Analysis of Climate Change and its Potential Impacts

Akash Iyengar
Computer Science
University of Colorado,
Boulder, United States
akiy0076@colorado.edu

Keerthika Rajvel
Computer Science
University of Colorado,
Boulder, United States
kera5806@colorado.edu

Swarnalatha Natarajan
Computer Science
University of Colorado,
Boulder, United States
swna2675@colorado.edu

Vandana Sridhar
Computer Science
University of Colorado
Boulder, United States
vasr6141@colorado.edu

ABSTRACT

Countries all over the world experience several climate change impacts such as floods, droughts, heat waves and hurricanes. Most of these natural disasters occur as a result of Global Warming. Carbon emissions affect around seven million people per year on average as a result of air pollution. As per the statistics presented by NOAA,the percentage of green houses gases in the atmosphere have gone up by 41%. As a result of greenhouse emissions, heat trapped in the atmosphere results in reduction of atmospheric moisture which causes droughts. In turn droughts affect the crop yields in countries. Another contributing factor that demands the supply of crops is population.With population explosion in countries, the supply demand problem is brought into focus. This project analyzed the changes in climatic conditions over time by investigating several parameters such as greenhouse gas emission rates, the SPEI index for droughts and the correlation coefficients between population and crop yields etc. Our focus was to analyze the trend of carbon footprint from the years 1990 to 2009 and predict the footprint values for the next 17 years for major carbon emission contributing countries. Similarly we utilized the SPEI to predict future drought occurrences. Finally, we analyzed the correlation between crop yields and the population changes for a span of 19 years using several statistical correlation parameters. We evaluated our time series model using the Mann Kendall monotonicity test, divided our dataset into a train test split and calculated the RMSE value for the difference in predictions performed by the model. Additionally, we compared our graphs with the models charted by World Bank. Based on the results obtained, appropriate recommendations are provided to mitigate the impacts of these factors.

INTRODUCTION

Climate change is a global phenomenon affecting the lives and livelihoods of not just humans but every single organism on the planet. Earth's climate has changed throughout history; however, the pace of change has exponentially increased in the last decade or so with evidence from NASA showing ice fronts retreating, rocky peaks becoming more exposed, resulting in lesser icebergs drifting in the oceans[1]. Scientists have overwhelmingly attributed climate change (over 95% probability)[2] to human activities. Humans have collectively destroyed several ecosystems resulting in an unprecedented increase in greenhouse gases that tend to trap the sun's heat, making the planet warmer. All this has resulted in global temperature rise, warming oceans, shrinking ice sheets, glacial retreat, decreased snow cover, sea level rise, extreme climatic events, ocean acidification, etc. The call of the hour is to increase awareness and tackle this on a global scale. So a system that is able to identify the primary/underlying factors that leads to climate change, extent of effect of each factor and also pinpoint to which regions are most severely affected as well as estimate the future would help take timely actions. In addition to this, identifying which countries/regions led to most degradation for academic purposes could help with framing country/region specific strategies to tackle this problem head on.

Most of the research done in this area consists of analysing spatial temporal data to estimate rainfall patterns, satellite data to evaluate the percentage of pollutants available in the atmosphere and studying soil samples to detect the level of moisture content.

Our approach involves analysis of various parameters that contribute to climate change. As mentioned in our abstract, we shed light on three factors namely: Carbon footprint, Drought and agricultural yield. Most of our work involves analysing the trend of the aforementioned factors. Trend analysis in our project provided valuable evidence on the environmental aspects of several countries that significantly contributed to global warming over the course of two decades. Based on the current trend, we estimated the future projections in terms of carbon footprint and drought occurrences. Additionally we studied other areas that are affected through the above factors namely: agricultural yield and population.

Our project was split into three components namely: calculating the carbon footprint of each country and building a time series model that predicts the carbon footprint value for future years, Drought prediction using SPEI index and analysing the correlation between population and agricultural yield for a filtered list of countries.

This report first discusses the abstract and the introduction to our project. Next we move on to the literature survey that highlights the related work done in this area. Third, we discuss the drawbacks of the related work and how our work differs from the work previously done. Next we move on to describing our dataset. We then discuss our approaches at length by stating the steps we undertook into each method. We then present our findings and how we evaluated our model. Finally we provide suitable recommendations and discussions about what can be furthered upon.

RELATED WORK

1.**Data Mining for Weather and Climate Studies**

K C Gouda and Chandrika M [3] propose an approach to provide a better understanding of the monsoon rainfall over India using spatio-temporal data mining at different scale i.e daily to decade. The aim of their work is to perform Spatial and Spatio-temporal data mining (SSTDM) on climate data.

In this work both Classification and Clustering were used to analyse data gathered from India over a period of 53 years (1951 - 2003), in order to develop classification rules for the weather parameters over the study period using available historical data. The targets for the prediction are those weather changes that affect us daily like changes in minimum and maximum temperature, rainfall, evaporation and wind speed.

Their developed system consists of several modules enlisted below :

Module 1: Fetching the satellite data, climate model output, gridded data and converting into system understandable format (like ASCII) .

Module 2: Providing the converted data as an input to the HPC system(data mining system).

Module 3: Grouping similar output data of different regions using clustering technique.

Module 4: Graphical representation of radical data.

The Clustering module has been implemented using K-means clustering algorithm with K set to 4. On clustering, Cluster 0 represented the largest amount of rain, lower temperatures, moderate humidity and faster wind speed. Hence, representing the Rainy season. The distribution of this cluster includes the monthly analysis over the periods June, July, August, and September. Cluster 1 represented the least amount of rain, higher temperatures, higher humidity and slower wind speed. Thereby representing the Summer season.The distributions of this cluster include: the end of December, January, February, March and April. Similarly, Cluster 2 and 3 represented Autumn and Spring seasons respectively.

K C Gouda and Chandrika M[3] believe that this understanding of the seasons based on rainfall is very important to many sectors as well as many industries which largely depend on weather conditions such as agriculture, vegetation, water resources and tourism.

2.**Data Mining for Climate Change and Impacts**

Auroop R Ganguly and Karsten Steinhaeuser's work [4] aims at mapping climate requirements to solutions available in temporal, spatial and spatiotemporal data mining.

The primary contributions of their work are three-fold: (a) introduction of climate challenges to the data mining (specifically the SSTDM community to motivate future research, (b) defining the climate data mining problem by comparing and contrasting with SSTDM, and (c) presenting a case study to demonstrate how even simple data mining applications can lead to novel insights in climate science.

This study has used data from multiple disparate sources. Climate projections have been based on IPCC SRES A1FI, the worst-case fossil fuel intensive scenario which nonetheless has started to look credible in recent years with increased trends in temperature observations. The model simulations have been performed at ORNL and NCAR and output data are available through the Earth System Grid (ESG). The observation data have been compiled by the National Centers for Environmental Prediction (NCEP) and are available for download on the NOAA Earth System Research Laboratory (ESRL) website. Population projections for Europe are based on the IPCC SRES A1FI population, followed by downscaling to country-levels by the Center for International Earth Science Information Network (CIESIN) at Columbia University, while the grid-based allocations have been done based on current LandScan data from ORNL [5]

SIGNIFICANCE & DIFFERENCE GIVEN PRIOR WORK

Most of the work mentioned above use knowledge discovery from spatial, temporal data to analyze climate patterns and changes.Such data was used to analyze rainfall patterns and estimate weather conditions. Most of the work done in the area of climate change are niche to certain element of climate. Some work done in the area of greenhouse gases is to estimate the amount of carbon emissions made by a country that year. Hence we devised a formula to calculate the carbon footprint per country based on parameters such as carbon -dioxide emissions, methane and nitrous oxide emissions etc. Additionally we built a drought prediction model considering parameters such as SPEI[the Standardized Precipitation Evapotranspiration Index] to determine the onset, duration and magnitude of droughts. Finally, we studied the correlation between agricultural yield and population to check if a country is behind in its yield production due to climatic conditions affecting its agricultural lands. We performed a multilayer analysis to check how the above parameters affect the planet. We utilized the World Bank dataset to perform our analysis.

DATASET

1.1 World Bank

The World Bank[6] collects community data and organises them based on country. They focus on major economic challenges such as poverty and they analyze parameters such as Gross Domestic Product (GDP) to track the nation's economy. One of the factors that affects the GDP of a country is the change in climatic conditions. Frequent droughts, floods and hurricanes can cause severe damage to property and people and the World Bank constantly monitors climate data in several countries to predict financial trends in a country. We are using this data to understand how climate might change in the future. This dataset has around 20,000 instances and 56 different climate aspects. The data set covers 233 countries over a period of 18 years from 1990 - 2008. Predominantly the data consists of numerical data. Features like NO2 emissions, Methane emissions, Carbon emissions which has been used for calculating carbon footprint are expressed in terms kilotons CO2 equivalent (ktCO2). The dataset also provides other important attributes like population, GDP and crop yield for all 233 countries.

1.2 CSIC - Spanish National Research Council

The CSIC[13] collects data mainly for the purpose of research in the field of science and Technology. CSCI provides a dataset containing Standardized Precipitation Evapotranspiration Index (SPEI) along with temporal and spatial information. SPEI is a drought index that allows comparison of drought severity through time and space. The dataset contains about one million instances of SPEI values computed for various locations over the period of 1990 to 2015.

METHODOLOGY

1 Subtasks

1.1 Carbon Footprint Prediction and Analysis

In 2007 carbon footprint was first used as a measure of carbon emissions. It was much focused measure as they precisely measure the Co2 emissions and how they cause climatic changes. Carbon footprint is the overall amount of greenhouse gas emissions for a country, state, household or for an individual. An increase in greenhouse gas emission and therefore in carbon footprint is one of the main reasons for global warming. Increase in carbon footprint is having profound effects on the environment. The rise in temperature and shifts in precipitation patterns are changing the growth rates of plants, in turn increasing the sea surface temperatures. Increase in weather patterns and shift in temperatures also affect the wildlife. According to the World Health Organization, climate change is projected to increase the percentage of people in Mali suffering from hunger from 34 percent to at least 64 percent 40 years from now. An increase in malnutrition is caused by the result of climate change on food crops, such as drought that interferes with the growing season.

1.1.1 Data Collection

We used the climate change dataset provided by world bank for analyzing the carbon footprint. Various attributes like CO2 emissions, NO2 emissions, Methane emissions and population are taken into consideration for calculating the carbon footprint. We collected five different attributes for all 233 countries and put together a formula for calculating the carbon footprint. After doing research we found that greenhouse gas emissions and population were the targeted attributes for our research.

1.1.2 Data Preprocessing

For cleaning and preprocessing the data, we predominantly used pandas. As the first step we cleaned the dataset by creating a dataframe and replacing empty fields with zero. Since we were handling time dependent data, it was important to average out the empty fields. We used Country name as the primary index and created a dataframe with five attributes for 233 countries for 18 years. Once we had values for all the entries we calculated the carbon footprint using the following formula:

$$\frac{CO2\ emissions + NO2\ emissions + Methane\ emissions + Other\ green\ hous\ gas\ emissions}{Population}$$

Figure 1: Carbon Footprint Formula

The calculated carbon footprint was then visualized to understand more about the data.

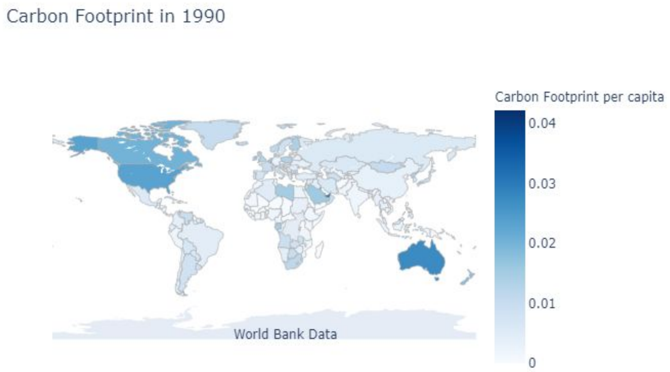

Figure 2: Distribution of carbon footprint in 1990

Once the carbon footprint was calculated we performed a dimension transformation to perform time series analysis and future projections.

1.1.3 Design

Our goal was to perform time series analysis and project carbon footprint for five major contributors. Time series analysis helps us to understand the trend. The trend could be positive or negative. A positive trend indicated that the value is gradually increased over a period of time, whereas a negative trend indicated that the value is gradually decreasing. Once we determine the trend it is easier to project future values based on the trend. Our design included multiple trend analysis models such as simple visualizations and moving averages. We have also used multiple projection methods like ARIMA and LSTM.

1.1.4 Implementation

Time Series Analysis

For the time series analysis we used two different models. The first model was to simply plot the data using matplotlib to understand the general trend. We plotted data for all five contributors.

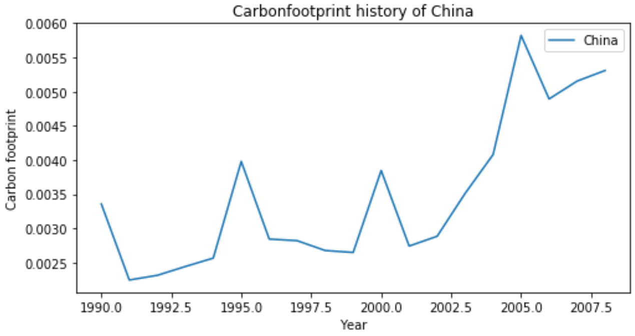

Figure 3: **Carbon footprint History of China**

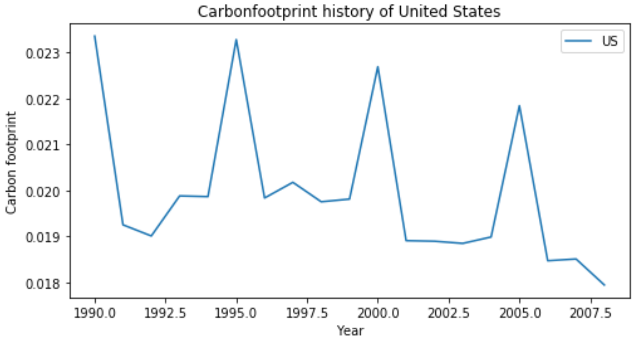

Figure 4: Carbon footprint History of United States

The second time series analysis we performed was the moving average analysis where we created a series of averages of different subsets of the full data set.

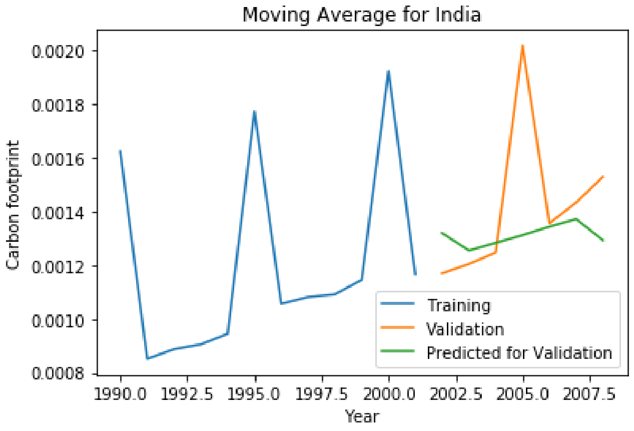

Figure 5 : Moving Average Value for India

We also calculated the RMSE values for the moving average analysis. The RMSE values were between 0.001 - 0.006 which indicated better fit.

Time Series Prediction

For the time series prediction we used multiple prediction models. The first step was to set Year as the primary index and created dataframes for every country. Once we created the data frame we split the data into training and validation set. We first tried the ARIMA model for predicting the value. The ARIMA model predicts the value based on its own value that is, own lags and lagged forecast errors. The ARIMA model had a RMSE of 0.001-0.004 which indicated a great fit for our data. Using the ARIMA Model we also projected carbon footprint for 17 years in the future.

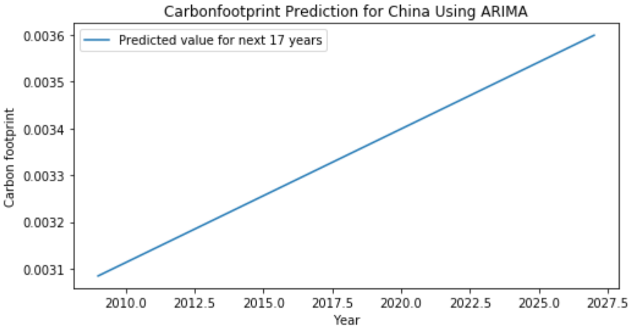

Figure 6: ARIMA Model Prediction for China

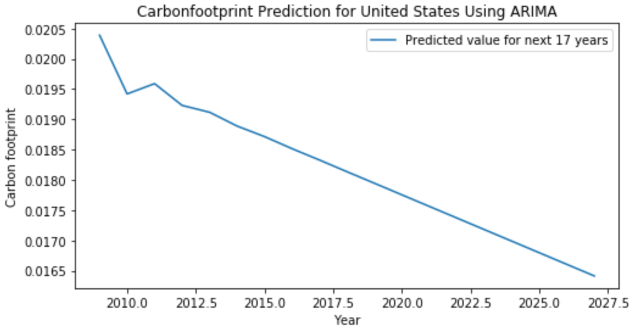

Figure 7: ARIMA model prediction for United States

The second model that we used was LSTM. The reason we chose LSTM was the fact that LSTM works perfectly for data which have long term dependencies. We used ADAM for optimizing our LSTM model. ADAM is an adaptive learning rate optimization algorithm.The reason why we used ADAM Optimization was to have learning rate for each parameter individually. Another huge advantage of ADAM was the fact that it can handle sparse gradient in a noisy problem. The RMSE were between 0.0001 - 0.006 which indicates that LSTM fits better than ARIMA model.

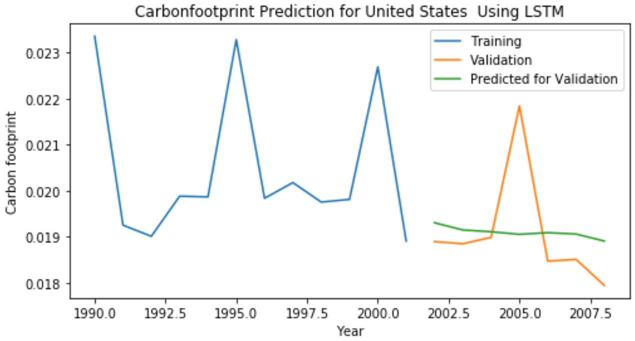

Figure 8: LSTM Model for United States

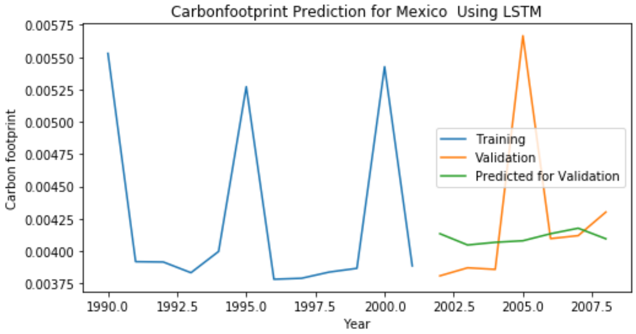

Figure 9: LSTM Model for Mexico

From the analysis we understood that for China and India the carbon footprint is likely to increase in the next 17 years whereas there is no certainty of an increase or decrease in Mexico. Based on the analysis, there is a definite decrease in carbon footprint for United States of America and the United Kingdom.

1.2 Drought Prediction and Analysis

Drought is a prolonged period of abnormally low rainfall, leading to an acute shortage of water. Droughts have historically resulted in conditions of famines, resulting in heavy loss of life. African countries like Ethiopia, Sudan, Somalia, Uganda and tropical places like China, Pakistan are most prone to droughts.

1.2.1 Data Collection

We utilized the dataset provided by CSIC for drought prediction. The dataset consists of Standardized Precipitation Evapotranspiration Index (SPEI) values computed for various regions over a period of years from 1990 to 2015. SPEI is a drought index that allows comparison of drought severity through time and space. SPEI is calculated using the following formula :

$$\textbf{SPEI = P - PET}$$

where P is precipitation and PET is Potential Evapotranspiration

Value Interpretation -

Extremely wet: SPEI >= 2

Very wet: 1.5 <= SPEI <= 1.99

Moderately wet: 1 <= SPEI <= 1.49

Near Normal: -0.99 <= SPEI <= 0.99

Moderately dry: -1 <= SPEI <= -1.49

Severely dry: -1.5 <= SPEI <= -1.99

Extremely dry: SPEI < -2

1.2.2 Data Preprocessing

The data collected from the previous step contained the features latitude, longitude, date and SPEI value. These features were pruned to obtain the SPEI values pertaining to Boulder and the final dataset consisted of the features : date and SPEI. On further analysis of this data, it was observed that most of the SPEI values were missing. So, the missing data had to be handled by removing them using Pandas. Also, the feature values of the feature date were transformed to Pandas's datetime data type to create a time series dataset with spatial information. The distribution of the

SPEI values in the final preprocessed dataset is as seen in Figure 10 and 11.

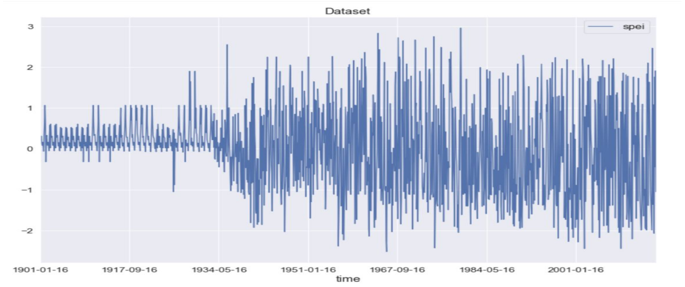

Figure 10 Variance of SPEI over time

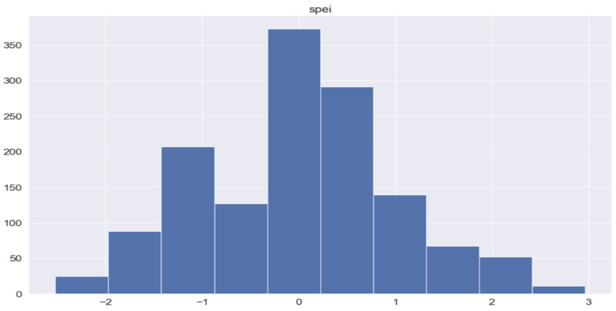

Figure 11 Distribution of SPEI in the preprocessed dataset

1.2.3 Design

Our goal was to perform time series analysis on Boulder's past SPEI values and incorporate that in drought prediction. Time series analysis of the SPEI values helps us to understand the trend of it, which in turn could help in predicting the SPEI values in future. We have used ARIMA and LSTM for this module.

1.2.4 Implementation

ARIMA

Similar to the ARIMA model used for the carbon footprint analysis, We implemented an ARIMA model to perform trend analysis of SPEI values. Initially the autocorrelation of the preprocessed dataset was performed and visualized as seen in Figure 12.

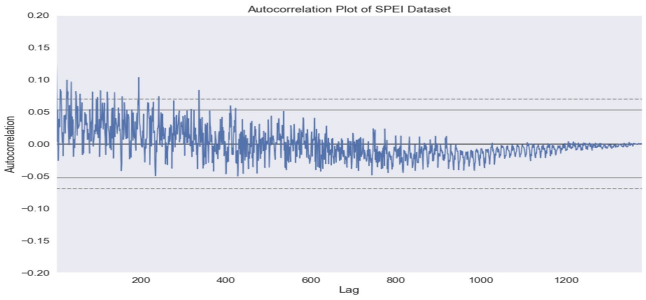

Figure 12 AutoCorrelation plot

This helped in understanding how the time series is correlated with itself. Further, the ARIMA model was fitted with the preprocessed dataset. The residual errors were calculated and visualized as seen in Figure 13.

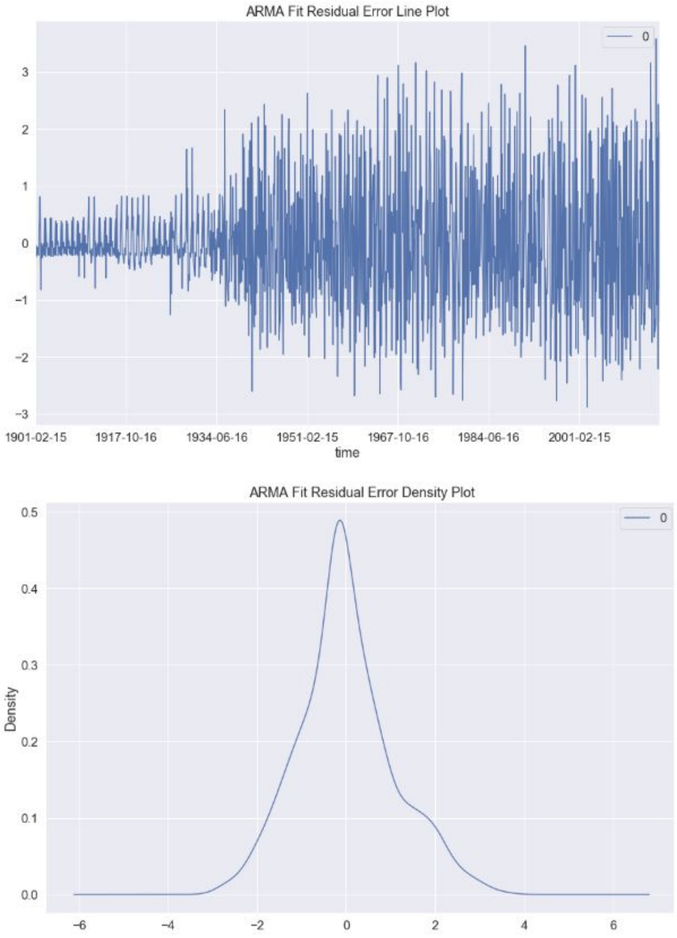

Figure 13 Visualization of the residual error analysis

This residual error from forecasts on the time series provides an insight into the forecast error, which in turn can be used to correct forecasts. Finally, The dataset was split into training and testing test. The former was used to fit the ARIMA model and rolling forecast was performed for the test set. The actual and the predicted SPEI values were visualized as seen in Figure 14.

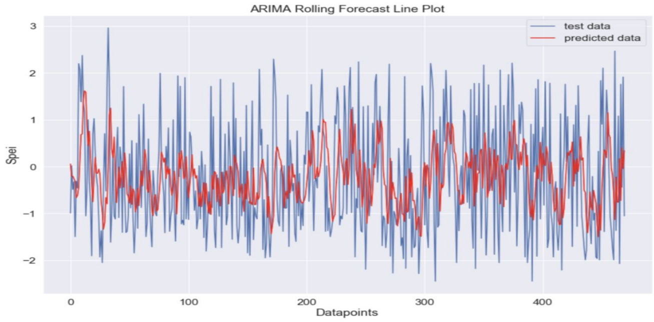

Figure 14 ARIMA Rolling Forecast Actual vs Predicted data visualization

LSTM

In order to perform drought predictions using the SPEI dataset, an LSTM model was implemented using mean squared loss function and the ADAM optimizer. The model uses an one-shot multi-step forecasting methodology to predict for a period of 12 months. The data preparation involved converting the raw spatial-temporal SPEI dataset into a supervised learning dataset by creating the train data including the labels and splitting the overall data into train and test. The Mann-Kendall test was performed to understand the trend seen in the raw SPEI data. Using this as a directional reference, the LSTM model was built to perform the multi-step forecasting for the greater Boulder area. The results obtained are as seen in Figure 15.

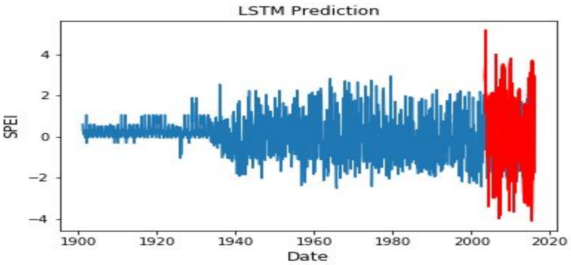

Figure 15 LSTM multi-step Prediction

1.3 Population and Yield Analysis

As population increases there is more demand for food,energy and resources. Increasing population with land degradation increases the challenges of crop production. Overpopulation results in scarce land resources and additional emissions into the atmosphere. This section covers the correlation between population and agricultural yield for major countries such as Australia, India, China, the United States, South Africa and smaller countries such as Sudan and El Salvador.

1.3.1 Preprocessing

The population and agricultural yield data was taken from the World Bank's climate change dataset. Out of 56 different series, the population and the cereal yield data was extracted out into 2 different data frames using the Pandas data processing framework of Python. Columns that were completely empty were dropped. Instances that had empty values were converted to zeroes through the numpy package. The countries that had feature values of zero were added into a separate list. Both the dataframes mentioned above were filtered out in such a way where they consisted of data with countries that had values available for all 19 years.

Next, the required country's population and yield values from 1990 to 2009 was filtered out from the modified population and yield dataframes. The values were transformed from row wise data to columnar data using transpose function and the values' data types were converted to numeric using Pandas.

1.3.2 Design

For performing analysis between the population and agricultural yield, we utilized the correlation coefficients to determine the strength of the relationship between the two parameters in focus.Correlation was chosen because it determines the strength of association between two parameters and the direction of the relationship. A value of +1/-1 determines a strong perfect relationship between the two variables. As the correlation value moves towards zero, the relationship gets weaker.

We utilized two correlation metrics to perform the correlation analysis namely: Pearson, Spearman and Kendall Tau correlation.The Spearman correlation function evaluates a monotonic relationship between ordinal values. They function based on the ranked value for each variable. The Pearson correlation function evaluates the linear relationship between two continuous variables. The Kendall Tau correlation function is similar to the Spearman correlation function.

1.3.3 Implementation

The three correlation metrics were calculated between population and agricultural yield for the countries namely Australia, India, China, the United States, South Africa, Sudan and El Salvador.

Australia had a correlation value of -0.218 between population and crop yields suggesting that it has a negative correlation. As the value is closer to zero, the two parameters are weakly correlation. However as the population increased, the crop yields decreased in Australia from years 1990 to 2009. Below is a table of the countries and their respective Pearson correlation value.

Aus	India	China	USA	SA	Sudan	El Salvador
-0.21	0.95	0.93	0.90	0.80	0.30	0.76

Table 1 - Pearson Correlation coefficients

However others countries apart from Australia has a positive correlation values. India has an almost perfect correlation value of 0.95, followed closely by China and the United States. Sudan has a weak positive correlation value of 0.3 which isn't significantly higher. It is more towards the neural zone which suggests that crop yields may or may not increase with population.

EVALUATION

1 Carbon Footprint

It is important to evaluate forecast accuracy using genuine forecasts. Consequently, the size of the residuals is not a reliable indication of how large true forecast errors are likely to be. To understand the performance of our model we performed multiple evaluation techniques.

1.1 Mann-Kendall Test

The purpose of the Mann-Kendall test was to statistically assess the presence of upward or downward trend for the variable of interest over a period of time. In our project we wanted to check for a monotonic trend for the carbon footprint over a period of 18 years. We performed Original Mann-kendall test instead of the modified version typically used for autocorrelated data. The test values for the five countries are as follows:

S.no	Country	Trend	Slope	Z score	Actual Prediction
1	China	Increasing	0.00013	3.70	Increasing
2	India	Increasing	3.148	3.428	Increasing
3	Mexico	No trend	7.67	0.69	Increasing & Decreasing
4	United States	Decreasing	-0.0017	-2.86	Decreasing
5	United	Decreasing	-0.00010	-3.84	Decreasing

Table 2 - Mann-Kendall Analysis for carbon footprint

So from the test we could verify the trend for our projections, we found that Mexico was one of those countries where the Carbon footprint doesn't have any trend. The values change erratically. It can have an increase and it can decrease too.

1.2 Cross Validation Analysis

We also performed cross validation analysis for all over models where we split our dataset into training set and validation set. Once the training was completed we evaluated our model by comparing the predicted values with the validation set. The Cross validation analysis were plotted to understand the fit.

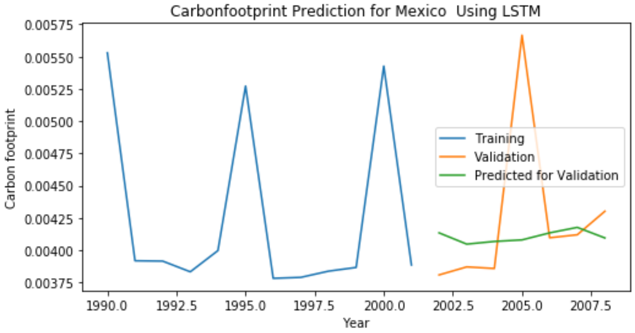

Figure 10: Cross Validation Analysis for LSTM Model

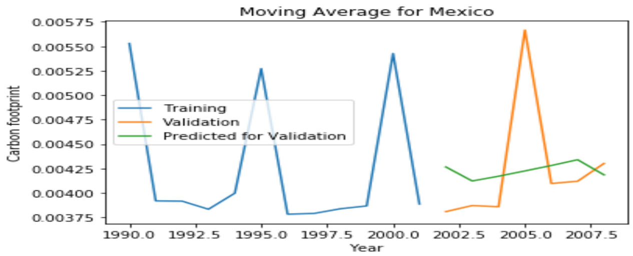

Figure 11: Cross validation Analysis for moving average

1.3 Root Mean Square Error

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. For our ARIMA model we got RMSE values between 0.001 and 0.006 and for LSTM we got in between 0.0001 to 0.0006, this indicated that LSTM has better fit then ARIMA for our data.

2 Drought Prediction

2.1 Mann-Kendall Test

Modelling the carbon footprint analysis, we wanted to check for a monotonic trend in the SPEI scores. We performed Original Mann-kendall test here also. The test values for the five countries are as follows:

Trend	Slope	Z score	Actual

			Prediction
Increasing	-0.00057	-8.151	Increasing & Decreasing

2.2 Root Mean Square Error

Similar to the ARIMA model for carbon footprint analysis, Root Mean Square Error (RMSE) is used for drought prediction in both the ARIMA and LSTM models as it will have a larger restriction on errors and results in a score that is in the same units as the forecast data. For both the ARIMA and LSTM models for drought prediction, we got RMSE with negligible delta indicating both fit the data well.

3 Population and Yield

Evaluating the correlation values between population and Yield was done through a series of different methods. This correlation analysis was done from the year 1990 to 2009. The values obtained for this time period does not imply that future correlation values may stay the same. We used the percentage change analysis, other correlation metrics and research articles to evaluate our correlation results.

3.1 Percentage Changes

While evaluating our countries, we found that Australia had a negative correlation and India had a near perfect strong correlation value between population and agriculture. Hence to understand why this was the case, a perchange change in population and agricultural yield was conducted for the two countries. The percentage change in population was calculated as the difference between the population of the country in 2009 and the population in 1990 divided by the population in 1990. A similar formula was used to calculate the percentage change in agricultural yield for the countries. Below the table of the percentage differences in both population and agricultural yield for all countries.

Country	Aus	India	China	US	SA	Sudan	El Salvador
% pop	28	36	17	22	40	60	15
% yield	2.7	35	26	52	135	28	40

Table 3:Percentage Change in Population and Agriculture from 1990 to 2009

From the above table, we can observe the increase the population for Australia is about 28% in the span of 19 years but the crop yields have grown about a scarce 2.7%. Hence the supply doesn’t meet the demand. The reasons for this low crop yield is outlined under the article evaluation. Hence this evaluates the fact that Australia has a negative correlation between population and agricultural yield. As the population increased, it was difficult to meet crop demands.. Out of the list of countries, India had a strong positive correlation value. Judging by its population and yield percentage increases, they appear to proportionally increasing which is why India has the highest positive correlation. The next closest proportionate increase is China followed by the united states. So in this way proportionately increasing yield and population values accounts to a strong positive correlation.

3.2 Correlation Metrics

Most of the correlation analysis was done using the Pearson correlation function. To evaluate the values obtained through this method, other metrics such as Kendall Tau and Spearman correlation metric was used. Out of the two metrics mentioned above, the correlation method that provided the nearest value to Pearson’s correlation was the Spearman’s rank correlation metric. The values obtained through Spearman was similar to ones presented in table 1. Kendall Tau provided smaller correlation values compared to the other two and the values obtained were a little far away from the expected correlations. This is due to the fact that Kendall Tau works with smaller value that Spearman’s rho correlation. Calculations are usually done based on concordant and discordant pairs. Moreover Kendall Tau correlation is insensitive to errors and their correlation values are more accurate for smaller sample sizes. However Spearman’s correlation works well on larger sample sets which is why it would fit really well with regard to the country's population and yield. Additionally Spearman and Pearson correlation metrics are sensitive to errors and discrepancies in data. They work well with deviations in data.

3.3 Research Articles

Apart from statistical analysis to evaluate our model, government projections were used to evaluate the correlation results. As per the government of Australia[9], the country experienced a millennium Drought between the years 1997 and 2009 where the annual rainfall was well below average. The poor conditions led to low agricultural yield and milk production. Australia’s population has grown about 1.5% per year in the last couple years [10] and continued adoption of technical change, improvements in technical efficiency, and structural adjustment within the farming sector is critically important to the ability of humankind to feed itself in the future. As per this wiki article[11], India has shown a steady nationwide increase in agricultural yield over the last 60 years. These gains have come mainly from India's green revolution, improving road and power generation infrastructure, knowledge of gains and reforms. As the population and crop yield increase proportionately, the correlation is a strong positive value.

FUTURE WORK

1.1 Carbon footprint

In future we would implement hybrid neural networks which give us better performance than LSTM. By using the Hybrid model we can reduce the time and memory required to train the model. We can also reduce the overfit we are currently experiencing with our LSTM Models.

1.2 Land and Ocean Temperature

Study the land temperature and ocean temperature to understand global temperature overtime. The surface temperatures can be converted to absolute temperature and temperature anomalies which provides the difference between the observed temperature and the long-term average temperature for each location. The increase in surface temperatures should be measured on account of greenhouse gas emissions which trap heat in the atmosphere.[12]

1.3 Drought Prediction

Implement an ensemble of ARIMA and LSTM or other neural networks to improve upon the accuracy of the models. Extend the scope of the model to make predictions for a longer window frame.

DISCUSSION / CONCLUSION

Being highly motivated by the current problems faced due to climate change happening at an accelerated pace, we ventured into looking at the problem holistically. We wanted to apply the strengths of Data Mining to answer some of the pressing questions related to climate change and its repercussions. These include the 3 main topics we have covered in this project - i) Carbon footprint analysis, ii) Drought predictions, and iii) population growth impact on agricultural yield. We started by wanting to understand how the nations of the world contributed to climate change through their greenhouse gas emissions. This gave us a direction for finding the dataset that would fit our needs. We came across the World Bank dataset which had more than 25 parameters related to climate change. We zeroed in on the ones that helped us answer our KPIs better which were the estimates of greenhouse gas emissions, population and yield metrics. From the NOAA, we deeply ventured into studying droughts through the SPEI metric discussed in this report. We performed detailed time series analysis on the carbon footprint and drought data using models such as LSTMs and ARIMA and we performed correlation analysis between population and crop yields. We evaluated our model accordingly using the Mann Kendall test that analyses the monotonic trend in model predictions. We additionally tested our model accuracy using train-test split. We evaluated the correlation results using the Spearman rank test and Kendall Tau correlation test. From this project, we analysed how human-induced climate change has contributed and affected the environment.

REFERENCES

[1]https://climate.nasa.gov/news/2908/landsat-illustrates-five-decades-of-change-to-greenland-glaciers/

[2] https://climate.nasa.gov/evidence/

[3] K C Gouda , Chandrika M (2016) “Data Mining for Weather and Climate Studies”, International Journal of Engineering Trends and Technology (IJETT) – Volume 32 Number 1- February 2016.

[4] Auroop R Ganguly and Karsten Steinhaeuser (2008) “Data Mining for Climate Change and Impacts”, IEEE International Conference on Data Mining Workshops.

[5] Bhaduri, B., E. Bright, P. Coleman, J. Dobson. “LandScan: Locating People is What Matters”, *Geoinformatics*, 4(2), 34-37 (2002).

[6] https://data.world/worldbank/climate-change-data

[7]https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data

[8]https://toolkit.climate.gov/tool/climate-resilience-evaluation-awareness-tool-creat

[9]https://www.environment.gov.au/system/files/resources/1342e31a-72b6-4b21-a75f-ace1a8318645/files/australian-agricultural-emissions-projections-2050.pdf

[10]https://www.acaresearch.com.au/australian-market-research-blog/bid/204656/Agriculture-Rising-to-the-Challenges-of-Future-Population-Growth

[11] https://en.wikipedia.org/wiki/Agriculture_in_India

[12]https://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature

[13]https://digital.csic.es/dc/contacto.jsp

APPENDIX

1.Honor Code Pledge

University of Colorado, Boulder Honor Code Pledge : On my honor, as a University of Colorado Boulder student, I have neither given nor received unauthorized assistance.”

2. Contribution of Team Members

2.1 Akash Iyengar

Akash worked on cleaning the world bank dataset and calculating the carbon footprint which was later transposed to time dependent data. He also designed and wrote the code for ARIMA model, LSTM model and performed trend analysis on the carbon footprint data. He also designed the moving average model for the carbon footprint data. He performed Mann-Kendall test to evaluate the model and the trend. He also evaluated the correlation analysis model for the population and yield component using Kendall-Tau test. He also performed cross-validation analysis for all models using in predicting carbon footprint.

2.2 Keerthika Rajvel

Keerthika worked on the carbon footprint visualization and drought prediction subtasks..She combined the calculated carbon footprints with the respective country codes. She then used this to create multiple choropleth interactive visualizations of the carbon footprint for each county over several years. Keerthika also cleaned the CSIC dataset to obtain localized SPEI scores for Boulder. She designed and built an LSTM model for multi-step drought predictions. She performed Mann-Kendall test to evaluate the model and the trend. She also performed error analysis using root mean squared error (RMSE) to evaluate the model. She created plots to visualize the results.

2.3 Swarnalatha Natarajan

Swarnalatha worked on the carbon footprint analysis and drought prediction subtasks. She initially collected the world bank data and performed data preprocessing and cleaning. Further, the dataset was split and aggregated by common country by her. The carbon footprint per country was then calculated. Pertaining to the drought prediction subtask, Swarnalatha cleaned the CSIC dataset to handle the missing values and obtained localized SPEI scores for Boulder. Further, She performed exploratory data analysis to get an insight into the dataset. She then designed and implemented an ARIMA model for drought predictions. The results of the ARIMA model were visualized. She performed Mann-Kendall test to evaluate the model and the trend. She also performed error analysis using root mean squared error (RMSE) to evaluate the model.

2.4 Vandana Sridhar

Vandana worked on the population and yield analysis component. She worked on preprocessing the World Bank Dataset by performing transformations and data type conversions for the population and yield data. Vandana performed Correlation analysis for the countries listed under topic 1.3 in the Methodology section. She evaluated the correlation values using percentage changes in population and agricultural yield. She performed a complete analysis of the Pearson correlation coefficient for the countries and cross matched it with the Spearman Rank correlation function. Additionally she evaluated the correlation results with the Kendall Tau correlation metric and the government projections provided by the countries. Vandana also worked on the moving average time series model for analysing carbon footprint values for various countries.