TIME SERIES DATA ANALYSIS ON AVERAGE TEMPERATURE IN KOLKATA FROM 1993-2012

A Project Submitted in Partial Fulfilment of the Requirements for the Degree of Bachelor of Science in Statistics

By

Debopriya Bose

Semester-6

Paper- DSE-B2

Registration No- 041-1211-0325-19

Roll No- 193041-11-0079

Under the supervision of

Dr. Soumita Modak

DEPARTMENT OF STSTISTICS

BASANTI DEVI COLLEGE

Estd. 1959

147B, Rash Behari Avenue, Kolkata – 700029, West Bengal, India

AFFILIATED TO THE UNIVERSITY OF CALCUTTA

Date: 14/07/2022

CONTENT

INTRODUCTION

Time Series

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus, it is a sequence of discrete-time data. Time series data is everywhere, since time is a constituent of everything that is observable. As our world gets increasingly instrumented, sensors and systems are constantly emitting a relentless stream of time series data. Such data has numerous applications across various industries.

Examples :-

- Electrical activity in the brain
- Rainfall measurements
- Stock prices
- Annual retail sales
- Monthly subscribers
- Heartbeats per minute

Uses of time series:-

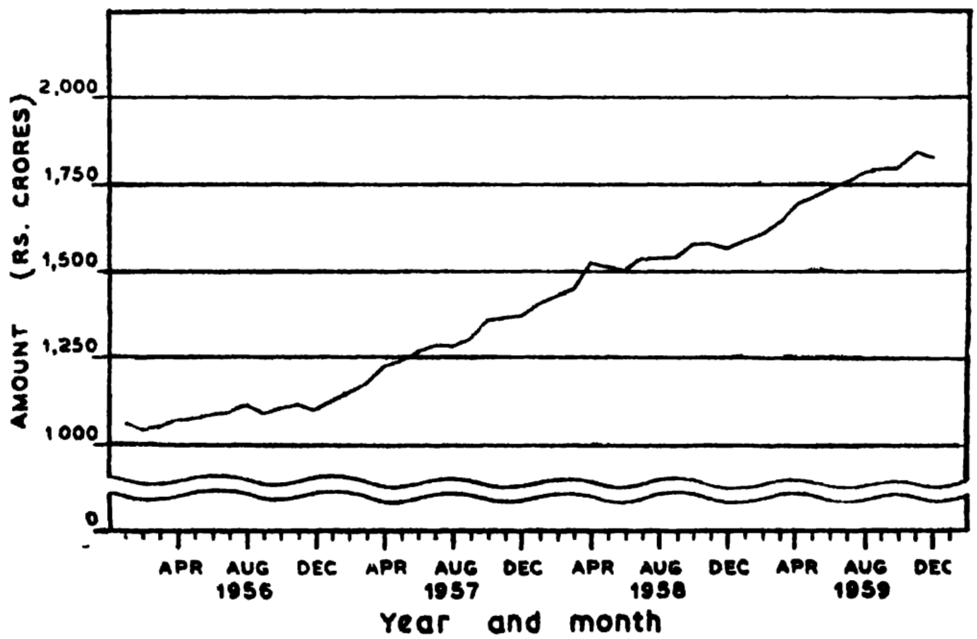
- Time series is used to predict future values based on previously observed values.
- Time series analysis is used to identify the fluctuation in economics and business.
- It helps in the evaluation of current achievements.
- Time series is used in pattern recognition, signal processing, weather forecasting and earthquake prediction.

Components of time series data:-

- **Trend**: By secular trend (or, simply, trend) of time series, we mean the smooth, regular, long-term movement of a series if observed long enough. Some series may exhibit an upward or a downward trend or may remain more or less at a constant level.

Example:

a. An aging population, which tends to have different spending and savings habits than a younger population
b. The expansion of a particular technology such as the internet
c. The clean-energy movement

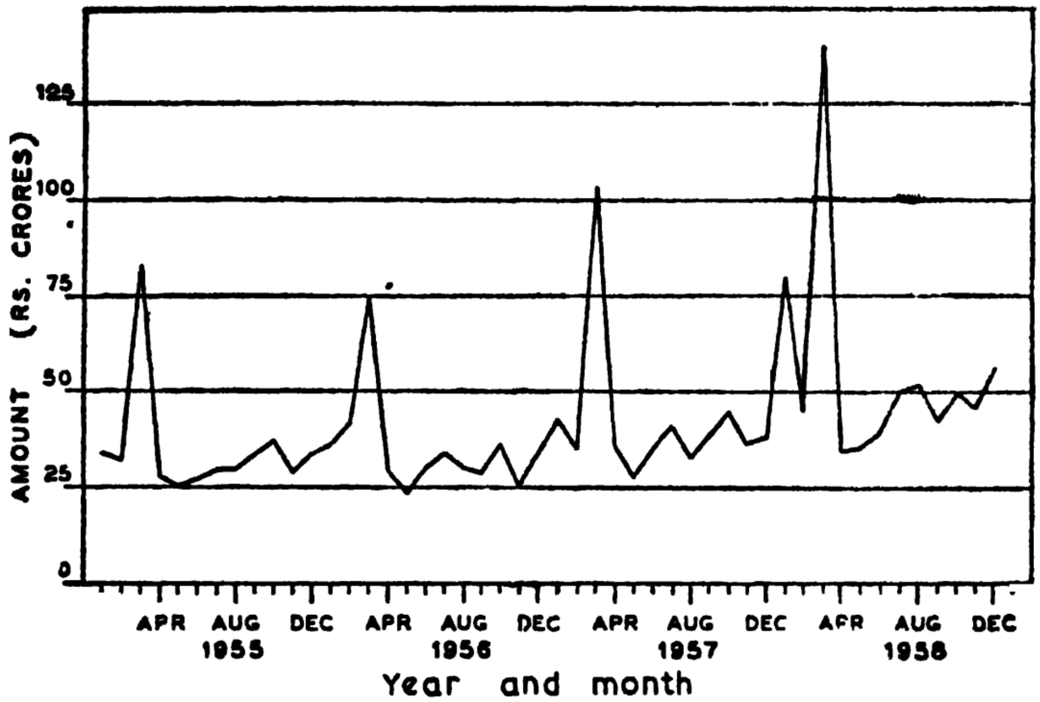


Graph showing deposit liabilities of scheduled banks in India

❖ **Seasonal fluctuation**: By seasonal fluctuations, we mean a periodic movement in a time series, where the period is no longer than one year. A periodic movement in a time series is one which recurs or repeats at regular interval of time (or periods). The factors which mainly cause this type of variation in economic time series are the climatic changes of the different seasons and the customs and habits which the people follow at different times.

Examples:

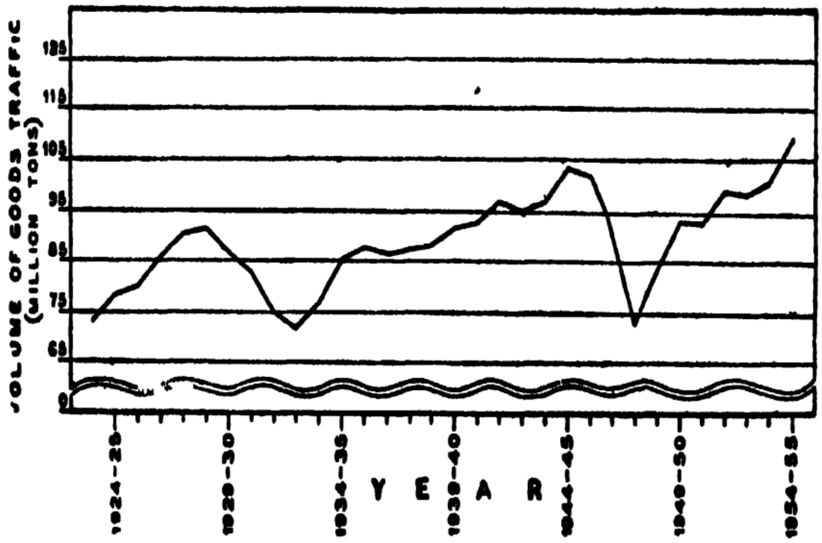
a. Occurrence of a festival in a particular month will increase the sale of certain consumer goods in that month.
b. Passenger traffic during the 24 hours of a day.
c. Issue of library books during the seven days of a week.



Graph showing revenue expenditure and defence drawings, Govt of India

❖ **Cyclical fluctuation**: By cyclical fluctuations, we mean the oscillatory movement in a time series, the period of oscillation being more than a year. One complete period is called a cycle. The cyclical fluctuations are not necessarily periodic, since the length of a cycle as also the intensity of fluctuations may change from cycle to another.

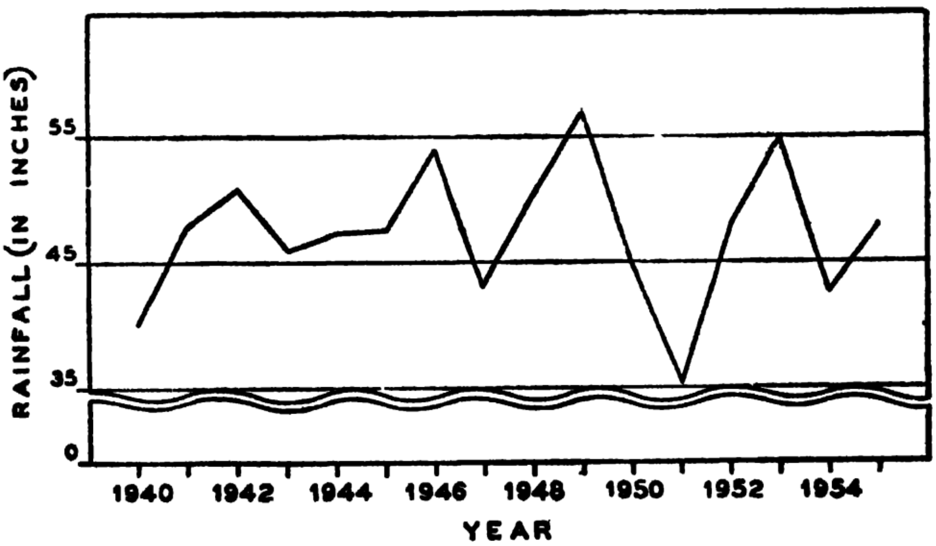
Example: Alternating periods of 'prosperity' (or 'boom') and 'depression' in business which follow one another in an irregular manner.



Graph showing volume of goods traffic carried by Indian Railways

❖ **Irregular components**: The irregular components (sometimes also known as the residual) is what remains after the seasonal and trend components of a time series have been estimated and removed. It results from short term fluctuations in the series which are neither systematic nor predictable. In a highly irregular series, these fluctuations can dominate movements, which will mask the trend and seasonality.

Examples: Wholly unaccountable events or unforeseen events such as wars, floods, strikes, etc.



Graph showing annual rainfall in Bihar

OBJECTIVE

Our time series data is completely based on the Average Temperature in Kolkata from 1992 to 2012. As we all know, climate change is one of the most concerning global threats at the moment, which basically refers to long-term shifts in temperatures and weather patterns. These shifts may be natural, but since the 1800s, human activities have been the main driver of climate change, primarily due to the burning of fossil fuels (like coal, oil, and gas), which produces heat-trapping gases.

Climate change threatens people with food and water scarcity, increased flooding, extreme heat, more disease, and economic loss. Human migration and conflict can be a result. The World Health Organization (WHO) calls climate change the greatest threat to global health in the 21st century.

Evidence of warming from air temperature measurements are reinforced with a wide range of other observations. There has been an increase in the frequency and intensity of heavy precipitation, melting of snow and land ice, and increased atmospheric humidity. Flora and fauna are also behaving in a manner consistent with warming; for instance, plants are flowering earlier in spring. Another key indicator is the cooling of the upper atmosphere, which demonstrates that greenhouse gases are trapping heat near the Earth's surface and preventing it from radiating into space.

Thus, the main objective of our study is to analyse our data and make useful predictions, which will help us to understand the extent to which global warming and climate change are affecting us and also predict that how the levels are going to increase or decrease in our near future.

TIME SEIES DATA

Data link: https://www.kaggle.com/code/leandrovrabelo/climate-change-forecast-sarima-model/data?select=GlobalLandTemperaturesByMajorCity.csv

Table – 1: Table showing Time series data on average temperature in Kolkata from 1993-2012

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1993	18.722	23.158	25.972	29.349	30.168	29.705	29.392	29.050	28.537	27.751	24.083	20.250
1994	19.662	21.457	27.809	29.782	31.489	29.794	29.221	29.178	29.139	27.617	23.762	19.577
1995	17.935	22.048	26.834	31.445	32.014	30.407	29.085	29.365	28.951	27.735	23.705	19.880
1996	19.283	22.374	28.349	30.415	31.965	29.567	29.385	28.631	29.695	27.139	23.594	19.395
1997	18.240	21.471	27.492	27.590	30.893	30.473	29.121	29.254	28.783	26.927	24.698	19.334
1998	17.639	22.461	25.213	29.418	31.205	31.690	29.514	29.546	29.296	28.591	25.269	20.410
1999	18.950	23.720	28.334	31.831	30.548	29.971	28.929	28.944	28.544	27.764	24.071	20.749
2000	19.196	21.275	26.881	30.165	30.245	29.861	29.320	29.694	28.846	28.017	24.668	19.728
2001	18.220	22.654	26.858	30.192	30.142	28.899	28.989	29.773	29.600	27.925	25.146	19.784
2002	19.676	22.625	27.406	29.798	30.845	30.001	29.972	29.083	28.866	27.616	24.046	20.513
2003	17.813	23.065	26.173	30.571	31.340	30.086	29.425	29.629	29.223	27.536	23.920	19.955
2004	18.661	22.416	28.452	29.587	32.045	30.153	29.123	29.399	28.883	26.720	23.552	20.621
2005	19.341	23.576	27.695	30.152	31.062	31.595	28.833	29.485	29.398	26.943	22.916	19.754
2006	19.269	24.980	27.602	30.323	30.762	30.101	29.523	29.163	29.391	28.040	24.175	20.506
2007	18.959	22.060	26.415	30.511	31.381	30.287	29.042	29.887	29.111	27.527	24.253	19.582
2008	19.065	20.805	27.880	30.538	31.134	29.111	28.852	29.284	29.087	27.422	24.227	20.863
2009	20.431	23.564	27.711	31.778	30.585	31.665	29.575	29.533	29.626	27.273	24.281	19.817
2010	17.503	22.959	29.451	32.318	31.308	31.075	29.979	30.226	29.536	28.071	25.437	19.603
2011	17.864	22.566	27.394	29.276	30.395	29.873	29.473	28.976	29.092	28.097	24.110	19.767
2012	18.815	22.408	27.878	30.299	32.232	31.959	29.680	29.762	29.653	27.427	23.487	19.621

The above table (Table – 1) consists of 20 years of data representing the average temperature in India from the year 1993-2012. We are going to analyse the given time series data and make useful predictions from it.

DATA ANALYSIS

Plotting of time series data:

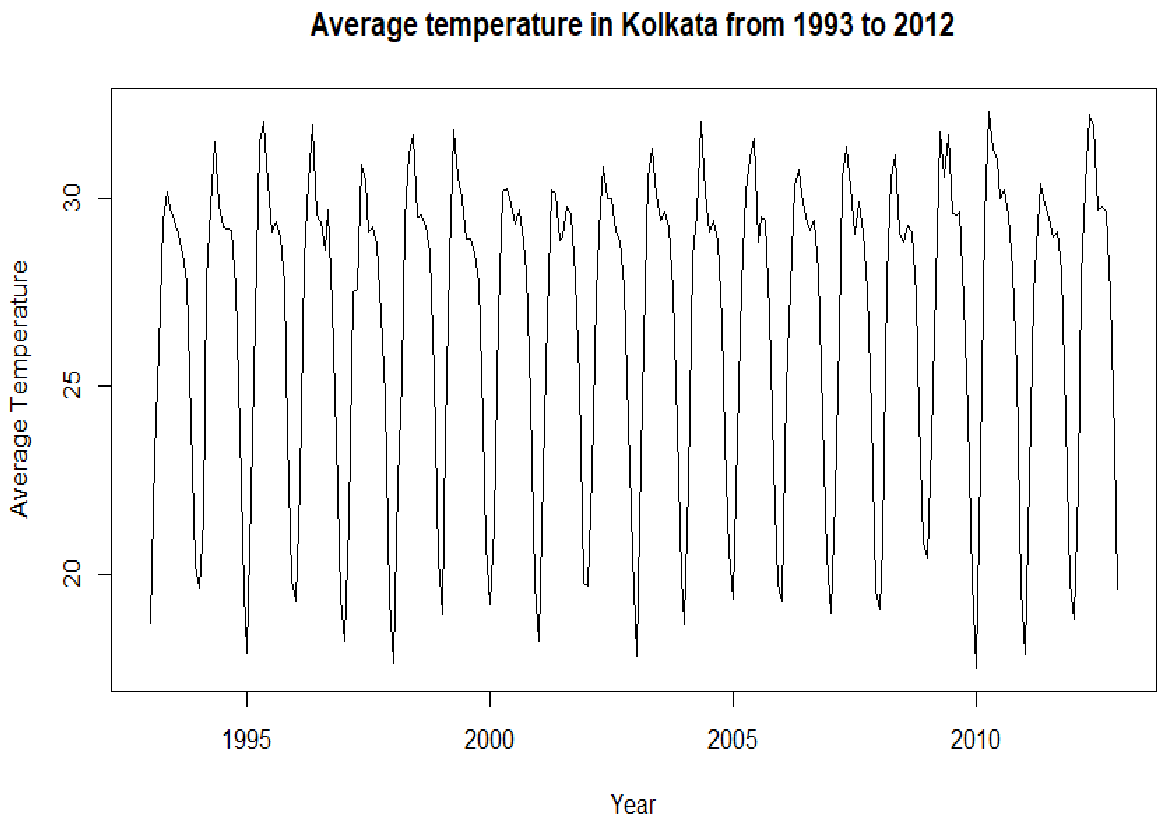


Diagram: 1

Interpretation:

The above diagram (Diagram: 1) is the graphical representation of the ‘average temperature in Kolkata’ from 1993-2012. Here, the x-axis represents ‘Year’ and the y-axis represents ‘Average Temperature (in ˚C)’. This time series plot demonstrates that the average temperature has seasonality pattern and it is not following any specific trend.

Decomposition of the time series data:

Time series data demonstrates a variety of patterns, and it is often helpful to split a time series into several components, each representing an underlying pattern category. Time series decomposition involves thinking of a series as a combination of trend, seasonality, and noise components.

Decomposition provides a useful abstract model for thinking about time series generally and for better understanding problems during time series analysis and forecasting. Each of these components are something we may need to think about and address during data preparation, model selection, and model tuning. We may address it explicitly in terms of modelling the trend and subtracting it from our data, or implicitly by providing enough history for an algorithm to model a trend if it may exist.

We may or may not be able to cleanly or perfectly break down our specific time series as an additive or multiplicative model. Real-world problems are messy and noisy. There may be additive and multiplicative components. There may be an increasing trend followed by a decreasing trend. There may be non-repeating cycles mixed in with the repeating seasonality components.

Nevertheless, these abstract models provide a simple framework that we can use to analyse our data and explore ways to think about and forecast our problem.

Now, we are going to decompose the original data into trend, seasonal and random part**.**

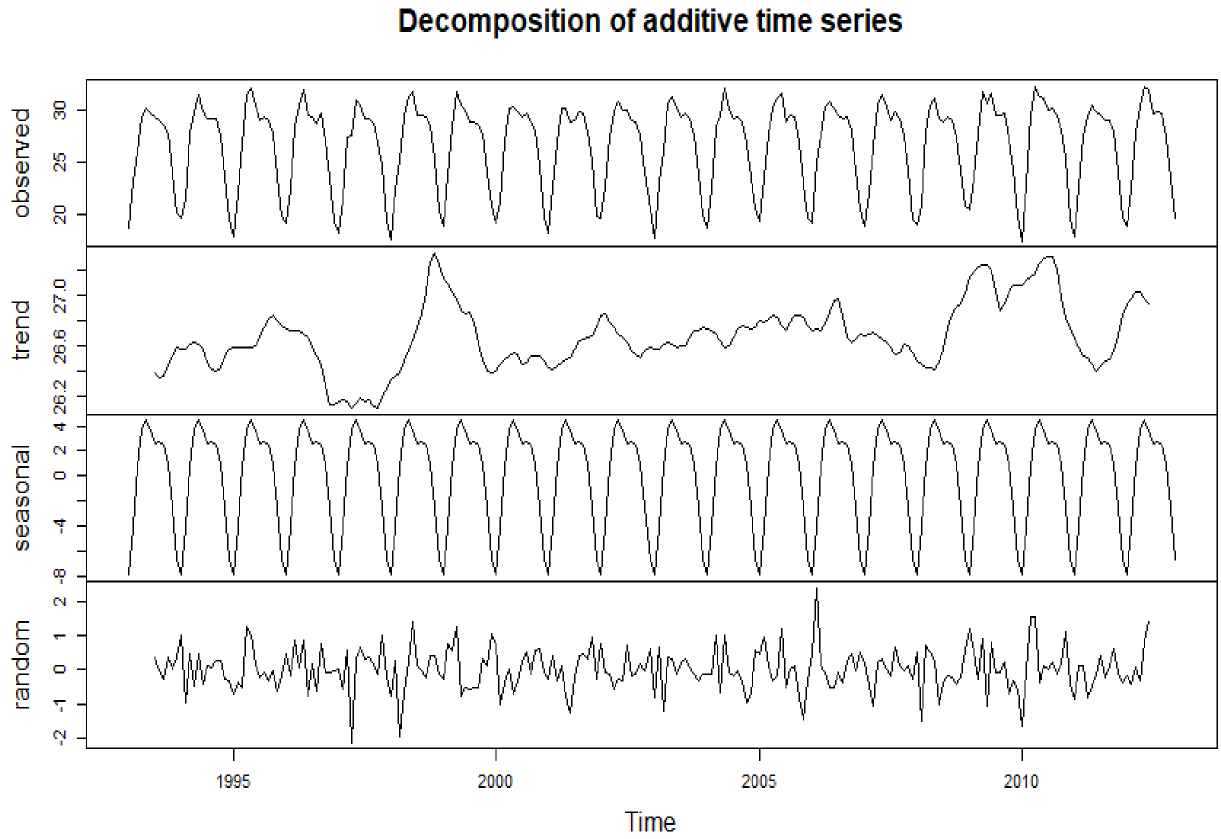


Diagram: 2

Interpretation:

The above diagram (Diagram: 2) demonstrates the decomposition of our time series data into trend, seasonality and random component. From this diagram, it is evident that our time series data has seasonality pattern but it is not following any specific trend.

Extract the random part:

Now, we are going to extract the random component.

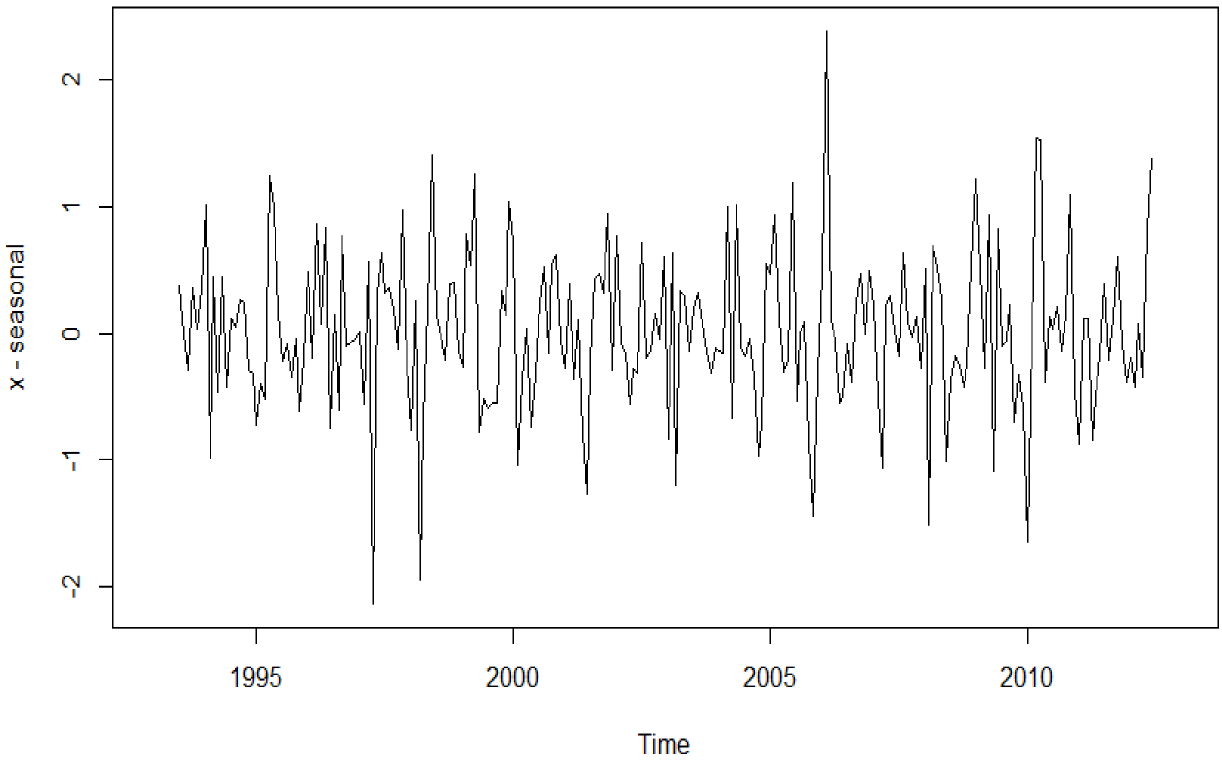


Diagram: 3

Interpretation:

From the above diagram (Diagram: 3), we can observe that the random component looks like white noise. Thus, it is stationary and we can confidently fit stochastic models on it.

Stationarity

A stationary time series is one whose properties do not depend on the time at which the series is observed. Thus, time series with trends, or with seasonality, are not stationary, the trend and seasonality will affect the value of the time series at different times. On the other hand, a white noise series is stationary as it does not matter when we observe it, it should look much the same at any point in time.

Some cases can be confusing like a time series with cyclic behaviour (but with no trend or seasonality) is stationary. This is because the cycles are not of a fixed length, so before we observe the series, we cannot be sure where the peaks and troughs of the cycles will be.

In general, a stationary time series will have no predictable patterns in the long-term. Time plots will show the series to be roughly horizontal (although some cyclic behaviour is possible), with constant variance.

The random part is stationary or not can be checked using the Augmented Dickey Fuller Test.

Augmented Dickey-Fuller Test (ADF)

In statistics and econometrics, an augmented Dickey–Fuller test (ADF) tests the null hypothesis that a unit root is present in a time series sample. The alternative hypothesis is different depending on which version of the test is used, but is usually stationarity or trend-stationarity. It is an augmented version of the Dickey–Fuller test for a larger and more complicated set of time series models

The augmented Dickey–Fuller (ADF) statistic, used in the test, is a negative number. The more negative it is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence.

Testing procedure:

The testing procedure for the ADF test is the same as for the Dickey–Fuller test but it is applied to the model

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t ,$$

where α is a constant, β is the coefficient on a time trend and p is the lag order of the autoregressive process. Imposing the constraints α=0 and β=0 corresponds to modelling a random walk and using the constraint β=0 corresponds to modelling a random walk with a drift. Consequently, there are three main versions of the test, analogous to the ones discussed on Dickey-Fuller test.

By including lags of the order p, the ADF formulation allows for higher-order autoregressive processes. This means that the lag length p has to be determined when applying the test. One possible approach is to test down from high orders and examine the t-values on coefficients. An alternative approach is to examine information criteria such as the Akaike information criterion, Bayesian information criterion or the Hannan–Quinn information criterion.

The unit root test is then carried out under the null hypothesis γ=0 against the alternative hypothesis of γ<0. Once a value for the test statistic

$$DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$$

is computed it can be compared to the relevant critical value for the Dickey–Fuller test. As this test is asymmetrical, we are only concerned with negative values of our test statistic DF_τ . If the calculated test statistic is less (more negative) than the critical value, then the null hypothesis of γ=0 is rejected and no unit root is present.

Intuition:

The intuition behind the test is that if the series is characterised by a unit root process, then the lagged level of the series (y_{t-1}) will provide no relevant information in predicting the change in y_t besides the one obtained in the lagged changes (Δy_{t-k}). In this case the γ=0 and null hypothesis is not rejected. In contrast, when the process has no unit root, it is stationary and hence exhibits reversion to the mean so the lagged level will provide relevant information in predicting the change of the series and the null of a unit root will be rejected.

Check stationary or not:

Now, we are going to check if our time series is stationary or not using Augmented Dickey-Fuller test. The outcome of ADF test for our time series is as follows:

Augmented Dickey-Fuller Test

data: Kolkata

Dickey-Fuller = -20.429, Lag order = 6, p-value = 0.01

alternative hypothesis: stationary

Interpretation:

We know, if p-value is less than 0.05, then the series is said to be stationary. Here, our p-value is 0.01, that is less than 0.05. Therefore, the series is stationary.

We know that our data has a seasonality pattern. So, to explore more about our rainfall data seasonality; seasonal plot, seasonal-subseries plot, and seasonal box plot will provide a much more insightful explanation about our data.

Seasonal plot:

Seasonal plots are a graphical tool to visualize and detect seasonality in a time series. Seasonal plots involve the extraction of the seasons from a time series into a subseries. Based on a selected periodicity, it is an alternative plot that emphasizes the seasonal patterns are where the data for each season are collected together in separate mini time plots.

Seasonal plots enable the underlying seasonal pattern to be seen clearly, and also shows the changes in seasonality over time. Especially, it allows to detect changes between different seasons, changes within a particular season over time.

However, this plot is only useful if the period of the seasonality is already known. In many cases, this will in fact be known. For example, monthly data typically has a period of 12. If the period is not known, an autocorrelation plot or spectral plot can be used to determine it. If there is a large number of observations, then a box plot may be preferable.

Seasonal sub-series plots are formed by

- Vertical axis: response variable
- Horizontal axis: time of year; for example, with monthly data, all the January values are plotted (in chronological order), then all the February values, and so on.

The horizontal line displays the mean value for each month over the time series.

The analyst must specify the length of the seasonal pattern before generating this plot. In most cases, the analyst will know this from the context of the problem and data collection.

It is important to know when analysing a time series if there is a significant seasonality effect. The seasonal subseries plot is an excellent tool for determining if there is a seasonal pattern. Practically, seasonal subseries plots are often inspected as a preliminary screening tool. They allow visual inferences to be drawn from data prior to modelling and forecasting.

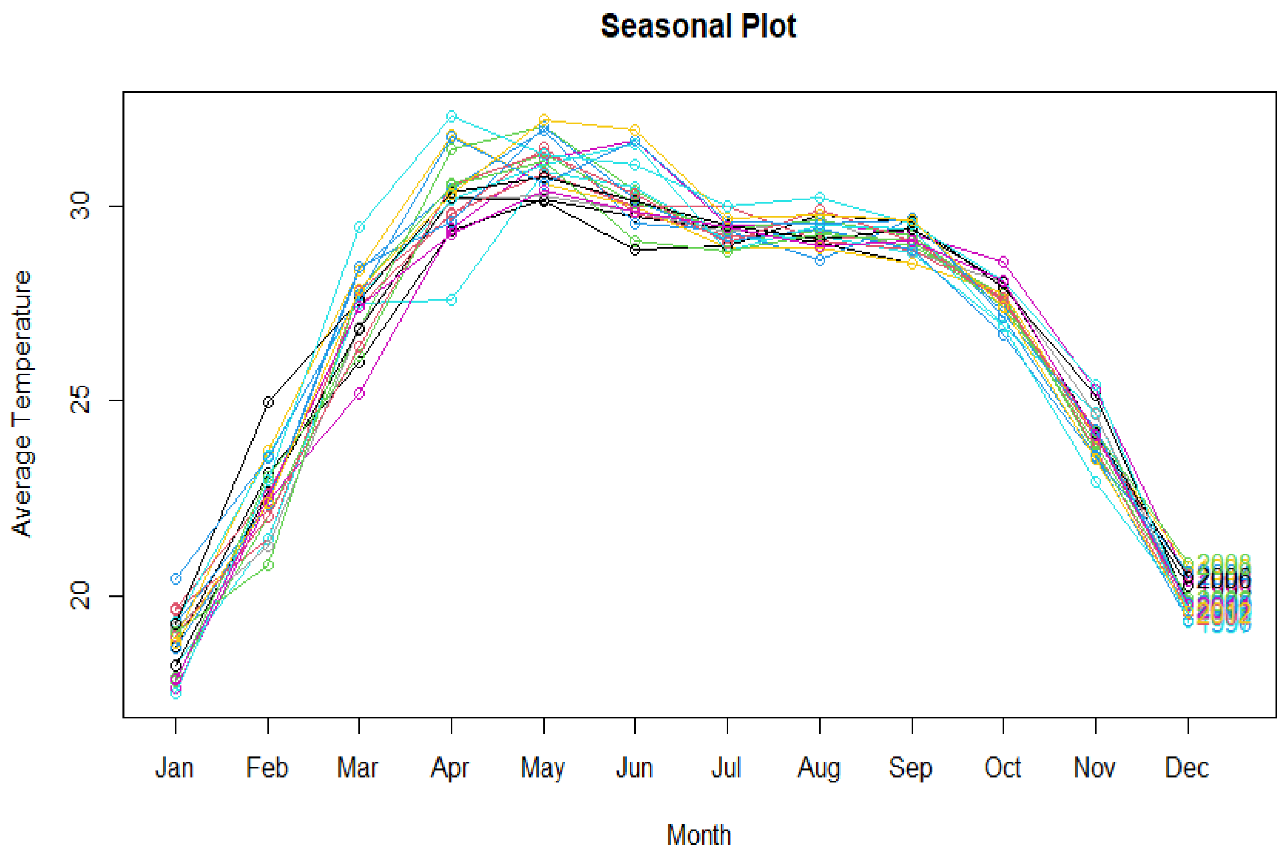


Diagram: 4

Interpretation:

From the above diagram (Diagram: 4), it is quite evident that this time series data has a seasonality pattern, which is occurring each year.

Seasonal Box plot:

Now, we are going to use seasonal box plot to get a better representation and understanding of our data pattern.

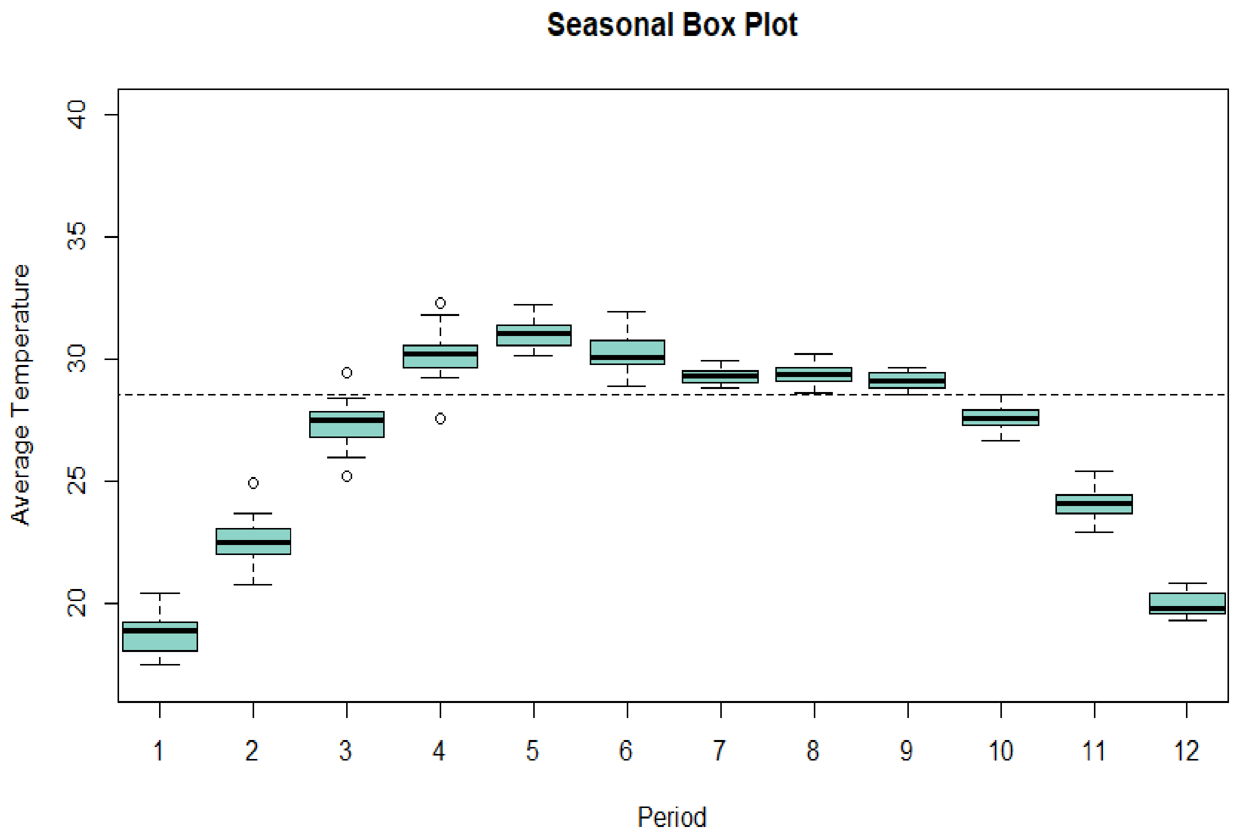


Diagram: 5

Interpretation:

From the above diagram (Diagram: 5), the horizontal line indicates the average temperature value, means grouped by month. With the help of this information, we can clearly observe that the average temperature (in ˚C) gradually starts to increase inn the month of April and it reaches its peak value in the month of May and its lowest value can be observed in the months of December and January.

STOCHASTIC PROCESS

A stochastic process, also known as a random process, is a collection of random variables that are indexed by some mathematical set. Each probability and random process are uniquely associated with an element in the set. Stochastic Process meaning is one that has a system for which there are observations at certain times, and that the outcome, that is, the observed value at each time is a random variable. Each random variable in the collection of the values is taken from the same mathematical space, known as the state space. This state-space could be the integers, the real line, or η-dimensional Euclidean space, for example. A stochastic process's increment is the amount that a stochastic process changes between two index values, which are frequently interpreted as two points in time. Because of its randomness, a stochastic process can have many outcomes, and a single outcome of a stochastic process is known as, among other things, a sample function or realization.

A stochastic process can be classified in a variety of ways, such as by its state space, index set, or the dependence among random variables and stochastic processes are classified in a single way, the cardinality of the index set and the state space.

When expressed in terms of time, a stochastic process is said to be in discrete-time if its index set contains a finite or countable number of elements, such as a finite set of numbers, the set of integers, or the natural numbers. Time is said to be continuous if the index set is some interval of the real line. Discrete-time stochastic processes and continuous-time stochastic processes are the two types of stochastic processes. The continuous-time stochastic processes require more advanced mathematical techniques and knowledge, particularly because the index set is uncountable, discrete-time stochastic processes are considered easier to study. If the index set consists of integers or a subset of them, the stochastic process is also known as a random sequence.

Auto-correlation Function (ACF):

The autocorrelation function (ACF) defines how data points in a time series are related, on average, to the preceding data points.

Autocorrelation, sometimes known as serial correlation in the discrete time case, is the correlation of a signal with a delayed copy of itself as a function of delay. Informally, it is the similarity between observations as a function of the time lag between them. The analysis of autocorrelation is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal obscured by noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies.

Uses of Autocorrelation function

- It helps to uncover the hidden patterns in our data and help us to select the suitable forecasting methods.
- It helps us to identify seasonality in our time series data.
- It helps us to identify the MA(q) value, which is very much essential for selecting appropriate ARIMA model.

ACF plot:

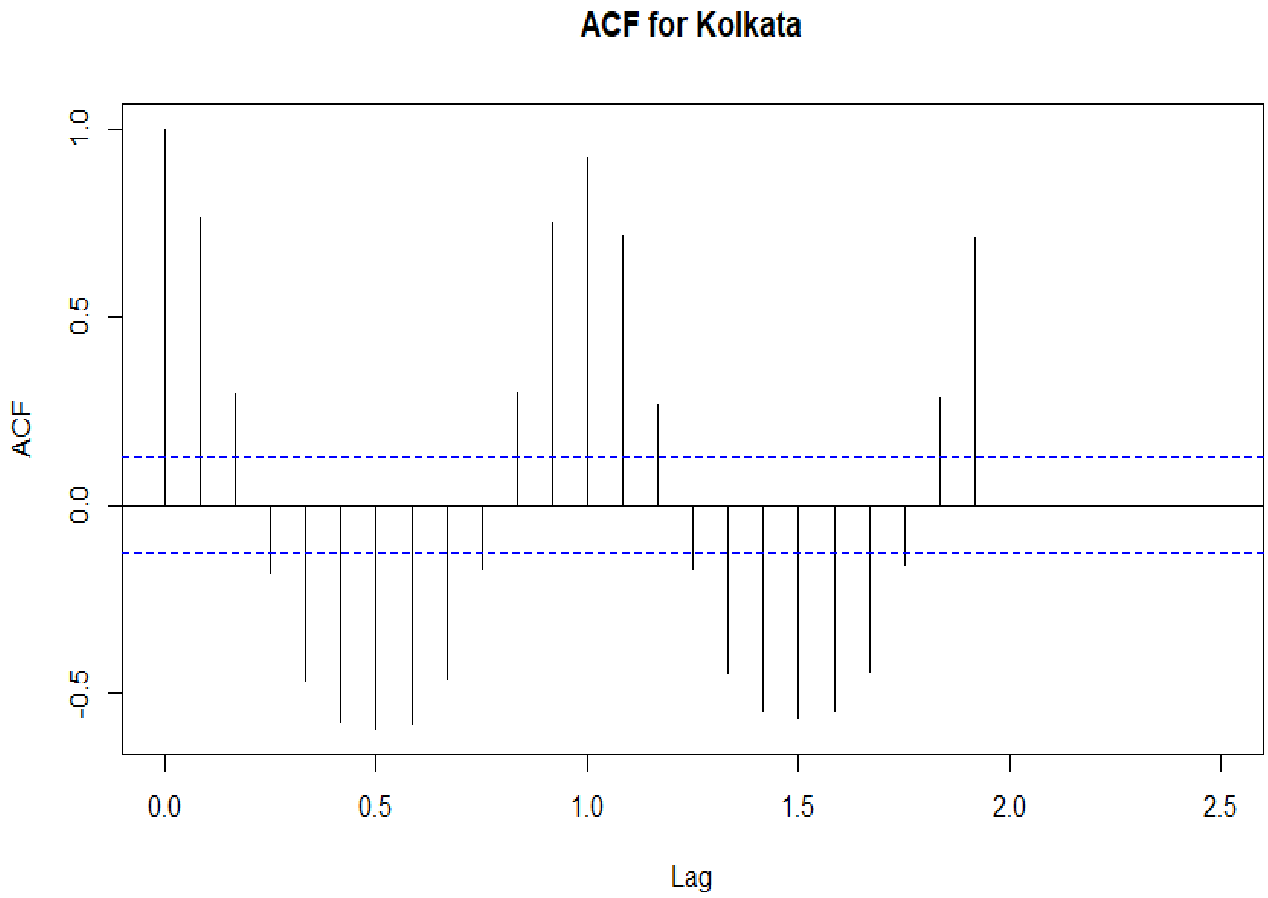


Diagram: 6

Interpretation:

From the above diagram (Diagram: 6), we can clearly observe the ACF plot, which enables us to identify the MA parameter q. In the ACF plot, there is a significant spike at lag 2.

The dashed blue lines indicate the 95% confidence interval, and for the correlations are significantly different from zero.

Partial Auto-correlation Function (PACF):

In time series analysis, the partial autocorrelation function (PACF) gives the partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags. It contrasts with the autocorrelation function, which does not control for other lags.

This function plays an important role in data analysis aimed at identifying the extent of the lag in an autoregressive model. The use of this function was introduced as part of the Box–Jenkins approach to time series modelling, whereby plotting the partial autocorrelative functions one could determine the appropriate lags p in an AR (p) model or in an extended ARIMA (p,d,q) model.

There are algorithms for estimating the partial autocorrelation based on the sample autocorrelations (Box, Jenkins, and Reinsel 2008 and Brockwell and Davis, 2009). These algorithms derive from the exact theoretical relation between the partial autocorrelation function and the autocorrelation function.

Partial autocorrelation plots are a commonly used tool for identifying the order of an autoregressive model. The partial autocorrelation of an AR(p) process is zero at lag p + 1 and greater. If the sample autocorrelation plot indicates that an AR model may be appropriate, then the sample partial autocorrelation plot is examined to help identify the order. One looks for the point on the plot where the partial autocorrelations for all higher lags are essentially zero. Placing on the plot an indication of the sampling uncertainty of the sample PACF is helpful for this purpose: this is usually constructed on the basis that the true value of the PACF, at any given positive lag, is zero.

PACF plot:

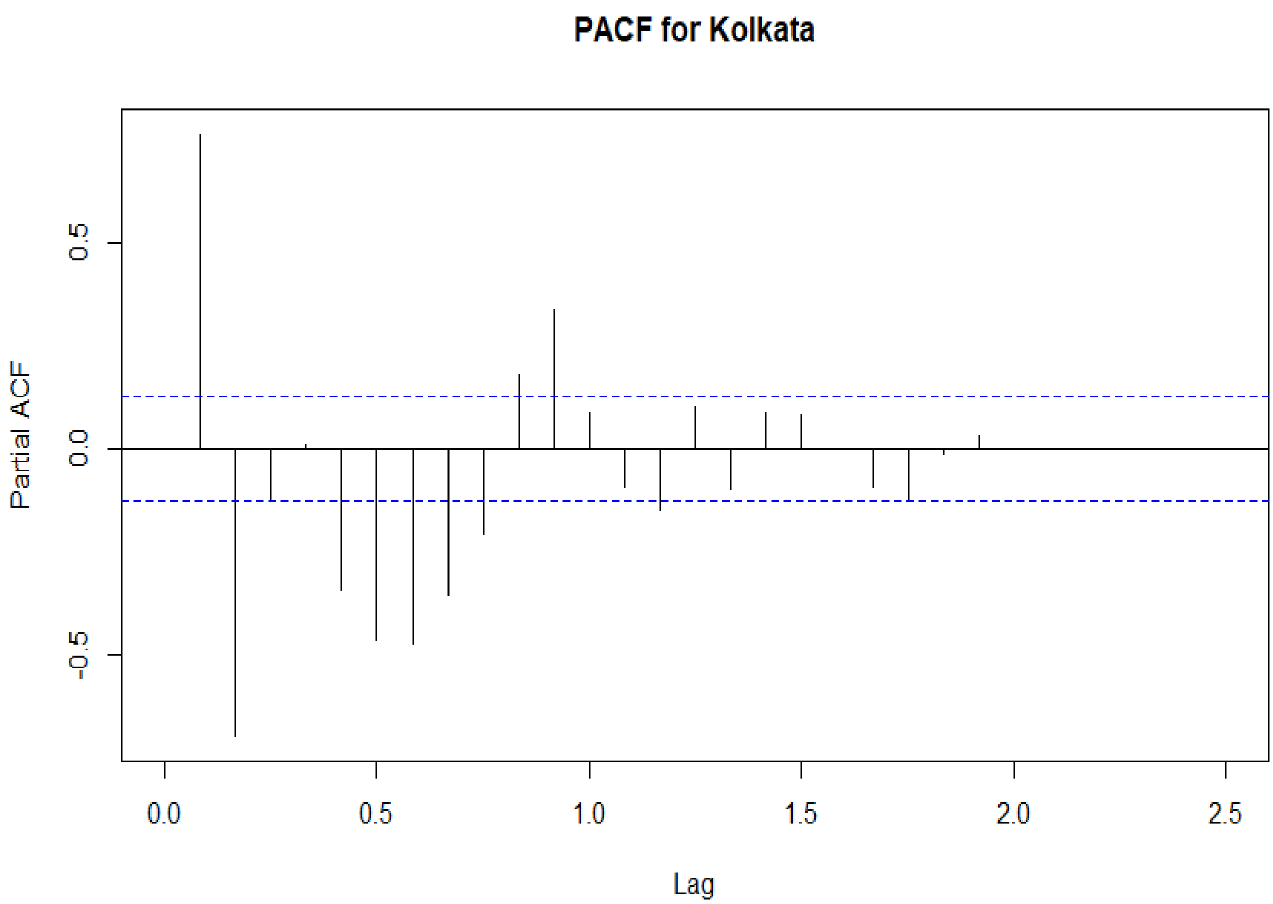


Diagram: 7

Interpretation:

From the above diagram (Diagram: 7), we can clearly observe the PACF plot, which enables us to identify the AR parameter p. In the ACF plot, there is a significant spike at lag 1.

The dashed blue lines indicate the 95% confidence interval, and for the correlations are significantly different from zero.

ARIMA

Autoregressive moving-average (ARMA) models provide a parsimonious description of a (weakly) stationary stochastic process in terms of two polynomials, one for the autoregression and the second for the moving average.ARIMA model is bascically an ARMA model fitted on d-th order differentiation time series such that th final differentiated time series is stationary.

We can split the Arima term into three terms, AR, I, MA:

❖ AR(p) stands for autoregressive model, the p parameter is an integer that confirms how many lagged series are going to be used to forecast periods ahead.
  For example: The average temperature of yesterday has a high correlation with the temperature of today, so we will use AR(1) parameter to forecast future temperatures.

  The formula for the AR(p) model is:

$$\widehat{y_t} = \mu + \theta_1 Y_{t-1} + \cdots + \theta_p Y_{t-p} ,$$

  where $\boldsymbol{\mu}$ is the constant term, the **p** is the periods to be used is the regression and $\boldsymbol{\theta}$ is the parameter fitted to the data.

❖ I(d) is the differencing part, the d parameter tells how many differencing orders are going to be used, it tries to make the series stationary.
  For example: Yesterday someone sold 10 items of a product, today that same person sold 14, the "I" in this case is just the first difference, which is +4, if we are using logarithm base this difference is equivalent to percentual difference.

- If d = 1, then

$$y_t = Y_t - Y_{t-1} ,$$

  where y_t is the differenced series and $Y_{t-period}$ is the original series.

- If d = 2, then
  $y_t = (Y_t - Y_{t-1}) + (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$

  Note that the second difference is a change-in-change, which is a measure of the local "acceleration" rather than trend.

❖ MA(q) stands for moving average model, the q is the number of lagged forecast errors terms in the prediction equation.
For example: It's strange, but this MA term takes a percentage of the errors between the predicted value against the real. It assumes that the past errors are going to be similar in future events.

The formula for the MA(p) model is:

$$\widehat{y_t} = \mu - \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} ,$$

where $\boldsymbol{\mu}$ is the constant term, **q** is the period to be used on the **e** term and $\boldsymbol{\theta}$ is the parameter fitted to the errors

The error equation is

$$e_t = Y_{t-1} - \widehat{y_{t-1}}$$

ARIMA models can be easily and accurately used for short-term forecasting with just the time series data, but it can take some experience and experimentation to find an optimal set of parameters for each use case.

In my project ,from the results of the ADF test, if the data is stationary then we can model the data by using an ARIMA(p,d,q) model. The order of the model is selected from the corresponding ACF & PACF plots. The ACF plot gives the order of the MA process while the PACF plot gives the order of the AR process. If there is a significant spike at lag 12 in the PACF plot, then we can expect seasonality in the data. The best model is selected using AIC criterion. After modelling the data using an appropriate order ARIMA model, we then check whether the residuals are random or not. If the residuals are random then the fit is good. To check whether the residuals are random or not, we are going to use Box-Ljung Test.

Our data has not been differentiated even once,so for our data the value of d will be equivalent to 0. Now, from ACF and PACF plot it can be observed that the apprppriate model for our data might be ARIMA(1, 0, 2).

Outcome after fitting the ARIMA model to our data:

```
Call:
arima(x = Kolkata, order = c(1, 0, 2))

Coefficients:
         ar1     ma1     ma2
      0.5591  0.8408  0.4951
s.e.  0.0642  0.0688  0.0514

sigma^2 estimated as 3.43:  log likelihood = -489.6,  aic = 989.19
```

Interpretation:

Here, we have used ARIMA(1,0,2) model. The coefficients of ar1, ma1 and ma2 are 0.5591, 0.8408 and 0.4951 respectively, the aic value is 989.19.

Residuals:

The "residuals" in a time series model are what is left over after fitting a model. For many (but not all) time series models, the residuals are equal to the difference between the observations and the corresponding fitted values:

$$et=yt-\hat{y}t$$

Residuals are useful in checking whether a model has adequately captured the information in the data. A good forecasting method will yield residuals with the following properties:

- The residuals are uncorrelated. If there are correlations between residuals, then there is information left in the residuals which should be used in computing forecasts.
- The residuals have zero mean. If the residuals have a mean other than zero, then the forecasts are biased.
- Any forecasting method that does not satisfy these properties can be improved. However, that does not mean that forecasting methods that satisfy these properties cannot be improved. It is possible to have several different forecasting methods for the same data set, all of which satisfy these properties. Checking these properties is important in order to see whether a method is using all of the available information, but it is not a good way to select a forecasting method.

If either of these properties is not satisfied, then the forecasting method can be modified to give better forecasts. Adjusting for bias is easy: if the residuals have mean mm, then simply add mm to all forecasts and the bias problem is solved.

In addition to these essential properties, it is useful (but not necessary) for the residuals to also have the following two properties:

- The residuals have constant variance.

- The residuals are normally distributed.

These two properties make the calculation of prediction intervals easier. However, a forecasting method that does not satisfy these properties cannot necessary be improved. Sometimes applying a Box-Cox transformation may assist with these properties, but otherwise there is usually little that you can do to ensure that your residuals have constant variance and a normal distribution. Instead, an alternative approach to obtaining prediction intervals is necessary. Again, we will not address how to do this until later in the book.

We also need to have residuals checked for this model to make sure this model will be appropriate for our time series forecasting.

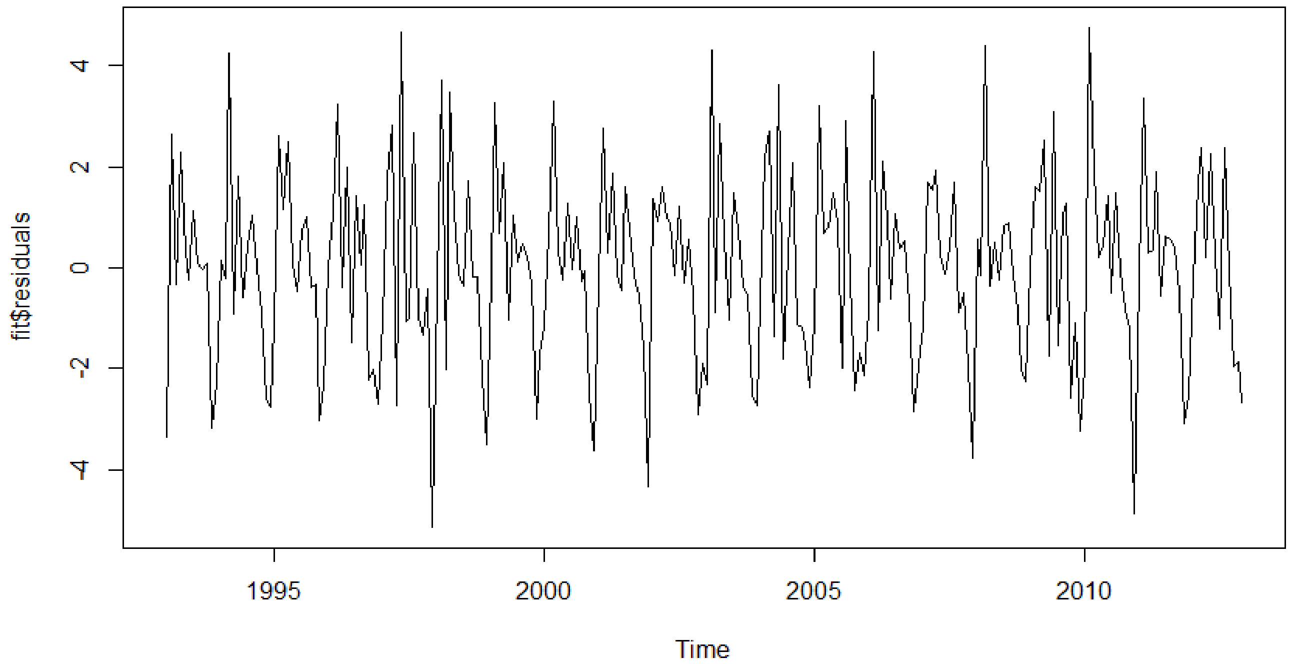


Diagram: 8

Interpretation:

From the above diagram (Diagram: 8), we can observe that it looks like white noise, but we cannot interpret anything else clearly yet.

Ljung-Box Test:

The Box-Ljung test (1978) is a diagnostic tool used to test the lack of fit of a time series model.

The test is applied to the residuals of a time series after fitting an ARMA (p, q) model to the data. The test examines m autocorrelations of the residuals. If the autocorrelations are very small, we conclude that the model does not exhibit significant lack of fit.

In general, the Box-Ljung test is defined as:

H_0 : The model does not exhibit lack of fit

H_α: The model exhibits lack of fit.

Test: Given a time series Y of length n.

Statistic: The test Statistic is defined as

$$Q = n(n+2) \sum_{k=1}^{m} (\hat{r}^2_k / n-k)$$

Where, $\hat{r}_k$ is the estimated autocorrelation of the series at lag k, and m is the number of lags being tested.

Significance Level: α

Critical Region: The Box- Ljung test rejects the null hypothesis (indicating that the model has significant lack of fit if,

$$Q > \chi^2_{1-\alpha,h}$$

Where, $\chi^2_{1-\alpha,h}$ is the chi-square distribution table value with h degrees of freedom and significance level α. Because the test is applied to residuals, the degrees of freedom must account for the estimated model parameters so that h = m – p – q, where p and q indicate the number of parameters from the ARMA (p, q) model fit to the data.

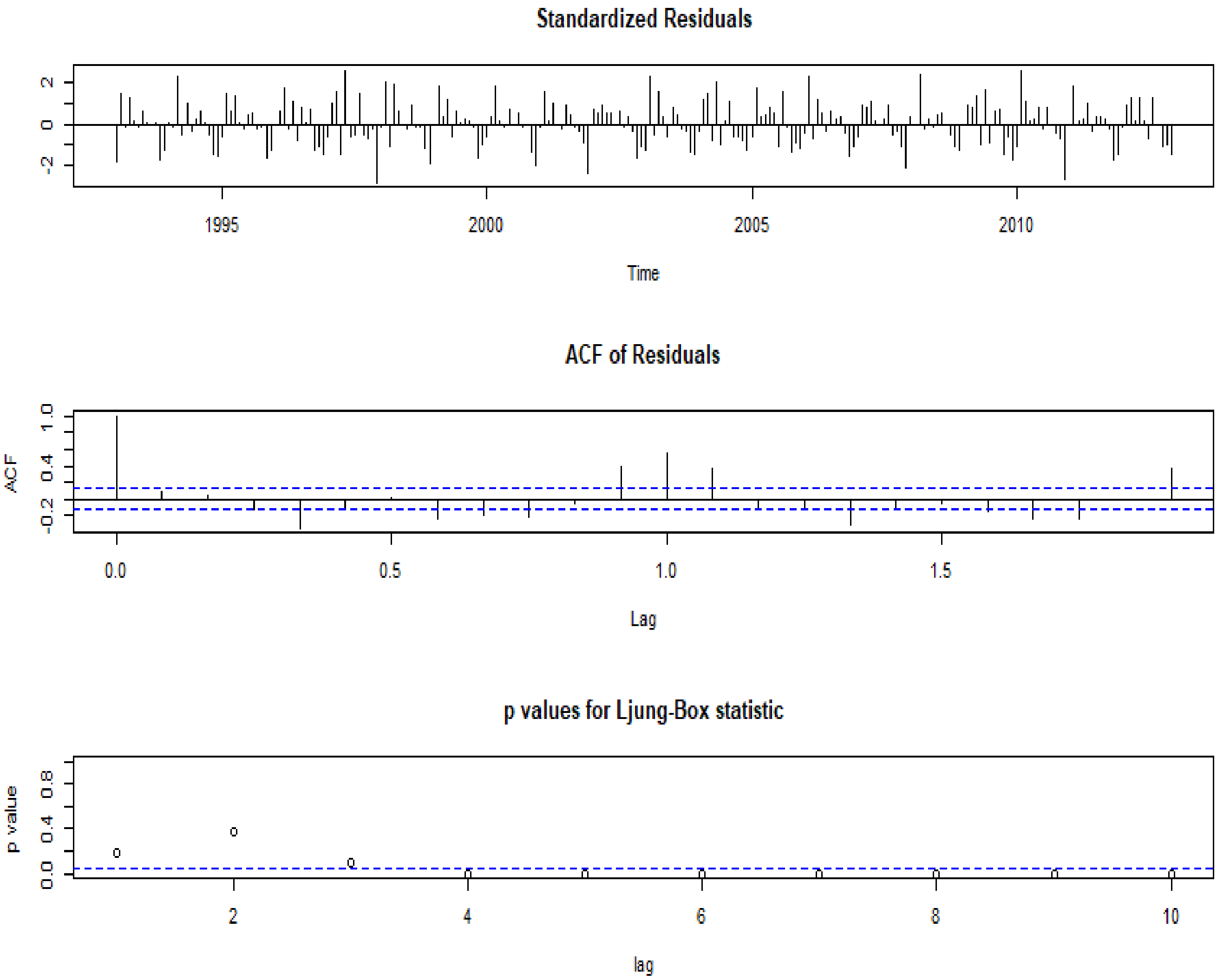


Diagram: 9

Interpretation:

From the above diagram (Diagram: 8), we can observe that the errors are random and from the ACF plot of residuals, we can conclude that this model is appropriate for forecasting, since its residuals show white noise behaviour and uncorrelated against each other.

FORECASTING

Time series forecasting is a technique for the prediction of events through a sequence of time. The technique is used across many fields of study, from the geology to behavior to economics. The techniques predict future events by analyzing the trends of the past, on the assumption that future trends will hold similar to historical trends.

Making predictions about the future is called extrapolation in the classical statistical handling of time series data. More modern fields focus on the topic and refer to it as time series forecasting.

Forecasting involves taking models fit on historical data and using them to predict future observations.

Descriptive models can borrow for the future (i.e., to smooth or remove noise), they only seek to best describe the data.

An important distinction in forecasting is that the future is completely unavailable and must only be estimated from what has already happened.

The purpose of time series analysis is generally two fold, such as

- It is usefull to understand or model the stochastic mechanisms that gives rise to an observed series.
- It is usefull to predict or forecast the future values of a series based on the history of that series.

The skill of a time series forecasting model is determined by its performance at predicting the future. This is often at the expense of being able to explain why a specific prediction was made, confidence intervals and even better understanding the underlying causes behind the problem.

Now, we are basically going to consider our time series data from 1993-2011 and try to predict the data for the year 2012, this will help us check whether or not our model and forecasting method is accurate.

Table – 2: Table showing Time series data on average temperature in Kolkata from 1993-2011

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1993	18.722	23.158	25.972	29.349	30.168	29.705	29.392	29.050	28.537	27.751	24.083	20.250
1994	19.662	21.457	27.809	29.782	31.489	29.794	29.221	29.178	29.139	27.617	23.762	19.577
1995	17.935	22.048	26.834	31.445	32.014	30.407	29.085	29.365	28.951	27.735	23.705	19.880
1996	19.283	22.374	28.349	30.415	31.965	29.567	29.385	28.631	29.695	27.139	23.594	19.395
1997	18.240	21.471	27.492	27.590	30.893	30.473	29.121	29.254	28.783	26.927	24.698	19.334
1998	17.639	22.461	25.213	29.418	31.205	31.690	29.514	29.546	29.296	28.591	25.269	20.410
1999	18.950	23.720	28.334	31.831	30.548	29.971	28.929	28.944	28.544	27.764	24.071	20.749
2000	19.196	21.275	26.881	30.165	30.245	29.861	29.320	29.694	28.846	28.017	24.668	19.728
2001	18.220	22.654	26.858	30.192	30.142	28.899	28.989	29.773	29.600	27.925	25.146	19.784
2002	19.676	22.625	27.406	29.798	30.845	30.001	29.972	29.083	28.866	27.616	24.046	20.513
2003	17.813	23.065	26.173	30.571	31.340	30.086	29.425	29.629	29.223	27.536	23.920	19.955
2004	18.661	22.416	28.452	29.587	32.045	30.153	29.123	29.399	28.883	26.720	23.552	20.621
2005	19.341	23.576	27.695	30.152	31.062	31.595	28.833	29.485	29.398	26.943	22.916	19.754
2006	19.269	24.980	27.602	30.323	30.762	30.101	29.523	29.163	29.391	28.040	24.175	20.506
2007	18.959	22.060	26.415	30.511	31.381	30.287	29.042	29.887	29.111	27.527	24.253	19.582
2008	19.065	20.805	27.880	30.538	31.134	29.111	28.852	29.284	29.087	27.422	24.227	20.863
2009	20.431	23.564	27.711	31.778	30.585	31.665	29.575	29.533	29.626	27.273	24.281	19.817
2010	17.503	22.959	29.451	32.318	31.308	31.075	29.979	30.226	29.536	28.071	25.437	19.603
2011	17.864	22.566	27.394	29.276	30.395	29.873	29.473	28.976	29.092	28.097	24.110	19.767

Now, we are going to try to forecast the average temperature data for the year 2012 and compare it with our original data.

Our predicted data:

$pred

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
2012	18.82360	20.87346	24.72979	26.59865	27.37390	27.08343	27.08543	26.79692

Sep	Oct	Nov	Dec
28.09835	25.52033	25.34123	24.44036

$se

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
2012	1.610598	2.487458	2.939024	3.025622	3.043562	3.047330	3.048120	3.048273

Sep	Oct	Nov	Dec
3.048239	3.047918	3.046361	3.044074

Our original data for the year 2012:

[1] 18.815 22.408 27.878 30.299 32.232 31.959 29.680 29.762 29.653 27.427 23.487 19.621

Forecasting plot:

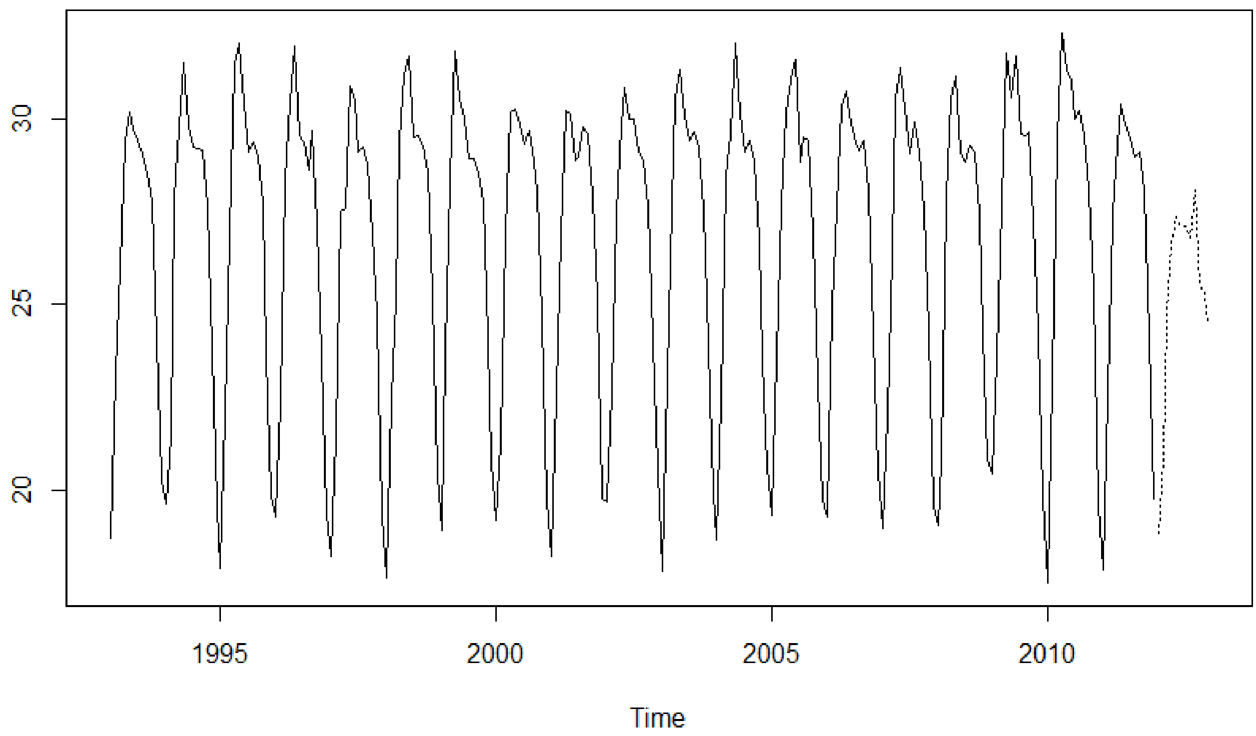


Diagram: 10

Interpretation:

I have tried to predict the average temperature in Kolkata in 2012 using the previously available data from 1993-2011. But clearly, we need to fix some further complex model for the data to get some better forecasting values, which is beyond our scope of study.

CONCLUSION

In this project, I analysed the time series data on average temperature in Kolkata from 1993-2012. Change in temperature is a major cause of climate change, so I tried to analyse the data and find an appropriate model to help us make useful predictions, which in turn may help us to take suitable precautions keeping the predictions in mind.

At first, I have analysed the time series data using various tools for analysis and then, obtained the appropriate model for our data, in order to make useful forecasts. Then, I compared the forecasted data with the original data to check the accuracy of our model.

The end results were somewhat satisfactory, but not completely up to our expectations, so we need to fix some further complex model for the data to get some better forecasting values, which is beyond our scope of study. I have mainly done this project to develop some idea about time series forecasting, which I consider to be a success.

R PROGRAMS

- Kolkata <- ts(Kolkata_temp, frequency = 12, start = c(1993,1))
  Kolkata
  #Converting our data to time series data

- plot(Kolkata, main = "Average temperature in Kolkata from 1993 to 2012", xlab = "Year", ylab = "Average Temperature")

- avg_temp <- decompose(Kolkata)

  plot(avg_temp)

  #Decomposition of our time series data

- install.packages("tseries")

  library(tseries)

- adf.test(Kolkata, alternative = c("stationary", "explosive"), k = trunc((length(kolkata)-1)^(1/3)))
  #Checking whether our time series data is stationary or not

- install.packages("forecast")

  library(forecast)

- seasonplot(Kolkata, year.labels = TRUE,col=1:13, main = "Seasonal Plot", ylab = "Average Temperature")
  #Obtaining the seasonal plot

- install.packages("tsutils")

  library(tsutils)

- seasplot(Kolkata,outplot=2,trend=FALSE, main="Seasonal Box Plot",ylab="Average Temperature")
  # Obtaining the seasonal box plot

- acf(Kolkata, main = "ACF for Kolkata", xlim = c(0,2.5))
  #Finding the MA(q) value

- pacf(Kolkata, main = "PACF for Kolkata", xlim = c(0,2.5))
  #Finding the AR(p) value

- fit <- arima(Kolkata, order = c(1,0,2))
  fit
  #Fitting the ARIMA model

- plot(fit$residuals)
  #Checking the residuals for our model

- tsdiag(fit)
  #Checking whether the errors are random or not

- data <- ts(Kolkata_temp, frequency = 12, start = c(1993,1), end = c(2011, 12))
  data
  #Obtaing time series data from 1993-2011

- pred <- predict(fit1, n.ahead = 1*12)
  pred
  # Predicting the data for 2012 using the data from 1993-2011

- original <- tail(kolkata, 12)
  original
  #Calling the original data

- ts.plot(data, pred$pred, lty = c(1,3))
  #Plotting the predicted data

ACKNOWLEDGEMENT

I, Debopriya Bose, would like to extend my gratitude to our respected Principal madam, Dr. Indrila Guha; HOD of my department, Mrs. Rusati Sen and to my supervisor, Dr. Soumita Modak for her expert guidance and constant supervision as well as for providing necessary information regarding the project. I am also grateful to all the other professors of the Statistics Department, Basanti Devi College for their kind co-operation and encouragement which helped me in completion of this project. I would also take this opportunity to thank R-Software for developing open-source software which helped me immensely in my dissertation. I would also like to thank all of my friends and my parents for always supporting me and for being by my side in the time of doing the project. This Project would not have been possible without their support and guidance. I am grateful to all of them for having faith in me and for cooperating with me.

…………………………………………..

Debopriya Bose

Semester – VI,

Department of statistics,

Basanti Devi College

BIBLIOGRAPHY

- Fundamental of Statistics (Volume-2) by A.M. Gun, M.K. Gupta & B. Dasgupta (Published by D. Chakraborty for The World Press Private Limited, Kolkata)
- https://en.wikipedia.org/wiki/Time_series
- https://www.influxdata.com/what-is-time-series-data/
- https://www.brainkart.com/article/Definition-and-Uses-of-Time-Series_39266/#:~:text=%C2%B7-,Time%20series%20analysis%20is%20used%20to,fluctuation%20in%20economics%20and%20business.&text=It%20helps%20in%20the%20evaluation%20of%20current%20achievements.&text=Time%20series%20is%20used%20in,weather%20forecasting%20and%20earthquake%20prediction
- https://www.investopedia.com/terms/s/secular.asp#:~:text=Examples%20of%20secular%20trends%20include,the%20growth%20in%20impact%20investing
- https://en.wikipedia.org/wiki/Climate_change#:~:text=Climate%20change%20threatens%20people%20with,health%20in%20the%2021st%20century.
- https://www.abs.gov.au/websitedbs/d3310114.nsf/home/time+series+analysis:+the+basics
- https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/
- https://otexts.com/fpp2/stationarity.html
- https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test
- https://en.wikipedia.org/wiki/Seasonal_subseries_plot
- https://en.wikipedia.org/wiki/Stochastic_process
- https://en.wikipedia.org/wiki/Partial_autocorrelation_function#:~:text=In%20time%20series%20analysis%2C%20the,not%20control%20for%20other%20lags.