Time Series Analysis of Monthly Highest Temperatures and Average Humidity in South Georgia (2018-2024)

Leslie Dawn Hunt
Topics in Data Science
Valdosta State University
July 31, 2024

Executive Summary

This report explores the time series data of monthly highest temperatures and average humidity percentages in South Georgia from July 2018 to June 2024. Seasonal Decomposition using STL and ARIMA forecasting are used to analyze the data in Python. The analysis provides insights into seasonal patterns and future temperature/humidity predictions. The Seasonal Decomposition identifies significant seasonal and trend components in the data, while the ARIMA model forecasts future values with notable accuracy. This analysis offers valuable insights for climate trend understanding and future forecasting.

Methodology

Data Sources and Collection

The dataset used for this analysis was sourced from the website Weather Underground, wunderground.com, and saved in a CSV file named dataset_.csv. The specific location the data was reported from is the Southwest Georgia Regional Airport Station in Albany, Georgia. The CSV file contains three key columns: Date, Avg Humidity, and Highest Temperature. The Date column spans from July 2018 to June 2024, while the Avg Humidity and Highest Temperature columns provide monthly records of average humidity and highest temperature, respectively. More specifically, the hottest temperatures out of each month were collected, and its corresponding average humidity for that day were recorded.

Data Cleaning and Preparation

Data was loaded using pandas, with date parsing to convert the Date column to a datetime format suitable for monthly intervals. Deviations from the mean of the highest temperature and average humidity were calculated for use in Seasonal Decomposition.

Algorithms and Models

For this analysis, all code and methods were done in Python. Seasonal Decomposition using STL (Seasonal-Trend decomposition using Loess) was used to decompose the deviation of temperatures into seasonal, trend, and residual components to analyze underlying patterns with a period of 12. ARIMA Forecasting was used with specified parameters (p, d, q), specifically p=12, d=1, and q=0, to model and forecast future values of highest temperatures and average humidity. Forecasts were generated for the next 30 months with confidence intervals.

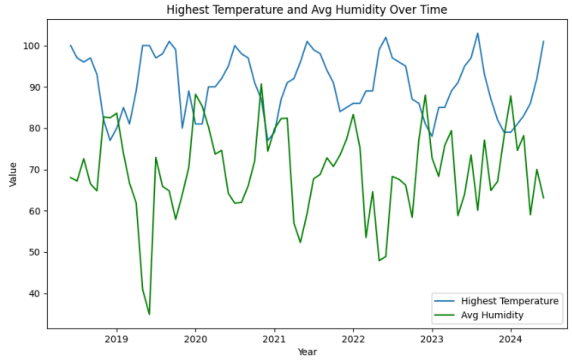
Changes in Data Collection

In real-world data science projects, research questions and analysis methods often need adjustments based on the data's availability and quality. For this project, the initial plan was to analyze the time series of the highest temperature and average precipitation in south Georgia. As data collection started, an observation was made that on the hottest days out of each month, precipitation was reported as zero for the majority of each corresponding temperature. This observation is consistent with the logic that less precipitation makes hotter temperatures.

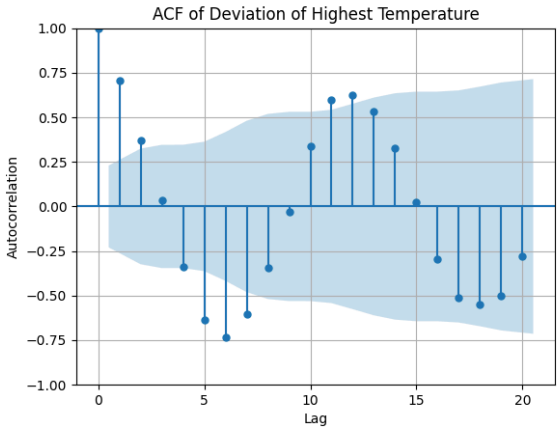
Data Exploration and Analysis

To begin the analysis, a graph was created to visualize the relationship between the highest temperature and the recorded average humidity. The graph indicates an inverse relationship, consistent with the understanding that humidity percentages tend to be lower on days with the highest recorded temperatures.

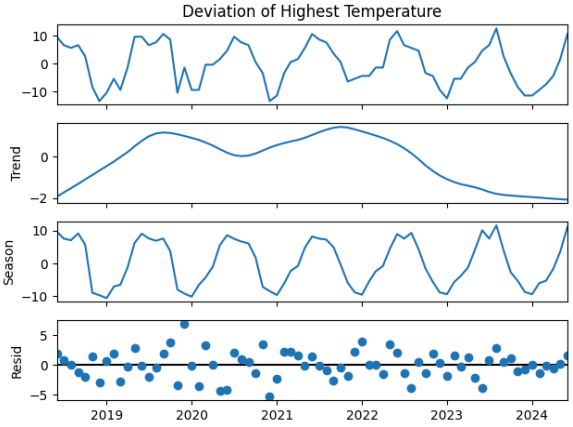
Time Series Analysis of Monthly Highest Temperatures and Average Humidity in South Georgia (2018-2024)



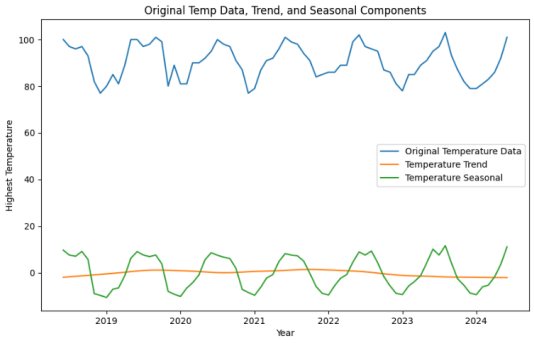
For the Seasonal Decomposition using STL analysis, the deviation of the highest temperature from its mean was calculated first. An ACF (Autocorrelation Function) plot was then generated to identify the period of the seasonal component, which was determined by observing the peak at lag 12.



Now, using a period of 12, the decomposition separates the time series into three components: trend, seasonal, and residual. The result.plot() function generates a plot to visualize these components, showing the underlying patterns in the data.



To further illustrate the findings, a combined plot of the original data (deviation of highest temperature), trend, and seasonal components is created. This plot clearly highlights the cyclical nature of the temperature data with its seasonal peaks, as well as the long-term trend indicating gradual changes over time.

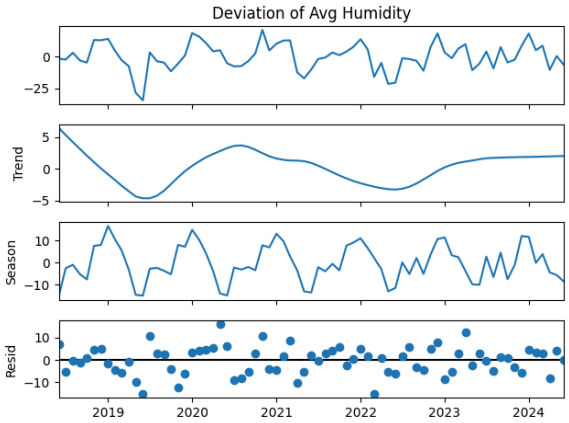


Given the plot, we can see there are trends in the data every 12 month period, and the highest temperature is seemingly increasing over time for both the summer and winter seasons. According to the trend line, there seems to be a steady and slowly increasing pattern for the temperatures. Forecasting and more data collection from previous years would help increase accuracy and visualize slow gradual increases of highest recorded temperatures in the coming years.

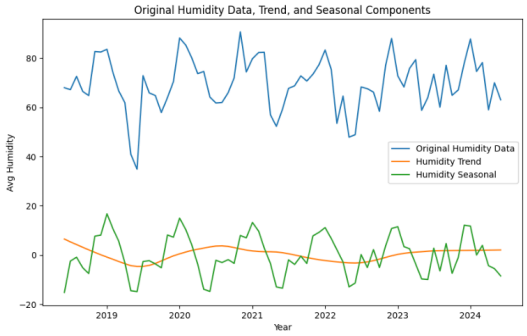
Now that Seasonal Decomposition using STL for highest temperature has been done, humidity will be

Time Series Analysis of Monthly Highest Temperatures and Average Humidity in South Georgia (2018-2024)

analyzed in the same way for comparison, using a period of 12. Once again, seasonal decomposition separates the time series into three components: trend, seasonal, and residual. The result.plot() function generates a plot to visualize these components, showing the underlying patterns in the data.



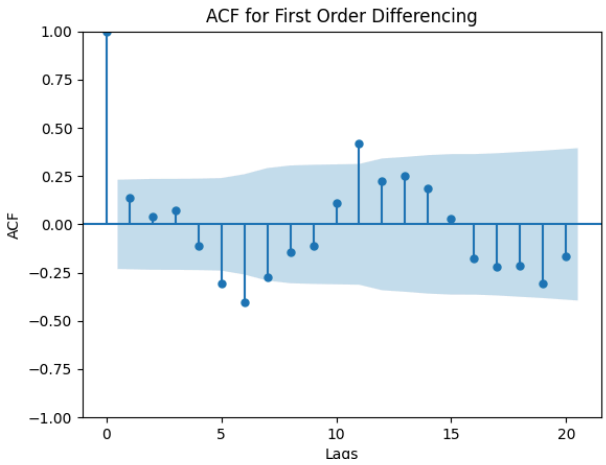
A combined plot of the original data (deviation of average humidity), trend, and seasonal components is created. This plot clearly highlights the cyclical nature of the temperature data with its seasonal peaks, as well as the long-term trend indicating gradual changes over time.



When comparing these to the plots found by temperature, it can be observed that they are inverse, where months with the hottest temperatures have the lowest recorded humidity. It can be assumed that humidity averages peak in the winter and drop lower in the summer. For temperature, it seems to have a seasonal component line that has slowly increasing peaks in the summer every year, and has slowly increasing temperatures in the winter each year. For humidity's seasonal component line, its average percentages are also increasing with every season.

For the ARIMA analysis, testing different parameter values (p, d, q) for optimal forecasting is required in order to find the best fitting model. In order to observe highest temperature vs average humidity, the code will use ARIMAX to add the humidity variable. To evaluate for d, testing if the original time series was stationary was done. In this context, stationary refers to a property where the statistical properties of a time series, such as mean, variance, and autocorrelation, are constant over time. We use the ADF Test, consisting of p-value (measure of stationary, $p < 0.05$; measure of nonstationary, $p>=0.05$) and the ADF statistic, in order to prove if a time series is stationary or not. When evaluating the data, it was determined through the stationary tests that the original time series was non-stationary, giving results of the ADF statistic being -1.428 and the p-value being 0.569. The p-value must be less than 0.05 to be stationary. To continue, first order differencing is calculated, and was proven to be stationary with the ADF statistic being -8.326 and the p-value being 3.47e-13. In this case, p-value is less than 0.05, giving us a stationary series. First order differencing being stationary tells us that for our ARIMA analysis, we use d = 1. Now that d is found, we can use this value to determine p and q. The parameter p, being the period or seasonal component, can be found by visualizing the ACF (AutoCorrelation Function) plot where lags seem to peak in certain patterns. According to the ACF plot, there is a lag at 12.

Time Series Analysis of Monthly Highest Temperatures and Average Humidity in South Georgia (2018-2024)

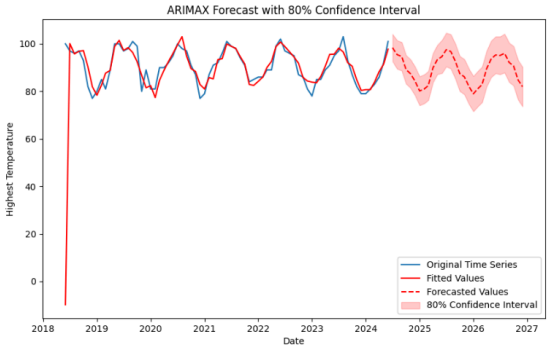


Given this, we can say that p = 12. For the parameter q, we must look for the optimal value given from the AIC (Akaike Information Criterion) scores based on what we have found for p and d.

AIC Scores:

- ARIMAX(p=12, d=1, q=0) - AIC: 395.716
- ARIMAX(p=12, d=1, q=1) - AIC: 397.639
- ARIMAX(p=12, d=1, q=2) - AIC: 401.848
- ARIMAX(p=12, d=1, q=3) - AIC: 400.533
- ARIMAX(p=12, d=1, q=4) - AIC: 402.837

The optimal q value based on the lowest AIC is 0, resulting in the best model: ARIMAX(p=12, d=1, q=0) - AIC: 395.716. Now that we have all parameters accurately determined, we can finalize our ARIMA analysis and forecast the highest temperature for the next 30 months with confidence intervals. This model seems to fit out data accurately with an 80 percent confidence interval.



Results and Discussion

Seasonal Decomposition

The STL decomposition revealed a clear seasonal pattern with annual fluctuations in the highest monthly temperatures. For each component, it can be seen that both the highest temperatures and average humidity percentages are slowly increasing each year. The seasonal decomposition indicated that the highest temperatures are gradually increasing over time, with distinct seasonal peaks during the summer months,while the average humidity levels tend to decrease during these hotter periods. This inverse relationship is forecasted to continue, with future temperature increases correlating with lower humidity levels during the hottest months.

ARIMA Forecasting

The ARIMA model provided a reliable forecast of future temperatures. The forecasts showed a continuation of the observed trends, with confidence intervals indicating a reasonable degree of certainty.

Results

These findings suggest that as temperatures rise, the average humidity will likely increase alongside it. This trend has important implications for understanding and predicting future climate conditions in South Georgia. Also, it can be observed that the correlation or relationship between these two

Time Series Analysis of Monthly Highest Temperatures and Average Humidity in South Georgia (2018-2024)

are inverse. As temperature increases, the relative humidity tends to decrease, and vice versa.

Strengths and Weaknesses

A strength of Seasonal Decomposition using STL and ARIMA Forecasting methods would be that they provided meaningful and accurate insights into temperature patterns and future forecasts.

A weakness for the ARIMA model's performance could be non-stationarity values and parameter tuning. In addition, the lack of data can cause issues in seeing clear results from the models. Future work may include collecting a larger dataset in order to see more accurate observations.

Potential Biases

Potential biases in the data could be measurement errors, external factors impacting temperature that were not included (precipitation, wind patterns, etc.), and specific variations (seasonal or daily). In the data collection method, each month was extracted manually and put into a CSV file. The date for each highest recorded temperature was set to the first of each month. Realistically, the date for the highest temperature in each month varied throughout the span of a month. This could cause bias or variation in the data. Recording the specific date of the month for the relative temperature and humidity would prove this as true or false.

Areas for Further Exploration

Future work for this project may involve incorporating additional variables to enhance model accuracy and exploring more complex models for better forecasting. In addition, recording specific recorded dates would help see different aspects of the data trends. To expand this project, further data collection with previous years would help give a more accurate visual of the trends between highest temperature and average humidity amongst the different seasons.

Conclusion

This project successfully analyzed and forecasted monthly highest temperatures in South Georgia using time series methods. STL decomposition provided valuable insights into seasonal and trend components. The ARIMA analysis offered accurate forecasts consistent within its confidence interval. These findings are useful for understanding highest temperature trends versus average humidity and making informed predictions about future climate conditions.

The analysis of time series data for monthly highest temperatures and average humidity in South Georgia revealed a significant inverse relationship between these variables. Using STL decomposition and ARIMA forecasting methods, it was observed that months with higher temperatures typically have lower average humidity levels, a trend consistent throughout the observed period. Over the years, the data indicates a gradual increase in both highest temperatures and average humidity percentages across different seasons each year.

References

[1] Ault, Shaun. "Chapter 05 - Time Series and Forecasting."

[2] "What Is an ARIMAX Model?" *365 Data Science*, 21 Apr. 2023, 365datascience.com/tutorials/python-tutorials/arimax/.

[3] "Albany, GA Weather History." *Weather Underground*, www.wunderground.com/history/daily/us/ga/albany/KABY/date.