CS598 - Foundations of Data Curation Tan Dong Xiang

End-to-end data curation workflow Project Proposal dxt2@illinois.edu

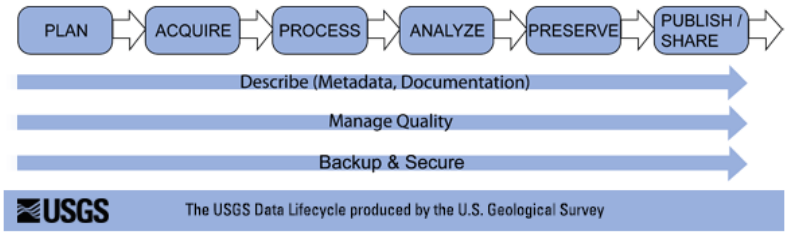
Effects of climate change on crop yield

Overview

Even though the threat of climate change looms over our heads, it is still difficult to understand its effects on the everyday lives of a typical household in the United States right now. I believe that finding a point of reference for an average person on the effects of climate change will greatly contribute to the education and message of the importance of slowing climate change. My hypothesis on an effect on the average person is: Rising temperatures and its effects on the weather has adverse effects on crop yields. In order to test this hypothesis, we will examine data of these two variables and test if they are correlated. For this exercise, we will focus on the temperature and crop yield over time of the state of Iowa, the state with the highest corn production in the country.

Plan

The data will be broken down into 6 main stages based on the USGS Science Data Lifecycle model. This section will describe what happens in each stage of the data lifecycle for this project.



Primary Model Elements

Plan: At this stage, a plan will be developed for the project. This includes the plan on how data will be acquired and managed throughout the lifecycle of the project and the expected outputs from the project.

Acquire: In this stage, data will be gathered through different data sources. The initial set of data sources for this project will be described in the “data sources” section of this proposal.

Process: This stage includes all activities that involve the de-identification, cleaning, transforming, and integrating of the data collected from the data sources to be analyzed.

Analyze: At this stage, the processed data will be analyzed to find patterns, and test the hypothesis proposed by the project.

Preserve: Actions and procedures to ensure the accessibility of the data and metadata used and produced by this project happens at this stage. This makes sure that provenance capture is implemented and documented.

Publish/share: This stage includes activities to make sure the work done in the project is understandable, reproducible, and reusable before it is submitted.

Cross-cutting activities

Describe: Create and maintain up-to-date documentation and metadata of data and processes throughout the project.

Manage Quality: Ensuring the quality of the data as well as the outputs of the project throughout the lifecycle of the project.

Backup & Secure: Manage risks of corruption and loss throughout the lifecycle of the project.

Data sources

This project will make use of data sources that data will be gathered from for the project.

NOAA Global Historical Climatology Network (GHCN)

This data source will provide historical data on changes in climate over time. Namely the temperature, precipitation, and other climate drivers affected by climate change. The data is available via NOAA's Climate Data Online API. A Python script can be written to scrap this data from the API and save it into a single CSV file so that it can be easily integrated with the other two sets of data.

However, if there are any limitations to the API or availability of historical data, alternate data sources can be used to supplement or replace missing climate data.

USDA NASS Quick Stats (Crop Data)

The data from this data source includes the yields and production for major crops on a county, state, and national level. This will be used to represent the agricultural production outcomes (crop yield) over time. The data is available on USDA's quick stats API, but it would require a signup to get an API key. Alternatively, data can be fetched from the quick stats portal but this only supports fetching 50,000 rows at a time or a bulk download option.

Integrating the data sources

In order to test the hypothesis, the two data sources have to be merged into a single dataset. Since we are interested in changes in the climate (e.g. temperature) and crop yield over time, the integration process must make use of the same time ranges (e.g. data in months/years) as well as the type of data (e.g. time-series vs cross-sectional data). This ensures that the data will be able to understand and draw conclusions from the data.

Summary

Data source	Data type	Coverage	Link
NOAA Global Historical Climatology Network (GHCN)	Daily and monthly climate data (temperature, precipitation, etc.)	Historical to present	https://www.ncei.noaa.gov/pub/data/ghcn/daily/ghcnd_all.tar.gz
USDA NASS Quick Stats (Crop Data)	Crop yields, acreage, production statistics	1909 to present	https://www.nass.usda.gov/datasets

Team

I will be working alone on this project.

Timeline

This section will describe the timeline of the project, the dates will be anchored around the three deadlines (proposal, progress report, and final submission). Since I will be working on the project alone, all tasks will be done by me.

Task	Due date	Description
Proposal submission	9/15	Submit this document
Project planning (Plan)	9/20	Create an outline of the project. This includes: - Sampling data to identify data cleaning requirements & pivot if project is not feasible - Outline the project, including documenting steps in each of the data lifecycle - Decide on the medium of presentation of the project. (e.g. Jupyter notebook, webpage) - Decide on the final outputs from data analysis to find relationships supporting hypotheses. (e.g. graph showing matching trends for the data over X years)
Data acquisition (Aquire)	10/20	Download or write scripts to scrape/download data from the data sources. Data will be stored as CSV files until the next step for easy transformation.

Progress report	10/27	Status updates on the progress of the project, if there are significant changes to the project (e.g. change in data source or goal) they must be communicated to the professors before this.
Data cleaning & transformation (Process)	11/14	Data is cleaned and transformed for integration. One transformation is that data should be based on the same timeframe (e.g. each row should represent a month)
Data integration & graph plotting (Analyze)	11/30	Merge the datasets into a single dataset, plot graphs and analyze data draw conclusions on the hypothesis
Data storage and automation (Preserve & publish/share)	12/3	Along with the initial set of data, scripts used in each step of the project should be kept and integrated into a single platform (e.g. Jupyter notebook) and uploaded onto an accessible location like GitHub
Final Submission	12/10	Project report describing the entire project & where to access supporting artifacts

Constraints

The main constraints in the project is that there might be some rate limits to accessing the data which may slow the progress of the project.

Gaps

There are no gaps at this moment.