one-out cross-validation (LOOCV) is employed for local models, while



(a) Number of features and corresponding R^2 score

(b) Number of features and corresponding RMSE score

Figure 3. Out-of-sample performance in ten-fold repeated random sampling validation: number of features and corresponding model performance (global).

Table 3. Descriptive statistics for local predictors. nightlight_450 = Nightlight 450m buffer, nightlight_4950 = Nightlight 4950m buffer, population_3000 = Population in 3km grid, road_class_1_5000 = Total length of highway 5000m buffer, road_class_2_1000 =Total length of primary roads 1000m buffer, road_class_2_5000 =Total length of primary roads 5000m buffer, road_class_3_100 =Total length of local roads 100m buffer, road_class_3_300 =Total length of local roads 300m buffer, trafbuf50 = Traffic count 50m buffer

Variable	Unit	25th	50th	75th	Mean	Median	Max	Min
nightlight_450	$\mathrm{Wcm^{-2}sr^{-1}}$	31.24	38.97	50.3	42.03	38.97	98.39	3.96
nightlight_4950	$\mathrm{Wcm^{-2}sr^{-1}}$	28.82	32.91	33.99	30.15	32.91	35.97	5.11
population_3000	count	106347.12	121186.11	128723.57	111157.01	121186.11	137546.05	20097.26
road_class_1_5000	m	83586.9	88910.18	96428.33	88821.13	88910.18	137238.88	24270.47
road_class_2_1000	m	2367.6	4032.19	5348.42	4018.63	4032.19	9596.1	0
road_class_2_5000	m	54638.17	61129.2	64151.61	58553.23	61129.2	71428.22	24435.52
road_class_3_100	m	182.82	359.38	548.96	374.29	359.38	1057.03	0
road_class_3_300	m	1774.06	2574.59	3433.76	2713.63	2574.59	6283.23	0
trafbuf50	count	0	0	132.67	294.66	0	3976.16	0

a 75/25 train-test split is used for global models. Additionally, the prediction patterns of the local and global models are analyzed. To benchmark the model performance, a mobile NO_2 map of the study area (Kerckhoffs et al., 2019; Yuan et al., 2023) is used for comparison. This map provides detailed spatial information collected by two Google Street View cars continuously measuring NO_2 at a frequency of 1 Hz in Amsterdam from May 25, 2019, to March 15, 2020 (stopped due to the COVID-19

lockdown policy). We acknowledge that the temporal resolution of this benchmark data differs from the coarser temporal scales used in our models. The Kerckhoffs et al. (2019) data represents measurements over specific, limited time periods, while our models address predictions over broader temporal spans. Despite this temporal inconsistency, the detailed spatial granularity of the Kerckhoffs et al. map provides valuable insights and remains an appropriate standard for assessing spatial prediction quality. Table 4 provides an overview of the global and local models, along with selected predictors and evaluation methods. The global models are applied to areas with varying demographic characteristics, including two large cities with populations exceeding 700,000 (Amsterdam and Hamburg), a mid-sized city with around 350,000 inhabitants (Utrecht), and a small city with approximately 70,000 inhabitants (Bayreuth). The resolution of the analysis is 100m, with TIF files of predictors converted into 100m grid cells for these regions. The influential predictor information (for global models, see Tables 2 and 4; for local models, see Tables 3 and 4) is recalculated at 100m resolution for the extent of the aforementioned regions. Thereafter 100m by 100m grid cells containing predictor information are used to predict NO_2 values for the respective 100m grids, based on the trained local and global models. The 100m grid resolution is consistently applied in the predictions for both local and global models. Local model predictions are applied exclusively to Amsterdam. Table 5 summarizes the complexity of the models and how spatial components are accounted for.

Table 4. Global and local models defined by selected predictors, models evaluated, and how models are evaluated.

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 LASSO Ridge	cross validation over the entire area cross validation over different land types comparing with Kerckhoffs et al. (2019)
Local model	nightlight_4950 nightlight_450 road_class_3_100 trafbuf50 road_class_3_300 road_class_2_1000 road_class_2_5000 population_3000 road_class_1_5000	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 5. Features of the global and local models regarding model complexity and how the spatial component is considered.

Model	Model complexity	Accounting for the spatial component
Linear regression	No regularization	Classifying between land types and fitting a model in each class.
LASSO	L2 regularization	Not explicitly
Ridge	L1 regularization	Not explicitly
Mixed-effect	No regularization	Classifying between land types and including the classes as a random variable.
Kriging	No regularization	Covariance matrix based on Euclidean distance (second-order stationarity); Classifying between land types and fitting a model in each class.
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

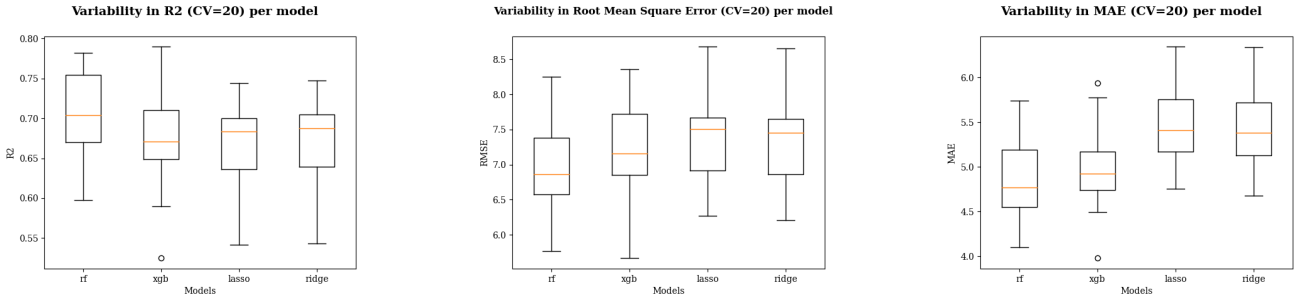
3 Results

3.1 Models

3.1.1 Global models

Evaluations of the different linear and non-linear models were carried out using repeated random sampling validation, performed 20 times. In each iteration, 75% of the data was used for training and the remaining 25% for testing. This approach allowed us to evaluate the variance and median statistics for each model in terms of R^2, MAE, and RMSE (Figure 4a, Figure 4b, and Figure 4c). The repeated sampling provided stable estimates.

When comparing out-of-sample performances via 20-fold repeated random sampling validation, the linear models (i.e., LASSO and Ridge) exhibited performances similar to those of the non-linear models, particularly in terms of R^2 and RMSE. Among the models, the random forest consistently outperformed others, with the highest median R^2, lowest RMSE, and lowest MAE. The robustness of the random forest model is further emphasized by its minimal standard deviation in R^2 and RMSE (Figure 4a and Figure 4b).



(a) R^2 (b) RMSE (c) MAE

Figure 4. Out-of-sample performances evaluated using 20-fold repeated random sampling validation: (a) R^2, (b) RMSE, and (c) MAE. Upper and lower quartiles indicate variability. RF = random forest, XGB = XGBoost.

Accounting for spatial information

We further investigated the influence of spatial heterogeneity by comparing model performance across different spatial groups using the global dataset. Descriptive statistics for NO_2 concentrations in each spatial group reveal distinct differences (Table 6).

Table 6. Descriptive statistics of NO_2 concentrations for each spatial group (in $\mu g/m^3$).

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
Suburban	138	27.601	9.769	7.872	19.876	26.876	34.407	56.706
Rural	259	16.653	8.341	2.122	10.331	15.892	22.518	48.887

Table 7 details the performance metrics (R^2, RMSE, MAE) for each spatial group. Non-linear models outperformed linear ones in suburban and rural areas, while performances were less distinguishable in urban areas, likely due to the smaller sample size. Ensemble tree-based methods, such as random forest, showed lower accuracy in urban areas, possibly due to the limited and heterogeneous nature of the data in this group.

			Urban			Suburban			Rural		
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
	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 7. Model performance per spatial group (CV = 20). RMSE and MAE are represented in NO_2 ($\mu g/m^3$).

Spatial prediction patterns

Figure 5 presents the spatial predictions of NO_2 concentrations across the Amsterdam area for each model. Panels (a) and (b) depict the predictions from non-linear models, while panels (c) and (d) illustrate the results from linear models. Generally, linear models exhibit a higher tendency for overfitting, as their prediction maps are more influenced by extreme values (i.e., concentrations below 15 $\mu g/m^3$ or above 50 $\mu g/m^3$) compared to the non-linear techniques. Interestingly, the linear models identify a significant NO_2 hotspot in the southwestern part of the study area, which is not captured by the non-linear models. Across all models, however, elevated pollution levels are consistently observed along major roads and in some urban areas, such as Haarlem (see Supplementary Figure 11).

Figures 6 show the spatial patterns of predicted NO_2 concentrations for Hamburg (a, b), Utrecht (c, d), and Bayreuth (e, f) using the random forest and Ridge Regression models. Predictions from other models (XGBoost, LASSO, LightGBM) for these cities, including both zoomed-in and zoomed-out views, are provided in Supplementary Figures 12a-c, 13a-c, 14a-c, and 15a-e.

Comparing the prediction maps of these cities reveals noticeable differences in spatial patterns. A key finding is that in Hamburg, the highest air pollution levels are concentrated around major roads, while in Utrecht, the urban center exhibits the highest NO_2 concentrations. This correlation between major roads and elevated air pollution in Hamburg can be reasonably explained by the city's high traffic congestion, as it ranks 69th among the most congested cities globally (Tomtom, 2021). Interestingly, there are also spatial differences in the predicted NO_2 concentrations along highways between the random forest and Ridge models. For instance, in Hamburg, the Ridge model predicts high NO_2 levels along highways in the southeastern and western parts of the city, whereas the random forest model provides a more nuanced spatial identification of these areas. The

random forest predictions highlight more pronounced air pollution along roads in the central and northern parts of Hamburg, compared to the Ridge model.

Furthermore, the magnitude of high pollution levels related to major roads is significantly greater in Hamburg than in Utrecht and Bayreuth. Nevertheless, the relationship between road presence and higher air pollution levels is evident in both Utrecht and Bayreuth, particularly in the predictions from the Ridge model. In Utrecht, the urban center is more prominently identified as a high NO_2 concentration area compared to Hamburg and Bayreuth. Additionally, the Ridge model for Utrecht shows more clusters of elevated NO_2 levels in the periphery, whereas the random forest model predicts a more scattered distribution of NO_2 concentrations in the urban center, similar to the pattern observed in the Amsterdam area.

Bayreuth, on the other hand, is characterized by moderate pollution levels, with very low NO_2 concentrations (<15 $\mu g/m^3$) in the rural areas surrounding the city. However, some clusters of higher NO_2 levels exceeding the 15 $\mu g/m^3$ benchmark are observed in the vicinity of other villages, suggesting that population or building density may influence air pollution levels in these areas (see also Supplementary Figures 15a-e). Supplementary Figure 16 provides a distribution of predicted NO_2 concentrations for each global model and location.

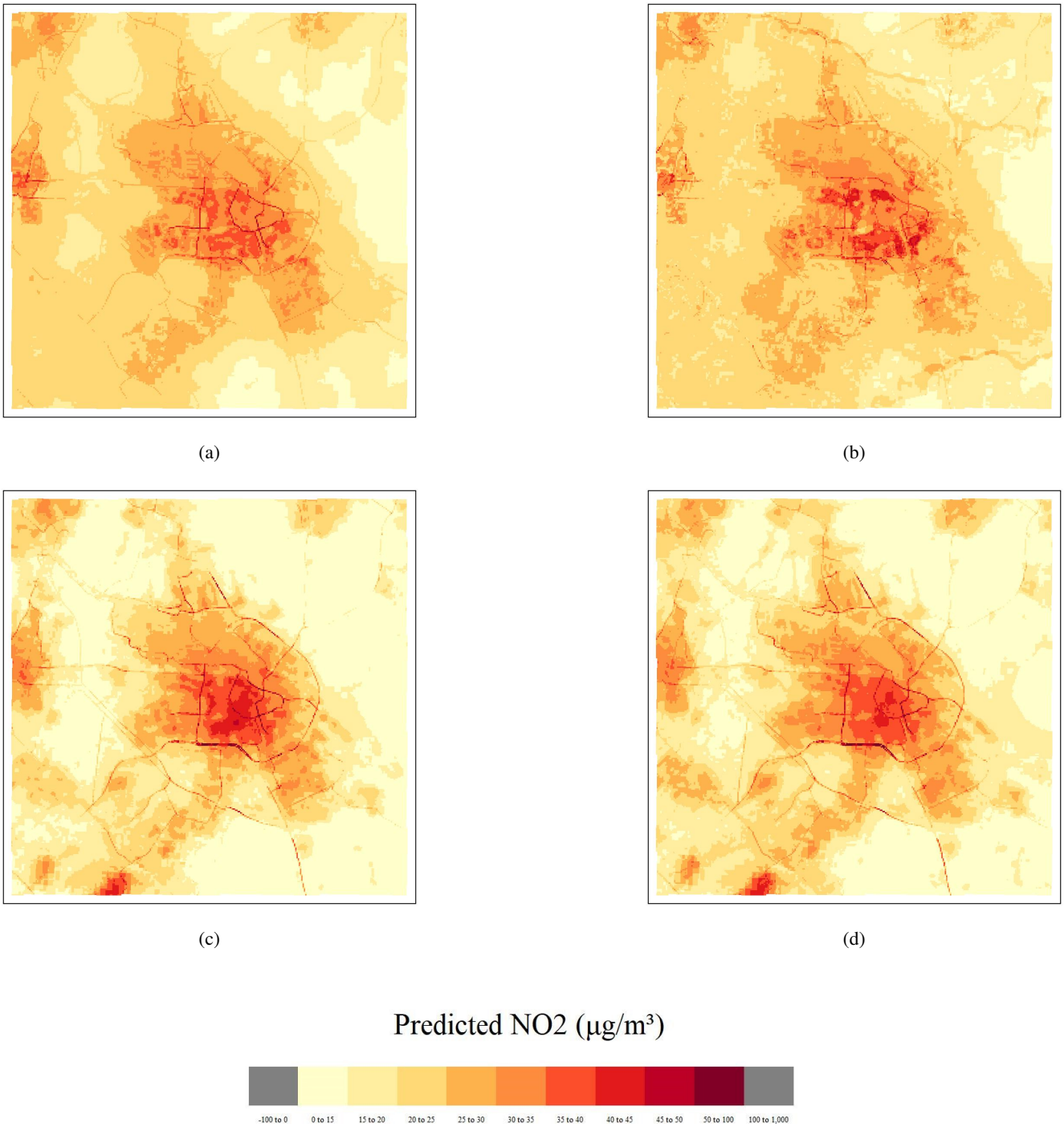


Figure 5. Spatial patterns of predicted NO_2 (100m), measured in μg/m3, per model for Amsterdam - non-linear models (top): (a) = random forest, (b) = XGBoost; linear models (bottom): (c) = LASSO, (d) = Ridge. Extent = 30km x 30km

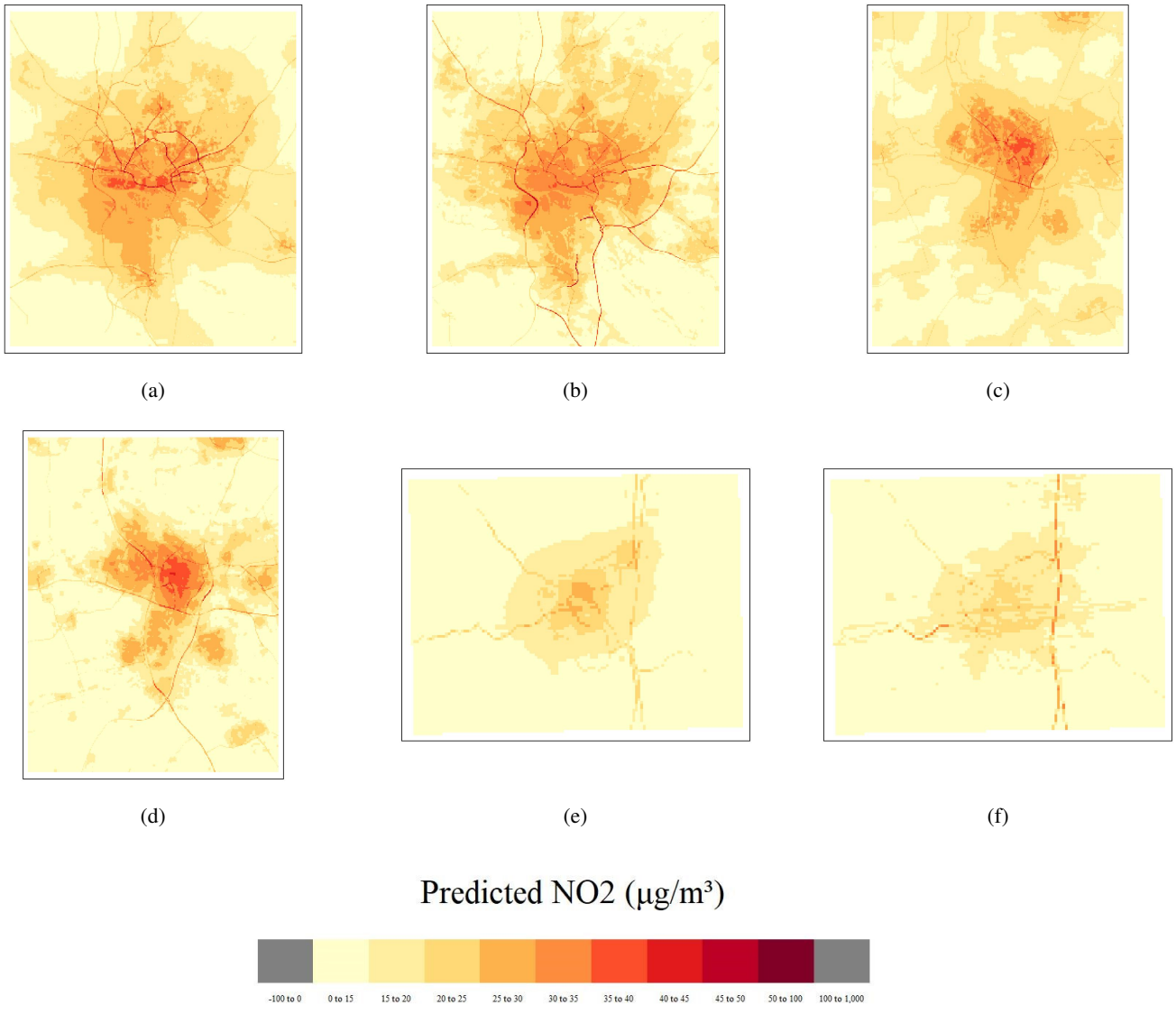


Figure 6. Spatial patterns of predicted NO_2 (100m), measured in μg/m3, per model for Hamburg (extent = 30km x 30km), Utrecht (extent = 25km x 25km) and Bayreuth (extent = 10km x 10km) - top: from left to right, random forest (Hamburg), Ridge (Hamburg), random forest (Utrecht); bottom: from left to right, Ridge (Utrecht), random forest (Bayreuth), Ridge (Bayreuth)

3.1.2 Local models

The performance of the local models was assessed using R^2, RMSE, and MAE metrics. Table 8 summarizes the performance of the linear model, mixed-effects model, ordinary kriging model, and universal kriging model, all evaluated using leave-one-out cross-validation. Among these, the ordinary kriging model exhibits the poorest performance. Figure 7 illustrates the spatial prediction patterns for each model. Notably, the universal kriging model outperforms the ordinary kriging model significantly. However, the simple linear model surpasses the universal kriging method in terms of prediction accuracy. Incorporating spatial groups as random effects in the mixed-effects model leads to a higher R^2, and lower RMSE and MAE, indicating improved model performance.

Table 8. Model Performance Using Leave-One-Out Cross-Validation

	R^2	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)
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
universal kriging (model + kriged residuals)	0.277	7.749	6.097

Table 9 provides model performance metrics for each spatial group, again using leave-one-out cross-validation. Consistent with the global model results, local models trained on urban observations tend to perform poorly. This poor performance is likely caused by an imbalance between the relatively few number of samples and the relatively high heterogeneity. This imbalance may hinder the models' ability to capture the variability within urban areas, contributing to their poorer performance in this group. Interestingly, proximity to roads does not necessarily correlate with model performance, as the suburban group exhibits a higher R^2 than the rural group. Unlike global models, which perform best in rural areas, local models perform best in suburban areas. This difference may arise because observations in rural areas within the local dataset are more similar to those in urban and suburban areas than in the global dataset, due to a more uniform distribution of predictor values.

Table 9. Model Performance Per Spatial Group (CV = Leave-One-Out Cross-Validation). RMSE and MAE in $\mu g/m^3$

	Urban			Suburban			Rural		
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	6.800	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
universal kriging *(model + kriged residuals)*	0.161	8.068	6.270	0.487	6.938	5.174	0.037	7.190	8.299

Spatial prediction patterns

Figure 7 displays the predicted NO_2 patterns based on the local dataset. The prediction map for the linear model (a) is quite similar to those for the mixed-effects (c) and universal kriging (e) models, with all identifying a high NO_2 concentration cluster in the northwestern part of Amsterdam. Further analysis suggests that this cluster is likely influenced by the predictor "road class 2 5000" (i.e., primary roads within 5000m), as this predictor exhibits a similar cluster in the same location (see Supplementary Figures 17, 18a-i).

The two models that account for spatial groups before the modeling process (mixed-effects and universal kriging) display comparable patterns where the influence of roads is evident, either through the predictors themselves or the spatial groupings (see also Supplementary Figure 19). The relatively low NO_2 values along roads in the outer Amsterdam area can be attributed to the spatial grouping divisions. High standard deviations in predictor values within a specific spatial group can affect that group's NO_2 predictions, potentially leading to overestimation or underestimation in certain areas.

The high NO_2 values along roads are primarily associated with the suburban spatial group, where observations are located within 100 meters of roads. Compared to the rural group, the data distribution for each predictor in the suburban group is substantially different, leading to distinct learning patterns that explain the relatively high prediction values along roads (see Supplementary Figures 20a-i). In some instances, negative predicted values are observed, albeit rarely. These may result from discrepancies in feature characteristics between the training and testing datasets.

Comparing local prediction patterns to global prediction patterns reveals that the local models identify a cluster of high air pollution in the northwestern part of Amsterdam that the global models do not detect. This discrepancy could be due to differences in the spatial distribution of NO_2 values between the local and global datasets, leading to distinct learning patterns in the respective models (Figure 1). Moreover, Figures 5 and 7 underscore the challenge of comparing spatial variations between global and local models, given their differing algorithms. Local models, with their focus on specific spatial groupings and detailed predictors, capture regional clusters that global models may overlook or underrepresent due to their broader scope.

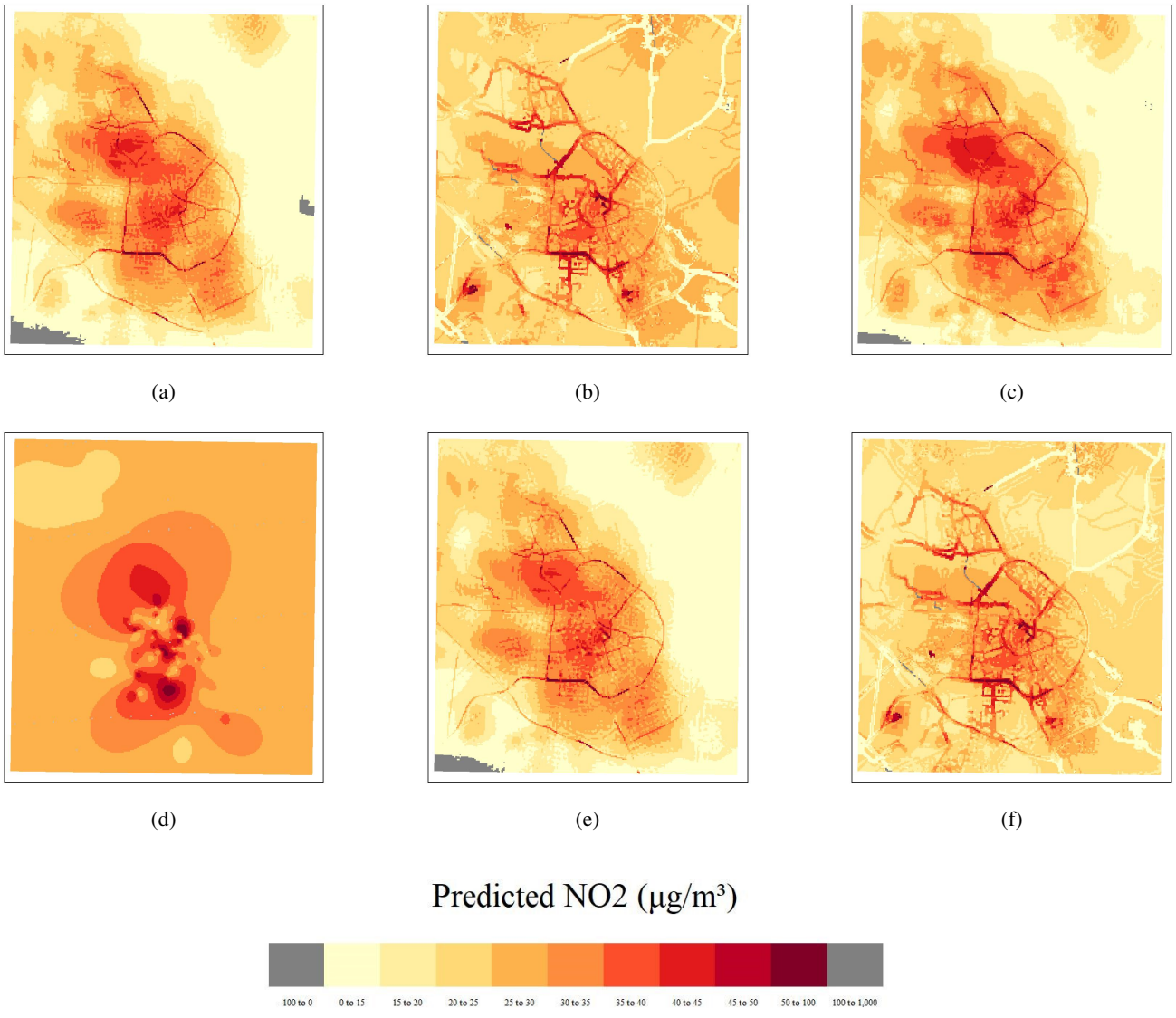


Figure 7. Spatial patterns of predicted NO_2 (μg/m3) at 100m resolution based on the local dataset - top: left = linear model, middle = linear model separating for spatial groups, right = mixed-effects model; bottom: left = ordinary kriging, middle = universal kriging, right = universal kriging separating for spatial groups

Model comparison

Figure 8 shows the correlation in predicted NO_2 values for the local and global models, as well as the mobile NO_2 map from Kerckhoffs et al. (2019) (referred to as the open NO_2 dataset), which was used as a benchmark (Supplementary Figure 22). To improve the clarity of the correlations between the models and the open NO_2 dataset, we addressed some extreme prediction values. These outliers were removed to prevent them from skewing the analysis and to provide a more accurate representation of the correlations. We selected a manual threshold of 85 as the upper bound, based on the maximum value observed across the ten models (excluding the two where outlier detection was applied first). The lower bound was set at 0. The correlation matrix with these extreme predictions filtered out (including LightGBM results) is shown in Supplementary Figure 23. The global models are highly correlated, with the LASSO model being the least correlated with other global models. The correlations between the ordinary kriging model and other models are low, which is expected as the covariance function has a small length scale. Another reason for this discrepancy is kriging's stationary assumption, which can lead to different results compared to models that do not rely on this assumption. When comparing the models with the open NO_2 dataset, local models generally show more similarity than global models. This is not surprising as the local model dataset is also from Amsterdam. Table 10 shows the residuals per global and local model. The ridge model emerged as the most accurate with the lowest mean residual (0.31), indicating it closely matched actual open NO_2 dataset values. Conversely, the LASSO model, despite its high internal correlation, had relatively higher residuals and showed less similarity in prediction patterns compared to other global models. XGBoost also performed well but with slightly higher residuals than the Ridge model. In contrast, the local linear models, mixed-effects model, ordinary kriging, and universal kriging generally displayed higher residuals, with ordinary kriging having the largest mean residual (4.71). This suggests that local models had greater prediction errors compared to global models. A spatial comparison of the predicted NO_2 concentration values between the open NO_2 dataset and the global and local models are shown in Supplementary Figures 24a-e and 25a-f respectively. A spatial comparison of the global and local model predictions with the measurement station data can be found in Supplementary Figures 26 and 27.

Table 10. Residual statistics for the difference between model predictions and open NO_2 dataset.

Model	Type	Mean	Median	SD	Min	Max
Random forest	Global	0.68	2.77	8.48	-54.54	17.98
LASSO	Global	1.24	2.50	9.03	-53.92	25.00
Ridge	Global	0.31	1.55	8.82	-53.70	25.06
XGBoost	Global	0.67	2.43	8.98	-57.94	24.48
Linear	Local	1.87	3.56	8.61	-55.16	28.17
Linear spatial groups	Local	2.25	3.09	15.22	-58.21	384.63
Mixed-effects model	Local	2.51	4.10	8.54	-53.75	26.70
Universal kriging	Local	1.83	3.46	8.30	-54.58	29.08
Universal kriging spatial groups	Local	1.99	2.76	14.56	-56.75	369.05
Ordinary kriging	Local	4.71	6.64	9.57	-57.21	30.71

Figure 8. Comparing model predictions whereby the numbers equal the Pearson correlation coefficient. RF: Random Forest, 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

Several studies have applied statistical modeling to ground station measurements and geospatial predictors for NO_2 mapping, but the impact of spatial heterogeneity has often been overlooked. In this study, we address this gap by comparing spatial and non-spatial modeling techniques across different spatial scales. Below, we discuss the key findings and provide our perspectives.

Relationship between predictors and other pollutants

For both global and local datasets, traffic and population density emerge as the most influential predictors, aligning with the findings of Beelen et al. (2013), which emphasize the importance of these variables for improving prediction accuracy. The strong influence of traffic on NO_2 concentrations also supports the conclusions of Lu et al. (2020) and Chen et al. (2019). However, since sources of different pollutants vary (Chen et al., 2019), the modeling results for NO_2 may not be directly applicable to other pollutants.

Accounting for spatial groups

Meyer and Pebesma (2021, 2022) argue that the growing popularity of global models, due to their ability to capture both linear and non-linear relationships, may lead to misinformation. Although global models can make accurate predictions in regions where the predictor variables are well-represented in the training data, their performance may degrade in areas with predictor values that deviate significantly from the training range, highlighting the risk of spatial bias in predictions.

In this study, the differences between linear and non-linear techniques are minimal when applied to the global dataset. Although the random forest model generally performs best (highest R^2, lowest MAE), the R^2 of the RIDGE model is higher than that of the XGBoost model. However, when accounting for spatial groups—urban, suburban, and rural—the differences in model performance between linear and non-linear techniques become more pronounced, with non-linear models generally outperforming linear models, particularly in rural areas where data are more homogeneous.

Although various cross-validation methods are available, with some researchers advocating for spatial cross-validation to better capture autocorrelation, we opted for random bootstrap cross-validation. According to Wadoux et al. (2021), standard cross-validation (i.e. ignoring autocorrelation) results in less bias than spatial cross-validation. They also argue that spatial cross-validation methods lack theoretical underpinning and should not be used for map assessment. Standard cross-validation is sufficient for clustered data scenarios (Wadoux et al., 2021; Lu et al., 2023).

A significant limitation of the study setup is the fact that the most heterogeneous group (urban) is the least represented in terms of number of data points, at least for the global dataset. In urban areas, the more heterogeneous nature of the data reduces the performance gap between linear and non-linear techniques, with both performing poorly. This poor prediction accuracy in urban areas is concerning, as the impact of air pollution is often more severe in these regions due to proximity to traffic-heavy roads and industrial areas (He et al., 2022). Although spatial grouping improves predictive reliability, it can

lead to counterintuitive patterns, such as lower predicted NO_2 concentrations along roads compared to surrounding rural areas. In the local dataset, the threshold for defining "urban" areas was adjusted from the upper 75% quartile (0.75) to the median (0.5). This adjustment was necessary due to the limited sample size, which required a broader definition to ensure sufficient data coverage for urban areas. However, this change also resulted in a less stringent definition of "urban," potentially including areas with lower population densities. While this adjustment expands the number of training samples available for the most heterogeneous group (urban), it introduces a limitation by diluting the urban group and affecting the comparability of results. This trade-off underscores the challenges of balancing data representation with statistical robustness in spatial analyses.

Global and local predictions

In comparing global and local models, each approach has distinct strengths and limitations. Local models, tailored to specific spatial groupings and incorporating detailed predictors, excel at capturing regional clusters and nuances. These models can identify patterns and variations that broader, global models might miss or inadequately represent. On the other hand, global models are designed to capture overarching trends across larger areas but often overlook the finer local details crucial for accurate predictions in specific regions.

The findings of Yuan et al. (2023) support this distinction, highlighting that integrating large-scale stationary measurements with local mobile data improves modeling performance in urban areas by accounting for finer spatial variations. Their study underscores the limitations of global models, which, while providing a broad overview, may fail to capture the detailed local variations necessary for precise predictions. By combining global and local data, a more accurate and nuanced depiction of air pollution can be achieved, particularly in complex urban environments where local details are critical.

Spatial variation in feature importance

While feature importance may be consistent across cities, the influence of specific predictors on NO_2 concentrations can vary significantly between cities. For example, building density and population are more significant contributors to air pollution in Utrecht, whereas traffic has a greater impact on high NO_2 concentrations in Hamburg. Applying global models with the same predictors across different cities may not yield optimal results; instead, models tailored to the specific conditions and dominant predictors of each city may provide better predictions. However, an important consideration is that each city must have a sufficient number of observations to avoid unreliable predictions.

Model quality

The limited number of observations in the local dataset poses challenges for fitting complex models. To address unreliable predictions, outliers were removed after model predictions. Transforming the original data could potentially avoid predictions falling outside the plausible range (e.g., below 0 $\mu g/m^3$). However, in this study, such transformations, like a log transforma-

tion, were not applied. Although airborne pollutant concentrations are often positively skewed (Maranzano et al., 2020), Lu et al. (2023) found that the best modeling results were obtained without data transformation and using Gaussian likelihood, even when other distributions like Gamma might better match the data distribution. Moreover, while the LASSO and Ridge models appear useful with the global dataset, their predictions were less satisfactory with the local dataset. In this study, traffic volumes were a significant feature, yet no distinction was made between different types of traffic (e.g., cars, buses, trucks), vehicle types (e.g., electric, diesel), or engine types, all of which are known to influence air pollution (Wong et al., 2021). For example, distinguishing between vehicle types could reveal that certain roads, such as those leading to or from the port of Hamburg, have a higher proportion of trucks, which might explain localized clusters of high NO_2 concentrations.

5 Conclusions

In this study, we investigate the spatial heterogeneity of NO_2 modeling by comparing various linear and non-linear statistical models at different scales (local vs. global). One of the key findings of this study is that the model performance varies little with models of different levels of complexity, but spatially in various population, traffic, and urban settings. Non-linear techniques predict better in rural and suburban areas, compared to linear models. Global model prediction accuracy is considerably higher in areas far from roads than in areas near roads. Methods preferred in global modeling appear to be unfavorable in local modeling. The relatively few NO_2 observations used in the local models could explain why non-linear models perform poorly. Using the local dataset of our study, we also found that explicitly accounting for spatial autocorrelation in the universal and ordinary kriging models does not improve accuracy; however, analyzing prediction performance across spatial groups provides valuable insights. Lastly, the spatial prediction patterns of global models indicate that non-linear methods are generally less sensitive to spatial variability than linear methods, and different modeling techniques lead to different NO_2 clusters in the prediction map. Our results suggest that only looking at the overall prediction accuracy is insufficient and can be misleading.

Code and data availability

Codes and data are 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 can be accessed in another repository: 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. ane M.L.; resources, F.B. and M.L.; data curation, F.B.; original draft preparation, F.B. and M.L.; revision 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. Both 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.

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.

Idir, Y. M., Orfila, O., Judalet, V., Sagot, B., and Chatellier, P.: Mapping urban air quality from mobile sensors using spatio-temporal geostatistics, Sensors, 21, 4717, 2021.

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.

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.

Khan, M., Almazah, M. M., Ellahi, A., Niaz, R., Al-Rezami, A., and Zaman, B.: Spatial interpolation of water quality index based on Ordinary kriging and Universal kriging, Geomatics, Natural Hazards and Risk, 14, 2190 853, 2023.

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.

Lee, J., Sun, Y., and Chang, H. H.: Spatial cluster detection of regression coefficients in a mixed-effects model, Environmetrics, 31, e2578, 2020.

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.

Mullen, R. S. and Birkeland, K. W.: Mixed effect and spatial correlation models for analyzing a regional spatial dataset, in: Proceedings of the 2008 International Snow Science Workshop, Whistler, British Columbia, pp. 421–425, 2008.

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.

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.

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.

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.

Yuan, Z., Kerckhoffs, J., Shen, Y., de Hoogh, K., Hoek, G., and Vermeulen, R.: Integrating large-scale stationary and local mobile measurements to estimate hyperlocal long-term air pollution using transfer learning methods, Environmental research, 228, 115 836, 2023.