Modelling and Forecasting of Monthly Temperature Time Series Data

Neha Chauhan - chneha@iitk.ac.in - 170428

Desarda Rushabh - rushabhd@iitk.ac.in - 180230

Venkatesh - pvenkat@iitk.ac.in - 180514

Sunil Dhaka - sunild@iitk.ac.in - 170735

Abstract

Time series modelling and forecasting – a method that predicts future values by analysing past values plays an important role in many practical fields.The forecasting of temperature on a seasonal time scale has been attempted by many researchers by different techniques at different times across the globe. It is a challenging task to forecast temperature on a monthly and seasonal time scale. In this paper, an attempt has been made to develop a Seasonal Autoregressive Integrated Moving Average (SARIMA) model to long term temperature data of **India** for the period of 117 years (1901-2017).

Introduction

The main goal of time series modelling is to collect and analyse past values to develop appropriate models that describe the inherent structure and characteristics of the series. Time series forecasting is the use of a certain model to forecast future values based on past observed values, and thus can be understood as a method for predicting future values by understanding past values. The importance of time series forecasting in countless practical fields, researchers should pay proper attention to fitting an appropriate model to the time series. Over the years, many intelligent time series models have been developed in the literature to improve the accuracy and efficiency of time series forecasting. One of the most widely used and recognised statistical forecasting time series models is the Autoregressive Integrated Moving Average (ARIMA) model. The ARIMA model is well-known for notable forecasting accuracy and efficiency in representing various types of time series with simplicity as well as the associated, Box–Jenkins methodology for optimal model construction. The basic assumption made in implementing this model is to assume the time series is linear and follows a statistical distribution, such as the normal distribution. For seasonal time series forecasting, Box and Jenkins proposed a quite successful variation of the ARIMA model called the Seasonal ARIMA (SARIMA) model.

Air temperature is a common meteorological variable indicative of how hot or cold the air is. It not only affects the growth and reproduction of plants and animals, but also has an influence on nearly all other meteorological variables, such as the rate of evaporation, the relative humidity, wind speed, wind direction and precipitation patterns. In this project , we analyse the monthly mean temperature of India, during 1901–2017.

Data

In this project we used monthly temperature data of India .We obtained this data from the National Government Data Website: data.gov.in

- The dataset contains - Temperature of each month over 117 years (1901 - 2017) of India in Celsius.

- Each row specifies a year(1901 - 2017), and each column specifies temperature for a particular month.
- There were no missing values in the dataset.

Methodology

In this project we are using the **Box Jenkins Method** for data analysis and forecasting.

The Box-Jenkins method refers to the iterative application of the following three steps:

- **Identification:** Using plots of the data, autocorrelations, partial autocorrelations, and other information, a class of simple ARIMA models is selected. This amounts to estimating appropriate values for p, d, and q.
- **Estimation:** The phis (ɸ) and thetas (Θ) of the selected model are estimated using least square estimation, maximum likelihood techniques etc.
- **Diagnostic Checking:** The fitted model is checked for inadequacies by considering the autocorrelations of the residual series (the series of residual or error values).

Main Outline of Paper

1. Visualise the time series
2. Stationarize the data
3. Plot ACF/PACF charts and find the optimal parameters
4. Build the ARIMA model
5. Make predictions
6. Conclusion

1 : Visualizing Data :

The first step is to visualize the data to understand what type of model we should use. We will check for the overall trend in our data. Also, look for any seasonal trends. This is important for deciding which type of model to use. We have used the predefined library functions to decompose the time series into trend, seasonal and residual components.

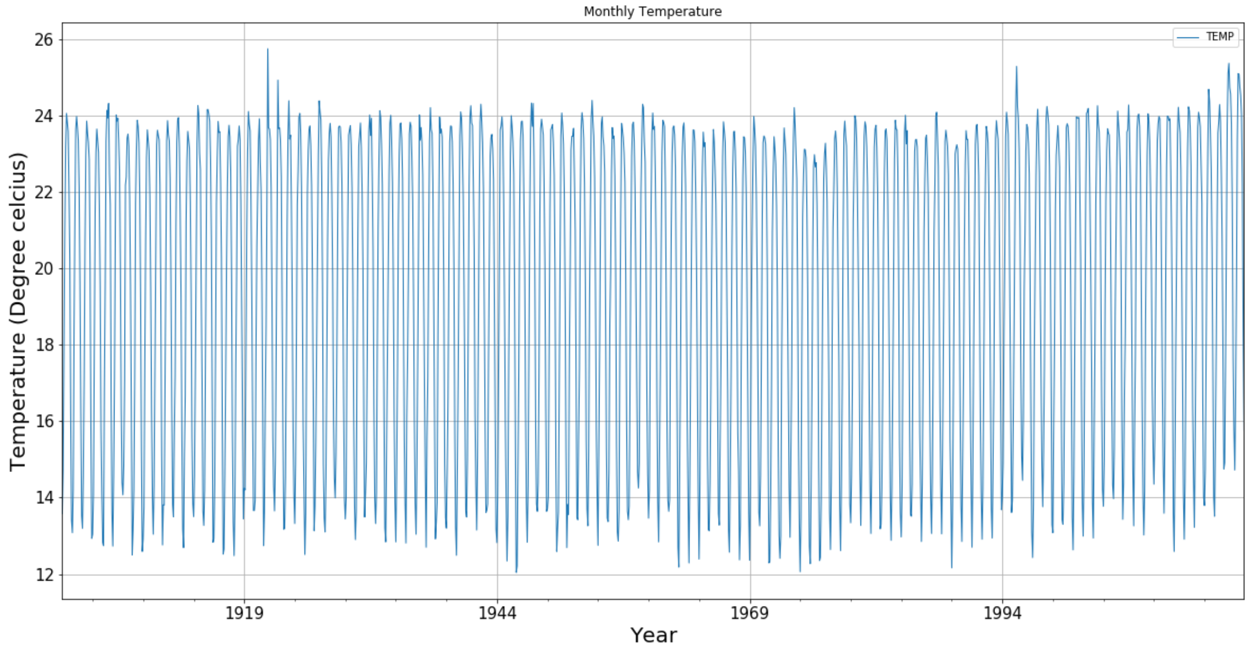


Figure 1. Monthly Temperature (normal Data)

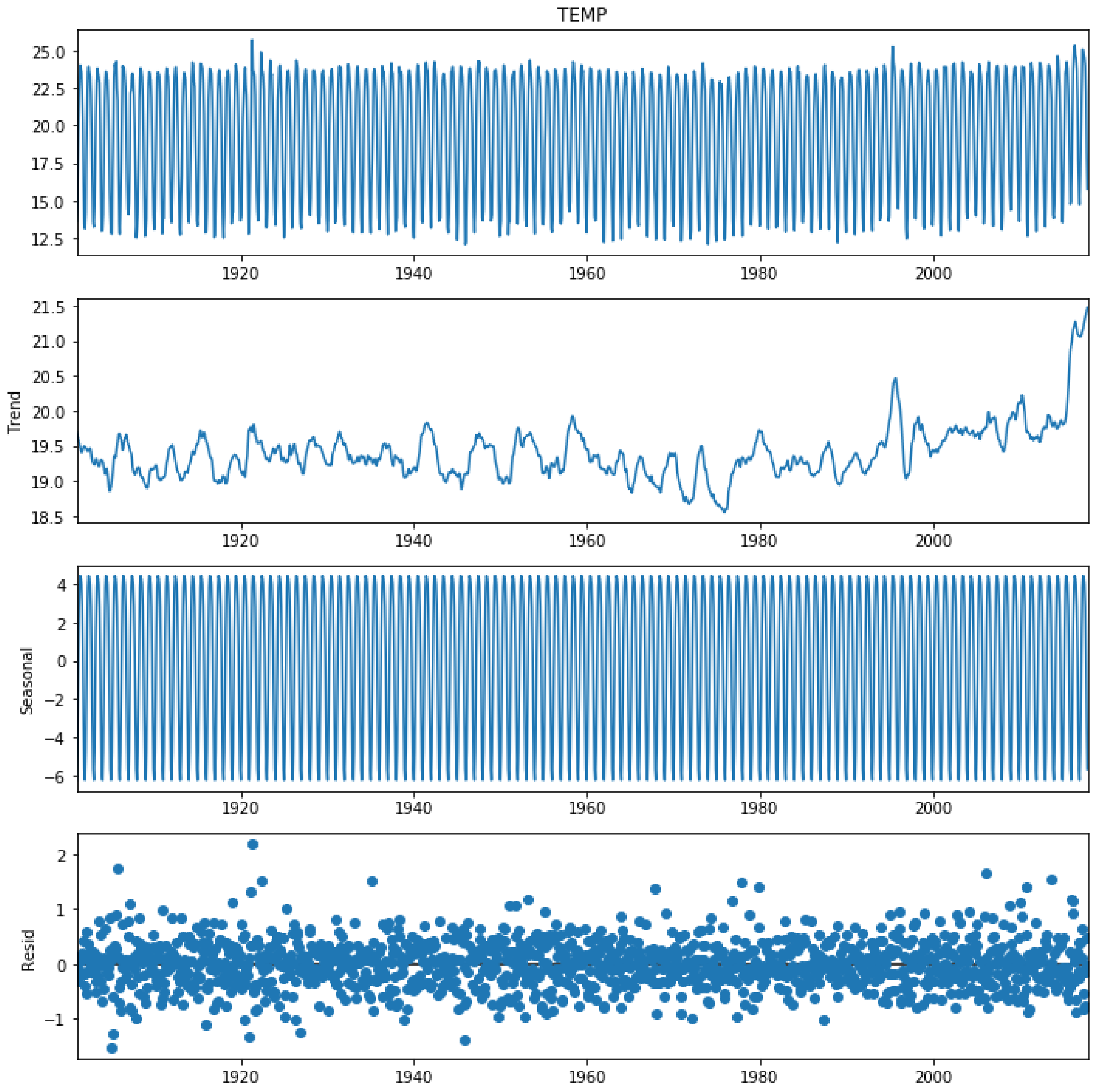


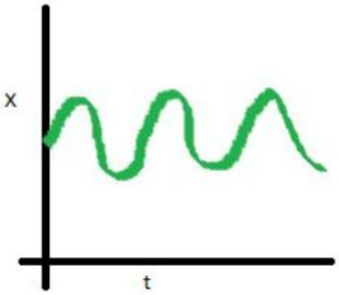
Figure 2. Decomposition of Data

2 : Stationarizing Data :

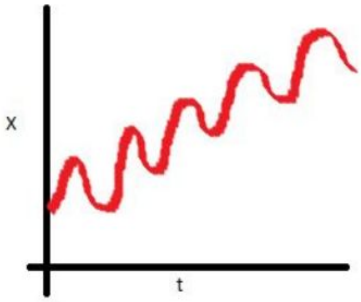
Definition-- $\{X_t\}$ is said to be a stationary process if for every n, and every admissible t_1 ,t_2 , ...t_n and any integer k, the joint distribution of $\{X_{t1}$,X_{t2} ,. . . .,$X_{tn}\}$ is identical to the joint distribution of $\{X_{t1+k}$,X_{t2+k} ,. . . .,$X_{tn+k}\}$.

Now, we will try to visualize what is meant by stationary time series process

1. The mean of the series should not be a function of time. The red graph below is not stationary because the mean increases over time.

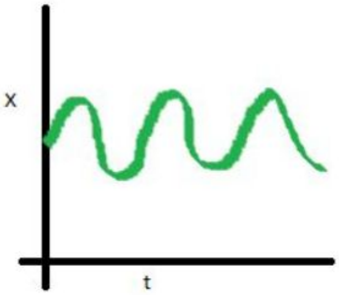


Stationary series

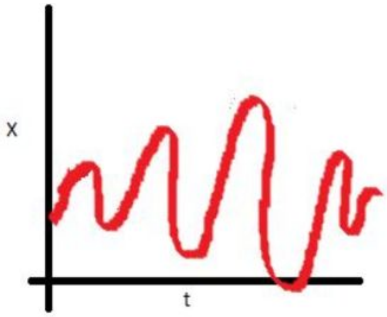


Non-Stationary series

2. The variance of the series should not be a function of time. This property is known as homoscedasticity. Notice in the red graph the varying spread of data over time.

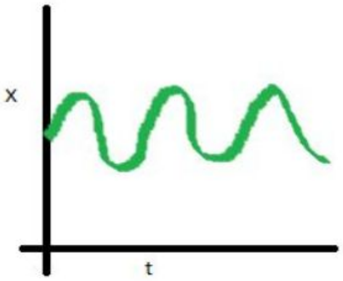


Stationary series

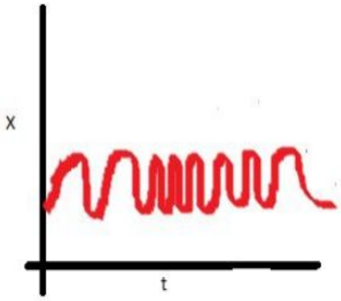


Non-Stationary series

3. Finally, the covariance of the i th term and the (i + m) th term should not be a function of time. In the following graph, you will notice the spread becomes closer as the time increases. Hence, the covariance is not constant with time for the 'red series'.



Stationary series



Non-Stationary series

Now a very natural question is that, why do we need a stationary time series? The reason being, when we run a linear regression the assumption is that all of the observations are all independent of each other. In a time series, however, we know that observations are time dependent. It turns out that a lot of nice results that hold for independent random variables (law of large numbers and central limit theorem to name a couple) hold for stationary random variables. So by making the data stationary, we can actually apply regression techniques to this time dependent variable.

There are two ways you can check the stationarity of a time series. The first is by looking at the data. By visualizing the data it should be easy to identify a changing mean or variation in the data. For a more accurate assessment there is the Dickey-Fuller test.

Dickey Fuller Test:

In statistics, the Dickey–Fuller test, tests the null hypothesis that a unit root is present in an autoregressive model. The alternative hypothesis is different depending on which version of the test is used, but is usually stationarity or trend-stationarity.

Hence our null hypothesis H0 is not stationary against our alternative hypothesis H1 which is data stationary .

The Dickey-Fuller test is testing if $\phi=1$ in this model of the data:

$y_t = \alpha + \beta t + \phi y_{t-1} + e_t$

which is written as: $\Delta y_t = y_t - y_{t-1} = \beta + \gamma y_{t-1} + e_t$

Where y_t is your data.It is written this way so we can do a linear regression of Δy_t against t and y_{t-1} and test if γ is different from 0. If $\gamma=0$, then we have a random walk process. If not and $-1<1+\gamma<1$, then we have a stationary process.

We applied the test to check whether our hypothesis is true or not, we look for a p-value in the test, and if the p-value is less than a specific significant level often 0.05 or 0.01, we reject our null hypothesis and thus making our time series stationary.

So now we need to transform the data to make it more stationary. There are various transformations you can do to stationarize the data.

1.Deflation: Merely applies a constant discount factor to the previous data. It usually helps in stabilizing variance.

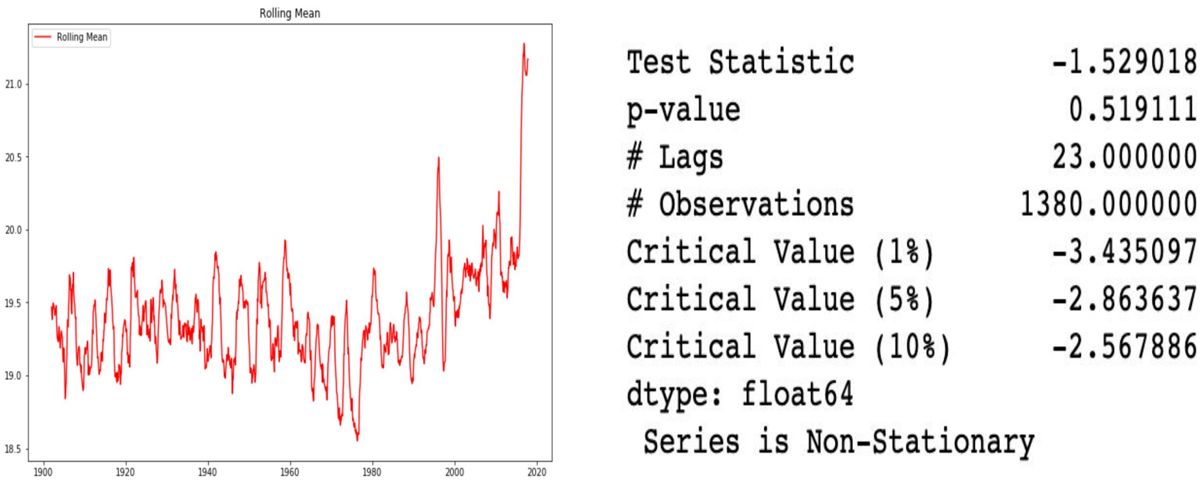
2.Logarithmic: Converts multiplicative patterns to additive patterns and/or linearized exponential growth. Converts absolute change to percentage changes. Often stabilizes the variance of data with compound growth, regardless of whether deflation is also used.

3.First Difference: Converts “levels” to “changes”. (X_t -X_{t-1})

4.Seasonal Difference: Convert “levels” to “seasonal changes”.(X_t -X_{t-s})

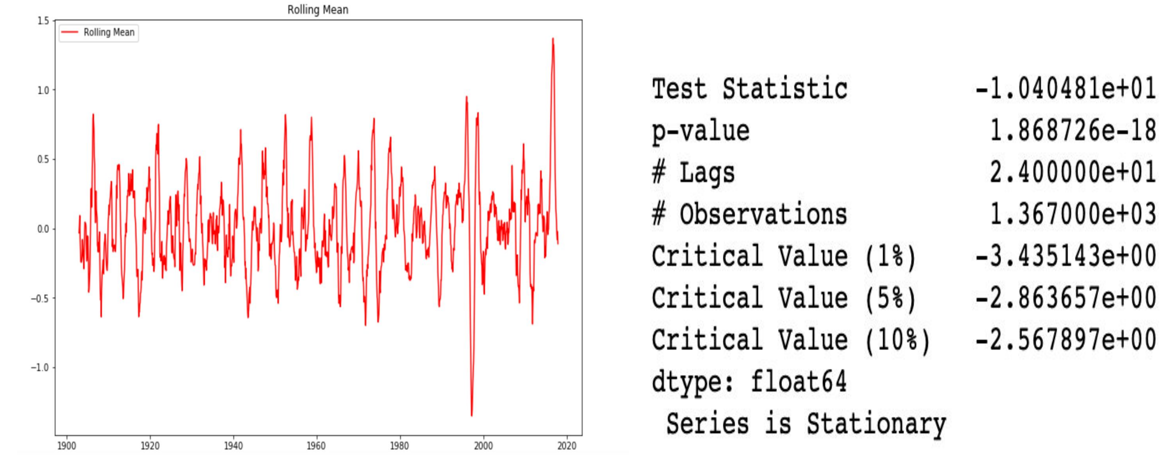
5.Seasonal Adjustment: Remove a constant seasonal pattern from a series (either multiplicative or additive).

Now we will apply various transformations recursively until we obtain a stationary time series according to the Dickey-Fuller test. In this project, we are just going to use ***Seasonal Differencing***.



```
Test Statistic              -1.529018
p-value                      0.519111
# Lags                      23.000000
# Observations            1380.000000
Critical Value (1%)         -3.435097
Critical Value (5%)         -2.863637
Critical Value (10%)        -2.567886
dtype: float64
 Series is Non-Stationary
```

Figure 3. D-F test on original data



```
Test Statistic         -1.040481e+01
p-value                 1.868726e-18
# Lags                  2.400000e+01
# Observations          1.367000e+03
Critical Value (1%)    -3.435143e+00
Critical Value (5%)    -2.863657e+00
Critical Value (10%)   -2.567897e+00
dtype: float64
 Series is Stationary
```

Figure 4. D-F test on seasonal differenced data

After performing a D-F test on original data, notice that the p-value>0.05, so we can’t reject our null hypothesis i.e. our original data is not stationary. We did seasonal differencing on original data of order 12. Then we performed a D-F test on seasonal differenced data and can notice the p-value<<0.05, so we can reject our null hypothesis that means our seasonal differenced data is stationary. So we use seasonal differenced data for the rest of the analysis.

3 : Plot the ACF and PACF charts and find the optimal parameters :

- ➔ **Autocorrelation Function (ACF).** The plot summarizes the correlation of an observation with lag values. The x-axis shows the lag and the y-axis shows the correlation coefficient between -1 and 1 for negative and positive correlation.
- ➔ **Partial Autocorrelation Function (PACF).** The plot summarizes the correlations for an observation with lag values that is not accounted for by prior lagged observations

Some useful patterns you may observe on these plots are:

● The model is AR if the ACF trails off after a lag and has a hard cut-off in the PACF after the lag. This lag value is taken as the value for p.

● The model is MA if the PACF trails off after a lag and has a hard cut-off in the ACF after the lag. This lag value is taken as the value for q.

● The model is ARIMA (mix of AR and MA) if both the ACF and PACF tail off. These are the trickiest because the order/lag value will not be particularly obvious. We might have to just guess one or two terms and then see what happens when you estimate the model.

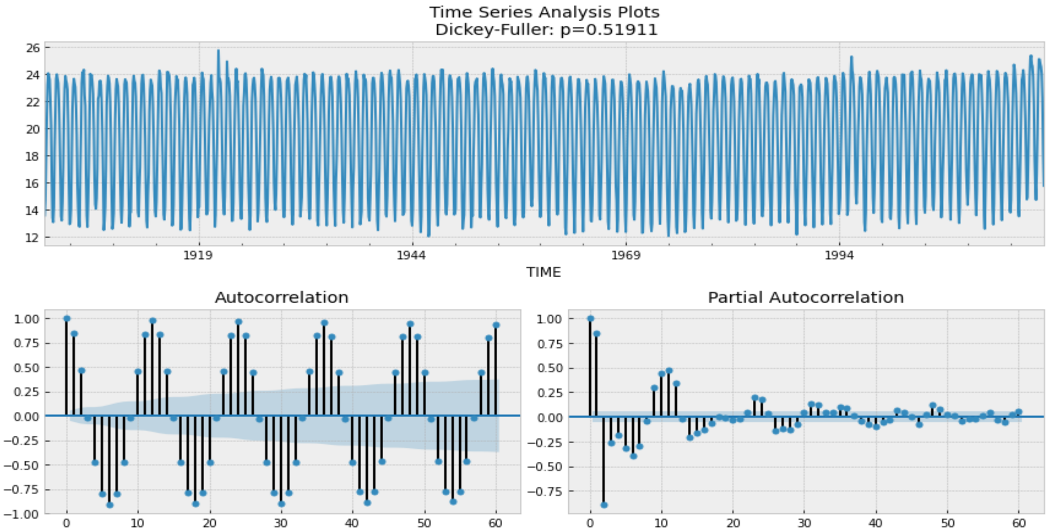


Figure 5. ACF and PACF of normal data

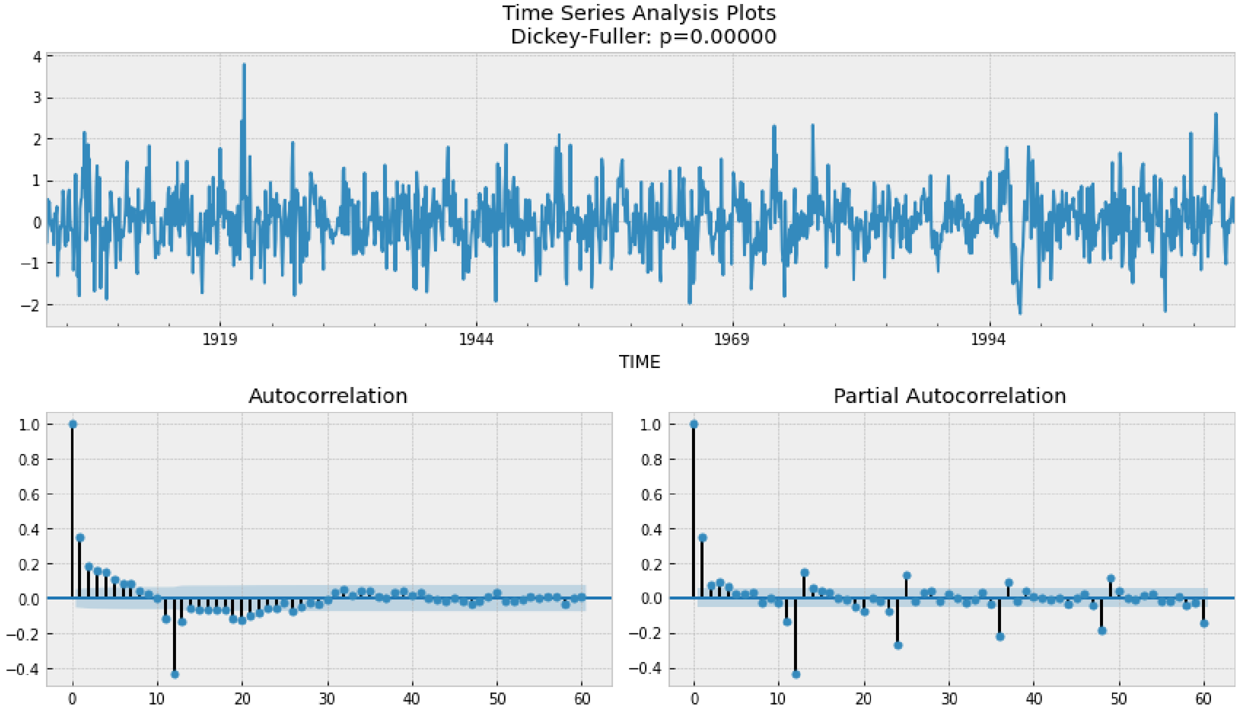


Figure 6. ACF and PACF of seasonal differenced data

Observations:

1. **Seasonality:** In the first plot notice that ACF and PACF do not tail off, instead it has values close to one over many lags. This supports the earlier assertion of seasonality in the data and differencing will be needed. So we try seasonal differencing with lag 12 (as this is monthly data). In the second plot of ACF and PACF of the differenced data do tail off, indicating that our data doesn't have significant seasonality, and is stationary.
2. **Seasonal terms:** From Figure, a spike at 12 in the ACF is significant but no other is significant at lags multiple of 12, the pacf shows an exponential decay in the seasonal lags; that is 12, 24, 36 etc. Thus, the seasonal part of the model has a moving average term of order 1 and an autoregressive term of 1. Hence,

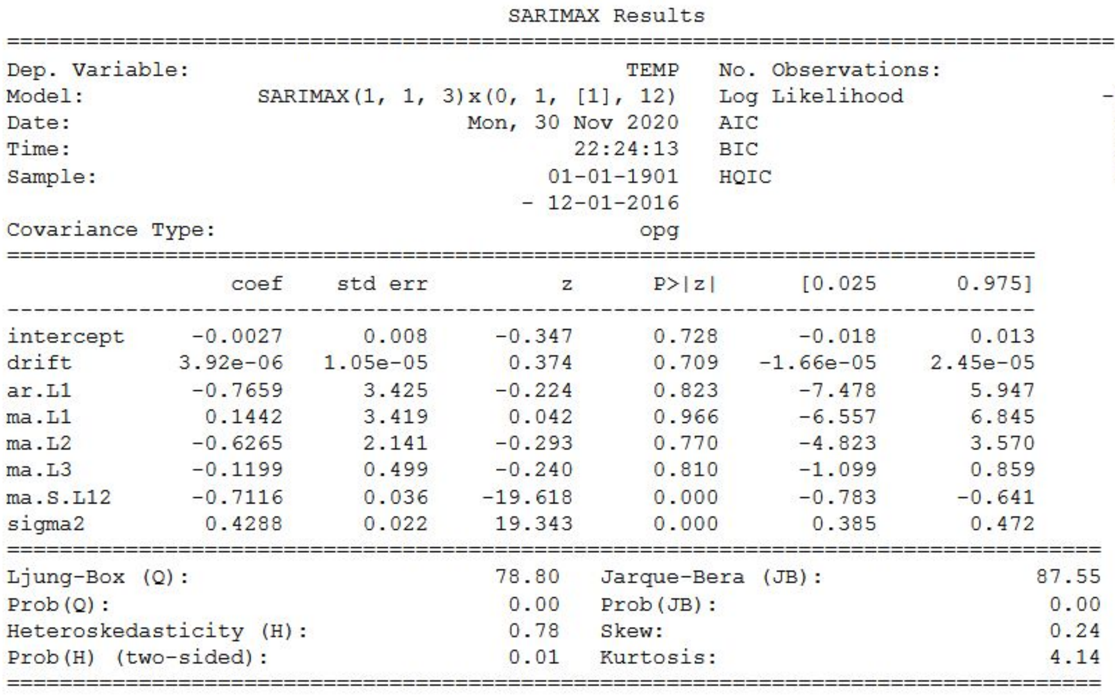
   P(seasonal AR order) = 1, Q(seasonal MA order) = 1, D(seasonal differencing) = 1, S(time span of repeating seasonal pattern) = 12

3. **Non-seasonal terms:** For the non-seasonal part, the ACF tails off after lag 2 and the PACF cuts off after lag 5. Therefore, the non-seasonal part has an autoregressive term of 2 and a moving average term of 5. Hence,

   p(non-seasonal AR order) = 2, q(non-seasonal MA order) = 5, d(non-seasonal differencing) = 0

4. **Model terms:** With general form of SARIMA (p,d,q) x (P,D,Q,S), first we have tried a set of parameters with p ranging in (0,4), q ranging in (0,6), P and Q ranging in (0, 2) with d = (0,1), D = 1 and s = 12.

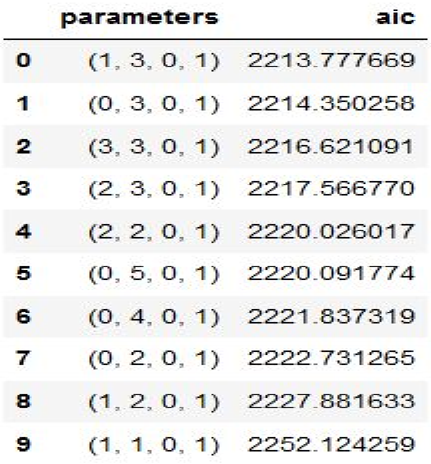
4 : Build Model :



```
                                      SARIMAX Results
============================================================================================
Dep. Variable:                                 TEMP   No. Observations:                 1392
Model:             SARIMAX(1, 1, 3)x(0, 1, [1], 12)   Log Likelihood               -1098.889
Date:                              Mon, 30 Nov 2020   AIC                           2213.778
Time:                                      22:24:13   BIC                           2255.611
Sample:                                  01-01-1901   HQIC                          2229.428
                                       - 12-01-2016
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     -0.0027      0.008     -0.347      0.728      -0.018       0.013
drift        3.92e-06   1.05e-05      0.374      0.709   -1.66e-05    2.45e-05
ar.L1         -0.7659      3.425     -0.224      0.823      -7.478       5.947
ma.L1          0.1442      3.419      0.042      0.966      -6.557       6.845
ma.L2         -0.6265      2.141     -0.293      0.770      -4.823       3.570
ma.L3         -0.1199      0.499     -0.240      0.810      -1.099       0.859
ma.S.L12      -0.7116      0.036    -19.618      0.000      -0.783      -0.641
sigma2         0.4288      0.022     19.343      0.000       0.385       0.472
===================================================================================
Ljung-Box (Q):                       78.80   Jarque-Bera (JB):                87.55
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.78   Skew:                             0.24
Prob(H) (two-sided):                  0.01   Kurtosis:                         4.14
===================================================================================
```

Figure 7. SARIMAX Model Statistics

As our aim is to find the model with the lowest value of the selected information criterion, So in our SARIMAX model with a set of integers, we tried for different values of p,q and i and found that for the above shown values, the AIC and BIC(which were our selected information criterions) values came out to be small and hence it is the required model which came out to be same as we anticipated using ACF and PACF plots.



	parameters	aic
0	(1, 3, 0, 1)	2213.777669
1	(0, 3, 0, 1)	2214.350258
2	(3, 3, 0, 1)	2216.621091
3	(2, 3, 0, 1)	2217.566770
4	(2, 2, 0, 1)	2220.026017
5	(0, 5, 0, 1)	2220.091774
6	(0, 4, 0, 1)	2221.837319
7	(0, 2, 0, 1)	2222.731265
8	(1, 2, 0, 1)	2227.881633
9	(1, 1, 0, 1)	2252.124259

The above table also displays the values of parameters. These parameter values are (1,1,3)×(0,1,1,12).

Residual Analysis:

Standardized Residual Plot: In the standardized residual plot we can see that the residuals appear to be just fluctuating about x-axis without any kind of trend or pattern and hence showing the randomness of these residuals.

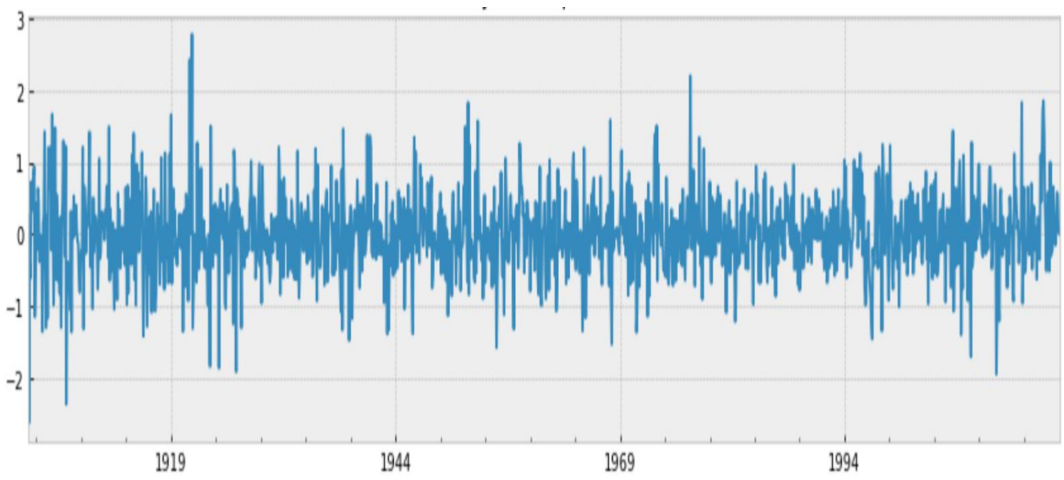


Fig 8. Standardized Residual plot

Histogram: Histogram of residuals is used to show whether variance is normally distributed or not, a symmetric well shaped histogram which is evenly distributed around 0, indicates that the normality assumption about the residuals is likely to be true. If the histogram indicates that random error is not normally distributed then it suggests that the model underlying assumptions may have been violated.

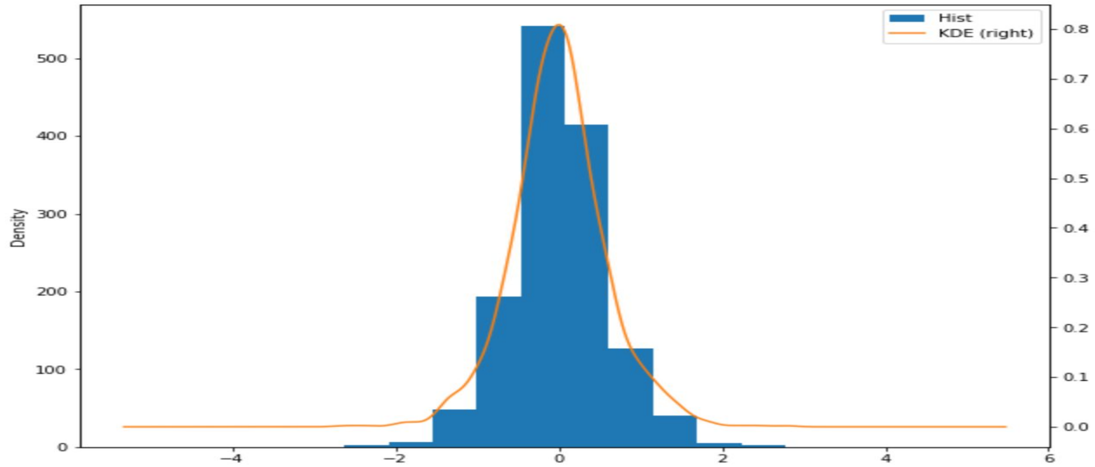


Fig 9. Histogram

Q-Q Plot: Normal Q-Q , or quantile-quantile plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. We can observe that for our Q-Q plots both the quantiles are overlapping.

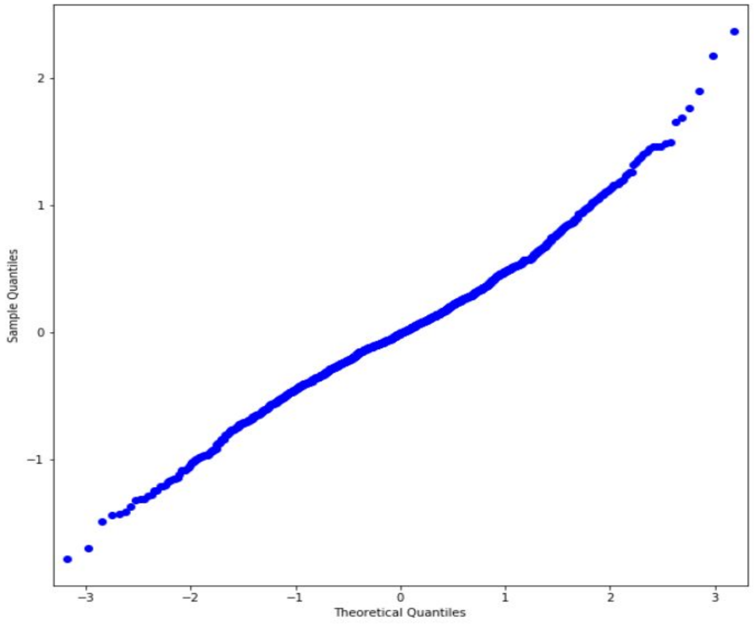


Fig 10. Q-Q plot

ACF Plot: A correlogram (also called Auto Correlation Function ACF Plot or Autocorrelation plot) is a visual way to show serial correlation in data that changes over time (i.e. time series data). Serial correlation (also called autocorrelation) is where an error at one point in time travels to a subsequent point in time. Correlograms can give you a good idea of whether or not pairs of data show autocorrelation. We can observe that our residual data follows a plot similar to WN sequence.

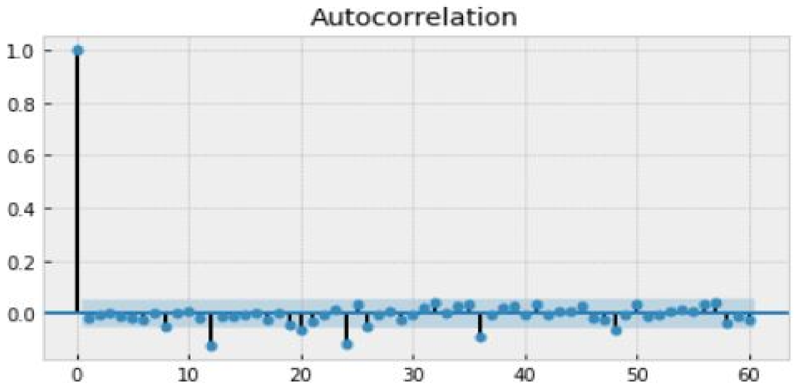


Fig 11. Correlogram

Hence based on these plots, we can conclude that there are no further trends that can be extracted from these residuals and that it follows a distribution similar to a normal WN process.

5 : Make Predictions :

Now that we have a model built, we want to use it to make forecasts. First we use the model to forecast for time periods that we already have data for, so we can understand how accurate the forecasts are.

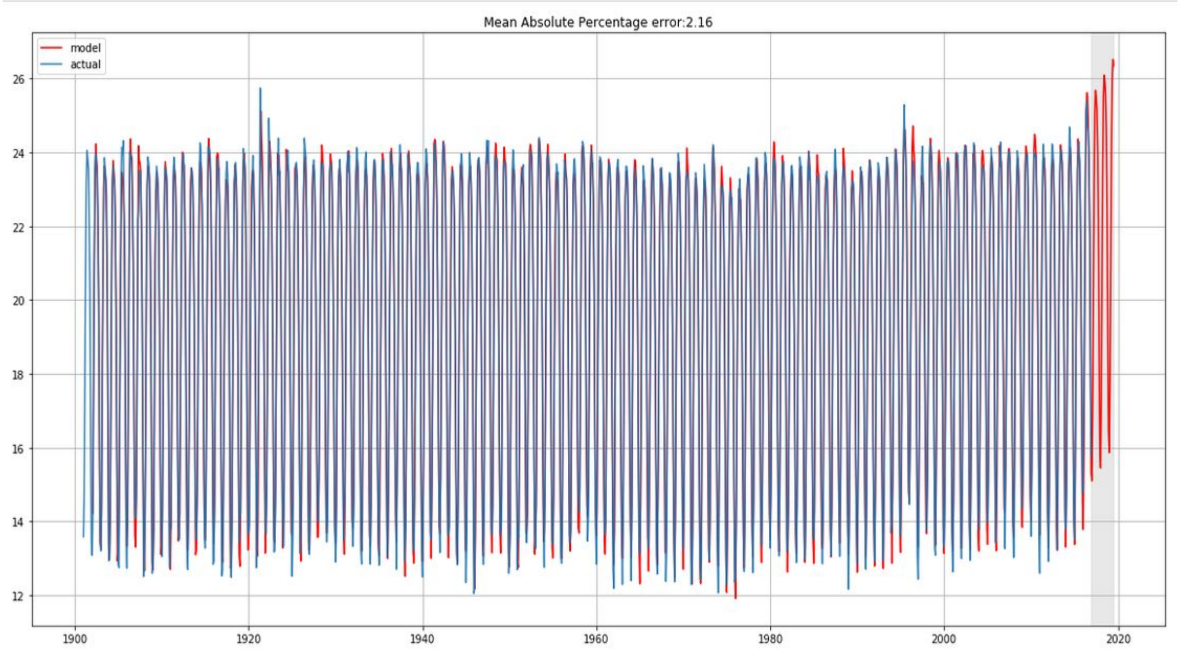


Figure 12. Model Fitting and Comparing

We have taken the dataset of year 1901 to 2016 as the train dataset and predicted the temperature of the next year i.e. 2017. Then we have calculated the prediction error of the model by comparing the predicted values with the actual 2017 temperature data.

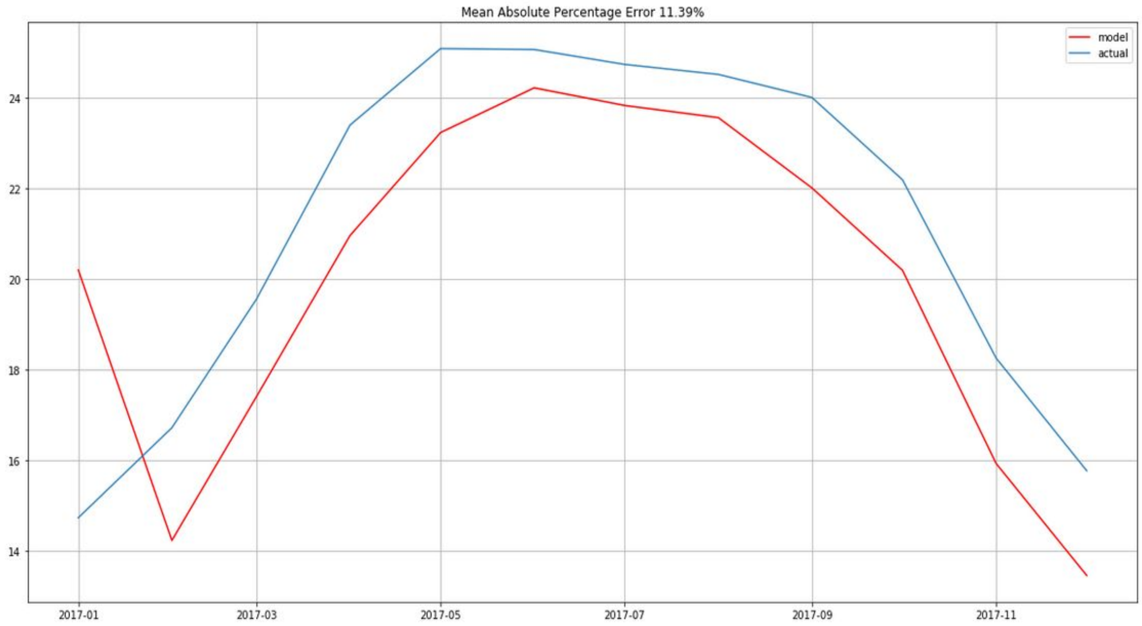


Figure 13. Model Fitting and Comparing

6 : Conclusion :

It is seen from the data analysis that the seasonal component is very persistent in the temperature data which has a period length equal to 12 and we can see a rising trend from the later years in the period of 1901 to 2017. We have presented a seasonal ARIMA model which has an absolute mean percentage error of 2.2 % on the training data and 11.4 % absolute mean percentage error on the unseen test data for one year ahead. This model can be used to get an estimate of the future mean monthly temperature of India.

7 : Acknowledgment :

We would like to thank **Prof. Amit Mitra**, our academic and project advisor, for his guidance throughout this project. The brainstorming ideas from each meeting, turned out to be a helpful part of this exploration. This project helped us in getting insight as to how Time Series forecasting can have significant real world applications.

8 : References :

- Class Notes of MTH 517A
- Wikipedia
- data.gov.in
- Time Series in Python — Part 2: Dealing with seasonal data [1]
- Time Series Forecasting with a SARIMA Model [2]