Kaden Rookus

CIS 635 Final Project - Water Flow Analysis

Introduction

For my Final Project, I wanted to utilize the provided dataset option that looks at Streamflow over time. I wanted to bring this data through the full data-mining pipeline to try and find patterns in the data to explore potential predictors for droughts and floods within streamflow to be able to understand how one could mitigate the impact. My reason for picking this dataset and project was for three reasons; firstly, I think that using the provided dataset is a good thing because of the extra credit provided. Secondly, having a chosen dataset reduces the risk of a poorly chosen dataset that is hard to work with. Thirdly, because I am working alone, I can put more effort into the data mining process itself with a dataset already chosen.

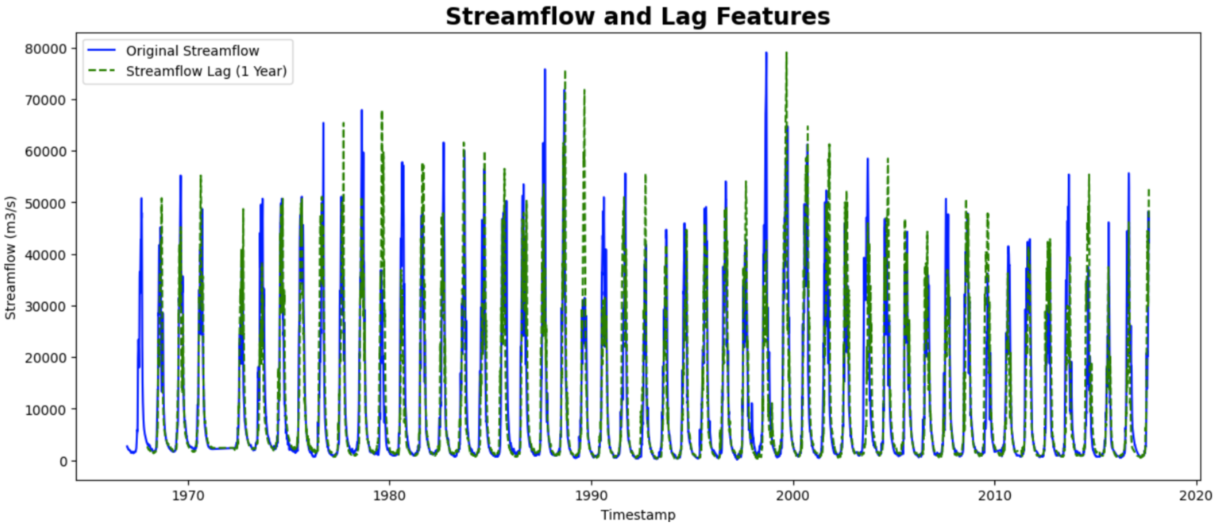
Overall I wasn't able to complete as much as I would have liked given the time restraints. However, I approached the project by investigating ARIMA and STL as potential ways to predict future time-series data. Because ARIMA's results were so poor, I did end up adding SARIMA as well, to try and capture the seasonality, this produced mediocre results, likely because of parameter constraints. Overall, I feel like my results were inconclusive, with the training data producing predictions that weren't able to fully capture the seasonality. I think more time and work could go into refining the models to potentially capture the data, since SARIMA seemed to have some promise.

Related Work

Since the provided dataset lacks a source website, I was not able to find a project on this exact dataset. However, there are similar studies done on droughts and floods with standing water reservoirs, such as “Challenges in modeling and predicting floods and droughts: A review” by Manuela I. Brunner. My project would be similar, but on stream flow, which is running water rather than standing water. The goal would be to reach a conclusive model that can help predict floods and droughts, similar to what they produced. Another study, “Using Satellite Data to Predict Floods and Droughts” does focus on finding trends in rivers, however, it focuses on using visual satellite data to reach its conclusions, rather than direct streamflow data. Lastly, I looked at “Evaluation of Statistical Methods for Estimating Missing Daily Streamflow Data” By Mustafa Utku YILMAZ. This paper looked at estimating streamflow observations through regression training as well as drainage area ratios and a lot of statistics. This actually seemed pretty on track with what I was attempting to do, just with a different approach and more personalized data production.

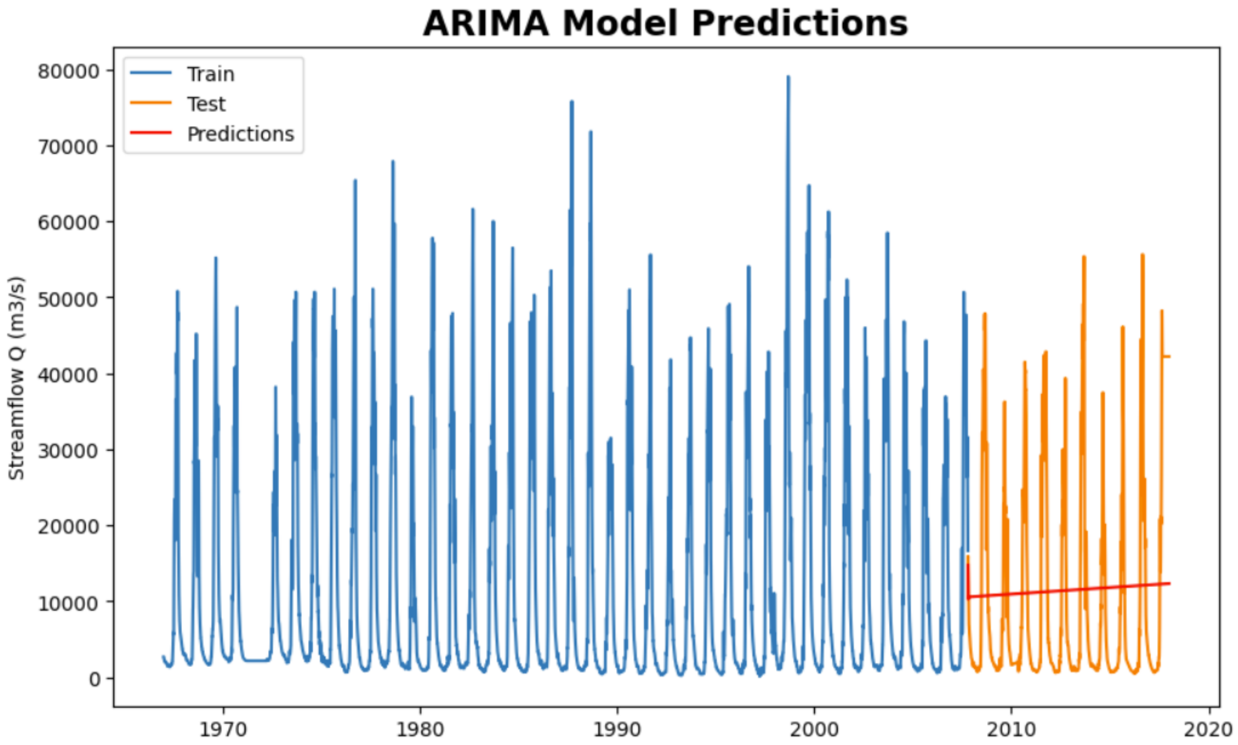
Methods

For this project, my methodology was fairly simple. My data selection was the streamflow dataset provided by Professor Zhuang. I loaded the dataset on Google Colab using Pandas and Numpy and performed basic analysis by looking at the head, info, and description. I then dropped all rows with missing values before visualizing the data through various charts; for example, I graphed the chart normally, then played around with lag features to overlap the data with itself 1 year apart to begin looking at yearly/seasonal patterns. One thing to note is I later learned some of the data was filled as NaN rather than actually missing, I resolved this by using forward-fill to fill in the missing data with the previous data. Next, I began aggregating the data in time intervals to try and notice general trends within the averaged data. Finally, for Model evaluation analyses, I used sklearn with ARIMA and Seasonal-Trend decomposition to try and train models to predict the data. Eventually, after the failure of ARIMA and our time-series assignment, I wanted to investigate how SARIMA would do on the model as well.

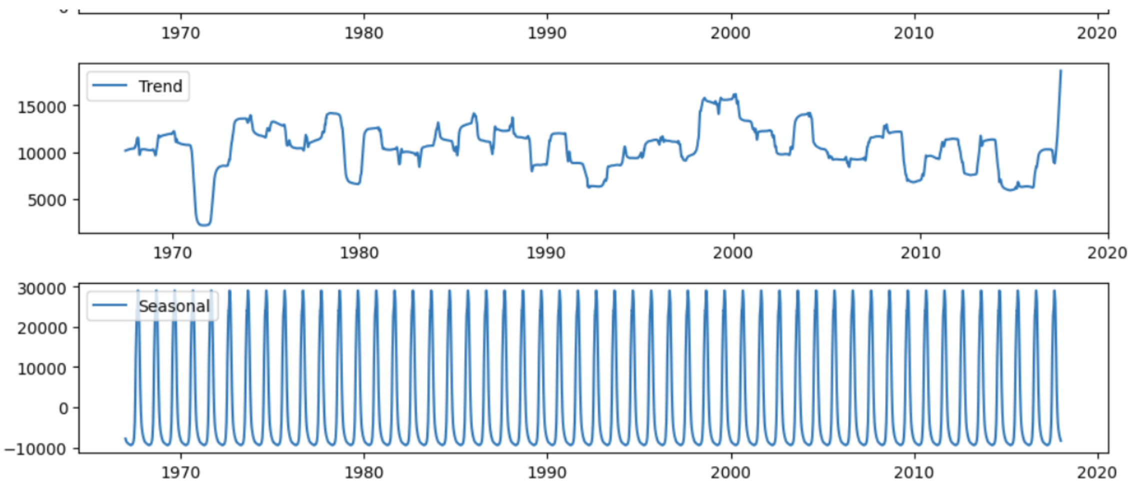


Results and Discussion

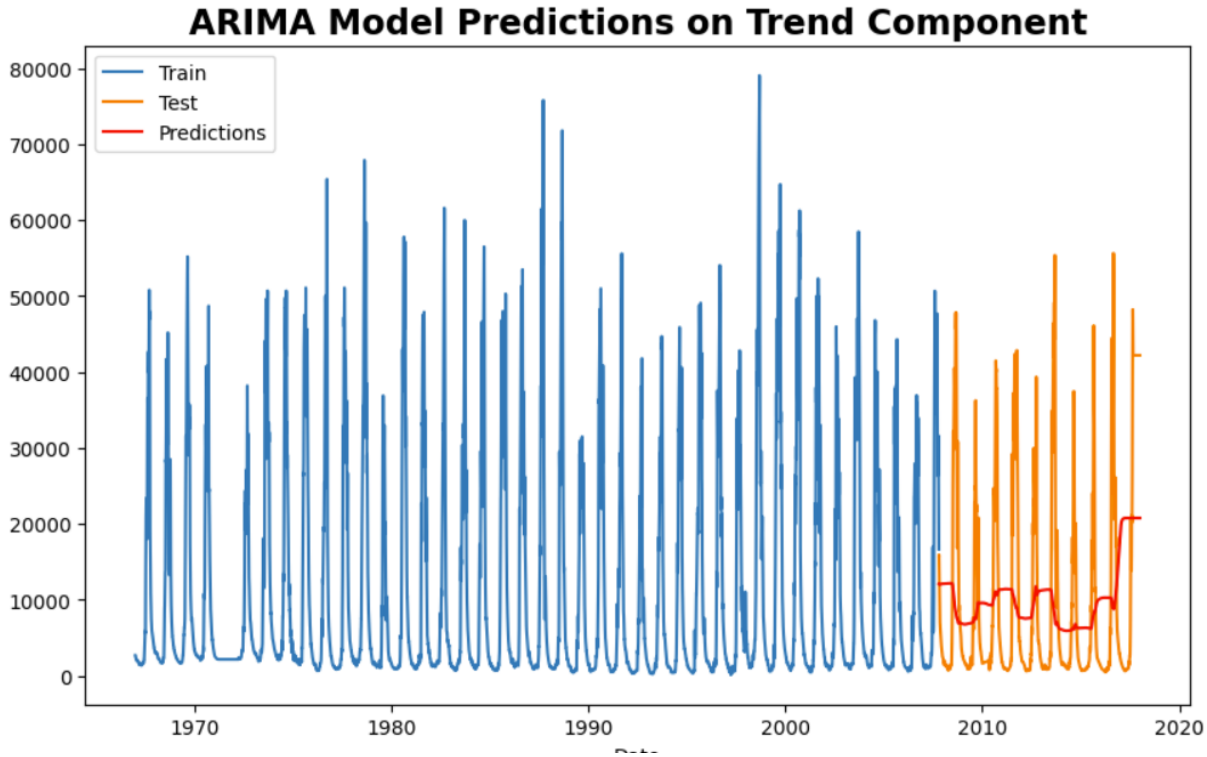
In my model creation and evaluation, initially, I tested ARIMA Model Predictions and Season-Trend decomposition with ARIMA, before later looking at SARIMA. The ARIMA model performance was extremely poor, with its prediction mostly flatlining. However, the ARIMA model has order variables that can be changed, and playing around with these I was able to get the model to slightly change from direct flatlines to showing minor trends in the general data, as shown below.



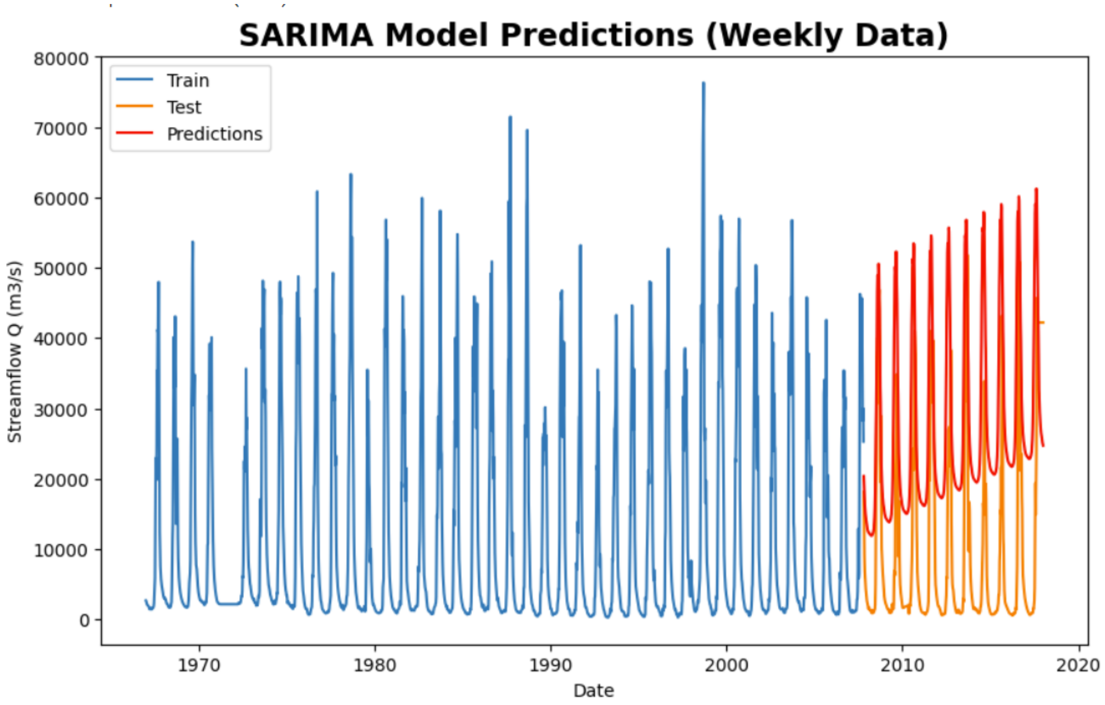
The Seasonal-Trend decomposition model was far more interesting. Just like before I loaded up the data and performed minor preprocessing(creating a date column, forward filling missing data). Next, I performed Seasonal Decomposition on the data, with an initial period value of 12, since that was the number of months. The data here was strange, with straight lines and poor trend graphs. I then changed my code to loop through with different period values, before noticing that the model seemed to function best with a period of 365, daily. This produced a graph able to capture data trends and seasonality of streamflow, shown below.



Even after capturing these trends, STL still failed to really capture the model and produce accurate predictions. In the following graph we see that it somewhat understood the ups and downs, but not the frequency or extremities actually required by the testing dataset. Since these predictions still used ARIMA, I did try and alter the order variables, however, the model remained roughly the same.



Because of ARIMA’s lackluster results, I did end up adding SARIMA as well, since it’s better known for capturing data seasonality. The results of this were far better at predicting the data, as seen below.



Due to Colab limitations, I was only able to run this on weekly averages rather than daily reports. This means that this model could definitely be better refined and potentially become even more accurate.

Overall, SARIMA was clearly the best model so far, however, I think a lot more time could go into testing different models and variables to try and fully capture the data.

Conclusion

Overall I got to experience the data mining process. I explored a dataset, executed preprocessing, and examined it using 3 different data mining algorithms. I didn’t find a conclusive way to predict droughts and floods, but I found general ways to predict the streamflow. I think this project had several significant limitations/shortcomings. Firstly, I feel that we learned and practiced the data mining process fairly late into the semester, limiting my time to actually write code and examine the data. Secondly, I think an expanded upon dataset would be necessary that includes different causations for floods/droughts, to help better predict data trends. Thirdly, I was partially limited by Google Colab's ability to handle large datasets and run complex algorithms on them. In the future, I think it would be interesting to continue to pursue this project by spending time to eliminate the above limitations.

Data and Software Availability

This dataset was obtained from the streamflow time series dataset provided by Professor Yong Zhuang, it can be found at:
https://yong-zhuang.github.io/gvsu-cis635/project-overview.html
The code used for analysis can be found on my Github here:
https://github.com/GVSU-CIS635/gvsu-cis635-term-project-flood-droughtpredictions

References:

Brunner, Manuela I., et al. “Challenges in Modeling and Predicting Floods and Droughts: A Review.” *WIREs Water*, vol. 8, no. 3, 2021, doi:10.1002/wat2.1520.

“Using Satellite Data to Predict Floods and Droughts.” *Earth Online*, The European Space Agency, 22 Dec. 2020, earth.esa.int/eogateway/news/using-satellite-data-to-predict-floods-and-droughts. Accessed 13 Dec. 2023.

YILMAZ, Mustafa Utku and Bihrat ÖNÖZ. “Evaluation of Statistical Methods for Estimating Missing Daily Streamflow Data”. Teknik Dergi, vol. 30, no. 6, 2019, pp. 9597-20, doi:10.18400/tekderg.421091.