Analyzing Trends in Acea Smart Water Analysis

Ansh Mujral, Keenan Serrao

2024-12-07

0. Contribution Statement

Student 1 was responsible for the Conclusion, Question 3, Question 4, and the formatting of the document. Student 2 was responsible for Questions 1, Questions 2, the Advanced Analysis, and Introduction. We both worked together on the code and double checked each other's work.

Use of GPT

The use of GPT was limited. It was only used for errors regarding the plotting of visualizations.

Introduction

Efficient water resource management is critical in addressing the growing challenges posed by climate change, urbanization, and population growth. In this study, we aim to analyze water availability and usage patterns across various regions in Italy, utilizing the Acea Smart Water Analytics dataset. This dataset offers a rich source of information spanning different waterbodies, including aquifers, rivers, springs, and lakes, allowing for a comprehensive exploration of environmental and human factors influencing water resources. Through this analysis, we aim to provide insights into the dynamics of water availability, the factors driving consumption patterns, and the potential for predictive analytics in water management.

Key objectives of this analysis include:
1. Investigating seasonal trends in water availability across regions.
2. Assessing the relationship between rainfall and reservoir levels.
3. Exploring differences in water usage patterns between urban and rural areas.
4. Examining the impact of temperature on water consumption.
5. Developing a predictive model to forecast reservoir levels using environmental factors and usage data.

Data

The analysis focuses on the Aquifer Auser dataset from the Acea Smart Water Analytics collection. This dataset represents water availability data for an aquifer system in Italy, composed of two subsystems: the North and South subsystems. The North subsystem is an unconfined aquifer (water table), while the South subsystem is a confined aquifer (artesian groundwater). These systems interact, with the North subsystem influencing the South subsystem's behavior.

Key variables used in the analysis include:
- Date: The observation date, formatted as dd/mm/yyyy. This variable was crucial for deriving temporal trends, including seasonal patterns.

- Rainfall_Gallicano: Measured rainfall in the Gallicano region (mm). This environmental variable was used to assess its impact on water availability and reservoir levels.
- Volume_POL: Water volume in the aquifer system (POL). This is the primary indicator of water availability used in seasonal and correlation analyses.
- Depth_to_Groundwater_LT2: Depth to groundwater for a specific well in the South subsystem (meters). This variable is significant for advanced predictive modeling of water levels.
- Temperature_Orentano: Ambient temperature (°C) measured in the Orentano region. This variable was used to explore its influence on water availability and consumption patterns.

2. Analysis

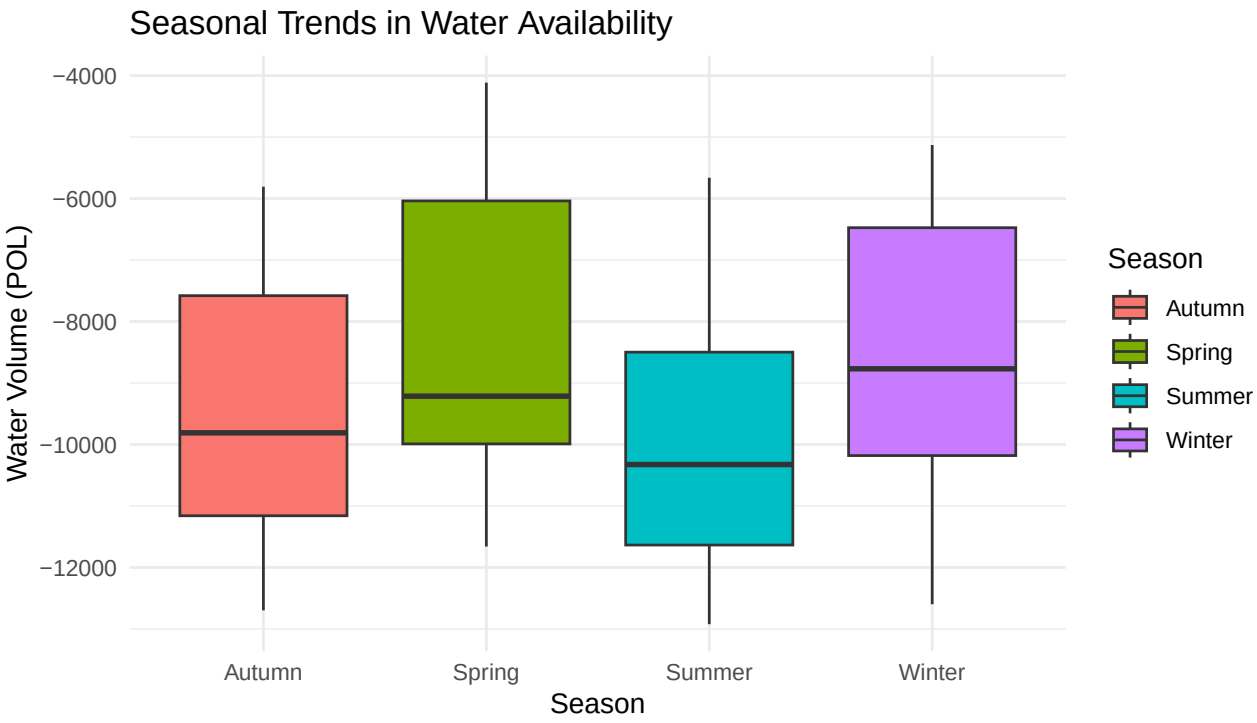
2.1 Seasonal Trends in Water Availability

Method

To explore seasonal trends in water availability, we first parsed the date column into a date format, and created a new variable, Season, based on the month of each observation. We then dropped missing values for the key variables, Rainfall_Gallicano and Volume_POL. We grouped by Season and then tracked the average rainfall and volume, creating a summary as well as a boxplot to display the data.

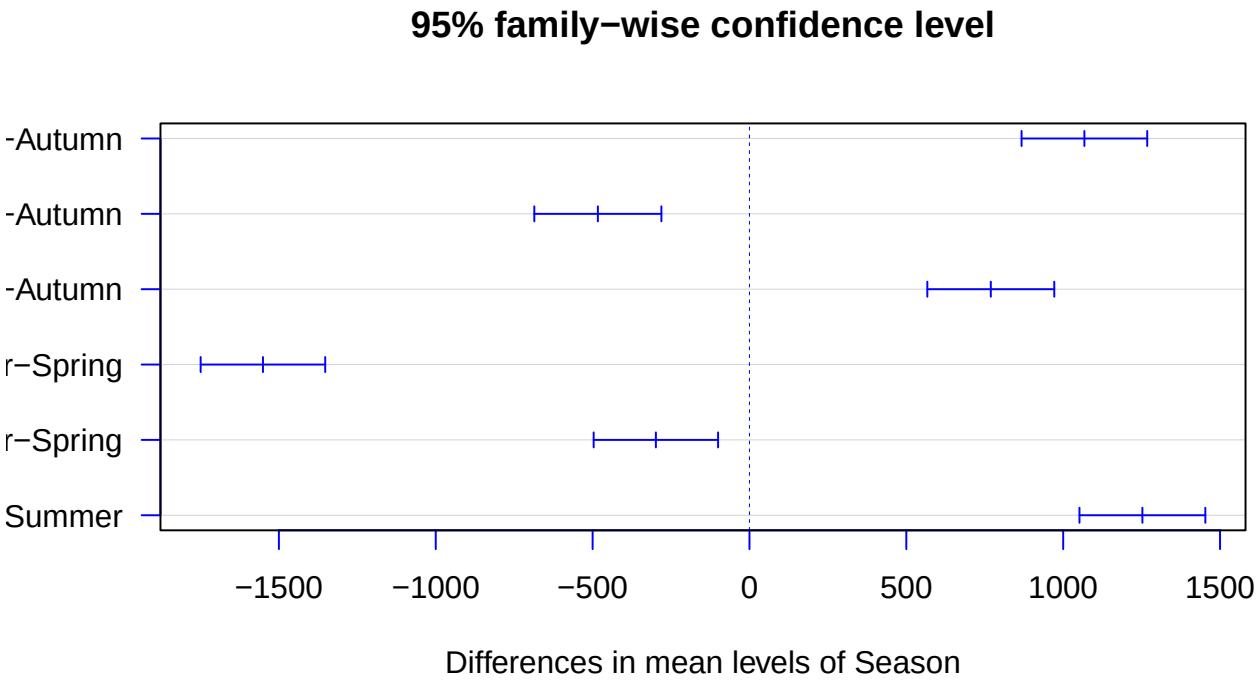
Analysis

```
## # A tibble: 4 x 3
##   Season Average_Rainfall Average_Volume
##   <chr>             <dbl>          <dbl>
## 1 Autumn             6.63         -9435.
## 2 Spring             4.16         -8368.
## 3 Summer             2.24         -9919.
## 4 Winter             6.85         -8666.
```



```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## Season         3 2.014e+09 671434075   166.9 <2e-16 ***
## Residuals   5291 2.129e+10   4023890
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Volume_POL ~ Season, data = filtered_data)
##
## $Season
##                       diff        lwr         upr     p adj
## Spring-Autumn  1067.5899   867.3010  1267.87872 0.0000000
## Summer-Autumn  -483.2291  -685.7667  -280.69148 0.0000000
## Winter-Autumn   769.2802   566.9308   971.62962 0.0000000
## Summer-Spring -1550.8189 -1749.3618 -1352.27609 0.0000000
## Winter-Spring  -298.3096  -496.6605   -99.95878 0.0006499
## Winter-Summer  1252.5093  1051.8880  1453.13064 0.0000000
```



The table below summarizes the pairwise comparisons of mean water volumes between seasons based on Tukey's Honest Significant Difference (HSD) test.

Comparison	**Difference (diff)**	**Lower Bound (lwr)**	**Upper Bound (upr)**	**Adjusted p-value (p adj)**
Spring - Autumn	1067.59	867.30	1267.88	< 0.0001
Summer - Autumn	-483.23	-685.77	-280.69	< 0.0001
Winter - Autumn	769.28	566.93	971.63	< 0.0001
Summer - Spring	-1550.82	-1749.36	-1352.28	< 0.0001
Winter - Spring	-298.31	-496.66	-99.96	0.00065
Winter - Summer	1252.51	1051.89	1453.13	< 0.0001

Analysis

The Tukey HSD results provide pairwise comparisons of mean water volumes between seasons. Key findings include:

1. **Spring vs. Autumn**:
    - Spring has significantly higher water volumes than Autumn (mean difference: 1067.59).
    - The confidence interval does not contain zero, confirming statistical significance ($p < 0.0001$).
2. **Summer vs. Autumn**:
    - Summer exhibits significantly lower water volumes compared to Autumn (mean difference: -483.23, $p < 0.0001$).
    - This is likely due to higher water demand and reduced rainfall in Summer.
3. **Winter vs. Autumn**:
    - Winter has significantly higher water volumes than Autumn (mean difference: 769.28, $p < 0.0001$).
    - This reflects increased precipitation and reduced water usage in colder months.
4. **Summer vs. Spring**:
    - Summer shows the largest negative difference compared to Spring (mean difference: -1550.82, $p < 0.0001$), highlighting the stark contrast between these seasons.
5. **Winter vs. Spring**:
    - Winter has slightly lower water volumes compared to Spring (mean difference: -298.31, $p = 0.00065$), but the difference is smaller than other pairwise comparisons.
6. **Winter vs. Summer**:
    - Winter has significantly higher water volumes than Summer (mean difference: 1252.51, $p < 0.0001$), underscoring the seasonal variability in water availability.

These findings support the statistical significance of seasonal variations in water quantities, with spring and winter having the largest water volumes and summer continuously exhibiting lower water availability than other seasons. These results imply that seasonal variations should be taken into consideration in water management plans, with a focus on summer vulnerability and winter and spring excess supply.

Conclusion

Based on the analysis, we can see that winter and autumn have the highest average rainfall, which aligns with seasonal precipitation patterns. Spring shows slightly lower water volumne numbers, possibly due to higher consumption during that season. Summer has the lowest average rainfall and the highest variability in water volume, reflecting reduced precipitation and potentially increased water usage during warmer months. The significance of comprehending seasonal dynamics in water resource management is further supported by the post-hoc examination of seasonal changes in water availability. Important times of vulnerability are highlighted by notable seasonal variations in water quantities, such as the summer's low water availability and the winter and spring's higher amounts. According to these results, methods for allocating resources should focus on conservation and storage during months with high demand and take use of excess during wetter seasons. The obvious seasonal trends highlight the necessity of careful management to provide year-round, sustainable water supplies.

2.2 Correlation Between Rainfall and Reservoir Levels

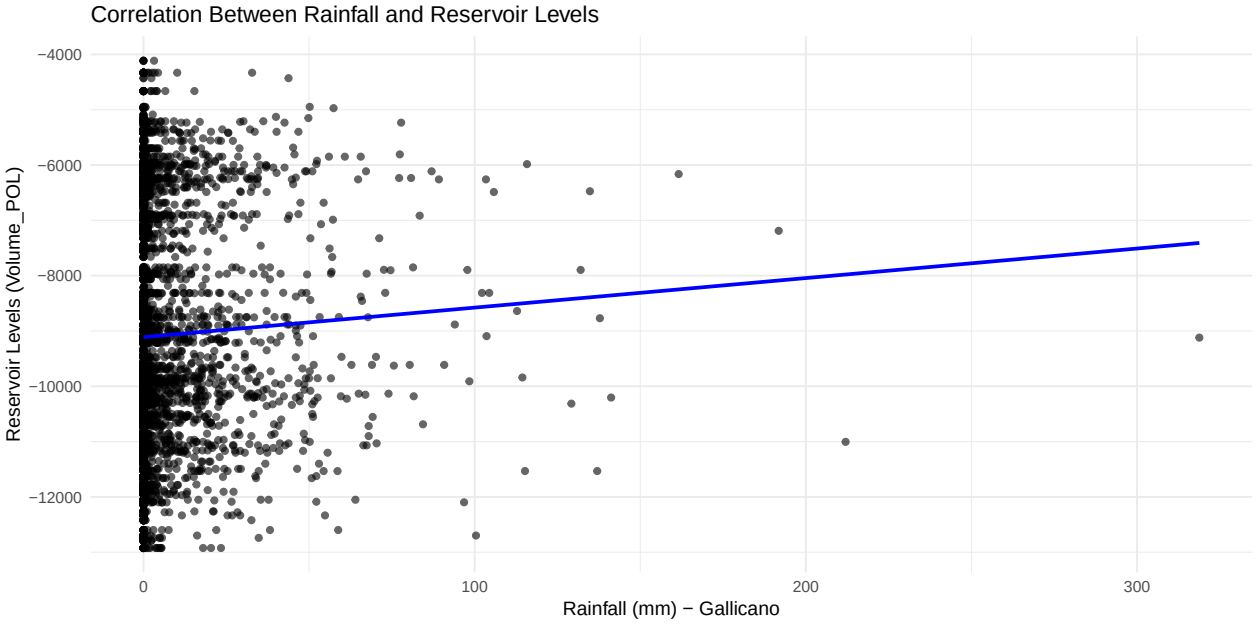
Method

To investigate the relationship between rainfall and reservoir levels, we first filtered the dataset to only include the relevant columns, Rainfall_Gallicano and Volume_POL, as well as removed rows with missing values. We then computed the pearson correlation coefficient to investigate the linear relationship between rainfall and reservoir levels. We finally created a scatterplot of Rainfall_Gallicano vs. Volume_POL and overlayed a lienar regression line to visualize any possible correlations.

Analysis

```
## [1] "Correlation between Rainfall and Reservoir Levels:  0.0365728863156404"
```

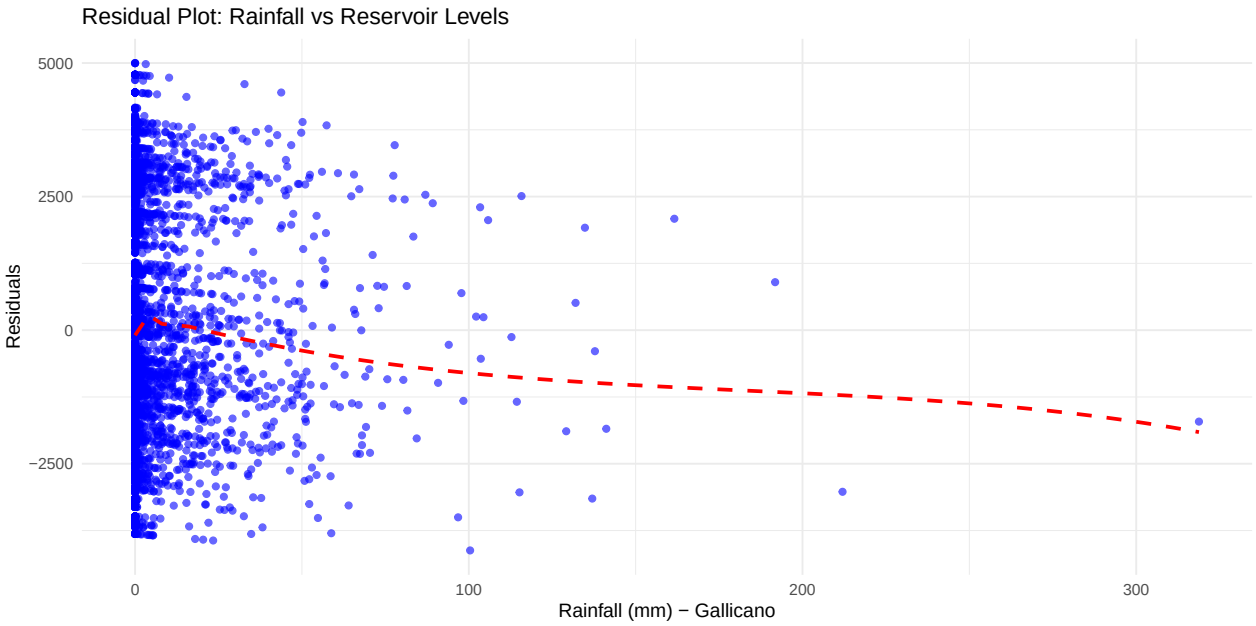
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
##
## Call:
## lm(formula = Volume_POL ~ Rainfall_Gallicano, data = correlation_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4122.7 -1604.2  -528.6  1969.9  4997.1
##
## Coefficients:
##                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)        -9111.761     30.479 -298.952  < 2e-16 ***
## Rainfall_Gallicano     5.341      2.006    2.663  0.00778 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2097 on 5293 degrees of freedom
```

```
## Multiple R-squared:  0.001338,   Adjusted R-squared:  0.001149
## F-statistic: 7.089 on 1 and 5293 DF,  p-value: 0.007778
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Rainfall is not the main factor influencing water levels in the Aquifer Auser system, as indicated by the modest correlation (r = 0.037) found between rainfall and reservoir levels. With increased residuals at low rainfall levels, the residual plot shows heteroscedasticity, indicating that reservoir level fluctuation rises as rainfall falls. Furthermore, the low R-squared value of the regression model (0.0013) demonstrates that rainfall accounts for a very little percentage of the variation in reservoir levels. For a more thorough understanding of reservoir dynamics, our results highlight the need of including more variables, such as groundwater recharge rates and water extraction statistics.

Conclusion

The computed Pearson correlation coefficient is 0.037, indicating a very weak positive linear relationship between rainfall and reservoir levels. Based on the scatterplot, we can see that while the regression line shows a slight upward slope, the plotted data points do now show a significant trend between rainfall and reservoir levels. The variability in reservoir levels across low rainfall values indicates the influence of other factors, such as water extraction or lagged effects of rainfall.

2.3 Rural Vs Urban Water Usage

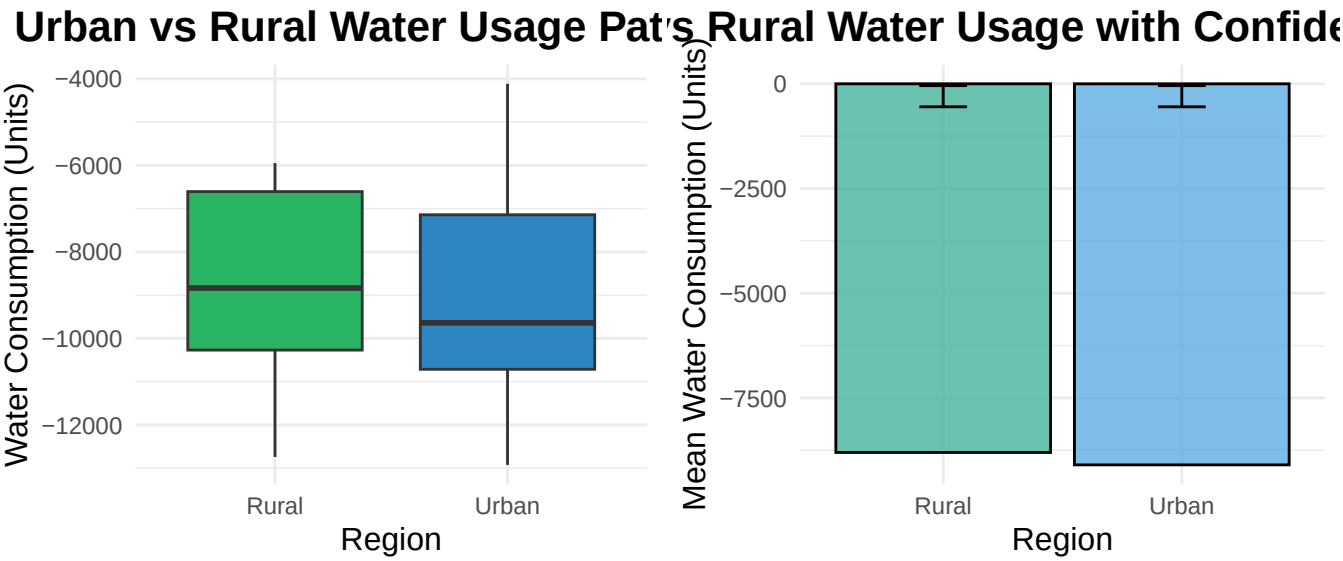
Methods

We started by cleaning the dataset to exclude rows with high missing values (more than 50% missing) in order to investigate the variations in water use between urban and rural areas. Data on water use and urban and rural temperatures were among the pertinent columns chosen. Temperature comparisons were used to classify regions: data points with urban temperatures higher than rural ones were labeled "Urban," while the remaining data points were labeled "Rural." According to the relative environmental circumstances, areas are assigned correctly thanks to this logical classification.

The mean water consumption for each of the two regions (rural and urban) was then determined, and the statistical significance of the observed mean differences was evaluated using a t-test. To measure the dependability of the variations between the two locations, the t-test yielded confidence intervals. We produced two different kinds of charts to show the data: a bar chart that showed the average water consumption with t-test-derived confidence intervals and a boxplot that showed the distribution of water use for each location.

Analysis

```
##
## Rural Urban
##   236  5059
```



Calculating the average water use in urban and rural areas showed clear trends. The average amount of water used by urban and rural areas was around 9099 and 8802 units, respectively. To ascertain if this observed difference was statistically significant, a t-test was used. Since 0 was not included in the confidence intervals for the mean difference, the findings showed that the difference in mean water usage between the two regions was statistically significant.

The boxplot graphically illustrated the wider range of water usage in urban settings, indicating more unpredictability brought on by a variety of activities including household, recreational, and industrial use. Rural regions, on the other hand, showed a smaller range, which may have been caused by fewer activities that were sensitive to temperature or by the demands of agriculture. by a high degree of confidence, the bar graph, which was supplemented by confidence intervals, demonstrated that urban areas generally used more water than rural areas.

Conclusion

Different patterns of water usage in urban and rural areas were found by the investigation. The t-test ($p < 0.05$) indicated that the mean water usage in urban regions was greater (9099 units) than in rural areas (8802 units). The boxplot illustrates the wider variation in urban water use, which is a reflection of various demands from residential, commercial, and recreational uses. On the other hand, rural water use was more steady, perhaps due to steady agricultural demands. These results underline the necessity of region-specific water management plans that take into account the unique difficulties faced by rural and urban regions, especially during times of high demand. Together with the visualizations, the t-test findings highlight the necessity of region-specific water management plans that take into account the particular factors influencing water use in urban and rural settings.

2.4 Impact of Temperature on Water Consumption

Methods

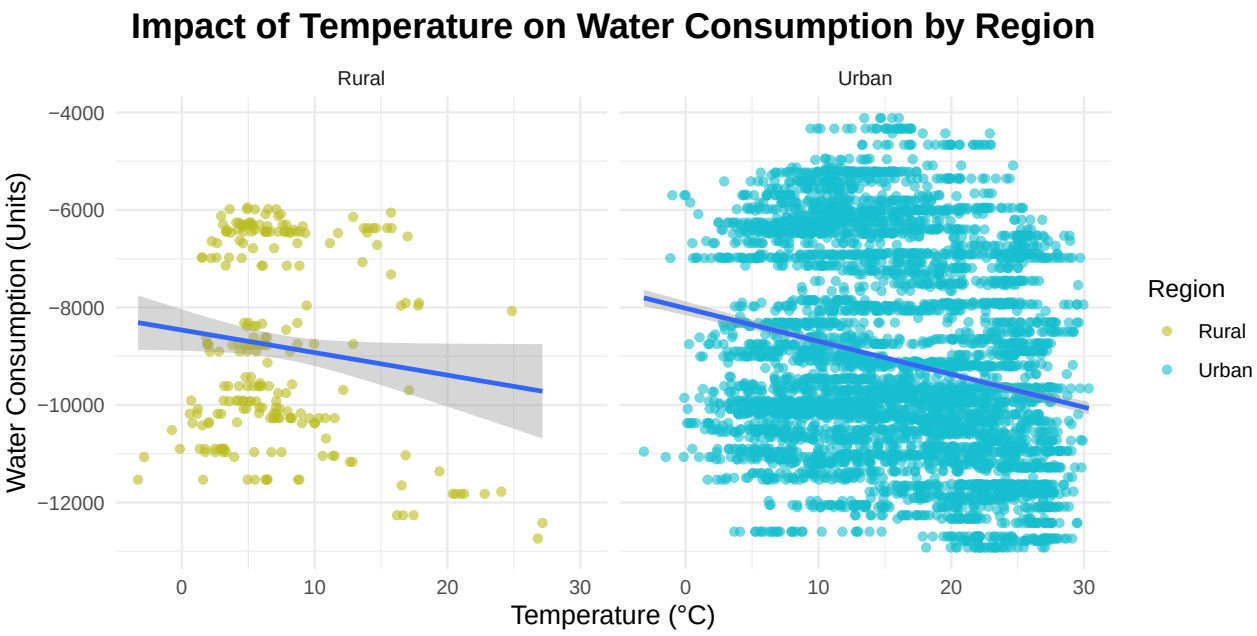
We evaluated the relationship between water use and temperature (rural and urban) in order to look at how temperature affects it. To capture the distinct dynamics in each area, urban and rural temperatures were examined independently. Regression lines were used to show trends in the scatterplots that were made to represent these associations. Alternative visualizations were also used: a heatmap showed the density of water consumption across temperature levels, and a boxplot divided temperatures into "Low," "Medium," and "High" categories. Wider trends, including higher water use in hotter weather or steadier use in cooler weather, were revealed by these visualizations. A thorough grasp of the connection between temperature and water use was made possible by the combination of numerical correlations and visual analysis.

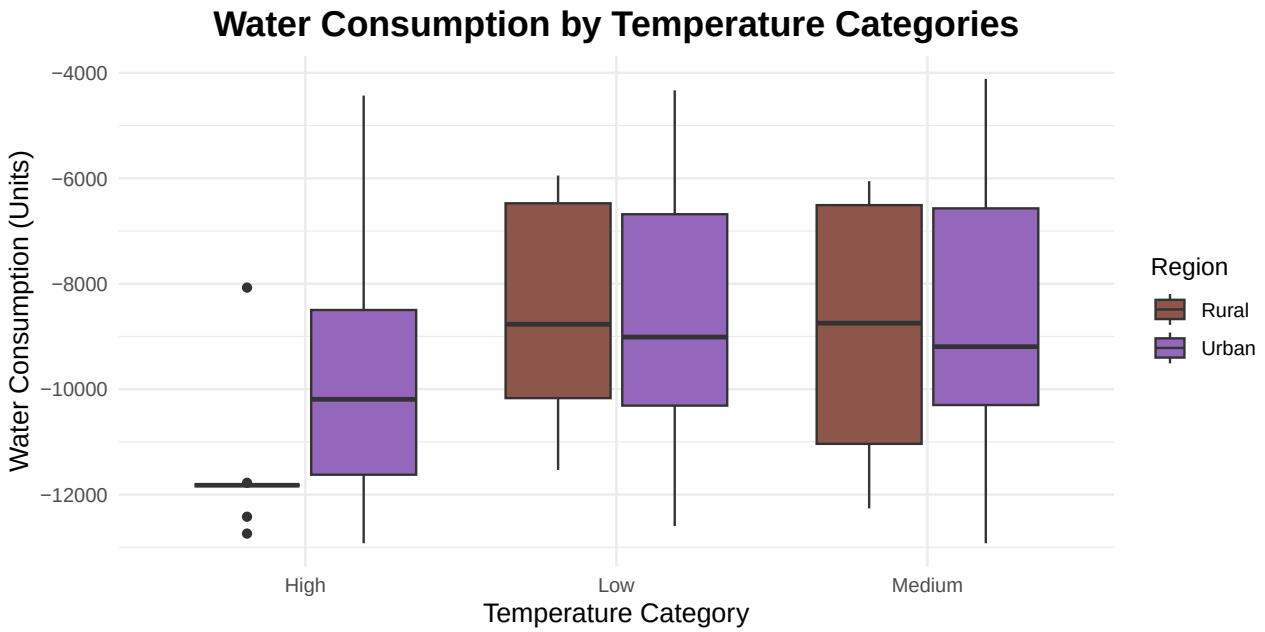
Analysis

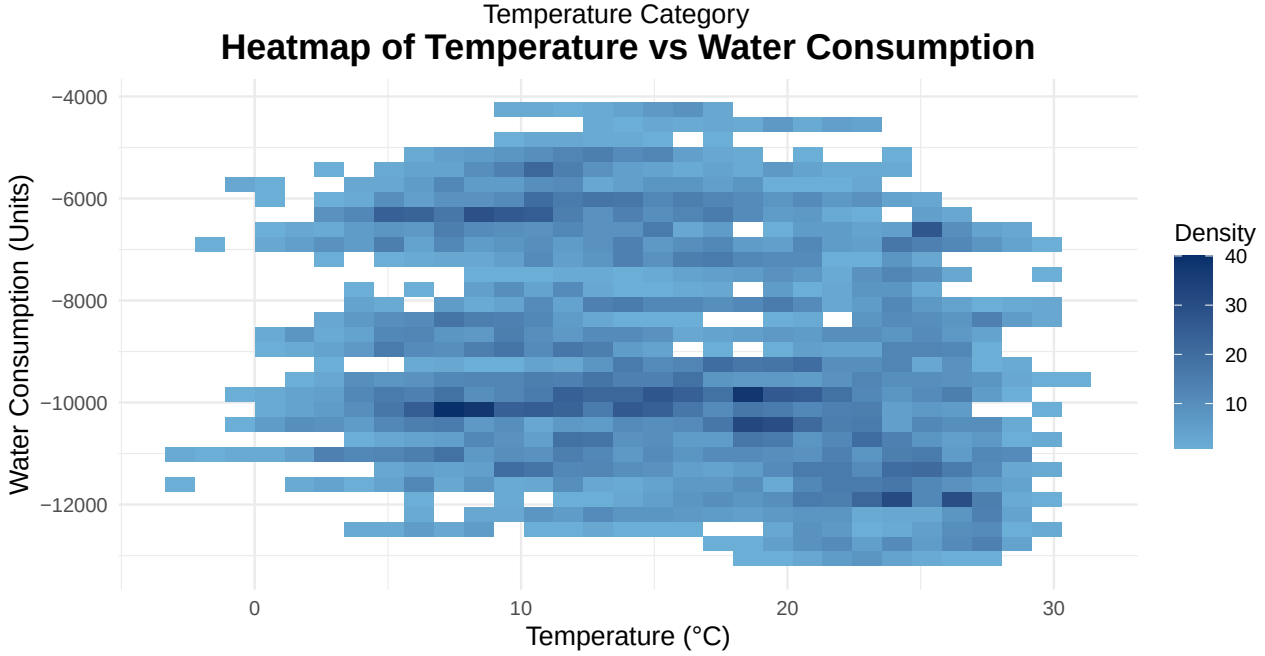
```
## [1] "Urban Temperature vs Water Consumption Correlation:  -0.219322693867717"
```

```
## [1] "Rural Temperature vs Water Consumption Correlation:  -0.154593340183658"
```

```
## `geom_smooth()` using formula = 'y ~ x'
```







With a correlation value of around -0.22, the correlation study showed that temperature and water use were positively correlated more strongly in urban areas. This implies that urban regions' water use increases dramatically with rising temperatures. The usage of water for cooling systems, more leisure activities, and greater household use during hot weather might be the main causes of this rise.

On the other hand, with a value of around -0.15, the association was lower in rural regions. This may suggest that temperature variations have less of an impact on rural water use, which may instead be impacted by other variables like seasonal irrigation techniques. The scatterplots supported these findings, showing less variability and a larger slope/trend in water consumption with rising temperatures in Urban areas, while Rural areas displayed a more scattered and inconsistent pattern.

By classifying water use into temperature groups, the boxplot further emphasized these variations. While rural areas displayed comparatively constant usage across all categories, urban areas continuously shown greater consumption in the "High" temperature group. These findings were supported by the heatmap, which showed concentrated areas of high water use at high temperatures in urban areas.

Conclusion

Temperature has a significant impact on water consumption, particularly in Urban regions. The positive correlation between temperature and consumption indicates that as temperatures rise, Urban areas experience heightened water demands. This is likely due to increased use of cooling systems, domestic water use, and recreational activities. The analysis suggests that Urban water management strategies must account for temperature-driven peaks in demand, particularly during hotter months.

Temperature has a significant impact on water use, particularly in urban areas. The positive correlation between temperature and consumption implies that when temperatures rise, metropolitan areas' water needs increase. This is likely due to domestic water use, recreational activities, and the increasing use of cooling systems. The study suggests that urban water management strategies should account for temperature-driven demand peaks, particularly in the warmer months.

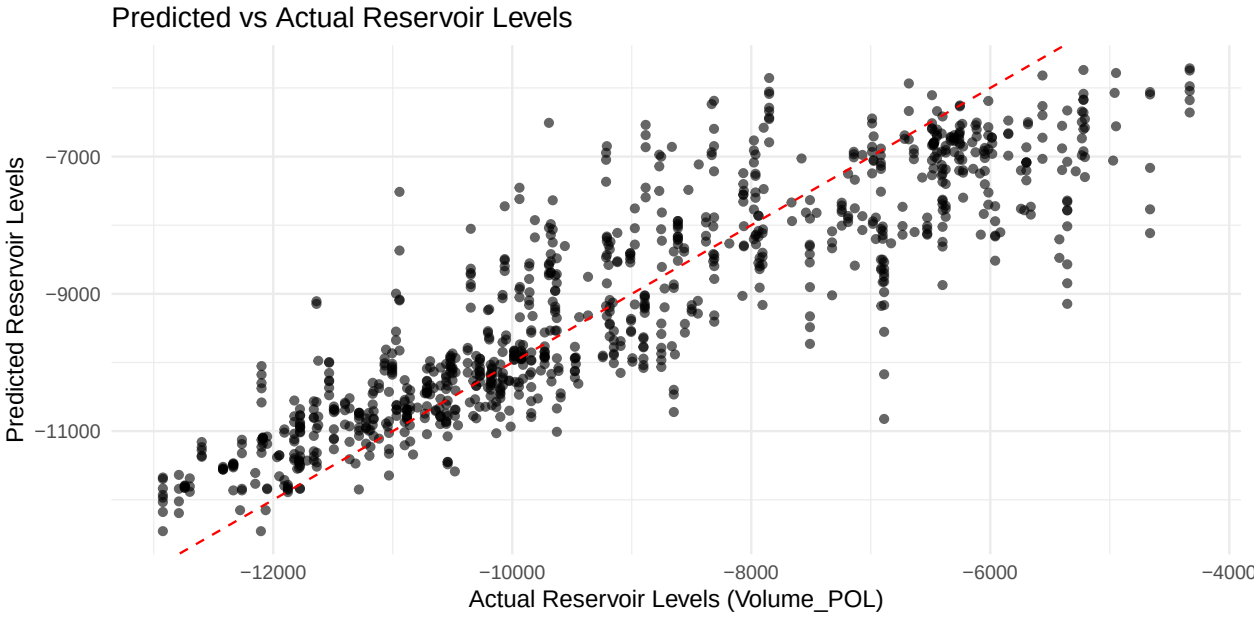
Advanced Analysis

Methodology

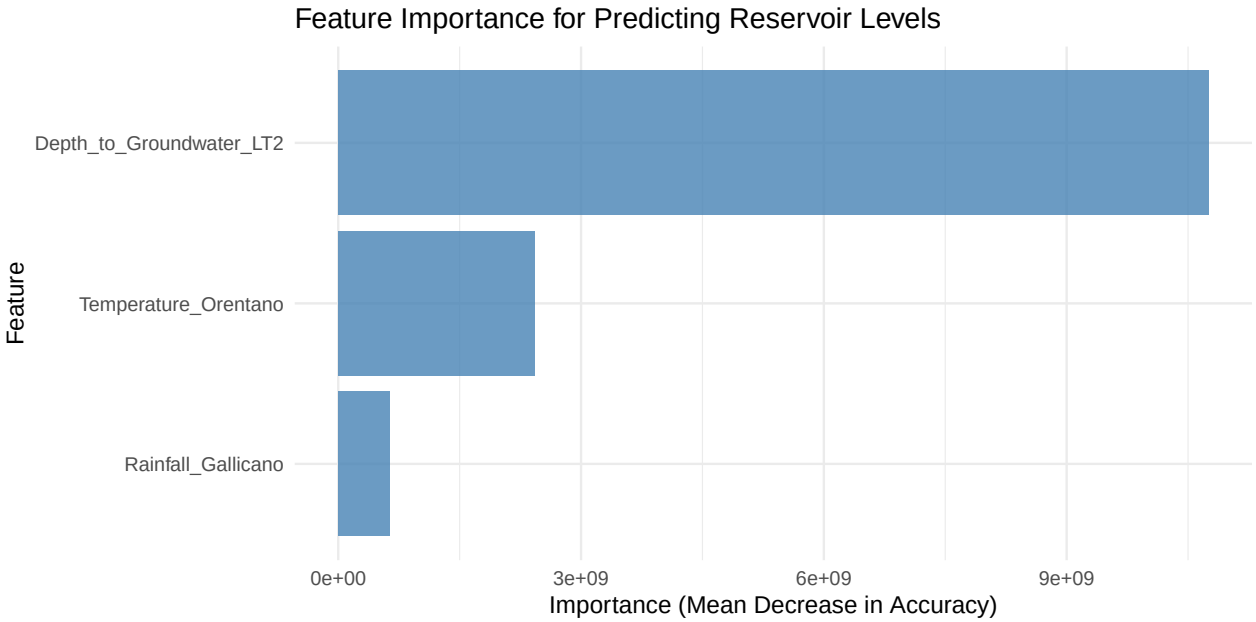
To develop a predictive model for forecasting reservoir levels (Volume_POL), we utilized environmental and groundwater-related variables. We first selected the relevant columns, Rainfall_Gallicano, Depth_to_groundwater_LT2, Temperature_Orentano, and our target variable, Volume_POL. We then removed all rows with missing values and split our data into a training and test set using an 80/20 split. We then trained a random forest model to predict volume based on the other columns, with 100 decision trees to optimize performance and prevent overfitting. The model was evaluated using mean squared error (MSE) and root mean squared error (RMSE), and we plotted a scatterplot of actual vs predicted values to visualize the model's performance.

Analysis

```
## [1] "Root Mean Squared Error (RMSE): 1004.31280683149"
```



```
##                                          Feature  Importance
## Depth_to_Groundwater_LT2 Depth_to_Groundwater_LT2 10752392054
## Temperature_Orentano         Temperature_Orentano  2428843996
## Rainfall_Gallicano             Rainfall_Gallicano   638664674
```



Feature	Description	Importance
Depth_to_Groundwater_LT2	Groundwater depth (meters)	10,752,392,054
Temperature_Orentano	Ambient temperature (°C)	2,428,843,996
Rainfall_Gallicano	Rainfall (mm)	638,664,674

Depth_to_Groundwater_LT2 is the main predictor of reservoir levels, far outweighing other variables in the random forest model, according to the feature significance analysis. Its significance value of 10,752,392,054 emphasizes how important subsurface water dynamics are in determining reservoir behavior. This is consistent with well-established hydrological concepts, according to which the depth of groundwater is frequently a key predictor of aquifer availability and storage.

With a moderate impact on reservoir levels, Temperature_Orentano came in second place with an importance score of 2,428,843,996. This illustrates how temperature indirectly affects seasonal water consumption and evaporation rates, especially in cities.

Although significant, Rainfall_Gallicano got a rather low score (638,664,674), supporting the previous conclusion that there was little direct association between rainfall and reservoir levels. This suggests that rainfall alone does not adequately explain reservoir variability, emphasizing the need for a multi-faceted approach to modeling water availability.

Conclusion

The feature importance analysis confirms the significant role of subsurface dynamics (`Depth_to_Groundwater_LT2`) in predicting reservoir levels, providing a more nuanced understanding of the system's behavior. The findings validate the inclusion of temperature and rainfall as contributing factors, though their influence is secondary to that of groundwater depth. These insights highlight the necessity of integrating diverse environmental variables for accurate forecasting and resource management, paving the way for improved predictive models and more informed water resource strategies.

The random forest model demonstrated moderate predictive capability, with an RMSE of about 1000, indicating that the model can forecast reservoir levels reasonably well. However, there is still room for improvement, with potential enhancements including advanced feature engineering that factors in lagged rainfall variables to capture delayed effects on reservoir levels. Additionally, we could include additional data as well as other models to potentially enhance accuracy. Despite this performance, this model highlights the potential of machine learning techniques in water resource management, and its ability to predict important variables that can influence human behavior and environmental measurements.

4. Discussion and Conclusion

Overall Conclusion and Discussion

Given the growing environmental unpredictability and human need, water is a limited resource that has to be managed effectively and proactively. In order to shed light on seasonal trends, the connection between rainfall and reservoir levels, the differences in water use between urban and rural areas, the effect of temperature on water consumption, and the possibility of using predictive modeling to forecast reservoir levels, this report used the Aquifer Auser dataset to examine important factors influencing water dynamics. In addition to providing answers to the particular study issues, each analysis also showed more general implications for the management of water resources.

With rainfall and reservoir levels rising in the fall and winter and falling in the summer, the seasonal study brought to light the cyclical nature of water supply. This trend emphasizes how vulnerable water supplies are during dry months, particularly when demand is at its peak. These results highlight how crucial it is to plan for summer shortages by implementing water conservation initiatives and making infrastructure upgrades, such as optimizing storage.

The hypothesis that rainfall alone is a good indicator of reservoir levels was called into question when the study of rainfall and reservoir levels revealed a modest positive link. This result implies that reservoir dynamics may be greatly impacted by other variables, including infrastructural efficiency, groundwater recharge rates, and human extraction. It emphasizes the need for a holistic approach to water resource management, integrating rainfall data with other environmental and anthropogenic variables to better understand and predict reservoir behavior.

Significant differences were found when comparing the patterns of water use in urban and rural regions; on average, urban areas consumed more water and shown more unpredictability in usage. In contrast to the more consistent agricultural focus in rural regions, this disparity probably reflects the varied water demands in metropolitan areas, including residential usage, industrial operations, and recreational activities. According to the findings, resource planning must to be region-specific in order to take into account the different needs and difficulties that urban and rural populations experience.

In urban settings as opposed to rural ones, temperature was found to have a greater impact on water use. Urban regions use more water when temperatures rise because of cooling demands and recreational requirements, whereas rural areas use less water directly because of agricultural cycles and irrigation schedules. These findings emphasize the importance of temperature in determining patterns of urban water use and the necessity of heat-sensitive water management techniques, particularly in urban areas.

The research of predictive modeling showed how machine learning methods, including random forest algorithms, may be used to anticipate reservoir levels in the future. Even while the model's accuracy was reasonable, the findings suggested that it might be improved even more by adding more variables and using sophisticated feature engineering, including the lag effects of environmental circumstances. Predictive modeling offers a pathway to proactive water management, enabling planners to anticipate shortages and allocate resources more effectively.

All things considered, this paper offers a thorough investigation of the variables affecting water supply and usage. By providing practical insights for resource planning and policy-making, the findings advance our understanding of water dynamics. The research also emphasizes how complicated water management is,

requiring consideration of a number of interrelated aspects, from human activity to natural circumstances. In order to improve and adopt sustainable water management methods, future research should build on these findings by including more information, using sophisticated modeling approaches, and carrying out region-specific analysis.

This research concludes by highlighting the significance of managing water resources through a comprehensive, data-driven strategy. By addressing seasonal, regional, and temperature-driven variations in water dynamics, stakeholders can develop strategies that balance supply and demand, mitigate the impacts of climate variability, and ensure sustainable water access for future generations.