Overview

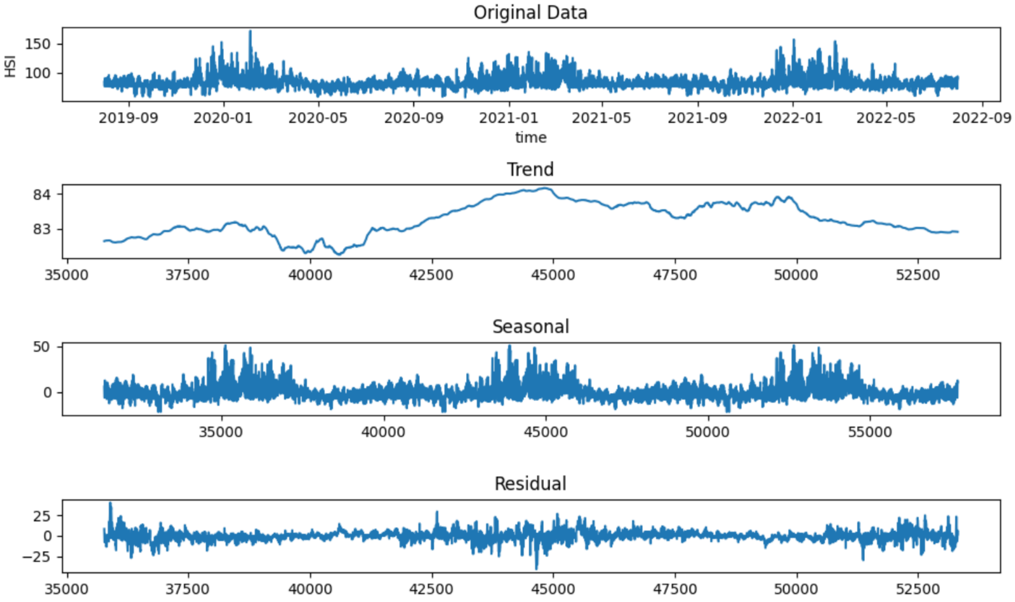
This report outlines 3 forecasting and decision-making approaches. Each includes a brief introduction, pros, cons, and potential next steps. I focused on HSI as the variable of interest. All exploration was conducted in colab.

Approach 1: SARIMA Forecast with a Threshold-Based Decision Model

Basic Idea

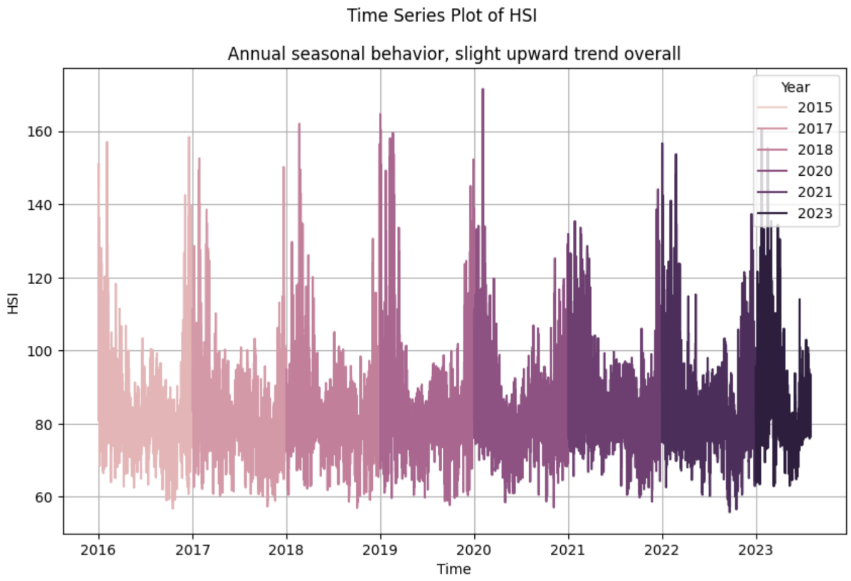
SARIMA stands for “Seasonal Autoregressive Integrated Moving Average” and is a modeling technique to capture complex time series patterns. It aims to remove the seasonal and trend components of time series and then model the remaining stationary series. Below is a visual example of these three components for the Ramona NASA data.

The idea here is that if we’re using a forecasting method that is highly accurate, we can simply flag points at which the 95% confidence interval crosses over a safe threshold. Our decision-making model would flag these as points of increased AC demand.

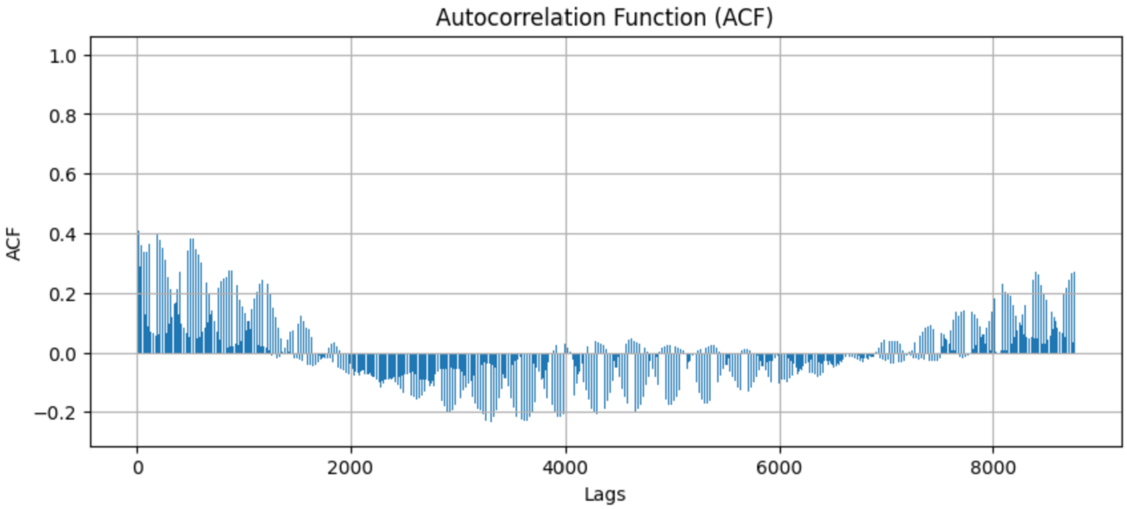


Pros

SARIMA is highly researched and, to my knowledge, is used in industry. As can be seen in the images below, there is annual seasonality. Theoretically, SARIMA would be able to capture this repetitive pattern that occurs at fixed intervals.

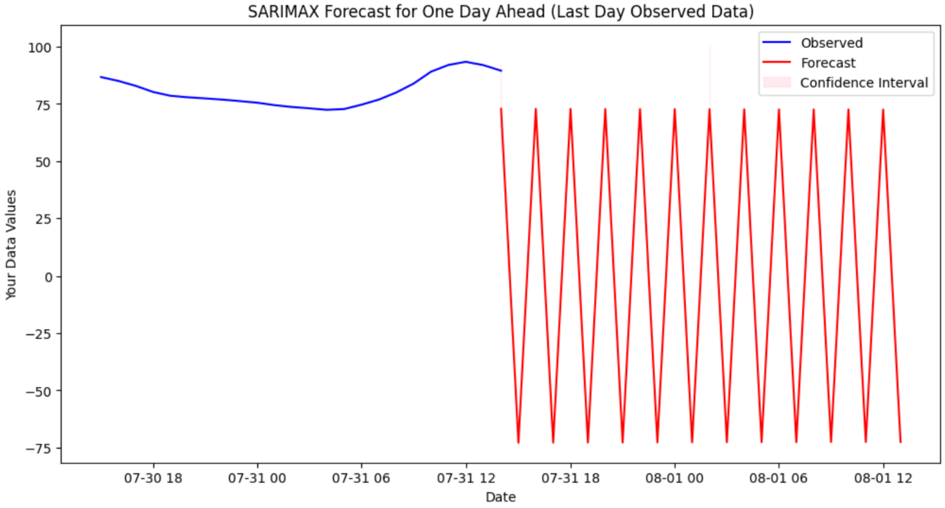


The oscillations in the ACF below are further proof of seasonality.



Cons

Fitting SARIMA models is computationally expensive, especially with such granular data and without a hyperparameter starting point. I was unsuccessful in producing a well-fit model in colab. I tried both brute force and auto_arima from pmdarima.arima. Many times I timed out or ran out of RAM. The image below shows a day of observed data in blue and the model's prediction one day ahead in red. The prediction doesn't look anything like the pattern of the actual data.



Potential Next Steps

- There is annual seasonality, but we could explore other intervals to see if there are patterns. For example, daily seasonality could exist that would help in hourly predictions.
- Conduct additional research to see if we can narrow the hyperparameters for SARIMA pdq and PDQ values. This would narrow the scope of models we'd have to test to find the best fitting instead of using brute force to try all possible models.
- Explore additional SARIMA packages that may have more efficient ways of selecting the best fit model. To my knowledge, R has good tools for SARIMA forecasting.
- Try different platforms with more compute power.
- Consider how the compute power needed would factor into a client-facing tool and whether a model would need to be refitted every time someone made a query.
- Test whether reducing data dimensionality (i.e. using daily highs instead of hourly data) helps with computation.

Additional Notes

- The "integrated" aspect of SARIMA modeling uses differencing. I believe this means you need at least 3 units of whatever level you're trying to model. For example, if you're modeling using annual seasonality, you would need at least 3 years of data.
- While it is a valid option, I did not take the log of HSI because its variance seems steady over time.

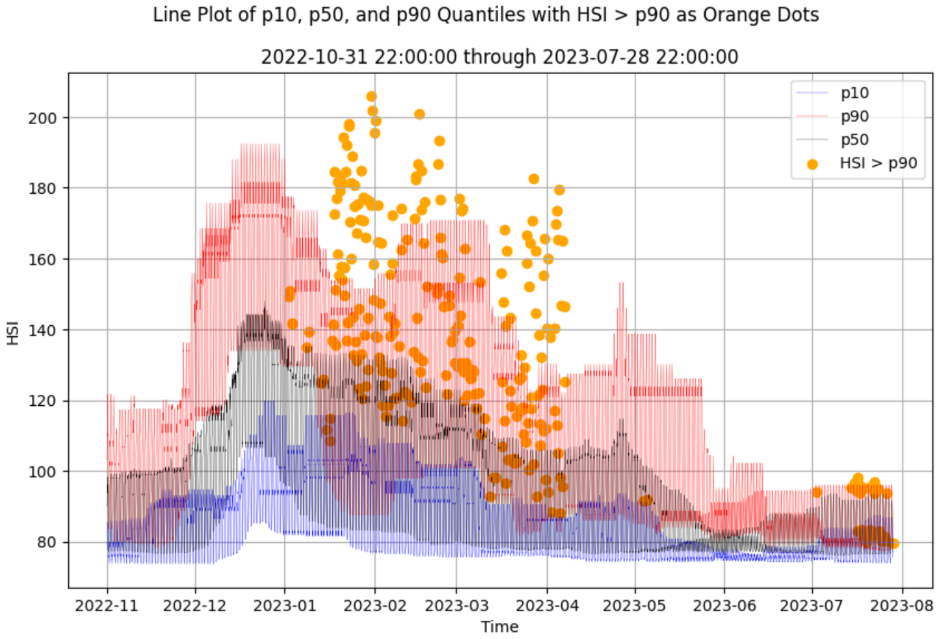
Approach 2: Smoothed Quantile Forecast with an Anomaly Detection-Based Decision Model using NOAA

Basic Idea

This is based on Divyam's method, which is to forecast by looking back 1 year and taking a 30-day moving average. This smoothing method reduces the effect of outliers and gives us an idea of normal behavior. A "normal" forecast is less likely to cross a dangerous threshold, given the assumption that danger is "not normal". We then need a "source of truth" to tell us when there's a data point outside of the normal range. In this case, we can flag points when the NOAA forecast is greater than our normal 95th quantile.

This is a form of anomaly detection for point estimates in univariate time series, in which you raise the alarm if your prediction is far off from the true value. In this case, our prediction is the 95th quantile and the NOAA forecast is our truth. Similar to what was described in the previous method, we could also flag points where the normal forecast crosses into a dangerous level.

The visual below shows the smoothed forecast for Mariposa. The 90th quantile is shown in red and the orange dots represent instances where the NOAA prediction is considered an anomaly.



Pros

Again, the moving average technique acts as a smoother to give us an idea of what's normal. This technique is also computationally efficient and can be easily modified for how far out you want to forecast.

Cons

We have to rely on external NOAA predictions as our source of truth. Also, the smoothing method may not pick up on more complex patterns, like the SARIMA theoretically should.

Potential Next Steps

- Look into different ways to incorporate the quantiles (i.e. uncertainty envelope) into a decision-making process.
- Try smoothing the forecast even more by incorporating 2 years of data and compare that to the 1-year smoother.

Approach 3: Sinusoidal and Climatological Decomposition Forecast with an Anomaly Detection-Based Decision Model using Probability Distributions

Basic Idea

Similar to the SARIMA approach, the idea here is to decompose the time series and then look at what's left. However, instead of fitting a SARIMA model, you use a sine wave to mimic the seasonal cycle and subtract the climatological mean as well.

Decomposed Series = Original Series - Seasonal Cycle - Climatological Mean

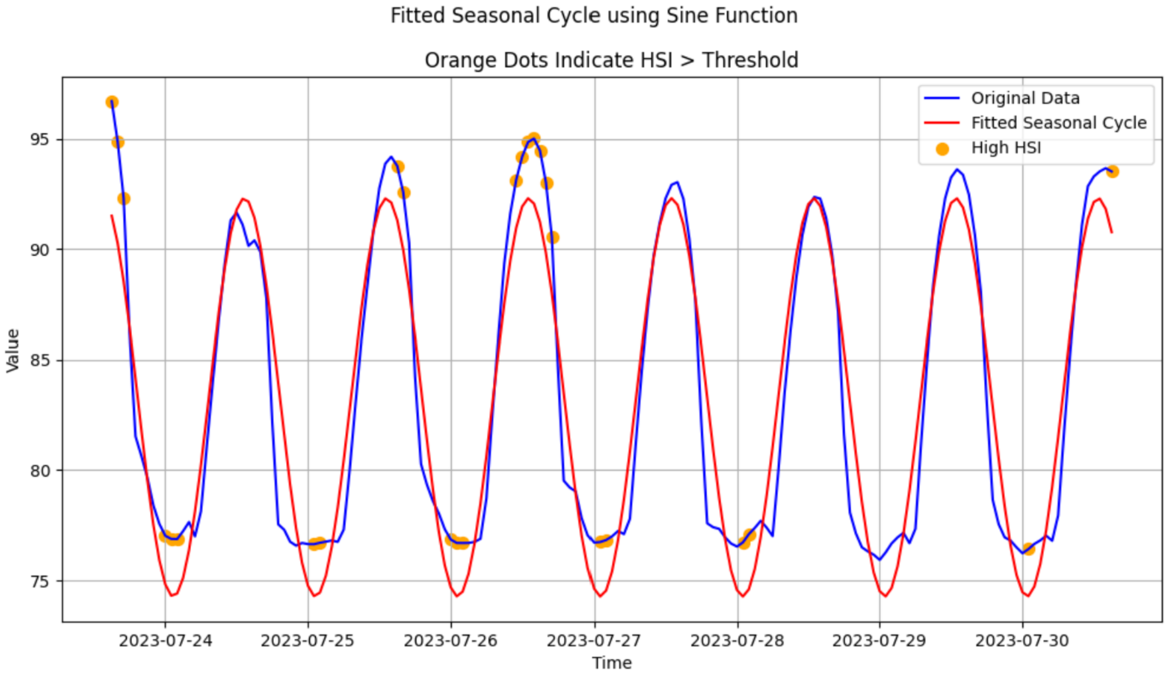
In a sense, this method also gives us a sense of normality so we can detect anomalies that fall outside of a decided threshold. Once the decomposed series is calculated, you determine its 95th percentile, and then flag points that are further away than expected behavior. In this case, the sine wave represents what's normal.

Pros

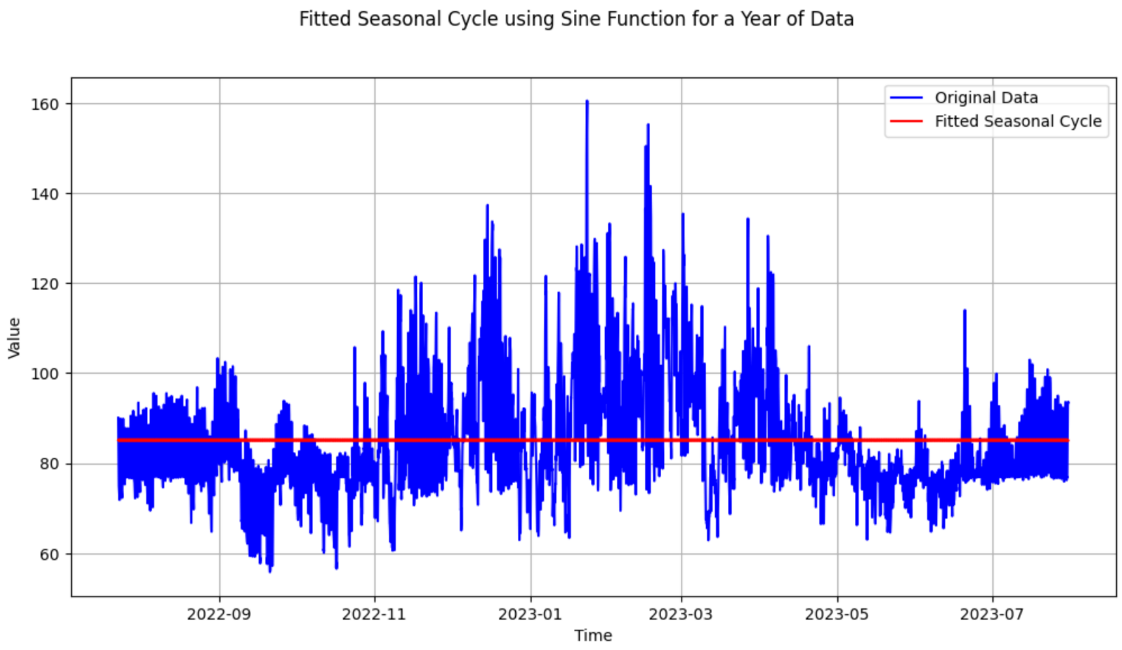
Again, this approach is more computationally feasible than SARIMA and is a straightforward way to detect anomalies.

Cons

A data point that's far from what's expected isn't necessarily bad. As you can see in the visual below, the flags capture HSI low-points as well as the highs. That said, this is an easy fix if we set a "danger threshold" that only returns the orange dots over a certain HSI value.



A sine wave may not be the most reliable forecasting method. I got a wave that seems to closely mimic the actual data's seasonality after some hyperparameter tuning, but it was not a very scientific process. The result of the initial curve_fit() without tuning did not produce great results. Also, the image shown above is for a week-long period. This approach may work for short-term predictions where the seasonal pattern is relatively simple, but I could not get a sine wave to even remotely fit longer periods of hourly data, as can be seen below. Note I did not try to tune this sine wave.



Potential Next Steps

- See if we can fit a sine wave to longer periods of data.
- For longer periods of data, try reducing data dimensionality to see if it more closely resembles a replicable pattern. For example, use daily highs instead of hourly data.
- Explore if functions other than sine are a better fit.
- Reduce data to daily time series and try fitting a sine over a few months.
- Identify high subsequent HSI values (i.e. heat waves).

Additional Notes

- From Detelina: “As a threshold, I would use 95% when the time series is longer e.g. 5y, if you have 2 years period use lower, maybe 80-90%”
- The climatological mean was calculated using the available Ramona data. Need to confirm whether 8 years of data is sufficient for this metric.

Conclusion

It may be worth exploring combinations of these methods. For example, flagging when the smoothed forecast is over a certain threshold and flagging abnormal NOAA data points. Ideally, we could build all forecasting methods, compare their flags, and dig into the differences.

Please feel free to send me any feedback!