Climate Change: An Investigation into Average Surface Temperature and Development Indicators by Country

James Hooper

I. INTRODUCTION

A. Domain: Climate Change

Climate Change is one of the biggest problems our society is facing in the 21st Century. The evidence seems conclusive that a drastic increase in the emissions of greenhouse gases is causing an unprecedented increase in the rate of global warming. The knock on effects of this warming are ice caps melting, rising sea levels and an increase in extreme weather events [1].

The most current example is the California wildfires, with 'nearly 200,000' resident having to be evacuated from their homes according to the BBC [2]. This is only one of an increasing number of devastating natural disasters linked to Climate Change. This years hurricane season has been the worst in recent memory with Hurricane Irma killing 134 people, destroying 90 percent of buildings in Barbuda and displacing roughly 6.5 million people in Florida alone [3]. Though Irma is not the largest hurricane in living memory (at a Category 4), the increased frequency of hurricanes is more troubling with Hurricane Maria and Harvey causing large amounts of destruction and taking a reported combined total of 143 although the actual number may be higher. This begs the question: What is being done to combat Climate Change?

The initial answer to this is the Paris Climate Agreement. The essence of the agreement is to limit warming from pre-industrial temperatures to below 2 degrees Celsius with a further effort to limit this warming to 1.5 degrees Celsius [4]. This agreement seems to ensure international cooperation when it comes to Climate Change and though there is no way to know if this goal will be achieved, it is at least a goal that all countries are working towards. With the recent signatures of Nicaragua and Syria, the only country left is the United States after Donald Trump opted to remove it from the agreement.

This leads to a further question: Is Climate Change natural or affected by the human race? Donald Trump appears to believe that CO2 emissions are not an important reason for the rise in Global Temperature. There has been enough work in this area to comfortably believe that rising CO2 emissions are affecting temperatures and that there is a man made component to Global Warming [5]. The Intergovernmental Panel on Climate Change (IPCC) forecast an increase of between roughly 1.4 to 5.5 degrees Celsius in the next century. The questions this report looks to answer are more focused on how countries are going to be able to achieve a maximum warming of 2 degrees Celsius:

- How will the temperature rise be kept to the lower boundary set out by the IPCC?
- What is a Country's warming profile and how is it forecast?
- How are Development Indicators affecting a Country's rate of warming?

B. Data

This analysis uses two datasets:

- The first is a collection of Time Series, which refer to Average Surface Temperatures, from http://berkeleyearth.org/data/. This dataset has a Time Series of Average Surface Temperatures in degrees Celsius for each country. There is a measurement on the first of each month starting in 1750, through to the final measurement on 1st August 2013.
- The second dataset that has been utilised, which contains Development Indicators by Country and Year, is from https://www.kaggle.com/worldbank/world-development-indicators/data. It has indicators for each country from 1960 to 2014. There are 1327 unique indicators, which can be roughly grouped into these categories:
  1) Social Attributes (e.g. Urban Population (%) and Rural Population (%)).
  2) Environmental Attributes (e.g. CO2 emissions (kt) and Electricity Production from Renewable Sources (kWh)).
  3) Economic Attributes (e.g. GDP growth (annual %) and Gross National Expenditure (% of GDP).

C. Initial Analysis Plan

After a number of steps to transform the data, which will be described in Data Pre-Processing, there will be a Time Series created for each Country that shows its Average Surface Temperature ending at 2013. The date that the Time Series commences depends on the completeness of the dataset. By utilising this data an initial analysis plan can be formed:

1) Calculate different profiles of temperature increase by clustering the Time Series into groups of similar profiles (the number of groups will depend on whether an increase in groups provides useful information).
2) Merge Cluster data to that of the Development Indicators to try and develop hypotheses for why each Country is in each Cluster.

3) Using Principle Component Analysis (PCA) the number of components in this dataset can be reduced and the Principle Components can be used in a Machine Learning model to predict which Cluster a Country will be in the following year depending on changes to its Development Indicators.
4) Using an ARIMA model, the warming profile of each cluster can be forecast to determine the proximity to the target of 2 degrees Celsius for each Cluster.

II. ANALYTICAL PROCESS

A. Data Pre-Processing

An initial review of Average Surface Temperature (AST) Time Series by Country shows that a majority of the countries have no data pre 1850 and therefore the data utilised for this exercise was restricted to the period from 1850 to 2013. After plotting a significant number of these AST Time Series, it became clear that temperature measurements before 1900 are quite volatile and seem to have a larger variance than should be reasonably expected. This could be explained by some variables such as the change from Mercury to Electronic Sensors, weather stations being moved and the time of day measurements are taken being changed from afternoon to morning [6]. Therefore each Time Series was further restricted from 1900 to 2013, thereby creating a more consistent dataset to analyse.

Now the data needs to be transformed so that useful insight can be gained from it and so it can answer the questions set out in the Introduction. This starts with filling in any missing data that remains. For this a pandas function was used to interpolate with a forward fill, which, since there were very few long gaps, seemed to best represent any seasonal trends in the data. For this analysis a change in temperature from 1900 seemed most appropriate so the temperatures from 1900 were re-based to show a change in temperature from 0 degrees Celsius. From there the data was smoothed using a rolling mean of 5 years worth of data, which made the general temperature trends more easy to identify as the seasonal trends were largely removed.

B. Time Series Clustering

To cluster the Time Series a package in Python called tslearn was used. The method used to cluster the Time Series was K-means clustering using a Euclidean distance measure for simplicity, though it is also supposed to be an effective distance measure for such Time Series [7]. The only hyperparameter the user can choose is the number of clusters. Once the number of clusters is chosen, the algorithm groups together time series that have a similar warming profile around a central warming profile for each cluster. In this case 6 clusters were chosen as each cluster is suitably unique. Using more than 6 clusters did not create any further insight.

Initially before the data was transformed to a change in temperature from 1900 the clusters were grouped on the temperature of the country rather than the actual profile of its warming, which prompted a transformation to temperature change. In Fig. 1, the six different clusters are on different graphs. The central warming profile for each cluster is the coloured Time Series on each graph. The black Time Series around the central profile are the Country's Time Series associated with that Cluster. To see where these Clusters appear on a world map, the colours of the central profiles in Fig. 1 matches the colour of each country in Fig. 2 (Tableau was used for this visualisation).

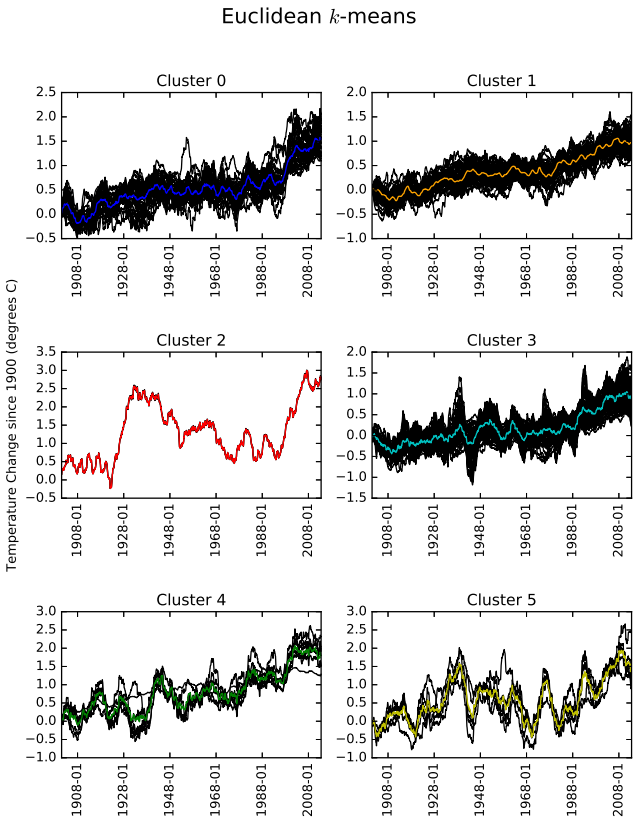

Fig. 1. Time Series Clusters

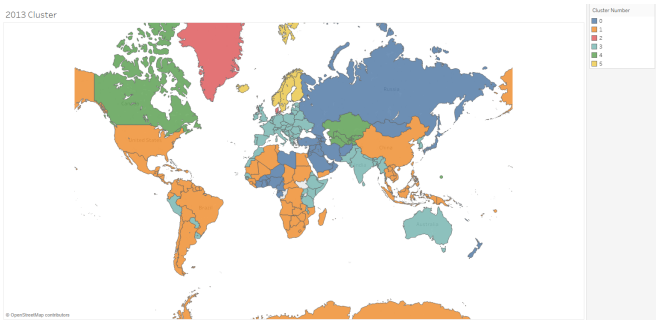

Fig. 2. Clusters Shown on World Map

C. PCA and Random Forests

At this point the Development Indicators dataset is used to gain more information about the clusters. Data was taken for each year from 2009 to 2014 for each country with attributes like Urban Population (%), C02 emissions (kt) and Access

to Electricity (%). If any attributes had missing values the previous year's data was utilised, assuming that over 1 year it was unlikely to change. However, if any attributes were missing more than 50% of entries, they were removed as being unreliable. This left a dataset with 900 rows and 539 numerical attributes to which the cluster number was merged as a classifier.

PCA produces a number of Principle Components, chosen by the user, which are a weighted linear combination of the attributes taken by the algorithm. It seemed appropriate to use PCA since it would help to simplify this large dataset and produce a more useful set of attributes for the Random Forest model that will be used to predict a Country's Cluster. Twenty Principle Components were used for the PCA as the explained variance ratio of the Principle Components dropped below 1 percent after the 20th component and it proved to provide only a marginal gain in information beyond this point. Using PCA before a Random Forest Classifier has the benefit of removing collinearity between attributes and also reducing the number of features to reduce the likelihood of overfitting. This will help to improve the accuracy of the Random Forest model.

A Random Forest Classifier was used in an attempt to see how a country's warming profile might move from one Cluster to another depending on its Indicators that year. Data from 2009-2013 was taken for training and testing, whilst data from 2014 was kept separate in order to predict what cluster a country will belong to. A Random Search Algorithm from sklearn.model_selection to find the most accurate parameter selection for the random forest, which had Number of Trees as 60 and Predictors to Sample as 1. A Random Search Algorithm was chosen as it is a more computationally efficient algorithm than other hyperparamater searches such as a Grid Search. After using 80 percent of the data from 2009-2013 to train the model, it was tested on the final 20 percent with an accuracy of 86.7%. With that accuracy score, it was appropriate to use the model to predict which cluster countries would be in in 2014 using the 2014 indicators that were kept separate previously. This led to Fig. 3 (Tableau was used for this visualisation), which shows the Clusters the Random Forest model has predicted each country will be in in 2014, given 2014's Development Indicators. The Cluster colours in Fig. 3 match the Figures that have come before.

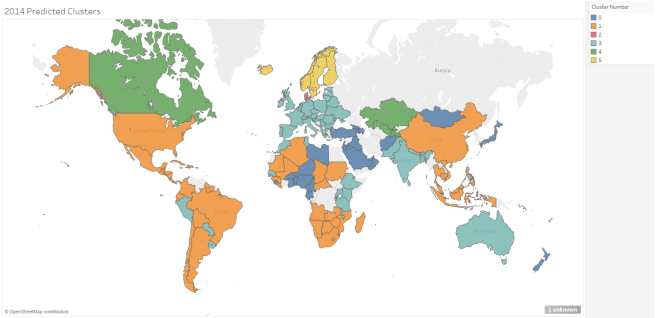

Fig. 3. 2014 Predicted Clusters

D. ARIMA Forecasting

The model can predict which warming profile a country belongs to depending on a years worth of Development Indicators and once this is known the next stage is to forecast how this countries temperature could change the following year. To do this an Autoregressive Integrated Moving Average Model (ARIMA) was used, which is good at forecasting Time Series that have stationarity. The centres of each Cluster will be used for modelling, since they do obey stationarity after they have been transformed to temperature change, whereas they did not obey stationarity when the value of the Time Series was average temperatures.

To train each ARIMA model data from 2004 - 2012 from each Cluster Centre's Time Series was used. ARIMA models have 3 hyperparameters to optimise the model (p, d, q). Again a Random Search Algorithm was used to calculate the Mean Absolute Percentage Error (MAPE), one of the most effective accuracy measures of an ARIMA forecast [8], in order to find the optimum parameters. This MAPE was calculated by testing the trained model on each Cluster Centre's Time Series from 2013, comparing the predicted warming profile against the actual warming profile. The ARIMA model with the lowest MAPE was ARIMA(12, 0, 1) at 1.81%. Therefore this model was used to forecast 12 months for each centre, which can be seen below in Fig. 4. Like before the colours in Fig. 4 match colours in previous Figures.

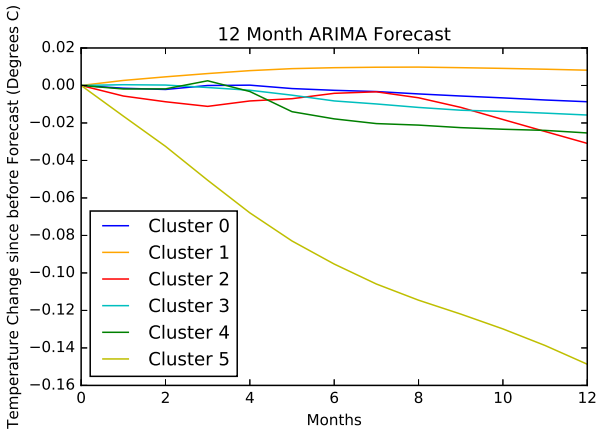

Fig. 4. 12 Month ARIMA Forecast of Clusters

III. FINDINGS AND REFLECTIONS

A. Cluster Analysis

Clusters		
Cluster No.	Major Countries in Cluster	Cluster Centre Temperature Change since 1900
0	Afghanistan, Cameroon, Russia	1.519 degrees Celsius
1	USA, China, Brazil	1.008 degrees Celsius
2	Denmark, Greenland	2.609 degrees Celsius
3	UK, India, Australia	0.912 degrees Celsius
4	Canada, Kazakhstan	1.799 degrees Celsius
5	Finland, Norway, Sweden	1.471 degrees Celsius

No countries were predicted to be in a different Cluster in 2014 compared to the actual Cluster in 2013. In one year the Development Indicators are not likely to vary enough for a country to be classed in a different Cluster. For this reason, utilizing the Clusters from Fig. 1 and Geography from Fig. 2, analysing Clusters in 2013 provides more insight.

In the table above, the Cluster Number has been colour coded so it is more easy to refer back to previous figures. The table provides two pieces of key information about each Cluster, the Major Countries belonging to each Cluster and the Cluster Centres Temperature Change since 1900. Some key points from the table are:

- Cluster 3 contains UK, India and Australia. This illustrates that the temperature change is not solely influenced by Geography.
- Cluster 5 contains Finland, Norway and Sweden. Whilst Geography is not the only reason Countries are clustered together, it is clear that Geography has an effect on a Country's warming profile.
- Clusters 1 and 3 cover a vast percentage of the world and contain the majority of the major world powers (USA, UK, China), see Fig. 2. These Clusters have the two smallest temperature increases at 1.008 and 0.912 degrees Celsius.

If this analysis is continually updated with new data each year, the Random Forest model will start to make predictions that vary from a Country's initial placement in a Cluster. Whilst a Country's Development Indicators are unlikely to vary dramatically year to year, over a span of decades Countries will have had more chance to enact policies to limit Global Warming and there are likely to be Countries that are classed in different Clusters than they were initially.

The ARIMA model was able to accurately represent a years worth of actual data using 2013 as its test. If these ARIMA models are updated as a month of new temperature data is introduced, a rolling 12 month forecast will be available for each Cluster. This will give each Country an effective way of tracking how close they will be to the temperature increase limit of 2 degrees Celsius 12 months into the future. The forecast for 2014, which can be seen in Fig. 4, indicates that only one warming profile has a forecast increase in temperature. However this forecast is for Cluster 1, which has the largest number of countries classed to it. It is worth noting that an ARIMA model forecast becomes much less accurate with more time, therefore these forecasts have been limited to 12 months for this purpose.

B. Future Work

An element of analysis that will give even more insight into how a Country could be classified in a different Cluster is extracting the importance of each Principle Component. Looking at what linear combination of attributes from the Development Indicators constitutes the most important Principle Components, the Development Indicators that have the strongest effect on a Country's Cluster can be extracted. Countries will then know what Development Indicators they need to be focused on when planning environmental policy.

C. Final Remarks

The Paris Agreement describes the target of limiting Global Warming by 2 degrees Celsius as 'ambitious' [4]. Being able to track and forecast the progress of each country is going to be of vital importance if staying below the 2 degree Celsius target is going to be achieved. The analysis that has been carried out in this report combined with a specialist knowledge in the field of environmental sciences could help shape policy for all members of the Paris Climate Agreement.

REFERENCES

[1] Global Climate Change: Effects. NASA, NASA, 3 Aug. 2017, climate.nasa.gov/effects/.

[2] California wildfires: Nearly 200,000 flee as new blaze spreads. BBC News, BBC, 8 Dec. 2017, www.bbc.co.uk/news/world-us-canada-42263237.

[3] Keneally, Meghan. Hurricane Irma: By the numbers. ABC News, ABC News Network, abcnews.go.com/US/hurricane-irma-numbers/story?id=49677062.

[4] United Nations Framework Convention on Climate Change. Status of ratification.The Paris Agreement - main page, 12 Oct. 2017, unfccc.int/paris_agreement/items/9485.php.

[5] Climate change evidence: How do we know? NASA, NASA, 10 Aug. 2017, climate.nasa.gov/evidence/.

[6] Hausfather, Z., Cowtan, K., Menne, M. J., & Williams, C. N. (2016). Evaluating the impact of U.S. Historical Climatology Network homogenization using the U.S. Climate Reference Network. Geophysical Research Letters, 43(4), 1695-1701. doi:10.1002/2015gl067640

[7] Kalpakis, K., Gada, D., & Puttagunta, V. (n.d.). Distance measures for effective clustering of ARIMA time-series. Proceedings 2001 IEEE International Conference on Data Mining. doi:10.1109/icdm.2001.989529

[8] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679-688. doi:10.1016/j.ijforecast.2006.03.001