Remote Sensing of Environment 190 (2017) 383–395

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

National-scale soybean mapping and area estimation in the United States using medium resolution satellite imagery and field survey

Xiao-Peng Song [a,*], Peter V. Potapov [a], Alexander Krylov [a], LeeAnn King [a], Carlos M. Di Bella [b,c,d], Amy Hudson [a], Ahmad Khan [a], Bernard Adusei [a], Stephen V. Stehman [e], Matthew C. Hansen [a]

[a] Department of Geographical Sciences, University of Maryland, College Park, MD 20742, USA
[b] INTA-Castelar, CNIA, Instituto de Clima y Agua, Las Cabañas y Los Reseros S/N, Hurlingham 1686, Buenos Aires, Argentina
[c] CONICET - Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina
[d] Departamento de Métodos Cuantitativos, Facultad de Agronomía, Universidad de Buenos Aires, Argentina
[e] College of Environmental Science and Forestry, State University of New York, Syracuse, NY 13210, USA

ARTICLE INFO

Article history:
Received 24 May 2016
Received in revised form 9 October 2016
Accepted 8 January 2017
Available online xxxx

Keywords:
Agriculture
Cropland
Sample
Remote sensing
Landsat
Image time-series
Classification
Decision tree

ABSTRACT

Reliable and timely information on agricultural production is essential for ensuring world food security. Freely available medium-resolution satellite data (e.g. Landsat, Sentinel) offer the possibility of improved global agriculture monitoring. Here we develop and test a method for estimating in-season crop acreage using a probability sample of field visits and producing wall-to-wall crop type maps at national scales. The method is illustrated for soybean cultivated area in the US for 2015. A stratified, two-stage cluster sampling design was used to collect field data to estimate national soybean area. The field-based estimate employed historical soybean extent maps from the U.S. Department of Agriculture (USDA) Cropland Data Layer to delineate and stratify U.S. soybean growing regions. The estimated 2015 U.S. soybean cultivated area based on the field sample was 341,000 km^2 with a standard error of 23,000 km^2. This result is 1.0% lower than USDA's 2015 June survey estimate and 1.9% higher than USDA's 2016 January estimate. Our area estimate was derived in early September, about 2 months ahead of harvest. To map soybean cover, the Landsat image archive for the year 2015 growing season was processed using an active learning approach. Overall accuracy of the soybean map was 84%. The field-based sample estimated area was then used to calibrate the map such that the soybean acreage of the map derived through pixel counting matched the sample-based area estimate. The strength of the sample-based area estimation lies in the stratified design that takes advantage of the spatially explicit cropland layers to construct the strata. The success of the mapping was built upon an automated system which transforms Landsat images into standardized time-series metrics. The developed method produces reliable and timely information on soybean area in a cost-effective way and could be applied to other regions and potentially other crops in an operational mode.

1. Introduction

Reliable and timely information on agricultural production is essential for ensuring world food security. Traditionally, agricultural data are acquired through census and ground survey. While ground-based data collection has the advantage of obtaining a wide range of variables related to the organizational structure of agriculture, such as land tenure, farm size, labor, crop area, irrigation, and fertilizer use, agricultural censuses are usually undertaken at a decadal frequency and thus they are most suitable to represent those aspects of agriculture that change slowly over time (FAO, 2015). Using census data across the globe would also encounter the data inconsistency problem, including inconsistent definitions of census variables, changing political or sampling units and various reporting protocols among different countries and census intervals (Portmann et al., 2010; Ramankutty et al., 2008).

Satellite observations, owing to their synoptic and repetitive nature, have the unique advantage of providing timely and spatially contiguous information on crop growth at regional to global scales. However, identification of crop type using satellite data remains a technical challenge due to the diversity of cropping systems, including crop types, crop varieties, management practices and field sizes. Thus, global cropland monitoring requires data of high spatial and temporal resolutions (Cihlar, 2000; Fritz et al., 2015; Thenkabail et al., 2010; Waldner et al., 2016). To date, satellite data of fine temporal resolution and coarse spatial resolution are predominantly used in agricultural research, especially over large areas. For example, data from the Moderate Resolution Imaging Spectroradiometer (MODIS) have been extensively used in cropland mapping (e.g. Chang et al., 2007; Lobell and Asner, 2004; Ozdogan, 2010; Wardlow and Egbert, 2008), as well as yield estimation

* Corresponding author.
E-mail address: xpsong@umd.edu (X.-P. Song).

http://dx.doi.org/10.1016/j.rse.2017.01.008

(e.g. Anderson et al., 2016; Becker-Reshef et al., 2010; Bolton and Friedl, 2013; Doraiswamy et al., 2005; Johnson, 2014; Lopresti et al., 2015). At the global scale, cropland is often characterized as one or a few aggregated land cover classes at moderate to coarse resolutions (30 m–1 km) (e.g. Friedl et al., 2002; Gong et al., 2013; Hansen et al., 2000). Crop-specific masks are available at even coarser resolutions (~10 km) (Portmann et al., 2010; Ramankutty et al., 2008; Thenkabail et al., 2009; You et al., 2014). Since the opening of the Landsat archive in year 2008, the recent launch of Landsat 8 and Sentinel-2, medium-resolution data have shown great potential of producing rich temporal information that is previously available only with coarse-resolution data (Hansen et al., 2014; Roy et al., 2010). Studies have begun to take this opportunity in research on crop classification (e.g. Zhong et al., 2014) and yield estimation (e.g. Lobell et al., 2015) albeit over small areas.

The idea of using satellite data in operational agricultural surveys was formulated in the 1970s and exploratory programs were soon initiated, such as the Large Area Crop Inventory Experiment (LACIE) (Macdonald and Hall, 1980), Agriculture and Resources Inventory Surveys through Aerospace Remote Sensing (AgRISTARS) and Monitoring Agriculture with Remote Sensing (MARS) (https://ec.europa.eu/jrc/en/mars). Today, two prominent examples of operational use of medium-resolution satellite data for national-scale crop type mapping are the Cropland Data Layer (CDL) generated by the United States Department of Agriculture (USDA) (Johnson and Mueller, 2010) and the Crop Inventory (CI) dataset generated by the Agriculture and Agri-Food Canada (AAFC). CDL has been produced annually since 2008 and CI has been produced annually since 2011, both with national coverage and at 30–56 m of spatial resolutions. Both CDL and CI are generated using supervised classification approaches and rely on comprehensive survey-based geospatial data for training. The sheer volume of the training data contributes to producing highly accurate maps, both over 85% for major crops in major agricultural states (AAFC, 2015; Boryan et al., 2011). However, not only are these training datasets not publically available, but they are also financially expensive and time-consuming to generate and update every year. Such datasets or the capacity to generate them may not be readily available in other countries, especially developing countries. Therefore, the methodology for producing CDL or CI is difficult to implement elsewhere.

Crop classification maps of high overall accuracy such as CDL and CI cannot be directly used through "pixel counting" for acreage estimation, because map products are usually biased due to misclassification and the existence of mixed pixels (Gallego, 2004). One way of deriving crop acreage and uncertainty estimates is to use a probability sample in an area sampling frame (AFS) (Carfagna and Gallego, 2005; Cotter and Tomczack, 1994; Pradhan, 2001). The classification map can be used as an ancillary variable and survey estimates as the dependent variable to perform regression analysis (Boryan et al., 2011; Gallego et al., 2014; Gonzáles-Alonso and Cuevas, 1993; Gonzáles-Alonso et al., 1991; Hill and Megier, 1988). Map-making and sample-based area estimation are often carried out as independent processes. But opportunities exist to reconcile the discrepancy of map-based and sample-based area estimates by closely integrating the two processes. Such a general approach has been successfully applied for mapping and estimating areas of forest cover change at continental to global scales (Broich et al., 2011; Hansen et al., 2010; Hansen et al., 2008b; Tyukavina et al., 2015).

The objective of this study is to develop a method applicable at national scales for estimating in-season cultivated area for a specific crop as well as to produce a spatially explicit crop cover map. Specifically, we estimate soybean cultivation area in the United States in year 2015 using a probability sample of field visits and we map soybean cover using all available Landsat data in the growing season of the year 2015. Soybean in the U.S. is chosen not only because the U.S. is the world's leading producer of this major commodity crop but also because independent data exist to provide a comparison with our results. The developed procedure is expected to be applicable to other crops and to other regions such as Brazil and Argentina where industrial monoculture dominates agricultural production.

2. Data and methods

2.1. Sample-based soybean area estimation

2.1.1. Study area and sampling design

We implemented a two-stage cluster sampling design for estimating national soybean area within the U.S. In the first stage, 20 km × 20 km blocks (clusters) were selected using a stratified random sampling technique. In the second stage, 30 m × 30 m Landsat pixels were selected using simple random sampling. The large agricultural fields in the U.S.—mean size 0.193 km^2 and median size 0.278 km^2 (Yan and Roy, 2016), ensured that our 20 km × 20 km blocks contained a large number of fields whereas the majority of our 30 m × 30 m pixels were pure pixels of a single crop. The cluster design was chosen to reduce the cost of field visits by spatially constraining the sample pixels to the selected clusters. The specific size of the cluster (20 km × 20 km) was chosen because it allowed the second-stage sample for each cluster to be completed in a single day.

The entire conterminous U.S. land area was divided into a regular grid of 20 km × 20 km blocks consisting of 20,371 blocks. We acquired the 30 m spatial resolution CDL for years between 2010 and 2014 and calculated the 5-year average soybean percentage for every block. For each block in each year, we counted the number of pixels labeled as soybean and divided the number by the total number of pixels within the block to compute the soybean percentage at the 20 km resolution. We then computed the arithmetic mean of soybean percentage over 2010–2014 for every block. After sorting the blocks based on the 2010–2014 mean soybean percentage from the largest to the smallest, the blocks that cumulatively accounted for 99.9% national soybean area were selected to create the fixed population ($N = 7028$ blocks).

Two-stage stratified random sampling was implemented, where the population was divided into four strata according to soybean intensity and a total of 70 sample blocks were randomly selected (Fig. 1, Table 1). The strata definitions and sample sizes were guided by evaluating the precision for estimating soybean area assuming the CDL soybean area was the population of interest. That is, we used the CDL soybean to obtain per-stratum means and variances and this allowed us to compute the standard error of estimated soybean area for various choices of strata and sample size allocation to strata (see Section 2.1.3 for estimation formulas). Based on these analyses, we determined that a sample size of 70 blocks was affordable while still likely to yield estimates that would have adequate precision to demonstrate the utility of the approach.

The second-stage sampling was constrained to the cropland region within each block to avoid visiting remote non-cropland sites, assuming no significant land use conversion between cropland and non-cropland within a year. We created a 2010–2014 maximum cropland mask for each block. A 30 m × 30 m pixel was included in this cropland mask if it was classified by CDL as cropland in any year between 2010 and 2014. Using simple random sampling, we selected 10 pixels within the cropland mask of each first-stage sample block. A total of 700 pixels were selected for field visit, 200 from each of the high, medium, and low strata, and 100 from the very low stratum.

2.1.2. Collecting field data

The crop cover type of every sample pixel was obtained through fieldwork conducted in middle-to-late August 2015. This time window was chosen because soybean plants in the U.S. typically reach reproductive stages with maximum canopy cover during this time of year. This optimal time was confirmed by satellite data. Soybean pixels exhibit maximum normalized difference vegetation index (NDVI) (Tucker, 1979) on Julian date 225 in the MODIS 16-day composites, corresponding to August 13–28 (Fig. 2).