Applied Geography 67 (2016) 140–149

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

Integrating OpenStreetMap crowdsourced data and Landsat time-series imagery for rapid land use/land cover (LULC) mapping: Case study of the Laguna de Bay area of the Philippines

Brian A. Johnson [a, *], Kotaro Iizuka [b]

[a] Institute for Global Environmental Strategies, 2108-11 Kamiyamaguchi, Hayama, Kanagawa, 240-0115, Japan
[b] Research Institute for Sustainable Humanosphere, Kyoto University, Gokasho, Uji-City, Kyoto, 611-0011, Japan

ARTICLE INFO

Article history:
Received 21 September 2015
Received in revised form
15 December 2015
Accepted 15 December 2015
Available online xxx

Keywords:
OpenStreetMap
Volunteered geographic information
Citizen science
Crowdsourced data
Random forest
Landsat 8
Google Earth Engine

ABSTRACT

We explored the potential for rapid land use/land cover (LULC) mapping using time-series Landsat satellite imagery and training data (for supervised classification) automatically extracted from crowdsourced OpenStreetMap (OSM) "landuse" (OSM-LU) and "natural" (OSM-N) polygon datasets. The main challenge with using these data for LULC classification was their high level of noise, as the Landsat images all contained varying degrees of cloud cover (causes of attribute noise) and the OSM polygons contained locational errors and class labeling errors (causes of class noise). A second challenge arose from the imbalanced class distribution in the extracted training data, which occurred due to wide discrepancies in the area coverage of each OSM-LU/OSM-N class. To address the first challenge, three relatively noise-tolerant algorithms – naïve bayes (NB), decision tree (C4.5 algorithm), and random forest (RF) – were evaluated for image classification. To address the second challenge, artificial training samples were generated for the minority classes using the synthetic minority over-sampling technique (SMOTE). Image classification accuracies were calculated for a four-class, five-class, and six-class LULC system to assess the capability of the proposed methods for mapping relatively broad as well as more detailed LULC types, and the highest overall accuracies achieved were 84.0% (four-class SMOTE-RF result), 81.0% (five-class SMOTE-RF result), and 72.0% (six-class SMOTE-NB result). RF and NB had relatively similar overall accuracies, while those of C4.5 were much lower. SMOTE led to higher classification accuracies for RF and C4.5, and in some cases for NB, despite the noise in the training set. The main advantages of the proposed methods are their cost- and time-efficiency, as training data for supervised classification is automatically extracted from the crowdsourced datasets and no pre-processing for cloud detection/cloud removal is performed.

1. Introduction

1.1. Land use/land cover mapping by supervised image classification

Land use/land cover (LULC) maps derived from remotely-sensed imagery are used for a wide range of applications, including land use planning (Dewan & Yamaguchi, 2009), population estimation (Fisher & Langford, 1996), natural resource inventory (Food and Agriculture Organization of the United Nations, 2010), and biodiversity modeling (Roy & Tomar, 2000). The number and types of LULC classes mapped vary from study to study depending on the intended application, and can range from a simple binary classification (e.g. to map buildings (Belgiu & Drăguţ, 2014), residential areas (Johnson, Bragais, Endo, Magcale-Macandog, & Macandog, 2015), or forests (Shimada et al., 2014)) to a detailed species-level vegetation classification with dozens of classes (Yu et al., 2006). LULC maps are often generated using supervised image classification algorithms; algorithms which utilize the spectral and/or contextual information of sample pixels with LULC class labels (i.e. training data) to classify the remaining unlabeled pixels in the image (Jensen, 2005). The accuracy of the maps generated using supervised classification techniques is affected by, among other things, the quality of the remotely-sensed imagery and the quality and quantity of the training data (Brodley & Friedl, 1999; Huang, Davis, & Townshend, 2002; Lippitt, Rogan, & Li, 2008).

* Corresponding author.
E-mail address: johnson@iges.or.jp (B.A. Johnson).

http://dx.doi.org/10.1016/j.apgeog.2015.12.006
0143-6228/

1.2. Image quality and LULC classification

The quality of remotely-sensed imagery for LULC classification is often affected by the level of cloud cover in the image, as clouds block a sensor's view of the ground at many electromagnetic wavelength regions (e.g. visible (VIS), near infrared (NIR), shortwave infrared (SWIR), and thermal infrared (TIR) wavelength regions). When clouds are present at a pixel location, pixel information is extracted from the cloud rather than from the feature on the ground, resulting in "attribute noise" (Nettleton, Orriols-Puig, & Fornells, 2010) because the information is not useful for classification of the pixel. In tropical regions and other areas with frequent cloud cover, this can make LULC mapping particularly challenging (Foody, Boyd, & Cutler, 2003; Hoan et al., 2013). To overcome this issue of attribute noise from cloud cover, a recent study by Schneider (2012) found that a so-called "brute force approach" of drastically increasing the data quantity (i.e. using many time-series images) could reduce the negative impact of noise in individual images. Schneider (2012) simply used all Landsat images with relatively little cloud cover, including "scan line corrector off" Landsat 7 images (http://landsat.usgs.gov/products_slcoffbackground.php) containing missing data, for mapping LULC change in several urban areas of China, and found that a boosted decision tree algorithm could efficiently make use of the high dimensional, noisy data for classification. Although more sophisticated methods exist for dealing with cloud contamination, e.g. various cloud-removal methods which use ancillary datasets to estimate the spectral and spatial properties of pixels located under cloud cover (Hoan et al., 2013; Jia et al., 2014; X. Zhu, Gao, Liu, & Chen, 2012; Zhu & Woodcock, 2012), the brute force approach has the advantages of speed and simplicity because it does not require additional image pre-processing or the use of ancillary datasets.

1.3. Training data quality/quantity and LULC classification

In addition to attribute noise, class labeling errors in the training data (i.e. pixels with wrong class assignments), i.e. "class noise" (Frenay & Verleysen, 2013), can also have an impact on classification accuracy. Unfortunately, high quality training data for LULC classification can be time-consuming, difficult, and/or expensive to obtain, particularly if ground surveys are needed to collect the data. LULC is rapidly changing in many countries due to urbanization, so up-to-date LULC maps are needed in these areas for effective land use management and planning. However, there is not always sufficient time or funding available to gather high quality, up-to-date training data.

Training data quantity is also an issue, particularly when the quantity of image data used for classification (e.g. number of time-series images or number of image bands) is high. This is because increasing the number of classification variables past a certain threshold (which varies depending on the classification algorithm) can lead to lower classification accuracy if the number of training samples is not also increased, which is known as the Hughes phenomenon (Hughes, 1968). Section 1.2. discusses that higher image quantity can help to overcome low image quality, so this issue of training data quantity is quite related.

1.4. Crowdsourced geographic datasets as a source of training data

Volunteered geographic information (VGI), geographic information on LULC features provided by citizen volunteers rather than official government agencies (Goodchild, 2007), is a relatively new source of freely-available crowdsourced information that may serve as a supplemental or even alternative source of training data (Estima & Painho, 2015; Jokar Arsanjani, Helbich, & Bakillah, 2013). In particular, this (class-labeled) VGI data could be quite useful in cases where higher quality training data cannot be collected in sufficient quantity. VGI is typically created by volunteers tracing of features onto georeferenced aerial/satellite images or by their collection of GPS data in the field (Neis & Zielstra, 2014). As one example, OpenStreetMap (OSM; https://www.openstreetmap.org/) lets volunteers create and edit geographic data online using Microsoft Bing Aerial Imagery as a backdrop. Although VGI is less accurate than geographic information from official sources in some cases due to many volunteers' lack of formal training (Estima & Painho, 2013; Haklay, 2010; Jokar Arsanjani, Mooney, & Zipf, 2015), it often provides the cheapest (and sometimes only) source of geographic information (Goodchild, 2007). OSM is one of the largest sources of VGI (Neis & Zielstra, 2014; Neis & Zipf, 2012), so it has the potential to provide a high quantity of training data in many areas (although data quality may be an issue). Some popular OSM datasets include "roads" (http://wiki.openstreetmap.org/wiki/Map_Features#Highway), "buildings" (http://wiki.openstreetmap.org/wiki/Buildings), "points of interest" (http://wiki.openstreetmap.org/wiki/Points_of_interest), "landuse" (http://wiki.openstreetmap.org/wiki/Map_Features#Landuse), and "natural" (http://wiki.openstreetmap.org/wiki/Map_Features#Natural).

As already stated, although these OSM datasets (and other crowdsourced datasets) can potentially be used to extract a large quantity of training data due to their relatively wide area coverage, a challenge with using them is the presence of errors in the OSM data, which would lead to class noise in the training dataset. One previous study investigated the class noise in the "landuse" (OSM-LU) and "natural" (OSM-N) datasets by comparing them with an official LULC map of Portugal, and found a 76% agreement between the OSM data and the official LULC map in areas where the datasets overlapped (Estima & Painho, 2013). Another study performed a similar comparison in four major German cities (Berlin, Frankfurt, Hamburg, and Munich) and found that the agreement between the OSM-LU/OSM-N and official datasets ranged from 64% (for Hamburg) to 77% (for Frankfurt) (Jokar Arsanjani et al., 2015). The results of these studies suggest that the OSM-LU and OSM-N datasets can be seen as at least moderately accurate.

In terms of past studies using OSM data to extract training data for LULC classification, there is only one that we are aware of, and it involved using OSM "points-of-interest" to extract training data for LULC classification (Jokar Arsanjani et al., 2013). In that study, the OSM data was visually pre-screened and all mislabeled points were removed prior to extracting training data (Jokar Arsanjani et al., 2013). Manual screening can be quite time- and labor-intensive though, particularly if there are many OSM polygons to inspect. On the other hand, some classification algorithms are relatively tolerant to mislabeled training data (i.e. class noise), so it may be possible to use the OSM data for classification even without manual pre-screening.

A previous study assessed the class noise-tolerance of four broad types of classification algorithms - probabilistic, decision tree (DT), instance-based, and support vector machines (SVM) algorithms – and found probabilistic methods, specifically the naïve bayes (NB) algorithm (John & Langley, 1995), best dealt with class noise, while DT and instance-based algorithms showed moderate performance, and SVM performed the worst due to its sensitivity to mislabeled training data located along the support vectors (Nettleton et al., 2010). Another study, which considered the effects of class imbalance (i.e. discrepancy in the number of training samples per class) in addition to class noise and attribute noise, found that the Random Forest (RF) algorithm (Breiman, 2001), an

ensemble DT learner, achieved the highest classification accuracies (Folleco, Khoshgoftaar, VanHulse, & Napolitano, 2009). Based on the results of these studies, naïve bayes (NB), DT, and RF algorithms can be considered as some possible candidates for dealing with class noise in OSM data. Since these three algorithms have been described in great detail in previous works (Breiman, 2001; John & Langley, 1995; Quinlan, 1993) and been applied relatively widely for LULC mapping (Breiman, 2001; Cracknell & Reading, 2014; Gislason, Benediktsson, & Sveinsson, 2006; Johnson & Xie, 2013; Pal, 2005; Reiche et al., 2013; Wang et al., 2014), here we skip the detailed descriptions and just briefly explain what makes them relatively noise-tolerant. NB assumes the predictive attributes of a class are conditionally independent (John & Langley, 1995), which reduces its sensitivity to attribute noise because the noisier attributes do not reduce the quality of less noisy attributes, and its use of conditional probabilities to derive posterior probabilities for classification is another benefit because the conditional probabilities are less sensitive to individual noisy training samples (Nettleton et al., 2010). DTs are built of decision nodes, where preliminary classification decisions occur, and leaves, where the final class assignment takes place (Quinlan, 1993). Unreliable leaves are typically pruned (i.e. removed) based on some evaluation criteria, e.g. the confidence level parameter for the C4.5 algorithm (Quinlan, 1993), to avoid overfitting to the training data, and the noise tolerance of DT algorithms comes from this pruning, as mislabeled training instances tend to generate some of the unreliable leaves (Clark & Niblett, 1989). RF consists of an ensemble of unpruned DTs, each generated using a random subset of the training and attribute data, and final classification is done by majority voting (Breiman, 2001). Since the DTs in the ensemble are typically unpruned, the noise tolerance of RF is instead due to its random subsetting of the training data and attribute data, which increases its diversity and reduces overfitting to the training data (Breiman, 2001; Folleco et al., 2009).

1.5. Objectives

The objective of our study was to evaluate the potential for using freely-available moderate spatial resolution (30 m) time-series imagery in combination with OSM-derived training data for LULC mapping. To test the performance of this proposed classification approach under noisy conditions, we chose a study area with frequent cloud cover (a cause of attribute noise) and undergoing rapid LULC change (a cause of class noise if outdated training data is used). Although noise is often removed or corrected in remote sensing studies prior to LULC classification, these noise removal/correction procedures can be computationally- and/or labor-intensive. To simplify the LULC mapping process, we instead explored the possibility of classifying the noisy data (i.e. without first removing/correcting the noise) using noise-tolerant classification algorithms. We tested three noise-tolerant classification algorithms – C4.5, RF, and NB – to evaluate their relative performance. Some significant benefits of the proposed approach are: (1) it uses freely- and publically-available image and training data, and (2) it does not require noise removal (i.e. cloud removal or manual correction of mislabeled training data) prior to classification.

2. Methods

2.1. Study area and data

The area surrounding the Laguna de Bay of the Philippines was chosen as the study site. This study area is approximately 125 km × 90 km, and includes the city of Manila as well as the surrounding suburban, agricultural, and forest areas. This area typically has at least some cloud cover, and it is often entirely covered by clouds during the rainy season. It is also undergoing rapid LULC change due to urbanization (Tongson & Faraon, 2012), so training data collected in the past may be outdated. These factors make it a good study site for evaluating the tolerance of different classification methods to attribute and class noise.

All 30 m resolution Landsat 8 Normalized Difference Vegetation Index (NDVI) images of the study area (World Reference System 2: Path 116, Row 50) acquired between 01/2014-07/2015 with less than 75% cloud cover (based on visual estimate) were downloaded using the Google Earth Engine (https://earthengine.google.org/#intro, last accessed 07/27/2015). NDVI images were calculated by:

$$\frac{band\_5_{TOA} - band\_4_{TOA}}{band\_5_{TOA} + band\_4_{TOA}}$$

where $band\_5_{TOA}$ is the NIR image band (0.85–0.88 μm) in top-of-atmosphere reflectance units and $band\_4_{TOA}$ is the red image band (0.64–0.67 μm) in top-of-atmosphere units. In total, 27 NDVI images were used, none of which were cloud-free (image dates and cloud cover estimates shown in Table 1). The rationale for using time-series NDVI data was to obtain detailed information on vegetation phenology for better discrimination between LULC types. A large number of time-series images was used to ensure that all pixels had at least one cloud-free observation from each season (as using only 1–2 images from each season would still leave many pixels without any cloud-free observation). Although additional Landsat 8 bands could also be used (e.g. bands 2–7 in the visible-to-shortwave infrared spectrum), download times/computation times would increase significantly (additional $27 \times 6 = 162$ image bands/classification variables if bands 2–7 are also included), so only NDVI data was used since the goal was rapid LULC mapping.

For the purpose of extracting training data for LULC classification at 30 m spatial resolution, the "landuse" (OSM-LU) and "natural" (OSM-N) datasets seem to be most relevant OSM datasets, as they represent a variety of manmade and natural LULC types and are usually large enough to detect at this resolution, whereas some roads, buildings, and points of interest may be too small for the sensor to detect. OSM-LU and OSM-N polygon data were downloaded in shapefile format from http://download.geofabrik.de/asia/ (last accessed 05/20/2015) and merged together using GIS software. OSM-LU and OSM-N classes with fewer than 10 polygons or which could not be clearly associated with a corresponding LULC class (described in Section 2.2.) were excluded, as were polygons with an area of less than 900 m^2 (since Landsat 8 pixels were 30 m × 30 m in size). Fig. 1 shows an overview of the study area imagery and OSM data.

2.2. Converting OSM-LU/OSM-N classes to LULC classes

The OSM-LU and OSM-N classes were converted into six LULC classes: "impervious", "farm", "forest", "grass", "orchard", and "water", as shown in Table 2. Since the "forest" and "orchard" classes both contained tree cover (orchards consisted of mainly coconut, mango, and banana trees), and it is often difficult to distinguish between forests and orchards even when high quality training data is used (Oetter, Cohen, Berterretche, Maiersperger, & Kennedy, 2001), we also tested a five-class LULC system in which "forest" and "orchard" classes were merged into a more general "tree" class. Finally, an even broader four-class LULC system was tested in which the "farm" and "grass" classes were merged into an "other vegetation" class. Aside from common types of errors in the

Table 1
Landsat 8 images used in this study. Values in parenthesis "()" are the cloud cover percentages for the entire Landsat scene based on the image metadata at http://earthexplorer.usgs.gov/(last accessed 12/14/2015). *The 7-Feb image had significantly more than 2% cloud cover in our study area based on visual inspection, as can be seen in Fig. 1.**The 18-Aug had significantly less than 85% cloud cover in our study area based on visual inspection.

2014			2015	
6-Jan (41%)	28-Apr (22%)	5-Oct (39%)	9-Jan (57%)	15-Apr (28%)
22-Jan (39%)	14-May (31%)	21-Oct (42%)	25-Jan (23%)	1-May (8%)
7-Feb* (2%)	30-May (22%)	6-Nov (47%)	10-Feb (24%)	17-May (19%)
23-Feb (55%)	15-Jun (21%)	22-Nov (23%)	26-Feb (23%)	2-Jun (14%)
27-Mar (43%)	1-Jul (51%)		14-Mar (23%)	20-Jul (10%)
12-Apr (38%)	18-Aug** (85%)		30-Mar (10%)	

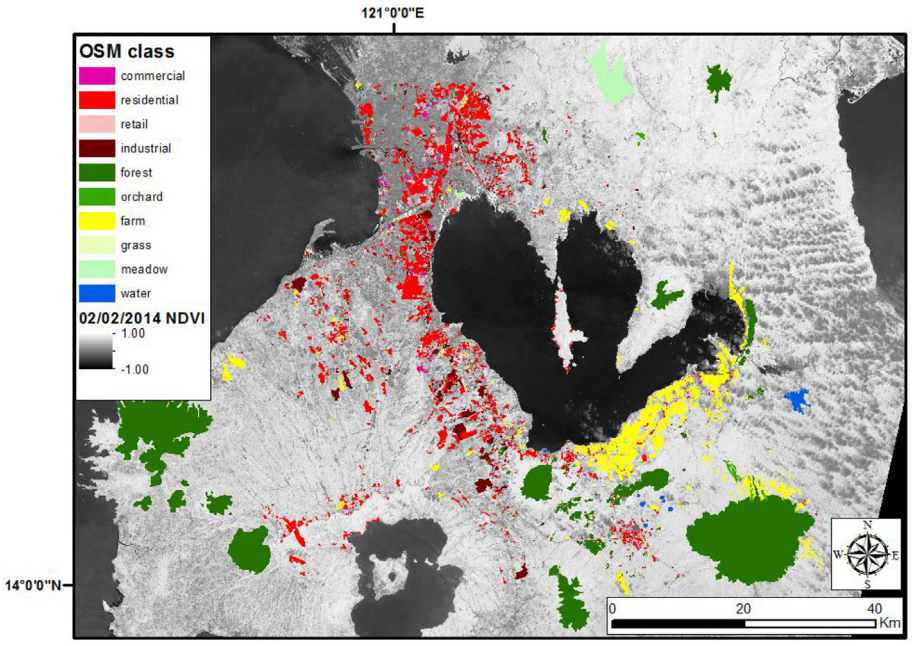

Fig. 1. Study area and data used. OpenStreetMap (OSM) "landuse" and "natural" classes used are shown as colored polygons (see electronic version for color reproduction) overlayed onto a gray-scale NDVI image (acquired on 02/02/2014). Gray areas in the East part of the image show scattered clouds. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Land use/land cover (LULC) classes derived from the OSM "landuse" (OSM-LU) and OSM "natural" (OSM-N) shapefiles.

OSM class	OSM shapefile	# of polygons	Area (ha.)	LULC class (six-class system)	LULC class (five-class system)	LULC class (four-class system)
Commercial	OSM-LU	165	735	Impervious	Impervious	Impervious
Residential	OSM-LU	2159	22,429	Impervious	Impervious	Impervious
Retail	OSM-LU	223	10,388	Impervious	Impervious	Impervious
Industrial	OSM-LU	442	4878	Impervious	Impervious	Impervious
Forest	OSM-N	187	45,868	Forest	Tree	Tree
Orchard	OSM-LU	31	565	Orchard	Tree	Tree
Farm	OSM-LU	751	13,681	Farm	Farm	Other veg.
Grass	OSM-LU	71	280	Grass	Grass	Other veg.
Meadow	OSM-LU	40	3125	Grass	Grass	Other veg.
Water	OSM-N	60	11,289	Water	Water	Water

OSM polygon datasets (e.g. inaccurate boundary delineations of OSM-LU/OSM-N features), another challenge in using the data for LULC mapping came from the fact the OSM polygons could contain multiple LULC types (e.g. an "industrial" OSM-LU polygon could potentially contain trees or grass in addition to impervious surfaces), so this type of mismatching between the OSM-LU/OSM-N classes and our selected LULC classes can be considered an additional source of class noise.

2.3. Training and validation data

The OSM polygons were randomly split 50/50 to create independent training and validation datasets. "Residential" polygons were deleted from the training dataset because residential areas contained a mixture of impervious and vegetated LULC types, and were thus not deemed useful for LULC classification purposes. For the remaining training polygons, random points were generated within each polygon (minimum point spacing of 90 m, or three

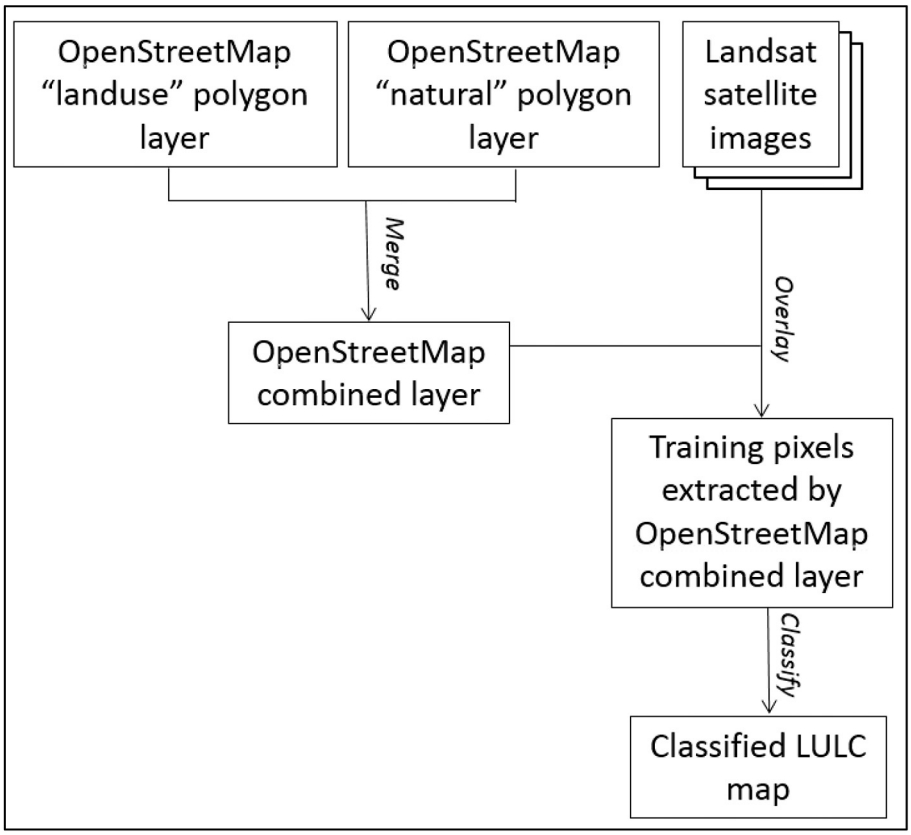

Fig. 2. Overview of the methodology workflow.

Landsat 8 pixels, to reduce the spatial autocorrelation in the training set (Congalton, 1991), and the pixels at these point locations were used as training pixels for classification. The 27 NDVI images and one maximum NDVI image (calculated from the 27 images) were used as classification variables. The number of extracted training pixels varied widely by LULC class, with 7431 pixels for "forest", 1441 for "farm", 969 for "impervious", 446 for "grass", 205 for "water", and 53 for "orchard".

For the validation dataset, 50 points were randomly generated for each LULC class inside the OSM polygons set aside for validation, and the actual LULC at each point was identified using high resolution Google Earth imagery from the years 2014–2015. Overall accuracy (OA), producer's accuracy (PA), and user's accuracy (UA) (Jensen, 2005) were used as the accuracy metrics.

In addition to performing accuracy assessment on our classified maps, we also performed an accuracy assessment for the OSM-derived training data itself to get estimates of class noise levels in the training data. This was done by converting the OSM classes to LULC classes (as in Table 2) at the 300 validation point locations, and comparing the result with the actual LULC.

2.4. Image classification

Image classification was performed using the NB, C4.5, and RF algorithms implemented in the Weka open source data mining software package (Hall et al., 2009). NB and C4.5 (called j48 in Weka) were run using their default parameters (NB set to assume a Gaussian distribution, and C4.5's confidence factor set to 0.25), as is consistent with past studies (Folleco et al., 2009; Nettleton et al., 2010). For RF, the number of randomly-selected attributes (*#a*) to consider at each decision node ($\#a = \log_2(A) + 1$, where A is the total number of attributes) and the number of DTs (*#dt*) to use for the ensemble ($\#dt = 500$) were set based on advice from past research (Breiman, 2001; Lawrence, Wood, & Sheley, 2006).

Table 3
Error matrix used to calculate the class noise levels in the OSM-LU and OSM-N datasets.

	True LULC								
		Impervious	Forest	Orchard	Farm	Grass	Water	Sum	UA (%)
OSM-derived LULC	Impervious	38	0	0	2	0	0	40	95.0
	Forest	5	41	2	0	29	1	78	52.6
	Orchard	0	4	42	1	0	0	47	89.4
	Farm	0	1	4	47	1	0	53	88.7
	Grass	7	4	2	0	20	3	36	55.6
	Water	0	0	0	0	0	46	46	100.0
	Sum	50	50	50	50	50	50	300	
	PA (%)	76.0	82.0	84.0	94.0	40.0	92.0		
									OA = 78.0

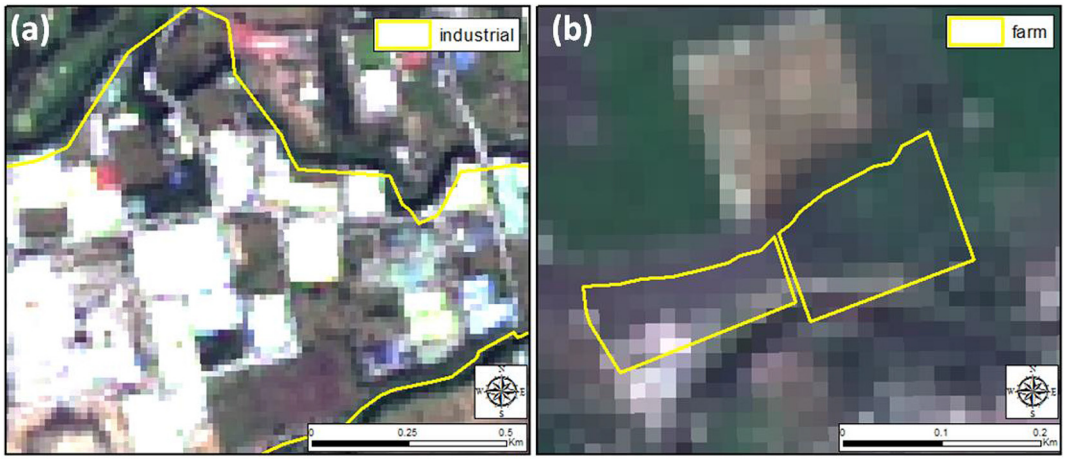

Fig. 3. Examples of class noise in the OSM datasets. (a) Class noise for "impervious" LULC class due to our conversion of an OSM-LU class ("industrial") to a LULC class ("impervious"), as other LULC types (mainly "grass" here) are mislabeled as "impervious". (b) Class noise for "farm" LULC class due to innacurate field delineations in the OSM-LU dataset. Basemap is a Landsat image (natural color composite) acquired 02/02/2014. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Classification was done using (a) the original imbalanced training dataset, and (b) a balanced training dataset in which all classes had the same number of training samples as the majority "forest" class. The synthetic minority over-sampling technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), which introduces artificial training samples into the feature space between existing samples in the minority classes, was used to generate the balanced training set in which all classes had the same number of training samples (n = 7431). SMOTE was chosen for this over-sampling because it was found to increase classification accuracy in past remote sensing studies (Johnson, Tateishi, & Hoan, 2013; Pozi, Sulaiman, Mustapha, & Perumal, 2015) and because it was available in the Weka software package.

For our five- and four-class LULC systems, classification was done using all six LULC classes, and then LULC classes ("forest" and "orchard", "farm" and "grass") were merged after classification. The general workflow of the study is shown in Fig. 2, although some of the optional steps (SMOTE and post-classification class merging) were excluded for illustration simplicity.

3. Results and discussion

3.1. Estimating class noise in the OSM-derived training data

The OA of the OSM-derived training data in our study area was 78.0%, as shown in Table 3, so the level of overall class noise in the training dataset can be estimated as 22.0%. Class noise varied quite widely by LULC class, with the "water" and "farm" classes having relatively low levels of noise (PA and UA values ~90% or higher), the "impervious" and "orchard" classes having moderate noise levels (PA and UA values ~80% or higher), and the "forest" and "grass" classes having the highest class noise levels and often being confused with one another. This high level of confusion between the "forest" and "grass" classes in the OSM datasets is not entirely surprising, as forests often contain some open spaces and grassy areas/meadows often contain some areas with tree cover. In heterogeneous areas with a mixture of forest and open grassy areas, it is likely that the volunteers digitizing the OSM data didn't view it as necessary (or worth the time/effort required) to delineate every boundary between these two LULC types. Our conversion of OSM classes to LULC classes was another cause of class noise, as although the conversion process was done based on the dominant LULC type for each OSM class (e.g. "impervious" is the dominant LULC for "industrial" OSM class), there is more than one LULC type that can be associated with some OSM classes (e.g. "industrial" OSM polygons can also contain some vegetation cover, as shown in Fig. 3 (a)). In general, we can expect that LULC classes with higher levels of class noise should be more difficult to classify accurately. Some visual examples of different types of class noise in the OSM datasets are shown in Fig. 3.

3.2. Image classification results

As shown in Table 4, for the six-class LULC system, SMOTE-NB had the most accurate image classification results (OA = 72.0%), followed by SMOTE-RF (OA = 70.6%), while for the five- and four-class LULC systems, SMOTE-RF produced the most accurate classifications (OA = 81.0% and OA = 84.0%, respectively), followed by NB (OA = 79.7% and OA = 81.3%, respectively). For all LULC systems,

Table 4
OA values of all classifications. Gray cells indicate the most accurate classification result.

Classification algorithm	OA (six-class system)	OA (five-class system)	OA (four-class system)
NB	68.3%	79.7%	81.3%
C4.5	49.0%	63.0%	66.0%
RF	63.3%	78.0%	80.3%
SMOTE-NB	72.0%	78.3%	80.0%
SMOTE-C4.5	60.7%	69.0%	71.3%
SMOTE-RF	70.6%	81.0%	84.0%

C4.5 gave the lowest OA values, indicating that it was not able to handle the class noise and attribute noise as well as NB and RF. As shown in Table 4, in most cases SMOTE led to higher classification accuracies, indicating that SMOTE could help deal with the noisy imbalanced training data, particularly for the DT-based algorithms (C4.5 and RF). Although previous remote sensing research also showed SMOTE could mitigate the negative effects of class imbalance (Johnson et al., 2013; Pozi et al., 2015), it was uncertain whether it could mitigate these effects when class noise existed. So, while SMOTE artificially creates additional noisy training samples when class noise is present, this generation of training samples can still improve classification results.

Table 5 shows the error matrices for the most accurate NB and RF classifications to allow for a comparison of their results for different LULC classes (C4.5 results not shown since its accuracies were always the lowest). In terms of the relationship between class noise and class-specific accuracies, there seemed to be some correlation, particularly for the RF results, as the classes with the most class noise ("forest" and "grass") had some of the lowest PA and UA values and were often confused with one another. The accuracies for the "orchard" class were also quite low, although this may have been due more to its spectral similarity to the "forest" class rather than its class noise, as orchards and forests can even be difficult to separate using multi-temporal Landsat imagery and high-quality training data (Oetter et al., 2001). While the class-specific accuracies of the NB and RF classifications were generally quite similar, there were some differences for the "impervious", "farm", "orchard", and "forest" classes. RF achieved higher accuracies for the "impervious" and "farm" classes, probably because NB assumes the spectral values of each LULC class follow a Gaussian distribution (John & Langley, 1995), and impervious surfaces (Weng, 2012) and mixed croplands do not typically fit this assumption due to their high spectral heterogeneity. NB achieved higher class-specific accuracies than RF mainly for the "orchard" and "forest" classes, possibly because the spectral values of these classes better fit a Gaussian distribution.

In terms of the classification accuracies that could be achieved for the different LULC classification systems, none of the methods evaluated in this study were very successful for LULC mapping under the six- and five-class systems due to the high level of confusion between the "orchard" and "forest" classes in the six-class system and the confusion between the "grass" and "farm" classes in the five-class system. However, for the four-class classification system, NB, SMOTE-NB, RF, and SMOTE-RF all achieved at least moderate OA, PA, and UA values (~80% or better). Figs. 4 and 5 show the most accurate six- and four-class LULC maps, respectively, produced by each algorithm for a subset of the study area to allow for a visual comparison. The lower accuracy and noisier result produced by SMOTE-C4.5 is clear in the Figs. 4 and 5 maps, as well as NB's greater confusion between the "impervious" and "other vegetation" classes compared to SMOTE-RF.

Based on our results, the classification methods presented in this study are recommended mainly for mapping relatively broad LULC classes similar to the ones in our four-class LULC classification

Table 5
OA values of the most accurate six-class (a)- (b), five-class (c)- (d), and four-class (e)- (f) classifications for NB and RF. Legend: I, "impervious"; FO, "forest"; O, "orchard"; FA, "farm"; G, "grass"; W, "water"; T, "tree"; V, "other vegetation".

(a) SMOTE-NB (six-class)

		True LULC							
		I	FO	O	FA	G	W	Sum	UA
Classified	I	29	1	0	0	6	4	40	72.5
	FO	0	60	5	5	8	0	78	76.9
	O	0	14	31	2	0	0	47	66.0
	FA	1	6	3	40	3	0	53	75.5
	G	3	9	2	2	19	1	36	52.8
	W	3	2	0	2	2	37	46	80.4
	Sum	36	92	41	51	38	42	300	
	PA (%)	80.6	65.2	75.6	78.4	50.0	88.1		
	OA (%)								72.0

(b) SMOTE-RF (six-class)

		True LULC							
		I	FO	O	FA	G	W	Sum	UA (%)
Classified	I	38	1	0	1	0	0	40	95.0
	FO	1	61	0	5	11	0	78	78.2
	O	0	31	12	4	0	0	47	25.5
	FA	0	5	0	48	0	0	53	90.6
	G	5	5	2	9	15	0	36	41.7
	W	3	2	0	2	1	38	46	82.6
	Sum	47	105	14	69	27	38	300	
	PA (%)	80.9	58.1	85.7	69.6	55.6	100		
	OA (%)								70.7

(c) NB (five-class)

		True LULC						
		I	T	FA	G	W	Sum	UA
Classified	I	33	1	1	5	0	40	82.5
	T	0	112	9	4	0	125	89.6
	FA	2	10	40	1	0	53	75.5
	G	3	11	4	17	1	36	47.2
	W	3	2	2	2	37	46	80.4
	Sum	41	136	56	29	38	300	
	PA (%)	80.5	82.4	71.4	58.6	97.4		
	OA (%)							79.7

(d) SMOTE-RF (five-class)

		True LULC						
		I	T	FA	G	W	Sum	UA (%)
Classified	I	38	1	1	0	0	40	95.0
	T	1	104	9	11	0	125	83.2
	FA	0	5	48	0	0	53	90.6
	G	5	7	9	15	0	36	41.7
	W	3	2	2	1	38	46	82.6
	Sum	47	119	69	27	38	300	
	PA (%)	80.9	87.4	69.6	55.6	100		
	OA (%)							81.0

(e) NB (four-class)

		True LULC					
		I	T	V	W	Sum	UA (%)
Classified	I	33	1	6	0	40	82.5
	T	0	112	13	0	125	89.6
	V	5	21	62	1	89	69.7
	W	3	2	4	37	46	80.4
	Sum	41	136	85	38	300	
	PA (%)	80.5	82.4	72.9	97.4		
	OA (%)						81.3

(f) SMOTE-RF (four-class)

		True LULC					
		I	T	V	W	Sum	UA (%)
Classified	I	38	1	1	0	40	95.0
	T	1	104	20	0	125	83.2
	V	5	12	72	0	89	80.9
	W	3	2	3	38	46	82.6
	Sum	47	119	96	38	300	
	PA (%)	80.9	87.4	75.0	100		
	OA (%)						84.0

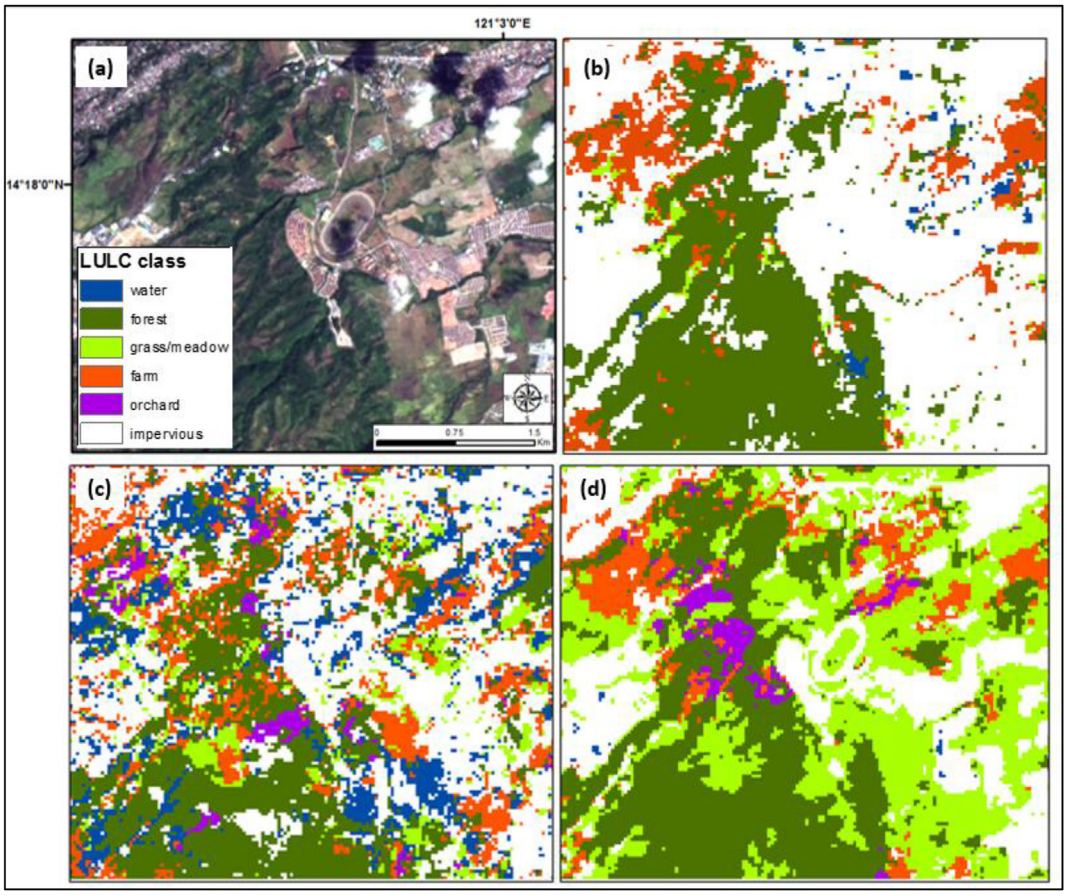

Fig. 4. Landsat image (natural color composite) acquired 02/02/2014 (a), and the six-class LULC maps produced by NB (b), SMOTE-C4.5 (c), and SMOTE-RF (d). Note: The area in this Fig. is a subset of the entire study area shown in Fig. 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

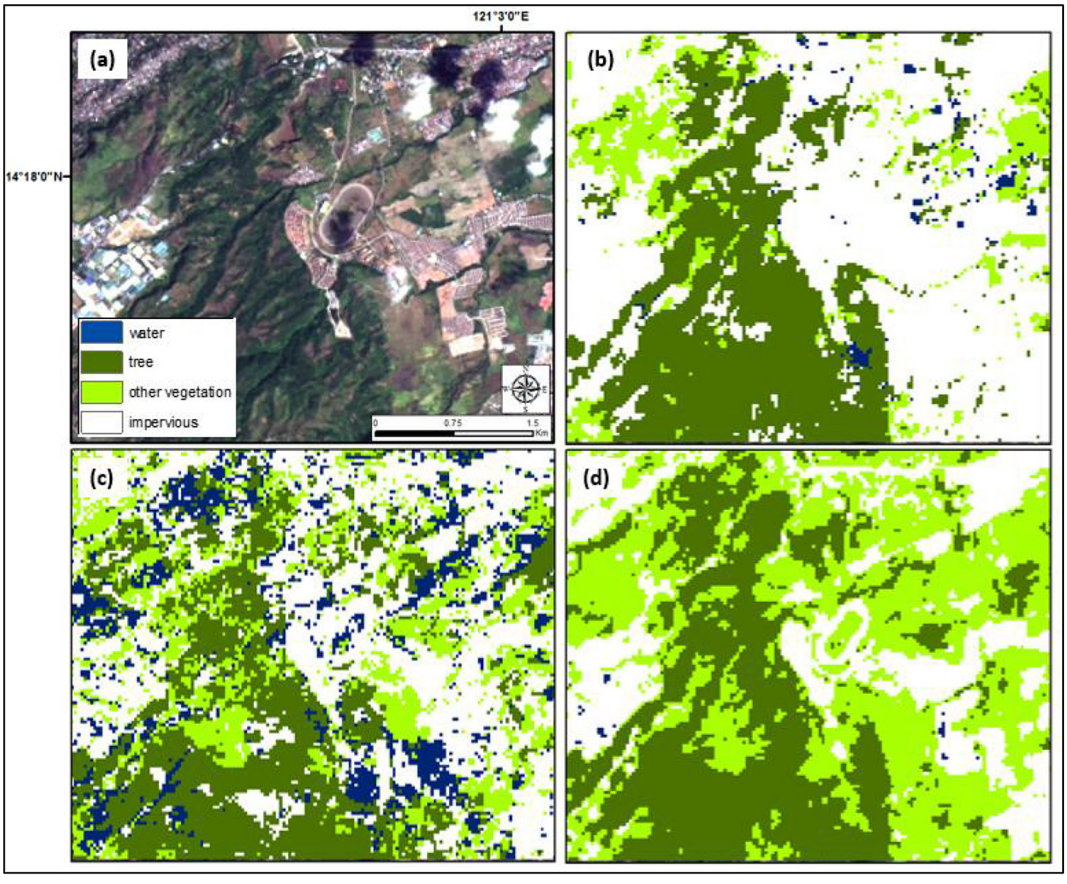

Fig. 5. Landsat image (natural color composite) acquired 02/02/2014 (a), and the four-class LULC maps produced by NB (b), SMOTE-C4.5 (c), and SMOTE-RF (d). Note: The area in this Fig. is a subset of the entire study area shown in Fig. 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

system. Of course, the usefulness of the LULC data for some applications (e.g. forest or crop monitoring) is limited when only a few broad LULC classes are mapped, but this can potentially be overcome if an older but more detailed LULC map is available. For example, a current tree cover map can be used for monitoring forest change over time if a historical map of forest boundaries exists (Johnson, 2015; Margono, Potapov, Turubanova, Stolle, & Hansen, 2014). Since NB and RF were relatively tolerant to class noise, it may even be possible to use the OSM-derived training data (representing the current LULC) to generate historical LULC maps for LULC change monitoring, despite some LULC changes occurring within the OSM polygons, and this could be an interesting direction for future research.

4. Conclusions

We evaluated the potential for rapid and automated land use/land cover (LULC) map production using Landsat time-series NDVI imagery and training data derived from OpenStreetMap (OSM) "landuse" and "natural" polygons. The Landsat data contained attribute noise due to cloud cover in all images, while the OSM data contained class noise due to inaccurate polygon boundary delineations and because the OSM classes did not exactly match our the LULC classes they were used to map. Extracting training data from the OSM data also led to a class imbalance in the training set since some OSM classes had much greater area coverage than others.

Naïve bayes (NB), decision tree (C4.5), and random forest (RF) classification algorithms were tested for image classification, and the synthetic minority over-sampling technique (SMOTE) was applied to the training data prior to classification in an attempt to reduce the effects of class imbalance. Classification accuracy was calculated for six-, five-, and four-class classification systems. SMOTE-NB achieved the highest overall accuracy (OA) for the six-class classification (OA = 72.0%), followed by SMOTE-RF (OA = 70.7%), while SMOTE-RF produced the highest OA for the five-class (OA = 81.0%) and four-class (OA = 84.0%) classifications, followed by NB (OA = 79.7% and OA = 81.3%, respectively). C4.5 had the lowest OA values for all classification systems.

This study showed that noise-tolerant classification algorithms like NB and RF have the potential to use OSM-derived training data and cloud-contaminated time-series imagery for rapid and automated mapping of relatively broad LULC classes (e.g. "impervious", tree", "other vegetation", and "water"). Some future research avenues to build on this study include: (i) evaluating the performance of other noise-tolerant classification algorithms, (ii) incorporating other image datasets containing less cloud contamination, e.g. Sentinel-1 synthetic aperture radar data, (iii) evaluating automated filtering approaches to remove mislabeled training samples from the training dataset (Brodley & Friedl, 1999; Frenay & Verleysen, 2013; Guan & Yuan, 2013). To support future research efforts and in the spirit of open data, the training and validation data used in this study can be downloaded at http://pub.iges.or.jp/modules/envirolib/view.php?docid=6201.

Acknowledgments

This paper is generally based upon outputs produced through a project on integrating climate change mitigation and adaptation, under the "Climate Change Resilient Low Carbon Society Network (CCR-LCSNet)" for the fiscal year 2015, commissioned work of the Japanese Ministry of the Environment.

References

Belgiu, M., & Drăguţ, L. (2014). Comparing supervised and unsupervised multi-resolution segmentation approaches for extracting buildings from very high resolution imagery. *ISPRS Journal of Photogrammetry and Remote Sensing, 96*, 67–75.

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

Brodley, C. E., & Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research, 11*, 131–167.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 341–378.

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning, 3*(4), 261–283.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment, 37*(1), 35–46.

Cracknell, M. J., & Reading, A. M. (2014). Geological mapping using remote sensing data: a comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences, 63*, 22–33.

Dewan, A. M., & Yamaguchi, Y. (2009). Land use and land cover change in Greater Dhaka, Bangladesh: using remote sensing to promote sustainable urbanization. *Applied Geography, 29*(3), 390–401.

Estima, J., & Painho, M. (2013). Exploratory analysis of OpenStreetMap for land use classification. *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, 39–46.

Estima, J., & Painho, M. (2015). Investigating the potential of OpenStreetMap for land Use/Land cover production: a case study for Continental Portugal. In J. J. Arsanjani, A. Zipf, P. Mooney, & M. Helbich (Eds.), *OpenStreetMap in GIScience* (pp. 273–293). Springer International Publishing.

Fisher, P. F., & Langford, M. (1996). Modeling sensitivity to accuracy in classified imagery: a study of areal interpolation by dasymetric mapping. *The Professional Geographer, 48*(3), 299–309.

Folleco, A., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2009). Identifying learners Robust to low quality data. *Informatica, 33*, 245–259.

Food and Agriculture Organization of the United Nations. (2010). *Global forest resources assessment 2010: Main report*.

Foody, G. M., Boyd, D. S., & Cutler, M. E. J. (2003). Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions. *Remote Sensing of Environment, 85*(4), 463–474.

Frenay, B., & Verleysen, M. (2013). Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems, 25*(5), 845–869.

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random forests for land cover classification. *Pattern Recognition Letters, 27*(4), 294–300.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal, 69*(4), 211–221.

Guan, D., & Yuan, W. (2013). A survey of mislabeled training data detection techniques for pattern classification. *IETE Technical Review, 30*(6), 524–530.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design, 37*(4), *682–703*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations, 11*(1), 10–18.

Hoan, N. T., Tateishi, R., Alsaaideh, B., Ngigi, T., Alimuddin, I., & Johnson, B. (2013). Tropical forest mapping using a combination of optical and microwave data of ALOS. *International Journal of Remote Sensing, 34*(1), 139–153.

Huang, C., Davis, L. S., & Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing, 23*(4), 725–749.

Hughes, G. P. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory, 14*(1), 55–63.

Jensen, J. R. (2005). *Introductory digital image processing - A remote sensing perspective* (3rd ed.). Upper Saddle River, USA: Pearson Prentice Hall.

Jia, K., Liang, S., Wei, X., Yao, Y., Su, Y., Jiang, B., et al. (2014). Land cover classification of landsat data with phenological features extracted from time series MODIS NDVI data. *Remote Sensing, 6*(11), 11518–11532.

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence* (pp. 338–345). Morgan Kaufmann Publishers Inc.

Johnson, B. A. (2015). Combining national forest type maps with annual global tree cover maps to better understand forest change over time: case study for Thailand. *Applied Geography, 62*, 294–300.

Johnson, B. A., Bragais, M., Endo, I., Magcale-Macandog, D. B., & Macandog, P. B. M. (2015). Image segmentation parameter optimization considering within- and between-segment heterogeneity at multiple scale levels: test case for mapping residential areas using landsat imagery. *ISPRS International Journal of Geo-Information, 4*(4), 2292–2305.

Johnson, B. A., Tateishi, R., & Hoan, N. T. (2013). A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees. *International Journal of Remote Sensing, 34*(20), 6969–6982.

Johnson, B. A., & Xie, Z. (2013). Classifying a high resolution image of an urban area using super-object information. *ISPRS Journal of Photogrammetry and Remote Sensing, 83*, 40–49.

Jokar Arsanjani, J., Helbich, M., & Bakillah, M. (2013). Exploiting volunteered geographic information to ease land use mapping of an urban landscape. In *International archives of the photogrammetry, remote sensing and spatial information sciences* (Vol. XL–4/W1, pp. 51–55).

Jokar Arsanjani, J., Mooney, P., & Zipf, A. (2015). Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets. In J. J. Arsanjani, A. Zipf, P. Mooney, & M. Helbich (Eds.), *OpenStreetMap in GIScience* (pp. 37–58). Springer International Publishing.

Lawrence, R., Wood, S., & Sheley, R. (2006). Mapping invasive plants using hyperspectral imagery and breiman cutler classifications (RandomForest). *Remote Sensing of Environment, 100*(3), 356–362.

Lippitt, C., Rogan, J., & Li, Z. (2008). Mapping selective logging in mixed deciduous forest: a comparison of machine learning algorithms. *Engineering & Remote, 74*(10), 1201–1211.

Margono, B. A., Potapov, P. V., Turubanova, S., Stolle, F., & Hansen, M. C. (2014). Primary forest cover loss in Indonesia over 2000–2012. *Nature Climate Change, 4*, 730–735.

Neis, P., & Zielstra, D. (2014). Recent developments and future trends in volunteered geographic information research: the case of OpenStreetMap. *Future Internet, 6*(1), 76–106.

Neis, P., & Zipf, A. (2012). Analyzing the contributor activity of a volunteered geographic information project — The case of OpenStreetMap. *ISPRS International Journal of Geo-Information, 1*(2), 146–165.

Nettleton, D. F., Orriols-Puig, A., & Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review, 33*(4), 275–306.

Oetter, D. R., Cohen, W. B., Berterretche, M., Maiersperger, T. K., & Kennedy, R. E. (2001). Land cover mapping in an agricultural setting using multiseasonal Thematic Mapper data. *Remote Sensing of Environment, 76*(2), 139–155.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing, 26*(1), 217–222.

Pozi, M. S. M., Sulaiman, M. N., Mustapha, N., & Perumal, T. (2015). A new classification model for a class imbalanced data set using genetic programming and support vector machines: case study for wilt disease classification. *Remote Sensing Letters, 6*(7), 568–577.

Quinlan, J. R. (1993). *C4.5 programs for machine learning*. San Mateo: Morgan Kaufmann.

Reiche, J., Souzax, C. M., Hoekman, D. H., Verbesselt, J., Persaud, H., & Herold, M. (2013). Feature level fusion of multi-temporal ALOS PALSAR and Landsat data for mapping and monitoring of tropical deforestation and forest degradation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6*(5), 2159–2173.

Roy, P. S., & Tomar, S. (2000). Biodiversity characterization at landscape level using geospatial modelling technique. *Biological Conservation, 95*(1), 95–109.

Schneider, A. (2012). Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach. *Remote Sensing of Environment, 124*, 689–704.

Shimada, M., Itoh, T., Motooka, T., Watanabe, M., Shiraishi, T., Thapa, R., et al. (2014). New global forest/non-forest maps from ALOS PALSAR data (2007-2010). *Remote Sensing of Environment, 155*, 13–31.

Tongson, E. E., & Faraon, A. A. (2012). *Hydrologic Atlas of Laguna de Bay 2012*. Quezon city, Philippines: Laguna Lake Development Authority and WWF-Philippines.

Wang, J., Zhao, Y., Li, C., Yu, L., Liu, D., & Gong, P. (2014). Mapping global land cover in 2001 and 2010 with spatial-temporal consistency at 250m resolution. *ISPRS Journal of Photogrammetry and Remote Sensing, 103*, 38–47.

Weng, Q. (2012). Remote sensing of impervious surfaces in the urban areas: requirements, methods, and trends. *Remote Sensing of Environment, 117*, 34–49.

Yu, Q., Gong, P., Clinton, N., Biging, G., Kelly, M., & Schirokauer, D. (2006). Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery. *Photogrammetric Engineering & Remote Sensing, 72*(7), 799–811.

Zhu, X., Gao, F., Liu, D., & Chen, J. (2012). A modified neighborhood similar pixel interpolator approach for removing thick clouds in landsat images. *IEEE Geoscience and Remote Sensing Letters, 9*(3), 521–525.

Zhu, Z., & Woodcock, C. E. (2012). Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sensing of Environment, 118*(July 2015), 83–94.