IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium | USGov | DOI: 10.1109/IGARSS53475.2024.10641079

CLOUD-POWERED AGRICULTURAL MAPPING: A REVOLUTION TOWARD 10M RESOLUTION CROPLAND DATA LAYERS

Zhe Li, Rick Mueller, Zhengwei Yang, David Johnson, and Patrick Willis

National Agricultural Statistics Service
United States Department of Agriculture
Washington DC 20250, USA

ABSTRACT

The transition from 30m to 10m resolution in the annual Cropland Data Layers (CDL) is imperative for enhanced national-scale assessments. This paper addresses the computational challenges posed by this transition, emphasizing the need for increased efficiency in data preparation, image processing and improved crop classification accuracy. Leveraging Google Earth Engine (GEE) cloud-based platform and advanced machine learning algorithms, this paper introduces a novel operational workflow to swiftly generate a 10-meter resolution CDL for 2022 across the conterminous United States (CONUS). The workflow includes the development and use of Sentinel-2 and Landsat 8/9 derived multi-sensor gap-filled 10-day image composites, additional ancillary variables, tile-based localized analysis, and stratified random sampling strategy, leading to improved accuracy for major and specialty crops. This workflow not only reduces labor requirements but also enhances decision-making processes for agriculture and policymaking stakeholders.

Index Terms— Cropland Data Layers, Google Earth Engine, Sentinel-2, Landsat 8/9, 10-meter resolution

1. INTRODUCTION

The annual Cropland Data Layers (CDL) [1], produced by the United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) at a 30-meter spatial resolution, have proven invaluable for national-scale cropland assessments across the conterminous United States. However, the recent surge in cloud computing power and widespread availability of higher resolution satellite imagery have underscored the need for a transition to 10-meter resolution national and global products [2]. This shift is exemplified by products like the 10m resolution maize and soybean map of China, the European Space Agency's (ESA) WorldCover 10m (2020-2021) products, ESA WorldCereal 10m 2021 product suite and Active Cropland 10m 2021 product suite, Esri's Sentinel-2 10m Land Use/Land Cover (2017-2021), and Google's Dynamic World (DW) 10m Land Use/Land Cover (2015-present). The finer spatial detail of 10-meter resolution offers substantial benefits for crop mapping, crop area estimation, and field size quantification, enabling a more precise identification of land cover features. Recognizing the imperative to empower a broader range of stakeholders and facilitate informed decision-making for sustainable and productive agricultural practices, there is a compelling demand for the creation of a 10-meter resolution CDL.

However, relying on federal centralized Citrix resources rather than leveraging commercial cloud infrastructure or supercomputing capabilities, the current 30m CDL process involves extensive data preparation, image processing, and requires extensive human capital resources. Transitioning from 30m to 10m resolution increases computational burden ninefold, exacerbating the challenges related to data preparation, processing, and storage capacity. While the existing 30m CDL captures major commodity crops nationwide, there is room for enhancing process efficiency and crop classification accuracy, particularly for small-area, unique, and specialty crops. To tackle these multifaceted challenges, this paper introduces a novel CDL operational workflow. This workflow incorporates Sentinel-Landsat multi-sensor gap-filled 10-day image composites, advanced machine learning algorithms and sampling strategies utilizing Google Earth Engine (GEE) [3] for the efficient generation of a 10-meter resolution CDL for the year 2022 across the CONUS.

2. DATA

GEE is a cloud-based platform revolutionizing remote sensing analysis by providing imagery processing libraries, machine learning algorithms, and access to extensive satellite imagery and geospatial datasets. Leveraging Google's computational infrastructure, GEE processes and analyzes large-scale Earth observation data seamlessly, eliminating the need to download and manage massive datasets on the computationally limited federal platform. This study utilized GEE-ingested surface reflectance data from harmonized Sentinel-2 MSI Level-2A, Landsat 8, and Landsat 9 Level 2

Collection 2 Tier 1 data. Spectral bands "GREEN," "RED," "NIR," "SWIR1," "SWIR2," and "RedEdge1" offered by the different sensors were selected. To generate cloud-free images and reduce computational intensity without sacrificing the ability to depict crop phenology, we generated 10-day median surface reflectance time-series image composites [4] and Normalized Difference Vegetation Index (NDVI) composites at 10-meter resolution covering the 2022 growing season (April to October) for crop classification. In each image composite, we filled gaps (if any) using the mean values of preceding and subsequent composites. To better distinguish between the agricultural and non-agricultural categories, additional ancillary variables including the "impervious" band from USGS NLCD 2021, the binary layers of grassland, shrubland, and tree cover extracted from the ESA WorldCover 10m 2021 product, and an elevation layer from US Geological Survey 3DEP 10m National Digital Elevation Program were included in the image classification. The ground reference data for model training and validation were derived by integrating USDA Farm Service Agency (FSA) Common Land Unit (CLU) data and additional crop-specific ground reference data obtained from non-FSA sources for agricultural categories, as well as USGS NLCD 2021 land cover data for non-agricultural categories.

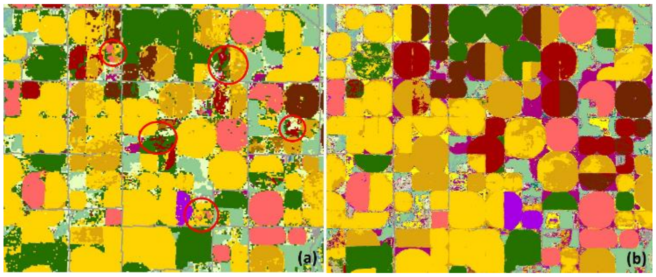

Figure 1. Crop types identified in Hale County, Illinois near Swan Lake by (a) Current 30m CDL, and (b) New 10m CDL.

3. METHODOLOGY

3.1. Sampling Strategies

In current operational CDL model training and classification, the See5 Decision Tree (DT) algorithm is employed and processed by state. However, the random sampling method adopted by the current CDL does not directly transfer it to a stratified manner and this may lead to over-sampling of predominant crops like corn and soybeans, while under-sampling or missing small-area/specialty crops. As depicted in Figure 1, crops within the red circles on 1(a) were misclassified as corn, soybeans, and others due to insufficient or missing samples representing local small-area crops such as Dry Beans (bright red), Potatoes (brown) and Sugar Beets (purple) in this example, as shown in 1(b). To address these issues more effectively, this study implemented a stratified random sampling (SRS) strategy and conducted localized analysis [5] by the Sentinel-2 tile.

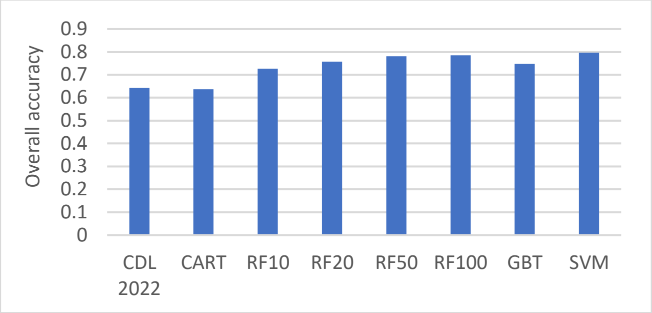

Figure 2. Performance of different machine learning algorithms for crop type classifications. The numbers following "RF" represent the number of trees configured for the Random Forest model.

3.2. Classification Algorithms

To identify the optimal classification algorithm for national-scale operational crop mapping, we evaluated several top-ranked machine learning algorithms: Random Forest (RF) [6], Gradient Boost Tree (GBT) [7], and Support Vector Machine (SVM). In Figure 2, we present the performance of these algorithms in classifying crop types using the 10-day 10m Sentinel-Landsat image composites in a pilot site in Illinois. Results indicate that CART (Classification and Regression Trees) achieved an accuracy equivalent to CDL 2022, as expected, given the same classification algorithm and training data were used. For RF, classification overall accuracies improved as the number of trees configured, peaking at 100 trees. Although SVM and GBT achieved similar accuracies to RF, our study selected the Random Forest algorithm due to its decent accuracies and significantly reduced processing time. Specifically, we utilized the "smileRandomForest" function in Google Earth Engine, configured with 50 trees and a bagFraction of 0.5 for crop classification.

3.3. Model Training and Validation

The entire CONUS was covered by 967 tiles. With the SRS strategy, individual sample datasets were created for each of the tiles, with up to 2,000 samples per crop type. Each sample dataset was divided into 80% for training and 20% for validation at the tile level. Then the 20% samples within each tile were merged by state and used for validation at the state level. The national-level validation used a total of 77,031 samples randomly drawn from the validation samples from all 967 tiles. This guaranteed that the validation samples used for accuracy assessment at all three levels were completely separated from the training samples. Metrics used for accuracy assessment included Overall Accuracy (OA), Kappa coefficient, Producer's Accuracy (PA), User's Accuracy (UA), and F1-score (F1).

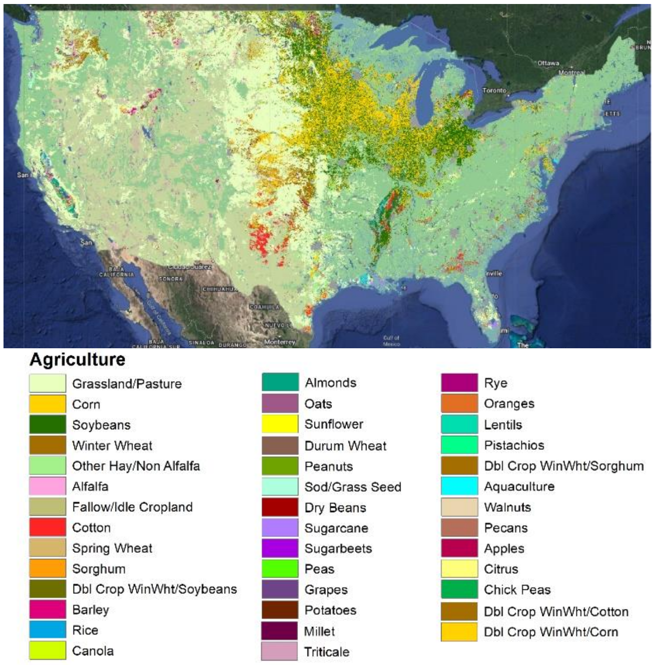

Figure 3. The new 10-meter resolution Cropland Data Layer 2022 generated by Google Earth Engine (The legend only shows the top 41 crop covers by decreasing acreage).

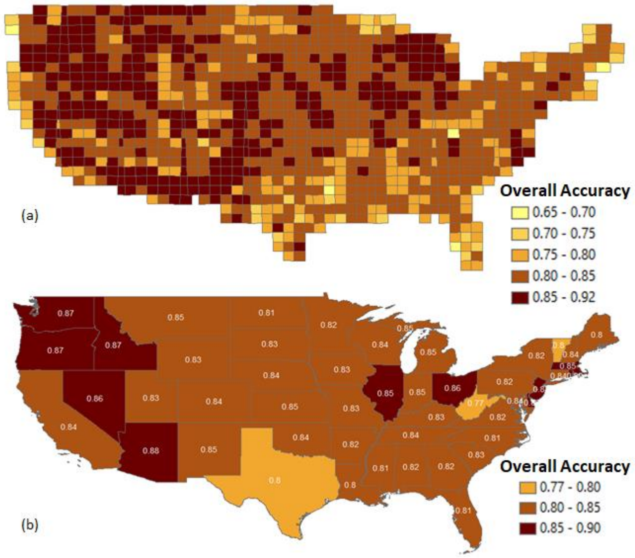

Figure 4. Overall accuracies of the new 10m CDL assessed by (a) tile, and (b) state.

4. RESULTS AND DISCUSSION

Figure 3 illustrates the new workflow-generated, 10-meter resolution CDL 2022 containing 110 agricultural and 14 non-agricultural categories using GEE. At the tile level, overall accuracies for all 967 tiles ranged from 65.9% to 92.1% (Figure 4a), averaging 83% (0.82 for Kappa) with a standard deviation (SD) of 0.04. State-level accuracies varied from 76.9% to 89.2% (Figure 4b), with an average of 83.6% and a SD of 0.02. The nation-level validation achieved an overall accuracy of 81.15% and a Kappa of 0.808 for 124 land cover types (110 agricultural categories and 14 non-agricultural categories). F-1 scores for the 110 crop types ranged from 0.692 to 1, with an average of 0.936 and SD of 0.1. F-1 scores for the 20 principal crops ranged from 0.825 to 0.976, averaging 0.923 (Figure 5), among which, corn, soybeans, wheat, cotton, and sorghum were 0.869, 0.825, 0.898, 0.852 and 0.904 respectively. F-1 scores for the rest of 88 crops (Alfalfa and Non-Alfalfa excluded) achieved an average score of 0.943 (Figure 5).

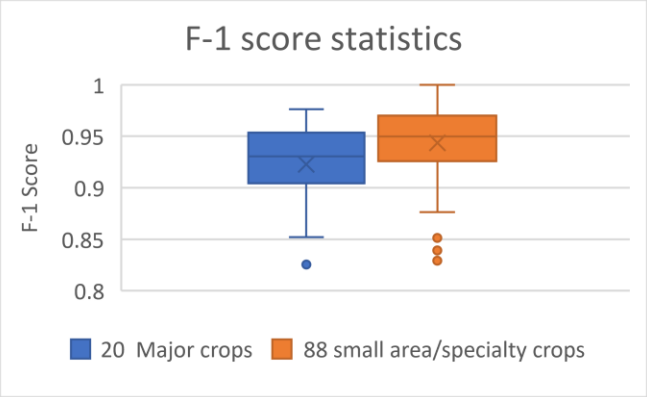

Figure 5. National-level validation (F-1 scores) of the new 10m CDL using 77,031 stratified samples for the 20 major crops and the 88 small-area/specialty crops.

To further assess the effectiveness of the new methodology, we selected three counties known for highly diversified crops: Mason County, Illinois; Hale County, Texas; and Monterey County, California. The assessment was conducted using all available ground-reference samples within each county. In Figure 6, we observe that the new 10m CDL achieved higher F-1 scores for many crop classes compared to the current 30m CDL in Mason and Hale counties. Classes such as Winter Wheat, Sugar Beets, Dry Beans, Double Crop Soybeans/Oats, Cabbage, Triticale, Double Crop Winter Wheat/Sorghum, Double Crop Winter Wheat/Cotton, and Turnips showed significantly higher scores. More importantly, some classes like Peppers, Greens, Squash, Cucumbers and Double Crop Winter Wheat/Corn were successfully mapped only by the new 10m CDL. In Monterey County, California, where ground-reference data was scarce, both the 30m and 10m CDLs yielded lower F-1 scores, with the former showing slightly higher scores for most classes. However, the new 10m CDL exhibited more homogenous patterns within crop parcels with cleaner defined boundaries in areas without ground-reference data, effectively reducing salt-pepper and boundary-aliasing effects (Figure 7i), compared to the current 30m CDL (Figure 7e). Further comparisons between the two data products are illustrated in

Figure 7 Columns 2, 3 and 4, showcasing highly diversified and complex cropping patterns at various locations.

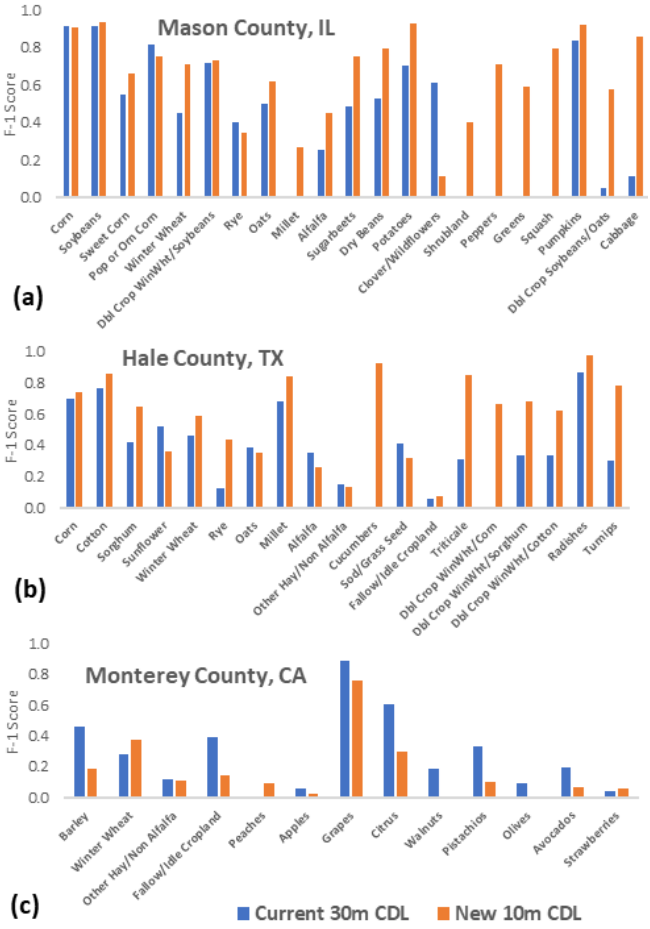

Figure 6. Comparisons of F-1 scores for classified crop types in (a) Mason County, Illinois, (b) Hale County, Texas, and (c) Monterey County, California.

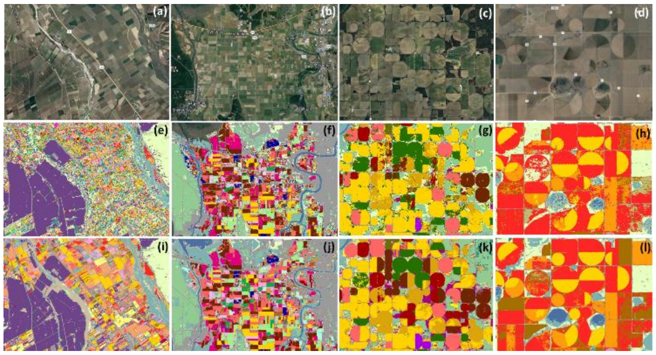

Figure 7. Comparisons of classified crop types at various locations. Row 1: Google Earth high-resolution background images; Row 2: Current 30m CDL 2022; and Row 3: New 10m CDL 2022. Column 1: Salinas Valley, California; Column 2: West of the Skagit River, Northwestern Washington; Column 3: Mason County Illinois close to Swan Lake; and Column 4: Northern Texas close to Amarillo.

5. CONCLUSIONS

By integrating Sentinel-Landsat multi-sensor gap-filled 10-day image composites, advanced machine learning algorithms and sampling strategies, this paper introduces a novel workflow for generating 10-meter resolution CDL using GEE, resulting in substantial labor and workload reductions with improved crop accuracy and spatial clarity for minor/specialty crops. The new GEE generated 10m CDL better captured the full diversity of crop types in regions with unique or specialty crops. The enhanced accuracy and detail facilitate improved decision-making for stakeholders in agriculture and policy making.

6. REFERENCES

[1] 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, pp. 341–358, 2011.

[2] Z.S. Venter, D.N. Barton, T. Chakraborty, T. Simensen, G. Singh, “Global 10 m Land Use Land Cover Datasets: A Comparison of Dynamic World, World Cover and Esri Land Cover”, *Remote Sensing,* 14, 4101. 2022.

[3] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, R. Moore, “Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone”, *Remote Sensing of Environment*, v. 202, pp. 18–27, 2017.

[4] A. Htitiou, A. Boudhar, A. Chehbouni, T. Benabdelouahab, “National-Scale Cropland Mapping Based on Phenological Metrics, Environmental Covariates, and Machine Learning on Google Earth Engine”, *Remote Sensing* (Basel), v. 13, no. 21, pp. 4378, 2021.

[5] F. Xuan, Y. Dong, J. Li, X. Li, W. Su, X. Huang, J. Huang, Z. Xie, Z. Li, H. Liu, W. Tao, Y. Wen, Y. Zhang, “Mapping crop type in Northeast China during 2013–2021 using automatic sampling and tile-based image classification”, *International Journal of Applied Earth Observation and Geoinformation*, vol. 117, 2023.

[6] L. Breiman, “Random Forests.” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] R. Lawrence, A. Bunn, S. Powell, M. Zambon, “Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis”, *Remote Sensing of Environment*, v. 90, no. 3, pp. 331–336, 2004.