GEOG788P | Fall 2021

Kha-Ai Tran

Project Proposal

# Data Sources

## Sources:

* NOAA – Climate data (raster and table data)
* USDA – Agriculture data (vector and table data)
* Cropscape (George Mason University) – Agriculture data (raster data)
* Earthstat – Agriculture and vegetation data
* U.S. Forest Service – Invasive species data (vector data)
* U.S. Fish and Wildlife Service – boundaries data (vector data)
* USGS – Land Cover data (raster data)

## Data Types:

* Raster: TIFF
* Vector: Polygons, points
* Data tables: CSV

## Data Size:

Data size ranges from approximately 100KB to 2GB.

## Data cleaning and processing:

Datasets originate from varying sources with different standards and formatting. Thus, data will need to be cleaned to ensure proper geolocation. Data attribute types will also need to be set appropriately to ensure accurate analysis (e.g., that attribute types for quantitative analysis are continuous variables, and categorical variables are coded as such).

Additionally, because data varies in terms of areal unit or by different methods, table joins and spatial joins will most likely be required for analysis.

Performance constraints may also require the need to clip data to a smaller area of interest, which should be selected based on the area with the greatest available and consistent data. Projections may also need to be set and all data will be projected to the most appropriate projection for the area of interest.

## Limitations:

Depending on collection method, there may be regions in the total area of interest that lack data due to poor collection or missing data. In this case, a limitation may be that these regions are ignored, or the analysis will be focused only on areas with available and consistent data.

Time of collection is also a limiting factor. Collection time is not consistent across all datasets in relation to each other, and some output variables may not be relevant until a certain amount of time as lapsed. Taking this potential lag factor into consideration, another limitation is that the analysis may not accurately reflect the reality or the magnitude of causative relationships.

# Study Area

The study area for this project is the contiguous United States. Data identified covers the entire United States, with varying areal units (such as state-level and ecological provinces for vector data). Depending on availability and spatiotemporal consistency across datasets, a smaller study area may be selected.

Among the environmental science and climate community, it is known that climate change has and will have significant impact on agriculture. Temperature and precipitation especially can have significant impact on annual crop yield. In 2017, 40% of land in the U.S. were farms[[1]](#footnote-1). This is a significant proportion and understanding climate change impacts on agriculture is critical to a major sector of the U.S.

Additionally, many invasive species have also been known to expand their territory and footprint, as well as grow in volume with the increase in globalization as well as changing climate. Changing climates change the ecosystem, sometimes detrimental to native species by pushing them out of their natural habitats (whether through migration or disease and death) and creating habitats supportive of growing invasive species. Invasive species can destroy forests, damage crops, spread diseases (some of which can devastate croplands), and more[[2]](#footnote-2).

# Investigative Question/Objective

**How does the relationship between agricultural health/productivity and temperature change over time?**

**Primary Outcome:** Spatial analysis of agricultural productivity and vegetation/land cover across the United States with relation to changes in air or surface temperature over time. This may result in one or more maps illustrating changes over time and potentially predictive model of how agriculture may change in the future if an existing trend in temperature continues over the next several years. Analysis and interpretation will accompany any produced maps.

*Potential Alternative (or bonus):* How does the relationship between agricultural health (or productivity) and invasive species change over time? In relation to temperature or another climate change metric?

# Python Tools

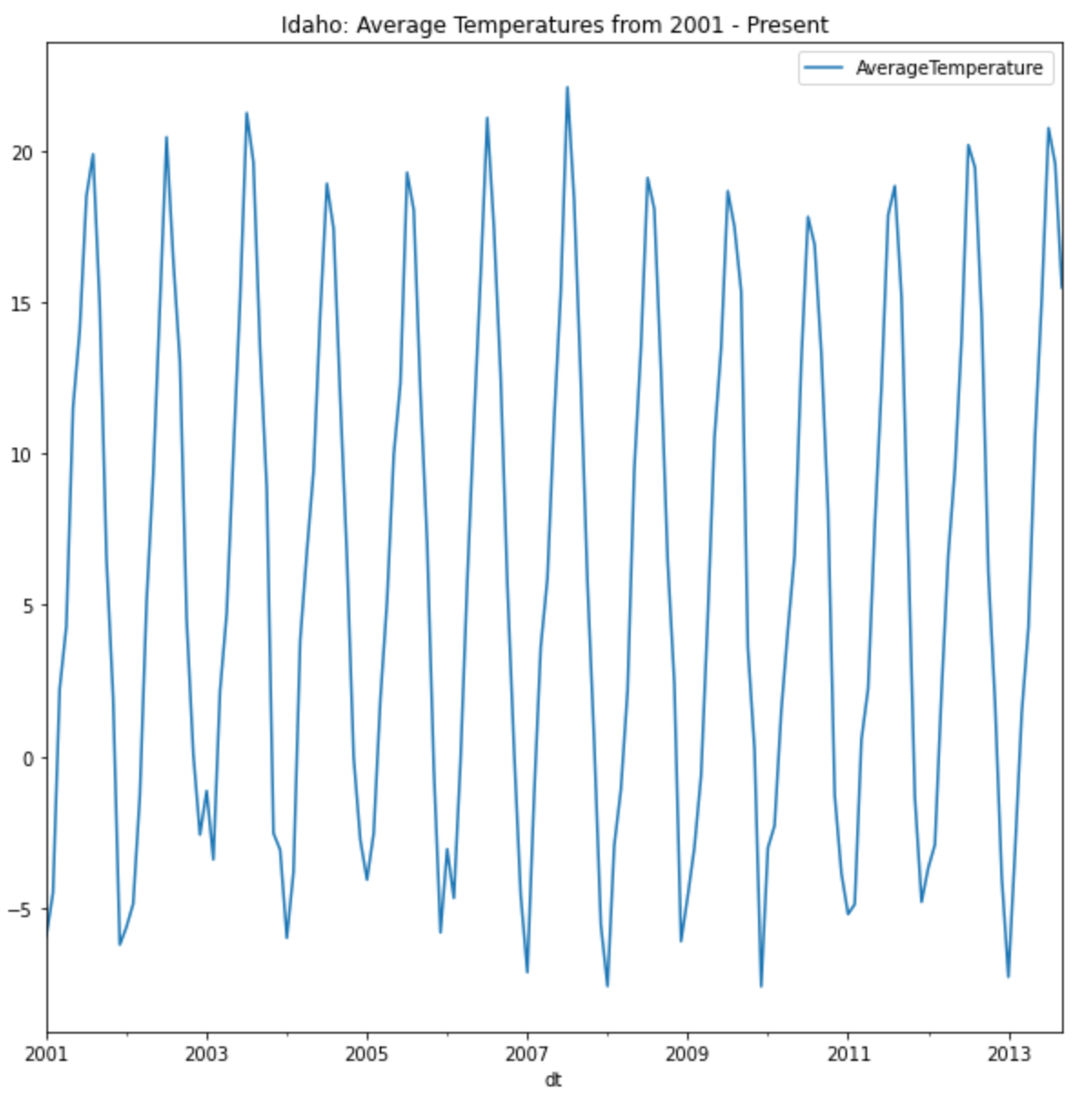
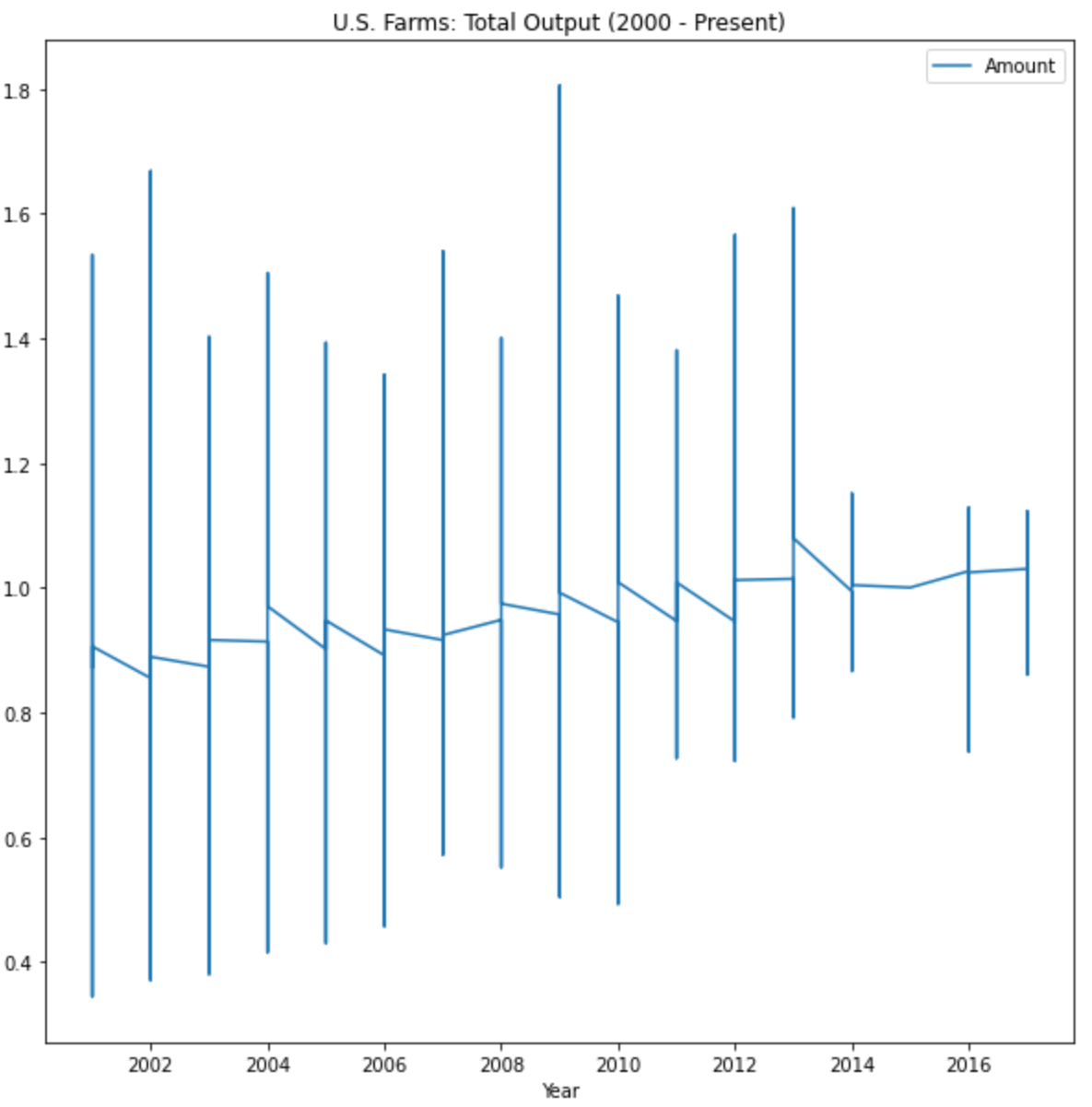
## Modules:

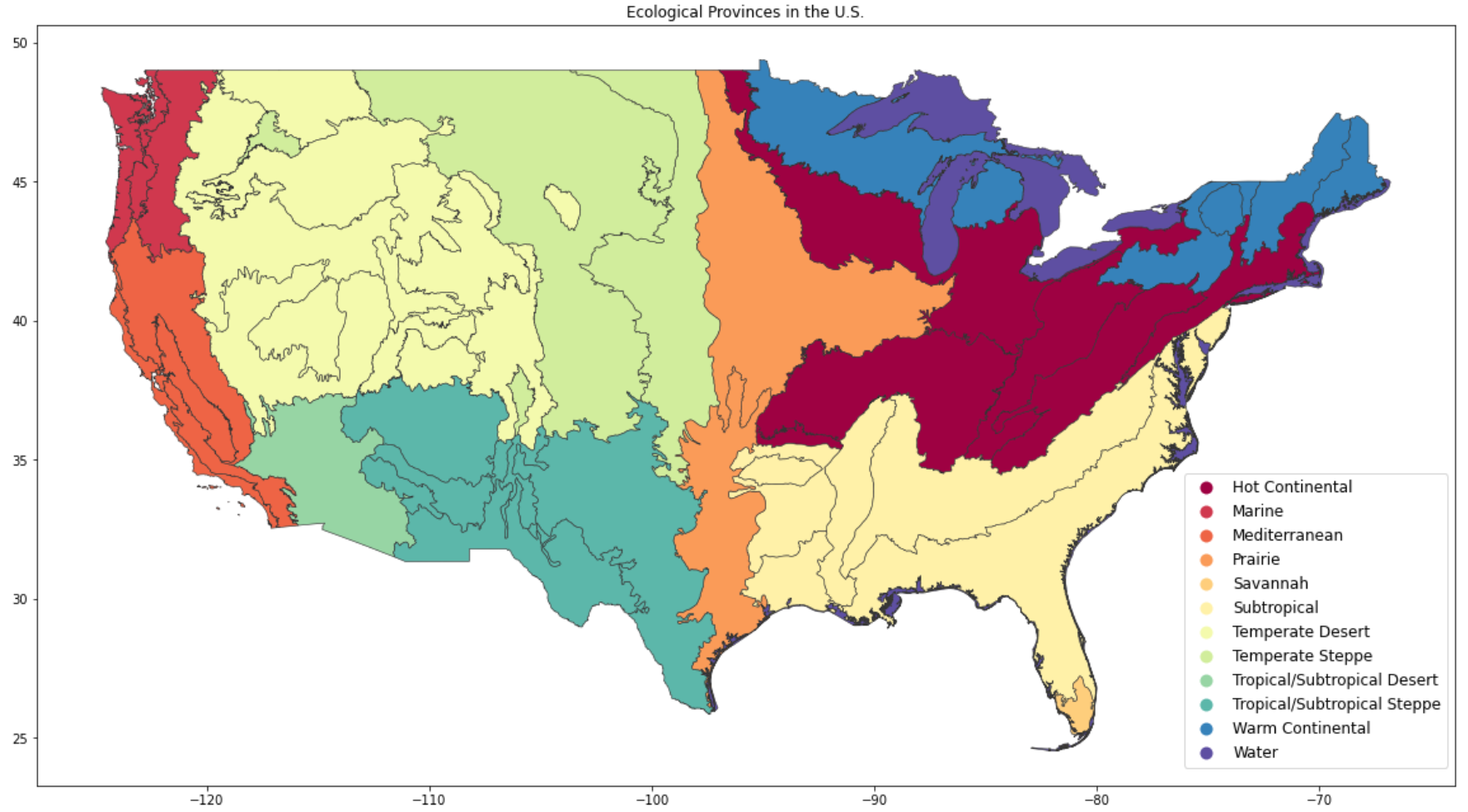
* Geopandas, Folium, PyProj, NumPy, Pandas, Matplotlib
* GeoRasters (for raster datasets)
* Potentially SciKit for classification and prediction models

# Challenges

Foreseen challenges include lack of data that overlaps temporally, or insufficient data to develop a robust model. For example, if invasive species data were to be included, this data is not as easily collected or as frequently collected as land cover, vegetation, and climate data. While the latter three datasets can also be collected in many known, multiple ways (such as via remote sensing, automated collection via physical stations, or manual collection), invasive species data tends to be collected manually, with much greater difficulty in automated collection. It is also much more difficult to collect accurate data or temporally relevant data because invasive species are usually identified after they have already impacted an ecosystem or community, or after they are gone (for example, rat droppings will indicate presence of the Norway rat, rather than people catching the rats themselves). Reconciliation between imagery and raster data with vector and table data may also be a challenge.

# Exploratory Data Analysis





1. https://www.nass.usda.gov/Publications/Highlights/2019/2017Census\_Farms\_Farmland.pdf [↑](#footnote-ref-1)
2. https://nifa.usda.gov/topic/invasive-pests-and-diseases [↑](#footnote-ref-2)