

Open land-use map: a regional land-use mapping strategy for incorporating OpenStreetMap with earth observations

Di Yang, Chiung-Shiuan Fu, Audrey C. Smith & Qiang Yu

To cite this article: Di Yang, Chiung-Shiuan Fu, Audrey C. Smith & Qiang Yu (2017) Open land-use map: a regional land-use mapping strategy for incorporating OpenStreetMap with earth observations, *Geo-spatial Information Science*, 20:3, 269-281, DOI: [10.1080/10095020.2017.1371385](https://doi.org/10.1080/10095020.2017.1371385)

To link to this article: <https://doi.org/10.1080/10095020.2017.1371385>



© 2017 Wuhan University. Published by
Taylor & Francis Group



Published online: 11 Sep 2017.



Submit your article to this journal



Article views: 1134



View Crossmark data



Citing articles: 1

Open land-use map: a regional land-use mapping strategy for incorporating OpenStreetMap with earth observations

Di Yang^{a,b} , Chiung-Shiuan Fu^{a,b} , Audrey C. Smith^a and Qiang Yu^c

^aDepartment of Geography, University of Florida, Gainesville, FL, USA; ^bLand Use and Environmental Change Institute, University of Florida, Gainesville, FL, USA; ^cBeijing Key Laboratory of Precision Forestry, Beijing Forestry University, Beijing, China

ABSTRACT

A land-use map at the regional scale is a heavy computation task yet is critical to most landowners, researchers, and decision-makers, enabling them to make informed decisions for varying objectives. There are two major difficulties in generating land classification maps at the regional scale: the necessity of large data-sets of training points and the expensive computation cost in terms of both money and time. Volunteered Geographic Information opens a new era in mapping and visualizing the physical world by providing an open-access database valuable georeferenced information collected by volunteer citizens. As one of the most well-known VGI initiatives, OpenStreetMap (OSM), contributes not only to road network distribution information but also to the potential for using these data to justify and delineate land patterns. Whereas, most large-scale mapping approaches – including regional and national scales – confuse “land cover” and “land-use”, or build up the land-use database based on modeled land cover data-sets, in this study, we clearly distinguished and differentiated land-use from land cover. By focusing on our prime objective of mapping land-use and management practices, a robust regional land-use mapping approach was developed by integrating OSM data with the earth observation remote sensing imagery. Our novel approach incorporates a vital temporal component to large-scale land-use mapping while effectively eliminating the typically burdensome computation and time/money demands of such work. Furthermore, our novel approach in regional scale land-use mapping produced robust results in our study area: the overall internal accuracy of the classifier was 95.2% and the external accuracy of the classifier was measured at 74.8%.

ARTICLE HISTORY

Received 9 June 2017
Accepted 14 August 2017

KEY WORDS

OpenStreetMap (OSM);
Volunteered Geographic
Information (VGI);
crowdsourced data; Random
Forest; Google Earth Engine
(GEE)

1. Introduction

1.1. Land-use mapping at regional scale

Land-use maps derived from remote sensing imagery play a vital role in monitoring human–environmental interactions such as landscape changes, ecological services (conservation), and urban planning and management (Lambin et al. 2001; Agarwal et al. 2002; Weng 2002; Abdullah and Nakagoshi 2007; Sumarga and Hein 2014; Hegazy and Kaloop 2015). Whereas, land cover maps represent the observed biophysical cover of the Earth’s surface, land-use maps describe the arrangements, activities, and inputs people undertake within a particular land cover type to produce, modify or maintain it. These specified and precise aspects of land-use patterns are indicative of the challenge in establishing distinctive use attributes and accurately mapping them. Ground-validated data are therefore essential to verify remotely sensed data used to infer land-use characteristics. In practice, there is a gap between remote sensing earth observations and their translation into subjective mapping products depicting how the land is used and

affected by human activity. Remotely sensed data-sets can be used to monitor land cover dynamics, but are insufficient on their own for deriving land-use characteristics – the manner in which people utilize, and thus modify the land on the ground. The latter is the concept of “land-use”, which is distinct from, and often confused with, the term “land cover”. Land-use reflects and results from anthropogenic activities on the Earth’s surface. The distinction between land cover and land-use poses a challenge and elicits a strategy, for mapping and monitoring the landscape changes and processes.

In modern history, ecosystems are strongly affected by anthropogenic factors (Folke, Holling, and Perrings 1996; Vitousek et al. 1997), including agriculture, building construction and urban expansion, forest timber extraction, and preservation systems (national and state parks) (Matson et al. 1997; Hartley 2002). Macrosystems of forests, croplands, and waterways are driven by human need and associated activities (Foley et al. 2005). The interactions between and the effects of different land-use types may be forces of landscape-wide importance. For example, in the Southeastern United States, significant

expansions of urban areas may convert forested land to urban uses, and pine plantation (forestry) may come from cropland (Wear and Greis 2002, 2013). These kinds of landscape conversions could represent macrosystem changes, depending on scale and extent, and can have immediate and local implications for landowners and management practices.

Although land-use mapping techniques are well studied and have been applied to mapping most developed areas from different perspectives and at local scales (Lambin et al. 2001; Bryan, Barry, and Marvanek 2009; Bateman et al. 2013; Lawler et al. 2014), it is challenging to map heterogeneous land-use at a regional scale with high spatial resolution. At the macrosystems scale, it becomes more difficult to map land-use due to the lack of training samples and validation points, and the necessity of heavy computing tasks. However, despite these obstacles, successful regional scale land-use mapping projects have been accomplished. The Australian Collaborative Land-use Mapping Program mapped land-use in the 17.3 million km² of Queensland using a Markov chain Monte Carlo machine learning technique (Lesslie, Barson, and Smith 2006; Bryan, Barry, and Marvanek 2009). The integration of remote sensing data with Volunteered Geographic Information (VGI) data platforms such as Open Street Map (OSM) and cloud computing platforms such as Google Earth Engine (GEE), provides a significant resource for land cover and land-use mapping and related research. Moreover, the combining of remote sensing data and OSM constitutes a powerful tool to monitor, characterize, and quantify the landscape. The product will be an important source of information for researchers and policy-makers for investigating land parcels.

Supervised classification has been proven to be an efficient tool for mapping land patterns and land changes accompanying the spatial and temporal configuration of landscape heterogeneity (Congalton 1991). However, it becomes ineffective/inapplicable at regional scale and for fine resolution mapping because the training data (i.e. ground-validated data) collection process is costly and time-consuming, and is difficult to automate. The processing and collecting time of classification *in situ* training sites significantly slows down the use of such methods in the applications of mapping regional land patterns.

1.2. Volunteered Geographic Information as a source of training data

VGI marks a new era in mapping and visualizing our world, and its data and applications have grown significantly during the last decade (Elwood 2008; Haklay, Singleton, and Parker 2008; Zook et al. 2010). OSM is one of the well-supported VGIs and has been studied and applied in multiple disciplines, but there remain large quantities of information that need to be investigated.

OSM has unique advantages, allowing its use to be fitting a wider range of applications as compared to the official geographic databases. OSM has the capability to create superior maps when considering temporal change trajectories by providing “up-to-date” data (Estima and Painho 2013). From a geographic perspective, this is significant because scaling has always been a major issue in the mapping and monitoring of the Earth. For the research undertaken in this study, OSM is an attractive choice to achieve the research objectives because of the growing coverage of big spatial data and cloud computing capabilities (Zook et al. 2010).

OSM contains not only road network data, but also land-use information, which might be derived from a combination of other ground-level features. Previous studies have shown that OSM is a valuable and structured data source for mapping land cover and land-use by providing high positional accuracy in comparison with the corresponding commercial data-sets (Estima and Painho 2013, 2015; Jokar et al. 2013; See et al. 2013, 2015; Johnson and Iizuka 2016). For improving the value and potential of collecting and mapping, OSM provides a VGI platform, to which many volunteers can contribute and collaborate. Moreover, the achieved OSM data-sets enable the study of past land changes based on multi-temporal trajectories (Neis, Zielstra, and Zipf 2012).

Recent studies have investigated the spatial distribution balance of OSM-trained classification maps. A total 76% agreement was found in a study based in Portugal, and 64% and 77% agreement was found in two cities in Germany as compared with the corresponding official land-use/cover databases (Estima and Painho 2013; Jokar, Helbich, et al. 2015; Jokar, Mooney, et al. 2015). However, most OSM land cover and land-use mapping applications have focused on urban areas at the local scale, and without using earth observation databases for cross-referencing. Moreover, to produce more accurate and reliable mapping of land-use patterns, the temporal features of the land should also be considered. There was no study that considered and analyzed the temporal land changes, which are critical for reflecting human activities. Different land-use types have distinctive time series signals that can be derived from earth observation images. For example, the land-use type of residential areas typically is not characterized by rapid or seasonal changes, whereas cropland and managed forestlands land-use types have a high temporal variation which can be observed in their respective time series signals.

1.3. Incorporating OpenStreetMap with earth observations

Remote sensing techniques have been widely used for observing and monitoring landscape changes, terrestrial features, and can extensively decrease the time and cost for large-scale mapping. Moreover, the Web 2.0 brings a revolution of how people collect, map, and analyze

geo-located data (Haklay, Singleton, and Parker 2008). For earth observation data-sets, GEE is a newly developed mapping and analysis platform which enables large-scale spatial analysis by its special infrastructure and automatically parallelized computation techniques (Johansen, Phinn, and Taylor 2015; Padarian, Minasny, and McBratney 2015; Patel et al. 2015). By storing and bringing together large amounts of earth observation data on Earth Engine's server, GEE can analyze big data "on-the-fly" (Yu and Gong 2012). Personal databases can also be uploaded to the server as additional assets to achieve time series generating, zonal statistics, spectral analysis, and many machine learning classification techniques.

GEE has been studied for its strong ability for mapping landscape at the regional scale in both time and space. Yang generated a Mayan forest land cover map by building OSM-derived training samples into a high-dimensional Landsat-based Random Forest classifier of the whole Yucatan Peninsula, and then delineated land patterns with road networks (Yang 2017). GEE was also applied to monitoring land changes at the regional scale based on Landsat 30-m spatial resolution (Soulard et al. 2016; Azzari and Lobell 2017). The major objective of this research is to incorporate OSM into regional land-use mapping using GEE to investigate and quantify land-use types in the Southeastern United States. The broader aim of this study is to develop an integrative and innovative regional scale mapping strategy that can be replicated for similar research purposes. Overall, the methodology presented here seeks to improve land-use mapping strategies by reducing economic and temporal burdens, and contribute to the science of monitoring human–environment interactions.

2. Study area

This study focuses on the land-use patterns of the heterogeneous Southeastern United States. The Southeastern United States has high landscape heterogeneity, with heavily managed forestlands, highly developed agriculture lands, and multiple metropolitan areas. Human activities are transforming and altering land patterns and structures in both negative and positive manners. Based on EPA ecoregion descriptions (Omernik and Griffith 2014), land cover in the Southeastern United States is a mosaic of cropland, pasture, woodland and forests, and wetlands. For this research, the study area consists of one Landsat scene of Worldwide Reference System II (WRS-2) Row 17 Path 39 (Figure 1). This Landsat scene is an illustrative example of land patterns of the Southeastern United States, which covers an area of approximately 185 km². The heterogeneous landscape in the study area consists of a mixture of highly developed cropland, plantation forests, multiple metropolitan areas, and industrial and commercial lands. Human land-use activities and management practices are transforming landscape

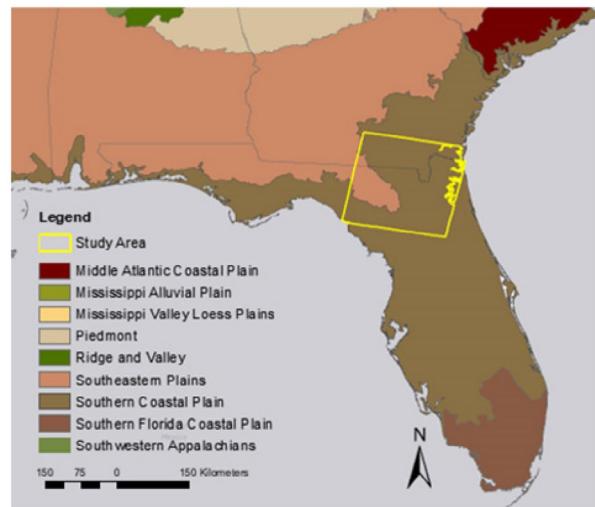


Figure 1. Study area in southern Georgia and north Florida delineated by the Landsat tile of WRS-II Path 17 Row 39. Underlying map shows the EPA level III Ecoregions (EPA). This map was created using ArcGIS® software by Esri. ArcGIS® and ArcMap™ are the intellectual property of Esri and are used herein under license. Copyright © Esri. All rights reserved. For more information about Esri® software, please visit <http://www.esri.com>

patterns and processes in the region, resulting in both negative and positive consequences for stakeholders and ecosystems.

3. Methodology

This research proposes a methodology to produce a regional scale land-use map by extracting training data from OSM. Random Forest is chosen as the classifier because it has become a widely used algorithm for remote sensing image classification (Breiman 2001). Random Forest has the ability to handle high-dimensional raster databases (Senf, Hostert, and van der Linden 2012), which suits the framework of this research.

3.1. Data sources

3.1.1. OpenStreetMap data structure

OSM is a fully free and openly accessible map of road network data. All OSM data were downloaded from the Geofabrik website <http://download.geofabrik.de>. Community volunteers collect and submit geographic information to the global OSM database (Ciepluch et al. 2009). OSM data quality has been broadly assessed with consistently positive reviews. Girres and Touya (2010) performed a spatial analysis of the quality of OSM street network representations in the UK and France, respectively, through a comparison to ground-truth data obtained from the corresponding national mapping agency. Both case studies found that on average, the quality of the data was reasonably good but exhibited a significant spatial heterogeneity. Neis, Zielstra, and Zipf (2012) analyzed how the quality of the OSM street

network in different regions of Germany had changed between the years 2007–2011. This multi-temporal study demonstrated that the quality and accuracy of OSM data have improved over time, and more recent studies confirm this assessment (Neis, Zielstra, and Zipf 2012).

The OSM database contains a full coverage of the area of interest for this research, with the following sub-data-sets: OSM-Places, OSM-Railways, OSM-Roads, OSM-Waterways, OSM-Buildings, OSM-Landuse, and OSM-Nature. In this study, the OSM feature classes applied for mapping land-use patterns include OSM-Places, OSM-Landuse, and OSM-Nature.

The OSM-Landuse feature class is used to describe the human use of land, presented as a polygon class consisting of forests, residential areas, and some industrial areas. Figure 2 shows the composition of OSM-Landuse labels. In the study area of Landsat scene Path 17 Row 29, there are 26 different types of land-use labels on land surface, which illustrate how human activities affect landscapes. Labels of “residential” contribute 43% (675 sites), followed by “commercial” (14% – 214 sites) and “farm” (11% – 175 sites). OSM-Nature is used to describe natural physical land features, including ones that have been modified by humans, which covers most waterbody and conservation areas. In OSM-Nature, all the waterbodies, parks, and recreation areas are represented as polygons. As shown in Figure 3, labels of “water” contribute 75% (2259 sites), follows by “forest” (17% – 503 sites) and “park” (8% – 228 sites). The OSM-Places feature class includes urban and suburb landmarks and attractions, with the following place labels for the study area shown in Figure 4: labels of “helmet” contribute 66% (585 sites); followed by “island” (24% – 211 sites;) and “village” (3% – 30 sites). Based on the general statistics of the data structure of OSM labels, it was determined that OSM-Nature, OSM-Places, and OSM-Landuse represent a

variety of human-made land-use types, and the quantity of data points is large enough to suffice for the necessary inputs for large-scale classification mapping.

The OSM-Landuse, OSM-Nature, and OSM-Places were reclassified and converted into five major land-use types: “residential”, “forestry”, “cropland”, “commercial/industrial”, and “waterbody”. Although it is noted that the “waterbody” class may not be a land-use type, it is included as it accounts for a significant proportion of Florida land cover. Furthermore, the inclusion of the “waterbody” class in the land-use map serves to avoid the classification bias. Users may also decide to apply different classification frameworks based on the specific research questions and understanding of the study area attributes.

3.1.2. Earth observations

Earth observation is one of the most essential tools for monitoring the earth’s surface and its dynamics at regional to global scales. Although earth observation data and remote sensing techniques allow for extensive mapping of characteristics of the land surface, it can be financially burdensome and time-consuming for mapping at large scale in both time and space. GEE provides online access to worldwide coverage of a vast of remote sensing data-sets. The GEE API is currently available upon request to a number of groups for testing and applying large-scale mapping on the cloud server.

The spatial and temporal covariates used in the land-use classification for this research include multi-temporal land reflectance, forest canopy height, DEM, mean-EVI, and the land ownership. The spatial and temporal covariates are important indicators of land-use, ecosystems, and the remotely sensed inputs are consistently and accurately derived. All the data used for the land-use map produced in this research were up-to-date, subset to the boundaries of the study area, and resampled to 30-m spatial resolution. Additionally, all the data used as spatial and temporal covariates cover the entire study area and can be extended to the global scale, thereby making the research methods reproducible for other regions of interest, given there are similar research objectives. The earth observation images used here consist of a singular time step Landsat images and multi-temporal composite satellite images. The temporal heterogeneity of the data-sets combines to reflect the characteristics of the landscape of the study area. Furthermore, the heterogeneous data allows learning a supervised algorithmic model (ie Random Forest) in order to extract relevant thematic classes of land-use/land cover from the satellite imagery. Classification is performed on a data-set consisting of the following five major sources (Table 1).

- (1) Landsat 8 Top-of-Atmosphere (TOA) (Chander, Markham, and Helder 2009) of 2016: this composite image reflects the greenest pixel during the whole period based on NDVI.

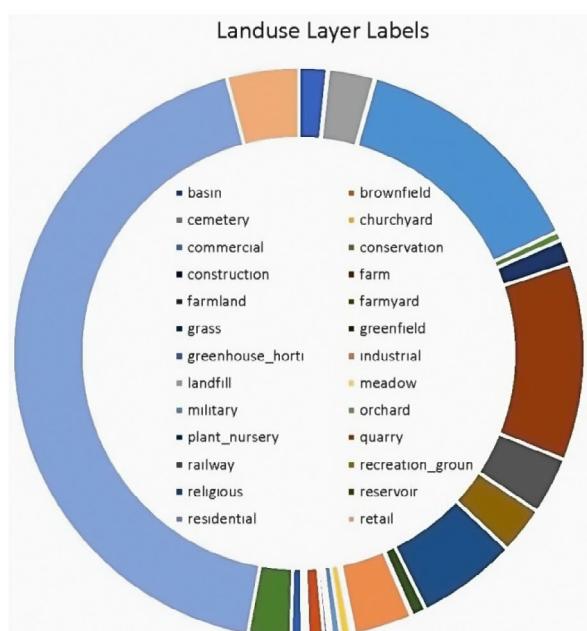
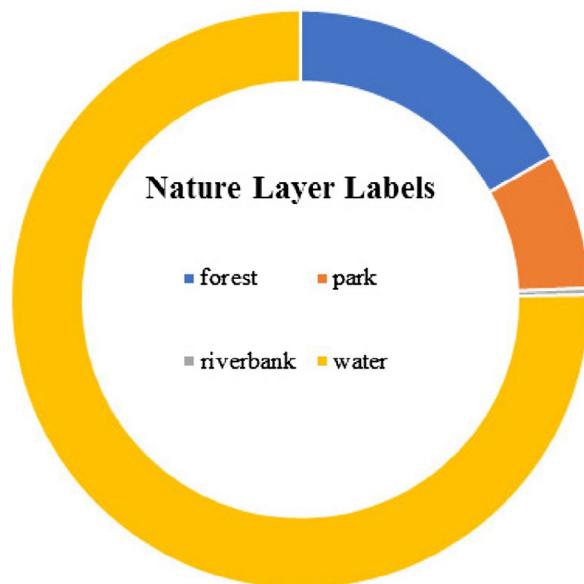
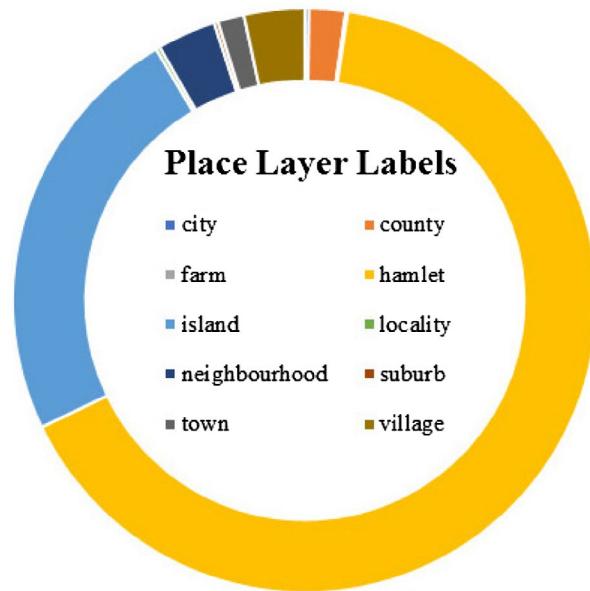


Figure 2. Breakdown of OSM-Landuse labels of study area.

Table 1. Spatial and temporal input covariates for Random Forest classifier.

Database name	Objective	Google Earth Engine ID	Spatial resolution	Temporal resolution
Landsat 8 TOP reflectance	Reflectance	LANDSAT/LC8_L1T_ANNUAL_GREENEST_TOA	30 m	16-day
Forest canopy height	Canopy height	NASA/JPL/global_forest_canopy_height_2005	1 km	Composite
MODIS 13 EVI	EVI	MODIS/MOD13Q1	250 m	16-day
SRTM (Shuttle radar topography mission)	DEM	CGIAR/SRTM90_V4	90 m	Composite
SEUS land ownership type	Ownership	users/yangdi1031/se_ownership_masked	250 m	Composite

**Figure 3.** Breakdown of OSM-Nature labels of study area.**Figure 4.** Breakdown of OSM-Place labels of study area.

(GEE ID: LANDSAT/LC8_L1T_ANNUAL_GREENEST_TOA). This is a multi-temporal data-set consisting of Landsat scenes with cloud removal by Fmask algorithm (Zhu and Woodcock 2012).

- (2) Global forest canopy height version of 2005 (Simard et al. 2011): this data-set represents the canopy height on global level by incorporating the Geoscience Laser Altimeter System (GLAS) and ancillary data. (GEE ID: NASA/JPL/global_forest_canopy_height_2005).
- (3) MOD13_EVI (Huete et al. 2002): the MODIS 13 EVI product was designed to minimize canopy background variations and maintain the sensitivity over dense vegetation conditions at a spatial resolution of 250 m every 16 days. (GEE ID: MODIS/MOD13Q1).
- (4) Shuttle Radar Topography Mission (SRTM): the DEM data from SRTM are produced to provide consistent, high-quality data with a resolution of 90 m (Jarvis et al. 2008). (GEE ID: CGIAR/SRTM90_V4).
- (5) Land Ownership Types: geospatial land ownership data sources from federal and

nongovernment agencies are integrated for landownership type mapping. Landowners are classified as public and private. There are six sub-types of public ownership, which are federally protected, federal, state-protected, state, military, and local. In addition, there are four sub-types of private ownership: non-government organization, private, family, and corporate. The ownership classification strategy is made, based on different management objectives, as well as landowner skills, budgets, and interests. The U.S. Protected Areas (PADUS) is the primary data source to identify public ownership.

To summarize, Figure 5 shows the examples of covariates used to generate land-use maps: including Shuttle Radar Topography Mission (SRTM) digital elevation at 30-m resolution (Figure 5(a)), MODIS13 Mean EVI of 2015 at 250-m resolution (Figure 5(b)), Forest canopy height at a resolution of 1 km (Figure 5(c)), and Landsat 8 calibrated Top-of-Atmosphere (TOA) with Fmask quality band to remove the cloud contaminations (Figure 5(d)).

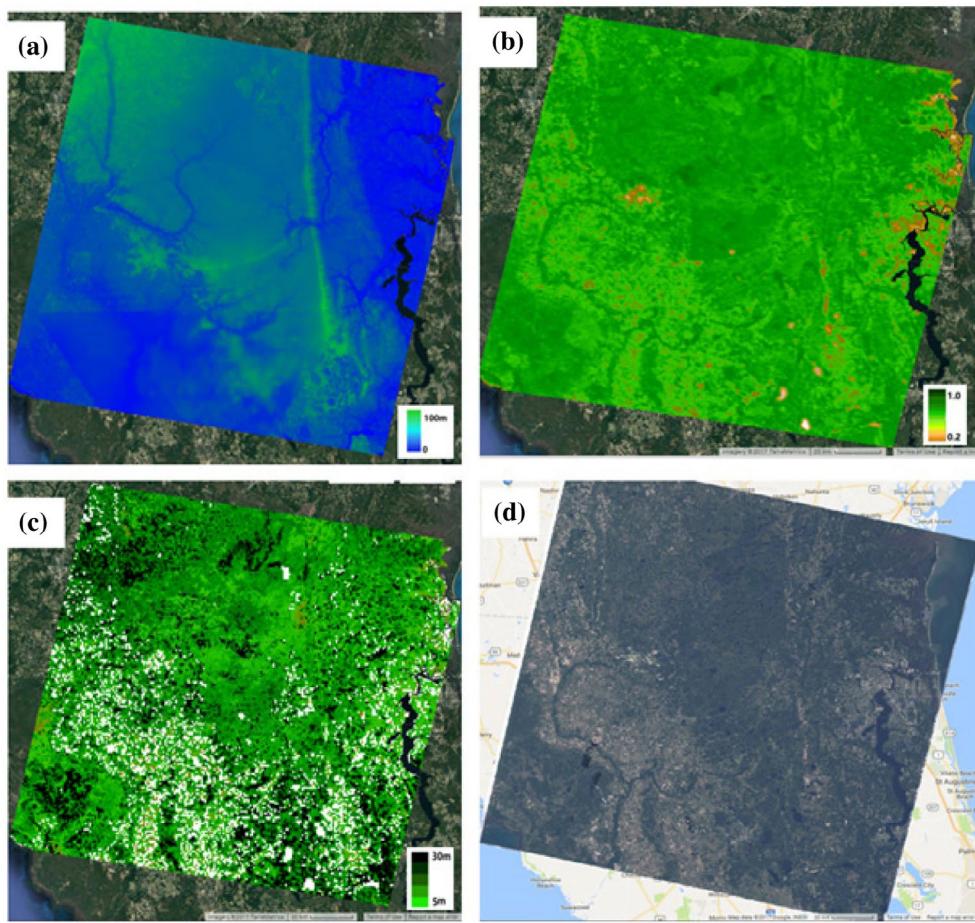


Figure 5. Examples of covariates used to generate land-use maps: (a) Shuttle radar Topography Mission (SRTM) digital elevation at 30-m resolution; (b) MODIS13 Mean EVI of 2015 at 250-m resolution; (c) Forest canopy height at 1-km resolution; (d) Landsat 8 calibrated Top-of-Atmosphere (TOA) with Fmask quality band to remove the cloud contaminations. Background image and map courtesy of Google Map (2017). “Florida, U.S.” Map. Google Maps. Google, Available from: <https://www.google.com/maps/@30.1671351,-82.8317191,8.92z?hl=en> [Accessed March 24, 2017].

3.2. Mapping strategy

3.2.1. Training samples setup

The randomly selected training points were converted into a Keyhole Markup Language file. Each chosen point centered at a 30×30 m polygon, one-pixel image size. The chosen points and associated polygons were overlaid in Google Earth virtual globe for visualization. The selected points were identified and assigned classes based on the OSM feature types and their surrounding areas. The points were then split into training and testing data-sets for model validation. To avoid the imbalance caused by polygon weight within the training sample (Visconti et al. 2013), OSM-Nature and OSM-Landuse polygons feature classes were converted into points features based on their respective geographic centroids. In Figure 6, the spatial distribution of the training samples of each feature including OSM-Nature, OSM-Landuse, OSM-Places, and Google Earth visualization samples are shown, and Figure 7 shows the visualization of all

training samples. There are total number of 3068 points from OSM and 380 from Google Earth visualization. It is found that the spatial distribution of OSM-Landuse and OSM-Nature clustered in a great degree, which might cause bias for the classifier (see Figures 6(a) and (b)). We therefore added the layer of OSM-Places.

Another challenge arose from the imbalanced class distribution of different extracted features from OSM. Thus, a layer of managed forest from Google Earth visual interpretation was added. Google Earth was used because it provides higher spatial resolution with a time span of more than 20 years. Therefore, more spatial information can be acquired about what types of land practices occur during the spatio-temporal validation of this study. Further details about the training sample setting strategy is depicted in Figure 8, which shows the workflow reflecting the conversion of OSM features to land-use labels. In this study, for our study area, there are 3068 training samples collected from OSM features, and 380 from visual interpretation from Google Earth over

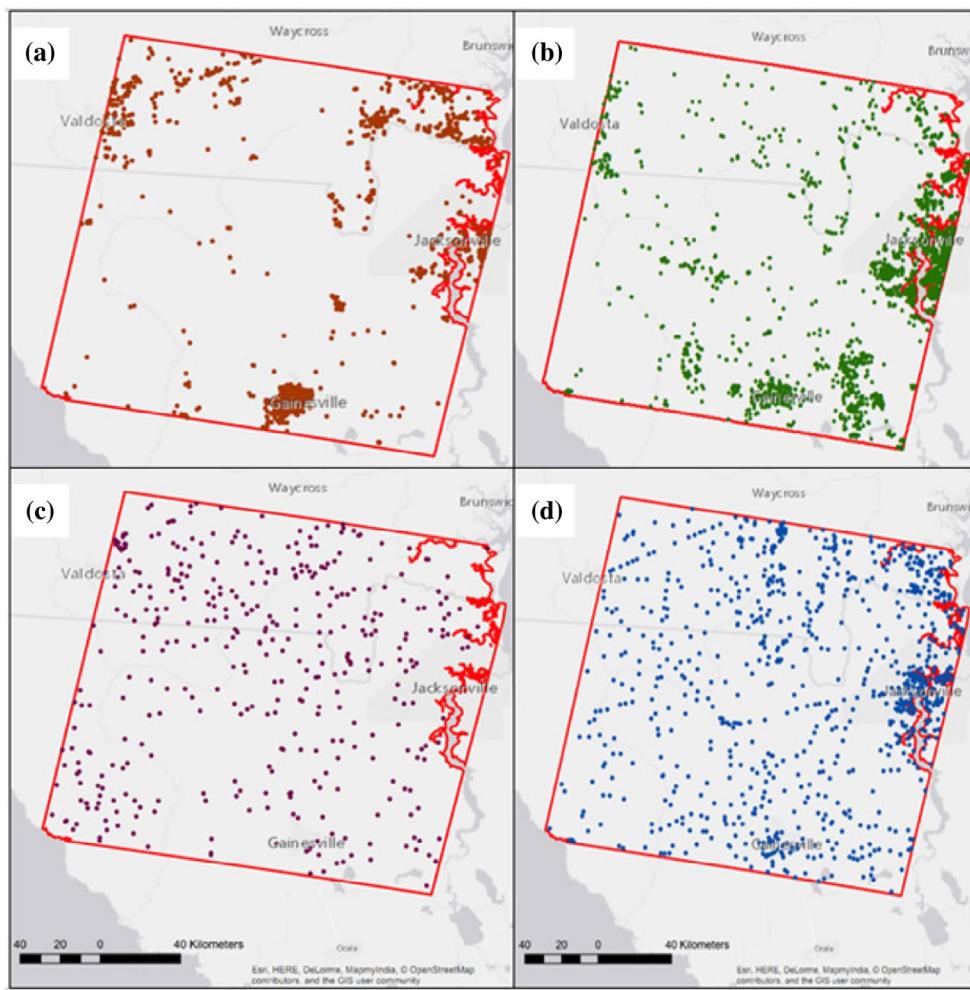


Figure 6. Spatial distribution of RF training from OSM and Google Earth Engine: (a) centroid center of OSM-Landuse feature; (b) centroid center of OSM-Nature feature; (c) managed forest sites from Google Earth visualization; (d) centroid center of OSM-Place feature. The base map was developed by Esri using HERE data, DeLorme basemap layers, OpenStreetMap contributors, Esri basemap data, and select data from the GIS user community. In North America coverage is provided from Level 14 (1:36k scale) through Level 16 (1:9k scale). http://goto.arcgisonline.com/maps/World_Light_Gray_Base. This map was created using ArcGIS® software by Esri. ArcGIS® and ArcMap™ are the intellectual property of Esri and are used herein under license. Copyright © Esri. All rights reserved. For more information about Esri® software, please visit <http://www.esri.com>

time (3448 in total). The training samples extracted from OSM were re-classified into five classes: (1) residential, (2) managed forest, (3) cropland, (4) commercial/industrial, and (5) waterbody. The training sample set of this study is visualized in Figure 7, which represents the spatial distribution of the training points geographically and illustrates the degree of urbanization.

3.2.2. Supervised classification

The classifier of Random Forest is considered for classification of multi-source remote sensing data mapping in this study because it was proved to perform relatively well for the integration of imagery data and collaborate the environmental covariates into classifier to improve its accuracy in such a complex landscape (Rodríguez-Galiano et al. 2012). For classifying land-use at the regional scale, Random Forest has several advantages over the others and the most notable ones (Breiman 2001) are:

- To implement advanced pixel-based supervised classifier with high accuracy.
- To estimate and measure the variable importance, and
- To compare out-of-bag error to independent validation to avoid classification bias.

All data used in this study as inputs for the classifier are completely embedded in GEE. In other words, the GEE IDs of each database used in this study are accessible by via GEE platform, and can be used for to reproduce the methodology presented here and for other study regions. Moreover, the use of GEE enables regional land cover mapping to be accomplished with greater ease than other platforms, as many spatial and temporal covariates are already available in the GEE server and the calculation itself is parallelized automatically to optimize the computing time cost. In addition, the spatial resolution can be resampled easily to match or fit

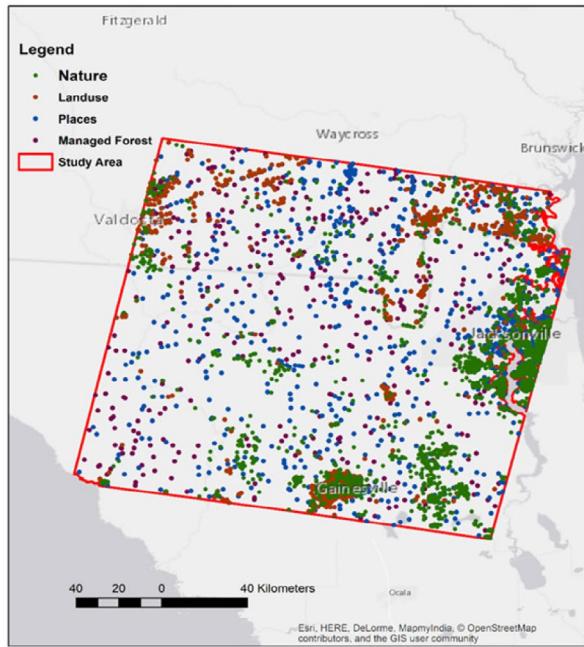


Figure 7. Spatial distribution of training sites over study area. The base map was developed by Esri using HERE data, DeLorme basemap layers, OpenStreetMap contributors, Esri basemap data, and select data from the GIS user community. In North America coverage is provided from Level 14 (1: 36k scale) through Level 16 (1:9k scale). http://goto.arcgisonline.com/maps/World_Light_Gray_Base. This map was created using ArcGIS® software by Esri. ArcGIS® and ArcMap™ are the intellectual property of Esri and are used herein under license. Copyright © Esri. All rights reserved. For more information about Esri® software, <http://www.esri.com>

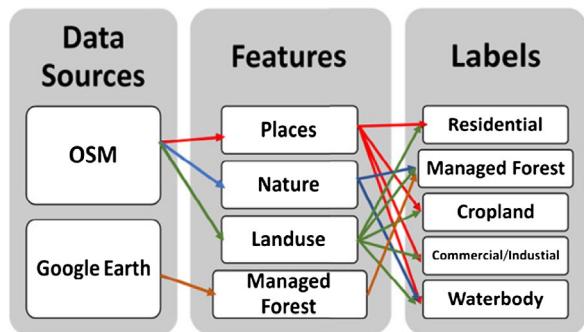


Figure 8. Workflow of training sample setting.

with other databases. To convert the uploaded training samples from OSM to GEE, a fusion table was generated representing the geometry and class types of the training samples extracted from OSM. Therefore, the OSM training sample used for this study are embedded in GEE for use by other users. The four-step procedure for mapping land-use is summarized below:

- (1) Explore the data-sets of OSM-Nature, OSM-Landuse, and OSM-Places, set the land-use classification strategy based on research objective, and reclassify and generate the training sample data-set.

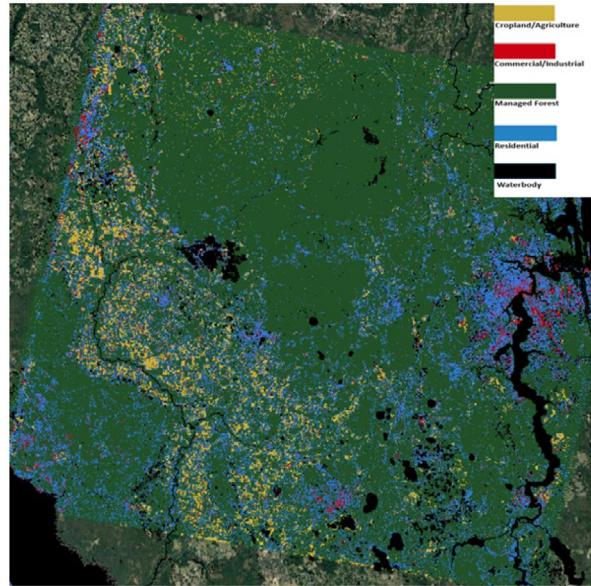


Figure 9. The dominant land-use types of study area derived from earth observation data-sets and self-automated training samples from OSM and Google Earth. Image is courtesy of Google Earth® (2017). Google Earth V 7.1.1.1888 (viewed March 12, 2017). Florida, USA. 30°05'46.33"N, 82°14'14.11"W, elev. -11 ft, Eye alt 125 mi. SIO, NOAA, U.S. Navy, NGA, GEBCO. Landsat/Copernicus [Accessed December 30 2014]. <https://www.google.com/earth/index.html>

- (2) Validate with Google Earth on both time and space for the training data-set.
- (3) Collaborate the spatial and temporal covariates as inputs for Random Forest classifier, run Random Forest classifier using GEE API, and analyze the classification accuracy for each classification type.
- (4) Extract the forest land from the resulting map at Step (3), superimpose rasterized OSM road networks, and run cluster analysis to generate land-use map.

4. Results and discussion

4.1. Land-use classification results

The results of the Random Forest classifier show land mosaics that describe the proportion of the study area that were classified into each land-use type. The landscape of the Landsat scene Path 17 Row 39, consisting of 185 km², is comprised of 62% forestland (2.41×10^6 ha), followed by 16% cropland, 3% commercial/industrial lands, and 8% residential areas. Using GEE, the total processing time was 126 s for the entire study area. This process could be extended extensively to be applied at a larger scale, such as continental and even global.

The map of the dominant land-use types in the study area is shown in Figure 9 and the internal validation error matrix is provided as Table 2. Figure 9 depicts the land-use patterns of the study area. Managed forest dominates much of the study region, surrounded by

Table 2. The internal confusion matrix used to calculate the class noise levels in Open Land-use map.

Classification pixels	Cropland	Commercial/ industrial	Forestry	Residential	Waterbody	Total	User's accuracy (%)
Cropland	347	0	17	2	0	366	94.8
Commercial	0	439	10	7	0	456	96.3
Forest	4	15	1380	40	12	1451	95.1
Residential	0	0	4	230	10	244	94.3
Waterbody	0	1	2	0	42	45	93.3
Total	351	455	1413	279	64	2562	—
Producer's accuracy (%)	98.9	96.4	97.7	82.4	65.6	—	95.2

residential areas and with some residential land penetrating forest cover. Urban build-up areas, including commercial/industrial, are seen scattered among other land-use types. Cropland/agriculture land-use is seen interwoven with residential land mainly in the western region of the image. However, smaller patches of agricultural land-use are notably evidenced to be encroaching on denser forest land cover. Figure 10 depicts a

representable zoomed-in view of land-use mosaics located in Alachua County, Florida. Figure 10(a) is the Google Earth base map and Figure 10(b) is the land-use map of the same area, showing the detailed differentiation of land-use types. The land-use mosaics in Figure 10 demonstrates the lack of detailed accuracy of land-use classifications in the absence of “ground-level” training samples, such as those that were derived from OSM for this study. This highlights the significance and applicability of the integration of OSM and earth observation remote sensing data-sets in achieving highly accurate land-use maps.

In this research, the out-of-bag error is 4.8%, yielding a total classification accuracy of 95.2% for the Random Forest classifier building technique. In comparison to the Google Earth base map, the results of this study indicate that the Random Forest classifier performed well in distinguishing residential and commercial areas, which often proves challenging. Furthermore, this land-use classification and mapping strategy is able to detect residential areas that are under heavy canopy cover. In the forthcoming steps, landscape ecology tools are used to analysis the land-use patterns of the study area and investigate the drivers of the land-use changes.

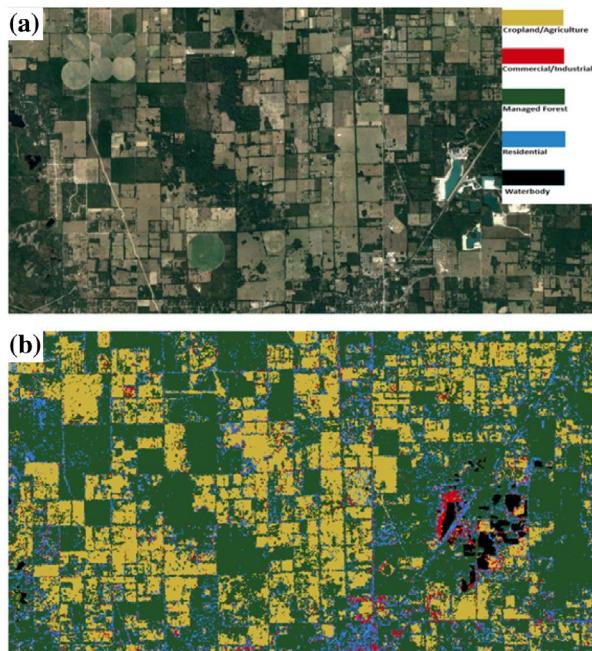


Figure 10. Zoomed-in view of Alachua County, Florida: (a) Google Earth base map; (b) land-use map. Image is courtesy of Google Earth® (2017). Google Earth V 7.1.1.1888 (viewed March 12, 2017). Alachua County, Florida, USA. 29°42'03.94"N, 82°38'16.24"W, elev. 76 ft, Eye alt 11.12 mi. Landsat/Copernicus [November 19, 2016]. <https://www.google.com/earth/index.html>

4.2. Accuracy assessment

To perform accuracy assessment on our open land-use maps, we also performed an accuracy assessment for the OSM-derived training samples as an out-of-bag validation to get estimates of class noise levels in the training data (see Table 2). We constructed an error matrix from the results of our land practice map, and calculated the commission, omission, and the overall errors (Table 2). Table 3 shows the confusion matrices for both internal out-of-bag accuracy and external accuracy. Producer's

Table 3. The external confusion matrix used to calculate the class noise levels in Open Land-use map.

Classification pixels	Cropland	Commercial/ industrial	Forestry	Residential	Waterbody	Total	User's accuracy (%)	Commission error (%)
Cropland	45	4	7	1	1	58	77.6	22.4
Commercial	3	27	10	7	0	47	57.4	42.6
Forest	4	0	43	6	2	55	78.2	21.8
Residential	4	4	3	30	1	42	71.4	28.6
Waterbody	3	1	2	0	42	48	87.5	12.5
Total	59	36	65	44	46	250	—	—
Producer's accuracy (%)	76.3	75.0	66.2	68.2	91.3	—	74.8	—
Omission error (%)	23.7	25.0	33.8	31.8	8.7	—	—	?
Kappa index (%)	70.7	50.3	70.5	65.3	84.7	69.7	—	?

accuracy (omission error) and user's accuracy (commission error) were calculated for each land-use class. The internal validation produced the out-of-bag error during the training of Random Forest classifier. Rows represent the reference data and columns represent the classification results. Producer's accuracy and user's accuracy are listed in the last row and column, respectively. The overall accuracy of the thematic map was 95.2%, and overall 95.1% of the forestry pixels were correctly identified as forest. Waterbody shows the lowest user's accuracy of 93.3%, which is because of the light weight in the training samples, and VGI contributors define waterbody differently in OSM-Nature feature layer.

Accuracy assessment was carried out to evaluate the classification results. The reference test pixels were randomly distributed in the study area. A total number of 250 test points were used for evaluation. The external validation was conducted with around 50 points per land-use class random derived samples from Google Earth. The assessment result is shown in Table 3, depicting the error matrix and condition coefficient for the five classes. In Table 3, Kappa index analysis along with per-class is used for accuracy assessment as well. According to Table 3, among the major land-use types, the Kappa index is 69.7% with the overall accuracy of 74.8%. This means the Open Land-use map and virtual validated Google Earth map match at a moderately rank. Class-based analysis of user's accuracies reveals that 77.6%, 78.2%, 71.4%, and 87.5% of the cropland, forestry, residential, and waterbody, respectively. Therefore, it was concluded that these major land-use classes could be used for land-use mapping purposes at a relatively good level of reliability. There was a low user's accuracy on commercial lands, which was 57.4% because of the total number of mills (confuse with land-use type of forestry) located in the study area of the Southeastern United States.

The mapping strategy presented in this study is recommended mainly for mapping relatively broad land-use classes at large scale. Based on the level of noise tolerance of random classifier, this mapping strategy even has the potential to generate historical multi-temporal land-use maps for change detection and monitoring.

4.3. Implications for land management

Based on the results of this study, the estimated area of forestry, cropland and the other major land-use types could be calculated based on the land-use patterns map. The quantitative estimate of the landscape is an important approach for the evaluation of the impact of the current land-use and might aid in the assessment of human impact on nature resources. Moreover, by incorporating earth observation imagery, it will provide a structured and valuable source of data to monitor land change at the regional scale. Land management practices, usually driven by land ownership types and vice versa, can reflect the objective of ownership types including varied

human land-use needs and associated activities. Land ownership is an important link between human and environmental factors, especially at the regional scale. At the macrosystems scale, forest ownership patterns explain different types of land management practices and trajectories of land-use change (Turner 1989).

In Southeastern United States, forest management is the main driving force of the forest structure, which affects forest ecosystem services (Becknell et al. 2015), and alters forest properties and processes. Southeastern forest system is a fire-dominated system with native trees adapted to short-period stand-clearing events. We classify forest management into four categories: production management, ecological management, passive management, and preservation management. The major forest management types are production and passive management due to the dominant ownership of private owners, logging companies, and investment institutions (e.g. Real Estate Investment Trusts (REITs) and Timber Investment Management Organizations (TIMOs)) (Zhang, Butler, and Nagubadi 2012). The land-use map produced at the regional scale will help landowners, governors, and decision-makers for better understanding the land patterns, processes, and consequences produced by diverse land management practices.

5. Conclusions and perspectives

In this study, we proposed a strategy to mapping land-use and management practices using cloud-computing and combining earth observations with OSM-derived training samples. OSM-Nature, OSM-Landuse, and OSM-Places were reclassified and converted into different major land-use class types. The mapping strategy can be designed based on your understanding of the land of your study of interest. Moreover, the map has more scientific sound, as it is ready for all those types of analysis from landscape ecology, biogeography. It can also be a base map for designated study as well.

OSM is an interesting new platform, which could be implemented in land cover mapping and assessing, because of its geo-located features and labels. The combination of earth observations with OSM features will be a tool for large-scale mapping and will present great opportunities in the creation ideas because of its unique workflow to fulfill specific mapping needs. The advantages of incorporating and deriving training samples from OSM can significantly enlarge the amount of training samples. The Random Forest was used for classification of a multi-source remote sensing based on geo-located supervised training data-set derived from OSM. In experiments, Random Forest performed well with the out-of-bag validation accuracy of 95.2% and external validation accuracy of 74.8%. Results also show that the Random Forest algorithm yields accurate land-use classifications at a regional scale, with 74.8% overall accuracy and a Kappa index of 69.7%. The time cost of

mapping a single Landsat scene is 126 s by GEE platform, which covers 1.08×10^9 pixels. This mapping strategy can be easily extended to a larger scale such as state level, continental level, and, even, global level.

According to the temporal availability of our covariates, the up-to-date land-use pattern mapping strategy can be extended to a multi-temporal changes database. It also has the capability to make predictions based on the spatio-temporal land-use pattern changes and patch size distributions. There is still a strong need for a continental coverage land-use map with a higher temporal resolution, which will reflect the land-use transition processes at shorter intervals. It is an urgent issue at present; On the other hand, in order to abate the ecological impacts caused by the rapidly increasing construction of road networks, we need the better road network designs combining both top-down and bottom-up processes, which would retain and preserve large forest patches and decelerate the parcellation rate. The use of open-data source will open our eyes of exploring the area that does not have an official classification database. Researchers and applicants can design their own mapping strategy based on their mapping objectives. The open land map can use single or multiple classifiers to fulfill the mapping goals as well.

Acknowledgments

We would like to thank the Google Earth Engine Development team for input and assistance with the API platform.

Notes on contributors

Di Yang is a PhD candidate at the Department of Geography, University of Florida. She is also a member of Land-use and Environmental Change Institute (LUECI), University of Florida. Di has an MSc degree in Environmental Engineering from Texas A&M University. The general theme of her work is to gain a quantitative understanding of the relationships between humans, landscape patterns, and ecological process by VGI applications.

Chiung-Shiuan Fu obtained her PhD degree from Geomatics program, Civil Engineering at Purdue University. She is a postdoctoral associate in Geography Department at the University of Florida. Her current research interests include: (1) land-use/cover mapping and analysis using remote sensing and GIS, (2) ensemble learning for remote sensing classification, and (3) temporal-spatial forest management monitoring.

Audrey C. Smith is a PhD student in the Department of Geography at the University of Florida. She has a BA degree in Anthropology and an MSc degree in Geography from the University of Florida. Her research focuses on African land cover and land-use change, largely in the context of human well-being, agriculture, and ecosystem services outcomes.

Qiang Yu is a PhD candidate at the College of Forestry, Beijing Forestry University. His main research focuses on the

applications of 3S technology in resources and environment management, landscape ecology, and land-use simulations.

ORCID

Di Yang  <http://orcid.org/0000-0002-4010-6163>
Chiung-Shiuan Fu  <http://orcid.org/0000-0003-0528-6231>

References

Abdullah, S. A., and N. Nakagoshi. 2007. "Forest Fragmentation and Its Correlation to Human Land Use Change in the State of Selangor, Peninsular Malaysia." *Forest Ecology and Management* 241 (1–3): 39–48.

Agarwal, C., G. M. Green, J. M. Grove, T. P. Evans, and C. M. Schweik. 2002. *A Review and Assessment of Land-use Change Models: Dynamics of Space, Time, and Human Choice*. UFS Technical Report NE-297. Burlington, VT: U.S. Department of Agriculture, Forest Service, Northeastern Research Station.

Azzari, G., and D. B. Lobell. 2017. "Landsat-based Classification in the Cloud: An Opportunity for a Paradigm Shift in Land Cover Monitoring." *Remote Sensing of Environment* (Online Published). doi:10.1016/j.rse.2017.05.025.

Bateman, I. J., A. R. Harwood, G. M. Mace, R. T. Watson, D. J. Abson, B. Andrews, A. Binner, A. Crowe, B. H. Dugdale, and C. Fezzi. 2013. "Bringing Ecosystem Services into Economic Decision-making: Land use in the United Kingdom." *Science* 341 (6141): 45–50.

Becknell, J. M., A. R. Desai, M. C. Dietze, C. A. Schultz, G. Starr, P. A. Duffy, J. F. Franklin, A. Pourmokhtarian, J. Hall, P. C. Stoy, and M. W. Binford. 2015. "Assessing Interactions Among Changing Climate, Management, and Disturbance in Forests: a Macrosystems Approach." *BioScience* 65 (3): 263–274.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Bryan, B. A., S. Barry, and S. Marvanek. 2009. "Agricultural Commodity Mapping for Land Use Change Assessment and Environmental Management: An Application in the Murray–Darling Basin, Australia." *Journal of Land Use Science* 4 (3): 131–155.

Chander, G., B. L. Markham, and D. L. Helder. 2009. "Summary of Current Radiometric Calibration Coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI Sensors." *Remote Sensing of Environment* 113 (5): 893–903.

Ciepluch, B., P. Mooney, R. Jacob, and A. C. Winstanley. 2009. "Using OpenStreetMap to Deliver Location-based Environmental Information in Ireland." *SIGSPATIAL Special* 1 (3): 17–22.

Congalton, R. G. 1991. "A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data." *Remote Sensing of Environment* 37 (1): 35–46.

Elwood, S. 2008. "Volunteered Geographic Information: Key Questions, Concepts and Methods to Guide Emerging Research and Practice." *GeoJournal* 72 (3–4): 133–135.

Estima, J., and M. Painho. 2013. "Exploratory Analysis of OpenStreetMap for Land use Classification." Paper presented at The Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, GEOCROWD'13, Orlando, FL, USA, November 5–8, 39–46.

Estima, J., and M. Painho. 2015. "Investigating the Potential of OpenStreetMap for Landuse/Land Cover Production: A

Case Study for Continental Portugal." In *OpenStreetMap in GIScience*, 273–293. Berlin, Heidelberg: Springer-Verlag.

Foley, J. A., R. DeFries, G. P. Asner, C. Barford, G. Bonan, S. R. Carpenter, F. S. Chapin, et al. 2005. "Global Consequences of Land Use." *Science* 309 (5734): 570–574.

Folke, C., C. S. Holling, and C. Perrings. 1996. "Biological Diversity, Ecosystems, and the Human Scale." *Ecological Applications* 6 (4): 1018–1024.

Girres, J. F., and G. Touya. 2010. "Quality Assessment of the French OpenStreetMap Dataset." *Transactions in GIS* 14 (4): 435–459.

Haklay, M., A. Singleton, and C. Parker. 2008. "Web Mapping 2.0: The Neogeography of the GeoWeb." *Geography Compass* 2 (6): 2011–2039.

Hartley, M. J. 2002. "Rationale and Methods for Conserving Biodiversity in Plantation Forests." *Forest Ecology and Management* 155 (1): 81–95.

Hegazy, I. R., and M. R. Kaloop. 2015. "Monitoring Urban Growth and Land Use Change Detection with GIS and Remote Sensing Techniques in Daqahlia Governorate Egypt." *International Journal of Sustainable Built Environment* 4 (1): 117–124.

Huete, A., K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira. 2002. "Overview of the Radiometric and Biophysical Performance of the MODIS Vegetation Indices." *Remote Sensing of Environment* 83 (1–2): 195–213.

Jarvis, A., H. I. Reuter, A. Nelson, and E. Guevara. 2008. "Hole-filled SRTM for the Globe Version 4." Available from the CGIAR-CSI SRTM 90m Database (<http://srtm.csi.cgiar.org>). 2008.

Jokar, A. J., M. Helbich, M. Bakillah, J. Hagenauer, and A. Zipf. 2013. "Toward Mapping Land use Patterns from Volunteered Geographic Information." *International Journal of Geographical Information Science* 27 (12): 2264–2278.

Jokar, A. J., M. Helbich, M. Bakillah, and L. Loos. 2015. "The Emergence and Evolution of OpenStreetMap: A Cellular Automata Approach." *International Journal of Digital Earth* 8 (1): 76–90.

Jokar, A. J., P. Mooney, A. Zipf, and A. Schauss. 2015. "Quality Assessment of the Contributed Land Use Information from OpenStreetMap versus Authoritative Datasets". In *OpenStreetMap in GIScience*, 37–58. Berlin, Heidelberg: Springer-Verlag.

Johansen, K., S. Phinn, and M. Taylor. 2015. "Mapping Woody Vegetation Clearing in Queensland, Australia from Landsat Imagery Using the Google Earth Engine." *Remote Sensing Applications: Society and Environment* 1: 36–49.

Johnson, B. A., and K. Iizuka. 2016. "Integrating OpenStreetMap Crowdsourced Data and Landsat Time-series Imagery for Rapid/Land Cover (LULC) Mapping: Case Study of the Laguna de Bay Area of the Philippines." *Applied Geography* 67: 140–149.

Lambin, E. F., B. L. Turner, H. J. Geist, S. B. Agbola, A. Angelsen, J. W. Bruce, O. T. Coomes, et al. 2001. "The Causes of Land use and Land-cover Change: Moving beyond the Myths." *Global Environmental Change* 11 (4): 261–269.

Lawler, J. J., D. J. Lewis, E. Nelson, A. J. Plantinga, S. Polasky, J. C. Withey, D. P. Helmers, S. Martinuzzi, D. Pennington, and V. C. Radeloff. 2014. "Projected Land use Change Impacts on Ecosystem Services in the United States." *Proceedings of the National Academy of Sciences of USA* 111 (20): 7492–7497. doi:10.1073/pnas.1405557111.

Lesslie, R., M. Barson, and J. Smith. 2006. "Land Use Information for Integrated Natural Resources Management – A Coordinated National Mapping Program for Australia." *Journal of Land Use Science* 1 (1): 45–62.

Matson, P. A., W. J. Parton, A. G. Power, and M. J. Swift. 1997. "Agricultural Intensification and Ecosystem Properties." *Science* 277 (5325): 504–509.

Neis, P., D. Zielstra, and A. Zipf. 2012. "The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011." *Future Internet* 4: 1–21.

Omernik, J. M., and G. E. Griffith. 2014. "Ecoregions of the Conterminous United States: Evolution of a Hierarchical Spatial Framework." *Environmental Management* 54 (6): 1249–1266.

Padarian, J., B. Minasny, and A. B. McBratney. 2015. "Using Google's Cloud-based Platform for Digital Soil Mapping." *Computers & Geosciences* 83: 80–88.

Patel, N. N., E. Angiuli, P. Gamba, A. Gaughan, G. Lisini, F. R. Stevens, and G. Trianni. 2015. "Multitemporal Settlement and Population Mapping from Landsat Using Google Earth Engine." *International Journal of Applied Earth Observation and Geoinformation* 35: 199–208.

Rodriguez-Galiano, V. F., B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez. 2012. "An Assessment of the Effectiveness of a Random Forest Classifier for Land-cover Classification." *ISPRS Journal of Photogrammetry and Remote Sensing* 67: 93–104.

See, L., A. Comber, C. Salk, S. Fritz, M. van der Velde, C. Perger, C. Schill, I. McCallum, F. Kraxner, and M. Obersteiner. 2013. "Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-experts." *PLOS ONE* 8 (7): e69958.

See, L., D. Schepaschenko, M. Lesiv, I. McCallum, S. Fritz, A. Comber, and M. A. Siraj. 2015. "Building a Hybrid Land Cover Map with Crowdsourcing and Geographically Weighted Regression." *ISPRS Journal of Photogrammetry and Remote Sensing* 103: 48–56.

Senf, C., P. Hostert, and S. van der Linden. 2012. "Using MODIS Time Series and Random Forests Classification for Mapping Land Use in South-East Asia." Paper presented at the 2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Munich, Germany, July 22–27, 6733–6736.

Simard, M., N. Pinto, J. B. Fisher, and A. Baccini. 2011. "Mapping Forest Canopy Height Globally with Spaceborne Lidar." *Journal of Geophysical Research: Biogeosciences* 116 (G4): G04021. doi:10.1029/2011JG001708.

Soulard, C. E., C. M. Albano, M. L. Villarreal, and J. J. Walker. 2016. "Continuous 1985–2012 Landsat Monitoring to Assess Fire Effects on Meadows in Yosemite National Park, California." *Remote Sensing* 8 (5): 371–387.

Sumarga, E., and L. Hein. 2014. "Mapping Ecosystem Services for Land Use Planning, the Case of Central Kalimantan." *Environmental Management* 54 (1): 84–97.

Turner, M. G. 1989. "Landscape Ecology: The Effect of Pattern on Process." *Annual Review of Ecology and Systematics* 20 (1): 171–197.

Visconti, P., M. Di Marco, J. G. Alvarez-Romero, S. R. Januchowski-Hartley, R. L. Pressey, R. Weeks, and C. Rondinini. 2013. "Effects of Errors and Gaps in Spatial Data Sets on Assessment of Conservation Progress." *Conservation Biology* 27 (5): 1000–1010.

Vitousek, P. M., H. A. Mooney, J. Lubchenco, and J. M. Melillo. 1997. "Human Domination of Earth's Ecosystems." *Science* 277 (5325): 494–499.

Wear, D. N., and J. Greis. 2002. "Southern Forest Resource Assessment: Summary of Findings." *Journal of Forestry* 100 (7): 6–14.

Wear, D. N., and J. Greis. 2013. *The Southern Forest Futures Project*. Technical Report (SRS-GTR-178). Asheville, NC: USDA-Forest Service, Southern Research Station. <https://www.srs.fs.usda.gov/futures/technical-report/>.

Weng, Q. 2002. "Land Use Change Analysis in the Zhujiang Delta of China Using Satellite Remote Sensing, GIS and Stochastic Modelling." *Journal of Environmental Management* 64 (3): 273–284.

Yang, D. 2017. "Mapping Regional Landscape by Using OpenStreetMap (OSM): A Case Study to Understand Forest Patterns in Maya Zone, Mexico." In *Volunteered Geographic Information and the Future of Geospatial Data*, edited by C. E. C. Campelo, M. Bertolotto and P. Corcoran, 138–157. Hershey, PA: IGI Global.

Yu, L., and P. Gong. 2012. "Google Earth as a Virtual Globe Tool for Earth Science Applications at the Global Scale: Progress and Perspectives." *International Journal of Remote Sensing* 33 (12): 3966–3986.

Zhang, D., B. J. Butler, and R. V. Nagubadi. 2012. "Institutional Timberland Ownership in the US South: Magnitude, Location, Dynamics, and Management." *Journal of Forestry* 110 (7): 355–361.

Zhu, Z., and C. E. Woodcock. 2012. "Object-based Cloud and Cloud Shadow Detection in Landsat Imagery." *Remote Sensing of Environment* 118: 83–94.

Zook, M., M. Graham, T. Shelton, and S. Gorman. 2010. "Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake." *World Medical & Health Policy* 2 (2): 7–33.