Remote Sensing of Environment 259 (2021) 112409

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Global land characterisation using land cover fractions at 100 m resolution

Dainius Masiliūnas [a,*], Nandin-Erdene Tsendbazar [a], Martin Herold [a], Myroslava Lesiv [b], Marcel Buchhorn [c], Jan Verbesselt [a]

[a] Wageningen University & Research, Laboratory of Geo-Information Science and Remote Sensing, Droevendaalsesteeg 3, 6708 PB Wageningen, the Netherlands
[b] International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, A-2361 Laxenburg, Austria
[c] Flemish Institute for Technological Research (VITO), Boeretang 200, BE-2400 Mol, Belgium

ARTICLE INFO

Keywords:
PROBA-V
Global land cover mapping
Land cover fraction mapping
Time series analysis
Machine learning
Random forest
Cubist
Support vector regression
Neural network
Spatial accuracy
Zero inflation

ABSTRACT

Currently most global land cover maps are produced with discrete classes, which express the dominant land cover class in each pixel, or a combination of several classes at a predetermined ratio. In contrast, land cover fraction mapping enables expressing the proportion of each pure class in each pixel, which increases precision and reduces legend complexity. To map land cover fractions, regression rather than classification algorithms are needed, and multiple approaches are available for this task.

A major challenge for land cover fraction mapping models is data sparsity. Land cover fraction data is by its nature zero-inflated due to how common the 0% fraction is. As regression favours the mean, 0% and 100% fractions are difficult for regression models to predict accurately. We proposed a new solution by combining three models: a binary model determines whether a pixel is pure; if so, it is processed using a classification model; otherwise with a regression model.

We compared multiple regression algorithms and implemented our proposed three-step model on the algorithm with the lowest RMSE. We further evaluated the spatial and per-class accuracy of the model and demonstrated a wall-to-wall prediction of seven land cover fractions over the globe. The models were trained on over 138,000 points and validated on a separate dataset of over 20,000 points, provided by the CGLS-LC100 project. Both datasets are global and aligned with the PROBA-V 100 m UTM grid.

Results showed that the random forest regression model reached the lowest RMSE of 17.3%. Lowest MAE (7.9%) and highest overall accuracy (72% ± 2%) was achieved using random forest with our proposed three-model approach and median vote.

This research proves that machine learning algorithms can be applied globally to map a wide variety of land cover fractions. Fraction mapping expresses land cover more precisely, and empowers users to create their own discrete maps using user-defined thresholds and rules, which enables customising the result for a diverse range of uses. The three-step approach is useful for addressing the zero-inflation issue and mapping 0% and 100% fractions more accurately, and thus has already been taken up in the operational production of global land cover fraction layers within the CGLS-LC100 project. Furthermore, this study contributes to the accuracy assessment of land cover fraction maps both thematically and spatially, and these methods could be taken up by future land cover fraction mapping efforts.

1. Introduction

Land cover, as one of the key variables for monitoring a number of Sustainable Development Goals (SDGs), has lately received more attention due to increased availability of higher spatial and temporal resolution satellite data. In this context, the capacity for land monitoring has increased, and new global land cover maps have emerged to better map the current land cover of the world, as well as to track land cover change. Some of the recent achievements have been the ESA Climate Change Initiative Land Cover (LC-CCI) product (ESA, 2017) that provides a long-term set of consistent global annual medium-resolution land cover maps aimed at the climate community, Copernicus Global Land

* Corresponding author.
E-mail addresses: dainius.masiliunas@wur.nl (D. Masiliūnas), nandin.tsendbazar@wur.nl (N.-E. Tsendbazar), martin.herold@wur.nl (M. Herold), lesiv@iiasa.ac.at (M. Lesiv), marcel.buchhorn@vito.be (M. Buchhorn), jan.verbesselt@wur.nl (J. Verbesselt).

https://doi.org/10.1016/j.rse.2021.112409
Received 6 August 2020; Received in revised form 19 February 2021; Accepted 19 March 2021
Available online 1 April 2021

Services 100 m Land Cover (CGLS-LC100) product (Buchhorn et al., 2019a, 2020) that provides a finer spatial resolution and higher quality with yearly updates since 2015, and the Finer Resolution Observation and Monitoring Global Land Cover (FROM-GLC10) product (Gong et al., 2019) that showcases the potential of land cover mapping at 10 m resolution.

Except for the cover fraction layers of the CGLS-LC100 product, all other global land cover products that include major land cover classes, such as the ones described by Bartholomé and Belward (2005); Friedl et al. (2010); Arino et al. (2007); See et al. (2015); Chen et al. (2015), are provided with discrete classes (also known as "hard" or "crisp" classification), where each pixel of the map can only represent a single land cover class. Such discrete classification oversimplifies reality, as mixed pixels that are covered by multiple land cover classes are a common occurrence. This issue is exacerbated at coarse resolutions and over heterogeneous landscapes. It may result in biases, for instance, a sparse forest may be classified as grassland, ignoring the relatively few trees in the area, and thus underestimate tree cover in the pixel. These systematic errors then add up when scaling the result to the entire region.

A potential solution to this issue is to characterise land cover using cover fractions. In this approach, instead of a single discrete class, the proportion of every class in the legend is reported for every pixel of the map. That way, the land cover models work not on pixel labels, such as "forest", but on land cover characteristics, such as tree cover, defined as the area of the pixel covered by tree canopies, or herbaceous cover, defined as the area not covered by woody vegetation. This is also called "fuzzy" or "soft" classification, and sometimes "subpixel" mapping or "linear mixture modelling" (Okeke and Karnieli, 2006).

Land cover fraction mapping has been attempted in the past. Most of the previous research has focused on deriving land cover fractions of 3–6 classes at a local scale (Adams et al., 1995; Foody, 1996; Walton, 2008; Hansen et al., 2011; Sharma et al., 2011; Uma Shankar et al., 2011; Dwivedi et al., 2012; Lizarazo, 2012; Gessner et al., 2013; Okujeni et al., 2018), less often at a regional scale and with more detailed classes (Colditz et al., 2011). The methods for assessing the accuracy of the results vary greatly between the different studies. In addition, global land cover fraction products have emerged, but focused on a particular class, such as tree cover (Hansen et al., 2003, 2013; Townshend, 2017), water (Schroeder et al., 2015; Pekel et al., 2016) and urban area (Corbane et al., 2019; Gao and O'Neill, 2020; Gong et al., 2020). To date, only CGLS-LC100 (Buchhorn et al., 2019b) provides global maps with fractions of every major land cover class (Tsendbazar et al., 2019).

Land cover fraction mapping can be performed using a variety of different approaches and algorithms. In its core, it is a regression rather than a classification problem, as the output is a fraction of a label rather than a label itself. Methods that have been tested in previous studies include fuzzy nearest centroid regression (Zhang and Foody, 2001), spectral mixture analysis (SMA) (e.g. Shimabukuro and Smith, 1991; Adams et al., 1995; Hobbs, 2003; Yang et al., 2012), random forest (RF) regression (e.g. Walton, 2008), support vector machine (SVM) regression (Walton, 2008), Cubist regression (Walton, 2008), multi-layer perceptron (MLP) neural networks (NNs) (Zhang and Foody, 2001), genetic algorithms (Stavrakoudis et al., 2011) and wavelet transformation (Uma Shankar et al., 2011). The previous studies have only used or compared a few of these methods at once, and never with a thematically complete set of land cover classes nor at global scale.

A common issue with the use of land cover fraction data as input into regression models is data imbalance. The more classes are mapped, the more likely it is that one or more classes are not present in a given pixel (have a 0% fraction), leading to zero inflation. This is especially the case in homogeneous areas, where we can find not only an inflation of 0% fractions, but also an inflation in 100% fractions. This data imbalance leaves little training data in the middle for the regression models to learn from. Conversely, regression models tend to favour predictions closer to the mean and rarely predict extreme values. In this study, we propose the use of a hierarchical multi-step modelling approach to better predict these extreme values.

Model accuracy assessment by itself is often challenging, especially at a global scale. It requires a comprehensive dataset across the globe that would be comparable with the training data, and yet independent of it. Accuracy assessment of land cover fractions is even more challenging, as it requires a dataset that either provides fraction information, or fine spatial resolution data from which it can be calculated. Because of these challenges, the information about the accuracy of the existing global land cover fraction products is often limited, which makes it difficult for the users to decide whether a given product suits their needs. Users would also benefit from knowing the spatial variation of accuracy, as models may be more accurate at certain locations of the world and less so at others.

In this study, after investigating and comparing a variety of methods for global land cover fraction mapping, we proposed an approach to enhance land cover fraction mapping by dealing with the inherent data imbalance issue. We assessed the accuracy of the result both thematically and spatially, by calculating model accuracy spatially across the globe. Lastly, we investigated the effect of input features on each mapped class fraction. Therefore the primary objectives of our study were to:

1. Investigate approaches for reducing bias in the model predictions with regards to zero inflation and predictions tending towards the mean.
2. Assess the accuracy of the models from a thematic and spatial point of view, comparing it to existing global land cover products.

2. Data and methods

2.1. Reference data

The reference data (for model training and validation) used in this study was collected as part of the CGLS-LC100 project (Buchhorn et al., 2020; Tsendbazar et al., 2019). The data includes over 150,000 training points and over 21,000 validation points across the globe, describing the fractions of 12 classes in the year 2015. The classification scheme follows the UN Land Cover Classification Scheme (LCCS) (Buchhorn et al., 2020). However, due to the limited number of observations for some rare classes, we merged them to get a total of seven: bare land (including snow and ice), cropland (including shifting cultivation), herbaceous vegetation (including wetland, lichen and moss), shrubs, trees, built-up and inland water. The "unknown" class was discarded: points with the dominant land cover class marked as unknown were not used. For points with a minority fraction of unknown, the remaining classes were linearly rescaled to add up to 100%. Thus in the end the training dataset size became 138,164 and the validation dataset size became 20,705. See Fig. 1 for the spatial distribution of the points.

These global datasets were generated by performing high-resolution satellite imagery interpretation by a team of experts, using the GeoWiki platform. Each sampled point corresponded to a single 100 m by 100 m pixel of the PROBA-V 100 m UTM grid (Buchhorn et al., 2020). The area of each of these sampled pixels was subdivided into 100 subpixels at 10 m by 10 m spatial resolution. The subpixels were labelled by the experts, and then converted into land cover fraction estimates by calculating the proportion of subpixels that each land cover class covers in the pixel. The satellite imagery that the experts interpreted corresponded to the year 2015.

The training set was generated by a team at the International Institute for Applied SystemsAnalysis (IIASA), whereas the validation dataset was generated by a team at Wageningen University& Research (WUR). The validation dataset was first developed over Africa (Tsendbazar et al., 2018), and was later expanded to cover the whole world (Buchhorn et al., 2020). The class definitions and tools used to collect the data was equivalent for both datasets, but performed independently by a separate group of regional experts to ensure independence of the data, and using a different sampling method. The validation dataset was generated

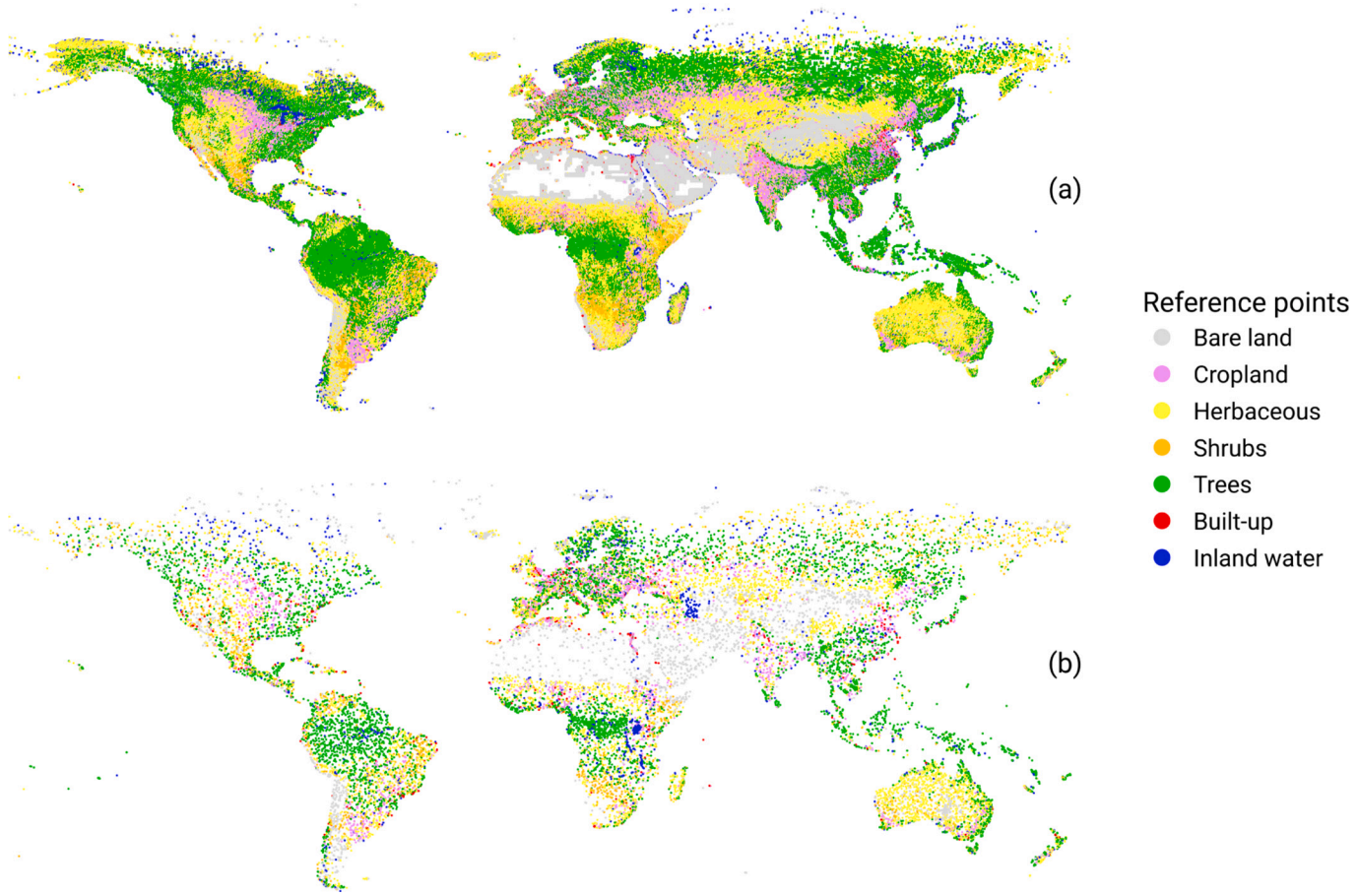

Fig. 1. Sample points representing 100 × 100 m areas at which land cover reference data was used in the study. The colours represent the dominant land cover class at each point. (a): training dataset, collected by IIASA, 138164 points used in this study. (b): validation dataset, collected by WUR, 20705 points used in this study. Both datasets were collected as part of the CGLS-LC100 project (Buchhorn et al., 2020).

separately using stratified random sampling and follows the CEOS Land Product Validation guidelines, which focus on independent and statistically rigorous accuracy assessment. In contrast, the training dataset uses a mix of systematic sampling and additional random sampling in areas that exhibit higher heterogeneity, so as to provide the algorithms with enough data to train in all areas of the world.

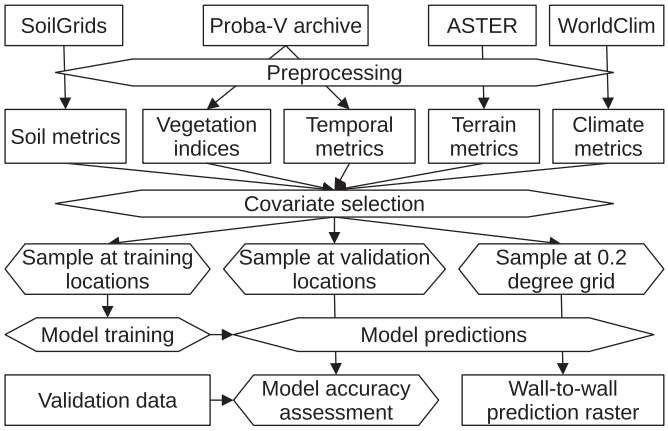

Fig. 2. Processing workflow, from the raw input data to model accuracy assessment and wall-to-wall map output.

2.2. Model training features

See Fig. 2 for an overview of the whole processing chain used in this study. The processing was carried out in R (R Core Team, 2019) and the resulting code has been made openly available in (Masiliūnas, 2020).

To train the models and predict land cover fractions in unsampled locations, six groups of features were used: vegetation indices, temporal metrics, terrain metrics, soil metrics, climate metrics and location data (see Appendix A, Table A.4). These features had to be preprocessed before they could be input into the models.

We chose to use the entire archive (2014-03-11 to 2019-07-16) of the PROBA-V 100 m Level 3 Top-of-Canopy 5-day composite product (Dierckx et al., 2014; Wolters et al., 2016) for this study. The PROBA-V archive provides a relatively long history of frequent (daily or 2-day) observations, which is beneficial for time series analysis, as there are more observations of the land surface in cloudy areas, and a dense time series of observations can be acquired to generate robust temporal metrics. While the reference data corresponded to the land cover at the year 2015 specifically, we chose to use the whole time series of PROBA-V data to obtain more robust statistics for the temporal metrics. The long time series makes the temporal outlier filtering step more reliable, increases the robustness of the fitted harmonic model, and removes the effects of interannual variability and seasonality in the calculated descriptive statistics, as described in the following sections. Land cover change is a relatively rare phenomenon, according to tests performed in the making of CGLS-LC100 collection 3, and so we expected that the difference caused by land cover change to the input data would be smaller than the margin of error of the model output. Nevertheless, this

may result in poorer model performance compared to if up-to-date reference data for each year would be available.

We masked out the clouds from the time series of each of the four PROBA-V spectral bands, first by applying the status mask provided with the product itself, and then by running a temporal cloud filter to remove the remaining outliers that were further than 2 standard deviations away from the locally estimated scatterplot smoothing (LOESS) curve fitted over the blue reflectance band. We then used the resulting cloud-free time series to generate the following vegetation indices (VIs): Normalised Difference Vegetation Index (NDVI), Normalised Difference Moisture Index (NDMI), Enhanced Vegetation Index (EVI), Optimised Soil-Adjusted Vegetation Index (OSAVI) and near infra-red (NIR) of vegetation (NIR_v). These VIs are commonly used in land cover mapping to aid in discerning the vegetation classes among each other. Next, we used the VI time series to calculate the descriptive statistics (median and interquartile range) over the whole time series, as well as for each phenological season separately. We included the resulting metrics as vegetation index features. Next, we ran harmonic analysis (Jakubauskas et al., 2001) on NDVI in order to decompose the time series into sine and cosine components for two frequency orders (annual and semiannual), as well as the trend and intercept of the model. The temporal metrics from the harmonic model quantify the seasonality of the area and allow differentiating between vegetation with different seasonality, such as crops. From this harmonic analysis we obtained the minimum and maximum values of the NDVI time series, which we included as additional vegetation index features. We also obtained the the trend and intercept components from the harmonic model. Lastly, we calculated the phase and amplitude for the two harmonic orders from the respective sine and cosine components of the model. We used these sine, cosine, trend and intercept components, as well as the phase and amplitude of the two harmonic orders, as temporal features in our study. For a list of features that were ultimately used as input to the models, see Appendix A, Table A.4.

To generate elevation features, we obtained the ASTER GDEM v003 (NASA et al., 2019) product (30 m) and resampled it to the PROBA-V 100 m grid. We used the result directly as the terrain elevation feature. In addition, we used the Geospatial Data Abstraction Library (GDAL) (GDAL/OGR contributors, 2020) to calculate terrain parameters out of elevation: slope, aspect and Terrain Position Index (TPI).

We chose the WorldClim 2.0 30 s product (Fick and Hijmans, 2017) as a source of climate features. It includes monthly temperature, precipitation, solar radiation, wind speed and water vapour pressure data. In addition to these features, we calculated 19 bioclimatic parameters from the data, using the dismo package (Hijmans et al., 2017). We also calculated some additional biophysical parameters, namely all of the climate variables during the coldest, warmest, driest and wettest months of the year at each location, as well as the yearly averages of the climate variables.

We used SoilGrids (Hengl et al., 2017) to obtain features related to soils. SoilGrids is based on a random forest model that predicts soil properties at various soil depths globally at a 250 m resolution. In the creation of SoilGrids, a land cover map (based on MODIS) had been used, and so, in order to avoid circular inference, the features that are significantly influenced by the land cover map as detailed in Hengl et al. (2017) were excluded. The soil taxonomy features were also excluded, since they are categorical derivatives from the numerical soil property data and thus do not contribute to land cover fraction prediction.

Lastly, we also included the latitude, longitude and absolute latitude of the reference points as location features when training the models, so that the models could learn spatial patterns.

2.3. Feature selection

In total, we generated 313 features in preprocessing. However, many of them were collinear with one another. Multicollinearity prolongs training time for machine learning models and leads to unreliable coefficient estimation and increased error variance for linear models. Thus, we employed variable selection to remove collinear features before predicting the land cover fractions. Features that had a Pearson's correlation coefficient r (Pearson, 1895) above 0.9 were excluded in an iterative process. After that, features with a Spearman's rank correlation ρ (Spearman, 1904) above 0.9 were likewise excluded.

We manually selected the features to exclude, to avoid interpretation difficulties that arise from automatic selection procedures. The majority of the collinear features were soil metrics predicted at different depths. Therefore, we left in the 10 cm depth features, and excluded the other depths, as long as r was above 0.9. Similarly, climate data was collinear between subsequent months. Thus January and July data was preferred, as these months represent different extremes of the year and are less correlated with data of the other months. Our initial correlation analysis also showed that the spectral bands of PROBA-V were highly correlated with each other as well as to VIs, so we only used VIs as features.

After the feature selection process, 67 features remained. These features include data from each of the feature categories. See Appendix A, Table A.4 for an overview of all of the features that remained and thus used in model training and prediction.

2.4. Land cover fraction mapping methods

We compared a wide array of machine learning regression methods for deriving land cover fractions. The tested methods can be broadly divided into three types: linear models, machine learning models based on classification and regression trees (CARTs), and machine learning models not based on CART. See Table 1 for the full list of methods that we compared in this study. In addition, as a baseline we also compared the results with a trivial equal proportion model (all fractions always predicted to be equal, namely $\frac{100\%}{7} \approx 14.29\%$). Any useful model has to perform better than the equal proportion model, and by comparison to it, it is possible to tell how much better a model performs in estimating the fraction of each class. We tuned each algorithm to select optimal parameters, and postprocessed the output of each algorithm as necessary to ensure that all land cover fractions in each pixel add up to 100%. Namely, if the model output for any class was outside of the 0–100% range, the values were clamped to that range, and if the values did not add up to 100%, they were linearly rescaled so that they would. All of

Table 1
List of regression methods for land cover fraction estimation tested in this study.

Category	Name	Reference	R package and authors
Linear models	Fuzzy nearest centroid (FNC)	Keller et al., 1985	GSIF (Hengl et al., 2004)
	General linear regression model (GLM)	Neter et al., 1996	stats (R Core Team, 2019)
	Partial least squares (PLS) regression	Wold et al., 2001	pls (Mevik et al., 2016)
	Lasso regression	Tibshirani, 1996	glmnet (Friedman et al., 2010)
	Multinomial logistic regression (MLR)	Theil, 1969	nnet (Venables and Ripley, 2002)
Machine learning models based on decision trees	Random forest (RF) regression	Breiman, 2001	ranger (Wright and Ziegler, 2017)
	Cubist regression	Quinlan, 1992	Cubist (Kuhn and Quinlan, 2020)
Other machine learning models	MLP neural networks (NNs)	Dreyfus, 1990	keras (Allaire and Chollet, 2018)
	Support vector machine (SVM) regression	Suykens and Vandewalle, 1999	liquidSVM (Steinwart and Thomann, 2017)
Ensemble learning	Super Learner	Laan et al., 2007	sl3 (Coyle et al., 2020)

the model building and data analysis was performed using the free and open-source statistical software R (R Core Team, 2019). The results showed that RF regression performed the best, with the lowest RMSE value of 17.3%. For more technical details and in-depth results of the regression model accuracy comparison, see Appendix B.

2.5. Multi-step approach to account for zero inflation

As RF regression performed the best by RMSE in the regression model comparison, we tested whether it could be improved further by attempting to solve the training dataset balance issue, namely the high frequency of 0% and 100% fractions. As the dataset describes fractions of each land cover class at each point, most of the locations consist of a mix of only a few classes, and the fraction of the rest of the classes is zero at that location. If the pixel is pure, then one land cover class will be 100% and the rest 0%, which is also a common case. This leads to the dataset getting dominated by zeroes. In that case, minimising the objective function of a machine learning model leads to prioritising predicting 0% fractions well, and ignoring the prediction errors in the middle of the 0–100% range. This is not desirable for users of land cover fraction data, as the added value of fraction information is the information about the middle of the range; otherwise, discrete classification would be just as good. Conversely, 0% by itself is rarely predicted precisely, because when the value is uncertain, predictions tend towards the mean. Therefore we tried several approaches to deal with data imbalance by employing a hierarchical combination of machine learning models.

We compared three approaches using RF models: (a): a single regression model trained on all data. (b): a two-model approach in two steps. Step 1: a binary classification model for each class to predict zeroes, trained on a generated dataset that, based on the land cover fraction values, had labels “zero” and “non-zero”. Step 2: a model to predict non-zeroes, trained on all of the non-zero fraction values. For the combined prediction, all points that were predicted as “zero” in step 1 were set to 0%, otherwise the value from the model in step 2 was used. (c): a three-model approach using three steps. Step 1: a binary classification model to predict pixel purity (i.e. whether we face a classification or a regression problem), trained on a generated dataset that had labels “pure” for points that had a single land cover fraction above a purity threshold, e.g. 95%, and “non-pure” for points that do not. Step 2: a model to perform regression on mixed pixels (as determined in step 1). Step 3: a model to perform classification on pure pixels (as determined in step 1), resulting in a prediction of 100% fraction of the predicted discrete class and 0% fractions for all other classes. The combined prediction is the combination of the results of steps 2 and 3, as determined by step 1.

For the three-step approach, we also tested the effect of the fraction threshold for when we consider a pixel “pure”. The lower the threshold, the more pixels are considered pure and the more often the classification model will be selected, as opposed to the regression model. We also evaluated the accuracy metrics of the separate steps of the multi-step models.

Lastly, we compared the results of our proposed multi-step approach with an approach that uses the median for ensembling tree votes, instead of the mean. The median vote leads to predicting the extreme fractions of 0% and 100% more often, since if the majority of the decision trees vote for one of the two extremes, it gets selected as the output value. Finally, we also investigated the combination of both approaches.

2.6. Accuracy assessment

To assess the performance of the models, we used a number of statistical measures. We started with the statistics of assessing land cover fraction model accuracy that are the most commonly used in this field: root mean squared error (RMSE), mean absolute error (MAE) and mean error (ME). MAE represents the average difference between the predicted and the reference land cover fractions. In our case, its unit is percentage points. RMSE squares the errors, therefore giving a larger penalty for large errors, and thus is always higher than MAE. These statistics are relatively straightforward to calculate and interpret. In the case of land cover fraction mapping, RMSE is very sensitive to errors where a pixel is entirely mapped as a different class (i.e. 100% instead of 0%). MAE is more lenient and not as influenced by a small number of such large misclassifications. Thus it is more indicative of the overall model accuracy, whereas RMSE is more indicative of the presence of large errors.

We calculated RMSE, MAE and ME, both separately per class, and also pooled overall. These overall measures were calculated by taking the mean of all class points pooled together, rather than taking a mean of the per-class means. In addition, we calculated the relative root mean squared error (RRMSE), relative mean absolute error (RMAE) and relative mean error (RME) for each class by dividing the absolute measures by the mean fraction of each class. The relative statistics give an extra penalty for poor predictions of rare land cover fractions (i.e. those that are absent from most pixels), to account for the issue that a prediction of constant 0% would lead to low RMSE and MAE for rare class fractions.

Next, we estimated the goodness of fit of the models by calculating the coefficient of determination R^2 of the models in two ways. The first way is Nash–Sutcliffe model efficiency cofficient (NSE) (Nash and Sutcliffe, 1970), which is equivalent to an R^2 of a linear regression model whose intercept and slope terms are predetermined and are equal to 0 and 1, respectively. This metric shows how far away the predicted values are from the 1:1 line with the reference values. The value range of NSE is $(-\infty, 1]$, where a value of 0 means that the predicted values are no better than predicting the mean value. The second way is to calculate the R^2 of an ordinary least squares (OLS) regression that estimates the intercept and slope, rather than presetting it. This method always gives higher R^2 values than NSE, as it allows the regression line to have more flexibility. The two ways of estimating the coefficient of determination are herewith represented as R_2^{NSE} and R_2^{OLS}, respectively.

Furthermore, we calculated a subpixel confusion-uncertainty matrix (SCM) (Silván-Cárdenas and Wang, 2008), and the metrics derived from it: overall accuracy (OA), as well as producer accuracy (PA) and user accuracy (UA) per class. The SCM is an adaptation of the confusion matrix concept to fractional data. We used the MIN-PROD composite operator as recommended by Silván-Cárdenas and Wang (2008). When using this operator, the diagonal of the matrix expresses the maximum overlap (agreement) of the target and predicted class fractions, and the off-diagonal is an expression of which classes the non-overlapping fractions should belong to by calculating the expected value of overlap (product of reference and predicted class fraction). For instance, a pixel of 60% grass and 40% shrub, when predicted as 40% grass and 60% shrub, would have $min(60\%, 40\%) = 40\%$ in the diagonals and $\frac{20\% \times 20\%}{20\%} = 20\%$ in the off-diagonals. For cases where the allocation of the overestimated and underestimated parts of the fraction do not have a unique solution, the SCM indicates the expected value and an associated uncertainty measure.

Moreover, to show the variation in predicted land cover fraction values depending on the magnitude of the fractions, we produced boxplots showing the distribution of the predictions. The boxplots are binned for each 10% of the predictions. The 1:1 line indicates how well do the distributions of the predictions and the reference data match. We also created additional plots showing how RMSE, MAE and ME change over these bins.

To show how the model accuracy varies in space, we produced spatial residuals, i.e. the overestimation and underestimation of each class fraction for each validation point. We rasterised the result into global maps for ease of view. If multiple points fell into the same raster cell, we reported the mean value.

To put our results in a wider context, we compared them with the

currently available global products for specific land cover classes. For this comparison, we used the same validation dataset as for our models. Since the validation dataset reflects land cover of the year 2015, we chose to compare our results with four products produced for or around the year 2015. Therefore we chose the Global Forest Cover Change (GFCC) forest cover product for 2015 (Townshend, 2017), the Global Surface Water (GSW) water occurrence history product for 2015 (Pekel et al., 2016), and two products for the comparison with the built-up class: Global Human Settlement Layer (GHSL) built-up for 2014 (Corbane et al., 2019) and FROM-GLC10 impervious surface change layer for 2015 (Gong et al., 2020). As the latter two products show the change in built-up cover over time, we reclassified them to have a value of 100 if the area was built-up in 2014 or 2015, respectively. For FROM-GLC10, we assumed that areas that had been covered by impervious surface at any time prior to 2015 continued to be covered by impervious surface in 2015. GSW data was also reclassified to 100 if the water was permanently present and 50 if it was seasonally present. Next, for validation purposes we resampled each product to the PROBA-V 100 m grid using the bilinear resampling method, so that the cover fractions get aggregated over the same areas as our own data. Then we compared the product values with our validation data point values.

Lastly, we investigated how features affect model accuracy. See Appendix C for more information on feature importance.

3. Results

3.1. Zero inflation adjustment with multi-step models and median voting

Regression model comparison showed that RF regression achieved the highest accuracy: by RMSE when using a single model and a mean vote (RMSE: 17.3%, MAE: 9.4%), and by MAE when using a median vote (RMSE: 20.7%, MAE: 7.9%). Therefore we chose RF regression to test our proposed multi-step approach. For more details about the results of the accuracy comparison between the tested regression models, see Appendix B.

The overall accuracy statistics of all of the RF regression models that we tested, including two-step, three-step, as well as median voting approaches and their combinations, can be seen in Table 2. Per-class results can be seen in Fig. 3 and the relative statistics in Fig. 4.

The median voting approach resulted in significantly more predictions of 0% and 100% class fractions, therefore the output looked closer to a discrete classification map compared to mean voting. Since these two most common values were predicted precisely much more often, MAE reduced. However, in cases of high uncertainty, the median vote was also much more likely to predict a 100% fraction compared to the mean vote, and thus was also more likely to predict 100% of the wrong class. This resulted in increased RMSE. Therefore the overall effect of using the median vote is a change in the balance of RMSE and MAE.

We observed a similar effect when using the multi-step approach for an RF regression that uses the mean statistic for ensembling the tree votes. Using a two-step approach decreased the MAE of the model, while increasing RMSE. This is because the false positives in the first step of the two-step model (the binary zero/non-zero classification) result in some high fractions nevertheless being predicted as zero. The two-step approach combined with binary relevance also had a drawback: in some cases all land cover class fractions were predicted to be 0% in the first step, making it impossible to determine the land cover fractions. Therefore, to make them sum up to 100%, we set these cases to equal proportions (100% / 7), which introduced further error.

The three-step approach solved this issue, since the first step predicts purity, rather than zeroes. In this approach, pure pixels are passed to RF classification, which always predicts the most likely pure class. When the three-step approach was applied to mean vote RF, the result was a slight decrease in RMSE across most classes compared to the two-step approach. However, there was also an increase in MAE of the predicted crop cover. Therefore the three-step model offsets the increase in RMSE as seen in the two-step model case, and does not cause a high increase in MAE, thus leading to a better balance between the two measures.

The median vote had a stronger effect on decreasing MAE (and increasing RMSE) compared to the multi-step approach with mean voting. When both approaches were combined, no further decrease in MAE could be achieved, however, the combination of the three-step model and median voting decreased RMSE compared to the single-step median model, therefore reducing large errors.

We also investigated the separate steps of the multi-step model in more detail to better understand the accuracy of each model. One parameter in the three-step model is the purity threshold: how high should the cover fraction be for a pixel to be considered “pure” and be subject to classification rather than regression. Decreasing the purity threshold led to a result closer to the one-step model, i.e. reduced RMSE and increased MAE, as the classification model was used less often. The purity binary classifier, when only 100% cover is considered “pure”, achieved 78% overall accuracy, and the accuracy decreased when the purity threshold was decreased. The classification model achieved 87% overall accuracy, but shrub and built-up classes had very low users’ accuracy. These classes are highly heterogeneous, therefore there were too few observations to train the classifier to identify these classes. Decreasing the purity threshold resulted in a lower overall accuracy, but an increase in the users’ accuracy in the classifier model for these particular classes. The regression step by itself had lower accuracy than the combined three-step model, with 21.38% RMSE and 10.49% MAE (using median voting), as the middle of the range is the most difficult to predict correctly. Decreasing the purity threshold led to a lower RMSE, as the value range of the regressor training data gets decreased, but a higher MAE, as the amount of training data for the regressor decreased. Overall for the whole multi-step model, lowering the purity threshold increased RMSE and slightly increased MAE as well.

All in all, both median voting and the multi-step approach successfully result in more correctly predicted 0% and 100% fractions, thus lowering MAE and increasing the SCM OA. While it also increases RMSE, combining the two concepts together leads to a lower increase in RMSE.

3.2. Spatial predictions and accuracy

We selected the three-step median vote RF model for further analysis, as it represents a good balance between the RMSE and MAE

Table 2
Accuracy statistics of multi-step models. Best performing statistics are highlighted. “slope” refers to the OLS-estimated slope, “int” refers to the OLS-estimated intercept.

Model	RMSE (%)	MAE (%)	R_2^{NSE}	R_2^{OLS} (slope/int)	OA (%)	Kappa
RF regression	**17.3**	9.4	**0.66**	**0.67** (1.09/−1.22)	67 ± 4	0.57 ± 0.05
″two-step	19.9	8.2	0.56	0.60 (0.78/3.48)	71 ± 2	0.62 ± 0.02
″three-step	19.4	8.4	0.58	0.60 (0.84/2.34)	71 ± 3	0.62 ± 0.04
″median vote	20.7	**7.9**	0.52	0.60 (0.74/3.80)	71 ± 1	0.63 ± 0.02
″″two-step	20.0	8.1	0.54	0.60 (0.77/3.67)	**72 ± 1**	0.63 ± 0.02
″″three-step	20.2	**7.9**	0.54	0.60 (0.77/3.34)	**72 ± 2**	**0.64 ± 0.02**

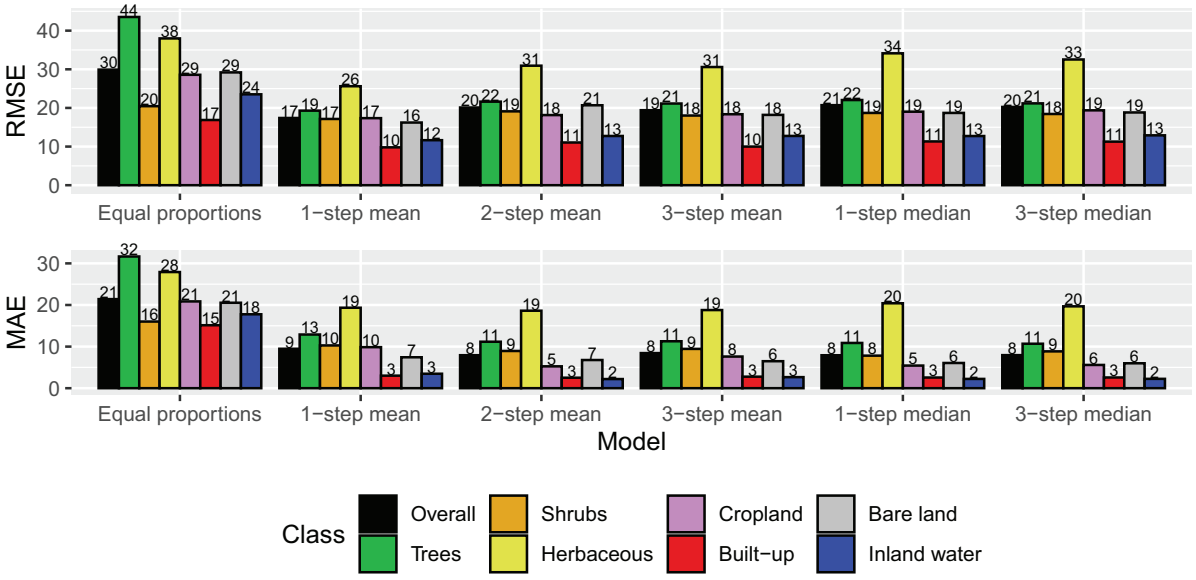

Fig. 3. Comparison of RF regression models (equal proportion model shown as a reference). 1-step models use no adjustment for zero inflation, 2-step models perform classification on zeroes and regression for non-zeroes, 3-step models perform a classification into pure and non-pure pixels, and then a regression or classification based on that. Mean and median are the tree vote summary statistics.

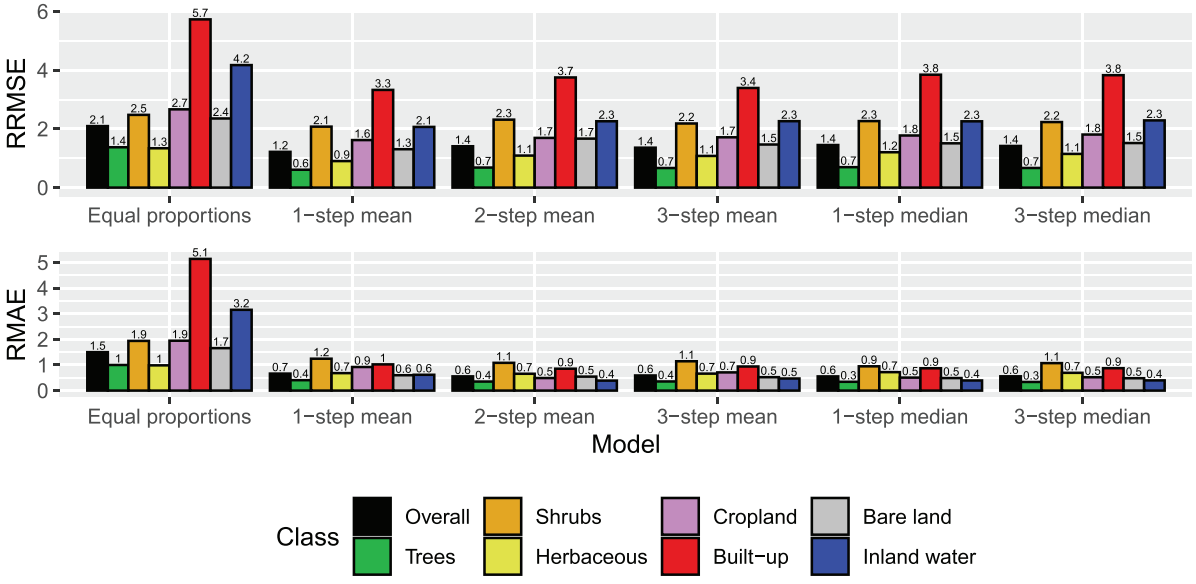

Fig. 4. Comparison of relative accuracy metrics of RF regression models (equal proportion model shown as a reference). 1-step models use no adjustment for zero inflation, 2-step models perform classification on zeroes and regression for non-zeroes, 3-step models perform a classification into pure and non-pure pixels, and then a regression or classification based on that. Mean and median are the tree vote summary statistics.

statistics, and its analysis helps further understand the three-step model. To visually demonstrate the model, we used it to predict land cover fractions at a global scale (100 m resolution, but sampled every 0.2 degrees). See Fig. 5 for a visualisation of all of the fraction layers separately, and the supplementary material for the output GeoTIFF file itself. The wall-to-wall fraction maps reveal how land cover fraction mapping is capable of expressing gradients and mixed land cover. In addition, the spatial accuracy maps that we produced based on the model predictions show the variation in the model accuracy globally, and are also presented in raster format in Fig. 5.

Biotic gradients can be seen in the global patterns of the land cover fractions. For instance, gradients are visible between communities dominated by shrubs and ones dominated by herbaceous vegetation, such as in south and east Africa. Likewise, the gradient of tree cover from 100% in the African tropics to 0% in the sub-Sahara region is evident. The tropical forest edge appears with a hard edge when using median voting and the three-step approach, as forest occurs in discrete patches due to human activity, rather than changing gradually over space. The tree cover in the transition zone towards savannah is much more mixed and gradual. Herbaceous cover in sub-Sahara shows an asymmetric gradient: the cover is highest at around 14–15° N and decreases quickly towards the north, becoming zero around 18° N; but decreases slowly towards the south, reaching all the way to 5° N. Inland water shows up as more discrete, as it naturally forms discrete patches. Mixed pixels that include water are uncommon. Built-up area is also relatively rare worldwide. It rarely forms a 100% fraction, as urban areas tend to

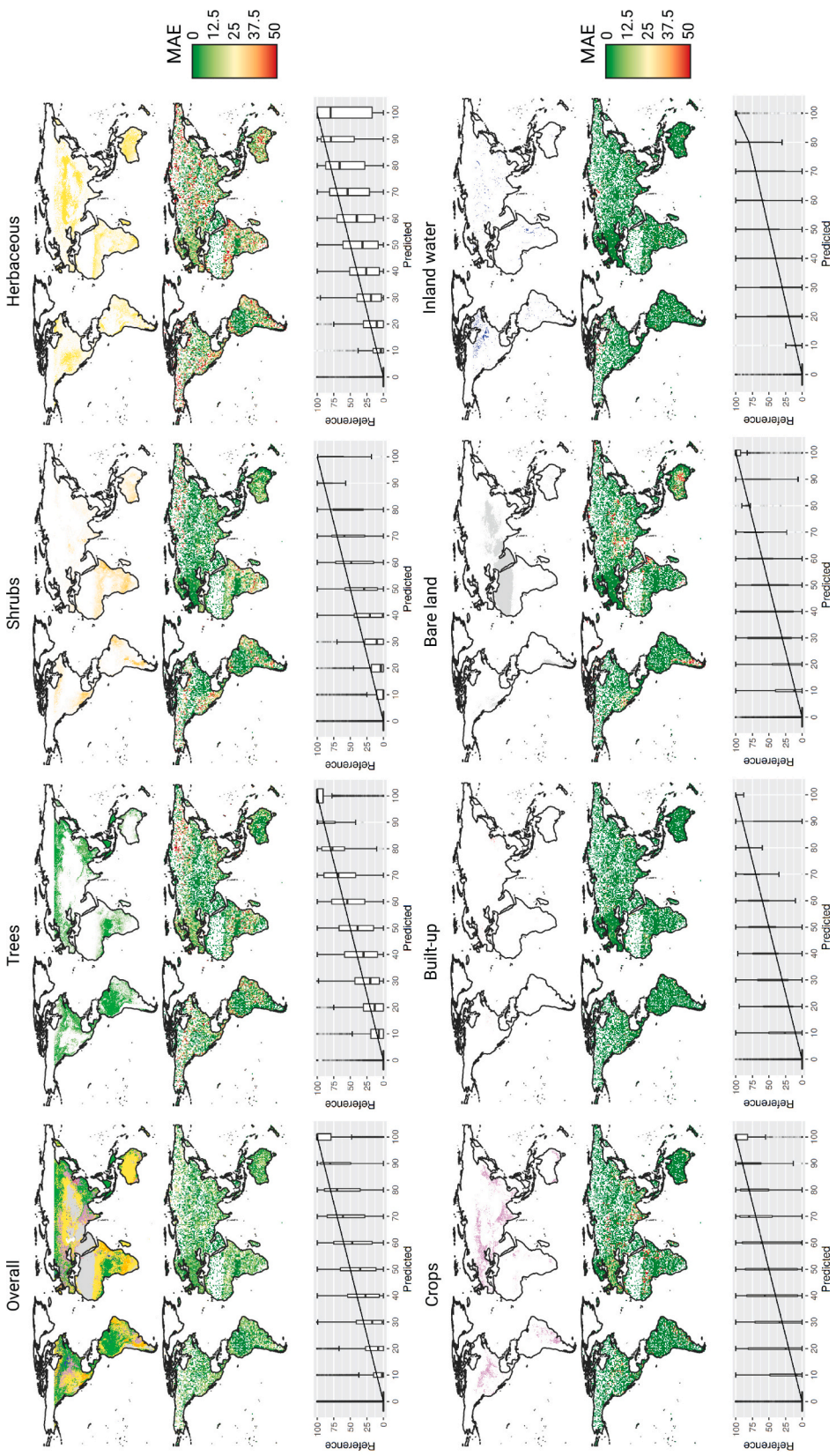

Fig. 5. Random Forest single model predictions per class. Top row: predicted fractions, each given its own colour ("Overall" shows a hardened map of dominant land cover using the same colour scheme). Middle row: absolute errors per class, based on predictions at point locations of validation data, displayed aggregated over 1 by 1 degree cells using the mean function. Bottom row: distribution of predicted versus true values, shown as box plots with bins each 10%, widths representing sample size.

include both built-up area and greenery within the footprint of the 100 m pixel.

The spatial pattern of map accuracy shows that the land cover fractions in areas with pure land cover, such as tropical forests (100% tree cover) and deserts (100% bare land), were predicted with the highest accuracy. Conversely, fractions in areas with mixed land cover were predicted less accurately, as isolating individual fractions from pixel-level information is more challenging. In addition, land cover fractions in the extreme latitudes were predicted less accurately as well. In these areas, less training data was available, owing to the lack of high-resolution imagery there.

As we can see from the box widths of the boxplots in Fig. 5, the distribution of the predictions was relatively even across the whole range for herbaceous vegetation and trees, but uneven for bare land, inland water and built-up area. The number of predictions was much more even across the entire range for herbaceous vegetation compared to shrubs. The model overestimated the fractions of trees and herbaceous cover, as the medians of each box are below the 1:1 line almost throughout the entire range, but it underestimated built-up and crop fractions. The overall ME is below 0.001, which means that overall the model is not biased. For an alternative visualisation showing the change of RMSE, MAE and ME overall and per class over each prediction bin, see Appendix D, Fig. D.9.

Our analysis of the effect of training features on model prediction accuracy showed that features obtained from remote sensing were the most important, but all of the feature categories contributed to improving the model predictions. For more detailed information, see Appendix C.

Lastly, the results of the comparison of the RF three-step median model predictions with existing global products that correspond to a particular land cover class in our classification are given in Table 3. The RF three-step median model performed worse than the GSW water occurrence history product, but better than the GFCC tree cover and GHSL built-up products. There was a tie when comparing the RF three-step median model with the FROM-GLC10 impervious surface product when it comes to MAE, but FROM-GLC10 was slightly more accurate according to RMSE and less biased according to ME.

4. Discussion

4.1. Multi-step approach for dealing with data imbalance

One challenge in land cover fraction prediction is the tendency of the models to favour the mean over the extremes, as that minimises RMSE that may arise from incorrect predictions. However, that leads to increased fuzziness of the result, where pixels with high uncertainty are marked as a mix of many classes, and fractions of 0% are rarely, if ever, predicted. Our proposed multi-step approach adjusts the balance the other way. As it is a combination of one or two classification models and one regression model, it predicts significantly more pure pixels compared to a single regression model.

Therefore, the multi-step approach was successful in reducing MAE (and improving the related SCM metrics), as the particularly common case of 0% fractions was captured better. On the other hand, it comes at a cost of higher RMSE and lower R^2. This is because in highly uncertain, but pure cases, the model makes a best guess of a pure class, and due to the high uncertainty, the prediction is often incorrect. This leads to 100% error in those cases, which is highly penalised by RMSE. Due to this effect, the resulting map is closer to a discrete classification map, with less expressive transition gradients between land cover classes.

A similar effect was seen when using techniques such as median voting for RF regression. Like the three-step model, median voting resulted in more correct 0% fraction predictions, but likewise increased the chances of predicting 100% of the wrong class. Our results showed that median voting has a stronger effect in reducing MAE than the multi-step approach. Therefore it may be a suitable choice in cases where it is computationally infeasible to train three models. However, this approach is only applicable to RF models, whereas the three-step approach is generic and can be used with any combination of models. In addition, when the median voting approach was combined with the three-step approach, the RMSE decreased without affecting MAE. Thus, three-step median RF achieved the best combination that is optimised towards reducing MAE, doing so without increasing RMSE as much as the single-step median vote model does.

These findings show that the use of a multi-step approach depends on what is more important for the user. If an occasional prediction of 100% of the wrong class is acceptable, then a multi-step model provides an overall more accurate result, especially for zero fractions. On the other hand, a single-step approach emphasises the strength of land cover fraction mapping by expressing gradual changes over space better, and avoids large errors. The latter is more likely to be useful for the modelling community that deals with uncertainty with probabilistic frameworks, and the former may be more useful for policymakers and land owners who are more concerned with what land cover is most likely to be present on the ground. In addition, we expect a multi-step approach to be more suitable for fine resolution mapping, where more pure pixels can be expected, and a single-step approach to be more useful for coarse resolution mapping, where mixed pixels are the norm.

It is also worth noting that the multi-step approach is flexible and can be used with any algorithm that provides both classification and regression modes, or with two separate unrelated classification and regression algorithms. Therefore there may be some combinations of models that have not been tested yet, but could achieve even higher accuracy.

4.2. Comparison with global land cover products

To gain insight into how well our proposed multi-step median vote RF model performs, we compared it to existing global land cover fraction products. These products, that only focus on a single land cover class, had varying accuracy compared to our model (see Table 3). Some products, like GSW water occurrence (Pekel et al., 2016), had a higher accuracy. Others, like GFCC forest cover (Townshend, 2017) and GHSL built-up (Corbane et al., 2019), had a lower accuracy. The accuracy of the impervious surface class fraction from FROM-GLC10 mostly matches that of the built-up cover fraction of our proposed model. This shows that our proposed method, coupled with the large training dataset that we used, achieves similar and sometimes even better performance

Table 3
Accuracy comparison between our results and existing global land cover fraction maps. Highest accuracy and lowest bias results are highlighted.

Fraction and source	RMSE (%)	MAE (%)	ME (%)	RRMSE	RMAE	RME	R_2^{NSE}	R_2^{OLS}
Inland water (RF 3-step median)	12.90	2.25	−1.17	2.29	0.40	−0.21	0.65	0.67
GSW, year 2015 (Pekel et al., 2016)	**10.22**	**1.94**	**−0.63**	**1.81**	**0.34**	**−0.11**	**0.78**	**0.78**
Trees (RF 3-step median)	**21.18**	**10.68**	**1.81**	**0.67**	**0.34**	**0.06**	**0.72**	**0.74**
GFCC, epoch 2015 (Townshend, 2017)	28.80	18.33	−12.57	0.91	0.58	−0.40	0.48	0.61
Built-up (RF 3-step median)	11.28	**2.57**	−2.33	3.83	**0.87**	−0.79	0.19	0.22
GHSL built-up, 2014 (Corbane et al., 2018)	18.67	5.43	4.80	6.34	1.84	1.63	−1.23	0.52
FROM-GLC10 impervious surface, 2015 (Gong et al., 2020)	**11.18**	**2.57**	**0.93**	**3.79**	**0.87**	**0.32**	**0.20**	**0.65**

compared to specialised land cover fraction products, but brings the advantage of producing fractions for a variety of land cover classes that sum up to 100%.

The accuracies of other global products reported in literature likewise varied compared to our results. However, these comparisons are much more difficult to make, as the validation methods and scope vary significantly between the studies. For example, our RF three-step median model had a higher RMSE for the tree class (21.1%) than the one reported by Sexton et al. (2013) for their vegetation continuous fields product (16.8%). However, Sexton et al. (2013) validated their data using lidar datasets within several local study areas, rather than using global image interpretation data as we did. The study by Montesano et al. (2009) that used a validation approach closer to ours to validate the MODIS tree cover product, reported an R^2 of 0.57, RMSE of 13.4%, RMSD of 21.3%, slope from a linear regression of 0.5 and intercept of 18.4. In comparison, our results for the three-step median RF model for the tree cover class had an R_2^{OLS} of 0.74, RMSE of 21.1%, slope of 0.87 and intercept of 5.82. Thus while our model appears to perform better, Montesano et al. (2009) only evaluated boreal regions rather than the entire globe, and used MODIS 500 m data rather than PROBA-V 100 m data.

4.3. Challenges and future outlook

Machine learning algorithms pose several challenges that are inherent to how the models are constructed. The trade-off between minimising RMSE and minimising MAE comes from the chosen loss function. Typically, in cases of high uncertainty, the loss function is minimised when the predictions tend towards the mean. In that case, the models predict in areas with a high cover of a fraction, such as mixed shrublands, a lower fraction of the class than expected, whereas in cases with a low cover, such as for fraction of built-up, higher fractions are predicted than expected. This is due to a prediction of the mean being less penalising than predicting the extremes incorrectly; e.g. for a case of 50% shrub cover, predicting 100% shrubs would be a larger mistake (and thus lead to higher RMSE) than predicting 15% shrubs. Likewise, predicting 0% built-up in areas covered by dark bare soils risks a case where it truly would be 100% built-up, so on average predicting 15% built-up in this case lowers the possible error. The three-step approach (or median vote) tilts the balance in the other direction, as e.g. the first step determines that the pixel is pure, and the classification step determines that it is more likely to be bare soil than built-up area. But there may also be cases when the classifier predicts the wrong class. If the pixel is not pure, the regression step still tends to predict towards the mean due to the loss function, so the challenge of underestimating large fractions and overestimating small ones remains. However, with the multi-step approach, it is now possible to influence the decision process of the model to tweak it towards the desired outcome. In addition, using a median vote in tree-based ensemble models makes the model tend towards the median, which is often one of the extreme values. In that case, we see the inverse pattern compared to mean vote: small fractions are underestimated (i.e. predicted as 0%), and large fractions are overestimated (i.e. predicted as 100%). On average, the model is still not biased, as the predicted values are more polarised but balance each other out. The multi-step approach could be improved further by exploring the option of using different machine learning models for each step, and by gathering more features that would increase the accuracy of the models at each of the steps.

Another challenge inherent in land cover mapping is discerning classes that are related, e.g. herbaceous vegetation and shrubs. These classes were particularly difficult to map, in part due to their heterogeneity, and in part due to confusion between herbaceous vegetation, cropland and shrubs. As shown in Appendix C, Fig. C.8, specific types of features can be used to discriminate between these classes better. Specifically, the identification of herbaceous vegetation is primarily based on vegetation indices, identification of shrubs relies mostly on climate data, whereas cropland identification is more data-intensive and makes use of vegetation indices, temporal metrics, climate, soil and location information. These three classes are challenging to discriminate between not just for regression algorithms, but also for expert interpreters, which may lead to higher uncertainties also in the training and validation data for these classes. It is even more challenging to discern between these classes if they mix within the area of a single pixel, which is common in grasslands and shrublands, as well as in smallholder agriculture. The difference between the definition of trees, shrubs and herbaceous vegetation largely comes down to plant height, therefore dynamic information about vegetation height would allow mapping these classes more accurately. However, this would require either photogrammetry techniques (that typically rely on much finer spatial resolution and more overlap between the scenes) to reconstruct vegetation height, or the use of non-optical sensors such as SAR interferometry or lidar data. Another way to differentiate between the vegetation classes could be to make use of hyperspectral data, which allows differentiation between different kinds of vegetation based on e.g. their photosynthesis processes or water content, which affect light absorption. Emerging new high level hyperspectral products, such as sun-induced chlorophyll fluorescence (SIF) or gross primary productivity (GPP), such as ones based on the upcoming FLEX satellite, could allow for a straightforward way to incorporate this extra information into land cover models. Hyperspectral data could be useful for better differentiation of non-vegetated classes as well, such as bare soil from urban, e.g. by separating the spectral signature of sand from concrete or asphalt. In addition, land cover time series information could help track land cover change over time, as the land cover at one time step depends on the land cover at the previous step. This information would allow the regressor to limit the predicted values to a smaller range, and thus reduce the noise in the predictions.

Another challenge is class imbalance. For example, the built-up class rarely forms a 100% fraction. That makes it simple to achieve a high prediction accuracy according to absolute statistics, as a fraction of 0% is in most cases not far off from the true value. However, a prediction of 0% in every pixel makes the fraction map not useful for user needs. This challenge is further exacerbated by the training dataset containing relatively few points in built-up areas to begin with. Therefore, having a more balanced training dataset may further increase the accuracy of the models. However, the issue of value imbalance within the class will always remain for land cover fraction mapping, therefore the multi-step model approach will be relevant, especially if the legend involves even more classes, or if the land cover is more homogeneous at the level of the mapping unit.

Several more challenges are yet to be tackled in this field, but doing so is becoming more and more feasible over time. Finer spatial resolution mapping, such as 10 m mapping using Sentinel-2 data, is a future research direction, where the pixel footprint will more likely cover homogeneous land. Therefore, due to an increase in 0%/100% fractions, such future developments would be more likely to benefit from a multi-step approach or optimisation for MAE. The multi-step approach is fully portable to finer scales, but more research is needed to determine the effect of the different scales on the purity of the pixels, and how much benefit does a combination of classification and regression bring compared to doing only classification or only regression. Finer spatial resolution sensors can also be used for mapping fractions at coarser resolutions in a more precise way, by performing aggregation of the finer resolution pixels to estimate the land cover fraction at a coarser resolution. This is likely to become the norm if even finer resolution data (e.g. 1 m) becomes available globally.

Another benefit of using different optical sensors is higher spectral resolution. As PROBA-V only measures four spectral bands, the amount of information that can be retrieved from them is limited. Sensors such as Sentinel-2 multi-spectral instrument (MSI) have a much wider range of spectral bands that could be used both directly as features, as well as

enable computing a more diverse range of VIs, such as ones based on the slope of the red edge. This could potentially improve the distinction between different land cover classes.

With more availability of such additional remote sensing data, it becomes increasingly more feasible to perform land cover monitoring and change mapping. Mapping land cover fractions is a great opportunity to track gradual changes, such as regrowth, better. The challenge here is that the higher uncertainty about fraction estimates may cause the time series of land cover fractions to fluctuate, making it difficult to determine robust trends.

5. Conclusions

We investigated ways to tackle the issue of accurately predicting the extreme fraction values of 0% and 100% by proposing a hierarchical multi-step approach combining classification and regression models. This approach was applied to an RF regression model, which our tests showed to have the highest accuracy (RMSE: 17.3%) for land cover fraction mapping compared to other regression algorithms. We also combined this approach with RF median voting. The combined RF median three-step approach obtained the best results for MAE (7.9%) and SCM OA (72% $\pm$ 2%). The proposed approach results in predictions of the most likely pure land cover class, when the class is uncertain. This is in contrast to predicting a mix of classes when a standard one-step model is used, and therefore is useful for users who are more interested in the most likely class, rather than class probabilities. Based on this model, we created a demonstration map showing the global distribution of land cover in separate land cover fraction layers. Remote sensing features were the most important for model accuracy, although all other types of features (climate, soil, terrain) also contributed significantly for some classes and thus could not be omitted without negatively affecting the overall model accuracy.

These findings directly contribute to the operationalisation of global land cover fraction mapping by analysing and advancing currently available methods for thematically exhaustive global land cover fraction mapping. Information on land cover fractions offers better precision than discrete land cover maps, and allows the users to manually define thresholds to generate discrete classifications of their own choosing, based on their classes on interest. Furthermore, given the recent advances in optical sensor spatial resolution and the resulting increase in pixel purity, the multi-step model approach may become more important in the future. Lastly, due to the advances in spatial and spectral resolution, longer imagery time series of existing sensors, and the availability of non-optical sensor data, this work paves the way towards operational land cover fraction change mapping, which would allow monitoring gradual land cover change.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2021.112409.

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

We would like to thank VITO and Terrascope for providing us with computational facilities to access and process PROBA-V data, and the team of Steffen Fritz (IIASA) for providing the model training dataset. In addition, we thank Bruno Smets for ideas and support from the CGLS-LC100 project, Devis Tuia for the initial idea for the multi-step approach, Benjamin Kellenberger for help on tuning NN parametrisation, as well as Linlin Li and Tomislav Hengl for suggestions on additional machine learning methods to test. We also thank the three anonymous reviewers for their valuable and detailed comments.

Appendix A. List of features used in regression models

Table A.4
List of features used as input for the models tested in this study, and their data sources.

Category & data source	Covariate
Location (3, intrinsic)	longitude
	latitude
	absolute latitude
Vegetation indices (22, derived from PROBA-V 100 m top-of-canopy reflectance v1.02 (Dierckx et al., 2014))	minimum NDVI
	maximum NDVI
	median NDMI
	NDMI yearly IQR
	NDMI March–May IQR
	NDMI June–August IQR
	NDMI September–November IQR
	NDMI December–February IQR
	OSAVI March–May IQR
	OSAVI June–August IQR
	OSAVI September–November IQR
	OSAVI December–February IQR
	EVI March–May IQR
	EVI June–August IQR
	EVI September–November IQR
	EVI December–February IQR
	median NIRv
	NIRv yearly IQR
	NIRv March–May IQR
	NIRv June–August IQR
	NIRv September–November IQR
	NIRv December–February IQR
Temporal metrics (9, derived from a harmonic model over time series of PROBA-V 100 m top-of-canopy reflectance v1.02 (Dierckx et al., 2014))	NDVI order 1 cosine
	NDVI order 1 sine
	NDVI order 2 cosine
	NDVI order 2 sine

(*continued on next page*)

Table A.4 (*continued*)

Category & data source	Covariate
	NDVI trend coefficient
	NDVI order 1 phase
	NDVI order 1 amplitude
	NDVI order 2 phase
	NDVI order 2 amplitude
Terrain (4, ASTER GDEM V003 (NASA et al., 2019))	elevation
	slope (log-transformed)
	aspect
	terrain position index
Climate (21, WorldClim 2.0 (Fick and Hijmans, 2017))	January precipitation (log)
	April precipitation (log)
	July precipitation (log)
	October precipitation (log)
	January solar irradiance
	July solar irradiance
	mean temperature
	temperature monthly range
	isothermality
	temperature annual range
	annual precipitation (log)
	temperature seasonality
	minimum solar irradiance
	maximum solar irradiance
	mean solar irradiance
	mean windspeed
	mean water vapour pressure
	coldest month precipitation (log)
	warmest month precipitation (log)
	wettest month solar irradiance
	driest month solar irradiance
Soil (8, SoilGrids (Hengl et al., 2017))	soil available water
	soil bulk density
	soil cation exchange capacity (log)
	soil clay fraction
	soil coarse fragments (log)
	soil pH
	soil sand fraction
	soil water wilting point

Appendix B. Regression model comparison

B.1. Regression models tested in this study

Before applying the multi-step approach, we tested the performance of various regression models for land cover fraction mapping. We tested four types of models: linear models, models based on decision trees, machine learning models not based on decision trees, and ensemble learning (for an overview, see Table 1).

First, we chose to compare five types of linear models, to have a baseline for a comparison with the nonlinear machine learning models. The most simple model we selected was the general linear regression model (GLM), also known as multivariate linear regression. It is an extension to the standard linear regression that allows for multiple outcomes. Next, we tested two linear models that include input data regularisation in the model itself: lasso regression and partial least squares (PLS) regression. We also tested multinomial logistic regression (MLR), which is usually used for classification and is fit using land cover class labels rather than fractions, but the output includes probabilities for each class that add up to 100%. We fit multinomial logistic regression (MLR) using the dominant land cover class as a label for the pixel and used the class probabilities as a proxy for land cover fractions. Lastly, we tested fuzzy nearset centroid (FNC) regression, also called fuzzy nearest prototype, fuzzy *c*-means or fuzzy *k*-means. It is a simple regression method, where the land cover fractions in a pixel are determined by the distance of the pixel from the centroids of each class in feature space.

The second group tested in this study was machine learning methods not based on decision trees. Neural networks (NNs) are a promising technique for land cover fraction mapping, as they allow both multiple inputs and multiple outputs, and, using the softmax activation function, also ensures that the result sums up to 100% with no need for additional postprocessing. In this study, after performing tuning, we ended up using a multi-layer perceptron (MLP) with three hidden layers with 128, 64 and 32 neurons per layer respectively. We used the Nadam optimiser (Dozat, 2016) with MAE as the loss function to optimise the NN, and softmax activation for the output. The models were trained using the keras package, built upon TensorFlow that enables the use of a graphics processing unit to accelerate the NN training process. The other method in this category that we tested was Support vector machines (SVMs), which are machine learning models that attempt to find the optimal boundary between the class clusters in feature space by constructing a dividing hyperplane. For land cover fraction classification, we used SVM regression based on Least Squares SVM. As SVM models are univariate, we used the binary relevance method (Karalas et al., 2016): training separate models per class that predict a single class, and then combining the results.

The third group we tested were tree-based machine learning models. Random forest (RF) regression is a popular method for land cover classification that works by building a number of CART decision trees based on random subsets of the input training data, and taking the mean or median of the "votes" of these individual decision trees. RF is univariate, therefore we again used the binary relevance approach. Next, we tested Cubist

regression. It is based on RF regression, but instead of using a threshold of a feature to split the decision tree, Cubist uses a linear regression based on a subset of the data relevant for the split in question. In addition, it features committees, a boosting technique that iteratively trains trees so as to learn from the previously generated ones. Model tuning led to us using 10 committees. Cubist predictions were also made using the binary relevance method.

Lastly, we made an ensemble from the two machine learning models that produced the lowest RMSE and MAE respectively: RF and Cubist regressions. We used the super learner algorithm (Laan et al., 2007) to create a hierarchical ensemble, where the two models were cross-validated using 10-fold cross-validation to obtain relative weights of each model for each land cover class. The predictions of the two models, along with the weights of the models, were then used as input features for another RF regression metalearner. The output from this metalearner is the final prediction of the whole ensemble. As the ensembled methods are univariate, we used the binary relevance method in this case as well.

B.2. Comparison results and discussion

The overall accuracy statistics of the compared models are reported in Table B.5, and per-class statistics in Fig. B.6. The results show that the R^2 statistics, especially R_2^{NSE}, are in agreement with the RMSE statistics, and the overall accuracy (OA) from the subpixel confusion-uncertainty matrix (SCM) is in agreement with the MAE statistics. R^2 is a representation of how closely the predictions correlate with the validation data, therefore big outliers have a large effect on the value. The SCM does not take this into account, as it does a comparison on the overlap of fractions in each pixel, which doesn't apply an extra penalty for large errors.

The baseline models that performed the best were the two tree-based machine learning models: RF regression and Cubist. Both of these models are univariate, therefore predictions per class had to be made using the binary relevance method (one model per class). This shows that the disadvantages of the binary relevance method, namely that each model is only trained on the fractions of its own class and has no knowledge of the fractions of the other classes, and the need for a rescale step to make sure the fractions sum up to 100%, are outweighed by the advantages of these machine learning models and the flexibility of training separate models for each class.

The super learner method that combines both RF regression and Cubist regression resulted in a model that is in between the two ensembled models in terms of accuracy. Its RMSE was below that of Cubist regression, but above that of RF regression. Conversely, the MAE was below that of RF regression, but above that of Cubist regression. This shows that the metalearner (another RF model) could not differentiate between the two ensembled models well enough to select the model that is the most accurate for a given class. Rather, it weighted in both of the models' predictions, losing the advantages of the models when taken separately.

Table B.5
Accuracy statistics of the tested regression models. Best performing statistics are highlighted. "slope" refers to the OLS-estimated slope, "int" refers to the OLS-estimated intercept. "Only RS features" stands for a model trained only with VI and temporal metrics.

Model	RMSE (%)	MAE (%)	R_2^{NSE}	R_2^{OLS} (slope/int)	OA (%)	Kappa
Equal proportions	29.9	21.4	0	–	26 ± 5	0.13 ± 0.07
FNC	24.4	13.5	0.33	0.35 (0.80/2.89)	53 ± 4	0.42 ± 0.06
GLM, PLS, Lasso	21.6	12.7	0.48	0.49 (1.10/−1.42)	56 ± 4	0.43 ± 0.05
MLR	21.6	12.1	0.48	0.48 (0.96/0.62)	58 ± 4	0.46 ± 0.06
MLP NN	22.7	9.2	0.43	0.52 (0.70/4.25)	68 ± 1	0.57 ± 0.02
SVM regression	20.7	8.9	0.52	0.56 (0.79/3.02)	69 ± 2	0.58 ± 0.03
Cubist regression	18.1	**8.1**	0.63	0.65 (0.88/1.77)	**72 ± 2**	**0.63 ± 0.03**
RF regression	**17.3**	9.4	**0.66**	**0.67** (1.09/−1.22)	67 ± 4	0.57 ± 0.05
″only RS features	18.4	10.3	0.62	0.64 (1.09/−1.25)	64 ± 4	0.54 ± 0.05
Cubist + RF ensemble	17.7	8.6	0.65	0.65 (0.95/0.73)	70 ± 3	0.61 ± 0.04

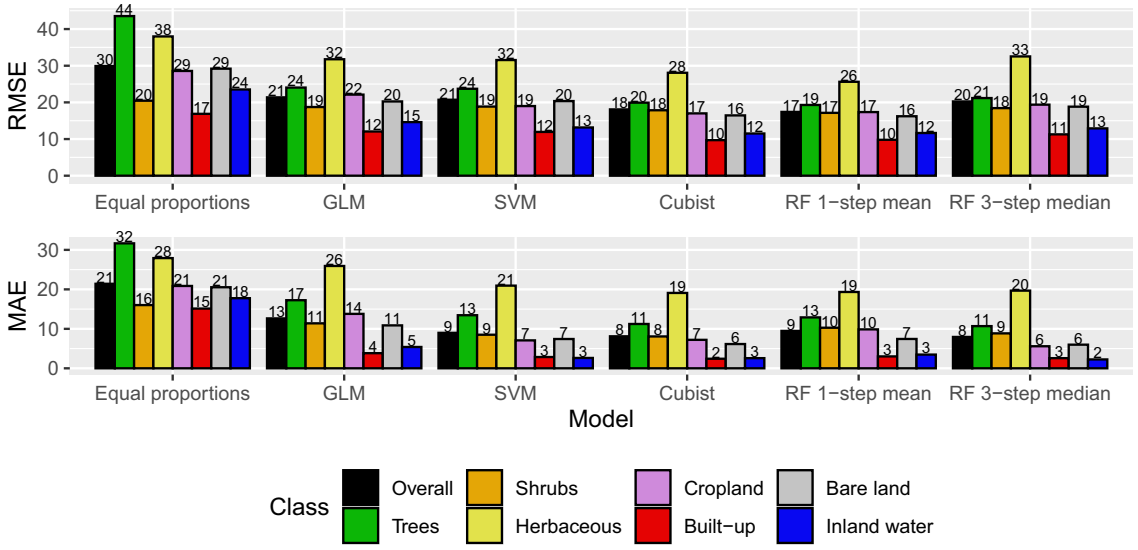

Fig. B.6. Comparison of absolute RMSE (top) and absolute MAE (bottom) per class of the best performing models in their category, and the equal proportions solution as a reference.

The linear statistical models (GLM, PLS, Lasso regression and MLR) performed the worst, as they are limited to a linear approach. The results of all

of the linear models were very similar. There was no added value to PLS and Lasso regressions over the basic GLM. This is likely due to the additional preprocessing step of feature selection, as detailed in section 2.3. Lasso and PLS regressions are extensions to GLM that include a regularisation step, which proved to be unnecessary if all of the highly correlated features are already manually removed.

Fuzzy nearest centroid (FNC) was the model that performed the worst out of the non-trivial models tested. This indicates that the shape of the feature data in feature space is too complex to capture merely with a centroid approach. This is further evidenced by a much better result obtained by the binary relevance SVM model, which works on a similar principle but can capture more complex shapes.

MLP NNs, despite being a multivariate machine learning model, performed worse than all of the tested binary relevance methods.

Nevertheless, it performed better than all of the multivariate linear models. All tested models had lower accuracy when estimating the cover fractions of vegetation classes (herbaceous, trees, crops, shrubs), compared to the non-vegetated land cover (inland water, built-up, bare land), as shown in Fig. B.6. This was exacerbated by the imbalance in class distribution: the built-up class is rare and very rarely forms a majority, therefore a prediction of 0% leads to a perfect prediction most of the times, and even if it does not, the error is low. In contrast, tree cover had a lot more balanced distribution of fractions, including a large number of pure pixels, which leaves no such trivial solution. The RRMSE and RMAE statistics show this (see Fig. B.7): the tree cover prediction has low relative error given its high mean value (32%), whereas the built-up class is the most challenging to predict according to RRMSE given its low mean value (3%). The shrub cover fraction is also very challenging to predict, since it had the highest RMAE and none of the models showed large improvements in RMSE compared to the equal proportions model. There were only moderate improvements in the prediction of the herbaceous vegetation fraction as well. In contrast, most models were a significant improvement in predicting tree, built-up, water and bare cover fractions compared to the equal proportion model, indicating that the features used to train the models were useful to distinguish these classes from the others and to quantify their proportions in each pixel.

To sum up, we tested a number of regression methods for characterising land cover by means of predicting thematically comprehensive land cover fractions at the global scale, and found that RF regression and Cubist regression produce output of the highest accuracy by RMSE and MAE, respectively. Our findings agree with those of Li et al. (2018), who compared Cubist and RF regression for water fraction classification and found that Cubist performs slightly better than RF regression for this particular class. Their Cubist regression result achieved 7.52% MAE and 10.39% RMSE. Our respective results using Cubist regression were 2.57% MAE and 11.51% RMSE. While Cubist works best for the inland water class, when considering all of the classes, RF regression nevertheless results in higher accuracy (see Table B.5). The difference in the reported numbers may be due to the differences in scale (regional vs global) and training data balance. As we focused on predicting multiple classes, our training dataset intrinsically had higher zero inflation.

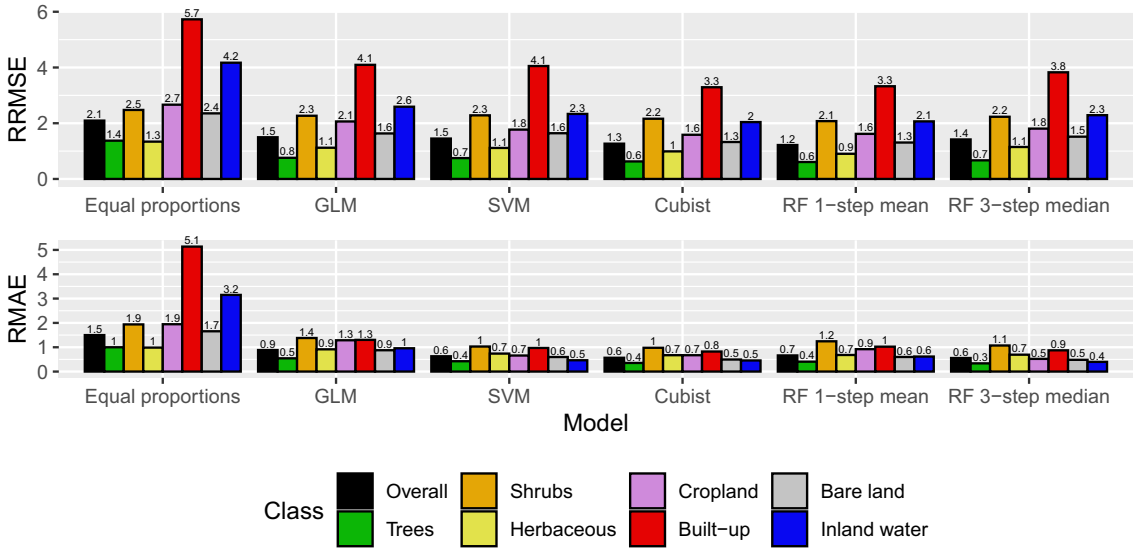

Fig. B.7. Comparison of relative RMSE (top) and relative MAE (bottom) per class of the best performing models in their category, and the equal proportions solution as a reference.

Our results showed that the binary relevance method works well to facilitate the use of univariate regression methods for the global land cover fraction mapping task, as the highest accuracy was achieved by algorithms that used the binary relevance method. This further expands the field of possible algorithms that could be used for the task. In addition, both the binary relevance method and the multi-step method can accept different algorithms for its submodels, i.e. for different classes in the case of binary relevance and for different steps in the case of the multi-step approach. There may be a combination of models that would work even better than the models we tested. We could not cover the entire range of methods in the scope of this study, therefore there is still room for improvement. For example, classical spectral unmixing methods, such as spectral mixture analysis, has been often used in the past (e.g. in Adams et al. (1995)). However, due to the limitations of the spectral mixture analysis methods (Somers et al., 2011), we could not use it with the features we selected, as they do not form a linear mixture representing the land cover fractions. Modern machine learning methods are advantageous for their ability to model complex nonlinear relationships between variables and thus make use of all of the features.

Appendix C. Feature importance

To gain insight on what features drive the RF three-step median model, we performed permutation importance on it: we shuffled the values of each feature in turn, and the model made predictions based on all of the features, including the shuffled one. This results in a decrease in the accuracy of predictions, as the feature in question no longer contains meaningful information. We then recorded the resulting increase in MAE compared to the validation set as a measure of feature importance, both per class and overall for all classes combined. The results are shown in Fig. C.8.

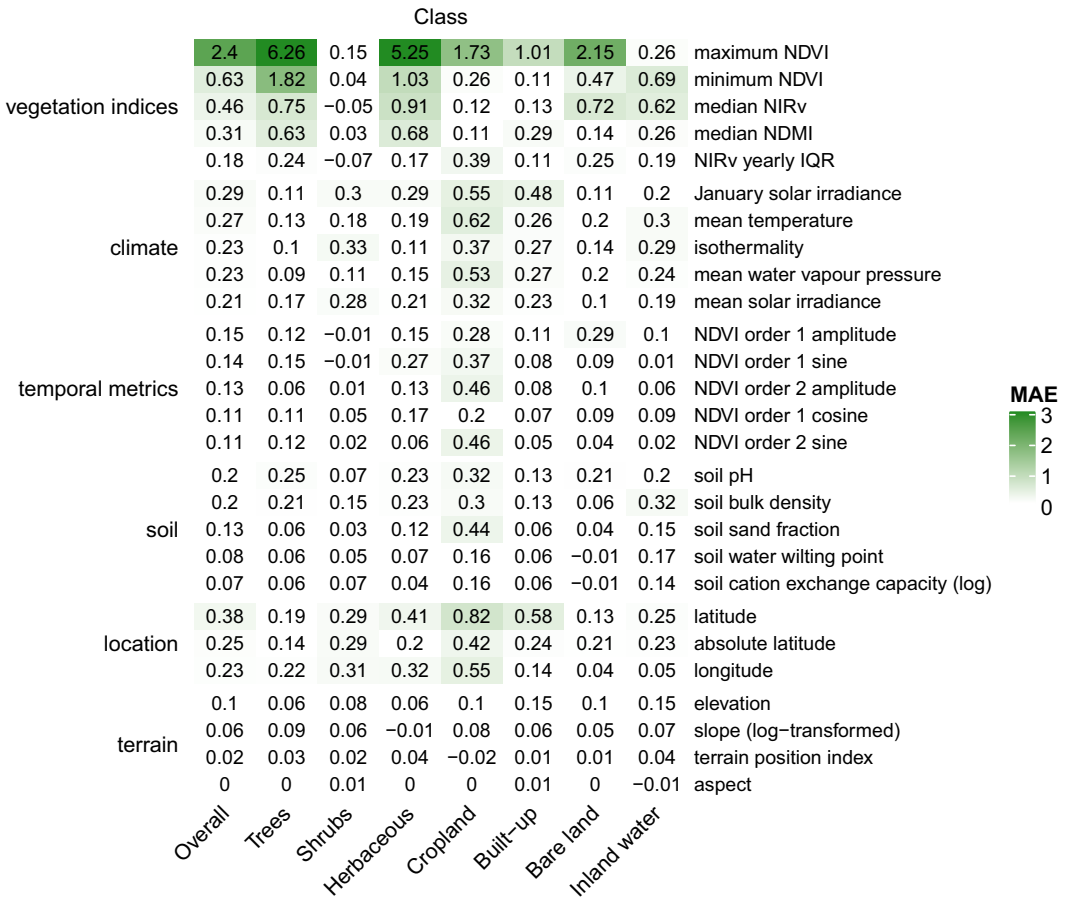

Fig. C.8. Random Forest three-step median approach variable importance, top 5 features per category. Categories are ordered by cumulative importance. The values represent increase in MAE for the given class when the given feature is permuted.

The overall most important variables were the maximum and minimum NDVI over the whole time series of PROBA-V imagery. They were followed by the median NIR_v and NDMI over the whole time series. The importance of remote sensing information can also be seen from the accuracy statistics of a single-step RF regression model, when only vegetation index and temporal features were input into the model (see Table 2, "only RS features"): RMSE and MAE increased only by around 1 percentage point.

The remote sensing data was of lower importance (and sometimes even confounding) only for the shrub class. This class is complicated to distinguish from other natural vegetation, and thus when trying to predict shrub cover fractions, the model benefited from a wide variety of additional features, including location and climate information.

To distinguish crops in particular, soil information, such as pH, bulk density and sand fraction, was useful for the model. Crops are usually grown in fertile soils that can sustain them, and due to the fact that cropland is further managed, the soil is also altered to be more fertile, which affects these soil properties. In addition, harmonic metrics derived from time series benefited crop cover fraction estimation the most of all the classes. The second order amplitude and sine of the harmonic model of the time series allows the model to detect areas with a double harvest throughout the year, which is indicative of crops.

Climate data, especially the mean temperature and the closely (exponentially) related mean of water vapour pressure, was also most useful for predicting crop cover fractions. It was also beneficial for estimating built-up fractions. Location features were the most important for predicting the cover fractions of these two classes as well. These classes tend to be spatially clustered, when looking at a large scale. While the models for predicting the cover fraction of classes made the best use of the location features, all of the classes could benefit from them to some extent, thus it is beneficial to include these intrinsic parameters in the model. Both absolute and regular latitude were used by the model to improve prediction accuracy, and latitude was more beneficial than longitude for increasing the prediction accuracy.

Terrain information was the least useful feature category. It was most useful for inland water fraction prediction, since it typically has little to no slope, and rarely occurs at high altitudes. The aspect was the one feature that does not appear to have contributed to model accuracy.

Remote sensing features were the most important features for the multi-step median RF model, especially the maximum and minimum NDVI over the entire time series. Note that these values are taken from a time series that has undergone temporal outlier removal, as detailed in section 2.2, and thus roughly correspond to the 5th and 95th percentiles of the data without additional temporal filtering. Multiple vegetation indices were useful for increasing the model accuracy: both NDMI and NIR_v median over the time series were much more important than any other feature from other groups. Nevertheless, even though a lot of the other features were of much lower importance, they contributed to prediction accuracy enough so that they could not be easily excluded from the models as redundant (after the removal of collinear features as explained in section 2.3). Methods based on decision trees help with effectively using features that may also have an overlap in the information that they provide, although linear models likewise tended to not exclude any features as non-informative. This is also due to the large variety of land cover classes in the study, since a feature is useful if it helps predict any of the land cover classes better.

The result that including more training features is beneficial, but remote sensing data is the most important, is in line with the conclusions of e.g. Li

et al. (2018) and Hengl et al. (2017). Remote sensing data is also unique in that it forms a time series, which enables us to both calculate additional temporal metrics and to monitor land cover change over time. In addition to remote sensing data, Hengl et al. (2017) also noted high importance of climate data, however, it is focused on soils, whereas climate has more effect on long-term processes such as soil formation than on land cover. Climate data was the second most important in our case, mostly for the crop fraction estimation, which is also closely linked with soils.

The feature importance results showed that all of the feature groups used in the study were useful, therefore leaving a feature out means sacrificing some predictive power. On the other hand, leaving out some features would be beneficial in that less time would be needed for processing, as that feature would no longer need to be downloaded, preprocessed and processed. This is an important consideration, given that for global land cover mapping, all features need to be available for the whole globe as well. In addition, model training performance (time for training and memory usage) may be an important consideration for global land cover mapping, as it may limit the scope of what models may be used for this task. For example, ensemble learning techniques like the super learner are very resource-intensive and may take weeks to train. Therefore for operational land cover product creation purposes, it is important to take into account not just the model accuracy, but also whether the improvement of the accuracy is worth the increase in processing time or computing resource usage.

Appendix D. Model accuracy changes per predicted fraction

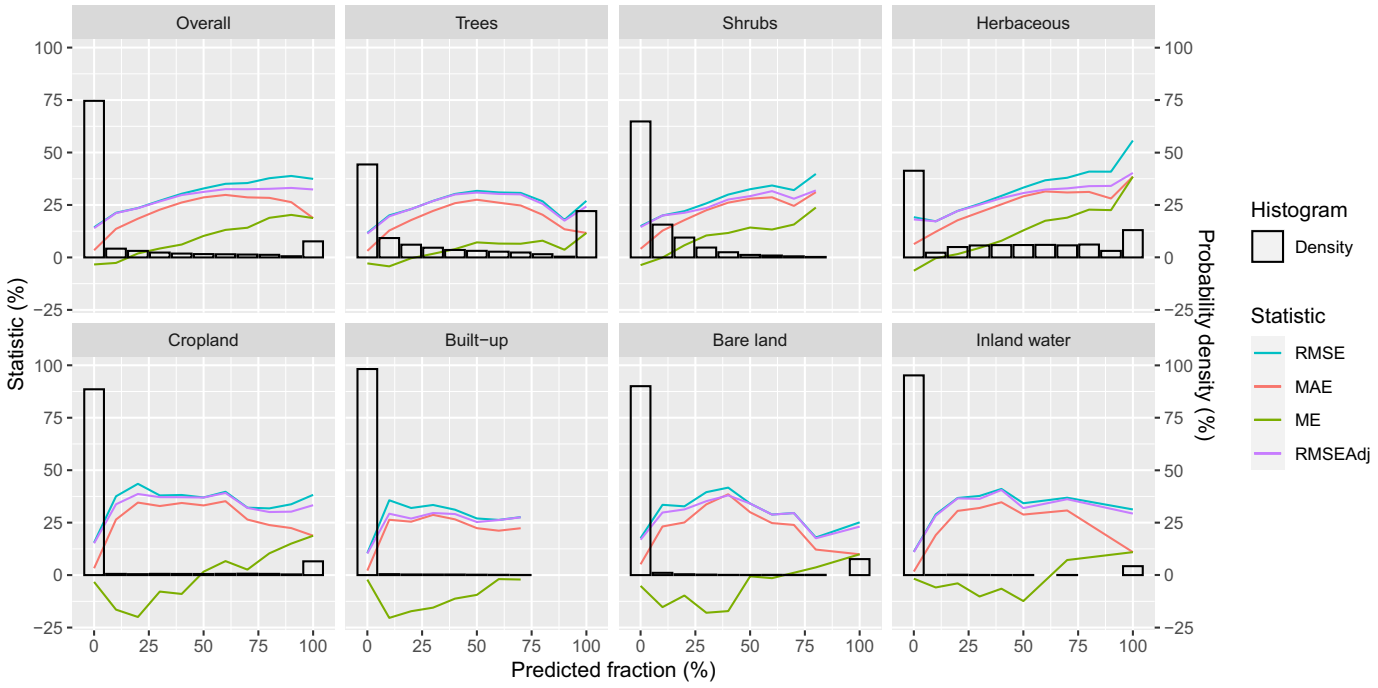

Fig. D.9. Change in accuracy statistics of the RF three-step median model per predicted land cover fraction, overlaid over the histogram of predictions. "RMSEAdj" stands for RMSE when the bias in predictions is adjusted for (ME is subtracted from the residuals). Bins containing less than 10 samples were omitted due to too low sample size leading to unreliable error statistics.

References

Adams, J.B., Sabol, D.E., Kapos, V., Almeida Filho, R., Roberts, D.A., Smith, M.O., Gillespie, A.R., 1995. Classification of multispectral images based on fractions of endmembers: application to land-cover change in the Brazilian Amazon. Remote Sens. Environ. *52*, 137–154. URL:

Allaire, J., Chollet, F., 2018. keras: R Interface to 'Keras'. URL: https://CRAN.R-project.org/package=keras.

Arino, O., Gross, D., Ranera, F., Leroy, M., Bicheron, P., Brockman, C., Defourny, P., Vancutsem, C., Achard, F., Durieux, L., Bourg, L., Latham, J., Gregorio, A.D., Witt, R., Herold, M., Sambale, J., Plummer, S., Weber, J.L., 2007. GlobCover: ESA service for global land cover from MERIS. In: 2007 IEEE International Geoscience and Remote Sensing Symposium, pp. 2412–2415. https://doi.org/10.1109/IGARSS.2007.4423328.

Bartholomé, E., Belward, A.S., 2005. GLC2000: a new approach to global land cover mapping from Earth observation data. Int. J. Remote Sens. 26, 1959–1977. https://doi.org/10.1080/01431160412331291297.

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

Buchhorn, M., Bertels, L., Smets, B., Lesiv, M., Tsendbazar, N.-E., . Moderate dynamic land cover 100 m algorithm theoretical basis document. URL: https://land.copernicus.eu/global/sites/cgls.vito.be/files/products/CGLOPS1_ATBD_LC100m-V2.0_I2.00.pdf https://doi.org/10.5281/zenodo.3606446.

Buchhorn, M., Smets, B., Bertels, L., Lesiv, M., Tsendbazar, N.-E., Herold, M., Fritz, S., . *Copernicus Global Land Service: Land Cover 100m: epoch 2015: Globe*. Zenodo. URL: https://zenodo.org/record/3243509#.XxbVajVc6V4 https://doi.org/10.5281/zenodo.3243509.

Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., Smets, B., 2020. Copernicus global land cover layers—collection 2. Remote Sens. *12*, 1044. URL:

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., Zhang, W., Tong, X., Mills, J., 2015. Global land cover mapping at 30m resolution: A POK-based operational approach. ISPRS J. Photogramm. Remote Sens. *103*, 7–27. https://doi.org/10.1016/j.isprsjprs.2014.09.002. URL: http://www.sciencedirect.com/science/article/pii/S0924271614002275.

Colditz, R.R., Schmidt, M., Conrad, C., Hansen, M.C., Dech, S., 2011. Land cover classification with coarse spatial resolution data to derive continuous and discrete maps for complex regions. Remote Sens. Environ. *115*, 3264–3275. URL:

Corbane, C., Florczyk, A., Pesaresi, M., Politis, P., Syrris, V., . GHS-BUILT R2018A - GHS built-up grid, derived from Landsat, multitemporal (1975-1990-2000-2014). URL: http://data.europa.eu/89h/jrc-ghsl-10007 https://doi.org/10.2905/jrc-ghsl-10007.

Corbane, C., Pesaresi, M., Kemper, T., Politis, P., Florczyk, A.J., Syrris, V., Melchiorri, M., Sabo, F., Soille, P., 2019. Automated global delineation of human settlements from 40 years of Landsat satellite data archives. Big Earth Data 3, 140–169. https://doi.org/10.1080/20964471.2019.1625528.

Coyle, J.R., Hejazi, N.S., Malenica, I., Sofrygin, O.. sl3: Pipelines for Machine Learning and Super Learning. URL. https://github.com/tlverse/sl3. r package version 1.3.7.

Dierckx, W., Sterckx, S., Benhadj, I., Livens, S., Duhoux, G., Achteren, T.V., Francois, M., Mellab, K., Saint, G., 2014. PROBA-V mission for global vegetation monitoring: standard products and image quality. Int. J. Remote Sens. 35, 2589–2614. https://doi.org/10.1080/01431161.2014.883097.

Dozat, T., 2016. Incorporating Nesterov Momentum into Adam. In: International Conference on Learning Representations. Puerto Rico.
Dreyfus, S.E., 1990. Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure. J. Guid. Control. Dyn. 13, 926–928. https://doi.org/10.2514/3.25422.
Dwivedi, R., Kumar, A., Ghosh, S.K., Roy, P.S., 2012. Optimisation of fuzzy based soft classifiers for remote sensing data. ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inform. Sci. *39B3*, 385–390. https://doi.org/10.5194/isprsarchives-XXXIX-B3-385-2012.
ESA, 2017. Land Cover CCI Product User Guide Version 2. European Space Agency. Technical Report. URL: maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf.
Fick, S.E., Hijmans, R.J., 2017. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315. https://doi.org/10.1002/joc.5086.
Foody, G.M., 1996. Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. Int. J. Remote Sens. 17, 1317–1340. https://doi.org/10.1080/01431169608948706.
Friedl, M.A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., Huang, X., 2010. MODIS collection 5 global land cover: algorithm refinements and characterization of new datasets. Remote Sens. Environ. *114*, 168–182. URL:
Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, *33*, 1–22. URL: https://www.jstatsoft.org/index.php/jss/article/view/v033i01. Doi: 10.18637/jss.v033.i01.
Gao, J., O'Neill, B.C., 2020. Mapping global urban land for the 21st century with data-driven simulations and shared socioeconomic pathways. Nat. Commun. *11*, 2302. URL:
GDAL/OGR contributors, 2020. GDAL/OGR Geospatial Data Abstraction Software Library. Open Source Geospatial Foundation. URL: https://gdal.org.
Gessner, U., Machwitz, M., Conrad, C., Dech, S., 2013. Estimating the fractional cover of growth forms and bare surface in savannas. A multi-resolution approach based on regression tree ensembles. Remote Sens. Environ. *129*, 90–102. https://doi.org/10.1016/j.rse.2012.10.026. URL: http://www.sciencedirect.com/science/article/pii/S0034425712004142.
Gong, P., Liu, H., Zhang, M., Li, C., Wang, J., Huang, H., Clinton, N., Ji, L., Li, W., Bai, Y., Chen, B., Xu, B., Zhu, Z., Yuan, C., Ping Suen, H., Guo, J., Xu, N., Li, W., Zhao, Y., Yang, J., Yu, C., Wang, X., Fu, H., Yu, L., Dronova, I., Hui, F., Cheng, X., Shi, X., Xiao, F., Liu, Q., Song, L., 2019. Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. Sci. Bull. 64, 370–373. https://doi.org/10.1016/j.scib.2019.03.002. URL: http://www.sciencedirect.com/science/article/pii/S2095927319301380.
Gong, P., Li, X., Wang, J., Bai, Y., Chen, B., Hu, T., Liu, X., Xu, B., Yang, J., Zhang, W., Zhou, Y., 2020. Annual maps of global artificial impervious area (GAIA) between 1985 and 2018. Remote Sens. Environ. 236, 111510. https://doi.org/10.1016/j.rse.2019.111510. URL: http://www.sciencedirect.com/science/article/pii/S0034425719305292.
Hansen, M.C., DeFries, R.S., Townshend, J.R.G., Carroll, M., Dimiceli, C., Sohlberg, R.A., 2003. Global percent tree cover at a spatial resolution of 500 meters: first results of the MODIS vegetation continuous fields algorithm. Earth Interact. 7, 1–15. https://doi.org/10.1175/1087-3562(2003)007<0001:GPTCAA>2.0.CO;2.
Hansen, M.C., Egorov, A., Roy, D.P., Potapov, P., Ju, J., Turubanova, S., Kommareddy, I., Loveland, T.R., 2011. Continuous fields of land cover for the conterminous United States using Landsat data: first results from the Web-Enabled Landsat Data (WELD) project. Remote Sens. Lett. 2, 279–288. https://doi.org/10.1080/01431161.2010.519002.
Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-resolution global maps of 21st-century forest cover change. Science *342*, 850–853. URL:
Hengl, T., Walvoort, D.J.J., Brown, A., Rossiter, D.G., 2004. A double continuous approach to visualization and analysis of categorical maps. Int. J. Geogr. Inf. Sci. 18, 183–202. https://doi.org/10.1080/13658810310001620924.
Hengl, T., Jesus, J.M.D., Heuvelink, G.B.M., Gonzalez, M.R., 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. https://doi.org/10.1371/journal.pone.0169748 e0169748. URL:
Hijmans, R.J., Phillips, S., Leathwick, J., Elith, J., 2017. dismo: Species Distribution Modeling. URL: https://CRAN.R-project.org/package=dismo.
Hobbs, S., 2003. *Linear mixture modelling solution methods for satellite remote sensing*. Report College of Aeronautics. Cranfield University. URL: http://dspace.lib.cranfield.ac.uk/handle/1826/94. ISBN: 9781871564839.
Jakubauskas, M.E., Legates, D.R., Kastens, J.H., 2001. Harmonic analysis of time-series AVHRR NDVI data. Photogramm. Eng. Remote. Sens. 67, 461–470.
Karalas, K., Tsagkatakis, G., Zervakis, M., Tsakalides, P., 2016. Land classification using remotely sensed data: going multilabel. IEEE Trans. Geosci. Remote Sens. 54, 3548–3563. https://doi.org/10.1109/TGRS.2016.2520203.
Keller, J.M., Gray, M.R., Givens, J.A., 1985. A fuzzy K-nearest neighbor algorithm. IEEE Trans. Syst. Man Cybern. SMC-15, 580–585. https://doi.org/10.1109/TSMC.1985.6313426.
Kuhn, M., Quinlan, R., 2020. Cubist: Rule- And Instance-Based Regression Modeling. URL: https://CRAN.R-project.org/package=Cubist. R package version 0.2.3.
Laan, M.J.V.D., Polley, E.C., Hubbard, A.E., 2007. Super learner. Stat. Appl. Genet. Mol. Biol. 6 https://doi.org/10.2202/1544-6115.1309. URL: https://www.degruyter.com/view/journals/sagmb/6/1/article-sagmb.2007.6.1.1309.xml.xml.
Li, L., Vrieling, A., Skidmore, A., Wang, T., Turak, E., 2018. Monitoring the dynamics of surface water fraction from MODIS time series in a Mediterranean environment. Int. J. Appl. Earth Obs. Geoinf. *66*, 135–145. https://doi.org/10.1016/j.jag.2017.11.007. URL. https://linkinghub.elsevier.com/retrieve/pii/S0303243417302660.
Lizarazo, I., 2012. Quantitative land cover change analysis using fuzzy segmentation. Int. J. Appl. Earth Obs. Geoinf. *15*, 16–27. https://doi.org/10.1016/j.jag.2011.05.012. URL. http://www.sciencedirect.com/science/article/pii/S0303243411000754.
Masiliūnas, D., 2020. *Code for global land cover fraction mapping on PROBA-V MEP*. Zenodo. URL:. https://doi.org/10.5281/zenodo.4281202. Version: v1.1. [dataset].
Mevik, B.-H., Wehrens, R., Liland, K.H., 2016. pls: Partial Least Squares and Principal Component Regression. URL: https://CRAN.R-project.org/package=pls.
Montesano, P.M., Nelson, R., Sun, G., Margolis, H., Kerber, A., Ranson, K.J., 2009. MODIS tree cover validation for the circumpolar taiga–tundra transition zone. Remote Sens. Environ. *113*, 2130–2141. URL:
NASA, METI, AIST, Japan Spacesystems, U.S./Japan ASTER Science Team, 2019. *ASTER Global Digital Elevation Model V003*. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/ASTER/ASTGTM.003.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. J. Hydrol. *10*, 282–290. URL:
Neter, J., Kutner, M., Nachtsheim, C., 1996. Applied Linear Statistical Models, 4th ed. McGraw-Hill/Irwin.
Okeke, F., Karnieli, A., 2006. Linear mixture model approach for selecting fuzzy exponent value in fuzzy c-means algorithm. Ecolog. Inform. *1*, 117–124. https://doi.org/10.1016/j.ecoinf.2005.10.006. URL: http://www.sciencedirect.com/science/article/pii/S1574954105000117.
Okujeni, A., Canters, F., Cooper, S.D., Degerickx, J., Heiden, U., Hostert, P., Priem, F., Roberts, D.A., Somers, B., van der Linden, S., 2018. Generalizing machine learning regression models using multi-site spectral libraries for mapping vegetation-impervious-soil fractions across multiple cities. Remote Sens. Environ. *216*, 482–496. URL:
Pearson, K., 1895. Notes on regression and inheritance in the case of two parents. In: Proceedings of the Royal Society of London, Vol. 58, pp. 240–242.
Pekel, J.-F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature *540*, 418–422. URL:
Quinlan, J.R., 1992. Learning with continuous classes. In: AI'92. World Scientific, Singapore. https://doi.org/10.1142/9789814536271.
R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria. URL. https://www.R-project.org/.
Schroeder, R., McDonald, K.C., Chapman, B.D., Jensen, K., Podest, E., Tessler, Z.D., Bohn, T.J., Zimmermann, R., 2015. Development and evaluation of a multi-year fractional surface water data set derived from active/passive microwave remote sensing data. Remote Sens. *7*, 16688–16732. URL:
See, L., Schepaschenko, D., Lesiv, M., McCallum, I., Fritz, S., Comber, A., Perger, C., Schill, C., Zhao, Y., Maus, V., Siraj, M.A., Albrecht, F., Cipriani, A., Vakolyuk, M., Garcia, A., Rabia, A.H., Singha, K., Marcarini, A.A., Kattenborn, T., Hazarika, R., Schepaschenko, M., van der Velde, M., Kraxner, F., Obersteiner, M., 2015. Building a hybrid land cover map with crowdsourcing and geographically weighted regression. ISPRS J. Photogramm. Remote Sens. *103*, 48–56. URL.
Sexton, J.O., Song, X.-P., Feng, M., Noojipady, P., Anand, A., Huang, C., Kim, D.-H., Collins, K.M., Channan, S., DiMiceli, C., Townshend, J.R., 2013. Global, 30-m resolution continuous fields of tree cover: Landsat-based rescaling of MODIS vegetation continuous fields with lidar-based estimates of error. Int. J. Digit. Earth 6, 427–448. https://doi.org/10.1080/17538947.2013.786146.
Sharma, C.S., Behera, M.D., Mishra, A., Panda, S.N., 2011. Assessing flood induced land-cover changes using remote sensing and fuzzy approach in eastern Gujarat (India). Water Resour. Manag. 25, 3219. https://doi.org/10.1007/s11269-011-9853-7.
Shimabukuro, Y., Smith, J., 1991. The least-squares mixing models to generate fraction images derived from remote sensing multispectral data. IEEE Trans. Geosci. Remote Sens. 29, 16–20. https://doi.org/10.1109/36.103288.
Silván-Cárdenas, J.L., Wang, L., 2008. Sub-pixel confusion–uncertainty matrix for assessing soft classifications. Remote Sens. Environ. 112 https://doi.org/10.1016/j.rse.2007.07.017. URL: http://www.sciencedirect.com/science/article/pii/S0034425707003434.
Somers, B., Asner, G.P., Tits, L., Coppin, P., 2011. Endmember variability in spectral mixture analysis: A review. Remote Sens. Environ. *115*, 1603–1616. URL:
Spearman, C., 1904. Rank's correlation. Am. J. Psychol. 15, 17–31.
Stavrakoudis, D.G., Theocharis, J.B., Zalidis, G.C., 2011. A Boosted Genetic Fuzzy Classifier for land cover classification of remote sensing imagery. ISPRS J. Photogramm. Remote Sens. *66*, 529–544. URL:
Steinwart, I., Thomann, P., 2017. liquidSVM: A fast and versatile SVM package. arXiv:1702.06899.
Suykens, J., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural. Process. Lett. *9*, 293–300. https://doi.org/10.1023/A:1018628609742.
Theil, H., 1969. A multinomial extension of the linear Logit model. Int. Econ. Rev. *10*, 251–259. URL:
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 58, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
Townshend, J., . Global Forest Cover Change (GFCC) Tree Cover Multi-Year Global 30 m V003. URL: https://lpdaac.usgs.gov/products/gfcc30tcv003/ https://doi.org/10.5067/MEASURES/GFCC/GFCC30TC.003.

Tsendbazar, N.-E., Herold, M., de Bruin, S., Lesiv, M., Fritz, S., Van De Kerchove, R., Buchhorn, M., Duerauer, M., Szantoi, Z., Pekel, J.F., 2018. Developing and applying a multi-purpose land cover validation dataset for Africa. Remote Sens. Environ. *219*, 298–309. URL:

Tsendbazar, N.-E., Herold, M., Lesiv, M., Fritz, S., 2019. Copernicus Global Land Service: Land Cover 100m: Version 2 Globe 2015: Validation Report. Technical Report Copernicus Global Land Service, URL.

Uma Shankar, B., Meher, S.K., Ghosh, A., 2011. Wavelet-fuzzy hybridization: feature-extraction and land-cover classification of remote sensing images. Appl. Soft Comput. *11*, 2999–3011. URL:

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, 4th ed. Springer, New York. URL: http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

Walton, J.T., 2008. Subpixel urban land cover estimation. Photogramm. Eng. Remote Sens. *74*, 1213–1222. URL:

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. *58*, 109–130. URL: http://www.sciencedirect.com/science/article/pii/S0169743901001551 https://doi.org/10.1016/S0169-7439(01)00155-1.

Wolters, E., Dierckx, W., Iordache, M.-D., Swinnen, E., 2016. *PROBA-V Products User Manual v2.0*. VITO N.V.

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

Yang, J., Weisberg, P.J., Bristow, N.A., 2012. Landsat remote sensing approaches for monitoring long-term tree cover dynamics in semi-arid woodlands: comparison of vegetation indices and spectral mixture analysis. Remote Sens. Environ. *119*, 62–71. URL:

Zhang, J., Foody, G.M., 2001. Fully-fuzzy supervised classification of sub-urban land cover from remotely sensed imagery: Statistical and artificial neural network approaches. Int. J. Remote Sens. 22, 615–628. https://doi.org/10.1080/01431160050505883.