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

**Oświadczenie promotora pracy**

Oświadczam, że niniejsza praca została przygotowana pod moim kierunkiem i stwierdzam, że spełnia warunki do przedstawienia tej pracy w postępowaniu o nadanie tytułu zawodowego.

**Declaration of the promoter**

I declare that this thesis was prepared under my supervision and I state that it meets the conditions for presenting such a body of work in the process of obtaining a professional title.

**Erklärung des Betreuers**

Hiermit erkläre ich, dass die vorliegende Arbeit, unter meiner Leitung erstellt wurde und ich bestätige, dass sie die Bedingungen zur Verleihung des Abschlussdiploms erfüllt.

| | | | |
|---|---|---|---|
| Data<br>Date<br>Datum | ...................... | Podpis promotora pracy<br>Signature of the promoter<br>Unterschrift des Betreuers | ...................... |

## Oświadczenie autora pracy

Świadom odpowiedzialności prawnej, w tym odpowiedzialności karnej za złożenie fałszywego oświadczenia, oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami prawa, w szczególności ustawą z dnia 4 lutego 1994 r.o prawie autorskim i prawach pokrewnych (Dz. U. Nr 90 poz. 631 z późn. zm.).

Oświadczam, że przedstawiona praca nie była wcześniej podstawą żadnej procedury związanej z nadaniem dyplomu lub uzyskaniem tytułu zawodowego test. Lorem ipsum dolores anno sacntum

Oświadczam, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną. Przyjmuję do wiadomości, że praca dyplomowa poddana zostanie procedurze antyplagiatowej.

## Declaration of the author

Aware of the legal liability, including criminal liability for submitting a false statement, I declare that this thesis was written by myself alone and does not contain content obtained in a manner breaking applicable laws, in particular the Act of February 4, 1994 on copyright and related rights (Journal of Laws, no. 90, item 631, as amended).

I certify that the work has not previously been the basis for any procedure in connection with obtaining a diploma or professional title.

I declare that this version of the work is identical with the attached electronic version.

I acknowledge that the thesis is subject to anti-plagiarism procedures.

## Erklärung des Autors

Gesetzlicher Haftpflicht, besonders der strafrechtlichen Verantwortlichkeit für Abgabe der falschen Erklärung bewusst, erkläre ich hiermit, dass vorliegende Diplomarbeit selbständig angefertigt wurde und keinen Inhalt enthält, der widerrechtlich erworben wurde, insbesondere nicht mit dem Gesetz über Urheberrecht vom 4. Februar 1994 (GB. Nr. 90, Pos. 631 mit späteren Änderungen) übereinstimmend.

Ich erkläre auch, dass die Arbeit bisher keiner anderen Prüfungsbehörde vorgelegt wurde.

Der Durchführung einer elektronischen Plagiatsprüfung stimme ich hiermit zu. Die eingereichte elektronische Fassung der Arbeit entspricht der eingereichten schriftlichen Fassung exakt.

| | | | |
|---|---|---|---|
| Data | ..................... | Podpis autora pracy | ..................... |
| Date | | Signature of the author | |
| Datum | | Unterschrift des Autors | |

## Streszczenie

**Tytuł: Text**

Text

Słowa kluczowe: Text

## Summary

**Title: Text**

Text

Keywords: Text

## Zusammenfassung

**Titel: Text**

Text

Schlüsselwörter: Text

# Contents

# 1 Data and methods

In this chapter we describe our approach to answer the scientific questions made in the introduction. The first section of this chapter introduces the datasets used during this study. For each dataset, we shortly describe by which approach it is derived and what are the fundamental properties of it. 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.

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 Standard Library (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.

## 1.1 Data

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. Each dataset is introduced in a extra section following the order of appearance in the table.

**Table 1.1. Datasets used during this study:** The source column contains the reference to the corresponding release publication. If the data is provided as a download the reference contains the corresponding 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 | |
| Soil Organic Carbon Change | empirical | Don et al. [2010]<br>de Groot et al. [2012] |
| Ecosystem Service Values | empirical | Costanza et al. [2014]<br>Siikamaki et al. [2015] |

### 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). 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 (logitude) 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, lossyear, and gain. The data acquisition is automatized with an Python script by using the 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 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.

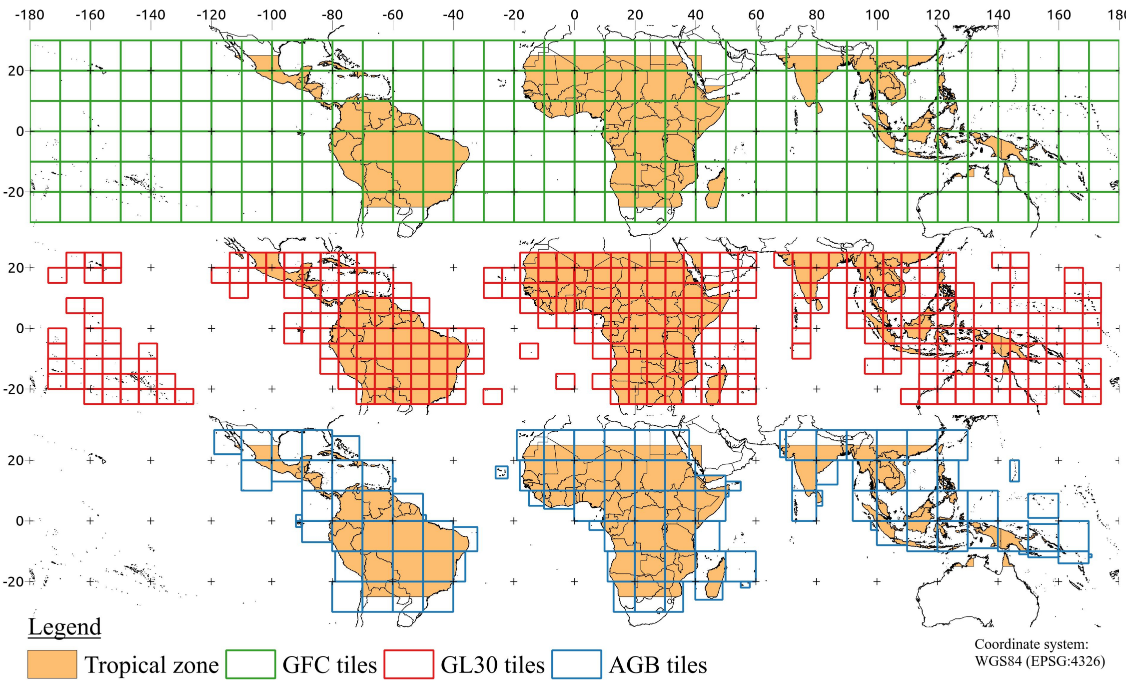


**Figure 1.1. Map of downloaded dataset tiles:** This map shows the acquired image tiles for this study. From top to bottom in green Global Forest Change (GFC) dataset tiles (Treecover2000, lossyear and gain), the land cover dataset GlobeLand30 (GL30) image tiles in red, and in blue the Aboveground Biomass (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 in 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 bareland 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 an 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 a 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 an 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 an 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 was automatized with a throw away 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

A Intact Forest Landscapes (IFL) is a mosaic of undisturbed forest patches and naturally treeless ecosystems without sings 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 of 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 trough 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 an SHP so you must only download a single compressed archive. The download is automatized with an 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 $ha^{-1}$ (mega gram 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 an mosaic of GTiff raster images where each tile of the mosaic has the Coordinate Reference System (CRS) WGS84 and is coded in 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 a 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 public available on the homepage of GFW. As mentioned, the dataset covers only the tropical zone, therefore we acquires 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 send to this API the server response 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 is 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 have 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 are defined as natural vegetation without human impacts which includes natural grassland and shrubland. Secondary forest are 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 corresponds to the GL30 and IFL classification schema. The actual values are shown in table 1.3.

**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 |

### 1.1.7 Ecosystem Service Values

Some text to ensure document flow.

**Table 1.4. Selection of Ecosystem Service Values (ESV) per biome used in this study:** ESV per biome connected with the corresponding GlobeLand30 land-cover class, 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 | Code | Type | Dg | Co | Wb |
|---|---|---|---|---|---|
| Cropland | 10 | Cropland | - | 5,567 | - |
| Forest tropical | 20 | Forest | 5,264 | 5,382 | 1,312 |
| Forest tropical | 25 | Regrowth | 5,264 | 5,382 | 1,312 |
| Grass/Rangelands | 30 | Grassland | 2,871 | 4166 | - |
| Wetlands | 50 | Wetland | 25,682 | 140,174 | - |
| Lakes/Rivers | 60 | Water bodies | 4,267 | 12,512 | - |
| Urban | 80 | Artificial | - | 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 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. A further analysis revealed that all tiles located in UTM zone 1 and 60 overflowing the maximum and minimum longitude coordinates of this 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 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 8-bit unsigned integer. 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 it pictured 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.

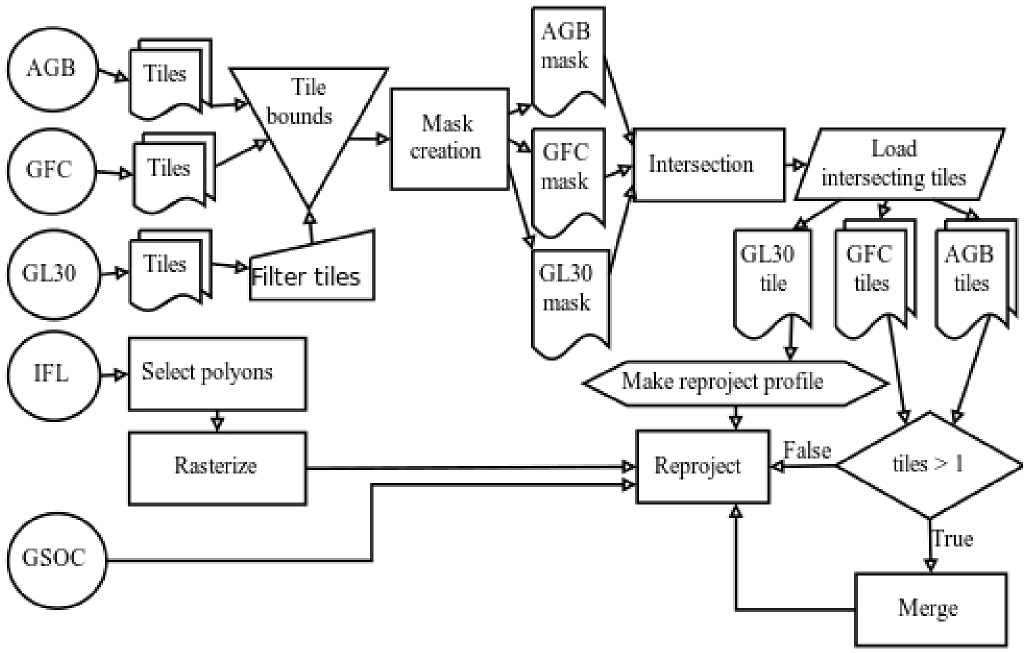


**Figure 1.2. Tile alignment algorithm:** For the multi-tiled datasets (multi-document symbols) a mask is created by extracting the tile bounds. Next, the intersection between theses masks is determined to identify superimposing data and the corresponding tiles are loaded from disk. GL30 tiles are used as template by creating 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.

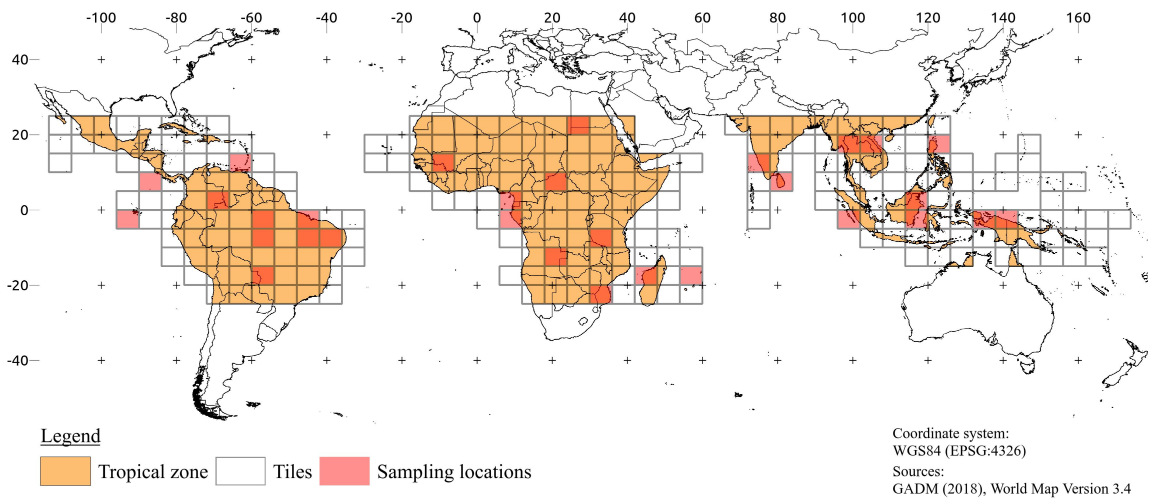


**Figure 1.3. Aligned raster images and sampling locations:** The map shows the location of the aligned multi-image stack tiles as black-framed square sized polygons, the sampling locations for accuracy assessment in red, and countries within the tropical zone in orange.

### 1.2.2 Deforestation

#### 1.2.2.1 Forest definition

We aggregate two layers. Both have slightly different definitions for tree cover. To successful aggregate the layers we must harmonize the tree cover situation between the gfc tree cover and the gl30 2000 treecover. Because we want to map changes both have to agree on tree cover as well on losses. Additionally we must develop an reliable process to compare the overall similarity between these images. Idea is if both treecover at 2000 are nearly similar the should capture the same loss at 2010. Also we compared the forest loss agreement between both layers.

To determine the similarity between GL30 2000 and GFC reference tree-cover we used Jaccard Index (JI). The JI or coefficient of community is a simple measure of similarity between two binary paired populations 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 number where both images have an agreement represented as a pixel value of one.

The JI is always within the closed interval $[0,1]$, where a 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].

First step is to filter gl30 map for tree cover and set all values to one. Also we must set all values of the gfc tree cover layer to one. Next we compute by equation bla the jaccard index. After we exclude in gfc tree cover the canopy density below 11 and compute another times the jaccard indexs. This we execute till we computed the following jaccard indexes: $\mathrm{JI}_0[0,100]$, $\mathrm{JI}_1(10,100]$, $\mathrm{JI}_2(20,100]$, $\mathrm{JI}_3(30,100]$ the interval describes the canopy densities.

$$\mathbf{JI_{coeff}} = \begin{bmatrix} |\mathbf{X_1} \wedge \mathbf{X_2}| & |(\mathbf{X_1} \wedge \mathbf{X_2}) \oplus \mathbf{X_2}| \\ |(\mathbf{X_1} \wedge \mathbf{X_2}) \oplus \mathbf{X_1}| & |(\neg\mathbf{X_1}) \wedge (\neg\mathbf{X_2})| \end{bmatrix} \tag{1.1}$$

$$JI = \frac{|\mathbf{X_1} \wedge \mathbf{X_2}|}{|\mathbf{X_1} \vee \mathbf{X_2}|} \tag{1.2}$$

Hypothesis testing comparing similarities in tree cover: After computing the jaccard index for each tile we tested if the tree cover similarity increases if we exclude a certain canopy density class. By using R and the Wilcoxon signed rank test we compared the similarities between the ji combinations.

Similarity treecover loss:

#### 1.2.2.2 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 haadamard 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. Figure 1.4 shows an overview of the classification process. The classification algorithm is implemented as a Python function which requires as 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 reclassify the misclassified pixels based on a approximated probability. We define mis-classified pixels as sites where the GFC annual loss data predicts a deforestation but the GL30 stratum still classifies them as forest. First step of our reclassification is to cluster the mis-classified pixels with the Hoshen-Kopelman algorithm. The clustering algorithm is implemented as a part of Geospatial Data Abstraction Library (GDAL) and can be called trough 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 side length of 500 x 500m around the polygon centroid (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 in equation 1.3 to determine the on-ground resolution in meter on pixel level. Where $d$ is the great-circle distance between two latitude, longitude pairs $\varphi_n, \lambda_n$ 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

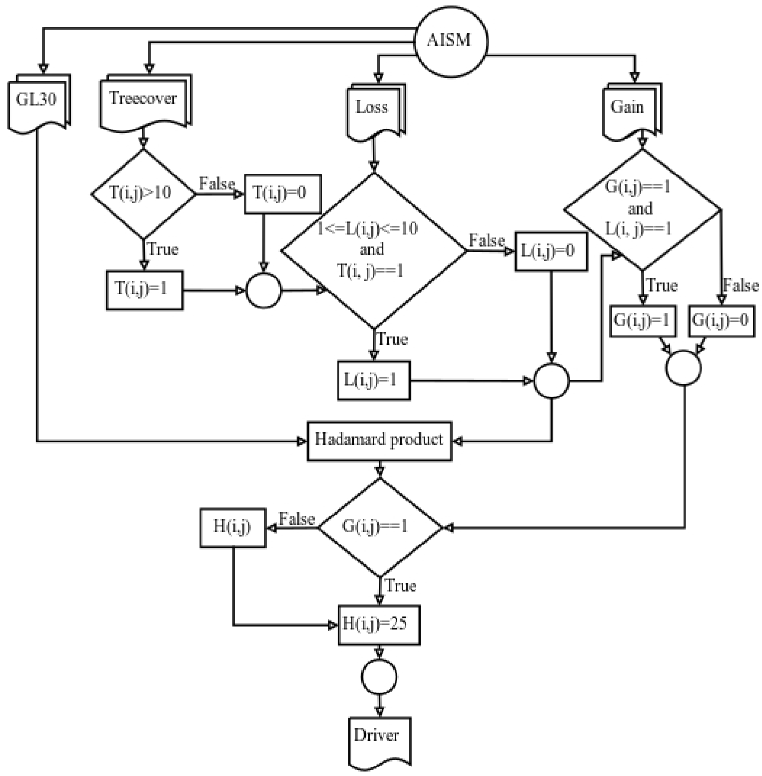


**Figure 1.4. Classification of proximate deforestation drivers:** For the classification of the proximate deforestation drivers the following layers are required GL30 2010, GFC treecover, annual losses, and gain. From treecover we select all pixels within the canopy density interval $(10, 100)]$. The treecover 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

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.

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right)+\cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right) \tag{1.3}$$

#### 1.2.2.3 Accuracy assessment

For examining the accuracy of our proximate deforestation driver prediction we used a confusion matrix (also known as two-way frequency tables, error matrix or contingency tables). These matrices are commonly used for an accuracy assessment of land cover classifications and enables the computation of marginal and conditional distributions [Congalton 1991; Foody 2002]. Table 1.5 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 1.5. 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 (Latin America, Africa, Asia/Oceania). 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 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 a 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, affine matrix of the corresponding raster image, and the seed is 42. The per tile samples are stored as an Comma

Separated Values (CSV) file.

For the collection of ground-truth data we used visual interpretation of Google Earth satellite and aerial imagery. 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 trough 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 trough the sample total $N$ (equation 1.4), the producer accuracy $PAc_{.n}$ by dividing the category agreement trough the column category total (equation 1.5), the error of commission $Com_{.n}$ (Type II error) by dividing the category disagreement trough the column category total (equation 1.6), the user accuracy $UAc_{n.}$ by dividing the category agreement trough the row category total (equation 1.7), the error of omission $Om_{.n}$ (Type I error) by dividing the category disagreement trough the row category total (equation 1.8), and the Cohens Kappa by substituting equation 1.9 and 1.4 into equation 1.10.

$$p_0 = OvAc = \frac{\sum_{i=1}^{n} x_{i,i}}{N} \tag{1.4}$$

$$PAc_{.n} = \frac{x_{i,i}}{x_{.n}} \tag{1.5}$$

$$Com_{.n} = \frac{FN_i}{x_{.n}} \tag{1.6}$$

$$UAc_{n.} = \frac{x_{i,i}}{x_{n.}} \tag{1.7}$$

$$Om_{n.} = \frac{FP_i}{x_{n.}} \tag{1.8}$$

$$p_c = \frac{1}{N^2} \sum_{i=1}^{n} x_{.i} \cdot x_{i.} \tag{1.9}$$

$$Kappa = \frac{p_0 - p_c}{1 - p_c} \tag{1.10}$$

### 1.2.3 Emissions

#### 1.2.3.1 Above ground biomass

For computing the loss and emissions from aboveground biomass we need the following layers: proximate deforestation drivers and the biomass map. From the proximate deforestation drivers map the total treecover loss is extracted by setting all values greater than zero to one. For converting carbon content to co2 we used the factor 3.7. The total aboveground biomass emissions from deforestation is computed by the following formula. We excluded from our calculation all pixels classified as forest.

$$AGBE_{tile} = 3.7A\sum_{i=0}^{N}\sum_{j=0}^{M} AGB(i,j)Px(i,j) \tag{1.11}$$

Where agbe is the total aboveground biomass emission, agb the carbon content per pixel in mg/ha, res the pixel resolution of the tile in ha, 3.7 for converting carbon in co2, and px the pixel value at this position. After totaling we aggregated the emissions per continent.

#### 1.2.3.2 Soil organic carbon change

For computing the soil organic carbon change we used the following layers from ou aism: gsoc, proximate deforestation drivers and intac forest landscape. We computed soil organic carbon change in two different scenarios: scenario one (sc1) assumes that all forest losses are primary forest and scenario two distincts between loss of primary and secondary forest. To compute the soc we used the following formular. For distinct between primary and secondary forest loss we used the ifl layer and assumed that forest loss within a intact forest path is a loss in primary forest and losses outside of the patch are loss in secondary forest. The soil organic carbon change can only be computed for land use change class which are represented by the values given in table bla. This means we linked the following proximate driver classes with the following soc change factors: cropland with cropland, grassland with tundra, grassland, shrubland, bareland. In result this means we excluded from calculation the following landuse change classes: water, artificial, and wetland.

$$SOCE_{tile} = 3.7A\sum_{i=0}^{N}\sum_{j=0}^{M} SOC(i,j) * SOCC(LUCC(i,j), F_{type}) \tag{1.12}$$

### 1.2.4 Ecosystem service values

For computing the ecosystem service balances we just relayed on our proximate deforestation drivers predictions. The loss of ecosystem service values is just the sum of deforested pixels multiplied by the ecosystem service value of tropical forest of a certain dataset. we excluded all pixels classified as forest from the loss sum. for computing the ecosystem service gain we used the deforestation driver prediction classes. now we depend on what the empirical data provides for the certain type.

### 1.2.5 Binning analysis

The previous sections were focused on the generation of large scale spatial data. Now, a feasible method must be developed for analyzing, aggregating, interpreting, and visualizing our results. For the development of a proper approach we have to generalize the problem domain. At first we are confronted with large N (many samples) which results in a high dimensionality and complexity of relationships among this variables [Carr 1990]. From a visual/analytical perspective georeferenced raster maps can be interpreted as a multivariate scatter plot of large datasets where longitude and latitude represent the x and y coordinate of an data point and the pixel values (in this case nominal scaled) representing the third dimension as an group coloring. Therefore we have a large multidimensional dataset combined with a scatter plot visualization which leads commonly to over plotting issues and hidden point densities [Carr et al. 1987]. Due to the spatial nature of your data we are also confronted with not equal distributed data some regions show high data densities and other regions have sparse to no data. Also a severe problem domain is the frame size of our representation. Goal is to present data on a continental level which intensifies visual problems. Each pixel has a resolution of approximately 30x30m, the continental representation of americas spanning approximately 1200000x120000km2. Therefore small scale isolated changes are hidden and only large scale changes are visual detectable. Which results in hidden details and not perceivable patterns of change.

Goal should be to develop an process who solve this issues and generates satisfying output for our multivariate data. In case of raster data a re-sampling to coarser resolution could solve over plotting and resolution issues as well normalize the unequal distributed data. But the nature of re-sampling (for nominal data a nearest neighbor or majority wins Reference) would negate important spatial patterns as well frequency distributions. Another well known approach is to use binning of the spatial explicit data with a certain kind of regular polygon that is tessellating the plane [Carr et al. 1992]. Polygon tessellations provide numerous opportunities for presenting multivariate statistical summaries. The scaling of the polygon could be used to represent pixel densities within the polygon area, a polygon filling color gradient is applicable to show nominal or ordinal scaled data. Also it is imaginable to use the polygon interior for a pie chart. To use regular tessellation it is important to mention there are only three types of regular polygons tessellate the plane: squares, equilateral triangles and hexagons [Carr et al. 1992]. Square tessellation is the most common approach used for binning in spatial visualization. A raster image is a square tessellation. In a square mosaic each polygon shares 4 edge neighbors and 4 vertex neighbors more explanation error distance disadvantages etc Hexagons properties, advantages disadvantages of both tessellations. Final goal is to show your analysis results of spatial explicit raster data in hexagonal binned form. For bivariate maps we choose a visual representation with scaled hexagons and colorization. For multivariate details we choose a pie chart alike visualization. We split the hexagons horizontal in regards of the presented ratio. The ratios should be ordered descending so that the greatest ratio is south oriented. It is following a general description how we created the hexagon grids and how we tackled the polygon split problem.

To be flexible at hexagon construction we accept 4 different parameters as construction arguments: $D$ long diagonal (Diameter of the circumscribing circle), $d$ short diagonal (diameter of the inscribed circle), $A$ area the hexagon should span and or $e$ the edge length. One selected parameter of these is used to compute $R$ the radius of the circumscribing circle with respect to input parameter as shown in equation 1.13. R is used to calculate the midpoint $< c_x, c_y >$ of the hexagon located in the first quadrant of the cartesian coordinate system Equation 1.14 and 1.15. Equation 1.16 shows the computation of the hexagon anti-clockwise vertex matrix. Whereas the two leftmost vertices (first and last row of the matrix $\mathbf{H}$) are located at koordinatenursprung, will sagen auf deutsch korridanten at x=0 und y=value of matrix. In summary equation 1.13 to 1.16 show the creation of an hexagon at the leftmost corner of first quadrant (Figure 1.5). The orientation is important for the subsequent mosaic creation.

$$R = \frac{\sqrt{2A}}{\sqrt[4]{27}} = \frac{D}{2} = \frac{d}{\sqrt{3}} = e \tag{1.13}$$

$$c_x = \frac{R\sqrt{3}}{2} \tag{1.14}$$

$$c_y = R \tag{1.15}$$

$$\mathbf{H} = \begin{bmatrix} 0 & c_x & 2c_x & 2c_x & c_x & 0 \\ R\sin\left(\frac{7\pi}{6}\right)+c_y & 0 & R\sin\left(\frac{11\pi}{6}\right)+c_y & R\sin\left(\frac{\pi}{6}\right)+c_y & 2R & R\sin\left(\frac{5\pi}{6}\right)+c_y \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{1.16}$$

A polygon tessellation needs several polygons to create a grid in case of the creation of one hexagon with the presented algorithm needs approximately benchmark but the creation of several N hexagons needs approximately benchmark. Therefore it is much simpler to create only one hexagon with the presented algorithm and to create the grid polygons by copying the coordinates of the source polygon and translating them to their target position with a affine transformation matrix shown in equation 1.17. To create the grid we get the rectangular bounds of the area to tessellate as a matrix $\mathbf{B} \in R^{2\times 2}$ (equation 1.18), where the first column of the matrix contains the lower left corner and the second column the upper right corner of the image. Each subsequent translation in regards of $x_{off}$ is $x_1 + d$ for even rows and bla bla for odd rows. $Y_{off}$ is computed by bla bla see figure 1.5.

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & x_{off} \\ 0 & 1 & y_{off} \\ 0 & 0 & 1 \end{bmatrix} \circ \mathbf{H} \tag{1.17}$$

$$\mathbf{B} = \begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \end{bmatrix} \tag{1.18}$$

Binning of raster data is easy we just have a point in polygon problem each points/pixels

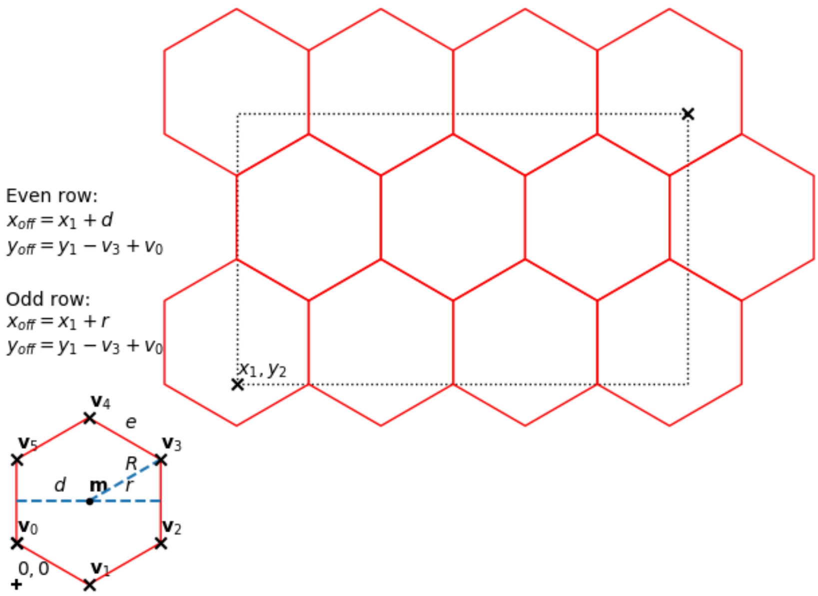


**Figure 1.5. Hexagon tessellation:** Located at the left bottom corner in red a hexagon defined by its geometric properties the 6 vertex vectors $\{\vec{v}_0, \ldots, \vec{v}_5\}$ (black crosses), with center vector $\vec{m}$, edge length $e$, $R$ radius of the circumscribing circle, $r$ radius of the inscribed circle and $d$ the length of the short diagonal. 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.

falling in hexagon are counted and aggregated through a function. In case of drivers of deforestation we count all driver classes per hexagon and compute ratios next we compute the sha describe for each map how you build it As mentioned before for the visualization of the drivers of deforestation map we want to segment the hexagons with horizontal lines and each segment should represent the share of the direct deforestation driver within the tessellated area. To compute the split line for a certain hexagon we need the hexagon R computeable from the area of the hexagon equation 1.19 and the rectangular bounds of the hexagon. We compute the relative share of an deforestation driver per hexagon this relative share can be used to compute the y-axis coordinate of an split line equation 1.20. A regular hexagon can not only be presented in it vertex form as shown above. We can also use functions to define the hexagon shape. A hexagon consist of 2 picewise functions where each function consist of 3 linear functions restricted to an intervall. If we invert these functions we can use these functions to compute the x-coordinate of the split line with the previous computed y-coordinate Equations 1.21 and 1.22. As a results we receive the solution matrix L which represents the horizontal line segment splitting the hexagon at the point where we want (driver ratio share) equation 1.23. The solution matrix can be plugged in to a polygon split function which separates the hexagon polygon in a upper and lower part to do so we iterate over the hexagon vertices and decide if they are above or under the split line and append to a lower upper polygon. These list are our results explain better split function.

$$R = \frac{\sqrt{2A}}{\sqrt[4]{27}} \tag{1.19}$$

$$y = \frac{P(y_2 - y_1)}{100} + y_1 \tag{1.20}$$

$$f^{-1}(y) = \begin{cases} -\frac{y-y_1}{\tan\left(\frac{\pi}{6}\right)} + \frac{x_1+x_2}{2} & \text{if } y_1 \leq y < y_1 + R\sin\left(\frac{5\pi}{6}\right) \\ x_1 & \text{if } y_1 + R\sin\left(\frac{5\pi}{6}\right) \leq y < R(\sin\left(\frac{5\pi}{6}\right) + 1) \\ \frac{y-y_2}{\tan\left(\frac{\pi}{6}\right)} + \frac{x_1+x_2}{2} & \text{if } R(\sin\left(\frac{5\pi}{6}\right) + 1) \leq y \leq y_2 \end{cases} \tag{1.21}$$

$$g^{-1}(y) = \begin{cases} \frac{y-y_1}{\tan\left(\frac{\pi}{6}\right)} + \frac{x_1+x_2}{2} & \text{if } y_1 \leq y < y_1 + R\sin\left(\frac{5\pi}{6}\right) \\ x_2 & \text{if } y_1 + R\sin\left(\frac{5\pi}{6}\right) \leq y < R(\sin\left(\frac{5\pi}{6}\right) + 1) \\ -\frac{y-y_2}{\tan\left(\frac{\pi}{6}\right)} + \frac{x_1+x_2}{2} & \text{if } R(\sin\left(\frac{5\pi}{6}\right) + 1) \leq y \leq y_2 \end{cases} \tag{1.22}$$

$$\mathbf{L} = \begin{bmatrix} f^{-1}(y) & g^{-1}(y) \\ y & y \end{bmatrix} \tag{1.23}$$

# 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*, 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 xxx*, pages 1–7, 2016b. doi: 10.1016/j.habitatint.2016.02.003.

Baccini A., Goetz S., Walker W., Laporte N., Sun M., Sulla-Menashe D., Hackler J., Beck P., Friedl M., Samanta S., and Houghton R. Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. *Nature Climate Change*, 2, March 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., and D. Sulla-Menashe R. A. H. Tropical forests are a net carbon source based on aboveground measumeasure of gain and loss. *Science*, 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), 1987.

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), 1992.

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. *International Journal of Geo-Information*, 6(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 the Environment*, 46:37–46, 1991.

Costanza R., Groot R., de, Sutton P., Ploeg S., van der, Anderson S. J., Kubiszewski I., Farber S., and Turner K. R. Change in the global value of ecosystem services. *Global Environmental Change*, 26:152–158, 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 the services in monetary units. *Ecosystem Services*, 1:50–61, 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*, 80:185–201, 2002.

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.

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 and Software*, 72:1–9, 2015.

Li Y., Sulla-Menashe D., Motesharrei S., Kalnay X.-P. S., 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(8748), August 2017. doi: 10.1038/s41598-017-07732-5.

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, 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), November 2009.

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, 1993.

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.

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, 2017.

# 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 |

# Appendix

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)*