INSTITUTE OF SCIENCE, NAGPUR

Project Report

On

TIME SERIES ANALYSIS AND FORECASTING OF OILSEEDS PRODUCTION IN INDIA

Submitted To

Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur

In Partial fulfillment for the degree of

Master of Science in Statistics

By

ANUSHRI PRASHANT GHATOLE

PAYAL KOMALDAS DHEKWAR

YASH NARAYAN FUNDE

Under the Guidance of

Dr. Sandhya Dhabe

Head,

Department of Statistics,

Institute Of Science,

Nagpur-440001

(2021-22)

CERTIFICATE

This is to certify that the project report entitled "TIME SERIES ANALYSIS AND FORECASTING OF OILSEEDS PRODUCTION IN INDIA", submitted by Ms. Anushri Prashant Ghatole, Ms. Payal Komaldas Dhekhwar and Mr. Yash Narayan Funde is a record of bonafied carried out under our guidance for the partial fulfilment of the degree of master of science in Statistics during the academic year 2021-22.

Date:-

Place:- Nagpur

Dr. Sandhya Dhabe
Professor and Head,
Department of Statistics
Institute of Science,
R.T.Road, Civil Lines,
Nagpur-440001

DECLARATION

We hereby declare that the work reported in the project entitled, "TIME SERIES ANALYSIS AND FORECASTING OF OILSEEDS PRODUCTION IN INDIA" has been carried out under the guidance of Dr. Sandhya Dhabe, Professor and Head, Department of statistics, Institute of science, Nagpur. The work has not been submitted as a whole or in part of any other university or institute for the award of degree or diploma or certificate.

ANUSHRI GHATOLE - _____

PAYAL DHEKWAR - _____

YASH FUNDE - _____

DATE:

PLACE: Nagpur

ACKNOWLEDGEMENT

It is our pleasure and privilege to express our sincere gratitude towards our guide, Dr. Sandhya Dhabe for her continuous and invaluable guidance, helpful suggestions, enthusiastic encouragement and faith in us during the entire course of our project.

We are greatly indebted to Dr. Anjali Rahatgaonkar, Director, Institute of Science, Nagpur for their inspiration.

We would like to express our sincere thanks to all the teaching and non-teaching staff of Department of Statistics for their timely help and support during the course of our project.

Lastly, we also wish to thank one and all who have been directly or indirectly instrumental in the completion of our project.

Place:

Date:

Anushri P. Ghatole

Payal K. Dhekwar

Yash N. Funde

M.Sc. II (Semester IV)

Department of Statistics,

Institute of Science, Nagpur.

INDEX

Abstract: Oilseeds create a very important group of commercial crops in India. As the oil extracted from oilseeds form an important item of our daily diet. Also they are used as raw materials for various items. Although India is the fourth largest oilseed producing economy in the world, it is also among the largest oil consuming countries. Hence the domestic edible oil production has not been able to keep pace with growth in consumption and country is incurring heavy cost owing to its dependence on imports. The need for addressing this deficit motivated a systematic study of the oilseeds economy to formulate appropriate strategies to bridge the demand-supply gap.

Statistical forecasting is used to provide assistance in decision making and planning the future more effectively and efficiently. Forecasting is a primary aspect of developing economy so that proper planning can be undertaken for sustainable growth of the country. In this study, an effort has been made to forecast the total oilseeds in India for the next four years by using Exponential smoothing method, Autoregressive Integrated Moving Average (ARIMA) model, and Group Method of Data Handling (GMDH).

Exponential smoothing method, ARIMA and GMDH are mathematical models well-known for time series forecasting. The results obtained by all methods are compared. The comparison results shows that the GMDH model perform better than the ARIMA model, Exponential smoothing method in terms of mean absolute percentage error (MAPE) and root mean square error (RMSE). The experimental results indicates that the GMDH model is an effective technique to handle the time series data and it provides a promising technique in time series forecasting methods.

Key words: Forecasting, Oilseeds production, Time Series analysis, Exponential Smoothing, ARIMA model, GMDH, India.

Chapter 1

INTRODUCTION

Introduction: Oilseeds create a very important group of commercial crops in India. They have been the backbone of India's agricultural economy. The oil extracted from oilseeds form an important item of our daily diet. And they are used as raw materials for manufacturing various number of items such as soaps, shampoo, cleaning products, toothpaste, candle, wax, paints, varnishes, hydrogenated oil, perfumery, lubricants, etc.

On the oilseeds map of world, India occupies a prominent position, not only in acreage and production but also in consumption. India is fourth largest oilseed producing economy in the world after USA, China, and Brazil. The major vegetable oilseeds production cultivated in our country are Groundnut, Castor oilseeds, Rapeseed (or mustard), Soyabean, Sunflower, Sesamum, Safflower and Niger seeds, Linseeds.

India contributes about 10% of worlds total oilseeds production and about 6-7% of the global production of vegetable oil. Although India has 20.8% of world's area under oilseeds crops, it accounts for only 10% of global production. This is because the low productivity of oilseed crops and year to year fluctuations in production in India. According to data during the period 2018-19 the productivity of oilseeds in India was 34.19 MMT (Million Metric Tonnes), whereas it was 61.00 MMT in USA, 60 MMT in Brazil and in China 58.6 MMT respectively. The reason of low and fluctuating productivity is mainly because cultivation of oilseed crops is mostly done on marginal lands of which 72% is confined to rainfed farming that means they are lacking in irrigation and using of low levels of input.

As a result, the domestic edible oil production has not been able to keep pace with growth in consumption and country is incurring heavy cost owing to its dependence on imports.

To improve the situation, Government of India is pursuing several development programs, such as National Food Security Mission (NSFM) has released some funds to National Seeds Corporation Ltd. (NSC). Various budget estimate and scheme released during the year 2007-08 to 2013-14 & NMOOP for 2014-15 to 2016-17 Integrated Scheme of Oilseeds, Pulses, Oil palm & Maize (ISOPOM), Integrated Scheme of Oilseeds, Pulses, Oil palm & Maize (ISOPOM), National Mission on Oilseeds and Oil Palm (NMOOP), National Mission on Oilseeds and Oil Palm NFSM (OS & OP), Oilseed Growers Cooperative Project, National Oilseed and Development Project, Technology Mission Oilseeds (TMO), Annual Action Plan for implementation of Front Line Demonstration

(FLD) and other related activities on oilseeds by Indian Council of Agricultural Research - Indian Institute of Spices Research (ICAR-IISR) during 2022-23 under National Food Security Mission Oilseeds (NFSM-OS) ,etc. The main objectives of National Food Security Mission Oilseeds & Oil Palm is to increase in production and productivity of vegetable oils sourced from oilseeds and Oil palm. It aims to augment the availability of vegetable oils and to reduce the import of edible oils by increasing the production and productivity of oilseeds.

The ongoing Ukraine-Russia war has disrupted global market. It has also affected India's edible oil market which gets more than 90% of its sunflower oil from these two countries. This war has once again highlighted that India needs to be self-sufficient in edible oils.

Statistical forecasting is used to provide assistance in decision making and planning the future more effectively and efficiently. Forecasting is a primary aspect of developing economy so that proper planning can be undertaken for sustainable growth of the country. In this study, an effort has been made to forecast the total oilseeds in India for the next four years using some statistical models. The model used for forecasting is an Exponential smoothing method. It was proposed by Prof. Brown, Prof. Holt and Prof. Winter. There are three types of exponential smoothing methods viz. single exponential smoothing method, double exponential smoothing method and triple exponential smoothing method. Holt's linear trend method, Brown's method and Damped trend methods have been used for exponential soothing and forecasting. An Autoregressive Integrated Moving Average (ARIMA) model also known as Box-Jenkins model was also used. This model was introduced by Box and Jenkins in 1960. This model is used for forecasting a single variable. The primary reason behind choosing ARIMA model for forecasting is that it assumes non-zero autocorrelation between the successive values of the time series data. But ARIMA model can only capture linear feature of time series data to deal with non-linearity of time series data, Group Method of Data Handling (GMDH) has also been used in our analysis for forecasting oilseeds production. This model was first used in 1966 by Prof. Alexey G. Ivakhnenko.

Objective: The objective of the study is to analyse the production of oilseeds in India by using Exponential Smoothing methods, Autoregressive Integrated Moving Average (ARIMA) model, and Group Method of Data

Handling (GMDH) model. An effort has been made to generate a short term forecast of the oilseeds production in India.

Chapter 2

REVIEW OF LITERATURE

Review of Literature:

There have been numerous studies and empirical researches on forecasting the production of oilseeds using variety of approaches. In this chapter we present a brief review of the same.

S. Pal, V. Ramsubramanian and S. C. Mehta (2007) have applied Double Exponential method and ARIMA model to forecast milk production in India. The validity of the models were verified with various model selection criteria such as minimum of AIC (Akaike Information Criteria) and lowest MAPE (Mean Absolute Percentage Error) values.

A. Shabri, R. Samsudin, and Z. Ismail (2009) have used Combined forecasting based on Artificial Neural Network (CANN) approach to forecast the rice yield in Malesia. Authors have compared the said model with the classical time series models (ARIMA, Exponential smoothing) and claimed that CANN performs reasonably well in predicting real life problems.

S. R. Krishnapriya, P. K. Bajpai and K. K. Suresh (2015) have done a comparative study between Double Exponential Smoothing method and Autoregressive Integrated Moving Average (ARIMA) method to forecast the sugarcane yield in Coimbatore.

A. Singh (2015) has also used Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Network to forecast the prices of Groundnut oil in Kolkata.

K. Karadas, S. Celik, E. Eyduran and S. Hopoglu (2017) attempts to forecast annual production of some oil seed crops (sesame, sunflower and soybean) using the three exponential smoothing methods, Holt, Brown and Damped Trend.

S. Ghosh (2017) has forecasted cotton exports in India using the ARIMA model. The goodness of fit of the model is observed through small values of RMSE. The study also statistically tested and validated the forecasted errors.

N. Vijay and G.C. Mishra (2018) examined the flexibility of Artificial Neural Network model (ANN) in time series forecasting by comparing with classical time series ARIMA model. The data regarding area and production of pearl millet (bajra) crop was used for the study.

D. Mithiya, L. Datta, and K. Mandal (2019) attempt to forecast the oilseeds production in India using Autoregressive Integrated Moving Average method (ARIMA) and Group Method of Data Handling (GMDH) model.

B. Dhyani, M. Kumar, P. Verma and A. Jain (2020) attempts to predict the future value of stock based on daily data of nifty 50 index using Autoregressive Integrated Moving Average (ARIMA) model.

E. R. Abraham, and et al (2020) studied the forecasting of soyabean production in Brazil. To analyse the data authors have used the Artificial Neural Network (ANN).

A. V. Akkaya (2020) has developed a forecasting model for the monthly electricity demand of turkey by using Group Method of Data Handling (GMDH)-type neural network. The author revealed that the GMDH- type neural network model was a better approach for forecasting monthly electricity demand in Turkey.

B. M. Nkurlu and et all (2020) attempted to forecast permeability using Group Method of Data Handling (GMDH) type neural network from well log data of the West arm of the East African Rift Valley. The authors further explored the comparative analysis of GMDH model, Back Propagation Neural Network (BPNN) and Radial Basic Function Neural Network (RBFNN) and found that GMDH model outperformed the BPNN model and RBFNN model.

Chapter 3

RESEARCH METHODOLOGY

3.1 Introduction:

In this chapter we give a brief description about the tools and method of data collection and analysis. We also give the systematic and theoretical analysis of the statistical tools and techniques applied to the field of study.

3.2 Data:

The study is based on secondary data. The data used for this study is the oilseeds production in India for the last 50 years, i.e. from 1970-71 to 2018-19 which is collected from "APY State Data", uploaded by Directorate of Economics and Statistics, Department of Agriculture, Cooperation and Farmers Welfare, Ministry of Agriculture and Farmers Welfare, Govt. of India (http://eands.dacnet.nic.in/latest_20011.htm).

3.3 Tools and Techniques:

The data has been analysed using IBM Statistical package for Social Sciences (SPSS), version 22, and GMDH shell software. The various data analysis techniques used in this study are Exponential Smoothing Methods, Autoregressive Integrated Moving Average (ARIMA) model and Group Method of Data Handling (GMDH) - one sub-model of Artificial Neural Network (ANN).

3.3.1 Exponential Smoothing Method:

Exponential smoothing method is a forecasting method of time series data. It is one of the time series forecasting method proposed by Sir Robert G Brown, Charles C. Holt and Peter Winters respectively. It is also called as "exponentially weighted moving average". Exponential smoothing is a method of prediction that uses weighted values of earlier sequence observations values to calculate future values. It allocates the highest weight to the latest observation and the weight decline exponentially in an organized way as observation gets older. The proposal of exponential smoothing is to smooth the original data and then make use of the smoothed sequence data to forecast upcoming values. This process is mainly helpful when the parameters relating the time series are varying gradually over time. Exponential smoothing method of prediction uses weighted average of past observations to predict future values. This method is convenient for forecasting series that reveal trend, seasonality or both in the data. The exponential smoothing method is classified into three methods based on trend and seasonal components of the time series data which is given as:

I. Single exponential smoothing

II. Double exponential smoothing

III. Triple exponential smoothing

Single exponential smoothing (SES) is a well-known method of forecasting for stationary time series data. However, it does not provide good result in the context of non-stationary time series data. Agricultural production data have a general tendency of increasing over the time. So, they contain a trend. Trend may be upward or downward. Upward trend will be seen in data containing agricultural production, population etc. So, SES will not be useful in such cases. Therefore, we take into account double exponential smoothing method.

Double exponential smoothing is usually more reliable for handling data that shows trends.

Here we are considering following three types of double exponential smoothing:

1. Holt's linear double exponential smoothing method
2. Brown's double exponential smoothing method
3. damped double exponential smoothing method

Corresponding estimates and smoothing parameter are used to estimate the forecasted results mentioned below:

$\mathbf{L_t}$ - level estimate

$\mathbf{T_t}$ - trend estimate

α - the smoothing parameter used for level.

γ - the smoothing parameter used for trend.

Ø - the smoothing parameter used for damped trend.

1. Holt's Linear Trend double exponential method

The Holt's method is one of the double exponential smoothing method used in the estimation of the series with trends use when data have linear trend. Holt's linear trend method includes two smoothing constants such as α and γ, two smoothing equations and one forecast equation. They are given as:

Forecast equation: $\bar{y}_{t+h} = L_t + hT_t$ (1)

Level equation: $L_t = \alpha y_t + (1-\alpha)*(L_{t-1} + T_{t-1})$ (2)

Trend equation: $T_t = \gamma(L_t - L_{t-1}) + (1 - \gamma)T_{t-1}$ (3)

Where,

y_t : t period actual value

L_t and T_t : estimate of the level and trend of time series respectively.

$\bar{y}_{t+h}$: forecasting the future value in forecast

α and γ values lie on the interval between 0 to 1 i.e. 0< α <1

2. Brown's double exponential smoothing method

Brown's double exponential smoothing method with one parameter is another exponential smoothing method for forecasting which is used when the data shows a linear trend and non-stationary data. The Brown's model is more suitable for increasing or decreasing trends in time series data. Brown's postulate of linear exponential smoothing is similar to linear moving average. The Brown's double exponential smoothing method has one parameter which lies between the numbers 0 and 1 (0<α<1). This parameter value is used to exponentially decrease the actual value in the older time period.

The equation used in Brown's linear double exponential smoothing method is:

- **Forecast equation:**

  $\bar{y}_{t+h} = a_t + b_t h$ (4)

- **Single exponential smoothing equation**

  $y'_t = \alpha y_t + (1 - \alpha) * y'_{t-1}$ (5)

  Where:

  y'_t : t period single smoothing

  y_t : t period actual value

- **double exponential smoothing equation**

  $y''_t = \alpha y'_t + (1 - \alpha) * y''_{t-1}$ (6)

  And

  $a_t = y'_t + (y'_t - y''_t) = 2y'_t - y''_t$ (7)

  $b_t = \frac{\alpha}{1-\alpha} + (y'_t - y''_t)$ (8)

y''_t: t period double smoothing

$a_t \; and \; b_t$: smoothing constants

$\bar{y}_{t+h}$: forecast value after h period

3. Damped trend double exponential smoothing method

The damped trend double exponential smoothing method is considered to perform an excellent forecasting. Gardner and McKenzie describe how a damping parameter (ø) can be used within Holt's method to give more control over trend in data. Smoothing parameter value ø lies between 0 and 1 (0<ø<1). Usually, the damping parameter ø is consider between 0.8 to 0.98. If ø = 1, the method is identical to the standard Holt method.

The damped method is expressed in the following equations.

- **Forecast equation:**

$$\bar{y}_t = L_t + (ø + ø^2 + \cdots + ø^h)T_t \quad (8)$$

- **Level equation:**

$$L_t = \alpha y_t + (1 - \alpha)(L_{t-1} + ø T_{t-1}) \quad (9)$$

- **Trend equation:**

$$T_t = \gamma * (L_t - L_{t-1}) + (1 - \gamma) * ø T_{t-1} \quad (10)$$

We can obtain the forecasted values by using the above three methods. The forecasted values are reported for a maximum of 5 years, as Exponential smoothing method is applicable for short term forecasting.

3.3.2 Autoregressive Integrated Moving Average (ARIMA) model:

The model used in this study is the autoregressive integrated moving average (ARIMA). ARIMA (p,d,q) is a linear model originating from the autoregressive model AR (p), the moving average model MA (q)and thus the combination of the two AR (p) and MA (q) is the ARIMA (p,d,q). The model is developed as follows-

Let Y_t is a discrete time series variable which takes different values over a period of time. The corresponding AR (p) model of Y_t series, Which is the generalizations of autoregressive model, can be expressed as:

AR (p): Yt

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Yt_{-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t \quad (11)$$

Where, Yt is the response variables at time t,

$Y_{t-1}, Y_{t-2}, \ldots \ldots Y_{t-p}$ is the respective variables at different time with lags;

μ is constant mean of the series, $\phi_1, \phi_2, \ldots, \phi_p$ are the coefficients; and ε_t is the error factor. ε_t is a white noise process, where $E(\varepsilon_t) = 0$, var $(\varepsilon_t) = \sigma^2 > 0$, cov$(\varepsilon_t, \varepsilon_{t-h}) = 0$, t, $h \neq 0$.

Similarly, the MA (q) model which is again the generalization of moving average model may be specified as:

$$\text{MA (q): } Y_t = \mu + \varepsilon_t - \delta_1 \varepsilon_{t-1} - \delta_2 \varepsilon_{t-2} - \ldots - \delta_q \varepsilon t_{-q} \quad (12)$$

Where, μ is the constant mean of the series;

$\delta_1, \delta_2, \ldots \delta_q$ is the coefficients of the estimated error term; ε_t is the error term.

By combining both the models, we get the Autoregressive Moving Average or ARMA models, which has general form as:

$$Y_t = \mu + \phi_1 y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t - \delta_1 \varepsilon_{t-1} - \delta_2 \varepsilon_{t-2} - \ldots - \delta_q \varepsilon_{t-q} \quad (13)$$

The complete procedure of model building and forecasting are described by Box and Jenkins 1976. There are basic steps viz. model identification, model selection and parameter estimation, diagnostic checking and forecasting.

(i) Model Identification:

The first step is to identify the order of ARIMA (p, d, q) model. That is selection of order of AR(p), MA(q) and I(d). d is estimated by taking differences (first order or second order) and applying unit root test (Augmented Dicky-Fuller test). The model specification and selection of order p and q involved plotting of autocorrelations functions (ACF) and partial autocorrelations functions (PACF) or correlogram of variables at different lag length. If the PACF displays a sharp cutoff while the ACF decays more slowly (i.e., has significant spikes at higher lags), we say that the series displays an AR

signature. However, if the ACF displays a sharp cutoff while the PACF decay more slowly, we say that the series displays an MA signature. The autocorrelation functions specify the order of moving average process, q and partial autocorrelations function select the order of autoregressive process p.

(ii) Estimation of the model:

ARIMA models are fitted and accuracy of the model has tested based on diagnostics statistics. Once the order of p, d, and q are identified, their statistical significance can be judged by t-distribution. The next step is to specify appropriate regression model and estimate it. ARIMA models are fitted and accuracy of the model was tested based on diagnostics statistics.

(iii) Diagnostic checking:

Now a question may arise that how we know whether the identified model is appropriate. One simple way to figure that out is by diagnostic checking the residual term obtained from ARIMA model by applying the same ACF and PACF functions. First obtaining the ACF and PACF of residual term up to certain lags of the estimated ARIMA model, and then checking whether the coefficients are statistically significant or not. The best model was selected based on the following diagnostics,

(i) Low Akaike Information Criteria (AIC): AIC is estimated by $AIC = -2\log_e(L) + 2m$, where m= p+q and L is the likelihood function.

(ii) Low Bayesian Information Criteria (BIC): The Bayesian information criterion is a criterion for model selection among a finite set of models. It is based, in part, on the likelihood function, and it is closely related to Akaike information criterion (AIC). Sometimes, Bayesian Information Criteria (BIC) is also used and estimated by $BIC = -2\log_e(L) + \log_e(N)m$. Where N is number of observation and m is the number of parameters.

a. The minimum Root Mean Square Error (RMSE) and Mean Absolute Percent Error (MAPE) are used as a measure of accuracy of the models. RMSE= $\sqrt{\sum_{t=1}^{n}\left(X_{Actual,t} - X_{Forecast,t}\right)^2 / n}$ and

$$MAPE=\frac{1}{n}\bar{\sum}_{t=1}^{n}\left[\frac{X_{Actual,t}-X_{Forecast,t}}{x_{Forecast,t}}\right]^2 x100$$ Where, $X_{Actual,t}$ and $X_{Forecast,t}$ are actual and forecast output at time t.

b. These may also be judged by Ljung-Box Q (LBQ) statistic2 under null hypothesis that autocorrelation co-efficient up to lag k is equal to zero. LBQ is used to assess assumptions after fitting a time series model (ARIMA), to ensure that the residuals are independent.

(iv) Forecasting:

Once the above three steps are over, then we can obtain the forecasted values by estimating the appropriate model. The forecasted values are reported for a maximum of 5 years, as ARIMA model is applicable for short term forecasting.

3.3.3 Group Method of Data Handling (GMDH) model:

Group method of data handling (GMDH) is a family of inductive algorithms for computer-based mathematical modelling of multi-parametric datasets that features fully automatic structural and parametric optimization of models.

GMDH is used in such fields as data_mining, knowledge_discovery, prediction, complex_____systems modelling, optimization and pattern recognition. GMDH algorithms are characterized by inductive procedure that performs sorting-out of gradually complicated polynomial models and selecting the best solution by means of the external criterion.

A GMDH model with multiple inputs and one output is a subset of components of the *base function* (1):

$$Y(x_1, \ldots, x_n) = a_0 + \sum_{i=1}^{m} a_i f_i$$

where f_i are elementary functions dependent on different sets of inputs, a_i are coefficients and m is the number of the base function components.

In order to find the best solution, GMDH algorithms consider various component subsets of the base function (1) called partial models. Coefficients of these models are estimated by the least squares method. GMDH algorithms gradually increase the number of partial model components and find a model structure with optimal complexity indicated

by the minimum value of an external criterion. This process is called self-organization of models.

As the first base function used in GMDH, was the gradually complicated Kolmogorov–Gabor polynomial (2):

$$Y(x_1, \ldots, x_n) = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j=i}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n}\sum_{j=i}^{n}\sum_{k=j}^{n} a_{ijk} x_i x_j x_k + \cdots$$

Usually more simple partial models with up to second degree

functions are used.

The inductive algorithms are also known as polynomial neural networks. Jürgen Schmidhuber cites GMDH as one of the first deep learning methods, remarking that it was used to train eight-layer neural nets as early as 1971.

External Criteria:

External criterion is one of the key features of GMDH. Criterion describes requirements to the model, for example minimization of Least squares. It is always calculated with a separate part of data sample that have not been used for estimation of coefficients. This makes it possible to select a model of optimal complexity according to the level of uncertainty in input data. There are several popular criteria:

- Criterion of Regularity (CR) – Least squares of a model at the sample
- Criterion of Minimum bias or Consistency – a squared error of difference between the estimated outputs (or coefficients vectors) of two models developed on the basis of two distinct samples A and B, divided by squared output estimated on sample B. Comparison of models using it, enables to get consistent models and recover a hidden physical law from the noisy data.

GMDH algorithm:

GMDH-type neural network algorithms are modelling techniques which learn the relations among the variables. In the perspective of time series, the algorithm learns the relationship among the lags. After learning the relations, it automatically selects the way to follow in algorithm. First, GMDH was used by Ivakhnenko (1966) to construct a high order polynomial. The following equation is known as the Ivakhnenko polynomial given by

$$y = a + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} d_{ijk} x_i x_j x_k + \cdots$$

where m is the number of variables and a, b, c, d, . . . are coefficients of variables in the polynomial, also named as weights. Here, y is a response variable, xi and xj are the lagged time series to be regressed. In general, the terms are used in calculation up to square terms as presented below,

$$y = a + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} x_i x_j$$

The GMDH algorithm considers all pairwise combinations of p lagged time series. Therefore, each combination enters each neuron. Using these two inputs, a model is constructed to estimate the desired output. In other words, two input variables go in a neuron, one result goes out as an output. The structure of the model is specified by Ivakhnenko polynomial in equation 4 where m = 2. This specification requires that six coefficients in each model are to be estimated.

The GMDH algorithm is a system of layers in which there exist neurons. The number of neurons in a layer is defined by the number of input variables. To illustrate, assume that the number of input variables is equal to p, since we include all pairwise combinations of input variables, the number of neurons is equal to h = $\binom{p}{2}$. The architecture of GMDH algorithm is illustrated in Figure 1 when there are three layers and four inputs.

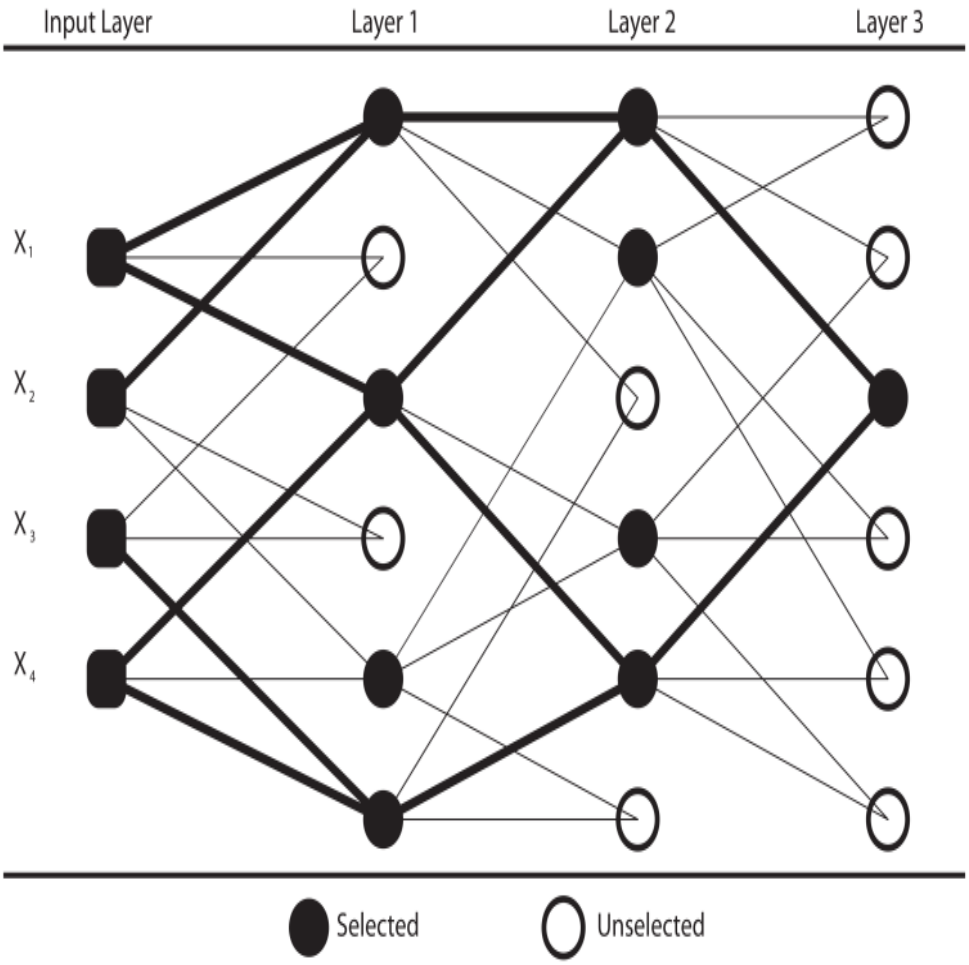

Figure 1: Architecture of GMDH algorithm

In the GMDH architecture shown in Figure 1, since the number of inputs is equal to four, the number of nodes in a layer is determined to be six. This is just a starting layer to the algorithm. The coefficients of equation 4 are estimated in each neuron. By using the estimated coefficients and input variables in each neuron, the desired output is predicted. According to a chosen external criterion, p neurons are selected and h−p neurons are eliminated from the network. In this study, prediction mean square error (PMSE) is used as the external criteria.

In Figure1, four neurons are selected while two neurons are eliminated from the network. The outputs obtained from selected neurons become the inputs for the next layer. This process continues until the last layer. At the last layer, only one neuron is selected. The obtained output from the last layer is the predicted value for the time series at hand. The flowchart of the algorithm is depicted in Figure 2.

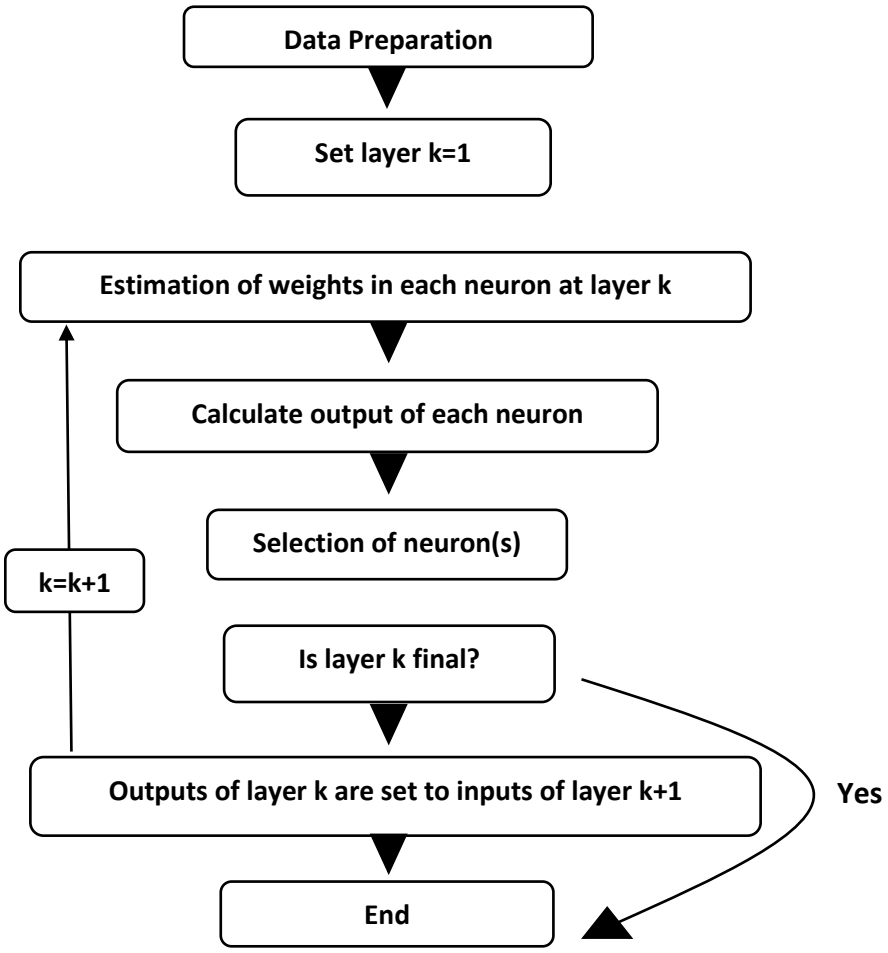

Figure 2: Flowchart of GMDH algorithms

In a GMDH algorithm, there exist six coefficients to be estimated in each model. Coefficients are estimated via RLSE.

Steps:

The main function of GMDH is based on the forward propagation of signal through nodes of the net similar to the principle used in classical neural nets. Every layer consists of simple nodes, each of which performs its own polynomial transfer function and passes its output to nodes in the next layer. The computation process comprises basic steps:

Step 1: Select input variables $\{x_1, x_2, x_k, \dots, x_n\}$ where n is the total number of inputs. The data are separated into training and testing data sets. The training data set is used to construct a GMDH model and the testing data set is used to evaluate the estimated GMDH model.

Step 2: Construct L numbers of new variables $Z = \{z_1, z_2, z_3, \ldots \ldots, z_L\}$ in the training data set for all

independent variables and choose a PD of the GMDH. Conventional GMDH has been developed using polynomial, PD of the following form

$$z_I = G(x_i, x_j) = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2$$

$$for\ I = 1,2,3, \ldots, L.$$

Where, $L = \frac{n(n-1)}{2}$

Select new variables as input of the next middle layer and truncate the subsequent computation. With the identification of the optimal output of partial polynomials at each layer, the selection of new variables enables the network to quickly converge to the target solution. Once the partial polynomial equations at each unit are selected, the residual error in each layer is further checked to determine whether the set of equations of the model should be further improved within the subsequent computation.

Step 3: Estimate the coefficient of the PD. The vectors of coefficients of the PDs are determined using the least square method.

Step 4: Determine new input variables for the next layer. There are several specific selection criteria to identify the input variables for the next layer. In our study, we used two criteria. The first criteria, the single best neuron out of these L neurons, Z′ identified according to the value of mean square error (MSE) of testing dataset. In second criteria, eliminate the least effective variables, replace the column of $\{x_1, x_2, x_k, \ldots, x_n\}$ by those columns $\{z_1, z_2, z_3, \ldots, z_I\}$ that best estimate the dependent variable y in the testing dataset.

Step 5: Build the final model and compute the predicted value. The final prediction model can be obtained with selected variables in each layer and the coefficients of partial polynomials between the connected layers. Check the stopping criterion. The lowest value of selection criteria using GMDH model at each layer obtained during this iteration is compared with the smallest value obtained at the previous one.

Table 1. An illustration of time series data structure in GMDH algorithms

Subjects	Y	X_1	X_2	X_3	X_p
1	y_t	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-p}
2	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-4}	y_{t-p-1}
3	y_{t-2}	y_{t-3}	y_{t-4}	y_{t-5}	y_{t-p-2}
.					
.					
.					
t-p	y_{p+1}	y_p	y_{p-1}	y_{p-2}	y_1

The GMDH algorithm is a system of layers in which there exist neurons. The number of neurons in a layer is defined by the number of input variables. To illustrate, assume that the number of input variables is equal to p; since we include all pair-wise combinations of input variables, the number of neurons is equal to $h = {}^{p}c_2$.

Time series prediction by GMDH:

A classical method for time series forecasting problem, the number of input nodes of nonlinear model, such as the GMDH is equal to the number of lagged variables $y_t = f(y_{t-1}, y_{t-2}, y_{t-3}, \dots, y_{t-p})$

where p is the number of chosen lagged. The outputs, yt, the predicted value of a time series defined a

$$y_t = f(y_{t-1}, y_{t-2}, y_{t-3}, \dots, y_{t-p})$$

However, there is no suggested systematic way to determine the optimum number of lagged p. The number of lagged p is chosen either in an adhoc basis or from traditional Box Jenkins methods. The lagged variables obtained from the Box-Jenkins analysis are the most important variables to be used as input nodes in the input layer of the GMDH model. In our study, a time series model is considered as nonlinear function of several past observations and random errors as follows:

$$y_t = f[(y_{t-1}, y_{t-2}, y_{t-3}, \dots, y_{t-p}), (a_{t-1}, a_{t-2}, a_{t-3}, \dots, a_{t-q})]$$

where f is a nonlinear function determined by the GMDH.

Data structure of GMDH: An illustration of time series data structure in GMDH algorithms is presented in Table 1. Since we have a time series data set with t time points and p inputs. We construct the model for the data with time lags, the number of observations presented under the subject

column in the table is equal to t-p; and the number of inputs i.e, lagged time series, is p. In this table, the variable called y is put in the models as a response variable, and the rest of the variables are taken into models as lagged time series xi , where i = 1,2,...,p. The notations in Table 1 are followed throughout this paper.

A better model which explains the relation between response and lagged time series is captured via transfer functions.

Chapter 4

RESULTS AND DISCUSSION

4.1 Exponential Smoothing Method

We have used SPSS software to plot, forecast time series trend and estimate the error rate. Figure 4.1 shows the time series graph of Oilseeds production in India for the period 1970-2019 and it can be seen that the data is non-seasonal with observable linear trend.

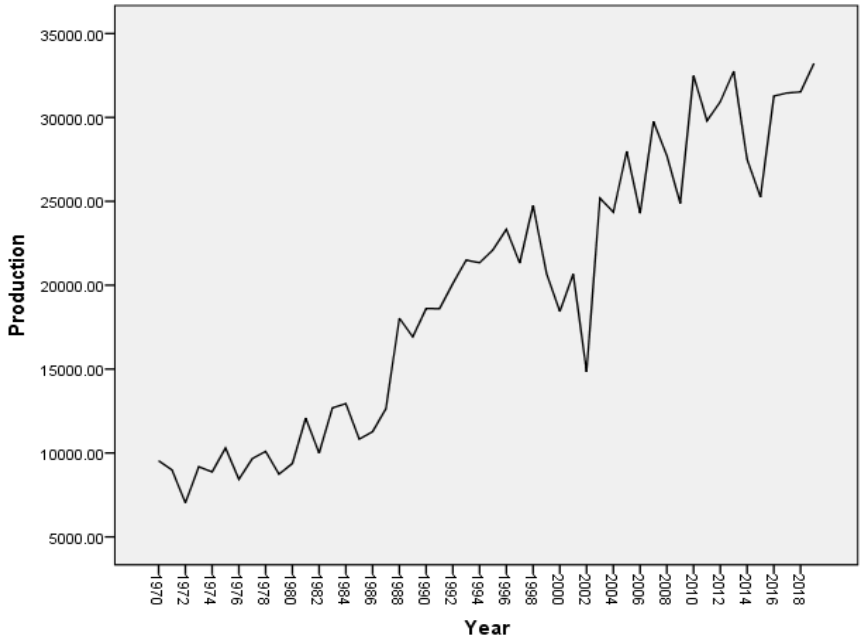

Figure 4.1 Actual time series plot of oilseeds production in India (1970-2019)

1. Holt's linear trend double exponential smoothing method

As our data is showing linear trend with no seasonality, we use holt's linear trend method. In Holt's linear trend model, results of model fit statistics Stationary R^2, R^2, MAPE, MAE and Normalized BIC criteria are evaluated which are summarized in Table 4.1.

In Table 4.2, Corresponding smoothing parameters of Holt's smoothing method were estimated as α =0.001 and γ=0.0001 respectively.

Table 4.1 Model Fit Statistics

Model	Number of Predictors	Model Fit statistics				
		Stationary R-squared	RMSE	MAPE	MAE	Normalized BIC
Production-Model_1	0	.774	2609.888	12.168	2008.373	15.891

Model Fit Statistics

Number of Predictors	Ljung-Box Q(18)			Number of Outliers
	Statistics	DF	Sig.	
0	40.617	16	.001	0

Table 4.2 Exponential Smoothing Model Parameters

Model			Estimate	SE
Production-Model_1	No Transformation	Alpha (Level)	.001	.064
		Gamma (Trend)	.0001	9.257

Thus, the equations given as

Forecast equation: $\bar{y}_t = L_t + hT_t$

Level equation: L_t**= (0.001)**y_t**+ (1− 0.001)*(** $L_{t-1} + T_{t-1}$**)**

Trend equation: T_t **= (0.0001)*(**$L_t - L_{t-1}$**)+(1-0.0001)**T_{t-1}

The measured autocorrelations among residuals at different lags using holt's linear model are shown in the table 4.3 and the measured partial autocorrelations among the residuals for several lags are shown in the table 4.4

Table 4.3 Estimated Residuals ACF for oilseeds production

Model		1	2	3	4	5	6	7	8	9	10
Production-Model_1	ACF	.270	.239	.114	-.115	-.099	-.106	-.109	-.326	-.207	-.313
	SE	.141	.151	.159	.160	.162	.163	.165	.166	.178	.183

11	12	13	14	15	16	17	18	19	20	21	22	23	24
-.328	-.128	.044	.048	.239	.063	.116	.072	.024	.101	.012	.083	.069	.011
.193	.204	.206	.206	.206	.212	.212	.213	.214	.214	.215	.215	.216	.216

Table 4.4 Estimated Residuals PACF for oilseeds production

Model		1	2	3	4	5	6	7	8	9	10
Production-Model_1	PACF	.270	.179	.014	-.207	-.065	-.006	-.022	-.330	-.094	-.176
	SE	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141

11	12	13	14	15	16	17	18	19	20	21	22	23	24
-.208	-.075	.130	-.052	.099	-.229	.023	-.119	-.139	-.063	-.062	-.035	.179	-.035
.141	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141

We will examine the autocorrelations and partial autocorrelations of the residuals of the fitted model. The ACF and PACF graphs of residuals are shown in the Figure 4.2.

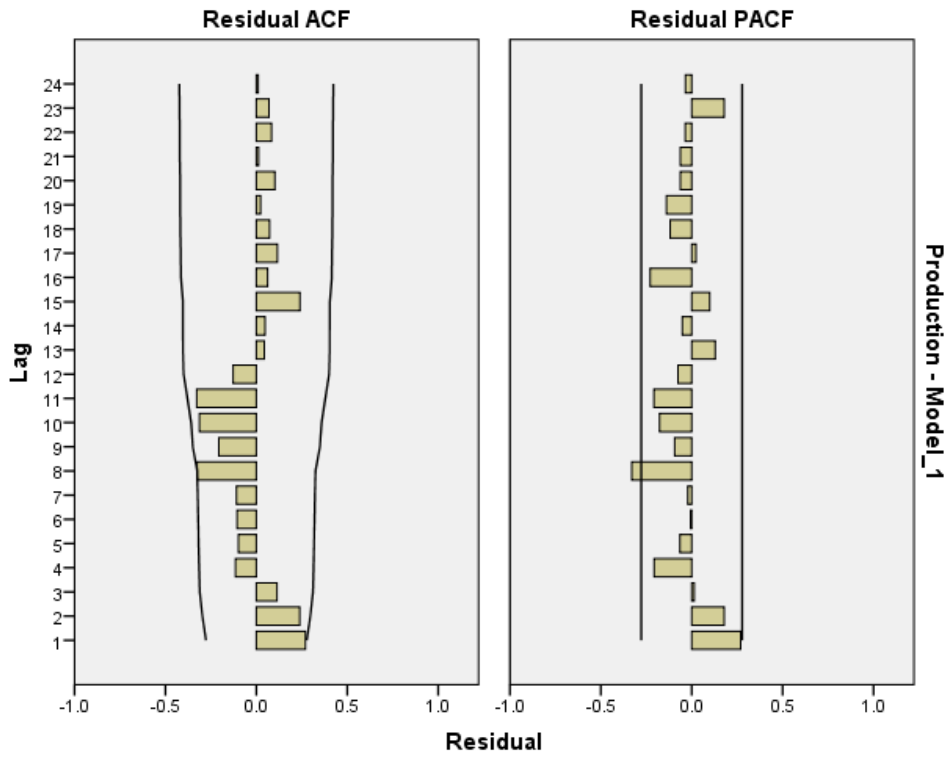

Figure 4.2 ACF and PACF graphs of residuals of oilseeds production

From Figure 4.2, it can be seen that 8th lag in the PACF graphs slightly exceeded confidence limit. Thus, results of Box-Ljung test used to find out whether the residuals have white noise.

At the next stage, the forecasting values were graphed together with observation values of the original series in Figure 4.3

After the results obtained above, oilseeds production can be forecasted. Forecasting results are given in Table 4.5. An increase in oilseeds production in the period 2020-2024 is expected.

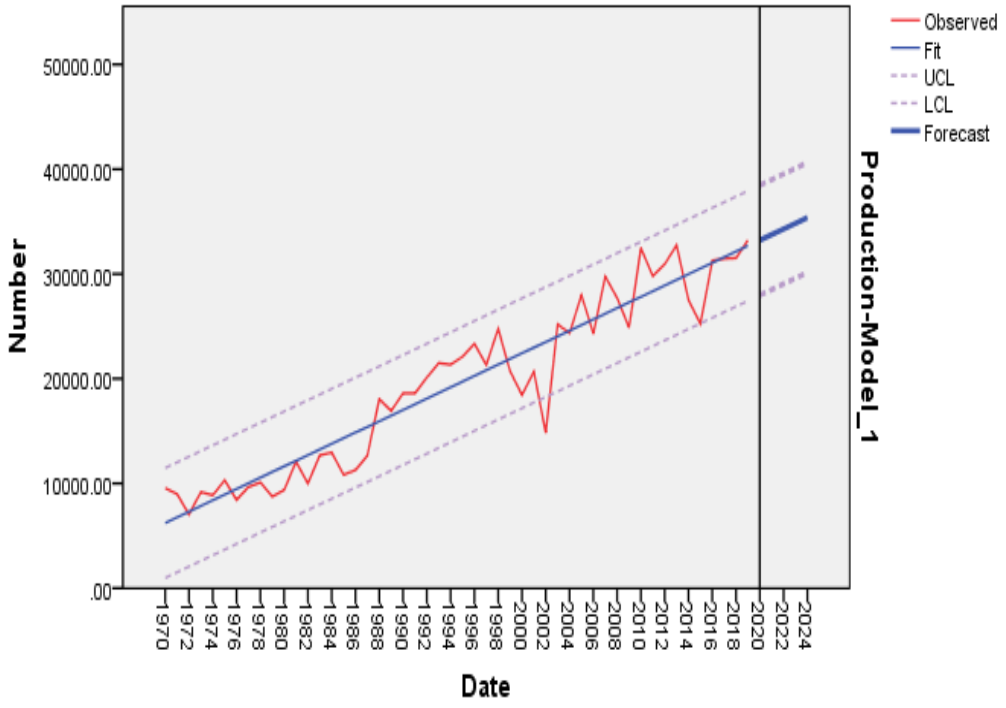

Figure 4.3 Graph of oilseeds production series and forecast values for four years by Holt's linear method

Table 4.5 Forecast

Model		2020-21	2021-22	2022-23	2023-24
Production-Model_1	Forecast	33222.30	33762.12	34301.93	34841.75
	UCL	38469.83	39009.65	39549.47	40089.29
	LCL	27974.77	28514.58	29054.40	29594.21

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

Table 4.5 shows the predicted values with 95% (low and high) prediction intervals.

2. Brown's double exponential smoothing method:

Brown's method with one parameter is use when data shows linear trend but no seasonality is present. Model fit statistics Stationary R^2, R^2, MAPE, MAE and Normalized BIC criteria were evaluated by using Brown's method which is summarized in Table 4.6

In Table 4.7, Smoothing parameters of Brown's method was estimated as α =0.226

Table 4.6 Model Fit Statistics

Model	Number of Predictors	Model Fit statistics				
		Stationary R-squared	RMSE	MAPE	MAE	Normalized BIC
Production-Model_1	0	.725	2784.734	11.075	2012.131	15.942

Model Fit Statistics

Number of Predictors	Ljung-Box Q(18)			Number of Outliers
	Statistics	DF	Sig.	
0	22.843	17	.154	0

Table 4.7 Exponential Smoothing Model Parameters

Model			Estimate	SE	t	Sig.
Production-Model_1	No Transformation	Alpha (Level and Trend)	.226	.043	5.207	.000

Thus, the equations given as

- **Forecast equation:** $\bar{y}_{t+h} = a_t + b_t h$
- **Single exponential smoothing equation**

  $y'_t = \alpha y_t + (1 - 0.226) * y'_{t-1}$
- **double exponential smoothing equation**

  $y''_t = 0.226 y'_t + (1 - 0.226) * y''_{t-1}$

  $a_t = 2y'_t - y''_t$

$$b_t = \frac{0.226}{1-0.226} + (y_t' - y_t'')$$

By Brown's method the measured autocorrelations among residuals at different lags are shown in the table 4.8 and the measured partial autocorrelations among the residuals for several lags are shown in the table 4.9.

Table 4.8 Estimated Residuals ACF for oilseeds production

Model		1	2	3	4	5	6	7	8	9	10
Production-Model_1	ACF	.048	.105	.007	-.204	-.091	-.009	.029	-.224	-.028	-.202
	SE	.141	.142	.143	.143	.149	.150	.150	.150	.157	.157

11	12	13	14	15	16	17	18	19	20	21	22	23	24
-.265	-.025	.117	.077	.306	.004	.027	-.060	-.087	.031	-.070	.033	.041	.005
.162	.170	.171	.172	.173	.183	.183	.183	.184	.185	.185	.185	.185	.186

Table 4.9 Estimated Residuals PACF for oilseeds production

Model		1	2	3	4	5	6	7	8	9	10
Production-Model_1	PACF	.048	.103	-.002	-.218	-.078	.049	.056	-.294	-.068	-.149
	SE	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141

11	12	13	14	15	16	17	18	19	20	21	22	23	24
-.257	-.121	.123	-.001	.179	-.159	.032	-.110	-.149	-.045	-.117	-.119	.186	.044
.141	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141

We will examinine autocorrelations and partial autocorrelations of the residuals of the fitted model. The ACF and PACF graphs of residuals are shown in the Figure 4.4

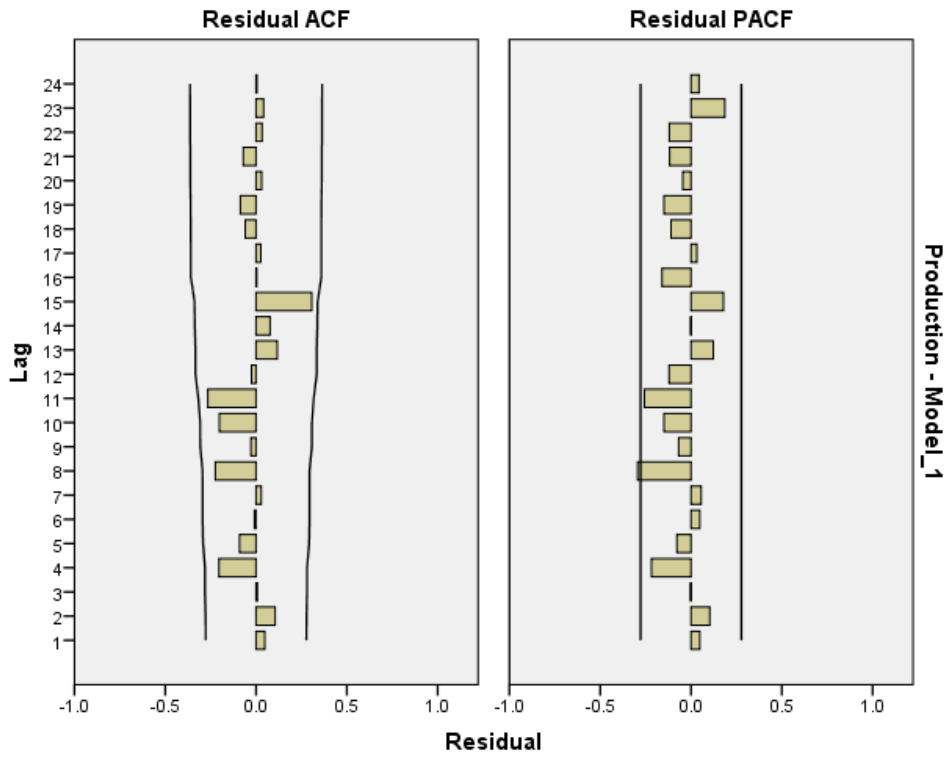

Figure 4.4 ACF and PACF graphs of residuals of oilseeds production

From Figure 4.4, we can conclude that all the ACFs and the PACFs of the residuals of the fitted model for lag 1 to 24 are within the significance limits. This means that there is no autocorrelation in the residuals of the fitted model.

At the next stage, the forecasting series were graphed together with observation values of the original series shown in Figure 4.5

After the results obtained above, oilseeds production can be forecasted. Forecasting results are given in Table 4.10

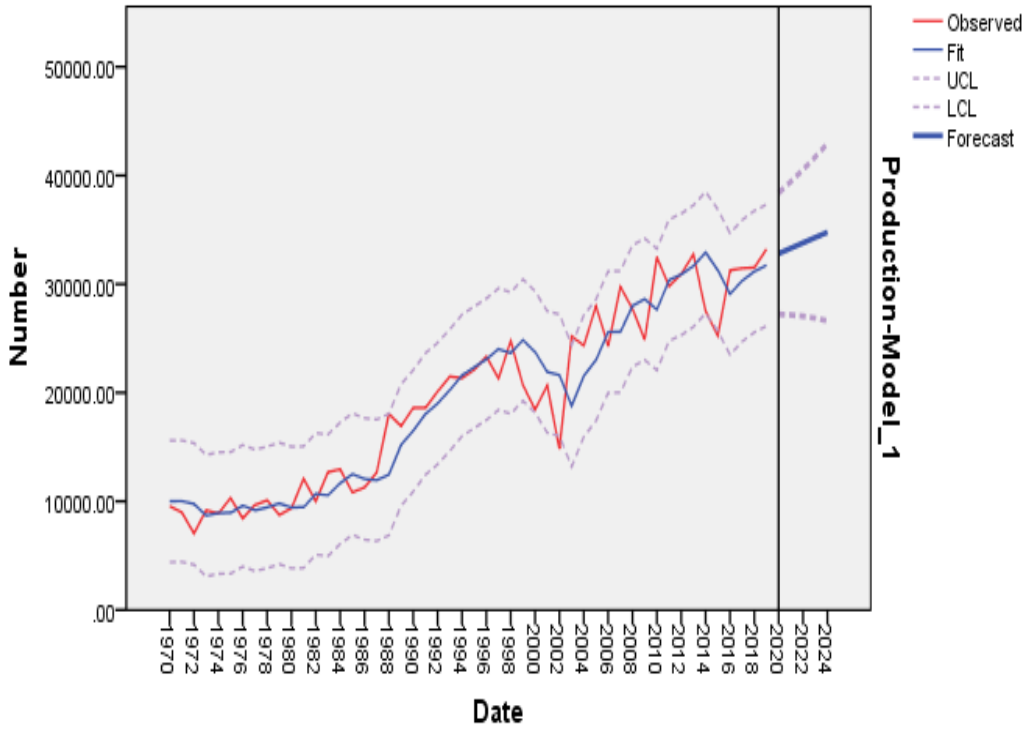

Figure 4.5 Graph of oilseeds production series and forecast values for four years by Brown's method

Table 1.10 Forecast

Model		2020-21	2021-22	2022-23	2023-24
Production-Model_1	Forecast	32814.15	33305.78	33797.41	34289.04
	UCL	38410.29	39445.33	40549.56	41716.79
	LCL	27218.02	27166.23	27045.26	26861.28

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

Table 4.10 shows the predicted values with 95% (low and high) prediction intervals.

3. Damped method:

We use the damped method because Holt's linear trend method overestimates the parameters. By using Damped method the model fit statistics Stationary R^2, R^2, MAPE, MAE and Normalized BIC criteria are evaluated which is summarized in Table 4.11.

Table 4.11 Model Fit Statistics

Model	Number of Predictors	Model Fit statistics					
		Stationary R-squared	R-squared	RMSE	MAPE	MAE	Normalized BIC
Production-Model_1	0	.276	.899	2686.074	11.164	1963.459	16.026

Number of Predictors	Ljung-Box Q(18)			Number of Outliers
	Statistics	DF	Sig.	
0	19.718	15	.183	0

In Table 4.12, Smoothing parameters of Damped method are α =0.397, γ= 0.0001 and we have set the damping parameter to a relatively low number (ø=0.999) to amplify the effect of damping for comparison.

Table 4.12 Exponential Smoothing Model Parameters

Model			Estimate	SE
Production-Model_1	No Transformation	Alpha (Level)	.397	.118
		Gamma (Trend)	.0001	.099
		Phi (Trend damping factor)	.999	.007

Thus, the equations given as

- **Forecast equation:**
  $\bar{y}_t = L_t + (0.999 + 0.999^2 + \cdots + 0.999^h)T_t$
- **Level equation:**
  L_t=0.397y_t+(1−0.397)*(L_{t-1}+0.999T_{t-1})
- **Trend equation:**
  T_t =0.0001*(L_t−L_{t-1})+(1−0.0001)*(0.999)T_{t-1}

By damped method the measured autocorrelations among residuals at different lags are shown in the table 4.13. And the measured partial autocorrelations among the residuals for several lags are shown in the table 4.14.

Table 4.13 Estimated Residuals ACF for oilseeds production

Model		1	2	3	4	5	6	7	8	9	10
Production-Model_1	ACF	-.014	.076	.011	-.202	-.074	.002	.052	-.217	-.008	-.185
	SE	.141	.141	.142	.142	.148	.149	.149	.149	.155	.155

11		12	13	14	15	16	17	18	19	20	21	22	23	24
-.253		-.028	.114	.046	.281	-.021	.035	-.029	-.072	.058	-.057	.056	.062	.018
.160		.167	.167	.169	.169	.178	.178	.179	.179	.179	.180	.180	.180	.181

Table 4.14 Estimated Residuals PACF for oilseeds production

Model		1	2	3	4	5	6	7	8	9	10
Production-Model_1	PACF	-.014	.076	.014	-.208	-.086	.034	.077	-.275	-.072	-.156
	SE	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141

11	12	13	14	15	16	17	18	19	20	21	22	23	24
-.257	-.159	.091	-.027	.172	-.158	.051	-.077	-.131	-.058	-.100	-.101	.189	.053
.141	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141	.141

We will examining autocorrelations and partial autocorrelations of the residuals of the fitted model. The ACF and PACF graphs of residuals are shown in the Figure 4.6.

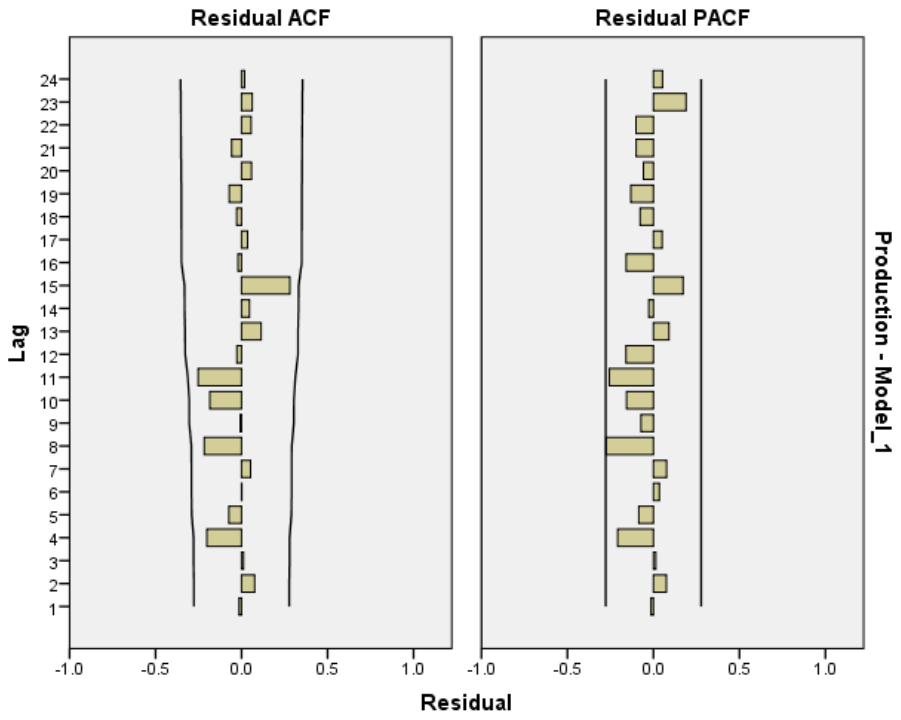

Figure 4.6 ACF and PACF graphs of residuals of oilseeds production

From Figure 4.6, we can conclude that all the ACFs and the PACFs of the residuals of the fitted model for lag 1 to 24 are within the significance limits. This means that there is no autocorrelation in the residuals of the fitted model.

At the next stage, in Figure 4.7 the forecasting series were graphed together with observation values of the original series.

After the results obtained above, oilseeds production can be forecasted. Forecasting results are given in Table 4.15.

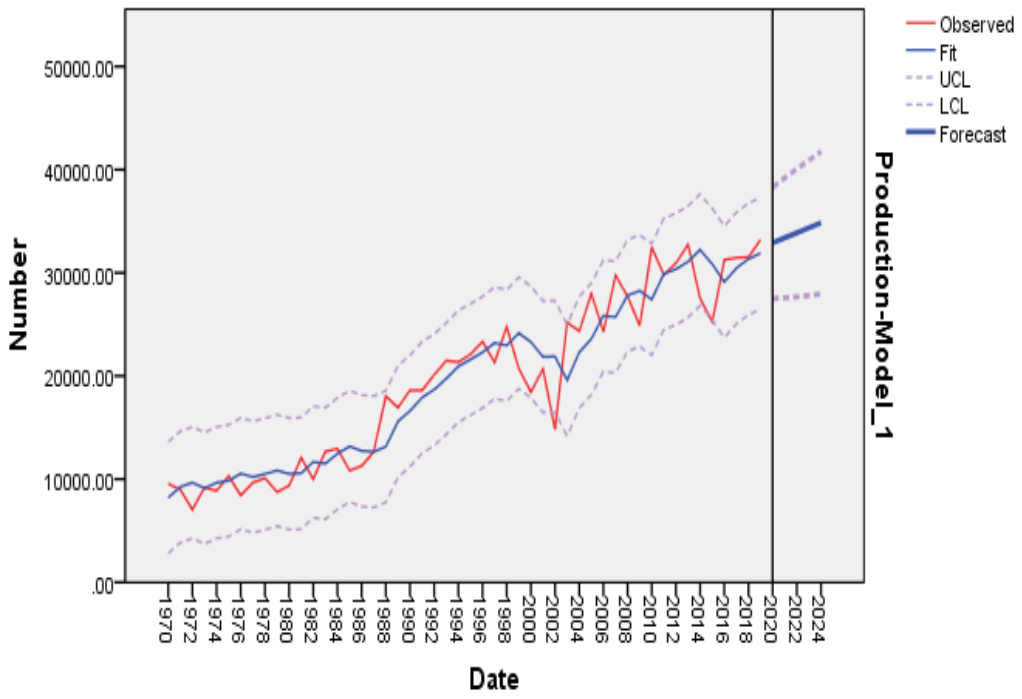

Figure 4.7 Graph of oilseeds production series and forecast values for four years by Damped method

Table 4.15 Forecast

Model		2020-21	2021-22	2022-23	2023-24
Production-Model_1	Forecast	32908.33	33393.35	33877.86	34361.87
	UCL	38312.02	39207.05	40074.69	40919.65
	LCL	27504.65	27579.64	27681.02	27804.08

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

Table 4.15 shows the predicted values with 95% (low and high) prediction intervals.

After this in table 4.16, the accuracy of the models: Holt, Brown and Damped Trend double exponential smoothing methods using Stationary R-squared, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Percentage

Error (MAE), Mean Absolute Percentage Error (MAPE) and normalized BIC; refers to as the magnitude of the error rate of an estimate. Smaller the value of these test statistics, the better the forecasts.

Model Fit statistics	Number of Predictors	Normalized BIC
Holt's linear method α =0.001, γ=0.0001	0	15.891
Brown's method α =0.226	0	15.942
Damped method α =0.397, γ= 0.0001, ø=0.999	0	16.026

In the statistical comparison of the models, it is meaningful to use statistics like Normalized BIC. From Table 4.16, Normalized BIC for holt's model is 15.891for brown's model is 15.942 whereas for damped model it is 16.026 respectively. It is well-understood that Holt smoothing method that yielded the lowest BIC value was the best method. Coefficients of Holt smoothing method were estimated as α =0.001 and γ=0.0001.

4.2 Autoregressive Integrated Moving Average (ARIMA) model

The figure 4.8 below represents the time series plot of oilseeds production in India.

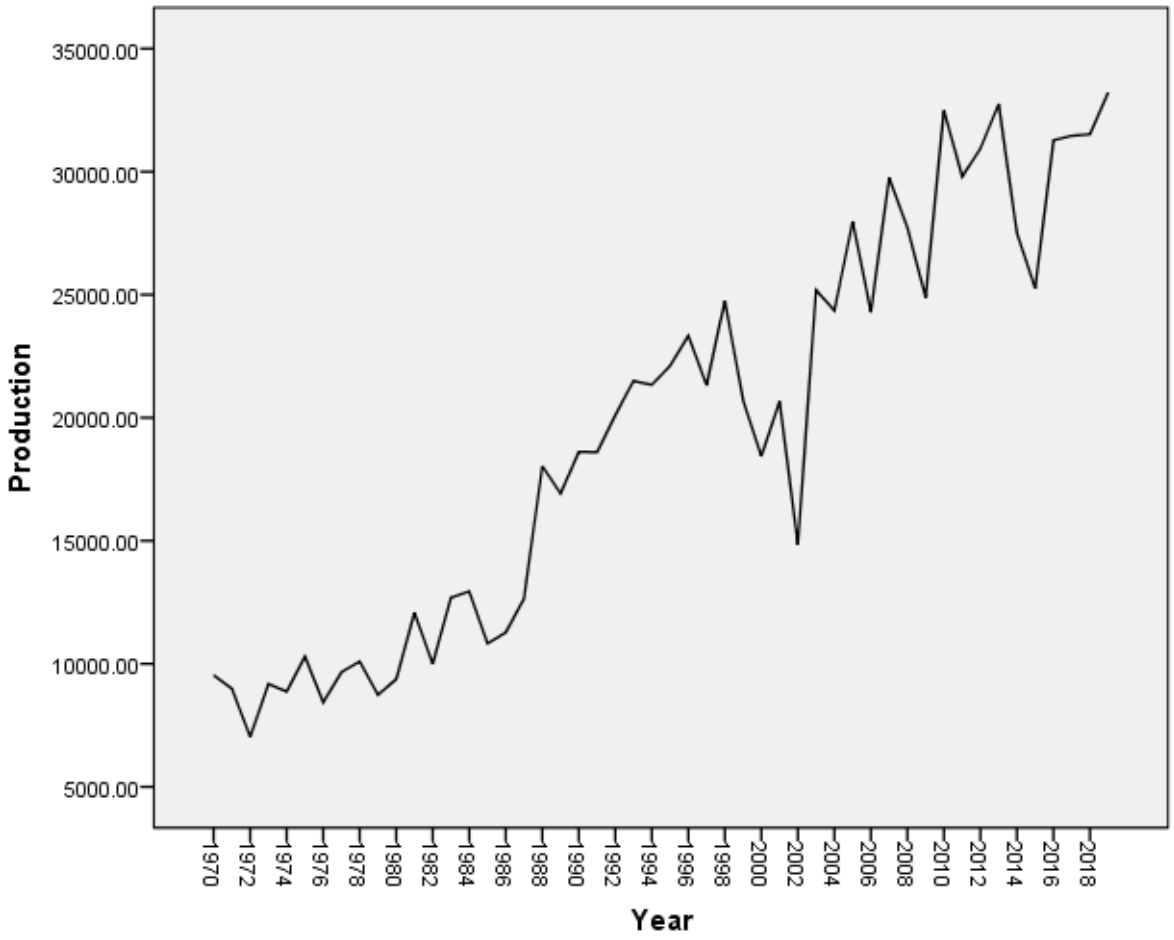

Figure 4.8 Original time series plot of oilseeds production in India

Since, we have already discussed that to build an ARIMA model for forecasting of a variable requires following steps:

1. Model Identification
2. Model selection and parameter estimation
3. Diagnostic Checking
4. Forecasting

1. Model Identification

One of the main assumption for ARIMA model is stationarity of the series. By stationarity we mean, the properties of time series should not change over time. In other words, the mean and variance of variable should not change over time. It is observed from the time series plot (Figure 4.8) that the given times series of oilseeds production in India is non-stationary. Moreover, the time series is showing an increasing trend.

The general procedure to convert a non-stationary series to a stationary series is through first difference or second difference. The below figure (Figure 4.9) is the time series plot of the first order differenced oilseeds production in India.

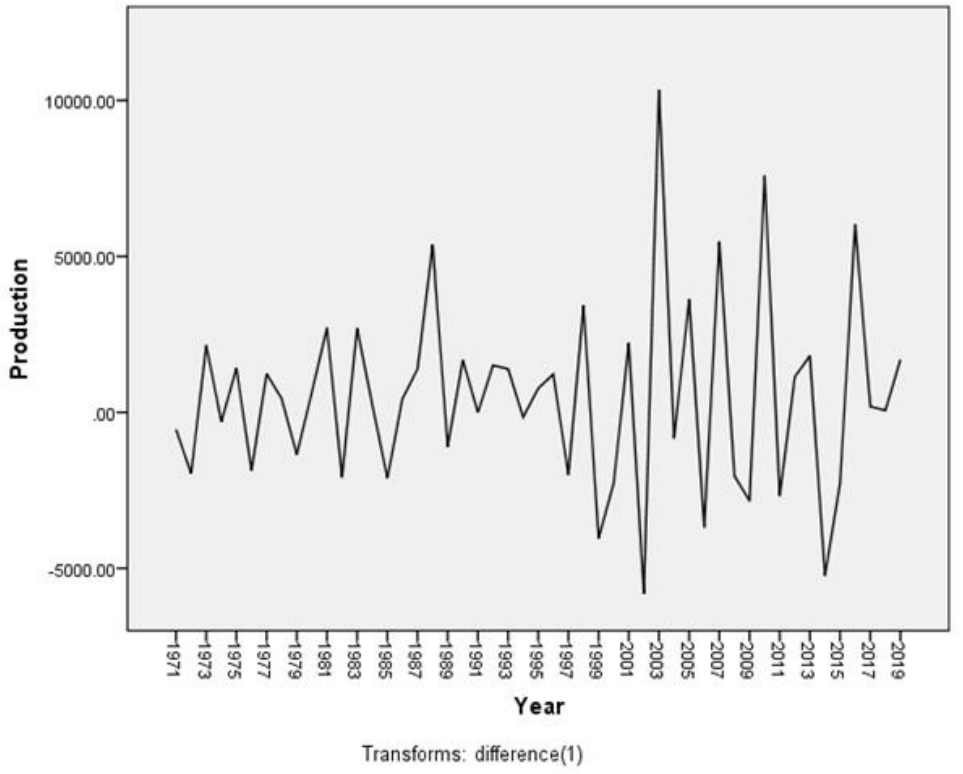

Figure 4.9 Time series plot of differenced oilseeds production data of first order

It can be easily observed from the above graph (Figure 4.9) that the time series appeared to be stationary both in mean and variance. Before moving further, we perform Augmented Dickey-Fuller test in order to check the stationarity of our data.

Test of stationarity: Augmented Dickey-Fuller test (ADF test)

To test the hypothesis

H_0: The given time series is not stationary

Vs

H_1: The given time series is not stationary

To test the above hypothesis, the first order differenced data was used. The result of ADF test is shown below:

Dickey-Fuller = -4.2196, Lag order = 3, p-value = 0.01

We, therefore fail to accept H_0 and thus we conclude that alternative hypothesis is true at 5% level of significance. Hence, the series is stationary at the first order difference (d=1).

After a time series has been stationarized by differencing, the next step in fitting an ARIMA model is to determine whether AR or MA terms are needed to correct any autocorrelation that remains in the differenced series. This leads us to select the values of p in AR and q in MA for our model. For this we make use of correlogram and partial correlogram of the first order differenced time series.

Correlogram and Partial Correlogram

The Figure 4.10 below represents the plot of correlogram (Auto correlation function, ACF) of the first order differenced time series.

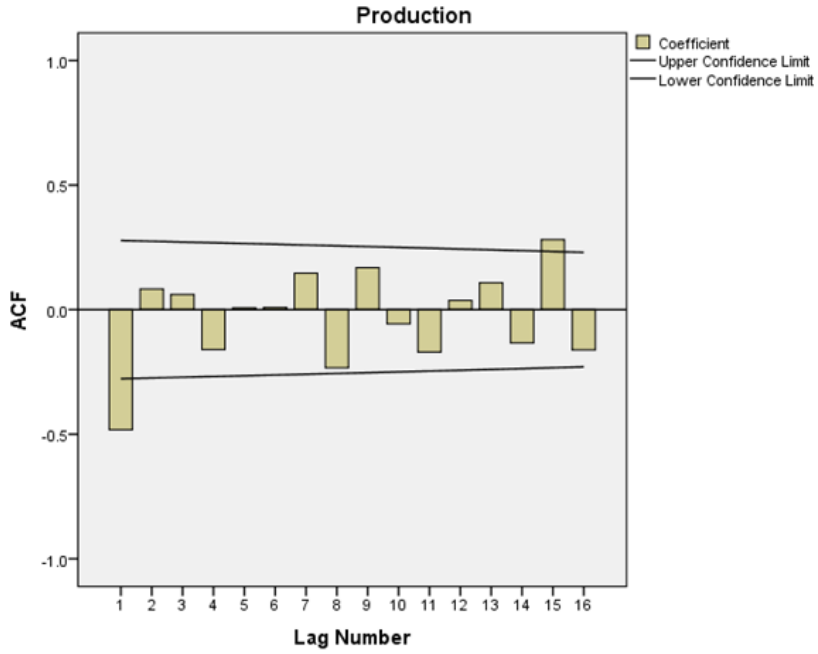

Figure 4.10 Autocorrelation (ACF) of first differenced series by lag

It can be inferred from the above correlogram that the auto-correlation at lag 1 exceeds the significance limits. Autocorrelations tail off to zero after lag 1. Although, autocorrelation at lag 15 just exceeds the upper confidence limit, rest all coefficients are within the significance limits. We can assume that lag 15 autocorrelation is an error.

Figure 4.11 below represents the plot of partial correlogram (Partial Autocorrelation Function, PACF) of the first order differenced time series.

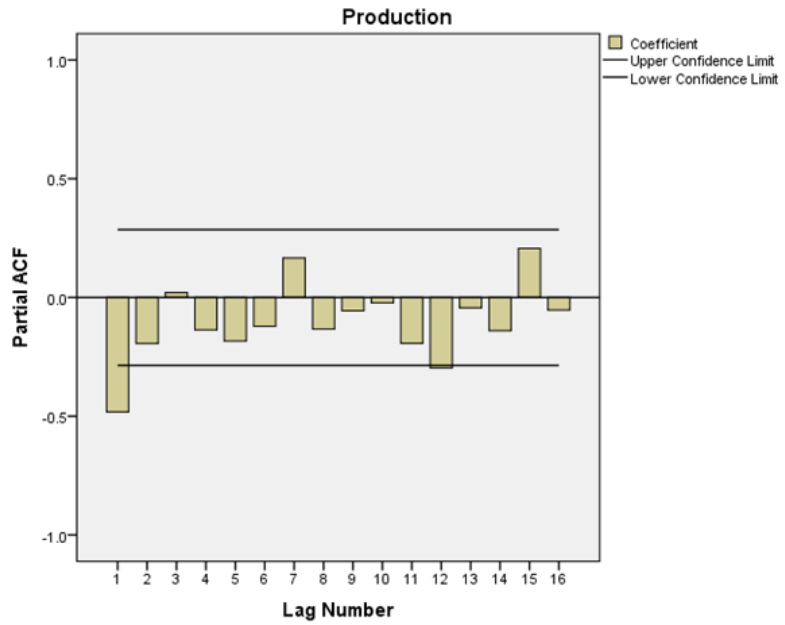

Figure 4.11 Partial Autocorrelation (PACF) of first differenced series by lag

The above partial correlogram indicates that the partial autocorrelation function exceeds the significant limits at lag 1. After lag 1 all the partial autocorrelation coefficients are within the significant limits.

Since the correlogram (ACF) tails off to zero after lag 1(omitting the outlier) and the partial correlogram (PACF) also tails off to zero after lag 1. Therefore, we select ARIMA(0,1,1), ARIMA(1,1,1) and ARIMA(1,1,0).

2. Model selection and parameter estimation

To select the best possible model for forecasting we select the model with lowest value of BIC (Bayesian Information Criterion).

The ARIMA (0,1,1) model is considered for forecasting oilseeds production, as the normalized BIC value for ARIMA (0,1,1) is 15.952 which is smaller as compared to other ARIMA (1,1,1) model having normalized BIC value 15.982 and ARIMA (1,1,0) model having normalized BIC value 16.003.

Table 4.17 Model statistics

Model	Number of Predictors	Model Fit statistics				
		Stationary R - squared	RMSE	MAPE	MAE	Normalized BIC
Production-Model_1	0	.271	2688.406	11.367	1998.079	15.952

Table 4.18 ARIMA model parameters

				Estimate	SE
Production-Model_1	Production	No Transformation	Constant	494.187	153.968
			Difference	1	
			MA lag 1	.612	.118

The model is thus given as

$$y_t = 494.187 + Y_{t-1} + 0.612\, e_{t-1}$$

Table 4.19 Model Statistics

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
Production-Model_1	0	.271	19.933	17	.278	0

3. Diagnostic Checking

Diagnostic checks have become a standard tool for identification of models before forecasting the data. This is done through examining autocorrelations and partial autocorrelations of the residuals of the fitted model. The ACF and PACF plots of residuals are shown in the Figure 4.12.

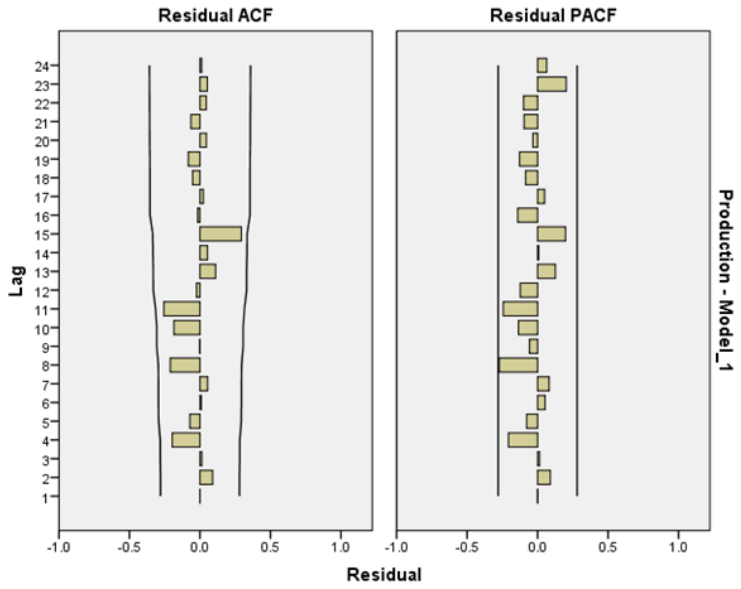

Figure 4.12 Graph showing ACF and PACF of residuals of fitted model

Since, all the ACFs and the PACFs or partial autocorrelation coefficients of the residuals of the fitted ARIMA model for lag 1 to 24 are within the significance limits. This means that there is no autocorrelation in the residuals of the fitted model.

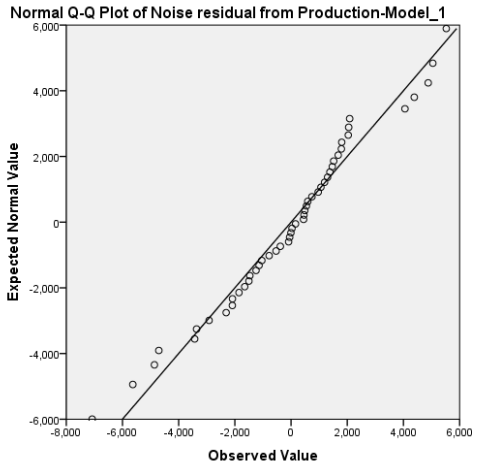

Figure 4.13 Q-Q plot of residuals of the fitted model

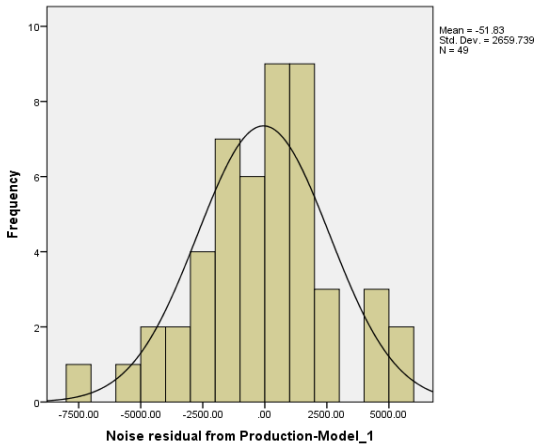

Figure 4.14 Histogram of residuals of the fitted model

The careful investigation from the above plots infers that the residuals of the fitted model are normally distributed.

Ljung-Box test: To test the hypothesis

H_0: The residuals are independently distributed or the model does not show lack of fit

H_1: The residuals are serially correlated or the model show lack of fit

The results of Ljung-Box test are shown in a below table.

Table 4.20 Results of Ljung-Box test

			Box-Ljung Statistic		
Lag	Autocorrelation	Std. Error[a]	Value	df	Sig.[b]
1	.001	.139	.000	1	.995
2	.090	.137	.431	2	.806
3	.014	.136	.441	3	.932
4	-.197	.134	2.587	4	.629
5	-.071	.133	2.876	5	.719
6	.010	.131	2.882	6	.824
7	.054	.130	3.057	7	.880
8	-.210	.128	5.751	8	.675
9	-.001	.127	5.751	9	.765
10	-.184	.125	7.925	10	.636
11	-.256	.123	12.232	11	.346
12	-.024	.122	12.271	12	.424
13	.110	.120	13.116	13	.439
14	.054	.118	13.321	14	.501
15	.293	.117	19.647	15	.186
16	-.017	.115	19.668	16	.236

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

The above result shows that none of the autocorrelations are significantly different from zero. Thus, we conclude that the model is good fit.

4. Forecasting

We now fit the chosen ARIMA (0,0,1) model to forecast for the future values of our time series. Following Table 4.21 shows the forecast for the next four years with 95% (low and high) prediction intervals.

Table 4.21 Forecast

Model		2020	2021	2022	2023
Production-Model_1	Forecast	33222.30	33762.12	34301.93	34841.75
	UCL	38469.83	39009.65	39549.47	40089.29
	LCL	27974.77	28514.58	29054.40	29594.21

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

Figure 4.14 below shows the plot for four years forecast of the oilseeds production by fitting ARIMA (0,1,1) model to our time series data:

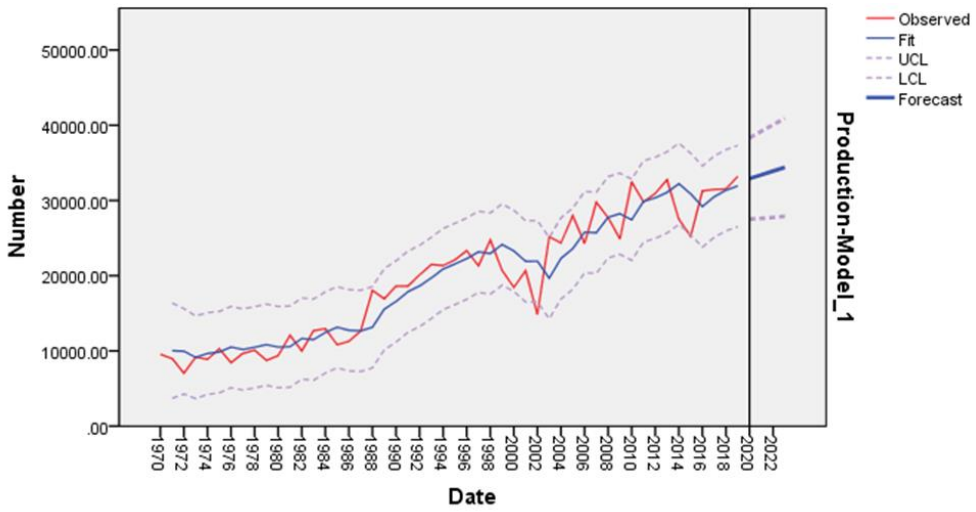

Figure 4.14: Forecast fitted with ARIMA (0,1,1)

The Figure 4.14 above shows the fitted ARIMA (0,1,1) along with upper and lower control limit of forecast.

4.3 Group Method of Data Handling (GMDH)

In this section we analyse the short-term forecasting results of oilseeds production through GMDH - neural network algorithms by using GMDH Shell software. GMDH-neural network selects the model of optimal complexity and such a selection depends on the form of external criterion realization. K-fold cross validation is one of such criteria. In our study, we used this k fold validation method. In this validation, original sample was randomly partitioned into k subsamples. A single subsample was taken as the validation data for testing model, and the other k – 1 sub-samples were used as training data. The cross-validation process was repeated k times using each of the k subsamples exactly once. The value of k obtained from the K folds can produce a single estimation.

The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. The experiment was carried out using RMSE validation criterion. Therefore, the optimal time series forecasting model was selected by minimum value of RMSE, calculated for the testing sample. In our time series analysis under GMDH-neural network model, based on k- cross validation criterion, our forecasting model is an optimal with k=2.

Accuracy of model shows different accuracy metrics for the model selected in the model browser. Model contains accuracy measures calculated for observations used to create the model. Error measure is used to choose a metric for calculation of the mean and the root mean errors. Available metrics are the absolute (MAE and RMSE), which outputs mean error values “as is” and the target percentage (MAPE), where for each model value we calculate percentage deviation from the actual value and then the percentage deviations are averaged. The model statistics of GMDH - neural network are presented in Table 4.22.

Calculation of magnitude of predicted variable involves only the observations that are used for training and testing. The forecasting values are presented in Table 4.23. In our study, GMDH - neural networks model forecasting oilseeds production will be 33778 thousand tonnes, 33271 thousand tonnes, 34858 thousand tonnes, 36084 thousand tonnes and 35751 thousand tonnes in 2020-2021, 2021-2022, 2022-23, 2023-24, respectively.

The diagrammatic presentation of forecast value of oilseeds production under GMDH- neural network has been shown in Figure 4.15. In this diagram the actual value is depicted by black line and the fitted value is shown in blue. The red line indicates the forecast value of oilseeds production whereas the confidence band has been presented by the shaded area. The time is measured along the horizontal axis and the vertical axis measures the level of production (thousand tonnes).

Table 4.22 Model Statistics

Sr.No.	Variable	Model	RMSE	MAE	R^2	MAPE
1	Production (000 tonnes)	GMDH	2414.89	1831.71	0.847608	9.3020

Table 4.23 Forecast

Variable	Year	Forecast	Upper Value	Lower Value
Production	2020	33778	28948.2109	38607.7890
	2021	33271	28441.2109	38100.7890
	2022	34858	30028.2109	39687.7890
	2023	36084	31254.2109	40913.7890

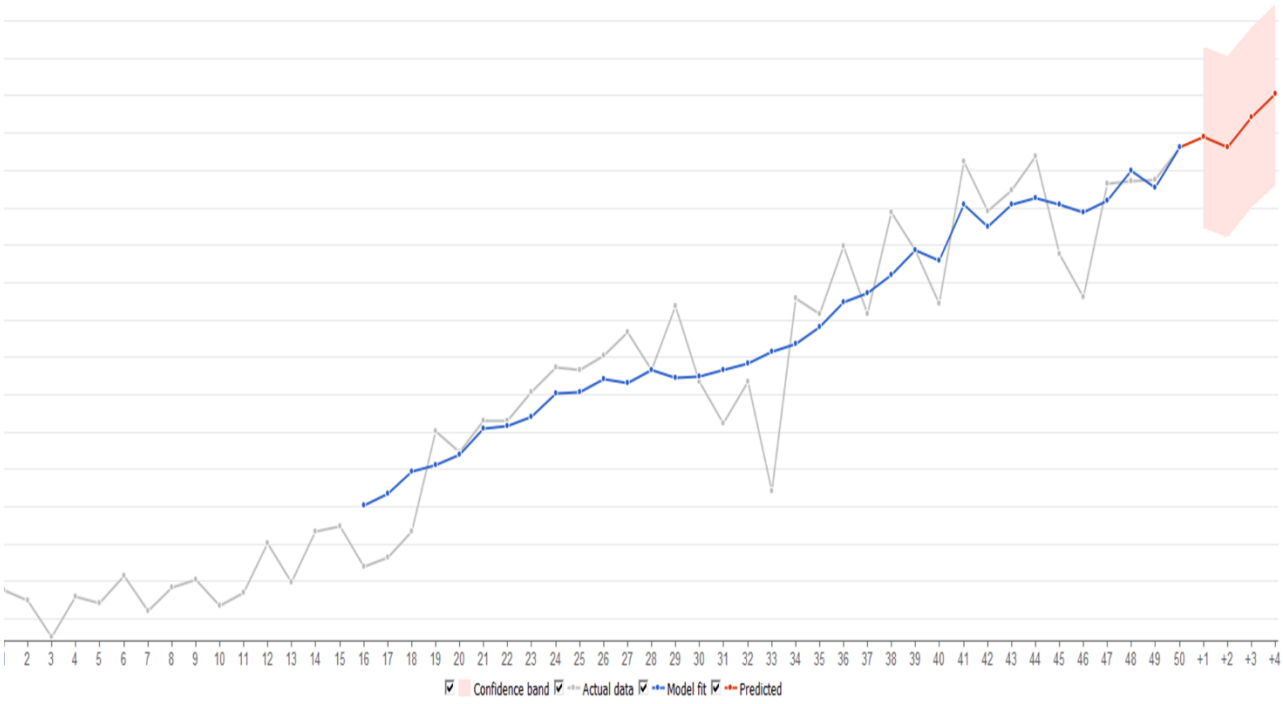

Figure 4.15 Actual value, fitted value and forecast value and confidence band in GMDH model

4.4 Comparison

Now the question that arises is which model is better and appropriate for forecasting the oilseeds production. To find the solution, we compare the model statistics of Exponential Smoothing Method, ARIMA and GMDH-neural network in terms of RMSE, MAE. Model with lower values of RMSE as compare to the other model, is better. The model statistics of Exponential Smoothing Method, ARIMA and GMDH-neural network are presented in Table 4.24. The table indicates that GMDH-neural network is performing better than Exponential Smoothing Method and ARIMA in all respect.

Table 4.24 MODEL COMPARISON:

Sr.No.	Variable	Model	RMSE	MAPE	R^2	MAE
1	Production	GMDH	2414.89	9.3020	0.847	1831.71
2		ARIMA(0,1,1)	2688.406	11.367	0.896	1998.079
3		HOLT	2609.888	12.168	0.903	2008.373

Chapter 5

CONCLUSION AND LIMITATIONS

5.1 Conclusion:

In this study, we have analyzed the data of oilseeds production in India for the last 50 years i.e. from 1970-1971 to 2018-2019. Time series analysis has been performed to analyze the data. An effort has been made to forecast the production of oilseeds by using Exponential Smoothing Method, ARIMA model and GMDH-Neural Network model.

The results obtained by all the three methods were compared. The comparison of the modelling results shows that the GMDH-neural network model performed well as compared to ARIMA model and Exponential Smoothing Method.

The experimental results shows that GMDH model is a powerful tool in modelling and forecasting of time series can increase the forecasting accuracy.

The results of forecasting in GMDH model reveals that India's oilseeds production will be 33778 thousand tones in 2020-21. It will decline to 33271 thousand tones in 2021-22 and thereafter it will slightly increase to 34858 thousand tones in 2022-23, 36084 thousand tones in 2023-24.

5.2 Limitations

- This study will help State & Central Government to improve policies and several development programs.
- The study is based on oilseeds production in India, this study will help to get idea of future oilseeds production.
- This study will take in account for calculations of import of oilseeds for next 4 years.

Chapter 6

BIBILOGRAPHY

- Abraham E R, Vendrametto o and et al. (2020), Time Series Prediction with Artificial Neural Networks: An Analysis Using Brazilian Soybean Production, Agriculture, p.p. 1-18.
- Akkaya (2021), GMDH-type neutral-based monthly electricity demand forecasting of Turkey, ISSN:2618-575X,Vol. 5, Issue 1.
- Bakar A N and Rosbi S (2016), Reliability of Exponential Smoothing Method for forecasting Islamic Share Price to oil and gas sector in Malaysian Stock Exchange, International Academic ResearchJournal of Business and Technology,vol. 2,issue 2 ,Page 38-44.
- Dag O, Yozgatligil C (2016), GMDH: An R Package for short Term Forecasting via GMDH-Type Neural Network Algorithms, The R Journal Vol 8/1, ISSN:2073-4859, p.p.379-386.
- Dash A, Mahapatra S(2020), Using ARIMA model for yield Forecasting of important pulse India, Amazonian Journal of Plant Research, Vol. 4 (3), p.p. 646-659.
- Dharmawan and Indradewi (2020), Double exponential smoothing brown method towards sales forecasting system with a linear and non-stationary data trend, Journal of Physics: Conference Series 1810 012026
- Dhyani B, Kumar M, Verma P, Jain A (2020), Stock Market Forecasting Technique using Arima Model, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878 (Online), Volume-8 Issue-6, P.P. 2694-2697.
- Fattah J, Ezzine L and et al.(2018), Forecasting of demand using ARIMA model, International Journal of Engineering Business Management, Volume 10,p.p 1–9
- Ghosh S (2017), Forecasting Cotton Exports in India using the ARIMA model, Amity Journal of Economics, vol. 2, issue 2, p.p. 36-52.
- Jahring, Pradani (2016), Forecasting with Double Exponential Smoothing Brown Method, ISSN: 2541-1756 Vol. 1, No. 2, p.p. 35-39.
- Kumar A., Deeepankar, P.K Jaslam Muhammed (2019), Wheat Yield Forecasting in Haryana: A Time Series Approach, Bull Evn. Pharmacol. Life Sci., ISSN 2277-1808, Vol. 8 [3], p.p. 63-69.
- Karadas K, Celik S, Eydurun E and Hopoglu S (2017), forecasting production of some oilseeds crops in turkey using exponential

smoothing methods, the journal of animal and plant sciences, ISSN:1018:7081, vol.5,p.p.1719-1729.

- Mithiya D, Datta L, Mandal K (2019), Time series analysis and forecasting of oilseeds production in India: using autoregressive moving average and group method of data handling-neural network, Asian journal of agricultural extension, economics and sociology, ISSN:2320-7027, Vol. 2, p.p.1-14.
- Oni and Akanle (2018), Comparison of Exponential Smoothing Models for Forecasting Cassava Production, International Journal of Scientific Research in Mathematical and Statistical Sciences, E-ISSN: 2348-4519, Volume-5, Issue-3, pp.65-68.
- Priya S, Bajpal P, Suresh K (2015), Stochastic Model For Sugarcane Yield Forecasting, Indian Journal Of Sugarcane Technology, Vol. 30(01), p.p. 1-5.
- Rathod S, Singh N K and Roy M (2018), Modeling and forecasting of oilseed production of India through artificial intelligence techniques, Indian Journal of Agricultural Sciences, p.p 22–27.
- Ravindra H(2013), Forecasting With Exponential Smoothing- What's The Right Smoothing Constant?, Vol 17, Number 3
- Saha A and Sinha K (2020), Usage of Holt's Linear Trend Exponential Smoothing for Time Series Forecasting in Agricultural Research , ISSN 2582-5437, Vol. 1,p.p.9-11.
- Satya Pal, Ramasubramanian v and s.c. mehta (2007), statistical models for forecasting milk production in India, journal of the indian society of agricultural statistics, vol.2, p.p. 80-83.
- Singh A (2015), Forecasting of Prices of Groundnut (Arachis hypogea) Oil using Time Series Models and Artificial Neural Networks, Research Journal of Agricultural Sciences, p.p. 1600-1604.
- Shabri A, Samsudin R (2014), Hybrid GMDH and Box-Jenkins Model in Time Series Forecasting, Vol. 8, pp-3051-3062.
- Shastri S, Sharm A , Mansotra V, Sharma A , Bhadwal A , Kumari M (2018), A Study on Exponential Smoothing Method for Forecasting International Journal of Computer Sciences and Engineering Vol.6(4), E-ISSN: 2347-2693.
- Shabri A, Samsudin R, and Ismail Z (2009), Forecasting of the Rice yields Time series forecasting using Artificial Neural Network and

Statistical Model, Journal of Applied Sciences, vol. 9, no.23, p.p 4168-4173.

- Tran T.T., Pham L. T., Ngo Q. X., Forecasting epidemic spread of SARS-CoV-2 using ARIMA (Case Study: Iran) (2020),GJIESM.
- Vijay N, Mishra C G (2018), Time Series Forecasting Using ARIMA and ANN Models for Production of Pearl Millet (BAJRA) Crop of Karnataka, India, International Journal of Current Microbiology and Applied Sciences, ISSN: 2319-7706, Volume 7, Number 12, p.p 880-889.

- https://in.investing.com/equities/tata-consultancy-services-historical-data?end_date=1652545376&st_date=1640995200
- https://www.javatpoint.com/artificial-neural-network
- https://www.ibm.com/in-en/cloud/learn/neural-networks
- 331395382_Time_Series_Analysis_and_Forecasting_of_Oilseeds_Production_in_India_Using_Autoregressive_Integrated_Moving_Average_and_Group_Method_of_Data_Handling_