Feeling the heat: Prediction of rise in average surface temperature using human-economic activity alternative data

# Abstract

Every one-degree increase in temperature will impact the planet adversely. Since the 1970s, global mean temperatures have increased by 0.2˚C per decade. The rise in temperatures could melt the ice, lead to rising sea levels, changing ocean currents and precipitation, as well as a potential risk to life on land and water. **Greenhouse gas emissions are the most significant contributor to the climate change issue**. Understanding the **human factors that cause greenhouse gas emissions** is imperative to take future action to combat climate change. Past work has considered many key contributors that include CO2 emissions, ocean, and atmospheric dynamics to predict the change in temperature, but did not research into the source of the issue – human activities. In this study, we **predict the rise in average surface temperature across countries** by tapping into **alternative data such as GDP, global car sales, land elevations**, etc. The result of the study would hold a **high impact on the decision-making of respective governments on prioritizing and working toward immediate concerns**.

# Goal

To predict anomalies in average surface temperature in different countries using various human activity stats as factors. This establishes a **causal relationship** between them.

# Significance

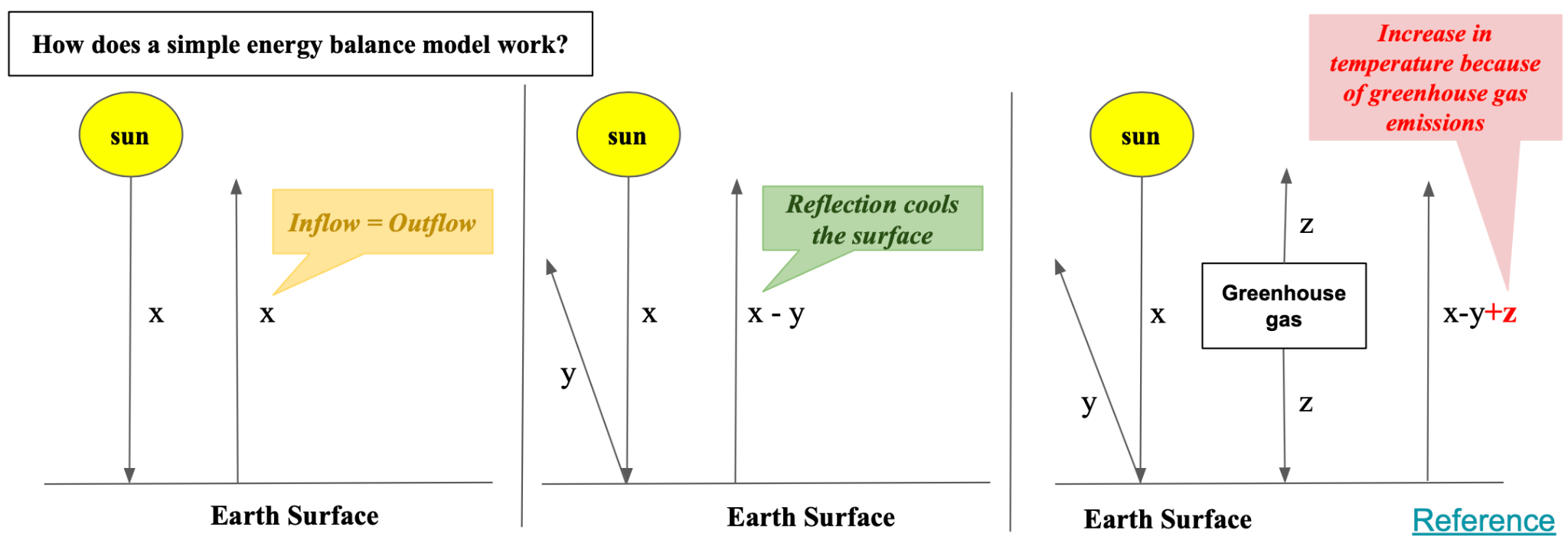
Past work on Climate Change Prediction focused on factors such as wind speed, rate of melting of ice, rate of rising sea level, etc. Such factors are uncontrollable and had to be tackled indirectly, which would be challenging. **Our work focuses on controllable human factors which would help the respective governments to make informed policy decisions, prioritize current issues and prepare for possible calamities.**

[Useful premise](https://www.science.org.au/learning/general-audience/science-climate-change/3-are-human-activities-causing-climate-change) to set the stage on why we focus on human activities mainly.

# Motivation

Our project took inspiration from Climate Modeling, which models the earth’s climatic factors into mathematical equations for future climate predictions.

## What is Climate Modeling?



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

The figure 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 ‘x’ amount of energy reaches the earth’s surface, it emits ‘x’ amount of energy.

Consider scenario 2 which is an **ideal world**. Here, **there is reflection of light but no greenhouse gas in the atmosphere**. If ‘x’ amount of energy is the inflow and ‘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 **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.**

Thus, it is important to address the **factors leading to increased emission of greenhouse gases** into the atmosphere, which are **predominantly human activities**.

# Research Questions

1. Is there a causal relationship between human activities and climate change?
   1. How do we measure climate change? *Average Surface Temperature*
   2. Which human activities might affect climate change?
      1. Which human activities lead to greenhouse gas emissions?
      2. Do we have relevant datasets?
      3. How do we preprocess the data?
   3. Which human activities are correlated with climate change?
      1. Which correlation tests should we employ? Why?
   4. How do we model the causal relationship?
      1. What should be the complexity of the model?
      2. What is the performance of an ensemble of models?
2. What do we infer from the result?
   1. How do we prioritize issues/factors? Do we set a threshold?

# Literature Review

1. Global Climate Models – https://www.ipcc.ch/report/ar4/wg1/ ➔ These complex models use atmospheric parameters like greenhouse gasses, ocean dynamics, cloud systems to predict the temperature.   
   Problems with GCM : There are many key processes that would affect the temperature change but it’s hard to add all of them. Also, there is a constant need for interpretability of these complex models.

1. European Multi Model Ensemble (EMME): A New Approach for Monthly Forecast of Precipitation -   
   ➔ The study combined individual forecasts from a group of european climate models to produce an ensemble forecast. Artificial neural networks, support vector regression, decision tree and random forests were used to predict the climate change.Neural networks and random forest methods performed better than the decision tree and the support vector machine and the ensemble result is better than the individual models.

Problems: There are few studies that have used an ensemble of models but still the accuracy is bad and we need to check what are the conditions that might affect the results.

1. Improving Predictions of Climate Change – Land Use Change Interactions -   
   ➔ This talks about how the human activities transform the natural landscape which in turn affects the climate change in an area.

# Analysis I - Is there a causal relationship between human activities and climate change?

## How do we measure climate change?

There are multiple ways to gauge climate change. The initial measure considered for this work was the **climate change risk index**, which is currently calculated based on the historic impact and vulnerability data of the region. **But it is calculated once in a decade while our goal is to predict climate change annually**. So, we dropped this measure.

The measure our work currently 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.

## Which human activities might affect climate change?

List summary of all the activities here. Explanations go under each sub-question.

### Which human activities lead to greenhouse gas emissions? Do we have relevant datasets?

Initially, data was collected for the time period of 1950-2020. But, a few of the factors considered as features to predict climate change were not available for a few countries in the early period. Imputation methods, such as Stochastic Regression, do not work well to fill a huge chunk of continuous missing data. Applying a machine learning model to predict the past using the present values was considered. But there is a huge possibility of error propagation which might severely affect the performance of the main predictive model.

Also, the significance of the kind of human activities is different in comparing the trends from the 1950s and the 2020s. Since the focus of the project is to analyze the current human factors to tackle climate change, the past data from the 1950s might act as noise rather than relevant information.

Having this in mind in addition to the availability of good quality data, a **time period of 1990-2020** was considered.

The following **12 features country-wise** would be the focus of the study after careful consideration of various aspects.

| **Category** | **Features** |
| --- | --- |
| Climate change prediction *(to predict)* | * [Average Surface Temperature](https://climateknowledgeportal.worldbank.org/download-data) |
| Demographics | * [Population](https://github.com/owid/co2-data/blob/master/owid-co2-data.csv) |
| Geography | * [Land elevation](https://www.atlasbig.com/en-us/countries-average-elevation) |
| Economics | * [GDP](https://github.com/owid/co2-data/blob/master/owid-co2-data.csv) * [Housing Market](https://www.nar.realtor/research-and-statistics/housing-statistics-and-real-estate-market-trends) |
| Motion | * [CO2 per capita](https://github.com/owid/co2-data/blob/master/owid-co2-data.csv) * [Migration Rate](https://www.migrationdataportal.org) * [Air Travel](https://data.worldbank.org/indicator/IS.AIR.PSGR) |
| Technology | * [Internet Usage](https://drive.google.com/file/d/10D5r9Vcf0rPYBgaaI9FxeRuqr1a1awxm/view?usp=share_link) |
| Agriculture and Food Habits | * [Fertilizer Consumption](https://ourworldindata.org/grapher/fertilizer-consumption-usda?country=~OWID_WRL-) * [Meat Consumption](https://drive.google.com/file/d/1MuphuHMz0a2oQklz3hp6B5ufA8uqN2zU/view?usp=share_link) |

### How do we preprocess the data?

#### Data Cleaning

The data collected was not in usable form. The following preprocessing techniques were employed for each dataset –

| **Dataset** | **Preprocessing Techniques employed** |
| --- | --- |
| Average Surface Temperature | * 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 eliminated the rest * Converted column data types * Extracted data from 1990-2021 * Checked for null values * Treated null values |
| Population | * Extracted required columns and eliminated the rest * Created a calculated column to convert measures to standard form (normalized by population) * Extracted data from 1990-2021 * Checked for null values * Treated null values |
| Land elevation | * Merged columns were split and renamed * Converted column data types * Checked for null values * Treated null values |
| GDP | * Extracted required columns and eliminated the rest * Created a calculated column to convert measures to standard form (normalized by population) * Extracted data from 1990-2021 * Checked for null values * Treated null values |
| Housing Market | * Extracted required columns and eliminated the rest * Extracted data from 1990-2021 * Extracted required rows as some columns were categorical * Checked for null values * Treated null values |
| CO2 per capita | * Extracted required columns and eliminated the rest * Created a calculated column to convert measures to standard form (normalized by population) * Extracted data from 1990-2021 * Checked for null values * Treated null values |
| Migration Rate | * Extracted required columns and eliminated the rest * Extracted data from 1990-2021 * Checked for null values * Treated null values |
| Air Travel | * Converted data format to the standard followed by other datasets * Country renaming in case of multiple names * Extracted required columns and eliminated the rest * Extracted required rows as some columns were categorical * Extracted data from 1990-2021 * Linear Imputation to fill missing data * Data type of columns set right * Checked for null values * Treated null values |
| Internet Usage | * Extracted required columns and eliminated the rest * Extracted data from 1990-2021 * Linear Imputation to fill missing data * Checked for null values * Treated null values |
| Fertilizer Consumption | * Extracted required columns and eliminated the rest * Extracted data from 1990-2021 * Linear Imputation to fill missing data * Checked for null values * Treated null values |
| Meat Consumption | * Extracted required columns and eliminated the rest * Extracted required rows as some columns were categorical * Extracted data from 1990-2021 * Linear Imputation to fill missing data * Checked for null values * Treated null values |

#### Data Fusion

There were many missing values in the datasets.

Data from 1990-1996 was missing in the migration dataset. Hence, extrapolation technique was employed to fill the gap.

Data was missing in between for the rest of the data. Linear interpolation 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 was filled, the features from all the datasets were fused together on the basis of country and year. Initially, 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 1022x21 with 40 countries (< > % of developing and < > % of developed countries).

### Trend Visualization



<Comment on trend before and during covid>

## Which human activities are correlated with climate change?

In the next step, feature selection is performed to eliminate redundant features and retain only those features with high predictive power for surface temperature prediction.

### Which correlation tests should we employ? Why?

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

* Feature Variance
* Pearson Correlation Matrix
* ANOVA f-score
* Univariate Linear Regression
* Feature Importance using sequential backward selection
* Mutual information based feature selection

Hence, these feature selection techniques were adapted to prepare the final list of features to feed into the predictive model.

#### 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 Matrix

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

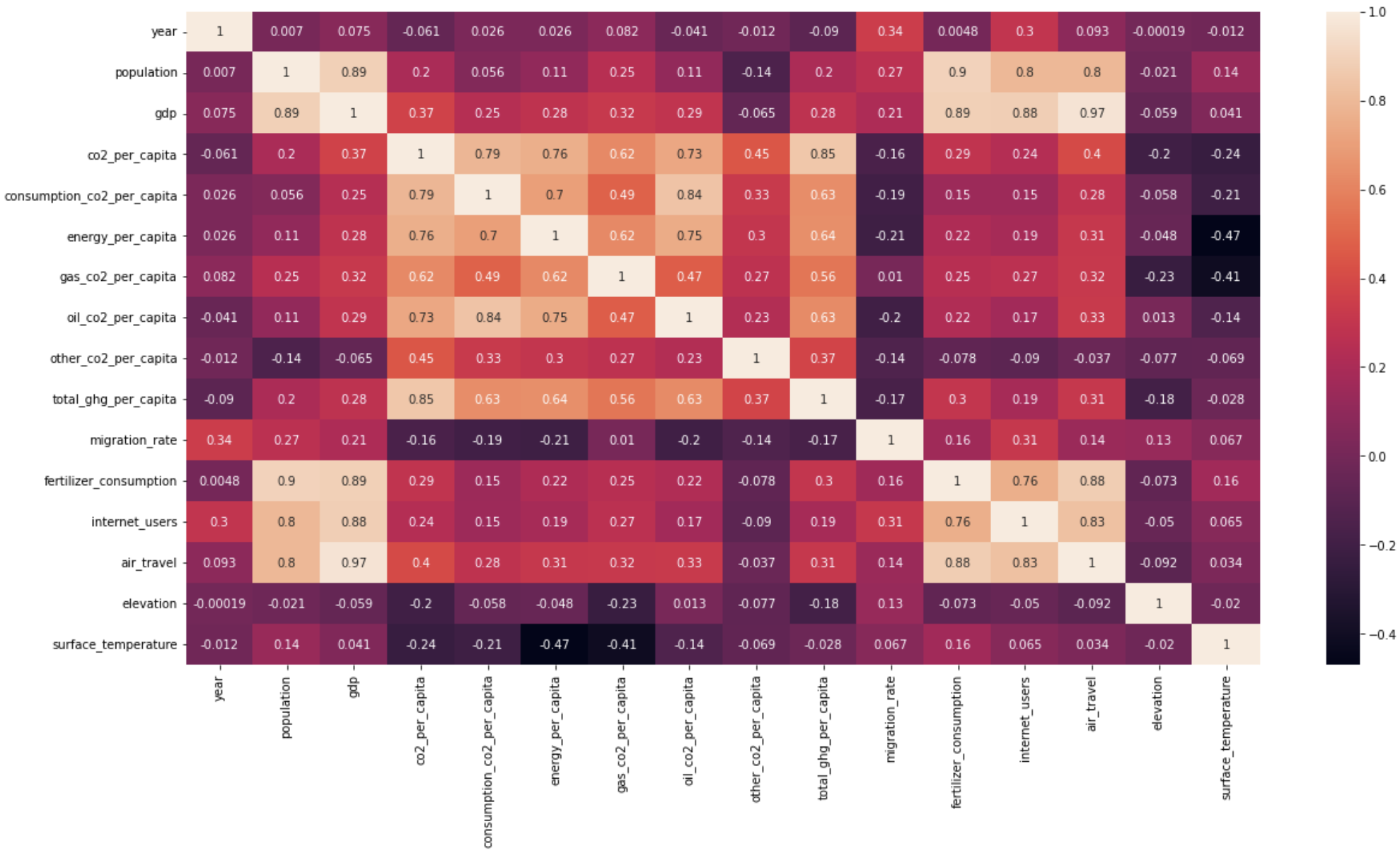


Table shows the highly correlated features and corresponding decisions made on feature elimination and retention.

| **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 the 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. Since GDP shows high correlation with quite a few features, it is eliminated as a redundant feature. |
| * Total\_ghg\_per\_capita * co2\_per\_capita | These two energy-related features show 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. |

#### ANOVA f-score

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

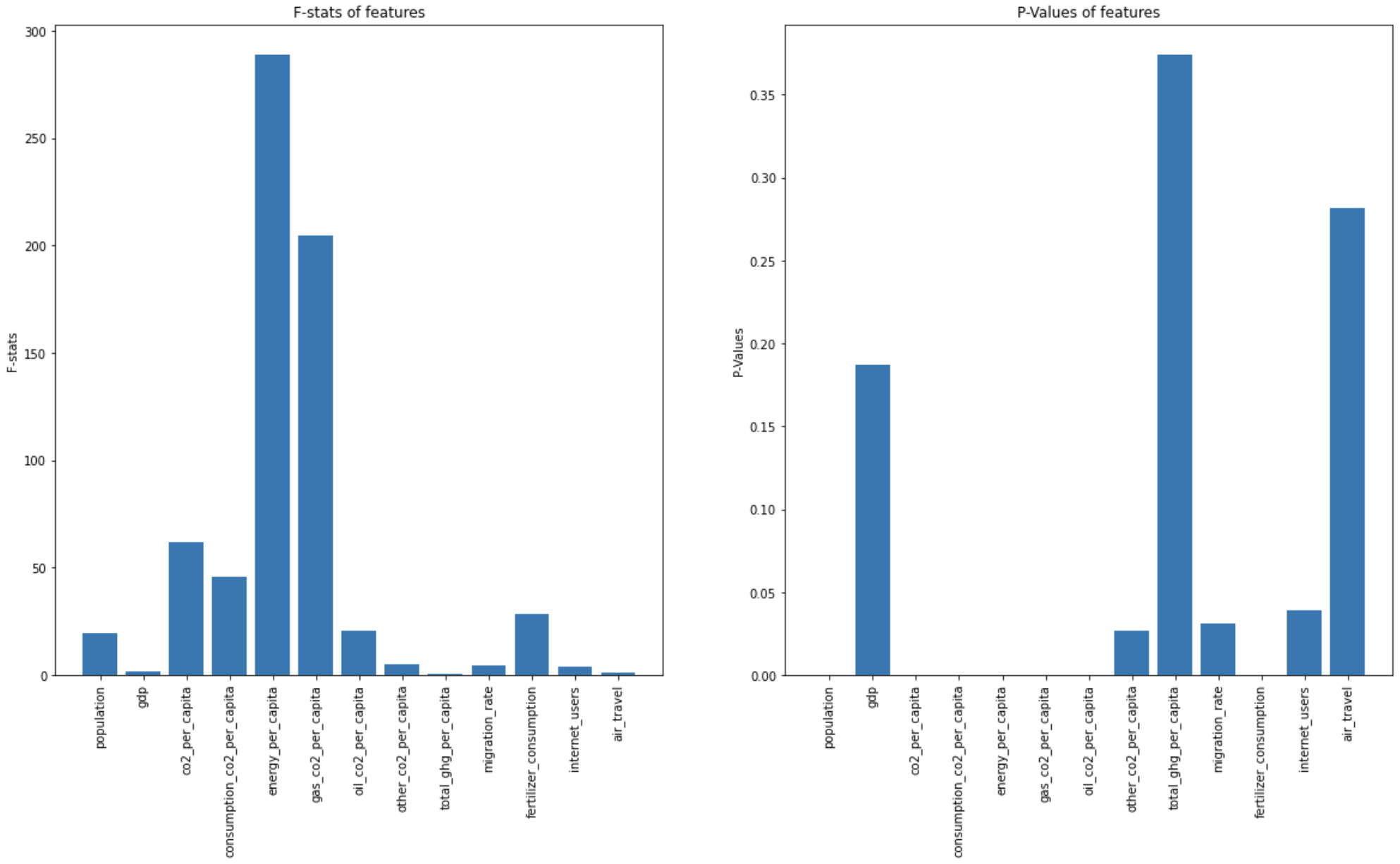
| **Feature** | **F-stats** | **P-values** |
| --- | --- | --- |
| Population | 3964.9362 | 0.00 |
| GDP | 635.757 | 0.00 |
| CO2\_per\_capita | 333.973 | 0.00 |
| Consumption\_CO2\_per\_capita | 172.404 | 0.00 |
| energy\_per\_capita | 771.395 | 0.00 |
| gas\_CO2\_per\_capita | 341.497 | 0.00 |
| oil\_CO2\_per\_capita | 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 |

It can be inferred from the table that F-values are high for most of the features, except migration rate and internet usage. Hence, the variance between the features and 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.

#### Univariate Linear Regression

Univariate Linear Regression was performed to understand the relationship between features and surface temperature.

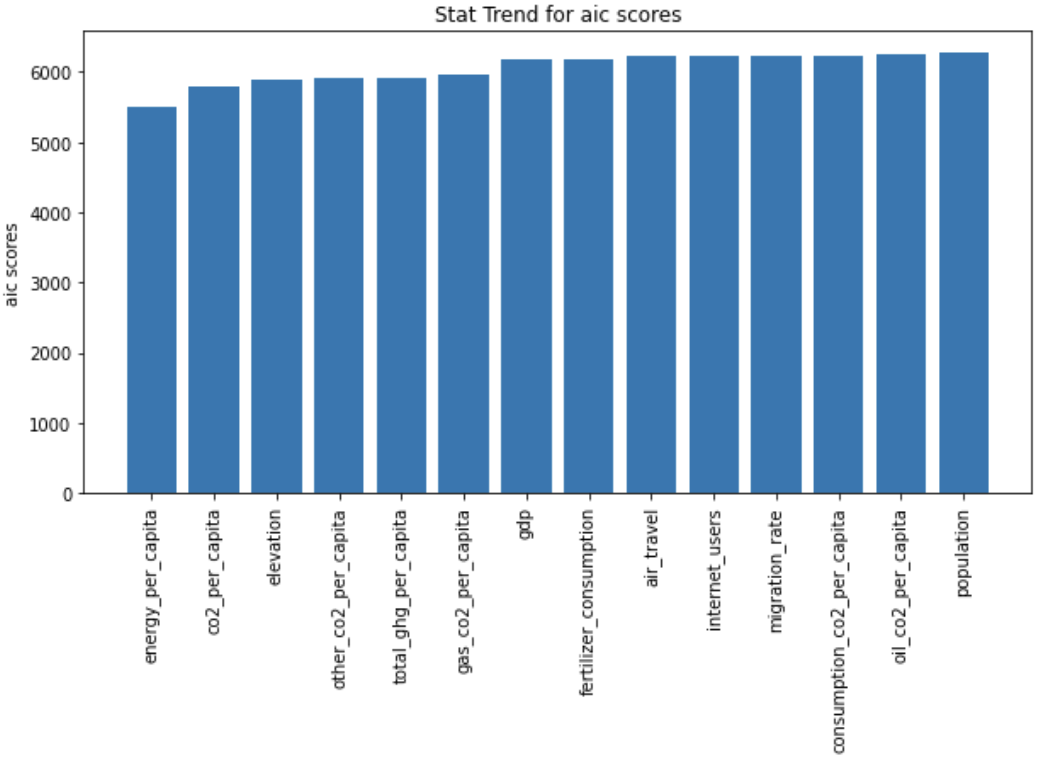


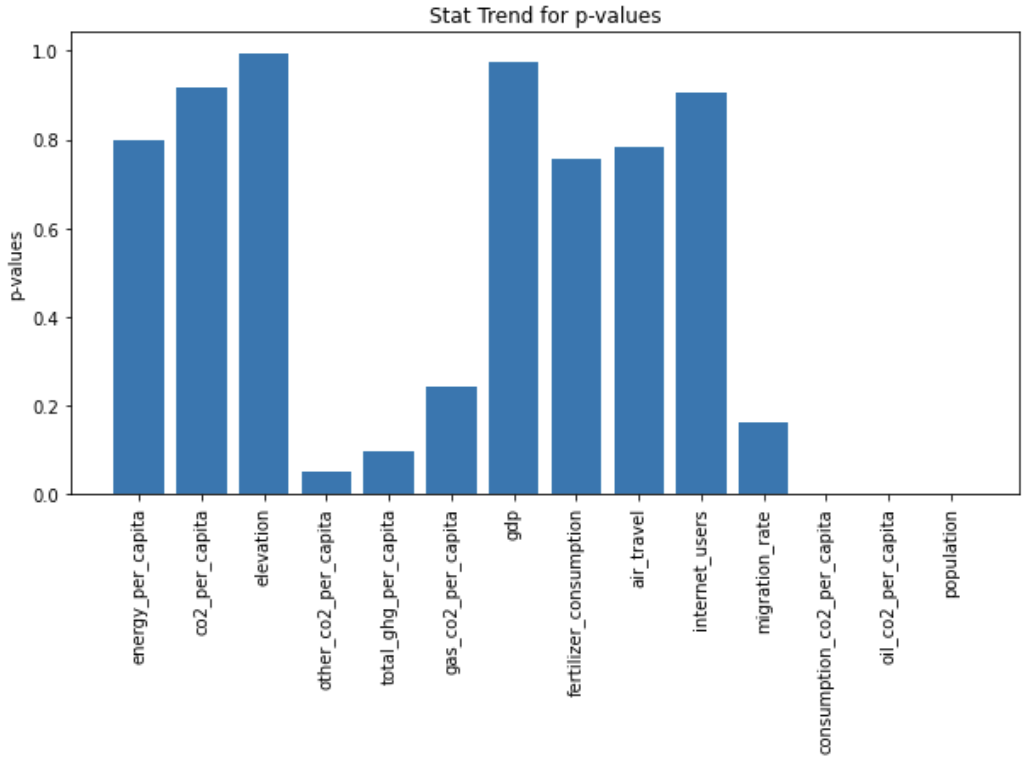
The first graph shows high f-values for most of the features, establishing strong relationship between features and surface temperature. The p-values of all the features except gdp, total\_ghc\_per\_capita and air\_travel are less than 0.05 in the second graph. Hence, this is good evidence to prove the strong relationship.

#### Feature Importance using Sequential Backward Selection

Features were also ranked by employing a sequential backward selection algorithm. The aic scores and p-values of results are visualized in the following charts.

The bar corresponding to a feature represents the stat obtained with the set of features to the right of the current feature inclusive. That is, the stat obtained after removing all the features to the left of the current one.





It can be observed in the first graph that the aic scores of models gradually increases as features are removed. However, the statistical significance of these scores is established by the corresponding p-values in the second graph. As can be seen, p-values are less than 0.05 for the following sets – (other\_co2\_per\_capita to population), (consumption\_co2\_per\_capita to population), (oil\_co2\_per\_capita to population) and (population).

Both co2\_per\_capita and energy\_per\_capita point to similar human activity results. Ignoring the elevation feature which is not human activity, most of the features together produce a high aic score that is statistically significant. Hence, the chosen human activities hold high predictive power.

#### 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.

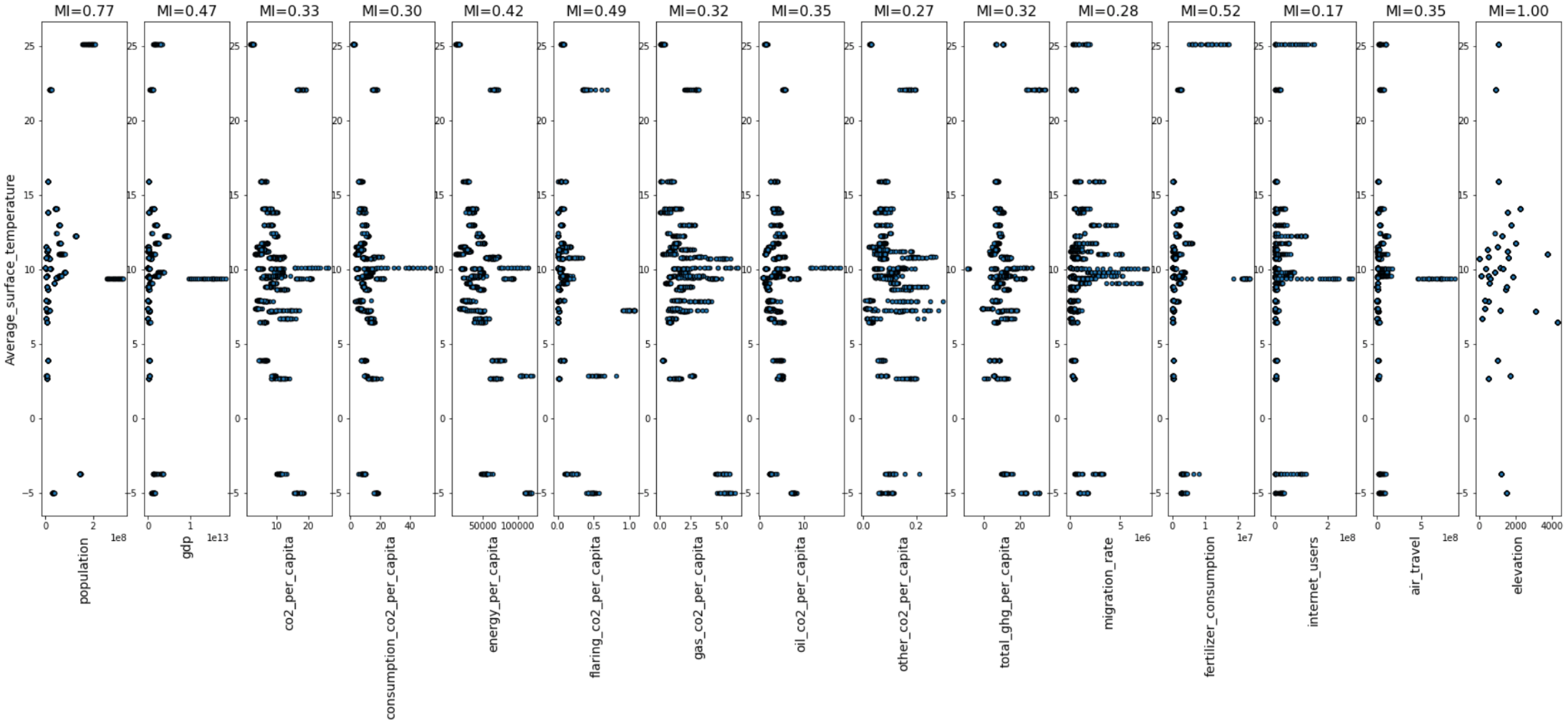


Table shows the mutual information scores of features in the ranked order.

| **Features** | **Mutual Information Score** |
| --- | --- |
| Elevation | 1.00 |
| 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 users | 0.17 |

Treating these various feature selection methods as an ensemble, the following features were removed:

* Flaring\_co2\_per\_capita
* Total\_ghc\_per\_capita
* energy\_per\_capita

#### Country-wise trend of features

The chart below shows the trend of a few features countrywise.

**Legend:**

Blue: CO2\_per\_capita

Orange: migration rate

Green: Air travel

Red: Fertilizer consumption

Violet: Surface Temperature



## How do we model the causal relationship?

Summary - explanations in sub-questions.

### What should be the complexity of the model?

The size of the final dataset at this point is 15,330x16. Complex models might overfit the data. Hence, we use basic statistical and machine learning models both individually and as an ensemble for average surface temperature prediction.

The models used are –

* Decision Tree for regression
* Random Forest Regressor
* Support Vector Regressor
* Lasso Regression

The evaluation metric used are **mean squared error and Mean Absolute Percentage Error**.

### What is the performance of models?

| **Model** | **Mean Squared Error** | **MAPE** |
| --- | --- | --- |
| Decision Tree for regression | 0.92 | 1.52 |
| Random Forest Regressor | 0.03 | 0.68 |
| Support Vector Regressor | 2.78 | 15.94 |
| Lasso Regression | 21.36 | 40.68 |

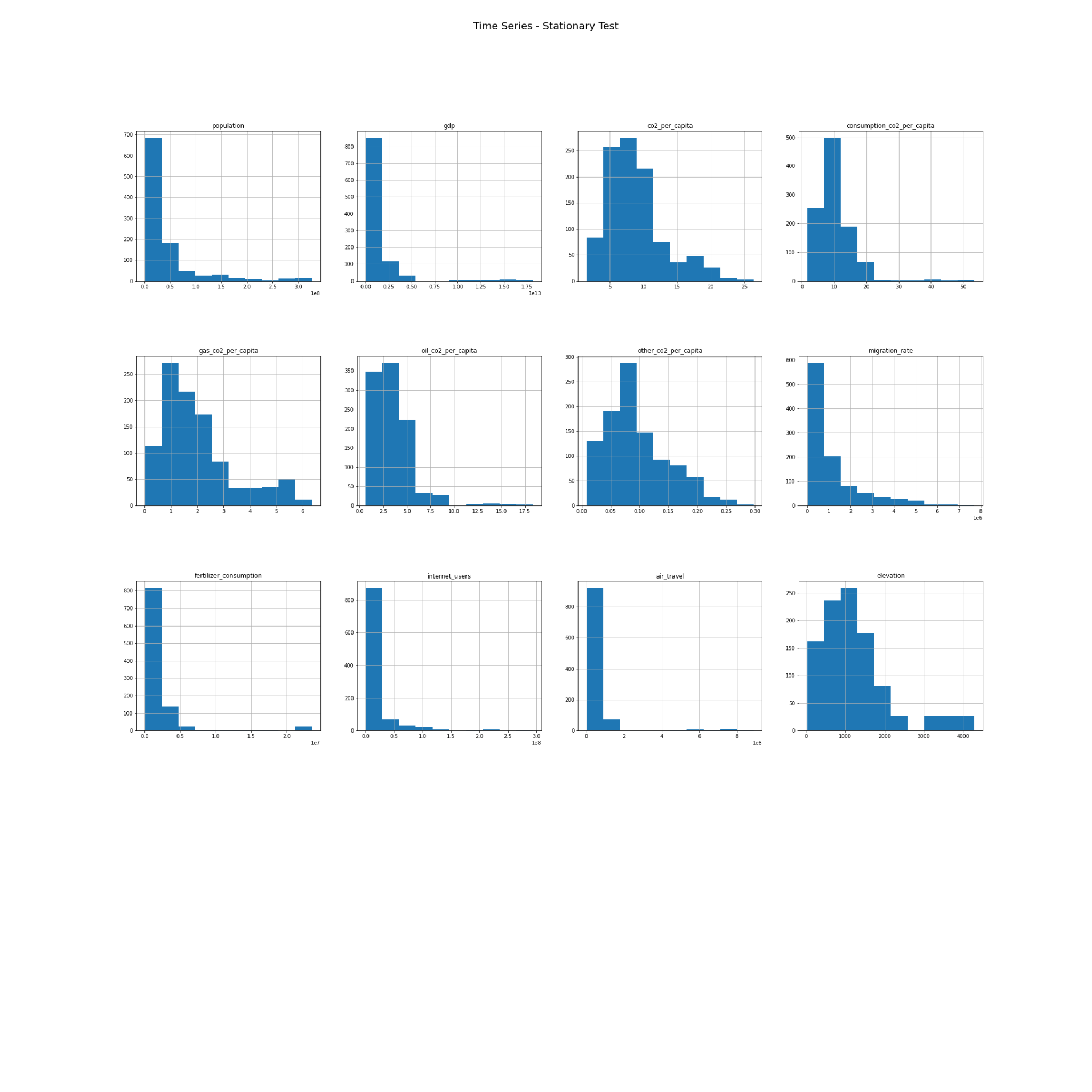
The two models, Decision Tree for regression and Random Forest Regressor, show low error values compared to the rest. So, we use these two models as our predictors.

### Test for Causation

We also performed Granger Causality and Reverse Causality between the features and average surface temperature. 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.

#### Qualitative Stationary Test

First, we tested if the time series for each feature is stationary by plotting a histogram.



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

#### 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. The ones highlighted in the table are non-stationary.

| **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 | 1497118484054293134442496.00 | 9835841353583598376058880.00 |
| 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 users | 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 |

#### Statistical Stationary Test

In addition to quantitative and qualitative, we also performed Augmented Dickey-Fuller test to test for stationarity of time series of features. The features with p-value > 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.

| 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 users | -1.18 | 0.68 |
| Air travel | 2.76 | 1.00 |
| elevation | -3.99 | 0.0015 |
| surface\_temperature | -4.67 | 0.000096 |

From the three tests, we conclude that the time series of population, gdp, fertilizer consumption, internet users and air travel are non-stationary.

#### Cointegration Test

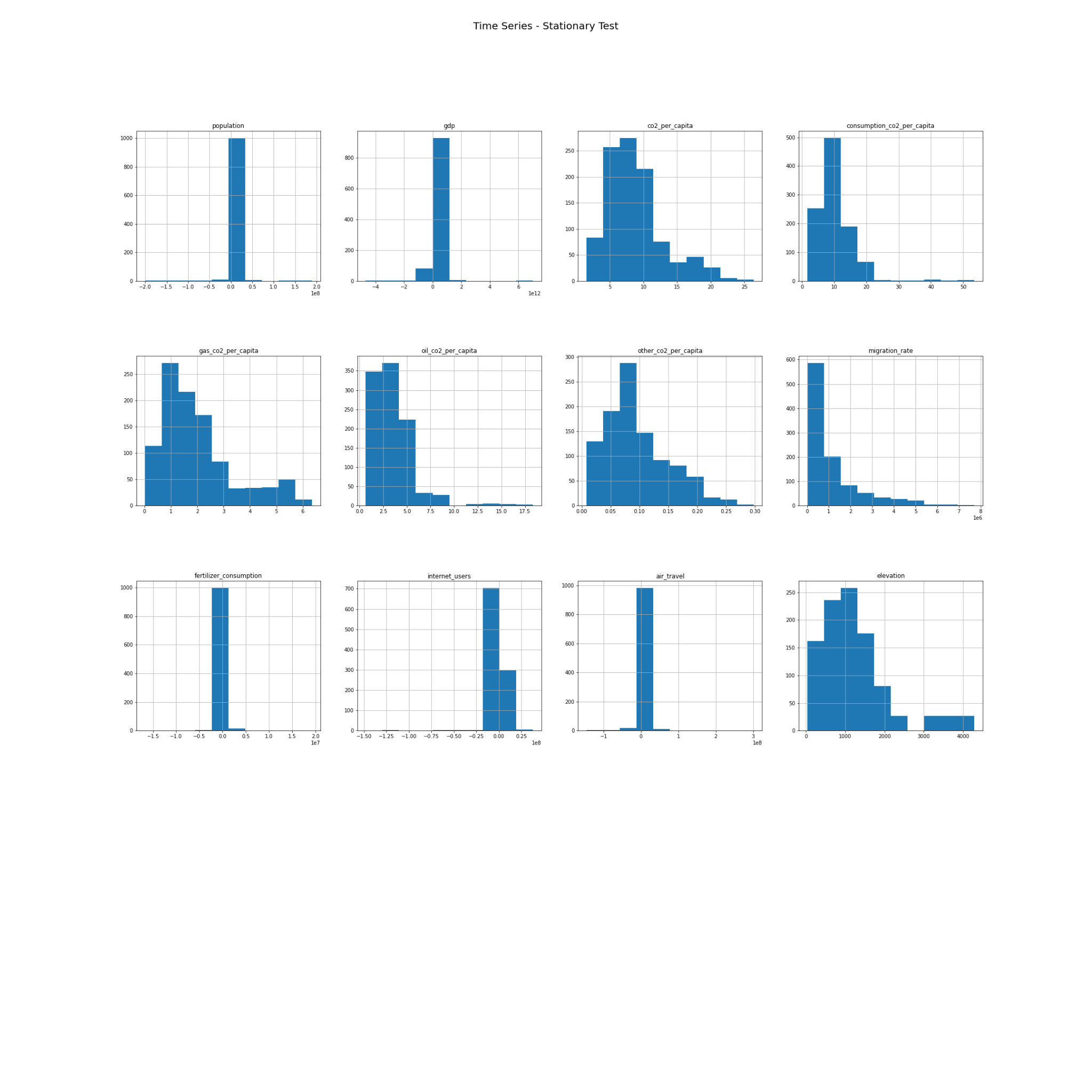
Cointegration test was performed to identify human activities with non-stationary time series that are cointegrated with the surface temperature.

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

| **Non-stationary time-series** | **Cointegration p-value** |
| --- | --- |
| population | 0.61 |
| GDP | 1.0 |
| Fertilizer consumption | 0.62 |
| Internet users | 0.85 |
| Air travel | 1.0 |

#### Non-stationary to Stationary

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. The histogram of the final data with stationary time series looks like below:



#### Granger Causality Analysis

Granger Causality Test was run on the stationary time series data of features. A maxlag of 1 was chosen as we are interested in the immediate causality. As can be seen from the table, both granger 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.

| 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 |

# Analysis II - What do we infer from the result?

## How do we prioritize issues/factors?

We observed a strong correlation in time series pattern between the average rise in temperature and CO2 related features. From observing the trends from the United States plots, oil CO2 per capita drops abruptly from 2008 due to the 2008 economic crisis. Around the same time, the average surface temperature dropped at similar rates. But, 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.

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

Surprisingly, the feature that was equally competent with CO2 per capita is Fertilizer consumption. From 2005 whenever there is a peak in fertilizer consumption, it is observed that there is a peak in average surface temperature. Between 2007 and 2009, there was a huge dip in fertilizer consumption, hence it is possible that the fertilizer consumption and CO2 related factors together have a significant effect on the average surface temperature. 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 the peaks in the average surface temperature. It is also interesting to note that between 2010 and 2018, the peaks in average surface temperatures follow the peaks of migration, with a lag of almost one year. Hence, 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 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 that is the possible effect of average surface temperature on air travel. If true, it could be used as an alternative stock price predictor in the aviation industry.

Internet usage pattern does not show any observable correlation with average surface temperature.

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 have 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.

# Future Work

Include Computer Vision on Satellite Image Processing to predict the rate of afforestation/deforestation.

# References

Citations and links

# Notes

Additional pointers if any.

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# Archives

## Proposed Project Ideas

### Prediction of S&P index using three kinds of alternate data

**Idea**

The stock market is a complex concept that could be potentially affected by a multitude of factors. Three main factors, which we treat as alternate data, in addition to historical data, include –

1. People: Social Media Sentiment Analysis (past literature work)
2. Economy: unemployment and mental health stats
3. Regulatory Institutions: SEC amendments

**Significance**

*Past Work considered only the people factor. Our project would extend the scope to the other two factors that could have a great impact on the S&P index.*

**Datasets**

1. [Stock Market – S&P index](https://finance.yahoo.com/quote/%5EGSPC/history/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAM9NWYtuPiVws73khkFLY0-eXc6_Di4syUI8McKAZ7NzqHIV4FzjXUYCbZ9r1-gL43RYq0T8-cQE7kpp2qmg80a0cppeL89iu9pQ8cutEVk0VdLQU-AntjslebAOWefdsEiYZM3q9XDJBSlcWxG_6crEF4jtkl0AaYw8olNyt08c) (check past study for window estomation)
2. People: (Media Data will be scraped from respective pages)
   1. Twitter (Informal Media)
   2. Reddit (Informal Media)
   3. News Articles (Formal Media)
3. Economy
   1. Suicide-related Search ([google trends](https://trends.google.com/trends/explore?q=suicide&geo=US))
4. Regulatory Institutions
   1. SEC Amendments ([SEC proposed rules](https://www.sec.gov/rules/proposed.shtml))

### Predicting Climate Change Risk Index of Countries based on CO2 and Greenhouse gas emissions and GDP

**Idea**

Carbon emission is one of the major contributors to the climate change issue. In any region, one of the indicators of the climate change risk index is GDP. If the GDP is high, it indicates the usage of more energy and fuel.

Changes in the GDP in the country provides the demand factor which indirectly points to high energy usage.

Mobility based on covid period less emission of gas. How to measure mobility?

Sea-level rise

Electricity consumption

Sales of cars (Electric vs Gas)

Factories

Wildfire rate

Satellite imagery - rate of afforestation/deforestation

**Significance**

*The Climate Change Risk Index is currently calculated based on historic impact and vulnerability data of the region. Our project taps into alternate data of potential contributing sources to global surface warming.*

**Datasets**

1. [Climate Change Risk Index](https://data.world/gpsdd/e1dcef1d-b9ca-4c22-8b78-f7b6703d2274) - http://berkeleyearth.org/data/
2. [GDP](https://data.worldbank.org/indicator/NY.GDP.MKTP.CD)
3. [CO2 and Greenhouse gas Emission](https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions)
4. Global WildFires - <https://gwis.jrc.ec.europa.eu/apps/country.profile/downloads>
5. [Global Car Sales](https://datasource.kapsarc.org/explore/dataset/world-motor-vehicle-sales-by-country-and-type/table/?disjunctive.date&disjunctive.country_name&disjunctive.indicator_name&refine.date=2005&refine.date=2021&sort=-date)
6. [Average elevat\ion of the country](https://www.atlasbig.com/en-us/countries-average-elevation)

**Climate Models References:**

1. <https://www.metoffice.gov.uk/research/approach/modelling-systems/unified-model/climate-models/hadcm3>

### Identifying the predictive features for US States GDP prediction in the post-covid period

**Idea**

Given the ‘Work From Home’ era, employees are not tied to staying in or near their work locations. Hence, we hypothesize that the impact of regional non-corporate related factors like crime rate, taxes etc. on the GDP of the region has increased (Personal Consumption Expenditure). Our project would answer the question of what are the main factors people prioritize while deciding a place to settle in. These factors, we believe, would help the government decide what to focus on to have a significant positive impact on the corresponding GDP.

**Significance**

*The impact on GDP before covid was highly biased towards employment opportunities in the region. Currently, every region has almost a fair chance at improving the quality of life for a significant impact on the GDP. Hence, our project would provide helpful suggestions for different regions on what to improve.*

**Datasets**

1. GDP of US States ([US Department of Commerce](https://www.bea.gov/data/gdp/gdp-state))
2. Crime Rate ([FBI Crime Data Explorer](https://crime-data-explorer.fr.cloud.gov/#))
3. Real-estate market ([US Census Bureau](https://www.census.gov/quickfacts/geo/chart/US/HSG860220))
4. Number of schools ([education](https://hifld-geoplatform.opendata.arcgis.com/datasets/87376bdb0cb3490cbda39935626f6604_0))
5. Healthcare ranking ([US News](https://www.usnews.com/news/best-states/rankings/health-care))
6. Restaurants - ([Inspection score](https://www.muni.org/Departments/it/Pages/opendata.aspx))

(This is an evolving list of factors to consider).

## Useful Links from Professor

<https://link.springer.com/article/10.1007/s13278-021-00723-5>

<https://covid19.apple.com/mobility>

<https://scied.ucar.edu/learning-zone/climate-change-impacts/predictions-future-global-climate>

<https://www.climate.gov/news-features/understanding-climate/climate-change-global-sea-level>

<https://github.com/KKulma/climate-change-data>

[Foresight study: Identification of datasets for impacts of climate change](https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5bedabc44&appId=PPGMS)

<https://portal.311.nyc.gov/article/?kanumber=KA-01017>

<https://www.sciencedirect.com/science/article/pii/S0160412019309341>

<https://www.liebertpub.com/doi/full/10.1089/big.2014.0026>

<https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=%22Climate+Change%22++relationship+dataset&btnG=>

<https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0002503>

Siyi Sun <[ss15463@nyu.edu](mailto:ss15463@nyu.edu)>

<https://www.liebertpub.com/doi/full/10.1089/big.2014.0026> -