Preidl, S., **Lange, M.**, **Doktor, D.** (2020):
Introducing APiC for regionalised land cover mapping on the national scale using Sentinel-2A imagery
Remote Sens. Environ. **240** , art. 111673

Introducing APiC for regionalised land cover mapping on the national scale using Sentinel-2A imagery

Sebastian Preidl, Maximilian Lange, Daniel Doktor

Helmholtz Centre for Environmental Research - UFZ
Department of Computational Landscape Ecology
Permoserstrasse 15, 04318 Leipzig, Germany
email: sebastian.preidl@ufz.de

Abstract

Overcoming the obstacle of frequent cloud coverage in optical remote sensing imageries is essential for monitoring dynamic land surface processes from space. APiC, a novel adaptable pixel-based compositing and classification approach, is especially designed to use high resolution spatio-temporal space-borne data. Here, pixel-based compositing is used separately for training data and prediction data. First, cloud-free pixels covered by reference data are used within adapted composite periods to compile a training dataset. The compiled training dataset contains samples of spectral reflectances for respective land cover class at each composite period. For land cover prediction, pixel-based compositing is then applied region-wide. Multiple prediction models are used based on temporal subsets of the compiled training dataset to dynamically account for cloud coverage at pixel level. Thus we present a data-driven classification approach which is applicable in regions with different weather conditions, species composition and phenology. The capability of our method is demonstrated by mapping 19 land cover classes across Germany for the year 2016 based on Sentinel-2A imageries. Since climatic conditions and thus plant phenology change on a large scale, the classification was carried out separately in six landscape regions of different biogeographical characteristics. The study could draw on extensive ground validation data provided by the federal states of Germany. For each landscape region, composite periods of different lengths have been established, which differ regionally in their temporal arrangement as well as in their total number, emphasising the advantage of a flexible regionalised classification procedure. Using random forest, an overall accuracy of 88% was achieved, with particularly high classification accuracy of around 90% for the major land cover types. We found that class imbalances has significant influence on classification accuracy. Based on multiple temporal subsets of the compiled training dataset, over 10,000 RF models were calculated and their performances varied considerably across and within landscape regions. The calculated importance of composite periods show that a high temporal resolution of the compiled training dataset is necessary to better capture the different phenology of land cover types. In this study we demonstrate that APiC, due to its data-driven nature, is a very flexible compositing and classification approach making efficient use of dense satellite time series in areas with frequent cloud coverage. Hence, regionalisation can be given greater focus in future broad-scale classifications in order to facilitate better integration of small-scale biophysical conditions and achieve even better results in detailed land cover mapping.

Keywords: land cover classification, Compositing, Crop mapping, Phenology, Sentinel-2, Random Forest

1. Introduction

Land cover has indeed become a force of global importance in recent years (Foley et al., 2005). Global demographic and economic developments are leading to an increase in anthropogenic land use and land cover change. Due to the ongoing transformation of natural ecosystems into agricultural land, 37% of the area is currently used for agriculture (`https://data.worldbank.org/indicator/AG.LND.AGRI.ZS` (accessed 5 April 2019)).

World-wide, the expansion of agriculture is often at the expense of forests (Hansen et al., 2013), contributing greatly to the negative trends in carbon stocks (DeFries et al., 2010; Houghton, 2010), climate change (Sombroek, 2001) and biodiversity (Billeter et al., 2008; Dormann et al., 2007; Newbold et al., 2015). On the local level, intensification and monocultures are responsible for the decline in soil fertility, which in turn contributes to an overuse of fertilisers (Smith et al., 2016). Land cover configuration is an important factor for reassessing nitrogen input into surface water or runoff, biodiversity loss due to the lack of animal corridors (Bleyhl et al., 2017) or changed pollination dynamics (Hadley and Betts, 2012).

Hence, there is an urgent need to gather information on how the land is being used at field level over time, so that land management can be improved. Remote sensing is a widespread tool for mapping land surfaces and has often been used to capture broad land cover categories such as forests, water bodies or agricultural land (Joshi et al., 2016).

However, mapping thematically detailed land cover classes - and crop types in particular - continues to be challenging. With the launch of Sentinel-2A, new classification approaches are conceivable, as the Earth observation instrument has relatively high resolutions in all three domains: (1) temporal: a revisit time of 2-3 days at mid-latitudes allows a better detection of dynamic vegetation processes; (2) spatial: a pixel size of 10 or 20 m allows the capture of smaller-scaled land cover configurations; and (3) spectral: 13 and 9 spectral bands at 10 m and 20 m ground resolution respectively allow plants with similar physiological and morphological characteristics to be better distinguished by their spectral traits.

The temporal resolution of a satellite system determines the number of available observations per time unit but says little about the usability of individual image pixels, which can be affected by cloud cover. In optical remote sensing, cloud removal techniques are required for large area land cover mapping or longer time series analysis (Cihlar, 2000), as the Earth's surface can only be reliably observed under cloud-free conditions. The detection and substitution of clouds for land cover mapping is usually done by pixel-based image compositing (Holben, 1986), where a contaminated pixel is replaced by the same pixel of a cloud-free satellite observation within a given time interval. The length and timing of these intervals should be well considered for the composites to be radiometric consistent.

Recently, Gomez et al. (2016) concluded that novel classification procedures which exploit the information in complex temporal data are not yet realized. In this sense, and in light of the high temporal and spatial resolution of Sentinel-2, we introduce a dynamic approach for adaptable pixel-based compositing and classification, called APiC.

We refrain from creating a seamless, cloud-free and artifact-free image composite of the entire study area (Lueck and van Niekerk, 2016; Roberts et al., 2017) and go beyond the original idea of pixel-based compositing where the best-available-pixel is selected by rule-based criteria (Lueck and van Niekerk, 2016; White et al., 2014). Instead, APiC distinguishes between two pixel-based compositing processes: (1) Compositing is exclusively applied to pixels covered by reference data. The aim is to compile spectral reflectances of different land cover types from different times of the year in a training dataset for analysis by a supervised classification algorithm. Within an iterative process, the availability of cloud-free pixels per land cover type determines the length and temporal localisation of each time interval. Due to the dynamic, data-driven process, we call our composite approach adaptable and the time intervals to be defined herein as (adaptable) composite periods. (2) Compositing is applied to all Sentinel-2A pixels, including those that are not part of the previously compiled training dataset. It is therefore likely that not all pixels can be compiled cloud-free in each composite period, so that here pixel-based compositing takes place within combinations of composite periods. Temporal subsets of the compiled training dataset are extracted accordingly, thereby requiring multiple prediction models for region-wide land cover mapping in APiC.

The dynamic, data-driven generation of composite periods is central to our approach, as the spectral trajectories of land cover's phenology are captured in more detail in high-resolution training data. This is in contrast to earlier studies, in which composites were created monthly wise (Roy et al., 2010) or around static (Griffiths et al., 2013) or adaptive seasonal target days-of-the-year (Frantz et al., 2017). Manually specified target days (White et al., 2014) require expert knowledge about the seasonal growth cycle in the study area and for each land cover type of interest. This knowledge can also be derived by spectral indices such as the normalized difference vegetation index (NDVI) to determine the season of main photosynthetic activity (Griffiths et al., 2013).

Within a thematically detailed land cover classification, however, few target days or long time intervals would disregard the different phenological patterns of the individual species. Cereals, for example, undergo a phenological cycle of nine growth stages from germination to senescence (Lancashire et al., 1991; Witzenberger et al., 1989). The distinction between cereal crops can only succeed if the temporal shifts in their growth phases can be identified. Once target days/intervals have been defined, their application in other regions may be undermined by changing climatic conditions and by the presence of plants with different vegetation dynamics. APiC is therefore less about when but rather how often growth phases can be captured in image composites without drawing on regional prior knowledge. That makes our data-driven approach easily applicable to regionalized studies. Accordingly, we have applied APiC not only once for the whole of Germany, but separately for six landscape regions.

One could argue that fixed, very narrowly defined time intervals would be even better suited to resolve plant dynamics. However, cloud-free pixel observations may be missing at these shorter intervals. Temporal data gaps in composites can be filled, for example, by regression imputation or mean imputation (Griffiths et al., 2019). In APiC, compositing is only based on available surface reflectance data, leading to very dense sequence of composite periods or - in times of persistently high cloud coverage - to larger temporal gaps between periods.

In summary, APiC differs from common classification methods in two main respects. First, APiC uses only available ground reference data and corresponding cloud-free Sentinel-2 pixels to define composite periods. Maximising their number requires composite periods be adaptable in length and temporal arrangement. Second, data imputation methods are not applied in APiC. Instead, multiple classification models are used for region-wide classification in order to account for cloud-free observation times on a pixel-by-pixel basis. The different prediction errors of the classification models allow a better understanding of the processes within APiC and a comprehensive evaluation of the results.

Our paper is structured as follows: The data used for regionalised land cover mapping are presented in section 2. The method section 3 first describes common methods used in APiC. Hereafter, central elements of APiC are defined: composite periods, the compiled training dataset, and the use of multiple prediction models for a region-wide classification. The classification result and other outcomes of APiC are presented and discussed in sections 4 and 5, respectively. Our concluding remarks are given in section 6.

2. Data

2.1. Satellite data

Sentinel-2A data were used to classify Germany's agricultural land for the year 2016 (Sentinel-2B was launched not until 2017). We opted for the higher spectral resolution (9 spectral bands) at 20 meter ground resolution to benefit from the spectral bands in the near-infrared (red edge) and shortwaveinfrared. The spatial resolution is suitable for our classification problem on the landscape level and resolves most field parcel sizes in Germany. In total 7200 Sentinel-2A tiles of the year 2016 were downloaded from the 'Copernicus Open Access Hub' (https://scihub.copernicus.eu), which were converted from radiance to bottom of atmosphere reflectances using ESA's processor Sen2Cor (Louis et al., 2016) in a (semi-) automatic processing routine. In addition, a so-called scene classification (SCL) image is generated by Sen2Cor, which identifies pixels that have been influenced by clouds or haze. For classification purposes, only pixels were used assigned to the classes "dark area pixels", "vegetation" or "bare soil" in the SCL image and thus identified as cloud-free. For the sake of simplicity, our definition of the term 'cloud-free' comprises all pixels showing land surface reflectances and therefore excludes not only cloud contaminated data but also missing data ('blackfilled areas'). Since winter crops of the following year are already sown in autumn, we have only used the satellite images from January to the end of October for the land cover classification in 2016.

A total of 470,578,123 Sentinel-2 pixels were classified, which is approximately 188231.2 square kilometers. With a total size of Germany of about 357578.2 square kilometers, this results in a relative proportion of 52.64%. This corresponds very well to the official figures according to which 50% of the land area is used for agriculture.

2.2. Ground observational data / Ancillary Data

Digital landscape model

The digital landscape model (DLM) of the official topographic-cartographic information system (ATKIS) from 2015 was used to differentiate between agricultural and non-agricultural areas. The numerous polygons of this vector data set were aggregated accordingly into the following categories: 1. urban, 2. water, 3. forest, 4. other vegetation and 5. farmland (including grassland, stone fruit plantations and hops). Subsequently, the shapefile was tailored to the geometric specifications of the Sentinel-2 tiles and rasterised to a 20 metre grid. Only the Sentinel-2A pixels matching the "farmland" class pixels were considered in the subsequent classification.

Landscape regions

Germany is characterized by different climate conditions and its landscape was influenced by different glacial-morphological and soil formation processes. Growth conditions vary respectively across the country. As a result our classification was separately performed in predefined landscape regions whose demarcation is based on those biogeographical conditions (Fig. 1, U.Hauke & A. Ssymank, Federal Agency for Nature Conservation (not published) based on IFAG (1979); Meynen et al. (1953-62)). The region Alps in the original dataset has been joined to the region Alpine Foreland for this study.

The sandy, hilly plains of the two lowland regions in the northwest (NW) and northeast (NE) part of Germany are closest to the sea and were mainly formed by ice age glaciers. The Upland regions are characterised by steeper and forested low mountain ranges. The Alpine Foreland is shaped by hilly meadows and forests in the north and end moraine landscapes in the south. In general, all western regions are more affected by the mild marine climate so that Germany's warmest places on average can be found in the southwest (SW-Upland region). A more continental climate characterises the eastern regions and the Alpine Foreland. Due to fertile loess deposits, the largest agricultural plains can be found in the NE-Lowland and E-Upland regions.

First, since phenology is the main driver for the differentiation of land cover types, we think that the consideration of landscape regions will improve the classification result. Second, we want to demonstrate that establishing composite periods is indeed an adaptable process to the given data availability and cloud coverage at the study site.

Integrated administration and control system

The EU Member States are accountable to maintain an integrated administration and control system (IACS) that was introduced to harmonise the agricultural policy between the countries and to support fair EU-payments to the landowners. This vector data set was provided by the state authorities we contacted and describes the geometry of individual field parcels, including the land cover types cultivated in the year 2016. These anonymised information was used for calibration (training) and validation of the land cover classification. The IACS data is distributed across Germany over about 25% of the total area, but differ in their extent within landscape regions (Fig. 1). The clustered data distribution in southern Germany is based on rectangular geometries, which we have provided for the states of Hesse, Baden-Wuerttemberg and Bavaria. This allowed us to cover the main agricultural areas in these regions. Similar to the DLM, the IACS shapefiles were tailored to the geometric specifications of the Sentinel-2 tiles and rasterised to a 20 metre grid. We have found that the land cover types in the reference data are unbalanced, meaning that for the most wide-spread land cover types, such as winter wheat or grassland, many millions of pixels are available, for others only a few thousand (Table 1). However, in order to include the most common crops (including smaller classes like spelt or spring oat) in the classification, we have set the minimum number of pixels to be available for each land cover type to the absolute threshold of 20,000 pixels. Due to its local relevance, we made an exception for the class stone fruits in the Alpine Foreland region, which was only represented by about 12,000 pixels. The strawberries class in the SW-Uplands was also included despite the lower 18,000 pixels. Given this threshold and including all landscape regions, a total of 19 land cover types were mapped: winter wheat, spelt, winter rye, winter barley, spring wheat, spring barley, spring oat, maize, legumes, rapeseed, leeks, potatoes, sugar beets, strawberries, stone fruits, vines, hops, asparagus and grassland (Table 1).

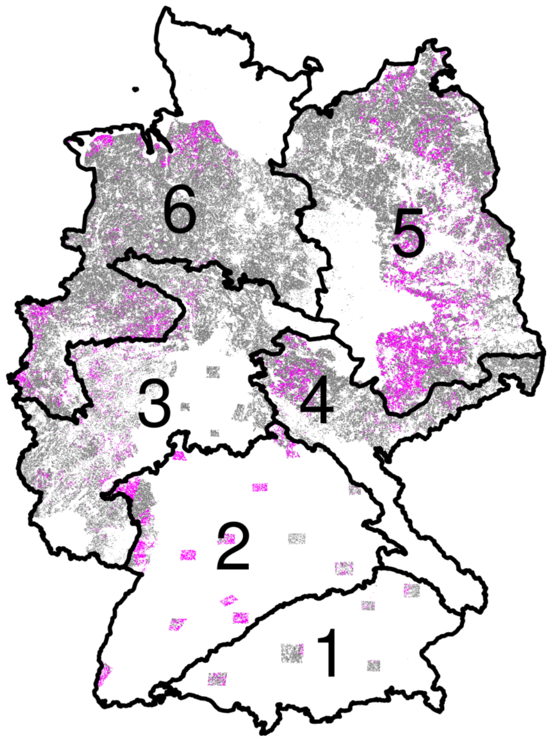

Fig. 1: The landscape regions of Germany (from South to North, black lines): Alpine Foreland (1), SW-Uplands (2), W-Uplands (3), E-Uplands (4), NE-Lowlands (5) and NW-Lowlands (6). Around 25% of Germany and thus approx. 50% of the total agricultural area is covered by reference data (IACS) (grey + magenta). Reference data used for pixel-based compositing of the training data is shown in magenta. The grey colored areas were used for validation.

Table 1: Number of pixels per landscape region and land cover class available in the given IACS data (reference data) (column *Ref*) and included in the compiled training dataset (column *Train*). The sample size of the compiled training dataset was reduced via LHS before being used to train prediction models (column *rTrain*).

	Alpine Foreland			SW-Uplands			W-Uplands			E-Uplands			NE-Lowlands			NW-Lowlands		
Land cover classes	Ref	Train	rTrain	Ref	Train	rTrain	Ref	Train	rTrain	Ref	Train	rTrain	Ref	Train	rTrain	Ref	Train	rTrain
Winter wheat	401607	76432	10815	2037420	1098796	16000	8394856	428771	4610	6532480	786206	16000	18231838	5447894	15353	15205694	1608466	8337
Spelt	0	0	0	64905	42177	1434	96676	7562	1017	92866	9269	1036	162441	73390	1165	104657	20348	1058
Winter rye	0	0	0	85052	42194	1434	321122	15602	1086	286381	21015	1262	6641181	1641235	5304	3237765	98529	1416
Winter barley	138533	23470	3492	483843	253169	4343	3147164	211170	2754	2285595	233741	5359	7189276	2176679	6717	5412589	633802	3870
Spring wheat	0	0	0	0	0	0	91475	5511	1000	116279	13897	1125	461955	83259	1191	253243	23245	1071
Spring barley	42894	8887	1475	571351	255142	4370	700811	46761	1351	1033502	89477	2580	699500	184654	1458	1400803	61877	1248
Spring oat	0	0	0	55982	33306	1312	255507	15989	1089	195945	17534	1195	645391	148373	1363	299814	34723	1124
Maize	375549	43080	6203	854315	461852	7219	3151602	162299	2337	1970309	127985	3322	11059563	2991749	8869	20508696	1467186	7689
Legumes	0	0	0	50160	28704	1248	195035	10740	1044	519647	49228	1805	1324832	384947	1987	373568	53904	1212
Rapeseed	41397	6126	1093	429483	245918	4243	3059858	176090	2455	3622931	385546	8283	11655107	3283103	9638	2748634	299936	2340
Leeks	0	0	0	70821	19943	1128	0	0	0	0	0	0	20728	13109	1005	133905	11427	1017
Potatoes	52188	10438	1690	173588	71809	1843	137273	6456	1008	73039	7394	1000	640004	207994	1520	3378061	261669	2164
Sugar beets	74170	17180	2622	470665	208023	3720	705745	21836	1139	203852	20987	1261	1151572	365771	1936	2846436	267361	2190
Strawberries	0	0	0	17956	10652	1000	33885	6070	1004	0	0	0	41470	10843	1000	177808	14286	1030
Stone fruits	12184	5447	1000	87868	55070	1612	60230	29267	1202	71702	14564	1138	104119	53165	1111	271146	64466	1260
Vines	0	0	0	1381951	652371	9846	256379	36040	1260	0	0	0	0	0	0	0	0	0
Hops	56348	6368	1127	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Asparagus	0	0	0	63893	22870	1168	0	0	0	0	0	0	132950	31812	1055	269912	7555	1000
Grassland	813100	113922	16000	1361984	774876	11534	16048472	1764047	16000	6909364	640864	13200	19971676	5692927	16000	22555561	3280278	16000
Sum	2007970	311350	45517	8261237	4276872	73454	36656090	2944211	40356	23913892	2417707	58566	80133603	22790904	76672	79178292	8209058	54026

3. Methods

3.1. Random forest classifier and validation

In APiC, a machine learning classifier, random forest (RF) (Breiman et al., 1984), is used for a supervised pixel-based land cover classification. RF is well-suited to solving high-dimensional problems and thus for the analysis of multispectral satellite time series. We applied Breiman and Cutler's RF implemented in R ('randomForest' package from Liaw and Wiener (2013)). Here, we set the internal RF parameters *ntree* (the number of internally grown trees) to 500 and *mtry* (the number of variables at each split) to the square root of the number of input variables.

Out-of-bag error

Besides its ability to work with numerous predictor variables, RF internally calculates estimates of the prediction error. Since RF trees are drawn by boostrapping it is referred to as the out-of-bag (OOB) error (Breiman, 2001). Due to our adaptable classification approach, in which multiple RF models are computed, we have used the OOB error to handle class imbalances in the compiled training dataset, to map the model prediction error at pixel level, and to determine the importance of composite periods.

Validation

An independent accuracy assessment of our classification result was performed based on the Sentinel-2A pixels and reference data that were not used for pixel-based compositing of the training data. For validation we have computed the confusion matrix, user accuracy (UA), producer accuracy (PA), overall accuracy and the Kappa coefficient (Congalton, 1991).

The calculated class-specific accuracy measures were also used to validate the class OOB error and gain insight into the general model behavior.

3.2. Latin hypercube sampling

For very large, multispectral datasets, such as the compiled training dataset in APiC, it would be beneficial to work only with samples that cover the original value range of each spectral band. Latin hypercube sampling (LHS) (McKay et al., 1979), a constrained Monte-Carlo sampling scheme is used to select samples which cover the hypercube of the feature space (Minasny and McBratney, 2006). We applied the R-package Conditioned Latin Hypercube Sampling (Roudier, 2011) that implemented LHS with a search algorithm based on heuristic rules combined with an annealing schedule (Metropolis et al., 1953; Minasny and McBratney, 2006).

3.3. Normalized difference vegetation index

The normalized difference vegetation index (NDVI) is considered as an indicator of vegetation activity. In a natural seasonal growth cycle, rising NDVI values up to +1 indicate vegetation with increasingly dense and greener leaves, while senescence is associated with declining NDVI values. Thus, NDVI has frequently been used for monitoring vegetation phenology and other ecological variables. It is calculated from reflectance values in the near infrared and the red visible range. The NDVI ratio for the Sentinel-2 bands is defined as:

$$\frac{Band\,8_{865} - Band\,4_{665}}{Band\,8_{865} + Band\,4_{665}}, \tag{1}$$

where the lower case number refers to wavelength in nm unit.

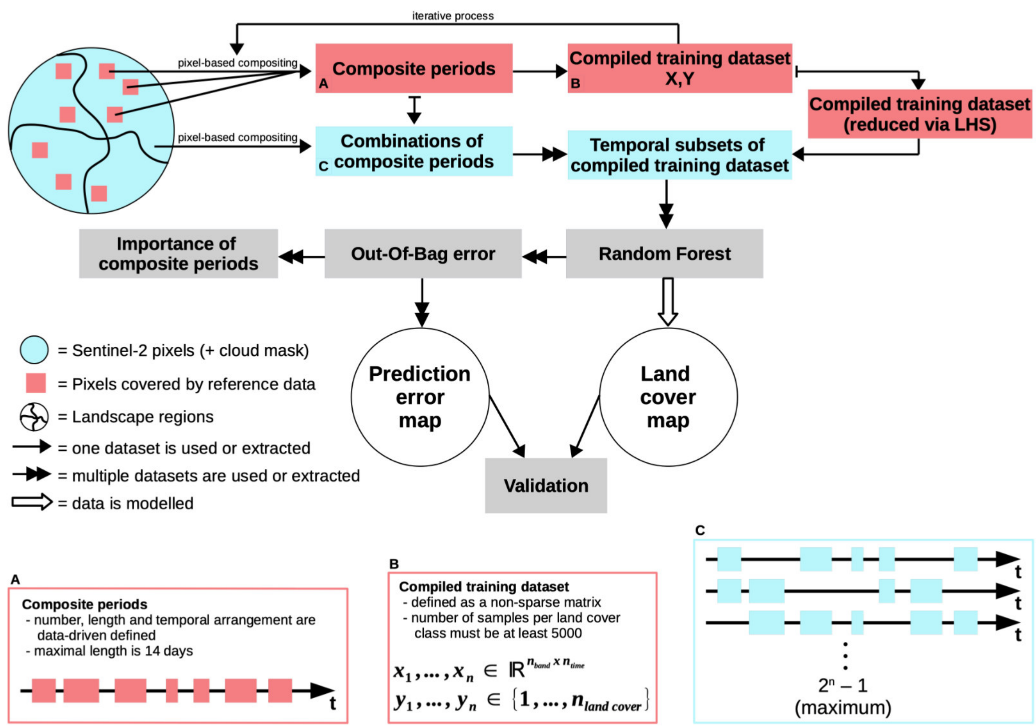

Fig. 2: Flow chart of the proposed adaptable pixel-based compositing and classification approach (APiC). During pixel-based compositing of training data (i.e. related to cloud-free Sentinel-2A pixels that are covered by reference data) composite periods are defined within an iterative process. At its end, composite periods were created whose length and temporal arrangement are adapted to the cloud cover in the satellite data and to the land cover information in the reference data. Additionally, a training dataset with a minimum sample size per land cover class has been compiled. For each composite period, reflectance values must be available in the compiled training data set (non-sparse). Due to class imbalances and excessive data volumes, the size of the training dataset is reduced via LHS. Pixel-based compositing of prediction data is based only on composite periods in which cloud-free pixel observations are available. According to these combinations of composite periods, temporal subsets are extracted from the compiled training dataset and passed to random forest. The number of prediction models to be computed therefore reflects the satellite observation density and/or temporal cloud coverage at pixel level. The OOB error output of each random forest model is used to create a map of prediction error estimates that complements the final land cover map. The windows marked with the letters A, B, C illustrate respective terms in the flow chart.

3.4. APiC

The following methodological description of APiC corresponds to the workflow shown in Fig. 2. Please note that in this study APiC was applied separately for each landscape region.

3.4.1. Pixel-based compositing of training data

In APiC, composite periods are used to compile a multitemporal and multispectral training dataset from cloud-free Sentinel-2A pixels that are covered by IACS reference data. A high temporal resolution of the compiled training dataset allows the vegetation phenology to be spectrally mapped more accurately. We therefore aim to maximise the number of composite periods within the classification year.

Composite periods

We consider a time period in which cloud-free pixels are compiled to be adaptable, since it is data-driven established, i.e. its length and temporal localisation are not fixed in advance. A composite period is defined by its maximal length, which must not exceed 14 days ($l_{CP} \leq 14$). Phenological studies have shown that, on average, there is no substantial progress in plant growth within two weeks and that the temporal shift between identical growth stages of different land cover types is - in most cases - more than two weeks (Xu et al., 2017). Thus, the spectral fingerprint of a growth stage should be well captured for each land cover type given this time window.

Compiled training dataset

Each sample (pixel) of the compiled training dataset is labelled with a land cover class from the reference data. In our classification context, land cover describes the outcome/dependent variable, while the associated spectral data of all composite periods represent the predictor/independent variables.

(1) The compiled training dataset is defined as a non-sparse matrix, that is, spectral values must be available for each composite period (NA values are not permitted). Hence, the number of predictor variables is given by the number of composite periods and the spectral resolution of the satellite system. For example, given the nine Sentinel-2A spectral bands, a compiled training dataset based on 12 established composite periods would have 108 predictor variables.

(2) It is defined, that the compiled training dataset consists of at least 5000 samples per land cover class ($n_{LC} \geq 5000$). Our empirical analyses showed that 5000 training pixels cover most of the spectral variance of a land cover class. This may be subject to modification depending on the size of the study area, land management and land cover types to be classified.

Iterative process

Compositing starts with analysing Sentinel-2A images from the first observation date of the year. The first composite period is established when 5000 cloud-free pixels per land cover class are available, otherwise the Sentinel-2A images of the next observation date are additionally included. In the latter case the same pixel may occur cloud-free in more than one satellite image. For compositing, this pixel is then taken from the image with the least total cloud coverage. If the length of a composite period reaches 14 days but the minimum 5000 pixels have not been found for all land cover classes, the second observation date of the year will be considered as the new start date for the compositing procedure. This process continues until the first composite period is established. All the following composite periods are created accordingly. Their earliest possible start date marks the first satellite observation after the end of the previous composite period.

Since the compiled training dataset must be non-sparse, samples with missing spectral values for any composite period are removed. As the number of composite periods increases, it is more likely that land cover classes will no longer be represented by at least 5000 samples. Therefore, maximising the number of composite periods becomes an indefinite iterative process:

$$n_i = min(n_{LC}) + I * i, \quad (2)$$

where n_i is the number of samples that must be contained in the compiled training dataset of the current iteration and only refers to the land cover class(es) that were underrepresented ($<$ 5000 samples) in the previous iteration, $min(n_{LC})$ equals 5000 and refers to the minimum number of samples per land cover class that a compiled training dataset must contain after completion of the iteration process, I is set to 1000 and defines the increment of n_i per iteration. i is initially set to zero and then increased by 1 for each iteration (0, 1, 2, . . .).

Starting the second iteration of pixel-based compositing ($i = 1$) with increased n_i forces some composite periods to be adjusted in length and rearranged in time, as more cloud-free pixels need to be found for certain classes. The iterative process is aborted once a compiled training dataset has been created that is non-sparse and contains at least 5000 samples per land cover class.

3.4.2. Class imbalances

Depending on the given reference data, the compiled training dataset can be affected by strong class imbalances, with some land cover classes being overrepresented by several orders of magnitude. These land cover classes inflate the compiled training dataset unnecessarily and increase the classifier's computational load. It is also known that class imbalances in the training data affect the classification result of RF and the validation outcome (Janitza and Hornung, 2018; Karpatne et al., 2016; Stumpf and Kerle, 2011).

Aiming at an operational classification framework, we automated the determination of appropriate class proportions in the compiled training dataset. Ten subsamples were created with increasing degrees of class imbalances using LHS. In the first subsample all classes are evenly represented with 1000 samples. In the next subsample, the size of the largest class was incremented by 5000 to 6000, 11000, ..., 46000. The other classes were sampled proportionally between 1000 and the respective maximum value. 1000 samples for the smallest class ensure the representation of its spectral variance and limit the size of the entire subsample. All ten subsamples were subsequently passed to RF to analyse the evolution of the OOB error geometrically. The error difference between a straight line connecting the OOB error value of the first (balanced) and last (most unbalanced) subsample and the OOB error curve was calculated. We expect the subsample where the calculated difference is largest to hold the best compromise between model performance and sample size. Its RF model will also provide more realistic class proportions in the land cover map and will therefore be used as the (reduced) compiled training dataset in our classification.

3.4.3. Pixel-based compositing of prediction data

The compiled training dataset would be best qualified for training a prediction model for land cover classification as it promises the highest temporal resolution. However, this means that each pixel to be classified would need to be observed cloud-free at least once in each composite period. In this case, the number of predictor variables of compiled prediction data would be identical to those in the compiled training dataset. For pixels not covered by reference data and therefore not considered during compositing of the trainings data this is unlikely. Rather pixel-based compositing leads to data gaps at different composite periods due to missing vegetation reflectance values. Theoretically, there are $2^n - 1$ possible combinations of how data gaps can occur across the compiled prediction data, where n refers to the number of composite periods. Assuming that our compiled training dataset is based on 12 composite periods, it may be that for some pixels to be classified, cloud-free observations are available only in the first six composite periods (to name just one possible combination of 4095). To classify this set of compiled prediction data while avoiding data imputation, we rather ignore respective periods in the compiled training dataset. This means that the corresponding temporal subset (in our example the first six composite periods) is extracted from the compiled training dataset and then passed to RF. The trained model is then applied to the particular set of compiled prediction data for land cover classification.

Prior to pixel-based composition of the prediction data, the length of composite periods is maximally extended to the permissible 14 days, so that potentially further satellite images can be taken into account. A temporal extension of composite periods includes both previous and subsequent days equally, but avoids temporal overlaps with other composite periods. Closely spaced composite periods may therefore be shorter than 14 days.

3.4.4. Using RF's OOB error

Multiple RF models are computed within APiC to dynamically account for different satellite observation densities and temporal cloud coverage at pixel level. Just as each model is based on different temporal subsets of the compiled training dataset, a different combination of predictor variables was used for each model. Since predictor variables of each composite period have different effects on model performance, corresponding changes in OOB error estimates also occur for each model run. Hence, pixels are now assigned different OOB error values and the land cover map can be interpreted taking model error estimates into account.

Importance of composite periods

To capture land surface phenology as accurately as possible, we were aiming to maximise the number of composite periods within the classification year. We then let RF decide on their importance. In contrast to the variable importance, which RF generates by default, namely Mean Decrease Gini or Mean Decrease Accuracy, we wanted to analyze the impact of individual composite periods on model performance instead of referring to individual predictor variables, namely the spectral bands. The importance for a particular composite period and land cover class was determined by the class OOB error difference between the RF model based on all composite periods and the models where data from a particular period was not included. The difference was then averaged and normalized by the standard deviation of the differences. We have addressed the land cover classes individually in order to take the different phenological behaviours into account.

4. Results

4.1. Composite periods

We applied APiC for each landscape region separately, which is reflected in the different temporal arrangement of the composite periods for each region (Fig. 3). The composite periods established within the iterative process of pixel-based compositing (black boxes) usually extend to the maximum of 14 days but may be shorter in periods of low cloud coverage. In many cases, it is a single, largely cloud-free observation at the end of a composite period from which samples of the compiled training dataset originate (visualized as a long red line on the right side of a black box). However, the first composite period of the NW-Lowlands, for example, shows that 5000 pixels per land cover class can also be compiled equally from several observation dates. The fourth composite period for W-Uplands, on the other hand, was established based on one observation only. This example illustrates that without the extension of this composite period to 14 days (black + white boxes), additional satellite images could not have been used for the compilation of prediction data. The arrangement of composite period varies in each landscape region, which is most evident in spring, with the number of composite periods being lower in the southern regions (SW-Uplands and Alpine Foreland) than in the northern regions. For SW-Uplands only six composite periods could be established during the year, less than half the number compared to W-

Uplands (14 composite periods). In all regions no composite periods could be identified in January and February 2016.

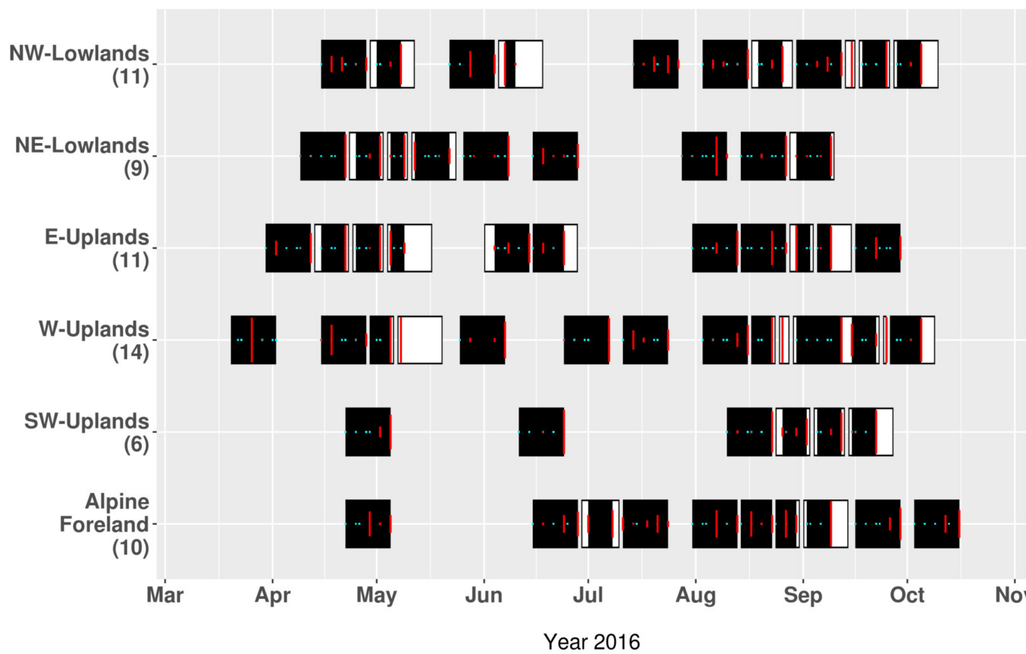

Fig. 3: Temporal arrangement of composite periods in six landscape regions. The total number of periods is given in brackets. Established composite periods are shown as black boxes. The cyan dots mark the date on which satellite observations were available during this period. The length of the red lines shows how many pixels from respective satellite observations of a composite period were included in the compiled training dataset. The black boxes + its adjacent white space correspond to the composite period length used for prediction. This only applies in cases where the black box comprises less than 14 days and could be extended to 14 days without a temporal overlap with subsequent and/or preceding composite periods. Since winter crops of the following year are already sown in autumn, possible composite periods for November and December are excluded from the classification and are not shown here.

4.2. Compiled training dataset and class imbalances

Table 1 lists the number of samples per land cover class of the compiled training dataset (column *Train*). Class imbalances become particularly evident between winter wheat or grassland as the most common land cover types and smaller classes such as leeks, strawberries or hops. Our analyses have shown that this would favor larger classes being classified at the expense of smaller classes. Therefore, it was our goal to systematically determine the appropriate class proportions in the compiled training dataset. Fig. 4 shows that the OOB error decrease exponentially as a function of increased imbalances (and increased sample size). Finally, we used the fourth subsample (marked by a vertical grey line) as the (reduced) compiled training dataset in our classification (third column in Table 1). The sample size has been reduced by at least 85% compared to the original training dataset, which accelerates the calculation of many RF models. The lower class imbalance in the reduced compiled training dataset leads to a more realistic class representation in the land cover map in average and to more balanced UA and PA results in the validation result.

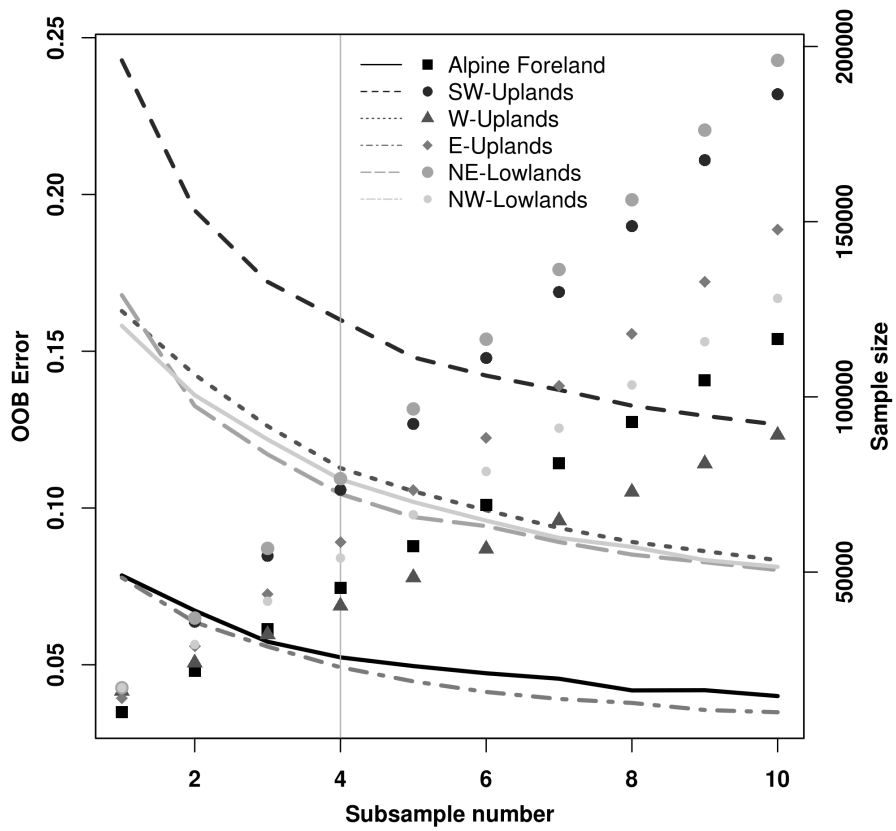

Fig. 4: The evolution of RF's OOB error (lines) and sample size (dots) between ten subsamples of the compiled training dataset with reduced class imbalances. The results for all six landscape regions are shown. At subsample number 1 all land cover classes are represented equally (1000 samples). The class imbalance of the original compiled training dataset is gradually approximated in the remaining 9 subsamples. The OOB error decrease exponentially as a function of increased imbalances. The vertical grey line marks the trade-off between reduced class imbalances and acceptable OOB error (corresponding approximately with the knee of the curves). This subsample will be used as the (reduced) compiled training dataset in our classification.

4.3. Multiple prediction models (= combinations of composite periods)

During pixel-based compositing of the prediction data it turned out that cloud-free pixel observations were often not available in all composite periods, but rather in different combinations of composite periods. For each combination, a temporal subset of the compiled training dataset was passed to RF to train individual prediction models. A comparison of the regional results thus shows that the number of computed prediction models grows exponentially with the number of established composite periods (Table 2). While the classification of the SW-Uplands region (six composite periods) is based on 63 prediction models (the maximum possible), 7291 models (45% of the maximum possible) are used to classify the region W-Uplands (14 composite periods).

Table 2: Number of RF models used per landscape region

Landscape region	**Number of RF models**
Alpine Foreland	1017
SW-Uplands	63
W-Uplands	7291
E-Uplands	1848
NE-Lowlands	511
NW-Lowlands	1990
Total	12720

4.4. Importance of composite periods

The calculated importance of composite period is shown in Fig. 5 for five landscape regions, four crop types (winter wheat, spring barley, rapeseed, sugar beets), stone fruits and the grassland class. We have also calculated the NDVI for each composite period and land cover type to interpret importance in relation to land cover phenology. For better illustration, composite periods are presented as single points in time in Fig. 5 (red and green dots) by calculating the weighted time average from respective observation dates.

For the classification of spring cereals, early observation periods in spring are most important, coinciding with the time of NDVI rise. In contrast, periods in early/mid summer when NDVI begins to decline are more relevant for winter cereals. This pattern can also be found in the model results of the other spring/winter cereal species. The plant growth of sugar beets and maize begins at about the same time, but thereafter the NDVI for sugar beets reaches its maximum values faster. The highest plant vitality for both crop types is reached in late summer, followed by a faster decline in NDVI for maize. The composite period at the beginning of plant growth is especially important for sugar beets, while for maize the periods a few weeks later are weighted higher. The NDVI for rapeseed usually drops briefly towards May. We explain this occurrence with the yellow rape blossom. According to this pronounced phenological event, the period is considered the most important. As expected, the NDVI values for grassland and stone fruits remain consistently high throughout the year. In contrast to other classes, here the estimated importance in relation to NDVI varies more strongly. Although higher importance values for both classes do not match any pronounced NDVI features, for stone fruits composite periods in spring are assigned usually more weight. For grassland, higher importance values are computed for periods in summer.

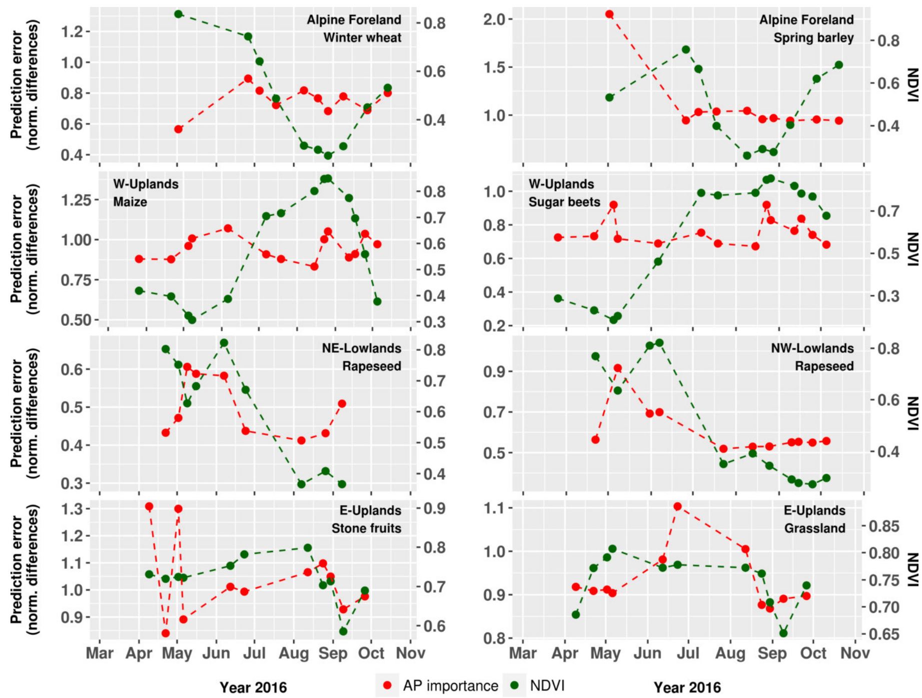

Fig. 5: Importance of composite periods (red dots) defined by the normalized differences of the prediction error between models for which respective composite period was omitted and the highest temporally resolved model. The higher the prediction error rates, the more important is the corresponding composite period for the overall model performance. The normalized difference vegetation index (NDVI) per composite period was also calculated (green dots). The evolution of importance and NDVI values over the year (red and green lines) are presented for five landscape regions, Alpine Foreland, W-Uplands, E-Uplands, NE-Lowlands, NW-Lowlands, and four crop types (winter wheat, spring barley, rapeseed, sugar beets), as well as the stone fruits and the grassland class.

4.5. Classification Accuracy

Fig. 6 shows the classification result for Germany and close-ups of four selected regions that were not covered by the IACS reference data. The classification map can be viewed at `http://ufz.maps.arcgis.com/apps/Styler/index.html?appid=84a36f4e815e4aa88f38a6d0f8382590` and downloaded at `http://PANGAEA...TO-BE-COMPLETED`. In general, single agricultural parcels are clearly identifiable in the land cover map, indicating that our classification well reproduces both, inter-field heterogeneity and intra-field homogeneity. Parcel sizes differ mainly between West- and East-Germany, while they are generally larger in the east (close-up 2). Winter wheat is predominantly cultivated in the Magdeburger Boerde (close-up 2) and in the Schleswig-Holstein Morainic Uplands (eastern part of close-up 1), whereas maize dominates the northwest and south regions (western part of close-up 1 and close-up 4). Sugar beets are mainly cultivated in the region of close-up 3.

A statistical validation of the classification result was performed by calculating PA and UA for individual land cover classes of the regions (Table 3). The overall accuracy and Kappa coefficient of the regions are also given in the table. The average overall accuracy over all regions accounts to around 88%. Despite the different number of composite periods, the overall accuracy differs only by a maximum of 4.38% between the regions.

Land cover classes with very high PA and UA (mainly $>=$ 90%) are grassland and winter wheat (with a tendency towards higher PA than UA) as well as maize, rapeseed and sugar beets (with the tendency towards higher UA than PA). Generally good results were achieved for the classes winter barley and hops (higher UA) and vines (higher PA). Good to moderate results

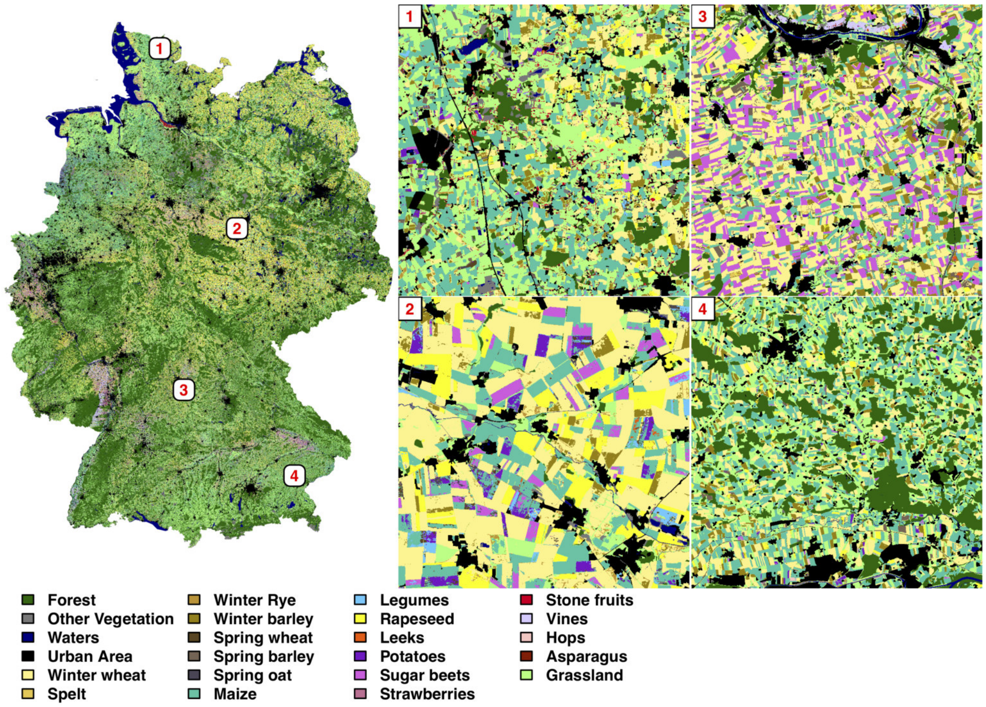

Fig. 6: Land cover map of Germany. In total 19 land cover classes were classified: winter wheat, spelt, winter rye, winter barley, spring wheat, spring barley, spring oat, maize, legumes, rapeseed, leeks, potatoes, sugar beets, strawberries, stone fruits, vines, hops, asparagus and grassland. The land-cover classes forest, other vegetation, urban area, and waters were taken from the ATKIS data base.

were achieved for spring barley and potatoes (rather higher UA) and without clear tendencies in UA/PA for legumes, leeks and asparagus. The stone fruits class received good UA (up to 87%) (but clearly worse in W-Uplands) with mostly lower PA (up to 65%). Spelt and winter rye also show much higher UA on average (up to 78%) than PA (rarely higher than 30% but with 65% quite high for winter rye in NE-Lowlands). On average, the classes spelt, spring wheat, spring oat and strawberries were assigned lowest accuracy. In regions where the classes stone fruits and grassland occur together, the former is usually classified as the latter, which is reflected in a low PA for stone fruits. The same applies to cereals, where low PA values of spelt, winter rye and spring wheat are mainly due to their misclassification as winter wheat.

Classification performance can vary substantially among the regions. Potatoes were classified with over 90% accuracy in the Alpine Foreland region, but only with 67% in the E-Uplands and mostly 70-80% in the other regions. Stone fruits' UA differs by almost 50% between the regions E-Uplands and W-uplands. Generally, the classes best reproduced show not only the most balanced results between PA and UA, but they are also very stable across all regions.

Table 3: Classification Accuracy

	Alpine Foreland		SW-Uplands		W-Uplands		E-Uplands		NE-Lowlands		NW-Lowlands	
land cover classes	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Winter wheat	92.89	89.77	91.16	86.43	89.63	86.90	90.76	89.37	89.61	80.52	88.45	83.79
Spelt			12.93	48.45	21.99	31.70	17.07	71.51	30.65	69.90	23.15	29.59
Winter rye			31.37	63.47	18.42	44.25	14.80	72.89	64.71	77.66	38.84	78.67
Winter barley	71.82	83.55	70.37	74.31	71.93	86.43	78.29	80.98	64.44	90.49	74.30	79.27
Spring wheat					18.05	23.03	14.88	74.52	39.82	33.46	10.43	25.39
Spring barley	49.89	84.73	84.22	85.63	59.93	63.03	83.79	74.42	50.14	56.18	50.82	61.04
Spring oat			48.97	57.59	44.15	32.12	40.46	57.75	30.51	40.03	30.79	24.69
Maize	89.56	92.12	84.33	89.34	89.70	94.28	93.96	95.78	94.38	91.97	95.54	94.42
Legumes			59.66	60.22	58.08	61.43	66.18	81.02	73.68	68.96	54.91	44.56
Rapeseed	72.64	84.13	90.79	91.81	92.50	92.81	95.55	95.97	94.36	98.43	90.76	96.30
Leeks			57.45	57.68					75.54	25.41	47.84	41.41
Potatoes	93.70	95.56	71.13	77.34	71.99	54.53	67.41	66.87	54.04	71.88	78.37	84.84
Sugar beets	93.21	96.19	93.83	90.14	91.57	90.17	86.83	92.15	85.36	88.67	83.35	94.17
Strawberries			32.05	28.61	77.30	22.58			55.02	13.33	55.11	37.06
Stone fruits	41.90	62.85	27.81	67.52	45.97	37.32	34.03	87.05	34.31	82.34	64.79	68.67
Vines			94.90	93.97	82.04	64.29						
Hops	77.41	85.86										
Asparagus			62.31	66.84					46.90	44.52	38.37	43.54
Grassland	96.40	90.45	90.94	84.19	97.00	94.94	97.56	91.11	96.68	88.70	96.99	89.69
Overall Accuracy	**90.41**		**86.14**		**89.44**		**89.35**		**86.03**		**87.39**	
Kappa Coefficient	**0.876**		**0.838**		**0.854**		**0.866**		**0.831**		**0.842**	

4.6. Model prediction error

Fig. 7 shows the spatial distribution of the OOB error over Germany. Some regions are characterized by generally lower (W-Uplands, E-Uplands) or higher (SW-Uplands) prediction errors and thus stand out clearly. Cross-regional features are three strips of high OOB errors, narrowing from southwest to northeast according to the satellite orbit. Adjacent swaths do not overlap in these areas and therefore only one image is captured per satellite's orbit cycle. On closer inspection, high OOB errors can be also attributed to individual cloud patterns (close-up 1 and 2).

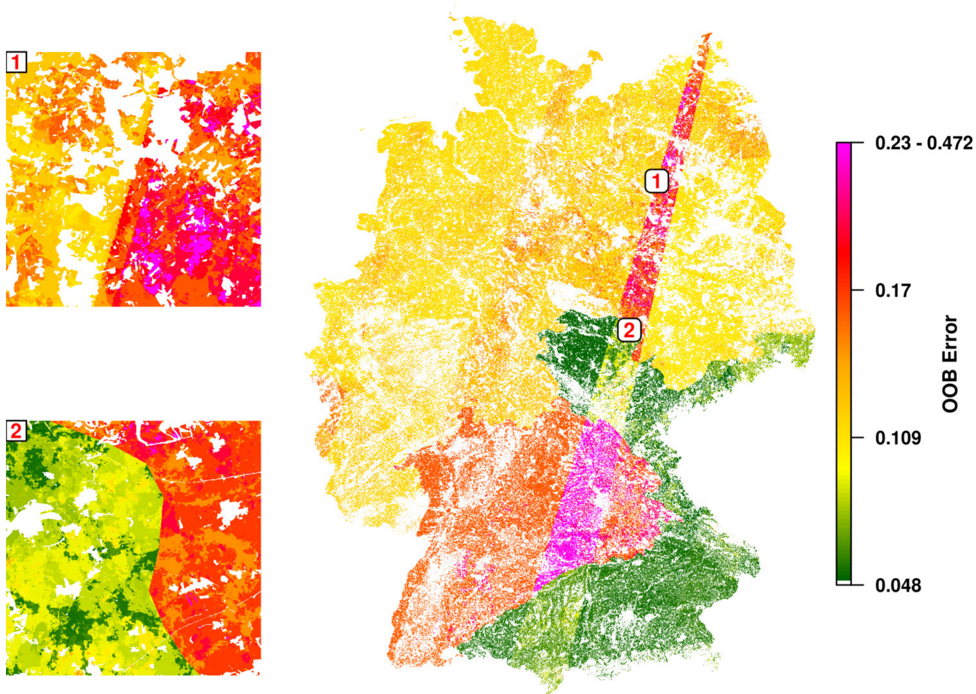

Fig. 7: The mapped averaged OOB (Out-Of-Bag) error for Germany. The OOB error originates from various random forest models, each trained with a different temporal subset of the compiled training dataset. Differences in the overall prediction error between the landscape regions, e.g. between Alpine Foreland and SW-Uplands, are clearly visible. The "stripes" with higher OOB errors running from northeast to southwest indicate areas with no overlap of adjacent satellite tracks and hence fewer satellite observations. Different OOB errors within regions are also due to clouds.

To better assess the significance of the class OOB error (based on modeling), we have investigated its relationship to classification accuracy (based on reference data) (Fig. 8). Here the class OOB errors of the different prediction models (grey dots) are compared with the producer accuracy from Table 3.

Per land cover class the value range of the OOB error is large (gray dots) except for the classes that achieved highest PA. However, the distribution of the class OOB error is strongly skewed (less so for SW-Uplands) to its lower values (higher accuracy) as shown by the averaged class errors (black dots). In all regions, higher PA is well represented by the averaged class OOB errors, while with decreasing PA the errors are usually underestimated (offset to the 1:1 line). The slope of the regression line (black line) for SW-Uplands corresponds most closely to the 1:1 line. Relative PA differences between classes are generally well reproduced by the class OOB error, as indicated by the given R^2 value, which represents the average coefficient of determination of each linear model.

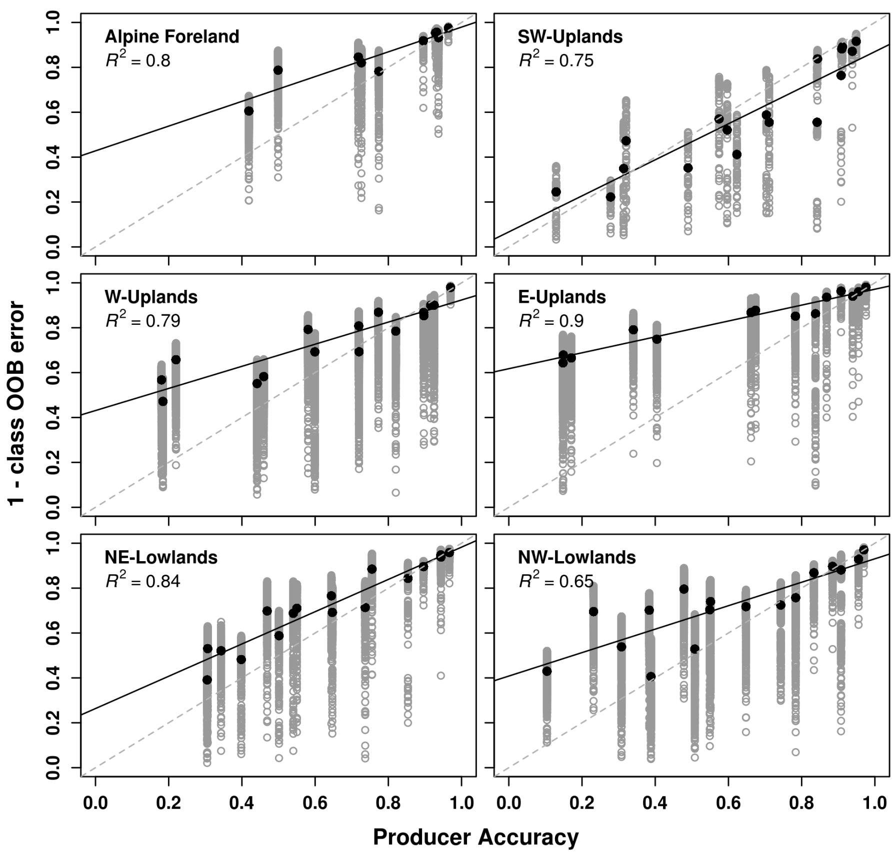

Fig. 8: Modelled class accuracy versus validation's producer accuracy for all landscape regions. Modelled class accuracy is expressed by $1 - class\ OOB\ error$. The grey dots represent the different class OOB errors of the random forest models. The averaged error values per land cover class are visualized as black dots and the corresponding linear model is shown as black line. The averaged value of the correlation coefficients (R^2) of the linear models of all model runs is given.

5. Discussion

The high spatial and temporal resolution of Sentinel-2 poses the challenge of how to use information from complex data for land cover classification, and specifically, how to deal with the obstacle of frequent cloud cover that hinders optical remote sensing worldwide. Therefore, our motivation was to develop a data-driven classification method based solely on measured reflectance values. Thus in APiC i) the establishment of composite periods is a dynamic process and involves the compilation of non-sparse training data, ii) the data availability at pixel level determines the total number of prediction models to be computed.

In previous studies, a single prediction model was used to classify a time series of seamless, cloud-free image composites of the entire study area. At shorter time intervals, cloud-free pixel observations are increasingly missing, which were then calculated by statistical data imputation methods. However, the effect of imputed data on the classifier's performance remains usually unclear, even though the number of clear-sky observations for each pixel has been reported in some studies (Frantz et al., 2017; Griffiths et al., 2019). To our knowledge, the number of interpolated data points per time interval and land cover class has not yet been reported, but would nevertheless limit the interpretation of the classification result (and the importance of time intervals used). The results of this study demonstrate, that a data-driven and dynamic approach at pixel level allows qualitative conclusions to be drawn about the predictive power of classification models, which go beyond mere data availability.

5.1. Regionalisation and composite periods

Especially in continental or global classification studies biogeographical characteristics of a region should be taken into account as they determine the phenology of a plant community. In other parts of the world, cloud cover may be more frequent, species composition more diverse, and phenological cycles more complex, contradicting a standardised classification procedure. As a result, there can be no general solution for fixed or predefined temporal intervals. For this reason, we have introduced (adaptable) composite periods that are tailored to respective cloud-free satellite observations and reference data availability of the study site.

The separate classification of six landscape regions has demonstrated that our methodology can be used in an operational framework for regionalised studies outside Germany, since the user only decides on the maximum length of composite periods and the minimum sample size of land cover classes in the training dataset. Thereafter, the definition of composite periods is automated. This flexibility of APiC was shown in Fig. 3, where for Alpine Foreland and SW-Uplands only one composite period was defined in spring and thus less than in other regions. This can be attributed to different weather conditions and/or the lower amount of reference data. Nevertheless, no decrease in classification accuracy was observed for both regions. It appears that phenological differences between land cover types, which are more pronounced in spring, have been well captured in this single composite period. It also shows that the results are not determined by the quantity but by the spatial distribution of the reference data and therefore by the regional representativeness of the compiled training dataset.

As with other classification approaches, APiC is expected to perform poorly in regions with very heavy cloud cover such as the tropics. In such cases, the region to be classified should be extended to less cloudy areas, even if they have different biogeographical characteristics. Depending on the availability of reference data, this can significantly increase the density of composite periods. In areas with high cloud coverage, higher model prediction errors are then assigned to the classification result.

5.2. Classification accuracy

Our classification result has been extensively validated against the large IACS dataset. Depending on the land cover class, between 2.8% (asparagus, NW-Lowlands) and and 65% (spelt, SW-Uplands) of the reference data were used for the compiled training dataset before it was reduced via LHS. The samples in the reduced compiled training dataset, which was finally passed to RF for the computation of prediction models, were based on only 0.04% (maize, NW-Lowlands) or 8% (stone fruits, Alpine Foreland) of the reference data. Compared to standard validation methods, which are rather based on 30% of the data while 70% are used for training, our classification performance has been reviewed more extensively.

The good validation results achieved for maize and sugar beets are certainly due to the relatively late sowing date between mid-April to mid-May and the late ripening phase. On the contrary, phenological and morphological similarities among the cereal types hamper their spectral differentiation, resulting in lower classification accuracy for the smaller classes spelt, winter rye, spring wheat and spring oat. Smaller parcel sizes may also have affected classification accuracy as the risk of mixed pixels is increased. Potatoes, strawberries, leeks and asparagus are often grown on fields that are not or barely larger than a Sentinel-2 pixel, mixing spectral properties of the adjacent land cover in the recorded signal. This hampers both, the compilation of representative training data and subsequent land cover prediction.

5.3. *Class imbalances*

Class imbalances in the reference data have a strong effect on classification accuracy, with the larger classes having the greatest impact on overall accuracy. Furthermore, dominant land cover classes in the training data are classified at the expense of smaller classes. Overrepresentation of larger classes in a land cover map mainly affects the validation results (lower PA) of the smaller class. We observed such effects, for example between grassland, the larger class, and stone fruits, the smaller class, which were often confused due to their related species composition and spectral similarities. In the case of balanced class proportions (not shown in the results section), PA could be improved for stone fruits, but only with a concurrent decrease in UA. For grassland, only minor changes in accuracy were noticed. We used an empirical approach to find a reasonable level of class imbalances in the training dataset that balances UA and PA well for most land cover classes (Janitza and Hornung, 2018; Stumpf and Kerle, 2011). The proposed procedure uses LHS to reduce the number of samples in the compiled training dataset while preserving the original spectral variance. The reduced dataset size accelerates the runtime of RF, which is advantageous for the calculation of multiple prediction models.

5.4. *Multiple prediction models*

Our dynamic classification approach uses multiple prediction models at pixel level, which is more computationally intensive than using a single model based on entire cloud-free image mosaics. For example, having 14 composite periods established for a region, a maximum of $2^{14} - 1 = 16383$ model runs (= combinations of composite periods) may be necessary to classify the total area. However, on a Linux-based computing cluster with a total of 2564 cores, 25.8 TB RAM and a parallel high performance file system, computing time was kept within days. We actually turned the alleged disadvantage to our advantage by relating the classified land cover to the mapped model prediction error. Additionally, we used the class OOB error for calculating the importance of composite periods.

5.5. Model prediction error

Comparing model performance with the PA from the validation revealed that our RF models mostly overfit. We assume that higher number of composite period come with increased (multi-) collinearity between the predictor variables and thus favoring overfitting (Dormann et al., 2013; Rodriguez-Galiano et al., 2012; Shih et al., 2019). The overfitting applies in particular to the smaller classes with lower PA and has therefore only slightly affected the overall accuracy of the validation result. Nevertheless, it could also be shown that there is a clear relationship between modelled class accuracy and PA. This can be useful for continental or global applications where validation data is insufficient or may even be missing. However, due to the different degree of overfitting, a cross-regional comparison of the OOB error is not always meaningful. Fig. 7 gives the impression that SW-Uplands has been classified worst, which could not be verified by our validation. Rather, it is the only region where almost no overfitting has been observed.

5.6. Importance of composite periods

The presence of highly correlated predictors impacts the importance measure of single variables (Gregorutti et al., 2017; Strobl et al., 2007). Likewise, in our study, the interpretation of importance becomes more difficult with a higher number of composite periods. Here, collinearity is certainly the main reason why consecutive composite periods often showed similar importance scores. A comparison of the importance measure between the regions should take into account the different composition of land cover classes, number of composite periods and their temporal arrangement. Nonetheless, we were able to show that i) class specific traits occur across regions, ii) closely spaced composite periods may have significant differences in their importance and iii) composite periods established at times of photosynthetic change are usually of higher importance. In this respect, we can draw the conclusion that it is indeed worth to maximise composite periods within the classification year to ensure the detection of important phenological events (such as rape flowering). Although composite periods in spring and early summer tend to have higher importance, there is no evidence that months of other seasons per se can be neglected. It should be left to the classifier how composite periods are weighted according to region-specific conditions. However, to keep the number of predictor variables low, composite periods with consistently low importance can be excluded successively in subsequent classification runs. Whether this counteracts overfitting and thus leads to better classification accuracy with more realistic OOB error estimates has to be analyzed in a follow-up study.

6. Conclusions and outlook

In this work we presented a highly automated pixel-based compositing and classification approach that was used to produce thematically detailed land cover maps in six landscape regions. The agricultural area of Germany was thus classified into a total of 19 land cover classes. APiC works largely data-driven, making it easily applicable to other study sites with different reference data (data extent and land cover composition), regional cloud coverage and satellite data availability. Time windows in which cloud-free satellite observations are used for classification adapt to these conditions and rely on only a few user-defined specifications. The classification result shown is based on more than 10,000 individual classification models, which allow the spatial representation of the estimated prediction error in addition to the actual land cover. While a high number of composite periods is necessary to detect relevant phenological phases, RF models might overfit with too many predictor variables (Karpatne et al., 2016), leading to highly optimistic OOB error estimates. The effect of collinearity has already been investigated for a large number of algorithms (Dormann et al., 2013). It should now be further investigated how other classifiers and their internally calculated prediction error estimates behave given a similar spectral data set.

The new high-resolution thematic map can be used to analyse land cover changes and intensities in more detail than before. The associated ecological issues such as nutrient fluxes, pollination and insect mortality could thus be addressed more comprehensively. An answer to these questions is urgent and must be given across borders, so that an upscale of the classification to continental level is necessary. Currently, such an approach is hampered by the lack of or limited access to reference data in landscape regions with different biogeographical characteristics. Upcoming German-wide classifications for the years 2017+ will differ in that additional observations from the Sentinel-

2B satellite will be available. Analyzes will show whether and to what extent denser time series have an impact on the establishment of composite periods and resulting classification accuracy. In future studies the APiC concept can also be applied to other types of land cover classifications. For example, a map of agricultural land cover classes in combination with the most common tree species would open up new opportunities in many scientific areas such as ecological modelling and ecosystem services and will certainly be of great interest to farmers, forest managers and policy makers.

Acknowledgements

This research was funded (FKZ: 3517860800) by the Federal Agency for Nature Conservation (BfN) and the integrated project initiative of the Helmholtz Centre for Environmental Research - UFZ. We thank the federal states for providing the IACS data sets.

References

Billeter, R., Liira, J., Bailey, D., Bugter, R., Arens, P., Augenstein, I., Aviron, S., Baudry, J., Bukacek, R., Burel, F., Cerny, M., De Blust, G., De Cock, R., Diekotter, T., Dietz, H., Dirksen, J., Dormann, C., Durka, W., Frenzel, M., Hamersky, R., Hendrickx, F., Herzog, F., Klotz, S., Koolstra, B., Lausch, A., Le Coeur, D., Maelfait, J.P., Opdam, P., Roubalova, M., Schermann, A., Schermann, N., Schmidt, T., Schweiger, O., Smulders, M.J.M., Speelmans, M., Simova, P., Verboom, J., van Wingerden, W.K.R.E., Zobel, M., Edwards, P.J., 2008. Indicators for biodiversity in agricultural landscapes: a pan-European study. JOURNAL OF APPLIED ECOLOGY 45, 141–150. doi:10.1111/j.1365-2664.2007.01393.x.

Bleyhl, B., Baumann, M., Griffiths, P., Heidelberg, A., Manvelyan, K., Radeloff, V.C., Zazanashvili, N., Kuemmerle, T., 2017. Assessing landscape connectivity for large mammals in the caucasus using landsat 8 seasonal image composites. Remote Sensing of Environment 193, 193 – 203. URL: https://doi.org/10.1016/j.rse.2017.03.001.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees. Chapman & Hall, London.

Cihlar, J., 2000. Land cover mapping of large areas from satellites: Status and research priorities. International Journal of Remote Sensing 21, 1093–1114. doi:10.1080/014311600210092.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37, 35 – 46. URL: https://doi.org/10.1016/0034-4257(91)90048-B.

DeFries, R.S., Rudel, T., Uriarte, M., Hansen, M., 2010. Deforestation driven by urban population growth and agricultural trade in the twenty-first century. NATURE GEOSCIENCE 3, 178–181. doi:10.1038/NGEO756.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., Marquez, J.R.G., Gruber, B., Lafourcade, B., Leitao, P.J., Munkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27–46.

Dormann, C.F., McPherson, J.M., Araujo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., Schurr, F.M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30, 609–628.

Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Coe, M.T., Daily, G.C., Gibbs, H.K., Helkowski, J.H., Holloway, T., Howard, E.A., Kucharik, C.J., Monfreda, C., Patz, J.A., Prentice, I.C., Ramankutty, N., Snyder, P.K., 2005. Global consequences of land use. Science 309, 570–574. 0036-8075.

Frantz, D., Rder, A., Stellmes, M., Hill, J., 2017. Phenology-adaptive pixel-based compositing using optical earth observation imagery. Remote Sensing of Environment 190, 331 – 347. URL: https://doi.org/10.1016/j.rse.2017.01.002.

Gomez, C., White, J.C., Wulder, M.A., 2016. Optical remotely sensed time series data for land cover classification: A review. ISPRS JOURNAL OF PHOTOGRAMMETRY AND REMOTE SENSING 116, 55–72. doi:10.1016/j.isprsjprs.2016.03.008.

Gregorutti, B., Michel, B., Saint-Pierre, P., 2017. Correlation and variable importance in random forests. Statistics and Computing 27, 659–678. doi:10.1007/s11222-016-9646-1.

Griffiths, P., van der Linden, S., Kuemmerle, T., Hostert, P., 2013. Pixel-Based Landsat Compositing Algorithm for Large Area Land Cover Mapping. IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING 6, 2088–2101. doi:10.1109/JSTARS.2012.2228167.

Griffiths, P., Nendel, C., Hostert, P., 2019. Intra-annual reflectance composites from sentinel-2 and landsat for national-scale crop and land cover mapping. Remote Sensing of Environment 220, 135 – 151. URL: https://doi.org/10.1016/j.rse.2018.10.031.

Hadley, A.S., Betts, M.G., 2012. The effects of landscape fragmentation on pollination dynamics: absence of evidence not evidence of absence.

Biological Reviews 87, 526–544. doi:10.1111/j.1469-185X.2011.00205.x.

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. doi:10.1126/science.1244693.

Holben, B.N., 1986. Characteristics of maximum-value composite images from temporal avhrr data. International Journal of Remote Sensing 7, 1417–1434. doi:10.1080/01431168608948945.

Houghton, R.A., 2010. How well do we know the flux of CO2 from land-use change? TELLUS SERIES B-CHEMICAL AND PHYSICAL METEOROLOGY 62, 337–351. doi:10.1111/j.1600-0889.2010.00473.x.

IFAG, 1979. Karte der Bundesrepublik Deutschland 1:1.000.000 - Landschaften (Namen und Abgrenzungen). IFAG - Institut für angewandte Geodäsie, Frankfurt/Main.

Janitza, S., Hornung, R., 2018. On the overestimation of random forests out-of-bag error. PLOS ONE 13, 1–31. doi:10.1371/journal.pone.0201904.

Joshi, N., Baumann, M., Ehammer, A., Fensholt, R., Grogan, K., Hostert, P., Jepsen, M.R., Kuemmerle, T., Meyfroidt, P., Mitchard, E.T.A., Reiche, J., Ryan, C.M., Waske, B., 2016. A Review of the Application of Optical and Radar Remote Sensing Data Fusion to Land Use Mapping and Monitoring. REMOTE SENSING 8. doi:10.3390/rs8010070.

Karpatne, A., Jiang, Z., Vatsavai, R.R., Shekhar, S., Kumar, V., 2016. Monitoring land-cover changes: A machine-learning perspective. IEEE Geoscience and Remote Sensing Magazine 4, 8–21. doi:10.1109/MGRS.2016.2528038.

Lancashire, P.D., Bleiholder, H., Boom, T.V.D., Langelueddeke, P., Stauss, R., Weber, E., Witzenberger, A., 1991. A uniform decimal code for growth stages of crops and weeds. Annals of Applied Biology 119, 561–601. doi:10.1111/j.1744-7348.1991.tb04895.x.

Liaw, A., Wiener, M., 2013. Package 'randomForest'. Breiman and Cutler's random forests for classification and regression. URL: http://cran.r-project.org/web/packages/randomForest/randomForest.pdf. last accessed 28 10 2014.

Louis, J., Debaecker, V., Pflug, B., Main-Knorn, M., Bieniarz, J., Mueller-Wilm, U., Cadau, E., Gascon, F., 2016. Sentinel-2 sen2cor: L2a processor for users, in: Ouwehand, L. (Ed.), ESA Living Planet Symposium 2016, Spacebooks Online. pp. 1–8. URL: https://elib.dlr.de/107381/.

Lueck, W., van Niekerk, A., 2016. Evaluation of a rule-based compositing technique for landsat-5 tm and landsat-7 etm+ images. International Journal of Applied Earth Observation and Geoinformation 47, 1 – 14. URL: https://doi.org/10.1016/j.jag.2015.11.019.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239–245. URL: http://www.jstor.org/stable/1268522.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21, 1087–1092. URL: https://doi.org/10.1063/1.1699114, doi:10.1063/1.1699114, arXiv:https://doi.org/10.1063/1.1699114.

Meynen, E., Schmidthüsen, J., Gellert, J., Neef, E., Müller-Miny, H., Schultze, J.H., 1953-62. Handbuch der naturräumlichen Gliederung Deutschlands. Bundesanstalt für Landeskunde und Raumforschung, Bad Godesberg.

Minasny, B., McBratney, A.B., 2006. A conditioned latin hypercube method for sampling in the presence of ancillary information. Computers Geosciences 32, 1378 – 1388. URL: http://www.sciencedirect.com/science/article/pii/S009830040500292X, doi:https://doi.org/10.1016/j.cageo.2005.12.009.

Newbold, T., Hudson, L.N., Hill, S.L.L., Contu, S., Lysenko, I., Senior, R.A., Boerger, L., Bennett, D.J., Choimes, A., Collen, B., Day, J., De Palma, A., Diaz, S., Echeverria-Londono, S., Edgar, M.J., Feldman, A., Garon, M., Harrison, M.L.K., Alhusseini, T., Ingram, D.J., Itescu, Y., Kattge, J., Kemp, V., Kirkpatrick, L., Kleyer, M., Correia, D.L.P., Martin, C.D., Meiri, S., Novosolov, M., Pan, Y., Phillips, H.R.P., Purves, D.W., Robinson, A., Simpson, J., Tuck, S.L., Weiher, E., White, H.J., Ewers, R.M., Mace, G.M., Scharlemann, J.P.W., Purvis, A., 2015. Global effects of land use on local terrestrial biodiversity. NATURE 520, 45+. doi:10.1038/nature14324.

Roberts, D., Mueller, N., Mcintyre, A., 2017. High-dimensional pixel composites from earth observation time series. IEEE Transactions on Geoscience and Remote Sensing 55, 6254–6264. doi:10.1109/TGRS.2017.2723896.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. Isprs Journal of Photogrammetry and Remote Sensing 67, 93–104. URL: https://doi.org/10.1016/j.isprsjprs.2011.11.002.

Roudier, P., 2011. clhs: a R package for conditioned Latin hypercube sampling.

Roy, D.P., Ju, J., Kline, K., Scaramuzza, P.L., Kovalskyy, V., Hansen, M., Loveland, T.R., Vermote, E., Zhang, C., 2010. Web-enabled Landsat Data (WELD): Landsat ETM plus composited mosaics of the conterminous United States. REMOTE SENSING OF ENVIRONMENT 114, 35–49. doi:10.1016/j.rse.2009.08.011.

Shih, H.C., Stow, D.A., Tsai, Y.H., 2019. Guidance on and comparison of machine learning classifiers for landsat-based land cover and land use mapping. International Journal of Remote Sensing 40, 1248–1274. doi:10.1080/01431161.2018.1524179.

Smith, P., House, J.I., Bustamante, M., Sobock, J., Harper, R., Pan, G., West, P.C., Clark, J.M., Adhya, T., Rumpel, C., Paustian, K., Kuikman, P., Cotrufo, M.F., Elliott, J.A., McDowell, R., Griffiths, R.I., Asakawa, S., Bondeau, A., Jain, A.K., Meersmans, J., Pugh, T.A.M., 2016. Global change pressures on soils from land use and management. Global Change Biology 22, 1008–1028. doi:10.1111/gcb.13068.

Sombroek, W., 2001. Spatial and temporal patterns of Amazon rainfall - Consequences for the planning of agricultural occupation and the protection of primary forests. AMBIO 30, 388–396. URL: https://doi.org/10.1579/0044-7447-30.7.388.

Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25. doi:10.1186/1471-2105-8-25.

Stumpf, A., Kerle, N., 2011. Object-oriented mapping of landslides using random forests. Remote Sensing of Environment 115, 2564 – 2577. URL: https://doi.org/10.1016/j.rse.2011.05.013.

White, J.C., Wulder, M.A., Hobart, G.W., Luther, J.E., Hermosilla, T., Griffiths, P., Coops, N.C., Hall, R.J., Hostert, P., Dyk, A., Guindon, L., 2014. Pixel-Based Image Compositing for Large-Area Dense Time Series Applications and Science. CANADIAN JOURNAL OF REMOTE SENSING 40, 192–212. doi:10.1080/07038992.2014.945827.

Witzenberger, A., Boom, T.v.d., Hack, H., 1989. Explanations of the bbch-decimal code for the development stages of cereals - with illustrations. Gesunde Pflanzen 41, 384–388.

Xu, X., Conrad, C., Doktor, D., 2017. Optimising Phenological Metrics Extraction for Different Crop Types in Germany Using the Moderate Resolution Imaging Spectrometer (MODIS). REMOTE SENSING 9. doi:10.3390/rs9030254.

List of Figures

Fig. 1:

The landscape regions of Germany (from South to North, black lines): Alpine Foreland (1), SW-Uplands (2), W-Uplands (3), E-Uplands (4), NE-Lowlands (5) and NW-Lowlands (6). Around 25% of Germany and thus approx. 50% of the total agricultural area is covered by reference data (IACS) (grey + magenta). Reference data used for pixel-based compositing of the training data is shown in magenta. The grey colored areas were used for validation.

Fig. 2:

Flow chart of the proposed adaptable pixel-based compositing and classification approach (APiC). During pixel-based compositing of training data (i.e. related to cloud-free Sentinel-2A pixels that are covered by reference data) composite periods are defined within an iterative process. At its end, composite periods were created whose length and temporal arrangement are adapted to the cloud cover in the satellite data and to the land cover information in the reference data. Additionally, a training dataset with a minimum sample size per land cover class has been compiled. For each composite period, reflectance values must be available in the compiled training data set (non-sparse). Due to class imbalances and excessive data volumes, the size of the training dataset is reduced via LHS. Pixel-based compositing of prediction data is based only on composite periods in which cloud-free pixel observations are available. According to these combinations of composite periods, temporal subsets are extracted from the compiled training dataset and passed to random forest. The number of prediction models to be computed therefore reflects the satellite observation density and/or temporal cloud coverage at pixel level. The OOB error output of each random forest model is used to create a map of prediction error estimates that complements the final land cover map. The windows marked with the letters A, B, C illustrate respective terms in the flow chart.

Fig. 3:

Temporal arrangement of composite periods in six landscape regions. The total number of periods is given in brackets. Established composite periods are shown as black boxes. The cyan dots mark the date on which satellite observations were available during this period. The length of the red lines shows how many pixels from respective satellite observations of a composite period were included in the compiled training dataset. The black boxes + its adjacent white space correspond to the composite period length used for prediction. This only applies in cases where the black box comprises less than 14 days and could be extended to 14 days without a temporal overlap with subsequent and/or preceding composite periods. Since winter crops of the following year are already sown in autumn, possible composite periods for November and December are excluded from the classification and are not shown here.

Fig. 4:

The evolution of RF's OOB error (lines) and sample size (dots) between ten subsamples of the compiled training dataset with reduced class imbalances.

The results for all six landscape regions are shown. At subsample number 1 all land cover classes are represented equally (1000 samples). The class imbalance of the original compiled training dataset is gradually approximated in the remaining 9 subsamples. The OOB error decrease exponentially as a function of increased imbalances. The vertical grey line marks the trade-off between reduced class imbalances and acceptable OOB error (corresponding approximately with the knee of the curves). This subsample will be used as the (reduced) compiled training dataset in our classification.

Fig. 5:

Importance of composite periods (red dots) defined by the normalized differences of the prediction error between models for which respective composite period was omitted and the highest temporally resolved model. The higher the prediction error rates, the more important is the corresponding composite period for the overall model performance. The normalized difference vegetation index (NDVI) per composite period was also calculated (green dots). The evolution of importance and NDVI values over the year (red and green lines) are presented for five landscape regions, Alpine Foreland, W-Uplands, E-Uplands, NE-Lowlands, NW-Lowlands, and four crop types (winter wheat, spring barley, rapeseed, sugar beets), as well as the stone fruits and the grassland class.

Fig. 6:

Land cover map of Germany. In total 19 land cover classes were classified: winter wheat, spelt, winter rye, winter barley, spring wheat, spring barley, spring oat, maize, legumes, rapeseed, leeks, potatoes, sugar beets, strawberries, stone fruits, vines, hops, asparagus and grassland. The land-cover classes forest, other vegetation, urban area, and waters were taken from the ATKIS data base.

Fig. 7:

The mapped averaged OOB (Out-Of-Bag) error for Germany. The OOB error originates from various random forest models, each trained with a different temporal subset of the compiled training dataset. Differences in the overall prediction error between the landscape regions, e.g. between Alpine Foreland and SW-Uplands, are clearly visible. The "stripes" with higher OOB errors running from northeast to southwest indicate areas with no overlap of adjacent satellite tracks and hence fewer satellite observations. Different OOB errors within regions are also due to clouds.

Fig. 8:

Modelled class accuracy versus validation's producer accuracy for all landscape regions. Modelled class accuracy is expressed by $1 - class\ OOB\ error$. The grey dots represent the different class OOB errors of the random forest models. The averaged error values per land cover class are visualized as black dots and the corresponding linear model is shown as black line. The averaged value of the correlation coefficients (R^2) of the linear models of all model runs is given.