REFINEMENT OF CROPLAND DATA LAYER USING MACHINE LEARNING

Chen Zhang [1,2], Zhengwei Yang [3], Liping Di [1,2,*], Li Lin [1,2], Pengyu Hao [1]

[1] Center for Spatial Science and Systems, George Mason University, Fairfax, VA 22030, USA - (czhang11, ldi, llin2, phao)@gmu.edu
[2] Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA 22030, USA
[3] Research and Development Division, U.S. Department of Agriculture National Agricultural Statistics Service, Washington, DC 20250, USA - Zhengwei.Yang@usda.gov

KEY WORDS: Cropland Data Layer, Machine Learning, Misclassification Correction, Crop Sequence Modeling, Raster Map Refinement

ABSTRACT:

As the most widely used crop-specific land use data, the Cropland Data Layer (CDL) product covers the entire Contiguous United States (CONUS) at 30-meter spatial resolution with very high accuracy up to 95% for major crop types (i.e., Corn, Soybean) in major crop area. However, the quality of early-year CDL products were not as good as the recent ones. There are many erroneous pixels in the early-year CDL product due to the cloud cover of the original Landsat images, which affect many follow-on researches and applications. To address this issue, we explore the feasibility of using machine learning technology to refine and correct misclassified pixels in the historical CDLs in this study. An end-to-end deep learning-based framework for restoration of misclassified pixels in CDL image is developed and tested. By feeding the CDL time series into the artificial neural network, a crop sequence model is trained and the misclassified pixels in an original CDL map can be restored. In the experiment with the 2005 CDL data of the State of Illinois, the misclassified pixels over Agricultural Statistics Districts (ASD) #1760 were corrected with a reasonable accuracy (>85%). The findings suggest that the proposed method provides a low-cost and reliable way to refine the historical CDL data, which can be potentially scaled up to the entire CONUS.

1. INTRODUCTION

Since its first release of a full state wide data product in 1997, the Cropland Data Layer (CDL) product of the U.S. Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) has been widely used by growers, agricultural industry, governments, educators and students, and researchers world-wide for crop production, agricultural production planning and management, government policy formulation and decision making, teaching, and various research activities (Liknes et al., 2009; Thompson, Prokopy; Hao et al., 2015; Lark et al., 2015; Di et al., 2017). Currently, the CDL data covers the entire conterminous United States (CONUS) at 30-meter spatial resolution with a high accuracy up to 95% for classifying major crop types (i.e., Corn, Soybean, and Wheat). However, the quality of the early-year CDL products 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. Moreover, only a few states of CDL data were produced before 2008. For example, the year 2000 CDL covers only Illinois, Indiana, Mississippi, North Dakota, and a part of Arkansas and Iowa. Obviously, the earlier year CDLs' availability and low quality issues affect many follow-on Land Use and Land Cover (LULC) related researches and applications. Therefore, an effective method for refining and correcting the old CDL data is badly needed to improve the quality and accuracy of the historical CDL data.

It is well known that monocropping will result in degradation of soil, build-up of diseases and pests, and decline in productivity. Thus crop rotation becomes a common farming practice in U.S. Corn Belt. The crop rotation can significantly improve the soil condition, such as fertility and soil physical/chemical properties (Pikul et al., 2001; Karlen et al., 2006; Govaerts et al., 2007; Karlen et al., 2013; Van Eerd et al., 2014). Meanwhile, the crop sequence and cropping decision also have significant impact on crop yields and profitability (Temperly, Borges; Parajuli et al., 2013; Farmaha et al., 2016). Based on this common cropping practice, many crop mapping and yield estimation models and approaches were developed. Secchi et al. (2011) constructed an prediction model of future land use scenario in the state of Iowa based on the corn-soybean rotation and production costs. Schönhart et al. (2011) developed a crop sequence model to generate crop rotations based on agronomic criteria and observed data. Sahajpal et al. (2014) detected the pronounced shifts from grassland to cultivated area by modelling crop rotation in the U.S. Western Corn Belt. Hao et al. (2016) explored the crop classification based on the previous-year crop knowledge. Zhang et al. (2019a) produced a crop cover map of Nebraska State based on the common crop rotation patterns of corn, soybeans, winter wheat, and alfalfa. They further implemented a crop sequence-based machine learning framework for prediction of crop cover maps (Zhang et al., 2019b).

In this paper, we present a machine learning-based crop sequence model to refine the historical CDL data. The proposed model utilizes artificial neural network (ANN) to automatically learn crop sequence information from the CDL time series. The misclassified pixels in the crop cover map can be automatically identified and corrected using the trained model on the historical CDL.

The rest of the paper is organized as follows. Section 2 introduces the CDL data, the study area, and an end-to-end machine learning framework for the historical CDL data refinement. Section 3 demonstrates the experiment results and as-

*Corresponding author

sesses the refinement performance. Section 4 discusses the limitation of the current implementation and gives the conclusion.

2. METHODS

2.1 Cropland Data Layer

CDL is a raster formatted, geo-referenced, crop-specific land cover map produced by USDA NASS. It is an annual product covering the entire CONUS at 30-meter spatial resolution from 2008 to present and some states from 1997 to 2007. The production of CDL is mainly based on moderate resolution satellite imagery and extensive agricultural ground truth (Boryan et al., 2011). The misclassified pixels in the CDL refer to the pixels that are covered with "clouds" or "no data". These pixels are mainly existing in the CDL products before 2006 due to lack of high-quality satellite data and the algorithm limitation back then. Examples of the misclassified pixels in the early-year CDL are shown in Figure 1.

The CDL data products are freely downloaded from CropScape (https://nassgeodata.gmu.edu/CropScape/), which is developed and maintained in cooperation with Center for Spatial Information Science and Systems of George Mason University (Han et al., 2012; Zhang et al., 2019c). It provides an easy-to-use Web GIS application to visualize, analyse, and download CDL data. All data hosted on CropScape are disseminated via the OGC standards-compliant geospatial Web services, such as Web Map Service (WMS), Web Coverage Service (WCS), Web Feature Service (WFS), and Web Processing Service (WPS).

2.2 Study Area

The Agricultural Statistics District (ASD) #1760 of Illinois state is selected as the study area. The study area lies on the Central Corn Belt Plains Ecoregion, which is mainly covered by corn, soybeans, grassland, and forest as shown in Figure 2. It can be seen that the 2005 CDL contains a considerable number of pixels are labelled as "clouds or no data" over the study area. The purpose of this study is to restore those misclassified pixels in the study area of 2005 CDL using the machine-learned crop sequence model.

2.3 Machine Learning Framework

To automatically correct the misclassified pixels in CDL, an end-to-end machine learning framework is proposed in this paper. The proposed framework is composed of four major components: data preparation, model training, classification, and evaluation.

2.3.1 Data Preparation: In data preparation, the CDLs from 2006 to 2018 are stacked sequentially to form CDL time series. 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. Training and validation data sets are randomly sampled from the "good pixels" in the study area and labelled with 2005 CDL. The experiment data set includes all pixels corresponding to those misclassified pixels in the study area without labels.

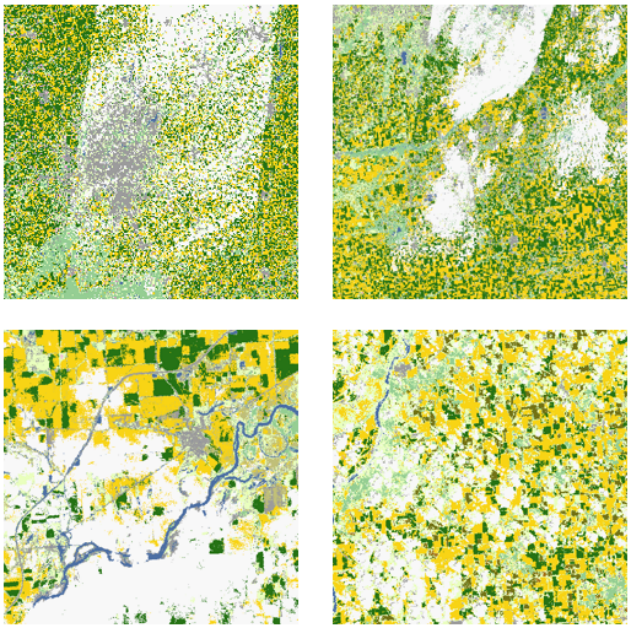

Figure 1. Examples of misclassified pixels in the early-year CDL data.

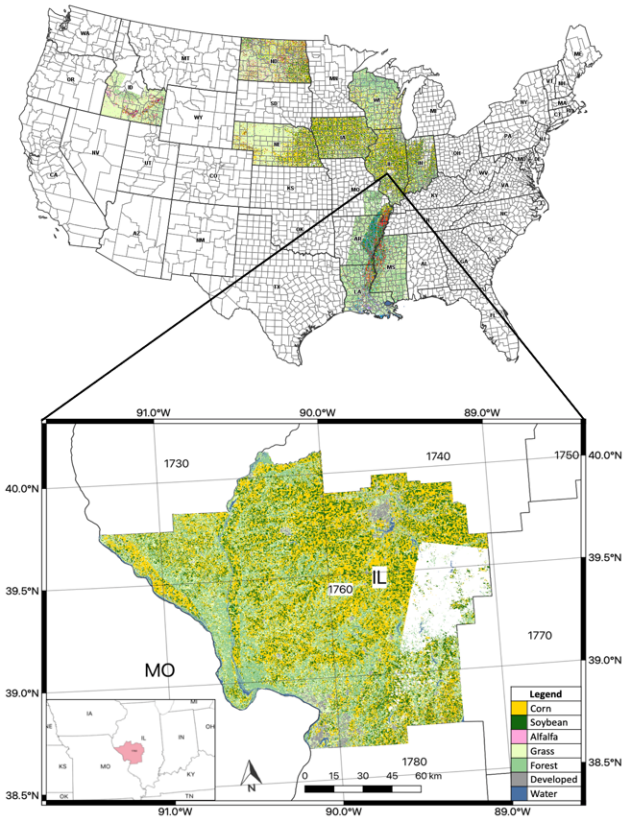

Figure 2. Study area with 2005 CDL as the experiment data (data available from CropScape).

2.3.2 Model Training: The crop sequence model is trained by feeding the training set into the artificial neural network, which contains one input layer, multiple hidden layers, and one output layer. The input layer contains a group of neurons corresponding to the same pixel of the CDL time series. Each input pixel represents a specific value of its crop type. There are mul-

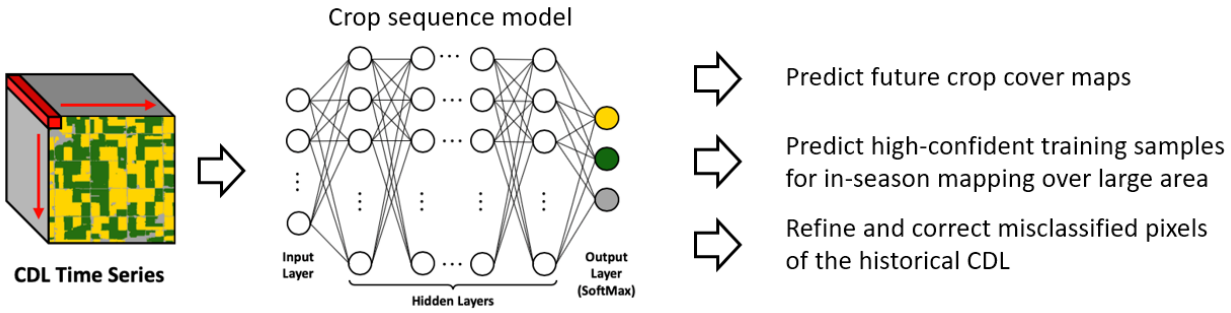

Figure 3. Applications of machine learning-based crop sequence model.

tiple hidden layers between the input layer and the output layer. The output layer uses SoftMax to estimate the probability of each crop type.

2.3.3 Classification and Validation: By feeding the experiment data set to the well-trained crop sequence model, the misclassified pixels in the original CDL can be refined. To validate the refinement performance, we applied the same crop sequence model to the validation set. Then we measured the model by calculating the agreement of the classified label and the original label of the validation set.

The applications of the proposed machine learning-based crop sequence model are illustrated in Figure 3. In this study, the crop sequence model is used to restore the historical crop cover map. This model, on the other hand, can be also applied to predict the future crop cover maps with the high-confident training samples for early-season and in-season crop mapping.

3. RESULTS

The refined 2005 CDL data of ASD #1760 is illustrated in Figure 4. Comparing the refined result with the original 2005 CDL data, we observed that the misclassified pixels had been corrected with the crop sequence information learned from the historical CDL time series.

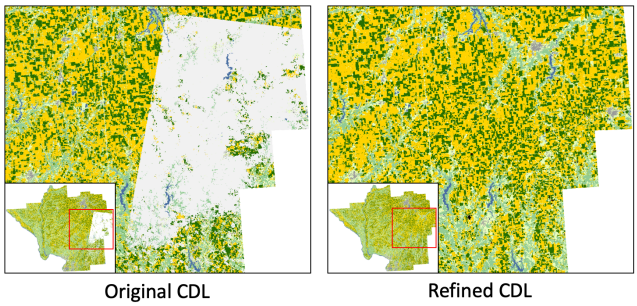

Figure 4. Refined 2005 CDL of ASD #1760.

The overall accuracy of the refined pixels is unable to be accessed directly due to lack of ground reference data. Instead, we utilized the validation data set, derived from the "good pixels" in the study area of 2005 CDL to indirectly measure the performance of the model. The overall accuracy of validation based on the validation sample set is over 85%. Therefore, the actual overall accuracy of the refined pixels may vary. To further validate the performance of refinement, the ground reference data are required.

4. CONCLUSION

This study investigated the feasibility of using machine learning technology to refine CDL data. An end-to-end ANN-based framework was proposed and tested to correct the misclassified pixels in the historical CDL data. The preliminary experiment result indicates that the misclassified pixels over the ASD #1760 could be corrected with reasonable accuracy (>85%). The findings suggest that the proposed machine learning approach is effective and low-cost for correcting the misclassified pixels, and has great potential for refining the historical CDL over large geographic area. More experiments and validation will be conducted in the future.

ACKNOWLEDGEMENTS

This work is supported by the U.S. Department of Agriculture National Agricultural Statistics Service.

REFERENCES

Boryan, C., Yang, Z., Mueller, R., Craig, M., 2011. Monitoring US Agriculture: The US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. *Geocarto International*, 26(5), 341–358. doi:10.1080/10106049.2011.562309.

Di, L., Yu, E. G., Kang, L., Shrestha, R., Bai, Y., 2017. RF-CLASS: A Remote-Sensing-Based Flood Crop Loss Assessment Cyber-Service System for Supporting Crop Statistics and Insurance Decision-Making. *Journal of Integrative Agriculture*, 16(2), 408-423. doi:10.1016/S2095-3119(16)61499-5.

Farmaha, B. S., Eskridge, K. M., Cassman, K. G., Specht, J. E., Yang, H., Grassini, P., 2016. Rotation Impact on On-Farm Yield and Input-Use Efficiency in High-Yield Irrigated Maize–Soybean Systems. *Agronomy Journal*, 108(6), 2313-2321. doi:10.2134/agronj2016.01.0046.

Govaerts, B., Mezzalama, M., Unno, Y., Sayre, K. D., Luna-Guido, M., Vanherck, K., Dendooven, L., Deckers, J., 2007. Influence of Tillage, Residue Management, and Crop Rotation on Soil Microbial Biomass and Catabolic Diversity. *Applied Soil Ecology*, 37(1), 18-30. doi:10.1016/j.apsoil.2007.03.006.

Han, W., Yang, Z., Di, L., Mueller, R., 2012. CropScape: A Web Service Based Application for Exploring and Disseminating US Conterminous Geospatial Cropland Data Products for Decision Support. *Computers and Electronics in Agriculture*, 84, 111–123. doi:10.1016/j.compag.2012.03.005.

Hao, P., Wang, L., Zhan, Y., Wang, C., Niu, Z., Wu, M., 2016. Crop Classification Using Crop Knowledge of the Previous-Year: Case Study in Southwest Kansas, USA. *European Journal of Remote Sensing*, 49(1), 1061-1077. doi:10.5721/EuJRS20164954.

Hao, P., Zhan, Y., Wang, L., Niu, Z., Shakir, M., 2015. Feature Selection of Time Series MODIS Data for Early Crop Classification Using Random Forest: A Case Study in Kansas, USA. *Remote Sensing*, 7(5), 5347-5369. doi:10.3390/rs70505347.

Karlen, D. L., Cambardella, C. A., Kovar, J. L., Colvin, T. S., 2013. Soil Quality Response to Long-Term Tillage and Crop Rotation Practices. *Soil and Tillage Research*, 133, 54-64. doi:10.1016/j.still.2013.05.013.

Karlen, D. L., Hurley, E. G., Andrews, S. S., Cambardella, C. A., Meek, D. W., Duffy, M. D., Mallarino, A. P., 2006. Crop Rotation Effects on Soil Quality at Three Northern Corn/Soybean Belt Locations. *Agronomy Journal*, 98(3), 484-495. doi:10.2134/agronj2005.0098.

Lark, T. J., Salmon, J. M., Gibbs, H. K., 2015. Cropland Expansion Outpaces Agricultural and Biofuel Policies in the United States. *Environmental Research Letters*, 10(4), 044003. doi:10.1088/1748-9326/10/4/044003.

Liknes, G. C., Nelson, M. D., Gormanson, D. D., Hansen, M., 2009. The Utility of the Cropland Data Layer for Forest Inventory and Analysis. *Proceedings of the Eighth Annual Forest Inventory and Analysis Symposium*, 259–264. doi:10.2737/WO-GTR-79.

Parajuli, P. B., Jayakody, P., Sassenrath, G. F., Ouyang, Y., Pote, J. W., 2013. Assessing the Impacts of Crop-Rotation and Tillage on Crop Yields and Sediment Yield Using a Modeling Approach. *Agricultural Water Management*, 119, 32-42. doi:10.1016/j.agwat.2012.12.010.

Pikul, J. L., Carpenter-Boggs, L., Vigil, M., Schumacher, T. E., Lindstrom, M. J., Riedell, W. E., 2001. Crop Yield and Soil Condition under Ridge and Chisel-Plow Tillage in the Northern Corn Belt, USA. *Soil and Tillage Research*, 60(1), 21-33. doi:10.1016/S0167-1987(01)00174-X.

Sahajpal, R., Zhang, X., Izaurralde, R. C., Gelfand, I., Hurtt, G. C., 2014. Identifying Representative Crop Rotation Patterns and Grassland Loss in the US Western Corn Belt. *Computers and Electronics in Agriculture*, 108, 173-182. doi:10.1016/j.compag.2014.08.005.

Schönhart, M., Schmid, E., Schneider, U. A., 2011. CropRota – A Crop Rotation Model to Support Integrated Land Use Assessments. *European Journal of Agronomy*, 34(4), 263-277. doi:10.1016/j.eja.2011.02.004.

Secchi, S., Kurkalova, L., Gassman, P. W., Hart, C., 2011. Land Use Change in a Biofuels Hotspot: The Case of Iowa, USA. *Biomass and Bioenergy*, 35(6), 2391-2400. doi:10.1016/j.biombioe.2010.08.047.

Temperly, R. J., Borges, R., 2006. Tillage and Crop Rotation Impact on Soybean Grain Yield and Composition. *Agronomy Journal*, 98(4), 999-1004. doi:10.2134/agronj2005.0215.

Thompson, A. W., Prokopy, L. S., 2009. Tracking Urban Sprawl: Using Spatial Data to Inform Farmland Preservation Policy. *Land Use Policy*, 26(2), 194-202. doi:10.1016/j.landusepol.2008.02.005.

Van Eerd, L. L., Congreves, K. A., Hayes, A., Verhallen, A., Hooker, D. C., 2014. Long-Term Tillage and Crop Rotation Effects on Soil Quality, Organic Carbon, and Total Nitrogen. *Canadian Journal of Soil Science*, 94(3), 303-315. doi:10.4141/cjss2013-093.

Zhang, C., Di, L., Lin, L., Guo, L., 2019a. Extracting Trusted Pixels from Historical Cropland Data Layer Using Crop Rotation Patterns: A Case Study in Nebraska, USA. *2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, 1–6. doi:10.1109/Agro-Geoinformatics.2019.8820236.

Zhang, C., Di, L., Lin, L., Guo, L., 2019b. Machine-Learned Prediction of Annual Crop Planting in the U.S. Corn Belt Based on Historical Crop Planting Maps. *Computers and Electronics in Agriculture*, 166, 104989. doi:10.1016/j.compag.2019.104989.

Zhang, C., Di, L., Yang, Z., Lin, L., Yu, E. G., Yu, Z., Rahman, M. S., Zhao, H., 2019c. Cloud Environment for Disseminating NASS Cropland Data Layer. *2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, 1–5. doi:10.1109/Agro-Geoinformatics.2019.8820465.