A comparison of INLA and machine learning-based methods in NO_2 modelling: prediction accuracy, uncertainty quantification, and interpretability

Meng Lu, Paula Moraga, Joaquin Cavieres

September 2020

Abstract

Ensemble tree-based statistical learning methods have been applied in NO_2 mapping and compared between each other and with data model methods including linear regression and Gaussian process models. However, the comparison is mainly based on prediction accuracy. In this study, we went a step further in systematically comparing between uncertainty quantification and latent space interpretability across algorithmic vs data models [2]. In this study, we focus on prediction intervals calculated using INLA [1] for latent Gaussian models (LGMs) and Random Forest models and their model interpretations. The findings are extensible to boosting ensemble-tree methods. Further, we compare a linear LGM, Lasso, Random Forest, and XGboost, and evaluated stack learning methods, with and without modeling spatial variations, in mapping NO_2 at the European country level, with several spatial validation methods. We used national ground station measurements of NO_2 in Germany and Netherlands, of the year 2017, and the NO_2 prediction grid is at 100 m by 100 m resolution.

1 Introduction

Ensemble tree-based statistical learning methods (e.g. random forest, boosting) have been introduced in statistical air pollution modelling and have been compared in [3, 4, 8]. However, in all the studies, comparisons are only based on the cross-validation accuracy of the prediction mean, ignoring the prediction distributions for uncertainty quantification. Also not discussed is the modeling of the spatial effects and error sources analysis (i.e., if the errors are introduced by missing co-variates or the improper specification of the distribution of priors, or the model structure, e.g. using a linear model to model non-linear relationships). Moreover, the accuracy assessment in all these studies are based on random bootstrapped ground station measurements, which do not provide an

error indication of the spatial areas where ground stations are not present for validation. Consequently, current model comparison studies may be one-sided. In this study, our objectives are to:

1. Compare machine learning and latent Gaussian models, for the reliability of the prediction intervals, and their potential in providing information for model improvement (indicating missing co-variates).

2. Evaluate stacked learning methods with and without modeling the spatial random effects.

3. Propose a framework for model comparison and spatial validation.

The first goal is achieved by comparing GLMs with three statistical learning methods: Lasso, XGBoost, and Random Forest (RF). The four methods are chosen for their dissimilarity: Lasso is a linear model without accounting for spatial dependency, which is modelled in INLA as spatial random effects. RF and XGBoost are non-linear and are not affected by dependent co-variates, with the later build tree models subsequently over the residuals of previous trees and has multiple routines to penalise model over-fittng, which has been reported in various studies to obtain the highest prediction accuracy. Among the ML methods, only the distribution prediction and interpretability from INLA and RF are compared, as INLA is representative to Gaussian process models and RF an ensemble tree model.

The second goal is achieved by using a Gaussian process to stack machine learners as a super-learner (cite), to compare the modelled spatial random effects with the super-learner that does not model the spatial random effects.

2 Data

We used average NO2 concentration of 2017 from ground stations in Netherlands and Germany.

2.1 Variables used in INLA modeling

The INLA is used in two models: applying to predictor variables as other methods and stacked modeling. Before applying INLA to predictor variables, we used Lasso to reduce the number of variables. The Lasso is used in contrast to ensemble tree-based methods as they are both linear models. We bootstrapped data 20 times, and used the variables that are selected more than 10 times to consider. The frequency that the Lasso selected a certain variable is shown in table 1.

Table 1: Variables selected by Lasso, frequency indicates the number of times the variables are selected in 20 times boot-strapping. (we can only select variables that appear more than 5 times to consider).

	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
10	road_class_3_100	14
11	road_class_3_5000	6
12	road_class_1_300	5
13	road_class_1_500	5
14	road_class_2_1000	2
15	nightlight_3150	1
16	road_class_2_300	1
17	road_class_3_1000	1
18	temperature_2m_7	1

3 Methods

3.1 Spatial modeling

3.1.1 Spatial random field

A spatial random field is a stochastic spatial process generally defined by $\{X(\boldsymbol{s}) : \boldsymbol{s} \in D \subset \mathcal{R}^d\}$, where $\boldsymbol{s}$ is the location in the space (e.g. latitude-longitude pairs) for the spatial process $X(\boldsymbol{s})$ and D is the spatial domain. A spatial random field is a Gaussian random field if $\{X(\boldsymbol{s}_1), ..., X(\boldsymbol{s}_n)\} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma})$, where N_n is a Normal multivariate distribution for the spatial process and is completely specified by it's mean $\mu = E(X(\boldsymbol{s}))$, and the covariance function $C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \text{Cov}(X(\boldsymbol{s}_1), X(\boldsymbol{s}_2))$. The Gaussian random field can be stationary and isotropic, where the covariance function depend only on the distance and not direction between points, that is $C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \text{Cov}(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|)$ and it's dependence is commonly modeled by a Matérn function ([16][17]). Since incorporating the spatial dependence directly with a large number of observations using a Gaussian random field is computationally expensive, [12] proposed the approximation of a Gaussian random field by a Gaussian Markov random field in order to do more efficient the computational process of estimation. The main property of the Gaussian Markov random field is that it uses a conditional dependency structure through the precision matrix $\boldsymbol{Q}$.

3.1.2 The SPDE (Stochastic partial differential equation) method

To modeling data indexed in space, [7] proposed a new methodology based mainly on the approximation of the Gaussian random field with the Matérn function using the Stochastic Partial Differential Equations (SPDE) as follow:

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau(\boldsymbol{s})x(\boldsymbol{s})) = \boldsymbol{\mathcal{W}(s)}, \tag{1}$$

where κ is a scale parameter, $x(\boldsymbol{s})$ is a spatial random field, Δ is the Laplacian, α is the parameter that controls the smoothness of the realizations, τ controls the variance and $\boldsymbol{\mathcal{W}(s)}$ is a Gaussian spatial white noise process ([6]). For the above we can use a Gaussian Markov random field that approximates to a Gaussian random field without specifying an explicit covariance structure through the SPDE method. This approximation leads to a decrease in the computational burden from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^{3/2})$.

3.1.3 Bayesian inference

The package `INLA` of the software `R` allow use the Integrated Nested Laplace Approximation (INLA) method to performs direct numerical calculation of the posterior distribution for a Bayesian hierarchical model ([13][10]). If we use $\boldsymbol{x}$ as a latent Gaussian field (a Gaussian Markov random field), $\boldsymbol{\theta}$ a vector of (hyper)parameters and $\boldsymbol{y}$ a vector of observations, assuming independent observations given the vector of the spatial latent field ($\boldsymbol{x}$) and the hyperparameters

($\boldsymbol{\theta}$), the likelihood can be expressed as:

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i \in \mathcal{I}} p(y_i | \eta_i, \boldsymbol{\theta}), \tag{2}$$

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

The main idea is to approximate the posterior density for the posterior of the spatial latent field and the hyperparameters, hence, the marginal densities can be obtained:

$$p(x_i \mid \boldsymbol{y}) = \int p(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}, \tag{3}$$

and

$$p(\boldsymbol{\theta}_j | \boldsymbol{y}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}_{-j}. \tag{4}$$

respectively ([6][5]).

4 Model for the data

The general structure for a Bayesian hierarchical model in INLA is as follows:

$$\begin{aligned}
\boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}) && (5)\\
\boldsymbol{x} \mid \boldsymbol{\theta} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}(\boldsymbol{\theta})^{-1}) && (6)\\
\eta_i &= \sum_j b_{ij} x_j && (7)\\
\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta} &\sim \prod_{i \in \mathcal{I}} p(\boldsymbol{y} \mid \boldsymbol{\eta}, \boldsymbol{\theta}), && (8)
\end{aligned}$$

where $\boldsymbol{\theta}$ is the vector of hyperparameters with $\log(\tau) = \theta_1$ and $\log(\kappa) = \theta_2$, $\boldsymbol{x}$ is the spatial latent field, $\boldsymbol{\eta}$ is the linear predictor with covariates b_{ij} and $\boldsymbol{y}$ is the vector of the response variable $f(\cdot \mid \boldsymbol{x}, \boldsymbol{\theta})$, commonly belonging from the exponential family of distributions.

4.0.1 Machine learning methods

4.0.2 Stacked learning

Stacked learning is trained based on the cross-validation of different learners, in this case the three learners are random forest, XGBoost, and Lasso. Two super learners are compared, one is the native optimisation, and the other using the gaussian process to also model the spatial random effects.

4.1 Prediction intervals

Prediction quantiles and coverage probability are calculated for each method for uncertainty quantification. The coverage probability is calculated as the ratio between the number of predictions within the upper and lower quantile and the total number of predictions. The prediction quantiles for each methods are calculated as following:

- INLA
- Random Forest The predictive quantiles of RF are calculated in two ways, one is uses all the observations in each terminal node and over all the trees as a predicted distribution for each terminal node, known as quantile regression forest [11]. The second method embeds the GAMLSS [15] into random forest, known as distributional forests [14], which for each tree uses standard maximum likelihood to fit distributions and recursively select and split co-variates according to the instability of the gradient of the likelihood at each observation along each co-variate. For interpretability, we compared the impact of a certain co-variate on model predictions, by analysing the coefficients estimated by INLA and Lasso and the tree SHAP [SHapley Additive exPlanations, 9] estimated by RF and XGBoost.

4.2 Spatial validation

Besides 20-times random bootstrapping, We also used four other spatial validation methods to comprehensively evaluate the models at different angles.

1. Probability sampling: higher probability is given in selecting observations that are isolated, to test how good the model is at predicting air pollution over the whole region. This is because with random sampling we would get more points in areas where observations are clustered and may not pick any observation in areas with few observations.
2. Spatial blocked cv: divide the data into spatial blocks, each time use one block for validation and other blocks for predicting.
3. Validation based on customised predictors: based on predictor values, we divided the study area so that we assess the prediction accuracy for certain areas. In this study, three subareas are defined: 1) close to traffic and high population: 2) close to traffic and middle low population. 3) far away from traffic. High population is defined as the variable population of 1000 m buffer that is in the last quartile. Low population is defined as the variable population of 1000 m buffer is below the median. Close to road is defined as:

```
road_class_2_100 > 0 |
```

```
road_class_1_100 > 0 |
road_class_3_100 > quantile(road_class_3_100, .7
    5))
```

Far away from road is defined as:

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

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

4. Validation based on known attribute: we know the air quality station types (traffic, background, industrial) and human settlement types (urban or rural), this allows us to quantify prediction accuracy for each type of air quality stations and separating between urban and rural areas.

5 Results

5.1 Prediction Accuracy

Table 2: Cross-validation results of 20 times boot-strapping.

	LA	RF	XGB	stacked	INLA	stacked INLA
RMSE	7.8	7.5	7.4	7.2	7.5	7.1
MAE	5.9	5.5	5.3	5.2	5.5	5.3
R^2	0.63	0.66	0.67	0.68	0.66	0.69

5.2 Distributional forest vs. quantile regression trees

Both the distributional forest and quantile regression trees reach the coverage probability higher than 0.9, but the distributional forest predict a more realistic prediction quantile, notably, it covers four observations that are not covered by the prediction quantile predicted by the quantile regression forest.

5.3 Interpretability

The INLA coefficients

SHAP variable impact for the RF and XGBoost, the variables are ranked by their variable importance. It can be observed from fig. 2 and fig. 3 that the variable rankings differ but the number of points that have positive or negative impact are similar.

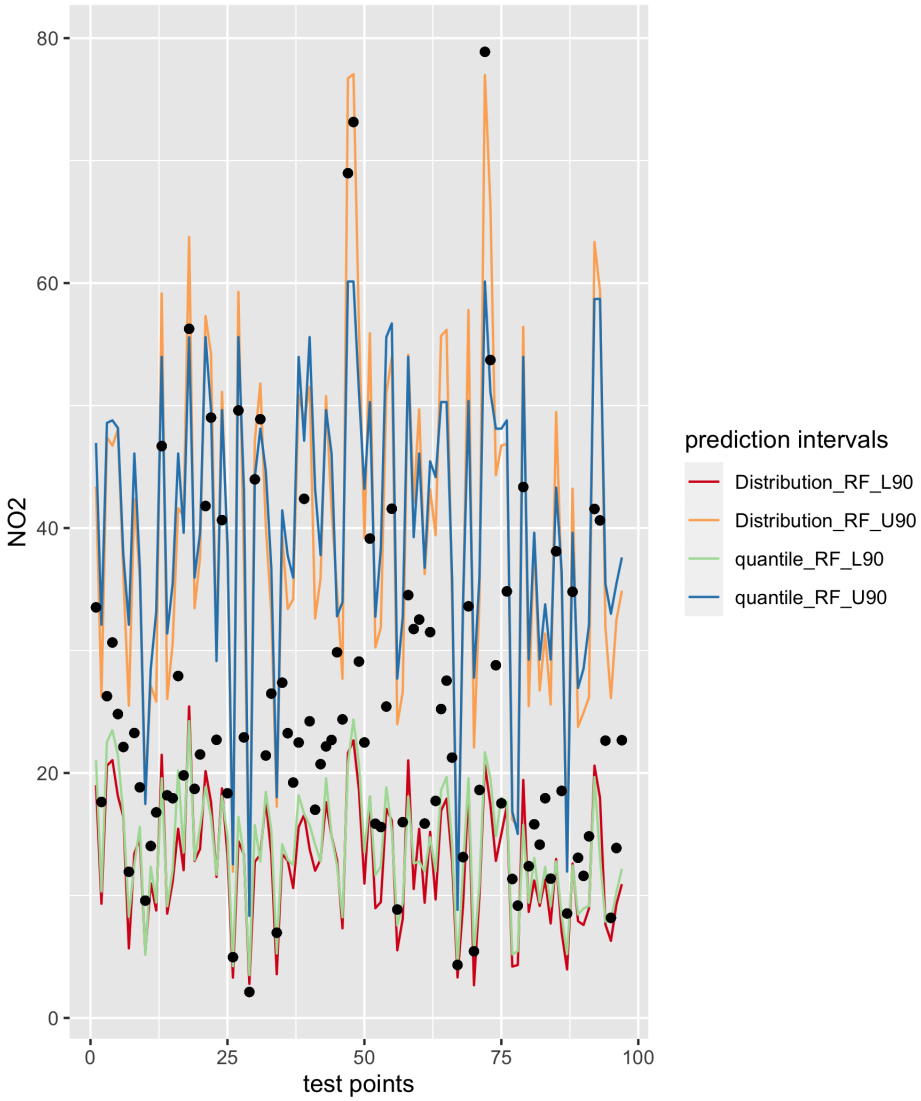

Figure 1: The 90% prediction interval predicted by distributional forest and quantile regression trees. The black dots indicate observations in the test dataset.

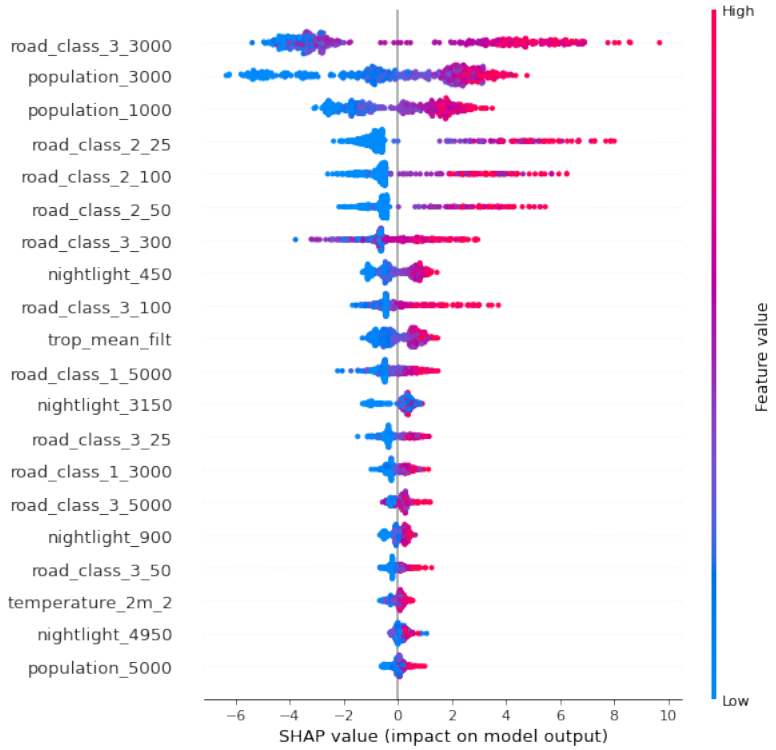

Figure 2: Variable impact calculated by SHAP, the RF model. The horizontal location shows whether the effect of that value is associated with a higher or lower prediction.

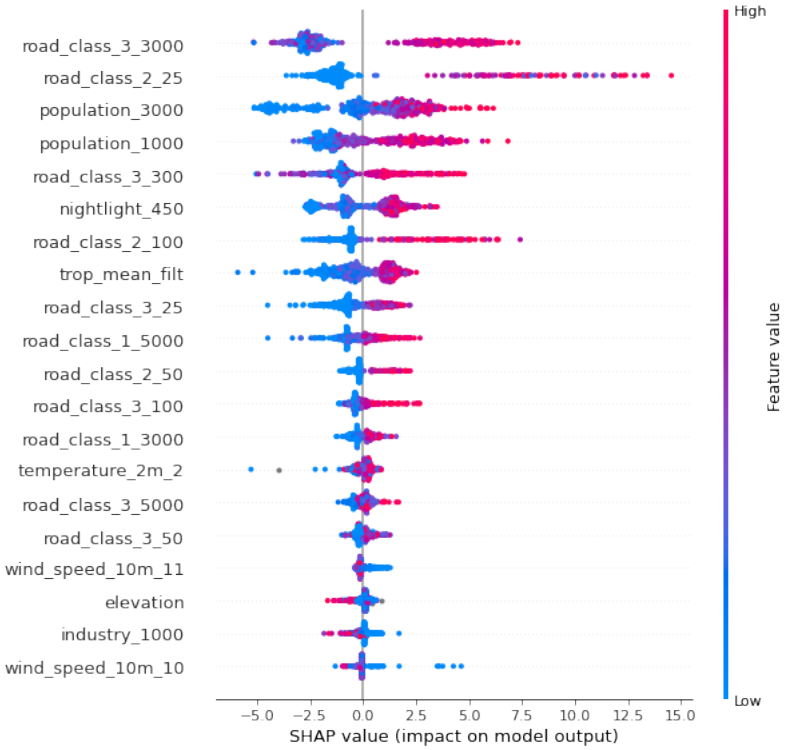

Figure 3: Variable impact calculated by SHAP, the XGBoost model. The horizontal location shows whether the effect of that value is associated with a higher or lower prediction.

6 Discussion

The advantage of INLA.

References

[1] Marta Blangiardo and Michela Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015.

[2] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[3] Jie Chen, Kees de Hoogh, John Gulliver, Barbara Hoffmann, Ole Hertel, Matthias Ketzel, Mariska Bauwelinck, Aaron van Donkelaar, Ulla A Hvidtfeldt, Klea Katsouyanni, et al. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. *Environment international*, 130:104934, 2019.

[4] Jules Kerckhoffs, Gerard Hoek, Lützen Portengen, Bert Brunekreef, and Roel CH Vermeulen. Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. *Environmental science & technology*, 53(3):1413–1421, 2019.

[5] Elias T Krainski, Virgilio Gómez-Rubio, Haakon Bakka, Amanda Lenzi, Daniela Castro-Camilo, Daniel Simpson, Finn Lindgren, and Håvard Rue. *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. CRC Press, 2018.

[6] Finn Lindgren, Havard Rue, et al. Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, 63(19):1–25, 2015.

[7] Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

[8] Meng Lu, Oliver Schmitz, Kees de Hoogh, Qin Kai, and Derek Karssenberg. Evaluation of different methods and data sources to optimise modelling of no2 at a global scale. *Environment international*, 142:105856, September 2020.

[9] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749, 2018.

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

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

[12] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005.

[13] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

[14] Lisa Schlosser, Torsten Hothorn, Reto Stauffer, Achim Zeileis, et al. Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *The Annals of Applied Statistics*, 13(3):1564–1589, 2019.

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

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

[17] Chengwei Yuan. Models and methods for computationally efficient analysis of large spatial and spatio-temporal data. 2011.