Our Changing Climate: A Geospatial Analysis of the Impact of Climate Change on Regional Cropland Health and Productivity

Authors:
- Sara Hunsberger (shunsbe2@jh.edu)
- Fernanda Montoya (fmontoy1@jh.edu)
- Meucci (Deerspring) Ilunga (milunga2@jh.edu)
- Meklit Yimenu (myimenu1@jh.edu)

Research Question and Importance

Key Research Question: How have variations in key climatic variables—such as temperature, precipitation, humidity, solar radiation, and cloud cover—impacted the health and productivity of croplands across different regions of the United States over time? And what potential does this sort of geospatial statistical modeling show for identifying regions with future risks for poor crop performance?

The objective of our project is to provide a basic analysis of how climate change has impacted regional cropland health and productivity in the United States over the last 25 years. This investigation is motivated by the significant role that agricultural output plays in food security and the broader economy. Understanding climate-driven changes in cropland productivity is essential for all community stakeholders, including policymakers and farmers, to make informed decisions and adapt effectively. To achieve these objectives, our project aims to integrate three key datasets: one for mapping US croplands, one serving as a proxy for vegetation health through metrics such as the Normalized Difference Vegetation Index (NDVI), and one for climate data, including temperature and precipitation.

Already Existing Work

There has been research done on the impact of climate change on vegetation in the United States. For instance: the article "Impacts of global change on peak vegetation growth and its timing in terrestrial ecosystems of the continental US" by Ying Liu et al. describes an analysis using similar factors that we will be looking at. These factors include how the Normalized Difference Vegetation Index (NDVI) changes with differences in temperature, precipitation and cloud cover. However, this research was done only on *all* vegetation in the United States, whereas our group is more interested in limitng our analsysi scope to the climate impact on croplands specifically. The Liu et al. study found that precipitation had the largest impact on NDVI, which tells us that geospatial precipitation data maybe be a prime data metric to look at for our analysis. (Liu et al. 2021)

In terms of looking at climate change and cropland specifically, the paper "Climate change and adaptation in agriculture: Evidence from US cropping patterns" by Xiaomeng Cui talks about climate change's effect on the types of crops being planted; for example in the US, soybean and corn production has increased, and this paper says that climate change has contributed to that increase. This means that in our analysis we can also possibly examine shifts in crops along with changes to their NDVI and other metrics of crop health. (Cui 2020)

Tentative Project Outline

Project Overview

This project aims to assess the impact of climate change on US cropland health and productivity by examining the relationship between key climate variables (temperature, precipitation, humidity, solar radiation, and cloud cover) and vegetation indices (NDVI, EVI) derived from satellite data. The project will utilize publicly available datasets like USDA NASS, MODIS, PRISM, and VIIRS. **Four potential statistical analyses are proposed, with an overall plan to implement at least two**; a brief justification is provided for each:

- **Feature Selection Analysis**: to identify the the potentially most influential climate variables on cropland health, helping to focus trend modeling efforts.
- **K-Means Clustering**: to potentially group regions with similar climate-crop health relationships, revealing spatial patterns and correlated regions of vulnerabilities. We think this would be useful because the climate datasets provide a rich set of variables, and a feature selection techniques might be able to effectively rank their importance in relation to NDVI/EVI, which are direct indicators of cropland health.
- **Multiple Linear Regression**: to model the relationship between climate variables and NDVI, allowing for future projections of cropland health under different climate scenarios. This is achievable because the data provides continuous time-based variables (climate parameters and NDVI) suitable for regression analysis, and the temporal component allows for trend analysis and future projections.
- **Decision Tree Analysis**: as an alternative to linear regression, this approach might be able to uncover more complex interactions between climate variables and their impact on cropland health, potentially providing more nuanced insights.

Two analyses of these four analyses should be reasonably achievable in four weeks. The data's publicly available and the project's focus remains on straightforward model implementation and interpretation and not complex development, which should make a subset of these proposed analyses workable.

Phase 1: Data Collection and Preparation

Objective: Acquire, clean, and prepare all relevant datasets to ensure a strong foundation for analysis.

Subgoals:

1. **Data Acquisition**:
    - Retrieve cropland data from USDA NASS Cropland Data Layers or Global Food Support Analysis Data.
    - Collect vegetation index data (NDVI, Enhanced Vegetation Index) from MODIS.
    - Gather climate data, including temperature, precipitation, humidity, solar radiation, and cloud cover from PRISM and VIIRS.
2. **Data Integration and Cleaning**:
    - Merge cropland maps with corresponding vegetation and climate data.

- Ensure temporal and spatial alignment across datasets.

3. **Data Validation**:
    - Check data quality (completeness, accuracy, consistency).
    - Implement basic statistical checks to identify outliers or inconsistencies.

Phase 2: Exploratory Data Analysis (EDA) and Initial Modeling

Objective: Trend identification to inform build initial modeling.

Subgoals:

1. **Exploratory Analysis**:
    - Perform EDA on climate and vegetation data to understand basic trends over time and across regions across various climatic variables.
    - Visualize relationships between key climate variables (e.g., temperature, precipitation) and vegetation indices.
2. **Correlation and Feature Selection**:
    - Identify significant climate variables influencing cropland health through correlation analysis.
    - Select features for predictive modeling based on their perceived importance.
    - **Possible Statistical Analyses: Feature Selection Analysis, K-Means Clustering**
3. **Basic Spatial and Temporal Analysis**:
    - Use simple mapping tools (e.g., QGIS or basic R/Python visualization packages) to create maps showing the average cropland health (e.g., NDVI) for major climate zones or states by year/decade. Highlight any clear visual patterns that suggest a link between climate variables and cropland health.
    - Create basic line plots or bar charts that show changes in average NDVI over the time period studied, segmented by region or major climate variable (e.g., annual temperature or precipitation changes). Look for straightforward trends such as an overall increase or decrease in productivity.

Phase 3: Basic Predictive Modeling and Final Analysis

Objective: Implement predictive models and interpret results.

Subgoals:

1. **Predictive Trend Analysis Using Linear Regression**:
   - Extend the temporal trend analysis by using a multiple linear regression model to project future cropland health based on key climate variables. Fit the model to the historical data to identify the relationship between variables such as temperature and precipitation with NDVI.
   - Use the fitted model to forecast NDVI values for the next 5-10 years, assuming current climate trends continue. Visualize these predictions in a line plot alongside historical data to show expected future changes in productivity.
   - Highlight regions where the model predicts significant positive or negative changes and discuss potential implications for agricultural planning and policy.
   - **Possible Statistical Analyses: Multilpe Linear Regression, Decision Tree Analysis**
   - Focus on fine-tuning basic parameters (e.g., regularization via penalization for linear models or max depth for decision trees).
2. **Brief Review of Advanced Models in Literature**:
   - Conduct a quick review of existing literature to identify promising models or methods used in similar climate-crop studies.
   - Summarize key findings and highlight any advanced approaches that could be worth considering for deeper analysis in future work, even if not implemented in this project.
3. **Result Interpretation and Visualization**:
   - Interpret model results to understand the general impact of key climate variables on cropland health. Identify which variables have the most significant influence and any notable regional or temporal patterns.
   - Create simple visualizations to present predictive model findings.
   - Prepare a concise one-page summary that discusses the interpretation of core results, emphasizing key findings and their potential implications.

Overall Project Deliverables:

1. **Predictive Trend Analysis**:
   - Linear regression model outputs projecting future NDVI trends.
   - Line plots showing historical and forecasted NDVI changes.
   - Brief notes on key regions with predicted significant changes and policy implications.
2. **Literature Review Summary**:
   - Short report on advanced models from literature worth future exploration.
3. **Result Interpretation and Visualization**:
   - Visuals (e.g., line graphs) summarizing key model findings.
   - One-page summary discussing main results, variable impacts, and policy recommendations.

Programming Paradigms

We will use the following program paradigms: Functional programming (R), Object oriented programming (Python for accessing data through API), explore parallel programming to deal with size of data set. We need to use python to retrieve the data because Earth Engine requires some use of python to get the data, and then we will do some data cleaning in r because it works very well with data wrangling. We will then explore parallel computing because some of our data is very large so it may be useful to take advantage of parallel computing.

Packages and Software

For the packages we can use the program R, use Python to get data from the earth engine, rgee (to work with API data through R). We then will use the following packages to process the crop land data QGIS/ARCGis and Geopandas.

Data Access and Collection

Data collection for this project will involve three primary data sources which we will then combine for our analysis. Regarding data access:

1. **Publicly Available:** All datasets are sourced from publicly accessible repositories (USDA NASS, MODIS, PRISM, VIIRS via Google Earth Engine).
2. **API Access:** Google Earth Engine provides APIs for programmatic data retrieval.

The three specific datasets we intend to use are explored in more detail here:

Crop land data: We will use one of the two following sources from Earth Engine to retrieve data on where cropland is in the United States. The first source is the USDA NASS Cropland Data Layers (source). The second source is global food support analysis data (source). We are not sure which data will be best to use when we actually pull the data so we will explore both. Both data sets are considered public domain and we are able to retrieve both data sets with API calls from Earth Engine.

Vegetation data: We will use MODIS data also retrieved with an API call from Earth Engine which has no restrictions on use (source). In this data we will use the Normalized Difference Vegetation Index, Enhanced Vegetation Index and Terra MODIS Vegetation Continuous Fields as metrics to understand how well the vegetation is doing in different areas.

Climate data: We will use the PRISM Daily Spatial Climate Dataset from Earth Engine which has no restrictions on use (source). This data includes data on temperature and precipitation for different areas in the US.

Data Product

Our final product will be a deployed Quarto website that walks through the research question and analysis, and summarizes the final project findings. We plan to include details on how the data was sourced and processed, along with visualizations and specific data analysis models used.

Timeline

The tentative project timeline is as follows:

- Week 1 (11/18): Meeting with Wenxuan to assess feasibility of project and make adjustments to project goals as needed. Data extraction and processing.
- Week 2 (11/25): Data cleaning and data analysis. Start of website deployment
- Week 3 (12/2): Continuation of data analysis. Implementation of data visualizations. Continuation of website development.
- Week 4 (12/9): Finalization of data analysis. Finalization and submission of project presentation. Group presentation to course. Changes to analysis following presentation feedback.
- Week 5 (12/16): Final project report finalization and submission. Group evaluation submission.

Task Split

We will split up the work in the following manner with everyone collaborating on problems that arise.

- Deerspring: Data extraction and processing, final presentation and report contributions, review of website and quality of analysis.
- Sara: Data analysis & visualization, final presentation and report contributions, review of website and quality of analysis.
- Fernanda: Data analysis & visualization, final presentation and report contributions, review of website and quality of analysis.
- Meklit: Website creation, final presentation and report contributions, review of website and quality of analysis.

Cui, Xiaomeng. 2020. "Climate Change and Adaptation in Agriculture: Evidence from US Cropping Patterns." *Journal of Environmental Economics and Management* 101: 102306. https://doi.org/https://doi.org/10.1016/j.jeem.2020.102306.

Liu, Ying, Chaoyang Wu, Xiaoyue Wang, Rachhpal S. Jassal, and Alemu Gonsamo. 2021. "Impacts of Global Change on Peak Vegetation Growth and Its Timing in Terrestrial Ecosystems of the Continental US." *Global and Planetary Change* 207: 103657. https://doi.org/https://doi.org/10.1016/j.gloplacha.2021.103657.