Computers and Electronics in Agriculture 109 (2014) 232–246

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Reconciling field size distributions of the US NASS (National Agricultural Statistics Service) cropland data

Yubin Yang *, Lloyd T. Wilson, Jing Wang

Texas A&M University System, AgriLife Research & Extension Center, 1509 Aggie Drive, Beaumont, TX 77713, USA

ARTICLE INFO

Article history:
Received 17 December 2013
Received in revised form 1 July 2014
Accepted 11 October 2014

Keywords:
Cropland Data Layer
Field size distribution
Pixel merging
Fishnet subdivision
Distribution optimization
Geographic Information System

ABSTRACT

Cropping system models and decision support tools are increasingly structured to provide the capability for simulation and analysis at multiple spatial scales. A major challenge is having access to large volumes of geo-referenced data. The Cropland Data Layer (CDL) products from the U.S. National Agricultural Statistics Service (NASS) provide digitized maps on the types and distributions of major agricultural crops and non-agricultural lands. However, the CDLs have too many single pixels or small pixel clusters that are classified as crop fields due to spectral artifacts, and unrealistically large polygons classified as fields due to CDL's limited ability to distinguish field boundaries. This paper presents a methodology to reconcile NASS CDLs with observed field size distribution data. The reconciliation algorithm consists of several automated steps, including field delineation by road networks, streams, and road extension; pixel merging; fishnet subdivision; and distribution optimization. The distribution optimization uses both pixel merging and fishnet subdivision to match the reconciled data with observed field size distribution. Observed field size data for rice and cotton (2009–2012), and corn and sorghum (2012) from selected counties in Texas were used to test the algorithm. The 2009–2012 CDLs underestimated total rice acreage by 2–12% for Colorado County and overestimated total cotton acreage by 1–10% for Dawson County, while the CDLs overestimated the number of fields by up to 714% for rice and 280% for cotton. In contrast, the 2009–2012 optimized CDLs underestimated the number of rice fields by 4–13% and overestimated the number of cotton fields by 6–13%. The corresponding acreage ranges from 1% increase to 4% reduction for rice and 5–12% reduction for cotton. Most of the reduction in acreage is due to areas that were classified as rice or cotton fields but now identified as road networks since the CDLs have limited capability to separate fields from road networks. Summary results were also provided for other land types. Field size distributions of the reconciled CDLs closely match observed data and are appropriate for use in cropping systems simulations and analyses such as crop rotation, land use change, and biorefinery site selection. The methodology has broad applicability to other digitized land cover data.

1. Introduction

Cropping system models and decision support tools have traditionally been used for point or site-specific applications. With the increasing availability of digital spatial data and advancement in Geographic Information Systems (GIS), crop models are increasingly structured to provide the capability for area-wide simulation and analysis at a range of spatial scales (Balkovič et al., 2013; Beccali et al., 2009; Gijsman et al., 2007; Graham et al., 2000; Hartkamp et al., 1999; Liu, 2009; Liu et al., 2007; Shi et al., 2008; Tan and Shibasaki, 2003). Hartkamp et al. (1999) provided a comprehensive review on interfacing GIS with agronomic models for simulation and analysis. Graham et al. (2000) developed a regional-scale, GIS-based modeling system for estimating biomass supplies from energy crops by integrating digital map data on soil, land use, road network, and watershed. Liu et al. (2007) developed a GIS-based EPIC Model (GEPIC), which integrates GIS with the EPIC (Environmental Policy Integrated Climate) bio-physical model, to simulate the spatial and temporal dynamics of major processes of a soil-crop-atmosphere-management system on a global scale (Liu, 2009).

A major challenge in supporting large-scale spatial simulation and analysis is lack of data management systems that provide dynamic access to large volumes of geo-referenced data (Boryan et al., 2012; Feng et al., 2009; Gärtner et al., 2013; Gijsman et al., 2007; Johnson, 2013; Liu, 2009; Resop et al., 2012; Secchi et al., 2011; Yan et al., 2013; Yang et al., 2011, 2012). The Integrated

* Corresponding author. Tel.: +1 409 752 2741.
E-mail address: yyang@aesrg.tamu.edu (Y. Yang).

http://dx.doi.org/10.1016/j.compag.2014.10.012
0168-1699/© 2014 Elsevier B.V. All rights reserved.

Agricultural Information Management System (*iAIMS*) was developed to address the challenge of both data consolidation and integration (Wilson et al., 2007, 2010; Yang et al., 2007, 2010b, 2011). *iAIMS* consists of climatic, soil, cropland, and road network databases, which serve as a foundation for applications that address different aspects of cropping system performance and management (Arthur et al., 2011; Yang et al., 2004, 2010a, 2012). The major source for *iAIMS* climatic data is from the U.S. National Oceanographic and Atmospheric Administration National Climatic Data Center (NCDC, 2013). *iAIMS* soil data is based on the Soil Survey Geographic (SSURGO) database (NRCS, 2013), while its cropland data is based on the Cropland Data Layer (CDL) products from the USDA National Agricultural Statistics Service (Boryan et al., 2011; NASS, 2013).

The Cropland Data Layer (CDL) products provide digitized maps on the types and distributions of major agricultural crops (e.g. rice, cotton, corn, and soybean) and non-agricultural lands (e.g. shrubland, wetland, and woodland) (Han et al., 2012; Wilson et al., 2010). The pixel-level accuracy of the CDLs is approximately 80% for most crops, but usually above 90% for major crops such as corn, soybean, wheat, cotton and rice (Johnson and Mueller, 2010).

A number of applications have been developed based on the CDL raster maps (Boryan et al., 2012; Han et al., 2012; Johnson, 2013; Resop et al., 2012; Yang et al., 2011). Secchi et al. (2011) used 2002–2006 CDLs to construct historic crop rotations to examine the impact of biofuels expansion on land use change in Iowa. Boryan et al. (2012) developed and evaluated different sets of rules for pixel merging to build US cultivated land data sets, using multi-year CDLs. Han et al. (2012) provided a detailed description of the design and implementation of CropScape – a web application for exploring and disseminating the CDL products. Johnson (2013) integrated multi-year CDLs to provide a detailed estimate and analysis on annually tilled cropland within the conterminous United States.

Liu et al. (2005) developed a stratified, multi-layer sampling framework to account for the fraction of non-cultivated land (e.g. narrow roads and footpaths, small rice paddy levees, irrigation channels, etc.) to rectify cropland area estimation in China. Their study indicates that the gross 2000 cropland area for China without rectification was 27.5% more than the rectified cropland area, with the gross area of paddy land and dry farming land overestimated by 32.7% and 25.7%, respectively. Their results highlight the need for methodologies to improve classification and estimation of cropland area.

A limitation of the CDL raster maps is its lack of information on individual field boundary and field size, which are needed for applications relying on spatial distributions of individual crop fields, such as crop rotation, land use change, area-wide pest management, and biorefinery site selection.

The CDL raster maps can be converted to vector maps with each vector object considered as an individual field. Our preliminary analysis indicates substantial differences in the number of fields and field size distributions between the converted vector maps and the observed data. As a result, analyses based on the converted vector maps can give greatly biased results. The objective of this paper is to present a methodology to reconcile NASS CDLs to match observed crop field size distributions.

2. Methodology

2.1. Data sets and sources

A number of data sets were used to evaluate the reconciliation algorithm (see Section 2.2), including original CDLs, county boundary, road networks, streams, county-level crop acreage, and actual crop-specific field size distributions. The CDL products for Texas from 2009 to 2012 were downloaded from USDA-NRCS Geospatial Data Gateway (NRCS, 2013). They are raster images in GeoTIFF (.TIF) format. Field size distributions for the original and reconciled CDLs were obtained based on the attribute table in the corresponding CDL feature class (see Section 2.2). Feature class data for county boundary, road networks and streams were obtained from ESRI Data and Maps (ESRI, 2013b). County-level crop-specific acreage data were downloaded from NASS Quick Stats (NASS, 2012).

Observed field size data for rice and cotton (2009–2012), and corn and sorghum (2012) from selected counties in Texas were used to evaluate the reconciliation algorithm. Area of individual rice fields from 2009 to 2012 for Colorado County were obtained from the Texas Rice Crop Survey (Wilson et al., 2013), which collects detailed data on rice production from rice producers, crop consultants, and rice mills. The total survey acreage for Colorado County accounts for 54–89% of the total rice acreage for 2009–2012. It was assumed that those fields included in the survey are representative of the rice fields for the county. The number of total rice fields for the county for a specific year was adjusted based on the percentage of the fields included in the survey. Observed field size data from 2009 to 2012 for cotton, corn, and sorghum were obtained from USDA Farm Service Agency (FSA, 2013) through a Freedom of Information Act request, with information on farm, track and field ID removed.

2.2. CDL reconciliation procedures

The GeoTIFF raster images from the CDL products can be converted to vector files consisting of polygons, using GIS software such as ESRI ArcGIS geoprocessing tool (ESRI, 2013a). These polygons are treated as fields, but may not represent actual fields, especially for those unrealistically small or large polygons. For the purpose of discussion, we use the term ‘fields’ to refer to either polygons derived from the raster images or observed field data. Preliminary analysis indicates substantial differences in field size distribution between the polygons and actual fields, with the NASS CDLs having too many single or small pixel-cluster polygons due to spectral artifacts (Trichilo et al., 1996), and unrealistically large field polygons due to CDL’s limited ability to distinguish roads and streams between fields, and field boundaries. We used line vectors of streams and road networks to separate individual fields, an approach similar to that used by Trichilo et al. (1996).

Even with delineation of field boundary by streams and road networks, some estimated fields remained unrealistically large. The next step was to extend a road ending inside a large field polygon to the boundary of the field, subdividing the field into two smaller fields. This was followed by a stepwise process of pixel merging of small polygons into neighboring fields and subdivision of excessively large polygons into realistically sized fields using ERSI *Fishnet* feature (ESRI, 2013a). For crops that have observed field size data, a final optimization step was used to minimize the deviation between the derived and observed field size distributions.

Fig. 1 provides a schematic representation of the steps of the field size reconciliation process, and Table 1 summarizes key operations that can be accessed from the interactive ArcGIS desktop (manual) or from Python script (automated). The reconciliation process was performed county by county. The entire process was fully automated using a combination of Microsoft Visual Studio 2008 and C# language (Microsoft, 2007), and ESRI ArcGIS and Python scripts (ESRI, 2013a).

2.2.1. CDL – Stage 0

The GeoTIFF CDL image was converted to a polygon feature class using Python script (ESRI, 2013a) (Table 1). Each group of