Prediction of Crop Phenology in the Northwest

HDSI Agri Datathon 2024
Team Butter Bros
October 6, 2024

Nora Amer
Boston University
SEC, 150 Western Ave
Allston, MA 02134
amernor@bu.edu

Jason Courtois
UMass Boston
SEC, 150 Western Ave
Allston, MA 02134
jcourtois1293@gmail.com

Jenny Tang
Boston University
SEC, 150 Western Ave
Allston, MA 02134
jetang@bu.edu

Abstract

The Northwestern US is a vital agricultural hub, contributing significantly to national and local food supplies. The drier climate and extreme temperatures allows for a diverse number of crops to grow. Despite the agricultural wealth of this area, food insecurity remains a pressing issue. This paper explores geospatial data informed predictive modeling, in order to inform communities on trends within crop phenology and it's implications on food insecurity.

1. Introduction

The Northwestern area of the United States, specifically the states of Washington, Oregon, and Idaho, is an extremely important agricultural region, not just serving local communities, but individuals across the country. Drier weather, hot summers, and colder winters can be located within areas east of the Cascade Mountains spanned through Washington and Oregon, providing a foundation for agricultural diversity. Crops such as potatoes in Idaho, apples in Washington, and various types of berries in Oregon are essential to the agriculture of the Northwest- in terms of consumption and economic activity. This sector makes up for around 13% of each of the three states' economic activity, providing thousands of jobs and resources to local communities [U.S24]. However, within the Northwest, the issue of food insecurity still stands. In Washington, 1 in 9 people face the issue of hunger [Fee24], an issue consistent throughout the Northwest, as well as the entire country.

As mentioned, the agricultural sector of the Northwest is a fundamental aspect of its own economy and food security. Thus, with the potential of inconsistencies or decreases in Northwest crop population, the following question is raised: How can one predict future crop phenology in order to inform a community on possible trends in agriculture?

To address this question, we integrate geospatial sources, time series data, and scoped insights. By employing linear regression and analyzing how crop production varies over the years in Gem County, Idaho, we aim to explore changes in crop production occurrence to ensure a better scope of the climate, of production and thus food insecurity.

2. Data and Methods

We utilized data focusing on crop phenology from 22 counties within the inland areas of Washington, Oregon, and Idaho. Using the geospatial processing services, VegScape and CropScape made by the USA NASS, we were able to access imaging of individual counties from 2006 to 2022. CropScape is a geospatial resource that indicates what a specific area of land is being used for- whether it would be a type of crop, a forest, or an industrial building. CropScape uses its own index and color to identify each type, allowing for comparison to other counties. It only releases data per year for each county. VegScape, another geospatial resource, assigns a Normalized Difference Vegetation Index (NDVI) areas within a county to measure plant health. Higher NDVI values represent healthier and denser vegetation. For every year that contains data for a county, there is one CropScape scan for the whole year, and up to 52 VegScape scans in the year. The general idea of our integration was to select a county and year that it has data, parse the CropScape file for the locations of crops, and find the corresponding NDVI values for those locations in VegScape's files to find the mean NDVI value for the month and year.

In order to perform an analysis on crop health among different counties, we needed to align the data from CropScape and VegScape. Both databases store their data in Tagged

Value	RGB	Class Name
1	(1.000, 0.827, 0.000)	Corn
2	(1.000, 0.149, 0.149)	Cotton
3	(0.000, 0.659, 0.894)	Rice
4	(1.000, 1.000, 0.000)	Sorghum

Table 1: Example of some value codes of a CropScape file, matching RGB values with different crops and vegetation

Image Files (.tif) where each pixel holds a specific value. In CropScape's files, these pixels hold a value representing the crop type and VegScape's files hold NDVI values. Since these two databases have different spatial resolutions for their files, they need to be scaled in order to be compared. All VegScape files had a spatial resolution of 250 meters [ZY13]. CropScape files from 2006 - 2009 had a spatial resolution of 56 meters, and from 2010 onwards had a resolution of 30 meters [WH12]. This led to VegScape's files having a smaller overall width and height compared to CropScape. The VegScape images for any given county were scaled up using bilinear interpolation, allowing for pixel-perfect alignment between datasets.

Once the two datasets were aligned, we first parsed through the CropScape data to get the locations of the crops for the whole year. VegScape scans had multiple scans within one month, as there are up to 52 scans per year for each week in the year. To calculate the NDVI mean, we took the average NDVI reading for each scan within the month for each crop location and took the mean of these values. This was repeated for each month in the year, and the yearly NDVI mean was found by taking the average of each month's total average.

In order to predict the average for the NDVI values for each month within the year 2023, we decided to create a predictive model using RandomForestsRegressor, a package made by sci-kit-learn. Essentially, we isolated the average NDVI values for each month of each year of Gem County. We scaled our NDVI values, using a scale we derived from the following NDVI equation.

$$NDVI = \frac{Band2 - Band1}{Band2 + Band1} \times 125 + 125$$

We interpreted this by subtracting each average by 125, and then dividing each value by 125 after, ensuring each value was between 0 and 1. Then, we used our data from 2007 to 2021 in our X variable and left out the 2022 data in our Y variable, in order to test our model using the Leave-One-Out Cross Validation method. This was important because we were working with a small dataset, only consisting of 9 years. We then split the data into training and test data,

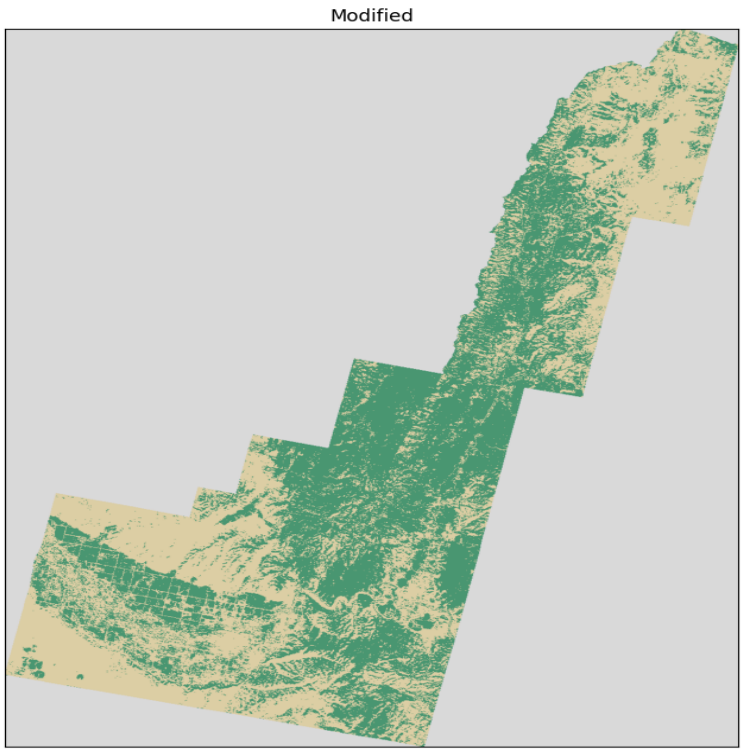

Figure 1: Modified CropScape data where crops were turned into binary. All crops are represented as green in the image

and ran our predictive model to generate our average NDVI values for 2023. We also evaluated our model using MSE, mean squared error, and RMSE, root mean squared error, essentially to gain insight on whether our predictions were close to the same trend that has historically been shown within the years, especially because we must consider seasons within our NDVI results. Finally, we visualized our results for each month of 2023 using a line plot with Seaborn, a data visualization tool. Finally, we plotted both our predictions for 2023 on a Seaborn line plot, alongside the actual documented 2023 data.

3. Results

After using our predictive model, we can see our results shown in Figure 2. Our 2023 prediction still follows the trend of higher NDVI values within the months ranging from March to July, and a much lower value within the winter months, around December, January, and February. Our MSE was equal to 0.013531718399999981, and our RMSE was equal to 0.11632591456764903, both being decently low values, signifying a decently low error within our regression model. When comparing our results to the actual NDVI values of 2023, we can see in Figure 2 that they are

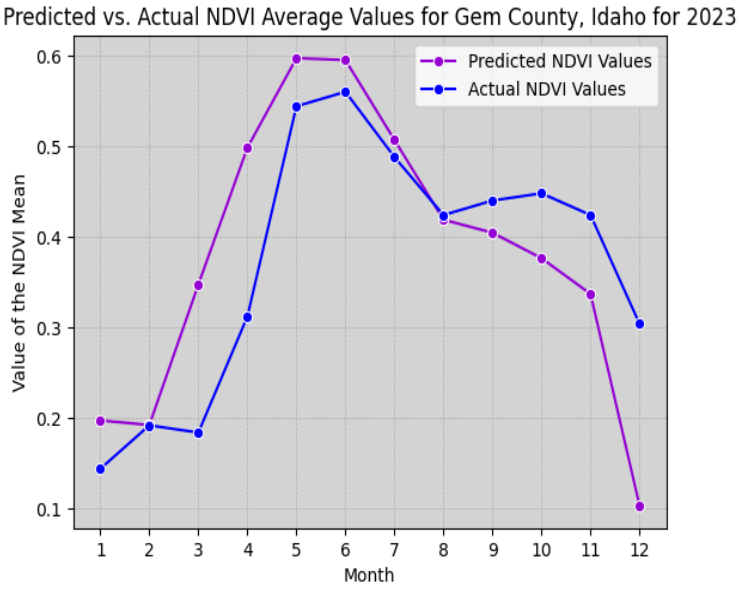

Figure 2: Our prediction of Gem County's NDVI values compared to the actual NDVI values for 2023

mostly similar. It can be acknowledged that when predicting the 2023 NDVI values of Gem, we did not have a lot of data to train our RandomForestsRegressor- which could lead to a possible case of overfitting, even though we split our data into train and test sets adequately, only having 9 non consecutive years makes a model hard to learn from. However, the predicted NDVI values were less than the real values in the first few months, but the opposite was true in the later months, as predicted values were higher instead, which could be explained by changes in climate or agricultural practices.

4. Conclusion

What we can take from this prediction is that predictive modeling of NDVI values across the northwest, is a valuable tool in anticipating what agricultural resources will possibly be produced within the next few years. This provides intel into the issue of food insecurity in the northwest. Although there are very real, systematic issues that get in the way of communities and their food security, we are able to at least know and predict the amount of vegetation and be able to anticipate that it will increase within specific areas as time goes on. Communities will be able to gain an understanding of when there is less vegetation or more vegetation and be able to prepare for it. When addressing systematic and economic impacts on food security, it is at the very least important that the food will be grown in the first place, which places all the more emphasis on supporting groups and nonprofits that reach out and make food more affordable and accessible to northwestern American communities.

Please see this link to our Submission Video.
Please see this link to our Google Colab Notebook.
Please see this link to our Other Colab Notebook.

References

[Fee24] Feeding America. Hunger in america: Washington. https://www.feedingamerica.org/hunger-in-america/washington, 2024. Accessed: 2024-10-06. 1

[U.S24] U.S. Department of Agriculture Climate Hubs. Agriculture in the northwest. https://www.climatehubs.usda.gov/hubs/northwest/topic/agriculture-northwest, 2024. Accessed: 2024-10-06. 1

[WH12] Liping Di Richard Mueller Weiguo Han, Zhengwei Yang. Cropscape: A web service based application for exploring and disseminating us conterminous geospatial cropland data products for decision support, March 2012. Received 29 August 2011; Revised 19 January 2012; Accepted 8 March 2012. 2

[ZY13] Liping Di Bei Zhang Weiguo Han Rick Mueller Zhengwei Yang, Genong (Eugene) Yu. Web service-based vegetation condition monitoring system - vegscape, July 2013. 2