**ANL 488 PROJECT PROPOSAL**

**Time series forecasting of Greenhouse Gas Production over time**



**Submitted by**

**Bryan Lim Tze Yuan**

**SCHOOL OF BUSINESS Singapore University of Social Sciences**

**Presented to Singapore University of Social Sciences in partial fulfillment of the requirements for the**

**Degree of Bachelor of Science in Business Analytics**

**2022**

**TABLE OF CONTENTS**

Table of Contents

[**Chapter 1.0** **Introduction** 3](#_Toc111452058)

[**Chapter 2.0** **Literature Review** 4](#_Toc111452059)

[**Chapter 3.0** **Data Understanding and Preparation** 7](#_Toc111452060)

[**Chapter 3.1** **Data Exploration** 8](#_Toc111452061)

[**Chapter 3.2** **Data Preparation** 9](#_Toc111452062)

[**Chapter 4.0** **Proposed Modelling and Evaluation** 15](#_Toc111452063)

[**Chapter 5.0** **Proposed Schedule** 17](#_Toc111452064)

[**References** 19](#_Toc111452065)

# **Chapter 1.0 Introduction**

With greenhouse gas (GHG) emissions being on an unprecedented rise in the recent decades, the earth is now the hottest that it’s ever been, with an average increase of 0.85 °C of land and ocean temperatures from 1880 to 2012 (Tang, 2019). Some reports have also stated that the worlds temperatures between 1983 and 2012 have been the warmest in 800 years (IPCC, 2018).

This increase in temperature caused by GHGs have a multitude of consequences, ranging from rising sea levels to increased acidity of the oceans, all of which have detrimental effects to the inhabitants of the planet. Currently, GHGs emissions can be seen as somewhat of a necessary evil, given the fact that it is a by-product of many imperative industries such as agriculture, which also happens to be one of the biggest contributors of GHG emissions, responsible for more than a quarter of global emissions (Homaira & Hassan, 2021).

As the population inevitably grows, we are faced with the challenge of sustaining our necessities such as food and transport, whilst keeping rising temperatures in check. Unfortunately, solving the problem of increasing GHG emissions is a complex matter which may take years of change due to its nuance. This report thus aims to discuss the future of GHG emissions, should we continue this upward trend, through the application of predictive modelling. More specifically, we will be conducting predictive modelling to forecast future GHG emissions for the 5 highest GHG producing countries for the next 50 years.

# **Chapter 2.0 Literature Review**

This chapter aims to review five literatures pertaining to time series forecasting and aims to summarize their predictive modelling methods and results. While they may not all strictly pertain to predicting GHG emissions, the performance of the models as well as their data mining methodology still largely apply to this project.

Radojević et al. (2013) conducted a study on GHG emissions in Serbia using Artificial Neural Networks (ANNs). Using the countries performance indicators such as GDP and energy consumption as input parameters, Radojević et al. (2013) used ANN to forecast GHG emissions.

The resulting model showed satisfactory results, having an R-Squared value of 0.9125, in essence having a relative error rate of less than 10%. Radojević et al. (2013) states that ANN models are suitable for predicting GHG emissions, given the non-linear nature of the problem as well as not having strict mathematical relationships between the variables (Radojević et al., 2013)

However, Radojević et al. (2013) also states that other ANN architectures, such as Recurrent Neural Networks (RNN) should be tested to improve the quality of the models, due to their ability to factor in previous values in a time series.

Siami-Namini et al. (2018) carried out a comparison between traditional forecasting methods and deep-learning algorithms for time series forecasting, specifically Auto-Regressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) respectively.

Siami-Namini et al. (2018) used the 2 techniques and conducted data mining on stocks from various indexes on Yahoo finance, predicting the “Adjusted Close” price based on previous closing prices in the time series, at a ratio of 70% training data to 30% testing data for both techniques. The following results from the 2 techniques were then assessed using metrics such as Root-Mean-Square Error (RMSE), comparing the accuracy of the 2 techniques.

Siami-Namini et al. (2018) found that LSTM performed vastly superior to ARIMA, with an estimated 84% - 87% reduction in error rates, likely due to the iterative nature of deep-learning algorithms like LSTM.

Similarly, Ludwig (2019) did a similar comparison of three different methods of predictive modelling: Adaptive Neuro-Fuzzy Inference System (ANFIS), Recurrent Neural Network (RNN) and LSTM. These 3 techniques were chosen thanks to their nonlinearity, as compared to traditional forecasting techniques such as ARIMA.

Using time series data of GHG concentrations in 2,921 different grid cells in California, Ludwig (2019) used the 3 aforementioned techniques to forecast GHG emissions. Using evaluation measures such as Root Mean Squared Error (RMSE) and Mean Absolute Error

(MAE), Ludwig (2019) found both ANFIS and LSTM to perform favourably, with RNN performing the worst amongst the three.

Ludwig (2019) attributed this difference in performance to RNNs shortcomings, that being the network only remembering the past few steps in the data sequence and is thus unsuitable for longer time series data (Ludwig, 2019). As for LSTM and ANFIS, it was found that ANFIS outperformed the other 2 when using RMSE as an evaluation measure, and LSTM outperformed the other 2 when using MAE as an evaluation measure.

However, one notable observation is that LSTM took significantly longer to execute, as compared to ANFIS, with a run time of 150.39 seconds as compared to ANFIS’ 6.87 seconds (Ludwig, 2019).

Manowska (2020) used LSTM to conduct predictive modelling on electricity consumption in Poland. Using electrical consumption rates based on each different sector, Manowska (2020) was able to forecast electricity consumption with satisfactory results, using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), as performance metrics.

More specifically, Manowska (2020) found that the LSTM model had a MAPE of 1%, lower than the assumed 5% error rate. Manowska (2020) also finds that certain industries such as the industry sector may actually decline in overall energy consumption, linking it to other factors such as legal regulations that are outside of the scope of the model.

Manowska (2020) also states that the model operates under many assumptions and its accuracy is subject to uncertainties such as technological development and the aforementioned legal regulations that may affect the credibility of the model.

Homaira & Hassan (2021) used ARIMA, LSTM, as well as a simple linear regression model to predict agricultural emissions in Malaysia.

Homaira & Hassan (2021) used datasets consisting of values such as temperature and total emissions from 2000 to 2017 to build their models. Once again, evaluations measures such as RMSE, MAE and MAPE were used to calculate the accuracy of the forecast. For the ARIMA model, evaluation measures such as Akaike Information Criterion (AIC) were used to find the best order of the model.

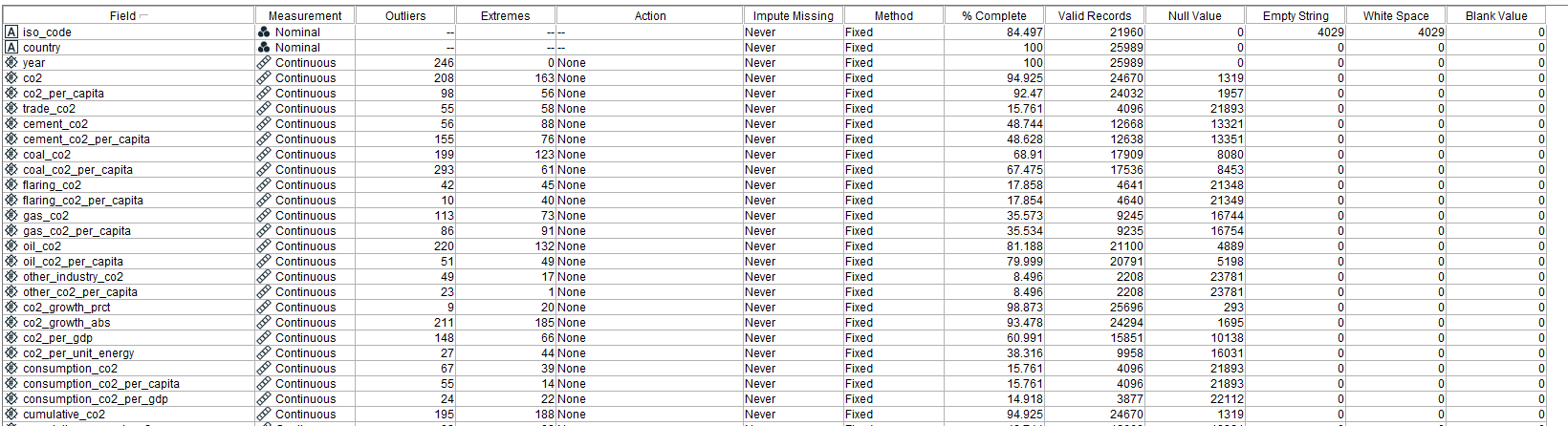
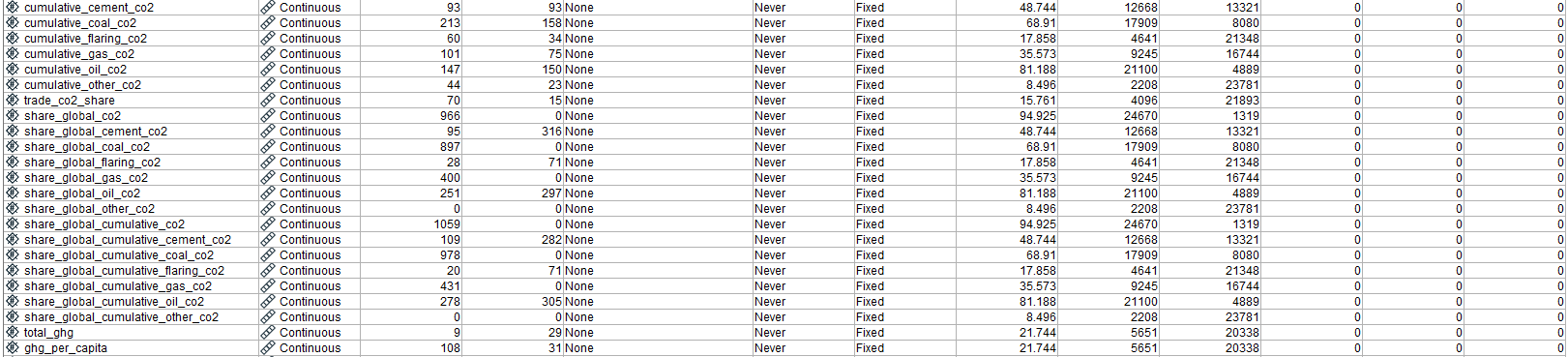
All models were found to have satisfactory results, all of which forecasted a rising trend in emissions for Malaysia.

# **Chapter 3.0 Data Understanding and Preparation**

This study will use primarily one dataset. The dataset named “owid-co2-data” contains global CO2 and GHG emissions data. The dataset is a repository, containing emission data from various sources such as government websites (Ritchie et al., 2022). The dataset contains several fields, including annual CO2 and other GHG emissions, GDP as well as CO2 emissions per industry of 248 countries and regions for that year. The dataset contains a total of 25,990 records with 60 columns each.

# **Chapter 3.1 Data Exploration**

While this project aims to use Python to use LSTM and predict future GHG emissions, preliminary data exploration will be done using SPSS Modeller.



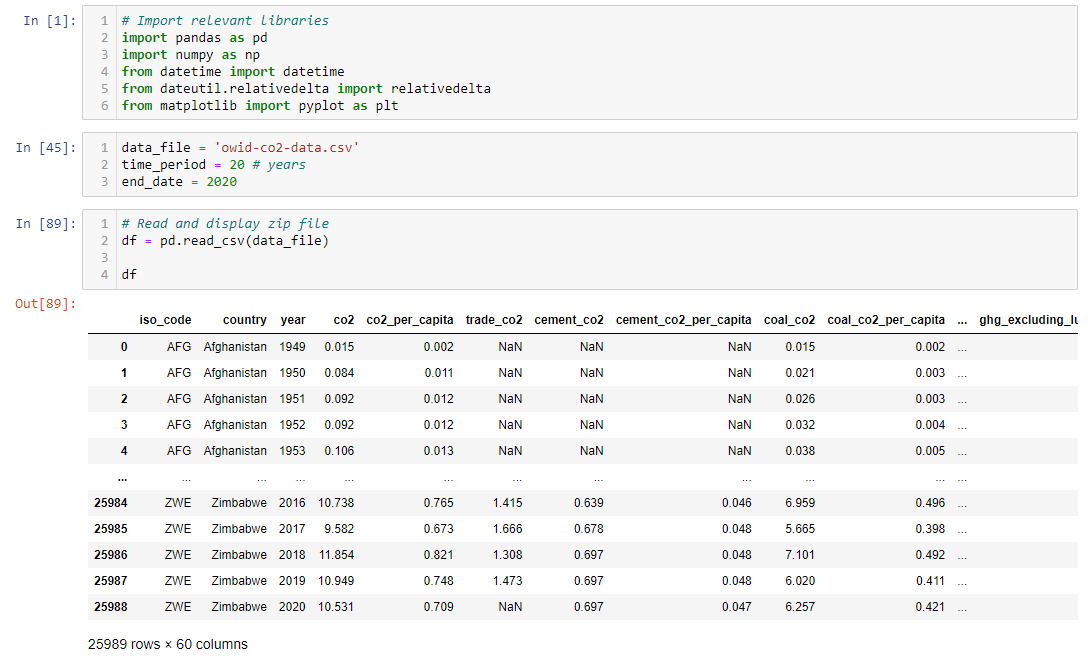
Using the Data Audit Node, we find that the dataset has only 3.33% of the fields being complete, with several missing, extreme, as well as outlier values. However, the purpose of this study is to forecast future GHG emissions (in this case CO2) based on the amount of GHGs emitted previously, hence, only 3 fields will be used, those being: “country”, which describes which country or region the CO2 emission comes from, “year” which is the date of the recorded CO2 level, and “co2” which measures each countries production-based CO2 emissions in million tonnes annually.

For the “co2” field, there are a number of outliers and extreme values (208 and 163 respectively). This is likely due to the fact that there is CO2 emission data for groups of countries, such as “Upper-middle-income countries”, as well as global CO2 emission levels. Emission data like these are a cumulation of multiple countries and will thus have much higher values than emissions from a single country.

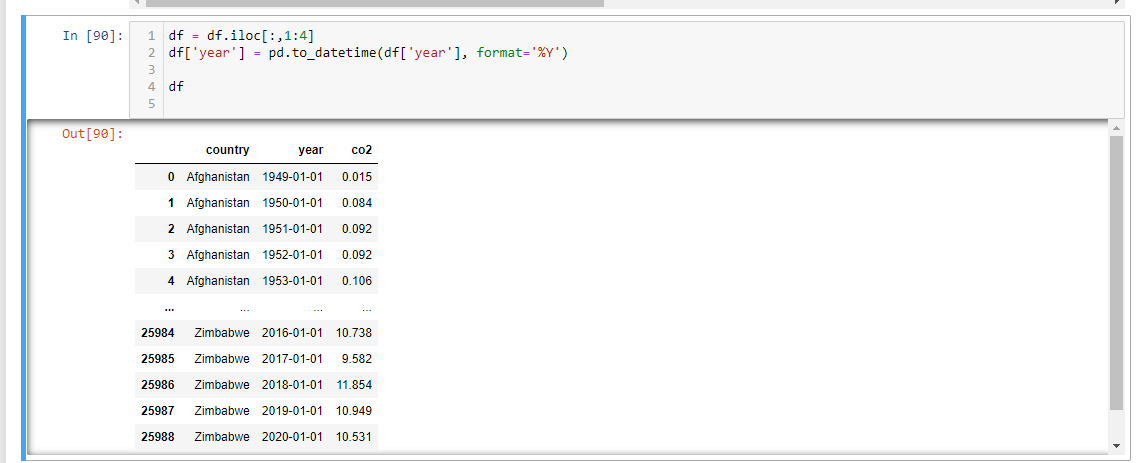
# **Chapter 3.2 Data Preparation**

Because the goal of this study is to forecast emissions for the top 5 GHG producing countries, the data will need to be prepared before conducting forecasting.

After first importing the relevant libraries, the csv file is read and put it into a dataframe, in this case “df”.



Currently, this study aims to conduct univariate time series forecasting of CO2 emissions. Hence, unneeded columns are filtered out as such:

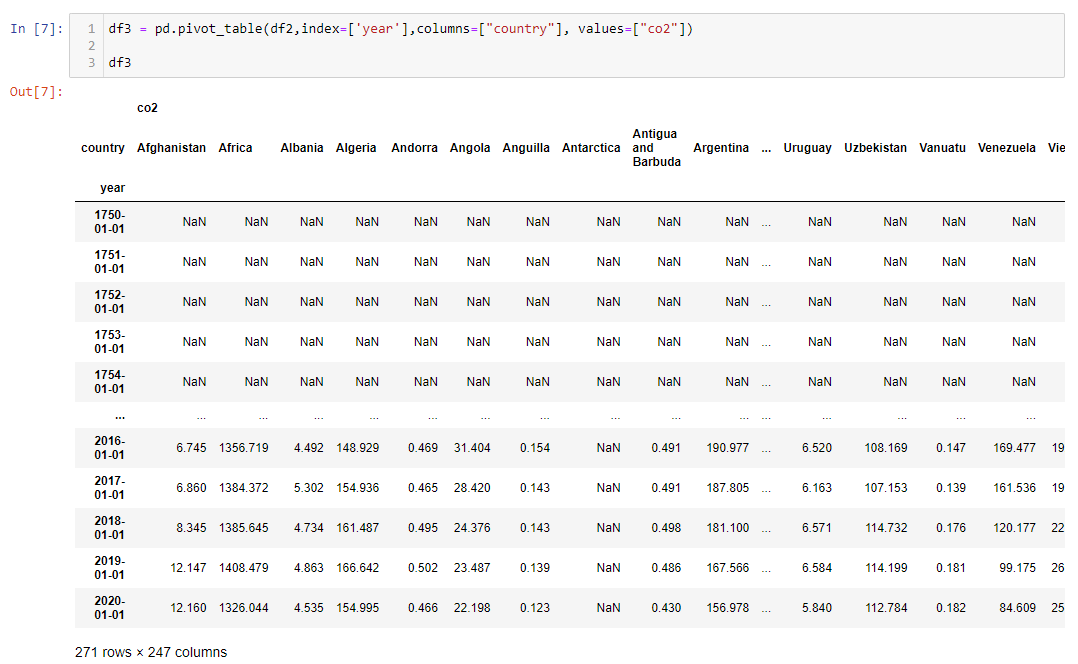


This leaves the country, the year of the CO2 emission level, as well as the CO2 emission level itself.

Rows that do not contain any CO2 emission data are then removed.

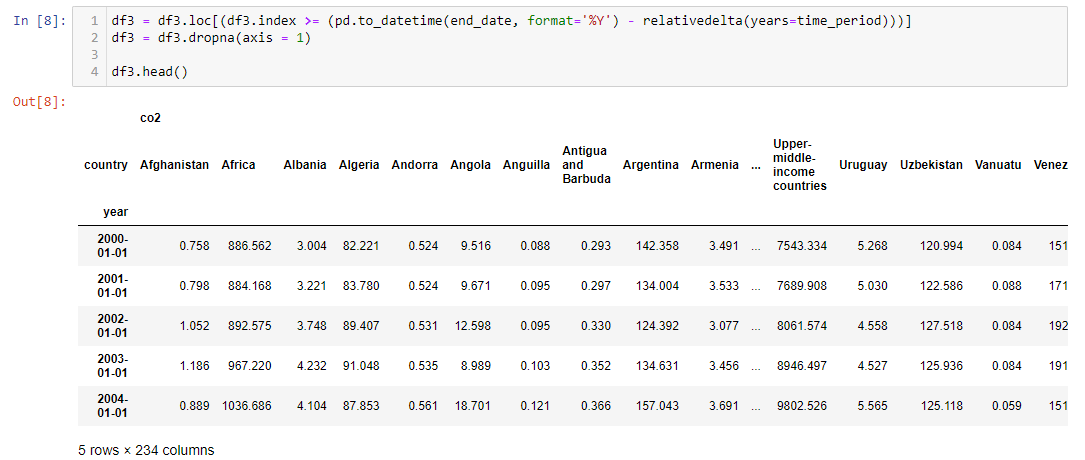


Afterwards, the dataframe is pivoted, giving a better look at the CO2 emissions from each country year over year:



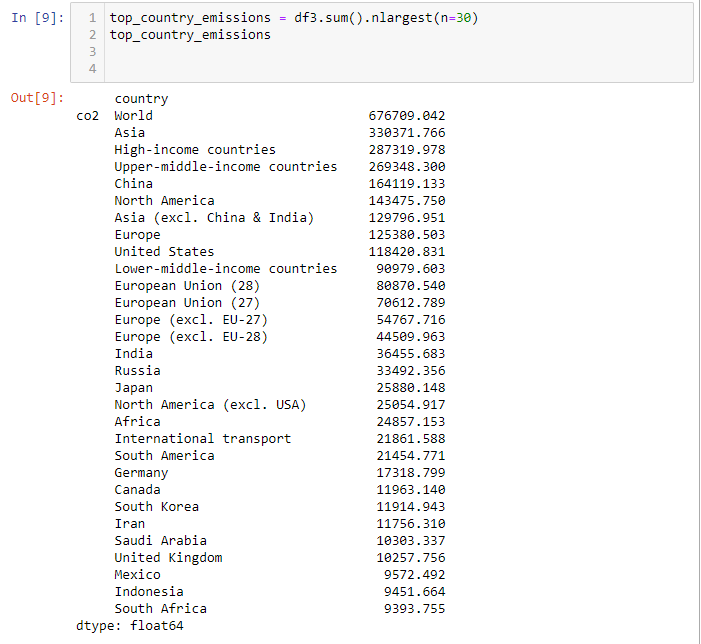
For now, there is no specific time frame of past datapoints used for forecasting. Hence, this study will be arbitrarily using datapoints from the last 20 years, up to the latest year of record (that being the year 2020).

Because this study is doing forecasting, null values may affect the accuracy of the final model. Hence, any countries that do not have a consecutive 20 years of data from the year 2000 to 2020 are dropped.



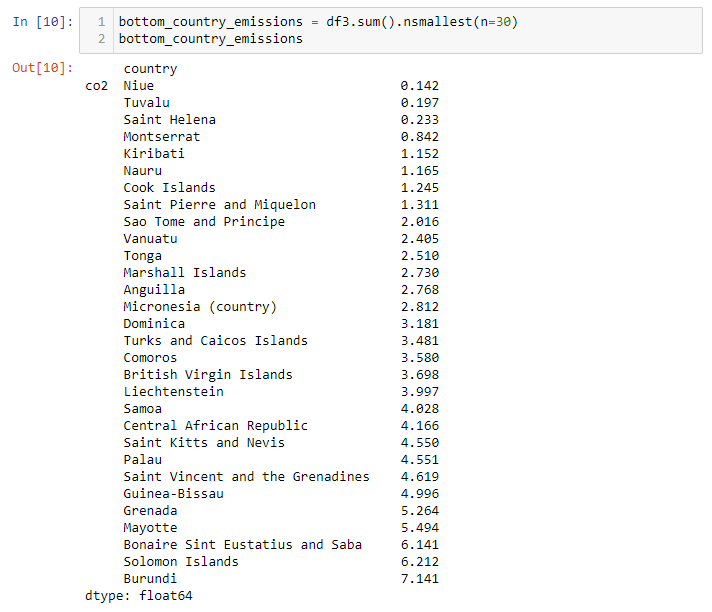
This leaves 234 countries and regions.

Finding the top 5 GHG emitters, we sum up the total CO2 emissions for the past 20 years and list the highest results.

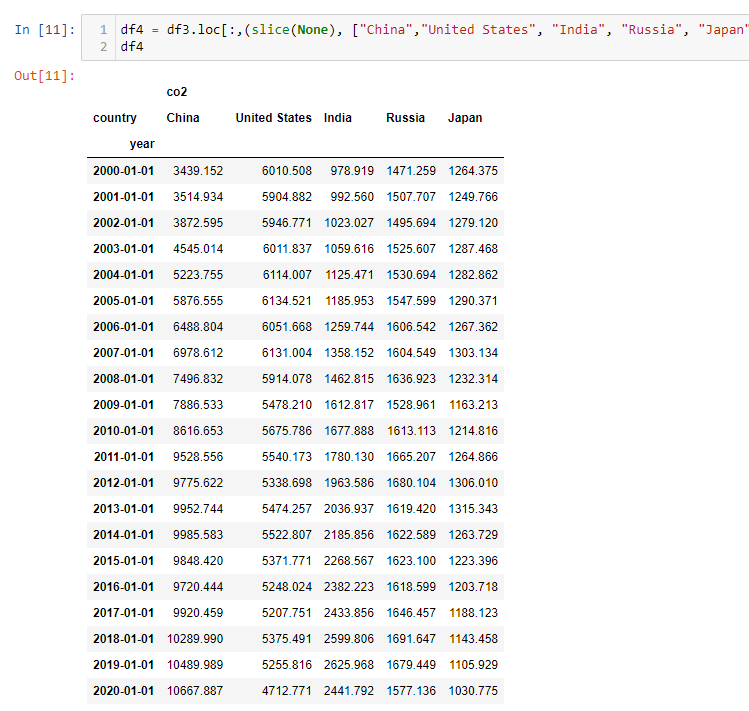


Unfortunately, due to the dataset containing global data as well as aggregated emissions, only rows that are part of the 195 countries (Worldometer, n.d.). will be selected. Doing this, the top 5 countries are: China, United States, India, Russia and Japan.

The bottom 5 countries are: Tuvalu, Kiribati, Nauru, Sao Tome and Principe, and Vanuatu.



Given this, the finalised dataset to conduct time series forecasting on looks as such:



# **Chapter 4.0 Proposed Modelling and Evaluation**

Given the literature reviews, LSTM has shown to be an appropriate predictive modelling technique to use, given the nonlinear nature of co2 emissions, as well as the LSTMs ability to handle sequential data.

LSTM is a type of Recurrent Neural Network (RNN), which aims to predict the next value based on a sequence of past datapoints, by learning from them. Different from traditional feed-forward neural networks, RNNs hidden layers store information of the earlier sequence, allowing them to leverage past datapoints to predict future values (Siami-Namini et al., 2018). However, typical RNNs suffer from what is called a vanishing gradient problem, whereby the network has trouble remembering datapoints the earlier on in the sequence, basing its predictions on only more recent datapoints (Kostadinov, 2017).

LSTM thus aims to solve this memory problem through additional features, such as each cell having gates that either drop, filter or add data for the subsequent cell (Siami-Namini et al., 2018). This forms something of a transport line between cells, transferring and gathering data. The three gates that alter each cells state are the Forget Gate, Memory/Input Gate, as well as the Output Gate.

The Forget Gate defines what information to forget and remember, and outputs a number between 1 and 0, with 1 being “remember everything” and 0 being “forget everything”. The Memory/Input Gate chooses which information to be added or updated in the cell, and the Output Gate determines which cell state information should be selected for the output of the cell.

Since the goal of this study is to forecast GHG emissions for the next 50 years, using a long time series to for model training is critical in achieving accurate and statistically significant results. Thus, using LSTM for time series forecasting is the best choice.

Model evaluation will be done in a similar manner to the previously discussed literature. Using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), as performance metrics, we should be able to ascertain the overall accuracy of the LSTM model.

# **Chapter 5.0 Proposed Schedule**

This chapter timelines the planned tasks for the current project, starting from the 15th of August:

|  |  |  |
| --- | --- | --- |
| Week | Date | Description |
| 1 | 15th August 2022 | Submission of proposal |
| 1 | 15th August – 21st August 2022 | Begin work on final report, based on supervisors’ feedback   * Adjust scope of project if necessary * Finalize objective of project * Build first proposed model |
| 2 | 22nd August – 28th August 2022 | * Read more literature reviews on different time series forecasting methods * Collection of new datasets (if necessary) * Data collection & data preparation (if necessary) |
| 3 | 29th August – 4th September 2022 | Build more models based on the project scope change   * Specifically, ARIMA |
| 4 | 5th September 2022 | 3rd Meeting with Supervisor |
| 4 | 6th September – 11th September 2022 | Adjust models or scope, based supervisor feedback |
| 5 | 12th September – 18th September 2022 | Prepare for oral presentation   * Finalize models by this week * Perform model evaluation |
| 6 | 19th September – 23rd September 2022 | Oral Presentation |
| 6 | 24th September – 25th September 2022 | Work on final report   * Introduction done |
| 7 | 26th September – 2nd October 2022 | Work on final report   * Literature review |
| 8 | 3rd October – 9th October 2022 | Work on final report   * Data understanding and data preparation |
| 9 | 10th October 2022 | 4th Meeting with Supervisor |
| 9 | 10th October – 16th October 2022 | Adjust final report based on supervisor feedback |
| 10 | 17th October – 23rd October 2022 | Work on final report   * Modelling * Evaluation and discussion |
| 11 | 24th October – 30th October 2022 | First draft of final report |
| 12 | 31st October 2022 | 5th Meeting with Supervisor |
| 12 | 31st October – 6th October 2022 | Adjust final report draft, make a second draft, contact supervisor and finalize report |
| 13 | 7th November 2022 | Final Report Submission |

# **References**

Homaira, M., & Hassan, R. (2021). Prediction of agricultural emissions in Malaysia using the arima, LSTM, and regression models. *International Journal on Perceptive and Cognitive Computing*, *7*(1), 33-40.

IPCC. (2018). *Climate Change 2013: The Physical Science Basis*. Retrieved August 14, 2022, from https://www.ipcc.ch/report/ar5/wg1/.

Kostadinov, S. (2017). *How recurrent neural networks work*. Medium. Retrieved August 14, 2022, from https://towardsdatascience.com/learn-how-recurrent-neural-networks-work-84e975feaaf7

# Ludwig, S. A. (2019). Comparison of time series approaches applied to greenhouse gas analysis: ANFIS, RNN, and LSTM. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-6). IEEE.

# Manowska, A. (2020). Using the LSTM network to forecast the demand for electricity in Poland. *Applied Sciences*, *10*(23), 8455.

Radojević, D., Pocajt, V., Popović, I., Perić-Grujić, A., & Ristić, M. (2013). Forecasting of greenhouse gas emissions in Serbia using artificial neural networks. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, *35*(8), 733-740.

Ritchie, H., Roser, M., Mathieu, E., Macdonald, B. & Rosado, P. (2022). Data on CO2 and Greenhouse Gas Emissions by Our World in Data. Github. Retrieved August 14, 2022, from <https://github.com/owid/co2-data>

# Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018). A comparison of ARIMA and LSTM in forecasting time series. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)* (pp. 1394-1401). IEEE.

# Tang, K. H. D. (2019). Climate change in Malaysia: Trends, contributors, impacts, mitigation and adaptations. *Science of the Total Environment*, *650*, 1858-1871.

Worldometer. (n.d.). *Alphabetical list of countries*. Worldometer. Retrieved August 14, 2022, from https://www.worldometers.info/geography/alphabetical-list-of-countries/