FINAL PROJECT

Stats 415

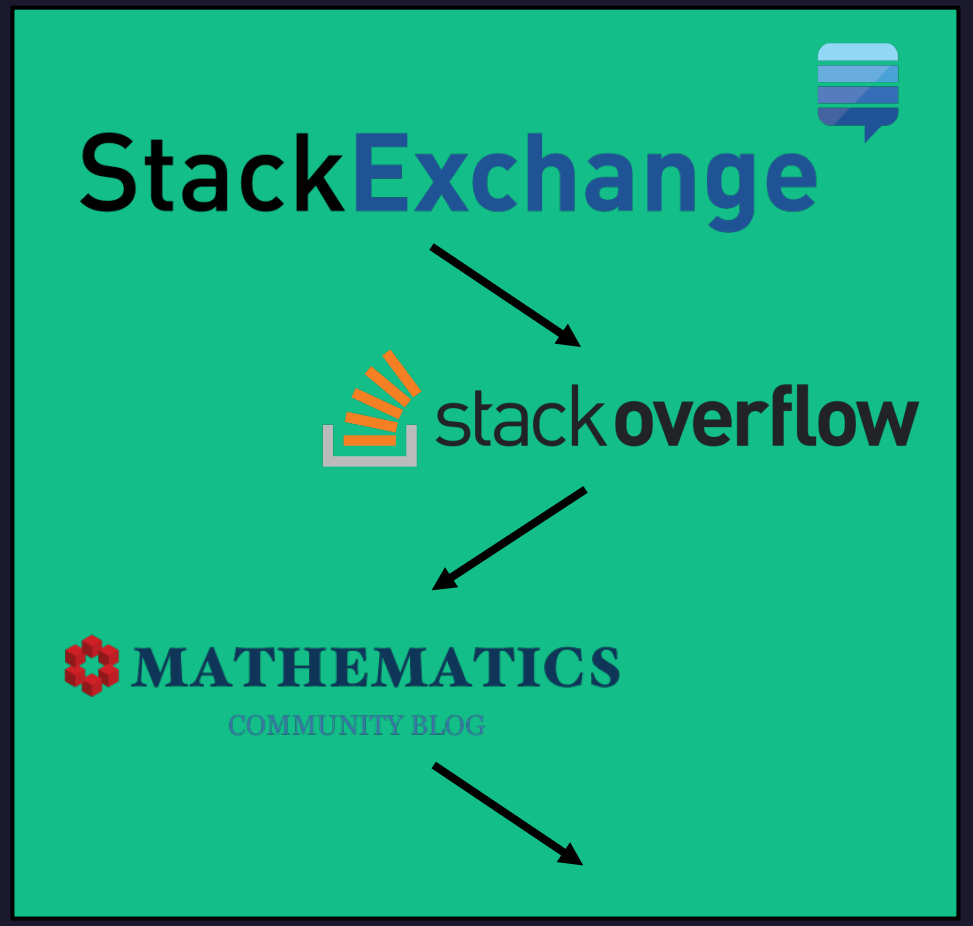
Andrew Mashhadi

Spring 2023

FORECASTING MONTHLY "TIME-SERIES" RELATED QUESTIONS ON THE STACK EXCHANGE

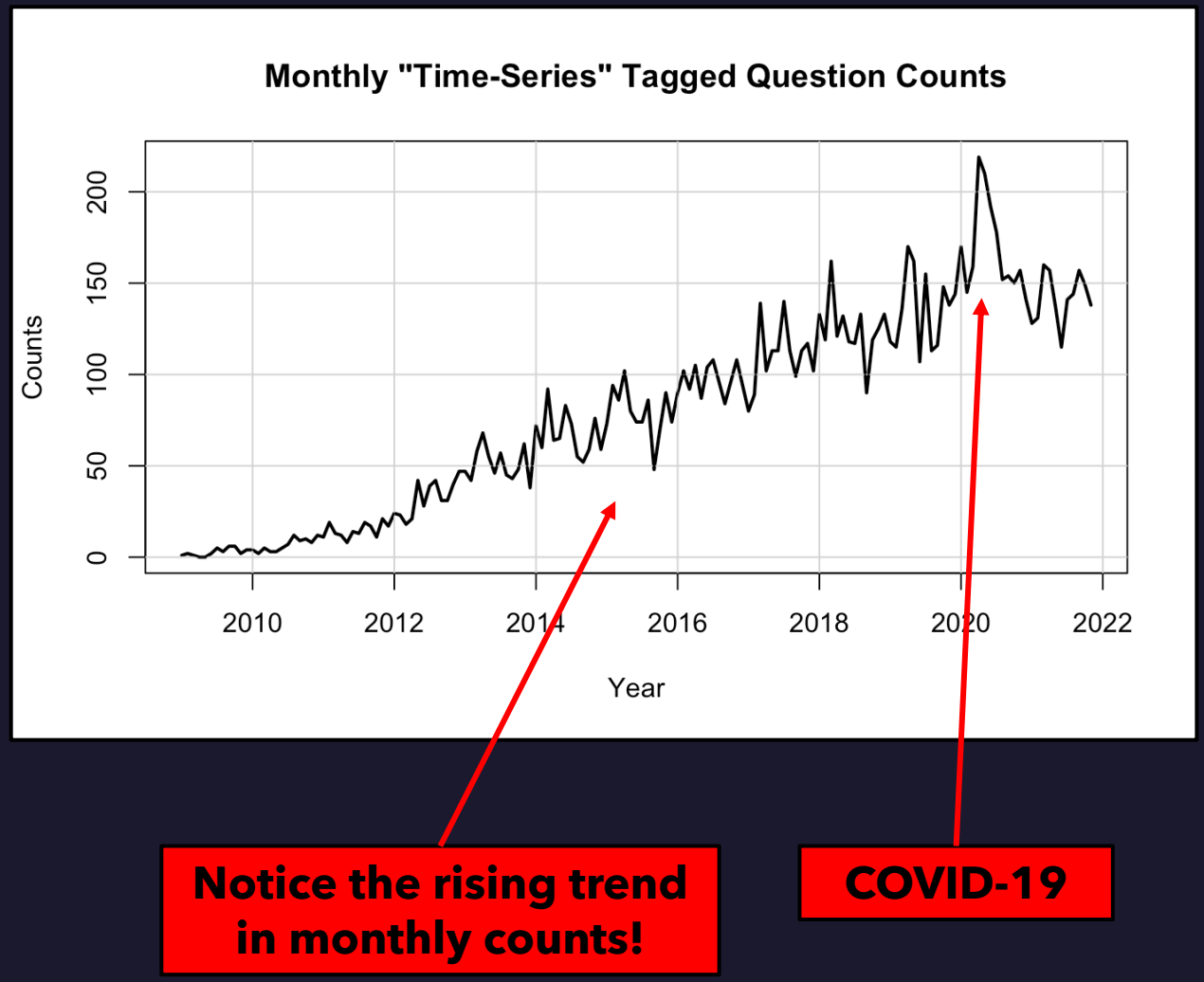
Introduction

- Stack Exchange is a network of Q & A websites.
- Since 2008, they have grown as one of the most used networks for programming and mathematics questions.
- Every posted question may contain various tags to help locate other interested users.
- **Time Series Analysis:**
  - Topic within mathematics and statistics that encompasses techniques for examining or modeling data points arranged in chronological order.
  - The recent surge in machine learning and data science has sparked a renewed interest in this subject.
  - Also happens to be the topic of this course ☺.
- In this study, we aim to analyze and model the monthly count of Stack Exchange questions tagged with "time-series".



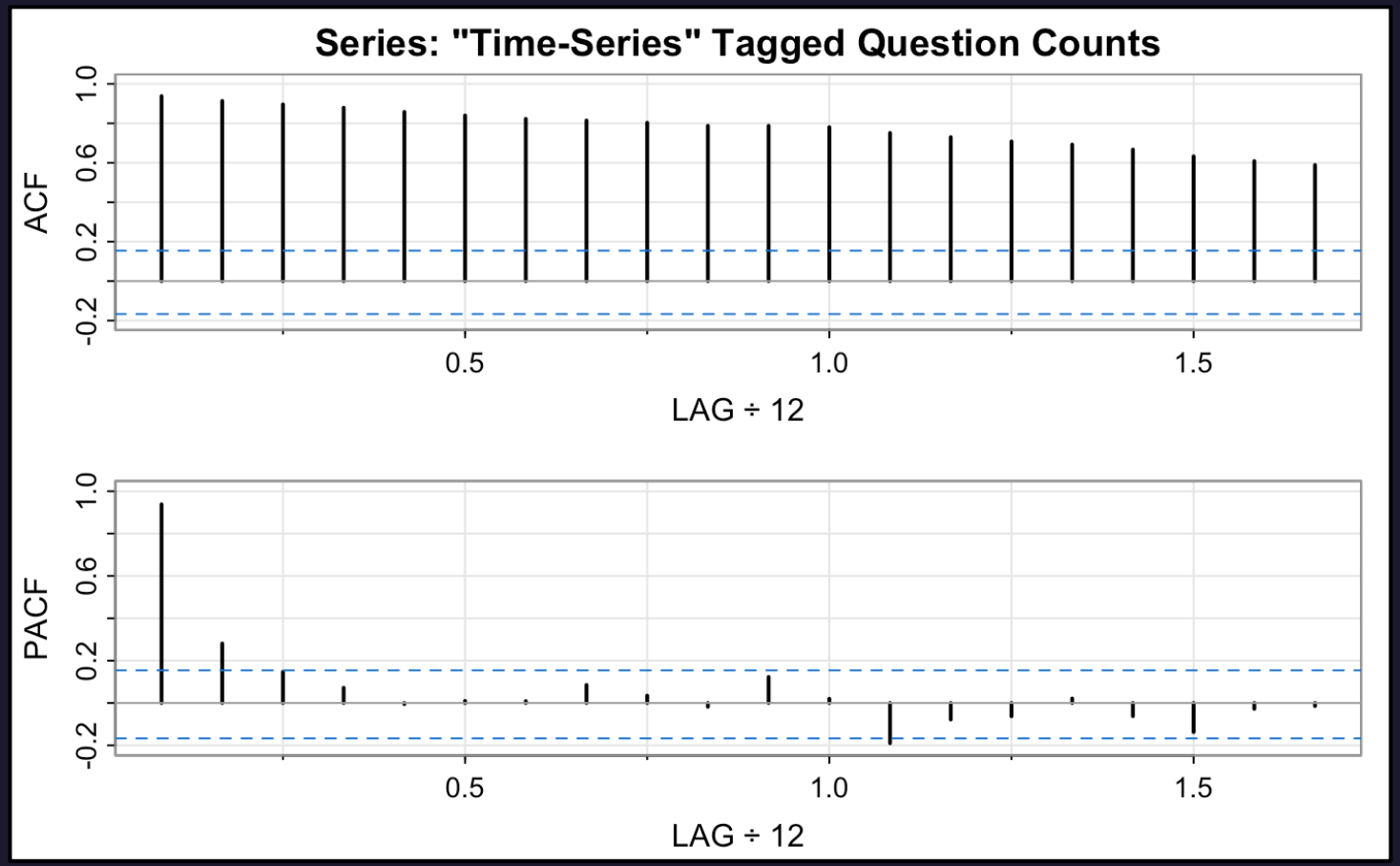
Data

- The Stack Exchange provides an open source tool known as the *Stack Exchange Data Explorer* for running arbitrary SQL queries against public data from the Stack Exchange network.
- We collected the monthly counts of Stack Exchange questions tagged with “time-series”. In other words, our data is the number of time series related questions for each available month.
- The data consisted of monthly counts from January 2009 to May 2023 (n=173), with no missing entries.
- First 90% (Jan. 2009 - Nov. 2021) used for training and remaining 10% (Dec. 2021 - May 2023) was used as ”unseen” testing set.



Data

- Large ACF values displayed in the top plot indicate that the data is currently being dominated by the trend.
- Assuming our data is trend stationary, the strong trend will likely obscure the behavior of the associated stationary process and introduce extremely low frequency components in a periodogram.



Removing Trend

- Trend must be removed before we can conduct much of our analysis and modeling.
- Fit a linear model and a 3rd degree polynomial.
- Found that the 3rd degree polynomial has a significantly better fit

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	78.948	1.263	62.525	≈ 0
time	651.295	15.720	41.431	≈ 0
time^2	-16.894	15.720	-1.075	0.284
time^3	-66.857	15.720	-4.253	≈ 0

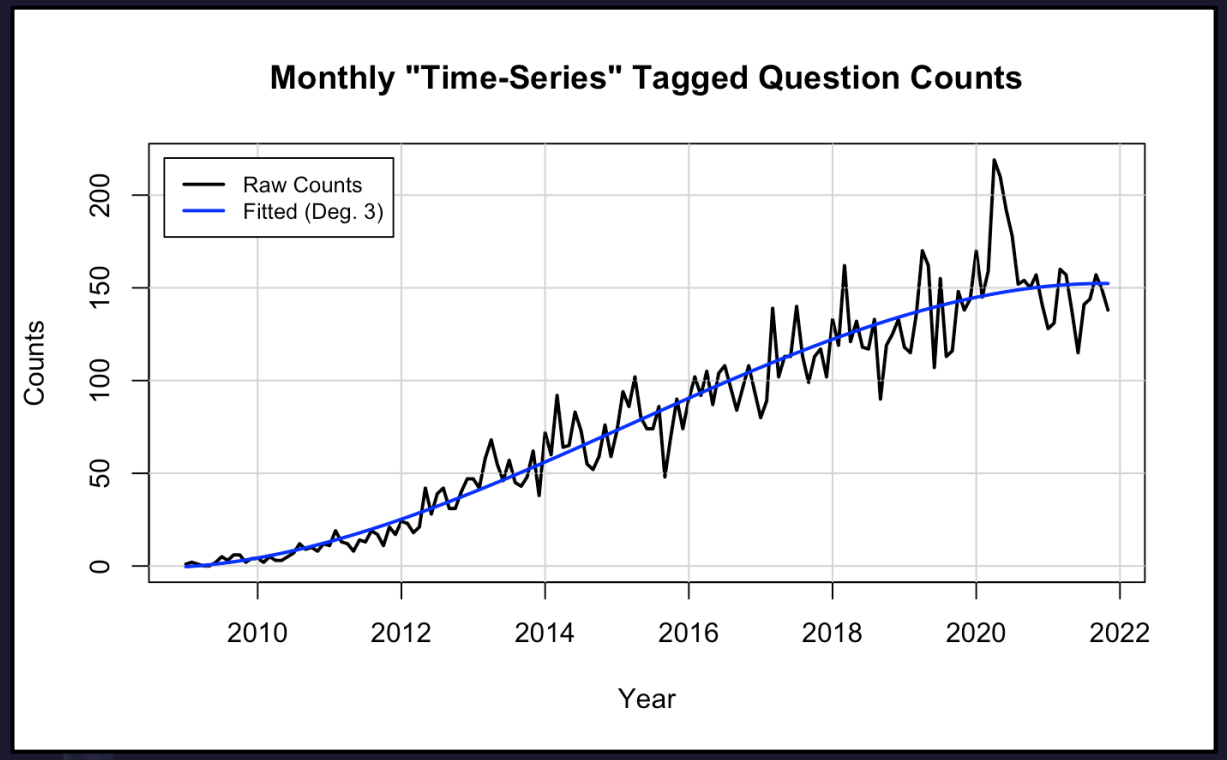
Residual standard error: 15.72 on 151 *df*
Multiple R-squared: 0.92
F-statistic: 578.6 on 3 and 151 *df*
p-value: ≈ 0

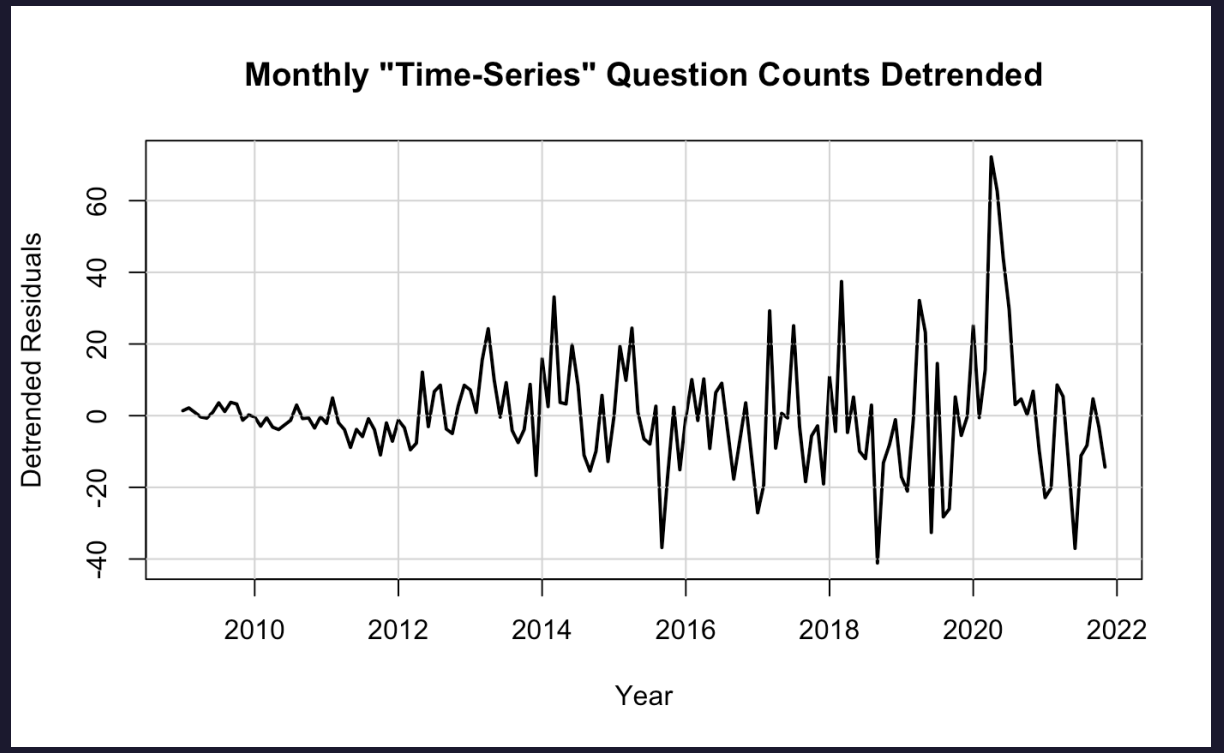
Polynomial Regression Summary

Fit	Resid. Df	RSS	DF	Sum. Sq.	F	$\Pr(>F)$
Linear Fit	153	42070				
Poly Deg. 3	151	37315	2	4755.2	9.621	< 0.001

Analysis of Variance Table

Removing Trend



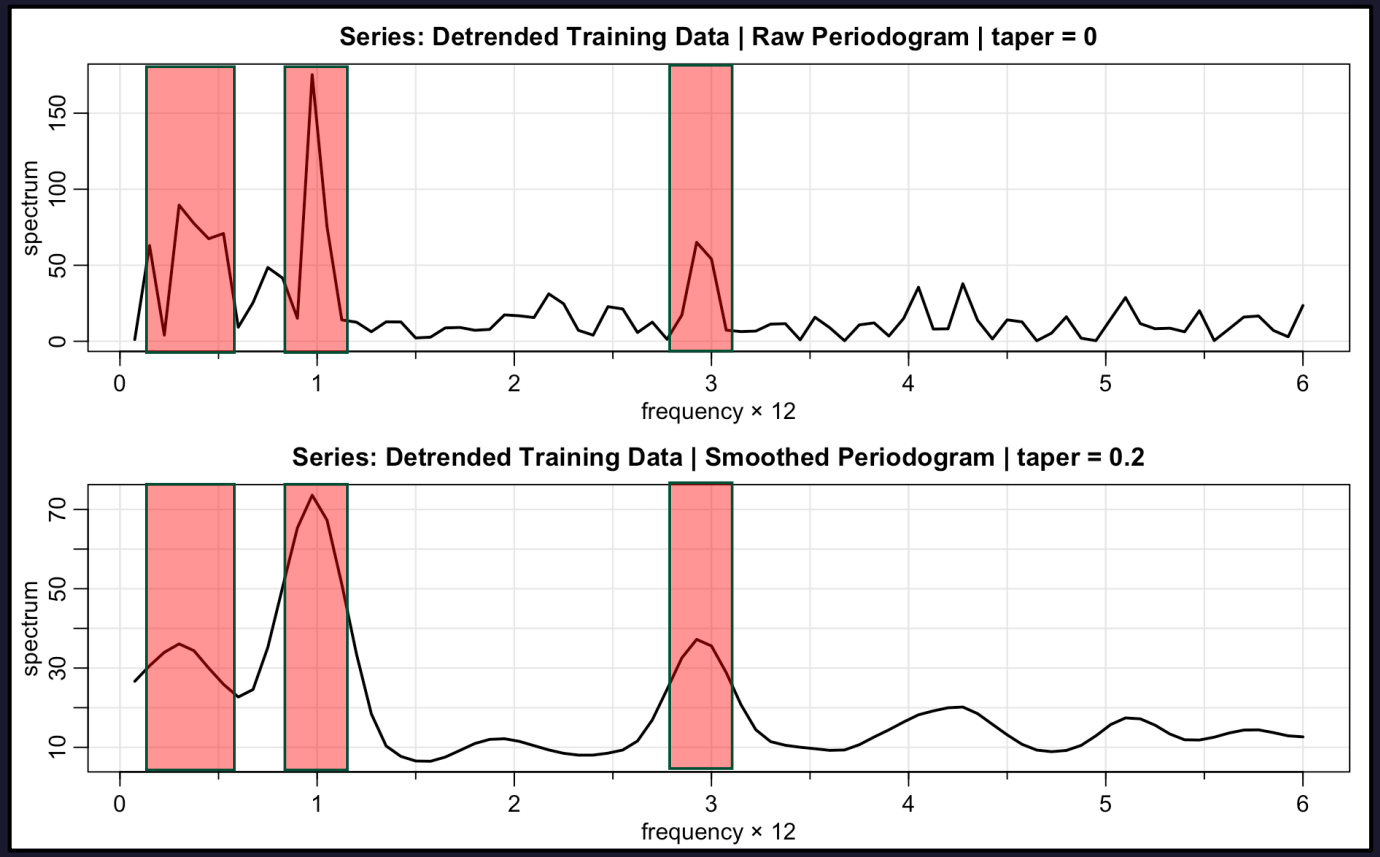


Spectral Analysis

- We estimated the spectrum of our detrended training data to help us determine predominant periods and to obtain approximate confidence intervals for the identified periods.

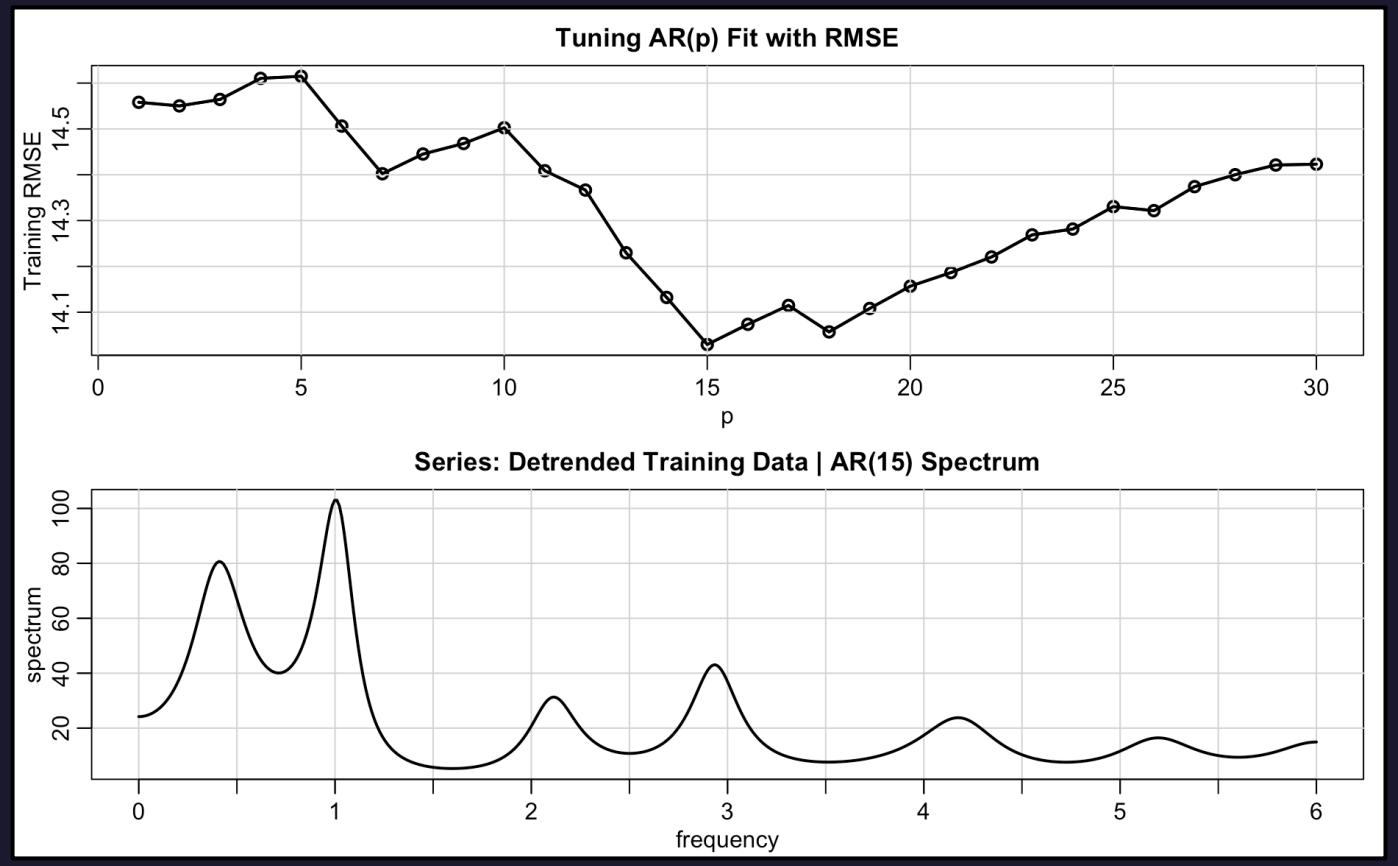
95% C.I.	Freq. × 12	Lower	Upper
Two-Sided	≈ 0.3	18.68	96.92
Two-Sided	≈ 1	38.08	197.61
Two-Sided	≈ 3	19.26	99.95
One-Sided	≈ 0.3	20.71	∞
One-Sided	≈ 1	42.23	∞
One-Sided	≈ 3	21.36	∞

Confidence Intervals (Non-Parametric)

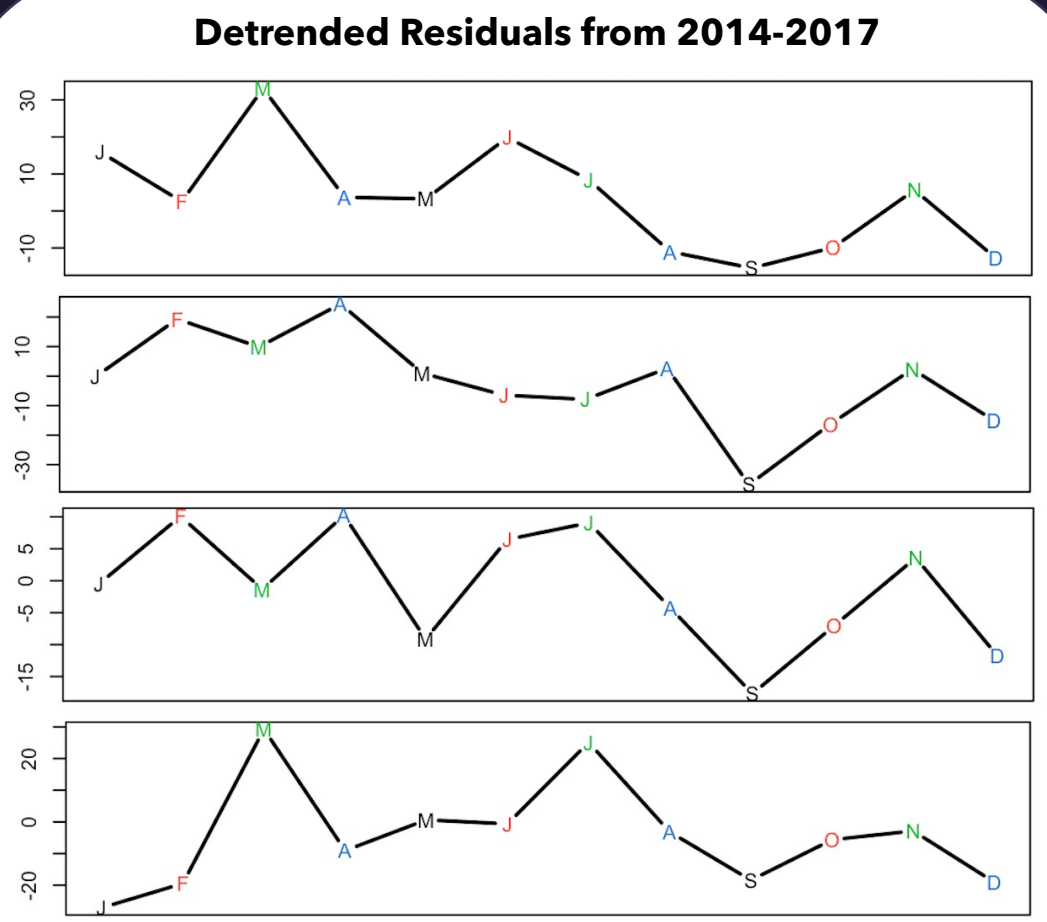


Spectral Analysis

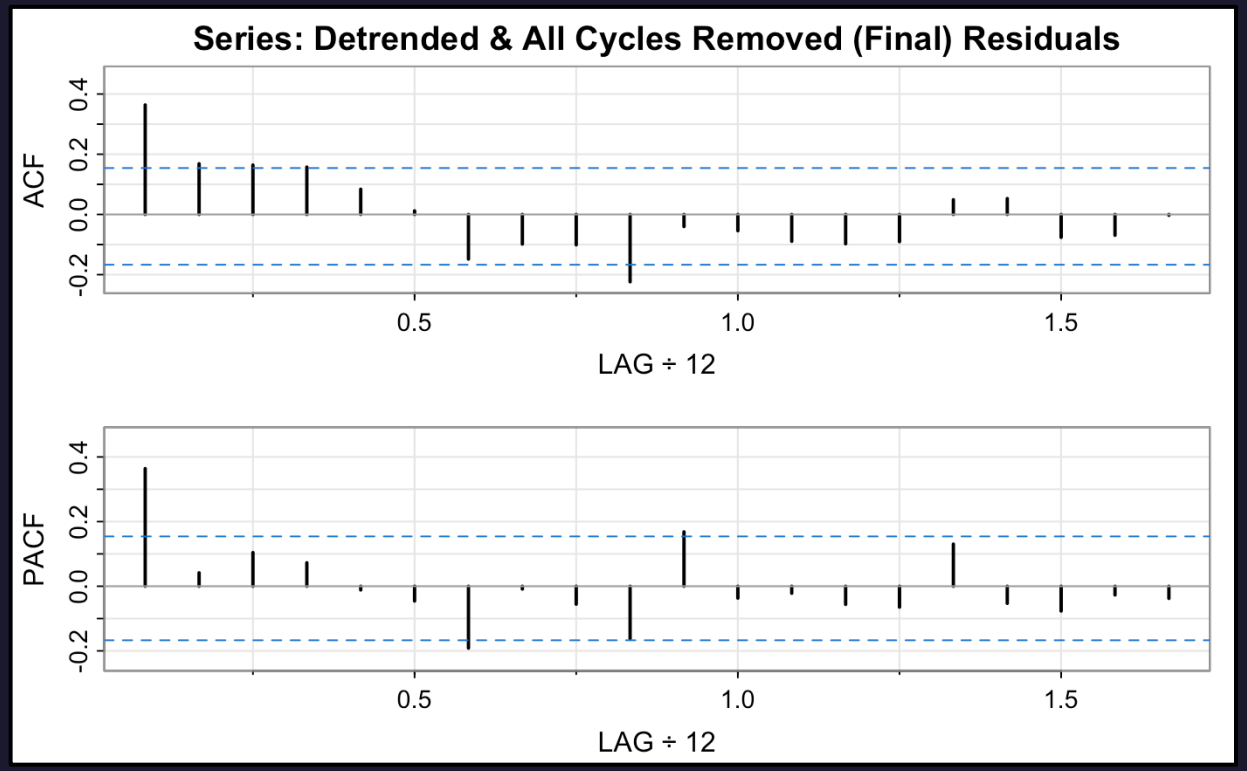
- Also used autoregressive spectral estimator, with an AR(15).
- Autoregressive spectral estimators generally have superior resolution in problems when several closely spaced narrow spectral peaks are present, giving us more confidence in our identification of the 40-month cycle.
- We assessed the possibility that the apparent 4-month cycle is a harmonic.



Spectral Analysis

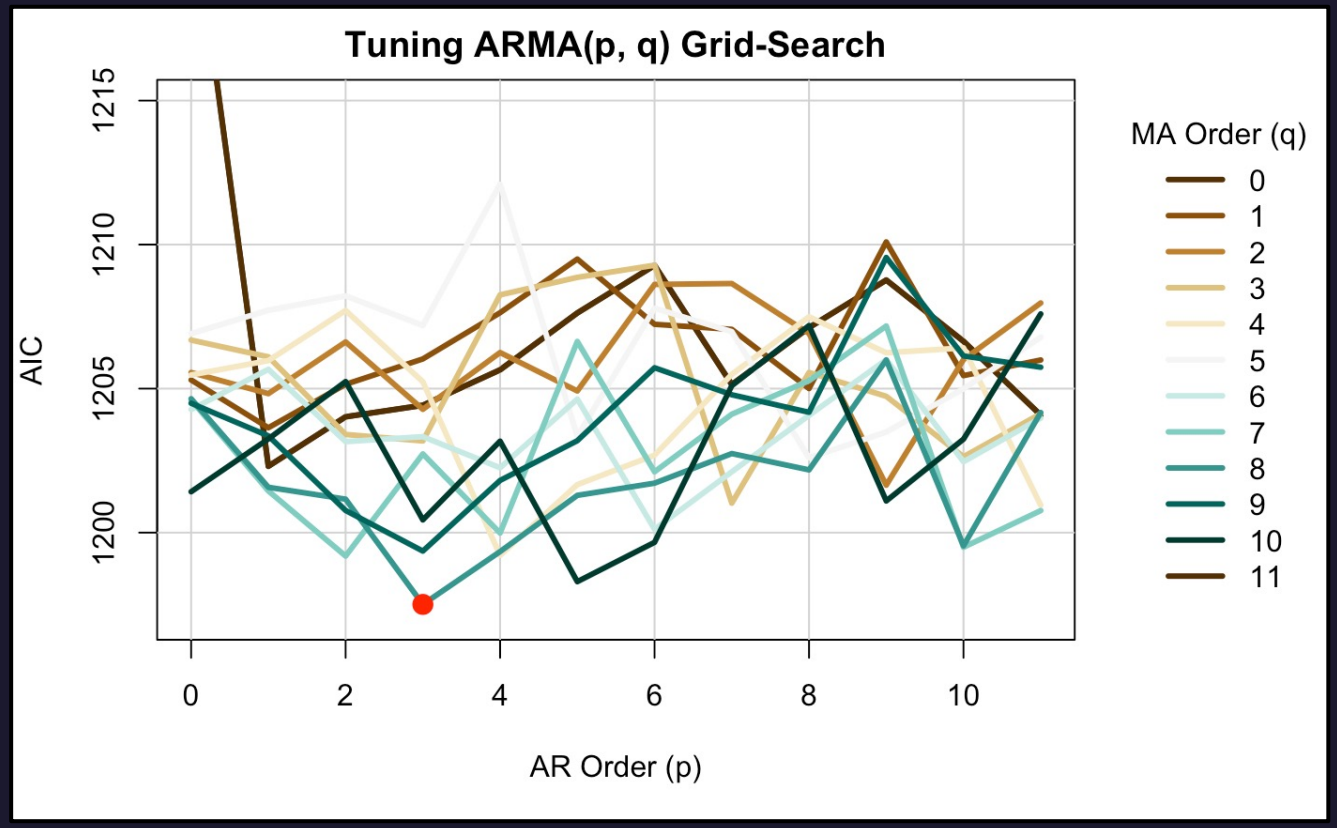


- Iteratively removed **all three** cycles by subtracting the associated monthly averages from the detrended data.



Modeling

- We proceeded to fit ARMA models on the detrended and cycle-removed training data.
- Used grid search with the *Akaike Information Criterion* (AIC) as the tuning metric.
- The ARMA(3, 8) model achieved the lowest AIC score of 1197.51.



Final Model:

$$r_t = 0.107 - 0.407r_{t-1} + 0.470r_{t-2} + 0.631r_{t-3} + w_t + 0.763w_{t-1} - 0.279w_{t-2} - 0.754w_{t-3} - 0.060w_{t-4} + 0.030w_{t-5} - 0.073w_{t-6} - 0.334w_{t-7} - 0.311w_{t-8}$$

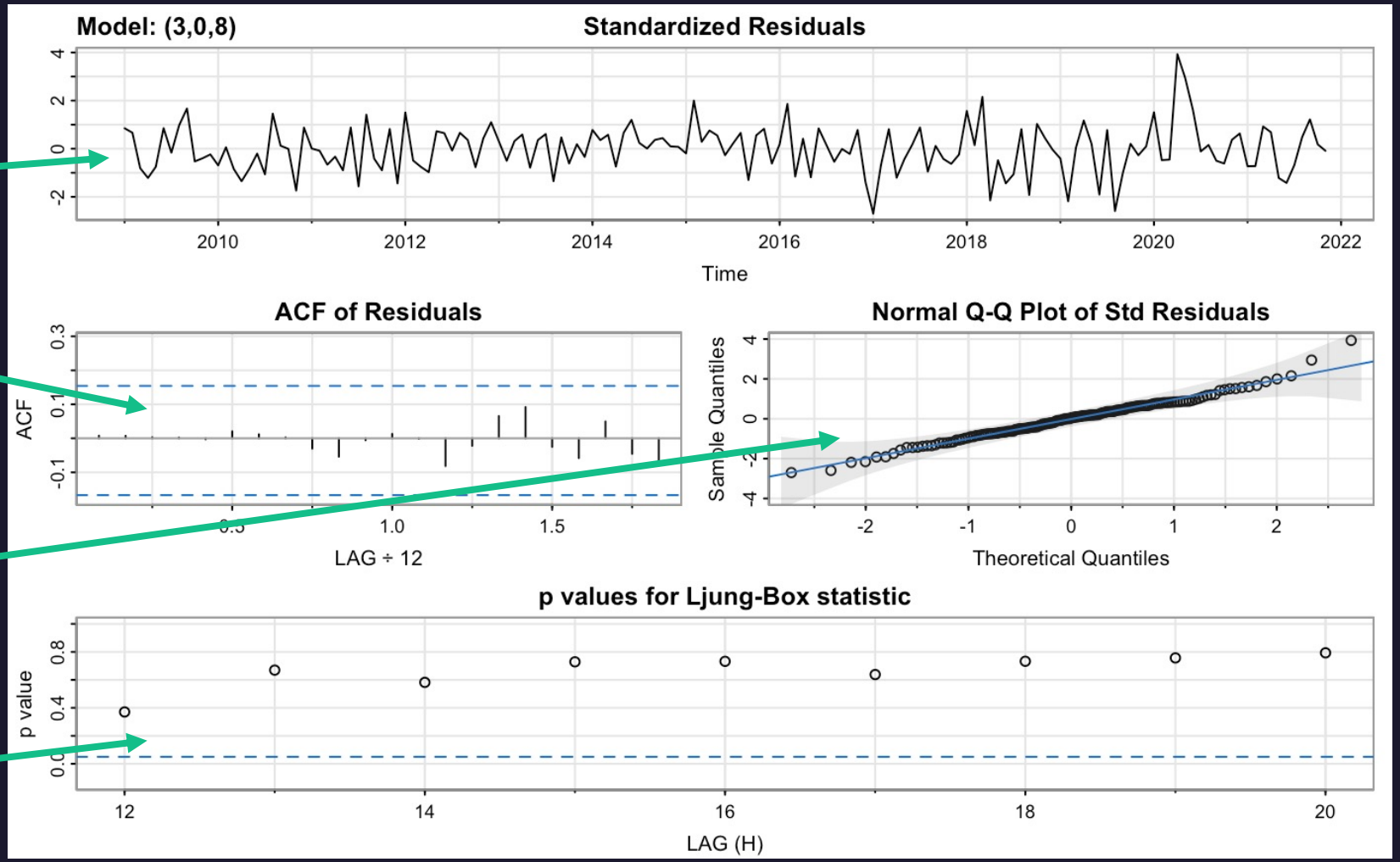
Modeling

No discernible patterns and only one outlier > 3 in 2020

No deviations from our model assumptions

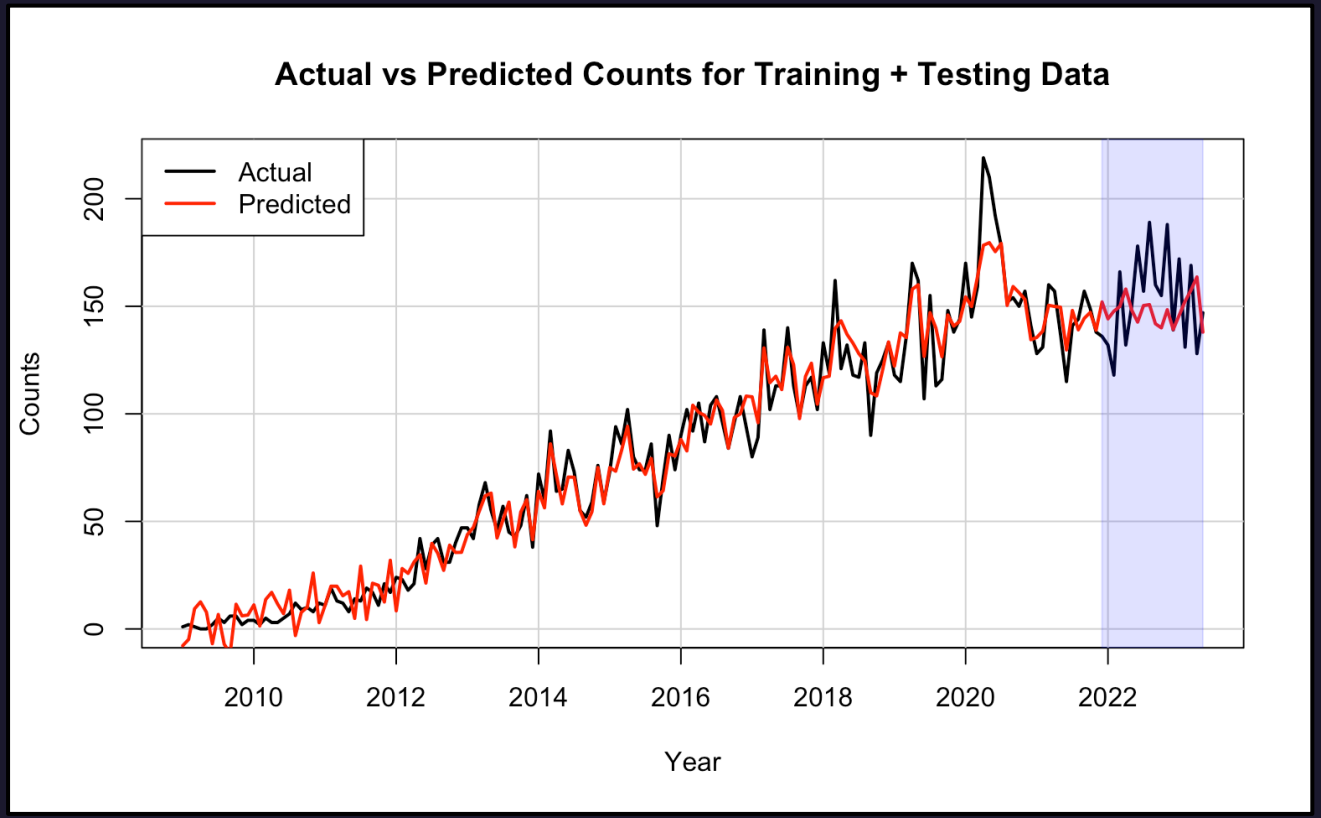
Reasonable adherence to the assumption of normality

Q-statistics remained non-significant across the lags



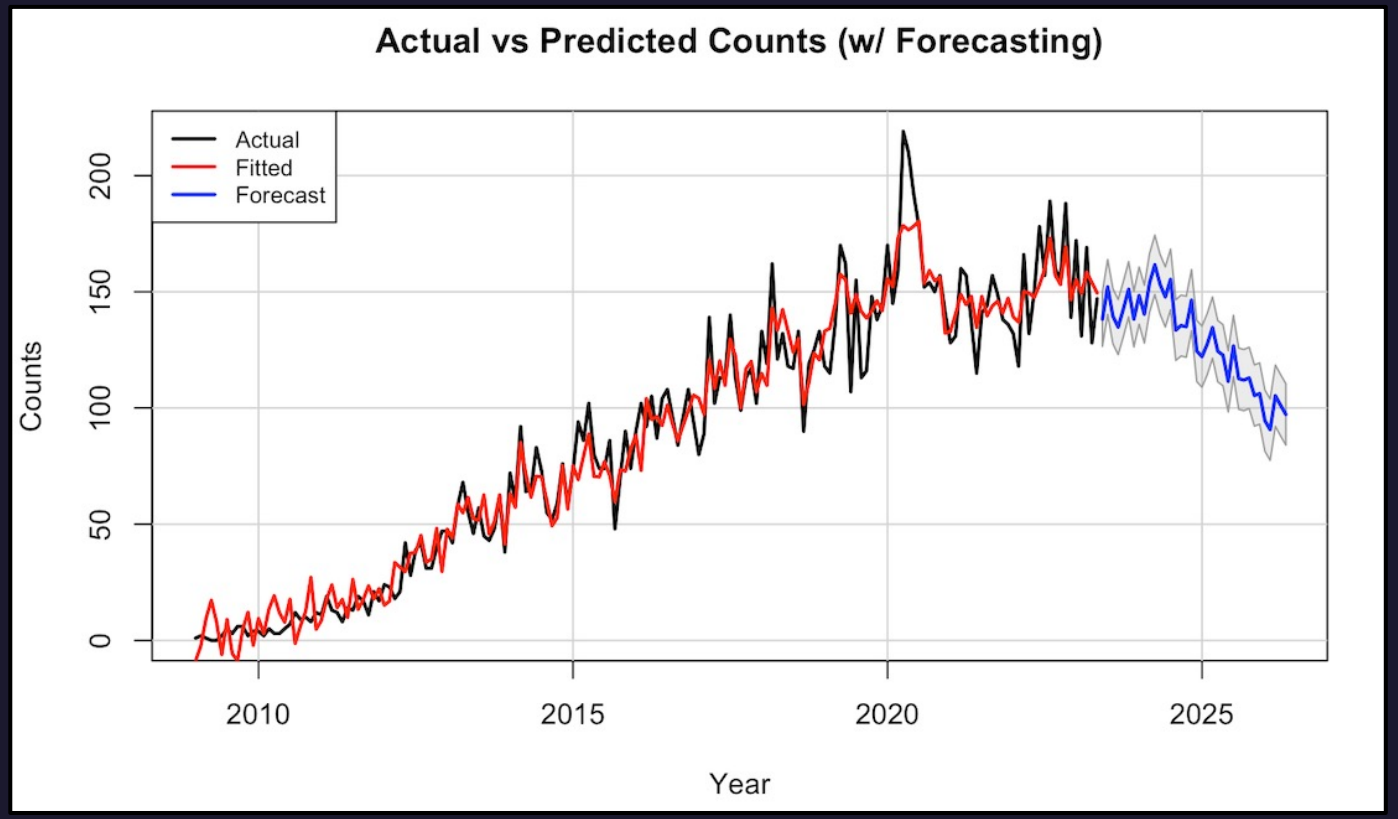
Predictions & Forecasting

- Generated predictions from our ARMA(3, 8) model for both the training data and the testing data, with trend and cycles added back.
- Predictions for the test set used an 18-step *long-term forecast horizon*.
- Predictions fit the data well, with slightly larger error over the testing range.
- Training RMSE = 10.34 counts
- Testing RMSE = 23.25 counts
- COVID-19 pandemic may have introduced an unanticipated increase in time-series interest.



Predictions & Forecasting

- Retrained model using all available data to generate forecasts extending up to May 2026.
- Retrained with testing data to maximize data utilization and ensure the model's up-to-dateness prior to forecasting.
- Another long-term, 36-step, forecast horizon.
- Our 3-year forecast indicates that the number of questions tagged as "time-series" will plateau within the next year and will gradually decline until May 2026.



Conclusion

- In summary, we collected the data from the Stack Exchange Data Explorer, removed trend, removed the **three** observed cycles, then found the optimal ARMA model using the minimum AIC.
- We assessed our fitted model's performance using a variety of plots, diagnostics, tests, and performance metrics. The RMSE scores reported for the training and testing data were 10.34 counts and 23.25 counts, respectively.
- These scores, coupled with the generated figures discussed earlier, indicated an excellent fit of our model and reasonable predictive capability for future counts of "time-series" tagged questions.

Future Work

1. Stack Exchange has only been around since 2008, limiting available data to just n = 173:
   - ❖ Future studies should consider exploring additional data sources with a longer time span or higher frequency of collection (e.g., daily or weekly).
2. COVID-19 clearly impacted the monthly counts of “time-series” tagged questions:
   - ❖ It would be interesting to conduct a similar analysis specifically for the period after 2020 in order to compare it with our existing model and to assess the impact COVID-19 had on the interest in “time-series”.
3. The slight downward trend observed in our forecasts may be attributed to the polynomial fit of the trend in our data:
   - ❖ Polynomial fits typically struggle with extrapolation, as they tend to diverge rapidly when extending beyond the observed time range.
   - ❖ Future work should consider alternative methods for trend fitting to mitigate this limitation.