Remote Sensing of Environment 165 (2015) 234–248

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

A global reference database from very high resolution commercial satellite data and methodology for application to Landsat derived 30 m continuous field tree cover data

Bruce Pengra [a,*], Jordan Long [b], Devendra Dahal [a], Stephen V. Stehman [c], Thomas R. Loveland [d]

[a] SGT, Inc., Contractor to USGS EROS Center, 47914 252nd Street, Sioux Falls, SD 57198 USA
[b] Inuteq, Contractor to USGS EROS Center, 47914 252nd Street, Sioux Falls, SD 57198, USA
[c] College of Environmental Science and Forestry, State University of New York, Syracuse, NY 13210, USA
[d] U.S. Geological Survey, EROS Data Center, Sioux Falls, SD 57198, USA

ARTICLE INFO

Article history:
Received 7 August 2014
Received in revised form 22 January 2015
Accepted 24 January 2015
Available online 26 February 2015

Keywords:
Validation database
Global land cover
High resolution satellite
Continuous field data
Landsat
Tree cover
Accuracy assessment

ABSTRACT

The methodology for selection, creation, and application of a global remote sensing validation dataset using high resolution commercial satellite data is presented. High resolution data are obtained for a stratified random sample of 500 primary sampling units (5 km × 5 km sample blocks), where the stratification based on Köppen climate classes is used to distribute the sample globally among biomes. The high resolution data are classified to categorical land cover maps using an analyst mediated classification workflow. Our initial application of these data is to evaluate a global 30 m Landsat-derived, continuous field tree cover product. For this application, the categorical reference classification produced at 2 m resolution is converted to percent tree cover per 30 m pixel (secondary sampling unit)for comparison to Landsat-derived estimates of tree cover. We provide example results (based on a subsample of 25 sample blocks in South America) illustrating basic analyses of agreement that can be produced from these reference data. Commercial high resolution data availability and data quality are shown to provide a viable means of validating continuous field tree cover. When completed, the reference classifications for the full sample of 500 blocks will be released for public use.

1. Introduction

Land-cover products are increasingly valuable inputs to a range of scientific studies and resource management activities including climate change, food security, fire science, habitat loss and population distribution (Foley et al., 2005; Hansen & Loveland, 2012; Ramankutty & Foley, 1999; Strahler et al., 2006; Tsendbazar, de Bruin, & Herold, in press). As an Essential Climate Variable (ECV), high quality land-cover data at a global scale are crucial for reducing uncertainties in our understanding of the global climate system (Bontemps, Defourny, Radoux, Kalogirou, & Arino, 2012). New global land-cover maps hold the promise of making important contributions to our understanding of global processes, so it is critical to obtain information on the degree and structure of error in these products. While validation is widely understood to be a crucial element of producing useful land-cover products (Olofsson et al., 2012; Strahler et al., 2006), the challenges, including a very large commitment of resources and time, have generally discouraged thorough validation efforts (Herold, Latham, Di Gregorio and Schmullius, 2006).

In this article, we describe the validation methodology used for producing reference data for a global sample of 500 sample blocks (5 km × 5 km) selected by stratified random sampling (Olofsson et al., 2012; Stehman, Olofsson, Woodcock, Herold, & Friedl, 2012). These reference data are thematically classified, commercial high-resolution satellite data for each of the 500 sample blocks. High resolution satellite data provide the only practical source of reference data for this global 500-sample-block dataset (Strahler et al., 2006). Such use of relatively finer resolution satellite data as the basis of or as a component of global and large area land cover accuracy assessment has been established (Bontemps et al., 2012; De Fries, Hansen, Townshend, & Sohlberg, 1998; Desclee, Simonetti, Mayaux, & Achard, 2013; Mayaux et al., 2006; Montesano et al., 2009; Morisette et al., 2003; Petropoulos, Partsinevelos, & Mitraka, 2013; Raši et al., 2011; Scepan & Estes, 2001; Small & Lu, 2006). Our purpose in this article is to document the identification, collection and preprocessing of reference source-data, the reference map production, and the methods used in applying these high resolution thematic data to assess 30 m continuous field data layers characterizing global tree cover. We illustrate the use of the reference data via an assessment of continuous field tree cover produced by collaborators at the University of Maryland. The product evaluated in the example provides a percent tree cover value (0–100%) at nominal

* Corresponding author. Tel.: +1 6055946265.
E-mail address: bpengra@usgs.gov (B. Pengra).

http://dx.doi.org/10.1016/j.rse.2015.01.018

30 m resolution from a temporally composited global mosaic of Landsat data.

1.1. Sampling design

The sampling design for this study was constructed to provide the foundation for developing a global validation database that could be used to assess multiple land cover products (Olofsson et al. (2012). A global sampling frame (i.e., the "population") was created by partitioning the earth's land surface into 5 km × 5 km grid cells (the "primary sampling units" of the design of the cluster sampling design). A stratified random sample of 500 validation sample blocks was selected. This design satisfies the criteria that define a probability sampling design and thus serves as the basis to support rigorous design-based statistical inference (Stehman, 2000). Further, the stratified design distributes the sample globally across a broad range of biomes. A total of 21 strata were defined based on modified Köppen Climate/Vegetation classification and population density (Olofsson et al., 2012). The 5 km × 5 km sample blocks are clusters (or primary sampling units) and the 30 m pixels within each cluster represent secondary sampling units. Therefore, the sampling design is a stratified one-stage cluster sample, where one-stage cluster sampling is defined as sampling all pixels within each 5 km × 5 km sample block (the terms "block" and "cluster" will be used interchangeably). Identifying the design as one-stage cluster sampling is important for estimating standard errors. Global estimates of agreement may be obtained from these 500 sample blocks. Moreover, the sampling design was constructed so that the base sample of 500 sample blocks can be augmented by additional sample

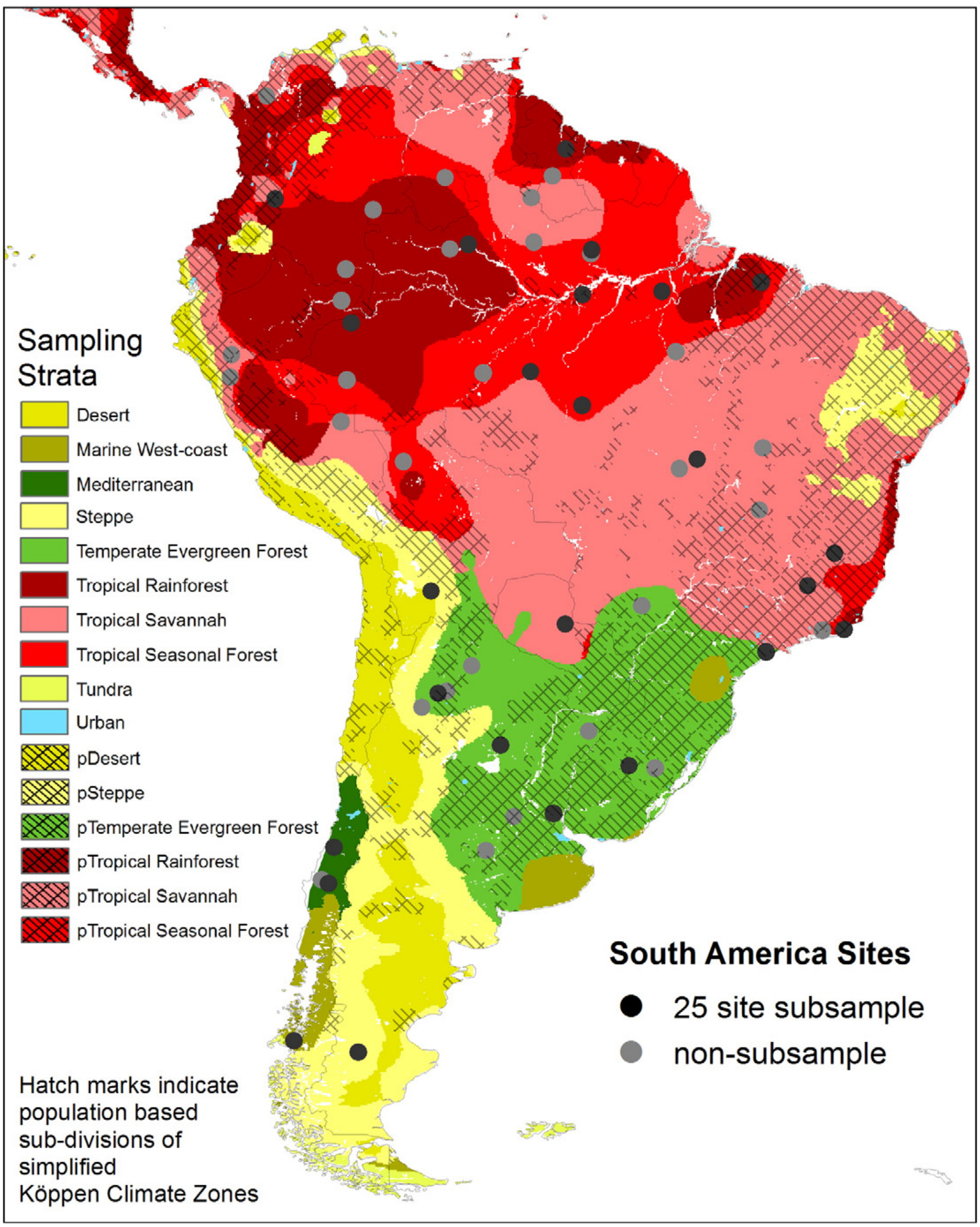

Fig. 1. The random subsample of 25 sample blocks (black dots) selected from the global sample blocks within South America. The basemap shows the modified Köppen zones with hatching indicating where those zones were sub-divided by population density, into the final strata.