

1 Introduction

² NO₂ is a traffic-related air pollutant and has been found in epidemiological time series analysis
³ to highly associated with respiratory (Luo et al., 2016) and cardiovascular (Chiusolo et al., 2011)
⁴ diseases. NO₂ values are measured using monitoring stations at certain locations (e.g. close to traffic)
⁵ and most of the epidemiological studies identified the relationships between NO₂ and diseases or
⁶ hospital admission using a single NO₂ monitoring station to represent the entire district. However,
⁷ NO₂ is highly dynamic over the district and the difference in NO₂ concentrations will reflect on
⁸ personal exposures to NO₂. Detailed spatial mapping of NO₂ is therefore required for more accurate
⁹ quantification of the relationships between NO₂ and health effects. In addition, detailed NO₂ maps
¹⁰ are necessary for scientific recommendations to be provided to policymakers and city planners.

¹¹ Statistical methods for NO₂ mapping have attracted a lot of attention with the burgeoning Ma-
¹² chine Learning (ML)¹ methods and availability of ground monitoring station networks, atmospheric
¹³ satellite products, and geospatial predictors. Geospatial predictors are variables that are included
¹⁴ as covariates in a statistical air pollution model. Commonly used geospatial predictors are air
¹⁵ emission- (e.g. road networks) and dispersion-related (e.g. wind speed) variables, numerical mod-
¹⁶ elling (e.g. with chemistry transport model) output, and atmospheric remote sensing measurements
¹⁷ or products. A most recent (data available from Jan-2018) atmosphere sensing instrument, Tropomi

¹list of abbreviations: CRPS: Continuous Ranked Probability Score; CV: Cross Validation; DF: Distributional Forest; GRF: Gaussian Random Field; GMRF: Gaussian Markov Random Field; GAMLS: 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

18 (Tropospheric monitoring instrument, NSO and ESA, 2019) onboard of Sentinel 5p satellite, mea-
19 sures column density of a variety of gaseous air pollutants, in particular with an unprecedentedly
20 high resolution for NO₂ (3.5 km x 5.5 km, across along track, since 06 August 2019).

21 Statistical methods applied for spatial air pollution prediction can be broadly classified depending
22 on whether the spatial dependency is explicitly modelled. If not modelled, we refer to the methods
23 "non-spatial" and otherwise "spatial". Most of the spatial air pollution models were developed to
24 predict at coarser resolutions, commonly 1 km or coarser (Young et al., 2016; Shaddick et al., 2018;
25 Beloconi and Vounatsou, 2020). Non-spatial methods are more dominant in air pollution mapping,
26 particularly in high-resolution (100 m resolution or higher) mapping. Among them, LUR (Land
27 Use Regression) models which assumes linear relationships between NO₂ and geospatial predictors
28 are the most studied (Briggs et al., 2000; Hoek et al., 2008). Most recently, statistical learning (in
29 this study, "statistical learning" is used interchangeably with "machine learning") methods (Hastie
30 et al., 2009), including regularised linear regression (e.g. Lasso and Ridge regression (James et al.,
31 2013)), kernel methods such as support vector machine (Suykens and Vandewalle, 1999), ensemble
32 tree-based methods such as random forest (RF, Breiman, 2001) and XGBoost (XGB, Chen and
33 Guestrin, 2016), have been applied for feature selection or capturing non-linear response-covariate
34 relationships (Lu et al., 2020a; Chen et al., 2019a). In air pollution (not restricted to NO₂) mapping,
35 several studies compared between statistical learning and conventional LUR methods (Chen et al.,
36 2019a; Kerckhoffs et al., 2019; Lu et al., 2020a; Ren et al., 2020; Rybarczyk and Zalakeviciute, 2018).

37 Geostatistical models (e.g. Kriging) and Geographically Weighted Regression (GWR) are the
38 most used spatial methods for air pollution prediction (Vicedo-Cabrera et al., 2013; Li et al., 2014;
39 Wang et al., 2021; Zou et al., 2016) and these methods have been combined with dimension reduction
40 Zhai et al. (2018) and RF (Zhan et al., 2018; Liu et al., 2020) to improve NO₂ prediction accuracy.
41 A Bayesian geostatistical model is developed in Beloconi and Vounatsou (2020) to predict NO₂ by

42 integrating Tropomi satellite instrument measurements and chemical transport models. A GWR
43 model naturally models spatial varying coefficients by fitting multiple local regressions depending
44 on the homogeneity in response-covariate relationships when a number of observations are involved.
45 A typical geostatistical model can be viewed as consisting of two components: a mean function,
46 commonly a linear model, capturing the response-covariate relationships and a covariance function
47 modelling dependency of residuals from the mean (Bhatt et al., 2017). Conventional Kriging methods
48 suffer from the "big n problem", i.e. it may become computationally intractable with a large number
49 of observations. To deal with this problem, Lindgren et al. (2011) propose to use Stochastic Partial
50 Differential Equations (SPDE) to approximate the Gaussian Random Field (GRF) to a Gaussian
51 Markov Random Field (GMRF, Rue and Held (2005)). The main advantage of this is that the GMRF
52 has a sparse structure of the precision matrix, which is the inverse of the covariance matrix of a
53 GRF. Along with this, Rue et al. (2009) propose to use the Integrated Nested Laplace Approximation
54 (INLA) in a Bayesian framework to achieve the computational scalability of a geostatistical model
55 using approximations for all the estimations. This is especially advantageous when modelling NO₂
56 over a larger scale e.g., continental or global-scale modelling when a large amount of observations
57 are modelled, and in spatiotemporal modelling.

58 As spatial models are typically more complex compared to their non-spatial counterparts, several
59 studies compared spatial and non-spatial models to understand if the spatial effects could be simply
60 modelled by including certain covariates in LUR models. Young et al. (2016) studied the use of
61 universal Kriging (UK), OMI (Ozone Monitoring Instrument) satellite instrument (Earthdata) and
62 LUR models for NO₂ prediction at 2.5 km resolution. Young et al. (2016) indicated that either
63 using UK or adding OMI in the LUR model improves a LUR model but adding OMI in a UK
64 model only trivially improves the performance. Bertazzon et al. (2015) shows that the inclusion of
65 the meteorological variables accounts for spatial effects similarly to the use of spatial autoregressive

⁶⁶ models(Anselin et al., 2001). However, even if the spatial dependency can be captured by involving
⁶⁷ certain covariates in a LUR model, we may still need geostatistical methods to understand the
⁶⁸ spatial structure present in the data. Linear models have been used for the mean function but the
⁶⁹ relationships between NO₂ and predictors have been shown to be better modelled with non-linear
⁷⁰ ML methods (Lu et al., 2020a). Most recent studies attempt to replace the linear mean function
⁷¹ with ML models. Liu et al. (2020) applied a geostatistical model to the residuals from an RF model
⁷² for the spatial prediction of PM_{2.5}. In disease mapping, Bhatt et al. (2017) proposes to stack ML
⁷³ models to replace the mean function in a geostatistical model.

⁷⁴ Few studies have compared between geostatistical and ML methods, possibly because the ML
⁷⁵ methods are still relatively less studied in air pollution mapping and in the field of geostatistics. It
⁷⁶ might be more interesting to compare between geostatistical methods and ML methods than geosta-
⁷⁷ tistical methods and LUR, because ML methods may be more capable of (implicitly) capturing the
⁷⁸ spatial dependency by integrating covariates, when the number of observations is sufficient. More-
⁷⁹ over, most comparison studies only compare the cross-validation accuracy of the prediction mean
⁸⁰ (e.g. using R-squared, mean absolute error, or root mean squared error), ignoring the prediction
⁸¹ intervals. Also not discussed is the cause of the prediction errors, are they caused by missing co-
⁸² variants, violation of the model assumptions (e.g. data distribution, non-linearity), or inconsistent
⁸³ distributions between training and validation sets. Also, different cross-validation strategies, e.g.,
⁸⁴ how do we split the train-test sets, may lead to different model validation results. Current studies
⁸⁵ typically solely rely on k-fold splitting (Kerckhoff et al., 2019; Larkin et al., 2017; Ren et al., 2020)
⁸⁶ or bootstrapping (Lu et al., 2020a) to randomly splitting between train-test sets, which may be
⁸⁷ one-sided and does not provide an indication of accuracy in spatial blocks (but only at the locations
⁸⁸ of ground stations).

⁸⁹ 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.

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

- 118 1. Optimising a set of spatial hierarchical and ML models for NO₂ prediction in Germany and
119 the Netherlands.
- 120 2. Developing a non-spatial and a geostatistical stacked ensemble model, i.e., a stack of various
121 ML learners.
- 122 3. Model comparison regarding the predicted mean, prediction interval, and model interpretation.

123 The spatial Hierarchical model incorporates the spatial random effect along with other covariates
124 and the estimation is performed using the R package **INLA** (Rue et al., 2009; Martins et al., 2013).
125 XGB, RF and Lasso are chosen for the comparison with the geostatistical model and they also
126 form the base learners in the two (geostatistical and non-spatial) stacked learning models. The ML
127 methods are chosen for their dissimilarity. Specifically, Lasso is a linear regression model without
128 accounting for spatial dependency. RF and XGB are non-linear models with regression trees as base-
129 learners and are not affected by dependent covariates. XGB is a highly scalable boosting method
130 that builds tree models subsequently over the residuals of previous trees and has multiple routines
131 to penalise model over-fitting (Chen et al., 2019b), which has been reported in various studies to
132 obtain the highest prediction accuracy Lu et al. (2020a).

133 2 Data

134 NO₂ concentration measurements of 2017 from national ground stations of Germany and the Nether-
135 lands are used. The original hourly data is downloaded from the EEA (European Environment

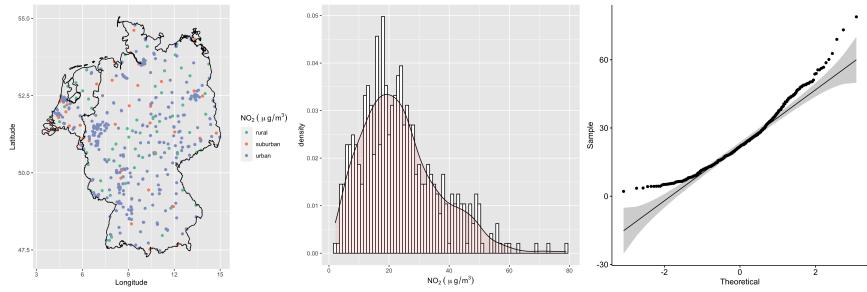


Figure 1: Spatial distribution of NO₂ stations, histogram and Q-Q plot of the NO₂ measurements.

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 NO₂ 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 ($p\text{-value}= 8.605\text{e-}12$, "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 ($p\text{-value} = 0.32$) 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 radiances 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 NO₂ column density (Copernicus, 2021) from 2019 (due to the increased resolution compared to 2018).

153 The buffered predictors of road and industry are calculated from OpenStreetMap (OpenStreetMap
154 contributors, 2019). For detailed descriptions of the processing of the geospatial predictors please
155 refer to Lu et al. (2020a).

156 3 Methods

157 The methods considered in this study are classified as spatial and non-spatial and are given the
158 names below in this study.

159 **Spatial models:**

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

164 **Non-spatial models:**

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

173 **3.1 Non-spatial methods**

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

180 QRF is a non-parametric prediction interval estimation method which keeps all the observations
181 in the terminal node for estimating the conditional probability function. Specifically, it samples
182 from all the response values in each terminal node and use the ratio between the number of samples
183 that is taken from each terminal node and the number of total observations in the terminal node as
184 weights to aggregate the samples. The weights of all the trees are summed. The summed weights
185 computed for each observation are then used to construct the empirical conditional cumulative
186 distribution function (Meinshausen, 2006). QRFLA uses Lasso as a post-processing of QRF (Hastie
187 et al., 2017, page 617). This method firstly preserves all the trees instead of aggregating them
188 (e.g. taking the mean of all the predictions) and then apply Lasso regression to all the trees for
189 aggregation. This leads to a shrinkage of the tree space and theoretically reduces model variance.
190 DF (Schlosser et al., 2019) firstly divide data into regions as homogeneous as possible with respect
191 to a parametric distribution, thus capturing changes in location, scale and shapes. For each tree,
192 maximum likelihood is used to fit distributions and recursively select and split covariates according
193 to the instability of the gradient of the likelihood at each observation along each co-variate. Then,
194 the distributional trees are ensembled for DF.

195 XGB is a variation of gradient boosting, which grows trees subsequently by fitting to model
196 residuals of the previous step. XGB is scalable to multiple threads. It enables multiple penalisation

197 paths to control model complexity to prevent model over-fitting, including regularisation (e.g. L1
198 regularisation) on tree width and terminal node values, as well as drop-out (dropping trees), sampling
199 observations (take a subset of observations in each run), and early stopping (stop iterating when after
200 a few rounds the loss does not decrease or the node does not meet the splitting rule). The default
201 objective function for regression assumes normal distribution of target variables (and the prediction
202 is the mean of the distribution). This assumption is used in all the air pollution mapping studies.
203 Here, we additionally fit a model with the objective function assuming the target variable follows a
204 Gamma distribution (XGB-G) as the distribution of NO₂ measurements is closer to Gamma than
205 normal distribution.

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

211 **3.2 Hyperparameter setting for XGB and RF**

212 To optimise the hyperparameters of XGB (known as ”model tuning”), we used grid search to optimise
213 hyperparameters in 5-fold cross-validation basing on the minimum RMSE (Root Mean Squared
214 Error) and additionally manual adjustment of the hyperparameters to look at the prediction patterns.
215 The grid search is used instead of more computationally efficient methods (e.g. Bayesian or random
216 search) as the optimal hyperparameter range is largely known from our previous experiences (Lu
217 et al., 2020a, 2021). The search grid for the number of iterations (nrounds) was from 200 to 3000,
218 with a step of 200; maximum tree depth (max-depth) from 3 to 6 with a step of 1, learning rate
219 (eta) from 0.001 to 0.1 with a step of 0.05, the penalty term Gamma (Chen et al., 2019b) from 1

220 to 5 with a step of 1, the subsample is set to 0.7, L1 norm penalisation (lambda) is set to 2 and L2
 221 norm penalisation (alpha) is set to 0. RF is not sensitive to hyperparameter tuning. We used the
 222 default setting of number of variables that are randomly drawn for each tree (Breiman, 2001), which
 223 is the integer part of the total number of variables divided by three. The number of trees is set to
 224 2000 for a safe choice as the high number of trees will not negatively affect model performance.

225 3.3 Geostatistical models

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

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

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

230 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
 231 coefficient vector, and $\mathbf{x}(\mathbf{s}_i)$ denotes a spatial Gaussian random field. That is, $\{\mathbf{x}(\mathbf{s}_1), \dots, \mathbf{x}(\mathbf{s}_n)\} \sim$
 232 $\mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma})$, where N_n is a Normal multivariate distribution for the spatial process specified by its
 233 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
 234 can be stationary and isotropic, where the covariance function depends only on the distance and not
 235 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
 236 using a Matérn function (Stein (2012); Yuan (2011); Diggle et al. (2013)). Since incorporating the
 237 spatial dependence directly with a large number of observations using a Gaussian random field is
 238 computationally expensive, Rue and Held (2005) proposed the approximation of a Gaussian random
 239 field by a Gaussian Markov random field for a more efficient computational process of estimation.

240 The main property of the Gaussian Markov random field is that it uses a conditional dependency
241 structure through the precision matrix \mathbf{Q} .

242 In this study, we compare two spatial hierarchical models with geospatial predictors as covariates,
243 one uses a Gaussian likelihood and the other a Gamma likelihood. The Gamma model has the same
244 hierarchical structure as the Gaussian model: the response variable in (1) can be represented by
245 $y_i \sim \text{Gamma}(\alpha, \beta)$ where α is the shape parameter and β the rate parameter. The INLA-SE model
246 uses a Gaussian likelihood.

247 **3.4 INLA and SPDE**

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

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

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

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

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

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

258 \mathbf{y} a vector of observations, assuming independent observations given the vector of the spatial latent
 259 field (\mathbf{x}) and the hyperparameters ($\boldsymbol{\theta}$), the likelihood can be expressed as:

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

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

261

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

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

264 and

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

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

266

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

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

270 where κ is a scale parameter, $x(\mathbf{s})$ is a spatial random field, Δ is the Laplacian, α is the parameter
 271 that controls the smoothness of the realizations, τ controls the variance and $\mathbf{W}(\mathbf{s})$ is a Gaussian
 272 spatial white noise process (Lindgren et al. (2015)). For the above we can use a Gaussian Markov
 273 random field that approximates to a Gaussian random field using a triangulation of the region of

274 study without specifying an explicit covariance structure through the SPDE method. This approx-
275 imation leads to a decrease in computational burden from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^{3/2})$.

276 3.5 Geospatial predictor selection for the INLA model

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

290 3.6 INLA model parameterisation

291 The triangulated mesh constructed in the SPDE approach is shown in supplementary material
292 (supfig. 1), with size of the inner and outer extensions around the data locations (*offsets*) 1/8 of
293 the maximum distance among all the observations for both the inner and outer extensions. The
294 maximum allowed triangle edge lengths in the region and in the extension (*max.edge*) are set
295 to respectively 1/30 and 1/5 times maximum distance among all the observations. The Matern

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

299 4 Model evaluation

300 4.1 Cross validation

301 We use RMSE, MAE (Mean Absolute Error), IQR (Interquartile Range) and R² (R-squared) to
302 compare model performance. RMSE is calculated as the square root of the differences between
303 predictions and observations; MAE is calculated as the absolute differences between predictions
304 and observations; IQR is the differences between the third and first quartiles of the prediction. R²
305 indicates the explained variance and is calculated as $R^2 = 1 - \text{var}(\text{error})/\text{var}(y)$, where var(.)
306 indicates variance, error indicates model residuals and y indicates observed response values. When
307 different data is used in CV (e.g. separating between close and far-away from roads), we additionally
308 calculated the RRMSE (relative RMSE), RMAE (relative MAE), RIQR (relative IQR) to account
309 for the differences in the magnitudes of response values. The RRMSE and RMAE are calculated by
310 dividing the RMSE and MAE, respectively, by the mean of observations. The RIQR was calculated
311 by dividing the IQR by the median of observations. The three CV methods we designed and used
312 to assess our model performance are:

- 313 1. Bootstrapped CV. 20-times randomly bootstrapped splitting of training and test sets (Lu et al.,
314 2020a).
- 315 2. Spatial-blocked CV. Dividing data into spatial blocks, each time use one block for test and
316 other blocks for training.

317 3. Customised CV. Splitting train-test based on values of certain covariates. In this study, three
318 sub-areas are defined, 1) close to traffic and with high population ("tr-hp"), 2) close to traffic
319 and with middle low population ("tr-lmp"), 3) far away from traffic ("far"). High population is
320 defined as the variable population of 1000 m buffer that is in the last quartile. Low population
321 is defined as the variable population of 1000 m buffer is below the median. Close to road is
322 defined as (please refer to table 1 for the definition of covariates):

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

326 Far away from road is defined as:

```
327       road_class_2_100 == 0 &  
328       road_class_1_100 == 0 &  
329       road_class_3_100 < quantile(road\class\_3\_100, .5)
```

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

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

337 **4.2 Prediction intervals**

338 CRPS (Continuous Ranked Probability Score) and coverage probabilities are used as quality indica-
339 tors of prediction intervals. CRPS is an uncertainty measure that assesses the similarities between

340 two distributions. We use it to indicate how the predicted distribution matches the observed dis-
341 tribution. The CRPS implemented as an R package **ScoringRules** (Jordan et al., 2017) is used.
342 CRPS is calculated for the INLA and QRF models. For the INLA model, the prediction intervals
343 are calculated by simulating from the response $Y \sim N(\theta, \sigma^2)$ where θ and σ^2 are the fitted mean and
344 variance. The mean of CRPS for all the points within each test block is calculated in spatial-blocked
345 CV. Coverage probabilities are calculated as the ratio between the number of predictions within
346 the upper and lower quantile and the total number of predictions (in the test set). The prediction
347 intervals are mainly compared between INLA, INLA-G, QRF and DF. The prediction interval for
348 QRFLA is compared with QRF to investigate the effects of Lasso tree-aggregation strategy on the
349 prediction intervals.

350 4.3 Model interpretation

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

358 5 Results

359 5.1 Accuracy assessment and uncertainty quantification

360 Non-spatial CV

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

372 SE-INLA improves prediction accuracy compared to SE and INLA, obtained the best results in
373 terms of root mean squared error (6.83, 24.5% of the mean of observations) and R^2 (0.71). This in-
374 dicates the spatial structures could further improve prediction accuracy despite flexible relationships
375 captured from ML models.

376 **Spatial-blocked CV**

377 Spatial-blocked CV provides information about prediction accuracy in spatial blocks. The R^2
378 map (fig. 2) shows that the XGB, RF and INLA predict relatively well in most parts of Germany
379 besides blocks at the boundaries. The R^2 for the block western the Netherlands is also relatively low
380 with all the three methods and especially for XGB (R^2 : 0.2). RF obtains the best result for the block
381 of western the Netherlands (R^2 : 0.5). The INLA model outperforms RF and XGB in the blocks
382 at south-east and north. The R^2 between blocks are the most heterogeneous with XGB, which is
383 consistent to the result of bootstrapped CV that the XGB falls short at predicting extremes.

384 The spatial-blocked CRPS fig. 3 is computed for QRF and INLA (the DF is not shown as it will

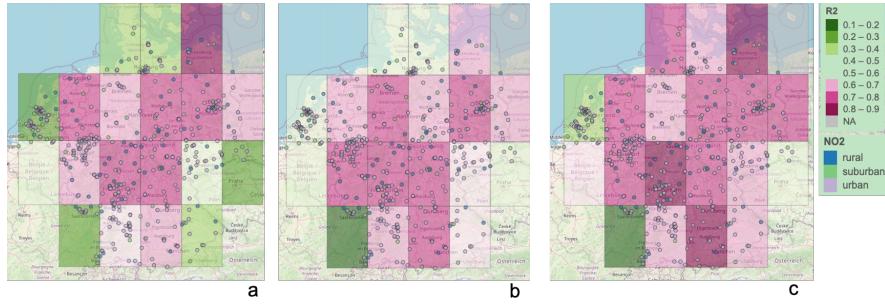


Figure 2: The R-squared of each block, using the rest of the blocks for training. The models are a) XGB, b) QRF, c) INLA.



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

385 be shown that the QRF and DF performed similarly in prediction interval prediction (section 5.2)).
 386 The INLA predicted prediction distribution deviates considerably from observed distribution for the
 387 block of western the Netherlands, as reflected by the high value of mean CRPS. This is consistent
 388 to the relatively low R^2 observed for the same block. However, some blocks with relatively high R^2
 389 (in the north and south) have high CRPS. This indicates that the prediction mean is well-predicted
 390 but not the prediction interval (too narrow).

391 **Customised CV**

392 There is a distinctive difference between model performance in areas close to traffic (i.e. *tr-hp*
393 and *tr-lmp*) and far away from traffic (i.e. *far*). The INLA model outperformed other non-spatial
394 methods in both *tr-hp* and *tr-lmp*, especially for the latter while the XGB model outperformed the
395 INLA model (and all the other models) in *far*. This indicates the importance of modelling spatial
396 dependency in areas close to traffic and possibly non-linear relationships far-away from roads. All the
397 ensemble tree-based methods obtained much worse results compared to linear regression methods in
398 *tr-lmp*. A linear regression model typically outperforms ensemble tree-based methods when there are
399 relatively few observations for a flexible relationship to be justified. As the number of observations
400 that are close to traffic and far away from traffic is balanced, the results indicate that the population
401 density alters relationships between NO₂ and road density (i.e. the relationships between NO₂ and
402 road density is different with different population density) in areas close to traffic.

403 5.2 Prediction interval

404 The 90% prediction intervals for INLA, INLA-G, DF, QRF and QRFLA are shown in figs. 4 to 6.
405 The RF-based methods, namely DF, QRF and QRFLA reach the coverage probability higher than
406 0.9, but the DF predicts a more realistic prediction quantile, notably, it covers four observations that
407 are not covered by the same prediction quantiles predicted by the QRF. The INLA 90% prediction
408 interval is too narrow. The coverage probability is 0.41 for INLA and 0.36 for INLA-G. The predicted
409 90th quantile of the INLA-G turned to better capture extreme high values but the model also turned
410 to miss more at lower values. The QRFLA predicted a slightly narrower prediction interval compared
411 to QRF. This indicates that Lasso reduced the variance of a QRF model by aggregating trees.

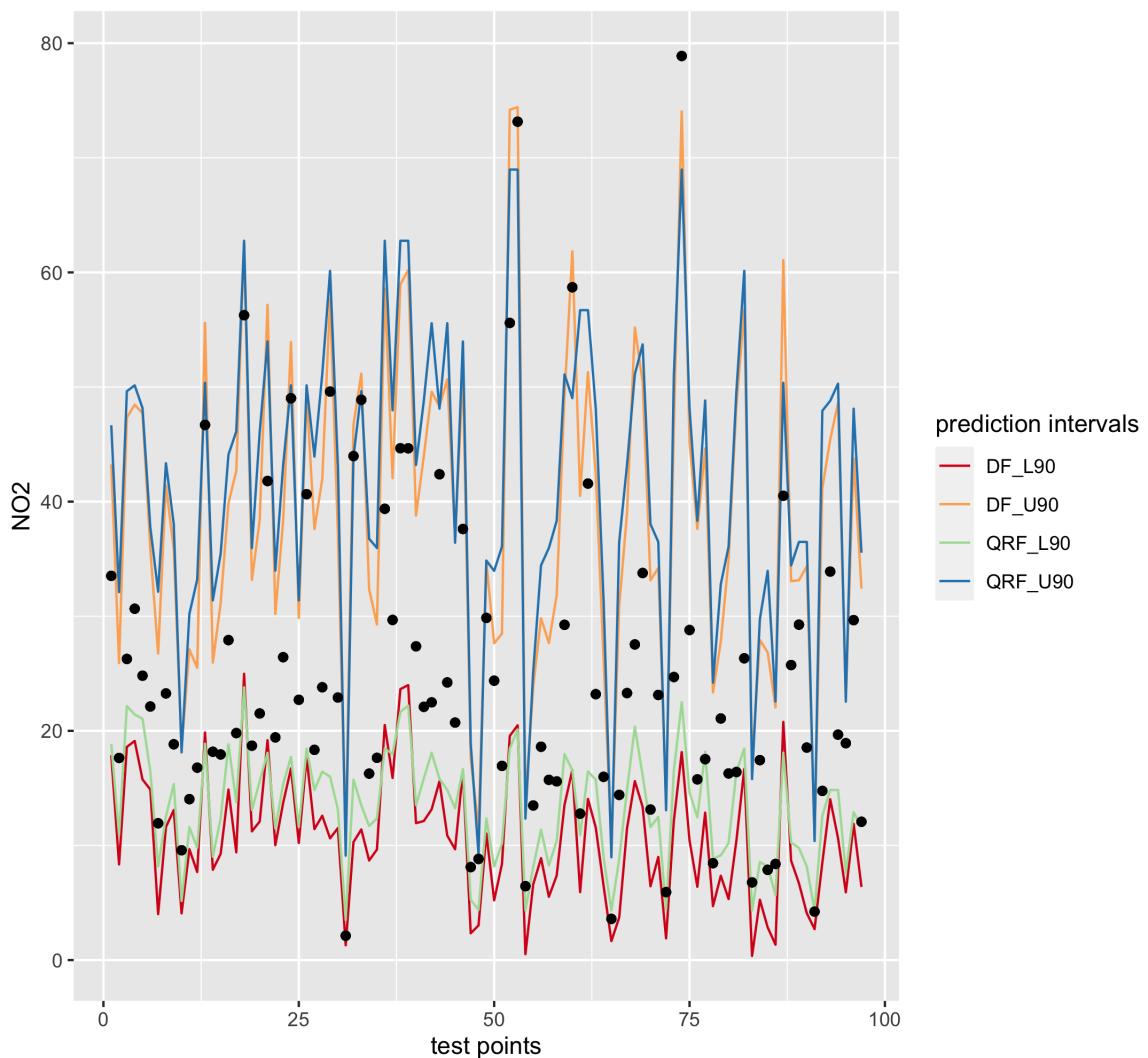


Figure 4: The 90% prediction interval predicted by DF and QRF. The black dots indicate observations in the test dataset.

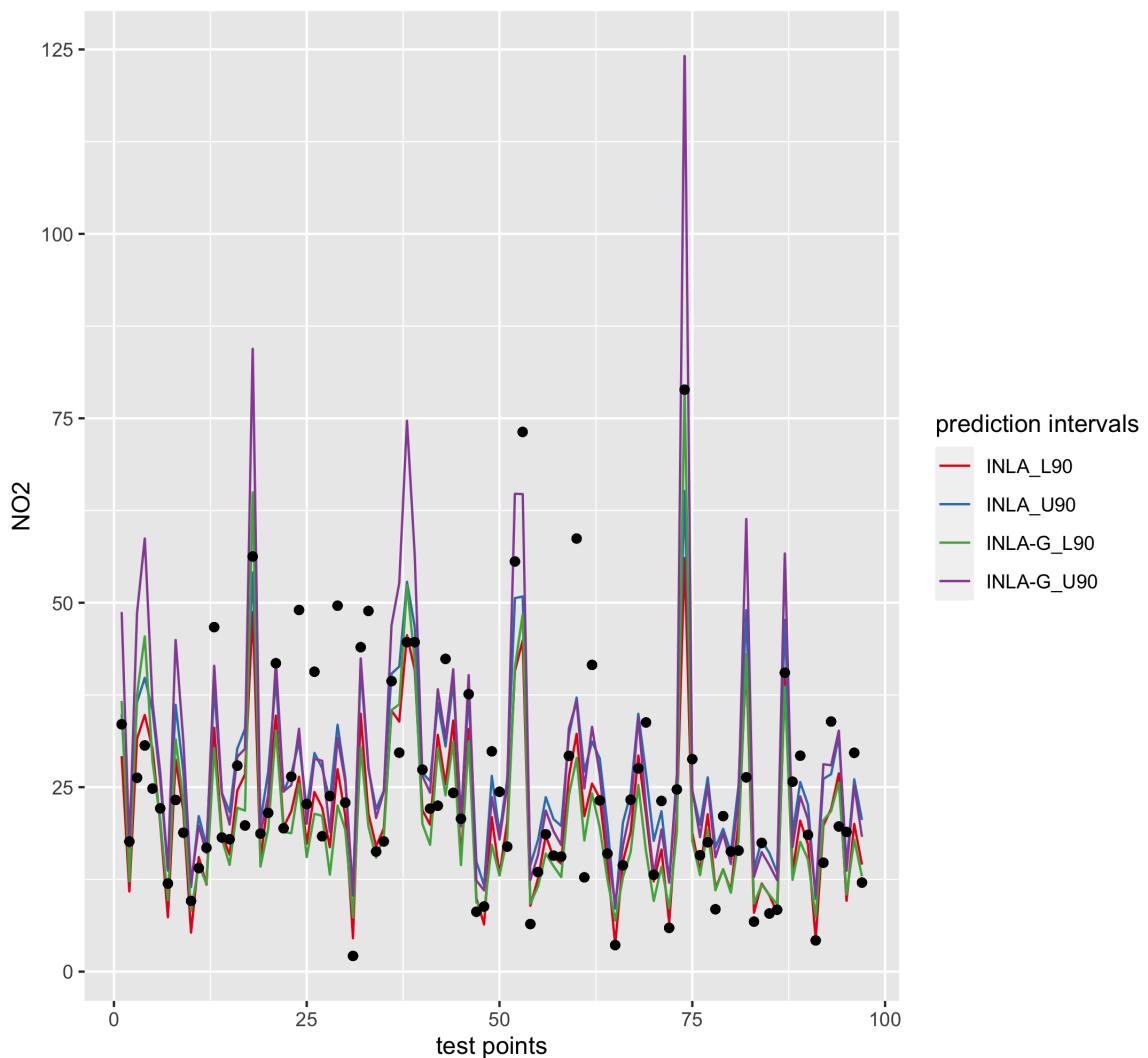


Figure 5: The 90% prediction interval predicted by INLA and INLA-G. The black dots indicate observations in the test dataset.

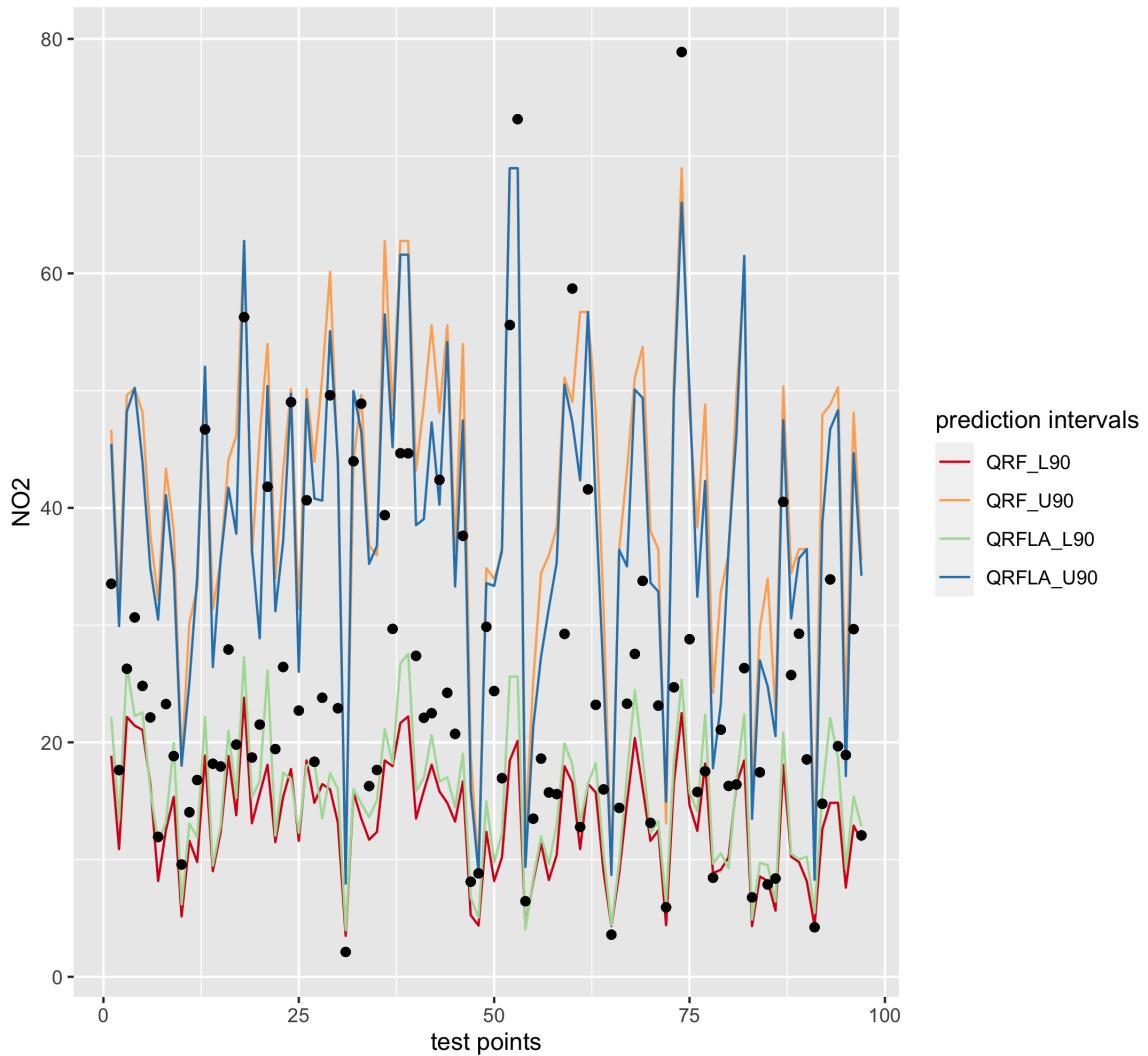


Figure 6: The 90% prediction interval predicted by QRF and QRFLA. The black dots indicate observations in the test dataset.

412 **5.3 Model Interpretation**

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

421 To analyse the effect of each covariate in the INLA model, we firstly normalised all the covariates
422 (by subtracting the mean and dividing the centred columns by their standard deviations) and used
423 all the data to fit the INLA model. road_class_2_100 has the highest effect (mean = 4.37), follows by
424 the population_3000 (3.08), these are consistent to the XGB variable importance (fig. 7b). Then,
425 the road_class_3_300 (3.00) has a notably higher effect (besides the top 2) than other covariates,
426 which has coefficients from 0.72 to 1.88. This differs from the XGB and RF variable importance
427 which ranked the population_1000 higher above, while in the INLA model the population_1000 has
428 the lowest effect (0.72). This may be because of the high correlation between population_1000 and
429 population_3000, as SHAP is a permutation test, it ignores the dependency between covariates.
430 In general, both geostatistical and ML methods estimated covariate effects match their physical
431 explanations. The statistics (mean, standard deviation, mode) and predicted quantiles of each
432 coefficient are shown in the supplementary material figure 3.

433 The differences between the predicted NO₂ and the mean of the spatial random field fig. 8
434 indicates the effects of covariates. The highest values of the mean of the spatial random field are
435 shown close to the Stuttgart region. Relatively high values can be observed in northern, southern

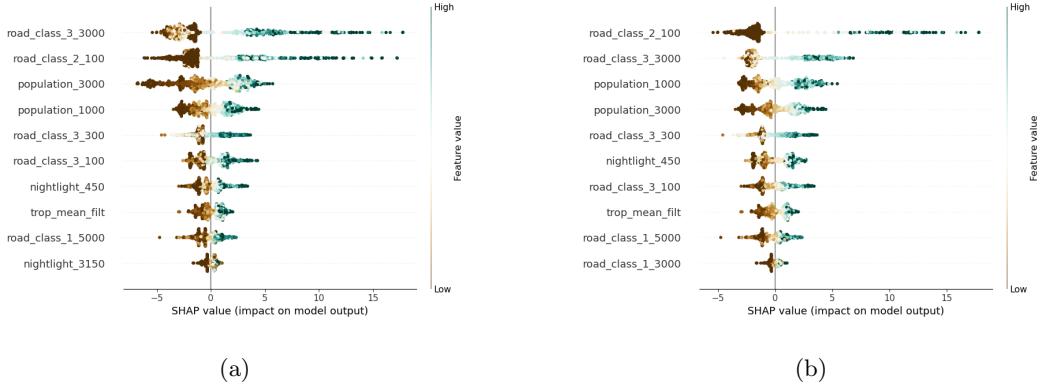


Figure 7: Variable impact calculated by SHAP (SHapley Additive exPlanations), a) the RF model, b) The XGB model. The horizontal location shows whether the effect of that value is associated with a higher or lower prediction. The covariate ranking is based on the sum of SHAP magnitudes over all the samples.

436 and western Germany. Compared to fig. 9, the areas close to the Stuttgart (Germany) region where
 437 the mean values of the spatial random field are high corresponds to the high magnitudes of NO₂
 438 concentrations. Also, the differences between the observations and predictions are relatively large
 439 in magnitudes in this region. To facilitate visualisation, we also calculated the differences between
 440 INLA model predictions and the observations (supplementary material, figure 2).

441 6 Discussion

442 In this study, we compared geostatistical methods with ML methods for spatial NO₂ prediction in
 443 Germany and the Netherlands. The comparison consists of the predicted mean, prediction inter-
 444 vals, and model interpretation. Spatial and non-spatial CV strategies are used to reveal prediction
 445 accuracy in different aspects. We also implemented the Lasso post-processed RF and geostatistical
 446 stacked learning for NO₂ mapping (which to our knowledge have not been applied in air pollution

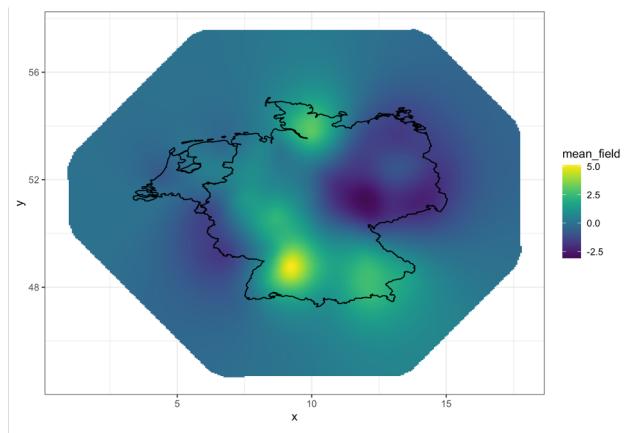


Figure 8: Mean of the spatial random field fitted by the INLA model.

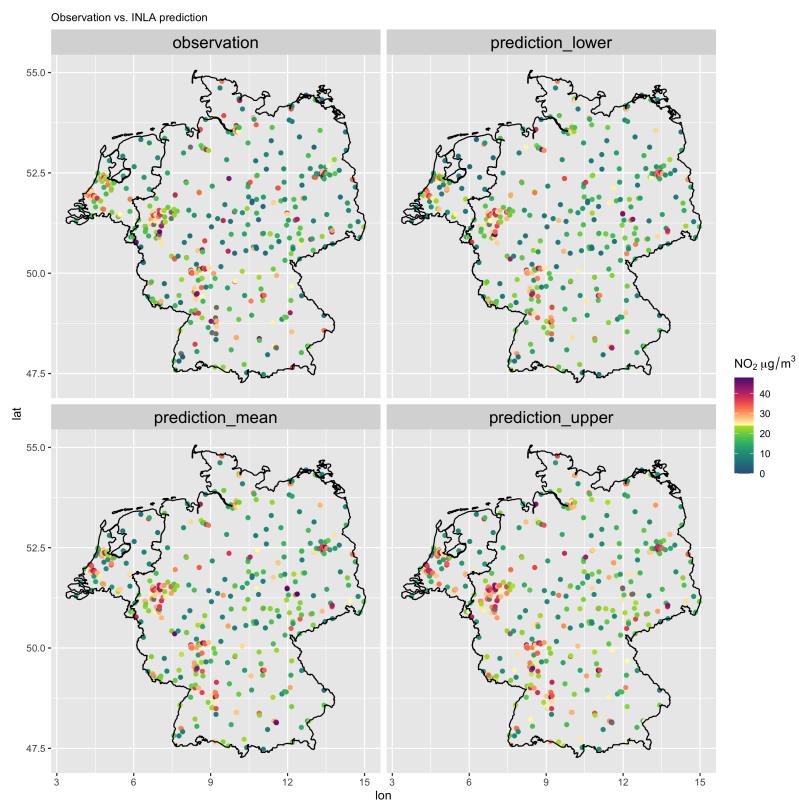


Figure 9: INLA predicted NO₂ at the ground stations with mean (prediction_mean), high (prediction_high, 0.975) and low (prediction_low, 0.925) quantiles and the observed NO₂ (observation).

⁴⁴⁷ mapping before) and these two methods considerably improve from the original RF and stacked
⁴⁴⁸ learning methods, respectively.

⁴⁴⁹ Several venues were attempted to further improve the geostatistical model fitted with INLA.
⁴⁵⁰ Firstly, as we observed in general worse results at the geographical boundaries (figs. 2 and 3), we
⁴⁵¹ inspected if different meshes with edge-effects fully accounted (e.g. the mesh is sufficiently large for
⁴⁵² observations at the edge) could improve the prediction accuracy. It turned out that the same perfor-
⁴⁵³ mance is obtained. Secondly, we suspected that the deviation from assumed distribution (Gaussian)
⁴⁵⁴ is the cause of narrow prediction intervals of the INLA model. However, assuming a Gamma distri-
⁴⁵⁵ bution likelihood did not improve the model performance (in terms of the accuracy matrix, CRPS
⁴⁵⁶ and coverage probability). We also experienced the square transformation of the observations and
⁴⁵⁷ the use of the log-normal likelihood but that also decreases the model performance. Thirdly, we
⁴⁵⁸ additionally added two factor variables, namely "country code" (country code, "DE" for Germany
⁴⁵⁹ and "NL" for the Netherlands) and "urban types" (rural, urban, city centre according to (Dijkstra
⁴⁶⁰ and Poelman, 2014)). However, that also does not increase the model performance. In future works,
⁴⁶¹ using a different spatial model (e.g. by specifying different hyperparameters), using the country and
⁴⁶² urban types as mixed-effects, and modelling spatial varying coefficients may improve the modelling
⁴⁶³ results. Major improvement may also be achieved by integrating mobile sensing measurements and
⁴⁶⁴ other geospatial predictors (e.g. traffic count, urban morphological matrix) (Moraga et al., 2017).

⁴⁶⁵ We implemented an INLA model without modelling the spatial random effect (called non-spatial
⁴⁶⁶ INLA) to deepen our understanding of the effect of modelling the spatial process in our INLA model.
⁴⁶⁷ The non-spatial INLA model obtained lower DIC (Information Criterion) 3286.66 vs. 3251.97 (with
⁴⁶⁸ spatial effects) and WAIC (Watanabe-Akaike information criterion) 3291.75 vs. 3253.93 (with spa-
⁴⁶⁹ tial effects). These suggest the advantage of modelling the spatial effects. We normalised covari-
⁴⁷⁰ ates before inputting into the spatial and non-spatial INLA models and compared the differences

471 between the fixed-effects obtained by the original and non-spatial INLA model (supplementary ma-
472 terial figure 3-4) and found the most notable change is on the increased effect on the covariate
473 population_1000 for the non-spatial INLA model. This can be explained by that part of the effects
474 of population_1000 is modelled in the spatial random field. The second most notable change is on
475 the decreased effect of nightlight_450 for the non-spatial INLA model. After the spatial process is
476 modelled, the nightlight_450 has a higher contribution to the model. Together with the decreased
477 effects of road_class_2_100 and road_class_3_300 for the non-spatial INLA model, these may indicate
478 that the spatial model could better account for traffic-related variables (i.e. road and nightlight in
479 smaller buffers).

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

485 Different from non-parametric models such as ensemble trees, a parametric geostatistical model
486 fitted with INLA as the one developed in our study requires feature selection and the assumption
487 of the distribution of the response. Several studies used the whole dataset for variable selection and
488 then use selected variables for CV (Lu et al., 2020b; Larkin et al., 2017). This may however lead to
489 an information leak as the validation data is also used in CV. To avoid this problem, one can include
490 the variable selection process in each CV (i.e. use the same training data for variable selection and
491 test). However, variable selection in each run added in additional error and uncertainty, therefore,
492 a determined set of covariates may be preferred. We obtain a fixed set of selected variables while
493 reducing information leakage to a negligible level by choosing only the variables that are selected
494 90% -100% times of all the bootstraps of Lasso.

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

504 7 Conclusion

505 We proposed a model comparison process to comprehensively compare between models considering
506 not only the predicted mean but also prediction intervals and model interpretation. We also showed
507 that the information provided by commonly single-used non-spatial CV may miss reflecting model
508 behaviours. With the model comparison process, we compared the use of geostatistical and ML
509 methods for the spatial prediction of NO₂ in Germany and the Netherlands and found noticeable
510 differences in their limitations and strength. The geostatistical models are preferred especially for
511 urban area prediction and provide the spatial process of observations and indicate the insufficient
512 modelling of spatial random-effects of fixed-effects. But the uncertainty assessment of geostatistical
513 methods, which is commonly known as strength, fails to provide a prediction interval that meets
514 the expectation. The QRF and DF obtained satisfying prediction intervals, with the DF slightly
515 more capable of predicting the extremes. Using Lasso to aggregate trees in random forest increase
516 model performance and reduce model variance. Using the geostatistical method to stack learners
517 obtained the highest accuracy in terms of the mean prediction. Despite the NO₂ observations follow

518 closer to a Gamma distribution than a Gaussian, the use of a Gamma likelihood in the geostatistical
519 model and Gamma objective in the XGBoost obtained much worse results than using a Gaussian
520 likelihood or objective. By comparing with the non-spatial stacking, geostatistical stacking suggests
521 the necessity of modelling the spatial process.

522 **References**

- 523 C. Alakus, D. Larocque, and A. Labbe. Rfpredinterval: An r package for prediction intervals with
524 random forests and boosted forests. *arXiv preprint arXiv:2106.08217*, 2021.
- 525 L. Anselin et al. Spatial econometrics. *A companion to theoretical econometrics*, 310330, 2001.
- 526 A. Beloconi and P. Vounatsou. Bayesian geostatistical modelling of high-resolution no₂ exposure
527 in europe combining data from monitors, satellites and chemical transport models. *Environment*
528 *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>.
- 530 S. Bertazzon, M. Johnson, K. Eccles, and G. G. Kaplan. Accounting for spatial effects in land use
531 regression for urban air pollution modeling. *Spatial and Spatio-temporal Epidemiology*, 14-15:9 –
532 21, 2015. ISSN 1877-5845.
- 533 S. Bhatt, E. Cameron, S. R. Flaxman, D. J. Weiss, D. L. Smith, and P. W. Gething. Improved
534 prediction accuracy for disease risk mapping using gaussian process stacked generalization. *Journal*
535 *of the Royal Society Interface*, 14(134):20170520, 2017.
- 536 M. Blangiardo and M. Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John
537 Wiley & Sons, 2015.
- 538 L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- 539 D. J. Briggs, C. de Hoogh, J. Gulliver, J. Wills, P. Elliott, S. Kingham, and K. Smallbone. A
540 regression-based method for mapping traffic-related air pollution: application and testing in four
541 contrasting urban environments. *Science of the Total Environment*, 253(1-3):151–167, 2000.
- 542 J. Chen, K. de Hoogh, J. Gulliver, B. Hoffmann, O. Hertel, M. Ketzel, M. Bauwelinck, A. van
543 Donkelaar, U. A. Hvidtfeldt, K. Katsouyanni, et al. A comparison of linear regression, regular-

- 544 ization, and machine learning algorithms to develop Europe-wide spatial models of fine particles
545 and nitrogen dioxide. *Environment international*, 130:104934, 2019a.
- 546 T. Chen and C. Guestrin. xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm*
547 *sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM,
548 2016.
- 549 T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano,
550 T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li. *xgboost: Extreme Gradient Boosting*, 2019b.
551 URL <https://CRAN.R-project.org/package=xgboost>. R package version 0.82.1.
- 552 M. Chiusolo, E. Cadum, M. Stafoggia, C. Galassi, G. Berti, A. Faustini, L. Bisanti, M. A. Vigotti,
553 M. P. Dessì, A. Cerniglio, et al. Short-term effects of nitrogen dioxide on mortality and sus-
554 ceptibility factors in 10 italian cities: the epiair study. *Environmental health perspectives*, 119(9):
555 1233–1238, 2011.
- 556 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-
557 cessed: Aug 3, 2021.
- 559 D. P. Dee, S. M. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda,
560 G. Balsamo, d. P. Bauer, et al. The era-interim reanalysis: Configuration and performance of the
561 data assimilation system. *Quarterly Journal of the royal meteorological society*, 137(656):553–597,
562 2011.
- 563 P. J. Diggle, P. Moraga, B. Rowlingson, and B. M. Taylor. Spatial and spatio-temporal log-gaussian
564 cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.
- 565 L. Dijkstra and H. Poelman. *A harmonised definition of cities and rural areas: the new degree of*

- 566 *urbanisation*, 2014. URL https://ec.europa.eu/regional_policy/sources/docgener/work/2014_01_new_urban.pdf. Last accessed: Aug 4, 2021.
- 568 T. Duan, A. Anand, D. Y. Ding, K. K. Thai, S. Basu, A. Ng, and A. Schuler. Ngboost: Natural
569 gradient boosting for probabilistic prediction. In *International Conference on Machine Learning*,
570 pages 2690–2700. PMLR, 2020.
- 571 Earthdata. *GES DISC*. URL "https://disc.gsfc.nasa.gov/datasets/OMN02d_003/summary?keywords=OMI%202017%20No2". last assessed May 21, 2019.
- 573 EEA. *Explore air pollution data*, 2021. URL <https://www.eea.europa.eu/themes/air/explore-air-pollution-data>.
- 575 F. Fouedjio and J. Klump. Exploring prediction uncertainty of spatial data in geostatistical and
576 machine learning approaches. *Environmental Earth Sciences*, 78(1):38, 2019.
- 577 J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*,
578 pages 1189–1232, 2001.
- 579 E. Gonzalez-Estrada and J. A. Villasenor-Alva. *goft: Tests of Fit for some Probability Distributions*,
580 2020. URL <https://CRAN.R-project.org/package=goft>. R package version 1.3.6.
- 581 T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference,
582 and prediction*. Springer Science & Business Media, 2009.
- 583 T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference,
584 and prediction, second edition*. Springer Science & Business Media, 2017.
- 585 G. Hoek, R. Beelen, K. De Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs. A review
586 of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric
587 environment*, 42(33):7561–7578, 2008.

- 588 G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume
589 112. Springer, 2013.
- 590 A. Jordan, F. Krüger, and S. Lerch. Evaluating probabilistic forecasts with scoringrules. *arXiv*
591 *preprint arXiv:1709.04743*, 2017.
- 592 J. Kerckhoffs, G. Hoek, L. Portengen, B. Brunekreef, and R. C. Vermeulen. Performance of pre-
593 diction algorithms for modeling outdoor air pollution spatial surfaces. *Environmental science &*
594 *technology*, 53(3):1413–1421, 2019.
- 595 E. T. Krainski, V. Gómez-Rubio, H. Bakka, A. Lenzi, D. Castro-Camilo, D. Simpson, F. Lindgren,
596 and H. Rue. *Advanced spatial modeling with stochastic partial differential equations using R and*
597 *INLA*. CRC Press, 2018.
- 598 A. Larkin, J. A. Geddes, R. V. Martin, Q. Xiao, Y. Liu, J. D. Marshall, M. Brauer, and P. Hystad.
599 Global land use regression model for nitrogen dioxide air pollution. *Environmental Science &*
600 *Technology*, 51(12):6957–6964, 2017.
- 601 J. J. Li, A. Jutzeler, B. Faltings, S. Winter, and C. Rizos. Estimating urban ultrafine particle
602 distributions with gaussian process models. *Research@ Locate14*, pages 145–153, 2014.
- 603 F. Lindgren, H. Rue, and J. Lindström. An explicit link between gaussian fields and gaussian
604 markov random fields: the stochastic partial differential equation approach. *Journal of the Royal*
605 *Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.
- 606 F. Lindgren, H. Rue, et al. Bayesian spatial modelling with r-inla. *Journal of Statistical Software*,
607 63(19):1–25, 2015.
- 608 Y. Liu, G. Cao, and N. Zhao. Integrate machine learning and geostatistics for high-resolution

- 609 mapping of ground-level pm2. 5 concentrations. In *Spatiotemporal Analysis of Air Pollution and*
610 *Its Application in Public Health*, pages 135–151. Elsevier, 2020.
- 611 M. Lu, O. Schmitz, K. de Hoogh, Q. Kai, and D. Karssenberg. Evaluation of different methods
612 and data sources to optimise modelling of no2 at a global scale. *Environment international*, 142:
613 105856, September 2020a. ISSN 1873-6750. doi: 10.1016/j.envint.2020.105856.
- 614 M. Lu, I. Soenario, M. Helbich, O. Schmitz, G. Hoek, M. van der Molen, and D. Karssenberg. Land
615 use regression models revealing spatiotemporal co-variation in no2, no, and o3 in the netherlands.
616 *Atmospheric Environment*, 223:117238, 2020b.
- 617 M. Lu, R. Dai, C. de Boer, O. Schmitz, I. Kooter, S. Cristescu, and D. Karssenberg. *Problems*
618 *in Statistical Modelling of Air Pollution Basing Solely on Ground Monitor Stations and a Novel*
619 *Mobile Sensing Instrument Solution*, 2021. submitted to Science of the Total Environment.
- 620 S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In
621 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and
622 R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Cur-
623 ran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>.
- 625 S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W.
626 Low, S.-F. Newman, J. Kim, et al. Explainable machine-learning predictions for the prevention
627 of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749, 2018.
- 628 K. Luo, R. Li, W. Li, Z. Wang, X. Ma, R. Zhang, X. Fang, Z. Wu, Y. Cao, and Q. Xu. Acute effects
629 of nitrogen dioxide on cardiovascular mortality in beijing: an exploration of spatial heterogeneity
630 and the district-specific predictors. *Scientific reports*, 6(1):1–13, 2016.

- 631 S. Martino and H. Rue. Implementing approximate bayesian inference using integrated nested laplace
 632 approximation: A manual for the inla program. *Department of Mathematical Sciences, NTNU,*
 633 *Norway*, 2009.
- 634 T. G. Martins, D. Simpson, F. Lindgren, and H. Rue. Bayesian computing with inla: new features.
 635 *Computational Statistics & Data Analysis*, 67:68–83, 2013.
- 636 N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999,
 637 2006.
- 638 P. Moraga. *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*. Chapman
 639 & Hall/CRC, 2019.
- 640 P. Moraga, S. M. Cramb, K. L. Mengersen, and M. Pagano. A geostatistical model for combined
 641 analysis of point-level and area-level data using inla and spde. *Spatial Statistics*, 21:27–41, 2017.
- 642 NASA. *Shuttle Radar Topography Mission*. URL [https://www2.jpl.nasa.gov/srtm/
 643 dataprelimdescriptions.html](https://www2.jpl.nasa.gov/srtm/dataprelimdescriptions.html). last assessed Aug 15, 2021.
- 644 D. A. Nelson. European environment agency. *Colo. J. Int'l Envtl. L. & Pol'y*, 10:153, 1999.
- 645 NOAA. Dmsp and viirs data download. "<https://ngdc.noaa.gov/eog/download.html>", 2021.
 646 Last Accessed: 11.03.2021.
- 647 OpenStreetMap contributors. Planet dump 7 Jan 2019 retrieved from <https://planet.osm.org>, 2019.
- 648 X. Ren, Z. Mi, and P. G. Georgopoulos. Comparison of machine learning and land use regression
 649 for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations
 650 across the contiguous united states. *Environment International*, 142:105827, 2020. ISSN 0160-
 651 4120. doi: <https://doi.org/10.1016/j.envint.2020.105827>. URL <https://www.sciencedirect.com/science/article/pii/S0160412020317827>.

- 653 H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005.
- 654 H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by
655 using integrated nested laplace approximations. *Journal of the royal statistical society: Series b*
656 (*statistical methodology*), 71(2):319–392, 2009.
- 657 Y. Rybarczyk and R. Zalakeviciute. Machine learning approaches for outdoor air quality modelling:
658 A systematic review. *Applied Sciences*, 8(12):2570, 2018.
- 659 L. Schlosser, T. Hothorn, R. Stauffer, A. Zeileis, et al. Distributional regression forests for prob-
660 abilistic precipitation forecasting in complex terrain. *The Annals of Applied Statistics*, 13(3):
661 1564–1589, 2019.
- 662 G. Shaddick, M. L. Thomas, H. Amini, D. Broday, A. Cohen, J. Frostad, A. Green, S. Gumy, Y. Liu,
663 R. V. Martin, et al. Data integration for the assessment of population exposure to ambient air
664 pollution for global burden of disease assessment. *Environmental science & technology*, 52(16):
665 9069–9078, 2018.
- 666 D. M. Stasinopoulos, R. A. Rigby, et al. Generalized additive models for location scale and shape
667 (gamlss) in r. *Journal of Statistical Software*, 23(7):1–46, 2007.
- 668 M. L. Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business
669 Media, 2012.
- 670 J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing*
671 *letters*, 9(3):293–300, 1999.
- 672 J. Velthoen, C. Dombry, J.-J. Cai, and S. Engelke. Gradient boosting for extreme quantile regression.
673 *arXiv preprint arXiv:2103.00808*, 2021.

- 674 A. M. Vicedo-Cabrera, A. Biggeri, L. Grisotto, F. Barbone, and D. Catelan. A bayesian kriging
675 model for estimating residential exposure to air pollution of children living in a high-risk area in
676 italy. *Geospatial health*, 8(1):87–95, 2013.
- 677 J. A. Villaseñor and E. González-Estrada. A variance ratio test of fit for gamma distributions.
678 *Statistics & Probability Letters*, 96:281–286, 2015.
- 679 S. Wager, T. Hastie, and B. Efron. Confidence intervals for random forests: The jackknife and the
680 infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651, 2014.
- 681 Q. Wang, H. Feng, H. Feng, Y. Yu, J. Li, and E. Ning. The impacts of road traffic on urban air
682 quality in jinan based gwr and remote sensing. *Scientific Reports*, 11(1):1–9, 2021.
- 683 M. T. Young, M. J. Bechle, P. D. Sampson, A. A. Szpiro, J. D. Marshall, L. Sheppard, and J. D.
684 Kaufman. Satellite-based no₂ and model validation in a national prediction model based on
685 universal kriging and land-use regression. *Environmental science & technology*, 50(7):3686–3694,
686 2016.
- 687 C. Yuan. Models and methods for computationally efficient analysis of large spatial and spatio-
688 temporal data. 2011.
- 689 L. Zhai, S. Li, B. Zou, H. Sang, X. Fang, and S. Xu. An improved geographically weighted regression
690 model for pm2. 5 concentration estimation in large areas. *Atmospheric Environment*, 181:145–154,
691 2018.
- 692 Y. Zhan, Y. Luo, X. Deng, K. Zhang, M. Zhang, M. L. Grieneisen, and B. Di. Satellite-based
693 estimates of daily NO₂ exposure in China using hybrid random forest and spatiotemporal kriging
694 model. *Environmental science & technology*, 52(7):4180–4189, 2018.

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

698 List of Tables

699 1	Geospatial predictors considered in this study. ”_mon” indicates months (mon = 1, 700 2,...,12). ”_buf” indicates buffer radius in meters. The road length and industrial 701 areas are calculated with buffer radii of 100 m, 300 m, 500 m, 800 m, 1000 m, 3000 702 m and 5000 m. The night lights digital numbers are calculated with buffer radii of 703 450 m, 900 m, 3150 m and 4950 m. The original resolution is provided for gridded 704 variables and data types for vector variables.	41
705 2	Frequency (number of times) of variables selected by Lasso in 20 times bootstrapping 706 and variables that are selected more than 90% times (i.e. 18) are listed below. These 707 variables are considered in INLA besides road_class_3_3000.	42
708 3	Prediction accuracy matrix for different models using 20 times bootstrapped cross- 709 validation. Non-spatial models: LA: Lasso; RF: random forest, XGB: XGBoost using 710 the default Gaussian loss; XGB-G: XGBoost using a Gamma loss; QRFLA: quantile 711 random forest with Lasso for shrinkage aggregation of regression trees; SE: stacked 712 ensembling. Spatial models: INLA: a latent Gaussian model implemented using INLA 713 assuming a Gaussian likelihood. INLA-G: a latent Gaussian model implemented using 714 INLA assuming a Gamma likelihood. SE-INLA, geostatistical stacked ensembling. .	43
715 4	Results with customised CV. tr-hp: close to traffic and high population, tr-lmp: close 716 to traffic and middle and low population, far: far away from traffic. RRMSE (relative 717 RMSE), RMAE (relative MAE), RIQR (relative IQR).	44

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

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

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

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