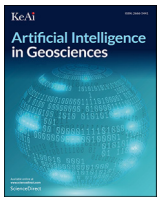
Artificial Intelligence in Geosciences 1 (2020) 11–23

Contents lists available at ScienceDirect

Artificial Intelligence in Geosciences

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



Exact Conditioning of Regression Random Forest for Spatial Prediction

Francky Fouedjio

AngloGold Ashanti Australia Ltd., Growth and Exploration, 140 St. Georges Terrace, Perth, WA, 6000, Australia

ARTICLE INFO

Keywords:
Exact conditioning
Monte Carlo sampling
Multi-Gaussian
Spatial prediction
Principal component analysis
Random forest

ABSTRACT

Regression random forest is becoming a widely-used machine learning technique for spatial prediction that shows competitive prediction performance in various geoscience fields. Like other popular machine learning methods for spatial prediction, regression random forest does not exactly honor the response variable's measured values at sampled locations. However, competitor methods such as regression-kriging perfectly fit the response variable's observed values at sampled locations by construction. Exactly matching the response variable's measured values at sampled locations is often desirable in many geoscience applications. This paper presents a new approach ensuring that regression random forest perfectly matches the response variable's observed values at sampled locations. The main idea consists of using the principal component analysis to create an orthogonal representation of the ensemble of regression tree predictors resulting from the traditional regression random forest. Then, the exact conditioning problem is reformulated as a Bayes-linear-Gauss problem on principal component scores. This problem has an analytical solution making it easy to perform Monte Carlo sampling of new principal component scores and then reconstruct regression tree predictors that perfectly match the response variable's observed values at sampled locations. The reconstructed regression tree predictors' average also precisely matches the response variable's measured values at sampled locations by construction. The proposed method's effectiveness is illustrated on the one hand using a synthetic dataset where the ground-truth is available everywhere within the study region, and on the other hand, using a real dataset comprising southwest England's geochemical concentration data. It is compared with the regression-kriging and the traditional regression random forest. It appears that the proposed method can perfectly fit the response variable's measured values at sampled locations while achieving good out of sample predictive performance comparatively to regression-kriging and traditional regression random forest.

1. Introduction

A common problem in geosciences is predicting over the entire study region a physical quantity of interest measured at a few sampled locations within the study region. The spatial prediction is used for impactful decision-making in many geoscience fields, including geology, geophysics, and geochemistry. There is an increasing interest in using regression random forest for spatial prediction in various geoscience fields, when auxiliary information is available everywhere within the study region. Regression random forest is a machine learning ensemble method based on a collection of randomized regression trees, which are combined through averaging (Breiman, 2001). Its popularity for spatial prediction relies on its ability to efficiently deal with many predictor variables, handle complex non-linear relationships and interactions, and require less data pre-processing. Regression random forest has proven relevant for spatial prediction in several research works, including Fouedjio and Klump (2019), Szatmári and Pásztor (2019), Veronesi and Schillaci (2019), Hengl et al. (2018), Vaysse and Lagacherie (2017), Vermeulen and Niekerk (2017), Barzegar et al. (2017), Kirkwood et al. (2016a), Taghizadeh-Mehrjardi et al. (2016), Ballabio et al. (2016), Khan et al. (2016), Wilford et al. (2016), Hengl et al. (2015), Appelhans et al. (2015), Li (2013), and Li et al. (2011).

Like other popular machine learning techniques for spatial prediction, regression random forest does not perfectly match the response variable's observed values at sampled locations during the training stage. This characteristic is often undesirable in many geoscience applications, including mining and petroleum exploration, where it is of interest for modeling ore bodies and petroleum reservoirs. Competitor methods like regression-kriging perfectly fit the response variable's observed values at sampled locations (Hengl et al., 2004; Chiles and Delfiner, 2012). In this article, a new methodology ensuring the exact conditioning of regression random forest is presented. The proposed method achieves the exact conditioning through a step-by-step approach. First, traditional regression random forest is performed on data as usual, and an ensemble of

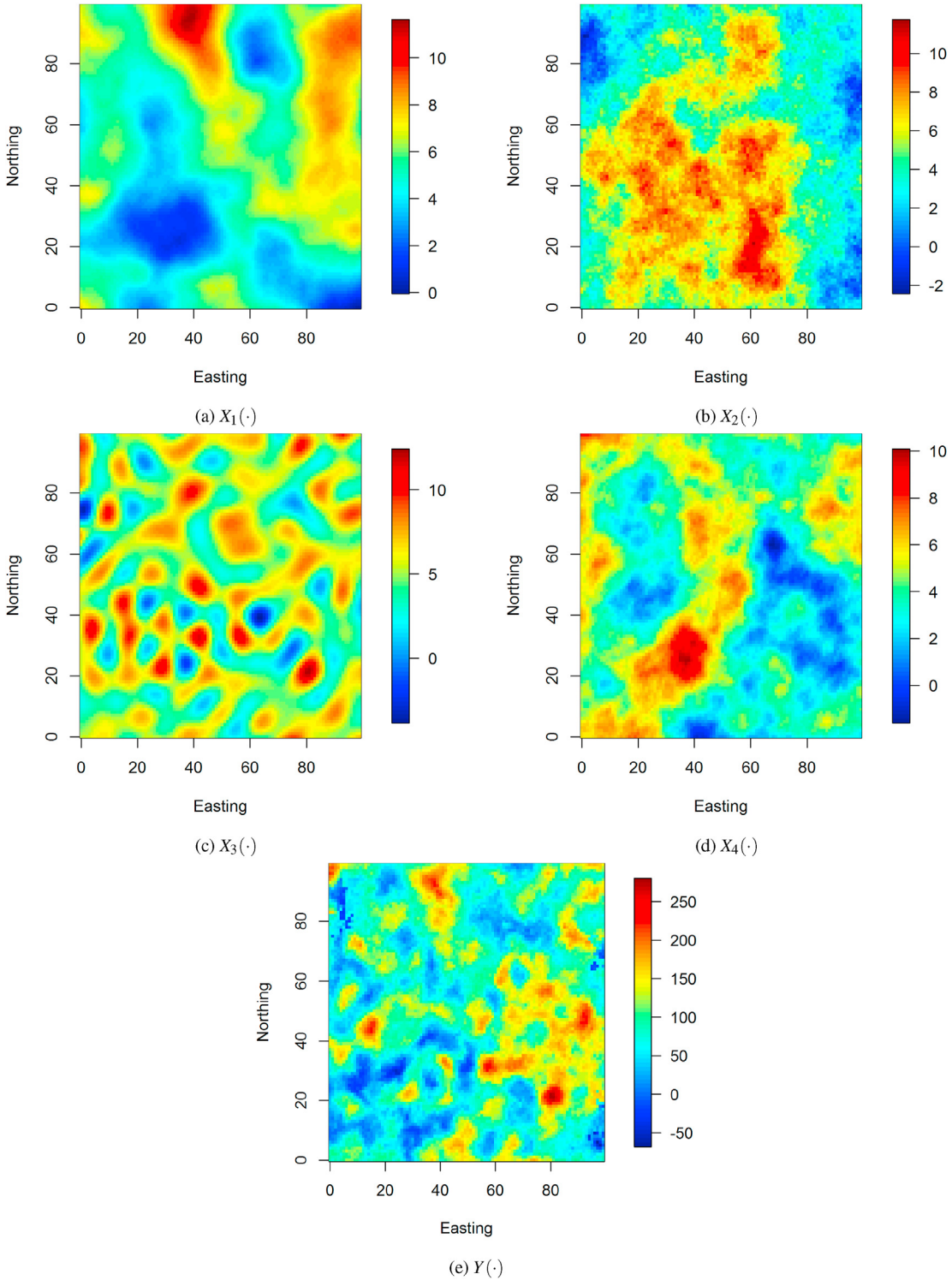
E-mail address: ffouedjio@anglogoldashanti.com.

https://doi.org/10.1016/j.aiig.2021.01.001
Received 18 November 2020; Received in revised form 7 January 2021; Accepted 9 January 2021
Available online 13 January 2021

Table 1
Synthetic data example - simulation parameters.

	Mean	Covariance function		
		Type	Scale	Sill
$X_1(\cdot)$	5	Cubic	40	5
$X_2(\cdot)$	5	Spherical	40	5
$X_3(\cdot)$	5	Cardinal Sine	2	5
$X_4(\cdot)$	5	K-Bessel (shape = 1)	11	5
$\varepsilon(\cdot)$	0	Exponential	6.5	175

regression tree predictors is produced. Second, the principal component analysis (PCA) is carried out on the ensemble of regression tree predictors. Third, the exact conditioning problem is reformulated as a Bayes-linear-Gauss problem on principal component scores. This problem has an analytical solution making it easy to carry out Monte Carlo sampling of new principal component scores and then reconstruct regression tree predictors that perfectly fit the response variable's measured values at sampled locations. A synthetic dataset where the ground-truth is available within the region under study and a real dataset

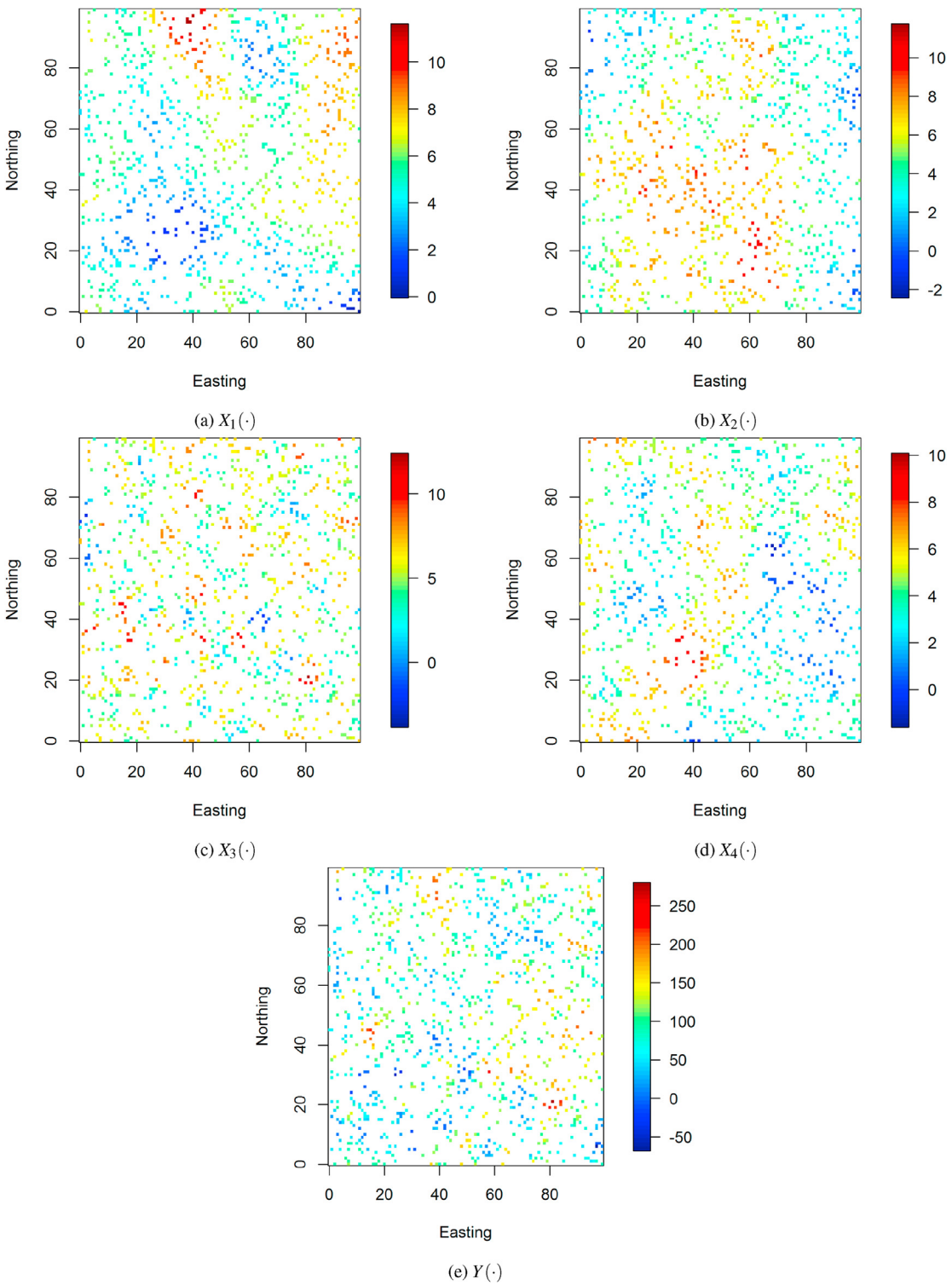


(a) $X_1(\cdot)$ (b) $X_2(\cdot)$

(c) $X_3(\cdot)$ (d) $X_4(\cdot)$

(e) $Y(\cdot)$

Fig. 1. Synthetic data example - (a), (b), (c), (d) explanatory variables, and (e) response variable.



(a) $X_1(\cdot)$ (b) $X_2(\cdot)$

(c) $X_3(\cdot)$ (d) $X_4(\cdot)$

(e) $Y(\cdot)$

Fig. 2. Synthetic data example - training dataset of $n = 1000$ observations.

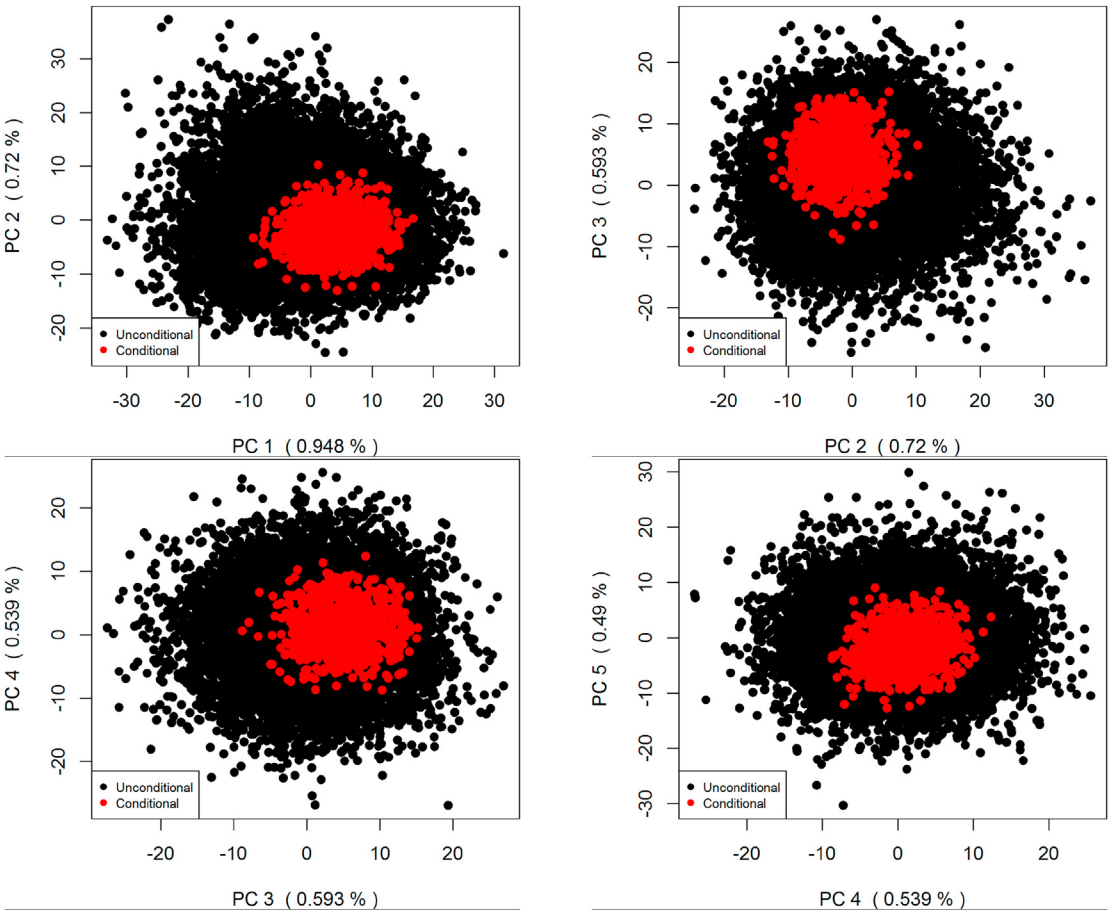


Fig. 3. Synthetic data example - $B = 10000$ unconditional first PC scores and $T = 1000$ conditional first PC scores.

including southwest England's geochemical concentration data are used to illustrate the proposed approach's ability to perfectly fit the response variable's observed values at sampled locations.

The remainder of the article is organized as follows. Section 2 describes the different ingredients required to apply the proposed approach. Section 3 illustrates the proposed method's effectiveness on a synthetic dataset as well as a real dataset. A comparison with the regression-kriging and the traditional regression random forest is carried out. Section 4 offers concluding remarks.

2. Methodology

Let $\{Y(\mathrm{s}) : \mathrm{s} \in D \subset R^d\}$ be a real-valued response variable defined on a spatial domain of interest $D \subset \mathbb{R}^d$. Let $\{\mathbf{x}^1(s), \ldots, \mathbf{x}^p(\mathrm{s}) : \mathrm{s} \in D \subset \mathbb{R}^d\}$ to be the set of predictor variables exhaustively known in the spatial domain D. The following terms are also used to refer to the response variable: target variable, dependent variable, outcome variable, explained variable. A predictor variable is usually also referred as explanatory variable or independent variable or covariate. Let $\{Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)\}$ be the response variable's measured values at sampled locations $\{\mathrm{s}_1, \ldots, \mathrm{s}_n\} \subset D$. The goal is to use the regression random forest for predicting the response variable over the spatial domain D represented by a number of grid locations N such that the response variable's predicted values are the same as the response variable's measured values at sampled locations, i.e., $\widehat{Y}(\mathbf{s}_i) = Y(\mathbf{s}_i),\ i = 1, \ldots, n$. Different ingredients required to implement the proposed method are described in this section. The implementation is carried out in the R platform (R Core Team, 2020).

2.1. Regression random forest

Regression random forest is an ensemble learning approach that creates many regression tree models built from bootstrap samples of the training data (Breiman, 2001). It also injects some randomness into the tree-growing process by randomly selecting only a subset of predictor variables to consider for split-point selection at each node. This operation reduces the chance of the same strong predictor variables to be selected when a split is to be carried out, thus avoiding regression trees from becoming overly correlated. The multiple regression tree predictors are knitted together to reduce the prediction variance and increase prediction accuracy. The method predicts the value that is the mean prediction of all individual regression tree predictors.

Like most machine learning techniques, regression random forest has some free parameters that can be optimized. 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. These free parameters are optimized through cross-validation. In practice, there is no need to tune the number of decision trees; it is usually recommend to set it to a large number, allowing the convergence of the prediction error to a stable minimum (Hengl et al., 2018). The implementation of the regression random forest is done using the R packages *ranger* (Wright and Ziegler, 2017) and *tuneRanger* (Probst et al., 2018).

Let $\{\tilde{Y}_b(\mathbf{s}) : \mathbf{s} \in D\}_{b=1,\ldots,B}$ be the ensemble of regression tree predictors resulting from the training of the traditional regression random forest. At this stage, individual regression tree predictors and their average do not necessarily match the response variable's measured

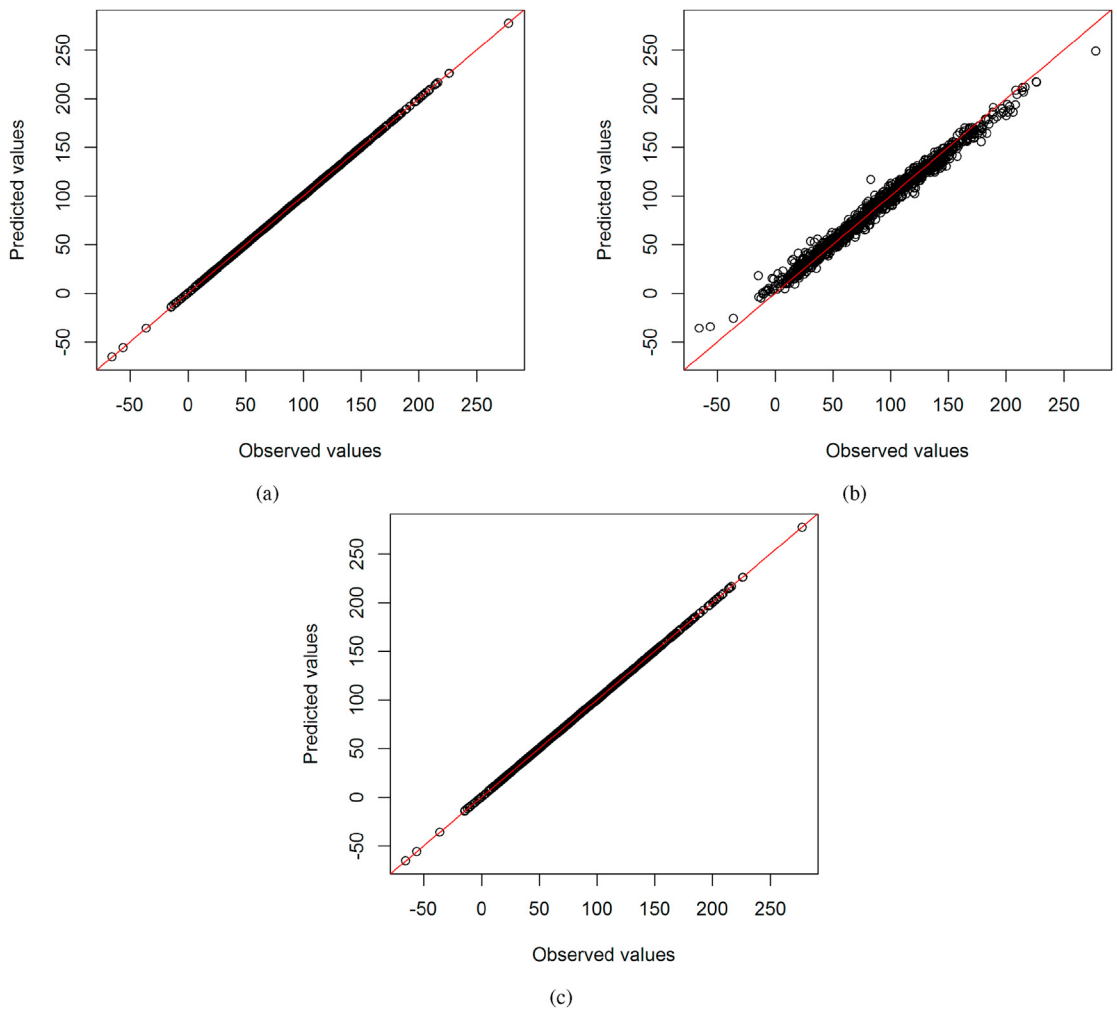


Fig. 4. Synthetic data example - response's observed values vs response's predicted values in the training dataset, for (a) regression-kriging, (b) traditional regression random forest, and (c) proposed regression random forest.

values at sampled locations. The next steps consist of using this ensemble of regression tree predictors $\{\tilde{Y}_b(\mathbf{s}) : \mathbf{s} \in D\}_{b=1,\ldots,B}$ to reconstruct individual regression tree predictors that perfectly match the response variable's observed values at sampled locations. By construction, their average also exactly fits the response variable's observed values at sampled locations.

2.2. Principal component analysis

After training of the traditional regression random forest model, the next step consists of performing principal component analysis on the ensemble of regression tree predictors $\{\tilde{Y}_b(\mathbf{s}) : \mathbf{s} \in D\}_{b=1,\ldots,B}$ arranged as a matrix of size $(B \times N)$ whose each row vector represents a single regression tree predictor $\{\tilde{Y}_b(\mathbf{s}) : \mathbf{s} \in D\}$. This results in the following decomposition in finite dimensions:

$$\tilde{Y}_b(\mathbf{s}) = \sum_{l=1}^{L} \alpha_{b,l}\psi_l(\mathbf{s}),\ \forall \mathbf{s} \in D,\ b = 1, \ldots, B; \tag{1}$$

where $\{\alpha_{b,l}\}_{l=1,\ldots,L}$ are principal component scores (or coefficients) and $\{\psi_l(\mathbf{s}) : \mathbf{s} \in D\}_{l=1,\ldots,L}$ are principal components factors (or eigenfunctions); $L = \min(B, N)$.

$\{\tilde{Y}_b(\mathbf{s}) : \mathbf{s} \in D\}$ can be interpreted as an image and $\{\tilde{Y}_b(\mathbf{s}) : \mathbf{s} \in D\}_{b=1,\ldots,B}$ as an ensemble of images as well. Thus, Eq. (1) provides a decomposition of an ensemble of images into a set of eigenimages $\{\psi_l(\mathbf{s}) : \mathbf{s} \in D\}_{l=1,\ldots,L}$ and a set of coefficients $\{\alpha_{b,l}\}_{l=1,\ldots,L}$. It is essential to highlight that the eigenfunctions are considered fixed in the PCA, while the coefficients are considered random. It is also important to emphasize here that the PCA is used more as an orthogonal decomposition method than a dimension reduction technique since all the eigenfunctions are kept, as shown in Eq. (1). The bijective nature of PCA allows the reconstruction of regression tree predictors from coefficients. In other words, an image can be reconstructed back once all the principal component factors and scores are used.

2.3. Bayes-linear-Gauss problem

Given the PCA decomposition of the ensemble of regression tree predictors as shown the previous section, the next step consists of finding new principal component scores such that the reconstructed regression tree predictors perfectly match the response variable's observed values at sampled locations. Let

$$\breve{Y}(\mathbf{s}) = \sum_{l=1}^{L} \beta_l \psi_l(\mathbf{s}),\ \forall \mathbf{s} \in D \tag{2}$$

where $\{\beta_l\}_{l=1,\ldots,L}$ are random coefficients; $\{\psi_l(\mathbf{s}) : \mathbf{s} \in D\}_{l=1,\ldots,L}$ are eigenfunctions defined in Eq. (1). It is important to note that all the eigenfunctions are considered; so there is no truncation.

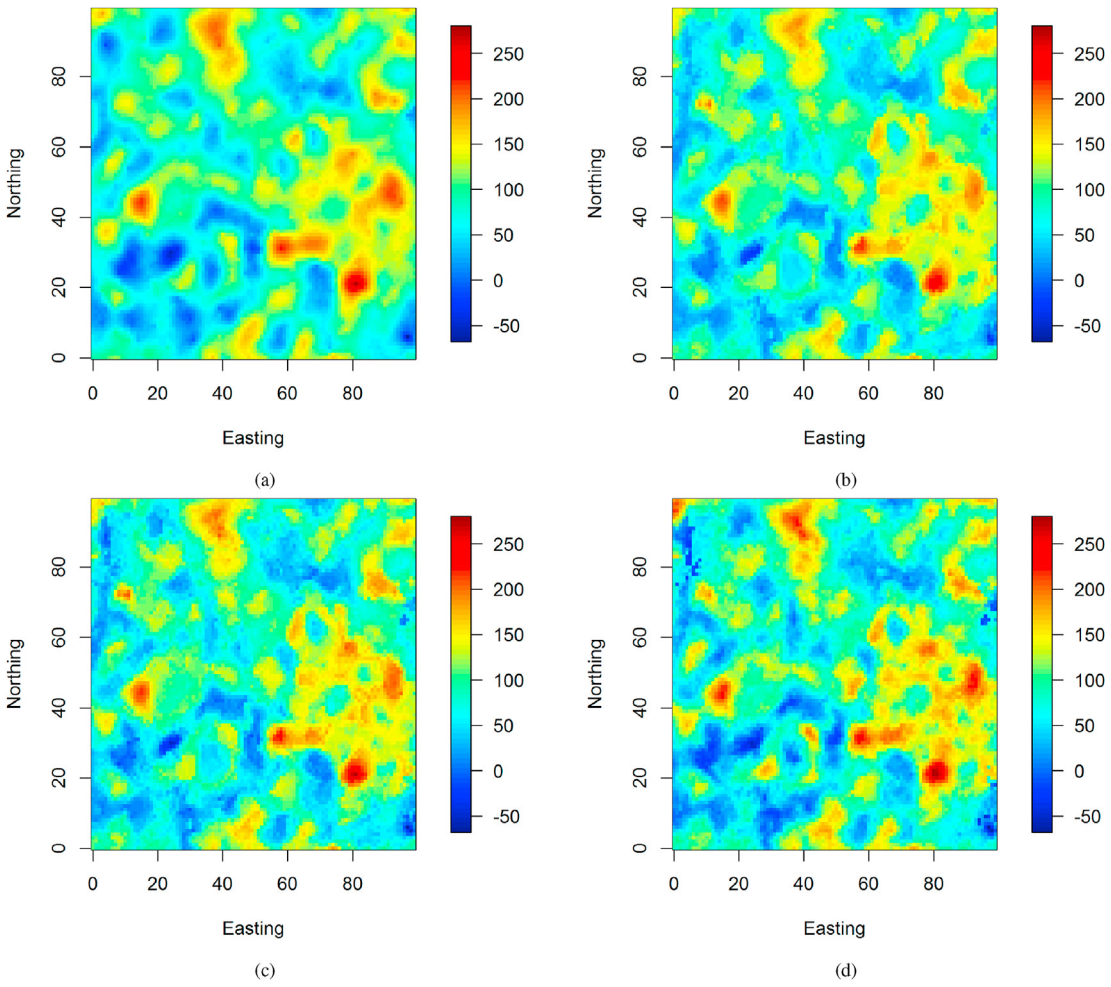


Fig. 5. Synthetic data example - spatial prediction map for (a) regression-kriging, (b) traditional regression random forest, (c) proposed regression random forest, and (d) ground-truth.

The objective here is to sample the vector of coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_L)^{'}$ such that $\breve{Y}(\mathbf{s}) = \sum_{l=1}^{L} \beta_l \psi_l(\mathbf{s})$ perfectly fits the response variable's observed values at sampled locations, i.e., $\breve{Y}(\mathbf{s}_i) = Y(\mathbf{s}_i)$, $i = 1, \ldots, n$. This exact conditioning can be rewritten as follows:

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\beta} \tag{3}$$

where $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^{'}$ is the response variable's measured values at sampled locations and $\mathbf{F} = [\psi_l(\mathbf{s}_i)]$ is a fixed matrix of size $(n \times L)$ whose elements are eigenfunctions at sampled locations.

To explore in a stochastic and convenient way the solution space associated with Eq. (3), the vector of coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_L)^{'}$ is assumed to follow a multivariate Gaussian distribution defined by:

$$\boldsymbol{\beta} \sim f(\boldsymbol{\beta}) = (2\pi)^{-L/2} \left|\boldsymbol{\Sigma}\right|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu})^{'} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}) \right) \tag{4}$$

where the mean $\boldsymbol{\mu} = E[\boldsymbol{\beta}]$ and the covariance matrix $\boldsymbol{\Sigma} = Var[\boldsymbol{\beta}]$ are computed using PC scores $\{\alpha_{b,l}\}$ derived from the PCA of regression tree predictors given in Eq. (1). Specifically,

$$\boldsymbol{\mu} = \left[\frac{1}{B} \sum_{b=1}^{B} \alpha_{b,l} \right]_{l=1,\ldots,L} \quad ;$$

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

Equations (3) and (4) define a Bayes-linear-Gauss problem (Scheidt et al., 2018; Tarantola, 2013). This problem has a solution subject that $L = \min(B,N) > n$. In practice, the number of grid locations is larger than the number of sampled locations ($N \gg n$). So there is a solution when $B > n$. It is important to highlight that given the number of sampled locations n, one can always have $B > n$ since B is free parameter in the regression random forest. Let $f(\boldsymbol{\beta}|\mathbf{Y})$ be the probability density distribution of $\boldsymbol{\beta}|\mathbf{Y}$. According to Bayes rule, $f(\boldsymbol{\beta}|\mathbf{Y})$ is also multivariate Gaussian with mean and covariance given by (Scheidt et al., 2018; Tarantola, 2013):

$$E[\boldsymbol{\beta}|\mathbf{Y}] = \boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{F}^{'} (\mathbf{F}\boldsymbol{\Sigma}\mathbf{F}^{'})^{-1} (\mathbf{Y} - \mathbf{F}\boldsymbol{\mu}) \tag{6}$$

$$Var[\boldsymbol{\beta}|\mathbf{Y}] = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{F}^{'} (\mathbf{F}\boldsymbol{\Sigma}\mathbf{F}^{'})^{-1} \mathbf{F}\boldsymbol{\Sigma} \tag{7}$$

Since $f(\boldsymbol{\beta}|\mathbf{Y})$ is a multivariate Gaussian distribution, it easy to draw Monte Carlo samples from this distribution. The generation of the samples $\{\boldsymbol{\beta}_t^c\}_{t=1,\ldots,T} \sim f(\boldsymbol{\beta}|\mathbf{Y})$ is carried out using the R package *rockchalk* (Johnson, 2019). Given the Monte Carlo samples $\{\boldsymbol{\beta}_t^c\}_{t=1,\ldots,T}$, the reconstructed regression tree predictors are given by:

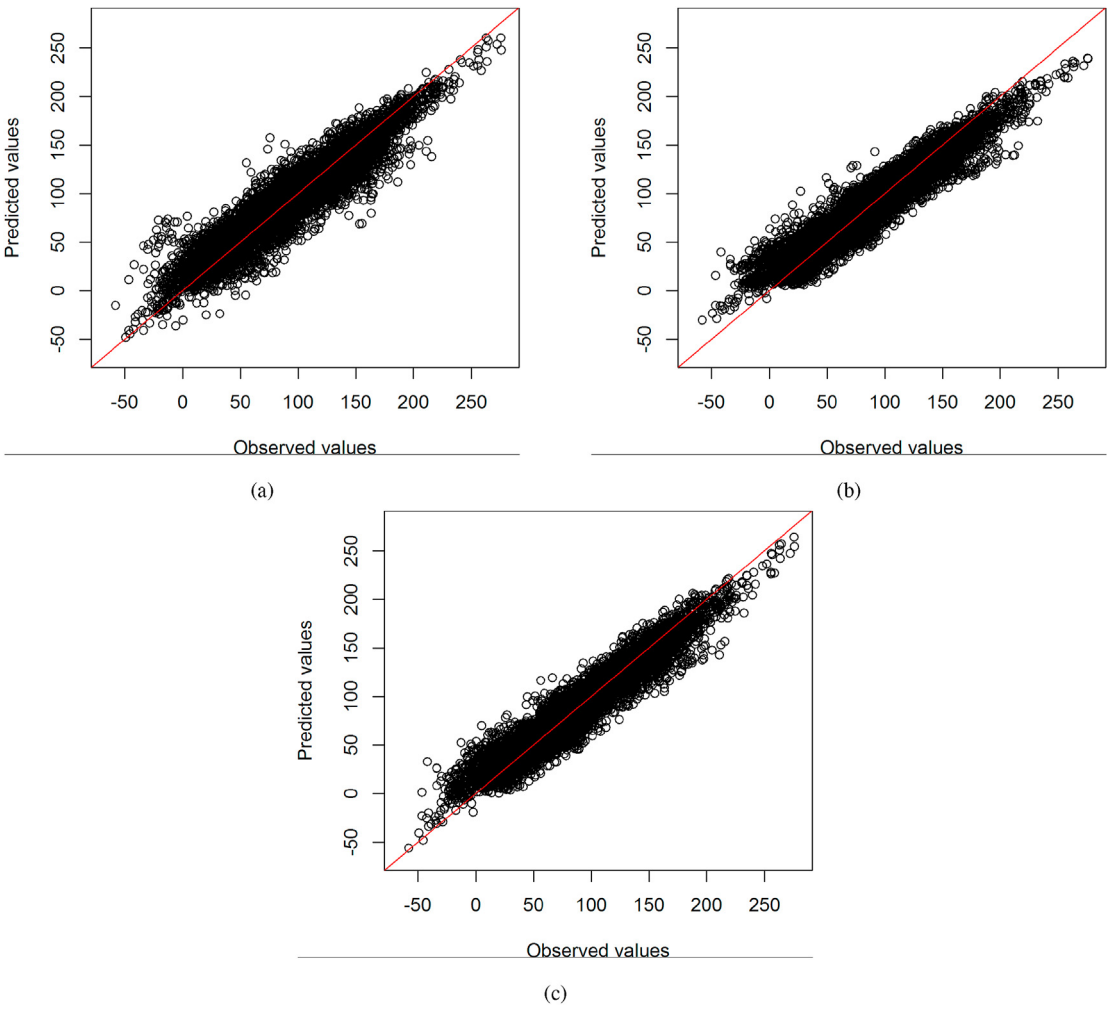


Fig. 6. Synthetic data example - response variable's observed values vs response variable's predicted values in testing dataset, for (a) regression-kriging, (b) traditional regression random forest, and (c) proposed regression random forest.

$$\text{with } \boldsymbol{\beta}_t^c = \left(\beta_{t,1}^c, \ldots, \beta_{t,L}^c\right);\ t = 1, \ldots, T. \qquad (8)$$

The prediction at an unsampled location $\mathbf{s}_0 \in D$ is made by averaging the predictions from all the individual reconstructed regression tree predictors:

$$\widehat{Y}(\mathbf{s}_0) = \frac{1}{T}\sum_{t=1}^{T} \breve{Y}_t(\mathbf{s}_0) \qquad (9)$$

Since all the individual reconstructed regression tree predictors $\{\breve{Y}_t(\mathbf{s}) : \mathbf{s} \in D\}_{t=1,\ldots,T}$ perfectly match the response variable's observed values at sampled locations, their mean $\{\widehat{Y}(\mathbf{s}) : \mathbf{s} \in D\}$ does also. PC scores $\{\boldsymbol{\alpha}_b\}_{b=1,\ldots,B}$ in Eq. (1) are called unconditional PC scores while those $\{\boldsymbol{\beta}_t^c\}_{t=1,\ldots,T}$ in Eq. (8) are called conditional PC scores. It is important to highlight that there no constraint on T relatively to B; T can be less or greater than B.

3. Practical examples

The proposed method's ability to perfectly match the response variable's observed values at sampled locations is illustrated using both synthetic and real datasets. Prediction performance comparison is carried out with the regression-kriging and the traditional regression random forest using some well-known prediction accuracy criteria (Hengl et al., 2018): the mean absolute error (MAE), the root mean square error (RMSE), the coefficient of determination (R-square), and Lin's concordance correlation coefficient (CCC). The lower are MAE and RMSE, the better is the model. The closer are R-square and CCC to 1, the better is the model.

Table 2
Synthetic data example - predictive performance statistics in the testing dataset.

Criteria	Regression-Kriging	Traditional Random Forest	Proposed Random Forest
MAE	12.69	12.01	11.05
RMSE	16.86	15.66	14.40
R-square	0.88	0.90	0.91
CCC	0.93	0.94	0.95

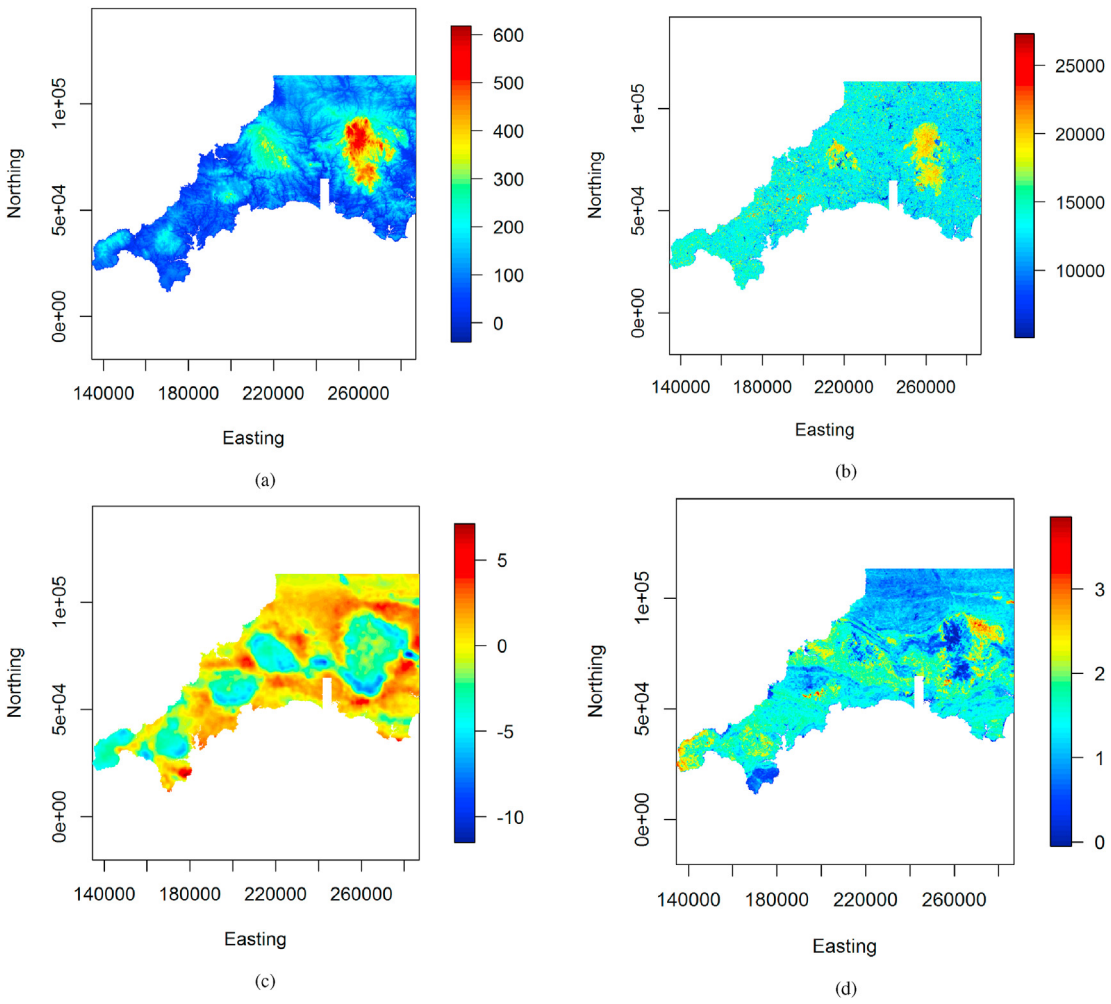


Fig. 7. Real data example - some predictor variables: (a) elevation, (b) landsat 8 band 6, (c) gravity survey high-pass filtered Bouguer anomaly, (d) potassium counts from gamma ray spectrometry.

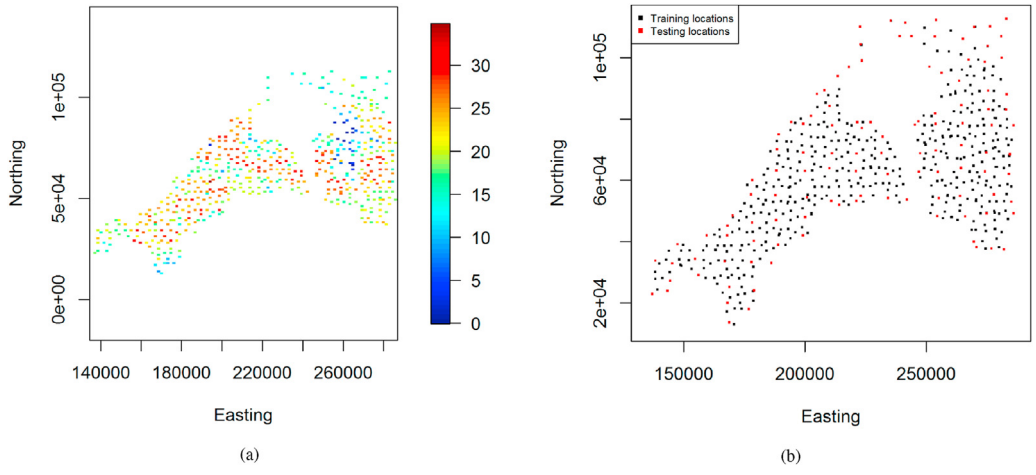


Fig. 8. Real data example - (a) response variable (Ga concentration), and (b) training and testing locations.

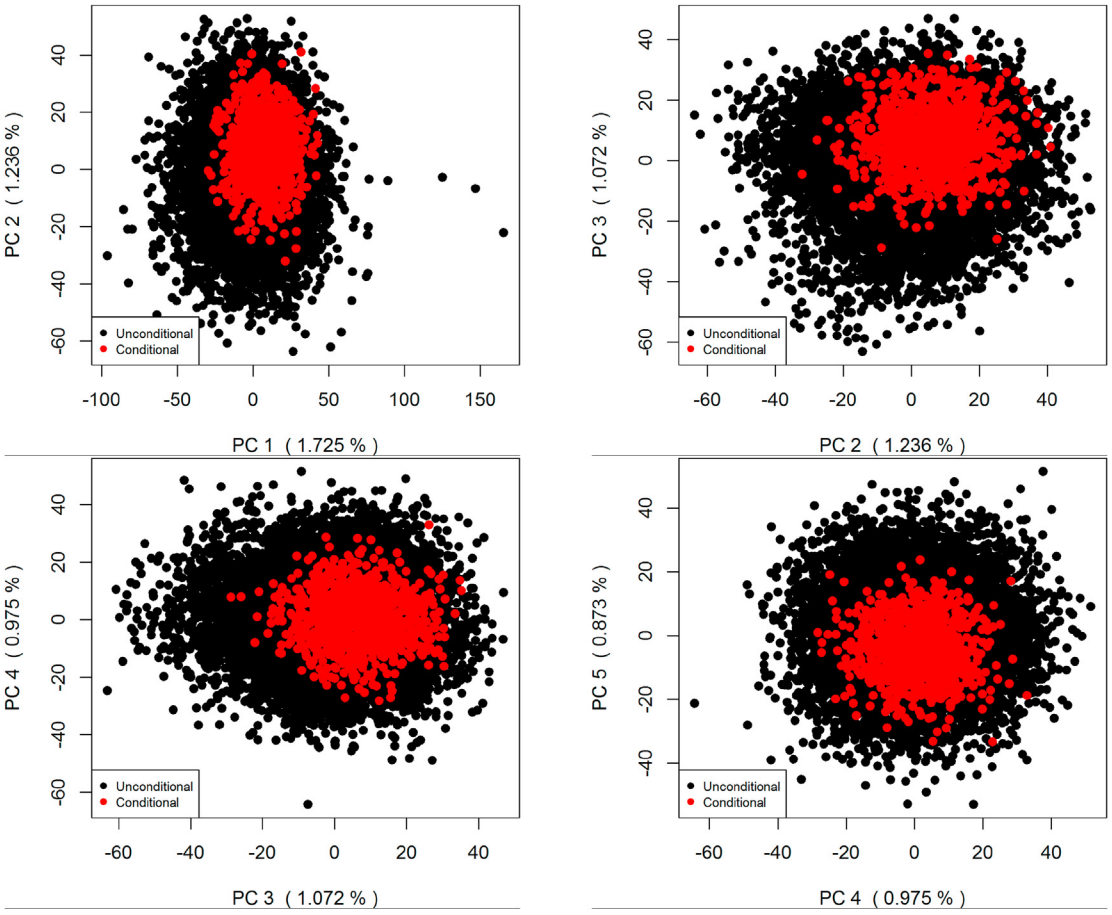


Fig. 9. Real data example - $B = 10000$ unconditional first PC scores and $T = 1000$ conditional first PC scores.

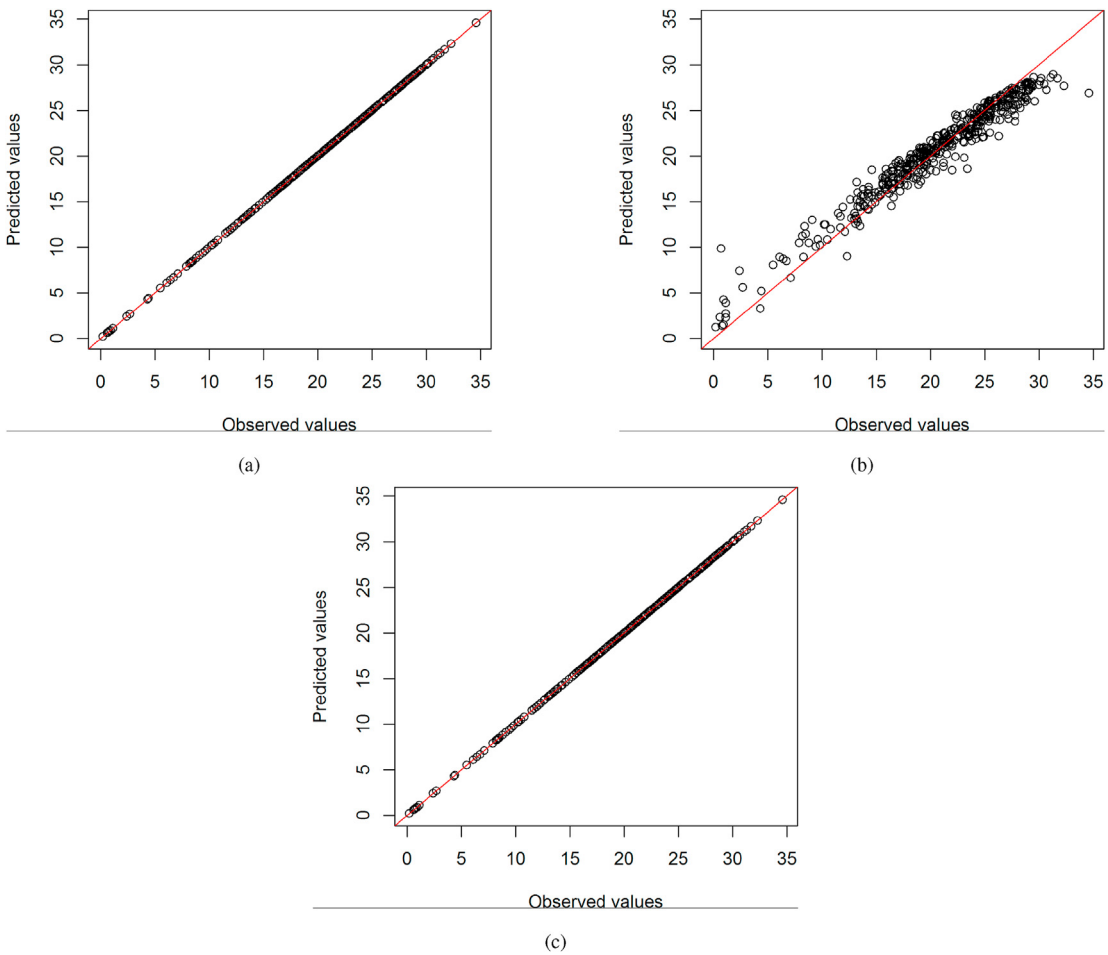


Fig. 10. Real data example - response's observed values vs response's predicted values in the training dataset, for (a) regression-kigring, (b) traditional regression random forest, and (c) proposed regression random forest.

3.1. Synthetic data example

The data-generating process is given by the following underlying model:

$$Y(\mathbf{s}) = X_1(\mathbf{s})^2 + 40\tanh(X_1(\mathbf{s})X_2(\mathbf{s})) + X_3(\mathbf{s})^2 + 50\sin(X_4(\mathbf{s})) + \varepsilon(\mathbf{s}), \quad \forall \mathbf{s} \in [0,100]^2 \tag{10}$$

where $Y(\cdot)$ is the response variable; $X_1(\cdot)$, $X_2(\cdot)$, $X_3(\cdot)$, and $X_4(\cdot)$ are predictor variables; and $\varepsilon(\cdot)$ is a latent (non-observed) variable.

Predictor variables are simulated on the spatial domain $[0,100]^2$ as Gaussian isotropic stationary random fields with mean and covariance function given in Table 1. For background on Gaussian random fields, see Chiles and Delfiner (2012). The simulation is performing using the R package *RGeostats* package (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 not Gaussian.

Fig. 1 presents the synthetic data over a 100×100 regular grid. $n = 1000$ observations are sampled randomly and taken as the training data as shown in Fig. (2). The rest of data (9000 observations) is kept aside for the testing. The regression random forest is performed on the training data with a large number of regression trees set to $B = 10000$. Thus, an ensemble of $B = 10000$ regression tree predictors $\{\tilde{Y}_b(\mathbf{s}) : \mathbf{s} \in [0,100]^2\}_{b=1,\ldots,10000}$ is constructed. According to the methodology described in Sect. 2, principal component analysis is performed on this ensemble, followed by the Monte Carlo sampling of new PC scores ensuring the exact conditioning. $T = 1000$ new PC scores are generated, thus giving an ensemble of $T = 1000$ reconstructed (new) regression tree predictors $\left\{Y_t^{\smile}(\mathbf{s}) : \mathbf{s} \in [0,100]^2\right\}_{t=1,\ldots,1000}$ that perfectly match the response variable's observed values at sampled locations. The mean prediction of reconstructed (new) regression tree predictors also exactly matches the response variable's observed values at sampled locations.

PC scores before the conditioning (unconditional PC scores $\{\boldsymbol{\alpha}_b\}_{b=1,\ldots,B}$) and after the conditioning (conditional PC scores $\{\boldsymbol{\beta}_t^c\}_{t=1,\ldots,T}$) are presented in Fig. 3. In this figure, the points cloud of conditional PC scores is less scattered than those from unconditional PC scores due effectively to the conditioning. Indeed, unconditional PC scores have no constraints on their possible values. They can be any point in the whole Euclidean space $\mathbb{R}^L$. However, the set of linear constraints defined by Eq. (3) produces a convex feasible region of possible values for conditional PC scores. So, conditional PC scores can not be any point in the Euclidean space $\mathbb{R}^L$.

Fig. 4 shows the response variable's observed values versus predicted values in the training data, under the regression-kriging, the traditional regression random forest, and the proposed one. One can effectively notice that the traditional regression random forest does not fit the training data perfectly while the regression-kriging and the proposed regression random forest do. Fig. 5 presents spatial prediction maps

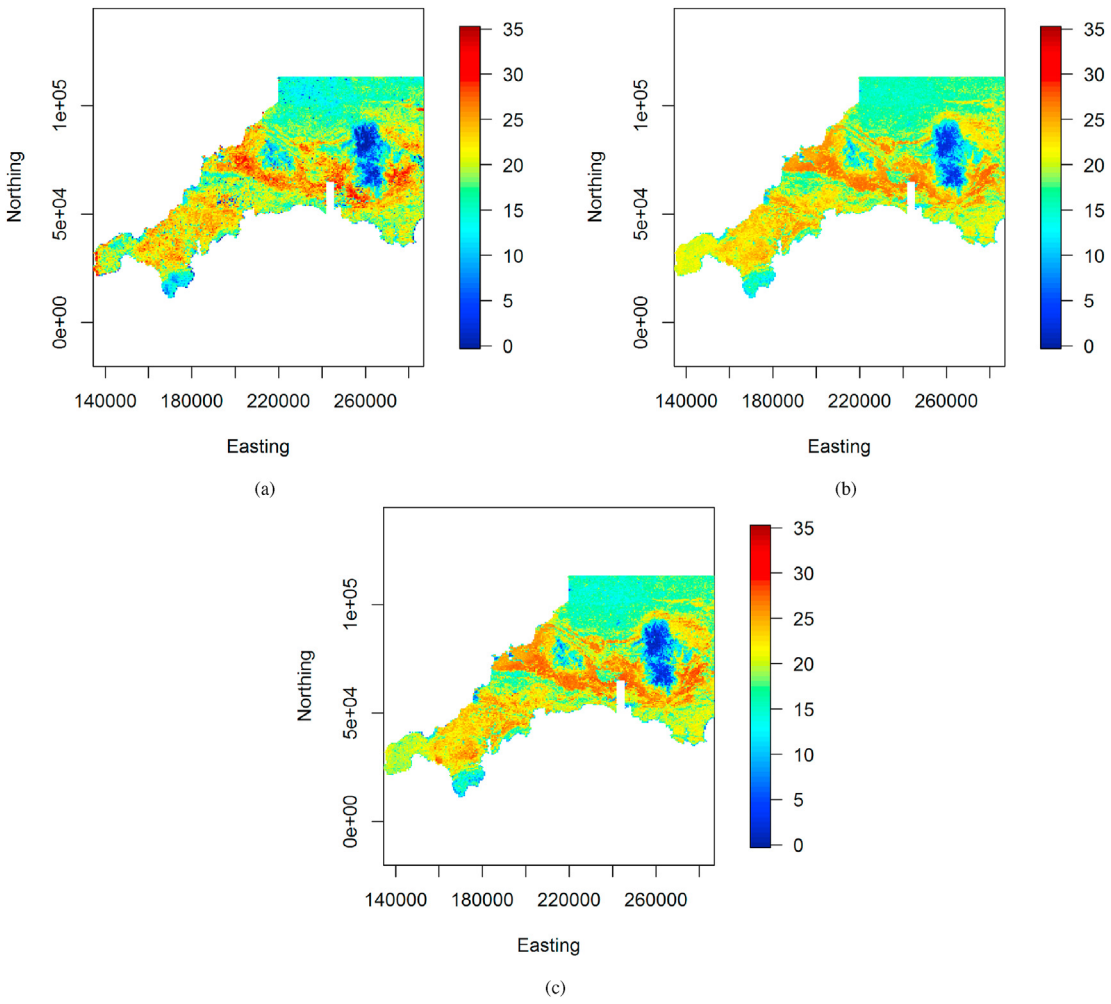


Fig. 11. Real data example - spatial prediction map for (a) regresssion-kriging, (b) traditional regression random forest, and (c) proposed regression random forest.

provided by the regression-kriging, the traditional regression random forest, and the proposed one. The spatial prediction map resulting from the regression-kriging differs from the ones provided by the traditional regression random forest and the proposed regression random forest. In particular, the spatial prediction map of regression-kriging is smoother than the two others.

The overall look of spatial prediction maps resulting from the traditional regression random forest and the proposed regression random forest looks very similar. However, there are some local differences due to the exact conditioning of the proposed regression random forest. The predictive performance in the testing set for the regression-kriging, the traditional regression random forest, and the proposed one is shown in Fig. 6 and Table 2. One can notice that the proposed regression random forest performs better than the regression-kriging and the traditional regression random forest.

3.2. Real data example

The real dataset of interest comprises geochemical concentration data of the southwest England (Kirkwood et al., 2016b). We are interested in the target variable Ga (Gallium) concentration which is measured at 568 locations over the spatial domain of interest. Predictor variables include elevation, gravity, magnetic, Landsat, radiometric, and their derivatives, totaling 26 predictor variables. Some predictor variables are shown in Fig. 7. Fig. 8a presents measurements of the response variable. The data are divided into a training set ($\approx$ 80%) and testing set ($\approx$ 20%) as shown in Fig. 8b.

The estimated regression random forest model is an ensemble of $B = 10000$ regression tree predictors. PCA applied to this ensemble, following by the Monte Carlo sampling result to unconditional and conditional PC scores as shown in Fig. 9; $T = 1000$ conditional PC scores are generated. Fig. 10 shows the conditioning performance in the training data of the regression-kriging, the traditional regression random forest, and the proposed regression random forest. As noticed in the simulated data example, the regression-kriging and proposed regression random forest perfectly fit the data while the traditional regression random forest does not.

Spatial prediction maps provided by the regression-kriging, the traditional regression random forest, and the proposed one are depicted in Fig. 11. The spatial prediction map of the regression-kriging differs from the two others. The general appearance of the spatial prediction maps of these later looks very similar. However, one notes some local differences due to the exact conditioning of the proposed regression random forest. Fig. 12 and Table 3 present the predictive performance of the regression-kriging, the traditional regression random forest, and the proposed regression random forest on the testing data. The proposed regression random forest shows better predictive performance than the two other methods. Thus, the proposed approach can exactly match the response variable's measured values at sampled locations while achieving good out of sample predictive performance.

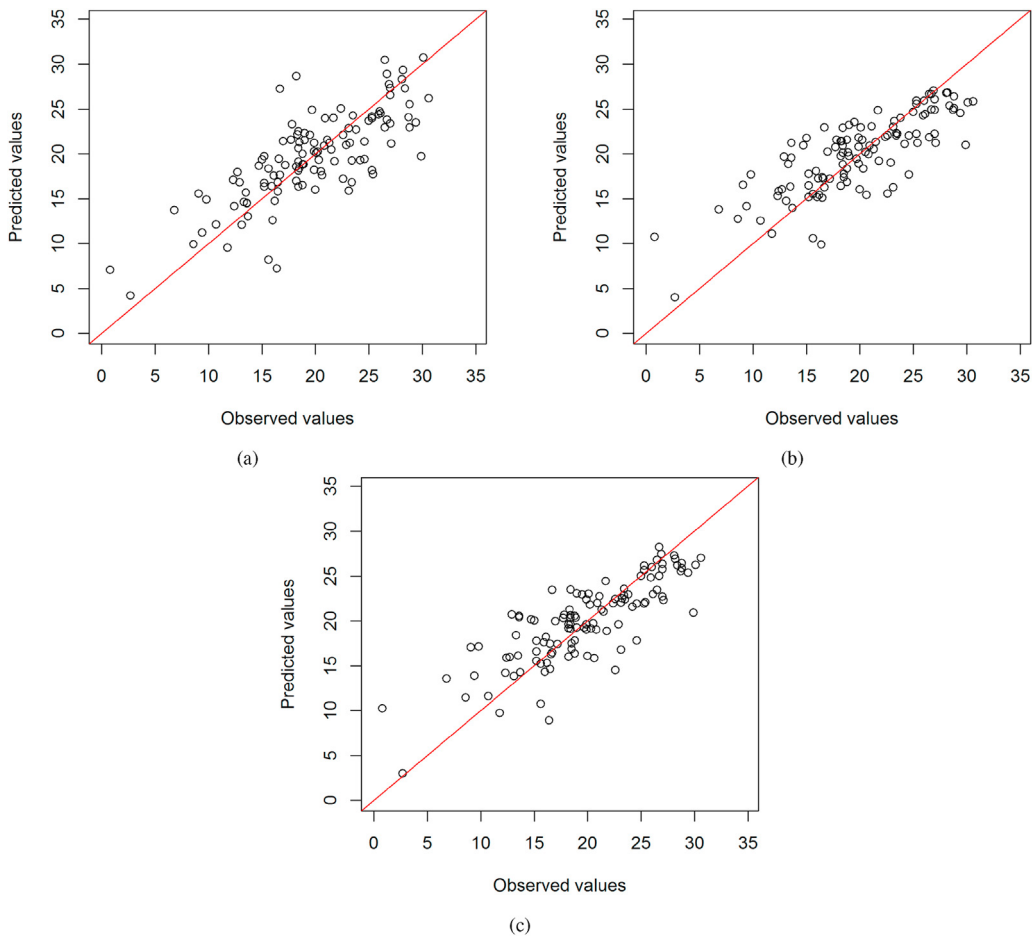


Fig. 12. Real data example - response variable's observed values vs response variable's predicted values in testing dataset, for (a) regression-kriging, (b) traditional regression random forest, and (c) proposed regression random forest.

Table 3
Real data example - predictive performance statistics on the testing dataset.

Criteria	Regression-Kriging	Traditional Random Forest	Proposed Random Forest
MAE	2.87	2.74	2.63
RMSE	3.79	3.59	3.46
R-square	0.57	0.61	0.64
CCC	0.74	0.75	0.78

4. Concluding remarks

This article presented a methodology guaranteeing that regression random forest perfectly matches the response variable's observed values at sampled locations like competitor techniques such as regression-kriging. Given the ensemble of regression tree predictors resulting from the traditional regression random forest, the exact conditioning is achieved by combining principal component analysis and Monte Carlo sampling. As a result, a new ensemble of regression tree predictors that perfectly fit the response variable's observed values at sampled locations is obtained. The average of this new ensemble of regression tree predictors also exactly matches the response variable's observed values at sampled locations by construction. The effectiveness of the proposed approach has been demonstrated on both synthetic and real datasets. It can perfectly fit the response variable's measured values at sampled locations while achieving good out of sample performance comparatively to regression-kriging and traditional regression random forest. It is easy to implement since it combines well-known existing machine learning and Monte Carlo sampling methods.

As highlighted previously, the exact conditioning is performed subject that the number of regression trees is greater than the number of sampled locations. Nonetheless, it will always be possible to meet this constraint because the number of regression trees is a free parameter. The number of regression trees should also be large enough to allow good coverage of the solution space when performing the exact conditioning. The proposed approach relies on the regression random forest to perform the exact conditioning. Thus, one expects the proposed method to provide better predictive performance in situations favorable to regression random forest like a non-linear relationship between the response variable and predictor variables and some interactions between predictor variables. In a situation where there is linear relationship between the dependent variable and predictor variables, competitor techniques like regression-kriging could perform better. The proposed method can be used in any spatial dimension (e.g., 2D and 3D). It can be extended in the case where the response variable is categorical. The exact conditioning can be achieved following the idea developed in Fouedjio et al. (2020).

Declaration of competing interest

There is no conflict of interest.

Acknowledgments

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

References

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., 2017. Comparison of machine learning models for predicting fluoride contamination in groundwater. Stochastic Environmental Research and Risk Assessment 31 (10), 2705–2718.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Chiles, J.P., Delfiner, P., 2012. Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons.

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

Fouedjio, F., Scheidt, C., Yang, L., Jef, C., 2020. 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. https://doi.org/10.1007/s00477-020-01925-7.

Hengl, T., Heuvelink, G.B., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120, 75–93.

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. https://doi.org/10.7717/peerj.5518.

Johnson, P.E., 2019. Regression estimation and presentation. URL. rockchalk. https://CRAN.R-project.org/package=rockchalk. r package version 1.8.144.

Khan, S.Z., Suman, S., Pavani, M., Das, S.K., 2016. Prediction of the residual strength of clay using functional networks. Geoscience Frontiers 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.

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 (AusDM´13). Australia, Canberra, 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.

Probst, P., Wright, M., Boulesteix, A.L., 2018. Hyperparameters and Tuning Strategies for Random Forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery doi:10.1002/widm.1301.

R Core Team, 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/. (Accessed 15 November 2020).

Renard, D., Bez, N., Desassis, N., Beucher, H., Ors, F., Freulon, X., 2020. Geostatistical package. URL. RGeostats. http://cg.ensmp.fr/rgeostats. r package version 12.0.1.

Scheidt, C., Li, L., Caers, J., 2018. Quantifying Uncertainty in Subsurface Systems. Geophysical Monograph Series. Wiley.

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.

Tarantola, A., 2013. Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation. Elsevier Science.

Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64.

Vermeulen, D., Niekerk, A.V., 2017. Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates. Geoderma 299, 1–12.

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.