National College of Ireland

The International Ecological Footprint Accounts

22 December 2020

BSHTM

Data Analytics

2020/2021

Brian McGrath

x15580167

x15580167@student.ncirl.ie

Technical Report

Contents

Executive Summary

This project aims to analyse the Ecological footprints of every country and use predictive analysis to conclude what countries are actively trying to lower their footprint and what countries are not. Through extensive analysis of this dataset, it should be clear to identify which countries are using more ecological resources and services than they have. This study should highlight which countries have the most cropland, fishing ground and built-up land, all measured in global hectares. These factors can affect how much carbon a country will produce; this study should find a link between these and how they affect carbon levels.

This project is based around sustainability as it has such a profound effect on how millions of people live their lives; sustainability is a conversion that is not going away and can help save wildlife and areas across the globe. Sustainability is a subject that is relevant now and always will be for as long as humans survive. As of now, 80% of countries on the planet have pledged to help lower carbon in the earth atmosphere; for this to work, each country must make improvements and sacrifices, learn how to live more sustainably, relying less on other countries resources.

The findings in this report will mainly be how countries compared to others of similar size and resources; for the first semester, findings will be based on continents as any further division and analysis of the dataset will take more time.

1.0 Introduction

1.1. Background

This project will understand how countries react to the current environmental crisis and how they have improved in recent years. The data contains 84,000 rows with tons of duplicate data; this was a significant benefactor to why this dataset was chosen. This dataset is challenging, and that is what is needed to achieve a high grade.

This dataset stood out when researching possible options as it contains a lot of numerical columns; numerical columns can be used in a lot of different ways compared to characters and factors. This project's finding should yield engaging and real-life conclusion and, hopefully, through data mining and machine learning, a glimpse into the future and how countries will continue to improve.

1.2. Aims

The project will extract new knowledge from the data and display the data using different visual software such as SPSS and Tableau. The dataset now is challenging to read, extract knowledge from and impossible to visualise. Once the project is complete, it is hoped that the data for the years before 2014 will be calculated using predictive analysis. Data Mining and Machine learning is something that will be incorporated during the latter stages of the project. By the end of the project, it should be possible to display which countries can ecologically maintain themselves on a map using ggplot2 (subject to change).

The goal is to display an in-depth knowledge using Excel, Rapid Miner, SPSS and Tableau, and an advanced understanding of R and Python's programming languages. A balanced mixture of the software's should lead to some fascinating statistical findings.

1.3. Technology

This project's primary programming language will be R, with Python being introduced in the later stages; this is subject to change as more research is done to discover what is possible with R and what is not. R Studio is the IDE used to write the R code; the dataset has been imported and cleansed using R studio, any missing values were removed, which reduced the dataset from 87,020 rows to 54,330 rows. The initial expectation was to cleanse all the data in R Studio to be ready to be transformed and visualised. However, this was not the case as the dataset contains a ton of duplicate data; it was easier to use Excel to divide the dataset into continents and countries.

Excel was used to divide each country into a table and calculate all numerical columns; this was done as they were needed to replace the 0 values scattered throughout the dataset. Every 0 value needed its own countries mean values, so it was necessary to create the 162 separate tables. Again, it was hoped that R could replace the 0 values with their mean values, but Excel was the only option. All 0 values were replaced using the If (function in Excel, e.g. =IF(K2=0,K232,K2), this code told excel to replace any 0 value with the mean value was situated at the bottom of each table. Excel was also used to summarise each continent and countries mean values.

SPSS will be used mainly for its chart building and analysis features; the data imported into SPSS will be snippets of the actual dataset as the original is too large and complex to achieve knowledge worthy visualisations.

Rapid Miner and Tableau have yet to incorporate technologies into the project; this work will take place after Christmas once there is more time to focus solely on this project. Rapid Miner is a software that is yet to be used.

2.0 Data

The dataset was found while researching Kaggle; the data was sourced from the National Footprint Accounts. The National Footprint accounts run a website that is updated every year by the National Footprint and Biocapacity Accounts, the primary provider of data needed for ecological analysis footprints worldwide (Data and Methodology - Global Footprint Network, 2021)

This data studies how much biologically productive area it takes to sustain people's demands; people need space to grow crops, feed livestock, deforestation to aid infrastructure and timber regeneration to absorb carbon emissions from things such as fossil fuel burning, methane and waste disposal. Imports are added and exports subtracted; this is how countries total consumption is measured. Each column in the dataset represents yields of primary products to measure the area needed to support said activity (grazing land, cropland, forest land and fishing grounds). If a country wanted to reduce carbon and decided to increase forest land that would decrease cropland, this might not be the correct decision as all activities need a particular area to produce for its population.

The original dataset contained 87,020 rows and had a considerable number of missing values predominately in the Per capita GDP column; the decision was made to remove all the column with missing values as they would have had too large of an effect on the study's findings.

The data has been used to conduct multiple studies on how humans affect the earth's ecosystem. In 2013, a paper was written that talked about the demand and supply of the biospheres regenerative capacity using the National Footprint Accounts data. The paper outlined how humans high demand on the earth's biosphere is far greater than the earth's regenerative and absorptive capacity. The paper also documents the latest techniques for calculating countries biocapacity and Ecological Footprint (Borucke et al., 2013).

A paper written in 2009 focused on how to improve the National Footprints accounts by suggested that it should be used more by governments and multinational corporations across the globe, rather than the accounts to be perceived as purely an academic exercise. The core message delivered from the National Footprint accounts is that the world is currently living

well beyond its means and that the situation is worsening. The study also concludes that residents live in higher-income countries and demand more productive capacity than low-income countries.

3.0 Methodology

It was decided at the beginning that the project would follow the Knowledge Discovery in Databases (KDD) methodology. This methodology was chosen for its core data analytics and ties to data mining. The project aims to extract functional, structured patterns from the dataset; KDD methodology will aid the smooth running of this project with good, knowledgeable, and understandable data patterns.

The selection of the chosen dataset was down to the relevant interest in the data and the potential findings, and the high number of numerical data, leading to more accurate statistics and patterns later. The dataset chosen allows for the years following 2014 to be predicted using the KDD methodology.

Pre-processing the dataset was a big task as the data contained many missing values, 0 values and duplicate rows. The dataset needs to be broken down into hundreds of subsections to gather accurate findings on individual countries as each country has multiple values per column each year. The pre-processing was done using R Studio and Excel, the CSV file was imported into R Studio, and any missing values were identified and then omitted from the dataset. The columns UN_region, sub-UN_region and record, were all changed to their correct datatype and are now factors; this allows the data to be easier to read when summarised. The data still contained 0 values through many of the columns; it was decided to change the 0 values to that country's mean value for the mid-point presentation. Once we begin to study machine learning, those figures will be changed to their expected values. Excel was used to separate the data into tables for each country so that their mean values could be identified and exchanged for 0 values.

Transformation of the data has already begun as the first few valid and understandable visualisations are created. Dimensionality reduction will play a significant role in this project as the original dataset is far too large to extract meaningful statistics or visualisations. The process of dimensional reduction has already begun, new datasets have been created to make sense of the massive volume of data, each continent now has its datasets with all their mean values uploaded in R Studio and SPSS with the hope of extracting some good knowledge (Maduranga, 2020). The process of dividing the dataset into their own countries and records

will begin after Christmas; graphs and statistics presented before then may not be accurate, but it is a steppingstone in the right direction.

4.0 Analysis

Many obstacles were faced when conducting meaningful analysis on the dataset; after pre-processing the data, it needed to be separated into smaller sections of more defined data. Firstly, the data was divided into 162 tables for each country where the mean value was found for each numerical column to replace the 0 values. The countries were split using Excel; tables were inserted, and the AVERAGE(function was used to find the mean values and insert them into the dataset. This data was then summarised and imported into R Studio as a separate dataset to see if any exciting statistics could be discovered. The primary purpose of finding the mean values was to replace the 0 values within the dataset. However, the additional mean dataset will present findings in the latter stages of the project.

The additional dataset was then separated to be one for each continent (Africa, Asia, Europe, Latin America and the Caribbean, North America, and Oceania). Separating the data like this makes the visualisation more comfortable to read and the statistics easier to calculate. The data was split into continents using the filter function in Excel; once separated, worksheets were created and then imported into R Studio and SPSS for further analysis.

At this point, the data captured is not accurate, as many rows contain mean values instead of their expected values. Analysis has taken place only on the separated datasets for the mid-point as the readability is much better and results are more localised.

After the mid-point presentation, it was time to analyse the original dataset, the 0 values in the dataset signifies that there is no build-up of carbon in that specific area. Keeping the 0 values within the data for the mid-point presentation was not possible as the 0 values hindered the analysis. The ARIMA model is the time-series technique chosen to analyse how countries ecological footprint will evolve from 2015 to 2022. The results will also indicate whether the highlighted countries are ecological debtors or creditors. A country's ecological footprint is measured by analysing their annual demand for goods and services, and their biocapacity is measured by analysing the resources they have available. A countries ecological footprint is subtracted by their biocapacity to determine whether they are an ecological debtor or creditor. Each country had its data separated into five datasets: biocapacity per capita, eco-footprint of

consumption, eco-footprint of exports, eco-footprint of imports and eco-footprint of production. This was done as all five datasets needed to be forecasted to produce the results.

An Auto-Regressive Integrated Moving Average (ARIMA) model was performed on the data of 10 countries to predict and analyse the trend of their biocapacity levels and ecological footprints. The ARIMA contains within it an autoregressive model and a moving average model. Firstly, the was analysed using the Augmented Dickey-Fuller test to determine whether the data was stationary; the Null Hypothesis HO is that the time-series is non-stationary, which will be accepted when $p > 0.05$, the Alternate Hypothesis is that the time-series is stationary, which will be accepted when $p < 0.05$. The parameters of the ARIMA models were chosen based on the value of the Akaike Information Criterion (AIC). Usually, the ARIMA model that outputs the lowest AIC value is the most accurate.

Accuracy is essential when choosing an appropriate time-series model for forecasting; each ARIMA output runs six performance errors measures to determine the model's accuracy.

- **Mean Error (ME)** – The mean error is a statistical test that refers to the average error within the dataset, any uncertainty in measurement is classified as an error (Stephanie, 2016).
- **Root Mean Square Error (RMSE)** – This statistical test calculates the standard deviation of the residuals. The residuals are the prediction errors that measure the distance from the data points to the regression line (Glen, 2016).
- **Mean Absolute Error (MAE)** – This statistical test represents the average absolute values of the individual prediction errors (Sammut and Webb, 2010).
- **Mean Absolute Percentage Error (MAPE) -** This statistical test measures how accurate the forecast is; the error within the forecast is measured as a percentage (Stephanie, 2021).
- **Mean Absolute Scaled Error (MASE) –** This statistical test outputs each error within the forecast as a ratio to the average error (Stephanie, 2019).

.

Performance of ARIMA to forecast Irelands biocapacity & footprint.						
	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	0.0025	0.0794	0.0616	0.0544	1 .5208	0.7963

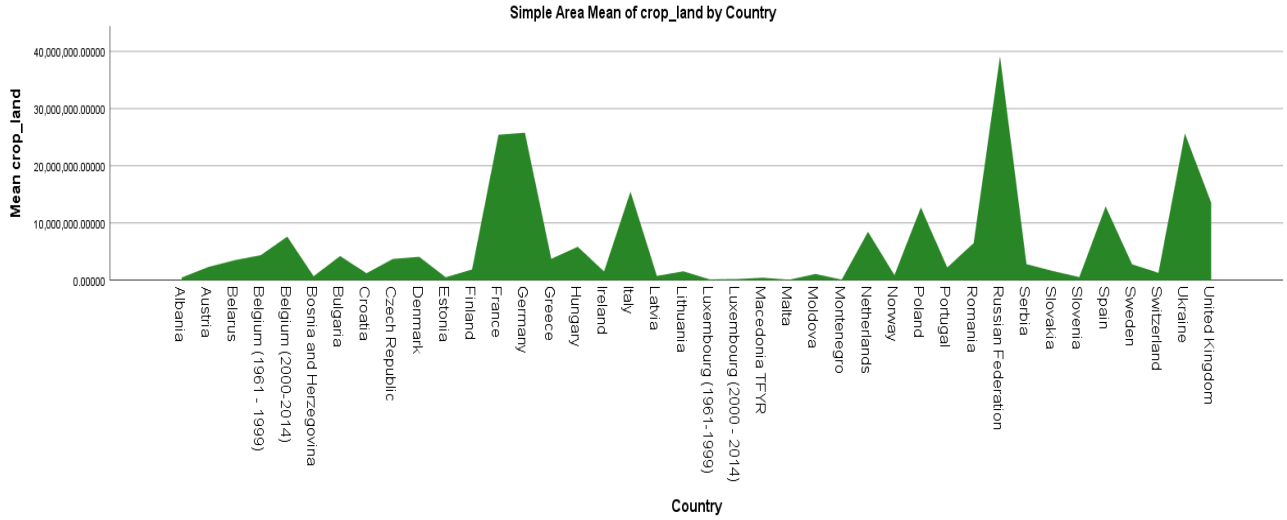
Table 1

The lower the values represented in Table 1, the more accurate the time-series model, ARIMA model performed remarkably for all countries data. Ireland's ARIMA had a tiny 1.52% of mean absolute percentage error, and only 0.0794 for root mean square error, meaning that data is concentrated around the line of best fit.

5.0 Results

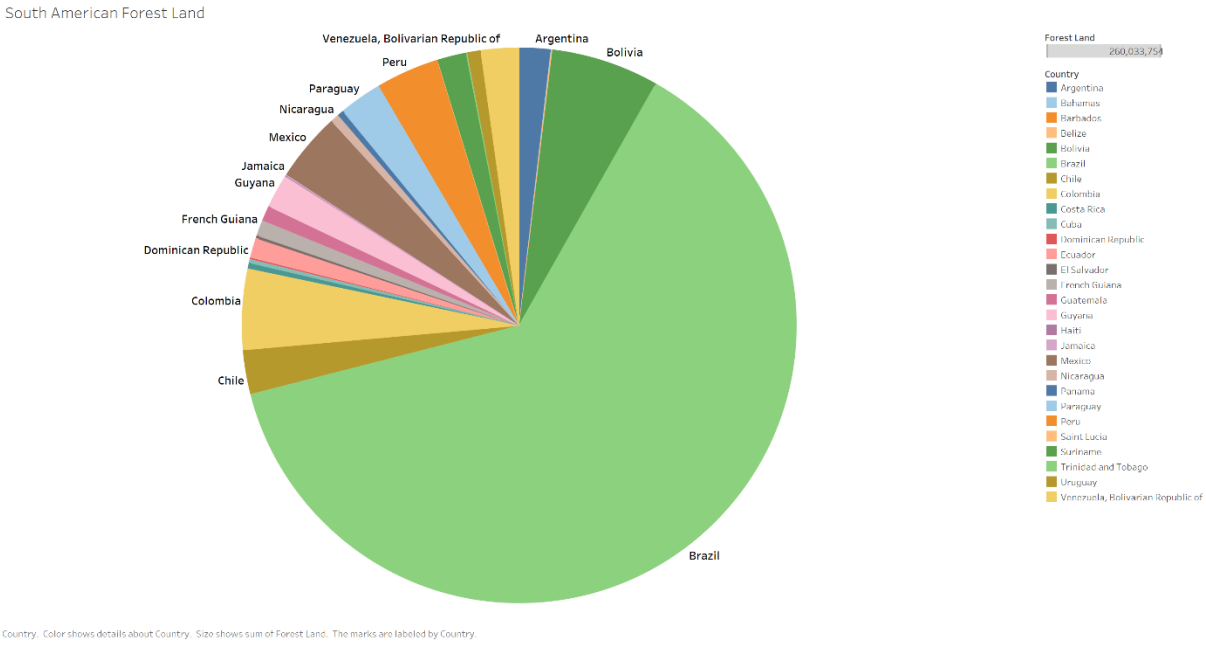
Summary of Continents

This section of the results is analysing Europe, South America, Asia, and Africa using their mean values. The mean values of each of the continents were calculated so that the data could be visualised. Separating the data into continents allows every country to be analysed and compared against others in the same region. Many countries within a continent live under similar circumstances, yet there can be significant differences in the level of carbon produced when compared against one another.



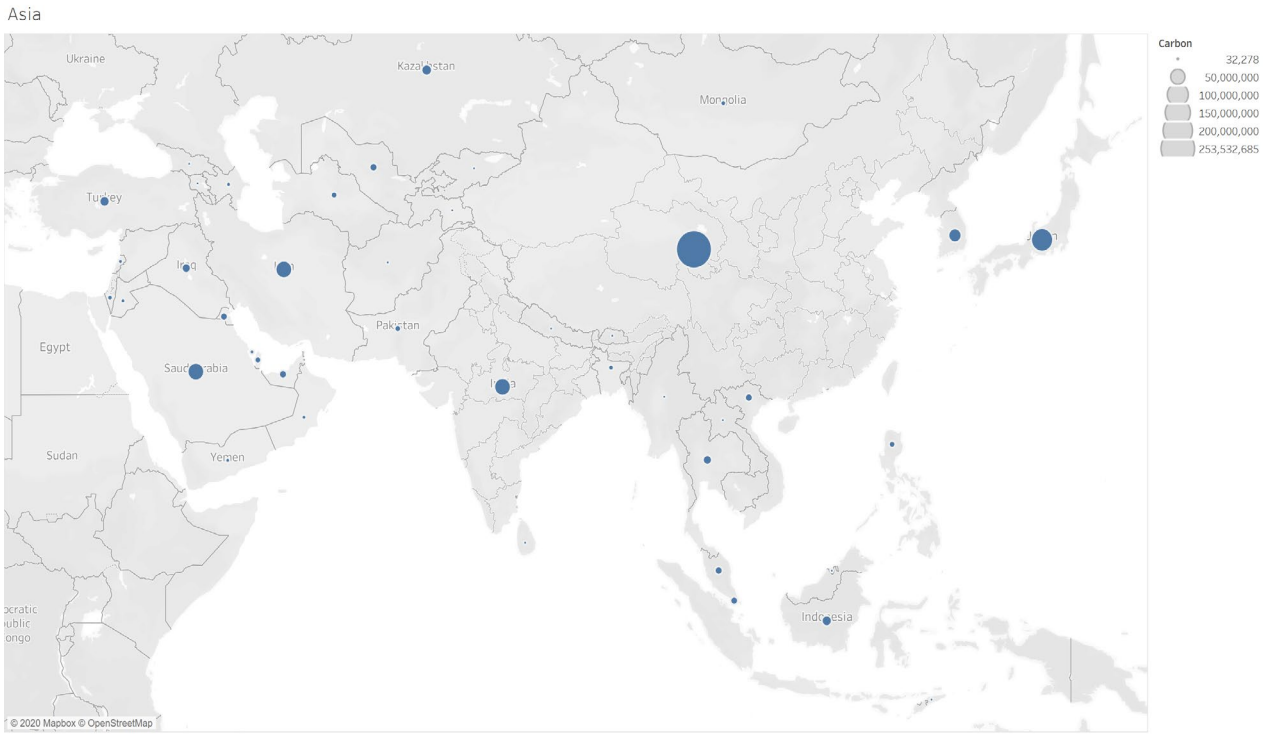
Cropland in Europe

Above is an area chart created using SPSS; the chart represents all the cropland in Europe for each country; the chart clearly shows that larger nations have more global hectares of cropland than smaller nations. Nations such as Russia, Germany and France lead the way in Europe for cropland. Russia contains a massive average of 39,151,927GHA of cropland, while Malta only contained 69,659.34GHA of cropland, a range of 39,082,268GHA between Europe's largest and smallest countries.



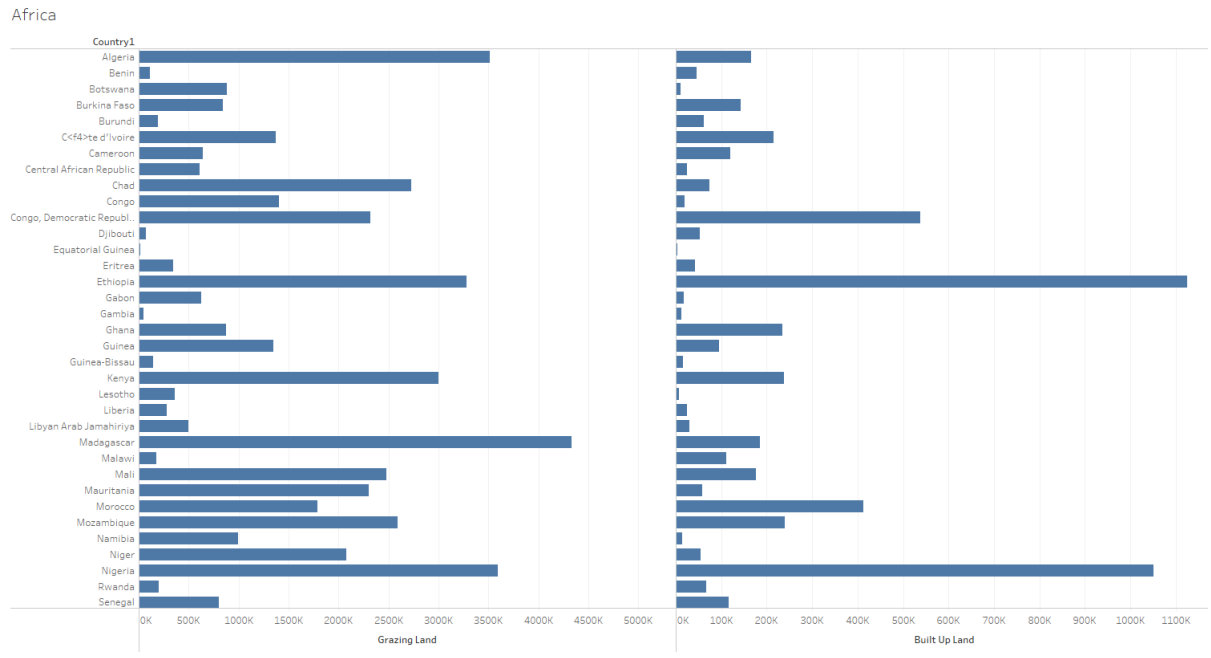
South American Forest Land

The chart below was created in Tableau and represents all the forest land in South America; forest land is known for absorbing carbon in the atmosphere. South America is full of dense jungles and rainforest and is home to the Amazon, the world's largest rainforest. The Amazon predominantly covers Brazil, which is evident in the pie chart below. However, the Amazon rainforest extends into Columbia, Peru, Bolivia, Ecuador, Venezuela, Guyana, Suriname, and French Guiana. Brazil owns 163369086GHA of South America forest land compared to Saint Lucia's 7577.681983GHA of forest land. That is a difference of 163361508GHA between South America largest and smallest forest land.



Carbon Map of Asia

The interactive map above was created using Tableau, and blue circles represent the carbon levels in each country; the more significant the circle, the higher levels of carbon that country has. China is leading the way in Asia and worldwide with 3,688,446,407, and Japan produced the second-highest carbon levels in Asia with 89,252,892.



Grazing vs Built-up Land Africa

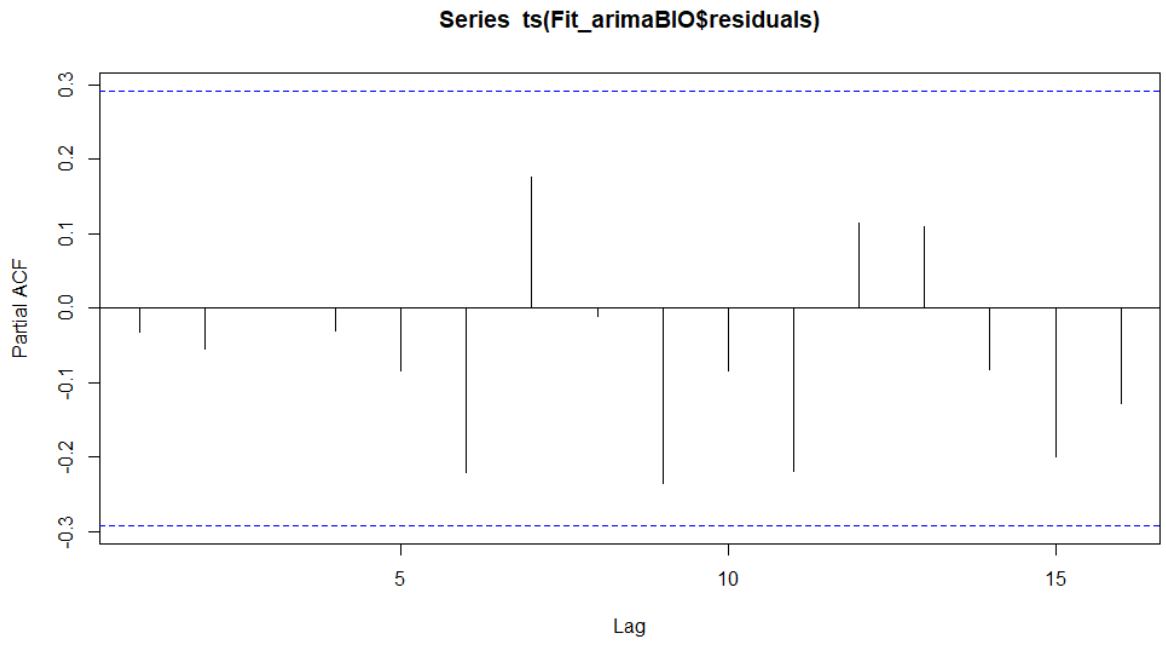
The above chart was created in Tableau and represents the grazing land and every African nation's built-up land. Africa has a most considerable amount of grazing land than any other continent and has much scarce land that may be marked as grazing land; as shown in the chart above, Africa has minimal built-up land compared to other continents. Africa's small amount of built-up land may be a factor for its large volume of built-up land.

Time-Series Results

An Auto-Regressive Integrated Moving Average (ARIMA) model was employed for eight countries to analyse how their biocapacity, consumption, production, imports, and exports have evolved from 2015 to 2022 variables possible the calculate the countries ecological footprint and their ecological balance. Conducting a time-series model on these countries will provide dynamic information and changes occurring with their land usage and carbon levels. Five ARIMA models had to be run on each country to calculate their results.

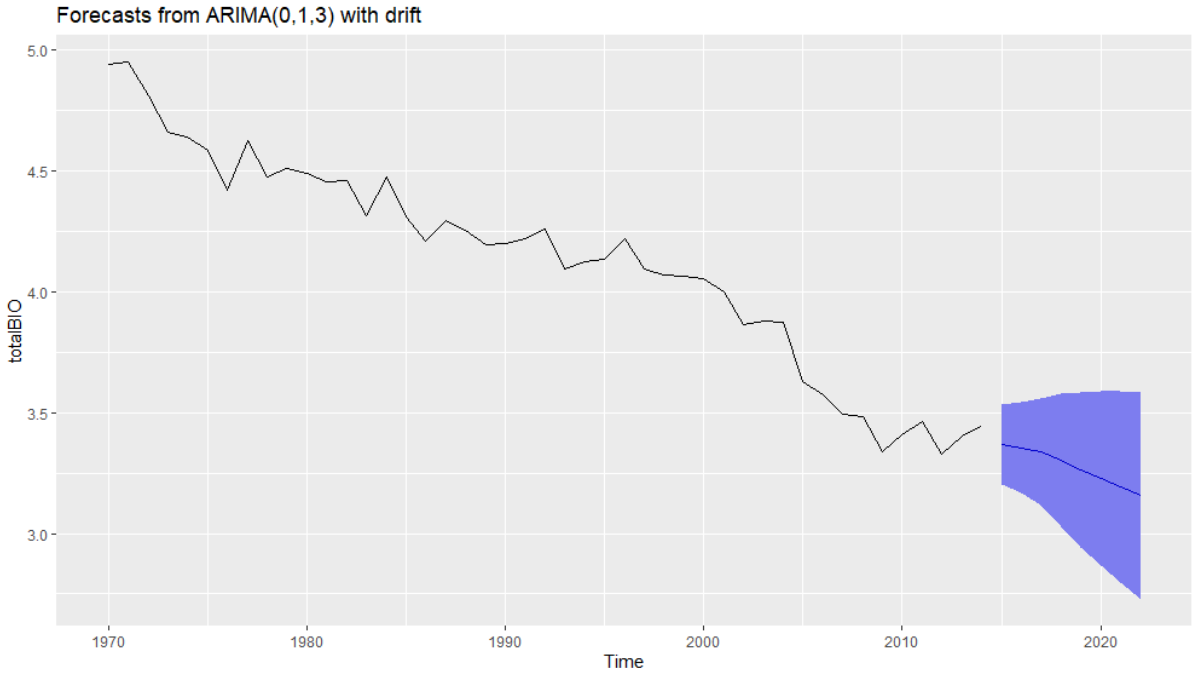
Ireland

Ireland is the smallest country that has been chosen for analysis; Ireland's results are focused on the biocapacity levels. As of 2014, Ireland's most significant biocapacity resource is its fishing ground with 1.476 global hectares per capita, and their most minor resource is their built-up land with only 0.101 global hectares per capita. The ARIMA model chosen to forecast Ireland's biocapacity was (0,1,3) as it outputted the lowest AIC value.



PACF of ARIMA residuals

In the partial autocorrelation function plot above, the broken blue line represents the significance threshold; no spikes exceed the significance threshold. Considering that no spikes have surpassed the significance threshold, it is assumed the ARIMA model is best suited for the data. The spikes represent the residuals (errors), and the line represents the biocapacity forecast. The residual standard deviation for the ARIMA (0,1,3) model is 0.084; a low standard deviation value indicates a high performing model.



Time plot of Ireland's Biocapacity

The graph above plots all the historical data and the forecast, Ireland has steadily decreasing biocapacity levels since 1970. The point forecast line surrounded in blue is the ARIMA models forecast, which predicts that Ireland's biocapacity will continue to decline until 2022. The blue area represents the upper and lower 95% intervals; this is where the prediction can deviate from the point forecast line.

Year	BIO	EFC	EFE	EFI	EFP	Trade	Footprint	C/D
2015	3.3696	4.7067	4.4609	4.2329	4.8152	-0.2279	4.5873	-1.2177
2016	3.3552	4.7067	4.5224	4.2329	4.8360	-0.2895	4.5465	-1.1913
2017	3.3375	4.7067	4.5840	4.2329	4.8355	-0.3511	4.4844	-1.1469
2018	3.3014	4.7067	4.6456	4.2329	4.8181	-0.4126	4.4055	-1.1042
2019	3.2652	4.7067	4.7071	4.2329	4.8295	-0.4742	4.3553	-1.0901
2020	3.2290	4.7067	4.7687	4.2329	4.8227	-0.5358	4.2870	-1.0580
2021	3.1928	4.7067	4.8303	4.2329	4.8200	-0.5973	4.2227	-1.0299
2022	3.1566	4.7067	4.8918	4.2329	4.8233	-0.6589	4.1644	-1.0078

Table 2

The table above are the outputs of the ARIMA model results; the Trade, Footprint, and C/D columns were calculated using Excel. The Trade column is Ireland's imports (EFI) minus their exports (EFE), the Footprint column is Trade plus production (EFP), and the C/D column, which stands for creditor or debtor, is Ireland's biocapacity (BIO) minus their footprint.

Trade = EFI – EFE

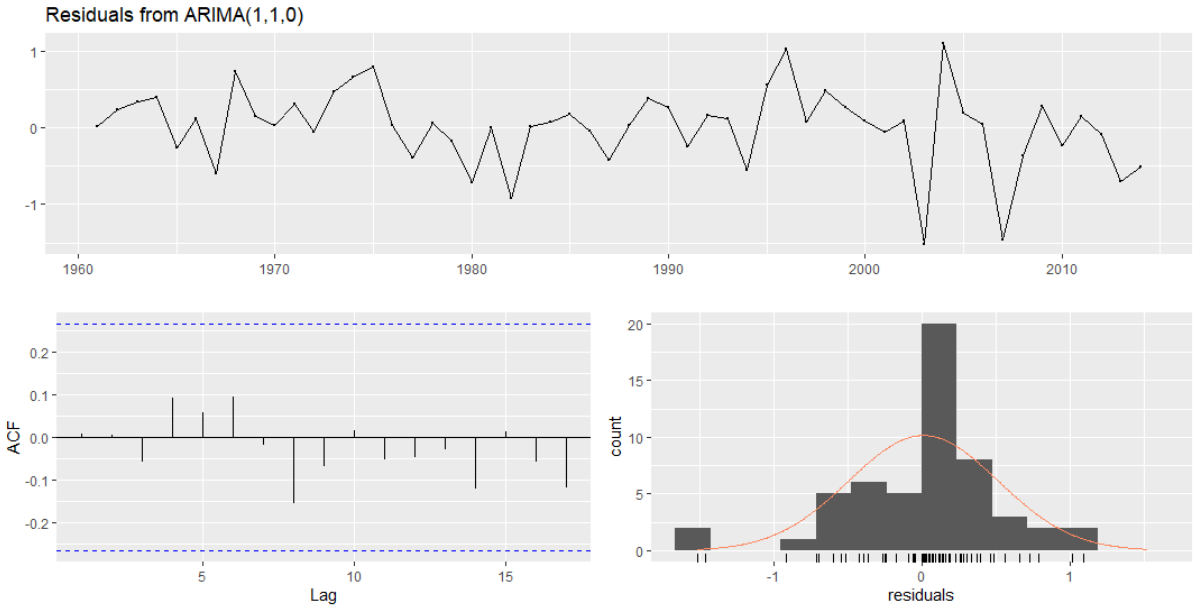
Footprint = Trade + EFP

C/D = BIO – Footprint

Ireland is currently an ecological debtor; however, the ARIMA model forecast that Ireland is to steadily improve its ecological balance from 2015 to 2022. Ireland's ecological footprint is also improving. Like many governments worldwide, Ireland has pledged to create a greener future for its residents, and the results show that its efforts are paying off.

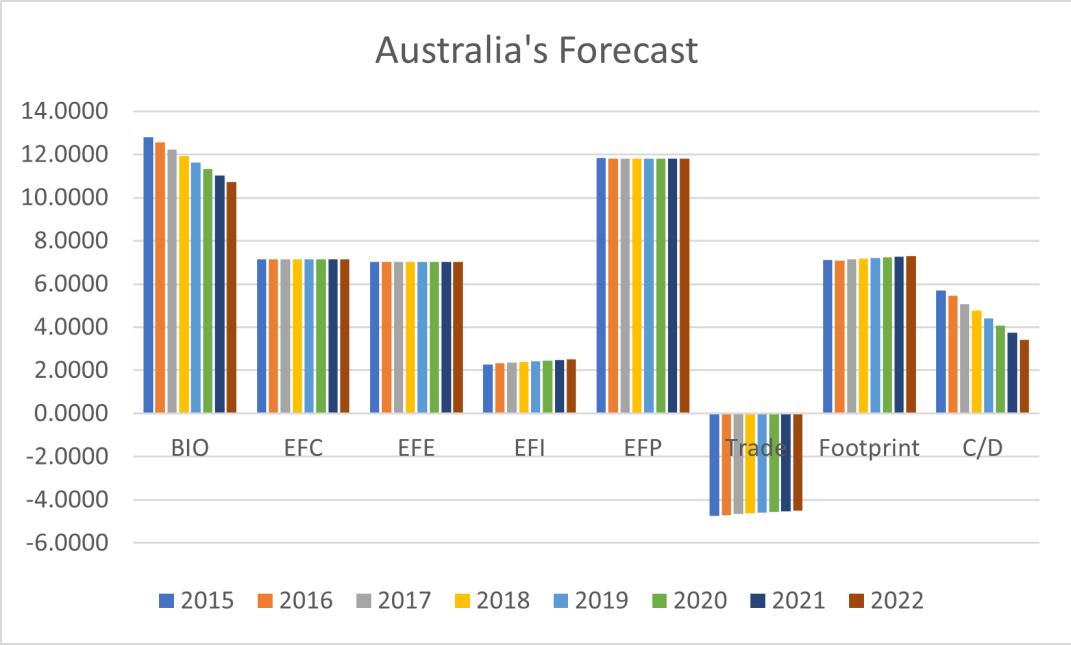
Australia

Australia's results are focused on their ecological footprint of production (EFP). As of 2014, Australia's most valuable production resource was their forest land with 0.8627 of global hectares per capita, and their least valuable production resources were their built-up land with 0.1217 global hectares per capita. In 2014, Australia's production activities created 4.700 global hectares per capita of carbon in the atmosphere. The ARIMA model best suited to forecast Australia's production is ARIMA (1,1,0). The residual standard deviation for this model is 0.509.



Residuals of Production Australia

The graph above plots all the residual visualisations of ARIMA (1,1,0). All spikes with the auto-correlation function are within the significance threshold, meaning there is no additional modelling needed. The time plot on the top half of the graph is the difference time plot; difference time plots remove the trend from the data and highlight the change and fluctuations that occur. The difference time plot shows that 2000 to 2010 were high volatile years in Australia in terms of production.



Australia's Forecast

Above summary of Australia's forecast from 2015 to 2022. The graph shows that Australia's biocapacity is steadily decreasing. Australia's consumption and exports are to remain level, and their imports are to increase slightly. Australia is exporting more than double the amount of goods they are importing; this trend is set to continue up until 2022 with minimal fluctuations in between.

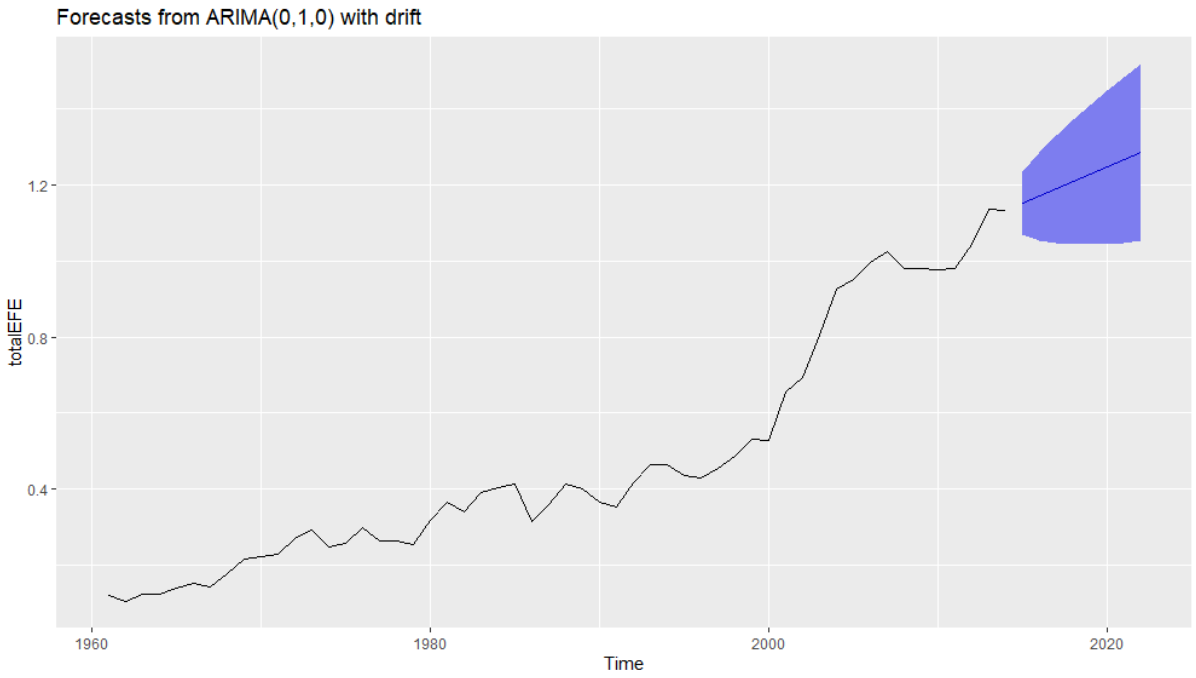
Year	BIO	EFC	EFE	EFI	EFP	Trade	Footprint	C/D
2015	12.7934	7.1314	7.0142	2.2833	11.8342	-4.7308	7.1034	5.6900
2016	12.5587	7.1314	7.0142	2.3169	11.7937	-4.6972	7.0965	5.4622
2017	12.2265	7.1314	7.0142	2.3505	11.8092	-4.6637	7.1456	5.0810
2018	11.9370	7.1314	7.0142	2.3841	11.8033	-4.6301	7.1732	4.7637
2019	11.6288	7.1314	7.0142	2.4177	11.8056	-4.5965	7.2091	4.4197
2020	11.3288	7.1314	7.0142	2.4513	11.8047	-4.5629	7.2418	4.0870
2021	11.0252	7.1314	7.0142	2.4849	11.8050	-4.5293	7.2758	3.7494
2022	10.7232	7.1314	7.0142	2.5185	11.8049	-4.4957	7.3092	3.4139

Table 3

Australia is an ecological creditor with a value of 5.690 at the beginning of 2014. The balance value will decrease by 2.276 with a forecasted value of 3.4139 in 2022. The table also shows that Australia's production is forecast to decrease slightly over the eight years.

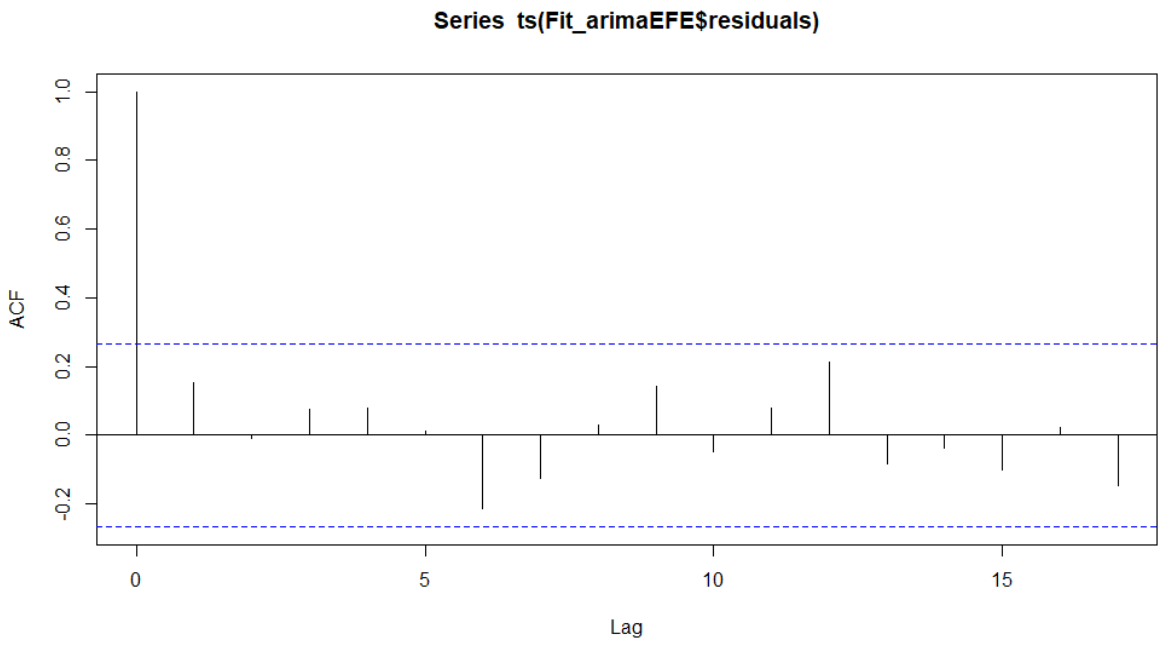
Brazil

Brazil is home to the Amazon rainforest, which is the worlds largest intact forest. The Amazon is home to 10% of all animal and plant life on Earth and plays an essential role in managing the planets carbon levels (‘Brazil and the Amazon Forest’, 2019). The ARIMA model conducted on Brazil’s ecological footprint of exports is analysed. Surprisingly, Brazil’s largest export resource is its crop land at 0.498 global hectares per capita and its second largest is forest land at 0.217 global hectares per capita. The ARIMA (0,1,0) model that had the lowest AIC value was chosen to forecast exports for Brazil.



Time plot of Brazil's Exports

Brazil’s exports have increased by 1.1619 global hectares per capita since 1961 with the 2022 forecast value at 1.2838. The ARIMA model is forecasting that exports are expected to increase beyond 2022. A Box-Ljung test for the forecast produced a p-value of 0.8429, it is assumed that the forecast is accurate as the p-value > 0.05. Any p-value above 0.05 indicates that there are auto correlation issues with the data.



ACF of ARIMA Residuals

The ACF chart above shows that one spike has passed the significance threshold, this means that the ARIMA model is not forecasting all the data; however, this is not enough to consider utilising a different technique. If two or three spikes had surpassed the significance threshold then it may be worthwhile attempting the Auto-Regressive (AR) model.

Year	BIO	EFC	EFE	EFI	EFP	Trade	Footprint	C/D
2015	8.7589	3.0777	1.1504	0.4164	3.8186	-0.7341	3.0845	5.6743
2016	8.6617	3.0777	1.1695	0.4164	3.8445	-0.7531	3.0914	5.5703
2017	8.5609	3.0777	1.1885	0.4164	3.8704	-0.7722	3.0982	5.4627
2018	8.4565	3.0777	1.2076	0.4164	3.8963	-0.7912	3.1051	5.3515
2019	8.3487	3.0777	1.2266	0.4164	3.9222	-0.8103	3.1119	5.2367
2020	8.2373	3.0777	1.2457	0.4164	3.9481	-0.8293	3.1188	5.1185
2021	8.1226	3.0777	1.2647	0.4164	3.9740	-0.8484	3.1256	4.9970
2022	8.0045	3.0777	1.2838	0.4164	3.9999	-0.8674	3.1325	4.8720

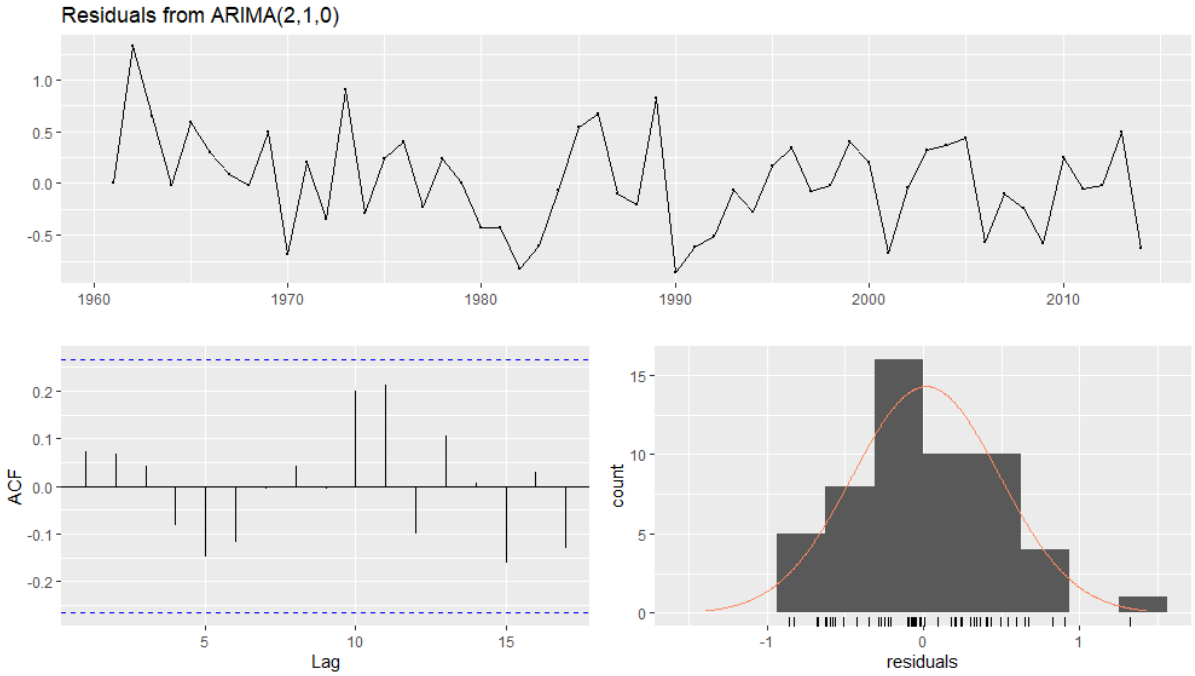
Table 4

The forecasted data for Brazil shows that its ecological footprint will grow by 0.048 global hectares per capita from 2015 to 2022. As Brazil's ecological footprint grows its ecological balance drops by 0.8023 global hectares; however, according to the forecasts Brazil will still be an ecological creditor with a value of 5.6743 in 2022. Brazil's biocapacity is expected to continue in a downward trajectory, with the forecast predicting a decrease 0.7544 from 2015 to 2022. Brazil's biocapacity level was 22.78 in 1961 and the predict value for 2022 is 8.00,

that is a decrease of 14.78 global hectares per capita, an assumption can be made that this is a result of deforestation in the Amazon rainforest.

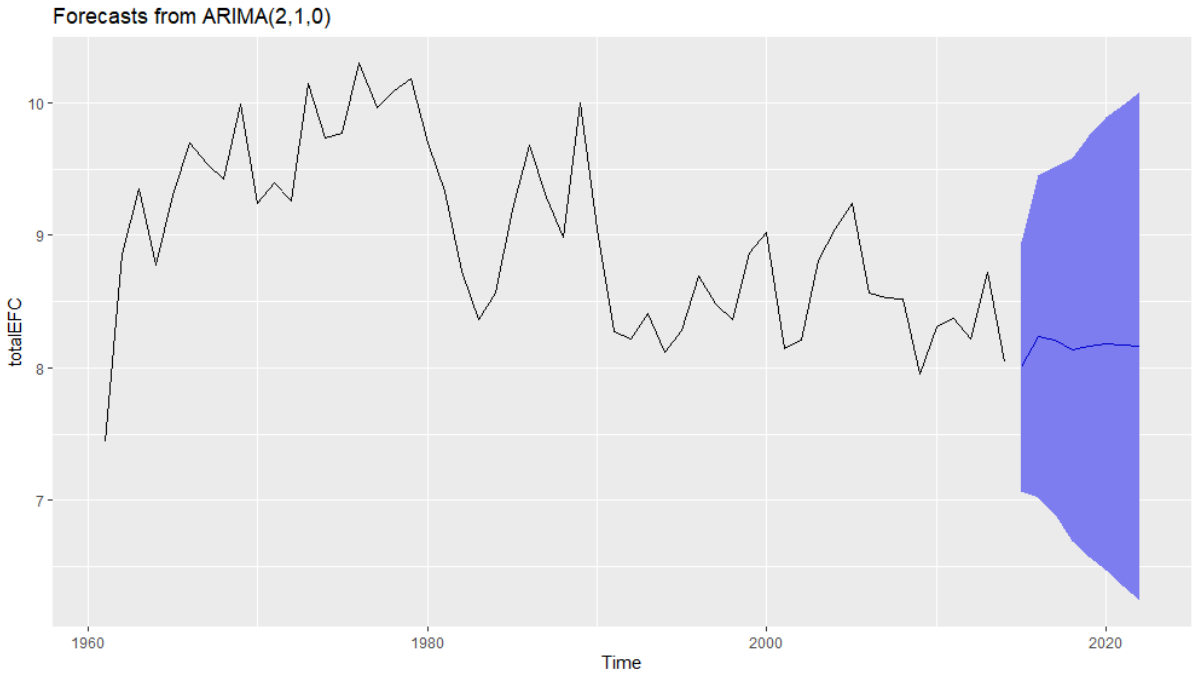
Canada

Canada's analysis is focused on its ecological footprint of consumption (EFP). Canada is well known for its dense forests which correlates to its largest resource being forest land with 1.07 global hectares per capita. Canada's resource that produces the least amount of consumption is its fishing ground at 0.137 global hectares per capita.



Residuals of Consumption Canada

The best ARIMA model to forecast Canada's consumption is ARIMA (2,1,0). The ACF plot above shows that there are no spikes surpassing the significance threshold and the mean absolute percentage error (MAPE) outputs only 4.13%. Taking these considerations into account it can be assumed that the ARIMA (2,1,0) model is accurate. The residual standard deviation for this model is 0.48.



Time plot of Canada Consumption

Above is a time plot of Canada's consumption from 1961 to 2022. Canada's consumption spiked dramatically from 1961 to 1980, from then, there has been big fluctuations in the data but ultimately the plot does not follow any trend. The ARIMA model forecasts that Canada's consumption is to level out from 2015 to 2022.

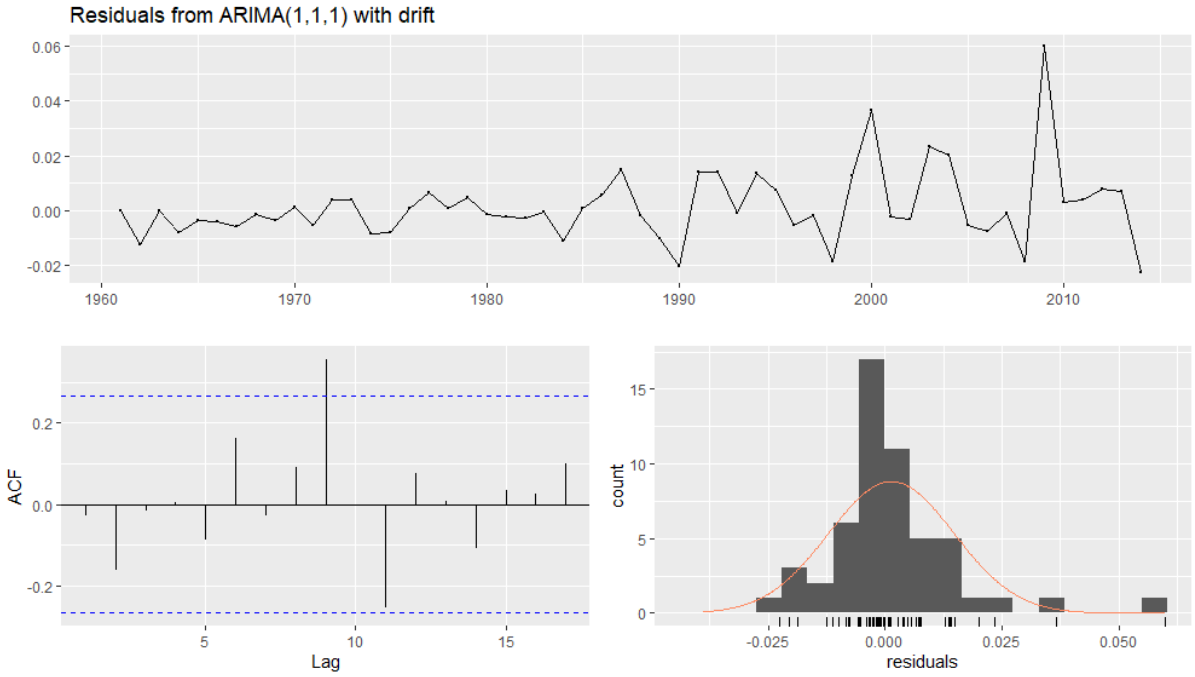
Year	BIO	EFC	EFE	EFI	EFP	Trade	Footprint	C/D
2015	14.8581	8.0073	7.1372	3.4927	11.7420	-3.6445	8.0975	6.7606
2016	14.8647	8.2366	7.1372	3.5406	11.7420	-3.5966	8.1454	6.7193
2017	14.6857	8.2087	7.1372	3.5884	11.7420	-3.5488	8.1932	6.4924
2018	14.3811	8.1381	7.1372	3.6363	11.7420	-3.5009	8.2411	6.1400
2019	14.1721	8.1602	7.1372	3.6841	11.7420	-3.4531	8.2890	5.8831
2020	13.9962	8.1795	7.1372	3.7320	11.7420	-3.4052	8.3368	5.6594
2021	13.7766	8.1687	7.1372	3.7798	11.7420	-3.3574	8.3847	5.3920
2022	13.5515	8.1643	7.1372	3.8277	11.7420	-3.3095	8.4325	5.1190

Table 5

Canada's forecasts show its ecological footprint is expected to increase until 2022 with 8.4325 global hectares per capita. Although Canada produces such a large ecological footprint it is still an ecological creditor, this is due to their high biocapacity level with a forecasted value of 13.5515 in 2022. Although, Canada's biocapacity is forecasted to decrease by 1.3066 from 2015 to 2022, factors that cause this may be the increase in population and a heavy reliance on cars and buses for transportation.

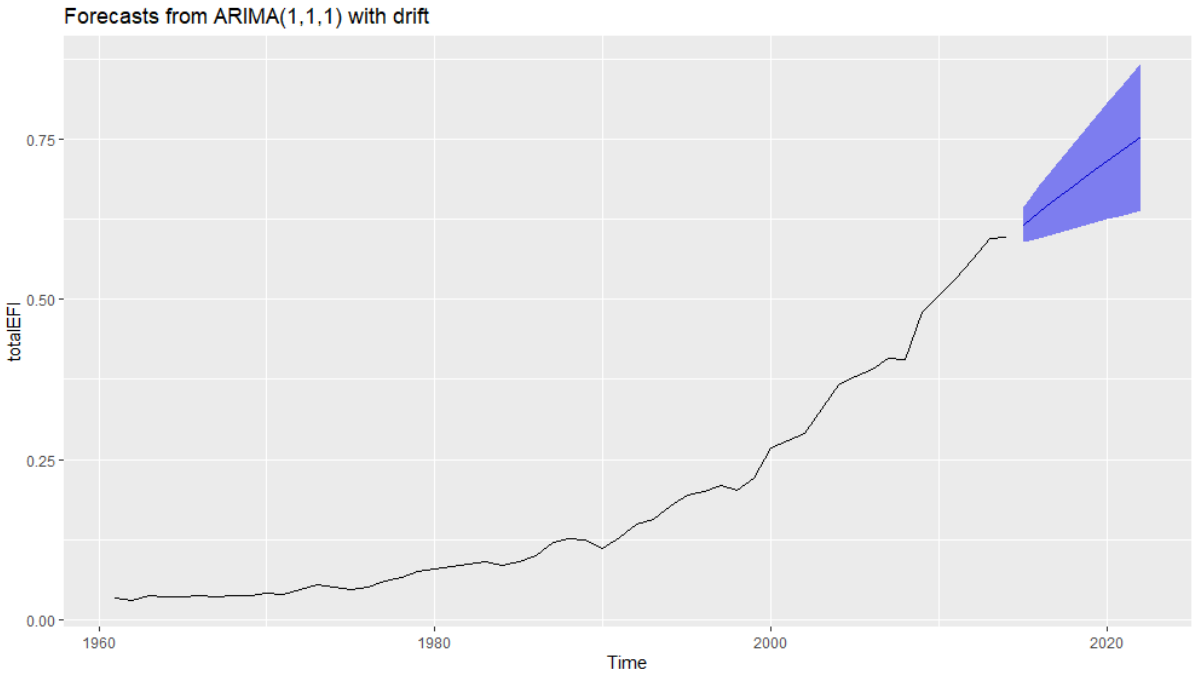
China

China is the largest producer of carbon on Earth. China produced 10.06 gigatons in 2018 (Scientists, 2020), this massive volume of carbon is due to China having a population of 1,400,602,000 in 2014. This study analyses China's ecological footprint of imports and how they have evolved from 2015 to 2022. China's largest number of imports are goods that are produced from cropland with 0.1167 global hectares per capita.



Residuals of Imports China

To forecast China's imports the ARIMA (1,1,1) model is used, this model produced the lowest AIC value; however, this model is the least accurate yet. A Dicky-Fuller test was conducted and produced a p-value of 0.99, any p-value above 0.05 indicates that the data is not stationary. The ACF plot shows that one spike has surpassed the significance threshold and the MAPE value output was 6.925045%, the highest percentage so far. After attempting an Auto-Regressive (AR) model and a Moving-Average (MA) model on the data, it was concluded that the ARIMA model produced the most accurate results.



Time plot of Imports China

The plot above represents China's imports. The time series plot shows a clear upward trend and that tend is forecasted to continue from 2015 to 2022. China's imports have increased by 0.563 from 1961 to 2014 and is forecasted to increase by another 0.136 from 2015 to 2022. China's time-series data on imports from 1961 show no sudden shifts or changes.

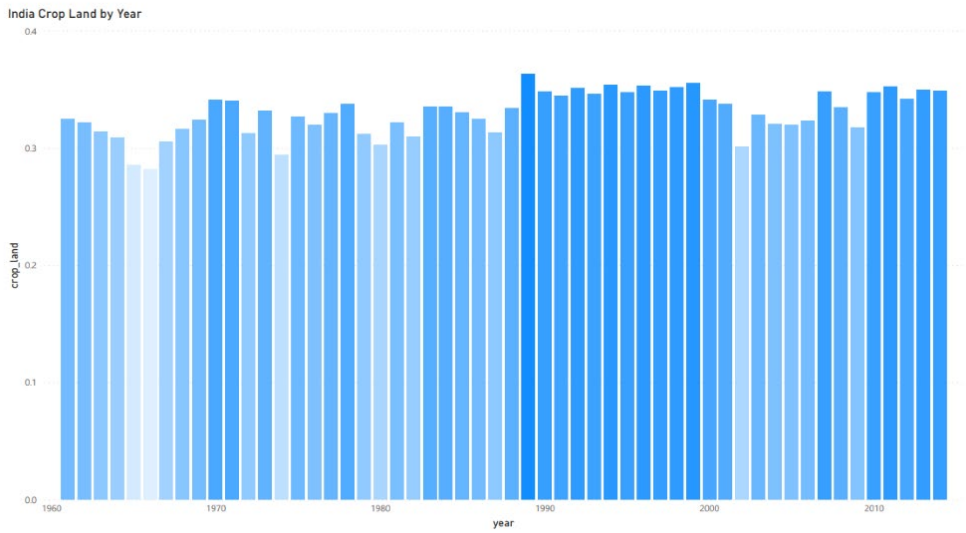
Year	BIO	EFC	EFE	EFI	EFP	Trade	Footprint	C/D
2015	0.9753	3.7107	0.4216	0.6168	3.5294	0.1951	3.7245	-2.7492
2016	0.9753	3.7300	0.4286	0.6371	3.5470	0.2085	3.7555	-2.7802
2017	0.9753	3.7605	0.4306	0.6571	3.5759	0.2265	3.8024	-2.8271
2018	0.9753	3.7976	0.4323	0.6769	3.6112	0.2446	3.8558	-2.8805
2019	0.9753	3.8388	0.4337	0.6963	3.6504	0.2627	3.9131	-2.9378
2020	0.9753	3.8823	0.4348	0.7155	3.6918	0.2807	3.9725	-2.9972
2021	0.9753	3.9273	0.4358	0.7345	3.7345	0.2987	4.0332	-3.0579
2022	0.9753	3.9732	0.4366	0.7532	3.7780	0.3166	4.0946	-3.1193

Table 6

China's ecological footprint is following an upward trend as their biocapacity is forecasted to flatline. China is an ecological debtor and is expected to worsen by 0.37 global hectares per capita from 2015 to 2022.

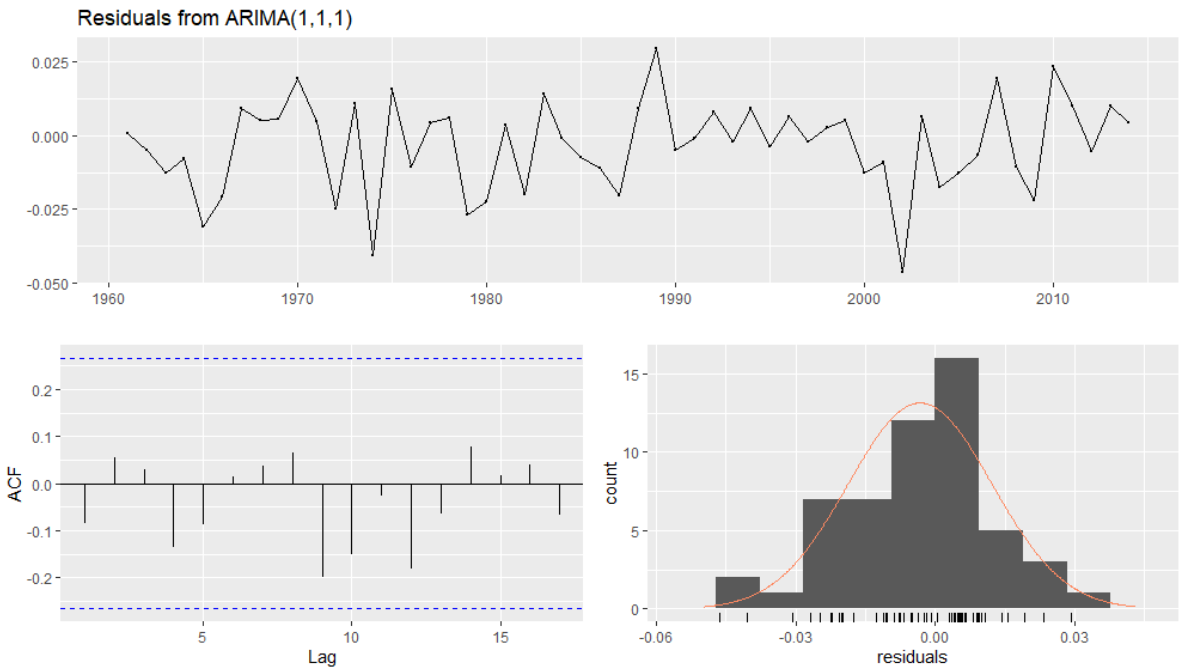
India

India's analysis is focused on its biocapacity level and how it is evolving since 1961. India's largest biocapacity resource is its cropland, with 0.349 global hectares per capita. India's most minor resource for biocapacity is its grazing land, with only 0.003 global hectares per capita.



Crop Land of India

The graph above was created using Microsoft BI; it represents India's biocapacity levels produced by its cropland. The darker the bars are, the higher the biocapacity for that year. India has experienced a lot of sudden shifts in its cropland and follows no trend.



Residuals of Biocapacity India 1

The ARIMA model best suited to forecast India's biocapacity is ARIMA (1,1,1); this model outputs an AIC value of 282.12 and a residual standard deviation of 0.017. The difference time plot removes the trend from the data and shows that India's biocapacity highly volatile and has experienced many sudden shifts since 1961. The ACF graph shows that no spikes have surpassed the significance threshold. The MAPE value for ARIMA (1,1,1) is 2.28%, so it is assumed that the model is accurate.

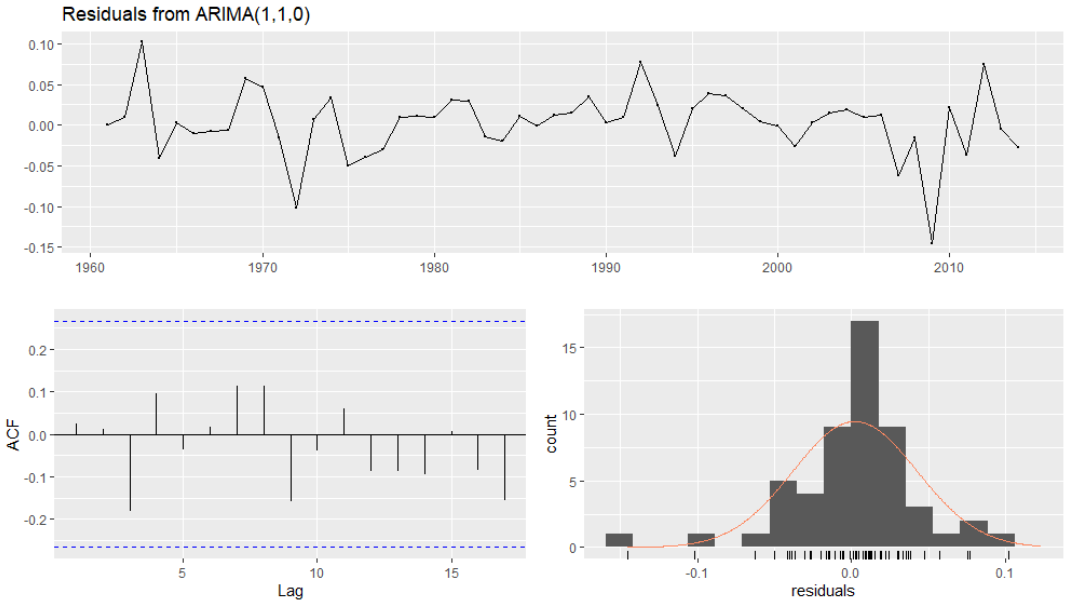
Year	BIO	EFC	EFE	EFI	EFP	Trade	Footprint	C/D
2015	0.4481	1.1296	0.0913	0.1338	1.0910	0.0425	1.1335	-0.6854
2016	0.4462	1.1391	0.0926	0.1382	1.1002	0.0457	1.1459	-0.6997
2017	0.4452	1.1487	0.0940	0.1426	1.1094	0.0486	1.1579	-0.7127
2018	0.4448	1.1582	0.0956	0.1468	1.1186	0.0513	1.1698	-0.7251
2019	0.4446	1.1677	0.0971	0.1510	1.1278	0.0539	1.1816	-0.7371
2020	0.4445	1.1773	0.0986	0.1550	1.1370	0.0564	1.1933	-0.7489
2021	0.4444	1.1868	0.1002	0.1589	1.1462	0.0587	1.2049	-0.7605
2022	0.4444	1.1963	0.1017	0.1627	1.1554	0.0610	1.2164	-0.7720

Table 7

Above is a table of India's forecast from 2015 to 2022; India's biocapacity is forecasted to increase by 0.0037, and its footprint is forecasted to increase by 0.0829 global hectares per capita. India is an ecological debtor; it is expected that its balance will continue to decrease by 0.087 in 2022. India's ecological footprint of production is expected to increase by 0.064; this may be a crucial reason why India's ecological balance worsens.

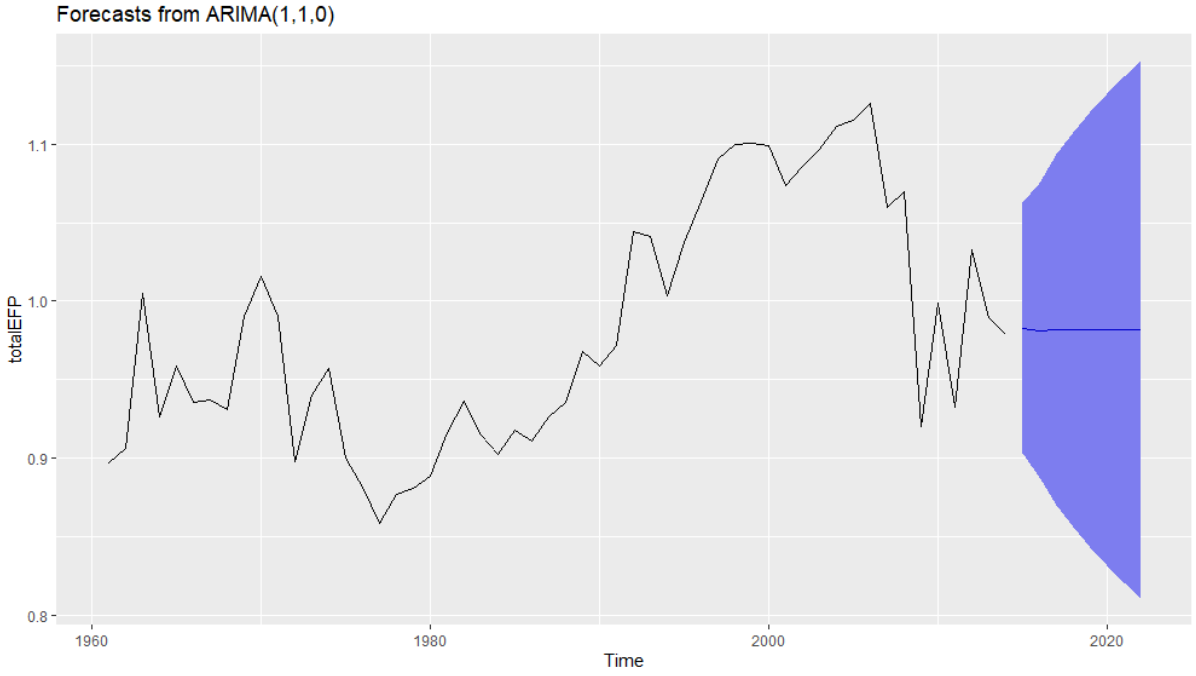
Nigeria

Nigeria is the only African country that has been forecasted for the study; Nigeria was chosen to investigate the effect of having Lagos, the city with the highest population in Africa, on production. Nigeria's largest production resource is its crop and with 0.51, and its second largest resources is forest land with 0.18 global hectares per capita.



Residuals of Production Nigeria

The ARIMA model best suited to forecast Nigeria's ecological footprint of production is ARIMA (1,1,0); this model produces an AIC value of 185.74 and a MAPE value of 2.88%. All spikes are within the significance threshold, which means we can assume that ARIMA (1,1,0) is accurate. The difference time plot shows that Nigeria's production was volatile from 1970 to 1975 and 2005 to 2014.



Time Plot of Production Nigeria

The time plot for Nigeria's ecological footprint of production is highly volatile and shows no trend. The blue section represents the ARIMA forecast which predicts that production will stabilise from 2015 to 2022. The high to low forecast range is significant as the data is volatile and difficult to predict; this can be seen by the large blue area surrounding the point forecast line.

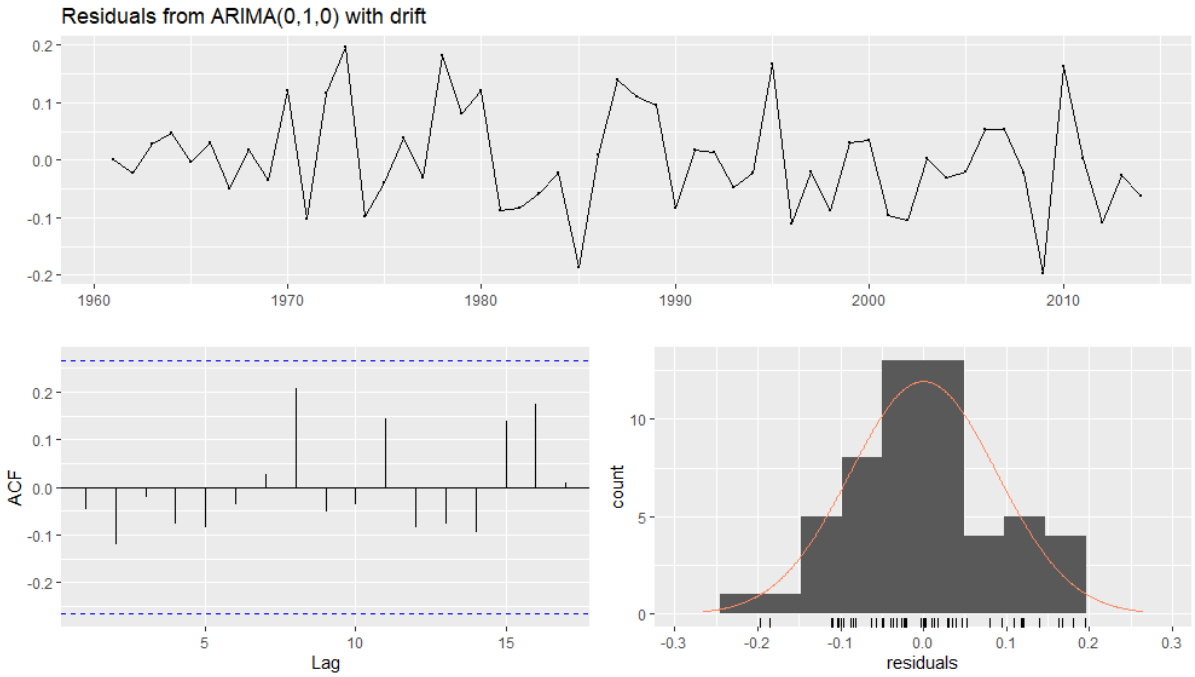
Year	BIO	EFC	EFE	EFI	EFP	Trade	Footprint	C/D
2015	0.7042	1.1153	0.0371	0.1533	0.9829	0.1162	1.0992	-0.3950
2016	0.7042	1.1153	0.0323	0.1533	0.9814	0.1210	1.1023	-0.3982
2017	0.7042	1.1153	0.0347	0.1533	0.9820	0.1186	1.1005	-0.3964
2018	0.7042	1.1153	0.0354	0.1533	0.9817	0.1179	1.0996	-0.3955
2019	0.7042	1.1153	0.0323	0.1533	0.9818	0.1210	1.1028	-0.3987
2020	0.7042	1.1153	0.0363	0.1533	0.9818	0.1171	1.0988	-0.3947
2021	0.7042	1.1153	0.0332	0.1533	0.9818	0.1201	1.1019	-0.3977
2022	0.7042	1.1153	0.0342	0.1533	0.9818	0.1192	1.1009	-0.3968

Table 8

Nigeria's ecological footprint of production decreases from 2015 to 2022 and is forecasted to decline by 0.012 from 2015 to 2022; however, its ecological footprint is forecasted to rise by 0.0017 global hectares per capita. Nigeria is an ecological debtor, but only by -0.3950 in 2015, which is expected to decrease further to -0.3968 in 2022.

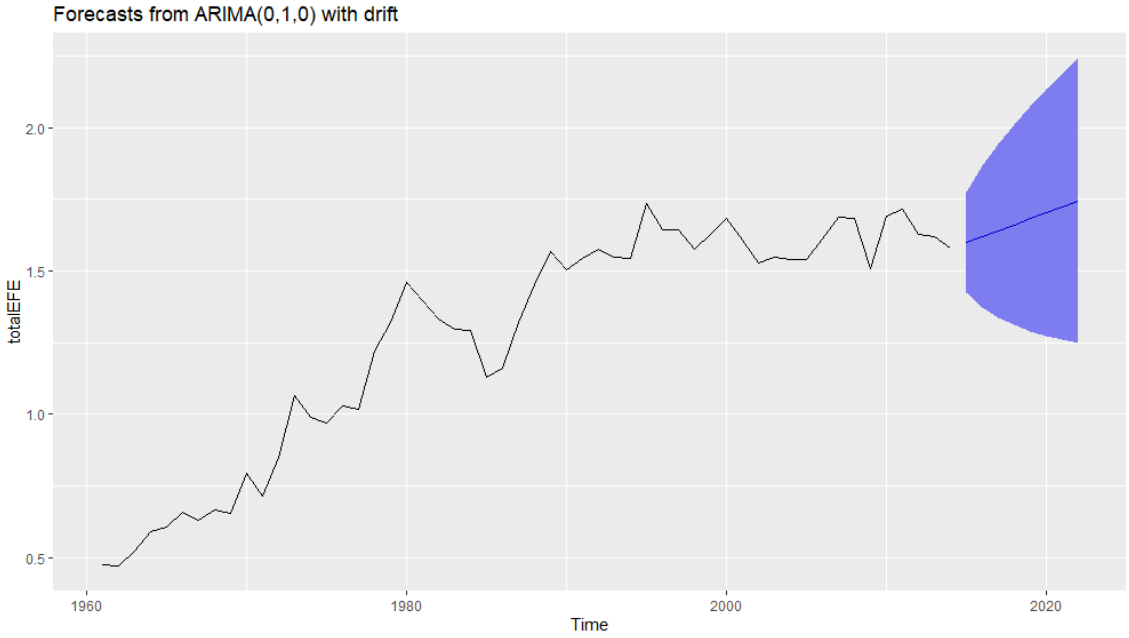
United States

The United States is the second-largest contributor of carbon on the planet, producing 5.41 gigatons in 2018 (Scientists, 2020). The ARIMA model is used to forecast the United States ecological footprint of exports; United States largest resource for exports is its cropland with 0.505 global hectares per capita. The United States most minor resource for exports is its fishing ground, with only 0.042 global hectares per capita.



Residual of Exports Unites States

The ARIMA (0,1,0) model is best suited for forecast the United States exports; this model produced an AIC value of 102.47. The ARIMA (0,1,0) produced a MAPE value of 5.47% and a residual standard deviation of 0.089. The difference time plot shows that the data is volatile, and there are several sudden shifts through the time series. All the spikes in the ACF graph remain within the significance threshold, meaning that the ARIMA model was the best fit to forecast the data.



Time Plot of Exports United States

The time plot above represents United States export data from 1961 to 2014 and the forecast data from 2015 to 2022. The United States ecological footprint of exports has increased by 1.105 from 1961 to 2014, and further increases of 0.146 are forecasted from 2015 to 2022. The time plot demonstrates that exports from 1990 to 2014 were volatile and shows no obvious trend pattern. The forecast shows that the United States is expected to experience an upward trend in exports from 2015 to 2022.

Year	BIO	EFC	EFE	EFI	EFP	Trade	Footprint	C/D
2015	3.4887	8.3661	1.5983	1.6260	8.3377	0.0277	8.3654	-4.8767
2016	3.4321	8.3661	1.6192	1.6462	8.3377	0.0270	8.3647	-4.9326
2017	3.4181	8.3661	1.6400	1.6664	8.3377	0.0263	8.3640	-4.9459
2018	3.4183	8.3661	1.6609	1.6865	8.3377	0.0257	8.3634	-4.9450
2019	3.4002	8.3661	1.6817	1.7067	8.3377	0.0250	8.3627	-4.9625
2020	3.3523	8.3661	1.7026	1.7269	8.3377	0.0243	8.3620	-5.0097
2021	3.3189	8.3661	1.7234	1.7470	8.3377	0.0236	8.3613	-5.0424
2022	3.2977	8.3661	1.7442	1.7672	8.3377	0.0230	8.3606	-5.0629

Table 9

The table above shows all the United States forecasted values; the US biocapacity is expected to decrease by 0.191 from 2015 to 2022. The ecological footprint of imports and exports is forecasted to increase by 0.1412 and 0.1459, respectively. An increase in imports and exports negatively impact countries ecological footprint; however, the United States ecological footprint is forecasted to decrease by 0.0048 global hectares per capita.

6.0 Conclusions

Of the eight countries analysed, three are ecological creditors. Australia, Brazil, and Canada are all forecasted to have ecological balances of 3.4139, 4.8720 and 5.1190 as of 2022, respectively. These countries ecological balances are decreasing. Australia's imports and exports have been increasing since 1961 by 2.026 and 2.578766, respectively. Australia's biocapacity has also decreased by 18.47 since records began. An increase in trade due to high population and demand for foreign goods and a decrease in biocapacity due to deforestation from bush fires will negatively impact Australia's ecological balance.

Brazil's biocapacity also suffers from deforestation; it has decreased by 14.77 global hectares per capita since records began in 1961. Brazil will have a greater ecological balance than Australia by 1.4581 in 2022. Brazil will have a smaller biocapacity than Australia, but it also consumes, trades, and produces a lot less than Australia. Brazil will have a smaller ecological footprint than Australia by 4.1767; these factors all contribute to Brazil's greater ecological balance.

Canada will have the most significant ecological balance of all the creditors in this paper, with a forecasted balance of 5.1190 in the green in 2022. Canada has the most significant amount of biocapacity, trade and consumption of the three creditors. Of the three creditors, Canada also has the most significant ecological footprint; however, it has managed to maintain a better ecological balance than Australia and Brazil due to its massive biocapacity.

Ireland, China, India, Nigeria, and the United States are all ecological debtors meaning that their ecological footprints are more significant than their biocapacity. Of the eight countries selected, Ireland is the only country with an increasing ecological balance. Ireland's biocapacity has decreased by 1.783 since records began, and its imports and exports have increased by 1.93 and 3.20, respectively. Having a decreasing biocapacity and increasing trade should negatively impact Ireland's ecological balance; however, it manages to lower its footprint and balance from 2015 to 2022.

China is the most significant ecological debtors highlighted for analysis. China's consumption, trade, production, and ecological footprint are increasing, as little is done to tackle climate change. China's biocapacity is steady; however, its ecological balance is expected to decline by 0.3701 global hectares per capita from 2015 to 2022. China has a population of 1,400,602,000, and with production and trade increasing, there is little hope that its ecological balance will improve.

India has a similar population as China, with 1,366,000,000 as of 2019; having such a large population makes it challenging to maintain biocapacity and decrease its ecological footprint. Like China, India's consumption, trade, and production are forecasted to increase; however, it has impressively managed to decrease its ecological footprint with a decline of 0.0829 from 2015 to 2022. A decrease in India's footprint is not enough to improve its balance as that is expected to decrease by 0.087 from 2015 to 2022.

Nigeria shares little similarity with India and China; Nigeria's ecological balance is forecasted slightly to increase from 2015 to 2022. Nigeria has always had a small biocapacity with only 1.118 global hectares per capita in 1961, which is 23.5 times smaller than Canada and 20.4 times smaller than Brazil at that time. Considering that Nigeria biocapacity has always been low, it is difficult for them to remain ecological creditors, as its population grows.

Like India, the United States ecological footprint is decreasing but its ecological balance is worsening. Like Ireland, the United States trade is decreasing and is forecasted to fall by 0.0047 from 2015 to 2022. Every country highlight for analysis has steady or decreasing biocapacity

and the United States in no different, biocapacity has decreased there by 1.71 global hectares per capita since records began in 1961.

Like India, the United States ecological footprint is decreasing, but its ecological balance is worsening. Like Ireland, the United States trade decreases and is forecasted to fall by 0.0047 from 2015 to 2022. Every country highlight for analysis has steady or decreasing biocapacity, and the United States is no different; biocapacity has decreased by 1.71 global hectares per capita since records began in 1961.

7.0 Further Development or Research

Given more time to continue researching the national footprint accounts dataset, it would have been possible to analyse a more significant number of countries. Having five countries from each continent analysed would have provided insight into how countries manage their ecological balances with similar resources to their neighbours. Also, having countries analysis separated according to continents would identify which continent has the most biocapacity and has the most significant ecological footprint. With current time restrictions conducting meaningful analysis on more than eight countries would not have been possible.

Another goal that had been hindered by time was attempting to calculate the world's data. The original dataset is complex, with over 84,000 rows, 15 columns and 54 observations for 196 countries. Data from different countries often display no correlation, so combining countries data to analyse and understand would be a strenuous and time-consuming activity. Had this been possible exciting insights and theories could have been explored as to how the world can decrease its ecological footprint.

Selecting a suitable time-series forecasting model that is accurate is time-consuming and had the ARIMA model been tested on the data sooner, perhaps more analyses would have taken place on the highlighted countries.

8.0 References

Borucke, M. *et al.* (2013) 'Accounting for demand and supply of the biosphere's regenerative capacity: The National Footprint Accounts' underlying methodology and framework', *Ecological Indicators*, 24, pp. 518–533. doi: 10.1016/j.ecolind.2012.08.005.

'Brazil and the Amazon Forest' (2019) *Greenpeace USA*. Available at: https://www.greenpeace.org/usa/issues/brazil-and-the-amazon-forest/ (Accessed: 14 May 2021).

Data and Methodology - Global Footprint Network (2021) *Global Footprint Network*. Available at: https://www.footprintnetwork.org/resources/data/ (Accessed: 21 December 2020).

Glen, S. (2016) *RMSE: Root Mean Square Error*, *Statistics How To*. Available at: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/ (Accessed: 14 May 2021).

Maduranga, U. (2020) *Dimensionality Reduction in Data Mining*, *Medium*. Available at: https://towardsdatascience.com/dimensionality-reduction-in-data-mining-f08c734b3001 (Accessed: 22 December 2020).

Sammut, C. and Webb, G. I. (eds) (2010) 'Mean Absolute Error', in *Encyclopedia of Machine Learning*. Boston, MA: Springer US, pp. 652–652. doi: 10.1007/978-0-387-30164-8_525.

Scientists, C. (2020) *Each Country's Share of CO2 Emissions | Union of Concerned Scientists*. Available at: https://www.ucsusa.org/resources/each-countrys-share-co2-emissions (Accessed: 15 May 2021).

Stephanie (2016) *Mean Error: Definition*, *Statistics How To*. Available at: https://www.statisticshowto.com/mean-error/ (Accessed: 14 May 2021).

Stephanie (2019) *Mean Absolute Scaled Error: Definition, Example*, *Statistics How To*. Available at: https://www.statisticshowto.com/mean-absolute-scaled-error/ (Accessed: 14 May 2021).

Stephanie (2021) *Mean Absolute Percentage Error (MAPE)*, *Statistics How To*. Available at: https://www.statisticshowto.com/mean-absolute-percentage-error-mape/ (Accessed: 14 May 2021).

9.0 Appendices

This section should contain information that is supplementary to the main body of the report.

9.1. Project Plan

National College of Ireland

Project Proposal

The International Ecological Footprint Accounts

8 November 2020

BSHTM

Data Analytics

2020/2021

Brian McGrath

x15580167

x15580167@student.ncirl.ie

1.0 Objectives

I have chosen to do my project on the Ecological Footprint of Countries across the globe. When choosing a dataset, I wanted to base the project around the subject of sustainability. I found this dataset interesting as it gives an insight into which countries can naturally produce all the resources they consume and absorb all the waste they generate. My goal is to learn which countries have the lowest ecological footprint and which countries are the fastest improving. As the world becomes more environmentally conscious, countries are likely to reduce their ecological footprint in years to come.

Each unit of measurement within the dataset is recorded using global hectares; the dataset measures each countries biocapacity, with approximately 15,000 data points per country.

Analysing the data should provide results that prove countries are trying to improve the biocapacity; countries can choose to do this by increasing factors such as forest land will, in turn, will decrease cropland which may decrease food production which would negatively affect their biocapacity. Countries will struggle to find the right formula to join the fight against climate change; many factors need to be considered; this study will indicate which countries are outperforming others and what techniques they are deploying.

The dataset is quite large, so that most comparisons will take place between continents rather than countries. For each country and year, the Ecological Footprint of Consumption is recorded, their biocapacity (BiocapPerCap, BiocapTotGHA) and their Ecological Footprint (EFConsPerCap, EFConsTotGHA). Their Ecological Footprint of Production (EFProdPerCap, EFProdTotGHA) and their trade imports (EFImportsPerCap, EFImportsTotGHA) and exports (EFExportsPerCap, EFExportsTotGHA) are also recorded in order to be able to calculate whether not that country is an Ecological debtor or an Ecological creditor (Data and Methodology - Global Footprint Network, 2021)

2.0 Background

I plan to centre my project around the data collected on The Global Ecological footprint per capita of every country from 1961 – 2014 in global hectares. This dataset captures data from 196 countries. The ecological footprint is the measure of how much area of biologically productive land and water an individual requires to naturally produce all the resources they consume and absorb the waste they generate (Kriebel, 2018). Countries are actively trying to lower their Ecological Footprint through international committees like the UN and environmental activists worldwide. Having a lower ecological footprint means that countries are less reliant on foreign produce and increase domestic suppliers' business. Countries with lower footprints are less affected by global disasters, similarly to Covid-19. This pandemic has taught the world how vital it is to source your produce locally as international trade breaks down; Covid-19 has changed governments' perspectives worldwide, which will be represented in the data over the next few years.

My dataset was created two years ago by Andy Kriebel, it contains twelve columns, and they are as follows: country, ISO Alpha-3 code (which abbreviates countries into three letters), UN_region, UN_subregion, year, record, crop_land, grazing_land, forest_land, fishing_ground, built_up_land, carbon, total, per capita GDP, and population. These columns excellently cover the bases needed to examine countries ecological footprint.

I have sourced my data from a website called Kaggle, and the dataset is accessible to the public. The dataset calculations are based on United Nations data, including the UN Commodity Trade Statistics Database, the UN statistics division, the International Energy Agency, and the Food and Agriculture Organisation (Kriebel, 2018).

I will examine the ecological footprints of countries worldwide and conduct predictive analysis to assess which country will reach their biocapacity sooner. This project can help ecologists and governments around the globe to implement new sustainability measures.

3.0 Technical Approach

The knowledge discovery in databases methodology will be incorporated throughout this project. A strict methodology or plan must be followed to derive any knowledge from the data (Peng, Yang and Ren, 2009). KDD is a popular methodology chosen for many projects as civilisation proceeds into the age of digital transformation; data overload is inevitable; KDD gives us the ability to analyse, extract, and understand massive volumes of data.

My first objective was to source an accessible dataset that matched my interests and is suitable for the knowledge discovery process. I was careful to choose a dataset that allowed me to clean, analyse and predict outcomes.

My next objective is the study the dataset as there are components and columns that I need to grasp a better understanding of before I can move to the following stages; having an in-depth knowledge of the data will help me with various decisions during the early stages of the project. Also, extracting as much knowledge as possible will allow me to make more accurate predictions on how the data might transform in the years to come.

After analysing, the next step would be to begin cleansing the data. To improve data reliability, I will have to incorporate statistical techniques and possibly some data mining algorithms. My goal is to handle any missing data by creating a prediction model that identifies and predicts any missing data. By cleansing the data, I will be able to remove any attributes which lack reliability.

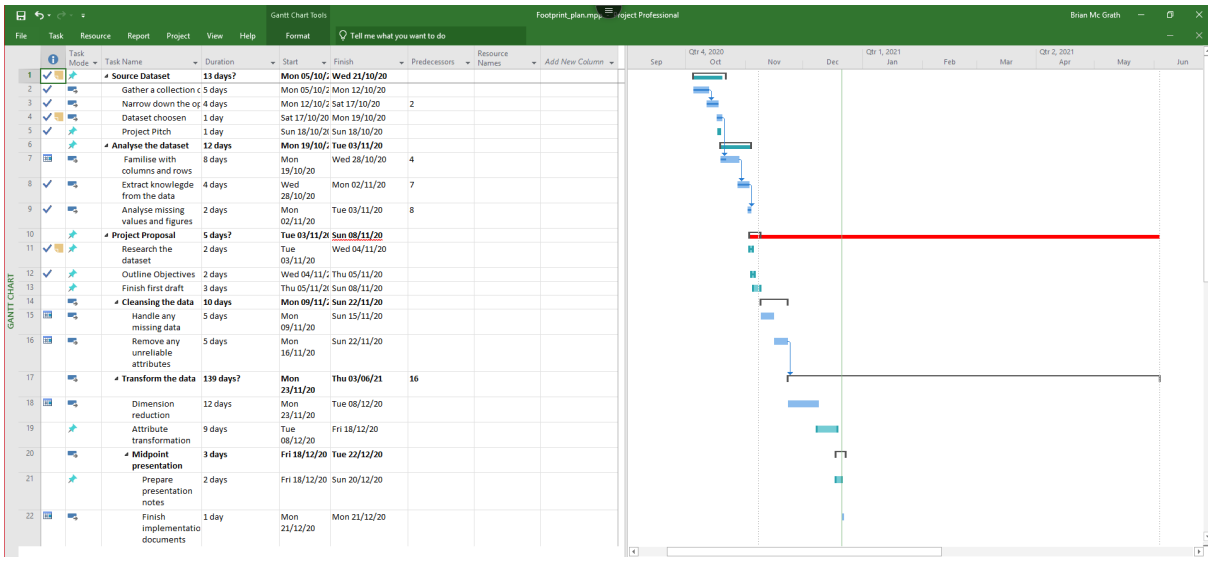
At this stage, my data should be ready for dimension reduction and attribute transformation. This process should lead me to conclude which type of data mining to use (regression, clustering).

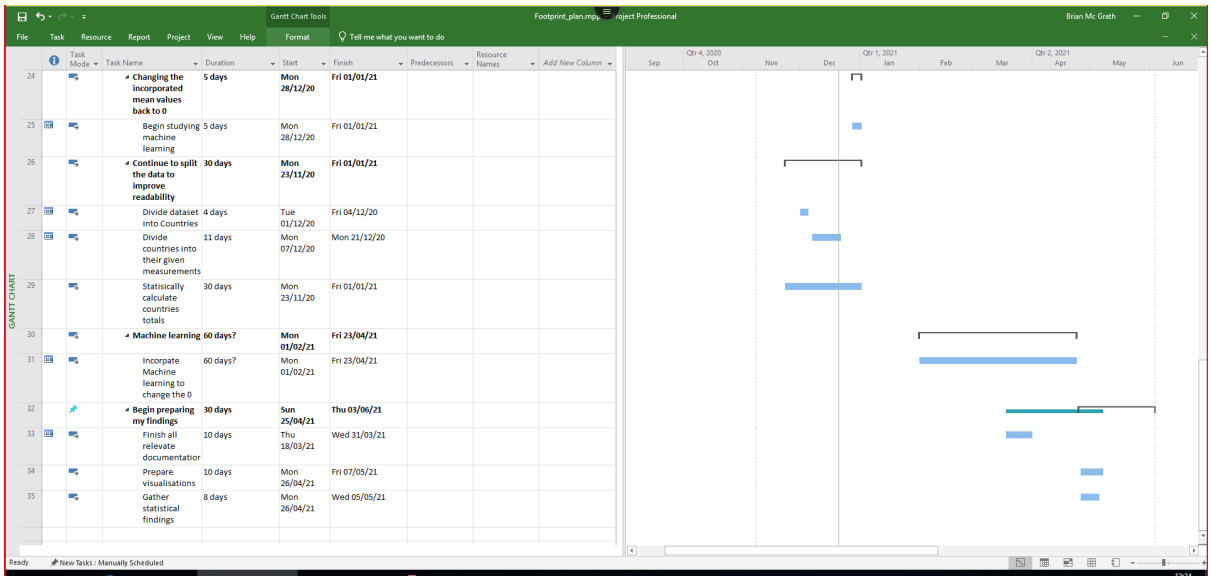
4.0 Special Resources Required

This study does not require any special hardware or books to complete. The programming language R and Python will be taught throughout the course, and the year and machine learning will be introduced during our second semester; other than that, the skills required to complete this project have been taught and put into practice in previous projects.

Project Plan

The project plan is a Gantt Chart for this study, which was created using MS Project; I have added all the processes used throughout the project. The Gantt chart depicts the timeline in which these tasks will be completed; this is not a finished product and is subject to change as the project's progression is made.





5.0 Technical Details

The majority of the work carried out on the data will be done using R in RStudio. RStudio will be used to pre-process the data, get rid of any missing values, change the data to their correct data types and rename any variables where necessary. Once a new clean data set has been created, the data can then be exported locally and ready to use software such as SPSS, Excel, Python and Tableau.

Phyton is the programming language that will be taught after reading week so that Python will be incorporated into the project during the latter stages of the semester. The dataset will be downloaded as a CSV file and imported from Excel into R Studio. R Studio will be used to pre-process and cleanse the data, including changing row and column names if necessary, changing the data to the correct data type to carry out statistical analysis and identifying and removing NA values. The dataset chosen has a massive 157,910 NA values; once omitted, the dataset will be cut from 87,020 rows to 54,330 rows. An obstacle that will be encountered after removing the NA values is to alter the zero values to their country's mean values; the dataset contains 36,122 zero values. If the zero values are altered to NA values and then omitted from the dataset, there will only be 21,144 rows left to analyse, which would comprise the study results.

Excel may be used to divide the dataset into tables, one for each countries data. That should allow for the mean to be calculated for each numeric column and transferred to a separate CSV file; once the file is complete, it can be imported into R Studio to alter the zero values; this method is subject to change as more options are explored. The dataset contains tons of duplicate row names with the Country, ISO alpha-3, UN_region, UN_subregion and year columns; this may make analysing in R studio difficult.

6.0 Evaluation

The dataset chosen will be broken down using excel and R Studio; once the data is cleaned and separated, analysis can begin using SPSS and Python. Separating the data will allow for more accurate results, and countries can then be pinned against one another. Transformation of the data will be done using all software's Excel, R Studio, Python, SPSS, Rapid Miner and Tableau for visualisations.

Time series analysis will be used to fill in all 0 values in the dataset; once the 0 values are replaced, the final analysis can begin as the finding will be more accurate.

7.0 Bibliography

Data and Methodology - Global Footprint Network (no date) *Global Footprint Network*. Available at: https://www.footprintnetwork.org/resources/data/ (Accessed: 21 December 2020).

Peng, Z., Yang, B. and Ren, H. (2009) 'Research on KDD Process Model and an Improved Algorithm', in *2009 International Joint Conference on Artificial Intelligence*. *2009 International Joint Conference on Artificial Intelligence*, pp. 113–115. doi: 10.1109/JCAI.2009.15.

Kriebel, A., 2018. *2018/W17: Ecological-Footprint-Per-Capita*. [online] Data.World. Available at: <https://data.world/brianmg96/software-project/workspace/project-summary?agentid=makeovermonday&datasetid=2018w17-ecological-footprint-per-capita> [Accessed 7th November 2020].

www.javatpoint.com. 2020. *KDD Process In Data Mining - Javatpoint*. [online] Available at: <https://www.javatpoint.com/kdd-process-in-data-mining> [Accessed 8 November 2020].

7.1. Reflective Journals

Reflective Journal of October

Brian McGrath

X15580167

I plan to centre my project around the data collected on every country's Global Ecological footprint over the last 50 years. I have sourced my data from a website called Kaggle. I will examine the ecological footprints of countries worldwide and conduct predictive analysis to assess which country will reach their biocapacity sooner. This project can help ecologists and governments around the globe to implement new sustainability measures.

I have researched the dataset and have 50% of my project proposal complete.

The month of November will be spent working on the dataset to learn ways to manipulate the data through a suitable IDE.

Reflective Journal of November

Brian McGrath

X15580167

The month of November has been hectic in assignment submissions, which has made it challenging to get properly stuck into my software project.

I have managed to upload my data into R studio, and start the pre-processing steps, removed all the missing values from the dataset to prepare to begin to make some visualisations.

I have installed SPSS and Tableau and have begun to teach myself to use the software before I upload my dataset.

I will focus on my documentation in the coming weeks considering December will be less hectic.

We will be taught to code in Python in the coming weeks, so I hope that I will incorporate some Python into my mid-point presentation.

Reflective Journal December

Brian McGrath

X15580167

December began by replacing all the 0 values in the dataset with their mean values, this was done using the IF(function in Excel and took a considerable amount of time.

Once the mean values were placed in the dataset they were divided into continents, the continents were then uploaded to RStudio, Tableau, and SPSS.

The first four sections of the report were finished for the mid-point presentation, this included Executive Summary, Introduction, Data, Methodology and Analysis.

Canva was then used to create a slideshow for the video presentation.

Reflective Journal January

Brian McGrath

X15580167

The month of January has been hectic in terms of assignment submissions that have been put in place instead of exams.

Once the exams had finished, I had the opportunity to review my mid-point submission and plan how to move forward.

I have discovered a few mistakes within the dataset that I have begun to rectify. My next challenge is to prepare my dataset to be used in new software such as Tableau and Rapid Miner

We will be taught Data mining and machine learning this semester, so I hope that I will be able to incorporate these techniques over the coming months.

Reflective Journal February

Brian McGrath

X15580167

The month of February was primarily spent researching time series analysis techniques; some time was spent conducting the prophet model on some sample data.

It is still unknown how the data will be divided so that meaningful analysis can be conducted. The visualisations being outputted from the original dataset are too dense and impossible to derive any useful information.

A few aspects of the report needed to be shuffled around as they were in the incorrect locations, and research into which countries would be most interesting to analyse has also begun.

Reflective Journal March

Brian McGrath

X15580167

The month of March has been hectic in terms of assignments for Advanced Business Data Analysis and Data Mining.

March was spent dividing my data into tables for each of the of 192 countries in order to conduct time series analysis.

The last couple of weeks I have been exploring different time series analysis techniques to find the best suitable option.

The last week of March was spent planning how to properly report my findings.

April will be spent learning how to incorporate Phyton into my project.

Reflective Journal April
Brian McGrath

X15580167

The beginning of April was spent finishing the remainder of CA's and TABA's that needed to be completed.

Countries were being prepared for time-series analysis. First, the data would be divided according to country; each country contains ten excel worksheets, five for global hectares per capita and five for global hectares.

Time-series analysis cannot be conducted on every country as the workload would be too large, and the result would be impossible to squeeze into the final report; it was decided to conduct an in-depth analysis of ten countries.

Once the selected countries had been separated from the original dataset, it was time to upload them into RStudio; only the global hectares per capita worksheet were selected to be uploaded.

After testing multiple models such as AR, MA and ARMA, it was clear that the ARIMA model was outputting the most accurate forecasts. ARIMA will be the time-series model chosen the analyse all the data for this project.

7.2. Other materials used

Any other reference material used in the project for example evaluation surveys etc.