Artificial Intelligence in Geosciences 2 (2021) 115–127

Contents lists available at ScienceDirect

Artificial Intelligence in Geosciences

journal homepage: www.keaipublishing.com/en/journals/artificial-intelligence-in-geosciences

# Random forest for spatial prediction of censored response variables

Francky Fouedjio

*Kaplan Business School Pty Ltd, Perth Campus, 1325 Hay St, West Perth, WA, 6005, Australia*

ARTICLE INFO

*Keywords:*
Censored observations
Exact observations
Principal component analysis
Quadratic programming
Spatial prediction

ABSTRACT

The spatial prediction of a continuous response variable when spatially exhaustive predictor variables are available within the region under study has become ubiquitous in many geoscience fields. The response variable is often subject to detection limits due to limitations of the measuring instrument or the sampling protocol used. Consequently, the response variable's observations are censored (left-censored, right-censored, or interval-censored). Machine learning methods dedicated to the spatial prediction of uncensored response variables can not explicitly account for the response variable's censored observations. In such cases, they are routinely applied through ad hoc approaches such as ignoring the response variable's censored observations or replacing them with arbitrary values. Therefore, the response variable's spatial prediction may be inaccurate and sensitive to the assumptions and approximations involved in those arbitrary choices. This paper introduces a random forest-based machine learning method for spatially predicting a censored response variable, in which the response variable's censored observations are explicitly taken into account. The basic idea consists of building an ensemble of regression tree predictors by training the classical regression random forest on the subset of data containing only the response variable's uncensored observations. Then, the principal component analysis applied to this ensemble allows translating the response variable's observations (uncensored and censored) into a linear equalities and inequalities system. This system of linear equalities and inequalities is solved through randomized quadratic programming, which allows obtaining an ensemble of reconstructed regression tree predictors that exactly honor the response variable's observations (uncensored and censored). The response variable's spatial prediction is then obtained by averaging this latter ensemble. The effectiveness of the proposed machine learning method is illustrated on simulated data for which ground truth is available and showcased on real-world data, including geochemical data. The results suggest that the proposed machine learning technique allows greater utilization of the response variable's censored observations than ad hoc methods.

## 1. Introduction

With the increasing development of geoscience data collection platforms, the spatial prediction of a continuous response variable using predictor variables everywhere available within the study area is arousing much interest in many geoscience disciplines. Machine learning methods are increasingly used for this purpose. Indeed, the number of predictor variables that can help explain the spatial variation in the response variable has grown dramatically, making other methods cumbersome to use. Kirkwood et al., 2016a, 2022, Taghizadeh-Mehrjardi et al. (2016), Ballabio et al. (2016), Barzegar et al. (2016), Khan et al. (2016), Wilford et al. (2016), Hengl et al. (2015), Appelhans et al. (2015), Li (2013), Li et al. (2011) demonstrated the relevance of machine learning methods (e.g., random forest, support vector machines, and, neural networks) for spatial prediction in geoscience applications (e.g., geochemical mapping, soil mapping, hydrological mapping, and environmental mapping). Talebi et al. (2021), Sekulić et al. (2020), Hengl et al. (2018) developed machine learning approaches for spatial prediction, in which the spatial correlation is accounted. This latter plays a crucial role in the realm of geoscience data. Fouedjio (2020) introduced a machine learning technique for spatial prediction, where the response variable is exactly conditioned to data.

In many geoscience applications, the response variable's observations often arise below or above the instrument's detection limit (DL). These observations are referred to as censored observations (left-censored, right-censored, or interval-censored). Censoring refers to a condition in which the value of a measurement or observation is only partially known. Left censoring denotes that an observation is below a certain value but it is unknown by how much. Right censoring means that an observation is above a certain value but it is unknown by how much. Interval censoring indicates that an observation is somewhere on an interval between two values. Censored data are a well-known problem when dealing, for

*E-mail address:* francky.fouedjio@kbs.edu.au.

https://doi.org/10.1016/j.aiig.2022.02.001
Received 31 December 2021; Received in revised form 13 February 2022; Accepted 13 February 2022
Available online 23 February 2022

example, with geochemical data (Sanford et al., 1993). Many analytical results for some geochemical elements are reported under (or above) the detection limit (DL). That is, concentrations for some samples are reported as "less than" or "greater than" a specific value. Censoring typically results when analytical methods are not sensitive enough to detect small quantities of an element or when the technique is so sensitive that large concentrations overwhelm the detection system. Censored observations create difficulties for classical machine learning methods for spatial prediction as these latter require a complete set of uncensored observations. Approaches to tackle this problem have been relatively ad hoc.

The deletion and substitution methods are the ad hoc approaches to handle the response variable's censored observations when using traditional machine learning techniques for spatial prediction. The response variable's censored observations are discarded in the deletion procedure, and only uncensored observations are used. The substitution method replaces the response variable's censored observations with arbitrary values. Typically, censored observations are set equal to some constant value. This constant value is some function of the detection limit (e.g., DL, DL/2, 2DL) and depends on the censoring type (left censoring, right censoring, interval censoring). Thus, censored observations are treated as uncensored observations under the substitution method. Traditional machine learning techniques for spatial prediction are usually applied to censored response variables through these ad hoc strategies. Consequently, the response variable's spatial prediction may be imprecise and sensitive to the assumptions and approximations involved in those subjective choices. In particular, the response variable's spatial prediction may be inconsistent with censored observations as the response variable's predicted values at censored sampling locations are not honoring censored observations. In other words, the response variable's predicted values at censored sampling locations could be outside constraint intervals. Although ad hoc methods are easy to implement, they can have difficulties to accurately predict values below (or above) the detection limit (DL).

So far, no alternative methods to the ad hoc techniques have been proposed for spatially predicting a censored response variable when spatially exhaustive predictor variables are available within the study area. There currently exist spatial prediction methods dedicated to censored response variables only in the univariate context, where no predictor (auxiliary) variables are available within the study region. These methods are based among other things on kriging with inequality constraints, data augmentation approaches, and Markov chain Monte Carlo (MCMC) algorithms (Ordoñez et al., 2018; Schelin and Luna, 2014; Toscas, 2010; Fridley and Dixon, 2007; Rathbun, 2006; Abrahamsen and Benth, 2001; De Oliveira, 2005; Militino and Ugarte, 1999; Journel, 1986; Kostov and Dubrule, 1986; Dubrule and Kostov, 1986). This paper presents a machine learning-based method for spatially predicting a censored response variable, in which the response variable's censored observations (left-censored, right-censored, and interval-censored) are explicitly taken into account, that is to say, as they are. Under the proposed machine learning method, the response variable's spatial prediction is carried out such that the response variable's predicted values exactly honor the response variable's observations (uncensored and censored) at sampling locations. The proposed machine learning approach naturally accounts for the response variable's censored observations (inequality data).

The proposed machine learning method starts with constructing an ensemble of regression tree predictors by training the classical regression random forest on the subset of data containing only the response variable's uncensored (exact) observations. Next, the principal component analysis is carried out to create an orthogonal decomposition of the ensemble of regression tree predictors in terms of principal component coefficients and factors. Then, the response variable's observations (uncensored and censored) are converted into a system of linear equalities and inequalities with principal component coefficients as unknown variables. The system of linear equalities and inequalities is then solved through randomized quadratic programming, which allows sampling new principal component coefficients and then reconstructing regression tree predictors that exactly honor the response variable's observations (uncensored and censored) at sampling locations. The response variable's spatial prediction is then obtained by averaging the reconstructed regression tree predictors ensemble. The resulting response variable's spatial prediction effectively honors the response variable's observations (uncensored and censored) at sampling locations. As a byproduct, the response variable's prediction uncertainty is provided. The proposed machine learning method for spatial prediction of a censored response variable is illustrated and compared to ad hoc methods on simulated and real-world data.

The rest of the article is organized as follows. In Sect. 2, different ingredients required to apply the proposed machine learning technique are described. Section 3 demonstrates the proposed machine learning method's effectiveness on simulated and real-world data. A comparison with ad hoc methods is considered. Finally, in Sect. 4 concluding remarks are given.

## 2. Methodology

Let $\{Z(\mathbf{x})\colon \mathbf{x} \in G\}$ represents the continuous response variable defined on a fixed continuous geographical domain $G$ subset of $\mathbb{R}^p (p \geq 1)$. In addition to the response variable, there is a set of $q$ predictor variables $\{f_1(\mathbf{x}), \ldots, f_q(\mathbf{x})\colon \mathbf{x} \in G\}$ exhaustively known in the geographical domain $G$. We consider the situation where the data collection mechanism is such that the response variable $Z$ is not fully quantified due to limitations of the measuring device or the sampling protocol used. For any sampling location, the response variable $Z$ may or may not be fully measured, wherein in the latter case, the response variable is known only up to a set of values. Thus, the response variable's observed data consist of "exact observations" (hard data) measured at some of sampling locations and "interval observations" (inequality data) measured at the other sampling locations as the result of censoring.

The response variable's observed data is denoted by $\{Z(\mathbf{x}_i) \in A_i, i = 1, \ldots, n\}$, with $\{\mathbf{x}_i \in G\}_{i=1,\ldots,n}$ representing sampling locations where exact (uncensored) observations and interval (censored) observations are obtained. The two following cases are of common use: $A_i$ is reduced to a single value $z_i$ (exact or uncensored observations) or $A_i$ is an interval where $Z(\mathbf{x}_i)$ is known to belong (interval or censored observations). Three types of inequality constraints can be considered for the response variable which cover many of the censoring mechanisms encountered in practice. When $A_i$ is an interval, it would be equal to either $(-\infty, u_i]$, $[l_i, +\infty)$, or $[l_i, u_i]$ for, respectively, left, right, and interval censoring; $l_i \in \mathbb{R}$, $u_i \in \mathbb{R}$. $A_i$ can vary with location (multiple censoring).

The goal is to predict the response variable $\{Z(\mathbf{x})\colon \mathbf{x} \in G\}$ over the geographical domain $G$ represented as a grid of $N$ locations, using the response variable's observations (uncensored and censored) and predictor variables data. In addition, the response variable's predicted values must honor the response variable's observations (uncensored and censored) at sampling locations, i.e., $\hat{Z}(\mathbf{x}_i) \in A_i, i = 1, \ldots, n$. The basic ingredients required to implement the proposed machine learning method for spatial prediction are described in this section. The implementation is carried out in the R platform (R Core Team, 2021).

### 2.1. Regression random forest

The starting point of the proposed machine learning method for spatially predicting a censored response variable in the presence of spatially exhaustive predictor variables is the regression random forest (Breiman, 2001). Regression random forest is a type of ensemble machine learning method that constructs a multitude of regression tree models on various subsets of the training dataset (bootstrap samples) using different subsets of available predictor variables, followed by aggregation. Under random forest, each built regression tree model is

unique (less correlated with others) due to the bootstrapping of the training data and the random selection of subsets of predictor variables. The multiple regression tree models knitted together reduce the prediction variance and increase prediction accuracy. The regression random forest's prediction is obtained by averaging all of the regression tree's predictions.

The random forest popularity for spatial prediction relies on its ability to efficiently deal with many predictor variables, handle complex nonlinear relationships and interactions, require less data pre-processing, and be a non-parametric method (model-free). Regression random forest has some tuning parameters that can be optimized via a cross-validation procedure. There are, among others, the number of trees, number of predictor variables randomly selected at each node, proportion of observations to sample in each regression tree, and minimum number of observations in a regression tree's terminal node. It is usually advocated to set the number of trees to a large number, allowing the convergence of the prediction error to a stable minimum (Hengl et al., 2018). The R packages *ranger* (Wright and Ziegler, 2017) and *tuneRanger* (Probst et al., 2018) implement the regression random forest.

The proposed machine learning method for spatial prediction starts by training the classical regression random forest on the subset of data containing only the response variable's uncensored (exact) observations. The outcome is an ensemble of regression tree predictors $\{\tilde{Z}_b(\mathbf{x}) : \mathbf{x} \in G\}_{b=1,\dots,B}$, where $B$ is the number of regression trees. At this stage, the response variable's censored (interval) observations are not yet considered. Also, individual regression tree predictors do not exactly honor the response variable's observed values at uncensored sampling locations. Thus, $\{\tilde{Z}_b(\mathbf{x}) : \mathbf{x} \in G\}_{b=1,\dots,B}$ will be called "unconditional regression tree predictors". The next steps aim to generate conditional regression tree predictors that perfectly honor the response variable's observations at censored and uncensored sampling locations.

### 2.2. Principal component analysis

The second step of the proposed machine learning method consists of performing principal component analysis (PCA) on the ensemble of unconditional regression tree predictors $\{\tilde{Z}_b(\mathbf{x}) : \mathbf{x} \in G\}_{b=1,\dots,B}$ arranged as a matrix $\mathbf{\Gamma}(B \times N)$ whose each row represents a single regression tree predictor $\{\tilde{Z}_b(\mathbf{x}) : \mathbf{x} \in G\}$. We obtain the following decomposition in finite dimensions:

$$\tilde{Z}_b(\mathbf{x}) = \sum_{l=1}^{L} \alpha_{b,l}\psi_l(\mathbf{x}), \ \forall \mathbf{x} \in G, \ b = 1, \dots, B, \tag{1}$$

where $\{\alpha_{b,l}\}_{l=1,\dots,L}$ are principal component (PC) scores (coefficients) and $\{\psi_l(\mathbf{x}) : \mathbf{x} \in G\}_{l=1,\dots,L}$ are principal components (PC) factors (eigen-functions); $L = \min(B, N)$.

Eq. (1) can be interpreted as a decomposition of a set of images $\left\{\tilde{Z}_b(\mathbf{x}) : \mathbf{x} \in G\right\}_{b=1,\dots,B}$ into a set of eigen-images $\{\psi_l(\mathbf{x}) : \mathbf{x} \in G\}_{l=1,\dots,L}$ and coefficients $\{\alpha_{b,l}\}_{l=1,\dots,L}$. The PC factors are considered fixed, while the PC coefficients are considered random. As one can note, PCA is utilized here as an orthogonal decomposition method rather than a dimension reduction technique. All PC factors are kept, as shown in Eq. (1). The bijective property of PCA allows reconstructing regression tree predictors from PC coefficients. In other words, an image can be reconstructed back once all the PC factors and coefficients are used.

### 2.3. Randomized quadratic programming

The third step of the proposed machine learning method consists of generating new principal component (PC) coefficients under the PCA decomposition depicted by Eq. (1) such that regression tree predictors exactly honor the response variable's observations (uncensored and uncensored) at sampling locations. Let

$$Z(\mathbf{x}) = \sum_{l=1}^{L} \theta_l\psi_l(\mathbf{x}), \quad \forall \mathbf{x} \in G, \tag{2}$$

where $\{\theta_l\}_{l=1,\dots,L}$ are random coefficients and $\{\psi_l(\mathbf{x}) : \mathbf{x} \in G\}_{l=1,\dots,L}$ are PC factors derived from the PCA of unconditional regression tree predictors as given in Eq. (1). All PC factors are considered, so there is no truncation.

We want to generate coefficients $\{\theta_l\}_{l=1,\dots,L}$ such that $\{Z(\mathbf{x}) : \mathbf{x} \in G\}$ exactly honors the response variable's observations (uncensored and censored) at sampling locations, i.e., $Z(\mathbf{x}_i) \in A_i, i = 1,\dots,n$. To achieve that, the response variable's observations (uncensored and censored) at sampling locations are translated into a set of equality and inequality constraints using Eq. (2). Hence, the following system of equalities and inequalities:

$$\begin{cases} \theta_1\psi_1(\mathbf{x}_1) + \theta_2\psi_2(\mathbf{x}_1) + \cdots + \theta_L\psi_L(\mathbf{x}_1) \in A_1 \\ \cdots \\ \theta_1\psi_1(\mathbf{x}_n) + \theta_2\psi_2(\mathbf{x}_n) + \cdots + \theta_L\psi_L(\mathbf{x}_n) \in A_n \end{cases}, \tag{3}$$

where $\{\theta_l\}_{l=1,\dots,L}$ are the unknown variables. $A_i (i = 1, \dots, n)$ is equal to either a single value $z_i$ (uncensored observations), or an interval of the type $(-\infty, u_i]$, $[l_i, +\infty)$, or $[l_i, u_i]$ (censored observations). Thus, the system of equalities and inequalities defined in Eq. (3) is induced by the response variable's observations (uncensored and censored). In that way, the response variable's censored observations are naturally taken into account. Conditional PC coefficients $\boldsymbol{\theta} = (\theta_1, \dots, \theta_L)^T$ that exactly honor the response variable's observations (uncensored and censored) are generated by solving the following randomized quadratic optimization problem (Fouedjio et al., 2021a; Fouedjio, 2021):

$$\min_{\boldsymbol{\theta} \in \mathbf{R}^L} \left((\boldsymbol{\theta} - \boldsymbol{\beta})^T \mathbf{\Sigma}^{-1} (\boldsymbol{\theta} - \boldsymbol{\beta})\right) \quad \text{subject to} \quad \{\mathbf{\Psi}_i \boldsymbol{\theta} \in A_i\}_{i=1,\dots,n}; \mathbf{\Psi}_i = [\psi_l(\mathbf{x}_i)]_{l=1,\dots,L}; \tag{4}$$

where $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$. The mean $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{\Sigma}$ of the multivariate normal distribution are computed using unconditional PC coefficients $\{\alpha_{b,l}\}_{l=1,\dots,L}$ derived from the PCA of unconditional regression tree predictors as shown in Eq. (1). Especially,

$$\boldsymbol{\mu} = \left[\frac{1}{B}\sum_{b=1}^{B} \alpha_{b,l}\right]_{l=1,\dots,L} ; \ \mathbf{\Sigma} = \frac{1}{B-1}\sum_{b=1}^{B} (\boldsymbol{\alpha}_b - \boldsymbol{\mu})(\boldsymbol{\alpha}_b - \boldsymbol{\mu})^T, \quad \text{with } \boldsymbol{\alpha}_b = [\alpha_{b,l}]_{l=1,\dots,L}. \tag{5}$$

For each Monte Carlo sample $\boldsymbol{\beta}_t \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}) (t = 1, \dots, T)$, quadratic programming (Goldfarb and Idnani, 1983) is performed to find a solution $\boldsymbol{\theta}_t$ that satisfies the composite constraints (equality and inequality) and minimizes the quadratic objective function defined in Eq. (4). The covariance matrix $\mathbf{\Sigma}$ in Eq. (4) is a diagonal matrix because the PC coefficients are uncorrelated by construction. Conditional PC coefficients $\boldsymbol{\theta}_t$ can be also generated via the Gibbs sampling method (Fouedjio et al., 2021b). However, this approach can be time-consuming for very large datasets since Gibbs sampler generate samples that are highly correlated.

As one can note in Eq. (3), the number of response variable' observations (uncensored and censored) defines the number of equalities and inequalities constraints. The number of unconditional regression tree predictors $B$ should be large enough to allow good coverage of the solution space when solving the system of linear equalities and inequalities defined in Eq. (3). Indeed, the more significant is the number of unconditional regression tree predictors, the wider is the solution space of the system of linear equalities and inequalities defined in Eq. (3). Also, too many composite constraints (hard and inequality data) relative to too few unconditional regression tree predictors will lead to low uncertainty. It is worth mentioning that the number of conditional regression tree predictors $T$ does not depend on the number of unconditional regression tree predictors $B$ under randomized quadratic programming. That is to say, $T$ can be smaller or greater than $B$.

Given conditional PC coefficients $\{\boldsymbol{\theta}_t\}_{t=1,\ldots,T}$, regression tree predictors that exactly honor the response variable's observations (uncensored and censored) at sampling locations are obtained by reconstruction:

$$Z_t(\mathbf{x}) = \sum_{l=1}^{L} \theta_{t,l}\psi_l(\mathbf{x}),\ \forall \mathbf{x} \in G. \tag{6}$$

The prediction of the response variable over the geographical domain $G$ is obtained by averaging the predictions from all the individual reconstructed regression tree predictors:

$$\hat{Z}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} Z_t(\mathbf{x}), \forall \mathbf{x} \in G. \tag{7}$$

It is important to highlight that in Eq. (7), all the individual reconstructed regression tree predictors $\{Z_t(\mathbf{x}) : \mathbf{x} \in G\}_{t=1,\ldots,T}$ exactly honor the response variable's observations (uncensored and censored) at sampling locations. Thus, the mean $\{\hat{Z}(\mathbf{x}) : \mathbf{x} \in G\}$ does also. In addition to providing predictions, the proposed machine learning method naturally delivers a quantification of the uncertainty associated with the prediction as a byproduct. The prediction uncertainty represents the uncertainty around the prediction at a target location, reflecting the inability to exactly define the unknown value. Assessing the uncertainty about the value of the response variable at a target location and of the need to incorporate this assessment in subsequent studies or to support decision making is becoming increasingly crucial (Fouedjio and Klump, 2019; Szatmári and Pásztor, 2019; Veronesi and Schillaci, 2019). Under the proposed machine learning method, an ensemble of conditional regression tree predictors is produced at any target location. Thus, the conditional distribution for the response variable at any target location is available. Hence, predictions using, e.g., the expectation, the mode, or the median can be evaluated and prediction uncertainty using the interquartile range or the variance of the ensemble conditional regression tree predictors can be obtained.

To summarize, the proposed machine learning method for spatially predicting a censored response variable in the presence of spatially exhaustive predictor variables is performed using the following pseudo algorithm:

**Algorithm 1**. Random Forest for Spatial Prediction of Censored Response Variables

**Input**: response variable's observations $\{Z(\mathbf{x}_i) \in A_i, i = 1,\ldots,n\}$ and spatially exhaustive predictor variables $\{\mathrm{f}_1(\mathbf{x}),\ldots,\mathrm{f}_q(\mathbf{x}) : \mathbf{x} \in G\}$; $G$ is represented by a grid of $N$ locations.

*1.* Perform the classical regression random forest on the subset of data containing only the response variable's uncensored observations; the outcome is an ensemble of unconditional regression tree predictors $\{\widetilde{Z}_b(\mathbf{x}) : \mathbf{x} \in G\}_{b=1,\ldots,B}$;

*2.* Apply PCA to the matrix $\boldsymbol{\Gamma}(B \times N)$ arranged as a set of $B$ row vectors, each representing a single unconditional regression tree predictor $\left\{\widetilde{Z}_b(\mathbf{x}) : \mathbf{x} \in G\right\}$; the PCA outcomes are unconditional PC coefficients $\{\alpha_{b,l}\}_{l=1,\ldots,L}$ and PC factors $\{\psi_l(\mathbf{x}) : \mathbf{x} \in G\}_{l=1,\ldots,L}$; $L = \min(B,N)$;

*3.* Generate conditional PC coefficients $\{\theta_{t,l}\}_{l=1,\ldots,L}(t = 1,\ldots,T)$ that exactly honor the response variable's observations (uncensored and censored) using the randomized quadratic programming;

*4.* Derive conditional regression tree predictors by reconstruction $\{\check{Z}_t(\mathbf{x}) = \sum_{l=1}^{L} \theta_{t,l}\psi_l(\mathbf{x})\}_{t=1,\ldots,T}$ and average them to obtain the response variable's prediction $\{\widehat{Z}(\mathbf{x}) : \mathbf{x} \in G\}$;

**Output**: predicted response variable $\{\widehat{Z}(\mathbf{x}) : \mathbf{x} \in G\}$.

**Table 1**
Simulated data example - simulation parameters.

| | Mean | Covariance function | | |
|---|---|---|---|---|
| | | Type | Scale | Sill |
| $f_1(\cdot)$ | 10 | Gaussian | 11.5 | 1 |
| $f_2(\cdot)$ | 10 | Exponential | 6.5 | 1 |
| $f_3(\cdot)$ | 10 | Cardinal Sine | 1.5 | 1 |
| $f_4(\cdot)$ | 10 | Cubic | 20 | 1 |
| $\eta(\cdot)$ | 0 | Spherical | 30 | 100 |

## 3. Application examples

The proposed machine learning method's ability to spatially predict a censored response variable is illustrated using simulated and real-world data. The prediction performance is assessed using some well-known prediction accuracy statistics: mean absolute error (MAE), root mean square error (RMSE), and Lin's concordance correlation coefficient (CCC). The lower are MAE and RMSE, the better is the prediction method. The closer is CCC to 1, the better is the prediction technique. A prediction performance comparison of the proposed method with two ad hoc methods is carried out.

### 3.1. Simulated data example

The data-generating process of the response and predictor variables is given by the following model:

$$Z(\mathbf{x}) = 50\sin(f_1(\mathbf{x})) + 3f_1(\mathbf{x})f_2(\mathbf{x}) + 0.5f_3(\mathbf{x})^2 + 10\sin(f_4(\mathbf{x})) + \eta(\mathbf{x}),\ \forall \mathbf{x} \in [0,100]^2, \tag{8}$$

where $Z(\cdot)$ is the response variable. The predictor variables $f_1(\cdot)$, $f_2(\cdot)$, $f_3(\cdot)$, and $f_4(\cdot)$, and the latent variable $\eta(\cdot)$ are independent Gaussian isotropic stationary random functions (Chiles and Delfiner, 2012) with mean and covariance function specified in Table 1.

The predictor, latent, and response variables are simulated over a 250 $\times$ 250 regular grid in the geographical domain $[0,100]^2$. For background on Gaussian random functions, see Chiles and Delfiner (2012). The

simulation is performing using the turning bands method in the R package *RGeostats* (Renard et al., 2020). This simulated data example for which the ground truth is available everywhere within the study domain refers to a situation where there is a non-linear relationship between the response variable and predictor variables with some interactions between predictor variables. Also, the response variable shows some spatial auto-correlation, and its distribution is non- Gaussian.

Fig. 1 displays the simulated data over a 250 x 250 regular grid (62500 observations). $n = 300$ observations are sampled randomly and taken as the training data (Fig. 1f) as follows. The response variable's observations are divided in three groups: $(-\infty, \lambda]$, $[\gamma, +\infty)$, and $[\lambda, \gamma]$, where $\lambda = 220.86$ and $\gamma = 442.40$ are the respectively, 1st and 99th percentiles. 45 observations are sampled randomly from the group $(-\infty, \lambda]$ and taken as left-censored observations ($Z \leq \lambda$). 45 observations are sampled randomly from the group $[\gamma, +\infty)$ and taken as right-censored observations ($Z \geq \gamma$). 210 observations are sampled randomly from the group $[\lambda, \gamma]$ and considered as uncensored observations. Thus, the proportion of censored data (left-censored and right-censored) is 30% in the training data. The rest of data (62200 observations) is kept aside for the testing.

The ad hoc method 1 discards the response variable's censored observations and considers only uncensored observations. The ad hoc method 2 replaces the response variable's censored observations by the bounds ($\lambda$ and $\gamma$). In ad hoc methods 1 and 2, classical regression random forest is performed with the number of trees equaling 5000. The other hyper-parameters have been optimized through cross-validation. Under

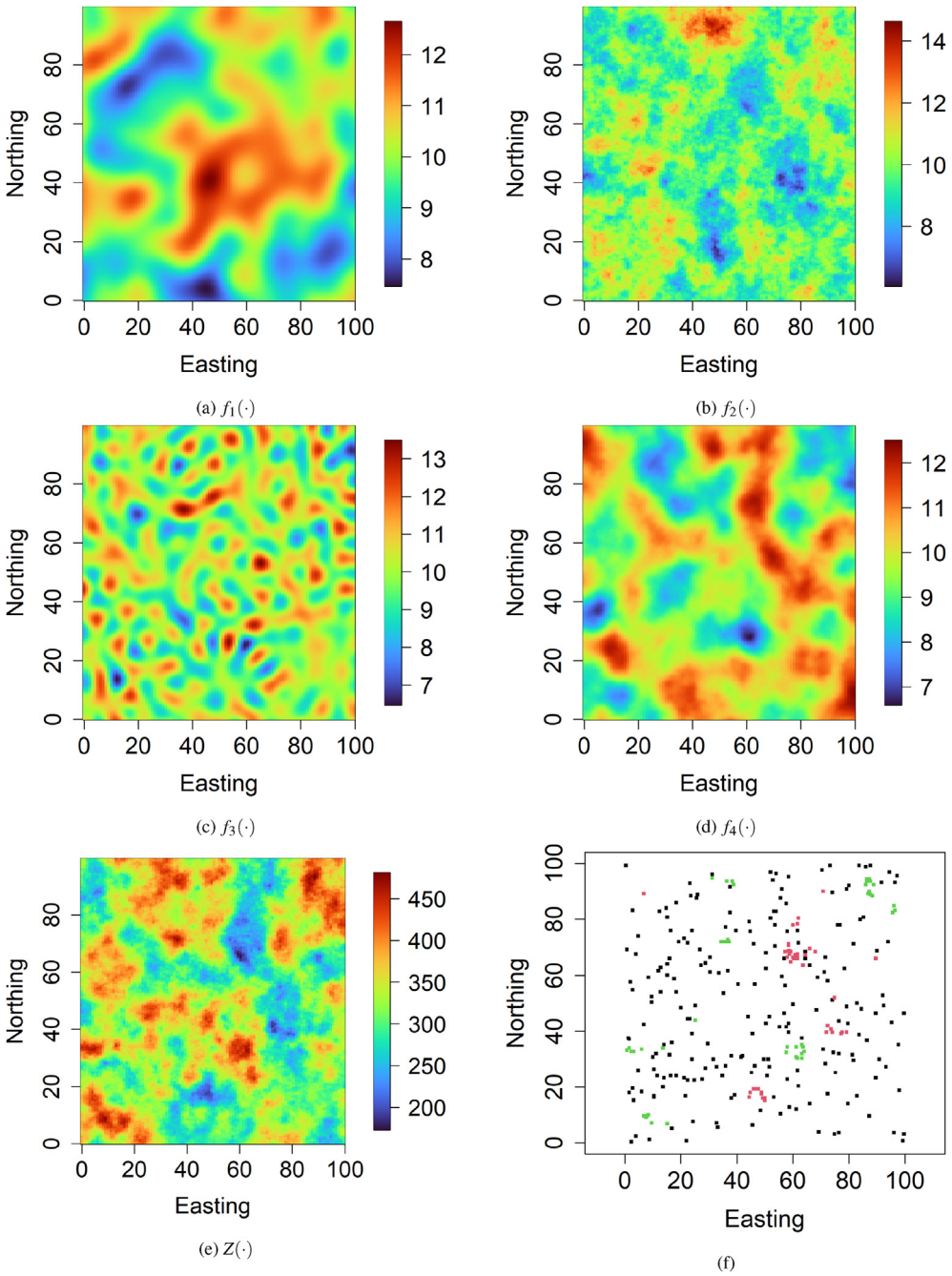


**Fig. 1.** Simulated data example - (a), (b), (c), (d) predictor variables, (e) response variable, and (f) sampling locations. black, red, and green points in (f) represent respectively, uncensored, left-censored, and right-censored sampling locations.

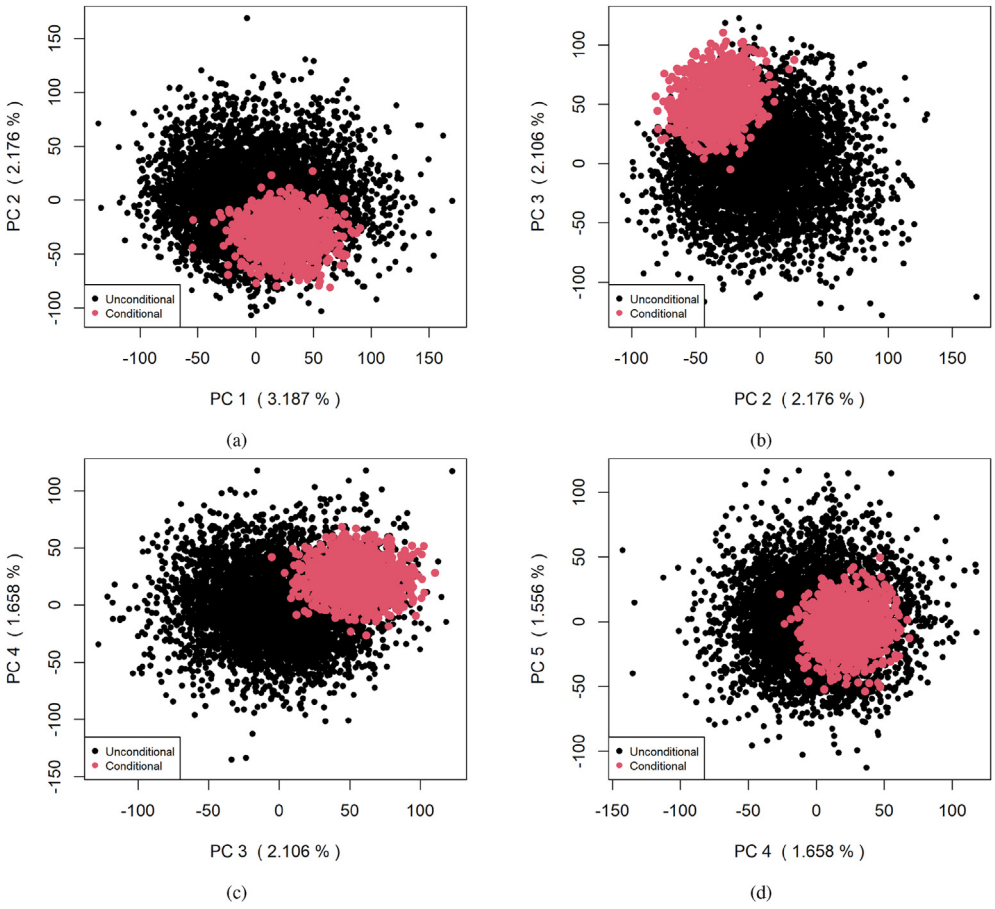


**Fig. 2.** Simulated data example - $B = 5000$ unconditional first four PC scores and $T = 1000$ conditional first four PC scores.

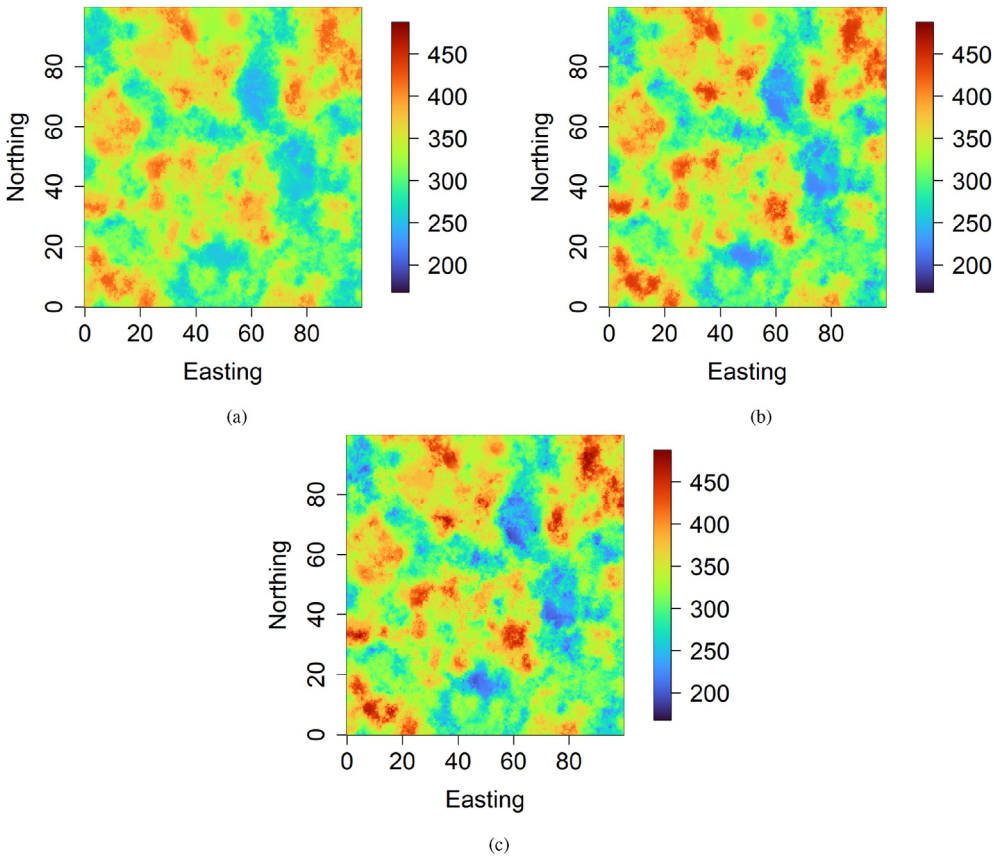


**Fig. 3.** Simulated data example - prediction maps provided by (a) ad hoc method 1, (b) ad hoc method 2, and (c) proposed method.

the proposed machine learning method, the regression random forest model built using only the response variable's uncensored observations consists of an ensemble of $B = 5000$ unconditional regression tree predictors $\{\tilde{Z}_b(\mathbf{x}) : \mathbf{x} \in [0,100]^2\}_{b=1,\ldots,5000}$. According to the methodology described in Sect. 2, PCA is performed on this ensemble, followed by randomized quadratic programming. $T = 1000$ new PC scores are generated, thus giving an ensemble of $T = 1000$ reconstructed (new) regression tree predictors $\{Z_t(\mathbf{x}) : \mathbf{x} \in [0,100]^2\}_{t=1,\ldots,1000}$ that perfectly honor the response variable's observations (uncensored and censored) at sampling locations. The response variable's spatial prediction, defined as the average of reconstructed (new) regression tree predictors, also honors the response variable's observations (uncensored and censored). It is essential to highlight that $B = 5000$ unconditional regression tree predictors are the same as those generated in the ad hoc method 1. Unconditional PC scores $\{\boldsymbol{\alpha}_b\}_{b=1,\ldots,5000}$ and conditional PC scores $\{\boldsymbol{\theta}_t\}_{t=1,\ldots,1000}$ are presented in Fig. 2. The points cloud of conditional PC scores is less scattered than those from unconditional PC scores due effectively to the exact conditioning to the response variable's observations (uncensored and censored).

Fig. 3 presents prediction maps provided by ad hoc methods 1 and 2 and the proposed method. The prediction map resulting from the ad hoc method 1 differs from the two other methods as the former only uses the response variable's uncensored observations. The general appearance of prediction maps resulting from the ad hoc method 2 and proposed method looks similar. However, there are some local differences in areas dominated by censored observations due to the exact conditioning characteristic of the proposed method. The prediction uncertainty (interquartile range) maps for ad hoc methods 1 and 2 and the proposed method are given in Fig. 4. The prediction uncertainty map resulting from the proposed method differs significantly from the others. This is explained by the exact conditioning nature of the proposed method. Under the proposed machine learning method, the response variable's spatial prediction exactly honors the response variable's observations (uncensored and censored) at sampling locations. Consequently, the prediction uncertainty is zero at uncensored sampling locations (exact observations) by construction which is not the case for ad hoc methods. As ad hoc methods can provide the response variable's predicted values that are outside constraint intervals at censoring sampling locations, they tend to overestimate the prediction uncertainty.

Fig. 5 shows the response variable's observed values versus predicted values in the testing data, under ad hoc methods 1 and 2 and the proposed method. One can notice that ad hoc methods have difficulties predicting values below (resp. above) the lower (resp. upper) detection limit, which is not the case for the proposed machine learning method. Fig. 6 provides the histogram of predictions ensemble of the response variable at a training location (left-censored) for ad hoc methods 1 and 2 and the proposed method. One observes that many predictions are greater than the lower detection limit (220.86) under ad hoc methods 1 and 2. In comparison, all predictions are less than the lower detection limit under the proposed method. Fig. 7 depicts the histogram of predictions ensemble of the response variable at a training location (right-censored) for ad hoc methods 1 and 2 and the proposed method. Similarly, ad hoc methods 1 and 2 provide predictions less than the upper detection limit (442.40), while the proposed method offers predictions greater than the upper detection limit. Thus, the proposed machine learning method is more consistent with the response variable's censored observations than ad hoc methods. In Figs. 6 and 7, one can see that the proposed method provides a more reliable confidence interval (prediction uncertainty) than the ad hoc methods.

Table 2 shows the predictive performance of ad hoc methods 1 and 2 and the proposed method on the testing data (62200 observations). The same experiment is repeated for different proportions of censored data (30%, 40%, 50%, 60%, and 70%). One can observe that the proposed

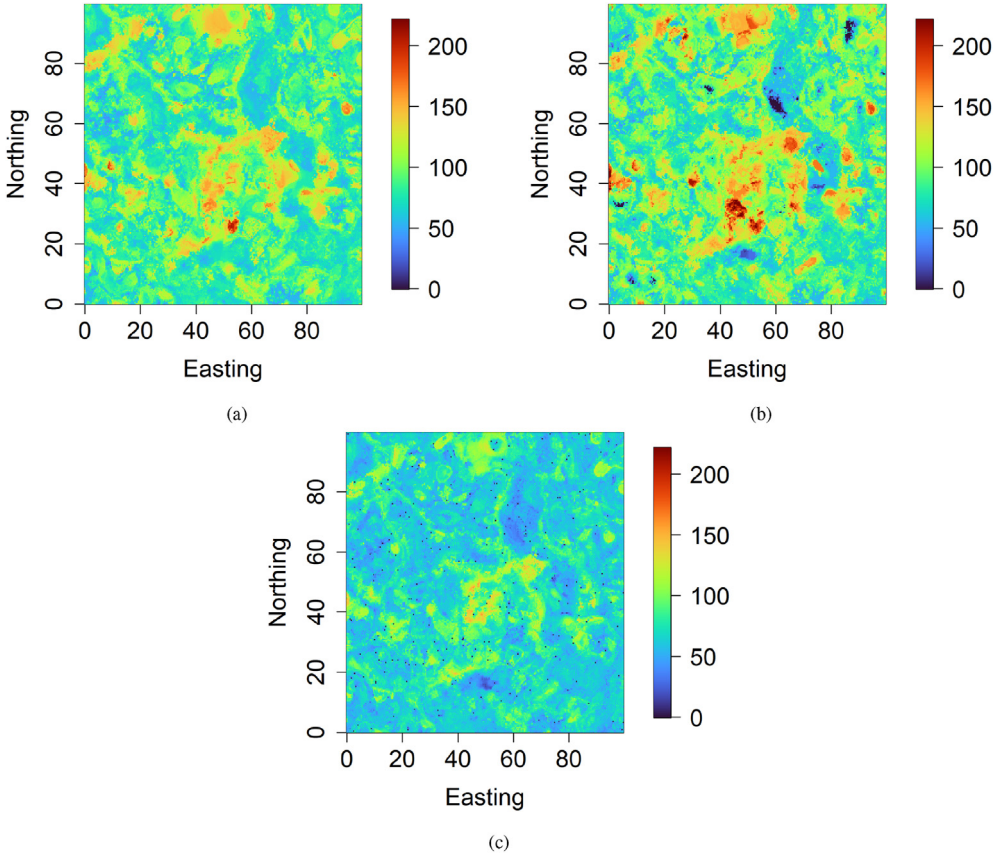


**Fig. 4.** Simulated data example - prediction uncertainty maps provided by (a) ad hoc method 1, (b) ad hoc method 2, and (c) proposed method.

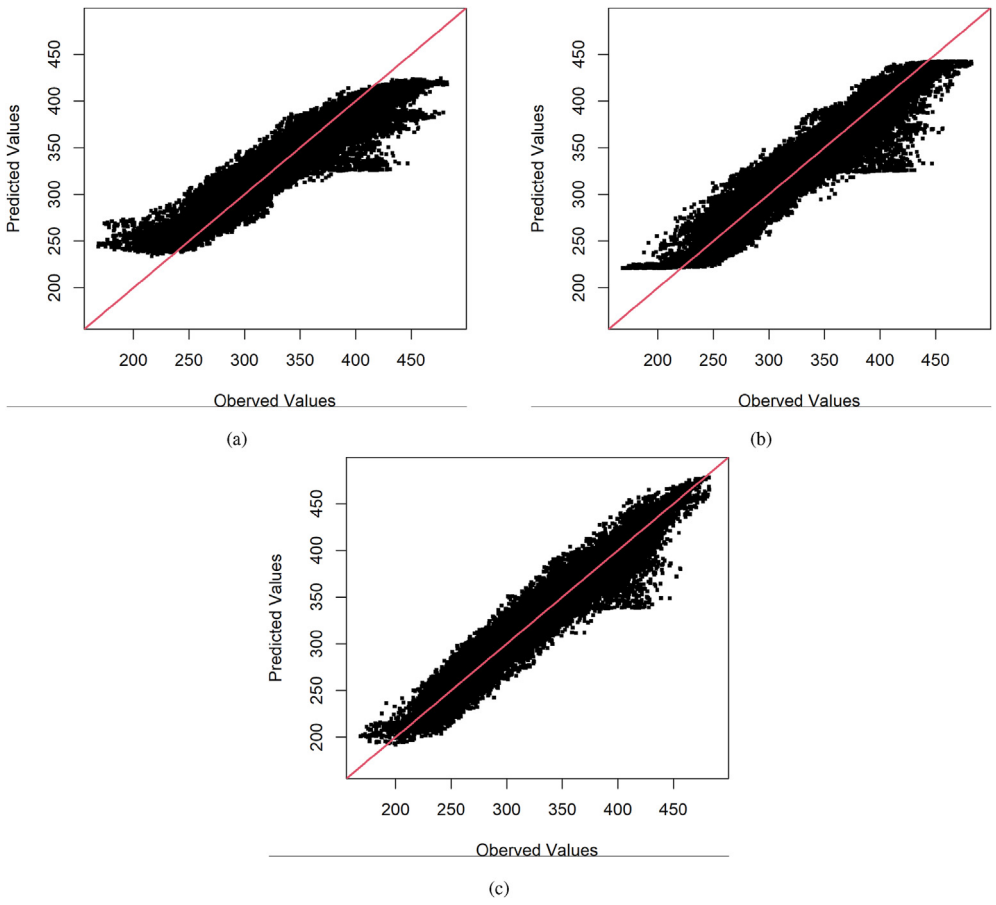


**Fig. 5.** Simulated data example - response variable's observed values vs response variable's predicted values in testing dataset, for (a) ad hoc method 1, (b) ad hoc method 2, and (c) proposed method.

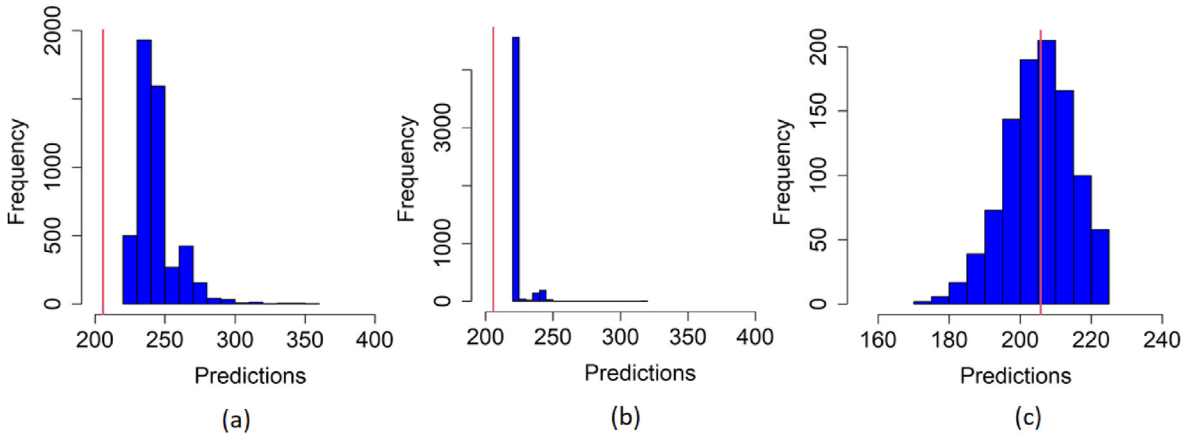


**Fig. 6.** Simulated data example - histogram of predictions ensemble at a training location (left-censored, $Z \leq 220.86$) provided by (a) ad hoc method 1, (b) ad hoc method 2, and (c) proposed method. The red line represents the response variable's true value.

method provides better predictive performance than ad hoc methods. In particular, the ad hoc method 1 is the worst as it only uses a fraction of the data available. Table 3 shows the predictive performance of ad hoc methods 1 and 2 and the proposed method on the testing data for the response values below the lower detection limit (625 observations) and the response values above the upper detection limit (625 observations). One can note the predictive performance between the proposed and ad hoc methods is significantly large. Thus, the proposed method handles better the response variable's censored observations than ad hoc methods.

### 3.2. Geochemical data example

In this application example, the response variable is Scandium (Sc) geochemical concentration observed at 568 sampling locations across the study region in southwest England (Kirkwood et al., 2016b). The response variable is subject to a lower detection limit of 3 mg/kg (left--censoring). The response variable's censored observations represent $\sim$ 6% of total observations. The observations are divided into a training set ($\sim$ 75%) and a testing set ($\sim$ 25%). As we are more interested in the added-value of censored observations, the testing set consists of

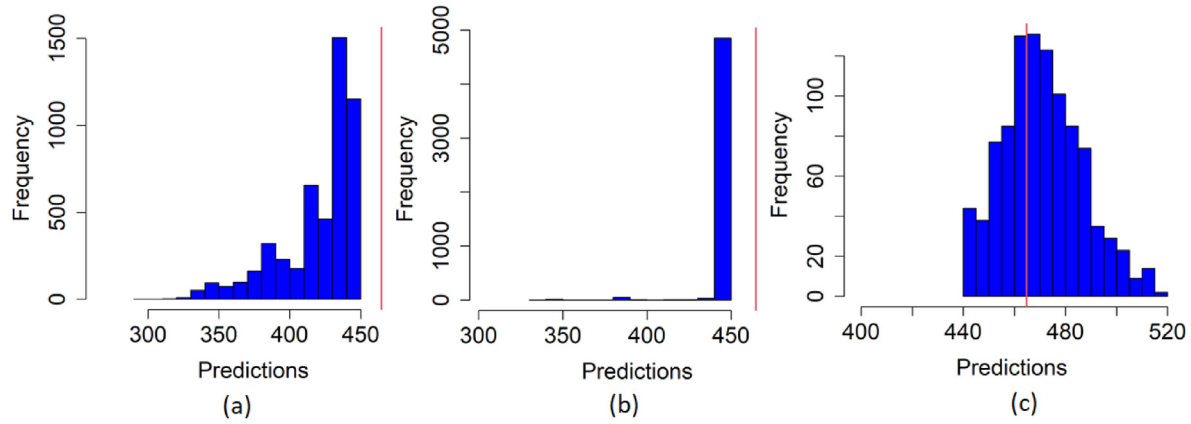


**Fig. 7.** Simulated data example - histogram of predictions ensemble at a training location (right-censored, $Z \geq 442.40$) provided by (a) ad hoc method 1, (b) ad hoc method 2, and (c) proposed method. The red line represents the response variable's true value.

**Table 2**
Simulated data example - predictive performance statistics in the testing dataset containing 62 200 observations, under ad hoc methods 1 and 2, and the proposed method.

| | Proportion of Censored Data | | | | |
|---|---|---|---|---|---|
| | 30% | 40% | 50% | 60% | 70% |
| MAE 1 | 13.15 | 14.15 | 14.29 | 14.45 | 15.03 |
| MAE 2 | 11.27 | 12.13 | 11.97 | 12.67 | 13.58 |
| MAE | 10.32 | 11.11 | 11.36 | 11.61 | 11.32 |
| RMSE 1 | 17.55 | 18.78 | 18.93 | 19.18 | 19.81 |
| RMSE 2 | 14.84 | 16.07 | 15.95 | 16.83 | 17.54 |
| RMSE | 13.33 | 14.42 | 14.72 | 14.94 | 14.42 |
| CCC 1 | 0.924 | 0.911 | 0.909 | 0.908 | 0.901 |
| CCC 2 | 0.952 | 0.944 | 0.945 | 0.944 | 0.939 |
| CCC | 0.963 | 0.957 | 0.956 | 0.957 | 0.960 |

**Table 3**
Simulated data example - predictive performance statistics in the testing dataset containing 625 observations that are below the lower detection limit (220.86) and 625 observations that are above the upper detection limit (442.40), under the ad hoc methods 1 and 2, and the proposed method.

| | Proportion of Censored Data | | | | |
|---|---|---|---|---|---|
| | 30% | 40% | 50% | 60% | 70% |
| MAE 1 | 45.72 | 46.46 | 47.24 | 49.98 | 49.55 |
| MAE 2 | 19.46 | 18.70 | 17.80 | 17.24 | 16.51 |
| MAE | 8.67 | 8.66 | 9.55 | 9.18 | 8.68 |
| RMSE 1 | 47.87 | 48.54 | 49.15 | 51.54 | 51.14 |
| RMSE 2 | 22.97 | 22.52 | 21.86 | 21.89 | 21.22 |
| RMSE | 12.41 | 12.49 | 13.01 | 12.72 | 12.03 |
| CCC 1 | 0.893 | 0.899 | 0.886 | 0.872 | 0.874 |
| CCC 2 | 0.980 | 0.981 | 0.982 | 0.982 | 0.983 |
| CCC | 0.995 | 0.995 | 0.994 | 0.995 | 0.995 |

uncensored observations that are geographically close to the censored observations as shown in Fig. 8a. Fig. 8b provides the spatial plot of the response variable's uncensored observations. Figs. 8c and 8d display respectively, the histogram and the variogram of the response variable's uncensored observations. Predictor variables comprise elevation, gravity, magnetic, Landsat, radiometric, and their derivatives, totaling 26 predictor variables. Some predictor variables are displayed in Fig. 9.

In this application example, the ad hoc method 1 ignores the response variable's censored observations and considers only uncensored observations. The ad hoc method 2 replaces the response variable's censored observations by half the lower detection limit as it is common for geochemical data. In ad hoc methods 1 and 2, classical regression random forest is performed with the number of trees set to 5000. The other hyperparameters have been optimized through cross-validation. The proposed machine learning method generates an ensemble of $B = 5000$ unconditional regression tree predictors, followed by an ensemble of $T = 1000$ conditional regression tree predictors. Unconditional and conditional PC scores are shown in Fig. 10. As mentioned in the simulated data example, the size of the envelop containing conditional PC scores is smaller than the one containing unconditional PC scores because of the exact conditioning to the observations (uncensored and censored). It is important to highlight that $B = 5000$ unconditional regression tree predictors are the same as the ones generated from the ad hoc method 1.

Prediction maps provided by ad hoc methods 1 and 2 and the proposed method are depicted in Fig. 11. The prediction map produced by the ad hoc method 1 differs from the two others, especially in areas dominated by censored observations. The general appearance of prediction maps generated by the ad hoc method 2 and the proposed method looks similar. However, one notes some local differences in regions dominated by censored observations due to the exact conditioning nature of the proposed method. Fig. 12 presents the prediction uncertainty (interquartile range) map under ad hoc methods 1 and 2 and the proposed method. The prediction uncertainty map resulting from the proposed method differs significantly from the others. As highlighted in the simulated data example, this is explained by the exact conditioning property of the proposed method. Under the proposed machine learning method, the response variable's spatial prediction exactly honors the response variable's observations (uncensored and censored) at sampling locations. In contrast, ad hoc methods can provide the response variable's predicted values that are outside constraint intervals at censoring sampling locations. Consequently, they tend to overestimate the prediction uncertainty.

Fig. 13 depicts the histogram of predictions ensemble of the response variable at a training location (left-censored) for ad hoc methods 1 and 2 and the proposed method. One can notice that many predictions are greater than the lower detection limit (3 mg/kg) under ad hoc methods 1 and 2. In contrast, all predictions are smaller than the lower detection limit under the proposed method. Fig. 14 shows the histogram of predictions ensemble of the response variable at one testing location (uncensored) for ad hoc methods 1 and 2 and the proposed method. One can see that the proposed method provides a more reliable confidence interval (prediction uncertainty) than the ad hoc methods.

Table 4 presents the predictive performance of ad hoc methods 1 and 2 and the proposed method on the testing data. The proposed machine learning method shows better predictive performance than the other methods. Thus, the proposed method can exactly honor the response variable's observations (uncensored and censored) at sampling locations while achieving good out-of-sample predictive performance. Even in the case of a small proportion of censored data, the difference can be substantial between the proposed and ad hoc methods.

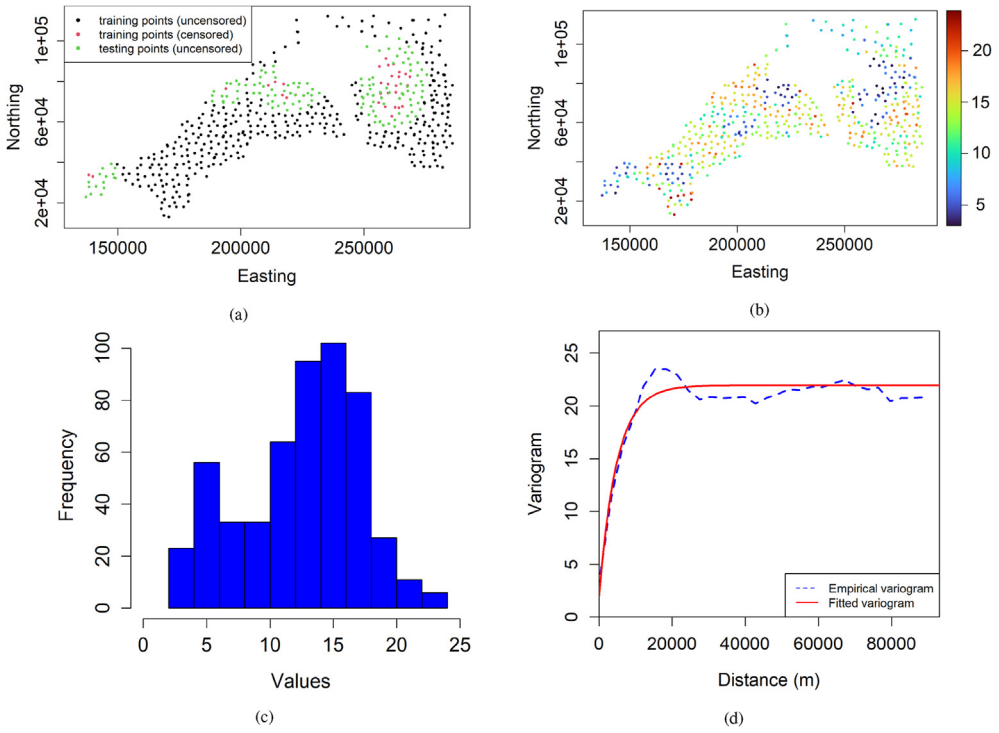


**Fig. 8.** Geochemical data example: (a) uncensored vs censored sampling locations, (b) spatial plot of the response variable's uncensored observations, (c) histogram of the response variable's uncensored observations, (d) variogram of the response variable's uncensored observations.

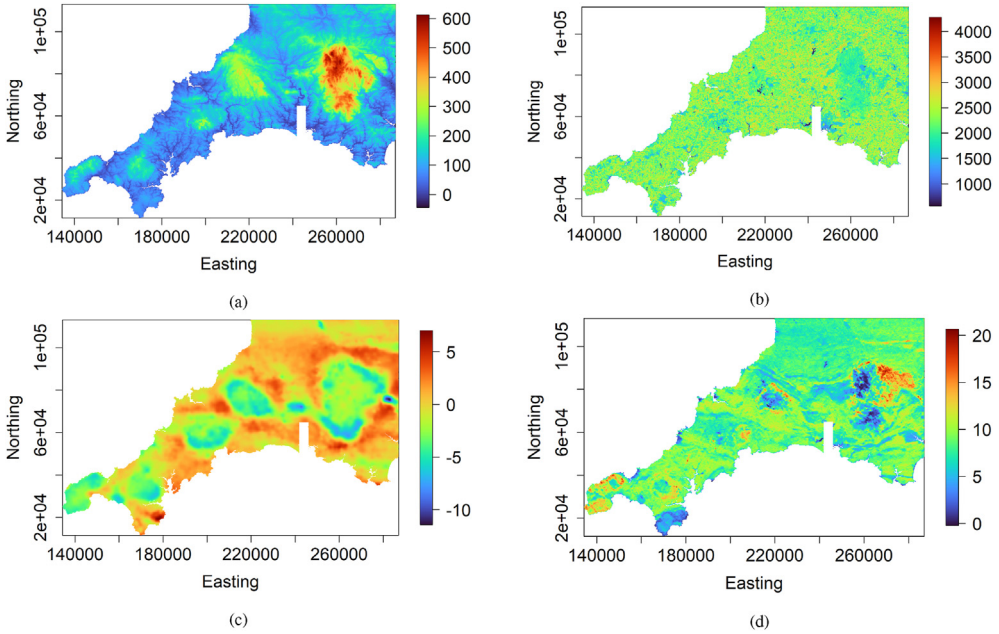


**Fig. 9.** Geochemical data example - some predictor variables: (a) elevation, (b) Landsat 8 band 5, (c) gravity survey high-pass filtered Bouguer anomaly, (d) Thorium counts from gamma ray spectrometry.

## 4. Concluding remarks

This paper presented a random forest-based machine learning method for spatially predicting a censored response variable in which the response variable's censored observations are explicitly taken into account. Under the proposed machine learning method, the response variable's spatial prediction exactly honors the response variable's observations (uncensored and censored) at sampling locations. This is achieved by combining traditional regression random forest, principal component analysis, and randomized quadratic programming. The effectiveness of the proposed machine learning method has been showcased on simulated and real-world data. The proposed method allows better use of censored data than ad hoc methods.

The proposed machine learning method has the advantage of naturally incorporating the response variable's censored observations compared to ad hoc methods (e.g., deletion and substitution methods). There is no substitution, imputation, or discarding of censored observations, as with ad hoc methods. It can perfectly honor the response

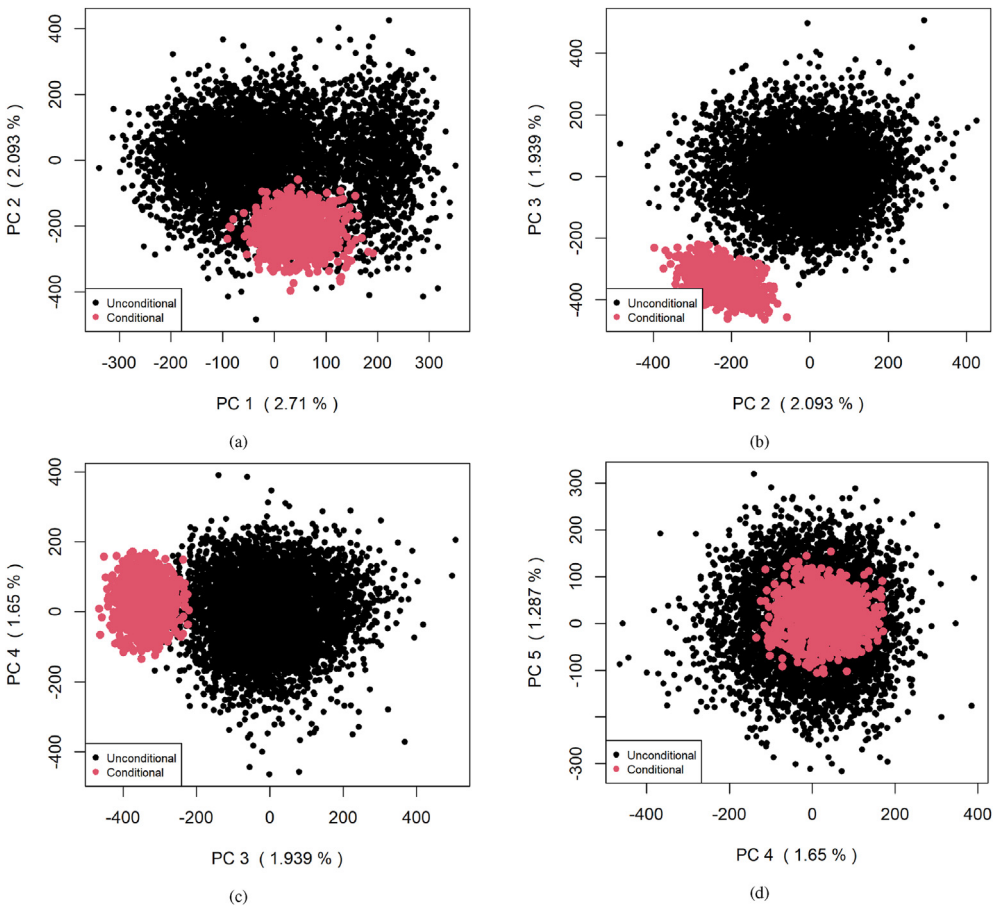


**Fig. 10.** Geochemical data example - unconditional and conditional first four PC scores.

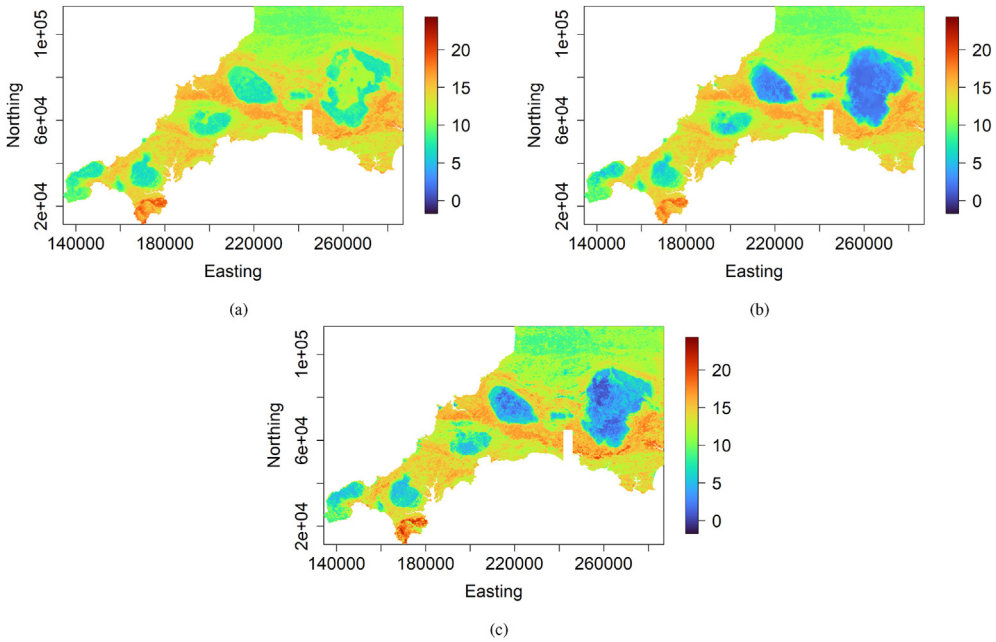


**Fig. 11.** Geochemical data example - prediction maps provided by (a) ad hoc method 1, (b) ad hoc method 2, and (c) proposed method.

variable's observations (uncensored and censored) at sampling locations while achieving good out-of-sample predictive performance compared to ad hoc methods. It also provides realistic prediction uncertainties of the response variable compared to ad hoc techniques. It has the advantage of allowing a fast updating of the response variable predictive map when a few observations (uncensored and censored) are added. Only the last part of the proposed method, i.e., the randomized quadratic programming, should be performed. The proposed machine learning method is easy to implement since it combines well-known existing machine learning, Monte Carlo sampling, and optimization techniques. It handles any

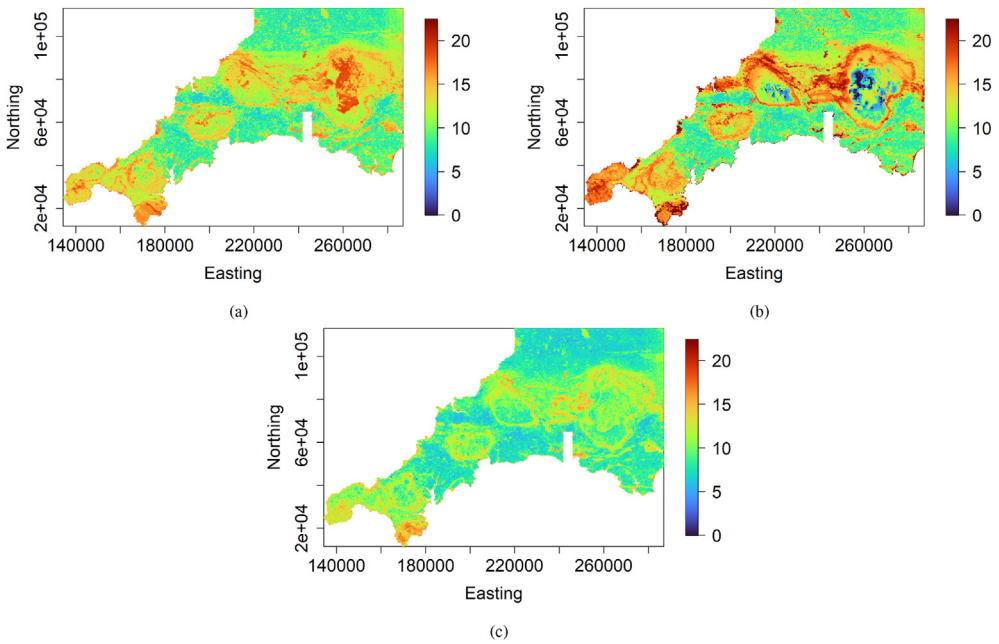


**Fig. 12.** Geochemical data example - prediction uncertainty maps provided by (a) ad hoc method 1, (b) ad hoc method 2, and (c) proposed method.

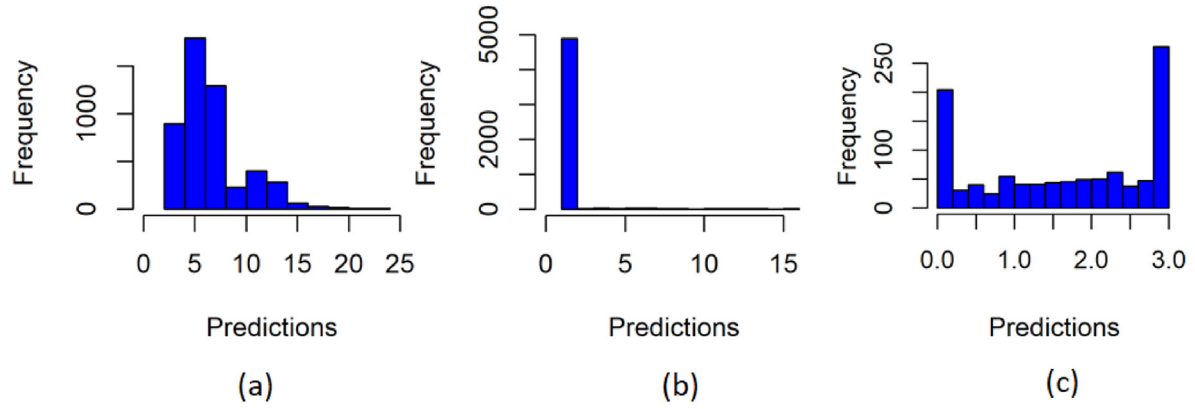


**Fig. 13.** Geochemical data example - histogram of predictions ensemble of the response variable at a training location (left-censored) provided by (a) ad hoc method 1, (b) ad hoc method 2, and (c) proposed method. The lower detection limit of the response variable is equal to 3.

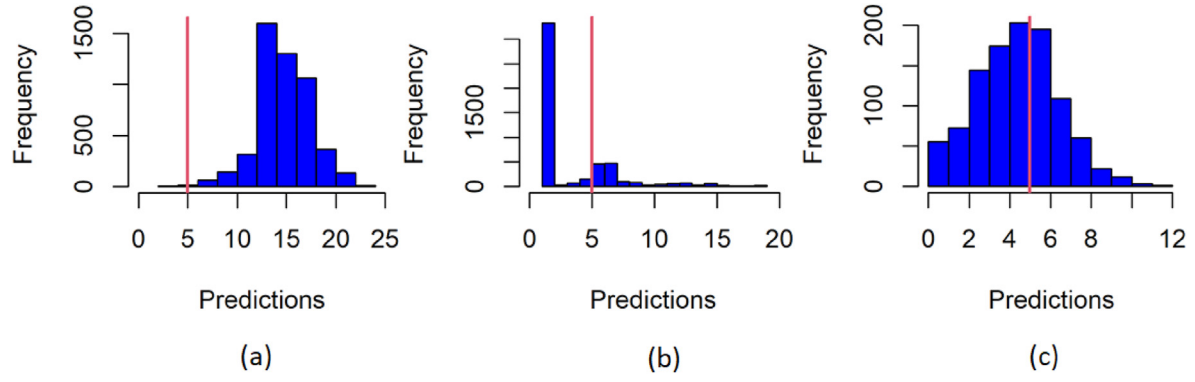


**Fig. 14.** Geochemical data example - histogram of predictions ensemble of the response variable at a testing location (uncensored) provided by (a) ad hoc method 1, (b) ad hoc method 2, and (c) proposed method. The red line represents the response variable' observed value.

censoring type (right censoring, left censoring, and interval censoring observations) and allows multiple censoring.

The proposed machine learning method is built from the regression random forest. The number of unconditional regression tree predictors should be large enough for good coverage of the solution space when performing the exact conditioning of the ensemble of unconditional regression tree predictors to the response variable's observations (uncensored and censored). Indeed, the response variable's observations

**Table 4**
Geochemical data example - predictive performance statistics in the testing dataset containing 147 observations.

| Criteria | Ad hoc Method 1 | Ad hoc Method 2 | Proposed Method |
|---|---|---|---|
| MAE | 3.13 | 2.89 | 2.54 |
| RMSE | 3.91 | 3.97 | 3.49 |
| CCC | 0.65 | 0.73 | 0.78 |

(uncensored and censored) define the number of equalities and inequalities constraints. The more significant is the number of unconditional regression tree predictors, the wider the solution space is. Thus, too many constraints relative to a few unconditional regression tree predictors will lead to too small uncertainty. Generating a large number of unconditional regression tree predictors is not a problem as this parameter is free.

## Conflict of interest

There is no conflict of interest.

## Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

The author is grateful to the anonymous reviewers and the editor for their helpful and constructive comments that helped improve the manuscript.

## References

Abrahamsen, P., Benth, F.E., 2001. Kriging with inequality constraints. Math. Geol. 33, 719–744.
Appelhans, T., Mwangomo, E., Hardy, D.R., Hemp, A., Nauss, T., 2015. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. Spatial Statistics 14, 91–113.
Ballabio, C., Panagos, P., Monatanarella, L., 2016. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma 261, 110–123.
Barzegar, R., Asghari Moghaddam, A., Adamowski, J., Fijani, E., 2016. Comparison of machine learning models for predicting fluoride contamination in groundwater. Stoch. Environ. Res. Risk Assess. 1–14.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Chiles, J.P., Delfiner, P., 2012. Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons.
De Oliveira, V., 2005. Bayesian inference and prediction of Gaussian random fields based on censored data. J. Comput. Graph Stat. 14, 95–115.
Dubrule, O., Kostov, C., 1986. An interpolation method taking into account inequality constraints: I. methodology. Math. Geol. 18, 33–51.
Fouedjio, F., 2020. Exact conditioning of regression random forest for spatial prediction. Artif. Intelligen. Geosci. 1, 11–23.
Fouedjio, F., 2021. Classification random forest with exact conditioning for spatial prediction of categorical variables. Artif. Intelligen. Geosci. 2, 82–93.
Fouedjio, F., Klump, J., 2019. Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. Environ. Earth Sci. 78, 38.
Fouedjio, F., Scheidt, C., Yang, L., Achtziger-Zupančič, P., Caers, J., 2021a. A geostatistical implicit modeling framework for uncertainty quantification of 3d geo-domain boundaries: application to lithological domains from a porphyry copper deposit. Comput. Geosci. 157, 104931.
Fouedjio, F., Scheidt, C., Yang, L., Wang, Y., Caers, J., 2021b. Conditional simulation of categorical spatial variables using Gibbs sampling of a truncated multivariate normal distribution subject to linear inequality constraints. Stoch. Environ. Res. Risk Assess. 35, 457–480.
Fridley, B.L., Dixon, P., 2007. Data augmentation for a Bayesian spatial model involving censored observations. Environmetrics: Off. J. Int. Environ. Soc. 18, 107–123.
Goldfarb, D., Idnani, A., 1983. A numerically stable dual method for solving strictly convex quadratic programs. Math. Program. 27, 1–33.
Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. PLoS One 10, 1–26.
Hengl, T., Nussbaum, M., Wright, M., Heuvelink, G., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518.
Journel, A., 1986. Constrained interpolation and qualitative information—the soft kriging approach. Math. Geol. 18, 269–286.
Khan, S.Z., Suman, S., Pavani, M., Das, S.K., 2016. Prediction of the residual strength of clay using functional networks. Geosci. Front. 7, 67–74.
Kirkwood, C., Cave, M., Beamish, D., Grebby, S., Ferreira, A., 2016a. A machine learning approach to geochemical mapping. J. Geochem. Explor. 167, 49–61.
Kirkwood, C., Everett, P., Ferreira, A., Lister, B., 2016b. Stream sediment geochemistry as a tool for enhancing geological understanding: an overview of new data from south west England. J. Geochem. Explor. 163, 28–40.
Kirkwood, C., Economou, T., Pugeault, N., Odbert, H., 2022. Bayesian deep learning for spatial interpolation in the presence of auxiliary information. Math. Geosci. https://doi.org/10.1007/s11004-021-09988-0.
Kostov, C., Dubrule, O., 1986. Interpolation method taking into account inequality constraints: II. practical approach. Math. Geol. 18, 53–76.
Li, J., 2013. Predictive modelling using random forest and its hybrid methods with geostatistical techniques in marine environmental geosciences. In: 11-th Australasian Data Mining Conference (AusDMÍ3), Canberra, Australia, pp. 73–79.
Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Model. Software 26, 1647–1659.
Militino, A.F., Ugarte, M.D., 1999. Analyzing censored spatial data. Math. Geol. 31, 551–561.
Ordoñez, J.A., Bandyopadhyay, D., Lachos, V.H., Cabral, C.R., 2018. Geostatistical estimation and prediction for censored responses. Spatial Statistics 23, 109–123.
R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.
Probst, P., Wright, M., Boulesteix, A.L., 2018. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Min. Knowl. Discov. https://doi.org/10.1002/widm.1301.
Rathbun, S.L., 2006. Spatial prediction with left-censored observations. J. Agric. Biol. Environ. Stat. 11, 317–336.
Renard, D., Bez, N., Desassis, N., Beucher, H., Ors, F., Freulon, X., 2020. RGeostats: geostatistical package. URL: http://cg.ensmp.fr/rgeostats.r.package.version.12.0.1.
Sanford, R.F., Pierson, C.T., Crovelli, R.A., 1993. An objective replacement method for censored geochemical data. Math. Geol. 25, 59–80.
Schelin, L., Luna, S., 2014. Spatial prediction in the presence of left-censoring. Comput. Stat. Data Anal. 74, 125–141.
Sekulić, A., Kilibarda, M., Heuvelink, G., Nikolić, M., Bajat, B., 2020. Random forest spatial interpolation. Rem. Sens. 12, 1687.
Szatmári, G., Pásztor, L., 2019. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. Geoderma 337, 1329–1340.
Taghizadeh-Mehrjardi, R., Nabiollahi, K., Kerry, R., 2016. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in baneh region, Iran. Geoderma 266, 98–110.
Talebi, H., Peeters, L.J., Otto, A., Tolosana-Delgado, R., 2021. A truly spatial random forests algorithm for geoscience data analysis and modelling. Math. Geosci. 1–22.
Toscas, P.J., 2010. Spatial modelling of left censored water quality data. Environmetrics 21, 632–644.
Veronesi, F., Schillaci, C., 2019. Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. Ecol. Indicat. 101, 1032–1044.
Wilford, J., de Caritat, P., Bui, E., 2016. Predictive geochemical mapping using environmental correlation. Appl. Geochem. 66, 275–288.
Wright, M.N., Ziegler, A., 2017. ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Software 77, 1–17.