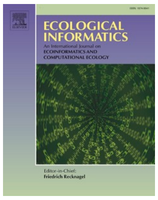
Ecological Informatics 69 (2022) 101665

Contents lists available at ScienceDirect

Ecological Informatics

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



Dealing with clustered samples for assessing map accuracy by cross-validation

Sytze de Bruin [a,*], Dick J. Brus [b], Gerard B.M. Heuvelink [c], Tom van Ebbenhorst Tengbergen [a], Alexandre M.J-C. Wadoux [d]

[a] Wageningen University & Research, Laboratory of Geo-information Science and Remote Sensing, PO Box 47, 6700AA Wageningen, the Netherlands
[b] Wageningen University & Research, Mathematical and Statistical Methods – Biometris, PO Box 16, 6700AA Wageningen, the Netherlands
[c] Wageningen University & Research, Soil Geography and Landscape Group, PO Box 47, 6700AA Wageningen, the Netherlands
[d] Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Australia

ARTICLE INFO

Keywords:
Machine learning
Spatial cross-validation
Spatial autocorrelation
Soil organic carbon
Above-ground biomass

ABSTRACT

Mapping of environmental variables often relies on map accuracy assessment through cross-validation with the data used for calibrating the underlying mapping model. When the data points are spatially clustered, conventional cross-validation leads to optimistically biased estimates of map accuracy. Several papers have promoted spatial cross-validation as a means to tackle this over-optimism. Many of these papers blame spatial autocorrelation as the cause of the bias and propagate the widespread misconception that spatial proximity of calibration points to validation points invalidates classical statistical validation of maps. We present and evaluate alternative cross-validation approaches for assessing map accuracy from clustered sample data. The first method uses inverse sampling-intensity weighting to correct for selection bias. Sampling-intensity is estimated by a two-dimensional kernel approach. The two other approaches are model-based methods rooted in geostatistics, where the first assumes homogeneity of residual variance over the study area whilst the second accounts for heteroscedasticity as a function of the sampling intensity. The methods were tested and compared against conventional *k*-fold cross-validation and blocked spatial cross-validation to estimate map accuracy metrics of above-ground biomass and soil organic carbon stock maps covering western Europe. Results acquired over 100 realizations of five sampling designs ranging from non-clustered to strongly clustered confirmed that inverse sampling-intensity weighting and the heteroscedastic model-based method had smaller bias than conventional and spatial cross-validation for all but the most strongly clustered design. For the strongly clustered design where large portions of the maps were predicted by extrapolation, blocked spatial cross-validation was closest to the reference map accuracy metrics, but still biased. For such cases, extrapolation is best avoided by additional sampling or limitation of the prediction area. Weighted cross-validation is recommended for moderately clustered samples, while conventional random cross-validation suits fairly regularly spread samples.

1. Introduction

Maps of environmental variables such as above-ground biomass, soil carbon stock and land cover change are essential information sources for assessing global carbon fluxes and to support climate change mitigation actions (Fitts et al., 2021; Harris et al., 2021). Such maps are commonly produced by machine learning approaches using spatially exhaustive Earth observation imagery, climatic data and terrain variables derived from digital elevation models as predictors (e.g. Du et al., 2020; Grabska et al., 2020; Li et al., 2020; Morais et al., 2021; Poggio et al., 2021; Sanderman et al., 2018). It is widely acknowledged that maps resulting from model predictions are not error-free and need proper accuracy assessment (Ploton et al., 2020; Stehman and Foody, 2019; Wadoux et al., 2021).

Classical map accuracy assessment is rooted in sampling theory wherein an unbiased estimate of map accuracy (e.g. mean squared map error) is obtained by design-based inference from a probability sample (de Gruijter and ter Braak, 1990; Stehman, 2009). In practice, post-

[*] Corresponding author at: Wageningen University & Research, Laboratory of Geo-information Science and Remote Sensing, PO Box 47, 6700AA Wageningen, the Netherlands.

E-mail address: sytze.debruin@wur.nl (S. de Bruin).

https://doi.org/10.1016/j.ecoinf.2022.101665
Received 12 February 2022; Received in revised form 29 April 2022; Accepted 29 April 2022
Available online 5 May 2022

mapping probability samples that are exclusively used for map evaluation are often not available and therefore alternative methods have been proposed. In machine learning, if data are abundant, a common approach is to randomly divide the full dataset used for modelling into three parts: a training set, a validation set, and a test set (Hastie et al., 2009, Chapter 7). The training set is used for fitting the models, the validation set is used to estimate prediction error for model selection and hyperparameter tuning, while the test set is used for assessing the accuracy of the final model. This paper addresses this latter testing phase, with the specific aim to assess the accuracy of a thematic map produced by a calibrated statistical prediction method. Data availability is often limited so that setting aside a test set cannot always be afforded and therefore resampling methods are used (Hastie et al., 2009; Steele et al., 2003). To this end, the widely used k-fold cross-validation method splits the full dataset into k approximately equally-sized disjoint subsets or folds, where repeatedly (i.e. k times) the model is calibrated on k-1 folds, whilst the remaining fold is used for assessing prediction accuracy. The overall cross-validation accuracy is estimated by aggregating the (squared) residuals over the k folds. In conventional k-fold cross-validation, the folds are chosen randomly.

If the full sample dataset is acquired by simple random sampling and if k equals the sample size (i.e. leave-one-out cross-validation, hereafter LOOCV), estimation from cross-validation is known to be nearly unbiased (Bengio and Grandvalet, 2004; Krzanowski, 2001; Steele et al., 2003). Since the computational burden of LOOCV is heavy, k is conventionally set to five or ten, in which case bias is no longer negligible but can be parametrically corrected (Fushiki, 2011). However, the sample datasets used for mapping environmental variables generally are not acquired by simple random sampling. Rather, they are amalgamations of several datasets, each with its own formal or opportunistic sampling design, which strictly impedes design-based inference. The latter also applies to probability samples whose inclusion probabilities are not published, such as the LUCAS dataset (d'Andrimont et al., 2020). If the data points are fairly uniformly distributed in space, conventional k-fold cross-validation likely produces reasonable results but the estimates of the map accuracy metrics may be biased and no confidence intervals can be derived. In contrast, strongly clustered points in the compound datasets may not be representative of the entire study area as they overrepresent some regions while underrepresenting or even missing others. This implies that the machine learning models are most intensely trained on the densely sampled areas which also have the largest impact on the estimated map accuracy. Conventional cross-validation map accuracy estimates based on such preferential samples tend to be too optimistic and methods are needed to correct for that (Steele et al., 2003).

Several papers including Brenning (2012), Le Rest et al. (2014), Roberts et al. (2017), Just et al. (2020) and Ploton et al. (2020) address the over-optimistic accuracy estimates by promoting methods collectively known as spatial cross-validation. These methods start from the premise that spatial proximity of data points in the calibration and test data folds is to be avoided. This is commonly achieved by spatial blocking in k-fold CV or spatial buffering in LOOCV (Le Rest et al., 2014; Roberts et al., 2017). However, the underlying assertion is incorrect as it ignores that there are many prediction points close to calibration points. Hence, spatial cross-validation tends to produce overly pessimistic validation results. Milà et al. (2022) proposed a spatial buffering filtering for LOOCV that matches the nearest neighbour distance distribution function between the test and training data to that between the prediction and training points.

The motivation for spatial cross-validation further seems to be rooted in a persistent confusion about the meaning of independence in design-based and model-based approaches (Brus, 2021; de Gruijter and ter Braak, 1990). In the case of design-based accuracy estimation from a probability sample, the estimators and their associated variance estimators are unbiased, regardless of the magnitude of spatial correlation and test locations being close to calibration locations (Gregoire and Valentine, 2007; Stehman and Foody, 2019). With Wadoux et al. (2021) we therefore object against rejecting long-standing, statistically valid methods for assessing map accuracy.

Since the distribution of most —if not all— environmental properties is spatially structured, having zero sampling density in portions of geographic space often implicates the risk of failure to cover portions of feature space. Therefore, we concur with Meyer et al. (2019), Just et al. (2020), de Sousa et al. (2021), Helmstetter et al. (2021) and others that strong data clustering in geographic space may complicate the identification of predictive models and may result in making predictions beyond the feature space covered by the sample. Undeniably, conventional cross-validation cannot assess the accuracy of extrapolations but blocked spatial cross-validation will not solve this problem either. Instead, one may consider assessing the disparity between feature data at prediction sites and those in the training dataset to delineate a prediction model's area of applicability (AOA), as proposed by Meyer and Pebesma (2021). In contrast, the current work aims to tackle bias in map accuracy estimates from cross-validation by balancing the impact of residuals in the regions with different sampling intensities. Otherwise, residuals in densely sampled regions dominate the accuracy estimates and this effect arises even if the clustered sample comprises the full feature space and, hence, no extrapolation occurs.

The objectives of this work are to present and evaluate alternative cross-validation approaches for assessing thematic map accuracy when the sample data are clustered. The first proposed method is a quasi-randomization approach (Elliott and Valliant, 2017) using inverse sampling-intensity weighting to correct for selection bias by giving more weight to observations in sparsely sampled areas and less weight to observations in densely sampled areas. Sampling intensity is estimated from the observational data points by a two-dimensional kernel approach (cf. Cronie and Van Lieshout, 2018; Deutsch, 1989). The two other approaches are model-based methods rooted in geostatistics (e.g. Goovaerts, 1997; Isaaks and Srivastava, 1990). These account for redundant information of spatially clustered residuals by using spatial correlation functions (variograms). Estimates of the global map accuracy metrics are obtained by simulating residuals on every node of a grid covering the entire study area. The first variant assumes homoscedasticity of the residuals, whilst the second accounts for heteroscedasticity of the residuals, again as a function of sampling intensity. We explain how these methods work and compare their map accuracy estimates and those from conventional k-fold cross-validation and blocked spatial cross-validation against reference map accuracy metrics. Genuine demonstration of our approach would require true values of the target variable to be known everywhere, which is unfeasible in reality. To mimic a situation of omnipresent reference data, the target environmental variables were sampled from existing above ground biomass (AGB) and soil organic carbon stock (OCS) maps and the acquired samples were used for fitting and prediction with random forest models.

2. Methods

2.1. Cross-validation methods

2.1.1. Conventional (method 1)

Five cross-validation methods were compared using different samples, the details of which are provided in Section 2.2. In our implementation of *conventional* 10-fold cross-validation (k = 10), the full sample dataset was randomly split 100 times into ten equally sized disjoint folds, each time providing a different folding of the full sample. Each time, nine of the folds were used for calibrating the model, and the remaining fold for testing. The map accuracy metrics (see Section 2.3) were computed from the n pairs of reference observations and map predictions (n being the sample size) distributed over the ten folds (Hastie et al., 2009). The metrics were averaged over the 100 foldings.

2.1.2. Spatial (method 2)

Blocked *spatial* cross-validation was implemented using Brennings' (2012) sperrorest package for the R language and environment for statistical computing (R Core Team, 2021). In this method, spatial partitions of the sample are created by *k*-means ($k = 10$) clustering based on the spatial coordinates. Following Brenning's (2012) implementation, these sample partitions were allowed to differ somewhat in size. Computation of the map accuracy metrics over the ten folds proceeded as in the conventional cross-validation. Again, we averaged over 100 repetitions, each producing a somewhat different spatial partitioning. The maximum number of iterations for the *k*-means coordinate clustering procedure was set to 50, which was found to be sufficient for convergence.

2.1.3. Inverse sampling-intensity weighted (method 3)

Thirdly, we used a heuristic, quasi-randomization method (Elliott and Valliant, 2017) employing inverse sampling-intensity weighting to give more weight to observations in sparsely sampled areas and less weight to observations in densely sampled areas to correct for estimation bias. Sampling intensities of the dataset were estimated using a two-dimensional kernel approach where the kernel width was computed from the sample using the Cronie and van Lieshout (2018) criterion as implemented in the spatstat package for R (Baddeley et al., 2016). The map accuracy metrics were computed by weighting the squared residuals obtained from conventional random cross-validation by the inverse of the estimated sampling intensity (details are in Section 2.3). The RMSE estimator bears some cursory similarity with the Horvitz–-Thompson estimator of map accuracy metrics in a stratified sample (Cochran, 1977; Lohr, 2019) but note that here the weights are obtained from estimated sampling intensities, which can largely differ from the inclusion probabilities of stratified sampling designs. From here on we refer to this method as the weighted cross-validation method.

2.1.4. Homoscedastic model-based (method 4)

Next, we used two variants of a model-based method, which were again applied to the residuals obtained from conventional cross-validation (Section 2.1.1). Using kriging weights, these methods account for redundant information in spatially clustered residuals based on their spatial configuration and autocorrelation as characterised by a variogram model. The first variant assumes stationarity of the variance (*homoscedasticity*) of the residuals $r(s_i) = z(s_i) - \widehat{z}_m(s_i)$, with $z(s_i)$ denoting a reference observation of the target variable at sample location s_i, and $\widehat{z}_m(s_i)$ being the map prediction of the target variable at s_i. The residuals $r(s_i)$ are seen as local realizations of a random field $R = \{R(s), s \in D\}$ over the entire study area D, modelled by (Eq. (1)):

$$R(s) = \mu + \varepsilon_{hom}(s) \tag{1}$$

where μ is a fixed but unknown spatial mean error of the target variable map (map bias) and ε_{hom} denotes a stationary, zero mean spatially correlated Gaussian random field that is conditioned on the residuals at the sample locations and whose spatial correlation is characterised by a variogram model. The variogram models were acquired by fitting a permissible function through experimental semivariances computed from the residuals obtained by conventional cross-validation. Next, 500 maps of the residual fields R were generated by sequential conditional Gaussian simulation on a dense grid with an ordinary kriging model, using the gstat package (Gräler et al., 2016; Pebesma, 2004). Note that ordinary kriging includes prediction of the unknown μ in Eq. (1) by $\widehat{\mu}_{OK}$. The procedure was repeated for each sample (see Section 2.2) and for both target variables (see Section 2.4). For each of the 500 gridded simulated fields the map accuracy metrics were computed (see Section 2.3), after which the means were computed from the sampling distributions.

2.1.5. Heteroscedastic model-based (method 5)

The *heteroscedastic* model-based method employs a fairly simple approach for modelling non-stationarity of the variance of the residuals using (Eq. (2)):

$$R(s) = \mu + \sigma(s) \bullet \varepsilon_{\text{het}}(\text{s}) \tag{2}$$

where μ denotes the unknown map bias as in Eq. (1), σ is a deterministic field mapping the standard deviation of the residuals $r(s_i)$ over the entire study area s, and ε_{het} is a zero mean, unit variance spatially correlated Gaussian random field. The underlying rationale is that σ varies over the study area depending on the strength of model fit, which varies with sampling intensity.

In contrast to earlier work (Lark, 2009; Wadoux et al., 2018) that assumes the standard deviation to depend linearly on covariates and jointly estimated the parameters of that linear model and the variogram by restricted maximum likelihood, here we modelled heteroscedasticity separately from variogram modelling. The map bias μ (Eq. (2)) was predicted by $\widehat{\mu}_{OK}$ from the homoscedastic model (Section 2.1.4). Acknowledging the potentially non-linear impact of sampling intensity on the variance of the residuals, σ was modelled by a smooth function of the sampling intensity. To that end, zero-degree (constant) locally estimated scatterplot smoothing (LOESS) models, as implemented in R's stats library, were fitted through standard deviations computed from binned random cross-validation residuals. The bins were delimited by the quantiles (0, 0.01, 0.02, …, 1) of the sampling intensities estimated by a two-dimensional kernel approach (see Section 2.1.3). The LOESS smoothing parameter was set to 0.5. The observed conventional cross-validation residuals were next transformed through division by the local $\sigma(s_i)$ after subtracting $\widehat{\mu}_{OK}$ (Eq. (3)):

$$r^{'}(\text{s}_i) = (z(s_i) - \widehat{z}_m(s_i) - \widehat{\mu}_{OK}) / \sigma(\text{s}_i) \tag{3}$$

where $r'(s_i)$ is the transformed residual at sample location s_i. Variogram modelling and conditional sequential Gaussian simulation proceeded as in the homoscedastic model but now using the transformed residuals, $r'(s_i)$. The residuals simulated at the nodes of the simulation grid (s_0) were subsequently back-transformed to $r(s_0)$ by multiplication with the standard deviations predicted by the LOESS model and addition of $\widehat{\mu}_{OK}$ (Eq. (4)):

$$r(s_0) = \widehat{\mu}_{OK} + r^{'}(\text{s}_0) \bullet \sigma(\text{s}_0) \tag{4}$$

2.2. Explored samples

To allow evaluation of the cross-validation methods, our analyses needed populations which were sampled and used for computing reference map accuracy metrics of random forest models fitted on the samples. The used populations are proxies of target environmental variables. They cover the entire study area where both the target variables and the covariates (Section 2.4) are available. The explored samples which represent different degrees of clustering in a reproducible way are described below.

1. *Simple random sample*. This corresponds to a simple random sample —without replacement— of the study area, where each location (i.e., grid cell) has equal inclusion probability. Note that by nature, a spatial simple random sample exhibits some degree of clustering and hence differentiation in sampling intensity. This sample was added as a reference case, and it was analysed the same way as the other samples.
2. *Systematic random sample.* This refers to sampling on a regular grid producing equal density over the study area (cf. Su et al., 2020). The grid nodes were obtained by randomly shifting an initial square sampling grid that was spaced so as to achieve the desired sample size (here 5000). Next, shifts in x and y directions were applied, where the shifts were sampled from a uniform distribution between

minus half and plus half the spacing of the sampling grid. Only nodes hitting grid cells within the study area were retained, which implies there can be some spatial gaps in the sample. Apart from these gaps, the sampling intensity is uniform. Owing to the shifts, area boundaries and some no data areas, the actual realized sample sizes differed to some degree from the expected value.

3. *Moderately clustered sample*. This sample was produced by stratified sampling. First, the study area was divided into 100 compact geo-strata using the spcosa package (Walvoort et al., 2010). In each sample, 20 of the 100 geo-strata were randomly selected to form a stratum and 50% of the total sample size was randomly chosen within this stratum. The locations of the remaining 50% of the sample were randomly chosen from the stratum formed by the other 80 geo-strata.
4. *Strongly clustered sample*. This sample was produced similarly to the moderately clustered sample. However, here ten of the spcosa geo-strata were randomly selected to form the first stratum in which 90% of the sample grid cells were randomly allocated. Additionally, 10% of the sample was allocated to the stratum composed of the remaining 90 geo-strata. An example of a real-life dataset represented by this sample is the mixture of forest inventory plots and research plots used for evaluating global AGB maps (de Bruin et al., 2020).
5. *Strongly clustered sample with gaps*. This sample is based on the spatial configuration of observed AGB pixels in central Africa used by Ploton et al. (2020). The data file belonging to that paper was downloaded from https://doi.org/10.6084/m9.figshare.11865450. The centre of the dataset was shifted to the centre of our study area. The 1 km^2 pixels were expanded by a factor two so that the spatial extent matched our study area. The pixels were next sampled at a regular point spacing of 0.5 km. The sampling points obtained in this way were shifted by a random (uniformly distributed) shift of $+/-$ 300 km in x-direction and $+/-$ 70 km in y-direction. After intersecting with a mask of the study area, the remaining points were randomly subsampled to achieve the intended sample size.

Each of the sampling designs 1–5 was repeated 100 times, every time producing a different sample. The size of samples 1, 3, 4 and 5 corresponded to 5000 each. The systematic random samples' size (2) varied between 4998 and 5056; upon trying several grid spacings. This range was the closest we got to the intended sample size of 5000.

2.3. Map accuracy metrics

The map accuracy metrics used in this work are the square root of the mean squared prediction error (RMSE) and the Nash and Sutcliffe (1970) model efficiency coefficient (MEC), which quantifies the improvement made by the model (in this case the map) over using the mean of the observations as the prediction. The metrics are defined in Eqs. (5) and (6):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(z(s_i) - \widehat{z}_m(s_i))^2}{N}} \tag{5}$$

$$MEC = 1 - \frac{\sum_{i=1}^{N}(z(s_i) - \widehat{z}_m(s_i))^2}{\sum_{i=1}^{N}(z(s_i) - \overline{z}_p)^2} \tag{6}$$

where N denotes the size of the sampling frame (i.e. all units (grid cells) in the population) and $\overline{z}_p$ is the population mean of the target variable reference values. Since we had population data at our disposal, reference *population* values for these metrics were computed by applying the above equations.

The four samples described in Section 2.2 have n units, with $n \ll N$. Therefore, they only *estimate* the map accuracy metrics for the entire population. Additionally, the inverse sampling-intensity weighted cross-validation method applies case weights to the residuals at different locations. For the random, the blocked spatial and the weighted cross-validation methods, estimates of the map accuracy metrics were hence computed by Eqs. 7 and 8:

$$\widehat{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} w_i \bullet (z(s_i) - \widehat{z}_m(s_i))^2}{\sum_{i=1}^{n} w_i}} \tag{7}$$

$$\widehat{MEC} = 1 - \frac{\sum_{i=1}^{n} w_i \bullet (z(s_i) - \widehat{z}_m(s_i))^2}{\sum_{i=1}^{n} w_i \bullet (z(s_i) - \overline{z}_s)^2} \tag{8}$$

where w_i are the weights applied to individual observations (i.e. inverse sampling intensities for cross-validation method 3, otherwise the weights are constant and set to 1), $\overline{z}_s$ is the sample mean of reference values of the target variable and $\widehat{RMSE}$ and $\widehat{MEC}$ are the estimated map accuracy metrics. Note that the pairs of reference observations and map predictions were collected over ten folds.

In the two model-based methods (methods 4 and 5), the map accuracy metrics are computed from simulated residuals at the N nodes of the grid covering the entire study area. For each simulation, estimates of RMSE and MEC were computed by Eqs. (9) and (10):

$$\widehat{RMSE}_{sim} = \sqrt{\frac{\sum_{i=1}^{N} r(s_i)^2}{N}} \tag{9}$$

$$\widehat{MEC}_{sim} = 1 - \frac{\sum_{i=1}^{N} r(s_i)^2}{\sum_{i=1}^{N}(r(s_i) + \widehat{z}_m(s_i) - \overline{\mu}_{sim})^2} \tag{10}$$

where $\overline{\mu}_{sim}$ is the mean of $r(s_i) + \widehat{z}_m(s_i)$ over all nodes of the simulation grid.

To allow comparison of accuracy estimates across different samples, relative deviations of the accuracy estimates from their reference metrics were expressed as percentages. For example, the relative RMSE (rRMSE) was computed by Eq. (11). The relative MEC was computed similarly.

$$rRMSE = 100 \bullet \frac{\widehat{\mathrm{RMSE}} - \mathrm{RMSE}}{\mathrm{RMSE}} \tag{11}$$

2.4. Case study implementation

The following two maps provided population reference data for the target variables AGB and OCS (see Fig. 1), which were used for drawing the sample datasets (for model training) and for computing the reference map accuracy metrics:

- AGB: version 3 of the 2017 CCI-Biomass product (https://catalogue.ceda.ac.uk/uuid/5f331c418e9f4935b8eb1b836f8a91b8), which is a follow up of the Globbiomass product (Santoro et al., 2021);
- OCS (0–30 cm soil depth): Soilgrids (https://www.isric.org/explore/soilgrids; Poggio et al., 2021).

The maps were spatially resampled and aligned to a common grid with a resolution of 0.5 km that is compatible with the resolution of the covariates. The covariates used for predicting AGB and OCS comprise:

- Seven terrain properties derived from the digital elevation model EU-DEM version 1.1 (Copernicus Land Monitoring Service - EU-DEM — European Environment Agency (europa.eu));
- GEDI forest height (Potapov et al., 2021);
- Seven CHELSA V2.1 climate variables (Karger et al., 2020);

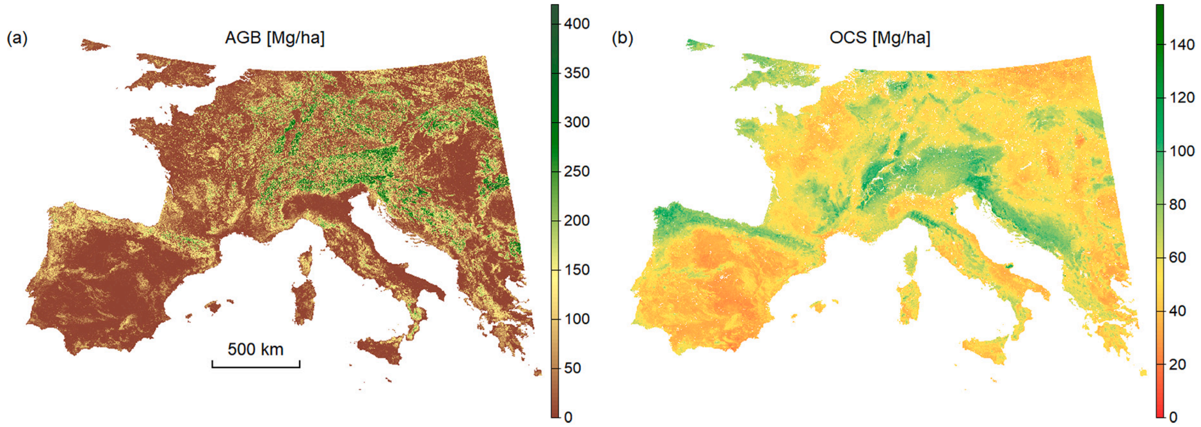


Fig. 1. Reference above-ground biomass from CCI Biomass (a) and soil organic carbon stock in the 0–30 cm layer from Soilgrids (b) in the study area.

- Seven generalized land cover classes derived from the 2017 Copernicus land cover map (Buchhorn et al., 2020);
- Three soil properties from SoilGrids (only used for predicting AGB);
- Two spatial coordinates (x, y) and distance from the coast, the latter computed using a land mask of the study area that was derived from the other covariates.

The full list of covariates and data sources is provided in the appendix.

The study area is western Europe, constrained in the north at 52° latitude because GEDI forest height is only available up to that latitude and at −10° and 24° longitude mainly because of the availability of EU-DEM. The chosen projection was IGNF:ETRS89LAEA (Lambert azimuthal equal area projection). The elevation-derived terrain properties were computed at the original resolution of EU-DEM (25 m) and next aggregated by a factor 20 to 0.5 km resolution. The target variables and the covariates were sampled at the sites selected by the designs described in Section 2.2 for model fitting and map accuracy estimation. All raster calculations were done using the terra package (Hijmans, 2021).

Random Forest models as implemented in the ranger package (Wright and Ziegler, 2017) predicting the target variables as a function of the covariates were fitted using out-of-the-box hyperparameter settings. The only exception was the `respect.unordered.factors` option, which was set to `TRUE` to ensure the categorical land cover variable was properly used in the regression trees. Maps of AGB and OCS were obtained by predicting with the fitted models using the covariate maps.

In the two model-based approaches, for the AGB residuals automated variogram fitting first attempted to fit a nested model composed of a short-range spherical structure, a longer-range exponential structure, and a nugget. If that failed —as indicated by a model singularity warning— an exponential structure with a nugget was tried and, as an ultimate resort, a pure nugget model was fitted (a single case). For the OCS residuals, first a nested model composed of a spherical structure, a Gaussian structure, and a nugget was tried, which in case of failure was followed by fitting an exponential structure with nugget. The model structures were chosen upon visual inspection of a sample of experimental variograms computed from the data. To reduce the computational burden of the model-based approaches, the number of foldings used in the two model-based approaches was reduced to ten out of the 100 foldings from random cross-validation. Furthermore, the residual fields were computed at a grid spacing of 5 km, and the maximum number of nearby points on which local simulations were conditioned (`nmax`) was set to 75. The geostatistical models were implemented using the gstat package for R (Gräler et al., 2016; Pebesma, 2004).

3. Results and discussion

3.1. Samples

A subset of the 100 realizations of the five sampling designs described in Section 2.2 is shown in Fig. 2. Note that for the systematic random sample (Fig. 2(c)) the entire regular grid was randomly shifted in horizontal and vertical direction. Since the effect of such shifts is difficult to appreciate visually, here only a single systematic random sample is shown. Non-sampled areas within the European continent (e.g. around the Alps) correspond to no data areas for some of the covariates owing to bare rock surfaces, water bodies or urban areas. These sites were excluded from the study area.

3.2. Random forest predictions and reference accuracy metrics

Fig. 3 shows example maps of AGB predictions made by random forest models trained on samples as per four of the five designs compared in this study. At first sight, the maps (a-c) are very similar to the proxy of exhaustive ground truth provided in Fig. 1(a) but closer inspection reveals some dissimilarities. Perhaps the most apparent difference is a reduction of the largest AGB values along Europe's mountain ranges in Fig. 3(c). The latter map shows predictions based on a strongly clustered sample whose locations are shown in Fig. 2(f). It can be observed that this sample had few sampling sites in the high biomass areas, which may partly explain the observed tendency. The predictions in Fig. 3(d) differ considerably from those in Fig. 3(a-c). These are based on the strongly clustered, gapped sample shown in Fig. 2(h) which fails to cover the major part of the study area including the regions with large AGB values (cf. Fig. 1(a)). These higher AGB sites are thus predicted by extrapolation rather than interpolation. Moreover, the population map accuracy metrics shown in Fig. 4 reveal that the strongly clustered, gapped design had the smallest MEC of the five compared sampling designs. Weaker correlation between predictions and reference data (as indicated by MEC) generally implies overprediction of small reference values and underprediction of greater values. This effect is commonly observed for AGB maps (e.g. Avitabile and Camia, 2018; Santoro et al., 2021) and in statistics it is referred to as reversion or regression toward the mean (Samuels, 1991). A similar pattern was observed for the OCS predictions, which for reason of brevity are not shown here.

The map accuracy metrics depicted in Fig. 4 show that models trained on the systematic samples covering the study area with uniform density on average produced the greatest accuracy (i.e. smallest RMSE and greatest MEC) for both AGB and OCS. The models trained on the simple random samples were on average slightly less accurate but the ranges of their map accuracy metrics largely overlap with those of the systematic samples. Models trained on the strongly clustered samples

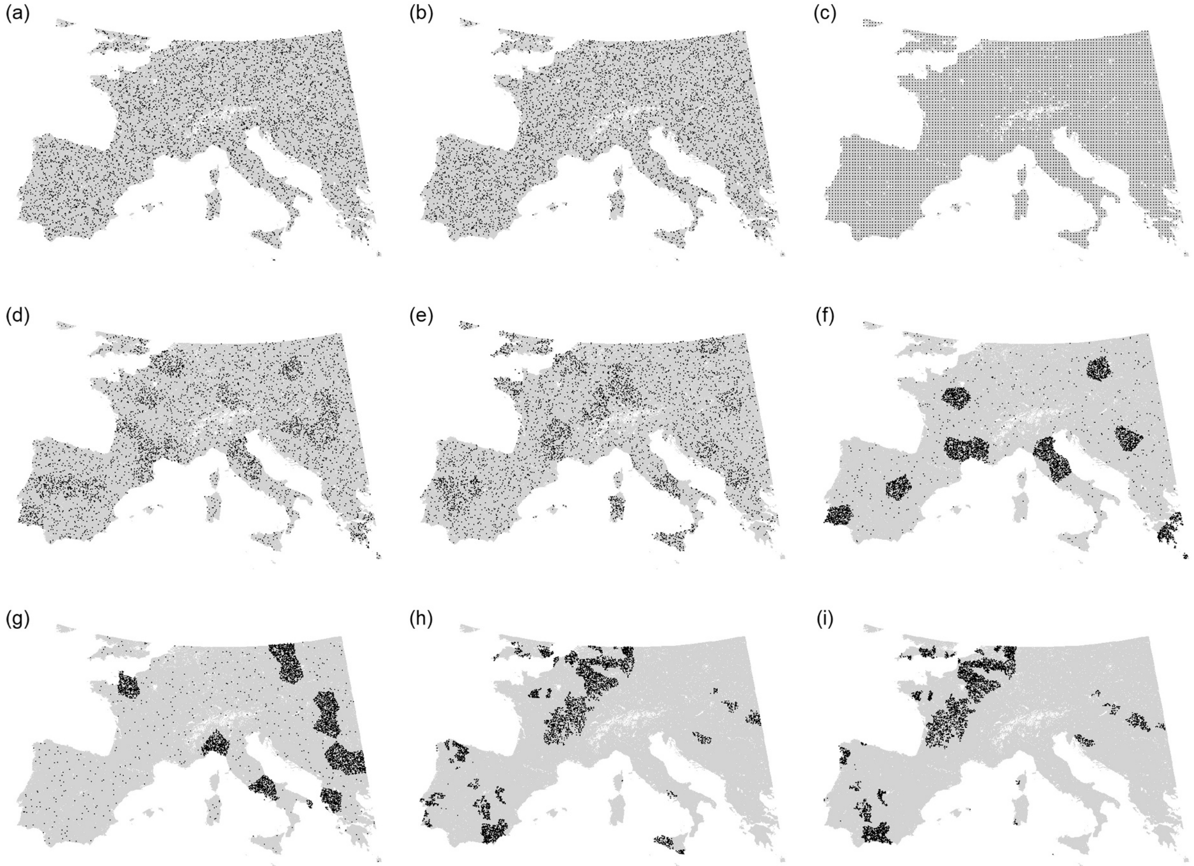


Fig. 2. Examples of studied spatial samples. (a-b) simple random samples; (c) systematic random sample; (d-e) moderately clustered samples; (f-g) strongly clustered samples; (h-i) strongly clustered, gapped samples. Except for the systematic sample (c), the sample size always amounted to 5000. The systematic sample had an expected size of 5000 but realized samples varied in size between 4998 and 5056.

were least accurate and also varied the most over the samples, while the ranges of their map accuracy metrics did not overlap those of the other samples. These effects are as expected, as strongly clustered samples overrepresent some geographic regions while underrepresenting others, which is likely to also impact representation of feature space. The large spread in the map accuracy metrics can be explained by major differences in the coverage of geographic space and feature space amongst different realizations of the strongly clustered design. Note that all random forest models had decent predictive skill as judged by the population MEC values, yet the AGB models were more accurate than the OCS models.

3.3. Sampling intensities and model identification

Both the weighted cross-validation method (Section 2.1.3) and the heteroscedastic model-based method (Section 2.1.5) require sampling intensities to be estimated from the dataset. For each spatial sampling design, Fig. 5 shows an example of the intensities computed with the spatstat two-dimensional kernel approach (Baddeley et al., 2016) with the bandwidth selection according to the Cronie and Van Lieshout (2018) criterion. Spatial clustering of the simple random samples produced some variation in the estimated sampling intensity as can be observed in Fig. 5(a). As expected, systematic sampling resulted in a homogeneous sampling intensity, except for areas near gaps in the sampling frame, such as around the Alps (Fig. 5(b)). The two clustered samples exhibit most differentiation in sampling intensity, with the strongly clustered sample having the greatest spatial variation. The latter is evident in Fig. 5(d), whose pattern emerged from the sample depicted in Fig. 2(f).

In the heteroscedastic model-based method, the estimated sampling intensities were used for modelling non-stationarity of the variance of the AGB and OCS residuals from reference data by a smooth function of sampling intensity. Fig. 6 presents examples of the LOESS heteroscedasticity models fitted to standard deviations computed from binned cross-validation residuals for the four sampling intensities depicted in Fig. 5 (a, c-e). The systematic design lacks sampling intensity variation and therefore the corresponding plot is not shown here. Note that for a single sample, owing to their similarity, the semi-transparent red curves representing the LOESS models from different foldings can hardly be discriminated. However, different samples of the same design and samples from different designs produced widely divergent models. The latter can be appreciated by comparing the curves of Fig. 6(a-d). To facilitate comparison of the data ranges over which the models were fitted, these were kept constant along the x-axes and the y-axes of Fig. 6 (a-d). For the simple random sampling design, the standard deviation of the residuals was fairly constant over the observed range of sampling intensities. The curves for the moderately clustered sample and the strongly clustered, gapped sample shows the anticipated smooth reduction in the standard deviation with increasing sampling intensity. In the case of a strongly clustered sample, the modelled standard deviation of the residuals of the AGB residuals decreases with sampling intensities up to approximately $7.5 \bullet 10^{-9}$, beyond which it increases again. This behaviour seems counterintuitive but note that the two patches with the largest sampling intensities (in Poland and Bulgaria) coincide with areas having heterogeneous AGB values at relatively short distances (cf. Fig. 1(a)). This spatial pattern appears not to be accurately reproduced by the random forest model predictions.

Figs. 7 and 8 show example variograms of AGB residuals and of

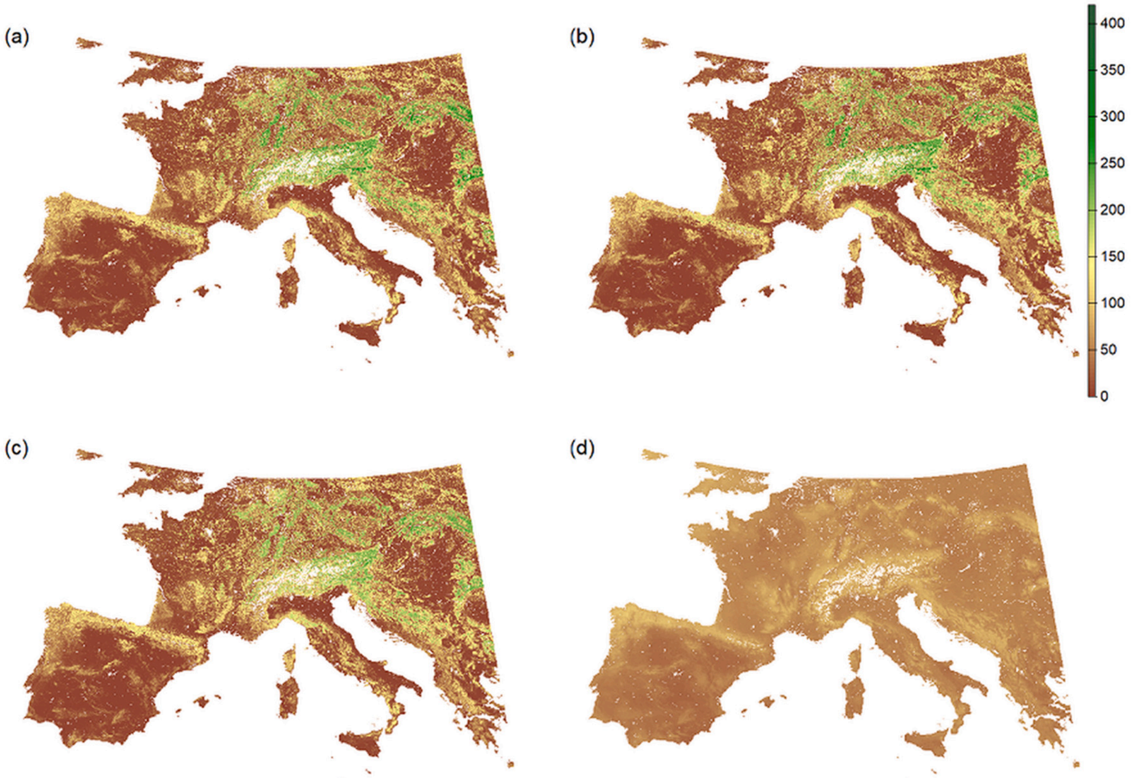


Fig. 3. AGB predictions using (a) simple random sample; (b) moderately clustered sample; (c) strongly clustered sample; (d) strongly clustered, gapped sample.

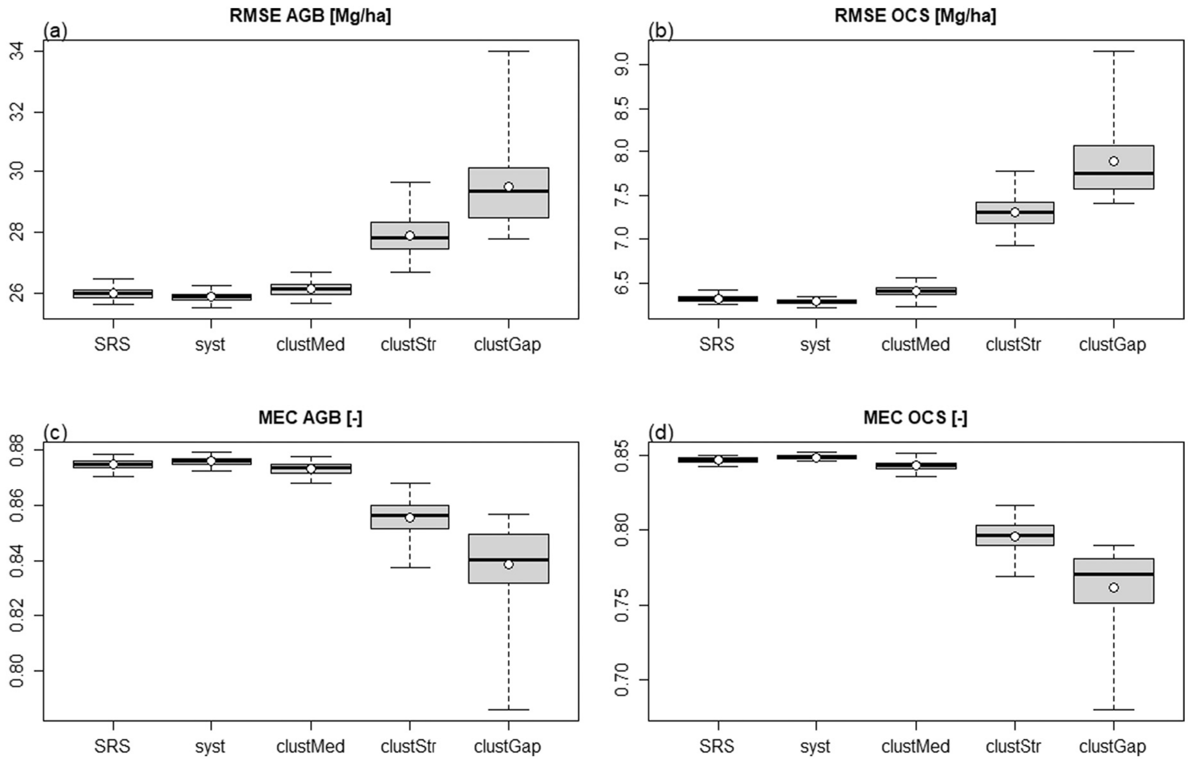


Fig. 4. Reference map accuracy metrics computed from the population data for models trained on 100 realizations of each sampling design. Boxes denote the interquartile range; thick horizontal line inside boxes are medians; whiskers mark the full range and points indicate the means. SRS = simple random; syst = systematic random; clustMed = moderately clustered; clustStr = strongly clustered; clustGap = strongly clustered, gapped sample.

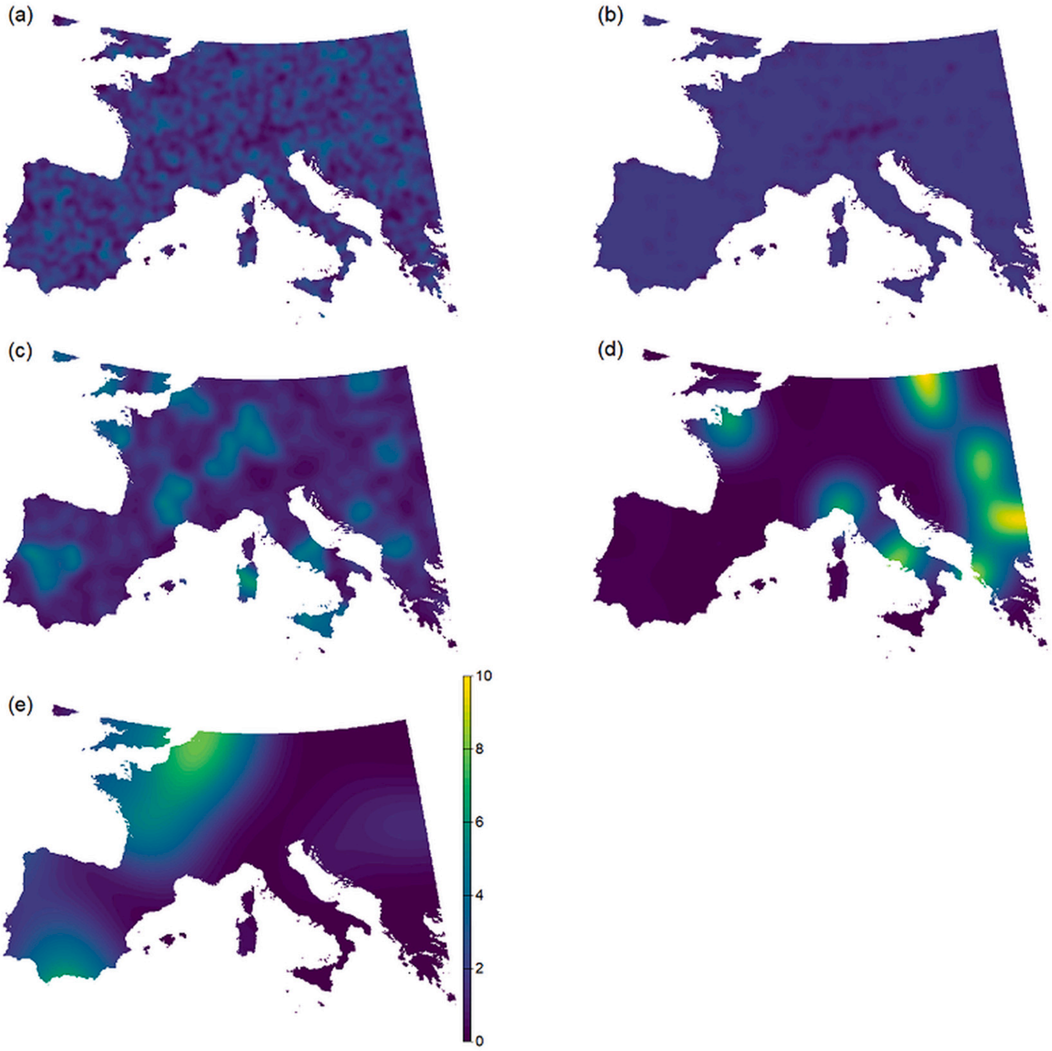


Fig. 5. Examples of estimated sampling intensities (multiplied by 10^9) for five samples shown in Fig. 2: (a) simple random (cf. Fig. 2(a)); (b) systematic random (cf. Fig. 2(c)); (c) moderately clustered (cf. Fig. 2(e)); (d) strongly clustered (cf. Fig. 2(g)); (e) strongly clustered, gapped sample (cf. Fig. 2(i)).

transformed AGB residuals as used in the homoscedastic and the heteroscedastic model-based methods, respectively. The sill of the AGB residual variogram computed from the strongly clustered, gapped sample in Fig. 7(e) is substantially lower than that of the other samples (Fig. 7(a-d)). Hence, the sampled sub-space is internally relatively homogeneous in terms of AGB residuals whilst these residuals are hardly spatially correlated (no remaining spatial structure) as witnessed by the (almost) pure nugget variograms. The random forest models effectively characterise AGB variability within the densely sampled sub-space (as discussed below) but beyond that region their performance is weak as indicated by the reference map accuracy metrics shown in Fig. 4(a, c).

The transformation of the residuals to account for heteroscedasticity as a function of sampling intensity rescaled the variograms to approximately unit sill (Fig. 8), as it should, while it also impacted the shape of the variogram model to some degree. For example, for the moderately clustered sample it led to a reduction of the long-range spatial correlation structure of AGB residuals for distances up to 200 km (compare Fig. 7(c) and Fig. 8(c)) while for the strongly clustered, gapped sample heteroscedasticity modelling introduced some weak spatial structure (compare Fig. 7(e) and Fig. 8(e)). Note that Figs. 7-8 only show the variograms of AGB residuals from a single sample of each sampling design; however, we assessed 100 samples per sampling design for both AGB and OCS residuals.

3.4. Map accuracy

Boxplots of the relative deviation of the map accuracy metrics from their reference values for maps produced by models trained on 100 samples are shown in Fig. 9. Blocked spatial cross-validation systematically overestimated the RMSE while underestimating the MEC for all designs except for the strongly clustered gapped AGB samples, which subscribes its pessimistic bias anticipated in the Introduction. In blocked spatial cross-validation, the *k* assessed models are each trained on a subsample lacking coverage of a spatially contiguous part of geographic space where the validation points are located. Since this is likely to have an impact on the coverage of feature space as well, these *k* models tend to be inferior to the model trained on the full dataset. In fact, blocked spatial cross-validation aggravates the impact of clustered sampling patterns by removing parts of geographic space in each fold on which the model is tested. This effect has been claimed useful for assessing a model's extrapolation error (Roberts et al., 2017). However, the feature space of regions not covered by the sample can be quite different from those of the blocks folded out in spatial cross-validation so that blocked spatial cross-validation can also be overly optimistic about map accuracy. Examples are the RMSE estimates for the strongly gapped AGB samples (Fig. 9(a)), where the cross-validation results seem to have been slightly optimistic while there was no remaining spatial correlation in the AGB residuals that would be removed by spatial blocking (see Fig. 7

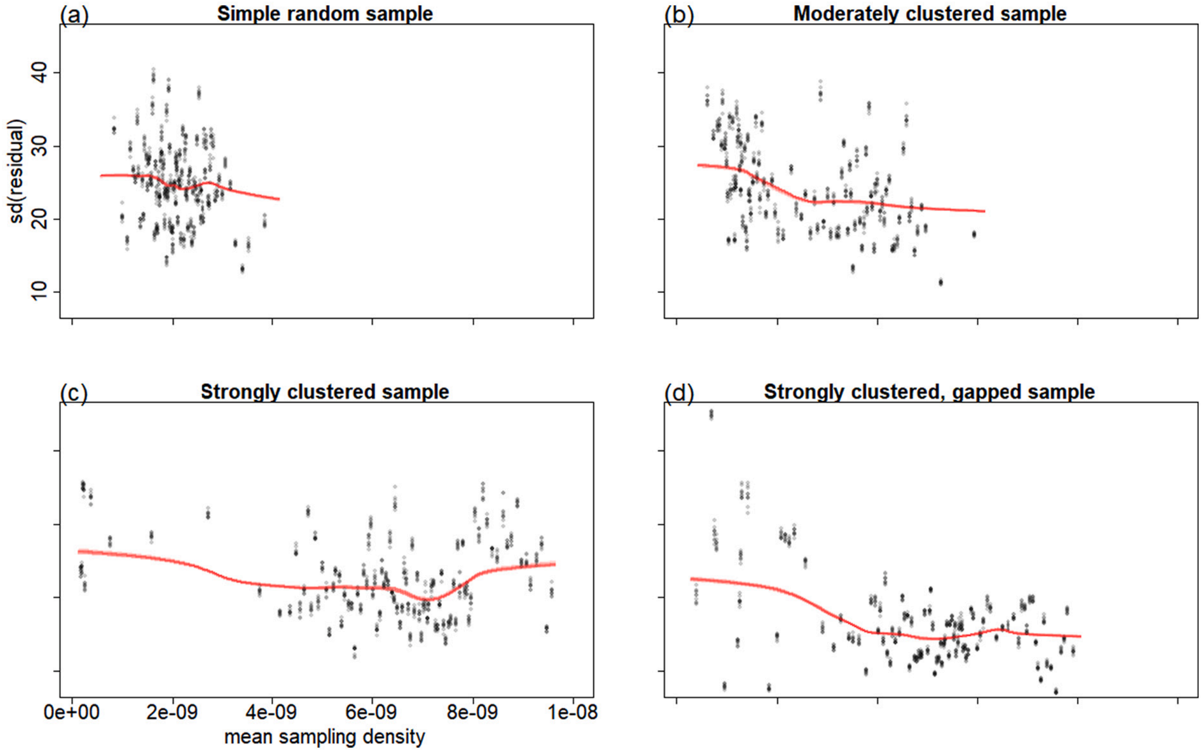


Fig. 6. Average standard deviation of AGB residuals per density percentile (dots) and fitted LOESS models (curves) for 10 foldings of instances of each of the four sampling designs (a-d; same as in Fig. 5(a, c-e). Semi transparency is used for plotting dots and curves corresponding to different data foldings. Per sample, the fitted LOESS models are nearly identical over the different foldings and therefore their curves can hardly be distinguished.

(e)).

For the non-clustered sampling designs, the range of the metrics over the 100 samples obtained by blocked spatial cross-validation did not include the reference metrics. In contrast, the other four cross-validation methods estimated the map accuracy metrics within 10% from their reference values for nearly all samples of these designs and design-bias was substantially less than with the blocked spatial cross-validation method. The conventional random cross-validation results were very close to those obtained by weighted cross-validation and the heteroscedastic model-based method. It is noted that for the simple random samples (which have equal inclusion probability) and the systematic random samples (which in addition to equal inclusion probability have uniform sampling intensity), accounting for differences in sampling intensity is irrelevant. Nevertheless, doing so hardly if at all affected the map accuracy metric estimates compared to their non-intensity-adjusted counterparts (cf. conventional versus intensity weighted, and heteroscedastic versus homoscedastic in Fig. 9).

The weighted cross-validation (quasi-randomization) and the heteroscedastic model-based methods —which both account for the effects of differences in sampling intensity— tended to be the least biased for the clustered sampling designs without gaps. However, for the strongly clustered, non-gapped design, bias remained substantial and should be further reduced. This specifically concerns OCS, which compared to AGB was predicted with weaker model fits as judged from the reference MEC (Fig. 4). The overoptimistic accuracy estimates can be attributed to estimation from geographically disproportional samples from regions that were also most intensely sampled for fitting the random forest models. In contrast, the reference metrics were computed over the entire sampling frame. Weighted cross-validation and the model-based methods attempt to correct for selection bias but only accomplished this to a limited extent. Potential enhancements may be attained with improved sampling intensity estimation and heteroscedasticity modelling.

For the strongly clustered designs with gaps, most map accuracy estimates were strongly optimistically biased for all cross-validation methods except for blocked spatial cross-validation. The latter may be attributed to blocked cross-validation's hypothetical ability to estimate a model's extrapolation error (Roberts et al., 2017) if the areas of the map not covered by the sample have similar accuracy characteristics as the folded out cross-validation blocks, which is not guaranteed. The weighted cross-validation and model-based methods proposed in this work lack information for filling gaps in this design.

Elliott and Valliant (2017) discussed quasi-randomization and superpopulation modelling as generic approaches for making inferences from non-probability samples. Here instances of the two approaches were considered within a *k*-fold cross-validation context. The weighted cross-validation method is an implementation of a quasi-randomization approach. It was found to be less biased than blocked spatial cross-validation for the target variables and sampling designs without gaps assessed in this study. The two geostatistical model-based methods presented in this paper are examples of the superpopulation modelling approach discussed by Elliott and Valliant (2017). Accounting for heteroscedasticity in the AGB and OCS residuals was found to reduce bias compared to the homoscedastic model for the non-gapped designs. However, it did not mitigate bias for the strongly clustered gapped design, which is no surprise as the non-covered areas on the map lack data points to inform the geostatistical model. The best the geostatistical model can do in gaps where the nearest sample points are beyond the range of the residual variogram is to predict the spatial mean of the residuals, which is optimistically biased as there is no information about potential extrapolation beyond the feature space the model is trained on.

None of the *k*-fold cross-validation methods evaluated in this paper is guaranteed to be capable of assessing map extrapolation error for areas not covered by the sample. To support mapping the entire study area without extrapolation, additional sampling is needed if the current sample lacks coverage of feature space. Alternatively one may confine the mapped area to the part covered by the sampled feature space, e.g., by excluding sites exceeding a critical distance between feature data at

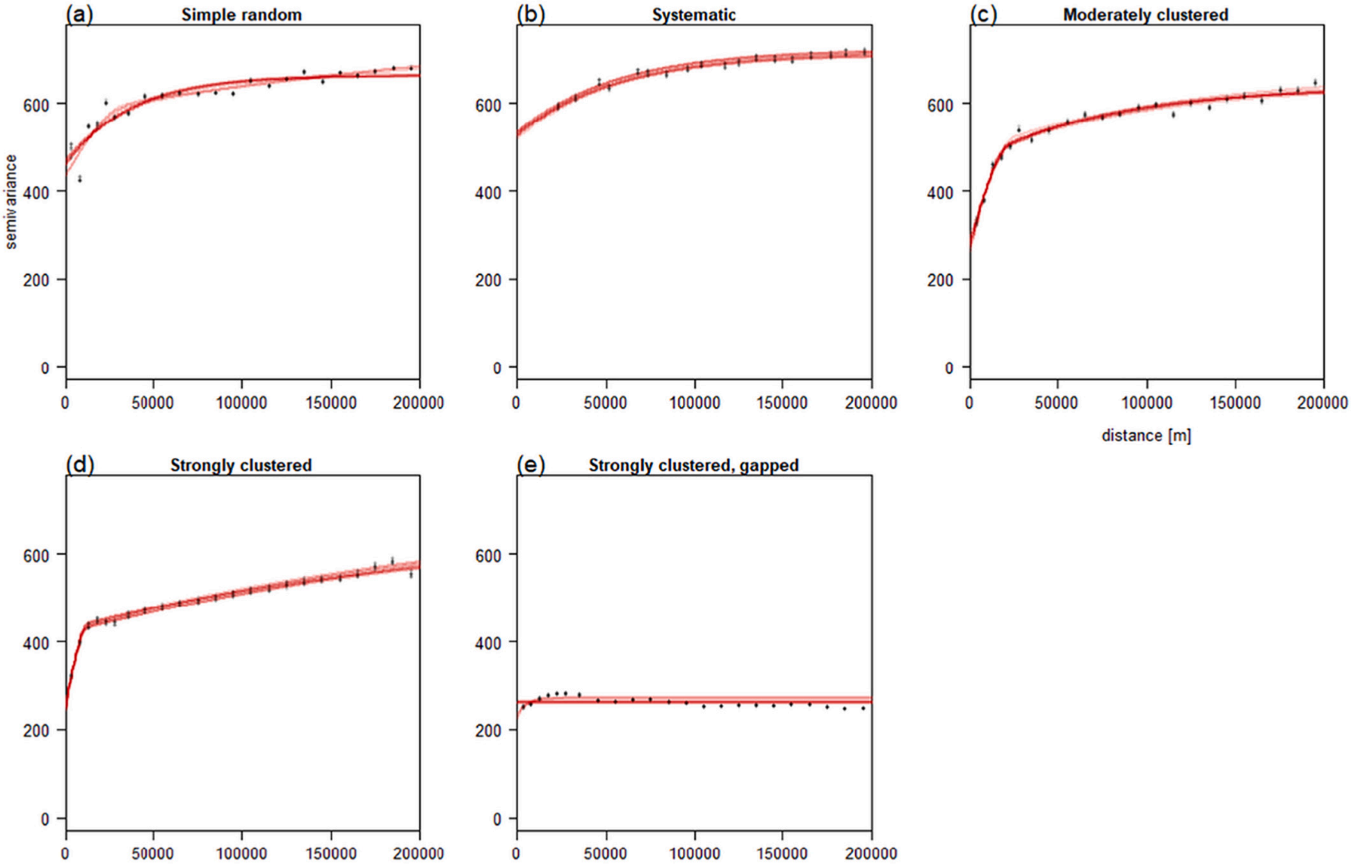


Fig. 7. Experimental semivariances (dots) and fitted variogram models (curves) of AGB residuals for 10 foldings of instances of each of the five sampling designs (a-e; same as in Fig. 5). Semi transparency is used for plotting dots and curves corresponding to different data foldings.

the prediction sites and those in the training dataset (Meyer and Pebesma, 2021).

To support statistically convincing conclusions, it is important to also estimate the uncertainty of map accuracy estimates. Currently, there is no universal unbiased estimator of the variance of *k*-fold cross-validation (Bengio and Grandvalet, 2004; Yousef, 2021). A model-based approach allows computing the model-variance of the predicted RMSE and MEC and use these to compute prediction intervals for these metrics. However, the model-variance of a predictor and the sampling variance of an estimator are fundamentally different, as the source of randomness differs between the approaches. In design-based estimation, randomness concerns the selection of sampling units, whereas in model-based prediction randomness is introduced via the statistical model of the spatial variation (Brus, 2021). The latter randomness depends on the model's specification and identification. In general, estimation of confidence intervals from *k*-fold cross-validation is a yet unsolved puzzle requiring further research.

We acknowledge that our case studies relied on proxies of the spatial distribution of environmental properties rather than the true spatial distribution as encountered in the field. The used AGB and OCS maps smooth out part of the natural spatial variability and prediction models fitted on samples from the maps therefore tend to be more accurate than models fitted on samples collected in the real world. However, since our demonstration required an entire population to derive reference values for RMSE and MEC, it would have been infeasible to accomplish without reliance on proxy data. Moreover, besides the case of absence of spatial structure in the target variable, we cannot think of any reason why the general bias trends observed in our case studies would diverge from those observed with real field samples.

4. Conclusions

We have proposed an inverse sampling-intensity weighted and two geostatistical model-based cross-validation approaches as alternatives to the broadly propagated spatial cross-validation method that tends to produce pessimistically biased map accuracy estimates. The proposed methods can be characterised as quasi-randomization and super-population modelling approaches, respectively. Like blocked spatial cross-validation, our weighted approach is a heuristic method. However, in contrast to the former it explicitly addresses the spatial clustering problem rather than the incorrectly posed problem of spatial proximity of test and training data. In our case studies, bias in the map accuracy metrics assessed over multiple realizations of the sampling designs by weighted cross-validation was much smaller than that of blocked spatial cross-validation for non-clustered to moderately clustered samples. For the strongly clustered design where large portions of the maps were predicted by extrapolation, blocked spatial cross-validation was closest to the reference map accuracy metrics. However, blocked spatial cross-validation may still yield biased estimates of the map accuracy metrics, because it is impossible to tell how large the blocks should be to adjust for the deterioration of map accuracy metrics caused by extrapolation. Rather, extrapolation is to be avoided by additional sampling or limitation of the prediction area.

The proposed model-based approaches are rooted in geostatistics. An initial homoscedastic model was expanded by modelling heteroscedasticity of residuals in the target variable as a smooth function of observed sampling intensity. The resulting heteroscedastic model's map accuracy metrics were similar to those obtained by weighted cross-validation for the samples without spatial gaps. Hence, for reasons of parsimony we recommend conventional random cross-validation for

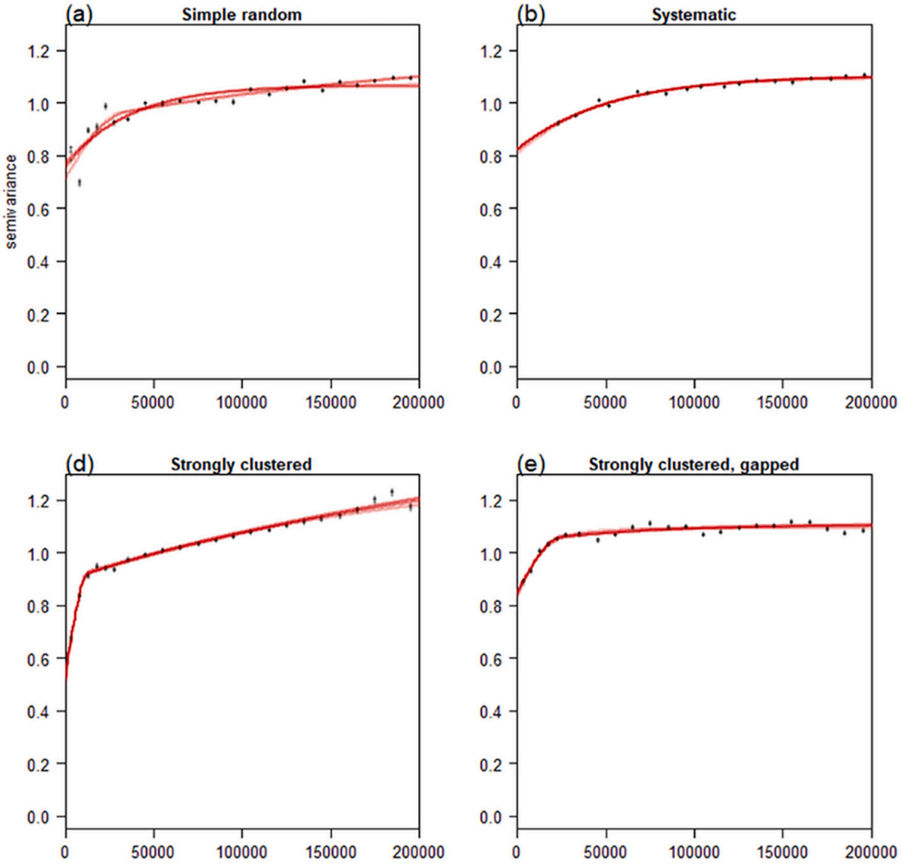


Fig. 8. Experimental semivariances (dots) and fitted variogram models (curves) of *transformed* AGB residuals for 10 foldings of instances of each of the four sampling designs (a-e; same as in Fig. 5). Transformation was done using the local regression curves shown in Fig. 6. Semi transparency is used for plotting dots and curves corresponding to different data foldings.

non-clustered samples and weighted cross-validation for moderately clustered samples.

Further research is needed to improve accuracy assessment by cross-validation from strongly spatially clustered samples and for estimating confidence intervals for map accuracy metrics.

Authorship

All authors conceived the ideas and designed the methodology; SB analysed the data; SB, DB, GH and AW led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Declaration of Competing Interest

None.

Appendix A. Data sources

Name	Description	Source	Note
ai	Aridity Index	https://chelsa-climate.org/downloads/	Version 2.1
bio1	Mean annual air temperature [°C]		
bio5	Mean daily maximum air temperature of the warmest month [°C]		
bio7	Annual range of air temperature [°C]		
bio12	Annual precipitation [kg/m^2]		
bio15	Precipitation seasonality [kg/m^2]		
gdd10	Growing degree days heat sum above 10 °C		
clay	Clay content [g/kg] of the 0-5 cm layer	https://soilgrids.org/	Only used for AGB
sand	Sand content [g/kg] of the 0-5 cm layer		
pH	Acidity (Ph(water)) of the 0-5 cm layer		
glc2017	Landcover 2017	https://land.copernicus.eu/global/products/lc, reclassified to: closed forest, open forest, natural non-forest veg., bare & sparse veg. Cropland, built-up, water	Categorical variable
dem	Elevation	https://www.eea.europa.eu/data-and-maps/data/copernicus-land-monitoring-service-eu-dem	

(continued on next page)

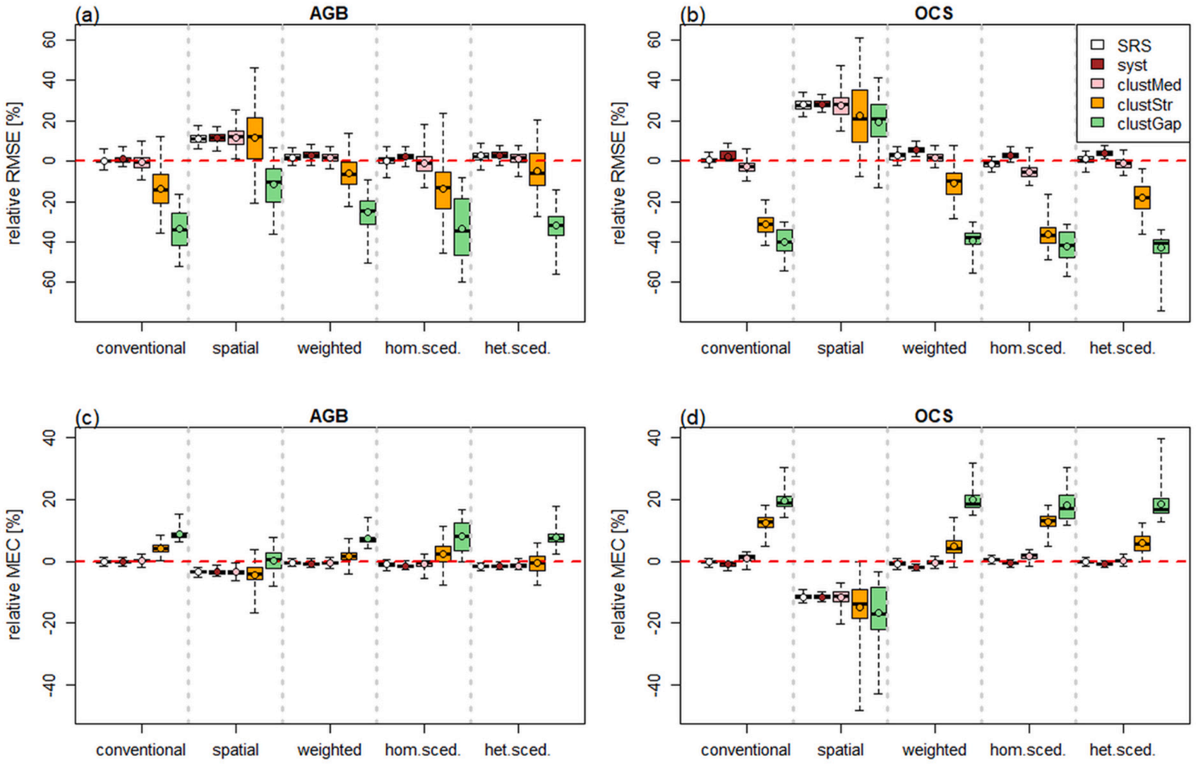


Fig. 9. Relative deviation of $\widehat{RMSE}$ (a, b) and $\widehat{MEC}$ (c, d) from their reference metrics for AGB (a, c) and OCS (b, d) for models trained on 100 realizations of the explored sampling designs, according to the cross-validation methods listed along the x-axes (hom.sced = homoscedastic model-based; het.sced = heteroscedastic model-based). Boxplot symbology as in Fig. 4.

(*continued*)

Name	Description	Source	Note
cosasp	Cosine of slope aspect	Computed with the terra package from elevation	Computed @25 m resolution; next aggregated to 0.5 km
sinasp	Sine of slope aspect		
slope	Slope		
TPI	Topographic position index		
TRI	Terrain ruggedness index		
TWI	Topographic wetness index	Computed with SAGA from 500 m resolution (aggregated) dem	
gedi	Forest height	https://glad.umd.edu/dataset/gedi	Zone: NAFR
xcoord	X coordinate	Using a mask created from the other covariates	
ycoord	Y coordinate		
Dcoast	Distance from coast	Using a land mask created from the other covariates	
clustGap	Spatial configuration of observed AGB pixels in central Africa	doi:https://doi.org/10.6084/m9.figshare.11865450	

A compiled version of the datasets used in this paper can be downloaded from: https://doi.org/10.5281/zenodo.6513429. The analysis scripts are available at: https://doi.org/10.5281/zenodo.6514923.

References

Avitabile, V., Camia, A., 2018. An assessment of forest biomass maps in Europe using harmonized national statistics and inventory plots. For. Ecol. Manag. 409, 489–498. https://doi.org/10.1016/j.foreco.2017.11.047.

Baddeley, A., Rubak, E., Turner, R., 2016. Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC. https://doi.org/10.1201/b19708.

Bengio, Y., Grandvalet, Y., 2004. No unbiased estimator of the variance of K-fold cross-validation. J. Mach. Learn. Res. 5, 1089–1105.

Brenning, A., 2012. Spatial Cross-Validation and Bootstrap for the Assessment of Prediction Rules in Remote Sensing: The R Package Sperrorest, pp. 5372–5375. https://doi.org/10.1109/IGARSS.2012.6352393.

de Bruin, S., Araza, A., Herold, M., Kay, H., Lucas, R.M., 2020. CCI Biomass Product Validation and Intercomparison Report, Year 2, Version 2. Wageningen University, Aberystwyth University, Wageningen. https://climate.esa.int/media/documents/Biomass_D4.1_Product_Validation_Intercomparison_Report_V2.0.pdf.

Brus, D.J., 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.

Buchhorn, M., Smets, B., Bertels, L., De Roo, B., Lesiv, M., Tsendbazar, N.-E., Herold, M., Fritz, S., 2020. Copernicus Global Land Service: Land Cover 100m: Collection 3: Epoch 2017: Globe (V3.0.1), Zenodo. https://doi.org/10.5281/zenodo.3518036.

Cochran, W.G., 1977. Sampling Techniques. John Wiley & Sons.

Cronie, O., Van Lieshout, M.N.M., 2018. A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. Biometrika 105 (2), 455–462. https://doi.org/10.1093/biomet/asy001.

d'Andrimont, R., Yordanov, M., Martinez-Sanchez, L., Eiselt, B., Palmieri, A., Dominici, P., Gallego, J., Reuter, H.I., Joebges, C., Lemoine, G., van der Velde, M., 2020. Harmonised LUCAS in-situ land cover and use database for field surveys from 2006 to 2018 in the European Union. Sci. Data 7 (1), 352. https://doi.org/10.1038/s41597-020-00675-z.

Deutsch, C., 1989. DECLUS: a fortran 77 program for determining optimum spatial declustering weights. Comput. Geosci. 15 (3), 325–332. https://doi.org/10.1016/0098-3004(89)90043-5.

Du, P., Bai, X., Tan, K., Xue, Z., Samat, A., Xia, J., Li, E., Su, H., Liu, W., 2020. Advances of four machine learning methods for spatial data handling: a review. J. Geovisual. Spatial Anal. 4 (1), 13. https://doi.org/10.1007/s41651-020-00048-5.

Elliott, M.R., Valliant, R., 2017. Inference for nonprobability samples. Stat. Sci. 32 (2), 249–264, 216. https://doi.org/10.1214/16-STS598.

Fitts, L.A., Russell, M.B., Domke, G.M., Knight, J.K., 2021. Modeling land use change and forest carbon stock changes in temperate forests in the United States. Carbon Balance Manag. 16 (1), 20. https://doi.org/10.1186/s13021-021-00183-6.

Fushiki, T., 2011. Estimation of prediction error by using K-fold cross-validation. Stat. Comput. 21 (2), 137–146. https://doi.org/10.1007/s11222-009-9153-8.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press.

Grabska, E., Frantz, D., Ostapowicz, K., 2020. Evaluation of machine learning algorithms for forest stand species mapping using Sentinel-2 imagery and environmental data in the polish Carpathians. Remote Sens. Environ. 251, 112103 https://doi.org/10.1016/j.rse.2020.112103.

Gräler, B., Pebesma, E., Heuvelink, G.B.M., 2016. Spatio-temporal interpolation using gstat. R J. 8 (1), 204–218. https://doi.org/10.32614/RJ-2016-014.

Gregoire, T.G., Valentine, H.T., 2007. Sampling Strategies for Natural Resources and the Environment. Chapman and Hall/CRC, Boca Raton, FL. https://doi.org/10.1201/9780203498880.

de Gruijter, J.J., ter Braak, C.J.F., 1990. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. Math. Geol. 22 (4), 407–415. https://doi.org/10.1007/BF00890327.

Harris, N.L., Gibbs, D.A., Baccini, A., Birdsey, R.A., de Bruin, S., Farina, M., Fatoyinbo, L., Hansen, M.C., Herold, M., Houghton, R.A., Potapov, P.V., Suarez, D.R., Roman-Cuesta, R.M., Saatchi, S.S., Slay, C.M., Turubanova, S.A., Tyukavina, A., 2021. Global maps of twenty-first century forest carbon fluxes. Nat. Clim. Chang. 11 (3), 234–240. https://doi.org/10.1038/s41558-020-00976-6.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. In: Data Mining, Inference, and Prediction, Second edition. Springer-Verlag. https://doi.org/10.1007/978-0-387-84858-7.

Helmstetter, N.A., Conway, C.J., Stevens, B.S., Goldberg, A.R., 2021. Balancing transferability and complexity of species distribution models for rare species conservation. Divers. Distrib. 27 (1), 95–108. https://doi.org/10.1111/ddi.13174.

Hijmans, R.J., 2021. Terra: Spatial Data Analysis. https://cran.r-project.org/web/packages/terra/terra.pdf.

Isaaks, E.H., Srivastava, R.M., 1990. An Introduction to Applied Geostatistics. Oxford University Press.

Just, A.C., Arfer, K.B., Rush, J., Dorman, M., Shtein, A., Lyapustin, A., Kloog, I., 2020. Advancing methodologies for applying machine learning and evaluating spatiotemporal models of fine particulate matter (PM2.5) using satellite data over large regions. Atmos. Environ. 239 https://doi.org/10.1016/j.atmosenv.2020.117649.

Karger, D.N., Schmatz, D.R., Dettling, G., Zimmermann, N.E., 2020. High-resolution monthly precipitation and temperature time series from 2006 to 2100. Sci. Data 7 (1), 248. https://doi.org/10.1038/s41597-020-00587-y.

Krzanowski, W.J., 2001. Data-based interval estimation of classification error rates. J. Appl. Stat. 28 (5), 585–595. https://doi.org/10.1080/02664760120047915.

Lark, R.M., 2009. Kriging a soil variable with a simple nonstationary variance model. J. Agric. Biol. Environ. Stat. 14 (3), 301–321. https://doi.org/10.1198/jabes.2009.07060.

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.

Li, Y., Li, M., Li, C., Liu, Z., 2020. Forest aboveground biomass estimation using landsat 8 and sentinel-1A data with machine learning algorithms. Sci. Rep. 10 (1), 9952. https://doi.org/10.1038/s41598-020-67024-3.

Lohr, S.L., 2019. Sampling Design and Analysis. Chapman and Hall/CRC. https://doi.org/10.1201/9780429296284.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12 (9), 1620–1633. 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. Ecol. Model. 411, 108815 https://doi.org/10.1016/j.ecolmodel.2019.108815.

Milà, C., Mateu, J., Pebesma, E., Meyer, H., 2022. Nearest neighbour distance matching leave-one-out cross-validation for map validation. Methods Ecol. Evol. https://doi.org/10.1111/2041-210X.13851.

Morais, T.G., Teixeira, R.F.M., Figueiredo, M., Domingos, T., 2021. The use of machine learning methods to estimate aboveground biomass of grasslands: a review. Ecol. Indic. 130, 108081 https://doi.org/10.1016/j.ecolind.2021.108081.

Nash, J.E., Sutcliffe, J.V., 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.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30 (7), 683–691. https://doi.org/10.1016/j.cageo.2004.03.012.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat. Commun. 11 (1), 4540. https://doi.org/10.1038/s41467-020-18321-y.

Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7 (1), 217–240. https://doi.org/10.5194/soil-7-217-2021.

Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M.C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H., Silva, C.E., Armston, J., Dubayah, R., Blair, J. B., Hofton, M., 2021. Mapping global forest canopy height through integration of GEDI and Landsat data. Remote Sens. Environ. 253, 112165 https://doi.org/10.1016/j.rse.2020.112165.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

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. https://doi.org/10.1111/ecog.02881.

Samuels, M.L., 1991. Statistical reversion toward the mean: more universal than regression toward the mean. Am. Stat. 45 (4), 344–346. https://doi.org/10.1080/00031305.1991.10475833.

Sanderman, J., Hengl, T., Fiske, G., Solvik, K., Adame, M.F., Benson, L., Bukoski, J.J., Carnell, P., Cifuentes-Jara, M., Donato, D., Duncan, C., Eid, E.M., Ermgassen, P.Z., Lewis, C.J.E., Macreadie, P.I., Glass, L., Gress, S., Jardine, S.L., Jones, T.G., Nsombo, E.N., Rahman, M.M., Sanders, C.J., Spalding, M., Landis, E., 2018. A global map of mangrove forest soil carbon at 30 m spatial resolution. Environ. Res. Lett. 13 (5) https://doi.org/10.1088/1748-9326/aabe1c.

Santoro, M., Cartus, O., Carvalhais, N., Rozendaal, D.M.A., Avitabile, V., Araza, A., de Bruin, S., Herold, M., Quegan, S., Rodríguez-Veiga, P., Balzter, H., Carreiras, J., Schepaschenko, D., Korets, M., Shimada, M., Itoh, T., Moreno Martínez, Á., Cavlovic, J., Cazzolla Gatti, R., da Conceição Bispo, P., et al., 2021. The global forest above-ground biomass pool for 2010 estimated from high-resolution satellite observations. Earth Syst. Sci. Data 13 (8), 3927–3950. https://doi.org/10.5194/essd-13-3927-2021.

de Sousa, K., van Etten, J., Poland, J., Fadda, C., Jannink, J.-L., Kidane, Y.G., Lakew, B.F., Mengistu, D.K., Pè, M.E., Solberg, S.Ø., Dell'Acqua, M., 2021. Data-driven decentralized breeding increases prediction accuracy in a challenging crop production environment. Commun. Biol. 4 (1), 944. https://doi.org/10.1038/s42003-021-02463-w.

Steele, B.M., Patterson, D.A., Redmond, R.L., 2003. Toward estimation of map accuracy without a probability test sample. Environ. Ecol. Stat. 10 (3), 333–356. https://doi.org/10.1023/A:1025111108050.

Stehman, S.V., 2009. Sampling designs for accuracy assessment of land cover. Int. J. Remote Sens. 30 (20), 5243–5272. https://doi.org/10.1080/01431160903131000.

Stehman, S.V., Foody, G.M., 2019. Key issues in rigorous accuracy assessment of land cover products. Remote Sens. Environ. 231 https://doi.org/10.1016/j.rse.2019.05.018.

Su, H., Shen, W., Wang, J., Ali, A., Li, M., 2020. Machine learning and geostatistical approaches for estimating aboveground biomass in Chinese subtropical forests. Forest. Ecosystems 7 (1). https://doi.org/10.1186/s40663-020-00276-7.

Wadoux, A.M.J.C., Brus, D.J., Heuvelink, G.B.M., 2018. Accounting for non-stationary variance in geostatistical mapping of soil properties. Geoderma 324, 138–147. https://doi.org/10.1016/j.geoderma.2018.03.010.

Wadoux, A.M.J.C., Heuvelink, G.B.M., de Bruin, S., Brus, D.J., 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.

Walvoort, D.J.J., Brus, D.J., de Gruijter, J.J., 2010. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Comput. Geosci. 36 (10), 1261–1267. https://doi.org/10.1016/j.cageo.2010.04.005.

Wright, M.N., Ziegler, A., 2017. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77 (1), 1–17. https://doi.org/10.18637/jss.v077.i01.

Yousef, W.A., 2021. Estimating the standard error of cross-validation-based estimators of classifier performance. Pattern Recogn. Lett. 146, 115–125. https://doi.org/10.1016/j.patrec.2021.02.022.