Big Data Continual Assessment

Time Series Prediction for Climate Change with ARIMA

De Jong Yeong

Computing Department, Institute of Technology, Tralee, Dromtacker, Tralee, Co. Kerry, Ireland

Abstract

A climate change: earth surface temperature data sourced from Kaggle, originally from Berkeley Earth containing several data source files were used to perform prediction and forecasting on climate change including visualizations on the data. Two data source files were selected for data visualization and climate change prediction and forecasting process. First the GlobalLandTemperaturesByCountry.csv file was analyzed by converting the data index to a datetime index and uses the visualization and descriptive techniques on the average land temperature and country. The descriptive and visualization techniques were also used to analyze the GlobalTemperatures.csv data source file. Predictive and forecasting analysis was performed only on the land and ocean average temperature data column in the global temperature data source file, using the Autoregressive Integrated Moving Average (ARIMA) model.

Keywords: ARIMA (Autoregressive Integrated Moving Average); Time-Series Prediction; Time-Series Forecast

1. Introduction

Mark Kaufman stated that the planet earth is the warmest that has been recorded in about 120,000 years and NASA announced that July was the third warmest month in 2018 (Kaufman, 2018). Climate change, also known as global warming disrupts national economies and health of the nation like allergy and asthma due to pollution and causes an imbalance in an ecosystem itself (Rosenberg, 2018). The predicted and forecasted values can be treated as an alert to bring climate action top of the international agenda. The climate change datasets were originally compiled by Berkeley Earth and was repackaged by Kaggle. The datasets consist of 5 distinct source files and only the GlobalLandTemperaturesByCountry.csv file and GlobalTemperatures.csv file was chosen. The GloabalTemperatures.csv file contains 3193 rows and 9 columns whereas the GlobalLandTemperaturesByCountry.csv file contains 577462 rows and 4 columns. The research should perform descriptive and visualizations of data in both selected datasets and should predict future global land and ocean average temperature column in GlobalTemperatures.csv file (Kaggle, 2017).

2. Literature Review

An accelerated warming trend began in the early 1980s due to influence of carbon dioxide (CO_2) and was gradually increased during the twentieth century. The benefits of time series prediction and forecasting of climate change help to bring climate action to the top of the international agenda. Climate forecasts are valuable to societies in which the forecasted values are intended to help societies to cope better with climate variations (Stern, Paul C.; Easterling, William E.;, 1999). A combination of visualizations, descriptive analysis, and the predictive and forecasting analysis on global average temperature would aid climate scientists in discovering the climate pattern. Predicting and forecasting global land and ocean future temperature could vastly benefit societies on planning suitable climate change action and strategy and aids to bring climate action to top of the international agenda. Analysis of the global temperature by country dataset concluded that countries in the Africa continent was the warmest and the analysis of the global temperature dataset concluded that the global temperature was gradually increasing in the twentieth century. This paper aims to identify the climate pattern and using the time series data mining techniques model to predict and forecast future average global temperature of land and ocean.

3. Methods

The KDD (Knowledge Discovery in Database) process was utilized to perform visualization and descriptive techniques on both the GlobalLandTemperaturesByCountry.csv and GlobalTemperatures.csv data source files and to accomplish the predictive and forecasting analysis on the global land and ocean average temperature data column in GlobalTemperatures.csv file. The KDD process includes the selection, preprocessing and data transformation process. Time-series prediction and forecasting techniques such as the Autoregressive Integrated Moving Average (ARIMA) model was implemented for time series prediction and forecasting of the data.

3.1. Selection & Pre-processing

The selected datasets for the KDD process analysis were the Berkeley Earth's datasets which contains numerous records about the Earth's surface temperature (berkeleyearth.org, 2012)and were repackaged by Kaggle and is available in 2016. It contains five distinct data source files and only the GlobalLandTemperatureByCountry.csv file and GlobalTemperatures.csv file were selected from Kaggle website: https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data. The GlobalTemperatures.csv file that will be used for both descriptive analysis, data visualizations, and predictive and forecast analysis contains information about the date since 1750, minimum and maximum land temperature, global land temperature and global land and ocean average temperature and its corresponding 95% confidence interval. The CSV file was read into a pandas data frame and a total of 100 null objects (1750-1849) were removed. The date column was set as an index of the data frame.

The GlobalLandTemperatureByCountry.csv dataset that will be used only for descriptive analysis and visualizations of data contains information about the date, average temperature, 95% confidence interval (CI) around the average, and country. The CSV file was read into a pandas data frame, and the 95% confidence interval around the average column was removed. The continent will not be considered during data visualization process and were removed from dataset along with duplicated countries. The date column was set as an index of the data frame and the country column was set as the column index. The data frame was interpolated to estimate temperature within two known values in a sequence of values across each column for handling null values. Null values that were not handled with the use of interpolated method will be ignored during data visualization process.

3.2. Transformation

The selected .csv files were read into a pandas frame and data pre-processing techniques were executed on the pandas data frame which exists in persistent memory of Jupyter Notebook.

3.3. Data Mining - Descriptive Analysis and Data Visualizations

The selected datasets were read in with the use of *read_csv(filename)* function in the pandas module to a pandas data frame for data visualization process and descriptive analysis. Size of the selected datasets in terms of total number of rows and columns were determine with the *df.shape* function. The GlobalLandTemperatureByCountry.csv dataset consists of 3,239 total rows and 232 columns whereas the GlobalTemperatures.csv dataset consists of 1,992 total rows and 8 columns after data-processing process. The *df.head(5)* function was executed to display the first 5 rows of the datasets.

3.3.1. Monthly Average Land Temperature in Ireland and Malaysia and Mean Temperature in each Country

A comparative analysis on the monthly average land temperature in Ireland and Malaysia was illustrated on line plot. The line plot in figure 1(a) illustrates the average land temperature in Ireland and Malaysia. The difference of average land temperature in Malaysia was insignificant, in which the highest temperature was 23.86°C and the lowest was 28.26°C. However, the average land temperature in Ireland shows significant difference, in which the highest temperature was 17.98°C and the lowest was 0.27°C. The mean temperature in each country was computed, ignoring null values and was plotted as seen in the map chart in figure 1(b). Greenland has the lowest temperature of -18.59°C and countries lies within the intertropical zone has the highest temperature.

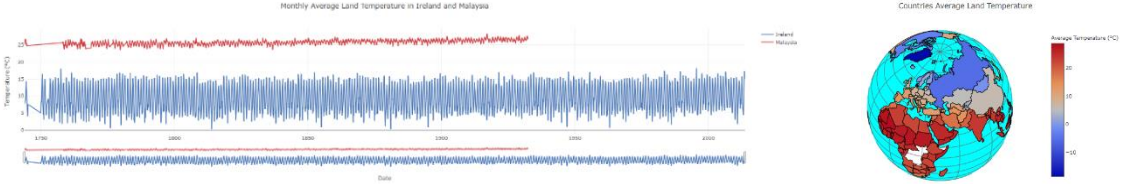


Fig. 1. (a)(b) Left - Monthly Average Temperature in Ireland and Malaysia, Right - Countries Mean Average Land Temperature in Map Chart

3.3.2. Distribution of Data in GlobalTemperatures.csv dataset

An analysis on the distributions of selected dataset was plotted on a boxplot with the use of Plotly visualization tool. The boxplot in figure 3 illustrates the distributions of data in each column of the dataset. It shows information about the minimum, lower quartile, upper quartile, inter-quartile range (IQR) and information on the whiskers. No outliers were shown and were positively skewed on each feature/column.

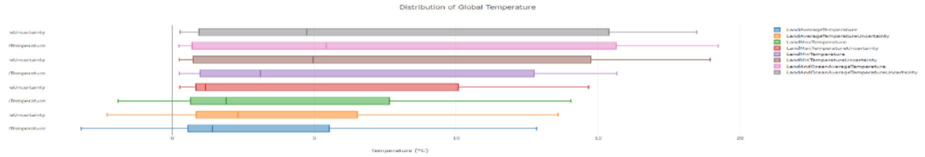


Fig. 3. Distribution of Data in GlobalTemperatures.csv dataset.

3.3.3. Global Land Temperature & Average Temperature of Land and Ocean

An analysis on the climate pattern of land temperature and average temperature of land and ocean features in GlobalTemperatures.csv dataset was plotted on a scatter line plot with Plotly visualization tool. The line plot in figure 4(a) illustrates the climate pattern of global land temperature from 1950 to 2015 with its calculated upper and lower limit using its uncertainty value shown in the dataset. The line plot in figure 4(b) illustrates the climate pattern of global land and ocean average temperature from 1950 to 2015 with its calculated upper and lower limit using its uncertainty value available in the dataset. The temperature in both line plot shows a seasonal fluctuation.

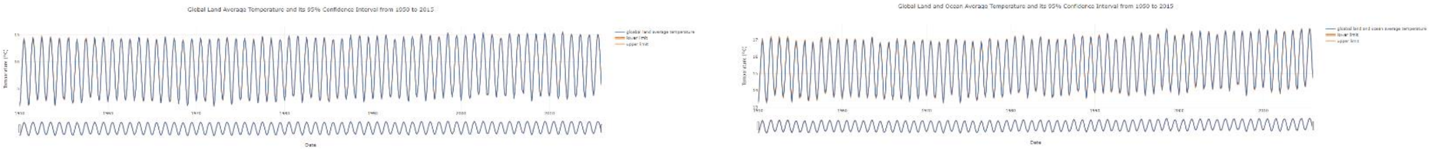


Fig. 4. (a)(b) Left - Global Land Temperature and its 95% CI. Right - Global Land and Ocean Average Temperature and its 95% CI

3.3.4. Global Land and Ocean Average Temperature Time Series Decomposition

The scatter line plot in figure 5 illustrates the time series decomposition of global land and ocean average temperature feature in the GlobalTemperatures.csv dataset. The LandAndOceanAverageTemperature feature and date instances will only be utilized for predictive and forecasting analysis. The *seasonal_decompose(dataframe, model)* in statesmodel module is used to decompose the time-series, this helps to deconstruct time series into trend, seasonality and noise which each represents the underlying categories of patterns (Brownlee, 2017). It can be seen from the scatter line plot that the data have a seasonal fluctuation, random residuals and a gently fluttered trend.

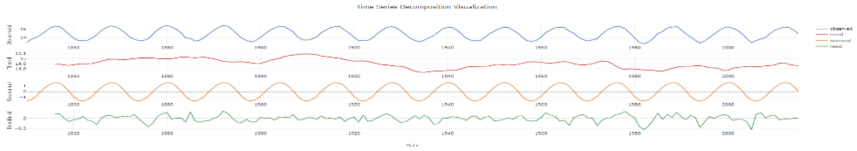


Fig. 5. Global Land and Ocean Average Temperature Time Series Decomposition

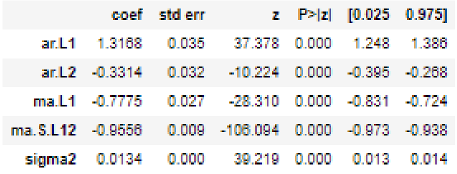
3.4. Data Mining - Time Series Prediction and Forecasting

Time series prediction was executed on the land and ocean average temperature feature in GlobalTemperatures.csv dataset with the use of ARIMA model. The data frame was preprocessed to only include the selected feature and date instance and the yearly average of temperature. The preprocessed data frame contains the time-series data was split into 80% training set and 20% testing set for model evaluation. The train set contains 1593 rows of data and the test set contains 399 rows of data. The 'grid search' method in conjunction with the train set was used for exploring the different combinations of ARIMA parameters and uses the combination of parameters to fit into a new seasonal model with the *SARIMAX()* function in the *statsmodels* module. The lowest Akaike Information Criterion (AIC) was identified and is fitted into an ARIMA time series model.

The predicted values and its associated CI of the forecasted value is retrieved with the *get_prediction()* and *conf_int()* attribute. The predicted values were compared with observed values to help understanding of the forecasted accuracy. The *get_forecast()* attribute is used to compute forecasted values for a specified number of steps ahead. The value of root mean squared error (RMSE) was calculated to determine the absolute fit of the ARIMA model.

4. Results/Data Findings/Discussion

The 'grid search' method returns list of possible combinations and an AIC value. It was indicated that the SARIMAX(2, 0, 1)x(0, 1, 1, 12) yields the lowest AIC value of -2203.6924056. The *summary* attribute is used to return information on the results of the output as shown in figure 6(a). It can be seen from the summary report that each weight has a p-value of lower than the common significance levels of 0.05, 0.1 and 0.01 (admin, 2018). An analysis on the model diagnostics were plotted with the *plot_diagnostics()* method as shown in figure 6(b). It can be seen from the plotted graph that the model residuals were normally distributed and have a low correlation with its lagged version.



	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3168	0.035	37.378	0.000	1.248	1.386
ar.L2	-0.3314	0.032	-10.224	0.000	-0.395	-0.268
ma.L1	-0.7775	0.027	-28.310	0.000	-0.831	-0.724
ma.S.L12	-0.9556	0.009	-106.094	0.000	-0.973	-0.938
sigma2	0.0134	0.000	39.219	0.000	0.013	0.014

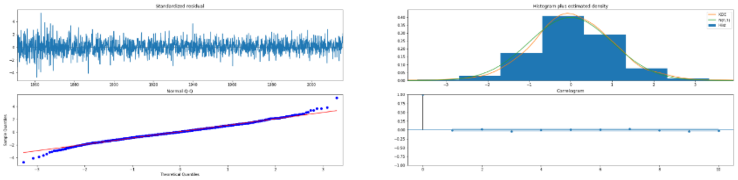


Fig. 6. (a)(b) Left - Summary information of ARIMA model. Right - ARIMA model diagnostics visualization.

Forecast of the global land and ocean average temperature was set to start at 2000-01-01 to 2015-12-01 and the predicted temperature was compared to observed value of the time series to better understand of the predicted values. A comparison of the monthly predicted values and observed values were plotted on line plot as shown in figure 7(a). The predicted values were aligned with observed values very well. The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) was calculated to determine the fitness of the model. An MSE or RMSE of 0 indicates an absolute fitness of the model. The predicted values obtained yield an MSE of 0.01032 and RMSE of 0.10161. The predicted monthly average temperatures from 2016-01-01 to 2020-12-01 were plotted on a line plot as shown in figure 7(b). The result shows a seasonal fluctuated of the forecasted global land ocean average temperature. This is because the global average temperature was recorded based on seasonal temperature in each country. Countries that are located within the intertropical zone have higher temperature compared to temperature of other countries. It was expected to have a gradually increasing predicted temperatures, but the predicted temperatures shown in figure 7(b) is gradually decreasing in January from 2016 to 2020. This is because the recorded average land and ocean temperature is affected by the seasonal temperature in each country, e.g. temperature during Spring, Summer, Autumn and Winter. Some countries may have lower temperature throughout the four seasons, e.g. Greenland, while countries that are located within the intertropical zone may have higher temperature throughout the four seasons, e.g. Africa.

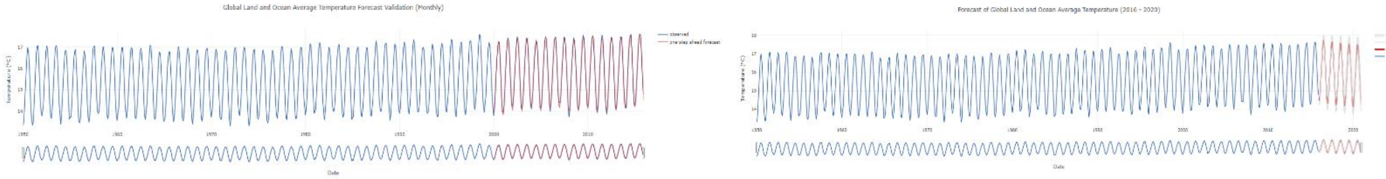


Fig. 7. (a) (b) Left - Comparison of predicted and true differenced values. Right - Forecast average temperature from 2016 to 2020.

ARIMA is a time series technique in which the time series historical data is used as an explanatory variable for time series prediction. It combines autoregression method to fit current data point to linear function of some data points and moving average method to retrieve the average of sum of several consecutive data points. It uses the output of the combination method to calculate the estimations of future value (gandalf, 2018). ARIMA model works on an assumption that properties of series do not depend on time when series is captured, and time series data should be univariate (Jassika, 2017). ARIMA model is hard to automate and is large amount of data is required to provide more accurate prediction result. It provides high interpretability and realistic confidence intervals for time series predictive analysis (Siva, 2018).

5. Conclusion

Climate change remains as a global issue and the global land and ocean average temperature is gradually increasing. ARIMA time series prediction model offers an insight for planning the suitable climate change action and strategy and aids to bring climate action to top of the international agenda. The result of the forecast validation yields an MSE of 0.01032 and an RMSE of 0.1016. It was concluded that the global land and ocean average temperature from 2016 to 2020 is forecasted to have seasonal fluctuated temperature. However, it was expected to have a gradually increasing predicted temperature due to an increase in emission of CO_2. Seasonality of the selected dataset for time series prediction can be examined and removed that causes data to be non-stationary which violates the assumptions of the ARIMA model before executing time series prediction in future work.

6. Source of Evidence

A GitHub repository was created containing Plotly API key text file, data source files and Python notebook of the predicted results can be seen at the following link: https://github.com/dejongyeong/climate-change-big-data. The implementation of ARIMA model for time series prediction and forecasting were executed on Azure notebook and was uploaded to GitHub after successful implementation.

References

admin, 2018. *Here is How to Interpret a P-Value of 0.000.* [Online]
Available at: http://www.statology.org/here-is-how-to-interpret-a-p-value-of-0-000/
[Accessed 22 April 2019].

Brownlee, J., 2017. *How to Decompose Time Series Data into Trend and Seasonality.* [Online]
Available at: https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/
[Accessed 22 April 2019].

Brownlee, J., 2017. *How to Difference a Time Series Dataset with Python.* [Online]
Available at: https://machinelearningmastery.com/difference-time-series-dataset-python/
[Accessed 23 April 2019].

gandalf, 2018. *Advantage and disadvantage of ARIMA methodology?.* [Online]
Available at:
https://answers.yahoo.com/question/index?qid=1006030814106&guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAMKNLXl4Q7PIAw2jjVSO5uSr0NJlFMxn2raf0Y1G5BEaBq2ZqV3DZqKzj7QNhiqHeoeYwMR21J2tEJVEsJ8lOkLjfeFQy692XljJDcbM2PH9GPhvLAd
[Accessed 24 April 2019].

Jassika, 2017. *Arima model.* [Online]
Available at: https://www.slideshare.net/rekhagill/arima-model-71750711
[Accessed 23 April 2019].

Kaggle, 2017. *Climate Change: Earth Surface Temperature Data.* [Online]
Available at: https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data#GlobalTemperatures.csv
[Accessed 21 April 2019].

Kaufman, M., 2018. *Earth is the warmest it's been in 120,000 years.* [Online]
Available at: https://mashable.com/article/earth-warmest-temperatures-climate-change/?europe=true
[Accessed 21 April 2019].

Murchú, C. Ó., 2018. *What is climate change and what causes it?.* [Online]
Available at: https://spunout.ie/opinion/article/climate-change-what-causes-it
[Accessed 21 April 2019].

Rosenberg, M., 2018. *Advantages and Disadvantages of Global Warming.* [Online]
Available at: https://www.thoughtco.com/advantages-and-disadvantages-of-global-warming-1434937
[Accessed 21 April 2019].

Siva, C., 2018. *Time Series Forecasting.* [Online]
Available at: https://dzone.com/articles/time-series-forecasting
[Accessed 23 April 2019].

Stern, Paul C.; Easterling, William E.;, 1999. *Making Climate Forecasts Matter.* 1st ed. Washington: National Research Council.