Math Geosci (2022) 54:1–22
https://doi.org/10.1007/s11004-021-09946-w

A Truly Spatial Random Forests Algorithm for Geoscience Data Analysis and Modelling

Hassan Talebi[1,2,4] · Luk J. M. Peeters[1,3] · Alex Otto[2] · Raimon Tolosana-Delgado[5]

Received: 14 September 2020 / Accepted: 20 April 2021 / Published online: 14 July 2021

Abstract Spatial data mining helps to find hidden but potentially informative patterns from large and high-dimensional geoscience data. Non-spatial learners generally look at the observations based on their relationships in the feature space, which means that they cannot consider spatial relationships between regionalised variables. This study introduces a novel spatial random forests technique based on higher-order spatial statistics for analysis and modelling of spatial data. Unlike the classical random forests algorithm that uses pixelwise spectral information as predictors, the proposed spatial random forests algorithm uses the local spatial-spectral information (i.e., vectorised spatial patterns) to learn intrinsic heterogeneity, spatial dependencies, and complex spatial patterns. Algorithms for supervised (i.e., regression and classification) and unsupervised (i.e., dimension reduction and clustering) learning are presented. Approaches to deal with big data, multi-resolution data, and missing values are discussed. The superior performance and usefulness of the proposed algorithm over the classical random forests method are illustrated via synthetic and real cases, where the remotely sensed geophysical covariates in North West Minerals Province of Queensland, Australia, are used as input spatial data for geology mapping, geochemical prediction, and process discovery analysis.

✉ Hassan Talebi
Hassan.Talebi@csiro.au

[1] Deep Earth Imaging FSP, CSIRO, 26 Dick Perry Avenue, Kensington, WA 6151, Australia

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

[3] Land and Water, CSIRO, Locked Bag 2, Urrbrae, SA 5062, Australia

[4] School of Science, Edith Cowan University, Joondalup, WA 6027, Australia

[5] Helmholtz Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resources Technology, Chemnitzerstrasse 40, 09599 Freiberg, Saxony, Germany

Keywords Random forests · Higher-order spatial statistics · Spatial correlation · Spatial data · Geostatistical learning

1 Introduction

Spatial data mining reveals hidden and previously unknown but potentially informative patterns from big and high-dimensional geoscience data. It takes advantage of the ever-growing availability of geographically referenced data and their potential abundance (Sellars 2018). Many geomatic applications benefit from spatial data mining in several stages, including data collection, data storage, exploratory data analysis, data processing, prediction, and uncertainty quantification. For instance, spatial data mining techniques can be used for splitting the study area to account for different behaviours of natural phenomena, which is useful for discovering earth processes and simplifying subsequent modelling steps (Rolnick et al. 2019). However, understanding the particularities of geosystems and geoscience data is critical for obtaining accurate and physically consistent inferences and predictions via data mining approaches (Reichstein et al. 2019; Talebi et al. 2020).

Geoscience processes vary significantly through time and space. Such heterogeneity and non-stationarity are related to the spatial and/or temporal variation of soil types, rock types, land uses, vegetation types, climatic conditions, and tectonic activities. Geographical observations that are located close to each other in space and time tend to share similar characteristics. This phenomenon is known as auto-correlation and provides additional information to inform statistical models (Matheron 1962; Cliff and Ord 1973). Remotely sensed data are examples of earth observations that show spatial and/or temporal auto- and cross-correlations. Generally, pixels that are located close to each other in satellite images are more likely to have similar values compared to those that are spaced further apart (Woodcock et al. 1988). Traditional data mining algorithms treat observations as independent values (i.e., independent and identically distributed data) and exclude useful information from the analysis by disregarding space and time dependencies. Consequently, predictions and inferences from non-spatial learners can be misleading when applied to geoscience data (Reichstein et al. 2019; Bergen et al. 2019; Karpatne et al. 2019).

Spatial data mining techniques should be able to capture multivariate spatial and/or temporal patterns of different scales and types. Either current machine learning (ML) algorithms can be amended to be consistent with the nature of geoscience data or new algorithms need to be developed (Karpatne et al. 2019; Talebi et al. 2020). Among the statistical learners, tree-based techniques are very useful for geoscience data analysis and modelling due to their transparency, simplicity of implementation, ability to capture non-linear relationships, and ability to handle big and high-dimensional data, mixed data, and missing values (Kuhn and Johnson 2013). The tree-based learners can be improved further by considering the heterogeneity and spatial dependency of geoscience data.

Hengl et al. (2018) showed that adding covariates such as Euclidean buffer distances to observations to a random forests (RF) model produces estimates as accurate and unbiased as those from the common geostatistical method, kriging. Liu et al. (2018) used a combination of RF and regression kriging to measure particulate matter concentration derived from remotely sensed and ground observations. In their proposed methodology, a non-linear trend between the dependent variable and covariates is predicted by RF. Subsequently, kriging is used to estimate residuals of the predicted trend (Hengl et al. 2015). Meyer et al. (2019) argued that the nonspatial validation procedures based on the random sampling of a dataset (e.g., k-fold cross-validation and bootstrapping) are inappropriate for learning from spatial data. Nonspatial sampling results in over-optimistic predictive models (caused by the spatial correlations) that can predict the input training data accurately but have marginal performance in terms of extrapolation (i.e., predicting patterns that have not been seen during the training process). Moreover, using geolocation variables such as latitude and longitude may generate several artefacts in the final predicted maps. Talebi et al. (2019) introduced a new method for spatial uncertainty quantification of geoscience data based on a combination of geostatistical simulation and random forests predictive models. Approaches to deal with compositional data were also discussed. Georganos et al. (2019) introduced a locally varying RF predictive model and concluded that the performance of the technique can be improved when an appropriate spatial scale is selected. Finally, Mitchell and Sheppard (2019), inspired by convolutional neural networks, demonstrated how the performance of the RF algorithm for recognising spatial patterns can be improved by introducing a local spatial bias during the learning process.

The previously mentioned possibilities to improve the spatial awareness of the tree-based learners are not suitable for recognising complex spatial patterns, objects, and structures of different scales and their spatial distributions across the domain of study. Most of these techniques only use the information available at single points or rely on two-point geostatistics (e.g., variogram models and kriging techniques) to capture spatial dependency. Considering only two points or pixels is not sufficient for capturing complex spatial patterns. Multiple point geostatistics were developed to address this limitation by considering more than two points simultaneously (Mariethoz and Caers 2015). In addition, approaches for accurately handling multi-resolution spatial data and missing values must be improved further.

The objective of this study is to develop a spatial random forests (SRF) technique based on nonparametric higher-order spatial statistics for spatial data analysis and modelling. The proposed model can be applied to the high-dimensional and nonlinear phenomenon with a small number of observations and multi-resolution predictors of mixed type (e.g., continuous and categorical). In the proposed technique, the dimension of input data is increased to capture local information and to learn intrinsic heterogeneity, spatial dependencies, and complex spatial patterns. Compared with the previous attempts to improve the spatial awareness of tree-based learners, the proposed technique can be considered as a truly spatial predictive model. Approaches to handle big data, multi-resolution data, and missing values are discussed. Algorithms for supervised (i.e., regression and classification) and unsupervised (i.e., dimension reduction and clustering) learning are presented.

The following sections discuss the basics of the proposed methodology and provide the implementation details of the SRF model. To demonstrate the applicability of the proposed technique, extensive use of one synthetic and one real dataset is made. The objective of the synthetic case is to identify complex spatial patterns using unsupervised SRF, while the focus of the real case is to reproduce the interpreted geological map and to predict the geochemistry of the formations in the North West Minerals Province (NWMP) of Queensland, Australia. The paper is concluded with a short review of the pros and cons of the proposed technique and recommendations for future research.

2 Methodology

The methodology is based on the classical binary recursive partitioning tree (Breiman et al. 1984) and the classical random forests algorithm (Breiman 2001). In the following subsections, the classical decision tree and random forests algorithm are extended to account for the complex spatial dependencies of regionalised variables. The local spatial-spectral information is captured by extracting and vectorising multivariate spatial patterns. Such local information is used to learn intrinsic heterogeneity, spatial dependencies, and complex spatial patterns during the training process of regressors and classifiers. Approaches to account for multi-resolution data and missing values are discussed. Finally, a novel approach to generate synthetic multivariate spatial patterns is proposed for unsupervised modelling of spatial data.

2.1 vectorised Spatial Patterns

Supervised and unsupervised learning models are trained using a set of N observations at locations u_i $i = \{1, ... N\}$ (Fig. 1(a)). These observations can be multivariate (e.g., of dimension R) and mixed (e.g., continuous and categorical, Fig. 1(b) to (d)). It is assumed that the observations are located on regular grids. Non-gridded observations should first be migrated to the closest nodes of a regular grid of suitable resolution or be rasterised using geostatistical techniques. For each location u_i and regionalised variable $x^r(u)$, $r = \{1, \ldots, R\}$, a local spatial pattern is extracted and stored as a vector $pat_r\left(u_i\right) = \left[x^r\left(u_{i1}\right), \ldots, x^r\left(u_{iE_r}\right)\right]$ which consists of the values of the rth variable at E_r nodes around the location u_i. The parameter E_r controls the order of spatial statistics (i.e., number of nodes in the extracted pattern) for each regionalised variable. The univariate spatial patterns can have different shapes and scales. In this illustration, the spatial patterns have square shapes with different sizes and numbers of nodes (Fig. 1(e)). To store square patterns as $pat_r(u_i)$ vectors, they are stored pixel by pixel from left to right and row by row downward. The overall multivariate spatial pattern (Fig. 1(e)) for each location u_i is stored as a long vector $pat(u_i) = \left[pat_1\left(u_i\right), \ldots, pat_R(u_i)\right]$. The total number of predictors is $P = \sum_{r=1}^{R} E_r$. To make the technique invariant in terms of rotation, scaling, and distortion, the transformed patterns should be learned during the training process. For instance, the rotated multivariate patterns, $rot\left(pat\left(u_i\right)\right)$, can be added to the information about location u_i to be used later during the training process. In the proposed method, only a simplified version of rotation invariance is used by rotating multivariate patterns through angles of 90 degrees, which quadruples the number of observations N

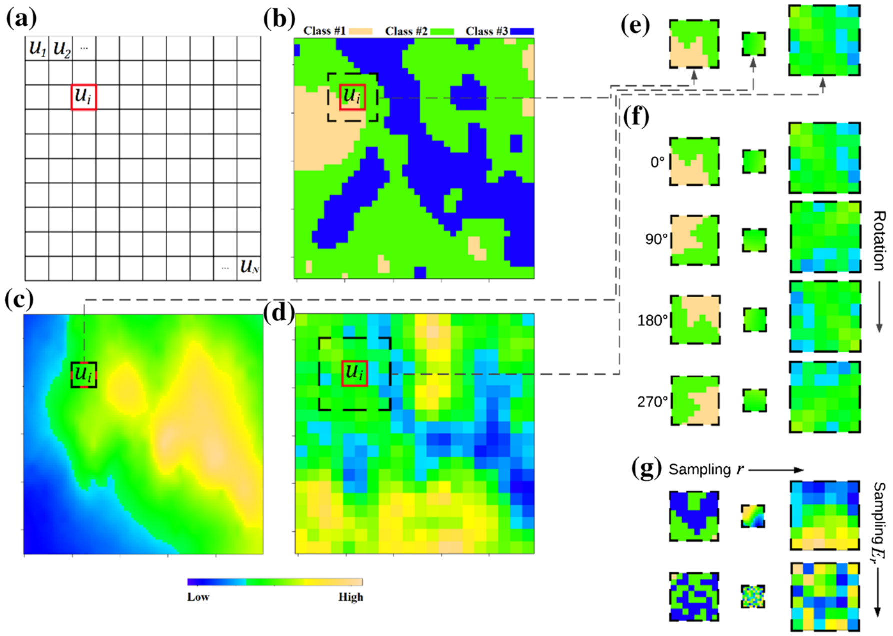

Fig. 1 Preprocessing the input spatial data, **a** prediction grid, **b** categorical variable, **c** high-resolution continuous variable, **d** low-resolution continuous variable, **e** extracted multivariate pattern for location u_i, **e** rotated patterns for invariant training, and **g** the proposed process for generating synthetic patterns for unsupervised learning

(Fig. 1(f)). To train a supervised spatial model (i.e., classification and regression), the response values $y(u_1), \ldots, y(u_N)$ must be stored with the predictors. The final training dataset $\mathcal{D} = \{(pat(u_1), y(u_1)), \ldots, (pat(u_N), y(u_N))\}$ will be used to train the SRF model.

2.2 Spatial Decision Trees

A spatial tree partitions the predictor space via a sequence of binary splits on individual predictors $x^p(u), p = \{1, \ldots, P\}$. The root node of the tree comprises the entire vectorised multivariate patterns $pat(u_i), i = \{1, \ldots, N\}$. The final partitions of the predictor space are associated with the terminal nodes (i.e., the nodes that are not split). The internal nodes are split into two descendant nodes (i.e., right and left nodes) according to the value of one of the predictors. If a predictor is continuous, a split is determined by a cut-off value; multivariate patterns with the selected predictor smaller than the cut-off go to the left, the remainder go to the right (Fig. 2). Similarly, for a categorical predictor with finite levels S_i, a binary split is obtained by defining a subset of levels $S \subset S_i$.

A splitting criterion must be defined to find the best predictor and the best split. For a continuous response variable with values $y(u_1), \ldots, y(u_n)$ at the node to be split, the mean of the squared residuals is typically used as the splitting criterion

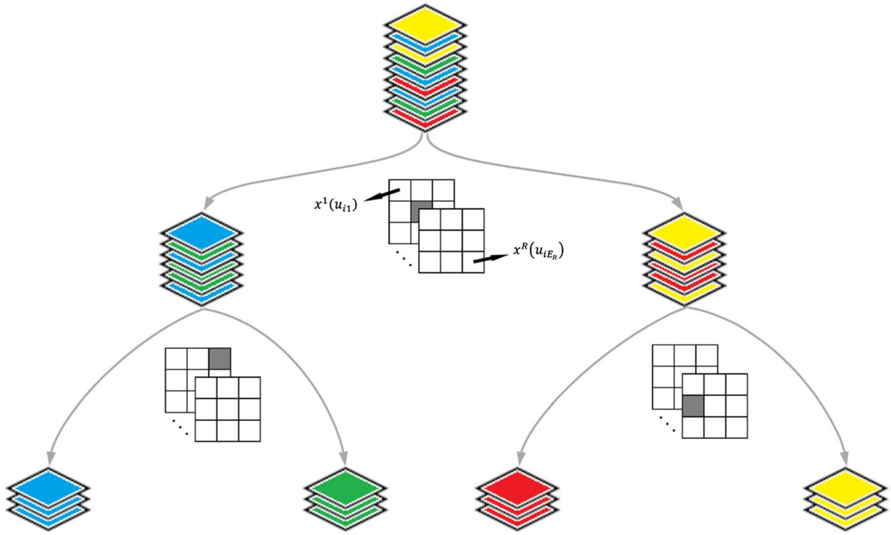

Fig. 2 Spatial decision tree for learning multivariate spatial patterns

$$\mathcal{Q} = \frac{1}{n}\sum_{i=1}^{n}\left(y\left(u_i\right) - \overline{y}\right)^2, \tag{1}$$

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y\left(u_i\right)$ is the local mean of the response variable and n is the number of patterns at the node to be split. For a categorical response variable with K levels, a typical splitting criterion is the Gini impurity index

$$\mathcal{Q} = 1 - \sum_{k=1}^{K}\widehat{p}_k^2, \tag{2}$$

where $\widehat{p}_k = \frac{1}{n}\sum_{i=1}^{n} I\left(y\left(u_i\right) = k\right)$ is the proportion of class k in the node to be split, and $I\left(y\left(u_i\right) = k\right) = 1$ if $y\left(u_i\right) = k$ and 0 otherwise. The splitting criteria for left and right descendent nodes, $\mathcal{Q}_L$ and $\mathcal{Q}_R$, and the associated sample sizes, n_L and n_R, are calculated for each split candidate. The best split is that which minimises $\mathcal{Q}_{split} = n_L\mathcal{Q}_L + n_R\mathcal{Q}_R$. The splitting process continues until a stopping criterion is met. For instance, the recursive splitting stops when the sample size of the node is less than a predefined threshold. Predicted values of the response variable at the terminal nodes are given by $\widehat{h}(x(u)) = \frac{1}{n}\sum_{i=1}^{n} y\left(u_i\right)$ for regression and $\widehat{h}(x(u)) = argmax_{y(u)}\sum_{i=1}^{n} I\left(y\left(u_i\right) = y(u)\right)$ for classification. The predictors $x^p(u)$, $p = \{1, \ldots, P\}$, of a new multivariate pattern $pat(u)$ are used to determine a terminal node, and $\widehat{h}(x(u))$ is used as the prediction.

2.3 Spatial Random Forests

The SRF model uses the proposed spatial trees as the base learners. Using the input data $\mathcal{D} = \left\{ \left(pat\left(u_1\right), y\left(u_1\right) \right), \ldots, \left(pat\left(u_N\right), y\left(u_N\right) \right) \right\}$, the base learners $\widehat{h}_j\left(x(u), \theta_j, \mathcal{D}_j\right)$, $j = 1, \ldots, J$, are trained to learn the input multivariate spatial patterns. The random component θ_j is used to inject randomness to the base learners by fitting each spatial tree to an independent bootstrap sample (i.e., $\mathcal{D}_j$) of $\mathcal{D}$ and finding the best split over m randomly selected predictors. Sampling the predictors for node splitting is particularly relevant in the case of high-dimensional data, where data consist of a very large number of predictors compared to the small number of observations. The SRF prediction for a new multivariate pattern $pat(u)$ is given by $\widehat{f}(x(u)) = \frac{1}{J} \sum_{j=1}^{J} \widehat{h}_j(x(u))$ for regression and $\widehat{f}(x(u)) = argmax_{y(u)} \sum_{j=1}^{J} I\left(\widehat{h}_j(x(u)) = y(u)\right)$ for classification. In

the case of classification, instead of the abstract class prediction based on majority voting for a $pat(u)$, one can look at the fractions of the total number of trees $p_k(u)$ which vote for each class k, $k = \{1, \ldots, K\}$. These predicted fractions can be used to calculate the local entropy (Shannon 1948; Goovaerts 1997) as a measure of local uncertainty

$$H(u) = -\frac{1}{ln(K)} \sum_{k=1}^{K} p_k(u).\left[ln\left(p_k(u)\right) \right], \tag{3}$$

where $H(u) \in [0, 1]$. Locations where the entropy is close to 1 have high uncertainty, and the entropy is greatest when all fractions $p_k(u)$ are equal.

For each bootstrap sample $\mathcal{D}_j$ and tree prediction$\widehat{h}_j(x(u))$, some patterns $pat\left(u_i\right)$ do not make it into the sample. These remaining patterns are called out-of-bag (OOB) and can be used to estimate the generalisation error of the SRF model and to measure predictor importance. For each pattern$pat\left(u_i\right)$,$i = 1, \ldots, N$, a prediction of the response variable is obtained using the J_i spatial trees for which $pat\left(u_i\right)$ is OOB,$d_i = \left\{ j : \left(pat\left(u_i\right), y\left(u_i\right) \right) \notin \mathcal{D}_j \right\}$. The OOB predictions are given by $\widehat{f}_{OOB}(x(u)) = \frac{1}{J_i} \sum_{j \in d_i} \widehat{h}_j(x(u))$ for regression and

$\widehat{f}_{OOB}(x(u)) = argmax_{y(u)} \sum_{j \in d_i} I\left(\widehat{h}_j(x(u)) = y(u)\right)$ for classification. The generalisation

error for a regression model is typically estimated via the out-of-bag mean squared error

$$MSE_{OOB} = \frac{1}{N} \sum_{i=1}^{N} \left(y\left(u_i\right) - \widehat{f}_{OOB}\left(x\left(u_i\right)\right) \right)^2. \tag{4}$$

Similarly, the generalisation error for a classification model is estimated using the OOB error rate

$$E_{OOB} = \frac{1}{N} \sum_{i=1}^{N} I\left(y\left(u_i\right) \neq \widehat{f}_{OOB}\left(x\left(u_i\right)\right) \right). \tag{5}$$

The strategies used for tuning the hyperparameters of a classical RF model (Probst et al. 2019) can also be implemented for the proposed SRF model. However, due to the underlying structure of the SRF model that increases the dimension of the input data to capture spatial patterns, automated hyperparameter optimisation is computationally demanding.

Measuring the importance of predictors is useful for variable selection and for interpreting the SRF model. For the input patterns $pat\left(u_i\right)$, $i = 1, \ldots, N$, the OOB predictions $\widehat{f}_{OOB}(x(u))$ are obtained. The values of a predictor of interest $x^p(u)$ are randomly permuted in the OOB patterns and the OOB predictions are recalculated. The difference between the error rates (classification) or mean squared errors (regression) of the predictions obtained from the original OOB patterns and those obtained using the permuted data gives a measure of importance for the predictor $x^p(u)$. The same procedure is used to measure the importance of other predictors. The predictor importance estimates the increase of error rate or mean square error of a predictive model on a test set when values of a predictor of interest are randomly permuted. The spatial approach to measure predictor importance not only ranks the predictors but also measures the zone of influence for each regionalised variable $x^r(u)$, $r = \{1, \ldots, R\}$. The zone of influence for a regionalised variable $x^r(u)$ shows the importance of local information at E_r nodes surrounding a central node u_i. The zone of influence can be used to assess the anisotropic behaviour and scale of the spatial patterns and structures. The provided information regarding the significance of the E_r surrounding nodes can be used to define the structure or geometry of data events in multiple point geostatistics and also to assess training images (Mariethoz and Caers 2015).

Missing data occur frequently and require appropriate handling. For instance, the spatial patterns at the margins of a study area may contain many missing values. In the proposed algorithm, missing data imputation is conducted temporarily at each splitting node. To assess the splitting criterion for a predictor $x^p(u)$, the existent missing values are imputed by sampling from the known in-bag values at the node to be split. The imputed values are only used to assign the patterns with missing values to the left or right descendent nodes. The imputed values are removed from the descendent nodes after splitting (Ishwaran et al. 2008).

2.4 Unsupervised Learning

Geoscience data typically hold informative statistical and spatial patterns. Consequently, such statistical and spatial patterns should be distinguishable from a randomly generated version of themselves. A synthetic set of multivariate patterns is generated by introducing two steps of randomisation. First, for each regionalised variable $x^r(u)$, $r = \{1, \ldots, R\}$, the N univariate patterns $pat_r\left(u_i\right)$ are randomly sampled (without replacement). Then, the E_r nodes of each sampled pattern $pat_r\left(u_i\right)$ are permuted (Fig. 1(g)). The random sampling of the univariate patterns removes the spatial cross-correlations while the permutation of the E_r locations removes the remaining auto-correlations in the synthetic patterns. The synthetic pattern for each location u_i is stored as a long vector $pat^*\left(u_i\right) = \left[pat_1^*\left(u_i\right), \ldots, pat_R^*(u_i)\right]$. The final dataset for unsupervised learning is constructed by concatenating the synthetic vectorised multivariate patterns below the original ones and consists of $2 \times N$ rows of vectorised multivariate patterns and P columns of predictors. A binary SRF classifier (introduced in Sect. 2.3) is trained to discriminate the original multivariate patterns from the synthetic ones. An OOB error rate lower than 50\% indicates the presence of coherent statistical and spatial patterns in the input spatial data and the capability of the SRF model to capture those patterns accurately.

The N patterns $pat\left(u_i\right)$ are modelled by each spatial tree, and if a pair of multivariate patterns, $pat\left(u_i\right)$ and $pat\left(u_j\right)$, i, j, $(1, 2, \ldots, N)$, ends in the same terminal node, their proximity is increased by 1. The proximities are averaged over J trees in the SRF classifier to generate the spatially aware $N \times N$ proximity matrix. The principle is that any two similar multivariate spatial patterns will follow the same paths in different spatial decision trees. The proximity between two patterns is a measure of how close together they are in spatial predictor space, but it automatically gives more weight to important spatial predictors that are useful for identifying the underlying statistical and spatial structures of geoscience data. The proximities between multivariate patterns are real numbers bounded between 0 and 1. By subtracting 1 from the proximities, a positive semi-definite dissimilarity matrix D is constructed that can be used for unsupervised analysis.

To understand various patterns and spatial structures in the study area, the spatially aware dissimilarity matrix D is used to perform a classical multidimensional scaling (MDS) and to obtain a simplified two- or three-dimensional plot (Mead 1992). Each point on such a plot represents one of the multivariate patterns $pat\left(u_i\right)$, and the distances between the points reproduce, as closely as possible, the proximity-based distances in D. Several subgroups of patterns or outlier patterns (e.g., gold mineralisation) are easily captured from the MDS plots. The spatially aware dissimilarity matrix D can also be used to cluster the geoscience data and define the underlying natural domains. In this study, the partitioning around medoids technique (Kaufman and Rousseeuw 1990) is used for this purpose.

Several techniques can be used to assess the quality of clustering and to select the optimum number of clusters (Kassambara 2017). In this study, the silhouette width is used to assess the performance of the proposed unsupervised SRF. For each location u_i, the silhouette width is calculated as

$$S\left(u_i\right) = \frac{b\left(u_i\right) - a\left(u_i\right)}{max\left(a\left(u_i\right), b\left(u_i\right)\right)}, \tag{6}$$

where $a\left(u_i\right) = \frac{1}{\left|C_p\right|-1} \sum_{u_j \in C_p, i \neq j} d\left(pat\left(u_i\right), pat\left(u_j\right)\right)$ is the average distance between the multivariate pattern located at u_i and all other patterns at locations u_j of the cluster C_p to which u_i belongs. The smaller the value of $a\left(u_i\right)$, the better the assignment of location u_i to a cluster C_p. The parameter $b(u_i) = \min_{q \in (1,2,\ldots,M) \text{ and } q \neq p} \left(\frac{1}{\left|C_q\right|} \sum_{u_j \in C_q} d\left(pat\left(u_i\right), pat\left(u_j\right)\right)\right)$ is the minimum of the average distances between the pattern at location u_i, and all patterns at locations

u_j belong to the other clusters in which u_i is not a member. This parameter can be seen as the distance between the location u_i and the closest cluster to which u_i does not belong. These parameters can be easily calculated from the spatially aware dissimilarity matrix D. Locations with a $S(u_i)$ close to 1 are well clustered. As $S(u_i)$ approaches −1, the location u_i is probably in the wrong cluster. Finally, $S(u_i)$ close to 0 means that the location u_i is between two clusters. The silhouette width can be averaged over the study area A for different numbers of clusters M. The optimal number of clusters is the one that maximises this average.

In the case of large datasets such as high-resolution satellite images, the dissimilarity matrix D can be enormous. Calculating, storing, and analysing such large matrices are challenging. To handle such large datasets, a representative sample of the vectorised multivariate patterns is taken, and the proposed method is applied to this subset to define the cluster numbers. Subsequently, an SRF classifier is trained using the sampled vectorised multivariate patterns as input predictors and the cluster numbers as the new categorical response variable. This multiclass classifier is later used to predict the cluster numbers across study area A.

3 Implementation and Results

3.1 SRF Unsupervised Learning

The aim of this experiment is to investigate the capability of the unsupervised SRF model for capturing complex spatial patterns. A synthetic dataset is used in this experiment which consists of two continuous variables with the same resolution (Fig. 3a, b). In this case study, the resolution of the clustering grid is the same as that of the input variables ($N = 10,000$). Unconsolidated sediments, marshes, and channels are the three spatial structures that need to be captured by unsupervised learning (Fig. 3a). The second variable is a Gaussian noise and is used to assess the performance of the proposed SRF model in the presence of noisy variables without any coherent spatial structure. The same order of spatial statistics was selected for the two input variables, $E_1 = E_2 = 81$ (Fig. 3(d) and (e)). The default value was selected for the parameter $m = ceiling\left(\sqrt{\sum_{r=1}^{2} E_r}\right) = 13$. The number of spatial decision trees was set to $J = 1{,}000$. The error rate $E_{OOB} = 0.024$ demonstrates the presence of coherent statistical and spatial structures in the input spatial data and the ability of the SRF model to discriminate the original spatial patterns from the synthetic ones. The average silhouette width suggests three spatial clusters (Fig. 3c), which is consistent with the prior knowledge. Comparing the zone of influence for the two input variables reveals the robustness of the SRF model in terms of automatically ignoring the noisy variables without any coherent spatial structures (Fig. 3d, e). The spatially aware dissimilarity matrix D was used to visualise the level of similarity between the multivariate spatial patterns in the study area via the multidimensional scaling approach. The first coordinate in Fig. 3f discriminates marshes and channels from the unconsolidated sediments, whereas the second coordinate discriminates marshes from channels. Classical RF

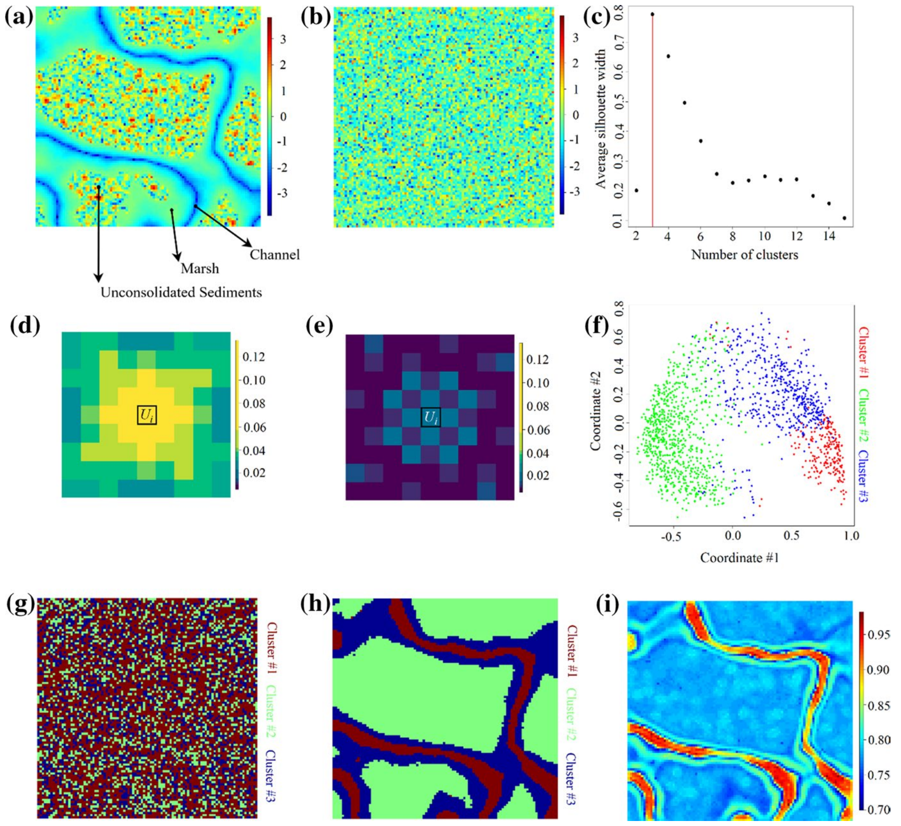

Fig. 3 **a** and **b** input variable #1 and #2 respectively, **c** optimum number of clusters based on average silhouette width, **d** and **e** spatial predictor importance for the input variable #1 and #2 respectively, **f** multidimensional scaling plot coloured by the predicted spatial clusters via the SRF model, **g** final clusters generated via classical RF model, **h** spatial clusters generated via the SRF model, and **i** the estimated silhouette width $S(u)$ across the study area

($E_1 = E_2 = 1$) was also implemented to assess the gain in the implementation of a spatially aware learner. The final clustered map via a classical RF model and the proposed SRF model can be seen in Fig. 3g, h respectively. The SRF model recognised the underlying geological structures, while the nonspatial RF model only captured unstructured noise. Finally, Fig. 3i shows the silhouette width $S(u)$ across the study area. According to this map, the channels are very well clustered, whereas the transitions between channels and marshes show lower silhouette width.

3.2 Automated Geology Mapping and Geochemical Prediction

3.2.1 Study Area and Initial Data

The overall aim of the experiments with real geoscience data is to investigate how aspects of the SRF model can support geoscientists to make quicker and more informed interpretations by semi-automated or automated analysis of multi-source spatial datasets. The study area covers a small part of the North West Minerals Province (NWMP) in Queensland (Fig. 4). The first experiment aims at digital geological mapping using geophysical covariates and an interpreted geology map (SRF classification), while the second experiment attempts to predict the concentration of Cu as an indicator of possible mineralisation in rock formations using in situ geochemical soil samples (SRF regression).

All the geophysical covariates were acquired from the Geological Survey of Queensland (GSQ, https://geoscience.data.qld.gov.au). The digital elevation model

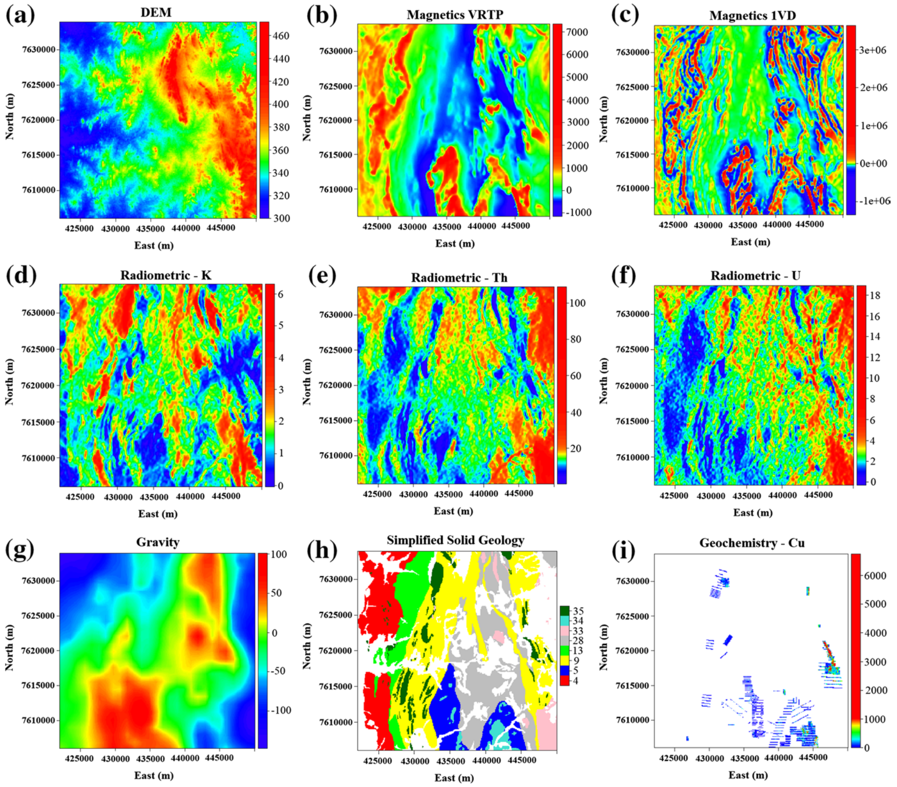

Fig. 4 The input spatial data to train predictive models. **a** to **g** Continuous regionalised variables, **h** categorical response variable, and **i** continuous response variable. To enhance the visualisation of spatial variables, the colour scales for the continuous variables are defined according to quantile (equal probability) classes

(DEM, Fig. 4(a)) was obtained from the 1 arc second DEM of bare-earth with attempted removal of vegetation and man-made structure effects (Gallant et al. 2011). The variable reduction to the pole (VRTP) correction was applied to the merged total magnetic intensity grid (Fig. 4b). Apart from the VRTP map, the first vertical derivative (IVD) of the magnetic field (Fig. 4c) was also used to enhance the resolution of closely spaced and superposed anomalies (Greenwood 2018). The gamma-ray spectrometry data (Fig. 4d–f) derived from Geoscience Australia's Third Edition Radiometric Map of Australia (Minty et al. 2010). The Bouguer anomaly gravity grid (Fig. 4(g)) represents a gridded compilation of all freely available gravity observations from exploration companies and state and federal regional surveys maintained by the GSQ (Cant 2014). The simplified solid geology (interpreted by GSQ, Fig. 4(h)) is used as the categorical response variable for the classification experiment. The in situ soil geochemistry data were obtained from the Queensland Government Department of Natural Resources and Mines (QDEX 2018). The Cu samples from 3,755 locations across the study area are used to train the random forests regressors (Fig. 4(i)). Although dependent and independent layers can have different spatial resolutions in the proposed model, the resolution of the predictors and the response variables is 90 m in this case.

3.2.2 SRF Classification

Geological maps are normally products of field observations and interpretation of geophysical and remote sensing datasets, as well as expert background knowledge about the geological history of the map area. This experiment investigates the suitability of the SRF model to provide a repeatable and faster method for automated geology mapping. The model utilises the existing geological interpretations (Fig. 4(h)) by drawing samples from it and demonstrates how the published geology can be replicated when all possible geological classes are known a priori and training samples are available from each class. A balanced training dataset was obtained by randomly sampling 84 pixels from each geological class ($N = 672$). The overall training pixels constitute 1\% of the interpreted geology map (Fig. 4(h)). The remaining 99\% of the pixels will be used to validate the predictions.

The number of bootstrap samples (number of spatial decision trees) is set to $J = 1{,}000$. The number of regionalised variables is R = 7 (Fig. 4a–g), and $E_r = 100$ for all covariates. The overall number of spatial predictors is $P = \sum_{r=1}^{7} E_r = 700$. The default value was selected for the parameter $m = ceiling\left(\sqrt{\sum_{r=1}^{7} E_r}\right) = 27$ in the classification framework. The OOB error rate for the SRF classifier is $E_{OOB} = 0.134$. The SRF model returns a measure of predictor importance that reveals which input covariates provide the most value (Fig. 5). The gravity layer and thorium concentrations from the gamma-ray spectrometric data are more important than the other covariates for predicting the true underlying geology at the selected scale (900 m × 900 m). The zone of influence for each regionalised covariate can be investigated further. For instance, the gravity zone of influence shows long-range spatial correlations, which means that this covariate captures large geological

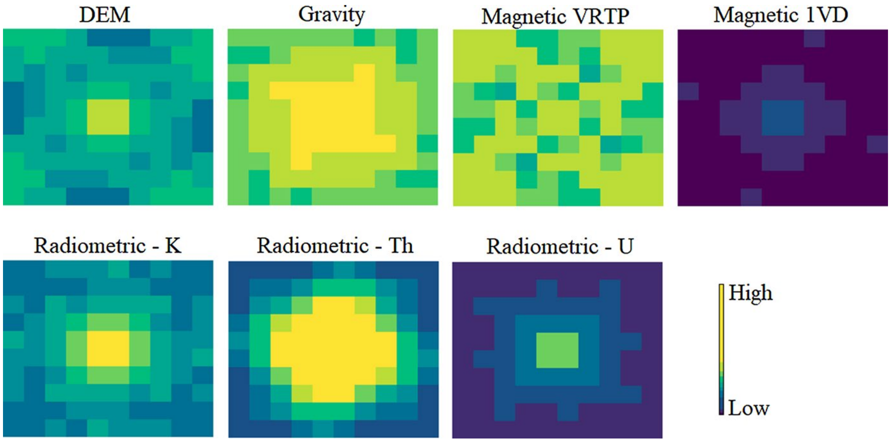

Fig. 5 Spatial predictor importance and the zone of influence for the input covariates obtained from the SRF classifier

structures with specific gravity features. Unlike the gravity layer, the zone of influence for magnetic VRTP is anisotropic and shows a clear orientation of the geological patterns with specific shallow magnetic responses at the selected scale. Finally, the thorium layer captures isotropic small-scale geological structures with exceptionally high thorium concentration.

A classical RF classifier ($E_r = 1; r = \{1, \ldots, 7\}$) was also implemented to compare the results of the SRF model with a nonspatial ensemble leaner. Figure 6a and b show the predicted surface geology by the SRF and RF classifiers respectively. The misclassified pixels (based on the 99\% validation pixels) are shown in Fig. 6c. There is a higher proportion of misclassified pixels for the RF predicted geological map compared to the SRF results. The SRF predicted geology map shows greater spatial continuity, whereas the RF map displays a distinct lack of spatial continuity and a noisy appearance. Figure 6d shows the omnidirectional experimental variograms for geological class #9. The larger nugget effect and shorter variogram range generated by the RF model show the limitations of nonspatial learners in reproducing the spatial continuity of geoscience data. The SRF model was able to correctly reproduce the geological picture with similar details as the interpreted geology map (Fig. 4h). The class-wise and overall error rates for the two models are shown in Table1. All the geological classes were predicted with higher accuracy using the SRF model. Geological classes #13, #4, and #33 show the lowest error rates, while class #9 shows the highest error rate in the SRF model. The local uncertainty of the geological class predictions can be measured by the standardised entropy (Eq. 3). The entropy maps can be used interactively during the mapping process to gain valuable information during fieldwork through the identification of areas of high variability and uncertainty. Figure 6e and f show the entropy maps for the SRF and RF models respectively. Compared to the RF model, geological predictions show lower uncertainty in the SRF model. Geological class #33 shows the lowest level of uncertainty, which might be related to the physical homogeneity of this unit. The

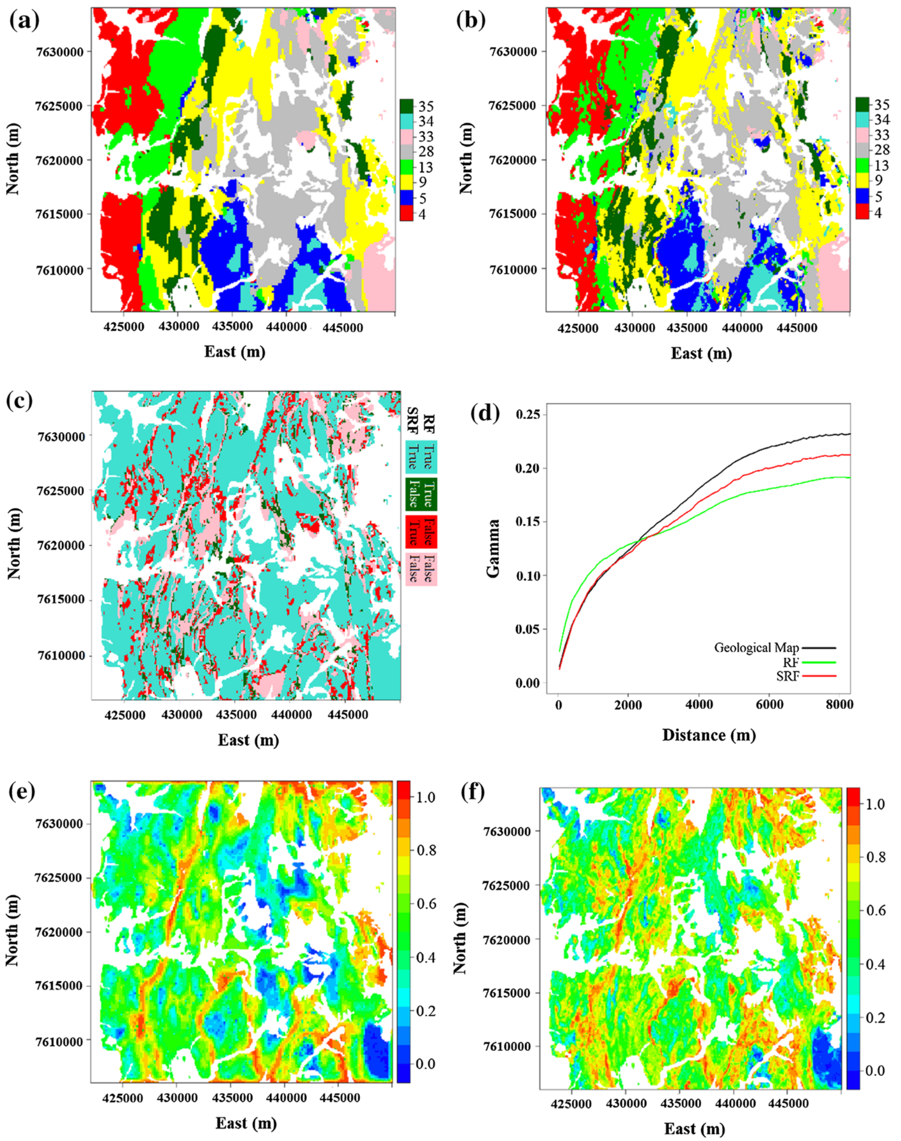

Fig. 6 Comparing the results of the SRF and RF classifiers, **a** and **b** predicted geology maps via SRF and RF models respectively, **c** misclassified pixels, **d** omnidirectional experimental variograms for the geological class #9, e and **f** entropy maps for the SRF and RF models respectively

Table 1 Class-wise and overall error rate for the SRF and RF classifiers

Class error	#4	#5	#9	#13	#28	#33	#34	#35	#All
RF	0.1708	0.2543	0.4154	0.1746	0.2672	0.1608	0.1645	0.2822	0.2837
SRF	0.0795	0.1952	0.3762	0.0756	0.2052	0.0801	0.1229	0.2268	0.2211

transitions from class #13 to classes #9 and #35 and boundaries of class #5 show high uncertainty (Fig. 6e). High uncertainty of transition zones is partly related to the fact that geological units are also interpreted models themselves utilising the same geophysical covariates. Deterministic and subjective approaches to generate geological maps can be efficiently replaced with objective approaches capable of modelling uncertainty such as SRF classification. This case study also revealed that the traditional geological maps can be revisited to highlight the areas with the highest level of uncertainty.

This experiment suggests that if a field geologist was able to identify all potential geological classes within a given spatial domain successfully and collect a reasonable amount of ground truth observations, then an SRF model could be trained (using the relevant spatial covariates) to predict the spatial distribution of the geological classes. The SRF model can provide a meaningful automated prediction of geological units which could assist in geological mapping and help to guide field work while minimising the time needed to produce a geological map.

3.2.3 SRF Regression

This experiment describes an approach for the prediction of in situ soil geochemistry (Cu concentration) even under deeper regolith that covers the bedrocks. The relationships between surface geochemical measurements and subsurface geophysical data are investigated. Mineral exploration in such covered areas strongly depends on the use of subsurface geophysical data as they can indirectly map the deep and covered geological structures that control mineralisation. However, the influence of cover on geophysical data was not considered in this study. The SRF regressor was used to model Cu concentration across the study area based on subsurface geophysical covariates only (i.e., magnetic and gravity layers).

The training samples consist of $N = 3{,}755$ Cu observations that are mostly located in the southern part of the study area (Fig. 4(i)). The number of spatial decision trees in the SRF regressor is set to $J = 1{,}000$. The number of regionalised variables is $R = 3$ (Fig. 4(b), (c) and (g)), and $E_r = 100$ for all covariates. The overall number of spatial predictors is $P = \sum_{r=1}^{3} E_r = 300$. The default value was selected for the parameter $m = ceiling\left(\frac{\sum_{r=1}^{3} E_r}{3}\right) = 100$ in the regression framework. The SRF regressor returned a measure of predictor importance that reveals the most significant covariates for Cu prediction (Fig. 7).

The gravity layer is the most important covariate (Fig. 7) for predicting Cu concentration. The zone of influence for gravity covariate shows isotropic behaviour at small scales (less than 300 m) and anisotropic behaviour (NW–SE and NE–SW trends) at larger scales (greater than 900 m). The first vertical derivative (IVD) of the magnetic field is more important than the original magnetic VRTP layer for predicting Cu concentration. This may be partly related to the complex relationships between shallow magnetic features and Cu mineralisation. The zone of influence for magnetic IVD covariate is anisotropic, and the anisotropy directions change with scale.

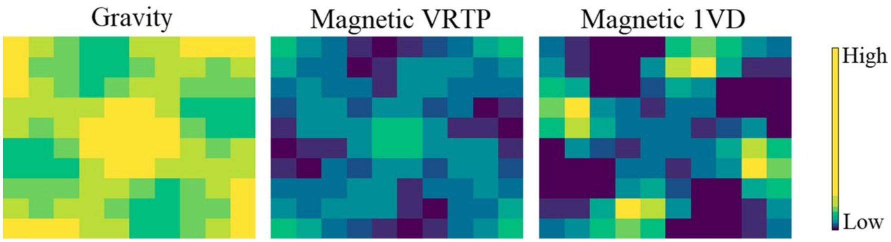

Fig. 7 Spatial predictor importance and the zone of influence for the input covariates obtained from the SRF regressor

The results of the SRF regressor were compared to those from a classical RF regressor, RF with buffer distances to observations as extra covariates (RFsp, Hengl et al. 2018), and the geographical random forests (GRF, Georganos et al. 2019). The classical RF regressor is based on three geophysical covariates ($E_r = 1; r = \{1, \ldots, 3\}$) in Fig. 4(b), (c) and (g). In addition to these three covariates, the RFsp model uses Euclidean buffer distances to 375 (10\% of the total observations) randomly selected samples. Multiple RFsp models were built by iterating the random sampling of the Euclidean buffer covariates, and the most accurate RFsp model (i.e., lowest MSE_{OOB}) was selected for prediction. Finally, the GRF model predicts the response values by fusing the results from multiple local sub-RFs and a global RF to account for spatially heterogeneous data. The local sub-RFs are trained using a number of nearby observations (i.e., adaptive kernel). The most accurate GRF model was obtained by setting the number of nearest neighbours in the adaptive kernel to 400 and the fusing weights to 0.4 and 0.6 for the local and global models respectively. To implement the GRF analyses, the R package ‘SpatialML’ (https://cran.r-project.org/web/packages/SpatialML/index.html) was used.

The number of trees in all the experiments was set to 1,000. The default value (one third of the covariates) was selected for the number of covariates randomly sampled as candidates at each split in the trees.

The generalisation errors for the four random forests experiments were estimated via the out-of-bag mean squared error. The SRF regressor showed superior performance (MSE_{OOB} = 132,391) compared to other models in terms of predicting OOB Cu concentrations. The GRF regressor was the second most accurate model (Table 2).

Figure 8 shows the Cu prediction via the four regressors. The models are different, especially in terms of extrapolation under cover. For instance, unlike the SRF model, the RF regressor predicted an anomalous area in the southwestern part of the study area. Several unrealistic artefacts can be seen in the maps generated by RFsp and GRF models. Consequently, these models are not recommended for extrapolation under deeper regolith that is the main goal of this case study.

Table 2 Out-of-bag mean squared error for the four random forests regressors

	SRF	RF	RFsp	GRF
MSE_{OOB}	132,391	180,348	167,234	159,412

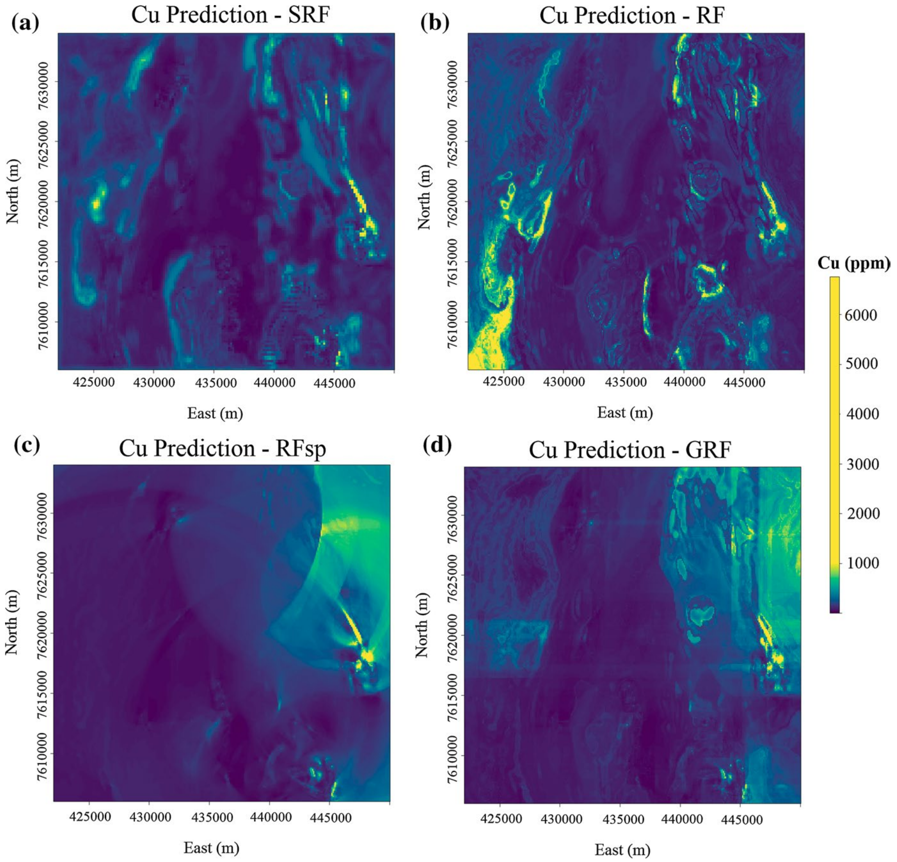

Fig. 8 Cu prediction via four random forests regressors

The correlation between true Cu concentrations and Cu prediction (OOB cases) via the RF model is 0.518, while this correlation is 0.685 for the SRF model (Fig. 9(a) and (b)). The correlations for the RFsp and GRF models are 0.491 and 0.502 respectively (Fig. 9(c) and (d)). The conditional bias assessment for the four models can be seen in Fig. 9(e) to (h), where the errors ($\widehat{f}_{OOB}\left(x(u_i)\right) - y\left(u_i\right)$) are plotted against the observed Cu concentrations. The lowest conditional bias was achieved by the SRF model.

Based on the implemented performance analyses, the predicted Cu map obtained from the SRF model is more accurate and reliable for Cu exploration. This experiment showed the potential of using geophysical datasets and soil assay samples to gain insight into the surface and subsurface geochemistry.

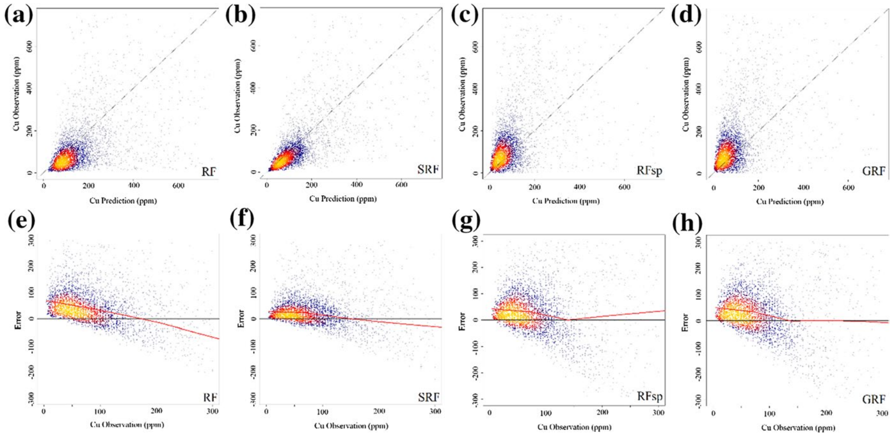

Fig. 9 **a** to **d** Scatterplots of the true Cu concentrations and random forests predictions (OOB cases). **e** to **h** Distribution of errors against true Cu concentrations. The red lines are the locally weighted polynomial regressions and the scatterplots are coloured by kernel density estimates

4 Discussion

Unlike other attempts to improve the spatial awareness of the tree-based learners, the proposed spatial random forests model is capable of learning complex spatial patterns and using such knowledge during the classification, regression, and unsupervised modelling tasks. Multi-resolution covariates can be used as spatial predictors without any preprocessing (e.g., upscaling all covariates to a consistent resolution). The spatial framework for measuring the predictor importance provides valuable knowledge about the underlying spatial structures of predictors that control the patterns in the response variable. For instance, anisotropic behaviour at various scales can be captured by calculating the zone of influence for each regionalised variable. The unsupervised SRF model is very robust against noisy covariates without meaningful auto- and/or cross-correlations. The synthetic case study reveals the robustness of the SRF model in terms of automatically ignoring such noisy covariates.

Multi-resolution gridded data can be used as input to the SRF model. However, if the input data do not lie on a regular grid, they should first be migrated to the closest nodes of a regular grid of suitable resolution or be rasterised using geostatistical techniques. The SRF model is not suitable for sparse scattered data unless representative training images are available for training. Missing values are allowed, and no separate imputation step is necessary.

Compared to the classical nonspatial random forests algorithm, the proposed spatial model is computationally demanding. However, the SRF algorithm is easily parallelisable, and parallel computing with multi-core processors can speed up such intensive calculations.

Although the spatial continuity of the predicted response variable is more realistic in the SRF model compared to the noisy outcomes predicted by nonspatial learners such as RF, some artefacts or spatial discontinuity can be seen. Such spatial discontinuities are partly related to the prediction function. The response values are predicted for each pixel independent of the predictions in nearby pixels. The pixelwise and independent prediction is not consistent with the spatial dependency of response values. The other limitation is the nonspatial bootstrapping or random sampling for training the base learners. Since the spatial and/or temporal response variables are correlated through space and time, nonspatial bootstrapping or random sampling leads to over-optimistic learners. These limitations are the topics of future studies.

5 Conclusions

A truly spatial random forests technique based on higher-order spatial statistics was introduced in this study. The proposed non-parametric technique is capable of learning multivariate spatial patterns of mixed type (i.e., continuous and categorical spatial patterns) and capturing complex nonlinear relationships. The algorithm measures the predictor importance internally and estimates the zone of influence for each regionalised variable. Unlike deep learning techniques, the high-dimensional systems in which the number of predictors is much larger than the number of observations can be handled appropriately via the SRF model. The algorithm accepts multi-resolution covariates without any preprocessing (i.e., interpolation, upscaling, and imputation). The algorithm can be used for both supervised and unsupervised learning. The case studies demonstrate applications where a spatial learner has the potential to improve, aid, or automate existing processes around the preparation and validation of geoscience data and interpretations.

Acknowledgements This study was funded by the Deep Earth Imaging Future Science Platform, CSIRO. We greatly thank Professor Ute Mueller (School of Science, Edith Cowan University) for her constructive comments on a previous version of this manuscript.

References

Bergen KJ, Johnson PA, de Hoop MV, Beroza GC (2019) Machine learning for data-driven discovery in solid earth geoscience. Science. https://doi.org/10.1126/science.aau0323

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC Press, New York

Cant R (2014) Queensland gravity grid. Accessed 9 Aug 2018

Cliff A, Ord J (1973) Spatial autocorrelation. Pion, London

Gallant JC, Dowling TI, Read AM, Wilson N, Tickle P, Inskeep C (2011) 1 second SRTM derived digital elevation models user guide. Geoscience Australia. www.ga.gov.au/topographic-mapping/digital-elevation-data.html

Georganos S, Grippa T, Gadiaga AN, Linard C, Lennert M, Vanhuysse S, Mboga N, Wolff E, Kalogirou S (2019) Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int. https://doi.org/10.1080/10106049.2019.1595177

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York

Greenwood M (2018) Queensland merged RTP, Queensland merged magnetic 1VD. Accessed 9 Aug 2018

Hengl T, Heuvelink GBM, Kempen B, Leenaars JGB, Walsh MG, Shepherd KD, Sila A, MacMillan RA, de Jesus J, Tamene L, Tondoh JE (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

Hengl T, Nussbaum M, Wright MN, Heuvelink GBM, Gräler B (2018) Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6:e5518. https://doi.org/10.7717/peerj.5518

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. Ann Appl Stat 2:841–860. https://doi.org/10.1214/08-AOAS169

Karpatne A, Ebert-Uphoff I, Ravela S, Babaie HA, Kumar V (2019) Machine learning for the geosciences: challenges and opportunities. IEEE Trans Knowl Data Eng 31:1544–1554. https://doi.org/10.1109/TKDE.2018.2861006

Kassambara A (2017) Practical guide to cluster analysis in R: unsupervised machine learning, 1st edn. Create Space, North Charleston

Kaufman L, Rousseeuw PJ (eds) (1990) Finding groups in data. Wiley, Hoboken

Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, New York

Liu Y, Cao G, Zhao N, Mulligan K, Ye X (2018) Improve ground-level PM25 concentration mapping using a random forests-based geostatistical approach. Environ Pollut 235:272–282. https://doi.org/10.1016/j.envpol.2017.12.070

Mariethoz G, Caers J (2015) Multiple-point geostatistics: stochastic modeling with training images. Wiley, New York

Matheron G (1962) Traite´ de Ge´ostatistique Applique´e. Technip, Paris

Mead A (1992) Review of the development of multidimensional scaling methods. J R Stat Soc 41:27–39

Meyer H, Reudenbach C, Wöllauer S, Nauss T (2019) Importance of spatial predictor variable selection in machine learning applications—moving from data reproduction to spatial prediction. Ecol Modell 411:108815. https://doi.org/10.1016/j.ecolmodel.2019.108815

Minty BR, Franklin R, Milligan P, Richardson L, Wilford J (2010) Radiometric map of Australia (2nd Edition), scale 1:15 000 000, Geoscience Australia, Canberra

Mitchell B, Sheppard J (2019) Spatially biased random forests. In: FLAIRS conference

Probst P, Wright MN, Boulesteix A (2019) Hyperparameters and tuning strategies for random forest. Wiley Interdiscip Rev Data Min Knowl Discov. https://doi.org/10.1002/widm.1301

QDEX (2018) Department of natural resources queensland government and mines. QDEX Data. http://qdexdata.dnrm.qld.gov.au

Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat (2019) Deep learning and process understanding for data-driven earth system science. Nature 566:195–204. https://doi.org/10.1038/s41586-019-0912-1

Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, Ross AS, Milojevic-Dupont N, Jaques N, Waldman-Brown A, Luccioni A, Maharaj T, Sherwin ED, Mukkavilli SK, Kording KP, Gomes C, Ng AY, Hassabis D, Platt JC, Creutzig F, Chayes J, Bengio Y (2019) Tackling climate change with machine learning. ArXiv Preprint, arXiv:1906.05433

Sellars SL (2018) “Grand challenges” in big data and the earth sciences. Bull Am Meteorol Soc 99:95–98. https://doi.org/10.1175/BAMS-D-17-0304.1

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Talebi H, Mueller U, Tolosana-Delgado R, Grunsky EC, McKinley JM, de Caritat P (2019) Surficial and deep earth material prediction from geochemical compositions. Nat Resour Res 28:869–891. https://doi.org/10.1007/s11053-018-9423-2

Talebi H, Peeters LJM, Mueller U, Tolosana-Delgado R, van den Boogaart KG (2020) Towards geostatistical learning for the geosciences: a case study in improving the spatial awareness of spectral clustering. Math Geosci 52:1035–1048. https://doi.org/10.1007/s11004-020-09867-0

Woodcock CE, Strahler AH, Jupp DLB (1988) The use of variograms in remote sensing: II. Real digital images. Remote Sens Environ 25:349–379. https://doi.org/10.1016/0034-4257(88)90109-5