

¹ A comparison of geostatistical and non-spatial machine
² learning methods in NO_2 modelling: prediction accuracy,
³ uncertainty quantification, and model interpretation

⁴ Meng Lu¹, Joaquin Cavieres², and Paula Moraga³

⁵ ¹Department of Geography, University of Bayreuth, Universitaetsstrasse 30, 95447
⁶ Bayreuth, Germany

⁷ ²Instituto de Estadística, Facultad de Ciencias, Universidad de Valparaíso,
⁸ Valparaíso, Chile

⁹ ³Computer, Electrical and Mathematical Sciences and Engineering Division, King
¹⁰ Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900,
¹¹ Saudi Arabia

¹² **Abstract**

¹³ NO_2 is a traffic-related air pollutant that is strongly associated with cardiovascular and
¹⁴ respiratory diseases. Ground NO_2 monitoring stations measure NO_2 concentrations at certain
¹⁵ locations and statistical predictive methods have been developed to predict NO_2 as a continuous
¹⁶ surface to inform decision-making. Among them, machine learning methods are the most power-
¹⁷ ful in capturing non-linear relationships between NO_2 measurements and geospatial predictors,

18 but it is unclear if the spatial structure of NO₂ is also captured in the response-covariates rela-
19 tionships. In addition, most model comparison studies only compare accuracy in the prediction
20 mean at ground stations, but do not consider prediction intervals and model interpretation and
21 the effects of different model evaluation methods. In this study, we dive into the comparison
22 between spatial and non-spatial data models accounting for the above-mentioned aspects. More-
23 over, we implemented a spatial and a non-spatial methods that have not been applied to air
24 pollution mapping before and evaluated stack learning methods with and without modelling
25 the spatial process. We implemented our study using national ground station measurements of
26 NO₂ in Germany and Netherlands of the year 2017, predicting NO₂ to 100 m resolution grid.
27 Our results indicate the importance of modelling the spatial process especially in areas close to
28 traffic. The geostatistical stack learning method obtained the best results in terms of root mean
29 squared error (6.83, 24.5% of the mean of observations) and R-squared (0.71). The prediction
30 intervals predicted with ensemble tree-based methods are satisfactory but too narrow with the
31 geostatistical methods. Compared to ensemble tree-based methods, the geostatistical methods
32 provide important spatial information for analysing emission sources and the spatial process of
33 observations.

34 1 Introduction

35 NO₂ is a traffic-related air pollutant and has been found in epidemiological time series analysis
36 to highly associated with respiratory (Luo et al., 2016) and cardiovascular (Chiusolo et al., 2011)
37 diseases. NO₂ values are measured using monitoring stations at certain locations (e.g. close to traffic)
38 and most of the epidemiological studies identified the relationships between NO₂ and diseases or
39 hospital admission using a single NO₂ monitoring station to represent the entire district. However,
40 NO₂ is highly dynamic over the district and the difference in NO₂ concentrations will reflect on
41 personal exposures to NO₂. Detailed spatial mapping of NO₂ is therefore required for more accurate

42 quantification of the relationships between NO₂ and health effects. In addition, detailed NO₂ maps
43 are necessary for scientific recommendations to be provided to policymakers and city planners.

44 Statistical methods for NO₂ mapping have attracted a lot of attention with the burgeoning Ma-
45 chine Learning (ML) methods and availability of ground monitoring station networks, atmospheric
46 satellite products, and geospatial predictors. Geospatial predictors are variables that are included
47 as covariates in a statistical air pollution model. Commonly used geospatial predictors are air
48 emission- (e.g. road networks) and dispersion-related (e.g. wind speed) variables, numerical mod-
49 elling (e.g. with chemistry transport model) output, and atmospheric remote sensing measurements
50 or products. A most recent (data available from Jan-2018) atmosphere sensing instrument, Tropomi
51 (Tropospheric monitoring instrument, NSO and ESA, 2019) onboard of Sentinel 5p satellite, mea-
52 sures column density of a variety of gaseous air pollutants, in particular with an unprecedentedly
53 high resolution for NO₂ (3.5 km x 5.5 km, across along track, since 06 August 2019).

54 Statistical methods applied for spatial air pollution prediction can be broadly classified depending
55 on whether the spatial dependency is explicitly modelled. If not modelled, we refer to the methods
56 "non-spatial" and otherwise "spatial". Most of the spatial air pollution models were developed to
57 predict at coarser resolutions, commonly 1 km or coarser (Young et al., 2016; Shaddick et al., 2018;
58 Beloconi and Vounatsou, 2020). Non-spatial methods are more dominant in air pollution mapping,
59 particularly in high-resolution (100 m resolution or higher) mapping. Among them, LUR (Land
60 Use Regression) models which assumes linear relationships between NO₂ and geospatial predictors
61 are the most studied (Briggs et al., 2000; Hoek et al., 2008). Most recently, statistical learning (in
62 this study, "statistical learning" is used interchangeably with "machine learning") methods (Hastie
63 et al., 2009), including regularised linear regression (e.g. Lasso and Ridge regression (James et al.,
64 2013)), kernel methods such as support vector machine (Suykens and Vandewalle, 1999), ensemble
65 tree-based methods such as random forest (RF, Breiman, 2001) and XGBoost (XGB, Chen and

⁶⁶ Guestrin, 2016), have been applied for feature selection or capturing non-linear response-covariate
⁶⁷ relationships (Lu et al., 2020a; Chen et al., 2019a). In air pollution (not restricted to NO₂) mapping,
⁶⁸ several studies compared between statistical learning and conventional LUR methods (Chen et al.,
⁶⁹ 2019a; Kerckhoffs et al., 2019; Lu et al., 2020a; Ren et al., 2020; Rybarczyk and Zalakeviciute, 2018).

⁷⁰ Geostatistical models (e.g. Kriging) and Geographically Weighted Regression (GWR) are the
⁷¹ most used spatial methods for air pollution prediction (Vicedo-Cabrera et al., 2013; Li et al., 2014;
⁷² Wang et al., 2021; Zou et al., 2016) and these methods have been combined with dimension reduction
⁷³ Zhai et al. (2018) and RF (Zhan et al., 2018; Liu et al., 2020) to improve NO₂ prediction accuracy.

⁷⁴ A Bayesian geostatistical model is developed in Beloconi and Vounatsou (2020) to predict NO₂ by
⁷⁵ integrating Tropomi satellite instrument measurements and chemical transport models. A GWR
⁷⁶ model naturally models spatial varying coefficients by fitting multiple local regressions depending
⁷⁷ on the homogeneity in response-covariate relationships when a number of observations are involved.

⁷⁸ A typical geostatistical model can be viewed as consisting of two components: a mean function,
⁷⁹ commonly a linear model, capturing the response-covariate relationships and a covariance function
⁸⁰ modelling dependency of residuals from the mean (Bhatt et al., 2017). Conventional Kriging methods
⁸¹ suffer from the "big n problem", i.e. it may become computationally intractable with a large number
⁸² of observations. To deal with this problem, Lindgren et al. (2011) propose to use Stochastic Partial
⁸³ Differential Equations (SPDE) to approximate the Gaussian Random Field (GRF) to a Gaussian
⁸⁴ Markov Random Field (GMRF, Rue and Held (2005)). The main advantage of this is that the GMRF
⁸⁵ has a sparse structure of the precision matrix, which is the inverse of the covariance matrix of a
⁸⁶ GRF. Along with this, Rue et al. (2009) propose to use the Integrated Nested Laplace Approximation
⁸⁷ (INLA) in a Bayesian framework to achieve the computational scalability of a geostatistical model
⁸⁸ using approximations for all the estimations. This is especially advantageous when modelling NO₂
⁸⁹ over a larger scale e.g., continental or global-scale modelling when a large amount of observations

90 are modelled, and in spatiotemporal modelling.

91 As spatial models are typically more complex compared to their non-spatial counterparts, several
92 studies compared spatial and non-spatial models to understand if the spatial effects could be simply
93 modelled by including certain covariates in LUR models. Young et al. (2016) studied the use of
94 universal Kriging (UK), OMI (Ozone Monitoring Instrument) satellite instrument (Earthdata) and
95 LUR models for NO₂ prediction at 2.5 km resolution. Young et al. (2016) indicated that either
96 using UK or adding OMI in the LUR model improves a LUR model but adding OMI in a UK
97 model only trivially improves the performance. Bertazzon et al. (2015) shows that the inclusion of
98 the meteorological variables accounts for spatial effects similarly to the use of spatial autoregressive
99 models(Anselin et al., 2001). However, even if the spatial dependency can be captured by involving
100 certain covariates in a LUR model, we may still need geostatistical methods to understand the
101 spatial structure present in the data. Linear models have been used for the mean function but the
102 relationships between NO₂ and predictors have been shown to be better modelled with non-linear
103 ML methods (Lu et al., 2020a). Most recent studies attempt to replace the linear mean function
104 with ML models. Liu et al. (2020) applied a geostatistical model to the residuals from an RF model
105 for the spatial prediction of PM_{2.5}. In disease mapping, Bhatt et al. (2017) proposes to stack ML
106 models to replace the mean function in a geostatistical model.

107 Few studies have compared between geostatistical and ML methods, possibly because the ML
108 methods are still relatively less studied in air pollution mapping and in the field of geostatistics. It
109 might be more interesting to compare between geostatistical methods and ML methods than geosta-
110 tistical methods and LUR, because ML methods may be more capable of (implicitly) capturing the
111 spatial dependency by integrating covariates, when the number of observations is sufficient. More-
112 over, most comparison studies only compare the cross-validation accuracy of the prediction mean
113 (e.g. using R-squared, mean absolute error, or root mean squared error), ignoring the prediction

114 intervals. Also not discussed is the cause of the prediction errors, are they caused by missing co-
115 variants, violation of the model assumptions (e.g. data distribution, non-linearity), or inconsistent
116 distributions between training and validation sets. Also, different cross-validation strategies, e.g.,
117 how do we split the train-test sets, may lead to different model validation results. Current studies
118 typically solely rely on k-fold splitting (Kerckhoffs et al., 2019; Larkin et al., 2017; Ren et al., 2020)
119 or bootstrapping (Lu et al., 2020a) to randomly splitting between train-test sets, which may be
120 one-sided and does not provide an indication of accuracy in spatial blocks (but only at the locations
121 of ground stations).

122 In this study, we focus on ensemble tree-based methods (e.g. RF and boosting) in the ML
123 category and a hierarchical spatial model (Lindgren et al., 2015; Blangiardo and Cameletti, 2015;
124 Moraga, 2019) called latent Gaussian model in the geostatistics category. Additionally, we invest in
125 stacked models in integrating ML and geostatistical models and develop a LUR model using Lasso for
126 comparison. Ensemble trees are nonparametric models, deriving prediction intervals is therefore less
127 straightforward than a parametric model (e.g. a linear regression model) but has been studied and
128 shown satisfactory results with simulated data. Prediction intervals have been most well studied for
129 RF (Meinshausen, 2006; Wager et al., 2014; Stasinopoulos et al., 2007; Alakus et al., 2021) and more
130 recently for boosting (Duan et al., 2020; Velthoen et al., 2021). Comparing probabilistic methods
131 (i.e. prediction interval calculation) of RF and boosting is beyond the scope of this study and
132 we focus on prediction intervals derived for RF to compare with geostatistical methods. Possibly,
133 one of the most widely recognisable methods to derive RF prediction intervals is Quantile Random
134 Forest (QRF) (Meinshausen, 2006). QRF has been shown to estimate middle quantiles well but may
135 fall short at the extremes due to the limited number of observations in the tail regions (Velthoen
136 et al., 2021). Velthoen et al. (2021) proposed to use extreme quantile regression to estimate for
137 data outside the range of observations. Another well-recognised method is distributional regression

¹³⁸ forests (DF) (Schlosser et al., 2019), which embeds the GAMLSS (Generalised Additive Models for
¹³⁹ Location Scale and Shape) (Stasinopoulos et al., 2007) into RF.

¹⁴⁰ Fouedjio and Klump (2019) compared prediction accuracy and uncertainty quantification be-
¹⁴¹ tween KED (Kriging with external drift) and QRF by simulating data with various levels of spatial
¹⁴² dependency. It concluded that an optimal model choice depends on the level of spatial dependency
¹⁴³ and response-covariate relationships. However, it does not account for the fact that in practice, as
¹⁴⁴ an ensemble tree-based method can make use of a large number of (possibly correlated) predictors
¹⁴⁵ without being constrained to certain (e.g. linear) relationships, the spatial dependency may be
¹⁴⁶ explained by the covariates despite not being explicitly modelled.

¹⁴⁷ The objective of our study is to compare geostatistics and non-spatial ensemble tree-based models
¹⁴⁸ for NO₂ mapping, in terms of their prediction accuracy, uncertainty quantification, and model inter-
¹⁴⁹ pretation and to understand effect of modelling spatial structures. More specifically, the following
¹⁵⁰ sub-objectives are reached:

¹⁵¹ 1. Optimising a set of spatial Hierarchical and ML models for NO₂ prediction in Germany and
¹⁵² the Netherlands.

¹⁵³ 2. Developing a non-spatial and a geostatistical stacked ensemble model, i.e., a stack of various
¹⁵⁴ ML learners.

¹⁵⁵ 3. Model comparison regarding the predicted mean, prediction interval, and model interpretation.

¹⁵⁶ The spatial Hierarchical model incorporates the spatial random effect along with other covariates
¹⁵⁷ and the estimation is performed using the R package INLA (Rue et al., 2009; Martins et al., 2013).
¹⁵⁸ XGB, RF and Lasso are chosen for the comparison with the geostatistical model and they also
¹⁵⁹ form the base learners in the two (geostatistical and non-spatial) stacked learning models. The ML
¹⁶⁰ methods are chosen for their dissimilarity. Specifically, Lasso is a linear regression model without

₁₆₁ accounting for spatial dependency. RF and XGB are non-linear models with regression trees as base-
₁₆₂ learners and are not affected by dependent covariates. XGB is a highly scalable boosting method
₁₆₃ that builds tree models subsequently over the residuals of previous trees and has multiple routines
₁₆₄ to penalise model over-fitting (Chen et al., 2019b), which has been reported in various studies to
₁₆₅ obtain the highest prediction accuracy Lu et al. (2020a).

₁₆₆ 2 Data

₁₆₇ NO₂ concentration measurements of 2017 from national ground stations of Germany and the Nether-
₁₆₈ lands are used. The original hourly data is downloaded from the EEA (European Environment
₁₆₉ Agency, Nelson, 1999; EEA, 2021). Negative values are considered as missing. The data is ag-
₁₇₀ gregated to annual concentrations by taking the mean and omitting missing values. The spatial
₁₇₁ distribution of NO₂ stations and the station types, histogram and Q-Q plot for normality are shown
₁₇₂ in fig. 1. We conducted a Shapiro test for normality, with the result implying the distribution of data
₁₇₃ being significant different from normal distribution (p-value= 8.605e-12, "normal distribution" and
₁₇₄ "Gaussian distribution" are used interchangeably in this study). A Gamma distribution test was
₁₇₅ conducted using the method proposed in Villaseñor and González-Estrada (2015) and implemented
₁₇₆ in Gonzalez-Estrada and Villasenor-Alva (2020). The test result (p-value = 0.32) indicates that the
₁₇₇ data distribution is not significantly different from Gamma distribution.

₁₇₈ The geospatial predictor grids (table 1) are calculated or re-sampled at 100 m resolution. They
₁₇₉ are either spatial attributes aggregated in a circular ring centred at each sensor or prediction location,
₁₈₀ called buffered predictors, or values of the spatial attribute at the observation or prediction location,
₁₈₁ called gridded variables. The buffered predictors include total road length, total industry areas, VI-
₁₈₂ IRS (Visible Infrared Imaging Radiometer Suite) Nighttime Day/Night Band (DNB) radiance values
₁₈₃ (nightlight, NOAA, 2021) and population. Variables that are originally grids include wind speed

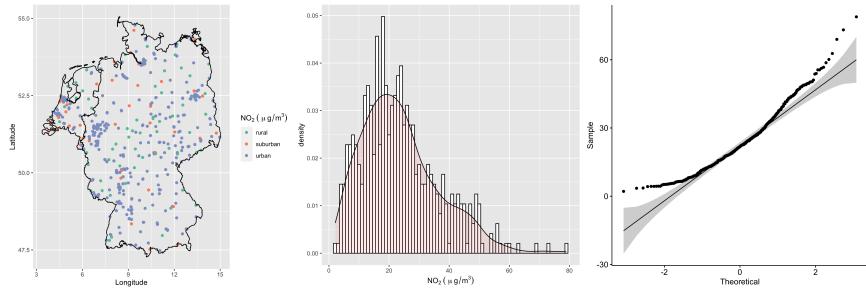


Figure 1: Spatial distribution of NO₂ stations, histogram and Q-Q plot of the NO₂ measurements.

184 and temperature (Dee et al., 2011), elevation (NASA), monthly TROPOMI level 3 product of NO₂
 185 column density (Copernicus, 2021) from 2019 (due to the increased resolution compared to 2018).
 186 The buffered predictors of road and industry are calculated from OpenStreetMap (OpenStreetMap
 187 contributors, 2019). For detailed descriptions of the processing of the geospatial predictors please
 188 refer to Lu et al. (2020a).

189 **3 Methods**

190 The methods considered in this study are classified as spatial and non-spatial and are abbreviated
 191 below.

192 **Spatial models:**

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

197 **Non-spatial models:**

- 198 1. LA: A Lasso regression model;

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

¹⁹⁹ 2. RF: A RF model;

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

206 **3.1 Non-spatial methods**

207 Lasso is a linear regression algorithm with the L1 regularisation to shrink variable coefficients to
208 zero, which enables "feature selection". In the cost function, the absolute value of coefficient is added
209 to the original least squares as a penalty term. RF and XGB in this study use trees as base learners
210 and ensemble them to reduce variability of single trees (Friedman, 2001). RF firstly randomly draws
211 a subset of features, and then choose features from this subset to build the tree. RF (Breiman, 2001)
212 grows trees independently and then take the mean of the predictions of each tree.

213 QRF is a non-parametric prediction interval estimation method which keeps all the observations
214 in the terminal node for estimating the conditional probability function. Specifically, it samples
215 from all the response values in each terminal node and use the ratio between the number of samples
216 that is taken from each terminal node and the number of total observations in the terminal node as
217 weights to aggregate the samples. The weights of all the trees are summed. The summed weights
218 computed for each observation are then used to construct the empirical conditional cumulative
219 distribution function (Meinshausen, 2006). QRFLA uses Lasso as a post-processing of QRF (Hastie
220 et al., 2017, page 617). This method firstly preserves all the trees instead of aggregating them

221 (e.g. taking the mean of all the predictions) and then apply Lasso regression to all the trees for
222 aggregation. This leads to a shrinkage of the tree space and theoretically reduces model variance.
223 DF (Schlosser et al., 2019) firstly divide data into regions as homogeneous as possible with respect
224 to a parametric distribution, thus capturing changes in location, scale and shapes. For each tree,
225 maximum likelihood is used to fit distributions and recursively select and split covariates according
226 to the instability of the gradient of the likelihood at each observation along each co-variate. Then,
227 the distributional trees are ensembled for DF.

228 XGB is a variation of gradient boosting, which grows trees subsequently by fitting to model
229 residuals of the previous step. XGB is scalable to multiple threads. It enables multiple penalisation
230 paths to control model complexity to prevent model over-fitting, including regularisation (e.g. L1
231 regularisation) on tree width and terminal node values, as well as drop-out (dropping trees), sampling
232 observations (take a subset of observations in each run), and early stopping (stop iterating when after
233 a few rounds the loss does not decrease or the node does not meet the splitting rule). The default
234 objective function for regression assumes normal distribution of target variables (and the prediction
235 is the mean of the distribution). This assumption is used in all the air pollution mapping studies.
236 Here, we additionally fit a model with the objective function assuming the target variable follows a
237 Gamma distribution (XGB-G) as the distribution of NO₂ measurements is closer to Gamma than
238 normal distribution.

239 Different from the ensembling in RF or XGB,SE (Stacking Ensemble) refers to a class of al-
240 gorithms that trains a second-level “meta-learner” to optimise the combination of a collection of
241 prediction algorithms (base-learners). The base-learners are preferably diverse to capture different
242 relationships or patterns. In this study, Lasso, RF, and XGB are the base-learners. Cross-validated
243 predicted values (commonly known as level-one data) are used to train the meta-learner.

244 **3.2 Hyperparameter setting for XGB and RF**

245 To optimise the hyperparameters of XGB (known as "model tuning"), we used grid search to optimise
246 hyperparameters in 5-fold cross-validation basing on the minimum RMSE (Root Mean Squared
247 Error) and additionally manual adjustment of the hyperparameters to look at the prediction patterns.
248 The grid search is used instead of more computationally efficient methods (e.g. Bayesian or random
249 search) as the optimal hyperparameter range is largely known from our previous experiences (Lu
250 et al., 2020a, 2021). The search grid for the number of iterations (nrounds) was from 200 to 3000,
251 with a step of 200; maximum tree depth (max-depth) from 3 to 6 with a step of 1, learning rate
252 (eta) from 0.001 to 0.1 with a step of 0.05, the penalty term Gamma (Chen et al., 2019b) from 1
253 to 5 with a step of 1, the subsample is set to 0.7, L1 norm penalisation (lambda) is set to 2 and L2
254 norm penalisation (alpha) is set to 0. RF is not sensitive to hyperparameter tuning. We used the
255 default setting of number of variables that are randomly drawn for each tree (Breiman, 2001), which
256 is the integer part of the total number of variables divided by three. The number of trees is set to
257 2000 for a safe choice as the high number of trees will not negatively affect model performance.

258 **3.3 Geostatistical models**

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

$$y_i \sim N(\mu_i, \sigma^2), \quad i = 1, 2, \dots, n \quad (1)$$

$$\mu_i = \mathbf{d}_i \boldsymbol{\beta} + \mathbf{x}(s_i) \quad (2)$$

263 Here, $\mathbf{d}_i = (d_{i1}, \dots, d_{ip})$ is the vector of covariates at location \mathbf{s}_i , $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is the
 264 coefficient vector, and $\mathbf{x}(\mathbf{s}_i)$ denotes a spatial Gaussian random field. That is, $\{\mathbf{x}(\mathbf{s}_1), \dots, \mathbf{x}(\mathbf{s}_n)\} \sim$
 265 $\mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma})$, where N_n is a Normal multivariate distribution for the spatial process specified by its
 266 mean $\mathbb{E}(\mathbf{x}(\mathbf{s}))$, and covariance function $C(\mathbf{s}_1, \mathbf{s}_2) = \text{Cov}(\mathbf{x}(\mathbf{s}_1), \mathbf{x}(\mathbf{s}_2))$. The Gaussian random field
 267 can be stationary and isotropic, where the covariance function depends only on the distance and not
 268 direction between points, that is $C(\mathbf{s}_1, \mathbf{s}_2) = \text{Cov}(\|\mathbf{s}_1 - \mathbf{s}_2\|)$ and its dependence is commonly modeled
 269 using a Matérn function (Stein (2012); Yuan (2011); Diggle et al. (2013)). Since incorporating the
 270 spatial dependence directly with a large number of observations using a Gaussian random field is
 271 computationally expensive, Rue and Held (2005) proposed the approximation of a Gaussian random
 272 field by a Gaussian Markov random field for a more efficient computational process of estimation.
 273 The main property of the Gaussian Markov random field is that it uses a conditional dependency
 274 structure through the precision matrix \mathbf{Q} .

275 In this study, we compare two spatial hierarchical models with geospatial predictors as covariates,
 276 one uses a Gaussian likelihood and the other a Gamma likelihood. The Gamma model has the same
 277 hierarchical structure as the Gaussian model: the response variable in (1) can be represented by
 278 $y_i \sim \text{Gamma}(\alpha, \beta)$ where α is the shape parameter and β the rate parameter. The INLA-SE model
 279 uses a Gaussian likelihood.

280 3.4 INLA and SPDE

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

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

$$\mathbf{x} \mid \theta_2 \sim \text{GRF}(\mathbf{0}, \mathbf{Q}(\boldsymbol{\theta}_2)^{-1}) \quad (4)$$

$$\boldsymbol{\theta} = \{\theta_1, \theta_2\} \quad (5)$$

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

288 The R package INLA can be used to perform direct numerical calculation of the posterior distri-
 289 bution for a Bayesian hierarchical model (Rue et al. (2009), Martino and Rue (2009)). If we use \mathbf{x}
 290 as a latent Gaussian field (a Gaussian Markov random field), $\boldsymbol{\theta}$ a vector of (hyper)parameters and
 291 \mathbf{y} a vector of observations, assuming independent observations given the vector of the spatial latent
 292 field (\mathbf{x}) and the hyperparameters ($\boldsymbol{\theta}$), the likelihood can be expressed as:

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

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

294

295 The main aim is to approximate the posterior density for the posterior of the spatial latent field
 296 and the hyperparameters. The marginal densities can be obtained:

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

297 and

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

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

299

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

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

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

309 3.5 Geospatial predictor selection for the INLA model

310 As involving too many covariates (e.g. more than 12) in the INLA model brings problems in model
311 inferencing and multicollinearity, we used Lasso to reduce the number of variables. The Lasso was
312 used instead of ensemble tree-based methods for feature selection because it is also a linear model
313 (same as the INLA and INLA-G models in our study). Variables are selected with the L1 norm
314 penalty that returns a model with errors that are within one standard error of the minimum mean
315 cross-validated error. Lasso is applied to 80% data randomly sampled from all the observations
316 and this process is repeated 20 times. Variables that are selected more than 90% of the times (i.e.
317 18) will be considered as covariates in INLA. The times that the Lasso selected certain variables is
318 shown in table 2. The INLA modelling process applies the same bootstrapped samples for training

319 and validation. In addition, AIC (step-wise) model selection is applied to the entire dataset to
320 suggest a model as a further reference. The variables selected by AIC are almost the same as
321 Lasso selected variables, besides it does not choose road_class_3_3000, which is highly correlated
322 with road_class_1_5000. Based on this, the road_class_3_3000 is not used as a covariate in INLA.

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

323 **3.6 INLA model parameterisation**

324 The triangulated mesh constructed in the SPDE approach is shown in supplementary material
325 (supfig. 1), with size of the inner and outer extensions around the data locations (*offsets*) 1/8 of
326 the maximum distance among all the observations for both the inner and outer extensions. The
327 maximum allowed triangle edge lengths in the region and in the extension (*max.edge*) are set
328 to respectively 1/30 and 1/5 times maximum distance among all the observations. The Matern
329 SPDE model is constructed with $\alpha = 2$. The SE-INLA model has the same specification (i.e.
330 mesh structure, likelihood, objective function, priors, optimisation process) as the INLA model
331 parameterisation described above.

332 **4 Model evaluation**

333 **4.1 Cross validation**

334 We use RMSE, MAE (Mean Absolute Error), IQR (Interquartile Range) and R² (R-squared) to
335 compare model performance. RMSE is calculated as the square root of the differences between
336 predictions and observations; MAE is calculated as the absolute differences between predictions
337 and observations; IQR is the differences between the third and first quartiles of the prediction. R²
338 indicates the explained variance and is calculated as $R^2 = 1 - \text{var}(\text{error})/\text{var}(y)$, where $\text{var}(\cdot)$
339 indicates variance, error indicates model residuals and y indicates observed response values. When
340 different data is used in CV (e.g. separating between close and far-away from roads), we additionally
341 calculated the RRMSE (relative RMSE), RMAE (relative MAE), RIQR (relative IQR) to account
342 for the differences in the magnitudes of response values. The RRMSE and RMAE are calculated by
343 dividing the RMSE and MAE, respectively, by the mean of observations. The RIQR was calculated
344 by dividing the IQR by the median of observations. The three CV methods we designed and used

345 to assess our model performance are:

- 346 1. Bootstrapped CV. 20-times randomly bootstrapped splitting of training and test sets (Lu et al.,
347 2020a).
- 348 2. Spatial-blocked CV. Dividing data into spatial blocks, each time use one block for test and
349 other blocks for training.
- 350 3. Customised CV. Splitting train-test based on values of certain covariates. In this study, three
351 sub-areas are defined, 1) close to traffic and with high population ("tr-hp"), 2) close to traffic
352 and with middle low population ("tr-lmp"), 3) far away from traffic ("far"). High population is
353 defined as the variable population of 1000 m buffer that is in the last quartile. Low population
354 is defined as the variable population of 1000 m buffer is below the median. Close to road is
355 defined as (please refer to table 1 for the definition of covariates):

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

359 Far away from road is defined as:

```
360     road_class_2_100 == 0 &  
361     road_class_1_100 == 0 &  
362     road_class_3_100 < quantile(road\class\_3\_100, .5)
```

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

365 This yields 85, 65, and 177 samples in each category. This ensures a balanced number of samples
366 between close to traffic and far-away from traffic. Each time, 30 samples (7% of the entire dataset)

367 are drawn from the corresponding category for CV. For example, each time, 30 samples are drawn
368 from the 85 samples as the test set to obtain the prediction accuracy CV for the situation "tr-hp"
369 and the rest is used for training.

370 4.2 Prediction intervals

371 CRPS (Continuous Ranked Probability Score) and coverage probabilities are used as quality indica-
372 tors of prediction intervals. CRPS is an uncertainty measure that assesses the similarities between
373 two distributions. We use it to indicate how the predicted distribution matches the observed dis-
374 tribution. The CRPS implemented as an R package **ScoringRules** (Jordan et al., 2017) is used.
375 CRPS is calculated for the INLA and QRF models. For the INLA model, the prediction intervals
376 are calculated by simulating from the response $Y \sim N(\theta, \sigma^2)$ where θ and σ^2 are the fitted mean and
377 variance. The mean of CRPS for all the points within each test block is calculated in spatial-blocked
378 CV. Coverage probabilities are calculated as the ratio between the number of predictions within
379 the upper and lower quantile and the total number of predictions (in the test set). The prediction
380 intervals are mainly compared between INLA, INLA-G, QRF and DF. The prediction interval for
381 QRFLA is compared with QRF to investigate the effects of Lasso tree-aggregation strategy on the
382 prediction intervals.

383 4.3 Model interpretation

384 We inspect fixed and spatial random effects modelled by INLA and compare the spatial random field
385 with modelled prediction intervals and model residuals to understand the contribution of spatial
386 random effects. Different from linear regression methods, which themselves are the best models for
387 interpretation, interpreting ensembling tree-based methods requires external models (Lundberg and
388 Lee, 2017). We use SHAP (SHapley Additive exPlanations, Lundberg et al., 2018; Lundberg and

³⁸⁹ Lee, 2017), a unified method based on additive feature attribution, to estimate variable influence in
³⁹⁰ RF and XGB models.

³⁹¹ 5 Results

³⁹² 5.1 Accuracy assessment and uncertainty quantification

³⁹³ Non-spatial CV

³⁹⁴ Both ensemble tree-based methods with a Gaussian objective function and INLA with a Gaussian
³⁹⁵ likelihood function obtain higher prediction accuracy than Lasso (table 3), indicating the necessity of
³⁹⁶ using a more flexible model and modelling spatial random fields. Among individual methods, in terms
³⁹⁷ of R^2 and RMSE, INLA with Gaussian likelihood obtained the highest prediction accuracy, followed
³⁹⁸ by XGB-G and QRFLA. QRFLA greatly improves from original RF. Despite the distribution of
³⁹⁹ response being closer to Gamma distribution compared to Gaussian distribution, using Gamma
⁴⁰⁰ regression in XGB and specifying Gamma likelihood in INLA both decrease the prediction accuracy
⁴⁰¹ considerably. Compared to INLA, XGB obtained lower RMSE and R^2 despite it obtained lower
⁴⁰² MAE and IQR, indicating that the XGB model predicts less well at more extreme ranges. The
⁴⁰³ QRF and DF results are not shown in table 3 as the results are very similar to RF. Their prediction
⁴⁰⁴ intervals are compared.

⁴⁰⁵ SE-INLA improves prediction accuracy compared to SE and INLA, obtained the lowest RMSE
⁴⁰⁶ and the highest R^2 among all the models. This indicates the spatial structures could further improve
⁴⁰⁷ prediction accuracy despite flexible relationships captured from ML models.

⁴⁰⁸ Spatial-blocked CV

⁴⁰⁹ Spatial-blocked CV provides information about prediction accuracy in spatial blocks. The R^2
⁴¹⁰ map (fig. 2) shows that the XGB, RF and INLA predict relatively well in most parts of Germany

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

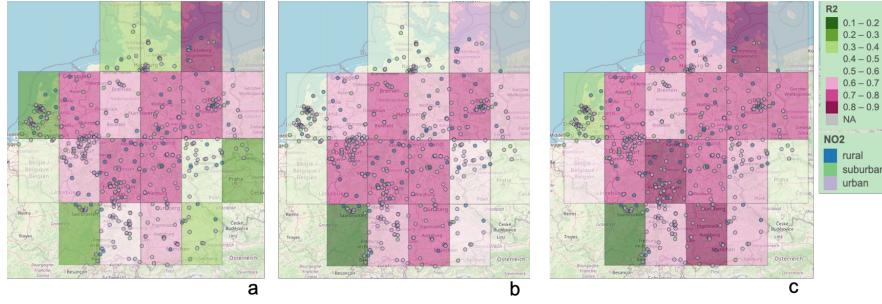


Figure 2: The R-squared of each block, using the rest of the blocks for training. The models are a) XGB, b) QRF, c) INLA.

411 besides blocks at the boundaries. The R^2 for the block western the Netherlands is also relatively low
 412 with all the three methods and especially for XGB (R^2 : 0.2). RF obtains the best result for the block
 413 of western the Netherlands (R^2 : 0.5). The INLA model outperforms RF and XGB in the blocks
 414 at south-east and north. The R^2 between blocks are the most heterogeneous with XGB, which is
 415 consistent to the result of bootstrapped CV that the XGB falls short at predicting extremes.

416 The spatial-blocked CRPS fig. 3 is computed for QRF and INLA (the DF is not shown as it will
 417 be shown that the QRF and DF performed similarly in prediction interval prediction (section 5.2)).
 418 The INLA predicted prediction distribution deviates considerably from observed distribution for the
 419 block of western the Netherlands, as reflected by the high value of mean CRPS. This is consistent
 420 to the relatively low R^2 observed for the same block. However, some blocks with relatively high R^2
 421 (in the north and south) have high CRPS. This indicates that the prediction mean is well-predicted
 422 but not the prediction interval (too narrow).

423 Customised CV

424 There is a distinctive difference between model performance in areas close to traffic (i.e. *tr-hp*
 425 and *tr-lmp*) and far away from traffic (i.e. *far*). The INLA model outperformed other non-spatial



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

426 methods in both *tr-hp* and *tr-lmp*, especially for the latter while the XGB model outperformed the
 427 INLA model (and all the other models) in *far*. This indicates the importance of modelling spatial
 428 dependency in areas close to traffic and possibly non-linear relationships far-away from roads. All the
 429 ensemble tree-based methods obtained much worse results compared to linear regression methods in
 430 *tr-lmp*. A linear regression model typically outperforms ensemble tree-based methods when there are
 431 relatively few observations for a flexible relationship to be justified. As the number of observations
 432 that are close to traffic and far away from traffic is balanced, the results indicate that the population
 433 density alters relationships between NO₂ and road density (i.e. the relationships between NO₂ and
 434 road density is different with different population density) in areas close to traffic.

435 5.2 Prediction interval

436 The 90% prediction intervals for INLA, INLA-G, DF, QRF and QRFLA are shown in figs. 4 to 6.
 437 The RF-based methods, namely DF, QRF and QRFLA reach the coverage probability higher than
 438 0.9, but the DF predicts a more realistic prediction quantile, notably, it covers four observations that
 439 are not covered by the same prediction quantiles predicted by the QRF. The INLA 90% prediction
 440 interval is too narrow. The coverage probability is 0.41 for INLA and 0.36 for INLA-G. The predicted

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

441 90th quantile of the INLA-G turned to better capture extreme high values but the model also turned
442 to miss more at lower values. The QRFLA predicted a slightly narrower prediction interval compared
443 to QRF. This indicates that Lasso reduced the variance of a QRF model by aggregating trees.

444 **5.3 Model Interpretation**

445 SHAP values are calculated for RF and XGB methods using all the data. The variables are ranked by
446 their variable importance, which is calculated as the sum of SHAP magnitudes over all the samples. It
447 can be observed from fig. 7 that the variable rankings and the pattern of variable impacts on model
448 output are similar. Both methods ranked road_class_2_100 at the top. The variable importance
449 calculated by the SHAP indicates a pattern that matches well with our expectation in the emission
450 sources (e.g. high pollution close to primary roads). To illustrate, we observe a positive trend of
451 SHAP values along with road_class_2_100 values, this matches with the explanation that areas with
452 higher primary road density generally experience higher NO₂ concentrations.

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

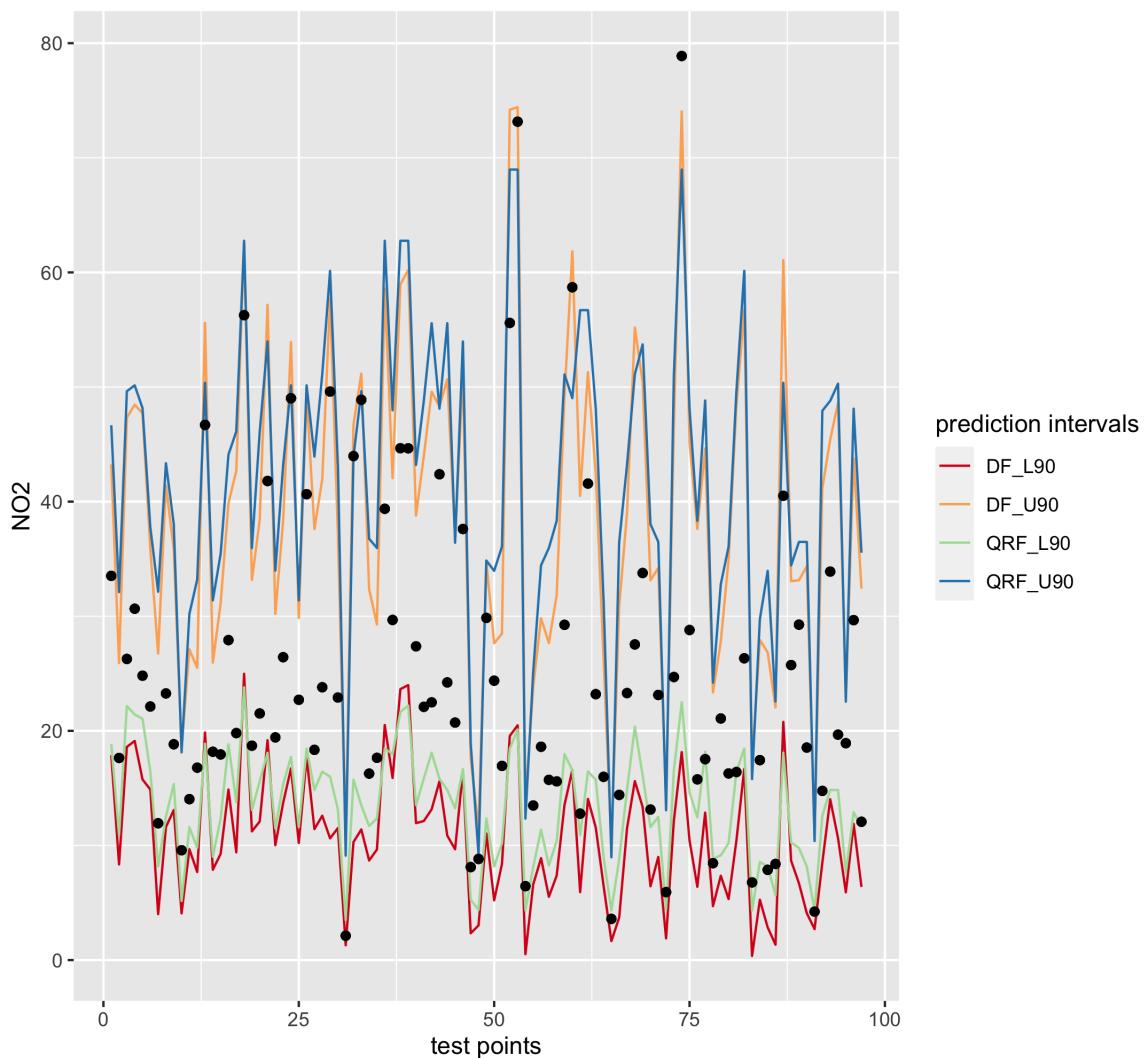


Figure 4: The 90% prediction interval predicted by DF and QRF. The black dots indicate observations in the test dataset.

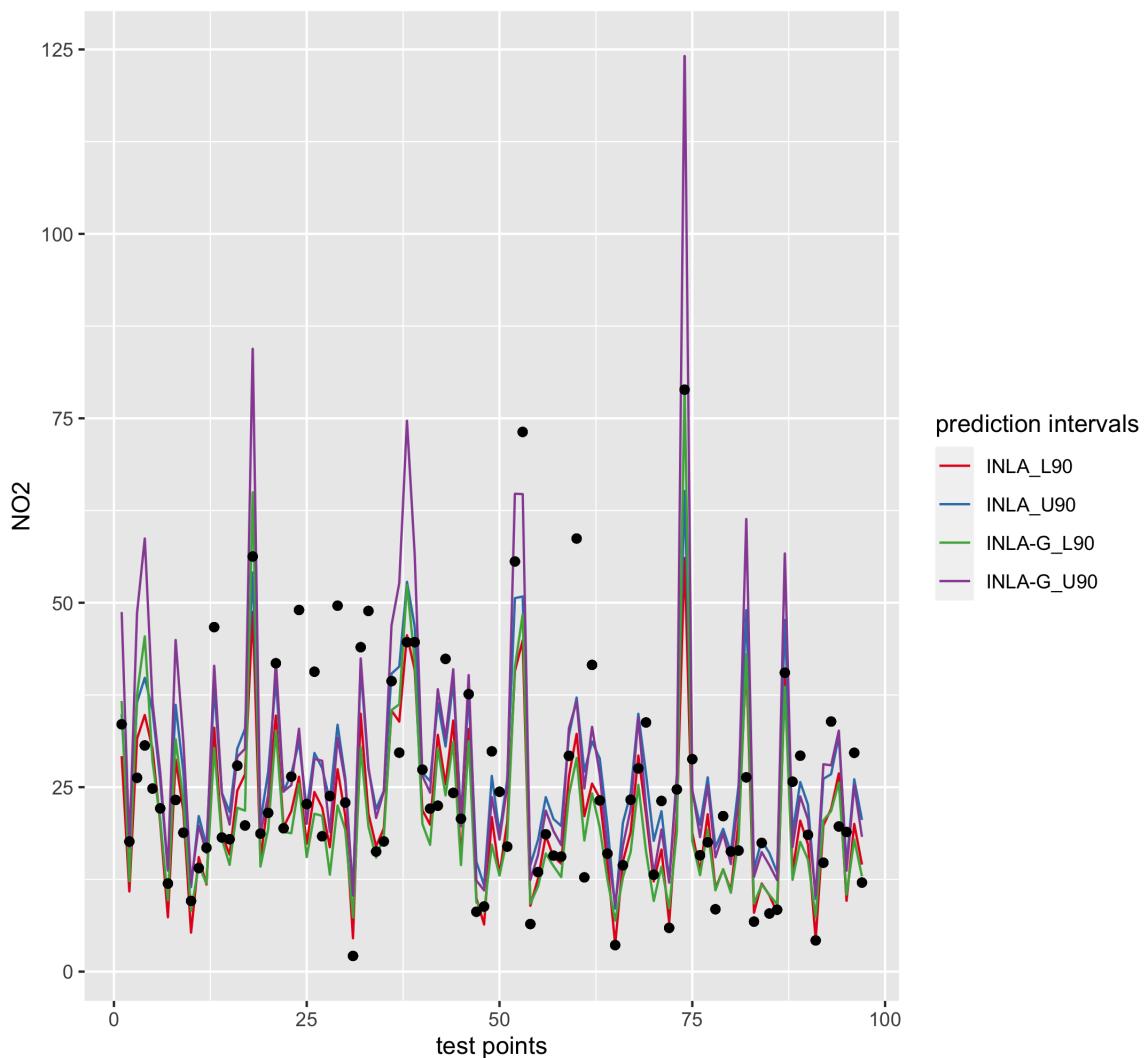


Figure 5: The 90% prediction interval predicted by INLA and INLA-G. The black dots indicate observations in the test dataset.

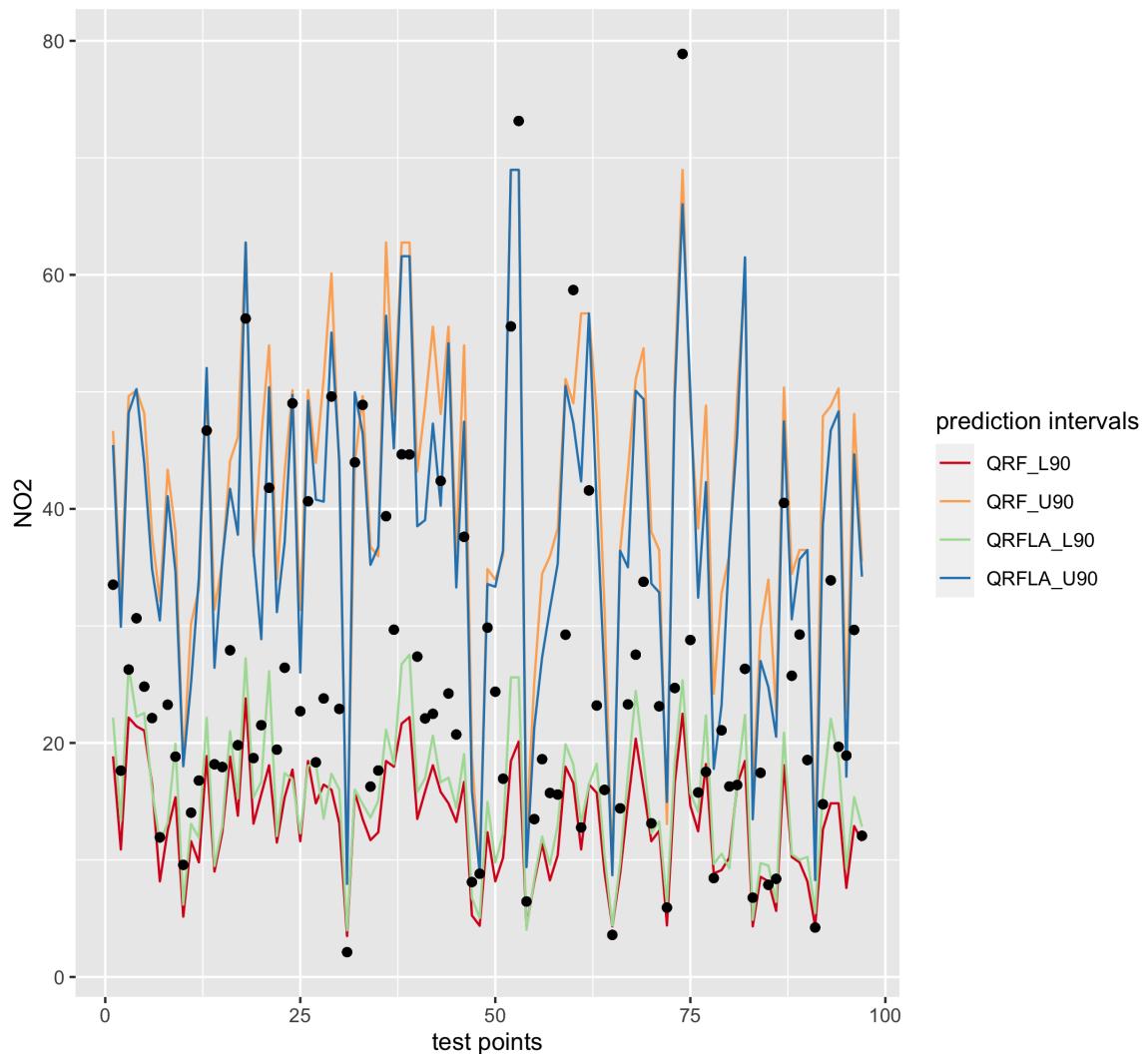


Figure 6: The 90% prediction interval predicted by QRF and QRFLA. The black dots indicate observations in the test dataset.

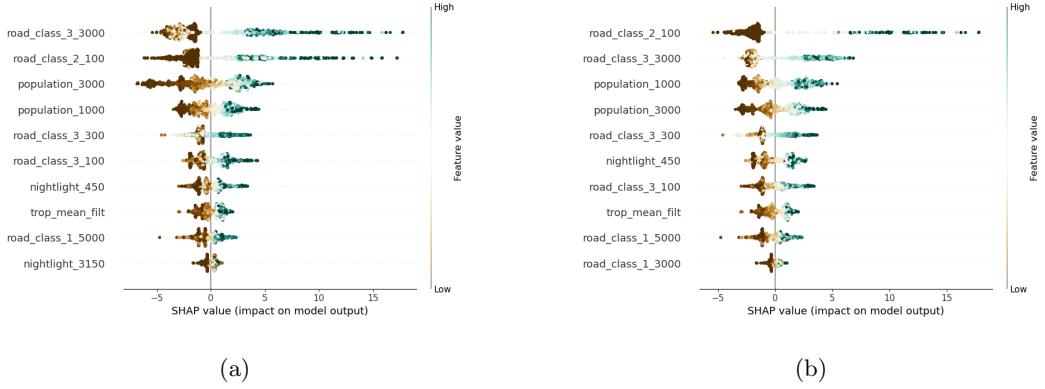


Figure 7: Variable impact calculated by SHAP (SHapley Additive exPlanations), a) the RF model, b) The XGB model. The horizontal location shows whether the effect of that value is associated with a higher or lower prediction. The covariate ranking is based on the sum of SHAP magnitudes over all the samples.

464 coefficient are shown in the supplementary material figure 3.

465 The differences between the predicted NO₂ and the mean of the spatial random field fig. 8
 466 indicates the effects of covariates. The highest values of the mean of the spatial random field are
 467 shown close to the Stuttgart region. Relatively high values can be observed in northern, southern
 468 and western Germany. Compared to fig. 9, the areas close to the Stuttgart (Germany) region where
 469 the mean values of the spatial random field are high corresponds to the high magnitudes of NO₂
 470 concentrations. Also, the differences between the observations and predictions are relatively large
 471 in magnitudes in this region. To facilitate visualisation, we also calculated the differences between
 472 INLA model predictions and the observations (supplementary material, figure 2).

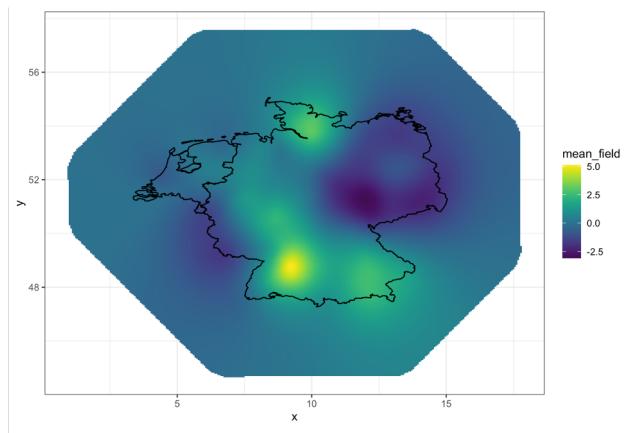


Figure 8: Mean of the spatial random field fitted by the INLA model.

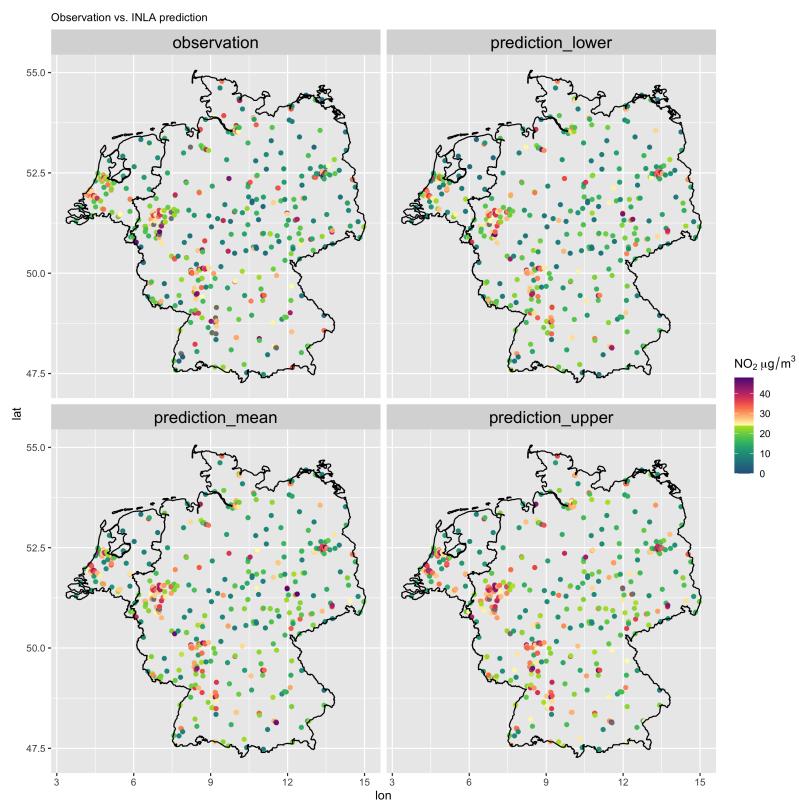


Figure 9: INLA predicted NO₂ at the ground stations with mean (prediction_mean), high (prediction_high, 0.975) and low (prediction_low, 0.925) quantiles and the observed NO₂ (observation).

473 6 Discussion

474 In this study, we compared geostatistical methods with ML methods for spatial NO₂ prediction in
475 Germany and the Netherlands. The comparison consists of the predicted mean, prediction inter-
476 vals, and model interpretation. Spatial and non-spatial CV strategies are used to reveal prediction
477 accuracy in different aspects. We also implemented the Lasso post-processed RF and geostatistical
478 stacked learning for NO₂ mapping (which to our knowledge have not been applied in air pollution
479 mapping before) and these two methods considerably improve from the original RF and stacked
480 learning methods, respectively.

481 Several venues were attempted to further improve the geostatistical model fitted with INLA.
482 Firstly, as we observed in general worse results at the geographical boundaries (figs. 2 and 3), we
483 inspected if different meshes with edge-effects fully accounted (e.g. the mesh is sufficiently large for
484 observations at the edge) could improve the prediction accuracy. It turned out that the same perfor-
485 mance is obtained. Secondly, we suspected that the deviation from assumed distribution (Gaussian)
486 is the cause of narrow prediction intervals of the INLA model. However, assuming a Gamma distri-
487 bution likelihood did not improve the model performance (in terms of the accuracy matrix, CRPS
488 and coverage probability). We also experienced the square transformation of the observations and
489 the use of the log-normal likelihood but that also decreases the model performance. Thirdly, we
490 additionally added two factor variables, namely "country code" (country code, "DE" for Germany
491 and "NL" for the Netherlands) and "urban types" (rural, urban, city centre according to (Dijkstra
492 and Poelman, 2014)). However, that also does not increase the model performance. In future works,
493 using a different spatial model (e.g. by specifying different hyperparameters), using the country and
494 urban types as mixed-effects, and modelling spatial varying coefficients may improve the modelling
495 results. Major improvement may also be achieved by integrating mobile sensing measurements and
496 other geospatial predictors (e.g. traffic count, urban morphological matrix) (Moraga et al., 2017).

497 We implemented an INLA model without modelling the spatial random effect (called non-spatial
498 INLA) to deepen our understanding of the effect of modelling the spatial process in our INLA model.
499 The non-spatial INLA model obtained lower DIC (Information Criterion) 3286.66 vs. 3251.97 (with
500 spatial effects) and WAIC (Watanabe-Akaike information criterion) 3291.75 vs. 3253.93 (with spa-
501 tial effects). These suggest the advantage of modelling the spatial effects. We normalised covari-
502 ates before inputting into the spatial and non-spatial INLA models and compared the differences
503 between the fixed-effects obtained by the original and non-spatial INLA model (supplementary ma-
504 terial figure 3-4) and found the most notable change is on the increased effect on the covariate
505 population_1000 for the non-spatial INLA model. This can be explained by that part of the effects
506 of population_1000 is modelled in the spatial random field. The second most notable change is on
507 the decreased effect of nightlight_450 for the non-spatial INLA model. After the spatial process is
508 modelled, the nightlight_450 has a higher contribution to the model. Together with the decreased
509 effects of road_class_2_100 and road_class_3_300 for the non-spatial INLA model, these may indicate
510 that the spatial model could better account for traffic-related variables (i.e. road and nightlight in
511 smaller buffers).

512 Model performance differs between the three road and population situations. The "far" situation
513 obtained the best modelling accuracy while the "tr-hp" the worst. This is likely due to the fact that
514 the urban NO₂ process is more complex due to urban forms and traffic conditions. This may also
515 indicate that more detailed traffic counts and meteorological data are needed for modelling the NO₂
516 emission sources.

517 Different from non-parametric models such as ensemble trees, a parametric geostatistical model
518 fitted with INLA as the one developed in our study requires feature selection and the assumption
519 of the distribution of the response. Several studies used the whole dataset for variable selection and
520 then use selected variables for CV (Lu et al., 2020b; Larkin et al., 2017). This may however lead to

521 an information leak as the validation data is also used in CV. To avoid this problem, one can include
522 the variable selection process in each CV (i.e. use the same training data for variable selection and
523 test). However, variable selection in each run added in additional error and uncertainty, therefore,
524 a determined set of covariates may be preferred. We obtain a fixed set of selected variables while
525 reducing information leakage to a negligible level by choosing only the variables that are selected
526 90% -100% times of all the bootstraps of Lasso.

527 Using the geostatistical method to stack learners obtained higher prediction accuracy in terms
528 of the mean prediction compared to the non-spatial stacking. This suggests the complex response-
529 covariate relationships modelled by the ML learners do not fully capture the spatial process. The
530 geostatistical stacked models obtained the highest prediction accuracy and with high-performance
531 computation, it is possible to apply them to a large-scale and at a high resolution. The limitation of
532 such stacked methods is that they cannot be used to analyse the effects of covariates and therefore
533 NO₂ emission sources. But these models could be a reference to the level of accuracy a statistical
534 predictive model could reach with the data available and the characteristics of the base learners
535 (here: if the base learners are global or local models).

536 7 Conclusion

537 We proposed a model comparison process to comprehensively compare between models considering
538 not only the predicted mean but also prediction intervals and model interpretation. We also showed
539 that the information provided by commonly single-used non-spatial CV may miss reflecting model
540 behaviours. With the model comparison process, we compared the use of geostatistical and ML
541 methods for the spatial prediction of NO₂ in Germany and the Netherlands and found noticeable
542 differences in their limitations and strength. The geostatistical models are preferred especially for
543 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 NO₂ 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.

554 **References**

- 555 C. Alakus, D. Larocque, and A. Labbe. Rfpredinterval: An r package for prediction intervals with
556 random forests and boosted forests. *arXiv preprint arXiv:2106.08217*, 2021.
- 557 L. Anselin et al. Spatial econometrics. *A companion to theoretical econometrics*, 310330, 2001.
- 558 A. Beloconi and P. Vounatsou. Bayesian geostatistical modelling of high-resolution no₂ exposure
559 in europe combining data from monitors, satellites and chemical transport models. *Environment*
560 *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>.
- 562 S. Bertazzon, M. Johnson, K. Eccles, and G. G. Kaplan. Accounting for spatial effects in land use
563 regression for urban air pollution modeling. *Spatial and Spatio-temporal Epidemiology*, 14-15:9 –
564 21, 2015. ISSN 1877-5845.
- 565 S. Bhatt, E. Cameron, S. R. Flaxman, D. J. Weiss, D. L. Smith, and P. W. Gething. Improved
566 prediction accuracy for disease risk mapping using gaussian process stacked generalization. *Journal*
567 *of the Royal Society Interface*, 14(134):20170520, 2017.
- 568 M. Blangiardo and M. Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John
569 Wiley & Sons, 2015.
- 570 L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- 571 D. J. Briggs, C. de Hoogh, J. Gulliver, J. Wills, P. Elliott, S. Kingham, and K. Smallbone. A
572 regression-based method for mapping traffic-related air pollution: application and testing in four
573 contrasting urban environments. *Science of the Total Environment*, 253(1-3):151–167, 2000.
- 574 J. Chen, K. de Hoogh, J. Gulliver, B. Hoffmann, O. Hertel, M. Ketzel, M. Bauwelinck, A. van
575 Donkelaar, U. A. Hvidtfeldt, K. Katsouyanni, et al. A comparison of linear regression, regular-

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

578 T. Chen and C. Guestrin. xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm*
579 *sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM,
580 2016.

581 T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano,
582 T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li. *xgboost: Extreme Gradient Boosting*, 2019b.
583 URL <https://CRAN.R-project.org/package=xgboost>. R package version 0.82.1.

584 M. Chiusolo, E. Cadum, M. Stafoggia, C. Galassi, G. Berti, A. Faustini, L. Bisanti, M. A. Vigotti,
585 M. P. Dessì, A. Cerniglio, et al. Short-term effects of nitrogen dioxide on mortality and sus-
586 ceptibility factors in 10 italian cities: the epiair study. *Environmental health perspectives*, 119(9):
587 1233–1238, 2011.

588 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 ac-
589 ccessed: Aug 3, 2021.

591 D. P. Dee, S. M. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda,
592 G. Balsamo, d. P. Bauer, et al. The era-interim reanalysis: Configuration and performance of the
593 data assimilation system. *Quarterly Journal of the royal meteorological society*, 137(656):553–597,
594 2011.

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

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

- 598 *urbanisation*, 2014. URL https://ec.europa.eu/regional_policy/sources/docgener/work/2014_01_new_urban.pdf. Last accessed: Aug 4, 2021.
- 600 T. Duan, A. Anand, D. Y. Ding, K. K. Thai, S. Basu, A. Ng, and A. Schuler. Ngboost: Natural
601 gradient boosting for probabilistic prediction. In *International Conference on Machine Learning*,
602 pages 2690–2700. PMLR, 2020.
- 603 Earthdata. *GES DISC*. URL "https://disc.gsfc.nasa.gov/datasets/OMN02d_003/summary?keywords=OMI%202017%20No2". last assessed May 21, 2019.
- 605 EEA. *Explore air pollution data*, 2021. URL <https://www.eea.europa.eu/themes/air/explore-air-pollution-data>.
- 607 F. Fouedjio and J. Klump. Exploring prediction uncertainty of spatial data in geostatistical and
608 machine learning approaches. *Environmental Earth Sciences*, 78(1):38, 2019.
- 609 J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*,
610 pages 1189–1232, 2001.
- 611 E. Gonzalez-Estrada and J. A. Villasenor-Alva. *goft: Tests of Fit for some Probability Distributions*,
612 2020. URL <https://CRAN.R-project.org/package=goft>. R package version 1.3.6.
- 613 T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference,
614 and prediction*. Springer Science & Business Media, 2009.
- 615 T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference,
616 and prediction, second edition*. Springer Science & Business Media, 2017.
- 617 G. Hoek, R. Beelen, K. De Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs. A review
618 of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric
619 environment*, 42(33):7561–7578, 2008.

- 620 G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume
621 112. Springer, 2013.
- 622 A. Jordan, F. Krüger, and S. Lerch. Evaluating probabilistic forecasts with scoringrules. *arXiv*
623 *preprint arXiv:1709.04743*, 2017.
- 624 J. Kerckhoffs, G. Hoek, L. Portengen, B. Brunekreef, and R. C. Vermeulen. Performance of pre-
625 diction algorithms for modeling outdoor air pollution spatial surfaces. *Environmental science &*
626 *technology*, 53(3):1413–1421, 2019.
- 627 E. T. Krainski, V. Gómez-Rubio, H. Bakka, A. Lenzi, D. Castro-Camilo, D. Simpson, F. Lindgren,
628 and H. Rue. *Advanced spatial modeling with stochastic partial differential equations using R and*
629 *INLA*. CRC Press, 2018.
- 630 A. Larkin, J. A. Geddes, R. V. Martin, Q. Xiao, Y. Liu, J. D. Marshall, M. Brauer, and P. Hystad.
631 Global land use regression model for nitrogen dioxide air pollution. *Environmental Science &*
632 *Technology*, 51(12):6957–6964, 2017.
- 633 J. J. Li, A. Jutzeler, B. Faltings, S. Winter, and C. Rizos. Estimating urban ultrafine particle
634 distributions with gaussian process models. *Research@ Locate14*, pages 145–153, 2014.
- 635 F. Lindgren, H. Rue, and J. Lindström. An explicit link between gaussian fields and gaussian
636 markov random fields: the stochastic partial differential equation approach. *Journal of the Royal*
637 *Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.
- 638 F. Lindgren, H. Rue, et al. Bayesian spatial modelling with r-inla. *Journal of Statistical Software*,
639 63(19):1–25, 2015.
- 640 Y. Liu, G. Cao, and N. Zhao. Integrate machine learning and geostatistics for high-resolution

- 641 mapping of ground-level pm2. 5 concentrations. In *Spatiotemporal Analysis of Air Pollution and*
642 *Its Application in Public Health*, pages 135–151. Elsevier, 2020.
- 643 M. Lu, O. Schmitz, K. de Hoogh, Q. Kai, and D. Karssenberg. Evaluation of different methods
644 and data sources to optimise modelling of no2 at a global scale. *Environment international*, 142:
645 105856, September 2020a. ISSN 1873-6750. doi: 10.1016/j.envint.2020.105856.
- 646 M. Lu, I. Soenario, M. Helbich, O. Schmitz, G. Hoek, M. van der Molen, and D. Karssenberg. Land
647 use regression models revealing spatiotemporal co-variation in no2, no, and o3 in the netherlands.
648 *Atmospheric Environment*, 223:117238, 2020b.
- 649 M. Lu, R. Dai, C. de Boer, O. Schmitz, I. Kooter, S. Cristescu, and D. Karssenberg. *Problems*
650 *in Statistical Modelling of Air Pollution Basing Solely on Ground Monitor Stations and a Novel*
651 *Mobile Sensing Instrument Solution*, 2021. submitted to Science of the Total Environment.
- 652 S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In
653 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and
654 R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Cur-
655 ran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>.
- 657 S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W.
658 Low, S.-F. Newman, J. Kim, et al. Explainable machine-learning predictions for the prevention
659 of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749, 2018.
- 660 K. Luo, R. Li, W. Li, Z. Wang, X. Ma, R. Zhang, X. Fang, Z. Wu, Y. Cao, and Q. Xu. Acute effects
661 of nitrogen dioxide on cardiovascular mortality in beijing: an exploration of spatial heterogeneity
662 and the district-specific predictors. *Scientific reports*, 6(1):1–13, 2016.

- 663 S. Martino and H. Rue. Implementing approximate bayesian inference using integrated nested laplace
664 approximation: A manual for the inla program. *Department of Mathematical Sciences, NTNU,*
665 *Norway*, 2009.
- 666 T. G. Martins, D. Simpson, F. Lindgren, and H. Rue. Bayesian computing with inla: new features.
667 *Computational Statistics & Data Analysis*, 67:68–83, 2013.
- 668 N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999,
669 2006.
- 670 P. Moraga. *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*. Chapman
671 & Hall/CRC, 2019.
- 672 P. Moraga, S. M. Cramb, K. L. Mengersen, and M. Pagano. A geostatistical model for combined
673 analysis of point-level and area-level data using inla and spde. *Spatial Statistics*, 21:27–41, 2017.
- 674 NASA. *Shuttle Radar Topography Mission*. URL [https://www2.jpl.nasa.gov/srtm/
675 dataprelimdescriptions.html](https://www2.jpl.nasa.gov/srtm/dataprelimdescriptions.html). last assessed Aug 15, 2021.
- 676 D. A. Nelson. European environment agency. *Colo. J. Int'l Envtl. L. & Pol'y*, 10:153, 1999.
- 677 NOAA. Dmsp and viirs data download. "<https://ngdc.noaa.gov/eog/download.html>", 2021.
678 Last Accessed: 11.03.2021.
- 679 OpenStreetMap contributors. Planet dump 7 Jan 2019 retrieved from <https://planet.osm.org>, 2019.
- 680 X. Ren, Z. Mi, and P. G. Georgopoulos. Comparison of machine learning and land use regression
681 for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations
682 across the contiguous united states. *Environment International*, 142:105827, 2020. ISSN 0160-
683 4120. doi: <https://doi.org/10.1016/j.envint.2020.105827>. URL <https://www.sciencedirect.com/science/article/pii/S0160412020317827>.

- 685 H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005.
- 686 H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by
687 using integrated nested laplace approximations. *Journal of the royal statistical society: Series b*
688 (*statistical methodology*), 71(2):319–392, 2009.
- 689 Y. Rybarczyk and R. Zalakeviciute. Machine learning approaches for outdoor air quality modelling:
690 A systematic review. *Applied Sciences*, 8(12):2570, 2018.
- 691 L. Schlosser, T. Hothorn, R. Stauffer, A. Zeileis, et al. Distributional regression forests for prob-
692 abilistic precipitation forecasting in complex terrain. *The Annals of Applied Statistics*, 13(3):
693 1564–1589, 2019.
- 694 G. Shaddick, M. L. Thomas, H. Amini, D. Broday, A. Cohen, J. Frostad, A. Green, S. Gumy, Y. Liu,
695 R. V. Martin, et al. Data integration for the assessment of population exposure to ambient air
696 pollution for global burden of disease assessment. *Environmental science & technology*, 52(16):
697 9069–9078, 2018.
- 698 D. M. Stasinopoulos, R. A. Rigby, et al. Generalized additive models for location scale and shape
699 (gamlss) in r. *Journal of Statistical Software*, 23(7):1–46, 2007.
- 700 M. L. Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business
701 Media, 2012.
- 702 J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing*
703 *letters*, 9(3):293–300, 1999.
- 704 J. Velthoen, C. Dombry, J.-J. Cai, and S. Engelke. Gradient boosting for extreme quantile regression.
705 *arXiv preprint arXiv:2103.00808*, 2021.

- 706 A. M. Vicedo-Cabrera, A. Biggeri, L. Grisotto, F. Barbone, and D. Catelan. A bayesian kriging
707 model for estimating residential exposure to air pollution of children living in a high-risk area in
708 italy. *Geospatial health*, 8(1):87–95, 2013.
- 709 J. A. Villaseñor and E. González-Estrada. A variance ratio test of fit for gamma distributions.
710 *Statistics & Probability Letters*, 96:281–286, 2015.
- 711 S. Wager, T. Hastie, and B. Efron. Confidence intervals for random forests: The jackknife and the
712 infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651, 2014.
- 713 Q. Wang, H. Feng, H. Feng, Y. Yu, J. Li, and E. Ning. The impacts of road traffic on urban air
714 quality in jinan based gwr and remote sensing. *Scientific Reports*, 11(1):1–9, 2021.
- 715 M. T. Young, M. J. Bechle, P. D. Sampson, A. A. Szpiro, J. D. Marshall, L. Sheppard, and J. D.
716 Kaufman. Satellite-based no₂ and model validation in a national prediction model based on
717 universal kriging and land-use regression. *Environmental science & technology*, 50(7):3686–3694,
718 2016.
- 719 C. Yuan. Models and methods for computationally efficient analysis of large spatial and spatio-
720 temporal data. 2011.
- 721 L. Zhai, S. Li, B. Zou, H. Sang, X. Fang, and S. Xu. An improved geographically weighted regression
722 model for pm2. 5 concentration estimation in large areas. *Atmospheric Environment*, 181:145–154,
723 2018.
- 724 Y. Zhan, Y. Luo, X. Deng, K. Zhang, M. Zhang, M. L. Grieneisen, and B. Di. Satellite-based
725 estimates of daily NO₂ exposure in China using hybrid random forest and spatiotemporal kriging
726 model. *Environmental science & technology*, 52(7):4180–4189, 2018.

⁷²⁷ B. Zou, Q. Pu, M. Bilal, Q. Weng, L. Zhai, and J. E. Nichol. High-resolution satellite mapping
⁷²⁸ of fine particulates based on geographically weighted regression. *IEEE Geoscience and Remote*
⁷²⁹ *Sensing Letters*, 13(4):495–499, 2016.