International Journal of Applied Earth Observation and Geoinformation 121 (2023) 103364

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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

Spatial+: A new cross-validation method to evaluate geospatial machine learning models

Yanwen Wang [*], Mahdi Khodadadzadeh, Raúl Zurita-Milla

Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7514AE Enschede, The Netherlands

ARTICLE INFO

Dataset link: https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:278047

Keywords:
Data-driven models
Model evaluation
Cross-validation
Spatial autocorrelation
Feature space

ABSTRACT

Random cross-validation (CV) is often used to evaluate geospatial machine learning models, particularly when a limited amount of sample data are available, and collecting an extra test set is unfeasible. However, the prediction locations can be substantially different from the available sample, leading to over-optimistic evaluation results. This has fostered the development of spatial CV methods. Yet these methods only focus on spatial autocorrelation and cannot sufficiently guarantee that the validation subset is a good proxy of the test set with significant differences. In this paper, we propose the spatial+ cross-validation (SP-CV) method. This method, which considers both the geographic and feature spaces, is composed of two stages. The first stage addresses spatial autocorrelation issues by using agglomerative hierarchical clustering to divide the available sample into blocks. The second stage deals with multiple sources of differences. It uses cluster ensembles to split the blocks into training and validation folds based on the locations of the sample data and the values of the covariates and target variable. The proposed method is compared against random and block CV methods in a series of experiments with Amazon basin above ground biomass and California houseprice datasets. Our results show that SP-CV provided the smallest error differences with respect to the reference error. This means that SP-CV produced more representative splits and led to more reliable model evaluations. It suggests that a reliable model evaluation requires to consider both the geographic and the feature spaces in a comprehensive manner.

1. Introduction

Spatially continuous variables are needed in many geoscience studies. However, due to economic and time constraints, many variables are just collected from a limited set of locations (Zhu et al., 2015). At the same time, the availability of fundamental geographic datasets and products, such as digital elevation models, gridded climatic data, and Earth observation images, is constantly increasing. These datasets and products can be used as covariates to create spatially continuous variables. For this, the sampled target variable and the corresponding covariates are used to build a predictive model. This approach to generate spatially continuous variables, called geospatial prediction, is now widely embedded in many geoscience studies and applications (Meyer et al., 2019).

Because of its remarkable performance in solving complex problems, machine learning (ML) is commonly used to build geospatial prediction models (Gao et al., 2022; Wei et al., 2022). For instance, ML has been used for soil (Hengl et al., 2015) and crop mapping (Aguilar et al., 2018), ecological modeling (Dang et al., 2019), mineral studies (Khodadadzadeh and Gloaguen, 2019), crime forecasting (Kounadi et al., 2020) and geo-health studies (Garcia-Marti et al., 2018). More recently, ML has played an important role in forecasting the spatial distribution of the COVID-19 pandemic (Pourghasemi et al., 2020).

Model evaluation is a crucial step in geospatial prediction (Pohjankukka et al., 2017). To obtain reliable evaluation results, a test set that unbiasedly represents the prediction locations (e.g., the test set collected by probability sampling, Brus et al. (2011)) is needed (Wadoux et al., 2021). However, collecting new sample data to create a test set is usually unfeasible due to many practical challenges, such as sampling cost, research urgency, and sample data shortage (Valavi et al., 2019). Therefore, the available sample are customarily split into two subsets: the training subset, which is a proxy of the available sample data used to build the model, and the validation subset, which is a proxy of the test set, as such, it is used to evaluate the model. In this context, random k-fold cross-validation (CV) is frequently used for such a data splitting. The key idea of k-fold CV is to randomly split the available sample data into k non-overlapping folds that are then used as the validation subset one by one (in an iterative manner) with the remaining k-1 folds

[*] Corresponding author.
E-mail address: y.wang-4@utwente.nl (Y. Wang).

https://doi.org/10.1016/j.jag.2023.103364
Received 5 February 2023; Received in revised form 15 May 2023; Accepted 18 May 2023
Available online 29 May 2023

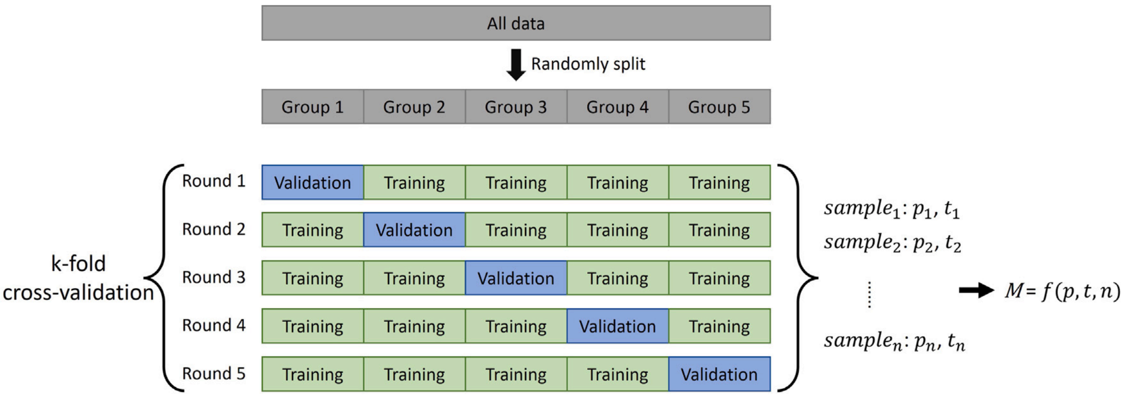


Fig. 1. The workflow of 5-fold random cross-validation.

used as the training subset. Fig. 1 shows the workflow of a typical 5-fold CV (i.e., $k = 5$). In the end, because every sample ($sample_n$) gets a prediction value (p_n) and a true value (t_n), a model's performance metric (M) can be calculated based on all n prediction-true values pairs. Random CV has long been proven accurate (Efron, 1983) and efficient for model's evaluation and is often used by the statistical and computer science communities. It has also been used in plenty of geospatial prediction ML modeling studies (Chen et al., 2018; Nesha et al., 2020). In recent years researchers have showed that, when the available sample data represent the prediction locations well (e.g., the samples are uniformly distributed in the prediction area), random CV provides accurate evaluations (Wadoux et al., 2021).

However, in practical geospatial prediction situations, there is no guarantee that the available sample data represent the prediction locations well. In fact, in many cases, samples and prediction locations are different. For example, samples are usually spatially clustered (i.e., unevenly distributed (Li et al., 2020)) due to multiple data sources (Meyer and Pebesma, 2022; de Bruin et al., 2022) or other reasons. This phenomenon is frequently observed in ecological (Ploton et al., 2020), air pollution (Xiao et al., 2018), and soil (Hengl et al., 2015) data, especially on a global scale (Hooker et al., 2018; Meyer and Pebesma, 2022). When the available sample data are clustered, they tend to over-represent the specific regions where samples are distributed, resulting in under-representation of other regions in the prediction locations (de Bruin et al., 2022). Another case is extrapolation, i.e., the geospatial prediction model should be applied in a new area (Roberts et al., 2017). For instance, when working on landslides (Brenning, 2005; Wei et al., 2022) or invasive-species diffusion (Cheng et al., 2018), the sample can only be collected from areas where landslides or species invasion already occurred, but the prediction locations are the new areas where this phenomenon has not yet occurred. In this circumstance, similar to the clustering samples case mentioned above, the prediction locations are different and cannot be represented by the available sample data.

The actual cases that sample data are different from prediction locations bring challenges to the evaluation of geospatial ML models (Roberts et al., 2017). While recent studies have justified the use of random CV in certain evaluations, it should be noted that this approach may not be suitable for all situations. These studies themselves admit that situations such as strong clustering of samples can still pose challenges for model evaluation using random CV (de Bruin et al., 2022). In particular, a good model evaluation needs to check the model's generalization ability (Beigaitė et al., 2022), which requires the selection of a validation subset that can be representative of a possible test set, especially when the test set (at the prediction locations) is different from the available sample. However, in random CV, validation sample could be very close to the training sample because of the random split. In this case, both the target variable and the corresponding covariates could be very similar (Gao et al., 2022). As a result, the validation sample is actually "(pseudo) replica" of the training sample instead of a proxy of the test sample. This leads to the situation, where the built models in traditional k-fold CV method are over-fitted and the derived evaluation results are over-optimistic (Brenning, 2005; Wiens et al., 2008; Xu et al., 2021).

Since the early 2000s, a series of spatial CV methods have been proposed to avoid the limitations of random CV. All the spatial CV methods originate from the natural and straightforward idea — avoiding or mitigating spatial autocorrelation when splitting training and validation samples (Oliveira et al., 2021; Beigaitė et al., 2022). For example, the buffer leave-one-out cross-validation (BLOOCV) removes spatially autocorrelated training sample data by considering a spatial buffer around the selected validation samples (Le Rest et al., 2014; Valavi et al., 2019). The weighted CV reduces the importance of high-density sample data in CV to decrease the influence of spatial autocorrelation in evaluation (Sarafian et al., 2021; de Bruin et al., 2022). Another example is block CV that first divides all samples into contiguous blocks, and then avoids the selection of samples within the same block as both training and validation samples (Brenning, 2005; Valavi et al., 2019; Kollert et al., 2021).

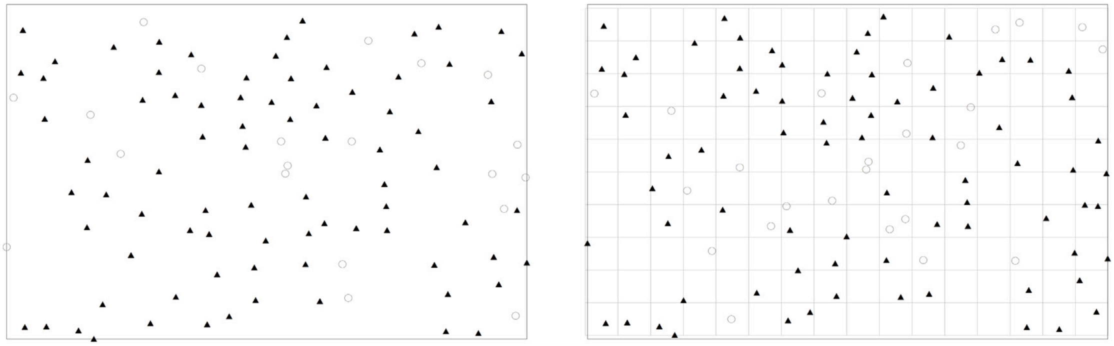
Spatial CV methods, as mentioned earlier, may not be sufficient to account for differences between available sample data and prediction locations when these differences are significant (Meyer and Pebesma, 2021; Milà et al., 2022). This is because spatial autocorrelation is only one source of differences, while there are many other sources from both the geographic and feature spaces (Roberts et al., 2017). Especially in the feature space as covariates and target variable, there are often remarkable differences between available sample data and prediction locations (Wadoux et al., 2021; Meyer and Pebesma, 2022). When splitting the validation subset to represent the test set, these multiple sources of differences should be considered comprehensively.

In this paper, we propose a new CV method to evaluate ML models used for geospatial prediction. Our method addresses situations where the sample data are different from the prediction locations. It also addresses the multiple sources that constitute such differences (especially the differences in the feature space) along with the consideration of spatial autocorrelation. By doing this, we guarantee that the validation subset split better reflects the differences between the training and test sets.

The remainder of this paper is organized as follows. In Section 2, we introduce the main steps of the proposed method, and briefly describe the traditional and block CV methods, which are used to benchmark our proposed method. In Section 3 we describe the experimental setup designed to assess the added value of the proposed method. In Section 4 we present and discuss our results and, finally, Section 5 contains our main conclusions and recommendations for future research.

2. Methods

In this section, we further elaborate on the traditional random k-fold CV (RDM-CV), and a typical spatial CV — the block k-fold CV (BLK-CV). Then, we discuss the detailed design of the proposed method.



(a) Random k-fold cross-validation (RDM-CV)

(b) Block k-fold cross-validation (BLK-CV)

Fig. 2. A graphical representation of compared CV methods (k = 5). Legends: hollow circles (∘) are validation samples, black triangles (▲) are training samples, squares (□) in (b) are divided blocks.

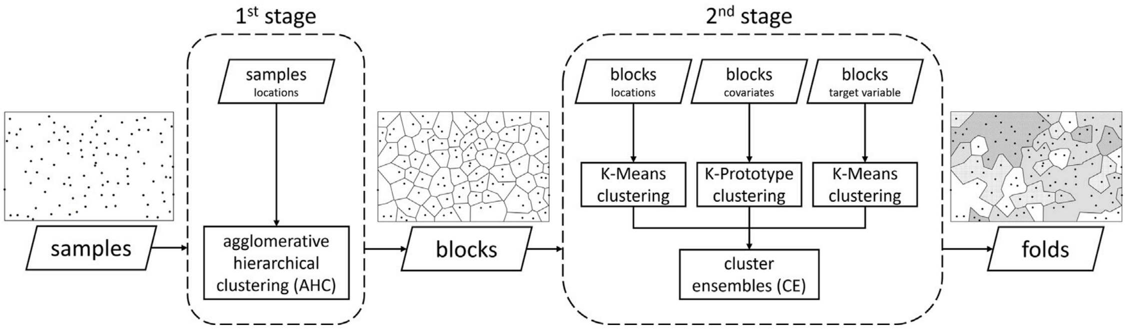


Fig. 3. The flowchart of the proposed method. (The subfigures above the samples, blocks, and folds are their examples.).

2.1. Random and block cross-validation methods

Just as its name implies, RDM-CV split the available sample data into k equal size folds randomly. Fig. 2(a) shows an example of RDM-CV split for a spatial dataset. In this figure, 1/5 of the samples are collected in one fold and used as the validation subset, and the remaining samples are used as the corresponding training subset. We can also see that the validation sample data are surrounded by and close to a lot of training sample data, which will lead to autocorrelation.

BLK-CV tries to avoid the spatial autocorrelation by dividing the sample data into contiguous blocks. The squares shown in Fig. 2(b) are the most common block shape (Brenning, 2005; Lyons et al., 2018; Valavi et al., 2019). The block size is typically equal to the spatial autocorrelation threshold (Roberts et al., 2017). In the BLK-CV method, every fold is created by randomly selecting blocks instead of samples. In this way, training and validation samples are forced to be in different blocks. As shown in Fig. 2(b), within the same block, the sample data belong to either the validation or the training subsets.

2.2. The proposed method: spatial+ cross-validation

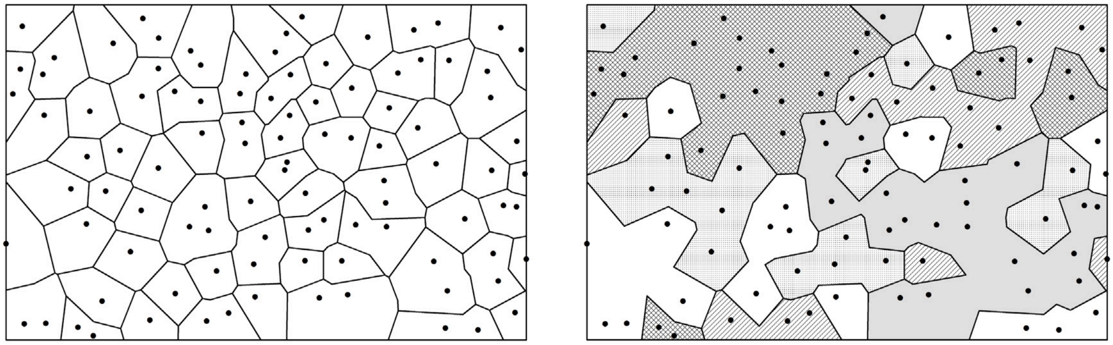
In this section, we present the proposed CV method called spatial+ cross-validation (SP-CV). Fig. 3 summarizes the complete methodology by a detailed flowchart. As shown in Fig. 3, the proposed method is composed of two stages. The first stage is similar to BLK-CV. It adopts the idea of considering spatial autocorrelation in the process of samples splitting. The second stage complements the first stage by considering the multiple sources of differences. It splits the blocks from the first stage into k folds based on locations, the values of the target variable and covariates. These two stages are explained in detail in the following sections.

2.2.1. The first stage: dividing samples into blocks

The detailed steps of the first stage are shown in Algorithm 1. In this stage we adopt the idea of BLK-CV to divide sample data into blocks. However, for generating blocks, we use agglomerative hierarchical clustering (AHC) (Arabie et al., 1996) instead of using square blocks to avoid its problems as unbalanced amount sample and sample close to block boundary (Ploton et al., 2020), which are also shown in Fig. 2(b). AHC is a “bottom-top” clustering method that merges the closest samples or sub-clusters. It integrates the consideration of sample data’s spatial distribution in the blocks division process.

Algorithm 1 1st stage: using AHC to divide samples into blocks

```
Input: samples
Output: blocks
Create an empty list $waiting-clusters$ and add all $samples$ into it
Create an empty list $finished-clusters$
(Here is the start of AHC)
while $waiting-clusters$ has more than 1 cluster do
    Find clusters pair $A\&B$ whose linkage value is the smallest
    if $A\&B$'s linkage $\leqslant$ threshold then
        Merge $A\&B$ as a new cluster $C$
        Add $C$ into $waiting-clusters$
        Delete $A\&B$ from $waiting-clusters$
    else
        Add $A\&B$ into $finished-clusters$
        Delete $A\&B$ from $waiting-clusters$
    end if
end while
if $waiting-clusters$ has 1 cluster then
    Add the remaining only 1 cluster into $finished-clusters$
end if
(Here is the end of AHC)
The clusters of $finished-clusters$ are divided blocks
```



(a) The 1st stage result of AHC blocks

(b) The 2nd stage result of CE folds

Fig. 4. Example of SP-CV. These two figures are also shown in Fig. 3.

An important element in AHC is the so-called linkage, which represents the distance of sub-clusters (Murtagh, 1983). By its two functions, the linkage determines how blocks are divided. The first is determining which two sub-clusters should be merged, and the second is determining when AHC should stop, i.e., determining the size of blocks. We use the maximum linkage because it represents the maximum distance of samples within the same cluster. The maximum linkage can determine the size of clusters (blocks) in the first stage. Hence, it is adopted here.

Similar to BLK-CV, in SP-CV, the block's size is set equal to the spatial autocorrelation threshold (Roberts et al., 2017). And the spatial autocorrelation threshold is calculated by the sample data's semivariogram (Gasch et al., 2015).

Fig. 4(a) shows an example of AHC blocks produced by the first stage of SP-CV. As this figure shows, compared with the blocks of BLK-CV in Fig. 2(b), the amount of samples in the AHC blocks is more balanced and all blocks have at least one sample. Moreover, the samples are all far away from the blocks' borders, which means that AHC can better avoid spatial autocorrelation.

2.2.2. The second stage: splitting blocks into folds

The lack of spatial autocorrelation is not the only factor that can cause the differences between samples and prediction locations. These differences can arise from multiple sources in the geographic and feature spaces. As such, in SP-CV, a second stage is included to account for these factors.

The data used for spatially predicting a target variable using an ML model involves three main components: locations, covariates, and the target variable. The location information pertains to the geographic space, while covariates and the target variable are part of the feature space. To account for differences from multiple sources, all three components should be considered when splitting samples in the CV process. Clustering is a suitable tool to capture such differences in the splitting of the training and validation subsets. For example, Schratz et al. (2019) used K-Means clustering to split samples into five folds based on sample data's locations. Thus, in the second stage, we suggest using a clustering approach based on all three components (locations, covariates, and the target variable) to split the blocks. This approach reflects the differences that can arise from multiple sources in both the geographic and feature spaces.

Due to the fact that the number of covariates is typically much larger than the number of the target (one) and location (usually two, i.e., x and y) variables, using a single clustering approach on a combined set of features from all three sources may not be appropriate. This might lead to the over-representation of the differences in covariates space in the clustering. In addition, each of the three sources contains distinct types of information and patterns, a single clustering approach is unable to capture this diversity. Therefore, we suggest first performing clustering on locations, covariates, and the target variable separately, and then, combining the clustering results.

For combining three clustering results mentioned above to produce the final clusters as the split k folds, we use cluster ensembles (CE) (Strehl and Ghosh, 2002) here. CE is a method that can combine different clustering results to obtain a single comprehensive clustering result. This means, by CE, the multiple sources of differences can be considered simultaneously. The differences between sample data and prediction data can be better reflected by training and validation subsets.

Algorithm 2 2nd stage: using CE to split blocks into folds

```
Input: blocks
Output: k clusters (k folds)
For each block, by averaging contained samples' corresponding values, produce its unique values of coordinates, covariates, and target variable.
(Here is the start of CE)
For all blocks,
    • For considering the differences of locations, using K-Means based on coordinates to produce $k$ clusters – $clusters^{(L)}$.
    • For considering the differences of covariates, using K-Prototypes based on covariates to produce $k$ clusters – $clusters^{(C)}$.
    • For considering the differences of target variable, using K-Means based on target variable to produce $k$ clusters – $clusters^{(T)}$.
Input $clusters^{(L)}$, $clusters^{(C)}$, $clusters^{(T)}$ to consensus function (here use Hybrid Bipartite Graph Formulation (HBGF)), produce a final $k$ clusters result – $final - cluster$
(Here is the end of CE)
The $k$ clusters of $final - cluster$ are split $k$ folds of SP-CV
```

The detailed steps of the second stage is shown in Algorithm 2. The first stage divides samples into blocks for considering spatial autocorrelation. These blocks serve as the operational unit in the second stage. To begin, the values of the coordinates, covariates, and the target variable for each block should be calculated by averaging the values of all samples contained within the block.

Then, clustering is performed on the blocks for the three components – locations, covariates and the target variable – respectively. Three clustering results ($clusters^{(L)}$ (i.e., locations clusters), $clusters^{(C)}$ (i.e., covariates clusters), and $clusters^{(T)}$(i.e., target variable clusters)) are obtained separately for each source. As to the respective clusterings, for locations and target variable, K-Means clustering is used. For the covariates, we suggest applying K-Prototypes clustering method (Huang, 1998) because it can deal with mixed data types and there could be categorical variables in the list of covariates.

For the combination process using CE, consensus function (Strehl and Ghosh, 2002; Alqurashi and Wang, 2019) is the key. Consensus function can find the maximum consistency between different clustering results. In this way, the final produced clustering results can guarantee comprehensiveness as much as possible. Three respective clustering results from the previous step are input to the consensus function to acquire the final clustering result. Here, we select the Hybrid Bipartite Graph Formulation (HBGF) (Fern and Brodley, 2004) as the consensus function, because it shows reliable performances and guarantees both the similarity among instances and clusters when forming the final clustering result.

Additionally, it should be noted that the value of k in CE (both initial respective clusterings and final multiple clustering) has to be set to the number of folds. As such, when the second stage finishes, the final k clusters produced by CE, are the k folds of SP-CV.

Fig. 4(b) shows an example of CE folds produced after the second stage of the proposed SP-CV, where each region with an identical texture represents a fold. These folds are not only split by their locations, but also by the other aspects associated with the feature space.

3. Experiments

The experiments were implemented on two datasets so that we can assess our CV method corresponding to the actual cases that sample data are obviously different from the prediction locations. Both datasets are sufficiently large for our experiments and have been previously used in spatial CV studies (Wadoux et al., 2021; Agarwal et al., 2021). In the following subsections, we provide more details on the datasets, the ML model, and our experimental setup.[1]

3.1. Datasets

3.1.1. Brazil Amazon basin above ground biomass dataset

The Brazil Amazon basin above ground biomass (AGB) dataset was adopted from Wadoux et al. (2021). This dataset contains 28 covariates and one target variable, i.e., AGB. Fig. 5(a) shows the distribution of Amazon AGB dataset. All data of covariates and the target variable are based on a 928*1642 raster layer with 1 km^2 resolution.

Using this dataset, the experiments are designed to simulate the actual case of prediction with strongly clustered samples. Therefore, the selection of the samples in experiments needs to reflect the characteristic of clustering. The detailed steps will be introduced in the following experiment design subsection.

3.1.2. California houseprice dataset

The well-known California housing dataset (Pace and Barry, 1997), which contains information from the 1990 California census. It comprises 20640 records and nine covariates next to the house price, which is the target variable. This dataset covers the entire state of California (United States).

The experiments of this dataset are used to simulate the actual case of extrapolation. Therefore, the prediction locations should be based on a distinct area to reflect that model is applied in a totally new area. For defining what is a distinct area, regional information is required for this dataset. An individual region from this dataset can serve as a suitable simulation for a new area because the concept of a region involves grouping similar data and distinguishing it from other data. Hence, if we know how regions are separated in this dataset, the prediction locations can be constructed by selecting one region's data to simulate extrapolation. Considering that house price is highly related to population activities and distribution, we mainly refer to the regional stay-at-home map,[2] made for COVID-19 pandemic, and additional information from the government regions division map,[3] to obtain the regions information. Fig. 5(b) shows the California houseprice dataset, and also the regions' labels.

3.2. Machine learning method

A large variety of ML methods have been used for spatial prediction. However, the focus of this paper is on the evaluation of CV methods. Thus, similar to previous CV studies (Roberts et al., 2017; Wadoux et al., 2021), we selected just one ML method for our experiments: random forest (RF). This ML method is popular and has shown good general performance (Hengl et al., 2018; de Bruin et al., 2022; Milà et al., 2022). Additionally, RF has many advantages such as reliability (Chen et al., 2018; Filippi et al., 2019), robustness and stability (Garcia-Martí et al., 2017), and being user-friendly (Breiman, 2001; Hengl et al., 2018).

3.3. Experiments design

Fig. 6 illustrates our four-step experimental workflow. Step 1 deals with the construction of sample data and prediction locations from the entire dataset. In step 2 the prediction locations are applied to calculate the reference prediction error, i.e., e_{ref}. This reference error is used to check which CV method can provide a more accurate evaluation result. In step 3 we implement each of CV methods on the sample data. This implementation provides estimated prediction error (i.e., evaluation result) of each CV method (i.e., e_{CV}). In the final step we calculate the absolute difference (d_{CV}) of the two error values, e_{ref} and e_{CV}.

3.3.1. Step 1: Construct sample data and prediction locations

To quantify which CV method can provide a more accurate evaluation result, the reference prediction error (e_{ref}) is an indispensable element in experiments. Hence, CV methods cannot be implemented on the entire dataset. The dataset should be divided into two parts, one containing sample data to implement CV methods, and the other containing prediction locations to provide a standard prediction error i.e., e_{ref} (Oliveira et al., 2021). Thus, the first step of the experiments is constructing sample data and prediction locations.

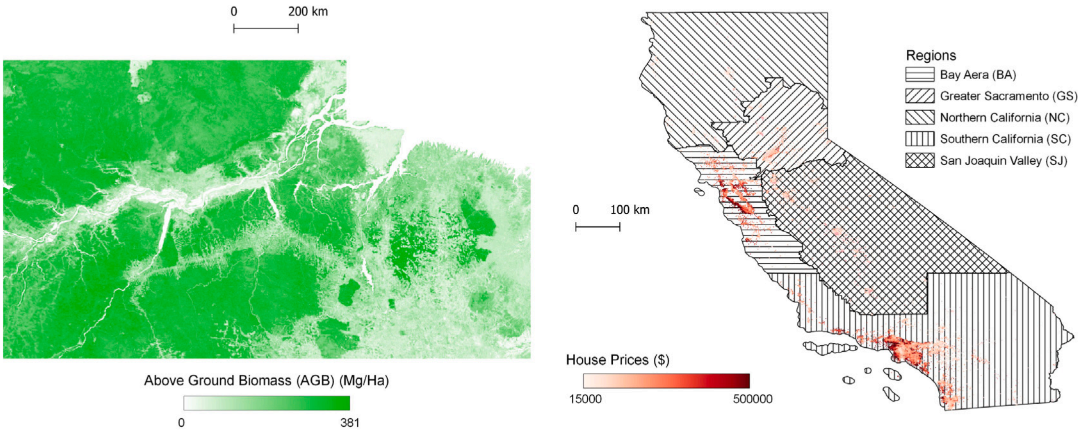
For the Amazon AGB experiments, as introduced in the description of the dataset, the strongly clustered samples should be selected first. The selection of clustered samples was adopted from Wadoux et al. (2021). The specific selection method is as follows: first, all grid points with available data are clustered into 100 sub-regions. Then, 10 sub-regions are chosen at random manner. Next, from each of 10 selected sub-regions, 100 samples are randomly selected, resulting in a total of 1000 strongly clustered samples from the study area. The remaining grid points that were not selected are considered as prediction locations. Fig. 7(a) shows an example of selected samples (green points) and prediction locations (red points) in the Amazon AGB experiment. For reducing the random effect, the sample selection process was repeated 10 times. It means all the experiments using the Amazon AGB dataset was repeated 10 times in this research.

Since California houseprice experiments are used to reflect the case of extrapolation, the prediction locations with a new area should be determined first. As explained in the subsection of datasets, an individual region's data can be used as prediction locations. Fig. 7(b) shows an example where, the data in "Greater Sacramento" region are used as the prediction locations, i.e., red points. The remaining data cannot be directly used as samples. Because there are still some data close to the prediction locations that are spatially autocorrelated

[1] The datasets and the necessary code to reproduce our results can be downloaded from: https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:278047 and https://zenodo.org/record/7796221#.ZCsjZ3bP2Uk

[2] https://www.gov.ca.gov/2020/12/03/california-health-officials-announce-a-regional-stay-at-home-order-triggered-by-icu-capacity/

[3] https://census.ca.gov/regions/



(a) Brazil Amazon basin above ground biomass dataset

(b) California houseprice dataset

Fig. 5. Samples and regions distribution of two datasets.

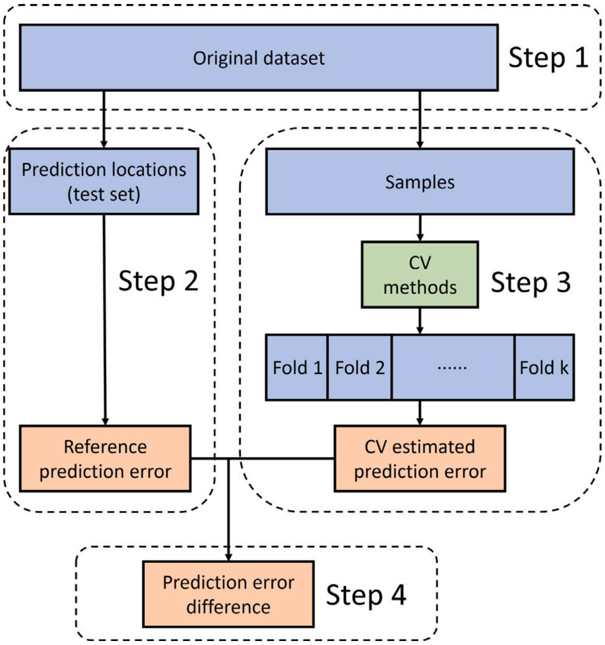

Original dataset
Step 1
Prediction locations (test set)
Samples
Step 2
CV methods
Step 3
Fold 1
Fold 2
......
Fold k
Reference prediction error
CV estimated prediction error
Prediction error difference
Step 4

Fig. 6. The workflow of the experiment.

(i.e., black points in Fig. 7(b)). These data within the autocorrelation buffers of the prediction locations should be removed, to guarantee that the prediction locations can reflect a totally new area. As Fig. 5(b) shows, there are five regions in California houseprice dataset. Every region's data will be used as prediction locations in turns. Thus, there were five California houseprice experiments in this research.

3.3.2. Step 2: Calculate reference prediction error

In step 2, all prediction locations are directly used as the test set. Therefore, the calculated value of reference prediction error (e_{ref}) can completely represent the model's performance. First, all samples (e.g., green points in Figs. 7) are used to build an ML model. Then, this model is applied to the prediction locations (e.g., red points in Figs. 7) to obtain the reference error based on the prediction of the ML model at the prediction locations and the true values.

There are various error metrics options, such as mean absolute error (MAE) (Oliveira et al., 2021), model efficiency coefficient (MEC) (de Bruin et al., 2022), and root-mean-square error (RMSE) (Wadoux et al., 2021) to calculate e_{ref}. In this paper, we used the RMSE metric, not only because it is commonly used as the sole metric in spatial CV studies (Roberts et al., 2017; Ploton et al., 2020; Wadoux et al., 2021), but also because previous studies have demonstrated that other metrics have produced similar results and conclusions to those obtained using RMSE (Oliveira et al., 2021; de Bruin et al., 2022).

3.3.3. Step 3: Calculate the prediction error of each CV method

In this step, we adopted the same strategy as in the traditional RDM-CV to calculate the prediction error (e_{CV}, i.e., evaluation result) of the BLK-CV and the proposed SP-CV. As shown in Fig. 1, after splitting k folds, one fold is taken as the validation subset and the remaining k-1 folds are used to train the ML model. Then, the trained model is validated using the validation subset and the error is calculated. This process is repeated k times until each of the folds is used as the validation subset. Finally, after all the k rounds, e_{CV} is calculated based on each sample's prediction and true values. In the Amazon AGB experiments, k was set to 10; and in the California houseprice experiments, k was set to 5. Both 10 and 5 are the most commonly used values in CV (Nesha et al., 2020; Carvalho et al., 2022). Because random selections exist in RDM-CV and BLK-CV, on the same sample set, each CV method was implemented 10 times. The final e_{CV} was computed by averaging the results of 10 times repetition to account for random errors (Ploton et al., 2020).

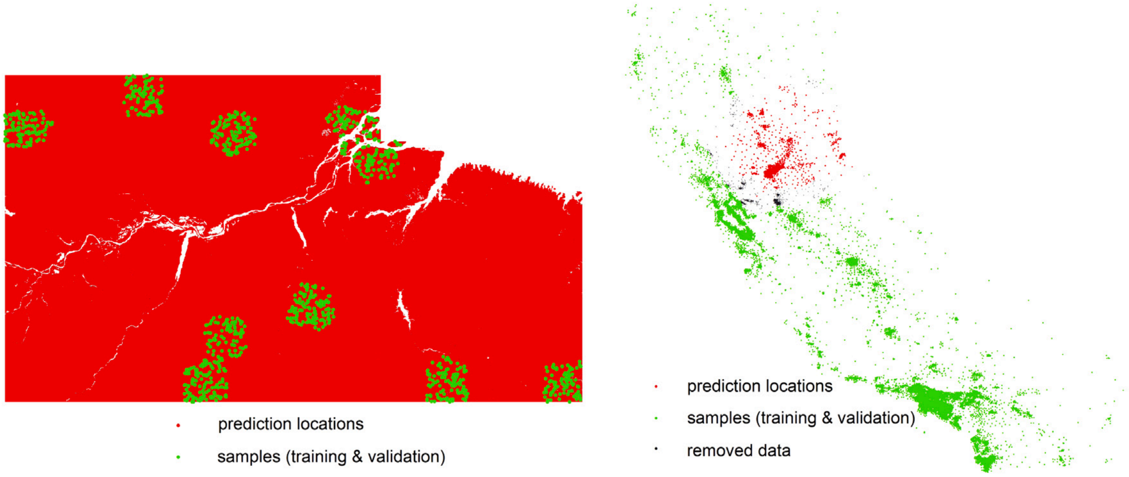
3.3.4. Step 4: Calculate CV method's prediction error difference

After step 2 and step 3, we obtain the reference prediction error – e_{ref} – and the evaluation result of every CV method (i.e., e_{CV}). By comparing them, we can find out which CV method performs better in evaluation. For this purpose, we use the CV method's prediction error difference (i.e., d_{CV}) as a quantitative metric. We calculate d_{CV} by subtracting e_{CV} from e_{ref} and getting the absolute value (i.e., $d_{CV} = |e_{CV} - e_{ref}|$). When d_{CV} is closer to zero, the corresponding CV method's performance is considered better.

Since there were 10 Amazon AGB experiments and 5 California houseprice experiments, the final performance of each CV method in these two series of experiments was calculated by averaging all the d_{CV} results.

4. Results and discussion

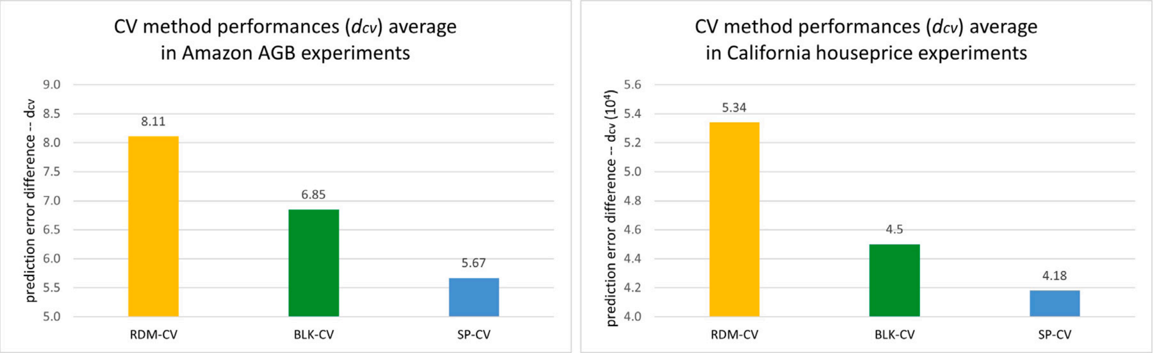
Fig. 8 shows the averaged d_{CV} results for all three CV methods. First and foremost, it is remarkable that both BLK-CV and the proposed SP-CV, which consider spatial properties of the data, produce the closer evaluation results to the reference prediction error than RDM-CV. This suggests that spatial CV may work for evaluating spatial prediction ML models when sample data and prediction locations are different.


prediction locations
samples (training & validation)
prediction locations
samples (training & validation)
removed data

(a) An example of sample data and prediction locations in Amazon AGB experiments.

(b) An example of sample data and prediction locations in California houseprice experiments.

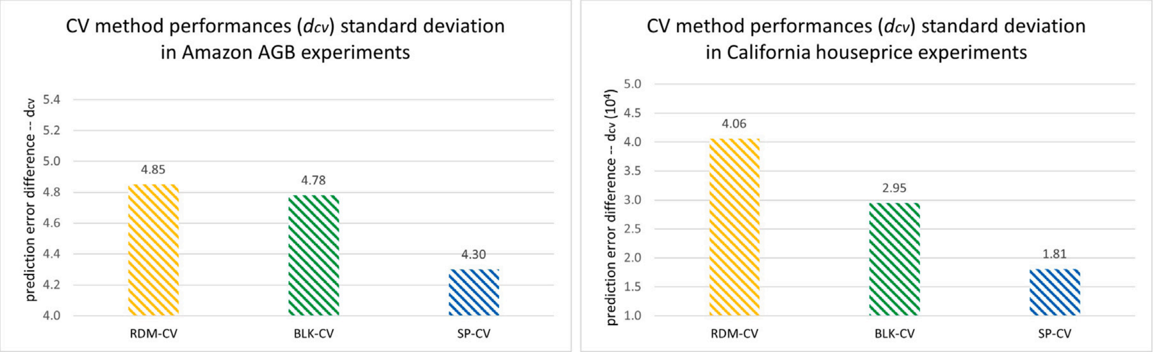
Fig. 7. The examples of sample data and prediction locations in experiments.



(a) The Amazon AGB experiments results

(b) The California houseprice experiments results

Fig. 8. The final results (prediction error difference – d_{CV}) averages of experiments.



(a) The Amazon AGB experiments results

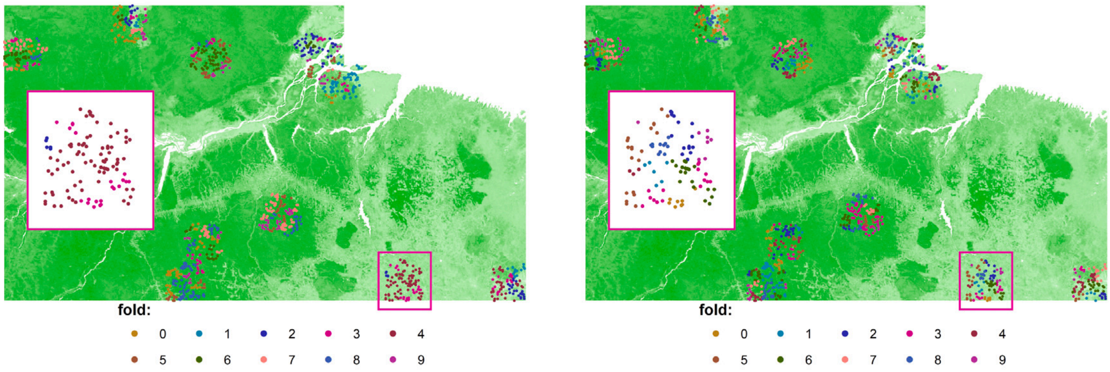
(b) The California houseprice experiments results

Fig. 9. The final results (prediction error difference – d_{CV}) standard deviations of experiments.

Furthermore, SP-CV's evaluation results were much closer to the reference prediction error than the results of BLK-CV for both datasets experiments. This shows that the proposed method, which considers both the spatial autocorrelation and the multiple sources of differences in data, can indeed provide a reasonable evaluation result. Fig. 9 shows the standard deviations of d_{CV} in two series of experiments. SP-CV still obtained the lowest standard deviation. To summarize, in experiments, SP-CV was closer to the reference metric.

For a CV method, the way folds are split determines the produced evaluation result. In order to better understand why SP-CV obtained closer evaluations in our experiments, we visually analyze the folds splitting results of this method in comparison with the BLK-CV method. Figs. 10 and 11 show two examples of folds splitting using SP-CV and BLK-CV methods respectively.

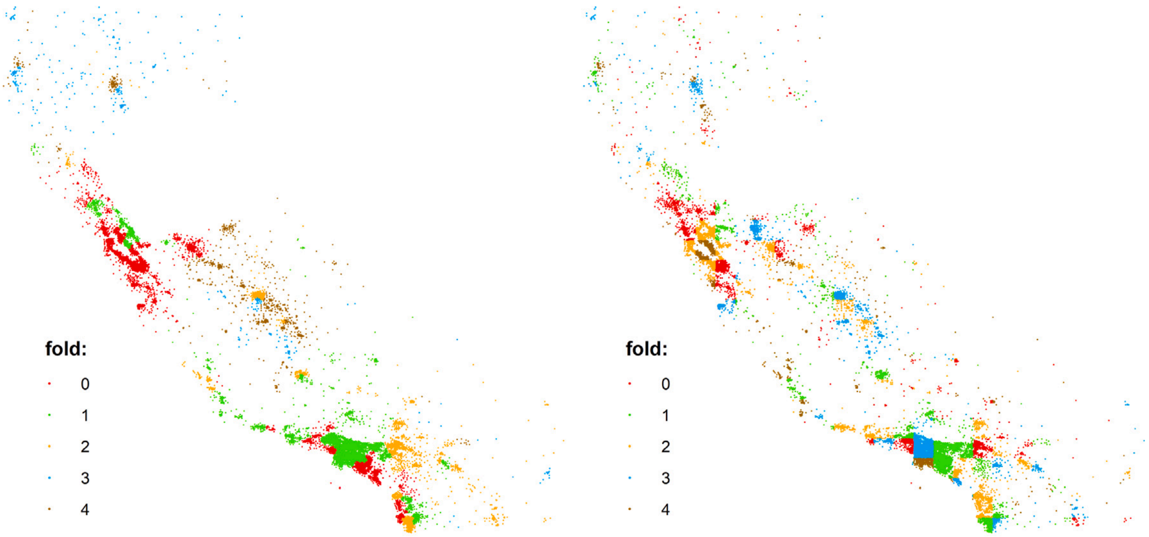
If we first focus on the sub-figures with pink borders in Fig. 10 as the Amazon AGB experiments, we can observe the details of folds' samples in SP-CV and BLK-CV methods. In the sub-figure of Fig. 10(a),



(a) Spatial+ cross-validation, SP-CV

(b) Block cross-validation, BLK-CV

Fig. 10. Examples of folds split in Europe OCS experiments.



(a) Spatial+ cross-validation, SP-CV

(b) Block cross-validation, BLK-CV

Fig. 11. Examples of folds split in California houseprice experiments.

it is noticeable that almost all samples are split into the same fold (three folds in total) by SP-CV. While, in the subfigure of Fig. 10(b), BLK-CV split the same samples into eight different folds, where these samples are located quite close to each other. This suggests that SP-CV could have a better consideration of spatial autocorrelation in the folds splitting process.

For the California houseprice experiments, Fig. 11 shows the generated folds using SP-CV and BLK-CV methods respectively. As shown in Fig. 11(b), the folds generated by the BLK-CV method are randomly distributed, and the square shape of the blocks can be observed at the borders of the folds. Considering that the target variable is house price, this folds split result is meaningless, especially since urban and rural regions are not distinguishable. On the contrary, as shown in Fig. 11(a), SP-CV provides a more reasonable split result, following the spatial patterns of the pre-defined regions shown in Fig. 5(b). For example, in Fig. 11(a), Fresno and Bakersfeld (fold 2, orange points) can be distinguished from the rural areas (fold 3 and fold 4, brown points and blue points). They can also be distinguished from much larger cities, i.e., San Francisco and Los Angeles, which are located in fold 0 and fold 1 (red points and green points). This suggests that SP-CV was able to consider the spatial autocorrelation and the multiple sources of differences in experiments.

5. Conclusion and future research

In geospatial prediction tasks, the available sample data for building a model are usually different from the data in prediction locations. This common problem poses many challenges for model evaluation, making traditional cross-validation (CV) impractical. Current spatial CV methods fail to fully address this challenge as they neglect to holistically account for the various sources of these differences between sample data and prediction locations. Especially, many differences are attributed to the feature space. As a result, the validation subsets generated by spatial CV methods do not always reflect the test set (prediction locations) with obvious differences well.

In this paper, we proposed a new CV method — spatial+ cross-validation (SP-CV). The primary advantage of SP-CV is its ability to split the training and validation subsets by taking into account the various differences present between sample data and prediction locations, including locations, the target variable, and particularly the covariates. Furthermore, SP-CV enhances the consideration of spatial autocorrelation present in the data. SP-CV was compared to traditional RDM-CV and of the most widely used spatial CV, BLK-CV, in a series of experiments using two datasets. According to the experiments simulating the actual cases of strongly clustered samples and extrapolation, the results indicated that SP-CV outperformed these methods by producing the closest evaluations to the reference prediction errors.

At present, we already know that the differences between available sample data and prediction locations will influence the models' performances. For example, in recent years, researchers (Wadoux et al., 2021; de Bruin et al., 2022) found that when the sample data and prediction locations are similar or only slightly different, random CV could provide good evaluation, whereas spatial CV tended to be overly pessimistic. However, when the sample data and prediction locations are significantly different (which is frequently the case in practical geospatial predictions), the performances of random CV and spatial CV are reversed, with random CV being overly optimistic and spatial CV providing more reliable evaluations. Therefore, in the future, further research is needed to understand the varying levels of the differences between sample data and prediction locations, how to quantitatively measure them, and the impact they have on model prediction error magnitude. By effectively utilizing the information and relationships between covariates and sample data, along with the aforementioned future research, it may be possible to develop a "generic" evaluation method for assessing geospatial prediction ML models in various scenarios and applications.

CRediT authorship contribution statement

Yanwen Wang: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Mahdi Khodadadzadeh:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Raúl Zurita-Milla:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision.

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.

Data availability

The datasets and the necessary code to reproduce our results can be downloaded from: https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:278047

Acknowledgments

The corresponding author acknowledges the financial support from the China Scholarship Council, no. 201804910723.

The corresponding author offers his sincerest gratitude to Dr. Ourania Kounadi. Dr. Kounadi gave plenty of help and guidance during the early stages of the corresponding author's Ph.D. candidate studies, as well as the pre-research phase of this study.

References

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., Hinton, G.E., 2021. Neural additive models: Interpretable machine learning with neural nets. In: Advances in Neural Information Processing Systems, NeurIPS 2021, vol. 34, Curran Associates, pp. 4699–4711.

Aguilar, R., Zurita-Milla, R., Izquierdo-Verdiguier, E., de By, R.A., 2018. A cloud-based multi-temporal ensemble classifier to map smallholder farming systems. Remote Sens. 10 (5), 729. http://dx.doi.org/10.3390/RS10050729.

Alqurashi, T., Wang, W., 2019. Clustering ensemble method. Int. J. Mach. Learn. Cybern. 10 (6), 1227–1246. http://dx.doi.org/10.1007/s13042-017-0756-7.

Arabie, P., Hubert, L.J., De Soete, G., Gordon, A.D., 1996. Hierarchical classification. In: Clustering and Classification. World Scientific, pp. 65–121. http://dx.doi.org/10.1142/9789812832153_0003.

Beigaitė, R., Mechenich, M., Žliobaitė, I., 2022. Spatial cross-validation for globally distributed data. In: International Conference on Discovery Science 2022. Springer, Cham, Montpellier, pp. 127–140. http://dx.doi.org/10.1007/978-3-031-18840-4_10.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Brenning, A., 2005. Spatial prediction models for landslide hazards: Review, comparison and evaluation. Nat. Hazards Earth Syst. Sci. 5 (6), 853–862.

Brus, D.J., Kempen, B., Heuvelink, G.B., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62 (3), 394–407. http://dx.doi.org/10.1111/J.1365-2389.2011.01364.X.

Carvalho, M.d.A., Marcato, J., Martins, J.A.C., Zamboni, P., Costa, C.S., Siqueira, H.L., Araújo, M.S., Gonçalves, D.N., Furuya, D.E.G., Osco, L.P., Ramos, A.P.M., Li, J., de Castro, A.A., Gonçalves, W.N., 2022. A deep learning-based mobile application for tree species mapping in RGB images. Int. J. Appl. Earth Obs. Geoinf. 114, 103045. http://dx.doi.org/10.1016/J.JAG.2022.103045.

Chen, G., Wang, Y., Li, S., Cao, W., Ren, H., Knibbs, L.D., Abramson, M.J., Guo, Y., 2018. Spatiotemporal patterns of PM10 concentrations over China during 2005–2016: A satellite-based estimation using the random forests approach. Environ. Pollut. 242, 605–613. http://dx.doi.org/10.1016/J.ENVPOL.2018.07.012.

Cheng, Y., Tjaden, N.B., Jaeschke, A., Lühken, R., Ziegler, U., Thomas, S.M., Beierkuhnlein, C., 2018. Evaluating the risk for Usutu virus circulation in Europe: Comparison of environmental niche models and epidemiological models. Int. J. Health Geogr. 17 (1), 1–14. http://dx.doi.org/10.1186/s12942-018-0155-7.

Dang, A.T.N., Nandy, S., Srinet, R., Luong, N.V., Ghosh, S., Senthil Kumar, A., 2019. Forest aboveground biomass estimation using machine learning regression algorithm in Yok Don National Park, Vietnam. Ecol. Inform. 50, 24–32. http://dx.doi.org/10.1016/J.ECOINF.2018.12.010.

de Bruin, S., Brus, D.J., Heuvelink, G.B., van Ebbenhorst Tengbergen, T., Wadoux, A.M.-C., 2022. Dealing with clustered samples for assessing map accuracy by cross-validation. Ecol. Inform. 69, 101665. http://dx.doi.org/10.1016/J.ECOINF.2022.101665.

Efron, B., 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. J. Amer. Statist. Assoc. 78 (382), 316–331. http://dx.doi.org/10.1080/01621459.1983.10477973.

Fern, X.Z., Brodley, C.E., 2004. Solving cluster ensemble problems by bipartite graph partitioning. In: ICML '04: Proceedings of the Twenty-First International Conference on Machine Learning. Association for Computing Machinery, pp. 281–288. http://dx.doi.org/10.1145/1015330.1015414.

Filippi, P., Jones, E.J., Wimalathunge, N.S., Somarathna, P.D.S.N., Pozza, L.E., Ugbaje, S.U., Jephcott, T.G., Paterson, S.E., Whelan, B.M., Bishop, T.F.A., 2019. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precis. Agric. 20, 1015–1029. http://dx.doi.org/10.1007/s11119-018-09628-4.

Gao, B., Stein, A., Wang, J., 2022. A two-point machine learning method for the spatial prediction of soil pollution. Int. J. Appl. Earth Obs. Geoinf. 108, 102742. http://dx.doi.org/10.1016/J.JAG.2022.102742.

Garcia-Marti, I., Zurita-Milla, R., Harms, M.G., Swart, A., 2018. Using volunteered observations to map human exposure to ticks. Sci. Rep. 8 (1), 15435. http://dx.doi.org/10.1038/s41598-018-33900-2.

Garcia-Martí, I., Zurita-Milla, R., Swart, A., van den Wijngaard, K.C., van Vliet, A.J., Bennema, S., Harms, M., 2017. Identifying environmental and human factors associated with tick bites using volunteered reports and frequent pattern mining. Trans. GIS. 21 (2), 277–299. http://dx.doi.org/10.1111/tgis.12211.

Gasch, C.K., Hengl, T., Gräler, B., Meyer, H., Magney, T.S., Brown, D.J., 2015. Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D + T: The cook agronomy farm data set. Spat. Stat. 14, 70–90. http://dx.doi.org/10.1016/j.spasta.2015.04.001.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. Plos One. 10 (6), e0125814. http://dx.doi.org/10.1371/journal.pone.0125814.

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

Hooker, J., Duveiller, G., Cescatti, A., 2018. A global dataset of air temperature derived from satellite remote sensing and weather stations. Sci. Data. 5 (1), 1–11. http://dx.doi.org/10.1038/SDATA.2018.246.

Huang, Z., 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discov. 2 (3), 283–304. http://dx.doi.org/10.1023/A:1009769707641.

Khodadadzadeh, M., Gloaguen, R., 2019. Upscaling high-resolution mineralogical analyses to estimate mineral abundances in drill core hyperspectral data. In: International Geoscience and Remote Sensing Symposium. IGARSS 2019, Institute of Electrical and Electronics Engineers Inc., pp. 1845–1848. http://dx.doi.org/10.1109/IGARSS.2019.8898441.

Kollert, A., Bremer, M., Löw, M., Rutzinger, M., 2021. Exploring the potential of land surface phenology and seasonal cloud free composites of one year of Sentinel-2 imagery for tree species mapping in a mountainous region. Int. J. Appl. Earth Obs. Geoinf. 94, 102208. http://dx.doi.org/10.1016/J.JAG.2020.102208.

Kounadi, O., Ristea, A., Araujo, A., Leitner, M., 2020. A systematic review on spatial crime forecasting. Crime Sci. 9 (1), 7. http://dx.doi.org/10.1186/s40163-020-00116-7.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., Bretagnolle, V., 2014. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. Glob. Ecol. Biogeogr. 23 (7), 811–820. http://dx.doi.org/10.1111/geb.12161.

Li, T., Shen, H., Zeng, C., Yuan, Q., 2020. A validation approach considering the uneven distribution of ground stations for satellite-based PM2.5 Estimation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 1312–1321. http://dx.doi.org/10.1109/JSTARS.2020.2977668.

Lyons, M.B., Keith, D.A., Phinn, S.R., Mason, T.J., Elith, J., 2018. A comparison of resampling methods for remote sensing classification and accuracy assessment. Remote Sens. Environ. 208, 145–153. http://dx.doi.org/10.1016/j.rse.2018.02.026.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12 (9), 1620–1633. http://dx.doi.org/10.1111/2041-210X.13650.

Meyer, H., Pebesma, E., 2022. Machine learning-based global maps of ecological variables and the challenge of assessing them. Nature Commun. 13 (1), 1–4. http://dx.doi.org/10.1038/s41467-022-29838-9.

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. Model. 411, 108815. http://dx.doi.org/10.1016/j.ecolmodel.2019.108815.

Milà, C., Mateu, J., Pebesma, .E., Meyer, H., 2022. Nearest neighbour distance matching Leave-One-Out Cross-Validation for map validation. Methods Ecol. Evol. 13 (6), 1304–1316. http://dx.doi.org/10.1111/2041-210X.13851.

Murtagh, F., 1983. A survey of recent advances in hierarchical clustering algorithms. Comput. J. 26 (4), 354–359. http://dx.doi.org/10.1093/comjnl/26.4.354.

Nesha, M.K., Hussin, Y.A., van Leeuwen, L.M., Sulistioadi, Y.B., 2020. Modeling and mapping aboveground biomass of the restored mangroves using ALOS-2 PALSAR-2 in East Kalimantan, Indonesia. Int. J. Appl. Earth Obs. Geoinf. 91, 102158. http://dx.doi.org/10.1016/J.JAG.2020.102158.

Oliveira, M., Torgo, L., Costa, V.S., 2021. Evaluation procedures for forecasting with spatiotemporal data. Mathematics 9 (6), 703–718. http://dx.doi.org/10.3390/math9060691.

Pace, R.K., Barry, R., 1997. Sparse spatial autoregressions. Statist. Probab. Lett. 33 (3), 291–297. http://dx.doi.org/10.1016/s0167-7152(96)00140-x.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Commun. 11, 4540. http://dx.doi.org/10.1038/s41467-020-18321-y.

Pohjankukka, J., Pahikkala, T., Nevalainen, P., Heikkonen, J., 2017. Estimating the prediction performance of spatial models via spatial k-fold cross validation. Int. J. Geogr. Inf. Sci. 31 (10), 2001–2019. http://dx.doi.org/10.1080/13658816.2017.1346255.

Pourghasemi, H.R., Pouyan, S., Heidari, B., Farajzadeh, Z., Fallah Shamsi, S.R., Babaei, S., Khosravi, R., Etemadi, M., Ghanbarian, G., Farhadi, A., Safaeian, R., Heidari, Z., Tarazkar, M.H., Tiefenbacher, J.P., Azmi, A., Sadeghian, F., 2020. Spatial modeling, risk mapping, change detection, and outbreak trend analysis of coronavirus (COVID-19) in Iran (days between February 19 and June 14, 2020). Int. J. Infect. Dis. 98, 90–108. http://dx.doi.org/10.1016/j.ijid.2020.06.058.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 40 (8), 913–929. http://dx.doi.org/10.1111/ecog.02881.

Sarafian, R., Kloog, I., Sarafian, E., Hough, I., Rosenblatt, J.D., 2021. A domain adaptation approach for performance estimation of spatial predictions. IEEE Trans. Geosci. Remote Sens. 59 (6), 5197–5205. http://dx.doi.org/10.1109/TGRS.2020.3012575.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., Brenning, A., 2019. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecol. Model. 406, 109–120. http://dx.doi.org/10.1016/J.ECOLMODEL.2019.06.002.

Strehl, A., Ghosh, J., 2002. Cluster ensembles-A knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3, 583–617.

Valavi, R., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G., 2019. BlockCV : An R package for generating spatially or environmentally separated folds for k -fold cross-validation of species distribution models. Methods Ecol. Evol. 10 (2), 225–232. http://dx.doi.org/10.1111/2041-210X.13107.

Wadoux, A.M., Heuvelink, G.B., de Bruin, S., Brus, D.J., 2021. Spatial cross-validation is not the right way to evaluate map accuracy. Ecol. Model. 457, 109692. http://dx.doi.org/10.1016/J.ECOLMODEL.2021.109692.

Wei, R., Ye, C., Sui, T., Ge, Y., Li, Y., Li, J., 2022. Combining spatial response features and machine learning classifiers for landslide susceptibility mapping. Int. J. Appl. Earth Obs. Geoinf. 107, 102681. http://dx.doi.org/10.1016/J.JAG.2022.102681.

Wiens, T.S., Dale, B.C., Boyce, M.S., Kershaw, G.P., 2008. Three way k-fold cross-validation of resource selection functions. Ecol. Model. 212 (3–4), 244–255. http://dx.doi.org/10.1016/j.ecolmodel.2007.10.005.

Xiao, Q., Chang, H.H., Geng, G., Liu, Y., 2018. An ensemble machine-learning model to predict historical PM2.5 concentrations in China from satellite data. Environ. Sci. Technol. 52 (22), 13260–13269. http://dx.doi.org/10.1021/acs.est.8b02917.

Xu, C., Hystad, P., Chen, R., Van Den Hoek, J., Hutchinson, R.A., Hankey, S., Kennedy, R., 2021. Application of training data affects success in broad-scale local climate zone mapping. Int. J. Appl. Earth Obs. Geoinf. 103, 102482. http://dx.doi.org/10.1016/J.JAG.2021.102482.

Zhu, A.X., Liu, J., Du, F., Zhang, S.J., Qin, C.Z., Burt, J., Behrens, T., Scholten, T., 2015. Predictive soil mapping with limited sample data. Eur. J. Soil Sci. 66 (3), 535–547. http://dx.doi.org/10.1111/EJSS.12244.