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 prone to overfitting than linear methods. Additionally, clusters of predicted air pollution differ between models within cities despite similar prediction accuracy. ~~The~~ Also, the influence of predictors on NO_2 concentrations varies across different cities. ~~Lastly, applying the same methods to the local dataset yields poor accuracymetrics, the linear methods outperform non-linear methods, contrary to the results of~~ For the local models, accounting for spatial autocorrelation does not improve accuracy, but modeling spatial groups does. 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 ~~estimation of NO_2 concentration levels are essential for understanding air pollution comprehensively, which is important for~~ estimating NO_2 concentration levels is essential for a comprehensive understanding of air pollution, which plays a critical role in urban planning and ~~making political decisions towards~~ policy-making to foster a healthy society. Air pollutants

have been modeled ~~at~~ across various spatial scales, ~~up to the globallevel. The~~ from local to global. These models can be ~~classified into~~ broadly classified into three categories: statistical models, chemical ~~transportation models~~transport models, and air dispersion models. ~~The chemical transportation models aim at~~ Chemical transport models are typically used for large-scale air pollution modeling~~. The~~ , while air dispersion models require detailed~~and spatially revolved~~ , spatially resolved emission data to ~~model~~ capture small-scale ~~spatial variations in air~~ variations in pollutants (Beelen et al., 2013). ~~Statistical modeling is becoming more popular in~~

In recent years, statistical modeling has gained popularity for high-resolution mapping at different spatial scales, ~~due to an increment~~ driven by the increase in available predictors (~~i.e.~~ e.g., GIS variables) and ~~computational capability~~advancements in computational capabilities. Land Use Regression (LUR) is the most well-known statistical approach for air pollution modeling~~approach. LUR builds linear regression models~~ , using linear regression to capture the spatial variability of traffic-related air pollution in urban areas. Most LUR models ~~are based on measurements~~ rely on data from ground monitoring stations (Hoek et al., 2008; Wang et al., 2020). Geostatistical methods ~~such as kriging could further capture the~~ like kriging can further account for spatial correlations between ~~the~~ observations. However, several studies ~~favor~~ have favored the simplicity of LUR~~and conclude that they outperform or are equivalent to~~ , 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). ~~However, most of these studies draw this conclusion purely~~ Notably, these conclusions are typically based on prediction accuracy ~~but not on~~ alone, without considering the models' ability ~~in uncertaintyquantification, scientific interpretation, and the integration of~~ 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 ~~of parameters in the geostatistical modelingare commonly not a priority of these studies, or are neglected~~in geostatistical modeling.

~~Although the linear models have the advantage of being highly interpretable and can be extrapolated~~While linear models are advantageous for their interpretability and ability to extrapolate, they may ~~not capture the complex~~ fall short in capturing the complex processes of air emission, dispersion, and deposition ~~processes (Wang et al., 2020). Data-driven,~~ (Wang et al., 2020). As a result, data-driven, non-parametric models~~(most commonly known~~ —commonly referred to as machine learning methods in air pollution mapping~~)~~—have become increasingly popular. These models, such as tree-based algorithms~~have been more on trend in air pollution mapping, as these methods may better capture~~ , 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 example, Brokamp et al. (2017) compare the~~ (Weichenthal et al., 2016; Reid et al., 2015; Lu et al., 2020). For instance, Brokamp et al. (2017) compared Land Use Random Forest ~~models~~ (LURF) ~~with the~~ models with LUR models for elemental components of PM~~2.5 in the urban area of~~ $_{2.5}$ in Cincinnati, Ohio, and ~~find that the LURF shows a~~ found that LURF models demonstrated lower prediction error variance ~~for each elemental model with cross-validation. Kerckhoffs et al. (2019) report~~ across all elemental models when cross-validated. Similarly, Kerckhoffs et al. (2019) reported that machine learning algorithms, such as bagging and random forest, ~~explain~~ explained more variability in ultra-fine particle concentrations than multiple linear regression and regularized regression techniques. Ameer et al. (2019) ~~advocate that the~~ advocated for random forest regression ~~is~~ as the best technique ~~, compared to~~ for pollution prediction across varying datasets, locations, and characteristics, outperforming decision tree re-

gression, ~~Multi-Layer Perceptron~~ multi-layer perceptron regression, and gradient boosting regression~~, for pollution prediction for data sets of varying size, location, and characteristics. Ren et al. (2020) conclude~~ . Ren et al. (2020) also concluded that non-linear machine learning methods achieve higher accuracy than ~~the~~ linear LUR, ~~thereby stressing that a careful design of~~ emphasizing the importance of careful hyperparameter tuning and ~~flexible~~ robust data splitting and ~~validations is important to acquire stable~~and validation to ensure stable, reliable results. Chen et al. (2019) ~~compare~~ compared 16 algorithms ~~to predict the~~ for predicting annual average fine particle (~~PM2.5~~$PM_{2.5}$) and nitrogen dioxide (NO_2) concentrations across Europe~~, and also conclude with a favor of the~~ . They found that ensemble tree-based methods ~~and the difference is more prevalent for the PM2.5 pollutant~~ were particularly effective for $PM_{2.5}$, while NO_2 ~~model predictions show a~~ models showed similar R^2 ~~. At the same time,~~ values across different methods. Importantly, they reported a high correlation ~~is reported~~ between the predicted values of ~~the various modelsused in the study. Furthermore, since they measure two pollutants,~~ various models, noting that the most influential predictors ~~differ~~ differed substantially between pollutants. ~~Satellite~~ For example, satellite observations and dispersion model estimates were ~~among the most influential predictors for PM2.5 concentrations, whereas the variation in~~ key predictors for $PM_{2.5}$ concentrations, while NO_2 ~~is primarily attributable to~~ variability was primarily driven by traffic-related variables. The ~~major~~ significant contribution of road traffic to ~~NO_2 concentrations is~~ $_2$ levels is further supported by Wong et al. (2021), ~~it was~~ who found that nitrogen ~~is produced particularly~~ emissions are particularly influenced by long-range transport ~~,~~ from gasoline-fueled passenger cars.

~~Over the past few years, there has been a notable rise in the utilization~~ In recent years, the use of statistical modeling for air pollution mapping ~~, leading to the emergence of~~ has surged, resulting in numerous local and global ~~air pollution maps . These maps are now being increasingly utilized~~ pollution maps that are increasingly applied in urban and health studies. However, evaluating ~~air pollution~~ these models and maps remains ~~a challenge. One reason is the lack~~ challenging. One challenge is the scarcity of air pollution measurements. ~~A second reason may be attributable to a different~~ Another is the varying focus on spatial heterogeneity in air pollution. For ~~instance~~example, He et al. (2022) acknowledge spatial heterogeneity in measurement stations ~~as they show~~ by demonstrating that the probability density ~~function~~ functions of concentrations (NO, ~~NO_2, PM10, PM2.5) of~~ $_2$, PM_{10}, $PM_{2.5}$) vary across different spatial categories (~~urban traffic;~~ e.g., urban traffic, suburban/rural traffic~~; urban industrial;~~ , urban industrial, suburban/rural industrial~~; urban background; suburban background;~~ , urban background, suburban background, rural background)~~show different patterns~~. However, ~~the~~ their study does not ~~focus on modeling~~ model potential differences in prediction accuracy ~~for every concentration type and spatial category~~across these categories. A third ~~reason~~ challenge is that most ~~of the current statistical modeling approaches only assess the overall accuracybut not the accuracy over space~~ current statistical approaches assess only overall accuracy, neglecting spatial variation (Hoek et al., 2008; Chen et al., 2019). Hoek et al. (2008) ~~state that a LUR model typically explains~~ reported that LUR models typically explain 60-70% of the variation in ~~NO_2. However, the~~ $_2$, but this explained variation may be ~~very low in areas~~ significantly lower near traffic. Chen et al. (2019) ~~argue that most of the previous~~ argued that many air pollution exposure ~~assessment studies make no distinction in~~ studies fail to account for the characteristics of ~~the~~ monitoring sites when performing cross-validation, potentially ~~leading to~~ misrepresenting model results. ~~Therefore, they opt to "evaluate~~ They suggest evaluating models using pollution data ~~collected~~

from monitoring sites ~~which represent~~ that reflect the application locations ~~" (Chen et al., 2019, p.3). Lastly~~(Chen et al., 2019) ~~.~~

Finally, a consistent and coherent ~~uncertainty quantification method is lacking~~ method for quantifying uncertainty in air pollution mapping ~~. Shaddick et al. (2020) argue that the~~ is lacking. Shaddick et al. (2020) pointed out that uncertainty in air pollutant measurements is ~~only discussedin limited studies. A consequence of the inadequate evaluation is the non-extrapolating property of most~~ rarely discussed. This inadequate evaluation can lead to overlooked biases, especially since non-parametric machine learning methods ~~is commonly ignored. Areas to be predicted could differ considerably in their~~ often lack extrapolation capabilities. When predicted areas differ significantly in societal and environmental ~~properties compared to the~~ characteristics from training data, ~~yielding~~ highly biased predictions may result, which are not ~~evaluated in multiple~~ adequately evaluated in many studies (Shaddick et al., 2020).

Given the ~~increasing~~ growing number of modeling and prediction techniques ~~, and the presence of~~ and the potential for misrepresented prediction maps due to heterogeneity issues, this study aims to ~~understand~~ investigate: ~~*to*~~ *To what extent can statistical models ~~be used in predicting~~ predict* NO_{2} ~~$_2$~~ *concentrations ~~given~~ using high-quality, ~~high-temporal resolution~~ high-temporal-resolution ground station measurements~~: how does~~ ? How do the performance of ~~statistical~~ these models ~~differ~~ and ~~how does it differ spatially~~their spatial accuracy vary?* The study ~~area is in~~ focuses on the Netherlands and Germany~~.~~ ~~Two datasetsare used. One is~~ , using two datasets: the official national ground station measurements ~~of the two countries (OpenAQ, 2017; EEA, 2021),~~ from both countries (referred to as ~~global dataset; and the other is the~~ the global dataset) (OpenAQ, 2017; EEA, 2021), and the more densely distributed ground station measurements ~~of~~ from the Amsterdam area ~~(Gemeente Amsterdam, 2022),~~ (referred to as ~~local dataset. The aims are~~ 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) ~~of~~ the two datasets, 2) ~~in~~ different spatial groups classified ~~based on the distances~~ by proximity to traffic and population ~~densities~~density, and 3) ~~in using different~~ various statistical models, to ~~understand~~ evaluate the added value of non-linear machine learning models and geostatistical ~~models~~approaches.

2 Methodology

2.1 Data

~~The global -~~

The global and local datasets ~~contain~~ include the annual mean ~~of~~ NO_2 concentrations ~~,~~ (measured in $\mu g/$~~m3, for~~ m^3) for the year 2017 (OpenAQ, 2017; EEA, 2021). Figure ~~?? shows the distributions~~ ?? presents the distribution of NO_2 concentrations at the global and local measurement stations.

The spatial distribution of NO_{2} ~~measurement stations are shown in~~ ~~$_2$~~ measurement stations is provided in the supplementary materials (~~figure~~ Figure 1a, 1b). Urban areas generally have a higher density of measurement stations. ~~The differences between the global -~~ This study focuses on the differences between global and local models~~are studied in Amsterdam. Less~~ , particularly

in Amsterdam, while also considering the city's less densely populated areas ~~around Amsterdam are included~~ to examine the ~~effect of the urban area on the~~ urban impact on predicted NO_2 ~~concentration levels per local model~~concentrations in the local models.

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

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

Variable	Dataset	Mean	Min
~~Road class 1 100m (total length of highway (m))~~ Road class 1 100m (total length of highways [m])	Local ~~data~~	2154.787	0
	Global ~~data~~	12.295	0
~~Road class 2 100m (total length of primary roads (m))~~ Road class 2 100m (total length of primary roads [m])	Local ~~data~~	4018.626	0
	Global ~~data~~	68.943	0
~~Road class 3 100m (total length of local roads (m))~~ Road class 3 100m (total length of local roads [m])	Local ~~data~~	25838.098	6483.43
	Global ~~data~~	272.059	0
~~Population 1000m~~ Population 1000m	Local ~~data~~	111157.013	20097.2
	Global ~~data~~	6154.486	0

The three spatial groups are defined as follows:

1. ~~Urban: areas that are~~ **Urban:** Areas within 100 meters of ~~either~~ road class 1 (~~highway) or~~ highways) and 2 (primary roads) and ~~the population 1000 (population density within 1000 meters of every measurement station) values are~~ with population density in the highest 25%; or ~~the~~ areas where both road class 3 (local roads) values and ~~the population 1000 values are both~~ population density are in the highest 25%.

2. ~~Low population: areas that are~~ **Suburban:** Areas within 100 meters of road class 1 and 2 ~~and the population 1000 values are~~ with population density in the lowest 75%; or ~~the~~ areas where road class 3 values are in the highest 25% and ~~the population 1000 values are~~ population density in the lowest 75%.

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

By applying this grouping [1], This classification resulted in 85 observations are classified being labeled as "urban", 138 as "low populationsuburban", and 259 as "far from roadsrural", together compromising the totaling 482 observations of in the global dataset. Since the local dataset contains fewer samples and is characterized by Given the higher population density in the local dataset and its smaller sample size, the threshold is adjusted from 0.75 to 0.5 (i.e. for defining "urban" is now related to the 50% highest values rather than the 75% highest values). For the local dataset, was adjusted from the 75th percentile to the 50th percentile. This adjustment was necessary to better capture the high population density in the local dataset and resulted in 56 observations are attributable to the spatial group being categorized as "urban"," 46 to as "low population",suburban," and 30 to as "far from roadsrural." . Supplementary, figure

While this adjustment introduces some inconsistency between the global and local definitions of "urban," it ensures that the local model accurately reflects the dense urban context. The unequal distribution of instances across groups could introduce bias into the statistical learning models, but this threshold adjustment was an initial step to mitigate such effects. Supplementary Figures 2 and 3 display the spatial distribution of observations through spatial groups , for the global- and local datasets, thereby including information on the spatial groups. Supplementary figure across these groups for both datasets, while Supplementary Figures 4 (global dataset) and 5 (local dataset) show the measured NO_2 concentrations per station.

Spatial predictors

A We utilized a set of variables with related data is already derived from Lu et al. (2020), including industrial areas from OpenStreetMaps, road length from OpenStreetMapsdata on industrial areas, road lengths, population densityfrom GHS-POP R2019A population grid, and Earth night light from VIIRS in various buffers, Earth night lights, wind speedand temperatureat 2 m altitude from ERA-LAND 5 climate re-analysis model, elevationfrom 30 m Radar global product, temperature, elevation, Tropomi level 3 NO_2 of 2018, and global radiation. An overview of the variables derived from Lu et al. (2020) can be found in supplementary material, table 1. We used the precipitation 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 conducted spatial interpolation interpolated using ordinary kriging to cover the NO_2 measurement stations. Kriging parameters can be found in supplementary, section parameters. The precipitation consists of average monthly precipitation data, measured in millimeters. The building density is derived from the "are detailed in the supplementary material.

Building density was obtained from the "World Settlement Layer 2015", which is publicly available on figshare (Marconcini et al., 2020) . Taking building density as an explanatory variable, multiple studies use several measurement scales, consisting of buffers that vary in size. Beelen et al. (2013) used building density in buffers of " 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, 300m, 500m, 1000m, and 5000m (sizes in radius) . Another study performed by Kheirbek et al. (2014) measures the correlation between air pollution and building density via 15 circular buffers, ranging from 50m) around measurement stations to 1000m. Lu et al. (2020) also use buffers for one explanatory factor to encourage comprehensiveness within the methodology. Varying

[1] The related code could be found in supplementary, Code 1

buffer sizes are implemented in our study too. As several measuring stations are close to each otheraccount for spatial proximity effects, especially in urban areas, the buffers are of sizes 100m, 500m, and 1000m. The Normalized Difference Vegetation Index (NDVI) values are obtained through NASA and are related to 2017 (NASA, 2017). The Dutch dataset for traffic volume is obtained via densely populated urban areas. NDVI values were obtained from NASA (NASA, 2017).

Traffic volume data was sourced from the "Nationaal Dataportaal Wegverkeer" (NDW) (Rijkswaterstaat, 2017) whereas the German dataset for traffic volume is obtained via in the Netherlands (Rijkswaterstaat, 2017) and "Bundesanstalt für Strassenwesen" (BAST) (Bundesanstalt für Strassenwesen, 2017). Both the NDW- and BAST-datasets are generated via in Germany (Bundesanstalt für Strassenwesen, 2017). This data, generated by automatic counting stations. The traffic volume is expressed in , is expressed as average hourly traffic , measured over 2017, and in buffers of sizes with buffer sizes of 25m, 50m, 100m, 400m, and 800m. The formula for calculating average hourly traffic can be found in supplementary, section equationsis provided in the supplementary material.

2.2 Modeling NO_2 globally and locally

2.2.1 Ensemble trees

The global models can be classified into use two types of statistical learning methods. The first group composes consists of ensemble tree-based approachesconsisting of , including random forest, Light Gradient Boosting (LightGBM), and Extreme Gradient Boosting (XGboostXGBoost). Hyperparameters are tuned based on the cross-validation error. For the random forest model, the number of estimators is set to 1000; the min_samples _split equals , with a minimum samples split of 10; the min_samples _leaf equals , minimum samples per leaf of 5; the maximum features used per tree is set to , maximum features per tree of 4; the maximum depth is , and a maximum depth of 10. For both the LightGBM and XGboost models , the number of estimators is set to Both LightGBM and XGBoost models use 50,000 ; the estimators, with a reg_alpha equals of 2; the , reg_lambda equals of 0; the , max_depth equals of 5; the learning rate is , and a learning rate of 0.0005. Additionally, the gamma of for the XGBoost model is set to 5. Further details can be found in supplementary material, section parameters. The equations for the ensemble trees can be found in supplementary material, section equationsthe supplementary material.

2.2.2 Multiple linear regressionLinear Regression

The key variables highlighted Key variables identified by the random forest model are chosen used as predictors in Multiple Linear Regression (MLS). MLR). Regularization techniques such as Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge regression regularize regression coefficients. The LASSO is different are employed to prevent overfitting. LASSO differs from Ridge in that the penalty equals the it uses the sum of the absolute values of the coefficients (Ren et al., 2020). Consequently, coefficients can be equal to 0 which leads to feature selection as a penalty, allowing some coefficients to be exactly zero, thus enabling feature selection (Ren et al., 2020). The alpha for the both LASSO and Ridge models are is tuned to 0.1, leading to the lowest MAEoptimizing for the lowest Mean Absolute Error (MAE), Root Mean Square Error (RMSE), RMSE, and highest R^2 out of among options ranging from 0.1 to 1 with a step in increments of

0.1. The parameters and equations for the linear regression, error term, Ridge regression, and LASSO regression can be found in supplementary material, section parameters and section equations, respectively.

2.2.3 Mixed-Effects Model and Kriging

2.2.4 ~~Mixed-effects model and kriging~~

~~To use the~~ Due to the poor performance of random forest, LightGBM, and XGBoost (Supplementary Table 4), and to a lesser extent LASSO and Ridge (Supplementary Table 5), alternative methods were selected for the local dataset.

To incorporate spatial information, we ~~add~~ also employ mixed-effects modeling and kriging methods. ~~With~~ In the mixed-effects model, fixed ~~and random effects are included. Fixed effects~~ effects consist of the most influential predictors, while random effects account for potential spatial trends~~in the data~~. The spatial ~~character of the observation, i.e. whether an observationis situated in an urbanarea, low-populated area, or far from roads area, accounts for the~~ context of an observation—whether it is urban, suburban, or rural—serves as a random effect in the model. In contrast, the linear model ~~composes all the fixed effectswhile neglecting~~ only includes fixed effects, thereby ignoring the possibility of observation clustering. ~~Ordinary and Universal kriging~~

Ordinary and universal kriging methods are used for local modeling. The ~~R package automap (Hiemstra et al., 2008) is used for the initialization of~~ `automap` package in R (Hiemstra et al., 2008) is employed to initialize the covariance parameters. Details are ~~found~~ provided in the supplementary ~~material section parameters and equations. For the~~ materials under the "Parameters" and "Equations" sections. Two separate models—one that accounts for spatial groups and one that does not—are created using universal kriging and linear modeling ~~method, two models are created with and without separating between spatial groups. When accounting for spatial groups, a model is created for each spatial group. Eventually, eleven models are~~ methods. This leads to a total of eleven models being fit and compared~~,~~: five using the global dataset and six using the local dataset. ~~The relevant equations can be found in supplementary material, section equations~~Relevant equations are included in the supplementary materials.

2.3 Feature selection

~~We first select features~~ Feature selection for global models ~~based on the Shapley value (Shapley, 1953). The variable selection removes irrelevant or strongly~~ 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 2 and 3). Feature selection aims to remove irrelevant or highly correlated predictors that ~~would otherwise~~ could generate unstable estimates (Araki et al., 2018). ~~Feature selection is based on examining the Shapley values of~~

Shapley values are calculated for each feature (i.e.~~predictor). The Shapley value for a feature value j is determined by the contribution ϕ_j of feature j~~, predictor) based on its contribution ϕ_i to the prediction ~~, in this case, NO_2~~ of NO_2 concentration

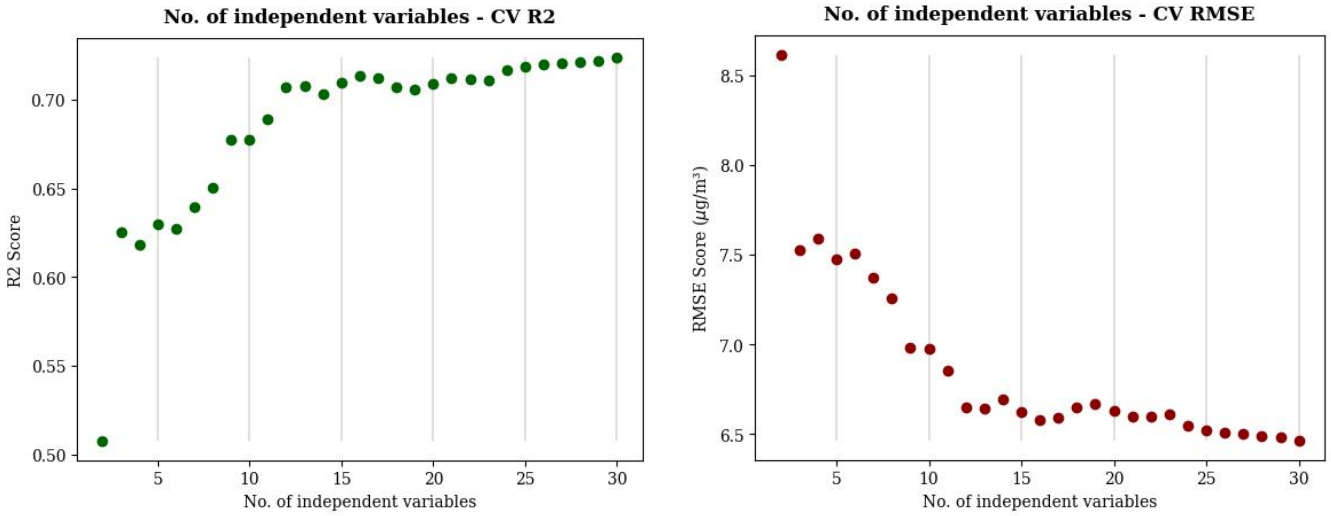
levels, compared to the average prediction ~~of~~ across the dataset (Shapley, 1953). The contribution of a feature is ~~calculated by examining the difference between~~ determined by comparing the difference in the response variable ~~that is obtained~~ when the feature is present ~~in comparison to the response variable that is obtained when the feature~~ versus when it is absent (i.e., marginal contribution) (Algaba et al., 2019; Shapley, 1953). The formula for calculating ~~the Shapley value~~ Shapley values can be found in ~~supplementary material, section equations. In our study, the~~ the supplementary materials.

In this study, feature selection is ~~based on an~~ guided by the out-of-sample performance ~~of~~ in a 10-fold ~~cross-validation, which~~ repeated random sampling validation, where Shapley values are calculated in each iteration ~~calculates shapely values in~~ of the random forest models. ~~The ranking of predictors is~~ Predictors are ranked based on the median ~~of the Shapley value in~~ Shapley value across all iterations. The relative positions of each predictor ~~for~~ using the median-based approach ~~can be found in supplementary material, figure 6. The~~ are illustrated in the supplementary materials (Figure 6), with the Shapley ranking of ~~one fold is shown in supplementary material, figure~~ a single fold shown in Figure 7. ~~To determine the preferred number of predictor variables, a~~ A random forest algorithm is applied ~~to each number of most influential features, based on the average median ranking, ranging from the~~ iteratively to determine the optimal number of predictors, starting with the two most influential predictors and extending to the thirty most influential features. ~~Thereafter, the RMSE and R^2~~ The RMSE and R^2 metrics are used to ~~determine~~ evaluate the optimal number of predictors~~used for modeling~~. The number of predictor variables and ~~the~~ their corresponding evaluation scores (R^2, RMSE) are shown in ~~figure~~ Figures 1a ~~, and figure~~ and 1b. ~~A remarkable prediction accuracy improvement is obtained~~ Notably, prediction accuracy significantly improves when considering at least twelve predictors~~. However, ~~, although the improvement is marginal ~~considering more than twelve predictors~~beyond this number.

Due to the ~~poor performance by the~~ random forest model~~over all the~~ 's poor performance across all local station measurements (~~supplementary material, figure 8 a-c) ,~~ Supplementary Figures 8a-c) and per spatial group (~~supplementary, table 2~~Supplementary Table 5), the random forest algorithm is ~~not applicable to identify the (number of)~~ deemed unsuitable for identifying the number of variables for the local models. ~~Rather, the~~ Instead, best subset regression is used for variable selection ~~for the~~ in local models. This approach ~~consists of testing~~ tests all possible combinations of predictor variables (Kassambara, 2018)~~. The maximum number of predictors considered is 30. ~~, with a maximum of 30 predictors considered.~~ The statistical criteria ~~considered are the~~ include adjusted R^2, Mallows CP, and Bayesian Information Criteria (BIC) scores. ~~Nine~~ As a result, nine features are identified for the local models.

2.4 Model comparison

In global modeling, comparisons are made ~~between~~ among tree-based models~~, namely, the ~~~~random forest, LightGBM, and XGBoost~~models; and the linear models, namely ~~~~and linear models~~~~LASSO and Ridge~~models. In ~~. For local modeling, ~~comparisons are made between~~ we compare linear models, ~~mixed effect~~ mixed-effect models, and kriging models~~; For every model , the spatial groups are compared in terms of~~ . Each model is evaluated based on R^2, RMSE, and MAE~~. These metrics are commonly applied ~~, which are standard metrics in the field (Rybarczyk and Zalakeviciute, 2018; Ameer et al., 2019; Chang et al., 2020). ~~Furthermore~~Additionally, the prediction patterns of the ~~local- and global mappings are examined. The mobile~~



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

(b) Number of features and corresponding RMSE score

Figure 1. Out-of-sample performance in ten-fold ~~cross-validation~~repeated random sampling validation: number of features and corresponding model performance (global).

~~NO_2 map (Kerckhoffs et al., 2019) of the same study area is used as a benchmark to further understand the performance of~~ ~~each model~~local and global models are analyzed. To benchmark the model performance, a mobile NO_2 map of the study area (Kerckhoffs et al., 2019) is used for comparison. Table 2 ~~shows~~ provides an overview of the global and local models, ~~relevant selected predictors in sequential order of importance, and relevant~~ along with selected predictors and evaluation methods. ~~Predictions based on the~~ The global models are ~~made for different areas with different~~ applied to areas with varying demographic characteristics, including two ~~big cities of more than~~ large cities with populations exceeding 700,000 ~~inhabitants,~~ (Amsterdam and Hamburg~~; a middle-sized city (~~), a mid-sized city with around 350,000 inhabitants ~~) Utrecht~~(Utrecht), and a small city ~~(around~~ with approximately 70,000 inhabitants ~~) Bayreuth~~(Bayreuth). Local model predictions are applied exclusively to Amsterdam. ~~Predictions derived from the local model were applied to Amsterdam only.~~ Table 3 ~~shows model complexity and the potential presence of a spatial component~~summarizes the complexity of the models and how spatial components are accounted for.

Table 2. 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 LightGBM LASSO Ridge	cross validation over the entir cross validation over different la comparing with Kerckhoffs et al
Local model	~~population_1000~~ nightlight_4950 nightlight_450 ~~nightlight_4950~~ road_class_3_100 ~~population_3000~~ trafbuf50 road_class_~~1_5000~~ 3_300 road_class_2_1000 road_class_2_5000 ~~road_class_3_100~~population_3000 road_class_~~3_300~~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 entir cross validation over different la comparing with Kerckhoffs et al

Table 3. 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~~ 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
LightGBM	Controlled by hyperparameters: number of estimators, alpha, lambda, learning rate, maximum depth, gamma	Not explicitly

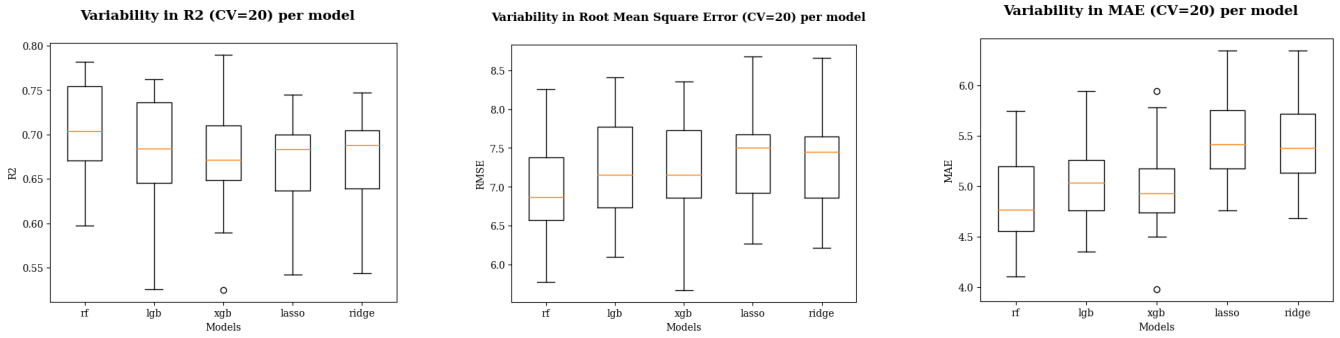
3 Results

3.1 Models

3.1.1 Global models

~~Evaluating the different linear~~ Evaluations of the different linear and non-linear models ~~is done by performing out-of-sample performances of 20-fold cross-validation, thereby examining~~ were carried out using repeated random sampling validation, performed 20 times. This approach enabled us to assess the variance and median statistics ~~per~~ for each model in terms of R^2, MAE, and RMSE (~~figure~~ Figure 2a, ~~figure~~ Figure 2b, and ~~figure~~ Figure 2c). ~~We found the 20 folds lead to stable estimations.~~ The repeated sampling provided stable estimates.

~~With~~ When comparing out-of-sample performances ~~of~~ via 20-fold ~~cross-validation~~repeated random sampling validation, the linear models (i.e.~~LASSO and Ridge) score similarly to~~ , LASSO and RIDGE) exhibited performances similar to those of



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

Figure 2. ~~The out-of-sample~~ Out-of-sample performances ~~of~~ evaluated using 20-fold ~~cross-validation~~repeated random sampling validation: ~~performance per model~~ (a) R^2, (b) RMSE, and (c) MAE~~(global)~~. ~~The upper-~~ Upper and lower ~~Quartiles~~ quartiles indicate variability~~:~~. RF = random forest, LGB = LightGBM, XGB = XGBoost.

the non-linear models, ~~especially the R^2~~ particularly in terms of R^2 and RMSE. ~~Generally~~Among the models, the random forest ~~model performs the best out of the considered global models according to the~~ consistently outperformed others, with the highest median R^2, ~~median~~ lowest RMSE, and ~~median~~ lowest MAE. The robustness of the random forest model is ~~relatively high too, this is implied by the lowest standard deviation of~~ further emphasized by its minimal standard deviation in R^2 and ~~RSME (figure~~ RMSE (Figure 2a and ~~figure~~ Figure 2b).

Accounting for spatial information

We ~~examine the differences between~~ further investigated the influence of spatial heterogeneity by comparing model performances across different spatial groups ~~when predicting~~ using the global model. ~~The descriptive statistics per spatial group also indicate interesting differences in terms of NO~~Descriptive statistics for NO_2 ~~concentration levels (table~~ concentrations in each spatial group reveal distinct differences (Table 4).

Table 4. Descriptive statistics of NO_2 ~~statistics~~ concentrations for each spatial group ~~, measured~~ (in $\mu g/$~~m3~~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
~~Low population~~ Suburban	138	27.601	9.769	7.872	19.876	26.876	34.407	56.706
~~far from roads~~ Rural	259	16.653	8.341	2.122	10.331	15.892	22.518	48.887

Table 5 ~~describes the performances, in terms of R^2~~details the performance metrics (R^2, RMSE, ~~and MAE,~~ MAE) for each spatial group~~per model. For each model, the observations far from roads perform considerably better than observations close to roads, for both urban and low-population groups. The non-linear models outperform the linear models when the data is trained~~

~~on observations in the groups "far away from roads" and "low population". For urban areas, the performances between the linear and non-linear methods are less distinguishable which might be explained by the relatively low number of observations~~.

315 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~~obtained poor prediction accuracy~~ , such as random forest, showed lower accuracy in urban areas, possibly due to the ~~relatively limited number of observations and heterogeneous character of data in the "urban" class~~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*~~ Mean	0.271	10.994	8.964	0.387	7.285	5.361	0.712	4.189	3.007
		~~*SD.*~~ SD	0.099	1.298	0.950	0.185	1.323	0.762	0.102	0.983	0.550
	LightGBM	~~*Mean*~~ Mean	0.175	11.631	9.477	0.367	7.381	5.468	0.725	4.075	2.872
		~~*SD.*~~ SD	0.145	0.226	0.955	0.226	1.513	0.739	0.120	1.040	0.537
	XGboost	~~*Mean*~~ Mean	0.228	11.230	9.147	0.426	7.060	5.228	0.737	3.991	2.774
		~~*SD.*~~ SD	0.150	1.014	0.807	0.183	1.340	0.687	0.116	1.096	0.530
Linear	Ridge	~~*Mean*~~ Mean	0.328	10.491	8.617	0.348	7.517	5.703	0.696	4.358	3.211
		~~*SD.*~~ SD	0.127	1.080	0.860	0.167	1.139	0.606	0.103	1.133	0.564
	LASSO	~~*Mean*~~ Mean	0.265	10.936	9.017	0.282	7.859	6.047	0.613	4.912	3.749
		~~*SD.*~~ SD	0.177	1.159	1.040	0.201	1.105	0.672	0.119	1.153	0.678

Table 5. Model performance per spatial group (CV = 20). RMSE and MAE are represented in NO_2 (μg/~~m3~~m^3)~~values~~.

Spatial prediction patterns

Figure 3 ~~shows the NO_2 spatial predictions for~~ presents the spatial predictions of NO_2 concentrations across the Amsterdam area ~~per model, a-c show the~~ for each model. Panels (a) to (c) depict the predictions from non-linear ~~spatial predictions , d and e show the linear techniques~~models, while panels (d) and (e) illustrate the results from linear models. Generally, ~~the linear models are shown to be more prone to overfittingas extreme values influence these prediction maps compared to non-linear techniques (i.e.~~linear models exhibit a higher tendency for overfitting, as their prediction maps are more influenced by extreme values (i.e., concentrations below 15 μg/~~m3~~ m^3 or above 50 μg/~~m3)~~ m^3) compared to the non-linear techniques. Interestingly, ~~linear techniques identify a high~~ the linear models identify a significant NO_2 ~~hot spot~~ hotspot in the southwestern part of the study area~~that is not identified~~ , which is not captured by the non-linear ~~techniques. Generally, high pollution~~ models. Across all models, however, elevated pollution levels are consistently observed along major roads ~~(highways, national roads) and~~ and in some urban areas~~(e.g. Haarlem) are obvious (supplementary, figure~~ , such as Haarlem (see Supplementary Figure 9).

Figures 4 show the spatial patterns of ~~the~~ predicted NO_2 concentrations for Hamburg (a~~and~~ , b), Utrecht (c~~and d)~~, d), and Bayreuth (e~~and f) for~~ , f) using the random forest and Ridge ~~models. The~~ Regression models. Predictions from other models (LightGBM, ~~XGboost~~XGBoost, LASSO) for ~~Hamburg, Utrecht, and Bayreuth (both zoomed in and out) can be found in supplementary sections figure~~ these cities, including both zoomed-in and zoomed-out views, are provided in the supplementary sections (Figures 10a-c, ~~figure~~ 11a-c, ~~figure~~ 12a-c, ~~figure~~ 13a-e~~respectively. Comparing~~).

Comparing the prediction maps of these cities ~~, there are~~ reveals noticeable differences in ~~prediction~~ spatial patterns. A ~~most important~~ key finding is that in Hamburg, the highest air pollution ~~seems to be situated~~ levels are concentrated around major

roads~~in Hamburg while~~ , while in Utrecht, the urban center ~~accounts for the highest air pollution concentrationsin Utrecht. The high~~ exhibits the highest NO_2 concentrations. This correlation between major roads and ~~high air pollution could~~ elevated air pollution in Hamburg can be reasonably explained ~~considering that Hamburg is the~~ by the city's high traffic congestion, as it ranks 69th ~~of~~ among the most congested cities ~~in the world~~ globally (Tomtom, 2021). Interestingly, ~~the highest~~ there are also spatial differences in the predicted NO_2 ~~concentration levels among highways differ in spatial patterns~~ concentrations along highways between the random forest and Ridge models~~, for example, the~~ . For instance, in Hamburg, the Ridge model predicts high NO_2 levels along highways in the southeastern and western ~~part of the Hamburg area contain high NO_2 levels for the Ridge models while a nuanced identification is related to the random forest prediction for the same area. In the random forest prediction map for Hamburg, air pollution among~~ 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 ~~center and northern part of the city is more pronounced~~ central and northern parts of Hamburg, compared to the Ridge model~~equivalence. Additionally~~.

Furthermore, the magnitude of high ~~air pollution~~ pollution levels related to major roads is ~~considerably higher for Hamburg , compared to~~ significantly greater in Hamburg than in Utrecht and Bayreuth. ~~Still~~Nevertheless, the relationship between ~~the presence of roads and heavier air pollution concentration is identifiable for~~ road presence and higher air pollution levels is evident in both Utrecht and Bayreuth, ~~especially with the Ridge model predictions. For~~ particularly in the predictions from the Ridge model. In Utrecht, the urban center is more ~~pronounced in terms of~~ prominently identified as a high NO_2 concentration ~~levels,~~ area compared to Hamburg and Bayreuth. ~~Moreover~~Additionally, the Ridge model ~~applied to Utrecht identifies more clusters (i.e. scattering) of~~ for Utrecht shows more clusters of elevated NO_2 ~~values~~ levels in the periphery~~. In comparison, the predicted NO~~, whereas the random forest model predicts a more scattered distribution of NO_2 ~~values are more scattered~~ concentrations in the urban center~~for the random forest when compared to the Ridge model. Again, this difference in prediction patterns between a linear and non-linear model is apparent for~~ , similar to the pattern observed in the Amsterdam area. ~~Bayreuth~~

Bayreuth, on the other hand, is characterized by moderate ~~air pollution and very low~~ pollution levels, with very low NO_2 concentrations (<15 μg/~~m3) pollution in~~ m^3) in the rural areas surrounding the city~~- some clusters~~ . However, some clusters of higher NO_2 levels exceeding the 15 μg/~~m3 benchmark are noticeable that correspond to other villages in the area, hinting to the influence of~~ m^3 benchmark are observed in the vicinity of other villages, suggesting that population or building density ~~on air pollution~~ may influence air pollution levels in these areas (see also ~~supplementary, figure~~ Supplementary Figures 13a-e). Supplementary ~~, figure~~ Figure 14 ~~shows the~~ provides a distribution of predicted ~~NO2 per global model for each~~ NO_2 concentrations for each global model and location.

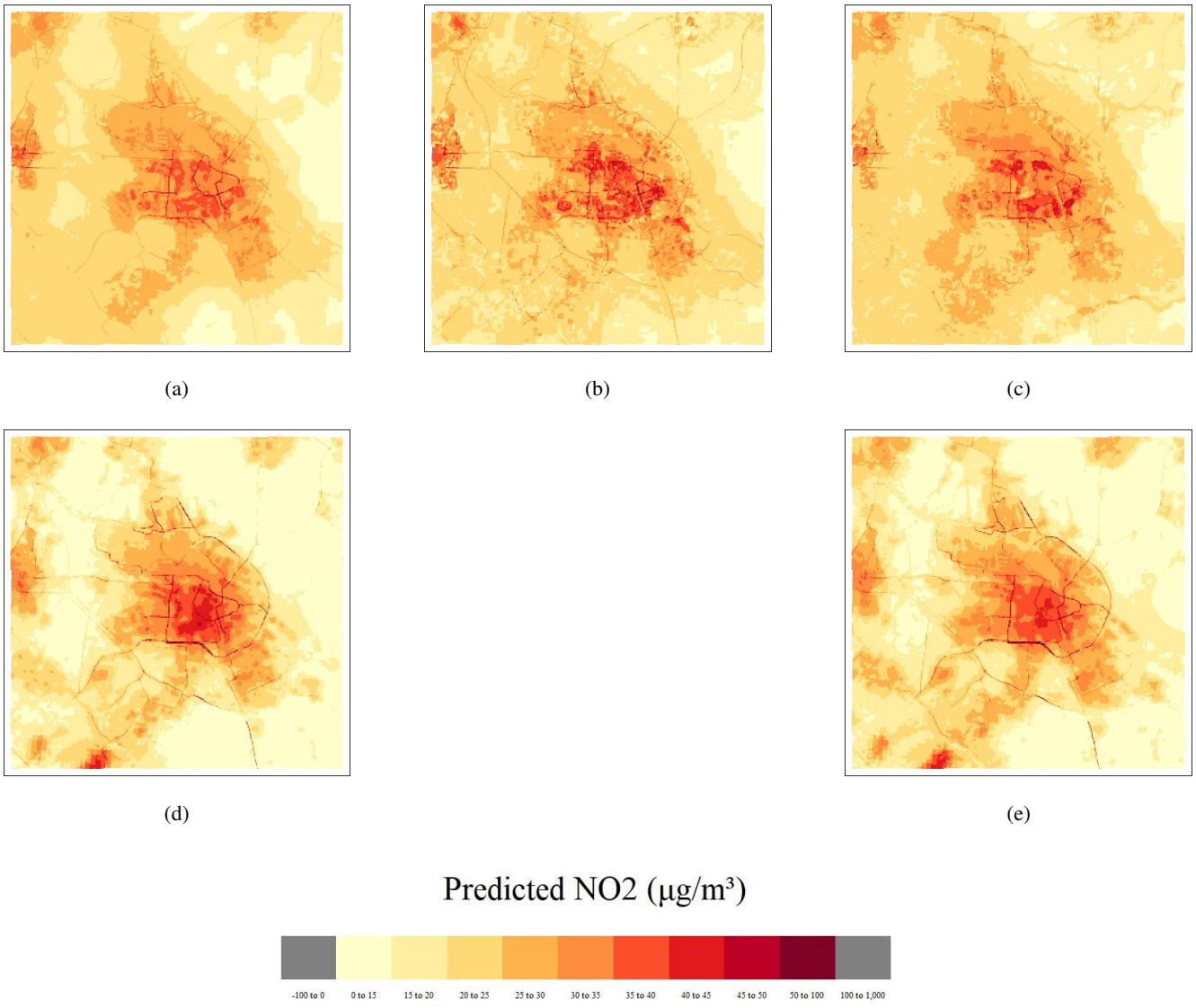


Figure 3. Spatial patterns of predicted NO_2 (100m), measured in μg/m3, per model for Amsterdam - non-linear models (top): (a) = random forest, (b) = LightGBM, (c) = ~~XGboost~~XGBoost; linear models (bottom): (d) = LASSO, (e) = Ridge. Extent = 30km x 30km

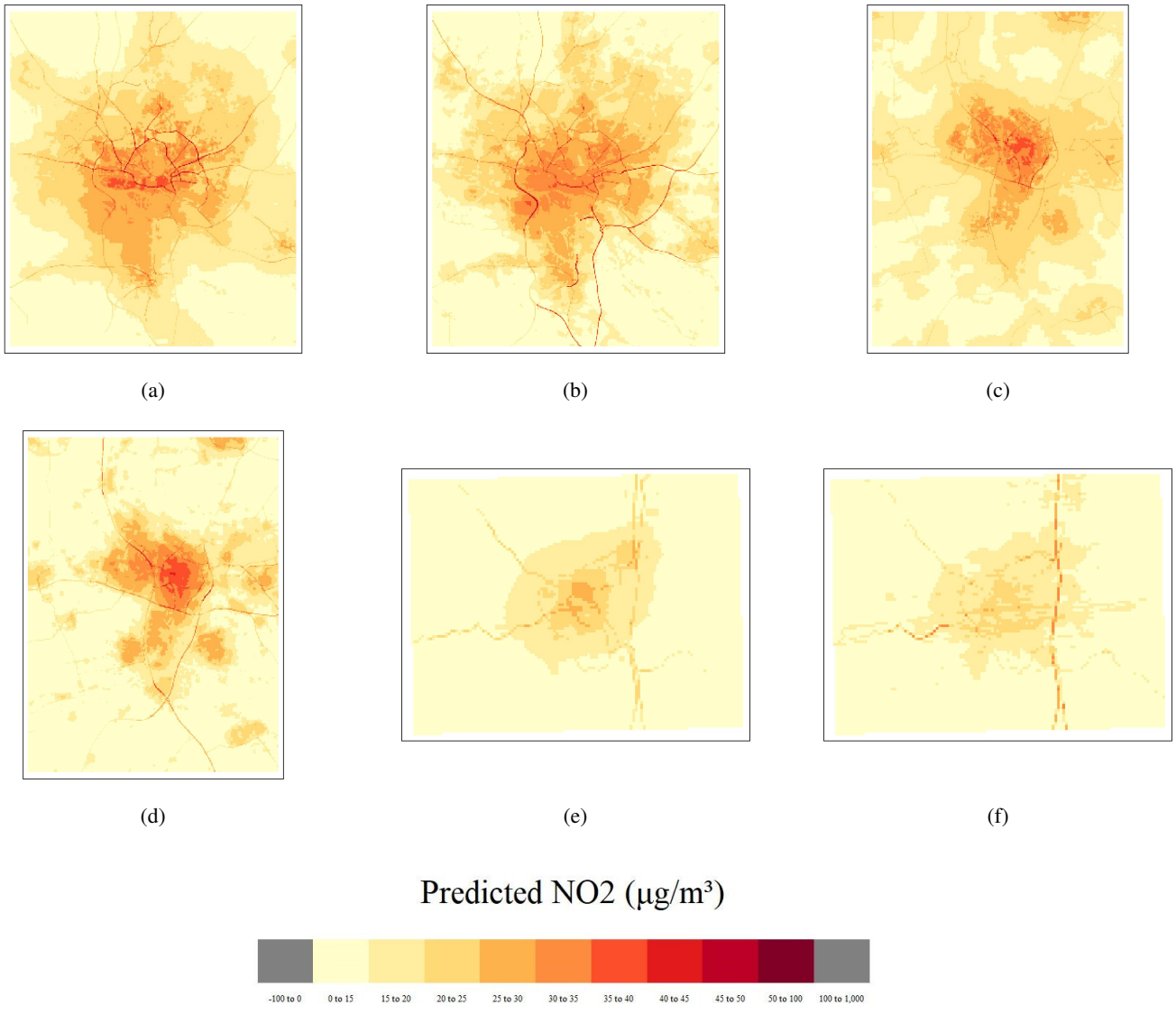


Figure 4. Spatial patterns of predicted NO_2 (100m), measured in μg/m3, per model for Hamburg (extent = 30km x 30km), Utrecht (extent = 25km x 25km) and Bayreuth (extent = 10km x 10km) - top: from left to right, random forest (Hamburg), Ridge (Hamburg), random forest (Utrecht); bottom: from left to right, Ridge (Utrecht), random forest (Bayreuth), Ridge (Bayreuth)

3.1.2 Local models

~~The performances~~

The performance of the local models ~~are composed of the~~ was assessed using R^2, RMSE, and MAE metrics. Table 6 ~~shows the model performances for~~ summarizes the performance of the linear model, ~~the~~ mixed-effects model, ~~the~~ ordinary kriging model, and ~~the~~ universal kriging model, ~~whereby the~~ all evaluated using leave-one-out cross-validation~~is applied. The~~ . Among these, the ordinary kriging model ~~shows~~ exhibits the poorest performance. ~~The~~ Figure 5 illustrates the spatial prediction patterns ~~are shown in figure 5). The~~ for each model. Notably, the universal kriging model ~~performs considerably better than the ordinal kriging model . The~~ outperforms the ordinary kriging model significantly. However, the simple linear model ~~outperforms~~ surpasses the universal kriging method in terms of prediction accuracy. ~~Accounting for~~ Incorporating spatial groups as random effects ~~yields~~ in the mixed-effects model leads to a higher R^2, ~~a lower RMSE , and a lower MAE~~and lower RMSE and MAE, indicating improved model performance.

Table 6. Model Performance Using Leave-One-Out Cross-Validation

	R^2	RMSE ($\mu g/$~~m3~~m^3)	MAE ($\mu g/$~~m3~~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
~~UK~~ universal kriging (model + kriged residuals)	0.277	7.749	6.097

Table 7 ~~shows the model results per~~ provides model performance metrics for each spatial group, again ~~based on~~ using leave-one-out cross-validation. ~~Similar to the results of the global model , the resultsfor the local models indicate that models trained on "urban " observations~~ Consistent with the global model results, local models trained on urban observations tend to perform poorly. ~~However, the proximity to the road~~ Interestingly, proximity to roads does not necessarily ~~influence the model performancesince the~~ correlate with model performance, as the suburban group exhibits a higher R^2 ~~of the "low population" class is higher than the R^2 of the "far from roads" class. In contrast to the models trained on the global dataset~~than the rural group. Contrary to global models, which perform best in ~~"far from roads" areas, the models trained on the local dataset perform the best in areas with low populations and proximity to roads. A plausible explanation is~~ rural areas, local models achieve their best performance in suburban areas. This difference may stem from the fact that observations in ~~"far from roads" areas for~~ rural areas within the local dataset are more similar to ~~observations in the urban and low-population areas when compared to~~ those in urban and suburban areas than in the global dataset, ~~as the predictor valuesare distributed more uniformly in the local dataset~~due to a more uniform distribution of predictor values.

Table 7. Model Performance Per Spatial Group (CV = Leave-One-Out Cross-Validation). RMSE and MAE in $\mu g/m^3$

	Urban			Suburban			
Models	R^2 ^{2}	RMSE	MAE	R^2 ^{2}	RMSE	MAE	R^2
ordinary kriging	0.072	8.257	6.772	0.223	8.558	6.575	0.0
linear model	0.140	7.890	6.360	0.509	6.8 6.800	5.301	0.1
mixed-effects model	0.141	7.874	6.316	0.524	6.505	5.298	0.1
UK (model + kriged residuals) 0.161 universal kriging *(model + kriged residuals)*	8.068	6.27 6.270	0.487	6.938	5.174	0.037	

Spatial prediction patterns

Figure 5 shows displays the predicted NO_2 patterns based on the local dataset. The prediction map of for the linear model (a) is fairly similar to the prediction maps of the quite similar to those for the mixed-effects (c) model and universal kriging (e) model: the modelsidentify models, with all identifying a high NO_2 concentration cluster at in the northwestern part of Amsterdam. Further examination reveals analysis suggests that this cluster is likely highly influenced by the predictor "road class 2 5000" (i.e., primary roads within 5000m), as this predictor shows exhibits a similar cluster at in the same location (supplementary, figure see Supplementary Figures 15, figure 16a-i). Two

The two models that account for the spatial groups first, spatial groups before the modeling process , show comparable patterns whereby (mixed-effects and universal kriging) display comparable patterns where the influence of roads is obvious via evident, either through the predictors themselves or the spatial groups groupings (see also supplementary, figure Supplementary Figure 17). The relative relatively low NO_2 values along the roads in the outer Amsterdam area can be attributed to the spatial grouping divisions. To extend, the presence of predictor values with high standard deviations can impact the NO_2 values for that 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 parts of the prediction area. The patterns that are along the roads belong to the spatial group "low population" whereby observations within this group are in the vicinity of roads(<100m). Comparing this spatial group to the spatial group"far from roads"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 every predictor in low population each predictor in the suburban group is substantially differentthan the data distribution for every predictor in the group "far from roads", leading to different learning patterns which explain the relative distinct learning patterns that explain the relatively high prediction values along the roads (supplementary, figure roads (see Supplementary Figures 18a-i). At some placesIn some instances, negative predicted values are apparent albeit few. This is likely a cause of the training dataset having different feature characteristics than the testing dataset. Comparing the observed, albeit rarely. These may result from discrepancies in feature characteristics between the training and testing datasets.

Comparing local prediction patterns to ~~the~~ global prediction patterns ~~,~~ reveals that the local models identify a cluster of high air pollution in the northwestern part of Amsterdam ~~is visible in some local models that is not visible in~~ that the global models ~~(as discussed, these models refer to the general linear-, mixed-effects-, and universal kriging models). A possible explanation for why the cluster is identified in the local dataset, as opposed to the global dataset, could be the difference in~~ do not detect. This discrepancy could be due to differences in the spatial distribution of NO_{2} ~~$_{2}$~~ values between the ~~local-~~ local and global datasets, ~~resulting in different learning patterns between the local- and global models (figure ??). Supplementary, figure 19 shows the distribution of predicted NO2 per local model.~~ leading to distinct learning patterns in the respective models (Figure **??**). Moreover, Figures 3 and 5 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.

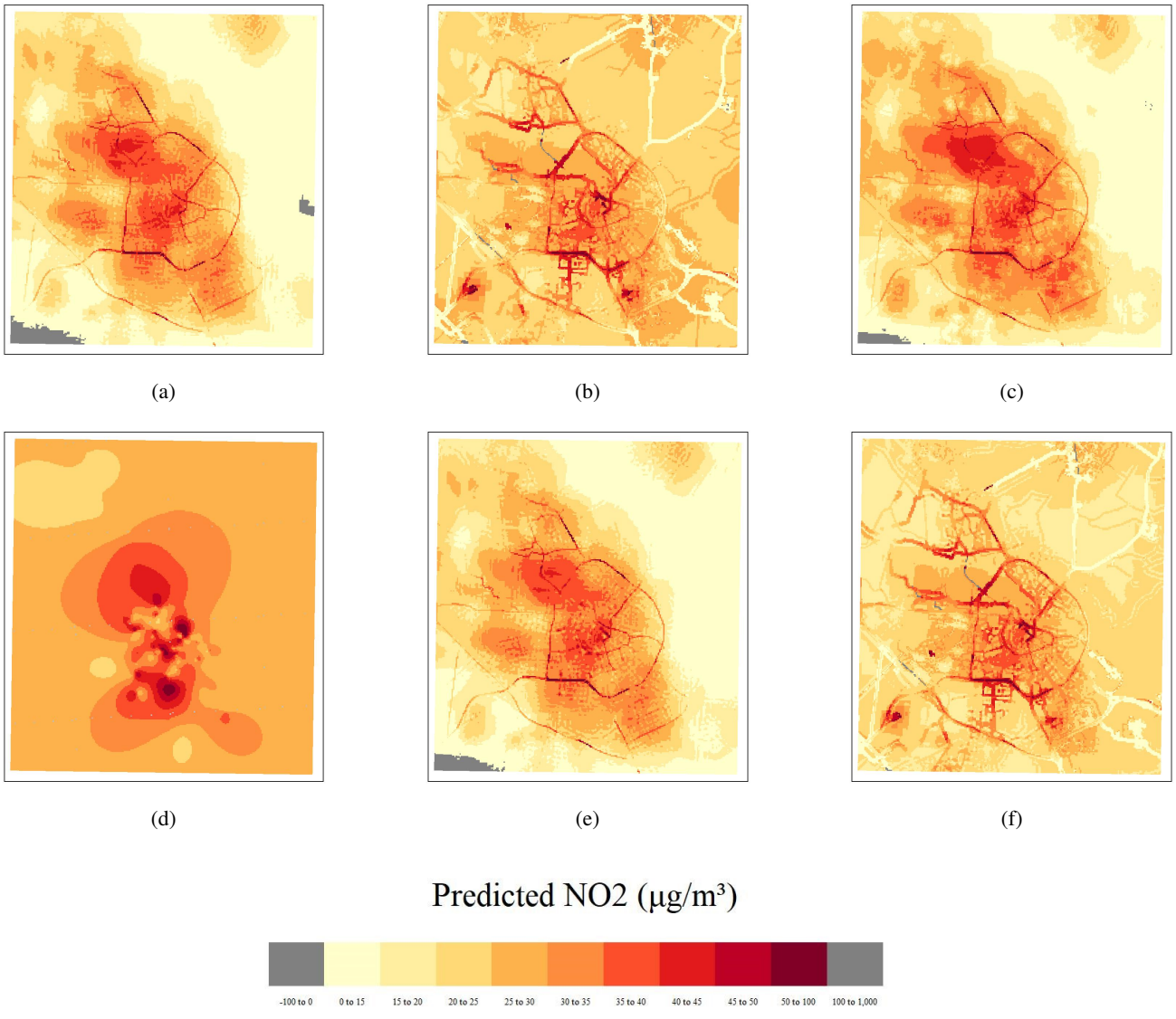


Figure 5. Spatial patterns of predicted NO_2 (μg/m3) at 100m resolution based on the local dataset - top: left = linear model, middle = linear model separating for spatial groups, right = mixed-effects model; bottom: left = ordinary kriging, middle = universal kriging, right = universal kriging separating for spatial groups

Model comparison

Figure 6 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 . Some extreme values of predictions are present in the local linear model that accounts for spatial groups and the universal kriging model that accounts for spatial groups, resulting in a relatively low correlation with the open NO_2 dataset. We removed these extreme values to further understand the correlations between the mentioned models and (supplementary figure 20). To improve the clarity of the open NO_2 dataset. A 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 is chosen as the upper boundsince this is , based on the maximum value of observed across the ten models (excluding the two where the outlier detection is outlier detection was applied first). The lower bound is set to was set at 0. The correlation matrix with extreme prediction these extreme predictions filtered out is visible in supplementary , figure 20. shown in supplementary figure 21. The global models are highly correlated, with the LASSO model being the least correlated with other global models. The correlations between ordinary kriging model and other models are also 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. Comparing the models to the open NO_2 dataset, the local models generally show more similarity than global models. This is not surprising as the local model dataset is also from Amsterdam. Table 8 shows the residuals per global and local model. The ridge 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. LightGBM and 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 materials figure 22a-e and 23a-f respectively.

Table 8. 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
LightGBM	Global	0.56	2.29	8.76	-55.83	22.85
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

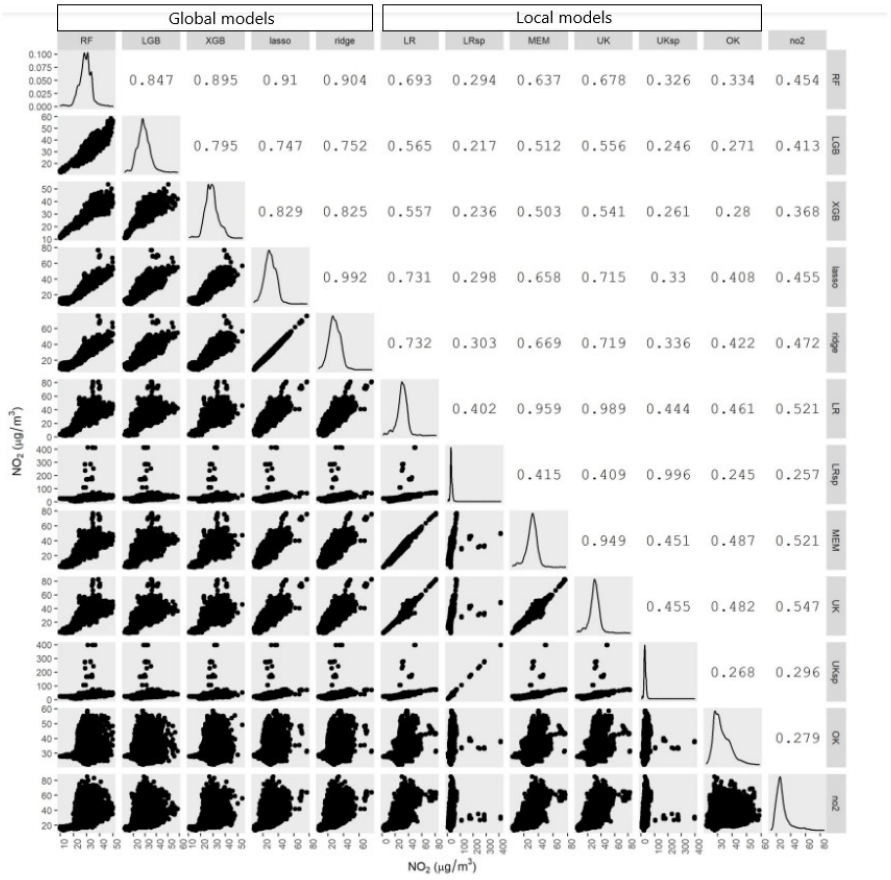


Figure 6. Comparing model predictions whereby the numbers equal the Pearson correlation coefficient. RF: Random Forest, LGB: LightGBM, XGB: XGBoost, LR: linear regression, LRsp: Linear Regression accounting for spatial groups, MEM: Mixed-Effects Model, UK: Universal Kriging, UKsp: Universal Kriging accounting for spatial groups, OK: Ordinary Kriging, no2: mobile NO_2 map.

4 Discussion

~~While several studies~~ Several studies have applied statistical modeling to ground station measurements and geospatial predictors for NO_2 ~~modeling and mapping, the influence~~ mapping, but the impact of spatial heterogeneity has ~~not been discussed in detail. This is pursued in our studythrough model comparisons between using~~ often been overlooked. In this study, we address this gap by comparing spatial and non-spatial ~~techniques and at~~ modeling techniques across different spatial scales. Below~~we discuss important findings of this study and~~ , we discuss the key findings and provide our perspectives.

Relationship between predictors and other pollutants

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

Accounting for spatial groups

~~For the global dataset, the differences between the~~ 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. While a well-trained global model can make accurate predictions where global predictors are available, it may perform poorly in regions where predictor values differ significantly from the training data.

In this study, the differences between linear and non-linear techniques are ~~negligible based on accuracy assessment matrices. The random forest models obtained the~~ minimal when applied to the global dataset. Although the random forest model generally performs best (highest R^2 ~~and the~~ , lowest MAE) ~~but~~ , the R^2 of the ~~Ridge~~ RIDGE model is higher than that of the LightGBM and XGBoost models. ~~When accounting for the spatial groups- urban, low population, and far from roads -~~ However, when accounting for spatial groups—urban, suburban, and rural—the differences in model performance between linear and non-linear techniques become more ~~distinguishable, whereby the~~ pronounced, with non-linear models ~~perform better for observations far away from roads, where data generally is~~ generally outperforming linear models, particularly in rural areas where data are more homogeneous. ~~However, in urban areas, the model performances between linear and~~ This finding confirms with studies by Weichenthal et al. (2016), Reid et al. (2015), Chen et al. (2019), and Lu et al. (2020), which suggest that non-linear ~~models are less pronounced, and the model performances are all unsatisfactory. These findings are against the conclusions of several studies~~ techniques typically provide better predictions. Our results support the argument by Meyer and Pebesma (2021) that non-linear models ~~could achieve better predictions ((Weichenthal et al., 2016), (Reid et al., 2015) , (Chen et al., 2019), (Lu et al., 2020)). In this study, we used~~ perform better in areas where the environmental variables are similar to those in the training data.

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 ~~instead of spatial cross-validation, following the arguments regarding spatial and~~ . According to Wadoux et al. (2021), standard cross-validation ~~in Wadoux et al. (2021) and Lu et al. (2023).~~ (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).

~~The heterogeneous data nature of urban areas renders poor statistical modeling performance , which has gone unnoticed in studies not controlling for spatial heterogeneity. The~~ 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 ~~worrisome given that~~ concerning, as the impact of air pollution ~~can depend on the surrounding environment, i.e. people who live in the vicinity of~~ is often more severe in these regions due to proximity to traffic-heavy ~~roads (which are often more present in , or around urban areas) and/or industries facing higher exposure to air pollution (He et al., 2022). Though spatial grouping greatly improves the predicting~~ roads and industrial areas (He et al., 2022). While spatial grouping improves predictive reliability, it can ~~present counter-intuitive patterns. For instance, in some areas, the~~ lead to counterintuitive patterns, such as lower predicted NO_2 ~~concentration levels are lower along roads than the concentration levels of the rural surroundings.~~ concentrations along roads compared to surrounding rural areas. Additionally, adjusting the threshold for defining "urban" in the local dataset from 0.75 to 0.5, due to higher population density and fewer samples, was necessary to more accurately represent urban areas. This adjustment, while affecting the classification of "urban" areas, is crucial for improving model relevance and accuracy in high-density regions. Therefore, while spatial grouping enhances prediction reliability, the definition of "urban" varies between datasets and can influence model performance and interpretation.

Moreover, Patelli et al. (2023) identify three main categories for integrating random forests with spatial data: pre-processing, in-processing, and post-processing. In our study, the link between random forest performance and spatial groups can be considered a form of post-processing. However, there is potential for better integration of spatial data into ensemble tree-based models, such as random forests, to further improve predictive performance (Patelli et al., 2023).

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.

~~*Spatially varying on*~~ *Spatial variation in feature importance*

While ~~the feature importance is equal between~~ feature importance may be consistent across cities, the influence of specific predictors on NO_2 concentrations ~~differs between the citiesstudied. For instance~~can vary significantly between cities. For example, building density and population are more ~~prevalent~~ significant contributors to air pollution in Utrecht, ~~compared to Hamburg, while~~ whereas traffic has a ~~higher influence~~ greater impact on high NO_2 concentrations in Hamburg~~, compared to~~

Utrecht. Additionally, global models are applied to different cities . Applying global models with the same predictors . As the case cities unravel that high NO_2 can be attributed to different predictors per city , applying models with different features may yield better prediction results. An important condition is that every city has enough 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 a problem in fitting complicated models. Outliers are omitted after the model predictionsto deal with unreliable predictionschallenges for fitting complex models. To address unreliable predictions, outliers were removed after model predictions. Transforming the original data could avoid data out of potentially avoid predictions falling outside the plausible range (e.g.<, below 0 mgμg/m^3). In our However, in this study, such transformation, e.g., transformations, like a log transformation, is were not applied. Airborne Although airborne pollutant concentrations are often positively skewed (Maranzano et al., 2020). However, Lu et al. (2023) examined several techniques such as transformations, likelihood functions, and loss functions to address the issue of non-Gaussian distributions but suggested the , Lu et al. (2023) found that the best modeling results were obtained without data transformation and using Gaussian likelihood(i.e. instead of using e.g. a Gamma likelihood, which matches the best with the data distributionin the study). , even when other distributions like Gamma might better match the data distribution. Moreover, while the LASSO and Ridge models seem appear useful with the global dataset, the predictions are unsatisfactory their predictions were less satisfactory with the local dataset. In this study, traffic volumes are a prevalent feature, however, no distinction is made between traffic types (e.g. were a significant feature, yet no distinction was made between different types of traffic (e.g., cars, buses, trucks), car vehicle types (e.g., electric, diesel), and engine typeswhile such aspects or engine types, all of which are known to be influential to influence air pollution (Wong et al., 2021). For instanceexample, distinguishing between vehicle types may show that relatively many trucks are on specific roads(e.g. going 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 certain localized clusters of high NO_2 concentrations. Further studies may attempt to integrate spatial dependence in random forest (Patelli et al., 2023)Future studies could explore integrating spatial dependence into random forest models (Patelli et al., 2023) to potentially enhance predictive performance.

5 Conclusions

In this study, we understand the spatial heterogeneity heterogeneity of NO_2 modeling through 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 ~~areas far from roads and in areasnear roads with low population density~~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. We also found that modeling the spatial autocorrelation does not improve the local modeling accuracy, but modeling spatial groups does. Lastly, prediction patterns show that ~~nonlinear~~ non-linear models are less prone to overfitting compared to 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.

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.

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

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

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

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

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

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

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

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

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

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

OpenAQ: Fighting air inequality through open data, 2017.

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

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

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

Ren, X., Mi, Z., and Georgopoulos, P. G.: Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States, Environment international, 142, 105 827, 2020.

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

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

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

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

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

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.