Time Series Analysis of Retail Gas Prices in Boston

David Custer

Department of Mathematics and Statistics

Coastal Carolina University

1. Introduction and Decomposition

Over 20 million people travel through Boston every year. At some point, locals and visitors will need to stop for gas. Our goal is to model the Energy Information Administration's collection of retail gas prices in Boston for all grades **from January 2010 to December 2019** and predict future gas prices over the next five years until 2025. Modeling and forecasting data over time requires methods beyond regression. For example, residual assumptions such as constant variance and Normality generally do not hold for time-dependent data. A linear relationship between the predictor and response rarely exists. Furthermore, the most significant issue is that autocorrelation between observations refutes the premise of mutual independence. We can employ time series analysis (a stochastic process with an indexed collection of random variables that randomly evolves) to perform thorough statistical inference. Our first step starts with plotting the data over time to look for seasonality, the periods of predictable cyclic patterns.

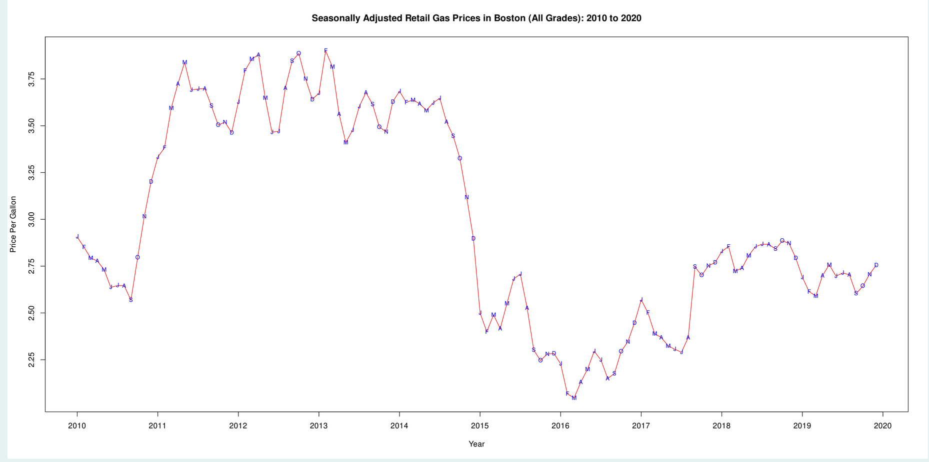

Figure 1: Seasonally Adjusted Data

- The seasonally adjusted retail gas prices in Boston range from $2.05 to $3.9 per gallon.

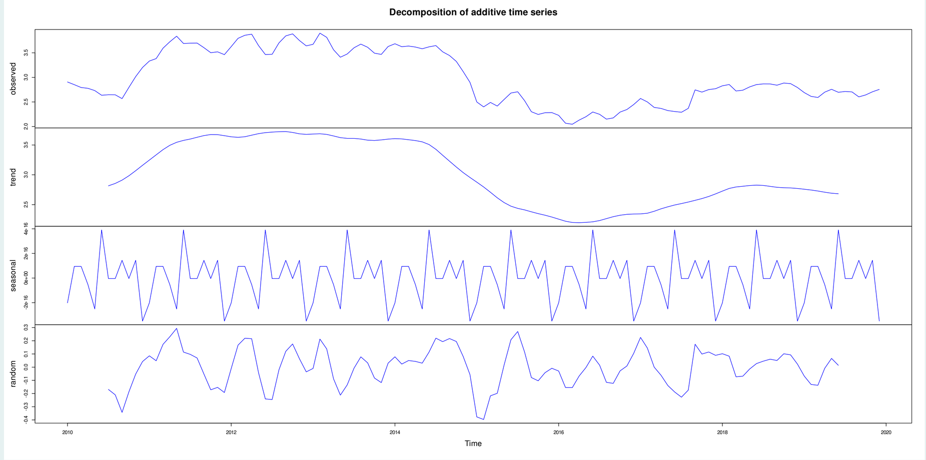

Figure 2: Additive Decomposition of Time Series

- This additive decomposition breaks down our series into four components used for abstraction.

2. Autocorrelation and Differencing

The Difference Process

- Differencing begins by subtracting the previous observation from the current observation.
- A process made stationary by differencing is an integrated time series [1].
- ACF: autocorrelation function $\rightarrow$ measures the linear relationship between current and previous observations with linear dependency
- PACF: partial autocorrelation function $\rightarrow$ measures the linear relationship between current and previous observations with the linear dependency removed
- MA(q): a moving average process of order q with parameter θ, where the current value relies on previous residuals
- AR(p): an autoregressive process of order p with parameter ϕ, where the current value relies on previous observations

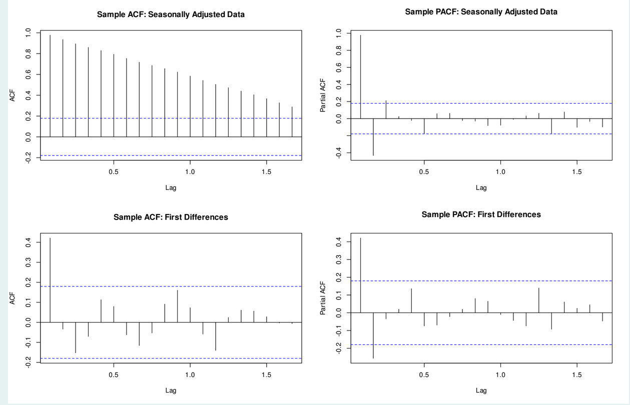

Figure 3: Sample Correlograms

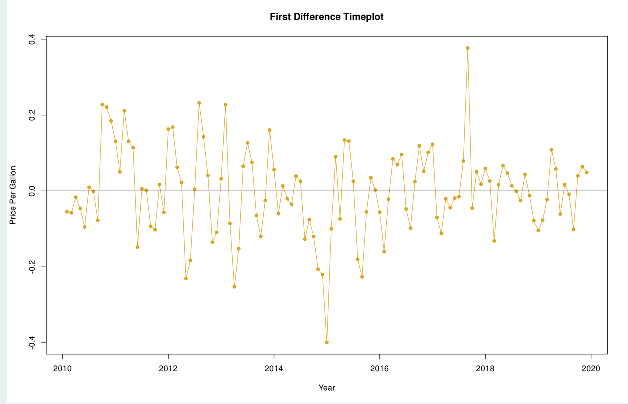

Figure 4: First Difference Process

- Before Differencing: sample ACF decays slowly, suggesting first differences should be taken, and the sample PACF cuts off after lag k = 3 (vertical bars fall within blue dashed lines)
- After Differencing: sample ACF cuts off after lag k = 1, and the sample PACF cuts off after lag k = 2

	AR(p)	MA(q)
ACF	Tails off	Cuts off after lag q
PACF	Cuts off after lag p	Tails off

Table 1: ACF and PACF Behavior

3. Stationarity Testing and Invertibility

Stationarity Testing

- Aside from the sample ACF and PACF correlograms in the previous section, we can use more formal tests for stationarity (a condition where the properties of the data are constant throughout time).
- A stationary process is a series that exhibits zero trends, no seasonality, and has constant variance.
- KPSS: determines if the trend of a series or the level (observed) values are stationary
- ADF: ascertains if a unit root is present in a series

Test	Test Statistic	p-value	Conclusion
ADF	-2.1538	0.5127	little to no evidence of stationarity
KPSS (Level)	0.4916	0.0436	strong evidence of nonstationarity
KPSS (Trend)	0.1192	0.0996	some evidence of nonstationarity

Table 2: Stationarity Test Results Before Differencing

Test	Test Statistic	p-value	Conclusion
ADF	-2.5536	0.3467	little to no evidence of stationarity
KPSS (Level)	0.1242	> 0.1	little to no evidence of nonstationarity
KPSS (Trend)	0.1165	> 0.1	little to no evidence of nonstationarity

Table 3: Stationarity Test Results After Differencing

- In the table before differencing, all three tests result in evidence against the stationarity of the gas price data.
- However, after differencing, both KPSS tests suggest that our integrated series is stationary.
- From the ADF test, a complex unit root exists due to the sinusoidal pattern seen after differencing in the previous section.

Invertibility Requirements

- A series is invertible if the residuals represent a linear function of current and past observations.
- Invertibility allows us to calculate the model's parameters [3].
- For an MA(q) process to be invertible, we need the roots of the MA characteristic polynomial to all exceed one in absolute value (or modulus, for this series).
- If $|\theta| > 1 \iff$ distant observations have a more significant effect on current observations.
- If $|\theta| = 1 \iff$ all observations have the same influence as the current observations.
- If $|\theta| < 1 \iff$ recent observations have a more significant effect on current observations.

4. Order Specification and Parameter Estimation

The Extended Autocorrelation Function

- Now that we have enough evidence that our series is stationary, we can specify the order and estimate the parameters of our model.
- Since the sample ACF and PACF correlograms show evidence that an AR or MA model is appropriate, we can use the EACF to determine the order of a mixed ARMA process that models the gas price data within our ten-year timeframe.
- Our EACF on the left forms a wedge with its tip at the top left point (0,1), suggesting an MA(1) model.
- The EACF on the right is an example that would suggest an ARMA(1,1) process.

$$\begin{array}{cc} & MA \\ AR & \left[\begin{array}{c|cccc} & 0 & 1 & 2 & 3 \\ \hline 0 & x & o & o & o \\ 1 & x & o & o & o \\ 2 & o & o & x & o \end{array}\right] \end{array} \qquad \begin{array}{cc} & MA \\ AR & \left[\begin{array}{c|cccc} & 0 & 1 & 2 & 3 \\ \hline 0 & x & x & o & x \\ 1 & x & o & o & o \\ 2 & o & x & o & o \end{array}\right] \end{array}$$

Akaike and Bayesian Information Criterion

- Given the output of the EACF and by the principle of parsimony, we will use an MA(1) process to model the gas prices.
- The auto.arima function in R selects the order of the series via the Hyndman-Khandakar algorithm.

Criterion	Seasonally Adjusted	First Differences
AIC	ARIMA(0,1,1) $\equiv$ IMA(1,1)	MA(1)
BIC	ARIMA(0,1,1) $\equiv$ IMA(1,1)	MA(1)

Table 4: Order Specification Before & After Differencing

- Both model selection methods agree that a first-order integrated moving average fits our data.
- The integrated moving average before differencing is equivalent to the moving average after differencing.
- Maximum Likelihood Estimation (MLE) maximizes the log-likelihood function of the ARIMA model to estimate its parameters.
- We will use the MLE from R to make forecasts.

5. Residual Diagnostics and Model Suitability

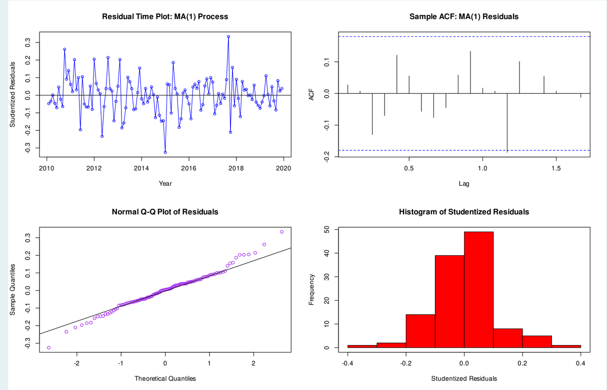

Figure 5: Residual Diagnostics

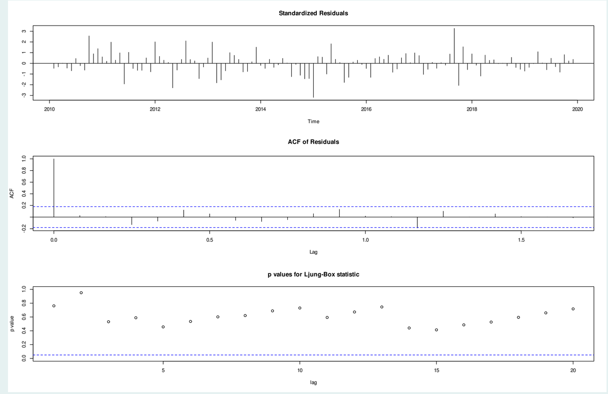

Figure 6: Box-Ljung Output

Test	p-value	Conclusion
Shapiro-Wilk	0.1948	little to no evidence of non-Normality
Runs	0.484	little to no evidence of non-independence
Portmanteau Q	p↑ as k↑	little to no evidence of heteroskedasticity
Lagrange Multiplier	p↑ as k↑	little to no evidence of heteroskedasticity
Box-Ljung	0.6581	little to no evidence an MA(1) process is unfit

Table 5: MA(1) Residual Test Results

- We now have substantial evidence based on our p-values above that our MA(1) model is well suited for our seasonally adjusted data. We can now attempt forecasting, as seen in the next section.
- Two notable outliers exist at the beginning of 2015 and near the end of 2017. "By 2016, after several years of increasing production, particularly from the Bakken shale oil fields, the market was flooded with oil. That was one of the main factors that pushed crude prices down more than 50% in four months. Moreover, gas prices followed, dropping to $1.72 on February 15, 2016" [2].

6. Forecasting

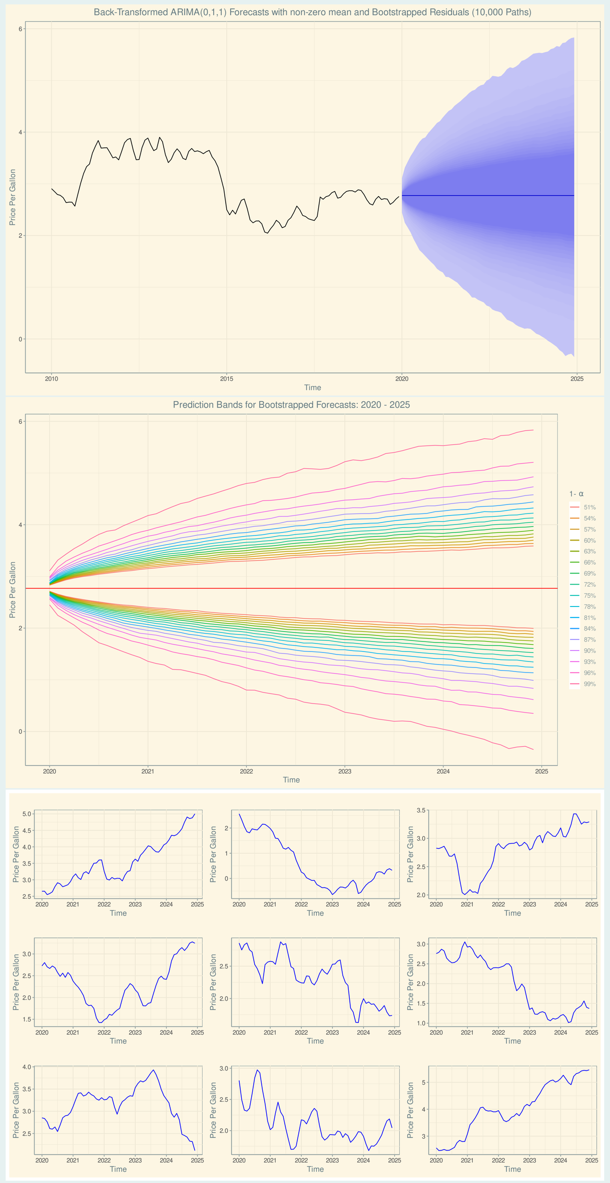

Figure 7: Bootstrap Extrapolation

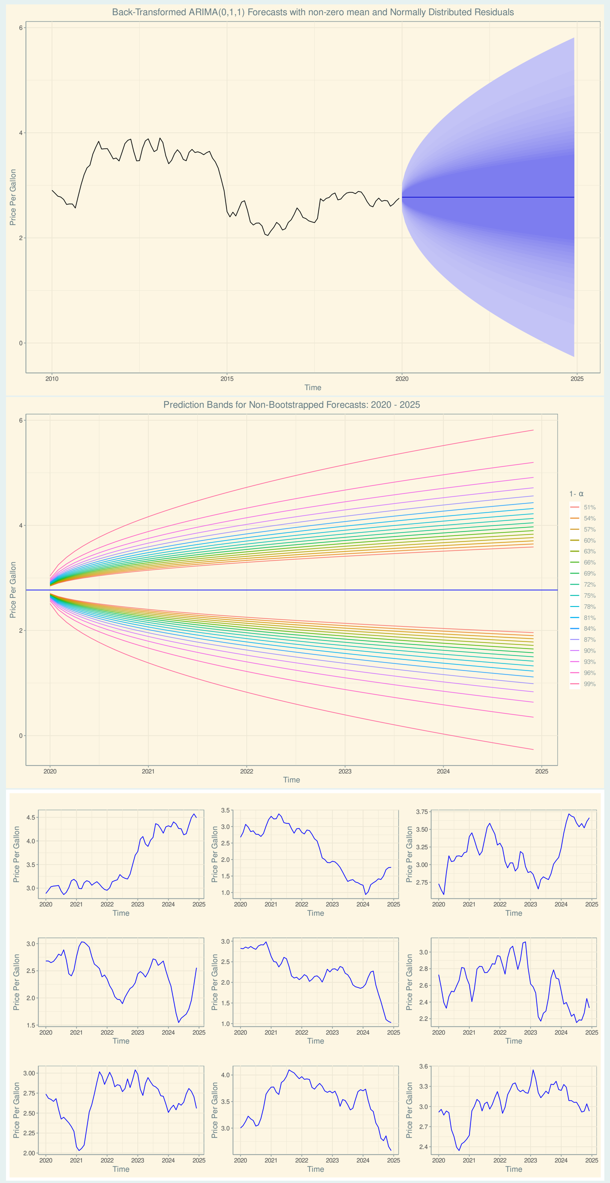

Figure 8: Normal Extrapolation

- For our forecasts and simulations, the left side consists of bootstrapped (resampled) residuals, unlike their Normally distributed counterparts on the right.
- The prediction bands are smoother for Normally distributed residuals since resampled residuals may not always form a bell curve.
- The forecast funnels at the top consist of horizontal lines after 2020. This line is the moving average's prediction of gas prices **until December 2024,** with a forecasted price of $\approx$ $2.77 per gallon.
- When looking at the simulations, we notice that moving averages do not predict future values well. A notable drawback is their reliability on residuals from previous data, with each observation given equal weight. Moving averages also tend to overlook volatility in prices and cyclic patterns. An ARMA process could forecast better but at the cost of a more complex model.

7. References

[1] Mathworks Deutschland. Trend-stationary vs. difference-stationary processes. 2022.

[2] Jeff Lenard. When were gas prices low? *National Association of Convenience Stores*, March 2022.

[3] PSU. 2.1: Moving average models. *Eberly College of Science: Statistics Online Courses*, 2022.