Analyzing and Predicting the Flow Rate of the Madonna di Canneto Spring

1. Introduction

The **Madonna di Canneto spring** is a natural spring water source located in the Canneto valley in Italy at an altitude of **1010 meters above sea level**. A spring is a natural source of water that emerges from the ground, typically as a result of water from an underground source reaching the surface. This spring is specifically fed by the water catchment area of the Melfa River. This means that the water flowing through the spring is directly sourced from precipitation and runoff collected in the river's surrounding catchment area, rather than being stored in underground reservoirs. This unique characteristic makes the spring highly dependent on seasonal and regional water availability from the river's catchment. Springs like Madonna di Canneto are vital sources of freshwater and are often influenced by climatic and environmental conditions such as rainfall and temperature. In this report, we will study the interactions between **rainfall**, **temperature**, and the **flow rate** of the spring through the following questions:

1. Section 2.1. [EDA] What are trends in waterflow, temperature, and rainfall?
2. Section 2.2. [Stat. Analysis] Does rainfall & temperature significantly vary across the four seasons?
3. Section 2.3. [Trends] Do rainfall and temperature vary significantly across years?
4. Section 2.4. [Linear Regression] How do rainfall and temperature in the Settefrati region influence the flow rate?
5. Advanced Analysis. [Time Series] How well can a machine learning model predict groundwater levels using features like rainfall, temperature, and water extraction?

Data

`Water_Spring_Madonna_di_Canneto.csv` provides four key columns of data:

- **Date**: Records the dates on which the measurements were taken, starting from January1, 2012 to June 30, 2020.
- **Rainfall_Settefrati**: Represents the amount of rainfall (in millimeters) recorded in the Settefrati region.
- **Temperature_Settefrati**: Contains temperature readings (in degrees Celsius) from the Settefrati region.
- **Flow_Rate_Madonna_di_Canneto**: Records the flow rate (in cubic meters per second) of the spring itself.

The `Date` feature indicates that the dataset is time-series in nature, meaning the explanatory and response variables are ***not independent and identically distributed (I.I.D.)***. Instead, the sampling process is influenced by time dependency, as observations are collected sequentially over time. This temporal structure implies that values at one point in time may be correlated with values at previous or subsequent points, making standard I.I.D. assumptions inappropriate for this dataset.

2. Analysis

2.1. Raw Data and Initial Examination

Methods

This section applies Exploratory Data Analysis (EDA) to the `Water_Spring_Madonna_di_Canneto` dataset to explore trends, relationships, and distributions among `Rainfall_Settefrati`, `Temperature_Settefrati`, and `Flow_Rate_Madonna_di_Canneto`. The methods include:

Missing Value Analysis: Compute missing value counts to guide data cleaning.

Quantitative Summary: Calculate key statistics (mean, median, range, quartiles).

Qualitative Analysis: Visualize variable distributions to detect patterns, skewness, and outliers.

Correlation Analysis: Assess relationships with correlation coefficients with a heatmap.

Trend Analysis: Plot flow rate over time, highlighting a 7-day rolling average to smooth fluctuations.

Analysis

```
## Summary Statistics:

##   Rainfall_Min Rainfall_Mean Rainfall_Max Temperature_Min Temperature_Mean
## 1            0      4.252444        140.8            -4.9         13.56173
##   Temperature_Max Flow_Rate_Min Flow_Rate_Mean Flow_Rate_Max
## 1            31.1      187.7532       263.1605       300.161
```

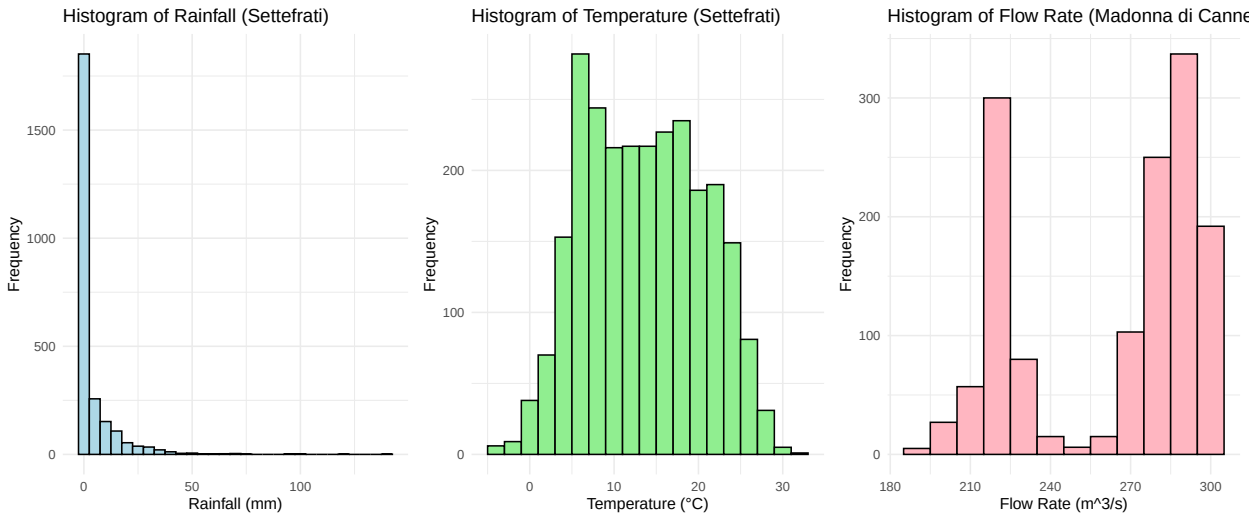

Figure 1: Distribution of rainfall, temperature, and flow rate.

Conclusion

The exploratory data analysis (EDA) provided several insights into the dataset, which includes variables for rainfall (`Rainfall_Settefrati`), temperature (`Temperature_Settefrati`), and flow rate (`Flow_Rate_Madonna_di_Canneto`).

The analysis revealed a notable amount of missing data in the dataset. Specifically, 556 out of 3,105 entries (~18%) are missing for `Rainfall_Settefrati` and `Temperature_Settefrati`, and 1,726 entries (~55.6%) are missing for `Flow_Rate_Madonna_di_Canneto`. To address this, appropriate strategies will be applied based on the variable. Based on future Sections 2.2 and 2.3, random imputation will be used for seasonal data after confirming significant differences across seasons and that these differences don't vary across years. For the response variable, `Flow_Rate_Madonna_di_Canneto`, entries with missing values will be excluded in Section 2.4 and the Advanced Analysis, as lagged values of the response will serve as features. Imputation methods, such as rolling averages, are avoided for the response variable because they could interfere with the temporal dependencies captured through lagged values.

The summary statistics highlighted distinct characteristics of each variable. The flow rate has a mean of approximately 263.2 m^3/s and a maximum of 300.2 m^3/s, indicating a range concentrated around the higher end. Rainfall values are heavily skewed toward zero, with a mean of 4.25 mm but a maximum of 140.8 mm. Temperature shows a relatively normal distribution, with a mean of 13.56 °C and a maximum of 31.1 °C.

The histograms provide visual confirmation of the patterns observed in the summaries: The rainfall histogram is highly skewed, with most data concentrated at or near zero. The temperature histogram demonstrates a roughly symmetric distribution, centered around its mean, indicating seasonality or natural variability. The flow rate histogram shows a bimodal distribution, with peaks near 210 m^3/s and 300 m^3/s, suggesting potential underlying seasonal or operational dynamics influencing the spring's flow rate.

Referring to Appendix Figure 8, the correlation matrix reveals weak relationships between variables. `Rainfall_Settefrati` and `Flow_Rate_Madonna_di_Canneto` exhibit a slight negative correlation (-0.036), suggesting minimal direct influence of rainfall on flow rate. Similarly, `Temperature_Settefrati` and `Flow_Rate_Madonna_di_Canneto` also show a weak negative correlation (-0.076), implying temperature has limited direct impact on flow rate.

Referring to Appendix Figure 9, the time-series analysis of the flow rate reveals abrupt drops and fluctuations over the years, suggesting the presence of seasonal or other external factors impacting flow dynamics. Incorporating a 7-day rolling average smooths these variations, highlighting underlying trends and long-term stability in the flow rate. However, this rolling average is applied only in this section to better visualize the data; it **will not** be applied before predictive task because lagged values in flow rate will be used as a predictor.

This initial analysis indicates that while rainfall and temperature may not directly correlate strongly with flow rate, additional investigations, such as lagged correlations or interactions, may uncover hidden relationships. The bimodal flow rate distribution and weak correlations highlight the potential for complex interactions among the variables, meriting further modeling and statistical analyses.

2.2 Rainfall and Temperature Distributions Across Seasons

Methods

To explore rainfall and temperature across the four seasons, each data entry was assigned to one of four seasons (`Spring`, `Summer`, `Fall`, or `Winter`) and years that have <95% data coverage are excluded to ensure reliability in reconstruction of seasonal distributions. Then rainfall was grouped into 2 mm bins (0–25 mm), and temperature into 5°C bins (0–40°C) so that the total number of days per bin was calculated for each season and normalized by dividing by the number of valid years, producing average annual counts for each bin. Using the Kruskal-Wallis Test, a non-parametric test, evaluated whether the distributions of rainfall and temperature differ significantly across seasons. Pairwise Wilcoxon Tests follow-uped pairwise comparisons to identify specific seasonal differences, with Bonferroni correction follows a specific distribution (e.g., normal distribution). It is particularly suitable for this time-series dataset, as it accounts for non-IID data.

By asking whether rainfall varies significantly across seasons and employing robust non-parametric methods, this analysis respects the time-dependent nature of the dataset. It provides insights into seasonal rainfall variability, informing future steps in addressing missing data or modeling rainfall-dependent phenomena.

Analysis

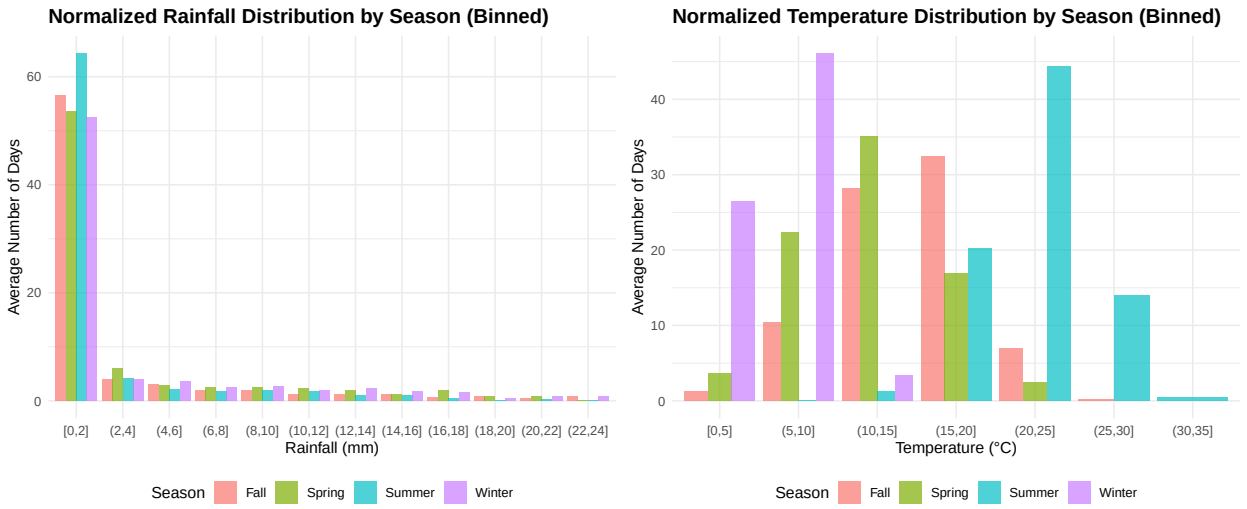

Figure 2: Normalized distribution of rainfall and temperature by season.

```
## Kruskal-Wallis Test for Rainfall:
```

```
## Chi-squared = 46.27, df = 3, p-value = 0.0000
```

```
## Significant pairwise differences for Rainfall (adjusted p-values):
##                Fall       Spring       Summer
## Spring 1.000000e+00           NA           NA
## Summer 1.204374e-05 2.267178e-08           NA
## Winter 1.000000e+00 1.000000e+00 1.073168e-08

## Kruskal-Wallis Test for Temperature:

## Chi-squared = 1826.87, df = 3, p-value = 0.0000

## Significant pairwise differences for Temperature (adjusted p-values):
##                 Fall        Spring       Summer
## Spring  2.375035e-21            NA           NA
## Summer 4.170449e-149 4.750595e-185           NA
## Winter 5.627196e-168 4.574305e-129 4.73787e-209
```

Conclusion

The analysis of seasonal rainfall and temperature distributions reveals notable differences across the four seasons, as demonstrated by both the visualizations and the statistical tests. The normalized rainfall distribution histograms show that low rainfall values (0–2 mm) dominate across all seasons, a pattern consistent with the Mediterranean climate. However, statistical tests reveal significant differences for higher rainfall bins. The Kruskal-Wallis Test test confirms that rainfall distributions vary significantly across seasons ($p < 0.05$). Adjusted for multiple comparisons using the Bonferroni method, the Pairwise Wilcoxon Tests tests indicate that **Summer** differs significantly from all other seasons: Spring ($p = 2.3 \times 10^{-8}$), Fall ($p = 1.2 \times 10^{-5}$), Winter ($p = 1.1 \times 10^{-8}$).

The normalized temperature distribution histograms illustrate clear distinctions among the seasons which **Summer** having the highest average temperature values and **Winter** the lowest. The Kruskal-Wallis Test test confirms significant differences in temperature distributions across all seasons ($p < 2.2 \times 10^{-16}$). The Pairwise Wilcoxon Tests, after Bonferroni adjustment, significant differences are observed between every pair of seasons ($p < 2 \times 10^{-16}$). These results emphasize the pronounced, distinct seasonal variability in temperature.

The findings highlight strong seasonal dynamics in both rainfall and temperature. Rainfall exhibits variability, with Summer standing out as a distinct season, while the other seasons share more similar patterns. Temperature distributions, on the other hand, are distinct for every season. These insights are critical for understanding hydrological and ecological processes influenced by seasonal climatic variations and provide a foundation for further modeling, such as seasonal water resource management or flow rate prediction. The distinct seasonal patterns could inform the design of seasonal water management strategies or the modeling of hydrological phenomena sensitive to rainfall variability.

The next section will examine whether these distributions remain consistent across years, in which the seasonal distribution of rainfall and temperature could be used for random imputation of explanatory variables.

2.3. Rainfall and Temperature Across Years

Methods

To investigate rainfall and temperature across years, annual trends in rainfall and temperature were analyzed using non-parametric statistical tests, which are particularly suitable for time-dependent data. The `Water_Spring_Madonna_di_Canneto` dataset was processed to calculate the annual averages of rainfall (`Rainfall_Settefrati`) and temperature (`Temperature_Settefrati`) for each year and missing values were excluded. The Mann-Kendall test, a non-parametric method used to detect trends in time-series data, will be used to compare the annual averages of rainfall or temperature across years to identify whether there is a significant monotonic trend. Sen's slope, a method to estimate the rate of change in a time series, will be used to calculate the median of all possible slopes between pairs of data points, offering a robust estimate of the trend's magnitude.

Analysis

```
##
## **Mann-Kendall Test for Rainfall:**
## Tau: 0.238 | S: 5 | p-value: 0.5480
```

```
##
## **Mann-Kendall Test for Temperature:**
## Tau: 0.143 | S: 3 | p-value: 0.7639
```

```
##
## **Sen's Slope for Rainfall (mm/year):**
## Slope: 0.0759 | CI: [-0.5300, 0.7553]
```

```
##
## **Sen's Slope for Temperature (C/year):**
## Slope: 0.0468 | CI: [-0.1314, 0.2341]
```

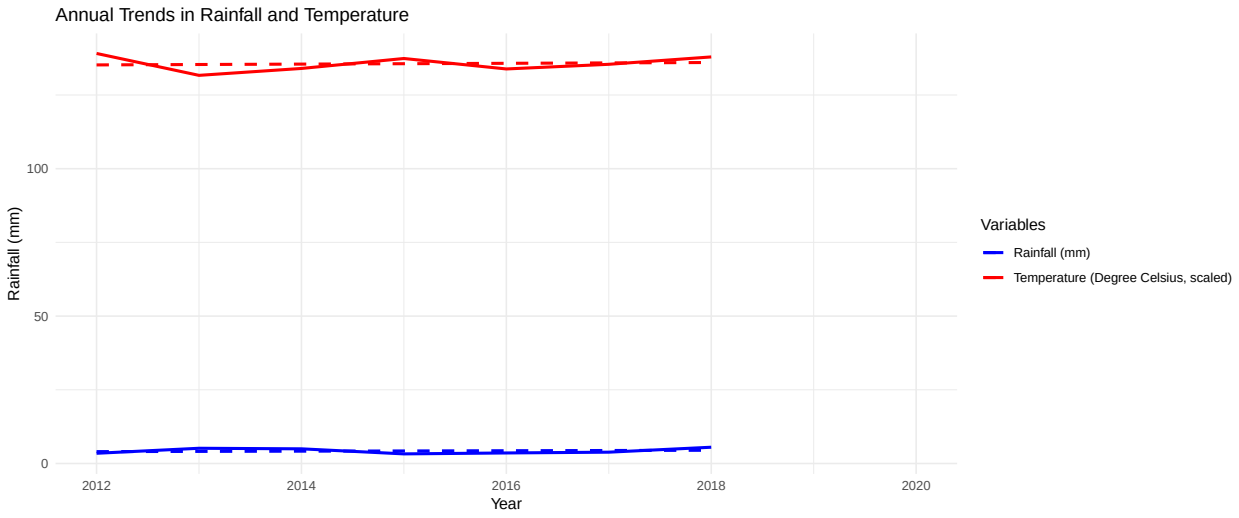

Figure 3: Annual trends in rainfall and temperature.

Conclusion

The results of the Mann-Kendall trend test and Sen's Slope estimation indicate no statistically significant trends in either variable over the analyzed period.

The Mann-Kendall test for annual rainfall produced a p-value of 0.548, exceeding the standard significance level of 0.05. This indicates no evidence of a monotonic trend in rainfall over the years. The test statistic $S = 5$ and Kendall's tau of 0.238 suggest a weak positive association, but it is not statistically significant. Sen's Slope estimation provided a slope of $0.0759\,\text{mm/year}$ with a 95% confidence interval of $[-0.53, 0.76]\,\text{mm/year}$. The inclusion of 0 within this interval further confirms the lack of a significant change in rainfall over time.

Similarly, the Mann-Kendall test for annual temperature yielded a p-value of 0.764, indicating no significant trend. The test statistic $S = 3$ and Kendall's tau of 0.143 point to a very weak positive association. Sen's Slope estimation showed a slope of $0.0468\,^{\circ}\text{C/year}$ with a 95% confidence interval of $[-0.13, 0.23]\,^{\circ}\text{C/year}$. The inclusion of 0 in this interval confirms no significant change in annual temperature during the study period.

The absence of significant trends in both rainfall and temperature suggests that, within the dataset's timeframe, environmental conditions have remained relatively stable. This stability implies that external factors other than climatic changes likely influence the flow rate of the Madonna di Canneto water spring. However, the small sample size ($n = 7\,\text{years}$) limits the statistical power of the tests, and longer-term data would be required to draw more robust conclusions about potential trends.

2.4. Linear Regression

Methods

In this section, we first tackled data cleaning by formatting the Date column and categorizing each data into their respective Spring, Summer, Fall or Winter season. Then we normalized rainfall and temperature distribution across seasons and performed random imputation based on seasonal distributions. Next we checked the data for conditions that would make it a good candidate for a linear regression:

1. Linearity - Is there a linear relationship between explanatory and response variables?
2. Homoscedasticity - Is there constant variance in residuals for the estimated flow rate?
3. Independence - Are the data points independent of each other?
4. Normality - Do the residuals follow a normal distribution?
5. Multicollinearity - Independent variables should not have strong correlations with each other.

Lastly, a multiple linear regression model is executed and its results interpreted.

Analysis

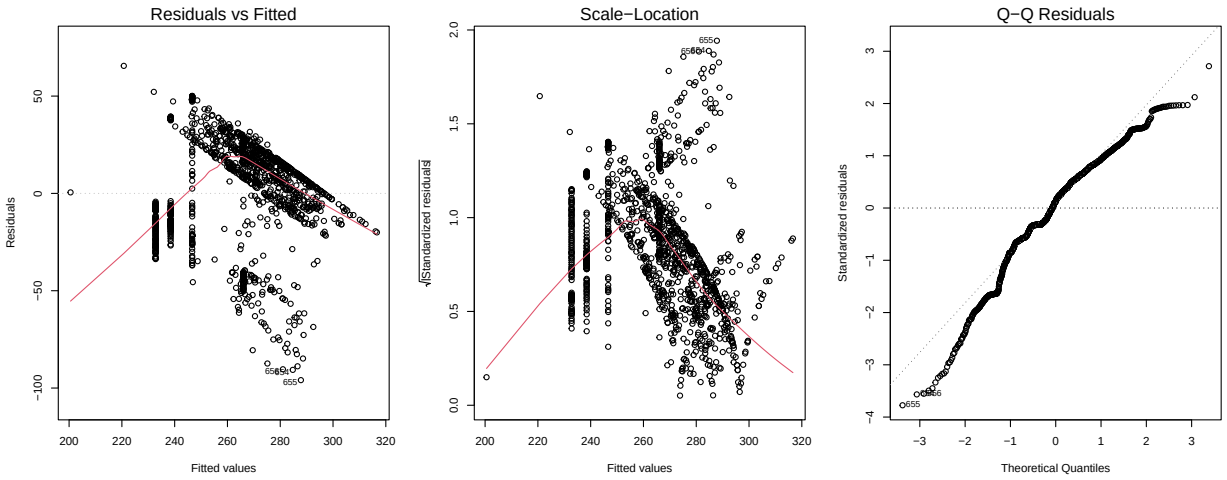

Figure 4: From left to right, criteria: linearity, homoscedasticity, normality.

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 0.1153, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## Rainfall_Settefrati     1.095269  1        1.046551
## Temperature_Settefrati 1.358798  1        1.165675
## Season                  1.304715  3        1.045328
```

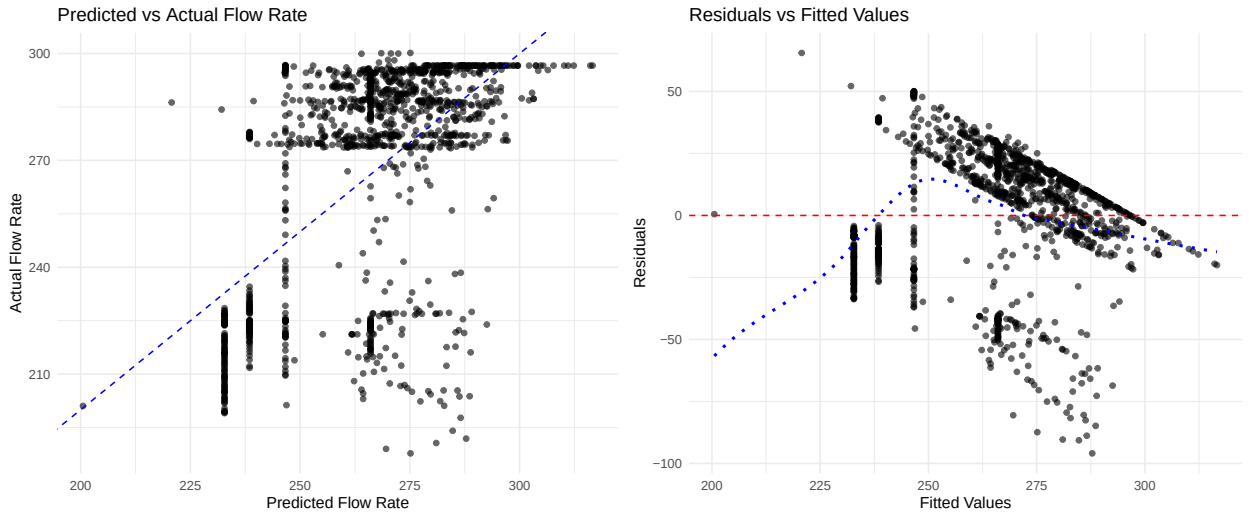

Figure 5: From left to right, scatterplot of Predicted vs Actual Flow Rate and Residuals vs Fitted.

Conclusion

The multiple linear regression analysis examined how rainfall, temperature, and season influenced the flow rate of the Madonna di Canneto spring. The overall model is statistically significant ($p < 2.2e^{-16}$), meaning the predictors collectively contribute to explaining the variability in flow rate. Temperature and seasonality significantly affect the flow rate, while short-term rainfall does not. The flow rate decreases with higher temperatures and varies across seasons, with Summer showing the highest flow rate and Spring the lowest relative to Fall. These findings align with the spring's dependence on seasonal water availability and environmental factors. However, the model is highly limited by the criteria violations below:

1. Linearity - The patterns in residuals in the Residuals vs. Fitted plot indicate large deviations from linearity.
2. Homoscedasticity - The shape of the graph indicates heteroscedasticity, which impacts the reliability of coefficient estimates.
3. Independence - The significant result of the Durbin-Watson Test indicates autocorrelation, as expected due to the time-series nature of the data. Time series data are not independent, the previous day has some impact on subsequent day.
4. Normality - Deviations from the diagonal line indicate departures from normality, and a following Shapiro-Wilk test (see Appendix) confirms non-normality.
5. Multicollinearity - All VIF values are below 2, indicating no significant multicollinearity.

The following section will tackle more advanced model techniques for more reliable results.

3. Advanced Analysis

Methods

XGBoost was chosen for its robustness and ability to iteratively refine predictions by optimizing decision trees. To prepare the dataset for modeling, missing values in `Temperature_Settefrati` were imputed using season-specific distributions, while rows with missing values in `Flow_Rate_Madonna_di_Canneto` were excluded. Lagged features (up to 7 days) were generated for `Rainfall_Settefrati`, `Temperature_Settefrati`, and `Flow_Rate_Madonna_di_Canneto` to account for temporal dependencies inherent in the data. The `Season` variable was one-hot encoded into binary columns (`Season_Spring`, `Season_Summer`, `Season_Fall`, and `Season_Winter`) to ensure compatibility with the model.

The data was split chronologically into training (80%) and testing (20%) sets to preserve the time-series structure. The model was assessed on the test set using **RMSE** to indicate prediction error magnitude, **MAE** to measures the average absolute difference between predictions and actual values, and R^2 to explains the proportion of variance in the response variable captured by the model.

Analysis

```
## [1]  train-rmse:245.539219   test-rmse:206.472497
## Multiple eval metrics are present. Will use test_rmse for early stopping.
## Will train until test_rmse hasn't improved in 10 rounds.
##
## [11] train-rmse:86.236309    test-rmse:72.341332
## [21] train-rmse:30.694634    test-rmse:25.643881
## [31] train-rmse:11.656381    test-rmse:9.180451
## [41] train-rmse:5.410067 test-rmse:4.033202
## [51] train-rmse:3.545140 test-rmse:2.926840
## [61] train-rmse:2.862080 test-rmse:2.680398
## [71] train-rmse:2.373514 test-rmse:2.642092
## [81] train-rmse:2.068286 test-rmse:2.628809
## Stopping. Best iteration:
## [78] train-rmse:2.131876 test-rmse:2.623884

## Number of data points: 1380

## Number of features: 29

## Mean Absolute Error (MAE): 1.632844

## R-squared (R^2): 0.9758832
```

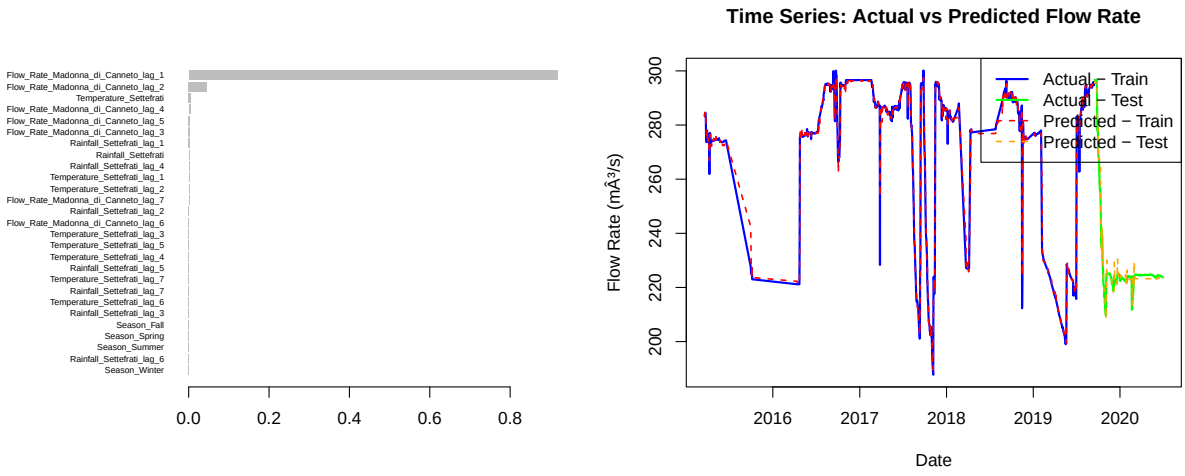

Figure 6: From left to right, Feature Importance Plot and Time Series of Flow Rate.

Conclusion

The XGBoost model applied to the `Water_Spring_Madonna_di_Canneto` dataset demonstrates strong predictive performance for forecasting the `Flow_Rate_Madonna_di_Canneto`. Key performance metrics include a root mean squared error (RMSE) of 2.737 on the test set and an R^2 value of 0.974, indicating that approximately 97.4% of the variance in the flow rate is explained by the model. The Mean Absolute Error (MAE) of 1.61 further highlights the model's accuracy in predicting the flow rate values. These results suggest that the model effectively captures the underlying patterns in the data.

The feature importance analysis reveals that the lagged values of `Flow_Rate_Madonna_di_Canneto` are the most influential predictors, particularly the first few lags. This result aligns with the time series nature of the dataset, where past flow rate values strongly influence future ones. Other features, such as `Rainfall_Settefrati` and `Temperature_Settefrati` (and their respective lags), showed comparatively lower importance, indicating limited direct impact on the flow rate under the conditions represented in this dataset.

The overlay time series plot further confirms the model's reliability, as the predicted values closely follow the actual flow rate values for both the training and test sets. The model's ability to capture sudden drops and peaks, as well as gradual trends, is evident, although slight deviations are observed in extreme or abrupt changes. This is a typical limitation of many machine learning models when applied to complex time series data with potential external influences not captured in the input features.

In conclusion, the XGBoost model effectively predicts the `Flow_Rate_Madonna_di_Canneto`, leveraging its capability to model nonlinear relationships and capture the influence of lagged features. However, the reliance on lagged flow rate values as dominant predictors suggests that external factors, such as operational dynamics or environmental conditions not included in the dataset, may also play a role in influencing the flow rate.

4. Discussion and Conclusion

Conclusion Summary

This report aimed to analyze the flow rate of the Madonna di Canneto spring and its relationship with rainfall and temperature in the Settefrati region. The primary goals were to understand seasonal and yearly variability, identify significant predictors, and assess predictive models for flow rate. Rainfall exhibited a skewed distribution with low average values, while temperature showed a more symmetric distribution. The flow rate displayed a bimodal distribution, suggesting seasonal influences. Weak correlations were observed between flow rate and both rainfall and temperature, implying minimal direct impact. Seasonal analysis confirmed significant differences in rainfall and temperature across seasons, with summer standing out due to distinct characteristics. However, no significant annual trends were identified for either variable, indicating environmental stability during the study period. The regression model highlighted temperature and seasonality as significant predictors of flow rate, with limited influence from short-term rainfall. Violations of assumptions such as autocorrelation and non-linearity were noted, limiting the model's reliability. The XGBoost model effectively predicted flow rate, achieving high accuracy ($R^2 = 0.974$). Lagged flow rate values were the most influential predictors, underscoring the importance of temporal dependencies. These findings align partially with initial expectations, confirming some seasonal patterns but highlighting complex interactions that limit simpler models' effectiveness.

Discussion

The results underscore the influence of seasonal dynamics on the flow rate of the Madonna di Canneto spring. While temperature and seasonal variables significantly impact flow rate, the weak correlations with rainfall suggest the interplay of additional environmental or operational factors. This is consistent with the spring's dependence on broader hydrological and climatic systems.

The XGBoost model's strong performance suggests machine learning methods can capture complex temporal patterns effectively. However, the reliance on lagged flow rate values indicates that unmeasured variables, such as aquifer dynamics or water management practices, may play critical roles.

Some limitations to our dataset include its limited time span (2012–2020) that restricts long-term trend analysis, as well as the large amount of missing data that required imputation, potentially introducing bias. For annual temperature and rainfall, while the yearly averages may not have changed, the analysis does not account for changes in the distribution.

Future work could include incorporating additional variables, such as land use, groundwater extraction, or upstream rainfall to improve model performance. Other avenues include expanding the dataset with more recent and comprehensive data to enhance trend analysis and investigating non-linear relationships and interactions among predictors could uncover hidden dynamics.

5. Appendix

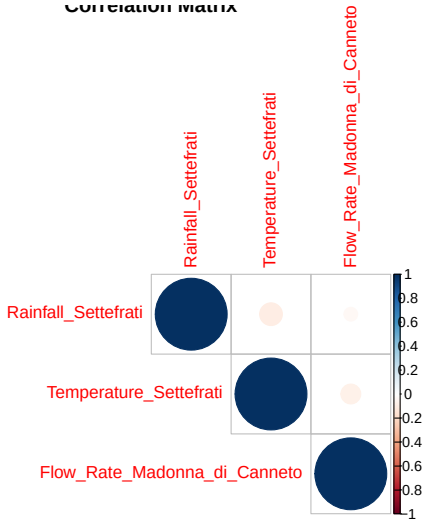

Figure 7: Correlation matrix.

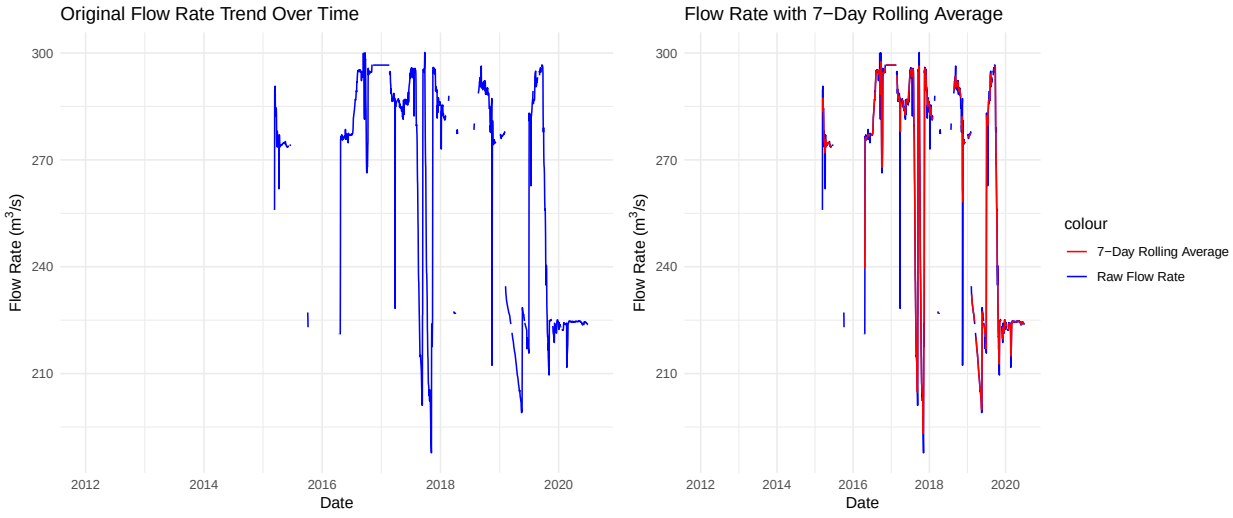

Figure 8: Flow rate trends.

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.96675, p-value < 2.2e-16

##
## Call:
```

```
## lm(formula = Flow_Rate_Madonna_di_Canneto ~ Rainfall_Settefrati +
##     Temperature_Settefrati + Season, data = WS_MdC_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -95.934 -14.212   3.214  18.251  65.524
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             313.15011    2.25946 138.595  < 2e-16 ***
## Rainfall_Settefrati      -0.46241    0.07033  -6.575 6.88e-11 ***
## Temperature_Settefrati   -2.58137    0.11584 -22.284  < 2e-16 ***
## SeasonSpring            -13.85948    1.85447  -7.474 1.38e-13 ***
## SeasonSummer             19.41656    2.12072   9.156  < 2e-16 ***
## SeasonWinter             -8.15172    2.01825  -4.039 5.66e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.47 on 1381 degrees of freedom
## Multiple R-squared:  0.3718, Adjusted R-squared:  0.3695
## F-statistic: 163.5 on 5 and 1381 DF,  p-value: < 2.2e-16
```