Regan Connell, Nathan Jeffries
ER131: Data, Environment and Society Final Project
University of California, Berkeley
reganconnell@berkeley.edu
njjeffries@berkeley.edu

Predicting Water Usage and Electricity from Water Usage based on Demographic, Climate, and Geographic Data

Abstract

As the U.S. climate becomes more volatile due to climate change and with the population ever-rising, water usage has become a more pressing topic than ever. The aim of our project is to predict water usage and the amount of electricity needed to produce water based on demographics, climate, and geographic data.
The type of model we set out to develop falls under the category of Land Use Regression, or spatial predicting. There is limited publicly available data about water and water processing energy usage across the United States, so we wanted to develop a predictor that could take in data about a specific location and output a prediction for the water and energy usage per capita for the location. In this way, we would be filling in the spatial gaps across the country which would help better inform things like policy decisions to help optimize two vital resources—water and energy.
We developed four types of models for both drinking water per capita predictions and electricity per capita needed to process water. These four models included K Nearest Neighbors, Linear regression, Ridge regression, and Lasso regression.
Through our model development process, we found that our KNN and Ridge model for water per capita prediction works best for larger, high water consumption cities and that no single feature gives a complete water consumption prediction. We found that none of the models we developed made adequate predictions for energy per capita usage, which we believe is due to fluctuating levels technology across the country—some cities have much more advanced, energy-saving technologies than others, so we would need to add more features to predict energy usage.

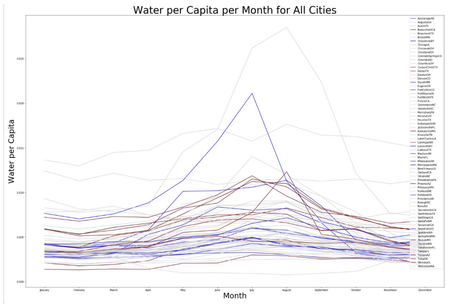
Background

Most of the U.S. lives within urban districts, and all utilities across the United States generate data on how much water is consumed in any given month, as well as the electricity needed to distribute and treat water. However, this information is not often collected or made public. The Consortium of Universities for the Advancement of Hydrologic Science published a paper with the aim of addressing this issue and set out to generate a primary data source that contains drinking and wastewater usage by month or year, as well as the energy needed to process each respective type of water usage. Their primary data set contains this information for 2012.

Upon reading the Advancement of Hydrologic Science's analysis of the water-energy nexus across the United States, we found a motivation behind generating spatial prediction models that would fill in the gaps of the missing locations in their existing primary data set. The implications of developing this type of prediction model would help inform better policy decisions, such as how to allocate resources to which cities in order to help decrease water consumption in inefficient locations.

One of the interesting conclusions that the Advancement of Hydrologic Science made was that often cities in hotter climates end up consuming the same amount across all months, while those in colder climates consume varying amounts according to the seasons. This led us to include climate data as part of our prediction model, assuming that climate data is usually a very readily available set of information for all locations across the United States.

With a changing climate, there is threat of increased temperatures globally. If temperature has an effect on drinking water consumption, there is a need to understand this effect in order to adequately address conserving this finite, vital resource. We intended to better understand this effect through developing our models that predict drinking water consumption and energy usage based off of climate, geographic, and demographics data. If water consumption could be predicted with this data with high accuracy, it would allow cities to know how much water they will need in coming years given the projected climate and population data. This would help predict how climate change may impact water consumption in varying locations.



Objective

1. The purpose of this project is to train and evaluate different models that are able to spatially predict water consumption per capita using features such as population, climate, and geographic data.

2. Another purpose of this project is to train and evaluate models that predict the amount of electricity per capita needed for drinking water consumption using features such as population, climate, and geographic data.

Both of these predictions are valuable because they would help fill in missing data across the United States about water usage and energy usage for locations that lack the capacity to collect this type of data. This would help address a resource allocation problem of which cities need help in decreasing water consumption per capita by providing a complete picture of the levels of consumption across the country. These models would also help cities across the United States predict what their future water consumption might look like given how their climate and populations are projected to change. It can give cities worst case scenarios or potential disaster scenarios to look out for. It could also be possible for cities to identify points at which the water consumption would exhaust the water supply. Should we achieve our purpose, it would build a stronger case for water conservation as a whole.

Interpretation and Conclusions

Our results have shown that the best possible model that fits our data would be a ridge regression for drinking water as well as K Nearest Neighbors with k=5 for our water per capita by month. The Ridge regression is accurate enough that it could be used to give a rough estimate of how much drinking water would be used for a given location with known population and climate data, however, we believe this model could be further refined if given more features as well as more data to develop a model. Our main limiting factor was the lack of data due to the availability of public water consumption data.

Our K Nearest Neighbors model for water per capita does a decent job of showing that geographically close locations have similar water per capita consumption for given months, which may be attributed to similar climates having similar consumption rates. This model would be made even more accurate if we had more cities to compare. With the results of these models, hopefully a policymaker would understand that as the climate proceeds to become more drastic, and regions begin to shift in average temperature, water use and electricity use are sure to be effected, so measures should be made to conserve as much water as possible to prepare for an uncertain climate future.

However, we failed to produce a model for electricity consumption based on population, geography, and climate data. Each city appears to have its own level of efficiency when it comes to water and neither K nearest neighbors nor our regressions could accurately model it. There are clearly more features we appear to be missing needed achieve a well-performing model for electricity per capita prediction.

We believe that the implications of uncovering that water consumption's relationship with population and climate should allow more individuals to consider the need for water conservation policy and a move towards not just mitigating climate change, but preparing to conserve whatever resources we have left, such as water. Otherwise, the U.S. could easily be looking at a water shortage by the end of this century, exacerbating our limited resource problems.

Our results are far from perfect, and more features and data points are needed to produce a more accurate model. The pleasant side of the story is that this model is inaccurate partly because some cities are overall more water-efficient than others, consuming less per capita because they have taken the initiative to implement water-efficient technologies and infrastructure. We assumed that water consumption is mostly determined by climate, location, and population, but other factors such as water conservation policy play a massive part in how much is consumed.

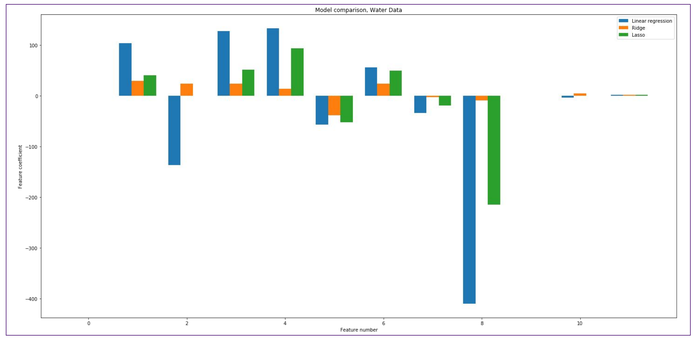


Figure 1. Linear Regression (Blue) gives the highest feature coefficients, as well as the highest Mean Squared Error(MSE) while Lasso (green) provides the second highest coefficients and the second highest MSE. Ridge performs the best as it's coefficients approach 0, and it's MSE is the smallest of all.

Data Description

We obtained our data sets from The Consortium of Universities for the Advancement of Hydrologic Science as well as U.S. Climate, Weather Underground, and the USGS for geographic data.

Structure— how are the data stored?: The energy and drinking water data that we merged into our final data frame came from individual csv files by city. The weather data originally was stored on a website that had filter features which helped get to each data point we needed in our data frame.
Granularity— how are the data aggregated?: Our data are aggregated by month and by city.
Scope— how much time, how many people, what spatial area?: Our data covers 67 cities in the United States from the year of 2012, including data from each month in each city.
Temporality— how is time represented in the data?: Time is represented in the data by month because the entire data set is for the year of 2012.
Faithfulness— are the data trustworthy?: The data are trustworthy based on our exclusion of cities with missing data, and due to our data exploration that revealed no significant outliers.

Forecasting and Prediction Modeling

Our main method of creating our model was to develop a K Nearest Neighbor approach as well as create a linear regression, ridge regression, and lasso regression model for water usage and a separate set of regressions and K Nearest Neighbors for Electricity from Water Usage. For our KNN model, k=5 predicts water consumption per capita most accurately, and ridge predicts water consumption per capita the best amongst the three regression models we built. We were unable to accurately predict electricity per capita consumption.

month number	citystate	electricity (kwh)	month	volume (mg)	population	T Max Avg	T Avg	T Min Avg	Precipitation (Inches)	Dew Point Avg	Snowdepth	Wind (Sea Level Pressure	Elevation	lats	lons
0	AnchorageAK	242089	Jan-12	629.485	226984	9.77	3.59	-3.29	0.75	-2.33	15.4	5.34	29.48	102	61.21738	-149.863
1	AnchorageAK	125174	Feb-12	600.465	226984	30.28	25.44	20.28	0.71	19.28	10	5.53	29.44	102	61.21738	-149.863
2	AnchorageAK	136445	Mar-12	620.445	226984	27.97	21.19	12.77	0.59	11.99	12.2	3.99	29.34	102	61.21738	-149.863
3	AnchorageAK	132454	Apr-12	656.376	226984	45.8	38.7	31.37	0.47	24.28	5.4	5.09	29.69	102	61.21738	-149.863
4	AnchorageAK	133798	May-12	700.916	226984	51.65	45.72	39.39	0.71	30.02	0.6	8.47	29.64	102	61.21738	-149.863
5	AnchorageAK	125531	Jun-12	731.679	226984	60.9	54.38	48.57	0.98	43.91	0	6.86	29.7	102	61.21738	-149.863
6	AnchorageAK	134204	Jul-12	742.507	226984	61.32	55.96	50.81	1.81	46.82	0	6.89	29.81	102	61.21738	-149.863
7	AnchorageAK	127764	Aug-12	692.234	226984	62.58	56.49	50.9	3.27	48.09	0	6.48	29.79	102	61.21738	-149.863
8	AnchorageAK	117449	Sep-12	614.592	226984	54.17	48.58	43.97	2.99	39.7	0.5	8.25	29.53	102	61.21738	-149.863
9	AnchorageAK	135852	Oct-12	610.043	226984	40.35	34.03	28.58	2.05	24.39	5.1	5.6	29.8	102	61.21738	-149.863
10	AnchorageAK	140042	Nov-12	594.91	226984	23.83	19.13	13.87	1.14	9.04	12.4	5.94	29.58	102	61.21738	-149.863
11	AnchorageAK	147553	Dec-12	615.117	226984	21.58	16.37	10.55	1.1	7.25	20.5	6.75	29.42	102	61.21738	-149.863
0	AugustaGA	1119318	Jan-12	916.51	160000	62.77	48.34	35.19	3.9	36.13	0.4	5.25	29.98	136	33.46667	-81.9667

Table 1. The first 13 points of our data frame, as you can see we had 21 points for every city, to total 804 points altogether.

References

- **The State of U.S. Urban Water: Data and the Energy-Water Nexus**, Christopher M. Chini and Ashlynn S. Stillwell
- **Weather Underground**
- **US Climate**
- **USGS Geographic Data**