UNIVERSITY OF CALIFORNIA

Los Angeles

Time Series Analysis and Forecasting of Monthly Coffeemaker Search Interest

A dissertation submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics and Data Science

by

William S Wang

ABSTRACT OF THE DISSERTATION

Time Series Analysis and Forecasting of Monthly Coffeemaker Search Interest

by

William S Wang
Master of Applied Statistics and Data Science
University of California, Los Angeles, 2024
Professor Frederic R. Paik Schoenberg, Chair

This study investigated coffeemaker search interest in the United States using the monthly time series data from Google Trends. The forecasting model developed can be utilized as a part of the coffeemaker market research since accurately forecasting user interest would enable whoever is intrigued to anticipate future developments and make informed decisions. To analyze the underlying pattern, the data was decomposed with STL into seasonal, trend, and residual components. We observed a consistent annual seasonality with a surge in interest every November and December. This pattern was attributed to the increase in user interest during the end-of-the-year holiday season sales. Anomaly detection using the STL residuals found two anomalies. The anomaly witnessed in December 2020 is best understood as the result of the demand surge during the holiday season compounded by the adoption of online shopping imposed by the COVID-19 lockdown. For the model selection process, ACF and PACF plots were used to make the initial judgments on the parameters of the time series model. The first round of model selection tested potential AR and MA orders. The second round of model selection tested potential seasonal AR and MA orders. SARIMA(0, 1, 2)×(1, 0, 1)$_{12}$ is the final model, chosen based on AIC and BIC scores. This model was able to capture the annual seasonal pattern and meet the stationary assumption with first-order differencing. The model has a MAPE of 4.3% and a RMSE of 3.841 with the rolling forecast origin prediction on the out-sample set. The residuals were confirmed to be white noise, which indicates the SARIMA model is a good fit for predicting the monthly coffeemaker search interest in the United States.

The dissertation of William S Wang is approved.

Nicolas Christou

Ying Nian Wu

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Coffee has become increasingly popular in the United States. The study by NCA's National Coffee Data Trends found that daily coffee consumption is at a 20-year high and the number of Americans who consumed coffee in the past day increased from 49% in 2004 to 67% in 2024. (Cadwalader, 2024) Another discovery is many respondents prefer specialty coffee; 57% reported drinking specialty coffee in the last year, a 7.5% increase from 2023. (Cadwalader, 2024) Speciality coffees are high-quality roasted coffee beans that score 80 or more in tasting on a point scale of 1-100 by the Specialty Coffee Association (SCA). Most of the time, specialty coffee is also associated with more sophisticated methods of brewing with an espresso machine or pour-over setup. Due to the increase in popularity of specialty coffee, many people are opting to brew their own at home using coffeemakers for better customization, comfort, and cost savings in the long run. According to (wor, 2023), "World Coffee Portal data estimates the global domestic coffee machine market for pod, filter and espresso units is worth $6.7bn. Today, the fast-developing home coffee market is redefining how roasters, equipment manufacturers and operators do business – and making high-end coffee more accessible than ever before". For people who want to dive into home brewing, many hours can be spent researching various coffeemakers by comparing prices, ease of use, and brew quality on the web. Given Google's dominance as the primary search engine in the United States, we can have a good understanding of the general public's interest with Google Trends data.

The first goal of the study is to analyze the monthly coffeemaker search interest using Google Trends data. We want to understand what are the most likely causes of the trend, seasonality, and anomaly events in the dataset. Such insights can be useful for coffeemaker companies aiming to align their strategies with consumer behavior. By understanding the trends and preferences of their targeted customers, companies can tailor their products, advertisements, and sales to match the demands of the customers. The second goal of the study is to develop an accurate time series forecasting model that can capture most of the patterns in the monthly coffeemaker search interest in the United States. The quality of the model would be examined with different metrics of prediction errors (overall forecast accuracy) and residual white noise tests (whether the model captures most of the pattern). Some common pitfalls of time series modeling would be addressed by meeting the said model's statistical assumptions and comparing different models' AIC and BIC scores for overfitting prevention.

CHAPTER 2

Methodology

2.1 Data Overview

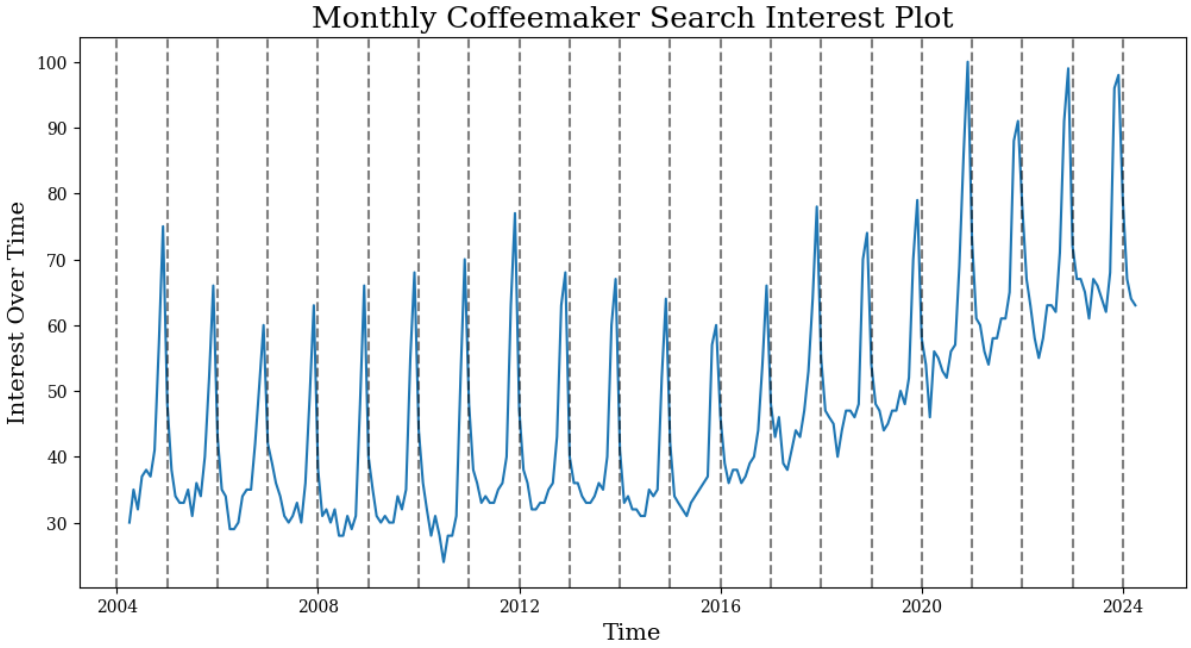


Figure 2.1: Monthly coffeemaker search interest from April 2004 to April 2024

The data for this study was obtained via Google Trends. The three datasets chosen are the topics of coffeemaker (cof, 2024b), coffee (cof, 2024a), and holiday (hol, 2024). In Google Trends, a topic is a broad category represent a group of related search terms that share similar meanings. The datasets are in the format of time series, including all categories' web searches from the United States. The time series data have monthly frequency ranging from April 2004 to April 2024 with a total of 241 data points. This time frame includes most of the available data on Google Trends at this time of the study. It also matches the time frame of the study by NCA's National Coffee Data Trend mentioned in the Introduction. The values in the datasets are the monthly interest over time. According to Google Trends, interest over time is, "Numbers represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term". The coffeemaker dataset is the main focus of this study as shown in Figure 2.1. The coffee and holiday datasets were for the data analysis on the trend and seasonality of the coffeemaker dataset. Simple data cleaning was performed to transform the X column (month) into date time format, set as the index, and changed to monthly frequency. Summary statistics of the count, mean, and standard deviation, and quantiles were calculated for data analysis.

2.2 Data Analysis

The components that make up a time series are trend, seasonality, cyclicality, and noise. (Box et al., 2015) Trend is the long-term movement in the data, it can be increasing, decreasing, or staying constant over time. Seasonality is the periodic fluctuations that occur at fixed intervals in the data. Common time series intervals are yearly, quarterly, and monthly. Cyclicality is the repeating pattern not fixed to a time interval. Noise are random fluctuations or irregular movement in the data that are not part of trend, seasonality, or cyclicality. Since time series can be separated into these components, we analyzed the data closer by decomposing a time series into the trend, seasonality, cyclicality, and noise components. The trend and cyclicality components were merged into a singular trend-cycle component, referred to trend for simplicity in this study. One popular method is STL decomposition. According to (Hyndman and Athanasopoulos, 2018), "STL is a versatile and robust method for decomposing time series. STL is an acronym for 'Seasonal and Trend decomposition using Loess', while Loess is a method for estimating nonlinear relationships. The STL method was developed by R. B. Cleveland, Cleveland, McRae, and Terpenning (1990)". The STL decomposition was calculated with

STL function from the “statsmodels.tsa.seasonal” library in the programming language Python, utilizing the default loess smoothing level of 0.25. The additive decomposition was used over multiplicative decomposition since in Figure 2.1 we can see the seasonal pattern remains constant regardless of the magnitude of the time series values. The additive decomposition equation is shown in Equation 2.1 and the residual displayed from STL decomposition is the difference between the actual data and the sum of the trend and seasonal components.

$$y_t = S_t + T_t + R_t \tag{2.1}$$

Then anomaly detection was performed to identify unusual data points deviating from the usual patterns. Identifying anomalies can highlight errors or events of interest in the datasets. Equations 2.2 and 2.3 were used to calculate the upper and lower bounds of the anomaly detection. We utilized the mean and standard deviation of the residual values calculated from STL decomposition in the equations. Any values outside the bounds were considered anomaly points.

$$\text{Lower Bound} = R_t - 3 \times \sigma_{Rt} \tag{2.2}$$

$$\text{Upper Bound} = R_t + 3 \times \sigma_{Rt} \tag{2.3}$$

STL is not the only way to decompose time series, by using spectral analysis, we can decompose the time series into constituent frequencies to understand the underlying patterns within the signals. Frequency refers to the rate at which a periodic or oscillatory phenomenon repeats over time. In the programming language R, the “spec.ar” function from the “astsa” library uses the autoregressive (AR) method to output the spectral density plot of an AR model. According to (pen), “This method is supported by a theorem which says that the spectral density of any time series process can be approximated by the spectral density of an AR model (of some order, possibly a high one)”. The X-axis of the plot represents the frequency spectrum, which indicates the range of frequencies the spectral density computed. The Y-axis of the plot represents the spectral density, which measures the distribution of power across different frequencies. The peaks in the plot indicate frequencies where the time series data exhibits significant variation. Time series data need to be de-trend prior to spectral analysis. A trend will cause such a dominant spectral density at a low frequency that other peaks would not be seen. (pen) The first-order differenced data (de-trended) was used instead of the raw data for the spectral analysis in R with a AR(15) model. If there is a consistent periodic pattern in the plot, the time series exhibits periodic phenomenon. On the other hand, the time series has an oscillatory phenomenon when we do not see the exact same periodic pattern in the plot. For example, a decline in peak frequency over time. After we identify the peaks, we can calculate the periodic cycle by using Equation 2.4, which is 1 divided by the peak frequency.

$$\text{Period} = \frac{1}{\text{Frequency}} \tag{2.4}$$

2.3 Model Assumptions

Stationarity is a key assumption for time series models. Time series is called stationary when the mean and variance of the time series are time invariant; the covariance of the time series is also time variant, but can be depended upon the lag length. (Mushtaq, 2011) If the above conditions do not hold, the series is non-stationary. (Mushtaq, 2011) Since all the time series models used in this study assume the data is stationary, non-stationary data makes the results biased. We can use the augmented Dickey Fuller test (Dickey and Fuller, 1979) to check the stationarity of the time series. The “adfuller” function from “statsmodels.tsa.stattools” library in Python was used. The null hypothesis of the test is the presence of a unit root (not stationary) and the alternative hypothesis is the absence of a unit root (stationary). Non-stationarity can be mitigated with differencing, see Equation 2.5. Differencing is a common technique used in time series analysis to transform a non-stationary time series into a stationary one by computing the difference between each observation and its preceding observation. The main drawback of differencing is that by transforming data into changes in values, original temporal relationships and context are obscured. So differencing should only be used when necessary. Stationary time series exhibits a similar bell shape as normal distribution. However, getting statistically significant results from the augmented Dickey Fuller test doesn't necessarily mean the absence of seasonality.

$$\nabla X_t = X_t - X_{t-1} \tag{2.5}$$

2.4 Model Selection

The most basic time series models are autoregressive (AR) and moving average (MA) models. AR model forecasts future values using a linear combination of past values of the variable. (Hyndman and Athanasopoulos, 2018) MA model forecasts future values using past forecasted error of the variable. (Hyndman and Athanasopoulos, 2018) They are the main components of more advanced time series models like autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA) models. An ARIMA model combines the AR and MA models while allowing differencing to make the time series stationary, see Equation 2.6. A SARIMA is an ARIMA model with additional seasonality terms, see Equation 2.7. $By_t = y_{t-1}$ represents the backshift operator or the lag of the time series. $(1 - B)^d y_t$ $By_t = y_{t-1}$ represent a dth-order differencing for the ARIMA and SARIMA models. An ARIMA model is written as ARIMA(p, d, q). p represent the AR order, d represent the differencing order, and q represent the MA order. A SARIMA model is written as SARIMA(p, d, q)×(P, D, Q)$_m$, where the additional parameters represents the seasonality part of the model. P represent the seasoanl AR order, D represent the seasonal differencing order, Q represent the seasonal MA order, and $_m$ represents the frequency of the time series. Based on the data analysis of the coffeemaker search interest time series, the SARIMA model is appropriate here since we can capture the seasonality in the time series and have the flexibility for different AR and MA terms, while making the data stationary with first-order differencing.

$$\left(1-\phi_1 B-\cdots-\phi_p B^p\right)(1-B)^d y_t=c+\left(1+\theta_1 B+\cdots+\theta_q B^q\right) \varepsilon_t \tag{2.6}$$

$$\left(1-\phi_1 B\right)\left(1-\Phi_1 B^m\right)(1-B)\left(1-B^m\right) y_t=\left(1+\theta_1 B\right)\left(1+\Theta_1 B^m\right) \varepsilon_t \tag{2.7}$$

We can start the model selection process with the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. The ACF measures all the correlations between a time series and its lagged values. The PACF only measures the direct correlation between a series and its lagged values without factoring in the intermediate lags. Since ACF and PACF assume stationary data, we need to make sure the data does not have a unit root, which can be identified with the augmented Dickey Fuller test. Lags that are outside of the statistically significant autocorrelation upper and lower bounds are considered significant. If we see significant lags at the same time interval, this tells us that there is seasonality in the time series and a SARIMA model might be appropriate. In general, by looking at the significant lags of ACF and PACF plots, we can make some initial judgments on which AR and MA orders to use for the model. Overfitting can happen in time series modeling when there are too many parameters that capture the noise instead of the underlying pattern. According to (Bianco, 2016), "AR and MA terms can come up as significant (low p-values) however these terms will be competing with other parts of the model when over-fitting occurs". AIC and BIC are effective for model selection and over-fitting prevention due to their ability to balance model complexity and goodness of fit by penalizing excessive parameters. AIC and BIC were calculated with Equations 2.8 and 2.9.

$$\mathrm{AIC}=2k-2L \tag{2.8}$$

$$\text{BIC} = \ln(n)k - 2\ln(L) \tag{2.9}$$

k represents the number of parameters and L represents the log-likelihood. The model would be rewarded with a high L (better model fit) and penalized with a high k (additional parameters). In general, a lower AIC and BIC indicate a better model. The AR and MA orders of the time series model were decided by comparing the AIC and BIC scores of multiple fitted models. Two rounds of model selection were performed to select the best-fitted SARIMA model. First round compared different AR and MA orders while keeping the seasonal component constant. The model with the lowest AIC and BIC was selected. The second round of model selection used the winning model from the first round, but now testing different seasonality parameters. Again, the model with the lowest AIC and BIC was selected.

2.5 Model Prediction

The data was split into 80% for the training set (in-sample) and 20% for the testing set (out-sample). The best practice is to use values at the end of the time series dataset for testing and the rest for training. This is because time series data has autocorrelation between consecutive data points. Static data have observations independent of each other, so creating the split with randomized sampling and order does not cause problems. The training set of monthly coffeemaker search interest is from April 2004 to April 2020 (16 years of data) and the testing set is from April 2020 to April 2024 (4 years of data). The in-sample data was used to train the time series model and the out-sample data was used to test the performance of it. The SARIMA time series model was fitted on the training set with one-step forecasts and the parameters decided during the model selection process. The SARIMAX model function from the "statsmodels.tsa.statespace.sarimax " library in Python was used for computation. According to (Hyndman and Athanasopoulos, 2018), "Typically, we compute one-step forecasts on the training data (the 'fitted values') and multi-step forecasts on the test data. However, occasionally we may wish to compute multi-step forecasts on the training data, or one-step forecasts on the test data". In this study, both multi-step forecast and one-step forecast were computed for comparison. There was a concern that the 4 years multi-step forecast horizon is too long, so the model's predictions are likely to be inaccurate because the ARIMA-type models are usually not ideal for predicting over a long forecast horizon. This problem can be remedied with the one-step forecast (rolling forecast origin) of the testing data. Rolling forecast origin is an evaluation technique according to which the forecasting origin is updated successively and the forecasts are produced from each origin (Tashman, 2000). In this study, a one-steps-ahead forecast was used, which means we only forecast one month out at a time and updated the training set after each forecast. The forecasting accuracy was measured with mean absolute percent error (MAPE), see Equation 2.10 and root mean squared error (RMSE), see Equation 2.11. A_i is the actual data and F_i is the prediction. MAPE penalizes over-prediction more and under-prediction less. RMSE provides easy-to-interpret results and penalizes larger errors more heavily.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{A_i} \times 100 \tag{2.10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - F_i)^2} \tag{2.11}$$

If the residuals of the time series is white noise (shows no autocorrelation in the error), the time series model would be considered a good fit and captures all the patterns in the data. According to (Hyndman and Athanasopoulos, 2018), "The prediction intervals for ARIMA models are based on assumptions that the residuals are uncorrelated and normally distributed. If either of these assumptions does not hold, then the prediction intervals may be incorrect. For this reason, always plot the ACF and histogram of the residuals to check the assumptions before producing prediction intervals". This is why ACF and PACF plots of the residuals were used to check for autocorrelation. Absence of autocorrelation means no significant lags and underlying patterns in the plots. We also visualized the residuals by plotting testing model errors against time. If the residuals are white noise, the values would all be close to the mean of 0. Finally, the Ljung-Box test (Ljung and Box, 1978) was used as a more robust way to check if there were any autocorrelation in the residuals. The null hypothesis is the absence of autocorrelation and the alternative hypothesis is the presence of autocorrelation. The “acorr_ljungbox” function from “statsmodels.stats.diagnostic” library in Python was used.

CHAPTER 3

Results

3.1 Data Analysis Results

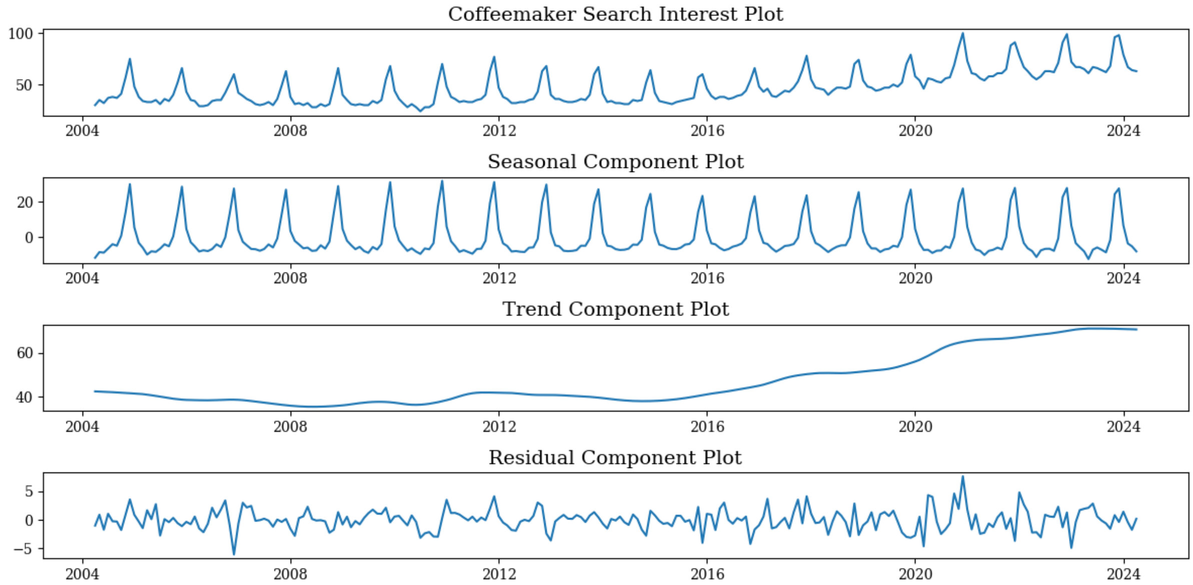


Figure 3.1: Trend, seasonal, and residual components of coffeemaker search interest

	Summary Statistics
Count	241.000
Mean	46.900
Standard Deviation	16.104
Minimum	24.000
25% Quantile	34.000
50% Qauntile	42.000
75% Qauntile	58.000
Maximum	100.000

Table 3.1: Summary statistics of coffeemaker search interest

Figure 3.1 displays the seasonal, trend, and residual components of the coffeemaker search interest calculated from STL decomposition. We can see a recent upward trend in coffeemaker search interest and a consistent annual seasonality. The results from STL decomposition were utilized for anomaly detection. We started by calculating the estimated interest over time by combining the trend and seasonal components of the time series. Then we visualized the difference between the actual vs the estimated interest over time in Figure 3.2. The consistency between the estimated in orange and the actual in blue suggests a close alignment, indicating minimal presence of anomalies that diverge from the established trends or seasonal patterns. In Figure 3.3, we employed a more robust method of identifying anomaly points by establishing upper and lower bounds on the STL residuals, computed from Equations 2.2 and 2.3. Any data points falling outside these bounds was flagged as anomalies. In Table 3.2, the two anomalies are both in the month of December: lower-than-expected interest in 2006 and higher-than-expected interest in 2020. It is hard to say what caused the anomaly in December 2006, but there is a clear reason for the spike in December 2020 search interest. The COVID-19 pandemic is widely considered a black swan event that disrupted and dramatically changed everyone's daily lives. Since the lockdown started in 2020, many people around the world have chosen to pick up home brewing with specialty coffee as a hobby due to a lack of access to commercial coffee shops and abundant free time. According to (con, 2021), "The growth of at-home gourmet coffee preparation existed before 2020, although it was accelerated by the pandemic". It is possible the onset of the COVID-19 lockdown during December 2020 magnified the annual surge in coffeemaker search interest typically observed during the end-of-the-year holiday season.

	Interest Over Time
2006-12	60
2020-12	100

Table 3.2: Anomaly interest over time

	Holiday Search Interest	Coffee Search Interest
Correlation	0.608	0.781

Table 3.3: Correlation of coffeemaker search interest with holiday and coffee search interest

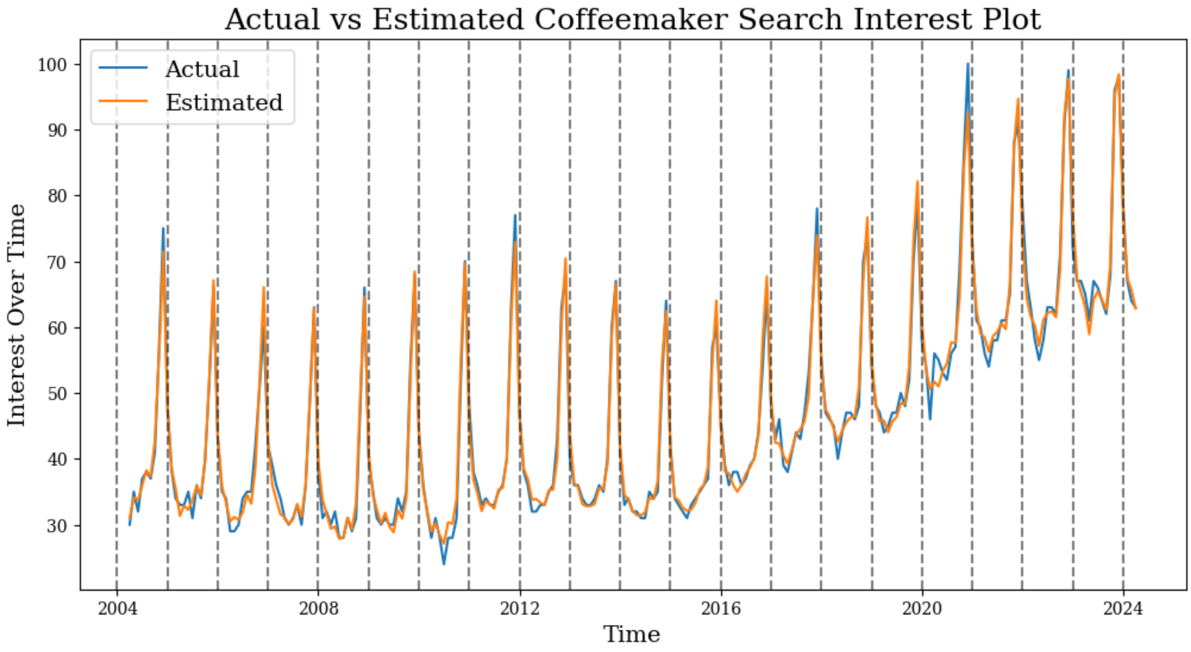


Figure 3.2: Actual vs estimated coffeemaker search interest

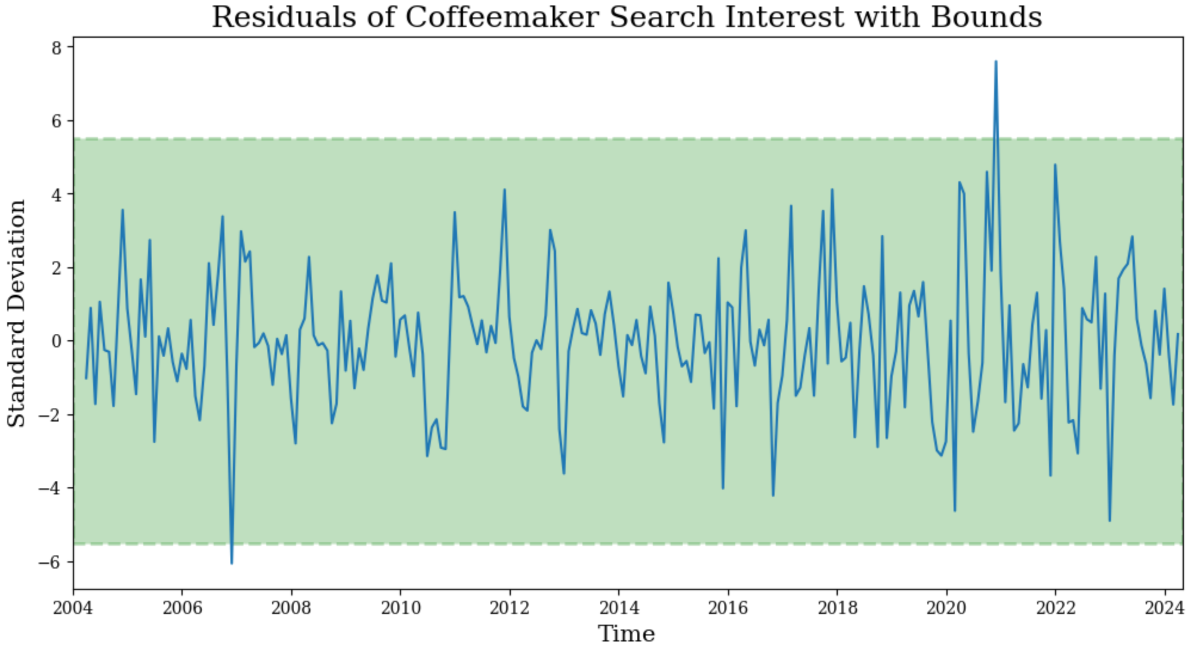


Figure 3.3: Anomaly detection of coffeemaker search interest

After an overview of the components of time series through STL decomposition and anomaly detection, further analysis was done on the seasonal pattern in the time series. Figure 3.4 shows the means of the interest over time of each month. This subseries

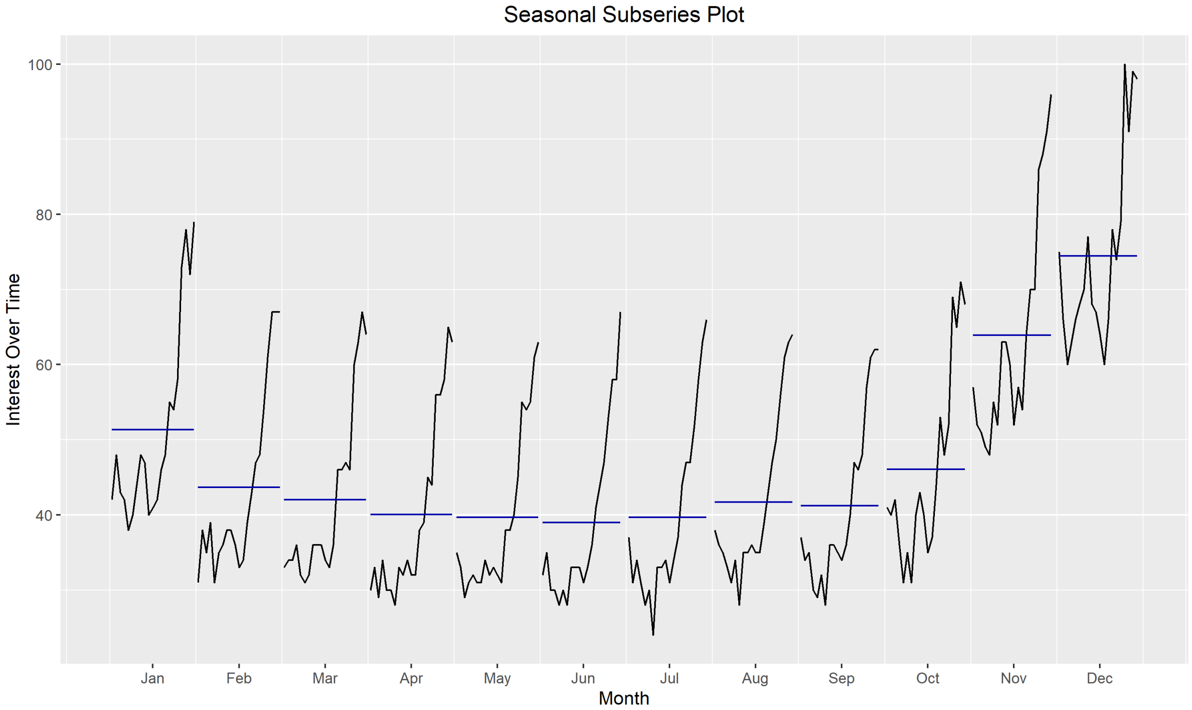


Figure 3.4: Monthly seasonal subseries plot

plot allows us to clearly see the seasonality in the data. Based on the plot, the mean was much higher in November and December compared to the rest of the months. This annual seasonality was also observed with certain product sales. According to the analysis of (Ensafi et al., 2022) on furniture sales, “It is not surprising that the sales of this category reached their peak in the winter holiday season. During this time of year, retailers offer big sales to start the season. In addition to that, other post-Thanksgiving sales events such as Black Friday and Cyber Monday deals are very encouraging to the customers”. Since this seasonal pattern alludes to a relationship between coffeemaker and holiday, the correlation between the search interest of coffeemaker and holidays was checked. Search interest of coffeemakers and holidays have a moderate positive correlation at 0.608, see Table 3.3, this tells us that as the search interest of holidays increases, the search interest of coffeemakers also tends to increase as well. Figure 3.5 is the plot of the coffeemaker and holiday search interest. This plot reveals a very similar seasonality between the two. Additionally, we also observed that both time series had an abnormal search interest of 100 in December 2020. One possible conclusion for this shared anomaly is that December 2020 marking the first holiday season amidst lockdown restrictions, since in-person shopping was not available, there was a surge of interest in online holiday shopping. Coffeemakers were among the products people were very interested in. Due to the moderate correlation between the search interest in coffeemakers and holidays, we observed the same anomaly point and similar seasonal patterns in both time series.

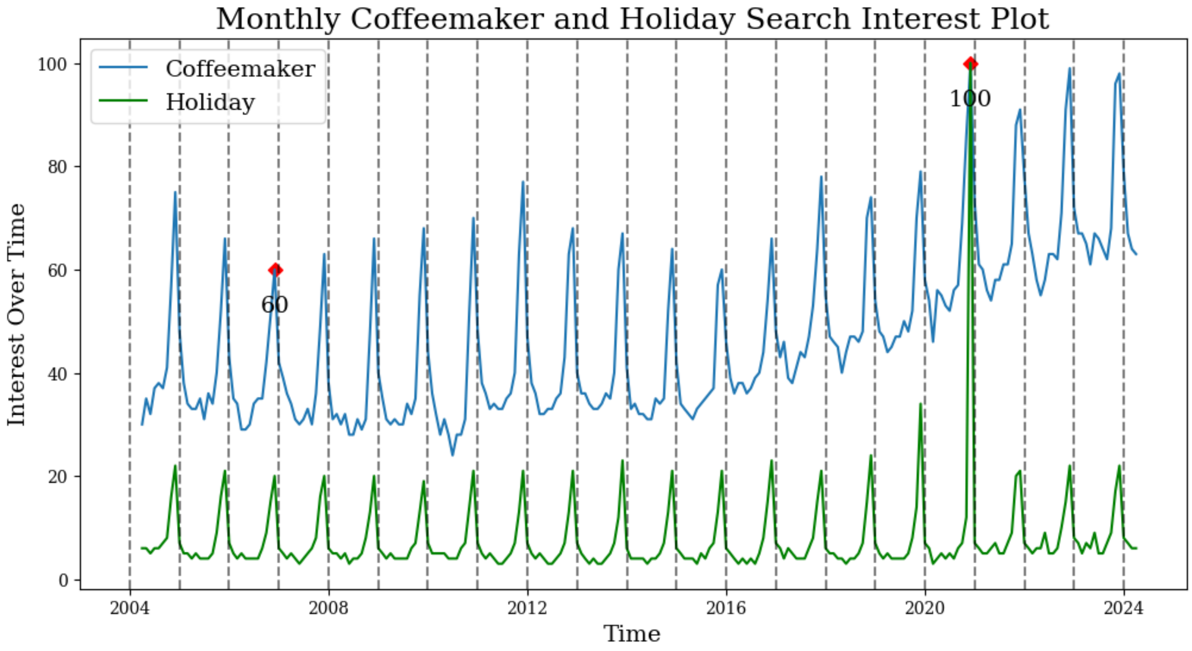


Figure 3.5: Monthly coffeemaker and holiday search interest with anomalies

We compared the trend of the coffeemaker and coffee search interest in Figure 3.6. The interest over time of coffeemakers increased from 42 to 70 from April 2004 to April 2024, which is a 66.7% increase. The interest over time coffee increased from 42 to 91, which is a 175% increase. Coffee and coffeemakers have a strong positive correlation of 0.781, see Table 3.3. This tells us that as the search interest in coffee increases, the search interest of coffeemakers also increases as well. Around 2015, we can see an upwards trend in both, this tells us that not until recently there is a big increase in user interest in coffee and coffeemakers. Around 2020, there is a much steeper upward trend, this lines up with the boom in consumer interest and demand for specialty coffee and coffeemakers starting from the COVID-19 lockdown in 2020.

The time series's periodic cycle was analyzed with spectral analysis. A fitted AR(15)

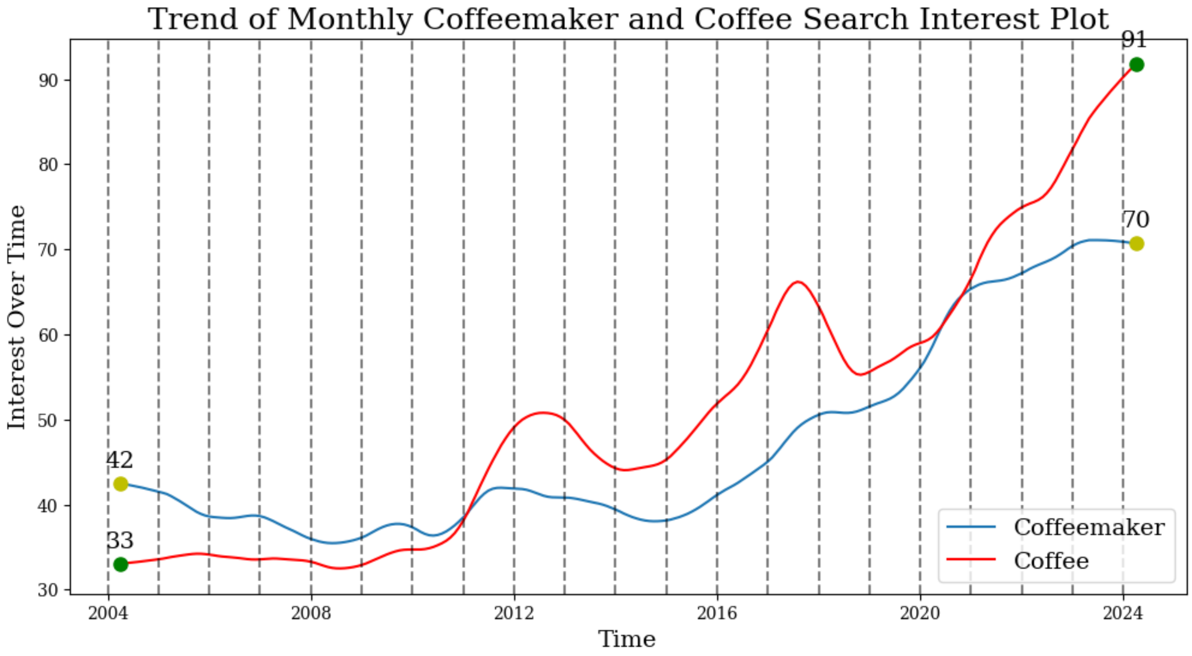


Figure 3.6: Trend of monthly coffeemaker and coffee search interest

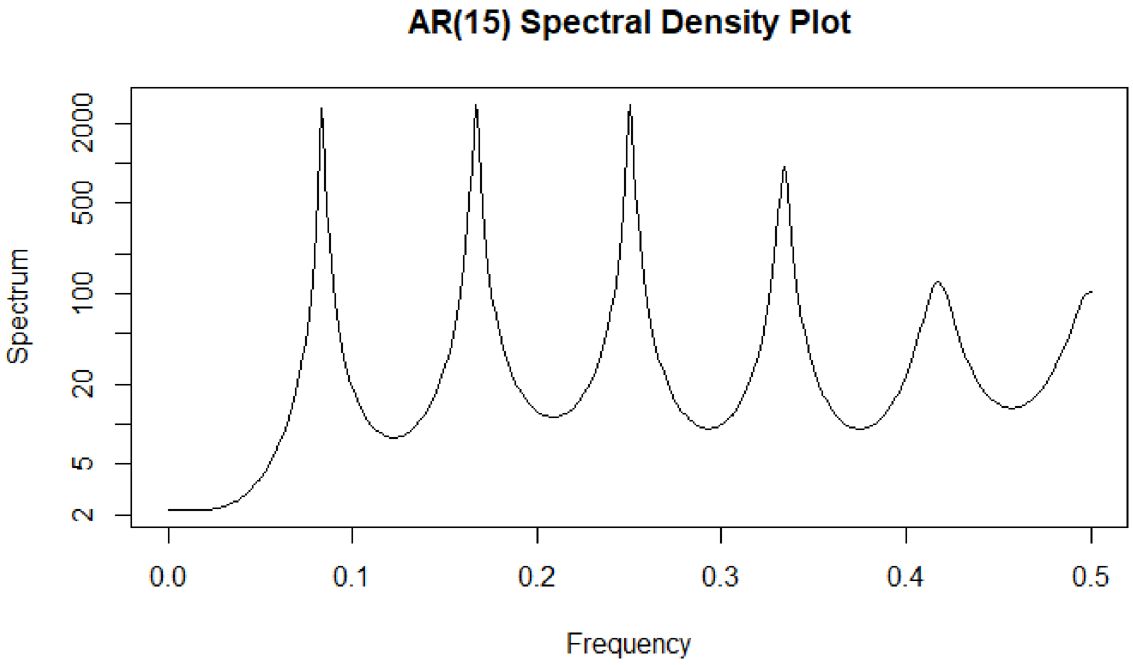


Figure 3.7: Spectral density with AR(15) model

model outputs the spectral density plot seen in Figure 3.7. Coffeemaker search interest exhibits an oscillatory phenomenon since there is a decline in peak frequency over time.

We also calculated the periodic cycle by using Equation 2.4 on the first peak of the spectral density plot. The first peak frequency is 0.084, so the cycle is about 1 / 0.084 = 11.905 (11.905 months per cycle). This matches up with the 12 monthly seasonality observed in Figure 3.1 and 3.4.

3.2 Model Assumptions Results

	No Differencing	First-order Differencing
ADF Statistics	0.941	-5.003
5% Critical Values	-2.874	-2.874
P-Value	0.994	0.000022

Table 3.4: Augmented Dickey Fuller test results

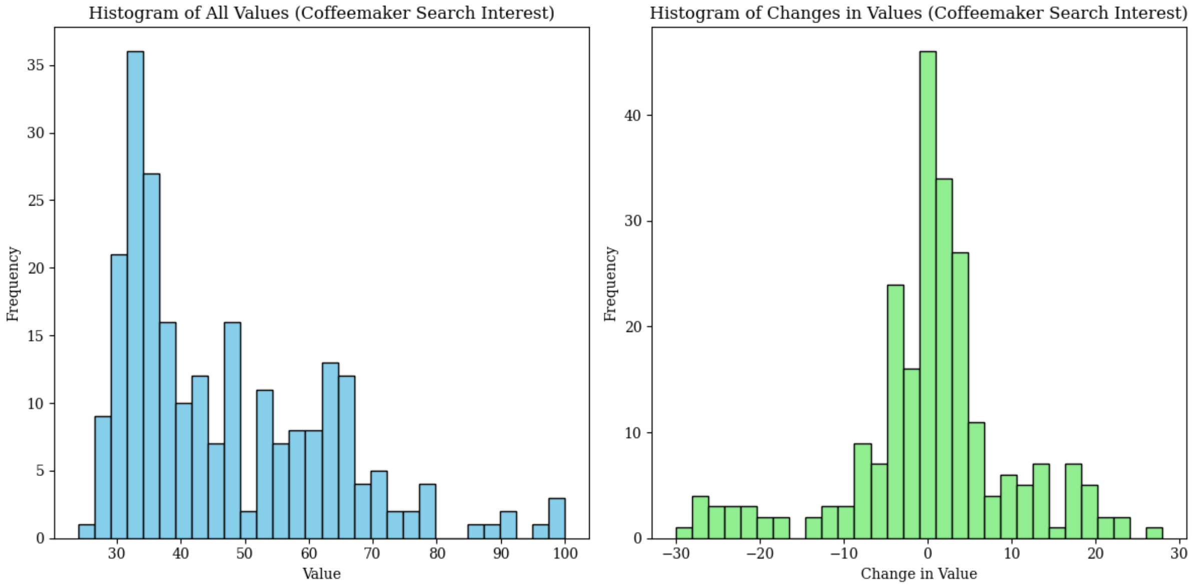


Figure 3.8: All values and change in values of coffeemaker search interest

In Figure 3.8 we can see the histogram of all values' mean is skewed to the left, with most of the values between 30 and 40. The histogram of change in values (differenced values) exhibits a much more normal distribution and most of the differenced values are around 0. The stationary of the time series data was checked with the augmented Dickey Fuller test. In Table 3.4, we can see the p-value is 0.99 before differencing, so we can not reject the null hypothesis that there is a unit root at the 5% significance level. After first-order

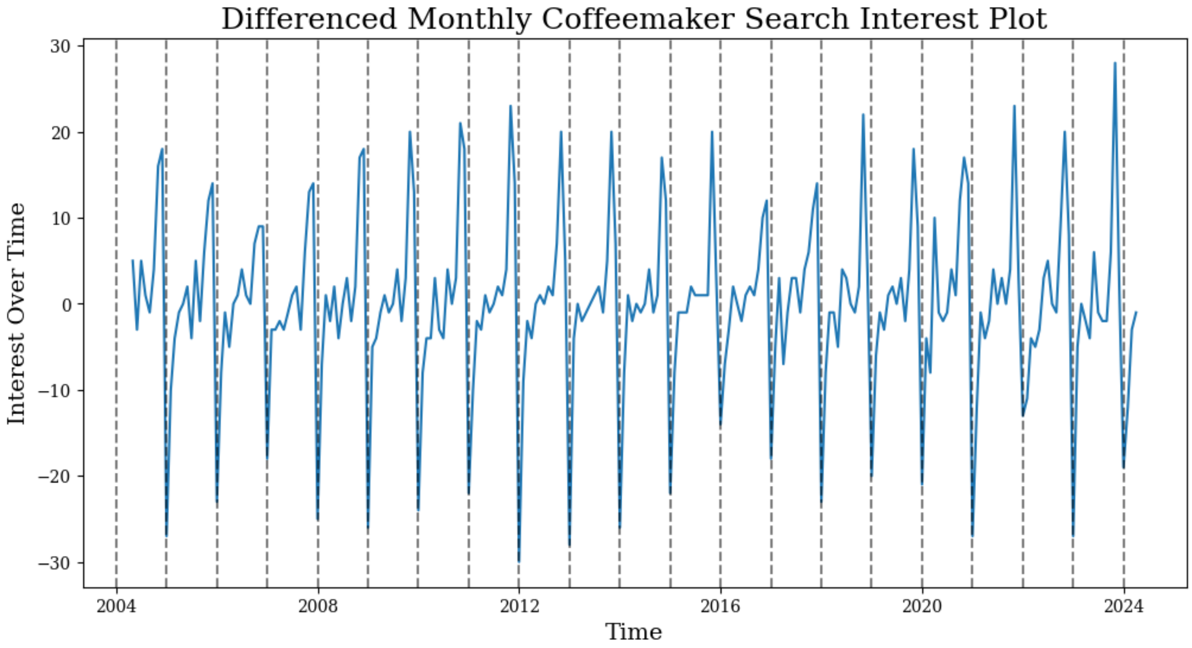


Figure 3.9: First-order differenced coffeemaker search interest

differencing in Equation 2.4, the p-value is 0.000022, so we reject the null hypothesis that there is a unit root at the 5% significance level. Since the test focuses on checking for unit root, which is the presence of a stochastic trend, seasonality can still be left in the data. As we can see in the first-order differenced plot shown in Figure 3.8, the seasonal component is still present.

3.3 Model Selections Results

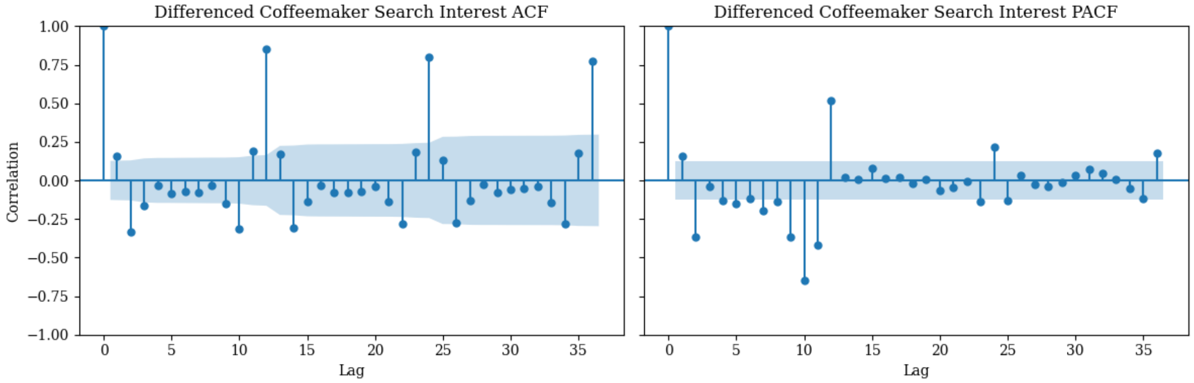


Figure 3.10: First-order differenced coffeemaker search interest ACF and PACF

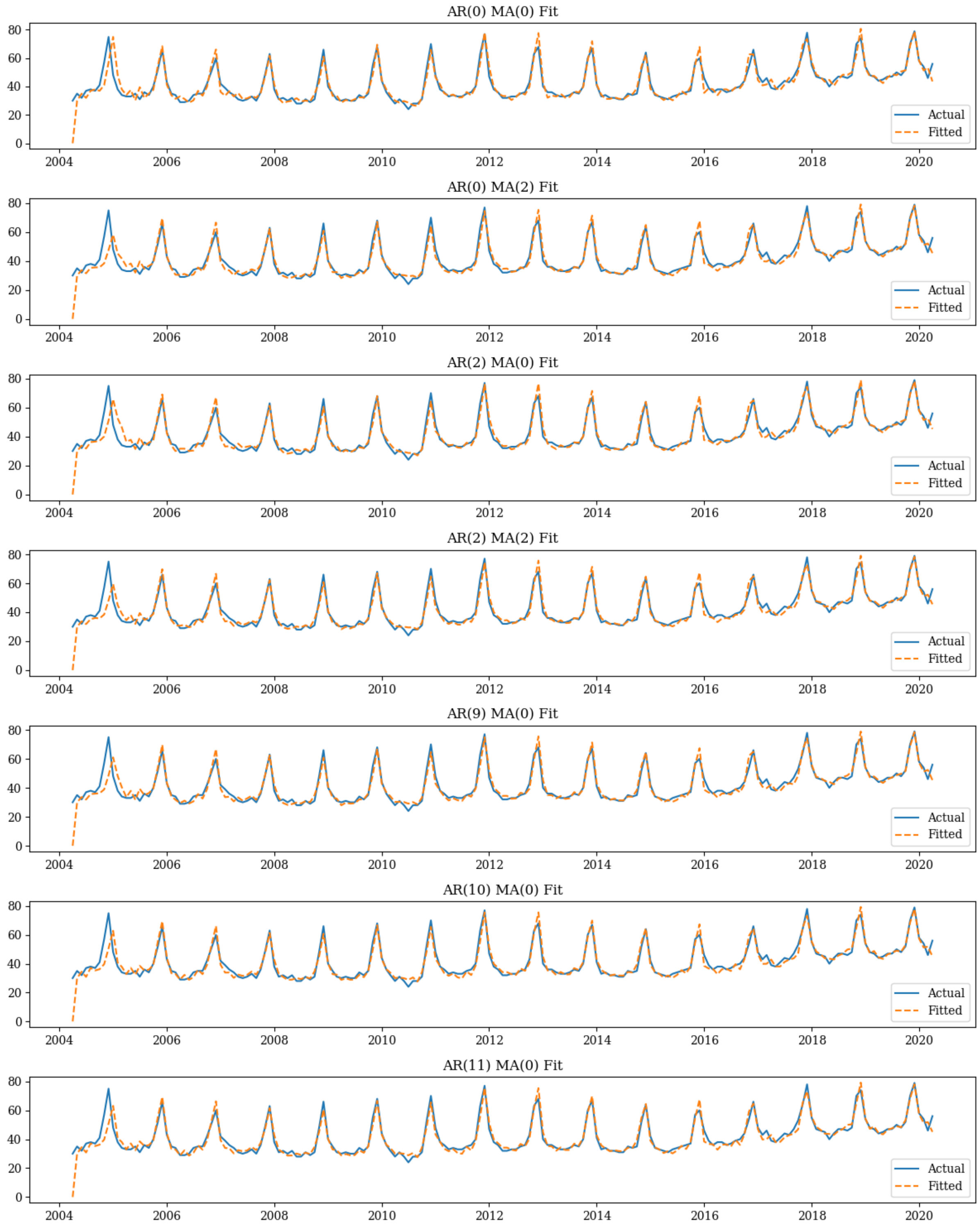
	AIC	BIC
$(0, 1, 0) \times (1, 0, 1)_{12}$	1030.941	1040.713
$(0, 1, 2) \times (1, 0, 1)_{12}$	999.622	1015.910
$(2, 1, 0) \times (1, 0, 1)_{12}$	1013.461	1029.748
$(2, 1, 2) \times (1, 0, 1)_{12}$	1001.959	1024.761
$(9, 1, 0) \times (1, 0, 1)_{12}$	1010.178	1049.268
$(10, 1, 0) \times (1, 0, 1_{12}$	1001.840	1044.188
$(11, 1, 0) \times (1, 0, 1)_{12}$	1003.606	1049.211

Table 3.5: Round 1 model selection: AR and MA parameters

	AIC	BIC
$(0, 1, 2) \times (1, 0, 1)_{12}$	999.622	1015.910
$(0, 1, 2) \times (1, 0, 0)_{12}$	1036.738	1049.768
$(2, 1, 2) \times (0, 0, 1)_{12}$	1245.594	1258.624

Table 3.6: Round 2 model selection: seasonal AR and MA parameters

The SARIMA model was selected because prior data analysis in the study shows the time series is stationary only after first-order differencing and there is a consistent annual seasonality. ACF and PACF plots were used to identify possible model fits based on the significant autocorrelations observed. Figure 3.10 shows the ACF and PACF plots with lag order 36, this configuration allow us to observe the pattern for 3 years (12×3). In Figure 3.10, we can see significant ACF lags at 2, 10, 12, and 24 and significant PACF lags at 2, 9, 10, 11, and 12. There is a clear annual seasonality since 12, 24, and 36 lags are all significant. We used AIC and BIC scores to compare model fit and performance. The final model had the lowest AIC and BIC in two rounds of model selection. In round one of model selection, 7 SARIMA models with different AR and MA orders were fitted while keeping the seasonality AR and MA order constant, see Figure 3.11. We can only see marginal improvement in the fit on the model with higher orders in Figure 3.11 and overfitting is definitely a concern when using so many terms. In Table 3.5, the MA(2) model was shown to have the lowest AIC score at 999.622 and the lowest BIC score at 1015.910. So the SARIMA$(0, 1, 2)\times(1, 0, 1)_{12}$ was selected in round one. As shown in Figure 3.12, the MA(2) model was used in round two of model selection and different combinations of seasonality order are fitted for comparison. Once again, SARIMA$(0, 1, 2)\times(1, 0, 1)_{12}$ has the lowest AIC and BIC, see Table 3.6. This tells us that this model is


AR(0) MA(0) Fit

AR(0) MA(2) Fit

AR(2) MA(0) Fit

AR(2) MA(2) Fit

AR(9) MA(0) Fit

AR(10) MA(0) Fit

AR(11) MA(0) Fit

Figure 3.11: Model fits of different AR and MA parameters

relatively simple and is a better fit for the data compared to the others. In conclusion, the final model has a seasonal AR component of order 1, no differencing, and a seasonal MA

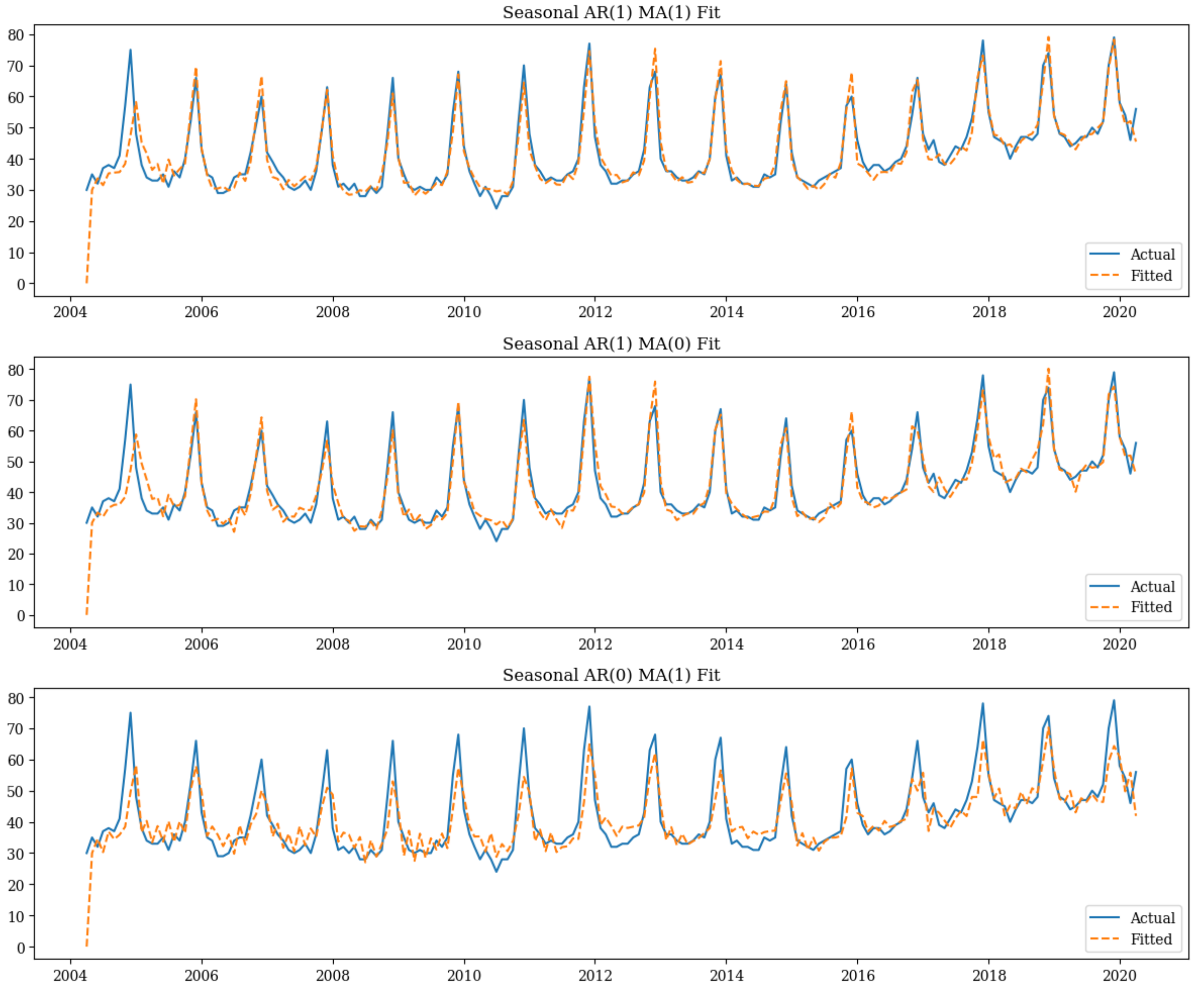


Figure 3.12: Model fits of different seasonal AR and MA parameters

component of order 1 on a seasonal frequency of 12 months. Alongside an MA component of order 1, an MA component of order 2, and first-order differencing.

3.4 Forecasting Results

	Coefficient	Standard Error	P-value	Confidence Interval
MA(1)	-0.437	0.066	0.000	(-0.567, -0.307)
MA(2)	-0.266	0.066	0.000	(-0.396, -0.137)
SAR(1)	0.996	0.002	0.000	(0.991, 1)
SMA(1)	-0.708	0.063	0.000	(-0.831, -0.586)

Table 3.7: SARIMA model fit summary

The SARIMA time series model was trained on the in-sample data from April 2004 to April 2020. We can examine the model summary in Figure 3.7. Every parameter is statistically significant, with p-values much smaller than 5% significance level. The final model is presented in Equation 3.1, it has a training RMSE of 4.482. Both MA(1) and MA(2) terms have negative coefficients, which suggests that for each unit increase in the past error term of coffeemaker search interest from one month ago, the current observation is expected to decrease by approximately 0.437 units, and for each unit increase in the past error term of coffeemaker search interest from 2 months ago, the current month is expected to decrease by approximately 0.266, all else being equal. The seasonal AR(1) term suggests a strong positive relationship, indicating that a 1 unit increase in coffeemaker search interest 12 months ago resulted in approximately a 0.996 unit increase in the interest of coffeemaker, all else being equal. The seasonal MA(1) suggests that for every unit increase in the seasonal past error term 12 months ago, the current month's observation is expected to decrease by approximately 0.708 units, all else being equal.

$$\left(1 - 0.996B^{12}\right)(1 - B)y_t = \left(1 - 0.437B - 0.266B^2\right)\left(1 - 0.708B^{12}\right)_{\epsilon_t} \tag{3.1}$$

	4 Years Forecast Error	Rolling Forecast Origin Error
MAPE	0.072	0.043
RMSE	6.630	3.841

Table 3.8: SARIMA model prediction error comparison

Multi-step forecast and single-step forecast were both used for model performance analysis. The first method is the 4 years multi-step forecast, which has the forecast horizon from April 2020 to April 2024. The second method is the single-step forecast (rolling forecast origin), which only predicts one month ahead at a time. For both the forecasting methods, the MAPE and RMSE were calculated and shown in Table 3.8. The 4 year forecast's MAPE is 7.2% and the RMSE is around 6.630. The rolling forecast's MAPE is 4.3% and the RMSE is around 3.841. The rolling prediction has much lower MAPE and RMSE.

This was to be expected since time series prediction gets worse the further you get from the last value in the dataset. According to (Hyndman and Athanasopoulos, 2018), “The forecast variance usually increases with the forecast horizon, so if we are simply averaging the absolute or squared errors from the test set, we are combining results with different variances”.

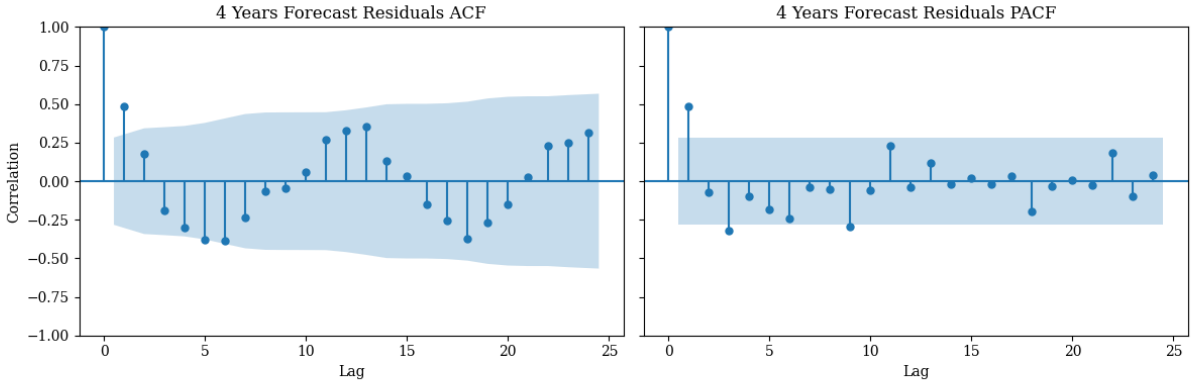


Figure 3.13: 4 years forecast residuals ACF and PACF

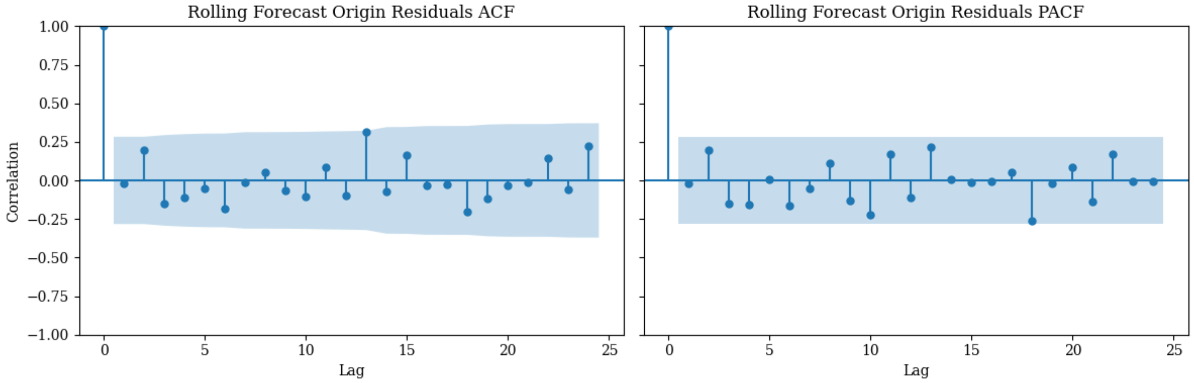


Figure 3.14: Rolling forecast origin residuals ACF and PACF

The residuals of both methods of forecasting were checked for white noise. White noise residuals are crucial since they indicate the model captured all available information (absence of autocorrelation) and made unbiased predictions. The ACF and PACF plots of the residuals gave us a visual aid on whether there is autocorrelation in the lags of the residuals. The Ljung-box test was used as a more precise way of testing for autocorrelation. If the test result is statistically significant, there is autocorrelation in the residuals. Finally, systemic bias was checked in the residual plots where we can visualize whether most of

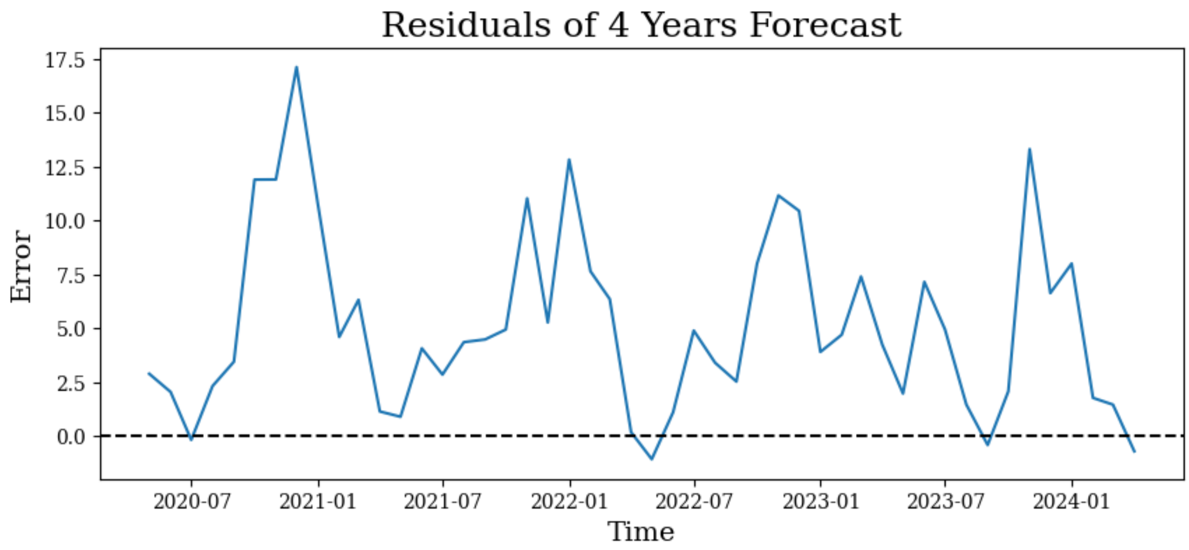


Figure 3.15: Residuals plot of 4 years forecast of out-sample coffeemaker search interest

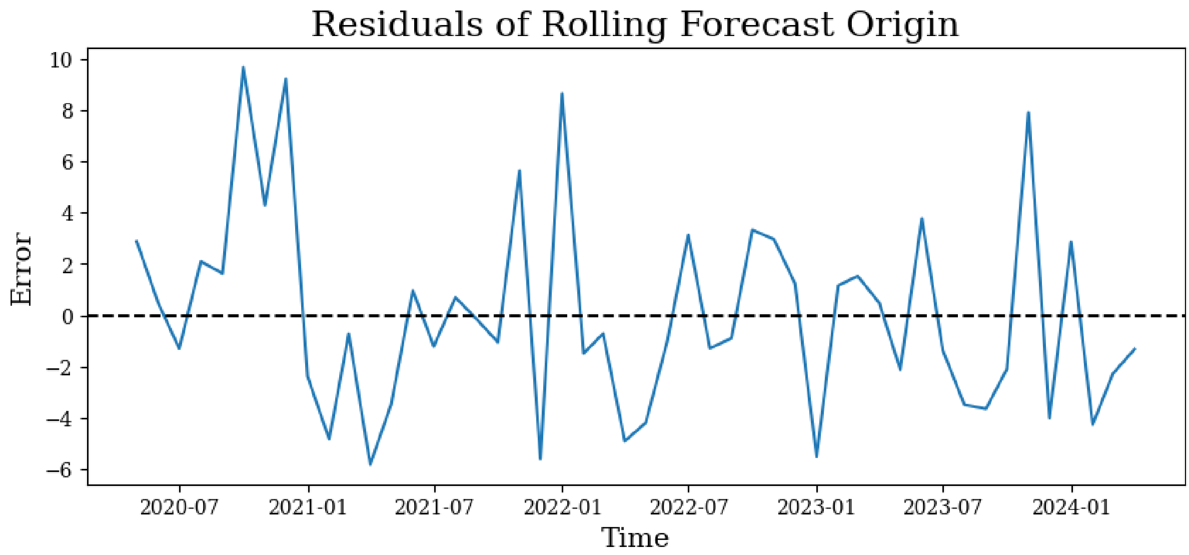


Figure 3.16: Residuals plot of rolling forecast of out-sample coffeemaker search interest

the errors fall near the mean of 0. Figure 3.13 displays the ACF and PACF of the 4 years forecast residuals. We can see there is a sine wave pattern in the ACF; lag 1 is significant in both ACF and PACF. In Table 3.9, from lag 1 to 9, all the p-values are smaller than the significance level of 5%. In Figure 3.15, we can see the residuals of the predictions are mostly above 0. These results tell us the model validated with the 4 years forecast was not capturing all the information in the data and was making systematically biased predictions. However, another possible explanation for these results is that the model

	4 Years Forecast P-value	Rolling Forecast P-value
Lag 1	0.000559	0.893
Lag 2	0.001148	0.370
Lag 3	0.001472	0.355
Lag 4	0.000424	0.414
Lag 5	0.000032	0.535
Lag 6	0.000002	0.420
Lag 7	0.000001	0.535
Lag 8	0.000003	0.624
Lag 9	0.000006	0.691

Table 3.9: Ljung-Box test on autocorrelation in residuals

with a 4 years forecasting horizon was not adapting to the changing pattern over time. This is a common pitfall of ARIMA-type models and long forecast horizons. Rolling forecast origin was used to remedy this problem. Figure 3.15 displays the ACF and PACF of the rolling forecast origin's residuals. There is no underlying pattern and no significant lags. This observation is confirmed with the results in Table 3.9, from lag 1 to 9, all the p-values are bigger than the significance level of 5%. This tells us there is no autocorrelation in the residuals. In Figure 3.16, we can see the residuals have a mean of around 0, indicating unbiased predictions. In conclusion, the fitted model has white noise residual in the rolling forecast origin, but not in the 4 years forecast. The prediction values are compared with the actual test set data in Figures 3.17 and 3.18. In general, the predictions are much closer to the actual in the rolling forecast origin compared to the 4 years forecast. We can see both models under predicted the search interest in December 2020 since they were not able to account for the COVID-19 lockdown holiday anomaly. Finally, we used the SARIMA$(0, 1, 2)\times(1, 0, 1)_{12}$ moel to forecast forward from April 2024. The forecast window is 6 months since we learned from the 4 years forecast that a long forecast window reduces the reliability of the prediction. The result is shown in Figure 3.19, where the forecasted value is in red with the prediction interval. There will be a big spike in coffeemaker search interest in November and December of 2024 and a continuous increasing upward trend for the next 6 months.

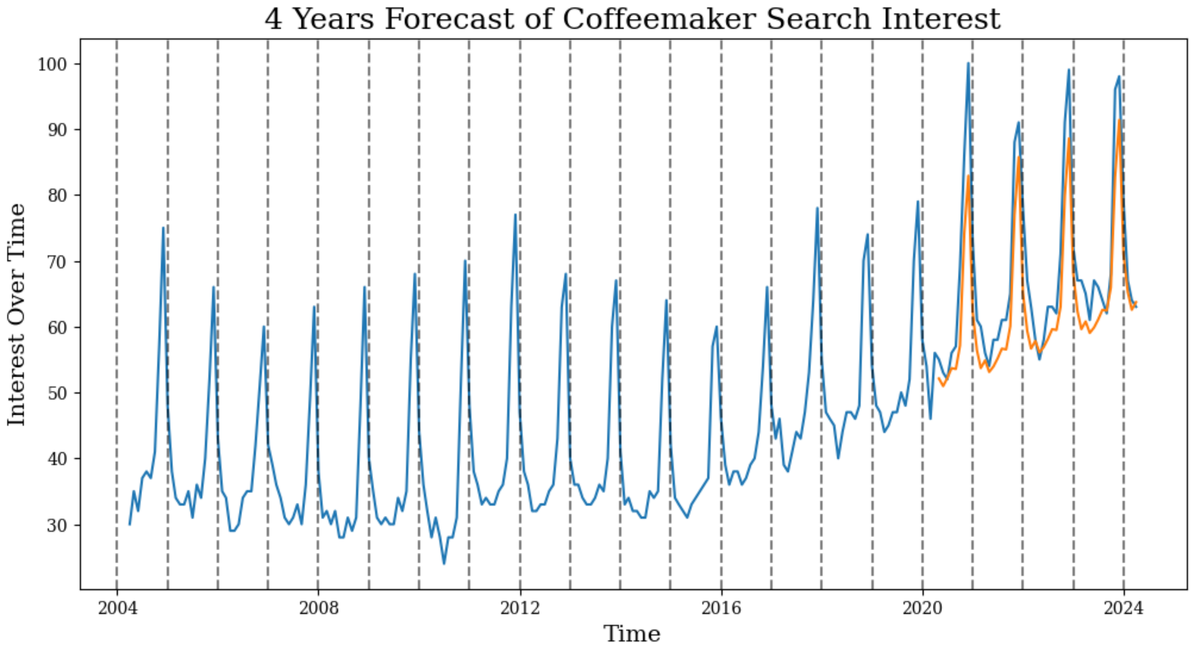


Figure 3.17: 4 years forecast of out-sample coffeemaker search interest

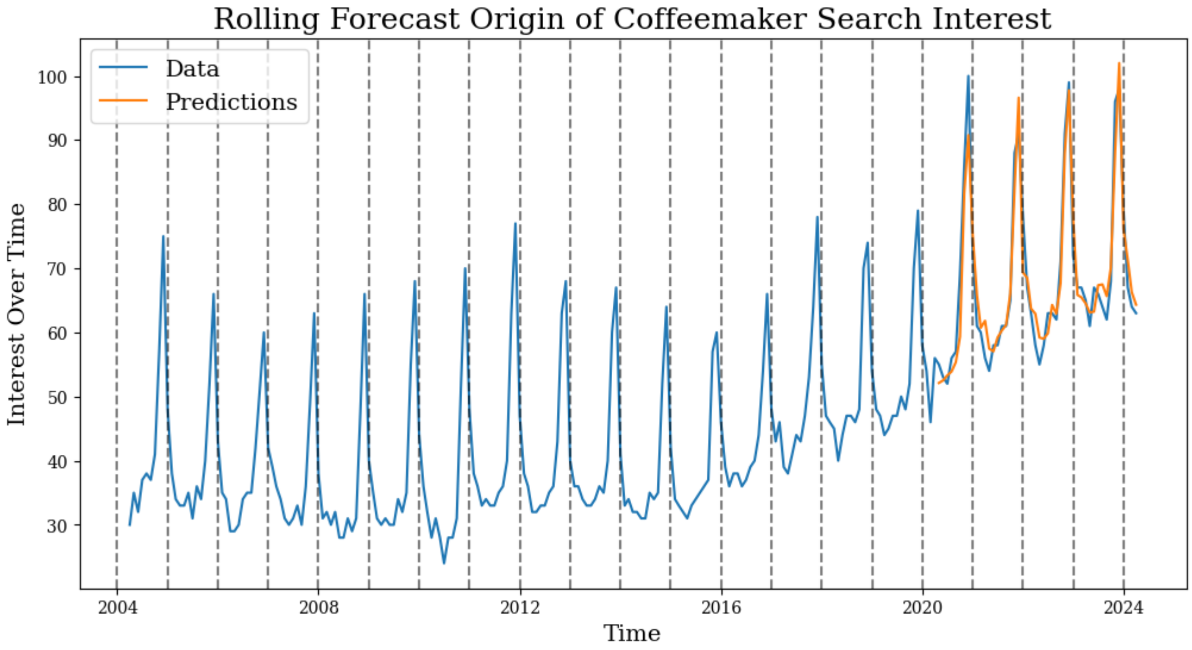


Figure 3.18: Rolling forecast of out-sample coffeemaker search interest

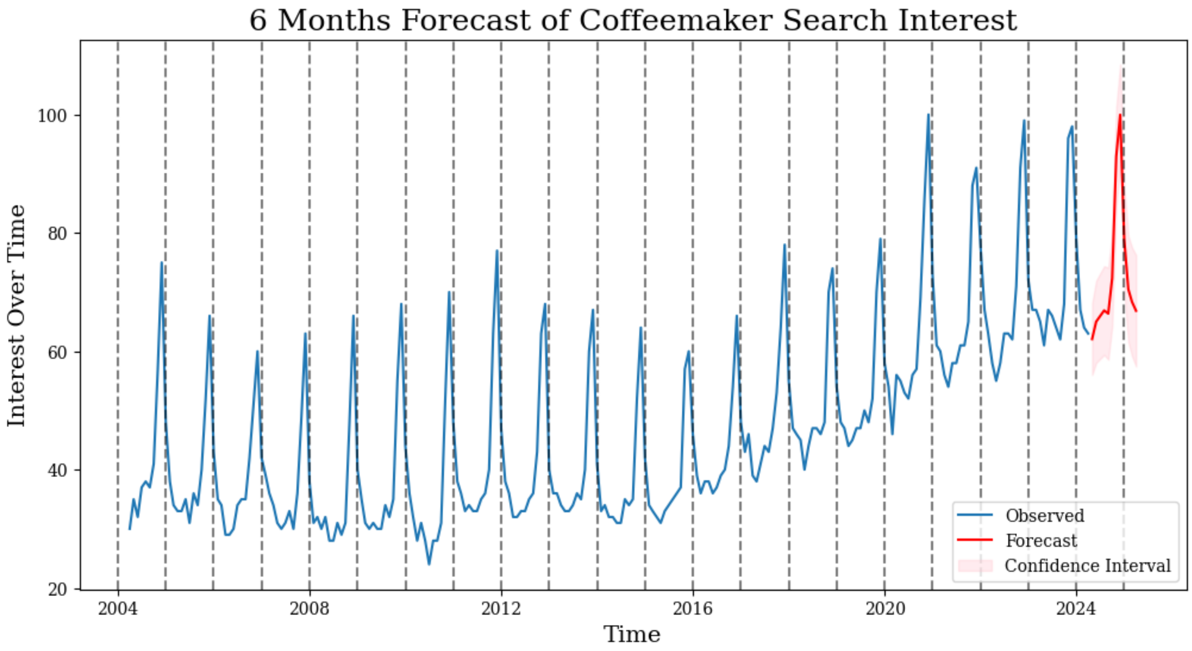


Figure 3.19: 6 months forecast of coffeemaker search interest after April 2024

CHAPTER 4

Conclusion

The data analysis segment of the study was able to find the reason behind the underlying trend and seasonal pattern of the monthly coffeemaker search interest with STL decomposition, anomaly detection, and spectral analysis. By plotting the trend component of the search interest in coffeemakers, we can clearly see a noticeably upward trend starting around 2015 and a much steeper increase around 2020. This shows how coffeemakers like pod and press machines have only gotten popular in the last 8 years with a big increase in interest during the COVID-19 lockdown due to the lack of access to local coffee shops. By analyzing the seasonal component of the data, we found the annual seasonality is the result of the increase in user interest during the holiday season due to sales from November to December. By analyzing the residuals component of the time series, two anomalies were found. The anomaly in December 2020 is due to the combined effect of seasonal holiday surge with the increase in online shopping during the COVID-19 lockdown.

The model development segment of the study determined the best model at capturing the underlying pattern in the monthly coffeemaker search interest is a SARIMA$(0, 1, 2)\times(1, 0, 1)_{12}$ model. This model was selected based on having the loweset AIC and BIC scores to ensure that it is a good fit and not over-fitted. 4 years forecast and rolling forecast origin were used to evaluate model performance. Rolling forecast origin had the lower MAPE and RMSE values compared to the 4 years forecast. It also had white noises residuals while the 4 years forecast did not pass the white noise tests. This is due to the 4 years forecast horizon being too long for accurate ARIMA-type model predictions. In conclusion, the SARIMA model was a good fit for the data.

In the future, additional variables relevant to coffeemaker search interest can be investigated and incorporated into a SARIMAX model, which is a SARIMA model with exogenous variables. Other statistical models can be tested and compared with the SARIMA model. Different lengths of forecasting windows can be used to compare with the 4 years forecast and the rolling forecast origin. Finally, it would be interesting to analyze and forecast the weekly or daily frequency of coffeemaker search interest in the United States.

REFERENCES

Applied time series analysis 12.1 estimating the spectral density. *Available at STATS 510 PennState Eberly College of Science.* 5, 6

(2021). At-home coffee consumption climbs to 81% amidst pandemic. *Available at Convenience Store News.* 13

(2023). The rising market for premium and specialty coffee at home. *Available at World Coffee Portal.* 1

(2024a). Google trends: beverage of coffee. Topic: holiday. Time Range: 04-2004 to 04-2024. Location: United States. Category: all. 3

(2024b). Google trends: topic of coffeemaker. Topic: coffeemaker. Time Range: 04-2004 to 04-2024. Location: United States. Category: all. 3

(2024). Google trends: topic of holiday. Topic: holiday. Time Range: 04-2004 to 04-2024. Location: United States. Category: all. 3

Bianco, D. (2016). Uses and abuses of arima in ppnr modeling and risk management: Why not to fear arima. *Available at SSRN 3653514.* 8

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control.* John Wiley & Sons. 4

Cadwalader, Z. (2024). Coffee's popularity in american is at a 20-year high. *Available at SPRUDGE.* 1

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431. 6

Ensafi, Y., Amin, S. H., Zhang, G., and Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning–a comparative analysis. *International Journal of Information Management Data Insights*, 2(1):100058. 15

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice.* OTexts. 4, 7, 9, 10, 24

Ljung, G. M. and Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303. 11

Mushtaq, R. (2011). Augmented dickey fuller test. 6

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting*, 16(4):437–450. 10