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 14 January 2016

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/© 2015 Elsevier Ltd. All rights reserved.

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