geographical analysis

Check for updates

Geographical Analysis (2023) **55**, 703–727

A Comparison of Spatial and Nonspatial Methods in Statistical Modeling of NO_2: Prediction Accuracy, Uncertainty Quantification, and Model Interpretation

Meng Lu[1], Joaquin Cavieres[2], and Paula Moraga[3]

[1]Department of Geography, University of Bayreuth, Universitaetsstrasse 30, Bayreuth, 95447, Germany, [2]Instituto de Estadística, Facultad de Ciencias, Universidad de Valparaíso, Valparaíso, Chile, [3]Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia

NO_2 is a traffic-related air pollutant. Ground NO_2 monitoring stations measure NO_2 concentrations at certain locations and statistical predictive methods have been developed to predict NO_2 as a continuous surface. Among them, ensemble tree-based methods have shown to be powerful in capturing nonlinear relationships between NO_2 measurements and geospatial predictors but it is unclear if the spatial structure of NO_2 is also captured in the response-covariates relationships. We dive into the comparison between spatial and nonspatial data models accounting for prediction accuracy, model interpretation and uncertainty quantification. Moreover, we implement two new spatial and a nonspatial methods that have not been applied to air pollution mapping. We implemented our study using national ground station measurements of NO_2 in Germany and the Netherlands of 2017. Our results indicate heterogeneous levels of importance of modeling the spatial process in different areas. The prediction intervals predicted with ensemble tree-based methods are more satisfactory than the geostatistical methods. The two new methods implemented each obtained better prediction accuracy compared to the original ensemble tree-based and stacking methods. The probabilistic distribution of the spatial random field estimated by the geostatistical methods could provide useful information for analyzing emission sources and the spatial process of observations.

Introduction

NO_2 is a traffic-related air pollutant highly dynamic over space. Detailed spatial mapping of NO_2 is needed in health cohort studies to understand the long-term health effects of NO_2 on individuals. Statistical methods for NO_2 mapping have attracted a lot of attention with the

Correspondence: Meng Lu, Department of Geography, University of Bayreuth, Universitaetsstrasse 30, 95447 Bayreuth, Germany. Email: meng.lu@uni-bayreuth.de

Submitted: August 31, 2021. Revised version accepted: November 23, 2022.

doi: 10.1111/gean.12356

burgeoning Machine Learning (ML)[1] methods and availability of ground monitoring station networks, atmospheric satellite products, and spatial data of our environment and atmosphere. Geospatial predictors of NO_2 factors relevant to this study and considered as covariates in air pollution models include the following. First, emission-related variables, like road networks, describe where and how NO_2 is generated. Second, dispersion-related variables represent where NO_2 goes after being generated, like wind speed and direction, tell us how the emitted NO_2 will drift once emitted. In addition to these process-based measures, outputs from physics-based numerical models and satellite measurements of atmospheric conditions have been shown to be useful in statistical modeling of NO_2. Numerical models are generally very different from statistical models as they are mechanism-based simulators of a physical system. Satellite measurements do not directly reflect surface NO_2 concentrations, but could provide important spatial information of NO_2. For example, since August 2019, the "Tropomi" instrument onboard the Sentinel 5p mission satellite provides the highest-resolution of NO_2 column density yet available, with 3.5 km by 5.5 km pixels across the satellite track.

Statistical methods for spatial air pollution prediction can be broadly classified depending on whether the spatial dependency is explicitly modeled. If not modeled, we refer to the methods "non-spatial" and otherwise "spatial." Most of the spatial air pollution models were developed to predict at coarser resolutions, commonly one kilometer or coarser (Young et al. 2016; Shaddick et al. 2018; Beloconi and Vounatsou 2020). Nonspatial methods are more dominant in air pollution mapping, particularly in high-resolution (100 m resolution or higher) mapping. Among them, LUR (Land Use Regression) models which assume linear relationships between air pollution observations and geospatial predictors are the most studied (Briggs et al. 2000; Hoek et al. 2008). Most recently, statistical learning[2], including regularized linear regression such as Lasso and Ridge regression (James et al. 2013), support vector machine (Suykens and Vandewalle 1999), ensemble tree-based methods such as Random Forest (RF, Breiman 2001) and XGBoost (XGB, Chen and Guestrin 2016), have been applied for feature selection and to capture the nonlinear response–covariate relationships (Chen et al. 2019a; Lu et al. 2020a). In air pollution mapping, several studies compared statistical learning and conventional LUR methods (Rybarczyk and Zalakeviciute 2018; Kerckhoffs et al. 2019; Chen et al. 2019a; Ren, Mi, and Georgopoulos 2020; Lu et al. 2020a).

Geostatistical models and Geographically Weighted Regression (GWR) are the most used spatial methods for air pollution prediction (Vicedo-Cabrera et al. 2013; Li et al. 2014; Zou et al. 2016; Wang et al. 2021) and these methods have been combined with dimension reduction (Zhai et al. 2018) and RF (Zhan et al. 2018; Liu, Cao, and Zhao 2020) to improve NO_2 prediction accuracy. A Bayesian geostatistical model is developed in Beloconi and Vounatsou (2020) to predict NO_2 by integrating Tropomi satellite instrument measurements and chemical transport

[1]List of abbreviations: CRPS: Continuous Ranked Probability Score; CV: Cross Validation; DF: Distributional Forest; GRF: Gaussian Random Field; GMRF: Gaussian Markov Random Field; GAMLSS: Generalized Additive Models for Location Scale and Shape; INLA: Integrated Nested Laplace Approximation; IQR: Interquartile range; GWR: Geographic Weighted Regression; KED: Kriging with external drift; LUR: Land Use Regression; MAE: Mean Absolute Error; ML: Machine Learning; RF: Random Forest; OMI: Ozone Monitoring Instrument; Quantile Random Forest; RMSE: Root Mean Squared Error; SE: stacked ensemble; SPDE: Stochastic Partial Differential Equations; Tropomi: Tropospheric monitoring instrument; UK: Universal Kriging (UK); OMI (Ozone Monitoring Instrument) VIIRS: Visible Infrared Imaging Radiometer Suite; XGB: XGBoost

[2]In this study, "statistical learning" is used interchangeably with "machine learning" methods (Hastie, Tibshirani, and Friedman 2009).

models. A GWR model naturally models spatial varying coefficients by fitting multiple local regressions depending on the homogeneity in response–covariate relationships when a number of observations are involved. A typical geostatistical model can be viewed as consisting of two components: a mean function, commonly a linear model, capturing the response–covariate relationships and a covariance function modeling dependency of residuals from the mean (Bhatt et al. 2017). Conventional Kriging methods suffer from the "big n problem," that is, it may become computationally intractable with a large number of observations. To deal with this problem, Lindgren, Rue, and Lindström (2011) propose to use Stochastic Partial Differential Equations (SPDE) to approximate the Gaussian Random Field (GRF) by a Gaussian Markov random field (GMRF, Rue and Held (2005)). The main advantage of this approach is that the GMRF has a sparse structure of the precision matrix which is the inverse of the covariance matrix of a GRF. The SPDE approach can be used in combination with the Integrated Nested Laplace Approximation (INLA, Rue, Martino, and Chopin 2009) in a Bayesian framework to achieve computational scalability of a geostatistical model by using approximations for all the estimations. This is especially advantageous when modeling NO_2 over a larger scale, for example, continental or global-scale modeling when a large amount of observations are modeled, and in spatiotemporal settings.

As spatial models are typically more complex compared to their nonspatial counterparts, several studies compared spatial and nonspatial models to understand if the spatial effects could be simply modeled by including certain covariates in LUR models. Young et al. (2016) studied the use of universal Kriging (UK), OMI (Ozone Monitoring Instrument) satellite instrument (Earthdata) and LUR models for NO_2 prediction at 2.5 km resolution. Young et al. (2016) indicated that either using UK or adding OMI in the LUR model improves a LUR model but adding OMI in a UK model only trivially improves the performance. Bertazzon et al. (2015) show that the inclusion of the meteorological variables accounts for spatial effects similarly to the use of spatial autoregressive models (Anselin et al. 2001). However, even if the spatial dependency can be captured by involving certain covariates in a LUR model, we may still need geostatistical methods to understand the spatial structure present in the data. Linear models have been used for the mean function but the relationships between NO_2 and predictors have been shown to be better modeled with nonlinear ML methods (Lu et al. 2020a). Most recent studies attempt to replace the linear mean function with ML models. Liu, Cao, and Zhao (2020) applied a spatial model to the residuals from an RF model for the spatial prediction of $PM_{2.5}$. Bhatt et al. (2017) propose to stack ML models to replace the mean function in a spatial model and applied the method to disease mapping.

Few studies have compared geostatistical and ML models, possibly because the ML models are still relatively less studied in air pollution mapping and in the field of geostatistics. It might be more interesting to compare geostatistical and ML models than geostatistical models and LUR, because ML models may be more capable of capturing the spatial dependency by integrating covariates, though implicitly, when the number of observations is sufficient. Moreover, most comparison studies only compare the Cross-Validation (CV) accuracy of the prediction mean, ignoring the confidence and prediction intervals, or the probability distributions of the parameters and predictions. If correctly derived, narrower intervals would be preferred. Also not discussed is the cause of the prediction errors, are they caused by missing covariates, violation of the model assumptions (e.g. data distribution, nonlinearity), or inconsistent distributions between training and validation sets. Also, different CV strategies, for example, how do we split the train-test sets, may lead to different model validation results. Current studies commonly ignore this problem

and did not discuss the consequence of applying k-fold splitting (Larkin et al. 2017; Kerckhoffs et al. 2019; Ren, Mi, and Georgopoulos 2020) or bootstrapping (Lu et al. 2020a). These train-test splitting methods also do not provide an indication of accuracy in spatial blocks but only at the locations of observations.

In this study, we focus on ensemble tree-based algorithms (e.g. RF and boosting) in the nonspatial modeling category and a hierarchical spatial model (Blangiardo and Cameletti 2015; Lindgren, Rue, et al. 2015; Moraga 2019) called latent Gaussian model in the spatial modeling category. Additionally, we invest in stacked models for integrating ML and geostatistical models. This model is treated as a spatial model and is to explore if, with ML methods to estimate the mean of a geostatistical model, it could obtain both the merits of spatial and nonspatial models. Lastly, we also develop a LUR model using Lasso as a base model for comparison.

Uncertainty is commonly quantified by confidence intervals, or credible intervals in Bayesian inference, for the estimated parameters and by prediction intervals for the predictions from the model. If the credible intervals could be estimated, we could estimate the prediction intervals. For the methods that are used in a nonparametric setting, only the prediction intervals are quantified. For parametric models, we quantified both the prediction intervals and the confidence interval.

Ensemble trees are nonparametric models, deriving prediction intervals is therefore less straightforward than a parametric model (e.g., a linear regression model) but has been studied and shown satisfactory results with simulated data. Prediction intervals have been most well studied for RF (Meinshausen 2006; Stasinopoulos, Rigby, et al. 2007; Wager, Hastie, and Efron 2014; Alakus, Larocque, and Labbe 2021) and more recently for boosting (Duan et al. 2020; Velthoen et al. 2021). Comparing probabilistic methods (i.e., prediction interval calculation) of RF and boosting is beyond the scope of this study and we focus on prediction intervals derived for RF to compare with geostatistical methods. Possibly, one of the most widely recognizable methods to derive RF prediction intervals is Quantile Random Forest (QRF) (Meinshausen 2006). QRF has been shown to estimate middle quantiles well but may fall short at the extremes due to the limited number of observations in the tail regions (Velthoen et al. 2021). Velthoen et al. (2021) proposed to use extreme quantile regression to estimate for data outside the range of observations. Another well-recognized method is distributional regression forests (DF) (Schlosser et al. 2019), which embeds the GAMLSS (Generalized Additive Models for Location Scale and Shape) (Stasinopoulos, Rigby, et al. 2007) into RF.

Fouedjio and Klump (2019) compared prediction accuracy and uncertainty quantification between KED (Kriging with external drift) and QRF by simulating data with various levels of spatial dependency. It concluded that an optimal model choice depends on the level of spatial dependency and response–covariate relationships. However, it does not account for the fact that in practice, as an ensemble tree-based method can make use of a large number of (possibly correlated) predictors without being constrained to certain (e.g., linear) relationships, the spatial dependency may be explained by the covariates despite not being explicitly modeled.

The objective of our study is to compare geostatistical and nonspatial ensemble tree-based models for NO_2 mapping, in terms of their prediction accuracy, uncertainty quantification, and model interpretation and to understand effect of modeling spatial structures. From here we will refer a geostatistical model simply as a "spatial model." More specifically, the following subobjectives are reached:

1. Optimizing a set of spatial hierarchical and ML models for NO_2 prediction in Germany and the Netherlands.

2. Developing a nonspatial and a spatial stacked ensemble model, that is, a stack of various ML learners.
3. Model comparison regarding the predicted mean, prediction interval, and model interpretation.

The spatial hierarchical model incorporates the spatial random effect along with other covariates and the estimation is performed using the R package `INLA` (Rue, Martino, and Chopin 2009; Martins et al. 2013). XGB, RF and Lasso are chosen for the comparison with the spatial model and they also form the base learners in the spatial and nonspatial stacked learning models. Base learners are individual learners or algorithms of the ensemble. The ML methods are chosen for their dissimilarity. Specifically, Lasso represents regularized linear regression models. RF and XGB represents nonlinear ensemble models, in our study the regression trees are the base learners. XGB is a highly scalable boosting method that builds tree models subsequently over the residuals of previous trees and has multiple routines to penalize model over-fitting (Chen et al. 2019b), which has been reported in various studies to obtain the highest prediction accuracy (Lu et al. 2020a).

Data

NO_2 concentration measurements of 2017 from national ground stations of Germany (416 stations) and the Netherlands (66 stations) are used (in total: 482). The original hourly data is downloaded from the EEA (European Environment Agency, Nelson 1999; EEA 2021). Negative values are considered as missing. The stations consisting of more than 25% of missing data according to the original hourly measurements are shown in Appendix S1. The data is aggregated to annual concentrations by taking the mean and omitting missing values. The spatial distribution of NO_2 stations and the station types, histogram and Q-Q plot for normality are shown in Fig. 1. We conducted a Shapiro test for normality, with the result implying the distribution of data being significantly different from normal distribution (P-value= 8.605e-12, "normal distribution" and "Gaussian distribution" are used interchangeably in this study). A Gamma distribution test was conducted using the method proposed in Villaseñor and González-Estrada (2015) and implemented in Gonzalez-Estrada and Villasenor-Alva (2020). The test result (P-value = 0.32) indicates that the data distribution is not significantly different from Gamma distribution.

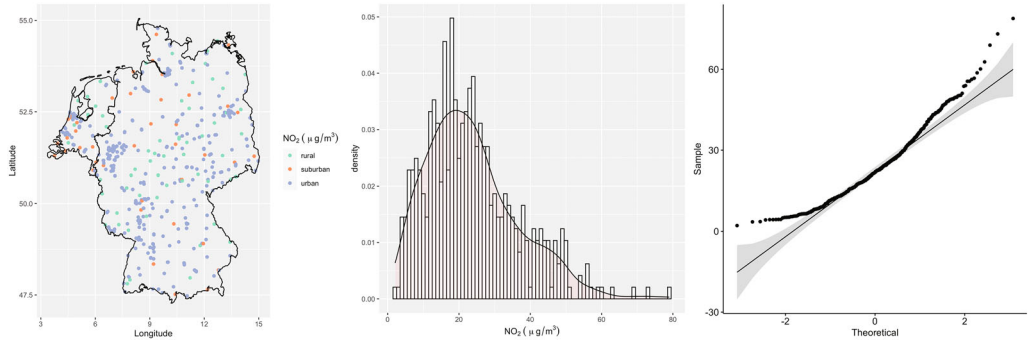

Figure 1. The geographical distribution of ground stations, histogram and Q-Q plot of the NO_2 measurements used in this study.

Table 1. Geospatial Predictors Considered in This Study. "_mon" Indicates Months (mon = 1, 2, … , 12). "_buf" Indicates Buffer Radius in Meters. The Road Length and Industrial Areas are Calculated with Buffer Radii of 100, 300, 500, 800, 1000, 3000, and 5000 m. The Night Lights Digital Numbers are Calculated with Buffer Radii of 450, 900, 3150, and 4950 m. The Original Resolution is Provided for Gridded Variables and Data Types for Vector Variables

Predictor	Variable name	Unit	Resolution/data type
Monthly wind speed at 10 m altitude.	Wind_speed_10m_mon	km/hr	10 km
Monthly temperature at 2 m altitude.	temperature_2m_mon	Celsius	10 km
TROPOMI 2018 mean vertical column density.	trop_mean_filt; Tropomi	mol/cm^2	0.01 arc degrees
Population in 5 km grid	population_5000	count	5 km
Population in 3 km grid	population_3000	count	3 km
Population in 1 km grid	population_1000	count	1 km
Nightlight	nightlight_bufnl	$Wcm^{-2}sr^{-1}$	500 m
Total length of highway	road_1_buf	m	polygon, lineString
Total length of primary roads	road_2_buf	m	polygon, lineString
Total length of local roads	road_M345_buf	m	polygon, lineString
Area of industry	I_1_buf	m^2	polygon, lineString

The geospatial predictor grids (Table 1) are calculated or resampled at 100 m resolution. They are either spatial attributes aggregated in a circular ring centered at each sensor or prediction location, called buffered predictors, or values of the spatial attribute at the observation or prediction location, called gridded variables. The buffered predictors include total road length, total industry areas, VIIRS (Visible Infrared Imaging Radiometer Suite) nighttime day/night band radiance values (nightlight, NOAA 2021) and population. Variables that are originally grids include wind speed and temperature (Dee et al. 2011), elevation (2021), annual mean Tropomi level 3 product of NO_2 column density (Copernicus 2021) from 2019 (due to the increased resolution compared to 2018). The buffered predictors of road and industry are obtained from OpenStreetMap (OpenStreetMap contributors 2019). For a detailed description of the processing of the geospatial predictors please refer to Lu et al. (2020a).

Methods

The methods considered in this study are classified as spatial and nonspatial and are given the names below in this study.

Spatial models:

1. INLA: A spatial hierarchical model fit using INLA with a Gaussian likelihood.
2. INLA-G: A spatial hierarchical model fit using INLA with a Gamma likelihood.
3. SE-INLA: using a spatial hierarchical model to stacked ensemble learning with Lasso, RF and XGB models as base learners;

Nonspatial models:

1. LA: A Lasso regression model;
2. RF: A RF model;
3. XGB: An XGB model assuming a Gaussian objective function;
4. XGB-G: An XGB model assuming a Gamma objective function;
5. QRFLA: using Lasso to aggregate QRF trees (Hastie, Tibshirani, and Friedman 2009);
6. SE: stacked ensemble learning with Lasso, RF and XGB models as base learners;
7. QRF: quantile regression forest (Meinshausen 2006);
8. DF: distributional regression forest (Schlosser et al. 2019).

To deepen our understanding of the effects of modeling the spatial process in our INLA model, we implemented an INLA model without modeling the spatial random effect (called nonspatial INLA).

Spatial model

We assume y_i (here: NO_2 observations at ground stations), measured at locations $\boldsymbol{s_i}$, $i = 1, \ldots, n$, follow a Gaussian distribution with mean μ_i and variance σ^2. The mean μ_i is expressed as a sum of covariates plus a spatial random effect, which is assumed as a Gaussian random field (Cressie 2015). We can describe a spatial statistical model:

$$y_i \sim N(\mu_i, \sigma^2), \quad i = 1, 2, \ldots, n \tag{1}$$

$$\mu_i = \boldsymbol{d_i}\boldsymbol{\beta} + x(\boldsymbol{s_i}), \tag{2}$$

where, $\boldsymbol{d_i} = (d_{i1}, \ldots, d_{ip})$ is the vector of covariates at location $\boldsymbol{s_i}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the coefficient vector, and $x(\boldsymbol{s_i})$ denotes a Gaussian random field. The Gaussian random field can be expressed as $\{x(\boldsymbol{s}_1), \ldots, x(\boldsymbol{s}_n)\} \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\mathcal{MVN}$ denotes a multivariate normal distribution with a zero-mean vector $\mathbf{0}$ and a covariance matrix $\boldsymbol{\Sigma}$. Furthermore, the Gaussian random field is specified completely by its mean $\mathbb{E}(x(\boldsymbol{s}))$ and covariance function $C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \mathrm{Cov}(x(\boldsymbol{s}_1), x(\boldsymbol{s}_2))$. The Gaussian random field can be stationary and isotropic, where the covariance function depends only on the distance and not direction between points, that is $C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \mathrm{Cov}(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|)$ and its dependence is commonly modeled using a Matérn function (Stein 1999; Yuan 2011; Diggle et al. 2013). Incorporating the spatial dependence by a Gaussian random field with a large number of observations "n" makes the estimation process computationally expensive due to the dense covariance matrix. Specifically, computing the Gaussian likelihood has the memory cost $\mathcal{O}(n^2)$ and the arithmetic cost $\mathcal{O}(n^3)$ (Chen and Stein 2021). To address this limitation, Rue and Held (2005) proposed the approximation of a Gaussian random field by a Gaussian Markov random field for a more efficient computational process of estimation. The main property of the Gaussian Markov random field is that it uses a conditional dependency structure through the precision matrix $\boldsymbol{Q}$.

In this study, we compare two spatial hierarchical models with geospatial predictors as covariates, one uses a Gaussian likelihood and the other a Gamma likelihood. The Gamma model has the same hierarchical structure as the Gaussian model: the response variable in (equation 1) can be represented by $y_i \sim \mathrm{Gamma}(\alpha, \beta)$ where α is the shape parameter and β the rate parameter. The SE-INLA model uses a Gaussian likelihood.

SPDE and INLA

To fit the spatial models, we use the `R` package `INLA`. Following the expression proposed in (equation 1), the structure for the hierarchical model is:

$$\mathbf{y}|\mathbf{x}, \theta_1 \sim N(\boldsymbol{D\beta} + \mathbf{Ax}, \theta_1), \tag{3}$$

$$\mathbf{x}|\theta_2 \sim \mathsf{GRF}(\mathbf{0}, \boldsymbol{Q}(\boldsymbol{\theta_2})^{-1}), \tag{4}$$

$$\boldsymbol{\theta} = \{\theta_1, \theta_2\}, \tag{5}$$

where $\boldsymbol{\theta}$ is the vector of hyperparameters with $\theta_1 = \sigma^2$, $\theta_2 = \{\log(\tau), \log(\kappa)\}$, where τ denotes the precision and κ the range. $\boldsymbol{x}$ is the Gaussian random field (commonly known as a spatial latent field), $\boldsymbol{Q}$ is the precision matrix, $\boldsymbol{A}$ is the projector matrix and $\boldsymbol{y}$ is the vector of the response variable $f(\cdot|\boldsymbol{x}, \boldsymbol{\theta})$, commonly from the exponential family of distributions. $\boldsymbol{D}$ is the design matrix and $\boldsymbol{\beta}$ a vector of coefficients associated with the covariates in the design matrix.

To model data indexed in space, Lindgren, Rue, and Lindström (2011) proposed a new methodology based mainly on the approximation of the Gaussian random field with the Matérn function using Stochastic Partial Differential Equations (SPDE method) as follows:

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau(\boldsymbol{s})x(\boldsymbol{s})) = \boldsymbol{\mathcal{W}}(\boldsymbol{s}), \tag{6}$$

where κ is a scale parameter, $x(\boldsymbol{s})$ is a spatial random field, Δ is the Laplacian, α is the parameter that controls the smoothness of the realizations, τ controls the variance and $\boldsymbol{\mathcal{W}}(\boldsymbol{s})$ is a Gaussian spatial white noise process (Lindgren, Rue, et al. (2015)). For the above, we can use a Gaussian Markov random field that approximates a Gaussian random field using a triangulation of the region of study without specifying an explicit covariance structure through the SPDE method. This approximation leads to a decrease in computational burden from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^{3/2})$.

The `R` package `INLA` is used to perform direct numerical calculation of the posterior distribution for a Bayesian hierarchical model (Rue, Martino, and Chopin (2009), Rue, Martino, and Chopin (2009)). If we use $\boldsymbol{x}$ to indicate a a Gaussian Markov random field (a latent Gaussian field), $\boldsymbol{\theta}$ a vector of hyperparameters and $\boldsymbol{y}$ a vector of observations, assuming independent observations given the vector of the spatial latent field ($\boldsymbol{x}$) and the hyperparameters ($\boldsymbol{\theta}$), the likelihood can be expressed as:

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i \in \mathcal{I}} p(y_i|\eta_i, \boldsymbol{\theta}), \tag{7}$$

where η_i is the linear predictor and $\mathcal{I}$ contains the indices of the observed values $\boldsymbol{y}$.

The main aim is to approximate the posterior distribution of the spatial random effects and the hyperparameters. The marginal densities can be obtained:

$$p(x_i|\boldsymbol{y}) = \int p(x_i|\boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}, \tag{8}$$

and

$$p(\boldsymbol{\theta}_j|\boldsymbol{y}) = \int p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-j}, \tag{9}$$

respectively (Lindgren, Rue, et al. 2015; Krainski et al. 2018).

Table 2. Frequency (number of times) of Variables Selected by Lasso in 20 Times Bootstrapping and Variables that are Selected more than 90% Times (i.e., 18) are Listed Below. These Variables are Considered in INLA Except the road_class_3_3000

	Variables	Frequency
1	nightlight_450	20
2	population_1000	20
3	population_3000	20
4	road_class_1_5000	20
5	road_class_2_100	20
6	road_class_3_300	20
7	trop_mean_filt	20
8	road_class_3_3000	19
9	road_class_1_100	18

Geospatial predictor selection for the INLA model

As involving too many covariates (e.g., more than 12) in the INLA model brings problems in model multicollinearity, we used Lasso to reduce the number of variables. The Lasso was used instead of ensemble tree-based methods for feature selection because it is also a linear model (same as the INLA and INLA-G models in our study). Variables are selected with the L1 norm penalty that returns a model with errors that are within one SE of the minimum mean cross-validated error. Lasso is applied to 80% data randomly sampled from all the observations and this process is repeated 20 times. Variables that are selected more than 90% of the times (i.e. 18) will be considered as covariates in INLA. The times that the Lasso selected certain variables is shown in Table 2. The INLA modeling process applies the same bootstrapped samples for training and validation. In addition, the AIC step-wise model selection is applied to the entire dataset to suggest a model as a further reference. The variables selected by AIC are almost the same as the Lasso selected variables, only that it does not choose the road_class_3_3000, which is highly correlated with road_class_1_5000. Based on this, the road_class_3_3000 is not used as a covariate in INLA.

INLA model parameterization

The triangulated mesh constructed in the SPDE approach is shown in Fig. S1. The mesh has an inner area with small triangles where precision is needed, and an outer extension with bigger triangles to avoid boundary effects (Moraga et al. 2021). Specifically, the size of the inner and outer extensions around the data locations (*offsets*) is set to 1/8 of the maximum distance among all the observations for both the inner and outer extensions. The maximum allowed triangle edge lengths in the region and in the extension (*max.edge*) are set to respectively 1/30 and 1/5 times maximum distance among all the observations. The Matérn SPDE model is constructed with $\alpha = 2$. The SE-INLA model has the same specification (i.e. mesh structure, likelihood, objective function, priors, Optimization process) as the INLA model parameterization described above.

Nonspatial methods

Lasso is a linear regression algorithm with the L1 regularization to shrink variable coefficients to zero, which enables "feature selection." In the cost function, the absolute value of coefficient is added to the original least squares as a penalty term. RF and XGB in this study use trees as

base learners and ensemble them to reduce variability of single trees (Friedman 2001). RF firstly randomly draws a subset of features, and then chooses features from this subset to build the tree. RF (Breiman 2001) grows trees independently and then takes the mean of the predictions of each tree.

QRF is a nonparametric prediction interval estimation method which keeps all the observations in the terminal node for estimating the conditional probability function. Specifical ly, it samples from all the response values in each terminal node and use the ratio between the number of samples that is taken from each terminal node and the number of total observations in the terminal node as weights to aggregate the samples. The weights of all the trees are summed. The summed weights computed for each observation are then used to construct the empirical conditional cumulative distribution function (Meinshausen 2006).

QRFLA uses Lasso as a postprocessing of QRF (Hastie, Tibshirani, and Friedman 2017, page 617). This method preserves all the trees instead of aggregating them (e.g., taking the mean of all the predictions) and then apply Lasso regression to all the trees for aggregation. This leads to a shrinkage of the tree space and theoretically reduces model variance. DF (Schlosser et al. 2019) firstly divides data into regions as homogeneous as possible with respect to a parametric distribution, thus capturing changes in location, scale, and shapes. For each tree, maximum likelihood is used to fit distributions and recursively select and split covariates according to the instability of the gradient of the likelihood at each observation along each covariate. Then, the distributional trees are ensembled for DF.

XGB is a variation of gradient boosting, which grows trees subsequently by fitting to model residuals of the previous step. XGB is scalable to multiple threads. It enables multiple penalization paths to control model complexity to prevent model over-fitting, including regularization (e.g., L1 regularization) on tree width and terminal node values, as well as drop-out (dropping trees), sampling observations (take a subset of observations in each run), and early stopping (stop iterating when after a few rounds the loss does not decrease or the node does not meet the splitting rule). The default objective function for regression assumes normal distribution of target variables (and the prediction is the mean of the distribution). This assumption has been used in all the air pollution mapping studies. Here, we additionally fit a model with the objective function assuming the target variable follows a Gamma distribution (XGB-G) as the distribution of NO_2 measurements is closer to Gamma than normal distribution.

Different from the ensembling in RF or XGB, SE (Stacked Ensemble) refers to a class of algorithms that trains a second-level "meta learner" to optimize the combination of a collection of base learners. The base learners are preferably diverse to capture different relationships or patterns. In this study, Lasso, RF, and XGB are the base learners. Cross-validated predicted values (commonly known as "level-1" data) are used to train the meta learner.

Hyperparameter setting for XGB and RF

To optimize the hyperparameters of XGB, we used the grid search to optimize hyperparameters in a fivefold CV based on the minimum RMSE (Root Mean Squared Error) and additionally manual adjustment of the hyperparameters to look at the prediction patterns. The grid search is used instead of more computationally efficient methods (e.g., Bayesian or random search) as the optimal hyperparameter range is largely known from our previous experiences (Lu et al. 2020a; Lu et al. 2021). The search grid for the number of iterations (nrounds) was from 200 to 3000, with a step of 200; maximum tree depth (max-depth) from three to six with a step of one, learning rate (eta) from 0.001 to 0.1 with a step of 0.05, the penalty term Gamma (Chen et al. 2019b)

from one to five with a step of one, the subsample is set to 0.7, L1 norm penalization (lambda) is set to two and L2 norm penalization (alpha) is set to 0. For RF, we used the default setting of number of variables that are randomly drawn for each tree (Breiman 2001), which is the integer part of the total number of variables divided by three. The number of trees is set to 2000 for a safe choice as the high number of trees will not negatively affect model performance. The minimum size of terminal node was optimized between 5 and 10, and was set to five.

Model evaluation

Cross validation

We use RMSE, MAE (Mean Absolute Error), IQR (InterQuartile Range) and Nash-Sutcliffe model Efficiency coefficient (NSE) to assess and compare model performance. RMSE is calculated as the square root of the differences between predictions and observations; MAE is calculated as the absolute differences between predictions and observations; IQR is the differences between the third and first quartiles of the prediction. NSE is calculated as $\mathrm{NSE} = 1 - \mathrm{MSE}/\mathrm{var}(y)$, where MSE indicates mean squared error, var(.) indicates variance, and y indicates observed response values. When different data is used in CV (e.g., separating between close and far-away from roads), we additionally calculated the RRMSE (relative RMSE), RMAE (relative MAE), RIQR (relative IQR) to account for the differences in the magnitudes of response values. The RRMSE and RMAE are calculated by dividing the RMSE and MAE, respectively, by the mean of observations. The RIQR was calculated by dividing the IQR by the median of observations. The three CV methods we designed and used to assess our model performance are:

1. Bootstrapped CV. 20-times randomly bootstrapped splitting of training and test sets (Lu et al. 2020a).
2. Spatial-blocked CV. Dividing data into spatial blocks, each time use one block for test and other blocks for training. In this study, the spatial grids are divided with the cell size two degree (around 222 km), which leads to 20 spatial blocks.
3. Customized CV. Splitting train-test based on values of certain covariates which definition is presented in Table 1. In this study, three subareas are defined (1) close to traffic and with high population ("tr-hp"), (2) close to traffic and with middle low population ("tr-lmp"), (3) far away from traffic ("far"). High population is defined as the variable population of 1000 m buffer is in the last quartile. Low population is defined as the variable population of 1000 m buffer is below the median. Close to traffic is defined as: the road_class_2_100 is larger than 0, or the road_class_1_100 is larger than 0, or the road_class_3_100 is within its 75th percentile. Far away from traffic is defined as: the road_class_2_100 is 0, and the road_class_1_100 is 0, and the the road_class_3_100 is below its median.

This yields 85, 65, and 177 samples in each category. This ensures a balanced number of samples between close to traffic and far-away from traffic. Each time, 30 samples (7% of the entire dataset) are drawn from the corresponding category to form a test sets for CV. To illustrate, each time, 30 samples are drawn from the 85 samples as the test set to obtain the prediction accuracy CV for the situation "tr-hp" and the rest is used for training.

Prediction intervals

CRPS (Continuous Ranked Probability Score) and coverage probabilities are used as quality indicators of prediction intervals. CRPS is an uncertainty measure that assesses the similarity between two distributions. We use it to indicate how the predicted distribution matches the

observed distribution. The CRPS implemented in the R package `ScoringRules` (Jordan, Krüger, and Lerch 2017) is used. CRPS is calculated for the INLA and QRF models. For the INLA model, the prediction intervals are calculated by simulating from the response Y $\sim N(\theta, \sigma^2)$ where θ and σ^2 are the fitted mean and variance. The mean of CRPS for all the points within each test block is calculated in spatial-blocked CV. Coverage probabilities are calculated as the ratio between the number of predictions within the upper and lower quantiles and the total number of predictions (in the test set). The prediction intervals are mainly compared between INLA, INLA-G, QRF, and DF. The prediction interval for QRFLA is compared with QRF to investigate the effects of Lasso tree-aggregation strategy on the prediction intervals.

Model interpretation

We inspect fixed and spatial random effects modeled by INLA and compare the spatial random field with modeled prediction intervals and model residuals to understand the contribution of spatial random effects. Different from linear regression methods, which themselves are the best models for interpretation, interpreting ensembling tree-based methods requires external models (Lundberg and Lee 2017). We use SHAP (SHapley Additive exPlanations, Lundberg et al. 2018), a unified method based on additive feature attribution, to estimate variable influence in RF and XGB models.

Results

Accuracy assessment and uncertainty quantification

Spatial-blocked CV

Spatial-blocked CV provides information about prediction accuracy in spatial blocks. We compare the spatial patterns of NSE as an indicator of model prediction accuracy and spatial patterns of CRPS as an indicator of the quality of the prediction intervals, of INLA, representing spatial method, and RF, representing nonspatial models. As the XGB outperforms RF in nonspatial CV, we also compares the spatial-blocked NSE for XGB.

The NSE map (Fig. 2) shows that the XGB, RF, and INLA predict relatively well in most parts of Germany besides blocks at the boundaries. The NSE for the block western the Netherlands is also relatively low with all the three methods and especially for XGB (NSE: 0.2). RF obtained the best result for the block of western the Netherlands (NSE: 0.5). The INLA model

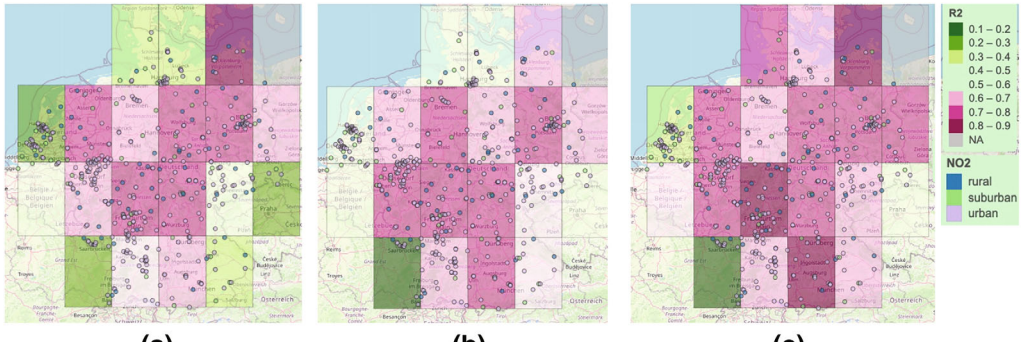

Figure 2. The NSE of each block, using the rest of the blocks for training. The models are (a) XGB, (b) QRF, (c) INLA.

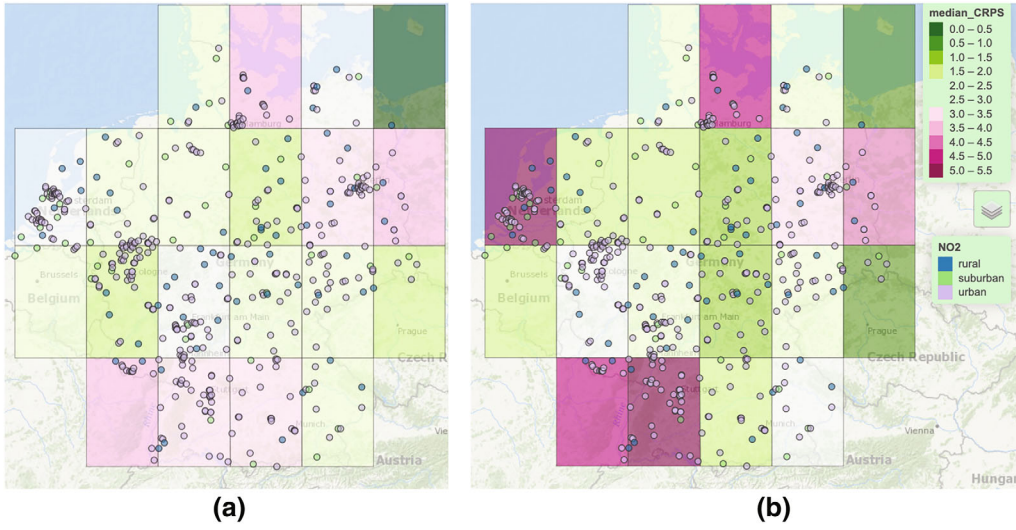

Figure 3. The CRPS (Continuous Ranked Probability Score) of each block, using the rest of the blocks for training. (a) RF, (b) INLA.

outperformed RF and XGB in the blocks at south-east and north. The NSE between blocks are the most heterogeneous with XGB, which is consistent to the result of bootstrapped CV that the XGB falls short at predicting extremes.

The spatial-blocked CRPS (Fig. 3) is computed for QRF and INLA. We did not show the results using DF as it will be seen that the QRF and DF performed similarly in predicting the intervals (section 5.2). The INLA predicted prediction distribution deviates considerably from observed distribution for the block of western the Netherlands, as reflected by the high value of mean CRPS. This is consistent to the relatively low NSE observed for the same block. However, some blocks with relatively high NSE (in the north and south) have high CRPS. This indicates that the prediction means are well-predicted but not the prediction interval (too narrow).

Nonspatial CV

Both ensemble tree-based methods with a Gaussian objective function and INLA with a Gaussian likelihood function obtained higher prediction accuracy than Lasso (Table 3), indicating the necessity of using a more flexible model and modeling the spatial random fields. Among individual methods, in terms of NSE and RMSE, INLA with Gaussian likelihood obtained the highest prediction accuracy, followed by XGB and QRFLA. QRFLA greatly improves from the original RF. Despite the distribution of response being closer to Gamma distribution compared to Gaussian distribution, using Gamma regression in XGB and specifying Gamma likelihood in INLA both decrease the prediction accuracy considerably. Compared to INLA, XGB obtained lower RMSE and NSE despite it obtained lower MAE and IQR, indicating that the XGB model predicts less well at more extreme ranges. The QRF and DF results are not shown in Table 3 as the results are very similar to RF. Their prediction intervals are compared.

SE-INLA obtained a higher prediction accuracy compared to SE and INLA. It obtained the best results in terms of root mean squared error (6.83, 24.5% of the mean of observations) and NSE (0.71). This indicates the explicit modeling of spatial structures could further improve the prediction accuracy despite flexible relationships captured from ML models.

Table 3. Prediction Accuracy Matrix for Different Models Using 20 Times Bootstrapped Cross-Validation. Nonspatial Models: LA: Lasso; RF: Random Forest, XGB: XGBoost Using the Default Gaussian Loss; XGB-G: XGBoost Using a Gamma Loss; QRFLA: Quantile Random Forest with Lasso for Shrinkage Aggregation of Regression Trees; SE: Stacked Ensembling. Spatial models: INLA: a Latent Gaussian Model Implemented Using INLA Assuming a Gaussian Likelihood. INLA-G: a Latent Gaussian Model Implemented Using INLA Assuming a Gamma Likelihood. SE-INLA, Geostatistical Stacked Ensembling

	LA	RF	XGB	XGB-G	QRFLA	SE	INLA	INLA-G	SE-INLA
RMSE	7.54	7.45	7.14	8.91	7.23	7.18	7.06	9.21	6.83
IQR	8.47	7.39	6.54	9.21	7.27	7.30	7.1	7.4	6.8
MAE	5.69	5.51	5.05	6.27	5.28	5.31	5.3	6.2	5.0
NSE	0.65	0.65	0.68	0.51	0.67	0.69	0.69	0.45	0.71

Compared to the INLA model, the nonspatial INLA model obtained lower DIC (Deviance Information Criterion, 3286.66 versus 3251.97 with spatial effects) and WAIC (Watanabe-Akaike information criterion, 3291.75 versus 3253.93 with spatial effects). These suggest the advantage of modeling the spatial effects. We normalized covariates before inputting into the spatial and nonspatial INLA models and compared the differences between the fixed-effects obtained by the original and nonspatial INLA model (Figs. S3 and S4) and found the most notable change being the increased effect of the population_1000 for the nonspatial INLA model. This can be explained by that part of the effects of population_1000 is modeled in the spatial random field. The second most notable change is on the decreased effect of nightlight_450 for the nonspatial INLA model. After the spatial process is modeled, the nightlight_450 has a higher contribution to the model. Together with the decreased effects of road_class_2_100 and road_class_3_300 for the nonspatial INLA model, these may indicate that the spatial model could better account for traffic-related variables (i.e., road and nightlight in smaller buffers).

Customized CV

There is a distinctive difference between model performance in areas close to traffic (i.e., *tr-hp* and *tr-lmp*) and far away from traffic (i.e., *far*). The INLA model outperformed other nonspatial methods in both *tr-hp* and *tr-lmp*, especially for the latter while the XGB model outperformed the INLA model (and all the other models) in *far*. This indicates the importance of modeling spatial dependency in areas close to traffic and possibly nonlinear relationships far-away from roads. All the ensemble tree-based methods obtained much worse results compared to linear regression methods in *tr-lmp*. A linear regression model typically outperforms ensemble tree-based methods when there are relatively few observations for a flexible relationship to be justified. As the number of observations that are close to traffic and far away from traffic is balanced, the results indicate that the population density alters relationships between NO_2 and road density (i.e., the relationships between NO_2 and road density is different with different population density) in areas close to traffic (Table 4).

Prediction interval

After examining the quality of the prediction intervals. We compare prediction intervals estimated from the spatial models with different likelihood functions (Fig. 4), the RF-based methods with

Table 4. Results with Customized CV. tr-hp: Close to Traffic and High Population, tr-lmp: Close to Traffic and Middle and Low Population, far: Far Away from Traffic. RRMSE (relative RMSE), RMAE (relative MAE), RIQR (relative IQR)

	RMSE	RRMSE	IQR	RIQR	MAE	RMAE	NSE
LA_tr-hp	12.4	0.3	17.3	0.4	10.2	0.3	0.11
RF_tr-hp	11.9	0.3	17.8	0.5	9.8	0.3	0.18
XGB_tr-hp	11.6	0.3	15.3	0.4	9.3	0.2	0.21
INLA_tr-hp	11.3	0.3	16.6	0.4	9.5	0.3	0.26
LA_tr-lmp	7.5	0.3	10.4	0.5	6.1	0.3	0.21
RF_tr-lmp	8.2	0.4	10.9	0.5	6.4	0.3	0.05
XGB_tr-lmp	8.2	0.4	10.5	0.5	6.4	0.3	0.04
INLA_tr-lmp	6.7	0.3	8.7	0.4	5.3	0.2	0.36
LA_far	5.0	0.4	4.9	0.4	4.2	0.3	0.47
RF_far	4.9	0.3	4.0	0.3	3.6	0.3	0.47
XGB_far	3.4	0.2	3.6	0.3	2.5	0.2	0.74
INLA_far	4.0	0.3	4.3	0.3	3.2	0.2	0.65

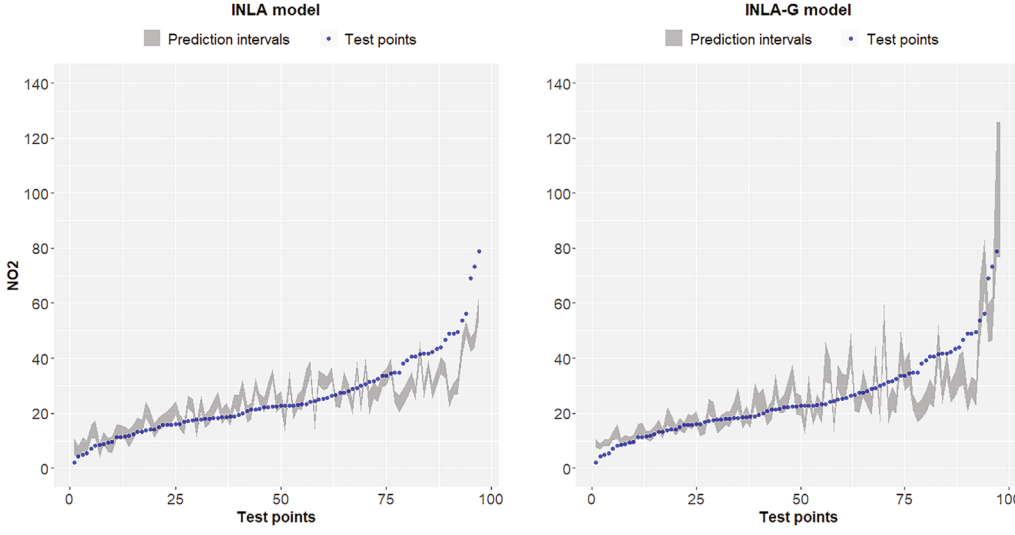

Figure 4. The 90% prediction intervals predicted by INLA and INLA-G.

prediction intervals estimated in two different methods, QRF and DF (Fig. 5), and the RF-based method with and without Lasso postprocessing, QRF and QRFLA (Fig. 6).

The RF-based methods, namely DF, QRF, and QRFLA reach the coverage probability higher than 0.9, but the DF predicts a more realistic prediction quantile, notably, it covers four observations that are not covered by the same prediction quantiles predicted by the QRF. The INLA 90% prediction interval is too narrow. The coverage probability is 0.41 for INLA and 0.36 for INLA-G. The predicted 90th quantile of the INLA-G turned out to better capture extreme high values but miss more at the lower values. The QRFLA predicted a slightly narrower prediction

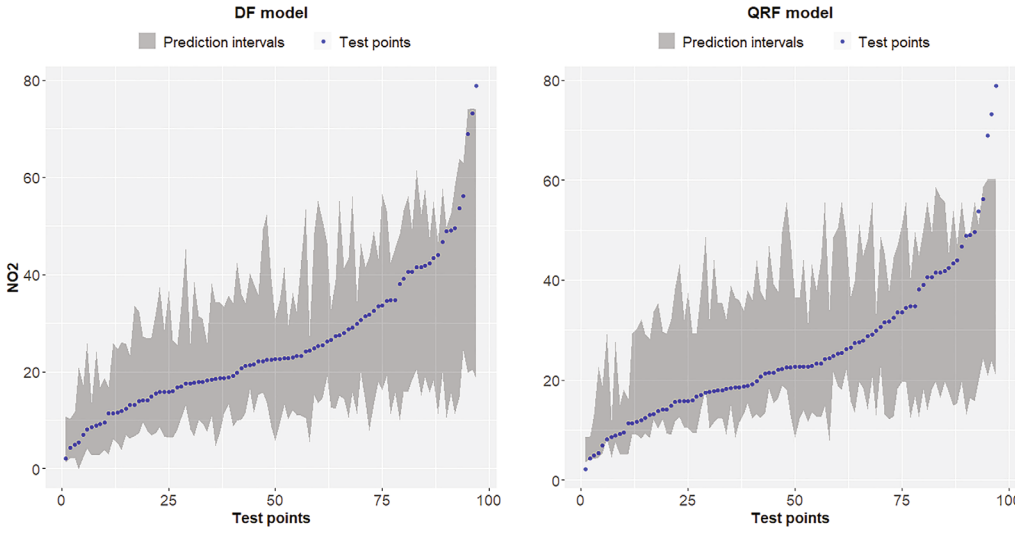

Figure 5. The 90% prediction intervals predicted by DF and QRF.

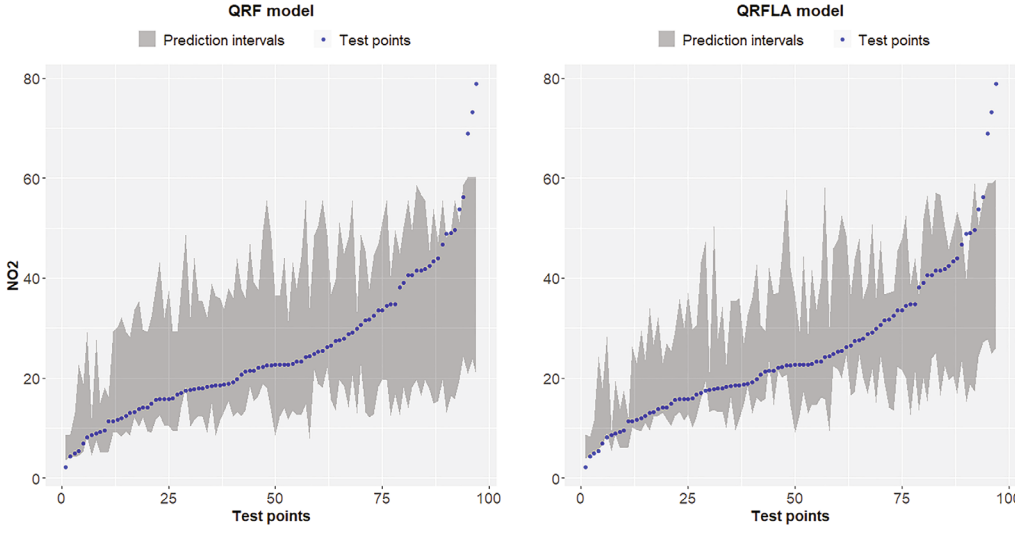

Figure 6. The 90% prediction intervals predicted by QRF and QRFLA.

interval compared to QRF. This indicates that the Lasso-based postprocessing could reduce the variance of a QRF model.

Model interpretation

SHAP values are calculated for RF and XGB methods using all the data. The variables are ranked by their variable importance, which is calculated as the sum of SHAP magnitudes over all the samples. It can be observed from Fig. 7 that the variable rankings and the pattern of variable impacts on model output are similar. Both methods ranked road_class_2_100 at the top.

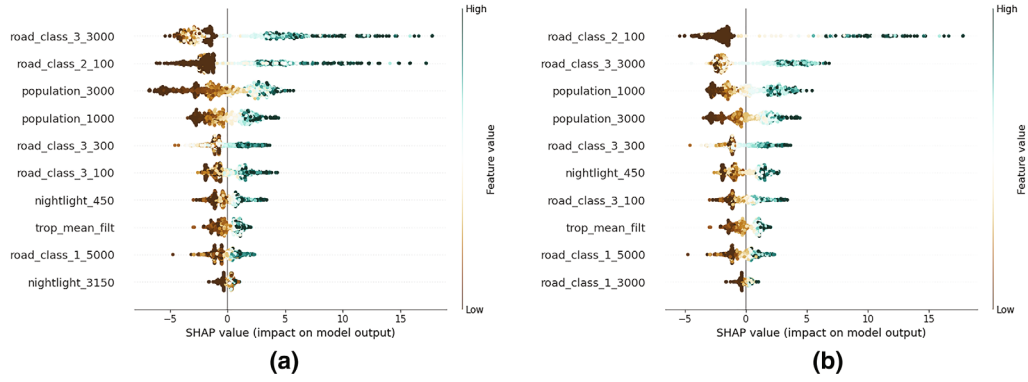

Figure 7. Variable impact calculated by SHAP (SHapley Additive exPlanations), (a) the RF model, (b) The XGB model. The covariate ranking is based on the sum of SHAP magnitudes over all the samples. The horizontal axis shows how the effect of a feature changes with the SHAP value of the feature (indicated by the brown to green color bar). The SHAP value is calculated as the conditional mean (conditional to the features that are not used for making the prediction) of the prediction. For example, the low values of the feature "road_class_2_100," shown in dark brown, correspond mostly with low SHAP values (points to the left). This can be explained that usually the less local road, the less pollution, considering other effects. High values in "road_class_2_100" also contribute greatly to the predictions.

The variable importance calculated by the SHAP indicates a pattern that matches well with our expectation in the emission sources (e.g., high pollution close to primary roads). To illustrate, we observe a positive trend of SHAP values along with road_class_2_100 values, this matches with the explanation that areas with higher primary road density generally experience higher NO_2 concentrations.

To analyze the effect of each covariate in the INLA model, we firstly normalized all the covariates (by subtracting the mean and dividing the centered columns by their standard deviations) and used all the data to fit the INLA model. road_class_2_100 has the highest effect (mean = 4.37), follows by the population_3000 (3.08), these are consistent to the XGB variable importance (Fig. 7b). Then, the road_class_3_300 (3.00) has a notably higher effect (besides the top 2) than other covariates, which has coefficients from 0.72 to 1.88. This differs from the XGB and RF variable importance which ranked the population_1000 higher above, while in the INLA model the population_1000 has the lowest effect (0.72). This may be because of the high correlation between population_1000 and population_3000, as SHAP is a permutation test, it ignores the dependency between covariates. In general, both geostatistical and ML methods estimated covariate effects match their physical explanations. The statistics (mean, standard deviation, mode) and predicted quantiles of each coefficient are shown in Fig. S3.

The differences between the predicted NO_2 and the mean of the spatial random field (Fig. 8) indicates the effects of covariates. The highest values of the mean of the spatial random field are shown in the south-west (48.7758° N, 9.1829° E, in and around the German city Stuttgart). Relatively high values can be observed in northern, southern and western Germany. Compared to Fig. 9, the areas close to the Stuttgart (Germany) region where the mean values of the spatial random field are high corresponds to the high magnitudes of NO_2 concentrations. Also,

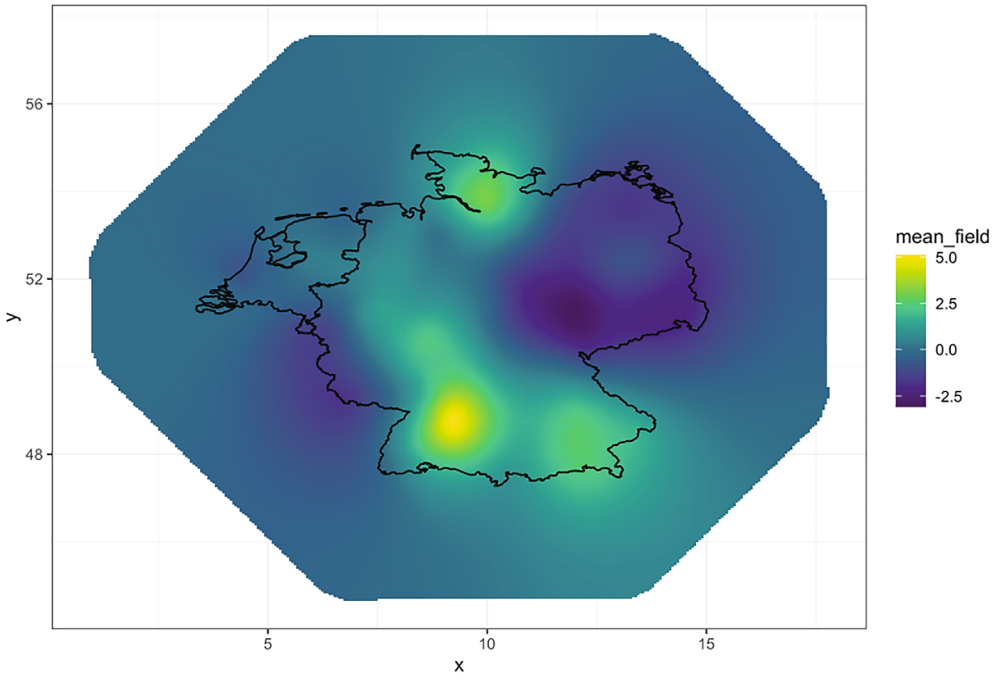

Figure 8. Mean of the spatial random field fitted by the INLA model. The polygons indicate Dutch and German boundaries.

the differences between the observations and predictions are relatively large in magnitudes in this region. To facilitate visualization, we also calculated the differences between INLA model predictions and the observations (Fig. S2).

Discussion

In this study, we compared spatial and nonspatial models for spatial NO_2 prediction in Germany and the Netherlands. The comparison consists of the predicted mean, prediction intervals, and model interpretation. Spatial and nonspatial CV strategies are used to reveal prediction accuracy in different aspects. We also implemented the Lasso postprocessed RF and spatial stacked learning for NO_2 mapping (which to our knowledge have not been applied in air pollution mapping before) and these two methods considerably improve from the original RF and stacked learning models, respectively.

Several venues were attempted to further improve the spatial model fitted with INLA. Firstly, as we observed in general worse results at the geographical boundaries (Figs. 2 and 3), we inspected if different meshes with edge-effects fully accounted (e.g., the mesh is sufficiently large for observations at the edge) could improve the prediction accuracy. It turned out that the same performance is obtained. Secondly, we suspected that deviating from the assumed Gaussian distribution causes narrow prediction intervals of the INLA model. However, assuming a Gamma likelihood did not improve the model performance in terms of the accuracy matrix, CRPS and coverage probability. We also experienced the square transformation of the observations and the use of the log-normal likelihood but that also decreases the model performance. Thirdly, we

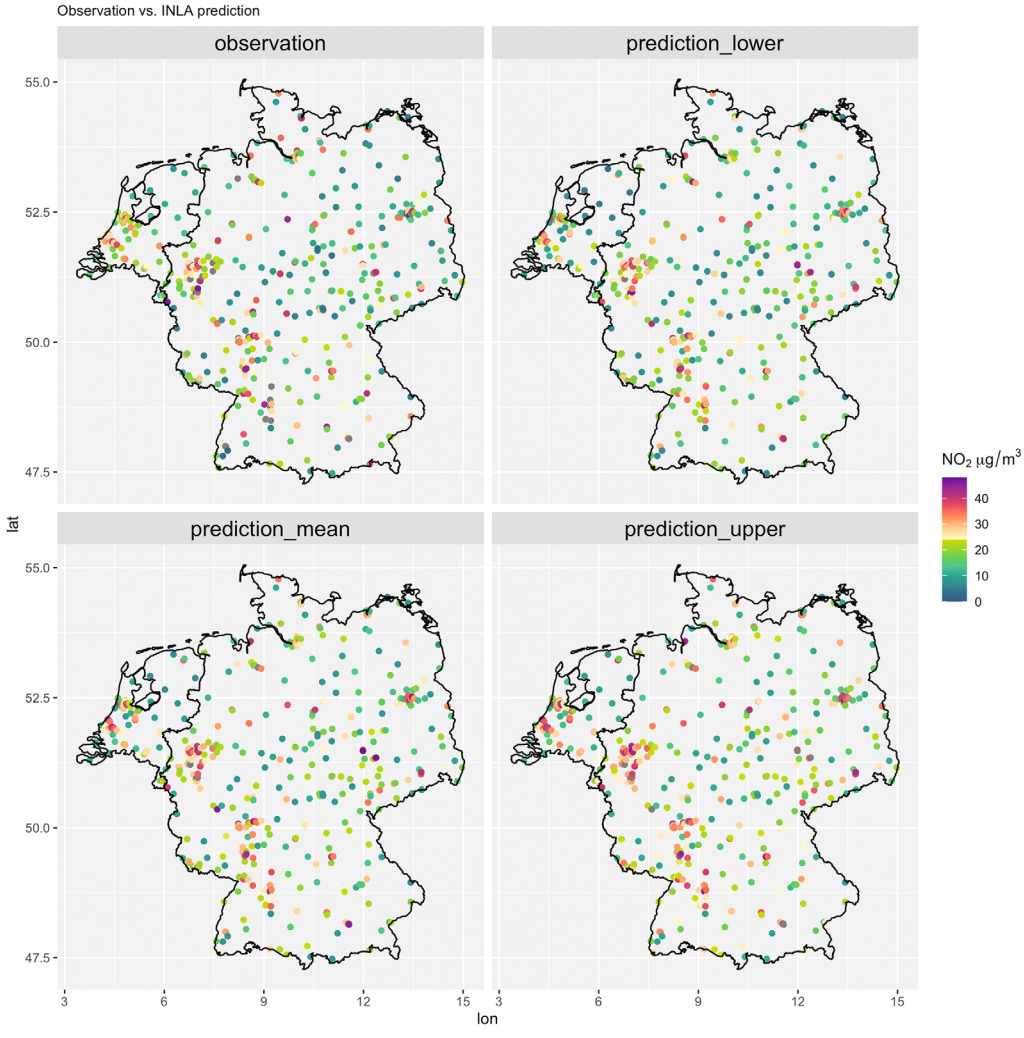

Figure 9. INLA predicted NO_2 at the ground stations with mean (prediction_mean), high (prediction_high, 0.975) and low (prediction_low, 0.925) quantiles and the observed NO_2 (observation). The polygons indicate Dutch and German boundaries.

additionally added two factor variables, namely "country code"[3] and "urban types"[4]. However, that also does not improve the model performance. In future works, using a different spatial model (e.g., by specifying different hyperparameters), including the country code and urban types as factor variables for random effects, and modeling spatial varying coefficients may improve the modeling results. Major improvement may also be achieved by integrating mobile sensing measurements and other geospatial predictors (e.g., traffic count, urban morphological matrix) (Moraga et al. 2017).

[3]DE for Germany and NL for the Netherlands
[4]Rural, urban, city center according to Dijkstra and Poelman (2014)

In our study, we designed our spatial-blocked CV to disaggregate the model accuracy metrics from global to spatial-block-wise error. This method has the benefit that we know exactly which data is used for training and which for testing, however, as Wadoux et al. (2021) pointed out, this type of method may experience the extrapolation problem. We reduce this problem by selecting a relatively small block size. Specifically, we chose the two-degree cell size as with it we have a relatively balanced number of stations and stations of different urban types and can better visualize the geographical pattern. With a smaller cell size, the information in each grid cell is closer to a single or a small combination of ground stations and gives less information about the areas surrounding each ground station. Applying the typical random k-fold or leave-one-out-cross-validation and then calculating the accuracy in each block could avoid the extrapolation problem. In Fig. S7, we show the NSE of the XGB in each spatial block after applying the random 10-fold CV and compared it with the spatial-blocked CV.

There has been a debate in how to faithfully assess the model accuracy with spatially correlated data. Several studies believe the typical k-fold CV causes overly optimistic accuracy assessment and developed "spatial CV" methods (Brenning 2012; Meyer et al. 2018) and they have been advocated in environmental mapping (Ploton et al. 2020). However, the spatial CV does not seem to solve the overly optimistic accuracy assessment problem as the problem is caused by insufficient samples in the feature space. The spatial CV causes extra extrapolation problem (Wadoux et al. 2021) without solving the real problem.

A sensible accuracy assessment approach is to look into the differences between the distribution of the feature values of the population and the samples. A progress is made by Meyer and Pebesma (2021), who proposes the area of applicability, which quantifies the differences between feature values used in training and for prediction and used the magnitude of it as an indicator for applicability. The method adds a useful diagnostic tool for applying machine learning models to spatial prediction. However, as is discussed in Meyer and Pebesma (2021), there are several limitations. One important limitation is that there is no quantitative associations between *how "applicable" an area is* and *the prediction errors or uncertainty*.

The model performance differs greatly between the three road and population situations. The "far" situation obtained the best modeling accuracy while the "tr-hp" the worst. This is likely due to the fact that the urban NO_2 process is more complex due to urban forms and traffic conditions. This may also indicate that more detailed traffic counts and meteorological data are needed for modeling the NO_2 emission sources.

Different from nonparametric models such as ensemble trees, a parametric spatial model fitted with INLA as the one developed in our study requires feature selection and the assumption of the distribution of the response. Several studies used the whole dataset for variable selection and then use selected variables for CV (Larkin et al. 2017; Lu et al. 2020b). This may, however, lead to an information leak as the validation data is also used in CV. To avoid this problem, one can include the variable selection process in each CV, that is, use the same training data for variable selection and test. However, variable selection in each run could introduce additional error and uncertainty, therefore, a determined set of covariates may be preferred. We obtain a fixed set of selected variables while reducing information leakage to a negligible level by choosing only the variables that are selected 90%–100% times of all the bootstraps of Lasso.

Using the geostatistical method to stack learners obtained higher prediction accuracy in terms of the mean prediction compared to the nonspatial stacking. This suggests the complex response–covariate relationships modeled by the ML learners do not fully capture the spatial process. The spatial stacked models obtained the highest prediction accuracy and

with high-performance computation, it is possible to apply them to a large-scale and at a high resolution. The limitation of such stacked methods is that they cannot be used to analyze the effects of covariates and therefore NO_2 emission sources. But these models could be a reference to the level of accuracy a statistical predictive model could reach with the data available and the base learners.

As ground stations of NO_2 are predominately close to traffic, the prediction might be biased, for example, toward higher NO_2 concentrations. This problem is referred to preferential sampling, which occurs when the sampling design process is dependent on the spatial process. A few methods have been developed to address this problem, using Monte Carlo estimates for the likelihood function (Diggle, Menezes, and Su 2010) or numerical methods (Dinsdale and Salibian-Barrera 2019). A much less technical approach to reduce the bias might be to include sufficient traffic-related covariates and a model capable of capturing the traffic-NO_2 relationships. In this sense, NO_2 mapping may experience a promising improvement when more traffic counts and emission data become available in future.

An advantage of the spatial model is that it could quantify the uncertainty of the spatial covariance matrix in terms of the hyperparameters of the covariance function, as well as block averages or block totals, at any aggregation level. Examples are block or point-to-area Kriging. In practice, the block averages are often more of interest than values at points, and we commonly would like to aggregate data at different spatial and spatiotemporal scales. For example, with mobile sensor measurements. Even though we could derive a prediction interval from the nonspatial method, the prediction interval is derived at the point level. At a block level, the uncertainty can only be formally quantified while accounting for the spatial dependency between points within a block. This is a distinction between nonspatial and spatial models, for example, blocked Kriging, whose Kriging variance is very sensitive to the behavior of the variogram at within-block distances.

Conclusion

We proposed a model comparison process to comprehensively compare between models considering not only the predicted mean but also prediction intervals and model interpretation. We also showed that the information provided by commonly single-used nonspatial CV may miss reflecting model behaviors. With the model comparison process, we compared the use of spatial models and ML models for the spatial prediction of NO_2 in Germany and the Netherlands and found noticeable differences in their limitations and strength. The spatial models are preferred especially for urban area prediction and provide the spatial process of observations and indicate the insufficient modeling of the fixed-effects. But the uncertainty assessment of spatial models, which is commonly known as a strength, fails to provide a prediction interval that meets the expectation when INLA is used to fit the models. The QRF and DF obtained satisfying prediction intervals, with the DF slightly more capable of predicting the extremes. Using Lasso to postprocess random forest increases model performance and reduces model variance. Using a spatial model to stack learners obtained the highest accuracy in terms of the mean prediction. Despite the NO_2 observations follow closer to a Gamma distribution than a Gaussian, the use of a Gamma likelihood in the spatial model and Gamma objective in the XGBoost obtained much worse results than using a Gaussian likelihood or objective. By comparing with the nonspatial stacked ensemble learning, spatial stacked ensemble learning suggests the necessity of modeling the spatial process.

Acknowledgement

Open Access funding enabled and organized by Projekt DEAL.

References

Alakus, C., D. Larocque, and A. Labbe. (2021). Rfpredinterval: An R Package for Prediction Intervals with Random Forests and Boosted Forests. *arXiv preprint arXiv:2106.08217*.

Anselin, L., C. Pin Tan, Y. Wang, and Z. Zhang. (2001). "Spatial Econometrics." *A Companion to Theoretical Econometrics*, 310–330.

Beloconi, A., and P. Vounatsou. (2020). "Bayesian Geostatistical Modelling of High-Resolution NO2 Exposure in Europe Combining Data from Monitors, Satellites and Chemical Transport Models." *Environment International* 138, 105578. https://doi.org/10.1016/j.envint.2020.105578

Bertazzon, S., M. Johnson, K. Eccles, and G. G. Kaplan. (2015). "Accounting for Spatial Effects in Land Use Regression for Urban Air Pollution Modeling." *Spatial and Spatio-temporal Epidemiology* 14-15, 9–21.

Bhatt, S., E. Cameron, S. R. Flaxman, D. J. Weiss, D. L. Smith, and P. W. Gething. (2017). "Improved Prediction Accuracy for Disease Risk Mapping Using Gaussian Process Stacked Generalization." *Journal of the Royal Society Interface* 14(134), 20170520.

Blangiardo, M., and M. Cameletti. (2015). *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. Hoboken, NJ: John Wiley & Sons.

Breiman, L. (2001). "Random Forests." *Machine Learning* 45(1), 5–32.

Brenning, A. (2012). "Spatial Cross-Validation and Bootstrap for the Assessment of Prediction Rules in Remote Sensing: The R package sperrorest." In *2012 IEEE International Geoscience and Remote Sensing Symposium* (pp. 5372–5375). IEEE, 2012.

Briggs, D. J., C. de Hoogh, J. Gulliver, J. Wills, P. Elliott, S. Kingham, and K. Smallbone. (2000). "A Regression-Based Method for Mapping Traffic-Related Air Pollution: Application and Testing in Four Contrasting Urban Environments." *Science of the Total Environment* 253(1-3), 151–67.

T. Chen and C. Guestrin. (2016). xgboost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. New York: ACM.

Chen, J., and M. L. Stein. (2021). "Linear-Cost Covariance Functions for Gaussian Random Fields." *Journal of the American Statistical Association*, 0, 1–18.

Chen, J., K. de Hoogh, J. Gulliver, B. Hoffmann, O. Hertel, M. Ketzel, M. Bauwelinck, A. van Donkelaar, U. A. Hvidtfeldt, K. Katsouyanni, et al. (2019a). "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, 104934.

T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li. (2019b). *xgboost: Extreme Gradient Boosting*. R Package Version 0.82.1. https://CRAN.R-project.org/package=xgboost.

Copernicus. (2021). *Sentinel-5P NRTI NO2: Near Real-Time Nitrogen Dioxide*. https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_NRTI_L3_NO2#bands. last accessed August 3, 2021.

Cressie, N. (2015). *Statistics for Spatial Data*. Hoboken, NJ: John Wiley & Sons.

Dee, D. P., S. M. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, D. P. Bauer, et al. (2011). "The Era-Interim Reanalysis: Configuration and Performance of the Data Assimilation System." *Quarterly Journal of the Royal Meteorological Society* 137(656), 553–97.

Diggle, P. J., R. Menezes, and T. L. Su. (2010). "Geostatistical Inference Under Preferential Sampling." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59(2), 191–232.

Diggle, P. J., P. Moraga, B. Rowlingson, and B. M. Taylor. (2013). "Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm." *Statistical Science* 28(4), 542–63.

Dijkstra, L. and H. Poelman. (2014). *A Harmonised Definition of Cities and Rural Areas: The New Degree of Urbanisation*. https://ec.europa.eu/regional_policy/sources/docgener/work/2014_01_new_urban.pdf. Last accessed August 4, 2021.

Dinsdale, D., and M. Salibian-Barrera. (2019). "Methods for Preferential Sampling in Geostatistics." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68(1), 181–98.

Duan, T., A. Anand, D. Y. Ding, K. K. Thai, S. Basu, A. Ng, and A. Schuler. (2020). "Ngboost: Natural gradient boosting for probabilistic prediction." In *International Conference on Machine Learning*, 2690–700. PMLR.

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

Fouedjio, F., and J. Klump. (2019). "Exploring Prediction Uncertainty of Spatial Data in Geostatistical and Machine Learning Approaches." *Environmental Earth Sciences* 78(1), 38.

Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, 29(5), 1189–232.

Gonzalez-Estrada E., and J. A. Villasenor-Alva. (2020). *goft: Tests of Fit for Some Probability Distributions*. R Package Version 1.3.6. https://CRAN.R-project.org/package=goft.

Hastie, T., R. Tibshirani, and J. Friedman. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer Science & Business Media.

Hastie, T., R. Tibshirani, and J. Friedman. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer Science & Business Media.

Hoek, G., R. Beelen, K. De Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs. (2008). "A Review of Land-Use Regression Models to Assess Spatial Variation of Outdoor Air Pollution." *Atmospheric Environment* 42(33), 7561–78.

James, G., D. Witten, T. Hastie, and R. Tibshirani. (2013). *An Introduction to Statistical Learning*, Vol 112. New York: Springer.

A. Jordan, F. Krüger, and S. Lerch. (2017). Evaluating Probabilistic Forecasts with Scoringrules. *arXiv preprint arXiv:1709.04743*.

Kerckhoffs, J., G. Hoek, L. Portengen, B. Brunekreef, and R. C. Vermeulen. (2019). "Performance of Prediction Algorithms for Modeling Outdoor Air Pollution Spatial Surfaces." *Environmental Science & Technology* 53(3), 1413–21.

Krainski, E. T., V. Gómez-Rubio, H. Bakka, A. Lenzi, D. Castro-Camilo, D. Simpson, F. Lindgren, and H. Rue. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Boca Raton: CRC Press.

Larkin, A., J. A. Geddes, R. V. Martin, Q. Xiao, Y. Liu, J. D. Marshall, M. Brauer, and P. Hystad. (2017). "Global Land Use Regression Model for Nitrogen Dioxide Air Pollution." *Environmental Science & Technology* 51(12), 6957–64.

J. J. Li, A. Jutzeler, B. Faltings, S. Winter, and C. Rizos. (2014). Estimating Urban Ultrafine Particle Distributions with Gaussian Process Models. Research@ Locate14, 145–53.

Lindgren, F., H. Rue, and J. Lindström. (2011). "An Explicit Link between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(4), 423–98.

Lindgren, F., H. Rue, et al. (2015). "Bayesian Spatial Modelling with R-INLA." *Journal of Statistical Software* 63(19), 1–25.

Liu, Y., G. Cao, and N. Zhao. (2020). "Integrate Machine Learning and Geostatistics for High-Resolution Mapping of Ground-Level pm2. 5 Concentrations." *Spatiotemporal Analysis of Air Pollution and Its Application in Public Health*, 135–51. Elsevier.

Lu, M., O. Schmitz, K. de Hoogh, Q. Kai, and D. Karssenberg. (2020a). "Evaluation of Different Methods and Data Sources to Optimise Modelling of NO2 at a Global Scale." *Environment International* 142, 105856. https://doi.org/10.1016/j.envint.2020.105856

Lu, M., I. Soenario, M. Helbich, O. Schmitz, G. Hoek, M. van der Molen, and D. Karssenberg. (2020b). "Land Use Regression Models Revealing Spatiotemporal Co-Variation in NO2, NO, and O3 in the Netherlands." *Atmospheric Environment* 223, 117238.

M. Lu, R. Dai, C. de Boer, O. Schmitz, I. Kooter, S. Cristescu, and D. Karssenberg. (2021). Problems in Statistical Modelling of Air Pollution Basing Solely on Ground Monitor Stations and a Novel Mobile Sensing Instrument Solution. submitted to Science of the Total Environment.

Lundberg, S. M., and S.-I. Lee. (2017). "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems* Vol 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Long Beach, CA: Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Lundberg, S. M., B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, et al. (2018). "Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia During Surgery." *Nature Biomedical Engineering* 2(10), 749.

Martins, T. G., D. Simpson, F. Lindgren, and H. Rue. (2013). "Bayesian Computing with INLA: New Features." *Computational Statistics & Data Analysis* 67, 68–83.

Meinshausen, N. (2006). "Quantile Regression Forests." *Journal of Machine Learning Research* 7(Jun), 983–99.

Meyer, H., and E. Pebesma. (2021). "Predicting Into Unknown Space? Estimating the Area of Applicability of Spatial Prediction Models." *Methods in Ecology and Evolution* 12(9), 1620–33.

Meyer, H., C. Reudenbach, T. Hengl, M. Katurji, and T. Nauss. (2018). "Improving Performance of Spatio-Temporal Machine Learning Models Using Forward Feature Selection and Target-Oriented Validation." *Environmental Modelling & Software* 101, 1–9.

Moraga, P. (2019). *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*. Boca Raton, FL: Chapman & Hall/CRC Press.

Moraga, P., S. M. Cramb, K. L. Mengersen, and M. Pagano. (2017). "A Geostatistical Model for Combined Analysis of Point-Level and Area-Level Data Using INLA and SPDE." *Spatial Statistics* 21, 27–41.

Moraga, P., C. Dean, J. Inoue, P. Morawiecki, S. Raja Noureen, and F. Wang. (2021). "Bayesian Spatial Modelling of Geostatistical Data Using INLA and SPDE Methods: A Case Study Predicting Malaria Risk in Mozambique." *Spatial and Spatio-Temporal Epidemiology* 39, 100440.

NASA. *Shuttle Radar Topography Mission* https://www2.jpl.nasa.gov/srtm/dataprelimdescriptions.html. Last accessed August 15, 2021.

Nelson, D. A. (1999). "European Environment Agency." *Colorado Journal of International Environmental Law and Policy* 10, 153.

NOAA. (2021). DMSP and VIIRS Data Download. https://ngdc.noaa.gov/eog/download.html. Accessed March 11, 2021.

OpenStreetMap contributors. (2019). Planet Dump January 7, 2019. https://planet.osm.org.

Ploton, P., F. Mortier, M. Réjou-Méchain, N. Barbier, N. Picard, V. Rossi, C. Dormann, G. Cornu, G. Viennois, N. Bayol, et al. (2020). "Spatial Validation Reveals Poor Predictive Performance of Large-Scale Ecological Mapping Models." *Nature Communications* 11(1), 1–11.

Ren, X., Z. Mi, and P. G. Georgopoulos. (2020). "Comparison of Machine Learning and Land Use Regression For Fine Scale Spatiotemporal Estimation of Ambient Air Pollution: Modeling Ozone Concentrations Across The Contiguous United States." *Environment International* 142, 105827. https://doi.org/10.1016/j.envint.2020.105827, https://www.sciencedirect.com/science/article/pii/S0160412020317827

Rue, H., and L. Held. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL: CRC Press.

Rue, H., S. Martino, and N. Chopin. (2009). "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71(2), 319–92.

Rybarczyk, Y., and R. Zalakeviciute. (2018). "Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review." *Applied Sciences* 8(12), 2570.

Schlosser, L., T. Hothorn, R. Stauffer, A. Zeileis, et al. (2019). "Distributional Regression Forests for Probabilistic Precipitation Forecasting in Complex Terrain." *The Annals of Applied Statistics* 13(3), 1564–89.

Shaddick, G., M. L. Thomas, H. Amini, D. Broday, A. Cohen, J. Frostad, A. Green, S. Gumy, Y. Liu, R. V. Martin, et al. (2018). "Data integration for the Assessment of Population Exposure to Ambient Air Pollution for Global Burden of Disease Assessment." *Environmental Science & Technology* 52(16), 9069–78.

Stasinopoulos, D. M., R. A. Rigby, et al. (2007). "Generalized Additive Models for Location Scale and Shape (GAMLSS) in R." *Journal of Statistical Software* 23(7), 1–46.

Stein, M. L. (1999). "Linear Prediction" *Interpolation of Spatial Data: Some Theory for Kriging*. SPRINGER Science & Business Media, 12.

Suykens, J. A., and J. Vandewalle. (1999). "Least Squares Support Vector Machine Classifiers." *Neural Processing Letters* 9(3), 293–300.

J. Velthoen, C. Dombry, J.-J. Cai, and S. Engelke. (2021). Gradient Boosting for Extreme Quantile Regression. *arXiv preprint arXiv:2103.00808*.

Vicedo-Cabrera, A. M., A. Biggeri, L. Grisotto, F. Barbone, and D. Catelan. (2013). "A Bayesian Kriging Model for Estimating Residential Exposure to Air Pollution of Children Living in a High-Risk Area in Italy." *Geospatial Health* 8(1), 87–95.

Villaseñor, J. A., and E. González-Estrada. (2015). "A Variance Ratio Test of Fit for Gamma Distributions." *Statistics & Probability Letters* 96, 281–6.

Wadoux, A. M.-C., G. B. Heuvelink, S. De Bruin, and D. J. Brus. (2021). "Spatial Cross-Validation is not the Right Way to Evaluate Map Accuracy." *Ecological Modelling* 457, 109692.

Wager, S., T. Hastie, and B. Efron. (2014). "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife." *The Journal of Machine Learning Research* 15(1), 1625–51.

Wang, Q., H. Feng, H. Feng, Y. Yu, J. Li, and E. Ning. (2021). "The Impacts of Road Traffic on Urban Air Quality in Jinan Based GWR and Remote Sensing." *Scientific Reports* 11(1), 1–9.

Young, M. T., M. J. Bechle, P. D. Sampson, A. A. Szpiro, J. D. Marshall, L. Sheppard, and J. D. Kaufman. (2016). "Satellite-Based NO2 and Model Validation in a National Prediction Model Based on Universal Kriging and Land-Use Regression." *Environmental Science & Technology* 50(7), 3686–94.

Yuan, C. (2011). Models and Methods for Computationally Efficient Analysis of Large Spatial and Spatio-Temporal Data.

Zhai, L., S. Li, B. Zou, H. Sang, X. Fang, and S. Xu. (2018). "An Improved Geographically Weighted Regression Model for pm2. 5 Concentration Estimation in Large Areas." *Atmospheric Environment* 181, 145–54.

Zhan, Y., Y. Luo, X. Deng, K. Zhang, M. Zhang, L. Grieneisen, and B. Di. (2018). "Satellite-Based Estimates of Daily NO_2 Exposure in China Using Hybrid Random Forest and Spatiotemporal Kriging Model." *Environmental Science & Technology* 52(7), 4180–9.

Zou, B., Q. Pu, M. Bilal, Q. Weng, L. Zhai, and J. E. Nichol. (2016). "High-resolution Satellite Mapping of Fine Particulates Based on Geographically Weighted Regression." *IEEE Geoscience and Remote Sensing Letters* 13(4), 495–9.