

Evolution of the representation of global vegetation by vegetation continuous fields

Charlene DiMiceli^{a,*}, John Townshend^a, Mark Carroll^b, Robert Sohlberg^a

^a Department of Geographical Sciences, University of Maryland, College Park, MD, USA

^b Code 606, NASA Goddard Space Flight Center, Greenbelt, MD, USA

ARTICLE INFO

Keywords:

Fractional vegetation cover
MODIS
Machine learning
Land cover
Vegetation continuous fields
Land remote sensing

ABSTRACT

Two decades of global annual fractional vegetation cover products have been derived using daily surface reflectance and land surface temperature data from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard NASA's Terra satellite. These MODIS Vegetation Continuous Fields (VCF) products are unique and a distinct advance over discrete (ordinal) land cover characterizations. VCF consists of three continuous fractional cover layers at 250 m resolution summing to 100%: tree cover, non-tree vegetation and non-vegetated surface plus surface water. This paper describes the fully-automated machine learning algorithm used to create MODIS VCF and details recent algorithm improvements. Comparisons of the latest VCF products in the MOD44B collection 6, with field-based fractional tree cover and other remote sensing fractional cover products demonstrate improved agreement: the root mean square (RMS) deviations range from 9 to 23%. A comparison of MODIS canopy cover values with those from the Global Ecosystem Dynamics Investigation (GEDI) gives an R^2 value of 0.53. Automation of the algorithm provides flexibility for readily porting the method to derive fractional cover products from additional instruments with similar radiometric characteristics such as the Visible Infrared Imaging Radiometer (VIIRS) payload aboard Suomi NPP and JPSS platforms, and successors.

1. Introduction

Accurate maps of vegetation cover are fundamental for environmental and climate-related applications. Vegetation is a key component of the biosphere. It is an important regulator of biogeochemical cycles including water and carbon and an integral part of Earth's energy balance. It is a dominant component of most ecosystems, providing critical ecosystem services such as wildlife habitat, food, fuel and timber production, reduction of erosion, and flood control. Vegetation maps provide the primary means of measuring global forest cover change (e.g., Hansen et al., 2013; Kim et al., 2014, 2015), providing parameters for climate and carbon land surface models (Lawrence and Chase, 2007; Tian et al., 2004; Jung et al., 2006), and depicting key variables for the determination of biomass and forest structure (Cartus et al., 2012; Lefsky, 2010).

The benefits of characterizing vegetation type using continuous variables rather than discrete classes were recognized in the 1990s (DeFries et al., 1995). Continuous variables better describe the mixture of trees, herbaceous vegetation and non-vegetated surfaces to be found in most pixels at all resolutions. Fractional vegetation cover maps assign

fractional values to each pixel rather than assigning a single class. They also eliminate the need to translate between differing vegetation class schemes when using data from different sources, and do not have the artificially sharp representation of boundaries found in maps of discrete land cover types.

MODIS Vegetation Continuous Fields (VCF) products are the descendants of the earliest global fractional vegetation cover products, developed by DeFries et al. (1995) and Hansen et al. (2000) using data from the Advanced Very High Resolution Radiometer (AVHRR) at one degree, 8 km and 1 km spatial resolutions, albeit with improved inputs and methods.

The first MODIS instrument, onboard NASA's Terra platform, launched in 2000, was designed to provide much improved data for observing vegetation. Taking advantage of near-daily looks and finer resolution down to 250 m, MODIS VCF products were created to provide a continuous record of fractional cover for three land surface types: trees, non-tree vegetation and bare ground. The introduction of machine learning algorithms for these products offered a substantial improvement in accuracy and allowed the algorithms to be automated. Automation, in turn, ensures an internally consistent time series suitable for

* Corresponding author.

E-mail address: cdimicel@umd.edu (C. DiMiceli).

users in many scientific fields and for parameterization of biogeochemical, hydrological and weather models.

The MODIS production system periodically reprocesses scientific products, resulting in an updated 'collection.' Reprocessing provides the science teams the opportunity to submit improved algorithms and to address drift in calibration and orbit through time. Major milestone VCF collections are collection 3, which introduced the basic post-launch algorithms; collection 5, which included important updates; and collection 6, which demonstrated improved results due to updates to the upstream input products. MODIS VCF collection 6 data products are currently available for 2000 to present, and represent the most accurate MODIS fractional cover products to date.

The value of MODIS VCF products has been demonstrated in their widespread use by varied research communities. They have been used as input in the creation of 30 m resolution tree cover products from Landsat data (Hansen et al., 2013; Sexton et al., 2013), which provide the ability to monitor deforestation and forest degradation in a consistent manner globally (Sexton et al., 2013; Kim et al., 2014, 2015; Hansen et al., 2008) (Fig. 1). They are also being used to parameterize climate, carbon and land surface models (Lawrence and Chase, 2007; Tian et al., 2004; Jung et al., 2006), and to determine biomass and forest structure (Cartus et al., 2012; Lefsky, 2010). Studies of many crucial phenomena now depend on use of MODIS VCF products, such as fire occurrence (Giglio et al., 2006; Wiedinmyer et al., 2006), ecosystem functioning (Nightingale et al., 2011; Villagra et al., 2009), biodiversity (Coops et al., 2009; Hernández and Pastor, 2008), conservation (Sexton et al., 2016; DeFries et al., 2005), and public health (Calhoun, 2013; Cecchi et al., 2008).

2. Methods

The MODIS Vegetation Continuous Fields products are maps of annual 250 m global fractional vegetation cover (FVC) in three layers: tree cover, non-tree vegetation and non-vegetated ground (such as bare ground, rock, ice/snow, and water). Fractional vegetation cover is defined as the ratio of the area of the vertical projection of green vegetation onto the ground to the total area; it captures the horizontal arrangement and density of vegetation across the landscape.

2.1. Pre-MODIS development

The earliest global VCF map was created by DeFries et al. (1999) from AVHRR data at 1 km resolution using linear mixture models. Thirty metrics describing spectral and phenological characteristics were calculated and used to decompose pixels into tree cover, non-tree vegetation and non-vegetated components. A similar linear model was then used to discriminate between evergreen and deciduous, and broadleaf and needleleaf tree cover. This linear mixture model method was subsequently applied to create annual 8 km AVHRR FVC maps for each of the years 1982 to 1994 (DeFries et al., 2000). The current MODIS VCF algorithm, described below, has evolved from this research.

2.2. MODIS VCF overview

The first MODIS VCF products (collection 3) were produced at 500 m globally (Hansen et al., 2003), an improvement spatially over the 1 km AVHRR product produced by Hansen et al., 2000. Collection 3 products included only the tree cover layer (Hansen et al., 2003). Completed in 2008, collection 5 improved resolution to 250 m and added bare ground and non-tree vegetation layers (Carroll et al., 2011). Collection 6 was completed in 2018. It used the same algorithm and training data as collection 5 but was improved due to higher quality in upstream surface reflectance and brightness temperature products developed by Vermote and Wolfe (2015a, 2015b) used as inputs. An outline of the current VCF method is shown in Fig. 2, and is detailed in the following paragraphs.

2.3. Training data

MODIS VCF production depends on a supervised machine learning algorithm (described below). To create the necessary training data for the machine learning models, imagery from Landsat 5 TM, Landsat 7 ETM+, and Landsat 8 from 1998 to 2013 were used. This range of sources represents the imagery available as the training library was compiled and augmented throughout the MODIS VCF project. Collection 5 included a major addition to the training data library, roughly doubling the number of tree cover training scenes over collection 3, resulting in 68 scenes for tree cover training and adding 45 scenes for bare ground. Scenes were chosen to provide a sample representative of global ecosystems and vegetation cover types; every continent is represented across a range of latitudes, and selections were compared to an ecosystem map (Olson et al., 2001). Scenes at peak greenness without cloud cover were preferentially chosen to ensure they represent vegetation at its maximum coverage. To estimate tree cover and non-tree vegetation fractions, we performed manual classification on individual scenes, aided by image segmentation software, to provide training sets containing percent cover for each output layer. Landsat scenes were manually classified, then high resolution data from the NGA High Resolution Data Archive (Neigh et al., 2013) and Google Earth™ were used to verify the classification. Classification was done in the Landsat UTM projection at native resolution (30 m), then reprojected to the MODIS sinusoidal projection at 30 m resolution. Exact area averaging was then applied to derive the final training dataset on the MODIS sinusoidal 250 m grid to match the gridding of the input surface reflectance products.

2.4. MODIS data

Atmospherically-corrected, gridded surface reflectance and brightness temperature data for 2000 to 2018 from the MODIS Terra instrument were used as inputs to the VCF machine learning algorithm (Vermote and Wolfe, 2015a, 2015b). Because MODIS 250 m bands were not processed until the end of the fourth year after launch, Collection 3 VCF products were made using bands 1–7 at 500 m. Collections 5 and 6 combine bands 1–2 at 250 m, bands 5–7 at 500 m, and band 32 at 1 km in order to produce an improved product at 250 m (Table 1). MODIS provides several observations for each pixel; multiple observations occur where observations intersect multiple grid cells and where overpasses overlap. Quality flags are also included in the MODIS inputs and are used extensively in compositing (described in Section 2.5). Improvements in the quality flags, particularly in cloud and aerosol detection (Vermote and Kotchenova, 2008; Vermote, 2019) have significantly benefited MODIS VCF collections by better eliminating pixels where the ground is obscured due to high aerosols and cloud cover.

2.5. Compositing

MODIS gridded land surface products are available at a daily time-step, with multiple observations available per pixel each day. Compositing is performed to eliminate clouds, aerosols, and other poor quality data where possible and to choose observations with good grid cell coverage and near-nadir viewing geometry. This compositing process reduces data to a manageable size, selects high-quality data inputs for the modeling algorithm, and retains the phenological characteristics of the underlying vegetation (Carroll et al., 2011).

Early MODIS VCF products were produced using 40-day composites. With improved surface reflectance and quality flag inputs, composites were reduced to 32-day composites and then to 16-day composites in collections 5 and 6, resulting in improved VCF products. Collection 3 performed compositing using only the normalized difference vegetation index and blue band; in collections 5 and 6 a more comprehensive set of quality tests are used, resulting in great improvement (Table 2). Currently, compositing is performed to choose the best 250 m observations, at the same time storing matching (same day, orbit and grid

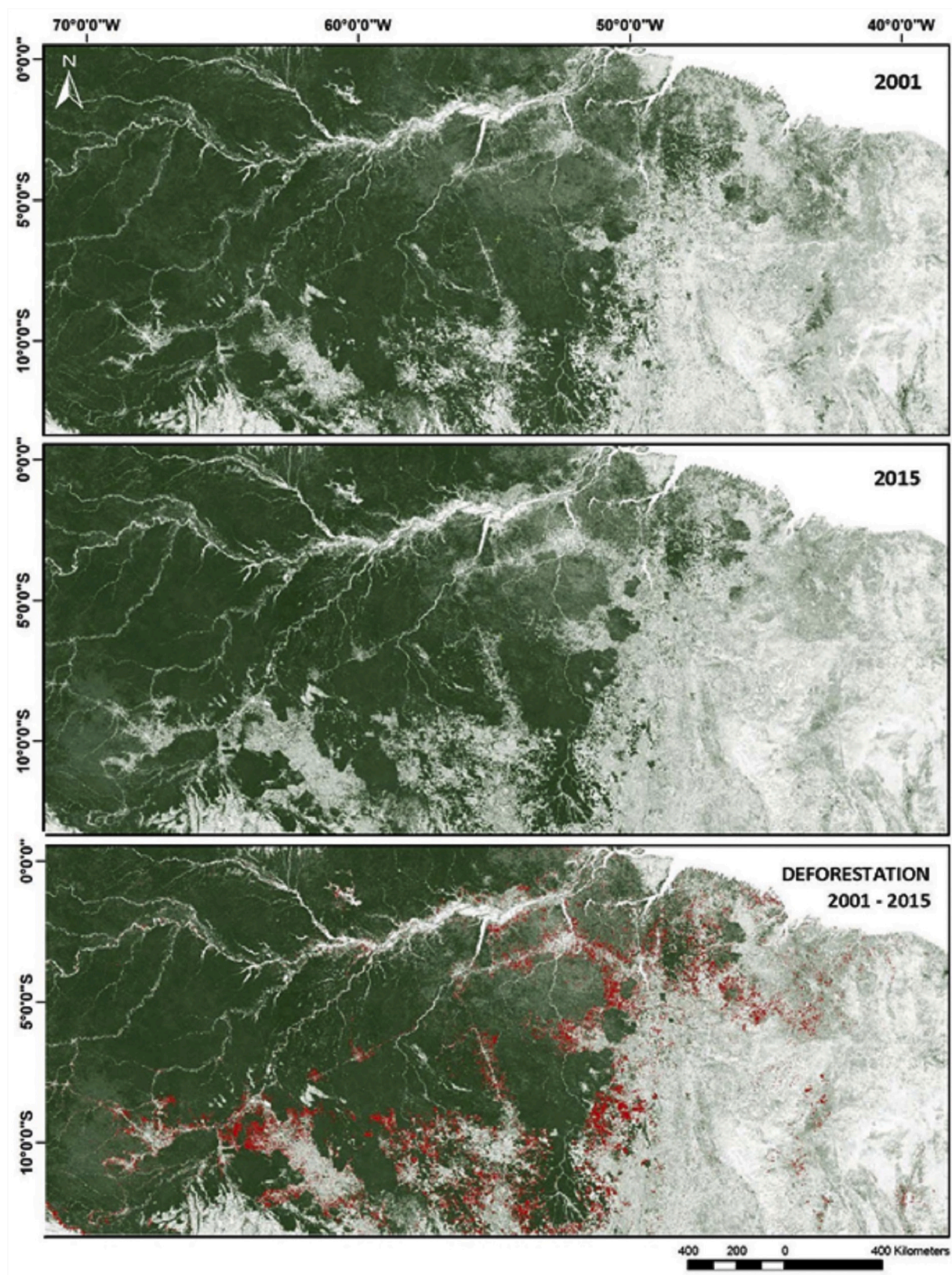


Fig. 1. Deforestation in the Amazon Basin, 2001–2015. MODIS Vegetation Continuous Fields percent tree cover, 2001 (top) and 2015 (center). Deforestation shown in red (bottom). Forest is defined as tree cover >30% and deforestation is defined as a 20% or greater decline in tree cover. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

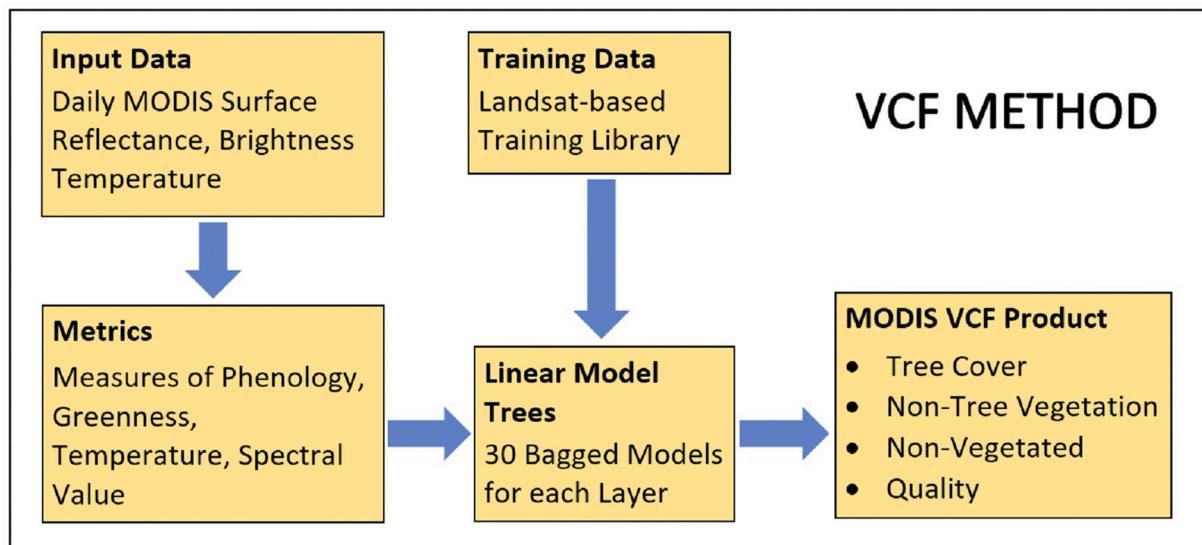


Fig. 2. Outline of machine learning methods used to create the MODIS Vegetation Continuous Fields product.

Table 1

MODIS bands used as input to the VCF machine learning algorithm.

Band	Bandwidth	Spatial Resolution [m]
Band 1 – Red	620–670 nm	250
Band 2 – Near Infrared	841–876 nm	250
Band 3 – Blue	459–479 nm	500
Band 4 – Green	545–565 nm	250
Band 5 – Short-wave Infrared	1230–1250 nm	500
Band 6 – Short-wave Infrared	1628–1652 nm	500
Band 7 – Short-wave Infrared	2105–2155 nm	500
Band 32 – Brightness Temperature	10.780–11.280 μ m	1000

Table 2

Quality tests for observation compositing.

Bit	Indicator	Value
0	Band 1 > 0 and QA non-fill value	1 = yes, 0 = no
1	Bands 2–7 > 0	1 = yes, 0 = no
2	Cloud	1 = no, 0 = yes
3	Solar zenith <65°	1 = yes, 0 = no
4	Cloud shadow	1 = no, 0 = yes
5	High aerosol (level 3)	1 = no, 0 = yes
6	Observation coverage >15% and view angle <45°	1 = yes, 0 = no
7	Observation coverage >10% and view angle <50°	1 = yes, 0 = no
8	Observation coverage >5% and view angle <55°	1 = yes, 0 = no
9	NDVI >90% of max NDVI	1 = yes, 0 = no
10–12	Cirrus level	0–3 (0 = highest, 3 = lowest)
13–15	Aerosol level	0–3 (0 = highest, 3 = lowest)

cell) observations for the 500 m and 1 km bands to produce 16-day composited data for input to the processing stream. The compositing algorithm assigns a score to each observation by setting quality bits (Table 2), then sorts observations based on the resulting 16-bit integer. The observation with the highest score is retained for each pixel.

2.6. Metrics

Metrics (or features) were created for each year using these pre-processed composites. These metrics are descriptors that capture

seasonality, maximum greenness, and minimum brightness temperature. Metrics are based on all bands listed in Table 1. Metrics include the time series of composited data in chronological order and data sorted by spectral value, greenness and temperature. We also include the means of several ranges of data in the sorted and unsorted datasets and take the differences of values occurring in various positions in the time series and sorted data. Metrics are used, together with the training data, to train the linear tree models. Once the models are built, the metrics are again fed into the models to predict vegetation cover for all pixels.

In collection 3, sixty-eight pre-selected metrics were used. In collections 5 and 6, in order to improve metrics, we began with a large set of candidate metrics and then used a semi-automatic process to reduce the number of metrics, eliminating strongly correlated variables, retaining only the metrics that best predicted the land surface. The metrics selection process is performed iteratively using both metrics and training data, by building the models, then eliminating metrics that are infrequently chosen during model creation and that do not occur in the highest levels of the regression trees. There is a balance between the number of metrics that can be selected and the number of training points that can be ingested by the data mining software, a function of physical memory limitations. At each step, as unnecessary metrics are eliminated, additional training is added, because model accuracy is improved with a larger training dataset. Error statistics produced during the model creation process are used to determine when the optimum balance between number of metrics and number of training data instances has been found. This balance is attained when removing metrics and increasing training points no longer improves model error statistics derived from cross-validation. Once a metrics set is chosen, metrics are computed for all pixels and used for both model training and prediction.

2.7. Regression tree models

The MODIS VCF machine learning algorithms are regression tree models (Quinlan, 1992). There are several advantages to using regression trees:

- Data prediction is fast.
- The tree is easily interpreted.
- Important variables are apparent.
- Linear Model Trees (LMTs) can accurately model nonlinear data with a ‘jagged’ regression surface.
- Fast, robust software libraries are available to create trees.

Collection 3 used a small regression tree model (109 terminal nodes) and was hand-pruned in a stepwise manner. In collection 5, linear regression functions were added at the nodes, a variation of decision trees known as a linear model tree, which more accurately captures variability in the data and reduces noisy speckle and patchiness. Collection 5 trees were also larger (approximately 4000 terminal nodes), which was made possible by a larger training dataset; this made models more accurate. Also in collection 5, bagging was added to improve the stability of predictions (Breiman, 1996); in the MODIS VCF process, the average of thirty LMTs was taken, using randomly sampled subsets of training data with corresponding metrics. Bagging also provides a confidence measure: a large standard deviation for the 30 predictions for a single pixel indicates poor predictive performance. Improvements in collection 5 were carried forward into collection 6.

For regression tree creation and prediction, the collection 3 algorithm used the Classification and Regression (CART) methodology introduced by Breiman et al., 1984. Regression trees are decision trees where the target variables are continuous. Breiman et al.'s monograph (1984) remains a good introduction to decision trees. CART has a number of advantages: 1) it is not dependent on the underlying statistical distribution of the variables, 2) it is not sensitive to outliers, 3) it can use each variable at multiple decision points in the tree, modeling complex non-linear relationships, and 4) cross-validation can be used to assess the accuracy of the model (Nisbet et al., 2009). Currently (collections 5 and 6), the M5' algorithm by Wang and Witten (1997) is used, an improved decision tree algorithm developed from an earlier M5 algorithm (Quinlan, 1992). This algorithm is similar to CART but incorporates methods to obtain linear model trees. Pruning is done by both algorithms using a stepwise process to eliminate overfitting. The M5' algorithm does this in an automated fashion; a pruning factor, controlling the severity of pruning, is adjusted experimentally to provide optimal model trees based on cross-validation statistics.

2.8. Software

An important update to MODIS VCF production is the full automation of the software. Early VCF products were created using S+ software, which was not amenable to insertion into the MODIS production stream. Collection 5 production takes advantage of newer software libraries to overcome this problem and provide full automation. MODIS VCF production software was written by the author to perform data management, compositing and metrics creation. Open source libraries were incorporated for model creation and prediction tasks. The software package was successfully ported to the MODIS Adaptive Processing System (MODAPS, resident at NASA Goddard Space Flight Center, Code 619) and is operational there. All of the libraries used by the code are open source, mature, and well maintained by large, active communities. Software is written in C and Java and uses the Java Native Interface to bridge between languages.

The WEKA data mining software package (Hall et al., 2009) was employed as the core data mining tool in the VCF software. WEKA can be accessed using a GUI front-end for easy data exploration, and individual functions can then be incorporated into user code. Written in Java, it is easily ported. Source code is available and well-organized, which provides users the opportunity to modify code for their own use. A small modification was made for this project to allow adjustments to the regression tree pruning severity, but otherwise the library was used as written. (Adjusting the pruning factor makes it possible to tune the model experimentally for better accuracy based on the cross-validation results.) WEKA is also well-documented, with a tutorial text in print in its 4th edition (Witten et al., 2016). Because users are encouraged to add new methods as they are developed by the machine learning community, WEKA includes a wide range of methods including decision tree inducers, rule learners, model tree generators, support vector machines, locally weighted regression, instance-based learning, bagging, boosting, and stacking, clustering methods, and an association rule learner. To be

included in the WEKA workbench, a method must be published in peer-reviewed literature and demonstrate increased performance over established methods. WEKA algorithms typically include solutions to common pitfalls such as finding global rather than local optimums and dealing with missing data. Software deals robustly with errors and is well-optimized for good runtime speed. Another important advantage of WEKA is that it can ingest the very large training and metrics data sets necessary to build accurate global land cover models.

2.9. Description of collection 6 VCF output files

One MODIS VCF product file was created for each year, containing tree cover, non-tree vegetation, and non-vegetated layers (Fig. 3) (DiMiceli et al., 2015). The output files are in the MODIS sinusoidal projection at 250 m resolution. The MODIS Land/Water Mask was used to flag inland water pixels in each layer (Carroll et al., 2009). The annual product also contains a quality layer and uncertainty layers corresponding to each vegetation layer, with values for each pixel. Layers are:

- Percent tree cover (woody plants greater than 5 m in height)
- Percent non-tree vegetation cover (all other green vegetation including shrubs and grasses)
- Percent non-vegetated (rock, soil, permanent ice and snow, and inert vegetable matter)
- Quality flags
- Continuous confidence value layers, one for each cover layer (3 layers total)

Layers are stored as scientific datasets (SDS) in tiled HDF format. Metadata is also stored in the file with version information, information about the input files, projection and file size, and overall quality. Data layers and metadata are accessible to users through a number of software packages (e.g., ENVI, ArcGIS, PCI Geomatica, GDAL, QGIS). MODIS VCF data products can be freely downloaded from the Land Processes Distributed Active Archive Center (<https://lpdaac.usgs.gov>). As products have improved, new MODIS VCF product versions have been released; version 6.0 is the recommended version at the time of writing.

2.10. Quantification of error and uncertainty

Two types of per-pixel quality variables are provided: a quality flag layer and confidence layers. The quality flag layer contains information about each of the 16-day composites that went into the model creation and prediction process for a single year. This information is contained in a bitfield for each pixel, the bits indicating whether there was good-quality data (i.e., cloud-free, good light conditions, etc.) in each composite. The confidence layers contain, for each pixel and layer, the standard deviation of the 30 values resulting from the bagged linear model trees. A large standard deviation for a pixel indicates poor agreement among the bagged models and therefore low confidence in the accuracy of the pixel value. The standard deviation is in units of percent cover. Further guidance for the use of these quality and confidence layers is provided in the MODIS VCF User's Guide (Carroll et al., 2017).

3. Product evaluation

In order to evaluate VCF product accuracy, we compared VCF tree cover outputs with canopy cover from the Global Ecosystem Dynamics Investigation L2B products. GEDI, aboard the International Space Station, collects waveform Lidar in 25 m footprint samples, spaced 60 m apart along transects spaced 600 m apart. Canopy cover for each footprint was derived by the GEDI team using a canopy directional gap probability algorithm (Tang and Armston, 2019). Taking these canopy estimates, we reprojected them to the MODIS grid, restricted the

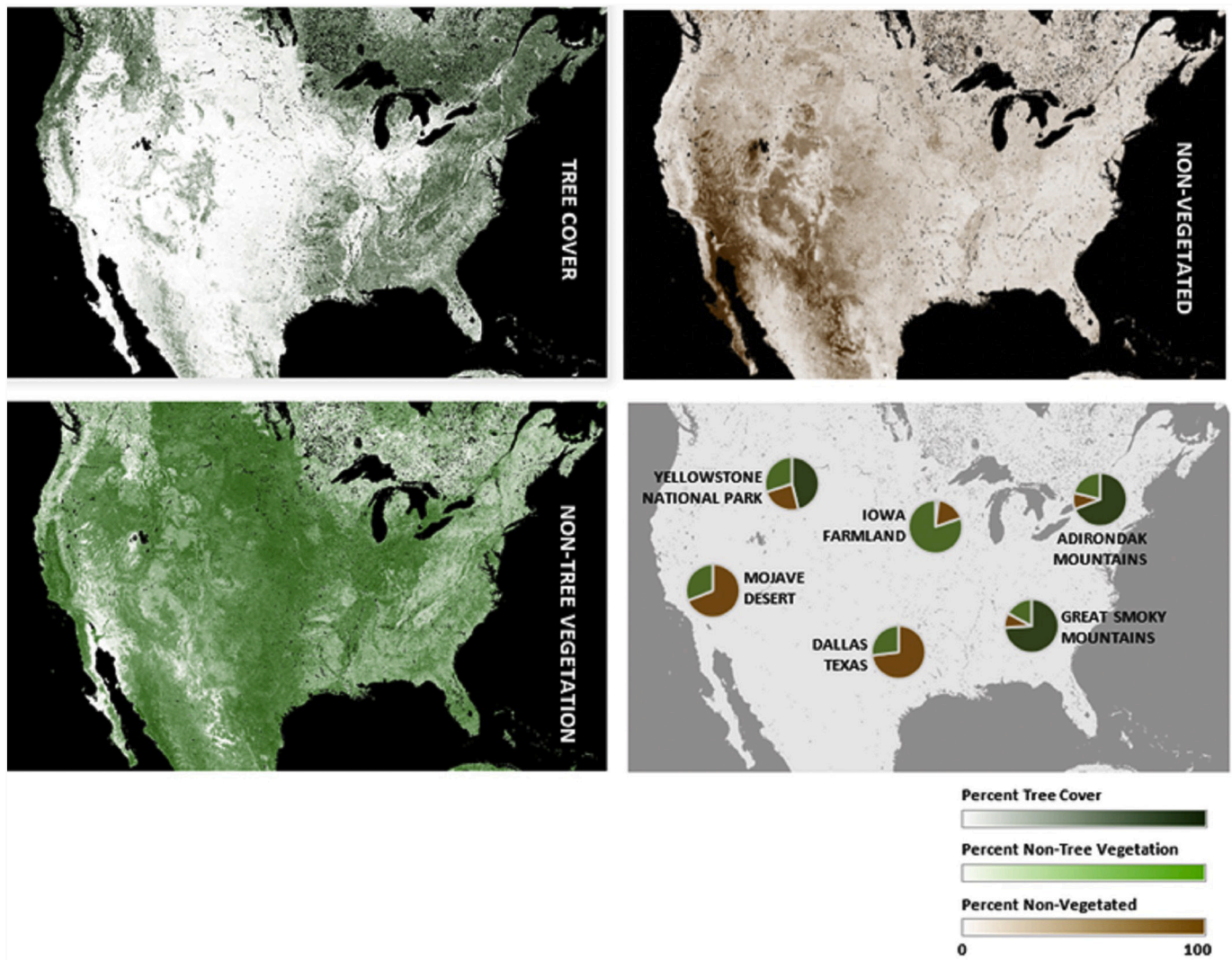


Fig. 3. MODIS VCF vegetation cover layers, US, 2010: percent tree cover (dark green), percent non-tree vegetation (bright green), and percent non-vegetated (brown). Lower right: proportions of each vegetation cover type for six sites. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

validation dataset to those MODIS pixels containing five or more high quality GEDI retrievals, then averaged the canopy cover samples falling into each MODIS pixel. This resulted in validation data for more than 1.5 million MODIS pixels. A linear regression comparing MODIS canopy cover values with those from GEDI gives an R^2 value of 0.53 (Fig. 4). The regression indicates a systematic difference, with GEDI's values on average somewhat higher at the higher end of the range and lower at the lower end of the range. Systematic difference in the higher values is at least partially due to GEDI retrievals over buildings which predict high canopy cover values, not distinguishing between trees and buildings. Variability about the regression line is also expected since GEDI footprints cover approximately 1% of each of the MODIS pixels in the validation dataset.

Other validation efforts include Carroll et al. (2017) who have carried out validation of the VCF products using field data from two sites in Maryland, and three sites in Brazil. As algorithms and upstream inputs have improved, the RMSE has dropped from 19.27 to 9.47 (in units of percent tree cover) for the Maryland sites and from 14.37 to 10.4 for the Brazil sites when compared to the field data.

Several other studies have compared MODIS VCF tree cover data to fractional cover values obtained from field data or from high resolution aerial and satellite data. Hansen et al. (2002) compared 250 m MODIS

tree cover results with two maps of western Zambia derived from aerial photos, field work, 'Landscape and grassland mapping of Western Province, Zambia' (Jeanes and Baars, 1991) and 'Vegetation Map of Zambia' (Edmonds, 1976). Land cover classes were correlated with continuous fields for this comparison; the resulting RMSE was 9.8% tree cover. Montesano et al. (2009) used Quickbird imagery to estimate crown cover across the circumpolar boreal region; 2005 VCF-Quickbird comparisons resulted in a RMSE of 13.4% tree cover. Sexton et al. (2013) compared MODIS VCF to FVC derived from four Lidar images in western North America, finding an RMSE of 16.8%.

Comparisons have also been made between VCF products and other land cover products. Song et al. (2014) conducted a comprehensive comparison of six land cover characterizations, including MODIS VCF tree cover, with a large collection of regional and country-level forest cover maps, supplemented by Landsat forest/no forest maps created with a clustering-labeling algorithm. All products were aggregated to 5 km before assessment. Resulting MODIS NDVI RMSE values were 21.73% tree cover for North America, 24.06% tree cover for South America, and 20.10% tree cover for Africa when compared to reference maps. When comparing land cover products, RMSE values may be higher than comparisons to field data because they incorporate errors from both products.

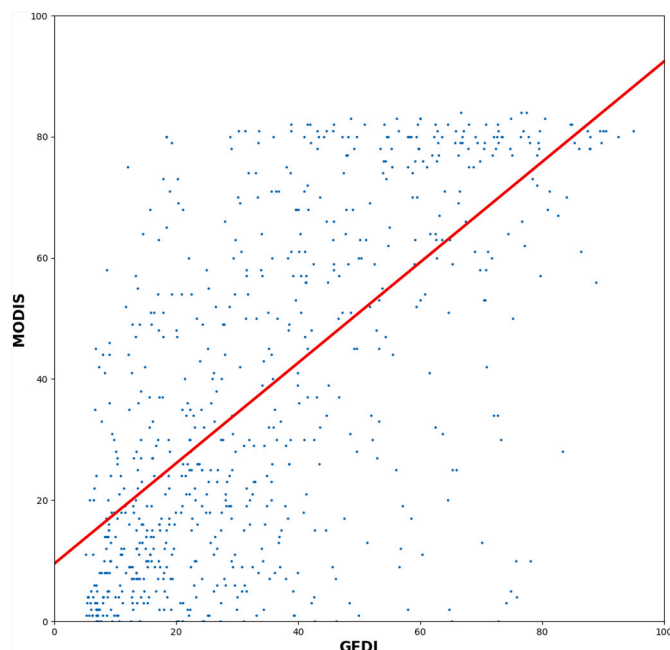


Fig. 4. Comparison of MODIS VCF percent tree cover with GEDI percent canopy cover. Red line is the linear regression line fit to the paired data points ($\text{MODIS_cover} = 12.501 + 0.682 * \text{GEDI_cover}$). Blue points show a subset of actual observations for purposes of visual clarity. The regression $R^2 = 0.53$; Mean absolute difference = 17.324; Root mean squared difference = 22.399; and Mean difference = 1.728. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Conclusions

The number of terabytes of satellite data available for land cover characterization continues to increase rapidly as data are gathered from new instruments at ever finer resolutions. These data have great potential to contribute new information about the Earth's surface, but to realize this potential, improved analytical methods are required, beyond the simple indices and image transforms of the past. As 'big data' have become ubiquitous, open source data mining tools have been created and are readily available resources to provide these new methods.

This paper demonstrates the success of a fully-automated data mining algorithm for production of global estimates of fractional vegetation cover. A detailed outline was given of the methods used to create the MODIS VCF products, including data compositing, metrics creation and selection, and model training; this description should provide valuable guidance to researchers desiring to use similar machine learning algorithms. Crucial aspects in building this data mining project were careful attention paid to data preparation and training data accuracy and the avoidance of model overfitting. In addition, reduction of dimensionality, in this case by removal of redundant variables, resulted in more accurate depictions, particularly as it allowed more training examples to be used without exceeding memory limitations.

Funding

The work was supported by NASA contracts and grants NAS596060, NNG04HZ16C, NNX08AH05A, NNX11AE77G, NNX14AJ33G, and 80NSSC18K0737.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgments

Many have contributed to this work in early years under the direction of John Townshend, especially Ruth DeFries, Matthew Hansen and Xiwu Zhan as investigators; Marc Steininger, Chengquan Huang, Eric Brown de Coulston and Jessica McCarty as a doctoral candidates; April Tompkins, Jenny Hewson-Scardaletti, Jill Eastman and Karen Schlee-weis as students; and James Tedrick, Bethany Semeiks and Peter Ma as technical staff. The precursor effort would not have been possible without the creation of the 8 km AVHRR Pathfinder under the direction of Martha Maiden.

References

- Breiman, L., 1996. Bagging predictors. *Mach. Learn.* 24 (2), 123–140.
- Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. *Classification and Regression Trees*. Wadsworth & Brooks.
- Calhoun, L.M., 2013. A quantitative assessment of the relationship between the malaria vector *Anopheles gambiae* sensu stricto and environment in Mali. PhD dissertation. University of California, Los Angeles.
- Carroll, M.L., Townshend, J.R.G., DiMiceli, C.M., Noojipady, P., Sohlberg, R.A., 2009. A new global raster water mask at 250 m resolution. *Int. J. Digit. Earth* 2, 291–308.
- Carroll, M.L., Townshend, J.R.G., Hansen, M.C., DiMiceli, C.M., Sohlberg, R.A., Wurster, K.W., 2011. Vegetative cover conversion and vegetation continuous fields. In: Ramachandran, B., Justice, C.O., Abrams, M. (Eds.), *Land Remote Sensing and Global Environmental Change: NASA's Earth Observing System and the Science of ASTER and MODIS*. Springer-Verlag, New York, pp. 725–745.
- Carroll, M.L., Townshend, J.R.G., Hansen, M.C., DiMiceli, C.M., Sohlberg, R.A., Huang, C., 2017. User guide for the MODIS Vegetation Continuous Fields Product, Collection 6, Version 1. https://lpdaac.usgs.gov/sites/default/files/public/product_documentation/mod44b_user_guide_v6.pdf (accessed 13 October, 2017).
- Cartus, O., Santoro, M., Kelldorfer, J., 2012. Mapping forest aboveground biomass in the northeastern United States with ALOS PALSAR dual-polarization L-band. *Proc. SPIE* 124, 466–478.
- Cecchi, G., Mattioli, R.C., Slingenberg, J., De La Rocque, S., 2008. Land cover and tsetse Fly distributions in sub-Saharan Africa. *Med. Vet. Entomol.* 22, 364–373.
- Coops, N.C., Wulder, M.A., Iwanicka, D., 2009. An environmental domain classification of Canada using earth observation data for biodiversity assessment. *Ecol. Inform.* 4, 8–22.
- DeFries, R.S., Field, C.B., Fung, I., Justice, C.O., Los, S., Matson, P.A., Matthews, E., et al., 1995. Mapping the land surface for global atmosphere-biosphere models: toward continuous distributions of vegetation's functional properties. *J. Geophys. Res.-Atmos.* 100, 20867–20882.
- DeFries, R.S., Townshend, J.R.G., Hansen, M.C., 1999. Continuous fields of vegetation characteristics at the global scale at 1-km resolution. *J. Geophys. Res.-Atmos.* 104, 16911–16923.
- DeFries, R.S., Hansen, M.C., Townshend, J.R.G., 2000. Global continuous fields of vegetation characteristics: a linear mixture model applied to multi-year 8 km AVHRR data. *Int. J. Remote Sens.* 21, 1389–1414.
- DeFries, R.S., Hansen, A., Newton, A.C., Hansen, M.C., 2005. Increasing isolation of protected areas in tropical forests over the past twenty years. *Ecol. Appl.* 15, 19–26.
- DiMiceli, C.M., Carroll, M.L., Sohlberg, R.A., Kim, D.H., Kelly, M., Townshend, J.R.G., 2015. MOD44B MODIS/Terra vegetation continuous fields yearly L3 global 250m SIN grid V006. In: NASA EOSDIS Land Processes DAAC. <https://doi.org/10.5067/MODIS/MOD44B.006>.
- Edmonds, A.C.R., 1976. *Vegetation Map of Zambia*. Forest Department/Survey Department, Lusaka.
- Giglio, L., Van der Werf, G.R., Randerson, J.T., Collatz, G.J., Kasibhatla, P., 2006. Global estimation of burned area using MODIS active fire observations. *Atmos. Chem. Phys.* 6, 957–974.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. *The WEKA Data Mining Software: An Update*. Assoc. for Computing Machinery SIG on Knowledge Discovery and Data Mining Explorations Newsletter, 11, pp. 10–18.
- Hansen, M.C., DeFries, R.S., Townshend, J.R.G., Sohlberg, R., 2000. Global land cover classification at 1 km spatial resolution using a classification tree approach. *Int. J. Remote Sens.* 21, 1331–1364.
- Hansen, M.C., DeFries, R.S., Townshend, J.R.G., Marufu, L., Sohlberg, R.A., 2002. Development of a MODIS tree cover validation data set for Western Province, Zambia. *Proc. SPIE* 83, 320–335.
- Hansen, M.C., DeFries, R.S., Townshend, J.R.G., Carroll, M.L., DiMiceli, C.M., Sohlberg, R.A., 2003. Global percent tree cover at a spatial resolution of 500 meters: first results of the MODIS vegetation continuous fields algorithm. *Earth Interact.* 7, 1–15.
- Hansen, M.C., Stehman, S.V., Potapov, P.V., Loveland, T.R., Townshend, J.R.G., DeFries, R.S., Pittman, K.W., et al., 2008. Humid tropical forest clearing from 2000 to 2005 quantified by using multitemporal and multiresolution remotely sensed data. *P. Nat. Acad. Sci. USA* 105, 9439–9444.
- Hansen, M.C., Potapov, P.V., Moore, R.T., Hancher, M.D., Turubanova, S.A., Tyukavina, A., Thau, D.M., et al., 2013. High-resolution global maps of 21st-century forest cover change. *Science* 342, 850–853.

- Hernández, A.J., Pastor, J., 2008. Relationship between plant biodiversity and heavy metal bioavailability in grasslands overlying an abandoned mine. *Environ. Geochem. Hlth.* 30, 127–133.
- Jeanes, K.W., Baars, R.M.T., 1991. Landscape and Grassland Mapping of Western Province, Zambia. Dept. of Agriculture, Western Province.
- Jung, M., Henkel, K., Herold, M., Churkina, G., 2006. Exploiting synergies of global land cover products for carbon cycle modeling. *Proc. SPIE* 101, 534–553.
- Kim, D.-H., Sexton, J.O., Noojipady, P., Huang, C., Anand, A., Channan, S., Feng, M., Townshend, J.R.G., 2014. Global, Landsat-based forest-cover change from 1990 to 2000. *Proc. SPIE* 155, 178–193.
- Kim, D.-H., Sexton, J.O., Townshend, J.R.G., 2015. Accelerated deforestation in the humid tropics from the 1990s to the 2000s. *Geophys. Res. Lett.* 42, 3495–3501.
- Lawrence, P.J., Chase, T.N., 2007. Representing a new MODIS consistent land surface in the Community Land Model (CLM 3.0). *J. Geophys. Res.-Biogeo.* 112 (G1).
- Lefsky, M.A., 2010. A global forest canopy height map from the moderate resolution imaging Spectroradiometer and the geoscience laser altimeter system. *Geophys. Res. Lett.* 37 (15).
- Montesano, P.M., Nelson, R., Sun, G., Margolis, H., Kerber, A., Ranson, K.J., 2009. MODIS tree cover validation for the circumpolar taiga–tundra transition zone. *Remote Sens. Environ.* 113 (10), 2130–2141.
- Neigh, C.S.R., Masek, J.G., Nickeson, J.E., 2013. High-resolution satellite data open for government research. *EOS Trans. Am. Geophys. Union* 94, 121–123.
- Nightingale, J.M., Phinn, S.R., Hill, M.J., 2011. Remote sensing for modeling biogeographic features and processes. In: Millington, A., Blumler, M., Schickhoff, U. (Eds.), *The SAGE Handbook of Biogeography*. Sage Publications Ltd., pp. 394–418.
- Nisbet, R., Elder, J., Miner, G., 2009. *Handbook of Statistical Analysis and Data Mining Applications*. Academic Press.
- Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., D'amico, J.A., et al., 2001. Terrestrial ecoregions of the world: a new map of life on earth. *BioScience* 51 (110), 933–938.
- Quinlan, J.R., 1992. Learning with continuous classes. In: Adams, A., Sterling, L. (Eds.), *AI '92: Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, Hobart, Tasmania, Australia, 16–18 November 1992. World Scientific, Singapore, pp. 343–348.
- Sexton, J.O., Song, X.-P., Feng, M., Noojipady, P., Anand, A., Huang, C., Kim, D.-H., et al., 2013. Global, 30-m resolution continuous fields of tree cover: Landsat-based rescaling of MODIS Vegetation Continuous Fields with Lidar-based estimates of error. *Int. J. Digit. Earth* 6, 427–448.
- Sexton, J.O., Noojipady, P., Song, X.-P., Feng, M., Song, D.-X., Kim, D.-H., Anand, A., et al., 2016. Conservation policy and the measurement of forests. *Nat. Clim. Chang.* 6, 192–196.
- Song, X.-P., Huang, C., Feng, M., Sexton, J.O., Channan, S., Townshend, J.R.G., 2014. Integrating global land cover products for improved forest cover characterization: an application in North America. *Int. J. Digit. Earth* 7, 709–724.
- Tang, H., Armston, J., 2019. Algorithm theoretical basis document for GEDI L2B footprint canopy cover and vertical profile metrics. In: NASA EOSDIS Land Processes DAAC. https://lpdaac.usgs.gov/documents/588/GEDI_FCCVPM_ATBD_V1.0.pdf.
- Tian, Y., Dickinson, R.E., Zhou, L., Shaikh, M., 2004. Impact of new land boundary conditions from moderate resolution imaging Spectroradiometer (MODIS) data on the climatology of land surface variables. *J. Geophys. Res.-Atmos.* 109 (D20).
- Vermote, E.F., 2019. MOD09A1 v006. Improvements/Changes from Previous Versions. <https://lpdaac.usgs.gov/products/mod09a1v006> (accessed 25 May 2019).
- Vermote, E.F., Kotchenova, S., 2008. Atmospheric correction for the monitoring of land surfaces. *J. Geophys. Res.-Atmos.* 113 (D23).
- Vermote, E.F., Wolfe, R.E., 2015a. MOD09GQ MODIS/Terra surface reflectance daily L2G global 250m SIN grid v006. In: NASA EOSDIS Land Processes DAAC. <https://doi.org/10.5067/MODIS/MOD09GQ.006>.
- Vermote, E.F., Wolfe, R.E., 2015b. MOD09GA MODIS/Terra surface reflectance daily L2G global 1 km and 500m SIN grid V006. In: NASA EOSDIS Land Processes DAAC. <https://doi.org/10.5067/MODIS/MOD09GA.006>.
- Villagra, P.E., Defossé, G.E., Del Valle, H.F., Tabeni, S., Rostagno, M., Cesca, E., Abraham, E., 2009. Land use and disturbance effects on the dynamics of natural ecosystems of the Monte Desert: implications for their management. *J. Arid Environ.* 73, 202–211.
- Wang, Y., Witten, I.H., 1997. Inducing model trees for continuous classes. In: van Someren, M., Widmer, G. (Eds.), *Poster Papers of the Proceedings of the 9th European Conference on Machine Learning*. Czech Republic, Prague, pp. 128–137.
- Wiedinmyer, C., Quayle, B., Geron, C., Belote, A., McKenzie, D., Zhang, X., O'Neill, S., Wynne, K.K., 2006. Estimating emissions from fires in North America for air quality modeling. *Atmos. Environ.* 40, 3419–3432.
- Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. *Data Mining: Practical Machine Learning Tools and Techniques*, fourth ed. Morgan Kaufmann, Cambridge.