Computers, Environment and Urban Systems

A comparison of geostatistical and non-spatial machine learning methods in NO2 modelling: prediction accuracy, uncertainty quantification, and model interpretation

--Manuscript Draft--

Manuscript Number:	
Article Type:	Research Paper
Keywords:	geostatistics; machine learning; spatial prediction; model comparison; prediction interval; model interpretation
Corresponding Author:	Meng Lu University of Bayreuth bayreuth, Bayern GERMANY
First Author:	Meng Lu
Order of Authors:	Meng Lu
	Joaquin Cavieres
	Paula Moraga
Abstract:	NO2 is a traffic-related air pollutant that is strongly associated with cardiovascular and respiratory diseases. Ground NO2 monitoring stations measure NO2 concentrations at certain locations and statistical predictive methods have been developed to predict NO2 as a continuous surface to inform decision-making. Among them, machine learning methods are the most powerful in capturing non-linear relationships between NO2 measurements and geospatial predictors, but it is unclear if the spatial structure of NO2 is also captured in the response-covariates relationships. In addition, most model comparison studies only compare accuracy in the prediction mean at ground stations, but do not consider prediction intervals and model interpretation and the effects of different model evaluation methods. In this study, we dive into the comparison between spatial and non-spatial data models accounting for the above-mentioned aspects. Moreover, we implemented a spatial and a non-spatial methods that have not been applied to air pollution mapping before and evaluated stack learning methods with and without modelling the spatial process. We implemented our study using national ground station measurements of NO2 in Germany and Netherlands of the year 2017, predicting NO2 to 100 m resolution grid. Our results indicate the importance of modelling the spatial process especially in areas close to traffic. The prediction intervals predicted with ensemble tree-based methods are satisfactory but too narrow with the geostatistical methods. Compared to ensemble tree-based methods, the geostatistical methods provide important spatial information for analysing emission sources and the spatial process of observations.
Suggested Reviewers:	Orietta Nicolis orietta.nicolis@unab.cl Spatial statistics and air quality modelling
	Jules Kerckhoffs j.kerckhoffs@uu.nl Statistical and machine learning methods for air quality mapping.
	Andrew Larkin Andrew.Larkin@oregonstate.edu Big data and air pollution exposure assessment
	Gavin Shaddick G.Shaddick@exeter.ac.uk Geostatistics, Bayesian modelling, air quality and environmental health.
Opposed Reviewers:	

Dear Prof. Tony H. Grubesic,

Please find attached to this letter our research paper

“A comparison of geostatistical and non-spatial machine learning methods in NO2 modelling: prediction accuracy, uncertainty quantification, and model interpretation,

to be considered for publication in *Computers, Environment and urban systems.*

NO2 is a traffic-related air pollutant that negatively affect our health. Machine learning (ML) methods have shown to be powerful in capturing non-linear relationships between NO2 measurements and geospatial predictors for spatial prediction of NO2. but it is unclear if the spatial structure of NO2 is sufficiently captured in the response-covariates relationships and how prediction intervals and model interpretation derived from ML models compare with geostatistical models. In addition, it is commonly not evaluated how different models behave in different geographical areas.

In this study, we compared geostatistical methods with ML methods in the spatial prediction of NO2. We developed a comparison process that comprehensively compare the predicted mean, prediction intervals, and model interpretation of different spatial and non-spatial models. Spatial and non-spatial CV strategies are used to reveal prediction accuracy in different aspects. We also implemented two methods that to our knowledge have not been applied in air pollution mapping, one post-processes quantile random forest with L1-norm shrinkage (Lasso) regression and the other geostatistical stacked learning. These two methods considerably improve from the original (quantile) random forest and stacked learning methods, respectively.

With geospatial predictors and ground observations becoming increasingly available, many statistical methods have been developed in NO2 mapping, but a study that comprehensively dive into models with different structures and complexity to understand the strength and limitations of each is lacking in air pollution mapping. Our comparison study is important in understanding different model behaviours and pointing out good practices in spatial prediction of air pollution and future directions for improvements. We also paid full attention to computational efficiency in the methods applied, all the methods we applied are highly scalable and the Lasso post-processing further reduces model redundancy. We therefore strongly believe our study is highly relevant to the domain of applied mathematics in *Computers, Environment and Urban Systems* .

Thank you very much for your consideration,

Dr. Meng Lu,
On behalf of Joaquin Cavieres and Dr. Paula Moraga

Highlight

- Geostatistical and machine learning methods show different strength and limitations.
- The model comparison concerns prediction intervals and model interpretation.
- Post-processing random forest with Lasso regression outperforms random forest.
- Geostatistical stacked learning outperforms stacked learning methods.
- It is important to model spatial structure in national NO2 mapping.
- Non-spatial machine learning methods may not fully capture the spatial process.
- Spatial and non-spatial CV strategies affect model comparison.

A comparison of geostatistical and non-spatial machine learning methods in NO_2 modelling: 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, 95447 Bayreuth, Germany

meng.lu@uni-bayreuth.de

[2]Instituto de Estadística, Facultad de Ciencias, Universidad de Valparaíso, Valparaíso, Chile

joaquin.cavieres@uv.cl

[3]Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

paula.moraga@kaust.edu.sa

*Corresponding Author

Abstract

NO_2 is a traffic-related air pollutant that is strongly associated with cardiovascular and respiratory diseases. 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 to inform decision-making. Among them, machine learning methods are the most powerful in capturing non-linear 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. In addition, most model comparison studies only compare accuracy in the prediction mean at ground stations, but do not consider prediction intervals and model interpretation and the effects of different model evaluation methods. In this study, we dive into the comparison between spatial and non-spatial data models accounting for the above-mentioned aspects. Moreover, we implemented a spatial and a non-spatial methods that have not been applied to air pollution mapping before and evaluated stack learning methods with and without modelling the spatial process. We implemented our study using national ground station measurements of NO_2 in Germany and Netherlands of the year 2017, predicting NO_2 to 100 m resolution grid. Our results indicate the importance of modelling the spatial process especially in areas close to traffic. The prediction intervals predicted with ensemble tree-based methods are satisfactory but too narrow with the geostatistical methods. Compared to ensemble tree-based methods, the geostatistical methods provide important spatial information for analysing emission sources and the spatial process of observations.

Keywords: geostatistics; machine learning; spatial prediction; model comparison; prediction interval; model interpretation

1 Introduction

NO_2 is a traffic-related air pollutant and has been found in epidemiological time series analysis to highly associated with respiratory (Luo et al., 2016) and cardiovascular (Chiusolo et al., 2011) diseases. NO_2 values are measured using monitoring stations at certain locations (e.g. close to traffic) and most of the epidemiological studies identified the relationships between NO_2 and diseases or hospital admission using a single NO_2 monitoring station to represent the entire district. However, NO_2 is highly dynamic over the district and the difference in NO_2 concentrations will reflect on personal exposures to NO_2. Detailed spatial mapping of NO_2 is therefore required for more accurate quantification of the relationships between NO_2 and health effects. In addition, detailed NO_2 maps are necessary for scientific recommendations to be provided to policymakers and city planners.

Statistical methods for NO_2 mapping have attracted a lot of attention with the burgeoning Machine Learning (ML)[1] methods and availability of ground monitoring station networks, atmospheric satellite products, and geospatial predictors. Geospatial predictors are variables that are included as covariates in a statistical air pollution model. Commonly used geospatial predictors are air emission- (e.g. road networks) and dispersion-related (e.g. wind speed) variables, numerical modelling (e.g. with chemistry transport model) output, and atmospheric remote sensing measurements or products. A most recent (data available from Jan-2018) atmosphere sensing instrument, Tropomi

[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: Generalised 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

(Tropospheric monitoring instrument, NSO and ESA, 2019) onboard of Sentinel 5p satellite, measures column density of a variety of gaseous air pollutants, in particular with an unprecedentedly high resolution for NO_2 (3.5 km x 5.5 km, across along track, since 06 August 2019).

Statistical methods applied for spatial air pollution prediction can be broadly classified depending on whether the spatial dependency is explicitly modelled. If not modelled, 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 1 km or coarser (Young et al., 2016; Shaddick et al., 2018; Beloconi and Vounatsou, 2020). Non-spatial 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 assumes linear relationships between NO_2 and geospatial predictors are the most studied (Briggs et al., 2000; Hoek et al., 2008). Most recently, statistical learning (in this study, "statistical learning" is used interchangeably with "machine learning") methods (Hastie et al., 2009), including regularised linear regression (e.g. Lasso and Ridge regression (James et al., 2013)), kernel methods such as 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 or capturing non-linear response-covariate relationships (Lu et al., 2020a; Chen et al., 2019a). In air pollution (not restricted to NO_2) mapping, several studies compared between statistical learning and conventional LUR methods (Chen et al., 2019a; Kerckhoffs et al., 2019; Lu et al., 2020a; Ren et al., 2020; Rybarczyk and Zalakeviciute, 2018).

Geostatistical models (e.g. Kriging) and Geographically Weighted Regression (GWR) are the most used spatial methods for air pollution prediction (Vicedo-Cabrera et al., 2013; Li et al., 2014; Wang et al., 2021; Zou et al., 2016) and these methods have been combined with dimension reduction Zhai et al. (2018) and RF (Zhan et al., 2018; Liu et al., 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 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 modelling dependency of residuals from the mean (Bhatt et al., 2017). Conventional Kriging methods suffer from the "big n problem", i.e. it may become computationally intractable with a large number of observations. To deal with this problem, Lindgren et al. (2011) propose to use Stochastic Partial Differential Equations (SPDE) to approximate the Gaussian Random Field (GRF) to a Gaussian Markov Random Field (GMRF, Rue and Held (2005)). The main advantage of this is that the GMRF has a sparse structure of the precision matrix, which is the inverse of the covariance matrix of a GRF. Along with this, Rue et al. (2009) propose to use the Integrated Nested Laplace Approximation (INLA) in a Bayesian framework to achieve the computational scalability of a geostatistical model using approximations for all the estimations. This is especially advantageous when modelling NO_2 over a larger scale e.g., continental or global-scale modelling when a large amount of observations are modelled, and in spatiotemporal modelling.

As spatial models are typically more complex compared to their non-spatial counterparts, several studies compared spatial and non-spatial models to understand if the spatial effects could be simply modelled 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) shows 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 modelled with non-linear ML methods (Lu et al., 2020a). Most recent studies attempt to replace the linear mean function with ML models. Liu et al. (2020) applied a geostatistical model to the residuals from an RF model for the spatial prediction of $PM_{2.5}$. In disease mapping, Bhatt et al. (2017) proposes to stack ML models to replace the mean function in a geostatistical model.

Few studies have compared between geostatistical and ML methods, possibly because the ML methods are still relatively less studied in air pollution mapping and in the field of geostatistics. It might be more interesting to compare between geostatistical methods and ML methods than geostatistical methods and LUR, because ML methods may be more capable of (implicitly) capturing the spatial dependency by integrating covariates, when the number of observations is sufficient. Moreover, most comparison studies only compare the cross-validation accuracy of the prediction mean (e.g. using R-squared, mean absolute error, or root mean squared error), ignoring the prediction intervals. Also not discussed is the cause of the prediction errors, are they caused by missing covariants, violation of the model assumptions (e.g. data distribution, non-linearity), or inconsistent distributions between training and validation sets. Also, different cross-validation strategies, e.g., how do we split the train-test sets, may lead to different model validation results. Current studies typically solely rely on k-fold splitting (Kerckhoffs et al., 2019; Larkin et al., 2017; Ren et al., 2020) or bootstrapping (Lu et al., 2020a) to randomly splitting between train-test sets, which may be one-sided and does not provide an indication of accuracy in spatial blocks (but only at the locations of ground stations).

In this study, we focus on ensemble tree-based methods (e.g. RF and boosting) in the ML

category and a hierarchical spatial model (Lindgren et al., 2015; Blangiardo and Cameletti, 2015; Moraga, 2019) called latent Gaussian model in the geostatistics category. Additionally, we invest in stacked models in integrating ML and geostatistical models and develop a LUR model using Lasso for comparison. 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; Wager et al., 2014; Stasinopoulos et al., 2007; Alakus et al., 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 recognisable 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-recognised method is distributional regression forests (DF) (Schlosser et al., 2019), which embeds the GAMLSS (Generalised Additive Models for Location Scale and Shape) (Stasinopoulos 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 modelled.

The objective of our study is to compare geostatistics and non-spatial ensemble tree-based models for NO_2 mapping, in terms of their prediction accuracy, uncertainty quantification, and model interpretation and to understand effect of modelling spatial structures. More specifically, the following sub-objectives are reached:

1. Optimising a set of spatial hierarchical and ML models for NO_2 prediction in Germany and the Netherlands.
2. Developing a non-spatial and a geostatistical stacked ensemble model, i.e., 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 et al., 2009; Martins et al., 2013). XGB, RF and Lasso are chosen for the comparison with the geostatistical model and they also form the base learners in the two (geostatistical and non-spatial) stacked learning models. The ML methods are chosen for their dissimilarity. Specifically, Lasso is a linear regression model without accounting for spatial dependency. RF and XGB are non-linear models with regression trees as base-learners and are not affected by dependent covariates. XGB is a highly scalable boosting method that builds tree models subsequently over the residuals of previous trees and has multiple routines to penalise model over-fitting (Chen et al., 2019b), which has been reported in various studies to obtain the highest prediction accuracy Lu et al. (2020a).

2 Data

NO_2 concentration measurements of 2017 from national ground stations of Germany and the Netherlands are used. The original hourly data is downloaded from the EEA (European Environment

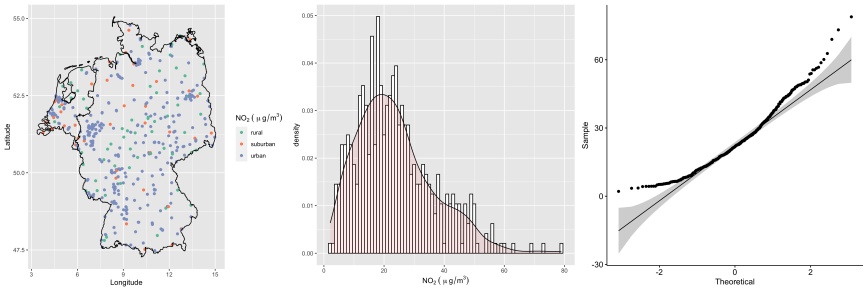


Figure 1: Spatial distribution of NO_2 stations, histogram and Q-Q plot of the NO_2 measurements.

Agency, Nelson, 1999; EEA, 2021). Negative values are considered as missing. 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 significant 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.

The geospatial predictor grids (table 1) are calculated or re-sampled at 100 m resolution. They are either spatial attributes aggregated in a circular ring centred 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 (NASA), 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 calculated from OpenStreetMap (OpenStreetMap contributors, 2019). For detailed descriptions of the processing of the geospatial predictors please refer to Lu et al. (2020a).

3 Methods

The methods considered in this study are classified as spatial and non-spatial 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 the spatial hierarchical model to stacked learning with Lasso, RF and XGB models as base learners;

Non-spatial 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 et al., 2009);
6. SE: stacked 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).

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 m, 300 m, 500 m, 800 m, 1000 m, 3000 m and 5000 m. The night lights digital numbers are calculated with buffer radii of 450 m, 900 m, 3150 m 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

3.1 Non-spatial methods

Lasso is a linear regression algorithm with the L1 regularisation 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 choose features from this subset to build the tree. RF (Breiman, 2001) grows trees independently and then take the mean of the predictions of each tree.

QRF is a non-parametric prediction interval estimation method which keeps all the observations in the terminal node for estimating the conditional probability function. Specifically, 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 post-processing of QRF (Hastie et al., 2017, page 617). This method firstly 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 divide 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 co-variate. 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 penalisation

paths to control model complexity to prevent model over-fitting, including regularisation (e.g. L1 regularisation) 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 is 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 (Stacking Ensemble) refers to a class of algorithms that trains a second-level "meta-learner" to optimise the combination of a collection of prediction algorithms (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-one data) are used to train the meta-learner.

3.2 Hyperparameter setting for XGB and RF

To optimise the hyperparameters of XGB (known as "model tuning"), we used grid search to optimise hyperparameters in 5-fold cross-validation basing 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, 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 3 to 6 with a step of 1, learning rate (eta) from 0.001 to 0.1 with a step of 0.05, the penalty term Gamma (Chen et al., 2019b) from 1

to 5 with a step of 1, the subsample is set to 0.7, L1 norm penalisation (lambda) is set to 2 and L2 norm penalisation (alpha) is set to 0. RF is not sensitive to hyperparameter tuning. 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.

3.3 Geostatistical models

Suppose we assume that NO_2 values y_i measured at locations $\boldsymbol{s_i}$, $i = 1, \ldots, n$, follows a Gaussian distribution with mean μ_i and variance σ^2, where the mean μ_i is expressed as a sum of covariates and a spatially structured random effect following a zero-mean Gaussian process with a spatial covariance function (Moraga, 2019).

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

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

Here, $\boldsymbol{d_i} = (d_{i1}, \ldots, d_{ip})$ is the vector of covariates at location $\boldsymbol{s_i}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the coefficient vector, and $\boldsymbol{x}(\boldsymbol{s_i})$ denotes a spatial Gaussian random field. That is, $\{\boldsymbol{x}(\boldsymbol{s}_1), ..., \boldsymbol{x}(\boldsymbol{s}_n)\} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma})$, where N_n is a Normal multivariate distribution for the spatial process specified by its mean $\mathbb{E}(\boldsymbol{x}(\boldsymbol{s}))$, and covariance function $C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \text{Cov}(\boldsymbol{x}(\boldsymbol{s}_1), \boldsymbol{x}(\boldsymbol{s}_2))$. 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 $C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \text{Cov}(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|)$ and its dependence is commonly modeled using a Matérn function (Stein (2012); Yuan (2011); Diggle et al. (2013)). Since incorporating the spatial dependence directly with a large number of observations using a Gaussian random field is computationally expensive, 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 (1) can be represented by $y_i \sim Gamma(\alpha, \beta)$ where α is the shape parameter and β the rate parameter. The INLA-SE model uses a Gaussian likelihood.

3.4 INLA and SPDE

To fit the geostatistical models, we use the `R` package `INLA` which facilitates the application of the INLA and the SPDE approaches. Following the expression proposed in (1), the structure for the hierarchical model is:

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

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

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

where $\boldsymbol{\theta}$ is the vector of hyperparameters with $\theta_1 = \sigma^2$, $\theta_2 = \{\log(\tau), \log(\kappa)\}$, $\boldsymbol{x}$ is the spatial latent field, $\boldsymbol{A}$ is the projector matrix and $\boldsymbol{y}$ is the vector of the response variable $f(\cdot \mid \boldsymbol{x}, \boldsymbol{\theta})$, commonly from the exponential family of distributions. $\boldsymbol{D}$ is a covariate matrix and $\boldsymbol{\beta}$ a coefficient matrix.

The `R` package `INLA` can be used to perform direct numerical calculation of the posterior distribution for a Bayesian hierarchical model (Rue et al. (2009), Martino and Rue (2009)). If we use $\boldsymbol{x}$ as a latent Gaussian field (a Gaussian Markov random field), $\boldsymbol{\theta}$ a vector of (hyper)parameters 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:

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i \in \mathcal{I}} p(y_i | \eta_i, \boldsymbol{\theta}), \tag{6}$$

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 density for the posterior of the spatial latent field and the hyperparameters. The marginal densities can be obtained:

$$p(x_i \mid \boldsymbol{y}) = \int p(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}, \tag{7}$$

and

$$p(\boldsymbol{\theta}_j | \boldsymbol{y}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}_{-j}. \tag{8}$$

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

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

$$(\kappa^2 - \Delta)^{\alpha/2} (\tau(\boldsymbol{s}) x(\boldsymbol{s})) = \boldsymbol{\mathcal{W}}(\boldsymbol{s}), \tag{9}$$

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 et al. (2015)). For the above we can use a Gaussian Markov random field that approximates to 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 $\mathcal{O}(n^{3/2})$.

3.5 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 inferencing and 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 standard error 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 modelling process applies the same bootstrapped samples for training and validation. In addition, 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 Lasso selected variables, besides it does not choose 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.

3.6 INLA model parameterisation

The triangulated mesh constructed in the SPDE approach is shown in supplementary material (supfig. 1), with size of the inner and outer extensions around the data locations (*offsets*) 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 Matern

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 besides 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

SPDE model is constructed with $\alpha = 2$. The SE-INLA model has the same specification (i.e. mesh structure, likelihood, objective function, priors, optimisation process) as the INLA model parameterisation described above.

4 Model evaluation

4.1 Cross validation

We use RMSE, MAE (Mean Absolute Error), IQR (Interquartile Range) and R^2 (R-squared) to 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. R^2 indicates the explained variance and is calculated as $R^2 = 1 - var(error)/var(y)$, where var(.) indicates variance, error indicates model residuals 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.

3. Customised CV. Splitting train-test based on values of certain covariates. In this study, three sub-areas 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 that is in the last quartile. Low population is defined as the variable population of 1000 m buffer is below the median. Close to road is defined as (please refer to table 1 for the definition of covariates):

```
road_class_2_100 > 0 |
road_class_1_100 > 0 |
road_class_3_100 > quantile(road_class_3_100, .75))
```

Far away from road is defined as:

```
road_class_2_100 == 0 &
road_class_1_100 == 0 &
road_class_3_100 < quantile(road\_class\_3\_100, .5)
```

where "`&`" indicates "and" and "`|`" indicates "or". The second variable of the function "`quantile(.)`" indicates the percentage quantile of the variables.

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 for CV. For example, 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.

4.2 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 similarities between

two distributions. We use it to indicate how the predicted distribution matches the observed distribution. The CRPS implemented as an R package `ScoringRules` (Jordan et al., 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 quantile 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.

4.3 Model interpretation

We inspect fixed and spatial random effects modelled by INLA and compare the spatial random field with modelled 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; Lundberg and Lee, 2017), a unified method based on additive feature attribution, to estimate variable influence in RF and XGB models.

5 Results

5.1 Accuracy assessment and uncertainty quantification

Non-spatial CV

Both ensemble tree-based methods with a Gaussian objective function and INLA with a Gaussian likelihood function obtain higher prediction accuracy than Lasso (table 3), indicating the necessity of using a more flexible model and modelling spatial random fields. Among individual methods, in terms of R^2 and RMSE, INLA with Gaussian likelihood obtained the highest prediction accuracy, followed by XGB-G and QRFLA. QRFLA greatly improves from 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 R^2 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 improves prediction accuracy compared to SE and INLA, obtained the best results in terms of root mean squared error (6.83, 24.5% of the mean of observations) and R^2 (0.71). This indicates the spatial structures could further improve prediction accuracy despite flexible relationships captured from ML models.

Spatial-blocked CV

Spatial-blocked CV provides information about prediction accuracy in spatial blocks. The R^2 map (fig. 2) shows that the XGB, RF and INLA predict relatively well in most parts of Germany besides blocks at the boundaries. The R^2 for the block western the Netherlands is also relatively low with all the three methods and especially for XGB (R^2: 0.2). RF obtains the best result for the block of western the Netherlands (R^2: 0.5). The INLA model outperforms RF and XGB in the blocks at south-east and north. The R^2 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 (the DF is not shown as it will

Table 3: Prediction accuracy matrix for different models using 20 times bootstrapped cross-validation. Non-spatial 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
R^2	0.65	0.65	0.68	0.51	0.67	0.69	0.69	0.45	0.71

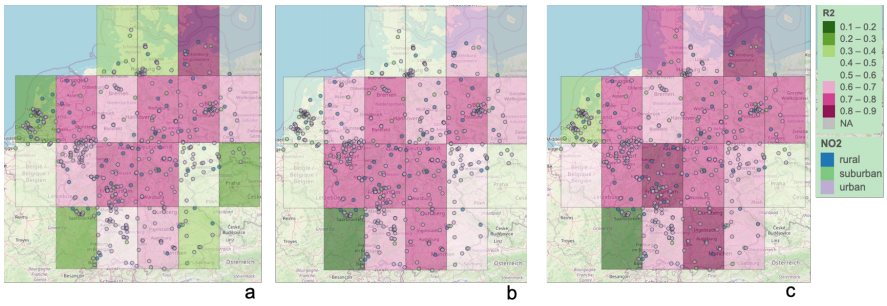


Figure 2: The R-squared of each block, using the rest of the blocks for training. The models are a) XGB, b) QRF, c) INLA.

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

be shown that the QRF and DF performed similarly in prediction interval prediction (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 R^2 observed for the same block. However, some blocks with relatively high R^2 (in the north and south) have high CRPS. This indicates that the prediction mean is well-predicted but not the prediction interval (too narrow).

Customised 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 non-spatial 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 modelling spatial dependency in areas close to traffic and possibly non-linear 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.

5.2 Prediction interval

The 90% prediction intervals for INLA, INLA-G, DF, QRF and QRFLA are shown in figs. 4 to 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 to better capture extreme high values but the model also turned to miss more at lower values. The QRFLA predicted a slightly narrower prediction interval compared to QRF. This indicates that Lasso reduced the variance of a QRF model by aggregating trees.

Table 4: Results with customised 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	R^2
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

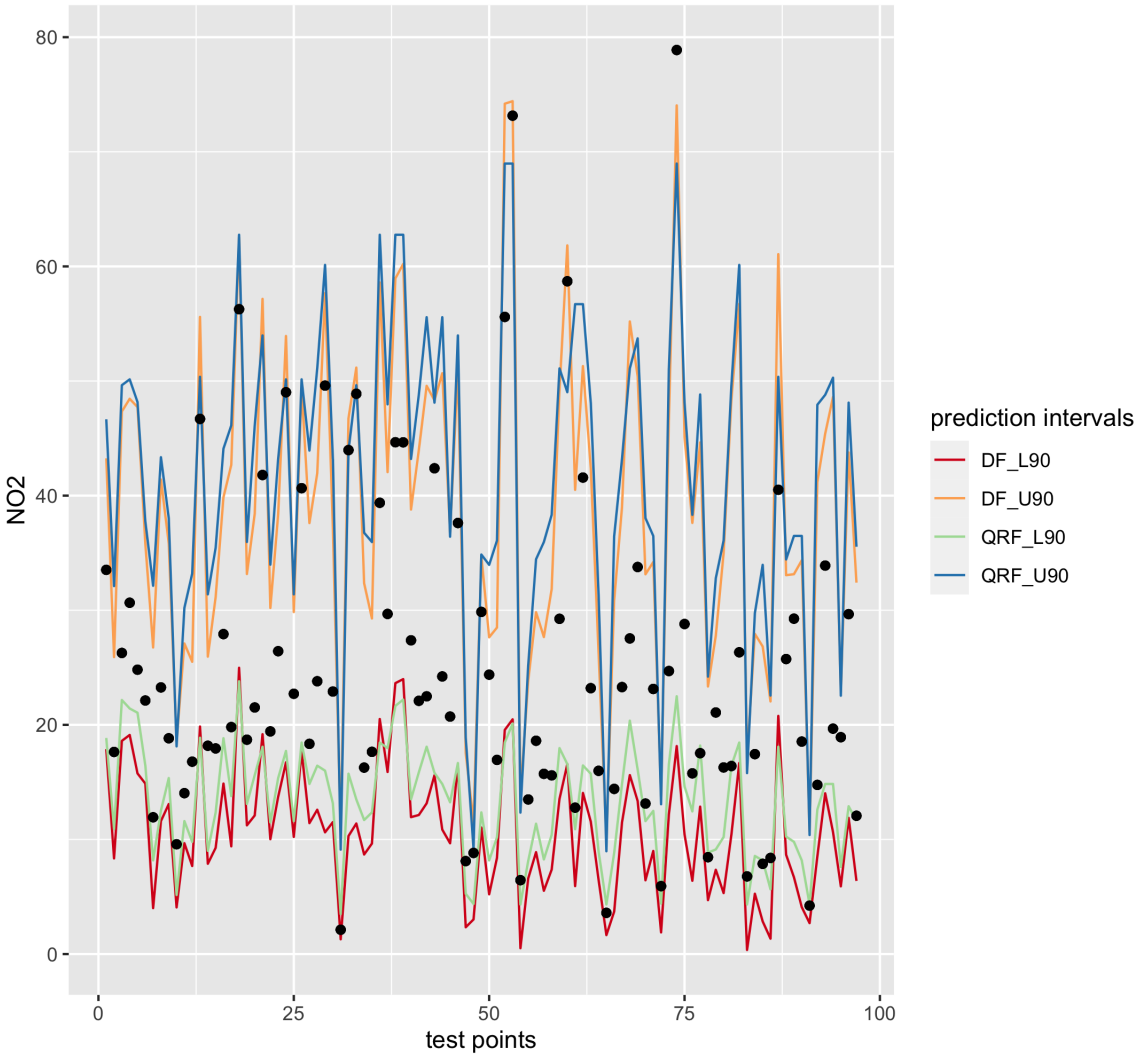


Figure 4: The 90% prediction interval predicted by DF and QRF. The black dots indicate observations in the test dataset.

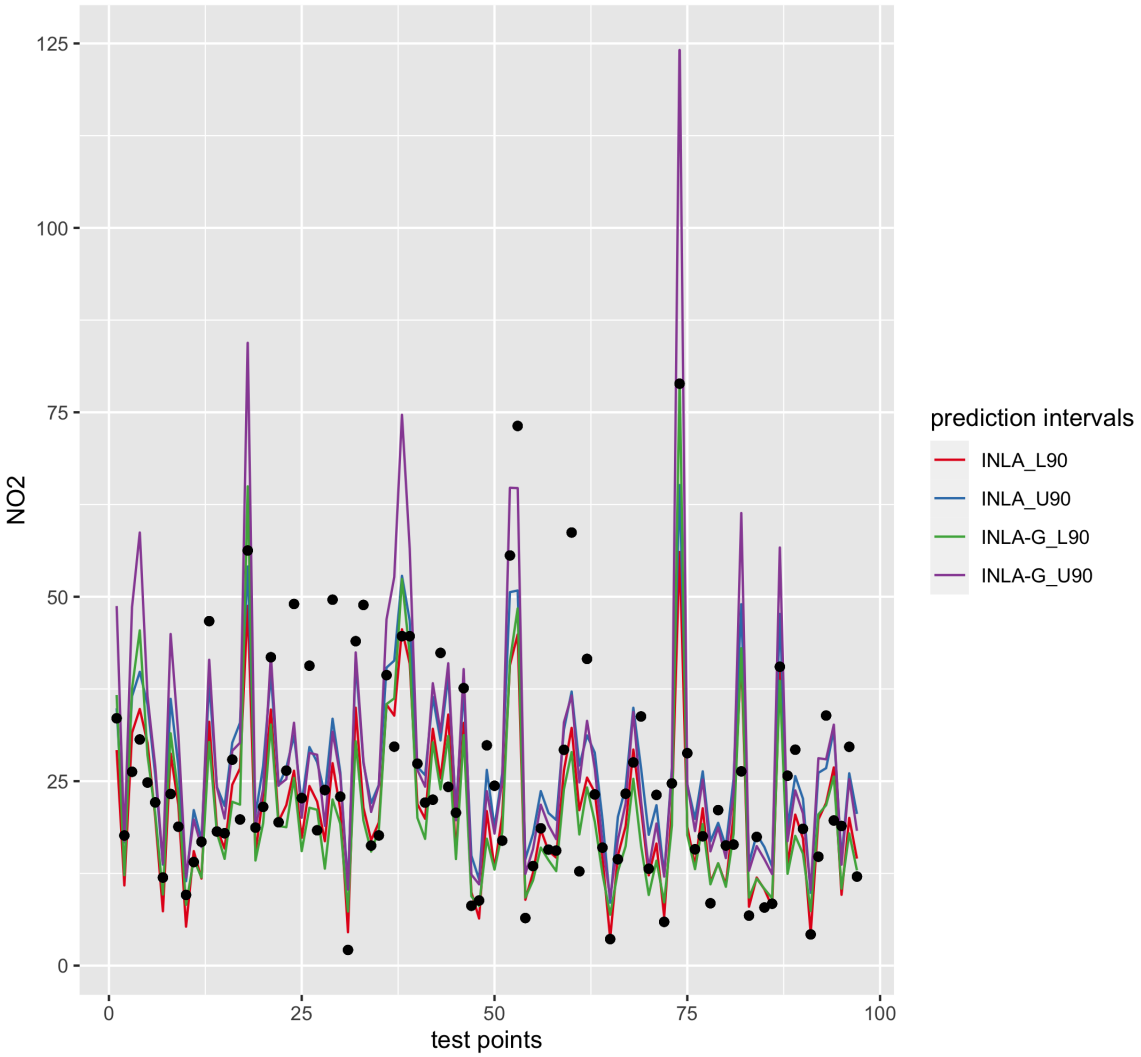


Figure 5: The 90% prediction interval predicted by INLA and INLA-G. The black dots indicate observations in the test dataset.

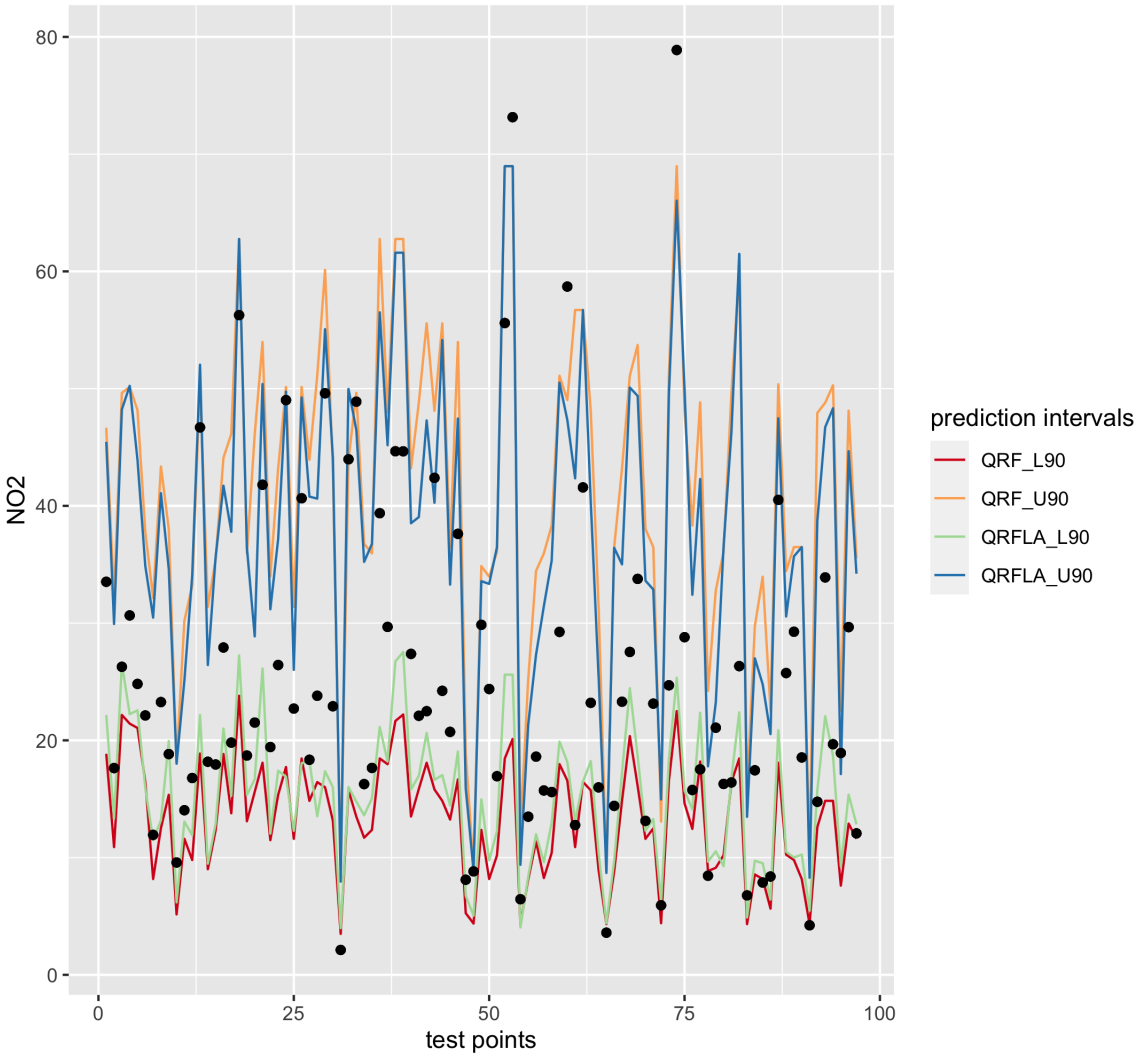


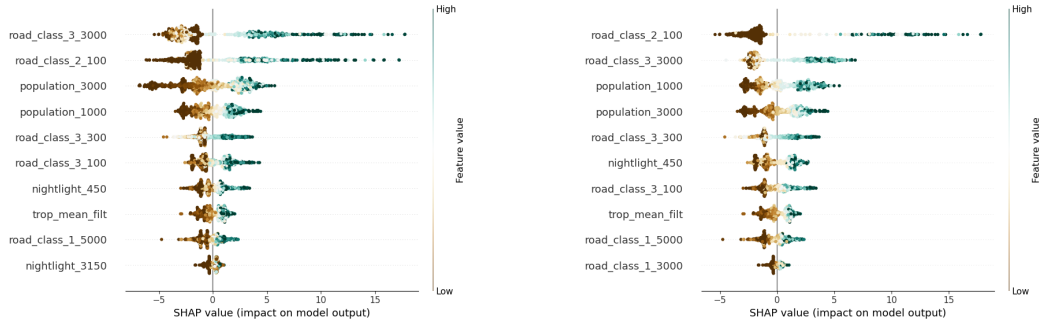
Figure 6: The 90% prediction interval predicted by QRF and QRFLA. The black dots indicate observations in the test dataset.

5.3 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. 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 analyse the effect of each covariate in the INLA model, we firstly normalised all the covariates (by subtracting the mean and dividing the centred 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 the supplementary material figure 3.

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 close to the Stuttgart region. Relatively high values can be observed in northern, southern



(a) (b)

Figure 7: Variable impact calculated by SHAP (SHapley Additive exPlanations), a) the RF model, b) The XGB model. The horizontal location shows whether the effect of that value is associated with a higher or lower prediction. The covariate ranking is based on the sum of SHAP magnitudes over all the samples.

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, the differences between the observations and predictions are relatively large in magnitudes in this region. To facilitate visualisation, we also calculated the differences between INLA model predictions and the observations (supplementary material, figure 2).

6 Discussion

In this study, we compared geostatistical methods with ML methods for spatial NO_2 prediction in Germany and the Netherlands. The comparison consists of the predicted mean, prediction intervals, and model interpretation. Spatial and non-spatial CV strategies are used to reveal prediction accuracy in different aspects. We also implemented the Lasso post-processed RF and geostatistical stacked learning for NO_2 mapping (which to our knowledge have not been applied in air pollution

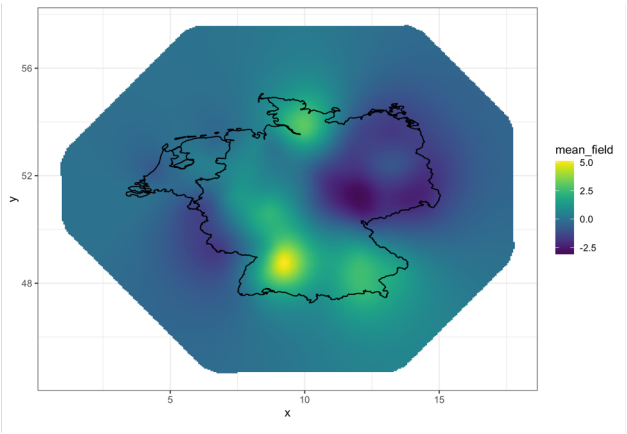


Figure 8: Mean of the spatial random field fitted by the INLA model.

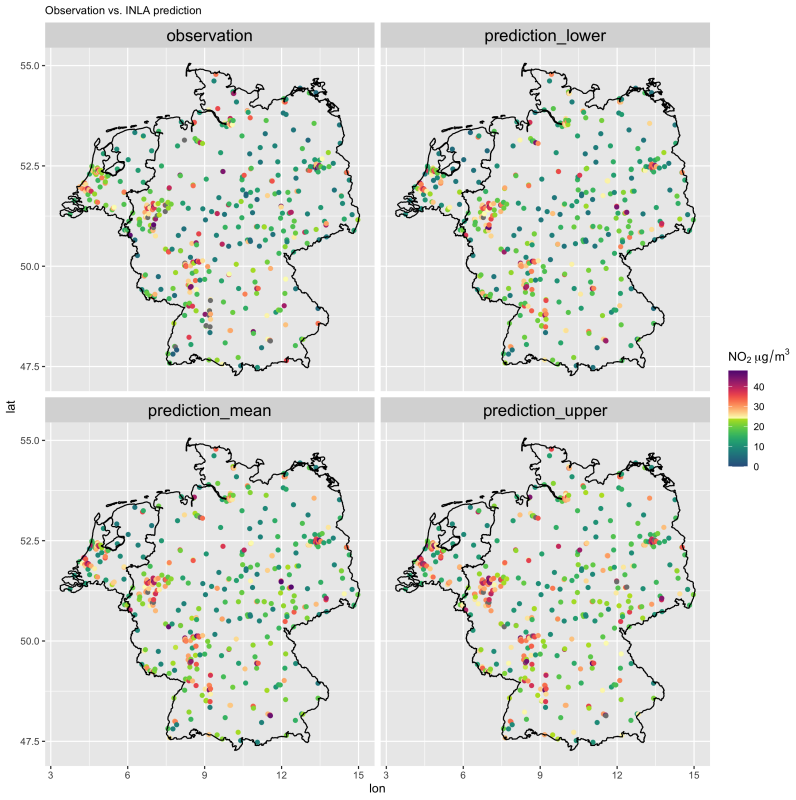


Figure 9: INLA predicted NO_2 at the ground stations with mean (prediction_mean), high (prediction_high, 0.975) and low (prediction_low, 0.925) quantiles and the observed NO_2 (observation).

mapping before) and these two methods considerably improve from the original RF and stacked learning methods, respectively.

Several venues were attempted to further improve the geostatistical 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 the deviation from assumed distribution (Gaussian) is the cause of narrow prediction intervals of the INLA model. However, assuming a Gamma distribution 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 additionally added two factor variables, namely "country code" (country code, "DE" for Germany and "NL" for the Netherlands) and "urban types" (rural, urban, city centre according to (Dijkstra and Poelman, 2014)). However, that also does not increase the model performance. In future works, using a different spatial model (e.g. by specifying different hyperparameters), using the country and urban types as mixed-effects, and modelling spatial varying coefficients may improve the modelling 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).

We implemented an INLA model without modelling the spatial random effect (called non-spatial INLA) to deepen our understanding of the effect of modelling the spatial process in our INLA model. The non-spatial INLA model obtained lower DIC (Information Criterion) 3286.66 vs. 3251.97 (with spatial effects) and WAIC (Watanabe-Akaike information criterion) 3291.75 vs. 3253.93 (with spatial effects). These suggest the advantage of modelling the spatial effects. We normalised covariates before inputting into the spatial and non-spatial INLA models and compared the differences

between the fixed-effects obtained by the original and non-spatial INLA model (supplementary material figure 3-4) and found the most notable change is on the increased effect on the covariate population_1000 for the non-spatial INLA model. This can be explained by that part of the effects of population_1000 is modelled in the spatial random field. The second most notable change is on the decreased effect of nightlight_450 for the non-spatial INLA model. After the spatial process is modelled, 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 non-spatial INLA model, these may indicate that the spatial model could better account for traffic-related variables (i.e. road and nightlight in smaller buffers).

Model performance differs between the three road and population situations. The "far" situation obtained the best modelling 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 modelling the NO_2 emission sources.

Different from non-parametric models such as ensemble trees, a parametric geostatistical 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 (Lu et al., 2020b; Larkin et al., 2017). 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 (i.e. use the same training data for variable selection and test). However, variable selection in each run added in 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 non-spatial stacking. This suggests the complex response-covariate relationships modelled by the ML learners do not fully capture the spatial process. The geostatistical 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 analyse 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 characteristics of the base learners (here: if the base learners are global or local models).

7 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 non-spatial CV may miss reflecting model behaviours. With the model comparison process, we compared the use of geostatistical and ML methods for the spatial prediction of NO_2 in Germany and the Netherlands and found noticeable differences in their limitations and strength. The geostatistical models are preferred especially for urban area prediction and provide the spatial process of observations and indicate the insufficient modelling of spatial random-effects of fixed-effects. But the uncertainty assessment of geostatistical methods, which is commonly known as strength, fails to provide a prediction interval that meets the expectation. The QRF and DF obtained satisfying prediction intervals, with the DF slightly more capable of predicting the extremes. Using Lasso to aggregate trees in random forest increase model performance and reduce model variance. Using the geostatistical method 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 geostatistical model and Gamma objective in the XGBoost obtained much worse results than using a Gaussian likelihood or objective. By comparing with the non-spatial stacking, geostatistical stacking suggests the necessity of modelling the spatial process.

References

C. Alakus, D. Larocque, and A. Labbe. Rfpredinterval: An r package for prediction intervals with random forests and boosted forests. *arXiv preprint arXiv:2106.08217*, 2021.

L. Anselin et al. Spatial econometrics. *A companion to theoretical econometrics*, 310330, 2001.

A. Beloconi and P. Vounatsou. Bayesian geostatistical modelling of high-resolution no2 exposure in europe combining data from monitors, satellites and chemical transport models. *Environment International*, 138:105578, 2020. ISSN 0160-4120. doi: https://doi.org/10.1016/j.envint.2020.105578. URL `https://www.sciencedirect.com/science/article/pii/S0160412019324109`.

S. Bertazzon, M. Johnson, K. Eccles, and G. G. Kaplan. Accounting for spatial effects in land use regression for urban air pollution modeling. *Spatial and Spatio-temporal Epidemiology*, 14-15:9 – 21, 2015. ISSN 1877-5845.

S. Bhatt, E. Cameron, S. R. Flaxman, D. J. Weiss, D. L. Smith, and P. W. Gething. Improved prediction accuracy for disease risk mapping using gaussian process stacked generalization. *Journal of the Royal Society Interface*, 14(134):20170520, 2017.

M. Blangiardo and M. Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

D. J. Briggs, C. de Hoogh, J. Gulliver, J. Wills, P. Elliott, S. Kingham, and K. Smallbone. 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):151–167, 2000.

J. Chen, K. de Hoogh, J. Gulliver, B. Hoffmann, O. Hertel, M. Ketzel, M. Bauwelinck, A. van Donkelaar, U. A. Hvidtfeldt, K. Katsouyanni, et al. A comparison of linear regression, regular-

ization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. *Environment international*, 130:104934, 2019a.

T. Chen and C. Guestrin. xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

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. *xgboost: Extreme Gradient Boosting*, 2019b. URL `https://CRAN.R-project.org/package=xgboost`. R package version 0.82.1.

M. Chiusolo, E. Cadum, M. Stafoggia, C. Galassi, G. Berti, A. Faustini, L. Bisanti, M. A. Vigotti, M. P. Dessì, A. Cernigliaro, et al. Short-term effects of nitrogen dioxide on mortality and susceptibility factors in 10 italian cities: the epiair study. *Environmental health perspectives*, 119(9): 1233–1238, 2011.

Copernicus. Sentinel-5p nrti no2: Near real-time nitrogen dioxide. `https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_NRTI_L3_NO2#bands`, 2021. last accessed: Aug 3, 2021.

D. P. Dee, S. M. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, d. P. Bauer, et al. The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society*, 137(656):553–597, 2011.

P. J. Diggle, P. Moraga, B. Rowlingson, and B. M. Taylor. Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.

L. Dijkstra and H. Poelman. *A harmonised definition of cities and rural areas: the new degree of*

urbanisation, 2014. URL https://ec.europa.eu/regional_policy/sources/docgener/work/2014_01_new_urban.pdf. Last accessed: Aug 4, 2021.

T. Duan, A. Anand, D. Y. Ding, K. K. Thai, S. Basu, A. Ng, and A. Schuler. Ngboost: Natural gradient boosting for probabilistic prediction. In *International Conference on Machine Learning*, pages 2690–2700. PMLR, 2020.

Earthdata. *GES DISC*. URL "https://disc.gsfc.nasa.gov/datasets/OMNO2d_003/summary?keywords=OMI%202017%20No2". last assessed May 21, 2019.

EEA. *Explore air pollution data*, 2021. URL https://www.eea.europa.eu/themes/air/explore-air-pollution-data.

F. Fouedjio and J. Klump. Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. *Environmental Earth Sciences*, 78(1):38, 2019.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

E. Gonzalez-Estrada and J. A. Villasenor-Alva. *goft: Tests of Fit for some Probability Distributions*, 2020. URL https://CRAN.R-project.org/package=goft. R package version 1.3.6.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction, second edition*. Springer Science & Business Media, 2017.

G. Hoek, R. Beelen, K. De Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric environment*, 42(33):7561–7578, 2008.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

A. Jordan, F. Krüger, and S. Lerch. Evaluating probabilistic forecasts with scoringrules. *arXiv preprint arXiv:1709.04743*, 2017.

J. Kerckhoffs, G. Hoek, L. Portengen, B. Brunekreef, and R. C. Vermeulen. Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. *Environmental science & technology*, 53(3):1413–1421, 2019.

E. T. Krainski, V. Gómez-Rubio, H. Bakka, A. Lenzi, D. Castro-Camilo, D. Simpson, F. Lindgren, and H. Rue. *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. CRC Press, 2018.

A. Larkin, J. A. Geddes, R. V. Martin, Q. Xiao, Y. Liu, J. D. Marshall, M. Brauer, and P. Hystad. Global land use regression model for nitrogen dioxide air pollution. *Environmental Science & Technology*, 51(12):6957–6964, 2017.

J. J. Li, A. Jutzeler, B. Faltings, S. Winter, and C. Rizos. Estimating urban ultrafine particle distributions with gaussian process models. *Research@ Locate14*, pages 145–153, 2014.

F. Lindgren, H. Rue, and J. Lindström. 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):423–498, 2011.

F. Lindgren, H. Rue, et al. Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, 63(19):1–25, 2015.

Y. Liu, G. Cao, and N. Zhao. Integrate machine learning and geostatistics for high-resolution

mapping of ground-level pm2. 5 concentrations. In *Spatiotemporal Analysis of Air Pollution and Its Application in Public Health*, pages 135–151. Elsevier, 2020.

M. Lu, O. Schmitz, K. de Hoogh, Q. Kai, and D. Karssenberg. Evaluation of different methods and data sources to optimise modelling of no2 at a global scale. *Environment international*, 142: 105856, September 2020a. ISSN 1873-6750. doi: 10.1016/j.envint.2020.105856.

M. Lu, I. Soenario, M. Helbich, O. Schmitz, G. Hoek, M. van der Molen, and D. Karssenberg. Land use regression models revealing spatiotemporal co-variation in no2, no, and o3 in the netherlands. *Atmospheric Environment*, 223:117238, 2020b.

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

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf`.

S. M. Lundberg, 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. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749, 2018.

K. Luo, R. Li, W. Li, Z. Wang, X. Ma, R. Zhang, X. Fang, Z. Wu, Y. Cao, and Q. Xu. Acute effects of nitrogen dioxide on cardiovascular mortality in beijing: an exploration of spatial heterogeneity and the district-specific predictors. *Scientific reports*, 6(1):1–13, 2016.

S. Martino and H. Rue. Implementing approximate bayesian inference using integrated nested laplace approximation: A manual for the inla program. *Department of Mathematical Sciences, NTNU, Norway*, 2009.

T. G. Martins, D. Simpson, F. Lindgren, and H. Rue. Bayesian computing with inla: new features. *Computational Statistics & Data Analysis*, 67:68–83, 2013.

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

P. Moraga. *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*. Chapman & Hall/CRC, 2019.

P. Moraga, S. M. Cramb, K. L. Mengersen, and M. Pagano. A geostatistical model for combined analysis of point-level and area-level data using inla and spde. *Spatial Statistics*, 21:27–41, 2017.

NASA. *Shuttle Radar Topography Mission*. URL `https://www2.jpl.nasa.gov/srtm/dataprelimdescriptions.html`. last assessed Aug 15, 2021.

D. A. Nelson. European environment agency. *Colo. J. Int'l Envtl. L. & Pol'y*, 10:153, 1999.

NOAA. Dmsp and viirs data download. `"https://ngdc.noaa.gov/eog/download.html"`, 2021. Last Accessed: 11.03.2021.

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

X. Ren, Z. Mi, and P. G. Georgopoulos. 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, 2020. ISSN 0160-4120. doi: https://doi.org/10.1016/j.envint.2020.105827. URL `https://www.sciencedirect.com/science/article/pii/S0160412020317827`.

H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005.

H. Rue, S. Martino, and N. Chopin. 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):319–392, 2009.

Y. Rybarczyk and R. Zalakeviciute. Machine learning approaches for outdoor air quality modelling: A systematic review. *Applied Sciences*, 8(12):2570, 2018.

L. Schlosser, T. Hothorn, R. Stauffer, A. Zeileis, et al. Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *The Annals of Applied Statistics*, 13(3): 1564–1589, 2019.

G. Shaddick, M. L. Thomas, H. Amini, D. Broday, A. Cohen, J. Frostad, A. Green, S. Gumy, Y. Liu, R. V. Martin, et al. Data integration for the assessment of population exposure to ambient air pollution for global burden of disease assessment. *Environmental science & technology*, 52(16): 9069–9078, 2018.

D. M. Stasinopoulos, R. A. Rigby, et al. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7):1–46, 2007.

M. L. Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

J. Velthoen, C. Dombry, J.-J. Cai, and S. Engelke. Gradient boosting for extreme quantile regression. *arXiv preprint arXiv:2103.00808*, 2021.

A. M. Vicedo-Cabrera, A. Biggeri, L. Grisotto, F. Barbone, and D. Catelan. 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, 2013.

J. A. Villaseñor and E. González-Estrada. A variance ratio test of fit for gamma distributions. *Statistics & Probability Letters*, 96:281–286, 2015.

S. Wager, T. Hastie, and B. Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651, 2014.

Q. Wang, H. Feng, H. Feng, Y. Yu, J. Li, and E. Ning. The impacts of road traffic on urban air quality in jinan based gwr and remote sensing. *Scientific Reports*, 11(1):1–9, 2021.

M. T. Young, M. J. Bechle, P. D. Sampson, A. A. Szpiro, J. D. Marshall, L. Sheppard, and J. D. Kaufman. Satellite-based no2 and model validation in a national prediction model based on universal kriging and land-use regression. *Environmental science & technology*, 50(7):3686–3694, 2016.

C. Yuan. Models and methods for computationally efficient analysis of large spatial and spatio-temporal data. 2011.

L. Zhai, S. Li, B. Zou, H. Sang, X. Fang, and S. Xu. An improved geographically weighted regression model for pm2. 5 concentration estimation in large areas. *Atmospheric Environment*, 181:145–154, 2018.

Y. Zhan, Y. Luo, X. Deng, K. Zhang, M. Zhang, M. L. Grieneisen, and B. Di. Satellite-based estimates of daily NO_2 exposure in China using hybrid random forest and spatiotemporal kriging model. *Environmental science & technology*, 52(7):4180–4189, 2018.

B. Zou, Q. Pu, M. Bilal, Q. Weng, L. Zhai, and J. E. Nichol. High-resolution satellite mapping of fine particulates based on geographically weighted regression. *IEEE Geoscience and Remote Sensing Letters*, 13(4):495–499, 2016.