VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

J Component report

Programme : B.Tech

Course Title : Essentials of Data Analytics

Course Code : CSE3506

Slot : G2

Title: Time Series Analysis On Climate Change Dataset

Team Members: Sharon Shibu George | 20BCE1033

Govind Anuj Kumar | 20BCE1526

Sanya Garg | 20BCE1853

G. Narayanee Nimeshika | 20BPS1111

Faculty: Dr. Vergin M

Sign:

Date: 11/4/2023

Abstract

This project compares the performance of Simple Exponential Smoothing (SES), Autoregressive Integrated Moving Average (ARIMA), Prophet, and Naive models in predicting climate change using monthly temperature datasets. The models' performance was evaluated using the Mean Absolute Percentage Error (MAPE) and the Root Mean Squared Error (RMSE). The ARIMA and Prophet models are suitable for predicting long-term temperature trends. The ARIMA model is effective in capturing long-term trends, while Prophet is good at predicting seasonal and yearly trends. These models need regular updates and refinements to ensure accuracy and effectiveness in predicting climate change.

Keywords: Climate Change, Naive, ARIMA, Prophet, SES Model

Introduction

Climate change is a critical global issue that requires accurate forecasting of future climate patterns to develop effective strategies to mitigate its impacts. Time series forecasting models are essential tools for predicting climate trends, including SES, ARIMA, Prophet, and Naive models. Each model has its strengths and weaknesses in predicting climate patterns.

Simple Exponential Smoothing (SES) is a commonly used time series forecasting model that uses a weighted average of past observations to predict future values. It is a straightforward model that assumes that future climate patterns will be similar to past patterns. However, it is not suitable for predicting trends that change rapidly, such as climate change.

Autoregressive Integrated Moving Average (ARIMA) is a popular time series forecasting model that is commonly used for climate change forecasting. It uses past observations to predict future trends, and it can capture long-term trends in climate patterns. ARIMA is a more complex model than SES and can handle more complex patterns in the data. However, it requires more data than SES to accurately predict trends and can be challenging to interpret.

Prophet is a relatively new time series forecasting model that has gained popularity due to its ability to capture seasonality and yearly trends. It is a flexible model that can handle data with missing values and outliers, making it suitable for climate data that may be incomplete or contain errors. Prophet uses a combination of additive and multiplicative models to capture the various components of climate patterns.

Naive is the simplest time series forecasting model and assumes that future climate patterns will be the same as the previous observation. While it is not a reliable model for long-term forecasting, it can be useful in predicting short-term trends.

The objective of this study is to compare the performance of the SES, ARIMA, Prophet, and Naive models in predicting climate change. The study aims to evaluate the performance of each model in predicting short-term temperature trends (up to 5 years) and long-term temperature trends (up to 50 years). Furthermore, the study seeks to determine which model is the most effective in predicting climate change patterns.

To achieve these objectives, we implemented the four models: SES, ARIMA, Prophet, and Naive, to forecast monthly temperature data.

The SES model is represented by the following equation:

$$Yt+1 = \alpha Yt + (1 - \alpha)Yt-1$$

where $Yt+1$ is the predicted value for the next time period, Yt is the observed value for the current time period, $Yt-1$ is the observed value for the previous time period, and α is the smoothing parameter ($0 < \alpha < 1$).

The ARIMA model is represented by the following equation:

$$Yt = \mu + \varphi 1Yt-1 + ... + \varphi pYt-p + \varepsilon t + \theta 1\varepsilon t-1 + ... + \theta q\varepsilon t-q$$

where Y_t is the observed value at time t, μ is the mean of the series, φ is the autoregressive parameter, ε is the error term, and θ is the moving average parameter.

The Prophet model is represented by the following equation:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

where y(t) is the observed value at time t, g(t) is the trend component, s(t) is the seasonal component, h(t) is the holiday component, and ε_t is the error term.

The Naive model is represented by the following equation:

$$Y_{t+1} = Y_t$$

To evaluate the performance of each model, we used the Mean Absolute Percentage Error (MAPE) and the Root Mean Squared Error (RMSE). The MAPE measures the percentage difference between the predicted and actual values. The RMSE measures the square root of the average of the squared differences between the predicted and actual values.

Through this project, we compare the performance of these four models in predicting climate change. We use global temperature and Indian monthly temperature datasets to train and test our models. The performance of each model is evaluated using the Mean Absolute Percentage Error (MAPE) and the Root Mean Squared Error (RMSE).

Our study aims to identify the strengths and weaknesses of each model and determine which model is best suited for predicting climate change patterns. Our findings can help us study more about climatic trends, climate change mitigation and adaptation strategies.

Literature Survey

Dimri, T., Ahmad, S. & Sharif, M. Time series analysis of climate variables using seasonal ARIMA approach. *J Earth Syst Sci* 129, 149 (2020) The authors utilized data spanning 100 years, from 1901 to 2000, for their study. The precipitation and temperature (maximum and minimum) information were sourced from the India Water Portal site. The ARIMA approach was applied for weather prediction in this basin region as no study of this kind has been conducted there yet. The seasonal ARIMA (SARIMA) model was utilized and forecasts were made for the next two decades (2001-2020). In the research paper, the evaluation metrics used were: AIC and BIC for assessing the quality of the statistical model, RMSE, and the Stationary R2.

Shad, M., Sharma, Y.D. & Singh, A. **Forecasting of monthly relative humidity in Delhi, India, using SARIMA and ANN models**. *Model. Earth Syst. Environ.* 8, 4843–4851 (2022). The objectives of the study were met using the monthly average relative humidity data from 2000 to 2016 collected by the India Meteorological Department in Pune. This data consists of the monthly average relative humidity expressed as a percentage. The trend in the forecast for relative humidity shows a decrease from 2017 to 2025. The performance of the models was evaluated using the root mean squared error (RMSE) and mean absolute error (MAE) metrics. The findings indicated that the SARIMA model had a forecasted relative humidity with an RMSE of 6.04 and an MAE of 4.56. In contrast, the MLP model reported a forecasted relative humidity with an RMSE of 4.65 and an MAE of 3.42. The study ultimately determined that the ANN model was more dependable in predicting relative humidity compared to the SARIMA model.

Ye, L. M., G. X. Yang, E. Van Ranst, and H. J. Tang, 2013: **Time-series modeling and prediction of global monthly absolute temperature for environmental decision making**. Adv. Atmos. Sci., 30(2), 382–396, doi: 10.1007/s00376-012

A deterministic-stochastic combined (DSC) approach was used to develop a generalized, structural, time series modeling framework for analyzing the monthly records of absolute surface temperature, a crucial environmental parameter. The framework characterizes the variation patterns in the temperature signal using both deterministic processes, such as polynomial functions and the Fourier method, to capture the global trend and cyclic oscillations, and stochastic processes, such as seasonal autoregressive integrated moving average (SARIMA) models, to account for any remaining patterns. The evaluation of prediction accuracy indicates that DSC models perform well compared to selected models of other authors and can be used as a supplementary tool for short-term environmental planning and decision-making when combined with other environmental models.

Romilly, Peter. (2005). **Time series modeling of global mean temperature for managerial decision-making. Journal of environmental management.** 76. 61-70. 10.1016/j.jenvman.2005.01.008. The authors in this paper have used univariate time series techniques to model the properties of a global mean temperature dataset to develop a cost-effective forecasting model for managerial decision-making over a short timeperiods. The statistical techniques used by the authors include seasonal and non-seasonal unit root testing with and without structural breaks, as well as ARIMA and GARCH modeling. The methodology employed was based on the autoregressive-integrated-moving average

(ARIMA) models popularized by Box and Jenkins (1970), together with more recent developments in the form of seasonal and non-seasonal unit root testing and generalized autoregressive conditional heteroscedasticity (GARCH) models. These modeling techniques are commonly used in a wide range of business and economic applications. Forecasting performance was compared in terms of three criteria: the root means square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE).

Prediction of climate change using ARIMA model - D Hebsiba beula, S Srinivasan C D Nanda kumar. This study focuses on climate prediction through the use of R-language algorithms. The time ARIMA (Autoregressive Integrated Moving Average) model is applied to forecast the climate. It is a predictive analysis designed to quantify events that take place throughout time. The dataset includes 56 years of past information for temperature, rainfall, crop productivity, and the area of land utilised for agriculture in India from 1961 to 2016. The data used is stationary which means the mean and variance are constant throughout the period. If the data are not stationary, they are transformed into stationary form and examined with the ACF and PACF (Partial Autocorrelation Function). R's built-in (auto.arima) function is used to make predictions. The report concludes by claiming that as the year progresses, both crop yield and temperature rise.

FORECATING OF TEMPERATURE BY USING TIME SERIES ANALYSIS - Asst.Prof. Mr. Amit Kumar Shubham Puri , Prajwal Selokar , Yash Talhar , Rushikesh Bhokre , Charchit Chakole , Department of Computer Technology, Priyadarshini College of Engineering, Nagpur, India.This research paper deals with the model that has been trained using past 137 years of data (1880-2017) and tested over 60 years to forecast maximum and minimum temperature. Forecasting for the next 60 years is done using the SARIMA (seasonal ARIMA) model. The Box Jenkins technique is the foundation for the auto-regressive (p) integrated (d) moving average (q) (ARIMA) model, which forecasts future trends by making the data stationary and eliminating seasonality.

Trends and variations in global temperature and precipitation from 1951 to 2010" by X. Zhang et al. (2015). This study examines the trend of extreme temperatures, heat waves, and cold spells in South Asian countries from 1951-2015, and the potential association with El Niño-Southern Oscillation events. The study concluded from 65 years of temperature data the frequency, duration, and intensity of heat waves also showed a consistent increase, with a significant positive correlation with annual AMO/IOBW.

Time series analysis of climate variables using seasonal ARIMA approach. T. Dimri, Shamshad Ahmad, Mohammad Sharif. December, 2020, Vol. 22. The study analyses precipitation and temperature data for the Bhagirathi river basin in Uttarakhand, India, from 1901 to 2002, obtained from the Climate Research Unit (CRU) TS2.1 dataset using the ARIMA and SARIMA model.

Proposed Work

Data Source:
https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

The data was sourced from the Berkeley Earth Surface Temperature Study, which combines 1.6 billion temperature reports from 16 archives. It is well-packaged and allows for easy slicing by country. The study publishes source data and transformation code, and uses methods to include weather observations from shorter time series, reducing the need to discard observations.

Data Analytics Models:
In this project we have used 4 models, Naive method, ARIMA Model, Prophet Model and Simple Exponential Smoothing (SES) using normal method.

1) Naive Model:

- Simple forecasting technique for benchmarking
- Predicts next value by adding a constant change to previous value
- Can be inaccurate and unreliable
- Assumes future value will be same as current value
- May not capture trend or seasonal patterns in data

Equation:
Y_t+1 = Y_t-11

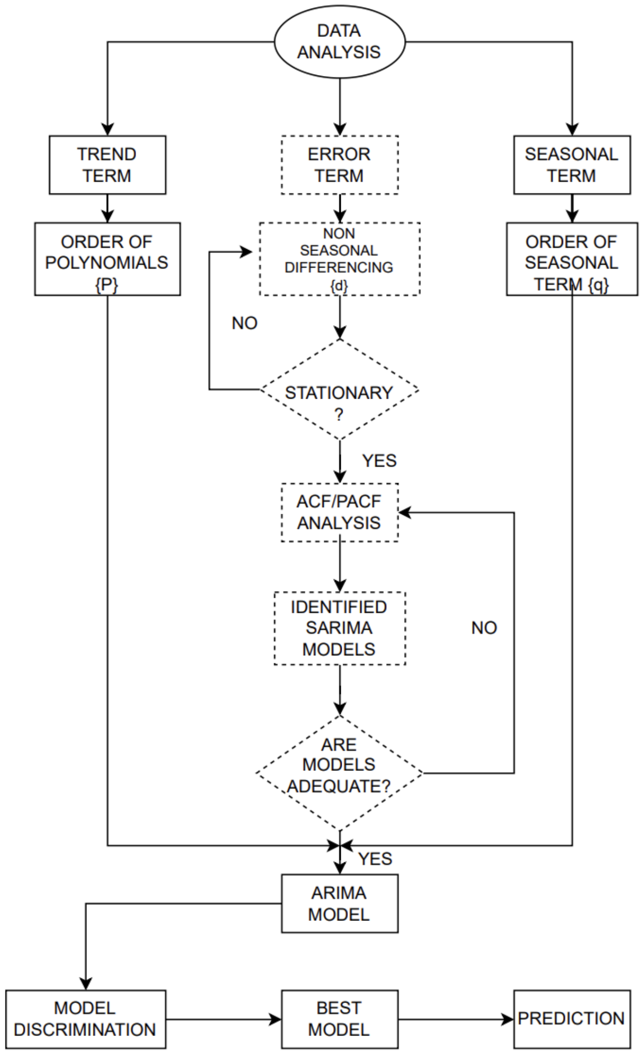
Pseudo-code:
-Take the corresponding observed value based on the month.
-Use that value as the forecast for the next time period.

Merits:
Easy to understand and implement.
Demerits:
May not capture trend or seasonal patterns in the data.

2) ARIMA Model:
ARIMA (Autoregressive Integrated Moving Average) is a popular time series forecasting technique that models the dependence of the variable on its past values, as well as its own previous errors.

Equation:
Y_t = c + Φ_1 Y_t-1 + Φ_p Y_t-p + ε_t + Θ_1 ε_t-1 + Θ_q ε_t-q

Flowchart:



Merits:

Can capture both trend and seasonal patterns in the data.

Demerits:

Requires the data to be stationary or made stationary through differencing.

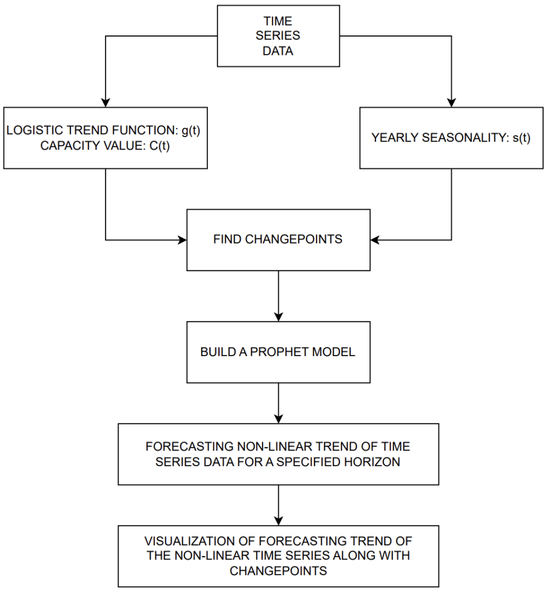
3) Prophet Model:

Prophet is a time series forecasting technique developed by Facebook that models the trend, seasonality, and holiday effects in the data. Prophet uses a decomposable time series model with three main components: trend, seasonality, and holidays.

Equation:

$y(t) = g(t) + s(t) + h(t) + \varepsilon(t)$

Flowchart



Merits:
Can capture both trend and seasonal patterns in the data.

Demerits:
Can be computationally intensive for large datasets.

4) Simple Exponential Smoothing:
Simple Exponential Smoothing is a popular time series forecasting technique that models the future value of a variable as a weighted average of its past values.
Equations:
$Y\_t+1 = \alpha Y\_t + (1 - \alpha)F\_t$

Pseudo-code:
-Initialize the forecast for the first time period as the first observed value
-For each subsequent time period, update the forecast using the Simple Exponential Smoothing equation
-Forecast future values of the time series using the fitted Simple Exponential Smoothing model
Merits:
Easy to understand and implement.
Demerits:

May not capture trend or seasonal patterns in the data.
Prophet may be the best model for predicting climate change due to its versatility. However, the choice of model depends on the data and specific problem.

Results and Discussions

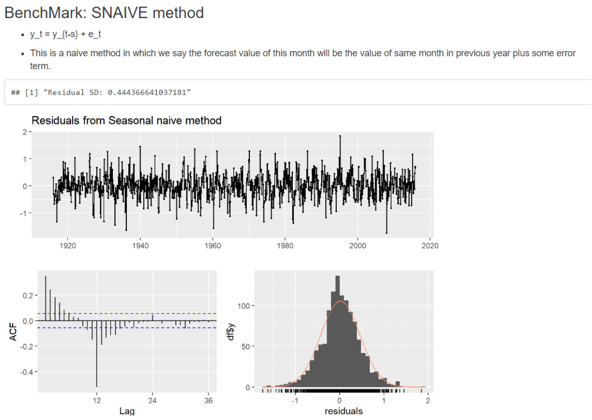
Global

Our problem being a diagnostic and predictive problem, we have analyzed our data and understood how the temperature changes on a monthly and yearly basis.
We explored the dataset and understood trends and patterns and found how the dataset shows anomalies. This goes in hand with the ongoing global warming hence proving that our inferences are correct.

Forecasting:

Naive Model

Arima Model



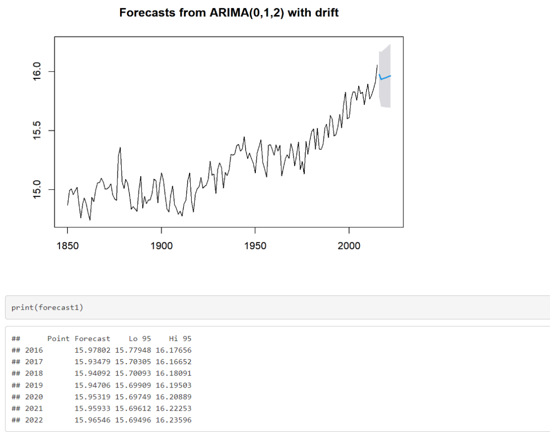
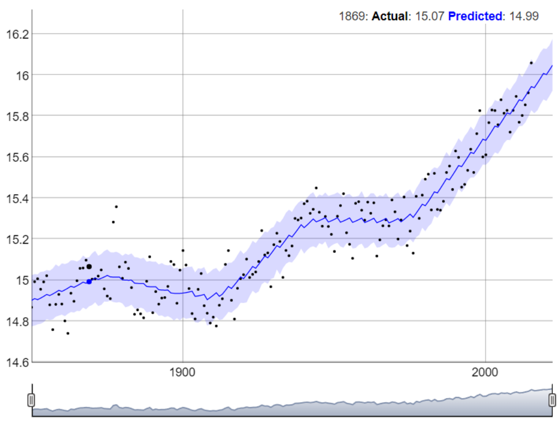


Fig 1.1

Fig 1.2

ARIMA model does a better job than Seasonal Naive method becuase the Residual SD of ARIMA is lower.

- **Prophet Model**



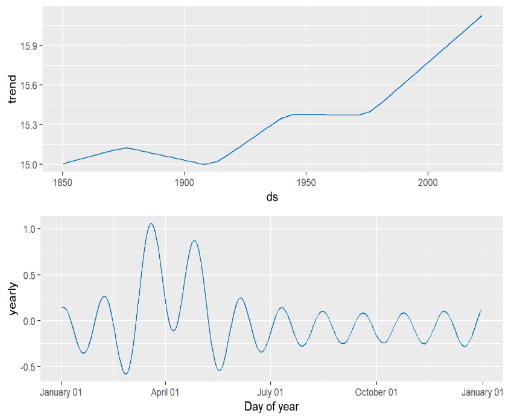


Fig 1.3

Fig 1.4

Prophet model gives a forecast which is almost similar to ARIMA model's forecast.

Exponential Smoothing

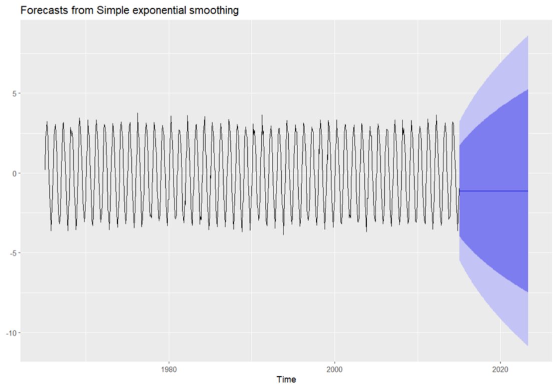


Fig 1.5

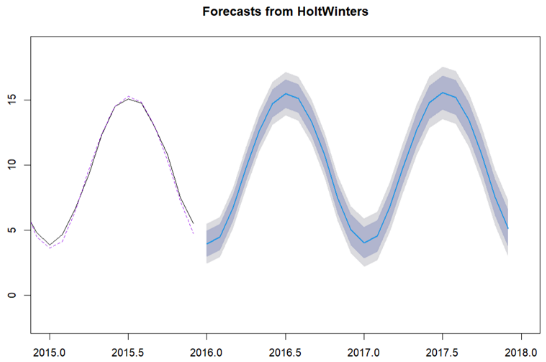


Fig 1.6

Using the Holt-Winters fit from before, we can use forecasts to make new predictions and include both 80% and 95% confidence intervals. The third graph shows the forecasted values from 2016 to 2018(grey region).

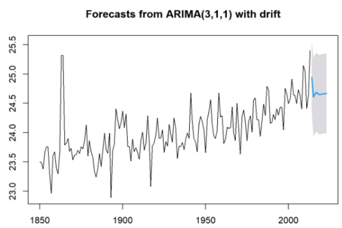
India

Forecasting :

Arima

Forecast next 10 years

```
forecast1 = forecast(model1, level=c(95),h = 10)
plot(forecast1) ## the trend continues as the avg temperature continues to increase on yearly basis
```




```
print(forecast1)
```

```
##      Point Forecast    Lo 95    Hi 95
## 2014       24.94554 24.34025 25.55082
## 2015       24.61017 23.93759 25.28275
## 2016       24.65194 23.97891 25.32497
## 2017       24.68896 24.01456 25.36336
## 2018       24.65972 23.98179 25.33764
## 2019       24.64791 23.96855 25.32726
## 2020       24.65733 23.97733 25.33732
## 2021       24.66455 23.98384 25.34525
## 2022       24.66846 23.98700 25.34993
## 2023       24.67397 23.99183 25.35610
```

```
observed = c(24.98,24.85,25.23,25.4,25.54,25.64,25.93,25.79,25.78,25.85)
forecasted = c(24.77,24.95,24.61,24.66,24.69,24.66,24.65,24.66,24.66,24.67)
err = abs(observed-forecasted)/observed
cat("Average error in prediction : ",round(sum(err)/10,4))
```

Fig 2.1 **Fig 2.2**

- ❖ The forecasts although not 100% accurate, are pretty close to real values.
- ❖ The forecast for 2021 only varies by 1 degree.

- **Prophet Model**

```
forecast1 = predict(prop_fit,Future1)
tail(forecast1[c('ds','yhat')],11)
```

```
##             ds     yhat
## 164 2013-04-06 24.76912
## 165 2014-04-06 24.78062
## 166 2015-04-06 24.79188
## 167 2016-04-06 24.81808
## 168 2017-04-06 24.82983
## 169 2018-04-06 24.84133
## 170 2019-04-06 24.85259
## 171 2020-04-06 24.87879
## 172 2021-04-06 24.89053
## 173 2022-04-06 24.90203
## 174 2023-04-06 24.91330
```

```
prop_fit$component.modes
```

Fig 2.3

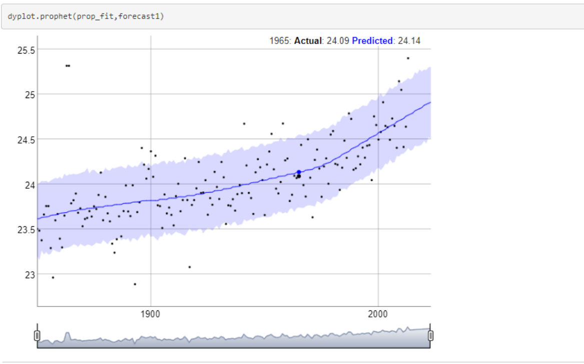


Fig 2.4

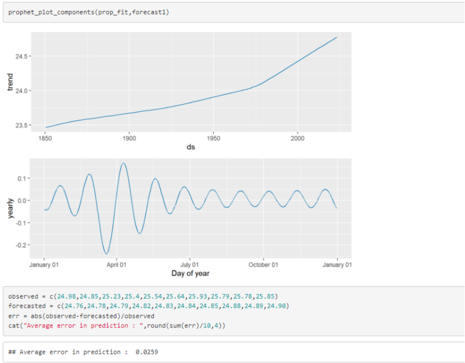


Fig 2.5

Prophet model gives a forecast which is almost similar to the ARIMA model's forecast.

Comparing the forecasted values:

YEAR	ARIMA	PROPHET	OBSERVED
2013	24.77	24.76	24.98
2014	24.95	24.78	24.85
2015	24.61	24.79	25.23
2016	24.66	24.82	25.4
2017	24.69	24.83	25.54
2018	24.66	24.84	25.64
2019	24.65	24.85	25.93
2020	24.66	24.88	25.79
2021	24.66	24.89	25.78
2022	24.67	24.90	25.85
2023	24.67	24.91	

Table 1

- **Exponential Smoothing**

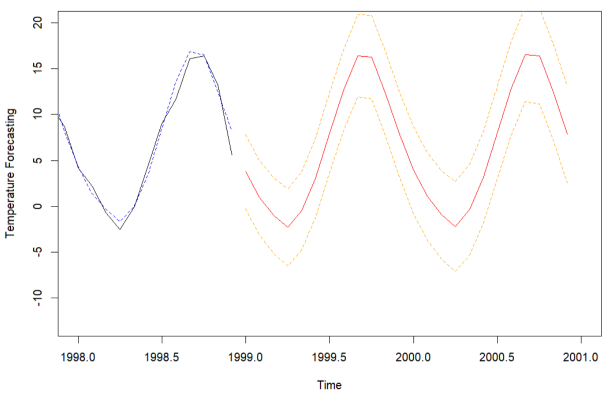


Fig 2.6

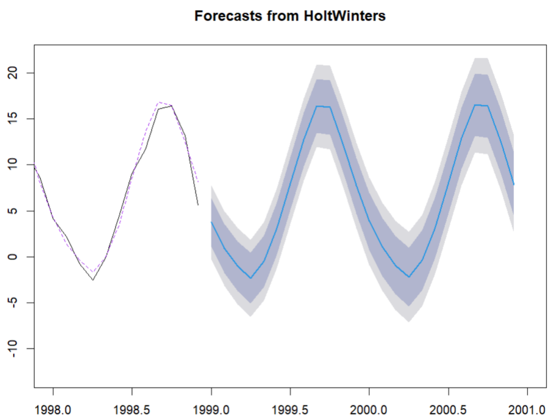


Fig 2.7

Using the Holt-Winters fit from before, we can use forecasts to make new predictions and include both 80% and 95% confidence intervals. The third graph shows the forecasted values from 1998 to 2001(grey region).

Discussions

- Time Series Analysis was done on Berkeley Earth Climate Change Earth Surface Temperature Data Repository.

- Seasonality and Trend test on these data showed they correspond to the natural facts about climate.

- The entire project was divided into different phases of data analytics life cycle.

- In analysis of the Global scale land temperature, we saw a slight trend of increasing temperature across the time period this corroborates the phenomenon of Global Warming.

- As the final result of the project, we forecasted the average temperature value for year 2021 and 2022.

- At global scale the value predicted is 15.959 °C for 2021 and 15.9655 °C for 2022.

- For India the forecasted temperature value is 24.664 °C for 2021 and 24.668 °C for 2022.

- All the visualizations done in this project showcases the validity of climate change and other known facts of yearly temperature distribution.

View the entire project :

https://drive.google.com/drive/folders/1dyI05RjPHjhl86mGO9uc_xxyG7mF1ZZ-?usp=sharing

Conclusion

We did a thorough and effective analysis of the Time Series provided by the Berkeley Earth organization, which had data points from the year 1750 till 2015. We did a comprehensive comparative analysis on average land and ocean temperature globally and at country-level scales – India. We implemented various models that were best fit for our datasets like the Naive model, Arima Model, Prophet Model, and Exponential smoothing. We were successful in finding an ARIMA model that best fits the time series taken for the analysis purpose. Using this model, we forecasted the temperature values for the years 2021 and 2022 globally and for India. In this project, we gave a heavy emphasis on the Data Analytics Life cycle and the principles of doing it which were taught to us in the course. All of them are documented in a punctilious manner. In conclusion, we were able to achieve our objective which was to learn and understand Time Series Analysis along with applying those concepts to discern the phenomenon of climate change. As part of future work, we will explore better complex models than the one presented in this project. We can also add multivariate variables like other climatic factors - precipitation, humidity, rainfall, etc, in the analysis to better gauge the scenario of Global Warming.

References

1) Dimri, T., Ahmad, S. & Sharif, M. Time series analysis of climate variables using seasonal ARIMA approach. *J Earth Syst Sci* 129, 149 (2020). https://doi.org/10.1007/s12040-020-01408-x

2) Shad, M., Sharma, Y.D. & Singh, A. Forecasting of monthly relative humidity in Delhi, India, using SARIMA and ANN models. *Model. Earth Syst. Environ.* 8, 4843–4851 (2022). https://doi.org/10.1007/s40808-022-01385-8

3) Ye, L. M., G. X. Yang, E. Van Ranst, and H. J. Tang, 2013: Time-series modeling and prediction of global monthly absolute temperature for environmental decision making. Adv. Atmos. Sci., 30(2), 382–396, doi: 10.1007/s00376-012

4) Romilly, Peter. (2005). Time series modeling of global mean temperature for managerial decision-making. Journal of environmental management. 76. 61-70. 10.1016/j.jenvman.2005.01.008.
5) Prediction of climate change using ARIMA model - D Hebsiba beula, S Srinivasan C D Nanda Kumar

6) FORECASTING OF TEMPERATURE BY USING TIME SERIES ANALYSIS - asst.Prof. Mr. Amit Kumar Shubham Puri, Prajwal Selokar, Yash Talhar, Rushikesh Bhokre, Charchit Chakole, Department of Computer Technology, Priyadarshini College of Engineering, Nagpur, India

7) Trends and variations in global temperature and precipitation from 1951 to 2010" by X. Zhang et al. (2015)

8) Time series analysis of climate variables using seasonal ARIMA approach. T. Dimri, Shamshad Ahmad, Mohammad Sharif. December 2020, Vol. 22.