Public Water Systems Data Integration Project

Spring 2019

Avena Cheng, Terry Kim, Yifan Ning, Emily Zhang

GitHub: https://github.com/richardnnn/PWS_Water_Integration

Abstract

Accurate information on the water supply is crucial to the development of communities, given that clean drinking water is essential to ensure public health, and a multitude of industries need access to water for various purposes. However, the California State Water Resources Control Board lacks the means to quickly and accurately determine the water supply of different counties across the state. The sources of water data was very unstructured and in some cases not well collected. Even so, we wished to identify water sources for a specific location, so we developed a tool to integrate data to allow users to look up water supply sources for Public Water Systems. The various datasets were cleaned, merged, and used to create a visualization. The final result was a map of California that shows the locations of wells, information of water systems for counties, and the distances between locations on the visualizations.

I. Introduction

Every year, more than a million Californians are exposed to unsafe drinking water a some point during the year. Some communities have been exposed to unsafe water for more than a decade. In the San Joaquin Valley alone, there were nearly 100,000 residents living without access to clean drinking water. A 2017 drinking water compliance report by the State Water

Resources Control Board shows that an estimated 592,000 Californians lived in a public water district that received a water quality violation in 2017.[1] Droughts, floods, aging infrastructure, and other human and natural causes can disrupt the water supply and limit—or eliminate—access to safe drinking water for days, months, or even years. Therefore, we are focusing on water supply information to provide data on the water supply sources of various counties in an attempt to mitigate this issue.

In this project, we worked in conjunction with the West Big Data Innovation Hub, the Berkeley Division of Data Sciences, and the California State Water Resources Control Board to create a visualization to display the water system information after clicking on a certain location. We utilized the given datasets and Jupyter notebooks, pandas, ipyleaflet, ipywidgets, requests, and matplotlib as our principle methods. Our visualization is especially powerful given the context of the current unsafe drinking water crisis, since being able to see the sources and destinations of contaminated water is the first step to mitigating the issue. Being able to pinpoint specific counties with unsafe drinking water also gives authorities critical information on which locations to prioritize.

II. Methods

II-1. Data Acquisition

At the start of this project, the team was provided with 2 main datasets: Main Dataset[2] and Stress Test Dataset[3]. The Stress Test Dataset contains various information including water demand, supply, and supplier, but since it only contains about 400 rows, the team did not utilize this dataset. The Public Potable Water Systems dataset, referred to as the Main Dataset from this point, contains information about different water systems and their unique water system number. This dataset was finalized during June of 2018 and published online. In the middle of the semester, the team was given the EAR Reports[4] to supplement any lacking information. However, EAR Reports were not used as well since these reports did not contain information necessary to create the map. The team instead used these reports to create case studies and find interesting points in the data. The team also found the Drinking Water Watch Dataset[5] online from the California Open Data Portal. This dataset, finalized by April of 2018, contains water systems with their respective wells and locations (longitude and latitudes). The created visualizatiobn mainly uses data from merging the Main Dataset with the Drinking Water Watch Dataset.

II-2. Data Cleaning

The prime merged table is named as "c". Some latitude and longitude values are either missing or have the value 'zero'. Such rows will give no useful information to us as we are trying to render the location, so we simply removed these rows. This filters out 3387 of the original 23277 rows. The table during this step is named "removed_null_na" However, some well names contain words like 'ABANDONED' or 'DISCARDED'. Therefore, we created both tables "c" and "removed_null_na" before and after filtering out the abandoned wells and did statistical analysis on the amount of wells abandoned. Numerically 3293 out of the 19890 remaining wells are abandoned or destroyed. Our final map did not include the abandoned or destroyed wells. The table after this cleaning process is named as "primary_merged".

II-3. Tools

The main tools utilized were Jupyter notebooks, pandas, ipyleaflet, ipywidgets, requests, and matplotlib.

Since the platform is limited to Jupyter notebooks, there aren't too many choices for interactive visualization. We considered Folium, but it does not interact well with taking user inputs and dynamically updating states. Ipyleaflet also tends to integrate with other ipywidgets much more seamlessly. Request is used for sending queries to an online site that converts the latitude, longitude tuple into corresponding county and state. Other tools are mainly used for data cleaning and explorative analysis. Here is the list of package versions:

pandas: 0.24.2

Ipyleaflet: 0.10.1

Ipywidgets: 7.4.2

requests: 2.21.0

Matplotlib: 2.1.2

II-4. Main Handle Logic

The main logic for the function "`handle_click(**kwargs):` " handling user clicks is described as follows:

1. Obtain the latitude and longitude point of the click through the argument of the function.
2. Render the marker for the user's click location.
3. For the location (lat, long) the user clicks, determine the corresponding state and county by sending a request query to a parsing website API.
4. Stop here if the result is not in California.

5. Filter out the rows of information by the county we just determined, convert the rows to markers, and put them into a marker cluster.

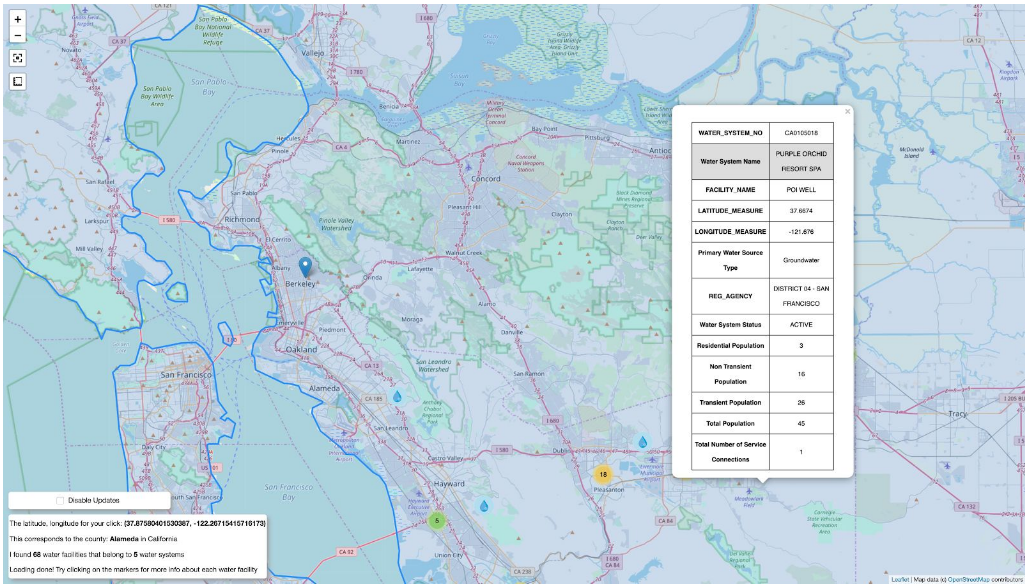
II-5. Extra Notes

1. Currently, it takes about 10+ seconds to render all the well locations if the found results are more than 800 rows. We tried to use multithreading or multiprocessing to speed up the rendering, but it seems in vain as the rendering process of leaflet.js does not interact with this well. We achieved the most speed through better organization of information retrieval from panda tables.
2. It seems ‘ipywidgets’ does not offer the “widgetControl” module in python3. Try both the python2 and python3 kernels on Jupyter.

III. Discussion

The purpose of this project is to be able to quickly visualize and analyze water systems in our state. One region we looked at was Central Valley (Fresno County) because it had one of the highest number of wells in California. An example of the information gained from clicking on the visualization is displayed in Figure 1. It displays information on the exact coordinates, whether or not it is active, number of connections, and the residential population it serves. The map can give us an idea of how far apart these wells are, which can be useful for understanding the density of water sources in an area. The distance will also be helpful in understanding how far water must travel to reach a neighborhood (assuming that we gather data on the locations of such neighborhoods).

Figure 1. Water supply information after clicking on a marker.



WATER_SYSTEM_NO	CA0105018
Water System Name	PURPLE ORCHID RESORT SPA
FACILITY_NAME	POI WELL
LATITUDE_MEASURE	37.6674
LONGITUDE_MEASURE	-121.676
Primary Water Source Type	Groundwater
REG_AGENCY	DISTRICT 04 - SAN FRANCISCO
Water System Status	ACTIVE
Residential Population	3
Non Transient Population	16
Transient Population	26
Total Population	45
Total Number of Service Connections	1

□ Disable Updates

The latitude, longitude for your click: **(37.87580401530387, -122.26715415716173)**

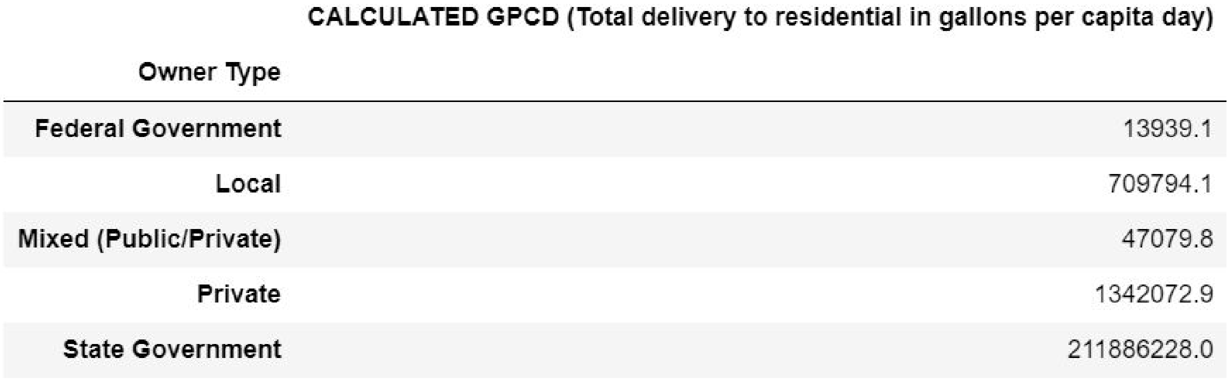
This corresponds to the county: **Alameda** in California

I found **68** water facilities that belong to **5** water systems

Loading done! Try clicking on the markers for more info about each water facility

We studied three locations (Central Valley, Bay Area, and Los Angeles) as case studies to provide an example for how to use the data and tools we integrated and created. For Central Valley, we focused on Fresno since it has the highest number of wells. We noticed that a lot of Fresno County's wells were abandoned or inactive which, after further research, makes sense as we found that they get their water from an underground aquifer that requires multiple wells to be drilled [6]. We also found that most of them were privately or locally owned (94%), as seen in Figure 2, which came as a surprise to us, since we assumed that water would be more publicly owned. However, we found that most of the delivered water from wells (99%) comes from state government.

Figure 2. Owner Type of Water in Fresno.

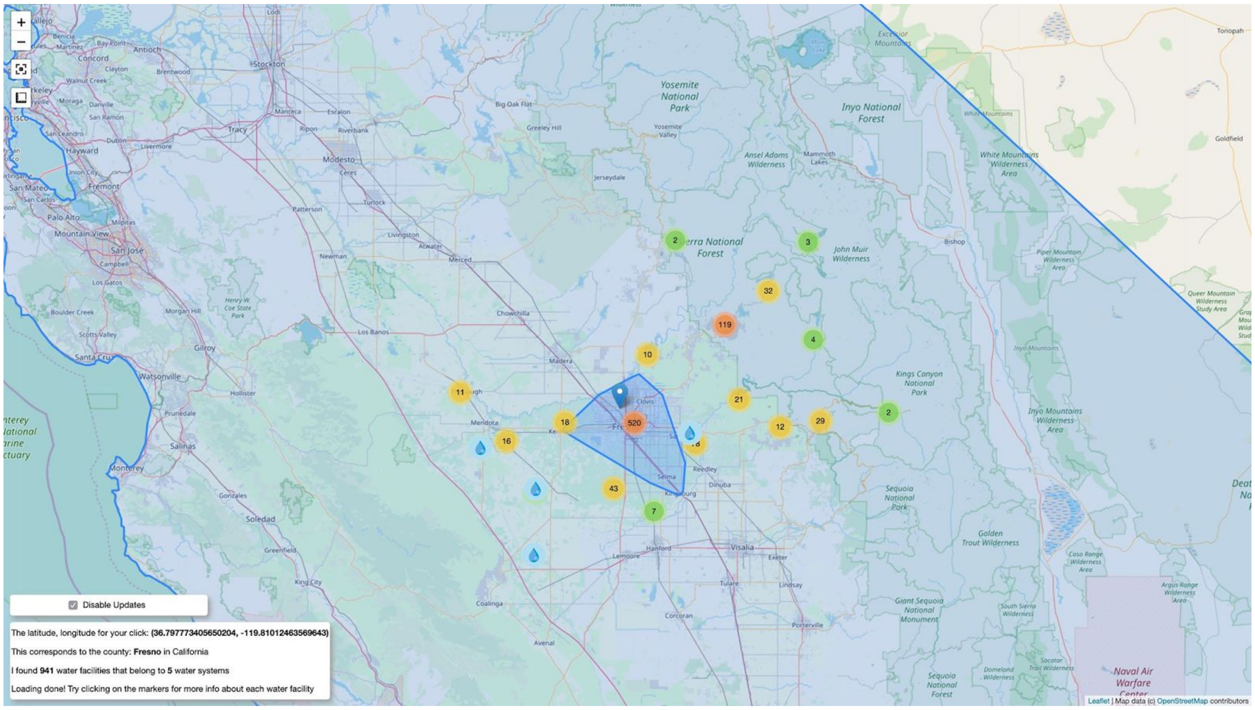


	CALCULATED GPCD (Total delivery to residential in gallons per capita day)
Owner Type	
Federal Government	13939.1
Local	709794.1
Mixed (Public/Private)	47079.8
Private	1342072.9
State Government	211886228.0

Next we looked at the Bay Area and found that San Francisco County only had two water systems (CA3810001 and CA381070), whereas the other eight counties (that consist of the Bay Area) have almost 2000 water systems supporting it. After further investigation, we found that San Francisco's water comes Hetch-Hetchy, a reservoir in Yosemite, that transports 85% of the water to the San Francisco Bay Area, whereas the other 15% come from the reservoirs in San Mateo and Alameda. We used a similar analysis from Central Valley, and found that most of the delivered water from wells (99%) comes from private or local organizations. Finally, we looked at Los Angeles (and adjacent areas), but did not find anything that encouraged further investigation; the numbers seemed standard. We also looked at number of service connections for all three areas and found that in each region, most connections come from local facilities.

We've made excellent progress in mapping out the specific locations for each water system, but there are a few limitations which, given more data, could greatly improve the quality of our map. For example, understanding which neighborhoods have access to what type of water could help us understand how the water is distributed (is there bias or correlation with cost?), but we only have data at the county level. Furthermore, the Drinking Water Set only contains coordinates for wells (reservoirs, dams, and other sources are not available), and the EAR dataset only contains information regarding residential locations, so water usage for commercial, agricultural, and any other usages are currently unavailable.

Figure 3. Fresno water supply information on map



☑ Disable Updates

The latitude, longitude for your click: **(36.797773405650204, -119.81012463569643)**

This corresponds to the county: **Fresno** in California

I found **941** water facilities that belong to **5** water systems

Loading done! Try clicking on the markers for more info about each water facility

IV. Conclusion

In the future, there is opportunity to create a more extensive and richer visualization that addresses the issues above and adds more functionality for further analysis. To improve our visualization, the next step is to add a search engine for the data. For example, it would be good to gather data on reservoirs and other water sources and have more specific information on where the water travels and its volume. One idea is to find pipeline data and show which ones have a higher volume through line thickness. In addition to this, we would collect information on water usage and distribution for commercial purposes as well. Another idea is to combine this research with another ongoing project in order to see how the cost of water varies among location. Ideally, we would want this project to eventually become an easy tool for users to quickly analyze the public water systems and find ways to make appropriate changes or develop meaningful insight on this essential resource.

V. References

[1] https://www.politifact.com/california/statements/2019/feb/14/gavin-newsom/true-more-million-californians-dont-have-clean-dri/

[2] https://data.ca.gov/dataset/drinking-water-public-water-system-information

3https://www.waterboards.ca.gov/water_issues/programs/conservation_portal/docs/emergency_reg/uw_self-cert_submittals.xlsx

[4] https://data.ca.gov/dataset/drinking-water-public-water-system-annually-reported-water-production-and-delivery

[5] https://data.ca.gov/dataset/drinking-water-public-water-system-information

[6] https://www.fresno.gov/publicutilities/water-quality-delivery-testing/water-source-distribution/