Illinois Institute of Technology

MATH546: Introduction to Time Series



Final Project

**Forecasting Earth Surface Temperature**

|  |  |
| --- | --- |
| **Group Members Names** | **CWID** |
| Dhanvanth Voona | A20543395 |
| Neeraj Vardhan Buneeti | A20545853 |
| Sai Charan Gangili | A20543155 |

**Contents**

|  |  |
| --- | --- |
| S. No. | Topic |
| 1 | Problem Definition |
| 2 | Dataset Details |
| 3 | Data Preparation |
| 4 | Exploratory Data Analysis (EDA) |
| 5 | Stationarity Check |
| 6 | Trend & Seasonality Analysis |
| 7 | ARIMA model fitting & Residual Analysis |
| 8 | SARIMA model fitting & Residual Analysis |
| 9 | Conclusion |

1. **Problem Definition**

As the global climate undergoes rapid changes, understanding and forecasting Earth's surface temperature have become paramount. The Earth's surface temperature plays a crucial role in various aspects of life on our planet, including agriculture, ecosystems, and human health.

Accurate temperature forecasts are essential for planning and mitigating the impacts of climate change, such as extreme weather events, sea level rise, and shifts in agricultural productivity. Moreover, understanding historical temperature trends helps in analyzing the effectiveness of current climate policies and guiding future actions.

In this project, we aim to forecast Earth's surface temperature using historical temperature data spanning two centuries. By employing state-of-the-art forecasting models, we seek to provide valuable insights into future temperature trends and contribute to the ongoing efforts to combat climate change.

**A diagram of a process

Description automatically generated**

**Figure 1.** Proposed methodology**.**

1. **Data handling and preprocessing:**

The dataset consists Average temperature, Average Temperature Uncertainty, of 159 countires for every month from the year 1743 to 2013 in total 270 years. Average Temperature: Average land temperature in Celsius of the particular city in the recorded in the whole month. Average Temperature Uncertainty: The 95% confidence interval around the average.

We saw presence of missing values in the dataset: We have cleaned the dataset with no missing values or outliers

There are total of 159 countries in our dataset with recordings from various cities in each country.

A screenshot of a computer

Description automatically generated

**Figure 2**. Overview of the number of temperatures recordings for each country in the dataset.

From Figure 2. we can observe the number of recordings vary for each country, this is because each country has different number of cities involved. For example, the dataset involves 391 cities in India while 248 cities in USA.

From 248 cities in United States, we have chosen Chicago city for further analysis and forecasting.

**A graph showing the temperature of a month

Description automatically generated**

**Figure 3**. Temperature by Month.

**A graph of different colored lines

Description automatically generated**

**Figure 4.** Average Temperature in Chicago by each season.

We can clearly observe the difference in average temperatures for each season with summer being the highest and winter being the least from Figure 4.

1. **Trend and Seasonality Analysis:**

A blue line graph with numbers

Description automatically generated with medium confidence

**Figure 5.** Temperature vs Date for Chicago city.

It is difficult to identify any presence of trend and seasonality in the data from Figure 5. Since the Data consists of more than 3000 observations, we have split the data into 3 parts and plotted 10 years data to have better observation for trend & seasonality.

A graph showing the temperature of a temperature

Description automatically generated

A graph showing the temperature of a temperature

Description automatically generated with medium confidence

A graph showing the temperature of a temperature

Description automatically generated

**Figure 6.** Exploring Trend and Seasonality: 10-Year Temperature Data from each split.

From Figure 6. It is very evident that there is presence of seasonality with approximate time-period of 12 months for each split indicating the time period of seasonality hasn’t changed over the last 3 centuries.

1. **Decomposition:**

Decomposition is a method used to break down a time series into its constituent components: trend, seasonality, and residual (or error) components.

Trend: The long-term progression or directionality of the time series. It represents the underlying pattern in the data, showing whether the series is increasing, decreasing, or staying relatively constant over time.

Seasonality: The repeating pattern or fluctuations in the data that occur at regular intervals. For example, temperature might exhibit seasonality with higher values in summer and lower values in winter.

Residual: The part of the data that cannot be explained by the trend or seasonality. It represents the random fluctuations or noise in the data.

A screenshot of a graph

Description automatically generated

**Figure 7.** Trend, Seasonal and Residual Components of the given data.

We can observe in trend component is mostly in a plateau or stagnant position without significant upward or downward movement. Whereas in Seasonal component we can clear see pattern or fluctuation that occur at a regular intervals.

**Need of separating these components:**

Understanding the trend and seasonality in the data is crucial for accurate forecasting. By separating these components, you can better model and predict future values. By separating out the trend and seasonality components, you can often make the data more stationary. Stationary data is easier to model and forecast because its statistical properties, such as mean and variance, remain constant over time. Models like ARIMA (Autoregressive Integrated Moving Average) require stationary data for accurate predictions.

Separating out the trend and seasonality can reduce the autocorrelation in the residual component. Autocorrelation occurs when a time series is correlated with its past values, which can lead to biased forecasts. By removing trend and seasonality, you can often reduce this autocorrelation, leading to more accurate forecasts.

1. **Stationarity Check**

**Stationarity:** Stationary data refers to a time series where the statistical properties such as mean, variance, and autocorrelation do not change over time. In simpler terms, a stationary time series is one whose behavior is consistent and predictable throughout its entire length

To check for stationarity in a time series, you can use the **Augmented Dickey-Fuller (ADF) test**.

In the Augmented Dickey-Fuller (ADF) test, the null hypothesis (Ho) is that the time series has a unit root and is non-stationary. The alternative hypothesis (Ha) is that the time series does not have a unit root and is stationary. So, if the p-value is less than a chosen significance level (e.g., 0.05), you reject the null hypothesis. In other words, if the p-value is less than 0.05, you have enough evidence to conclude that the data is stationary. If the p-value is greater than or equal to 0.05, you fail to reject the null hypothesis, indicating that the data is non-stationary

A white background with text

Description automatically generated

**Figure 8.** ADF test results.

We can see that the p-value <<< 0.05 which rejects the null hypothesis and indicates that time series does not have a unit root and is stationary. Since the time series is stationary we need not deseasonalize the time series.

1. **Autocorrelation and Partial Autocorrelation Analysis.**

A graph with blue dots and lines

Description automatically generated

**Figure 9.** Autocorrelation plot of given time series

The autocorrelation at lag k (ACF(k)) that is significantly different from zero and larger than the adjacent autocorrelations suggests an MA component of order k. From the Figure 9. We can infer the order of MA to be 2 or 3. The seasonal pattern in the ACF suggests that there is a repeating pattern in the data at regular intervals, indicating the presence of seasonality.

A graph with blue dots and numbers

Description automatically generated

**Figure 10.** Partial Autocorrelation of given time series.

The partial autocorrelation at lag k (PACF(k)) that is significantly different from zero and larger than the adjacent partial autocorrelations suggests an AR component of order k. From the Figure 10. We can infer the order of AR to be 2. The slow decay indicates that each observation is related to its seasonal neighbors.

**Model Selection:**

**ARIMA:**

* ARIMA, which stands for Autoregressive Integrated Moving Average, is a robust statistical method utilized for analyzing and forecasting time series data.
* Autoregression (AR):AR captures the dependency between an observation and its lagged observations. Put simply, the current value of the time series depends on its past values.
* Integrated (I):The integration process makes the time series stationary by differencing raw observations. Stationarity ensures consistent statistical properties over time, which is crucial for reliable analysis.
* Moving Average (MA):MA models the relationship between an observation and the residual errors from a moving average model applied to lagged observations. It captures the short-term fluctuations in the data.

**SARIMA:**

**Model Training:**

**ARIMA:**

We have performed Grid Search to find the optimal Hyper-parameters of the model.

We chose the range of p, d, and q to be (0, 5), (0,5), and (0, 5) respectively.

After performing Grid Search technique we found the optimal parameters to be (2, 0, 3).

Using these parameters we have trained the ARIMA model and forecasted the last 10% portion of the data. From Figure 11. We can observe the forecasted temperature in red and the original temperature in blue.

A graph showing a number of data

Description automatically generated

**Figure 11.** Forecasting with ARIMA using optimal parameters (2, 0, 3).

In the beginning, the forecast has shown minute difference with original values, but in the last few years the model was unable to exhibit the seasonal characteristics of the temperature and shown poor performance.

A graph of red lines

Description automatically generated

**Figure 12.** Forecasting future values with ARIMA using optimal parameters (2, 0, 3).

**Residual Analysis:**

**A blue sound wave graph

Description automatically generated**

**Figure 13.** Residual Plot of ARIMA forecasted values.

Here the residual plot follows the random process indicating that the model has captured the underlying patterns in the data.

A graph of a function

Description automatically generated with medium confidence

**Figure 14.** Histogram of Residuals.

We observe a normal distribution in the histogram of residuals indicating that, on average, the model's predictions are accurate. But the right skewed histogram suggests that the model tends to overestimate the values, leading to positive residuals (actual - predicted) being larger than expected. This could indicate that the model is missing important features or that the model assumptions are violated.

A graph of a graph with blue dots

Description automatically generated with medium confidence

**Figure 15.** Autocorrelation and Partial Autocorrelation of residuals.

We observe a tiny seasonal components in the ACF and PACF of residuals which suggests that there may be residual seasonal patters in the data that are not captured by ARIMA model. To capture the seasonal patterns, we have implemented SARIMA on our time series.

**SARIMA:**

**A graph showing a number of data

Description automatically generated**

**Figure 16.** Forecasting with SARIMA on test data(last 10% data).

Compared to ARIMA forecasting from Figure 11. SARIMA has given more significant forecasting result, which indicates SARIMA has captured the underlying patterns in the time series.

A graph showing a temperature

Description automatically generated with medium confidence

**Figure 17.** Forecasting temperature of next 20 years using SARIMA.

We can observe a downward trend along with seasonality with time period of 12 months in the next 20 years forecast.

**Residual Analysis:**

**A blue line graph with white text

Description automatically generated**

**Figure 18.** Residual plot of SARIMA forecasted values.

Here the residual plot follows the random process indicating that the model has captured the underlying patterns in the data.

A graph of a graph with a line and a green line

Description automatically generated with medium confidence

**Figure 19.** Histogram of residuals.

Compare to Figure 14. The histogram of residuals of SARIMA is lesser right skewed indicating the SARIMA model has proven to be better fit compared to ARIMA

A graph of a graph

Description automatically generated with medium confidence

**Figure 20.** Autocorrelation and Partial autocorrelation of residuals.

Well clearly the ACF and PACF of residuals has no signs of seasonal patterns which indicates that the seasonal component of the data has been effectively captured by the model. This is a positive sign, as it suggests that the model is able to explain the seasonal fluctuations in the data, and the residuals are mostly random with no remaining systematic patterns.

|  |  |  |
| --- | --- | --- |
| Metrics | ARIMA | SARIMA |
| MSE | 123.55 | 48.33 |
| RMSE | 11.115 | 6.95 |

**Table 1.** Performance metrics.

As observed from residual analysis the performance metrics has shown that SARIMA has outperformed ARIMA model in terms of forecasting accuracy for the given dataset.

**Conclusion:**

In this project, we aimed to forecast temperature using data spanning the last three centuries. The project involved an extensive exploratory data analysis (EDA) that yielded valuable insights into the dataset. Trend and seasonality analysis revealed clear patterns in the data, highlighting the importance of accounting for these factors in the modeling process.

Stationarity tests were conducted to ensure the suitability of the data for time series analysis. Both ARIMA and SARIMA models were trained on the dataset, with SARIMA demonstrating superior performance over ARIMA. This was evident from the residual analysis, which showed that SARIMA had lower residual errors compared to ARIMA.

Performance metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) further confirmed that SARIMA outperformed ARIMA in forecasting temperature. These results suggest that SARIMA is a more suitable model for forecasting temperature.

For future enhancements, it is recommended to include external factors such as greenhouse gas emissions, solar activity, volcanic eruptions, and ocean currents in decision making. Implementing ensemble modeling techniques, such as combining forecasts from multiple ARIMA and SARIMA models, can further improve the accuracy of predictions. Additionally, incorporating long-term trends in Earth's surface temperature, such as global warming, into the forecasting models can provide more comprehensive and insightful forecasts.