Ecological Modelling 457 (2021) 109692

Contents lists available at ScienceDirect

Ecological Modelling

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

Short communication

Spatial cross-validation is not the right way to evaluate map accuracy

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

[a] Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Australia
[b] Soil Geography and Landscape group, Wageningen University & Research, The Netherlands
[c] Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, The Netherlands
[d] Biometris, Wageningen University & Research, The Netherlands

ARTICLE INFO

Keywords:
Map quality
Model performance
Above-ground biomass
Sampling theory
Design-based
Model-based
Random forest
Design-unbiased

ABSTRACT

For decades scientists have produced maps of biological, ecological and environmental variables. These studies commonly evaluate the map accuracy through cross-validation with the data used for calibrating the underlying mapping model. Recent studies, however, have argued that cross-validation statistics of most mapping studies are optimistically biased. They attribute these overoptimistic results to a supposed serious methodological flaw in standard cross-validation methods, namely that these methods ignore spatial autocorrelation in the data. They argue that spatial cross-validation should be used instead, and contend that standard cross-validation methods are inherently invalid in a geospatial context because of the autocorrelation present in most spatial data. Here we argue that these studies propagate a widespread misconception of statistical validation of maps. We explain that unbiased estimates of map accuracy indices can be obtained by probability sampling and design-based inference and illustrate this with a numerical experiment on large-scale above-ground biomass mapping. In our experiment, standard cross-validation (i.e., ignoring autocorrelation) led to smaller bias than spatial cross-validation. Standard cross-validation was deficient in case of a strongly clustered dataset that had large differences in sampling density, but less so than spatial cross-validation. We conclude that spatial cross-validation methods have no theoretical underpinning and should not be used for assessing map accuracy, while standard cross-validation is deficient in case of clustered data. Model-free, design-unbiased and valid accuracy assessment is achieved with probability sampling and design-based inference. It is valid without the need to explicitly incorporate or adjust for spatial autocorrelation and perfectly suited for the validation of large scale biological, ecological and environmental maps.

1. Introduction

In recent years, mapping studies have provided new insights into continental and global patterns of biogeographical variables. Some recent examples include global scale maps of soil fungi (Tedersoo et al., 2014), bacteria (Delgado-Baquerizo et al., 2018) and nematodes (Van Den Hoogen et al., 2019), landcover change (Song et al., 2018), forest cover change (Hansen et al., 2013) and aboveground biomass (Baccini et al., 2012), among others.

Commonly in these studies, the map accuracy is evaluated using statistical measures that evaluate how close the predictions $\hat{z}(s)$ are to the reality $z(s)$ for a set of locations $s \in D$, where D is the area of interest, i.e. the population. In practice, the area of interest is often discretized by overlaying it with a fine grid. Popular population map accuracy indices are the mean error (ME), root mean squared error (RMSE), the squared Pearson's correlation coefficient (r^2) and the Nash–Sutcliffe model efficiency coefficient (Nash and Sutcliffe, 1970). These indices are computed either by collecting a new set of data or by using the existing dataset for both model calibration and validation.

Collecting a new set of data is ideally done by probability sampling. The sample data are used to estimate the map accuracy indices through design-based statistical inference based on classical sampling theory. This method is statistically sound and has been extensively described in the statistical (Cochran, 1977) and environmental science (e.g. De Gruijter et al., 2006; Gregoire and Valentine, 2007) literature. But when data are scarce, it is sensible to use the available data for both model calibration and validation (Burt et al., 2009). In such case, the existing dataset is split into two subsets called calibration and validation folds. The calibration fold is used to calibrate a mapping model and make predictions whereas the validation fold is used to estimate the map accuracy indices. This procedure can be repeated several times, as in bootstrapping when multiple bootstrap samples

* Correspondence to: Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, New South Wales, Australia.
E-mail address: alexandre.wadoux@sydney.edu.au (A.M.J.-C. Wadoux).

https://doi.org/10.1016/j.ecolmodel.2021.109692
Received 16 July 2021; Received in revised form 30 July 2021; Accepted 2 August 2021
Available online 12 August 2021

with replacement are used for calibration and prediction, or as in cross-validation (CV) (Hastie et al., 2009). In CV, the data are split randomly into K disjoint folds. Each fold is put aside in turn and used to evaluate the predictions obtained from a model calibrated on the remaining $K-1$ folds. Conventionally, K is set to 10 (10-fold CV), but K can also be equal to the sample size, as in leave-one-out CV. An advantage of CV and other data-splitting strategies is that the existing data are used to evaluate map accuracy, without the cost of additional sampling.

Several recent studies (Brenning, 2005; Le Rest et al., 2014; Roberts et al., 2017; Ploton et al., 2020), however, contend that statistical validation of maps should account for spatial autocorrelation between data points. Data collected at points that are geographically close generally are more similar than at points that are geographically distant. As a consequence, these studies claim that map accuracy indices as derived using standard CV are biased because calibration points are not statistically independent from validation points. This conception of model validation has led to the recent development of CV techniques that avoid spatial autocorrelation, such as spatial K-fold CV and buffered leave-one-out CV (B-LOO CV). The study of Ploton et al. (2020), for example, asserts that validation statistics should only be computed on validation points that are spatially independent from the calibration points, and found in an experiment on mapping the above-ground forest biomass in central Africa that this resulted in quasi-null predictive model performance.

Here we argue that spatial CV techniques presented by recent studies on biological, ecological and environmental mapping (such as in Le Rest et al., 2014; Ploton et al., 2020; Hengl et al., 2021, for example) give rise to a misconception of statistical validation of maps in a spatial context. There is a serious risk that practitioners misinterpret spatial CV techniques and reject long-standing, standard and statistically valid methods for assessing map accuracy. In the following, we explain why we should not use spatial CV techniques for estimating map accuracy indices and instead adhere to statistically rigorous methods of validation via probability sampling and design-based inference. Our theoretical argument is illustrated with a case study.

2. Evaluation of map accuracy with probability sampling

Map accuracy indices are defined as population parameters. The RMSE, for example, as a finite population parameter is defined as $\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{z}(\mathbf{s}_i) - z(\mathbf{s}_i))^2}$, i.e. the square root of the squared prediction errors averaged over all N units (i.e. grid cells) in the population. Note that for infinite populations the map accuracy indices are defined as an integral. In practice it is usually impossible to compute the map accuracy indices, because one would need to take a census of the whole population. Instead, a subset of n units is selected (where typically $n \ll N$), and this sample is used to *estimate* the map accuracy indices (i.e. the population validation parameters). If the sample is a probability sample, classical sampling theory can be used to estimate these population validation parameters, using design-based estimation (Stehman, 1999; De Gruijter et al., 2006; Gregoire and Valentine, 2007; Stehman and Foody, 2009; Brus et al., 2011). This has the important advantage that one can prove that the estimates are unbiased and that valid confidence intervals can be computed. Probability samples have two characteristics: all units in the population must have a positive probability of being selected, and these "inclusion" probabilities must be known for at least the selected population units (points).

The most basic probability sampling design is simple random sampling, in which case all samples of a given size (number of units) have equal probability of being selected, and a design-based estimate of the population RMSE is obtained by $\widehat{\text{RMSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{z}(\mathbf{s}_i) - z(\mathbf{s}_i))^2}$. There are also more complex probability sampling designs exploiting the spatial structure of the data, such as stratified random sampling (De Gruijter et al., 2015), balanced sampling (Deville and Tillé, 2004; Brus, 2015), and the local pivotal method (Grafström et al., 2012). These designs are often more efficient than simple random sampling to estimate population parameters. Design-based estimators of map accuracy indices are model-free and design-unbiased, i.e. over repeated sampling with the design used to select a validation sample, the average of the estimated map accuracy is equal to the true map accuracy.

Note that when validating by design-based inference, validation locations are allowed to be geographically close to calibration locations, also in a population with spatial structure. The prediction errors at two independently selected locations are design-independent (i.e. when the points are selected independently from each other, Brus, 2021), regardless of whether they are selected from a spatially structured population or not (Gregoire and Valentine, 2007; Stehman and Foody, 2009; Brus, 2021). This refutes the core message of several recent studies (e.g. Brenning, 2005; Ploton et al., 2020), that spatial autocorrelation invalidates map accuracy assessment. Spatial autocorrelation need not be explicitly incorporated in design-based estimation of the validation indices because in probability sampling the validation data are *design-independent*. Several studies have used a probability sampling strategy for map validation, for examples see Kempen et al. (2009), Olofsson et al. (2012) and Boschetti et al. (2016).

3. Evaluation of map accuracy with cross-validation

It is not always feasible to collect an additional probability sample for map validation, given the available resources and time (Gregoire and Valentine, 2007; Duncanson et al., 2019). In such case, cross-validation can be used to obtain estimates of the map accuracy (Steele et al., 2003). Cross-validation makes use of the calibration dataset, which typically is not a probability sample of the mapping area, so that there is no possibility to determine how close the estimates are to the population validation parameters.

In standard cross-validation using a non-probability calibration sample we cannot make use of classical sampling theory to derive the probability distribution of the estimation errors of the accuracy indices and, for example, cannot prove that the estimation is design-unbiased. In practice cross-validation may approximate the population validation parameters well, but this depends on the sample size and the distribution of the sample locations across the study area. For example, if sample locations tend to be spatially clustered and large parts of the study area have low to zero sampling density, then the map accuracy indices are likely to be over-optimistic. Computing map accuracy indices with standard CV is not ideal (ideally, probability sampling is used), but the map accuracy indices obtained this way can still be useful. In many cases, map accuracy indices estimated by standard CV strategies may be close to the (unknown) population indices, but this is an assumption that cannot be verified in practice.

In spatial CV strategies, validation points are forced to be geographically distant from calibration points, by selecting validation folds that are assumed statistically independent (i.e. model-independent) from calibration folds. Commonly spatial partitioning (also known as "blocking") and buffering, or a combination thereof, are used to achieve independence. In spatial partitioning (spatial K-fold CV), the geographic space is divided into K spatially disjoint subsets. The partitions can be determined by a coarse square grid of K cells or by clustering the spatial coordinates of the data set into K clusters. In buffered leave-one-out CV (B-LOO-CV), observations that are within a distance-based radius from a validation point are not considered for model calibration. This radius is usually taken to be larger than the range of a variogram computed on the whole data set, or as computed on the residuals of this data set after a model is calibrated and model predictions subtracted from the observations.

Spatial CV strategies remove entire portions of the geographic and hence also the covariate space, causing under-representation of environmental conditions similar to those at validation locations. This is particularly a problem for biological, ecological and environmental

variables which are geographically structured: the environmental conditions in the validation fold might be unseen in the calibration folds, and the model is likely to predict outside the feature space covered by the joint set of covariates. The map accuracy indices estimated in this way are likely systematically and potentially severely over-pessimistic, as would be the case if environmental conditions between validation and calibration folds are very different.

Below we investigate the degree in which this occurs and illustrate our theoretical arguments with a simple case study.

4. Case study

We built a random forest model for large-scale mapping of above-ground forest biomass (AGB) for an area in the Amazon basin (Fig. 1a) using a large set of ecological covariates as predictors. To do so, we selected a large rectangular area (size 928 km × 1642 km) of above-ground live woody biomass data (in Mg·ha^{-1}) from the "Baccini" dataset (Baccini et al., 2012) as the response variable and source of calibration data. This allowed us to compute the population map accuracy indices and evaluate the various cross-validation approaches. Note that there are no fundamental objections to using a proxy of the true AGB as a response variable because for our purposes the only requirement is that it must have spatial structure. The Baccini map was aggregated to a spatial resolution of 1 km × 1 km. We prepared a stack of 28 ecologically relevant covariates for the same extent, representing mean climatic conditions, climate seasonality and extreme conditions, relief, soil properties and six long-term average MODIS bands. The list of covariates, their unit and source is provided in the Supplementary Material. All covariates were either resampled using bilinear interpolation or aggregated to conform with the grid of the above-ground biomass map.

Considering the spatially exhaustive values of above-ground biomass (AGB) as our population of interest, we repeatedly (500 times) selected calibration samples of 500 grid cells ($n = 500$) using: (i) systematic random sampling; (ii) simple random sampling; and (iii) two-stage cluster random sampling. For each sample, we paired the corresponding 500 AGB values and the stack of 28 covariates. We applied three cross-validation (CV) strategies: (i) random K-fold CV; (ii) spatial K-fold CV;

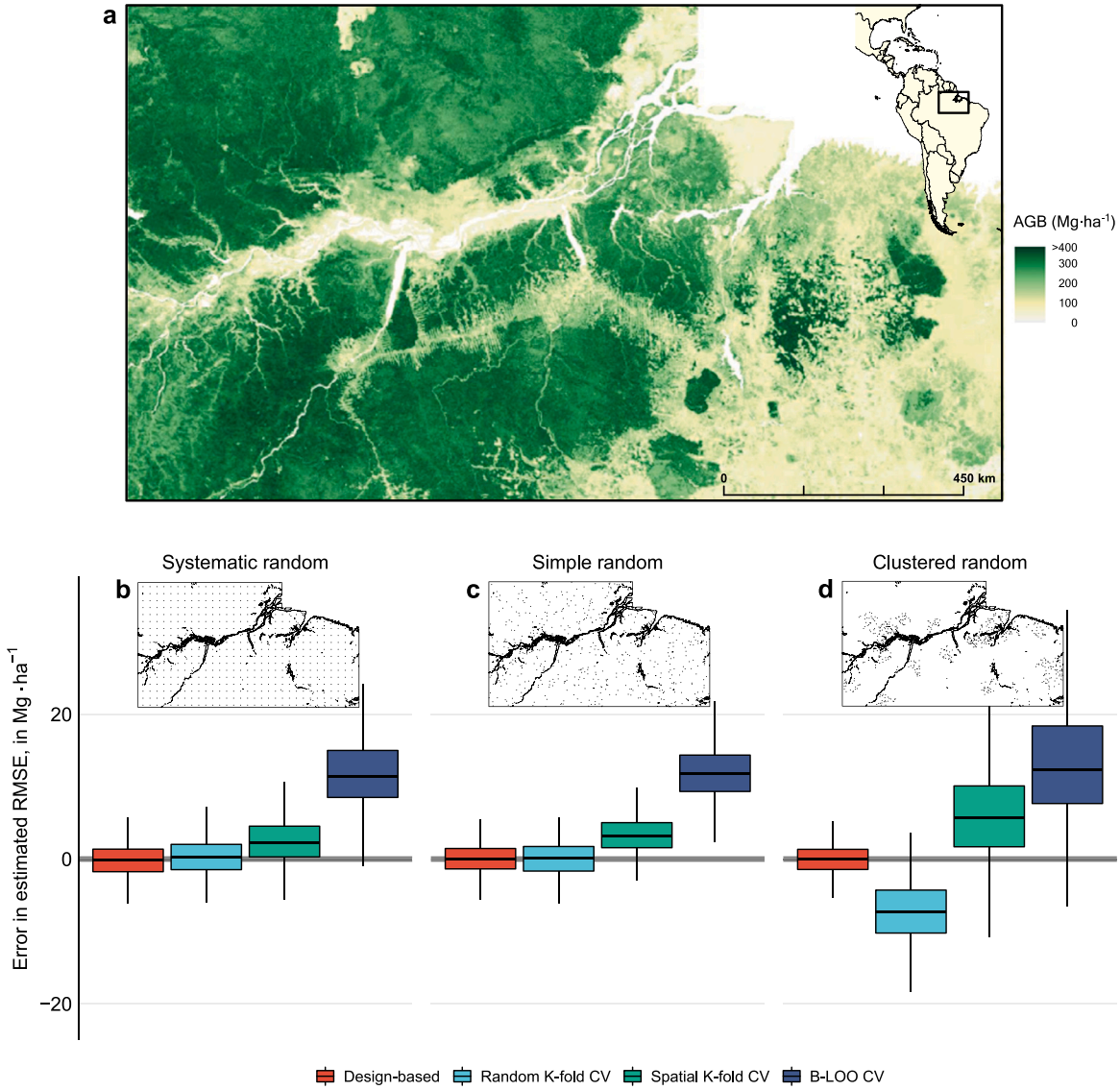

Fig. 1. Overview of the study area and results of the evaluation of validation strategies. **a** Study area in the Amazon basin with values of the above-ground biomass, according to the Baccini map (Baccini et al., 2012). **b–d** Error in estimates of the population RMSE (in Mg·ha^{-1}) for calibration samples collected by systematic random (**b**), simple random (**c**) and two-stage cluster random (**d**) sampling. Note that the horizontal grey line at 0 in **b–d** effectively refers to the population RMSE, because deviations from the population RMSE are plotted. The sampling locations shown in the maps in **b–d** are one realization out of 500.

and (iii) buffered leave-one-out CV (B-LOO CV). We also performed design-based validation by repeatedly selecting a separate probability sample of size 500 by simple random sampling without replacement. For the two spatial cross-validation strategies (i.e. spatial K-fold CV and B-LOO CV), we applied the same computational methodology as in Ploton et al. (2020) and used an exclusion radius slightly larger than the autocorrelation range of the AGB empirical variogram, in our case 350 km. The random forest (RF) algorithm was used for modelling and prediction, and the RMSE was estimated. Each population RMSE was subtracted from the RMSE estimate. The population RMSE was obtained by predicting the AGB by the RF model at all locations N in the area, subtracting these predictions from the "true" AGB (i.e., the Baccini map) and thus computing the population RMSE. We assessed the error in the RMSE estimated from each sample by computing the difference between the estimated RMSE and the population RMSE. We repeated the selection of the samples and the procedure for the four validation strategies 500 times to obtain the sampling distributions of the RMSE estimation errors.

Note that we adopted a model-based approach for mapping, as is nearly always done in practice. Selection of calibration locations using probability sampling designs was therefore not essential. We only did this to automate the repeated sampling and do so in a reproducible way. By assessing prediction performance for three very different sampling designs, our experiment evaluates the efficiency of different validation strategies for a range of spatial configurations of acquired calibration data. The results are illustrative for both dispersed and clustered sampling scenarios, also when the calibration sample is not a probability sample, as typically is the case in biological, ecological and environmental mapping.

Fig. 1b-d shows that design-based estimation of the population RMSE is unbiased and that the estimates have little variation. These are attractive properties that show the superiority of design-based validation to evaluate the prediction performance of the mapping model and estimate the map accuracy, but as noted before it requires probability sampling from the population. Cross-validation using standard random K-fold CV is nearly unbiased for the systematic and simple random sampling designs, but too optimistic in the case of clustered sampling. The two spatial cross-validation methods are too pessimistic, with B-LOO CV severely overestimating the RMSE in all cases.

The pessimistic results of the spatial cross-validation methods are likely caused by over-representation of environmental conditions distinct from the environmental conditions at the calibration points, and under-representation of environmental conditions similar to those at the calibration locations. Standard K-fold CV was too optimistic in case of clustered sampling because each validation point had nearby calibration points, while most points in the map did not. This could be remedied by using leave-cluster-out CV, which is another variant of spatial CV, but this would likely overestimate the RMSE.

5. Discussion

Existing methods of map validation using probability sampling and design-based inference are perfectly valid for map accuracy assessment, without the need to adjust for spatial autocorrelation. These methods provide design-unbiased estimates of the map accuracy, which cannot be guaranteed by CV strategies (i.e. neither standard nor spatial CV). Spatial CV strategies performed poorly and severely overestimated the population RMSE in all sampling design cases considered in the case study. The over-pessimistic results of spatial CV are essentially caused by the fact that in this method only those parts of the mapped area that are distant from calibration points are validated. Prediction performance will tend to be poorer in subareas that have no nearby calibration points. The map accuracy indices obtained this way are systematically off, because of under-representation of environmental conditions similar to those of the folds used in calibrating the model.

Spatial CV strategies are highly subjective and depend not only on the sampling design on which the folds are defined, but also on the method to determine the spatial partitioning (spatial K-fold CV) and the radius (the distance of exclusion, in B-LOO CV). Further, spatial CV strategies suffer from the fundamental problem of dealing with two conflicting objectives: exclude validation data that are geographically close to calibration data to achieve "spatial independence" (model-based independence under a geostatistical model, i.e. with distances larger than the variogram range), and avoid extrapolation in the geographic and covariate space. There seems to be no proper solution to this paradox. The only sound way to validate a map is the use of probability sampling and design-based inference. Scientists and practitioners can confidently proceed knowing that the map accuracy obtained by standard methods based on sampling theory are not invalidated by claims made in spatial CV studies. Spatial CV strategies (and standard CV in case of clustered data) are not appropriate for map validation. Spatial autocorrelation does not invalidate map accuracy assessment in a design-based inference framework. For map accuracy assessment, the design-based approach makes no assumptions on spatial autocorrelation of prediction errors, whereas such assumptions are required in a model-based approach, thus giving rise to discussions on the validity of the estimated map accuracy.

Some studies on spatial CV claim to focus on the validation of the mapping model instead of the validation of a map. We call for a better articulation of what validating a mapping model means. Ploton et al. (2020), for example, did not evaluate the strength of the relationships between the response variable and the covariates, nor did they attempt to explain the causal determinism of the spatial distribution of AGB. They assessed the accuracy of the AGB map, using the RMSE as an overall statistic of the map quality. It remains unclear what validation of a mapping model means and how this differs from validation of a map. This lack of definition causes confusion, so that there appears to be a mismatch between the claims (i.e. validation of the mapping model) and results (i.e. estimating map accuracy indices) in studies on spatial CV (e.g. Brenning, 2005; Meyer et al., 2019; Ploton et al., 2020). In mapping studies, the objective is to validate the map, i.e. to assess the map accuracy. Map accuracy assessment is relevant in many disciplines, some recent examples of which are validation of global scale maps of soil fungi (Tedersoo et al., 2014), bacteria (Delgado-Baquerizo et al., 2018) and nematodes (Van Den Hoogen et al., 2019), landcover change (Song et al., 2018), forest cover change (Hansen et al., 2013) and aboveground biomass (Baccini et al., 2012), among others.

Finally, we stress that the methodology communicated in this article for the validation of maps with probability sampling and design-based inference is directly applicable to large-scale mapping studies. Strahler et al. (2006) and Olofsson et al. (2012) give recommendations for estimating the accuracy of global maps with sampling designs satisfying the definition of a probability sample. Depending on the objective, these designs typically use stratification or clustering to balance precision of regional map accuracy assessment and cost associated to the collection of the validation data. The key aspect of large-scale map accuracy assessment relies on the use of probability sampling designs and design-based inference of the map accuracy indices. Examples of studies using these strategies are McRoberts et al. (2019) and Stehman et al. (2012) for the validation of a global biomass map and landcover map, respectively. More efforts should be invested in this direction.

6. Conclusion

We have shown that spatial cross-validation strategies resulted in a grossly pessimistic map accuracy assessment, and gave no improvement over standard cross-validation. Both standard and spatial cross-validation methods may provide biased estimates of map accuracy. In our case study, spatial cross-validation strategies severely underestimated the map quality, while standard CV overestimated it in case

of clustered data. More importantly, studies of spatial cross-validation propagate a widespread misconception on the statistical validation of maps. Map accuracy ideally should be estimated with probability sampling and design-based statistical inference. Such methodology and inference is valid without the need to adjust for spatial autocorrelation in the data. Scientists and practitioners can confidently proceed knowing that the map accuracy obtained by standard methods based on sampling theory are valid. Validation based on rigorous design-based inference is also feasible for large-scale and global maps of biological, ecological and environmental variables.

CRediT authorship contribution statement

Alexandre M.J.-C. Wadoux: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Gerard B.M. Heuvelink:** Conceptualization, Writing – original draft , Writing – review & editing. **Sytze de Bruin:** Conceptualization, Methodology, Writing – review & editing. **Dick J. Brus:** Conceptualization, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Code and data availability

All statistical analysis code, data and results of the case study are available via an open access link at https://github.com/AlexandreWadoux/SpatialValidation. For the spatial cross-validation strategies we used the original code provided in Ploton et al. (2020).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ecolmodel.2021.109692.

References

Baccini, A., Goetz, S.J., Walker, W., Laporte, N.T., Sun, M., Sulla-Menashe, D., Hackler, J., Beck, P.S.A., Dubayah, R., Friedl, M.A., et al., 2012. Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. Nature Clim. Change 2, 182–185.

Boschetti, L., Stehman, S.V., Roy, D.P., 2016. A stratified random sampling design in space and time for regional to global scale burned area product validation. Remote Sens. Environ. 186, 465–478.

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

Brus, D.J., 2015. Balanced sampling: A versatile sampling approach for statistical soil surveys. Geoderma 253, 111–121.

Brus, D.J., 2021. Statistical approaches for spatial sample survey: Persistent misconceptions and new developments. Eur. J. Soil Sci. 72, 686–703.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62, 394–407.

Burt, J.E., Barber, G.M., Rigby, D.L., 2009. Elementary Statistics for Geographers, third ed. Guilford Press, New York.

Cochran, W.G., 1977. Sampling Techniques, third ed. John Wiley & Sons, New York.

De Gruijter, J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer Science & Business Media, Berlin.

De Gruijter, J.J., Minasny, B., McBratney, A.B., 2015. Optimizing stratification and allocation for design-based estimation of spatial means using predictions with error. J. Surv. Stat. Methodol. 3, 19–42.

Delgado-Baquerizo, M., Oliverio, A.M., Brewer, T.E., Benavent-González, A., Eldridge, D.J., Bardgett, R.D., Maestre, F.T., Singh, B.K., Fierer, N., 2018. A global atlas of the dominant bacteria found in soil. Science 359, 320–325.

Deville, J.-C., Tillé, Y., 2004. Efficient balanced sampling: The cube method. Biometrika 91, 893–912.

Duncanson, L., Armston, J., Disney, M., Avitabile, V., Barbier, N., Calders, K., Carter, S., Chave, J., Herold, M., Crowther, T.W., et al., 2019. The importance of consistent global forest aboveground biomass product validation. Surv. Geophys. 40, 979–999.

Grafström, A., Lundström, N.L., Schelin, L., 2012. Spatially balanced sampling through the pivotal method. Biometrics 68, 514–520.

Gregoire, T.G., Valentine, H.T., 2007. Sampling Strategies for Natural Resources and the Environment. CRC Press, Boca Raton, USA.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., et al., 2013. High-resolution global maps of 21st-century forest cover change. Science 342, 850–853.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed. Springer Science & Business Media, New York.

Hengl, T., Miller, M.A., Križan, J., Shepherd, K.D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S.M., et al., 2021. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. Sci. Rep. 11, 1–18.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1: 50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma 151, 311–326.

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 Ecol. Biogeogr. 23, 811–820.

McRoberts, R.E., Næsset, E., Saatchi, S., Liknes, G.C., Walters, B.F., Chen, Q., 2019. Local validation of global biomass maps. Int. J. Appl. Earth Obs. Geoinf. 83, 101931.

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.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. J. Hydrol. 10, 282–290.

Olofsson, P., Stehman, S.V., Woodcock, C.E., Sulla-Menashe, D., Sibley, A.M., Newell, J.D., Friedl, M.A., Herold, M., 2012. A global land-cover validation data set, part I: Fundamental design principles. Int. J. Remote Sens. 33, 5768–5788.

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

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., et al., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929.

Song, X.-P., Hansen, M.C., Stehman, S.V., Potapov, P.V., Tyukavina, A., Vermote, E.F., Townshend, J.R., 2018. Global land change from 1982 to 2016. Nature 560, 639–643.

Steele, B.M., Patterson, D.A., Redmond, R.L., 2003. Toward estimation of map accuracy without a probability test sample. Environ. Ecol. Stat. 10, 333–356.

Stehman, S.V., 1999. Basic probability sampling designs for thematic map accuracy assessment. Int. J. Remote Sens. 20, 2423–2441.

Stehman, S.V., Foody, G.M., 2009. Accuracy assessment. In: The SAGE Handbook of Remote Sensing. SAGE, London, UK, pp. 297–309.

Stehman, S.V., Olofsson, P., Woodcock, C.E., Herold, M., Friedl, M.A., 2012. A global land-cover validation data set, II: Augmenting a stratified sampling design to estimate accuracy by region and land-cover class. Int. J. Remote Sens. 33, 6975–6993.

Strahler, A.H., Boschetti, L., Foody, G.M., Friedl, M.A., Hansen, M.C., Herold, M., Mayaux, P., Morisette, J.T., Stehman, S.V., Woodcock, C.E., 2006. Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps. European Communities, Luxembourg.

Tedersoo, L., Bahram, M., Põlme, S., Kõljalg, U., Yorou, N.S., Wijesundera, R., Ruiz, L.V., Vasco-Palacios, A.M., Thu, P.Q., Suija, A., et al., 2014. Global diversity and geography of soil fungi. Science 346.

Van Den Hoogen, J., Geisen, S., Routh, D., Ferris, H., Traunspurger, W., Wardle, D.A., De Goede, R.G.M., Adams, B.J., Ahmad, W., Andriuzzi, W.S., et al., 2019. Soil nematode abundance and functional group composition at a global scale. Nature 572, 194–198.