Received: 20 September 2021 | Accepted: 8 March 2022

DOI: 10.1111/2041-210X.13851

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Nearest neighbour distance matching Leave-One-Out Cross-Validation for map validation

Carles Milà[1] | Jorge Mateu[2] | Edzer Pebesma[3] | Hanna Meyer[4]

[1]Barcelona Institute for Global Health (ISGlobal), Universitat Pompeu Fabra, CIBER Epidemiología y Salud Pública, Barcelona, Spain

[2]Universitat Jaume I, Castellón, Spain

[3]Institute of Geoinformatics, Westfälische Wilhelms-Universität Münster, Münster, Germany

[4]Institute of Landscape Ecology, Westfälische Wilhelms-Universität Münster, Münster, Germany

Correspondence
Carles Milà
Email: carles.mila@isglobal.org

Funding information
C.M. was supported by a PhD fellowship of the *Severo Ochoa Centre of Excellence* program awarded to ISGlobal. The work was further supported by the Federal Ministry of Economic Affairs and Energy of Germany (project number 50EE2009).

Handling Editor: Phil J Bouchet

Abstract

1. Several spatial and non-spatial Cross-Validation (CV) methods have been used to perform map validation when additional sampling for validation purposes is not possible, yet it is unclear in which situations one CV method might be preferred over the other. Three factors have been identified as determinants of the performance of CV methods for map validation: the prediction area (geographical interpolation vs. extrapolation), the sampling pattern and the landscape spatial autocorrelation. In this study, we propose a new CV strategy that takes the geographical prediction space into account, and test how the new method compares with other established CV methods under different configurations of these three factors.
2. We propose a variation of Leave-One-Out (LOO) CV for map validation, called Nearest Neighbour Distance Matching (NNDM) LOO CV, in which the nearest neighbour distance distribution function between the test and training data during the CV process is matched to the nearest neighbour distance distribution function between the target prediction and training points. Using random forest as a machine learning algorithm, we then examine the suitability of NNDM LOO CV as well as the established LOO (non-spatial) and buffered-LOO (bLOO, spatial) CV methods in two simulations with varying prediction areas, landscape autocorrelation and sampling distributions.
3. LOO CV provided good map accuracy estimates in landscapes with short autocorrelation ranges, or when estimating geographical interpolation map accuracy with randomly distributed samples. bLOO CV yielded realistic error estimates when estimating map accuracy in new prediction areas, but generally overestimated geographical interpolation errors. NNDM LOO CV returned reliable estimates in all scenarios we considered.
4. While LOO and bLOO CV provided reliable map accuracy estimates only in certain situations, our newly proposed NNDM LOO CV method returned robust estimates and generalised to LOO and bLOO CV whenever these methods were the most appropriate approach. Our work recognises the necessity of considering the geographical prediction space when designing CV-based methods for map validation.

KEYWORDS
Cross-Validation, map accuracy estimation, map validation, spatial point patterns, spatial prediction

1 | INTRODUCTION

Prediction of spatially structured variables at unsampled locations is crucial in ecology and the environmental sciences (Schmidt-Traub, 2021), where sampling methods are often complex and costly (e.g. forest inventories, Ploton et al., 2020; fixed meteorological stations, Meyer et al., 2016). An important step in a predictive mapping workflow is map validation, where the quality of the spatial predictions generated by the trained model is estimated. Although unbiased map accuracy estimates can be obtained via probability sampling and design-based inference (Wadoux et al., 2021), in many instances additional sampling for validation purposes is not possible. Furthermore, in the majority of ecological studies, field surveys are often conducted in opportunistic locations (Yates et al., 2018), might be subject to access restrictions, and are often biased towards populated or protected areas (Martin et al., 2012). This makes design-based inference hardly possible. Instead, model validation measures are frequently used as a proxy for map accuracy.

Model validation strategies commonly used include (repeated) train/test data splits or resampling techniques such as bootstrapping and Cross-Validation (CV) (Lyons et al., 2018). All of these methods rely on the key assumption of independence between train and test data to assess model generalisation to new, unobserved observations (Hastie et al., 2009). However, in the presence of spatial autocorrelation, independence between train and test data points may not hold. Nonetheless, standard CV (e.g. Leave-One-Out [LOO] and random k-fold CV) methods that ignore the locations of the training samples have been largely applied to spatial prediction problems, although the risk of obtaining invalid results has been frequently discussed (Misiuk et al., 2019; Ploton et al., 2020; Pohjankukka et al., 2017; Roberts et al., 2017; Telford & Birks, 2009; Wenger & Olden, 2012).

To overcome this issue, variations of traditional CV techniques have been proposed to minimise the spatial dependence between training and test points. One of them is spatial block k-fold CV, a variation on k-fold CV where folds are defined as contiguous blocks in geographical space (Wenger & Olden, 2012) using many possible configurations (Roberts et al., 2017; Valavi et al., 2019). Another method is buffered Leave-One-Out (bLOO) CV, which was introduced in the ecological literature (Telford & Birks, 2009) as an extension of LOO CV. Briefly, when holding out each of the points, observations within a given geographical radius of that point are also excluded from the training data (Le Rest et al., 2014). Radii have been suggested to be defined based on the estimated semi-variogram range of the model residuals (Le Rest et al., 2014; Telford & Birks, 2009), outcome (Roberts et al., 2017) or predictors (Valavi et al., 2019); or otherwise based on Moran's I correlogram (Karasiak et al., 2021).

Though frequently used, the use of CV strategies for map validation purposes remains controversial. Recent studies have made an important distinction between model and map validation (Stehman et al., 2021; Wadoux et al., 2021): while the former concerns itself with model transferability to new, unobserved observations, the latter aims to estimate the accuracy of the predictions of a model for a fixed and clearly defined area (e.g. a defined population of pixels) which may or may not be independent from the existing training samples. While existing CV approaches can be useful for model validation, their use for map validation may be inappropriate since they ignore the prediction space. Furthermore, the election of the most appropriate CV method for map validation has been found to be complex and dependent on several factors: the prediction area, the spatial autocorrelation of the landscape (set of outcome and predictors) and the sampling pattern.

Regarding the prediction area, it is important to distinguish the case where the sampling and prediction areas overlap (i.e. geographical interpolation) and the case where they do not (i.e. geographical extrapolation). Provided no covariate shifts are found between the training and prediction areas, spatial CV methods that impose independence between train and test data may be suitable to estimate map accuracy of a new prediction area where all prediction points will also be independent from the training data. However, they may not be adequate for interpolation in the geographical space, where at least a subset of the prediction locations in the vicinity of sampling points will not be independent from the training data (Misiuk et al., 2019; Roberts et al., 2017).

In relation to spatial autocorrelation, it has been suggested that standard CV methods tend to overestimate model transferability in landscapes with long autocorrelation ranges, while yielding reasonable estimates when ranges are short (Rocha et al., 2018). While most spatial modellers focus on residual spatial autocorrelation as a possible source of bias during CV and design spatial CV strategies to remove it (Le Rest et al., 2014; Ploton et al., 2020; Telford & Birks, 2009), recent studies suggest that spatial overfitting with non-causal, spatially structured predictors can also lead to a decreased model transfer ability (Fourcade et al., 2018) and erroneous CV analyses (Roberts et al., 2017), and hence recommend considering the outcome (Roberts et al., 2017) and predictor (Valavi et al., 2019) autocorrelation as well.

The importance of the sampling pattern has been examined in the recent study by Wadoux et al. (2021), where the authors showed that random k-fold CV is able to reasonably estimate the error of machine learning-interpolated surfaces with regular and randomly distributed training samples, but not if training samples are clustered. Indeed, several authors have pointed out that spatially clustered samples may lead to an overestimation of map accuracy when using standard CV methods for spatial interpolation problems, and recommend using spatial CV variants in that case (Hengl et al., 2018; Meyer et al., 2019; Misiuk et al., 2019). This is the case of many large-scale prediction tasks, where samples come from composite databases, often lack a common sampling design and are frequently clustered in

the geographical space (e.g. Van Den Hoogen et al., 2019). However, Wadoux et al. (2021) warned that even with clustered samples, spatial CV may still overestimate errors.

Finally, we argue that in order to obtain reliable map accuracy estimates for a given spatial prediction task, not only must these three factors (prediction area, spatial autocorrelation and sampling pattern) be taken into account, but also the geographical prediction space, which is currently ignored by both spatial and non-spatial CV methods, needs to be considered. Under spatial autocorrelation, the prediction error at a given location has been repetitively observed to follow an increasing function of the distance to its nearest training point (Brenning, 2021; Park et al., 2020; Ploton et al., 2020; Pohjankukka et al., 2017; Roberts et al., 2017; Telford & Birks, 2009). Thus, and similarly to time series CV methods where the prediction horizon is taken into account (Hyndman & Athanasopoulos, 2018), distances from the prediction locations to its nearest training sample should be matched during the CV process to obtain realistic map accuracy estimates for a given prediction task. This is illustrated in Figure 1, where nearest neighbour distances between test and training data during LOO CV (i.e. nearest neighbour distances of sampling points) are much shorter than distances from the prediction points to the nearest training point, which may trigger an overestimation of LOO CV accuracy if the landscape is subject to significant spatial autocorrelation.

In this study, we propose a new CV strategy suited for map validation that takes the geographical space of the target of the prediction task into account, and test how it compares to other established non-spatial and spatial CV strategies under different scenarios of prediction areas, landscape autocorrelation and sampling distributions.

2 | MATERIALS AND METHODS

2.1 | Nearest neighbour distance matching Leave-One-Out Cross-Validation

We propose a variation of LOO CV for map validation, called *Nearest Neighbour Distance Matching* (NNDM) LOO CV, in which the nearest neighbour distance distribution function (see below) between the test and training data during the CV process is matched to the nearest neighbour distance distribution function between the prediction and training points. This is done for those distances in which spatial autocorrelation is present, that is, distances shorter than the autocorrelation range (similar accuracy is expected for prediction points beyond it).

To characterise the distribution of nearest neighbour distances, we use concepts from the spatial point pattern field. First, to characterise the distribution of nearest neighbour distances between target $\mathbf{x}^{(i)}$ (Figure 2b) and sampling $\mathbf{x}^{(j)}$ points (Figure 2a) that is found during prediction, we use the empirical multitype nearest neighbour distance distribution function $\hat{G}_{ij}(r)$ (Baddeley et al., 2015), which expresses the proportion of prediction points that have a sampling point at a distance equal or lower than r, and is defined as:

$$\hat{G}_{ij}(r) = \frac{1}{n_i} \sum_{i} 1\{d_{ij} \leq r\},$$

where r is a distance, n_i is the total number of prediction points and $d_{ij} = \min \| x_i^{(i)} - x^{(j)} \|$. Here, no edge correction or stationarity assumptions are needed since $\hat{G}_{ij}(r)$ is simply used as a means to describe the nearest neighbour distance distribution of a set of fixed locations, rather than to perform an inferential spatial point pattern analysis of a realisation of a random point process. Moreover, for the common case where prediction points consist of a regular grid spanning all the sampling area independently of the locations of the training samples, $\hat{G}_{ij}(r)$ will be roughly equivalent to $\hat{F}_j(r)$, that is, the empirical empty space function of the sampling points. In our example in Figure 2c, $\hat{G}_{ij}(r)$ can be interpreted as follows: 50% of the prediction points have at least one training point at a distance equal or lower than 19.8 units.

Second, to characterise the distribution of the nearest neighbour distances during LOO CV where each of the points is held out sequentially, we can resort to the $\hat{G}_j(r)$ empirical nearest neighbour distance distribution function, which expresses the proportion of sampling points that have another sampling point at a distance equal or lower than r, and can be calculated as:

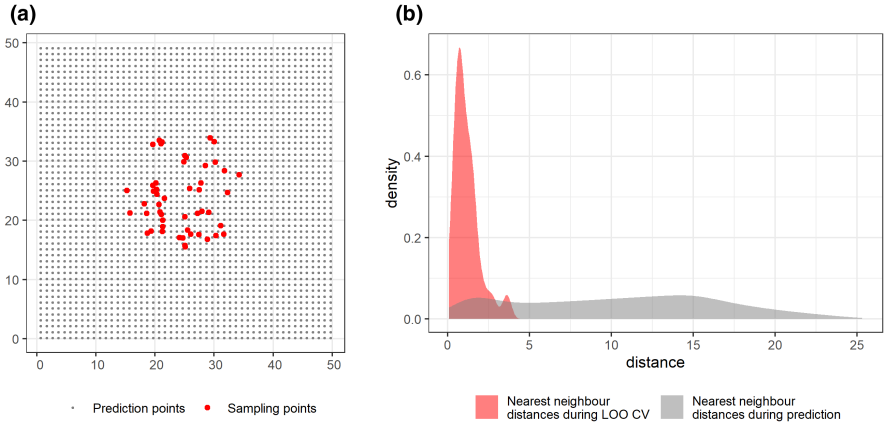


FIGURE 1 (a) 2,500 prediction points structured in a 50×50 regular grid, and 50 clustered sampling points. (b) Mismatch between the distribution of nearest neighbour distances between test and training data during LOO CV (i.e. nearest neighbour distances of sampling points) and prediction (i.e. distances from each prediction point to its nearest sampling point)

$$\hat{G}_j(r) = \frac{1}{n_j} \sum_j 1\{d_j \leq r\},$$

where r is a distance, n_j is the total number of sampling points and $d_j = \min_{k \neq j} \| x_j^{(j)} - x_k^{(j)} \|$. In our example in Figure 2c, $\hat{G}_j(r)$ can be interpreted as follows: 50% of the training points have at least another training point at a distance equal or lower than 7.9 units. Note that, in this example, there is a relevant mismatch between the nearest neighbour distance distribution function encountered during the prediction, that is, $\hat{G}_{ij}(r)$, and the one found during LOO CV, that is, $\hat{G}_j(r)$.

Our proposed NNDM LOO CV method retains the idea of LOO CV to validate each of the points in the training sample sequentially by holding them out one by one, but, similarly to bLOO CV, may exclude additional samples from the training set in each CV iteration. Let $\mathbf{L} = \left\{ \mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_{n_j} \right\}$ be a list of sets $\mathbf{l}_j$ containing the indices of the samples to fit the model to when holding out observation j during LOO CV. We can now define $\hat{G}_j^*(r, \mathbf{L})$ as the modified empirical nearest neighbour distance distribution function:

$$\hat{G}_j^*(r, \mathbf{L}) = \frac{1}{n_j} \sum_j 1\left\{ d_j^* \leq r \right\},$$

where $d_j^* = \min_{k \in \mathbf{l}_j} \| x_j^{(j)} - x_k^{(j)} \|$. Note that the only difference between d_j and d_j^* is that while the former considers all training points but j when computing the minimum distances, the latter considers the set of points included in $\mathbf{l}_j$. However, if $\forall j,\ \mathbf{l}_j = \{1, 2, \ldots, j-1, j+1, \ldots, n_j\}$, that is, $\mathbf{l}_j$ contains all training samples except the hold out point j, $d_j = d_j^*$, and hence $\hat{G}_j^*(r, \mathbf{L}) = \hat{G}_j(r)$. In Figure 2d, one can see how $\mathbf{L} = \{\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_2, \mathbf{l}_3, \mathbf{l}_4, \mathbf{l}_5, \mathbf{l}_6, \mathbf{l}_7, \mathbf{l}_8, \mathbf{l}_9, \mathbf{l}_{10}\}$ has been defined in this example, for example, $\mathbf{l}_3 = \{1,2,4,6,7,8,9\}$, that is, in the third iteration of NNDM LOO CV, we will fit a model with training samples 1,2,4,6,7,8,9 which will be used to assess out-of-sample accuracy of sample 3. To determine $\mathbf{L}$, we propose a simple greedy algorithm called *Nearest Neighbour Distance Matching* (NNDM) that, starting from LOO CV, excludes observations from $\mathbf{l}_j$ whenever by doing so $\hat{G}_j^*(r, \mathbf{L})$ matches better $\hat{G}_{ij}(r)$, for those r where autocorrelation is present (Algorithm 1). The inputs of the NNDM algorithm are the prediction points $\mathbf{x}^{(i)}$ and training points $\mathbf{x}^{(j)}$, and the autocorrelation range ϕ, which can be estimated as the range of the semi-variogram computed from the model residuals, or alternatively as the outcome range if the model is likely to be subject to spatial overfitting (Roberts et al., 2017).

Algorithm 1: Nearest Neighbour Distance Matching (NNDM) algorithm pseudo-code

Data: prediction points: $\mathbf{x}^{(i)}$, training points: $\mathbf{x}^{(j)}$, autocorrelation range: ϕ
Result: List of training point indices to be used in NNDM LOO CV: $\mathbf{L}$

initialize;
compute $\forall i, d_{ij} = \min \| x_i^{(i)} - x^{(j)} \|$;
define $\mathbf{L} = \{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_{n_j}\}$ where $\mathbf{l}_j = \{1, 2, \ldots, j-1, j+1, \ldots, n_j\}$;
compute $\forall j, d_j^* = \min_{k \in \mathbf{l}_j} \| x_j^{(j)} - x_k^{(j)} \|$;
compute $r_{\min} = \min d_j^*$;
compute $j_{\min} = \{j : d_j^* = r_{\min}\}$;
compute $k_{\min} = \{k : \| x_{j_{\min}}^{(j)} - x_{k \in \mathbf{l}_{j_{\min}}}^{(j)} \| = r_{\min}\}$;
while $r_{min} \leq \phi$ **do**
  if $(\hat{G}_j^*(r_{min}, \mathbf{L}) - 1/n_j) \geq \hat{G}_{ij}(r_{min})$ **then**
    remove $k_{\min}$ from $\mathbf{l}_{j_{\min}}$;
    update $\forall j, d_j^* = \min_{k \in \mathbf{l}_j} \| x_j^{(j)} - x_k^{(j)} \|$;
    update $r_{\min} = \min\{d_j^* : d_j^* \geq r_{\min}\}$;
    update $j_{\min} = \{j : d_j^* = r_{\min}\}$;
    update $k_{\min} = \{k : \| x_{j_{\min}}^{(j)} - x_{k \in \mathbf{l}_{j_{\min}}}^{(j)} \| = r_{\min}\}$;
  else
    update $r_{\min} = \min\{d_j^* : d_j^* > r_{\min}\}$;
    update $j_{\min} = \{j : d_j^* = r_{\min}\}$;
    update $k_{\min} = \{k : \| x_{j_{\min}}^{(j)} - x_{k \in \mathbf{l}_{j_{\min}}}^{(j)} \| = r_{\min}\}$;
return $\mathbf{L}$;
finalize;

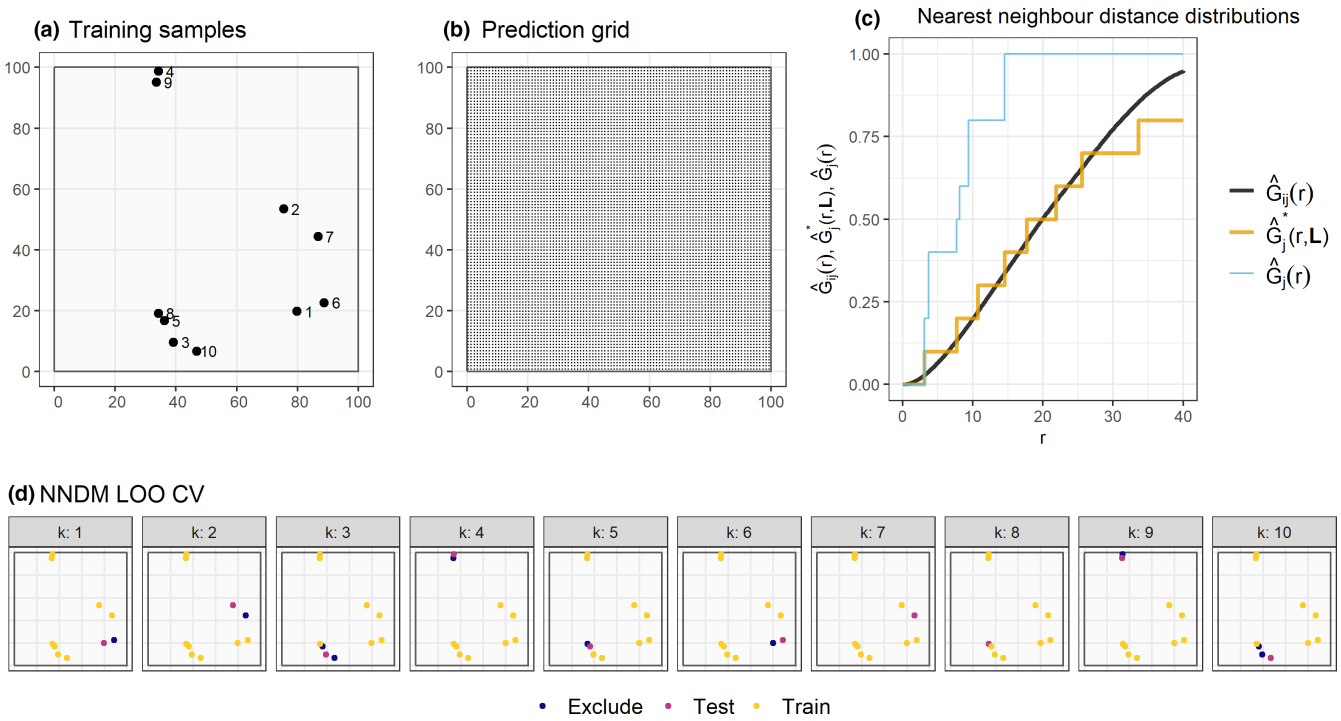


FIGURE 2 Simplified example of NNDM LOO CV: (a) 10 training samples, (b) 10,000 prediction points structured in a 100×100 grid, (c) $\hat{G}_{ij}(r)$ (i.e. nearest neighbour distance distribution function during prediction), $\hat{G}_j^*(r, \mathbf{L})$ (i.e. nearest neighbour distance distribution function during NNDM LOO CV), $\hat{G}_j(r)$ (i.e. nearest neighbour distance distribution function during LOO CV) functions and (d) view of NNDM LOO CV iterations. Here we assume the autocorrelation range parameter ϕ to be known and equal to 40

The NNDM algorithm starts by computing the distances d_{ij} between each prediction point and its nearest training point. It initially defines the set of training points to be used in each iteration of NNDM LOO CV, $\mathbf{L}$, as all but the hold out point (i.e. LOO CV). Then, it calculates the nearest neighbour distances d_j^* between each CV hold out sample and their respective CV training data $\mathbf{l}_j$, identifies the minimum among them $r_{\min} = \min d_j^*$, as well as the hold out point $j_{\min}$ and training sample $k_{\min} \in \mathbf{l}_{j_{\min}}$ where it is found. Then, we check if after removing $k_{\min}$ from $\mathbf{l}_{j_{\min}}$, $\hat{G}_j^*(r_{\min}, \mathbf{L})$ is greater or equal than $\hat{G}_{ij}(r_{\min})$. If it is, the point $k_{\min}$ is removed from $\mathbf{l}_{j_{\min}}$ and d_j^*, $r_{\min}$, $j_{\min}$, and $k_{\min}$ are updated; otherwise, we look for the next $r_{\min}$, update $j_{\min}$, $k_{\min}$ and keep iterating while $r_{\min}$ is smaller than or equal to the autocorrelation range ϕ.

Some points are worth noting: (a) If $\forall r \leq \phi$, $\hat{G}_j(r) \leq \hat{G}_{ij}(r)$ (e.g. regularly distributed samples), no point will be excluded from $\mathbf{l}_j$ and NNDM LOO CV will be equivalent to LOO CV; (b) if the autocorrelation range is very short or null $\phi \approx 0$, NNDM LOO CV will also be equivalent to LOO CV; and (c) if the minimum nearest neighbour distance between any of the prediction and training points, that is, $\min d_{ij}$ is larger than ϕ (e.g. the model is transferred to a new area), all points within a distance ϕ of each hold out point will be excluded as $\forall r <= \phi$, $\hat{G}_{ij}(r) = 0$, that is, NNDM LOO CV will be equivalent to bLOO CV with a radius equal to ϕ.

To illustrate these concepts, we simulated three sets of points with different distributions in a 100×100 prediction grid (i.e. 10,000 regular points) and estimated and compared the $\hat{G}_{ij}(r)$, $\hat{G}_j(r)$ and $\hat{G}_j^*(r, \mathbf{L})$ functions (Figure 3). Here, we assume that ϕ is known and equal to 20. A visual comparison of $\hat{G}_j(r)$ and $\hat{G}_{ij}(r)$ reveals that for random sampling patterns, the distribution of the nearest neighbour distances when performing LOO CV, that is, $\hat{G}_j(r)$, and when predicting the continuous surface, that is, $\hat{G}_{ij}(r)$, is very similar, which is not surprising as $\hat{G}_{ij}(r) \approx \hat{F}_j(r)$ in this case, and $G(r)$ and $F(r)$ are known to be equivalent under complete spatial randomness (Baddeley et al., 2015). On the other hand, we see that nearest neighbour distances during prediction are shorter than those in LOO CV for regularly distributed samples, while the opposite happens for clustered samples. Our proposed NNDM LOO CV method produces a distribution of nearest neighbour distances $\hat{G}_j^*(r, \mathbf{L})$ identical or very similar to LOO CV for random and regular sampling patterns and well matched to $\hat{G}_{ij}(r)$ in the clustered scenario.

2.2 | Simulation 1: Random fields

The first simulation compared the performance of LOO, bLOO and NNDM LOO CV for geographical interpolation and extrapolation map validation in simulated landscapes with varying autocorrelation range, and sampling patterns with different number of points and distributions (Figure 4). The input parameter of the simulation was the landscape autocorrelation range, which we set to a value of 1, 10, 20, 30 and 40 units. For each these values, we run 100 iterations of the simulation (i.e. 500 simulations iterations in total).

In each simulation iteration, we defined a 300×100 two-dimensional grid with a sampling area [0,100]×[0,100] and

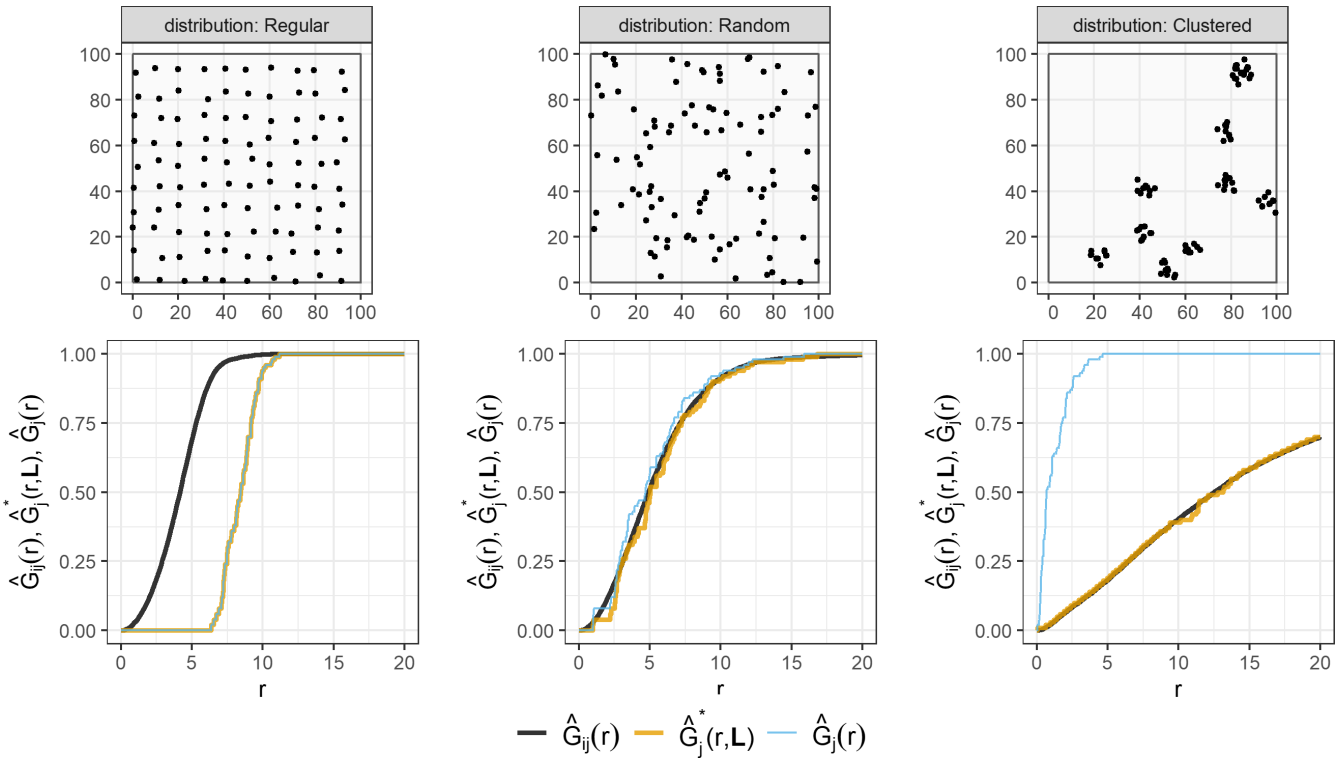


FIGURE 3 Three simulated sets of 100 points in a 100×100 grid (top) and their respective $\hat{G}_{ij}(r)$ (i.e. nearest neighbour distance distribution function during prediction), $\hat{G}^{*}_{ij}(r,\mathbf{L})$ (i.e. nearest neighbour distance distribution function during NNDM LOO CV) and $\hat{G}_{j}(r)$ (i.e. nearest neighbour distance distribution function during LOO CV) functions (bottom). Note that $\hat{G}^{*}_{j}(r,\mathbf{L}) = \hat{G}_{j}(r)$ in the regular case

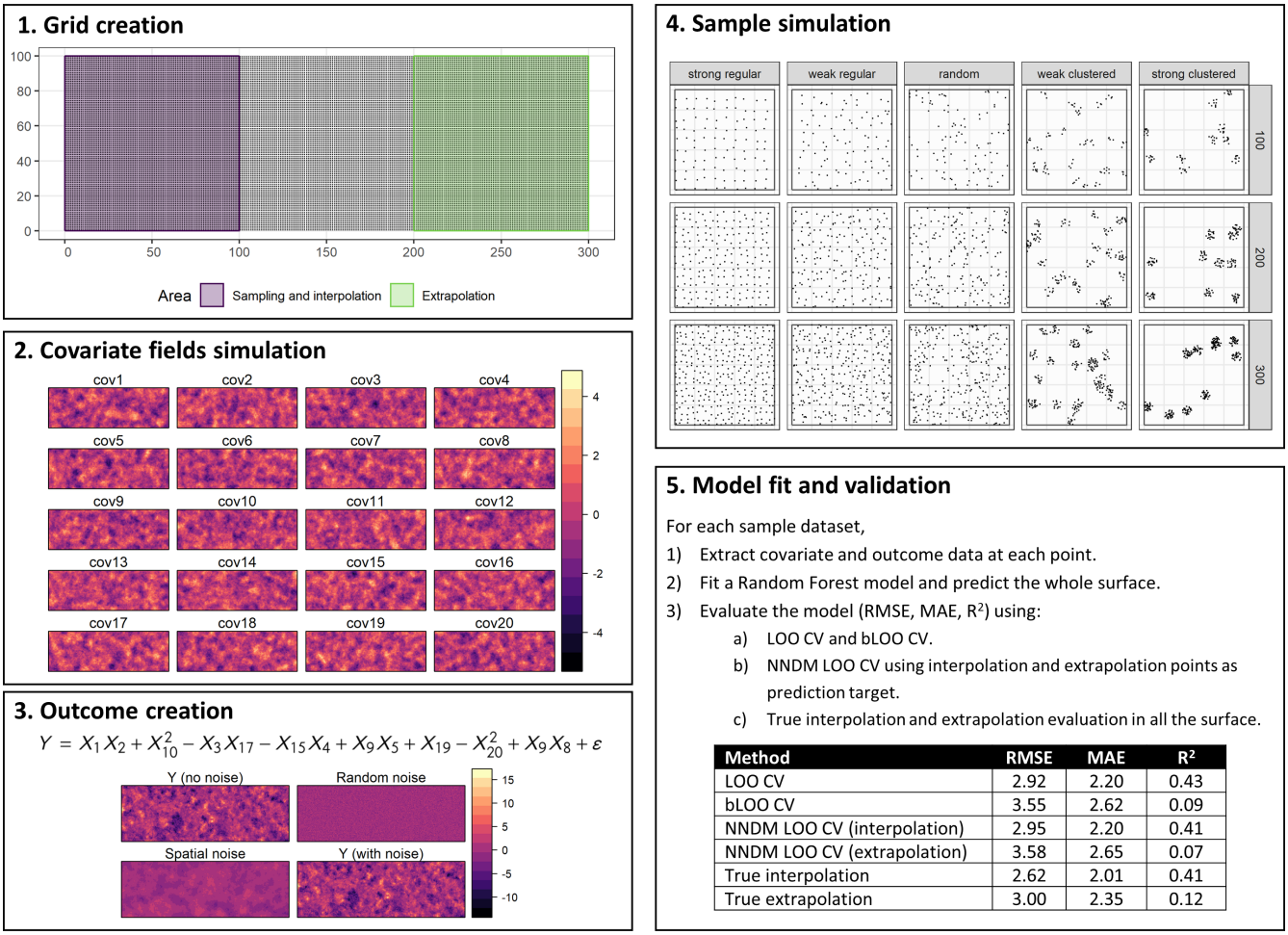
two distinct prediction areas: geographical interpolation [0,100]×[0,100], which coincided with the sampling area, and extrapolation [200,300]×[0,100]. We then simulated 20 independent covariate fields, defined as two-dimensional stationary and isotropic Gaussian random fields with a constant mean of 0 and subject to spatial autocorrelation with spherical semi-variogram with a sill of 1, a null nugget, and a range equal to the landscape autocorrelation range simulation parameter (Figure S1). Note that by assuming a constant mean in all the study area, we implicitly assumed that no significant covariate shifts were present.

To compute the outcome, we used the equation of simulation framework 2 from Van der Laan et al. (2007): $Y = X_1X_2 + X_{10}^2 - X_3X_{17} - X_{15}X_4 + X_9X_5 + X_{19} - X_{20}^2 + X_9X_8 + \varepsilon$ and defined a noise with a random and a spatially autocorrelated component $\varepsilon = \varepsilon_1 + \varepsilon_2$ where $\varepsilon_1 \sim N(0,1)$and ε_2 simulated using the same semi-variogram as the covariates. Random field simulations were carried out using the unconditional sequential Gaussian simulation algorithm (Gebbers & de Bruin, 2010).

Next, we simulated sampling points with three different sample sizes (N = 100,200,300) and five different sampling distributions (weak and strong regular, random, and weak and strong clustered), that is, 15 different sets of training points per simulation iteration, within the sampling area. Random samples were drawn by uniform random sampling; regular samples were obtained by jittering regular grids by 2 units (strong regular) or 5 units (weak regular); clustered samples were simulated by first drawing 10 (strong clustering) or 25 (weak clustering) n_{parents} points randomly, and then $\frac{n - n_{\text{parents}}}{n_{\text{parents}}}$ points in the 5-unit radius buffer of each parent point. Landscape data (outcome and 20 covariate fields) at the sampling points' locations were extracted.

With the extracted data at each of the 15 sampling patterns simulated in the previous step, a Random Forest (RF) model (Liaw & Wiener, 2002) was fit (mtry parameter fixed to 7) and used to predict the interpolation and extrapolation areas. We chose a RF model as it is one of the most popular ML models for spatial prediction (Wylie et al., 2019), and a large body of research on predictive mapping uses it (Hengl et al., 2018; Meyer et al., 2019; Meyer & Pebesma, 2021; Misiuk et al., 2019; Wadoux et al., 2021). Since the actual values of the outcome were known for the entire study area, we computed the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R^2 between all the gridded predicted and actual outcome values in the interpolation and extrapolation areas, separately (true map accuracy). Furthermore, we estimated map accuracy by computing the same three statistics when using (a) LOO CV, (b) bLOO CV, (c) NNDM LOO CV using the interpolation grid as prediction points and (d) NNDM LOO CV with the extrapolation grid as prediction points. Since in this simulation the landscape autocorrelation parameter was known, we used it as the buffer radius and ϕ parameter in bLOO and NNDM LOO CV, respectively.

To analyse the simulation results, we plotted the distribution of the RMSE, MAE and R^2 by validation method, sample size and distribution, landscape autocorrelation range, and prediction area across the 100 simulations iterations available for each combination of factors. To further assess whether CV methods were able to estimate the true map accuracy, we also subtracted the true from the CV RMSE, MAE and R^2 and plotted the distribution of these differences.



Method	RMSE	MAE	R²
LOO CV	2.92	2.20	0.43
bLOO CV	3.55	2.62	0.09
NNDM LOO CV (interpolation)	2.95	2.20	0.41
NNDM LOO CV (extrapolation)	3.58	2.65	0.07
True interpolation	2.62	2.01	0.41
True extrapolation	3.00	2.35	0.12

FIGURE 4 Simulation 1: Example of the workflow of one simulation iteration given a landscape autocorrelation of 20. bLOO: buffered LOO; CV, Cross-Validation; LOO: leave-one-out; MAE, Mean Absolute Error; NNDM, nearest neighbour distance matching; RMSE, Root Mean Square Error

2.3 | Simulation 2: Virtual species

We performed a second simulation study of a spatial prediction problem where this time covariate fields were observed rather than simulated, and bLOO and NDMM LOO CV autocorrelation range parameters were unknown and therefore needed to be estimated. The simulation used WorldClim bioclimatic variables (www.worldclim.org/data/bioclim.html) and the `R` package `virtualspecies` (Leroy et al., 2015) to create a virtual species suitability surface for the Iberian peninsula. Briefly, we downloaded grids for the 19 Worldclim bioclimatic variables at 2.5 min spatial resolution and cropped the data to the Iberian peninsula (defined as the area covered by continental Spain and Portugal, Figure 5a). We first generated a species suitability surface by taking the first and second dimensions of a principal component analysis on bioclimatic variables 1 (annual mean temperature), 2 (mean diurnal range), 4 (temperature seasonality), 5 (maximum temperature of warmest month), 6 (minimum temperature of the coldest month), 12 (annual precipitation), 13 (precipitation of the wettest month), 14 (precipitation of the driest month) and 15 (precipitation seasonality). Then, we assigned virtual species suitability values (scaled between 0 and 1) according to the product of two Gaussian probability density functions following a $N(0, 2)$ distribution in each axis (Figure S2). Data were projected into the ETRS89 UTM 30N coordinate reference system (EPSG: 25830). The resulting species suitability surface was treated as the outcome to be modelled.

Next, we simulated five sets of sampling points ($N = 100$) with different distributions (weak and strong regular, random, and weak and strong clustered; Figure 5b) with the same methods as in simulation 1 but changing the distance parameters for regular jittering (40 km for weak regular and 80 km for strong regular) and clustered sampling (80 km buffer radius). In all, 100 replications of the sampling process were done.

For each of the training points sets, we extracted the climate and suitability data and fitted a RF model to the habitat suitability outcome using all 19 bioclimatic covariates (mtry parameter fixed to 6), which was used to predict the suitability in the whole surface. Unlike simulation 1, since in this case the autocorrelation range parameters needed for both bLOO and NNDM LOO CV were not known, we estimated them by automatically fitting a spherical semi-variogram model to the

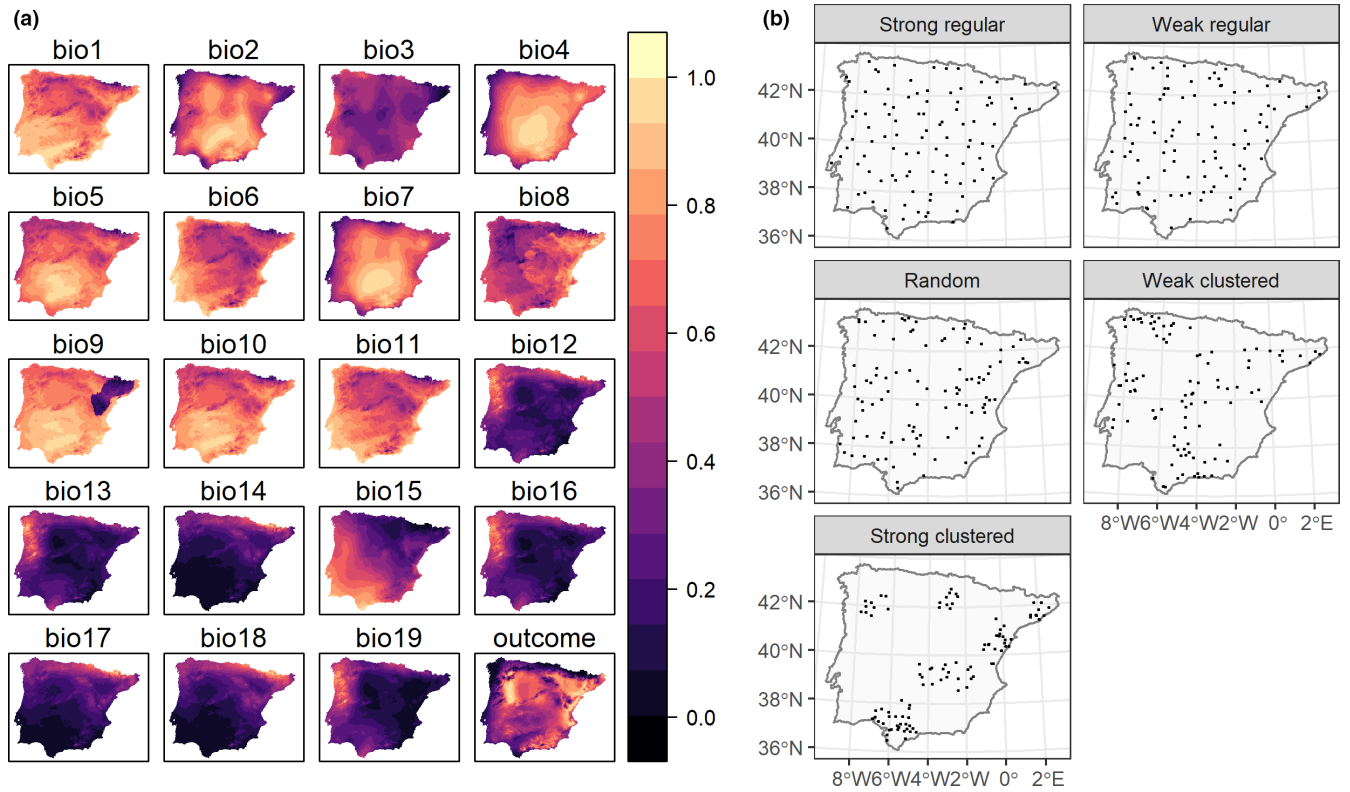


FIGURE 5 Simulation 2: (a) bioclimatic covariates and outcome species suitability (all linearly stretched to [0,1] for visualisation purposes); and (b) example of one iteration of the sample simulation

raw outcome and model residuals at the sample locations in each iteration, and then extracting the estimated ranges. Predictions were validated (a) comparing the actual vs. predicted suitability surfaces (i.e. true accuracy), (b) using LOO CV, and using (c) bLOO and (d) NNDM LOO CV with the estimated outcome and residual autocorrelation range as a parameter, that is, two different bLOO and NNDM LOO CV estimates were obtained. RMSE, MAE and R^2 were used as validation statistics. To analyse the results, we subtracted the true from the CV RMSE, MAE and R^2 and plotted the distribution of these differences by CV method and sample size and distribution.

2.4 | Implementation

All methods and simulation analyses were developed in R version 4.0 using several packages: `doParallel` (Corporation & Weston, 2019) for parallel computing, `randomForest` and `caret` (Kuhn, 2020; Liaw & Wiener, 2002) for RF model fitting, `sf` (Pebesma, 2018) and `raster` (Hijmans, 2019) for vector and raster data management, `gstat` (Pebesma, 2004) for random field simulation and semi-variogram fitting, and `tidyverse` (Wickham, 2017) for data management and graphics. Additional packages were used to execute other minor tasks. NNDM LOO and bLOO CV functions and classes have been implemented in an R package named `NNDM`, which is available at www.github.com/carlesmila/NNDM together with a worked example. The code used in the simulations included in this work, as well as the simulation data, are available at www.github.com/carlesmila/NNDMpaper.

3 | RESULTS

3.1 | Simulation 1: Random fields

Mean RMSEs across simulation scenarios (Figure 6, $N = 200$) showed that, while for short autocorrelation ranges most map validation strategies resulted in the same RMSE, as the range increased, RMSEs started to diverge, with increasing true extrapolation RMSEs and decreasing true interpolation RMSEs. bLOO and NNDM LOO CV (extrapolation area) had overlapping profiles and were considerably well aligned with the true extrapolation RMSE in all the scenarios, whereas LOO CV and NNDM LOO CV (interpolation area) were only equivalent for regular and random patterns but not clustered, where NNDM LOO CV approximated much better the true interpolation RMSE. Results for sample sizes equal to 100 and 300 points (Figures S3 and S4) and for the MAE and R^2 statistic (Figures S5 and S6) yielded similar conclusions.

Examination of the differences between the CV and the true RMSEs highlighted the fact that, for landscapes with almost null spatial autocorrelation, all CV methods were able to successfully estimate map accuracies (Figure 7). However, for longer autocorrelation ranges, distinct patterns were found in interpolation vs. extrapolation areas. In the interpolation area, LOO CV provided good estimates of the true RMSE under randomly distributed samples regardless of the range, yet slightly overestimated RMSEs for regularly distributed samples and strongly underestimated RMSEs for clustered samples; this pattern was more severe as the range

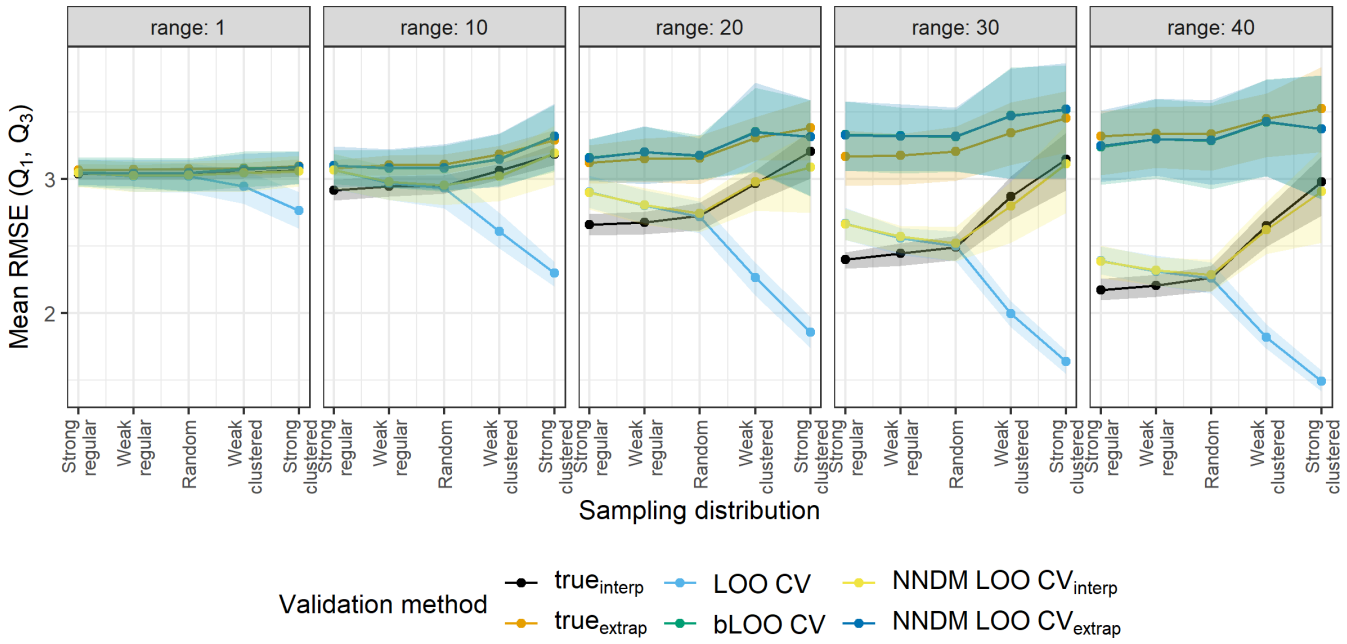


FIGURE 6 Simulation 1: Mean (Q_1, Q_3) RMSE by validation method, sampling distribution and landscape spatial autocorrelation range ($N = 200$). bLOO and LOO CV do not depend on the prediction area as they do not consider the geographical prediction space. Statistics have been calculated based on 100 simulation iterations. NNDM LOO CV_{extrap} and bLOO CV have the same profile and overlap in the graph

increased (mean (*SD*) differences for range = 30 were 0.27 (0.13), 0.01 (0.16) and −1.51 (0.32) for strong regular, random and strong clustered samples, respectively). bLOO CV overestimated RMSEs with larger variability in all sampling patterns (mean (*SD*) difference for range = 30 was 0.82 (0.34) for the random sampling pattern). NNDM LOO CV (interpolation area) gave identical or very similar results to LOO CV for regular and random patterns, yet corrected the error underestimation for clustered samples observed in LOO CV (NNDM LOO CV mean (*SD*) difference for range = 30 was −0.03 (0.63) for the strong clustered pattern).

In the extrapolation area, LOO CV underestimated the true RMSE for medium and long autocorrelation ranges regardless of the sample distribution. Both bLOO and NNDM LOO CV (extrapolation area) yielded identical and approximately correct RMSE estimates on average (mean (*SD*) differences for range = 30 and a weak clustered pattern were −1.35 (0.23), 0.13 (0.72) and 0.13 (0.72) for LOO CV, bLOO and NNDM LOO CV, respectively). Results were similar for samples sizes equal to 100 and 300 points (Figures S7 and S8), as well as when considering the MAE and R^2 statistic (Figures S9 and S10).

3.2 | Simulation 2: Virtual species

Results for LOO CV in simulation 2 (Figure 8) generally agreed with findings of simulation 1 for geographical interpolation: while LOO CV slightly overestimated the true RMSE in regular sampling patterns, it provided good estimates of the error in random patterns but considerably underestimated RMSEs for clustered samples (mean (*SD*) differences 0.01 (0.01), 0 (0.01), −0.07 (0.03), for strong regular, random and strong clustered sampling patterns, respectively). bLOO CV with radius equal to the outcome autocorrelation range yielded larger differences than bLOO CV with radius equal to the residual range (mean (*SD*) difference in weak clustered samples was 0.07 (0.04) for outcome range vs. 0.04 (0.04) for residual range), since estimated residual ranges were generally shorter than outcome ranges (median outcome range was 217 km vs. 127 km residual range). Nevertheless, both bLOO strategies generally overestimated true RMSEs.

Differences for NNDM LOO CV using either outcome or residual autocorrelation range were similar to each other and to LOO CV for regular and random samplings (mean (*SD*) difference in random samples was 0 (0.1) for both outcome and residual range). For weak clustered samplings, both provided reasonable estimates of the error while having smaller variability than their bLOO counterparts (NNDM LOO CV mean (*SD*) difference was −0.01 (0.02) for both outcome and residual range). In the strong clustered sampling patterns, slightly larger differences between the two could be observed (mean (*SD*) difference was −0.01 (0.04) for outcome range vs. −0.02 (0.04) in residual range). Similar findings were found for MAE and R^2 statistics (Figures S11 and S12).

4 | DISCUSSION

Our proposed NNDM LOO CV method takes the geographical prediction space into account by matching the distribution of nearest neighbour distances between test and training points during LOO CV to the distribution of nearest neighbour distances between the target and sampling points during prediction. In our two simulations with different prediction areas, landscape autocorrelation ranges and sampling distributions, our newly proposed NNDM LOO CV

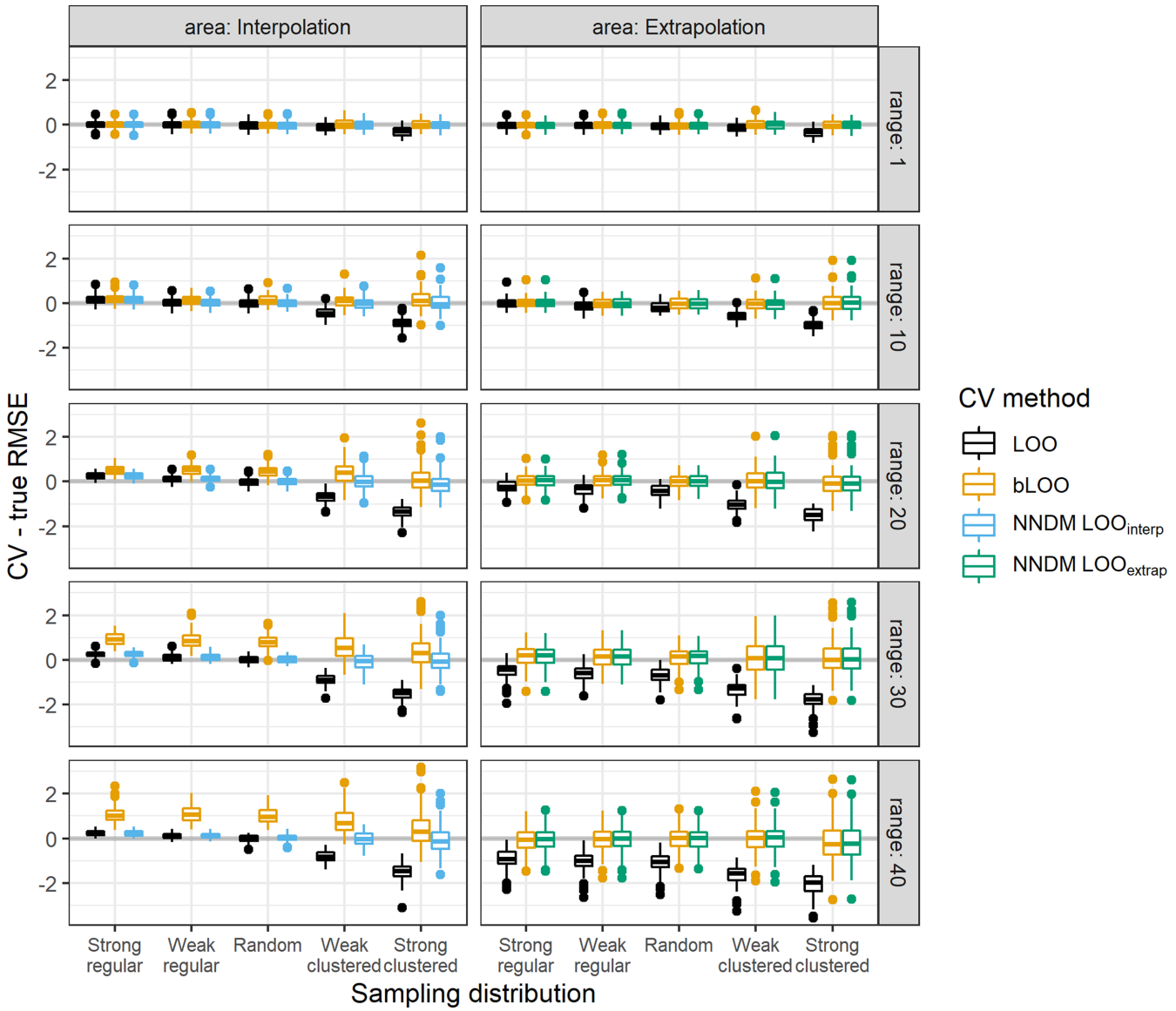
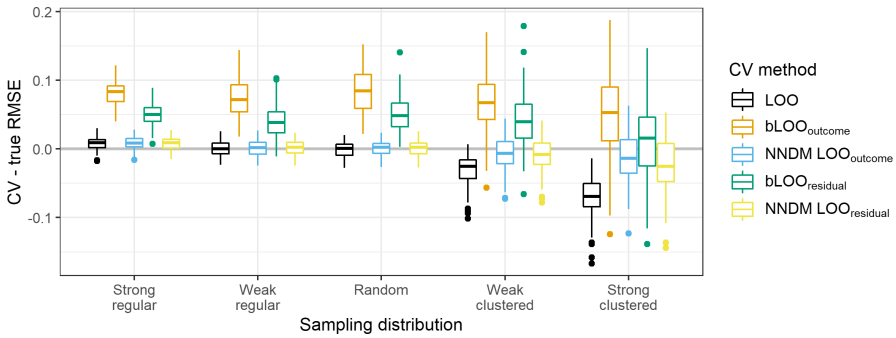


FIGURE 7 Simulation 1: Differences between CV and true interpolation and extrapolation RMSE by CV method, sampling distribution and landscape spatial autocorrelation range ($N = 200$). bLOO and LOO CV do not depend on the prediction area as they do not consider the geographical prediction space. Each boxplot consists of 100 data points resulting from 100 simulation iterations

FIGURE 8 Simulation 2: Differences between CV and true RMSE by CV method and sampling distribution. Two estimates for bLOO and NNDM LOO CV are provided depending on whether they use the residual or the outcome estimated autocorrelation range as a parameter. Each boxplot consists of 100 data points resulting from 100 simulation iterations



method provided robust estimates across all scenarios we considered while LOO and bLOO CV yielded reliable map accuracy estimates only in certain situations.

We discovered that LOO CV returned unbiased map accuracy estimates when (a) estimating geographical interpolation accuracy with random samples and (b) in landscapes with very short

autocorrelation range regardless of the sampling pattern and the prediction area. Regarding the first case, we found that, in geographical interpolation problems with random samples, the distribution of nearest neighbour distances during LOO CV matched well the distribution of nearest neighbour distances between prediction and sampling points. Nearest neighbour distances for regular/clustered samples during LOO CV were longer/shorter than nearest neighbour distances during prediction, resulting in map error over/underestimation. Similar patterns were found by Wadoux et al. (2021), where the RMSE estimated using random 10-fold CV under a random sampling pattern resulted in correct error estimates on average, while for clustered samples errors were strongly underestimated. As to the second case, in landscapes with almost null autocorrelation, the independence assumption of standard CV methods between train and test data, as well as between train and prediction data, will mostly hold, therefore returning reliable CV estimates. This is aligned with the findings of Rocha et al. (2018), who found that random k-fold CV yielded reasonable RMSE estimates only for short ranges when assessing transferability to new landscape realisations.

bLOO CV yielded reliable map accuracy estimates for prediction areas different than the sampling area, under the assumption of similar covariate space in the sampling and prediction areas. In agreement with the discussions of Roberts et al. (2017) and Misiuk et al. (2019), we found that spatial CV methods that force train and test data to be independent, such as bLOO, may be suited to assess map accuracy when significant geographical extrapolation is required, since the independence between the training and prediction points is required to be matched during the CV process. Following our simulation results and in agreement to the results of Wadoux et al. (2021), we cannot generally recommend using bLOO for map validation in spatial interpolation problems as it mostly overestimated prediction errors.

Our proposed NNDM LOO CV method was able to address the different prediction areas, landscape autocorrelation and sampling patterns we considered, and converged to LOO CV and bLOO whenever these were the most appropriate method: for interpolation problems with regular or random samples, as well as for short autocorrelation ranges, NNDM LOO CV was similar to, if not exactly the same, as LOO CV; for extrapolation problems, NNDM LOO CV produced the same estimates as bLOO CV. NNDM LOO CV was superior to all other methods in spatial interpolation problems with clustered samples. Some recommendations regarding NNDM LOO CV must be noted:

1. Although not so extreme as in bLOO CV, if training points are very clustered and the autocorrelation range is long, NNDM LOO CV may remove a large fraction of training data during CV and make models unstable. To address this, we recommend inspecting the summaries provided by the functions to detect such cases and use the minimum sample size parameter we have included in our implementation of the algorithm. Nevertheless, in extreme clustering cases, the nearest neighbour distance distribution function between the prediction and sampling points $\hat{G}_{ij}(r)$ may not be able to be matched during NNDM CV, which will be revealed with an examination of the estimated $\hat{G}$ functions.
2. As in bLOO, NNDM LOO can only correct instances where map accuracy is overestimated by removing points in the CV process. Therefore, methods to correct for map accuracy underestimation such as the case of regular samples in interpolation problems (other than collecting more samples) remain to be investigated. In such cases, we recommend researchers to consider their CV statistics to be conservative estimates.
3. Though not so critical as in bLOO CV, the estimation of the autocorrelation range parameter is still important in NNDM LOO CV. Given our results in simulation 2, we recommend researchers that, in case of doubt, they opt for a longer range estimate for geographical interpolation problems such as the outcome autocorrelation range as, unlike bLOO CV, NNDM LOO CV will still approximate the prediction nearest neighbour distance distribution function for all ranges below that threshold rather than excluding all data points.
4. The NNDM algorithm is greedy and prioritises matching the CV to the prediction nearest neighbour distance distribution function at short distances: starting from the shortest distance found during LOO CV to ϕ, it chooses whether to remove a training point during CV based only on a given step, which may lead to mismatches between the two functions for distances close to ϕ. Although having small deviations for long distances is not so crucial since spatial autocorrelation generally decays with distance, we always recommend a close examination of the estimated $\hat{G}$ functions provided in our implementation.
5. Similarly to bLOO, NNDM LOO CV only considers geographical distances to estimate map accuracy, but ignores the directions of those distances and the actual locations of the sampling and prediction points, and therefore may fail to account for anisotropy or lack of stationarity of the errors (Brenning, 2021).

As an additional recommendation, we suggest researchers to assess the spatial distribution of the sampling points to correctly interpret their CV results, especially if LOO or bLOO CV methods are used (NNDM LOO CV takes the sample distribution into account explicitly). Visual analyses (e.g. point and density mapping) and dedicated spatial point pattern tests (in particular those based on Monte Carlo envelopes) can be used to assess departure from complete spatial randomness under different null hypotheses and assumptions. An overview of these methods can be found in Baddeley et al. (2015).

The main limitation of our proposed method is that it only considers the geographical space and therefore implicitly assumes a similar covariate space in the training and prediction data (Roberts et al., 2017), that is, no large covariate shifts are to be found (Hoffimann et al., 2021). Even though this assumption may be reasonable in many instances where the sampling and prediction areas coincide, it is likely it does not hold when models are used to predict unsampled new areas, for example, spatial transfer of ecological models with large environmental dissimilarity

(Yates et al., 2018). This is the case of global mapping exercises with samples concentrated in some parts of the globe (e.g. Van Den Hoogen et al., 2019), wherein models fitted using data from a restricted set of locations are transferred to completely new environments. In that sense, spatially explicit methods able to detect extrapolation in the covariate space (Bouchet et al., 2020; Meyer & Pebesma, 2021) can be used to assess whether this is indeed the case. If it is, CV methods that consider the covariate space can be a useful alternative to purely spatial CV methods (Hoffimann et al., 2021; Roberts et al., 2017).

Our work has a number of aspects up for ongoing research. First, we only considered regression problems; however, evidence suggests that classification problems are as susceptible, if not more, to some of the same issues such as map error underestimation in geographical interpolation problems with clustered samples (Meyer et al., 2019; Misiuk et al., 2019). Second, our results are based on RF models and, while we believe that our results are applicable to other ML models, this needs to be confirmed. Third, we did not compare our new method to spatial CV strategies other than bLOO, for example, spatial block CV methods might offer a higher flexibility in the partition of the geographical space when compared to bLOO (Roberts et al., 2017). Fourth, and similarly to LOO CV, our proposed NNDM LOO CV method may not be computationally feasible for large samples. Therefore, a variation of k-fold CV that takes into account the prediction geographical space would be highly desired.

Our work benefits the predictive mapping community by providing a new CV method that, whenever additional probability sampling and design-based inference for validation purposes is not possible, can be used in a wide range of spatial prediction tasks to produce more reliable estimates of the quality of the generated predictions. Furthermore, it sheds light onto the factors that can affect the CV of a spatial prediction model and in which instances established spatial and non-spatial models can be used for map accuracy estimation. More generally, our work recognises the necessity of considering the geographical prediction space when designing CV-based methods for map validation.

ACKNOWLEDGEMENTS

We thank the three reviewers for their useful comments. We acknowledge support from the Spanish Ministry of Science and Innovation and State Research Agency through the 'Centro de Excelencia Severo Ochoa 2019-2023' Program (CEX2018-000806-S), and support from the Generalitat de Catalunya through the CERCA Program.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHORS' CONTRIBUTIONS

C.M. and H.M. conceived the ideas and designed the study, C.M. carried out the analysis, wrote the code and drafted the manuscript; J.M., E.P. and H.M. critically contributed to discussions and drafts and gave final approval for publication.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13851.

DATA AND CODE AVAILABILITY

NNDM LOO and bLOO CV functions and classes, and a worked example of their use, are available in the NNDM package available at www.github.com/carlesmila/NNDM and archived in Zenodo (Milà, 2022b) at https://doi.org/10.5281/zenodo.6366981. The code used in the simulations included in this work, as well as the simulation data, is available at www.github.com/carlesmila/NNDMpaper and archived in Zenodo (Milà, 2022a) at https://doi.org/10.5281/zenodo.6366985.

ORCID

Carles Milà https://orcid.org/0000-0003-0470-0760
Jorge Mateu https://orcid.org/0000-0002-2868-7604
Edzer Pebesma https://orcid.org/0000-0001-8049-7069
Hanna Meyer https://orcid.org/0000-0003-0556-0210

REFERENCES

Baddeley, A., Rubak, E., & Turner, R. (2015). *Spatial point patterns: Methodology and applications with R*. CRC Press.

Bouchet, P. J., Miller, D. L., Roberts, J. J., Mannocci, L., Harris, C. M., & Thomas, L. (2020). Dsmextra: Extrapolation assessment tools for density surface models. *Methods in Ecology and Evolution*, *11*(11), 1464–1469. https://doi.org/10.1111/2041-210X.13469

Brenning, A. (2021). Spatial machine-learning model diagnostics: A model-agnostic distance-based approach. *arXiv preprint arXiv:2111.08478*.

Corporation, M., & Weston, S. (2019). *doParallel: Foreach parallel adaptor for the 'parallel' package*. R package version 1.0.15. Retrieved from https://CRAN.R-project.org/package=doParallel

Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, *27*(2), 245–256. https://doi.org/10.1111/geb.12684

Gebbers, R., & de Bruin, S. (2010). Application of Geostatistical simulation in precision agriculture. In M. A. Oliver (Ed.), *Geostatistical applications for precision agriculture* (pp. 269–303). Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

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

Hijmans, R. J. (2019). *Raster: Geographic data analysis and modeling*. R package version 2.9-23. Retrieved from https://CRAN.R-project.org/package=raster

Hoffimann, J., Zortea, M., de Carvalho, B., & Zadrozny, B. (2021). Geostatistical learning: Challenges and opportunities. *Frontiers in Applied Mathematics and Statistics*, *7*, 40. https://doi.org/10.3389/fams.2021.689393

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.

Karasiak, N., Dejoux, J.-F., Monteil, C., & Sheeren, D. (2021). Spatial dependence between training and test sets: Another pitfall of classification accuracy assessment in remote sensing. *Machine Learning*, 1–25. https://doi.org/10.1007/s10994-021-05972-1

Kuhn, M. (2020). *Caret: Classification and regression training*. R package version 6.0-86. Retrieved from https://CRAN.R-project.org/package=caret

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. *Global Ecology and Biogeography*, *23*(7), 811–820. https://doi.org/10.1111/geb.12161

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2015). Virtualspecies, an r package to generate virtual species distributions. *Ecography*, *39*, 599–607.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18–22.

Lyons, M. B., Keith, D. A., Phinn, S. R., Mason, T. J., & Elith, J. (2018). A comparison of resampling methods for remote sensing classification and accuracy assessment. *Remote Sensing of Environment*, *208*, 145–153.

Martin, L. J., Blossey, B., & Ellis, E. (2012). Mapping where ecologists work: Biases in the global distribution of terrestrial ecological observations. *Frontiers in Ecology and the Environment*, *10*(4), 195–201.

Meyer, H., Katurji, M., Appelhans, T., Müller, M. U., Nauss, T., Roudier, P., & Zawar-Reza, P. (2016). Mapping daily air temperature for Antarctica based on modis lst. *Remote Sensing*, *8*(9), 732–748.

Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, *12*, 1620–1633. https://doi.org/10.1111/2041-210X.13650

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. *Ecological Modelling*, *411*, 108815.

Milà, C. (2022a). Nearest neighbour distance matching leave-one-out cross-validation for map validation: Article code and data. *Zenodo*, https://doi.org/10.5281/zenodo.6366985

Milà, C. (2022b). Nndm r package. *Zenodo*, https://doi.org/10.5281/zenodo.6366981

Misiuk, B., Diesing, M., Aitken, A., Brown, C. J., Edinger, E. N., & Bell, T. (2019). A spatially explicit comparison of quantitative and categorical modelling approaches for mapping seabed sediments using random forest. *Geosciences*, *9*(6), 254.

Park, Y., Kwon, B., Heo, J., Hu, X., Liu, Y., & Moon, T. (2020). Estimating pm2.5 concentration of the conterminous United States via interpretable convolutional neural networks. *Environmental Pollution*, *256*, 113395.

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

Pebesma, E. J. (2004). Multivariable geostatistics in s: The gstat package. *Computers & Geosciences*, *30*(7), 683–691.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., & Bayol, N. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, *11*(1), 1–11.

Pohjankukka, J., Pahikkala, T., Nevalainen, P., & Heikkonen, J. (2017). Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, *31*(10), 2001–2019. https://doi.org/10.1080/13658816.2017.1346255

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

Rocha, A. D., Groen, T. A., Skidmore, A. K., Darvishzadeh, R., & Willemen, L. (2018). Machine learning using hyperspectral data inaccurately predicts plant traits under spatial dependency. *Remote Sensing*, *10*(8), 1263–1283.

Schmidt-Traub, G. (2021). National climate and biodiversity strategies are hamstrung by a lack of maps. *Nature Ecology & Evolution*, *5*, 1325–1327.

Stehman, S. V., Pengra, B. W., Horton, J. A., & Wellington, D. F. (2021). Validation of the u.s. geological survey's land change monitoring, assessment and projection (lcmap) collection 1.0 annual land cover products 1985–2017. *Remote Sensing of Environment*, *265*, 112646.

Telford, R., & Birks, H. (2009). Evaluation of transfer functions in spatially structured environments. *Quaternary Science Reviews*, *28*(13), 1309–1316.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2019). Blockcv: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, *10*(2), 225–232.

Van Den Hoogen, J., Geisen, S., Routh, D., Ferris, H., Traunspurger, W., Wardle, D. A., De Goede, R. G., Adams, B. J., Ahmad, W., Andriuzzi, W. S., Bardgett, R. D., Bonkowski, M., Campos-Herrera, R., Cares, J. E., Caruso, T., de Brito Caixeta, L., Chen, X., Costa, S. R., Creamer, R., … Crowther, T. W. (2019). Soil nematode abundance and functional group composition at a global scale. *Nature*, *572*(7768), 194–198.

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1). https://doi.org/10.2202/1544-6115.1309

Wadoux, A. M.-C., Heuvelink, G. B., de Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, *457*, 109692.

Wenger, S. J., & Olden, J. D. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, *3*(2), 260–267.

Wickham, H. (2017). *Tidyverse: Easily install and load the 'Tidyverse'*. R package version 1.2.1. Retrieved from https://CRAN.R-project.org/package=tidyverse

Wylie, B. K., Pastick, N. J., Picotte, J. J., & Deering, C. A. (2019). Geospatial data mining for digital raster mapping. *GIScience & Remote Sensing*, *56*(3), 406–429.

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., Fielding, A. H., Bamford, A. J., Ban, S., Barbosa, A. M., Dormann, C. F., Elith, J., Embling, C. B., Ervin, G. N., Fisher, R., Gould, S., Graf, R. F., Gregr, E. J., Halpin, P. N., … Sequeira, A. M. (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution*, *33*(10), 790–802.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Milà, C., Mateu, J., Pebesma, E. & Meyer, H. (2022). Nearest neighbour distance matching Leave-One-Out Cross-Validation for map validation. *Methods in Ecology and Evolution*, *13*, 1304–1316. https://doi.org/10.1111/2041-210X.13851