Mushagalusa *et al.*
Environmental Systems Research (2024) 13:23
https://doi.org/10.1186/s40068-024-00352-9

Environmental Systems Research

RESEARCH

Open Access

Random forest and spatial cross-validation performance in predicting species abundance distributions

Ciza Arsène Mushagalusa[1,2*], Adandé Belarmain Fandohan[1,3] and Romain Glèlè Kakaï[1]

Abstract

Random forests (RF) have been widely used to predict spatial variables. Several studies have shown that spatial cross-validation (CV) methods consistently cause RF to yield larger prediction errors compared to standard CV methods. This study examined the impact of species characteristics and data features on the performance of the standard RF and spatial CV approaches for predicting species abundance distribution. It compared the standard 5-fold CV, design-based validation, and three different spatial CV methods, such as spatial buffering, environmental blocking, and spatial blocking. Validation samples were randomly selected for design-based validation without replacement. We evaluated their predictive performance (accuracy and discrimination metrics) using artificial species abundance data generated by a linear function of a constant term (β_0) and a random error term following a zero-mean Gaussian process with a covariance matrix determined by an exponential correlation function. The model was tuned over multiple simulations to consider different mean levels of species abundance, spatial autocorrelation variation, and species detection probability. Here we found that the standard RF had poor predictive performance when spatial autocorrelation was high and the species probability of detection was low. Design-based validation and standard K-fold CV were found to be the most effective strategies for evaluating RF performance compared to spatial CV methods, even in the presence of high spatial autocorrelation and imperfect detection for random samples. For weakly or moderately clustered samples, they yielded good modelling efficiency but overestimated RF's predictive power, while they overestimated modelling efficiency, predictive power, and accuracy for strongly clustered samples with high spatial autocorrelation. Globally, the checkerboard pattern in the allocation of blocks to folds in blocked spatial CV was found to be the most effective CV approach for clustered samples, whatever the degree of clustering, spatial autocorrelation, or species abundance class. The checkerboard pattern in spatial CV was found to be the best method for random or systematic samples with spatial autocorrelation, but less effective than non-spatial CV approaches. Failing to take data features into account when validating models can lead to unrealistic predictions of species abundance and related parameters and, therefore, incorrect interpretations of patterns and conclusions. Further research should explore the benefits of using blocked spatial K-fold CV with checkerboard assignment of blocks to folds for clustered samples with high spatial autocorrelation.

Keywords Machine learning, Species distribution modelling, Imperfect detection, Spatial blocking

*Correspondence:
Ciza Arsène Mushagalusa
shaga.ciza@gmail.com
Full list of author information is available at the end of the article

Introduction

Measuring abundance and distribution is a primary goal of ecologists, conservationists and managers (Royle and Dorazio 2009). Understanding population dynamics, community structure and the effects of management depends on the knowledge of species abundance and distribution, as well as related parameters, which can shed light on less obvious aspects of a community (McGill et al. 2007; Kellner and Swihart 2014; Baldridge et al. 2016; Su 2018). Abundance trends can serve as an early indicator of population collapse (Clements et al. 2017; Ceballos et al. 2020; Hastings et al. 2020). Therefore, to better inform spatial conservation planning, one needs to improve species monitoring and abundance prediction (Pauly and Froese 2010; Mi et al. 2017).

Recent developments in big data analysis have led to the development of several high-computational statistical methods with the potential for ecological data mining, enabling the use of machine learning (ML) techniques, to make reliable predictions from noisy and incomplete datasets (Cutler et al. 2007). One of the most popular ML techniques is the Random Forest (RF) (Breiman 2001; Saha et al. 2023). RF is already widely used in ecological research and outperforms most commonly used methods (Lawler et al. 2006; Prasad et al. 2006; Cutler et al. 2007; Martín et al. 2021). Unlike many conventional statistical analysis techniques, RF makes no distributional assumptions and can handle complex and nonlinear relationships between abundance and its environmental factors. It is also able to handle scenarios where the number of predictors significantly exceeds the number of observations (Kuhn and Johnson 2013; Scornet 2016; Zurell et al. 2016; Zhang et al. 2020).

Temporal and spatial processes often cause autocorrelation in ecological data, leading to model errors and overfitting (Legendre and Fortin 1989; Miller et al. 2007; Roberts et al. 2017). The independence problem is theoretically solved by models that take into account the dependence structure, and this should allow model fit assessment and model selection using conventional parametric techniques. In practice, however, due to specification errors, structural overfitting, and other problems, parametric model evaluations may not perform as well as they should (Dormann et al. 2007; Miller et al. 2007). Reliable non-parametric approaches are required for ML models performance assessment, which represents a critical stage in the modelling process (Roberts et al. 2017).

In ecology, this typically involves the measurement of how well predictions match field observations (Franklin 2010; Peterson et al. 2012; Guisan et al. 2017). Ideally, prediction errors and model validation should be computed using independent data. However, in most cases, such independent data are not available (Araújo et al. 2005; Franklin 2010; Radosavljevic and Anderson 2014). Therefore, cross-validation (CV) is used to estimate the predicted error on a single dataset (Hastie et al. 2009).

Many studies show that the dependence structure can compromise the independence of the validation dataset, leading to overly optimistic estimates of prediction error (Wenger and Olden 2012; Bahn and McGill 2013; Roberts et al. 2017; Meyer et al. 2019; Ploton et al. 2020). Spatial dependence in modelling is a complex issue due to factors such as model misspecification, omission of important covariates and processes such as localised dispersal or social behaviour (Fletcher and Fortin 2018). It is important in ecological modelling and conservation because predictive models often rely on random and independent observations, violating the relationship between distance and similarity in geography and ecological theory (Tobler 1979; Legendre and Fortin 1989; Miller et al. 2007). Incorporating spatial dependence into models may enhance understanding of effects of different explanatory variables, resulting in better statistical inference and subsequent ecological interpretation of patterns observed (Miller et al. 2007; Fletcher and Fortin 2018).

In recent years, there has been a debate about the best way to separate training and validation datasets (Bahn and McGill 2013; Radosavljevic and Anderson 2014; Wenger and Olden 2012). Stone (1974), argued that random data partitioning does not generate independent validation datasets when dependence structures are present in the data because calibration points are not statistically independent from validation points. Consequently, many CV techniques have been developed and are now used in ecological studies to obtain supposedly unbiased error and parameter estimates (Shao 1993; Kohavi 1995; Rykiel 1996).

To correct for the bias in error estimates produced by these approaches, adjustments based on spatially separated training and testing datasets have been proposed to assess whether the model performs as well at closer as at more distant sites (Telford and Birks 2009; Amesbury et al. 2013). Their main focus is to increase the independence of CV by appropriately partitioning the data into blocks and accounting for the dependence structure to prevent overfitting (Dormann et al. 2007; Trachsel and Telford 2016). However, only a small number of studies have clearly shown that the estimates obtained from blocked CV are quite close to the 'true' error that would be expected for a data set that is truly independent (Roberts et al. 2017).

Numerous studies have shown that spatial CV approaches consistently produce higher prediction errors compared to the standard random CV (Araújo et al. 2005; Veloz 2009; Arlot and Celisse 2010; Lieske

and Bender 2011; Roberts and Hamann 2012; Wenger and Olden 2012; Bahn and McGill 2013; Radosavljevic and Anderson 2014; Wadoux et al. 2021). Although spatial block CV addresses spatial autocorrelation, a new validation problem may arise if block structures follow environmental gradients, which could prevent the use of large regions of predictor space (Snee 1977). As a result, to predict the hold-out data, the model must extrapolate beyond the ranges or into novel combinations of predictor values from those contained in the training folds (Zurell et al. 2012).

Wadoux et al. (2021) provided evidence that estimates based on both non-spatial and spatial CV can be biased, and spatial CV does not improve the predictive performance. They stated that recent research on spatial CV approaches in ecology has led to a misunderstanding of statistical validation in a spatial context. When validating via design-based inference, validation sites may be physically close to calibration sites, so they recommended sticking to rigorous statistical validation procedures using probability sampling and design-based inference. The design-based approach assumes that validation samples used to estimates performance metrics are based on classical sampling theory (Cochran 1977; Gruijter et al. 2006; Gregoire and Valentine 2007). Whether or not two randomly selected locations are drawn from a spatially structured population, prediction errors arise independently at each location (Gregoire and Valentine 2007; Brus 2021). Recent studies acknowledge that situations where samples are highly clustered, can still pose challenges for evaluating models using non-spatial CV methods (de Bruin et al. 2022).

However, many studies that have recently investigated the performance of spatial CV methods used random assignment of blocks to folds in modelling above-ground forest biomass (Brenning 2005; Lyons et al. 2018; Valavi et al. 2019; de Bruin et al. 2022; Wang et al. 2023). This research investigates how different methods of assigning blocks to folds may affect blocked spatial CV methods. Blocks to folds are one of the essential processes in species modelling, as species data are rarely evenly distributed across the landscape (Valavi et al. 2019). Considering species and data features (abundance class, imperfect detection, sampling method, sample size and spatial autocorrelation), this article provides practical guidance on using spatial CV to evaluate the predictive effectiveness, power and accuracy of the RF in ecological research. The standard 5-fold CV, the design-based validation, and three different blocking strategies: spatial blocking, spatial buffering, and environmental blocking were all tested to achieve this goal. Random, systematic, and checkerboard pattern blocks-to-folds allocation strategies were tested.

Methods

Data simulation

We used artificial species data to assess the performance of the standard RF and spatial CV approaches in predicting species abundance in a geographical area. The advantage of generating artificial data is that we have full knowledge and control over the factors under investigation, whereas we are often unfamiliar with the empirical distributions of the data when comparing models with real data (Hirzel et al. 2001; Austin et al. 2006; Meynard and Quinn 2007). We generated datasets as proposed by Guélat and Kéry (2018) to assess the impact of blocking strategies on the effectiveness of the random forest algorithm in predicting species abundance in ecology, taking into account imperfect detection and spatial autocorrelation. We generated 250 datasets within an 80*80 cell landscape of 6400 cells for each case. The dependent variable was the result of two random processes, while the independent variable was generated using a normal distribution. The following models were used to generate the datasets:

$$N_i \sim \text{Poisson}(\lambda_i), \tag{1}$$

$$\log(\lambda_i) = \beta_0 + \beta x_i + \gamma \rho_i, \tag{2}$$

$$\rho_i \sim \text{MVN}\left(0, \sigma^2 e^{-\theta d_{ij}}\right), \tag{3}$$

$$C_{it} \sim \text{Binomial}(N_i, p). \tag{4}$$

At each location, i, the first random process returns the true latent abundance N_i, λ is the expected abundance values. β_0 represents a constant term contributing to abundance, β is the growth rate coefficient of the exponential function, x_i is a continuous covariate, ρ_i is a spatially autocorrelated random effect generated by an exponential correlation function and γ represents the spatial autocorrelation's variation strength. We used $\beta = 0.8$ for all generated species abundance, $\beta_0 = -0.6$ and $\gamma = 0.5$ for rare species with low spatial autocorrelation; $\beta_0 = -1.5$ and $\gamma = 1.5$ for rare species with high spatial autocorrelation; $\beta_0 = 1.8$ and $\gamma = 0.5$ for common species with low spatial autocorrelation; $\beta_0 = 0.8$ and $\gamma = 1.5$ for common species with high spatial autocorrelation (Guélat and Kéry 2018).

The strength of pairwise correlation in the landscape is defined by the covariance matrix of the multivariate normal (MVN) distribution and the distance d_{ij} between

sites i and j. In addition, σ^2 denotes the spatial variance and θ is the scale parameter that determines the distance-dependent decay of spatial autocorrelation in expected species abundance. In other words, the distance-based strength of pairwise correlations in the landscape is determined by the covariance matrix of the multivariate normal distribution. The second random process produces the observed data, the counts C_{it} at site i during visit t, and is linked to the outcome of the first (i.e. conditional on N_i). This second process is a description of the error in the measurement of abundance and is governed by the probability of detection p per visit. We used $p = 0.3$ and $p = 0.8$ respectively for low and high species probability of detection.

We have generated a total of 28 different categories of species (cases). These have been grouped into three scenarios. Table 1 shows the 28 different types of abundance data that were generated.

Data analysis

We used the standard RF to test the influence of three sampling methods on the predictive performance of five spatial CV and the standard 5-fold CV strategies in modelling species abundance distribution accounting for species features (abundance class and detection probability) and data features (sample size and spatial autocorrelation variation strength).

Random forest

RF is a data-driven statistical technique, primarily used for classification or regression, which improves prediction accuracy by combining a large number of decision trees (Breiman 2001; Prasad et al. 2006; Biau and Scornet 2016). RF focuses on iterative training of the algorithm rather than formulating a statistical model, which simplifies its mathematical formulation (Hengl et al. 2018). Its algorithm uses two powerful techniques: random subspace selection at each split and bagging of unpruned decision tree learners (Breiman et al. 1984; Breiman 1996; Amit and Geman 1997; Ho 1998; Dietterich 2004). The RF decision rule uses averaging for regression or a majority vote to classify the results produced from separate trees (Biau 2012).

In regression, RF predictions ($\hat{\theta}$) are obtained by averaging results from a given number (B) of individual decision trees (t_b^*) based on generated bootstrap samples, such as (Breiman et al. 1984; Breiman 2001; Prasad et al. 2006; Biau and Scornet 2016; Hengl et al. 2018):

$$\hat{\theta}^B(x) = \frac{1}{B}\sum_{b=1}^{B} t_b^*(x), \tag{5}$$

We used the fast implementation of the RF (*ranger*) function (Wright and Ziegler 2017) in *R version 4.2.2* (R Core Team 2022) using default parameters for regression to predict observed abundance data (dependent variable) from a randomly generated standard normal distribution (independent variable). The focus of this work, however, is on the evaluation of blocking strategies in spatial CV approaches.

Explored samples

Sampling design is one of the most important factors determining the limits of results and interpretation of analyses in ecological research (Greig-Smith 1983; Kenkel et al. 1989; Goedickemeier et al. 1997). For spatially structured scenarios, we compared three common sampling techniques (Simple random sampling, systematic sampling, and two-stage cluster sampling) to better understand their influence on spatial CV methods.

- **Simple random sampling**

  The basic form of probability sampling is simple random sampling (SRS). In this sampling method, the probability of selecting any sample of a specific size remains unchanged for all possible samples. Therefore, each observation unit or individual in the population has the same probability, p, of being included

Table 1 Generated species abundance characteristics

Scenario	Spatial autocorrelation	Variation strength of autocorrelation	Probability of detection	Abundance class	Sample size (n)	Number of cases
1	No	$\gamma = 0$	$p = 1$	Rare, Common	Small (n=200), Large (n=1000)	4
2	Yes	$\gamma = 0.5$, $\gamma = 1.5$	$p = 1$	Rare, Common	Small, Large	8
3	Yes	$\gamma = 0.5$, $\gamma = 1.5$	High ($p = 0.8$), Low ($p = 0.3$)	Rare, Common	Small, Large	16

in the sample, such as ($p = 1/N$), where N is the population size (West 2016). Despite its simplicity, it can be an effective sampling method under the right circumstances and serve as a theoretical basis for more complex sampling techniques (Yang and Laven 2021).

However, the SRS can be costly and sometimes is not practical because it demands that all items be identified and named before the sampling. Additionally, a basic SRS design may provide samples that are dispersed over a wide geographic area since it gives each potential sample of n units an equal chance of being chosen. However, implementing such a sample's geographic distribution would be exceedingly expensive. In addition, subdomains are probably represented in the sample in proportion to the frequency they are found in the population. While this would be advantageous for some surveys, it would be problematic for those whose interest is centred on subgroups that comprise a small fraction of the population (Levy and Lemeshow 2013).

- **Systematic sampling** Systematic sampling is a probability sampling where every n^{th} case following a random start is chosen. Let N be the population size and n the sample size required. In the case of systematic sampling, the sampling interval, represented by k, is defined first: $k = \frac{N}{n}$, where k is the number of elements in the population to be skipped before moving on to the next. Then, randomly choose a starting number (r) between 1 and k. From r, select every k^{th} element until the required sample size is obtained. The sample is then given by: $\text{Sample} = r, r + k, r + 2k, \ldots, r + (n - 1)k,$ where r is the random starting element and k is the sampling interval. Systematic sampling is commonly used in practice because it is simple to implement and can be taught to people with minimal experience in survey methodology. However, in reality, systematic sampling is perhaps the most extensively used method, alone or in conjunction with another method.
- **Two-stage cluster sampling**
With this sampling technique, individuals from a limited population are grouped into clusters, which are bigger sampling units. A cluster sampling technique sampled a population of N clusters to create a set of (n) clusters. Typically, these clusters are referred to as primary sampling units, while the individuals inside each cluster are referred to as secondary sample units (Yang and Laven 2021). All secondary sampling units may be measured or observed within the original sampling units (one-stage cluster sampling), or the secondary sampling units may be sampled using SRS (two-stage cluster sampling). The chosen individuals then form a sample of the finite population within the chosen clusters (Cochran 1977).
The two-stage cluster sampling can be described as follows:

– A sample of primary sampling units (clusters) is selected.
– Select n clusters from the population, denoted $C_1, C_2, ..., C_n$.
– Within each cluster selected, sample individuals:
– For each selected cluster C_i, let m_i be the number of individuals in the cluster C_i.
– Select a random sample of m_i individuals from the cluster C_i, denoted as $x_{11}, x_{12}, \ldots, x_{1m_i}$.

The overall sample from the population consists of the individuals selected in the second stage, namely $x_{11}, x_{12}, \ldots, x_{1m_1}, x_{21}, x_{22}, \ldots, x_{2m_2}, \ldots, x_{n1}, x_{n2}, \ldots, x_{nm_n}$.

The two most fundamental reasons for the widespread use of cluster sampling in large-scale sample research are feasibility and economics (Levy and Lemeshow 2013). This study used a specific case where K-means spatial clustering was performed using grid coordinates. First, from a grid of 6400 units, 25 clusters were formed. We then randomly selected 20 clusters for weakly clustered samples, 10 clusters for moderately clustered samples and 5 clusters for strongly clustered samples. We collected 10 and 50 units for small and large sample sizes, respectively, in each cluster. Figure 1 shows the sampled sites within each of the sampling designs.

Model validation

We used a design-based validation approach, the standard 5-fold CV and three spatial CV approaches to assess the RF predictive performance. Figure 2 is a schematic representation of the different CV strategies under consideration in this study using a random sample.

Design-based probability sampling This study used simple random sampling, where each unit in a population has an equal probability of being selected in a sample of a given size (number of units). When sampling is probability-based, a design-based estimation can be used to estimate population validation parameters (Stehman 1999;

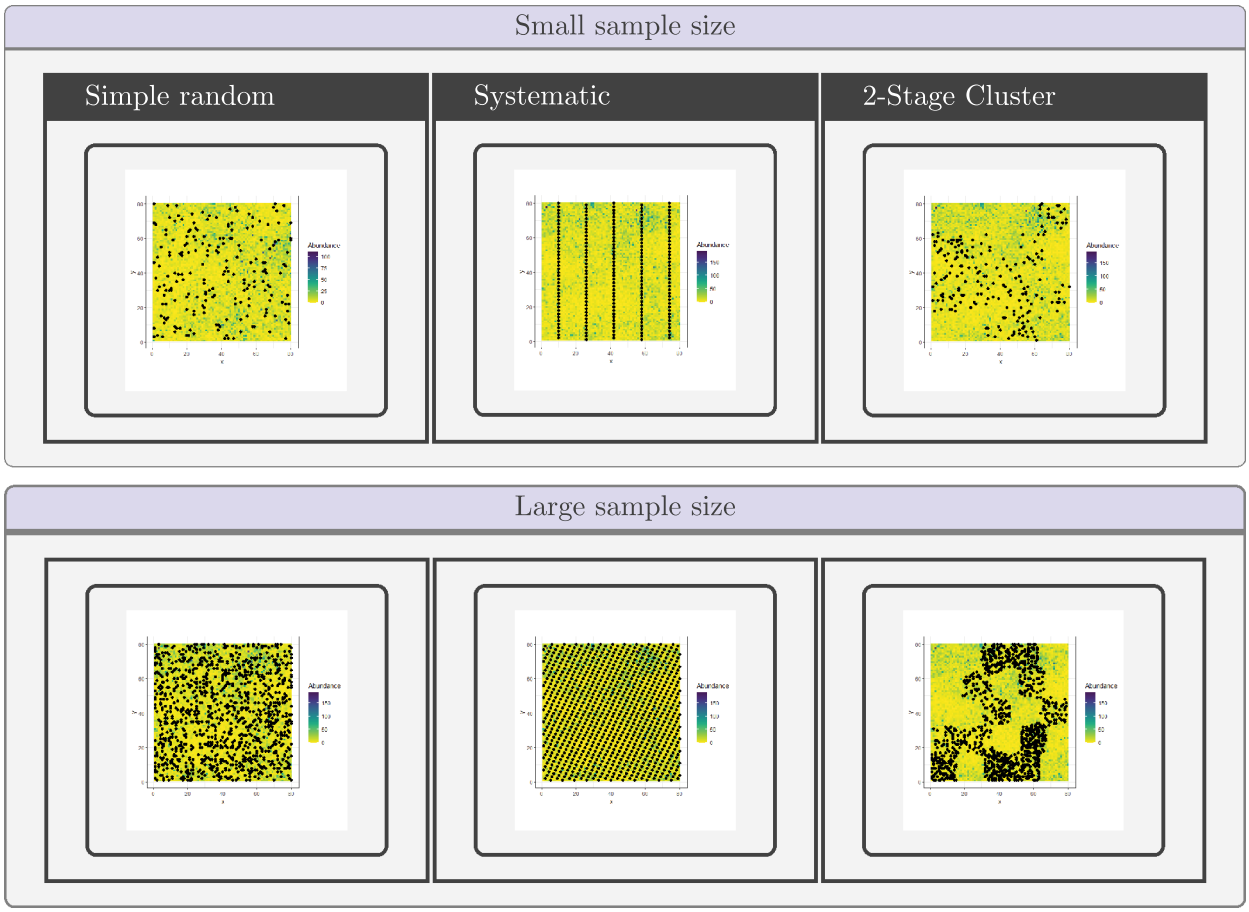


Fig. 1 Location of sample sites: Black dots represent sample sites; Coordinates x and y are measured in degrees, representing longitude and latitude

Stehman and Foody 2009; Brus et al. 2011). The expected performance and the design-based performance for each dataset were obtained by randomly selecting training (75%) and test (25%) samples using the entire population (6400 grids) and sampled grids (sites), respectively. First, the training set is used for model fitting, and the fitted model is used for prediction in the holdout set (James et al. 2013).

Cross-validation methods Cross-validation involves estimating the test's error rate by excluding a subset of training observations from the calibration process and applying the test to the excluded observations. Four approaches were used to cross-validate the model's performance.

- **Standard K-fold cross-validation**: In K-fold cross-validation, the data set is randomly divided into K-folds or groups of approximately equal size, where each group is used as the test set and the remaining groups ($K-1$) are used as the training set (Brownlee 2019), (Wadoux et al. 2021). This study used the five-fold CV. It has been empirically shown that $K = 5$ or 10 provides estimates of the test error rate that do not suffer from excessive bias or variance (James et al. 2013). The standard K-fold cross-validation does not account for spatial autocorrelation.
- **Spatial K-fold cross-validation**: The spatial K-fold CV differs from the random K-fold CV in the way in which the observations are divided into geographically structured sets (blocks). To ensure independence between CV folds, the aim is to group observations into spatially homogeneous clusters with sizes larger than the range of autocorrelation in the data (Valavi et al. 2019; Ploton et al. 2020; Wadoux et al. 2021). We assumed that each block size was 1.25 * the range of the variogram generated over the whole data set in the case of spatial autocorrelation, and considered the block size equal to the variogram range in the absence of spatial autocorrelation. Because species data are rarely evenly distributed across land-

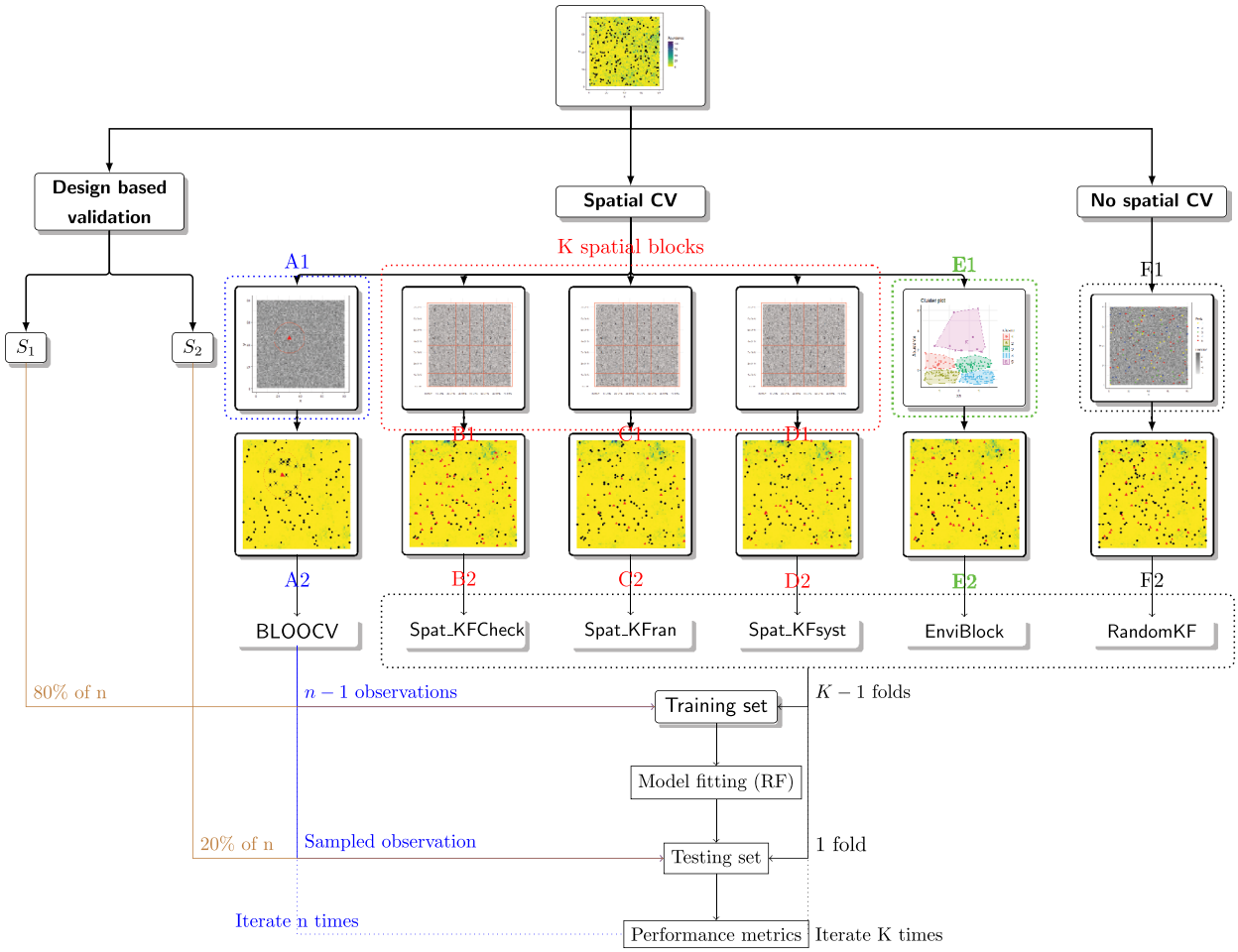


Fig. 2 Schematic representation of design based validation and the different CV methods. *RandomKF* Random K-folds, *BLOOCV* Buffered leave-one-out, *Spat_KFcheck* Checkboard spatial K-folds, *Spat_KFran* Random spatial K-folds, *Spat_KFsyst* Systematic spatial K-folds and *EnviBlock* Environmental blocking

scapes, one of the most important processes in modelling species is assigning blocks to folds (Valavi et al. 2019). We tested three techniques for the assignment of blocks to folds:

- *Random*: blocks are randomly assigned to folds,
- *Systematic*: blocks are numbered and sequentially assigned to folds,
- *In a checkerboard pattern*: Although the checkerboard design involves only two folds, it effectively ensures that there are no adjacent blocks in any one fold.

In all spatial blocking scenarios, all data in the test folds are excluded from the training data sets. We used square blocks because these are the most common block form (Brenning 2005; Lyons et al. 2018; Valavi et al. 2019; Wang et al. 2023).

- **Buffered leave-one-out cross-validation**: The buffering technique creates spatially distinct training and testing folds, considering a circular buffering of a predetermined radius around each observation point (Rest et al. 2014). During model calibration, observations that are close to a validation point by the defined radius, depending on the distance, are not taken into account (Ploton et al. 2020). The typical assumption is that this radius is larger than the range of a variogram that is performed on the whole data set or that is calculated on the residuals after the model calibration and the predictions (Ploton et al. 2020; Wadoux et al. 2021). In this study, r is equal to the size of the block as defined in the spatial K-fold cross-validation. In this method, only one observation is used for the validation

set in each model run. The remaining $n-1$ observations form the training set. This method is inspired by the 'leave-one-out' scheme of cross-validation. In this case, each point that is left out corresponds to a fold or a block (Valavi et al. 2019).

- **Environmental Blocking**: Based on the input variables, we used a K-means clustering technique to determine the number of clusters in environmental space, defining clusters of similar environmental conditions (Hartigan and Wong 1979). Within each cluster, one-fold was identified (Valavi et al. 2019).

With, S_1: random sample used to train the model, S_2: random sample used to test the model, n: sample size, A1: sample 1 observation and exclude neighbours within a radius r; B1, C1 and D1: Split the observation into K spatial blocks of length greater than the range of the variogram. The block length is equal to the radius in BLOOCV. B, C and D represent checkerboard, random and systematic patterns in the assignment of blocks to folds. E1: Clustering of observations into K groups (K-means clustering in this study). Since the 5-fold CV requires at least 5 clusters, datasets with less than 5 clusters were not considered in this study. F1: Splitting of observations into K random folds ($K=5$). A2, B2, C2, D2, E2 and F2 show how observations and/or blocks can be allocated in the training and test datasets. For blocked K-folds CV, $K-1$ folds are assigned to the training set and one fold is assigned to the test set. For BLOOCV, the selected observation is assigned to the test dataset and the remaining observations are assigned to the training dataset, except observations within the exclusion radius. Training and test observations are represented by black and red dots. Crossed black dots are observations within BLOOCV's exclusion buffer.

Predictive performance assessment

Performance metrics were averaged (mean) over 250 simulated data sets for each case to limit the influence of randomness associated with the different CV methods. Measures of accuracy and discrimination were used to assess the impact of the sampling strategy and the CV method on RF predictive performance. The CV approach that yields performance measures closest to the design-based method used on the whole population is the best. Results for all CV's across the 250 replicates are summarised in box plots.

- **Accuracy:** The bias and root mean square error (RMSE) were calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2}, \tag{6}$$

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i), \tag{7}$$

where O_i and P_i refer to observed and predicted species abundance at sampled locations i, respectively, and n is the number of sampled locations (number of observations).

- **Discrimination:** The mean squared Spearman rank correlation coefficient $R^2 = \rho_s^2$ between the predicted and observed abundance of species at the sampled sites and the modelling efficiency coefficient (MEC) of Nash and Sutcliffe (1970) were used.
The n raw observed O_i and predicted P_i are converted to ranks $R(O_i)$, $R(P_i)$ and (ρ_s) is defined as their Pearson correlation coefficient.

$$\rho_s = \frac{\text{cov}(R(O), R(P))}{\sigma_{R(O)}\sigma_{R(P)}}, \tag{8}$$

where $\text{cov}(R(O), R(P))$ represents the covariance between the rank variables of O and P, $\sigma_{R(O)}$ and $\sigma_{R(P)}$ are standard deviations of the rank variable of O and P.
The MEC is calculated as:

$$\text{MEC} = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2}, \tag{9}$$

where O_i and P_i refer to observed and predicted species abundance at sampled location i, n is the number of sampled locations and $\bar{O}$ represents the mean observed species abundance across all sampled locations.

Since the different performance measures were either not normally distributed according to the Shapiro and Wilk (1965) test for normality, or had heterogeneous variance according to the Fligner and Killeen (1976) test for homogeneity of variances, or both, the Kruskal and Wallis (1952) test was used to determine the significant difference in predictive performance between the different cross-validation techniques. To compare their predictive performance, Dunn (1964) post-hoc test for the Kruskal-Wallis test was used. The Benjamini and Hochberg (1995) method was applied to adjust P-values. The cross-validation strategy that produced performance statistics from a sample as close as possible to train the model on 80% of population observations and testing on the remaining 20% was considered the most effective.

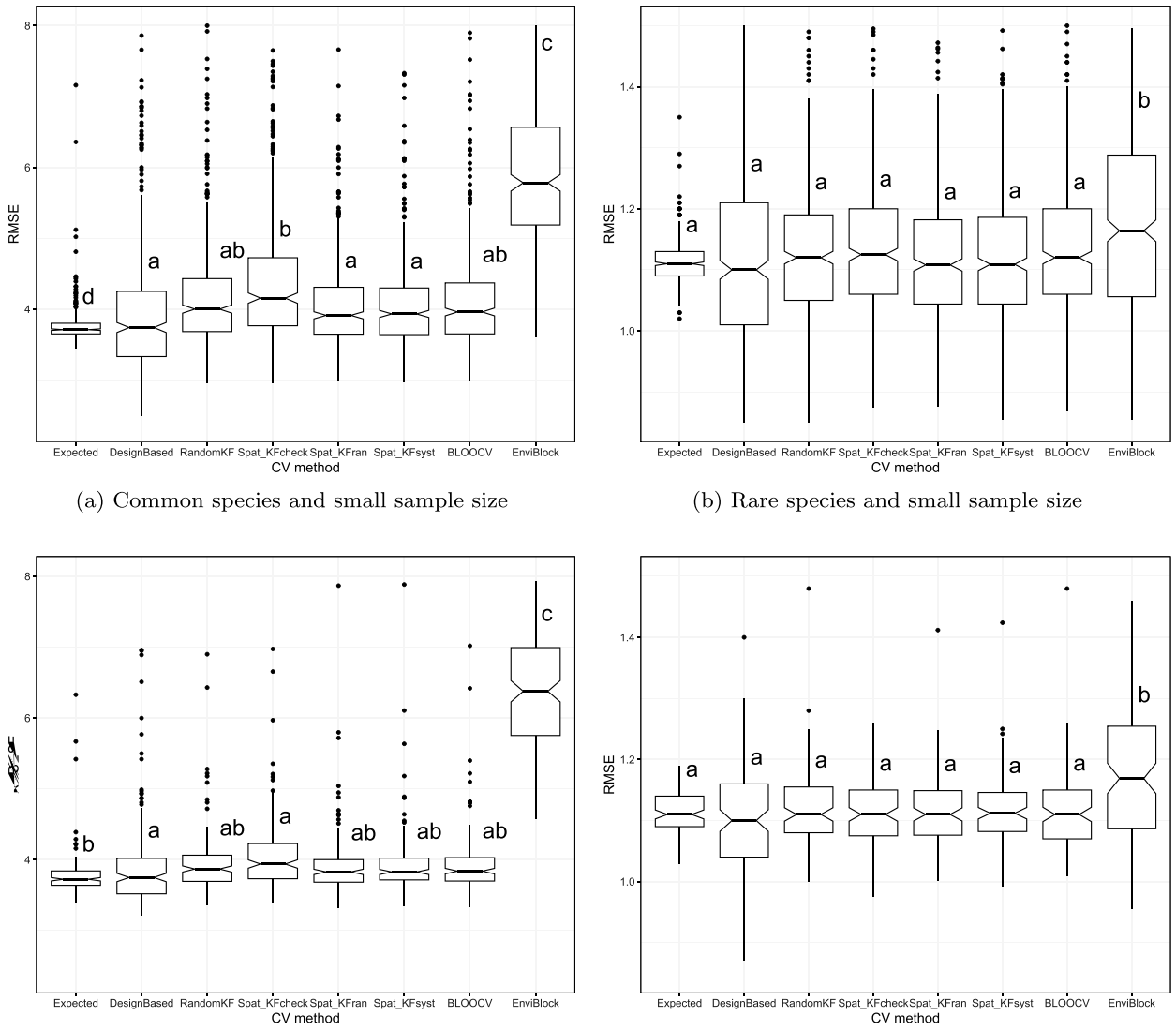
Table 2 Characteristics of the simulated distribution of species abundance (linear specification is used for overdispersion parameters and exponential model for the variogram. Min and Max are the minimum and maximum counts on the sampled sites after three visits)

Scenarios	Species detectability	Spatial autocorrelation	Abundance class	Descriptive statistics					Overdispersion parameters			Variogram parameters			
				Min	Mean	Median	Max	Sum	Statistic	Estimate	p value	Range	Partial sill	Nugget	Ratio partial sill/range
1	Perfect	No spatial	Common	0	11 ± 0	7 ± 0	153 ± 41	65237 ± 784	15.5 ± 2.8	9.1 ± 0.6	0.000 ± 0.000	396.3 ± 2330.5	34.7 ± 191.4	103.4 ± 6.6	0.1 ± 0.2
			Rare	0	1 ± 0	1 ± 0	16 ± 4	5914 ± 94	11.3 ± 1.7	0.8 ± 0.1	0.000 ± 0.000	459.7 ± 4921.6	0.3 ± 1.6	1.7 ± 0.1	0 ± 0.1
2	Perfect	High	Common	0	17 ± 19	3 ± 4	1870 ± ± ± 2734	103654 ± 118935	5.6 ± 2	214 ± 368	0.002 ± 0.008	133 ± 709	8881.4 ± 56395.8	4465.4 ± 16001.2	421.6 ± 5652.1
			Rare	0	2 ± 2	0 ± 0	191 ± 276	10403 ± 11944	5.3 ± 1.9	21.6 ± 37.1	0.002 ± 0.009	248.9 ± 1825.9	130.3 ± 811.6	46.7 ± 163.5	4 ± 76.1
		Low	Common	0	11 ± 3	6 ± 2	236 ± 118	64293 ± 18622	12.1 ± 3	15.4 ± 5.6	0.000 ± 0.000	469.8 ± 2427.1	365.6 ± 1917.2	140.3 ± 86.8	3.7 ± 12.4
			Rare	0	1 ± 0	0 ± 0	23 ± 10	5831 ± 1683	10.1 ± 2.2	1.4 ± 0.5	0.000 ± 0.000	598.9 ± 5819.4	8.3 ± 109.4	2 ± 0.9	0 ± 0.2
3	High	High	Common	0	14 ± 12	3 ± 2	1755 ± 3621	84271 ± 79377	8.2 ± 2.8	113.3 ± 349.1	0.000 ± 0.003	150.9 ± 1290.3	34895 ± 415072.7	4453.1 ± 31339.3	528.4 ± 3658.7
			Rare	0	2 ± 1	0 ± 0	179 ± 367	8999 ± 8341	7.7 ± 2.5	11.4 ± 34.8	0.000 ± 0.003	158.2 ± 789.7	342.4 ± 4384.9	47 ± 324.6	4.6 ± 31.9
		Low	Common	0	9 ± 2	6 ± 2	197 ± 94	56501 ± 14808	21.4 ± 3.9	2.7 ± 1.2	0.000 ± 0.000	565.1 ± 4900.4	344.8 ± 3589.7	96.8 ± 51.2	2.1 ± 5.2
			Rare	0	1 ± 0	0 ± 0	21 ± 9	5580 ± 1483	8.5 ± 2.4	0.2 ± 0.1	0.000 ± 0.001	342.1 ± 1162.8	2.1 ± 8.6	1.7 ± 0.7	0 ± 0.1
	Low	High	Common	0	6 ± 5	1 ± 1	671 ± 1369	35536 ± 31333	8.2 ± 2.8	41.1 ± 128	0.000 ± 0.003	1201.3 ± 17103.7	6325.6 ± 63464.6	630.2 ± 4437.3	−37.3 ± 1721.3
			Rare	0	1 ± 1	0 ± 0	71 ± 141	4403 ± 3643	7.2 ± 2.3	4 ± 12.4	0.001 ± 0.012	100 ± 429.7	54.3 ± 685	7 ± 47.4	0.2 ± 14.1
		Low	Common	0	5 ± 1	3 ± 1	79 ± 36	25847 ± 6214	12.8 ± 4.1	0.7 ± 0.5	0.004 ± 0.059	398.1 ± 2411.9	31.5 ± 149.7	17.4 ± 8.4	0.4 ± 0.8
			Rare	0	1 ± 0	0 ± 0	10 ± 4	3187 ± 776	−2.4 ± 2.4	0 ± 0	0.828 ± 0.302	678.5 ± 4187.5	1.1 ± 6.9	0.6 ± 0.2	0 ± 0.1

Results

Description of the generated species abundance distributions

Characteristics of simulated species abundance distribution parameters (descriptive statistics, variogram parameters and dispersion (overdispersion vs. equidispersion) and their variation between datasets are presented in Table 2. The simulated abundance distributions are skewed to the right for all species, as the medians are lower than the means. All of the generated species distributions are over-dispersed, except in the case where the species is rare, the probability of detection is low, and the spatial autocorrelation is low or absent. Datasets having outliers for the variogram's range parameter were not included in the analysis since the minimal number of blocks needed to assess the model's performance using the spatial fivefold CV approach was not fulfilled.



(a) Common species and small sample size

(b) Rare species and small sample size

(c) Common species and large sample size

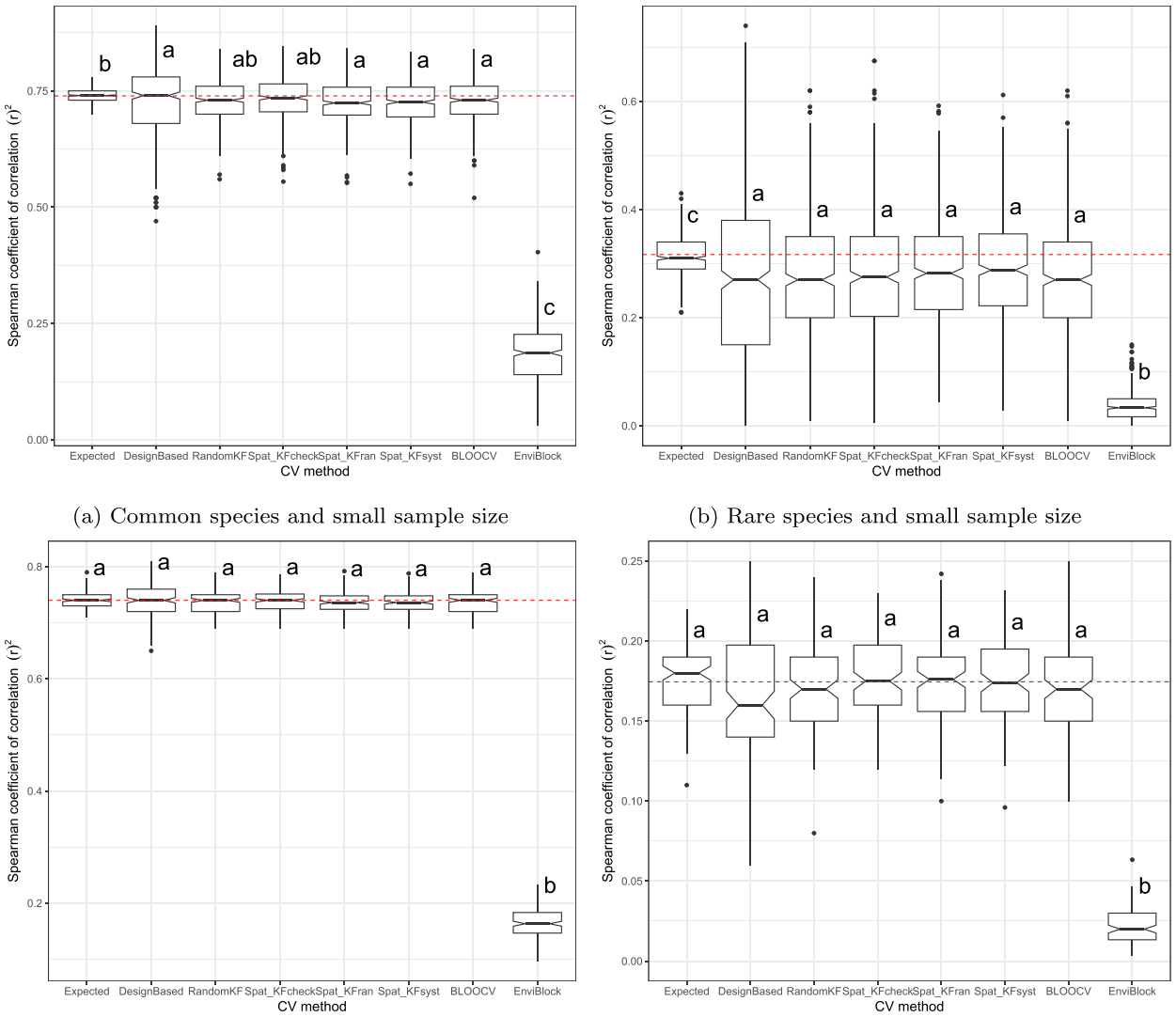
(d) Rare species and large sample size

Fig. 3 Forest accuracy in modelling species abundance data under optimal conditions (perfect detection and no spatial autocorrelation), using various cross-validation approaches. Expected: Design-based validation using all the population; *DesignBased* Design-based validation using sampled sites, *RandomKF* Standard random K-fold CV, *BLOOCV* Buffered leave-one-out CV, *Spat_KFran* Random spatial K-folds CV, *Spat_KFsyst* Systematic Spatial K-folds CV, *Spat_KFcheck* Checkerboard spatial K-folds CV, *EnviBlock* Environmental blocking. The red dashed line shows the expected average performance measures; There is no statistically significant difference between the means for CV methods sharing the same letter

RF performance using different CV strategies for species abundance data with perfect detection and no-spatial structure

Under ideal conditions, i.e. perfect detection and no spatial autocorrelation, Fig. 3 and Supplementary Figure 1 show that environmental blocking is the least accurate CV approach. This is true regardless of sample size or species abundance class. This is followed by spatial CV using a checkerboard pattern to assign blocks to folds for common species. This latter approach requires more observations to achieve the same level of accuracy as other CV approaches (spatial or no spatial). For relatively small sample sizes ($n = 200$), CV methods do not produce prediction estimates that are equivalent to those of the population from which the sample was taken, especially for common species. As rare species are less variable, they theoretically yield lower predictive errors and variability than common species.

Using discrimination metrics to evaluate the predictive performance of the RF, Fig. 4 and Supplementary



(a) Common species and small sample size

(b) Rare species and small sample size

(c) Common species and large sample size

(d) Rare species and large sample size

Fig. 4 The predictive power of the standard random forest for modelling species abundance data with perfect detection and no spatial autocorrelation, using different cross-validation methods. Expected: Design-based validation using all the population; *DesignBased* Design-based validation using sampled sites, *RandomKF* Standard random K-fold CV, *BLOOCV* Buffered leave-one-out CV, *Spat_KFran* Random spatial K-folds CV, *Spat_KFsyst* Systematic Spatial K-folds CV, *Spat_KFcheck* Checkerboard spatial K-folds CV, *EnviBlock* Environmental blocking. The red dashed line shows the expected average performance measures; There is no statistically significant difference between the means for CV methods sharing the same letter

Figure 2 show that, under ideal conditions, the environmental blocking method is the CV method with the least predictive power and modelling efficiency. This means the relationship between observed and predicted abundance is very weak when using this method. For common species, when the sample size is small, standard random CV and checkerboard spatial K-folds CV are the methods that produce R^2 values that are similar to the population R^2 but not significantly different from those obtained using random, systematic spatial K-folds CV or buffered leave-one-out CV. No CV method achieves the expected R^2 for rare species with a relatively small sample population. When sample sizes are sufficiently large for both common and rare species (see Fig. 4), the standard random K-folds CV method has the same predictive power as all other spatial CV methods except environmental blocking.

RF is effective (see Supplementary Figure 2) in modelling species abundance in absence of spatial dependence. It accurately reproduces observed data (MEC > 0.7) for common species. However, CV environmental blocking underestimates the modelling efficiency of RF. For rare species, Random Forest (RF) models perform moderately well (with MEC values between 0.5 and 0.7) using all CV methods except buffer-leave-one-out CV and environmental blocking, which tend to underestimate RF model efficiency when spatial structure in abundance data is absent. The standard random K-fold CV is the most efficient technique for all abundance groups and sample sizes.

Impact of sampling strategies on RF predictive performance using different CV strategies for spatially structured species abundance

This subsection summarise findings on the effect of the sampling strategy on RF predictive performance in spatially structured abundance data using various CV approaches. In the following sections, only large sample results are presented to avoid figures overloading. Findings show that the interpretation of RF and CV strategies performance depend on various parameters. These parameters include the chosen performance metrics, the sampling method used to collect data, and data characteristics. Compared to the expected predictive performance, results indicate that the best CV method may not always be the one that yields the highest values for discrimination performance metrics (R^2 and MEC) or the lowest accuracy metrics (RMSE and bias). This is because the above mentioned parameters may affect RF predictions, resulting in underfitting or overfitting for certain datasets.

RF predictive accuracy

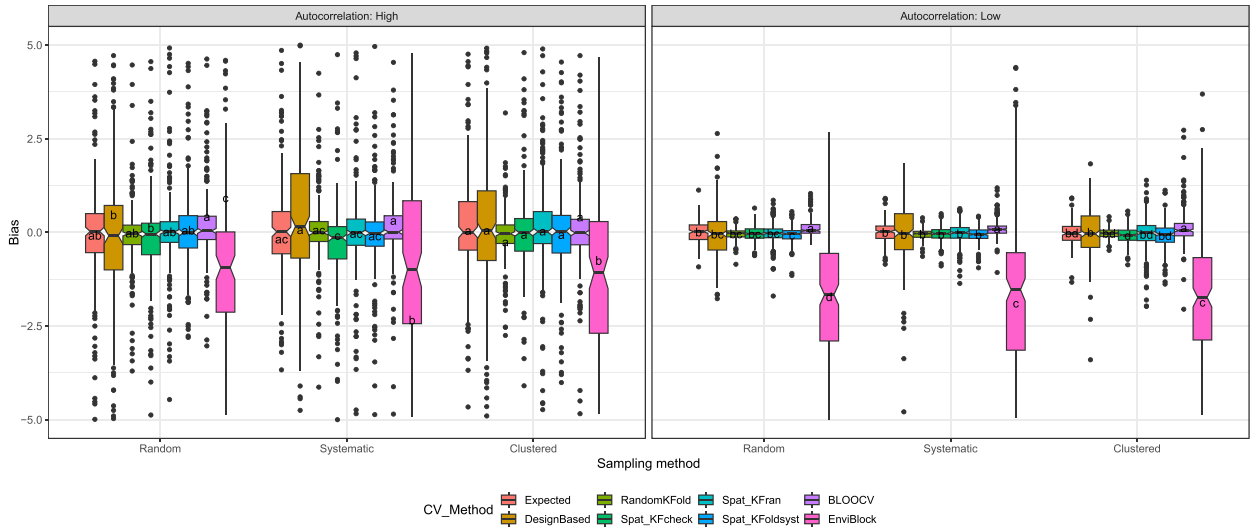
For low spatially autocorrelated abundance data, the results indicate that the choice of CV method does not have a significant effect on the RF's RMSE regardless of the species abundance class except for the two-stage clustered sampled abundance data when the species is common. In this case, design-based CV decreases the predictive accuracy while the buffered leave-one-out CV yields the lowest RMSE but is not significantly different from other methods. However, if the spatial autocorrelation in abundance data is high and the sampling is the two-stage clustering, the RF predictive accuracy is significantly overestimated using the environmental blocking (see Supplementary Figure 3) for both rare and common species. Log (RMSE) was used due to the high variability of RMSE values for abundance data with high spatial autocorrelation.

Bias

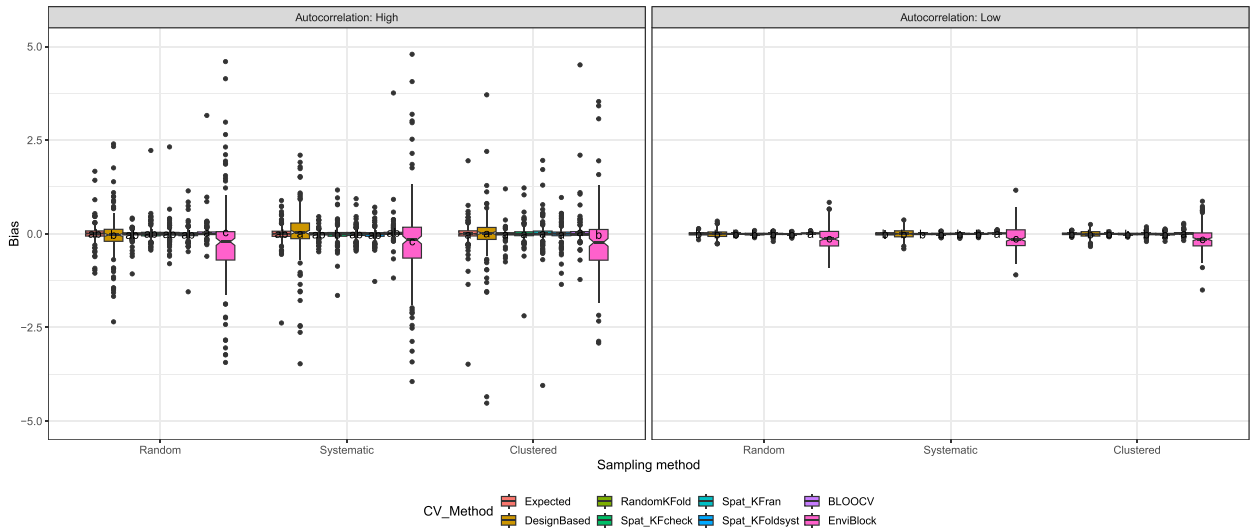
Figure 5 reveals that the choice of CV method significantly impacts RF predictive bias. Environmental blocking underestimates RF bias for all sampling methods, regardless of spatial autocorrelation variation strength, while buffered leave-one-out CV overestimates RF predictive bias for low spatially autocorrelated abundance data.

Predictive power

For low spatially autocorrelated abundance data, the analysis suggests that when the sampling is random or systematic, the choice of CV method has a significant impact on the R^2 between observed and RF predicted abundances and some methods especially Buffered leave-one-out CV and Environmental blocking have more substantial effects than others. If the sampling is systematic and the species is rare, the buffered leave-one-out CV yields RF predictive power, which is not significantly different from the best CV methods. Environmental blocking underestimates R^2 for both rare and common species for all sampling methods tested in this study. If the sampling is the two-stage clustering, the highest RF's R^2 are obtained using design-based validation and standard K-fold CV methods but they are not significantly different to random spatial K-fold CV and systematic spatial K-fold CV methods if the species is common. However, the most reliable R^2 is obtained using Buffered leave-one-out CV and spatial K-fold CV with a checkerboard pattern in the assignment of blocks to folds which are not statistically different to the expected population R^2 if the species is common. For rare species, the most reliable R^2



(a) Common species

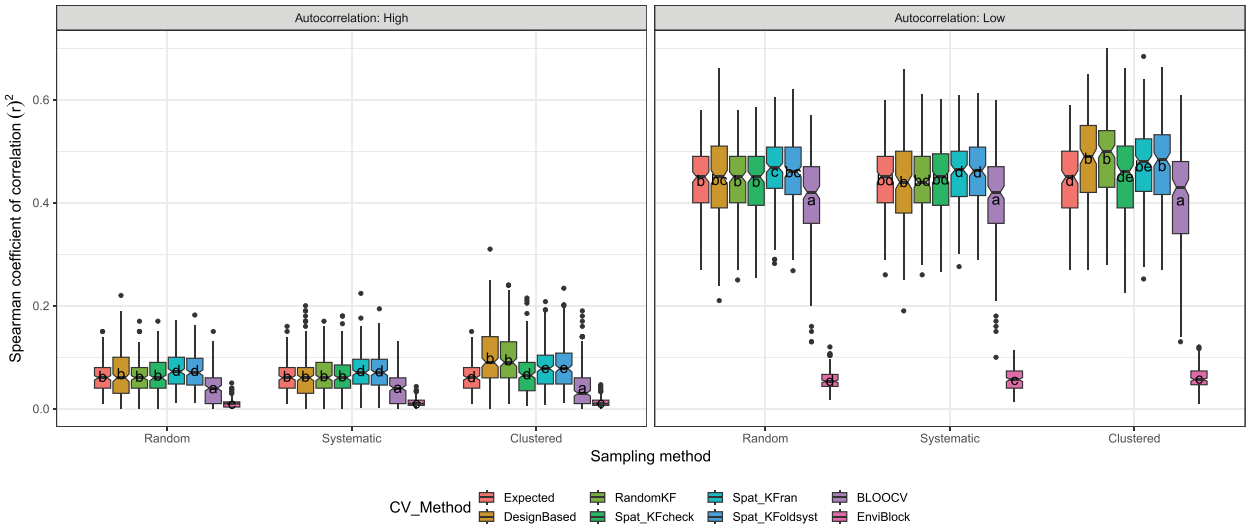


(b) Rare species

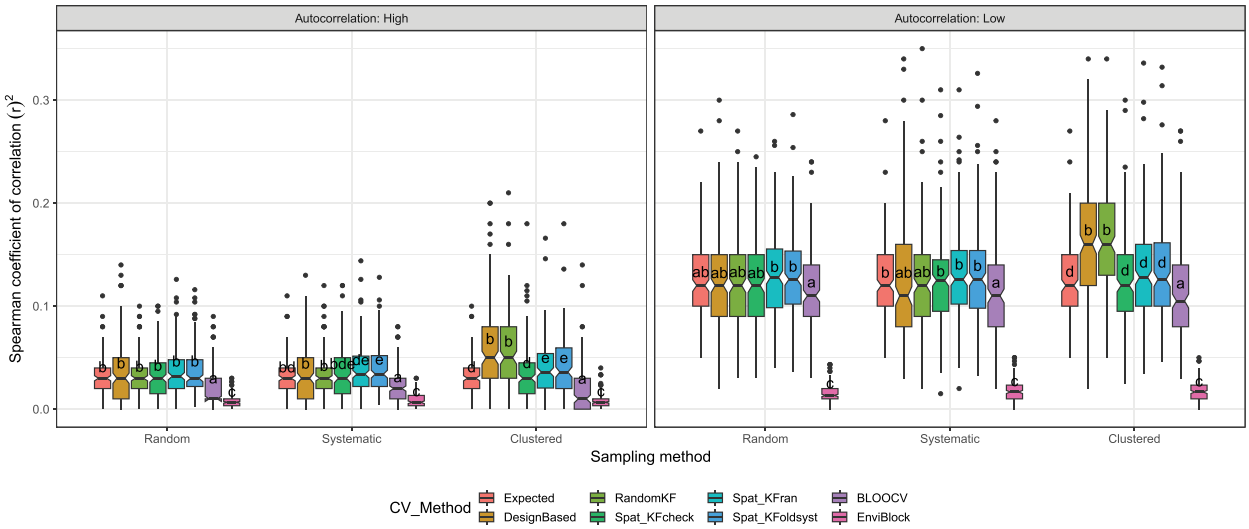
Fig. 5 RF predictive Bias using different sampling and cross-validation methods. Expected: Design-based validation using all the population; *DesignBased* Design-based validation using sampled sites, *RandomKF* Standard random K-fold CV, *BLOOCV* Buffered leave-one-out CV, *Spat_KFran* Random spatial K-folds CV, *Spat_KFsyst* Systematic Spatial K-folds CV, *Spat_KFcheck* Checkerboard spatial K-folds CV, *EnviBlock* Environmental blocking. The red dashed line shows the expected average performance measures; There is no statistically significant difference between the means for CV methods sharing the same letter

is obtained using spatial K-fold CV with a checkerboard pattern in the assignment of blocks to folds.

For high spatially autocorrelated abundance data, the analysis suggests that when the sampling is systematic or random the spatial K-fold CV with a checkerboard pattern in the assignment of blocks to folds yields expected population R^2 but it is not significantly different to other CV methods, except environmental blocking and buffered leave-one-out CV which underestimates the RF predictive power for both rare and common species. For high spatially autocorrelated abundance data, only the spatial K-fold CV with a checkerboard pattern in the assignment of blocks to folds allows the RF to yield a predictive power which is not significantly different to the expected population



(a) Common species



(b) Rare species

Fig. 6 Impact of sampling strategies on RF predictive power with different CV strategies. Expected: Design-based validation using all the population; *DesignBased* Design-based validation using sampled sites, *RandomKF* Standard random K-fold CV, *BLOOCV* Buffered leave-one-out CV, *Spat_KFran* Random spatial K-folds CV, *Spat_KFsyst* Systematic Spatial K-folds CV, *Spat_KFcheck* Checkerboard spatial K-folds CV, *EnviBlock* Environmental blocking. The red dashed line shows the expected average performance measures; There is no statistically significant difference between the means for CV methods sharing the same letter

predictive power if the species is common and the sampling is the two-stage clustering. If the species is rare (see Fig. 6), it is not significantly different from random spatial K-fold CV and systematic spatial K-fold CV methods.

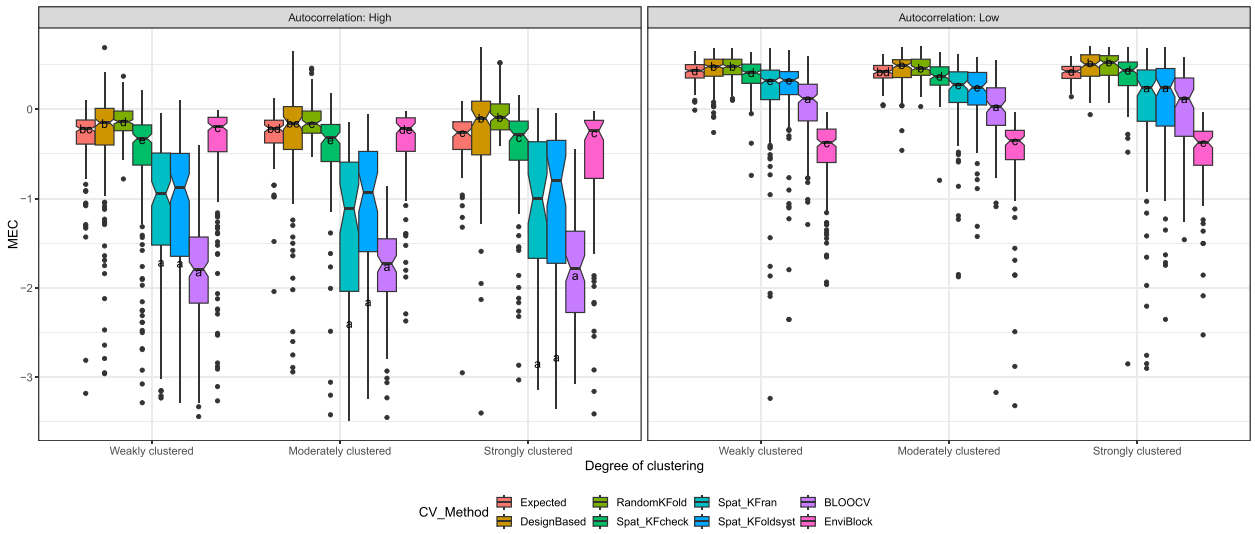
Modelling efficiency coefficient

For low spatially autocorrelated abundance data, whatever the sampling method, CV methods have a significant effect on the RF MEC. For both common and rare species, environmental blocking followed by buffered leave-one-out CV, random spatial K-fold CV and systematic spatial K-fold CV appear to have the most substantial effects on reducing RF MEC. Standard K-fold CV, spatial K-fold CV with a checkerboard pattern in the assignment of blocks to folds and design-based validation yield MECs similar or closest to expected RF MEC using all the population datasets.

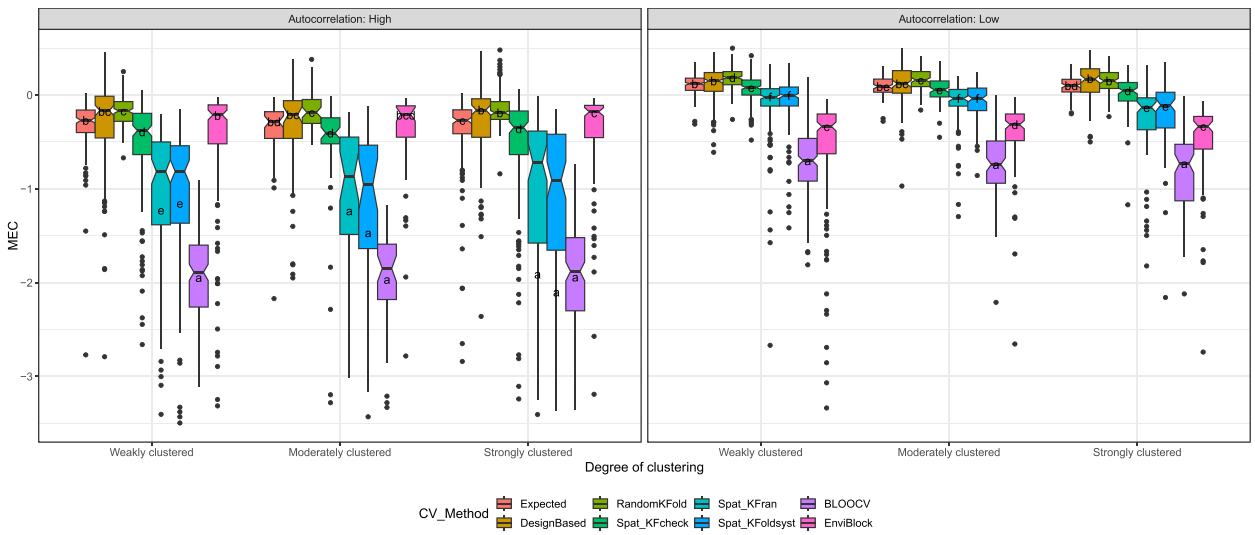
For high spatially autocorrelated abundance data, Supplementary Figure 4a reveal that for all sampling methods and CV methods, RF yields negative values for the MEC which indicates that the RF's predictions are poor and may be considered unacceptable. However, the analysis suggests that for randomly and systematically sampled abundance data, the best RF MEC are obtained using standard K-fold CV, environmental blocking, or design-based validation. If sampling is the two-sample clustering (see Supplementary Figure 4), RF's best MEC is obtained using environmental blocking followed by design-based validation and standard K-fold CV.

Effect of clustering level on spatial cross-validation performance

For low spatially autocorrelated abundance data, Fig. 7a, b show that for all degrees of clustering, the blocked



(a) Common species

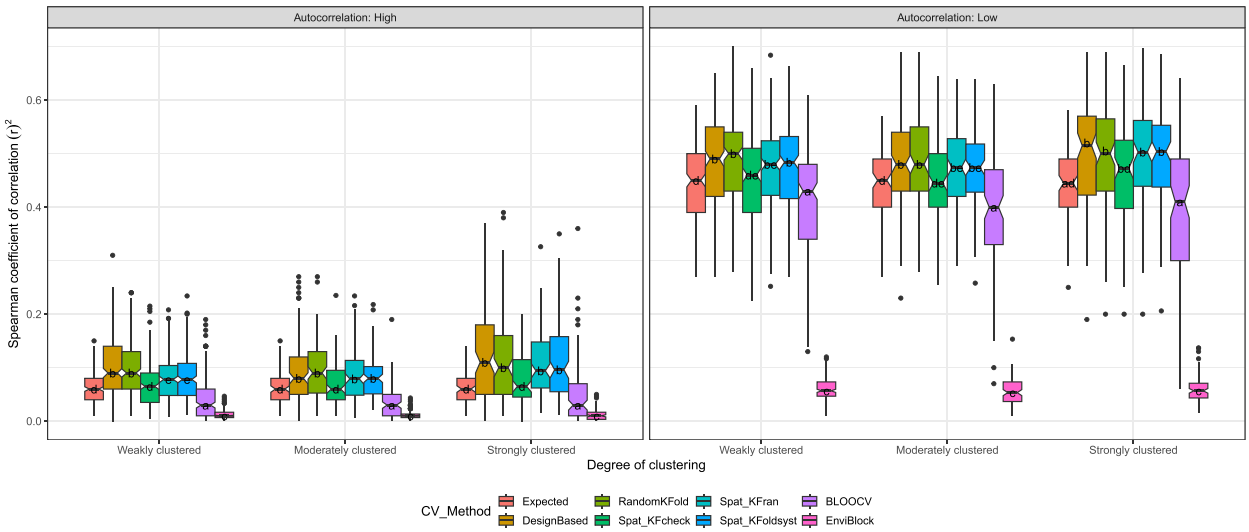


(b) Rare species

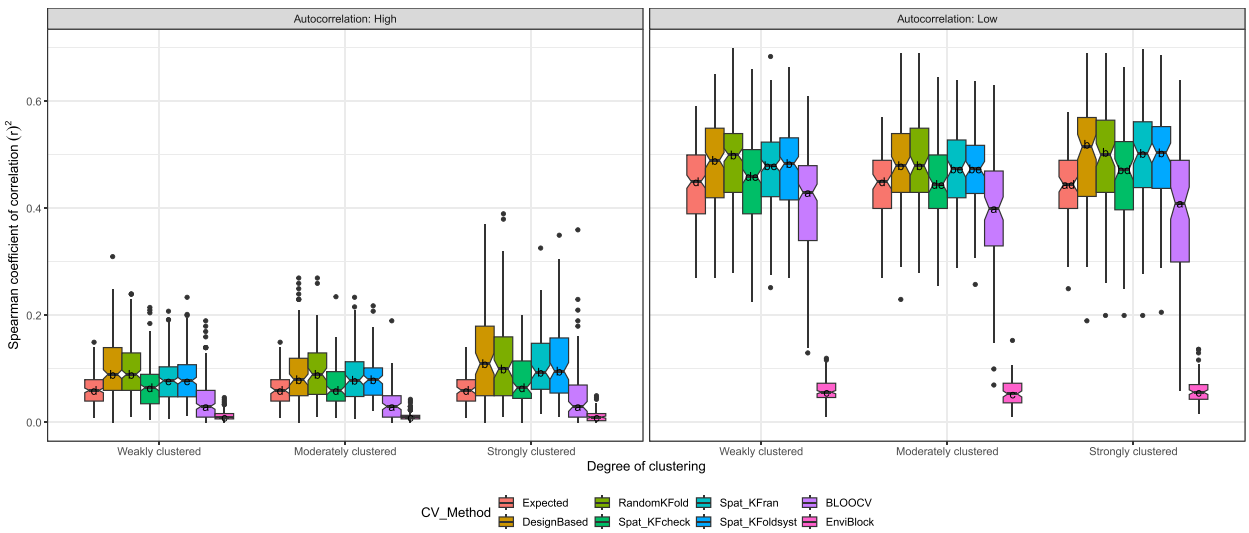
Fig. 7 Effect of the clustering and spatial autocorrelation on the RF modelling efficiency using different CV methods. Expected: Design-based validation using all the population; *DesignBased* Design-based validation using sampled sites, *RandomKF* Standard random K-fold CV, *BLOOCV* Buffered leave-one-out CV, *Spat_KFran* Random spatial K-folds CV, *Spat_KFsyst* Systematic Spatial K-folds CV, *Spat_KFcheck* Checkerboard spatial K-folds CV, *EnviBlock* Environmental blocking. The red dashed line shows the expected average performance measures; There is no statistically significant difference between the means for CV methods sharing the same letter

spatial CV with a checkerboard pattern in the assignment of blocks to folds yields from clustered samples the closest or similar MEC to the expected population MEC while the design-based validation and the standard K-fold CV significantly overestimate the MEC of the RF for both common and rare species. If the species is rare, the design-based validation and the standard K-fold CV are not significantly different from the expected MEC values. All other spatial CV approaches underestimate the MEC of RF. For high spatially autocorrelated abundance data, results suggest that for weakly and moderately clustered samples, design-based validation and environmental blocking CV yield the closest MEC to expected values for both common and rare species. However, for strongly clustered samples, if the species is common, environmental blocking CV and blocked spatial CV with a checkerboard pattern in the assignment of blocks to folds yield MEC similar to expected values, while only environmental blocking CV yields a similar MEC to expected values for rare species. Design-based validation significantly overestimates the MEC of the RF for strongly clustered samples for common species.

For all degrees of clustering, species abundance class, and spatial autocorrelation, Fig. 8a show that the most reliable R^2 is obtained using blocked spatial K-fold



(a) Common species



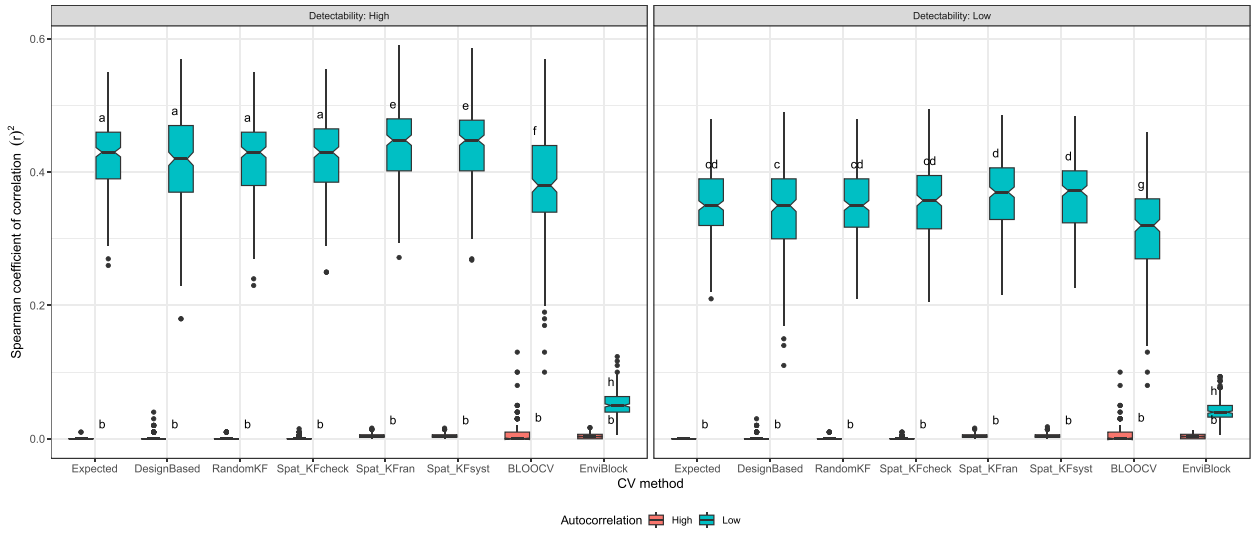
(b) Rare species

Fig. 8 Impact of clustering and spatial autocorrelation levels on RF predictive power and spatial CV approaches

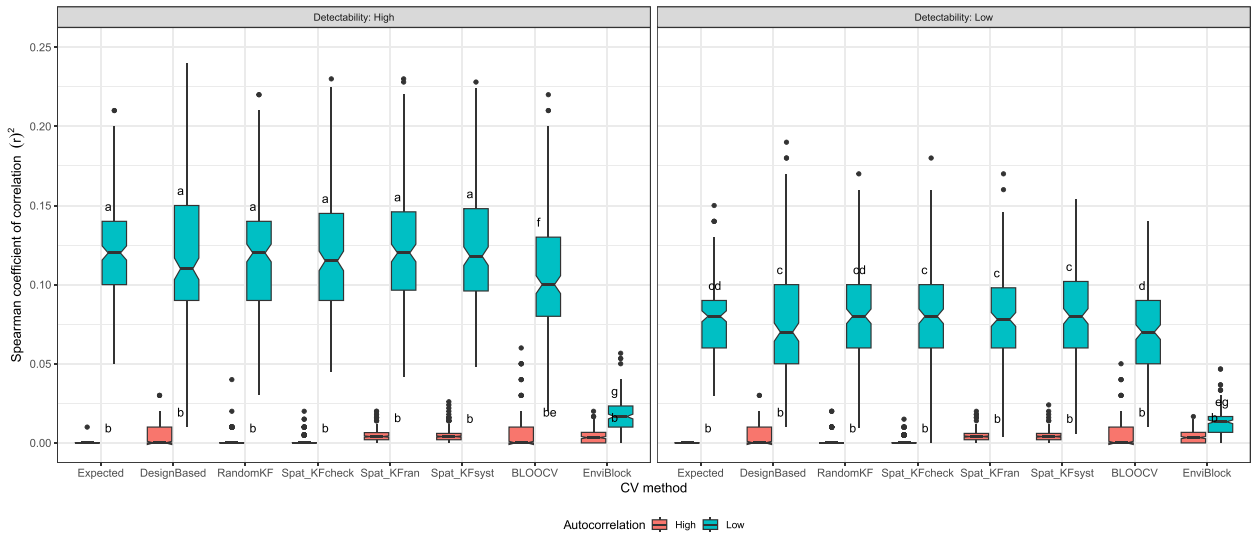
CV with a checkerboard pattern in the assignment of blocks to folds. Its prediction power from clustered samples is not significantly different from that expected assuming that the sample is representative of the population. However, if the species is rare and the spatial autocorrelation is low, its predictive power is not significantly different from the random and systematically blocked spatial CV K-fold CV.

Effects of spatial autocorrelation and imperfect detection on the performance of spatial CV approaches in RF using random sampling to predict species abundance

Given spatial autocorrelation and imperfect detection, CV methods behave differently depending on species type (abundance class) and sample size. The probability of species detection significantly affects the predictive accuracy and predictive discrimination of RF (see Figs. 9 and 10). When the probability of detecting a species is

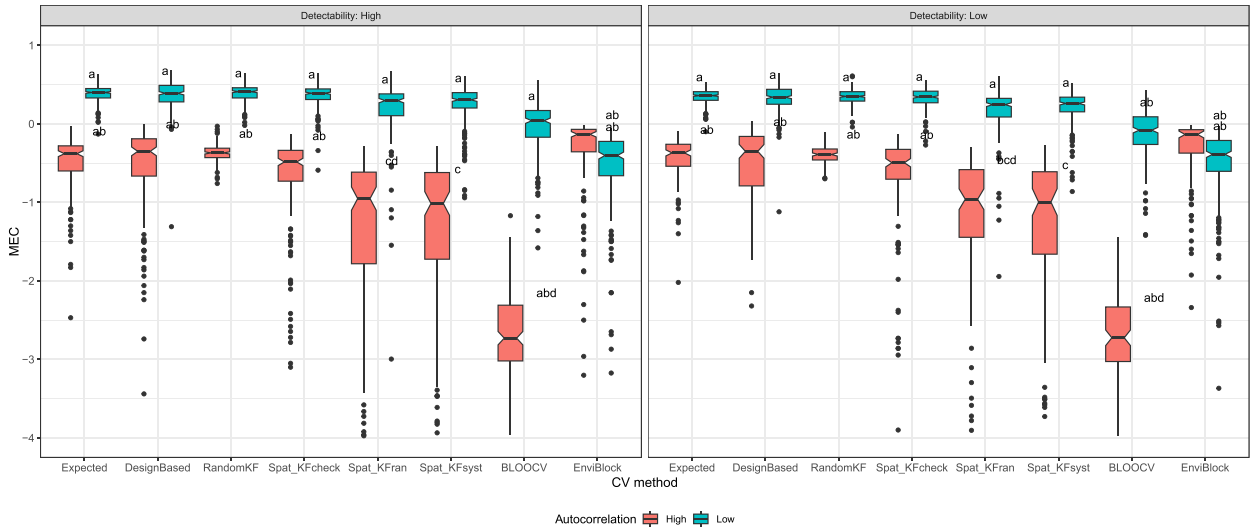


(a) Common species and large sample size

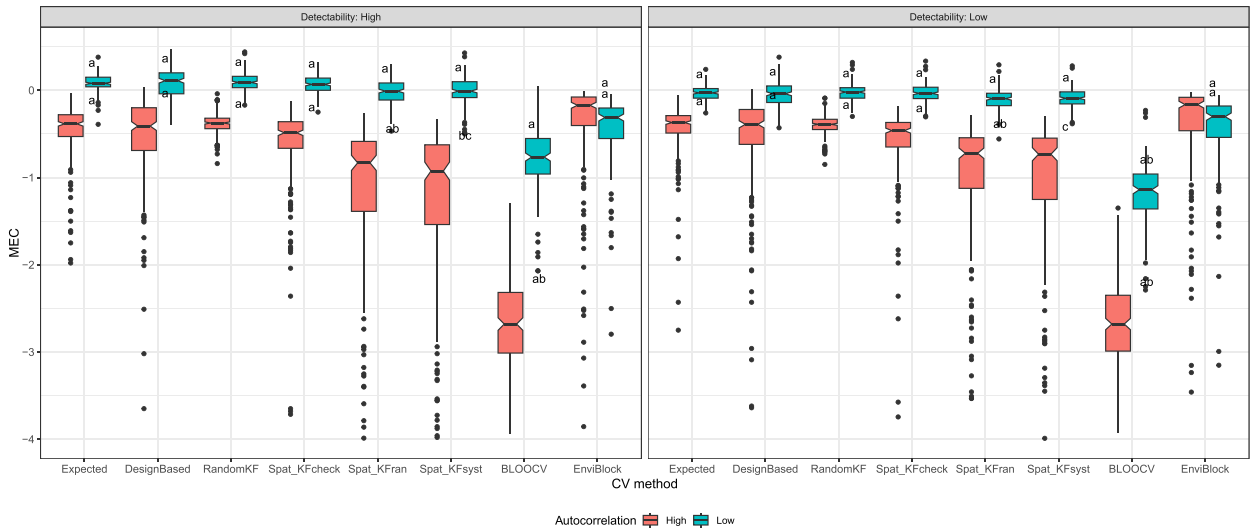


(b) Rare species and large sample size

Fig. 9 Effect of imperfect detection and spatial autocorrelation levels on the predictive power of spatial CV techniques in RF for different species abundance class. Expected: Design-based validation using all the population; *DesignBased* Design-based validation using sampled sites; *RandomKF* Standard random K-fold CV, *BLOOCV* Buffered leave-one-out CV, *Spat_KFran* Random spatial K-folds CV, *Spat_KFsyst* Systematic Spatial K-folds CV, *Spat_KFcheck* Checkerboard spatial K-folds CV, *EnviBlock* Environmental blocking. The red dashed line shows the expected average performance measures; There is no statistically significant difference between the means for CV methods sharing the same letter



(a) Common species and large sample size



(b) Rare species and large sample size

Fig. 10 Effect of imperfect detection and spatial autocorrelation levels on the MEC of RF for different species abundance classes using different spatial CV techniques. Expected: Design-based validation using all the population; *DesignBased* Design-based validation using sampled sites, *RandomKF* Standard random K-fold CV, *BLOOCV* Buffered leave-one-out CV, *Spat_KFran* Random spatial K-folds CV, *Spat_KFsyst* Systematic Spatial K-folds CV, *Spat_KFcheck* Checkerboard spatial K-folds CV, *EnviBlock* Environmental blocking. The red dashed line shows the expected average performance measures; There is no statistically significant difference between the means for CV methods sharing the same letter

high, RF provides low predictive accuracy but high predictive power, whereas when the probability of detecting a species is low, it provides high predictive accuracy but low predictive power.

Effect of species characteristics and spatial autocorrelation level on the predictive power of spatial CV techniques in RF

For a relatively small random sample, design-based validation and standard random CV are the best methods, but not significantly different from checkerboard spatial CV in assigning blocks to folds. However, design-based validation yields predictive accuracy with high variability. When spatial autocorrelation is low, random and systematic spatial CV overestimate the predictive power of the RF. In contrast, Fig. 9 shows that the buffer-leave-one-out CV and environmental blocking underestimate the predictive power of the RF. When spatial autocorrelation is high, the buffer-leave-one-out CV gives a similar

predictive ability to the most reliable methods, which are the standard random K-fold CV and the spatial CV with a checkerboard pattern in block-to-fold assignment methods for all species detection probabilities and abundance classes. They do not differ significantly from those estimated using the design-based validation, however, in contrast to the case of low spatial autocorrelation, when spatial autocorrelation is high and the sample size is relatively small, environmental blocking overestimates RF predictive ability more than random and systematic spatial CV, which overestimates RF predictive ability and prediction accuracy for both low and high spatial autocorrelation. For large sample sizes, when spatial autocorrelation is low, other spatial CV methods do not significantly affect RF performance to predict the abundance of a species with a low probability of detection, common or rare, except for environmental blocking, which is the least effective CV method in terms of predictive power, followed by buffer leave-one-out CV. When spatial autocorrelation is high, buffer-leave-one-out CV and environmental blocking are not statistically different from other methods. However, standard random CV and spatial CV with checkerboard assignment of blocks to folds are the methods with the lowest variability in predictive performance for both rare and common species.

Influence of species characteristics and spatial autocorrelation levels on the prediction accuracy of spatial CV techniques in RF using random sampling

For common species (see Supplementary Figure 4a), there is no significant difference in prediction accuracy between the different CV methods, regardless of sample size. However, environmental blocking has the lowest mean RMSE. The predictive accuracy of environmental blocking is not significantly different from that of systematic, random, and checkerboard spatial K-fold CV methods, given the variability in the accuracy of RF prediction. The different spatial CV methods do not correct for sampling biases and incomplete detection effects when spatial autocorrelation is present in small sample sizes. For rare species, box-plots in Supplementary Figure 4b depict the lack of a substantial effect for low or no spatially autocorrelated species abundance distribution. When spatial autocorrelation is high, the buffer-leave-one-out CV method allows RF to make accurate predictions with a relatively small sample size compared to other CV methods. However, when the spatial autocorrelation is low, this method is no longer more accurate than the other methods. For both rare and common species, where spatial autocorrelation is high, environmental blocking significantly overestimates the predictive accuracy of RF when sample sizes are small. When the probability of detecting a rare species is low and spatial autocorrelation is high, Figure Supplementary Figure 4b shows that the checkerboard spatial CV, buffer-leave-one-out CV methods and the standard K-fold CV methods yield RMSE no significantly different from expected values.

Influence of data features on the modelling efficiency coefficient of RF using different spatial CV methods

For high spatially autocorrelated species abundance and imperfect detection, even when the sample is representative, Fig. 10 demonstrates that the RF may perform very poorly regardless of the CV strategy used. RF fails to capture the variation and direction of observed abundance. These negative values suggest that alternative models, which take into account spatial autocorrelation and imperfect detection, should be considered to improve species abundance prediction. The standard random CV and the checkerboard spatial CV methods yield better MEC compared to other CV methods. However, they are not statistically different from other CV methods for common species due to the high variability in MEC.

Discussion

In the present study, we compared spatial CV strategies using the standard RF to predict species abundance in the presence of spatial autocorrelation and imperfect detection, two common characteristics of abundance data. Results showed that imperfect detection and spatial autocorrelation significantly affect the predictive accuracy and ability of the RF regardless of the CV technique. This is particularly the case when there is a cooccurrence of these factors.

Following the discussion of spatial CV methods, this section explores the conditions under which spatial or non-spatial CV methods may or may not be reliable in ecological studies, taking into account spatial autocorrelation level, sampling method, species abundance class and measurement error.

RF over or under-fitting using various CV methods for small sample size

Wadoux et al. (2021) found that the standard random K-fold CV method was less biased than spatial CV methods. Spatial CV methods are highly subjective and depend on various factors such as the sampling design used to define the folds, the method used to determine the spatial partitioning, and the exclusion distance. However, this study demonstrates that it is not prudent to draw such a general conclusion, as several factors can influence the performance of CV techniques. For example, except for environmental blocking which is the least accurate CV method in cases of perfect detection and no spatial autocorrelation, all methods provide predictive power, modelling efficiency, bias and accuracy that match

that of the underlying population, provided that the samples are large and representative of the population due to the large size of spatial blocs or exclusion distance. When the dependent variable is a count variable, such as the abundance of a species, the large variation in observations can mask the effect of the CV technique.

When the sample size is small and not representative of the population it was drawn from, either spatial or non-spatial CV methods may not produce prediction estimates equivalent to population validation parameters. Although RF can handle datasets with limited sample sizes (Ishwaran et al. 2010; Chen and Ishwaran 2012), its application requires a large sample to ensure that estimates reflect the population validation parameters. For the small sample size, the predictive performance of RF is overestimated. This suggests that in CV, the presence of spatial structure in the data is not the only factor influencing the over or underestimation of RF performance. When the species is common, spatial CV approaches that randomly and systematically allocate blocks to folds yield the highest RMSE in scenarios with imperfect detection even if data are not spatially autocorrelated, but they do not differ significantly from other CV methods.

However, as with all CV methods, they may overestimate the predictive accuracy and underestimate the predictive power of the RF, especially for rare species. Based on random sampling in the presence of high spatial autocorrelation, environmental blocking CV achieved the highest RMSE. However, the different spatial CV methods did not significantly improve prediction accuracy for the large sample size. Results from this study show that both spatial and non-spatial CV methods can significantly overestimate the accuracy of species abundance predictions with small samples and high spatial autocorrelation. This finding contrasts with those of Roberts et al. (2017) and Ploton et al. (2020), who showed that ignoring spatial dependence when cross-validating models can lead to severe underestimation of prediction error and false confidence in model predictions, masking model overfitting. However, Ploton et al. (2020) findings are likely to apply only to cluster samples and may not apply to all datasets and levels of spatial autocorrelation. Therefore, although the RF is robust to overfitting due to random noises (Breiman 2001), researchers should not assume that it is robust to overfitting caused by heterogeneity in predictor and response relationships (Wenger and Olden 2012) or other data features and selected validation method.

Performance of spatial CV methods in species abundance modelling

In terms of prediction ability, environmental blocking is the least efficient CV approach. It yields the poorest relationship between the observed abundance and the predicted values. When the probability of detecting a rare species is low and spatial autocorrelation is high, the checkerboard assignment of blocks to folds in the spatial K-fold CV methods, and the standard random K-fold CV method predict species abundance with an accuracy that is not significantly different from that of the population. However, the determination of the optimal block or buffer size is one of the challenges in the use of spatial blocks and buffered CV (Trachsel and Telford 2016). One of the approaches used to determine the block size is to fit a variogram to the raw species data and use the resulting range as the block size for spatial K-fold CV or the radius around each location in buffered leave-one-out CV (Roberts et al. 2017; Bio et al. 2002; Valavi et al. 2019).

Several studies Roberts et al. (2017); Telford and Birks (2009); Trachsel and Telford (2016) have used the spatial autocorrelation range in the model residuals to determine the block size for optimal separation between the training and test sets. Using the empirical variogram, which is an essential geostatistical tool for determining spatial autocorrelation, it is possible to define the range over which the residuals are independent of each other (Valavi et al. 2019). The empirical variogram is a measure of the variability between all pairs of points to explain the spatial autocorrelation pattern (O'Sullivan and Unwin 2010).

For spatial independence, (Ploton et al. 2020) recommend excluding validation data that are geographically close to the calibration data by a distance greater than the autocorrelation range of the empirical variogram. In the case of spatial autocorrelation, we assumed that each block size was 1.25 * the range of the variogram fitted to the raw species data. However, this proved insufficient to significantly improve the predictive performance of RF over what would be obtained using a randomly selected sample and standard random K-fold cross-validation. Wadoux et al. (2021) found similar results when assessing the accuracy of aboveground biomass (AGB) maps. Their results indicate that the design-based estimation of the population RMSE is unbiased and the estimates have little variation, but it requires probability sampling from the population.

For systematic and simple random sampling designs, CV using the standard random K-fold CV is close to unbiased but is too optimistic for clustered sampling. They also showed that the buffered leave-one-out and random spatial CV methods were too pessimistic, and

the former significantly overestimated the RMSE. Our results show that environmental blocking is more pessimistic about predictive ability and overestimates RMSE compared to other methods. These findings support a statement by Wadoux et al. (2021) that the pessimistic results of random spatial K-fold CV and buffered leave-one-out CV are likely caused by an overrepresentation of environmental conditions different from those at the calibration points and an underrepresentation of conditions comparable to those at the calibration points.

Checkerboard assignment of blocks to folds in the spatial K-fold CV methods for clustered samples

However, if the data are spatially structured, the performance of the RF using validation based on the sampling design or standard random K-fold CV varies significantly depending on the sampling method used. For example, standard K-fold CV is accurate even in the presence of spatial autocorrelation, regardless of species abundance, if the sample size is representative of the population and sampling is random or systematic. If the data are spatially structured, standard K-fold CV is unreliable for clustered samples. Over-fitting problems with structured data can be addressed by block cross-validation, which strategically splits data rather than randomly (Roberts et al. 2017). Bruin et al. (2022) used cross-validation to assess map accuracy and found that blocked spatial cross-validation was closest to the reference map accuracy metrics for highly clustered samples, where a large proportion of maps were predicted by extrapolation. However, there was still bias in the results. To avoid extrapolation in these cases, the best approach was shown to be either to restrict the predicted area or to perform additional sampling.

As alternatives to random blocked spatial CV, Bruin et al. (2022) proposed inverse sampling-intensity weighted and two geostatistical model-based CV methods. Sampling-intensity weighted CV directly addresses the problem of spatial clustering, rather than the erroneously described problem of physical proximity of test and training data. They found that, compared to blocked spatial cross-validation, sampling-intensity weighted CV was less biased and more reliable for unclustered to moderately clustered data, but more biased for highly clustered samples. They emphasise the need for further research to improve accuracy assessment using CV from highly spatially clustered samples and suggest the use of the standard random CV for unclustered data and weighted CV for moderately clustered samples.

Recently, Wang et al. (2023) proposed a new CV method for evaluating geospatial ML models by considering the geographical and feature spaces to split the samples into training and validation sets. This method resulted in a more rational split, which led to more accurate results when evaluating the models. However, they used random assignment of blocks to folds, like most studies evaluating spatial CV methods.

According to our findings, regardless of the degree of spatial autocorrelation, the probability of species detection, and the type of species, the checkerboard assignment of blocks to folds in the spatial CV method provides accurate predictive performance for clustered samples. Roberts et al. (2017) suggest using grouped sets of blocks, such as the checkerboard pattern, to select evaluation blocks in cases of irregular or unbalanced data sampling. As species abundance samples are rarely evenly distributed across the landscape, one of the most important processes in species modelling is the assignment of blocks to folds (Valavi et al. 2019).

The checkerboard pattern in assignment of blocks to folds in the spatial CV method provides an excellent opportunity to explore when samples are clustered. The checkerboard pattern is a technique used in ecological models to ensure spatial independence between the training and test sets. It prevents neighbouring blocks from sharing observations, allowing the model to be applied to distinct areas. This is crucial in ecology, where species distribution may vary. The checkerboard pattern exposes the model to various spatial patterns, enhancing its robustness to spatial heterogeneity. It ensures an even representation of spatial patterns across each fold, preventing biases due to unequal distributions. It also minimizes the risk of leakage by keeping neighbouring blocks separate during cross-validation, which occurs when information from the training set affects the test set.

This study explored the application of spatial CV for evaluating models with clustered and unclustered spatial data. While the different spatial CV approaches tested in this study do not offer advantages over traditional CV methods, even for high spatially autocorrelated data for large samples randomly or systematically collected, the checkerboard pattern in spatial CV gives better performances for clustered samples in most cases. Recent work suggest alternative methods to random blocked spatial CV such as inverse sampling-intensity weighted and geostatistical model-based CV methods that enhance predictive performance for clustered samples (Bruin et al. 2022). However, the performance of these methods depends on the clustering level, and to our knowledge, no approach performs well regardless of the data clustering level. Future research could explore the effectiveness of applying the checkerboard pattern, commonly used in blocked spatial CV, to these alternative methods for clustered data, contributing to a more comprehensive understanding of spatial CV strategies for evaluating models with spatial dependence.

Conclusion

Making accurate and reliable predictions of species abundance can help us respond to changing ecological conditions and improve scientific understanding. To select, validate and assess the predictive power of ecological models, cross-validation is often used. Ecological data often have internal dependence structures, so there is increasing interest in spatial cross-validation techniques that increase the independence between training and test data by dividing the data into 'blocks' at some central point(s) of the dependence structure, and avoid overfitting to produce unbiased errors and parameter estimates. The findings indicate that even in the presence of high spatial autocorrelation and imperfect detection, design-based validation and the standard K-fold CV remain effective strategies for evaluating the performance of machine learning methods for predicting species abundance, provided that sampling is random and, in some cases, systematic. Although RF can deal with data sets with limited sample sizes, for small sample sizes, both the standard and spatial CV approaches overestimate prediction accuracy and discrimination. Standard K-fold CV overestimates RF's predictive performance, particularly for clustered data. Checkerboard block-fold assignment in spatial CV is the spatial strategy that provides accurate prediction regardless of sampling design, whereas environmental blocking CV is overly pessimistic for predictive ability and overestimates RF predictive accuracy. Thus, in the case of clustered data, the checkerboard allocation of blocks to folds in the spatial CV method is an opportunity to explore. This is because, until proven otherwise, spatial CV methods have no theoretical basis.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40068-024-00352-9.

Supplementary material 1.

Acknowledgements

This work was supported by the DAAD In-Country/In-Region Scholarship Program FSA/UAC, International Development Research Centre (IDRC) and Swedish International Development Cooperation Agency (SIDA) through the Artificial Intelligence for Development (AI4D) Africa Programme, managed by Africa Center for Technology Studies (ACTS) Scholarship Program; and the Regional Universities Forum for Capacity Building in Agriculture through the Graduate Training Assistantship Program supported by the Carnegie Corporation in New York.

Author contributions

C.A.M. designed the study, analysed, and interpreted results, and drafted the manuscript, A.B.F. and R.G.K. supervised the study and revised the manuscript, and all authors approved the final version.

Funding

This study was funded by Deutscher Akademischer Austauschdienst (In-Country/In-Region Scholarship Program FSA/UAC), International Development Research Centre (Artificial Intelligence for Development (AI4D) Africa Programme), Styrelsen för Internationellt Utvecklingssamarbete (Artificial Intelligence for Development (AI4D) Africa Programme).

Availability of data and materials

The data and materials supporting this study's findings are available from the corresponding author on request.

Declarations

Competing interests

The authors declare that there is no Competing interests regarding the publication of this article.

Author details

[1]Laboratoire de Biomathématiques et d'Estimations Forestières, Faculté des Sciences Agronomiques, Université d'Abomey-Calavi, 04 PB 1525 Cotonou, Benin. [2]Faculty of Agriculture and Environmental Sciences, Université Evangélique en Afrique (UEA), P.O. Box: 3323, Bukavu, Democratic Republic of the Congo. [3]Unité de Recherche en Foresterie et Conservation des Bioressources, Ecole de Foresterie Tropicale, Université Nationale d'Agriculture, BP 43 Porto Novo, Bénin.

Received: 22 November 2023 Accepted: 7 June 2024
Published online: 28 June 2024

References

Araújo MB, Pearson RG, Thuiller W, Erhard M (2005) Validation of species-climate impact models under climate change. Glob Change Biol 11(9):1504–1513. https://doi.org/10.1111/j.1365-2486.2005.01000.x

Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Stat Surv 4(none):40–79. https://doi.org/10.1214/09-SS054

Austin MP, Belbin L, Meyers JA, Doherty MD, Luoto M (2006) Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. Ecol Model 199:197–216. https://doi.org/10.1016/j.ecolmodel.2006.05.023

Bahn V, McGill BJ (2013) Testing the predictive performance of distribution models. Oikos 122(3):321–331. https://doi.org/10.1111/j.1600-0706.2012.00299.x

Baldridge E, Harris DJ, Xiao X, White EP (2016) An extensive comparison of species-abundance distribution models. PeerJ 4:e2823. https://doi.org/10.7717/peerj.2823. (**ISSN 2167-8359**)

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc Ser B 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Biau G (2012) Analysis of a random forests model. J Mach Learn Res 13:1063–1095

Bio Ana MF, De Piet B, De Els B, Willy H, Martin W (2002) Prediction of plant species distribution in lowland river valleys in Belgium: modelling species response to site conditions. Biodivers Conserv 11(12):2189–2216. https://doi.org/10.1023/A:1021346712677

Breiman L (1996) Bagging predictors. J Mach Learn Res 24(2):123–40

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324. (**ISSN 1573-0565**)

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees, 1st edn. Chapman and Hall/CRC Press, Boca Raton. https://doi.org/10.1201/9781315139470

Brenning A (2005) Spatial prediction models for landslide hazards: review, comparison and evaluation. Nat Hazards Earth Syst Sci 5(6):853–862. https://doi.org/10.5194/nhess-5-853-2005

Brownlee J (2019) Statistical methods for machine learning: discover how to Transform Data into Knowledge with Python, volume 4 of machine learning mastery. Machine learning mastery, 1 edition. URL https://dokumen.pub/statistical-methods-for-machine-learning.html. Accessed 18 Jul 2023.

Brus DJ (2021) Statistical approaches for spatial sample survey: persistent misconceptions and new developments. Eur J Soil Sci 72(2):686–703. https://doi.org/10.1111/ejss.12988
Brus DJ, Kempen B, Heuvelink GBM (2011) Sampling for validation of digital soil maps. Eur J Soil Sci 62(3):394–407. https://doi.org/10.1111/j.1365-2389.2011.01364.x
Ceballos G, Ehrlich PR, Raven PH (2020) Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. Proc Natl Acad Sci USA 117(24):13596–13602. https://doi.org/10.1073/pnas.1922686117. (**ISSN 0027-8424**)
Chen X, Ishwaran H (2012) Random forests for genomic data analysis. Genomics 99(6):323–329. https://doi.org/10.1016/j.ygeno.2012.04.003
Clements CF, Blanchard JL, Nash KL, Hindell MA, Ozgul A (2017) Body size shifts and early warning signals precede the historic collapse of whale stocks. Nat Ecol Evol 1(7):0188. https://doi.org/10.1038/s41559-017-0188
Cochran WG (1977) Sampling techniques. Wiley, New York
Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88(11):2783–2792. https://doi.org/10.1890/07-0539.1
De Bruin S, Brus DJ, Heuvelink GBM, van Ebbenhorst TT, Wadoux AMJC (2022) Dealing with clustered samples for assessing map accuracy by cross-validation. Ecol Inform 69:101665. https://doi.org/10.1016/j.ecoinf.2022.101665
Dietterich TG (2004) An experimental comparison of three methods for constructing ensembles of decision trees. Mach Learn 40:139–157
Dormann FC, McPherson JM, Araújo MB, Bivand R, Bolliger J, Carl G, Davies RG, Hirzel A, Walter Jetz W, Kissling D, Kühn I, Ohlemüller R, Peres-Neto PR, Reineking B, Schröder B, Schurr FM, Wilson R (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30(5):609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x
Dunn OJ (1964) Multiple comparisons using rank sums. Technometrics 6(3):241–252. https://doi.org/10.1080/00401706.1964.10490181
Fletcher R, Fortin M (2018) Accounting for spatial dependence in ecological data. Springer International Publishing, Cham, pp 169–210
Fligner JM, Killeen TL (1976) Distribution-free two-sample tests for scale. J Am Stat Assoc 71(353):210–213
Franklin J (2010) Mapping species distributions: spatial inference and prediction. Ecology, biodiversity and conservation. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511810602
Gérard B, Erwan S (2016) A random forest guided tour. TEST 25(2):197–227. https://doi.org/10.1007/s11749-016-0481-7
Goedickemeier I, Wildi O, Kienast F (1997) Sampling for vegetation survey: some properties of a gis-based stratification compared to other statistical sampling methods. Coenoses 12(1):43–50
Gregoire TG, Valentine HT (2007) Sampling strategies for natural resources and the environment, 1st edn. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9780203498880
Greig-Smith P (1983) Quantitative plant ecology volume 9 of California series on social choice and political economy. University of California Press, Berkeley
Gruijter JJ, Bierkens FPM, Brus JD, Martin K (2006) Sampling for natural resource monitoring. Earth and environmental science, earth and environmental science, 1st edn. Springer, Berlin. https://doi.org/10.1007/3-540-33161-1
Guélat J, Kéry M (2018) Effects of spatial autocorrelation and imperfect detection on species distribution models. Methods Ecol Evol 9(6):1614–1625. https://doi.org/10.1111/2041-210X.12983
Guisan A, Thuiller W, Zimmermann NE (2017) Habitat suitability and distribution models : with applications in R. Ecology, biodiversity, and conservation. Cambridge University Press, Cambridge
Hartigan JA, Wong MA (1979) A k-means clustering algorithm. J Royal Stat Soc Ser C 28(1):100–108
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7
Hastings RA, Rutterford LA, Freer JJ, Collins RA, Simpson SD, Genner MJ (2020) Climate change drives poleward increases and equatorward declines in marine species. Curr Biol 30(8):1572-1577.e2. https://doi.org/10.1016/j.cub.2020.02.043
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. https://doi.org/10.7717/peerj.5518
Hirzel AH, Helfer V, Metral F (2001) Assessing habitat-suitability models with a virtual species. Ecol Model 145(2):111–121. https://doi.org/10.1016/S0304-3800(01)00396-9
Ho T (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20:832–844
Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS (2010) High-dimensional variable selection for survival data. J Am Stat Assoc 105(489):205–217. https://doi.org/10.1198/jasa.2009.tm08622
James G, Witten D, Hastie T, Tibshirani RJ (2013) An introduction to statistical learning: with applications in R. Springer, Berlin
Kellner KF, Swihart RK (2014) Accounting for imperfect detection in ecology: a quantitative review. PLOS ONE 9(10):1–8. https://doi.org/10.1371/journal.pone.0111436
Kenkel NC, Juhász-Nagy P, Podani J (1989) On sampling procedures in population and community ecology. Vegetatio 83:195–207. https://doi.org/10.1007/BF00031692
Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th international joint conference on Artificial intelligence, volume 2, pages 1137–1143, San Francisco, CA, USA. Morgan Kaufmann Publ
Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. J Am Stat Assoc 47(260):583–621. https://doi.org/10.1080/01621459.1952.10483441
Kuhn M, Johnson K (2013) Applied predictive modeling, 1st edn. Springer, New York. https://doi.org/10.1007/978-1-4614-6849-3
Lawler JJ, White D, Neilson RP, Blaustein AR (2006) Predicting climate-induced range shifts: model differences and model reliability. Glob Change Biol 12(8):1568–1584. https://doi.org/10.1111/j.1365-2486.2006.01191.x
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. https://doi.org/10.1111/geb.12161
Legendre P, Fortin MJ (1989) Spatial pattern and ecological analysis. Vegetatio 80(2):107–138. https://doi.org/10.1007/BF00048036
Levy PS, Lemeshow S (2013) Sampling of populations: methods and applications. Wiley series in survey methodology, 4th edn. John Wiley & Sons, Hoboken
Lieske DJ, Bender DJ (2011) A robust test of spatial predictive models: geographic cross-validation. J Environ Inform 17(2):91–101. https://doi.org/10.3808/JEI.201100191
Lyons MB, Keith DA, Phinn SR, Mason TJ, Elith J (2018) A comparison of resampling methods for remote sensing classification and accuracy assessment. Remote Sens Environ 208:145–153. https://doi.org/10.1016/j.rse.2018.02.026
Martín B, González-Arias J, Vicente-Vírseda JA (2021) Machine learning as a successful approach for predicting complex spatio-temporal patterns in animal species abundance. Animal Biodivers Conserv 44(2):289–301
Matthew JA, Gunnar M, Dan JC, Paul DMH, Robert KB, Timothy JD, Michelle G (2013) Statistical testing of a new testate amoeba-based transfer function for water-table depth reconstruction on ombrotrophic peatlands in north-eastern Canada and Maine, united states. J Quat Sci 28(1):27–39. https://doi.org/10.1002/jqs.2584
McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benecha HK, Dornelas M, Enquist BJ, Green JL, He F, Hurlbert AH, Magurran AE, Marquet PA, Maurer BA, Ostling A, Soykan CU, Ugland KI, White EP (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol Lett 10(10):995–1015. https://doi.org/10.1111/j.1461-0248.2007.01094.x
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. https://doi.org/10.1016/j.ecolmodel.2019.108815
Meynard CN, Quinn JF (2007) Predicting species distributions: a critical comparison of the most common statistical models using artificial species. J Biogeogr 34(8):1455–1469. https://doi.org/10.1111/j.1365-2699.2007.01720.x

Mi C, Huettmann F, Sun R, Guo Y (2017) Combining occurrence and abundance distribution models for the conservation of the great bustard. PeerJ 5:e4160. https://doi.org/10.7717/peerj.4160
Miller J, Franklin J, Aspinall R (2007) Incorporating spatial dependence in predictive vegetation models. Ecol Model 202(3):225–242. https://doi.org/10.1016/j.ecolmodel.2006.12.012
Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part i - a discussion of principles. J Hydrol 10(3):282–290. https://doi.org/10.1016/0022-1694(70)90255-6
O'Sullivan D, Unwin DJ (2010) Area objects and spatial autocorrelation, chapter 7. John Wiley & Sons, Ltd., Hoboken. 187–214. https://doi.org/10.1002/9780470549094.ch7
Pauly D, Froese R (2010) A count in the dark. Nat Geosci 3(10):662–663. https://doi.org/10.1038/ngeo973
Peterson TA, Soberén J, Pearson RG, Anderson RP, Martínez-Meyer E, Nakamura M, Araújo MB (2012) Ecological niches and geographic distributions (MPB-49). Princeton University Press, Princeton. https://doi.org/10.1515/9781400840670
Ploton P, Mortier F, Réjou-Méchain M, Barbier N, Picard N, Rossi V, Dormann C, Cornu G, Viennois G, Bayol N, Lyapustin A, Gourlet FS, Rl Pélissier (2020) Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat Commun 11(1):4540. https://doi.org/10.1038/s41467-020-18321-y
Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9(2):181–199. https://doi.org/10.1007/s10021-005-0054-1
R Core Team (2022) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. URL https://www.R-project.org/
Radosavljevic A, Anderson RP (2014) Making better maxent models of species distributions: complexity, overfitting and evaluation. J Biogeogr 41(4):629–643. https://doi.org/10.1111/jbi.12227
Roberts DR, Hamann A (2012) Method selection for species distribution modelling: are temporally or spatially independent evaluations necessary? Ecography 35(9):792–802. https://doi.org/10.1111/j.1600-0587.2011.07147.x
Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8):913–929. https://doi.org/10.1111/ecog.02881
Royle AJ, Dorazio RM (2009) Occupancy and abundance. In: Royle JA, Dorazio RM (eds) Hierarchical modeling and inference in ecology. Academic Press, San Diego, pp 127–157
Rykiel EJ (1996) Testing ecological models: the meaning of validation. Ecol Model 90(3):229–244. https://doi.org/10.1016/0304-3800(95)00152-2
Saha A, Basu S, Datta A (2023) Random forests for spatially dependent data. J Am Stat Assoc 118(541):665–683. https://doi.org/10.1080/01621459.2021.1950003
Scornet E (2016) Random forests and kernel methods. IEEE Trans Inform Theory 62(3):1485–1500. https://doi.org/10.1109/TIT.2016.2514489
Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc 88(422):486–494. https://doi.org/10.1080/01621459.1993.10476299
Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52(3–4):591–611. https://doi.org/10.1093/biomet/52.3-4.591
Snee RD (1977) Validation of regression models: Methods and examples. Technometrics 19(4):415–428. https://doi.org/10.1080/00401706.1977.10489581
Stehman SV (1999) Basic probability sampling designs for thematic map accuracy assessment. Int J Remote Sens 20(12):2423–2441. https://doi.org/10.1080/014311699212100
Stehman Stephen V, Foody Giles M (2009) Accuracy assessment. In: Warner TA, Foody GM, Nellis MD (eds) The SAGE handbook of remote sensing. SAGE Publications Inc, Thousand Oaks
Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J Royal Stat Soc Ser B 36(2):111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x
Su Q (2018) A general pattern of the species abundance distribution. PeerJ 6:e5928. https://doi.org/10.7717/peerj.5928. (**ISSN 2167-8359**)
Telford RJ, Birks HJB (2009) Evaluation of transfer functions in spatially structured environments. Quat Sci Rev 28(13):1309–1316. https://doi.org/10.1016/j.quascirev.2008.12.020
Tobler WR (1979) Cellular geography. In: Gale S, Olsson G (eds) Philosophy in geography. Springer, Dordrecht, pp 379–386
Trachsel M, Telford RJ (2016) Technical note: Estimating unbiased transfer-function performances in spatially structured environments. Clim Past 12(5):1215–1223. https://doi.org/10.5194/cp-12-1215-2016
Valavi R, Elith J, Lahoz-Monfort JJ, 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. https://doi.org/10.1111/2041-210X.13107
Veloz SD (2009) Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. J Biogeogr 36(12):2290–2299. https://doi.org/10.1111/j.1365-2699.2009.02174.x
Wadoux AMJ-C, Heuvelink GBM, de Bruin S, Brus DJ (2021) Spatial cross-validation is not the right way to evaluate map accuracy. Ecol Model 457:109692. https://doi.org/10.1016/j.ecolmodel.2021.109692
Wang Y, Khodadadzadeh M, Zurita-Milla R (2023) Spatial+: a new cross-validation method to evaluate geospatial machine learning models. Int J Appl Earth Obs Geoinform 121:103364. https://doi.org/10.1016/j.jag.2023.103364
Wenger SJ, Olden JD (2012) Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods Ecol Evol 3(2):260–267. https://doi.org/10.1111/j.2041-210X.2011.00170.x
West PW (2016) Simple random sampling of individual items in the absence of a sampling frame that lists the individuals. N Z J For Sci 46:15. https://doi.org/10.1186/s40490-016-0071-1
Wright MN, Ziegler A (2017) ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 77(1):1–17. https://doi.org/10.18637/jss.v077.i01
Yang DA, Laven RA (2021) Design-based approach for analysing survey data in veterinary research. Vet Sci. https://doi.org/10.3390/vetsci8060105
Yali A, Donald G (1997) Shape quantization and recognition with randomized trees. Neural Comput 9(7):1545–1588. https://doi.org/10.1162/neco.1997.9.7.1545
Zhang C, Chen Y, Xu B, Xue Y, Ren Y (2020) Improving prediction of rare species' distribution from community data. Sci Rep. https://doi.org/10.1038/s41598-020-69157-x
Zurell D, Elith J, Schröder B (2012) Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. Divers Distrib 18(6):628–634. https://doi.org/10.1111/j.1472-4642.2012.00887.x
Zurell D, Thuiller W, Pagel J, Cabral JS, Münkemüller T, Gravel D, Dullinger S, Normand S, Schiffers KH, Moore KA, Zimmermann NE (2016) Benchmarking novel approaches for modelling species range dynamics. Glob Change Biol 22(8):2651–2664. https://doi.org/10.1111/gcb.13251

Publisher's Note