Environmental Earth Sciences (2019) 78:38
https://doi.org/10.1007/s12665-018-8032-z

THEMATIC ISSUE

Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches

Francky Fouedjio[1] · Jens Klump[2]

Received: 30 March 2018 / Accepted: 24 December 2018 / Published online: 3 January 2019

Abstract
Geostatistical methods such as kriging with external drift (KED) as well as machine learning techniques such as quantile regression forest (QRF) have been extensively used for the modeling and prediction of spatially distributed continuous variables when auxiliary information is available everywhere within the region under study. In addition to providing predictions, both methods are able to deliver a quantification of the uncertainty associated with the prediction. In this paper, kriging with external drift and quantile regression forest are compared with respect to their ability to deliver reliable predictions and prediction uncertainties of spatial data. The comparison is carried out through both synthetic and real-world spatial data. The results indicate that the superiority of KED over QRF can be expected when there is a linear relationship between the variable of interest and auxiliary variables, and the variable of interest shows a strong or weak spatial correlation. In other hand, the superiority of QRF over KED can be expected when there is a non-linear relationship between the variable of interest and auxiliary variables, and the variable of interest exhibits a weak spatial correlation. Moreover, when there is a non-linear relationship between the variable of interest and auxiliary variables, and the variable of interest shows a strong spatial correlation, one can expect QRF outperforms KED in terms of prediction accuracy but not in terms of prediction uncertainty accuracy.

Keywords Auxiliary information · Prediction uncertainty · Kriging with external drift · Quantile regression forest · Spatial data

Introduction

Most of the time in mineral exploration, measurements of the spatially distributed variable of interest (e.g., geochemical element concentration) are expensive to obtain (Carranza 2008). In fact, both the samples and the associated chemical analyses are often laborious and difficult to obtain and, therefore, come at a high cost. As consequence, measurements of the spatially distributed variable of interest are relatively scarce over the region of interest. A simple interpolation of such relatively sparse spatial data always involves large uncertainties. With the increasing development of remote sensing platforms and sensor networks, large volumes of diverse geoscientific data (e.g., geological, geophysical) are becoming available everywhere within the region of interest. This vast amount of auxiliary spatial data has the potential to improve the prediction of the variable of interest over the region of interest, beyond interpolations based solely on point measurements of the variable of interest (Hengl 2009). The underlying assumption is that the spatially distributed variable of interest, which is known at only relatively few locations, is correlated to auxiliary spatial variables which are available everywhere within the region of interest.

Geostatistical methods such as kriging with external drift (KED) (Chiles and Delfiner 2012; Wackernagel 2013) and machine learning techniques such as quantile

This article is part of a Topical Collection in Environmental Earth Sciences on "Learning from spatial data: unveiling the geo-environment through quantitative approaches", guest edited by Sebastiano Trevisani, Marco Cavalli, Jean Golay, and Paulo Pereira.

✉ Francky Fouedjio
francky.fouedjio@stanford.edu

Jens Klump
jens.klump@csiro.au

[1] Department of Geological Sciences, Stanford University, 367 Panama Street, Stanford, CA 94305, USA

[2] CSIRO Mineral Resources, 26 Dick Perry Ave, Kensington, WA 6151, Australia

regression forest (QRF) (Meinshausen 2006) have been intensively used for modeling and mapping of spatially distributed continuous variables in the presence of secondary information available everywhere within the region under study (Hengl et al. 2004; Lado et al. 2008; Kanevski 2008; Li and Heap 2008; Kanevski et al. 2009; Foresti et al. 2010; Hengl 2009; Li et al. 2011; Li 2013; Tadic et al. 2015; Leuenberger and Kanevski 2015; Appelhans et al. 2015; Kirkwood et al. 2016; Taghizadeh-Mehrjardi et al. 2016; Ballabio et al. 2016; Barzegar et al. 2016; Khan et al. 2016; Wilford et al. 2016; Vaysse and Lagacherie 2017; Vermeulen and Niekerk 2017; Hengl et al. 2018). In addition to providing a prediction, both approaches can deliver a quantification of the uncertainty associated with the prediction. Geostatistical approaches such as KED are, by essence, designed to provide such prediction uncertainties. However, they frequently require significant data pre-processing, can handle only linear relationships, and make some assumptions about the underlying spatial distribution of data (e.g., stationarity, isotropy, and normality) which are rarely met in practice. In contrast to geostatistical methods, machine learning techniques such as QRF, often require less data pre-processing, can handle complex nonlinear relationships, make no assumption about the underlying spatial distribution of the data though relying on the independence assumption of the data. This assumption is often unrealistic, especially, when the sampling density is very dense in some areas and very sparse in others.

Though there have been numerous studies comparing geostatistical and machine learning methods in terms of their accuracy in making point predictions, very little attention has been paid to their ability to provide reliable prediction uncertainties (Coulston et al. 2016; Kirkwood et al. 2016; Vaysse and Lagacherie 2017). Coulston et al. (2016) provided an approach for approximating prediction uncertainty for random forest regression models in a spatial framework. Kirkwood et al. (2016) compared the capability of ordinary kriging and quantile regression forest to provide reliable prediction uncertainties of various geochemical mapping products in south west England. Vaysse and Lagacherie (2017) performed the same comparison for digital soil mapping products in France.

The prediction uncertainty represents here the uncertainty around the prediction at a target location, and it reflects the inability to exactly define the unknown value. Assessing the uncertainty about the value of the variable of interest at target locations, and of the need to incorporate this assessment in subsequent studies or to support decision making is becoming increasingly important. Uncertainty about any particular unknown value is modeled by a probability distribution of that unknown value conditional to available related information. Their determination should be done prior and independently of the predictor(s) retained, and accounts for the data configuration, data values, and data quality.

Thus, the aim of the present work is twofold: Kriging with external drift (KED) and quantile regression forest (QRF) are compared (1) with respect to their accuracy in making point predictions and (2) their success in modeling prediction uncertainty of spatial data. For this comparison, we used both simulated and real-world spatial data. Apart from classical performance indicators, comparisons make use of accuracy plots, probability interval width plots, and the visual examinations of the prediction uncertainty maps provided by the two methods.

Methods and data

Methods

In this section, KED (Chiles and Delfiner 2012; Wackernagel 2013) and QRF (Meinshausen 2006) are described, as well as performance measures used to compare them. All modeling was conducted in R (R Core Team 2018). KED is performed using RGeostats package (Renard et al. 2018) and QRF is carried out with quantregForest package (Meinshausen 2017).

Kriging with external drift

KED is a particular case of universal kriging (Chiles and Delfiner 2012; Wackernagel 2013). It allows the prediction of a spatially distributed variable of interest (target variable or dependent variable or response variable) $\{Y(\mathbf{s}),\ \mathbf{s} \in D \subset \mathbb{R}^d\}$, known only at relatively small set of locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ of the study region D, through spatially distributed auxiliary variables (explanatory variables or independent variables or covariates) $\{\mathrm{x}_l(\mathbf{s}),\ \mathbf{s} \in D\}_{l=1,\ldots,L}$, exhaustively known in the same area. It assumes that the spatially distributed variable of interest can be modeled as a second-order random field of the form (Chiles and Delfiner 2012; Wackernagel 2013):

$$Y(\mathbf{s}) = m(\mathbf{s}) + e(\mathbf{s}),\ \mathbf{s} \in D \subset \mathbb{R}^d, \tag{1}$$

where $m(\cdot)$ is the mean function (drift) assumed to be deterministic and continuous, and $e(\cdot)$ is a zero-mean second-order stationary random field (residual) with covariance function $\mathrm{Cov}(e(\mathbf{s}_1), e(\mathbf{s}_2)) = \mathrm{C}(\mathbf{s}_1 - \mathbf{s}_2)$.

Under the model defined in Eq. (1), the large-scale spatial variation is accounted through the mean function $m(\cdot)$, and the small-scale spatial variation (spatial dependence) is accounted through the second-order random field $e(\cdot)$. Under KED, it is assumed that the mean function $m(\cdot)$ should vary smoothly in the spatial studied domain D. KED assumes a linear relationship between the variable of interest and auxiliary variables at the observation points of the variable of interest. More specifically, the mean function is expressed as follows:

$$m(\mathbf{s}) = \beta_0 + \sum_{l=1}^{L} \beta_l \mathrm{x}_l(\mathbf{s}), \tag{2}$$

where $\{\beta_0, \dots, \beta_L\}$ are unrestricted parameters.

The prediction of the unknown value Y_0 of the variable of interest at a new location $\mathbf{s}_0 \in D$ is given by

$$\widehat{Y}_0 = \boldsymbol{\Lambda}^T \mathbf{Y}, \tag{3}$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ is the vector of observations at sampling locations $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$, and $\boldsymbol{\Lambda}^T = (\lambda_1(s_0), \dots, \lambda_n(s_0))$ are weights solution of the following system of equations obtained by minimizing the mean squared prediction error under unbiasedness constraints (Chiles and Delfiner 2012; Wackernagel 2013):

$$\begin{pmatrix} \mathbf{C} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda} \\ \boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_0 \\ \mathbf{x}_0 \end{pmatrix}, \tag{4}$$

where $\mathbf{C}$ is the covariance matrix associated with the observations, i.e., with entries $C_{ij} = \mathrm{C}(\mathbf{s}_i - \mathbf{s}_j),\ (i,j) \in \{1, \dots, n\}^2$, $\mathbf{C}_0$ is the vector of covariances between Y_0 and the observations, $\mathbf{0}$ is the matrix of zeroes $\mathbf{x}_0 = (1, \mathrm{x}_1(\mathbf{s}_0), \dots, \mathrm{x}_L(\mathbf{s}_0))^T$,

$$\mathbf{X} = \begin{pmatrix} 1 & \mathrm{x}_1(\mathbf{s}_1) & \dots & \mathrm{x}_L(\mathbf{s}_1) \\ \vdots & & \ddots & \vdots \\ 1 & \mathrm{x}_1(\mathbf{s}_n) & \dots & \mathrm{x}_L(\mathbf{s}_n) \end{pmatrix},$$

and $\boldsymbol{\mu} = (\mu_0, \mu_1, \dots, \mu_L)^T$ is the vector of Lagrange multipliers accounting for the unbiasedness constraints.

The KED predictor at a new location $\mathbf{s}_0 \in D$ is computed as (Chiles and Delfiner 2012; Wackernagel 2013)

$$\widehat{Y}_0 = \mathbf{C}_0^T \mathbf{C}^{-1} \mathbf{Y} + (\mathbf{x}_0^T - \mathbf{C}_0^T \mathbf{C}^{-1} \mathbf{X})(\mathbf{X}^T \mathbf{C}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{C}^{-1} \mathbf{Y}, \tag{5}$$

and the associated prediction error variance or kriging error variance is given by

$$\begin{aligned} \widehat{\sigma}_0^2 &= C(0) - \mathbf{C}_0^T \mathbf{C}^{-1} \mathbf{C}_0 \\ &\quad + (\mathbf{x}_0 - \mathbf{X}^T \mathbf{C}^{-1} \mathbf{C}_0)^T (\mathbf{X}^T \mathbf{C}^{-1} \mathbf{X})^{-1} (\mathbf{x}_0 - \mathbf{X}^T \mathbf{C}^{-1} \mathbf{C}_0), \end{aligned} \tag{6}$$

where $C(0) = \mathbb{V}(Y(\mathbf{s}))$ corresponding to the punctual variance. Here the interpolation is carried out using all data in the domain of interest (unique neighborhood).

Thus, KED naturally generates uncertainty estimates for interpolated values via the kriging variance. The first two terms on the right-hand side of Eq. (6) quantify the prediction error variance of the residuals, while the last term which is always non-negative is the estimated drift prediction error variance representing the penalty for having to estimate $\{\beta_0, \dots, \beta_L\}$. It is important to point out that KED is equivalent to optimum drift estimation followed by simple kriging of the residuals from this drift estimate, as if the mean were estimated perfectly. This property only holds when the mean is estimated in a statistically consistent manner–that is, by generalized least squares (GLS) and not by ordinary least squares (OLS). The GLS method itself requires a covariance function for the residuals, so an iterative procedure is followed. The OLS estimates are obtained, and a covariance function is fitted to the residuals. This covariance function is then used in GLS to re-estimate the spatial trend parameters, and the procedure is repeated until the estimates stabilize (Hengl et al. 2004). However, it may happen that this iterative process does not converge. KED is often termed regression kriging (RK) (Hengl et al. 2004). In KED, the estimation of the drift coefficients and the kriging of the residuals are performed in an integrated way, while in RK, the regression and kriging are carried out separately.

By assuming a Gaussian distribution of the kriging error $Y_0 - \widehat{Y}_0$, its distribution is completely specified by its mean (zero) and its variance (kriging variance). Thus, the conditional distribution function (cdf) of the variable of interest at $\mathbf{s}_0$ $F(\mathbf{s}_0; y | Y_1, \dots, Y_n) = \mathbb{P}(Y_0 < y | Y_1, \dots, Y_n)$ is estimated as $\widehat{F}(\mathbf{s}_0; y | Y_1, \dots, Y_n) = \mathcal{N}\left(\frac{y - \widehat{Y}_0}{\widehat{\sigma}_0}\right)$, where $\mathcal{N}(\cdot)$ is the standard Gaussian distribution. Hence, the predicted value and the kriging variance can be used to derive a Gaussian-type confidence interval centered on the predicted value. A $100(1-\alpha)\%$ $(0 < \alpha < 1)$ prediction interval for Y_0 is given by (Chiles and Delfiner 2012; Wackernagel 2013)

$$[\widehat{Q}_{\alpha/2}(\mathbf{s}_0), \widehat{Q}_{(1-\alpha/2)}(\mathbf{s}_0)] = \left[\widehat{Y}_0 + z_{\alpha/2}\widehat{\sigma}_0, \widehat{Y}_0 + z_{(1-\alpha/2)}\widehat{\sigma}_0\right], \tag{7}$$

where $\widehat{Q}_\alpha(\mathbf{s}_0)$ denotes the α-quantile of the cdf of Y_0 defined as $\widehat{Q}_\alpha(\mathbf{s}_0) = \inf\{y : \widehat{F}(\mathbf{s}_0; y | Y_1, \dots, Y_n) \geq \alpha\}$; z_α is the α-quantile of the standard Gaussian distribution $\mathcal{N}(\cdot)$. The interquartile range $\widehat{Q}_{(1-\alpha/2)}(\mathbf{s}_0) - \widehat{Q}_{\alpha/2}(\mathbf{s}_0) = \left(z_{(1-\alpha/2)} - z_{\alpha/2}\right)\widehat{\sigma}_0$ can be used as a measure of uncertainty as well as the kriging variance.

To use the Gaussian error model for prediction uncertainty quantification, the multi-gaussianity should be checked a-priori. However, some non-parametric geostatistical methods such as indicator kriging are available to model non-Gaussian errors as well as data transformation approaches. It is important to note that to model the prediction uncertainty in KED, the Gaussian assumption is assumed for the kriging error and not for the target variable.

Quantile regression forest

QRF (Meinshausen 2006) is an extension of regression random forest (Breiman 2001). This latter is an ensemble method based on the averaged outputs of multiple decision trees (Breiman et al. 1984). A Decision tree is a non-parametric regression model that works on non-linear situations. A decision tree model partitions the data into subsets of leaf nodes and the prediction value in each leaf node is taken as the mean of the response values of the observations in that leaf node. Decision tree model is unstable in high-dimensional data because of the large prediction variance. This problem can be overcome using an ensemble of decision trees (e.g., regression random forest) built from the bagged samples of data.

From n independent observations $\{(Y_i, \mathbf{X}_i)\}_{i=1,\dots,n}$, regression random forest grows an ensemble of decision trees to learn the model $Y = f(\mathbf{X}) + \epsilon$, where $Y \in \mathbb{R}$ is the variable of interest (target variable or dependent variable or response variable); $\mathbf{X} = [X_1, \dots, X_L] \in \mathbb{R}^L$ is the vector of covariates (explanatory variables or auxiliary variables or independent variables); ϵ is error that is independent of the covariates $\mathbf{X}$. Each decision tree is grown from a separate sub-sample (roughly two-third) of the full data (bagged version of the data). Regression random forest takes the average of multiple decision tree predictions to reduce the prediction variance and increase the accuracy of prediction.

For each decision tree and each node, regression random forest employs randomness when selecting a covariate to split on. In addition, only a random subset of covariates is considered for split-point selection at each node. This reduces the chance of the same very strong covariates being chosen at every split and, therefore, prevents trees from becoming overly correlated. Every node in the decision trees is a condition on a single covariate, designed to split the data set into two so that similar response values end up in the same set. The split (optimal condition) is determined by the impurity reduction at the node; impurity being measured by the variance. For every leaf of every decision tree, the average of all response values end up in this leaf is taken as the prediction value of the leaf.

QRF follows the same idea as described above to grow trees. However, for every leaf of every decision tree, it retains all observations in this leaf, not just their average. Therefore, QRF keeps the raw distribution of the values of the target variable at leaf. For a given new point $\mathbf{X} = \mathbf{x}_0$, regression random forest models the conditional mean $\mathbb{E}(Y|\mathbf{X} = \mathbf{x}_0)$, while QRF models the full conditional distribution function $F(y|\mathbf{X} = \mathbf{x}_0) = \mathbb{P}(Y < y|\mathbf{X} = \mathbf{x}_0)$.

For a given new data point $\mathbf{X} = \mathbf{x}_0$, let $l_k(\mathbf{x}_0)$ be the leaf of the k-th decision tree containing $\mathbf{x}_0$. All $\mathbf{X}_i \in l_k(\mathbf{x}_0)$ are assigned to an equal weight $w_{ik}(\mathbf{x}_0) = 1/n_{lk}$ and $\mathbf{X}_i \notin l_k(\mathbf{x}_0)$ are assigned 0 otherwise, where n_{lk} is the number of observations in $l_k(\mathbf{x}_0)$.

For a single decision tree prediction, given $\mathbf{X} = \mathbf{x}_0$, the prediction value is given by (Meinshausen 2006)

$$\begin{aligned}\hat{Y}_0^k = \sum_{i=1}^{n} w_{ik}(\mathbf{x}_0) Y_i &= \sum_{\mathbf{x}_0, \mathbf{X}_i \in l_k(\mathbf{x}_0)} w_{ik}(\mathbf{x_0}) Y_i \\ &= \frac{1}{n_{lk}} \sum_{\mathbf{x}_0, \mathbf{X}_i \in l_k(\mathbf{x}_0)} Y_i.\end{aligned} \tag{8}$$

Let $w_i(\mathbf{x}_0)$ be the average of weights over all decision trees, that is: $w_i(\mathbf{x}_0) = K^{-1} \sum_{k=1}^{K} w_{ik}(\mathbf{x}_0)$, K being the number of decision trees. The prediction of regression random forest is given by (Meinshausen 2006)

$$\hat{Y}_0 = \sum_{i=1}^{n} w_i(\mathbf{x}_0) Y_i = \mathbf{W}^T \mathbf{Y}, \tag{9}$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, and $\mathbf{W}^T = (w_1(\mathbf{x}_0), \dots, w_n(\mathbf{x}_0))$.

Given an input $\mathbf{X} = \mathbf{x}_0$, we can find leaves $\{l_k(\mathbf{x}_0)\}_{k=1,\dots,K}$ from all decision trees and the sets of observations belonging to these leaves. Given $\{Y_i\}_{i=1,\dots,n}$ and the corresponding weights $\{w_i(\mathbf{x}_0)\}_{i=1,\dots,n}$, the conditional distribution function of Y given $\mathbf{X} = \mathbf{x}_0$ is estimated as (Meinshausen 2006)

$$\hat{F}(y|\mathbf{X} = \mathbf{x}_0) = \sum_{i=1}^{n} w_i(\mathbf{x}_0) \mathbb{1}(Y_i \leq y), \tag{10}$$

where $\mathbb{1}$ is the indicator function that is equal to 1 if $Y_i \leq y$ and 0 if $Y_i > y$.

The α-quantile $Q_\alpha(\mathbf{x}_0)$ which is defined such that $\mathbb{P}(Y < Q_\alpha(\mathbf{x}_0)|\mathbf{X} = \mathbf{x}_0) = \alpha$ is estimated as follows (Meinshausen 2006): $\hat{Q}_\alpha(\mathbf{x}_0) = \inf\{y : \hat{F}(y|\mathbf{X} = \mathbf{x}_0) \geq \alpha\}$. A $100(1 - \alpha)\%$ prediction interval of Y given $\mathbf{X} = \mathbf{x}_0$ is expressed as follows:

$$\begin{aligned}&[\hat{Q}_{\alpha/2}(\mathbf{x}_0), \hat{Q}_{(1-\alpha/2)}(\mathbf{x}_0)] \\ &\quad = [\inf\{y : \hat{F}(y|\mathbf{X} = \mathbf{x}_0) \geq \alpha/2\}, \inf\{y : \hat{F}(y|\mathbf{X} = \mathbf{x}_0) \\ &\quad \geq (1 - \alpha/2)\}].\end{aligned} \tag{11}$$

Thus, QRF specifies quantiles from the outputs of the ensemble of decision trees, providing a quantification of the uncertainty associated with each prediction.

Performance criteria

KED and QRF are compared not only for their ability to accurately predict the spatially distributed variable of interest but also for their ability to deliver an accurate estimate of the associated uncertainty. The comparison is carried out using validation data, i.e., data kept aside for the whole analysis. A prediction accuracy measure helps to evaluate the overall match between observed and predicted values of the variable of interest. A prediction uncertainty accuracy measure helps to assess the overall match between expected coverage probabilities and observed coverage probabilities.

The criteria used to assess the prediction accuracy of KED and QRF are the root mean square error (RMSE), and the mean rank of each method (MR). The RMSE should be close to 0 and the MR value should be close to 1 for accurate prediction. They are computed as

$$RMSE = \sqrt{\frac{1}{m}\sum_{j=n}^{m}[\widehat{Y}_j - Y_j]^2}, \tag{12}$$

$$MR_k = \frac{1}{m}\sum_{j=1}^{m} r_{kj},\ k = 1, 2, \tag{13}$$

where $\{Y_j\}_{j=1,\ldots,m}$ are validation measurements of the variable of interest at locations $\{\mathbf{s}_j\}_{j=1,\ldots,m}$. r_{kj} is the rank of the kth method to predict the target variable at the jth validation location.

The criterion used to assess the prediction uncertainty accuracy is the goodness statistic (Deutsch 1997; Papritz and Dubois 1999; Papritz and Moyeed 2001; Goovaerts 2001; Moyeed and Papritz 2002). It consists to compare the proportion of values of a validation data set falling into the symmetric p-probability intervals (PI) computed from the conditional distribution function (cdf) of the variable of interest. By construction, there is a probability p $(0 < p < 1)$ that the true value of the variable of interest falls into a given symmetric p-interval bounded by the $(1-p)/2$ and $(1+p)/2$ quantiles of the cdf (e.g., 0.5-p-interval is bounded by lower and upper quartiles). Therefore, given validation measurements of the variable of interest $\{Y_j\}_{j=1,\ldots,m}$ at locations $\{\mathbf{s}_j\}_{j=1,\ldots,m}$, the fraction of true values falling into a given symmetric p-PI interval is computed as

$$\bar{\kappa}(p) = \frac{1}{m}\sum_{j=1}^{m}\kappa_j(p), \tag{14}$$

with

$$\kappa_j(p) = \begin{cases} 1, & \text{if } \widehat{Q}_{\frac{(1-p)}{2}}(j) < Y_j < \widehat{Q}_{\frac{(1+p)}{2}}(j) \\ 0, & \text{otherwise,} \end{cases}$$

where $\widehat{Q}_{\frac{(1-p)}{2}}(j)$ and $\widehat{Q}_{\frac{(1+p)}{2}}(j)$ are the $\frac{(1-p)}{2}$ and $\frac{(1+p)}{2}$ quantiles of the estimated cdf of the variable of interest at validation location $\mathbf{s}_j$.

The scatter plot of the estimated proportion $\bar{\kappa}(p)$ versus the expected proportion p is called "accuracy plot", and the estimated cdf is considered accurate when $\bar{\kappa}(p) > p$ for all $p \in [0, 1]$. The closeness of the estimated and theoretical proportions can be quantified using the goodness statistic (G) (Deutsch 1997):

$$G = 1 - \int_0^1 [3a(p) - 2][\bar{\kappa}(p) - p]dp, \tag{15}$$

where $a(p)$ is an indicator variable set to 1 if $\bar{\kappa}(p) > p$ and 0 otherwise.

The G-statistic corresponds to the closeness of points to the bisector of the accuracy plot. $G = 1$ for maximum goodness corresponding to the case $\bar{\kappa}(p) = p,\ \forall p \in [0, 1]$. $G = 0$ when no true values are contained in any of the PIs, i.e. $\bar{\kappa}(p) = 0,\ \forall p \in [0, 1]$. Twice more importance is given to deviations when the proportion of true values falling into the p-PI is smaller than expected $\bar{\kappa}(p) < p$. The weight $|3a(p) - 2| = 2$ rather than 1 for the accurate prediction.

Not only should the true value of the variable of interest should fall into the p-probability interval, but this interval should be as narrow as possible to reduce the uncertainty about that value. In other words, among two methods with similar goodness statistics, one would privilege the one with the smallest spread (less uncertain). Thus, a method that consistently provides narrow and accurate PIs should be preferred to a method that consistently provides wide and accurate PIs. A complimentary tool to the G-statistic is the average width of the PIs that include the true values of the variable of interest for various probabilities p. For a probability p, the average width $\bar{W}(p)$ is computed as

$$\bar{W}(p) = \frac{1}{m\bar{\kappa}(p)}\sum_{j=1}^{m}\kappa_j(p)\left[\widehat{Q}_{\frac{(1+p)}{2}}(j) - \widehat{Q}_{\frac{(1-p)}{2}}(j)\right]. \tag{16}$$

Data

It is difficult to know whether one method outperforms an another one without being able to compare the results against a ground truth. Given the inherent uncertainties of real-world data, we chose to generate synthetic data to eliminate the uncertainties inherent in real-world data. To compare the ability of KED and QRF to provide reliable predictions and prediction uncertainties, several synthetic spatial data sets with known characteristics were generated through simulations. In addition to synthetic spatial data, this study also uses real-world spatial data from a geochemical soil survey.

Synthetic spatial data

Simulated spatial data presented here do not cover all possible scenarios. However, some common situations encountered in practice are considered: (a) linearity or non-linearity between the target variable and explanatory variables; (b) strong or weak spatial correlation of the target variable; (c) presence

Table 1 Simulation ingredients

	Y	X_1 (mean = 5)			X_2 (mean = 5)			X_3 (mean = 5)			X_4 (mean = 5)			ϵ (mean = 0)		
		V	R	S	V	R	S	V	R	S	V	R	S	V	R	S
Case 1	Model 1	Cub.	20	5	Sph.	20	5	Sin.	1.47	5	Bes.	2.74	5	Nug.	–	10
Case 2	Model 1	Cub.	20	5	Sph.	20	5	Sin.	1.47	5	Bes.	2.74	5	Exp.	13.4	10
Case 3	Model 2	Cub.	20	5	Sph.	20	5	Sin.	1.47	5	Bes.	2.74	5	Nug.	–	2000
Case 4	Model 2	Cub.	20	5	Sph.	20	5	Sin.	1.47	5	Bes.	2.74	5	Exp.	13.4	2000
Case 5	Model 1	Cub.	20/3	5	Sph.	20/3	5	Sin.	1.47/3	5	Bes.	2.74/3	5	Nug.	–	10
Case 6	Model 1	Cub.	20/3	5	Sph.	20/3	5	Sin.	1.47/3	5	Bes.	2.74/3	5	Exp.	13.4/3	10
Case 7	Model 2	Cub.	20/3	5	Sph.	20/3	5	Sin.	1.47/3	5	Bes.	2.74/3	5	Nug.	–	2000
Case 8	Model 2	Cub.	20/3	5	Sph.	20/3	5	Sin.	1.47/3	5	Bes.	2.74/3	5	Exp.	13.4/3	2000

V variogram type, *R* variogram range parameter, *S* variogram sill parameter, *Cub.* cubic variogram, *Sph.* spherical variogram, *Sin.* cardinal Sine variogram, *Bes.* K-Bessel Variogram with shape parameter equal to 1, *Nug.* nugget effect variogram, *Exp.* exponential variogram

or absence of noise in the target variable; and (d) Normal or non-Normal distribution of the target variable.

Simulated spatial data are generated according to models described in Eqs. (17) and (18) and parameters given in Table 1:

$$\text{Model 1:} \quad Y(\mathbf{s}) = X_1(\mathbf{s}) + X_2(\mathbf{s}) + X_3(\mathbf{s}) + X_4(\mathbf{s}) + \epsilon(\mathbf{s}), \quad \forall \mathbf{s} \in [0, 100]^2 \tag{17}$$

$$\text{Model 2:} \quad Y(\mathbf{s}) = 1.5X_1(\mathbf{s})X_2(\mathbf{s}) + X_3(\mathbf{s})^2 + 100\sin(X_4(\mathbf{s})) + \epsilon(\mathbf{s}), \quad \forall \mathbf{s} \in [0, 100]^2, \tag{18}$$

where Y is the spatially distributed variable of interest (target variable or response variable or dependent variable); X_1, X_2, X_3, and X_4 are spatially distributed auxiliary variables (explanatory variables or covariates or dependent variables); ϵ is a spatially distributed latent variable (unobserved).

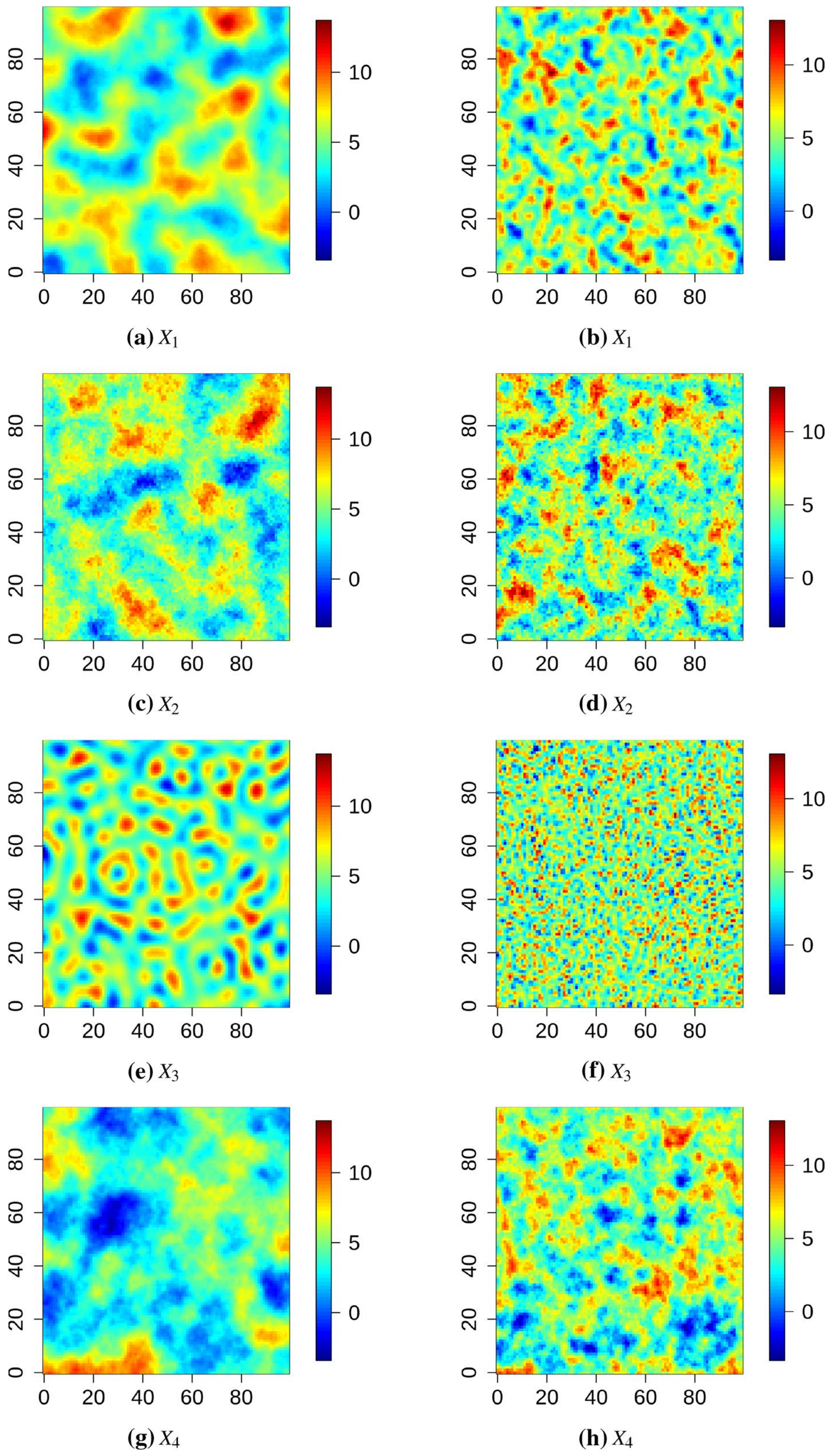
Explanatory variables are simulated on the spatial domain $[0, 100] \times [0, 100]$ as follows. They are simulated in two situations: strong and weak spatial correlation (small and large spatial correlation length). X_1 is generated according to a Gaussian random field with mean 5 and cubic isotropic stationary variogram model with sill parameter 5 and range parameter 20 (respectively, 20/3); X_2 is generated according to a Gaussian random field with mean 5 and spherical isotropic stationary variogram model with sill parameter 5 and range parameter 20 (respectively, 20/3). X_3 is generated according to a Gaussian random field with mean 5 and sine cardinal isotropic stationary variogram model with sill parameter 5 and range parameter 1.47 (respectively, 1.47/3). X_4 is generated according to a Gaussian random field with mean 5 and K-Bessel isotropic stationary variogram model with shape parameter 1, sill parameter 5, and range parameter 2.74 (respectively, 2.74/3). For background on isotropic stationary variogram models, see Chiles and Delfiner (2012). Gaussian random fields are simulated via the turning bands method implemented in RGeostats package (Renard et al. 2018).

Given models described in Eqs. (17) and (18), the latent variable is simulated on the spatial domain $[0, 100] \times [0, 100]$ as follows: (a) ϵ is generated with respect to a Gaussian random field with zero mean and nugget effect model with sill parameter 10 (respectively, 2000) and (b) ϵ is generated with respect to a Gaussian random field with zero mean and exponential isotropic stationary variogram model with sill parameter 10 (respectively, 2000) and range parameter 13.4 (respectively, 13.4/3).

In total, eight simulation cases are considered, as described in Table 1. The first four cases correspond to the situation, where the target variable shows a strong spatial correlation (large spatial correlation length). In the first case, there is a linear relationship between the target variable and auxiliary variables, and the target variable shows a strong spatial correlation with a nugget effect. In the second case, there is a linear relationship between the target variable and auxiliary variables, and the target variable presents a strong spatial correlation without a nugget effect. In the third case, there is a non-linear relationship between the target variable and auxiliary variables, and the target variable exhibits a strong spatial correlation with a nugget effect. In the fourth case, there is a non-linear relationship between the target variable and auxiliary variables, and the target variable exhibits a strong spatial correlation without a nugget effect.

The last four cases correspond to the situation, where the target variable shows a weak spatial correlation (small spatial correlation length). In the fifth case, there is a linear relationship between the target variable and auxiliary variables, and the target variable shows a weak spatial correlation with a nugget effect. In the sixth case, there is a linear relationship between the target variable and auxiliary variables, and the target variable presents a weak spatial correlation without a nugget effect. In the seventh case, there

Fig. 1 Example of simulated explanatory variables for: **a**, **c**, **e**, and **g** cases 1, 2, 3, and 4 (strong spatial correlation) and **b**, **d**, **f**, **h** cases 5, 6, 7, and 8 (weak spatial correlation)



(a) X_1

(b) X_1

(c) X_2

(d) X_2

(e) X_3

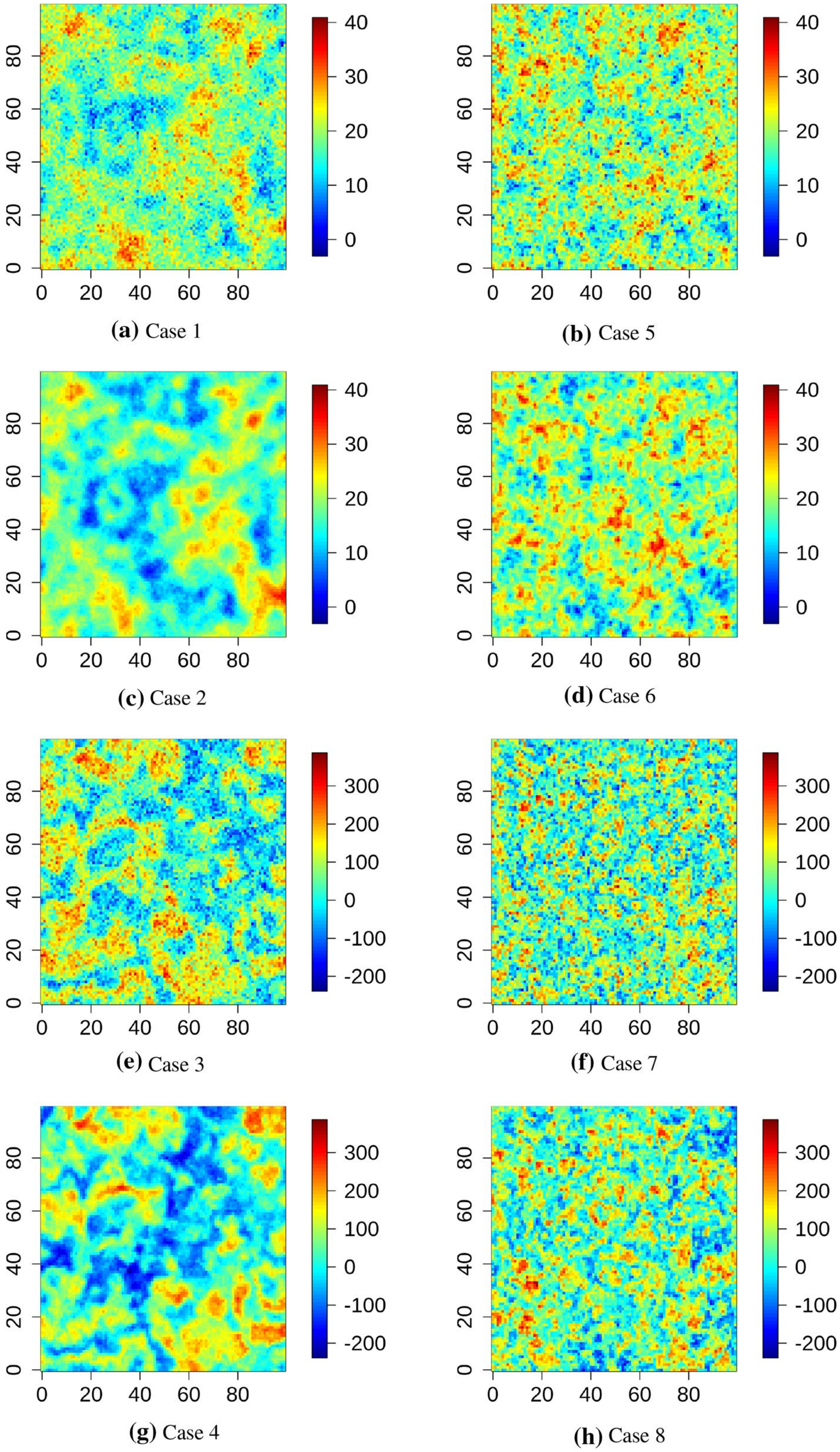
(f) X_3

(g) X_4

(h) X_4

is a non-linear relationship between the target variable and auxiliary variables, and the target variable exhibits a weak spatial correlation with a nugget effect. In the last case, there is a non-linear relationship between the target variable and auxiliary variables, and the target variable exhibits a weak spatial correlation without a nugget effect.

Fig. 2 Example of the simulated target variable for each case: **a**, **c**, **e**, **g** strong spatial correlation and **b**, **d**, **f**, **h** weak spatial correlation



(a) Case 1

(b) Case 5

(c) Case 2

(d) Case 6

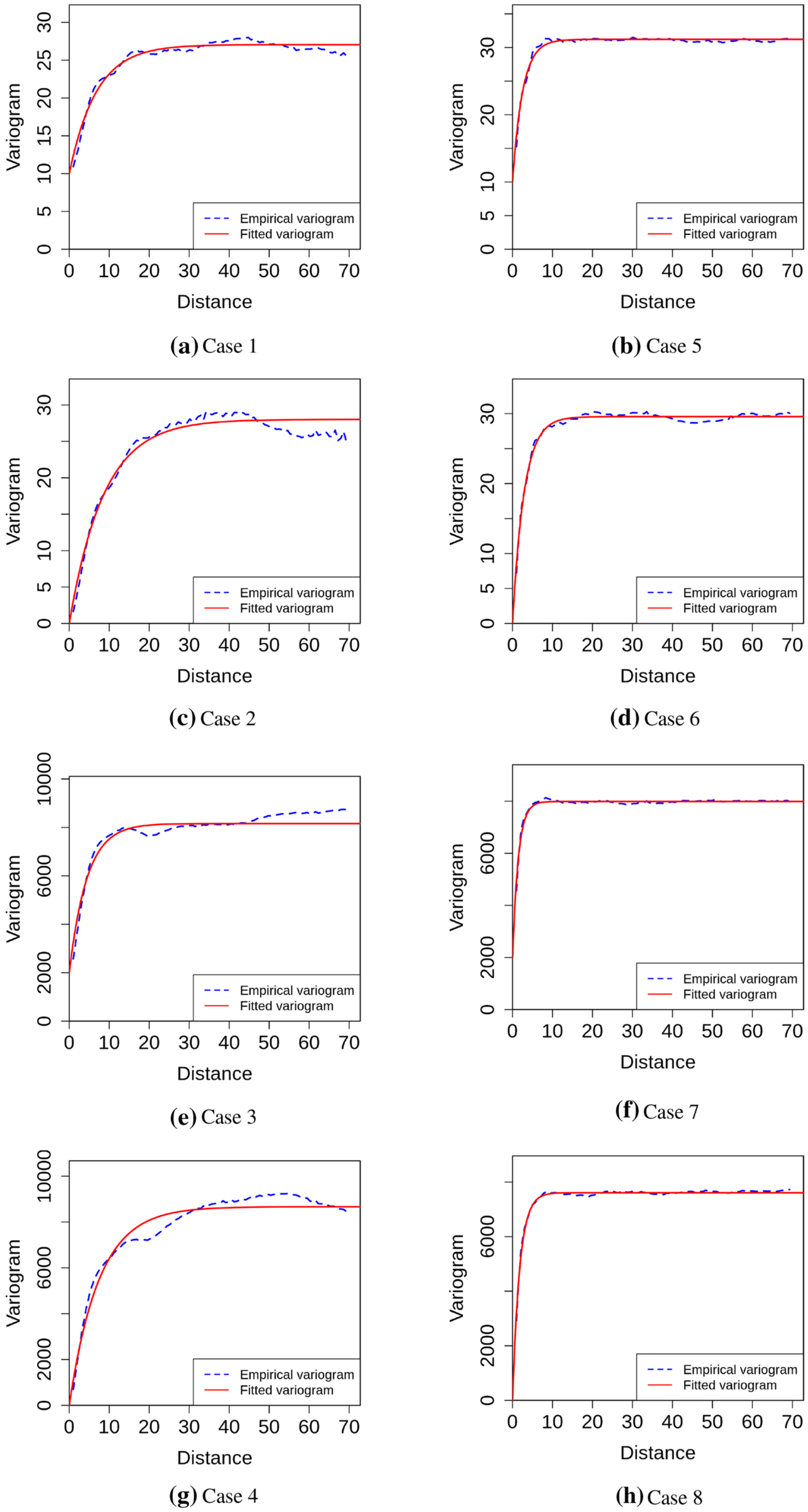
(e) Case 3

(f) Case 7

(g) Case 4

(h) Case 8

Under cases 1, 2, 5, and 6 (Table 1), the target variable is normally distributed, while in cases 3, 4, 7, and 8 (Table 1), the target variable is non-Normal. Figure 1 shows one simulation of explanatory variables over a 100 x 100 regular grid. A representation of one simulated target variable in each case is given in Fig. 2. Figure 3 shows the variogram associated with one simulated target variable in each case. Figure 4 presents the histogram and Normal QQ plot of one simulated



(a) Case 1

(b) Case 5

(c) Case 2

(d) Case 6

(e) Case 3

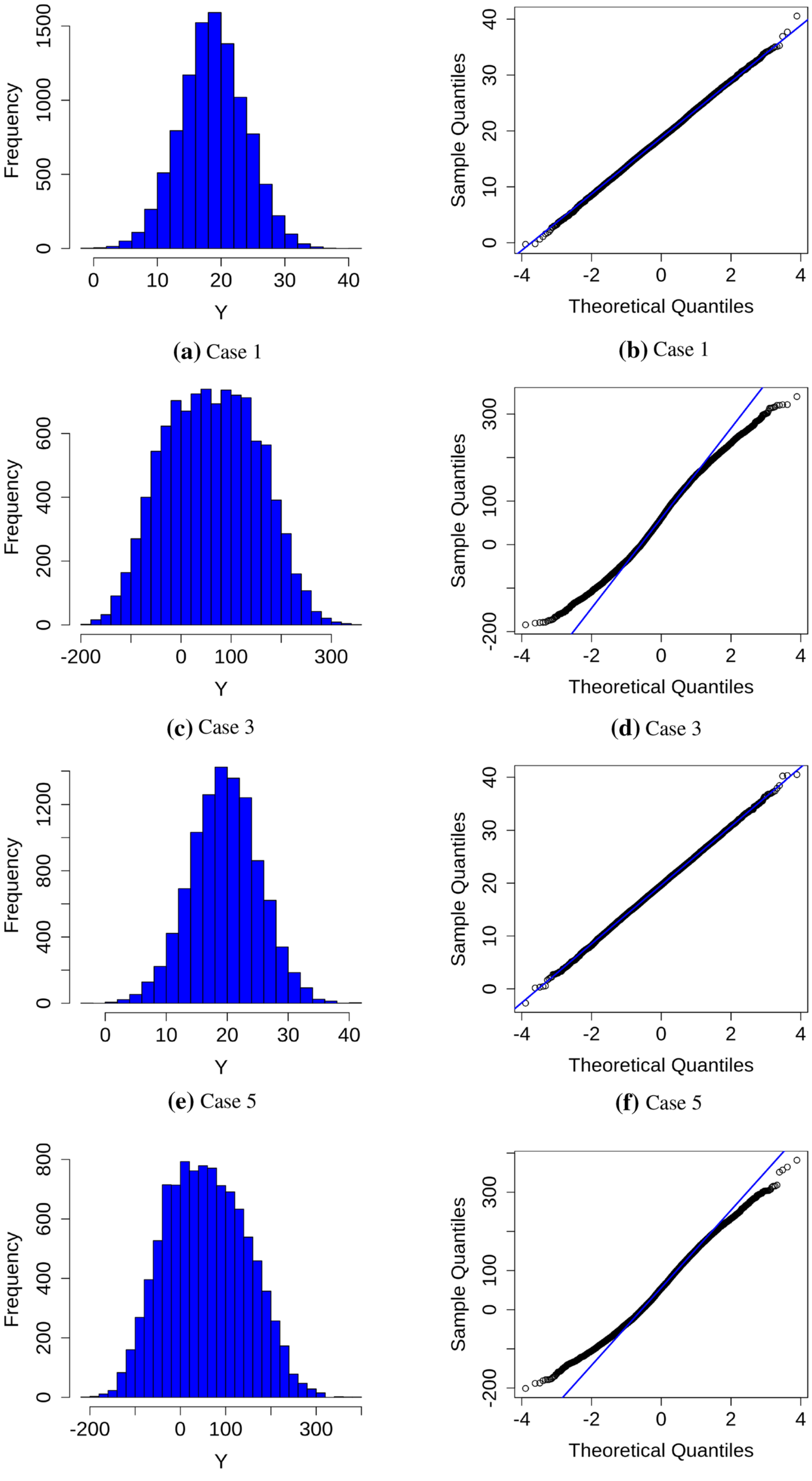
(f) Case 7

(g) Case 4

(h) Case 8

Fig. 3 Example of the variogram of the simulated target variable for each case: **a**, **c** ,**e**, **g** strong spatial correlation and **b**, **d**, **f**, **h** weak spatial correlation

Fig. 4 Example of the histogram and Normal QQ plot of the simulated target variable for **a**, **b** case 1 (Normal distribution); **c**, **d** case 3 (non-Normal distribution); **e**, **f** case 5 (Normal distribution); and **g**, **f** case 7 (non-Normal distribution)



(a) Case 1

(b) Case 1

(c) Case 3

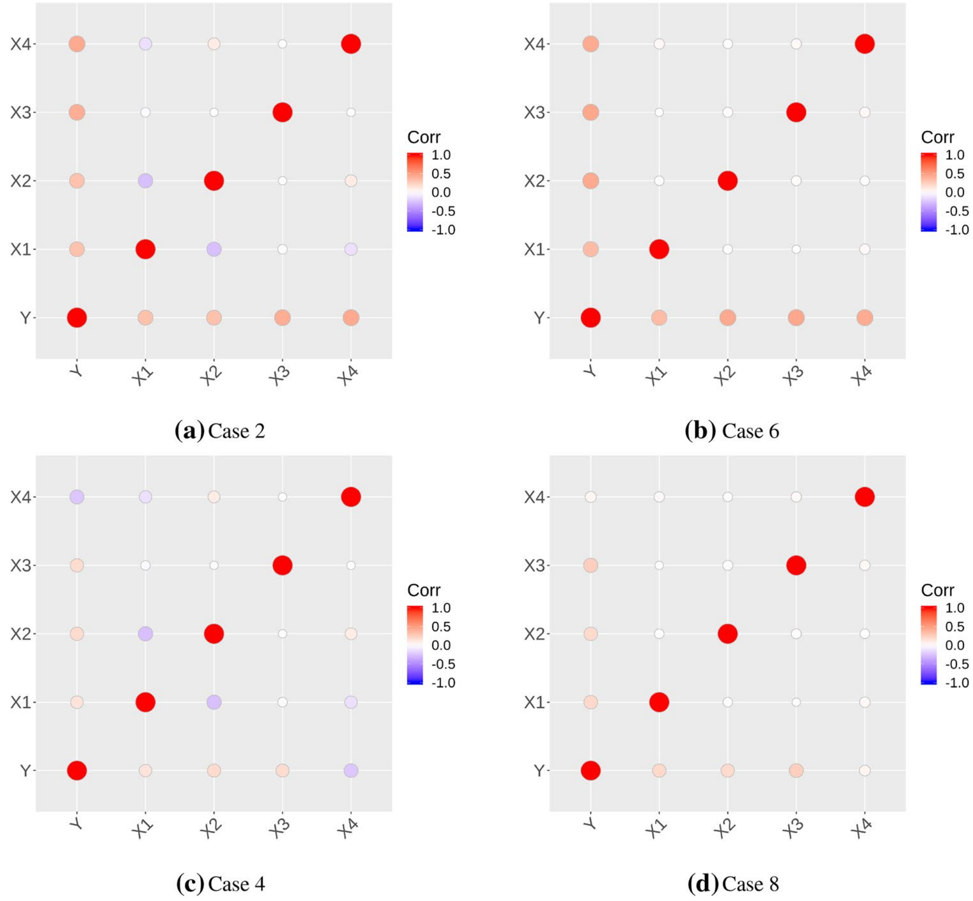
(d) Case 3

(e) Case 5

(f) Case 5

(g) Case 7

(h) Case 7



(a) Case 2

(b) Case 6

(c) Case 4

(d) Case 8

Fig. 5 Example of the correlation plot between the simulated target variable (Y) and simulated explanatory variables (X_1, X_2, X_3, X_4) for **a** case 2 and **b** case 6 (linear relationship); **c** case 4 and **d** case 8 (non-linear relationship)

target variable. Figure 5 provides one example of correlation plot between the target variable and explanatory variables.

Real-world spatial data

The real-world spatial data used in this study are derived from samples collected across south west England by the British Geological Survey (Kirkwood et al. 2016). There are 50 target variables which are element concentrations (in mg/kg) measured at 568 locations. The available auxiliary information comprise magnetic data, radiometric data, gravity data, Landsat data, elevation data, and their derivatives, in total 26 covariates. A detailed description of the region of interest and data can be found in Kirkwood et al. (2016). We focus on two target variables Ba (Barium) and Tl (Thalium). Figure 6 shows measurements of these two variables as well as their corresponding histogram and variogram. Some covariates are depicted in Fig. 7. Figure 8 provides the correlation plot between the two target variables and some explanatory variables.

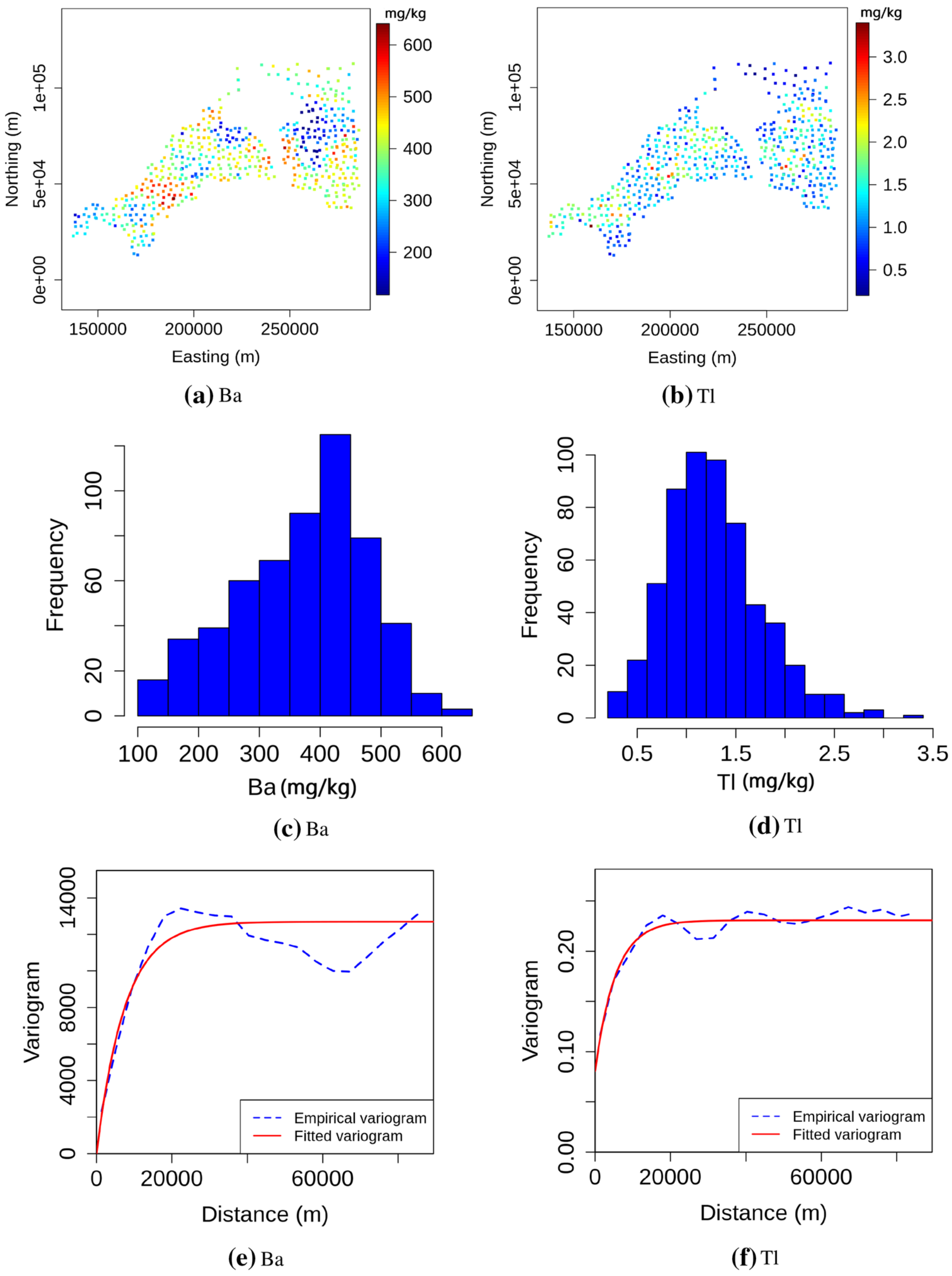
Results and discussion

This section juxtaposes results derived by KED and QRF on synthetic and real-world spatial data, as described in “Data” section.

Simulated data example

The process of generation of synthetic spatial data described in “Synthetic spatial sata” section is repeated one hundred times. Thus, one hundred independent realizations are generated in the same way as the realization, as depicted in Figs. 1 and 2. For each realization, a training data set of $n = 1000$ observations sampled randomly is formed. Thus, the remainder of data ($m = 9000$ observations) is set aside for the validation. KED and QRF models are trained on each of these 100 training data sets and their performances are evaluated on each of these 100 validation data sets. QRF is designed as follows. It contains 1001 decision trees—a sufficient number to allow convergence of error to a stable minimum. The odd number of decision trees prevents possible ties in variable importance. Each decision tree is grown until the terminal nodes contained 8 observations to reduce over-fitting to outliers. Geographical coordinates are

Fig. 6 **a**, **b** Spatial distribution maps; **c**, **d** histograms and **e**, **f** variograms of the two target variables. Variograms are fitted using an exponential model with or without nugget effect



(a) Ba

(b) Tl

(c) Ba

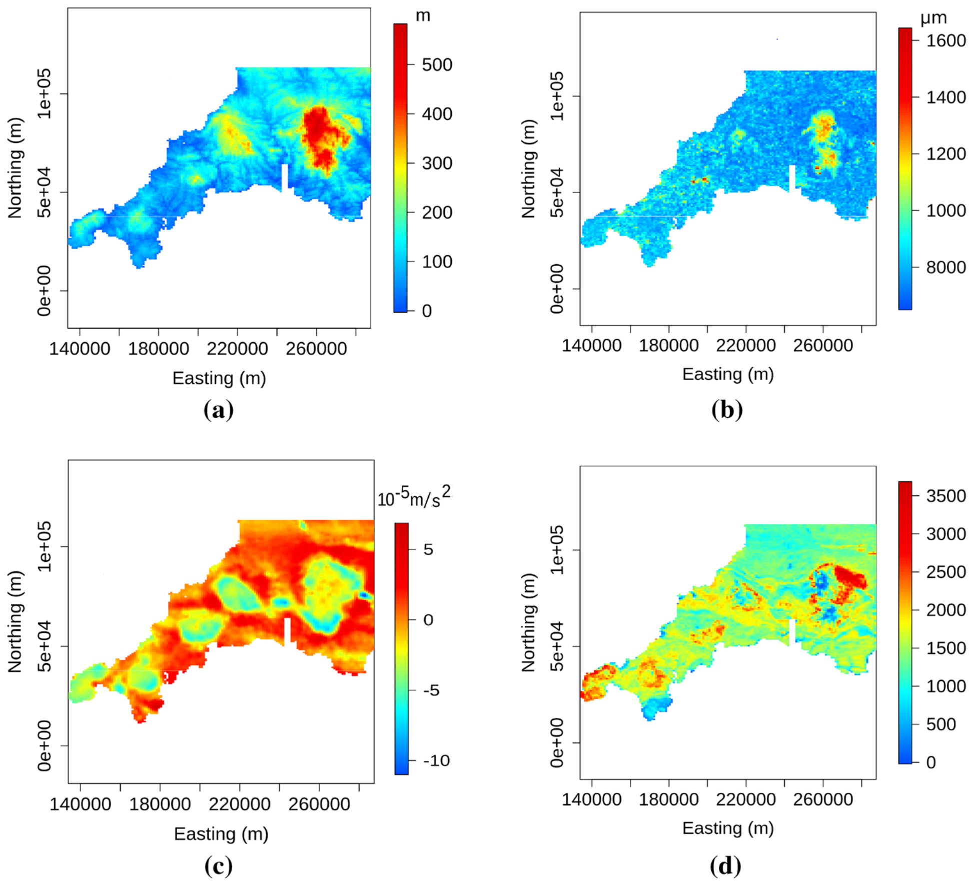
(d) Tl

(e) Ba

(f) Tl

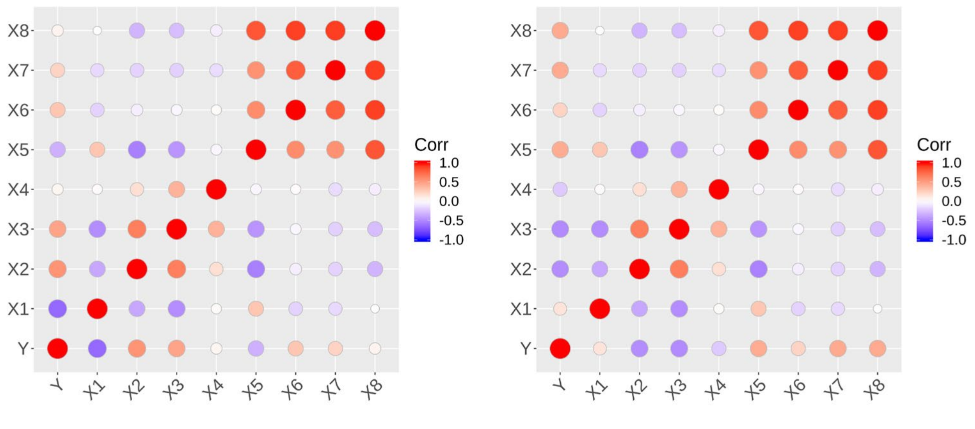
considered as auxiliary variables. Results from the first four cases, where the variable of interest exhibits a strong spatial correlation (large spatial correlation length), are provided in Figs. 9 and 10. Results from the last four cases, where the variable of interest shows a weak spatial correlation (small spatial correlation length), are given in Figs. 11 and 12.

In the first case, there is a linear relationship between the target variable and auxiliary variables, and the target variable shows a strong spatial correlation with a nugget effect. Results from this case show that KED outperforms QRF in terms of prediction accuracy and prediction uncertainty accuracy. Figure 9a, b presents Box-plot statistics of the RMSE and MR for each method. There is a marked difference between KED and QRF methods, the former giving the best prediction performance. With regard to the modeling prediction uncertainty using goodness statistic, accuracy plot, and probability interval width plot, Fig. 10a–c shows that KED performed better in terms of the modeling prediction uncertainty than QRF. Similar findings are obtained in the second case, where there is a linear relationship between the target variable and auxiliary variables, and the target variable shows a strong spatial correlation without a nugget effect (Figs. 9c, d, 10d–f). Moreover, the difference between KED and QRF is more marked in the second case than the



(a) (b) (c) (d)

Fig. 7 Spatial distribution map of some explanatory variables: **a** elevation, **b** landsat 8 band 4, **c** gravity survey high-pass-filtered Bouguer anomaly, and **d** total count of unmixed gamma ray signal



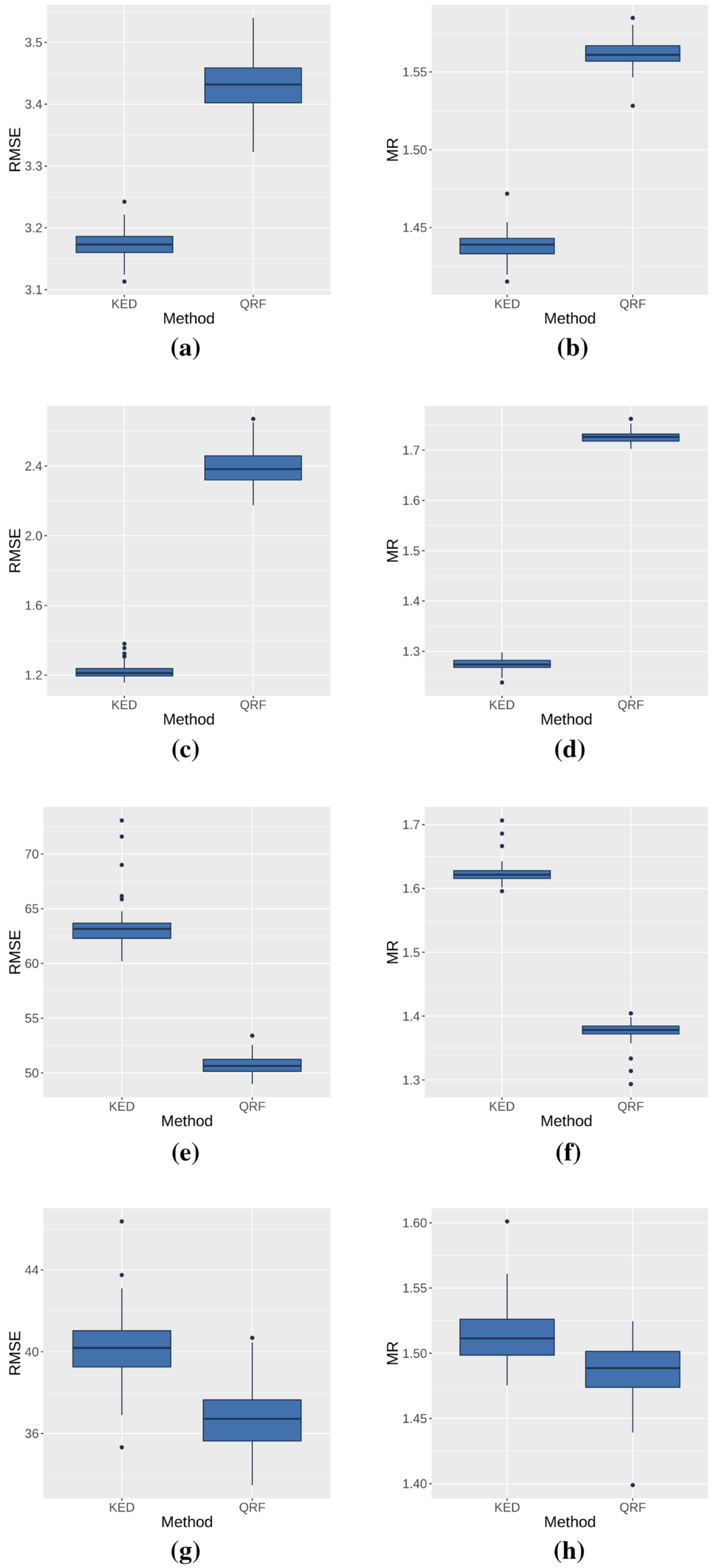
(a) Ba **(b)** Tl

Fig. 8 Correlation plot between **a** Ba (respectively, **b** Tl) and explanatory variables. *Y* (Ba, respectively, Tl) X_1 (Elevation), X_2 (gravity survey high-pass-filtered Bouguer anomaly), X_3 (gravity survey Bouguer anomaly), X_4 (reduction to the pole of TMI), X_5 (uranium counts from gamma ray spectrometry), X_6 (thorium counts from gamma ray spectrometry), X_7 (potassium counts from gamma ray spectrometry), X_8 (total count of unmixed gamma ray signal)

first case, both in terms of prediction accuracy and prediction uncertainty accuracy.

In the third case, where there is a non-linear relationship between the target variable and auxiliary variables, and the target variable exhibits a strong spatial correlation with a nugget effect, results indicate that QRF outperforms KED in terms of prediction accuracy but not in terms of prediction uncertainty accuracy. Figure 9e, f exhibits the distribution of RMSE and MR for each method. KED and QRF differ notably, the latter giving the best prediction performance. When considering the goodness statistic, the accuracy plot, and the probability interval width plot (Fig. 10g–i), KED shows better performance in prediction uncertainty than QRF. The KED accuracy plot is generally close to the 1:1 line, while QRF accuracy plot is above the 1:1 line with medium *p* value points further than the extreme *p* values one. This shows an overestimation of uncertainty with QRF. The KED probability interval width

Fig. 9 KED and QRF in terms of prediction accuracy in case 1 (1st row), case 2 (2nd row), case 3 (3rd row), and case 4 (4th row): **a**, **c**, **e**, **g** RMSE, and **b**, **d**, **f**, **h** MR



(a)

(b)

(c)

(d)

(e)

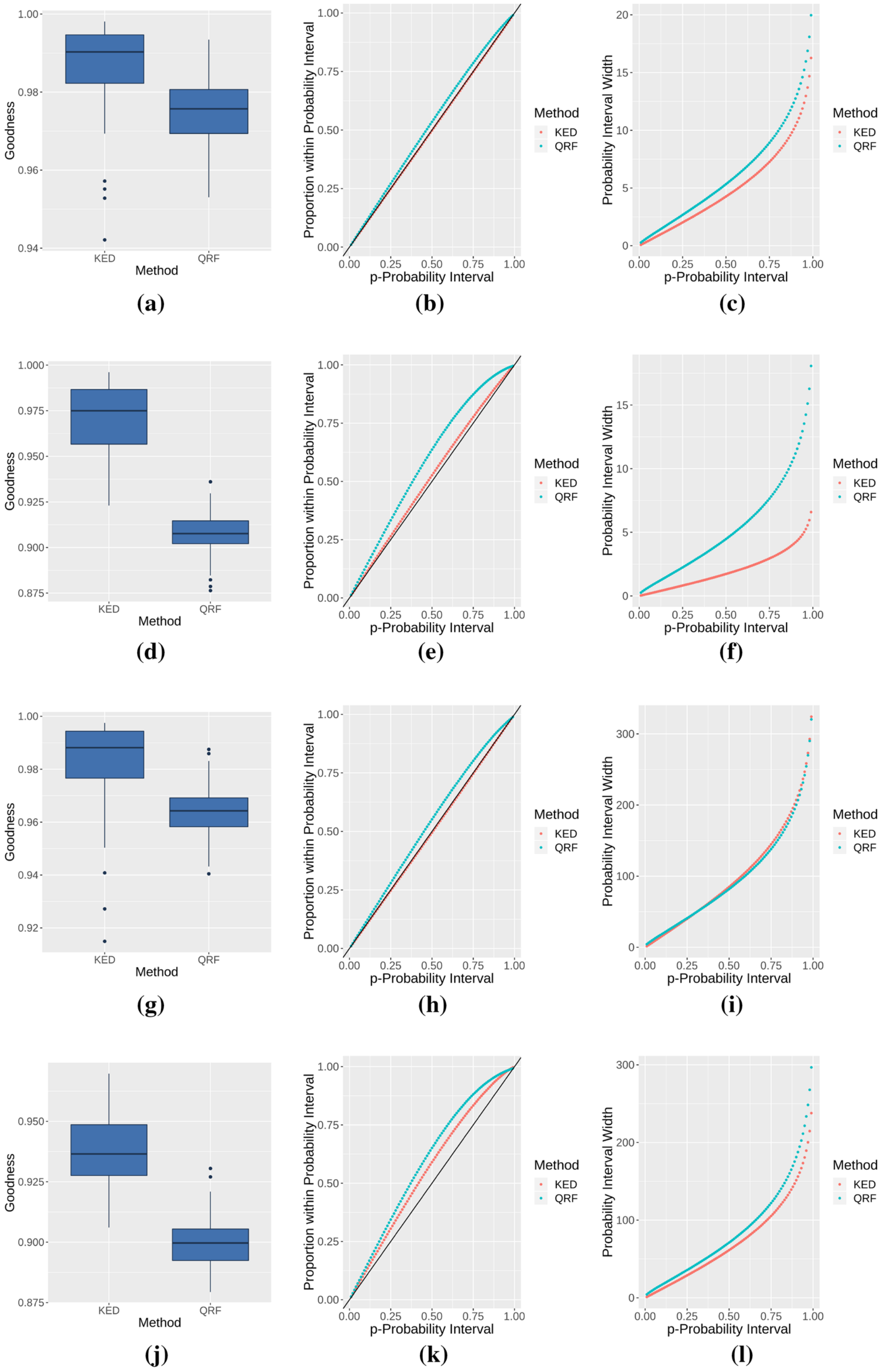
(f)

(g)

(h)

plot is below the QRF probability interval width plot. Similar results are observed in the fourth case, where there is a non-linear relationship between the target variable and auxiliary variables, and the target variable exhibits a strong spatial correlation without a nugget effect (Figs. 9g, h, 10j–l). In addition, the difference between QRF and KED

Fig. 10 KED and QRF in terms of prediction uncertainty accuracy in case 1 (1st row), case 2 (2nd row), case 3 (3rd row), and case 4 (4th row): **a**, **d**, **g**, **j** goodness statistic, **b**, **e**, **h**, **k** accuracy plot, and **c**, **f**, **i**, **l** probability interval width plot



(a) (b) (c)

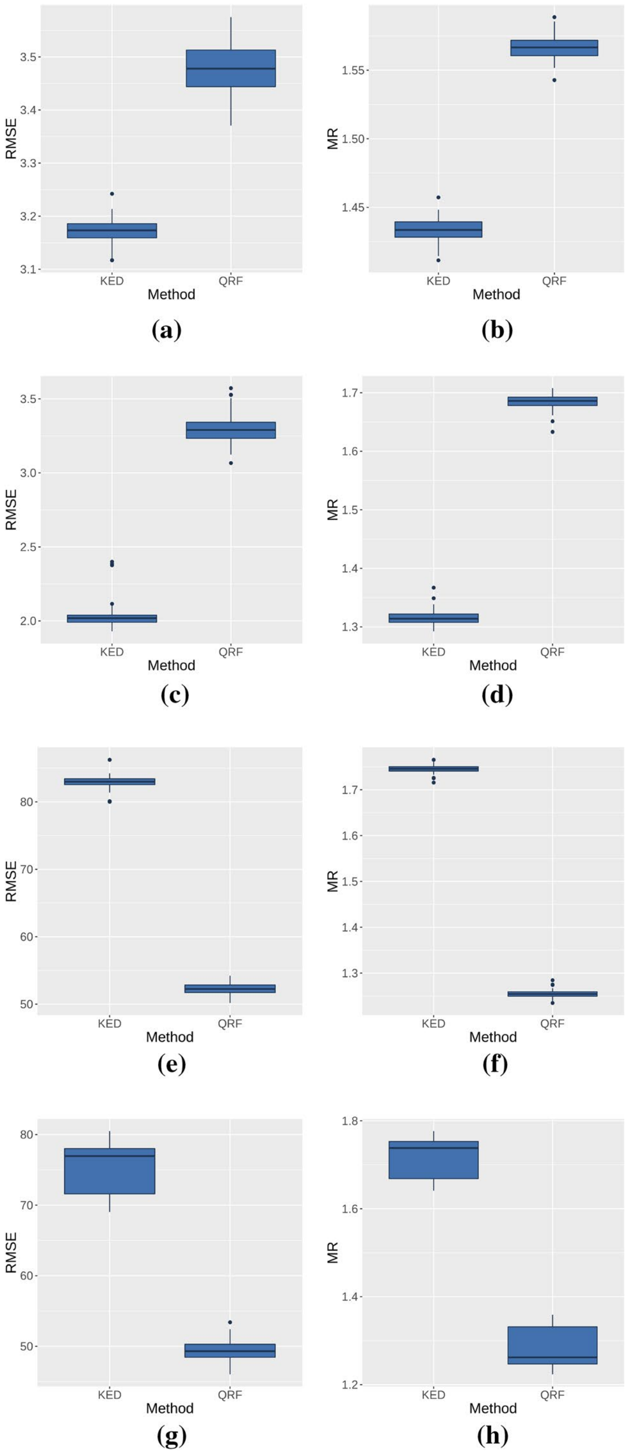
(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

is more marked (resp. less marked) in the third case than the fourth case in terms of prediction accuracy (resp. prediction uncertainty accuracy).

In the fifth case, where there is a linear relationship between the target variable and auxiliary variables, and the target variable shows a weak spatial correlation with a nugget effect, results are similar to ones in the first case (Figs. 11a, b, 12a–c). Likewise, in the sixth case, where there is a linear relationship between the target variable and auxiliary variables, and the target variable shows a weak spatial correlation without a nugget effect, results are similar to ones in the second case (Figs. 11c, d, 12d–f).



(a) (b) (c) (d) (e) (f) (g) (h)

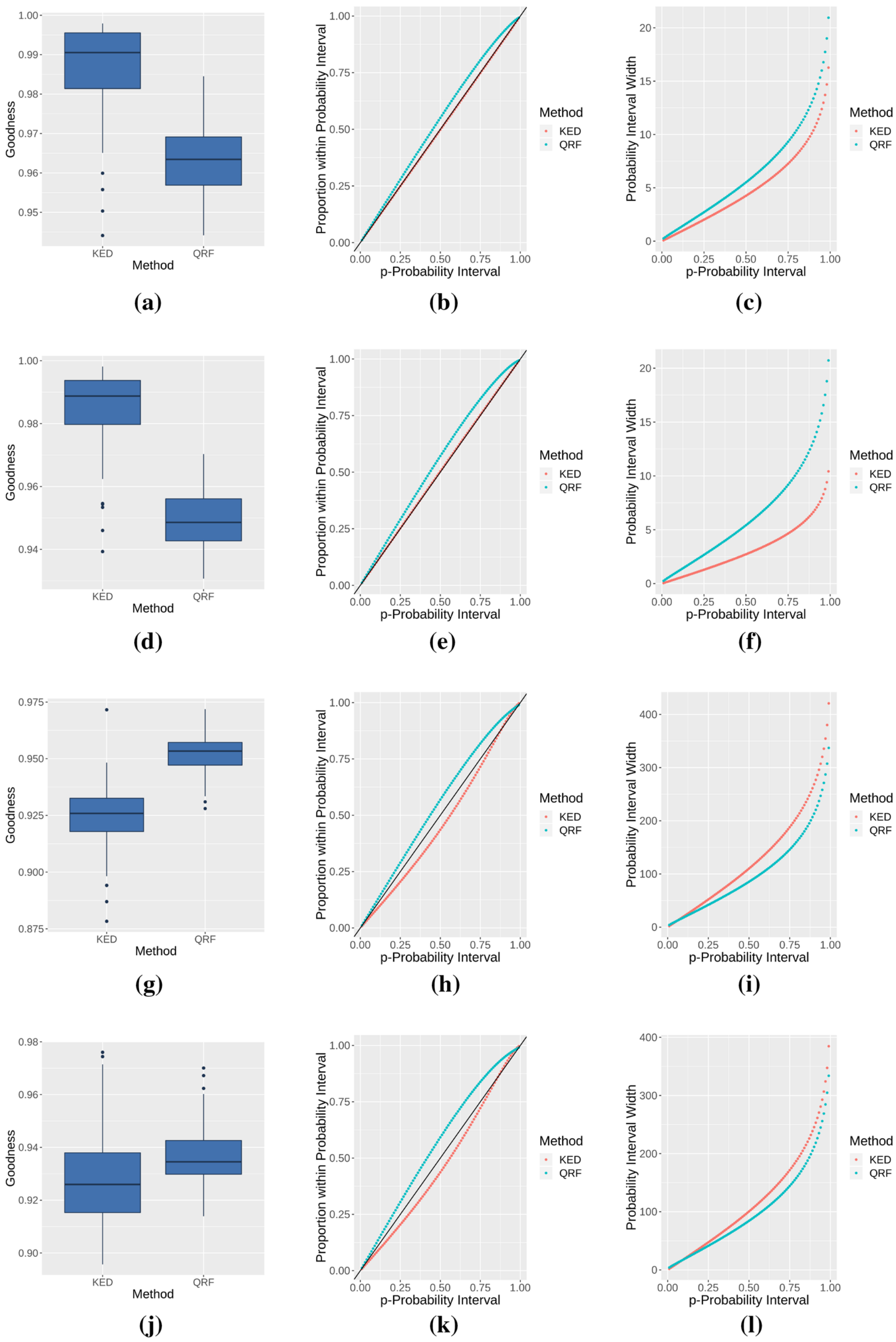
Fig. 11 KED and QRF in terms of prediction accuracy in case 5 (1st row), case 6 (2nd row), case 7 (3rd row), and case 8 (4th row): **a**, **c**, **e**, **g** RMSE, and **b**, **d**, **f**, **h** MR

In the seventh case, where there is a non-linear relationship between the target variable and auxiliary variables, and the target variable shows a weak spatial correlation with a nugget effect, results show that QRF outperforms KED in terms of prediction accuracy and prediction uncertainty accuracy (Figs. 11e, f, 12g–i). In particular, the KED accuracy plot is below the bisector line with medium p value points further than the extreme p values one. This shows an underestimation of uncertainty by KED. A wider range of PIs is observed for KED than QRF according to probability interval width plots. Similar findings are obtained in the last case with a non-linear relationship between the target variable and auxiliary variables, and where the target variable presents a weak spatial correlation without a nugget effect (Figs. 11g, h, 12j–l).

Figure 13 shows one variogram of validation errors for each case under KED and QRF. In the first and fifth cases, errors show no spatial correlation both in KED and QRF. Thus, in these cases with a linear relationship between the target variable and auxiliary variables, and the target variable shows a strong or weak spatial correlation with a nugget effect, both KED and QRF are optimal. In the second and sixth cases, errors show a very weak spatial correlation in KED, while they exhibit a strong spatially correlation in QRF. Thus, in these cases, where we have a linear relationship between the target variable and auxiliary variables, and the target variable shows a strong or weak spatial correlation without a nugget effect, QRF is sub-optimal. In the third and seventh cases, errors show a very weak spatial correlation in KED, while they exhibit no spatial correlation in QRF. Thus, in the these cases, where we have a non-linear relationship between the target variable and auxiliary variables, and the target variable exhibits a strong or weak spatial correlation with a nugget effect, QRF is optimal. In the fourth and eighth cases, errors exhibit a very weak spatial correlation in KED, while they show strong spatially correlation in QRF. Hence, QRF is sub-optimal in these cases, where we have a non-linear relationship between the target variable and auxiliary variables, and the target variable exhibits a strong or weak spatial correlation without a nugget effect.

Geochemical data example

To assess the predictive ability of KED and QRF in this data set, a pseudo-cross validation is considered instead of an external validation due to the relatively small size of the data set. The pseudo-cross validation consists in leaving out a randomly selected 10% of observations (∼ 57 observations), and predict the remaining observations at those locations set aside for validation. This procedure is repeated 500 times. In KED, the prediction at each iteration is carried out using the same variogram of residuals fitted on the whole set of observations. QRF is built similarly as described in the simulation study “Performance criteria” section. Thus, it contains 1001 decision trees, enough to allow convergence of error to a stable minimum. The odd number of decision trees prevents possible ties in variable importance. To reduce over-fitting



(a) (b) (c)

(d) (e) (f)

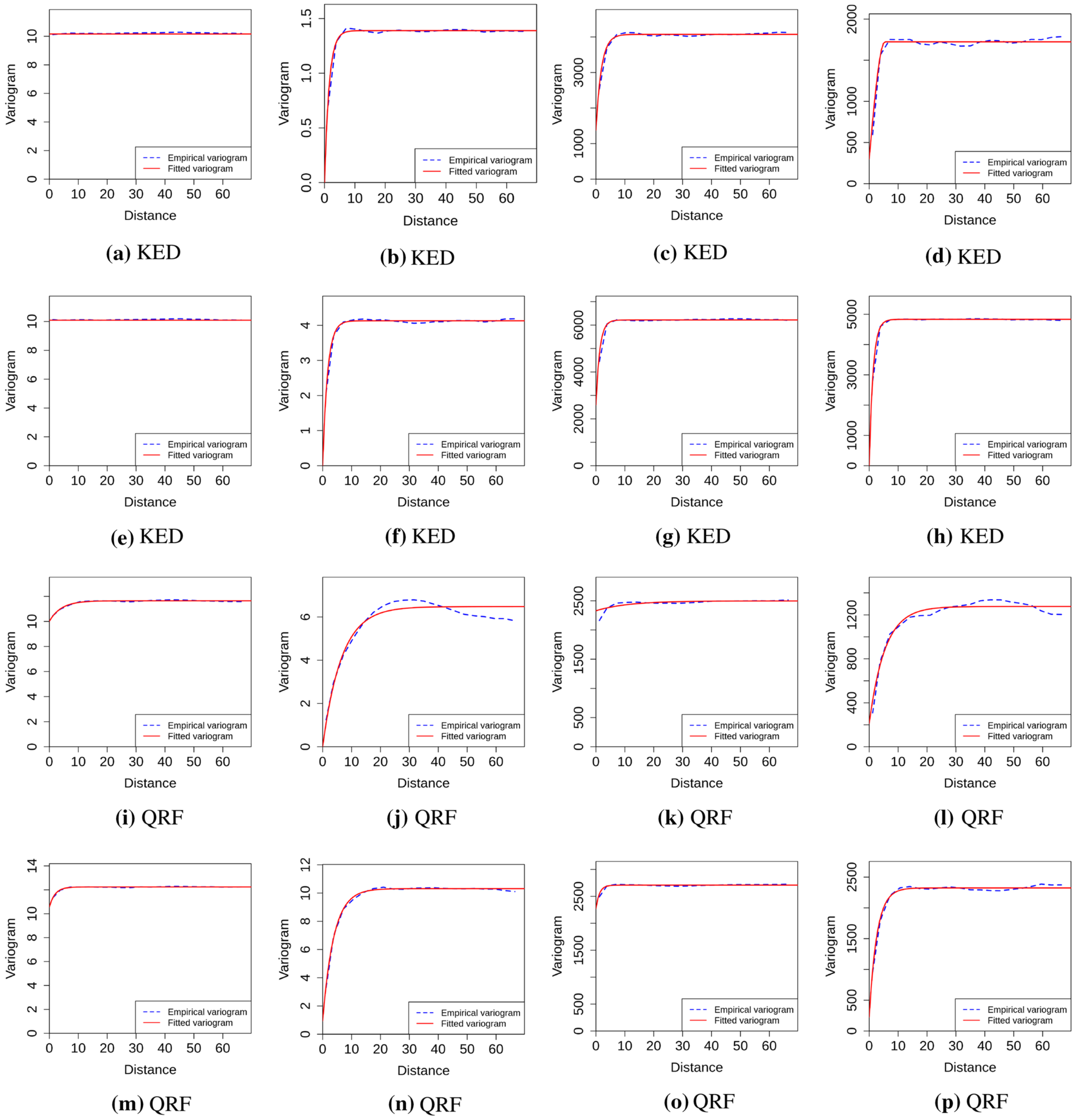
(g) (h) (i)

(j) (k) (l)

Fig. 12 KED and QRF in terms of prediction uncertainty accuracy in case 5 (1st row), case 6 (2nd row), case 7 (3rd row), and case 8 (4th row): **a**, **d**, **g**, **j** goodness statistic, **b**, **e**, **h**, **k** accuracy plot, and **c**, **f**, **i**, **l** probability interval width plot

to outliers, each decision tree is grown until the terminal nodes contained 8 observations. Geographical coordinates are accounted as auxiliary variables.

Figures 14 and 15 present the predictive performance for the two target variables. For each target variable, KED and QRF are fairly similar in terms of prediction accuracy. Regarding the prediction uncertainty accuracy, KED gives slightly better results than QRF for Ba concentration and performs similar to QRF for Tl concentration. Figures 18 and 19 show prediction and prediction uncertainty maps, respectively, for Ba and Tl. The prediction uncertainty map corresponds to the width of the 95% prediction interval. For each target variable, the overall appearance of the prediction maps associated with each method is very



(a) KED (b) KED (c) KED (d) KED

(e) KED (f) KED (g) KED (h) KED

(i) QRF (j) QRF (k) QRF (l) QRF

(m) QRF (n) QRF (o) QRF (p) QRF

Fig. 13 Example of variogram of KED validation errors (1st and 2nd rows) and QRF validation errors (3rd and 4th rows) for each case: **a**, **i** case 1, **b**, **j** case 2, **c**, **k** case 3, **d**, **l** case 4, **e**, **m** case 5, **f**, **n** case 6, **g**, **o** case 7, and **h**, **p** case 8. Variograms are fitted using either the nugget effect model either an exponential model with or without nugget effect

similar. However, the general appearance of the prediction uncertainty maps associated with each method differs notably.

Prediction uncertainty maps provided by KED vary much less across the study area compared to the ones provided by QRF. The largest prediction uncertainties given by KED are concentrated in those areas not surveyed or where the sampling was too sparse. Thus, the prediction uncertainty provided by KED depends mainly on the data configuration. Prediction uncertainty maps provided by QRF show spatial patterns which are not related to the density of sampling locations but rather to the distribution map of some auxiliary variables. In particular, prediction uncertainty maps of Ba and Tl are most strongly

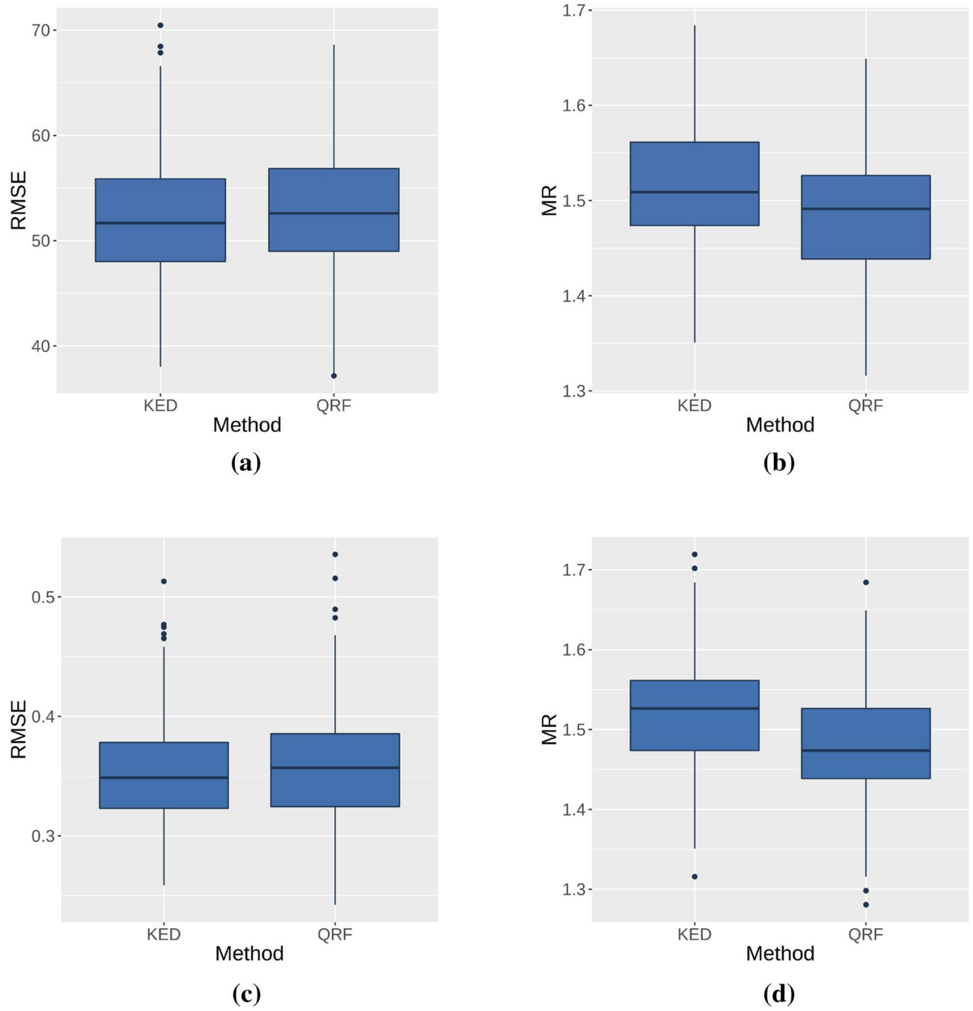


Fig. 14 KED and QRF in terms of prediction accuracy for Ba concentration (top) and Tl concentration (bottom): **a**, **c** RMSE and **b**, **d** MR

correlated to the gravity survey high-pass filter Bouger anomaly presented in Fig. 7, with a Spearman correlation coefficient of −0.49 and −0.42, respectively. This latter auxiliary variable is the most important variable in the trained QRF models.

Figure 16 displays one variogram of pseudo-cross-validation errors for Ba and Tl concentrations under KED and QRF. One can see that errors show no spatial correlation. Figure 17 presents one histogram and Normal QQ plot of KED pseudo-cross-validation errors for Ba and Tl concentrations. It appears that errors are relatively symmetrical and globally less deviated from the Gaussian distribution.

Discussion

Results from the simulation study showed that the superiority of KED over QRF can be expected when there is a linear relationship between the variable of interest and auxiliary variables, and the variable of interest shows a strong or weak spatial correlation. In other hand, the superiority of QRF over KED can be expected when there is a non-linear relationship between the variable of interest and auxiliary variables, and the variable of interest exhibits a weak spatial correlation. Moreover, when there is a non-linear relationship between the variable of interest and auxiliary variables, and the variable of interest presents a strong spatial correlation, one can expect QRF outperforms KED in terms of prediction accuracy but not in terms of prediction uncertainty accuracy.

The results of this comparison point out that a non-parametric regression method which has good prediction performance in a non-linear framework does not necessary have a good prediction performance in linear framework. The inability of QRF to provide reliable prediction uncertainties in the context of a strong spatial correlation with the variable of interest could be explained by the fact that under the QRF approach, the spatial dependency of data is ignored. In essence, QRF is a non-spatial method that ignores the general sampling pattern in the training of the QRF model when applied to spatial prediction. This can potentially lead to sub-optimal predictions, especially where the target variable exhibits a strong spatial correlation and where point patterns

Fig. 15 KED and QRF in terms of prediction uncertainty accuracy for Ba concentration (top) and Tl concentration (bottom): **a**, **d** goodness statistic, **b**, **e** accuracy plot, and **c**, **f** probability interval width plot

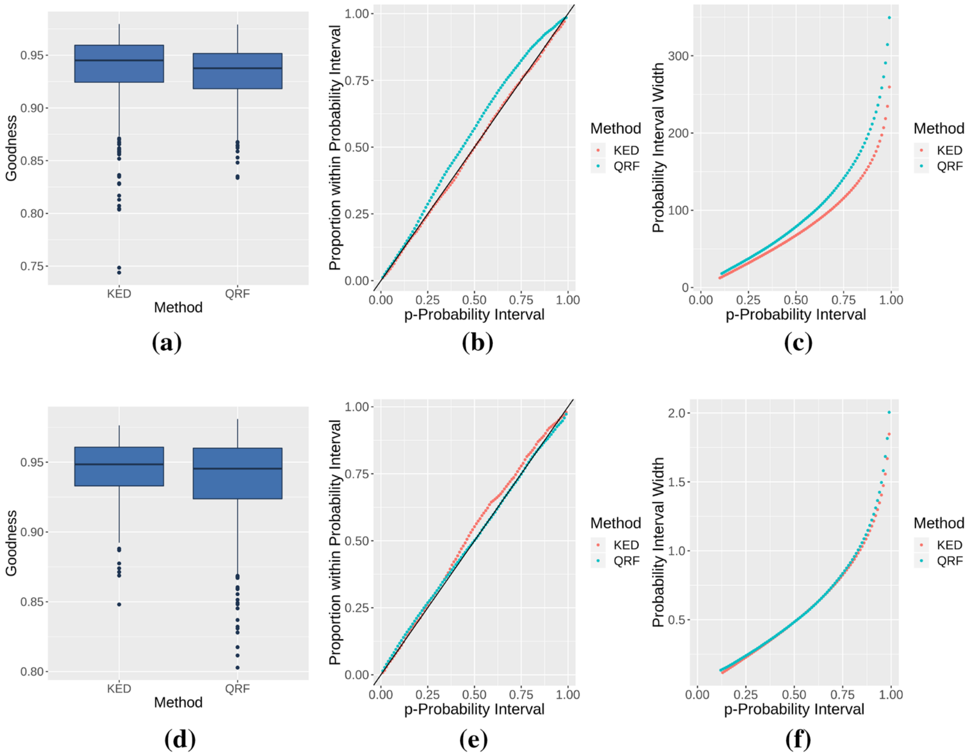


Fig. 16 Variogram of **a**, **c** KED and **b**, **d** QRF pseudo-cross-validation errors for Ba and Tl concentrations. Variograms are fitted using nugget effect or/and spherical models

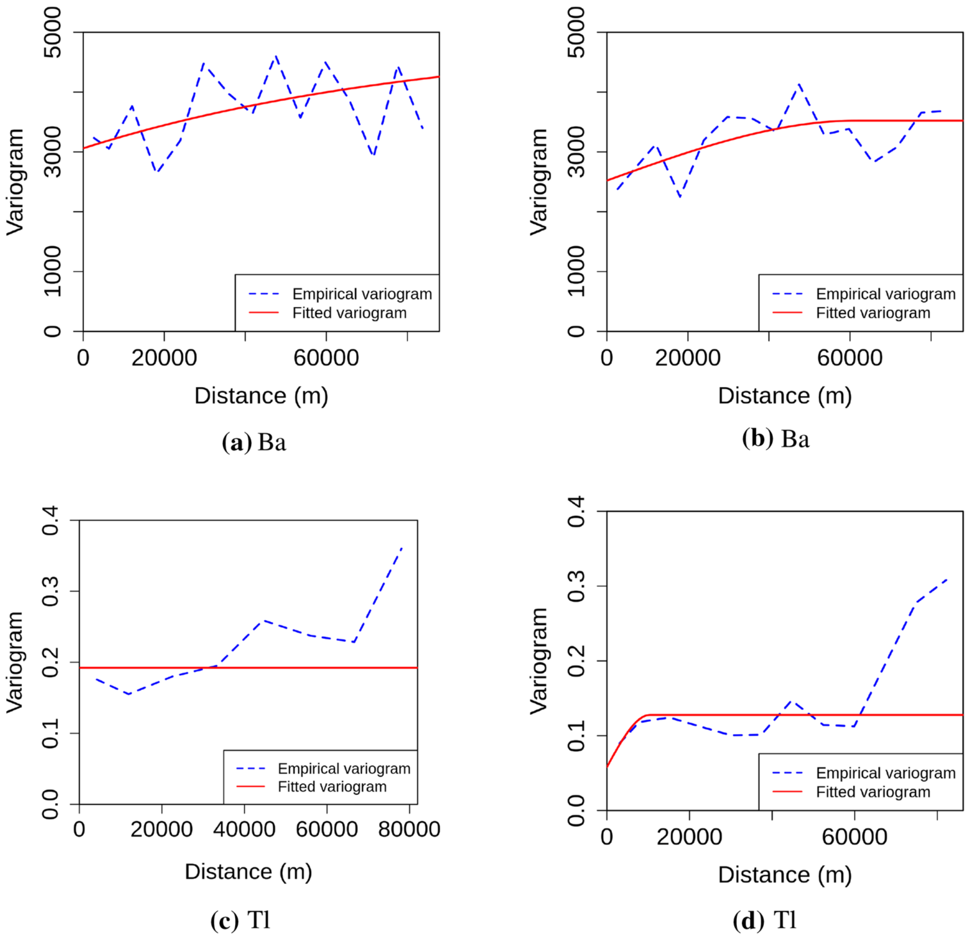
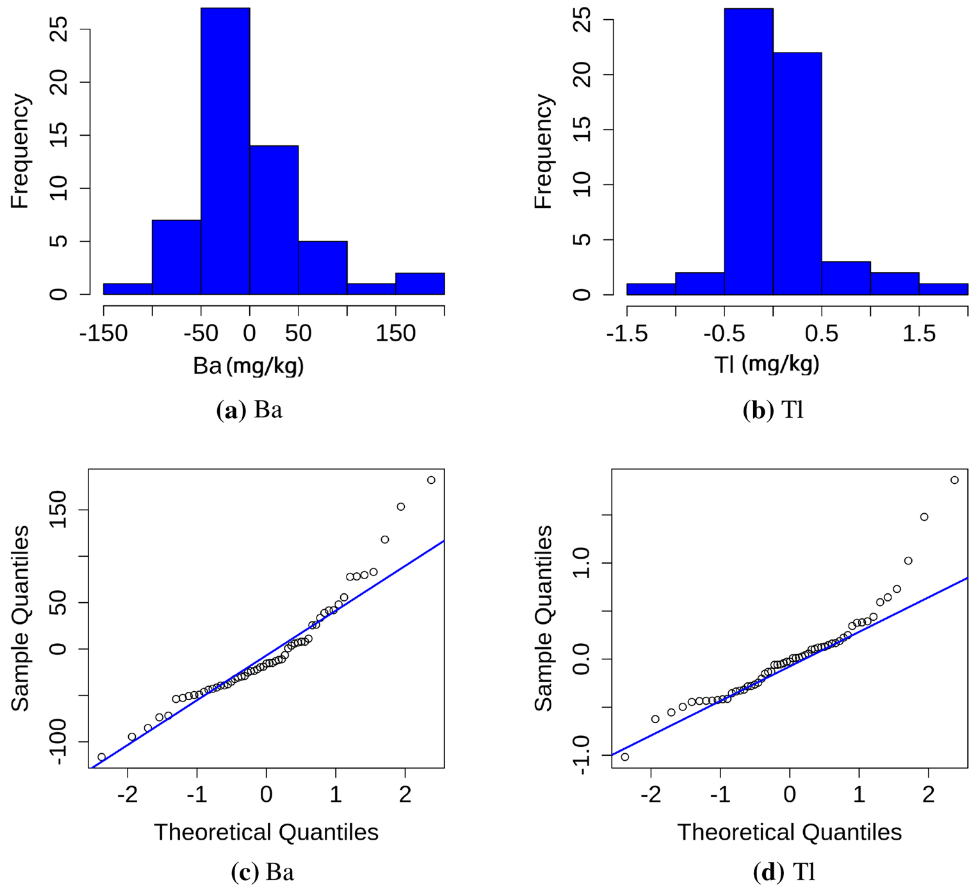


Fig. 17 Histogram and Normal QQ plot of KED pseudo-cross-validation errors for Ba and Tl concentrations



show clear sampling bias (the sampling density is very dense in some areas and very sparse in others). Under QRF, the continuous variation of the spatially distributed of interest is placed all into the mean function, i.e., assume that the observations equal a true but unknown continuous mean function plus independent and identically distributed errors. Whereas under KED, the continuous variation of the spatially distributed of interest is decomposed into the large-scale spatial variation (accounted through the mean function) and the small-scale spatial variation (accounted through the spatial-dependence structure). QRF assumes independent samples to compute classification rules. This assumption is very practical for estimating quantities involved in the algorithm and for assessing asymptotic properties of estimators. Unfortunately, in the realm of spatial data, data under study may present some amount of spatial correlation. When the sampling scheme is very irregular, a direct application of QRF could lead to biased discriminant rules due, for example, to the possible oversampling of some areas.

Results from the real case study demonstrated that QRF can be used to generate prediction maps comparable to those generated using KED (Figs. 18 and 19). This similarity could be explained by the fact that KED and QRF predictors are expressed as a linear combination of observations [Eqs. (3) and (9)]. The difference between KED and QRF predictors relies on the way that coefficients are estimated. Under KED, coefficients are estimated using a parametric approach, while under QRF, they are estimated via a non-parametric approach. Moreover, with QRF, there no need to consider any stationarity or normality conditions or any other transformation. QRF demands far less interventions and data preparation than using KED.

The KED approach of modeling the uncertainty consists of computing a kriging estimate and the associated error variance, which are then combined to derive a Gaussian-type confidence interval. However, the main limitation of kriging variance is that when it is used to calculate the confidence interval, it relies on the assumptions of normality of the distribution of prediction errors, and the variance of prediction errors is independent of the actual data values and depends only on the data configuration. Thus, the kriging variance tends to under-estimate the prediction uncertainty. Gaussian distributions of errors are more likely to arise when data used in estimations have a Gaussian distribution; as most distributions of variable of interests encountered in practice are skewed, it is unlikely that error distributions will be Gaussian or even symmetrical.

Fig. 18 Prediction and prediction uncertainty maps provided by **a**, **c** KED and **b**, **d** QRF for Ba concentration

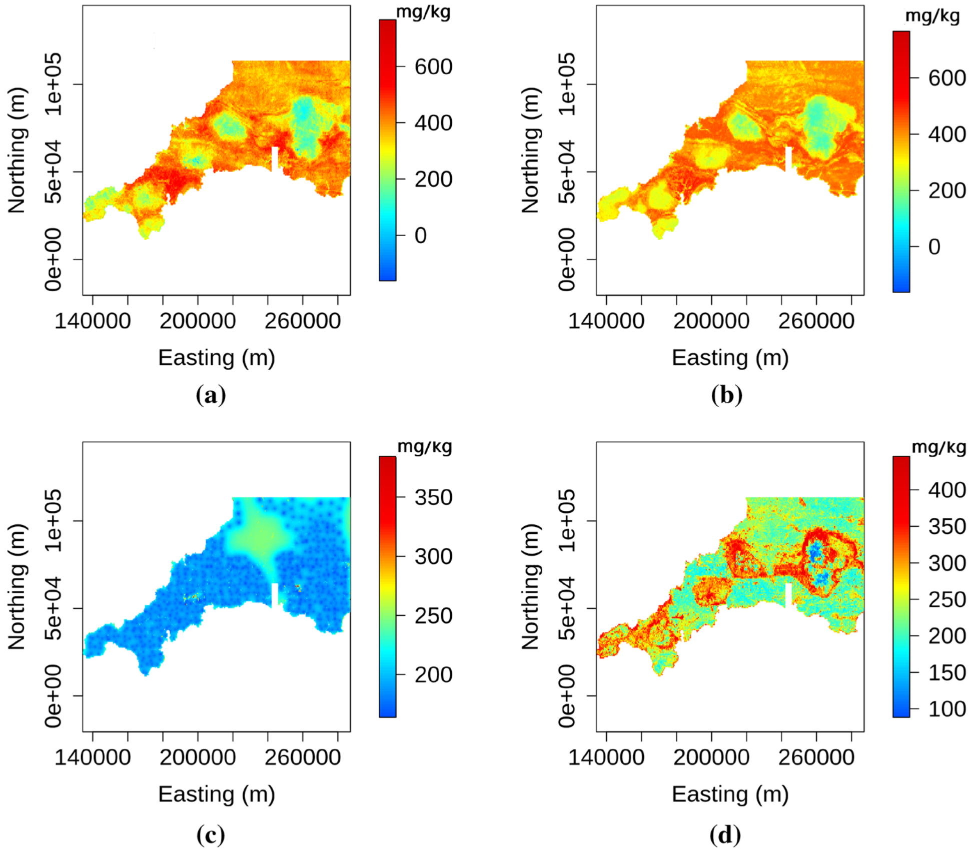
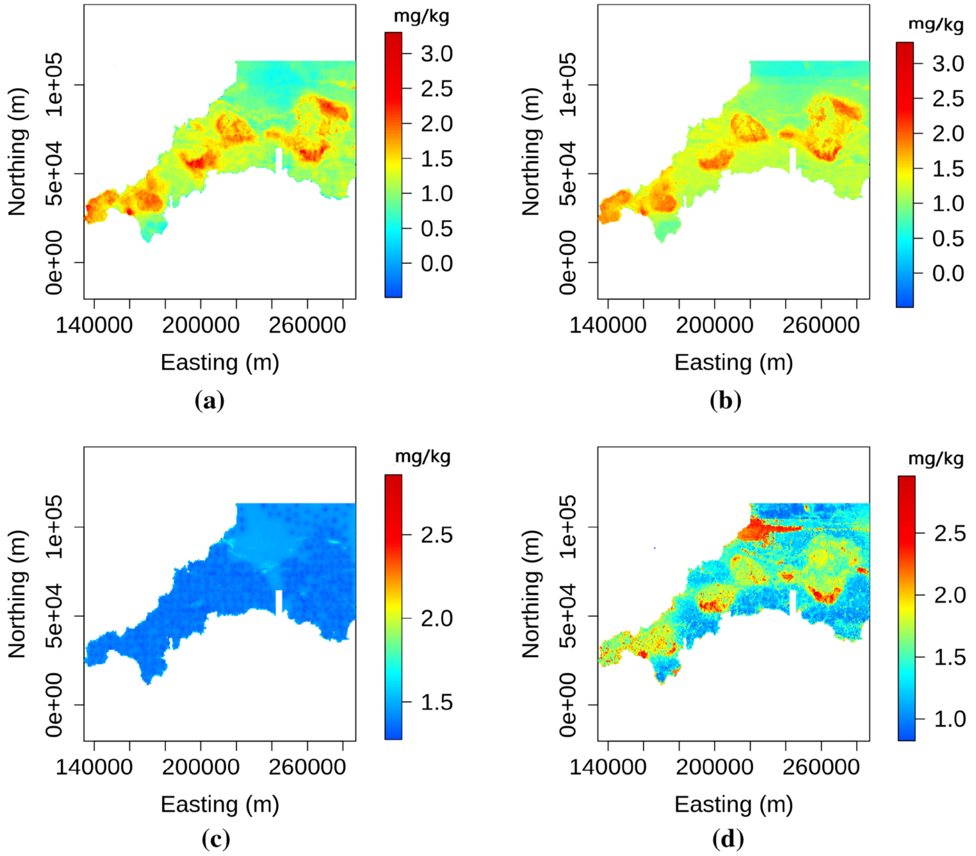


Fig. 19 Prediction and prediction uncertainty maps provided by **a**, **c** KED and **b**, **d** QRF for Tl concentration



It is argued that uncertainty about an unknown is intrinsic to the level of information available and to a prior model for the relations between that information and the unknown. Thus, the assessment of uncertainty should not be done around a particular estimate, because many optimality criteria can be defined resulting in different "optimal" estimates.

Contrarily to KED, under QRF, the conditional distribution function of the variable of interest is not related to any particular prior multivariate distribution model, such as Gaussian. Likewise, QRF is not able to provide a quantification of the predictor error in terms of the prediction error variance, which could also be used as a measure of uncertainty. QRF puts as priority not the derivation of a "optimal" estimator, but the modeling of the uncertainty. That uncertainty model takes the form of a probability distribution of the unknown rather than that of an estimation error.

Conclusions

This study compared kriging with external drift (KED) and quantile random forest (QRF) with respect to their performance in making point predictions and modeling prediction uncertainty of spatial data, where spatial dependence plays an important role. The study showed examples designed in such a way that one can expect them to show the superiority of one method over another in terms of prediction accuracy and prediction uncertainty accuracy. Given these distinct examples, it seems unlikely that there would be a best method for all applications. Nonetheless, QRF appears as a promising method for mapping spatially distributed continuous variables when auxiliary information are available everywhere within the region under study. However, there is a need to develop prediction uncertainty approaches for QRF in a context of relatively strong spatial correlation of the variable of interest. As suggested by Hengl et al. (2018), more specific geographical measures of proximity and connectivity between observations should be used during the training of QRF models (e.g., Euclidean distances to sampling locations, Euclidean distances to reference points in the study area). One approach could be to select random samples taking into account their clustering via spatial correlation during the training of QRF models. For example, the random samples could be picked with a probability inversely proportional to the clustering of samples. This strategy could resolve the oversampling in clustered areas.

Acknowledgements The authors are grateful to the anonymous reviewers for their helpful and constructive comments on earlier versions of the manuscript. The authors would like to thank Charlie Kirkwood at the British Geological Survey for providing the real-world spatial data set used in this paper.

References

Appelhans T, Mwangomo E, Hardy DR, Hemp A, Nauss T (2015) Evaluating machine learning approaches for the interpolation of monthly air temperature at mt. kilimanjaro, tanzania. Spat Stat 14(Part A):91–113

Ballabio C, Panagos P, Monatanarella L (2016) Mapping topsoil physical properties at european scale using the lucas database. Geoderma 261(Supplement C):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 31:1–14

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Breiman L, Friedman J, Stone C, Olshen R (1984) Classification and regression trees. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, Abingdon

Carranza EJM (2008) Geochemical anomaly and mineral prospectivity mapping in GIS. Handbook of Exploration and Environmental Geochemistry. Elsevier, Amsterdam

Chiles J-P, Delfiner P (2012) Geostatistics: modeling spatial uncertainty. Wiley, Hoboken

Coulston JW, Blinn CE, Thomas VA, Wynne RH (2016) Approximating prediction uncertainty for random forest regression models. Photogramm Eng Remote Sens 82(3):189–197

Deutsch C (1997) Direct assessment of local accuracy andprecision. In: Baafi, EY, Schofield NA (Eds), 5th International Geostatistics Congress, Wollongong '96. KluwerAcademic Publishers, London, pp 115–125

Foresti L, Pozdnoukhov A, Tuia D, Kanevski M (2010) Extreme precipitation modelling using geostatistics and machine learning algorithms. In: Atkinson PM, Lloyd CD (eds) geoENV VII—geostatistics for environmental applications. Springer, Dordrecht, pp 41–52

Goovaerts P (2001) Geostatistical modelling of uncertainty in soil science. Geoderma 103(1):3–26

Hengl T (2009) A practical guide to geostatistical mapping. University of Amsterdam, Amsterdam

Hengl T, Heuvelink GB, Stein A (2004) A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120(1):75–93

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

Kanevski M (2008) Advanced mapping of environmental data: geostatistics, machine learning and B ayesian maximum entropy. Wiley, Hoboken

Kanevski M, Pozdnoukhov A, Timonin V (2009) Machine learning for spatial environmental data: theory, applications, and software. EPFL press, Lausanne

Khan SZ, Suman S, Pavani M, Das SK (2016) Prediction of the residual strength of clay using functional networks. Geosci Front 7(1):67–74

Kirkwood C, Cave M, Beamish D, Grebby S, Ferreira A (2016a) A machine learning approach to geochemical mapping. J Geochem Explor 167(Supplement C):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

Lado LR, Hengl T, Reuter HI (2008) Heavy metals in european soils: a geostatistical analysis of the foregs geochemical database. Geoderma 148(2):189–199

Leuenberger M, Kanevski M (2015) Extreme learning machines for spatial environmental data. Comput Geosci 85(Part B):64–73

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). Canberra, Australia, pp 73–79

Li J, Heap AD (2008) A review of spatial interpolation methods for environmental scientists. Geoscience Australia, Canberra

Li J, Heap AD, Potter A, Daniell JJ (2011) Application of machine learning methods to spatial interpolation of environmental variables. Environmen Modell Softw 26(12):1647–1659

Meinshausen N (2006) Quantile regression forests. J Mach Learn Res 7(Jun):983–999

Meinshausen N (2017) *quantregForest: Quantile Regression Forests*. https://CRAN.R-project.org/package=quantregForest. R package version 1.3-7

Moyeed RA, Papritz A (2002) An empirical comparison of kriging methods for nonlinear spatial point prediction. Math Geol 34(4):365–386

Papritz A, Dubois JR (1999) Mapping heavy metals in soil by (non-)linear kriging an empirical validation. In: Gómez-Hernández J, Soares A, Froidevaux R (eds) geoENV II—geostatistics for environmental applications. Springer, Dordrecht, pp 429–440

Papritz A, Moyeed RA (2001) Parameter uncertainty in spatial prediction: checking its importance by cross-validating the wolfcamp and rongelap data sets. In: Monestiez P, Allard D, Froidevaux R (eds) geoENV III—geostatistics for environmental applications. Springer, Dordrecht, pp 369–380

R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Accessed 11 Nov 2018

Renard D, Bez N, Desassis N, Beucher H, Ors F, Freulon X (2018) RGeostats: geostatistical package. R package version 11.2.4. http://cg.ensmp.fr/rgeostats. Accessed 11 Nov 2018

Tadic JM, Ilic V, Biraud S (2015) Examination of geostatistical and machine-learning techniques as interpolators in anisotropic atmospheric environments. Atmos Environ 111:28–38

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(Supplement C):98–110

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

Vermeulen D, Niekerk AV (2017) Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates. Geoderma 299(Supplement C):1–12

Wackernagel H (2013) Multivariate geostatistics: an introduction with applications. Springer, Berlin

Wilford J, de Caritat P, Bui E (2016) Predictive geochemical mapping using environmental correlation. Appl Geochem 66(Supplement C):275–288