Natural Gas Price Analysis

Jongbae Yoon

Applied Mathematics, University of Colorado Boulder, Boulder, CO, USA

jongbae.yoon@colorado.edu

Abstract

Natural gas has various advantages over other types of fossil fuels as an energy source, and numerous studies have been conducted to understand not only its supply and demand, but also its price movements. The aim of this study is to comprehend the key predictors of natural gas prices and build a forecasting model using both frequentist and Bayesian methods. The Henry Hub natural gas price serves as the response variable, with 11 different datasets of supply and demand factors selected as predictors. The dataset covers monthly data from January 1997 to January 2023. To identify the crucial variables, both Vector Autoregression (VAR) and Bayesian VAR (BVAR) methods are utilized. For the forecasting task, Autoregressive Integrated Moving Average (ARIMA) and Bayesian Structural Time Series (BSTS) methods are employed. The results shows that the previous natural gas price, heating oil price, monthly gas export, degree days, and the heating oil price of 2 previous time stamp have the significant correlation with the current natural gas price. In addition, the prediction results show reasonable evaluation criteria values, but overall, the frequentist method yields better results than the Bayesian methods.

1. Introduction

Natural gas is a fossil fuel that is primarily composed of methane, along with other hydrocarbons such as ethane, propane, and butane. It is formed deep beneath the earth's surface by the decomposition of organic matter over millions of years. Natural gas is a colorless and odorless gas in its natural state but is typically processed with added odorants for safety reasons. It is typically found in reservoirs beneath the earth's surface and is extracted through drilling wells. Natural gas is commonly used as a fuel for heating, cooking, and electricity generation, among other industrial applications. In 2022, about 4.24 trillion kWh of electricity were generated at utility-scale electricity generation facilities in the United States. About 40% of this electricity generation was from natural gas [1] (EIA, 2022). Burning natural gas for energy results in fewer emissions of nearly all types of air pollutants and carbon dioxide (CO_2) than burning coal or petroleum products to produce an equal amount of energy. About 117 pounds of CO_2 are produced per million British thermal units (MMBtu) equivalent of natural gas compared with more than 200 pounds of CO_2 per MMBtu of coal and more than 160 pounds per MMBtu of distillate fuel oil. The clean burning properties of natural gas have contributed to increased natural gas use for electricity generation and as a transportation fuel for fleet vehicles in the United States [2] (EIA, 2022). Consequently, comprehending the factors driving natural gas prices, as well as the associated demand and supply, is essential for a wide range of stakeholders. Nevertheless, natural gas prices can be challenging to predict due to various factors that are closely interrelated, as well as unpredictable events such as geopolitical issues and unexpected weather patterns. For example, Russia is the second-largest natural gas producer in the world, only behind the United States [3] (EIA, 2022), and many European countries heavily rely on Russian gas. The Russo-Ukrainian war caused widespread concern about potential natural gas shortages, which could lead to an increase in prices further. However, due to an unexpectedly warm winter, the natural gas price has remained below expectations.

This paper primarily focuses on comprehending the driving factors of natural gas prices and constructing a model to forecast the future movement of the Henry Hub natural gas spot price. Of the many factors influencing natural gas prices, I select 11 major factors that have a significant impact on prices as predictors. For understanding the driving factors task, I apply two methods: Vector Autoregression (VAR) and Bayesian VAR. VAR is widely used to analyze the relationship between multiple time series variables. For the forecasting the future price movement task, I employ traditional Autoregressive integrated moving average (ARIMA) models and Bayesian Structural Time Series models. ARIMA is one of the most widely used time series modeling technique for the forecasting future values based on past observations. The performances of each model will be compared.

The rest of this paper is organized as follows: Section 2 introduces related works and literature reviews. Section 3 explains the entire dataset in detail. The section 4 explains key methods, techniques and background algorithms applied in this paper. Section 5 shows the results and explains insights and limitations of the study. Section 6 summarizes the overall paper and explains potential implications of this study.

2. Related works

Given its environmentally friendly characteristics and abundant availability, natural gas has a is expected to play an extended role in the global energy landscape. As a result, it has become increasingly important to accurately forecast the supply, demand, and price of natural gas. With the availability of vast datasets and advanced computing capabilities, researchers and scholars have been able to construct complex mathematical and computational models, leading to a surge in studies dedicated to analyzing the dynamics of natural gas prices. [4] (Jia Ding et al., 2023) built a neural network model to forecast natural gas consumption with multiple seasonal patterns. They found that the demand for natural gas often exhibits different seasonal patterns, and proposed a novel method, Dual Convolution with Seasonal Decomposition Network (DCSDNet), for natural gas consumption forecasting. Seasonal-trend decomposition was carried out on the original time series dataset and DCSDNet was then performed on the trend and residual, followed by the self-attention module. They concluded that the proposed DCSDNet method outperforms several state-of-the-art methods under various circumstances. [5] (Jolanta Szoplik et al., 2015) used an artificial neural network Multilayer perceptron (MLP) model to forecast the actual natural gas consumption in Szczecin in Poland. They found that the MLP model with 22 inputs per a batch, 36 hidden layers and 1 output layer performs the best among other models. They also argued that this model can predict the natural gas consumption on any day of the year and any hour of the day.

Lots of studies have shown that Artificial Neural Network (ANN) models or Recursive Neural Network (RNN), specifically Long Short Term Memory (LSTM) models outperform traditional time series models on stochastic dataset. [6] (Kexian et al., 2022) tried to compare the performance of ARIMA, ANN and LSTM models on crude oil price. They built traditional ARIMA model, and two of neural network models, ANN model and LSTM model. They compared the performance of the models and concluded that the LSTM model demonstrated higher forecasting accuracy and better forecasting stability for different timescales than ARIMA and ANN based models. [7] (Qihang Ma, 2020) compared the ARIMA, ANN, and LSTM on stock price. The paper mentioned that ARIMA and ANN have been widely used in time-series data forecasting, but these models cannot measure the continuity of the trends. However, due to its characteristics of feedback connection, LSTM makes it easier to find development trends through the back propagation of historical prices and the current prices. They conclude that the LSTM model performs better than ANN and ARIMA models. They assume the reason of the superiority of LSTM is the improvement of the LSTM model on the problem of vanishing gradient. [8] (Moting Su et al., 2019) investigated to develop a predictive model for natural gas price using various machine learning models, including Artificial Neural Network (ANN), Support Vector Machine (SVM), Gradient Boosting Machines (GBM), and Gaussian Process Regression (GPR). The study identified strong correlations between the number of rotary rigs, production, and import, and the natural gas price. After evaluating the performance of each model using various test metrics, the authors concluded that the ANN model outperformed the other models tested.

However, in this paper, ANN, RNN, LSTM models are not considered. These models perform best when the dataset is large enough, but the dataset for this paper is relatively small. In addition, one of the goals of this paper is finding the relationships between variables and Vector Autoregression models are superior on such a task. [9] (Xuejun Zhao et al., 2021) argued that Bayesian regression analysis is a reliable model for investigating variables having a significant impact on the output of a particular process, such as financial stock market. They also concluded that the VAR and the classical frequentist approach achieve a higher probability accuracy than non-Bayesian methods such as the Auto-Regressive and Moving Average Model time series models. [10] (Sebastian Nick et al., 2013) tried to find the driving factors of the natural gas price of the German natural gas market using structural vector autoregressive model (VAR). The study shows that the natural gas price is affected by temperature, storage, and supply shortfalls in the short term, while the long-term development is closely tied to both crude oil and coal prices, capturing the economic climate and the energy specific demand. [11] (Dixon Domfeh et al., 2019) examined the dynamic economic relationships between the fundamental variables that influence natural gas prices within the U.S. market. They implemented a structural vector autoregressive and Markov switching models to investigate the impact and stability of regime switches between the main drivers of natural gas prices. Their results show that the U.S. gas market is sensitive to temperature deviations in the short term, and Crude oil and coal prices have long-run effects on natural gas prices.

3. Dataset

Like most commodities, the supply and demand dynamics are the fundamental drivers of the price of natural gas. Increases in natural gas supply generally results in lower natural gas prices and decreases in supply tend to lead to higher prices. Three major supply-side factors affect prices are amount of natural gas production, level of natural gas in storage, and volumes of natural gas imports and exports. Three major demand-side factors affect prices are variations in winter and summer weather, level of economic growth, and availability and prices of other fuels [12] (EIA, 2022). The entire dataset is the monthly U.S. data from January of 1997 to the January of 2023. Natural log is applied on the entire dataset to stabilize the time series datasets.

3.1. Henry hub natural gas spot price

Henry Hub is a natural gas trading point located in Erath, Louisiana, USA. It is a key benchmark for natural gas prices in North America and represents the price at which natural gas is traded on the New York Mercantile Exchange (NYMEX). The Henry Hub is connected to several major natural gas pipelines that serve markets throughout the United States, making it a critical distribution point for natural gas in the country. It is quoted in US dollars per million British thermal units (USD/MMBtu). 1 MMBtu is equivalent to the amount of heat required to raise the temperature of one pound of water by one degree Fahrenheit at a constant pressure of one atmosphere. By expressing the price of natural gas in USD/MMBtu, market participants can compare the cost of natural gas with other sources of energy, such as crude oil, coal, and electricity, which are typically priced in different units.

3.2. Cushing, OK WTI Spot Price FOB

The Cushing, OK WTI (West Texas Intermediate) Spot Price FOB is a benchmark price for crude oil in the United States. Cushing, Oklahoma is a major crude oil storage hub and serves as the delivery point for the WTI futures contract traded on the NYMEX. As such, the Cushing, OK WTI Spot Price FOB is a key reference point for crude oil prices in the United States and it is based on the price of crude oil that is available for immediate delivery at the Cushing, OK storage hub and quoted in US dollars per barrel. Historically, crude oil prices have been thought to be one of the most critical factor affecting natural gas prices in the long run.

3.3. New York Harbor No.2 Heating Oil Spot Price FOB

The New York Harbor No.2 Heating Oil Spot Price FOB is a benchmark price for heating oil in the United States. Heating oil, also known as No.2 fuel oil, is a distillate fuel oil that is commonly used for space heating and hot water heating in residence and commercials. The New York Harbor No.2 Heating Oil Spot Price FOB is based on the price of heating oil that is available for immediate delivery in the New York Harbor area. It is quoted in US dollars per gallon. It serves as a key reference point for heating oil prices in the Northeastern United States. It is widely known that competition with other fuels influences natural gas prices. The heating oil spot price is only available from October to March of each year. To make it a stationary dataset, I used interpolation for the rest of the months of a year.

3.4. U.S. Natural Gas Rotary Rigs in Operation

US natural gas rotary rigs in operation refer to the number of active drilling rigs in the US. that are currently drilling for natural gas using rotary drilling equipment. This equipment is commonly used for drilling oil and gas wells, and it is a critical component of the upstream oil and gas industry. The number of natural gas rotary rigs in operation is a key indicator of the level of activity in the natural gas exploration and production industry. It provides insights into the supply and demand dynamics of the natural gas market.

3.5. Heating Degree-days Cooling Degree-days

The natural gas demand in the residential heating market is susceptible to temperature. Heating Degree Days (HDD) is a metric used to estimate the amount of energy required to heat a building or a region during the winter months. It is calculated by subtracting the average outdoor temperature for a day from a base temperature (usually around 65°F or 18°C) and summing up the differences for a given period of time, usually a month. The resulting value represents the number of degree-days that are accumulated, and it is used by energy analysts, utilities, and governments to track heating demand. Similarly, Cooling Degree Days (CDD) is a metric used to estimate the amount of energy required to cool a building or a region during the summer months. It is calculated by subtracting a base temperature from the average outdoor temperature for a day and summing up the differences for a given period. The resulting value represents the number of degree-days that are accumulated. I construct degree day deviations, DD. I add HDD and CDD of each month and subtract the average of degree days for the entire duration to find out whether each month has an extreme weather.

3.6. U.S. Natural Gas Marketed Production and consumption.

US natural gas marketed production refers to the volume of natural gas that is extracted from the ground and sold to customers. This is the major supply factor of the natural gas and have significant impact on the natural gas price. US natural gas total consumption is the total amount of natural gas that is used by consumers in the United States for various purposes, such as heating, electricity generation, industrial processes, and transportation. This is the major demand factor of the natural gas and is an important indicator of energy demand and economic activity. The unit of both production and consumption is Million Cubic Feet (MCF).

3.7. U.S. Total Natural Gas Underground Storage Volume

US total natural gas underground storage volume is the total volume of natural gas stored in underground facilities, such as depleted natural gas reservoirs, salt caverns, and aquifers, across the United States. Natural gas is stored in underground facilities to ensure a reliable supply of natural gas during periods of high demand, such as during cold winters or hot summers when natural gas consumption for heating or cooling is high. Storage also allows for natural gas to be supplied to markets that are not directly connected to natural gas production areas via pipelines. The unit of the storage volume is measured as MCF.

3.8. U.S. Natural Gas Imports and exports

US natural gas imports refer to the volume of natural gas that is imported into the United States from other countries. Natural gas imports can come in the form of pipeline imports from neighboring countries such as Canada or via liquefied natural gas (LNG) shipments from countries such as Qatar or Australia. US natural gas exports refer to the volume of natural gas that is exported from the US to other countries. The US has become a major exporter of natural gas in recent years due to the growth in domestic natural gas production, particularly from shale gas resources. The unit of the import and exports of the natural gas is also MCF.

3.9. Monthly Real GDP Index

To indicate the level of economic growth, US GDP index is utilized. The monthly Real Gross Domestic Product (GDP) index of the US is a measure of the total value of goods and services produced in the United States in each month, adjusted for inflation. The real GDP index is used as an indicator of the overall health and performance of the US economy. The unit of the GDP is US. Dollar. The entire dataset is summarized as below in the table.1, and the movement of each dataset is plotted in the Appendix#1.

Category		Variable	Unit
Response		Henry Hub Natural Gas spot Price	USD / MMbtu
Supply	Production	Natural Gas Marketed Production Natural Gas Rotary Rigs in Operation	MCF EA
	Storage	Total Natural Gas Underground Storage Vol.	MCF
	Imports & Exports	Natural Gas Imports and exports	MCF
Demand	Weather	Degree days deviation	EA
	Economic growth	Monthly Real GDP Index	USD
	Other fuels	Cushing, OK WTI Spot Price New York Harbor 2 Heating Oil Spot Price	USD / bbl USD / gal
	Consumption	Natural Gas Total Consumption	MCF

Table 1. Dataset

4. Methodology

As mentioned in the introduction, there are two goals in this paper. The first is comprehending the driving factors of natural gas price. To perform this task, VAR is applied, since it is widely utilized to analyze the relationship between multiple time series variables. Similarly, Bayesian VAR (BVAR) is also applied as a Bayesian method to find out the impacts of each predictor. For the second goal, which is building a forecasting model for the future movement of the natural gas price, also two different methods are utilized, traditional ARIMA and BSTS. The performance of each model will be evaluated using various evaluation criteria MAE, MSE, RMSE, and MAPE.

4.1. Vector Autoregression (VAR)

Vector autoregression is a statistical model used to analyze the dynamic relationship between multiple time series variables. In a VAR model, each variable is expressed as a linear function of its own past values, as well as the past values of the other variables in the model.

A lag is the value of a variable in a previous time period. So, in general a p^{th} order VAR refers to a VAR model which includes lags for the last p time periods. A p^{th}-order VAR is denoted "VAR(p)". A p^{th}-order VAR model is written as:

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t$$

Where, c is intercept and e is error terms.

To find the lag 'p', I used Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). AIC returns 5 while BIC returns 2. In general, AIC is preferred when the emphasis is on model prediction accuracy and the sample size is relatively small. AIC tends to select more complex models that may fit the data better and make better predictions. However, these more complex models may overfit the data and perform poorly when applied to new data. On the other hand, BIC is preferred when the emphasis is on model parsimony and the sample size is relatively large. BIC tends to select simpler models that may not fit the data as well as more complex models but are less likely to overfit the data and have better generalization properties. BIC is especially useful when the number of parameters in the model is large compared to the sample size, as it penalizes model complexity more strongly than AIC. BIC is chosen over AIC to avoid overfitting. Because the model has relatively large number of parameters compared to the sample size, using the result of AIC of p = 5, is likely to return overfitting. Thus, lag p is set to 2 for VAR analysis.

4.2. Bayesian VAR

The Bayesian approach to VAR estimation involves placing prior distributions on the model parameters, which allows for uncertainty to be quantified and incorporated into the analysis. By incorporating prior knowledge and uncertainty into the analysis, the Bayesian VAR model can provide more accurate and robust predictions than traditional VAR models. BVAR has the same mathematical form with VAR model. To approximate the posterior distribution, MCMC sampling with 10,000 iterations and 200 burn-in periods are considered. After iterations, 9,800 samples of forecast were created (9,800 x 63(test_size) x 11(variables)), and found the forecast value using colMeans (9,800 x 63) of the first variable (y).

4.3. Performance Evaluation Criteria

To evaluate the performance of each model, first 80% of the data are selected as training data, and the remaining 20% of the data are selected as test data. There exists many model evaluation criteria and some classic statistical criteria for testing model performance are selected.

4.3.1. Mean Absolute Error (MAE)

MAE is commonly used metric to evaluate the performance of regression models. It measures the average absolute difference between the predicted and actual values of a dataset.

$$MAE = \frac{1}{N} \times \sum_{t=1}^{N} \left| \widetilde{y_t} - y_t \right|.$$

4.3.2. Mean Squared Error (MSE)

MSE measures the average squared difference between the predicted and actual values.

$$MSE = \frac{1}{N} \times \sum_{t=1}^{N} \left(\widetilde{y_t} - y_t \right)^2.$$

4.3.3. Root Mean Squared Error (RMSE)

RMSE is the square root of MSE.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \times \sum_{t=1}^{N} \left(\widetilde{y_t} - y_t \right)^2}.$$

4.3.4. MAPE (Mean Absolute Percentage Error)

MAPE measures the average percentage difference between the predicted and the actual values.

$$MAPE = \frac{1}{N} \times \sum_{t=1}^{N} \left| \frac{\widetilde{y_t} - y_t}{y_t} \right|.$$

4.4. Auto Regressive Integrated Moving Average (ARIMA)

To perform forecasting task of the natural gas price, ARIMA model is utilized. ARIMA is one of the most popular time series modeling techniques for the forecasting future values based on past observations. For ARIMA {X}, write as

$$\phi(\mathrm{B})(1-\mathrm{B})^{\mathrm{d}}\mathrm{X}_{\mathrm{t}} = \theta(\mathrm{B})\mathrm{Z}_{\mathrm{t}}, Z_t \sim WN(0,\sigma^2)$$

Since the natural gas price is heavily rely on the previous price, so the autoregressive term (p =1) and the first order differencing (d=1) are selected for the model. The auto correlation and partial auto correlation of the residuals from the ARIMA model shows near stationarity property as below plot.

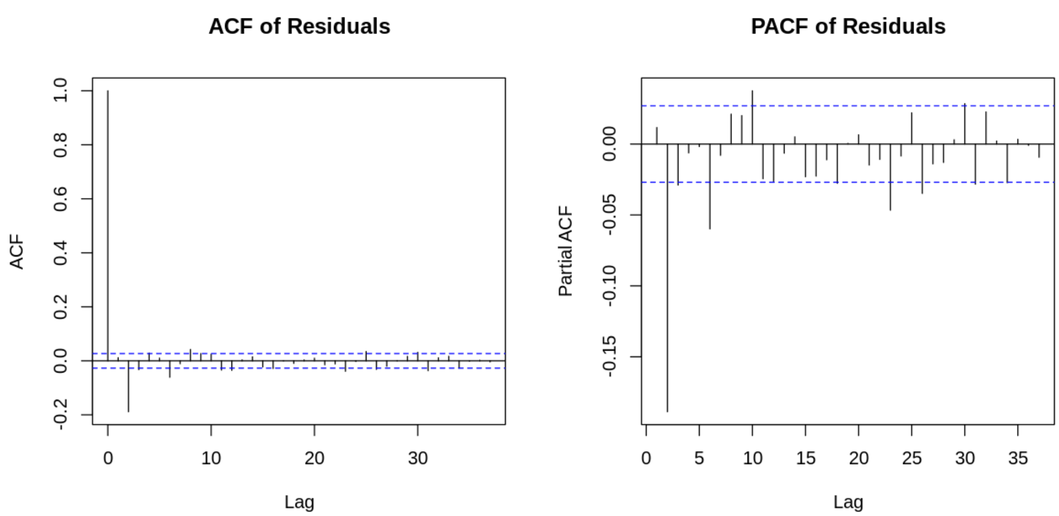

Figure 1. ACF and PACF

4.5. Bayesian Structural Time Series (BSTS)

Bayesian Structural Time Series (BSTS) is a modeling framework that uses Bayesian methods to estimate the parameters of a time series model and make predictions. The BSTS framework is a generalization of the classical state space models used in time series analysis, which allows for more flexible modeling of the time series data. In a BSTS model, the time series is decomposed into multiple components, including a local level, local trend, seasonality, and regression components. The model parameters are estimated using Bayesian methods, which allows for the incorporation of prior knowledge and uncertainty into the model. The posterior distribution of the model parameters is obtained using Markov Chain Monte Carlo (MCMC) methods, which allows for the calculation of Bayesian credible intervals for the parameter estimates.

Generally, we can write a Bayesian structural model like this:

$$Y_t = \mu_t + X_t + S_t + \varepsilon_t, \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

$$\mu_{t+1} = \mu_t + \nu_t , \nu_t \sim N(0, \sigma_\nu^2)$$

Where X_t denotes a set of regressors, S_t represents seasonality, and μ_t is the local level term. BSTS has some advantages over ARIMA model, specifically flexibility and forecasting accuracy. The BSTS framework allows for more flexible modeling of the time series data, with the ability to include multiple components such as trend, seasonality, and covariates. This can be useful when the time series data has complex patterns that cannot be adequately captured by a simple ARIMA model. In addition, the BSTS sometimes provide more accurate forecasts than ARIMA, especially when the time series has complex patterns. The bsts model iterates 10,000 times and I consider 10% as burn-in period.

5. Results

5.1. Understanding of the impact of each predictor

To find the impact of each predictor on the natural gas price, total 11 predictors are selected. VAR and BVAR are utilized to find the relationship between multiple time series variables. VAR method returns 5 predictors that are statistically significant – the previous natural gas price, the previous and 2step before heating oil price, the previous natural gas export, and the degree days as below figure.2. The previous heating oil price has positive correlation with the natural gas price by approximately 0.9, and the previous natural gas price has also positive correlation with the natural gas price by approximately 0.75.

```
              Estimate Std. Error t value Pr(>|t|)
y.l1          0.751169   0.067118  11.192  < 2e-16 ***
WTI.Price.l1 -0.034257   0.104810  -0.327 0.744017
HO.Price.l1   0.898338   0.329011   2.730 0.006717 **
Drill.Act.l1 -0.238200   0.186796  -1.275 0.203272
NG.prod.l1   -0.057932   0.226688  -0.256 0.798476
NG.cons.l1    0.050113   0.059953   0.836 0.403929
NG.store.l1  -0.214939   0.355430  -0.605 0.545836
NG.import.l1 -0.219205   0.135103  -1.623 0.105794
NG.export.l1  0.173786   0.061756   2.814 0.005230 **
GDP.l1        0.893532   0.979851   0.912 0.362584
DD.l1        -0.046179   0.019005  -2.430 0.015717 *
y.l2          0.045963   0.061918   0.742 0.458503
WTI.Price.l2  0.160706   0.116872   1.375 0.170188
HO.Price.l2  -0.724673   0.301903  -2.400 0.017016 *
Drill.Act.l2  0.052826   0.176137   0.300 0.764458
NG.prod.l2   -0.275457   0.202823  -1.358 0.175494
NG.cons.l2   -0.102533   0.063535  -1.614 0.107671
NG.store.l2  -0.098109   0.374573  -0.262 0.793568
NG.import.l2 -0.020713   0.130978  -0.158 0.874455
NG.export.l2 -0.030609   0.064143  -0.477 0.633588
GDP.l2        0.709890   1.005316   0.706 0.480676
DD.l2        -0.036233   0.019065  -1.901 0.058369 .
const        -1.636377   6.790781  -0.241 0.809750
trend        -0.005177   0.001527  -3.390 0.000798 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.1324 on 287 degrees of freedom
Multiple R-Squared: 0.9236,     Adjusted R-squared: 0.9175
F-statistic: 150.8 on 23 and 287 DF,  p-value: < 2.2e-16
```

Figure 2. VAR results

BVAR returns different results. It shows that the previous natural gas price is positive correlated with the current natural gas price by 0.89, but this time the previous heating oil has negative correlation with the natural gas price. Instead, the previous WTI price has the positive correlation with the current natural gas price. The results are as below.

```
Numeric array (dimensions 12, 11) of coefficient values from a BVAR.
Median values:
                    y WTI.Price HO.Price Drill.Act NG.prod NG.cons NG.store
constant        5.781     5.479   -1.740     0.588   4.894 -40.699    6.694
y-lag1          0.888    -0.036   -0.001     0.018   0.015   0.337   -0.017
WTI.Price-lag1  0.198     0.923    0.085     0.223  -0.021   0.058   -0.015
HO.Price-lag1  -0.242     0.217    0.878    -0.244  -0.001  -0.609    0.041
Drill.Act-lag1 -0.056    -0.096   -0.022     0.927  -0.010   0.193   -0.005
NG.prod-lag1   -0.255    -0.209    0.006     0.037   0.665   0.221   -0.055
NG.cons-lag1   -0.019    -0.012    0.017     0.001   0.005   1.076   -0.081
NG.store-lag1  -0.061    -0.089    0.051    -0.055  -0.039   2.847    0.675
NG.import-lag1  0.014     0.105    0.039    -0.006  -0.213  -0.504   -0.030
NG.export-lag1  0.079     0.030    0.007    -0.011   0.009   0.048    0.001
GDP-lag1       -0.167    -0.187    0.005    -0.018   0.339  -0.385    0.073
DD-lag1        -0.047    -0.007    0.001     0.008  -0.009   0.009    0.000
```

Figure 3. BVAR results

To find the performance of each model, I split the dataset 80% for training and the 20% for test. Statistical evaluation criteria MAE, MSE, RMSE, and MAPE are all computed, and the result is as below.

	VAR	BVAR
MAE	0.34	0.36
MSE	0.19	0.19
RMSE	0.44	0.44
MAPE	0.30	0.39

Table 2. Performance Evaluation

5.2. Forecasting the natural gas price

To build a forecasting model, the traditional ARIMA model and BSTS models are considered. The same 80% of the training dataset is used for the training and the 20% of the test dataset is used for the evaluation. For the ARIMA model, the autoregressive term (p =1) and the first order differencing (d=1) are selected. The grey area is confidence interval.

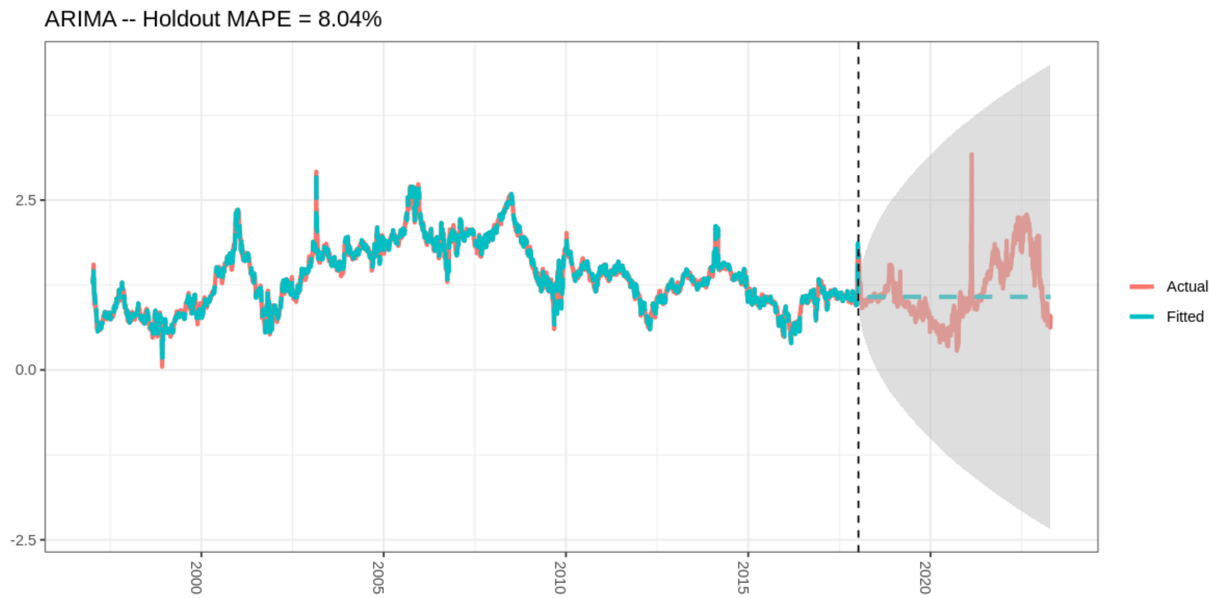

Figure 4. ARIMA model prediction

For BSTS model, the result is as below. The number of iterations was 10,000 and 10% of burin-in period was considered. The predicted value is the mean value of each time stamp and the BSTS library automatically computes the credible interval.

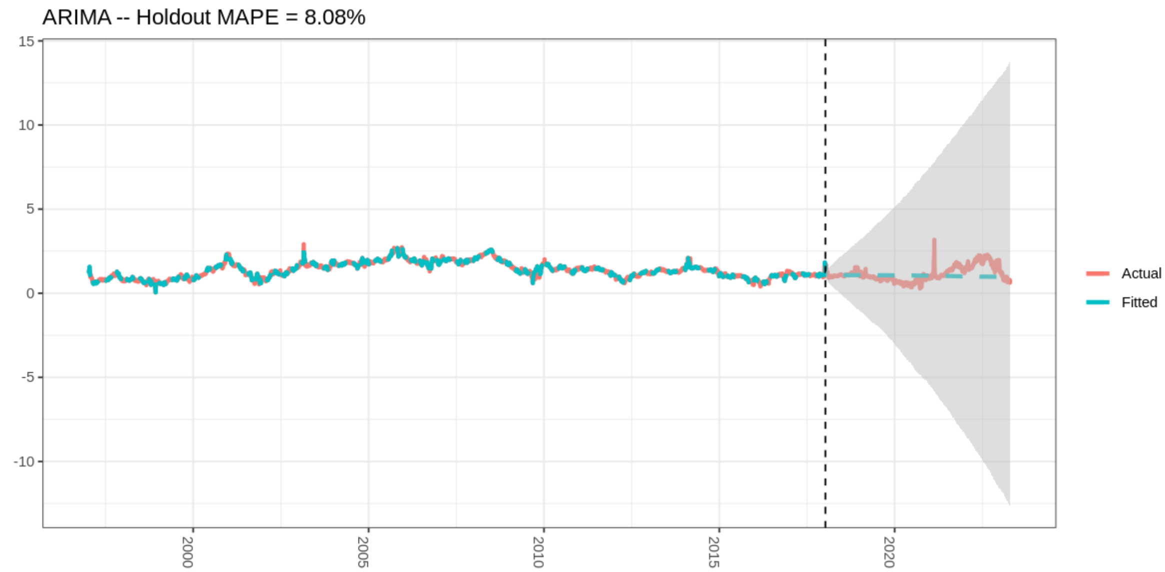

Figure 5. BSTS model prediction

The same performance evaluation criteria are applied for the forecasting models, and the result is as below.

	ARIMA	BSTS
MAE	0.09	0.09
MSE	0.04	0.04
RMSE	0.20	0.20
MAPE	0.08	0.08

Table 3. Performance Evaluation

The VAR and BVAR returns very similar evaluation values with each other. Compared to the similar study done by [8] (Moting Su et al.,), these values seem reasonable.

6. Conclusion and limitations

6.1. Summary and Conclusions

Natural gas is an essential energy resource in today's world, and it is vital for various stakeholders to understand its movements. The goal of this study is to comprehend the primary drivers of natural gas prices and construct a forecasting model using both frequentist and Bayesian methods. To determine the impact of each predictor, VAR and BVAR were used. The VAR analysis revealed that five factors significantly impact current natural gas prices, including the previous natural gas and heating oil prices, exports, degree days, and heating oil prices two time-stamps before. The previous heating oil price and natural gas price had the strongest correlation. The BVAR analysis returned slightly different results. The five most crucial predictors are the previous natural gas price, WTI crude oil price, heating oil price, natural gas production, and GDP. The VAR analysis had smaller evaluation criteria values than BVAR, indicating that the VAR results were more reliable.

To forecast natural gas prices, ARIMA and BSTS were used, with ARIMA producing smaller evaluation criteria values and a better confidence interval range at the same significant level of 0.05. Overall, the evaluation criteria values were acceptable when compared to similar studies. For both tasks, frequentist methods performed better than Bayesian methods. This could be because there was no prior distribution available for the natural gas price and other predictors used in this study, and all variables were assumed to have an uninformative prior. One advantage of Bayesian methods is having prior knowledge, but since no such knowledge was available in this case, the Bayesian method did not perform well.

6.2. limitations and future works

There are a few limitations in this study that should be acknowledged. Firstly, the dataset used is monthly, which may not capture the movement of gas prices precisely. Furthermore, some data have missing values, which could have a negative impact on the study's results. Secondly, as previously mentioned, the Bayesian method did not work as well because an uninformative prior distribution was assumed for all variables. For future studies, more advanced time series models could be used to improve predictions. Additionally, artificial neural network algorithms have recently shown great promise in time series prediction problems. Therefore, applying algorithms such as ANN or RNN (LSTM) to natural gas prediction could be explored in future research.

7. Bibliography

[1] EIA.(U.S. Energy Information Administration), What is U.S. electricity generation by energy source? https://www.eia.gov/

[2] EIA.(U.S. Energy Information Administration), Natural Gas and the environment, "Natural gas is a relatively clean burning fossil fuel", https://www.eia.gov/energyexplained/natural-gas

[3] EIA.(U.S. Energy Information Administration), Natural Gas and the environment, "Factors affecting natural gas prices", https://www.eia.gov/energyexplained/natural-gas/

[4] Jia Ding, Yuxuan Zhao, Junyang Jin, Forecasting natural gas consumption with multiple seasonal patterns, https://doi.org/10.1016/j.apenergy.2023.120911

[5] Jolanta Szoplik, Forecasting of natural gas consumption with artificial neural networks, http://dx.doi.org/10.1016/j.energy.2015.03.084

[6] Kexian Zhang and Min Hong, "Forecasting crude oil price using LSTM neural networks", DOI: 10.3934/DSFE.2022008

[7] Qihang Ma, "Comparison of ARIMA, ANN and LSTM for Stock Price Pre- diction", https://doi.org/10.1051/e3sconf/202021801026

[8] Moting Su, Zongyi Zhang, Ye Zhu, Donglan Zha and Wenying Wen, Data Driven Natural Gas Spot Price Prediction Models Using Machine Learning Methods, Energies 2019, 12, 1680; doi:10.3390/en12091680

[9] Xuejun Zhao, Application of Bayesian Regression Model in Financial Stock Market Forecasting, http://creativecommons.org/licenses/by-nc/4.0/.

[10] Sebastian Nicka, Stefan Thoenesb, What Drives Natural Gas Prices? – A Structural VAR Approach, Institute of Energy Economics at the University of Cologne (EWI)

[11] Dixon Domfeh, DETERMINANTS OF NATURAL GAS PRICES IN THE UNITED STATES – A STRUCTURAL VAR APPROACH, https://ssrn.com/abstract=3905560

[12] Matthew D. Hoffman, Andrew Gelman, The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, Journal of Machine Learning Research 15 (2014) 1351-1381

Appendix#1.

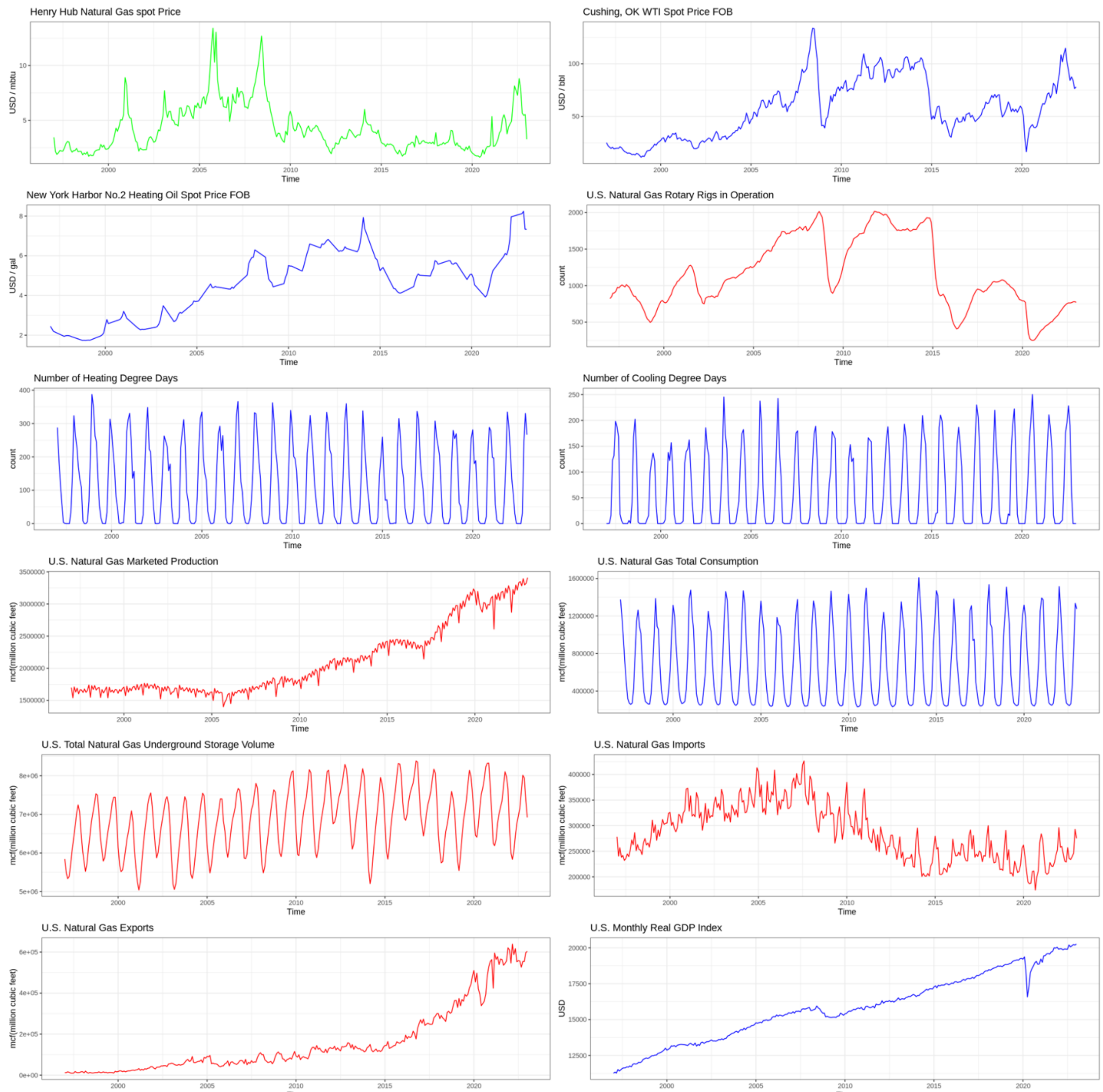

Figure 6. Dataset plot

Appendix#2

In [ ]:

```
# Install packages
install.packages('ggplot2')
install.packages('gridExtra')
install.packages("vars")
install.packages("astsa")
install.packages("googledrive")
install.packages('bayesplot')
install.packages('posterior')
install.packages('lubridate')
install.packages('bsts')
install.packages('forecast')
install.packages('dplyr')
install.packages('BVAR')

library(ggplot2)
library(gridExtra)
library(vars)
library(astsa)
library(googledrive)
library(readr)
library(bayesplot)
library(posterior)
library(lubridate)
library(bsts)
library(dplyr)
library(ggplot2)
library(forecast)
library(BVAR)
```

In [ ]:

```
# Import dataset
data = read.csv('/content/sample_data/dataset.csv')
head(data)

# Data Preprocessing
# Remove NA values
data = na.omit(data)

# Convert date type from chr to date
data$date = as.Date(data$date)
# summary of the data
summary(data)
```

In [ ]:

```
# Plot features and the response

options(repr.plot.width = 20, repr.plot.height = 20)
par(mfrow = c(4,3))

p1 = ggplot(data) +
        geom_line(aes(x = date, y = NG.Price)) +
        xlab("Time") +
        ylab("USD / mbtu") +
        ggtitle("Henry Hub Natural Gas spot Price") +
        theme_bw()

p2 = ggplot(data) +
        geom_line(aes(x = date, y = WTI.Price)) +
```

```
        xlab("Time") +
        ylab("USD / bbl") +
        ggtitle("Cushing, OK WTI Spot Price FOB") +
        theme_bw()

p3 = ggplot(data) +
        geom_line(aes(x = date, y = HO.Price)) +
        xlab("Time") +
        ylab("USD / gal") +
        ggtitle("New York Harbor No.2 Heating Oil Spot Price FOB ") +
        theme_bw()

p4 = ggplot(data) +
        geom_line(aes(x = date, y = Drill.Act)) +
        xlab("Time") +
        ylab("count") +
        ggtitle("U.S. Natural Gas Rotary Rigs in Operation") +
        theme_bw()

p5 = ggplot(data) +
        geom_line(aes(x = date, y = HTDD)) +
        xlab("Time") +
        ylab("count") +
        ggtitle("Number of Heating Degree Days") +
        theme_bw()

p6 = ggplot(data) +
        geom_line(aes(x = date, y = CLDD)) +
        xlab("Time") +
        ylab("count") +
        ggtitle("Number of Cooling Degree Days") +
        theme_bw()

p7 = ggplot(data) +
        geom_line(aes(x = date, y = NG.prod)) +
        xlab("Time") +
        ylab("mcf(million cubic feet)") +
        ggtitle("U.S. Natural Gas Marketed Production") +
        theme_bw()

p8 = ggplot(data) +
        geom_line(aes(x = date, y = NG.cons)) +
        xlab("Time") +
        ylab("mcf(million cubic feet)") +
        ggtitle("U.S. Natural Gas Total Consumption") +
        theme_bw()

p9 = ggplot(data) +
        geom_line(aes(x = date, y = NG.store)) +
        xlab("Time") +
        ylab("mcf(million cubic feet)") +
        ggtitle("U.S. Total Natural Gas Underground Storage Volume") +
        theme_bw()

p10 = ggplot(data) +
        geom_line(aes(x = date, y = NG.import)) +
        xlab("Time") +
        ylab("mcf(million cubic feet)") +
        ggtitle("U.S. Natural Gas Imports") +
        theme_bw()

p11 = ggplot(data) +
        geom_line(aes(x = date, y = NG.export)) +
        xlab("Time") +
```

```
        ylab("mcf(million cubic feet)") +
        ggtitle("U.S. Natural Gas Exports") +
        theme_bw()

p12 = ggplot(data) +
        geom_line(aes(x = date, y = GDP)) +
        xlab("Time") +
        ylab("USD") +
        ggtitle("U.S. Monthly Real GDP Index") +
        theme_bw()

grid.arrange(p1, p2, p3, p4, p5, p6, p7, p8, p9, p10, p11, p12, ncol = 2, wid
```

In [ ]:

```
# Heating Degree Days
htdd = matrix(data$HTDD, ncol = 12, byrow = TRUE)
htdd = htdd[1:nrow(htdd)-1,]

# Cooling Degree Days
cldd = matrix(data$CLDD, ncol = 12, byrow = TRUE)
cldd = cldd[1:nrow(cldd)-1,]

# Total number of Heating + Cooling degree days each month
dd = htdd + cldd

# Average of Heating + Cooling days per each month
dd_avg = colMeans(htdd) + colMeans(cldd)

# Create a dataset total degree days - average degree days per month
dd_diff = matrix(NA, nrow = 26, ncol = 12)
for (i in 1:nrow(dd)){
    for (j in 1:ncol(dd)){
        dd_diff[i,j] = dd[i,j] - dd_avg[j]
    }
}

# Append dd_diff in the dataset
data$DD[dim(data)[1]] = data$HTDD[dim(data)[1]] + data$CLDD[dim(data)[1]] - d
data$DD[1:dim(data[1])-1] = dd_diff

# Remove HTDD, CLDD and date columns
data_final = data[-c(1,6,7)]; head(data_final)

# Shift DD values to make all positives to take log

data_final$DD = data_final$DD - min(data_final$DD) + 1
min(data_final)

# Take log on the final dataset
data_log = log(data_final)
head(data_log)
```

In [ ]:

```
y = data_log$NG.Price # response. No scailing for the response
X = data_log[,2:ncol(data_log)] # predictors
```

In [ ]:

```
# Use AIC, BIC to find the optimal lag 'p'
final_data = cbind(y,X) # Final dataset after taking log and scailing.

# Separate the dataset, first 80% for training, last 20% for the test
train_index = dim(final_data)[1]*0.8

final_data_train = final_data[1:train_index,];
```

```
dim(final_data_train)
final_data_test = final_data[-(1:train_index),];
dim(final_data_test)

y_act = final_data[,1]; length(y_act) # actural y_value
y_train = y_act[1:train_index]; length(y_train)
y_test = y_act[train_index:length(y_act)]; length(y_test)

maxlag = 10  # maximum number of lags to consider

AIC_value = 0
BIC_value = 0

for (i in 1:maxlag){
    model = VAR(final_data, p = i, type = 'both')
    AIC_value[i] = AIC(model)
    BIC_value[i] = BIC(model)
}

# Find the order that minimizes AIC and BIC
AIC_min = which(AIC_value == min(AIC_value))
BIC_min = which(BIC_value == min(BIC_value))

# Print the optimal orders
cat("AIC optimal order: VAR(", AIC_min, ")", "\n")
cat("BIC optimal order: VAR(", BIC_min, ")", "\n")
```

In [ ]:

```
# USE VAR to find the correlation among predictors
fitvar = VAR(final_data, p=BIC_min, type = 'both')
summary(fitvar)

# # Forecasting
forecast = predict(fitvar, n.ahead = dim(final_data_test)[1])
y_pred = forecast$fcst$y[,1];length(y_pred)
```

In [ ]:

```
# MAE(Mean Absolute Error)
mae = sum(abs(y_test - y_pred)) / length(y_test)

# MSE(Mean Square Error)
mse = sum((y_test - y_pred)^2) / length(y_test)

# RMSE(Root Mean Square Error)
rmse = sqrt(mse)

# MAPE(Mean Absolute Percentage Error)
mape = sum(abs((y_test - y_pred)/y_test)) / length(y_test)

mae
mse
rmse
mape
```

In [ ]:

```
# BVAR
bvar_model = bvar(final_data, lags = 1, n_draw = 10000L, n_burn = 200L, verbo

# Print the summary of the model
summary(bvar_model)

# Forecasting
fitbvar = predict(bvar_model, bv_fcast(horizon = dim(final_data_test)[1]), da
```

```
y_pred2 = colMeans(fitbvar$fcast[,,1])
```

In [ ]:

```
# MAE(Mean Absolute Error)
mae2 = sum(abs(y_test - y_pred2)) / length(y_test)

# MSE(Mean Square Error)
mse2 = sum((y_test - y_pred2)^2) / length(y_test)

# RMSE(Root Mean Square Error)
rmse2 = sqrt(mse2)

# MAPE(Mean Absolute Percentage Error)
mape2 = sum(abs((y_test - y_pred2)/y_test)) / length(y_test)

mae2
mse2
rmse2
mape2
```

In [ ]:

```
# Technical Analysis with ARIMA
price_data = read.csv('/content/sample_data/henryhub.csv')
price_data = na.omit(price_data)
price_data$Date = as.Date(price_data$Date)
head(price_data)

index_train = dim(price_data)[1] * 0.8

price_act = price_data[,2];length(price_act)
price_train = price_data[1:index_train,2];length(price_train)
price_test = price_data[index_train:dim(price_data)[1],2];length(price_test)

# Logarithm
price_act = log(price_act)
price_train = log(price_train)
price_test = log(price_test)

# ARIMA (1,1,0)
arima = arima(price_train, order = c(1,1,0))

d1 = data.frame(c(as.numeric(fitted(arima)), as.numeric(predict(arima, n.ahea
  price_act, price_data$Date)
names(d1) <- c("Fitted", "Actual", "Date")

residuals = residuals(arima)

options(repr.plot.width = 10, repr.plot.height = 5)
par(mfrow = c(1,2))

# ACF and PACF plot
acf(residuals, main = "ACF of Residuals")
pacf(residuals, main = "PACF of Residuals")

p = predict(arima, n.ahead = length(price_test), interval="prediction", level

lb = p$pred - 1.96*p$se
ub = p$pred + 1.96*p$se

d1 = data.frame(c(as.numeric(fitted(arima)), as.numeric(predict(arima, n.ahea
  price_act, price_data$Date)
names(d1) <- c("Fitted", "Actual", "Date")
```

```
### 95% forecast credible interval
confidence_interval = cbind.data.frame(
  lb, ub, price_data$Date[index_train:dim(price_data)[1]])
names(confidence_interval) <- c("LL", "UL", "Date")

dd1 <- left_join(d1, confidence_interval, by="Date")

# MAE(Mean Absolute Error)
mae = sum(abs(dd1[,1] - dd1[,2])) / dim(dd1)[1]

# MSE(Mean Square Error)
mse = sum((dd1[,1] - dd1[,2])^2) / dim(dd1)[1]

# RMSE(Root Mean Square Error)
rmse = sqrt(mse)

# MAPE(Mean Absolute Percentage Error)
mape = sum(abs((dd1[,1] - dd1[,2])/dd1[,2])) / dim(dd1)[1]

mae
mse
rmse
mape

# Plot
ggplot(data=dd1, aes(x=Date)) +
  geom_line(aes(y=Actual, colour = "Actual"), size=1.2) +
  geom_line(aes(y=Fitted, colour = "Fitted"), size=1.2, linetype=2) +
  theme_bw() + theme(legend.title = element_blank()) +
  ylab("") + xlab("") +
  geom_vline(xintercept = price_data$Date[index_train], linetype = 2)+
  geom_ribbon(aes(ymin=LL, ymax=UL), fill="grey", alpha=0.5)+
   ggtitle(paste0("ARIMA -- Holdout MAPE = ", round(100*mape,2), "%")) +
  theme(axis.text.x=element_text(angle = -90, hjust = 0))
```

In [ ]:

```
# BSTS
ss = AddLocalLinearTrend(list(), price_train)
# ss = AddSeasonal(ss, price_train, nseasons = 365)
bsts.model = bsts(price_train, state.specification = ss, niter = 10000, ping

burn = SuggestBurn(0.1, bsts.model)

p2 = predict.bsts(bsts.model, horizon = length(price_test), burn = burn, quan

d2 <- data.frame(
    c(as.numeric(-colMeans(bsts.model$one.step.prediction.errors[-(1:burn),])
names(d2) <- c("Fitted", "Actual", "Date")

# 95% forecast credible interval
credible.interval <- cbind.data.frame(
  as.numeric(p2$interval[1,]),
  as.numeric(p2$interval[2,]),
  price_data$Date[index_train:length(price_act)])
names(credible.interval) <- c("LL", "UL", "Date")

dd2 <- left_join(d2, credible.interval, by="Date")

# MAE(Mean Absolute Error)
mae2 = sum(abs(dd2[,1] - dd2[,2])) / dim(dd2)[1]

# MSE(Mean Square Error)
mse2 = sum((dd2[,1] - dd2[,2])^2) / dim(dd2)[1]
```

```
# RMSE(Root Mean Square Error)
rmse2 = sqrt(mse2)

# MAPE(Mean Absolute Percentage Error)
mape2 = sum(abs((dd2[,1] - dd2[,2])/dd2[,2])) / dim(dd2)[1]

mae2
mse2
rmse2
mape2

# Plot
ggplot(data=dd2, aes(x=Date)) +
  geom_line(aes(y=Actual, colour = "Actual"), size=1.2) +
  geom_line(aes(y=Fitted, colour = "Fitted"), size=1.2, linetype=2) +
  theme_bw() + theme(legend.title = element_blank()) +
  ylab("") + xlab("") +
  geom_vline(xintercept = price_data$Date[index_train], linetype = 2)+
  geom_ribbon(aes(ymin=LL, ymax=UL), fill="grey", alpha=0.5)+
   ggtitle(paste0("ARIMA -- Holdout MAPE = ", round(100*mape2,2), "%")) +
  theme(axis.text.x=element_text(angle = -90, hjust = 0))
```