Predictive Data Analysis on Average Global Surface Temperature

Ravi Aishwarya
Department of Computer Science and Engineering
PES University
PES2UG19CS322
aishwarya2447@gmail.com

K Keerthana
Department of Computer Science and Engineering
PES University
PES2UG19CS170
keshavakeerthana@gmail.com

Keerthi Joshi K
Department of Computer Science and Engineering
PES University
PES2UG19CS181
keerthi.joshi2822@gmail.com

Naveen S Nelogal
Department of Computer Science and Engineering
PES University
PES2UG19CS249
naveenelogal143@gmail.com

Abstract:

The data for the Average Global Surface Temperature time series is analyzed and forecasted in this research. There are two types of ARIMA models: basic ARIMA and trend based ARIMA. We utilized simple ARIMA to forecast the world average temperature. It has been discovered that the Trend based ARIMA approach outperformed the basic ARIMA method among all linear models. MAPE (Mean Absolute Percentage Error), RMSE (Root Mean Squared Error), and MSE (Mean Squared Error) have all been used to evaluate the model's accuracy.

Keywords: Time Series Forecasting, Trend-based, data preprocessing, data forecasting, Average global temperature, ARIMA, SARIMA.

I. INTRODUCTION:

Prediction Data Mining, Data Analysis are the predominant phases in any time series model. Descriptive data mining is a technique for extracting knowledge from a set of data. Predictive data mining is a technique for forecasting data that may be utilized in a variety of ways. Time series forecasting is an important field of machine learning that is sometimes neglected. It's significant since there are several prediction issues with a time component. Such data may be found in a variety of applications, including climatic changes, commodity production, geographical data, sensor data, stock market data, inventory control, and so on.

For pre-processing the raw time-series data, methods such as Data Cleaning (imputing missing values with mean for numerical data and mode for categorical data), Data Normalization, Data Standardization for Data Transformation, and Dimensionality Reduction (Principal Component Analysis) for

Data Reduction can be used. Then predictive techniques can be applied to the same data to forecast the model. A model's accuracy after applying these strategies is higher than raw data prediction.

ARIMA models are widely used in a variety of applications. Electricity consumption, heart rate monitoring for medicinal purposes, and regional weather conditions. It's also used to analyze mobile user data and forecast employment in financial applications like the stock market.

Time series are not just things that is completely relied on time. There are a handful of components that make them the way they are:

- In the **Trend**, are the values in the time series getting higher or lower?
- If there is a pattern in the dataset that is determined by seasons, this is known as **Seasonality**. Are online sales, for example, higher during the holidays, festivals, or ice cream sales higher in the summer and lower in the winter?
- A **Cycle** allows us to see a certain changing pattern that is out of the usual for the time span. What other arbitrary variables may impact the data?
- **Irregularity** is what other arbitrary factors might influence the data?

Artificial neural networks (ANNs) have been employed in the literature for forecasting time series in recent years. ANN (Artificial Neural Networks) and other nonlinear approaches have been used in a variety of applications. Nonlinear approaches, on the other hand, are more complicated. Although ANNs can simulate both linear and nonlinear patterns in time series, they are not equally capable of operating both structures.

II. RELATED WORK:

Climate change is a science that has been researched for a long time. The many research articles propose several models and approaches produced and examine the best model among them for predicting and forecasting weather as well as acting as a warning system to notify people to hazardous weather conditions.

The main issue with climate change models is that producing all the results for global temperatures takes a long time. Weather may change at any time, making it impossible to predict the future properly. If anything happens that the model didn't predict, it can't estimate since there's no guarantee that it'll remain stationary, which is one of the primary drawbacks. Humans can respond to weather changes and make forecasts in response, whereas the weather prediction should re-run to integrate new data, which takes time.

Operational meteorology is the practice of weather forecasting. As a result, forecasting issues may be divided into two categories: meteorological and operational.

The availability, timeliness, and quality of observational data, time limitations on forecast development, and the nature and reliability of observational data are all issues.

III. APPROACH OVERVIEW AND METHODOLOGY:

1. Data Collection

   The Berkeley Earth, which is partnered with Lawrence Berkeley National Laboratory, provided the data for this study. The Berkeley Earth Surface Temperature Study brings together 1.6 billion temperature reports from 16 repositories. It's well-packaged and easy to slice into interesting subsets (for example by country, city). Average land temperature

changes from 1743 to 2015 have been analyzed in this literature. The forecast of this time series data is done using the above-said techniques and the performance is calculated. The following procedures were adopted at this stage of the research: Data Cleaning, Data Selection, Data Transformation.

2. Data Selection

Data for the analysis was determined on and retrieved from the dataset at this point. Table 1 shows the kind and description of the seventeen (17) attributes in the meteorological dataset, while Table 2 shows an analysis of the numeric values.

Table 1: Meteorological Dataset Attributes

Attributes	Type	Description
dt	Object	Starts in 1743 for average land temperature and 2015 for maximum and minimum land temperatures and global ocean and land temperatures
AverageTemperature	Float	Average temperatures in celsius degrees
AverageTemperatureUncertainty	Float	95% confidence interval around the average temperatures
Country	Categorical	Country considered
city	Categorical	City considered
Latitude	Categorical	Latitude considered
Longitude	Categorical	Longitude considered
state	Categorical	State considered
LandAverageTemperature	Float	Global average land temperature in celsius
LandAverageTemperatureUncertainty	Float	95% confidence interval around the land average temperatures
LandMaxTemperature	Float	global average maximum land temperatures in celsius
LandMaxTemperatureUncertainty	Float	95% confidence interval around the maximum land temperatures
LandMinTemperature	Float	global average minimum land temperatures in celsius degrees
LandMinTemperatureUncertainty	Float	95% confidence interval around the minimum land temperatures
LandAndOceanAverageTemperature	Float	Global average land and ocean temperatures in celsius degrees
LandAndOceanAverageTemparature uncertanint y	Float	95% confidence interval around the average land and ocean temperatures

Table 2: Analysis of Numeric data values

Variable	Min	Max	Mean	SD	Missing Values
AverageTemperature	-4.538	3.884	1.369	1.287	72493
LandAverageTempature	-2.08	19.021	8.3747	4.3813	1462326
LandMaxTemperature	5.9	21.32	14.35	4.309	1463514
LandMinTemperature	-5.407	9.715	2.743595	4.1558	1463514
LandAndOceanAverageTemperature	12.475	17.611	15.2125	1.274	1463514

3. Data preprocessing

i. Data Cleaning

At this step, a data model compatible format was created, which handles of missing data, detecting duplicate data, and removing bad data. Finally, the cleansed data was transformed into a format that could be used for data mining.

a) Removing duplicates or irrelevant observations:
Remove any unnecessary observations from our dataset, such as duplicates or irrelevant observations. Duplicate observations are most likely to occur during data collection. There is a risk of generating duplicate data when we merge data sets from various files or sources, scrape data, or receive data from several sectors. One of the far zones to be examined in this procedure is de-duplication. When we see observations that don't fit the problem we're trying to solve, we call them irrelevant observations.

b) Filter unwanted outliers:
There will frequently be one-off observations that do not materialize to suit within the data we're examining. If we have a good cause to remove an outlier, such as incorrect data input, we can improve the performance of the data we're working with. However, the presence of an outlier is less likely to validate a hypothesis we're working on.
Any point of data in a data collection that is more than 1.5 IQRs below the first quartile (Q1) or above the third quartile (Q3) is considered an outlier.
High = (Q3) + 1.5 IQR
Low = (Q1) – 1.5 IQR

c) Handling missing values:
In the dataset, missing values are imputed using the mean for numeric data and the mode for categorical data.

ii. Data Transformation

Data consolidation is another term for this. It's the process of transforming the chosen data into a format appropriate for data mining process. The data were

normalized to reduce the effect of scaling on the data, and the data file was saved as a Commas Separated Value (CSV) file format.

a) Normalization

Normalization is the process of rescaling data from its original range so that all values fall between 0 and We must know or be able to properly assess the lowest and maximum observable values to normalize. We may be able to predict these numbers based on the information you provide. If the time series is moving up or down, it may be difficult to forecast these predicted values, and normalizing may not be the best solution.

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

b) Standardization

Rescaling the distribution of values so that the mean of observed values is 0 and the standard deviation is 1 is the process of standardizing a dataset. This is similar to removing the mean value or centering the data.

$$Z = \frac{x - \mu}{\sigma}$$

c) Smoothening

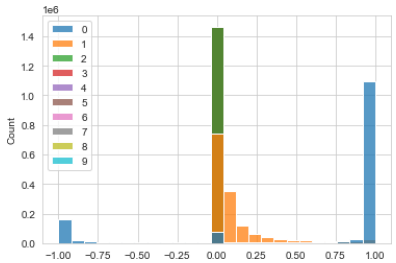
Data smoothing is a techniquefor detecting trends in noisy data without knowing the nature of the trend.

The Exponentially Weighted Moving Average (EWMA) is a statistical or quantitative metric that may be used to represent or characterize a time series. The parameter alpha is the only choice an EWMA user must make. The value of this parameter determines how relevant the current observation is in the EWMA computation. The greater the alpha value, the closer the EWMA follows the original time series

$$\text{EWMA}_t = \alpha * r_t + (1 - \alpha) * \text{EWMA}_{t-1}$$

iii. Data Reduction

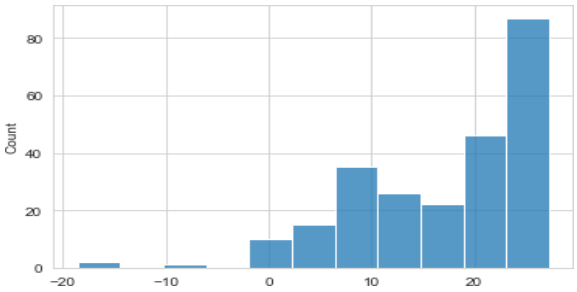
PCA is one of the most effective strategies for constructing a new collection of variables from a large number of existing ones.Principal Components are the new variables that have been extracted. The original variables are combined into a principal component, which is a linear combination of them.



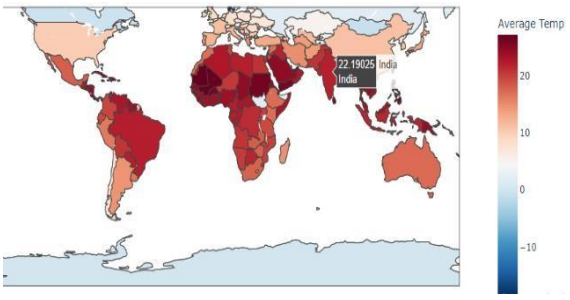
- The first principal component is extracted in such a way that it indicates the maximum variation in the dataset.
- The second principal component, which is uncorrelated to the first, aims to explain the remaining variance in the dataset.

4. Data Visualization

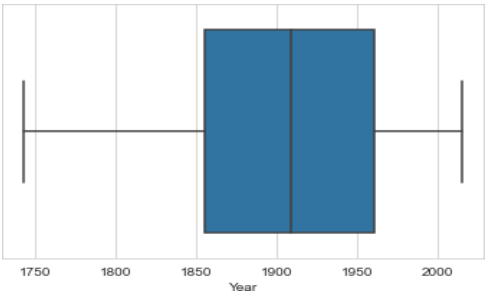
A Data visualisation is the graphical depiction of information and data. Techniques for data visualisation make it simple to identify trends, outliers, and patterns in data by using visual components like piecharts,graphs, and globalmaps.



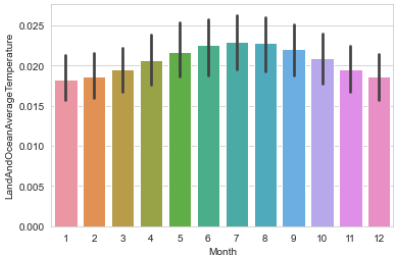
Average Temperature for each country



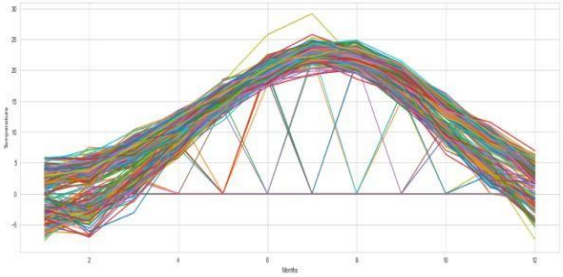
Average Land Temperatures in Countries



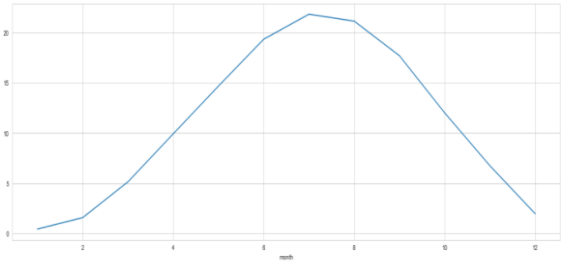
Year vs FirstTempDifference



Month vs LandOceanAverageTemperature



Month vs Temperature in the period (Seasonal)

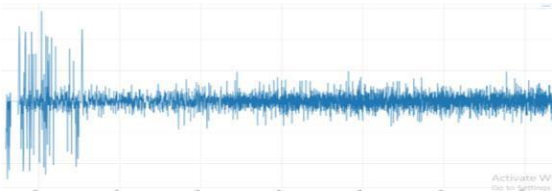


Month vs AverageTemperature(Seasonal)

IV. BUILDING AND FORECASTING MODEL:

1) Stationary- Differencing and Augmented Dickey Fuller Test

The time-series data does not have to be stationary all the time. An ARIMA model cannot be fitted in this case. So, first, the data is made stationary by executing the Dickey-Fuller test to see if it is stationary, and then the differencing operation is performed. The letter 'd' stands for the number of times differencing is performed.

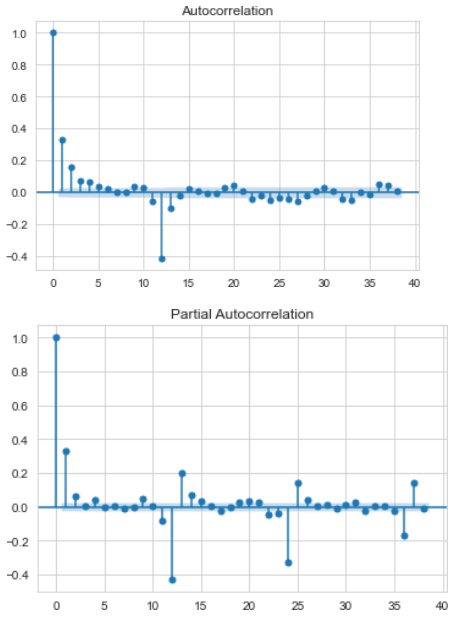


2) ACF and PACF plots

The autocorrelation function and the partial autocorrelation function are used to determine the values of p and q on the differenced data. The autocorrelation function (ACF) is a statistical approach for determining how closely values in a time series are related to one another. The partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags, is known as the PACF.

It is a pure AR process of order p if the ACF shows a sinusoidal decay and the PACF gets zero after a lag p.

If the ACF is zero after a lag q and the PACF decays sinusoidally, the process is an MA of order q. It becomes an ARMA if both ACF and PACF have sinusoidal decay and become zero after discrete delays q and p.



3) Model Implementation

a) Basic ARIMA:

Yt=rY(t1)+et is the form of the AR(1) model ARIMA(1,0,0), where r is the autoregressive parameter and et is the pure error component at time t.

The equation is **Yt=rY(t1)+et+ae(t1)** for ARIMA(1,0,1),

where an is the moving average parameter.

After obtaining the model order (p, d, q), the parameters a1,a2,...ap, b1,b2,...bp are determined using estimate methods. The time series data may be forecasted using methods such as the Least Squares Method, Maximum Likelihood, and others, utilizing these parameter estimations and error variance. This method may be used to a wide range of time series data. After the time series data is made stationary, it has a large variance, thus a logarithmic transformation is applied to the raw data to reduce the error variance. To go back to the raw data predictions, the

differenced data must be merged. The Auto Regressive "Integrated" Moving Average (ARIMA) model is the term given to this model.

b) Trend Based ARIMA:

Smoothening is a pre-processingstep before predicting data mining on the provided time series data in this approach, which is a composite methodology. Initially, a moving average linear filter is utilized on the raw data to make it smooth. This smoothing filter can be found in the (2). As a result, the derived smoothed data is the data's trend.

$$s_t = \frac{1}{2p+1}\sum_{i=-p}^{p} y_{t+i} \qquad (2)$$

In (2), S_t denotes the trend component, Y_t denotes the raw data, and (2p + 1) denotes the filter length. The residual data is extracted after subtracting the trend data from the original data. The raw data is now divided into two categories: trend and residual data components. This equation can be found in (3).

$$\mathbf{R_t = Y_t - S_t} \qquad (3)$$

c) Seasonal ARIMA:

SARIMA (Seasonal ARIMA) is an extension of ARIMA that explicitly allows univariate time series data with a seasonal component.

It includes three new hyperparameters for the seasonal component of the series: autoregression (AR), differencing (I), and moving average (MA), and an extra parameter for the seasonality period.

The parameters for these sorts of models are as follows:

SARIMA(p,d,q)x(P,D,Q,s)

p and the season P denote the number of autoregressive terms in the model (lags of the stationarized series)

d as well as seasonal D: specify the amount of differencing required to stationarize series q and seasonal Q: indicate the number of moving average terms (lags of the forecast errors)

s: denotes the data's seasonal duration.

V. EXPERIMENTAL RESULTS AND ANALYSIS

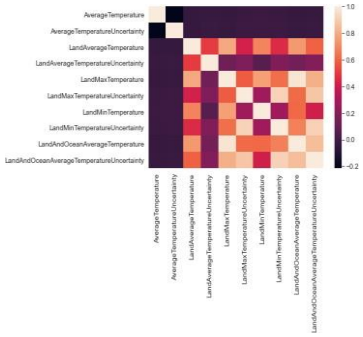
Evaluation metrics:

The following performance measures were utilized to identify the suitable algorithms and parameters that best model the weather forecasting variable.

Model	RMSE	MSE
ARIMA	2.264	5.126
SARIMA	4.643	21.563

a. Correlation coefficient:

The statistical correlation between the expected and actual values is measured here. This approach is unique in that it does not alter when the test cases' values are scaled up or down. A higher value indicates a more accurate model, with 1 indicating perfect statistical correlation and 0 indicating no correlation at all.



b. Mean Squared Error:
One of the most commonly used measurements of numeric prediction performance is mean-squared error. This number is calculated by taking the average of the squared differences between each estimated value and its corresponding true value.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Y- the vector of observed values of the variable being predicted
$\hat{Y}$ – predicted values

c. Root Mean Squared Error:
It is the square root of the mean-squared error. This root mean-squared error gives the error value the same dimensionality as the actual and predicted values

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\|y(i) - \hat{y}(i)\|^2}{N}}$$

N – No of data points

y(i) – i^{th} measurement

$\hat{y}$ (i) - corresponding prediction

d. Mean Absolute Percentage Error: The mean absolute percentage error (MAPE) is a measure of a forecast model's accuracy. It is expressed as the average absolute percent error for each time period minus actual values divided by actual values, and it may be calculated as the average absolute percent error for each time period minus real values divided by actual values.

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

A_t = actual value

F_t – Forecast value

VI. CONCLUSION:

The data used was for The Berkeley Earth, which is associated with Lawrence BerkeleyNational Laboratory metropolis obtained from the meteorological station between 1743 and 2015. The results show during the study period, these variables had an impact on the weather in these months.Given enough data the observed trend over time could be examined and important deviations which show changes in climatic patterns identified. The primary goal of this study is to create a low-cost, trustworthy, and efficient weather forecast model that may be used in our daily life. The importance of forecasting the weather conditions and its different patterns is observed using time series analysis.
In forecasting the model, ARIMA is a really helpful and widely used statistical method. These studies may have certain limits due to a variety of factors such as a huge dataset, missing data, harsh

circumstances, extreme conditions and so on.

The acquired findings were compared to the test data set prepared in combination with the training data and determined to be satisfactory given the minimal amount of data available for training and testing. A bigger data collection, consisting of data collected over several decades, will be required to achieve a better result. Neuro-fuzzy models will be used in future research projects to predict the weather. The variation in weather conditions in terms of temperature, rainfall, and wind speed may be analysed using these data mining techniques, which is significant for climate change research.

VII. REFERENCES:

[1] J. Li, K. Zhang and Z. Meng, "Vismate: Interactive visual analysis of station-based observation data on climate changes," 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), 2014, pp. 133-142, doi: 10.1109/VAST.2014.7042489.

[2] Ahrens, C. D., 2007, "Meteorology" Microsoft® Student 2008 [DVD], Redmond, WA: Microsoft Corporation, 2007.

[3] Bregman, J.I., Mackenthun K.M., 2006, Environmental Impact Statements, Chelsea: MI Lewis Publication.

[4] Casas D. M, Gonzalez A.T, Rodrígue J. E. A., Pet J. V., 2009, “Using Data-Mining for Short-Term Rainfall Forecasting”, Notes in Computer Science, Volume 5518, 487-490

[5] W. Li, L. Ni, Z. -L. Li, S. -B. Duan and H. Wu, "Evaluation of Machine Learning Algorithms in Spatial Downscaling of MODIS Land Surface Temperature," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 7, pp. 2299-2307, July 2019, doi: 10.1109/JSTARS.2019.2896923.

[6] Due R. A., 2007, A Statistical Approach to Neural Networks for Pattern Recognition, 8th edition. New York: John Wiley and Sons publication.

[7] O'Gorman, P. A., & Dwyer, J. G. (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. Journal of Advances in Modeling Earth Systems, 10(10), 2548-2563.

[8]Olaiya, F., & Adeyemo, A. B. (2012). Application of data mining techniques in weather prediction and climate change studies. International Journal of Information Engineering and Electronic Business, 4(1), 51.

[9]An, X., Ganguly, A.R., Fang, Y., Scyphers, S.B., Hunter, A.M. and Dy, J.G., 2014, August. Tracking climate change opinions from twitter data. In Workshop on Data Science for Social Good (pp. 1-6)

[10]Malik, R. and Pande, S., 2020. Artificial Intelligence and Machine Learning to Assist Climate Change Monitoring. Journal of Artificial Intelligence and Systems, 2(1), pp.168-190.

[11] Elia G. P., 2009, “A Decision Tree for Weather Prediction”, Universitatea Petrol-Gaze din Ploiesti, Bd. Bucuresti 39, Ploiesti, Catedra de Informatică, Vol. LXI, No. 1

[12] Fairbridge R. W., 2007, "Climate" Microsoft® Student 2008 [DVD], Redmond, WA: Microsoft Corporation, 2007