Time Series Analysis to Prices of Gold, Crude Oil, and Bitcoin

Manuel León Hoyos

Webster University

MATH 3220 - Data Mining Methods

December 19, 2017

Problem Description

The purpose of this project is to collect three time series datasets on the prices of crude oil, gold, and Bitcoin. Then, perform a time series analysis on these datasets using the R programming software and implementing the Autoregressive Integrated Moving Average (ARIMA) model to forecast future values of the series.

Background

Time series data is a collection of observations of a particular variable made chronologically. Preferably, the data should be large in size, with numerical observations, and preferably same time intervals. Time series analysis is the use of techniques to make sense of time series data. Some common uses of time series analysis include: to describe or monitor data, explain the behavior of the series, forecast future values, improve certain trends of the variable or variables analyzed (Brockwell-Davis, 2016; Hyndman, 2017).

Time series patters exhibited in the sustained movements of the variable of interest can be seasonal, cyclical or trend depending on the directions or waves the variable displays. The seasonal pattern represents increases or decreases of the variable due to the frequency of certain events that could be the season of the year, the weather, festivities, unexpected or sporadic major events, or natural disasters. An example of a seasonal pattern could be an increase of umbrella sales due to unusual high rain in the city. A trend involves a long-term increase or decrease of the variable analyzed. For example, there is an upward trend of world population which was 600 million in 1700 and over 7 billion in 2012 (Piketty, 2017). A cyclical pattern is seen when the data shows multiple combination of a rise and a fall. It is often referred in the field of economics as the ‘business cycle.’ The difference between cyclical patterns and seasonal patterns is that seasonal has constant length while cyclical has a variable length. Some uses of time series analysis include: interpretation, forecasting, hypothesis testing, trend analysis, control response, and simulations (Brockwell-Davis, 2016, Hyndman, 2017; Wikipedia, 2017).

Methodology

This project performs a time series analysis on three different time series datasets of prices. The first dataset is of Brent crude oil prices since 1986; The second, bitcoin prices since 2011; and third, gold prices since 1968. All of these financial datasets are available for download in csv format at the St. Louis Federal Reserve FRED database (https://fred.stlouisfed.org) and at the Quandl website (www.quandl.com).

The time series analysis is performed in the open source R programming software using the packages ‘forecast’, ‘ggplot2’, and ‘Quandl’ (R-project, 2017). Time series analysis is a growing field of practice and there are multiple other popular programs and packages. Other popular packages are the ‘xts’ and ‘zoo’. The purpose of these time series analyses is to have a better idea of the properties and behaviors of the selected financial datasets. The Methodology implements the autoregressive integrated moving average (ARIMA) model.

The ARIMA model is a generalization of an autoregressive moving average (ARMA) model. Both models are designed to help understand the data better and predict future values (Hyndman, 2017; Wikipedia, 2017). The model combines the autoregressive (AR) method with the moving average (MA) method. According to Bakar and Rosbi (2017), “An autoregressive (AR) model is a representation of a type of random process. It is used to describe certain time-varying processes in time series data. The autoregressive model specifies that the output variable depends linearly on its own previous values and on a stochastic term.”

As cited in Wikipedia, the autoregressive model, AR (p) refers to the autoregressive model of order p.

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t.$$

Where $\varphi_1, \ldots, \varphi_p$ are parameters, C is a constant, and the random variable ε_t is white noise.

In regards to the moving-average model, MA (q) refers to the moving average model of order q:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

Where the $\theta_1, \ldots, \theta_q$ are the parameters of the model, μ is the expectation of X_t and the $\varepsilon_t, \varepsilon_{t-1}$ are again white noise.

In regards to the ARMA model, ARMA(p, q) refers to the model with p autoregressive terms and q moving-average terms. This model contains the AR(p) and MA(q) models,

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}.$$

These models are helpful not only in forecasting but also in predicting the range of the prediction estimates. By analyzing the volatility of changes in the data the model, it helps understand what are the ranges plausible for the variable.

In terms of growth in GDP (Gross Domestic Product), institutions such as the International Monetary Fund and the Federal Reserve (Central Bank of the United States), usually release monthly reports and give an estimate of either increasing or decreasing GDP growth rates.

Some of the expected results for these time series analyses include: first, a better description of the datasets in terms of patterns and behaviors, and second, an understanding of the models for volatility, variance, and forecasting. With the ARIMA model the expectation is to first detect the patterns of the data in order to predict future values.

It is important to point out that prices of oil, gold, and bitcoin, are dependent on multiple external factors. At the same time, changes of prices on any of these subjects may have direct effects on other variables. For example, the prices of oil are subject to supply and demand, but also politics. The Organization of the Petroleum Exporting Countries (OPEC) holds regular meetings in which their members decide how much oil to produce or at what prices they should sell for. For example, in 1973, OPEC proclaimed an oil embargo which resulted in high inflation in multiple countries, including the United States (Office of the Historian, 2017). Another key factor that could reflect a decrease in demand for oil could improvements in renewable energy. If renewal energy becomes affordable, it may replace gasoline and bring the prices of oil down due to substitution.

In the case of gold, this precious metal is considered a reserve currency. Often, when people are uncertain about fiat currencies, the prices of gold increase abruptly. According to Ewing and Malik (2013), gold has traditionally served as a hedge against inflation. Melvin and Sultan (1990) concluded that oil price changes and political unrest are important determinants of volatility in gold prices. In recent decades, the governments of China and Russia have been notorious for accumulating large reserves of gold (Kirkulak, & Lkhamazhapov, 2017). Global demand of gold resulted in an increase of gold prices (Kirkulak, & Lkhamazhapov, 2016).

In the case of bitcoin, this crypto-currency has merely over five years in circulation. It is a decentralized type of currency and a “digital asset designed to work as a medium of exchange using cyptography to secure transactions, to control the creation of additional units, and to verify the transfer of assets” (Bakar & Rosbi, 2017). Demand for bitcoin only started to rise significantly this year. In May 2017, bitcoin was trading for less than $2,000 and by November 2017, bitcoin had surpassed the $8,000 mark. Bitcoin is not only a new type of currency but also an innovative technology. It is also running without significant government regulation. Therefore, it is not only possible that governments step in or improved technologies replace bitcoin.

In forecasting, it important to take into consideration the uncertainty of the future and that it is easier to forecast in the short term than in the long term. According to Hyndman (2014), “it is invalid to look at how well a model fits the historical data; the accuracy of forecasts can only be determined by considering how well a model performs on new data that were not used when estimating the model.

Assumptions

During the project it was assumed that all the numerical values for the observations in the datasets were correct.

Experimental Design

This project performed a time series analysis using the R programming software using the packages ‘forecat’, ‘ggplot2’, and ‘Quandl’ of prices of crude oil, gold, and bitcoin. First, by installing the R packages ‘forecast’, ‘ggplot2’, and ‘Quandl’. The datasets of crude oil, gold, and bitcoin prices used for this project were obtained in csv format from the website www.quandl.com. While conducting the analysis, it appeared that the datasets of crude oil and gold obtained from the St. Louis FRED database contained multiple missing values, therefore, the datasets obtained by Quandl were given preference. By using the package ‘Quandl’ in R it is possible to load datasets without downloading from the website. However, the Quandl code is needed for each dataset to then execute the command Quandl(“Code of Dataset”) to load the datasets. Also, before loading the datasets in the R global environment, the datasets were constrained to only two columns, dates and prices (at close). This, because the datasets contained multiple columns for prices such as prices at opening, close, and highest or lowest in the day.

After loading the datasets in the R environment, the time series objects were created for each dataset using the ts() command. In this case, the dataset for oil was named “oil”, for gold “gold”, and for bitcoin “bc”. The time series objects instead: “tsoil”, “tsgold”, and “tsbc”. It is important to set the objects as time-series datasets, as R does not immediately recognize the nature of the datasets even though there is a column of pure dates. Also, while conducting the time series data and setting the “frequency” it is important to note that while Bitcoin has a frequency of 365 as it is traded every single day, crude oil and gold do not trade everyday and have approximately 255 days of trading per year. Therefore, the “frequency” needs to be controlled. The Then, training datasets are created by taking out an end portion of the datasets. The training dataset should be larger than the test datasets to be compared. The window() command serves to create training and test datasets. The which.max() shows the number of the observations with the highest value or outlier. In this case, the highest values were for Bitcoin $19,187 on December 16, 2017; for gold $1,895 on May 5, 2011; and oil $145 on July 14, 2008.

For the ARIMA model in this project, the R code performed was auto.arima() which is a Hyndman-Khandakar algorithm (Hyndman, 2017). The purpose of this algorithm is to select the best possible ARIMA model after analyzing the given data. To run the auto.arima() the datasets for gold and oil were constrained to cover the years from 2000 to 2017 without the last 30 days. While for bitcoin, the dataset was constrained to only the observations of 2017 and 12 days prior.

Results and Discussion

In this project, it was evidenced that gold and oil prices possess much rich history than Bitcoin. While gold contains 12,574 observations since 1968 and oil 8,724 since 1983, bitcoin contains only 2,289 with the largest volatility for bitcoin in the last 100 observations. The upward trend of December of 2017 makes Bitcoin even harder to forecast as it never had any history of passing the $10,000 benchmark and almost no downturns. The day prior to this presentation (December 18, 2017) Bitcoin was valued at over $19,000. While gold and oil reflected results more similar to recent trends. The results showed an almost exponential rise of Bitcoin prices while for gold and oil the possible range of change was much smaller.

Issues

Throughout this project, several resources were used in addition to books and journal articles. There are multiple websites that offer online tutorials such as *Data Camp, Coursera*, and *edX*. Also, the *Stack Overflow* website proved to be of great utility as during the project several questions were asked in the forums regarding issues running the R code. Often, responses were received within 24 hours.

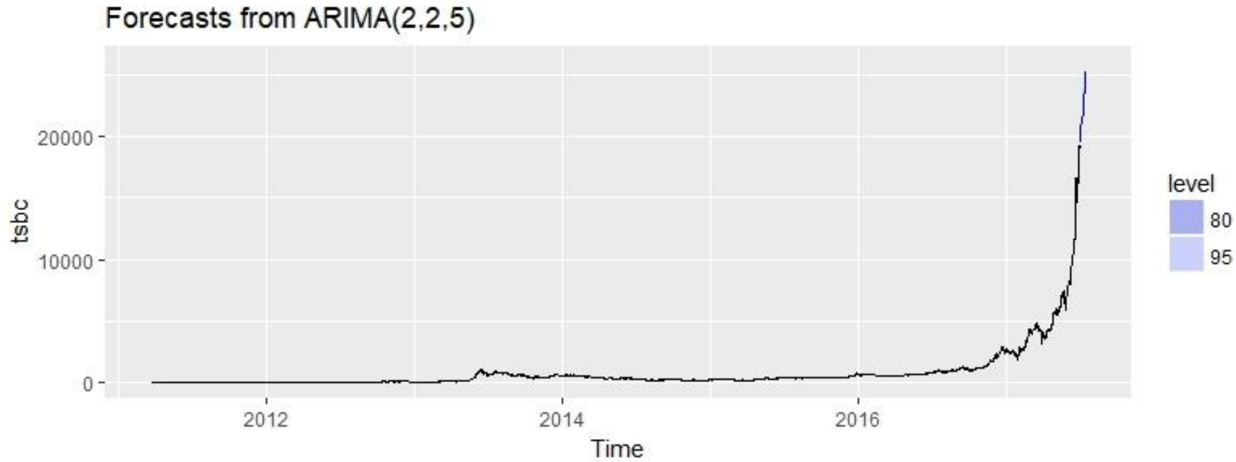
Conclusions and Future Work

While gold and oil prices in some way are still volatile they are more reliable than Bitcoin in a sense that variance of their fluctuations are much smaller. In this specific case of forecasting Bitcoin prices, the degree of uncertainty is much higher because Bitcoin has not showed any visible downward trends yet.

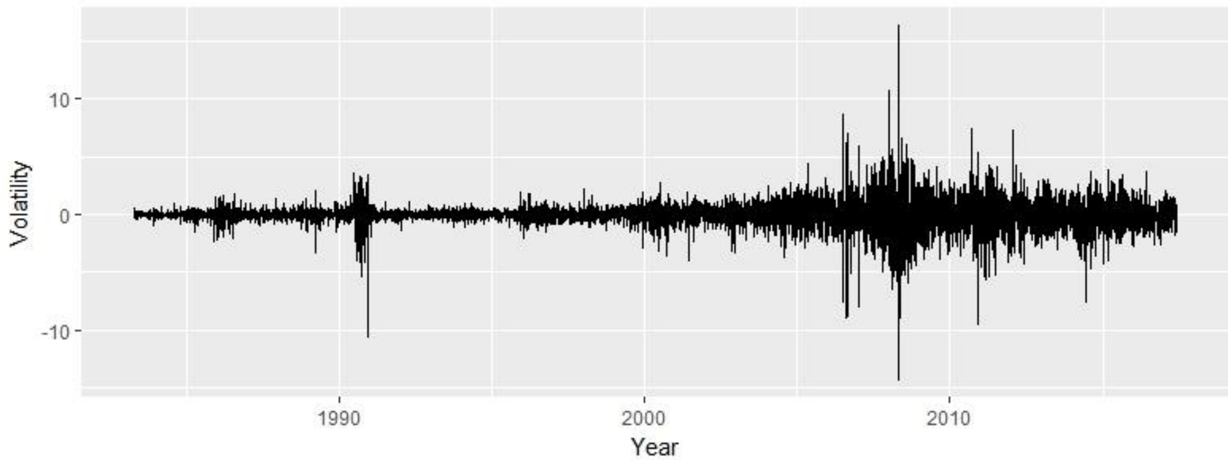
Future work is aimed to learn other forecasting models in addition to the ARIMA and ARMA models and conduct time series analyses including more than one variable of observations. Regressions are important to understand possible relationships between variables.

Appendices

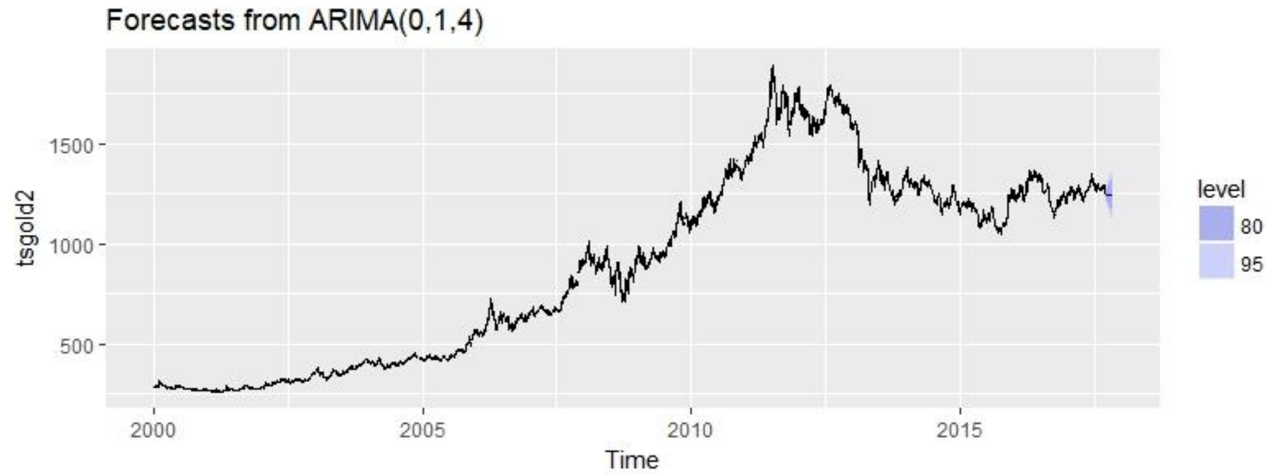
Forecast of Bitcoin including prices for December 18, 2017 shows Bitcoin passing the $25,000 mark in 12 days.



Volatility of oil prices.



Forecast of gold prices for a period of 30 days shows a range of values.



Bibliography

Bakar, N. A., & Rosbi, S. Autoregressive Integrated Moving Average (ARIMA) Model for Forecasting Cryptocurrency Exchange Rate in High Volatility Environment: A New Insight of Bitcoin Transaction. https://dx.doi.org/10.22161/ijaers.4.11.20

Bennett, R. & Hugen, D. (2016). Financial Analytics with R. Cambridge University Press.

Brockwell-Davis. (2016). Introduction to Time Series and Forecasting. Springer.

Ewing, B. T., & Malik, F. (2013). Volatility transmission between gold and oil futures under structural breaks. International Review of Economics & Finance, 25, 113-121.

Hyndman, R. (2017). Forecasting Using R. DataCamp. Retrieved from: https://campus.datacamp.com/courses/forecasting-using-r/exploring-and-visualizing-time-series-in-r

Hyndman, R. J., & Khandakar, Y. (2007). *Automatic time series for forecasting: the forecast package for R* (No. 6/07). Monash University, Department of Econometrics and Business Statistics.

Hyndman, R. J., & Koehler, A. B. (2014). Measuring forecast accuracy.

Kirkulak, B., & Lkhamazhapov, Z. (2016). The Volatility Dynamics of Spot and Futures Gold Prices: Evidence from Russia.

Kirkulak, B., & Lkhamazhapov, Z. (2017). Volatility Dynamics of Precious Metals: Evidence from Russia. Czech Journal of Economics and Finance.

Melvin, M., & Sultan, J. (1990). South African political unrest, oil prices, and the time varying risk premium in the gold futures market. Journal of Futures Markets, 10(2), 103-111.

Office of the Historian. (2017). Oil Embargo, 1973–1974. Retrieved from: https://history.state.gov/milestones/1969-1976/oil-embargo

Piketty, T. (2017). *Capital in the twenty-first century*. Harvard University Press.

Quandl. (2017). Financial datasets. Retrieved from: www.quandl.com

R-project. (2017). Package 'forecast'. Retrieved from: https://cran.r-project.org/web/packages/forecast/forecast.pdf

Singh, A. & Allen, D (2017). R in Finance and Economics A Beginner's Guide. World Scientific.

St. Louis Federal Reserve. (2017). FRED Economic Research Data. Retrieved from: https://fred.stlouisfed.org

Wikipedia. (2017). Time Series. Retrieved from: https://en.wikipedia.org/wiki/Time_series

Wikipedia. (2017). Stochastic Process. Retrieved from: https://en.wikipedia.org/wiki/Stochastic_process

Wikipedia. (2017). Autoregressive integrated moving average. Retrieved from: https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average

Woodward, W. A., Gray, H. L., & Elliott, A. C. (2017). *Applied Time Series Analysis with R*. CRC press.