Crop Phenology Analysis: A Scalable Framework for Agricultural Monitoring

HDSI Agri Datathon 2024
Three Guys
October 6th, 2024

Yuandi Tang
Northeastern University
440 Huntington Ave
Boston, MA 02115
tang.yuand@northeastern.edu

Rongjia Sun
Northeastern University
440 Huntington Ave
Boston, MA 02115
sun.rongj@northeastern.edu

Feng-Jen Hsieh
Northeastern University
440 Huntington Ave
Boston, MA 02115
hsieh.fe@northeastern.edu

Abstract

This report outlines the methodology and results from a project focused on integrating two USDA NASS geospatial services, VegScape and CropScape, to assess crop phenology in 23 counties located in the Western United States. These counties, located within Washington, Oregon, and Idaho, share similar environmental conditions shaped by the Cascade Mountains' rain shadow effect. By combining satellite data from VegScape and CropScape, we developed a scalable Python framework to isolate agricultural activity, extract NDVI values, and generate time series crop phenology reports. Our predictions of 2023 crop phenology trends were compared with actual data to better understand the dynamics of crop growth in these counties. The results shed light on the potential impacts of environmental factors on crop yield and provide insights into scalable data integration methods for agricultural research.

1. Introduction

As global concerns about food security rise, understanding crop phenology—how crops grow and develop in relation to environmental conditions—has become critical. Satellite imagery provides an efficient and accurate means of tracking agricultural activity, allowing researchers to assess how factors such as climate and land management practices affect crop yield. The USDA's VegScape and CropScape web services offer a wealth of remote sensing data, but they remain largely disjointed, limiting their utility in holistic crop phenology studies.

This project, part of the HDSI Agri Datathon 2024, aimed to integrate data from both VegScape and CropScape to create comprehensive crop phenology reports for 23 counties across the Western U.S. The region's agricultural landscape is shaped by the Cascade Mountains' rain shadow, which results in a unique climate marked by dry conditions, hot summers, and cold winters. By developing a scalable framework, our goal was to facilitate better crop monitoring and yield predictions using historical satellite imagery and NDVI (Normalized Difference Vegetation Index) values.

This report details our approach, from data preprocessing and pipeline development to the visualization of time series phenology reports, and predictions for the 2023 growing season. Ultimately, this project serves as a foundation for future work in crop monitoring and sustainable agriculture.

2. Data and Methods

The dataset used in this analysis consists of geospatial raster images and tabular data files describing cropland coverage for 23 counties over 13 years. The `.tif` files for cropland represent crop layout and classification of each year. The associated `.dbf` files contain mapping values of the land surface. The `.tif` files for NDVI images contain NDVI values for a specific county during a particular week of a given year.

2.1 Cropland Data Preparation:

The .tif files from CropScape were preprocessed to isolate active agricultural areas using binary classification: agricultural land versus non-agricultural land. We utilized the crop classification provided in the `.dbf` files for this step. Since cropland images are not available for every year, the missing years were inferred using the mean of the data from the two closest years.

2.2 NDVI Data Normalization:

NDVI values were extracted from VegScape images for areas classified as active agriculture. These NDVI values reflect the intensity of vegetation growth. NDVI values were

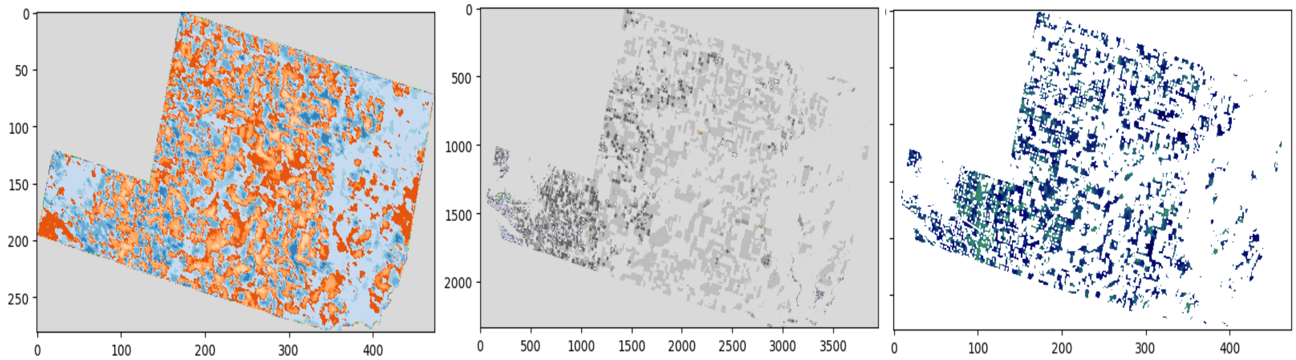

Figure 1: Using Adams County as an example to demonstrate the image preprocessing. **Left image**: Cropland image with different classifications represented by various colors. **Center image**: Filtered cropland area. **Right image**: Normalized NDVI values to a range of 0 to 1

scaled to a range of 0 to 1 by the formula below to ensure consistent time series representation. This scaling followed the guidelines provided in the VegScape documentation, where values closer to 1 indicate denser vegetation.

$$\text{Normalized NDVI} = \frac{\text{NDVI} - 125}{125} \tag{1}$$

2.3 Prediction Model:

We applied the **KMeans clustering algorithm** to group counties with similar historical NDVI trends (2001-2022) and predict future crop phenology patterns. By clustering counties based on past NDVI values, we identified regions with similar vegetation growth, allowing us to forecast the mean NDVI for each cluster in 2023. To evaluate the clustering approach, we compared the predicted NDVI values with the actual 2023 data, measuring the model's accuracy. KMeans was chosen because it allows us to identify similar crop growth patterns by clustering counties that share environmental characteristics and agricultural practices. Additionally, it simplifies prediction by grouping counties with similar data, reducing variability and improving the robustness of the prediction model.

2.4 Visualization:

For visualizing the NDVI values, we used matplotlib library in Python. Additionally, time series graphs showed the evolution of crop phenology for each county over the 13-year period.

The visualization effectively illustrates the yearly progression of vegetation, from lower NDVI values during the winter months to a peak during the summer, followed by a gradual decline. The comparison between actual and predicted NDVI for 2023 helps evaluate the accuracy of the KMeans model, providing insight into the effectiveness of historical data in predicting vegetation health.

This graphical approach allows us to not only track changes in NDVI over the years but also to identify anomalies or shifts in seasonal patterns, which could be significant for understanding long-term vegetation changes or crop phenology.

3. Results

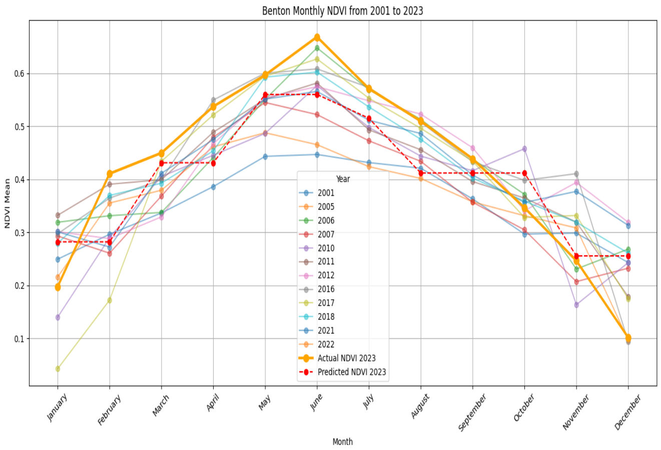

Figure 2: This line graph illustrates the monthly mean Normalized Difference Vegetation Index (NDVI) for Benton from 2001 to 2023. NDVI is a measure used to assess vegetation health and density.

The graph above illustrates the monthly Normalized Difference Vegetation Index (NDVI) mean values from the years 2001 to 2023, we take Benton county as example with the least $MSE = 0.00304$. Each line represents the average NDVI of a given year, highlighting the seasonal changes in vegetation across the dataset.

Key Elements:

1. **Historical NDVI Data (2001-2022):**

The colored lines represent the monthly NDVI values for each year from 2001 to 2022, showcasing the typical pattern of increasing vegetation from winter to summer, followed by a decline toward the end of the year.

These historical trends serve as the basis for predicting future NDVI values, enabling us to observe the variability in vegetation patterns over time.

2. **2023 NDVI Actual vs Predicted:**

The orange line represents the actual NDVI values for 2023, showing how vegetation progressed in the current year.

The red dashed line represents the predicted NDVI values for 2023, based on the KMeans clustering model, which was trained on the historical data from 2001-2022.

By comparing these two lines, we can visually assess the model's performance. The alignment of the predicted values with the actual ones, especially during peak vegetation months (e.g., May to July), indicates that the model performs reasonably well in forecasting NDVI trends.

4. Conclusion

This report analyzes cropland coverage and vegetation health across 23 counties from 2001 to 2023 using machine learning techniques and geospatial data. By implementing the KMeans clustering model, we successfully predicted crop phenology for 2023, comparing the predicted NDVI values to actual observations to assess model performance.

Our findings reveal significant seasonal variability in NDVI trends, capturing critical vegetation growth periods. The model's predictions align closely with actual data, though minor discrepancies suggest potential for improvement by incorporating additional environmental variables like precipitation and temperature.

This research provides valuable insights for agricultural stakeholders, enabling informed decisions regarding resource allocation and crop management.

Future work should focus on refining the model with alternative algorithms, higher temporal resolution data, and climatic factors to improve predictive accuracy and support sustainable agricultural practices. Overall, this study highlights the importance of data-driven approaches in addressing agricultural challenges and promoting environmental stewardship.

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

References

[1] Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2007)*, 1027-1035. Retrieved from https://arxiv.org/abs/2007.00134.

[2] USDA. (n.d.). VegScape: A National Landscape for Vegetation Analysis. *USDA Forest Service*. Retrieved from https://nassgeo.csiss.gmu.edu/VegScape/.

[3] USDA. (2021). CropScape: Cropland Data Layer. *National Agricultural Statistics Service (NASS)*. Retrieved from https://www.nass.usda.gov/Research_and_Science/Cropland/Release/index.php.

[4] Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment, 8*(2), 127-150. https://doi.org/10.1016/0034-4257(79)90013-0.