Artificial Intelligence in Geosciences 3 (2022) 162–178

Contents lists available at ScienceDirect

Artificial Intelligence in Geosciences

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

Original research articles

Geostatistical semi-supervised learning for spatial prediction

Francky Fouedjio [a,*], Hassan Talebi [b]

[a] Rio Tinto, Data & Analytics, 152-158 St Georges Terrace, Perth, WA 6000, Australia
[b] Rio Tinto, Development & Technology, 152-158 St Georges Terrace, Perth, WA 6000, Australia

ARTICLE INFO

Keywords:
Labeled spatial data
Unlabeled spatial data
Spatial autocorrelation
Pseudo labeling
Spatial prediction

ABSTRACT

Geoscientists are increasingly tasked with spatially predicting a target variable in the presence of auxiliary information using supervised machine learning algorithms. Typically, the target variable is observed at a few sampling locations due to the relatively time-consuming and costly process of obtaining measurements. In contrast, auxiliary variables are often exhaustively observed within the region under study through the increasing development of remote sensing platforms and sensor networks. Supervised machine learning methods do not fully leverage this large amount of auxiliary spatial data. Indeed, in these methods, the training dataset includes only labeled data locations (where both target and auxiliary variables were measured). At the same time, unlabeled data locations (where auxiliary variables were measured but not the target variable) are not considered during the model training phase. Consequently, only a limited amount of auxiliary spatial data is utilized during the model training stage. As an alternative to supervised learning, semi-supervised learning, which learns from labeled as well as unlabeled data, can be used to address this problem. However, conventional semi-supervised learning techniques do not account for the specificities of spatial data. This paper introduces a spatial semi-supervised learning framework where geostatistics and machine learning are combined to harness a large amount of unlabeled spatial data in combination with typically a smaller set of labeled spatial data. The main idea consists of leveraging the target variable's spatial autocorrelation to generate pseudo labels at unlabeled data points that are geographically close to labeled data points. This is achieved through geostatistical conditional simulation, where an ensemble of pseudo labels is generated to account for the uncertainty in the pseudo labeling process. The observed labels are augmented by this ensemble of pseudo labels to create an ensemble of pseudo training datasets. A supervised machine learning model is then trained on each pseudo training dataset, followed by an aggregation of trained models. The proposed geostatistical semi-supervised learning method is applied to synthetic and real-world spatial datasets. Its predictive performance is compared with some classical supervised and semi-supervised machine learning methods. It appears that it can effectively leverage a large amount of unlabeled spatial data to improve the target variable's spatial prediction.

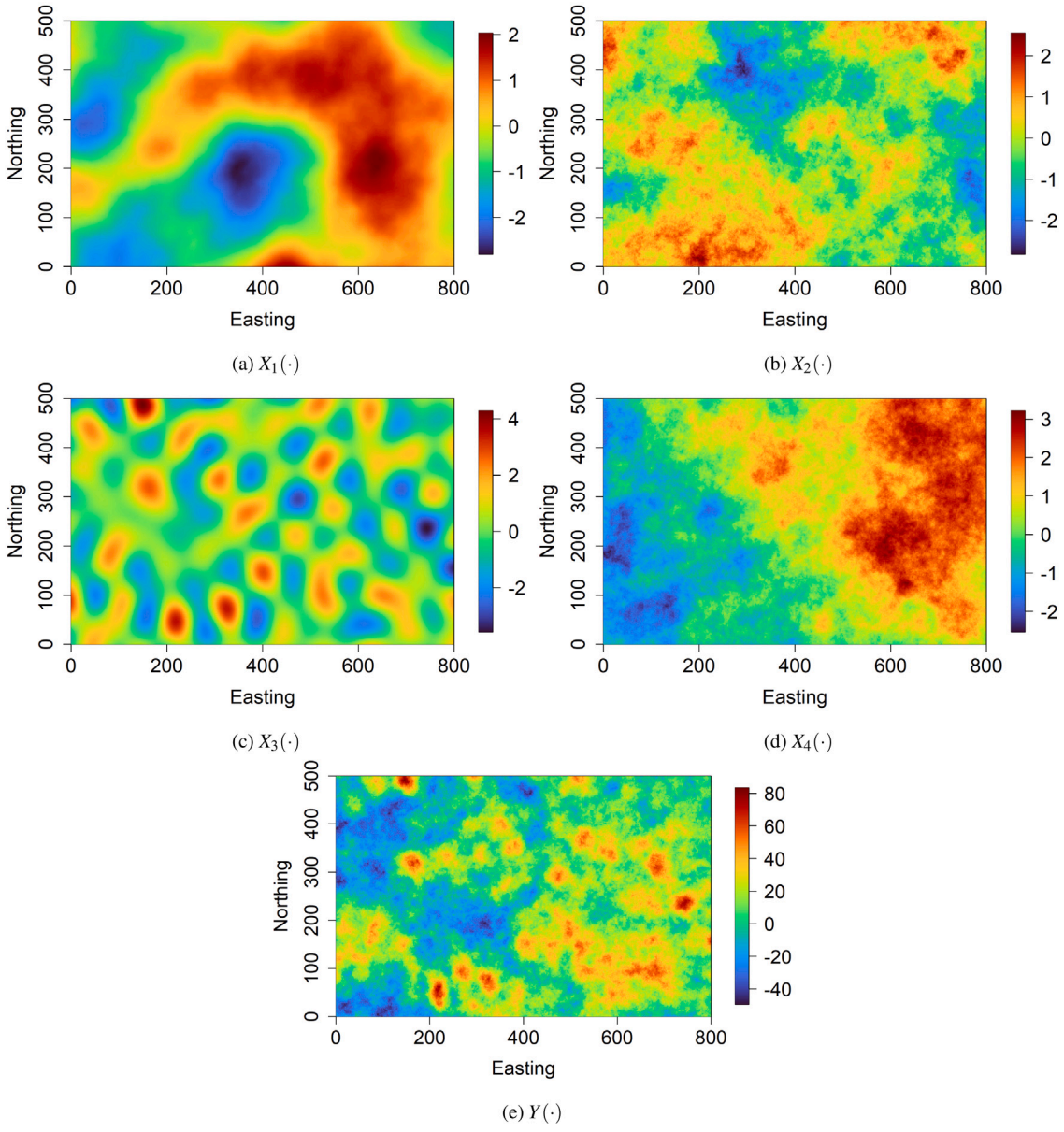
1. Introduction

The prediction of a variable of interest over the whole study region, in the presence of auxiliary variables observed everywhere within the area of interest, has become a ubiquitous task in various geoscience fields (Giaccone et al., 2021; Du et al., 2020; Maxwell et al., 2018; Kanevski, 2008; Kanevski et al., 2009). Supervised machine learning methods (e.g., random forest, support vector machines, neural networks) have been extensively used for this purpose in a variety of geoscience applications like geochemical mapping (Kirkwood et al., 2022, 2016a; Wilford et al., 2016), soil mapping (Wadoux et al., 2020; Taghizadeh-Mehrjardi et al., 2016; Ballabio et al., 2016; Khan et al., 2016; Hengl et al., 2015), hydrological mapping (Barzegar et al., 2016; Appelhans et al., 2015), environmental mapping (Li, 2013; Li et al., 2011), and geological mapping (Albrecht and González-Álvarez, 2021; Kumar et al., 2020; Latifovic et al., 2018; Sahoo and Jha, 2017; Othman and Gloaguen, 2017; Cracknell and Reading, 2015, 2014; Yu et al., 2012). In particular, supervised machine learning methods tailored to spatial data have gained much interest due to their ability to account for the specificities of spatial data, such as spatial autocorrelation (Fouedjio, 2021b; Talebi et al., 2021; Fouedjio, 2021a; Sekulić et al., 2020; Hengl et al., 2018; Fouedjio and Klump, 2019; Fouedjio, 2020). This latter plays a crucial role in the realm of spatial data.

The target variable's observations are commonly available at a few sampling locations due to relatively high acquisition costs and time to obtain measurements. In contrast, auxiliary variables are often available everywhere within the study region through the wider

* Corresponding author.
E-mail address: migrainefrancky.fouedjiokameni@riotinto.com (F. Fouedjio).

https://doi.org/10.1016/j.aiig.2022.12.002
Received 3 November 2022; Received in revised form 21 December 2022; Accepted 21 December 2022
Available online 24 December 2022

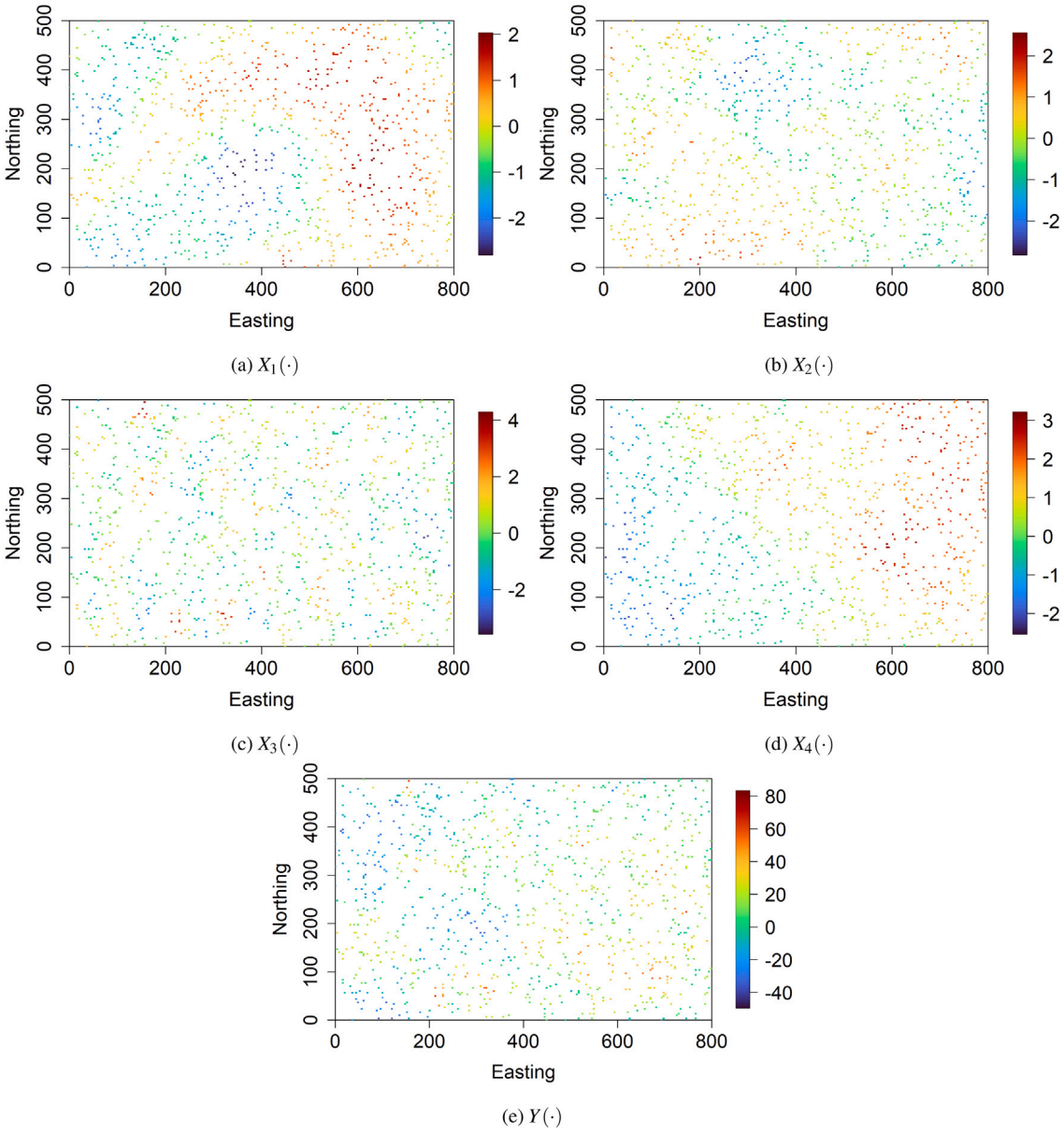


(a) $X_1(\cdot)$ (b) $X_2(\cdot)$ (c) $X_3(\cdot)$ (d) $X_4(\cdot)$ (e) $Y(\cdot)$

Fig. 1. Synthetic case study — (a)–(d) spatially exhaustive auxiliary variables and (e) spatially exhaustive target variable.

development of remote sensing platforms and sensor networks. Thus, spatial data available consists of a very small amount of labeled spatial data (primary spatial data) and a very large amount of unlabeled spatial data (secondary or auxiliary spatial data). This situation is common as acquiring auxiliary spatial data is relatively cheap while collecting primary spatial data is expensive. Supervised machine learning techniques do not fully exploit this vast amount of auxiliary spatial data. Indeed, in these techniques, the training dataset includes only labeled data locations (i.e., data locations at which both the target variable's observations and auxiliary variables' observations are available). At the same time, unlabeled data locations (i.e., data locations at which the target variable's observations are not available but auxiliary variables' observations are known) are not considered during the model training stage. In other words, only colocated auxiliary spatial data are considered during the training phase. Thus, the information provided by auxiliary spatial data beyond those colocated is ignored. As a consequence, the amount of auxiliary spatial data (unlabeled spatial data) used during the model training stage is limited.

Machine learning approaches that effectively exploit large quantities of unlabeled data are gaining interest. One such approach is semi-supervised learning, which learns from both labeled and unlabeled data to build better predictive models (Chapelle et al., 2010; Zhu and Goldberg, 2009). It contrasts supervised learning (data all labeled) and unsupervised learning (data all unlabeled). Semi-supervised learning is of great interest in machine learning because, in addition to labeled data, it can use readily available unlabeled data to improve supervised learning tasks when labeled data are scarce or expensive. One can distinguish two general forms of semi-supervised learning: inductive and transductive semi-supervised learning. They differ by the data availability for which one wants to obtain predictions during the training phase. The goal of inductive semi-supervised learning is to find the function that maps auxiliary variables to the target variable from a combined set of labeled and unlabeled data. This function can then be used to make predictions on new data points that are not available during the training phase. Transductive semi-supervised learning aims to infer the correct labels for the given unlabeled data, which is already present during the training phase. Semi-supervised learning methods can be classified into six categories : self-training, co-training, multi-view learning, expectation–maximization with generative mixture models, graph-based methods, and transductive support



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

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

(e) $Y(\cdot)$

Fig. 2. Synthetic case study — (a)–(d) auxiliary variables and (e) target variable at $n = 1,000$ sampling locations.

vector machines (Chapelle et al., 2010; Zhu and Goldberg, 2009). In particular, self-training combines information from unlabeled data with labeled data to iteratively identify labels for unlabeled data within the dataset. Thus, the labeled training dataset is enlarged at each iteration until the entire dataset is labeled. The self-training algorithm can be applied as a wrapper to any given supervised base learner. The co-training method basically trains two models, while self-training trains one. Thus, self-training can be seen as a particular case of co-training where just one model is trained.

In order to make any use of unlabeled data, some relationship to the underlying distribution of data must exist. A necessary condition of semi-supervised learning is that the underlying marginal data distribution over the attribute space contains information about the posterior distribution. Semi-supervised learning algorithms make use of at least one of the following assumptions: the smoothness assumption (if two data points are close in the attribute (feature) space, their labels should be similar), the low-density assumption (the decision boundary should not pass through high-density areas in the attribute space), and the manifold assumption (data points on the same low-dimensional manifold should have the similar label). These assumptions are the foundation of most, if not all, semi-supervised learning algorithms, which generally depend on one or more of them being satisfied, either explicitly or implicitly. Thus, if sufficient unlabeled data is available and under certain assumptions about the distribution of the data, the unlabeled data can help in the construction of a better predictive model. If, on the other hand, this condition is not met, it is inherently impossible to improve the accuracy of predictions based on the additional unlabeled data (Zhu and Goldberg, 2009). It is worth pointing out that blindly selecting a semi-supervised learning method for a specific task will not necessarily improve performance over supervised learning. In fact, unlabeled data can lead to worse performance with the wrong link assumptions. For a detailed review on semi-supervised learning, refer to Van Engelen and Hoos (2020), Pise and Kulkarni (2008).

While semi-supervised learning for non-spatial data has been gaining much attention, semi-supervised learning dedicated to spatial data has so far been scarce (Kobs et al., 2021; Asghar et al., 2020; Vatsavai et al., 2007). Classical semi-supervised learning techniques have been developed in a non-spatial context. As a consequence, they do

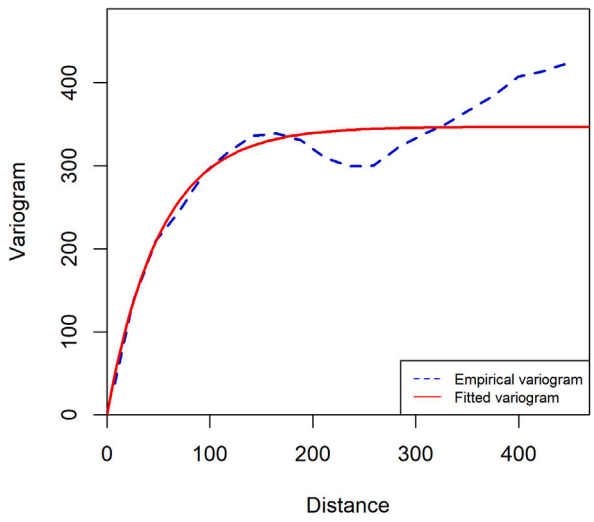


Fig. 3. Synthetic case study — experimental and fitted variograms of the target variable Y in the original training dataset ($n = 1,000$). The fitted variogram model corresponds to an isotropic exponential model with a practical range and sill respectively equal to 155.010 and 346.489.

not account for the characteristics of spatial data, especially spatial autocorrelation, which is widely ignored. In the spatial framework, in addition to the attribute space, the geographical space needs to be taken into account. Indeed, spatial data often show the property of spatial dependency over the study region. Observations located close to one another in the geographical space might have similar characteristics. Considering geographical coordinates as other auxiliary variables is not the best way to account for the spatial information, as shown in many research works (Hengl et al., 2018). Applying traditional semi-supervised learning methods in the realm of spatial data would be liable to produce less accurate predictions. Indeed, the success of semi-supervised learning depends critically on some underlying assumptions (Chapelle et al., 2010; Zhu and Goldberg, 2009). A basic approach to exploit unlabeled spatial data consists of using unlabeled spatial data as additional predictor variables in a supervised learning task. Thus, predictor variables' observations surrounding (at the nearest locations of) a labeled data location are defined as additional covariates. However, under this approach, the number of covariates increases drastically, while the number of observations (samples) is still the same and limited. This situation may lead to the well-known curse of dimensionality (Keogh and Mueen, 2017). As the dimensionality increases, the number of data points required for a good performance of any supervised machine learning algorithm increases exponentially. Thus, a marginal rise in dimensionality also requires a significant increase in the data volume to maintain the same level of performance.

This paper presents a geostatistical semi-supervised learning framework that allows harnessing a large amount of unlabeled spatial data available in many geoscience use cases in combination with typically smaller sets of labeled spatial data. It focuses on semi-supervised regression (i.e., where the target variable is continuous), but the proposed framework can be easily adapted to semi-supervised classification (i.e., for a categorical target variable). It follows the general idea of pseudo labeling (label generation) inherent to classical semi-supervised learning methods such as self-training and co-training. The core idea consists of leveraging the target variable's spatial autocorrelation to generate pseudo labels at unlabeled data points that are geographically close to labeled data points. This is achieved through geostatistical conditional simulation, where an ensemble of pseudo labels is generated to account for the uncertainty in the pseudo labeling process. This ensemble of simulated (pseudo) labels is augmented by the observed (true) labels to create an ensemble of pseudo training datasets. A supervised machine learning model is then trained on each pseudo training dataset, followed by an aggregation of trained models. As a byproduct, the target variable's prediction uncertainty is provided. The proposed geostatistical semi-supervised learning method is illustrated on synthetic spatial data for which ground truth is available everywhere within the study region. It is applied to real-world spatial data for geochemical mapping. Its predictive performance is compared with some classical supervised and semi-supervised learning methods.

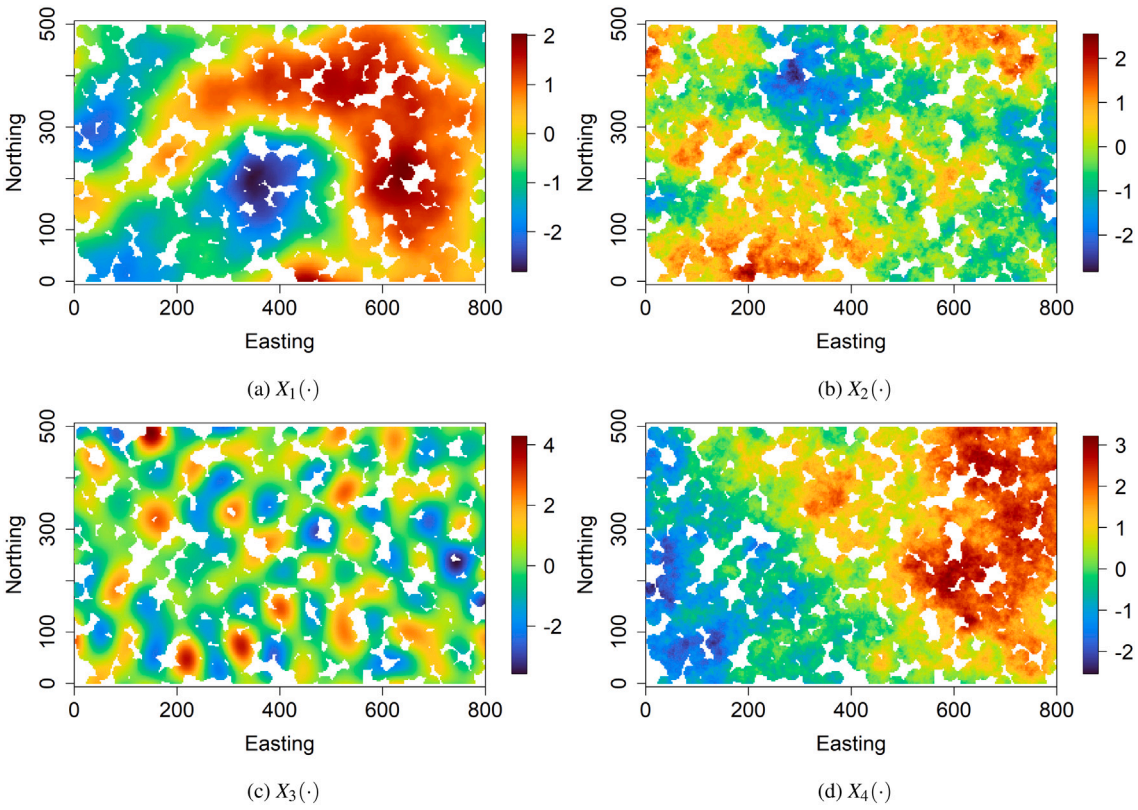
The rest of the paper is organized as follows. Section 2 describes the different ingredients and steps of the proposed geostatistical semi-supervised learning approach. Section 3 illustrates the proposed semi-supervised machine learning method on synthetic spatial data. An application example with real-world spatial data is given in Section 4. Sections 3 and 4 also include a comparison with classical supervised and semi-supervised learning methods. Concluding remarks are summarized in Section 5.

2. Methodology

Let $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ be the target variable (continuous) defined on a fixed continuous spatial domain of interest $D \subset \mathbb{R}^d (d \in \mathbb{N}^*)$. There are p predictor (auxiliary or explanatory) variables $\{\mathbf{X}(\mathbf{s}) = (X^1(\mathbf{s}), \dots, X^p(\mathbf{s})) : \mathbf{s} \in D\}$ exhaustively known in the spatial domain D. Suppose that we have labeled spatial data $\mathcal{L}(\mathbf{s}_1, \dots, \mathbf{s}_n) = \left\{ (\mathbf{X}(\mathbf{s}_1), Y(\mathbf{s}_1)), \dots, (\mathbf{X}(\mathbf{s}_n), Y(\mathbf{s}_n)) \right\}$, with $\{\mathbf{s}_i \in D\}_{i=1,\dots,n}$ representing the target variable's sampling locations. $\mathcal{L}(\mathbf{s}_1, \dots, \mathbf{s}_n)$ depicts the original training dataset. In addition to labeled spatial data, unlabeled spatial data $\mathcal{U}(\mathbf{s}_{n+1}, \dots, \mathbf{s}_N) = \left\{ \mathbf{X}(\mathbf{s}_{n+1}), \dots, \mathbf{X}(\mathbf{s}_N) \right\}$ are available. Data locations $\{\mathbf{s}_i \in D\}_{i=1,\dots,n}$ will refer to labeled data locations and $\{\mathbf{s}_j \in D\}_{j=n+1,\dots,N}$ will denote unlabeled data locations. We are dealing with the situation where relatively few labeled spatial data are available, but a large number of unlabeled spatial data are given ($N \gg n$). The goal is to leverage labeled and unlabeled spatial data in an attempt to improve the target variable's spatial prediction over the spatial domain D. This section describes the different steps and ingredients to implement the proposed geostatistical semi-supervised learning for spatial prediction. The implementation is carried out in the R platform (R Core Team, 2021).

2.1. Generating Pseudo training data

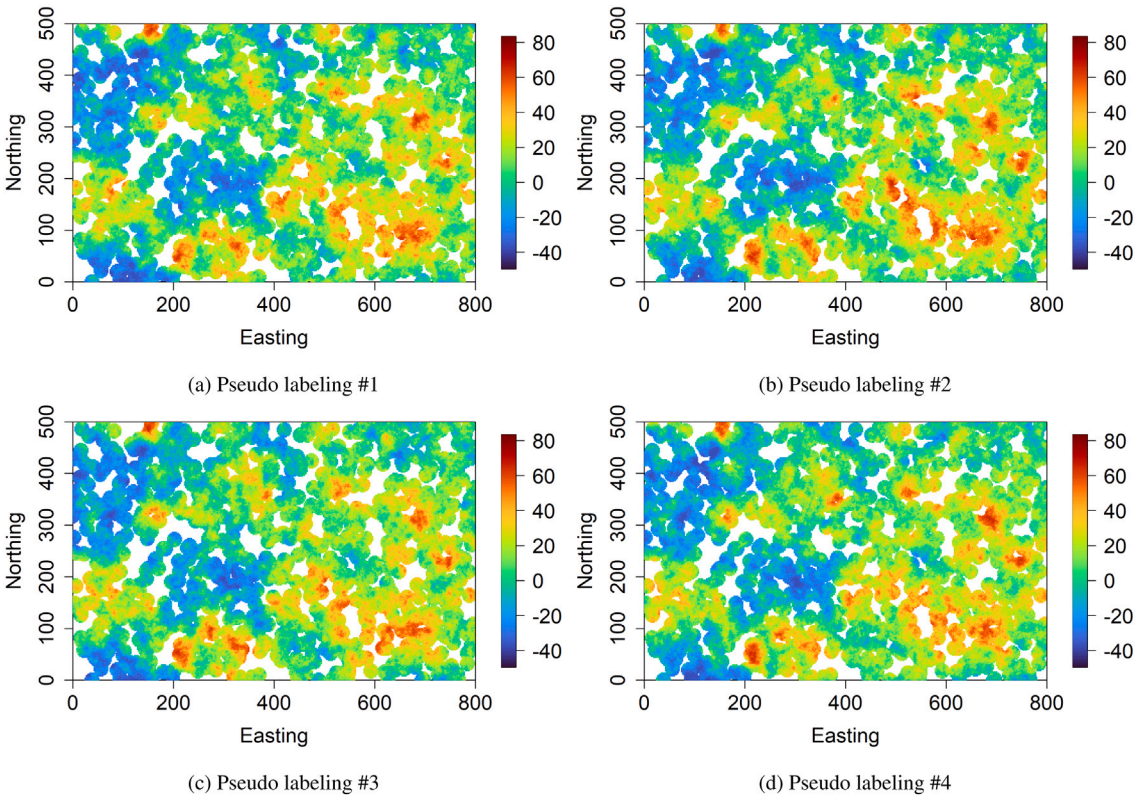
A basic approach to take advantage of unlabeled data is first to predict their labels and add the most confident predicted labels back to the labeled data. This process is known as pseudo labeling in classical semi-supervised learning methods such as self-training and co-training (Van Engelen and Hoos, 2020). In these later, pseudo labeling consists of first training a supervised machine learning model on labeled data. Then, use the predictions of the trained machine learning model to generate additional labeled data. Finally, the original labeled data (observed) and pseudo labeled data (generated) are combined for a final model retraining. The term pseudo labeling is utilized because these pseudo labels are not real (observed) labels. We adopt a different strategy to generate pseudo labels. In the spatial context, one can leverage the target variable's spatial autocorrelation to generate more confidently pseudo labels at unlabeled data locations. The spatial autocorrelation refers to Tobler's first law of geography (Tobler, 1970), meaning that everything is related to everything else, but near things are more related than distant things. In the presence of spatial autocorrelation, closer things tend to be more predictable and have less variability. In contrast, distant things tend to be less predictable and are less related. Thus, not all unlabeled data locations can be labeled with high confidence. The inaccuracy of pseudo labels for unlabeled data locations that are too far from labeled data locations may introduce errors into the machine learning model and cause its degradation. So, not all unlabeled data locations are beneficial for the model training. Our approach



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

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

Fig. 4. Synthetic case study — auxiliary variables at selected unlabeled data locations for pseudo labeling.



(a) Pseudo labeling #1 (b) Pseudo labeling #2

(c) Pseudo labeling #3 (d) Pseudo labeling #4

Fig. 5. Synthetic case study — four randomly selected realizations of pseudo labeling through geostatistical conditional simulation at selected unlabeled data locations.

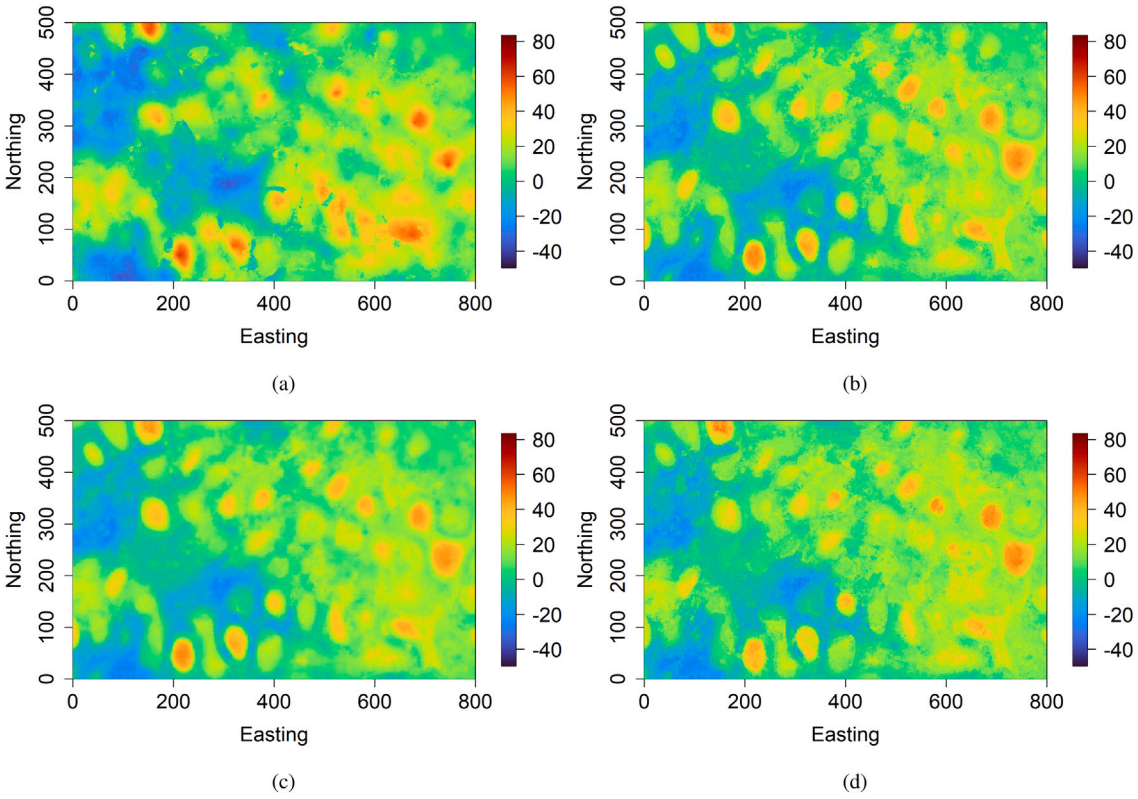


Fig. 6. Synthetic case study — prediction maps provided by (a) geostatistical semi-supervised random forest, (b) classical random forest, (d) random forest with unlabeled spatial data as additional covariates, and (e) self-training random forest.

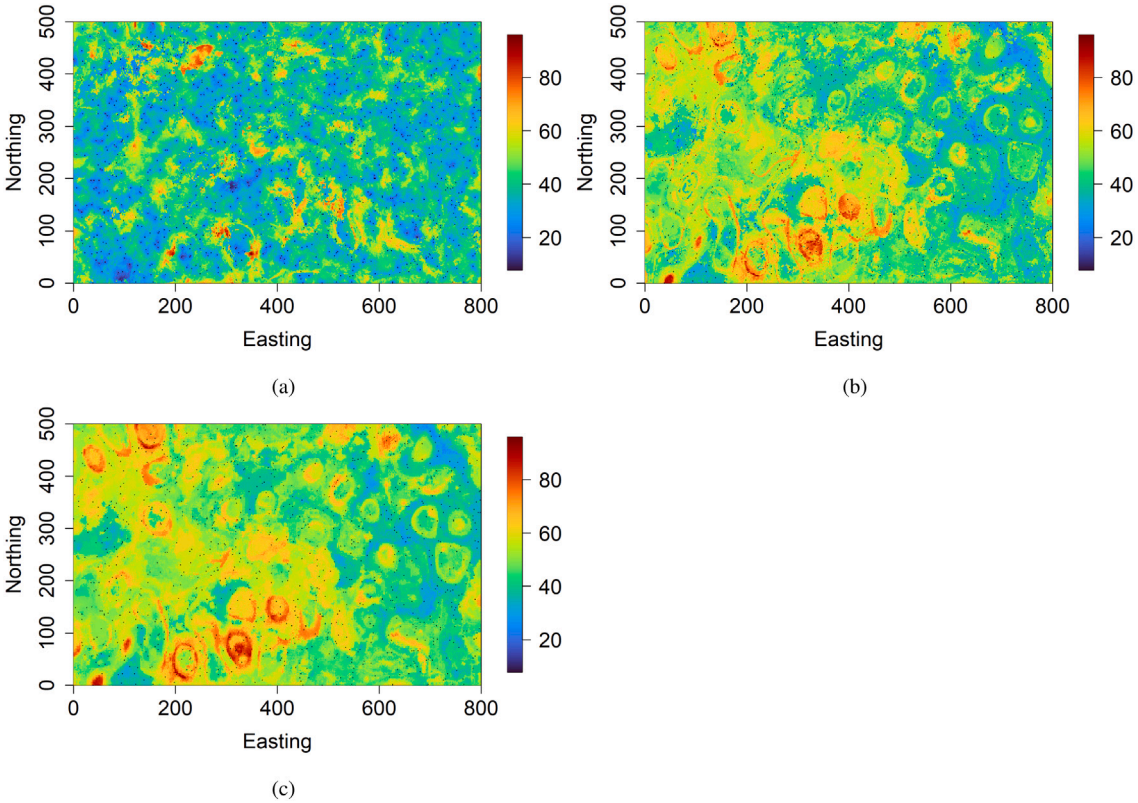


Fig. 7. Synthetic case study — prediction uncertainty maps provided by (a) geostatistical semi-supervised random forest, (b) classical random forest, (c) random forest with unlabeled spatial data as additional covariates. The black dots represent training data locations. The prediction uncertainty corresponds to the width of the 95% prediction interval.

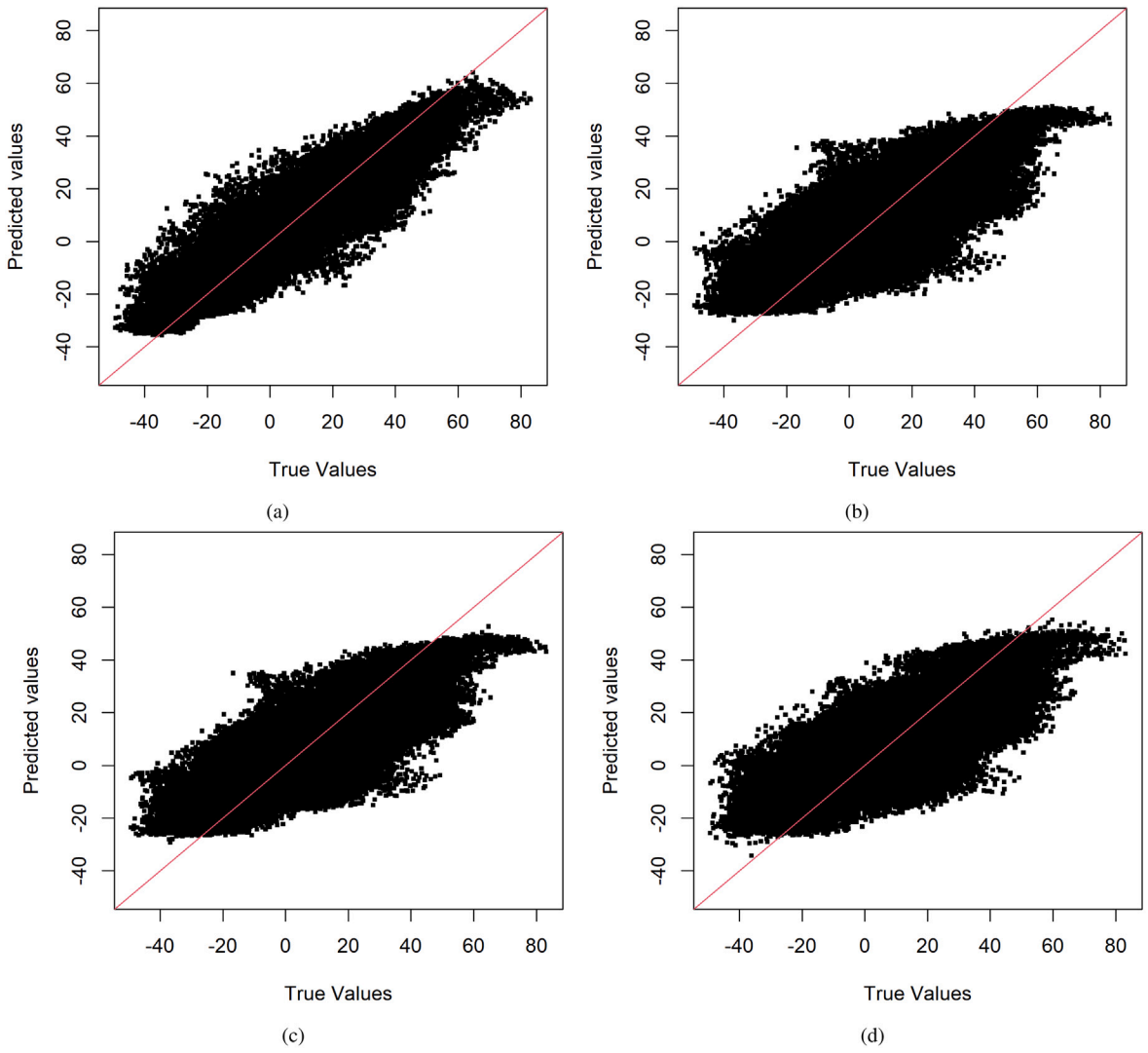
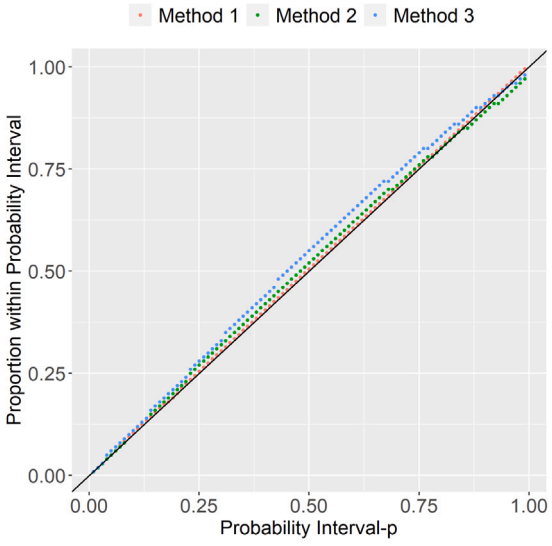


Fig. 8. Synthetic case study — target variable's observed values vs. target variable's predicted values at unlabeled data locations, for (a) geostatistical semi-supervised random forest, (b) classical random forest, (c) random forest with unlabeled spatial data as additional covariates, and (d) self-training random forest. The red line indicates the 1:1 line.

involves leveraging the target variable's spatial autocorrelation to generate pseudo labels at unlabeled data locations that are geographically close to labeled data locations. The target variable's spatial dependence structure is used to define the proximity to the labeled data locations. This latter is usually described using geostatistical tools such as the variogram (Chiles and Delfiner, 2012). One of the main features of the variogram, the range, is used to define the proximity to the labeled data locations. The range is the distance after which the variogram levels off. The physical meaning of the range is that pairs of locations that are this distance or greater apart are not spatially correlated. In other words, points further apart than the range are spatially independent.

The starting point of the geostatistical semi-supervised learning approach consists of estimating and modeling the target variable's spatial dependence structure (variogram) using the target variable's observations. This is achieved by calculating the sample (experimental) variogram through the method of moments, followed by the fitting (e.g., manual or automated fitting) to a class of theoretical parametric variogram models (refer to Chiles and Delfiner (2012), for more details about variogram estimating and modeling). Given labeled data locations, any unlabeled data location that falls within a neighborhood centered at a labeled data location with a radius equal to a quarter of the variogram's range of the target variable is selected for the pseudo labeling. Thus, only unlabeled data locations that have a strong spatial dependency with labeled data locations are considered. Let $\{\mathbf{s}_{t_k} \in D\}_{k=1,\ldots,M;t_k\in\{n+1,\ldots,N\}}$ be the set of unlabeled data locations selected for the pseudo labeling. The pseudo labeling is carried out through geostatistical conditional simulation. Geostatistical simulation is a spatial extension of the Monte Carlo simulation concept. In addition to reproducing the data histogram, geostatistical simulation also honors the data's spatial dependence structure (variogram). A simulation method that honors the data in the sense that the simulated value at a sampling location exactly matches the observed value is termed a conditional simulation. In contrast, a simulation method that does not honor the data is called unconditional simulation. It is worth highlighting that geostatistical simulation accounts for non-Gaussian data. In this latter case, a Gaussian transformation of the target variable is part of the simulation process (refer to Chiles and Delfiner (2012) and Lantuejoul (2013) for more details about geostatistical simulation).

The geostatistical conditional simulation method used for generating the pseudo labels is the well-known conditioning by kriging, which is the combination of a non-conditional simulation method (e.g., spectral turning bands method) and kriging (Chiles and Delfiner, 2012). Let $\left\{Y^l(\mathbf{s}_{t_1}),\ldots,Y^l(\mathbf{s}_{t_M})\right\}_{l=1,\ldots,L}$ be the ensemble of pseudo (plausible) labels generated at the selected unlabeled data locations. Geostatistical simulation allows obtaining multiple pseudo labels at each selected unlabeled data location instead of a single pseudo label.



(a)

Fig. 9. Synthetic case study — prediction interval coverage probability plot (accuracy plot). Method 1 refers to the geostatistical semi-supervised random forest; Method 2 refers to the classical random forest; Method 3 refers to the random forest with unlabeled spatial data as additional covariates.

Thus, the uncertainty in the pseudo labeling process is taken into account. Given the set of pseudo labels $\left\{Y^l(\mathbf{s}_{t_1}),\ldots,Y^l(\mathbf{s}_{t_M})\right\}_{l=1,\ldots,L}$, one forms the following ensemble of pseudo labeled spatial data $\left\{\left(\mathbf{X}(\mathbf{s}_{t_1}),Y^l(\mathbf{s}_{t_1})\right),\ldots,\left(\mathbf{X}(\mathbf{s}_{t_M}),Y^l(\mathbf{s}_{t_M})\right)\right\}_{l=1,\ldots,L}$. This latter ensemble is used to augment the original training dataset $\mathcal{L}(\mathbf{s}_1,\ldots,\mathbf{s}_n)$. This results in the following ensemble of pseudo training datasets: $\left\{\mathcal{L}_l(\mathbf{s}_1,\ldots,\mathbf{s}_n,\mathbf{s}_{t_1},\ldots,\mathbf{s}_{t_M})\right\}_{l=1,\ldots,L}$ where each $\mathcal{L}_l(\mathbf{s}_1,\ldots,\mathbf{s}_n,\mathbf{s}_{t_1},\ldots,\mathbf{s}_{t_M}) = \left\{(\mathbf{X}(\mathbf{s}_1),Y(\mathbf{s}_1)),\ldots,(\mathbf{X}(\mathbf{s}_n),Y(\mathbf{s}_n)),\left(\mathbf{X}(\mathbf{s}_{t_1}),Y^l(\mathbf{s}_{t_1})\right),\ldots,\left(\mathbf{X}(\mathbf{s}_{t_M}),Y^l(\mathbf{s}_{t_M})\right)\right\}, l = 1,\ldots,L$. Thus, the size of the original training dataset is increased by M, the number of selected unlabeled data locations. The training of a supervised machine learning model can be improved by exposing it to as much data as possible. The pseudo labeled spatial data provide context that will aid the training of a supervised machine learning model. It is important to note that each pseudo training dataset contains the original training dataset. Geostatistical conditional simulation is used here as a label generator and data augmentation approach. In many cases, at least $L = 50$ simulations will provide sufficient information to account for the uncertainty in the pseudo labeling. A larger number of simulations allows a better account for the uncertainty but requires more computing time. However, it is essential to note that the generation of pseudo labels can be performed in parallel.

2.2. Training supervised machine learning model

Given the ensemble of pseudo training datasets $\left\{\mathcal{L}_l(\mathbf{s}_1,\ldots,\mathbf{s}_n,\mathbf{s}_{t_1},\ldots,\mathbf{s}_{t_M})\right\}_{l=1,\ldots,L}$ generated at the previous section, this step consists of training a supervised machine learning model on each pseudo training dataset. We opt for the regression random forest, but any other supervised machine learning method can be used as well. The regression random forest popularity for spatial prediction relies on its ability to efficiently deal with many predictor variables, handle complex nonlinear relationships and interactions, require less data preprocessing, and be a non-parametric method (model-free). Regression random forest is a type of ensemble machine learning method that constructs a multitude of regression tree models on various subsets of the training dataset (bootstrap samples) using different subsets of available predictor variables, followed by aggregation. Under random forest, each built regression tree model is unique (less correlated with others) due to the bootstrapping of the training data and the random selection of subsets of predictor variables. The multiple regression tree models knitted together reduce the prediction variance and increase prediction accuracy. The regression random forest's prediction is obtained by averaging all of the regression tree's predictions.

For each pseudo training dataset $\mathcal{L}_l(\mathbf{s}_1,\ldots,\mathbf{s}_n,\mathbf{s}_{t_1},\ldots,\mathbf{s}_{t_M})$, a random forest regressor $\{\hat{f}_l(\mathbf{X}(\mathbf{s})) : \mathbf{s}\in D\}$ is built $(l = 1,\ldots,L)$. The number of trees is set to 1,000, and the other hyper-parameters are optimized through cross-validation. They include the number of predictor variables randomly selected at each node, the proportion of observations to sample in each regression tree, and the minimum number of observations in a regression tree's terminal node. The ensemble of random forest regressors $\left\{\hat{f}_l(\mathbf{X}(\mathbf{s})) : \mathbf{s}\in D\right\}_{l=1,\ldots,L}$ is aggregated to provide the final model as follows:

$$\hat{Y}(\mathbf{s}) = \hat{f}(\mathbf{X}(\mathbf{s})) = \frac{1}{L}\sum_{l=1}^{L}\hat{f}_l(\mathbf{X}(\mathbf{s})), \quad \forall \mathbf{s}\in D. \tag{1}$$

The ensemble of random forest regressors $\left\{\hat{f}_l(\mathbf{X}(\mathbf{s})) : \mathbf{s}\in D\right\}_{l=1,\ldots,L}$ can be also used to generate the prediction uncertainty. This can be achieved by calculating the inter-percentile range of the predictions ensemble. Let $\hat{F}(\cdot;\mathbf{s})$ be the predictive distribution of the target variable Y at the target location $\mathbf{s}\in D$, obtained from the predictions ensemble $\left\{\hat{f}_l(\mathbf{X}(\mathbf{s}))\right\}_{l=1,\ldots,L}$. A $100(1-\alpha)\%$ prediction interval of the target variable Y at location $\mathbf{s}\in D$ is expressed as $[\hat{Q}_{\alpha/2}(\mathbf{s}),\hat{Q}_{1-\alpha/2}(\mathbf{s})](0 < \alpha < 1)$; where $\hat{Q}_{\alpha/2}(\mathbf{s})$ denotes the α-quantile of $\hat{F}(\cdot;\mathbf{s})$ and is defined as $\hat{Q}_{\alpha/2}(\mathbf{s}) = \inf\{y : \hat{F}(y;\mathbf{s}) \geq \alpha\}$. Thus, the inter-percentile range $\hat{Q}_{1-\alpha/2}(\mathbf{s})-\hat{Q}_{\alpha/2}(\mathbf{s})$ can be viewed as a measure of prediction uncertainty. In particular, $\hat{Q}_{0.975}(\mathbf{s})-\hat{Q}_{0.025}(\mathbf{s})$ which corresponds to the width of the 95% prediction interval $[\hat{Q}_{0.025}(\mathbf{s}),\hat{Q}_{0.975}(\mathbf{s})]$ will be used as a measure of prediction uncertainty in the case studies presented at Sections 3 and 4.

The following R packages are used to perform the regression random forest: *ranger* (Wright and Ziegler, 2017) and *tuneRanger* (Probst et al., 2018). It is important to remark that the construction of the ensemble of random forest regressors $\left\{\hat{f}_l(\mathbf{X}(\mathbf{s})) : \mathbf{s}\in D\right\}_{l=1,\ldots,L}$ can be conducted in parallel. To summarize, the geostatistical semi-supervised learning approach is implemented using the pseudo algorithm 1.

2.3. Evaluating machine learning model performance

To assess the proposed spatial semi-supervised learning approach's ability to predict the target variable accurately, a comparison will be conducted with classical supervised and semi-supervised machine learning methods. The comparison will be performed using testing data, i.e., data kept aside for the whole analysis. The criteria used to evaluate the prediction accuracy quantitatively are: the mean absolute error (MAE), the root mean square error (RMSE), the coefficient of determination (R-square), and Lin's concordance correlation coefficient (CCC) (Steichen and Cox, 2002). The lower are MAE and RMSE, the better the model is. R-square and CCC close to 1 indicate a perfect model. These criteria are computed as follows:

$$MAE = \frac{1}{r}\sum_{j=1}^{r}|Y_j-\hat{Y}_j|, \tag{2}$$

$$RMSE = \sqrt{\frac{1}{r}\sum_{j=1}^{r}(Y_j-\hat{Y}_j)^2}, \tag{3}$$

$$R^2 = 1-\frac{\sum_{j=1}^{r}(Y_j-\hat{Y}_j)^2}{\sum_{j=1}^{r}(Y_j-\mu_Y)^2}, \tag{4}$$

$$CCC = \frac{2\times\rho\times\sigma_{\hat{Y}}\times\sigma_Y}{\sigma_{\hat{Y}}^2+\sigma_Y^2+(\mu_{\hat{Y}}-\mu_Y)^2} \tag{5}$$

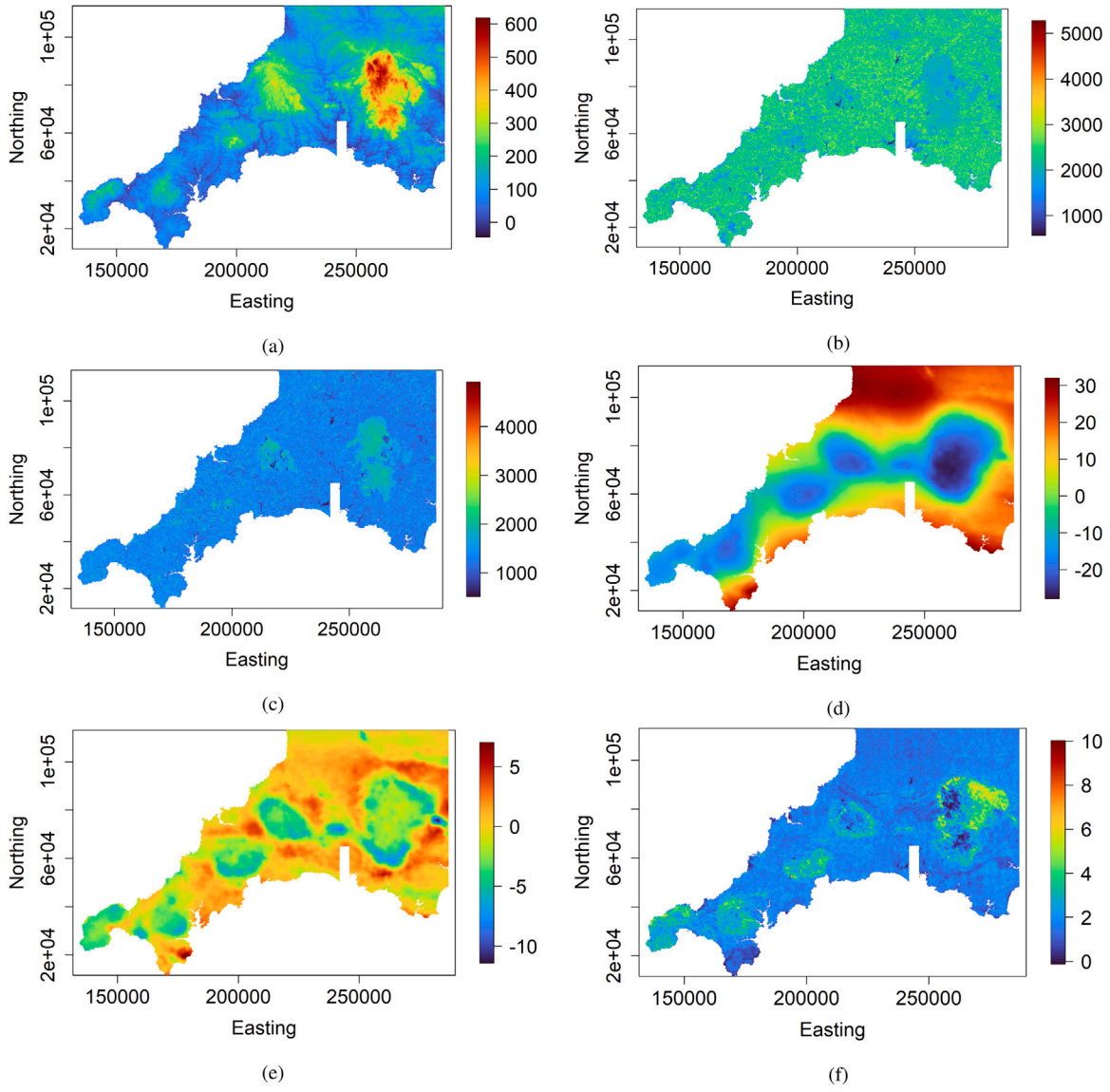


Fig. 10. Real case study — spatially exhaustive auxiliary variables: (a) elevation, (b) Landsat 8 band 5, (c) Landsat 8 band 6, (d) gravity survey Bouguer anomaly, (e) gravity survey high-pass filtered Bouguer anomaly, and (f) Uranium counts from gamma-ray spectrometry.

Algorithm 1 Geostatistical Semi-Supervised Learning for Spatial Prediction

Input: labeled spatial data $\mathcal{L}(\mathbf{s}_1, \ldots, \mathbf{s}_n) = \left\{ \left(\mathbf{X}(\mathbf{s}_1), Y(\mathbf{s}_1)\right), \ldots, \left(\mathbf{X}(\mathbf{s}_n), Y(\mathbf{s}_n)\right) \right\}$ and unlabeled spatial data $\mathcal{U}(\mathbf{s}_{n+1}, \ldots, \mathbf{s}_N) = \left\{ \mathbf{X}(\mathbf{s}_{n+1}), \ldots, \mathbf{X}(\mathbf{s}_N) \right\}$; $\mathbf{s}_i \in D \subset \mathbb{R}^d, \quad i = 1, \ldots, N;\ N \gg n$.

1. Estimate and model the target variable's variogram using observations $\{Y(\mathbf{s}_i)\}_{i=1,\ldots,n}$;

2. Select unlabeled data locations for which the distance to the closest labeled data location is less than the quarter of the variogram range of the target variable; the outcome is a set $\{\mathbf{s}_{t_k} \in D\}_{k=1,\ldots,M; t_k \in \{n+1,\ldots,N\}}$;

3. Generate pseudo labels at the selected unlabeled data locations through geostatistical conditional simulation; the output is an ensemble of pseudo labeled spatial data $\left\{ \left(\mathbf{X}(\mathbf{s}_{t_1}), Y^l(\mathbf{s}_{t_1})\right), \ldots, \left(\mathbf{X}(\mathbf{s}_{t_M}), Y^l(\mathbf{s}_{t_M})\right) \right\}_{l=1,\ldots,L}$;

4. Form the ensemble of pseudo training datasets $\left\{ \left(\mathbf{X}(\mathbf{s}_1), Y(\mathbf{s}_1)\right), \ldots, \left(\mathbf{X}(\mathbf{s}_n), Y(\mathbf{s}_n)\right), \left(\mathbf{X}(\mathbf{s}_{t_1}), Y^l(\mathbf{s}_{t_1})\right), \ldots, \left(\mathbf{X}(\mathbf{s}_{t_M}), Y^l(\mathbf{s}_{t_M})\right) \right\}_{l=1,\ldots,L}$;

5. Train a supervised machine learning model for each pseudo training dataset; this results to an ensemble of regressors $\left\{ \hat{f}_l(\mathbf{X}(\mathbf{s})) : \mathbf{s} \in D \right\}_{l=1,\ldots,L}$;

6. Aggregate the ensemble of regressors through averaging: $\widehat{f}(\mathbf{X}(\mathbf{s})) = \frac{1}{L} \sum_{l=1}^{L} \hat{f}_l(\mathbf{X}(\mathbf{s})), \ \forall \mathbf{s} \in D$;

Output: predicted target variable $\{\widehat{Y}(\mathbf{s}) = \widehat{f}(\mathbf{X}(\mathbf{s})) : \mathbf{s} \in D\}$ over the spatial domain D.

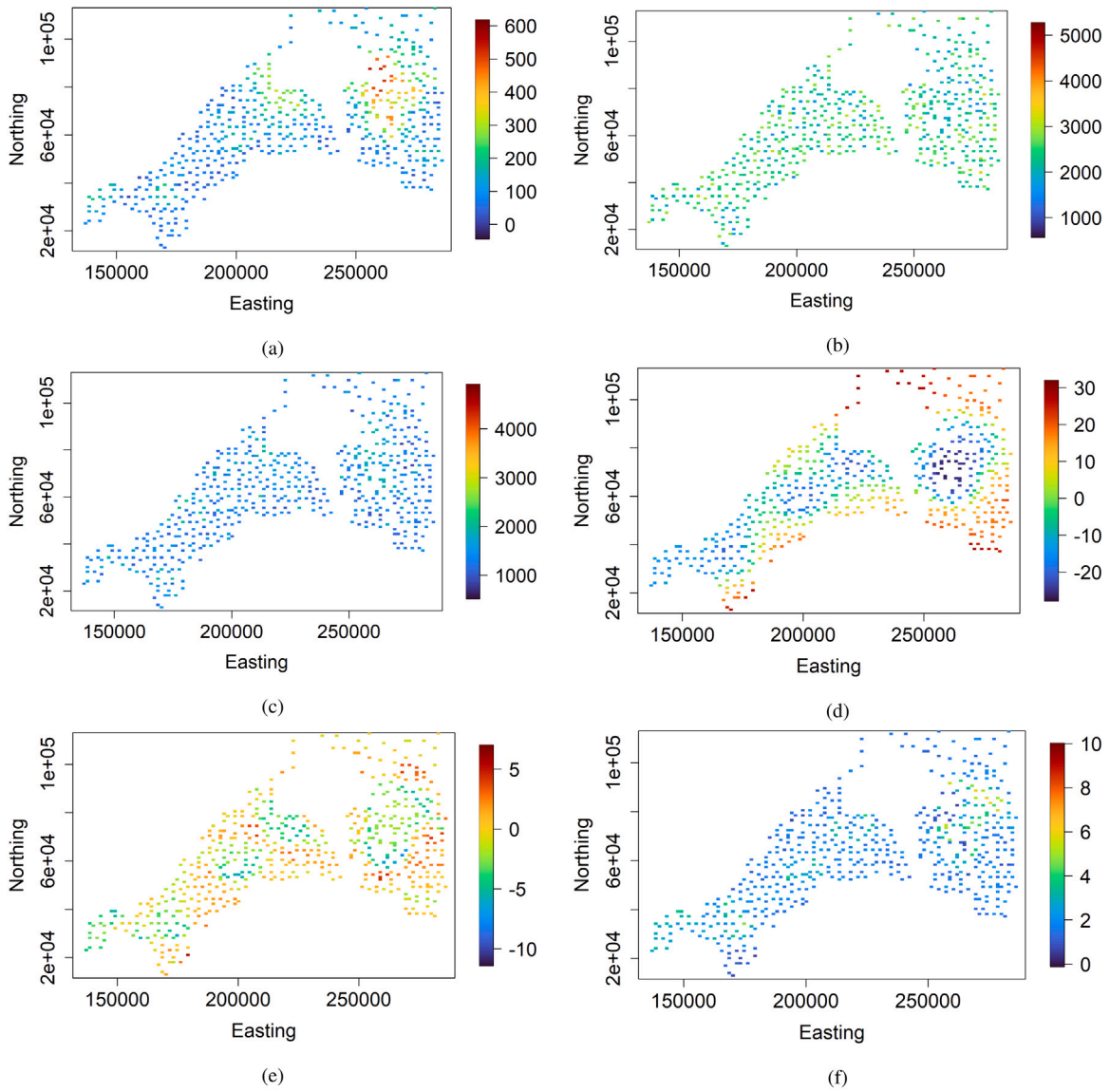


Fig. 11. Real case study — auxiliary variables at sampling locations: (a) elevation, (b) Landsat 8 band 5, (c) Landsat 8 band 6, (d) gravity survey Bouguer anomaly, (e) gravity survey high-pass filtered Bouguer anomaly, and (f) Uranium counts from gamma-ray spectrometry.

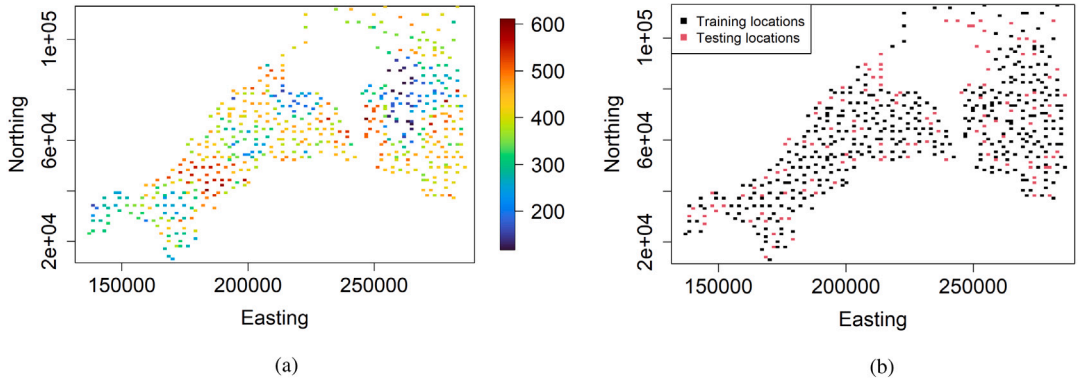
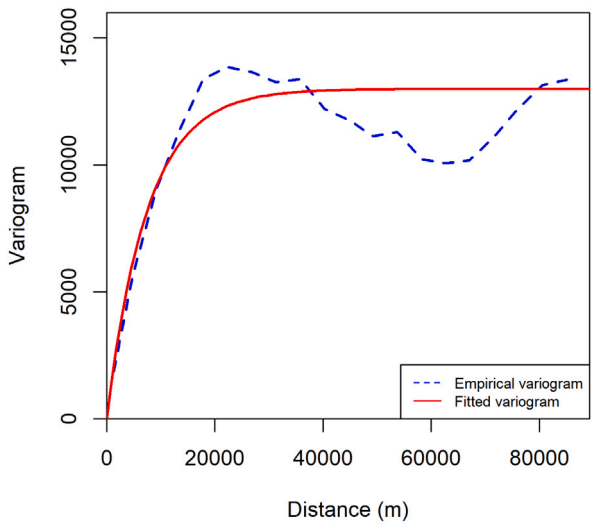


Fig. 12. Real case study — (a) target variable at sampling locations; (b) training and testing data locations.

where $\{Y_j\}_{j=1,\ldots,r}$ are actual values of the target variable at testing locations $\{\mathbf{u}_j \in D\}_{j=1,\ldots,r}$; $\{\widehat{Y}_j\}_{j=1,\ldots,r}$ are predicted values of the target variable at testing locations $\{\mathbf{u}_j \in D\}_{j=1,\ldots,r}$; r is the total number of testing locations; μ_Y and σ_Y^2 are respectively, the mean and variance of actual values of the target variable at testing locations; $\mu_{\widehat{Y}}$ and $\sigma_{\widehat{Y}}^2$ are respectively, the mean and variance of predicted values of the target variable at testing locations; ρ is the correlation coefficient between predicted and observed values at testing locations $\{\mathbf{u}_j \in D\}_{j=1,\ldots,r}$.



(a)

Fig. 13. Real case study — variograms (experimental and fitted) of the target variable in the original training dataset. The fitted variogram model corresponds to an isotropic stationary exponential model with a practical range and sill respectively equal to 22,551.842 and 12,983.605.

The tool used to evaluate the prediction uncertainty accuracy is the prediction interval coverage probability plot, also referred to as the accuracy plot (Fouedjio and Klump, 2019). Given the target variable's measurements $\{Y_j\}_{j=1,\dots,r}$ at testing locations $\{\mathbf{u}_j \in D\}_{j=1,\dots,r}$, the accuracy plot (which is a scatter-plot) compares the proportion of values of the testing dataset falling into the symmetric p-probability interval (PI) with the expected probability p. By construction, there is a probability p $(0 \leq p \leq 1)$ that the true value of the target of variable falls into the symmetric p-probability interval $\left[\widehat{Q}_{\frac{(1-p)}{2}}(j), \widehat{Q}_{\frac{(1+p)}{2}}(j)\right]$; $\widehat{Q}_{\frac{(1-p)}{2}}(j)$ and $\widehat{Q}_{\frac{(1+p)}{2}}(j)$ are the $\frac{(1-p)}{2}$ and $\frac{(1+p)}{2}$ quantiles of the predictive distribution of the target variable at testing location $\mathbf{u}_j$. The fraction of target variable's true values falling into the symmetric p-probability interval is given as: $\bar{\kappa}(p) = \frac{1}{r}\sum_{j=1}^{r} \kappa_j(p)$, where $\kappa_j(p) = 1$ if $\widehat{Q}_{\frac{(1-p)}{2}}(j) < Y_j < \widehat{Q}_{\frac{(1+p)}{2}}(j)$ and 0 otherwise. The more points are closer to the 45 degree line in the accuracy plot the well-calibrated the model is. Specifically, the closeness of points to the bisector of the accuracy plot is quantified using the goodness statistic (G-statistic): $G = 1 - \int_0^1 [3a(p) - 2][\bar{\kappa}(p) - p]dp$, with $a(p) = 1_{\bar{\kappa}(p)>p}$. The value of G lies in the interval $[0, 1]$. $G = 1$ 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. $\bar{\kappa}(p) = 0$, $\forall p \in [0, 1]$. The higher is the value of G the well-calibrated the model is.

3. Synthetic case study

The geostatistical semi-supervised learning based on random forest is first applied to synthetic spatial data for which the ground truth is exhaustively available over the study region. The comparison is carried out with classical random forest, random forest with unlabeled spatial data treated as additional covariates, and self-training random forest. The first two machine learning methods are supervised, while the last one is semi-supervised. All competing machine learning methods are based on random forest. The number of decision trees (1,000) is the same for all the competing machine learning methods. The other hyper-parameters are optimized through cross-validation. The classical random forest is implemented in the R packages *ranger* (Wright and Ziegler, 2017) and *tuneRanger* (Probst et al., 2018). The self-training random forest is implemented in the R package *ssr* (Garcia-Ceja, 2019).

Table 1
Synthetic case study — simulation parameters.

Variables	Mean	Variogram type	Variogram scale	Variogram sill
$X_1(\cdot)$	0	Cubic	300	1
$X_2(\cdot)$	0	Spherical	300	1
$X_3(\cdot)$	0	Cardinal Sine	15	1
$X_4(\cdot)$	0	Linear	0.003	–
$\zeta(\cdot)$	0	Exponential	130	400

The synthetic spatial data is generated using the following model:

$$Y(\mathbf{s}) = 5X_1(\mathbf{s}) \times X_2(\mathbf{s}) + 5X_3(\mathbf{s})^2 + 15\sin\left(X_4(\mathbf{s})\right) + \zeta(\mathbf{s}), \quad \forall \mathbf{s} \in [0, 800] \times [0, 500], \tag{6}$$

where $X_1(\cdot)$, $X_2(\cdot)$, $X_3(\cdot)$, and $X_4(\cdot)$ are auxiliary variables, $\zeta(\cdot)$ is the latent (unobserved) variable, and $Y(\cdot)$ is the target variable. $X_1(\cdot)$, $X_2(\cdot)$, $X_3(\cdot)$, and $X_4(\cdot)$, and $\zeta(\cdot)$ are independent Gaussian random fields with mean and isotropic stationary variogram model specified in Table 1. For background on Gaussian random fields, refer to Chiles and Delfiner (2012). The auxiliary, latent, and target variables are generated over a 400 × 250 regular grid in the spatial domain $[0, 800] \times [0, 500]$. The simulation is carried out via the turning bands method implemented in the R package *RGeostats* (Renard et al., 2021). This simulated example depicts a situation where there is a non-linear relationship between the target variable and auxiliary variables with some interactions between auxiliary variables. Also, the target variable shows some spatial autocorrelation, and its distribution is non-Gaussian.

Fig. 1 displays the target and auxiliary variables maps. There are $N = 100,000$ data locations. Fig. 2 depicts $n = 1,000$ data locations sampled randomly to form labeled spatial data $\mathcal{L}(\mathbf{s}_1, \dots, \mathbf{s}_n) = \left\{\left(\mathbf{X}(\mathbf{s}_1), Y(\mathbf{s}_1)\right), \dots, \left(\mathbf{X}(\mathbf{s}_n), Y(\mathbf{s}_n)\right)\right\}$ (original training dataset). The remainder of the data locations ($N - n = 99{,}000$) is considered as unlabeled spatial data $\mathcal{U}(\mathbf{s}_{n+1}, \dots, \mathbf{s}_N) = \left\{\mathbf{X}(\mathbf{s}_{n+1}), \dots, \mathbf{X}(\mathbf{s}_N)\right\}$, after hiding the corresponding target variable's observations $\{Y(\mathbf{s}_{n+1}), \dots, Y(\mathbf{s}_N)\}$. Thus, labeled and unlabeled spatial data represent respectively 1% and 99% of the data. So, we are in a situation where $N \gg n$. The goal is to predict the target variable over the same 400 × 250 regular grid using the labeled and unlabeled spatial data. The predictions will be compared to the hidden target variable's observations $\{Y(\mathbf{s}_{n+1}), \dots, Y(\mathbf{s}_N)\}$.

The first step of the geostatistical semi-supervised learning method consists of estimating and modeling the target variable's variogram from the target variable's observations. The corresponding variograms (experimental and fitted) are presented in Fig. 3. One can clearly see that the target variable shows some spatial autocorrelation. The fitted variogram model corresponds to an isotropic exponential model with a practical range and sill respectively equal to 155.010 and 346.489. The second step consists in selecting unlabeled data locations for which the distance to the closest labeled data location is less than a quarter of the variogram range of the target variable. The selected unlabeled data locations ($M = 82{,}912$) are shown in Fig. 4. They represent approximately 84% of the ensemble of unlabeled data locations. In the third step, pseudo labels at the selected unlabeled data locations are generated through geostatistical conditional simulation. Pseudo labels are generated not once but several times ($L = 50$) to account for the pseudo labeling process's uncertainty. The pseudo labeling performed through geostatistical conditional simulation is such that pseudo labels at labeled data locations match exactly the true (observed) labels. Four randomly selected realizations of pseudo labels at the selected unlabeled data locations are given in Fig. 5. The resulting ensemble of pseudo labels is augmented to the labeled spatial data (original training dataset) to create an ensemble of pseudo training datasets. The next step consists in training a random forest model for each pseudo training dataset, followed by an aggregation of the models through averaging.

Fig. 6 presents prediction maps provided by the geostatistical semi-supervised random forest, the classical random forest, the random

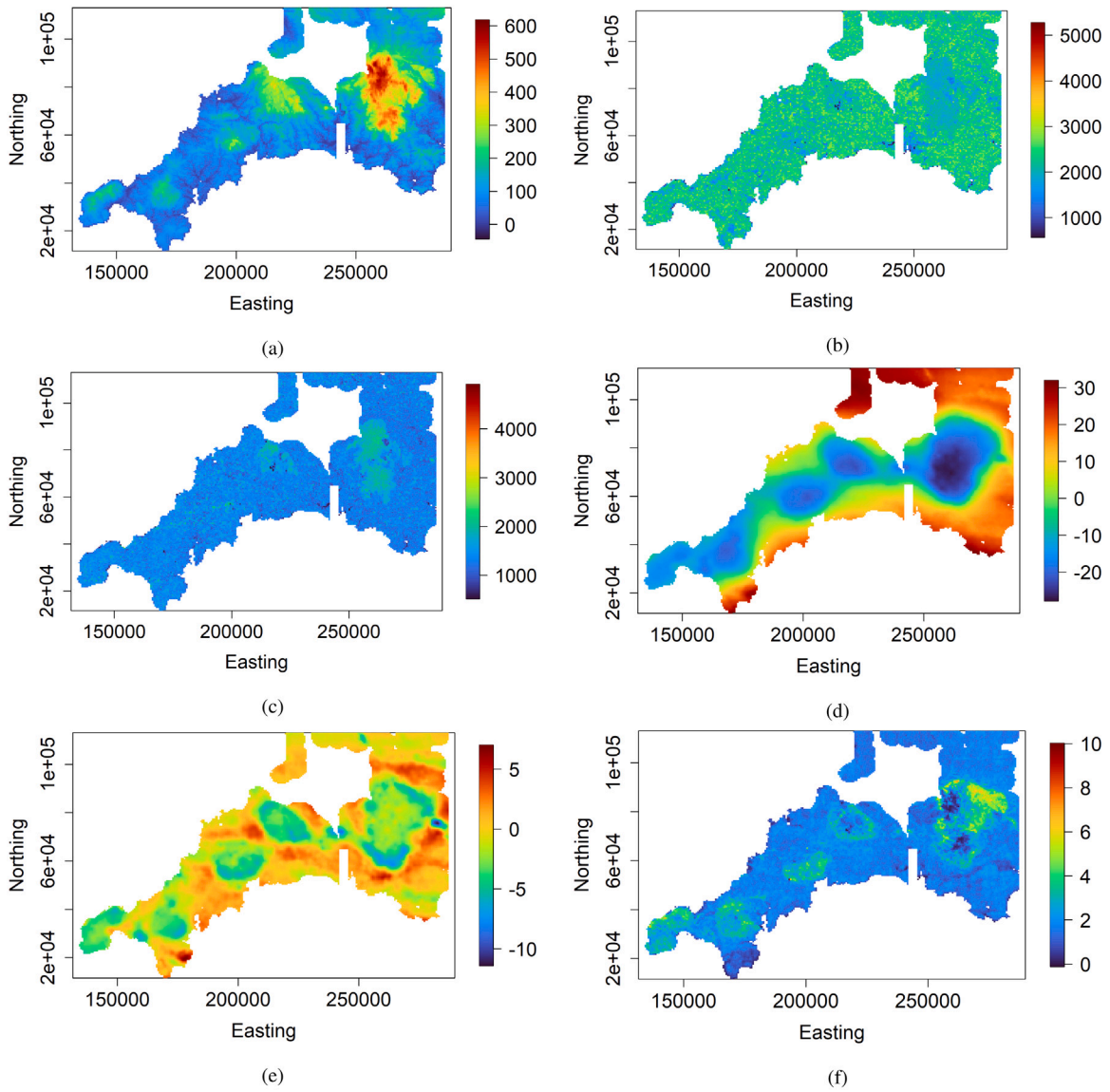


Fig. 14. Real case study — auxiliary variables at selected unlabeled data locations for pseudo labeling: (a) elevation, (b) Landsat 8 band 5, (c) Landsat 8 band 6, (d) gravity survey Bouguer anomaly, (e) gravity survey high-pass filtered Bouguer anomaly, and (f) Uranium counts from gamma-ray spectrometry.

forest with unlabeled spatial data defined as additional covariates, and the self-training random forest. It is important to highlight that all competing machine learning methods rely on random forest with the same number of decision trees set to 1,000. The other hyper-parameters have been selected via cross-validation. In the random forest with unlabeled spatial data treated as additional predictor variables, the ten closest predictor variables' observations surrounding a labeled data location are defined as additional predictor variables. So, there are $10 \times 4 = 40$ additional covariates. The general appearance of prediction maps resulting from the classical random forest, the random forest with unlabeled spatial data defined as additional covariates, and the self-training random forest looks similar. The proposed method's prediction map differs from the other methods. In particular, it is more similar to the ground truth than the other methods. The proposed method's prediction map shows some patterns existing in the target variable's reference map that are missing in other methods' prediction maps. Fig. 7 shows the prediction uncertainty maps (corresponding to the width of the 95% prediction interval) of the different machine learning methods except for the self-training random forest. Available R packages for self-training random forest do not return individual predictions for each tree which is necessary to compute the prediction uncertainty; only aggregated predictions for all trees are provided. The prediction uncertainty map resulting from the proposed method differs notably from the others. In particular, under the proposed method, the prediction uncertainty tends to be lower at the original training data locations than at other locations, which is not the case for the other methods.

Table 2, Figs. 8 and 9 show the predictive performance of the different machine learning methods on the unlabeled spatial data (99,000 observations). One can observe that the geostatistical semi-supervised random forest provides better predictive performance than the other methods. In particular, the self-training random forest (classical semi-supervised learning) is the worst. The random forest with unlabeled spatial data treated as additional predictor variables performs slightly better than the classical random forest. The underperformance of the classical semi-supervised learning method (self-training random forest) can be explained by the nature of the data. Classical semi-supervised learning methods have been designed for non-spatial data. As highlighted in Section 1, unlabeled data can lead to worse performance if the underlying assumptions are not met. In particular, the smoothness assumption as defined in the classical semi-supervised learning may not be valid in the spatial framework. Two data points that are close in the

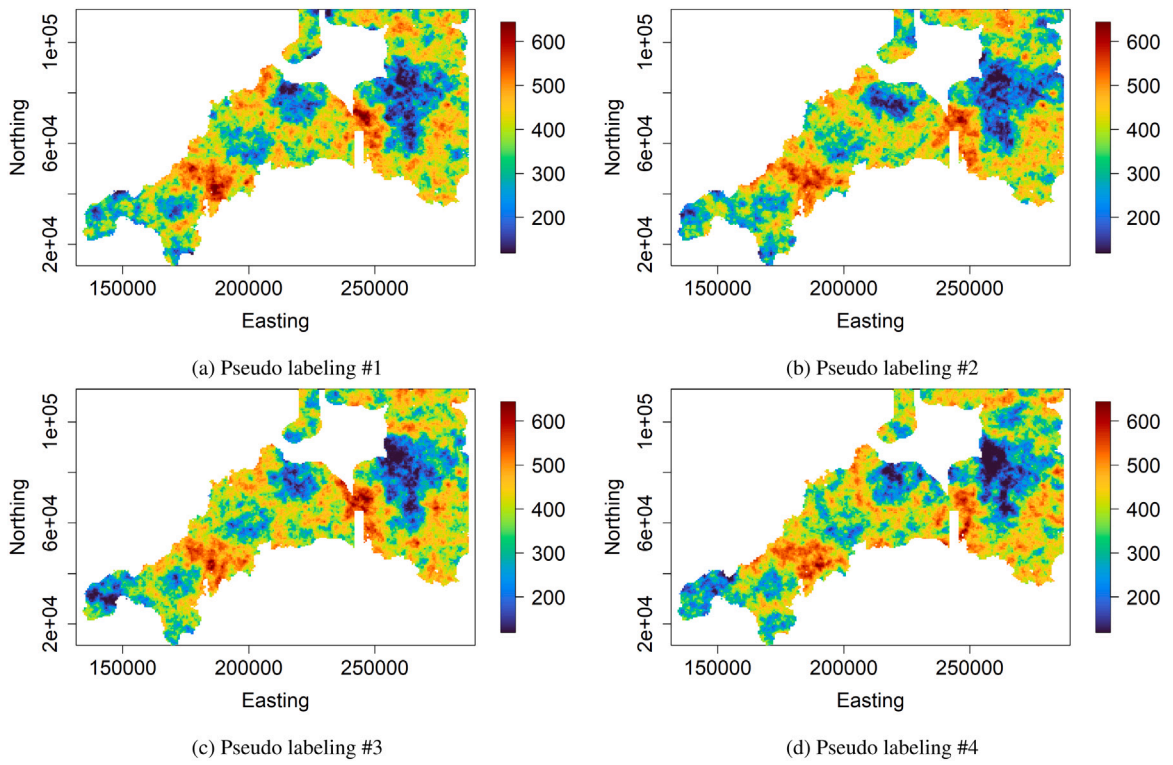


Fig. 15. Real case study — examples of pseudo labeling through geostatistical conditional simulation at selected unlabeled data locations.

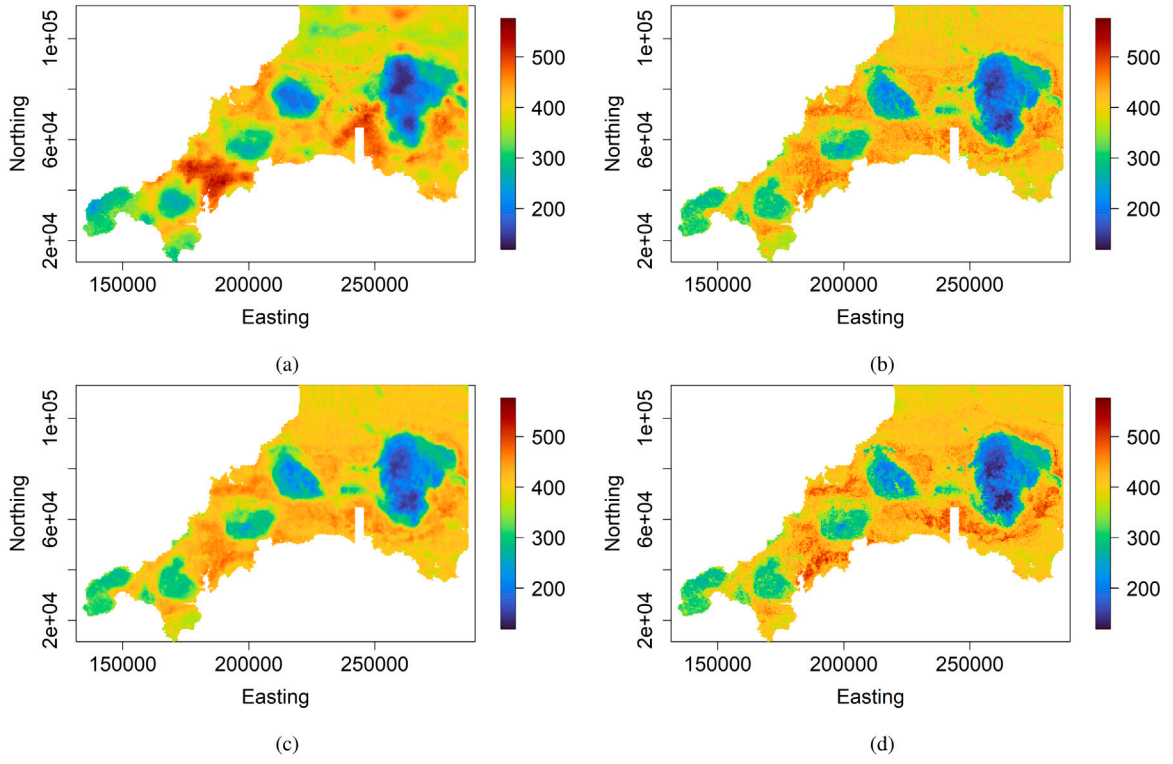


Fig. 16. Real case study — prediction maps provided by (a) geostatistical semi-supervised random forest, (b) classical random forest, (d) random forest with unlabeled spatial data treated as additional covariates, and (e) self-training random forest.

attribute space can have dissimilar labels as they may be geographically distant. In terms of prediction uncertainty accuracy, the accuracy plot in Fig. 9 and the G-statistic in Table 2 show that the geostatistical semi-supervised regression provides a better uncertainty estimates than the

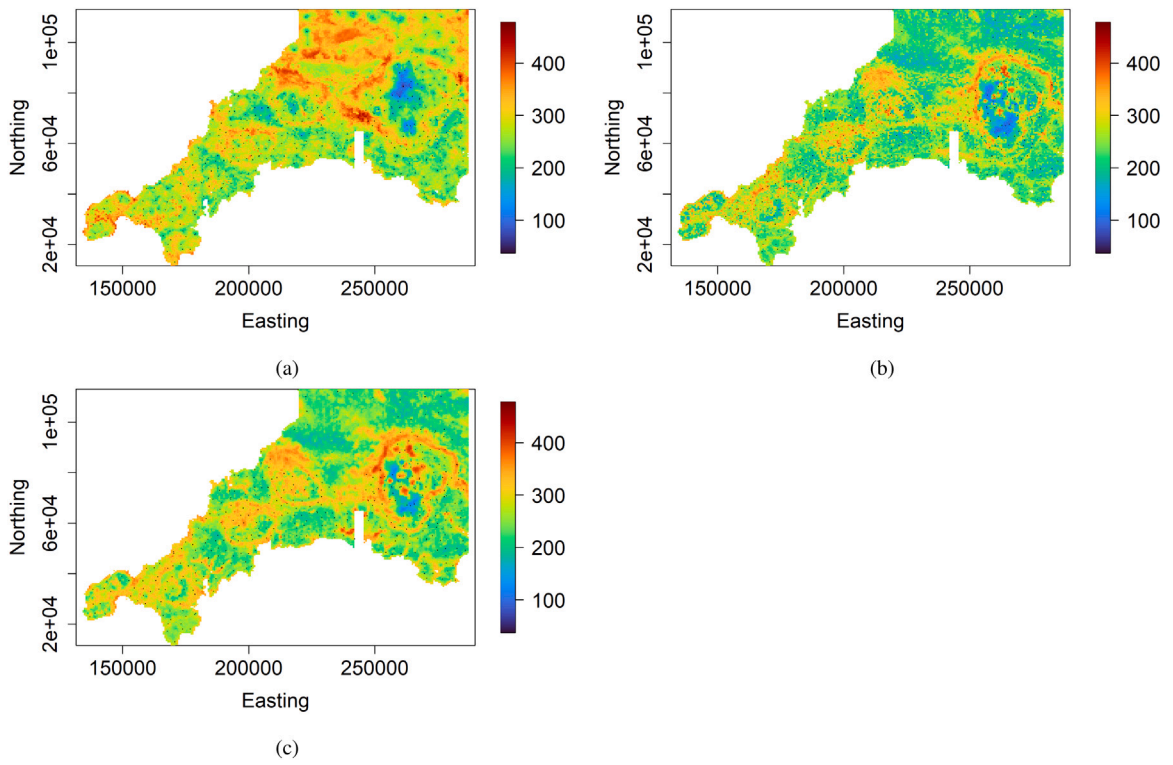


Fig. 17. Real case study — prediction uncertainty maps provided by (a) geostatistical semi-supervised random forest, (b) classical random forest, and (c) random forest with unlabeled spatial data treated as additional covariates. The black dots represent training data locations. The prediction uncertainty corresponds to the width of the 95% prediction interval.

Table 2
Synthetic case study — predictive performance statistics for unlabeled data locations ($N - n = 99\,000$). Method 1 refers to the geostatistical semi-supervised random forest; Method 2 refers to the classical random forest; Method 3 refers to the random forest with unlabeled spatial data as additional covariates; Method 4 refers to the self-training random forest.

Criteria	Method 1	Method 2	Method 3	Method 4
MAE	6.169	9.696	9.542	10.163
RMSE	8.121	12.256	12.008	12.796
R-SQUARE	0.823	0.597	0.613	0.5613
CCC	0.897	0.735	0.739	0.700
G-statistic	0.995	0.984	0.969	–

other methods. Thus, the spatial prediction performance of the geostatistical semi-supervised regression suggests that the spatial prediction task should be considered in a semi-supervised learning process instead of a conventional supervised learning process to exploit a larger set of unlabeled spatial data.

4. Real case study

The geostatistical semi-supervised learning with random forest is applied to real-world spatial data for geochemical mapping. The target variable is Barium (Ba) concentration (in mg/kg) measured at 568 sampling locations across the region of interest in southwest England (Kirkwood et al., 2016b). A detailed description of the study area and data can be found in Kirkwood et al. (2016a). The auxiliary variables selected include elevation, Landsat 8 band 5, Landsat 8 band 6, gravity survey Bouguer anomaly, gravity survey high-pass filtered Bouguer anomaly, and Uranium counts from gamma-ray spectrometry. The spatially exhaustive predictor variables are displayed over a grid containing 689,065 unsampled locations (Fig. 10). Fig. 11 shows predictor variables at sampling locations. The target variable at sampling locations (labeled spatial data) are shown in Fig. 12. Labeled spatial data represent less than 0.1% of the available data (unlabeled and labeled data). The target variable's observations are split into a training set (~ 75%) and a testing set (~ 25%). This latter set is used to evaluate the different machine learning methods: geostatistical semi-supervised random forest, classical random forest, random forest with unlabeled spatial data defined as additional predictor variables, and self-training random forest. The ten closest predictor variables' observations at a labeled data location are defined as additional covariates in the random forest with unlabeled spatial data treated as additional predictor variables. So, there are $10 \times 6 = 60$ additional covariates. All of the machine learning methods are based on random forest with the same number of decision trees set to 1,000. The other hyper-parameters have been selected via cross-validation.

Empirical and adjusted variograms of the target variable in the training dataset are given in Fig. 13. The target variable variogram is estimated from the isotropic stationary exponential family. The practical range and sill respectively correspond to $22,551.842$ and $12,983.605$. The selected unlabeled data locations for pseudo labeling are shown in Fig. 14, according the selection method described in Section 2.1. They represent approximately 88% of the unlabeled data locations. Note that unlabeled data locations that are more than a quarter of the variogram range away from the labeled data locations are ignored. Fig. 15 shows examples of pseudo labeling of the target variable at the selected unlabeled data locations. As described in Section 2.1, pseudo labeling is done through geostatistical conditional simulation, allowing for generating not just a single pseudo label at a selected data location but an ensemble of pseudo labels ($L = 50$). Thus, the uncertainty in the pseudo labeling process is accounted for. The ensemble of pseudo labels is combined with the original training dataset to form an ensemble of pseudo training datasets in which the size of each pseudo training dataset is much larger than the original training dataset. Then, a random forest model is trained for each pseudo training dataset, followed by an aggregation of the models through averaging.

Prediction maps associated with the different machine learning methods are shown in Fig. 16. The classical random forest, the random

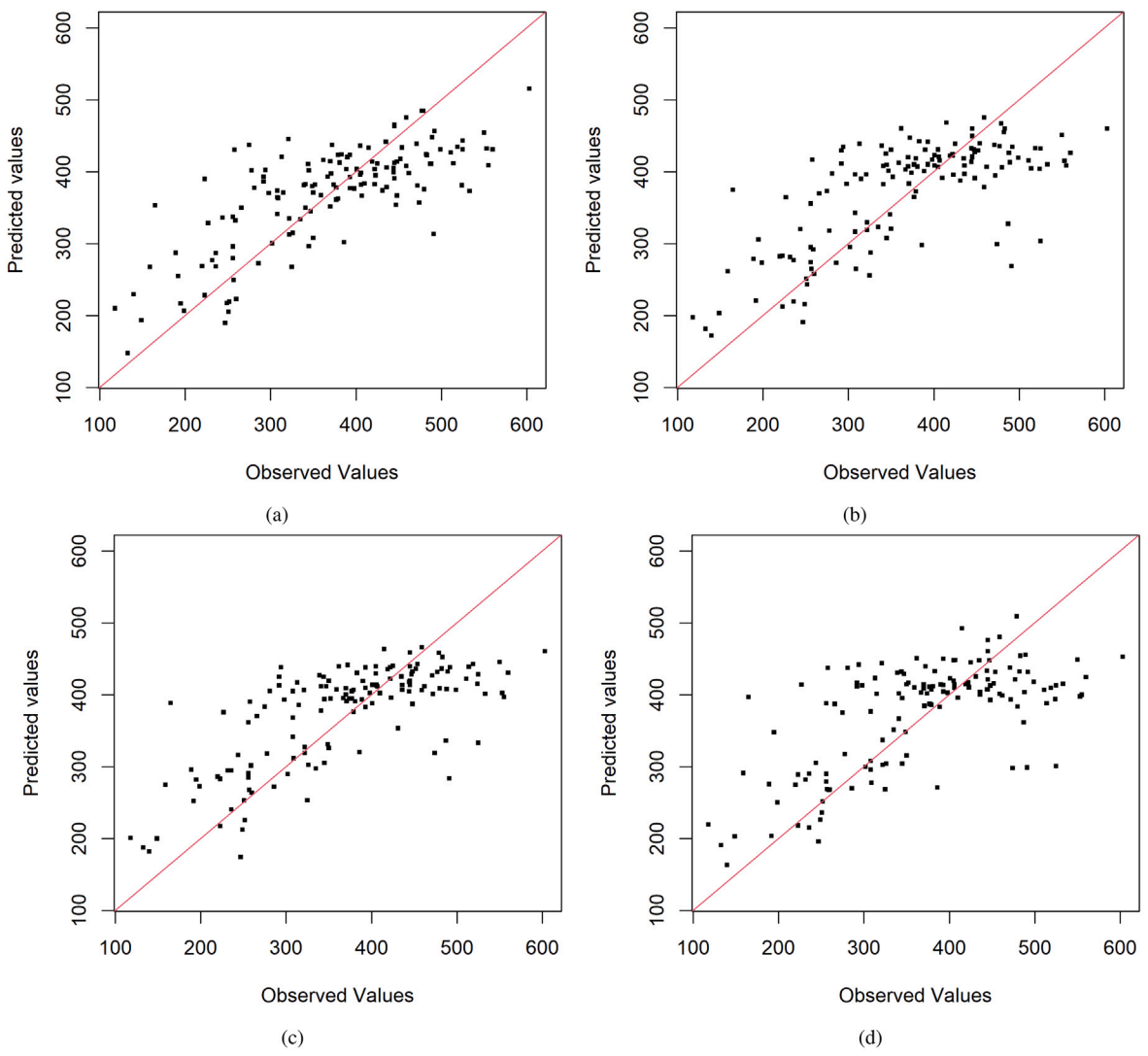
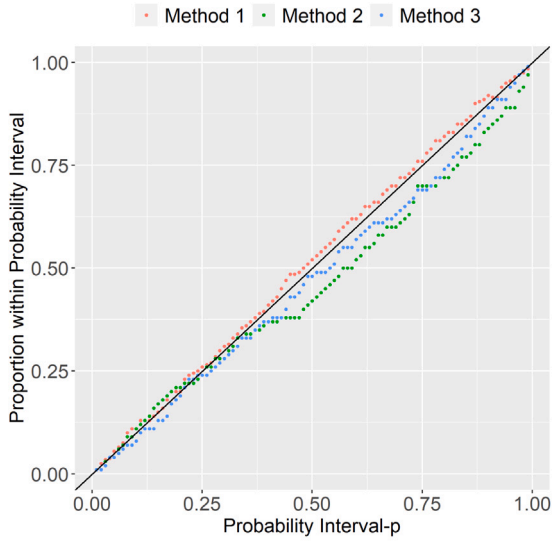


Fig. 18. Real case study — target variable's observed values vs. target variable's predicted values at testing data locations, for (a) geostatistical semi-supervised random forest, (b) classical random forest, (c) random forest with unlabeled spatial data treated as additional covariates, and (d) self-training random forest. The red line indicates the 1:1 line.

forest with unlabeled spatial data defined as additional covariates, and the self-training random forest provide similar prediction maps. Whereas there is a marked difference between the prediction maps provided by the geostatistical semi-supervised random forest and the other techniques. Prediction uncertainty maps of the different machine learning techniques except for the self-training random forest are given in Fig. 17. The prediction uncertainty is derived as the width of the 95% prediction interval. As mentioned earlier, available R packages for self-training random forest do not return individual predictions for each tree which is necessary to calculate the prediction uncertainty. The general appearance of the prediction uncertainty map associated with the proposed method differs from the others. As pointed out in the synthetic case study, the proposed approach's prediction uncertainty tends to be lower at the original training data locations than at other locations. In particular, one can note that the highest prediction uncertainties are concentrated in those areas not surveyed, which is more realistic than the prediction uncertainty map provided by the other methods. These latter provide relatively low prediction uncertainties in not surveyed areas.

To evaluate the predictive ability of the different machine learning methods, their predictions are compared with the target variable's observed values at 142 testing locations (Fig. 18). Predictive performance measures described in Section 2.3 are calculated and summarized in Table 3. One can note that the geostatistical semi-supervised random forest performs the best while the self-training random forest provides the worst prediction performance. The random forest with unlabeled spatial data treated as additional predictor variables and the classical random forest have similar prediction performances. As highlighted in the synthetic case study in Section 3, the poor prediction performance of the conventional semi-supervised learning technique (self-training random forest) can be explained by the specificities of spatial data and the underlying assumptions related to classical semi-supervised learning. Classical semi-supervised learning methods, such as self-training, are based on the smoothness assumption that data points that are close by in the input space should have similar labels. This assumption, though works for non-spatial data, is not valid in the realm of spatial data. Two data points that are close in the feature space can have dissimilar labels as they may be far away in the geographical space. Concerning the prediction uncertainty accuracy, the accuracy plot in Fig. 19 and the G-statistic in Table 3 show that the geostatistical semi-supervised regression provides a more reliable uncertainty estimates than the other methods. In particular, the uncertainty tend to be underestimated under the other methods.



(a)

Fig. 19. Real case study — prediction interval coverage probability plot (accuracy plot). Method 1 refers to the geostatistical semi-supervised random forest; Method 2 refers to the classical random forest; Method 3 refers to the random forest with unlabeled spatial data as additional covariates.

Table 3
Real case study — predictive performance statistics in the testing dataset containing 142 observations. Method 1 refers to the geostatistical semi-supervised random forest; Method 2 refers to the classical random forest; Method 3 refers to the random forest with unlabeled spatial data treated as additional covariates; Method 4 refers to the self-training random forest.

Criteria	Method 1	Method 2	Method 3	Method 4
MAE	54.124	56.393	56.168	59.138
RMSE	70.473	73.852	72.797	78.333
R-SQUARE	0.552	0.492	0.506	0.428
CCC	0.692	0.661	0.662	0.622
G-statistic	0.955	0.908	0.943	–

5. Concluding remarks

This article proposed a geostatistical semi-supervised learning method that leverages a large amount of unlabeled spatial data in combination with typically a smaller set of labeled spatial data to improve the target variable's spatial prediction. This is achieved by combining aspects of geostatistics and machine learning such as spatial autocorrelation, geostatistical conditional simulation, and supervised machine learning. The central idea consists of taking advantage of the target variable's spatial autocorrelation to generate pseudo labels at unlabeled data locations that are geographically close to labeled data locations. This mechanism allows for the selection of unlabeled data locations that are useful for the learning task. Different synthetic and real-world spatial datasets used in this paper showed the ability of the proposed spatial semi-supervised learning method to outperform classical supervised and semi-supervised machine learning methods. The proposed spatial semi-supervised learning technique allows better use of unlabeled spatial data than classical semi-supervised learning techniques. It also provides more realistic spatial prediction uncertainty than classical supervised machine learning methods.

The proposed geostatistical semi-supervised learning approach has some attractive features. It is easy to implement as it combines existing techniques in geostatistics and machine learning. Methods of geostatistics are used for pseudo labeling (label generation) and data augmentation, while machine learning methods focus on the supervised learning task. It can be used with any given supervised base learner, allowing unlabeled spatial data to be introduced in a straightforward manner. In particular, it allows a better training of data-hungry supervised base learners such as neural networks as the size of pseudo training datasets is much larger than the original training dataset. It can be easily adapted to semi-supervised classification (i.e., when the target variable is categorical). In this latter case, the pseudo labeling will be performed using a geostatistical conditional simulation method for categorical data (e.g., sequential indicator simulation, plurigaussian simulation) (Chiles and Delfiner, 2012). The aggregation of models can be done through the majority vote. Contrary to classical semi-supervised machine learning approaches (self-training and co-training), the proposed spatial semi-supervised learning technique accounts for uncertainty in the pseudo labeling process.

The proposed spatial semi-supervised learning technique assumes the presence of spatial autocorrelation in the target variable. Thus, one should expect the proposed semi-supervised learning method to perform best when the target variable's observations depict some spatial dependency. However, the proposed modeling method will not be suitable if the target variable shows a very weak or no spatial autocorrelation. Classical supervised and semi-supervised learning methods may be appropriate in this latter case as the target variable's observations can be considered independent. There are many tools that can be used to check spatial autocorrelation in the target variable, including the variogram. A variogram with a monotonic increasing slope indicates that the target variable is spatially dependent or autocorrelated. In the proposed spatial semi-supervised learning, the supervised base learner used was the traditional regression random forest. Still, any supervised machine learning model can also be used, including the existing ones dedicated to spatial data. In the proposed spatial semi-supervised learning framework, modeling the target variable's variogram and the geostatistical conditional simulation have been performed under the stationarity setting (Fouedjio, 2021c). This can also be carried out in the non-stationarity framework for better accounting for local characteristics of the target variable subject to having enough data (Fouedjio, 2017). Like traditional semi-supervised learning methods, the proposed one is computationally more intensive than classical supervised learning methods as it accounts for a large amount of unlabeled spatial data and the uncertainty in the pseudo labeling process. However, it has easily parallelizable components such as generating pseudo training datasets and training supervised machine learning models.

Declaration of competing interest

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

References

Albrecht, T., González-Álvarez, J., 2021. Using machine learning to map Western Australian landscapes for mineral exploration. ISPRS Int. J. Geo-Inf. 10.
Appelhans, T., Mwangomo, E., Hardy, D.R., Hemp, A., Nauss, T., 2015. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. Spat. Stat. 14, 91–113.
Asghar, S., Choi, J., Yoon, D., Byun, J., 2020. Spatial pseudo-labeling for semi-supervised facies classification. J. Pet. Sci. Eng. 195, 107834.
Ballabio, C., Panagos, P., Monatanarella, L., 2016. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma 261, 110–123.
Barzegar, R., Moghaddam, A.Asghari., Adamowski, J., Fijani, E., 2016. Comparison of machine learning models for predicting fluoride contamination in groundwater. Stoch. Environ. Res. Risk Assess. 1–14.
Chapelle, O., Scholkopf, B., Zien, A., 2010. Semi-supervised learning. In: Adaptive Computation and Machine Learning Series. MIT Press.
Chiles, J.P., Delfiner, P., 2012. Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons.
Cracknell, M.J., Reading, A.M., 2014. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. Comput. Geosci. 63, 22–33.

Cracknell, M., Reading, A., 2015. Spatial-contextual supervised classifiers explored: A challenging example of lithostratigraphy classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 1–14.

Du, P., Bai, X., Tan, K., Xue, Z., Samat, A., Xia, J., Li, E., Su, H., Liu, W., 2020. Advances of four machine learning methods for spatial data handling: a review. J. Geovisualization Spat. Anal. 4.

Fouedjio, F., 2017. Second-order non-stationary modeling approaches for univariate geostatistical data. Stoch. Environ. Res. Risk Assess. 31, 1887–1906.

Fouedjio, F., 2020. Exact conditioning of regression random forest for spatial prediction. Artif. Intell. Geosci. 1, 11–23.

Fouedjio, F., 2021a. Classification random forest with exact conditioning for spatial prediction of categorical variables. Artif. Intell. Geosci. 2, 82–93.

Fouedjio, F., 2021b. Random forest for spatial prediction of censored response variables. Artif. Intell. Geosci. 2, 115–127.

Fouedjio, F., 2021c. Stationarity. In: Daya Sagar, B., Cheng, Q., McKinley, J., Agterberg, F. (Eds.), Encyclopedia of Mathematical Geosciences. In: Encyclopedia of Earth Sciences Series, pp. 1–5.

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

Garcia-Ceja, E., 2019. ssr: semi-supervised regression methods. https://cran.r-project.org/web/packages/ssr/index.html. r package.

Giaccone, E., Oriani, F., Tonini, M., Lambiel, C., Mariéthoz, G., 2021. Using data-driven algorithms for semi-automated geomorphological mapping. Stoch. Environ. Res. Risk Assess. 1–17.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Jesus, J.Mendes.de., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLoS One 10, 1–26.

Hengl, T., Nussbaum, M., Wright, M., Heuvelink, G., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518.

Kanevski, M., 2008. Advanced Mapping of Environmental Data: Geostatistics, Machine Learning and Bayesian Maximum Entropy. John Wiley & Sons.

Kanevski, M., Pozdnoukhov, A., Timonin, V., 2009. Machine Learning for Spatial Environmental Data: Theory, Applications, and Software. EPFL Press.

Keogh, E., Mueen, A., 2017. Curse of dimensionality. In: Sammut, C., Webb, G. (Eds.), Encyclopedia of Machine Learning and Data Mining. pp. 314–315.

Khan, S.Z., Suman, S., Pavani, M., Das, S.K., 2016. Prediction of the residual strength of clay using functional networks. Geosci. Front. 7, 67–74.

Kirkwood, C., Cave, M., Beamish, D., Grebby, S., Ferreira, A., 2016a. A machine learning approach to geochemical mapping. J. Geochem. Explor. 167, 49–61.

Kirkwood, C., Economou, T., Pugeault, N., Odbert, H., 2022. Bayesian deep learning for spatial interpolation in the presence of auxiliary information. Math. Geosci. 54, 507–531.

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.

Kobs, K., Schäfer, M., Krause, A., Baumhauer, R., Paeth, H., Hotho, A., 2021. Semi-supervised learning for grain size distribution interpolation. In: Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R. (Eds.), Pattern Recognition. ICPR International Workshops and Challenges. Springer International Publishing, Cham, pp. 34–44.

Kumar, C., Chatterjee, S., Oommen, T., Guha, A., 2020. Automated lithological mapping by integrating spectral enhancement techniques and machine learning algorithms using AVIRIS-NG hyperspectral data in gold-bearing granite-greenstone rocks in Hutti, India. Int. J. Appl. Earth Obs. Geoinf. 86, 102006.

Lantuejoul, C., 2013. Geostatistical Simulation: Models and Algorithms. Springer Berlin Heidelberg.

Latifovic, R., Pouliot, D., Campbell, J., 2018. Assessment of convolution neural networks for surficial geology mapping in the South Rae geological region, Northwest territories, Canada. Remote Sens. 10.

Li, J., 2013. Predictive modelling using random forest and its hybrid methods with geostatistical techniques in marine environmental geosciences. In: 11-Th Australasian Data Mining Conference. AusDMÍ3, Canberra, Australia, pp. 73–79.

Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Model. Softw. 26, 1647–1659.

Maxwell, A.E., Warner, T.A., Fang, F., 2018. Implementation of machine-learning classification in remote sensing: an applied review. Int. J. Remote Sens. 39, 2784–2817.

Othman, A.A., Gloaguen, R., 2017. Integration of spectral, spatial and morphometric data into lithological mapping: A comparison of different machine learning algorithms in the Kurdistan region, NE Iraq. J. Asian Earth Sci. 146, 90–102.

Pise, N.N., Kulkarni, P., 2008. A survey of semi-supervised learning methods. In: 2008 International Conference on Computational Intelligence and Security. pp. 30–34. http://dx.doi.org/10.1109/CIS.2008.204.

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

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.r-project.org/.

Renard, D., Bez, N., Desassis, N., Beucher, H., Ors, F., Freulon, X., 2021. RGeostats: Geostatistical package. http://rgeostats.free.fr/. r package version 12.1.0.

Sahoo, S., Jha, M.K., 2017. Pattern recognition in lithology classification: Modeling using neural networks, self-organizing maps and genetic algorithms. Hydrogeol. J. 25, 311–330.

Sekulić, A., Kilibarda, M., Heuvelink, G., Nikolić, M., Bajat, B., 2020. Random forest spatial interpolation. Remote Sens. 12 (1687).

Steichen, T.J., Cox, N.J., 2002. A note on the concordance correlation coefficient. Stata J. 2, 183–189.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Kerry, R., 2016. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma 266, 98–110.

Talebi, H., Peeters, L.J., Otto, A., Tolosana-Delgado, R., 2021. A truly spatial random forests algorithm for geoscience data analysis and modelling. Math. Geosci. 1–22.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46, 234–240.

Van Engelen, J.E., Hoos, H.H., 2020. A survey on semi-supervised learning. Mach. Learn. 109, 373–440.

Vatsavai, R.R., Shekhar, S., Burk, T.E., 2007. An efficient spatial semi-supervised learning algorithm. Int. J. Parallel Emergent Distrib. Syst. 22, 427–437.

Wadoux, A.M.C., Minasny, B., McBratney, A.B., 2020. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth Sci. Rev. 210, 103359.

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

Wright, M.N., Ziegler, A., 2017. ranger: A fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77, 1–17.

Yu, L., Porwal, A., Holden, E.J., Dentith, M.C., 2012. Towards automatic lithological classification from remote sensing data using support vector machines. Comput. Geosci. 45, 229–239.

Zhu, X., Goldberg, A., 2009. Introduction to semi-supervised learning. In: Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool.