Predicting Agricultural Water Usage in California*

An Examination of the Use of Urban Water usage to predict Agricultural Water Usage

Catharina Castillo

October 29, 2025

Agricultural water use is harder to measure than urban water use due to reduced accuracy associated with the remote sensing techniques used. This paper seeks to explore if measurements of urban water use can be used to predict measurements of agricultural water use with simple linear regression to see if recorded urban water use can be used to assess confidence in future agricultural water use measurements. Unfortunately, it was discovered that measured urban water use cannot accurately predict measured agricultural water use through a simple linear regression, though we still found evidence of a relation between urban water use and agricultural water use. We hope future research may be able to add more variables to create a model that can be used to assess measured agricultural water use.

1 Introduction

Historically, it has been harder to measure agricultural water usage than urban water usage. The infrastructure used in urban settings has allowed water to be measured directly using metering. While many farms do use sensors for direct measurements, the cost and time needed to deploy them across all agricultural fields for direct measurement is too great. Instead, land use surveys are used to calculate water use based on factors such as acreage of active fields and crops grown (Cooley 2020). Irrigation alone has been noted as one of the more unreliable measurements ("California Water Use" 2018).

With recent technological advancements, such as micro sprinklers, more accurate measurements have been able to be taken. These have all arisen as a result of pushes for efficiency caused by drought concerns in California (PBSNewsHour 2025). The use of satellite and aerial

*Project repository available at: https://github.com/NettleHook/Groundwater_Project.git.

imagery has also added another method of tracking water use across all sectors (ABC10 and Knight 2023).

It is important to have accurate measurements of water usage across sectors, as this information helps to predict future water usage and allows us to track progress towards sustainability. ("Groundwater Sustainability Plans" 2018)

This paper seeks to explore if we can use the measured urban water usage to estimate the agricultural water usage using linear regression. While this can't replace current methods of measuring agricultural water usage, an estimate can help assess the confidence in the measurements taken.

In Section 2, we will explain the data used. In Section 3, we will introduce the linear regression model. Finally, in Section 4 we will lay out our results and in Section 5 we will discuss the implications and future steps.

2 Data

We will begin by using the Total Water Use dataset provided by the Groundwater Sustainability Plan Annual Report datasets submitted to data.ca.gov

The data provided is sourced through Groundwater Sustainability Agencies and Alternative Agencies as part of the Groundwater Sustainability Plans. These agencies are local agencies developed to sustainably manage local groundwater. These agencies are required in areas with high and medium priority groundwater basins as defined by the Sustainable Groundwater Management Act, which aims to foster sustainable management of groundwater basins and facilitate data collection about groundwater. Areas with low priority basins may also form their own agencies. ("Groundwater Sustainability Agencies," n.d.)

The Groundwater Sustainability Plans are plans for how long-term sustainability will be attained for a basin. These plans are developed and outlined by the Groundwater Sustainability Agencies and Alternative Agencies. As part of this, every year, each agency submits a report to the Department of Water Resources through the Sustainable Groundwater Management Act's Portal. These reports make up the observations in the datasets. While the organizations are concerned with increasing sustainability of groundwater sources, the Total Water Use dataset they provide takes account of all water sources and water usage sectors, which makes it a good fit for our research.

The reported values are measured rather than the true values, so we are limited by the accuracy of the measurement tools and methods being used, but hopefully the large number of samples should reduce the effect of potential errors on the final result. Additionally, due to where the Groundwater Sustainability Agencies and Alternative Agencies are required, the data gathered primarily comes from high and medium priority basin areas. We'll have to consider this fact in the application of our model when making predictions.

Each record represents one annual report from one basin. All water usage and water sources are measured in acre-feet. The variables of interest for this paper are labelled "WUS_URBAN" and "WUS_AGRICULTURAL".

"WUS_URBAN" stands for "Water Use Sector Urban", and represents the total volume of water in acre-feet that was applied to the urban sector.

"WUS_AGRICULTURAL" stands for "Water Use Sector Agriculture", and represents the total volume of water in acre-feet that was applied to the agricultural sector.

There are other variables in this dataset tracking the volume of water applied to other sectors. These were used in aid of cleaning. 24 annual reports were excluded as reported volume of water use was 0 for all sectors. 23 of these observations were cross-referenced with groundwater source and use measurements from the Groundwater Extractions dataset provided as part of the Groundwater Sustainability Plan. Entries with measurements for water use in the Groundwater Extractions dataset but not in the Total Water Use dataset were dropped. This included:

- 2021-2024 reports for basin 4-002
- 2022-2024 reports for basin 4-013
- 2022-2024 reports for basin 5-022.16
- 2021 and 2024 reports for basin 6-005.01
- 2022-2024 reports for basin 7-021.04
- 2017-2024 reports for 7-024.01

Finally, the 2024 report for basin 5-004 was excluded because it had no entries for water use in either dataset, which didn't match the reports of previous years.

The basin numbering system referenced is standard in California.

After removing these entries, we still have 458 annual basin water usage reports from 2017-2024 across 99 sub-basins.

A scatter plot between agricultural water usage and urban water usage is shown in Figure 1.

Already we can see there are some significant deviations from a linear pattern These are likely caused by basin areas that are primarily agricultural, or primarily urban.

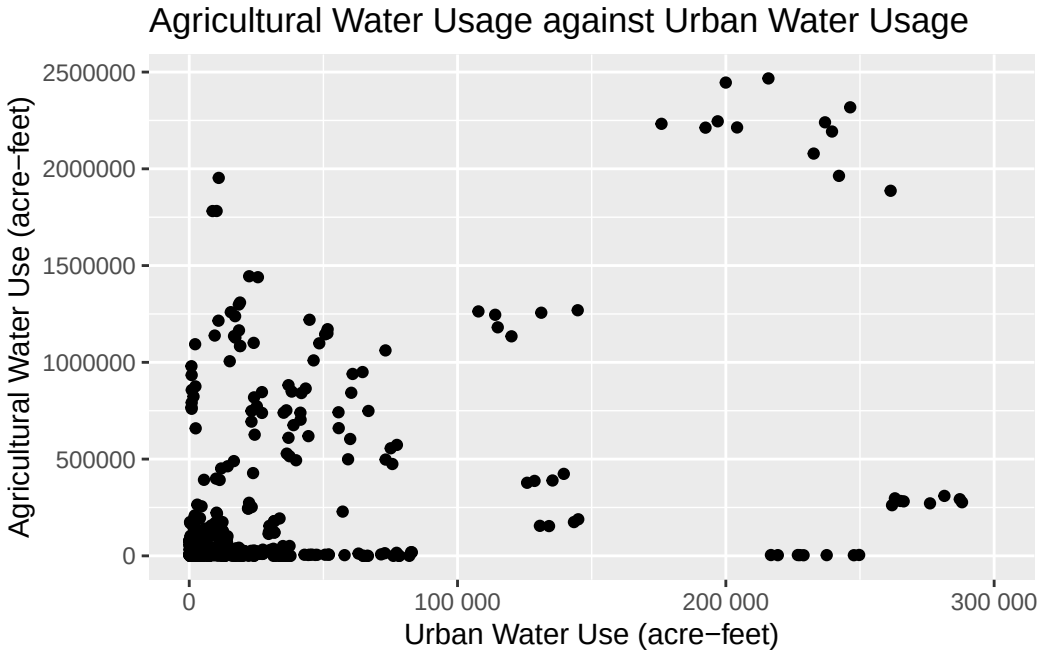

Figure 1: Scatter plot of urban water usage (x-axis) and agricultural water usage (y-axis). Both are measures of volume in units of acre-feet. The plot shows that there are many points in a vaguely linear fashion, though there are also many points at higher urban water use that may give our fit some trouble.

3 Methods

We will be using a linear regression model with equation:

$$Y_i = \beta_0 + \beta_1 * X_i + \epsilon_i$$

Where Y_i represents the estimated value for the agricultural water usage in acre-feet, X_i represents the measured value for the urban water usage in acre-feet, β_0 is the expected value of the agricultural water usage in acre-feet when we measure no urban water usage, and β_1 is the expected change in agricultural water usage when we see a unit change in urban water usage(X_i). ϵ_i represents the error term.

We will also be conducting a hypothesis test to affirm that there is a statistically significant relationship between the agricultural water usage and urban water usage recorded in our sample. We will use the null hypothesis $\beta_1 = 0$ and alternate hypothesis $\beta_1 \neq 0$. We will the test statistic $t^* = \frac{b_1}{s\{b_1\}}$. Using a 95% confidence level, our test statistic will need to be compared with $t(1 - \frac{\alpha}{2}; n - 2) = t(0.975; 456) = 1.965$.

We will also be examining the distribution of residuals against the fitted values, as the hypothesis test is only valid under certain assumptions We will be assuming that the relationship between errors and fitted values is linear, errors between each report are independent from each other, and that the errors have constant variance across all fitted values. The distribution of residuals against fitted values should reveal if any of these assumptions are broken.

The analysis has been implemented using the R-language (R Core Team 2025).

4 Results

After building our linear regression model, we end up with the following results:

$b_0 = 192974.5$ and $b_1 = 1.644$

From our model, we would expect to measure around 192974.5 acre-feet of agricultural water use when no urban water use is measured. For every acre-foot of urban water use measured, we should expect to see an additional 1.644 acre-feet of agricultural water use measured.

In Figure 2, we overlay the linear regression equation line over the scatter plot from Figure 1. We can see that the points showing high urban water usage but low agricultural usage have had a significant effect on the line fit to our model.

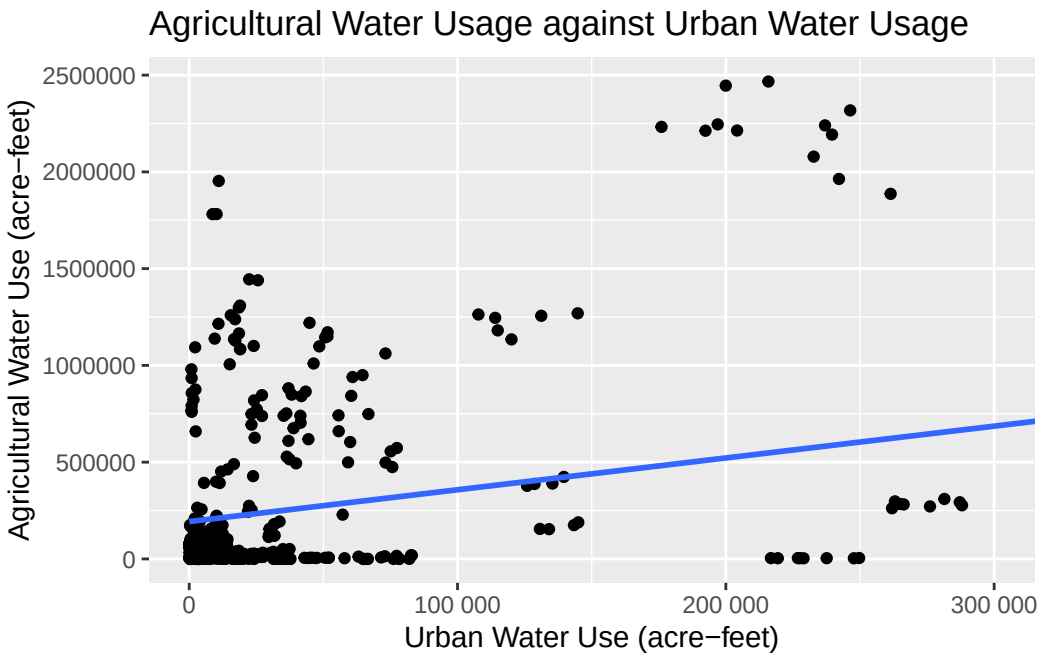

Figure 2: Scatter plot of urban water usage (x-axis) and agricultural water usage (y-axis) once again, with the linear equation found by the regression plotted. Both axes measure volume in units of acre-feet.

Because the slope is so small relative to the scale of our variables and the equation obtained from the linear regression is a poor fit for the points (as seen in Figure 2), we performed a hypothesis test using the null hypothesis $\beta_1 = 0$ and alternate hypothesis $\beta_1 \neq 0$.

We calculated our test statistic $t^* = \frac{b_1}{s\{b_1\}} = \frac{1.644}{0.272} = 6.043$. We can confirm that $|t^*| = 6.043 > 1.965$, so we reject the null hypothesis $\beta_1 = 0$.

By plotting the model residuals against the fitted values, we receive Figure 3.

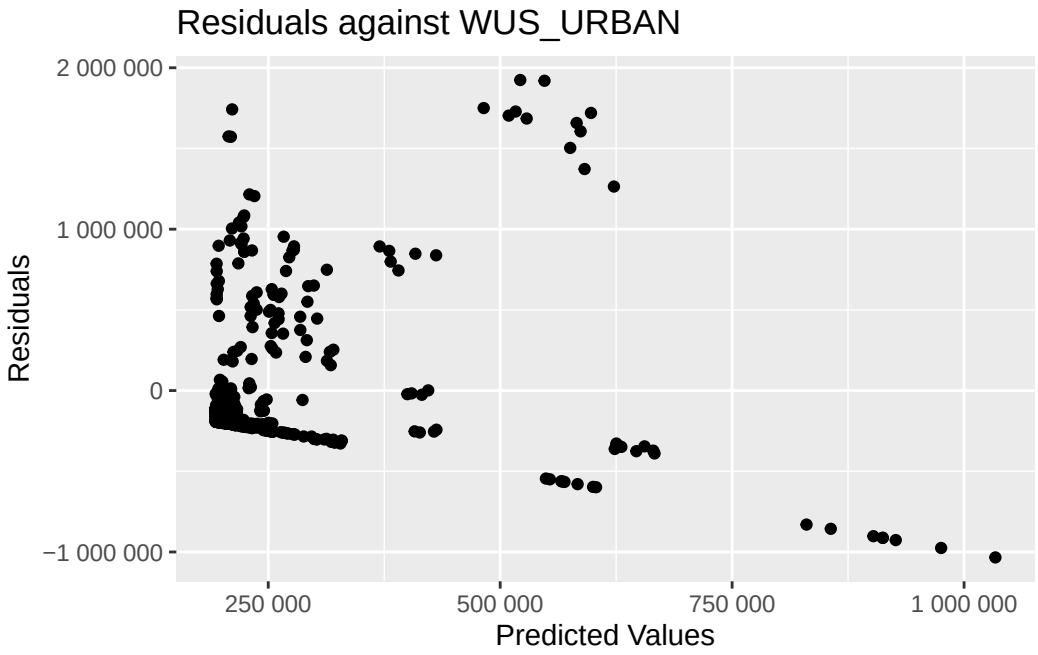

Figure 3: Scatter plot of the fitted values (x-axis) and residuals (y-axis) for a simple linear regression with urban water usage as the predictor variable and agricultural water usage as the response variable.

If our linear regression model is appropriate for the data, we should see a fairly even spread of the data points around the residuals = 0 line. However, there are several oddities:

1. Most of the points trend towards smaller values of the predictor variable.
2. There is a strong linear trend among some of the residuals.
3. Most of the points are negative. This is especially noticeable at the higher values for urban water usage.
4. There are some distinct clusters among the residuals.

All of these are indicators that the model is a poor predictor of agricultural water usage.

Additionally, as there is a clear lack of linearity in the residuals, the validity of our hypothesis test is also called into question.

5 Discussion

In this paper, we sought to determine if we could use measurements of urban water use to predict measurements of agricultural use. We started with data provided by the Department of Water Resources as part of the Sustainable Groundwater Management Plan. Then, we fit a linear regression model with urban water usage as the predictor variable and agricultural water usage as the response variable.

When examining the results, we could note a general positive relation between the two variables, confirmed by our hypothesis test. However, an examination of the residuals has shown that this is not a good model to use if we want to predict the agricultural water usage with the urban water usage.

In Figure 3, we noted many oddities, such as the significant number of negative residuals and a linear trend among the negative residuals. Additionally, the scatter plot of the residuals against the predictor variable almost looks like a skewed version of Figure 1, which might imply that little variance is explained by urban water use.

We did also observe some clusters. In the future, we can add variables to our model to try to explain some of the grouping in the residuals. Research into the potential relation between urban water usage and agricultural usage has brought up several potential variables not included in this dataset, such as population and attitudes and other legislation related to drought. Trends in urban water use and agricultural water use have been found in relation to both population and cultural attitudes post-drought. Urban water use has stayed relatively constant, despite growing populations (Mount, Hanak, and Peterson 2019). Agricultural water use also hasn't seen as much change as expected (Peterson et al. 2023).

There are also some limitations with the data we used. While total water use is reported regardless of source, the primary focus of the Groundwater Sustainability Agencies and Alternative Agencies that provide the data is to create a long-term sustainability for the groundwater basin they're responsible for. As a result, this data may not be directly taking into account other variables we may be interested in adding such as population and whether the area of concern is more urban or agricultural proportionally. If we want to take further steps into creating a model that can estimate agricultural water usage from urban water usage, we will have to find other data that can provide the variables we may be interested in adding.

Additionally, as the data is primarily gathered from areas with high and medium priority groundwater basins, it only reflects select portions of California, as opposed to the state as a whole.

In conclusion, the measured urban water usage alone is not enough to predict agricultural water use. However, we may be able to leverage other variables such as drought-related attitudes or urban-to-agricultural proportion classification to create a model that can give us a more accurate estimate of the agricultural water usage we can expect to see.

References

ABC10, and Rosemary Knight. 2023. "California's Groundwater Systems, Explained | Extended Interview." *YouTube*. ABC10. https://youtu.be/RylAOmJfMHk?si=CNx3xjbP6RJyrPem.

"California Water Use." 2018. USGS. https://www.usgs.gov/centers/california-water-science-center/science/california-water-use.

Cooley, Heather. 2020. "Urban and Agricultural Water Use in California, 1960-2015." Pacific Institute. https://pacinst.org/wp-content/uploads/2020/06/PI_Water_Use_Trends_June_2020.pdf.

"Groundwater Sustainability Agencies." n.d. California Department of Water Resources. https://water.ca.gov/Programs/Groundwater-Management/SGMA-Groundwater-Management/Groundwater-Sustainable-Agencies.

"Groundwater Sustainability Plans." 2018. water.ca.gov. https://water.ca.gov/Programs/Groundwater-Management/SGMA-Groundwater-Management/Groundwater-Sustainability-Plans.

Mount, Jeffrey, Ellen Hanak, and Caitlin Peterson. 2019. "Water Use in California." Public Policy Institute of California. https://www.ppic.org/publication/water-use-in-california/.

PBSNewsHour. 2025. "California Farms Face Pressure to Boost Efficiency as Water Supply Declines." *Youtube*. PBS NewsHour. https://youtu.be/9NvxwnhJS4s?si=iXps00QrI6dUPfJ1.

Peterson, Caitlin, Alvar Escriva-Bou, Josue Medellin-Azuara, and Spencer Cole. 2023. "Water Use in California's Agriculture." Public Policy Institute of California. https://www.ppic.org/publication/water-use-in-californias-agriculture/.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.