APPLYING MACHINE LEARNING TO CROPLAND DATA LAYER FOR AGRO-GEOINFORMATION DISCOVERY

Chen Zhang[†‡], *Zhengwei Yang*[§], *Liping Di*[†‡*], *Li Lin*[†‡], *Pengyu Hao*[†], *Liying Guo*[†]

[†] Center for Spatial Information Science and Systems, George Mason University, Fairfax, VA 22030, USA
[‡] Department of Geography and Geoinformation Sciences, George Mason University, Fairfax, VA 22030, USA
[§] Research and Development Division, U.S. Department of Agriculture National Agricultural Statistics Service, Washington, DC 20250, USA
czhang11@gmu.edu, zhengwei.yang@usda.gov, {ldi*, llin2, phao, lguo2}@gmu.edu

IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9554628

ABSTRACT

The Cropland Data Layer (CDL) is currently the only subfield level high resolution crop-specific land cover data product over the entire conterminous United States (CONUS). It has been widely used in agricultural industry, business decision support, research, and education worldwide. However, CDL data has its limitations. It is an end-of-season land cover map which is not available within growing season. Moreover, CDLs in early years have many misclassified pixels (relatively low accuracy) due to cloud cover and lack of satellite images. This paper will present the studies of using machine learning technique to address these issues in CDL data. Specifically, we will present the design and implementation of a machine learning model for agro-geoinformation discovery from CDL. Several application scenarios of the proposed model, including prediction of crop cover, crop acreage estimation, in-season crop mapping, and refinement of the early-year CDL data, are demonstrated and discussed.

***Index Terms*—** Machine learning, Cropland Data Layer, Agro-geoinformatics, Crop type classification

1. INTRODUCTION

Agro-geoinformatics is a new interdisciplinary that enabled geoinformatics in the study of advanced science and technology in agriculture [1]. As one of the state-of-the-art technologies in agro-geoinformatics, machine learning is efficient and effective to automatically discover intricate patterns and structures in agro-geoinformation data. A variety of machine learning-based approach has been developed and applied to agricultural applications and researches, such as land use and land cover (LULC) mapping [2], crop type classification [3], crop yield prediction [4], drought monitoring [5], agricultural sustainability [6], climate change assessment [7]. The accurate and reliable geospatial data is the key for the success of applying machine learning algorithms and methods in these applications. Among various open access geospatial data sources, the Cropland Data Layer (CDL) of the U.S. Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) has been widely used as reference data set in agricultural and environmental research as well as geosciences and remote sensing studies. The CDL is a raster, geo-referenced, crop-specific, 30-meter spatial resolution land cover data layer created annually for the continental United States (CONUS) using moderate resolution satellite imagery and extensive agricultural ground truth. It contains over 140 land cover classes with around 95% accuracy for major crop types [8]. Although CDL provides detailed land use information for the entire CONUS, there are two limitations that could affect many follow-on research and applications. First, in-season CDL data are not available for applications and research since the current year CDL is usually released to the public in early next year. Second, the accuracies of the early-year CDL products are relatively lower than recent years' due to cloud cover or lack of the original Landsat images. This paper presents several application scenarios of using the machine learning approach and historic CDL data for agro-geoinformation discovery.

The rest of the paper is organized as follows. Section 2 describes the data and study areas of this study. Section 3 introduces the design of the proposed machine learning model. Section 4 demonstrates several application scenarios of the method, including prediction of crop cover, crop acreage estimation, in-season crop mapping, and refinement of historical crop cover map. Section 5 gives the conclusion and future research recommendation.

2. DATA AND STUDY AREAS

This paper utilizes available high confidence pixels of the historic CDL data for machine learning model training. The full archive of CDL data are hosted on CropScape (`https://nassgeodata.gmu.edu/CropScape`), which is a geospatial web application for visualization, dis-

*Corresponding Author

semination, and analytics of on-demand CDL data [9]. The NASS CDL National Confidence Layers are used for selecting high confidence pixels. This data layer spatially represents the predicted confidence that is associated with that output pixel, based upon the rules that were used to classify it. Each layer provides the spatial representation of distribution and magnitude of error or confidence of the classification of CDL. The high confidence pixels are selected by thresholding the confidence layers with a high confidence threshold (e.g., 90% confidence). The Confidence Layers are available at USDA NASS website (`https://www.nass.usda.gov/Research_and_Science/Cropland/Release/index.php`). This study mainly focuses on the U.S. Corn Belt region. The region of interest may vary for the different study scenarios as shown in Section 4.

3. METHODOLOGY

To automatically recognize the crop sequence information from the CDL time series, we developed a machine learning workflow. First, we processed and retrieved the CDL data using the AgKit4EE toolkit through Google Earth Engine [10]. The data set are converted into a stack of CDL time series for the study area, which could be a county, an Agricultural Statistics District (ASD), or any study area. All pixels of the CDL time series are arranged into a 2-D array of samples. Each row of the data set array represents a pixel consisting of a sequence of crop type values of different years.

A machine learning model based on artificial neural network (ANN), which has been extensively used for remote sensing image interpretation [11, 12], was then developed to recognize the crop sequence pattern from the prepared CDL time series. Figure 1 illustrates the architecture of the crop sequence model. The proposed ANN has a multilayer perceptron (MLP) structure including one input layer, three hidden layers, and one output layer. The input layer has a group of input nodes corresponding to the crop type of each year in the historical CDL time series for individual pixels. The output layer uses SoftMax to estimate the probability of each crop type. Based on the probability distribution, the crop type of the target pixel will be assigned. Then we use the machine-learned crop sequence model to derive additional agro-geoinformation which could be used for solving problems in the present CDL data.

4. MODEL APPLICATIONS

4.1. Pre-season Crop Mapping

The pre-season crop mapping aims to predict the spatial distribution of crop cover before the beginning of a growing season. The proposed machine learning model can be used to predict the crop cover map from the historical CDL time series. Our study has shown that the prediction result of the U.S. Corn Belt is expected to reach 88% agreement with the future CDL [13]. Figure 2 illustrates an example of machine-learned prediction of 2018 crop cover. The probability map represents the spatial distribution of the highest probability from the SoftMax function. The crop cover map is similar with CDL data where each pixel is categorized as one of land cover categories. The 2018 CDL data are used as reference data to evaluate the prediction result. The predicted map of 12 land cover classes achieved the overall accuracy (OA) of 90% with Kappa value of 0.86.

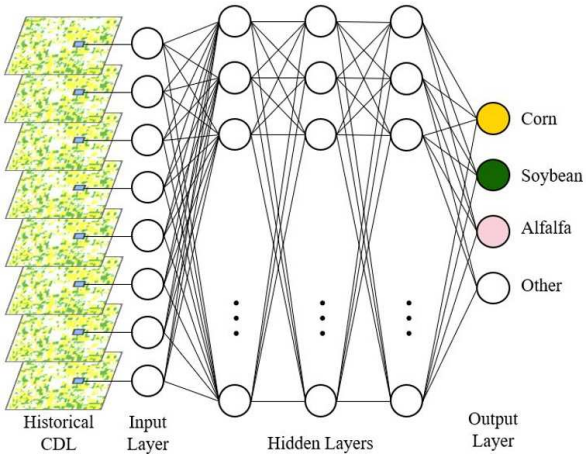


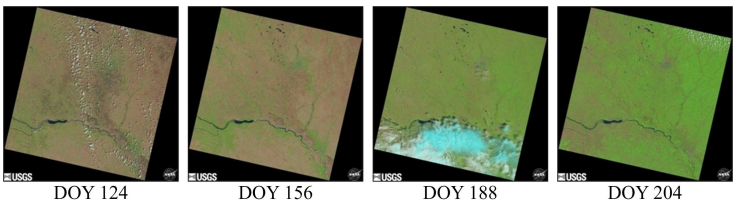
Fig. 1: Structure of the MLP-based crop sequence model.

(a) Probability of prediction (b) Predicted crop cover (c) Reference data (2018 CDL)

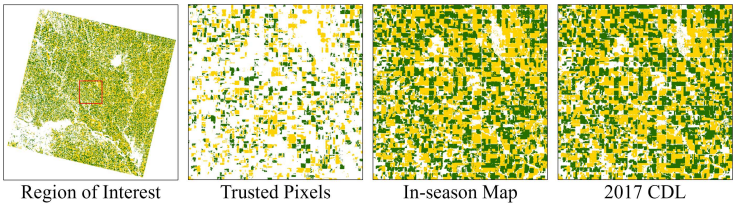
Fig. 2: Prediction of 2018 crop cover. The bright pixels in the probability map indicate high confidence pixels. The yellow and green pixels indicate corn and soybeans.

4.2. Crop Acreage Prediction

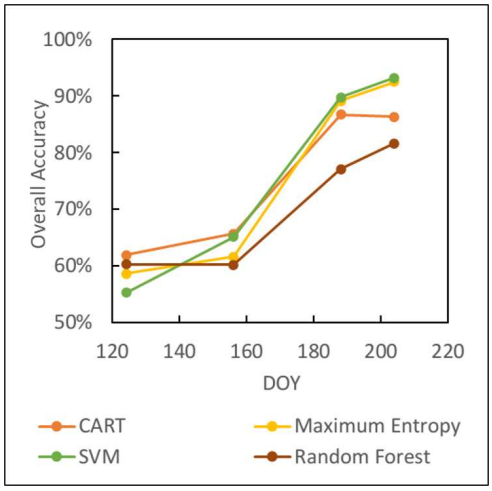
Crop acreage is one of the most critical information in agricultural decision making. With the predicted crop cover map derived from the crop sequence model, we can also estimate the future crop acreage. To assess the performance of the crop acreage prediction, we compared ASD-level crop acreage prediction with the official statistics by USDA NASS Iowa Field Office. Table 1 summarizes the total acreage of each crop type for Iowa. This result suggests the machine-



(a) Landsat-8 images (path: 029, row: 030) at specific day of year (DOY) from May 2017 to July 2017.



(b) Comparison of trusted pixels, in-season crop cover map (produced by July 2017), and 2017 CDL (released in February 2018).



(c) Classification performance at DOY within the growing season.

Fig. 3: Crop type classification using Landsat-8 images and machine-learned trusted pixels as training samples.

learned crop acreage predictions of corn is very close to the CDL data. The machine-learned crop acreage estimates of soybeans, on the other hand, is a little bit lower but still close to the CDL data. The crop acreage of both machine-learned result and CDL are less than the official statistics.

Table 1: Crop acreage estimates for Iowa, USA.

	Year	Prediction (acre)	CDL (acre)	Field Office (acre)
Corn	2016	13,607,650	13,628,727	13,900,000
	2017	13,420,096	13,216,879	13,300,000
	2018	13,704,565	13,537,935	13,200,000
Soybeans	2016	8,971,998	9,232,107	9,500,000
	2017	9,401,789	9,785,308	10,000,000
	2018	9,535,296	9,959,717	9,996,000

4.3. In-season Crop Mapping

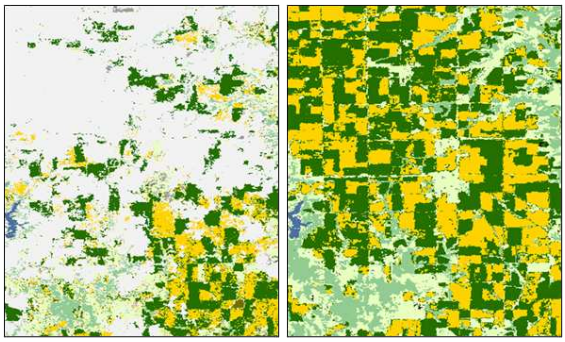
One important issue that needs to be addressed for the pre-season mapping is that the prediction result only relies on prior knowledge from historical data, which could be problematic while mapping for the year with disasters or large market volatility or major policy changes. Since remote sensing data contain abundant spectral signature information of different crop growth stages, remote sensing images acquired at the early growing season can be combined with trusted historical crop rotation patterns to improve the mapping result. Based on this idea, we assume the trusted pixels automatically learned from the historical CDL time series, whose crop types have been identified with high confidence by the crop sequence model, can be used as pseudo ground truth data to label training samples on the remote sensing data for in-season crop mapping. Figure 3 shows some preliminary results of crop type classification with multi-temporal Landsat-8 images and trusted pixels as training samples using common supervised classifiers including classification and regression tree (CART), maximum entropy, random forest, and support vector machine (SVM). The result indicates that the in-season crop cover map of corn and soybeans can reach over 90% agreement with the official CDL data by the end of July.

4.4. Refinement of CDL

Another application scenario of the proposed crop sequence model is the refinement of historical CDL. The quality of the early-year CDL data was not as good as recent years. In early years, there are many misclassified pixels in the CDL products because of cloud cover and lack of satellite images. To address this issue, we used the proposed machine learning model to refine and correct misclassified pixels in the historical CDLs. Our study showed that the proposed machine learning model can automatically correct most of misclassified pixels in an original CDL map [14]. Figure 4 illustrates the comparison of the original CDL data with the refined CDL data. It can be found that the misclassified pixels, especially the cloud pixels, had been corrected with the crop sequence information learned from the historical CDL time series.

5. CONCLUSIONS

This study presented a low-cost and effective machine learning approach for agro-geoinformation discovery from the CDL data and demonstrated several application scenarios. More experiments and validation will be conducted in the



(a) Original CDL (b) Refined CDL

Fig. 4: Comparison of original CDL and refined CDL.

future. Meanwhile, we will improve the current machine learning model and explore other applications. For example, the proposed method can be potentially used to identify more crop types and scale up to the entire CONUS. The machine learning-based crop sequence model also has great potential to be integrated with other artificial intelligence (AI) technique to discover spatial and temporal trend of cropping across the U.S. from CDL.

6. ACKNOWLEDGMENT

This research is supported by grants from National Science Foundation INFEWS program (Grant #: CNS-1739705, PI: Dr. Liping Di) and National Agricultural Statistics Service of U.S. Department of Agriculture (Grant #: 58-3AEU-7-0080, PI: Dr. Liping Di).

7. REFERENCES

[1] L. Di and Z. Yang, "Foreword to the Special issue on Agro-Geoinformatics—The Applications of Geoinformatics in Agriculture," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 11, pp. 4315–4316, Nov. 2014.

[2] N. Kussul, A. Shelestov, M. Lavreniuk, I. Butko, and S. Skakun, "Deep learning approach for large scale land cover mapping based on remote sensing data fusion," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Jul. 2016, pp. 198–201.

[3] P. Hao, L. Di, C. Zhang, and L. Guo, "Transfer Learning for Crop classification with Cropland Data Layer data (CDL) as training samples," *Science of The Total Environment*, vol. 733, p. 138869, Sep. 2020.

[4] P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 163, p. 104859, Aug. 2019.

[5] S. Park, J. Im, E. Jang, and J. Rhee, "Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions," *Agricultural and Forest Meteorology*, vol. 216, pp. 157–169, Jan. 2016.

[6] R. Sharma, S. S. Kamble, A. Gunasekaran, V. Kumar, and A. Kumar, "A systematic literature review on machine learning applications for sustainable agriculture supply chain performance," *Computers & Operations Research*, vol. 119, p. 104926, Jul. 2020.

[7] A. Crane-Droesch, "Machine learning methods for crop yield prediction and climate change impact assessment in agriculture," *Environmental Research Letters*, vol. 13, no. 11, p. 114003, Oct. 2018.

[8] C. Boryan, Z. Yang, R. Mueller, and M. Craig, "Monitoring US agriculture: The US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program," *Geocarto International*, vol. 26, no. 5, pp. 341–358, 2011.

[9] W. Han, Z. Yang, L. Di, and R. Mueller, "CropScape: A Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support," *Computers and Electronics in Agriculture*, vol. 84, pp. 111–123, 2012.

[10] C. Zhang, L. Di, Z. Yang, L. Lin, and P. Hao, "AgKit4EE: A toolkit for agricultural land use modeling of the conterminous United States based on Google Earth Engine," *Environmental Modelling & Software*, vol. 129, p. 104694, Jul. 2020.

[11] P. M. Atkinson and A. R. L. Tatnall, "Introduction Neural networks in remote sensing," *International Journal of Remote Sensing*, vol. 18, no. 4, pp. 699–709, Mar. 1997.

[12] J. Li, X. Huang, and J. Gong, "Deep neural network for remote-sensing image interpretation: Status and perspectives," *National Science Review*, vol. 6, no. 6, pp. 1082–1086, Nov. 2019.

[13] C. Zhang, L. Di, L. Lin, and L. Guo, "Machine-learned prediction of annual crop planting in the U.S. Corn Belt based on historical crop planting maps," *Computers and Electronics in Agriculture*, vol. 166, p. 104989, Nov. 2019.

[14] C. Zhang, Z. Yang, L. Di, L. Lin, and P. Hao, "Refinement of Cropland Data Layer Using Machine Learning," in *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-3-W11, Baltimore, Maryland, USA, Feb. 2020, pp. 161–164.