Geoderma xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms

Gábor Szatmári*, László Pásztor

Department of Soil Mapping and Environmental Informatics, Institute for Soil Sciences and Agricultural Chemistry, Centre for Agricultural Research, Hungarian Academy of Sciences, H-1022 Budapest, Herman Ottó út 15., Hungary

ARTICLE INFO

Keywords:
Digital soil mapping
Uncertainty
Kriging variance
Geostatistical simulation
Machine learning
Bootstrapping

ABSTRACT

We compared the suitability of several commonly applied digital soil mapping (DSM) techniques to quantify uncertainty with regards to a survey of soil organic carbon stock (SOCS) in Hungary. To represent the wide range of DSM techniques fairly, the followings were selected: universal kriging (UK), sequential Gaussian simulation (SGS), random forest combined with kriging (RFK) and quantile regression forest (QRF). For RFK two different uncertainty quantification approaches were adopted based on kriging variance (RFK-1) and bootstrapping (RFK-2). The selection of the potential environmental covariates was based on Jenny's factorial model of soil formation. The spatial predictions of SOCS and their uncertainty models were evaluated and compared using a control dataset. For this purpose, we applied the most common measures (i.e. mean error and root mean square error), furthermore, accuracy plot and G statistic. According to our results, QRF and SGS produced the best uncertainty models. UK and RFK-2 overestimated the uncertainty whereas RFK-1 produced the worst uncertainty quantification according to the accuracy plots and G statistics. We could draw the general conclusion that there is a need to validate the uncertainty models. Furthermore, great attention should be paid to the assumptions made in uncertainty modelling.

1. Introduction

Predictive soil maps suffer from different types of errors, where the most common error sources could be the measurements, digitization, typing, interpretation, classification, generalization and interpolation (Heuvelink, 2014). Therefore, the quantification, visualization and communication of the uncertainty of the digital soil mapping (DSM) products would be indispensable to stakeholders (e.g. policy makers, society etc.) as it has already been stressed by the GlobalSoilMap.net initiative (Arrouays et al., 2014).

Nowadays, various approaches (e.g. geostatistical and machine learning) are in use to model and quantify the uncertainty of DSM products. Most of these approaches apply a probabilistic framework within which the soil attribute of interest at a single location is regarded as a realization of a random variable. The most commonly applied geostatistical approach is the kriging variance that is jointly computed with the kriging prediction (Webster and Oliver, 2007). Vaysse and Lagacherie (2017) applied the kriging variance among others to construct uncertainty models to various DSM products in France. Kempen et al. (2014) applied the regression kriging variance to produce the 90% prediction interval for topsoil clay map in the Netherlands.

Another frequently applied approach is the family of geostatistical simulations that generates alternative and equally probable stochastic realizations from a random function model to assess uncertainty (Goovaerts, 1997). Via the generated stochastic realizations one is able to assess and quantify uncertainty. For example, Heuvelink et al. (2016) applied sequential Gaussian simulation to model the spatial variability of various soil properties over Europe. Szatmári et al. (2015) tested a sequential stochastic simulation approach based on regression kriging to model the spatial uncertainty of soil organic matter content in a small catchment area, Hungary. Poggio and Gimona (2014) used a 3D GAM + GS algorithm (i.e. generalized additive models with Gaussian simulation) to create a 3D soil organic carbon stock model for Scotland. We have to note if the same model is assumed, the predictions and the prediction intervals produced by a simulation approach should converge to those produced by kriging as the number of simulated realizations is increased.

Machine learning algorithms (MLA) are becoming more common in DSM because of the computational power availability (Rossiter, 2018). However, the quantification of uncertainty by MLA is quite novel. Vaysse and Lagacherie (2017) applied quantile regression forest to model the uncertainty of various DSM products in France. Furthermore,

* Corresponding author.
E-mail addresses: szatmari@rissac.hu (G. Szatmári), pasztor@rissac.hu (L. Pásztor).

https://doi.org/10.1016/j.geoderma.2018.09.008
Received 1 November 2017; Received in revised form 28 August 2018; Accepted 4 September 2018

Rudiyanto et al. (2016) applied the same technique to produce the 90% prediction interval to their peat thickness maps over some Indonesian peatlands. At the same time, a great effort was made to model and quantify the prediction uncertainty by alternative approaches. For example, Viscarra Rossel et al. (2015) elaborated an uncertainty quantification algorithm based on bootstrapping, where the resulting prediction realizations were used to predict and then quantify the uncertainty for the Australian 3D soil grid. Padarian et al. (2017) applied this approach for the Chilean soil grid. Malone et al. (2011) elaborated an empirical method whereby the prediction intervals are defined from the distribution of model errors. The feature space was partitioned into clusters (with a fuzzy k-means routine) which share similar error model.

When a geostatistical model is used to represent the residual variation, a parametric model must be assumed. The most common choice is a multivariate normal model but the normality assumption can be relaxed by applying a parametric (e.g. logarithmic) or non-parametric (e.g. normal scores) transformation to the data prior to model estimation.

The aim of our study was to evaluate and compare the uncertainty modelling capabilities of some frequently applied approaches in detail. For this purpose we applied an independent control dataset, which was not used in DSM. In this study we applied the following DSM techniques: (1) universal kriging, (2) sequential Gaussian simulation, (3) random forest combined with kriging and (4) quantile regression forest. All the selected techniques apply a probabilistic framework to model and quantify the uncertainty at a prediction location. However, the selected algorithms use different approaches to do that. For example, universal kriging uses its kriging variance to model the uncertainty, whereas sequential stochastic simulation applies simulated values to produce the model of uncertainty. Therefore, an additional objective of our paper was to discuss the pros and cons of the selected approaches in the light of our results.

We selected the soil organic carbon stock (SOCS) as the target variable. In this study, we applied the specifications of the Global Soil Organic Carbon (GSOC) mapping campaign (Yigini et al., 2018), which was launched by the Global Soil Partnership. In brief, the main goal of the campaign was to develop a global soil organic carbon stock map for the topsoil layer. The GSOC concept builds on official national datasets, therefore, a bottom-up (country-driven) approach is pursued. The area of Hungary (93,030 km^2) was the target domain of our study.

2. Theory

2.1. Spatial prediction and its uncertainty

The spatial variation of soil properties can be described and modelled in terms of a deterministic component and a stochastic component:

$$Z(\mathbf{u}) = m(\mathbf{u}) + \varepsilon(\mathbf{u}), \quad (1)$$

where Z is the soil property, m is the deterministic part describing structural variation, ε is the stochastic part consisting of random variation that could be spatially correlated and $\mathbf{u}$ is the vector of the geographical coordinates. DSM techniques are in use to predict the values of a given soil property in an area of interest. However, no map is error free (Heuvelink, 2014) (i.e. the predicted values could be slightly different from the true values), where the error is defined as the difference between the true and predicted value of a soil property. In fact, the error is not known spatially exhaustively (Heuvelink, 2014). Actually, we are uncertain about the error (and the true value). It is not the given soil property that is uncertain, it is our knowledge that is uncertain about the given soil property. Hence, uncertainty is a term expressing our imperfect knowledge in describing an environmental object, property or process and we aware of that (Bárdossy and Fodor, 2004). In this study, we adopt a probabilistic way to model and quantify the uncertainty at a prediction location, where we will consider the unknown value $z(\mathbf{u})$ as a realization of a random variable $Z(\mathbf{u})$. The (cumulative) distribution function of the random variable $Z(\mathbf{u})$ fully models the uncertainty because it gives the probability that the unknown is no greater than any given threshold z, that is

$$F(\mathbf{u}; z) = \text{Prob}\{Z(\mathbf{u}) \leq z\}, \quad (2)$$

We will consider Eq. (2) as the model of uncertainty at the prediction location $\mathbf{u}$. The aim is to produce such a model for each prediction location. For this purpose, either parametric or non-parametric approaches can be applied. In the case of parametric approaches, an analytical model defined by a few parameters is commonly adopted, whereas in the cases of non-parametric ones, the model of uncertainty is described by an empirical distribution function.

2.2. DSM algorithms for prediction and uncertainty quantification

In this study, we applied the following DSM techniques: (1) universal kriging, (2) sequential Gaussian simulation, (3) random forest combined with kriging and (4) quantile regression forest. The listed techniques are well-known and frequently applied algorithms in DSM. Therefore, just a brief introduction of the listed algorithms will be provided here. More details on them are given in the cited papers and textbooks.

Universal kriging (UK), also termed regression kriging or kriging with external drift (Hengl et al., 2004), combines regression of the target soil variable on environmental covariates with kriging of the regression residuals (Hengl et al., 2007). In terms of Eq. (1) the deterministic component is modelled by a multiple linear regression whereas the stochastic part of variation is modelled by kriging using the regression residuals. For the residuals we assume that they are multivariate normal. The parameters of the UK estimator and the variogram of the stochastic component are estimated by REML (residual maximum likelihood) (Lark, 2012). The prediction variance of UK is the sum of the estimation variance of the deterministic component and the prediction variance of the kriged residuals (i.e. kriging variance) (Hengl et al., 2007, Eq. (6)). Therefore, it reflects the position of unsampled locations in both geographic and feature space. By means of UK prediction and its variance a parametric model of uncertainty can be produced with the assumption of normality. It is a full and mathematically concise model since a normally distributed random variable is fully determined by its mean and variance.

As opposed to any kriging techniques, the main aim of *sequential Gaussian simulation (SGS)* is to generate alternative and equally probable stochastic realizations, which reproduce the model statistics (e.g. histogram and variogram) rather than to minimize the local prediction variance (Goovaerts, 1997). The SGS algorithm involves sequential sampling of the N-point conditional cumulative distribution function of the random function model that models the joint uncertainty at N locations (Goovaerts, 1997):

$$F(\mathbf{u}_1, \ldots, \mathbf{u}_N; z_1, \ldots, z_N \mid (n)) = \text{Prob}\{Z(\mathbf{u}_1) \leq z_1, \ldots, Z(\mathbf{u}_N) \leq z_N\}, \quad (3)$$

where N is the number of the prediction locations and n is the number of the observations. In practice, a one-point conditional cumulative distribution function is modelled and sampled at each of the prediction locations visited along a random path (Deutsch and Journel, 1998; Goovaerts, 1997). The kriging prediction and its variance are in use to construct the one-point distribution function at each prediction location. To ensure the reproduction of model statistics each one-point distribution function is made conditional not only to the observations but all previously simulated values visited along a random path. If each prediction location is visited and each has been given a simulated value, then the resulting set of simulated values represents one stochastic realization. Other realizations can be obtained by repeating the entire sequential sampling process with possibly different random paths (Deutsch and Journel, 1998; Goovaerts, 1997). A common approach in

DSM is to use SGS for the stochastic component (i.e. second term on the right-hand side of Eq. (1)) and add the generated realizations back to the deterministic component (i.e. first term on the right-hand side of Eq. (1)) (Poggio and Gimona, 2014). In that case, we assume that the residuals are multivariate normal but we do not have any assumption about the distribution of the target variable. However, Goovaerts (1997, 388 p.) suggests to use SGS for the whole spatial modelling. We applied the later approach in this study. In that case, the adopted random function model to $Z(\mathbf{u})$ is multivariate normal. This calls for a priori transformation of the original z-data into y-data with a standard normal cumulative distribution function (Deutsch and Journel, 1998; Goovaerts, 1997). For this purpose normal score transform is commonly applied, which is a type of quantile transformation based on Gaussian anamorphosis. UK was connected to SGS, i.e. the UK algorithm was used for characterizing the conditional cumulative distribution function at each prediction location (Szatmári et al., 2015). The model of uncertainty for a prediction location is given by the empirical distribution function of the back-transformed simulated values at that location. The spatial prediction for a prediction location is commonly identified by the mean of the simulated values.

Random forest combined with kriging (RFK) can be considered as a new "workhorse" in DSM (Keskin and Grunwald, 2018). In terms of Eq. (1) the deterministic component is modelled by random forest (RF) whereas the stochastic part of variation is modelled by kriging using the computed residuals. The variogram of the stochastic component is estimated by Matheron's (1963) method-of-moments estimator. RF provides a prediction for the target soil variables via an ensemble of classification or regression trees. The RF prediction is the conditional mean that is approximated by the averaged prediction of the generated trees. According to Hengl et al. (2015), some of the advantages of RF over linear regression are as follows: it can fit complex non-linear relationships and the correlation between the environmental covariates is not a limiting factor. However, we must assume that the observations are independent and the residuals are multivariate normal. In this study, we applied two approaches to model and quantify the uncertainty of the RFK prediction. These are based on (1) kriging variance (Vaysse and Lagacherie, 2017) and (2) bootstrapping (Malone et al., 2017; Viscarra Rossel et al., 2015). In the first approach, the kriging variance of the stochastic part of variation and the RFK prediction (RFK-1) described above are in use to model the uncertainty as in the case of UK. However, this uncertainty model does not account for the uncertainty in estimating the deterministic component. Therefore, there is a need to assume that the varying local mean is exactly equal to the RF prediction. In addition, we have to assume normality to be able to construct a parametric model of uncertainty. For bootstrapping, the approach involves repeated random sampling with replacement of the observations. Using the bootstrap sample a RFK model is fitted and a digital soil map is generated. By repeating the process of bootstrap sampling and applying the RFK model, we are able to generate probability distribution of the prediction realizations from each model at each prediction location (Malone et al., 2017). To produce a robust prediction for the target soil variable the mean of the prediction realizations (RFK-2) is commonly computed and mapped. According to Viscarra Rossel et al. (2015, Eq. (6)), the overall variance of the RFK prediction is approximated by the sum of the mean squared error of the spatial model, the average kriging variance of the residuals and the variance of the generated prediction realizations. By means of the average prediction realizations and the variance of the RFK prediction a parametric model of uncertainty can be produced with the assumption of normality.

Quantile regression forest (QRF) is a quite novel approach in DSM. According to Meinshausen (2006), RF can give valuable information not only about the conditional mean but also about the conditional distribution of the target variable. The key difference between RF and QRF can be summarized as follows (Meinshausen, 2006): for each node in each tree, RF keeps only the mean of the observations that fall into this node and neglects all other information, whereas QRF keeps not just their mean but the value of all observations in this node. Based on this information QRF can give the empirical distribution function, which will be the model of uncertainty. QRF keeps the advantages of RF. On the other hand, we have to assume that the observations are independent as in the case of RF. In this study, we did not model the stochastic part of variation (i.e. second term on the right-hand side of Eq. (1)) by residual kriging.

2.3. Derivation of the 90% prediction interval

In DSM a common way to spatially explicitly visualize the uncertainty of a spatial prediction is to map the upper and lower limit of the 90% prediction interval (PI) (Arrouays et al., 2014; Heuvelink, 2014). This PI reports the range of values within which the true value is expected to occur 9 times out of 10. If the uncertainty model is parametric and normally distributed, the lower and upper limit of the 90% PI can be readily computed by subtracting and adding 1.64 times the prediction standard deviation to the prediction. This can be used for UK and both types of RFK. If the uncertainty model is non-parametric (i.e. an empirical distribution function is given), the lower and upper limit of the 90% PI can be identified by the 5th and 95th percentiles of the empirical distribution function. This can be used for SGS and QRF.

2.4. Validation of uncertainty models

The uncertainty models can be validated by computing the actual fraction of true values falling within symmetric PIs of varying width p. A series of PIs can be readily derived by the $\frac{(1-p)}{2}$ and $\frac{(1+p)}{2}$ quantiles of the distribution function (Goovaerts, 2005). If a set of control data and independently derived distribution functions are available at some control locations, the fraction is computed by

$$\bar{\xi}(p) = \frac{1}{m}\sum_{i=1}^{m}\xi(\mathbf{u}_i;p) \quad \forall\, p \in [0,1] \tag{4}$$

with

$$\xi(\mathbf{u}_i;p) = \begin{cases} 1, & \text{if } z(\mathbf{u}_i) \in (p_{\text{lower}}, p_{\text{upper}}] \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where is the fraction for the PI of width p, $\xi(\mathbf{u}_i;p)$ is the indicator function, $z(\mathbf{u}_i)$ is the true value, p_{upper} is the upper limit of PI, p_{lower} is the lower limit of PI and m is the number of control points. A graphical way to check the performance of the uncertainty models is to plot against p that is frequently referred to as accuracy plot (Deutsch, 1997; Goovaerts, 2001), but also known as prediction interval coverage probability plot (Malone et al., 2011; Shrestha and Solomatine, 2006). Ideally, the observed fractions are equal to the expected fractions. If they are lower than the expectations, then the uncertainty has been underestimated. If they are higher, the uncertainty has been too liberally estimated (i.e. overestimated). The closeness of the observed and expected fractions can be assessed by the G statistic (Deutsch, 1997) defined as

$$G = 1 - \int_0^1 [3a(p) - 2][\bar{\xi}(p) - p]dp \tag{6}$$

with

$$a(p) = \begin{cases} 1, & \text{if } \bar{\xi}(p) \geq p \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $a(p)$ is the indicator function. The G value can be interpreted as the higher the value the closer the observed and expected fractions. Ideally, it is equal to 1.

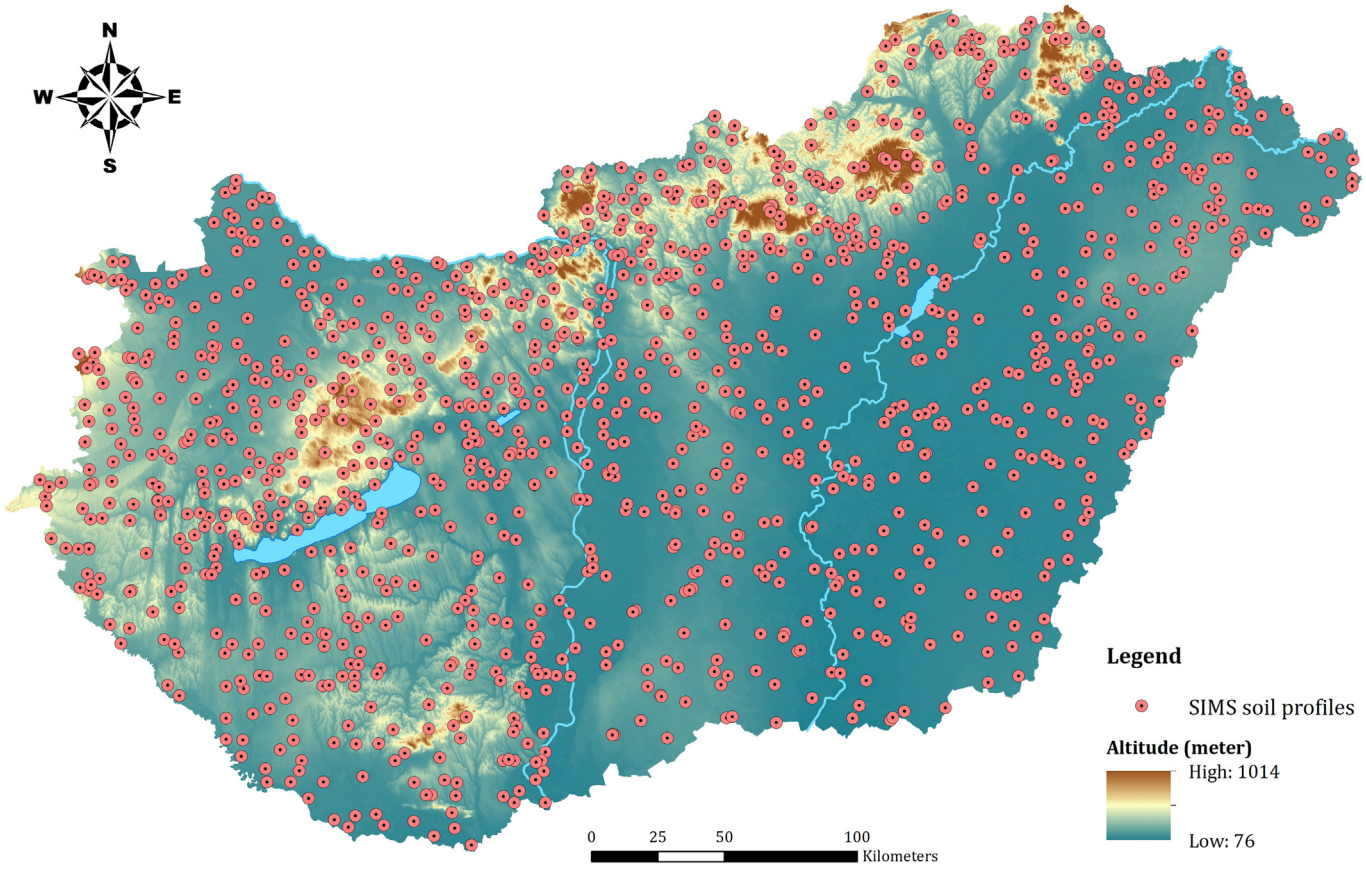

Fig. 1. Spatial position of the monitoring sites (n = 1236) of the Hungarian Soil Information and Monitoring System (SIMS).

3. Materials and methods

3.1. Soil data

In this study, we derived the reference soil data from the Hungarian Soil Information and Monitoring System (SIMS). SIMS contains 1236 monitoring sites (i.e. soil profiles) (Fig. 1). These soil profiles have 4859 genetic soil horizons altogether that have been defined according to the Hungarian genetic soil classification system. We applied the measured soil organic carbon content [%], bulk density [g·cm^{-3}] and the "on-the-field" estimated volume of stones [%] of each soil genetic horizon determined in the starting year (i.e. 1992). These primary soil properties were used for computing SOCS at the level of soil profile. The summary statistics of the applied soil properties are presented in Table 1 regardless of their vertical origins from the soil profile.

3.2. SOCS computation at the level of soil profile

We computed SOCS according to the GSOC specifications (Yigini et al., 2018). The mandatory mapping depth was 0–30 cm, however, there was an optional extension to organic soils (including peats), where the recommended mapping depth was 0–100 cm. SOCS was computed for each SIMS soil profile with the following equation:

$$SOCS_d = SOC_d \cdot BD_d \cdot (1 - VST) \cdot TL, \quad (8)$$

where $SOCS_d$ [g·cm^{-2}] is the soil organic carbon stock to given depth, SOC_d [%] is the soil organic carbon content for given depth, BD_d [g·cm^{-3}] is the bulk density for given depth, VST [%] is the volume of stones and TL [cm] is the thickness of soil layer. We transformed the [g·cm^{-2}] unit to [tons·ha^{-1}] because later one is a more common and convenient unit to express and interpret SOCS. Henceforth, the [tons·ha^{-1}] unit will be applied. Due to the different mapping depth specification of GSOC, we carried out the computation of SOCS for mineral and organic soils separately. In the case of mineral soils SOCS was computed for the mandatory depth (i.e. TL was 30 cm), whereas for organic soils SOCS was calculated for the recommended depth (i.e. TL was 100 cm). We selected randomly 200 SIMS soil profiles as control dataset that was not used in DSM. The aim of the control dataset was to validate and compare the resulting uncertainty models.

Table 1
Summary statistics of soil organic carbon (SOC), bulk density (BD), volume of stones (VST) and the computed soil organic carbon stock (SOCS).

Soil properties	Unit	Min	Max	Mean	Median	Std. dev.
SOC	[%]	0.005	12.993	0.692	0.440	0.695
BD	[g·cm^{-3}]	0.700	1.910	1.430	1.440	0.148
VST	[%]	0	99	0.025	0.000	0.118
SOCS	[tons·ha^{-1}]	0.000	367.800	50.260	46.100	34.100

3.3. Environmental covariates

The selection of the potential environmental covariates was based on Jenny's (1941) factorial model of soil formation, which has been formulated by McBratney et al. (2003). The applied environmental covariates are summarized in Table 2 according to the *scorpan*'s factors (i.e. *s*: other soil properties, *c*: climate, *o*: organisms, *r*: topography, *p*: parent material, *a*: age and *n*: geographical position). We applied the genetic soil type map of Hungary as environmental covariate that includes 9 (higher order) soil types according to the Hungarian soil classification system. We also used the available climatic data layers, such as the long-term mean annual precipitation and temperature (Table 2). Organisms were represented by satellite images that were acquired by moderate-resolution imaging spectroradiometer (MODIS)

Table 2
Summary of the applied environmental covariates.

Scorpan's factors	Name	Resolution	Type
Soil	Soil type map of Hungary	100 m	Categorical
Climate	Long-term mean annual evapotranspiration	100 m	Continuous
	Long-term mean annual evaporation	100 m	Continuous
	Long-term mean annual precipitation	100 m	Continuous
	Long-term mean annual temperature	100 m	Continuous
Organism	Normalized difference vegetation index (MODIS)	250 m	Continuous
	Near infrared (MODIS)	250 m	Continuous
	Red (MODIS)	250 m	Continuous
Relief	Altitude	100 m	Continuous
	Cross-sectional curvature	100 m	Continuous
	Diffuse insolation	100 m	Continuous
	Direct insolation	100 m	Continuous
	Diurnal anisotropic heating	100 m	Continuous
	Downslope curvature	100 m	Continuous
	Local curvature	100 m	Continuous
	Local downslope curvature	100 m	Continuous
	Local upslope curvature	100 m	Continuous
	Longitudinal curvature	100 m	Continuous
	LS factor	100 m	Continuous
	Relative slope position	100 m	Continuous
	Slope	100 m	Continuous
	Surface area	100 m	Continuous
	Topographic position index	100 m	Continuous
	Topographic wetness index	100 m	Continuous
	Upslope curvature	100 m	Continuous
	Vertical distance to channel network	100 m	Continuous
Parent material	Geological map of Hungary	1:100,000	Categorical

in 2012 and 2013. We applied the normalized difference vegetation index, as well as the near infrared and red bands. The environmental covariates related to the relief were derived from the digital elevation model (DEM) of Hungary, such as the slope, topographic wetness index, vertical distance to channel network and diurnal anisotropic heating (Table 2). The parent material was represented by the geological map of Hungary that includes 13 classes according to Bakacsi et al. (2014). The geological map was converted to raster layer.

Due to the various data sources, the selected environmental covariates had different spatial resolutions (Table 2). Therefore, we resampled them into a common reference system with 500 m resolution.

3.4. Implementation of DSM algorithms

We applied each of the DSM techniques listed in Section 2.2. In Fig. 2 we summarized our DSM activity. The computed SOCS has a positively skewed distribution. The summary statistics of SOCS are presented in Table 1.

In the case of SGS we used the normal scores for spatial modelling. For UK, RFK and QRF the SOCS data was directly applied. We carried out a principal component analysis on the continuous environmental covariates (Table 2) because we could suspect there is a correlation between the covariates. The resulting principal components were used for UK and SGS to avoid multicollinearity in regression analysis. We applied indicator transform to the categorical covariates (Table 2) to be able to use them in regression analysis (Goovaerts, 1997; Hengl et al., 2004). For RFK and QRF we applied the original environmental covariates listed in Table 2 and we generated 500 regression trees, respectively. For each variogram we fitted a nested variogram model. The first structure was a nugget model to describe discontinuity at the origin (i.e. lag zero). The second structure was an isotropic spherical model, which model type is frequently applied in soil science (Webster and Oliver, 2007). By each DSM technique we performed spatial prediction. By the SGS algorithm we generated 1000 alternative and equally probable stochastic realizations that were back-transformed to the original scale. For bootstrapping we generated 1000 bootstrap samples and using these samples we derived 1000 prediction realizations. For each spatial prediction we produced the model of uncertainty at each prediction location then we derived and mapped the lower and upper limit of the 90% PI.

3.5. Comparison of spatial predictions and uncertainty models

We divided the evaluation and comparison procedure into two parts. In the first part we evaluated and compared the errors of the spatial predictions using the most common measures, i.e. mean error (ME) and root mean square error (RMSE). The first one is commonly referred to as bias whereas the second one is frequently referred to as the spread of the error distribution. A reasonable goal for any DSM work is to produce map with ME close to zero and RMSE as low as possible. We applied the control dataset to compute these measures. In the second part we validated the uncertainty models. First of all, we compared the uncertainty models at a randomly selected control point, where the true SOCS value was known. In the next step, we compiled the accuracy plots for each DSM technique using the control dataset and we computed the G statistics.

4. Results

4.1. DSM and uncertainty quantification

The R-squared values of the fitted multiple linear regression models for UK and SGS are 0.24 and 0.31, respectively. In the case of RFK, the R-squared value of the fitted RF model is 0.71, which is higher than the previous ones. This can be explained by the suitability of RF for modelling complex non-linear relationships.

In Fig. 3 we present the histograms of the computed residuals (i.e. SOCS minus the deterministic model predictions). The computed residuals for SGS are in the transformed units. In the case of SGS the residual histogram shows a normal distribution that is appropriate for the required normality assumption. For UK and RFK this assumption does not appears to be appropriate because of the outliers.

We plot the omnidirectional variograms of the residuals, as well as the fitted models in Fig. 4. The range values of the fitted models for UK, SGS and RFK are 18.6 km, 38.1 km and 9.6 km, respectively. The nugget to sill ratios are quite high (0.76, 0.83 and 0.78 for UK, SGS and RFK, respectively). However, such high values are not rare in DSM (e.g. Hengl et al., 2015; Vaysse and Lagacherie, 2017). We have to note that the fitted models for UK and SGS differ significantly from each other, which can be attributed to the normal score transform (Deutsch, 2002).

For each DSM technique we present spatial prediction, as well as the upper and lower limit of the 90% PI in Fig. 5. We plot the width of the 90% PI for each DSM technique to inspect the different patterns of uncertainty (Fig. 6). It is apparent that for RFK-1 the uncertainty is only related to the configuration of SIMS points whereas for SGS and QRF the uncertainty is related to size of the predictions and the covariates, respectively. In the case of UK and RFK-2 the uncertainty accounts for both the unexplained stochastic variation and the uncertainty in estimating the deterministic model.

4.2. Performance of spatial predictions

In general, the computed biases are lower than zero (Table 3), i.e. the applied DSM techniques a little bit underestimate SOCS. For RFKs (i.e. RFK-1 and RFK-2) the computed biases are the closest to zero, in addition the RMSE values are almost the lowest. This can be attributed to the fact that RF often outperforms the regression techniques (Hengl et al., 2015). Furthermore, the residuals were modelled by kriging that minimizes the local error variance (Webster and Oliver, 2007). According to the error measures, RFKs outperformed UK, SGS and QRF.

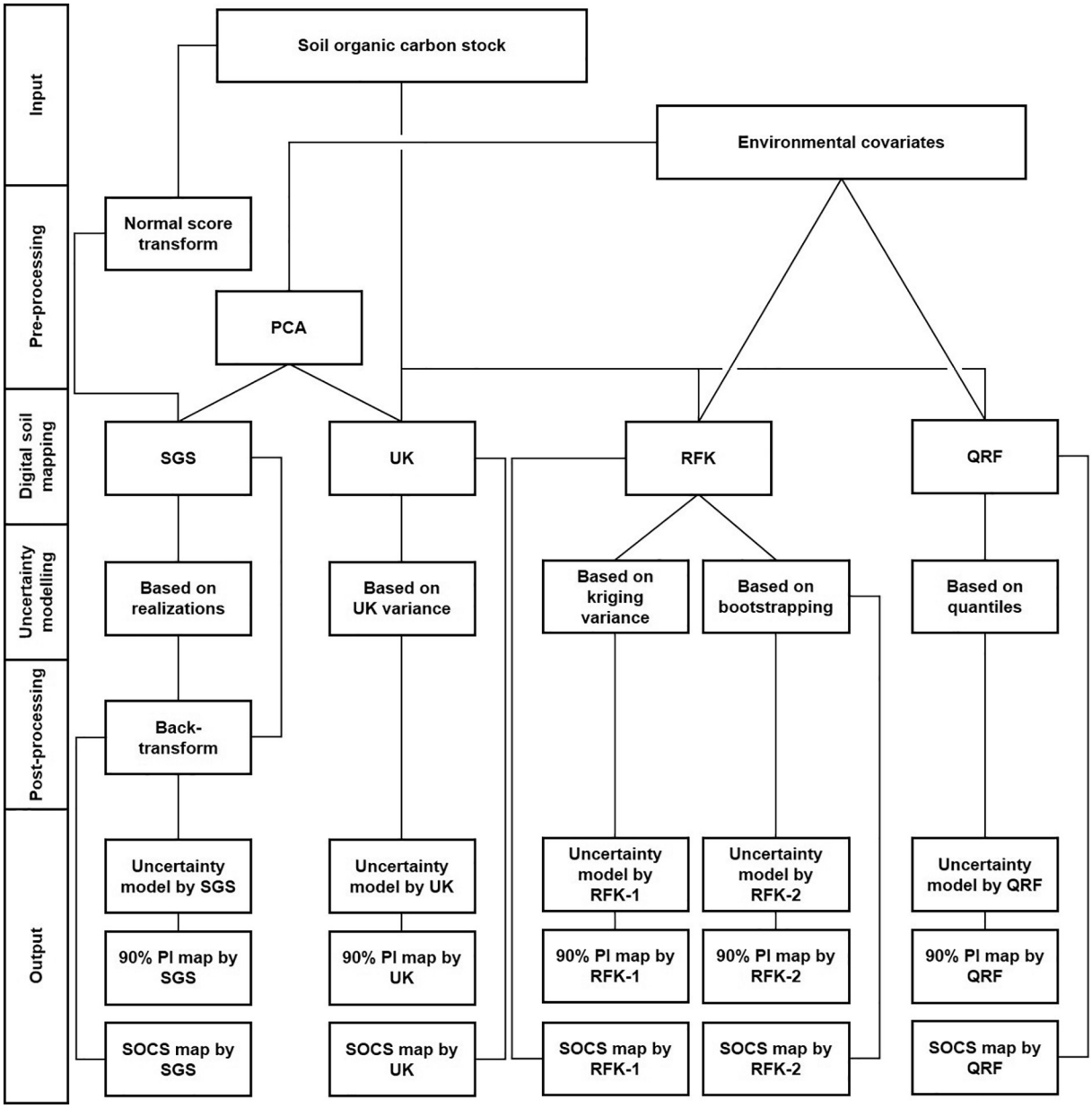

Fig. 2. Workflow. Abbreviations: PCA: principal component analysis, SGS: sequential Gaussian simulation, UK: universal kriging, RFK: random forest combined with kriging, QRF: quantile regression forest and PI: prediction interval.

4.3. Comparison and validation of uncertainty models

At the randomly selected control point the uncertainty models for UK and RFKs show a normal distribution (Fig. 7) that comes from our assumption about the error distribution. For SGS and QRF the models do not follow a normal distribution. In the case of QRF the distribution is slightly negatively skewed whereas for SGS the distribution has a longer tail for the high SOCS values (Fig. 7), which is an evidence of positive skewness. Each of the computed 90% PIs encapsulates the true SOCS value (Fig. 7). The width of the 90% PIs increases in the order RFK-1 (42.4) < RFK-2 (84.9) < QRF (88.8) < UK (98.0) < SGS (110.7).

In the next step, we validated the compiled 90% PI map (Fig. 5) for each DSM technique. We examined at the control points that how many times SOCS falls within the 90% PI. In Table 3 we summarizes the observed fraction for each DSM technique. SGS properly estimates the uncertainty because it yields the expectation. UK, QRF and RFK-2 too liberally estimate the uncertainty and therefore the observed fractions are higher than the expectation whereas RFK-1 underestimates the uncertainty.

At the majority of the control points RFK-1 underestimates the uncertainty and therefore the accuracy plot is below the $x = y$ line (Fig. 8). This is because only the kriging variance prevails the uncertainty since we assumed that the trend prediction is certain. UK and RFK-2 overestimate the uncertainty and therefore the accuracy plots are above the $x = y$ line (Fig. 8). In addition, we can state that they are too far from the $x = y$ line. SGS and QRF properly estimate the uncertainty and therefore the accuracy plots are quite close to the $x = y$ line (Fig. 8). For QRF the computed fractions in $p \in [0.25, 0.35]$ are somewhat lower than the expectations but the differences are almost negligible.

The G statistics increase in the order of RFK-1 < UK < RFK-2 < SGS < QRF (Table 3). For SGS and QRF the G statistics are close to the expectation and therefore they yield the most reliable uncertainty models. According to Goovaerts (2005), between two uncertainty models with similar G statistics, the one with the smallest spread would

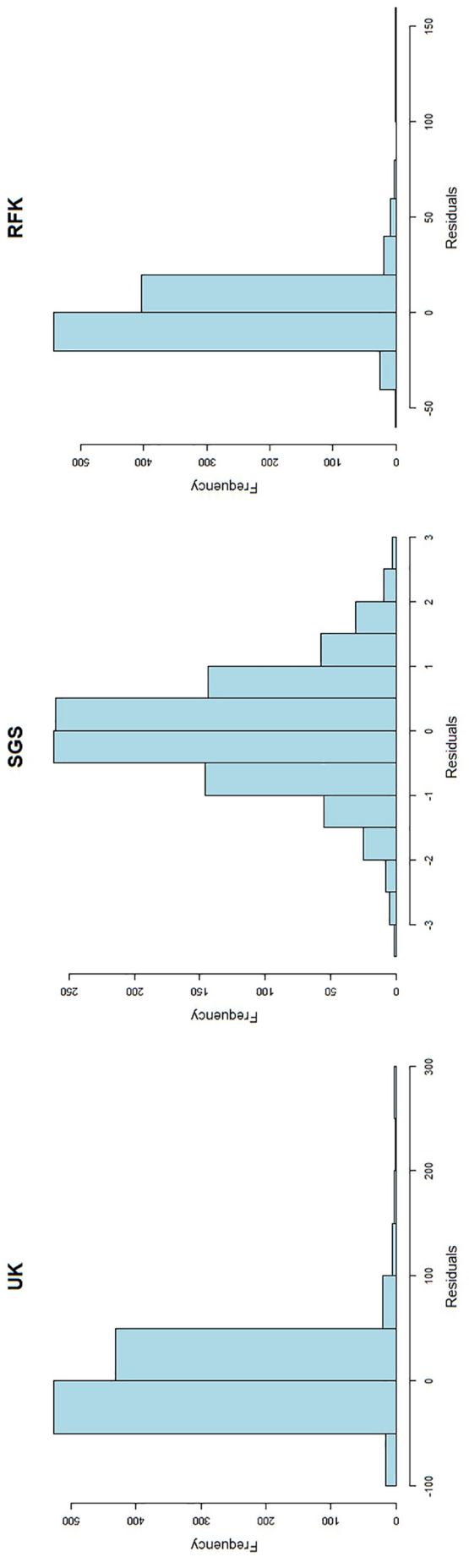

Fig. 3. Histograms of the residuals. Abbreviations: UK: universal kriging, SGS: sequential Gaussian simulation and RFK: random forest combined with kriging.

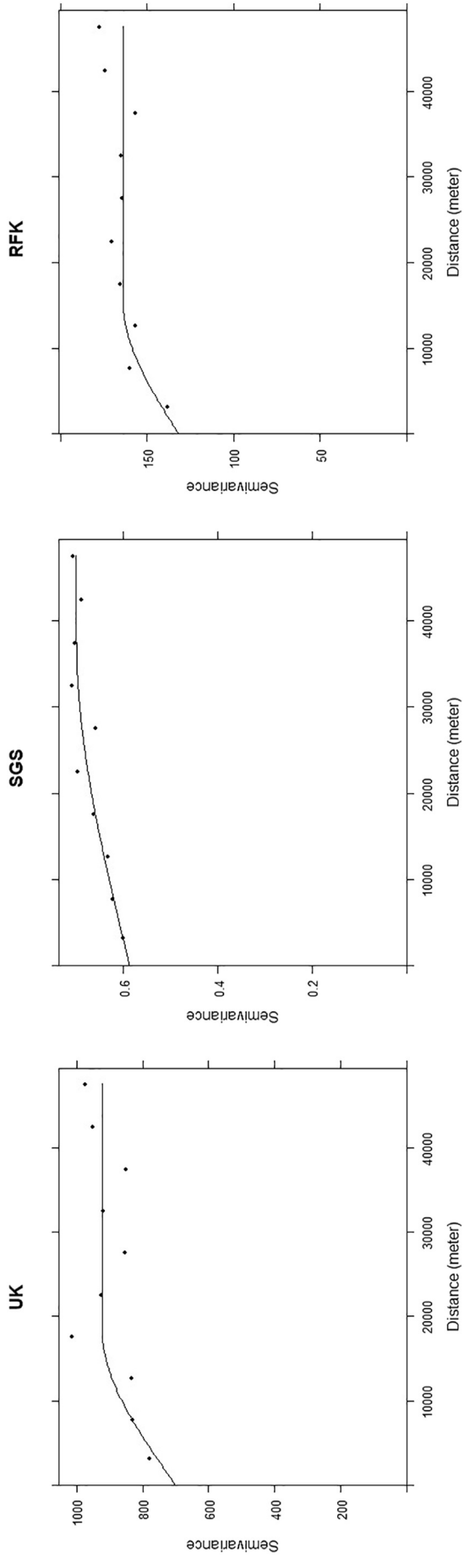

Fig. 4. Variograms and fitted models. Abbreviations: UK: universal kriging, SGS: sequential Gaussian simulation and RFK: random forest combined with kriging.

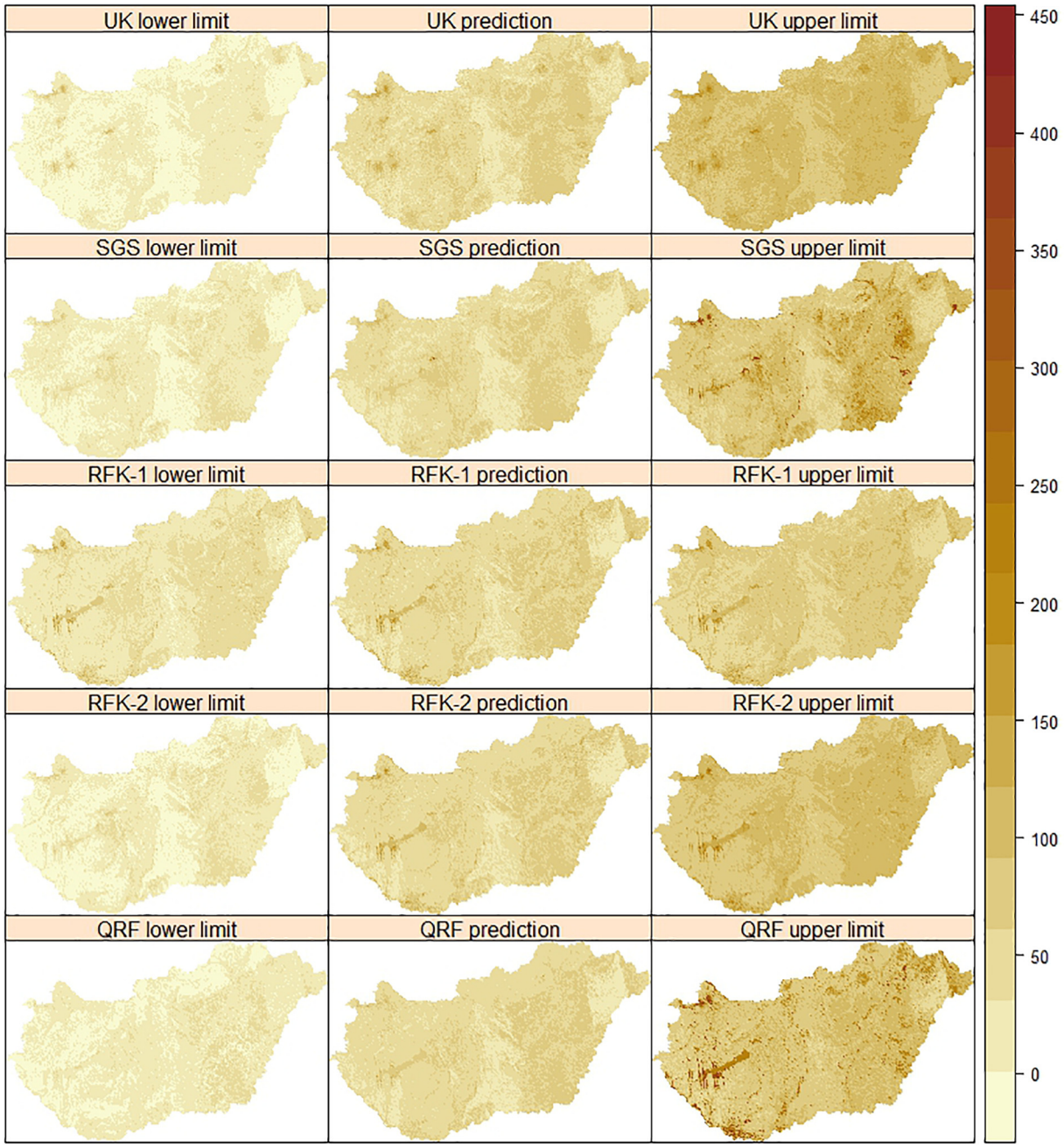

Fig. 5. Spatial predictions of soil organic carbon stock, as well as the upper and lower limit of the 90% prediction intervals. The unit of the maps is [tons·ha^{-1}]. Abbreviations: UK: universal kriging, SGS: sequential Gaussian simulation, RFK-1 random forest combined with kriging (using kriging variance), RFK-2: random forest combined with kriging (using bootstrapping) and QRF: quantile regression forest.

be preferred. Hence, we plot the average range of the PIs that include the true values for a series of probability values (Fig. 9). SGS provides smaller spread for the lowest (i.e. p ∈ [0.01, 0.1]) and highest (i.e. p ∈ [0.8, 0.99]) probability values, whereas QRF gives smaller spread for the middle ones. The highest difference between them occurs at $p = 0.9$, where SGS yields a shorter average range with 4.79 [tons·ha^{-1}].

5. Discussion

Many papers and textbooks on geostatistics (e.g. Goovaerts, 1997, 1999; Journel and Rossi, 1989) do not recommend to apply the kriging variance as a general measure of local accuracy because it is data-value independent. Indeed, “we have seen that the kriging variance does not directly depend on the data values used for the estimation: It is an unconditional variance” (Chilès and Delfiner, 2012, 176 p.). This independence calls for the stringent assumption of homoscedasticity, i.e. the error variance has to be independent from the actual data values and it depends only on the data configuration (Goovaerts, 1997). This could be unrealistic for some variables where the variance increases according to the measured value (Lark and Lapworth, 2012; Manchuk et al., 2009; Marchant et al., 2011), a situation referred to as proportional effect.

In this study, RFK-1 produced contradictory results. On the one hand, RFK-1 gave one of the best spatial predictions according to the error measures (Table 3), on the other hand, RFK-1 underestimated the uncertainty according to the accuracy plot and G statistic (Table 3 and Fig. 8). RF often outperforms the most commonly applied trend

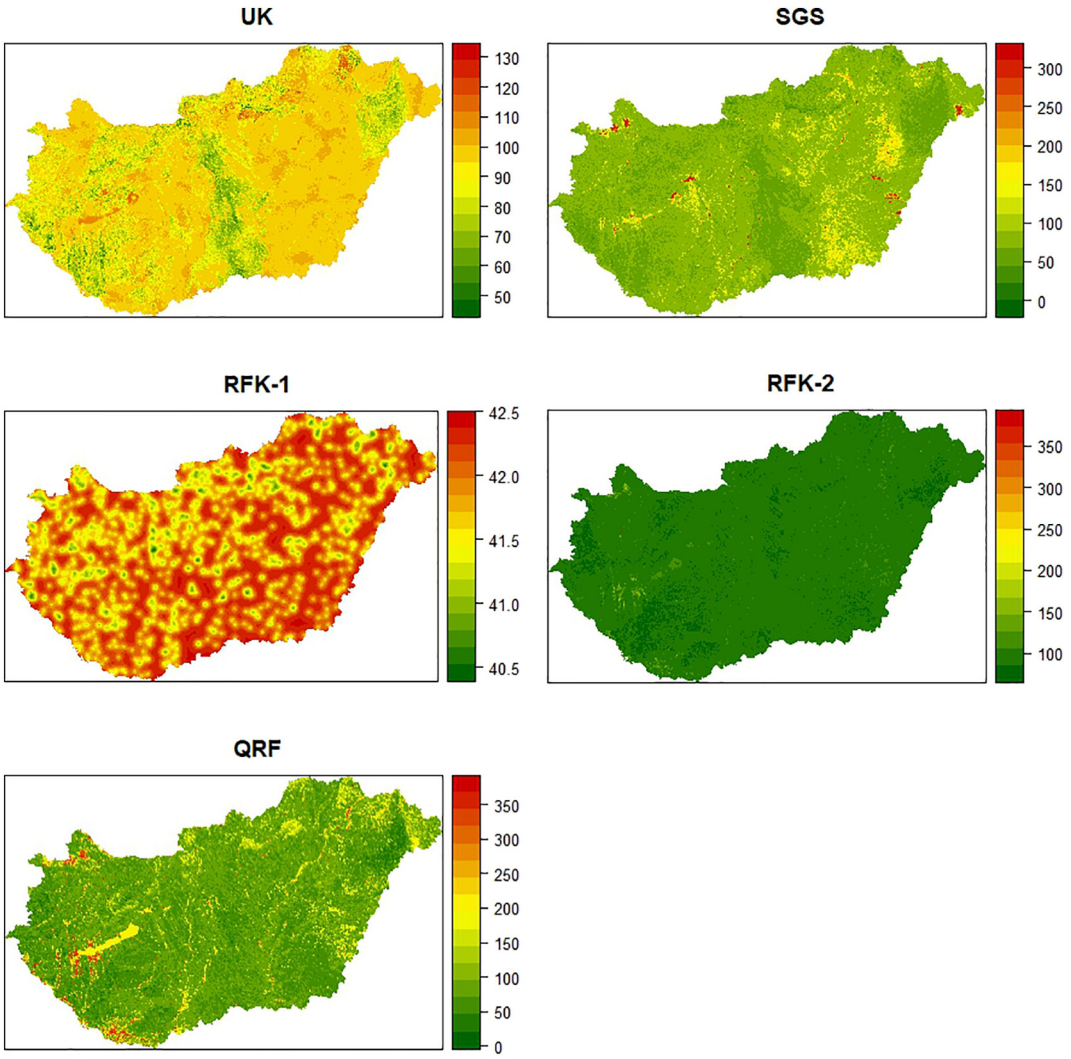

Fig. 6. Width of the 90% prediction intervals. The unit of the maps is [tons·ha^{-1}]. Abbreviations: UK: universal kriging, SGS: sequential Gaussian simulation, RFK-1: random forest combined with kriging (using kriging variance), RFK-2: random forest combined with kriging (using bootstrapping) and QRF: quantile regression forest.

Table 3
Performance of the applied digital soil mapping techniques for spatial prediction and uncertainty quantification. Abbreviations: ME: mean error, RMSE: root mean square error, PI: prediction interval, UK: universal kriging, SGS: sequential Gaussian simulation, RFK-1: random forest combined with kriging (using kriging variance), RFK-2: random forest combined with kriging (using bootstrapping) and QRF: quantile regression forest.

	UK	SGS	RFK-1	RFK-2	QRF
Error					
ME	−0.62	−0.62	−0.28	−0.19	−0.41
RMSE	25.63	25.53	25.05	24.89	24.86
Uncertainty					
Observed fraction in the 90% PI	0.96	0.90	0.76	0.94	0.93
G statistics	0.87	0.95	0.80	0.89	0.97

estimation techniques (Hengl et al., 2015). Furthermore, the residuals are modelled by kriging that minimizes the local error variance (Webster and Oliver, 2007). As a consequence, RFK-1 is designed to provide the most accurate prediction at an unsampled location. However, the uncertainty was prevailed by the kriging variance because we assumed that the trend prediction is certain. The kriging variance reflects only the position of unsampled locations in geographical space without any reflection about their position in feature space. Therefore, the assumption about the certainty of the RF prediction is too optimistic, which is a serious shortcoming of RFK-1 in point of uncertainty quantification. In addition, the average width of the 90% PI for RFK-1 was the lowest (Fig. 6). This could be misleading if one is looking for that algorithm, which provides the lowest uncertainty around a prediction.

SGS modelled properly the uncertainty (Table 3 and Fig. 8). However, all geostatistical simulations are computationally intensive and the generated stochastic realizations require massive storage capacity. Furthermore, using a simulation algorithm one will face a lot of pre- and post-processing steps (Geiger, 2012; Goovaerts, 2005) that make them not so attractive. In this study the pre- and post-processing of SGS was the most extensive (Fig. 2). In theory, SGS should provide equivalent result with UK if we apply SGS (1) directly or (2) to the stochastic part of variation (i.e. second term on the right-hand side of Eq. (1)). In both cases, the original (i.e. untransformed) data can be used to generate stochastic realizations because there is no assumption about the distribution of $Z(\mathbf{u})$. However, we applied Goovaerts' (1997, 388 p.) approach in this study, i.e. we relaxed the normality assumption by applying a normal scores transform prior to estimating our model. This is the reason why SGS and UK did not produce equivalent results.

QRF estimated properly the uncertainty (Table 3 and Fig. 8). One of the main drawbacks of QRF is that it is computationally intensive and

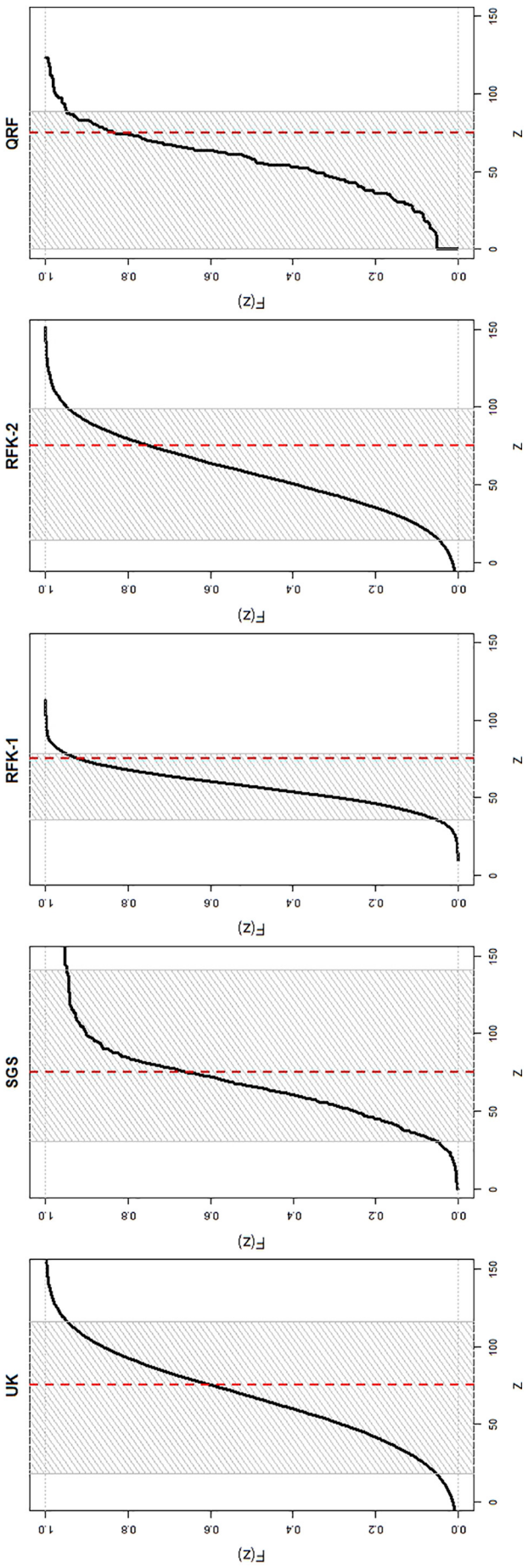

Fig. 7. Uncertainty models at a randomly selected control point. Abbreviations: UK: universal kriging, SGS: sequential Gaussian simulation, RFK-1: random forest combined with kriging (using kriging variance), RFK-2: random forest combined with kriging (using bootstrapping) and QRF: quantile regression forest. Legend: red dashed line: observed soil organic carbon stock value in [tons·ha^{-1}], grey shaded area: 90% prediction interval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

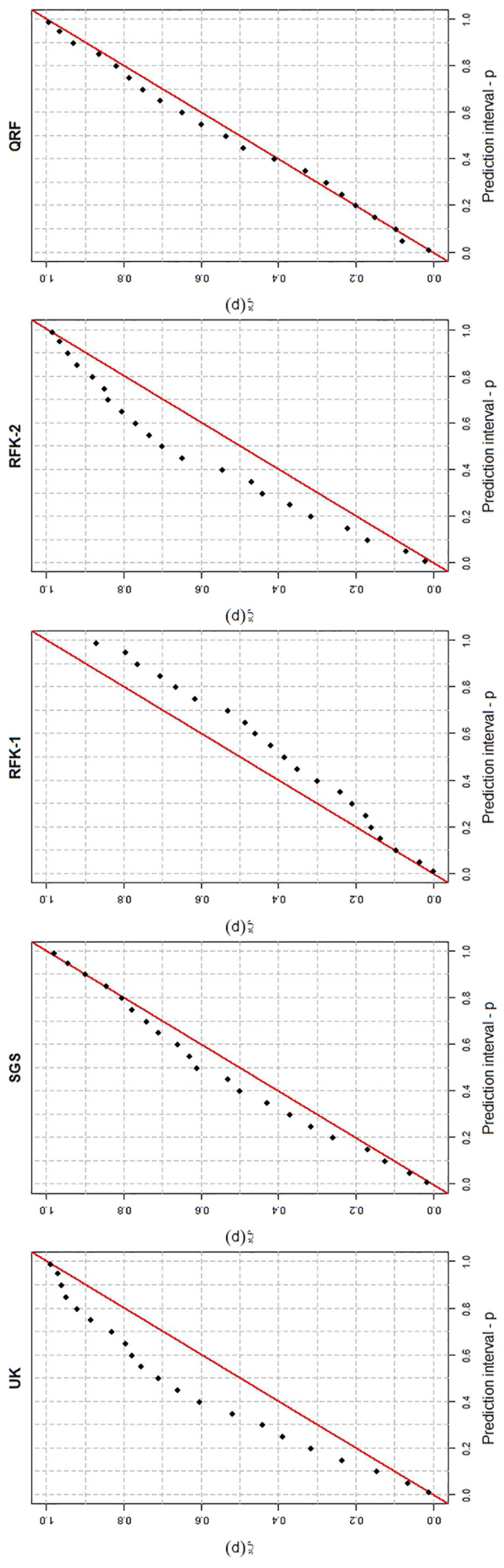

Fig. 8. Accuracy plots. Abbreviations: UK: universal kriging, SGS: sequential Gaussian simulation, RFK-1: random forest combined with kriging (using kriging variance), RFK-2: random forest combined with kriging (using bootstrapping) and QRF: quantile regression forest.

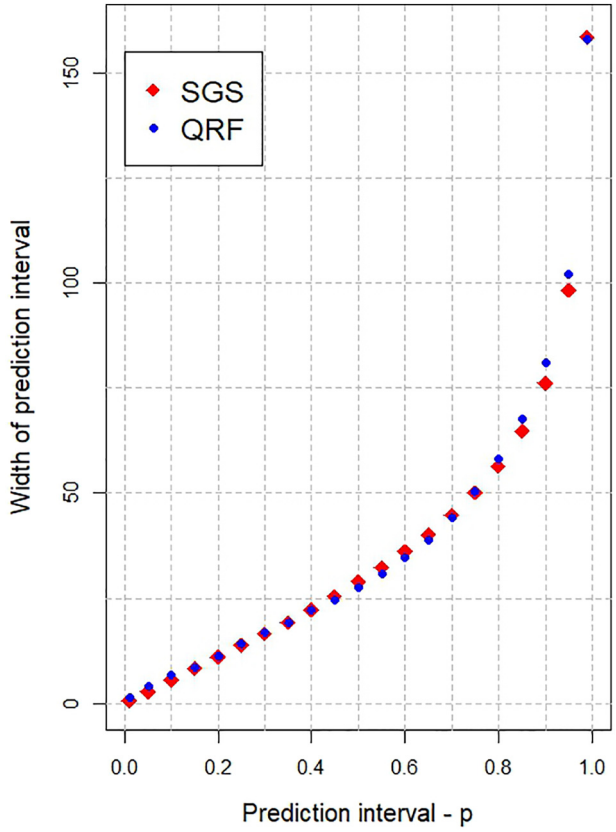

Fig. 9. Prediction interval-width plot. Abbreviations: SGS: sequential Gaussian simulation and QRF: quantile regression forest.

therefore, a great attention must be paid to the optimization of the application of QRF (Vaysse and Lagacherie, 2017). As opposed to the geostatistical algorithms, the modelling effort (e.g. pre-processing of environmental covariates to decrease multicollinearity, normal score transform, variogram modelling etc.) is reduced (Fig. 2) that makes QRF attractive. For QRF we did not model the stochastic part of variation (i.e. second term on the right-hand side of Eq. (1)) by residual kriging because this was beyond the scope of this study. Therefore, the applied QRF algorithm does not capture the auto-correlated error.

6. Conclusions

Our case study illustrated the importance of confirming that the assumptions made in uncertainty modelling and quantification are appropriate. Furthermore, there is a need to validate the resulting uncertainty models. For this purpose, accuracy plot and G statistic can be applied. In addition, we pointed out that the methods (i.e. UK and RFKs) which based upon a multivariate normal model were not appropriate to model and quantify the uncertainty of SOCS spatial prediction in Hungary. For this purpose, the two methods (i.e. SGS and QRF) which supported non-normal variation were more appropriate.

Acknowledgement

Our work was supported by the National Research, Development and Innovation Office (NKFIH; Grant No. KH-126725) and by the Széchenyi 2020 programme, the European Regional Development Fund - “Investing in your future” and the Hungarian Government (GINOP-2.3.2-15-2016-00028). The authors thank the two anonymous reviewers and B.P. Marchant for their helpful and constructive comments that helped to improve the manuscript.

References

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.d.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., Zhang, G.L., 2014. GlobalSoilMap. Toward a fine-resolution global grid of soil properties. Adv. Agron. 125, 93–134. https://doi.org/10.1016/B978-0-12-800137-0.00003-0.

Bakacsi, Z., Laborczi, A., Szabó, J., Takács, K., Pásztor, L., 2014. Az 1:100 000-es földtani térkép jelkulcsának és a FAO rendszer talajképző kőzet kódrendszerének javasolt megfeleltetése. Agrokém. Talajt. 63, 189–202.

Bárdossy, G., Fodor, J., 2004. Evaluation of Uncertainties and Risks in Geology. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-07138-0.

Chilès, J.-P., Delfiner, P., 2012. Geostatistics: Modeling Spatial Uncertainty, second ed. Wiley Blackwellhttps://doi.org/10.1002/9781118136188.

Deutsch, C.V., 1997. Direct assessment of local accuracy and precision. In: Baafi, E.Y., Schofield, N.A. (Eds.), Geostatstics Wollongong '96. Kluwer Academic Publishers, pp. 115–125.

Deutsch, C.V., 2002. Geostatistical Reservoir Modeling. Oxford University Press.

Deutsch, C.V., Journel, A.G., 1998. GSLIB: Geostatistical Software Library and user's Guide. Oxford University Press.

Geiger, J., 2012. Some thoughts on the pre- and post-processing in sequential gaussian simulation and their effects on reservoir characterization. In: Geiger, J., Pál-Molnár, E., Malvic, T. (Eds.), New Horizons in Central European Geomathematics, Geostatistics and Geoinformatics. GeoLitera, Szeged, pp. 17–34.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press.

Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89, 1–45. https://doi.org/10.1016/S0016-7061(98)00078-0.

Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma 103, 3–26. https://doi.org/10.1016/S0016-7061(01)00067-2.

Goovaerts, P., 2005. Geostatistical modeling of the spaces of local, spatial, and response uncertainty for continuous petrophysical properties. In: Coburn, T.C., Yarus, J.M., Chambers, R.L. (Eds.), Stochastic Moceling and Geostatistics: Principles, Methods, and Case Studies. Volume II. pp. 1–21. https://doi.org/10.1306/1063807CA53229.

Hengl, T., Heuvelink, G.B.M., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120, 75–93. https://doi.org/10.1016/j.geoderma.2003.08.018.

Hengl, T., Heuvelink, G.B.M., Rossiter, D.G., 2007. About regression-kriging: from equations to case studies. Comput. Geosci. 33, 1301–1315. https://doi.org/10.1016/j.cageo.2007.05.001.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., De Jesus, J.M., 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. https://doi.org/10.1371/journal.pone.0125814.

Heuvelink, G., 2014. Uncertainty quantification of GlobalSoilMap products. In: GlobalSoilMap. CRC Press, pp. 335–340. https://doi.org/10.1201/b16500-62.

Heuvelink, G.B.M., Kros, J., Reinds, G.J., De Vries, W., 2016. Geostatistical prediction and simulation of European soil property maps. Geoderma Reg. 7, 201–215. https://doi.org/10.1016/j.geodrs.2016.04.002.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York.

Journel, A.G., Rossi, M.E., 1989. When do we need a trend model in kriging? Math. Geol. 21, 715–739. https://doi.org/10.1007/BF00893318.

Kempen, B., Heuvelink, G., Brus, D., Walvoort, D., 2014. Towards GlobalSoilMap.net products for The Netherlands. In: Arrouays, D., McKenzie, N., Hempel, J., Richer de Forges, A.C., McBratney, A. (Eds.), GlobalSoilMap. CRC Press, pp. 85–90. https://doi.org/10.1201/b16500-19.

Keskin, H., Grunwald, S., 2018. Regression kriging as a workhorse in the digital soil mapper's toolbox. Geoderma 326, 22–41. https://doi.org/10.1016/j.geoderma.2018.04.004.

Lark, R.M., 2012. Towards soil geostatistics. Spat. Stat. 1, 92–99. https://doi.org/10.1016/j.spasta.2012.02.001.

Lark, R.M., Lapworth, D.J., 2012. Quality measures for soil surveys by lognormal kriging. Geoderma 173–174, 231–240. https://doi.org/10.1016/j.geoderma.2011.12.008.

Malone, B.P., McBratney, A.B., Minasny, B., 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160, 614–626. https://doi.org/10.1016/j.geoderma.2010.11.013.

Malone, B.P., Minasny, B., McBratney, A.B., 2017. Using R for Digital Soil Mapping, Progress in Soil Science. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-44327-0.

Manchuk, J.G., Leuangthong, O., Deutsch, C.V., 2009. The proportional effect. Math. Geosci. 41, 799–816. https://doi.org/10.1007/s11004-008-9195-z.

Marchant, B.P., Saby, N.P.A., Jolivet, C.C., Arrouays, D., Lark, R.M., 2011. Spatial prediction of soil properties with copulas. Geoderma 162, 327–334. https://doi.org/10.1016/J.GEODERMA.2011.03.005.

Matheron, G., 1963. Principles of geostatistics. Econ. Geol. 58. https://doi.org/10.2113/gsecongeo.58.8.1246.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma. https://doi.org/10.1016/S0016-7061(03)00223-4.

Meinshausen, N., 2006. Quantile Regression Forests. J. Mach. Learn. Res. 7, 983–999. https://doi.org/10.1111/j.1541-0420.2010.01521.x.

Padarian, J., Minasny, B., McBratney, A.B., 2017. Chile and the Chilean soil grid: a contribution to GlobalSoilMap. Geoderma Reg. 9, 17–28. https://doi.org/10.1016/j.geodrs.2016.12.001.

Poggio, L., Gimona, A., 2014. National scale 3D modelling of soil organic carbon stocks with uncertainty propagation - an example from Scotland. Geoderma 232–234, 284–299. https://doi.org/10.1016/j.geoderma.2014.05.004.

Rossiter, D.G., 2018. Past, present & future of information technology in pedometrics. Geoderma 324, 131–137. https://doi.org/10.1016/j.geoderma.2018.03.009.

Rudiyanto, Minasny, B., Setiawan, B.I., Arif, C., Saptomo, S.K., Chadirin, Y., 2016. Digital mapping for cost-effective and accurate prediction of the depth and carbon stocks in Indonesian peatlands. Geoderma 272, 20–31. https://doi.org/10.1016/j.geoderma.2016.02.026.

Shrestha, D.L., Solomatine, D.P., 2006. Machine learning approaches for estimation of prediction interval for the model output. Neural Netw. 19, 225–235. https://doi.org/10.1016/J.NEUNET.2006.01.012.

Szatmári, G., Barta, K., Farsang, A., Pásztor, L., 2015. Testing a sequential stochastic simulation method based on regression kriging in a catchment area in southern Hungary. Geol. Croat. 68. https://doi.org/10.4154/gc.2015.21.

Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64. https://doi.org/10.1016/j.geoderma.2016.12.017.

Viscarra Rossel, R.A., Chen, C., Grundy, M.J., Searle, R., Clifford, D., Campbell, P.H., 2015. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. Soil Res. 53, 845–864. https://doi.org/10.1071/SR14366.

Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists, second ed. Wiley.

Yigini, Y., Olmedo, G.F., Reiter, S., Baritz, R., Viatkin, K., Vargas, R., 2018. Soil Organic Carbon Mapping Cookbook, second ed. FAO, Rome.