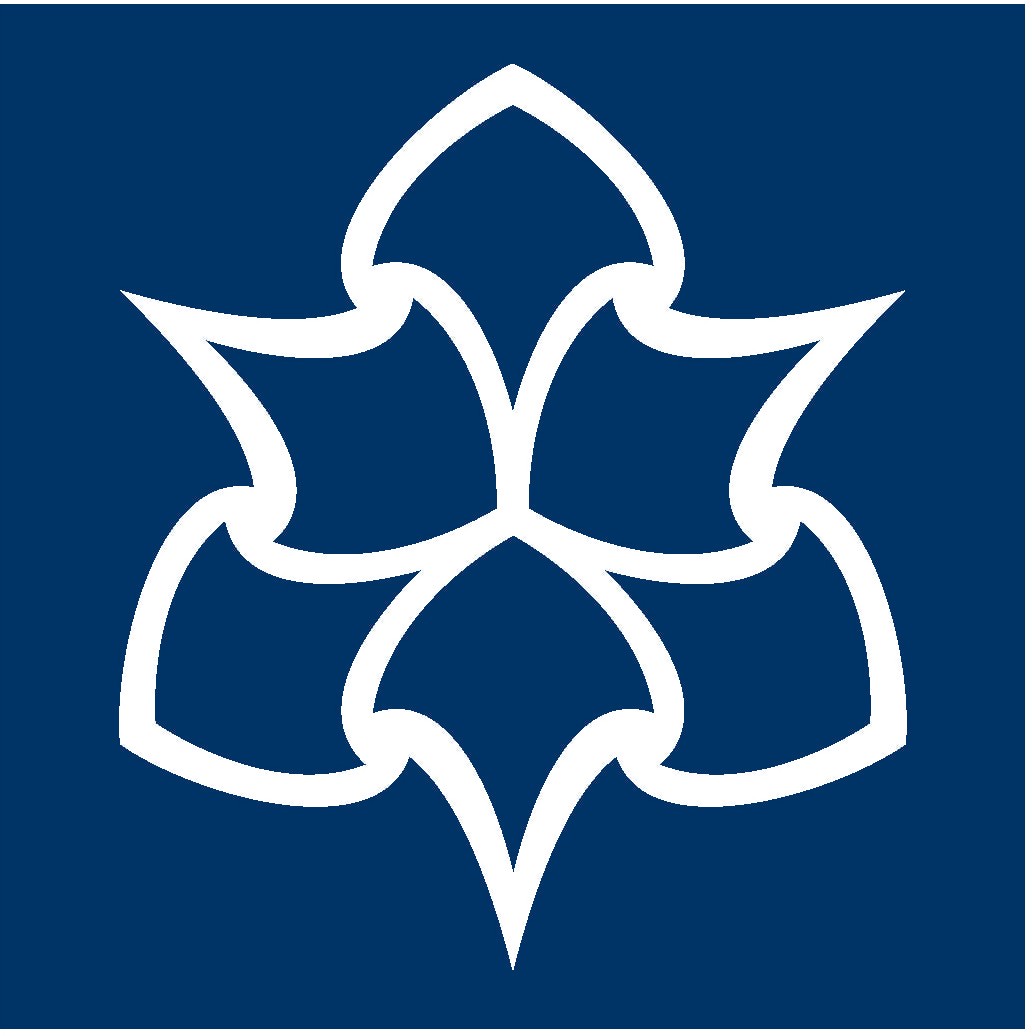
TIME SERIES ANALYSIS AND DEVELOPMENT OF MACHINE LEARNING MODELS FOR FORECASTING MULTIVARIATE DATA

A DISSERTATION SUBMITTED TO MANCHESTER METROPOLITAN UNIVERSITY FOR THE DEGREE OF MASTER OF SCIENCE

IN THE FACULTY OF SCIENCE AND ENGINEERING



2019

By Nikhil Jagatia

Department of Computing and Mathematics

# Contents

[**Abstract**](#_bookmark0) **v**

[Declaration](#_bookmark1) vi

[Acknowledgements](#_bookmark2) vii

[Abbreviations](#_bookmark3) viii

1. [First Chapter - probably “Introduction”](#_bookmark4) 1

[1.1 Referencing](#_bookmark5) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

[1.2 Figures and Tables](#_bookmark6) . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

[1.3 Equations](#_bookmark9) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

1. [Second Chapter - probably “Background”](#_bookmark11) 4

[References](#_bookmark12) 5

* 1. [The first appendix](#_bookmark16) 6
  2. [The second appendix](#_bookmark17) 7

# List of Figures

2.1 ARMA Equation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

2.2 Image of a LSTM Neural Network. . . . . . . . . . . . . . . . . . . . . . 21

4.1 Resampled data for yearly average temperatures showing an upwards trend temperature increase . . . . . . . . .. . . . . . . . . . . . . . . . . . . . . .38

4.2 A Histogram showing the range and density of temperatures recorded globally

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 39

4.3 Boxplots of the different continents and how their data variance differs . . 41

4.4 Resampled yearly Temperatures across time for the different continents. . 43

4.5 Scatter plots of residual values of different continents. . . . . . . . . . . 45

4.6. FITTED MLR MODELS FOR AFRICA AND ASIA. . . . . . . . . . . 47

4.7 FITTED MLR MODELS FOR OCEANIA AND EUROPE. . . . . . . . . 48

4.8. FITTED MLR MODELS FOR NORTH AND SOUTH AMERICA. . . . 49

4.9 BEST FORECASTED ARIMA MODELS 1 . . . . . . . . . . . . . . . . 53

4.10 BEST FORECASTED ARIMA MODELS 2. . . . . . . . . . . . . . . . . . 54

4.11 LOSSES FOR VARIOUS LSTM MODELS. . . . . . . . . . . . . . . . . . . .58

4.12. LOSSES FOR VARIOUS LSTM MODELS . . . . . . . . . . . . . . . . . . .59

4.13. PREDICTED VS ACTUALS FOR AFRICA AND ASIA LSTMS. . . . . . 60

4.14. PREDICTED VS ACTUALS FOR EUROPE AND OCEANIA LSTMS . . 61

4.15. PREDICTED VS ACTUALS OF NORTH AND SOUTH AMERICA LSTMS 62

5.1. COMPARISON OF THE DIFFERENT MODELS’ RMSE VALUES . . . . 68

APPENIX D FIGURE 1. RESIDUAL DISTRIBUTION PLOTS FOR DIFFERENT COUNTRIES WITH ARIMA MODEL BASED ON THE GAUSSIAN MODEL. . 102

# List of Tables

5.1 Best RMSE for the various continents and the models which had the best score . . 69

APPENDIX B

TABLE 1 SHOWING THE CORRELATIONS OF VALUES . . . . . . . . . . . . . . 96

TABLE 2. VIF VALUES FOR MULTICOLINEARITY CHECK. . . . . . . . . . . . . 96

TABLE 3. MLR EQUATION VALUES (3 D.P.) . . . . . . . . . . . . . . . . . . . . . 97

TABLE 4. P VALUES FOR THE DATA WITHIN THE DIFFERENT CONTINENTS. 97

TABLE 5. MEAN OF RESIDUALS FOR THE MODEL OF RESPECTIVE COUNTRIES . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 97

TABLE 6. SUMMARISED METRICS OF THE MLR MODELS FOR DIFFERENT CONTINENTS . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .98

TABLE 7. SUMMARISED METRICS OF ARIMA MODELS FOR DIFFERENT CONTINENTS . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98

TABLE 8. EXPERIMENTAL RMSE VALUES FOUND FROM DIFFERENT TUNING METHODS (NUMBER OF NODES AND EPOCHS) . . . . . . . . . . . . . . . . . . . 99

TABLE 9. SHOWING THE TEMPERATURE CHANGE PERCENTAGE BETWEEN JANUARY 1870 AND JANUARY 2013. . . . . . . . . . . . . . . . . . . . . . . . . . 100

# Abstract

This project was performed with the intent to investigate the use of machine learning models to be developed to enable forecasting using time series data. The Earth is at a critical point where we have the impact of global warming generating a cascade of events that we need to try and mitigate. The analysis of Earth surface temperatures enables us to obtain a historical snapshot and therefore model from what we know. The use of algorithms such as MLR, ARIMA and LSTMs are known for the independent qualities. Therefore, using these as a basis to attempt modelling is seen. Using an array of statistical and hypothetical investigations such as correlation testing, and hypothesis testing e.g. White’s test we were able to confirm the suitability of transformed Berkeley Earth data by means of separation into continents to allow a better modelling. After the analysis and transformations, it was found that certain attributes carried a greater weight than others. It was indeed key to make transformations to allow the models to be fed clean and satisfactory data that could be worked with. Furthermore, that the risk of over transformation and instantiating averages leads to a too wider generalisability and underfitting. This was understood to be improved by further training. On the whole LSTM models performed better and MLR models performed better than expected. While there was greater uncertainty surrounding the use of ARIMA models. The significance of machine learning models and its applications are expansive and when combined with expertise on global climate, we should be able to see how changes over time can be affected by some changes more than others. Nevertheless, the space of Earth’s global temperatures which are on the rise give way to further machine learning models to be developed for this cause. With greater computational power will hopefully be able to add to what we already know on how to best preserve our planet.

# Declaration

No part of this project has been submitted in support of an application for any other degree or qualification at this or any other institute of learning. Apart from those parts of the project containing citations to the work of others, this project is my own unaided work.

Signed: NIKHIL JAGATIA

Date:25/09/2019

# Acknowledgements

Thanks to

Luciano Gerber

&

N.P.

# Abbreviations

MLR Multiple Linear Regression

ARIMA Autoregressive Moving Average Model

LSTM Long Short Term Memory

MSE Mean Squared Error

RMSE Root Mean Squared Error

LST Land Surface Temperatures

OLS Ordinary Least Squares

NN Neural Network

ANN Artificial Neural Network

BP Backpropagation

MR Multiple Regression

PCA Principle Component Analysis

ELM Extreme Learning Machines

AWS Amazon Web Services

MODIS Moderate Resolution Imaging

ACF Autocorrelation Function

PACF Partial Autocorrelation Function

GAN Graph Attention Networks

GRU Gated Recurrent Unit

VIF Variance Inflation Factor

AIC Akaike Information Criterion

BIC Bayesian Information Criterion

MAPE Mean Absolute Percentage Error

SMAPE Symmetric Mean Absolute Percentage Error

SWM South West Monsoon

PSO Particle Swarm Optimisation

# Chapter 1

**Introduction**

## Introduction and Background

The importance of climate change has been highlighted as more of significance in the last 50 years. The preservation of our planet and its future has been taken on ourselves and its responsibility is becoming a greater burden as time passes. With the resources being used, our understanding of non-renewable fuels has increased and therefore it has been correlated as a contributing factor to the increase in the Earth’s global temperature. However weather prediction in general has been one of the holy grails of predictions, that date back to as far as the Stone age (Geetha and Nasira (2016)). We have explored the depths of our oceans and ventured into the stars but much of how our Earth works and trying to predict events that occur still remain under investigation.

This project aims to address the concern of climate change, given data, to generate a models which allows forecasting of global surface temperatures. The data obtained will be pre-processed and investigated to see if there are any pre-occurring relationships and statistical information that will be deemed useful. After substantial cleaning, the data will be subject to extensive feature engineering that focuses on the problem with respect to the dataset. Furthermore, hypothesis testing will be conducted to allow the correct inferences to be drawn from the findings generated.

The models that will be discussed in this project each have their advantages and disadvantages. The following models will be used: Multiple Linear Regression (MLR), ARMA and Long-Short-Term-Memory (LSTM) Neural Network.

Following the implementation of the models, the critical manner of feature engineering will occur followed swiftly by parameter tuning to achieve the best model possible. Each model respectively will be evaluated based on its performance.

The format of this paper and its outline will seek to use data from Berkley Earth which is enabled for public use and for such explorative studies. The structure will be presented as sections that critically examine the different impacts of society on global earth surface temperatures, models, data, findings and the discussion of the future prospects.

The Key aspects of this papers are as follows:

1. The significance of the dataset and the importance of why research and this field are necessary
2. Determining the best machine learning model for the data provided whereby the declaration will be based on metrics and usability.
3. The findings of the results and what they indicate for the future.

The hypotheses that are outlined, aim to use statistical inference to measure suitability and suggests which model performs the best. This is part of the experimental design.

# Chapter 2

**Literature Review**

**2.1 Background of Weather**

The outline of this literature review is to examine the basis of the paper by looking not only at mathematical and computing elements of the issue on hand, but also to consider the different elements that combine to impact the Earth surface temperatures the way that they do.

The weather shapes our days and how we go about it. The ecosystem being diverse as it is has adapted to such changes that have occurred, to a point where we think we understand what is normal. The perception of normal weather and temperature is relative to the location of the measurements being taken.

Some would say that we take nature and the change of weather as granted. Whereby changes allow growth and cycles to be completed.

The cycle of weather and its unpredictability or predictability can directly affect the measurements taken from the Earth’s surface and in some respects the Air temperature. Whereby for example a warm day would generate a given spike on the surface temperature recordings.

SEASONS

The word change is an important term when referring to weather and temperatures. This is because there is a difference seen during different periods of the year: seasons. This variability can be monitored and therefore contributes to the norm at any given time of the year. The periodic pattern of season allows us to recognise the trends, this is especially true over larger periods of time. This is due to reoccurring events that is interpreted in terms of the bigger picture and thus larger time span.

LATITUDE AND LONGITUDE

These are the measurements and co-ordinates as reference points of locations on the globe. The locality of these points can determine the average temperature readings for that area. Where equatorial regions will have a higher average temperature than those that are situated closer to the poles in their respective hemisphere. An interesting form of analysis could be the difference in rate of temperature increase over time for different regions of the earth, i.e. Could the average rate of temperature increase be greater in Europe than in Africa? Nevertheless, the prospect of the different temporal changes of temperature should not be discounted.

TEMPERATURE VARIABILITY

The study by Albright *et al* (2015) attempted to identify the relationship and affects between the land surface temperatures and the air temperatures given different contributing attributes. The Land Surface Temperature (LST) is a core element to various environmental processes and combines characteristic changes between the atmosphere and biosphere. The measurement of LST is made possible typically from thermal measurements via sensors. The most common source from sensors is the MODIS.

Gridded air sets have been seen as a medium to model and predict these temperature changes. while there are various factors that look into the dependency of the temperature based on different attributes Huband and Monteith (2015) identified the attributes that now can be postulated to have an impact on LST and its air component. It was also identified that during autumn there was a noticeably strong relationship. But this was later considered due to the increase in frequency of apparent clearer skies.

Logically it is known that there could be a distinct correlation between altitude of reading measurement and the temperature, whereby those at higher altitudes would be subject to colder temperatures, therefore in terms of the terrain aspect this is a significant feature.

Errors of temperature measurement can be due to a wide variety of reasons. Such examples are interference of carbon dioxide and other greenhouse gases from satellite measurements. Here different parts of the earth are more prone to emitting greater number carbon dioxide particles which can interact altering air composition which inevitably will be measured. Such examples are large cities where due to the increase in emissions and topography the surface temperatures are noticeably higher than rural areas. However, in different circumstances where for example there is volcanic activity, false positives can be seen as the surrounding land while it may be cold, the proximity of the temperature measurement given with respect to the natural elements of earth’s landscape may potentially generate false readings. Furthermore, the addition of air temperature can be influenced by continentality. This means the temperature brought in as a process of large bodies of water such as Oceans and Seas can have an impact on the coastal temperatures. These are features that would need to be considered during the model and predictions.

**2.2 Weather Prediction History**

In the past the primitive method of collecting temperature data and for weather prediction had been by using mercury-based thermometers to measure the air and surface. Modern day technology has enabled us to use these principles when measuring and recording temperature changes.

It was interestingly showed how the LST data can be highly correlated to near surface air temperature. Therefore the role of the air can have an impact on the surface temperatures as mentioned by Zhang, H. (2017).

The paper by Ceccato *et al* (2010) highlighted the significance of analysis of LST monitoring. In specific regions the importance of forecasting such climactic events can play a role in the distribution risk and transmission of vector borne diseases. Therefore, not only is it significant to predict for the land but also for the population that live in such areas and the possible implications on human health. By monitoring and then predicting such events we can potentially plan to mitigate the damage to human health. Such is the added importance of monitoring and forecasting temperature.

There are many models that can be tested but finding one specifically given a unique dataset is significant. In Himika and Randhawa (2018) 3 out of 15 models were evaluated to predict the changes of temperatures. It was tested that the best 3 models would undergo ensemble formation of which would further be the best model. This was concluded by using cross validation methods which concluded that in general the best model for the temperature prediction was the Decision Tree model. Given its simplicity and by using the information gain mechanism the best model after consideration of its accuracy was found.

Today we are fortunate to have the supercomputing power to enable us to look at features such as 215 billion weather observations across the globe which allows the formation of an atmospheric model, this is the MET Office’s supercomputer Cray XC40 which can use 24 Petabytes of data and is a massively parallelised. This is one of the biggest supercomputers in the world that is dedicated to weather and climate monitoring such is the requirement with the large number of readings being processed (Met Office (2018)).

Furthermore, added efforts will be made from the European Centre for Medium Range forecast which from 2020 will be situated in Bologna, Italy. This is a dedicated research institute that is used for weather prediction. The supercomputers used to generate the models allow global examination and insight into the shape of the upcoming weather and thus temperatures as outlined by the European Centre for Medium-Range Weather Forecasts (2019). Such is the significance of simulation-based predictions.

**2.3 Time Series Data**

A fundamental aspect of time series data is extrapolation, as defined by Vajpai and Vaishnav (2018). Time series is a distinct type of data that is attributed to the standard time element. This could be in the basis of Years, Months, Hours, Seconds, ect. The properties of time series data allow forecasting to occur. The forecasting is made possible by looking at the historical values and then performing computation alongside logical observations and knowledge of the domain to be able to predict future incidences of events.

It can be used to predict a wide variety of occurrences such as prices in the stock market whereby the unpredictability element can be reduced by mathematically looking at trends and behaviours of certain assets. By contrast a significant use of time series can be done by observation and prediction on meteorological elements. The importance of this can be of extreme significance as it can dictate economical and physical changes that have effects on our lives. Most notably is the prediction of weather and in this case temperature. Modelling weather events and predictions allow us to prepare for the extreme and dangerous occurrences of some of the Earths deadliest events, such as Hurricanes. As mentioned by Medar *et al* (2017) weather forecasts can never be fully accurate due to the nature of data analysis and the interpretation of findings for future results.

Machine learning algorithms are able to use the time series data alongside a mix of attributes that add to the information gain and mechanism to learn and train the prospective model (Liu, Z *et al* (2017)). Here some parameters and attributes have a greater weight on the final outcome than others. In these cases, significance testing is required.

Due to the variability and the uncertainty in time series forecasting the stochastic models that are represented in order to predict values are attempted by using the historical data points an extrapolation is made possible due to algorithmic learning.

STATIONARITY

The concept of stationarity is whereby the mean variance of the data from which is can be plotted and seen is constant. Due to the nature of time series data there can be factors that cause a proximal change in the data, therefore the essence of non-stationarity is generated. Non-Stationary data is harder to work with and requires altering to allow modelling to occur.

The way in which the data can be changed to satisfy the stationarity is by differencing. The differencing is achieved by observing next immediate values rather than the whole data that is being trained. This allows the variance between the data points to be more consistent. A method of statistical determination is by the Augmented Dickey-Fuller test which as shown in InsightsBot (2018) with code, which uses the concept of unit root testing and hypothesis testing to allow determinations of stationarity.

TREND AND SEASONALITY

A trend in time series data is one which over time there is a changing tendency of the data, this can be linear or non-linear. In addition, trends are a characteristic that is not repeated as it is a gradual change over the period in question. This allows inferences to be made.

Furthermore, another component of time series data is Seasonality. This is a regular pattern and change over a period of time that is consistent to time intervals. Combined, trend and seasonality are key in prediction models as they allow us to extrapolate using the historic data in a logical way by means of observations over k lags this is broken down in the text at Statsoft (2019).

NOISE

This feature is a common aspect of time series data. With unclean and messy data there will always be discrepancies or data which does not fit into the norm. It can help explain the random variation in the dataset and can be addressed by instantiating moving average or normalisation procedures.

**2.4 Multiple Linear Regression**

One of the most common and important models is regression. The principle of regression is to form a line that best fits the data points. The equation of the line can then be used to extrapolate the data and therefore predict values for future instances such is the advantageous simplicity of the model. The multiple linear regression models indicate the relationship between a dependent variable and two or more independent variables using historical data as mentioned in Zhang, J. and Cao (2011).

The equation for the line can be seen as shown below from Feng and Wang (2017):

**Y = β0 + β1 X1+ β 2 X2 + ⋯ + βn Xn + ε 2.4.1**

This multiple regression equation can be identified and decomposed by ε as the added random error term. The β values represent the coefficients that are unknown and weighted onto the models for n variables of the different X attributes, therefore generating an equation where Y is the predicted value.

The estimation for the coefficients is typically performed by using the least squares method outlined in Brito and Mendes-Moreira (2014). Here the aim is to limit the error from the training data to the line that will be formed. It can be calculated using the *i*th x values, y values and their averages respectively:

**2.4.2 2.4.3**

The equation uses a combination of coefficients multiplied by the input instances of the attributes. Additionally, the added part which is the y intercept shows what we see as the 0 mark for when the independent variable is 0. The premise of this model looks at multiple features that are correlated against each other. The article by Swalin (2018) decomposes the metrics that are R2 and Root Mean Squared Error (RMSE) and how they could be used for evaluation, alongside generation of the defined RMSE equation.

The RMSE is a useful metric that can be used to evaluate the goodness of the model in question using each estimated observation against the actual predicted. It is defined as:

**2.4.4**

Here for *n* variables the sum of the squares for observed vs estimated is taken and then square root, to account for negative values. This shows how close the model is from the true values, however it can be susceptible to outliers.

By using MLR was can consider the following as assumed: firstly, there is a linear relationship between the outcome and the independent variables. Furthermore, that there should be no, or limited multicollinearity. This factor is how highly correlated at least 2 features are. The problem with having this issue is that if there are more than 1 factors influencing the behaviours of the dependent variable then it becomes difficult to distinguish between the two variables. If the correlations are too high, then the data structure can be looked at further to minimise this and to potentially eliminate it. However, the problem is that although there can be multicollinearity shown, the results of this could be due to random chance effects. Therefore, editing the selection and use of certain features can harm the integrity and the true validity of the results.

The study conducted by Gupta and Priya (2015) which used time series components for the use of examining the criminal activities in India and forecasting using the predictors. This understood that it is possible to analyse trends and patterns to predict crime and perception of it in a noble attempt to mitigate the tragic occurrences that still exist even with questionable policing in India. They found using Pearson’s’ Correlation Coefficient that an increase in crime was highly correlated to the sex-ratio. Spearman’s rank which analyses statistical dependence between two variables was checked, and therefore this correlation coefficient validated the Pearson’s ranks as the crime rate and Sex-Ratio was high. Amongst the current data used the domain knowledge of Historical predictors were used to formulate the assumptions and relationships. However, while using Sex -Ratio seemed like an important factor in crime rate it can definitely be attributed to certain cases such as Rape, however it is questionable as to whether other offences are correlated to a difference in the Sex-Ratio. Therefore, although regression has been used to look at real-world problems happening based of historic measurements it is also significant to include the validity of the models based off the specified and targeted variables.

IMPORTANCE OF ATTRIBUTE SELECTION FOR MLR

During the model selection and using the specific attributes the selection of certain features and their significance needs to evaluate. An example is the study conducted on Stockholm housing prices and the trade-offs that have to be taken into consideration by Gustafsson and Wogenius (2014). Here since Stockholm is a very community driven place they require that most valuable priced houses are those that are in near proximity to public transportation. There is a lack of housing in Stockholm the questions raised with addition to socio economic indices make apartment values an interesting concept. Given that MLR is used there were a multitude of different attributes that were considered as variables for this study. Therefore, there could be an investigation of what is actually important and is being close to public transport actually valuable?

Naturally with any MLR model there are different assumptions that are made such as: linearity, error terms defined, variances were spread and are not correlated to each other (heteroscedacity) and importantly no multicollinearity. With the assumption that was made with the MLR assumptions the OLS method was deemed unbiased and therefore the estimation can equal the end beta value. While the heteroscedacity which is the spread of the variance was monitored visually, there was a need for another more complex method to verify that the certain attributes met the required standard. This was using White’s Robust Estimate.

What was also identified with respect to multicollinearity was that a large variance means that the estimates aren’t precise and cannot be considered during the hypothesis testing. Hence the verification performed is so important. The measurement of this was computed using VIF. Where a solution that was defined stated the prospect of the inclusion of more data. However, this would then result in an increase in variance but also more inappropriate data. In terms of feature engineering all the parameters were considered to be useful. The conduction of the data used an observation of the expected value of the error term equals to zero which was found by the formulation of the regression line with a confidence interval of 95%. This shows that any errors from now on meant that it should be probabilistically 0. Consequently 5% of the data was the testing data which was moved during extraction and then later progressed to validation using cross-validation. The results indicated after thorough screening where the R2 was 0.91 for the regression which therefore explains the 91% of the variation. The run through of the different attributes and their statistical significance is found for the conclusions where area and monthly price were key. Therefore, while the regression uses the maximum number of attributes to give an informed equation for prediction, there are occurrences where not all the attributes are required as some may hinder the model. Yet it is a requirement to be thought about that if there is too much removal of logical attributes there becomes a gateway for overfitting which needs to be taken into consideration.

MLR COMBINATIONS

While there are some advantages in using MLR for a wide number of problems, some require further learning of the inputs and their weights. Such is shown in the study which observed Flood Prediction in Mohawk River by Kapetanakis *et al* (2018). Using Kolmogorov-Zurbenko filtering for the time series decomposition allowed good observation of the significance of noise and its impact on the accuracy of the models defined. It was consequently found that while the MLR is justifiably generates ok results the ANN accounts for more of the realistic combinations and their effect. This can be due to the different weightings on the values from the inputs to the hidden layers of the perceptron. This then reflects how basic the MLR algorithm is in comparison even though the relevant attributes are included. Therefore, the use of the particular algorithm can rely on the dataset and the problem in question.

There have been several studies that have used the concept of MLR but with the additions of models that further use the concepts that engage in the development of a better model.

Such an example was the combinatory use of the Grey Model with Multiple Regression Model by Wang, Q. *et al* (2009). Meteorological data was used as an independent variable along with holidays. These are two very key factors in energy consumption. Where each respective model was used to determine a predicted value, using analysis of the variance the weights were calculated and then independent values generated predictions as expected. However, the combined method saw slightly higher accuracy, this was reflected in the overall smaller relative error measurements. This showed that there are applications of using combined Models.

REGRESSION AS A BIG DATA PROBLEM

When regression is seen to be used in large scale problems there is a requirement of how to manage the data accordingly. When data is very large some algorithms may suffer and the lag is significant. Such is the need for inspection for Multiple Linear Regression for scalability purposes mentioned in Boufares and Rehab (2015). For such big problems a framework like MapReduce can be used to perform the training in a parallel way. Enter grid computing on multiple clusters. This enables no limitations on the normal disk shared memory. However, the load to change to a MapReduce framework to allow efficiency is seen post 100,000 data points. It would have been beneficial to have a better understanding of the standalones initially rather than comparison immediately to the MLR-MR function which was found in this study. However, what was obvious is the way as the number of cluster nodes increased, the training time reduced making a greater speedup. In any case the accuracy was shown to drop as the training data size increased. It can be speculated whether this was because of the mixed values or if it was the sheer scale of the whole set.

In terms of using time series to fit to regression models, it can be defined that different periods of the time series can fit the model better or potentially worse than the whole on the average. There is potential to decompose the data delivered as time series into smaller fragments whereby multiple models are generated for different instances as subsets, this is defined as the Multiple Model Estimation in Cherkassky and Ma (2005). It shows that by having more than one regression model that there is more room for noisy data that can be attributed. The robustness of the models was evaluated where the best models allowed better accommodation of the noise found. However, a disadvantage to this study in terms of the general model that is required for simplicity and definitive prediction is that by having many smaller models this raises questions on overfitting the data. Since There will be natural variability in the data, having models that overfit not only for a particular year, but for a window of the year reduces the overall reliability for forthcoming events.

The type of time series data obtained to be used for the MLR can be understood to be developed from temporal information too. This is highlighted in Dorville and Williams (2015) whereby recognition of the impact of light energy spent and its relationship as a contributing factor to global temperatures increase. While the Earth’s natural resource is the Sun which lights and heats the planet we understand that there is a relationship between light and heat.

Light pipes that are used in buildings allow light to be better distributed. The illuminance of the light pipes can be calculated by meteorological data and then the adaptation can be altered to find the energy efficiency by analysing the relationship. In addition, the waste heat that dissipates into the atmosphere can contribute to warmer conditions. In this study, where it was conducted in Jamaica, this is notoriously a hot and humid region and therefore the right steps needed to be taken to ensure accurate predictions where 9 different variables and measurement mechanisms were used to record and manage the data. They tested the configurations with sunny and cloudy days so that the variance can be explained for the different circumstances. They found that should be correlation and closeness of data from each date or even intervals to allow accurate and explainable readings and findings to occur. This meant that to be able to draw relationships such that depend on weather and climate there needs to be a consistent period formed.

The significance of having MLR models working alongside geographical information is key because not only does it enhance our understanding, but we are able to use historic data that dates back to effective record keeping. A study conducted which aims to use multiple regression to be able to predict the and forecast the date of the South West Monsoon (SWM) for a particular state in India over a 41 year window in Ancy *et al* (2014). Awareness of the various parameters and factors that influence the onset of monsoons such as winds, humidity etc. While it is good to bear in mind these factors, there will be instances that are unexplainable, such is the science behind the investigations into weather prediction. Training was set as a 29-year window leaving the following 13 years up for forecasting and thus testing, typically the error was deemed to be found by Forecast – Actual. They managed to generate with a low error, a month prediction in advance however the main reason that this prediction isn’t usually performed until around the expected time is due to weather anomalies that can occur within that time period. Even though a long timeframe justifies the claims, the room for error must only be increased due to the uncertainty and the change in conditions over time.

This paper confirms that while MR systems are useful in forecasting such problems especially with weather and where we have many different parameters and attributes to consider, perhaps the firmness of the correlation had an influence on the overall accuracy and therefore overfitted more than it is realistic.

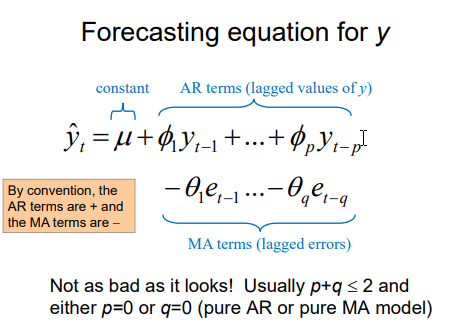
As shown the applications of the MLR remain elastic and can fit a wide range of data in its relatively simple functionality. The context of the data is important though, in addition the rationale on how pre-processing and selection of attributes is indicative of the proposed models. The aspect of obtaining multiple attributes to derive a suitable model is not simple as there needs to be a weight associated with each attribute. Furthermore, if there is an opportunity to initially use the MLR as an initial mode, and then progress to learn the weights of certain attributes to then be used on a secondary model it may yield more reliable results. Plus, there would be a greater explanation of the variance in the data. However, the computational cost and time spent doing this remains an issue that is somewhat dependant on the problem itself.

**2.5 Autoregressive Moving Average model (ARIMA)**

As working to forecast in advance using data requires knowledge on the past, there is a need to observe the historic incidences within the makeup of the data. While there may be relationships that aid the search for the best predictive model, there can be overestimation as to the type of relationships such regressive models have which may or may not affect the final outcome of what the dependent variable is found as.

In addition, having the correct amount of white noise for such data generated a predictive power in that the model is deemed more reliable. This is because when observing a data point in reflection to its past; while trends overall and seasonality may be relevant, the difference in variation which attributed to random variable change can become significant.

The ARIMA model was first established in 1970 by Box and Jenkins where the premise is based on the prediction intervals and the correlation association between the variables noted in Pintelin and Schoukens (2006). The AR element of the model was defined as the autoregressive element and is noted as the variable p. Secondly the MA element or the moving average element is denoted by the term q. Typically the values found by means of p, d and q respectively are derived by observing the autocorrelation function (ACF) and partial autocorrelation function (PACF) and checking the image output so that the limits can be manipulated effectively . Here as defined by IBM (2012) the ACF relates to correlations that are k intervals apart from lag k. Whereas PACF also accounts for the data intervals between.

 Fig 2.1. Formula taken from Gupta. N (2018) of the ARMA equation

The figure 1 shows the ARIMA model and the way both the AR and MA element contribute to the integrated form for those coefficients tagged to each value in their prediction of **y.**

The calculation phases of the ARIMA method was defined by firstly checking the stationarity of the time series. Followed by generation of the ACF and PACF functions to determine the p and q levels. Then this allows the form of the ARIMA model to be composed. Once we have the models parameters we are then able to implement the model and consequently form a fitting curve for the results that are calculated so that we can decide how well the prediction is as fit as per Hu, N. *et al* (2012). From which the RMSE will mainly be looked at alongside the AIC and BIC which are penalising scores that are found from analyses. Here the BIC is more critical of the parameters that are fed into the system. As mentioned in Arshanapalli *et al* (2014).

The study by Paz *et al* (2000) used rainfall data points across 37 years to be used alongside the ARIMA model. Since rainfall is an event that is readily observed our understanding of it can become unstable. Such was the variance. Therefore, the variance was unequal but the way to transform the data so that it complied to the ARIMA requirement was to square root the values of the data points thus stabilising the variance.

From this since the data closely matches it is concluded that for rainfall forecasts the ARIMA models can be a reliable source of experimentation for models even where some climates have little data.

The study by Adewumi *et al* (2014) conducted experiments for the time series modelling of the Nokia stock share prices from different exchanges. It was found that while ARIMA had been a useful tool, that it was particularly effective for short term models. The use of ARIMA in short term models as the stock market prediction is that where stock market has limited information on the movement and characteristics of the stock the use of this model allows good forecasting, but does it have the right type of explanation of why the values are as they are?

In our situation where we have the earths weather temperature we can identify and estimate dependencies, this is something which is difficult due to the nature of the stocks. Furthermore, in relation to the problem statement, given that the data obtained for this study is dated as far back as 16 years and considered as a short period does that infer that for time series that extend further than 16 years it is inadequate? It raises the question because of where the limit lies in the threshold of achieving good and significant prediction.

The study conducted by Babu  *et al*  (2012) attempted to look at how variants of ARIMA models can be used to aid the prediction of global temperatures.

Along with using basic ARIMA the trend based ARIMA uses a smoothed approach by using a filter which uses the trend component and leverages the raw data do develop its residual data, therefore the data is then predicted using a combination of the trend components and the residual data components.

Similarly, the wavelet transformations were used configure the residuals alongside the raw data into different components. With the inclusion of the wavelet filter it performs a processing step that aids in the predictive accuracy.

The data period chosen which was a 120 year time span had been used. Where the predicted had been from 2001 to 2010. This is a large period of time and could be considered as a long-term type forecast as a decade worth of data is attempted to be forecasted.

What was interestingly found is that where the basic ARIMA is used, it was trumped by the trend based ARIMA and then Wavelet based scored even better on the metrics. This all could be due to the filtering mechanism used within the Wavelet complex. However, this shows that over a long span of time the predictions by using ARIMA can be justified.

Another study that was conducted supports the claims of ARIMA’s suitability in short term forecasting models Raj *et al* (2017). Here this paper claims that ARIMA is useful for short term predictions and not accurate enough for long term prediction. Comparison between ARIMA and ANN and the potency of using both models. A 6-month short term forecast was made and then a 12-month long term forecast was made followed by a 24-month forecast. While the shorter-term forecast had a much less broad prediction for the next 6 months than the 24 month It is noticeably seen that the 95% confidence interval showed a greater variation. This may have been due to the variation that could potentially be seen with a longer period of time. There was a clear upwards trend and an obvious seasonality element was included. This reflects the suitability of the ARIMA models for such datasets. However, the use on larger scale datasets with many more data points is suspicious due to even further increase in variation.

ARIMA TESTED EXPERIMENTALLY

The study conducted by Kayvanpour and Mehrmolaei (2016) looked at the way we can use ARIMA as a basis to become more developed. This firstly uses the typical ARIMA model for forecasting. Secondly there is a proposed improved method to which errors are calculated and then the model’s average mean error is used to update the system. The proposition is good and uses a somewhat backpropagation style method to enable better learning and therefore prediction of the results. There were three different studies that were conducted to verify the proposed method.

The first test uses monthly births in NYC. Here where ¾ of the dataset was used for training purposes over a 13-year window. There was defined metrics as RMSE, MAE and MSE. Firstly, the different models using p, q, d were used and then to generate the best model given the parameters Then the proposed method was performed. There was a noticeable difference in the metrics when compared to standard ARIMA. These types of results were also true for immigration data and monthly number of measles that were conducted. While there was difference in the basic versus the proposed it needed to consider if the results are significantly different. However, the test was conducted for an array of different cases and therefore the weight of these should be taken into consideration. There seems to be a concept for further research on the relearning given weights of ARIMA however whether there would be a computational cost associated depending on the research and the consideration of the different socio-economic influences that can be affected. More so does the reiteration of the error learning cause a fitting of the model that is too tight? Nevertheless, there is an indication that perhaps learning can be a viable option. However, this could be the case with most models. It shows the different ways ARIMA can be used and means to show that the baseline models can be enhanced further depending on the errors generated.

It was found in Lo (2011) that there are great dependencies in the requirement to have fault free and reliable software. And so being able to effectively examine reliability using statistical methodologies can prove to be a valuable insight. Furthermore, there is dependency with having code updates and software updates as the testing phases and the error or in this case fault frequency can fluctuate. So, this model was used because it is widely been associated with time series data.

The ARIMA as outlined determined the estimated residuals to have a mean of zero and a variance which is constant, as per typical ARIMA algorithms.

While ANNs can be used to fit nonlinear data we can see that amongst other problems such as speed, there is unlikely to be significant overfitting when using SVM therefore it is used.

The aims of this paper specific was to propose a novel attempt at using SVM in conjunction with ARIMA. Where ARIMA is good for the linear data, we can see SVM which can typically be used for non-linear data. Therefore, the idea of using both where each type of model can complement the other. Furthermore, the hybridization of such models would in theory be beneficial.

It is concluded that each strength seems to have been utilised and the hybrid model generated the best and closest to the y values as seen. This reflects a way in which the strengths of ARIMA when applied can yield an informative result depending on the dataset.

Performed over 10 years ago Aguera *et al* (2009) as another comparative study, focused on using ARIMA but with NNs. Here the second lag was assumed to be the order of the model due to the ACF and the PACF functions being plotted and observed.

The derivatives were taken and then the most appropriate models were concluded to be found. There were 2 different conditions to be tested and short- long predictions.

As the predictions of higher estimates increases the indexes degenerate slowly.

After generating subset of models including those of NN it was found that all models differed on their averages for their given metric for those that satisfied each criterion.

Consequently, it was found that the ARIMA (2,0,0) with the conditional attributes was the easiest and less time consuming to implement. Therefore, whether the more diverse models from NN are advantageous they also lack the speed. However, another aspect of ARIMA that was verified by this study was that ARIMA is not a model that is best suited for long term predictions.

While understanding that hybrid and combinatory models can be effective it is also important to take a step back and reevaluate whether they should be used. The paper on the Ensemble method used for ANFIS-ARIMA for the rainfall prediction Faulina *et al* (2012) assessed the capability and compared not only the use of the ARIMA model to ANFIS (Adaptive Neuro Fuzzy Inference System) but tested whether it is necessary to do this.

The comparisons were made between the way ANFIS generated a better prediction of Pujon’s rainfall data whereas ARIMA performed better for Wagir’s rainfall data.

Here in this context the ensemble method uses a combined approach to multiple forecasting models. It could be seen as reliable however the integrity of the models may be compromised.

What was found by performing the tests individually and combined is that when the RMSE was analysed, the individual models performed better, with ARIMA generating the lowest RMSE and therefore the better of the two models. This is in line and corroborates the M3 competition where just because a complex model is available does not mean it should be implemented consistently in Hibon and Makridakis (2000).

Furthermore to (Lo. J., 2011) where nonlinearity is raised in question to the dataset and this was mitigated as a problem by the introduction of other models. Here LSTMs were examined and identified as a mechanism to use alongside ARIMA in Lou and Wang (2019) which was wavelet based using the Daubechies10 wavelet function.

Error as a natural process is addressed whereby the uncertainty of weather and prediction for small datasets can explain why some models might overfit. Another factor that has a seasonal dependency is that there is less noise in the dry season due to lack of rain and the variability escalates during wet occasions.

The experimental performance was where the ARIMA model forecasted the linear part of the data the residuals were then obtained. The LSTM learnt the nonlinear relationship which it is deemed good at and therefore once the LSTM was complete the residuals were calculated and the ARIMA was then updated.

Yet what was found was the benefits were seen by the forecast result where the predicted values accurately fitted the real values. In addition, the de-noising with the combinatory effects of the time series data ensured that while typically ARIMA is good for short term prediction LSTMs contribution and its learning mechanism for residuals is effective.

**2.6 Long Short Term Memory Models (LSTM)**

The architecture of LSTM models was first proposed by Hochreiter and Schmidhuber in 1997 written in Hu *et al* (2017). Here the basis of the problem to be solved was the issue of ‘Vanishing’ gradients. This is achieved by a safeguard from using backpropagation. The LSTM cell network implements a set of gated pathways which learn through either a sigmoid function or a Tanh function.

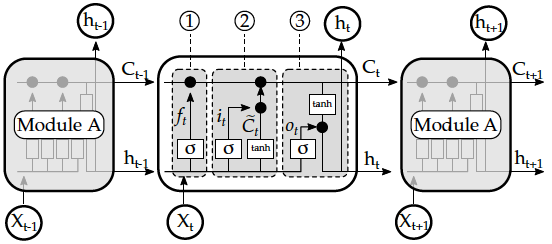


Fig 2.2 Image of a LSTM Neural Network. Source: Celeda *et al* (2019)

In these instances, the Sigmoid function allows the vector being passed through to be refined to between 0 and 1 whereby those that tend to 0 are more subject to be forgotten. This is so that only the most useful information is retained, and thus it applied integrally in the ‘forget gate’ (labelled 1 from Fig 2).

The Sigmoid function is one that is present throughout the LSTM system and one that is present in the input gate where once combined with the Tanh function to allow regularisation, is performed by the product of the previous element that has passed through the Forget Gate (from gate 2). The final calculation of the error is taken by:

***Ct = ƒt \* Ct-1 + it \* t* 2.6.1**

Backpropagation learning style is now seen by taking this newly formed weight and multiplication of it to the vector that attempts to pass through the Output Gate which forms the output or the next value of the hidden layer to be combined with the new input (output gate 3).

The paper by Choi *et al* (2018) developed their LSTM by means of trial and error with the use of the MAPE %. An initial number of hidden nodes was chosen as 300 with the number of relevant LSTM cells as 2,4,6,8 respectively. Once the best number of LSTM cells had been identified the same number was run for the different number of nodes ranging from 100 to 300. This incremental and logical approach is intuitive because it aims to check whether other models with different parameters perform better.

The selection of the LSTM model for the number of nodes and hidden layers is important and while there is a known relationship between accuracy and those elements there is no scientific and distinct way for determining the hidden layers. Furthermore, while an increase in layers might reduce training errors, the time taken to execute the layers is computationally expensive and may overfit the data, discussed in Liu. R. (2018).

A factor that should also be considered generically for all models is the sliding window. The relevancy was found in Liu, X. and Zeng (2018). Here while similar to Liu. R., 2018 whereby the suitability of LSTMs was considered for stock prediction purposes, here the paper identified that while the use of LSTMs is good for such variation and scale, the window and its accuracy are dependent on each other. If the monthly window was taken and compared to the accuracy of daily it was found that the overall accuracy of the monthly window matched greater. This could have perhaps been due to the consistent trend in the data, even though due to the nature of stocks they can be random and unpredictable. Yet it shows the significance where there can be such variation in a particular window even though the accuracy on a bigger scale can be higher.

Furthermore, to time series data where there are aspects of seasonal impacts on natural occurrences we need to take into consideration the different features that exhibit such features. Clustering can be deployed to divide the data into subsets of different weather conditions and then the LSTM model could then train given its respective weather state. This would increase reliability in the long run because one model may not account for specific seasonal changes. However, there is a potential to overfit the data by having such specific data points and characteristics set in place. Additionally, a question of whether the amount of data available would be representative of the model and therefore be generalisable. When compared LSTMs are able to perform better than ELM and ANNs even for looking at seasonal data via clustering mechanisms unique to the period in question which was identified in He *et al* (2018).

Interestingly Aparna (2018) was Trying to accurately predict the power demand by using NN techniques. There is appraisal of NN and mentions LSTM and RNN while mentioning regression which is unsuitable. The use of NNs was a key concept, along with the benefits and how they compare to other methods such as regression. LSTMs are good for long standing and dependant data. Can update weights which is useful with bigger data points and larger variances in time series data methods involved used LSTM, RNN which then used a modelling pipeline that gives feedback per iteration as neural networks do naturally and tries to implement them using TensorFlow and Keras. The rolling windows seemed like a sure way of ensuring that the data that is predicted is tested against historic data. Methods involved using meters where the values are aggregated using AWS. The smart meters stored characteristics of the electricity. Modbus protocol converted the data and allowed it to be transferred to MCU which then pings the data to the cloud. Easily AWS has support for Keras so when models are build using python on Jupyter then every time new data is received it automatically updates the model. This was an interesting way to automate and improve the accuracy and relevancy of the model by giving it real time inputs from data received from different sites. The MSE was performed until the losses per epoch were at a minimum. From the findings of the actual vs predicted; while there was some variance the MSE for the rolling window implementation would have been a lot lower and the accuracy much greater without seeming to overfit. There is a suggestion that with a larger dataset the LSTM would probably perform better inevitably. So, to combine the use of rolling window with automation seems like a good way to develop the model. In addition the use of techniques such as rolling window allow the data to be approached and then be delivered to a higher degree of accuracy thus ensuring a greater yield of validity and greater generalisability.

A min-max scaler was used to transform the data into a range which is more user friendly. This ensured that the pipeline was run. Because the data was taken every month there could be swings that could not be explained and therefore phenomena which can alter the accuracy, such is nature. Perhaps the data that was ran until losses are at a minimum meant that time and resources were affected, this is because NNs aren’t the fastest predictors. Also, it is questionable if the number of layers used to be experimented more. The beneficial automation element was how there is nice interaction between the meters, aggregators and AWS to streamline the process. Although it would’ve been interesting to see more of the data over a longer span. The findings from 5 years would be different to that of 10 years.

LSTM COMPARISONS

Further to the use of LSTMs, Al-Fuqaha *et al* (2017) concluded that from the paper that attempted to compare the ARIMA vs the LSTM model to see which would yield the better accuracy as defined by the RMSE. It was concluded that LSTMs did perform better however they suffered when it came to training time. Furthermore, due to the complex nature of the LSTM models there were many input parameters to tune. Not only did this mean more complex tuning but it also added to the overall train time. Therefore, there seems to be a significant tradeoff between the time taken to achieve the lower RSME value.

Furthermore Fu, Y. *et al* (2018) compared the accuracy of the different models from LSTM, GRU, ARIMA and SVM. They understood that the prediction accuracy from LSTMs to GRUs were close. Therefore, new developed cases were generated with different sizes of training sets and it was said that the difference will not be significant when the training size is small however when there are many data points and the set is large the LSTMs greater prediction accuracy as defined by the SMAPE will ensure it is the better model. This highlights the significant role the size of the dataset plays in the LSTM model forecasting.

In addition, Huang *et al* (2018) used elements of the fuzzy rough set to explain ambiguous data incidences. Then once this had been implemented on the attribute reduction there was a notable reduction in the accuracy metrics and was shown to have a good improvement when compared with the BP Neural Networks. This is useful for where there are many complex attributes and given the typical ‘long’ time taken to run a LSTM model it may be a beneficial and novel way to address the multifaceted dataset when it is given.

For the use of LSTMs we see in Chen *et al* (2018) that while LSTMs may be useful in time-series data sometimes the models requires encoding to allow unfolding of the vectors to achieve the forecasting necessary. Such an example here is where GAT is used. This principle could be used for complex time series with an irregular graph structure by generation of fitting matrices.

Due to the nature of LSTMs the data requires normalisation to be able to build a model. This is something required as part of the preprocessing specified in Di *et al* (2017).

Multi-step iteration is used for the forecasting here. For each predicted value there the following value to be predicted is gained from the newly predicted value. This over a long period of time series we are able to observe the growth or retardation rate.

When compared to ARIMA during the multistep prediction the LSTM model generated a lower MSE which therefore yields a better accuracy however the resource consumption of the LSTM is greater. Yet the benefits of using LSTMs are clear in the sense that modelling process is streamlined in comparison to ARIMA whereby the stability checking, ACF and PACF checking is not required.

LSTM APPLICATIONS

It has been brought forward by Hu, Y. *et al* (2019) that LSTMS are highly recognised as a very good way to deal with time series data. Furthermore, there was reference to the genetic algorithm that observes that the stochastic optimisation methods and markedly that there is not a need to have gradient calculation as opposed to such LSTM models. An example of this is the PSO algorithm. However, while using PSOs are good using them in conjunction with large NN is not ideal. Nevertheless, bearing in mind the problems faced by PSO and highlighting the benefits of using it this paper aimed to develop a novel approach to the complementary models. This paper used firstly the model that was performed with the LSTM.. Following this, parameters and information was then passed onto the PSO combined model which then performs further iteration without the gradient descent problem vanishing. The usefulness and context of this study was for the analysis of occurrences of disasters and if they can be observed from the data, the true information dynamic of safety forecasting can be then concluded by looking at the overall variation. This is significant for those data for which particulate matter can be measured. The particulate matter can have an impact on the Earth's natural climate and therefore the forecasted dynamic is important to be understood. It was concluded that while the PSO-GD can be beneficial in terms of accuracy the PSO-Adam (optimised) - LSTM is better in terms of accuracy. Yet the way this can be developed is by implementing some form of trend normalisation/extraction whereby the model can be implemented on less parameters.

The postulation of using genetic algorithms to improve hyperparameter tuning was also raised in Campbell *et al* (2018).

LSTM AND THE ENVIRONMENT

Gang and Teng-Fei (2018) proposed that due to the ever-changing environmental changes there is a growing and more inevitable dependency on renewable energy sources. Such an example is the prediction of wind speed which is integral for the generation of wind power from turbine mechanisms. By using NNs we can gain a deeper and more complexified insight on how prediction allows us to better understand something as dynamic as weather which plays an important part in temperature.

Another renewable source of energy is Solar Power which alleviates the pressure placed on coal mining. However, the agenda for Solar power is dependent on the weather Chi *et al* (2018) therefore the prediction of energy use with respect to solar power can also be predicted by using the history. PCA is a good way of eliminating noise and to eliminate redundancies. This study looked at how using PCA with LSTMs compared to using LSTMs with affected the forecasting accuracy of the solar power demand. By using PCA there was an observed reduction in error. This could be due to the aspect of dimension reduction and dataset preprocessing. As well as reducing the error, even though the results from LSTMs were close the PCA showed and improvement. Furthermore, it was found that the run time of the PCA added attribute was computationally cheaper. Whereby typically where LSTMs and general neural networks suffer in time to execute the PCA streamlines this process. However, caution should be considered because there could be elements where the altered data that PCA uses would overfit, affecting the reliability. This shows a valid suggestion of using PCA with LSTM networks, however the type of PCA performed and how the data will be affected would be unique to each dataset.

While there are many factors that can affect the Earth’s surface temperature there is a way we could use machine learning to predict components of the air which can indicate various areas for concern. Gu *et al* (2017) looked at how the PM10 particulates are associated with adverse health effects but also detrimental to climate change. From conversion of time series data to sequence data the use of LSTMs was compared to that of Linear Regression and RNN, while considering non-stationary nature of time-series that can exist. The Adagrad and RMSProp optimisation functions were used within the model, which are actually typically used for non-stationary time series.

While the findings showed a 500% improvement between LSTM and Linear Regression there was a 100% better performance by the RMSProp optimised model than the Adagrad RNN and LSTM. Yet this was performed on univariate data. The applications could be used to investigate a higher dimension of data whereby other factors that affect PM10 levels. It is likely that due to the nature of LSTMs there could be an even more improved score with relevant PCA. However, it will undoubtedly result in a much longer execution time. It shows the significant of the LSTM models on such important data that can have different variations. It also highlights the different optimisation tools that can be used to enhance the model.

There are further areas to explore for LSTMs such as how high seasonality and multiple dependent attributes can have an effect on the models’ outcomes. Furthermore, whether a vast scale of data would be handled effectively by such a model. However, there is uncertainty on its efficiency due to aspect of runtime and conversion throughout its layered pipeline.

# Chapter 3

**Design**

* 1. **Problem Analysis**

This section looks to determine the aims and also address some of the potential problems that could be encountered during the project phases. It is significant to plan in advance to make contingencies especially with data that has a lot of variability. Plus, it is key to evaluate which packages and environment might be needed.

Due to the nature of the project there is a requirement to be able to work independently by using programming tools to be able to manufacture the Data Science pipeline. This process includes becoming familiar with not only the appropriate language but the relevant packages that are associated to the project. Such a challenge required the development of transferrable skills from the programming language Python that will be used. The language includes various libraries e.g. Pandas. Furthermore, it was a requirement to install the relevant Tensorflow library so that Keras would be able to be implemented in the later machine learning models. The development is progressive in nature and with continual usage the ability of using programming and mathematics alongside Data Science methodologies allows the pipeline to be as logical as possible.

Since the objective of this project is to analyse and develop machine learning models for time series data. In this context the data in question is global land temperatures.

An identified goal of this project is to determine if global temperatures can be modelled in such a way that allows effective exploration of the model with respect to different regional temperatures. Furthermore, it is important to consider the benefits of using certain models whilst also observing the drawbacks of each respective machine learning model. While there is a question of which model performs better with the time series data, it is also dependent on the data itself due to the nature of the dataset.

In addition, as part of the data pre-processing the cleaning may need to be performed as some data might be unavailable and therefore missing. This can be addressed by analysis of whether measures of central tendencies are required to fill in the missing values. However, the impact of the missing data would need to be considered and whether there will be an impact on the end result gained.

The questions that can also be raised are how the data will be processed given the different data types that are included in the dataset. Once the data has been processed appropriately another element that needs to be considered are the metrics that will be used to evaluate the models. This is key because to draw similarities between different types of models a standard has to be used as a comparison tool. Given the model the explanatory power and relevance of the metrics will be key. From this the measurement of success for each model will be able to be validated.

The size of the data may cause a problematic instance due to the range and the models used. The level of generalisability may be either too broad or specific. More so the determination of the actual model will need to be investigated. It is important to consider the model with respect of the data as there is a fine line between an accurate model and one which is over fitted or under fitted, both of which decrease the validity of the model that is been developed.

The following section aims to address how the project will be performed and the structure of the process that will be carried out. This includes the breakdown of how the goals and objectives will be investigated while also explanation of the justifications of such actions.

**3.2 Dataset Description**

The dataset in question is from Berkeley Earth Surface Temperature which comprises of 8235082 data points which span from November 1743 to 2013 which was repackaged from Kaggle that had been derived from the Lawrence Berkeley National Laboratory. The data provided is intended to be of an open platform type which allows users to perform further analysis.

Berkeley Earth prides itself in addressing issues regarding data and its selection which includes data quality, data adjustments and biases. This is important because having the best quality data allows a better amount of reliability and accuracy in models. Furthermore, it is important that the quantity of data observed is of a representative sample whereby we are able to generalise. Berkeley Earth take a data driven approach to the analysis of temperature and its impact on the Earth.

The date is provided in a format of YYYY/MM/DD whereby the first day of each month is recorded for the respective data points and in a monthly fashion. Due to the timeseries nature of the data it is clear that the element of forecasting would be at play.

Secondly is the target variable which is the AverageTemperature which units were degrees Celsius, this is preferential over degrees Fahrenheit due to a much larger range in scale of Fahrenheit. Following this attribute was the AverageTemperatureUncertainty is the 95% confidence interval around the average temperature. Where accuracy is concerned it is important to be able to have this as a way to measure the amount of error with the temperature recorded.

The locality of each temperature reading is key were certain regions experience different norms to others, this is described by the City and Country columns respectively, each with its latitude and longitude.

It is clear initially that there are missing values which will have to be addressed and therefore inconsistencies with the number of temperature recordings between different cities, perhaps due to environmental and economic viability reasons.

* 1. **Planning**

The importance of planning is integral for any project as it outlines the basis of how the development and the experimentation will be set up and performed. Furthermore it is key when defining how we are attempting to measure the success of such as project as it allows us to state the structure and process and then reflect on what was planned, what was implemented and what was found from the process.

A benefit of using Python is the element of Object Oriented Programming (OOP) which in the case of data analysis, mining, modelling is good as the execution of elements of code can be structed in grouped frames. This will be implemented using Jupyter Notebook which is an open source online interactive development environment where code can be executed and facilitates the performance of data driven insights e.g. visualisations with the appropriate libraries that are imported. Hence combined with the necessary libraries will give way to the fetching of the dataset.

Before any kind of models can be instantiated or goals can be defined it is important to become familiar with the data available. Following this it allows questioning of certain aspects of the data and then consequently questions on what could potentially be found. Furthermore, it allows familiarity of the type of data available and its usefulness because naturally once the information has been investigated we think into the future on what the data means and what we would like to test for. Therefore, the first element that needs to be investigated is the data exploration.

The performance of the data exploration is core to the project development. Such an example of the exploration that will be performed is the number of total data points that will be worked with, followed by the data types of each of the attributes, this is so that it is understood what kind of transformations would need to take place as part of the feature engineering element of the data. Additionally, some of the exploratory information gained by the raw data is a check of if and how many of the missing values there are, this is useful for the pre-processing and cleaning of the dataset so that before any models can be implemented an appropriate collection of data is being fed into the system.

Another step that is part of the exploratory investigation is the statistical observations that can be found for the attributes. The statistical values identified would be of the locations of the minimum and maximum temperatures recorded. Furthermore the average temperatures, uncertainties minimum and maximum values too found for each country. Importantly a test that will need to be performed as part of the screening process for the dataset is the test for correlations between the attributes and if there is any multicollinearity in the cleaned data alongside the distribution of the data.

A significant feature of the dataset that is also key is the number of countries that exist. This is because analysis of the number of countries can indicate the representativeness of certain values found within averages and this could impact the findings.

Looking at the dataset as a whole is important. As the target attribute in this case is temperature it is likely that there will be some form of trend. It will be interesting to observe the direction of the trend (if any) because the prediction element of the task will reflect some form of trend it will be intriguing for speculative purposes of what could be expected. Furthermore, the findings and importance of them and what they mean to society would be based off this and the reflective explanatory power would be justified by the observed changes over time. This can be shown as line plots and specific observations depending on region by boxplot visualisation showing the averages, interquartile range, min, max values plus those that could be outside the logical representation as outliers.

Once exploratory analysis has been completed the prospective models in question (MLR, ARIMA and LSTMs) can be validated as appropriate and then be looked at being implemented and thus pre-processing to shape the data appropriately can be actioned.

**3.3.1 Preprocessing**

The dataset in question will need to be transformed in the aspect that from the imported original datatypes in particular the Date element. To be able to use the dataset’s time series component it would need to be ensured that the type is of datetime64 type. Observations over time can then be observed once it has been set as the index.

Given the data attributes the way in which the missing value are to be handled need to be considered. While the use of central tendency can be implemented to align values that would be relevant the particular data that is missing should be inspected before executed, this includes omission of the row.

Due to the vastness of the number of data points (over 8 million) there may be a requirement to compress the data for the sake of standardisation and to potentially meet the prerequisites of the respective models, therefore certain transformations may be necessary. Furthermore, since the data originates from the year of 1743 to 2013 the representation of data points and their significance may not be as valid as those recorded temperatures closer to the current period. This could be due to reasons such as accuracy of the recording instruments present at the time and also due to the historic nature of the time period. For this reason, the time frame from 1870 to 2013 will only be selected on a yearly basis. The reason for this is due to not only will the compressed dataset be easier to develop a model for; the accuracy potentially would be better as from the year 1870 we see the boom of the industrial revolution. Given the impact of fossil fuels and the impact of industry on temperature the fittingness of using such a data range poses a more viable question of the model’s explanatory power.

The problem of the variation of countries poses questions on how the data from one region will be relevant to a different region when composing a model. While it is interesting to propose a model that is generated for the whole globe, holding the conditions of all the countries it needs to be considered how the relevancy of how a localised average temperature determines the whole model, this is significant due to the number of instances of each country especially if there are not an equal number of instances. Therefore, after exploration of the number of countries and the variation in numbers after removal of missing values and the use of the allocated time window the sectioning of the countries will be investigated to allow a more realistic model.

**3.3.2 MLR model development**

The formation of the MLR model will be via the library of SkLearn which is a well-known library that is used for conventional machine learning models to be implemented. Knowing that the prerequisite for the generation of the MLR model are that the model must have no multicollinearity, there must be an aspect of the absence of heteroscedacity and the data must be linear in nature. Since time series data is fitting of regression type models it can be logical to assume as there are multiple attributes, the MLR model can be used. The Ordinary least squares method will be used to test and fit the model given the testing values. Plus the White’s test will be used to test the following hypothesis as a pre-screen of the data:

H0 = The variance is equal within the data presented. H0 = σ²

H1= There is a significant difference between the variances, H1  ≠ σ²

Here a 5% significance p-value will be examined to assess the level of heteroscedacity.

Following this the normality of the residuals can be analysed whereby the level of normally distributed values can be examined. Using SkLearn the method of train-test split would be performed with a test percentage of 30% and the rest being used as training data.

The main metrics that will be used to assess the model in general is the R2, AIC, BIC and the RMSE. These will then give a good indication of the explanatory power of the model produced and therefore the overall accuracy. Particularly the RMSE will be compared to the other machine learning models as it is a standard metric used to evaluate the performance of the proposed models whereby the target attribute will be the AverageTemperature, this is true for all models. This is because an appropriate model that can accurately predict and fit data that enables average temperature to be forecasted raises applications where the future can be proposed given similar parameters to the model. Here given an average temperature uncertainty, latitude and longitude, the model will be able to approximate the potential average temperature.

**3.3.3 ARIMA model development**

The use of the ARIMA model will be performed by using the module from within the imported Statsmodels library. For the ARIMA models a pre-requisite is that the data needs to be stationary along with time series. Methods of imposing stationarity are rolling averages and differencing. These take sections and windows of time periods and the data to show a generalised trend of the data over a identified period of time. However, when using rolling windows there will be a time frame of the data that will be missing due to the window, this would consequently need to be addressed when building the model and defining the time period.

Once the data has been transformed to be stationary it needs to be assessed on whether the element of stationarity of the dataset is significant enough to be used to be a part of the model. The 2 modes of assessment that will be used are identification by visual plots and by using the Augmented Dickey-Fuller (AdFuller) test. The value generated from the test alongside the p-value will be used in conjunction to be compared with the critical values. If the value found from the test statistic is lower than the critical values we would then be able to say with a good level of confidence that the dataset is appropriately transformed and is stationary.

Following the confirmation of the dataset’s new suitability then the next step is to perform tests to try and generate the most appropriate model for the given data. The way this is achieved is by running autocorrelation and partial-autocorrelation for the data. This consequently generates the indicated lags and therefore attributes for the autoregressive and moving average element of the data. The fitted results for the whole data are then observed while looking at the AIC, BIC and the significance of the attributes being measured in the model. Observations that are checked are the residuals and the distribution of the residuals.

Once this measurement has been executed the final training and testing of the data is performed and predicted values are generated from a windowed section of time. The predicted values are then compared to the test values to see the RMSE which is one of the primary evaluators.

The parameter that will be used for tuning will be the ‘trend’ element. This would add a form of bias and as a constant would affect the final model by its use in conjunction with the defined model coefficients.

After tuning the final RMSE values will be analysed to see whether the model has performed better or worst after being tuned and whether the baseline model is better.

**3.3.4 LSTM model development**

As with the other machine learning models that are being analysed the use of LSTMs will be by using the Keras library which will be extracted through Tensorflow. Similarly, to the ARIMA model, a window of data will be used to train and test the data.

Due to the nature of the LSTMs the expected form of the data that the Neural Network requires is of the 3-dimensional type. Furthermore, the implemented training of the data will require a period of past verification to lookback on the previous data points to allow a model of the future data points, therefore a method will have to be developed to allow this to occur and be a part of the training set. Once this has occurred the training and testing data can be passed through, consequently enabling model formation.

Some of the key parameters that will be tuned will be the number of nodes (to enable the learning) and the number of Epochs (for the number of passes through the dataset). Neural Networks notoriously have a longer than usual train time and this has been a factor in some of the studies, therefore it will be interesting to consider how well the data is used. It is important to consider amount of training that is required and number of nodes which is representative but does not over or underfit the data fed to the system. As a baseline experimentation of 100, 50 and 200 neurons will be conducted and also epochs of 500 and 1000 respectively will be evaluated. Another observation that will be investigated will be the loss function that the LSTM will generate. This will be by using the mean squared error and visualised to assess the over/underfittings.

If there has been sufficient learning by the LSTM then a stopping method will be invoked, this is key because it will aid in time saved and constant running of Epochs which in turn will reduce the computational cost. As the target in question is the AverageTemperature, this will be the aim of the LSTM therefore a Dense parameter of 1 will be given as this is the output that will be given from the linear activation.

Once the model has been constructed the prediction phase can begin. Once predicted, the comparison between the test and predicted can be evaluated by RMSE. Then graphical visualisations can be formed.

After all of the models have been executed and the results have been obtained there will be an evaluation and comparison of each model and how well they performed. In addition, there will be analysis into how suited the dataset was given the challenge plus what the results infer. Then there will be a review on drawbacks of using the certain types of models and where some models exceeded or didn’t exceed expectations. Finally, there will be a overview of what the findings mean for the data and how it can be implemented to benefit society.

The next section is implementation and how the data was handled, models were constructed and changes that needed to be made. Given the complexity of the challenge and the data in general there will notably be a need for a contingency made, this is particularly true during the transformation phase because naturally data may not always be clean and ‘user friendly’ therefore room for changes will be discussed in the next chapter.

# Chapter 4

**Implementation**

**4.1 Exploration and Preprocessing**

From the initial design the processes carried out for the analysis and the development of the machine learning models would require scrutiny of the dataset in question. The following information is regarding the data, what was performed, found, and the results obtained.

The initial exploration was such as the number of entries found which was 8235082 data points. Furthermore, from this the target value which was the ‘AverageTemperature’ was inspected. Here it was found that the minimum temperature recorded was from a Russian city called Norilsk recorded at -42.7 °C (1 d.p.) which was found on the 1st of February 1979. By contrast the maximum temperature was found to be 39.7°C (1 d.p.) which was in Warqla, Algeria on 1st July 1761. These values found at the specific data points seem logical due to the location of the cities and their proximity in relation to the equator.

Following this the project direction and the use of time series data needed to be addressed therefore the data types were identified where although the AverageTemperature and AverageTemperatureUncertainty were of type float64 the rest remained as objects and therefore the date attribute needed to be converted to an acceptable type and so this was datetime64, a check was performed to ensure that the type had been changed. Secondly another transformation was required, this was to make the index of the DataFrame the date. This was so that the time periods can be analysed. Another important element that needed to be investigated were the missing values. It was found that the missing values were from the attributes that had temperature instances. Typically, where the central tendencies may be used to fill these values it was deemed necessary to remove these values. This was because as is the uncertain nature of weather, to model it and replace values that might not be representative of the time period (months) would distort the dataset and therefore yield unrealistic models. It was discovered that only 4% of the overall data was missing therefore given the range and quantity of values, the difference of 4% wouldn’t be as impactful as compared to a percentage such as 10%.

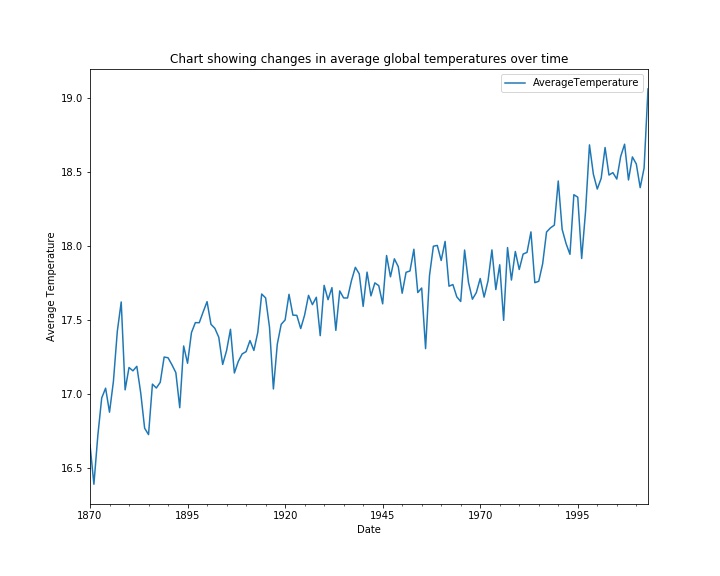
Following the elimination of the missing values another exploratory examination that is observed is the whole dataset which is resampled on a yearly basis that reflects the overall picture of how the temperature changes have occurred over the observed 143 years. The resulting visualisation indicates that while there were peaks and troughs in the average temperature change there is an overall linear trend and a positive increase of average temperature over time. There can be a multitude of explanations of these findings from global warming and the triggers that humanity has caused such as the extensive use of burning fossil fuels which may have been a contributing factor. Nevertheless, the growing trend is one that is alarming therefore the sooner it is addressed and actioned the better the gradient of temperature change will be (Figure 4.1.).

Figure 4.1 Resampled data for yearly average temperatures showing an upwards trend in temperature increase.

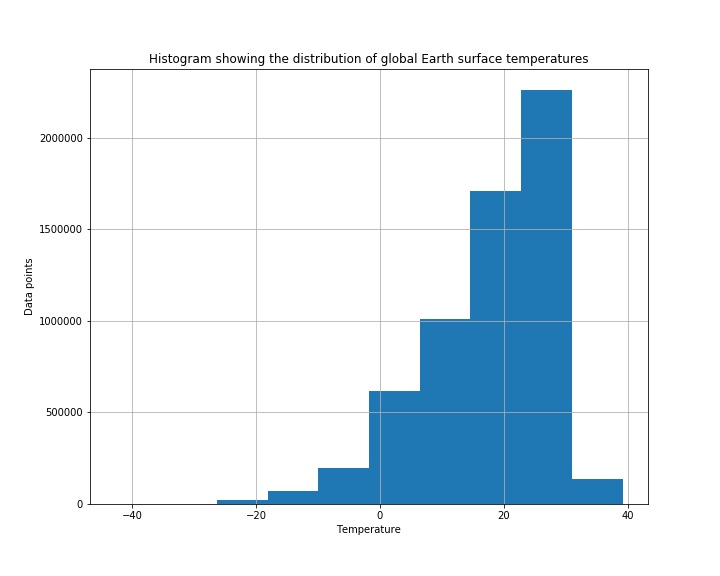
Another useful way of examining the differences that are occurring in the data is by performing the describe function which showed the differences of all the countries including how many values that were from the respective regions, the average for that specific country and the critical values. Such a feature was useful to gain further insight for example the average temperature uncertainty was found to be somewhat higher in those countries that aren’t as well developed. This could be due to technologies that are made available in recording the temperatures. However, another factor could be the fluctuations in that given region which shows greater swings in temperature. A quick observation of the distribution of the data shows a negative (left) skewness (Figure 4.2).

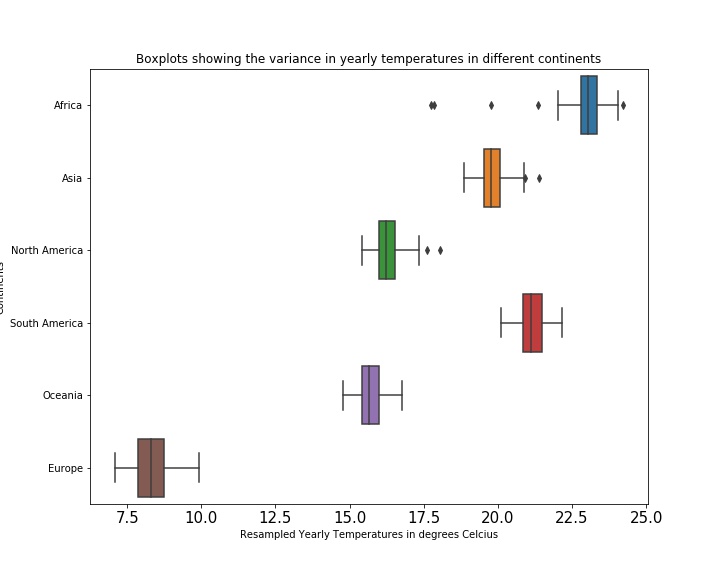
Figure 4.2 Histogram showing the range and density of temperatures recorded globally.

Typically, the latitude is seen to be associated with warmer countries that are situated more equatorially however to use the longitude and latitude features the cardinal direction of the latitude and longitude would need to be removed to allow its relevancy to be taken into consideration. For the sake of models such as MLR the pre-requisite was using data with as minimal multicollinearity as possible. Due to this the independent correlations between the target variable of AverageTemperature to Latitude, Longitude, and AverageTemperatureUncertainty respectively was analysed and it was found that the biggest correlation was between AverageTemperature and Latitude. This was an expected result due to the relationship and nature of the attributes (Table 1, APPENDIX B). The correlations between Longitude and Latitude were not performed due to the nature of the independent locator attributes.

To check if any prior transformations are before moving on the multicollinearity between the variables was measured where the following values from Table 2 (APPENDIX B) were found. Where all values are below 2 it can be suggested there is minimal multicollinearity (if any) using the VIF values found, where anything over 2 and limit of up to 5 would be concerning.

Due to the high variability the key transformative step that was taken was sectioning the 159 countries into respective continents. This would then allow model development for more localised regions. To be able to perform this operation the step that needed to be taken is to identify the different countries available and then take sections of the cleaned data frame that were from the specific countries. Therefore, there would be 6 separate data frames that would be trained and tested on which were: Africa, Asia, Europe, North America, Oceania and South America. Once this had been completed the removal of the cardinal points could be executed and therefore the next stages by which model development is performed is able to be carried out.

The target attribute of AverageTemperature was then explored for the respective continents and the boxplots were made to reflect the range and averages of the different countries’ resampled temperatures. As Figure 4.3 shows there are differing variances in the various continents. A feature that all the boxplots seem to have are symmetrical properties. This shows that there is a varied spread of data by having approximate ranges of 2 degrees Celcius.



Continents

Figure 4.3 The boxplots of the different continents and how their data variance differs.

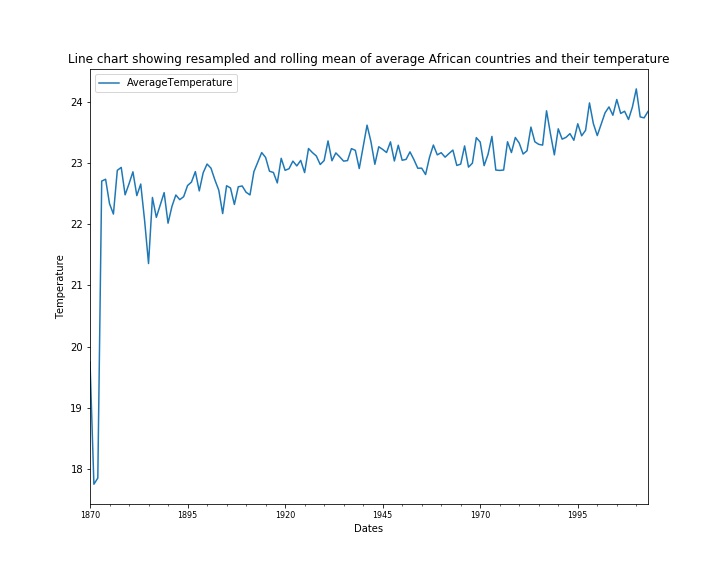
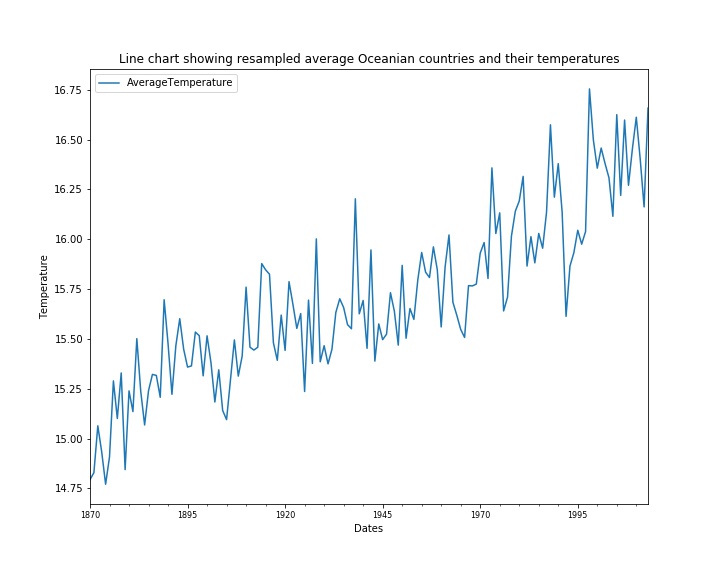
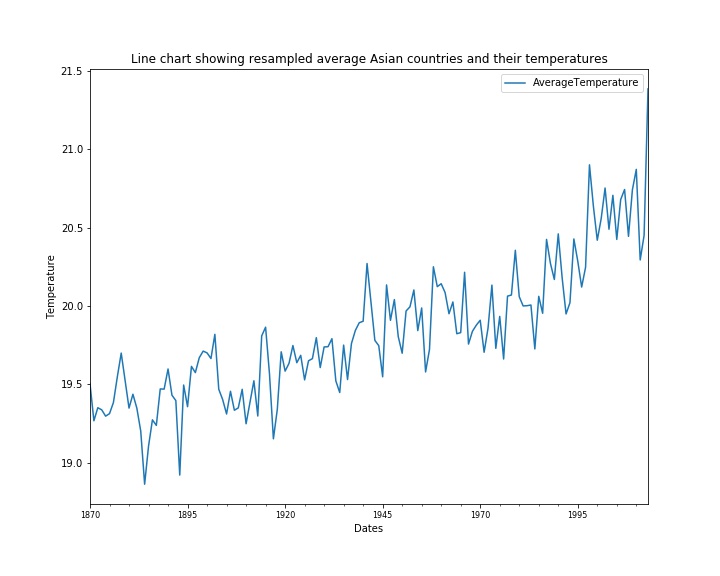
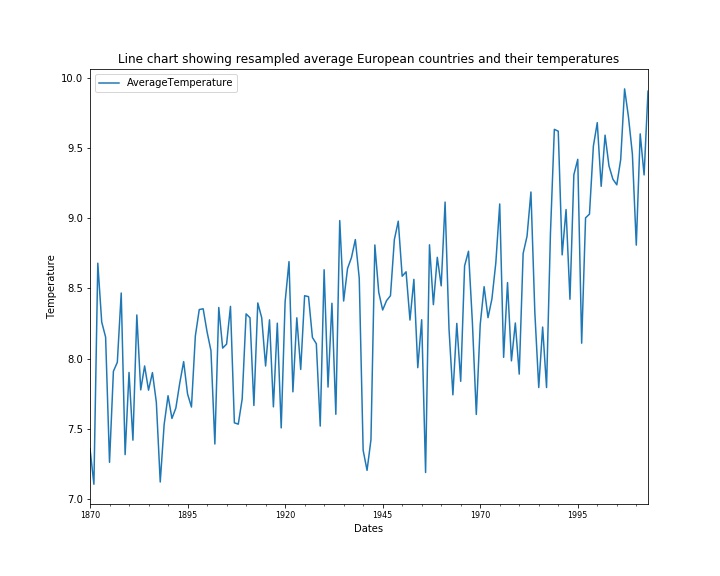
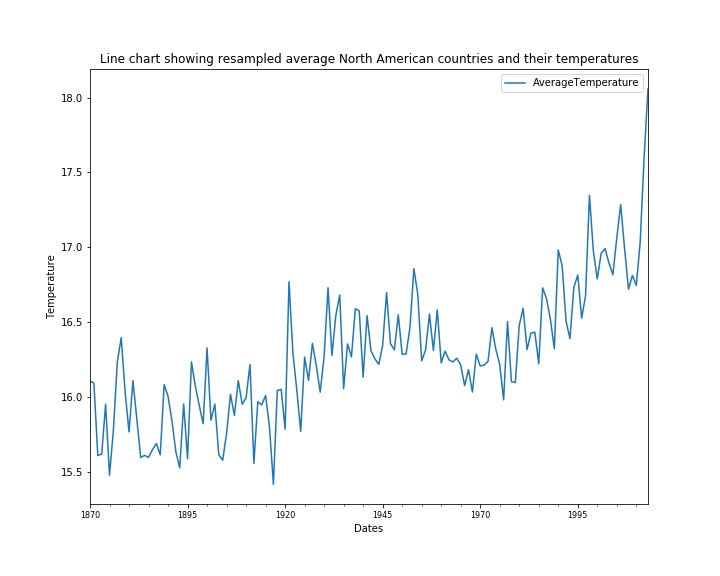
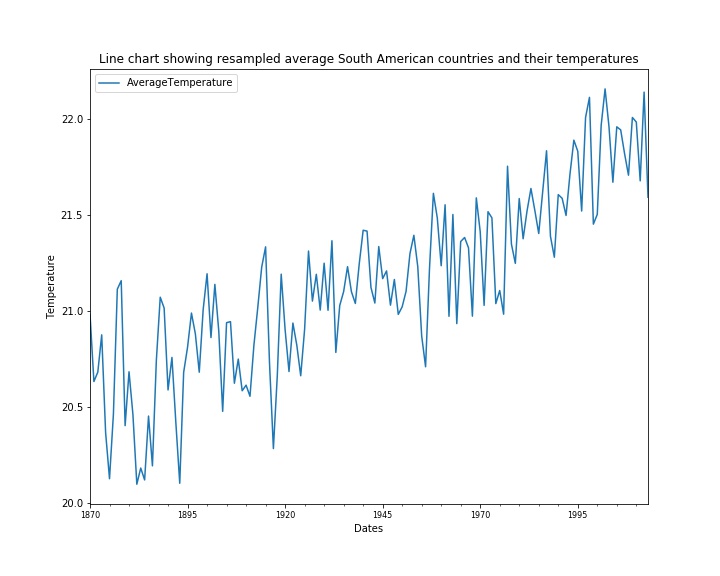
Noticeably the boxplot for Europe is much toward the colder end of the temperature spectrum, while Africa due to its locality has a boxplot which is based towards the warmer end however it has greater outliers. This could be due to places with higher altitude where the temperature may be cooler or places closer to the South Pole than others such as the difference between Cameroon and South Africa where South Africa at the tip of Africa is based furthest away from the equator in Africa.

**4.2 MLR Model Execution**

The rolling mean for this was taken to account for the varying fluctuations in temperatures and therefore to generate a more accurate model specifically for Africa. This is enabled due to there were not any seasonal effects of having yearly modelled data therefore the moving average implemented allowed for smoothing. Interestingly this method will be implemented and used against other models generated to compare its affects which just used the resampled data.

As per the full data (FIGURE 4.1) there is a clear upward trend across time for all continents where North America sees the slowest rise in average yearly temperatures and Oceania sees the greatest rise over the same time period. We notice a large trough in the Africa data which could signify a window of 5 years where temperatures recorded were abnormally low or potentially the large increase was from human interactions with that region. However it can be noted that the greatest fluctuations in the temperature data is seen from Europe. Europe shows to have the largest swings in the data recorded. This could be due to the varying climates whereby the where it is seen typically the extreme cold winters and extremely prolonged warm summers is over a large period of time with many varying countries that add to the array of recordings which is reflected where the greatest number of instances of data is from Europe.

Figure 4.4 Resampled yearly Temperatures across time for the different continents



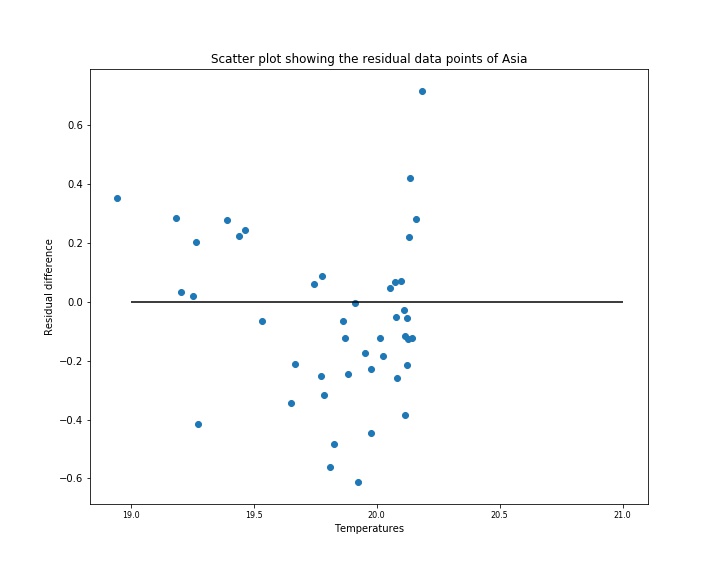
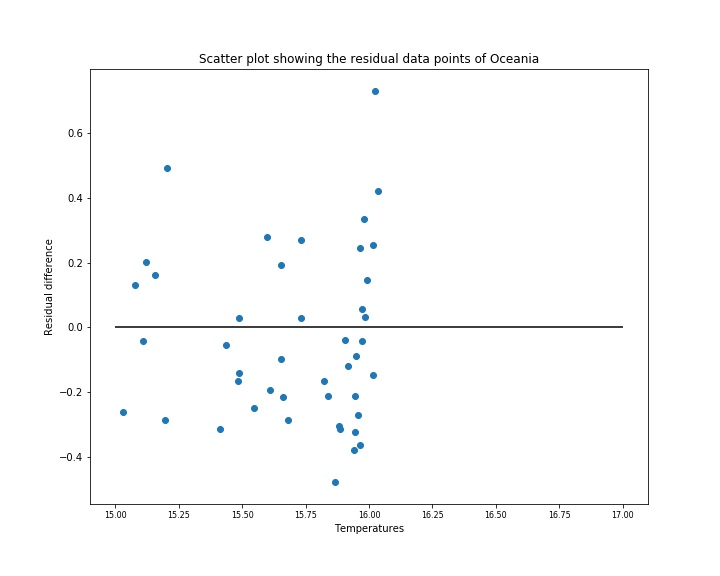
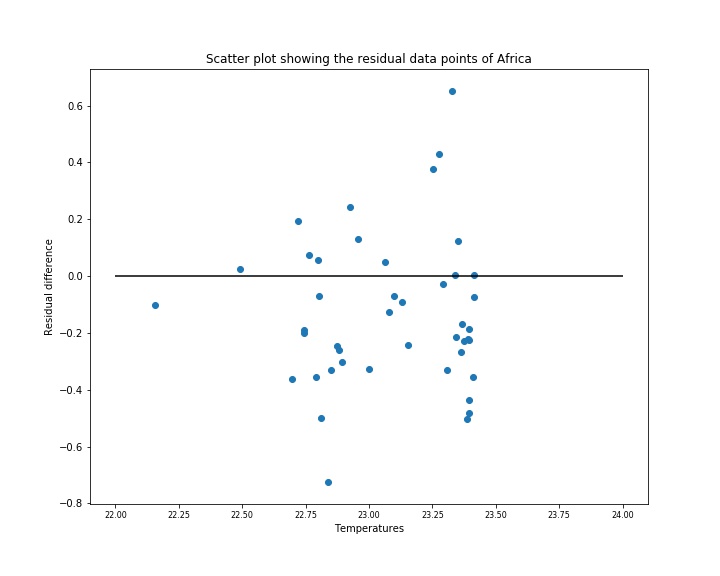
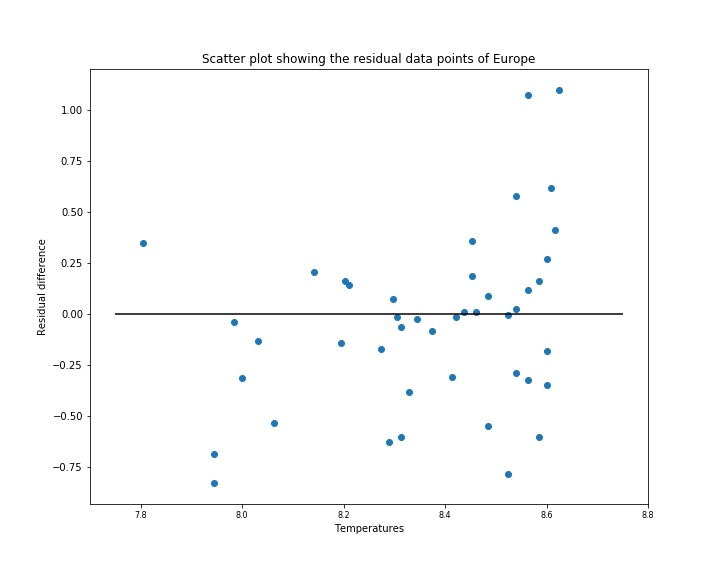
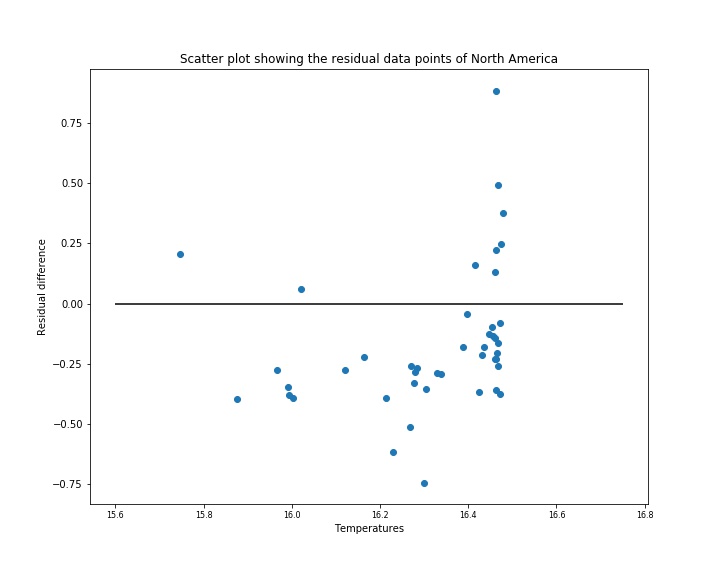
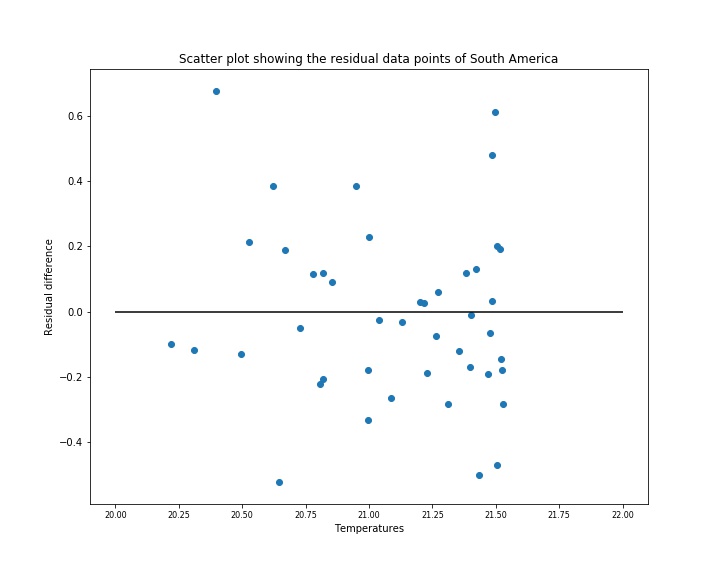
The training and testing split method of 30% as was used throughout all models was used with the incorporation of the True normalisation parameter such was the requirement of MLR. The resulting coefficients and intercepts which define the MLR equation are given in Table 3 (APPENDIX B), thus allowing the predictions to occur and then analysed by the Ordinary Least Squares estimation method. Which takes the regression line and analyses the fitting of the line to the data points thus generating a measurement of the accuracy of the overall model.

The key measure of the performance and generation of the model was to analyse the scatter of residuals to see the difference between the observed values minus the predicted values. The absence of heteroscedacity whereby the distribution of data is equal was analysed which checked the equality of variances. This is where it was seen that the distribution of data points for temperature and the differences were scattered around the 0 mark (Figure 4.2).

However, this visual separation while useful statistically a hard value would be required to assess for definite that the data meets the requirements and to check how precise the coefficient values are. This can be analysed by hypothesis testing using White’s test where:

H0 = There is no significant difference between the variances , H0 = σ²

H1= There is a significant difference between the variances, H1  ≠ σ²

Figure 4.5 Scatter plots of residual values of different continents

From the scatter plots shown and from the p values found for the continents we were able to fail to reject the null hypothesis of equal variances for most of the continents therefore we can statistically confirm the absence of heteroscedasticity (Table 4, APPENDIX B).

Interestingly while most continents met the 5% significance level set for the evaluation of the hypothesis the Europe continent had a much lower significance level however sparseness can be seen visually. For this reason, the model developed for Europe may not be as accurate in comparison to the other models developed for the other countries.

Another check of the model developed is to evaluate whether the mean of the residuals equals 0. This is important as the average error found needs to be minimal to ensure a good level of accuracy. The sample quantiles versus the theoretical quantiles were plotted as an additive measure. The findings showed that the mean of the residuals for all of the continents were extremely low and therefore the values found validated the assumption of the distribution of the errors of the residuals for their mean (Table 5, APPENDIX B.).

Following the model formation the end RMSE value was found which was compared on the final model assessment between all models developed and then visualised showing how well the fitted tested values compared to the actuals. (Figures 4.3-4.5).

Figure 4.6. FITTED MLR MODELS FOR AFRICA AND ASIA

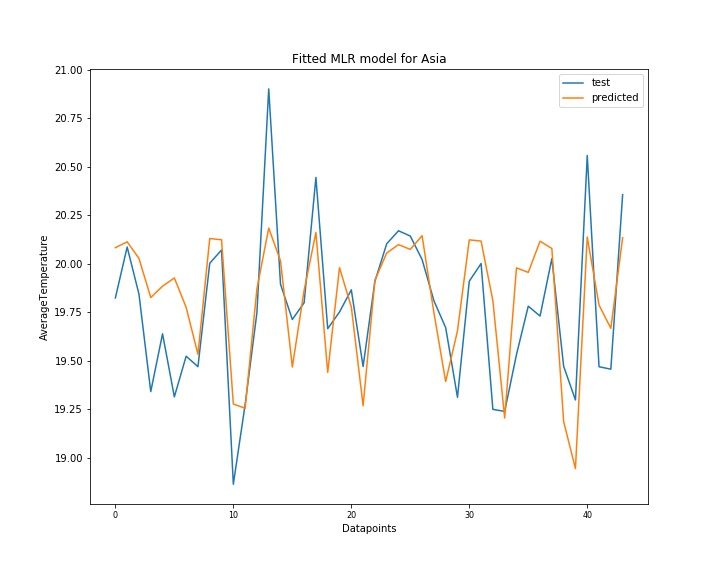
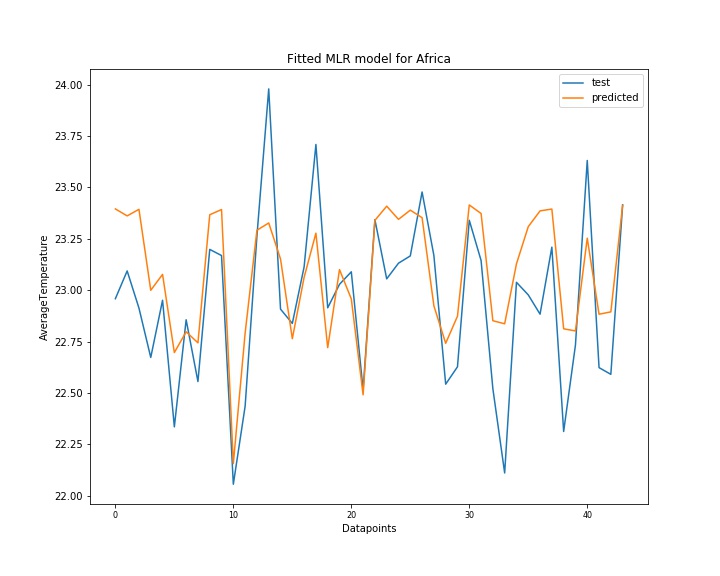


Figure 4.7. FITTED MLR MODELS FOR OCEANIA AND EUROPE.

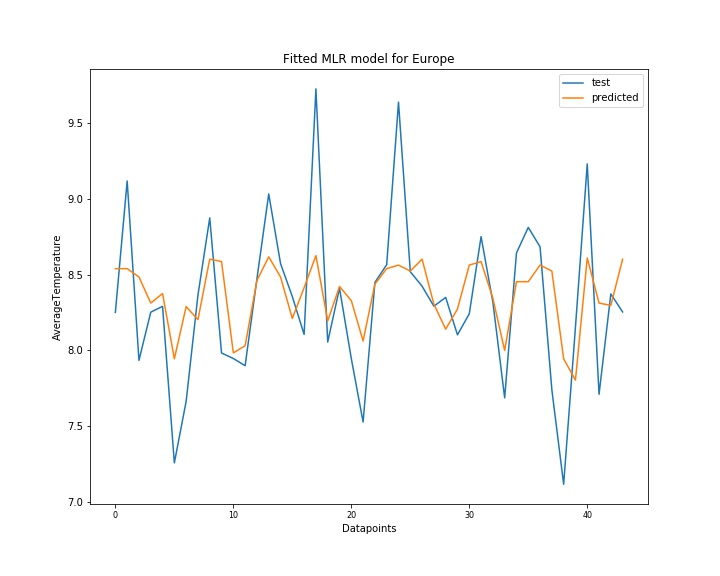
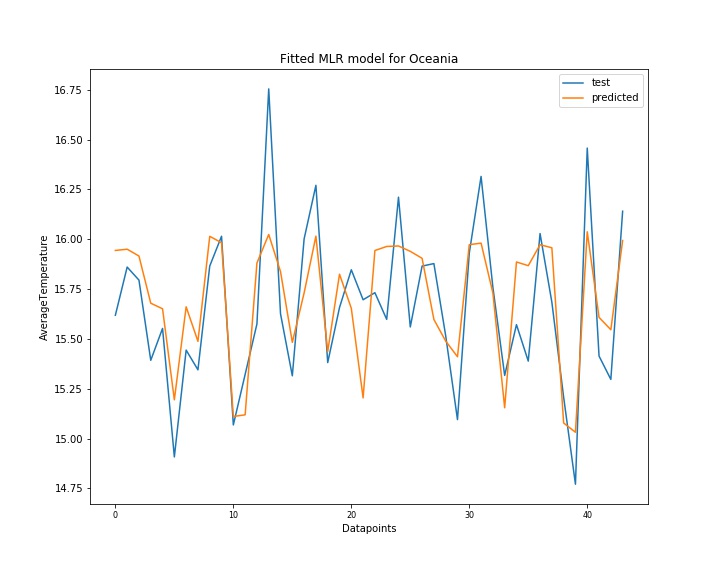
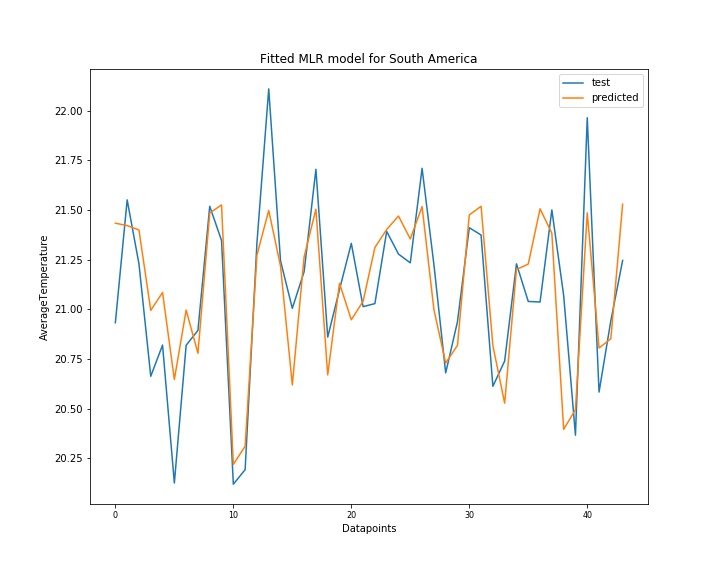
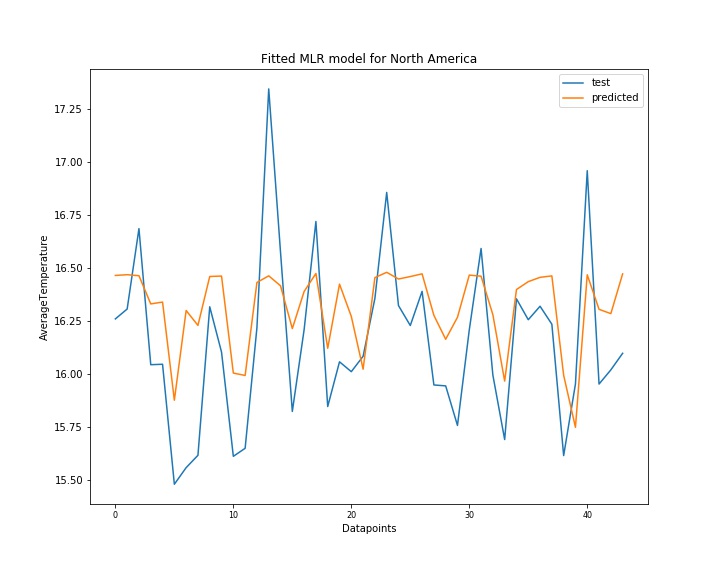


Figure 4.8. FITTED MLR MODELS FOR NORTH AND SOUTH AMERICA.



The AIC and BIC criteria’s which evaluates the model simplicity and assesses the goodness of fit using penalisation of parameters shows that the better and lowest values are for the Oceania model whereas the maximum are for the Europe model. This could coincide with the levels of potential heteroscedacity and the variability of the particular data. This assessment which aims to find the lowest value is an indicator of information loss whereby the higher the score the greater the loss and consequently the worse the model which could indicate elements of model overfitting which in turn is not generalisable and less accurate. These are the additional features that are taken into consideration for model evaluation. Similarly, Europe has a lower R2 than most. However, Africa and Oceania have the better values which gives a better understanding of the level of prediction and explanation achieved for the model for those respective continents. An explanation of the results could stem from the scale and ranges of temperatures collected and given the size of the land masses. An example would be since Oceania has countries that are located in similar regions the weather conditions may have similarities and therefore the differences in the errors from predictions are lower which seemingly indicates a better model. However further investigation by different models can be used to evaluate this claim.

**4.3 ARIMA Model Execution**

Following the similar basis, the ARIMA models that were developed used the concept of individual continental models to be generated. Specifically, for Africa due to the vast variability seen from the plots the data was smoothed y means of using the resampled as moving averages. This required a window of time to be used to look back on the previous data to model the moving average was placed at 10. This 10-year window for each data point was used to consequently form a more stable line of fit.

Due to the non-stationary nature of the data for the respective countries differencing was applied which used consecutive points and found the difference between the values. This consequently formed more stationary data. To assess the level of how suited the newly differenced data was the Augmented Dicky-Fuller test was performed which for all continents showed the ADF statistic to be lower than the critical values. The hypotheses for the testing of this newly structured data was:

H0 = There is time dependency

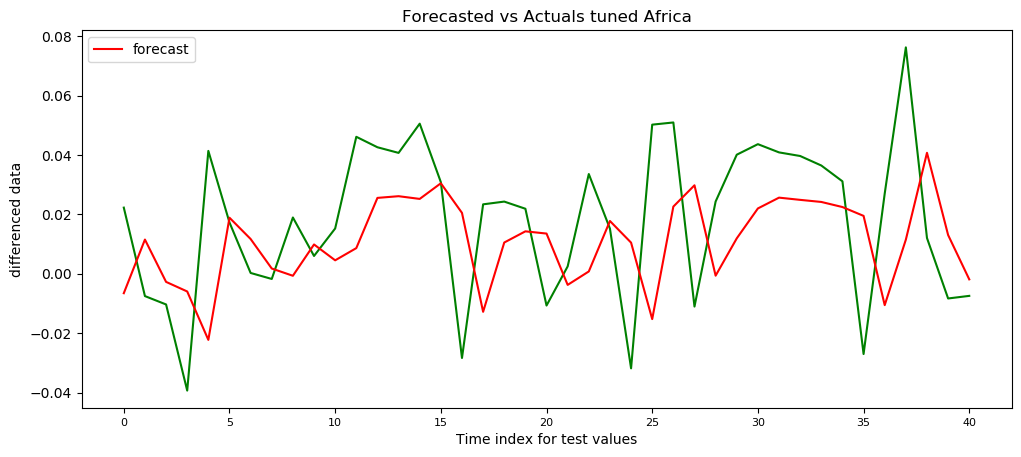
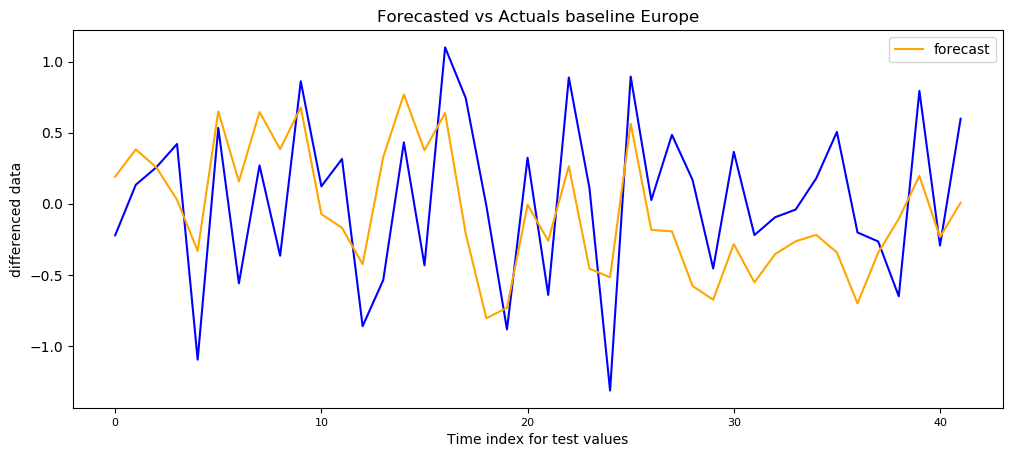
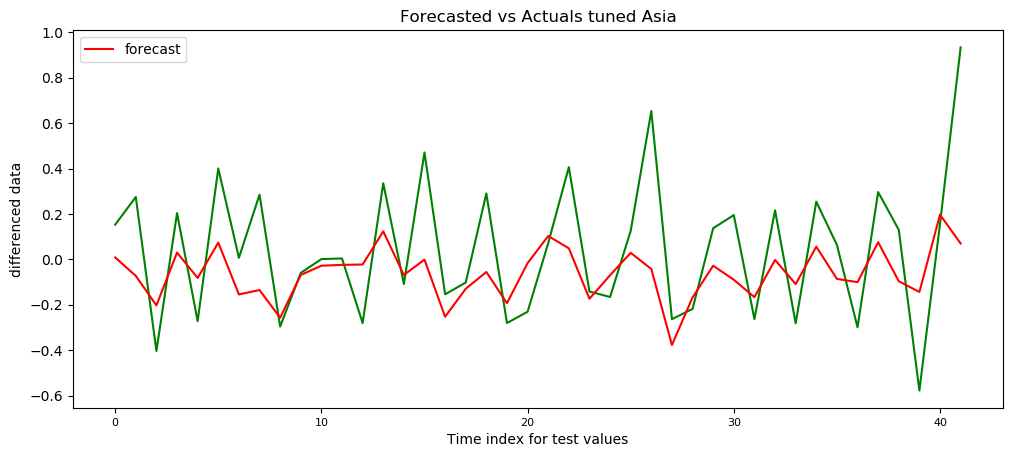
H1= There is a no time dependency

The continents where differenced maintained a p-value significance of less than 5%. This meant that we could reject the null hypothesis and the data was made stationary to statistical significance.

Following the stationarity check the next step was to determine the model using the autocorrelation and partial autocorrelation methods. The lags determined indicated the correlation between the data with different lags. The lags in turn specified the dimensions and order of the ARIMA model. For the D value this was left at 0 for the model, firstly because some of the data had already been transformed in such a way that moving averages were already established. Furthermore, given that the data had already been condensed a lot to further implement another average model may underfit the true values and findings that could arise. Consequently, on consistent occasions the featured lags that protruded past the 95% confidence interval were of lag 1. This is for the ACF and PACF which constituted the AR and MA elements of the model. This generated a model of (1,0,1).

Due to the nature of the rolling window introduced there was 10 years of the data missing from the original 1870 start point. For this reason, 30% of the years was taken as the testing proportion and 70% to be trained on. This incurred a testing date of 1973 to 2013. Once the model had been trained the performance by means of analysis were observed by the distribution of the residuals and how they were centralised unimodally (Figure 1, APPENDIX D).

The predicted differenced temperature values for the respective continents were made and then using the predictions visualisations were enabled to show how well the models fitted to what the expected data showed. In addition, the key metric RMSE was found. The tuning required addition of a constant to the models. Some models performed better with the addition of the ‘nc’ trend component of tuning whereas some baseline models actually performed better giving a lower RMSE. The results summarised in TABLE 7 show the different values found after analysis and tuning. The best models found were visually interpreted so that the fitness could be understood better (Figures 4.9 & 4.10).

FIGURE 4.9. BEST FORECASTED ARIMA MODELS 1

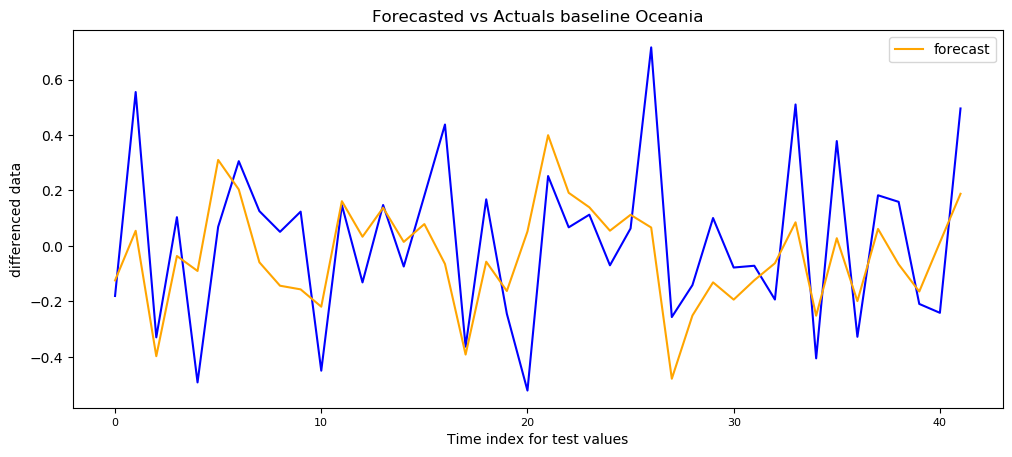
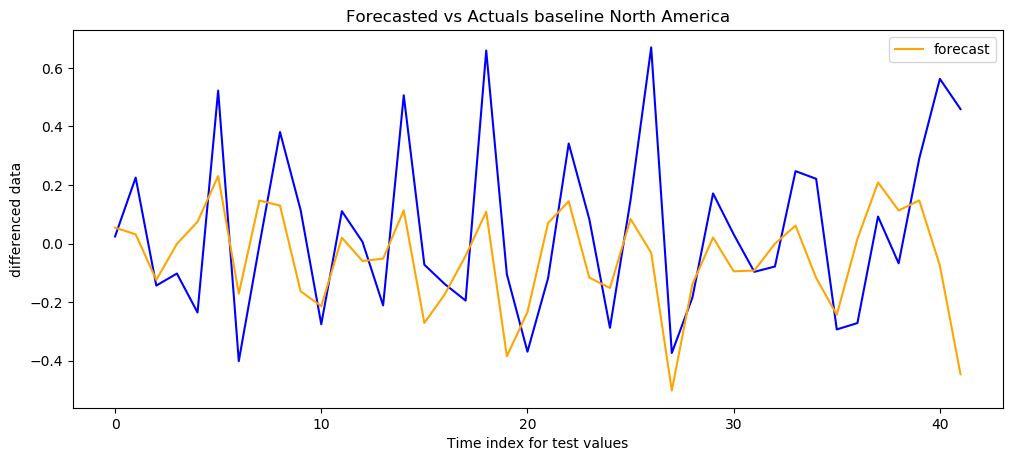
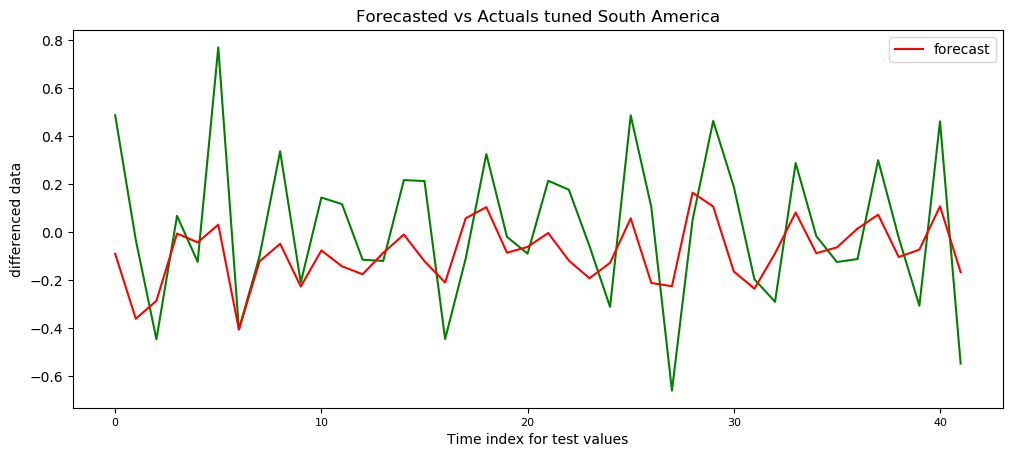


Figure 4.10. BEST FORECASTED ARIMA MODELS 2

The results show that while Africa had the biggest swing in the AIC and BIC values the Americas performed well in this respect. However only a third of the models generated a model that had a lower RMSE as tuned over the baseline. These were South America and Asia. This showed that the majority did not require a constant to be added to the models.

Similarly, to the MLR the worst performing model was for Europe which the highest RMSE however it could be suggested that give the volume and variation of the data the performance might be adequate relatively speaking whereby if other continents had such scale their RMSE’s might’ve been similar. Nevertheless, with the best RMSE coming from Africa, again potentially explainable due to the smoothing that was performed prior to differencing. With the second best RMSE generated by Oceania, again like the MLR results is similarly good at performance with the models.

**4.4 LSTM Model Execution**

For the use of LSTM models there was a requirement to ensure that the correct libraries were installed and this was primarily Tensorflow which held Keras as the library being used from the Tensorflow backend. Yet similarly to ARIMA models the test cases were from 1973 onwards to 2013, maintaining the 30% split of train to test data.

The sequential model that is used is a linear stacked implementation of the LSTM formation where different conditions and parameters are added to it. In this case the Vanilla LSTM that is developed is using a single layer. Experimentation was performed with number of LSTM nodes and number of Epochs. These changed how much the data was trained and the number of runs through the dataset was made. The dropouts that were specified for the input and the recurrent states was maintained at 0.2. This was to enable a sufficient amount of learning to be applied.

The Dense layer was added which specified that end output shape give the single target which therefore a linear activation was used. For the compilation the optimiser ‘Adam’ was used due to the elasticity of the optimiser being able to adapt the learning rate with respect to the parameters fed, as opposed to a more traditional stochastic gradient descent. In addition, the loss was identified by analysis of the mean squared error, while the actual model’s performance was to be measured by the RMSE. A separate parameter was used to control the length of time taken to execute the LSTM this was via a call-back function of EarlyStopping which was developed and enhanced to allow enough training to occur, but once an improvement is no longer seen to halt the network and stop. This is significant due to given the capacity of running the experimental Epochs it aided in model execution.

Due to the nature of the new size of the data being analysed a batch examination was performed and passed through the fitting of the model for the training data. Here a size of 5 was used which given the overall sizes of the training sets approximated to 5% of the training size. This increment of testing along with the lookback feature allowed reflection and learning by the network by previous observations, this again was a step taken that also decreased the training time.

Once the predictions were made the MSE and RMSE were generated to see how well the model fitted and how much of the data had been explained by the model. This was then to be used when examining all models and their performance. Furthermore, using the predicted values, it was then made possible to observe the losses and the fitted model to the data. On different occasions that the LSTM is trained the different RMSE values are generated due to the different learning instances this is a reason where multiple runs can generate slightly different RMSE values.

The model losses were plotted so that it could be visualised on how the performance of the model so that the fitness of the model could be evaluated (Figures 4.11 & 4.12.). The plotted loss visualisations show that the Asia and Oceania models have some degrees of being underfitted where North America and the model for Europe had performed better in this respect. The strength of these losses indicates even though there are some degrees of change between training and testing sets while they are small there is room for improvement, yet none of the losses increased over the epochs which indicated there had been an element of sufficient learning that had taken place.

The ways to use this as information would be to evaluate whether enough training has occurred. If not, then to increase the number of Epochs and potentially nodes trained with. Additionally, perhaps by adding more layers it allows further training to occur.

Figure 4.11 LOSSES FOR VARIOUS LSTM MODELS

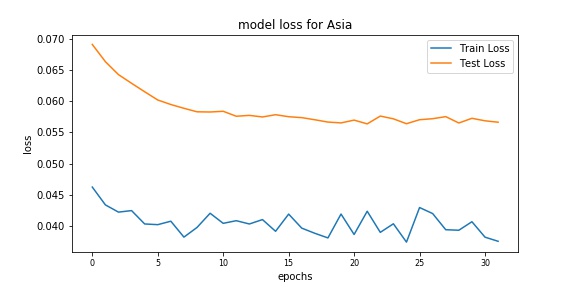
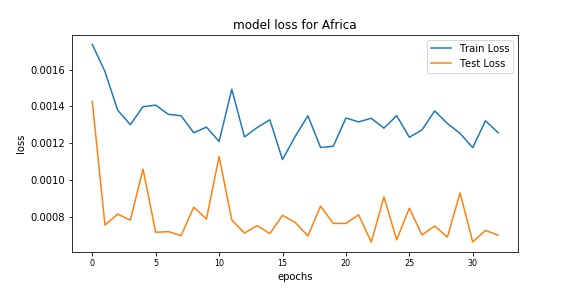


Figure 4.12. LOSSES FOR VARIOUS LSTM MODELS

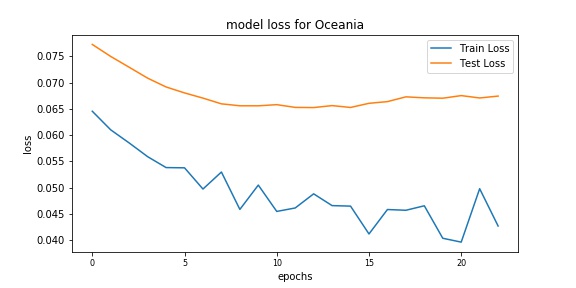
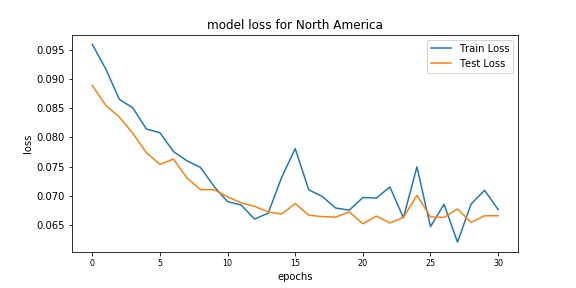
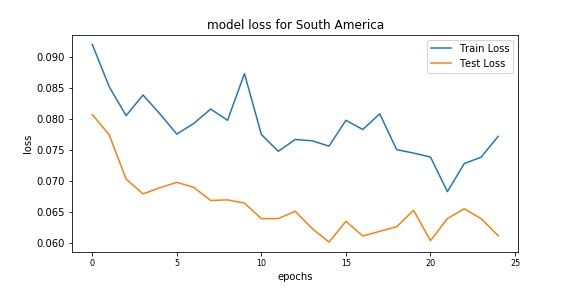


Figure 4.13. PREDICTED VS ACTUALS FOR AFRICA AND ASIA LSTMS

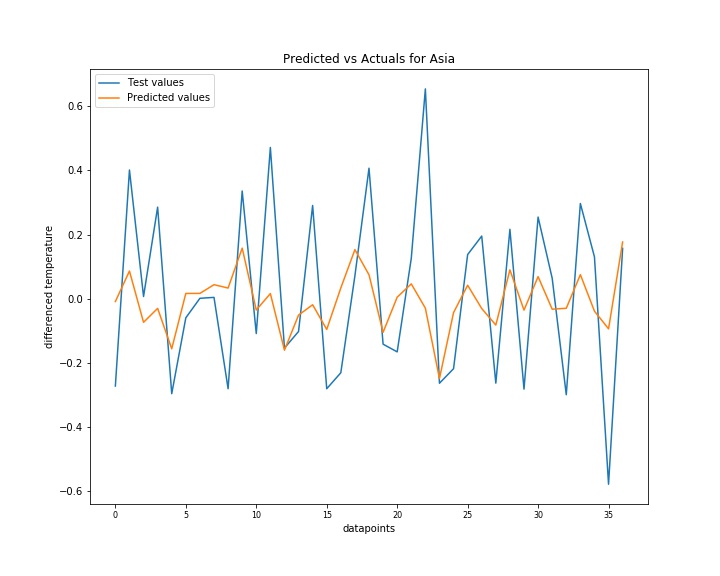
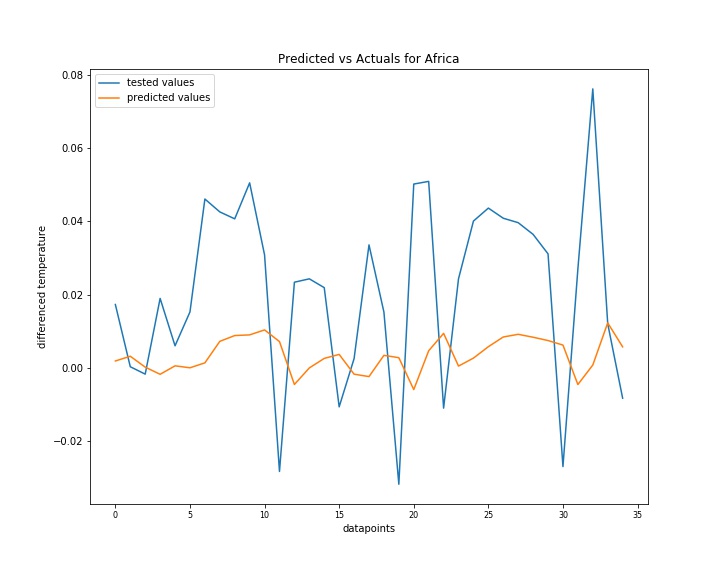


Figure 4.14. PREDICTED VS ACTUALS FOR EUROPE AND OCEANIA LSTMS

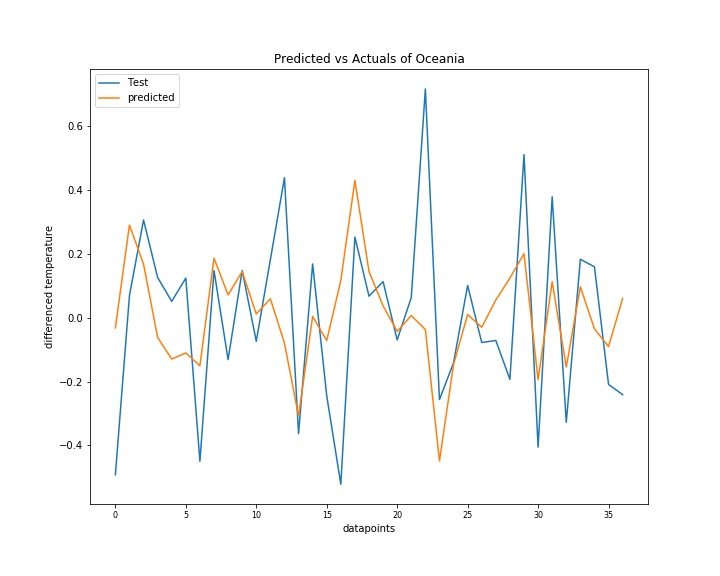
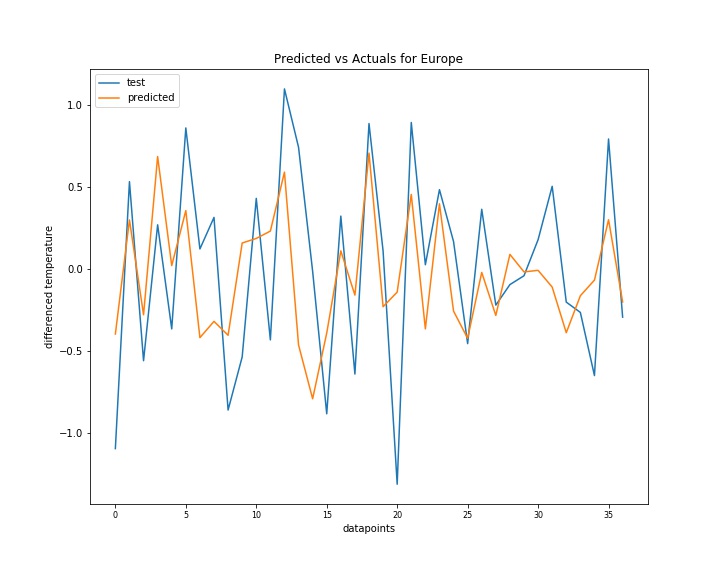
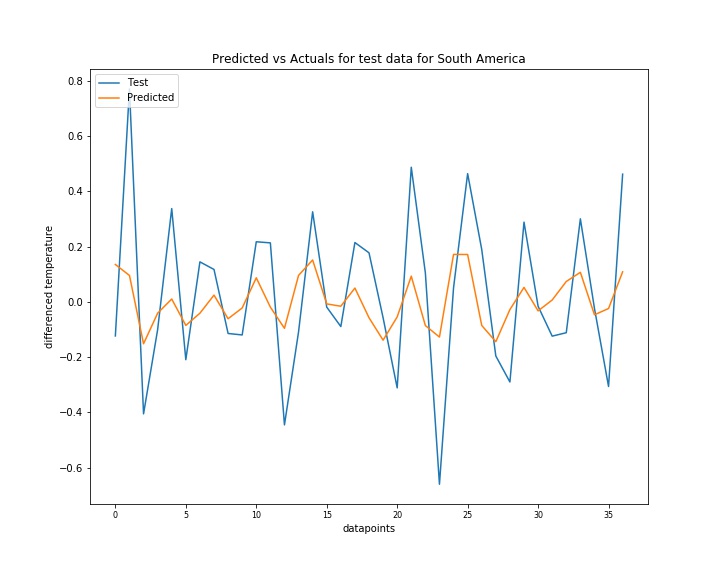
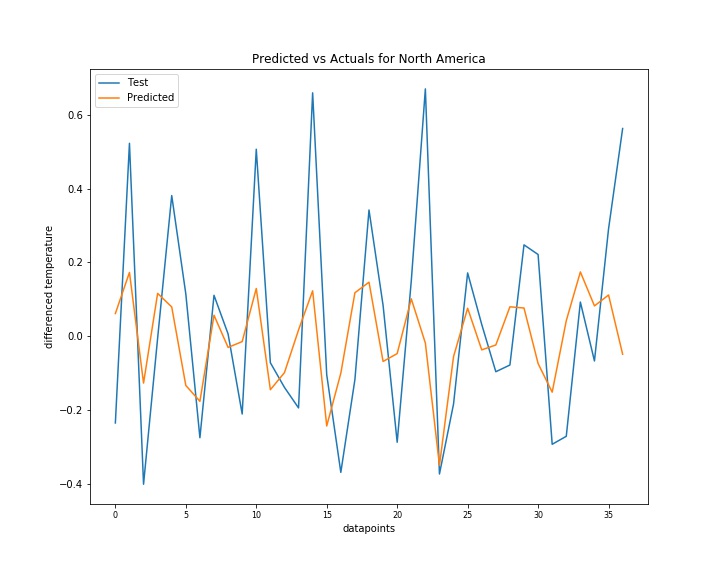


Figure 4.15. PREDICTED VS ACTUALS OF NORTH AND SOUTH AMERICA LSTMS



The images from Figure 4.13 shows the models developed for Africa and Asia. It is clear to see how generalised the model is. The variation in the peaks and troughs of the tested values is difficult to evaluate since the past data summarised is not as informative when modelled in regard to the tested values. Whereas Asia’s model was a much more predictive model that is developed for the data points. However, both required further training. It shows how using the rolling mean for data points in Africa has impacted the model. By contrast the models in Figure 4.14 showed a detailed level of modelling, however while the model for Europe seemingly was ok the RMSE found was one of the worst found. However, the model for Oceania generated a well fitted model with a low RMSE which indicates a decent performing model. Interestingly there was a lot of variation in the North American differenced test values. This therefore meant on average the model developed could not account for as much change as shown in the peaks and troughs in Figure 4.15. However South America had slightly less variation and produced a fitted model which minimised the predicted differenced temperatures. The TABLE 8 (APPENDIX B) which shows the comparison of the experimental efforts in finding a best fitting RMSE given changes made to number of nodes and Epochs. Interestingly the models which performed the worst (Africa and Europe) had their best training parameters as the ones with the greatest number of nodes and the greatest number of Epochs. This suggests that there may have been potential for further Epochs run to improve the models. Furthermore, the value of 50 nodes was favourable for the best performing models which showed that it might not always be a requirement to excessively train the model which would lead to overtraining and overfitting.

On the whole the LSTMs modelled the data in a mixed variety of ways. Yet it was clear to see the LSTM models use the test data and learn about the changes across time. Given the lookback and batch size features the shaped models are representative of the data for the specific continents. The element of variation within each continent was clear to see, this perhaps due to the vastness of such continents.

# Chapter 5

**Evaluation**

**5.1 Examination of Results**

From the aims of this project whereby the modelling of global temperatures was examined. The deviation to assign the whole dataset into smaller datasets that are region relative was necessary. While it could be considered to underfit the raw data, however it is representative to those specific regions.

The aspect of using multiple machine learning algorithms for the model development was beneficial as it enabled different approaches which varied from complexity. The flow of work done from pre-processing, exploration, feature manipulation and tuning allowed an incremental breakdown of the prior work and screening that needed to be performed to allow model building to occur. Following this the least complex model (MLR) was developed first and the most complex (LSTMs) last. This showed the gradual improvement of the results. The tuning that was performed generated a mixed array of results, where some performed better than others, this is especially the case with LSTMs where although the RMSE’s found were lower than MLR and ARIMA models.

The MLR models developed which used multiple elements of the data were the most simplistic model developed. However, it can be questioned of how relevant the certain elements of the data are. Most importantly is the weight of each attribute. An improvement of this model developed is to add on a feature of weights to each element. Here there would be much greater weight added to the AverageTemperatureUncertainty and Latitude as these are key indicators that allow the model to take its shape. Furthermore, this might’ve yielded better coefficients and thus a more accurate model overall. Plus, the element of multicollinearity which had been examined of the whole dataset was carried forward whereby none was observed was used in examination therefore it was assumed that this was still true. Due to the simplistic nature of the MLR the run-time performed was quicker and did not place a heavy computational burden.

The ARIMA models that were made were slightly more complex however since the yearly data was observed over monthly there was no need to decompose the data however the differencing provided a way to consequently remove the stationarity by differencing which in turn allowed the new data to be analysed.

The interpretability of the ACF and PACF was challenging because the decision made as to whether the lags that just about did not pass the 95% confidence interval was concrete enough or due to chance needed to be established, with the certainty being at lag 1 for the AR and MA components respectively.

By using the residuals, it was good when analysing the model’s predictive capacity due to the variations in the difference between the predicted vs actuals. Additionally, understanding the distribution of the variation was significant as they were unimodally centralised. Once the model had been fitted to the given parameters the tuned element could occur. This could have also been another improvement whereby if further parameters had been altered there could have possibly been a better model generated. However, with over tuning comes the problem of overfitting the data. Whereas if multiple tuning occurred while improved RMSE might have been seen the overall representativeness of the model would have suffered. The effects of altering the trend parameter showed some differences where some models were improved, and others were not. This showed the differences in the type of data contained within each continent.

The use of LSTMs involved the use of the Keras package which was a separate library that was learnt and use throughout this development which facilitated the model to be built. By using the interaction of the Tensorflow back end it was possible to perform what would otherwise be complex training models simply.

The construction of the observations by using the lookback was a familiar feature which took previous observations to help model the data therefore the concept of how to implement this was familiar. However, a critique of using the lookback feature was that whether the impact of using a value of 5 was sufficient. Given the new size of the resampled data for the continents it could be suggested that may have been a difference if the lookback was smaller, thus looking at more closely related intervals, whereby a larger lookback would generalise a lot more. Additionally, this is also true for the training batch size. Bigger datasets and series containing more data points would logically require a larger size of batch and lookback to reduce overfitting and improve speedup. This could have been a change that could be implemented to improve the quality and specificity of the models however experimentation would be required to establish how small the lookback and batch size should be.

Due to the nature of the deep neural networks to form a model to train and become developed enough there is a requirement of having enough data to learn from. Another improvement that can be addressed is the frequency of data points used to develop the model. Having many data points when combined with a sufficient number of nodes and additional layers will allow a good amount of training to occur. However, notably the key issue with such deep learning neural networks is the issue of computational cost. The time taken to train hundreds of thousands and in certain cases millions of data points across multiple layers over many Epochs would be extremely time consuming and the cost reward of developing such a model may not be feasible. In such cases operating on a shared memory distributed system would possibly facilitate this. Hence a positive aspect of the developed LSTMs used here are the time taken to execute the models whereby all took less than a minute to execute and allowed predictions to occur. Yet perhaps less nodes would have been more appropriate to use given the simplistic nature of the transformed dataset. However, the loss made was not excessive and remained less than 2 significant figures.

In addition, due to the differing number of values within each continent the instance of developing different models was beneficial to obtaining lower RMSE scores. As the specific nature of the different continents had different properties as outlined in their ranges and means the evaluation and compilation of separate models for the specific continents improves the credibility of the different models. This is as seen in the variation of RMSE scores for different tuning parameters.

**5.2 Comparisons**

In general, the different models had their different benefits and therefore it was interesting to evaluate their behaviours for each continent. The impact of the rolling mean included in the Africa data which was required to be implemented due to computational load and also due to the requirements of the respective models was interesting. While this might have indicated a potential improvement for the other continents the interpretability of generating a representative model that did not underfit by means of averages rather than specific values used, therefore the frequency of the level of how underfitted the model was due to over generalisation, here an improvement could have been made by addition of more Epochs to facilitate more training. However, it can be observed that Europe had performed the worst in model RMSE values, this could indicate the variability in European counties such as the weather differences between Sweden and Greece. Here the latitudes are different as their locations differ and experience different climates. By contrast Asia, which performed one of the best have counties located in a closer related latitude, this signifies a lower variance in the data and thus a model that was able to fit better. This is supported by the Oceania models which also performed well.

Where LSTMs largely performed better than ARIMA models but the MLR models generated satisfactory results where in some instances it outperformed even the ARIMA models. This could be due to the greater number of attributes that were used in model generation. While MLR was the simplest model to develop the most naïve model would potentially be the ARIMA model. This is because not only did the MLR model take into account multiple attributes, the ARIMA model based off the lags was more conservative given the uncertainty of the lag’s protrusion by chance. However, the overall fitness of the models is seen in the figure below (Figure 5.1) and values in Table 5.1.

|  |  |
| --- | --- |
|  | Best RMSE (3 d.p.) |
| AFRICA (LSTM) | 0.026 |
| ASIA (LSTM) | 0.239 |
| EUROPE (MLR) | 0.426 |
| OCEANIA (ARIMA) | 0.248 |
| NORTH AMERICA (LSTM) | 0.256 |
| SOUTH AMERICA (LSTM) | 0.247 |

Table 5.1 Best RMSE for the various continents and the models which had the best score.

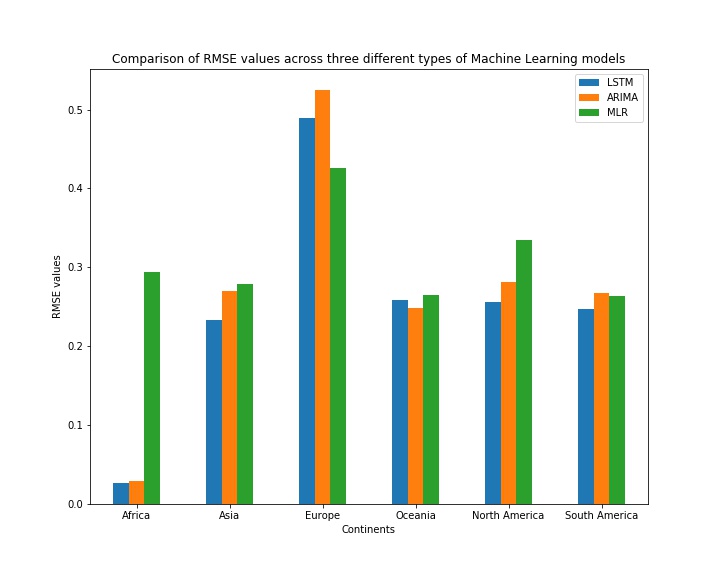


Figure 5.1. COMPARISON OF THE DIFFERENT MODELS’ RMSE VALUES

# Chapter 6

**Conclusion**

The following conclusion looks at the project as a whole and reflects upon the challenge and how the project had been followed through given the challenges and also including the various skills presented and developed over the course of the study period. As is any piece of scientific writing to conclude is to summarise and raise points which contain key aspects of what was found.

The achievements made in this project were that from taking accessible data from an open source and creating a full Data Science pipeline. This was to an extent that various transformations and analysis were performed by means of performing code using various libraries, some of which were originally unfamiliar however after research and practice alongside exploration and familiarity of the data set is was possible to perform the pipeline. Such challenges were investigated by means of effective documentation by the relevant packages. Other challenges such as the data handling and how the appropriate predictive task will be performed was very much a deep thought process. This is because it is key to gain relevant insights importantly and ask the right questions about the data. This led to the transformational processes to occur. Alongside the computational load it was a necessary thought process to embark on before such testing, manipulation and analysis could occur.

In order to carry out the particular machine learning processes there were pre-requisites that needed to be met. With the transformed data into continents analysis using hypothesis testing and visualisations aided with the clarity of how to best build the most appropriate model. The significance of running the pre-screening tests on the datasets ensured that the data was prepared so that the models that were formed satisfied the machine learning models and how they were to use the data.

The suggestions that some of the results inferred indicate there is potential for further development with the models and their uses with the different attributes made available due to the multivariate nature of the data. While some of the models indicated where certain techniques such as smoothing and rolling averages might not be necessary, it was also found how important certain features can be to the Data Science pipeline whereby not all fields are equal. This showed that while aspects such as the Africa transformation could have been approached differently such as using a broader range of data which when smoothed may have generated data which could have allowed more information gain. In comparison perhaps, another angle could be where Europe’s overall RMSE could also be improved by taking a rolling mean. Therefore, to develop the best models it is important to consider to minimise the amount of transformation by use of averages, but to also consider the variation and uncertainty of the given set too.

Another area for future investigation could be using the monthly components rather than the yearly components. This would allow a better understanding of how the changes yearly would occur. However, to process this would require a medium with a greater memory capacity. Following this further decomposition would be able to produce greater insights and assess the trend elements of the temperature of the different regions better.

Since the models were performed based on the continents, another area of development that could be investigated in the future is the difference between individual countries. Yet there may be difficulties in developing multiple machine learning models to 159 different countries. Not only would this be very time consuming the range of tuning that could occur would be vast.

Due to the dataset being of global land surface temperatures it would be also interesting to attempt to use the similar concept to combine land and sea/ocean temperatures over time so that analysis can occur of our whole global ecosystem and its changes over time.

Looking at the changes in temperature over the years across the different continents it is clear that for society there is an alarming increasing in global land temperatures across all continents. This is something that has been a growing issue and initiatives such as the Kyoto Protocol which looks at reducing the greenhouse gas effect which would then gradually cause a drop in the rate of increase of land temperatures. By using concepts like machine learning we are able to take and map where we see the future to be in regards to the health of our planet. This is vital as even changes by a few percent can trigger a multitude of catastrophic events. Therefore, the growing frequency of natural disaster occurrences could be partially attributed to the growing temperature trend. Such changes have been notably drastic in such cases in Europe which has seen a global increase of temperature by 20% since 1870 (Table 9, APPENDIX B). This could be due to lack of continentality because of many land-locked countries but also because of the boom in the economy which has seen more production and use of fossil fuels to drive the economies.

Machine learning has a place in predictive power it is key to use it to an advantage where we can look to the future where unless there are more drastic changes the trends will seemingly continue to rise. However, the term modelling is purely an approximation in itself. From what we know about global events we are only able to ascertain a certain amount of confidence in our predictability which is indicative of the planets weather system and changes that occur over time. Nevertheless, there is a place in society for modelling using machine learning. With improvements in science and data collection we will surely be able to gain more improved ways that we can handle such data to benefit and shape our future.

Finally, this project and its goals which were used to evaluate the different machine learning models that could be used to model the Earth’s global land temperatures and has allowed insights into the trends and the shape of how the Earth has changed over time. Facing challenges, as with any project is an occurrence such was the additive time taken to develop the project. However, the learnt aspect of how to manipulate and undertake such methods by critical thinking by implementation of the constant curiosity of what information can be gained from the data was key. It has been a significant platform to develop skills which uniquely contribute to personal evolution and frameworks understood in the constantly growing field of Data Science.

# References

Adewumi, A. O., Ariyo, A. A. and Ayo, C. K. (2014) *Stock Price Prediction Using the ARIMA Model*. 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. IEEE. PP. 106-112.

Aguera, A., de la Rosa, J. J. G., Melgar, J., Moreno, A., Palomares-Salas, J. C. and Ramiro, J. G.(2009) *ARIMA vs. Neural networks for wind speed forecasting*. IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, Hong Kong, 2009, IEEE. pp. 129-133.

Al-Fuqaha, A., Benhaddou, D., Gupta, A and Qolomany, B. (2017)"Role of Deep LSTM Neural Networks and Wi-Fi Networks in Support of Occupancy Prediction in Smart Buildings," 2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Bangkok, pp. 50-57.

Albright, T., Mutiibwa, D. and Strachan, S. (2015) 'Land Surface Temperature and Surface Air Temperature in Complex Terrain.' *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 8(10). IEEE. pp. 4762-4774.

Ancy, S., Asokan, R. Kumar, R. and Subhashini, R. (2014) *Prediction of onset of south west monsoon using multiple regression.* Proceedings of IEEE International Conference on Computer Communication and Systems ICCCS14. IEEE. pp. 170-175

Aparna, S. (2018) "Long Short Term Memory and Rolling Window Technique for Modeling Power Demand Prediction," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 1675-1678.

Arshanapalli, B.G., Fabozzi, F.J., Focardi, S.M., Rachev, S.T. (2014) *Model Selection Criterion: AIC and BIC. The Basics of Financial Econometrics:Tools, Concepts, and Asset Management Applications.* John Wiley & Sons, Inc. pp. 399-403.

Babu, C. N. and Reddy, B. E.(2012) *Predictive data mining on Average Global Temperature using variants of ARIMA models.* IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012) IEEE. PP256-260.

Boufares, F. and Rehab, M. A. (2015) *Scalable Massively Parallel Learning of Multiple Linear Regression Algorithm with MapReduce.* 2015 IEEE Trustcom/BigDataSE/ISPA. Vol. 2. IEEE. pp.41-47

Brito, J. and Mendes-Moreira, J. (2014 )*Instance ranking with multiple linear regression: Pointwise vs. listwise approaches.* 2014 9th Iberian Conference on Information Systems and Technologies (CISTI). IEEE. pp. 1-6.

Campbell, W., Shah, D. and Zulkernine, F.H. (2018) "A Comparative Study of LSTM and DNN for Stock Market Forecasting," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, pp. 4148-4155.

Ceccato, P., Temimi, M. and Vancutsem, C. (2010) "Monitoring air and Land Surface Temperatures from remotely sensed data for climate-human health applications," 2010 IEEE International Geoscience and Remote Sensing Symposium. IEEE. pp. 178-180.

Celeda, P., Jirsik, T., Trčka, Š. (2019)"Quality of Service Forecasting with LSTM Neural Network," 2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), Arlington, VA, USA, pp. 251-260.

Chen, F., Wan, Y. and Wu, T. (2018) "Graph Attention LSTM Network: A New Model for Traffic Flow Forecasting," 2018 5th International Conference on Information Science and Control Engineering (ICISCE), Zhengzhou, pp. 241-245.

Cherkassky, V. and Ma, Y. (2005) 'Multiple model regression estimation.' *IEEE Transactions on Neural Networks*, 16(4). IEEE. pp. 785-798

Chi, Y. & Xiao, L. and Zhang, J. (2018). Solar Power Generation Forecast Based on LSTM. (2018) IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, pp. 869-872.

Choi, H., Kim, H. and Ryu, S. (2018) "Short-Term Load Forecasting based on ResNet and LSTM," 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Aalborg, pp. 1-6.

Di, H., Junpeng, B., Yong, Q. and Yunpeng, L. (2017) "Multi-step Ahead Time Series Forecasting for Different Data Patterns Based on LSTM Recurrent Neural Network," 2017 14th Web Information Systems and Applications Conference (WISA), Liuzhou, pp. 305-310.

Dorville, J.-F. and Williams, D.(2015) *Predicting thermal impact of Light pipe on air temperature in room using OLS- Multiple Regression*. 2015 International Conference on Renewable Energy Research and Applications (ICRERA). IEEE. pp. 913-916.

European Centre for Medium-Range Weather Forecasts (2019) About Us. [Online] [Accessed on 19th August 2019] <https://www.ecmwf.int/en/about>

Faulina, R., Kuswanto, H., Lusia, D. A., Otok, B. W., Suhartono and Sutikno (2012) *Ensemble method based on ANFIS-ARIMA for rainfall prediction*. 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE), Langkawi. IEEE. pp. 1-4.

Feng, Y. and Wang, S. (2017) *A forecast for bicycle rental demand based on random forests and multiple linear regression*. 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). IEEE. pp. 101-105.

Fu, Y., Hu, W., Liu, B., Tang, M. and Yu, R.(2018). Multi-step Ahead Wind Power Forecasting Based on Recurrent Neural Networks. PES Asia-Pacific Power and Energy Engineering Conference (APPEEC). IEEE.

Gang, L. and Teng-Fei, W. (2018)"The Deep Neural Network Algorithm Based on Meteorological Features is in the Medium and Long Term," 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), Lanzhou, pp. 227-232.

Geetha, A. and Nasira, G. M. (2016) *Time series modeling and forecasting: Tropical cyclone prediction using ARIMA model.* 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE. pp. 3080 – 3086.

Gu, J., Jin, Z., Liu, C. and Qiu, C. (2017) *Short-term load forecasting using a long short-term memory network*. IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe). IEEE. pp. 1-6.

Gu, Y-H., Kim, K-J., Lee, K-H., Park, J-H., Son, U-H., Yoo, S-J. (2017) "PM10 density forecast model using long short term memory," 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN), Milan, pp. 576-581.

Gupta, L. and Priya, S.S. (2015) "Predicting the future in time series using auto regressive linear regression modeling," 2015 Twelfth International Conference on Wireless and Optical Communications Networks (WOCN), Bangalore, pp. 1-4.

Gupta. M (2018) Time Series Forecasting (Univariate vs Multivariate) [Online] [Accessed 22 August 2019] <https://medium.com/@nishantthewriter/time-series-forecasting-12a15dd0c64c>

# Gustafsson. A. and Wogenius. S. (2014) “Modelling Apartment Prices with the Multiple Linear Regression Model” KTH Royal Institute Of Technology SCI School of Engineering Sciences, Stockholm, pp 1-42.

He, H. Hu, R., Jiao, R., Zhang, Y. and Zhang, Y. (2018) A Power Forecasting Approach for PV Plant based on Irradiance Index and LSTM. Proceedings of the 37th Chinese Control Conference. IEEE.

Hibon, M and Makridakis, S. (2000). The M3-Competition: results, conclusions

and implications. International Journal of Forecasting, 16(4), pp 451–476.

Himika, Kaur, S. and Randhawa, S. (2018) *Global Land Temperature Prediction by Machine Learning Combo Approach.* 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE. pp. 1-8.

Hu, N., Li, G., Li, Y. and Yao, X. (2012) *Forecasting Mineral Commodity Prices with ARIMA-Markov Chain.* 2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics. Vol. 1. IEEE. PP. 49-52.

Hu, Y., Li, Y., Liu, L., Nie, X. and Sun, X. (2019) "An Enhanced LSTM for Trend Following of Time Series," in IEEE Access, vol. 7, pp. 34020-34030.

Huang, P., Jia, Z. and Yao, W. (2018) "Multidimensional LSTM Networks to Predict Wind Speed," 2018 37th Chinese Control Conference (CCC), Wuhan, 2018, pp. 7493-7497.

Huband, N.D.S. and Monteith, J.L. (1986) 'Radiative surface temperature and energy  balance of a wheat canopy.' *Boundary-Layer Meteorology*, 36(No. 1) pp. 1-17.

IBM. (2012) *Autocorrelation and Partial Autocorrelation Functions* [Online] [Accessed on 10th September 2019] <https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/timeseries_acf_pacf.htm>

InsightsBot (2018) *Augmented Dickey-Fuller Test in Python* [Online] [Accessed on 2nd September 2019] <http://www.insightsbot.com/blog/1MH61d/augmented-dickey-fuller-test-in-python>

Kapetanakis, S., Marsellos, A. and Tsakiri, K. (2018) 'Artificial Neural Network and Multiple Linear Regression for Flood Prediction in Mohawk River, New York.' *Water*, 10(9) p. 1158.

Keyvanpour, M. R. and Mehrmolaei, S. (2016) *Time series forecasting using improved ARIMA.* 2016 Artificial Intelligence and Robotics (IRANOPEN), Qazvin. IEEE pp. 92-97

Liu, R., Liu, Y., Wang, M. and Wang, Y. (2018) LSTM model optimization on stock price forecasting. 17th International Symposium on Distributed Computing and Applications for Business Engineering and Science. IEEE. pp173 – 177.

Liu, X and Zeng, Y. (2018) "A-Stock Price Fluctuation Forecast Model Based on LSTM," 2018 14th International Conference on Semantics, Knowledge and Grids (SKG), Guangzhou, China, pp. 261-264.

Liu, Z. Wang, T. and Xiao, Z. (2017) 'Performance Evaluation of Machine Learning Methods for Leaf Area Index Retrieval from Time-Series MODIS Reflectance Data.' *Sensors (Basel, Switzerland)*, 17(1) pp. 81.

Lo, J.-H. (2011) *A study of applying ARIMA and SVM model to software reliability prediction.* 2011 International Conference on Uncertainty Reasoning and Knowledge Engineering, Bali. IEEE. pp. 141-144.

Lou, Y. and Wang, Z.(2019) *Hydrological time series forecast model based on wavelet de-noising and ARIMA-LSTM.* IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China. IEEE. pp. 1697-1701

Medar, R., Rajpurohit, V. S. and Rashmi, B.(2017) *Impact of Training and Testing Data Splits on Accuracy of Time Series Forecasting in Machine Learning.* 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA). IEEE. pp. 1-6.

Met Office (2018) The Cray CX40 supercomputer. [Online] [Accessed on 19th August 2019] <https://www.metoffice.gov.uk/about-us/what/technology/supercomputer>

Paz, B, Sosa, C. and Sosa, J.. (2000) *ARIMA models in the rain attenuation prediction in a Mexican tropical area.* IEEE Antennas and Propagation Society International Symposium. Transmitting Waves of Progress to the Next Millennium. 2000 Digest. Held in conjunction with: USNC/URSI National Radio Science Meeting. Vol. 2. IEEE. PP. 546-549.

Pintelon, R. and Schoukens, J. (2006) Box–Jenkins identification revisited—Part I: Theory, Automatica, Volume 42, Issue 1. ScienceDirect. pp. 63-75

Raj, R. J. R., Rizwan, M. O. D. and Vasudev, M. (2017) *A novel approach for time series data forecasting based on ARIMA model for marine fishes*. 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), Chennai. IEEE. pp. 1-4

Swalin, A. - Medium (2018) Choosing the Right Metric for Evaluating Machine Learning - Part 2. [Online] [Accessed on 19th August 2019] <https://medium.com/usf-msds/choosing-the-right-metric-for-evaluating-machine-learning-models-part-2-86d5649a5428>

Statsoft (2019) *How to Identify Patterns in Time Series Data: Time Series Analysis.* [Online] [Accessed 21st September 2019]

<http://www.statsoft.com/textbook/time-series-analysis#analysis>

Vaishnav, V. and Vajpai, J. (2018) *Seasonal Time Series Forecasting by Group Method of Data Handling.* 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). IEEE. pp. 1-5.

Wang, Q., Wang, X. and Xia, F. (2009) *Integration of Grey Model and Multiple Regression Model to Predict Energy Consumption* 2009 International Conference on Energy and Environment Technology*.* Vol. 1 2009. IEEE. pp. 194-197.

Zhang, H. (2017) *Estimation of daily average near-surface air temperature using MODIS and AIRS data.* 2017 2nd International Conference on Frontiers of Sensors Technologies (ICFST).. IEEE. pp. 377-381.

Zhang, J., Li, Y. and Cao, J. (2011) *Sensor situation based on the multiple linear regression forecast.* 2011 IEEE International Conference on Computer Science and Automation Engineering. Vol. 2. IEEE. pp. 47-50

# Bibliography

Brownlee, J. (2019) *How to Backtest Machine Learning Models for Time Series Forecasting. Machine Learning Mastery.* [Online] [Accessed on 1st September 2019]

<https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/>

Brownlee, J. (2019) *How to Develop LSTM Models for Time Series Forecasting. Machine Learning Mastery.* [Online] [Accessed on 10th September 2019]

<https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/>

Brownlee, J. (2019) *How to Diagnose Overfitting and Underfitting of LSTM Models. Machine Learning Mastery.* [Online] [Accessed on 15th September 2019] <https://machinelearningmastery.com/diagnose-overfitting-underfitting-lstm-models/>

Brownlee, J. (2019) *How to Tune ARIMA Parameters in Python. Machine Learning Mastery.* [Online] [Accessed 8th September 2019] <https://machinelearningmastery.com/tune-arima-parameters-python/>

Brownlee, J. (2019) *How to Tune LSTM Hyperparameters with Keras for Time Series Forecasting. Machine Learning Mastery.* [Online] [Accessed 15th September 2019]

<https://machinelearningmastery.com/tune-lstm-hyperparameters-keras-time-series-forecasting/>

Brownlee, J. (2019) *Multivariate Time Series Forecasting with LSTMs in Keras. Machine Learning Mastery.* [Online] [Accessed on 14th September 2019]

<https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/>

Colah (2015) *Understanding LSTM Networks. GithubPages.* [Online] [Accessed 9th September 2019] <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Chatterjee, S. (2019) *lstm time series. Towards Data Science. Medium.* [Online] [Accessed on 7th September 2019] <https://medium.com/search?q=lstm%20time%20series>

Jain, A. (2016) *A comphrensive beginner’s guide to create Time Series Forecast (with Codes in Python and R). Analytics Vidhya.* [Online] [Accessed 1st September 2019]

<https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/>

Keras Documentation (2019) *Getting started with the Keras Sequential Model.* [Online] [Accessed on 14th September 2019] <https://keras.io/getting-started/sequential-model-guide/>

Khandelwal, R. (2019) *Multivariate Time Series using RNN with Keras. Medium.* [Online] [Accessed on 10th September 2019]

<https://medium.com/datadriveninvestor/multivariate-time-series-using-rnn-with-keras-7f78f4488679>

Kompella, R. (2018) *Using LSTMs to forecast time-series. Towards Data Science. Medium.* [Online] [Accessed on 9th September 2019]

<https://towardsdatascience.com/using-lstms-to-forecast-time-series-4ab688386b1f>

Li, S. (2018) *An End-to-End Project on Time Series Analysis and Forecasting with Python. Towards Data Science. Medium.* [Online] [Accessed 29th August 2019]

<https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b>

Li, S. (2019) *Time Series Analysis, Visualization & Forecasting with LSTM. Towards Data Science. Medium.* [Online] [Accessed 8th September 2019]

<https://towardsdatascience.com/time-series-analysis-visualization-forecasting-with-lstm-77a905180eba>

Prabhakaran, S. (2019) *ARIMA Model – Complete Guide to Time Series Forecasting in Python. Machine Learning Plus.* [Online] [Accessed on 6th September 2019]

<https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/>

Regression Analysis in NCSS (2019) *Multiple Regression. NCSS.* [Online] [Accessed on 2nd September 2019] <http://ncss.wpengine.netdna-cdn.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Multiple_Regression.pdf>

Sangarshanan (2018) *Time series Forecasting – ARIMA models. Towards Data Science.* [Online] [Accessed on 8th September 2019] <https://towardsdatascience.com/time-series-forecasting-arima-models-7f221e9eee06>

Sergeev, D. (2018) *Open Machine Learning Course. Topic 9. Part 1. Time series analysis in Python. Medium.* [Online] [Accessed 30th August 2019]

<https://medium.com/open-machine-learning-course/open-machine-learning-course-topic-9-time-series-analysis-in-python-a270cb05e0b3>

Sigma Coding (2019) *Multiple Regression Analysis in Python | Part 5. Youtube.* [Online] [Accessed 1st September 2019] <https://www.youtube.com/watch?v=j1sjqEStECQ&t=133s>

Singh, A, (2018) *A Multivariate Time Series Guide to Forecasting and Modeling (with Python codes). Analytics Vidhya.* [Online] [Accessed on 2nd September 2019]<https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide-forecasting-modeling-python-codes/>

Srivastava, P. (2017) *Essentials of Deep Learning: Introduction to Long Short Term Memory. Analytics Vidhya.* [Online] [Accessed on 7th September 2019]

<https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/>

XLSTAT (2011) *Ordinary Least Squares regression or Linear regression. Youtube.*  [Online] [Accessed 3rd September 2019] <https://www.youtube.com/watch?v=VZkXjh0m1Rg>

# Appendix A

**Terms of Reference**

**Terms Of Reference**

**Time series analysis and development of machine learning models for forecasting multivariate data**

**Nikhil Jagatia 18055146**

**MSc Data Science**

**Learning outcomes**

1. The composition of an effective and meaningful Data Science report
2. Judgement and sanctioning of a suitable time series dataset to be used
3. Effective use of research and appropriate literature
4. Effective manipulation of time series data
5. Development of machine learning models
6. Critical analysis of outcomes and the impact of experimentation

**Project Objectives/aims**

To be able to carefully prepare and carry out extensive machine learning tasks using Data Science techniques by applying thorough experimentation and critical analysis of a chosen dataset which is under investigation. Furthermore to be able to obtain and use a combination of domain knowledge, statistical and mathematical processes along with extensive programming skills in delivering and executing a full Data Science project. Furthermore the suitability of the dataset should be assessed and selected that is appropriate and allows meaningful findings to be discovered. Such examples of some research questions will be:

* What is the importance of Time Series data?
* How suitable is the dataset in question?
* What is the requirement of forecasting?
* Which machine learning model perform better?
* Which machine learning model has statistically the most significant result?
* Which machine learning model has the best accuracy in prediction, with metrics such as Mean Squared Error?
* Which methods of experimentation worked?
* How realistic are the models produced with the inclusion of real life future scenarios that will occur?
* Where is future work and investigation required?

To also use researching skills to apply relevant literature to the chosen problem while also demonstrating novel and innovating ideas to be used within the project

The assessment of potential risks and ethical issues are to be screened to enable relevancy and applicability of such findings and analysis to society. In addition the Data Science report produced should be logical, fluid and presented in an appropriate format which is consistent and objectively evaluated using evidence.

**Project Overview and Description**

The focus of this project is to use a combination of data science and machine learning techniques in building models which can be used to forecast and predict the values that would be able to learn and show insights. The process will involve using a full data science pipeline where there will be extensive pre-processing of data where data will be cleaned so that is prepared for the algorithms to be used. Exploratory data analysis will be performed on the data so that insights of the fundamental characteristics of the data are discussed, it will be achieved using visualization and statistical libraries. This will aid in the feature engineering that will be performed various times whereby the models will be assessed on their fitness. Special events will be investigated within the time series. This is so that findings and results for certain periods are understood.

The dataset that is going to be used is still under research. This is due to finding an appropriate dataset that provides the opportunity to perform, investigate and generate developmental findings using the data science method. The characteristics that the data should have are that it is of multivariate time series where the data range is significant enough to be able to detect changes and sufficient so that future forecasting can be of meaningful value.

The programming langue that will be used is Python with use of Jupyter Notebook with libraries such as Sci-kit learn, Pandas, Numpy, Matplotlib (for visualization), Seaborn, Pylab, Datetime, Statsmodels and Scipy. These modules will enable manipulation of the dataset so that there can be configurations, analysis and models built.

Working with timeseries data raises questions with the integrity and the characteristics that it presents. By decomposing the data the trend in data and its seasonality is visualized. Rolling and moving windows will be highlighted to see the pattern based differences. Furthermore moving averages of the data will be taken so that the normal behavior of the data is observed. With time series data and using domain knowledge you can infer that there is relationship between the data and natural occurrences. The importance of time series will also be discussed and its need in society along with its applications, while also evaluating the significance of forecasting and its usefulness. Moreover the specific dataset chosen will undergo thorough self-scrutiny and evaluation of the legal and ethical significance of not only of its use but what the findings could infer and represent in society.

**Models and Evaluation Plan**

For the model development the process will involve processing the data through autocorrelation and association functions. This will be to highlight linearity, stationarity and relationships. Once this has been completed there will be model development and once the baseline runs are performed there will be significant hyperparameter tuning so that the best fitting model is found. Through regression models and Deep Learning.

Such models that will be used are the ARIMA model which is a autoregressive method that uses moving averages and lags where metrics can be investigated so that the clarity of fitness of each model can be compared. Some of the metrics that will be involved are: Accuracy, Correlation, Akaike Information Criteria (AIC), Bayesian Information Criterion (BIC) and Mean Squared Error.

Deep learning methods can be used with time series, in such cases RNN and LSTMs and their fitting to the problem and dataset can be assessed so that there is variation in different types of models.

**Challenges**

While working on such a project there will be challenges that will be faced and addressed. Firstly the question of the dataset is the biggest and most important issue that requires to be resolved. This is so that a worthwhile and significant research is performed.

Another challenge that will be faced is the familiarization and review of the different tpyes of machine learning algorithms. By using metrics and measurable statistics such as mean squared error there are ways to limit and control this but it is a feature that will require investigation.

The time series data needs to of good quality and representative. The importance of such quality is so that when exploratory analysis and eventual modelling is performed the most accurate representation of the time with respect to the events is required. In addition the relevancy of findings from others can provide a foundation of evidence however there can be contradictory findings. However will sufficient proof this will be assessed as appropriate.

Due to time series data being readily available there will need to be checks in place so that there are no breaches in confidentiality where data that is not to be used is used by mistake. Typically datasets and competitions are publicly available therefore finding the appropriate dataset will consequently mean a well angled approach to the questions asked of the data. Similarly this is true of the problem description which will be dependent on the data.

**Product to be delivered**

A full Data Science pipeline report where analysis and modelling is performed on time series data.

A clear framework which is logically organized into systematic steps that shows thorough flow and thought into the data

Investigation of data and analysis by composition of meaningful data science questions

Showing a clear understanding of the machine learning methodologies and its relevancy and use of working data science knowledge to show extensive parameter tuning.

Assessment of challenges, problems and delays whereby the identification of underlying issues that affect final outcomes.

Accurately recorded performance of algorithms and their suitability.

Clear communication of findings using appropriate terminology and areas for further research.

The report will be delivered in the appropriate format and fit to a scale which is within the constraints of 20,000 words.

**Resources required**

A suitable dataset which meets the requirements of being of multivariate time-series will be obtained.

The programming language of Python and its libraries used as a toolkit for development of machine learning models, analysis and processing of the data.

The use of databases to collect information such as IEEE and ACM Digital Library and the MMU Library with the appropriate documentation for informative and literature information.

The deep learning aimed to be investigated, if implemented would require a GPU machine or a facilitator which holds sufficient memory loads such as Poincare. Solutions such as Google Colab may be suitable.

**Reporting to**

Project Supervisor: Mr Luciano Gerber

Reporting Mechanism: Communication will be achieved by use of frequent meetings to discuss ideas and to resolve any issues around the project and support if required. The use of Skype will also be used as a mode of communication to ensure regular contact for feedback and the status of the project with communication of further topics to be discussed.

Documents will be reviewed and conveyed over email to show the status of the project and any relating information or resources.

**Activity Schedule**



References

[1] Hyndman, R., Athanasopoulos, G. (2018) Forecasting: Principles and Practice [Online] [Accessed in May 2019] <https://otexts.com/fpp2/case-studies.html>

[2] Rohrer, B. (2018) How autocorrelation works. Youtube. [Online] [Accessed in May 2019] <https://www.youtube.com/watch?v=ZjaBn93YPWo>

[3] Brownlee, J. (2017) Time Series Forecasting with Python 7-Day Mini-Course [Online] [Accessed in May 2019] <https://machinelearningmastery.com/time-series-forecasting-python-mini-course/>

[4] Mueller, A. (2019) Applied Machine Learning 2019 – Lecture 23 – Basics of Time Series. Youtube. [Online] [Accessed in May 2019] <https://www.youtube.com/watch?v=d34aLzKP0rY&list=PL_pVmAaAnxIQGzQS2oI3OWEPT-dpmwTfA&index=23>

[5] McKinney, W. (2017) *Python for Data Analysis*. 2nd ed., O’Reilly Media Inc, Sebastopol., Chapter 11.

[6] Brownlee, J. (2018) How to Develop LSTM Models for Time Series Forecasting [Online] [Accessed in May 2019] <https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/>

[7] Asthana, A., (2018) Bitcoin Price Prediction Using Time Series Forecasting [Online] [Accessed in May 2019] <https://towardsdatascience.com/bitcoin-price-prediction-using-time-series-forecasting-9f468f7174d3>

[8] Li, Y., Ying, F., (2011) *Multivariate Time Series Analysis in Corporate Decision-making Application.* Business School, HoHai University . 2011 International Conference of Information Technology, Computer Engineering and Management Sciences. IEEE Computer Society. Pp 374 -376.

[9] Chen, S., *et al* (2015) *A Similarity Model Based On Trend For Time Series.* College of Computer and Information Hohai University. 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science. Conference Publishing Services. Pp 435-438.

[10] Li, M., Quan, D., (2017) *A Time Series Approach for Risk Forecasting of Individual Stocks in the New Three Board Market.* China Center for Industrial Security Research Beijing Jiaotong University. Special Funding Project of Ministry of Education and the Fundamental Research Funds for the Central Universities. [2017 4th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS)](https://ieeexplore-ieee-org.ezproxy.mmu.ac.uk/xpl/mostRecentIssue.jsp?punumber=8053939). Kyoto, Japan. IEEE. **Electronic ISBN:** 978-1-5386-0995-8

[11] Shahabi, C., Yang, K., Yoon, H., (2005) *Feature Subset Selection and Feature Ranking for Multivariate Time Series.* Department of Computer Science, University of Southern California. IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 9. IEEE Computer Society. Pp 1186-1198.

[12] Haratizadeh, S., Hoseinzade, E., (2019) *CNNpred: CNN-based stock market prediction using a diverse set of variables.* Faculty of New Sciences and Technologies, University of Tehran. Expert Systems With Applications 129 (2019). Lin, B., ScienceDirect. ELSEVIER, pp 273 – 295.

**Signature:**

Nikhil Jagatia

7Z10SS Masters Project NPC ToR Coversheet

Department of Computing and Mathematics Computing and Digital Technology Postgraduate Programmes Terms of Reference Coversheet

|  |  |
| --- | --- |
| Student name: | Nikhil Jagatia |
|  |  |
| University I.D.: | 18055146 |
|  |  |
| Academic supervisor: | Luciano Gerber |
|  |  |
| External collaborator (optional): |  |
|  |  |
| Project title: | Time series analysis and development of machine learning models for forecasting multivariate data |
|  |  |
| Degree title: | MSc Data Science |
|  |  |
| Project unit code: | 6G7Z1015\_1819\_9Z6 |
|  |  |
| Credit rating: | 60 |
|  |  |
| Start date: | 14/06/19 |
|  |  |
| ToR date: | 14/06/19 |
|  |  |
| Intended submission date: | 09/09/19 |
|  |  |
|  |  |
| Signature and date student: | Nikhil Jagatia 13/06/19 |
|  |
|  |  |
| Signature and date external |  |
| collaborator (if involved): |  |
|  |  |

This sheet should be attached to the front of the completed ToR and uploaded with it to Moodle.

**START HERE - Basic Information**

This form must be completed for all student projects.

**Before you proceed**

Some activities inherently involve increased risks or approval by external regulatory bodies, so a proportional ethics review is not recommended and a full ethical review may be required.

These may include:

1. Approval from an external regulatory body (including, but not limited to: NHS (HRA), HMPPS etc.);
2. Misleading participants;
3. Research without the participants' consent;
4. Clinical procedures with participants;
5. The ingestion or administration of any substance to participants by any means of delivery;
6. The use of novel techniques, even where apparently non-invasive, whose safety may be open to question;
7. The use of ionising radiation or exposure to radioactive materials;
8. Engaging in, witnessing, or monitoring criminal activity;
9. Engaging with, or accessing terrorism related materials;
10. A requirement for security clearance to access participants, data or materials;
11. Physical or psychological risk to the participants or researcher;
12. The project activity takes place in a country outside of the UK for which there is currently an active travel warning issued by the authorities (see info button);
13. Animals, animal tissue, new or existing human tissue, or biological toxins and agents.

**If any of these activities are fundamental to your project, please contact your supervisor to determine if a full application is required.**

This form must be completed for each research project which you undertake at the University. It must be approved by your supervisor (where relevant) PRIOR to the start of any data collection.

In completing this form, please consult the University's [ACADEMIC ETHICAL FRAMEWORK](http://www.mmu.ac.uk/policy/pdf/policy_ref_Academic_Ethical_Framework.pdf) for ethical research.

A1 Please confirm that you will abide by the University's Academic Ethical Framework in relation to this project.

 Yes

 No

A2 Are you submitting this application as a learning experience, for a unit which already has ethical approval? (please confirm with your supervisor)

 Yes

 No

A3 Student details

Title First Name Surname

Nikhil Jagatia

Email nikhil.jagatia@stu.mmu.ac.uk

A3.1 Manchester Metropolitan University ID number

18055146

A4 Supervisor

Title First Name Surname

Mr Luciano Gerber

Faculty Science and Engineering

Telephone +44 (0)161 247 1694

Email l.gerber@mmu.ac.uk

A5 Which Faculty is responsible for the project?

Science and Engineering

A6 Course title

MSc Data Science

A7 Project title

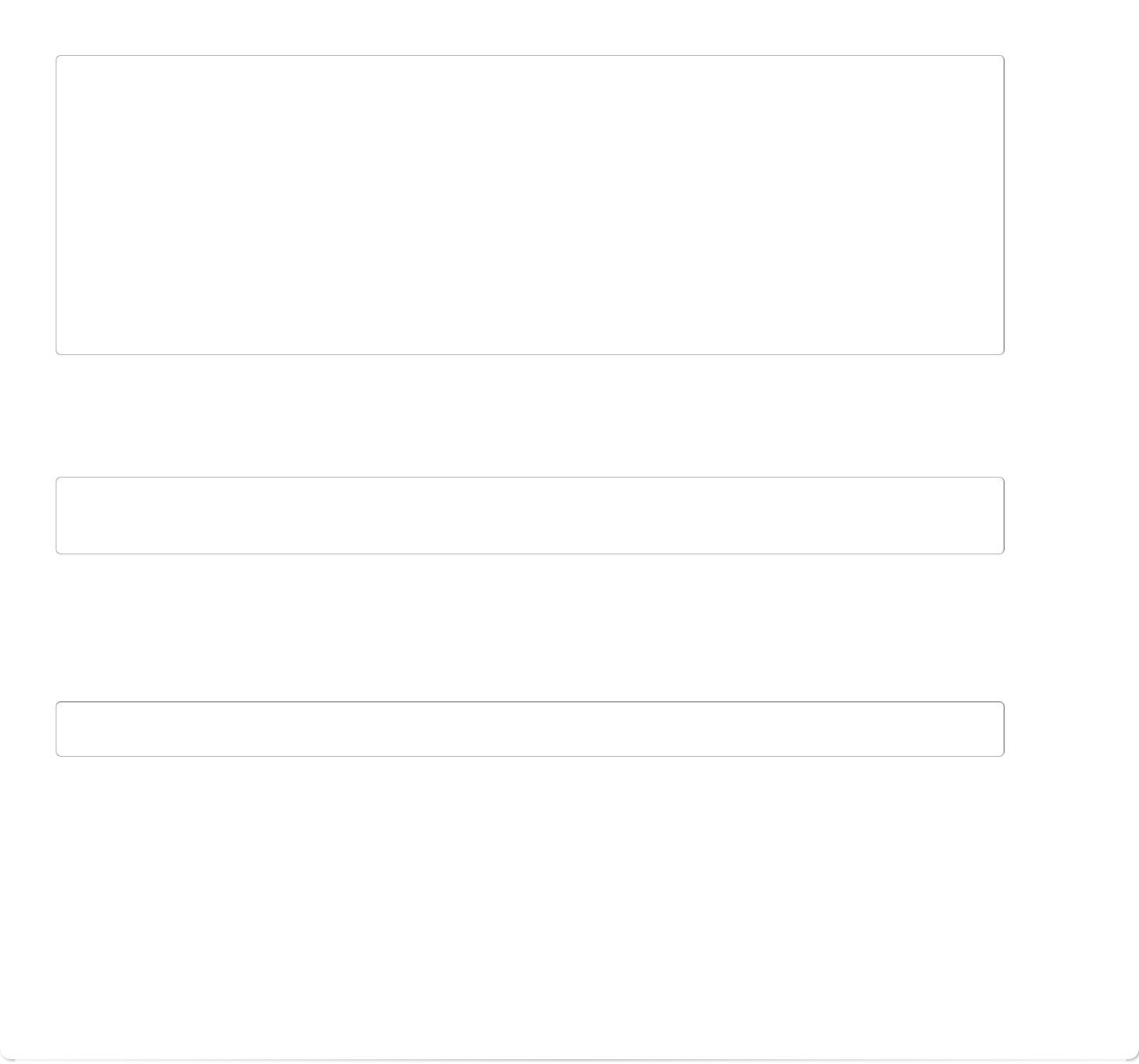
Time series analysis and development of machine learning models for forecasting multivariate data

A8 What is the proposed start date of your project?

 14/06/2019

A9 When do you expect to complete your project?

 29/09/2019

A10 Please describe the overall aims of your project (3-4 sentences). Research questions should also be included here.

The focus of this project is to use a combination of data science and machine learning techniques in building models which can be used to forecast and predict the values that would be able to learn and show insights. The process will involve using a full data science pipeline where there will be extensive pre-processing of data where data will be cleaned so that is prepared for the algorithms to be used.

* What is the importance of Time Series data?
* How suitable is the dataset in question?
* What is the requirement of forecasting?
* Which machine learning model perform better?
* Which machine learning model has statistically the most significant result?
* Which machine learning model has the best accuracy in prediction, with metrics such as Mean Squared Error?
* Which methods of experimentation worked?
* How realistic are the models produced with the inclusion of real life future scenarios that will occur?
* Where is future work and investigation required?

A11 Please describe the research activity

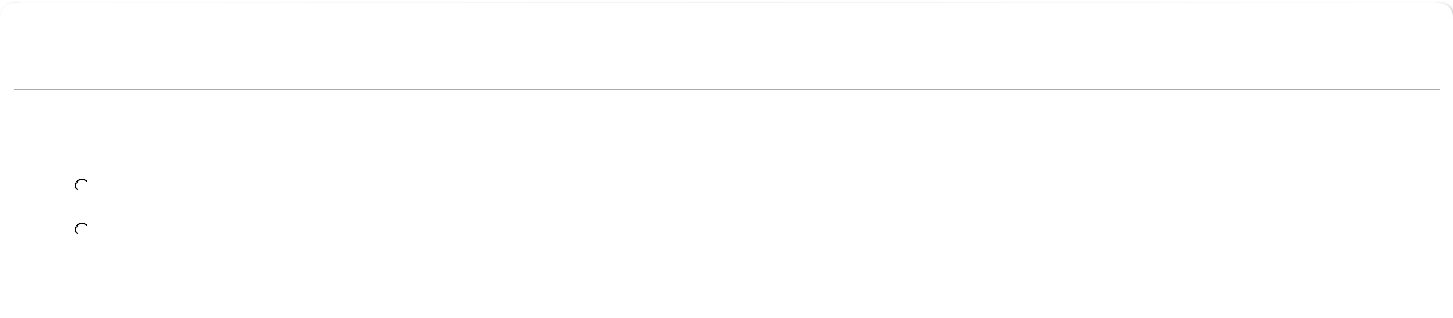
Research of an appropriate time series dataset for the use of forecasting. In addition to perform a full Data Science pipeline project on the data whereby machine learning models and algorithms will be compared and tested to see which performs better for the given task.

A12 Please provide details of the participants you intend to involve (please include information relating to the number involved and their demographics; the inclusion and exclusion criteria)

No animals or participants will be used for experimental purposes on this project.

A13 Please upload your project proposal

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type** | **Document Name** | **File Name** | **Version Date** | **Version** | **Size** |
|  |  |  |  |  |  |
| Project Proposal | nikhil-student-proposal (4) | nikhil-student-proposal (4).pdf | 22/02/2019 | 1 | 355.4 KB |

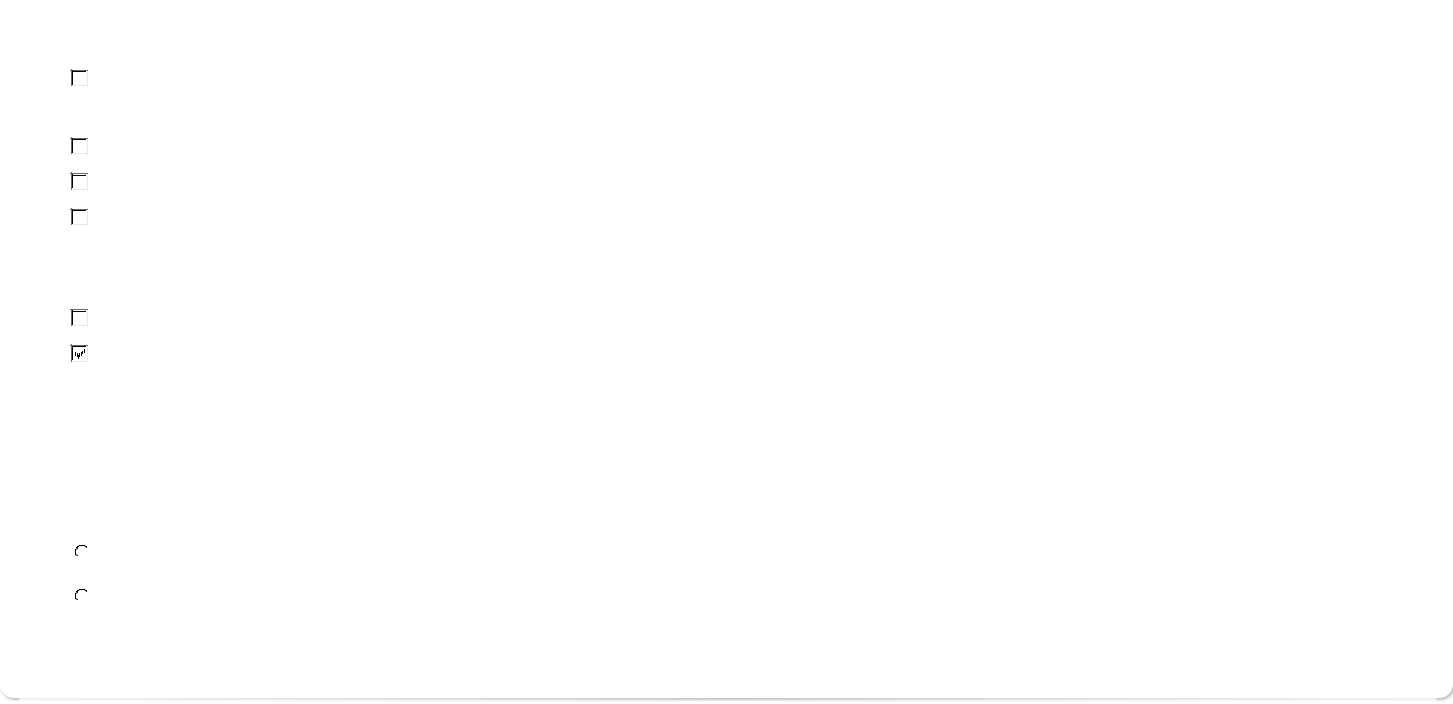


**Project Activity**

B1 Are there any Health and Safety risks to the researcher and/or participants?

 Yes

 No

B2 Please select any of the following which apply to your project

Aspects involving human participants (including, but not limited to interviews, questionnaires, images, artefacts and social media data)

Aspects that the researcher or participants could find embarrassing or emotionally upsetting Aspects that include culturally sensitive issues (e.g. age, gender, ethnicity etc.)

Aspects involving vulnerable groups (e.g. prisoners, pregnant women, children, elderly or disabled people, people experiencing mental health problems, victims of crime etc.), but does not require special approval from external bodies (NHS, security clearance, etc.)

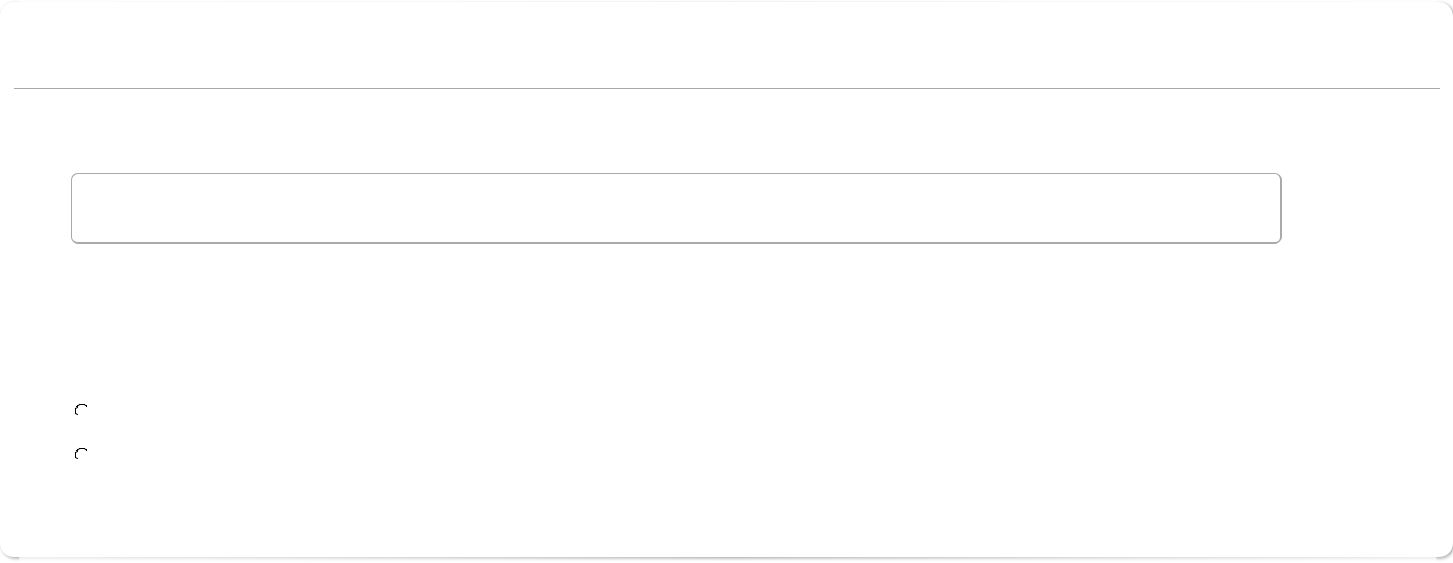
Project activity which will take place in a country outside of the UK

None of the above

B2.4 Is this project being undertaken as part of a larger research study for which a Manchester Metropolitan application for ethical approval has already been granted or submitted?

 Yes

 No



**Data**

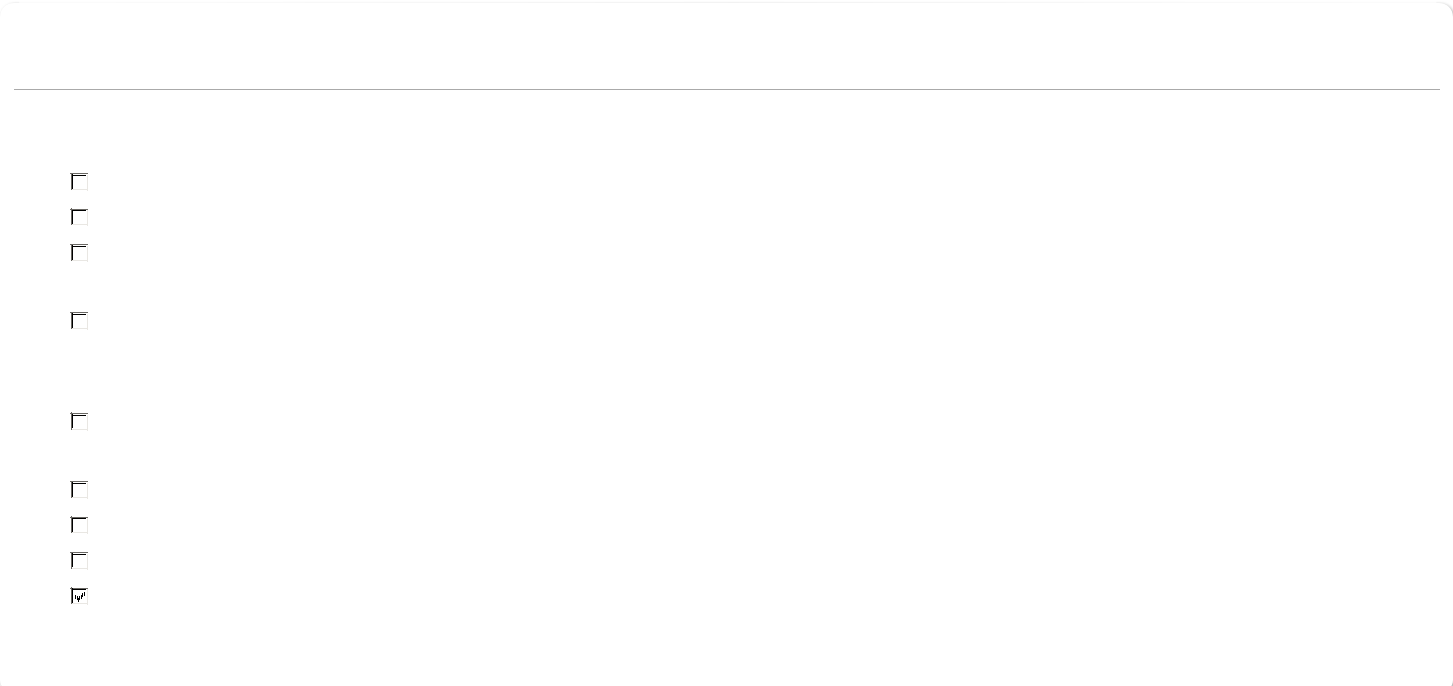
F1 How and where will data and documentation be stored?

Password protected machine, within appropriate folder arrangement

F2 Will you be collecting personal data or sensitive personal data as part of this project?

 Yes

 No



**Insurance**

F3 Does your project involve:

Pregnant persons as participants with procedures other than blood samples being taken from them? (see info button)

Children aged five or under with procedures other than blood samples being taken from them? (see info button)

Activities being undertaken by the lead investigator or any other member of the study team in a country outside of the UK as indicated in the info button? If ‘Yes’, please refer to the ‘Travel Insurance’ guidance on the info button

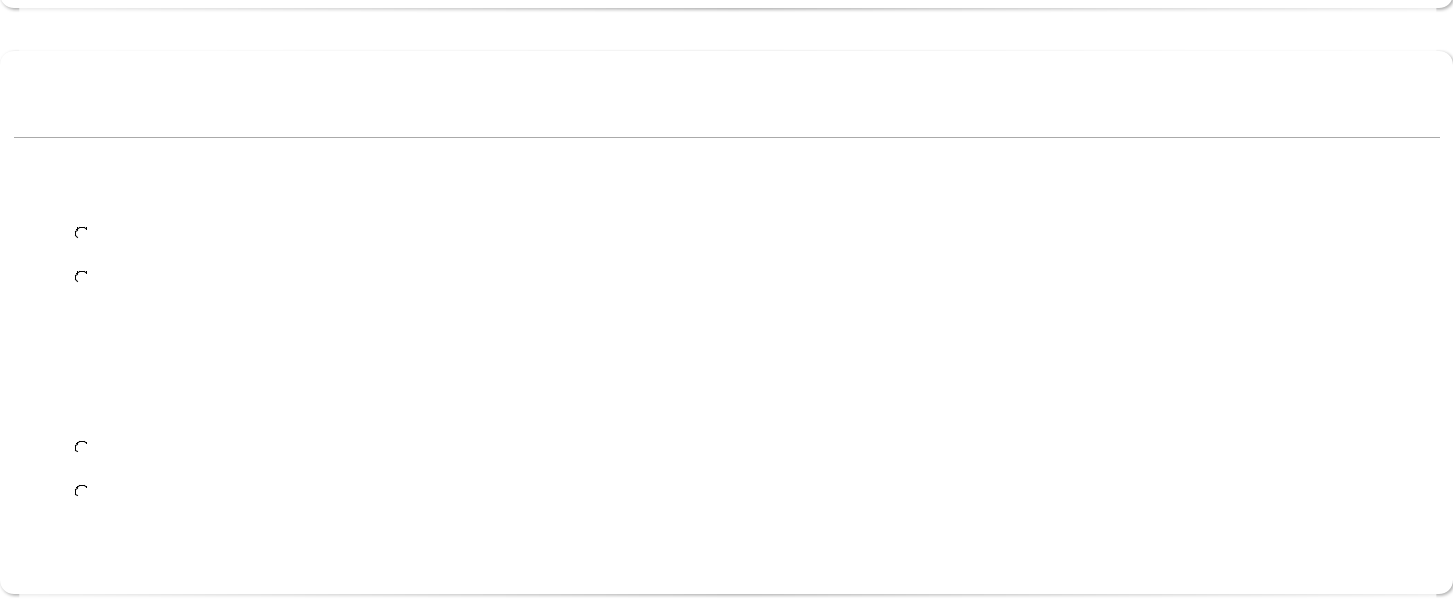
Working with Hepatitis, Human T-Cell Lymphotropic Virus Type iii (HTLV iii), or Lymphadenopathy Associated Virus (LAV) or the mutants, derivatives or variations thereof or Acquired Immune Deficiency Syndrome (AIDS) or any syndrome or condition of a similar kind?

Working with Transmissible Spongiform Encephalopathy (TSE), Creutzfeldt-Jakob Disease (CJD), variant Creutzfeldt-Jakob Disease (vCJD) or new variant Creutzfeldt-Jakob Disease (nvCJD)?

Working in hazardous areas or high risk countries? (see info button) Working with hazardous substances outside of a controlled environment?

Working with persons with a history of violence, substance abuse or a criminal record?

None of the above

**Additional Information**

G1 Do you have any additional information or comments which have not been covered in this form?

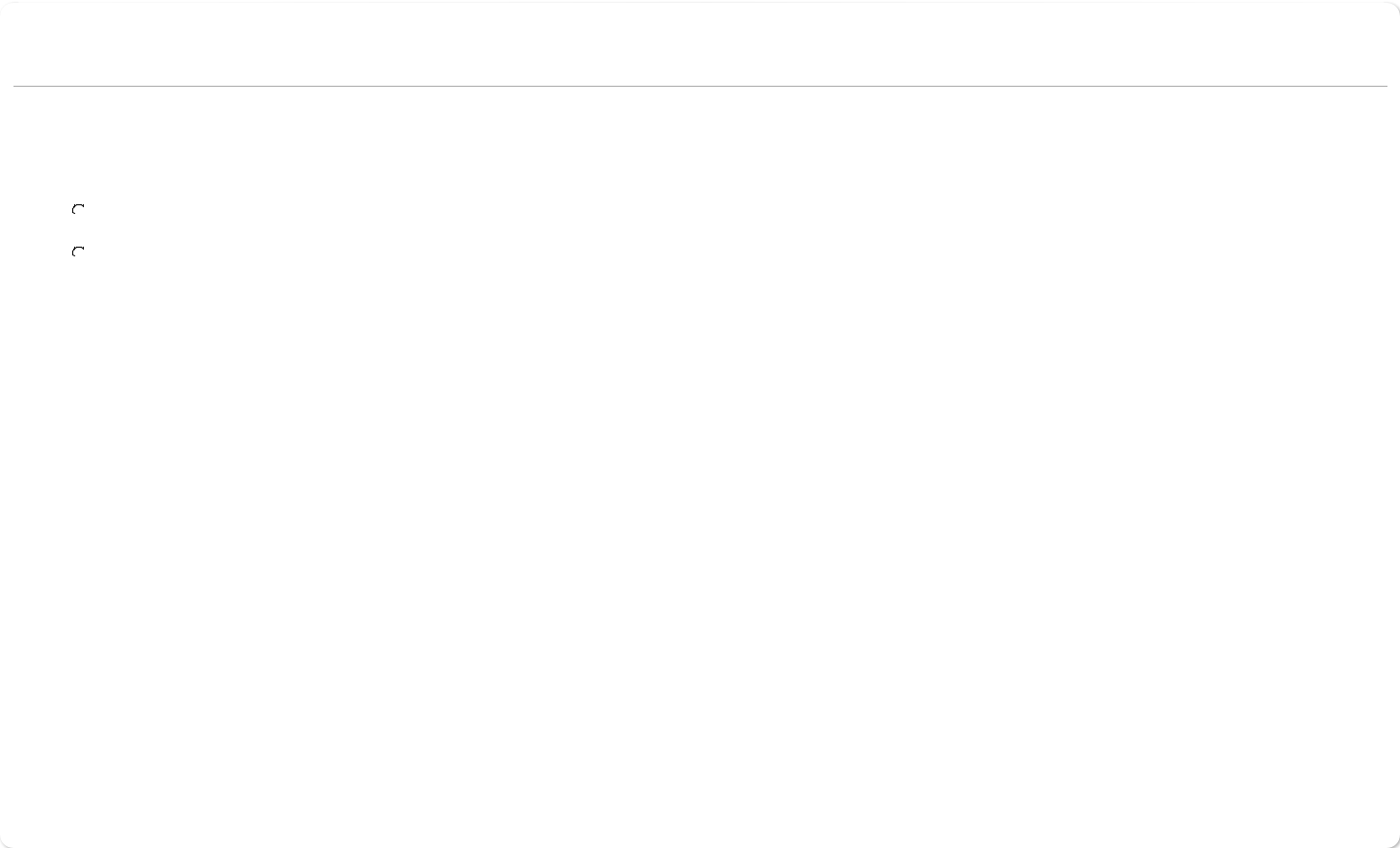
 Yes

 No

G2 Do you have any additional documentation which you want to upload?

 Yes

 No



**Signatures**

H1 I confirm that all information in this application is accurate and true. I will not start this project until I have received Ethical Approval.

 I confirm

 I do not confirm

H2 Please notify your supervisor that this application is complete and ready to be submitted by clicking "Request" below. Do not begin your project until you have received confirmation from your supervisor - it is your responsibility to ensure that they do this.

**Signature Request:** Signature requested from Luciano Gerber on 13/06/2019 14:08

H3 By signing this application you are confirming that all details included in the form have been completed accurately and truthfully.

# Appendix B

**List of Tables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AverageTemperature | AverageTemperatureUncertainty | Latitude | Longitude |
| AverageTemperature | / | -0.06480681362388918 | -0.6686662613391109 | 0.07174540425286098 |
| AverageTemperatureUncertainty | -0.06480681362388918 | / | -0.08995596884590952 | -0.0052325441137999465 |
| Latitude | -0.6686662613391109 | -0.08995596884590952 | / | / |
| Longitude | 0.07174540425286098 | -0.0052325441137999465 | / | / |

TABLE 1 SHOWING THE CORRELATIONS OF VALUES

TABLE 2. VIF VALUES FOR MULTICOLINEARITY CHECK.

|  |  |
| --- | --- |
|  | Variance Inflation Factor (VIF) values |
| AverageTemperature | 1.869569 |
| AverageTemperatureUncertainty | 1.038763 |
| Longitude | 1.035822 |
| Latitude | 1.925990 |

TABLE 3. MLR EQUATION VALUES (3 D.P.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AverageTemperatureUncertainty Coefficeint | Latitude Coefficient | Longitude Coefficient | Intercept |
| Africa | -1.085 | -0.231 | -0.176 | 30.365 |
| Asia | -1.664 | 6.237 | -1.626 | 15.054 |
| Europe | -1.366e+00 | -1.0158e+12 | 3.143e+10 | 49673226456069.92 |
| Oceania | -1.426 | 0.273 | -0.120 | 25.884 |
| North America | -0.834 | 236.843 | -127.639 | 4798.820 |
| South America | -1.536 | 0.253 | -0.238 | 30.996 |

|  |  |
| --- | --- |
|  | P Values |
| Africa | 98.7668722685223 |
| Asia | 99.58342469516704 |
| Europe | 2.2339780513294827 |
| Oceania | 5.123926511000654 |
| North America | 65.48571792483145 |
| South America | 31.43253861578322 |

TABLE 4. P VALUES FOR THE DATA WITHIN THE DIFFERENT CONTINENTS.

TABLE 5. MEAN OF RESIDUALS FOR THE MODEL OF RESPECTIVE COUNTRIES

|  |  |
| --- | --- |
|  | Mean of residuals |
| Africa | 5.445027145217157e-14 |
| Asia | -2.479004655874127e-13 |
| Europe | -1.3175263349454428e-12 |
| Oceania | 2.9702166652138356e-13 |
| North America | 2.7112065678933404e-11 |
| South America | 7.944262531762231e-14 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R2 | AIC | BIC | RMSE |
| Africa | 0.84 | 92.47 | 104.4 | 0.29345387 |
| Asia | 0.475 | 81.95 | 93.83 | 0.279189233 |
| Europe | 0.241 | 247.6 | 256.5 | 0.425709444 |
| Oceania | 0.575 | 45.2 | 57.08 | 0.26441528 |
| North America | 0.41 | 105.8 | 117.7 | 0.334876623 |
| South America | 0.574 | 74.09 | 85.97 | 0.26412479 |

TABLE 6. SUMMARISED METRICS OF THE MLR MODELS FOR DIFFERENT CONTINENTS

TABLE 7. SUMMARISED METRICS OF ARIMA MODELS FOR DIFFERENT CONTINENTS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | best RMSE | baseline RMSE | tuned RMSE | AIC | BIC |
| Africa (1,0,1) | 0.029 | 0.029 | 0.03 | -246.901 | -236.771 |
| Asia | 0.27 | 0.271 | 0.27 | -51.841 | -41.42 |
| Europe | 0.525 | 0.525 | 0.529 | 128.006 | 138.427 |
| Oceania | 0.248 | 0.248 | 0.258 | -35.575 | -25.154 |
| North America | 0.282 | 0.282 | 0.29 | 10.908 | 21.328 |
| South America | 0.268 | 0.278 | 0.268 | 19.04 | 29.46 |

TABLE 8. EXPERIMENTAL RMSE VALUES FOUND FROM DIFFERENT TUNING METHODS (NUMBER OF NODES AND EPOCHS)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 100N, 1000E | 50N, 1000E | 200N, 1000E | 100N, 500E | 50N, 500E | 200N, 500E |
|  | RMSE | RMSE | RMSE | RMSE | RMSE | RMSE |
| Africa | 0.026720887 | 0.030761858 | 0.026126872 | 0.02787925 | 0.0266334 | 0.03114028 |
| Asia | 0.238866573 | 0.238611979 | 0.242049133 | 0.23939906 | 0.2400097 | 0.24069785 |
| Europe | 0.496145872 | 0.50277535 | 0.489619811 | 0.50084465 | 0.5039723 | 0.49218285 |
| Oceania | 0.259583678 | 0.25864847 | 0.260477148 | 0.26222992 | 0.2586114 | 0.25895243 |
| North America | 0.258040322 | 0.259890579 | 0.258571653 | 0.25860428 | 0.2558186 | 0.25686828 |
| South America | 0.261163405 | 0.261284622 | 0.247325094 | 0.25556246 | 0.258847 | 0.25161855 |

TABLE 9. SHOWING THE TEMPERATURE CHANGE PERCENTAGE BETWEEN JANUARY 1870 AND JANUARY 2013

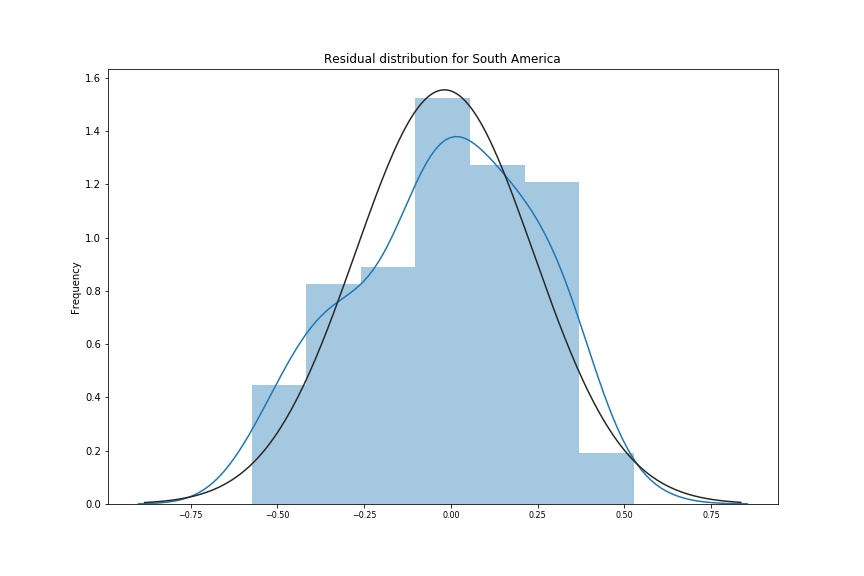
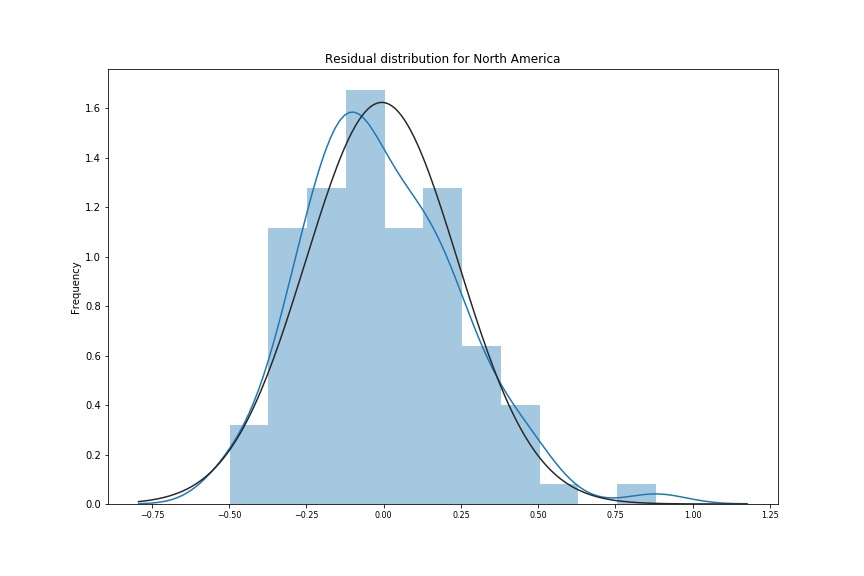
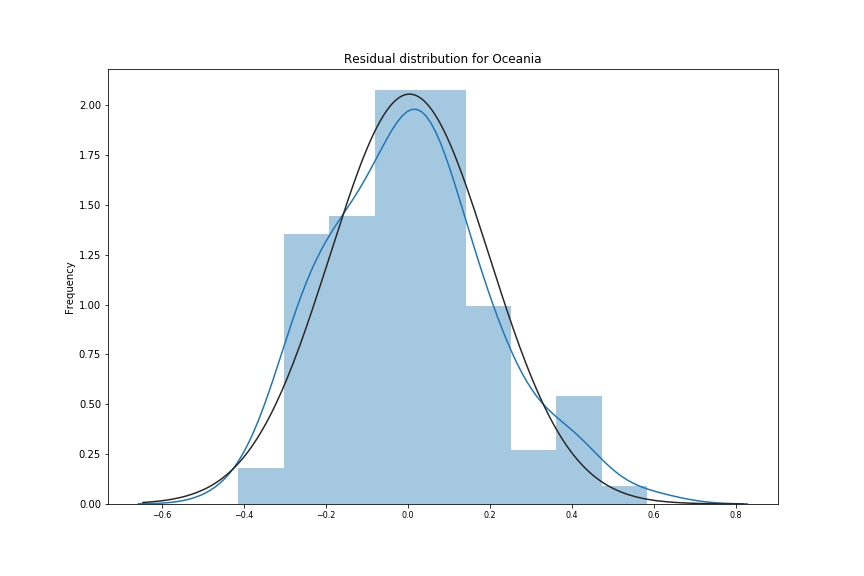
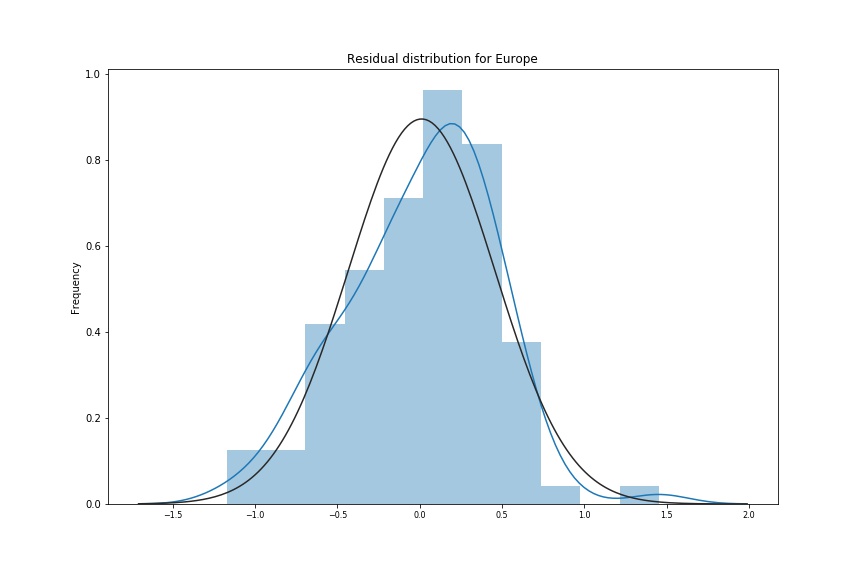
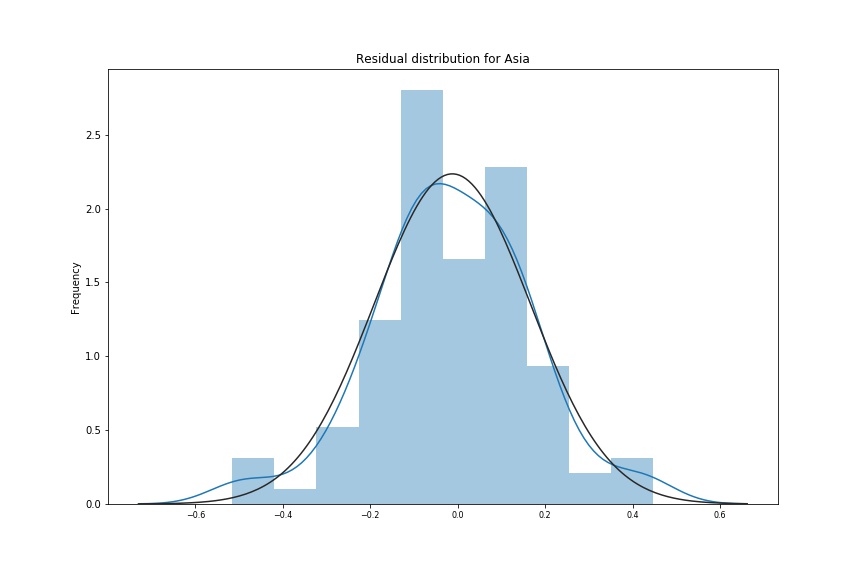
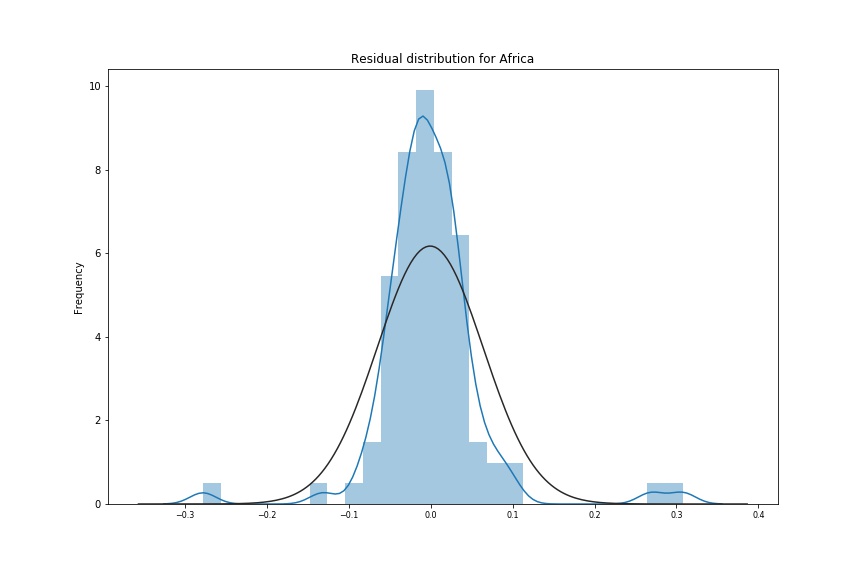
|  |  |
| --- | --- |
|  | Temperature change % (1870-2013) – 3 d.p. |
| Africa | 2.139 |
| Asia | *7.413* |
| Europe | *20.219* |
| Oceania | *4.568* |
| North America | *11.417* |
| South America | *0.837* |

# Appendix C

**Link to OneDrive for Jupyter Notebook with analysis and code:**

# Appendix D

**Additional Graphs**



APPENDIX C FIGURE 1. RESIDUAL DISTRIBUTION PLOTS FOR DIFFERENT COUNTRIES WITH ARIMA MODEL BASED ON THE GAUSSIAN MODEL.