Warsaw University of Life Sciences WULS – SGGW
in Warsaw
Faculty of Forestry

Eberswalde University for Sustainable Development – HNEE
University of Applied Sciences
Faculty of Forest and Environment

Tobias Seydewitz
Album number SGGW: 178311
Album number HNEE: 15210024

Kompleksowa analiza wylesiania w krajach tropikalnych - bezpośrednie czynniki wylesiania, emisje dwutlenku węgla i równowaga wartości usług ekosystemów

A comprehensive study on deforestation in the tropics - direct deforestation drivers, carbon emissions and ecosystem service value balance

Master's Thesis
on the course of - Forestry

Thesis written under the supervision of
Dr. Prajal Pradhan
Potsdam Institute of Climate Impact Research
Research Domain II - Climate Climate Impacts & Vulnerabilities

Potsdam, 2019

Contents

1 Data and methods

In this chapter, we describe our approach to answer the scientific questions stated during the introduction. The first section of this chapter introduces the datasets used during this study the underlying causes of using it. For each dataset, we shortly describe by which approach it is derived and what are the fundamental meta-data properties. Additionally, if possible we try to give for each dataset an accuracy assessment ideally prepared by other scientists or by the research group itself. Finally, we describe our idea behind using the data and how we acquired and filtered it. The second and last section of this chapter is focused on the applied methodology to prepare our analysis and results. For each processing step we give a short description of the methodical background and describe the core functionality of our processing algorithms as well we describe the required steps to achieve the results. For implementing our processing algorithms and visualizing our results we selected individually the technology which fulfills best the requirements. These approaches are encapsulated in a reusable software design to easily reproduce, alter or reuse our algorithms and findings.

1.1 Data

Table 1.1. Datasets used during this study: The source column contains a reference to the corresponding dataset launch publication. If the data is provided as a download the bibliographic reference contains the download URL.

Data	Type	Source
Global Forest Change	spatial	Hansen et al. [2013]
GlobeLand30	spatial	Chen et al. [2015]
Intact Forest Landscape	spatial	Potapov et al. [2017]
Aboveground Woddy Biomass	spatial	Baccini et al. [2015]
Global Soil Organic Carbon Content	spatial	FAO and ITPS [2018]
Global Administrative Areas	spatial	Hijmans et al. [2018]
Soil Organic Carbon Change	empirical	Don et al. [2010]
		de Groot et al. [2012]
Ecosystem Service Values	empirical	Costanza et al. [2014]
		Siikamaki et al. [2015]

Table 1.1 shows a comprehensive overview of the applied datasets for this study. Spatial datasets comprises vector as well raster data and empirical data is extracted from the cited

publications. The subsequent sections describe detailed each dataset following the order of appearance in the table.

1.1.1 Global Forest Change

Global Forest Change (GFC) 2000-2012 Version 1.0 is the first high-resolution dataset that provides a comprehensive view of the annual global forest cover change between 2000 and 2012 [Hansen et al. 2013; Li et al. 2017]. The initial GFC dataset released by Hansen et al. is extended by recent releases which encompass the annual forest cover changes between 2000-2013, 2000-2014, 2000-2015 and 2000-2016, respectively. All versions of this dataset have in common, that they are derived from growing season imagery captured by the remote sensing satellite Landsat 7 Enhanced Thematic Mapper Plus (ETM+) enhanced by band metrics of other sensors like Quickbird imagery, existing percent tree cover layers from Landsat data, and global Moderate Resolution Imaging Spectroradiometer (MODIS) percent tree cover [Hansen et al. 2013]. On the satellite imagery, a time-series spectral metrics analysis is applied to gather the global forest extent at 2000 as well as the annual forest loss and the accumulated gain for the period 2001 till 2012. Hence, GFC comprises three independent data layers tree cover, annually forest loss and forest gain divided into 10x10 degree tiles by the geodetic coordinate system World Geodetic System 1984 (WGS84) (EPSG:4326) with a spatial resolution of 1 arc-second per pixel (approximately 900 Km^2 or 30x30 m). Furthermore, across the provided Geo-Tiff (GTiff) layers the pixel data is coded in unsigned 8-bit integers. Hansen et al. defined trees as all vegetation taller than 5 meters for their study. For each pixel covered by trees, a canopy density ranging from 0 t0 100 % is computed. Forest loss is defined as a stand displacement disturbance leading from a forest state to a non-forest state (e.g. canopy density >50 % to 0). Therefore the underlying cause of forest loss ranges from anthropogenic impacts to natural causes. Tree cover gain is defined as the inverse of loss where the canopy density must exceed 50 % to get recognized.

Hansen et al. reports as an accuracy assessment of tree cover loss a producers accuracy of approximately 83 % for the tropical region. The mapped tree cover gain is probably an underestimation of the true gain with a producers accuracy of 48 % and a user's accuracy of 81 %.

This dataset is publicly available for download without any constraint. For a convenient bulk download, the dataset homepage provides a "`*.txt`" files comprising the Uniform Resource Locator (URL) of the tiles for each sub-dataset. The spatial location of an image can be directly determined from the file name within the URL. Each file name has a common pattern shown by the following expression: "`Hansen_VERSION_LAT[NS]_LNG[WE]`". LAT (latitude) and LNG (longitude) refer to the top left corner coordinates of a raster image,

whereas these coordinates are only given in natural numbers. The orientation of the image on the hemisphere is determined by the four cardinal directions N (north), S (south), W (west) and E (east). For this project, we require all three sub-datasets, namely: Treecover2000, loss-year, and gain. The data acquisition is automatized with a Python script by using the Standard Library (stdlib) modules urllib and re. At first, the Python script downloads the provided "`*.txt`" files and creates a list data structure, where each URL is an element of this list. After, it cycles through the list and extracts the corner coordinates from the file name by means of a Regular Expression (REGEX). These corner coordinates and cardinal directions are converted to valid latitude and longitude coordinates between $[-90, 90]$ and $[-180, 180]$, respectively. Now, an image is only downloaded if it is within the study extent between $[-20, 30]$ latitude. The acquired image tiles in total 678 are shown in the top panel (green squares) in figure 1.1.

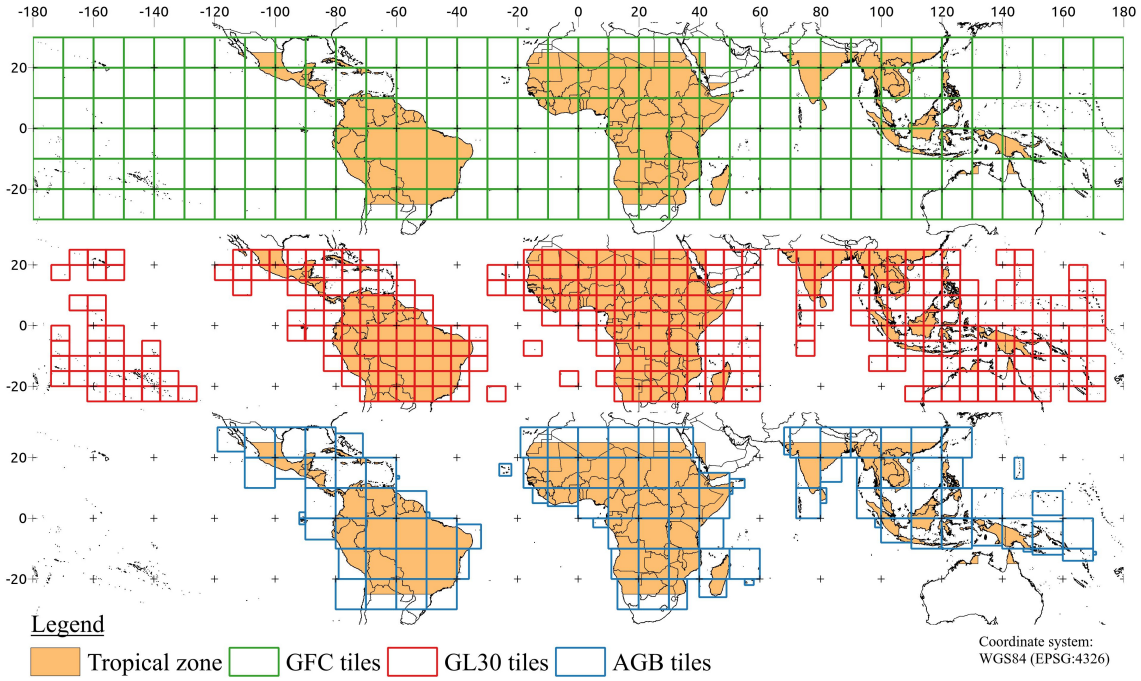

Figure 1.1. Map of downloaded dataset tiles: This map shows the acquired image tiles for this study. From top to bottom in green Global Forest Change (GFC) dataset tiles (Treecover2000, loss-year and gain), the land cover dataset GlobeLand30 (GL30) image tiles in red, and in blue the Aboveground Biomass (AGB) dataset tiles. The orange filled shapes highlight countries within the tropical zone.

1.1.2 GlobeLand30

GlobeLand30 (GL30) is the first global land cover dataset with a 30 meter per pixel spatial resolution that provides a comprehensive view on the distribution of 10 different land cover classes (table 1.2) over the entire globe [Chen et al. 2017]. Currently, this dataset is available for two different time periods 2000 and 2010 [Chen et al. 2015]. The pixel values of

this dataset are coded in unsigned 8-bit integers and as coordinates system, it uses WGS84 in Universal Transverse Mercator (UTM) projection. GL30 can be downloaded as a GTiff raster mosaic where each image covers 6x5 degrees [Chen et al. 2014]. For detecting the land cover classes Chen et al. used a so-called Pixel-Object-Knowledge (POK) oriented approach and satellite imagery from Landsat ETM+ [Chen et al. 2015]. Chen et al. divided the mapping process into different stages where each land cover type is detected separately and deleted subsequently from the source satellite image. The applied mapping order is water bodies, wetland, snow and ice, cultivated land and forest, shrubland, grassland and bare land synchronous. To detect the pixels of a selected land cover type the following pixel level classifiers are used: Decision Tree (DT), Support Vector Machine (SVM) or Maximum Likelihood Classifier (MLC). After pixel detection, the adjacent pixels are grouped as an aggregated land use object. These objects are subsequently validated by expert knowledge and the gained knowledge is used as a feedback loop to improve the automatized classification.

Chen et al. estimates an overall mapping accuracy of 80.33 % and 78.6 % for 2000 (only validated in Shaanxi, China) and 2010 (global), respectively [Chen et al. 2015]. Several research groups besides Chen et al. validated the mapping accuracy of GL30 at different regions and scales. Arsanjani et al. estimates an overall accuracy of 77.9 % for Iran and an accuracy >80 % for Germany [Arsanjani et al. 2016a,b]. Yang et al., Cao et al. and Jacobson et al. estimate the accuracy of 82.4 %, 80.1 % and 83.1 % for China, Nepal, and East Africa, respectively [Yang et al. 2017; Cao et al. 2016; Jacobson et al. 2015]. Unfortunately, no study focused on validating the mapping accuracy for regions exclusively within the tropical zone.

Chen et al. donated the GL30 land cover mapping to the United Nations (UN) but it is not accessible for public download. The download is restricted to users who register on the dataset homepage but the registration process is not working properly. Fortunately, the supervisor of this work had already an account otherwise it would be impossible to receive a copy of the dataset. A registered user must fill an order application to get access to the image tiles. The application form must contain the tile identifiers and the selected time period. Tile identifiers have the following common pattern: `"[NS]ZONE_LAT_NAME"` where zone refers to the UTM zone between $[1,60]$, N (north) or S (south) to the cardinal direction, and LAT (latitude) to the latitude coordinate of the top left corner. For better usability the homepage provides an interface for selecting the required image tiles but the selection of multiple tiles did not work. As well a vector file is provided which contains the dataset tile polygons with assigned identifiers. This file was used to select all required tiles within the tropical zone between approximately $[-23,23]$ degrees (WGS84). Figure 1.1 presents the selected images in red at middle panel. The corresponding image identifiers are converted to a single line string and copied to the application form. After submitting the form the order will be checked and approved within two weeks. After one week we received a two weeks limited

access to a password protected File Transfer Protocol (FTP) server where we downloaded 716 raster images. Due to the several restrictions, this process of selecting and downloading could not be automatized with one pipeline. Only the selection and string conversion were automatized with a throwaway script.

Table 1.2. Classification schema of the GlobeLand30 product: The code column is the assigned pixel value, type the corresponding land cover type and definition explains in broad terms which types of surfaces fall into the land cover type [Chen et al. 2017].

Code	Type	Definition
10	Cultivated land	used for agriculture, horticulture and gardens, including paddy fields, irrigated and dry farmland, vegetable and fruit gardens, etc.
20	Forest	covered by trees, vegetation covers over 30 %, including deciduous and coniferous forest, and sparse woodland with cover 10-30 %, etc.
30	Grassland	covered by natural grass with cover over 10 %, etc.
40	Shrubland	covered by shrubs with cover over 30 %, including deciduous and evergreen shrubs, and desert steppe with cover over 10 %, etc.
50	Wetland	covered by wetland plants and water bodies, including inland marsh, lake marsh, river floodplain wetland, forest/shrub wetland, peat bogs, mangrove and salt marsh, etc.
60	Water bodies	in land area, including river, lake, reservoir, fish pond, etc.
70	Tundra	covered by lichen, moss, hardy perennial herb and shrubs in the polar regions, including shrub-, herbaceous-, wet- and barren-tundra, etc.
80	Artificial surfaces	modified by anthropogenic influence, including all kinds of habitation, industrial and mining area, transportation facilities, and interior urban green zones and water bodies, etc.
90	Bareland	with vegetation cover lower 10 %, including desert, sandy fields, Gobi, bare rocks, saline and alkaline land, etc.
100	Snow and ice	covered by permanent snow, glacier and icecap

1.1.3 Intact Forest Landscapes

An Intact Forest Landscapes (IFL) is a mosaic of undisturbed forest patches and naturally treeless ecosystems without signs of human activity and large enough to maintain all native biological diversity [Potapov et al. 2017]. Due to the fact that IFL comprises different intact natural landscape patterns like primary forests, non-forest ecosystems, temporary treeless areas after a natural disturbance, and water bodies the term is not congruent to the term primary forest defined by the Food and Agriculture Organization of the United Nations (FAO) [FAO 2012]. But as mentioned IFLs includes large patches of primary forests with a minimum extent of 500 Km^2, therefore, primary forests can be extracted from the layer. Still, there are

smaller fragments of primary forest outside of the IFLs. In regards to the extent an IFL has a minimum size of 500 Km^2, a minimum width of 10 Km, and a minimum corridor/appendage width of 2 Km. Further an IFL should not contain any of the following: ecosystem alternation, fragmentation by infrastructure and disturbance, and areas altered or managed through agriculture, logging, and mining. For mapping and detecting IFLs Potapov et al. used Landsat imagery and several auxiliary data sources like GFC, and national transportation maps. The dataset can be downloaded as a Shapefile (SHP) file with the coordinate reference system WGS84. Each polygon in the SHP represents an IFL patch at a certain location on our planet at the time period 2000.

Data acquisition is pretty straight forward the IFL dataset public accessible for download. As mentioned it is a SHP so you must only download a single compressed archive. The download is automatized with a Python script by using the stdlib modules urllib and threading [van Rossum and Development 2018].

1.1.4 Aboveground Woody Biomass

The Aboveground live woddy Biomass density (AGB) raster dataset is prepared by Global Forest Watch (GFW) by an adapted approach of Baccini et al. [Baccini et al. 2012, 2015, 2017]. For the year 2000, this dataset estimates the aboveground biomass density per pixel in Mg C ha^{-1} (megagram carbon per hectare), and the confidence per pixel at a spatial resolution of approximately 1 arc-second (approximately 900 Km^2 or 30x30 meter). The dataset covering the global tropical zone as a mosaic of GTiff raster images where each tile of the mosaic has the Coordinate Reference System (CRS) WGS84 and is coded in a float. For deriving biomass density GFW used canopy metrics from Geoscience Laser Altimeter System (GLAS) Light Detection and Ranging (LIDAR) footprints and several regional and forest-specific allometric equations. The resulting GLAS AGB estimates are used as labels to train regional specific Random Forest (RF) models based on Landsat 7 ETM+ top-of-atmosphere reflectance, tree canopy density of GFC, elevation data, and climate data as predictor variables. After these models are subsequently applied to the entire study extent to predict the biomass content for each pixel. Additional an uncertainty layer is prepared accounting for the errors from allometric equations, the LIDAR based model, and the random forest model.

The AGB raster mosaic is publicly available on the homepage of GFW. As mentioned, the dataset covers only the tropical zone, therefore we acquire the entire mosaic. The GFW homepage provides an Geographic JavaScript Object Notation (GeoJSON) Application Programming Interface (API) to receive the actual URL of each raster image. If a request is sent to this API the server responded with a GeoJSON feature collection. The collection contains

as attributes the URLs of the biomass images, the URL of the uncertainty layers, and the rectangular bounds of each image. The data acquisition is automatized by means of Python and the stdlib modules urllib, threading, and the open source library geopandas [van Rossum and Development 2018; McKinney 2010]. At first, the GeoJSON is downloaded via an API call and eventually stored on disk. Next, we iterate the features of the GeoJSON collection and extract the URLs (biomass and uncertainty) of each tile. These URLs are downloaded and subsequently stored on disk. During the downloads of the uncertainty layers, the GFW server answered repeatedly with a 404 (Not found). Therefore the uncertainty layers are not available. In total we downloaded 105 different image tiles, their extent and spatial location are shown in blue at the bottom panel of figure 1.1.

1.1.5 Global Soil Organic Carbon

The Global Soil Organic Carbon map (GSOCmap) is a joint project between Global Soil Partnership (GSP) and Intergovernmental Technical Panel on Soils (ITPS) to produce a global Soil Organic Carbon (SOC) content map by a country-driven approach. In the year 2018, the first iteration of this map in version 1.0 was released and shortly followed by 1.1 (new country submission by Rwanda) and 1.2 (new country submissions by Chile and Colombia). As the short release cycle suggests the mapping project is intended as a long-lived dataset which will improve over time and by new country submissions. Till now 67 (approximately 63 % of the global land mass) different countries submitted their country based SOC estimates. To foster the national SOC mappings the International Soil Reference and Information Center (ISRIC) provides several covariate datasets like national Digital Elevation Map (DEM) maps, annual spectral remote sensing data or national soil type grids. Additionally, the contributors can join a mapping training and use the GSOCmap cookbook as guidance for their mapping efforts. As an exchange, each country shares its national GSOCmap by compliance of several criteria e.g. reporting of the Meta-data of the SOC sampling (sample timeline, sample depth, bulk density etc.), uncertainty assessment, and the applied methods for estimating and interpolation of the SOC content. For interpolating the guide organizations suggest the following approaches: simple geo-matching, class-matching, Multiple Linear Regression (MLR), RF or SVM. The national maps are aggregated to the final GSOCmap with a target resolution of 30 arc-seconds (approximately 1 Km^2) in the CRS WGS84. The dataset is one single raster image as GTiff coded in float covering the entire globe where each pixel value is the SOC content in Mg C ha^{-1} at a soil depth of 0-30 cm [FAO and ITPS 2018].

The product is validated by comparing the pixel level estimates with soil sampling data from various soil databases (WoSIS, HWSD, etc.). In total 312122 samples where divided into three sub-levels (<150 Mg C ha^{-1}, >150 Mg C ha^{-1}, and all samples) and subsequently

computed the Mean Error (ME). The ME of the entire sample space and <150 Mg C ha^{-1} suggests that the mean Soil Organic Carbon Content (SOCC) estimate is an overestimate of 1.6 and 4.5 Mg C ha^{-1} respectively. All samples with a SOCC content >150 Mg C ha^{-1} show an underestimate by approximately 165 Mg C ha^{-1} in the mean. Additionally, an uncertainty assessment was prepared to estimate a Standard Deviation (SD) between $\pm$ 0-16 t ha^{-1} for the tropical zone. Unfortunately, is this assessment pretty rough and till now not available as a product. The GSOCmap in comparison with other global SOC products has the lowest Root Mean Square Error (RMSE). In summary, the prepared validations show evidence that the GSOCmap is a conservative data product with a tendency to underestimate the SOCC.

The dataset is publicly available at the homepage of the FAO. As mentioned it consists of one raster image, therefore we download it by means of a Python script without any additional steps.

1.1.6 Soil Organic Carbon

Don et al. performed the first study of tropical SOC change for soil depth between 0 and 30 cm. For the study, a global meta-analysis is applied by using 358 (153 published an peer-reviewed) different studies to estimate SOC change for 12 major Land-use Change (LUC) types. The base date is derived from 39 different tropical countries covering all continents. Unfortunately, Africa and East-Asia are under-sampled whereas South-America has the best data coverage. The meta-analysis is restricted to mineral soils therefore all wet soil types are excluded from the analysis Don et al.. The 12 Land Cover (LC) transitions encompass the following LC types: primary forest, secondary forest, grassland, cropland, and perennial crops. Primary forest is defined as natural vegetation without human impacts which includes natural grassland and shrubland. Secondary forest is managed forests and regrown forests after partial destruction of the old stand. Grassland comprises pastures for livestock but excludes natural grasslands. Cropland comprises annual crops like maize or beans and perennial crops could be coffee or sugar cane. For our study we used only the SOC change estimates for these LUC types which correspond to the GL30 and IFL classification schema. The actual values are shown in table 1.3.

1.1.7 Ecosystem Service Values

Pending, which type of ecosystem services are included in the biome types

Table 1.3. Relative soil organic carbon change for certain land-use change types: The Land-use change columns from and to define the LUC type with the corresponding relative Soil Organic Carbon (SOC) change and the Standard Error of the Mean (SEM) [Don et al. 2010].

LUC type	Relative SOC change	
From→To	[%]	SEM
Primary forest→Grassland	-12.1	±2.3
Primary forest→Cropland	-25.2	±3.3
Primary forest→Secondary forest	-8.6	±2.0
Secondary forest→Grassland	-6.4	±2.5
Secondary forest→Cropland	-21.3	±4.1

Table 1.4. Selection of Ecosystem Service Values (ESV) per biome used in this study: ESV per biome and its monetary value in Int.$ ha^{-1}. Dg refers to data from de Groot et al., Co from Costanza et al., and Wb from Siikamaki et al..[de Groot et al. 2012; Costanza et al. 2014; Siikamaki et al. 2015]

Biome	Dg	Co	Wb
Cropland	-	5,567	-
Forest tropical	5,264	5,382	1,312
Grass/Rangelands	2,871	4166	-
Wetlands	25,682	140,174	-
Lakes/Rivers	4,267	12,512	-
Urban	-	6,661	-

1.1.8 Auxiliary

As auxiliary data for country boundaries we downloaded with Python the Global Administrative Areas Map (GADM) layers as SHP files [Hijmans et al. 2018; van Rossum and Development 2018].

1.2 Methods

Figure flowchart and reference shows an overview of the entire processing pipeline. The following sections describe detailed the applied approach for each step in figure. The order of appearance is from left to right.

1.2.1 Software, design and technology

For implementing our processing algorithms we selected for each task the technology which fulfills best the requirements. Python is our core language for implementing our processing algorithms because it supports a easy implementation of multiprocessing which is heavily used in this project. From the Python stdlib we used the following libraries: urllib, re, unittest, time, math, logging, collections, bisect, and enum. Additionally for geo-processing

we used the following open source libraries: numpy, pandas, fiona, geopandas, shapely, matplotlib. The entire frontend of our Python source code is aggregated in a Jupyter Notebook and available on GITHUB. This ensures that everyone interested can easily reproduce the findings of our project. JavaScript and the additional modules papaparse and the google maps api are used for programming a small web app for cross validation of land use predictions. R is used for hypothesis testing. Bash is used to aggregate large raster datasets as vrt files. To prepare map visualizations we used QGIS. Dia is used for preparing flowcharts and GIMP is used for image post-processing.

1.2.2 Preprocessing

Before we apply further analysis, we have to generalize the used datasets. As introduced in the data section do we use datasets which differ largely in their metadata properties, for example, single-tiled or multi-tiled images, used CRS, spatial resolution, and file type. Therefore, our goal should be to develop a process which creates an image stack of equal meta-data for each location in our study extent. In further descriptions, we will refer to this stack as Aligned Image Stack Mosaic (AISM). As target CRS for our AISM we chose WGS84 and as target extent for the mosaic, we use the bounding box of the GL30-2010 tiles. The following paragraph explains how we developed the alignment algorithm by means of Python and the additional open source libraries rasterio, geopandas, and shapely [van Rossum and Development 2018; McKinney 2010].

The first exercise of the preprocessing algorithm is to detect all tiles covering the extent of our template tiles. At first, we create for each multi-tiled dataset a polygon mask as SHP. This mask contains the spatial extent of each tile within a dataset and as attribute the corresponding file identifier. If the dataset tiles are not in WGS84 the extracted bounds are subsequently reprojected to this CRS. During the masking process, we recognized that the raster mosaic bounds of both GL30 datasets (2000 and 2010) generate re-projection errors. These errors showed up as polygons spanning the entire globe but one tile can only fill its UTM zone extent. Further analysis revealed that all tiles located in UTM zone 1 and 60 overflowing the maximum and minimum longitude coordinates of these zones. As solution we excluded all tiles within UTM zone 1 and 60 from further processing, namely: n01_00, s01_00, s01_10, s01_15, s01_20, s60_00, s60_05, s60_10, s60_15, and n53_00. The described steps can be found as well in figure 1.2. Now, as the figure suggests we determine the intersection between these mask layers and group the intersecting tiles by our template tile. Next, we create for the template tile a re-projection profile (warp profile) and apply it subsequently to all intersecting tiles based on the following rules: if from one dataset more than one tile intersects merge them followed by re-projection or if only one tile intersects

just re-project it. As introduced the GSOCmap consist only of one single tile with a spatial resolution of approximately 1 Km^2, so it must only re-project and re-sampled by the nearest-neighbor approach. We select from the IFL layer all polygons within our template warp profile and convert them to a raster layer where intact forest patches are coded by a one in an 8-bit unsigned integer. The last step of the alignment process is the rounding of the AISM bounds to full integer degrees and a subsequent clipping of each tile to this rounded bounds. The entire work-flow is presented in figure 1.2 and results in a AISM shown in figure 1.3. Finally, we create a polygon mask of our AISM and store for each polygon the corresponding dataset tiles. This mask is used as a file index for the next algorithms.

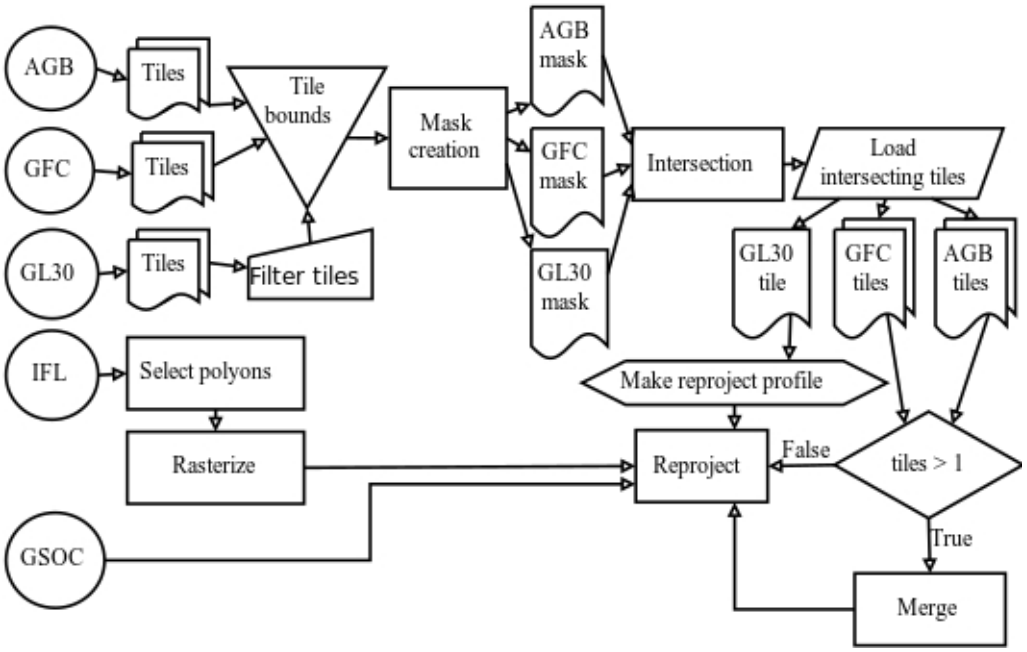

Figure 1.2. Tile alignment algorithm: For the multi-tiled datasets (multi-document symbols) a mask is created by extracting the tile bounds. Next, the intersection between these masks is determined to identify superimposing data and the corresponding tiles are loaded from disk. GL30 tiles are used as a template by creating the re-project profile and subsequently applying it to intersecting tiles. From the IFL layer only polygons within the re-project area are selected and subsequently converted to a raster layer.

1.2.3 Deforestation

1.2.3.1 Forest definition

To determine the proximate driver of deforestation we applied the two datasets GFC and GL30 but both differ in their definition of tree cover by canopy cover threshold as introduced in section 1.1. GFC detects tree cover over the entire canopy density interval of $(0, 100]$ while the GL30 threshold is set to > 10 %. To successful extract stable land cover transformation by superimposing both layers we must harmonize the tree cover definition of both strata.

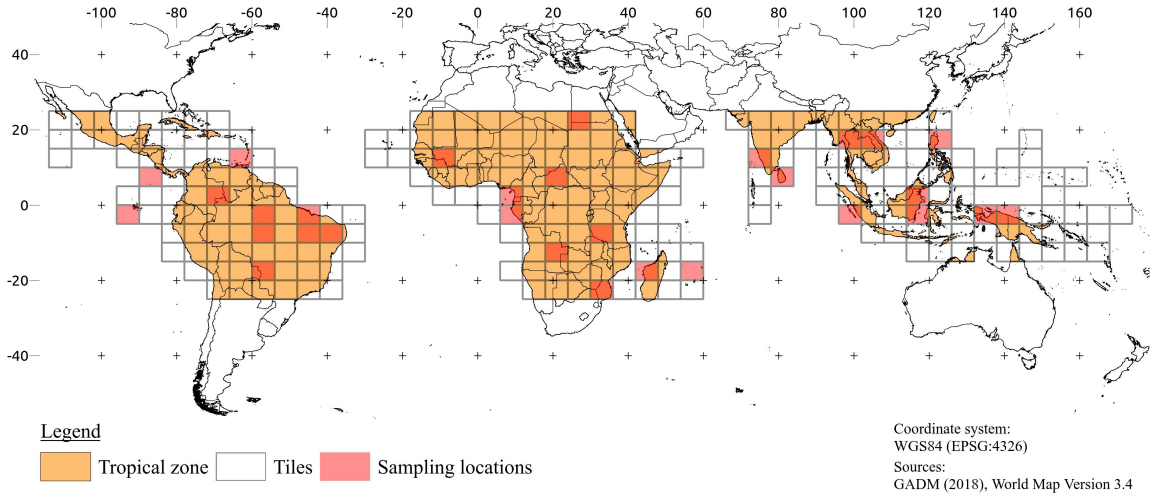

Figure 1.3. Aligned raster images and sampling locations: The map shows the location of the aligned multi-image stack tiles as black-framed square sized polygons, the sampling locations for accuracy assessment in red, and countries within the tropical zone in orange.

Our hypothesis is if both layers agree on tree cover they should also agree if a transition to a non-forest state occurs. To harmonize both definitions we have the opportunity to vary the canopy density of GFC to determine at which density class the similarity between them is at its maximum. After we use the examined max similarity canopy density to filter the tree cover loss and gain layer.

To determine the similarity between GL30 2000 and GFC reference tree cover we used the Jaccard Index (JI). The JI or coefficient of community is a simple measure of similarity between two pairs of a binary population or the measure of the degree of spatial overlap between two images [Sampat et al. 2009]. This index was first applied by Jaccard to compare distributions of rare alpine flora, in 1912 [Jaccard 1912]. If we compare two binary images, let a be the magnitude where both images (Img_1, Img_2) have an agreement represented as a pixel value of one. Let b the magnitude where Img_1 is zero and Img_2 is one and c the inverse of this expression. Assume that d is the magnitude of elements where both images are zero. The matrix in table 1.5 shows that the computation of this coefficients a, b, c, and d can be expressed as a set of boolean operations. Equation 1.1 shows how the JI is computed by substitute integer values for the variables. This computation can be reduced to two boolean operations for a major performance increase. The JI is always within the closed interval $[0, 1]$, where an index of one or zero means a complete similarity between both populations or a complete disagreement, respectively. The relationship between a and JI is near linear [Shi 1993]. The first step to compute the JI for our raster images is to extract the tree cover from the GL30 2000 land cover by setting all pixels with values $\neq 20$ to zero and values $= 20$ to one. Next, we extract from the GFC reference tree-cover pixel values within the half-opened interval of the following canopy density classes and set them to one: $(0, 100]$, $(10, 100]$, $(20, 100]$, and $(30, 100]$. We will refer to this JI of different canopy density classes

as JI_0, JI_1, JI_2, and JI_3. Now we compute for the entire study extent of 269 tiles and each tile combination of GL30 tree cover and GFC tree cover canopy density subset the JI by using equation 1.1. The algorithm is implemented in Python by using numpy's ability to perform boolean operations between large matrices. As parameters, the function expects two matrices with the same dimensionality in R^{n*m} and a boolean indicating if the function should return the coefficient matrix as well. The described preprocessing steps are implemented as an extra function which accepts to raster layers and optionally a list of integer values to consider as GFC tree-cover as well the lower interval threshold of the copy densities.

Table 1.5. Jaccard Index coefficient matrix: *a* is the magnitude of agreement, *d* is the magnitude of disagreement, *b* and *c* are the magnitudes of partial disagreements among both images. The computation of this coefficients can be expressed as boolean operations on matrices as shown in this table.

		Img_1	
	State	1	0
Img_2	1	$a = \lvert\mathbf{X_1} \wedge \mathbf{X_2}\rvert$	$b = \lvert(\mathbf{X_1} \wedge \mathbf{X_2}) \oplus \mathbf{X_2}\rvert$
	0	$c = \lvert(\mathbf{X_1} \wedge \mathbf{X_2}) \oplus \mathbf{X_1}\rvert$	$d = \lvert\neg(\mathbf{X_1} \vee \mathbf{X_2})\rvert$

$$JI = \frac{a}{a+b+c} = \frac{|\mathbf{X_1} \wedge \mathbf{X_2}|}{|\mathbf{X_1} \vee \mathbf{X_2}|} \tag{1.1}$$

To optimize the overall tree cover similarity between both datasets we must test which canopy density class yields the highest agreement over our study extent. For this exercise of testing the significance of the difference between two correlated samples, we decided to apply the non-parametric Wilcoxon signed-rank test [Wilcoxon 1945]. This test requires paired data from the same population, at least an ordinal scale of measurement, each sample pair is independent, and the dependent variable can be expressed as a continuous probability [Lowry 2019]. Further, an advantage of this test is that we don't have to assume a normal distribution for our sample population. Our sample population fulfills these requirements. The test procedure is implemented in R because this language is mainly intended for this kind of statistical analysis. We exported the computed JI from our Python environment as a Comma Separated Values (CSV) and applied a cross-testing in R. In our case cross testing is defined as the test of all two pair combinations of canopy density classes. Further, we applied a two and one-sided Wilcoxon test because we wanted to examine if there is a significant difference and which direction has the similarity distribution. Before we applied the examination of distribution we separated our population into three independent regions Americas, Asia, and Africa highlighted by the vertical blue lines in figure 1.1. Americas comprises 82 tiles, Asia 86 tiles, and Africa 101 image tiles. Additionally, we excluded from the analysis all samples where JI_0 is zero over the entire population which comprises 76 for Americas, 73 for Asia, and 86 for Africa. We decided to exclude these values because this tiles from our AISM did not contain any pixels covered by trees.

1.2.3.2 Proximate deforestation driver

Based on our forest definition developed in the previous section we want to classify all the tropical deforestation within a canopy density of $(10, 100]$ percent between 2001 till 2010. Additionally, we must consider the mean miss-classification rate of 52 % by previous findings of Seydewitz [Seydewitz 2017]. Therefore we have to develop a feasible method to resolve this issue.

For classifying the proximate drivers of deforestation we selected the following raster images from our AISM: GFC reference tree-cover, GFC annual losses, GFC gain, and the GL30 LC classification of 2010. Now, we apply to each raster image stack the following described operations. From the reference tree-cover images, we select all pixels where the canopy density is within the half-open interval of $(10, 100]$ percent and set them to one (true). The same exercise is applied on the annual losses stratum by setting all forest loss pixels within the time period 2001 till 2010 to one (true). After, both layers are combined with a logical AND operation to select our target deforestation pixels. Finally, we build the Hadamard product (element-wise multiplication) of the target deforestation layer and the GL30 LC stratum to classify the pixels with a deforestation event. For classifying forest regrowth we filtered the GFC gain layer to consider only tree-cover gain within our target temporal resolution and target canopy density. After, the filtered stratum is aggregated with our classified deforestations by using the Hadamard product of both layers. We will refer to this proximate deforestation driver layers as PDD. Figure 1.4 shows an overview of the classification process. The classification algorithm is implemented as a Python function which requires a parameter the previously named raster layers. Additionally the target canopy density and time period is freely selectable for experimental variations. The described filtering and aggregation steps are implements as binary matrix operations for fast processing of large data sizes by means of numpy.

After classifying the proximate deforestation drivers we developed an approach to smooth the misclassified pixels based on an approximated probability. We define misclassifications as sites where the GFC annual loss data predicts deforestation but the GL30 stratum still classifies them as forest. The first step of our reclassification is to cluster the misclassified pixels with the Hoshen-Kopelman algorithm [Hoshen 1998]. The clustering algorithm is implemented as a part of the Geospatial Data Abstraction Library (GDAL) library and can be called through the rasterio interface. For this project, we used the following parameters: connectivity 4 and a boolean mask where only pixel values $= 20$ are true. Now the algorithm creates for each pixel cluster a polygon. After we created a squared sized buffer with a side length of 500 x 500m around the polygon centroid (the geometric midpoint of the polygon). Because WGS84 is not an equal area CRS we must compute for each tile the buffer size separately. To compute the buffer size in image coordinates we used the Haversine formula

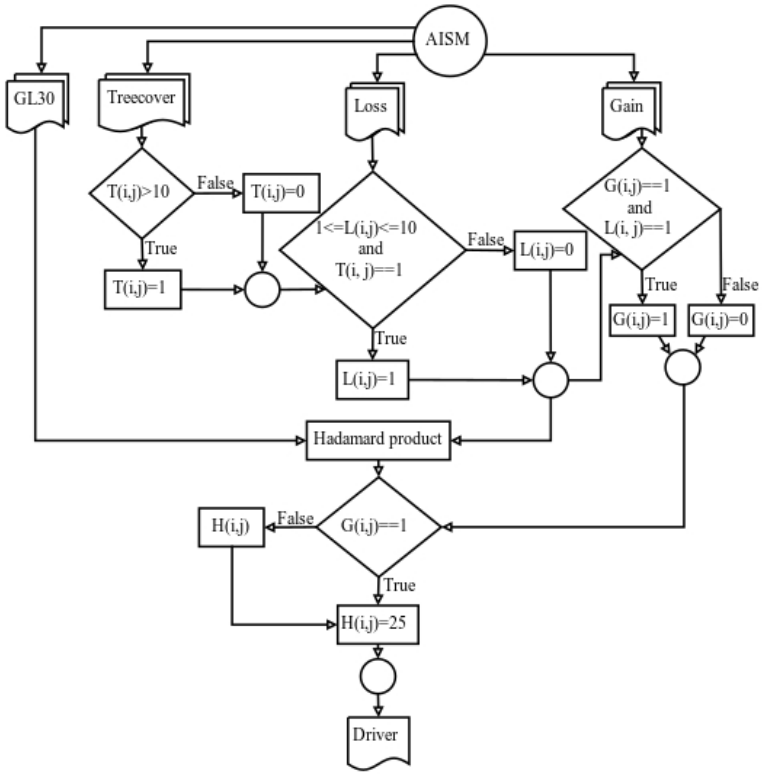

Figure 1.4. Classification of proximate deforestation drivers: For the classification of the proximate deforestation drivers the following layers are required GL30 2010, GFC tree cover, annual losses, and gain. From tree cover we select all pixels within the canopy density interval $(10,100)]$. The tree cover mask is used to select the appropriate annual losses within the time interval $[2001,2010]$. To predict a land cover change after a deforestation event we use the Hadamard product.

in equation 1.2 to determine the on-ground resolution in meter on pixel level. Where d is the great-circle distance between two latitude, longitude pairs φ_n, λ_n and r is the earth radius of approximately 6378137m. Because this computation is expensive we assumed that the pixel resolution is equal for an entire raster tile. After extracting the buffer we counted the most frequent class under exclusion of pixels with a value of 0, 20 or 255 within the buffer. Finally, if the most frequent class is defined we reassigned this class value to the cluster. The reclassification algorithm is implemented as a Python function which requires as parameters a proximate driver raster image, a list of elements which should be interpreted as occupied cells for the clustering, pixel values which should be excluded from counting, the side length of the buffer, and the on-ground resolution.

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right) \tag{1.2}$$

1.2.3.3 Accuracy assessment

For examining the accuracy of our Proximate Deforestation Driver (PDD) predictions we used a confusion matrix (also known as two-way frequency tables, error matrix or contin-

gency tables). These matrices are commonly used for an accuracy assessment of land cover classifications and enable the computation of marginal and conditional distributions [Congalton 1991; Foody 2002]. Table 1.6 shows a general model of a confusion matrix. Foundation for an accuracy assessment by means of a confusion matrix is a collection of ground-truth samples which can be compared with the class predictions for these samples produced by a classification algorithm. For the preparation of our accuracy assessment, we have to extract a collection of pixel samples with a deforestation occurrence from our proximate driver maps (further also called predictions). Next, we compose a set of ground-truth for these predictions (further also called references).

Table 1.6. A general model of a confusion matrix: X_1, ... , X_n denote classification categories of two independent raters. $x_{n,n}$ are the actual samples sorted into the categories where the values in the diagonal show the agreement between both raters. The remaining cell values account for the disagreement between the two raters. $\sum$ column and row show the marginal distribution and N is the total number of samples.

		Reference			
	Cls	X_1	$\cdots$	X_n	$\sum$
Predict	X_1	$x_{1,1}$	$\cdots$	$x_{1,n}$	$x_{1.} = \sum_{i=1}^{n} x_{1,i}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	X_n	$x_{n,1}$	$\cdots$	$x_{n,n}$	$x_{n.} = \sum_{i=1}^{n} x_{n,i}$
	$\sum$	$x_{.1} = \sum_{i=1}^{n} x_{i,1}$	$\cdots$	$x_{.n} = \sum_{i=1}^{n} x_{i,n}$	$\sum\sum = N$

To create our collection of ground-truth data we draw randomly 10 image tiles from all three continental regions (Americas, Africa, Asia/Oceania) namely the following tiles: 25N 024E, 20N 096E, 20N 102E, 20N 120E, 15N 066W, 15N 012W, 15N 072E, 10N 090W, 10N 018E, 10N 078E, 05N 072W, 05N 006E, 05N 114E, 00N 096W, 00N 060W, 00N 048W, 00N 006E, 00N 096E, 00N 114E, 00N 132E, 00N 138E, 05S 048W, 05S 042W, 05S 060W, 05S 030E, 10S 018E, 15S 060W, 15S 042E, 15S 054E and 20S 030E where the first part is the latitude coordinate and the last part is the longitude coordinate of the upper left corner. From each tile, we sampled by random 200 pixels which total to 6000 samples over the entire study region. The sampling is realized with our own raster sampling algorithm build in Python by means of the open source libraries numpy and rasterio. As mentioned in the previous section do we superimpose two datasets and only a certain amount of pixels per tile is classified as a proximate driver. Therefore, the sampling algorithm should only draw samples from occupied/classified pixels without replacement. The algorithm expects as parameters a raster image, the total number of samples to draw, a list of pixel values which should be interpreted as occupied cells, the affine transformation matrix of the raster image, and a seed for the random number generator. If occupied cells are set the algorithm will create a binary mask

where each occupied cell is set to one relative to the input raster image. Otherwise, it sets all pixel values greater or less than zero to one. After, the row and column coordinates of each one are extracted from the mask and converted to a flat list of coordinate tuples. Next, it draws the predefined number of samples from the list by a random order and uses the image coordinates to get the pixel value from the raster image. If an affine transformation matrix is provided the image coordinates are converted to real-world coordinates. The seed argument ensures that on every algorithm rerun the samples are drawn. For our sampling we set the parameters to the following values: samples 200, occupied pixels GL30 class values and 25 for regrowth, the affine matrix of the corresponding raster image, and the seed is 42. The per tile samples are stored as a CSV file.

For the collection of ground-truth data, we used a visual interpretation of satellite and aerial imagery provided by Google Maps. We developed a small JavaScript web application to access the imagery via the Google Maps API. The application expects as input a CSV file with the sampling coordinates. After upload of a sample file the user can cycle through the entries and the map jumps automatically to the coordinates of the sample. Now a reference label can be assigned to the coordinates by visual interpretation of the imagery. We subsequently assigned to all 6000 samples a reference label and downloaded the results as CSV.

Finally, we developed a Python class to compute the confusion matrix. The constructor of the class requires a list of reference and prediction labels. With the provided arguments it creates the confusion matrix. Further, it computes the following marginal and conditional distributions: overall accuracy *OvAc* by dividing the sum of classification agreements by the sample total N (equation 1.3), the producer accuracy $PAc_{.n}$ by dividing the category agreement by the column category total (equation 1.4), the error of commission $Com_{.n}$ (Type II error) by dividing the category disagreement by the column category total (equation 1.5), the user accuracy $UAc_{n.}$ by dividing the category agreement by the row category total (equation 1.6), the error of omission $Om_{.n}$ (Type I error) by dividing the category disagreement by the row category total (equation 1.7), and the Cohens Kappa by substituting equation 1.8 and 1.3 into equation 1.9.

$$p_0 = OvAc = \frac{\sum_{i=1}^{n} x_{i,i}}{N} \tag{1.3}$$

$$PAc_{.n} = \frac{x_{i,i}}{x_{.n}} \tag{1.4}$$

$$Com_{.n} = \frac{FN_i}{x_{.n}} \tag{1.5}$$

$$UAc_{n.} = \frac{x_{i,i}}{x_{n.}} \tag{1.6}$$

$$Om_{n.} = \frac{FP_i}{x_{n.}} \tag{1.7}$$

$$p_c = \frac{1}{N^2} \sum_{i=1}^{n} x_{.i} \cdot x_{i.} \tag{1.8}$$

$$Kappa = \frac{p_0 - p_c}{1 - p_c} \tag{1.9}$$

1.2.4 Emissions

Land cover change respectively deforestation releases carbon emissions. These emissions can be grouped to different categories like emissions from transportation, biomass removal, changes of soil carbon dynamics, processing of certain kind of commodities etc.. During the previous sections we developed an approach to predict the change of tree cover driven by proximate causes like conversion to cropland or else. Now we can use these predictions to approximate the CO_2 emissions uprising from this land cover transitions. For this study, we focus on the emissions emitted by biomass removal and from changes of soil carbon stock. The first paragraph is focused on the estimation of emissions from biomass removal and the second section tries to approximate the impact of land cover change on the soil organic carbon content.

To obtain the gross CO_2 emissions through proximate deforestation driver we selected the following raster tiles from our AISM: the AGB stratum and our classification of the PDD. By means of Python, we implemented a function which accepts as parameter two raster images, the area a pixel covers in m^2, a factor to convert carbon to CO_2, and a list of proximate driver classes to consider as deforestation. We considered the following driver classes as deforestation for the computation: 10 (cropland), 25 (regrowth), 30 (grassland), 40 (shrubland), 50 (wetland), 60 (water bodies), 70 (Tundra), 80 (artificial), and 90 (bareland). The function computes the gross emissions by using equation 1.10. Let Y_{ij} be the AGB in Mg C ha^{-1} and X_{ij} the PDD at an pixel index i, j obtained from a raster image matrix in R^{N*M}. Let A be the area in ha a pixel covers for a certain image tile. This area is calculated by using the Haversine function from equation 1.2. Factor 3.7 converts Carbon to CO_2. Let $AGBE_{tile}$ be the cumulative emissions emitted from the removal of tree cover. Then this value can be obtained by taking the sum of the product of Y_{ij} and $f(X_{ij})$. Whereas the piecewise function f only evaluates to one if the proximate deforestation driver is within our set of classes we want to consider as deforestation. To obtain the gross AGB emissions through the deforestation by proximate deforestation driver we aggregated the sum of $AGBE_{tile}$ for the regions Latin America, Asia, and Africa.

$$AGBE_{tile} = 3.7A \sum_{i=0}^{N} \sum_{j=0}^{M} f(X_{ij}) Y_{ij} \tag{1.10}$$

To obtain the gross CO_2 emissions emitted by the change of soil organic carbon content we

selected the following raster tiles from our AISM: the IFL stratum, the GSOCmap, and our prediction of PDD. We decided to predict the SOC emissions for two different scenarios. In scenario one SC_1 we assume that all tree covered areas concerned by a land cover change are primary forest. For scenario two SC_2 we used IFL stratum to determine the forest type. If land cover changes within an IFL patch it concerns primary forest otherwise it is secondary forest. The SOC emissions of both scenarios can be computed by equation 1.11. Let X_{ij} be the PDD from our prediction, Y_{ij} the forest type determined by the IFL stratum, and Z_{ij} the SOC Mg C ha^{-1} determined by GSOCmap at an pixel with index i, j obtained from a raster image matrix in R^{N*M}. Let A be the area in ha a pixel covers for a certain image tile. This area is calculated by using the Haversine function from equation 1.2. Factor 3.7 converts Carbon to CO_2. Let $SOCC_{tile}$ be the cumulative soil organic carbon emissions emitted by the change of forest to another land cover type. Then this value can be obtained by taking the sum of the product of Z_{ij} and $h(X_{ij}, Y_{ij})$. Whereas the piecewise function h returns the mean soil organic carbon change and the standard error in respect to the forest type and proximate driver class. The mappings of drive classes and forest type for both scenarios are shown in table 1.7 and 1.8. This algorithm is implemented by means of Python. The function needs as parameter the required layers whereas the IFL stratum is optional, the area a pixel covers in m^2, a conversion factor for carbon to CO_2, an identifier for the forest type, and if the standard error should be included during the computation of the emission. If the IFL stratum is provided the algorithm will rely on this layer to determine the forest type otherwise it uses forest type identifier. To obtain the gross SOC emissions by the transition of land cover we aggregated the sum of $SOCE_{tile}$ for the regions Latin America, Asia, and Africa.

$$SOCE_{tile} = 3.7A \sum_{i=0}^{N} \sum_{j=0}^{M} h(X_{ij}, Y_{ij}) Z_{ij} \tag{1.11}$$

Table 1.7. Scenario one mapping of soil organic carbon change to proximate driver: In scenario one we assume that deforestation always occurs in primary forest. Refer to table 1.2 for the description of the proximate driver class. Standard errors of the soil organic carbon change factors are denoted in table 1.3. The symbols in superscript denote the following transitions: † Primary forest→Cropland, ‡ Primary forest→Secondary forest, and ⋄ Primary forest→Grassland

	Proximate driver class					
Forest type	10	25	30	40	70	90
Primary	$.252^{\dagger}$	$.086^{\ddagger}$	$.121^{\diamond}$	$.121^{\diamond}$	$.121^{\diamond}$	$.121^{\diamond}$

1.2.5 Ecosystem service values

Ecosystems have an impact on the well being and subsistence of current future generation of humanity by providing regulatory, habitat, provisioning, and cultural services. For the

Table 1.8. Scenario two mapping of soil organic carbon change to proximate driver: In scenario two we use the Intact Forest Landscape stratum to distinguish between deforestation in primary and secondary forest. Refer to table 1.2 for the description of the proximate driver class. Standard errors of the soil organic carbon change factors are denoted in table 1.3. The symbols in superscript denote the following transitions: † Primary forest→Cropland, ‡ Primary forest→Secondary forest, ⋄ Primary forest→Grassland, § Secondary forest→Cropland, and ∗ Secondary forest→Grassland

	Proximate driver class					
Forest type	10	25	30	40	70	90
Primary	$.252^{\dagger}$	$.086^{\ddagger}$	$.121^{\diamond}$	$.121^{\diamond}$	$.121^{\diamond}$	$.121^{\diamond}$
Secondary	$.213^{\S}$	-	$.064^{*}$	$.064^{*}$	$.064^{*}$	$.064^{*}$

quantification of these ecosystem services a economic process is applied to assess the monetary value of each service per ecosystem also refereed as biome. These Ecosystem Service Valuess (ESVs) can be a strong tool to determine the impact of certain management practices on ecosystem structures. Especially as impact analysis for our study of tropical tree cover transitions. For a comprehensive insight of the ESV dynamics, we quantified the loss of ESV from forest cover degeneration within the tropical zone. This degeneration of forest cover is frequently followed by a transition to other land cover types expressed through our PDDs. These transitions can be interpreted as the gain of ESV and are computed subsequently. Finally, to give an insight into the overall trend of both ESV dynamics we determined the balance among the monetary loss and gain. The first paragraph describes our approach to determine the ESV loss, followed by the exercise to obtain the gain in monetary units and finally we explain how to derive the balance between both values.

By applying equation 1.12 we compute the gross ESV loss from the loss of tropical tree cover for the three continental regions Latin America, Asia, and Africa. Let X_{ij} be the PDD from our prediction at an pixel with index i, j obtained from a raster image matrix in R^{N*M}. Let ESV_{Forest} be the ESV of tropical forest from one of our selected source datasets from table 1.9. Let A be the area in ha a pixel covers for a certain image tile. The pixel area is calculated by using the Haversine function from equation 1.2. Let $ESV_{loss,tile}$ be the cumulative loss in ESV for a certain tile form you AISM. Then this value can be determined by adding the product of $f(X_{ij})$ and $ESV_{Forest,Dataset}$. Whereas the function f returns only one if the PDD is considered as deforestation by the mapping in table 1.9. The computation of ESV loss is implemented as a Python function. Whereas the function accept as parameters a raster image of PDD predictions or a pandas data frame object. Further, the function requires as parameter the area a pixel cover in ha and the monetary value of tropical forest. Additionally, the function requires a list of PDD classes considered as the loss of tropical forest cover. We considered the following PDD classes as forest loss: 10, 25, 30, 40, 70, 80 and 90 as table 1.9 suggests.

$$ESV_{loss,tile} = A \sum_{i=0}^{N} \sum_{j=0}^{M} f(X_{ij}) ESV_{Forest,Dataset} \tag{1.12}$$

Table 1.9. ESV biome types mapped to PDD classes: The monetary values are given in Int.$ ha^{-1}. Mapping of biome types to PDD classes have the following schema: 10 to cropland biome, 25 to tropical forest biome, 30 to grassland biome, and 80 to the urban biome. The abbreviations in the ESV dataset column refer to the following publications: Dg is de Groot et al., Co Costanza et al., and Siikamaki et al.

	Proximate driver class						
ESV dataset	10	25	30	40	70	80	90
Dg	-	5,264	2,871	-	-	-	-
Co	5,567	5,382	4,166	-	-	6,661	-
Wb	-	1,312	-	-	-	-	-

To estimate the gain in ESV from the transition of tropical forest to other land cover classes per continental region we applied equation 1.13. Let X_{ij} be the PDD from our prediction at an pixel with index i, j obtained from a raster image matrix in R^{N*M}. Let A be the area in ha a pixel covers for a certain image tile. The pixel area is calculated by using the Haversine function from equation 1.2. Let $ESV_{gain,tile}$ be the cumulative gain of ESV per tile. Then this value can be determined by taking the sum of $h(X_{ij})$. Whereas the function h returns for a selected PDD class the corresponding monetary value. The algorithm is implemented in Python. Whereas the function accept as parameters a raster image of PDD predictions or a pandas data frame object. Further, the function requires as parameter a mapping of ESVs to PDD classes from table 1.9. Additionally, the function can be called with a exclude list of PDD classes.

$$ESV_{gain,tile} = A \sum_{i=0}^{N} \sum_{j=0}^{M} h(X_{ij}) \tag{1.13}$$

By applying equation 1.14 we compute the ESV balance for the three continental regions Latin America, Asia, and Africa. Let ESV_gain be the total ESV gain per continental region and ESV_{loss} the total ESV loss per region. Then the ESV balance $ESV_{balance}$ can be obtained by the difference of ESV_gain and ESV_{loss}.

$$ESV_{balance} = ESV_{gain} - ESV_{loss} \tag{1.14}$$

1.2.6 Binning analysis and visualization

During the previous sections, we were focused on the exercise of creating large scale spatial explicit predictions for land cover transitions and the following consequences for humanity. Now, an appropriate method must be developed to analyze and visualize these spatial explicit datasets by generalizing the problem domain. By the nature of fine resolution raster images and the large area of our study extent we must handle a large N (many samples) and the resulting high dimensionality respectively complexity of relationships among the samples [Carr 1990]. Raster image maps can be interpreted as multivariate scatter plots. In our

case this scatter plot has the three dimensions x is the longitude, y the latitude coordinate of a pixel, and z is the nominal scaled pixel value in case of the PDD prediction. Drawing scatter plots with large multidimensional N commonly leads to overplotting and hidden point densities [Carr et al. 1987]. Additionally, it is to assume that the distribution of PDD is not equally distributed over the entire study extent. Hence, there should be regions with sparse data densities and with high densities but our goal is to visualize land cover changes on a continental level. As mentioned the ground resolution of one pixel covers an area of approximately 30x30 m and as an example, the bounding box of Latin America covers an area of $5 * 10^7$ Km^2. The large frame size as well the unequal distributed data leads to the issue that only large scale land cover changes are representable and small scale isolated changes stay hidden.

By referring to the latter paragraph our goal should be to develop a process to solve the representation issues and generates satisfying maps. In the case of raster data, one opportunity could be a re-sampling to a coarser on-ground resolution. This approach may solve the overplotting as well resolution issues and normalize unequal distributed data. For nominal scaled data the commonly used re-sampling methods are nearest neighbor or majority wins Reference. Both approaches are not appropriate because they would negate spatial patterns and eliminate important land cover class frequency distributions. Another well-accepted method is binning of spatially explicit data with a regular polygon that can tessellate the plane [Carr et al. 1992]. Polygon tessellations provide numerous opportunities for presenting multivariate statistical and visual summaries. The scale of a polygon may be used to visualize pixel densities within the bounds and a color gradient may be used to prepare a choropleth map for nominal or ordinal scaled data. Additionally, the interior of a polygon may be used to prepare a pie chart. Hence, binning enables convenient visualization of multidimensional data. For preparing a regular tessellation only three types of convex polygons can be used to tessellate the plane: squares, equilateral triangles, and hexagons [Carr et al. 1992]. Square tessellations are the most common method in comparison with hexagons for binning and visualizing spatial data. Every raster image is already a square tessellation of the mapped object and most of the image processing algorithms are focused on squares. Hexagon mosaic maps have two major advantages over square tessellations: visual appeal and representational accuracy. Binning of data by a square or hexagon mosaic creates visual lines. These lines compete with the data generated patterns. Especially humans have a strong visual response to horizontal and vertical lines. Hence, the line artifacts of square tessellations are distracting and should be avoided. Thus, we decided to use hexagon mosaic maps to represent the visual and statistical results of our study. For bivariate representations, we select the combination of scaling and color gradient. Multivariate data is visualized by hexagonal pie charts. The following paragraphs describe our algorithmic approach to create these mosaic maps. We used Python and the open source library shapely to implement our algorithms.

The first step to construct a hexagon tessellation is to define the vertices of the polygon. There are two common orientations of hexagons in R^2 flat topped and pointy topped. For our hexagon construction we decided to use pointy topped polygons. For flexibility our algorithm accepts one out of four parameter to construct a hexagon polygon. The unit of the parameter is always in map units. Let D be the long diagonal (diameter of the circumscribing circle), d the short diagonal (diameter of the inscribed circle), A the area the hexagon should cover, and e the edge length of a hexagon. Let R be the radius of the circumscribing circle. Then R can be obtained by applying equation 1.15 with one out of the parameter set D, d, A, or e. R is used to compute the center vector $\vec{m} =< c_x, c_y >$ of the polygon by applying equation 1.16 and 1.17. The polygon center is always located in the first quadrant of the Cartesian coordinate system. Now, by using R, c_x, and c_y we can obtain $\mathbf{H}$ the anti-clockwise orientated vertex matrix of a hexagon. The construction of a hexagon by using the introduced method is shown in the left bottom corner of figure 1.5. The next paragraph describes how we derive a tessellation from the constructed hexagon.

$$R = \frac{\sqrt{2A}}{\sqrt[4]{27}} = \frac{D}{2} = \frac{d}{\sqrt{3}} = e \tag{1.15}$$

$$c_x = \frac{R\sqrt{3}}{2} \tag{1.16}$$

$$c_y = R \tag{1.17}$$

$$\mathbf{H} = \begin{bmatrix} 0 & c_x & 2c_x & 2c_x & c_x & 0 \\ R\sin\left(\frac{7\pi}{6}\right) + c_y & 0 & R\sin\left(\frac{11\pi}{6}\right) + c_y & R\sin\left(\frac{\pi}{6}\right) + c_y & 2R & R\sin\left(\frac{5\pi}{6}\right) + c_y \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{1.18}$$

To create a polygon grid of a plane image we must align several hexagons to cover the image. For our tessellation algorithm we use the vertex matrix $\mathbf{H}$ computed by the previously described approach and subsequently translate it to its position within the grid. We expect to receive the bounds matrix $\mathbf{B}$ of the raster image which should be tessellated by hexagons, equation 1.19. Let x_1, y_1 be the left bottom corner coordinates and x_2, y_2 the right top corner coordinates of an image. Let $x_{off}(0)$, $y_{off}(0)$ in equation 1.20 and 1.21 be the initial coordinates for creating a polygon grid over a plane. Then we can obtain $x_{off}(n+1)$ the x coordinates for even rows by applying equation 1.22 and $x_{off}(n+1)$ the x coordinates for odd rows by equation 1.23. Where r is the radius of an inscribed circle in a hexagon and can be obtained by dividing d by 2. Then $\mathbf{H}$ can be translated to the vertex matrix $\mathbf{T}$ by applying the dot product of an affine transformation matrix and $\mathbf{H}$, equation 1.25.

$$\mathbf{B} = \begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \end{bmatrix} \tag{1.19}$$

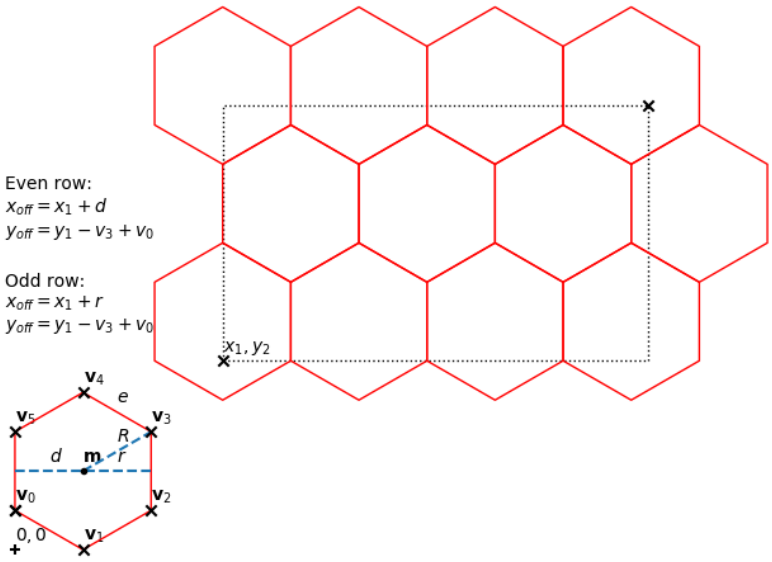

Figure 1.5. Hexagon tessellation: Located at the left bottom corner in red a hexagon defined by its geometric properties the 6 vertex vectors $\{\vec{v}_0, ..., \vec{v}_5\}$ (black crosses), with center vector $\vec{m}$, edge length e, R radius of the circumscribing circle, r radius of the inscribed circle and d the short diagonal (diameter of the inscribed circle). Top right black dotted box are the bounds of an area which is tessellated by a hexagon grid in red. Each grid cell is translated from the origin hexagon at its position by computing the x_{off} and y_{off} offset with the presented equations at the left-hand side of the grid.

$$x_{off}(0) = x_1 \tag{1.20}$$

$$y_{off}(0) = y_1 \tag{1.21}$$

$$x_{off}(n+1) = x_{off}(n) + d \tag{1.22}$$

$$x_{off}(n+1) = x_{off}(n) - r + d \tag{1.23}$$

$$y_{off}(n+1) = y_{off}(n) - v_{0,2} + v_{3,2} \tag{1.24}$$

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & x_{off}(n) \\ 0 & 1 & y_{off}(n) \\ 0 & 0 & 1 \end{bmatrix} \circ \mathbf{H} \tag{1.25}$$

Goal: A pie chart within the area of a hexagon, split the hexagon in horizontal pieces which represent a ratio

How: Compute the y coordinate of the split line from the ratio, the distance between y1 and y2, compute from this the x coordinates

$$y = \frac{P(y_2 - y_1)}{100} + y_1 \tag{1.26}$$

$$f^{-1}(y) = \begin{cases} -\frac{y-y_1}{\tan(\frac{\pi}{6})} + \frac{x_1+x_2}{2} & \text{if } y_1 \leq y < y_1 + R\sin(\frac{5\pi}{6}) \\ x_1 & \text{if } y_1 + R\sin(\frac{5\pi}{6}) \leq y < R(\sin(\frac{5\pi}{6})+1) \\ \frac{y-y_2}{\tan(\frac{\pi}{6})} + \frac{x_1+x_2}{2} & \text{if } R(\sin(\frac{5\pi}{6})+1) \leq y \leq y_2 \end{cases} \tag{1.27}$$

$$g^{-1}(y) = \begin{cases} \frac{y-y_1}{\tan(\frac{\pi}{6})} + \frac{x_1+x_2}{2} & \text{if } y_1 \leq y < y_1 + R\sin(\frac{5\pi}{6}) \\ x_2 & \text{if } y_1 + R\sin(\frac{5\pi}{6}) \leq y < R(\sin(\frac{5\pi}{6})+1) \\ -\frac{y-y_2}{\tan(\frac{\pi}{6})} + \frac{x_1+x_2}{2} & \text{if } R(\sin(\frac{5\pi}{6})+1) \leq y \leq y_2 \end{cases} \tag{1.28}$$

$$\mathbf{L} = \begin{bmatrix} f^{-1}(y) & g^{-1}(y) \\ y & y \end{bmatrix} \tag{1.29}$$

Goal: All use hexagons with an area of 0.5 degrees i square, aggregated tiles per continent americas, africa, asia

Treecover: Count tree cover pixels within hexagon in canopy interval (10,100], count total pixels within hexagon, compute pixel area with haversine, divide tree covered area by total area, 5 ratio bins 0.2 0.4 0.6 0.8 1.0 for scaling, store the area for interval, mean canopy density over occupied pixels, use mean canopy density for choropleth map

PDD: count frequencies within a hexagon, compute frequency ratios, segement hexagons by the ratio of pdd driver, order the drivers in decreasing order, most common is first, tried with scaling but not feasible cause few big sized and many small sized, just scale them a bit down cause visual appeal,

Loss: count frequency of pdd recognized as deforestation

2 Results

IN PROGRESS

2.1 Deforestation

2.1.1 Forest definition

Goal (review): Our goal was to determine at which canopy cover class the similarity between both layers is greatest to get the subsequent proximate deforestation driver for stable land cover changes optimal by anthropogenic causes by keeping the largest number of pixels from the gfc dataset. We applied the jaccard index for searching the similarity. We grouped our analysis by continental regions americas, asia, africa. Americas accounts for 82 tiles, Asia 86 tiles and Africa 101. We excluded from the analysis all tiles where the initial jaccard index is zero because theses tiles does not contain any tree cover in both tile pairs. This results in 76 Americas , 73 asia and 86 africa. Further we determined the optimal canopy density class for all regions and for single regions by applying the non parametric two and one sided wilcoxon signed rank test. Our initial hypothesis was that the agreement is max between gl30 and hansen when the selected canopy density is between 30 and 100. Because then both datasets should agree by their authors definition of tree cover. The following paragraphs present the results of the analysis for each continental region.

Americas (review): Figure 2.1 shows the quartile distribution of the computed jaccard index for each tile pair for each canopy density class over the three continental regions. Plotted on the x-axis is the canopy density class identifier where JI_0 accounts for (0,100], JI_0 (10,100], JI_0 (20,100], and JI_0 (30,100]. The y-axis is the corresponding jaccard index between 0 and 1 for the corresponding tile pair where 1 means total agreement and 0 total disagreement. The sample mean highlight by red crosses in the boxplot for the Americas does not change significantly within the different canopy density classes. For all experiments it is approximately 0.62. While the sample median decreases from 0.68 to 0.66 from the first canopy density class the last canopy density class. For the first canopy class the upper 25 % of the samples have tree cover similarity ranging between approximately 0.8 and 1.0. This behavior can be

observed at the other canopy density classes to only the maximal similarity increases slightly from 0.9787 to 0.9798. As the figure appendix suggests the change of the canopy density have only little impact on the tiles where already the similarity is high for the upper 25 percent. The similarity range of the first two canopy density classes for the lower 25 percent of the samples ranges between approximately 0.0003 and 0.47. Whereas the range for last to canopy classes ranges between 0.0 and 0.5. This suggests that the exclusion of higher canopy densities decreases the similarity at samples where the similarity is already low also shown in figure appendix. 50 percent of the samples have a jaccard index between approximately 0.5 and 0.8 where here the highest mobility of similarity increase and decrease can be observed. To deduce which canopy density class yield the highest similarity in the distribution overall all samples from americas we applied a wilcoxon test. Tabel 2.1 and 2.1 shows the results from these tests. The two sided test reveals that only the similarity distribution between JI_0 and JI_1 has a significant ($p<0.01$) difference in distribution. The other Jaccard Index pairs show now significant difference in distribution. The tree cover similarity distribution of JC_1 is significantly greater than JC_0 ($p<0.005$) as the results from the one sided test show in the second table. Therefore the exclusion of canopy densities < 11 fosters the overall agreement between both tree cover datasets. Further the test also reveals that the similarity distribution of JC_2 is significantly greater than JC_1 ($p<0.05$) but by the comparison of JC_1 and JC_2 shows no clear trend in a certain direction. It is to assume that JC_2 improves only the tree cover agreement for certain tile pairs and not in general. The same accounts for JC_3. For studies targeting Americas it is to recommended to use from the Global Forest Change dataset data which lays within the canopy densities greater than 10 percent.

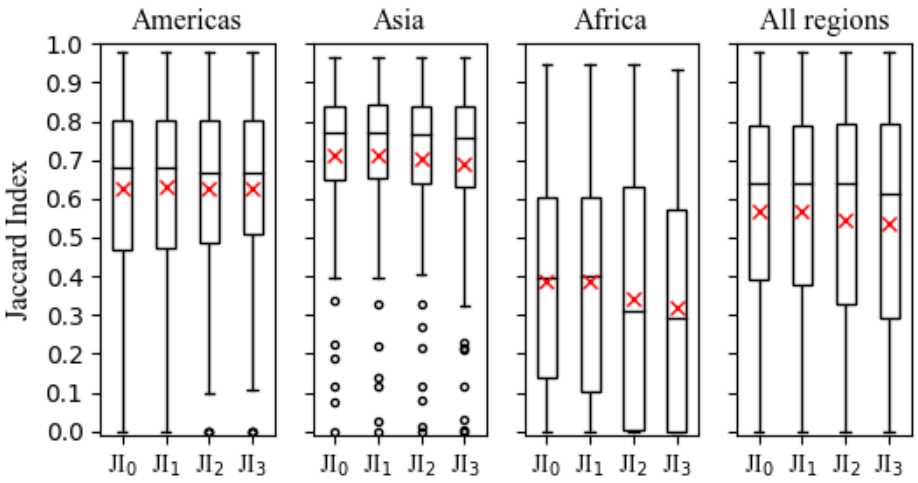

Figure 2.1. Tree cover similarity distribution over the continental regions: This boxplot shows the distribution of computed Jaccard Index for each raster image tile pair of GlobeLand30 and Global Forest Change tree cover from 2000. The labels JI_0, JI_1, JI_2, and JI_3 on the x-axis account for the canopy density classes (0,100], (10,100], (20,100], and (30,100], respectively. The y-axis is the computed Jaccard Index for the corresponding raster image pair, where 0 is a total disagreement and 1 a total agreement. Red crosses within the Q_{25}, Q_{50}, and Q_{75} boxes highlight the sample mean. Whiskers are 1.5 times the IQR.

Table 2.1. Regional two-sided Wilcoxon signed-rank test: This table shows, regional differences in the tree cover agreement by considering different canopy densities between GlobeLand30 and Global Forest Change at 2000. The classes JI_0, JI_1, JI_2, and JI_3 as row and column headings account for the canopy density classes (0,100], (10,100], (20,100], and (30,100], respectively. The test hypothesis is H_0: $X_1 = X_2$ where X_1 is the column JI_n class and X_2 the row JI_n class. The significance is indicated by $p^* < 0.05$, $p^{**} < 0.02$, and $p^{***} < 0.01$.

	Americas			Asia			Africa		
Cls	JI_0	JI_1	JI_2	JI_0	JI_1	JI_2	JI_0	JI_1	JI_2
JI_0	-	-	-	-	-	-	-	-	-
JI_1	$.00^{***}$	-	-	.72	-	-	.22	-	-
JI_2	.06	.36	-	$.00^{***}$	$.00^{***}$	-	$.03^{*}$	$.03^{*}$	-
JI_3	.16	.50	.60	$.00^{***}$	$.00^{***}$	$.00^{***}$	$.00^{***}$	$.00^{***}$	$.00^{***}$

Table 2.2. Regional one-sided Wilcoxon signed-rank test: This table shows, the direction of regional differences in the tree cover agreement by considering different canopy densities between GlobeLand30 and Global Forest Change at 2000. The classes JI_0, JI_1, JI_2, and JI_3 as row and column headings account for the canopy density classes (0,100], (10,100], (20,100], and (30,100], respectively. The test hypothesis is H_0: $X_1 \leq X_2$ and H_0: $X_2 \geq X_1$ where X_1 is the column JI_n class and X_2 the row JI_n class. The significance is indicated by $p^* < 0.05$, $p^{**} < 0.025$, $p^{***} < 0.01$, and $p^{\dagger} < 0.005$.

	Americas				Asia				Africa			
Cls	JI_0	JI_1	JI_2	JI_3	JI_0	JI_1	JI_2	JI_3	JI_0	JI_1	JI_2	JI_3
JI_0	-	$.00^{\dagger}$	$.03^{*}$	.08	-	.64	1.	1.	-	.11	.98	1.
JI_1	1.	-	.18	.25	.36	-	1.	1.	.89	-	.99	1.
JI_2	.97	.82	-	.30	$.00^{\dagger}$	$.00^{\dagger}$	-	1.	$.02^{**}$	$.01^{**}$	-	1.
JI_3	.92	.75	.70	-	$.00^{\dagger}$	$.00^{\dagger}$	$.00^{\dagger}$	-	$.00^{\dagger}$	$.00^{\dagger}$	$.00^{\dagger}$	-

Asia (review): As figure 2.1 suggest does the sample mean is approximately 0.7 for all canopy classes in asia. It decreases slightly at higher canopy density intervals. The median is approximately 0.8 by showing also a slight decrease when the canopy density class is raised. The similarity of the upper 25 percent of the samples is approximately 0.85 and 0.96 and the maximum similarity decreases slightly from 0.9654 to 0.9634 by increasing canopy density interval. This suggests as the appendix figure shows that high ranking tiles are not largely impacted by changes in the canopy density. For Asia the range of the lower 25 percent of the samples is quite large. It ranges between approximately 0.65 and 0.0 and as appendix shows the mobility of the samples show an overall downward trend. The iqr for the first two canopy density classes ranges between 0.65 and 0.85. The exclusion of lower canopy densities have not an large impact on the distribution of tree cover similarity. For last two canopy density classes the range between q1 and q3 is increasing. The exclusion of higher canopy densities leads to overall downwards trend within this intervals figure appendix. For asia the two sided wilcoxon test in table 2.1 reveals that the similarity distribution between every sample pair is significantly different (p<0.01) except the pair of JI_1 and JI_0. The directional test in table 2.2 shows that the overall tree cover agreement of JI_2 and JI_3 is significantly smaller than JI_0 and JI_1 (p<0.005). The direction of distribution differences between JC_0 and JC_1 is not

clearly deduce able. This could be explained over a large variability within regional tree cover agreement also a clearer picture could be achieved by applying smaller canopy density exclusion stepping. For studies in asia it is to recommended to use from the Global Forest Change dataset data which lays within the canopy densities greater than 10 percent or the entire data range. Here it must be decided if moving the canopy density threshold below the GlobaLand30 forest cover definition is a good trade for increased sample size.

Africa (review): As figure 2.1 for Africa suggests is the similarity distribution mobility of the samples in Africa at highest. The first two similarity distributions have an comparable mean and median at 0.38 and 0.4. The last two classes show a strong decline in mean and median to approximately 0.33 and 0.3. The mobility of the upper 25 percent of the first two canopy density classes is already quite strong as figure appendix suggests. While the iqr range for JI_0 is between 0.15 and 0.6 the range increases for the second JI_1 by connecting to more agreement downwards trend. The tile pairs in africa are characterized by 17 (approx. 20 percent) where the tree cover agreement is smaller than 0.1. 11 of these tiles have already a agreement of 0.0 when the first canopy density is excluded. These trend continues as more canopy density is excluded as more samples have a agreement of 0.0. This explains the high iqr range for the canopy densities exclusion greater than 10 percent. At JI_3 already two fifth (42 percent) of the samples have an tree cover agreement lower than 0.1. The two sided wilcoxon test shows that the distribution of JI_2 and JI_3 significantly differs ($p<0.05$ and $p<0.01$). As the table suggest is the distribution of JI_0 and JI_1 nearly the same. The one sided test reveals that reducing the canopy density below 10 percent decreases the tree cover agreement between sample distributions. For JI_0 and the JI_1 the test reveals that no clear trend is detectable. Appendix figure shows that some samples benefit from the exclusion of canopy density and as mentioned some samples show a strong decrease in tree cover similarity. As the data shows is the regional dependency of tree cover agreement at largest. To maximize the similarity on continental level for africa it should the entire data range of Global Forest Change should be slected. It could also be also a solution to select canopy densities in the interval 10 to 100 but here the trade is to have tiles where no tree cover agreement is detectable.

Comparison between regions (review): Table appendix shows that Asia has the highest tree cover similarity distribution over all regions within all tested canopy classes followed by the Americas for GlobeLand30 and Hansen 2000. Africa has the poorest tree cover agreement within our tested regions. As discussed in the previous paragraphs the reason could be that Americas and Asia have mainly core tropical forest zones where the forest cover is dense and the canopy density is above 30 percent. This can be highlighted by section tropical deforestation and the shown tree cover maps. Africa in comparison to Asia and Americas has large zones where the tree cover is high but the canopy density is low also described by the term sparse woodland. As it looks these sparse woodlands compete the forest detection

methods of both datasets. This fact could lead within the hansen dataset to ghost deforestation because it is hard to detect sparse woodland it could be the annual deforestation does not detect it as forest and it is recognized as deforestation. Therefore in africa the rationality of tree cover agreement must be considered during preparation and validation of studies. Also it is to suggest to optimize tree cover agreement on regional scale and not over the entire continent. Overall regions the upper 25 percent of the samples benefit or shown only small changes if the canopy density is increased.

All regions (review): To deduce in which canopy density interval we use the Global Forest Change data for our global study on proximate deforestation driver we analyzed the change of tree cover agreement over all samples. Figure 2.1 shows on the right hand side of the image the distribution of tree cover agreement over the entire sample range. The mean and median shown only a slight decline overall canopy density test classes. As deduced for the regions also on global scale the upper 25 percent commonly benefit or show no change if the canopy density is increased. As the figure shows is the range of them between 0.8 and 1.0. When the canopy density is increased the range of the mid 50 percent increases which highlights a more similarity downwards trend within this group. This downwards trend of the mid is connected with a decrease in the range of the lower 25 percent. The first canopy density class is between 0.4 and 0.0 and the last is between 0.3 ad 0.0. The distribution comparison in table 2.3 shows that each sample class has significant differences (p<0.02 and p<0.01) except JI_0 and JI_2 where the similarity distribution could be the same. Table 2.4 shows the direction of the distribution differences. Globally the tree cover agreement is highest when we consider only pixels within the canopy density interval (10,100] as the table shows. The distributions of tree cover agreement of JI_0, JI_2, and JI_3 are all significantly smaller than JI_1 (p<0.005 and p<0.01). By knowing this we decided to proceed for our study only with data within this tree cover interval. Therefore we filtered forest loss and gain of Global Forest Change to lay in this interval and classified them subsequently by superimposing the GlobeLand30 land cover layer. The result of this process are shown in section 2.1.3.

Table 2.3. Regional two-sided Wilcoxon signed-rank test: This table shows, global differences in the tree cover agreement by considering different canopy densities between GlobeLand30 and Global Forest Change at 2000. The classes JI_0, JI_1, JI_2, and JI_3 as row and column headings account for the canopy density classes (0,100], (10,100], (20,100], and (30,100], respectively. The test hypothesis is H_0: $X_1 = X_2$ where X_1 is the column JI_n class and X_2 the row JI_n class. The significance is indicated by $p^* < 0.05$, $p^{**} < 0.02$, and $p^{***} < 0.01$.

Cls	JI_0	JI_1	JI_2
JI_0	-	-	-
JI_1	$.00^{***}$	-	-
JI_2	.08	$.02^{**}$	-
JI_3	$.00^{***}$	$.00^{***}$	$.00^{***}$

Table 2.4. Global one-sided Wilcoxon signed-rank test: This table shows, the direction of global differences in the tree cover agreement by considering different canopy densities between GlobeLand30 and Global Forest Change at 2000. The classes JI_0, JI_1, JI_2, and JI_3 as row and column headings account for the canopy density classes (0,100], (10,100], (20,100], and (30,100], respectively. The test hypothesis is H_0: $X_1 \leq X_2$ and H_0: $X_2 \geq X_1$ where X_1 is the column JI_n class and X_2 the row JI_n class. The significance is indicated by $p^* < 0.05$, $p^{**} < 0.025$, $p^{***} < 0.01$, and $p^{\dagger} < 0.005$.

Cls	JI_0	JI_1	JI_2	JI_3
JI_0	-	$.00^{****}$	.96	1.
JI_1	1.	-	.99	1.
JI_2	$.04^{*}$	$.01^{***}$	-	1.
JI_3	$.00^{****}$	$.00^{****}$	$.00^{****}$	-

2.1.2 Tree cover and deforestation

Goal: Americas: Asia: Africa:

2.1.3 Proximate deforestation driver

2.1.4 Accuracy assessment

Goal (review): Goal is the assessment of the accuracy of our proximate deforestation driver predictions. We created a set of ground truth data by sampling our proximate deforestation driver layers. In each region we select per random 10 tiles and draw 200 samples per tile. The 200 samples comprises pixels over the full value range of our proximate deforestation driver classes. We imported the prepared sample to our JavaScript application and subsequently classified each sample with a label. To determine the accuracy we used a confusion matrix and the derived metrics like producers accuracy, overall accuracy, kappa coefficient etc.

Results (review): Table 2.5 shows the confusion matrix to determine the accuracy of our predictions where the term reference refers to the labeling of pixel by our visual interpretation and predictions refer to the labeling of our proximate driver predictions. The abbreviations PAc, UAc, OvAc, Com, Om, Tot, and Kappa refer to the terms Producers-Accuracy, Users-Accuracy, Overall-Accuracy, Error of Commission, Error of Omission, row or column total, and Kappa Coefficient. From the 6000 samples we draw from our study extent 14 %, 20 %, 22 %, 32 %, 8 %, 2 %, 0.5 %, 2 %, and 0.5 % account for cultivated land (10), tree cover (20), regrowth (25), shrubland (40), wetland (50), water (60), artificial land (80), and bareland (90), respectively. Our method predicts a distribution of 15 %, 18 %, 27 %, 31 %, 7 %, 1 %, 0.8 %, 1 %, and 0.2 % for the land cover classes 10, 20, 25, 30, 40, 50, 60, 80, and 90, respectively. Highest producers accuracy is achieved the regrowth class with 88 percent. The prediction of this class was achieved by including global forest change gain data within our target canopy density. During building our reference data set per visual interpretation

we determined this class by following these rules. We checked the surroundings and the corresponding pixel for signs of road networks and infrastructure. Further we checked if the canopy shows signs of age class forest and line patterns which show the establishment of artificial introduced forest cover. During the visual interpretation we recognized that a large portion of the regrowth class is occupied by plantations especially in Americas and Asia. For Asia a major share was occupied by palm oil or other plantations. It must be mentioned that here the deforestation within our temporal frame is not coercively the clearing of natural cover it shows it shows also rotational cycles of deforestation reforestation for commodities. The second highest accuracy with 85 (producers accuracy) was achieved in the prediction of cultivated land where only 15 percent could be identified as error of commission. 8.2 percent are classified as forest or regrowth which could reveal zones of shifting agriculture or temporal issues.

- grassland in majority with anthropogenic influence at many sides at water hole was detectable
- decision between regrowth and natural tree cover, relayed on expert knowledge how homogeneous is canopy
- shrub land comprises natural landscapes, young plantations, and areas for cattle ranching
- strong variability of images some high res some landsat quality

Table 2.5. Confussion matrix for accuracy assessment: We draw 6000 samples from 10 random selected tiles from the three regions Americas, Asia and Africa. Labels refer to our proximate deforestation driver classes which correspond to GlobeLand30 classification schema in table 1.2. Reference refers to the samples we classified by visual interpretation of external imagery and predictions refer to the label the sample has in our proximate driver product. The abbreviations PAc, UAc, OvAc, Com, Om, Tot, and Kappa refer to the terms Producers-Accuracy, Users-Accuracy, Overall-Accuracy, Error of Commission, Error of Omission, row or column total, and Kappa Coefficient.

		Reference											
	Cls	10	20	25	30	40	50	60	80	90	Tot	UAc	Om
Prediction	10	730	37	62	15	16	2	3	5	0	870	.84	.16
	20	41	744	56	189	31	12	0	15	4	1092	.68	.32
	25	29	202	1155	172	22	10	5	11	4	1610	.72	.28
	30	36	187	32	1466	73	21	0	17	0	1832	.80	.20
	40	14	21	4	41	352	1	1	2	1	437	.81	.19
	50	0	5	3	10	4	50	0	1	0	73	.68	.32
	60	2	1	0	3	0	2	18	2	0	28	.64	.36
	80	3	3	0	1	1	1	0	40	0	49	.82	.18
	90	0	0	0	1	0	0	0	3	5	9	.56	.44
	Tot	855	1200	1312	1898	499	99	27	96	14	6000		
	PAc	.85	.62	.88	.77	.71	.51	.67	.42	.36	Kappa		OvAc
	Com	.15	.38	.12	.23	.29	.49	.33	.58	.64	.69		.76

2.2 Emissions

2.3 Ecosystem service values

3 Discussion and Conclusion

3.1 Software, design and technology

3.2 Deforestation

3.2.1 Forest definition

- For a regional approach a better solution could be to select for each region independently the right canopy density. For America a good agreement between the tree cover could be achieved by selecting the second class. For Asia by selecting the first class and for africa the second. Even better would be to decide per tile individually which canopy density should be selected. This would eliminate regional effects of different forest densities.
- discuss regions independently asia and america have large tree cover agreement
- africa has the lowest agreement only core forest zones show high similarity
- We can see at which regions Chen et al switched their tree cover definition to 10
- To improve we should apply for each tile a canopy class decision based on our analysis
- This could improve the improve the similarity (accuracy) by maximizing the sample count
- Algorithm draft for single similarity: Compute jaccard indexes for tile pair at different canopy densities, put results in a list, sort the list in decreasing order, pick the class where jaccard index is max
- Use this jaccard method to exclude tiles where the tree cover similarity fall below a certain threshold

3.2.2 Tree cover and deforestation

3.2.3 Proximate deforestation driver

- Reclassification is not a good idea cause this approach leads not to consistent results

3.2.4 Accuracy assessment

- Is largely subjective because it is prepared from the study author
- Better if someone independent does it
- Even better if you have ground truth prepared by field studies
- Class variability errors source the reclassification
- Time frame of our classification we classified our ground truth data at google image data
- GL30 classifies with data from 2010
- It is hidden what happened between 2001 till 2010 except that it was deforested

3.3 Emissions

3.4 Ecosystem service values

- resilience of esv loss could be achieved over optimizing total value of the new land-use
- target optimization is use the clearcut by maximizing profit and minimizing the esv loss

3.5 Binning analysis and visualization

- Cut polygon by line the Scala (1992) approach explained with parametric separation function, and bezier
- A approach where ratio is also ratio of the hexagon area

Bibliography

Arsanjani J. J., See L., and Tayyebi A. Assessing the suitability of GlobeLand30 for mapping land cover in Germany. *International Journal of Digital Earth*, 9(9):873–891, March 2016a. doi: 10.1080/17538947.2016.1151956.

Arsanjani J. J., Tayyebi A., and Vaz E. GlobeLand30 as an alternative fine-scale global land cover map: Challenges, possibilities, and implications for developing countries. *Habitat International*, 55:1–7, July 2016b. doi: 10.1016/j.habitatint.2016.02.003.

Baccini A., Goetz S. J., Walker W. S., Laporte N. T., Sun M., Sulla-Menashe D., Hackler J., Beck P. S. A., Dubayah R., Friedl M. A., Samanta S., and Houghton R. A. Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. *Nature Climate Change*, 2(3):182–185, January 2012. doi: 10.1038/nclimate1354.

Baccini A., Walker W., Carvahlo L., Farina M., Sulla-Menashe D., and Houghton R. Tropical forests are a net carbon source based on new measurements of gain and loss. Online accessed through Global Forest Watch, 2015. URL `https://www.globalforestwatch.org`.

Baccini A., Walker W., Carvalho L., Farina M., Sulla-Menashe D., and Houghton R. A. Tropical forests are a net carbon source based on aboveground measurements of gain and loss. *Science*, 358(6360):230–234, September 2017. doi: 10.1126/science.aam5962.

Cao X., Li A., Lei G., Lei G., Tan J., Zhang Z., Yan D., Xie H., Zhang S., and Yang Y. Land cover mapping and spatial pattern analysis with remote sensing in Nepal. *Journal of Geo-information Science*, 18:1384–1398, 2016.

Carr D. B. Looking at large data sets using binned data plots. resreport, U.S. Department of Energy, 1990.

Carr D. B., Littlefield R. J., Nicholson W. L., and Littlefield J. S. Scatterplot matrix techniques for large N. *Journal of the American Statistical Association*, 82(398):424–436, June 1987. doi: 10.1080/01621459.1987.10478445.

Carr D. B., Olsen A. R., and White D. Hexagon mosaic maps for display of univariate and bivariate geographical data. *Cartography and Geographic Information Systems*, 19(4):228–236, January 1992. doi: 10.1559/152304092783721231.

Chen J., Chen J., Liao A., Cao X., Chen L., Chen X., He C., Han G., Peng S., Lu M., Zhang W., Tong X., and Mills J. *30-meter Global Land Cover Dataset - Product Description*. National Geomatics Center of China, May 2014.

Chen J., Chen J., Liao A., Cao X., Chen L., Chen X., He C., Han G., Peng S., Lu M., Zhang W., Tong X., and Mills J. Global land cover mapping at 30 m resolution: A POK-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103:7–27, 2015. doi: 10.1016/j.isprsjprs.2014.09.002. URL `http://www.globallandcover.com`.

Chen J., Cao X., Peng S., and Ren H. Analysis and applications of GlobeLand30: a review. *ISPRS International Journal of Geo-Information*, 6(8):230, July 2017. doi: 10.3390/ijgi6080230.

Congalton R. G. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37(1):35–46, July 1991. doi: 10.1016/0034-4257(91)90048-b.

Costanza R., Groot R., de, Sutton P., Ploeg S., van der, Anderson S. J., Kubiszewski I., Farber S., and Turner R. K. Changes in the global value of ecosystem services. *Global Environmental Change*, 26:152–158, May 2014. doi: 10.1016/j.gloenvcha.2014.04.002.

Groot R., de, Brander L., Ploeg S., van der, Costanza R., Bernard F., Braat L., Christie M., Crossman N., Ghermandi A., Hein L., Hussain S., Kumar P., McVittie A., Portela R., Rodriguez L. C., Brink P., ten, and Beukering P., van. Global estimates of the value of ecosystems and their services in monetary units. *Ecosystem Services*, 1(1):50–61, July 2012. doi: 10.1016/j.ecoser.2012.07.005.

Don A., Schumacher J., and Freibauer A. Impact of tropical land-use change on soil organic carbon stocks - a meta-analysis. *Global Change Biology*, 17(4):1658–1670, November 2010. doi: 10.1111/j.1365-2486.2010.02336.x.

FAO. FRA 2015 terms and definitions. resreport, Food and Agriculture Organization of the United Nations, 2012.

FAO and ITPS. Global Soil Organic Carbon Map. resreport, FAO and ITPS, 2018.

Foody G. M. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80(1):185–201, April 2002. doi: 10.1016/s0034-4257(01)00295-4.

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., and Townshend J. R. G. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, 342(6160):850–853, November 2013. doi: 10.1126/science.1244693. URL `https://earthenginepartners.appspot.com/science-2013-global-forest/download_v1.0.html`.

Hijmans R., Garcia N., Kapoor J., Rala A., Maunahan A., and Wieczorek J. GADM database of Global Administrative Areas. Online, May 2018. URL `https://www.gadm.org`. Version 3.6.

Hoshen J. On the application of the enhanced Hoshen-Kopelman algorithm for image analysis. *Pattern Recognition Letters*, 19(7):575–584, May 1998. doi: 10.1016/s0167-8655(98)00018-x.

Jaccard P. The distribution of the flor in the alpine zone. *The New Phytologist*, 11(2), February 1912.

Jacobson A., Dhanota J., Godfrey J., Jacobson H., Rossman Z., Stanish A., Walker H., and Riggio J. A novel approach to mapping land conversion using Google Earth with an application to East Africa. *Environmental Modelling & Software*, 72:1–9, October 2015. doi: 10.1016/j.envsoft.2015.06.011.

Li Y., Sulla-Menashe D., Motesharrei S., Song X.-P., Kalnay E., Ying Q., Li S., and Ma Z. Inconsistent estimates of forest cover change in China between 2000 and 2013 from multiple datasets: differences in parameters, spatial resolution, and definitions. *Scientific Reports*, 7(1), August 2017. doi: 10.1038/s41598-017-07732-5.

Lowry R. *Concepts and applications of inferential statistics*. Vassar College, 2019. URL `http://www.vassarstats.net/textbook/`.

McKinney W. Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 2010.

Potapov P., Hansen M. C., Laestadius L., Turubanova S., Yaroshenko A., Thies C., Smith W., Zhuravleva I., Komarova A., Minnemeyer S., and Esipova E. The last frontiers of wilderness: Tracking loss of intact forest landscapes from 2000 to 2013. *Science Advances*, 3(1), January 2017. doi: 10.1126/sciadv.1600821. URL `http://www.intactforests.org/`.

Sampat M. P., Wang Z., Gupta S., Bovik A. C., and Markey M. K. Complex wavelet structural similarity: a new image similarity index. *IEEE Transactions on Image Processing*, 18(11): 2385–2401, November 2009. doi: 10.1109/tip.2009.2025923.

Seydewitz T. Applicability of GlobeLand30 and Global Forest Change data products for forest land cover change studies on global and regional scales. resreport, Potsdam Institute for Climate Impact Research, 2017. Internship - report.

Shi G. R. Multivariate data analysis in palaeoecology and palaeobiogeography - a review. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 105(3-4):199–234, November 1993. doi: 10.1016/0031-0182(93)90084-v.

Siikamaki J., Santiago-Avila F. J., and Vail P. Global assessment of non-wood forest ecosystem services. resreport, Program on Forests (PROFOR), 2015.

Rossum G., van and Development T. The Python language reference: release 3.5.6. Online, 2018. URL `https://docs.python.org/3.5/download.html`. Python Software Foundation.

Wilcoxon F. Individual comparisions by ranking methods. *Biometrics Bulletin*, 1(6):80–83, December 1945.

Yang Y., Xiao P., Feng X., and Li H. Accuracy assessment of seven global land cover datasets over China. *ISPRS Journal of Photogrammetry and Remote Sensing*, 125:156–173, March 2017. doi: 10.1016/j.isprsjprs.2017.01.016.

List of Figures

List of Tables

List of Abbreviations

AGB	Aboveground live woddy Biomass density
AISM	Aligned Image Stack Mosaic
API	Application Programming Interface
CRS	Coordinate Reference System
CSV	Comma Separated Values
DEM	Digital Elevation Map
DT	Decision Tree
ESV	Ecosystem Service Values
ETM+	Enhanced Thematic Mapper Plus
FAO	Food and Agriculture Organization of the United Nations
FTP	File Transfer Protocol
GADM	Global Administrative Areas Map
GFC	Global Forest Change
GFW	Global Forest Watch
GIS	Geographic Information System
GL30	GlobeLand30
GLAS	Geoscience Laser Altimeter System
GSOCmap	Global Soil Organic Carbon map
GSP	Global Soil Partnership
GTiff	Geo-Tiff
GeoJSON	Geographic JavaScript Object Notation
IFL	Intact Forest Landscapes
IPCC	Intergovernmental Panel on Climate Change
ISRIC	International Soil Reference and Information Center
ITPS	Intergovernmental Technical Panel on Soils
LC	Land Cover
LCC	Land Cover Change
LIDAR	Light Detection and Ranging
LU	Land Use
LUC	Land-use Change
LULC	Land Use/Land Cover
ME	Mean Error
MLC	Maximum Likelihood Classifier
MLR	Multiple Linear Regression
MODIS	Moderate Resolution Imaging Spectroradiometer
NDVI	Normalized Difference Vegetation Index
POK	Pixel-Object-Knowledge

REGEX	Regular Expression
RF	Random Forest
RMSE	Root Mean Square Error
SD	Standard Deviation
SHP	Shapefile
SOC	Soil Organic Carbon
SOCC	Soil Organic Carbon Content
SVM	Support Vector Machine
UN	United Nations
URL	Uniform Resource Locator
UTM	Universal Transverse Mercator
WGS84	World Geodetic System 1984
stdlib	Standard Library
GDAL	Geospatial Data Abstraction Library
JI	Jaccard Index
PDD	Proximate Deforestation Driver
WSRT	Wilcoxon signed-rank test
WRST	Wilcoxon rank-sum test

Appendix

Table 3.1. Comparison of tree cover agreement between regions: This table shows, a comparison of tree cover agreement between regions. The classes JI_0, JI_1, JI_2, and JI_3 as row and column headings account for the canopy density classes (0,100], (10,100], (20,100], and (30,100], respectively. The test hypothesis is H_0: $X_1 = X_2$ where X_1 is the column JI_n class and X_2 the row JI_n class. The significance is indicated by $p^* < 0.05$, $p^{**} < 0.02$, and $p^{***} < 0.01$.

		Americas				Asia			
	Cls	JI_0	JI_1	JI_2	JI_3	JI_0	JI_1	JI_2	JI_3
Asia	JI_0	$.02^*$	-	-	-	-	-	-	-
	JI_1	-	$.02^*$	-	-	-	-	-	-
	JI_2	-	-	$.03^*$	-	-	-	-	-
	JI_3	-	-	-	$.05^*$	-	-	-	-
Africa	JI_0	$.00^{***}$	-	-	-	$.00^{***}$	-	-	-
	JI_1	-	$.00^{***}$	-	-	-	$.00^{***}$	-	-
	JI_2	-	-	$.00^{***}$	-	-	-	$.00^{***}$	-
	JI_3	-	-	-	$.00^{***}$	-	-	-	$.00^{***}$

Table 3.2. Comparison of tree cover agreement between regions: This table shows, a comparison of tree cover agreement between regions and the direction of differences. The classes JI_0, JI_1, JI_2, and JI_3 as row and column headings account for the canopy density classes (0,100], (10,100], (20,100], and (30,100], respectively. The test hypothesis is H_0: $X_1 \leq X_2$ where X_1 is the column JI_n class and X_2 the row JI_n class. The significance is indicated by $p^* < 0.05$, $p^{**} < 0.025$, $p^{***} < 0.01$, and $p^{\dagger} < 0.005$.

		Americas				Asia			
	Cls	JI_0	JI_1	JI_2	JI_3	JI_0	JI_1	JI_2	JI_3
Asia	JI_0	.99	-	-	-	-	-	-	-
	JI_1	-	.99	-	-	-	-	-	-
	JI_2	-	-	.99	-	-	-	-	-
	JI_3	-	-	-	.99	-	-	-	-
Africa	JI_0	$.00^{\dagger}$	-	-	-	$.00^{\dagger}$	-	-	-
	JI_1	-	$.00^{\dagger}$	-	-	-	$.00^{\dagger}$	-	-
	JI_2	-	-	$.00^{\dagger}$	-	-	-	$.00^{\dagger}$	-
	JI_3	-	-	-	$.00^{\dagger}$	-	-	-	$.00^{\dagger}$

Wyrażam zgodę na udostępnienie mojej pracy w czytelniach Biblioteki SGGW w tym w Archiwum Prac Dyplomowych SGGW.

I agree to share my work in the reading rooms of the SGGW Library, including the SGGW Theses Archive.

Ich erteile meine Zustimmung zur Veröffentlichung meiner Arbeit in der Bibliothek der SGGW (Warschauer Naturwissenschaftliche Universität), einschließlich des Archivs der Diplomarbeiten.

. . . . . .

(czytelny podpis autora pracy)
(legible signature of the author)
(lesbare Unterschrift des Autors der Arbeit)