Concept of Time Series Analysis and Use of Time Series Data in Agricultural Research for Validating and Forecasting

[1*]Sandeep Kumar, Assistant Professor, Department of Basic Science (Statistics), College of Horticulture, Sardar Vallabhbhai Patel University of Agriculture and Technology, Modipuram, Meerut.

ABSTRACT

The present research focuses on the application of time series data analysis and its significance in understanding and predicting agricultural trends. Time series analysis is a crucial tool for agricultural students and researchers, as it provides insights into past patterns, helps forecast future values, and assists in making informed decisions for crop management and productivity enhancement. Through time series methods, historical data can be examined to identify underlying trends, seasonal variations, and fluctuations, which are vital for accurate forecasting. this study emphasizes the fundamental concepts of time series analysis, including the identification of its components (trend, seasonality, cyclical variations, and irregular fluctuations) and the decomposition process used to better understand these elements. Additionally, both linear non-stationary models and linear stationary models are explored to determine their suitability in modeling agricultural data.

Introduction

Time Series data

The variable containing observations over time is called a time series variable, and the dataset is referred to as time series data. Each observation is referenced with a point of time, say, date or month, or year, or even in seconds and microseconds. Time series data may be evenly spaced, like daily sales data or unevenly spaced, e.g., measuring the weight of animals at different periodicities, say, the first few observations are taken daily, the next few observations every week, and subsequently on a monthly and annual basis based on the variations in the data.

A time series is a set of statistics, usually collected at regular intervals. Time series data occur naturally in many application areas.

According to *Ya-lun Chou*, "A time series may be defined as a collection of readings belonging to different periods, of some economic variable or composite of variables."

Mathematically, a time series is defined by the functional relationship

$$y_t = f(t)$$

Where y_t is the value of the phenomenon (or variable) under consideration at time *t.*

Some examples of typical time series data in agricultural research

- Annual yield of a particular crop in a particular location over years
- Consumption of foodgrains over months
- Sale of pesticides over the years
- Private investment in agriculture (annual data)
- Monthly data on employment in the tea garden

Components of a Time Series Data

Any time series can contain some or all of the following components:

1. Trend (T_t)
2. Cyclical (C_t)
3. Seasonal (S_t)
4. Irregular (I_t)

Trend component

The trend is the long-term pattern of a time series. A trend can be positive or negative, depending on whether the time series exhibits an increasing or decreasing long-term pattern. If a time series does not show an increasing or decreasing pattern, then the series is stationary in the mean.

E.g., Population growth in India

Cyclical component

Any pattern showing an up-and-down movement around a given trend is identified as a cyclical pattern. The duration of a cycle depends on the type of business or industry being analyzed.

Seasonal component

Seasonality occurs when the time series exhibits regular fluctuations during the same month (or months) every year, or the same quarter every year. For instance, retail sales peak during December.

E.g., Sales in festive seasons

Irregular component

This component is unpredictable. Every time series has some unpredictable component that makes it a random variable. In prediction, the objective is to model all the components to the point that the only component that remains unexplained is the random component. E.g.: Earthquake

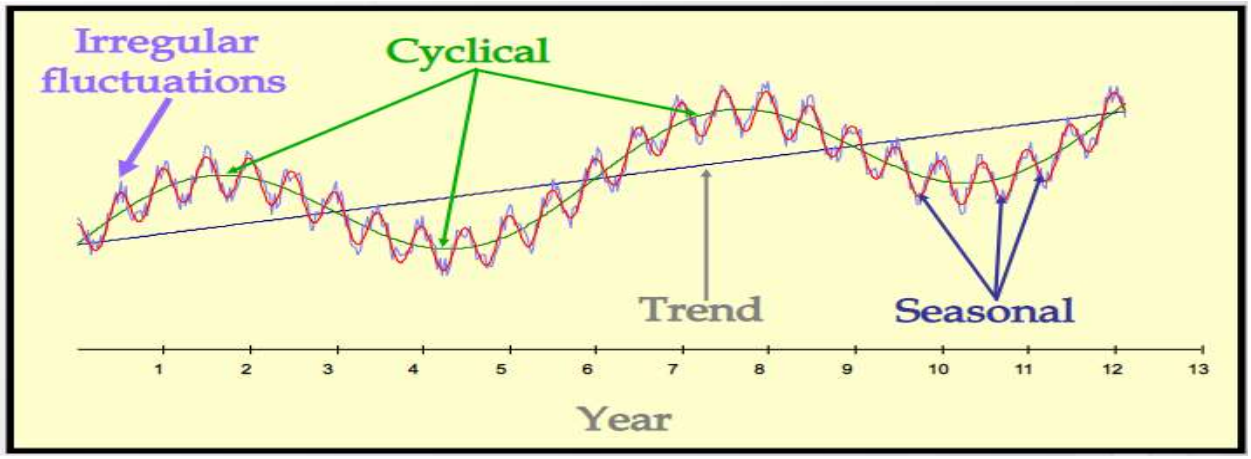

Fig.1 Components of Time series

Time Series Decomposition Model

For analysis of time series data, a model is essential. Generally, two broad approaches are resorted to. One is a multiplicative model, and the other is an additive model.

Let the original observation at the time point to be denoted by Y_t and the four components, *viz.,* Trend, seasonal, cyclical, and irregular variations by (T_t), (S_t), (C_t) and (I_t) respectively, for a time period t (where t = 1, 2, 3…).

The following two structures are considered for basic decomposition models:

1. Additive: Y_t = Trend (T_t) + Seasonal (S_t) + Cyclical (C_t) + Irregular (I_t)
2. Multiplicative: Y_t = Trend (T_t) x Seasonal (S_t) x Cyclical (C_t) x Irregular (I_t)

Time Series Model

1. **Linear Stationary Time Series Models**
   - i. Auto-Regressive Model
   - ii. Moving Average Model
   - iii. ARMA Model (Mixed Model)
2. **Linear Non-Stationary Time Series Models**
   - i. ARIMA Model

Stationary Series

A series x_t is said to be stationary if it satisfies the following properties:

- The mean $E(x_t)$ is the same for all t.
- The variance of x_t is the same for all t.
- The covariance (and also correlation) between x_t and x_{t-1} is the same for all t.

The Dickey-Fuller test is the most widely used statistical test for stationarity. To carry out the test, estimate using OLS and a regression model. If the series is non-stationary, it can be converted to a stationary series by differencing.

Linear Stationary Time Series Models

Assume we have a time series without trends or seasonal effects. That is, if necessary, any trends or seasonal effects have already been removed from the series. How might we construct a linear model for a time series with autocorrelation?

(1) **Autoregressive model of order *p***: AR(p), which has the general form

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

Where,

y_t = Response (dependent) variable at time t

y_{t-1}, y_{t-2}, ……,y_{t-p} = Response variable at time lags t-1, t-2, ……,t-p, respectively

μ = Constant mean of the process

$\phi_1, \phi_2, \ldots\ldots\ldots, \phi_p$ = Coefficients to be estimated

ε_t = Error term at time t

(2) **Moving Average model of order *q***: MA(q), which has the general form

$$y_t = \delta + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

Where,

y_t = Response (dependent) variable at time t

μ = Constant mean of the process

$\theta_1, \theta_2, ..., \theta_q$ = Coefficients to be estimated

ε_t = Error term at time t Error in previous periods that are incorporated in the response y_t

(3) **Autoregressive-moving average model of order *p* and *q*:** ARMA (*p,q*), which has the general form

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

Where, $\varepsilon_t's$ are independently and normally distributed with zero mean and constant variance σ^2 for t = 1,2,…..,n.

Autocorrelation function (ACF)

The coefficient of correlation between two values in a time series is called the autocorrelation function (ACF). The ACF for a time series y_t is given by:

$$\text{Cor}\,(y_t, y_{t-k})$$

The value of k is the time gap being considered and is called the lag.

Partial Autocorrelation Function (PACF)

In general, a partial correlation is a conditional correlation. It is the correlation between two variables under the assumption that we know and takes into account the values of some other set of variables.

For instance, consider a regression context in which y is the response variable and x_1 ,x_2 and $x_3$are predictor variables. The partial correlation between y and$x_3$is the correlation between the variables determined taking into account how both y andx_3 are related to$x_1$andx_2 .

$$\frac{\text{Covariance}(y, x_3 \,|x_1, x_2)}{\sqrt{\text{Variance}(y|x_1, x_2)\,\text{Variance}(x_3\,|x_1, x_2)}}$$

Correlogram

Graphical approaches to assessing the order of an autoregressive (AR) and moving average (MA) model include looking at the ACF and PACF values versus the lag. The correlogram is a two-dimensional graph between the lag k and the autocorrelation coefficient $\rho_s$which is plotted as lag on the X-axis and ρ_s on the Y-axis.

The PACF is most useful for identifying the order of an autoregressive model, and the ACF is useful for identifying the order of a moving average model. A correlogram gives a summary of correlation at different periods of time. The plot shows the correlation coefficient for the series lagged (in distance) by one delay at a time. For example, at x=1 you might be comparing January to February or February to March. The horizontal scale is the time lag and the vertical axis is the autocorrelation coefficient (ACF).

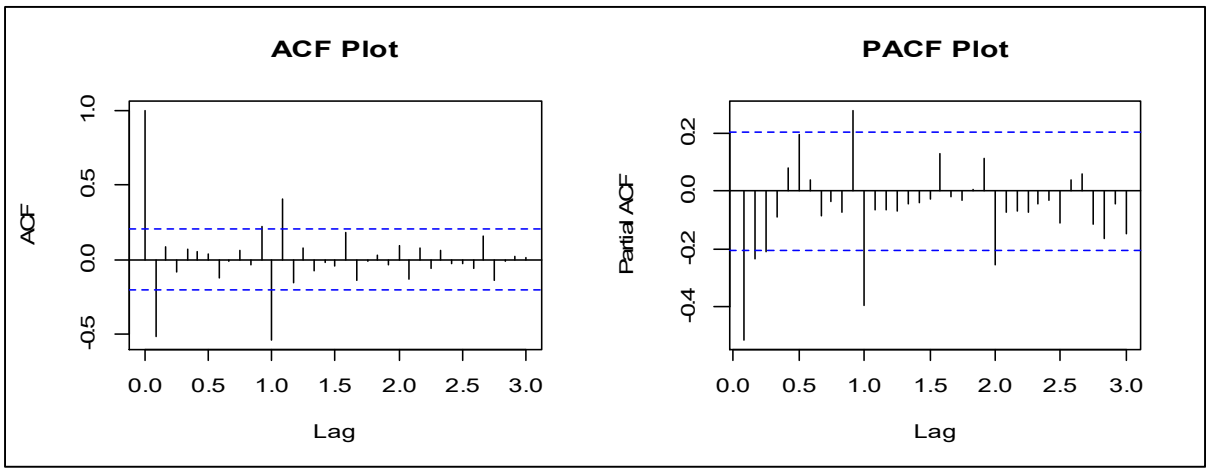

Fig:2 ACF and PACF Plot

In a plot of ACF versus the lag, if you see large ACF values and a non-random pattern, then likely the values are serially correlated. In a plot of PACF versus the lag, the pattern will usually appear random, but large PACF values at a given lag indicate this value as a possible choice for the order of an autoregressive model. It is important that the choice of the order makes sense.

Tests for Error Normality

Many of the statistical procedures, including correlation, regression, t tests, and analysis of variance, namely parametric tests, are based on the assumption that the data follows a normal distribution or a Gaussian distribution that it is assumed that the populations from which the samples are taken are normally distributed. The assumption of normality is especially critical when constructing reference intervals for variables. Normality and other assumptions should be taken seriously, for when these assumptions do not hold, it is impossible to draw accurate and reliable conclusions about reality.

Visual Methods

Visual inspection of the distribution may be used for assessing normality, although this approach is usually unreliable and does not guarantee that the distribution is normal. The frequency distribution (histogram), stem-and-leaf plot, boxplot, P-P plot (probability-probability plot), and Q-Q plot (Quantile-Quantile plot) are used for checking normality visually.

The frequency distribution that plots the observed values against their frequency provides both a visual judgment about whether the distribution is bell-shaped and insights about gaps in the data and outlier values. The stem-and-leaf plot is a method similar to the histogram, although it retains information about the actual data values. The P-P plot plots the cumulative probability of a variable against the cumulative probability of a particular distribution.

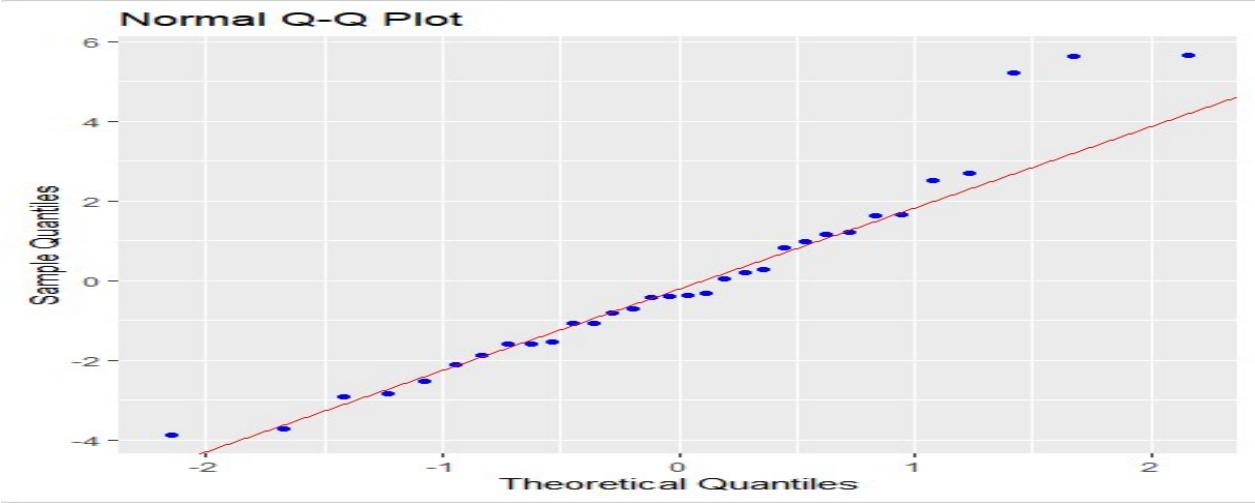

Fig.3 Normal Quantile-Quantile for Error Diagnostics plot

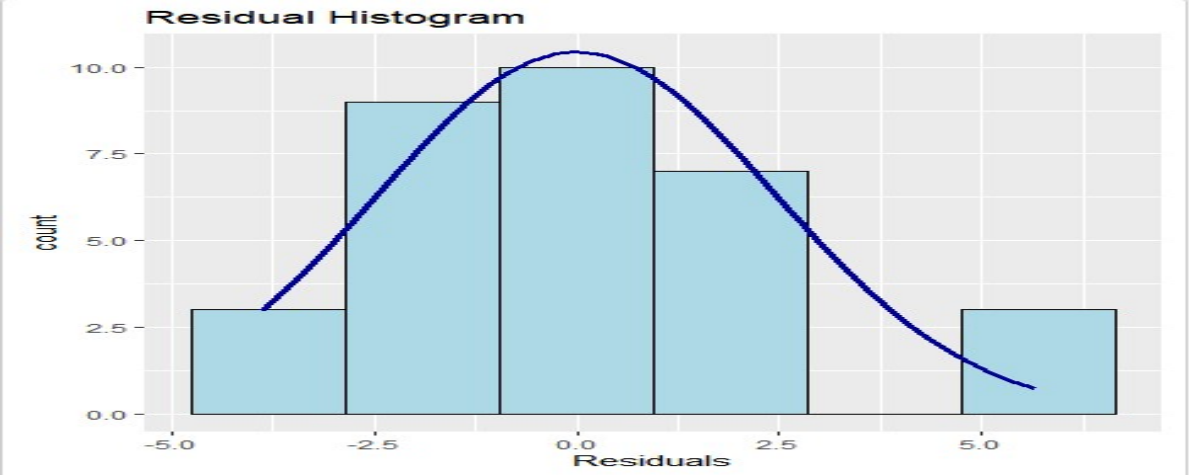

Fig.4 Residual Histogram plot

Normality Tests for testing the Normality of the series

For each test discussed below, the formal hypothesis test is written as:

$\boldsymbol{H_0}$: The errors follow a normal distribution

$\boldsymbol{H_1}$: The errors do not follow a normal distribution.

While hypothesis tests are usually constructed to reject the null hypothesis, this is a case where we hope we fail to reject the null hypothesis, as this would mean that the errors follow a normal distribution.

1. Anderson-Darling Test

The Anderson-Darling Test measures the area between a fitted line (based on the chosen distribution) and a nonparametric step function (based on the plot points). The statistic is a squared distance that is weighted more heavily in the tails of the distribution. Smaller Anderson-Darling values indicate that the distribution fits the data better. The test statistic is given by:

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} (\log \mathrm{F}(e_i) + \log(1 - \mathrm{F}(e_{n+1-i})))$$

Where F(·) is the cumulative distribution of the normal distribution. The test statistic is compared against the critical values from a normal distribution in order to determine the p-value.

2. Shapiro-Wilk Test

The Shapiro-Wilk Test uses the test statistic

$$W = \frac{\left(\sum_{i=1}^{n} a_i e_{(i)}\right)^2}{\sum_{i=1}^{n}(e_i - \bar{e})^2},$$

Where $e_i$pertains to the i^{th} largest value of the error terms and the a_i values are calculated using the means, variances, and covariance's of the e_i. W is compared against tabulated values of this statistic's distribution. Small values of W will lead to rejection of the null hypothesis.

3. Ryan-Joiner Test

The Ryan-Joiner Test is a simpler alternative to the Shapiro-Wilk test. The test statistic is actually a correlation coefficient calculated by

$$R_p = \frac{\sum_{i=1}^{n} e_{(i)} z_{(i)}}{\sqrt{s^2(n-1)\sum_{i=1}^{n} z_{(i)}^2}},$$

Where the $z_{(i)}$ values are the z-score values (i.e., normal values) of the corresponding $e_i$value and s^2 is the sample variance. Values of R_p closer to 1 indicate that the errors are normally distributed.

Identification

At the identification stage, we use two graphical devices to measure the correlation between the observations within a single data series. These devices are called an estimated Autocorrelation function (ACF) and an estimated partial autocorrelation function (PACF). The estimated ACF and PACF are used to measure the statistical relationships within a data series in a simple way but they give an idea about the patterns in the available data. Once we have an idea about the relationship between the observations in a time series. This relationship is expressed in the form of an equation. The basic thinking about the technique is that each process which occurs on a time scale has its own theoretical ACF and PACF. As the time series understudy is a

particular realization of the process the theoretical ACF and PACF must resemble the estimated ACF and PACF of the data series under study.

Table 1: Pattern of ACF and PACF for AR, MA, and ARMA processes

Process	**ACF**	**PACF**
AR (Auto Regressive)	Decays Towards zero	Cut off to zero (lag length of last spike is the order of the process)
MA (Moving Average)	Cut off to zero (lag length of last spike is the order of the process)	Decays towards zero
ARMA (Auto regressive and Moving Average)	Tails off towards zero	Tails off towards zero

Before the identification stage, some basic concepts of linear time series analysis, such as stationarity, non-stationarity, seasonality, and differencing, are also covered for any model building. Here we are discussing all the basic terminologies that are used for model identification.

Stationarity

The basis of time series analysis is to check whether the data is stationary or not. The time series is said to be stationary if the mean, variance and auto-covariance (at various lags) does not change regardless of what is the point measure, *i.e.* it fixed over time (Fig).

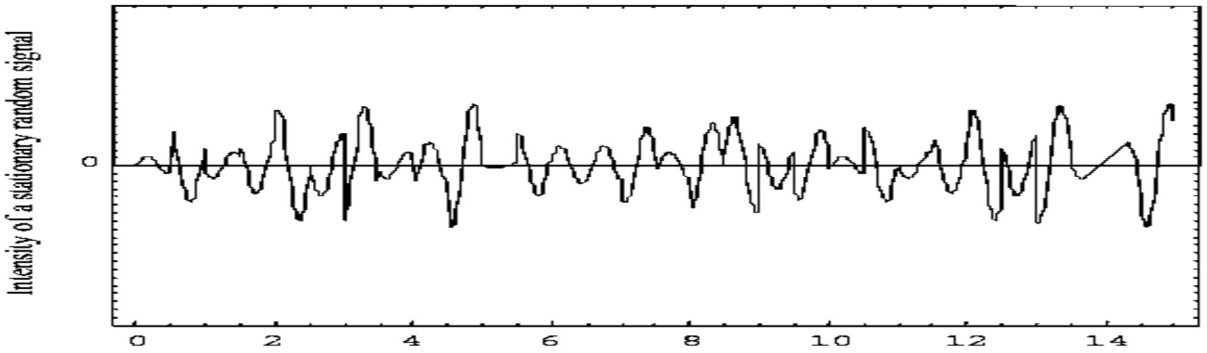

Fig. 5 Sample path of a stationary process

Moreover, the time series $\{r_t\}$ is said to be strictly stationary if the joint distribution of r_{t1} ,..., r_{tk} is identical to that of r_{t1-s},..., r_{tk-s} for all choices of t_1, t_2,., t_k and all choices of time lag s. In other words, strict stationarity requires that the joint distribution of r_{t1},..., r_{tk} is constant under a time shift. A weaker version of stationarity is often assumed.

A time series $\{r_t\}$ is weakly stationary if both the mean of r_t and the covariance between r_t and r_{t-s} are time-invariant, where s is an arbitrary integer. More specifically, $\{r_t\}$ is weakly stationary if:

1) $E(r_t) = \mu$, which is a constant, for all t.
2) $Cov(r_t, r_{t-s}) = \gamma_s$, which only depends on all-time t and lag s.

Non-stationary

A time series exhibits non-stationarity if the underlying generating process does not have a constant mean and/or a constant variance. As an example, the series given below displays considerable variation, especially since 2001, and a stationary model does not seem to be reasonable (Fig).

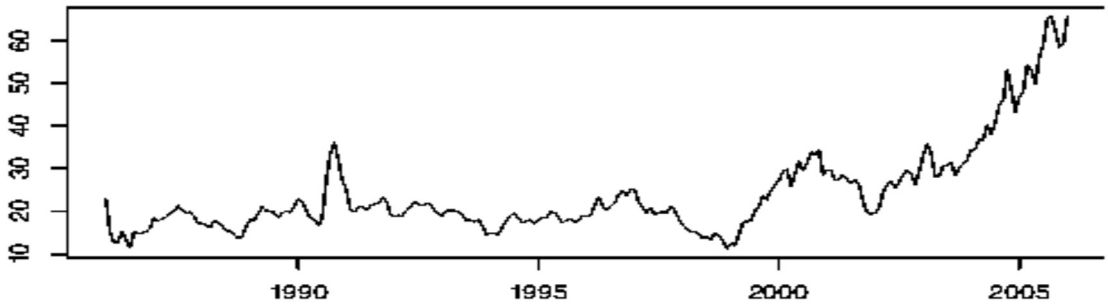

Fig.6 Sample path of a non-stationary process

Seasonality

In addition to trend, which has now been provided for, stationary series quite commonly display seasonal behavior where a certain basic pattern tends to be repeated at regular seasonal intervals. The seasonal pattern may additionally display constant change over time as well. In the figure given below, there is a strong upward trend but also a seasonality that can be seen.

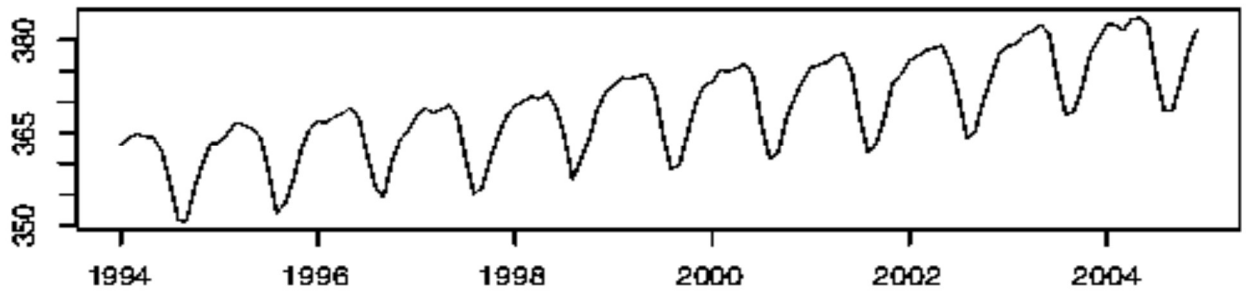

Fig.7 Sample path of a seasonal process

Just as regular differencing was applied to the overall trending series, seasonal differencing (SD) is applied to seasonal non-stationarity as well. And as autoregressive and moving average tools are available with the overall series, so too, they are available for seasonal phenomena using seasonal autoregressive parameters (SAR) and seasonal moving average parameters (SMA).

Autocorrelation Function (ACF)

The most important tool for studying dependence is the sample autocorrelation function. The correlation coefficient between any two random variables X, Y, which measures the strength of linear dependence between X, Y, always takes values between -1 and 1. If stationarity is assumed and the autocorrelation function ρ_k for a set of lags K = 1,2, ... is estimated by simply computing the sample correlation coefficient between the pairs, k units apart in time. The correlation coefficient between Y_t and Y_{t-k} is called the lag-k autocorrelation or serial correlation coefficient of Y_t and is denoted by the symbol ρ_k, which, under the assumption of weak stationarity, is defined as[4,5]

$$\rho_k = \frac{\sum_{t-k+1}^{T}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t-1}^{T}(Y_t - \overline{Y})^2} = \frac{\gamma_k}{\gamma_0} \; ; \; for \; k = 1,2,.....where, \gamma_k = \mathrm{cov}(Y_t, Y_{t-k})$$

Since ρ_k is a correlation, it has the simple properties:

a) $-1 \leq \rho_k \leq 1$, $\rho_k = \rho_{-k}$, $\rho_0 = 1$

Partial Autocorrelation Function (PACF)

`The correlation coefficient between two random variables Y_t and Y_{t-k} after removing the impact of the intervening Y_{t-1}, Y_{t-2} ,..., Y_{t-k+1} is called (PACF) at lag k and denoted by ϕ_{kk} .

$\phi_{00} = 1 \quad \phi_{11} = p_1$

$$\phi_{kk} = \frac{p_k - \sum_{j=1}^{k-1} \phi_{k-1,j} p_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j} p_j}, k = 2,3,.....where \phi_{k,j} = \phi_{k-1,j} - \phi_{k,k}\phi_{k-1,k-1}$$

A linear time series model can be tentatively identified by its Autocorrelation function (ACF), and Partial Autocorrelation Function (PACF) as follows [6]

- if ρ_1 is non-zero, this indicates that the series is first-order correlated.
- If ρ_k tails off geometrically with increasing lags, and the PACF cuts off after a certain lag, which means that the model is an autoregressive process.
- If ρ_k cut off after a small number of lags, and PACF tails off geometrically with increasing lags it means that the model is a moving-average process.

A plot of ρ_k versus lag k is often called a correlogram.

White Noise (WN)

A very important case of a stationary process is called white noise. For a white noise series, all the ACFs are zero or close to zero. If $\{r_t\}$ is normally distributed with zero mean and variance σ^2 and no autocorrelation, then it is said to be Gaussian white noise.

Diagnostic Checking and Forecasting

After having estimated the parameters of a tentatively identified ARIMA model, it is necessary to do diagnostic checking to verify that the model is adequate. The basic way of analyzing the goodness of the model is to check the residuals of the fitted model and the tool for analyzing the residuals is the residual ACF. It is assumed that they are independent of each other. Therefore, the residual [1]ACF for a properly built ARIMA model will ideally have autocorrelation coefficients that are all statistically zero or close to zero. Since the model has been estimated from a realization the ACF of the residuals will be subjected to sampling error. To test whether the estimated coefficients are statistically zero or not, a t-test is used. If the t-values for residual autocorrelations are significant, a reformulation of the model has to be done.

The final model is used to generate prediction values and then calculate the errors for the values obtained by the developed model.

Box and Jenkins gave the following characteristics of a good ARIMA model -

1. It is parsimonious (uses the smallest number of coefficients needed to explain the given data)
2. It is stationary (has *AR* coefficients which satisfy some mathematical inequalities).
3. It is invertible (has MA coefficients which satisfy some mathematical inequalities).
4. It has uncorrelated residuals.
5. It fits the available data (the past) well enough to satisfy the analyst:
   Root-mean-squared error (RMSE) is acceptable
6. It forecasts the future satisfactorily.

Forecast Performance Measures

Making Real-Time Forecasts: A Few Points

We have studied various useful and popular techniques for time series forecasting. For the implementation of the model, apply these methods for generating forecasts. While applying a particular model to some real or simulated time series, first the raw data is divided into two parts, viz., the Training Set and Test Set. [5]The observations in the training set are used for constructing the desired model. Often, a small subpart of the training set is kept for validation purposes and is known as the Validation Set. Sometimes, a preprocessing

is done by normalizing the data or taking logarithmic or other transforms. One such famous technique is the Box-Cox Transformation. Once a model is constructed, it is used for generating forecasts. The test set observations are kept to verify how accurately the fitted model performed in forecasting these values. If necessary, an inverse transformation is applied to the forecasted values to convert them to the original scale. To judge the forecasting accuracy of a particular model or to evaluate and compare different models, their relative performance on the test dataset is considered.

Due to the fundamental importance of time series forecasting in many practical situations, proper care should be taken while selecting a particular model. For this reason, various performance measures are proposed in the literature to estimate forecast accuracy and to compare different models. These are also known as performance metrics. Each of these measures is a function of the actual and forecasted values of the time series.

Description of Various Forecast Performance Measures

Now we shall discuss about the commonly used performance measures and their important properties. In each of the forthcoming definitions, y_t is the actual value, f_t is the forecasted value, $e_t = y_t - f_t$ is the forecast error and n is the size of the test set. Also,

$$\text{Test mean } \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \text{ , test variance } \sigma = \frac{\sum_{i=1}^{n} (y-\bar{y})^2}{n}$$

The Mean Forecast Error (MFE)

This measure is defined as MFE = $\frac{\sum_{t=1}^{n} e_t}{n}$. The properties of MFE are:

- It is a measure of the average deviation of forecasted values from actual ones.
- It shows the direction of error and is thus also termed as the *Forecast Bias*.
- In MFE, the effects of positive and negative errors cancel out, and there is no way to know their exact amount.
- A zero MFE does not mean that forecasts are perfect, i.e. contain no error; rather, it only indicates that forecasts are on the proper target.
- MFE does not panelize extreme errors.

- It depends on the scale of measurement and also affected by data transformations.
- For a good forecast, i.e., to have a minimum bias, the MFE should be as close to zero as possible.

The Mean Squared Error (MSE)]

Mathematical definition of this measure is MSE $=\frac{\sum_{t=1}^{n}|e_t|}{n}$

- It measures the average absolute deviation of forecasted values from original ones.
- It is also termed as the *Mean Absolute Deviation (MAD)* It shows the magnitude of overall error, occurred due to forecasting.
- In MAE, the effects of positive and negative errors do not cancel out. Unlike MFE, MAE does not provide any idea about the direction of errors.
- For a good forecast, the obtained MAE should be as small as possible.
- Like MFE, MAE also depends on the scale of measurement and data transformations.
- Extreme forecast errors are not panelized by MAE.

The Mean Squared Error (MSE)

Mathematical definition of this measure is MSE = $\frac{\sum_{t=1}^{n} {e_t}^2}{n}$.

- It is a measure of average squared deviation of forecasted values.
- As here the opposite signed errors do not offset one another, MSE gives an overall idea of the error occurred during forecasting.
- It panelizes extreme errors that occurred while forecasting.
- MSE emphasizes the fact that the total forecast error is much affected by large individual errors, i.e., large errors are much expensive than small errors.
- MSE does not provide any idea about the direction of the overall error.
- MSE is sensitive to the change of scale and data transformations.
- Although MSE is a good measure of overall forecast error, it is not as intuitive and easily interpretable as the other measures discussed before.

The Root Mean Squared Error (RMSE)

Mathematical definition of this measure = $\sqrt{\frac{\sum_{t=1}^{n} e_t{}^2}{n}}$.

RMSE is nothing but the square root of the calculated MSE. All the properties of MSE hold for RMSE as well. We have discussed ten important measures for judging the forecast accuracy of a fitted model. Each of these measures has some unique properties, different from others. In experiments, it is better to consider more than one performance criterion. This will help to obtain a reasonable knowledge about the amount, magnitude, and direction of overall forecast error. For this reason, time series analysts usually use more than one measure for judgment.

References

1. Kumar, S., Buragohain, R., Mohammad, C., Akram, S. K. P., and Michelle, C. Statistical Modeling of Nanomaterial Efficacy in Agricultural Applications.
2. Kumar, S., Lakhera, M. L., Tamrakar, M., and Ray, D. 2023. Impact of insects (YSB, BPH, RLF) in rice crop yield in Raipur district of Chhattisgarh.
3. Kumar, S., Mehta, V., Kumar, A., Singh, A., and Kumar, H. 2025. AREA VISUALIZATION OF SOME KHARIF CROPS DIVERSIFICATION IN CHHATTISGARH STATE OF INDIA. *Plant Archives*, *25*(1), 269-281.
4. Kumar, S., Mehta, V., Mourya, K. K., and Kumar, A. 2024. A Comprehensive Study on Trend Analysis of Area, Production and Productivity of Major Millets in India. *Environment and Ecology*, *42*(4C), 2030-2036.
5. Kumar, S., Mehta, V., Mourya, K. K., and Kumar, A. 2024. Comparative Study of Forecasting for Finger Millet (Ragi) According to their Area and Production in Different States of India through ARIMA Model. *Environment and Ecology*, *42*(4C), 1994-2003.
6. Sahu, T., Chouksey, N., Kumar, S., Rana, S. K., Kanauajia, S., and Yadav, R. 2023. Forecasting of honey bee population by ARIMAX model using weather variables.
7. Singh, A., Kumar, S., Agrahari, R. K., Pandey, A., Kumar, H., and Kumar, A. 2024. Study on Crop Diversification through Area Status of Crops in Kharif Season of Chhattisgarh. *International Journal of Environment and Climate Change*, *14*(4), 795-814.