# 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 promote public health. 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 lists 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 increasing availability of 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, employing 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 solely on prediction accuracy, without considering the models' ability to quantify uncertainty, offer scientific interpretation, or integrate known physical mechanisms (Lu et al., 2023). Specifically, many studies neglect the optimal estimation of the covariance function and the specification of priors in geostatistical modeling.

Although linear regression models are advantageous for their interpretability and ability to extrapolate, they may not capture 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 ensemble 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 had 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 in 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), whose modeling results implied 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 and they are increasingly applied to urban and geohealth studies. However, evaluating these models and maps remains challenging. One challenge is the scarcity of air pollution measurements. Another is the neglect of the spatial heterogeneity in air pollution mapping. 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 the prediction accuracy across these categories. Most current statistical approaches assess only overall accuracy (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 could 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 still lacking. As noted by Shaddick et al. (2020), uncertainty in air pollutant measurements is rarely discussed. This lack of evaluation can lead to overlooked biases, particularly because non-parametric machine learning methods often lack extrapolation capabilities. When prediction areas differ significantly in their societal and environmental characteristics from the training data, this can result in highly biased estimates that are rarely examined in many studies (Shaddick et al., 2020).

Given the growing number of modeling and prediction techniques, and the risk of misrepresenting spatial patterns due to data heterogeneity, this study seeks to answer the following research questions: *To what extent can statistical models predict $NO_2$ concentrations using high-quality, high-temporal-resolution ground station measurements? How do the performance and spatial accuracy of these models vary?*

This study focuses on the Netherlands and Germany and uses two datasets: the official national ground station measurements from both countries (referred to as the global dataset) (OpenAQ, 2017; EEA, 2021), and the more dense long-term measurements collected by Palmes tubes in Amsterdam from the Amsterdam area (referred to as the local dataset) (Gemeente Amsterdam, 2022). Palmes tubes are passive samplers used in the routine monitoring network that measure $NO_2$ on street lanterns and building facades in Amsterdam. The global dataset includes 482 measurement stations covering 398,000 $km^2$ (0.0012 points per $km^2$), while the local dataset contains 132 stations across 196 $km^2$ (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 the Palmes data in the Amsterdam region.

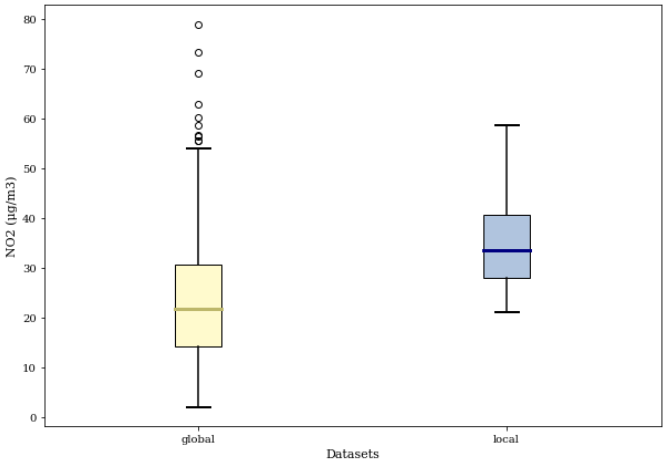



**Figure 1.** Distribution of $NO_2$ concentrations in the global (yellow) and local (blue) datasets.

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 characteristics of spatial patterns (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:

**Table 1.** Descriptive statistics of variables determining spatial groups for the local and global datasets. The statistics are derived from the station measurement locations. The distances in the "Variable"-column represent different buffer radii around measurement stations.

| Variable | Dataset | Mean | Min | 25% | 75% | Max |
|---|---|---|---|---|---|---|
| Road class 1 100 m (total length of highways [m]) | Local | 2154.787 | 0 | 0 | 3001.109 | 12950.676 |
| | Global | 12.295 | 0 | 0 | 0 | 982.912 |
| Road class 2 100 m (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 3 100 m (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 1000 m | Local | 111157.013 | 20097.258 | 106347.117 | 128723.570 | 137546.047 |
| | Global | 6154.486 | 0 | 2204.520 | 9036.756 | 20300.887 |

1. **Urban:** Areas within 100 meters of road class 1 (highways) or 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 or 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 or 2; or areas where road class 3 values are in the lowest 75%.

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, which had a converging effect on the relative group sizes. Moreover, the increase in samples classified as "urban" is desirable, as this group exhibits relatively high heterogeneity. The local dataset consists of 56 observations 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 toward mitigating 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 among 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 (100 m, 500 m, 1000 m) 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 25 m, 50 m, 100 m, 400 m, and 800 m. 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 10,000 estimators, with a reg_alpha of 2, reg_lambda of 0, max_depth of 5, and a learning rate of 0.0005 (see also supplementary 6). 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 7a-c, 8).

#### 2.2.2 Multiple Linear Regression

Key variables identified by the random forest model are used as predictors in Multiple Linear Regression (MLR). The Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge regularizations 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, optimized using the lowest Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and highest $R^2$. The search grid ranges 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 essentially a form of Gaussian process regression, developed and applied in Geosciences with a focus on spatial prediction. It typically uses spatial coordinates as covariates and places strong emphasis on variogram modeling to capture spatial dependence. The residuals of a linear regression model are treated as realizations of a spatial stochastic process, and their covariance is modeled to make predictions. Kriging is particularly suitable for estimating $NO_2$ concentrations, as $NO_2$ tends to vary smoothly over space.

In this study, ordinary and universal kriging are applied. Ordinary kriging assumes that the mean of the variable predicted is constant but unknown. It assumes stationary without trend removal. 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 effect of modeling spatial correlation on the prediction accuracy (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. The 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 and affect model interpretation(Araki et al., 2018).

Shapley values are calculated for each feature (i.e. predictor) based on its contribution $\phi_j$ 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 con-

tribution) (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-times 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 9 and 10, with the Shapley ranking of a single run shown in Figure 2. The most influential predictors for the global models include nightlight intensity (450 m and 3150 m buffers), population density (1000 m and 3000 m buffers), road class (class 2 within 25 m and class 3 within 300 m and 3000 m buffers), the annual mean $NO_2$ column density of 2018 measured by the TROPOMI instrument on-board of Sentinel-5p (trop mean filter 2018), building density in 100 m buffer, NDVI, and traffic buffers (25 m and 50 m 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 but the improvement is marginal beyond this number.

Due to the random forest model's poor performance across all local station measurements (Supplementary Figures 11a-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 (450 m and 4950 m buffer), population density (3000 m buffer), road class (class 1 within 5000 m; class 2 within 1000 m and 5000 m buffers; road class 3 within 100 m and 300 m buffers), and traffic buffer (50 m 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 with LASSO and Ridge penalization (the models are also called LASSO and Ridge). For local modeling, we compare linear models, mixed-effect models, and kriging models. Each model is evaluated based on the standard matrix of $R^2$, RMSE, and MAE (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 90/10 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

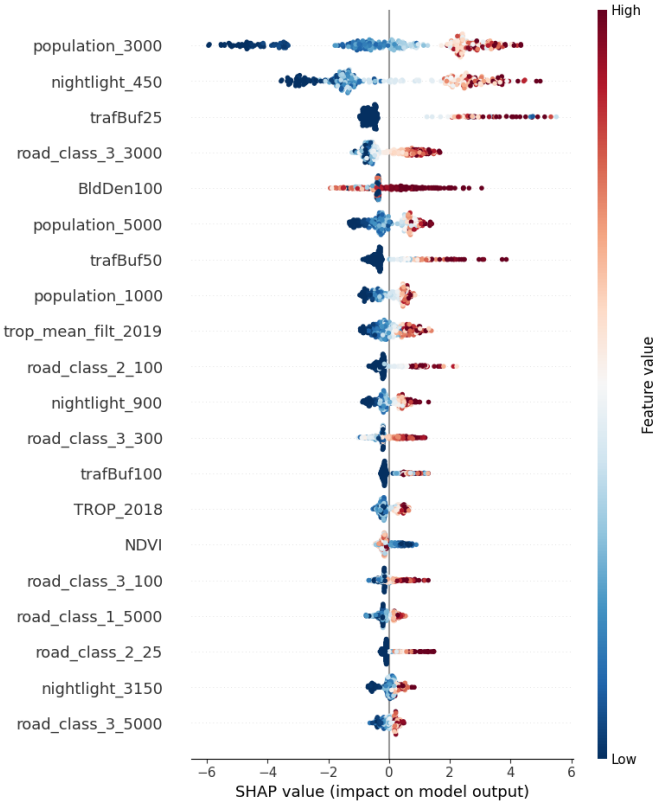



**Figure 2.** Variable importance ranked by Shapley values in a single run using the global dataset.

collected by two Google Street View cars that continuously measured $NO_2$ 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 data used in Kerckhoffs et al. (2019) are over specific, limited time periods, while our models address predictions over broader temporal spans. Despite this temporal inconsistency, the detailed spatial granularity of the annual map from Kerckhoffs et al. (2019) 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 100 m, with raster files of the most important predictors resampled and converted into 100 m grid cells for these regions, through spatial extraction methods. The influential predictor information (for global models, see Tables 2 and 4; for local models, see tables 3 and 4) is recalculated at 100 m resolution for the extent of the aforementioned regions. The 100 m by 100 m grid cells containing predictor information are used to predict $NO_2$ values for the respective 100 m grids, based on the trained local and global models (Lu et al., 2020, 2023).

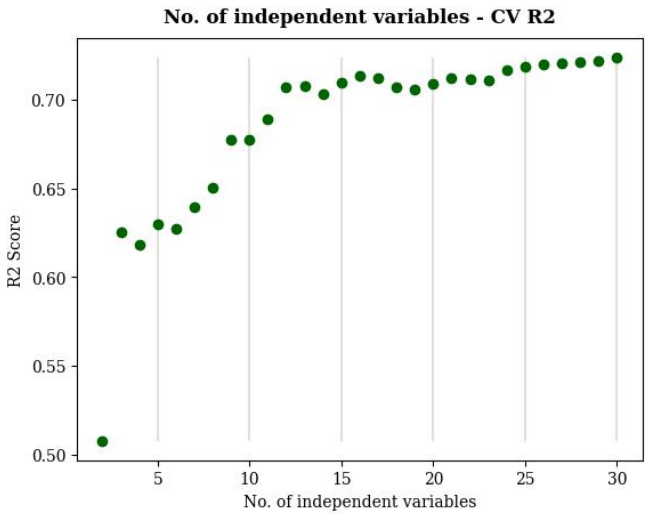
**Table 2.** Descriptive statistics for global predictors. bldden100 = Built area 100 m buffer, NDVI = Normalized Difference Vegetation Index, nightlight_3150 = Nightlight 3150 m buffer, nightlight_450 = Nightlight 450 m buffer, population_1000 = Population in 1 km grid, population_3000 = Population in 3 km grid, road_class_2_25 = Total length of primary roads 25 m buffer, road_class_3_300 = Total length of local roads 300 m buffer, road_class_3_3000 = Total length of local roads 3000 m buffer, trafbuf25 = Traffic count 25 m buffer, trafbuf50 = Traffic count 50 m buffer, trop_mean_filt_2018 = TROPOMI 2018 mean vertical column density. Note that the NDVI has a scale factor of 0.0001.

| Variable | Unit | 25th | 50th | 75th | Mean | Median | Max | Min |
|---|---|---|---|---|---|---|---|---|
| bldden100 | % | 0.4 | 0.88 | 0.99 | 0.68 | 0.88 | 1 | 0 |
| NDVI (scaled) | - | 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 |

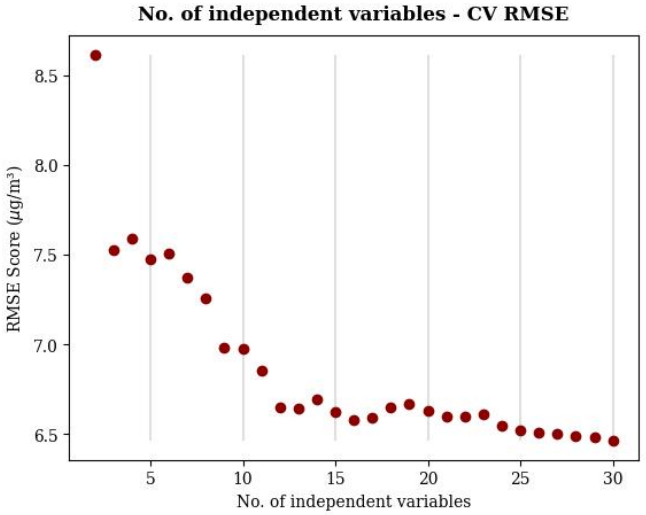
* The values for *trop mean filt 2018* are very small, on the order of $10^{-5}$.

**Table 3.** Descriptive statistics for local predictors. nightlight_450 = Nightlight in a 450 m buffer, nightlight_4950 = Nightlight in a 4950 m buffer, population_3000 = Population in a 3 km grid, road_class_1_5000 = Total length of highway a 5000 m buffer, road_class_2_1000 =Total length of primary roads in a 1000 m buffer, road_class_2_5000 =Total length of primary roads in a 5000 m buffer, road_class_3_100 =Total length of local roads 100 m buffer, road_class_3_300 =Total length of local roads in a 300 m buffer, trafbuf50 = Traffic count in a 50 m 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) Number of features and the corresponding $R^2$ (R-squared).




(b) Number of features and the corresponding RMSE (Root Mean Squared Error)

**Figure 3.** Out-of-sample performance in 20-times repeated random sampling validation: number of features and corresponding model performance, with the global model.

The 100 m 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<br>road_class_3_3000<br>trafbuf25<br><br>population_1000<br>nightlight_450<br>nightlight_3150<br>trafbuf50<br>road_class_3_300<br>bldden100<br>ndvi<br>road_class_2_25<br>trop_mean_filt_2019 | Random Forest<br>XGBoost<br>LASSO<br><br>Ridge | cross validation over the entire area<br>cross validation over different land types<br>compared with the final map of Kerckhoffs et al. (2019) |
| Local model | nightlight_4950<br>nightlight_450<br>road_class_3_100<br><br>trafbuf50<br>road_class_3_300<br>road_class_2_1000<br><br>road_class_2_5000<br>population_3000<br>road_class_1_5000 | linear model<br>linear model separating for spatial groups<br>mixed-effects model<br><br>ordinary kriging<br>universal kriging<br>universal kriging separated for different spatial groups | cross validation over the entire area<br>cross validation over different land types<br>compared with the final map of Kerckhoffs et al. (2019) |

Evaluations of the different linear and non-linear models were carried out using repeated random sampling validation, performed 20 times. In each iteration, 90% of the data was used for training and the remaining 10% for testing. For testing model performances on spatial groups, 30 testing samples are used every time. 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.

**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); Fitting a model in each land group. |
| 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, L1 and L2 norms, learning rate, maximum depth | Not explicitly |

# 3 Results

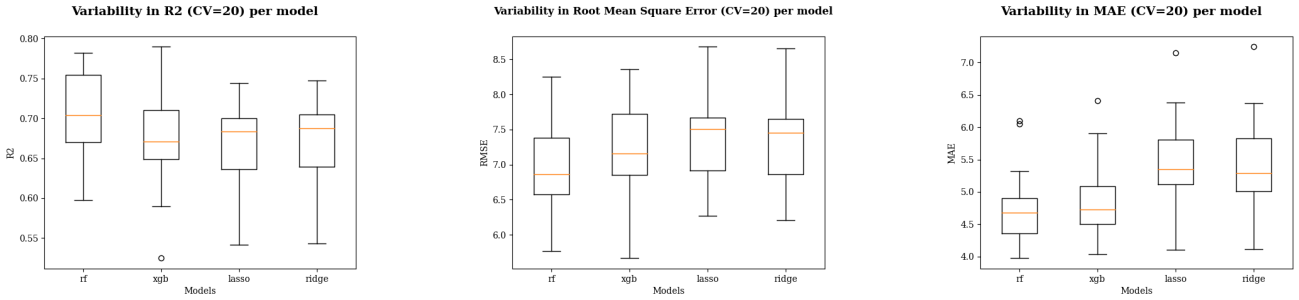
## 3.1 Global models

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 $R^2$. 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 MAE (Figure 4a and Figure 4c).

*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 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.




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

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

**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 |

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 | 0.270 | 10.991 | 8.955 | 0.378 | 7.345 | 5.426 | 0.718 | 4.147 | 2.990 |
| | | SD | 0.104 | 1.249 | 0.955 | 0.185 | 1.293 | 0.744 | 0.103 | 0.968 | 0.525 |
| | XGB | Mean | 0.212 | 11.356 | 9.165 | 0.410 | 7.159 | 5.292 | 0.739 | 3.986 | 2.745 |
| | | SD | 0.160 | 1.172 | 0.907 | 0.174 | 1.283 | 0.713 | 0.116 | 1.181 | 0.533 |
| Linear | RIDGE | Mean | 0.290 | 10.754 | 8.921 | 0.274 | 7.891 | 6.117 | 0.612 | 4.919 | 3.743 |
| | | SD | 0.163 | 1.070 | 0.969 | 0.206 | 1.071 | 0.676 | 0.122 | 1.164 | 0.677 |
| | LASSO | Mean | 0.263 | 10.946 | 9.003 | 0.255 | 7.996 | 6.171 | 0.613 | 4.911 | 3.749 |
| | | SD | 0.178 | 1.167 | 1.049 | 0.212 | 1.107 | 0.675 | 0.119 | 1.152 | 0.678 |

**Table 7.** Model performance per spatial group (20 times bootstrap, 30 samples used for testing per time). 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 $NO_2$ is consistently observed along major roads and in some urban areas, such as Haarlem (see Supplementary Figure 12).

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 13a-c, 14a-c, 15a-c, and 16a-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 $NO_2$ pollution, 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 the relationship between population or building density and air pollution (see also Supplementary Figures 16a-e). Supplementary Figure 17 provides the distribution of predicted $NO_2$ concentrations for each global model and location.

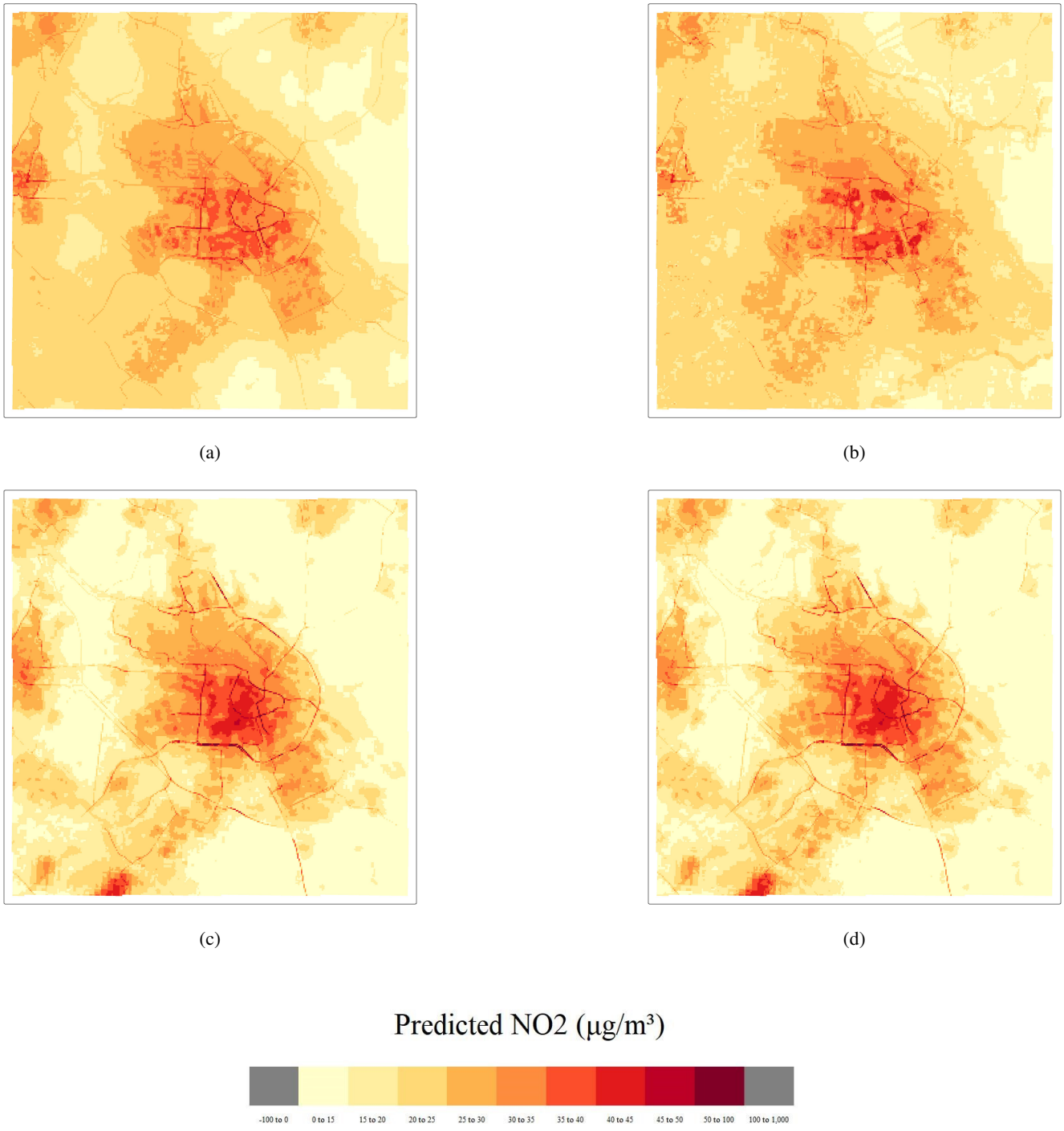



**Figure 5.** Spatial patterns of predicted $NO_2$ (100 m), measured in $\mu$ g/m3, non-linear global models for Amsterdam. (a) random forest, (b) XGBoost; linear models (bottom): (c) LASSO, (d) Ridge. The area is 30 km x 30 km.

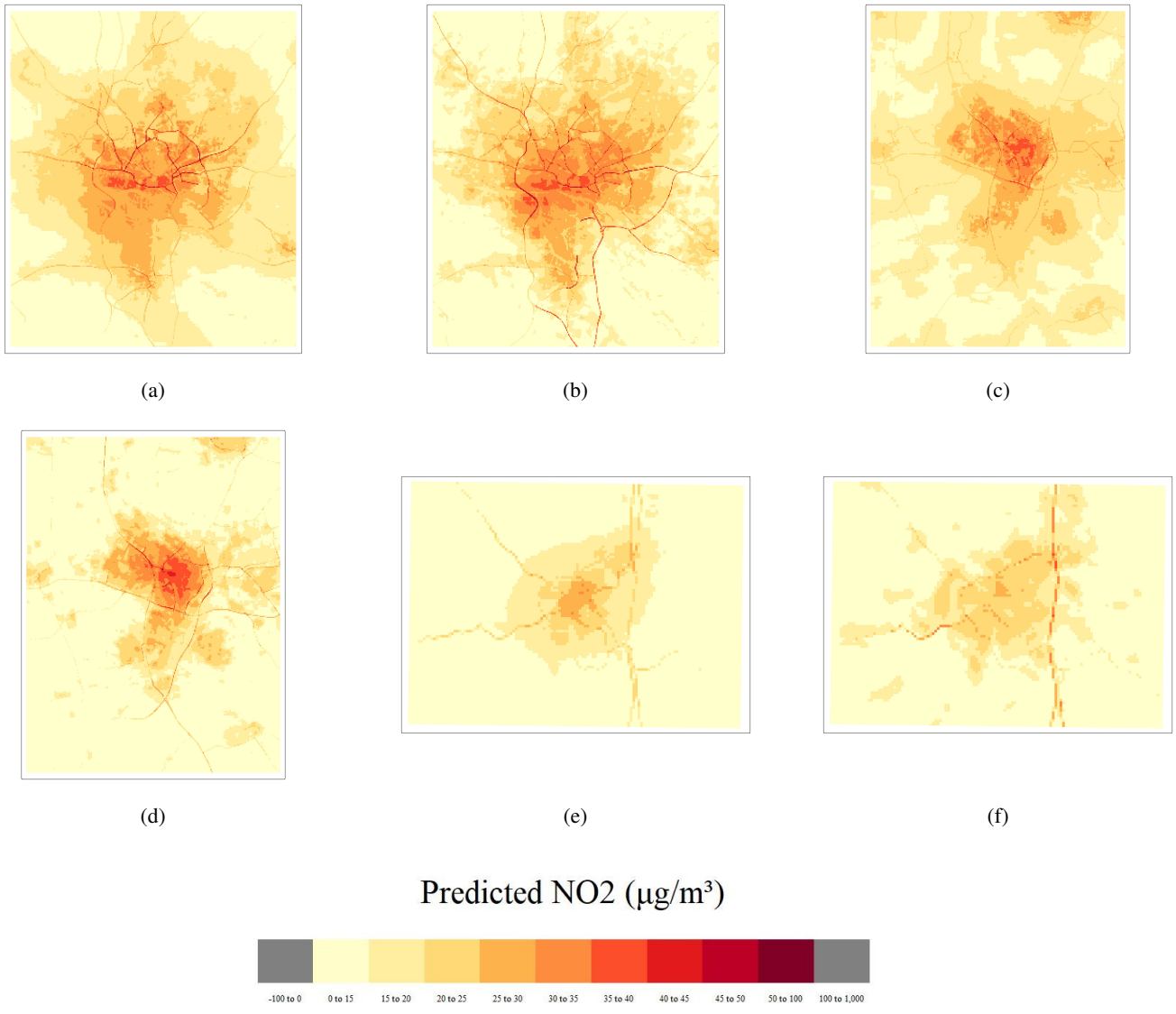



**Figure 6.** Spatial patterns of predicted $NO_2$ (100 m), measured in $\mu$g/m3, per global model for Hamburg (area: 30 km x 30 km), Utrecht (area: 25 km x 25 km) and Bayreuth (area: 10 km x 10 km) - top: from left to right, random forest (Hamburg), Ridge (Hamburg), random forest (Utrecht); bottom: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

### 3.2 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, note that with the ordinary kriging model, we could see the smoothed spatial patterns of the air pollution measurements. 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 the importance of accounting for spatial heterogeneity.

**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 |

### 3.3 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., the number of primary roads within each 5000 m buffer), as this predictor exhibits a similar cluster in the same location (see Supplementary Figures 18, 19a-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 20). The relatively low $NO_2$ values along roads in the outer Amsterdam area can be attributed to the spatial grouping. 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 the observations are located within 100 meters of the 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 21a-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.

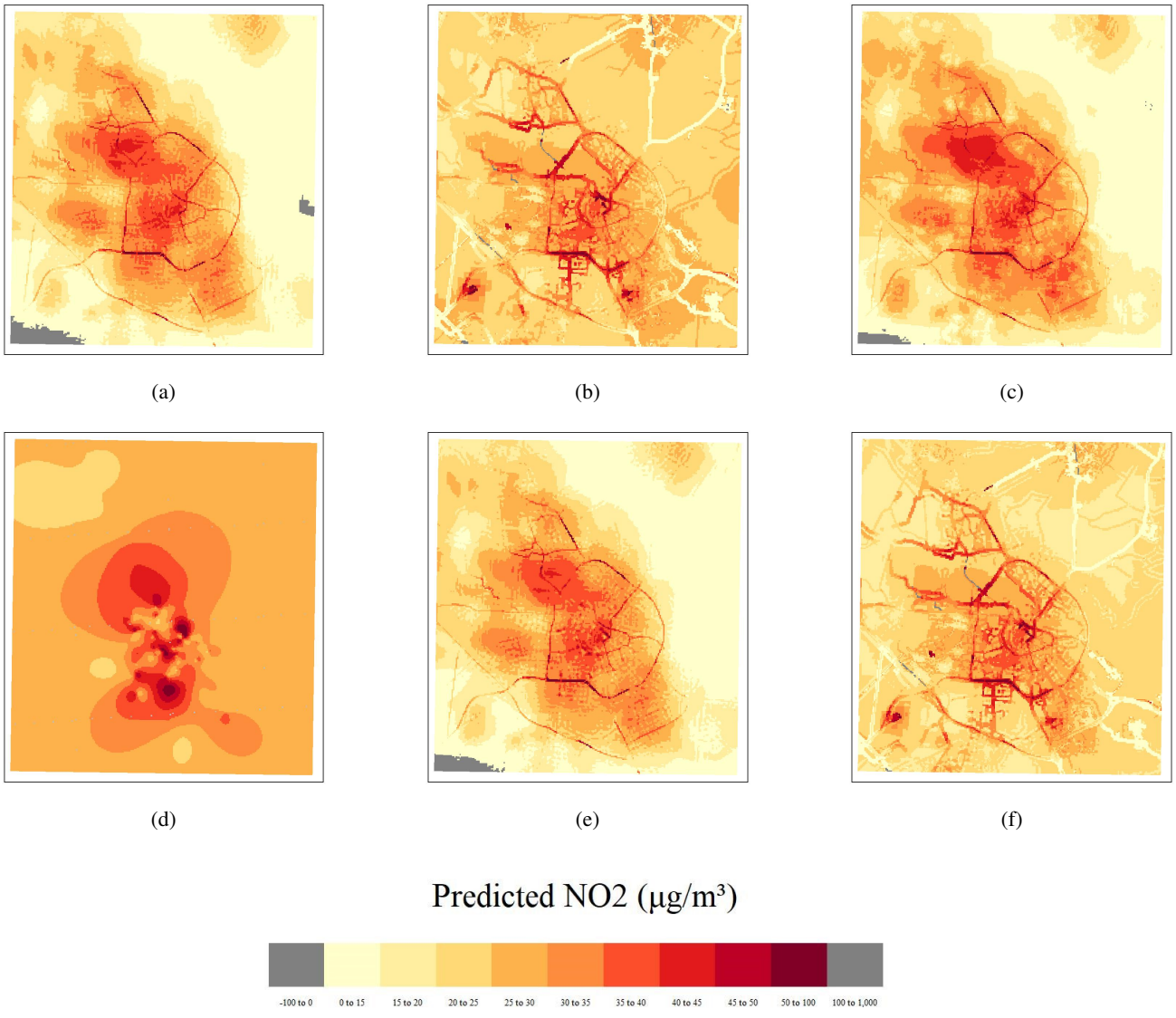



**Figure 7.** Spatial patterns of predicted $NO_2$ ($\mu$g/m3) at 100 m resolution based on the local dataset - top: left = linear model, middle = linear model separating for spatial groups, right = mixed-effects model; bottom: left = ordinary kriging, middle = universal kriging, right = universal kriging 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 23). 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 removed (including LightGBM results) is shown in Supplementary Figure 24. 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. Although the assumption of second-order stationarity is likely violated in ordinary kriging due to the strong influence of local traffic on $NO_2$ concentrations, the method remains applicable. It still provides the best linear unbiased prediction (BLUP) based on Euclidean distances between spatial coordinates. When comparing the models with the open $NO_2$ dataset, some local models show more similarity than global models. This is reasonable as the local model dataset is also from Amsterdam. Table 10 shows the residuals per global and local model. The XGBoost model emerged as the most accurate among global models, with the lowest mean residual (1.36), indicating it closely matched open $NO_2$ values. The LASSO model also demonstrated higher residuals compared to XGBoost and Ridge, suggesting less consistency in its predictions. In contrast, local models exhibited greater variability in residuals. The mixed-effects model and universal kriging had relatively moderate mean residuals (2.51 and 1.83, respectively), while the linear spatial groups and universal kriging spatial groups models had significantly higher standard deviations, indicating more extreme residuals. Ordinary kriging retained the highest mean residual (4.71), reinforcing the trend that local models generally 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 25a-e and 26a-f respectively. A spatial comparison of the global and local model predictions with the measurement station data can be found in Supplementary Figures 27 and 28.

**Table 10.** The summary statistics of the difference between model predictions and the mobile measurement-derived $NO_2$ map from Kerckhoffs et al. (2021).

| Model | Type | Mean | Median | SD | Min | Max |
|---|---|---|---|---|---|---|
| Random forest | Global | 1.65 | 3.44 | 8.94 | -54.72 | 19.83 |
| LASSO | Global | 1.85 | 3.19 | 9.27 | -54.65 | 24.75 |
| Ridge | Global | 1.75 | 3.16 | 9.35 | -54.53 | 24.47 |
| XGBoost | Global | 1.36 | 3.02 | 9.41 | -58.24 | 23.18 |
| 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 |

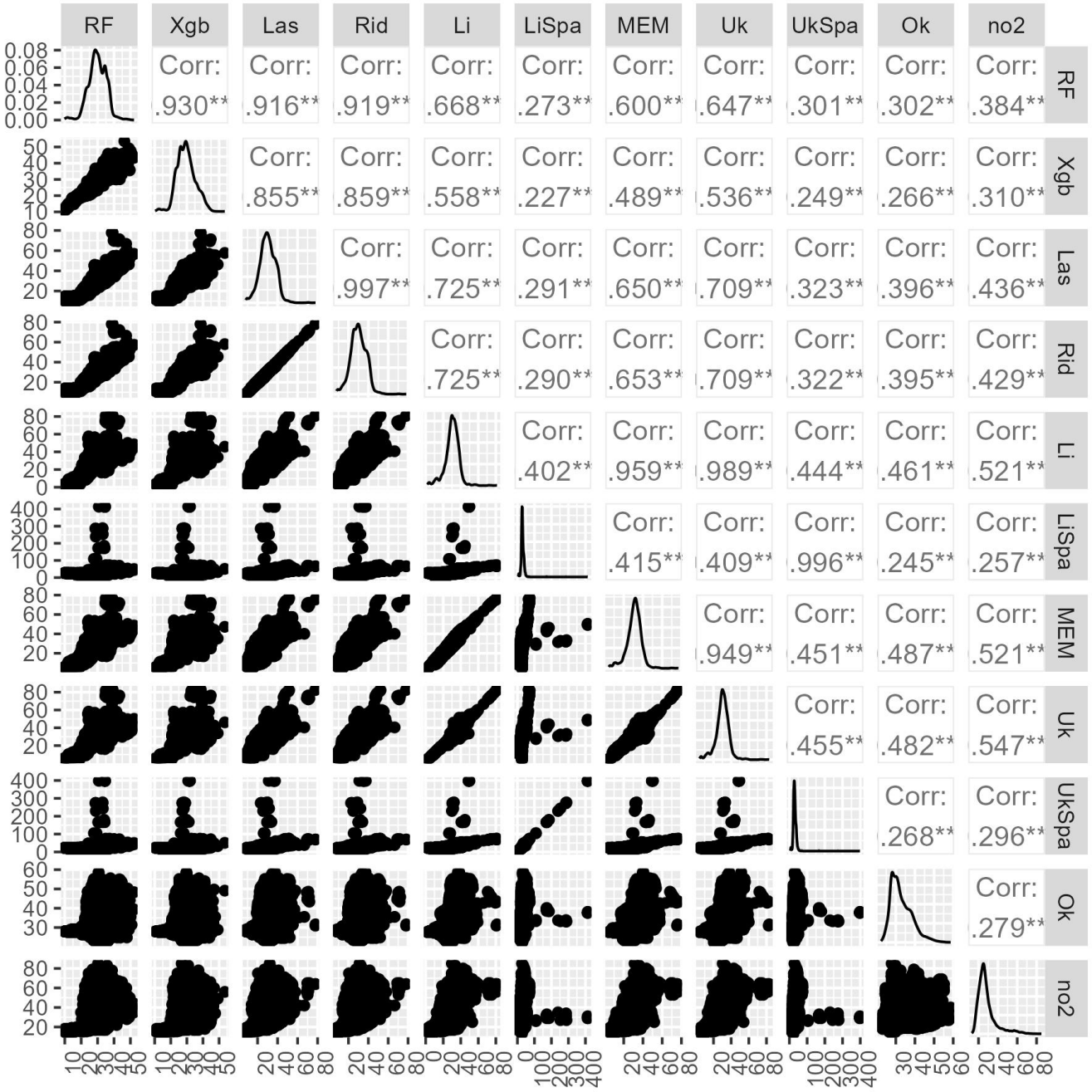



**Figure 8.** Comparing model predictions whereby the numbers equal the Pearson correlation coefficient. RF: Random Forest, XGB: XGBoost, LR: linear regression, LRsp: Linear Regression accounting for spatial groups, MEM: Mixed-Effects Model, UK: Universal Kriging, UKsp: Universal Kriging accounting for spatial groups, OK: Ordinary Kriging, no2: mobile $NO_2$ map from Kerckhoffs et al. (2021).

## 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, as well as a thorough analysis of the prediction patterns in different areas using different models has often been under-stressed. In this study, we address this gap by comparing spatial and non-spatial models 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 the pollutant sources vary (Chen et al., 2019), the modeling results for $NO_2$ may not be directly applicable to other pollutants.

*Accounting for spatial groups*

Without accounting for the spatial groups, the differences in terms of the accuracy assessment matrice between linear and non-linear techniques are minimal. The random forest model generally performs with the highest $R^2$ and lowest MAE and the $R^2$ of the Ridge model is higher than that of the XGBoost model. When accounting for spatial groups, 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.

A limitation of the data 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.

*Influence of Cross-Validation Techniques*

Cross-validation (CV) strategy plays a crucial role in model performance and generalizability. In this study, we opted for a 90/10 train-test split, repeated 20 times, to ensure model stability while maintaining sufficient training data. For testing on spatial groups, we further limited the testing samples to 30 to account for spatial variability and avoid data imbalance. This approach allows us to evaluate model performance across different spatial settings while mitigating the risk of overfitting.

However, the random split employed in this approach can lead to biased performance estimates, particularly with small datasets. Some points may be used multiple times, while others might not be used at all, which can skew results. To address this issue, alternative strategies like 5-fold CV are often employed, as they provide a good balance between bias and variance. In 5-fold CV, the dataset is divided into five partitions, ensuring that each point is used for validation exactly once. This method can be particularly useful when the dataset is small, as it ensures more comprehensive utilization of the data.

In our study, we chose to use bootstrapped CV because it provides a robust measure of uncertainty, particularly for relatively small datasets. Bootstrapping works well when data is limited, as it generates multiple datasets through resampling, allowing us to estimate variability and model performance more effectively. The bootstrapped CV aims to reduce bias by randomly drawing training points for each iteration, thus providing a more reliable estimate of model accuracy. While increasing the number of bootstraps and reducing the test set size could further reduce bias, leaving more data points in the test set offers additional information about the model's performance.

Alternative techniques, such as 5-fold CV, remain popular because they offer a straightforward balance between training and validation data. Nonetheless, bootstrapping may provide better estimates of uncertainty, especially in cases with limited data, which could make it a preferable choice for heterogeneous spatial datasets. Although increasing the number of folds or bootstraps may improve predictive reliability, the trade-off between computational efficiency and statistical robustness must be carefully considered.

Meyer and Pebesma (2021, 2022) criticized the imprudent use of global-scale models, particularly highlighting the issue that model performance cannot be validated in regions without observational data. They advocate for the use of spatial cross-validation to address this limitation. After careful consideration, we opted for randomly bootstrapped cross-validation to mitigate the bias introduced by spatial cross-validation (Wadoux et al., 2021; Lu et al., 2023). We would like to emphasize that the core issue in cross-validation for spatial modeling lies in the inclusion of spatial coordinates or distances as predictors. From this perspective, it is evident that sampling should be random in the predictor space. Therefore, we argue that cross-validation in spatially correlated data is not fundamentally different from standard cross-validation, provided that the predictor space sampling is handled appropriately.

*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*

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 conceal this finding and yield sub-optimal results. It is therefore important to consider the spatial heterogeneity and at the same time ensure a consistent uncertainty assessment. However, the number of official ground stations as for now may not be sufficient to characterise the spatial heterogeneity and ensure a detailed and reliable prediction.

*Further perspectives on model improvement*

The limited number of observations in the local dataset poses challenges for fitting complex models. 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, a transformation was not applied for the following reason. 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 perform well 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 matrices varies trivially with models of different levels of complexity, but significantly when consider spatial heterogeneity and modeling

in various population, traffic, and urban settings. The non-linear techniques predict better in rural and suburban areas, compared to linear models. Global model prediction accuracy is considerably higher in rural, homogeneous areas, the influence of which lead to high overall performance without accounting for spatial groups. Methods preferred in global modeling could be unfavorable in local modeling. The relatively few $NO_2$ observations and potentially lower quality in the local (Palmes) dataset compared to the official ground station measurement are important reasons of the unsatisfactory performance of non-linear models. Using the local dataset, we also found that explicitly accounting for spatial autocorrelation in the universal and ordinary kriging models does not improve accuracy; however, analyzing predictions across spatial groups provides valuable insights. Also, different modeling techniques lead to different $NO_2$ clusters in the prediction map despite the similar performance matrices they received. Last but importantly, our results suggest that focusing solely on overall prediction accuracy can lead to overconfidence and an underestimation of the further efforts required in statistical air pollution mapping.

*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.15008748

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.

Kerckhoffs, J., Hoek, G., Gehring, U., and Vermeulen, R.: Modelling nationwide spatial variation of ultrafine particles based on mobile monitoring, Environment International, 154, 106 569, 2021.

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.