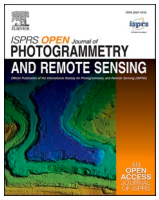
ISPRS Open Journal of Photogrammetry and Remote Sensing 5 (2022) 100018

Contents lists available at ScienceDirect

ISPRS Open Journal of Photogrammetry and Remote Sensing

journal homepage: www.journals.elsevier.com/isprs-open-journal-of-photogrammetry-and-remote-sensing



Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks

Teja Kattenborn [a,b,*], Felix Schiefer [c], Julian Frey [d], Hannes Feilhauer [a,b,e], Miguel D. Mahecha [a,b,e], Carsten F. Dormann [f]

[a] Remote Sensing Centre for Earth System Research (RSC4Earth), Leipzig University, Germany
[b] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany
[c] Institute of Geography and Geoecology (IfGG), Karlsruhe Institute for Technology (KIT), Germany
[d] Forest Growth and Dendroecology, University of Freiburg, Germany
[e] Helmholtz Centre for Environmental Research-UFZ, Leipzig, Germany
[f] Biometry and Environmental System Analysis, University of Freiburg, Germany

ARTICLE INFO

Keywords:
Spatial autocorrelation
Convolutional neural networks
Deep learning
Machine learning
Mapping
Reference data

ABSTRACT

Deep learning and particularly Convolutional Neural Networks (CNN) in concert with remote sensing are becoming standard analytical tools in the geosciences. A series of studies has presented the seemingly outstanding performance of CNN for predictive modelling. However, the predictive performance of such models is commonly estimated using random cross-validation, which does not account for spatial autocorrelation between training and validation data. Independent of the analytical method, such spatial dependence will inevitably inflate the estimated model performance. This problem is ignored in most CNN-related studies and suggests a flaw in their validation procedure. Here, we demonstrate how neglecting spatial autocorrelation during cross-validation leads to an optimistic model performance assessment, using the example of a tree species segmentation problem in multiple, spatially distributed drone image acquisitions. We evaluated CNN-based predictions with test data sampled from 1) randomly sampled hold-outs and 2) spatially blocked hold-outs. Assuming that a block cross-validation provides a realistic model performance, a validation with randomly sampled holdouts overestimated the model performance by up to 28%. Smaller training sample size increased this optimism. Spatial autocorrelation among observations was significantly higher within than between different remote sensing acquisitions. Thus, model performance should be tested with spatial cross-validation strategies and multiple independent remote sensing acquisitions. Otherwise, the estimated performance of any geospatial deep learning method is likely to be overestimated.

1. Introduction

In recent decades, our ability to image the Earth's land surface advanced due to several technological developments in remote sensing, including citizen science applications, Unmanned Aerial Vehicles (UAV) and space-borne high-resolution sensors (Colomina and Molina, 2014; Brandt et al., 2020; Ferreira et al., 2021; Schiller et al., 2021). Machine learning, and recently particularly deep-learning approaches are revolutionizing Earth system research using such data (Tuia et al., 2021). For empirical prediction tasks based on high resolution remote sensing images, Convolutional Neural Networks (CNN) have proven to be particularly suitable (Zhu et al., 2017; Brodrick et al., 2019; Kattenborn et al., 2021). The uptake of CNN by remote sensing researchers is driven by the synergy of fine-scaled spatial patterns revealed through high-resolution images and the effectiveness of CNN to distill them. CNNs are composed of a series of sequential filter functions (convolutional layers), which are iteratively optimized over observations with respect to the target variable (Goodfellow et al., 2016). For remote sensing data, this iterative training process is commonly performed on equal-sized spatial subsamples referred to as tiles or crops. Thereby, the neural network learns the informative spatial patterns in such tiles, e.g. those apt to identify a plant species. The fundamental advantage of such CNN-based pattern recognition in remote sensing data compared to previous approaches (e.g. texture metrics) is that it minimizes feature design and variable

* Corresponding author. Remote Sensing Centre for Earth System Research (RSC4Earth), Leipzig University, Leipzig, Germany.
E-mail address: teja.kattenborn@uni-leipzig.de (T. Kattenborn).

https://doi.org/10.1016/j.ophoto.2022.100018
Received 4 February 2022; Received in revised form 3 June 2022; Accepted 12 June 2022
Available online 21 June 2022

selection and automatically learns image patterns with an effectiveness that enables mapping continuous gradients and discrete entities at unprecedented spatial detail and accuracy (Zhu et al., 2017; Kattenborn et al., 2021).

CNNs are typically purely data-driven, so that the generalization of such models in the application domain is determined by the representativeness of the training data. In most cases, the generalization of such models is constrained due to limited availability of remote sensing data with associated reference observations (Kattenborn et al., 2021). This may not be an issue when a model is being used to interpolate within the spatial or environmental range of the training data set (Wadoux et al., 2021; Stehman et al., 2021; Meyer and Pebesma, 2021; Mila et al., 2022). However, remote sensing applications typically aim to extrapolate (predict) to unseen observations within the application domain for which reference data for validation is not available. These unseen observations may depart from the training data in terms of site conditions or properties of the remote sensing data.

The expected predictive performance of a model to unseen observations can be estimated using independent samples. For this, models are most often cross-validated, meaning that available samples are used alternatingly to either train or validate the model. However, the fact that an observation was not used for training a model does not necessarily imply that this observation is truly independent from the training data. A dependence between observations can already arise through their spatial proximity, since usually *nearby things are more related than distant things* (Tobler, 1970). A remote sensing signal from a tree, for example, is likely to be more similar to that of an immediate neighbour than to that of a distant tree. This phenomenon is referred to as spatial autocorrelation and can be observed across all spatial scales (Legendre, 1993; Dormann, 2007). This does not affect the estimation of *map accuracy* as long as the validation sample reflects the population of the map area (Wadoux et al., 2021; Brus, 2021; Mila et al., 2022). However, remote sensing applications often do not strive to evaluate a method for producing a map for a specific area, but to estimate the expected *model performance* for the entire application domain (i.e. model generalization). For the latter, spatially autocorrelated training and validation samples will inevitably result in an overly optimistic model performance, which in turn will not reflect the generalization of a model across the application domain (Bahn and McGill, 2013; Le Rest et al., 2014; Pohjankukka et al., 2017; Roberts et al., 2017; Ploton et al., 2020).

Such inflation of model performance can be circumvented with spatial cross-validation strategies, which create spatially independent training and validation folds through spatial blocking or buffering observations, and have been applied in various forms and contexts (Veloz, 2009; Wang et al., 2010; Wenger and Olden, 2012; Le Rest et al., 2014; Roberts et al., 2017; Valavi et al., 2018; Mahecha et al., 2021), including in the specific context of remote sensing and machine learning (Brenning, 2012; Rocha et al., 2018; Schratz et al., 2019; Ploton et al., 2020; Meyer and Pebesma, 2021). However, most studies that used CNNs for geoscience-oriented remote sensing applications used random cross-validation and, thus, did not ensure spatial independence between training and validation data (e.g. more than 90% of the studies reviewed in Kattenborn et al., 2021). It can be assumed that a large share of these studies (unintentionally) report overly optimistic model performance. Possible explanations are limited awareness of the problem and its relevance for deep learning or even CNN applications.

Here, we thus investigate to what degree spatially autocorrelated training and validation data can lead to over-optimistic evaluation of CNN models. For that, we chose a case study on CNN-based tree species segmentation with a data set composed of numerous spatially widely distributed UAV image acquisitions and wall-to-wall reference observations. We compare the predictive performance estimated with observations (image tiles) sampled in a random and a spatial block cross-validation. The effect of these two cross-validation strategies is further explained by quantifying the spatial dependence between image tiles. Finally, we discuss the relevance of these findings for any type of predictive deep learning approach applied using Earth observations.

2. Methods

We demonstrated the effect of spatially autocorrelated training and validation samples with a case study of a tree species classification (semantic segmentation) in RGB orthoimages acquired with UAVs. The application domain is the Black Forest region, which was covered with 47 sites of 100 × 100 m, each covered by an orthoimage acquisition. Thus, in this case study it is aimed to estimate the performance of CNN models to segment tree species in UAV images across the Black Forest region. The underlying experimental approach is based on the assumption that samples, i.e. image tiles extracted from the orthoimages for CNN training and prediction, are spatially dependent when extracted from the same site. Consequently, an optimistic model performance is expected when using such data for both training and validation (random cross-validation). In contrast, a more realistic model performance can be derived from samples that are extracted from sites from which the model has not seen any samples during training, assuming that these are spatially independent (block cross-validation). The discrepancy between these two modes of validation (dependence vs. independence of training and validation samples) must be regarded as optimism.

2.1. Remote sensing predictors and reference data (labels)

The 47 individual sites are distributed across more than 3600 km^2, with a minimum, average and maximum distance of 0.9, 31.7, and 82.0 km, respectively (Fig. 1). The area is mostly covered by mixed and coniferous forests and covers a wide range of forest types and age classes. The tree species that are most frequent in these sites and that are targeted for the CNN-based semantic segmentation are *Picea abies* L., *Fagus sylvatica* L., *Abies alba* Mill., *Quercus robur* L., *Acer pseudoplantus* L., *Larix decidua* Mill., *Pinus sylvestris* L., *Betula pendula* Roth, *Fraxinus exelsior* L., and *Pseudotsuga menziesii* Mirbel. Further details on the tree stands are given in Storch et al. (2020) and in the Appendix (Table 1).

For each of the 47 sites, the predictors in form of RGB orthoimages with a ground resolution of approximately 1 cm were created using

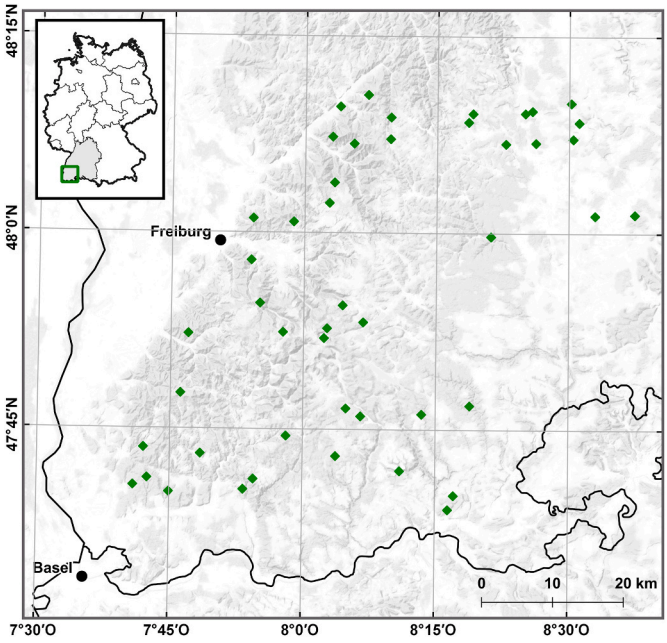


Fig. 1. Locations of the UAV orthomosaics (n = 47, green dots) acquired between 2017 and 2019 in the Black Forest region, Southwest Germany. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

aerial images from a multicopter (Mikrokopter, OktoXL 6S12) equipped with a Sony Alpha 7R and the software Agisoft Photoscan (for details see Frey et al., 2018). The imagery was acquired between 07:40 a.m. to 03:40 p.m. (CEST) from April to November in the years of 2017, 2018, and 2019 (for details see Appendix, Table 1). The orthoimages, hence, cover a wide range of possible variations, such as:

- illumination conditions induced by varying sun zenith and azimuth angles or the ratio of direct and diffuse solar radiance;
- vegetation status, in terms of phenology and health condition;
- forest structural characteristics, such as tree density, species composition, stand age, or management;
- site characteristics, including soil background, topography, or understory.

All orthoimages were cropped into non-overlapping and directly adjacent tiles. Based on the results of Schiefer et al. (2020), we used a tile size of 256 × 256 pixels (approx. 2.56 × 2.56 m). This resulted in around 1500 tiles per orthoimage and site.

A semantic segmentation aims at an area-wide classification of the target classes. For this case study, this implies that the CNN-based segmentation assigns each pixel of a tile to one of the tree species mentioned above. Training common CNN-based segmentation algorithms requires spatially explicit and wall-to-wall reference data, meaning that each image tile (predictors) used for training is associated with a mask containing species information for each pixel in that tile (response). The masks were created from polygons available for all targeted species, which were created with visual interpretation from imagery aided with ground observations. Further information on the site, data acquisition, and visual interpretation is given in Frey et al. (2018) and Schiefer et al. (2020). The entire data set, including orthoimagery, tree-species delineations and its metadata are openly accessible (https://dx.doi.org/10.35097/538).

2.2. Demonstrating optimistic model evaluation induced by spatial autocorrelation

The degree of optimism was assessed with multiple model setups. Firstly, we aimed to illustrate that optimism induced by spatially autocorrelated training and test data occurs across small and large sample sizes. Therefore, we varied the number of orthoimages used for model training with n = 10, 25, and 40. Secondly, we test if model regularization via a data augmentation, which is commonly applied to reduce model overfitting, can enhance the generalization on the independent test data. For this, each model setup was trained with and without augmented training data. The data augmentation was applied in three different ways that are commonly used in the literature (Kattenborn et al., 2021): 1) geometric modifications (horizontal and vertical flipping), 2) radiometric modifications (random change of brightness between 90 and 110%, contrast between 80 and 120%, and saturation between 80 and 120%), and 3) geometric and radiometric modifications in combination.

Each of these different model setups was evaluated with a random and a block cross-validation with five repetitions (Fig. 2). For this, the available orthoimages (n = 47) were randomly assigned to the random (n = 10, 25 or 40) or block cross-validation procedure (n = 7). In the random cross-validation procedure, 80% of the image tiles of the assigned site were used for model training and the remaining 20% for estimating the performance of the trained model objects (considered as non-independent samples). For the block cross-validation, we used exactly the same model instance, but the model performance was estimated with tiles that were extracted from those sites (n = 7) from which the model has not seen any image tiles during training (independent samples). The number of sites sampled for the block cross-validation (n = 7) was set arbitrarily as a compromise between sufficient data for model training and validation.

The CNN-based tree species segmentation was based on the U-net architecture (Ronneberger et al., 2015), which is frequently applied in vegetation remote sensing (Wagner et al., 2019; Schiefer et al., 2020; Kattenborn et al., 2019, 2021). The implemented U-net featured four encoding and four decoding blocks. Root Mean Squared Propagation (RMSprop) was selected as optimizer with a learning rate of 0.0001 and the F_1-score (= Sørensen index or Dice coefficient) was used as loss function. For each model setup, differing in the number of sites for training and mode of data augmentation, the U-net was trained in 50 epochs and the best performing model parametrization of these epochs was selected for the final prediction (selected with a 20% holdout of the training data). Model performance was finally compared using F_1-scores (eqn. (1)), i.e. the harmonic mean of precision and recall:

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}, \quad (1)$$

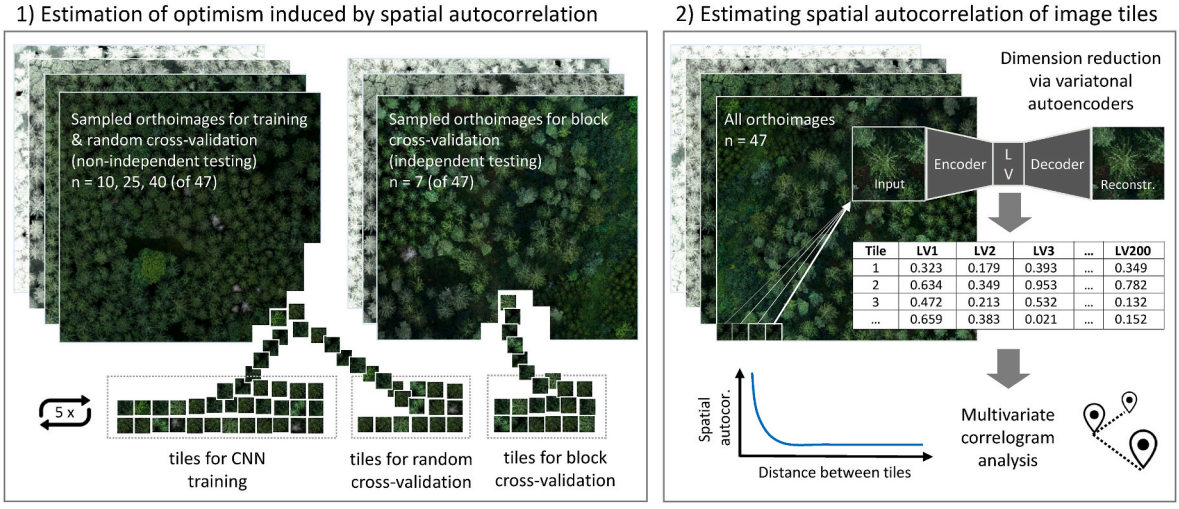


Fig. 2. Workflow used to demonstrate and explain optimistic model evaluation resulting from spatially autocorrelated image tiles used for training and validation. The non-independent and independent model evaluations (left) are implemented using a random and block cross-validation, respectively. The variational autoencoder (right) is based on a MMD-VAE (Zhao et al., 2017).

where TP stands for true positives, FP for false positives, and FN for false negatives. The significance of differences between F_1-scores was assessed with one-tailed t-tests. Further details on the U-net implementation are given in Schiefer et al. (2020). The code and preprocessed data are available at https://github.com/tejakattenborn/cnn_rs_optimism.

We tested if the optimism was not due to sampling effects using metadata available for each site. For this, based on the repeated cross-validations, we subtracted the site-specific F_1-scores obtained when the sites were included in model training (random cross-validation results) with F_1-scores obtained when the sites were excluded from model training (block cross-validation results). We calculated the R^2 for these site-specific differences in F_1-scores and the metadata. The latter included the image acquisition date, the image acquisition time, stand density [trees/ha], and the tree species composition in the form of the first three components of a principal component analysis (PCA) of the tree species reference data.

2.3. Spatial autocorrelation between image tiles - towards explaining optimistic model evaluation

The above-described experiment on tree species segmentation in image tiles aims to demonstrate predictive optimism induced by randomly sampled and potentially spatially autocorrelated image tiles. To support the results and assumptions in this experiment, the below described procedure was applied to determine the mean spatial autocorrelation between image tiles as a function of their geographic distance.

A common method for assessing spatial autocorrelation are correlograms, which can be used to quantify the similarity of observations at given spatial distances. Commonly, such methods require tabular data, where an observation is structured as vector. However, CNN-based image analysis is a pattern-oriented problem, where each observation (image tile) is a high dimensional and structured array (of rank 2 for grey-level, or of rank 3 for multi-channel images, respectively). Assessing the spatial autocorrelation of such higher-dimensional image-type observations with correlograms is not directly possible and requires a dimension reduction that respects the spatial structure of data. One suitable approach is CNN-based variational autoencoders (VAE), which use variational inference to generate a latent representation (vector) from the input arrays in an unsupervised way requiring few heuristics. Compared to classical dimensionality reduction techniques such as Principal Component Analysis (PCA), variational autoencoders enable the representation of complex and non-linear dependencies (Goodfellow et al., 2016; Fournier and Aloise, 2019). Moreover, autoencoders can integrate convolutional layers, which, instead of slicing and stacking the image data into vectors, retain the spatial relationships for an efficient detection of features and patterns therein (Pu et al., 2016). We choose variational autoencoders over vanilla autoencoders as these enable to define priors that constrain the modelling of the latent variables. This does not only result in a regularization of the latent space, but also facilities that the latter holds for new observations (Kingma and Welling, 2013).

Autoencoders are composed of an encoder block transforming the input data into a low-dimensional latent representation and a decoding block that is meant to reconstruct the input array from the latent representation (Fig. 2). A successfully trained variational autoencoder is able to represent the underlying patterns by a few latent variables. This can be verified by decoding these latent variables: If a VAE can reconstruct the input image from the encoded latent variables with only few deviations (decoding), it can be assumed that most image patterns have been preserved (learned) with negligible loss of information (Kingma and Welling, 2019).

We implemented a maximum mean discrepancy variational autoencoder (MMD-VAE), which was demonstrated to generate robust latent representations while being computational efficient (Zhao et al., 2017). We used a common encoder-decoder structure, with five convolutional and deconvolutional layers each. These layers featured a stride of two, kernel-size of three and were connected with GeLu activation functions (Gaussian Error Linear Units). The loss of MMD-VAE is determined by the sum of 1) the reconstruction result (squared difference of the input image and the decoder output) and 2) the maximum mean discrepancy (MMD) of the distribution of the latent space and a prior - here a Gaussian distribution. MMD compares the latent and prior distribution by means of their moments derived using Gaussian kernels. Details on the MMD-VAE are given in Zhao et al. (2017). The R-based Tensorflow implementation and image tiles are available online (https://github.com/tejakattenborn/cnn_rs_optimism).

The MMD-VAE was trained on regularly spaced and non-overlapping tiles of all orthoimages with 256 × 256 pixels size (same settings as for the tree-species segmentation, section 2.2). The tiles were encoded into a vector of 200 latent variables. A holdout of 10% of the tiles were used to monitor the training progress (loss) and to evaluate the reconstruction error (see Appendix). The training of the MMD-VAE was stopped with the convergence of the loss after 236 epochs. Subsequently, the trained MMD-VAE was used to predict the 200 latent variables for all tiles. The latent variables for all available tiles of the orthoimagery (predictors) were then used to quantify the spatial autocorrelation by means of multivariate correlograms. These estimate the spatial dependence across discrete distance classes (lags) using the centred Mantel statistic (Bjørnstad et al., 1999, 2001). The correlograms were created with *correlog* function in the R-package *ncf* (Bjørnstad, 2020), with a distance interval (lag) of 1 m and a minimum distance of 2.56 m. To compare the spatial autocorrelation of images tiles (predictors) with the tree species cover (response), we created additional correlograms for tree species cover values (%), which were derived from the masks available for each image tile (cf. section 2.1). For visual comparability of the predictor and response spatial autocorrelation, both correlograms were scaled between the maximum and the arithmetic mean.

3. Results

Model performance (F_1-score) estimated from non-independent test data, i.e. tiles sampled with a random cross-validation, significantly exceeded the model performance derived from independent test data based on the block cross-validation ($t = 22.734$, $P < 0.01$, Fig. 3). This optimism, i.e. the difference between the independent and the non-independent validation, increased when using fewer training data. It amounted to 14.6%, 17.7% and 27.9% when using 40, 25, and 10 UAV-orthoimages for model training, respectively.

The site-specific F_1-scores obtained from both the random and block cross-validation did not show notable correlations with site-specific metadata (acquisition date or time, stand density or tree species cover), indicating that the models generalized over these properties. The repeated cross-validation scheme also enabled the quantification of the site-specific optimism and its comparison with the corresponding metadata. This revealed that the optimism is largely independent of sampling effects emerging from the cross-validation scheme. The R^2 between optimism and site-specific metadata across the different model setups amounted to 0.0 for acquisition date, 0.0 for the acquisition time, 0.09 for the stand density, and 0.13, 0.04, and 0.0 for the three PCA axis synthesizing the tree species composition.

Augmenting the training data, i.e., radiometrically and geometrically modifying the input tiles, did not significantly reduce optimism for

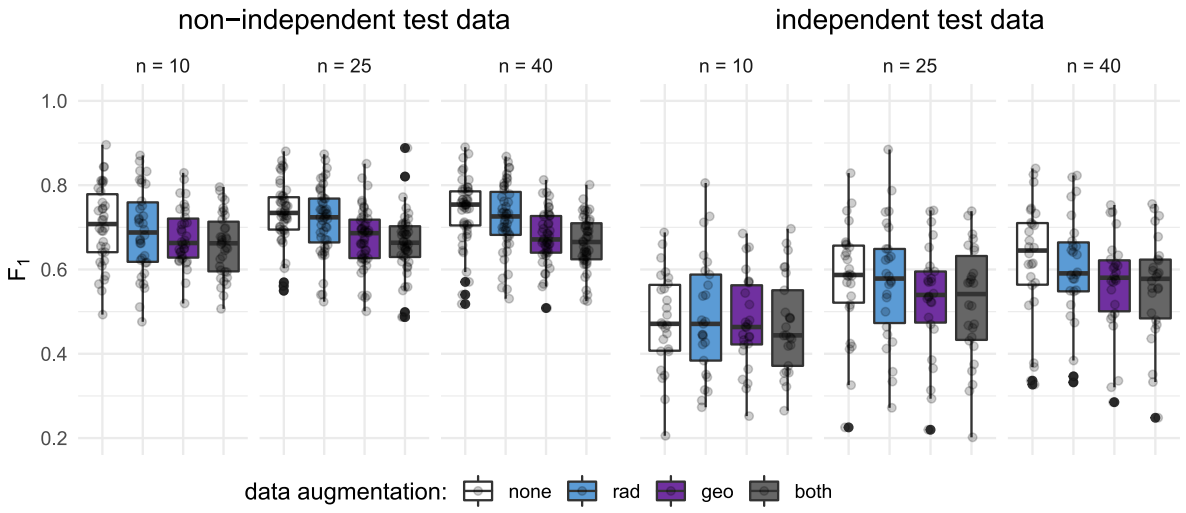


Fig. 3. Predictive performance (F_1-score) per site evaluated with potentially non-independent, spatially auto-correlated observations derived with a random cross-validation (left) and independent validation samples derived from a block cross-validation (right). Results are shown for models trained with different numbers of orthoimages (n = 10, 25, 40). The colours of the boxes indicate whether training data were augmented with radiometric, geometric or both modifications. Grey dots show F_1scores for individual sites sampled during the five repetitions of the cross-validation procedures. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

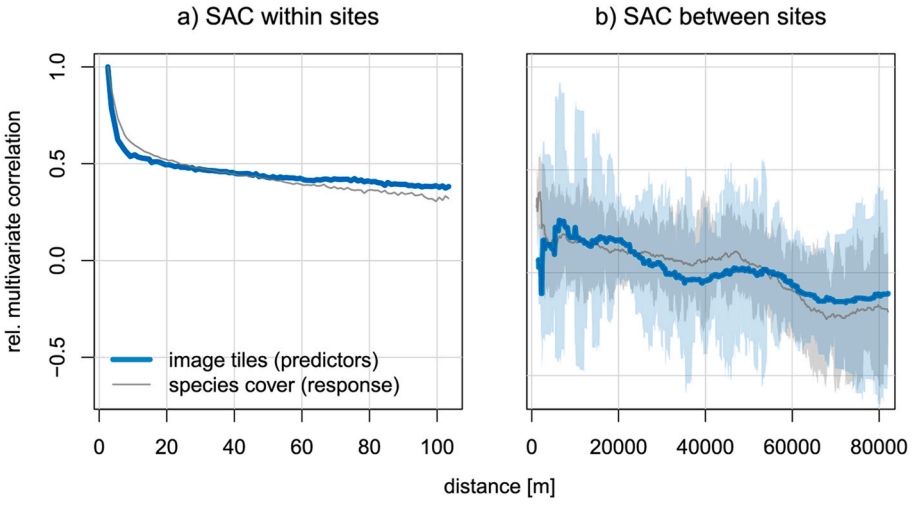


Fig. 4. Multivariate correlograms quantifying the spatial autocorrelation (SAC) in terms of the Mantel statistic within (a) and between (b) sites. Items in blue corresponds to the predictors (tiles extracted from the orthoimagery) and grey to the response (species cover). The correlogram for the predictors was calculated from 200 latent variables derived from tiles of the RGB orthoimages using a CNN-based variational autoencoder (MMD-VAE). In contrast to a), the correlations between sites (b), was highly variable (cf. discussion) and were, hence, visualized by a rolling mean (width = 10 km) and polygons depicting the area within the rolling standard deviation (width = 1 km). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

any number of sites used during training, i.e. 10 ($t = -0.301$, $P = 0.764$), 25 ($t = -0.232$, $P = 0.818$) or 40 ($t = 0.573$, $P = 0.572$). For both independent and non-independent validation, model performance was even significantly lower when incorporating a data augmentation by means of geometric modifications ($t = -2.37$, $P = 0.014$), radiometric modifications ($t = -2.60$, $P = 0.009$) or both ($t = -5.34$, $P < 0.001$).

The spatial autocorrelation of the predictors, i.e. the similarity of tiles of the orthoimages as a function of distance, steeply increased with distances below < 20 m and flattened for distances > 20 m (Fig. 4). Still, the overall autocorrelation within sites was still larger than between sites: The Mantel statistic for observations within sites (Fig. 4a) was significantly higher than for the Mantel statistic calculated from observations of different orthoimages (Fig. 4b; $t = 40.036$, $P < 0.001$). Moreover, the correlogram revealed that spatial autocorrelation of tiles was more heterogeneous across sites (Fig. 4b) than within sites (Fig. 4a). The spatial autocorrelation of the response, i.e. the tree species cover in each tile, showed overall a very similar pattern. Yet, we found a lower variation of the species cover per distance bin than for the image tiles.

4. Discussion

Our results are in line with previous studies that have shown with other machine learning methods that predictive performance estimates can be severely inflated if training and validation are spatially auto-correlated (Meyer et al., 2018; Rocha et al., 2018; Schratz et al., 2019; Meyer and Pebesma, 2021). For instance, Ploton et al. (2020) exemplified with random forest models overly optimistic model evaluation at the example of predicting above ground biomass from multispectral satellite imagery (MODIS). They interpret their findings that due to the fact that reflectance data is not strongly related to biomass, models are, thus, likely to learn non-invariant relations between the response (biomass) and predictor variables (reflectance), which do not hold across new, unseen domains. For the present case study on tree species segmentation with CNN-based pattern recognition in centimetre-scale

orthoimagery, however, one may expect that models are predominantly learning mechanistic and, hence, readily transferable relationships: Canopy and branching patterns as relatively unique indicators for certain species. Yet, our experiment demonstrated that even in such case, violating the independence among training and validation data severely inflates model accuracy.

To be clear, fitting a CNN to spatially autocorrelated data does not necessarily invalidate its prediction. Unless the input data themselves are unbalanced, the mean prediction of the CNN may be unbiased. However, the comparison between prediction and validation data is compromised, so that the quality of prediction is overestimated by the non-independence. Why is this critical? Firstly, optimistic estimates of model performance can lead to false impressions about the value of a method or predictors. For instance, in the above-mentioned example on satellite-based biomass mapping (Ploton et al., 2020), a high accuracy was found with a random cross-validation, while a spatial cross-validation resulted in no predictive power (R^2 of virtually 0). Using null-models, the authors showed that the models indeed did not learn satellite-signal-biomass relations, but indirectly learned the geographic space from the satellite data and how biomass values are distributed therein. Secondly, optimistic model performance may inflate the reliability of prediction maps for regions where predictor relationships depart from the relationships learned during training (further discussed below and in Rocha et al., 2018; Meyer and Pebesma, 2021).

The present study exemplified that spatial autocorrelation can induce optimistic model evaluations at the example of segmenting tree species in UAV imagery. However, this fundamental problem obviously exists equally for other application domains (e.g, precision agriculture or land cover classification), remote sensing data of any type, scale and quality (Dormann, 2007; Rocha et al., 2018), and modelling task, such as image classification and regression, object detection, or instance segmentation. Also, the modelling method is incidental with respect to an optimistic model evaluation (dependent data remain dependent data) and the problem also persists regardless of the model complexity and depth (cf. Meyer et al., 2018; Rocha et al., 2018; Ploton et al., 2020; Meyer and Pebesma, 2021; Schratz et al., 2019).

Additionally, our results show that a spatial dependence of training and validation data cannot be circumvented by increasing the sample size or applying model regularization techniques, as the fundamental issue is manifested in the raw data. Our results suggest a higher inflation of model performance with decreasing sample size. However, this does not necessarily imply that increasing the sample size can compensate effects induced by spatial autocorrelation. Rather, this means that model evaluations based on small sample sizes are even more likely to report optimistic performance, if spatial independence is violated (cf. Rocha et al., 2018). Regularizing models with different data augmentation techniques did not have an ameliorating effect on the observed performance inflation (Fig. 3). The different data augmentation schemes even decreased the overall model performance in the random and block cross-validation, although similar approaches are used in a wide range of studies (Kattenborn et al., 2021). We assume that the heterogeneity of the dataset already leads to a high model bias, while a synthetic inflation of the latter using a data augmentation resulted in a relative model underfitting (contrary to the common rationale of reducing overfitting via data augmentation strategies, cf. Wong et al., 2016; Shorten and Khoshgoftaar, 2019). The geometric augmentation may have even introduced unrealistic images. For instance, directions of cast shadows are naturally constrained by sun azimuth angles.

Quantifying the spatial autocorrelation of the predictors using correlograms elucidated the cause of optimistic model evaluation (Fig. 4): Highest spatial autocorrelation among tiles was found for distances < 20 m. In such close proximity, tiles may even correspond to the same tree crown and, thus, merely represent pseudo-replicates that hence are not at all suitable for model evaluation. Although the spatial autocorrelation decreases steeply after a few meters distance and appears to flatten, it should be noted that spatial autocorrelation within sites (Fig. 4a) is still significantly higher than between sites (Fig. 4a). Overall, we found very comparable patterns in spatial autocorrelation between tree species cover and the image tiles. Yet, the variation per distance was considerably higher for image tiles (predictors). This may be explained by the varying appearance of trees within a species due to varying phenological states or environmental conditions at the time of the image acquisition. For instance, the appearance of plant canopies might vary considerably due to seasonal leaf development, inflorescence and the branching structure may even show diurnal changes induced by acute water availability (Junttila et al., 2021; Schiefer et al., 2021). Moreover, higher variation of the spatial autocorrelation of the image tiles may be a result of image acquisition settings. For instance, the texture of tree canopies may largely depend on the ratio of diffuse and direct radiance, the sun zenith angle and topography (Lopatin et al., 2019). In the presented experiment the models appeared to generalize well over the tested image acquisition settings (time and date). However, it is likely that spatial autocorrelation in predictors may not only result from continuous gradients of environmental variables per se, but also of (spatially varying) data acquisition properties. Thus, for testing the generalization of predictive methods with remote sensing data, multiple acquisitions appear to be a compelling necessity.

The spatial autocorrelation derived of the predictors showed a stronger variability between orthoimages than within orthoimages (Fig. 4b). This can firstly be explained by the many times lower number of observation pairs (tiles) at a given distance between orthoimages than within orthoimages (mean of 2713 pairs for distance intervals > 100 m; 42,062 pairs for distance intervals < 100 m). Secondly, the increased heterogeneity in spatial autocorrelation between sites can also be attributed to distinct scene-specific characteristics (cf. previous paragraph). This in turn highlights that models should not only be evaluated using independent remote sensing data acquisitions, but several independent acquisitions covering the expected variation in site and acquisition conditions within the application domain. As demonstrated here, multiple acquisitions can be iteratively and alternately used for training and validation, e.g. using a block cross-validation (Roberts et al., 2017; Ploton et al., 2020), where, for example, each block corresponds to a single acquisition. Considering the computational load of most deep-learning applications, such a spatial cross-validation can be certainly challenging, but in our opinion, it is the only way to reliably assess the performance and transferability of such models.

We want to strongly emphasize that the experiments conducted here aim to evaluate model *extrapolations*, specifically how dependence between training and validation samples results in an overestimation of a model's ability to predict to new, unseen observations. This should not be confused with assessing the accuracy of prediction maps derived from model *interpolations* within the parameter and spatial range of the training data (Wadoux et al., 2021; Brus, 2021; Stehman et al., 2021). Evaluating the generalization of a model during extrapolation and evaluating a map product are two different questions: A map product may be evaluated with a probability sample drawn from the respective population, for instance with simple random cross-validation scheme using randomly distributed samples or more sophisticated stratified or balanced sampling approaches (Wadoux et al., 2021; De Gruijter et al., 2015; Mila et al., 2022). In contrast, if a method is to be evaluated in terms of its predictive performance and generalization to new, unseen observations - and this is the scope of most remote sensing applications -

a sampling scheme must maximize the independence of training and validation samples to realistically approximate the predictive performance of such an extrapolation (Ploton et al., 2020).

A considerable part of the studies in the geosciences and machine learning context, especially recent ones based on deep learning methods, intended and concluded that their methods have large predictive performance (Kattenborn et al., 2021), while extrapolations to truly independent data have in fact rarely been tested. Thus, there is often a discrepancy between the intention of such a study (commonly to evaluate a method to predict something from remote sensing data), the corresponding suitability of the conducted experiments and the conclusions drawn from them. In other words, one cannot conclude that a method has an expected accuracy of 80% for a domain if this has not been explicitly tested. Therefore, an appropriate assessment of the generalization of methods should be conducted to communicate reliable uncertainties of spatial predictions and thereby anticipate mistrust emerging from (unintended) exaggerations of model performances.

When can spatial autocorrelation lead to optimistic model evaluation and how can we approximate reliable predictive performance instead? As demonstrated by our results, optimism occurs when training and validation samples are dependent (very similar). In such case, the approximated predictive performance does not represent a model's generalization to unseen observations, for instance a remote sensing acquisition with different illumination conditions or a forest with a different stand structure. There is not a standard approach to approximate the generalization of a model as this very much depends on the purpose of the application, the structure of the data and requirements of the user (e.g., a desired performance, targeted locations; cf. Meyer and Pebesma, 2021). Commonly, variants of spatial cross-validation strategies are used to constrain the geographical proximity of training and validation samples, including block cross-validation, which clusters observations into spatially disjoint training and validation subsets, or spatial leave-one-out cross-validation, where observations within the geographic vicinity of a validation sample are excluding during training (Brenning, 2012; Wenger and Olden, 2012; Roberts et al., 2017; Ploton et al., 2020; Mila et al., 2022). It has also been suggested to consider, in addition to the spatial distance, distances in environmental space (Valavi et al., 2018; Mahecha et al., 2021). Here, we spatially cross-validated our models using a block cross-validation strategy, where each individual site represents a block, as we did not find a strong dependence of observations (tiles) between different orthoimage (cf. Fig. 4). Note, however, that dependencies between holdouts of a data set can only be minimized and never completely avoided.

Also note that the approximated model performance can only be assumed to hold within the predictor space (or domain) in which the model has been evaluated, aptly referred to as 'area of applicability' in Meyer and Pebesma, (2021). A promising option to approximate the uncertainty of predictions derived from new observations is to compare the similarity of the predictor space of these new observations with the predictor space of the observations that have been used for training the model. Meyer and Pebesma (2021) showed that a clear positive correlation can be expected between the actual prediction errors and the dissimilarities in the predictor space between training and new observations. Based on this relationship, they demonstrated a transferable method for identifying the 'area of applicability' in which the uncertainty approximated by spatial cross-validation can be expected to hold. However, applying such a method to data with high-dimensional predictor space is non-trivial. A variational autoencoder-based approach as used in this study, and the associated possibility of almost lossless transformation of high-dimensional information into a reduced latent space, could provide an efficient tool for comparing high dimensional predictor spaces between new observations and those used in training (cf. Janet et al., 2019).

5. Conclusions

Convolutional Neural Networks in concert with remote sensing observations are paving new avenues for predictive modelling in the geosciences. Although a series of studies has presented seemingly outstanding potentials, problems arising from spatial autocorrelation of input data are frequently ignored. Our results suggest that violating spatial independence between training and test data can severely inflate model apparent performance (up to almost 30%) and, hence, lead to an overly optimistic evaluation of the generalization of such models. While potential optimism induced by spatial dependence can generally not be prevented by larger sample sizes, model performance estimates from small sample sizes are more strongly inflated.

CNN are typically applied to higher dimensional data, such as tiles from orthoimagery or point clouds, which are not directly compatible with typical methods for assessing spatial autocorrelation, as the latter commonly require tabular data. We presented an effective, unsupervised, and transferable approach to quantify spatial autocorrelation between image tiles using a dimension reduction from higher ranked arrays to vector data using variational autoencoders. Such an approach may facilitate to reveal spatial autocorrelation in image data and may serve as an effective way to implement spatial cross-validation strategies.

The spatial autocorrelation of the data set used in this study, composed of numerous UAV orthomosaics, highlighted that not only tiles in close proximity are spatially autocorrelated, but that also far-distant tiles within the same acquisition are generally higher autocorrelated than tiles between different orthoimages. A robust model evaluation should therefore include multiple independent remote sensing acquisitions with a spatial cross-validation strategy.

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.

Acknowledgements

The study has been supported by the German Aerospace Centre (DLR) on behalf of the Federal Ministry of Economics and Technology (BMWi) under the project *UAVforSAT* (FKZ 50EE1909A) and the German Research Foundation (DFG) under the project *BigPlantSens* (Project number 444524904). The data acquisition within the Black Forest was founded by the German Research Foundation DFG (GRK 2123). Open Access funding enabled and organized by Project DEAL. We acknowledge support from Leipzig University for OpenAccess Publishing. We would like to thank Sebastian Schmidtlein, who provided a crucial impetus for initiating this study.

Appendix

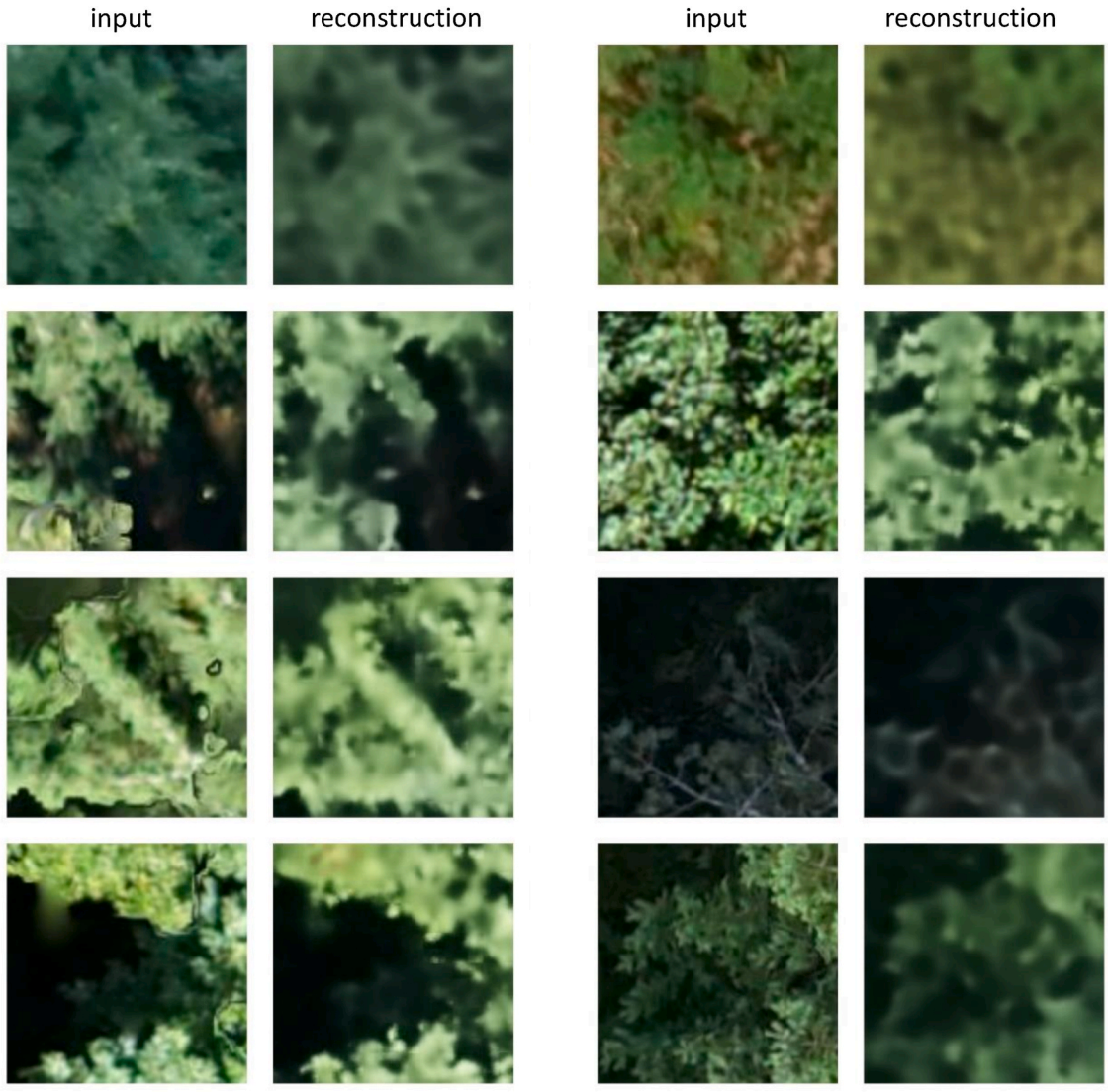


Fig. A1. Input tiles and the corresponding reconstruction (decodings) derived from the vector-based latent representation of length 200 (encodings). The average reconstruction error was 0.0046% (root mean squared difference between input image and decoded image).

Table A1
Overview on the species composition (*n* individuals, see Storch et al., 2020 for details) and the acquisition date and time (CEST) of the orthoimagery for each site.

Site-ID	*P. abies*	*A. alba*	*F. sylvatica*	*A. pseudoplatanus*	*P. menziesii*	*P. sylvestris*	*L. decidua*	*F. excelsior*	*Q. spec*	*B. pendula*	Other	Sum trees	Number of species	Acquisition date	Acquisition time
008	420		43									463	2	05.10.2017	12:50
014	36	47	83		90		19					275	5	11.07.2017	14:25
019	26	105	213	7								351	4	14.06.2017	14:15
030	17	60	116	1	171		56		8	35		464	8	14.06.2017	09:30
031	95	6	125	1			1					228	5	14.06.2017	08:05
035	38	17	141					1		1		198	5	03.07.2017	07:40
037		33	298									331	2	09.07.2018	12:45
044	347	31	22	1					1			402	5	11.07.2017	11:55
045	171	12	91	4	1	75		1			6	361	7	04.07.2017	15:25
050	222	287		54				4				567	4	04.07.2017	11:15
053	377	98	3			16						494	4	28.09.2017	10:10
056	134	308	129		3	7	3					584	6	11.07.2017	13:00
057	181	179	82			65		1				508	5	13.07.2017	11:10
061		22	142	10			24				11	209	4	10.07.2018	09:40
071	232	3	15	2		149	1	1				403	7	24.10.2017	08:20
084	666	90	17			15						788	4	24.10.2017	09:35
085	490	60	3			86	5					644	5	13.07.2017	14:40

(*continued on next page*)

Table A1 (*continued*)

Site-ID	*P. abies*	*A. alba*	*F. sylvatica*	*A. pseudoplatanus*	*P. menziesii*	*P. sylvestris*	*L. decidua*	*F. excelsior*	*Q. spec*	*B. pendula*	Other	Sum trees	Number of species	Acquisition date	Acquisition time
089	103	85	40	5			1					234	5	13.07.2017	12:20
091	64	26	316	27								433	4	03.07.2017	09:35
096	36	231	399			11			28			705	5	04.07.2017	13:05
106	28	1	265	4	2		1		1			302	7	03.07.2017	14:50
110	242	1	9	27			4					283	5	10.07.2018	11:00
111	79	8	44	79					2	9		221	6	14.06.2017	12:00
117	543	22	18	1		14						598	5	28.09.2017	13:50
121	7		148		84	2			49			290	5	03.07.2017	13:45
122	285	5	185		13	9	33					530	6	14.06.2017	10:40
124	345	13	9		13	9	33			2		369	4	03.11.2017	10:00
125	68	4	48	32			3	10	1			166	7	14.06.2017	07:05
129	5	157	125	2	2			5	12			308	7	11.06.2017	09:15
133	47	95	121						1			264	4	06.07.2017	08:55
134	339	52	2	1		105				3		502	6	03.11.2017	12:50
140	313	27	28		20	29	18					435	6	13.07.2017	15:00
151	170	91	11		37	2	2					313	6	20.04.2018	15:41
156	216	5	61	2		1						285	5	06.07.2017	10:35
162	653					180						833	2	04.10.2017	13:55
163	182	114	14	1		1	1	2				315	7	04.10.2017	12:00
167	469	3	73									545	3	05.07.2017	13:20
171	197	43	444									684	3	04.07.2017	11:25
173	274	2	52									328	3	04.07.2017	10:05
184	172	152	10									334	3	11.07.2017	09:40
003	78	11	30	6								125	4	10.09.2019	14:25
021	249	8	25	2						3		287	5	02.09.2019	14:50
073	141	89	281									511	3	28.09.2019	14:35
114	51	36	7	11	53		3	1				162	7	28.08.2019	11:00
128	270		109	21	1							401	4	10.09.2019	15:30
130	133	81	131	2	16					1		364	6	02.09.2019	12:20
153	237	2	28									267	3	29.08.2019	15:25

References

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

Bjørnstad, O.N., 2020. Ncf: Spatial Covariance Functions. R package version 1.2-9]. CRAN. https://CRAN.R-project.org/package=ncf.

Bjørnstad, O.N., Falck, W., 2001. Nonparametric spatial covariance functions: estimation and testing. Environ. Ecol. Stat. 8 (1), 53–70. https://doi.org/10.1023/A:1009601932481.

Bjørnstad, O.N., Ims, R.A., Lambin, X., 1999. Spatial population dynamics: analyzing patterns and processes of population synchrony. Trends Ecol. Evol. 14 (11), 427–432. https://doi.org/10.1023/A:1009601932481.

Brandt, M., Tucker, C.J., Kariryaa, A., Rasmussen, K., Abel, C., Small, J., Chave, J., Rasmussen, L.V., Hiernaux, P., Diouf, A.A., et al., 2020. An unexpectedly large count of trees in the West African Sahara and Sahel. Nature 587 (7832), 78–82. https://doi.org/10.1038/s41586-020-2824-5.

Brenning, A., 2012. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: the r package sperrorest. In: 2012 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 5372–5375. https://doi.org/10.1109/IGARSS.2012.6352393.

Brodrick, P.G., Davies, A.B., Asner, G.P., 2019. Uncovering ecological patterns with convolutional neural networks. Trends Ecol. Evol. 20, 1–12. https://doi.org/10.1016/j.tree.2019.03.006.

Brus, D.J., 2021. Statistical approaches for spatial sample survey: persistent misconceptions and new developments. Eur. J. Soil Sci. 72 (2), 686–703. https://doi.org/10.1111/ejss.12988.

Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S., 2021a. Review on Convolutional Neural Networks(CNN) in vegetation remote sensing. ISPRS J. Photogrammetry Remote Sens. 173, 24–49. https://doi.org/10.1016/j.isprsjprs.2020.12.010. November 2020 cit. on pp. 2, 3, 6, 7, 13, 15.

Colomina, I., Molina, P., 2014. Unmanned aerial systems for photogrammetry and remote sensing: a review. ISPRS J. Photogrammetry Remote Sens. 92, 79–97. https://doi.org/10.1016/j.isprsjprs.2014.02.013.

De Gruijter, J., Minasny, B., McBratney, A., 2015. Optimizing stratification and allocation for design-based estimation of spatial means using predictions with error. J.Surv.Stat. Methodol. 3 (1), 19–42. https://doi.org/10.1093/jssam/smu024.

Dormann, C.F., 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Global Ecol. Biogeogr. 16 (2), 129–138. https://doi.org/10.1111/j.1466-8238.2006.00279.x.

Ferreira, M.P., Lotte, R.G., D'Elia, F.V., Stamatopoulos, C., Kim, D.-H., Benjamin, A.R., 2021. Accurate mapping of Brazil nut trees (*Bertholletia excelsa*) in Amazonian forests us- ing Worldview-3 satellite images and convolutional neural networks. Ecol. Inf., 101302 https://doi.org/10.1016/j.ecoinf.2021.101302.

Fournier, Q., Aloise, D., 2019. Empirical comparison between autoencoders and traditional dimensionality reduction methods. In: 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). IEEE, pp. 211–214 doi: 10.1109/AIKE.2019.00044.

Frey, J., Kovach, K., Stemmler, S., Koch, B., 2018. UAV photogrammetry of forests as a vulnerable process. A sensitivity analysis for a structure from motion RGB-image pipeline. Rem. Sens. 10 (6) https://doi.org/10.3390/rs10060912.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. http://www.deeplearningbook.org.

Janet, J.P., Duan, C., Yang, T., Nandy, A., Kulik, H.J., 2019. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. Chem. Sci. 10 (34), 7913–7922. https://doi.org/10.1039/C9SC02298H.

Junttila, S., Holttä, T., Lindfors, L., El Issaoui, A., Vastaranta, M., Hyyppä, H., Puttonen, E., 2021. Why Trees Sleep? - Explanations to Diurnal Branch Movement. Research Square preprint doi: 10.21203/rs.3.rs-365866/v1.

Kattenborn, T., Eichel, J., Fassnacht, F.E., 2019. Convolutional Neural Networks enable efficient, accurate and fine-grained segmentation of plant species and communities from highresolution UAV imagery. Sci. Rep. 9 (1), 7. https://doi.org/10.1038/s41598-019-53797-9.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint 9. https://doi.org/10.48550/arXiv.1312.6114.

Kingma, D.P., Welling, M., 2019. An introduction to variational autoencoders. arXiv preprint 9. https://doi.org/10.48550/arXiv.1906.02691.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., Bretagnolle, V., 2014. Spatial leave-oneout cross-validation for variable selection in the presence of spatial autocorrelation. Global Ecol. Biogeogr. 23, 811–820. https://doi.org/10.1111/geb.12161.

Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74 (6), 1659–1673. https://doi.org/10.2307/1939924.

Lopatin, J., Dolos, K., Kattenborn, T., Fassnacht, F.E., 2019. How canopy shadow affects invasive plant species classification in high spatial resolution remote sensing. Remote Sensing in Ecology and Conservation 5 (4), 302–317. https://doi.org/10.1002/rse2.109.

Mahecha, M.D., Rzanny, M., Kraemer, G., Mader, P., Seeland, M., Wäldchen, J., 2021. Crowd-sourced plant occurrence data provide a reliable description of macroecological gradients. Ecography 44 (8), 1131–1142. https://doi.org/10.1111/ecog.05492.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12 (9), 1620–1633. https://doi.org/10.1111/2041-210X.13650.

Meyer, H., Reudenbach, C., 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. Software 101, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001.

Mila, C., Mateu, J., Pebesma, E., Meyer, H., 2022. Nearest neighbour distance matching leave-' one-out cross-validation for map validation, 00 Methods Ecol. Evol. 1–13. https://doi.org/10.1111/2041-210X.13851.

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

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. Inf. Sci. 31 (10), 2001–2019. https://doi.org/10.1080/13658816.2017.1346255.

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., Carin, L., 2016. Variational autoencoder for deep learning of images, labels and captions. Adv. Neural Inf. Process. Syst. 29, 2352–2360. https://doi.org/10.48550/arXiv.1609.08976.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz- Monfort, J.J., Schroder, B., Thuiller, W., et al., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40 (8), 913–929. https://doi.org/10.1111/ecog.02881.

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. Rem. Sens. 10 (8), 1263. https://doi.org/10.3390/rs10081263.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention 234–241. https://doi.org/10.1007/978-3-319-24574-4 28.

Schiefer, F., Kattenborn, T., Frick, A., Frey, J., Schall, P., Koch, B., Schmidtlein, S., 2020. Mapping forest tree species in high resolution UAV-based RGB-imagery by means of convolutional neural networks. ISPRS J. Photogrammetry Remote Sens. 170 (January), 205–215. https://doi.org/10.1016/j.isprsjprs.2020.10.015.

Schiefer, F., Schmidtlein, S., Kattenborn, T., 2021. The retrieval of plant functional traits from canopy spectra through rtm-inversions and statistical models are both critically affected by plant phenology. Ecol. Indicat. 121, 107062 https://doi.org/10.1016/j.ecolind.2020.107062.

Schiller, C., Schmidtlein, S., Boonman, C., Moreno-Martínez, A., Kattenborn, T., 2021. Deep learning and citizen science enable automated plant trait predictions from photographs. Sci. Rep. 11 (1), 1–12. https://doi.org/10.1038/s41598-021-95616-0.

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.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. Journal of big data 6 (1), 1–48. https://doi.org/10.1186/s40537-019-0197-0.

Stehman, S.V., Pengra, B.W., Horton, J.A., Wellington, D.F., 2021. Validation of the us geological survey's land change monitoring, assessment and projection (lcmap) collection 1.0 annual land cover products 1985–2017. Remote Sens. Environ. 265, 112646 https://doi.org/10.1016/j.rse.2021.112646.

Storch, I., Penner, J., Asbeck, T., Basile, M., Bauhus, J., Braunisch, V., Dormann, C.F., Frey, J., Gartner, S., Hanewinkel, M., et al., 2020. Evaluating the effectiveness of retention forestry tö enhance biodiversity in production forests of central europe using an interdisciplinary, multiscale approach. Ecol. Evol. 10 (3), 1489–1509. https://doi.org/10.1002/ece3.6003.

Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. Economic Geography 46 (Suppl. 1), 234–240. https://doi.org/10.2307/143141.

Tuia, D., Roscher, R., Wegner, J.D., Jacobs, N., Zhu, X., Camps-Valls, G., 2021. Toward a collective agenda on ai for earth science data analysis. IEEE Geoscience and Remote Sensing Magazine 9 (2), 88–104. https://doi.org/10.1109/MGRS.2020.3043504.

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, 357798. https://doi.org/10.1111/2041-210X.13107.

Veloz, S.D., 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presenceonly niche models. J. Biogeogr. 36 (12), 2290–2299. https://doi.org/10.1111/j.1365-2699.2009.02174.x.

Wadoux, A.M.-C., Heuvelink, G.B., De Bruin, S., Brus, D.J., 2021. Spatial cross-validation is not the right way to evaluate map accuracy. Ecol. Model. 457, 109692 https://doi.org/10.1016/j.ecolmodel.2021.109692.

Wagner, F.H., Sanchez, A., Tarabalka, Y., Lotte, R.G., Ferreira, M.P., Aidar, M.P., Gloor, E., Phillips, O.L., Aragao, L.E., 2019. Using the U-net convolutional network to map forest types~ and disturbance in the Atlantic rainforest with very high resolution images. Remote Sensing in Ecology and Conservation 5 (4), 360–375. https://doi.org/10.1002/rse2.111.

Wang, J., Haining, R., Cao, Z., 2010. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. Int. J. Geogr. Inf. Sci. 24 (4), 523–543. https://doi.org/10.1080/13658810902873512.

Wenger, S.J., Olden, J.D., 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods Ecol. Evol. 3 (2), 260–267. https://doi.org/10.1111/j.2041-210X.2011.00170.x.

Wong, S.C., Gatt, A., Stamatescu, V., McDonnell, M.D., 2016. In: Understanding Data Augmentation for Classification: when to Warp? 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, pp. 1–6.

Zhao, S., Song, J., Ermon, S., 2017. Infovae: information maximizing variational autoencoders. arXiv preprint arXiv:1706.02262.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine 5 (4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307.