DEVELOPING TRANSFERABLE SPATIAL PREDICTION MODELS: A CASE STUDY OF SATELLITE BASED LANDCOVER MAPPING

Marvin Ludwig[1,2]*, Jonathan Bahlmann[1,2] , Edzer Pebesma[2], Hanna Meyer[1]

[1] Institute of Landscape Ecology, University of Münster, Münster, Germany - (marvin.ludwig, hanna.meyer, jbahlmann)@uni-muenster.de
[2] Institute for Geoinformatics, University of Münster, Münster, Germany - edzer.pebesma@uni-muenster.de

KEY WORDS: Machine Learning, Transferability, Spatial Prediction, Mapping, Spatial Variable Selection, openEO

ABSTRACT:

The mapping of environmental information based on remote sensing requires a workflow that involves image processing, model training usually based on machine learning, as well as model application and validation. Remote sensing data processing capabilities are nowadays simplified by cloud computing platforms. State of the art machine learning methods for spatial data which involve a reduction of spatial overfitting, handling of extrapolation situations and a spatially explicit error assessment, however, are currently mostly implemented in local computation frameworks. Here we present a workflow that combines the improved processing capabilities of the cloud computation platform openEO with state-of-the-art machine learning model development in R. OpenEO is used for standardized imagery acquisition and preprocessing to provide predictors for model training. To reduce overfitting, predictors which are meaningful for the mapping are identified via spatial variable selection as implemented in R packages. The mapping accuracy is assessed via spatial cross-validation and predictions are limited to the 'Area of Applicability' of the model. The workflow is designed to enhance and assess the spatial transferability of machine learning models which is demonstrated by a case study of a landcover classification based on Sentinel-2 imagery.

1. INTRODUCTION

Machine Learning (ML) models and their associated predictions have become a key component in environmental science to contribute to major contemporary challenges like achieving sustainable development goals (Holloway and Mengersen, 2018) or biodiversity monitoring (Reddy et al., 2021). Especially in the field of remote sensing ML emerged as a indispensable tool for the large-scale mapping of environmental information (e.g. soil properties, (Hengl et al., 2017), species occurrence, (van den Hoogen et al., 2019), or landuse (Venter and Sydenham, 2021)). Most mapping studies follow a similar logic. A ML algorithm learns the statistical relations between the target variable and predictors from the location of available reference data. Once trained and validated, the model is applied to the spatially continuous predictors to map the target variable for the entire area of interest. While the modelling might be straightforward, in practice, the development of ML models from remote sensing data faces several challenges.

Reference data are often heavily clustered in geographic space (e.g. due to opportunistic field-sampling campaigns (Yates et al., 2018)) which bears the risk of training spatially overfitted models and a low ability of the model to make predictions for new areas (i.e. low transferability Meyer et al. (2019); Meyer and Pebesma (2021)). Hence, sub-optimal validation strategies for spatial predictions can lead to incorrect conclusions about the statistical model performance (observed by e.g. Roberts et al. (2017); Meyer et al. (2018); Ploton et al. (2020)) which ultimately leads to incorrect conclusions about the mapping error. Spatial cross-validation strategies are therefore proposed to assess model performances at unknown locations instead of commonly used random cross-validation approaches (Meyer et al., 2018; Ploton et al., 2020; Mila et al., 2022).

While changing the cross-validation strategy allows for a more reliable error assessment, it does not solve the problem of overfitting. Reducing the risk of overfitting usually involves some form of predictor selection (Ying, 2019). Simpler models - i.e. models that utilize fewer predictors to map the target variable - can be assumed to represent more generalized relations. Meyer et al. (2018, 2019) have suggested a spatial variable selection to identify only those predictors that are most useful for a spatial prediction task, usually leading to higher prediction performances when the model is transferred to new locations. This model transfer often requires that predictions are made for environments that are different from those used for model training. Machine learning models, however, can only make reliable predictions for new areas if the values of the predictor variables involved are comparable to those encountered in the training data. To assess the area to which this applies, Meyer and Pebesma (2021) recently suggested a method to compute the "Area of Applicability" (AOA) of prediction models and suggest that this should become common practice for spatial predictive mapping.

A second major challenge is that ML models require large amounts of training data in order to adequately learn (non-linear) relations between the target variable and predictors. Moreover, the model is usually applied to an even larger amount of data for the desired prediction (e.g. Europe wide Sentinel-2 imagery, Venter and Sydenham (2021)). This is especially the case for the development of global maps such as in Ma et al. (2021); Hengl et al. (2017); Moreno-Martinez et al. (2018) or van den Hoogen et al. (2019). Naturally, using ML in the context of remote sensing depends on large amounts of earth observation data which usually requires extensive preprocessings such as atmospheric correction, cloud masking or the computation of a composite. Especially if imagery from multiple time periods or different sensors are required for the mapping, the amount of data and computation resources quickly exceeds what most researchers have available on their local machine.

* Corresponding author

Ongoing improvements in spatial, temporal and spectral resolution of satellite imagery with a nearly global coverage further increases the amount of computation needed.

To approach this problem, cloud computing frameworks (e.g. Google Earth Engine (Amani et al., 2020) or openEO (Schramm et al., 2021)) emerged as a promising possibility to process large scale satellite imagery. In general, the idea of these platforms is to provide products and processes for analysis ready earth observation data without the need of downloading the satellite imagery. As a consequence, mapping studies utilizing cloud computing have seen a rise in popularity leading to attempts of global scale predictions of e.g. soil nematodes (van den Hoogen et al., 2019), plant biomass (Ma et al., 2021) or landcover (Venter and Sydenham, 2021).

Automated cloud computing workflows are in development to increase the accessibility of ML based mapping to users who lack a deep understanding of the underlying methods (van den Hoogen et al., 2021). While cloud computing platforms provide implementations of ML algorithms, the possibilities for adequate model development and validation are currently still limited due to lacking features (e.g. variable selection) that are, however, relevant for spatial mapping. This is not an issue of the platforms itself since they are meant for the processing of earth observation data and not the development of ML models. Consequently, the initial development of the ML model is often done locally (e.g. in van den Hoogen et al., 2019) with an established ML framework in Python (sklearn, Pedregosa et al. (2011)) or R (caret, Kuhn (2008)). This way, the model training and validation can incorporate already established methods to overcome the aforementioned challenges of spatial predictive modelling.

Here, we outline a workflow that utilizes the open-source cloud computing platform openEO for standardized earth observation imagery acquisition in combination with an R based ML model development that is designed to improve the spatial transferability of prediction models. By doing so, we combine the computational advantage of openEO with the possibilities of case specific model development and validation strategies. The usage of the workflow is demonstrated with a case study of a landcover classification of Sentinel-2 imagery. We explicitly show the benefits of spatial variable selection and the need for a transferability assessment with the AOA by applying the model to a different region.

2. METHODS

The idea of the suggested workflow is to utilize the benefits of the cloud computing platform openEO for the processing and acquisition of earth observation data with advanced predictive modelling methods provided by R (Fig. 1). Potential predictor layers are first computed only for areas with available training data. We then use spatial variable selection (Meyer et al., 2018, 2019) in order to find a set of predictors that, in combination, are most suitable to map the target variable beyond the training data locations. Model performances are validated with spatial cross-validation. Only the selected predictors that are regarded as relevant by the spatial variable selection are then acquired via openEO and used for mapping the area of interest. To prevent low quality and invalid model extrapolations, predictions are finally limited to the AOA of the trained model (Meyer and Pebesma, 2021).

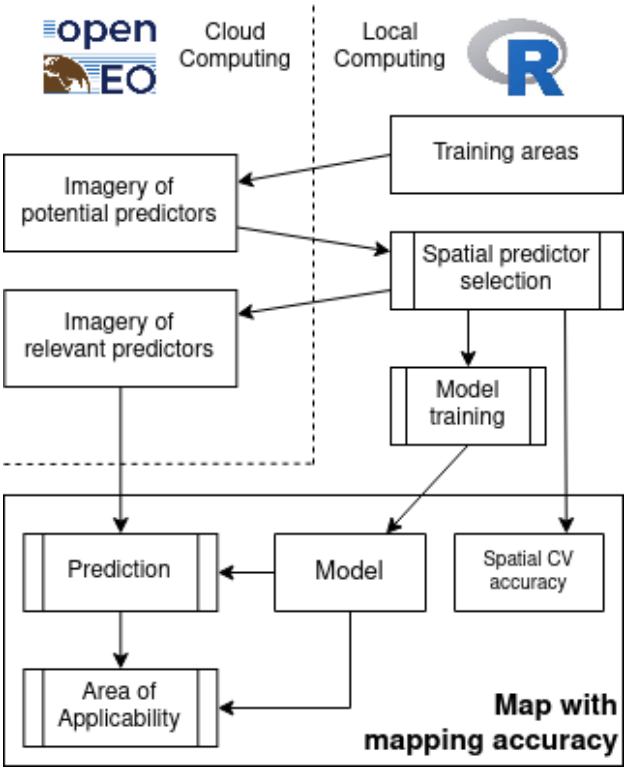

Figure 1. Outline of the Workflow

2.1 Acquisition of predictors with openEO

OpenEO is an emerging cloud computing platform that aims at harmonizing the access to earth observation data from different providers (Schramm et al., 2021). Following an open-source paradigm, openEO enables a community-driven, transparent and reproducible alternative to closed-source alternatives such as Google Earth Engine™. Using the openeo R client (Lahn, 2021) we developed a processing chain for the acquisition of analysis-ready Sentinel-2 L2A composites. Users define the area of interest and a time interval for which the median composite of all available Sentinel-2 scene with less than 20% overall cloud cover is computed. In addition, the Sentinel-2 Scene Classification Layer (SCL) is utilized to mask remaining clouds, shadows and low quality pixels in each time step. The resulting composite may then be used for the calculation of vegetation indices as additional potential predictor variables.

2.2 Modelling

The acquired predictors are then matched with the available reference data that contain the information of the target variable and serve as the training data of the ML model. We use an R-based modelling framework consisting of a spatial variable selection, training of a random forest model, prediction and the computation of the AOA (Fig. 1). However, ML models can be very case specific since the quality of the outcome is heavily dependent on the quality of training data, the used algorithm and its parameters or tuning (Maxwell et al., 2018). Further, each modelling task aims at different target variables, deals with different spatial units and might require different preprocessing steps. Hence, the framework is flexible enough and can easily be modified to the needs of a specific modelling task.

2.2.1 Spatial cross-validation One main challenge when dealing with machine learning models is the prevention of overfitting (Ying, 2019). In the geo-spatial context, this means that

the model has to be generalized enough to make valid predictions for new geographic locations. Hence, the model evaluation also has to account for the situation when the model is applied to locations that are not present in the model training. To do so, we use a spatial cross-validation approach for the assessment of the model error (Meyer et al., 2018; Ploton et al., 2020).

Spatial cross-validation and subsequently the spatial variable selection requires the definition of suitable spatial units used to define cross-validation folds. There is an ongoing discussion on how these spatial units should look like. Here we suggest that users of the workflow should strife for an optimal method for their specific case (e.g. discussed in Mila et al., 2022; Meyer and Pebesma, accepted).

2.2.2 Spatial variable selection and model tuning The risk of overfitting is greatly increased if the model utilizes a high number of predictors (Hassine et al., 2019), since this leads to a high probability that new locations contain combinations of predictor values that are not similar to the training data. Eliminating irrelevant predictors is further advantageous for a models computation time and interpretation (Maxwell et al., 2018).

In the context of spatial mapping, an adequate feature selection limits predictors to those that can be meaningfully used to make predictions for new geographic locations (Meyer et al., 2019; Le Rest et al., 2014). We therefore use spatial cross-validation in conjunction with a forward variable selection approach as described in Meyer et al. (2018). By doing so, we select predictors based on the described cross-validation strategy. Predictors therefore get automatically reduced to those that lead to the highest performance when making predictions for new regions. We assume that this set of predictors should also minimize extrapolation situations since the reduced feature space leads to a more generalized representation of the environment.

The resulting set of spatial predictors is used to train a random forest model (Breiman, 2001) as implemented in ranger (Wright and Ziegler, 2017). We use caret (Kuhn, 2008) for hyperparameter tuning in a grid search approach. Again, spatial cross-validation is used to determine the optimal set of hyperparameters (Schratz et al., 2019) namely mtry, minimum node size and the splitrule. The model internal variable importance is computed by random permutation of each individual predictor and measuring the effect on the model outcome. The tuned model can then finally be applied to new areas to predict the target variable. For this, the aforementioned openEO process is used to obtain the selected predictors for the entire area of interest.

2.3 Area of Applicability

Predictions in novel geographic areas might require model extrapolation if predictor values differ from the training data. This, however, is technically possible but not meaningful for random forests and similar algorithms. To detect these areas, we limit the prediction to the Area of Applicability (AOA) of the model according to Meyer and Pebesma (2021). The AOA is estimated for each pixel by calculating a "dissimilarity index" (DI). The DI of a new location is its Euclidean distance to the nearest training data point in the multidimensional predictor space, with predictors being weighted by their respective importance in the model. The AOA is then derived by applying a threshold on the DI. The threshold is the (outlier-removed) maximum DI of the training data derived from the spatial cross-validation. Hence, a new data point is outside of the AOA if it is more dissimilar in its predictor properties than the dissimilarity observed in the training data set.

2.4 Case study

To demonstrate the proposed workflow we applied it in a typical satellite-based landcover classification (LCC) scenario. We choose the case study of a LCC since landcover is one of the most important drivers for environmental processes and also widely used as a predictor for subsequent modeling. Further, the reasoning and effects of the predictor selection and AOA are depicted very clearly when applied to the use case of landcover mapping. For example, the spectral properties within certain classes might differ (Hermosilla et al., 2022), i.e. a deciduous forest in Germany might look different from a deciduous forest in Italy. Further, certain landcover types might be completely missing in the training set, leading to the relevance of accounting for the AOA, especially when a trained model is transferred to new geographic regions. This also holds true for models that are applied to different time periods. Spectral properties of landcover classes (e.g. deciduous forests) strongly depend on the time of observation and a model trained with imagery from spring might fail if it is faced with spectral data from a summer scene. This further supports the need for a standardized earth observation data acquisition workflow since predictors in the area of interest have to undergo the same processing steps as the predictors used as the training data.

We collected 40 reference polygons for each of the landcover classes agriculture, forest, grassland, roads, settlement and water in the German state North Rhine-Westphalia (Fig. 2). Additionally, we collected 20 reference polygons per class from three geographically distinct areas that are used for independent validation indicating the transferability of the model. One of the new regions is a coastal area on the Eiderstedt peninsula in northern Germany. Here, the additional class "sand" was sampled which was not present in the training set.

For the training areas, we acquired the Sentinel-2 median composite from all scenes between 2021-04-01 and 2021-10-01 (Bands 2, 3, 4, 5, 6, 7, 8, 8A, 11, and 12, NDVI and EVI) with a spatial resolution of 10m. Bands with a native resolution other than 10m were resampled. As a service provider we use the VITO backend through the openEO Platform early adopters program. For model training and testing, we sampled 2500 pixels per class from the Sentinel-2 composite at the training and test polygon locations respectively.

As spatial cross-validation strategy that was used during hyperparameter tuning and variable selection we applied a 20-fold "leave-polygon-out" cross-validation, where all training polygons were randomly divided into 20 groups. Hence we avoid that reference pixel in the training and test sets are in spatial proximity since they stem from different polygons.

We trained a random forest model that utilizes all 12 acquired Sentinel-2 predictors for the landcover mapping and trained a second model using spatial variable selection. We compared both models in terms of classification accuracy using the independent test data in the three regions (Fig. 2) as well as the transferability of the model using the method to estimate the AOA.

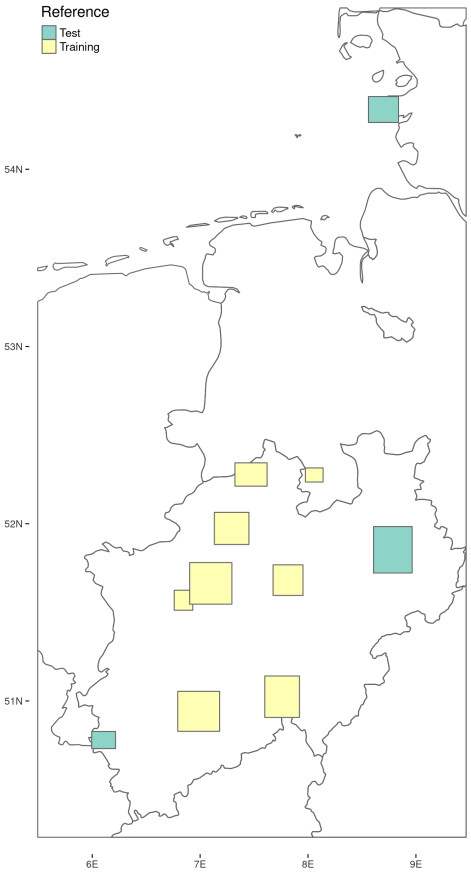

Figure 2. Study Area, The most northern test area is depicted in more detail in Figures 3 and 4

The code to reproduce the presented case study can be accessed at: https://github.com/LOEK-RS/ISPRS2022LCC. The methods for spatial variable selection and AOA estimation are implemented in the R-package CAST (Meyer and Ludwig, 2022).

3. RESULTS AND DISCUSSION

3.1 Effect of Spatial variable selection

The spatial variable selection identified bands 3, 4, 11 and 12, along with NDVI as relevant to predict the landcover classes with a spatial cross-validation accuracy of 0.83. Any further predictor variable could not increase the accuracy. The model that includes all 12 predictors led to nearly the same spatial cross-validation accuracy (0.84, Tab. 2). This marginal difference of the cross-validation accuracy is arbitrary since it stems from the randomness during model training. Thus, the reduction of predictors did not harm the model's ability to accurately fit held-back training data. Further, the prediction outcome of both models was identical for 96% of all classified pixels. Hence, there is no benefit of including more predictors than needed in the model. We observed slight confusions between the classes grassland, agriculture and settlement, (Tab. 1), which is expected as these classes share similar spectral features. Arguably, the within-class spectral properties of these classes is also diverse which makes it more difficult for the random forest model to define characteristic split rules.

Table 1. Cross-validated confusion matrix of the simplified model as the average percentage over the holdout data from the 20 folds (overall accuracy: 0.83).

	agri.	forest	grass	roads	sett.	water
agri.	13.59	0.37	2.40	0.65	0.71	0.05
forest	0.52	15.81	0.65	0.03	0.12	0.26
grass	1.06	0.23	13.34	0.18	0.25	0.02
roads	0.37	0.00	0.03	11.79	2.45	0.20
sett.	1.12	0.14	0.21	3.84	13.03	0.33
water	0.00	0.11	0.03	0.17	0.11	15.80

The validation with independent test data of geographically distinct areas revealed that the variable reduction increased the accuracy of the predictions from 0.71 (all predictors) to 0.77 (reduced predictors). This indicates that the spatial variable selection improved the transferability of the model, since the prediction accuracy in regions without training data increased. As a positive side-effect, the reduced number of variables led not only to a higher prediction accuracy but also to reduced computational requirements since only the selected variables had to be processed in openEO Platform. The smaller predictor space also drastically reduced the computation time of the AOA.

Table 2. Effects of the spatial variable selection on the classification accuracy and AOA. (*) The class "sand" was neglected in these calculations since it was not present in the training data.

	Full model	Simplified model
Predictors	12	5
Spatial cross-validation Accuracy	0.84	0.83
Testdata Accuracy*	0.71	0.77
False classifications masked by AOA*	25%	35%
Correct classifications masked by AOA*	4%	7%
Class "sand" identified as outside AOA	37%	76%

Applying the model to the coastal area resulted in the LCC depicted in Fig. 3. By comparison with the independent test data, we observed slight confusions of the classes agriculture and grassland which was already observed during spatial cross-validation (Tab. 1). The main observations in Fig. 3 however are the large settlement areas predicted near the coastline, which are evidently erroneous classifications of the beach visible in Fig. 4. Since the class "sand" was absent in the training data, the model can never predict these cases correctly. The model accuracy alone (or any model related quality metric) gives no guidance about such cases where the model was not able to learn about a certain class. Hence, the overall model accuracy – even from spatial cross-validation – is not sufficient to represent the mapping accuracy. In this case study, we can identify the incorrect classifications by using the test polygons from the area of interest. Independent and well distributed test data, however, are most likely not available for many mapping projects since the sole purpose of developing a spatial prediction model is its application to areas were no data are available. In a more complicated prediction task than a LCC (which can still be evaluated visually), noticing such prediction errors is challenging. Here, the AOA of the model can serve as a tool to identify possibly erroneous predictions based on distances in the predictor space.

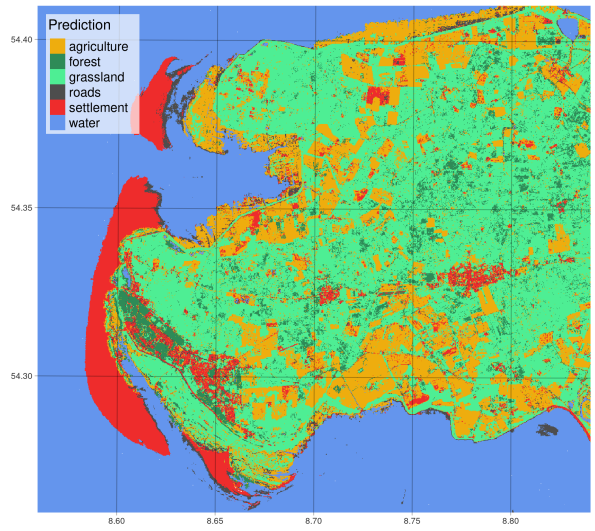

Figure 3. Landcover classification of the coastal area depicted in Fig. 4 as predicted by the simplified model after spatial variable selection

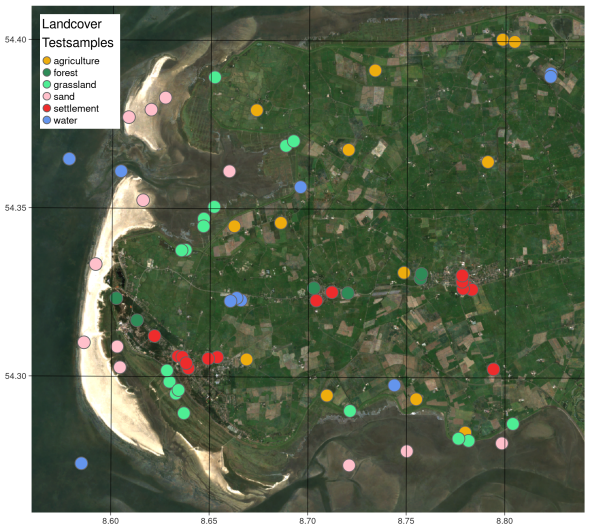

Figure 4. Sentinel-2 true color composite of a coastal area in Germany with the location of test polygons and the associated landcover class. Polygon locations are shown as centroid points for visualization purposes.

3.2 Area of Applicability comparison

The AOA of the model indicates locations where the predictor values are similar enough to the predictor values of observations in the training data. We computed the AOA for both models for the three test regions (compared in Tab. 2). The AOA of the model using all 12 spectral variables allowed avoiding 25% of incorrectly classified pixels (as estimated using the test data). Only a small amount of correctly classified pixels (4%) were outside of the AOA. Reducing the amount of predictors with spatial variable selection allowed avoiding 35% of false classifications by limiting predictions to the AOA. Only marginally more correctly classified pixels were outside the AOA of the simplified model (7%) compared to the full model (4%).

Assessing the transferability of the simplified model in the coastal area shown in Fig. 4 results in the AOA depicted in Fig. 5. Besides some minor patches in the agriculture / grassland areas, the entire coastline is outside the AOA. This can be expected since the coastline consists of spectral properties not present in the training data. From the pixels that are declared as ”sand” in the independent test polygons, the AOA was able to mask off 76%. This is a major improvement compared to the 37% masked off sand pixels from the AOA of the model without spatial feature selection (Tab. 2). The beach areas are not applicable for both models. The AOA of the simplified model also shows ambiguous shallow areas as not applicable that are declared as ”sand” in the training area, but predicted as water by the model (Fig. 5).

This case study shows that defining and visualising the AOA for a spatial prediction model is a useful tool to prevent low quality predictions. The AOA is therefore a crucial part of the spatially explicit mapping error estimation since it can depict where the estimated model performance can be expected to hold because the model was enabled to learn about such environments. The AOA can further give new insights on missing training data and where new sampling are required to adequately represent the entire prediction domain.

4. CONCLUSION

Spatial predictive modelling heavily benefits from novel methods that are so far developed for local use only. The results of the case study e.g. show the benefits of spatial variable selection and consideration of the AOA to increase and assess the transferability of predictive mapping models. We therefore regard the combination of cloud based earth observation data processing and local model development with established frameworks as currently the best compromise to produce high quality spatial prediction models. Our presented open source workflow streamlines the access to satellite based training data for the purpose of model development. In a next step, the locally developed model should be re-implemented or used directly in the cloud computing platform. In openEO Platform, this functionality is currently in development. Besides shareable and reproducible access to homogenized satellite data, the open source aspect of openEO Platform will also enable the implementation of the AOA in cloud environments to further reduce computational costs for users and enhance the spatial mapping workflow overall.

5. ACKNOWLEDGEMENT

The work was supported by the Federal Ministry for Economic Affairs and Climate Action of Germany (project number 50EE2009). We further want to thank the openEO Early Adopters Program for the help and free access to the openEO Platform for cloud computing.

References

Amani, M., Ghorbanian, A., Ahmadi, S. A., Kakooei, M., Moghimi, A., Mirmazloumi, S. M., Moghaddam, S. H. A., Mahdavi, S., Ghahremanloo, M., Parsian, S., Wu, Q., Brisco, B., 2020. Google Earth Engine Cloud Computing Platform for Remote Sensing Big Data Applications: A Comprehensive Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 5326–5350.

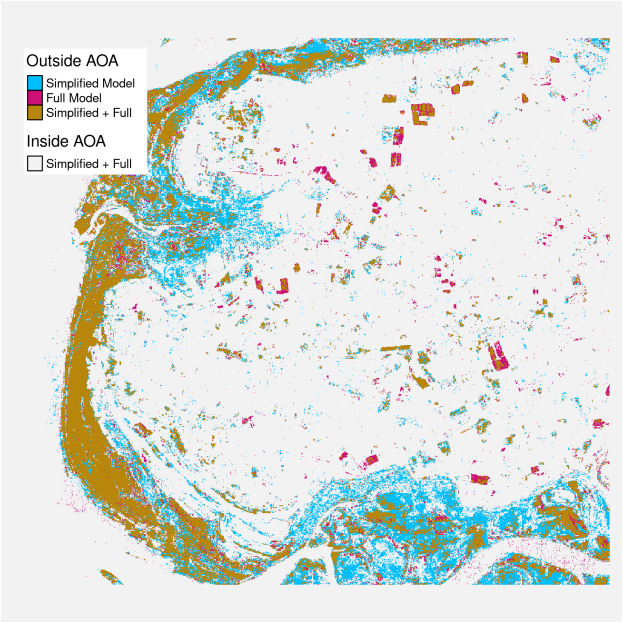

Figure 5. Area of Applicability

Breiman, L., 2001. Random Forests. *Machine Learning*, 45(1), 5–32.

Hassine, K., Erbad, A., Hamila, R., 2019. Important Complexity Reduction of Random Forest in Multi-Classification Problem. *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, IEEE, Tangier, Morocco, 226–231.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global Gridded Soil Information Based on Machine Learning. *PLOS ONE*, 12(2), e0169748.

Hermosilla, T., Wulder, M. A., White, J. C., Coops, N. C., 2022. Land Cover Classification in an Era of Big and Open Data: Optimizing Localized Implementation and Training Data Selection to Improve Mapping Outcomes. *Remote Sensing of Environment*, 268, 112780.

Holloway, J., Mengersen, K., 2018. Statistical Machine Learning Methods and Remote Sensing for Sustainable Development Goals: A Review. *Remote Sensing*, 10(9), 1365.

Kuhn, M., 2008. Building Predictive Models in *R* Using the Caret Package. *Journal of Statistical Software*, 28(5).

Lahn, F., 2021. Openeo: Client Interface for 'openEO' Servers.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., Bretagnolle, V., 2014. Spatial Leave-One-out Cross-Validation for Variable Selection in the Presence of Spatial Autocorrelation. *Global Ecology and Biogeography*, 23(7), 811–820.

Ma, H., Mo, L., Crowther, T. W., Maynard, D. S., van den Hoogen, J., Stocker, B. D., Terrer, C., Zohner, C. M., 2021. The Global Distribution and Environmental Drivers of Aboveground versus Belowground Plant Biomass. *Nature Ecology & Evolution*.

Maxwell, A. E., Warner, T. A., Fang, F., 2018. Implementation of Machine-Learning Classification in Remote Sensing: An Applied Review. *International Journal of Remote Sensing*, 39(9), 2784–2817.

Meyer, H., Ludwig, M., 2022. CAST: 'caret' Applications for Spatial-Temporal Models.

Meyer, H., Pebesma, E., 2021. Predicting into Unknown Space? Estimating the Area of Applicability of Spatial Prediction Models. *Methods in Ecology and Evolution*, 2041-210X.13650.

Meyer, H., Pebesma, E., accepted. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature*.

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. *Environmental Modelling & Software*, 101, 1–9.

Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of Spatial Predictor Variable Selection in Machine Learning Applications – Moving from Data Reproduction to Spatial Prediction. *Ecological Modelling*, 411, 108815.

Mila, C., Mateu, J., Pebesma, E., Meyer, H., 2022. Nearset Distance Matching Cross-Validation for Spatial Prediction. *Methods in Ecology and Evolution*.

Moreno-Martinez, A., Camps-Valls, G., Kattge, J., Robinson, N., Reichstein, M., van Bodegom, P., Kramer, K., Cornelissen, J. H. C., Reich, P., Bahn, M., Niinemets, U., Peñuelas, J., Craine, J., Cerabolini, B. E. L., Minden, V., Laughlin, D. C., Sack, L., Allred, B., Baraloto, C., Byun, C., Soudzilovskaia, N. A., Running, S. W., 2018. A Methodology to Derive Global Maps of Leaf Traits Using Remote Sensing and Climate Data. *Remote Sensing of Environment*, 218, 69–88.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial Validation Reveals Poor Predictive Performance of Large-Scale Ecological Mapping Models. *Nature Communications*, 11(1), 4540.

Reddy, C. S., Kurian, A., Srivastava, G., Singhal, J., Varghese, A. O., Padalia, H., Ayyappan, N., Rajashekar, G., Jha, C. S., Rao, P. V. N., 2021. Remote Sensing Enabled Essential Biodiversity Variables for Biodiversity Assessment and Monitoring: Technological Advancement and Potentials. *Biodiversity and Conservation*, 30(1), 1–14.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., Dormann, C. F., 2017. Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography*, 40(8), 913–929.

Schramm, M., Pebesma, E., Milenković, M., Foresta, L., Dries, J., Jacob, A., Wagner, W., Mohr, M., Neteler, M., Kadunc, M., Miksa, T., Kempeneers, P., Verbesselt, J., Gößwein, B., Navacchi, C., Lippens, S., Reiche, J., 2021. The openEO API–Harmonising the Use of Earth Observation Cloud Services Using Virtual Data Cube Functionalities. *Remote Sensing*, 13(6), 1125.

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. *Ecological Modelling*, 406, 109–120.

van den Hoogen, J., Geisen, S., Routh, D., Ferris, H., Traunspurger, W., Wardle, D. A., de Goede, R. G. M., Adams, B. J., Ahmad, W., Andriuzzi, W. S., Bardgett, R. D., Bonkowski, M., Campos-Herrera, R., Cares, J. E., Caruso, T., de Brito Caixeta, L., Chen, X., Costa, S. R., Creamer, R., Mauro da Cunha Castro, J., Dam, M., Djigal, D., Escuer, M., Griffiths, B. S., Gutiérrez, C., Hohberg, K., Kalinkina, D., Kardol, P., Kergunteuil, A., Korthals, G., Krashevska, V., Kudrin, A. A., Li, Q., Liang, W., Magilton, M., Marais, M., Martín, J. A. R., Matveeva, E., Mayad, E. H., Mulder, C., Mullin, P., Neilson, R., Nguyen, T. A. D., Nielsen, U. N., Okada, H., Rius, J. E. P., Pan, K., Peneva, V., Pellissier, L., Carlos Pereira da Silva, J., Pitteloud, C., Powers, T. O., Powers, K., Quist, C. W., Rasmann, S., Moreno, S. S., Scheu, S., Setälä, H., Sushchuk, A., Tiunov, A. V., Trap, J., van der Putten, W., Vestergård, M., Villenave, C., Waeyenberge, L., Wall, D. H., Wilschut, R., Wright, D. G., Yang, J.-i., Crowther, T. W., 2019. Soil Nematode Abundance and Functional Group Composition at a Global Scale. *Nature*, 572(7768), 194–198.

van den Hoogen, J., Robmann, N., Routh, D., Lauber, T., van Tiel, N., Danylo, O., Crowther, T. W., 2021. A geospatial mapping pipeline for ecologists. Preprint, Ecology.

Venter, Z. S., Sydenham, M. A. K., 2021. Continental-Scale Land Cover Mapping at 10 m Resolution Over Europe (ELC10). *Remote Sensing*, 13(12), 2301.

Wright, M. N., Ziegler, A., 2017. Ranger : A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. *Journal of Statistical Software*, 77(1).

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., Fielding, A. H., Bamford, A. J., Ban, S., Barbosa, A. M., Dormann, C. F., Elith, J., Embling, C. B., Ervin, G. N., Fisher, R., Gould, S., Graf, R. F., Gregr, E. J., Halpin, P. N., Heikkinen, R. K., Heinänen, S., Jones, A. R., Krishnakumar, P. K., Lauria, V., Lozano-Montes, H., Mannocci, L., Mellin, C., Mesgaran, M. B., Moreno-Amat, E., Mormede, S., Novaczek, E., Oppel, S., Ortuño Crespo, G., Peterson, A. T., Rapacciuolo, G., Roberts, J. J., Ross, R. E., Scales, K. L., Schoeman, D., Snelgrove, P., Sundblad, G., Thuiller, W., Torres, L. G., Verbruggen, H., Wang, L., Wenger, S., Whittingham, M. J., Zharikov, Y., Zurell, D., Sequeira, A. M., 2018. Outstanding Challenges in the Transferability of Ecological Models. *Trends in Ecology & Evolution*, 33(10), 790–802.

Ying, X., 2019. An Overview of Overfitting and Its Solutions. *Journal of Physics: Conference Series*, 1168, 022022.