Ecological Modelling 411 (2019) 108815

Contents lists available at ScienceDirect

Ecological Modelling

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

Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction

Hanna Meyer[a,*], Christoph Reudenbach[b], Stephan Wöllauer[b], Thomas Nauss[b]

[a] Institute for Geoinformatics, Westfälische Wilhelms-Universität Münster, Heisenbergstr. 2, 48149 Münster, Germany
[b] Faculty of Geography, Philipps-Universität Marburg, Deutschhausstr. 10, 35037 Marburg, Germany

ARTICLE INFO

Keywords:
Cross-validation
Environmental monitoring
Machine learning
Overfitting
Random Forests
Remote sensing

ABSTRACT

Machine learning algorithms find frequent application in spatial prediction of biotic and abiotic environmental variables. However, the characteristics of spatial data, especially spatial autocorrelation, are widely ignored. We hypothesize that this is problematic and results in models that can reproduce training data but are unable to make spatial predictions beyond the locations of the training samples. We assume that not only spatial validation strategies but also spatial variable selection is essential for reliable spatial predictions.

We introduce two case studies that use remote sensing to predict land cover and the leaf area index for the "Marburg Open Forest", an open research and education site of Marburg University, Germany. We use the machine learning algorithm Random Forests to train models using non-spatial and spatial cross-validation strategies to understand how spatial variable selection affects the predictions.

Our findings confirm that spatial cross-validation is essential in preventing overoptimistic model performance. We further show that highly autocorrelated predictors (such as geolocation variables, e.g. latitude, longitude) can lead to considerable overfitting and result in models that can reproduce the training data but fail in making spatial predictions. The problem becomes apparent in the visual assessment of the spatial predictions that show clear artefacts that can be traced back to a misinterpretation of the spatially autocorrelated predictors by the algorithm. Spatial variable selection could automatically detect and remove such variables that lead to overfitting, resulting in reliable spatial prediction patterns and improved statistical spatial model performance.

We conclude that in addition to spatial validation, a spatial variable selection must be considered in spatial prediction models of ecological data to produce reliable results.

1. Introduction

A key task in ecology is studying the spatial or spatio-temporal patterns of ecosystem variables, e.g. climate dynamics (Appelhans et al., 2015), variability of soil properties (Gasch et al., 2015) or distribution of vegetation types (Juel et al., 2015). Spatially continuous datasets of ecosystem variables are needed to analyze the spatial patterns and dynamics. However, ecological variables are typically acquired through field work, which only provides data with a limited spatial extent, such as from climate stations, soil profiles or plot-based vegetation records. These data do not provide spatially continuous information about the variable of interest. Predictive modelling is a method commonly used to derive spatially continuous datasets from limited field data (e.g. Lary et al., 2016). In predictive modelling, field data is used to train statistical models using spatially continuous predictor variables derived from remote sensing imagery. The resulting model is then used to make predictions in space, i.e. beyond the locations used for model training.

Most contemporary predictive modelling approaches use flexible machine learning algorithms, which can approximate the nonlinear and complex relationships found in nature. Recent software developments have simplified the application of machine learning algorithms (e.g. for R see Kuhn and Johnson, 2013). Noteworthy, however, is that machine learning is applied to ecological spatial modelling the same way as it is in other disciplines, while ignoring the unique characteristics of spatial environmental data. Yet, spatial (and temporal) dependencies differentiate spatial data from "ordinary" data and complicate the use of machine learning – due to the nature of the data, we cannot assume samples are identically and independently distributed (i.i.d assumption) (Xie et al., 2017). This is especially true when data are sampled in spatial clusters, which is a common design for providing ground truth data used in predictive modelling of ecological data.

* Corresponding author.
E-mail address: hanna.meyer@uni-muenster.de (H. Meyer).

https://doi.org/10.1016/j.ecolmodel.2019.108815
Received 26 March 2019; Received in revised form 29 August 2019; Accepted 29 August 2019

Previous studies in spatial applications of machine learning algorithms have widely ignored the spatial dependencies in the data. One problem of ignoring spatial dependencies in prediction methods becomes obvious in the error assessment of spatial predictive models. Many authors have shown that the commonly used random cross-validation provides considerably overoptimistic error estimates due to the problem of autocorrelation (Bahn and McGill, 2013; Micheletti et al., 2014; Juel et al., 2015; Gasch et al., 2015; Gudmundsson and Seneviratne, 2015; Roberts et al., 2017; Meyer et al., 2018). Hence cross-validation strategies based on random data splitting fail to assess a model's performance in terms of spatial mapping and only validate its ability to reproduce the sampling data. Several methods for spatial cross-validation have been proposed to account for spatial dependencies in the data (Brenning, 2005; Le Rest et al., 2014; Pohjankukka et al., 2017; Roberts et al., 2017; Meyer et al., 2018; Valavi et al., 2018). While spatial cross-validation can provide objective and meaningful error estimates, the algorithms' strong performance with random subsets and complete failures when predictions are made beyond the spatial extent of the training samples still remains an issue.

Meyer et al. (2018) have shown for spatio-temporal data that spatial (or spatio-temporal) dependencies can cause a misinterpretation of certain predictor variables which makes flexible algorithms fail when predicting beyond the location of the training data. Spatial dependencies in predictor variables are most apparent in "geolocation" predictors that describe the spatial location of the training samples (e.g. coordinates, elevation, euclidean distances and all derivations of these data). Hence, we assume that including predictor variables that describe the spatial location are problematic and prevent spatial models from making meaningful contributions to ecological research. However, predictor variables that describe the spatial location rather than the environmental properties are commonly included. Spatial coordinates are used especially often (Li et al., 2011; Langella et al., 2010; Shi et al., 2015; Janatian et al., 2017; Walsh et al., 2017; Jing et al., 2016; Wang et al., 2017; Georganos et al., 2019). Distances to certain points (e.g. Hengl et al., 2018) or Euclidean distance to the corner coordinates of the model domain (e.g. Behrens et al., 2018) have also been suggested as predictors and included in models.

This study uses autocorrelated spatial data to investigate the sensitivity of machine learning applications to commonly applied geolocation predictors and shows pathways towards an automatic selection of predictors that cannot be incorporated in spatial prediction tasks. We assume that spatial models cannot handle predictor variables that are highly autocorrelated in space (e.g. geolocation) due to spatial dependencies in the training data. Algorithms can easily misinterpret such variables, leading the model to make erroneous predictions outside of the locations of the training data. The problem becomes obvious in the limited spatial performance of the model as well as in visually obvious artefacts in the spatial predictions. We therefore assume that spatial variable selection is essential for automatically removing variables counterproductive to spatial mapping to provide scientifically valuable results.

We use two examples of classic prediction tasks in environmental science to investigate our hypotheses. First, we perform a Land Use/Land Cover (LULC) classification, which is a common field for applying machine learning-based predictive modelling in the context of ecology and remote sensing. The study area is located around the "Marburg Open Forest", an open research and education site owned by Marburg University in Hessen, Germany. Second, we model the Leaf Area Index (LAI) for the same region. Spectral, terrain-related as well as geolocation variables are used as potential predictor variables in both examples. We study the effect of spatial and non-spatial cross-validation on the estimated model performance with the frequently applied machine learning algorithm Random Forest. A spatial variable selection is suggested to analyze the importance of the potential predictor variables for spatial mapping and their effect on the prediction outcomes.

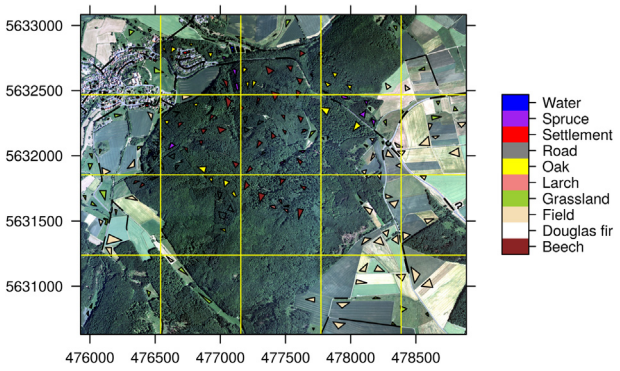

Fig. 1. Study area represented by the true color composite of the aerial image to be classified. Polygons indicate the training areas of the different LULC classes used for model training. The yellow grid represents spatial blocks used for spatial cross-validation. Reference system: UTM 32N (WGS84). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Methods

The following sections describe the two case studies, the data, all processing steps, modelling as well as validation. Data processing and modelling were performed in R Version 3.4 (R Core Team, 2018). The scripts performing processing and analysis can be retrieved from https://github.com/HannaMeyer/EcoMod_SpML.

2.1. Prediction task I: Land use/land cover classification

The first prediction task is to classify different types of forest, as well as adjacent LULC for the "Marburg Open Forest" http://nature40.org in Hessen, Germany. The basis for the classification is an aerial image that covers approx. 3000 × 2500 m (Fig. 1).

2.1.1. Reference data

A set of manually digitized polygons covering typical LULC classes are used as reference data which were selected by a combination of visual image inspection and knowledge firsthand from field work. In total, 10 different LULC classes were assigned (Table 1).

2.1.2. Predictor variables

Spectral, terrain-related and geolocation variables were all prepared as potential predictor variables (Fig. 2). Spectral variables come from a 20 cm resolution aerial image (Hessische Verwaltung für Bodenmanagement und Geoinformation, 2018a) taken at the 30th of September 2015. For this study, the image was resampled to a spatial resolution of 1 m. The spectral predictors were the three channels of the aerial image (red, green, blue). Further, the Visible Vegetation Index (VVI, Planetary Habitability Laboratory, 2015), Triangular Greenness

Table 1
Summary of the different land use/land cover classes and the size of training data used for training of the classification model.

Type	Polygons	Pixels
Beech	34	31,306
Douglas fir	20	13,241
Field	40	59,663
Grassland	85	27,134
Larch	4	1568
Oak	23	17,804
Road	38	18,461
Settlement	40	4722
Spruce	14	7521
Water	66	3261

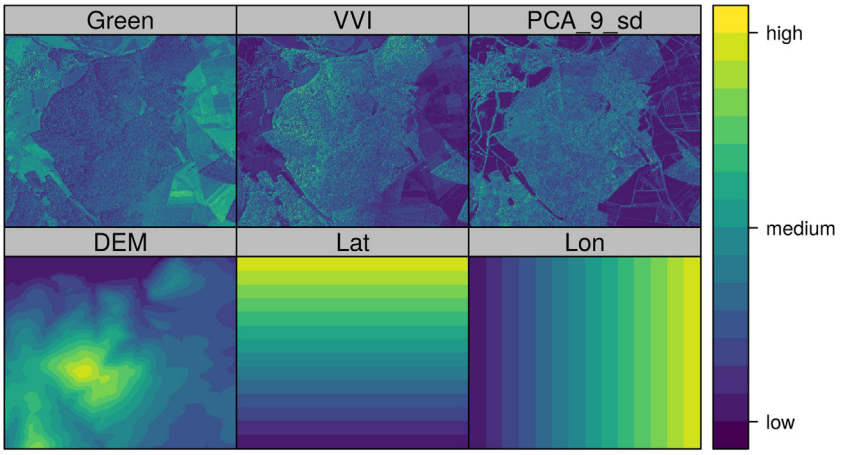

Fig. 2. Example of spectral, terrain-related and geolocation predictor variables: reflectance in the green band (green), Visible Vegetation Index (VVI), standard deviation in a 9 × 9 pixel environment of the first Principal Component of all spectral variables (PCA_9_sd), Digital Elevation Model (DEM), latitude (Lat) and longitude (Lon). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

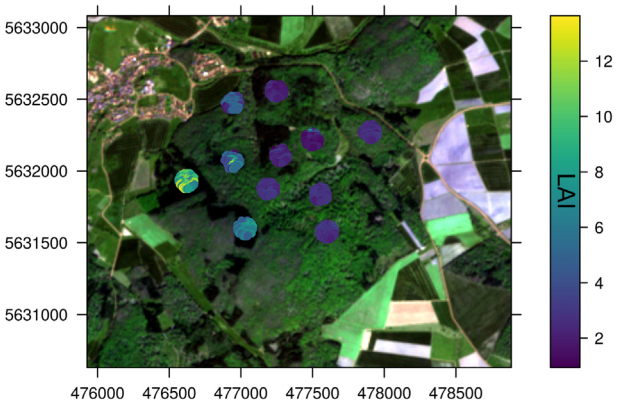

Fig. 3. Study area represented by the true color composite of the Sentinel-2 scene, which serves as the baseline for the LAI predictions. Points represent the location of the training samples. The color indicates the LAI values at these locations as derived from the lidar. The clear spatial clusters are the baseline for the spatial cross-validation. Reference system: UTM 32N (WGS84). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Index (TGI, Hunt et al., 2013), Normalized Green Red Difference Index (NGRDI), and Green Leaf Index (GLI, Hunt et al., 2013)) were derived from the channels. A Principal Component Analysis (PCA) was performed on the three channels of the visible spectrum and the vegetation indices; the first component of the PCA was included as an additional potential predictor variable. In addition, the standard deviation of the first principal component was calculated in 3x3 (PCA_3_sd), 5 × 5 (PCA_5_sd) and 9 × 9 (PCA_9_sd) pixel environments to account for the spectral variability of LULC classes. A lidar-derived 1 m Digital Elevation Model (DEM) was used as a terrain-related predictor. Elevation in the study area ranges from 210 to 415 m. Slope and aspect were calculated from the DEM in radians. The geolocation variables considered as potential predictors were latitude (Lat) and longitude (Lon). This results in a total set of 16 potential predictor variables for LULC prediction.

2.1.3. Compilation of the training data set

The set of predictor variables was extracted for each training polygon. The model considered each pixel related to the polygons (e.g. within, intersecting) as an individual training sample (Table 1). This resulted in a set of approx. 185,000 training samples. Each training sample contained the information about every potential predictor variable as well as about the LULC class based on the information from the polygons.

2.2. Prediction task II: Leaf Area Index modelling

The second prediction task aimed at modelling the LAI for the forested area of the Marburg Open Forest, a classic example of a regression task in environmental science.

2.2.1. Reference data

In this case study, the LAI reference was derived from lidar data taken in the vegetation period 2010 (Hessische Verwaltung für Bodenmanagement und Geoinformation, 2018b) that have 15 cm vertical and 30 cm spatial accuracy. The LAI was calculated from the lidar point cloud according to Getzin et al. (2017). Since no major management was present in the forest, the LAI from the lidar data was regarded as a reference for this study despite the time lag between lidar derived reference and the Sentinel-2 based predictor variables. Especially since this study focus on the effect of validation and variable selection strategies this time lag was neglected. The calculated LAI data was then rasterized with 10 m spatial resolution to match the geometry of the Sentinel-2 data that were used as predictors. To do this, the mean of all LAI values located in the extent of a Sentinel-2 pixel was calculated. 11 spatially distinct clusters were then assigned in homogenous areas of the forest. Every pixel in a 60 m radius around the center of each cluster was used as training data, resulting in clear spatial clusters of training samples (Fig. 3). In total, 824 training pixels distributed across the 11 clusters were used. The minimum, maximum and mean LAI in this training data set were 0.9, 13.6 and 4.2, respectively. A LAI below 1 means that the area is not fully covered by leaves. Values larger than 1 mean that more than one layer of leaves are present.

2.2.2. Predictor variables

A Sentinel-2 scene as Level-1C product from 2017/05/10 was used to derive spectral predictor variables. Sentinel-2 is the optical system from the earth observation mission from the EU Copernicus Programme and has channels in the visible (bands 2–4), red edge (bands 5–7), near infrared (band 8 and 8A) and short-wave infrared (bands 11 and 12) part of the spectrum. The Sentinel-2 bands 1, 9 and 10 were not considered in this study because they do not include relevant information for this prediction task. Hereafter, the used channels are referred to as B01, B02, …, B12. Channels B02-B04 and B08 have a spatial resolution of 10 m. The other channels have a resolution of 20 m and were resampled to match the geometry of the 10 m channels. In addition to the spectral channels, elevation, slope, aspect, as well as latitude and longitude (as described in the description of the previous prediction task but resampled to a 10 m spatial resolution) were used as potential predictors. This results in a total set of 15 potential predictor variables for LAI prediction.

2.2.3. Compilation of the training data set

Values for each predictor variable was extracted from the locations of the training samples. Each training sample contained the extracted information from all potential predictor variables as well as the information about the LAI based on the information from the lidar-derived reference points.

2.3. Model training and prediction

The Random Forest algorithm (Breiman, 2001) was chosen as the machine learning algorithm for predictive modelling due to its prominence in ecological modelling. Random Forest bases on the concept of regression and classification trees: a series of nested decision rules for the predictors determine the response (also called reference, i.e. LULC or LAI). Random forest repeatedly builds trees from random samples of the training data. Each tree is a separate model of the ensemble. The predictions of all trees are averaged to produce the final estimate. To overcome correlation between trees, a number of predictors (mtry) are randomly selected at each split. The best predictor from the random subset is used at this split to partition the data.

In this study, the Random Forests implementation of the randomForest package (Liaw and Wiener, 2002) in R was applied and accessed via the caret package (Kuhn, 2016). Throughout the study, each Random Forest model consisted of 500 trees after no increase in performance could be observed using a higher number of trees. The number of randomly chosen predictor variables at each split of the tree ("mtry") was tuned between two and the number of predictor variables (16 for LULC predictions and 15 for LAI predictions). See Kuhn and Johnson (2013) for a more detailed description on the Random Forest algorithm and mtry tuning.

To study the effect of spatial validation as well as spatial variable selection the following models were compared for both case studies:

1. Model using **all** potential predictor variables. Performance was estimated by **random** cross-validation (see 1a in Fig. 4). The results of this model are used to show the outcome of a "default" modelling approach. The performance was further estimated by **spatial** cross-validation (see 1b in Fig. 4). The results of this validation are used to show how spatial cross-validation affects the estimated error of a "default" model.
2. Model using **selected** variables only. Variable selection was based on the commonly used recursive feature elimination with **spatial** cross-validation. Performance was estimated by **random** cross-validation (see 2a in Fig. 4) and **spatial** cross-validation (see 2b in Fig. 4). The results of this model are used to show how "default" variable selection affects the spatial model performance.
3. Model using **selected** variables only. Variable selection was based on a forward feature selection with **random** cross-validation. Performance was estimated by **random** cross-validation (see 3a in Fig. 4) and **spatial** cross-validation (see 3b in Fig. 4). The results of this model are used to show how random variable selection affects the spatial model performance.
4. Model using **selected** variables only. Variable selection was based on a forward feature selection with **spatial** cross-validation. Performance was estimated by **random** cross-validation (see 4a in Fig. 4) and **spatial** cross-validation (see 4b in Fig. 4). The results of this model are used to show how spatial variable selection affects the spatial model performance.

The following sections describe the different cross-validation and variable selection strategies in more detail.

2.3.1. Cross-validation strategies

This study applied two cross-validation strategies: a standard random k-fold cross-validation and a spatial k-fold cross-validation. Each strategy first splits the data into k folds and then repeatedly trains the models (k times) using the data from all but one fold. The models are evaluated based on how they perform with the left-out data (see Fig. 5). See (Kuhn and Johnson, 2013) for more detailed description on cross-validation in general.

While the cross-validation procedure is the same for random and spatial cross-validation, the major difference is how the data points are split into folds (see Fig. 5). For the standard random cross-validation, each data point was **randomly** assigned to one of the k folds.

For the spatial cross-validation, we chose a spatial block approach as suggested in Roberts et al. (2017) and also Valavi et al. (2018) for the LULC case study. Therefore, we divided the spatial domain into 20 equally sized spatial blocks (yellow grid in Fig. 1). For each training sample, the spatial block affiliation was identified by the spatial location of the corresponding training polygon where the sample belongs to. If a training polygon lay within two spatial blocks, the sample was only assigned to the one block in which the greater proportion of the polygon lay. This precluded that training pixels from one (usually homogeneous) polygon were present in two spatial blocks. Analogous

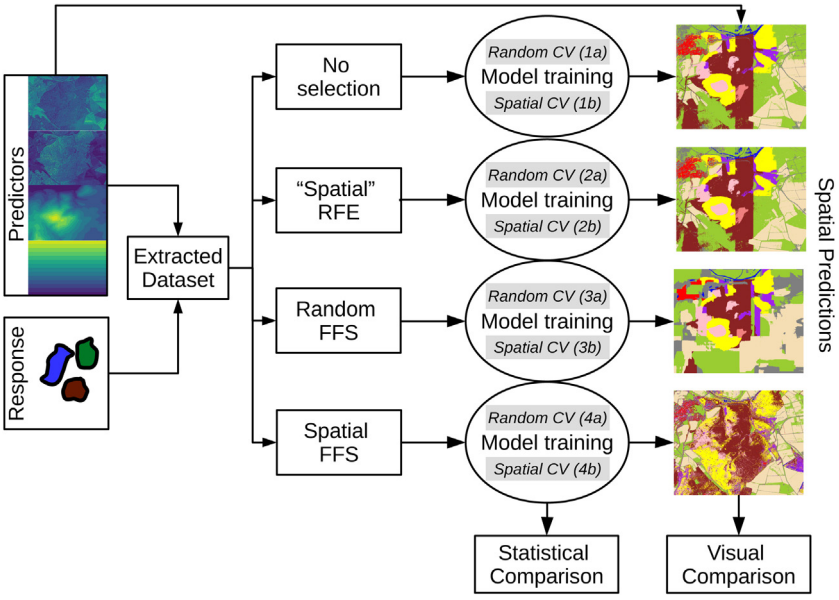

Fig. 4. Overview on the models compared in this study. Spectral, terrain related as well as geolocation variables were used as predictors (only examples shown here). The response variable was derived from training polygons for the case study of land cover classification or from lidar-derived Leaf Area Index (LAI) values for the case study of LAI prediction. Models were trained using either no variable selection or either recursive feature elimination, default random forward feature selection (FFS) or spatial FFS. Model performance was compared with random cross-validation (CV) or spatial CV. The entire modelling procedures were performed for the case study of land cover classification as well as for prediction of LAI.

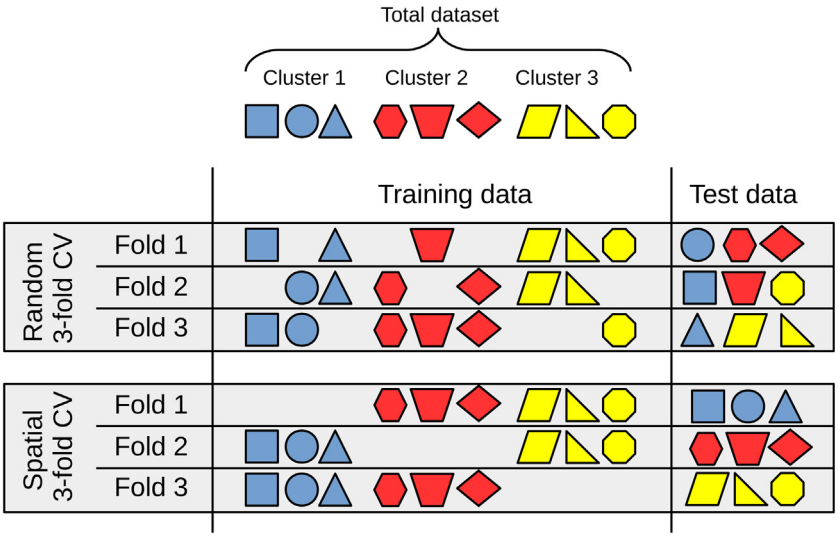

Fig. 5. Concept of random and spatial cross-validation (CV): A total dataset (here: 9 different data points represented by different shapes) is split into k folds (here: k = 3). Models are then repeatedly trained by always leaving one of the folds out and use it for model validation and not for model training. Random CV means that the data are randomly split into folds. Spatial CV means that the data are split into folds according to spatial location (e.g. a spatial cluster or a spatial block, here represented by unique color). Figure modified from Meyer et al. (2018) and Kuhn and Johnson (2013). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to the random cross-validation, models were then repeatedly trained using data from all but one spatial block (= fold) and their performance estimated using the left-out data, e.g. the spatial block left out of model training. Hence, models were assessed for their performance in making spatial predictions beyond the locations of the training data. For the case study of the LAI predictions, a leave-one-cluster-out cross-validation was applied. This method is similar to the concept described above but instead of spatial blocks, in each iteration one of the 11 clusters Fig. 3) was left out during model training.

The number of spatial blocks or clusters, k, equaled the number of spatial blocks (20) or the number of spatial clusters (11), depending on the case study. Random cross-validation was performed with the same number for k.

The validation measure for performance assessment for the LULC classification during cross-validation was the Kappa Index (Cohen, 1960) and the Accuracy. A Kappa of 0 (or lower) is associated with a random classification result, while a Kappa of 1 indicates a perfect classification. Since it accounts for chance agreement, it was used as the prior validation measure for the classification task rather than the Accuracy. For the LAI predictions, the performance was assessed by the Root-Mean-Square Error (RMSE) and the coefficient of determination (R^2).

Models were compared by fold, which gives the average performance (e.g. Kappa) over all k folds from cross-validation. Models were also compared by their global performance, which results from a comparison between every data point predicted during cross-validation, independently of the fold.

2.3.2. Spatial predictor variable selection

Forward Feature Selection (FFS) as described in Meyer et al. (2018) and implemented in the CAST package (Meyer, 2018) for R was used for spatial variable selection. This FFS implementation works in conjunction with user-defined cross-validation, hence it allows to select variables that lead to the highest spatial performance (if run in conjunction with spatial CV). First, the FFS trains models using every combination of two predictor variables. Based on the best performing variables (as identified by cross-validation), FFS increases the number of variables to test which (if any) variables further improve the (spatial) model's performance. Every variable that does not improve (or even decreases) the model's performance is excluded. See Meyer et al. (2018) for a more detailed description on this FFS.

This study used FFS with spatial cross-validation to test which variables are significant to spatial mapping and which ones have no spatial meaning or are even counterproductive (spatial FFS, Fig. 4). For comparison, FFS was also run with random cross-validation (random FFS, Fig. 4) to check that improvements were indeed due to the spatial selection and not the pure reduction of variables that lead to changes in performance. As FFS is very time consuming, mtry was not tuned but set to 2 for feature selection. Once the variables were selected, models were re-trained using the selected variables and either spatial or random cross-validation with mtry being tuned between 2 and the number of selected predictor variables.

In addition to FFS, we also used recursive feature elimination (RFE, explained in Kuhn and Johnson, 2013) to compare state-of-the art procedures (see e.g. Brungard et al., 2015; Meyer et al., 2017a,b; Ghosh and Joshi, 2014; Stevens et al., 2013, in the field of environmental mapping). However, we argue that the backward RFE selection fails to address the issue of overfitting. RFE relies on variable importance scores, which are only calculated using the training subset (Kuhn and Johnson, 2013). If a variable leads to considerable overfitting, it is highly significant in the models. Therefore, the RFE process will keep it as important and not remove it, even if it results in a high spatial error.

3. Results

3.1. Statistical performance

Using the “default” way of spatial prediction (using all potential variables) and the “default” random cross-validation, both Accuracy and Kappa index were higher than 0.99 for the classification task (Table 2). Random cross-validation indicated that the LAI could be predicted with a RMSE of 0.96 and a R^2 of 0.87 (Table 3, Fig. 6a).

When these models were validated using a spatial cross-validation, the performance was considerably lower (Kappa value of 0.55 for the LULC classification and RMSE of 1.25 for the LAI regression model, Table 3, Fig. 6b). A prominent source of high error estimates is that by leaving entire clusters out for validation, a held back cluster that has higher LAI values than all other clusters could not adequately be modelled since such high LAI values are unknown from the training data (Fig. 6b). Note that low per-fold R^2 values for the LAI regression models in Table 3 result from low variabilities within spatial folds (Fig. 6b) so that the per-fold RMSE or especially the global R^2/RMSE present the more reliable performance estimates here.

Using an RFE-based variable selection in conjunction with a spatial cross-validation during the variable selection does not (LULC classification Kappa = 0.55) or only marginally (LAI regression RMSE = 1.22) improve the spatial performances. The same is true for an FFS-based approach in conjunction with a random cross-validation during the variable selection (Kappa = 0.14, RMSE = 1.23). In both cases, the random model performance stayed high (Kappa > 0.99 for both RFE

Table 2
Statistical performance of the models for LULC classification. Models were compared by fold, which gives the average performance over all k folds from cross-validation (CV). Models were also compared by their global performance, which is the Accuracy or Kappa for every data point predicted during cross-validation. Bold numbers indicate a spatial validation that must be considered as the valid performance for the prediction task. For an overview on the model-ID see also Fig. 4.

ID	Variables	CV	By fold		Global	
			Accuracy	Kappa	Accuracy	Kappa
1a	All	Random	> 0.99	> 0.99	> 0.99	> 0.99
1b	All	Spatial	**0.71**	**0.55**	**0.68**	**0.61**
2a	Selected by RFE "spatial"	Random	> 0.99	> 0.99	> 0.99	> 0.99
2b	Selected by RFE "spatial"	Spatial	**0.71**	**0.55**	**0.69**	**0.61**
3a	Selected by FFS random	Random	> 0.99	> 0.99	> 0.99	> 0.99
3b	Selected by FFS random	Spatial	**0.43**	**0.14**	**0.41**	**0.30**
4a	Selected by FFS spatial	Random	0.89	0.87	0.78	0.82
4b	Selected by FFS spatial	Spatial	**0.71**	**0.56**	**0.70**	**0.62**

Table 3
Statistical performance of the models for LAI prediction. Models were compared by fold, which gives the average performance over all k folds from cross-validation (CV). Models were also compared by their global performance which is the RMSE or R^2 over every data point predicted during cross-validation. Bold numbers indicate a spatial validation that must be considered as the valid performance for the prediction task. For an overview on the model-ID see also Fig. 4.

ID	Variables	CV	By fold		Global	
			RMSE	R^2	RMSE	R^2
1a	All	Random	0.96	0.87	0.97	0.86
1b	All	Spatial	**1.25**	**0.07**	**1.75**	**0.58**
2a	Selected by RFE "spatial"	Random	0.93	0.88	0.95	0.87
2b	Selected by RFE "spatial"	Spatial	**1.22**	**0.06**	**1.73**	**0.58**
3a	Selected by FFS random	Random	0.91	0.88	0.92	0.88
3b	Selected by FFS random	Spatial	**1.23**	**0.04**	**1.80**	**0.58**
4a	Selected by FFS spatial	Random	1.12	0.83	1.14	0.81
4b	Selected by FFS spatial	Spatial	**1.20**	**0.06**	**1.64**	**0.63**

and random FFS, RMSE = 0.93 for RFE and 0.91 for FFS random). Hence, neither a FFS with random selection nor the RFE approach does prevent spatial overfitting even though spatial folds have been used for the latter.

When FFS was paired with spatial cross-validation (spatial FFS) for the variable selection task, the spatial performance slightly improved (Kappa = 0.56, RMSE = 1.20) compared to all other models. It is noteworthy that this type of variable selection reduces the model performance indicators in a random cross-validation (Kappa = 0.87, RMSE = 1.12) so that the differences in the error estimates between the validation strategies became smaller. This validation was based on the average performance for each fold that was left out during cross-validation. Similar patterns emerged when all independent predictions were simultaneously compared (global validation). Noticeable is that the R^2 of LAI predictions increased from 0.58 (all variables, spatial cross-validation) to 0.63 (spatial variable selection, spatial cross-validation).

3.2. Variable importance and selected variables

When all variables were presented to the algorithm to predict LULC and LAI, the most important variables were latitude, longitude and elevation (Fig. 7). The spectral predictor variables were considerably less important for these tasks.

The RFE based upon this variable importance ranking did not eliminate any variables for the LULC classification, hence the model is essentially identical to the initial full model. For the LAI prediction model, the RFE only dropped the Sentinel-2 band "B02" (blue band). The combination of FFS and random cross-validation selected latitude, longitude, DEM and aspect for LULC classification and DEM, longitude, latitude, B12, B07, aspect and slope for LAI predictions, in decreasing order of importance.

FFS with spatial cross-validation identified the geographic coordinates and elevation as irrelevant or even counterproductive and dropped them from the models. The final model used only a subset of the spectral variables and the slope for the LULC classification. Here, green, blue, red and the standard deviation of the pca in a 9 × 9 environment made the largest contributions (Fig. 8a). For LAI predictions, only the bands B05, B07 (both red edge), B03 (green) and B8A (narrow NIR) were identified as important. However, B03 and B8A only slightly decreased the RMSE compared to the model that used B05 and B07 only (Fig. 8b).

3.3. Visual assessment of the spatial prediction

The model that used all variables to predict LULC led to noticeably linear features when making spatial predictions for the full study area (Fig. 9 no selection; the RFE-based model produces a quasi identical

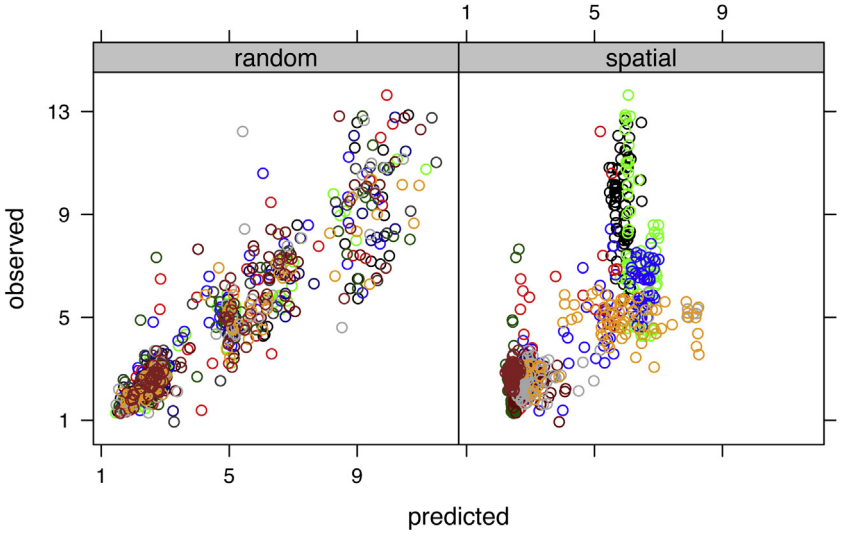

Fig. 6. Comparison between observed and predicted LAI values based on random (a) and spatial (b) cross-validation. Colors indicate the individual 11 folds. Note that the resulting models are quasi identical regardless of the validation strategy being used since cross-validation in this case serves mtry tuning and validation purposes only. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

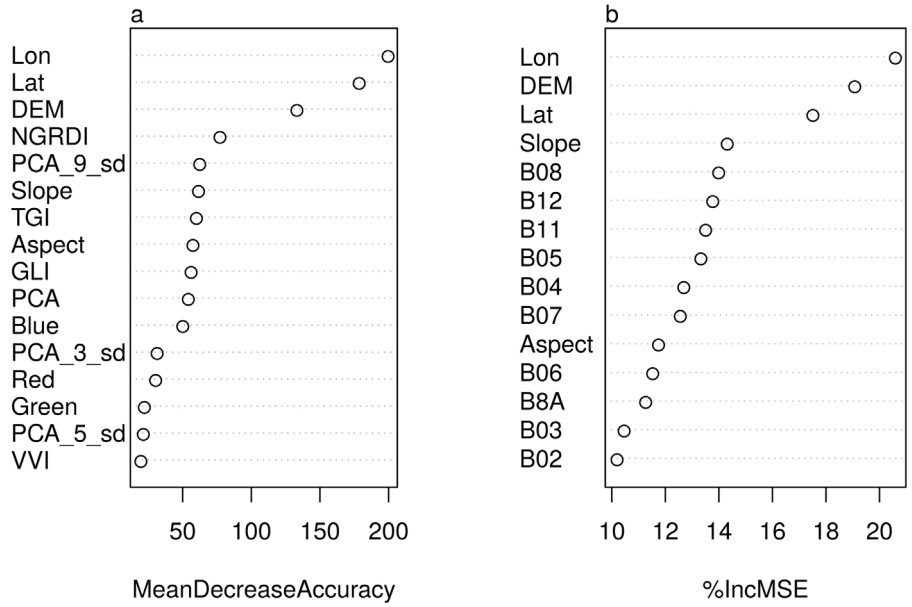

Fig. 7. Relative scaled importance of the predictor variables within the Random Forest models using all variables or using an RFE approach for the case study of predicting (a) LULC and (b) LAI. See Sections 2.1.2 and 2.2.2 for further explanations of the variables.

spatial prediction). A clear linear delineation was made between beech forest and grassland that does not correspond to the visual inspection of the underlying aerial image (Fig. 9 RGB). An obvious patch of forest in the southeastern part of the image was classified as grassland. Field and road were clearly confused for one another in the northwestern corner. A round patch of Douglas fir in the southwestern quarter of the image can be clearly associated to the highest elevation of the forest. Elevation appeared to be an important factor in the prediction of water, since parts of fields in the north of the image corresponding to the lowest elevations (Fig. 2) were falsely classified as such. These areas were visually distinguishable from water in the RGB. Several other patterns that did not correspond to a visual interpretation of the RGB were also present.

Random variable selection (FFS with random selection) enhanced the problem and linear features became more obvious (Fig. 9 random selection). The prediction has an overall smooth appearance as FFS removed spectral variables from the model.

When variables were selected by spatial FFS, the coordinates and elevation were removed by the algorithm and the classification was based on the spectral variables (Fig. 9 spatial selection). The result showed much greater local variability that was clearly driven by the underlying spectral information rather than gradual changes driven by geolocation. No linear artefacts were observed.

Similar though less striking patterns were found for the LAI predictions. Using all potential predictor variables led to a visible linear feature dividing generally lower LAI values to the east from generally higher values to the west (Fig. 10 no selection). Such features became more obvious when FFS with random cross-validation was applied (Fig. 10 random selection). When FFS selected variables with spatial cross-validation, no geolocation variables were selected and the results showed no obvious artefacts in the spatial prediction (Fig. 10 spatial selection).

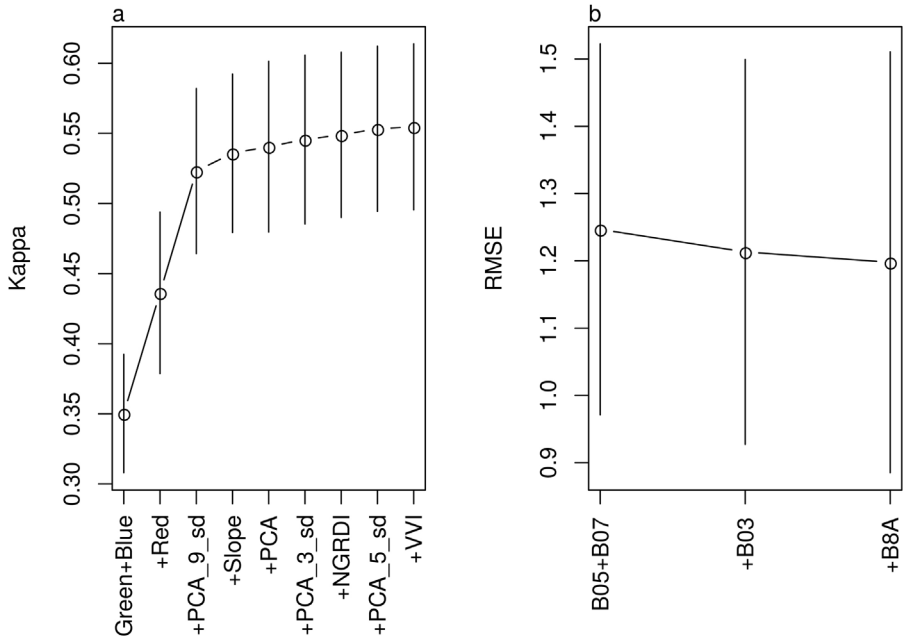

Fig. 8. Performance of FFS with variables selected by spatial cross-validation for prediction of (a) LULC and (b) LAI. The first point indicates the model's performance using the two variables that lead to the best spatial performance. Subsequent points indicate the model's performance with the addition of the next best variable, i.e. the third point represents the top four variables. Bars represent the standard deviation over the k spatial folds. See Sections 2.1.2 and 2.2.2 for further explanations of the variables.

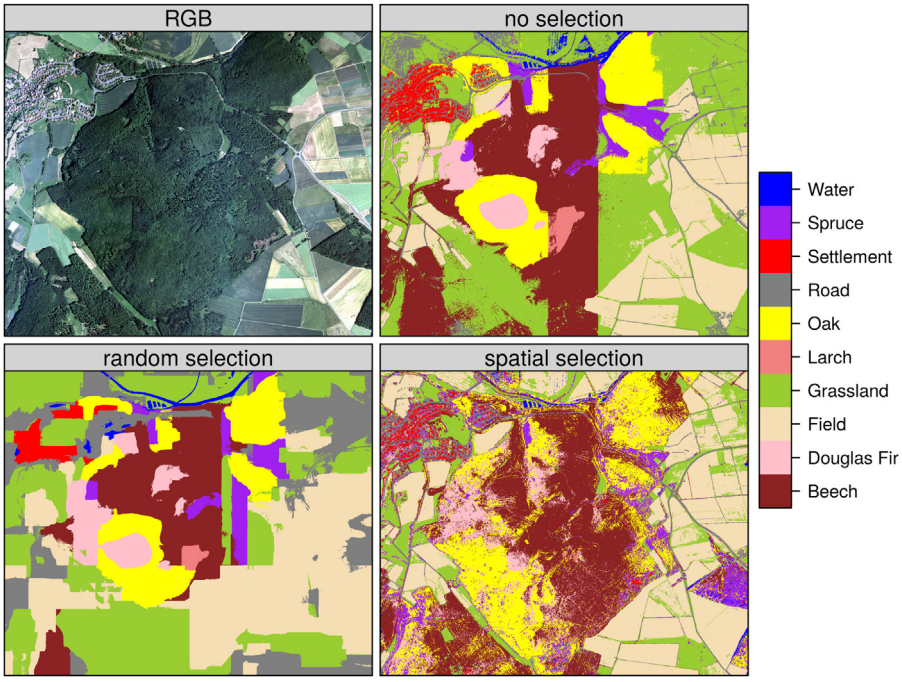

Fig. 9. RGB representation of the study area on the basis of the aerial image (RGB), spatial LULC predictions by the model that used all potential predictor variables (no selection), the model with variables being selected by random FFS (random selection), as well as spatial FFS (spatial selection).

4. Discussion

4.1. Importance of spatial validation

The results clearly highlight again the necessity of spatial validation for realistically assessing the performance of spatial prediction models. Standard validation procedures that use random subsets of the dataset (i.e. random k-fold cross-validation) produce overoptimistic estimates about the model performance (in this case "nearly perfect classification" of LULC and low errors for LAI predictions). These do not provide information about the actual model performance with respect to the prediction of any other place than the sampling locations (i.e. create a

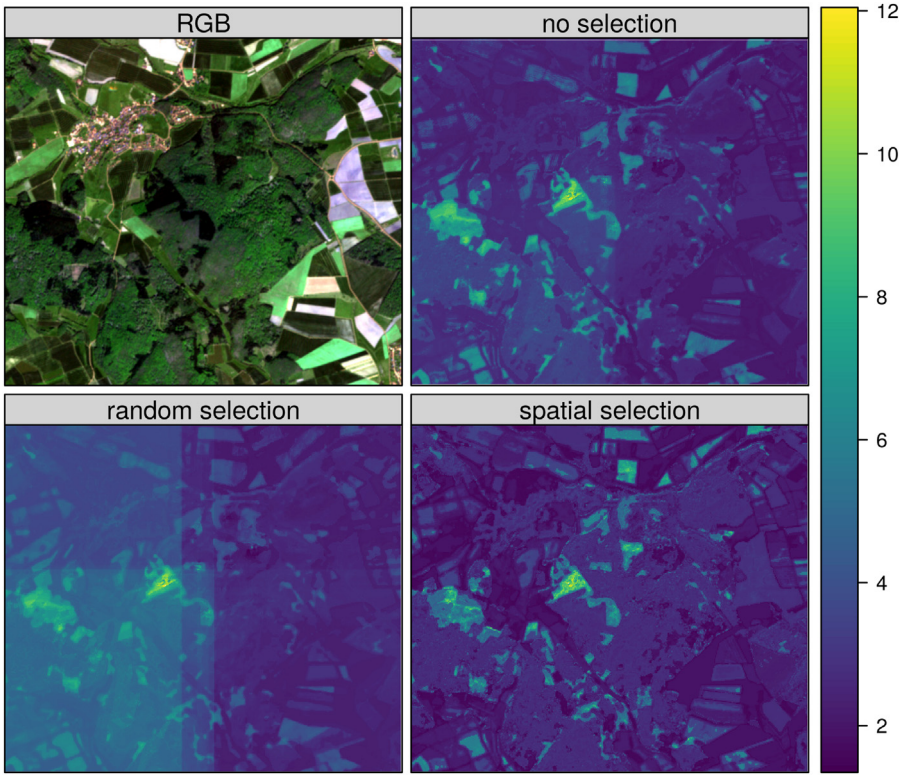

Fig. 10. RGB representation of the study area on the basis of the Sentinel-2 image (RGB), spatial LAI predictions by the model that used all potential predictor variables (no selection), the model with variables being selected by random FFS (random selection), as well as spatial FFS (spatial selection).

map of LULC or LAI). Predicting any and not just the training locations, however, is the aim of most spatial prediction models, and the performance estimates must account for that which is increasingly and consistently recommended in recent literature (e.g. Wenger and Olden, 2012; Juel et al., 2015; Roberts et al., 2017; Valavi et al., 2018; Pohjankukka et al., 2017; Cánovas-García et al., 2017). Random validation is not a meaningful strategy for spatial prediction tasks. This is especially essential when flexible algorithms are applied to highly clustered training samples that cause the risk of overfitting caused by the spatial autocorrelation of the predictor and response variables.

4.2. Relevance of spatial variable selection

It is not surprising that the model that uses all variable showed a strong performance in random cross-validation as compared to spatial cross-validation considering that within the polygons of digitized LULC or lidar-derived LAI training sites, the samples feature similar properties in their predictor variable space. Hence large parts of the training samples are not independent from one another. This leads to overfitting and incorrectly assigning high importance to the variables that represent the spatial location, which is especially clear for geolocation variables (i.e. latitude, longitude). Therefore, many studies have unsurprisingly identified coordinates as one of the, if not the, most important predictor, such as for tree species distribution (Attorre et al., 2011), monthly precipitation (Jing et al., 2016), deforestation (Zanella et al., 2017), phytoplankton abundance (Roubeix et al., 2016) and explaining the spatial variability of soil organic carbon (Yang et al., 2016). According to our results, spatial variable selection would have very likely removed geolocation variables from these studies' models. In addition to geolocation variables such as coordinates or Euclidean distance fields (Behrens et al., 2018), variables that are unique to a certain spatial cluster would be problematic. For example, Meyer et al. (2016) show that elevation can complicate models when it is clearly indicative of one spatial cluster – in this study, circular patterns of Douglas fir were predicted for the areas of the forest with the greatest elevation, which is a clear artefact caused by misinterpretation of elevation as a predictor variable. Spatial FFS removed elevation as it was identified as unimportant or counterproductive as evinced by improved visual and statistical results.

We show that including location variables does not solve the problem of autocorrelation but intensifies the problem, at least for spatially clustered data. This finding contrasts with recommendations from previous studies (Mascaro et al., 2013; Cracknell and Reading, 2014). Though Random Forests are known for being robust to uninformative predictor variables, this study clearly shows that misleading variables can have negative effects on the models. This notion is also supported by Rocha et al. (2018), who showed that spatial predictions of plant traits suffer when models include spatial relations. The phenomenon, however, can only be detected if spatial cross-validation is used. Spatial models evaluated in a default random way will still feign a high spatial performance.

Spatial cross-validation provides reliable performance measures for spatial models. However, it does not change the model itself. During the internal Random Forest training, variables are not selected by spatial cross-validation but by the out-of-bag (oob) error which is based on random bootstrapping. Hence in a Random Forest training, variables are not selected by their spatial importance. A spatial variable selection is therefore required to remove misleading variables from the models. Using geolocation variables, the algorithm could reproduce training samples with highest performance but only a selection of spatially meaningful variables allowed for predictions beyond the locations of the training samples.

Having a look at the internal variable importance ranking of the Random Forests algorithm (Fig. 7) also explains why recursive feature selection cannot help, even if spatial cross-validation determines the optimal variables: Since RFE is based on the importance ranking of the variables, those that are misleading and responsible for overfitting are often highly ranked (e.g. latitude, longitude in this study) and hence not removed by RFE.

This study automatically analyzes which variables are misleading, counterproductive and cause overfitting and hence must be removed from the models. We show that removing these variables from the models improves the statistical spatial performance; a visual inspection also confirmed reliable patterns compared to the common model that uses all variables. The increase in statistical performance was less obvious than in Meyer et al. (2018), who used spatio-temporal data that can be explained by a stronger autocorrelation due to the application to long time series. Therefore, improvements in performance will likely track with increasing degrees of autocorrelation if the sampling design includes clear spatial clusters.

The results of this study strongly suggest that spatial cross-validation needs to be considered not only for model validation and model tuning (see Schratz et al., 2019, for a study on the relevance of spatial validation for hyperparameter tuning in machine learning applications) but also for variable selection, hence during all steps of model building.

4.3. Need for visual assessment in addition to statistical validation

The results also show that statistical validation alone is insufficient to validate spatial prediction models. Both the model using all potential predictors and the one using spatially selected variables perform statistically similar in spatial cross-validation. Hence, we conclude that the models perform equally well, statistically speaking. A visual assessment reveals that this assumption does not hold true, however. Removing misleading variables dramatically changes the actual outcome as patterns in the LULC and LAI predictions are considerably different. Again, this highlights the need for spatial variable selection. In several other studies, artefacts are mainly visible as clear linear features and can most certainly be traced back to the geolocation variables. Jing et al. (2016) and Shi et al. (2015) used coordinates for downscaling precipitation, which the Random Forest algorithm identified as the most important variable, but which resulted in visible linear patterns in the spatial prediction. Mud content prediction by Li et al. (2011) also shows linear patterns that are most likely caused by the inclusion of latitude and longitude as predictors. In a study by Fox et al. (2017) the Random Forest algorithm also ranked latitude and longitude as important, yet the resulting marine bird distribution along the Canadian coast shows clear linear cuts. These examples visually highlight the issues that including geolocation variables can cause. They also underline the importance of spatial cross-validation for spatial error assessment in conjunction with spatial variable selection to ensure that only variables with actual predictive power beyond the training locations are included.

5. Conclusions

This study underlines the necessity of spatial validation strategies in spatial machine learning applications. Results will likely be overoptimistic if these strategies are ignored. This is especially the problem when there is strong spatial autocorrelation in the data and when training samples are clustered in space.

However, spatial machine learning applications should not be restricted to the usage of spatial validation. This study shows that certain variables are responsible for overfitting that causes strong random performance but a failure in predicting any other than the training samples location. This is especially evident for geolocation variables (e.g. latitude, longitude). When such variables are used in spatial modelling where training samples are highly clustered in space, they lead the algorithms to effectively reproduce the training data but lead the model to fail predicting on new samples. Hence, the applied Random Forest algorithm cannot interpret such variables in a meaningful way. Spatial variable selection is required to automatically select

variables that are useful in a Random Forest setup, for example the suggested forward feature selection that selects variables according to their contribution for spatial predictions. Spatial validation should hence be considered during all steps of modelling, from hyperparameter tuning, variable selection to performance estimation.

Like most other machine learning algorithms, Random Forests have the reputation that no assumptions about the data distribution are necessary. However, the results of this study show that it might be necessary to revisit this idea and general guidelines should be formulated to make applications more objective.

Finally, the results of this study allow the conclusion that ignoring spatial dependencies in machine learning applications for spatial predictions carries a high risk of developing models that can reproduce training data well but do not make reliable spatial predictions. Reliable spatial predictions can only be achieved if spatial dependencies are taken into account during the modelling process, i.e. not only for the purpose of model validation, but also for the selection of appropriate predictor variables.

Acknowledgments

This work was conducted within the Natur 4.0 | Sensing Biodiversity project funded by the Hessian state offensive for the development of scientific-economic excellence (LOEWE).

References

Appelhans, T., Mwangomo, E., Hardy, D.R., Hemp, A., Nauss, T., 2015. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. Spat. Stat. 14 (Part A), 91–113. https://doi.org/10.1016/j.spasta.2015.05.008.

Attorre, F., Alfò, M., Sanctis, M.D., Francesconi, F., Valenti, R., Vitale, M., Bruno, F., 2011. Evaluating the effects of climate change on tree species abundance and distribution in the Italian peninsula. Appl. Veg. Sci. 14, 242–255. http://www.jstor.org/stable/41058163.

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

Behrens, T., Schmidt, K., Viscarra Rossel, R.A., Gries, P., Scholten, T., MacMillan, R.A., 2018. Spatial modelling with euclidean distance fields and machine learning. Eur. J. Soil Sci. 69, 757–770. https://doi.org/10.1111/ejss.12687.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

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

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., T.C.E Jr., , 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239–240, 68–83. https://doi.org/10.1016/j.geoderma.2014.09.019.

Cánovas-García, F., Alonso-Sarría, F., Gomariz-Castillo, F., nate Valdivieso, F.O., 2017. Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery. Comput. Geosci. 103, 1–11. https://doi.org/10.1016/j.cageo.2017.02.012. http://www.sciencedirect.com/science/article/pii/S0098300416303909.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46. https://doi.org/10.1177/001316446002000104.

Cracknell, M.J., Reading, A.M., 2014. Geological mapping using remote sensing data: a comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. Comput. Geosci. 63, 22–33. https://doi.org/10.1016/j.cageo.2013.10.008.

Fox, C.H., Huettmann, F.H., Harvey, G.K.A., Morgan, K.H., Robinson, J., Williams, R., Paquet, P.C., 2017. Predictions from machine learning ensembles: marine bird distribution and density on Canada's pacific coast. Mar. Ecol. Prog. Ser. 566, 199–216. https://www.int-res.com/abstracts/meps/v566/p199-216/.

Gasch, C.K., Hengl, T., Gräler, B., Meyer, H., Magney, T.S., Brown, D.J., 2015. Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D + T: the cook agronomy farm data set. Spat. Stat. 14 (Part A), 70–90.

Georganos, S., Grippa, T., Gadiaga, A.N., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., Kalogirou, S., 2019. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int. 1–16. https://doi.org/10.1080/10106049.2019.1595177.

Getzin, S., Fischer, R., Knapp, N., Huth, A., 2017. Using airborne lidar to assess spatial heterogeneity in forest structure on mount Kilimanjaro. Landsc. Ecol. 32, 1881–1894. https://doi.org/10.1007/s10980-017-0550-7.

Ghosh, A., Joshi, P., 2014. A comparison of selected classification algorithms for mapping bamboo patches in lower gangetic plains using very high resolution WorldView 2 imagery. Int. J. Appl. Earth Obs. Geoinf. 26, 298–311. https://doi.org/10.1016/j.jag.2013.08.011.

Gudmundsson, L., Seneviratne, S.I., 2015. Towards observation-based gridded runoff estimates for Europe. Hydrol. Earth Syst. Sci. 19, 2859–2879.

Hengl, T., Nussbaum, M., Wright, M., Heuvelink, G., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.26693v3.

Hessische Verwaltung für Bodenmanagement und Geoinformation, 2018a. Aerial Imagery.

Hessische Verwaltung für Bodenmanagement und Geoinformation, 2018b. Lidar data.

Hunt, E.R., Doraiswamy, P.C., McMurtrey, J.E., Daughtry, C.S., Perry, E.M., Akhmedov, B., 2013. A visible band index for remote sensing leaf chlorophyll content at the canopy scale. Int. J. Appl. Earth Obs. Geoinf. 21, 103–112. https://doi.org/10.1016/j.jag.2012.07.020. http://www.sciencedirect.com/science/article/pii/S0303243412001791.

Janatian, N., Sadeghi, M., Sanaeinejad, S.H., Bakhshian, E., Farid, A., Hasheminia, S.M., Ghazanfari, S., 2017. A statistical framework for estimating air temperature using MODIS land surface temperature data. Int. J. Climatol. 37, 1181–1194. https://doi.org/10.1002/joc.4766.

Jing, W., Yang, Y., Yue, X., Zhao, X., 2016. A comparison of different regression algorithms for downscaling monthly satellite-based precipitation over North China. Remote Sens. 8, 835. https://doi.org/10.3390/rs8100835.

Juel, A., Groom, G.B., Svenning, J.-C., Ejrnæs, R., 2015. Spatial application of Random Forest models for fine-scale coastal vegetation classification using object based analysis of aerial orthophoto and DEM data. Int. J. Appl. Earth Obs. Geoinf. 42, 106–114. https://doi.org/10.1016/j.jag.2015.05.008. http://www.sciencedirect.com/science/article/pii/S0303243415001178.

Kuhn, M., 2016. caret: Classification and Regression Training. R package version 6.0-68. https://CRAN.R-project.org/package=caret.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling, 1st ed. Springer, New York.

Langella, G., Basile, A., Bonfante, A., Terribile, F., 2010. High-resolution space-time rainfall analysis using integrated ANN inference systems. J. Hydrol. 387, 328–342.

Lary, D.J., Alavi, A.H., Gandomi, A.H., Walker, A.L., 2016. Machine learning in geosciences and remote sensing. Geosci. Front. 7, 3–10. https://doi.org/10.1016/j.gsf.2015.07.003.

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. https://doi.org/10.1111/geb.12161.

Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Model. Softw. 26, 1647–1659.

Liaw, A., Wiener, M., 2002. Classification and regression by Random Forest. R News 2, 18–22.

Mascaro, J., Asner, G.P., Knapp, D.E., Kennedy-Bowdoin, T., Martin, R.E., Anderson, C., Higgins, M., Chadwick, K.D., 2013. A tale of two "Forests": random forest machine learning aids tropical forest carbon mapping. PLOS ONE 9, e85993. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3904849/.

Meyer, H., 2018. CAST: 'caret' Applications for Spatial-Temporal Models. R Package Version 0.2.1. https://CRAN.R-project.org/package=CAST.

Meyer, H., Katurji, M., Appelhans, T., Müller, M.U., Nauss, T., Roudier, P., Zawar-Reza, P., 2016. Mapping daily air temperature for Antarctica based on MODIS LST. Remote Sens. 8, 732. https://doi.org/10.3390/rs8090732.

Meyer, H., Kühnlein, M., Reudenbach, C., Nauss, T., 2017a. Revealing the potential of spectral and textural predictor variables in a neural network-based rainfall retrieval technique. Remote Sens. Lett. 8, 647–656. https://doi.org/10.1080/2150704X.2017.1312026.

Meyer, H., Lehnert, L.W., Wang, Y., Reudenbach, C., Nauss, T., Bendix, J., 2017b. From local spectral measurements to maps of vegetation cover and biomass on the Qinghai-Tibet-Plateau: do we need hyperspectral information? Int. J. Appl. Earth Obs. Geoinf. 55, 21–31. https://doi.org/10.1016/j.jag.2016.10.001.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T., 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ. Model. Softw. 101, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001.

Micheletti, N., Foresti, L., Robert, S., Leuenberger, M., Pedrazzini, A., Jaboyedoff, M., Kanevski, M., 2014. Machine learning feature selection methods for landslide susceptibility mapping. Math. Geosci. 46, 33–57.

Planetary Habitability Laboratory, 2015. Visible Vegetation Index (vvi). http://phl.upr.edu/projects/visible-vegetation-index-vvi.

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

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

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

Rocha, A.D., Groen, T.A., Skidmore, A.K., Darvishzadeh, R., Willemen, L., 2018. Machine learning using hyperspectral data inaccurately predicts plant traits under spatial dependency. Remote Sens. 10https://doi.org/10.3390/rs10081263. http://www.mdpi.com/2072-4292/10/8/1263.

Roubeix, V., Danis, P.-A., Feret, T., Baudoin, J.-M., 2016. Identification of ecological thresholds from variations in phytoplankton communities among lakes: contribution

to the definition of environmental standards. Environ. Monit. Assess. 188, 246. https://doi.org/10.1007/s10661-016-5238-y.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., Brenning, A., 2019. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecol. Model. 406, 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002. http://www.sciencedirect.com/science/article/pii/S0304380019302145.

Shi, Y., Song, L., Xia, Z., Lin, Y., Myneni, R.B., Choi, S., Wang, L., Ni, X., Lao, C., Yang, F., 2015. Mapping annual precipitation across mainland China in the period 2001–2010 from TRMM3B43 product using spatial downscaling approach. Remote Sens. 7, 5849–5878. https://doi.org/10.3390/rs70505849.

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., 2013. Prediction of soil organic carbon at the European scale by visible and near infraRed reflectance spectroscopy. PLOS ONE 8, 1–13. https://doi.org/10.1371/journal.pone.0066409.

Valavi, R., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G., 2018. blockcv: an r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. BioRxiv. https://doi.org/10.1101/357798.

Walsh, E.S., Kreakie, B.J., Cantwell, M.G., Nacci, D., 2017. A random forest approach to predict the spatial distribution of sediment pollution in an estuarine system. PLOS ONE 12, 1–18. https://doi.org/10.1371/journal.pone.0179473.

Wang, Y., Wu, G., Deng, L., Tang, Z., Wang, K., Sun, W., Shangguan, Z., 2017. Prediction of aboveground grassland biomass on the Loess Plateau, China, using a random forest algorithm. Sci. Rep. 7, 6940. https://doi.org/10.1038/s41598-017-07197-6.

Wenger, S.J., Olden, J.D., 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods Ecol. Evol. 3, 260–267. https://doi.org/10.1111/j.2041-210X.2011.00170.x.

Xie, Y., Efteliogl u, E., Ali, R.Y., Tang, X., Li, Y., Doshi, R., Shekhar, S., 2017. Transdisciplinary foundations of geospatial data science. ISPRS Int. J. Geo-Inform. 6https://doi.org/10.3390/ijgi6120395. http://www.mdpi.com/2220-9964/6/12/395.

Yang, R.-M., Zhang, G.-L., Yang, F., Zhi, J.-J., Yang, F., Liu, F., Zhao, Y.-G., Li, D.-C., 2016. Precise estimation of soil organic carbon stocks in the northeast Tibetan plateau. Sci. Rep. 6, 21842. https://doi.org/10.1038/srep21842.

Zanella, L., Folkard, A.M., Blackburn, G.A., Carvalho, L.M.T., 2017. How well does random forest analysis model deforestation and forest fragmentation in the Brazilian Atlantic forest? Environ. Ecol. Stat. 24, 529–549. https://doi.org/10.1007/s10651-017-0389-8.