Remote Sensing of Environment 162 (2015) 67–83

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Generating synthetic Landsat images based on all available Landsat data: Predicting Landsat surface reflectance at any given time

Zhe Zhu [a,*], Curtis E. Woodcock [a], Christopher Holden [a], Zhiqiang Yang [b]

[a] Center for Remote Sensing, Department of Earth and Environment, Boston University, 685 Commonwealth Avenue, Boston, MA 02215, United States

[b] Department of Forest Ecosystems and Society, Oregon State University, 321 Richardson Hall, Corvallis, OR 97331, United States

A R T I C L E I N F O

Article history:
Received 5 October 2014
Received in revised form 8 January 2015
Accepted 7 February 2015
Available online 5 March 2015

Keywords:
Synthetic
Landsat
Surface reflectance
Time series model
Predict

A B S T R A C T

A new algorithm for generating synthetic Landsat images is developed based on all available Landsat data. This algorithm is capable of predicting Landsat surface reflectance for any desired date. It first excludes cloud, cloud shadow, and snow observations, and then uses the remaining clear observations to estimate time series models for each Landsat pixel. Three time series models (a simple model, advanced model, and full model) are used for estimating surface reflectance for each pixel, and the selection of a time series model is dependent on the number of clear observations available: the more clear observations, the more complex the model will be that is used. For each time series model there are three components (seasonality, trend, and breaks), that are used for modeling intra-annual and inter-annual differences and abrupt surface change. Abrupt surface changes are detected by differencing predicted and observed Landsat observations, and if the difference is larger than twice the Root Mean Square Error (RMSE) for six consecutive observations, it will be detected as a "break" in the time series model. The RMSE values are temporally adjusted to provide better threshold range. For each "synthetic" image, a Quality Assessment (QA) Band is provided that contains information on how the time series model was estimated and used for generating the synthetic data. We have applied this approach to six Landsat scenes within the United States. We visually compared the synthetic images with real Landsat images for different kinds of environments and they are similar for all image pairs. We also quantitatively assessed the accuracy of the synthetic data by calculating the RMSE value for all clear Landsat observations. The RMSE values for the three visible bands are the lowest (approximately 0.01), and the Short-wave Infrared (SWIR) bands are slightly higher in magnitude (between 0.01 and 0.02). The Near Infrared (NIR) band has the highest RMSE values (between 0.02 and 0.03). The goal of this paper is to provide Landsat images that are free of cloud, cloud shadow, snow, and Scan Line Corrector (SLC)-off gaps that can be used to derive land cover and bio-physical products.

1. Introduction

The Landsat satellite series provides the longest record of earth observations (Williams, Goward, & Arvidson, 2006). Due to its relatively high spatial resolution, accurate radiometric calibration, and high geometric precision, it has been widely used in many aspects of remote sensing activities. In January 2008, U.S. Geological Survey (USGS) started to provide Landsat data at no cost via the Internet, which makes Landsat data even more popular (Woodcock et al., 2008; Wulder, Masek, Cohen, Loveland, & Woodcock, 2012). With the freely available Landsat data, it is now possible to reconstruct the history of the Earth's surface back to 1972 (Pflugmacher, Cohen, & Kennedy, 2012).

Despite all these advantages, Landsat data also have limitations and issues. The most significant limitation is its relatively low temporal frequency (16 day revisit capability). For each Landsat sensor, if every overpass is acquired, only 22 or 23 acquisitions per year per Path/Row are collected (Ju & Roy, 2008). Moreover, due to the limited duty cycles, the lack of on-board data recording capabilities, and the constraints of international ground stations, the Landsat project does not acquire every acquisition globally (Arvidson, Goward, Gasch, & Williams, 2006). Additionally, the presence of cloud, cloud shadow, and snow further reduce the number of available clear Landsat observations (hereafter "clear" refers to observations that are free of cloud, cloud shadow, and snow). For example, the annual mean cloud cover for all

Table 1
The six optical Landsat TM/ETM+ spectral bands used for generating synthetic Landsat images.

TM bands (μm)	ETM+ bands (μm)
Band 1 (0.45–0.52)	Band 1 (0.45–0.515)
Band 2 (0.52–0.60)	Band 2 (0.525–0.605)
Band 3 (0.63–0.69)	Band 3 (0.63–0.69)
Band 4 (0.76–0.90)	Band 4 (0.75–0.90)
Band 5 (1.55–1.75)	Band 5 (1.55–1.75)
Band 7 (2.08–2.35)	Band 7 (2.09–2.35)

* Corresponding author. Tel.: +1 617 233 6031.
E-mail address: zhuzhe@bu.edu (Z. Zhu).

http://dx.doi.org/10.1016/j.rse.2015.02.009
0034-4257/

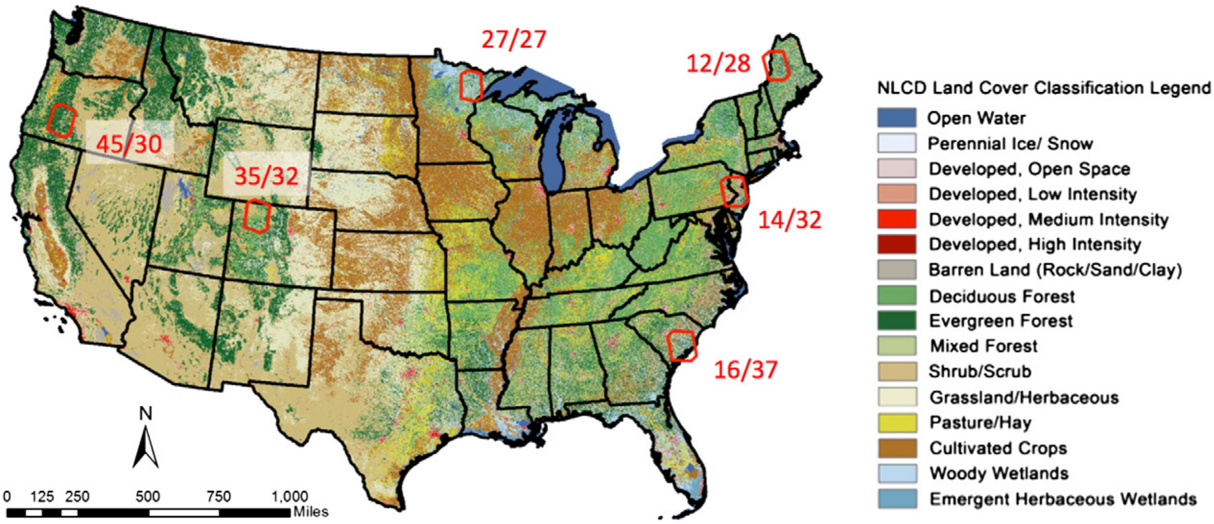

Fig. 1. Six study sites (red polygons are Landsat scenes by Path/Row), shown in the context of the 2006 National Land Cover Database (NLCD) cover map (Fry et al., 2011).

Landsat Enhanced Thematic Mapper Plus (ETM+) images stored in the U.S. Landsat archive is approximately 35% (Ju & Roy, 2008). What is more, the failure of the ETM+ Scan Line Corrector (SLC) that occurred in May 2003 reduces the total usable data in each Landsat ETM+ image by 22% (Maxwell, Schmidt, & Storey, 2007). Therefore, it is very difficult to find entire Landsat images that are free of cloud, cloud shadow, snow, and without SLC-off artifacts for a specified time period.

Image compositing has been shown to be a powerful tool for generating clear satellite images. There are many compositing methods available (Cihlar, Manak, & D'Iorio, 1994; Griffiths, van der Linden, Kuemmerle, & Hostert, 2013; Hansen et al., 2008; Holben, 1986; Luo, Trishchenko, & Khlopenkov, 2008; Roy et al., 2010; Stoms, Bueno, & Davis, 1997; White et al., 2014), either based on single criteria (e.g. maximum NDVI, minimum red Band, or maximum brightness temperature), or multiple criteria to select the "best" observation with minimum cloud, cloud shadow, and snow contamination. However, most of the existing image compositing methodologies are designed for satellite data with high temporal frequency, such as Moderate Resolution Imaging Spectroradiometer (MODIS) and Advanced Very High Resolution Radiometer (AVHRR), and only a few studies are applied for satellite data with relatively low temporal frequency like Landsat data (Griffiths et al., 2013; Hansen et al., 2008; Hermosilla, Wulder, White, Coops, & Hobart, 2015; Roy et al., 2010; White et al., in press). Due to the lack of frequent observations, it may take a few months or even years to provide

Table 2
The acquisition date for the first and the last Landsat images (month/day/year) and total number of Landsat images used to generate synthetic Landsat image for each Path/Row.

Path/Row	27/27	12/28	45/30	35/32	14/32	16/37
First image	12/08/1982	05/18/1984	06/26/1984	04/17/1984	11/27/1982	04/12/1984
Last image	11/08/2012	09/28/2012	11/22/2012	05/27/2013	06/25/2013	07/25/2013
# of images	312	257	486	447	482	617

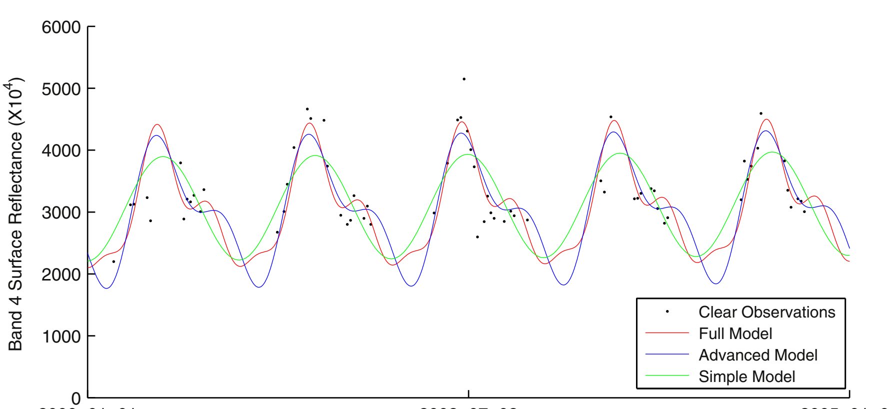

Fig. 2. Time series models estimated for Band 4 surface reflectance using all available Landsat observations between 2001 and 2004 for a crop pixel. The black points are the all available clear Landsat observations. The green line is the Band 4 surface reflectance estimated by the simple model. The blue line is the Band 4 surface reflectance estimated by the advanced model. The red line is the Band 4 surface reflectance estimated by the full model. Note that the more complex the time series model is, the better the performance in modeling the intra-annual differences in the time series data.