Agricultural Systems 141 (2015) 121–125

Contents lists available at ScienceDirect

Agricultural Systems

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

Short communication

Taken as a given: Evaluating the accuracy of remotely sensed crop data in the USA

Ashley E. Larsen *, Brandon T. Hendrickson, Nicholas Dedeic, Andrew J. MacDonald

Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara 93106-9620, USA

ARTICLE INFO

Article history:
Received 5 June 2015
Received in revised form 12 October 2015
Accepted 19 October 2015
Available online xxxx

Keywords:
Agriculture
Agricultural census
Cropland
Geospatial data

ABSTRACT

Quantifying the extent of agricultural land is important for addressing a large range of ecological, environmental, and economic questions. In many cases, answers focused on land use change require fine scale spatial data on the arrangement of crop types. Here we take advantage of the simultaneous availability of fine resolution, geospatial cropland data—the Cropland Data Layer, and comprehensive tabulated data—the USDA Census of Agriculture, to better understand the accuracy of geospatial data and thus, how geospatial data may be used in scientific research. We compared area estimates for cropland and major US crops (corn, soybeans, wheat and small grains) at the county level for the contiguous US in 2012 and for a subset of states in 2007. We find that accuracy of the Cropland Data Layer is high in regions dominated by a few crop types. However, elsewhere in the US accuracy is highly variable with common large areal overestimates and underestimates (+/−50% or more). Before employing the CDL and other geospatial data for applications such as measuring fine scale changes in land use, users should be wary of the potentially high misclassification error.

1. Introduction

Beyond food production, the extent and spatial arrangement of agricultural land is important for biodiversity, management of agricultural pests and disease (Larsen, 2013; Larsen et al., 2015; Rittenhouse et al., 2012; Tscharntke et al., 2012), carbon storage, bioenergy production, and agricultural policy (Lawler et al., 2014; Mosnier et al., 2013). Spatial arrangement of agricultural land has been shown to be particularly important for maintenance of on and off farm biodiversity (Landis et al., 2000), and for pest management in both ecological (Levins, 1969) and economic theory (Costello et al., 2014). However, a dearth of refined, spatially explicit data on cropland arrangement has largely limited investigations to either field studies or general analytical models thus hampering consensus across heterogeneous regions or crop types.

With the recent explosion of satellite data, large-scale studies are just now becoming feasible. This is especially exciting for investigations of agricultural processes in developing countries where traditional agricultural statistics have historically been unavailable. Yet, in order to understand how satellite data can and should be applied, scientists must understand the accuracy of such data relative to other agricultural statistics. To do so necessitates focusing on regions where satellite data can be compared to high quality tabulated data. Here we take advantage of the simultaneous availability of the USDA Census of Agriculture, the most comprehensive agricultural statistics tabulated in the US (USDA, 2014), and the National Agricultural Statistics Service (NASS) Cropland Data Layer (CDL), one of the only agriculture-focused satellite data layers available annually, to understand how satellite data may be employed more broadly in rigorous scientific investigations.

The purpose of the CDL is largely to produce seasonal area estimates for major crops or to inform the design of other NASS data products, such as the June Acreage Survey (Johnson, 2013). However, scientists have leveraged the disaggregated crop classification and refined spatial resolution of these data to address a much wider range of questions from grassland conversion to soy/corn (Wright and Wimberly, 2013), to predicting crop area in response to commodity price (Hendricks et al., 2014), to investigating land use change and conversion between specific crop types associated with the spatial location of ethanol refineries (Johnston, 2013). For these and similar studies, sub-county assessments, which rely on the accuracy of pixel data, are critical.

Area estimation from pixel counting, however, is thought to be biased downward, resulting in underestimates of cropland area (Johnson, 2013). While such bias in estimates of cropland extent could be corrected using a regression with other annual data (Boryan et al., 2011), spatial arrangement of different crop types could be distorted and is not easily corrected in such a manner. Furthermore, such corrective regression methods are not commonly used in scientific applications employing these data and would be impossible to replicate with other geospatial data in countries lacking accurate agricultural tabulations. Thus, understanding the accuracy of pixel counting is important for understanding the suitability of the CDL and other satellite data for the various applications to which they have been and could be employed.

* Corresponding author at: Department of Environmental Science, Policy & Management, University of California, Berkeley 94720-3114, USA.
E-mail address: Larsen.ashley@gmail.com (A.E. Larsen).

http://dx.doi.org/10.1016/j.agsy.2015.10.008
0308-521X/

To develop a comprehensive and scientifically relevant analysis of the accuracy of the CDL, we compare area estimates from the CDL to the USDA Census of Agriculture for 2012, the first year both data products are simultaneously available for the coterminous US. Using these datasets we address, (1) how different are county-level estimates of cropland and crop groups derived from pixel counting from the Census and are these differences statistically significant? and (2) how often and where are county-level CDL and Census estimates statistically similar for individual crops such as soybeans and wheat? For robustness, we also compare the CDL and Census estimates in 2007 for the subset of states available that year.

2. Methods & materials

2.1. Cropland Data Layer

NASS states the purpose of CDL is to produce area estimates for the Agricultural Statistics Board for major commodities within each state, and to "produce digital, crop-specific, categorized geo-referenced output products" (USDA and N.A.S.S., n.d.). The CDL was first produced for the Corn Belt in the late 1990s (USDA and N.A.S.S., n.d.). In 2007, the CDL was produced using a combination of satellite imagery (primarily AWiFS and Landsat TM) with rapid revisit (5 d, 16 d, respectively) for 21 states with a resolution of 56 m (Boryan et al., 2011). As satellite and computer technology increased in power and decreased in cost, NASS began to produce these high-resolution data freely available, annually, and for the contiguous US. The 2012 CDL dataset used an updated combination of satellite imagery (Deimos-1, UK-DMC2, and Landsat TM/ETM+) collected during the growing season to produce a cloud-free, 30 m data layer (USDA and N.A.S.S., n.d.). The CDL is ground truthed by the USDA Farm Service Agency (FSA) Common Land Unit (CLU) Program data (USDA and N.A.S.S., n.d.). An in-depth description of the program can be found in Boryan et al. (2011). State-level accuracy assessments are available in the metadata, but pixel accuracies for tilled crops are generally reported to be 70–95% at the state level (Boryan et al., 2011; Johnson, 2013).

2.2. USDA Census of Agriculture

The USDA Census of Agriculture is conducted every five years by the NASS, and is considered the most comprehensive agricultural data for every county in the US (USDA, 2014). The Census is conducted via questionnaires provided to every farm that produced or sold at least $1000 of agricultural products (or had the potential to) in a census year (USDA, 2014). The responses are generally tabulated at the county level. The Census provides extensive information regarding crop and livestock production, costs, inputs, and farmer demographics.

In 2012, the Census had a response rate of ~80% (USDA, 2014). The Census compensates for bias stemming from non-response or incomplete mailing lists at the country, state, and county level using a combination of weighted adjustments and other imputation measures to "produce agricultural census totals for publication that were fully adjusted for [mailing] list undercoverage, nonresponse and misclassification [of farm/nonfarm] at the county level" (USDA, 2014).

In 2012, the Census provides a measure of the uncertainty due to the above errors at the state and county level by means of a coefficient of variation. From the state level, generalized coefficient of variation, a 95% confidence interval around the census estimate can be easily computed (USDA, 2014).

2.3. Comparison

We compare the CDL to the Census of Agriculture for measures of total cropland area, and major crops and crop groups in the contiguous 48 US states in 2012 (Table 1). To do so, we use CDL crop pixels converted to acres and aggregated at the county level. These data are provided by NASS on the CropScape FAQs website (USDA and N.A.S.S., n.d.). We converted all measures to hectares. We compare the data sets using paired t-tests, and measure percent difference from the Census, provided the Census records at least 20 ha (~50 acres) of a given crop. We include this lower benchmark to avoid enormous percent difference resulting from trivially small differences in area.

We map percent differences for corn, soybeans, small grains (wheat, oats, barley) and cropland. Percent difference for crop i in county c was calculated as, $\%\text{Diff}_{\text{ic}} = \frac{\text{Census}_{\text{ic}} - \text{Nass}_{\text{ic}}}{\text{Census}_{\text{ic}}} * 100$.

We define cropland in the Census to be "cropland harvested" and "cropland on which all crops failed". The Census definition of cropland harvested includes area of hay, but not pasture. To construct as comparable a group as possible, our measure of cropland in the CDL includes crops as well as alfalfa and other hay, but not pasture or grassland (Table 1). We compute paired t-tests for each crop comparison to evaluate statistical significance of the observed differences in the two datasets. For a given crop, the t statistic was calculated as $t = \frac{\overline{X}_D}{S_D / \sqrt{n}}$ where $\overline{X}_D$ is the mean difference between the NASS and Census area estimates, S_D is the standard deviation of the sample difference, and n

Table 1
CDL and comparable Census crop categories and metrics of difference for observations in the continental US with at least 20 ha reported in the Census. The Census crops are measured in harvested area. The Census counts area for each crop harvested on the same plot of land as area for each crop, but only counts the area once in the overall cropland categories. To be consistent, we did the same for the CDL for grouped crops and overall cropland. The 2012 and 2007 CDL categories are consistent, but some categories present in the 2012 CDL are not present in the 2007 CDL. "Avg. % diff" and "Avg. Ha diff." indicate average percent difference and average difference in hectares between the Census and CDL at the county level. ** and * indicate significant differences between the two datasets at the 0.01 and 0.05 level, respectively, based on paired t-tests.

	CDL categories (grouped categories)	Census tables	Avg. % diff. 2012 (+/−SD)	Avg. Ha diff. 2012 (+/−SD)	Avg. % diff. 2007 (+/−SD)	Avg. Ha diff. 2007 (+/−SD)
Corn	001, 225, 226, 237, 241 ("Corn all")	Corn grain (Table 25), Corn silage (Table 26)	−16 (207)	−171* (3568)	9 (71)	2111** (5137)
Soybean	005, 026, 239, 240, 241, 254 ("Soybeans all")	Soybeans (Table 25)	−3 (65)	192* (3223)	10 (45)	1748** (3865)
Corn & soybeans	Corn + Soybeans	Corn + Soybeans	−23 (222)	−40 (5469)	6 (93)	3534** (7748)
Small grains	021, 022, 023, 024, 026, 028, 225, 226, 230, 234, 235, 236, 237, 238, 240, 254 ("Wheat Durum all", "Wheat Winter all", Spring Wheat, "Oats all", "Barley all")	Barley for grain, oats for grain, wheat for grain (all) (Table 25)	−33 (165)	−1129** (6320)	−4 (215)	230 (7302)
Total cropland	001–006, 010–014, 021–39, 041–057,059–060, 066–072, 074–077, 204–209, 211–214, 216–227, 229–232, 234–250, 254	Harvested cropland, cropland on which all crops failed (Table 8)	9 (93)	2234** (11,675)	−4 (1014)	12,375** (15,774)