Identifying Causes of Irrigated Cropland Misclassification

Cella Schnabel
Srikrishnan Research Group @ Cornell BEE

Spring 2024

Abstract

A reliable and complete global dataset of irrigated land would be ideal for understanding the impact of agricultural practices on sustainable water management – a crucial aspect of responding to shifts in land and water use due to climate change. While global agricultural land area has been largely well characterized, the distinction between irrigated and rainfed lands is often ignored. Several datasets to identify irrigated land in the United States have been developed using remote sensing, but it is unclear how well those methods might generalize to other regions. A critical question for expanding these datasets globally is what environmental conditions may cause misclassification. In this project, I conducted a thorough literature review to understand the limitations of irrigated cropland analysis, due primarily to distinguishing irrigation systems in coarser satellite imagery. My work follows the 2017 Landsat-based Irrigation Dataset for the United States (LANID-US) produced by Yanhua Xie et al. [1], in which they provide sample coordinates of ground-truth land identified as irrigated or rainfed in the conterminous United States (CONUS). I developed a random forest classifier in Google Earth Engine to identify irrigated vs. rainfed cropland across CONUS by attaching the LANID coordinates to a Landsat pixel to train the classifier with geographical data. In preliminary analysis, the classified image is highly consistent with published irrigation datasets for the U.S., and the classification is most influenced by geographic location and Short-Wave Infrared Band 1 (SWIR 1). This is reasonable because SWIR 1 is sensitive to soil moisture, and geographic location plays a pivotal role in shaping environmental conditions like climate and soil properties that impact irrigation strategies. This classifier may generalize poorly to extremely moist or arid regions, but more feature manipulation is required. Quantifying the probability of misclassification and systematic biases can point to breakdowns in generalizing irrigated cropland classification, to provide a starting point for a more skilled global dataset of irrigated cropland.

Introduction

Global change is putting extraordinary pressure on global croplands. While global agricultural land has been well distinguished from other land covers, the distinction between irrigated cropland and rainfed agricultural land– in this case, defined as cropland that relies solely on rainfall as a water source– is often ignored. Quantifying global irrigated cropland at high resolution can inform policy-making, minimize overuse or misuse of water, and contribute to overall environmental and economic sustainability and climate resilience of agricultural systems [2].

Irrigation consumes roughly 90% of global freshwater resources [3] and, as of 2018, approximately 40% of the world's food is cultivated on artificially irrigated land. This number is expected to increase by nearly 20% by 2050 [4]. Due to the increasingly limited availability of freshwater, a major challenge in global agriculture and global water use is to mitigate the misuse of water while continuing to meet increasing demands for food exacerbated by climate change and population growth [5]. High resolution information about where irrigated land is located, the distribution of local water demand, and how this pattern is changing over regions and time works to inform management policies and climate smart agriculture solutions.

Several datasets to distinguish irrigated land in the United States using census statistics [6] or remote sensing data [1] or both [7] have been developed. In this project, we are particularly interested in studying patterns of remote sensing based misclassification, which could help identify which areas of the world might be particularly challenging to classify in this manner, as it is currently unclear how well these methods may generalize when ground-based data isn't available. A critical question for generalizing remote sensing based classifiers is which environmental conditions result in misclassification. In equal and opposite force, this begs the question: what land features are most important in irrigated land classifying?

This study uses Shapley values to break down random forest classification of Landsat-based images. Shapley values yield class-level feature importance to identify the most prominent features that determine the class value [8]. I used the Shapley additive explanations (SHAP) [9] to better understand feature importance in cropland classification. The intention is to discover which variables consistently stand out in particular climates.

Materials and Methods

The region of interest in this study is the continental United States (CONUS), with more than 58 million acres of irrigated cropland, according to The 2017 Census of Agriculture [10].

Croplands can fluctuate immensely over a given growing season, but at any given instance, the heterogeneity of croplands in a given time stamp is a major barrier of accurate classification. It's crucial to include time series data to a classifier. However, due to the brief nature of this project, so far I have simply examined a single median composite Landsat 8 image from USGS Landsat 8 Collection 2 Tier 1 TOA Reflectance that is meant to optimize cloud coverage. I used data from June 15th to July 15th, 2017. This year was selected because the LANID-US map provides ground reference locations in the eastern CONUS until 2017 [1]. I cropped this median composite Landsat image to mask out all land covers except for cropland; using band 15 of the 2020 North American Land Cover 30-meter dataset [11]. I selected this land cover dataset despite the time inconsistency because of it uses Landsat at 30m resolution, instead of MODIS at 250m resolution.

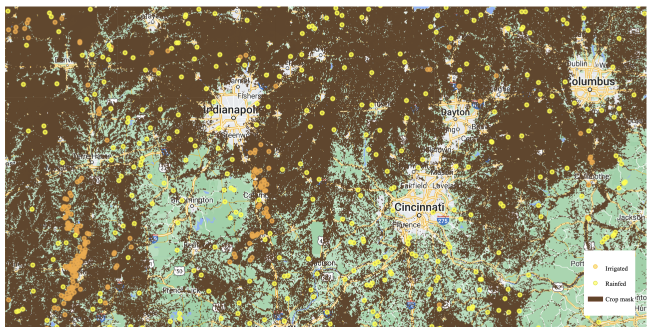

Figure 1: Sample of materials used

I then attached the Landsat bands to the sample coordinates provided in LANID-US of the eastern CONUS[12]. The sample locations include center pivot fields (irrigated) and stable rainfed fields. I labeled Landsat images within a 30m radius of the given coordinates a class: 1 for irrigated, 0 for rainfed. Using the data structure where the Landsat bands and geographic coordinates make up the features, I split the 30m buffered sample coordinates 80:20 and input the training 80 to a random forest classifier.

I selected

I attached these

Annual 30-m resolution irrigation maps, derivative products, and ground reference locations for the United States

Given the cryptic nature of irrigated croplands, a variety of remote sensing features and variables have been evaluated to characterize these areas, including band metrics

Using very crude minimal inputs and still getting pretty high accuracy is inspiring for this strategy. There seem to be

many avenues for adding data that may improve performance.

The heterogeneous nature of croplands in space and time is a major challenge for accurate classifications. In an effort to combat this

is to was to identify the most parsimonious combination of features that could reasonably map a particular class and enhance overall map accuracy. Shapley values were developed for used as

A prediction can be explained by assuming that each feature value of the instance is a "player" in a game where the prediction is the payout. Shapley values – a method from coalitional game theory – tells us how to fairly distribute the "payout" among the features.

This is intended to uncover the causes of classification. The limitations of irrigation cropland analysis primarily stem from the challenges of due primarily to distinguishing irrigation systems in coarser satellite imagery. My work follows the 2017 Landsat-based Irrigation Dataset for the United States (LANID-US) produced by Yanhua Xie et al., in which they provide sample coordinates of ground-truth land identified as irrigated or rainfed in the conterminous United States (CONUS). After developing a random forest classifier in Google Earth Engine, this study compares a naively classified image to published irrigation maps to identify sources of misclassification of Landsat-based image pixels between the images. This study proposes potential sources of misclassification based on Landsat-based imagery band properties.

Materials and Methods

IrrMapper [13]

Results

Discussion

Conclusion

$$NDMI = \frac{B5 - B6}{B5 + B6}$$

$$NDVI = \frac{B5 - B4}{B5 + B4}$$

References

[1] Yanhua Xie, Tyler J. Lark, Jesslyn F. Brown, and Holly K. Gibbs. Mapping irrigated cropland extent across the conterminous united states at 30·m resolution using a semi-automatic training approach on google earth engine. *ISPRS Journal of Photogrammetry and Remote Sensing*, 155:136–149, 2019. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2019.07.005. URL `https://www.sciencedirect.com/science/article/pii/S0924271619301728`.

[2] Prasad S. Thenkabail, Munir A. Hanjra, Venkateswarlu Dheeravath, and Muralikrishna Gumma. A holistic view of global croplands and their water use for ensuring global food security in the 21st century through ad-

vanced remote sensing and non-remote sensing approaches. *Remote Sensing*, 2(1):211–261, 2010. ISSN 2072-4292. doi: 10.3390/rs2010211. URL https://www.mdpi.com/2072-4292/2/1/211.

[3] Igor A. Shiklomanov. Appraisal and assessment of world water resources. *Water International*, 25(1):11–32, 2000. doi: 10.1080/02508060008686794. URL https://doi.org/10.1080/02508060008686794.

[4] Hervé Plusquellec and Sanjiva Cooke. A retrospective of lending for irrigation - reflections on 70 years of bank experience (english). *Discussion Paper Washington, D.C.*, 2024. URL http://documents.worldbank.org/curated/en/099757003282448162/IDU162ed994f17c69149bb1bcd21c19f282a40c2.

[5] Robert I. McDonald, Pamela Green, Deborah Balk, Balazs M. Fekete, Carmen Revenga, Megan Todd, and Mark Montgomery. Urban growth, climate change, and freshwater availability. *Proceedings of the National Academy of Sciences*, 108(15):6312–6317, 2011. doi: 10.1073/pnas.1011615108. URL https://www.pnas.org/doi/abs/10.1073/pnas.1011615108.

[6] Aaron Hrozencik, Grant Gardner, Nicholas Potter, and Steven Wallander. Irrigation organizations: Groundwater management. *Economic Research Service*, EB-34, 2023. doi: https://doi.org/10.32747/2023.7975545.ers.

[7] Jesslyn F. Brown and Md Shahriar Pervez. Merging remote sensing data and national agricultural statistics to model change in irrigated agriculture. *Agricultural Systems*, 127:28–40, 2014. ISSN 0308-521X. doi: https://doi.org/10.1016/j.agsy.2014.01.004. URL https://www.sciencedirect.com/science/article/pii/S0308521X14000134.

[8] Lloyd S Shapley et al. A value for n-person games. *Princeton University Press Princeton*, 1953.

[9] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Curran Associates, Inc.*, 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[10] USDA. Irrigation & water use. *Economic Research Service*, 2019.

[11] Commission for Environmental Cooperation. North american land cover, 2020 (landsat, 30m). *North American Land Change Monitoring System*, 2019.

[12] Yanhua Xie and Tyler Lark. LANID-US: Landsat-based Irrigation Dataset for the United States, October 2021. URL https://doi.org/10.5281/zenodo.5548555.

[13] David Ketchum, Kelsey Jencso, Marco P. Maneta, Forrest Melton, Matthew O. Jones, and Justin Huntington. Irrmapper: A machine learning approach for high resolution mapping of irrigated agriculture across the western u.s. *Remote Sensing*, 12(14), 2020. ISSN 2072-4292. URL `https://www.mdpi.com/2072-4292/12/14/2328`.