Timing Ozone

Predicting Daily Mean Ozone Levels with ARIMA

Ryker Dolese

Executive Summary

This report examines daily mean ozone levels in Houston, focusing on data collected from Moody Tower on the University of Houston campus. Ozone, a secondary pollutant, forms from precursor primary contaminants in the presence of sunlight and is a significant health concern for residents, particularly when levels exceed 80 parts per billion (ppb).

The analysis employs time series modeling techniques to forecast ozone levels. Initially, an ARIMA(3, 0, 0) model is fitted to the data, considering the autocorrelation and partial autocorrelation functions (ACF and PACF). We compare this with an AR(1) model to suggest potential improvements in modeling accuracy.

Despite limitations, the forecasting power of the models is adequate, particularly for short-term predictions. However, the forecasts tend to converge towards the mean over longer periods, indicative of an ARMA process. Some months are more accurately predicted than others, suggesting other variables, such as month, could improve our baseline. Including a seasonal component could ameliorate this issue as well.

The findings of this analysis provide valuable insights into the dynamics of ozone levels in Houston, indicating the importance of continued monitoring and potential interventions to mitigate ozone pollution in the city. Further research could explore additional factors influencing ozone concentration and refine forecasting models for improved accuracy in long-term predictions.

Exploratory Analysis

Time Series Plot

Something to note prior to the analysis is the data cleaning that was performed. There were approximately 30 data points that were missing. We used a linear interpolation technique to fill in these missing values.

Figure 1 showcases the time series we're analyzing, depicting daily mean ozone levels typically ranging between 10 and 50 PPB. Most days cluster around a daily average of 30 PPB. While no pronounced seasonality is evident, a noticeable cyclical pattern emerges. Based on these preliminary observations, an Autoregressive model appears well-suited for further investigation.

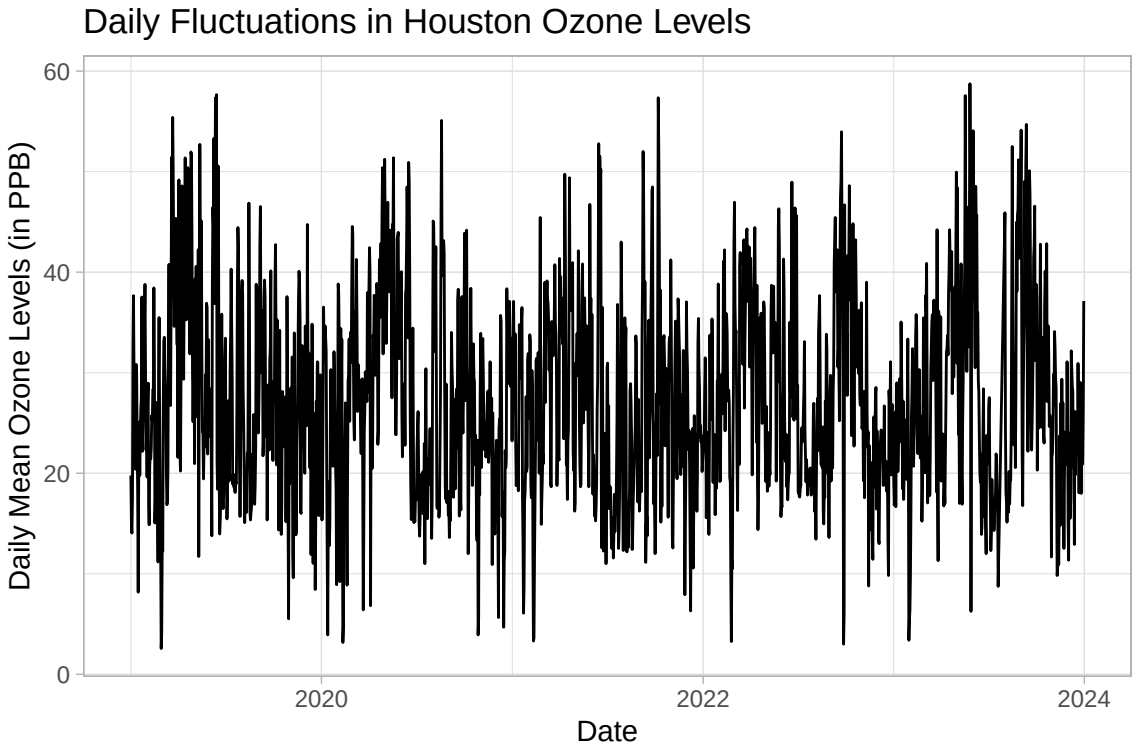

Figure 1: Basic Time Series Plot for Ozone Daily Means

Seasonal Plot: Any Trends Across Years?

Our next objective is to identify trends and patterns in our time series. In Figure 2, we have produced a seasonal plot capturing daily ozone levels across the 5 years in our data. While there is no clear pattern or differences from year to year, there does appear to be some seasonality within our data. Most notably, there appears to be a strong dip in ozone levels in and around the month of July. This could offer valuable insight into our modeling.

Subseries Plotting: Are There any Differences Across Months?

The trend we discovered in Figure 2 presents itself in Figure 3. Here, I have taken the mean ozone level for each month (this is to limit the convoluted nature of the daily data) over the 5 year period. We can see that ozone levels tend to peak around April and May and dip to an average of 20 ppb in July. Once again, there doesn't appear to be any pervasive increase or decrease across years. Rather, most variation occurs when comparing months.

Distribution of Ozone Levels

Figure 4 generalizes what we have discussed before: daily mean ozone levels tend to fall between 20 and 40 ppb. The distribution has a slight rightward skew, indicating very high levels of ozone do occur every so often. The box-plot reaffirms that ozone levels have not deviated substantially across years.

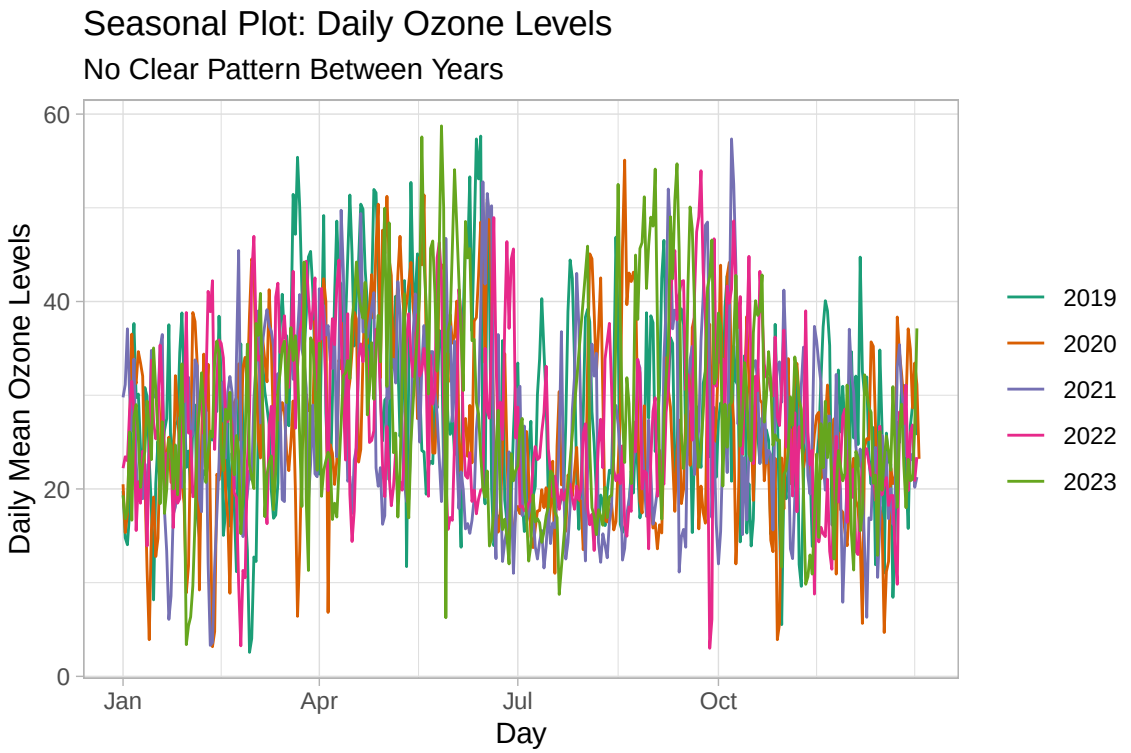

Figure 2: Seasonal Plot of Ozone Levels Across Years

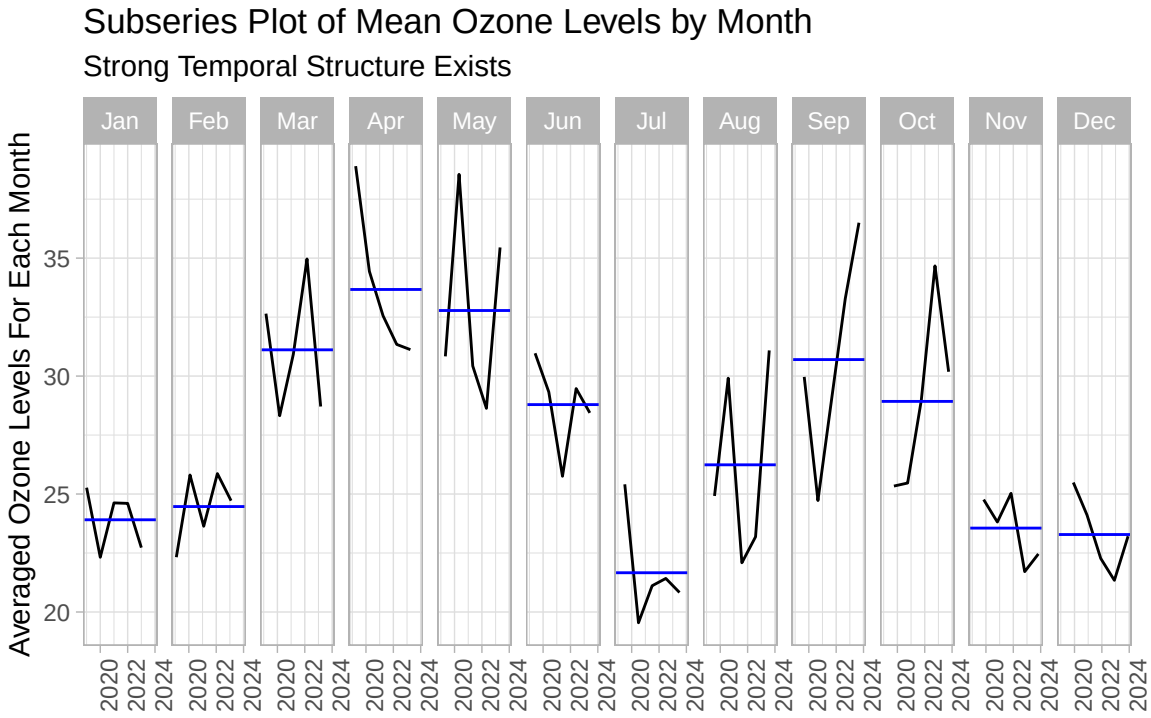

Figure 3: Subseries of Ozone Levels Across Months

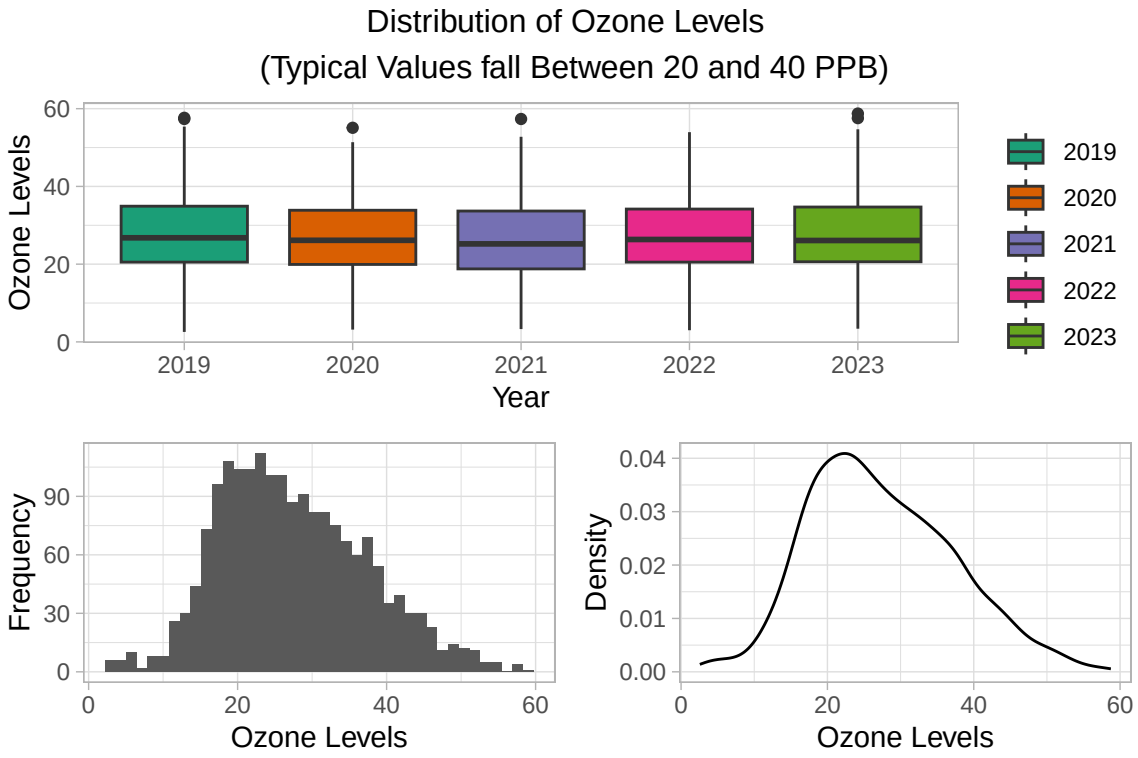

Figure 4: Distribution of Mean Ozone Levels

Modeling

Autocorrelation measures the similarity between a time series and a lagged version of itself. Partial autocorrelation measures the correlation between two time points after removing the effects of the intervening time points. Both are essential for identifying patterns and determining the appropriate lag order in time series analysis. In Figure 5, we have plotted both for our time series of interest.

Some important conclusions can be garnered from these two plots. With an exponential decrease in the ACF, this time series appears to have an auto-regressive component. The PACF is very high at lag 1, then tails off soon after. However, lags 2 and 3 could still be considered important. Based on this criteria, an AR(3), or auto-regressive process of order 3, could be a suitable model for our data.

A periodogram analyzes the frequency components of a time series, helping to identify dominant cycles or periodic patterns present in the data. Periodograms asymptotically estimate the spectral density. We have plotted a smoothed version of our periodogram in Figure 6. With a considerably low frequency, we can conclude that our time series has a long periodic component, suggesting the cycles in our data take a long time to complete. For an auto-regressive model, we would expect the coefficients to be primarily positive, indicating a less volatile time series. It will not jump back and forth across the mean.

Stationarity

Prior to modeling, especially for ARMA processes, it is crucial to determine whether the time series is stationary. There are two types of stationarity: strictly stationary, and covariance-stationary time

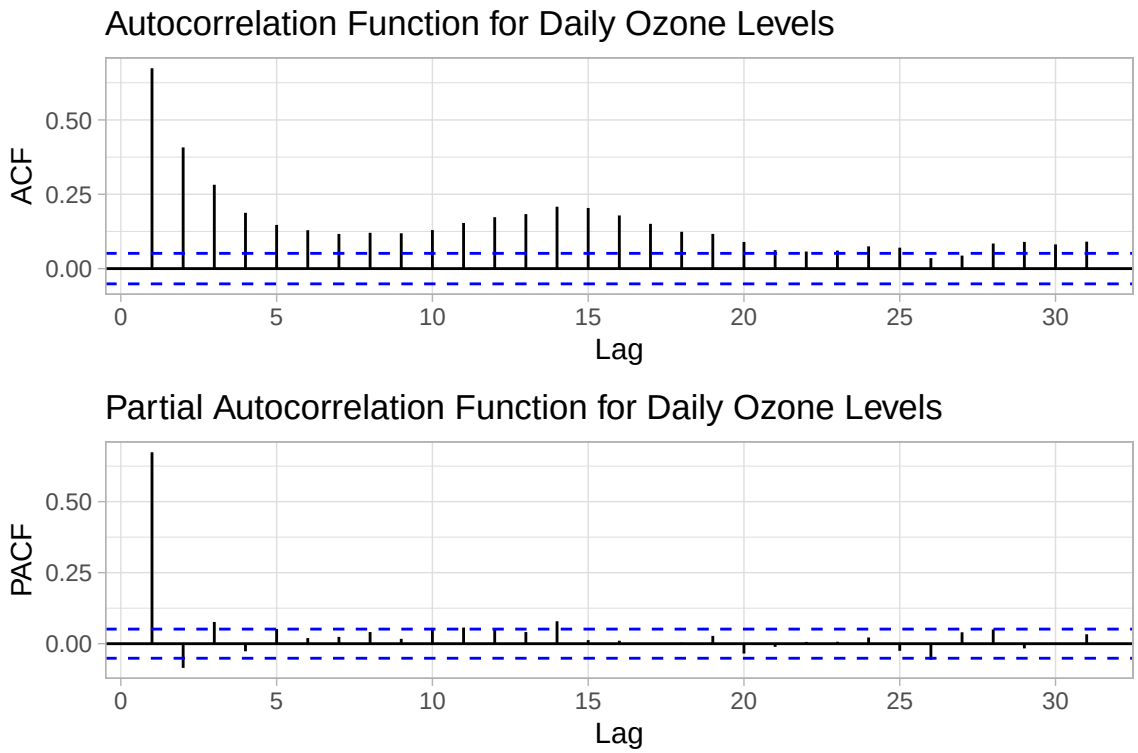

Figure 5: ACF and PACF for Ozone Time Series

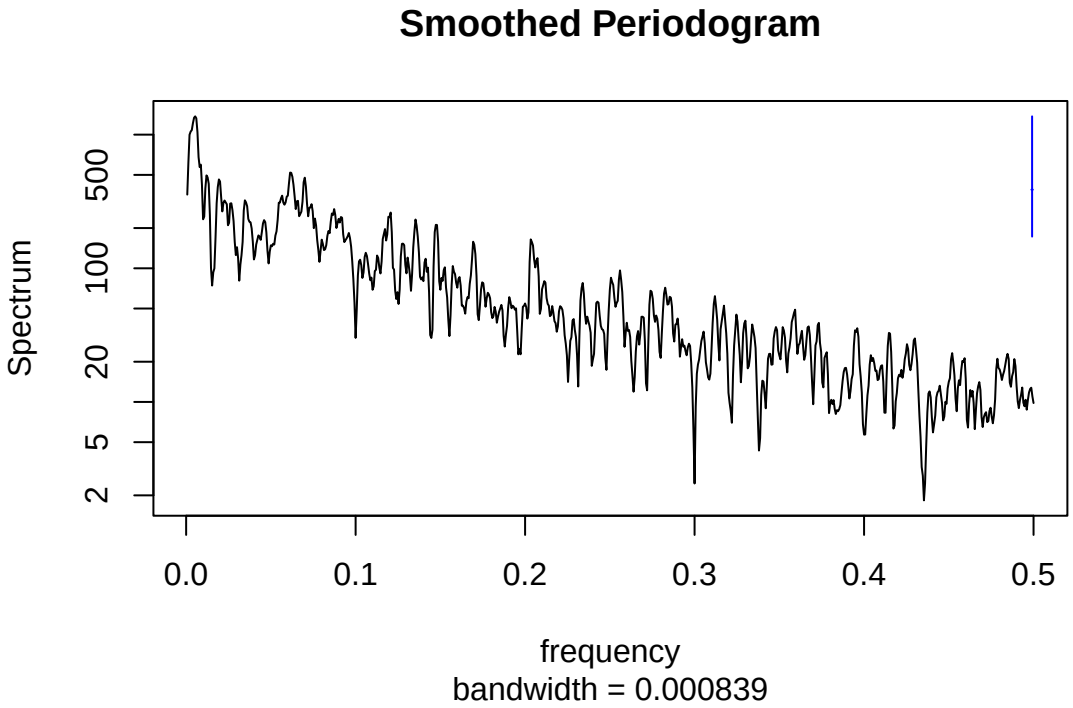

Figure 6: Estimating the Spectral Density

series. For our purposes, we will focus on the latter – where the covariance and the mean of our time series is not a function of time.

While our process appeared stationary in our previous plots, there are two unit-root tests we will implement to ensure this assumption. The Phillips-Perron test assumes a non-stationary time series under the null hypothesis, and the results in table 1 (a p-value of 0.01) suggest it is indeed stationary. The KPSS test assumes the opposite, and we fail to reject the null. Therefore, our time series appears stationary. We can continue with our implementation of an ARMA model.

Table 1: P-values for Unit-Root Tests

Phillips_Perron	KPSS
0.01	0.1

As mentioned above, an AR(3) could be a potential model. We will compare this to an AR(1) since there was some evidence to suggest that lags 2 and 3 are insignificant.

Our AR(3) model parameters and t-table can be found in table 2 while some model selection criteria are found in table 3. We have done the same for an AR(1) model in tables 4 and 5. A comparison of both models can be found in table 6. Notice that the MASE values and ACF1 values are lower for the ARIMA(3,0,0) model.

Table 2: AR(3) coefficients

term	estimate	std.error	statistic	p.value
ar1	0.74	0.03	28.3	0
ar2	-0.14	0.03	-4.4	0
ar3	0.08	0.03	2.9	0
constant	8.92	0.19	47.7	0

Table 3: Additional Model Selection Criteria

sigma2	log_lik	AIC	AICc	BIC	ar_roots	ma_roots
51	-4949	9908	9908	9934	1.4-0i, 0.2+3i, 0.2-3i	

Table 4: AR(1) coefficients

term	estimate	std.error	statistic	p.value
ar1	0.67	0.02	35	0
constant	8.91	0.19	47	0

Table 5: Additional Model Selection Criteria

sigma2	log_lik	AIC	AICc	BIC	ar_roots	ma_roots
52	-4959	9923	9923	9939	1.5+0i	

Table 6: Accuracy Measures for Both Models

.model	.type	ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
Arima300	Training	0	7.2	5.5	-11	26	0.54	0.55	0.00
Arima100	Training	0	7.2	5.6	-11	27	0.55	0.56	0.06

Although not much of a difference between models, the AR(3) model appears to have a lower AICc, implying a better model fit. The estimate of the variance of residuals is also lower (51 compared to 52). For this reason, we will move forward with the AR(3) model.

Something to note is that an automated model produces a SARIMA(2,0,2) which includes a seasonal component. For our analysis, we will only concern ourselves with ARIMA processes and not include a seasonal component to be modeled.

The model we will be evaluating can be written in the form below.

$$Y_t = 8.92 + 0.74Y_{t-1} - 0.14Y_{t-2} + 0.08Y_{t-3}$$

Residual Analysis

In figure 7, we have plotted the standardized residuals, their ACF, and a Q-Q plot. Overall, the model appears to be effective. The residuals seem to be white noise. There are a couple lag values that produce autocorrelations above our threshold, but that can also be attributed to random chance. The Q-Q plot also confirms that our residuals are approximately normally distributed (the tails only slightly deviate from what's expected). The Ljung-Box statistics could indicate some relationship among residuals, but there isn't enough to disregard our model assumptions.

A typical lag value that is considered for Ljung-Box is 10, which we have computed in table 7. A p-value of 0.55 suggests there is not autocorrelation among residuals.

Table 7: Ljung-Box Statistic

.model	lb_stat	lb_pvalue
Arima300	6	0.55

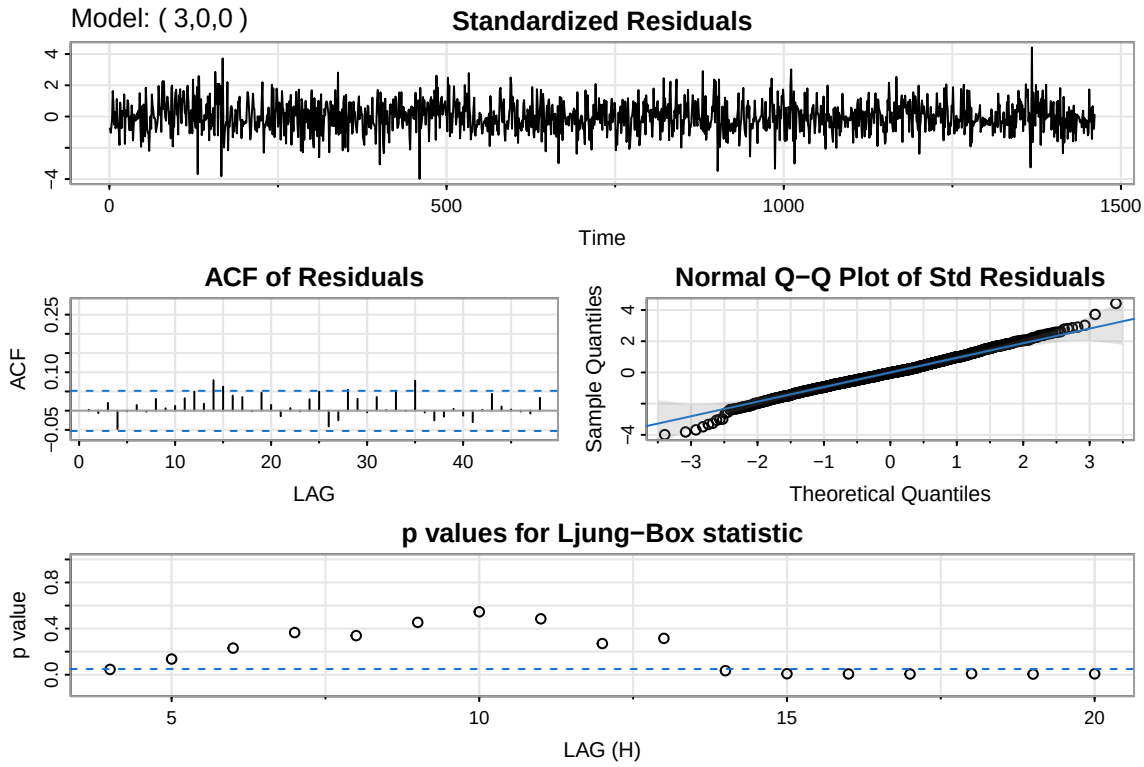

Figure 7: Residual Analysis

Fitted Values On Training Set

Figure 8 shows our fitted values plotted with our training set. Overall, our AR(3) model appears to perform well. It isn't great in predicting extreme values of mean ozone levels, but it performs well around the mean.

Our next goal is to see whether our model performs better at different times of the year. I have opted for a subseries plot where we can see the mean absolute error for the residuals for the given months. Therefore, the higher values suggest higher deviation from truth and lower values indicate good model predictions. In Figure 9, a noticeable pattern emerges wherein July is has more accurate predictions and several other months (such as June, April, and October) have worse prediction accuracy. This is something that could be improved with a seasonal component.

When are Estimates most Accurate?

Another consideration is whether our model performs better for certain ranges of ozone. To answer this question, I have included a scatter plot. In Figure 10, There appears to be a linear association between true ozone levels and residuals, suggesting that our model is best when predicting levels around the mean. It appears to over-predict low ozone values and under-predict high levels. This is to be expected but it is not ideal.

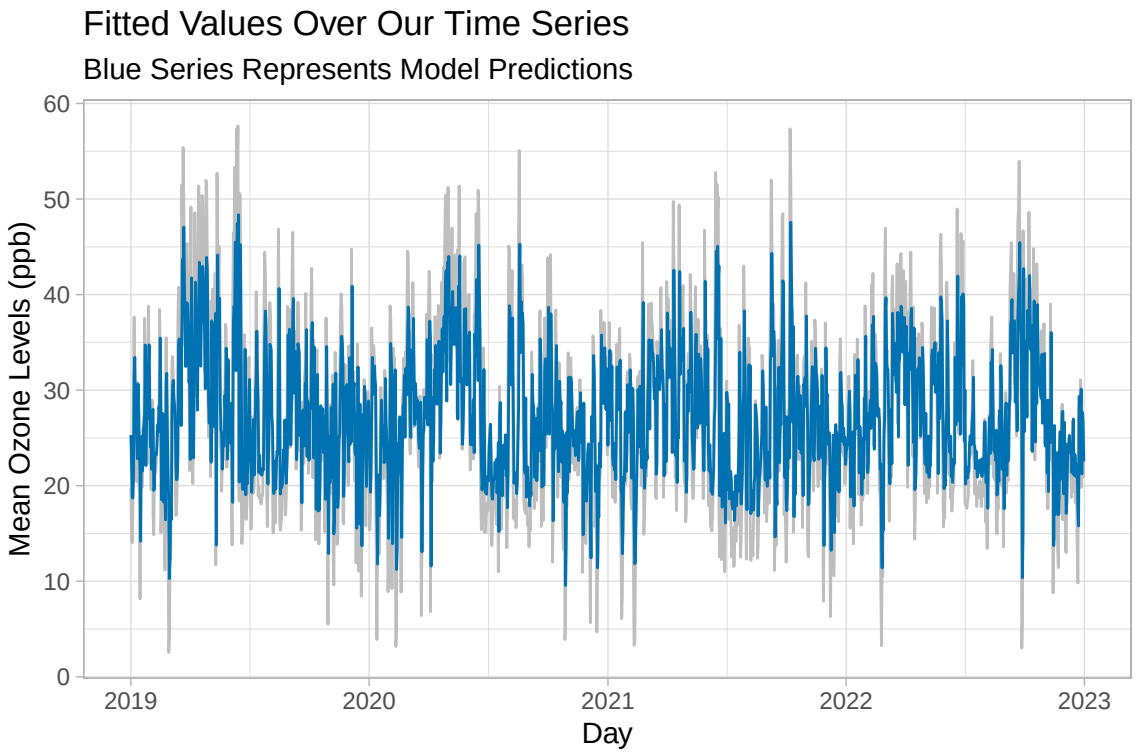

Figure 8: Plotting our Model

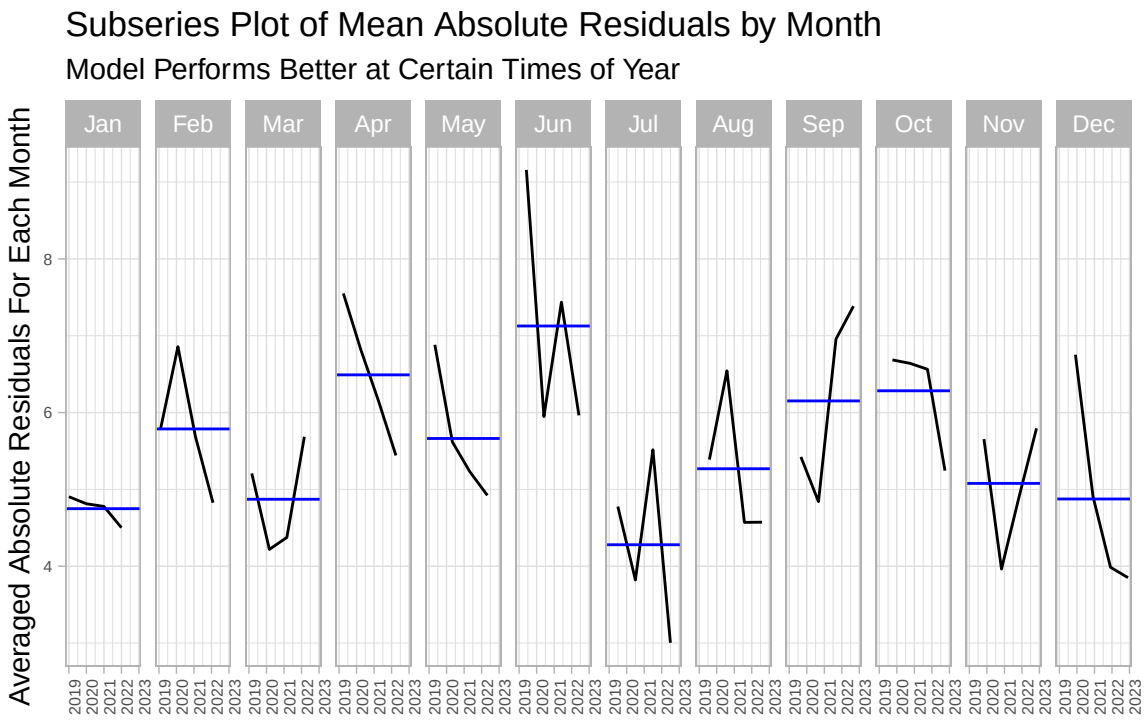

Figure 9: Model Accuracy by Month

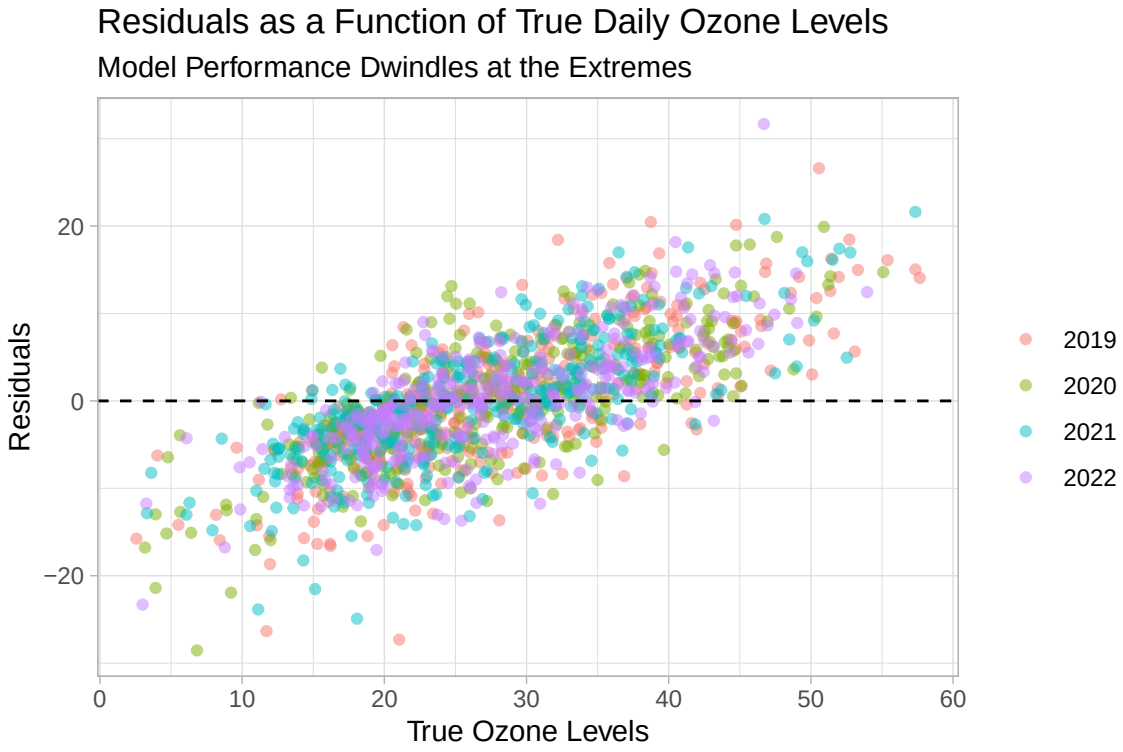

Figure 10: Ozone Levels and Model Performance

Examining Max Daily Ozone Levels

Another piece of data we haven't explored yet is the daily max ozone levels. Something to consider is whether the 99% confidence interval of our predictions are relatively close to the max daily ozone levels. While this isn't necessarily a fair comparison (they are entirely different measures), it is something to consider. In Figure 11, one can notice how the upper threshold of our CI aligns closely with the daily max. It is unreasonable to make any statistical conclusions but it is an interesting trend.

In Figure 12, we have summarized this distribution in a group of box-plots, a histogram, and a density plot. There appears to be a slight negative skew, indicating the upper 99% CI is often much lower than the daily max. Nonetheless, the median difference does appear to remain steadily around 0.

Forecasting

Our next objective is to forecast using our AR(3) model. While our dataset initially covered 2019-2023, we excluded 2023 to be used as test data. This will enable us to compare our model output to data that is was not trained on, further validating our assumptions. In Figure 13, we can observe the forecasts for 1, 2, 3, and 14 day horizons across the entire span of 2023.

Overall, our model seems accurate, especially for h = 1, 2, and 3 steps ahead. The 14 day forecast essentially estimates the mean ozone level, which is expected for a stationary ARMA process. Another concept to consider is that the confidence interval for the 14 day forecast is expected to be wider than if we falsely assumed our observations were independent and identically distributed (iid). With the presence of autocorrelation, this interval widens as a result of "effective sample size." Nonetheless, the

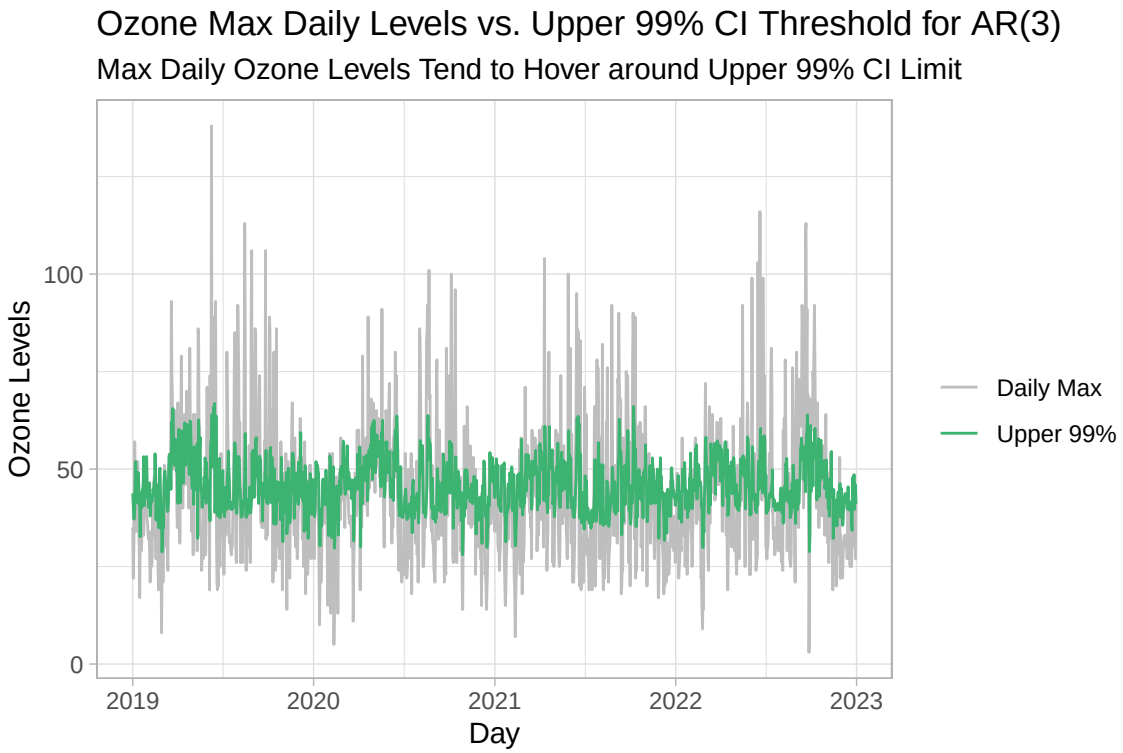

Figure 11: Analyzing Max Daily Ozone Levels

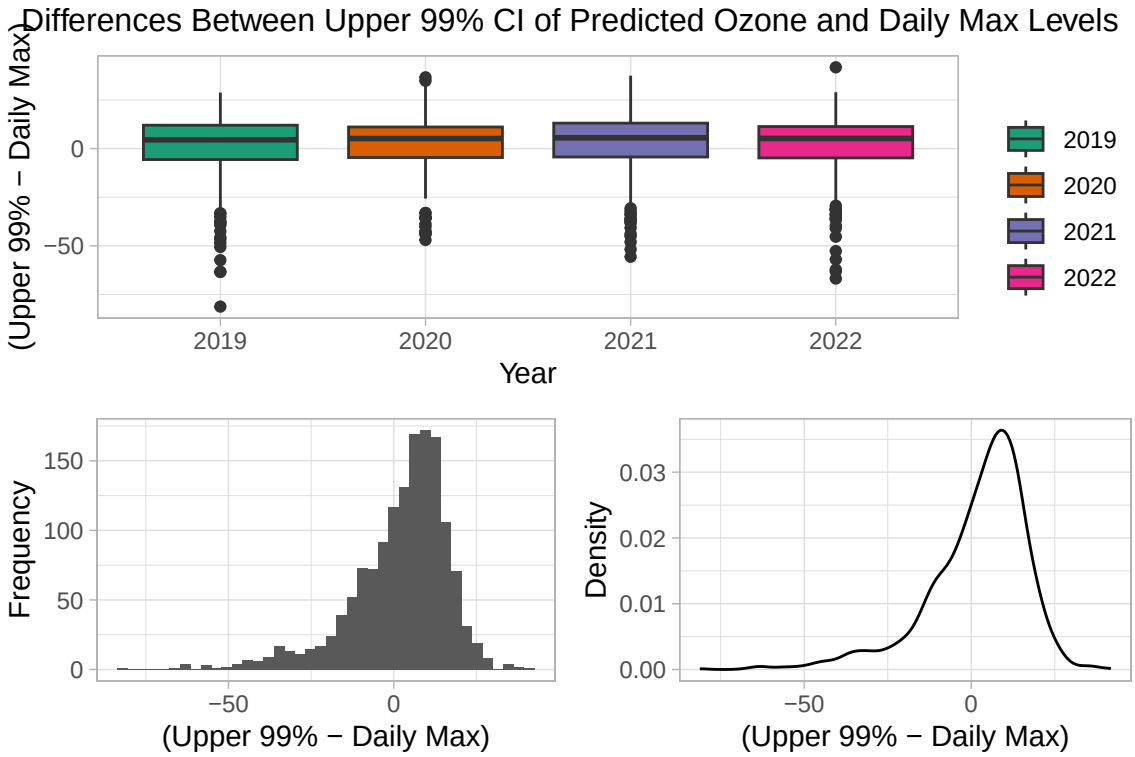

Figure 12: Upper 99% CI and Max Ozone

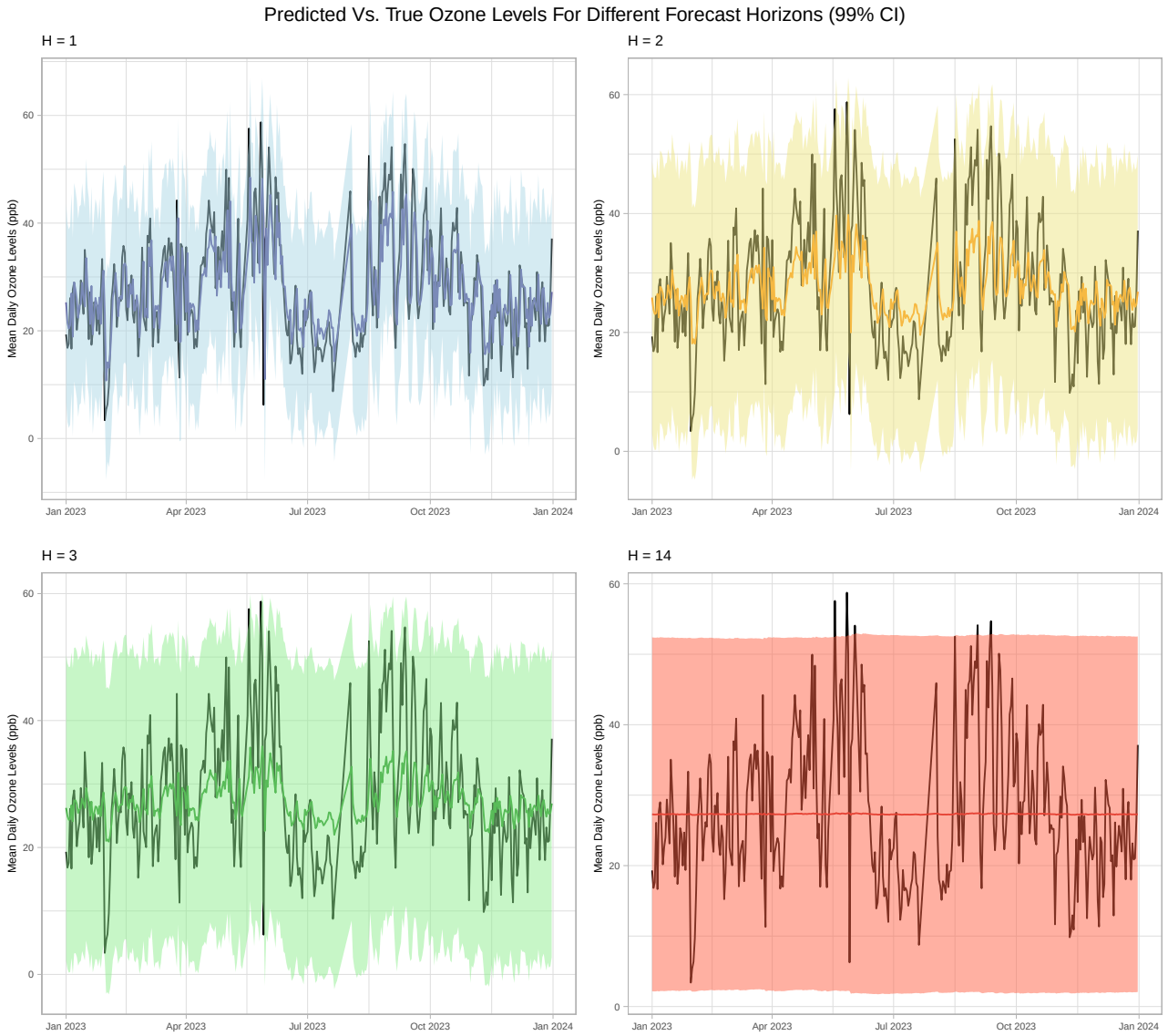

Figure 13: Forecasting 2023 with an ARIMA(3,0,0)

1 step-ahead forecast is especially pleasing, as it appears to fit the data pretty well. While the model doesn't have perfect point forecasts for extreme values, nearly all observed values in 2023 are captured by our 99% confidence intervals.

Another point to mention is the increasing nature of the CIs for further forecast horizons. This is typical of a time series process wherein we are less sure of our forecasts as we move further from the present.

Inspecting Forecast Errors

In Figure 14, we have produced qq-plots of our standardized forecast errors. Overall, they appear to generally follow normality with greater forecast horizons tailing off toward the more extreme quantiles. The model assumptions are generally satisfied.

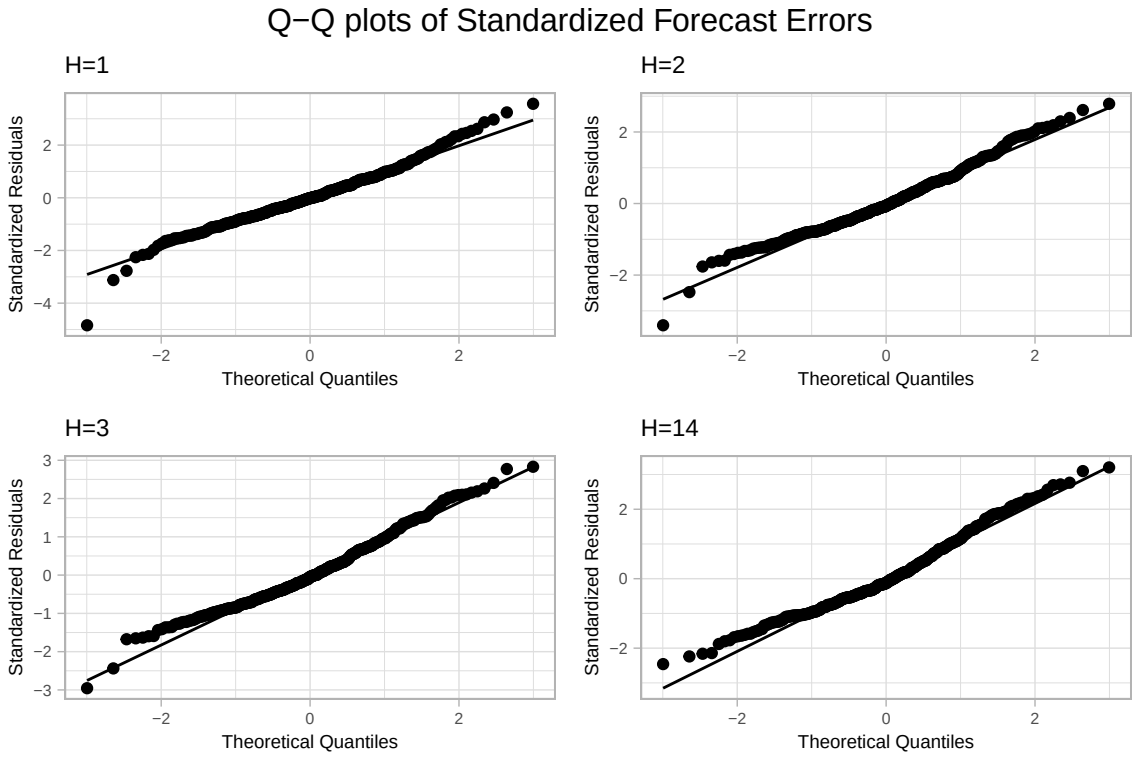

Figure 14: Standardized Error Analysis

Table 8: Forecast Error Statistics

	forecast	ME	RMSE	MAE	MPE	MAPE
Test set	H = 1	0.24	7.33	5.58	-9.14	24.84
Test set	H = 2	0.39	7.9	6.25	-11.69	28.36
Test set	H = 3	0.47	8.5	6.76	-12.84	30.64
Test set	H = 14	0.65	10.28	8.22	-15.28	36.83

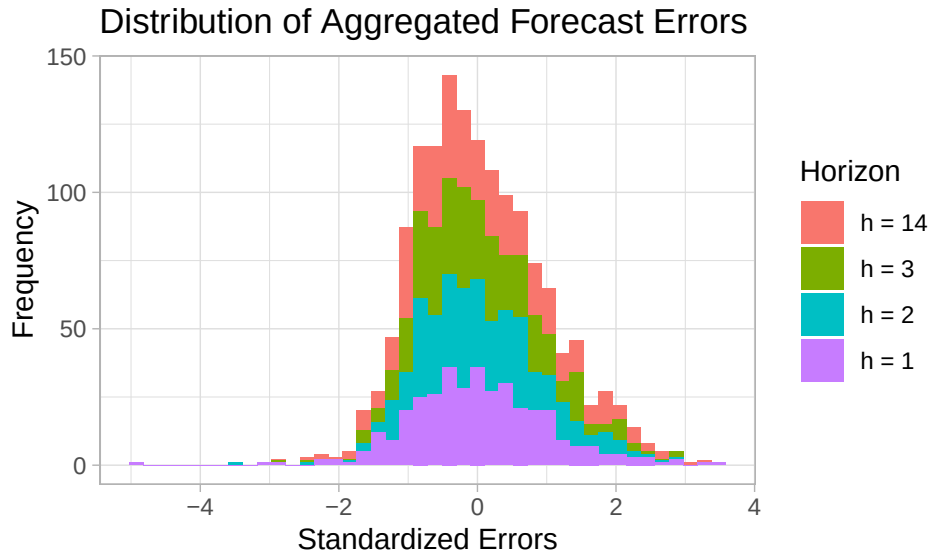

Figure 15: Standardized Error Analysis

Standardized Error Analysis

Figure 15 groups the standardized forecast errors and produces a histogram. Overall, the errors are approximately normally distributed with a slight leftward skew. Looking at our previous plots, it appears that one observation in May was significantly lower than expected. This is reflected in our plots.

Moreover, the forecast accuracy for each horizon has been shown in table 8. The RMSE is 7.3 for h=1, which is barely greater than our training RMSE of 7.2. As expected, the errors are magnified as the horizon extends.

Forecast Errors Across Months

Once again, there appears to be a discrepancy in forecast accuracy across months. June tends to be modeled the best (perhaps due to the overall lower levels of ozone) while May and September have poorer forecasts, suggesting month could be an exogenous variable in a more suitable model. Something to note is how the RMSE for the different horizons don't always align: for instance, while the RMSE for February is higher than December for h = 14, it is the opposite for the other forecast horizons. Additionally, in March specifically, our h = 1 forecasts performed worse than all other horizons; this is atypical.

Conclusion

It's clear from the analysis that our AR(3) model provides valuable insights into the daily fluctuations of ozone levels in Houston. By examining the time series plot, seasonal patterns, and subseries plots, we identified significant temporal structures within the data. Our model selection process, guided by autocorrelation and partial autocorrelation functions, led us to choose the AR(3) model as the most appropriate for forecasting.

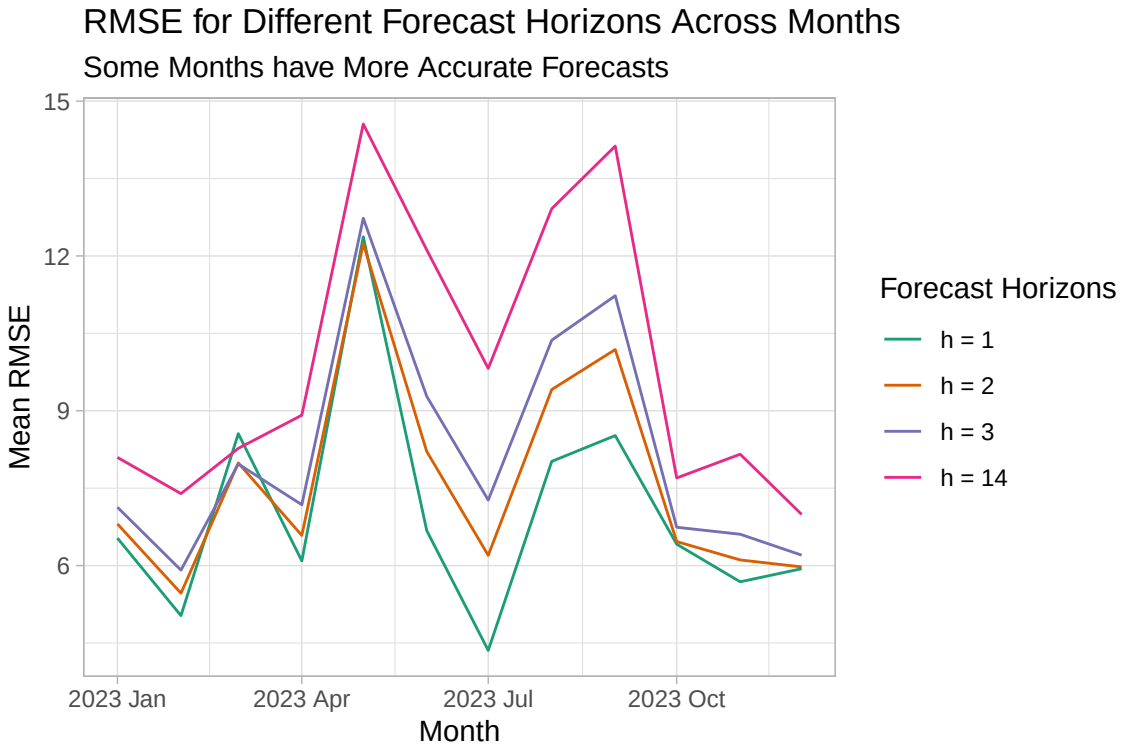

Figure 16: Forecast Accuracy Across Months

The stationarity tests confirmed that our time series is covariance-stationary, validating our modeling approach. We evaluated the model's performance through residual analysis, which indicated that our AR(3) model adequately captures the underlying patterns in the data.

Forecasting results showed promising accuracy for short-term predictions, with wider confidence intervals for longer forecast horizons, as expected in time series analysis. The qq-plots of standardized forecast errors demonstrated that our model assumptions are generally satisfied.

Overall, our study provides valuable insights into the dynamics of ozone levels in Houston and offers a reliable framework for short-term forecasting. Further research could explore incorporating additional variables or refining the model to improve long-term forecasting accuracy.

Link To Code