PRESIDENCY UNIVERSITY
KOLKATA

Project on Peleoclimate Temperature Backcasting
Presidency University
4th Semester

Shubhamoy Paul
Reg. No: 21414170024

Department of Statistics
Presidency University, Kolkata

Abstract. This project focuses on backcasting past temperatures in Greenland using paleoclimate data from ice cores. The objective is to understand historical changes in temperature in Greenland and develop a model to backcast past temperatures. The project includes data collection, univariate and multivariate analysis, and model selection and fitting. The exploratory data analysis includes descriptive measures and line plots to identify trends and patterns. Univariate modeling techniques were initially used but did not provide satisfactory results. Therefore, multivariate analysis was conducted using variables such as CO2 concentration, volcanic eruption temperature, and ice surface temperature. Linear regression was used to analyze the data, and the results were compared to actual values to evaluate the accuracy of the model. The final model achieved an accuracy of 96%. The project faced challenges in acquiring sufficient and comparable data and in analyzing the data due to the lack of prior knowledge in paleoclimate analysis. The bibliography includes sources on time series modeling of paleoclimate data and software for paleoclimate research and education. The Github link to the source code is provided.

Contents

Introduction:

According to recent studies, we have seen that most statisticians are busy predicting the current temperature, which is obviously most important for us. However, they have forgotten to look at our past while predicting the current temperature. Millions of years ago, when our Earth was formed, it was just a ball of fire; then, the temperature was much higher than today's. Then slowly, our Earth cooled down, and that period was the ice age when the temperature was freezing. And then, after many actions and reactions, we got this Earth. In this project, past temperature has played an important role. To introduce this project properly, we need to know the answer to some questions. These questions are

1. What is the meaning of paleoclimate? Paleoclimate refers to the study of past climates on Earth before it was possible to measure them directly. Paleoclimate research tries to figure out how the climate used to be, how it changed, and how it changed over time by looking at proxy records like ice cores, tree rings, sediment cores, and old documents. After taking a look at the way the climate has changed, scientists can predict more about the system's work and how human actions affect the climate now and in the future.

2. Why do we need to study paleoclimate or why scientist are interested to study paleoclimate? There are several reasons why scientists are interested in studying paleoclimate:

(1) Understanding natural climate change: Scientists can learn more about how the Earth's climate system changes on different time scales, from decades to millions of years, by looking at how the climate has changed in the past. This information is vital for figuring out how the climate will change and preparing for it.
(2) Evaluating the effects of human actions: Paleoclimate research can give a long-term view of the effects of human actions on the climate system so that they can be evaluated. Scientists can determine how much human activities contribute to global warming and climate change by looking at how the climate has changed in the past and how they think it will change.
(3) Creating climate models: Paleoclimate data are needed to test and improve climate models, which are used to simulate how the climate will be in the future. By comparing model simulations to paleoclimate data, scientists can determine how accurate their models are and find places to improve their models.
(4) Paleoclimate research can help determine the range of possible climate scenarios for the future, including extreme events like droughts, floods, and heat waves. This information is vital for figuring out how to deal with climate change's effects and what risks are involved.

Studying paleoclimate gives us essential information about the Earth's climate system and is necessary to understand and deal with the problems that climate change poses.

Data Collection:

There are commonly used proxies in paleoclimate analysis, including:**[2]**

(1) Ice cores: To get ice cores, scientists drill [0.0.1] deep into ice sheets in places like Antarctica and Greenland. These cores hold information about the temperature and chemical makeup of the atmosphere from hundreds of thousands of years ago.

FIGURE 0.0.1. Ice driller machines

Img[1]

[1] https://www.nsf.gov/news/news_images.jsp?cntn_id=112909&org=NSF.

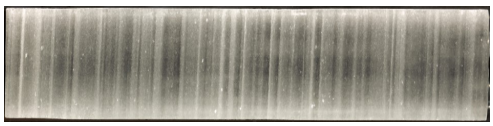


FIGURE 0.0.2. Ice Layers

2 Sediment cores: Scientists take sediment cores from the ocean floor and lake beds. These cores hold information about how the climate used to be and how it changed in the past. The layers of sediment can tell us about the temperature, the amount of rain, and the types of plants that were growing at the time.

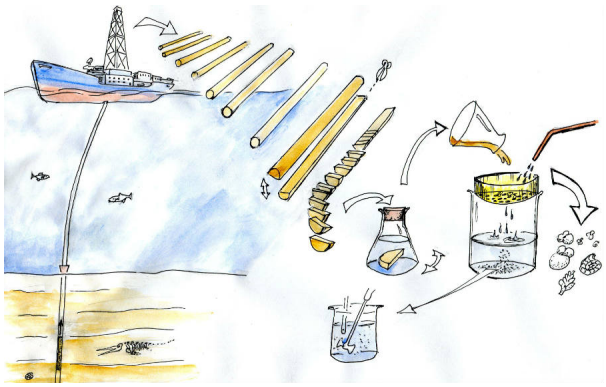


FIGURE 0.0.3. Sediment Core data extraction Process

FIGURE 0.0.4. Sediment Core Layers

Img[2] Img[3] Img[4]

[2]https://upload.wikimedia.org/wikipedia/commons/b/bc/GISP2D1837_crop.jpg.
[3]https://godwinlab.esc.cam.ac.uk/analytical-resources/research-laboratories/the-sedimentary-laboratory.
[4]https://www.researchgate.net/figure/Photographs-of-sediment-cores-from-reservoirs-A-B-Core-from-Grimselsee-Sediment-from_fig5_220013406.

3 Tree rings: Changes in temperature and rainfall cause trees to grow in different ways, which can be used to figure out about past weather. Scientists can determine how temperatures and rainfall have changed over time by looking at tree rings' width, density, and makeup.

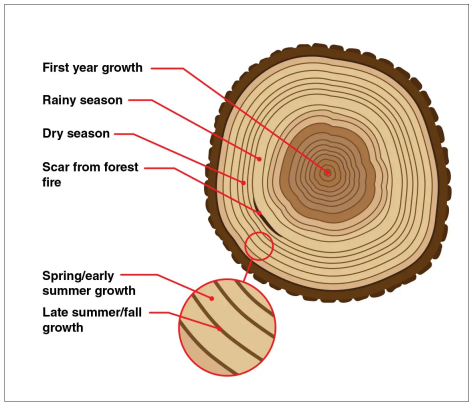


FIGURE 0.0.5. tree ring

4 Historical records: Things like diaries, ship logs, and weather reports from the past can tell us about weather patterns and extreme events from the past. These records can give us informations about temperature millions of years ago and are a great way to learn about the climate in the past, which is called paleoclimate.
5 Coral reefs: Coral reefs can tell us about the past temperatures and conditions of the ocean. By looking at coral skeletons, a scientist can quickly determine the under water chemical changes and the temperature change at how coral skeletons grew and what chemicals were in them.

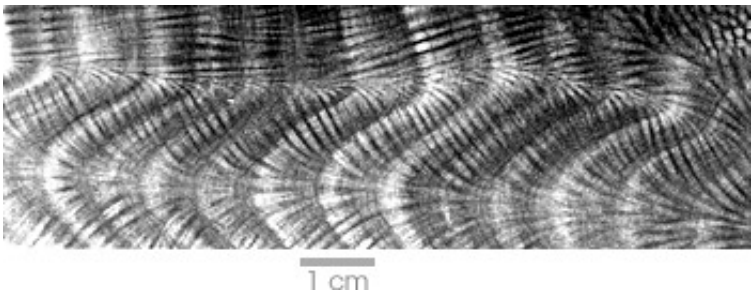


FIGURE 0.0.6. Coral reefs

Paleoclimate data can be obtained from various sources, and each method has strengths and limitations. By combining data from multiple sources, scientists can build a complete picture of past climate conditions and changes.

Img[5] Img[6]

4. Why we select ice-core to extract our data?:[**1**]

. Ice cores provide a high-resolution record of past climate conditions, with annual layers that can be used to track changes in temperature, precipitation, atmospheric composition, and more. Ice cores can provide records of past climate conditions dating back hundreds of thousands of years, providing a long-term perspective on climate variability. The layers in ice cores can be accurately dated using various methods, including counting layers, analyzing volcanic ash layers, and measuring isotopes. Ice cores contain information about

[5] https://climate.nasa.gov/news/2540/tree-rings-provide-snapshots-of-earths-past-climate/.
[6] https://earthobservatory.nasa.gov/features/Paleoclimatology_CloseUp/paleoclimatology_closeup_2.php.

past climate conditions, including atmospheric composition, temperature, precipitation, etc. Scientists can build a more comprehensive picture of past climate conditions by analyzing multiple proxies in ice cores. Ice cores have been collected from different world regions, including Antarctica, Greenland, and high-altitude glaciers. By studying ice cores from different regions, scientists can investigate regional and global patterns of climate change. Ice cores are an excellent way to learn about past climate conditions and changes. However, collecting paleoclimate data is expensive, dangerous, and takes a long time. The people who collect data have to go through hard training and take care of their health. They can die while collecting data. Because of this, these data are not open to the public. There are only studies. We have a dataset showing Greenland's temperature for the past 11500 years from the National Centers for Environments Information website.

Data: View

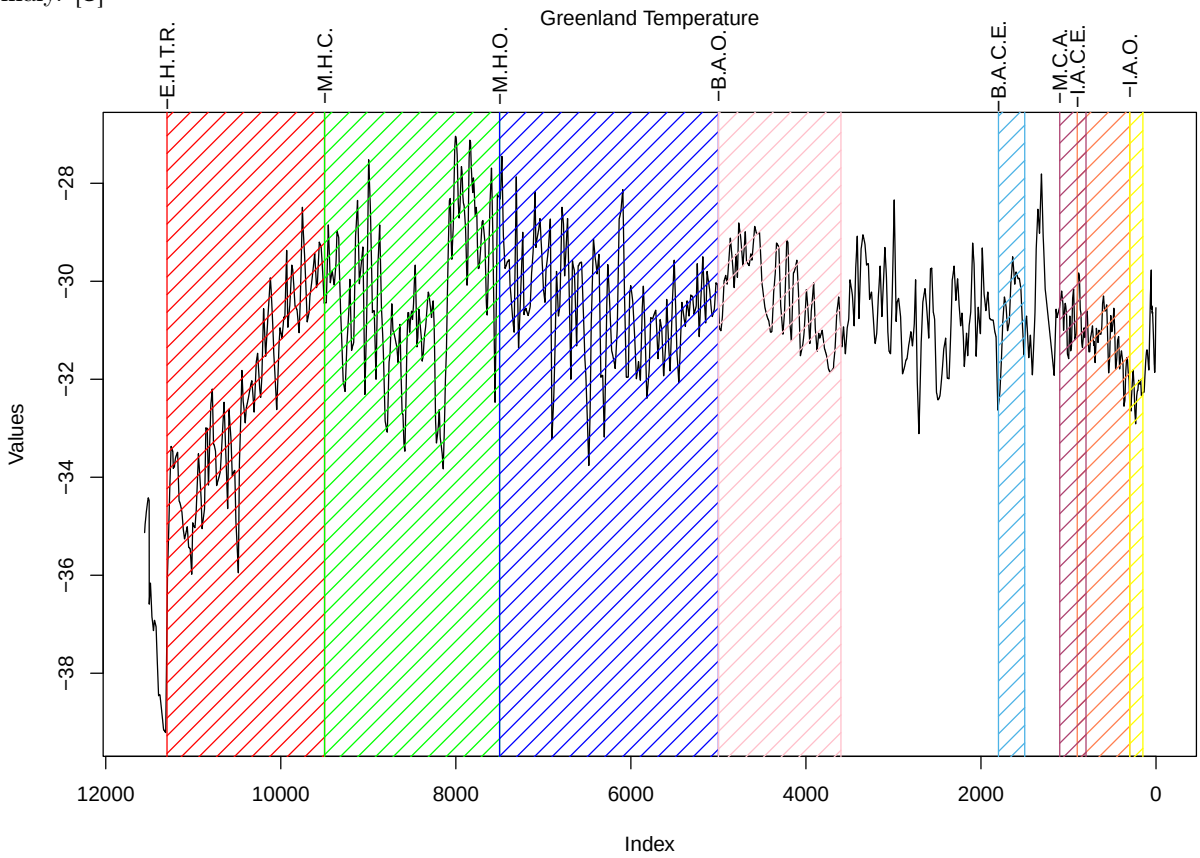
Univariate Analysis

Data Descriptions:

This data was collected from a Kobashi scientific research paper [**7**]. This paper is mainly about the reconstruction of Greenland's temperature over the past 11,500 years for different reasons. This paper has been approved by NCEI [**6**]. It talks about the temperature of the air and the ice age. The temperature of the air has two ranges, high and low. The data is taken for the GISP2(Green Land Ice Sheet Project) about the temperature over the past 11500 years. The data has been collected after drilling into Sumit on the Greenland Ice Sheet. This research paper was published on May 3, 2017. The place from which the data was collected.

Northernmost Latitude: 72.6
Southernmost Latitude: 72.6
Easternmost Longitude: -38.5
Westernmost Longitude: -38.5
Evaluation: 3200m

Here, the ice age is given in calibrated years before the present and in the 2005 Greenland Ice Core Chronology. It talks about a certain way to figure out how old ice cores from Greenland are. This timescale is based on the study of ice cores and other geological data, and it shows how events in Greenland's past can be linked to each other. The time scale is set up so that if you add one unit to the ice age, you go back one year. Radio Carbon-calibrated methods are used to figure out the ice age. A scale of degrees Celsius is used to measure the temperature here. The data covers the past 12000 years, allowing us to categorize this series into various Earth ages. These are E.H.T.R.=Early Holocene Temperature Rise, M.H.C.=Mid-Holocene Cooling, M.H.O.=Mid- Holocene Optimum, B.A.C.E.=Bronze Age Cold Epoch, B.A.O.=Bronze Age Optimum, I.A.C.E.=Iron Age Cold Epoch, I.A.O.=Iron Age Optimum, M.C.A.=Medieval Climate Anomaly. [**3**]



Project Plan (project Road map):[1]
. A time series project involves analyzing and modeling data to find trends, seasonality, patterns, and relationships that can help make predictions or decisions. Here is a general plan for a time series project or a roadmap:

(1) Set Objective: Figure out how the temperature of green land has changed. Moreover, try to create such a model to backcast past temperatures.
(2) Collect and preprocess data: We have to do cleaning, filtering, or scaling if it needs before processing. We have to ensure the data is in a format that can be analyzed.
(3) Exploratory data analysis: Conduct exploratory data analysis to understand the data and its characteristics better. This can involve visualizing the data, identifying trends and patterns, and performing statistical tests.
(4) Model selection: Select an appropriate model or model to analyze the data. Consider factors such as the nature of the data, the research question, and the desired outcomes.
(5) Model fitting and validation: Fit the chosen model to the data and validate its performance. Use statistical metrics such as R-squared, mean squared error, and root mean squared error to assess the model's accuracy.
(6) Backcasting: Use the model to predict past events or trends. Evaluate the quality of the backcasts using appropriate metrics.
(7) Interpretation: Interpret the analysis results and compare them with the actual temperature.

Objective
. This project aims to investigate and understand the historical changes in temperature in Greenland. Analyses will be performed to identify patterns and trends, and a model will be developed to backcast past temperatures. Validating the model against known historical data and comparing it with other reconstructions or paleoclimate records will assess its accuracy and performance. The developed model will be a valuable tool for backcasting past temperatures, contributing to our knowledge of climate change and its impacts on this region.

Collect and preprocess data
. Loading the Greenland Temperature dataset:

```
  Age Greenland_temperature Upper_band Lower_band
1 -43                -30.54     -30.34     -30.74
2 -42                -30.61     -30.46     -30.76
3 -41                -30.72     -30.58     -30.86
4 -40                -30.85     -30.73     -30.97
5 -39                -31.02     -30.88     -31.16
6 -38                -31.16     -31.00     -31.32
[1] "Rows: 11567"
[1] "Columns: 4"
```

There are 11556 rows and 4 columns in the data.

We have seen that each column has 11 empty rows. In the end, there are some outliers. During the step called "preprocessing," we have to get cleared of these "null" and "outlier" values.

```
[1] "Rows: 11200"
[1] "Columns: 4"
```

Exploratory Data Analysis (EDA).

Data Summary:

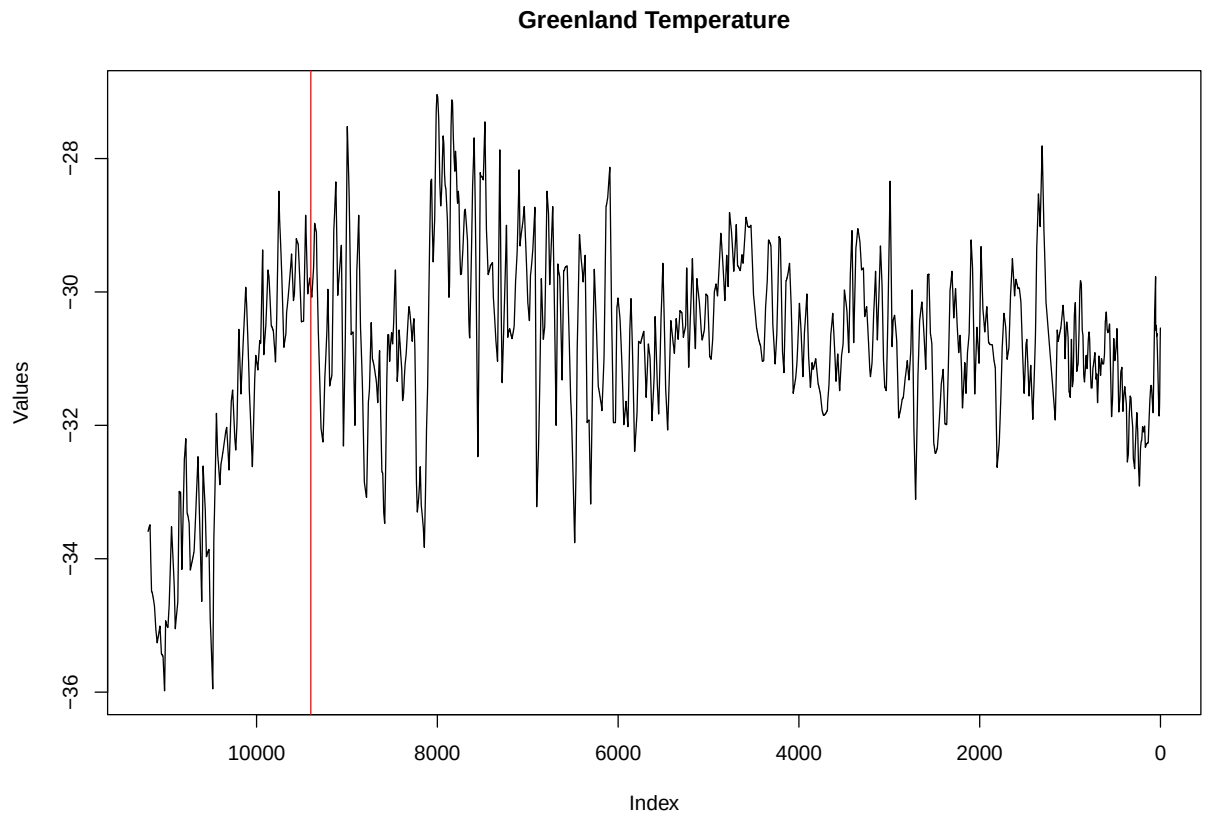
Let us see some descriptive measure studies.

```
      Age        Greenland_temperature   Upper_band       Lower_band
 Min.   :  -43   Min.   :-35.98        Min.   :-34.76   Min.   :-37.30
 1st Qu.: 2773   1st Qu.:-31.48        1st Qu.:-30.37   1st Qu.:-32.68
 Median : 5582   Median :-30.70        Median :-29.49   Median :-31.89
 Mean   : 5587   Mean   :-30.82        Mean   :-29.60   Mean   :-32.02
```

```
 3rd Qu.: 8403   3rd Qu.:-29.93        3rd Qu.:-28.66   3rd Qu.:-31.15
 Max.   :11223   Max.   :-27.04        Max.   :-25.48   Max.   :-28.19
```

The descriptive measures output shows that the youngest Age is -43(BP), and the oldest Age is 11580(BP). One unit is added to the age variable. In the data, the lowest temperature is at -39.21 degrees Celsius, and the highest is at -27.04 degrees Celsius. The average is -33.9 degrees Celsius, and the middle point is -30.74 degrees Celsius. Furthermore, the min, max, mean, and median values are almost the same for low and high values. It also shows 11 NULL(empty) rows in every column.-

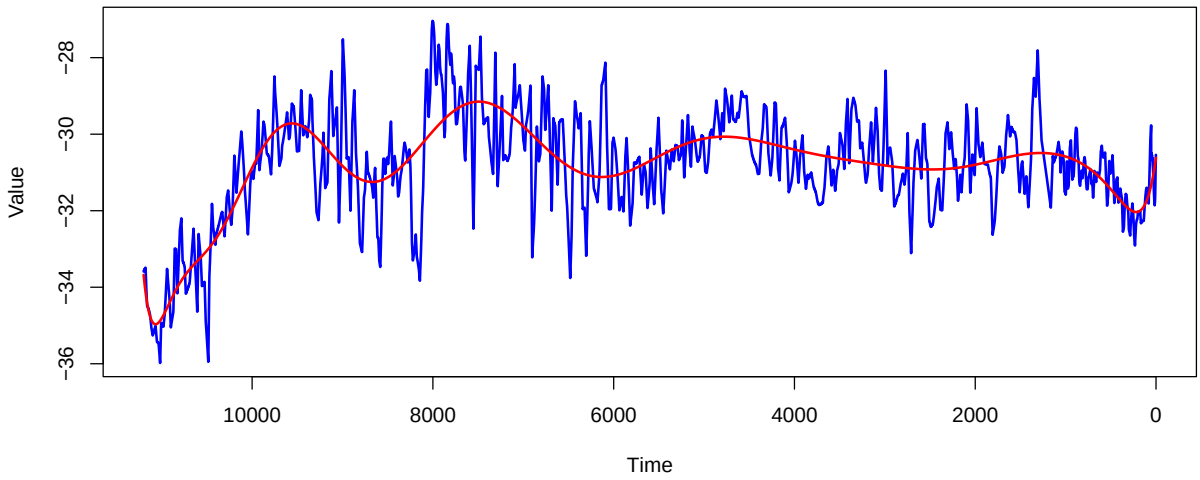
Line plot of Greenland Tempareture:

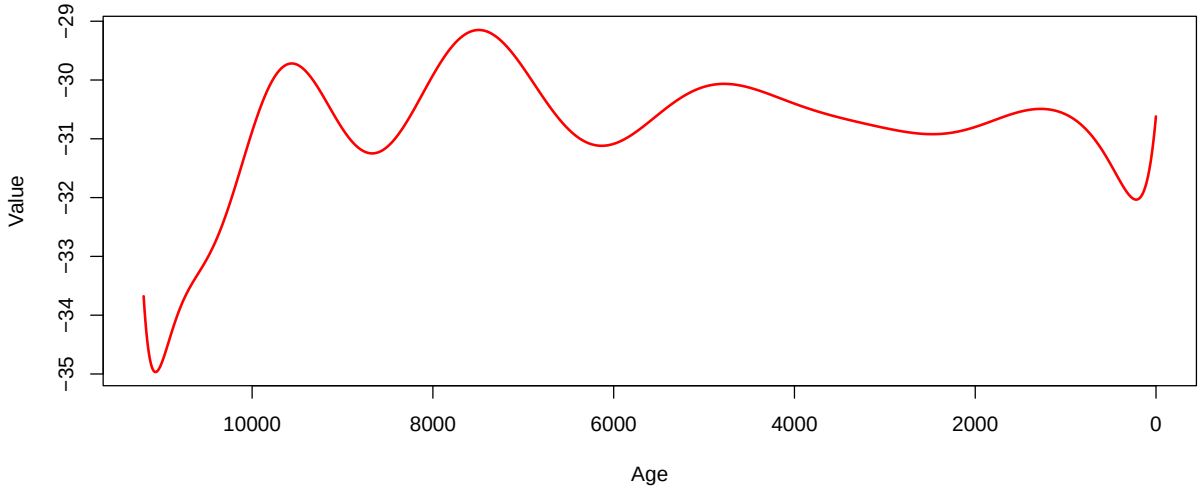


After If we look closely at the line plot, we can see that the Reverse X-indexing is done. It is not a mistake. On the left side of the plot, we can see the values from the past year, and on the right, we can see the current temperature. We have some outliers at the end of the time series.

Exploring Trend and Seasonality Patterns:

We used a 10-degree polynomial model to fit the data and then plotted it on the actual data. The primary purpose of this technique is to analyze time series data and identify any cyclic, seasonal, or trend patterns. This information can help conduct further analysis.





Based on the polynomial plot, it is evident that there is no discernible pattern between 0 and 9400. However, after 9400, there is a noticeable downward trend. Therefore, the data does not indicate seasonality or cyclic patterns.

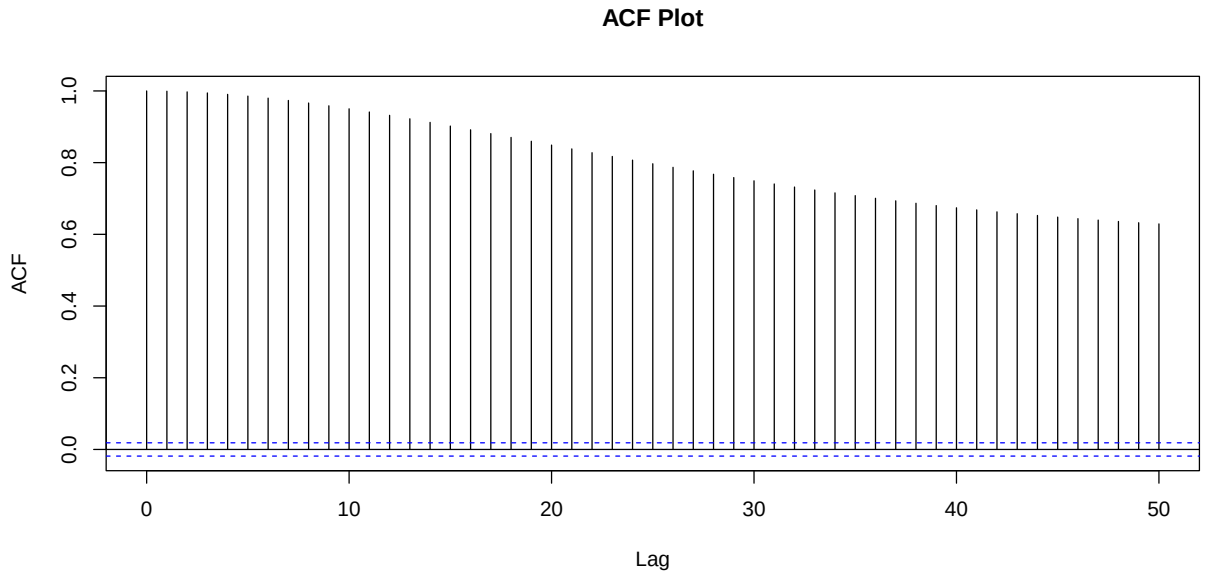
Augmented–Dickey–Fuller test:

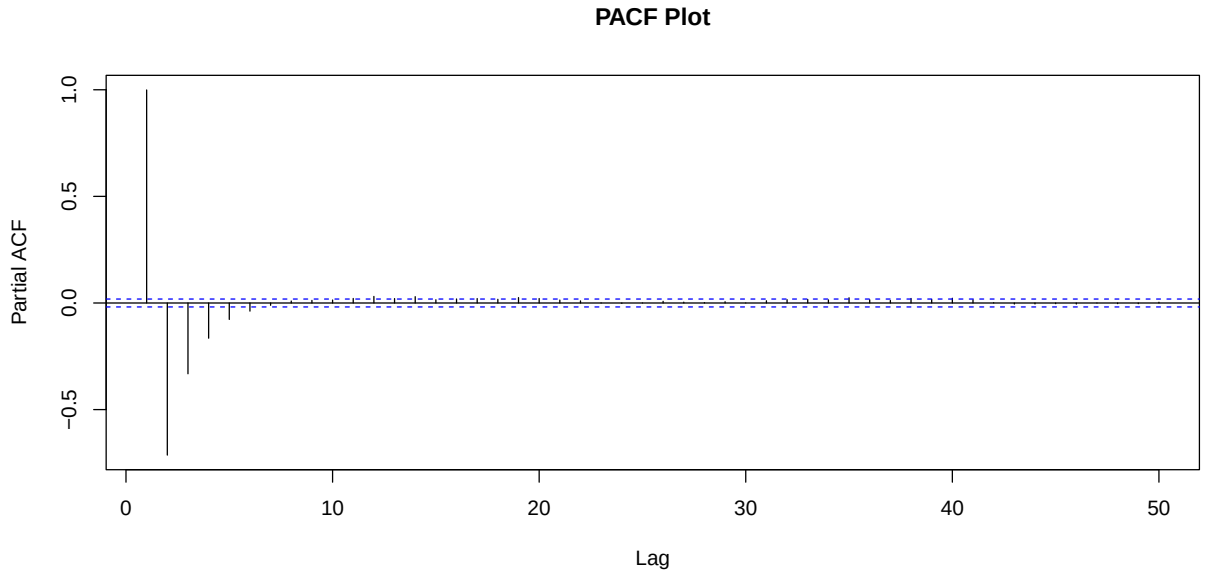
In time series analysis, "stationarity" means that the statistical properties of a series, like its mean, variance, and autocorrelation, stay the same over time. In other words, a stationary time series has a constant mean and variance, and its correlation with its lagged values (autocorrelation) does not change over time. On the other hand, a non-stationary time series is one in which the statistical properties change over time. Trends, cycles, or seasonal patterns can appear in time series that remain unchanged. If the time series is not stationary, it can cause problems while building a model, such as spurious correlations, instability, wrong parameter estimates, and bad forecasting. We need to check whether the time series is stationary to avoid these negative effects. There are several ways to determine whether a time series is stationary. we use the most common method, the Augmented–Dickey–Fuller (ADF) test.

```
Warning:  package 'tseries' was built under R version 4.2.3
Registered S3 method overwritten by 'quantmod':
method            from
as.zoo.data.frame zoo
Warning in adf.test(series):  p-value smaller than printed p-value

Augmented Dickey-Fuller Test

data:  series
Dickey-Fuller = -6.6162, Lag order = 22, p-value = 0.01
alternative hypothesis: stationary
```

Interpretation of given ADF test result:
Here -6.61 is the T statistic. This is a negative number; the more negative it is, the stronger the evidence against the null hypothesis of a unit root (nonstationary). Then we have p-value is used to determine the significance of the test. Here the value of p-values is 0.01, which is less than 0.05(significance level); we will reject the null hypothesis of a unit root and conclude that the time series is stationary. The function is determined with 22 lags. Overall, we fail to reject the null hypothesis of a unit root and conclude that the time series is stationary.

ACF and PACF plot
. In time series analysis, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are important tools that help us understand the nature of the time series data and choose the right models. The ACF plot shows the correlation between a time series and its lagged values. It helps us identify the order of the moving average (MA) component of the time series model. In particular, if the ACF plot shows a sharp drop-off after a certain number of lags, this means that the MA order of the model is equal to the number of lags where the drop-off happens. The PACF plot shows the relationship between a time series and its lags after intermediate lags have been taken into account. It helps us figure out the order of the autoregressive (AR) time series model. In particular, if the PACF plot shows a sharp drop-off after a certain number of lags, this means that the AR order of the model is equal to the number of lags where the drop-off happens. In conclusion, the ACF and PACF plots are important tools in time series analysis that can help us find the right models for the time series data and make accurate predictions.





From this ACF and PACF plot, we can tell that the ACF plot doesn't have a sharp drop-off. But the PACF plot clearly shows that there is a sharp drop-off after 2 or 3 lags, so we should choose the AR model with 2 or 3 orders.

Model selection

Data splitting:

Here the data is split into train and test datasets. The purpose of splitting the data is to evaluate the performance of the model on data that has not been used for model training, which helps to assess how well the model is likely to perform on new, unseen data. This can help us to assess the accuracy of the model and identify any potential issues, such as overfitting, which occurs when the model fits the training data too closely and does not generalize well to new data.

```
## [1] "Train: 8960"
## [1] "Test: 2240"
```

Here, the train and test series are split into 80-20. The train length is 8960, which is 80% of the actual series, and the test length is 2240, which is 20% of the actual series.

AR Model (Autoregressive Model)

. An Autoregressive (AR) model is a time series model that usually uses past values of the variable being modeled to predict future values. But here, we will perform a backcast using these model mean that the model will use present temperature and will predict past temperature. The basic idea behind an AR model is that the past value of the time series depends on its present values, and this dependence can be captured mathematically using a linear regression model.

an AR model can be represented as follows:

$$y_t = c + \varphi_1 y_{t+1} + \varphi_2 y_{t+2} + ... + \varphi_p y_{t+p} + \varepsilon_t$$

Where:

- y_t is the value of the time series at time t.
- c is a constant term.
- $\varphi_1, \varphi_2, ..., \varphi_p$ are the autoregressive coefficients, which represent the weights assigned to the present values of the time series.
- $y_{t+1}, y_{t+2}, ..., y_{t+p}$ are the present values of the time series up to p lags.
- ε_t is the error term, which represents the random variation in the time series that is not explained by the autoregressive model.

The order of the AR model is determined by the value of p, which represents the number of lags included in the model. For example, an AR(1) model uses only the present value of the time series at lag 1 to predict the past value, while an AR(2) model uses the present values at lags 1 and 2. To estimate the parameters of an AR model, maximum likelihood estimation or least squares estimation methods are commonly used. Once the model parameters are estimated, the model can be used to make predictions for past values of the time series. According to the PACF plot, the AR model is fitted with order p= 2.

```
Call:
arima(x = train, order = c(2, 0, 0))

Coefficients:
         ar1      ar2  intercept
      1.9155  -0.9174   -30.5777
s.e.  0.0042   0.0042     0.1085

sigma^2 estimated as 0.0003735:  log likelihood = 22640.62,  aic = -45273.25
```

Then the model will be,

$$y_t = -30.57 + 1.92y_{t+1} - 0.91y_{t+2} + \varepsilon_t$$

Intrepreation:
Interpret the coefficient estimates for the AR(1) and AR(2) terms. These represent the weights assigned to the present values of the time series. Here the estimated coefficient for AR(1) is 1.92; this means that each unit increase in the value of the time series at lag one is associated with a 1.92 unit increase in the past value of the time series, all else being equal. Similarly, if the estimated coefficient for AR(2) is -0.92, this means that each unit increase in the value of the time series at lag two is associated with a 0.92 unit decrease in the past value of the time series, all else being equal.

Measure	Value
AIC	-45273.248
BIC	-45244.846
HQIC	-45263.582

TABLE 1. Goodness of fit

Here we have AIC and BIC value , this measure commonly used measures of the goodness of fit of a model.

- $n =$ number of observations.
- $k =$ number of parameters to be estimated.
- $L_{max} =$ the maximized value of the log-Likelihood for the estimated model (fit the parameters by MLE).

$$AIC = \left(\frac{2n}{n-k-1}\right) k - 2ln\left[L_{max}\right]$$

$$BIC = kln(n) - 2ln(L_{max})$$

$$HQIC = 2ln[ln[n]]k - 2ln[L_{max}]$$

The lowest AIC, BIC, and HQIC values. This indicates that the model provides the best trade-off between model complexity and goodness of fit.

Back casting past temperature:
The total time series is split into 80-20 percentages. This means that 80% of time series data is considered train data, and remain dataset is considered test data.

After building an AR model based on train data, we want to use the model to predict past temperatures. Here, the size of the test data is the same size as the size of the back-calculated or predicted value because it will help to compare the predicted temperatures with the actual temperatures in the past.

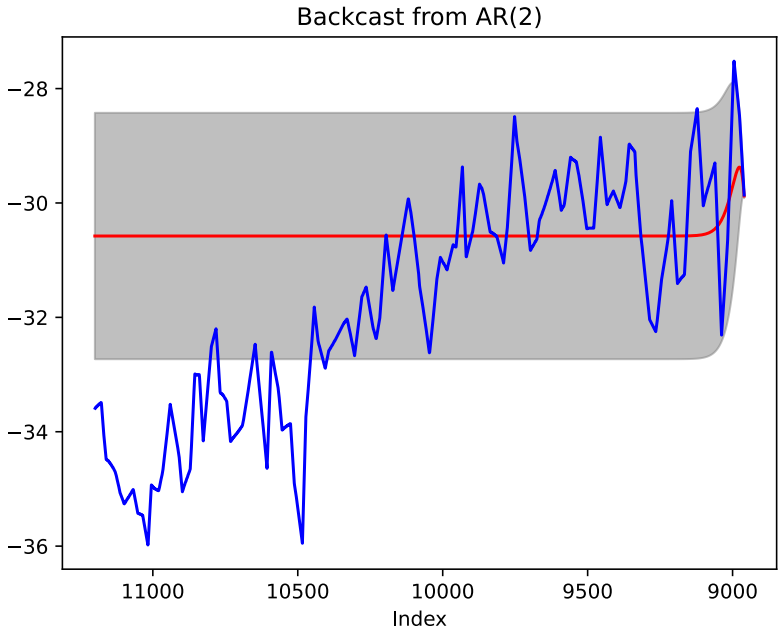


FIGURE 0.0.7. Backcast from AR(2)

Here plot backcasted value with actual temperatures and add a 95% confidence band. The red line is the backcasted temperature, the blue line is the actual past temperature, and the confidence band is denoted by the gray-shaded region. The diagram shows that the model is not performing well. The model is unable to capture trends at the end of the time series. Suppose we measure the goodness of fit of this AR model on test data. We got poor performance on test data.

Measures	Value
Mean Square Error	5.1
Root Mean Square Error	2.2
R-Squared	-0.3
AIC	3695.5
BIC	3706.9

TABLE 2. Goodness of fit measures

If we change the splitting percentage so that the train gets some trend parts, we can hope the model will do better.

So, The data is split with 85%,90%,95%, and 99% of the total series as training, and the rest of the data consider test data.

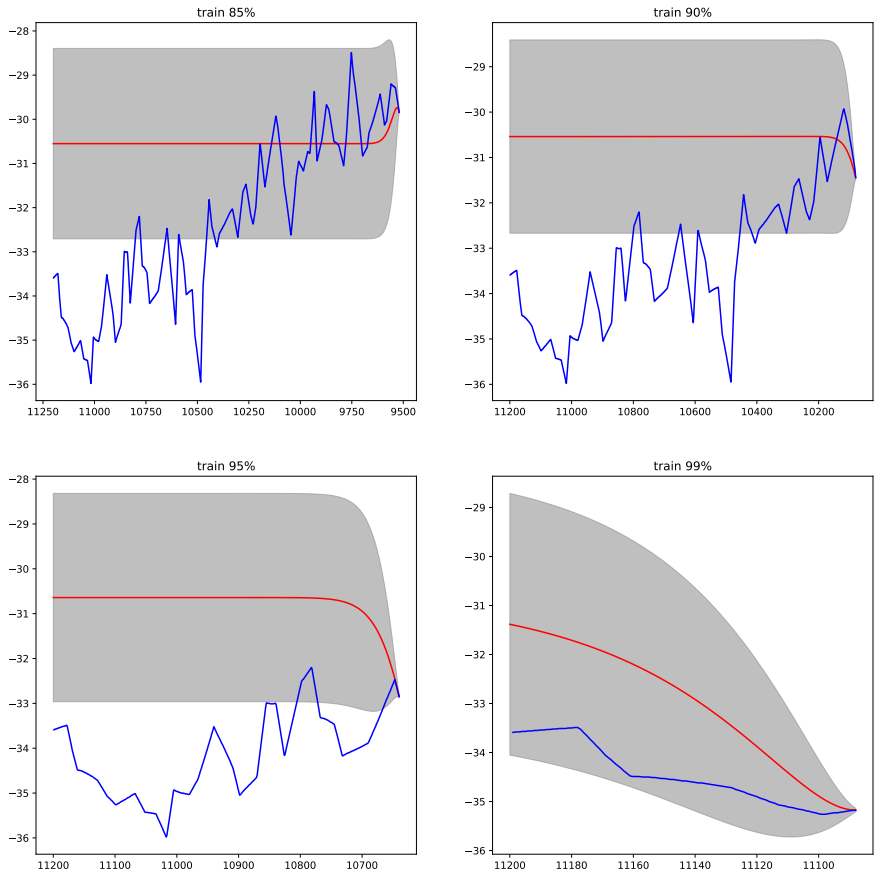


FIGURE 0.0.8. Backcast on multiple test data

Now the goodness of fit measure is used to check that the model performs better as we increase the test data size. But overall, this model was not performing well.

	MSE	RMSE	AIC	BIC	R-Squared
Train 85%	6.6	2.5	3181.6	3192.4	-0.8
Train 90%	9.6	3.1	2548.7	2558.7	-3.8
Train 95%	12.7	3.5	1429.1	1437.8	-14.9
Train 99%	2.5	1.6	110.6	116.1	-6.1

TABLE 3. Goodness fit for multiple train data

ARMA Model:

ARMA stands for Autoregressive Moving Average. It is a statistical model utilized frequently for time series analysis and forecasting. The autoregressive component (AR) and the moving average component (MA) are both included in the ARMA model. The ARMA model combines these two components. The general form of an ARMA model is represented as ARMA(p, q), where "p" represents the order of the autoregressive

component and "q" represents the order of the moving average component. The model can be written as follows:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}$$

where:

- Y_t represents the value of the time series at time t.
- c is a constant (intercept term).
- $\phi_1, \phi_2, ..., \phi_p$ are the autoregressive coefficients.
- ε_t represents the residual (prediction error) at time t.
- $\theta_1, \theta_2, ..., \theta_q$ are the moving average coefficients.

We attempted to fit an ARMA model here. A technique called grid search is used to determine the ARMA model's order (p,q). It entails analyzing numerous possible combinations of p and q values methodically and selecting the combination that results in the most accurate model. Here, We choose the (p,q) order with the smallest possible AIC value. This allows me to determine that the order is (3,5).

```
Call:
arima(x = train, order = c(3, 0, 5))

Coefficients:
         ar1      ar2     ar3      ma1     ma2     ma3      ma4     ma5
      2.8977  -2.8042  0.9064  -1.1259  0.1751  0.0194  -0.0482  0.0482
s.e.  0.0180   0.0344  0.0165   0.0208  0.0162  0.0163   0.0157  0.0122
      intercept
       -30.5619
s.e.     0.1337

sigma^2 estimated as 0.0003622:  log likelihood = 22777.75,  aic = -45535.51
```

Here the model will be,

$$Y_t = -30.56 + 2.9Y_{t+1} - 2.8Y_{t+2} + 0.9Y_{t+3} + \varepsilon_t - 1.12\varepsilon_{t+1} + 0.17\varepsilon_{t+2} + 0.02\varepsilon_{t+3} - 0.05\theta\varepsilon_{t+4} + 0.05\varepsilon_{t+5}$$

To evaluate the effectiveness of the ARMA(3,5) on the test data. If we backcast the model for test data, plot the predicted value against the actual test value and then evaluate the goodness of fit.

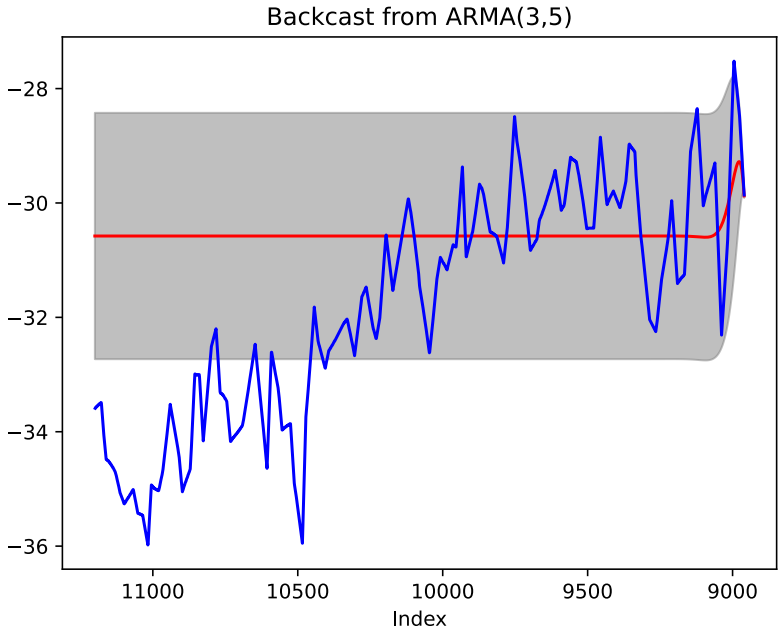


FIGURE 0.0.9. Backcast from ARMA(3,5)

Measures	Value
Mean Square Error	5.1
Root Mean Square Error	2.2
R-Squared	-0.3
AIC	3694.6
BIC	3706.1

TABLE 4. Goodness of fit measures

Based on the plot of the backcast and the measures of goodness of fits. This demonstrates that there has been no progress made since the AR(2) model was used.

So far, we have assumed that there is no trend or seasonality in the data. In that way, we tried to fit different models but could not find a good model. So there were trends and seasonality that didn't fit our models well. So now, using the Seasonal Decomposition method to do detrend and deseasonality. Then we fit the model with the residuals.

Desesonality and detrend

. Desesonality refers to the process of removing or reducing the seasonal component from a time series. Seasonality can introduce noise, making it difficult to analyze the underlying trends or make accurate forecasts. By deseasonalizing the time series, we can isolate the non-seasonal component and focus on the underlying patterns and trends.

Detrend, on the other hand, refers to the process of removing or estimating the trend component from a time series. The trend represents the long-term systematic change or directionality in the data. It reflects the underlying growth, decline, or stability of the phenomenon being observed, apart from any short-term fluctuations or seasonality. Removing the trend component can help in analyzing the residual fluctuations.

The use of the seasonal decomposition function in order to get rid of trends and seasonality. Decomposing a time series into its many components, such as its trend, seasonality, and residual (or error) component, can be accomplished through the use of a technique known as seasonal decomposition. It is useful for comprehending and evaluating the underlying patterns and structures that are contained within the data.

The process of decomposition involves the separation of the many components, which makes it possible to conduct a more in-depth investigation into the features and actions of each individual component.

The additive and multiplicative methods are the two primary methodologies that are utilized in seasonal decomposition.

1. **Additive Seasonal Decomposition:**

The time series is broken down into the sum of its trend, seasonal, and residual components by the use of the additive decomposition technique. The equation for the decomposition can be stated as follows:

$$Y(t) = \text{Trend}(t) + \text{Seasonal}(t) + \text{Residual}(t)$$

The trend component represents the long-term systematic change or directionality in the data.

2. **Multiplicative Seasonal Decomposition:**

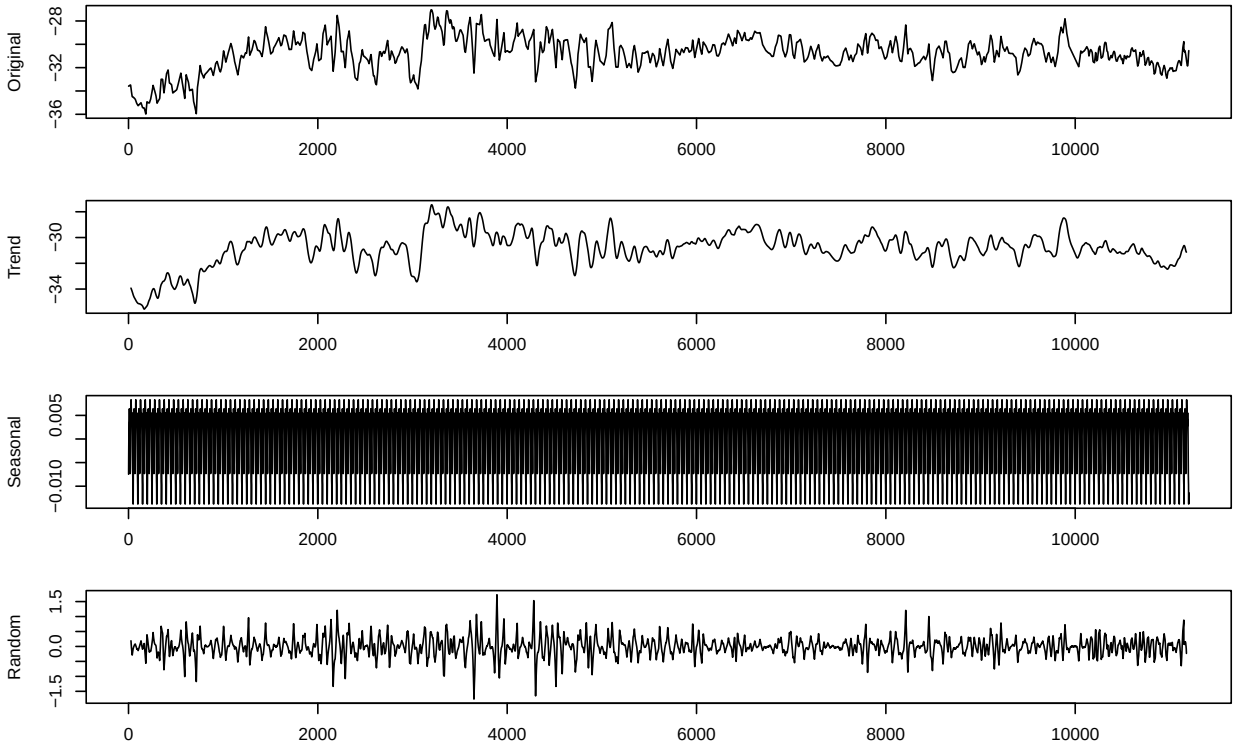
When using multiplicative decomposition, a time series is broken down into the product of its trend, seasonal, and residual components. This is one method of decomposing time series. The equation for the decomposition can be stated as follows:

$$Y(t) = \text{Trend}(t) * \text{Seasonal}(t) * \text{Residual}(t)$$

The trend and seasonal components are multiplied, and the residual component represents the remaining variations after removing the trend and seasonal patterns.

Calculation of the Residual Component can be calculated as the difference between the initial time series and the sum or product of the trend and seasonal components. This difference can be used in the calculation of the residual component.

Here, we worked with additive decomposition and then plotted the trend, seasonality, and residual.



ACF and PACF Plot:

Here, the ACF and PACF are plotted for the Residual component.

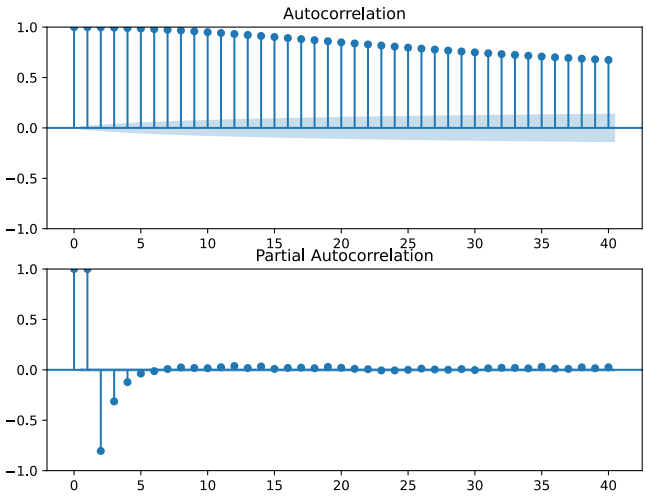


FIGURE 0.0.10. ACF & PACF Plot

From this ACF and PACF plot, we can tell that the ACF plot doesn't have a sharp drop-off. But the PACF plot clearly shows that there is a sharp drop-off after 2 or 3 lags, so we should choose the AR model with 2 or 3 orders.

Now,The residual components is split into train and test data. The train length is 8960, which is 80% of the actual series, and the test length is 2240, which is 20% of the actual series.

```
[1] "Train: 8960"
[1] "Test: 2240"
```

AR(2) model. According to the PACF plot, the AR model is fitted with order p= 2

```
Call:
arima(x = train, order = c(2, 0, 0))

Coefficients:
         ar1      ar2  intercept
      1.8513  -0.8671    -0.0004
s.e.  0.0053   0.0053     0.0123

sigma^2 estimated as 0.0003408:  log likelihood = 22988.23,  aic = -45968.46
```

Then the model will be,

$$y_t = 1.85y_{t+1} - 0.86y_{t+2} + \varepsilon_t$$

then plot the predicted value against the actual test value and evaluate the goodness of fit.

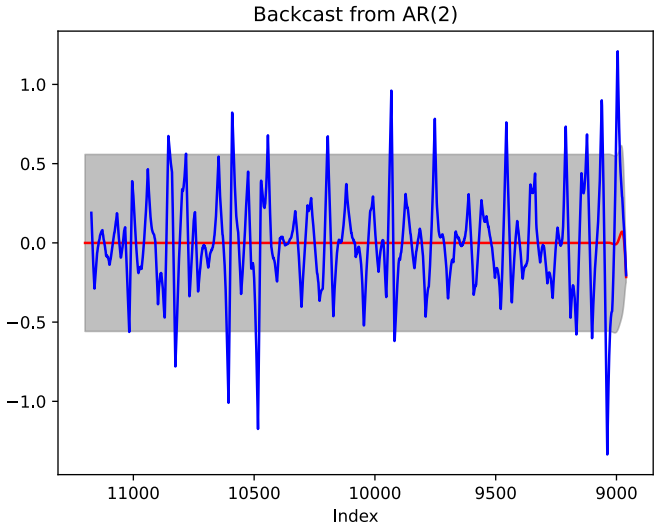


FIGURE 0.0.11. AR(2) on Residual Component

Measures	Value
Mean Square Error	0.1
Root Mean Square Error	0.3
R-Squared	0.004
AIC	-5474.4
BIC	-5462.9

TABLE 5. Goodness of fit measures

Here the goodness of fit between this and the previous one is nearly the same. There are no such differences from the previous one. So we have not tried any other model on this since there is no such difference in the outcomes.

Advanced Data Modeling Techniques: We have seen, using the ACF and PACF plots, that the value at the n^{th} position has a good or high correlation with the values at the $n+1^{th}, n+2^{th}, and\, n+3^{th}$ positions. So converted the series into a matrix like; if we have series like $(x_1, x_2, x_3, x_4, ...)$ it converted into a matrix that looks like

X			y
x_1	x_2	x_3	x_4
x_2	x_3	x_4	x_5
x_3	x_4	x_5	x_6
⋮	⋮	⋮	⋮
x_{n-3}	x_{n-2}	x_{n-1}	x_n

Here the first variables are considered dependent, and the rest y variable is as a predictor, which we have to predict.

```
Attaching package:  'dplyr'
The following objects are masked from 'package:stats':
      filter, lag
The following objects are masked from 'package:base':
      intersect, setdiff, setequal, union
```

```
       x1     x2     x3     x4      y
5  -30.54 -30.61 -30.72 -30.85 -31.02
6  -30.61 -30.72 -30.85 -31.02 -31.16
7  -30.72 -30.85 -31.02 -31.16 -31.32
8  -30.85 -31.02 -31.16 -31.32 -31.47
9  -31.02 -31.16 -31.32 -31.47 -31.58
10 -31.16 -31.32 -31.47 -31.58 -31.68
```

After converting, considered 80% of this data as a training dataset and the rest 20% as test data.

Because the independent variables have a strong correlation with the predictor variable, linear regression is an appropriate method to utilize when analyzing this data set. Therefore, we used a straightforward linear regression model to analyze the test data, and in order to evaluate the effectiveness of the model, we predicted its output for the test data. When we compared the actual value to the predicted value, we discovered an exciting phenomenon: the predicted values were so close to one another that they almost overlapped.

Now if we consider only 1st column of the matrix and fit a linear model on it, we also get better results. For this use of the linear model, the temperature takes reading at the n+kth time point and is used to make predictions about the nth time point data. Right now, we are analyzing the shift in the level of accuracy for k = 1,2,3,4,5,6,...

$x_t \sim x_{t+k}$	R-Squared	MSE	RMSE	AIC	BIC
k = 1	0.99	0.00	0.04	-13444.69	-13433.26
k = 5	0.98	0.05	0.23	-6388.43	-6377.01
k = 10	0.94	0.21	0.46	-3409.75	-3398.32
k = 15	0.89	0.4	0.63	-1998.6	-1987.17
k = 20	0.81	0.7	0.84	-766.61	-755.19
k = 25	0.72	1.04	1.02	111.8	123.22
k = 30	0.63	1.39	1.18	753.27	764.7

TABLE 6. Goodness of fit measures

Raising the k value shows that the accrual level is falling; Next, the linear model is fitted with k equal to 32.

	Values
Intercept	-6.22
Coefficient of x1	0.79
R-squared	0.63

TABLE 7. Linear Regression Model

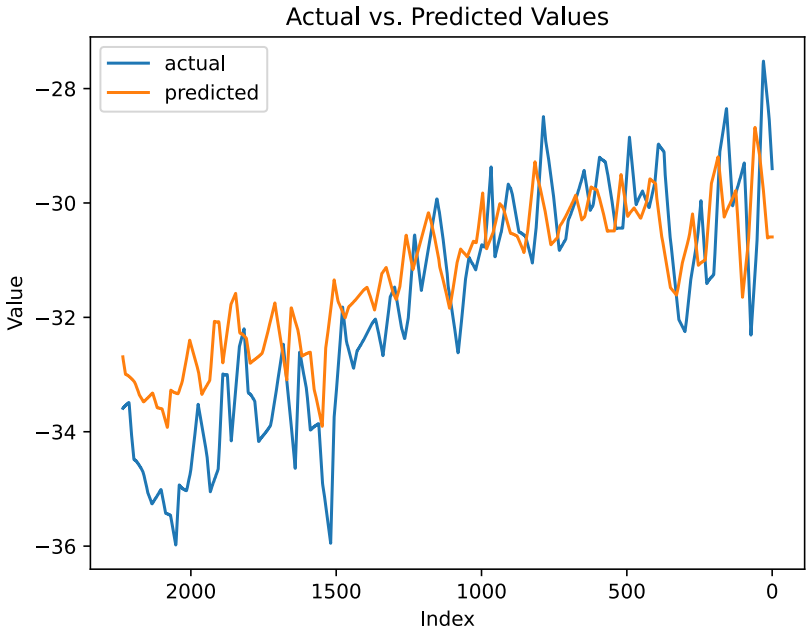


FIGURE 0.0.12. Plot of Actual vs. Predicted Values

In comparison to the other models that have been built, this one produced the best results. On the other hand, there are a few drawbacks. These drawbacks include the fact that if familiar with the nth phrase, one also must be familiar with the $(n+k)^{th}$ term. In that case, we won't be able to forecast the n^{th} term. Again, if we increase the value of k, the correlation between$x_{n+k}$and x_n will decrease, which will further lead to a reduction in the accuracy of the prediction.

A few different models are tried out in univariate modeling as methods for backcasting the temperature. However, we still need to achieve the level of accuracy that we want. After that, some variables factors are incorporated, so we transform the univariate data into a multivariate form. Finally, we use some modeling to determine the most accurate backcast.

Multivariate Analysis

In the pursuit of long-term backcasting through univariate analysis, challenges have emerged. Extensive exploration of research papers reveals that various studies have relied on multiple variables to reconstruct paleoclimate temperatures. Consequently, an investigation into processes that align with temperature has started, aiming to identify suitable variables for the primary objective. After thorough searching, several variables with potential have been discovered that are

- CO_2 (ppm)
- Volacnic Eruption Tempareture ($^{\circ}C$)
- Ice Surface Temperature ($^{\circ}C$)

Download dataset

Data Source:
Here the data taken from the Greenland 11,500 Year Temperature Reconstruction is a scientific study conducted by a team of investigators including Kobashi[**3**], T.; Menviel, L.; Jeltsch-Thömmes, A.; Vinther, B.M.; Box, J.E.; Muscheler, R.; Nakagawa, T.; Pfister, P.L.; Döring, M.; Leuenberger, M.; Wanner, H.; and Ohmura. The reconstruction of Greenland's temperature over the past 11,500 years is included in the dataset, in addition to the amount of carbon dioxide measured in parts per million and the time series of volcanic forcing, which illustrates the impact that volcanic activity and co2 have had on the climate. In addition to that, the climate model output for the temperature of the Northern Hemisphere is included

in the dataset. Both the study and the dataset were saved in the archives of the World Data Service for Paleoclimatology in Boulder, Colorado, as well as the NOAA Paleoclimatology Program at the National Centers for Environmental Information (NCEI). The following is a list of the site information that can be found for the Greenland 11,500 Year Temperature Reconstruction at the GISP2 site in Greenland:

- Site Name: GISP2
- Location: North America>Greenland
- Country: Denmark
- Northernmost Latitude: 72.6
- Southernmost Latitude: 72.6
- Easternmost Longitude: -38.5
- Westernmost Longitude: -38.5
- Elevation: 3200 m

The following information pertains to the Greenland 11,500 Year Temperature Reconstruction site, which is located at the GISP2 location in Greenland.

Download dataset

```
  Greenland_temperature Volcanic_Eruption_Temp Ice_Surface_Temp  Co2.ppm.
1                -30.54              0.5581339        0.6689402 0.7123848
2                -30.61              0.5581339        0.6689402 0.7066336
3                -30.72              0.5581339        0.6689402 0.6975958
4                -30.85              0.5581339        0.6689402 0.6869149
5                -31.02              0.5581339        0.6689402 0.6729475
6                -31.16              0.5581339        0.6689402 0.6614449
```

```
[1] "Rows: 11556"
[1] "Columns: 4"
 Greenland_temperature Volcanic_Eruption_Temp Ice_Surface_Temp
 Min.   :-39.21        Min.   :-1.570848      Min.   :-0.78926
 1st Qu.:-31.58        1st Qu.:-0.147097      1st Qu.:-0.02904
 Median :-30.74        Median :-0.006063      Median : 0.09097
 Mean   :-30.99        Mean   :-0.094403      Mean   : 0.01219
 3rd Qu.:-29.96        3rd Qu.: 0.071819      3rd Qu.: 0.12993
 Max.   :-27.04        Max.   : 0.558134      Max.   : 0.66894
    Co2.ppm.
 Min.   :-0.005795
 1st Qu.: 0.625216
 Median : 0.696053
 Mean   : 0.675617
 3rd Qu.: 0.760750
 Max.   : 1.202974
```

Data description:

Based on what was given, we can figure out the following:

Rows: 11556: There are 11,556 rows or observations in the data set.

Columns: 4: The dataset has four columns, which are also called variables.

These summaries give us a general understanding of the temperature and CO2 concentration measurements across the dataset. However, it's important to note that additional context and analysis are necessary to fully interpret and draw meaningful insights from the data. The summary provides information about the distribution of values within each variable. It includes measures such as minimum, first quartile, median, mean, third quartile, and maximum. These statistics allow you to understand the range, central tendency, and variability of the data in each column.

Explanatory data analysis:[5]

.

Line plot:

I'm going to plot these variables against the temperature of Greenland right now

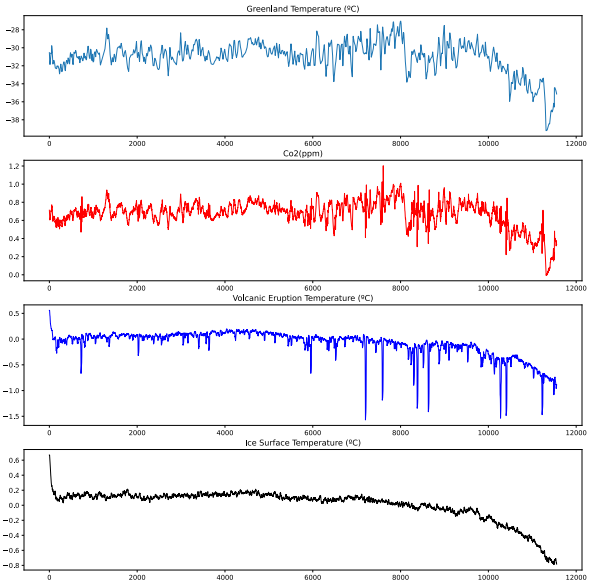


FIGURE 0.0.13. Line plot of variables

When looking at this figure, one may get the conclusion that each variable has patterns that are analogous to and similar to those seen in the other variables. Right now, I'm looking into how closely those two variables are correlated with one another.

```
                     Greenland_temperature Volcanic_Eruption_Temp
Greenland_temperature            1.0000000              0.5322202
Volcanic_Eruption_Temp           0.5322202              1.0000000
Ice_Surface_Temp                 0.6817255              0.8217696
Co2.ppm.                         0.9868701              0.5638183
                     Ice_Surface_Temp  Co2.ppm.
Greenland_temperature        0.6817255 0.9868701
Volcanic_Eruption_Temp       0.8217696 0.5638183
Ice_Surface_Temp             1.0000000 0.6734031
Co2.ppm.                     0.6734031 1.0000000
```

Correlation Matrix: The correlation coefficients between Greenland's temperature, volcanic eruptions, ice surface temperature, and concentration of carbon dioxide in the atmosphere are moderately positive. These correlations suggest that higher temperatures in Greenland are associated with higher temperatures during volcanic eruptions. The correlation between Volcanic Eruption Temp and Ice Surface Temperature is high, with 0.822 suggesting that higher temperatures during volcanic eruptions are directly tied to higher temperatures on the surface of the ice. Volcanic Eruption Temperature and CO2.ppm have a moderate association, with 0.564 showing a correlation between greater temperatures and higher quantities of carbon dioxide in the atmosphere. Ice Surface Temperature and Co2.ppm have a slightly positive correlation.

Scatter Plot: In addition to that, each variable is plotted alongside the corresponding temperature values.

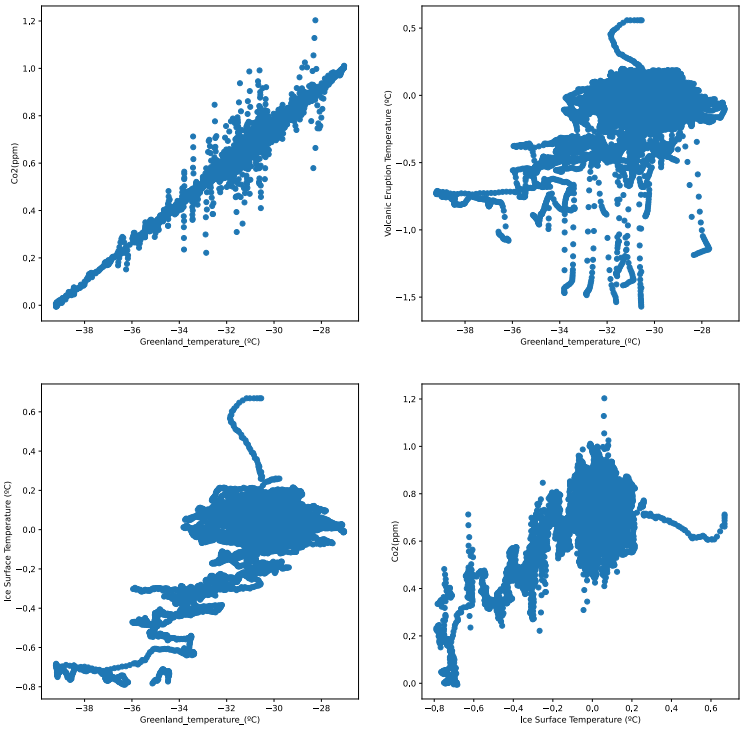


FIGURE 0.0.14. Scatter Plot of each variable

These scatter plots offer evidence that bolsters our correlation, which you can see here. Given that there is a positive correlation between the variables,It was able to determine that the basic linear model is the one that provides the best fit for this data.

Model Fitting:

MinMax Scaling:
In this instance, we have preprocessed the dataset using MinMax scaling before fitting models. MinMax scaling is a common method used to normalize the range of numeric characteristics to a certain range, which is commonly between 0 and 1. This transformation makes sure that all of the features have the same scale and prevents any one feature from taking over the learning process of the model.

Split the Dataset:
Before attempting to fit the model, the dataset was partitioned into 80-20 percentiles, with 80 percent of the data serving as the training set and 20 percent as the testing set, respectively, and then attempted to fit the model. We usually split the data set in such a manner to find the best-fitted model. This also helps prevent our model from overfitting.

```
Train rows:  9245
Test rows:  2311
```

Linear Regression Model:[8]
. Linear regression is a technique used to establish a relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to find the best-fitting line that represents the data. The line is defined by an intercept (bias) and coefficients (slopes) assigned to each independent variable.

Here are some key components and concepts related to linear regression:

(1) Dependent Variable: Also known as the target variable or response variable, it is the variable that we are trying to predict or explain based on the independent variables.
(2) Independent Variables: Also known as predictor variables or features, these are the variables used to predict or explain the dependent variable
(3) Assumptions: Linear regression assumes that there is a linear relationship between the independent and dependent variables. It also assumes that the residuals (the differences between the observed and predicted values) are normally distributed and have constant variance (homoscedasticity). Additionally, it assumes that the independent variables are not strongly correlated with each other (multicollinearity).
(4) Equation: The equation of a linear regression model is represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$$

Where:
- Y is the dependent variable.
- β_0 is the intercept or bias term.
- $\beta_1, \beta_2, ..., \beta_n$ are the coefficients (slopes) associated with the independent variables $X_1, X_2, ..., X_n. and \varepsilon$ is the error term or residual.

5 Coefficients: The coefficients $(\beta_1, \beta_2, ..., \beta_n)$ represent the impact or contribution of each independent variable on the dependent variable. They indicate the change in the dependent variable's value for a one-unit change in the corresponding independent variable, assuming all other variables are held constant.

Here is the Linear model after fitting Linear Regression model:

$$y = -0.1 - 0.09X_1 + 0.08X_2 + 1.24X_3$$

Where,

- y : Greenland Temperature
- X_1 :Volacnic Eruption Tempareture $(^\circ C)$
- X_2 :Ice Surface Temperature $(^\circ C)$
- X_3 :CO_2 (ppm)

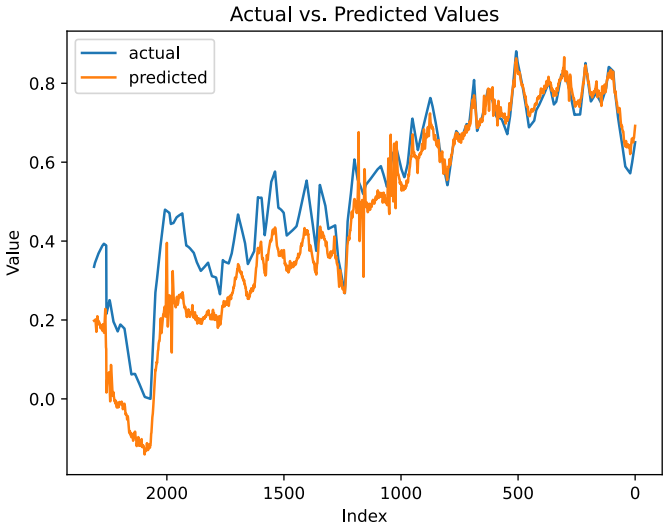


FIGURE 0.0.15. Model prediction data

Measures	Value
Mean Square Error	0.01
Root Mean Square Error	0.1
R-Squared	0.75
AIC	-10573.26
BIC	-10556.02

TABLE 8. Goodness of fit measures

The fitted model yielded an R-squared value of 0.75, a Root Mean Square of 0.1, an AIC value of -10573.26, and a BIC value of -10556.02, indicating a good fit. The accuracy of the model is approximately 75%, suggesting its reliability. However, numerous models were tested in an effort to enhance the accuracy rate.

Passive Aggressive Regression Model:[**8**]

.

The Passive Aggressive Regression model is particularly useful when dealing with large datasets or scenarios where the data distribution may change over time. It allows for incremental updates and can adapt to changes in the data distribution by adjusting its parameters in response to new observations. It is known for its efficiency and ability to converge quickly, making it suitable for real-time applications or situations where computational resources are limited. So, we are fitting the Aggrative Regression Model, which will update the coefficient of the linear model to get a good prediction.

Passive-aggressive regression is an algorithm used for regression tasks where the goal is to predict a continuous target variable based on a set of input features.

In passive-aggressive regression, the algorithm updates its model incrementally, adapting to new training samples one at a time. It is particularly useful in scenarios where the data distribution may change over time or when computational resources are limited.

The passive-aggressive regression algorithm works as follows:

(1) Initialize the model: Start with a regression model, such as a linear regression model, and set its coefficients to zero.
(2) Iterate through the training samples: For each training sample, do the following steps:

(a) Make a prediction: Use the current model to predict the target variable based on the input features.
(b) Calculate the error: Compute the error between the predicted and target values.
(c) Update the model: Adjust the model's coefficients based on the error and a predefined aggressiveness parameter.

(3) Repeat step 2 for all training samples until convergence or a maximum number of iterations is reached.

The aggressiveness parameter determines how much the model's coefficients should be updated based on each training sample. A higher aggressiveness value leads to larger updates, potentially allowing the model to adapt quickly to changes in the data. However, a high aggressiveness value may make the model more sensitive to noisy or irrelevant features.

Passive-aggressive regression provides a flexible and efficient approach for online learning and incremental model updates in regression tasks. It can be a useful tool in situations where data is continuously arriving or where computational resources are limited.

Here is the final model after passive-aggressive regression algorithm fitting:

$$y = -0.01 - 0.07X_1 + 0.06X + 1.08X_3$$

Where,

- y : Greenland Temperature
- X_1 :Volacnic Eruption Tempareture ($^{\circ}C$)
- X_2 :Ice Surface Temperature ($^{\circ}C$)
- X_3 :CO_2 (ppm)

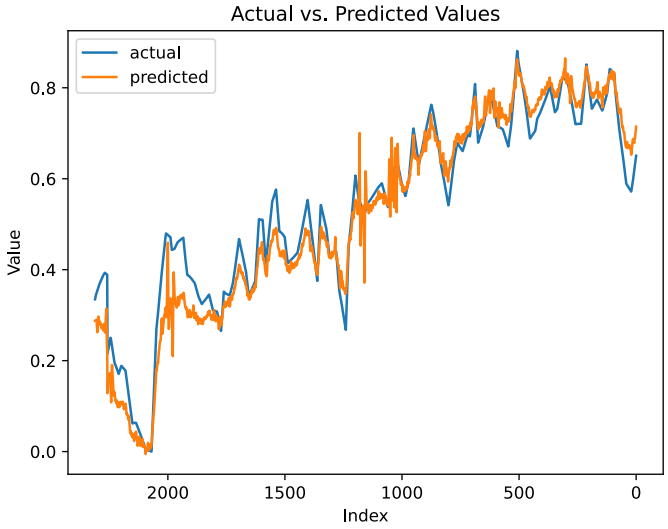


FIGURE 0.0.16. Model prediction data

Measures	Value
Mean Square Error	0.003
Root Mean Square Error	0.05
R-Squared	0.94
AIC	-13705.16
BIC	-13687.92

TABLE 9. Goodness of fit measures

The fitted model yielded an R-squared value of 0.94, a Root Mean Square of 0.05, an AIC value of -13705.16, and a BIC value of -13687.92, indicating a strong fit. The model accuracy is approximately 94%, suggesting its high level of accuracy. Despite achieving good accuracy in this model, further exploration will be conducted to ascertain if another model can yield even higher accuracy.

Our dataset has so many outliers that we are looking for a model that optimizes these outliers and updates the coefficients of the linear model. There exists a Huber regression model which can help us in this situation. The Huber regression model is a robust regression technique that can be advantageous in the presence of outliers or data points with a strong influence on the model. It assigns less weight to outliers, allowing for a more robust estimation of the regression parameters.

Huber Regression Model:[**8**]
. The Huber regression model, also known as Huber loss or Huber M-estimator, is a robust regression technique that combines the advantages of both ordinary least squares (OLS) regression and robust regression methods. It is designed to handle outliers and data points with a strong influence on the regMudelsee, M., 2010. Climate time series analysis. Atmospheric.ression model.

In traditional OLS regression, the squared difference between the observed and predicted values is minimized. However, OLS regression is sensitive to outliers, as the squared term amplifies the effect of large residuals. Robust regression methods, on the other hand, down-weight or ignore outliers to mitigate their impact on the model. However, these methods may not perform well when the proportion of outliers is high or when influential data points are present.

The Huber loss function compromises OLS and robust regression methods by combining squared and absolute differences. The Huber loss function is defined as:

$$L(y, y_{pred}) = \begin{cases} \frac{1}{2}(y - y_{pred})^2 & if\ |y - y_{pred}| \leq q \\ q(|y - y_{pred}| - \frac{q}{2}) & if\ |y - y_{pred}| > q \end{cases}$$

where:

- y is the observed value of the dependent variable.
- y_{pred} is the predicted value of the dependent variable.
- q is a threshold parameter determining the point at which the loss function transitions from squared to absolute.

In the Huber loss function, when the absolute difference between the observed and predicted values $|y - y_{pred}|$ is less than or equal to the threshold (q), the loss is quadratic, similar to the squared difference in OLS regression. When the absolute difference exceeds the threshold, the loss becomes linear, similar to the absolute difference in robust regression.

By using a threshold, the Huber loss function effectively handles outliers. When the absolute difference is small (within the threshold), the model is less influenced by those points, similar to robust regression. However, when the absolute difference is large (beyond the threshold), the model is more influenced by those points, similar to OLS regression.

The Huber regression model minimizes the sum of the Huber loss function across all data points to estimate the coefficients (slopes) of the linear regression equation. This can be done using various optimization algorithms, such as gradient descent or iteratively reweighted least squares (IRLS).

The threshold parameter q choice determines the trade-off between robustness and efficiency. A larger q value makes the model more robust to outliers but less efficient for non-outlying data points, while a smaller q value balances robustness and efficiency.

Linear Model for q = 1:

$$y = 0.09 - 0.04X_1 + 0.01X_2 + 0.95X_3$$

Where,

- y : Greenland Temperature
- X_1 :Volacnic Eruption Tempareture ($^\circ C$)
- X_2 :Ice Surface Temperature ($^\circ C$)
- X_3 :CO_2 (ppm)

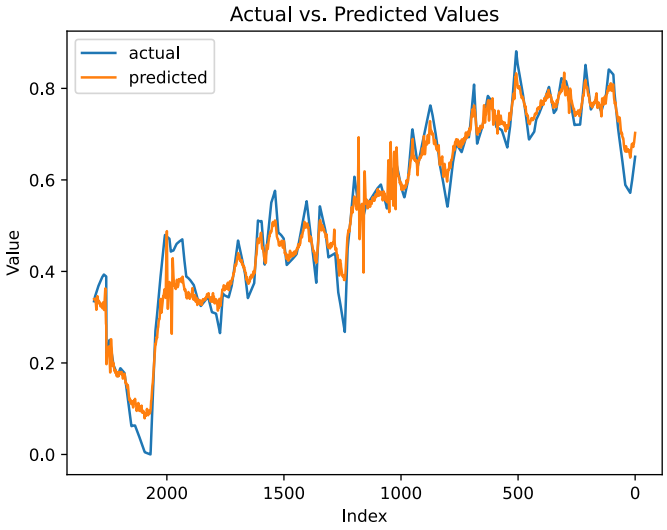


FIGURE 0.0.17. Model prediction data

Measures	Value
Mean Square Error	0.001
Root Mean Square Error	0.04
R-Squared	0.96
AIC	-15020.12
BIC	-15002.89

TABLE 10. Goodness of fit measures

We were able to determine that our fitted model had an R-squared value of 0.96, a Root mean square of 0.04, an AIC value of -15020.12, and a BIC value of -15002.89 from the goodness of fit. After completing the linear modeling process and calculating the goodness of measure, the current accuracy of this linear model is 96 percent, which is a significant improvement over the accuracy of the models I've used in the past.

Conclusion

. We have attempted to backcast temperature using a number of different models and methods that are univariate. However, We have not yet achieved the level of accuracy We desire, and there are issues with long-term prediction. After that, We decided to add some pricey factors, transform our univariate data into multivariate data, and finally use a linear model to obtain the backcast. This procedure yielded a satisfactory outcome for us in comparison. Rather than utilizing these two methods of analysis, we should try our hand at some other approaches. If we do so, we may have more success. Because of the constraints placed on our time, we have not attempted any other analyses; nonetheless, it is still possible to conduct a more thorough investigation.

Challenges and Limitations

During the course of working on this project, we ran into a number of challenges. Several of these issues include the following:

(1) Topic: Regarding the subject at hand, we have no prior experience with it. Regarding this project, we have no prior knowledge whatsoever. We are not familiar with the process of paleoclimate analysis. Therefore, in order to comprehend paleoclimate analysis. We have a lot of research papers that we need to read.

(2) Gathering information: Once more, to acquire data, we had to deal with many challenges. The collection of the dataset has evolved into a difficult challenge. In most cases, the collection of these data requires a significant amount of people in addition to a significant sum of money. A researcher needs to go to a challenging location, where there is a possibility that they will put their lives in danger, in order to acquire this kind of data. As a result of all of these factors, the internet does not make this kind of material easily accessible to the public. A few government websites publish research papers on the studies carried out on paleoclimate analyses. And additionally, publish the dataset that was utilized in their research. Finding raw data, as well as data that has been combined from a variety of variables, can be challenging.

(3) Preprocessing: The purpose for which we have decided to start this project, and the accomplishment of my objective, it has become quite tough to acquire the data. In the vast majority of instances, the quantity of the dataset was insufficient. There were a few instances in which we obtained sufficient data; however, there was a deficiency in comparable information. According to what we've read on the internet, Greenland has some number of research centers, each of which is distinctive from the others in a number of important respects. Each facility has applied a uniquely individualized approach to scaling. Due to the fact that we have employed a variety of variables, we face a significant challenge in aggregating all of the data because each variable uses a unique scale mechanism.

(4) Statistical analysis: We have got no similar objective in any of the previous study papers. If we had any studies on this topic, then it could be possible for me to make significant progress toward achieving the goal we have set for myself. At the time of analysis, the data cannot accurately capture whether there is a seasonality, trend, or cyclic pattern. Usual time series models do not provide reasonable predictions for this data

(5) Time Constraints: Because my available time is restricted, we have fitted some of the models. However, this data can be utilized in a variety of analytical directions using a wide variety of models and approaches. By doing so, we are able to do further and more in-depth studies by employing these data.

Source Code

Github Link : https://github.com/shubhamoypaul-svmy/PU_ Time_ Series_ Project/tree/master.

Bibliography

[1] Davidson, J.E., Stephenson, D.B. and Turasie, A.A., 2016. Time series modeling of paleoclimate data. Environmetrics, 27(1).

[2] Mudelsee, M., 2010. Climate time series analysis. Atmospheric and, 397.

[3] Kobashi, T. et al. (2017) Volcanic influence on centennial to millennial holocene greenland temperature change, Nature News. Available at: https://www.nature.com/articles/s41598-017-01451-7 (Accessed: 30 May 2023).

[4] Li, M., Hinnov, L. and Kump, L., 2019. Acycle: Time-series analysis software for paleoclimate research and education. Computers & Geosciences, 127.

[5] Davidson, J.E., Stephenson, D.B. and Turasie, A.A., 2016. Time series modeling of paleoclimate data. Environmetrics, 27(1).

[6] National Centers for Environmental Information (NCEI), National Centers for Environmental Information (NCEI). Available at: https://www.ncei.noaa.gov/access/paleo-search/study/17825 (Accessed: 30 May 2023).

[7] Kobashi, T., Box, J.E., Vinther, B.M., Goto-Azuma, K., Blunier, T., White, J.W.C., Nakaegawa, T. and Andresen, C.S., 2015. Modern solar maximum forced late twentieth century Greenland cooling. Geophysical Research Letters, 42(14)

[8] Supervised learning , scikit. Available at: https://scikit-learn.org/stable/supervised_learning.html (Accessed: 31 May 2023).