Modeling and Forecasting the Cost of Tomatoes in US Cities using ARIMA, SARIMA and Holt-Winters Models

Matt Ludwig

1 Introduction

As inflation and the cost of living rise across cities in the United States, the price of food has also increased. In this project, I will use several time series models to both model and forecast the price of tomatoes across cities in the United States. In particular, I will clean and impute missing values on Federal Reserve economic data as well as test for stationary with Augmented Dickey-Fuller Tests and implement differencing and seasonal differencing. I will also construct and elvaluate ARIMA and SARIMA models as well as implement a Holt-Winters smoothing-based forecast to forecast the price of tomatoes for the next few years.

2 Dataset

2.1 Overview

The dataset is from FRED (Federal Reserve Economic Data) of St. Louis and is the monthly average price per pound of field grown tomatoes across major cities in the United States from January 1st, 1980 until April 1st, 2024. The data is for fresh field grown and vine ripened round red tomatoes, both organic and inorganic.[U.S]

2.2 Cleaning the Data

Upon inspection of the data, there were several missing entries which had NA values. Using the R library "forecast" and the function "tsclean()" I found and removed outliers in the data, imputed missing values as well as stabilized the variance. Plots of the time series data before and after cleaning are shown below:

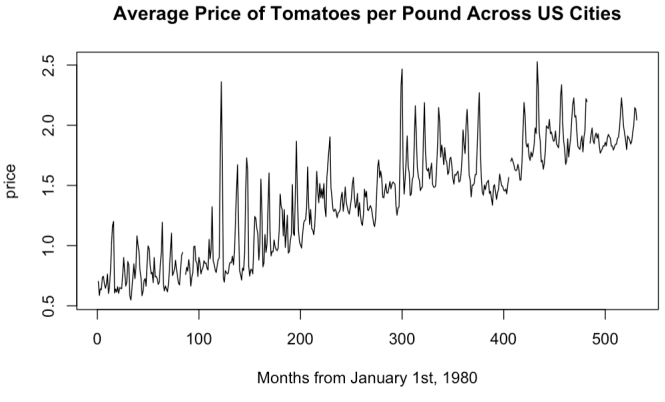

Figure 1: Plot of Original Data

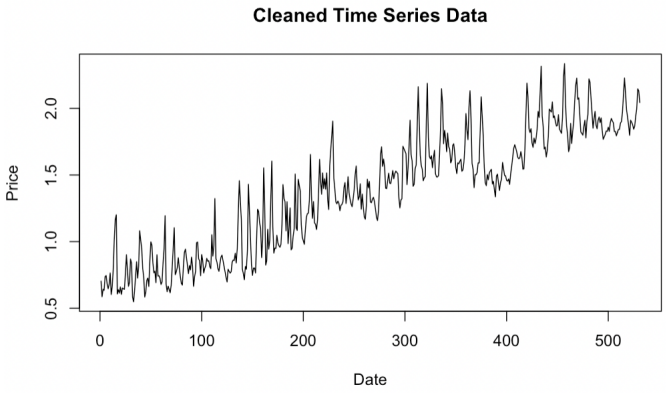

Figure 2: Plot of Cleaned Data

3 Results

3.1 Visualize, evaluate patterns of the data

After cleaning the data, used the R function "decompose" to create a classical additive decomposition of the time series data which is shown in the figures below:

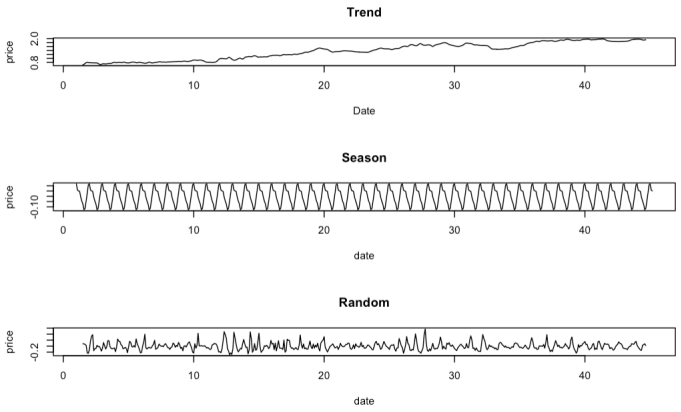

Figure 3: Classical Decomposition of the Data

From the plot of the classical decomposition of the data there is a clear increasing trend as well as seasonal pattern. To confirm that the data was non-stationary, I also performed a Dickey-Fuller test to check for a drift in the mean of the data (unit root) which produced a p-value of 0.0.4996, so I did not reject the null hypothesis that the time series is non-stationary.

3.2 Transformations

To correct for seasonality (deterministic part) and unit root, I applied a differencing as well as seasonal differencing with lag 12. I then reran the Dickey-Fuller test and found that the p-value was 0.01 indicating that the differenced data is stationary. A sample of this data is shown in the figure below:

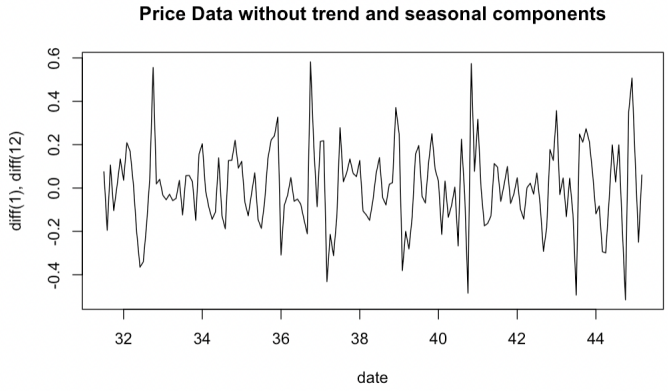

Figure 4: Random Component of the Differenced Data

3.3 Model-based Forecast: ARIMA

To forecast future values of the time series, I used a seasonal ARIMA model which I constructed from analyzing the ACF and PACF plots shown below:

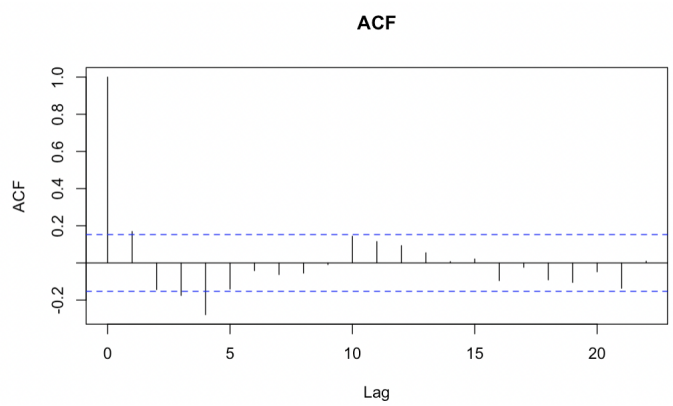

Figure 5: ACF Plot

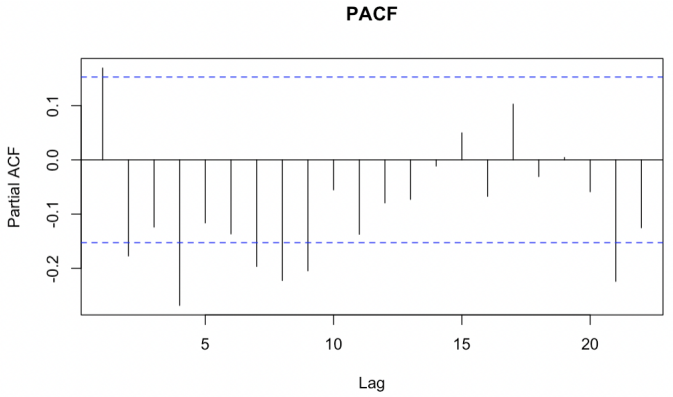

Figure 6: PACF Plot

These are the ACF and PACF plots for price data after conducting differencing and seasonal differencing. From these plots it is apparent that both the ACF plot and the PACF plot both appear to be decaying to zero. Therefore it seems reasonable to consider a low order ARMA model. It is also apparent that around lag 10 and lag 20, the ACF values again become significant before decaying to zero. Also, the PACF values increase around lag 10 and lag 20 as well. To try and capture these subtlties in the data, I decided to fit several ARIMA and SARIMA models to the original data and compare their AIC values which are shown in the table below:

Table 1: Model Comparison

Model	**AIC**
ARIMA(0,1,0)	-573.78
ARIMA(0,1,1)	-573.67
ARIMA(0,1,2)	-642.23
ARIMA(1,1,0)	-572.86
ARIMA(1,1,1)	-580.77
ARIMA(1,1,2)	-676.59
ARIMA(2,1,0)	-598.44
ARIMA(2,1,1)	-681.16
ARIMA(2,1,2)	-687.04
SARIMA(0,1,0)(0,1,0)10	-205.34
SARIMA(0,1,1)(0,1,0)10	-204.86
SARIMA(0,1,2)(0,1,0)10	-327.86
SARIMA(1,1,0)(0,1,0)10	-204.03
SARIMA(1,1,1)(0,1,0)10	-240.34
SARIMA(1,1,2)(0,1,0)10	-356.46
SARIMA(2,1,0)(0,1,0)10	-245.98
SARIMA(2,1,1)(0,1,0)10	-359.01
SARIMA(2,1,2)(0,1,0)10	-357.52

3.4 ARIMA Forecast for ARIMA Model with Lowest AIC

Since the ARIMA(2,1,2) model had the lowest AIC value, I used it to forecast the average price of tomatoes per pound for the next two years. The plot of the forecast with 90% and 95% confidence intervals is shown below:

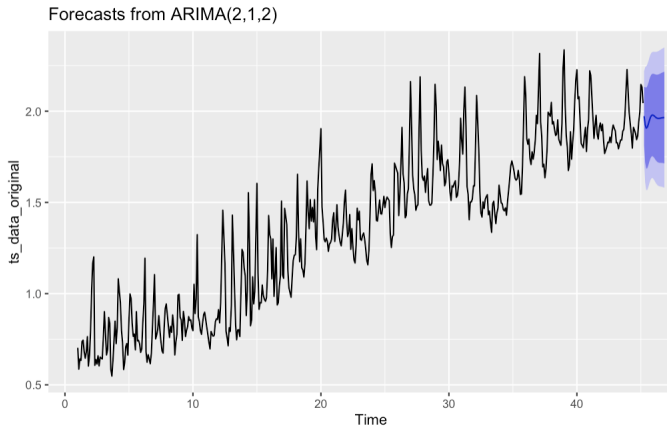

Figure 7: Forecast Prediction for ARIMA Model with Lowest AIC

3.4.1 Model Diagnostics for ARIMA Model with Lowest AIC

To ensure the forecast was accurate, I plotted the model's residuals and the ACF of the residuals which are shown in the figures below:

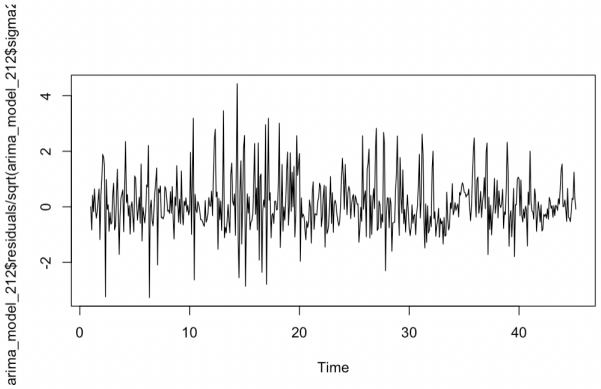

Figure 8: ACF Plot

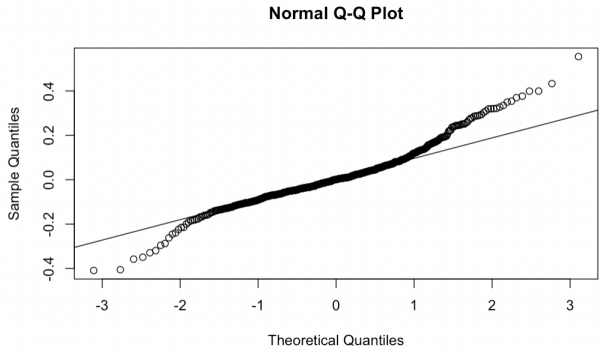

Figure 9: PACF Plot

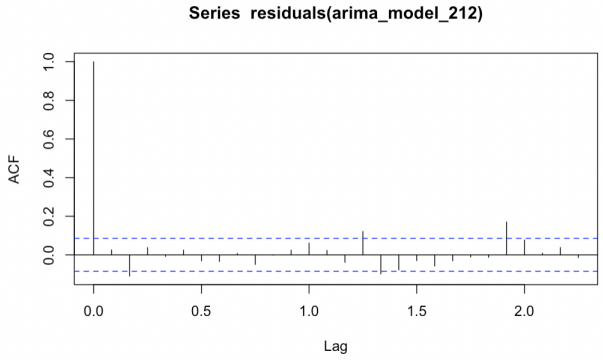

Figure 10: ACF Plot

3.5 ARIMA Forecast for Auto ARIMA Model

Following this, I compared the AIC selected model to the model selected by the "auto-arima" function in the forecast library. A plot of the forecast from the selected ARIMA(3,1,2)(2,0,0)[12] is shown in the figure below:

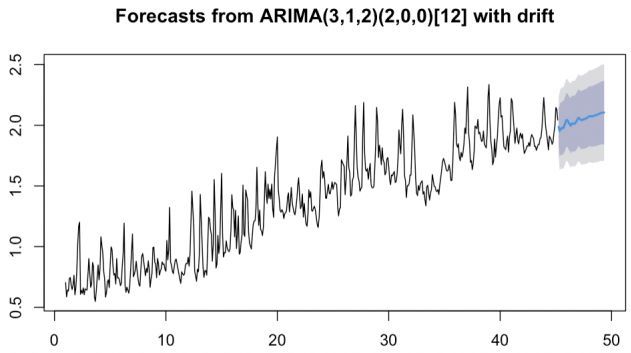

Figure 11: Forecast of Auto ARIMA Model

3.5.1 Model Diagnostics for Auto ARIMA Model

To determine how well the auto-arima forecast was, I plotted the model's residuals and the ACF of the residuals which were all very similar to the plots for the ARIMA model with the lowest AIC.

3.6 Smoothing-based Forecast: Holt-Winters

Another widely used and successful forecasting method is the seasonal Holt-Winters method. The smoothing-based method uses the pattern of the data to extrapolate the forecast using double exponential smoothing. The Holt-Winters forecast is shown in the figure below:

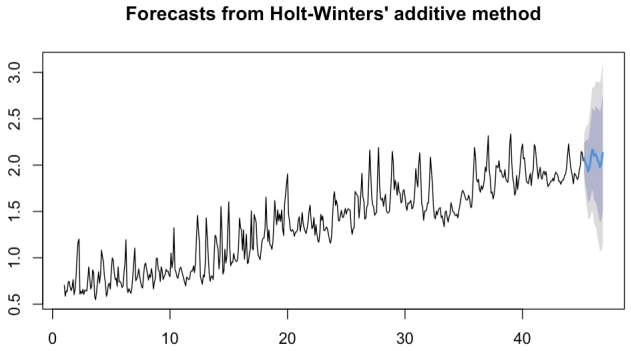

Figure 12: Holt-Winters Forecast

3.7 Evaluation of Forecast Accuracy

To judge forecast accuracy of the auto ARIMA and Holt-Winters forecasts, I used an out-of-sample forecast validation to compare the two methods. I also used the root mean square error (RMSE), Mean Absolute Error (MAE), and Mean absolute Percentage Error (MAPE) measures to evaluate the forecast. After running the metrics, I found that the auto-ARIMA forecast was better. A summary of the evaluation metrics is given in the table below:

Table 2: Model Comparison

Model	MAE	RMSE	MAPE (%)
ARIMA	0.1053	0.1282	5.4939
Holt-Winter	0.1115	0.1330	5.8162

4 Conclusions

The purpose of this time series analysis was to forecast the average tomato price per pound across cities in the United States. I constructed and compared several ARIMA and SARIMA models using ACF and PACF plots as well as AICc values to find a model for the price of tomatoes per pound since January 1st, 1980. After selecting a reasonable ARIMA model, I compared my finding against the auto arima function in R. I then examined the standardized residuals, the ACF of the residuals and the p-value for the Ljung-Box statistic for each model to ensure that they were both reasonable. Following this, I implemented ARIMA and Holt-Winters forecasts and compared their accuracy on a test set through RMSE, MAE, and MAPE and ultimatley found that the ARIMA forecast provided the best forecast for the price of tomatoes per pound.

For future analysis, it would be interesting to experiment with different data cleaning techniques as well as investigate outlier data points more closely to see if they were correlated or connected to real world events such as economic recessions or weather related events such as hurricanes or droughts which could affect the growing and price of tomatoes.

References

[U.S] U.S. Bureau of Labor Statistics. Average price: Tomatoes, field grown (cost per pound/453.6 grams) in u.s. city average [apu0000712311]. Retrieved from FRED, Federal Reserve Bank of St. Louis. Accessed on April 24, 2024.