Description of Project

Utilizing data provided by Irvine Ranch Water District, I will be conducting a causal analysis between water usage and climate variables including: precipitation, air temperature, relative humidity, wind speed, wind speed direction, solar radiation, evapotranspiration, and water usage. These climate variables are in a time series format, with six years of data. Each row represents one month of data per residential customer. There are over six million rows of data present in the dataset.

Determining a causal relationship is not a straightforward process, as correlation does not imply causation. Figuring out if there is a causal relationship between each climate variable and water usage will help bring insights into how the relationships are interpreted. If there are common trends, we can make actionable recommendations and derive important conclusions from them. Especially because saving water is extremely important, we can figure out the most efficient and optimal ways to save water, if we know the cause between seasonal changes and how much water people are using on a monthly basis. This project will allow us to quantify and gain causal insights between individual climate variables and water usage utilizing statistical analysis and machine learning models.

Steps

To estimate causal impacts of climate factors on water usage, we utilize causal forests. This method is motivated by the limitations of traditional methods such as Inverse Probability Weighting (IPW) and linear regressions. Linear regression with covariate adjustments requires specific assumptions of linearity and struggle with high dimensionality and nonlinear climate effects. Even with utilizing high-dimensional terms and interactions, there is a risk of overfitting or inconsistencies. In contrast, Causal Forests are a doubly robust method that combines propensity modeling and outcome modeling to reduce bias and variance . It uses machine learning to capture complex relationships and interactions and uses cross-fitting to reduce bias and obtain reliable causal estimates. This allows flexibility in climate and water relationships while maintaining the assumptions and methodologies of causal inference, providing accurate and interpretable effect estimates.

Using prior knowledge and correlation analysis, a Directed Acyclic Graph (DAG) is created. This visualizes causal pathways between variables. Our causal inference is testing for direct relationships between the treatment and the outcome variable, controlling for confounders. The DAG was created, keeping in mind that all of the climate variables are highly correlated with each other, except for precipitation. Each arrow represents a proposed causal pathway between variables, and allows the different causal pathways to be controlled for in the model. There are forks, mediators, and colliders. Since this study is observational, only forks and mediators are going to be controlled for in a direct causal estimate.

In this experiment, a separate causal forest for each climate variable is trained, treating each as the treatment variable, for each model respectively. The selected climate variable is the treatment, and the outcome is water usage, along with any other climate variables utilized as covariates to adjust for confounding. A large number of trees are in each forest to stabilize the estimates, and enabling confidence in splitting. We compute the average treatment effect for each climate variable, which is the mean of the estimate over the relevant sample. The average treatment effect represents the average causal impact on usage, of a one-unit increase in the climate variable, holding other factors constant.

Both a visual and numerical analysis of the outputs will be conducted, allowing for a deeper interpretation of the causal relationship between the climate variables and water usage. We can rank each of the variables in levels of causal strength and significance, if they are statistically significant and able to be concluded as a causal relationship.

References

Li, P., & Liu, Y. (2023, May 23). *Instability of Inverse Probability Weighting Methods and a Remedy for Nonignorable Missing Data*. Wiley Online LIbrary. https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12594

Athey, S., & Wager, S. (2019, February 20). *Estimating treatment effects with causal forests*. Arxiv. https://arxiv.org/pdf/1902.07409

Davis, J. M. V., & Heller, S. B. (2017, May 5). *Using causal forests to predict treatment heterogeneity: An application to summer jobs*. American Economic Review. https://www.aeaweb.org/articles?id=10.1257%2Faer.p20171000#:~:text=To%20estimate%20treatment%20heterogeneity%20in,interaction%20approaches%20would%20have%20missed

Chang, H., Praskievicz, S., Parandvash, H. (2014, August 27). *Sensitivity of Urban Water Consumption to Weather and Climate Variability at Multiple Temporal Scales: The Case of Portland, Oregon.* International Journal of Geospatial and Environmental Research. https://ijger-ojs-txstate.tdl.org/ijger/article/view/34

Marjanac, S., & Patton, L. (2018). Extreme weather event attribution science and climate change litigation: an essential step in the causal chain? *Journal of Energy & Natural Resources Law*, *36*(3), 265–298. https://doi.org/10.1080/02646811.2018.1451020

Deliverables

Part B -> Title page, abstract, introduction, literature review, methodology

Part C -> Completed final paper (Part B + results, discussion, conclusion, references, appendices)