Applying Predictive Analytics to Climate Change: Predicting Temperature Rise Using Human Behavior Alternative Data

Deepti Saravanan, Jahnavi Swetha Pothineni and Anasse Bari
Department of Computer Science
Courant Institute of Mathematical Sciences
New York University, New York, USA
{ds6812, jp5867, abari}@nyu.edu

***Abstract*—Climate change is a threat to humans and nature. Rise in temperature differs from one region to another around the world. In the general case, warming is relatively higher in land areas than in seas and oceans. Temperature rise is leading to sea level rise, ice melt, ocean precipitation, ocean currents, as well as, major risks to life on land and water. One of the most significant contributors to this threat is greenhouse gas emissions. Understanding the main human factors that cause greenhouse gas emissions is necessary to make data-driven decisions for future actions to combat climate change. In this study, we investigate human activities that contribute to increases in temperature by analyzing new alternative data sources that include internet usage, gas CO2, oil CO2, consumption CO2, air travel, meat consumption, population, car sales, GDP, and housing data, among others. Experimental results indicate that CO2-related features and fertilizer consumption showed a relationship with land temperature. While internet usage patterns did not show any significant correlation, the air travel data showed a reverse effect. The analytics approach and algorithms presented in this paper have the potential to serve as a supplement tool to track and detect emerging predictive features of temperature rise.**

***Index Terms*—predictive analytics, climate change, alternative data**

I. Introduction

Climate change as a result of global warming is one of the biggest concerns all over the world. Scientists are trying to investigate the future scenario of climate change by developing various forecasting models. Therefore, understanding the activities that would cause global warming is important for these climate models. The General circulation models (GCMs) [1] have been constructed by numerical representations of atmospheric physical conditions. Earth system models (ESMs) are advanced models based on GCMs and are mainly used for current climatic studies which consider features such as biogeochemical cycling and atmospheric chemistry. These models are based on the laws of physics such as the conservation of mass, energy, and momentum.

Consider a simple energy balance model. The basic intuition is that the inflow of energy into the Earth's surface should be equal to the outflow of energy from the surface.

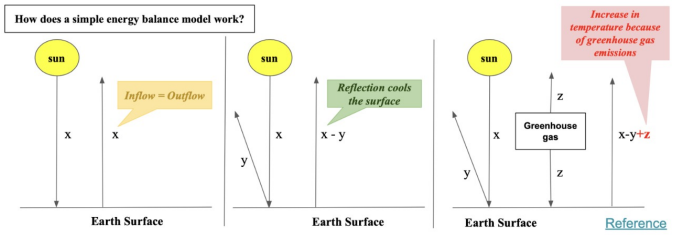

Fig. 1. Simple Climate Model

The net inflow of energy is defined by the amount of energy absorbed by the earth's surface. The outflow of energy determines the surface temperature.

Figure 1 shows three scenarios. In the first scenario where we assume that there is no concept of reflection of light and no greenhouse gas in the atmosphere if the 'x' amount of energy reaches the earth's surface, it emits an 'x' amount of energy.

Consider scenario 2 which is an ideal world. Here, there is a reflection of light but no greenhouse gas in the atmosphere. If the 'x' amount of energy is the inflow and the 'y' amount of that energy gets reflected, the net inflow absorbed by the earth's surface is 'x-y'. Hence, the outflow would be 'x-y'. We can infer that reflection cools the earth's surface.

Now consider the last scenario, where there is a greenhouse gas in the atmosphere. The greenhouse gas has the property of retaining heat back in the atmosphere. Say it retains back 'z' amount of energy. Then the net inflow now becomes 'x-y+z' which would be the outflow. The current outflow is greater than the ideal world scenario and is directly proportional to the amount of greenhouse gases in the atmosphere.

The climate models usually simulate these scenarios. To achieve long-term climate change mitigation and adaptation goals, there must be a global effort to decide and act upon effective and realistic factors. This requires an understanding

of the various human controllable factors, which are often diverse and involved in this phenomenon. Our work is motivated by the importance of addressing the human factors leading to increased emissions of greenhouse gases into the atmosphere. In this study, we apply simple models to predict the rise in average surface temperature using human economic activity alternative data that cover demographics, migration, geography, agriculture, and food habits.

II. Literature Review

Due to climate change and its impacts, natural disasters such as floods, droughts, and storms are increasing yearly. In the study, European Multi-Model Ensemble (EMME): A New Approach for Monthly Forecast of Precipitation by Morteza et al., [2] they have combined individual forecasts from a group of European climate models to produce an ensemble forecast for precipitation. Artificial neural networks, support vector regression, decision trees, and random forests were used for the monthly forecasts of precipitation. Henrike et al. [3] have proposed a risk-based framework to understand the Climate Change and Land use change interactions on biodiversity. Partha et al. [4] used Advanced Very High-Resolution Radiometer (AVHRR) infrared satellite sea surface temperature (SST) data and then fed this data to ANN and LSTM to predict SST. Takeshi Ise et al. [5] have constructed image grids using month mean temperature of regions for a period of 30 years and predicted a rise or fall in temperature using LeNet. In this study, they have found that the performance of a top-down approach like this is notably high in comparison to the conventional bottom-up physics-based models for decadal-scale forecasting. In the study, Forecasting The Air Temperature at a Weather Station Using Deep Neural Networks by Debneil Saha Roy et al. [6] they have used Multi-Layer Perceptron (MLP), LSTM, and an ensemble of CNN and LSTM to calculate the average air temperature as a function of its own historical values as well as other weather variables like average wind speed, precipitation, snowfall, snow depth, maximum temperature, and minimum temperature.

III. Methodology

A. Business Understanding

The measure our work focuses on is the **average surface temperature**, which is recorded annually. The country-wise availability of data further strengthens this to be our ideal choice.

B. Data Collection

The data used in this study consisted of data from **time period 1990 to 2020**. The **10 features country-wise**, shown in Table I, would be the focus of the study after careful consideration of various aspects.

The average surface temperature dataset recorded the mean temperature of the country over each year in Celsius. The GDP is calculated by multiplying GDP per capita with the population of the country in that year. The CO2 per capita dataset contains the annual total production-based CO2 emissions in tonnes/person. The land elevation denotes the average elevation of the country. The real housing price index of each country over the years was used in the housing market dataset. The migration rate dataset denotes the net migration rate of the country in a particular year. Air Travel data has the sum of air passengers carried by the air carriers registered in the corresponding countries. Internet Usage corresponds to the total number of Internet users in each country. Fertilizer Consumption data is the sum of the synthetic inputs of Nitrogen, Potassium, and Phosphorus along with organic nitrogen inputs in each country. Meat Consumption is measured in kg per capita for each country.

C. Data Preparation

1) Data Cleaning: The data collected was not in usable form. Table I shows the preprocessing techniques that were employed for each dataset –

2) Data Fusion: Data from 1990-1996 was missing in the migration dataset. Hence, the extrapolation technique was employed to fill the gap. Points were missing in between for the rest of the data. Linear interpolation [7] was employed to tackle missing data. This technique was chosen because this is a short-term forecast of 20 years, and the time series plot for the datasets did not result in any highly complex functions to employ more convoluted interpolation methods.

Once the missing data were filled, the features from all the datasets were fused together on the basis of country and year. Initially, the country name field was used to represent the country key. But the names were not unique with some datasets using the official names of the countries while some did not. Hence, the ISO Code of the countries was chosen instead as it was unique and uniform across datasets. Meat consumption and housing data had missing chunks of data for multiple countries. Due to the unavailability of data, these two features are dropped. The final dataset is of the size 1022×21 with 40 countries (22 $\%$ of developing and 78 $\%$ of developed countries).

Figure 2 shows the trend of human activities (features) in the United States over a period of 30 years.

3) Feature Selection: Several feature selection algorithms are applied to eliminate redundant features and retain only those features with high predictive power for surface temperature prediction.

Following a literature review of feature selection techniques, the following were best suited for the project –

- Feature Variance
- Pearson Correlation Analysis [8]
- ANOVA f-score [9]
- Mutual information based feature selection [10]

The results of each of the feature selection techniques are compiled here.

TABLE I
DATA COLLECTION AND CLEANING

Category	Features Data	Preprocessing Techniques employed
Prediction	Average Surface Temperature [17]	Extracted data was monthly – grouped by year based on mean Extracted data for in distributed between files for different countries – Data was consolidated into one CSV file Extracted required columns and rows from 1990-2021, Converted column data types Checked and treated null values
Demographics Economics Motion	Population [18] GDP [18] CO2 per capita [18]	Extracted required columns and rows from 1990-2021, Converted column data types Created a calculated column to convert measures to standard form (normalized by population) Checked and treated null values
Geography	Land Elevation [19]	Merged columns were split and renamed, Converted column data types Checked and treated null values
Economics	Housing Market [20]	Extracted required columns and rows from 1990-2021, Converted column data types Extracted required rows as some columns were categorical Checked and treated null values
Motion	Migration Rate [21]	Extracted required columns and rows from 1990-2021, Converted column data types Checked and treated null values
Motion	Air Travel [22]	Converted data format to the standard followed by other datasets Country renaming in case of multiple names Extracted required columns and rows from 1990-2021, Converted column data types Extracted required rows as some columns were categorical Linear Imputation to fill missing data Checked and treated null values
Technology Agriculture	Internet Usage [23] Fertilizer Consumption [24]	Extracted required columns and rows from 1990-2021, Converted column data types Linear Imputation to fill missing data Checked and treated null values
Food Culture	Meat Consumption [25]	Extracted required columns and rows from 1990-2021, Converted column data types Extracted required rows as some columns were categorical Linear Imputation to fill missing data Checked and treated null values

Feature Variance

The variance of the features was analyzed with a threshold of 10%. The following features showed variance less than the threshold –

- Flaring co2 per capita
- Other co2 per capita

On analyzing the trend of 'Other co2 per capita', there is a possible causal relationship with some lag. Hence, we just eliminate the feature 'Flaring co2 per capita' at this point.

Pearson Correlation Analysis

The Pearson correlation between pairs of features was computed and visualized using a heatmap in Figure 3.

Table II shows the highly correlated features and corresponding decisions made on feature elimination and retention.

ANOVA f-score

ANOVA f-test was performed to check whether the variance of means between features and surface temperature is more than expected.

It can be inferred from Table III that F-values are high for most of the features, except migration rate and internet usage. Hence, the variance between the features and the value to be predicted (surface temperature) is greater than the variance within the features, making these good predictor variables. In addition, the p-values of all the features are less than 0.05, making the f-stats statistically significant. This proves that the chosen population of features would be good predictors.

Mutual information based feature selection

This technique was employed to understand the uncertainty in surface temperature prediction given each of the features. The features were then ranked from lowest to highest uncertainty using the mutual information measure.

Table IV shows the mutual information scores of features in the ranked order.

The final set of features post feature selection techniques for modeling is: **Population, GDP, CO2 per capita, Consumption CO2 per capita, gas CO2 per capita, oil CO2 per capita, other CO2 per capita, Migration Rate, Air Travel, Internet Usage, Fertilizer Consumption, and Elevation**.

D. Test for Causation

To understand if there is any causal effect between the features and average surface temperature, we performed Granger Causality and Reverse Causality. The input time series is expected to be stationary. So, we first perform three kinds of tests on the features to check if the above assumption is obeyed.

1) Qualitative Stationary Test: First, we tested if the time series for each feature is stationary by plotting a histogram.

Fig. 2. Human Activities Trend in the United States

As can be seen in Figure 4, the histograms of most of the features are skewed. Hence, most of the feature time series are possibly non-stationary.

2) Quantitative Stationary Test: This can further be observed quantitatively by splitting the time series into two halves and comparing the mean and variances of the two parts. If the difference is too high, then those time series are non-stationary.

From Table V, we can infer that the following features are non-stationary – population, GDP, fertilizer consumption, internet users, air travel

3) Statistical Stationary Test: In addition to quantitative and qualitative, we also performed the Augmented Dickey-Fuller test [11] to test for stationarity of the time series of features. The features with a p-value greater than 0.05 do not reject the null hypothesis and thus are non-stationary. Note that the **average surface temperature time series that is to be predicted is stationary**.

TABLE II
CORRELATED FEATURES FROM PEARSON ANALYSIS

Correlated Features	Action Taken
GDP, Population, Fertilizer consumption	Even though the three features are correlated, the population, on the whole, is an important factor while studying climate change impact. Fertilizer consumption is a popular human activity that might affect multiple factors independent of the GDP. Hence, all three features are retained in the final list of features.
GDP, Air Travel	GDP and air travel are highly correlated (0.97). But air travel is not the only factor affecting GDP and might have a different level of impact when combined with other features. Both the features are retained in the final list of features.
Total GHG per capita, CO2 per capita	These two energy-related features show a high correlation (0.85). Hence, total GHG per capita is eliminated as a redundant feature. Note: this feature also showed low variance.
Fertilizer consumption, Air Travel	Fertilizer consumption and air travel are highly correlated. Since both these features cover different aspects of human activities, they are retained in the final list of features.

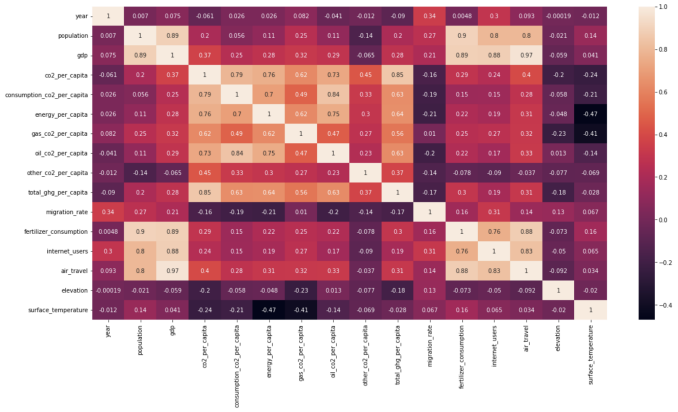

Fig. 3. Pearson Correlation Heatmap

TABLE III
ANOVA F-TEST RESULT

Features	F-stats	p-values
Population	3964.9362	0.00
GDP	635.757	0.00
CO2 per capita	333.973	0.00
consumption CO2 per capita	172.404	0.00
energy per capita	771.395	0.00
gas CO2 per capita	341.497	0.00
oil CO2 per capita	423.610	0.00
other CO2 per capita	115.555	0.00
total GHG per capita	211.241	0.00
Migration rate	51.70	1.94084945e-208
Fertilizer consumption	752.46	0.00
Internet usage	46.62	4.05802167e-194
Air Travel	526.349	0.00

TABLE IV
FEATURES RANKED BASED ON MUTUAL INFORMATION

Features	Mutual Information Score
Land Elevation	1.0
Population	0.77
Fertilizer consumption	0.52
Flaring CO2 per capita	0.49
GDP	0.47
Energy per capita	0.42
Air Travel	0.35
Oil CO2 per capita	0.35
CO2 per capita	0.33
Gas CO2 per capita	0.32
Total GHG per capita	0.32
Consumption CO2 per capita	0.30
Migration rate	0.28
Other CO2 per capita	0.27
Internet usage	0.17

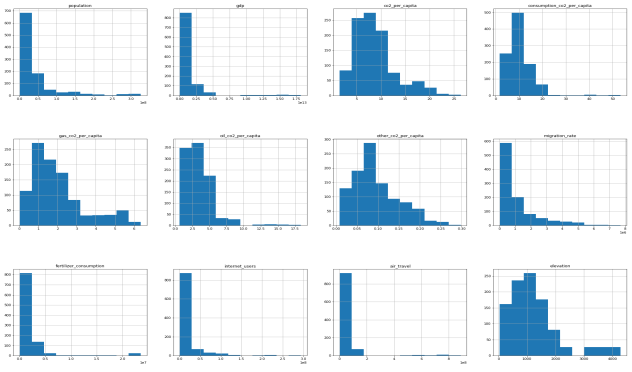

Fig. 4. Qualitative Stationary Test Result

From Table VI, we can infer that the following features are non-stationary – GDP, population, fertilizer consumption, internet users, air travel

Putting together the results from the three tests, we conclude that the **time series of population, GDP, fertilizer consumption, internet users and air travel are non-stationary**.

4) Cointegration Test: Cointegration test was performed to identify human activities with non-stationary time series that are cointegrated with the surface temperature.

All the p-values are greater than 0.05 in Table VII. Hence, **none of the non-stationary time series features are cointegrated with the surface temperature**.

5) Conversion of Non-stationary to Stationary Time series: Using difference operation with a shift of 1, we converted the non-stationary to stationary time series with a p-value of 0.00, which is less than 0.05 and hence, is statistically significant. Figure 5 shows The histogram of the final data with stationary time series.

6) Granger Causality Analysis: Granger Causality Test was run on the stationary time series data of features. A max lag of 1 was chosen as we are interested in the immediate causality. As can be seen from Table VIII, both Granger

TABLE V
QUANTITATIVE STATIONARY TEST RESULT

Time Series	Mean of first half	Mean of second half	Variance of first half	Variance of second half
Population	34617472.95	42221392.40	2321248699673917.50	4758800800554871.00
GDP	945379198495.06	1298135937817.55	14971184840542931344442496.00	98358413535835983760588880.00
CO2 per capita	9.29	8.33	15.39	20.59
consumption CO2 per capita	10.51	9.94	18.07	44.08
gas CO2 per capita	1.86	2.07	1.41	2.45
oil CO2 per capita	3.63	3.54	2.68	7.60
other CO2 per capita	0.10	0.09	0.003	0.003
Migration rate	913249.62	1325857.34	900942908988.37	2499518245216.96
Fertilizer consumption	1733557.28	2135399.22	6978800156148.64	23657388487941.37
Internet usage	14513292.84	18180074.18	675708054309899.38	1929653038294557.00
Air Travel	28124777.00	52169240.58	1185800799674121.50	22759106237130872.00
Elevation	1108.89	1335.76	520532.20	1172148.90

TABLE VI
STATISTICAL STATIONARY TEST RESULT

Time Series	ADF Statistic	p-value
Population	-1.82	0.37
GDP	2.95	1.00
CO2 per capita	-5.57	0.000001
consumption CO2 per capita	-6.04	0.00
gas CO2 per capita	-4.34	0.00038
oil CO2 per capita	-4.54	0.00017
other CO2 per capita	-5.55	0.000002
Migration rate	-3.95	0.0017
Fertilizer consumption	-1.83	0.37
Internet usage	-1.18	0.68
Air Travel	2.76	1.00
Elevation	-3.99	0.0015
Average surface temperature	-4.67	0.000096

TABLE VII
COINTEGRATION TEST RESULT

Non-stationary time-series	Cointegration p-value
Population	0.61
GDP	1.0
Fertilizer consumption	0.62
Internet usage	0.85
Air Travel	1.0

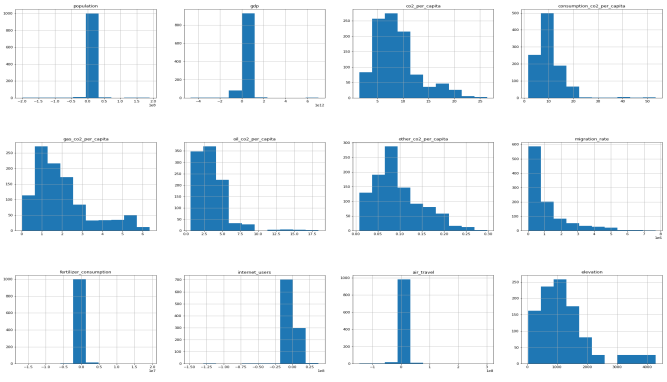

Fig. 5. Histogram after conversion to Stationary Time series

Causality and Reverse Causality have p-values greater than 0.05 for all the features. This indicates that there is no direct causal relationship between the features and average surface temperature.

E. Modeling

The final selected features for predicting climate change are **Population, GDP, CO2 per capita, Consumption CO2 per capita, gas CO2 per capita, oil CO2 per capita, other CO2 per capita, Migration Rate, Air Travel, Internet Usage, Fertilizer Consumption, and Elevation**. The size of the dataset at this point is 15,330x12. Complex models might overfit the data [26]. Hence, we use basic statistical and machine learning models for average surface temperature prediction. The models used are –

- Decision Tree for regression
- Random Forest Regressor [12]
- Support Vector Regressor [13]
- Lasso Regression [14]
- LSTM [15]

Hyperparameter tuning was performed using the random grid search method for all the models. This technique was chosen as it is flexible and efficient [27]. The search was performed across 100 different combinations to choose the best set of hyperparameters. Table IX shows the best set of hyperparameters used for the models.

IV. EXPERIMENTAL RESULTS

The evaluation metrics used are Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). Table X shows the performance of the machine learning models. The two models, the **Decision Tree for regression and Random Forest Regressor**, show low error values compared to the rest.

V. DISCUSSION

We observed a strong correlation in the time series pattern between the average rise in temperature and CO2-related features, as expected. From observing the trends from the United States plots, oil CO2 per capita drops abruptly from 2008 due to the economic crisis, and the average surface

TABLE VIII
GRANGER CAUSALITY TEST RESULT

Features	Granger Causality p-value	Reverse Causality p-value
Population	0.76	0.93
GDP	0.94	0.96
CO2 per capita	0.83	0.82
consumption CO2 per capita	0.44	0.35
gas CO2 per capita	0.74	0.36
oil CO2 per capita	0.77	0.55
other CO2 per capita	0.85	0.89
Migration rate	0.81	0.37
Fertilizer consumption	0.96	0.74
Internet usage	0.96	0.93
Air Travel	0.92	0.95
Elevation	0.72	0.65

TABLE IX
MODEL HYPERPARAMETERS

Model	Hyperparameters
Decision Tree for regression	Criterion = Squared Error Max Depth = 30 Max Features = sqrt Minimum Samples Leaf = 1 Minimum Samples Split = 5 n estimators = 1400
Random Forest Regressor	Criterion = Squared Error Max Depth = 50 Max Features = sqrt Minimum Samples Leaf = 2 Minimum Samples Split = 10 n estimators = 200
Support Vector Regressor	Criterion = Squared Error Max Depth = 50 Max Features = sqrt Minimum Samples Leaf = 2 Minimum Samples Split = 10 n estimators = 200
Lasso Regression	Criterion = Squared Error Max Depth = 50 Max Features = sqrt Minimum Samples Leaf = 2 Minimum Samples Split = 10 n estimators = 200
LSTM	Epochs = 15 Batch size = 32 Validation split = 2 Activation = tanh Optimizer = adam Learning rate = 0.001

TABLE X
EVALUATION OF MODELS

Model	MSE	MAPE
Decision Tree for regression	**0.92**	**1.52**
Random Forest Regressor	**0.62**	**0.66**
Support Vector Regressor	2.78	15.94
Lasso Regression	21.36	40.68
LSTM	13.23	1.41

temperature also dropped at similar rates. Later, even though there is a decrease in the Oil CO2 per capita, the surface temperatures started to increase due to the rise in gas CO2 per capita. It is also to be noted that other CO2-related features dropped during the same period. Similar results could be inferred from feature selection steps where CO2-related features were consistently ranked higher.

Around 2013 there was an abrupt decrease in the average surface temperature and the CO2-related features stayed low around that time. Also in 2013, Tesla sold a record of 22,477 cars and this could be an indicator of the high demand for electric vehicles.

Surprisingly, the feature that was equally competent with CO2 per capita is Fertilizer consumption. From 2005 both fertilizer consumption and average surface temperature have similar peaks. Fertilizer consumption is also the next high-ranked feature following CO2 features.

Between 2005 and 2017, the peaks in the migration time series almost match those in the average surface temperature. Between 2010 and 2018, the peaks in average surface temperatures follow the peaks of migration, with a lag of almost one year. This could be a good predictor in the future since the effect of migration on the average surface temperature in the current period of time is higher than in the past.

The number of passengers traveling by air travel increases following the peaks in average surface temperature. It is interesting to raise a question about the reverse effect. If true, it could be used as an alternative stock price predictor in the aviation industry.

Internet usage patterns did not show any significant correlation. This could be explained by the fact that the usage of the internet saw a steady increase in the early 2000s while the average surface temperature showed ups and downs in the time period considered.

Hence we can infer that the various human activities have distinctive effects on the average surface temperature. As of 2018, only sixteen countries out of the 197 that signed the Paris Agreement have defined a national climate action plan ambitious enough to meet their pledges. A deeper study along these lines of human factors by the countries would be immensely useful in tackling climate change and adopting policies that would best suit their interests.

VI. Conclusion

This systemic study used data from multiple sources for 10 human activities from 1990-2021 country-wise. This pooled data was then used as alternative data to predict average surface temperature. The collected data was cleaned using various preprocessing techniques. Missing data points were linearly imputed and the data was fused based on country and year. Various feature selection techniques were employed to understand the trend of different time series of features and their interplay with the time series of average surface temperature. The trend for the United States was plotted feature-wise with average surface temperature for comparison. Qualitative, quantitative, and statistical tests were performed to check if the features were stationary time series. The non-stationary features were found to be not cointegrated with the average surface temperature. These were then converted to stationary time series. Finally, Granger Causality Analysis was performed and no direct causal relationship was found. Four predictive models were trained on the dataset. The Decision Tree and Random Forest Regressor models performed the best with minimum mean squared and mean absolute percentage errors.

One major limitation is the scarcity of relevant and comprehensive data sources for various human activities. This is due to the fact that different countries have varying data collection capabilities and infrastructure. Future research should address these limitations by incorporating a wider range of data collection methods. In addition, the study could be extended to analyze the rate of deforestation over time by using computer vision techniques on satellite imagery that could help include new features in the predictive analytics framework.

VII. Acknowledgements

This study was partially supported by the 2023 Courant Institute Suzanne McIntosh Research Fellowship.

References

[1] https://www.ipcc.ch/report/ar4/wg1/

[2] Pakdaman, M., Babaeian, I., Javanshiri, Z. et al. European Multi Model Ensemble (EMME): A New Approach for Monthly Forecast of Precipitation. Water Resour Manage 36, 611–623 (2022). https://doi.org/10.1007/s11269-021-03042-8

[3] Henrike Schulte to Bühne, Joseph A. Tobias, Sarah M. Durant, Nathalie Pettorelli,Improving Predictions of Climate Change–Land Use Change Interactions, Trends in Ecology Evolution, Volume 36, Issue 1, 2021, Pages 29-38, ISSN 0169-5347, https://doi.org/10.1016/j.tree.2020.08.019.

[4] Sarkar, P.P., Janardhan, P. Roy, P. Prediction of sea surface temperatures using deep learning neural networks. SN Appl. Sci. 2, 1458 (2020). https://doi.org/10.1007/s42452-020-03239-3

[5] Ise Takeshi, Oba Yurika, Forecasting Climatic Trends Using Neural Networks: An Experimental Study Using Global Historical Data,Frontiers in Robotics and AI, Jouranl 6,2019, ISSN 2296-9144, https://www.frontiersin.org/articles/10.3389/frobt.2019.00032, 10.3389/frobt.2019.00032

[6] Debneil Saha Roy, Forecasting The Air Temperature at a Weather Station Using Deep Neural Networks, Procedia Computer Science, Volume 178, 2020, Pages 38-46, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.11.005.

[7] Mohamed Noor, Norazian and Abdullah, Mohd Mustafa Al Bakri and Yahaya, Ahmad Shukri and Ramli, Nor. (2014). Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set. Materials Science Forum. 803. 278-281. 10.4028/www.scientific.net/MSF.803.278.

[8] Schober, Patrick MD, PhD, MMedStat; Boer, Christa PhD, MSc; Schwarte, Lothar A. MD, PhD, MBA. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia Analgesia: May 2018 - Volume 126 - Issue 5 - p 1763-1768 doi: 10.1213/ANE.0000000000002864

[9] StudyCorgi. (2020, November 21). F Statistics as a Part of the Anova Test Results. Retrieved from https://studycorgi.com/f-statistics-as-a-part-of-the-anova-test-results/

[10] C. Qi, Z. Zhou, Q. Wang and L. Hu, "Mutual Information-Based Feature Selection and Ensemble Learning for Classification," 2016 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI), 2016, pp. 116-121, doi: 10.1109/IIKI.2016.81.

[11] Mushtaq, R. (2011). Augmented Dickey Fuller Test. Econometrics: Mathematical Methods Programming eJournal.

[12] Liaw, Andy Wiener, Matthew. (2001). Classification and Regression by RandomForest. Forest. 23.

[13] Basak, Debasish and Pal, Srimanta and Patranabis, Dipak. (2007). Support Vector Regression. Neural Information Processing – Letters and Reviews. 11.

[14] Kwon, Sunghoon and Han, Sangmi and Lee, Sangin. (2013). A small review and further studies on the LASSO. Journal of the Korean Data and Information Science Society. 24. 10.7465/jkdi.2013.24.5.1077.

[15] https://www.euractiv.com/section/climate-environment/news/only-16-countries-meet-their-commitment-to-paris-agreement-new-study-finds/

[16] S. Elsworth and S. Güttel, 'Time Series Forecasting Using LSTM Networks: A Symbolic Approach'. arXiv, 2020.

[17] https://climateknowledgeportal.worldbank.org/download-data

[18] https://github.com/owid/co2-data/blob/master/owid-co2-data.csv

[19] https://www.atlasbig.com/en-us/countries-average-elevation

[20] https://www.nar.realtor/research-and-statistics/housing-statistics-and-real-estate-market-trends

[21] https://www.migrationdataportal.org

[22] https://data.worldbank.org/indicator/IS.AIR.PSGR

[23] https://drive.google.com/file/d/10D5r9Vcf0rPYBgaaI9FxeRuqr1a1awxm/view?usp=share$_l ink$

[24] https://ourworldindata.org/grapher/fertilizer-consumption

[25] https://drive.google.com/file/d/1MuphuHMz0a2oQklz3hp6B5ufA8uqN2zU/view?usp=share$_l ink$

[26] Lever, J., Krzywinski, M. Altman, N. Model selection and overfitting. Nat Methods 13, 703–704 (2016). https://doi.org/10.1038/nmeth.3968

[27] J. Bergstra and Y. Bengio, 'Random Search for Hyper-Parameter Optimization', J. Mach. Learn. Res., vol. 13, no. null, pp. 281–305, Feb. 2012.