Irrigation Data Visualization and Analysis Tool

Kirin Mackey

November 22, 2024

1 Project Description, Background, and Motivation

Irrigation is an essential device in the U.S. agricultural industry, as it improves efficiency of farms and offers a way for states that have a dry and arid climate to sustain themselves. According to the 2017 Census of Agriculture, 58 million acres of cropland were irrigated, and farms using irrigation supported 54 percent of the total crop sales in the U.S. (Hrozencik, 2023). Irrigation has many components, such as energy, facilities and equipment, labor, practices, water, pumps, and wells. All of these components have further details, which can reveal the negative effects of irrigation. For instance, acres irrigated with groundwater in Oklahoma could indicate the decline of the Ogallala Aquifer, which accounts for one fourth of the water supply used for agricultural production in the U.S. (Hanrahan, 2024) and supplies more than 20 billion dollars worth of food (Little, 2009). On the other hand, these details can reveal sustainable practices and indicate future irrigation use, such as the amount of land irrigated with recycled water or the amount of land equipped for irrigation in a certain state. Learning about these details and finding statistics about them is an arduous task that involves searching through multiple research reports, especially if the user wants to look at a particular state or year.

To learn about these revealing details in an effective manner, this project will make a tool in which a user can either get visualizations or statistics about them using data at the state level from the United States Department of Agriculture (USDA). A user can specify an individual or multiple states, the specific data they want to visualize or analyze, and what years the data to be visualized or analyzed reflects through the use of dropdown lists and buttons. If they want a visualization, they will be given the option to choose a line graph or bar chart, each of which will affect the amount and which type of items the user can visualize. The user will also be prompted to choose which type of statistic they want displayed if applicable, such as minimum, maximum, average, and sum. The user will also have the option to save the visualization the tool creates as a png file. The tool can be used to aid in creating research reports, serve as reference material for government officials, and be used in classroom settings for students studying environmental science and its associated public policies and economics.

2 Data Description

The data comes from the United States Department of Agriculture and its Quick Stats Agricultural Database tool found at https://quickstats.nass.usda.gov/ . The data was found by filtering on "Group" as "IRRIGATION" and "Geographic Level" as "STATE." It was then saved as a csv file using the tool's "Spreadsheet" option and placed into a GitHub repository that can be found using https://github.com/krmackey/irrigation. Using Microsoft Excel, the first 5 out of 77,496 rows of data look like:

Program	Year	Period	Week Ending	Geo Level	State	State ANSI	Ag District	Ag District Code	County	County ANSI	Zip Code	Region	watershed_cod	Watershed	Commodity	Data Item	Domain	Domain Category	Value	CV (%)
CENSUS	2023	YEAR		STATE	ALABAMA	1							0		ENERGY	ENERGY, IRRIC	TOTAL	NOT SPECIFIED	6,343,000	17.8
CENSUS	2023	YEAR		STATE	ALABAMA	1							0		ENERGY	ENERGY, IRRIC	TOTAL	NOT SPECIFIED	2,220	18.1
CENSUS	2023	YEAR		STATE	ALABAMA	1							0		ENERGY	ENERGY, IRRIC	EXPENSE	EXPENSE: (1,000 TO 1,999	179	42.5
CENSUS	2023	YEAR		STATE	ALABAMA	1							0		ENERGY	ENERGY, IRRIC	EXPENSE	EXPENSE: (10,000 TO 19,9	115	54.6
CENSUS	2023	YEAR		STATE	ALABAMA	1							0		ENERGY	ENERGY, IRRIC	EXPENSE	EXPENSE: (2,000 TO 4,999	136	28.7

The columns and initial data types are: Program (string), Year (integer), Period (string), Week Ending (string), Geo Level (string), State (string), State ANSI (integer), Ag District (float), Ag District Code (float), County (float), County ANSI (float), Zip Code (float), Region (float), watershed_code (integer), Watershed (float), Commodity (string), Data Item (string), Domain (string), Domain Category (string), Value (string), and CV (%) (string).

Program describes whether the data was collected through the census or a survey, and Year details which year the data was collected in (2007 and 2012-2024). There will generally be more data available corresponding to 2013, 2018, and 2023 because the census is done every 5 years. Period describes whether the data relates to a year or a specific week, and Week Ending describes the end date of a week stated in the Period column. Geo Level describes the geographic level of the data, State gives the state name, and State ANSI gives the state's unique geographic code. Ag District and Ag District Code give the name and unique geographic code of an agricultural district if the geographic level is listed as agricultural district or county. County gives the name of a particular county and County ANSI gives the county's unique geographic code if county is listed as the geographic level. Similarly, Zip Code gives the zip code if the geographic level is listed as zip code. Watershed and watershed_code give the name and corresponding geographic code of a watershed if watershed is noted as the geographic level. The Commodity column describes what general component in irrigation (energy, facilities and equipment, labor, practices, water, pumps, and wells) each row in the data table relates to. Data Item gives a description of what the value in the Value column represents and its units. More details about some of the values in this column can be found at https://tinyurl.com/5v78uapr. Domain gives a general description of the item in Domain Category, such as "TOTAL," "WATER SOURCE," or "EXPENSE." Domain Category gives information about the entries in the Data Item column, such as "EXPENSE: (1,000 TO 4,999 $)." The Value column holds numeric values representing all the information stored in the other columns. CV(%) provides the coefficient of variation, which gives information on measurement errors, and is displayed as a percent.

To display the connections between the State, Year, Data Item, Domain, Domain Category, and Value columns, here is a representation of a row in the data, where the left column denotes the column names and the right denotes the corresponding values:

```
State                                                        DELAWARE
Year                                                             2023
Data Item          ENERGY, IRRIGATION, ON FARM PUMPING, GASOLINE ...
Domain                                                    WATER SOURCE
Domain Category                                WATER SOURCE: (SURFACE)
Value                                                              80
```

The full entry in the Data Item column is “ENERGY, IRRIGATION, ON FARM PUMPING, GASOLINE & ETHANOL BLENDS, IN THE OPEN - ACRES IRRIGATED.” The resulting interpretation of this row is that in 2023, Delaware had 80 acres in the open that were irrigated with surface water, which was pumped on farms using gasoline and ethanol blends.

In the project, this data is preprocessed when building a relational database. In the preprocessing, the Program, Week Ending, Geo Level, Ag District, Ag District Code, County, County ANSI, Zip Code, watershed_code, and Watershed columns are all dropped because they either have unnecessary information or hold no data since the USDA’s Quick Stats Agricultural Database was queried for data only at the state level. Additionally, the CV(%) column is dropped because the project is concerned with the raw values in the Value column rather than error estimations. The data table is also filtered to only display results where the time period is based on a year (70,612 rows). The Value column is further filtered on, as some entries have “(D)” (3,882 rows) to denote a state withheld information or “(Z)” (4 rows) to indicate values less than half the rounding unit specified by the USDA. All rows with either of these values are deleted. All entries in each column are also casted as string objects with the exception of the Value column, which has its values casted as floats.

The Domain and Data Item columns were further preprocessed to remove redundant information that is present in other columns. It should be noted that the entries in the Domain Category column are all listed “NOT SPECIFIED” when the entry in the Domain column is “TOTAL.” The final tool will address this by not allowing users to see “NOT SPECIFIED” when filtering for data and rather offer other data items using the same units if the user wants to compare across data items rather than state. The new table holds 66,726 records, where the first 5 rows look like the following after being loaded into a Pandas DataFrame:

	Year	State	State ANSI	Commodity	Data Item	Domain	Domain Category	Value
0	2023	ALABAMA	1	ENERGY	EXPENSE, MEASURED IN $	TOTAL	NOT SPECIFIED	6343000.0
1	2023	ALABAMA	1	ENERGY	NUMBER OF PUMPS	TOTAL	NOT SPECIFIED	2220.0
2	2023	ALABAMA	1	ENERGY	OPERATIONS WITH EXPENSE	EXPENSE	1,000 TO 1,999 $	179.0
3	2023	ALABAMA	1	ENERGY	OPERATIONS WITH EXPENSE	EXPENSE	10,000 TO 19,999 $	115.0
4	2023	ALABAMA	1	ENERGY	OPERATIONS WITH EXPENSE	EXPENSE	2,000 TO 4,999 $	136.0

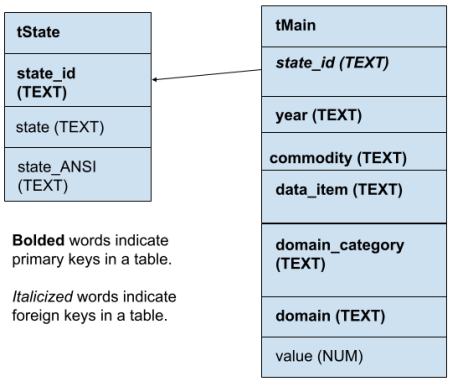
When setting up a relational table for state information (State, State ANSI), the state abbreviation was also added as a column to aid in visualizing the data and called "state_id." This was done by using data provided by CDC, provided by this link: https://www.cdc.gov/wcms/4.0/cdc-wp/data-presentation/data/data-map-state-abbreviations.csv In this dataset, any items that were not U.S. states were deleted in Microsoft Excel. After being loaded into a relational table called tState, the first 5 rows of the state data (both from the CDC and USDA) are:

	state_id	state	state_ANSI
0	AL	ALABAMA	1
1	AK	ALASKA	2
2	AZ	ARIZONA	4
3	AR	ARKANSAS	5
4	CA	CALIFORNIA	6

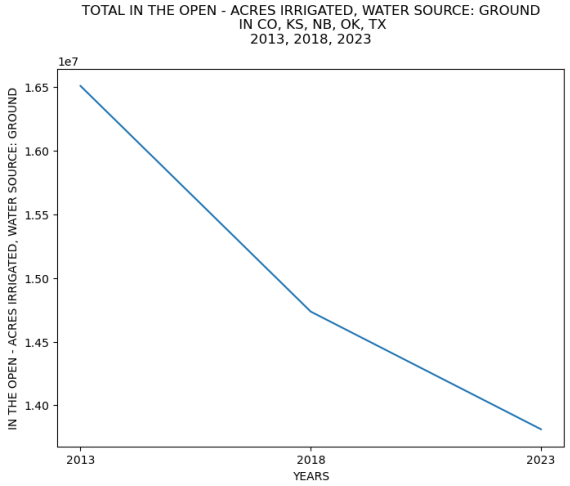
All the other columns from the pandas DataFrame after preprocessing, as well as the state abbreviation column, are then loaded in a relational table called tMain, where the first five rows are:

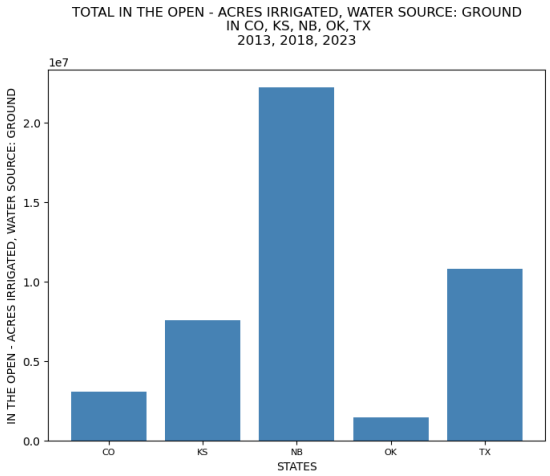
	state_id	year	commodity	data_item	domain	domain_category	value
0	AL	2023	ENERGY	EXPENSE, MEASURED IN $	TOTAL	NOT SPECIFIED	6343000.0
1	AL	2023	ENERGY	NUMBER OF PUMPS	TOTAL	NOT SPECIFIED	2220.0
2	AL	2023	ENERGY	OPERATIONS WITH EXPENSE	EXPENSE	1,000 TO 1,999 $	179.0
3	AL	2023	ENERGY	OPERATIONS WITH EXPENSE	EXPENSE	10,000 TO 19,999 $	115.0
4	AL	2023	ENERGY	OPERATIONS WITH EXPENSE	EXPENSE	2,000 TO 4,999 $	136.0

The database can be further represented by the following ERD:



Using the discussion from the introduction of this project and the described data, the following are preliminary graphs detailing the use of the groundwater in Colorado, Kansas, Nebraska, Oklahoma, and Texas, all of which use the Ogallala Aquifer:





They overall indicate that Nebraska has, over 15 years, irrigated more than 2 million acres of land with groundwater even though it is smaller than Texas, who irrigated slightly more than 1 million acres of land. Nonetheless, the number of acres irrigated potentially using water from the aquifer has slightly decreased over time, which may indicate the supply itself is dwindling or more conservation efforts have been applied. The final tool will have the visualizations interactive, where once hovering a user can specifically look at the values associated with the y axis for each state.

3 Progress and Next Steps

Currently, there are functions that create a relational database, clean the data accordingly, and insert the cleaned data into relational tables. There is also a series of functions that, provided a dictionary, query the database and give results the user can choose from to further specify the data they want to visualize and analyze. This is done for the user to specify commodity, domain, data item, domain category, and year or years. In the final tool, these functions will run in that order, and the option of specifying them only appears if the user made a selection for the previous item. These functions will be executed in the background, with items added to the dictionary when the user either clicks a button, checks a box in a checklist, or makes a selection in a dropdown list. Each of these functions querying the database return a list, which will be used in the formulation of buttons or a dropdown list for the next item to be specified in the final tool. State in the final tool will be the first specification the user makes, and does not affect querying until retrieving the available years to visualize and analyze.

In the case of a user choosing “TOTAL” as the domain, and the user moves on to specifying domain category, they will not be prompted with “NOT SPECIFIED.” The user will

rather, if they want to compare across data items rather than states or years, be returned other data items using the same units and that also have “TOTAL” listed as their domain. However, if they wish to compare across states or years using only 1 data item, they will not get returned these options of additional data items. It should be noted that the function retrieving years looks at years common to all the states listed in the dictionary. There is also a function that takes in an operation (max, min, average, or sum which are selected by the user), and formulates a query that correctly groups the data depending on the amount of items stored with each key in the final dictionary. It also takes into account whether the user specified a line graph for their desired visualization. After this is executed, the query held as a string is passed into a different function that queries the database and returns the desired values to be analyzed and visualized.

The next steps are to start writing generalized functions that visualize the results from a user’s final (meaning all required inputs are filled out) database query either in a line graph or bar chart. Further preprocessing of the items stored in the data item column may be necessary, as some such as “SOURCE = OFF FARM, SUPPLIER = NON-FEDERAL, IN THE OPEN” may be too long for 1 line in a tick label. A get_statistics function will also be made so that the user can look at the statistics corresponding to their chosen inputs in a text format rather than in a visualization. Additionally, there will be a save_results function that will be called if a user clicks a button, and will save either the text or visualization results to the user’s computer. These functions, along with the ones described above, will be called in an interface coded with Dash. The visualizations will then be created using plotly, as it is complementary to Dash.

References

Hanrahan, R. (2024, January 31). *Ogallala Aquifer depletion threatening rural communities & Ag*. Farm Policy News. https://farmpolicynews.illinois.edu/2024/01/ogallala-aquifer-depletion-threatening-rural-communities-ag/

Hrozencik, A. R. (2023, September 8). *Irrigation & water use*. USDA ERS - Irrigation & Water Use. https://www.ers.usda.gov/topics/farm-practices-management/irrigation-water-use/

Little, J. B. (2009, March 1). *The Ogallala Aquifer: Saving a vital U.S. water source*. Scientific American. https://www.scientificamerican.com/article/the-ogallala-aquifer/