TuijinJishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 3 (2024)

Research Article on:WushWush Tea Plantation Forecasting Sales Using the Box-Jenkins Approach

[1]Dr.Shimels Zewdie Werke (Associate professor)

[2]Dagim Dessalegn Malleto (PHD Candidate)

Jimma University College of Business and Economics Department of Management (Specialization in Marketing Management)

Abstract

This study is conducted to facilitate the decision making process of Wushwush Tea Planation Factory by developing an appropriate forecasting model. It aims at providing accurate sales forecasts for future sales, which is a vital input in decision-making. Particularly, the focus is given for forecasting monthly Tea sold by the factory. Towards achieving its objective the study considered the Box-Jenkins approach to time series analysis. A total of 60 monthly sales data has been taken for analysis or model building purpose. Moreover, additional 5 months sales data has been used for forecasting purpose. The analysis of the data, which is carried out using S-Plus 2000 package, suggested that ARIMA (3,1,0) model represents the pattern of monthly Tea sales data. According to this model, forecasting current sales essentially requires the inclusion or consideration of the previous four consecutive sales data occurring at the four successive lags. Moreover, it is found that the sales data recorded in the first lag has greater influence or contribution in forecasting current sales volume. On the other hand, it is observed that the sales data involves a seasonal component that turns out to affect the sales volume approximately in 3.5 months. In other words, the analysis indicated that there is a seasonal component that occurs with 3 or 4 months periodicity. This in turn resulted in attaching high importance to the third and fourth lag coefficients as compared to the coefficient of the third lag. The results obtained led to the conclusion that the time factor is the major but not the only relevant factor in forecasting sales. Other considerations in relation to promotional activities, competitors' action, seasonal factors, etc should be kept in mind.

Key Words: Forecasting, Sales, WushWush, Box-Jenkins approach, Auto Regressive

Introduction

Background of the study

Tea (Assam kind)was first brought to Ethiopia in 1927, Grown in the Oromia Region, Ilu Aba Bora Zone, Alle District, close to Gore Town, Commercial tea production in Ethiopia started in 1989 with the establishment of the Wushwush and Gumero tea estates, and the tea business swiftly expanded throughout the nation. Large-scale investors own Ethiopian tea plantations, with the exception of emerging out-growers surrounding tea firms. Ethiopia grows tea in a monoculture devoid of shade trees, unlike other countries like Assam, Ceylon, and Indonesia. The crop thrives in regions with five hours of sunlight each day, mean air temperatures of 18 to 20°C, average humidity of 70 to 90%, and at least 1,500 mm of annual rainfall.

The WushWush Tea Plantation, renowned for its unique, high-quality teas, plays a significant role in the specialty tea market. As consumer interest in premium teas continues to grow, accurate sales forecasting becomes essential for meeting market demand, managing production, and optimizing resource allocation. Forecasting sales for the plantation involves not only analyzing historical sales data but also accounting for various factors like seasonality, economic conditions, and changing consumer preferences. Given the intricate nature of tea cultivation and its sensitivity to external factors, such as weather and supply chain dynamics,

precise forecasting enables WushWush to maintain competitive edge, align production schedules, and ensure a steady revenue stream.

In one case, marketing (sales) managers need sales forecasting data to plan ahead and make well-informed decisions regarding potential circumstances. The amount of expected sales determines a lot of other resource allocation choices that managers have to make. In addition to providing inputs for financial, marketing, and personnel planning, a company's sales forecasts drive its production, capacity, and scheduling systems. (Lester Massingham and Geoff Lancaster, 2017)

Despite being crucial to effective company planning, sales forecasting seems to be underutilised, especially by businesses in developing countries. Kracmar claims that traditionally, the most underdeveloped economic sectors in developing countries have been marketing and marketing research (Charls W. Chase Jr., 2018). The situation is much more dire when it comes to the growth and availability of forecasting research in these countries.

In today's intensely competitive corporate environment, executives (managers) are expected to make appropriate and prudent decisions about the future. This is basically because bad decisions made by an organisation will let its competitors win. The situation containing uncertainty is the most difficult of the several situations where decisions must be made since it requires forecasting the most likely future event.

It goes without saying that future sales are not known with precision, which makes decision-making and planning difficult. Making precise sales projections is one way to lessen the problem's severity. However, in the history of business in general and marketing in particular, getting sales predictions for the purpose of choosing an appropriate marketing strategy on its own is becoming a subject of greater significance than ever before. This is mostly because the corporate climate is becoming more and more dynamic. Today's corporate climate is distinguished by its rapid transformation.

There is no established method or technique for sales forecasting at the WushWush tea plantation. As a result, there are no sales projections that are sufficiently accurate, which is crucial for creating marketing strategies. As a result, the management's decision-making process is significantly impacted. Making decisions on future activities in relation to the need for raw materials, operations financing, human resources, etc., typically presents challenges for management.

In addition to that, WushWush Tea Plantation, a prominent tea production company, has experienced fluctuating sales in recent years, influenced by seasonal patterns, market demand variations, and economic factors. Accurate sales forecasting is essential for the plantation to manage inventory, optimize production schedules, and make informed strategic decisions. However, existing forecasting methods have not provided the precision needed for effective planning and resource allocation, leading to periods of overproduction or underproduction, ultimately impacting profitability.

This study seeks to develop and validate a Box-Jenkins model tailored to the sales data of WushWush Tea Plantation, addressing the problem of inaccurate forecasting and enhancing the plantation's decision-making capabilities.

Objective of the study

The core objective of this research is to develop an appropriate time series model specifically tailored for forecasting the monthly sales of tea produced by WushWush Tea Plantation. The model development will involve a systematic process of model identification, parameter estimation, and diagnostic checking using the Box-Jenkins methodology, aiming to establish a robust and precise forecasting tool.

Significance of the Study

This study on forecasting sales for WushWush Tea Plantation using the Box-Jenkins approach is significant for several reasons: Enhanced Forecasting Accuracy, Strategic Decision-Making, contribution to Industry Knowledge, Seasonal and Trend Analysis,Foundation for Future Research.In summary, this study is valuable

not only for WushWush Tea Plantation's operational efficiency but also for the agricultural industry, contributing practical knowledge on leveraging time series models to forecast sales accurately.

Delamination of the Research

The scope of the research is limited in terms of two aspects. First, it is only the time factor that is considered as an influential factor that determines future sales. Another reason is that, it is assumed that future sales can be adequately forecasted depending only on historical data of sales (time factor).

Data Collection and Preliminary Analysis

Data Collection: Collect the time series data, ensuring that it is evenly spaced over time. For this research, monthly sales data from WushWush Tea Plantation spanning 60 months is used, with the first 50 months dedicated to model development and the remaining 10 months reserved for validation. Exploratory Data Analysis (EDA): Conduct preliminary analysis to understand the general characteristics of the data, such as trends, seasonality, and cyclical patterns. Visualization tools like line plots, autocorrelation plots (ACF), and partial autocorrelation plots (PACF) help reveal these characteristics.

Parameter Estimation

Model Fitting: Fit the tentative ARIMA model to the data based on the identified parameters (p, d, q) from the previous step. Use statistical software to estimate the parameters through maximum likelihood estimation (MLE) or least squares estimation. Parameter Optimization: Adjust the model parameters to minimize the error (e.g., residual sum of squares or AIC, Akaike Information Criterion) and ensure the model fits well to the data. Seasonal Component: If seasonality is present, a Seasonal ARIMA (SARIMA) model may be used, adding seasonal parameters (P, D, Q, m) to the model to capture seasonal patterns.

Tools and Software

Statistical Software: Common software for ARIMA modeling includes R, Python (statsmodels), EViews, and specialized forecasting packages that provide functions for ACF/PACF analysis, differencing, and ARIMA model fitting.Automated Model Selection: Automated procedures such as autoarima (in R) can suggest ARIMA models based on data-driven criteria like AIC, though results should be carefully validated.

Review of Literatures

Empirical Review

Sales forecasting has been an essential area of research in fields such as operations management, marketing, and data science. Accurate sales forecasts are crucial for effective inventory management, production planning, and strategic decision-making. Over the years, researchers have developed and applied various forecasting models, each with unique strengths and limitations. The following review outlines empirical findings on traditional statistical models, time series methods, and recent advancements in machine learning applied to sales forecasting.

Classical Time Series Models, Box-Jenkins Methodology for Sales ForecastingSeasonal ARIMA (SARIMA).Machine Learning and Hybrid Models, Hybrid Models:,Advanced Machine Learning and Deep Learning Approaches, Prophet Model:

Empirical Comparisons and Practical Implications

Multiple comparative studies, such as those by Makridakis et al. (2020), indicate that no single model consistently outperforms others across all settings. ARIMA models generally excel in short-term univariate forecasting and are preferred when data is limited. In contrast, machine learning and deep learning models tend to perform better when complex, high-dimensional, or exogenous factors impact sales data.The empirical evidence suggests that model selection should align with the nature of the increase in the interest of sales forecasting is evident from the marketing literatures.

Geurts and Ibrahim [1975] conducted a research to compare the performance of Box-Jenkins approach in forecasting Hawaii's tourists with the exponential smoothing forecasting technique developed by Brown. Box-Jenkins time series analysis, as stated by Geurt and Ibrahim, is suggested as an effective tool for short-term forecasting. For the research purpose, data used was the monthly number of tourists visiting Hawaii for the period 1952-1971 inclusive. Moreover, both the Box-Jenkins and exponential smoothing time series analyses were considered as methodological tools for forecasting the monthly number of tourists that visit Hawaii.

Data and Methodology

Data

This research study relies on secondary data collected from WushWush Tea Plantation's internal monthly sales reports. Specifically, this data spans from January 2018 to January 2023, covering a period of 60 consecutive months. This comprehensive dataset includes a variety of sales figures that reflect the factory's monthly tea production and sales activity over the five-year period. The dataset provides insights into seasonal patterns, monthly variations, and underlying trends that are essential for building an effective forecasting model.

To implement the forecasting model, the first 50 observations (from January 2018 to February 2022) have been allocated to the model's analysis and training phase. This subset will be used to identify patterns, estimate model parameters, and conduct diagnostic checks essential for developing a reliable Box-Jenkins model. The final 10 observations (from March 2022 to January 2023) are set aside to assess the forecasting model's performance by comparing predicted values against actual sales figures. This structure not only helps in

Methodology

Definition of some important terms

Time series:A set of observations generated sequentially in time.

Deterministic time series models: These are models where in future values of a time series are exactly determined by some mathematical function.

Stochastic process: A statistical phenomenon that evolves in time according to probabilistic laws.

Stationary time series: A stochastic processes that remain on equilibrium about a constant mean level. A strictly stationary process properties are unaffected by a change of time origin.

Backward shift operator (B): Let Xt be an observed variate value at time t, then BXt is given by;$BX_t = X_{t-1}$

More generally, the n^{th} order backward shift operator is given by:$B^nX_t = X_{t-n}$

Backward difference operator (V): Suppose Xt and Xt-l are two consecutive observed values. Thus, the backward difference operator applied to Xt is defined as follows.$X_t = X_t - X_{t-1}$

It can be shown that $v X_t = (1 - B)_{Xt} <=> v = 1 - B$

Autoregressive and Moving average processes

Autoregressive Model

In an autoregressive model, the current value of the process is expected as a finite, linear aggregate of previous values of the process and a shock et. That is, let Xt, Xt-1, Xt-2, ... ,Xt-p denote the observed values corresponding to equally spaced points t,t-l,t-2, ... ,t-p in time. Moreover, let be the mean level of the process (assuming the time series is stationary). Thus,

$$X_t - ù = Ø_1(X_{t-1} - ù) + Ø_2(X_{t-2} - ù) + \ldots + Ø_p(X_{t-p} - ù) + e_1$$

let us denote $X_{t-i} - ù$ by X_{t-i} ; i=O,I,2, ... , P

$$X_{t=Ø1(X\,t-1)} + Ø_2(X_{t-2}) + \ldots.. + Ø_p(X_{t-p}) + e_1$$

$$X_t = \sum_{i=1}^{p} \emptyset X_{t\_1} + e_t$$

Where e_t is the shock or error term distributed according to normal with mean zero and constant variance. It is also called white noise process.

The above presentation is called an autoregressive process of order p, i.e., AR (p). In its more compact appearance the above expression may be written as

$\emptyset(B)X_t$ (= e_t where $\emptyset$ (B) = 1- $\emptyset$ B^I - $\emptyset$ $2B^2$ - ... - $\emptyset$ $_pB^P$

Moving Average Model

A moving average model is a linear model that expresses the deviation Xt-f.! as a linear combination of the previous shocks (errors). Symbolically,

$X_t = e_t - B_1 e_{t\_1} - \emptyset_2 e_{t-2} - \ldots \emptyset_q e_{t-2}$

$$e_t - \sum_{i=1}^{q} B_i e_{t-i}$$

The expression given above is called moving average of order q or MA(q). One can write this expression in a more compact manner using the moving average operator 8(B).

X, = B(B)e , where B (B) I-BIB -B 2 B 2 -BqBq

Mixed Autoregressive moving Average

Sometimes a time series data my exhibit the feature of both autoregressive and moving average processes. In such a case we employ a model called mixed autoregressive-moving average, which is given by:

$$\emptyset (B) X_{t = \emptyset} (B)e_t$$

This equation is referred to as mixed autoregressive movmg average model of order p and q, ARMA (p,q). Box and Jenkins suggested that the adequate representation of actually occurring stationary time series can be obtained with autoregressive, moving average, or mixed models, in which p and q are not greater than 2, often less than 2.

Nevertheless, it is not always possible to obtain a stationary time series with constant mean level. As stated by Box and Jenkins many time series referring to business and economics are non-stationary. Such series may nevertheless exhibit homogenous behavior of a kind. Homogenous non-stationary behavior can be represented by a model that calls for the dth differencing of a process to be stationary. In practice, d is usually 0, 1 or at most 2. It can be noted that the whole purpose of differencing IS to convert a non-stationary time senes to that of stationary. A general model, which can represent homogenous nonstationary behaviour, is the difference equation form given by:

¢(B)(1- B)d X t = B(B)e

where d is the level of differencing. The above equation is called an autoregressive integrated moving average (ARIMA) process of order (p,d,q) or ARIMA(p,d,q).

Stages of the model building process

The overall Box-Jenkins approach to model building consists of three well-defined stages, namely

- Model identification
- Parameter estimation
- Model diagnostic checking

The following flow chart best represents the model building process.

Postulate General Class of Models
Identify model to be tentatively entertained

Estimate parameters in the tentative model
Diagnostic checking
Use the model for forecasting or control.

Table1. Steps in model building(Source: Box and Jenkins (1976:19))

Stage 1 Model Identification

The whole objective in model identification is to obtain some idea of the values of p,d, and q needed in the general ARIMA(p,d,q) model and to obtain initial guesses for the parameters. The identification of the values of p, d, and q is usually carried out using the plots of autocorrelation and partial autocorrelation functions.

a, Autocorrelation function,

clearly, the covariance between Xt and Xt+k, separated by k intervals of time, is given by$Cov[X_t,X_{t+k}] = ((Xt – U) (X_{t+k – U) = E (}X_tX_{t+k)}$

The above covariance is usually referred to as the auto covariance at lag k and denoted by Yk. Moreover, it is known that for any two variables X and Y, the correlation between them is given by:

$$\underline{Cov(X, Y)\ E((X - X)(Y - Y))}$$

Px,y = o-Xo-y = ~E(X -xy E(Y -YY

In time series analysis, we substitute X by Xt and Y by Xt+k and call the resulting expression the autocorrelation of X at lag k.

Pk =$\underline{E((X_1\text{-}U)(X_{t+k}\text{-}u)}$

$$\sqrt{E((X1 - U)E(Xt + k - u)}$$

Note that the word auto in autocorrelation refers to the fact that we are trying to find the correlation between any two observed values of a single variable separated by k periods. In short, X is correlated to itself at two different time points. For a stationary process, the autocorrelation at lag k is

P _ Yk k - Yo where Yo = a x 2 = E(X t - Jl) 2 = Var(X)

The plot of the autocorrelation coefficient as a function of the lag k is called the autocorrelation function of the process. Due to the fact that the autocorrelation function is necessarily symmetric about zero (as Pk=P-k) in practice we plot only the positive half of the function. Sometimes, the autocorrelation function is also called correlogram.

b,Partial autocorrelation function

A discussed by Box and Jenkins, partial autocorrelation function is a device that exploits the fact that whereas an AR(p) process has an autocorrelation function which is infinite in extent, it can by itself be described in terms of p non-zero functions of the autocorrelations. For an AR(p), the partial autocorrelation function will be non-zero for k less than or equal to p and zero for k greater than p. That is, it has a cutoff after lag p. The use of the autocorrelation and partial autocorrelation functions for identifying an appropriate model involves two stages.

Stage 2identifying the degree of differencing

It is failure of the estimated autocorrelation function to die out rapidly that suggests non-stationary. In order to achieve stationary a given nonstationary time series must be differenced d times. The value of d is reached, when the autocorrelation function dies out fairly quickly. The order of differencing should be the minimum possible number for which the plot of the autocorrelation function dies out fairly quickly. That means, over differencing is not appropriate. Therefore, by observing the pattern of the autocorrelation function we identify the degree of differencing d. In practice, as Box and Jenkins indicated, d is normally either 0,1, or 2 and it is

usually sufficient to inspect the first 20 or so estimated autocorrelations of the original series and of its first and second differences.

Identification of resulting stationary ARMA process

The aim of this stage of the identification process is to provide tentative values of p and q based on the tentatively determined value of d and the appearance of the autocorrelation and partial autocorrelation functions of the appropriately differenced series. To that end, whereas the autocorrelation function of an autoregressive process of order p tails off, its partial autocorrelation has a cutoff after lag p. Conversely, the autocorrelation function of a moving average process of order q has cutoff after lag q, while its partial autocorrelation tails off. Finally, if both the autocorrelation and partial autocorrelation tail off, a mixed process, containing a pth order autoregressive and a qth order moving average components. The partial autocorrelation function for a mixed process is dominated by a mixture of exponentials and damped sine waves after the first p-q lags. It can be clearly observed that the behavior of autoregressive process as measured by the autocorrelation function is the same as the behavior of the moving average process as measured by the partial autocorrelation function, and vice versa.

i. Model Estimation

After identifying the type of the model, the next step is to estimate the parameters of the model. Nowadays this model estimation process IS completely performed using statistical packages like S-Plus.

ii. Model Diagnostic Checking

In model diagnostic checking stage the adequacy of the tentative model is evaluated as a representative of the data on hand. If diagnostic checks, which have been thoughtfully devised, are applied to a model and fail to show serious discrepancies then we shall rightly feel more comfortable about using that model.

Box and Jenkins presented four methods for checking model adequacy.

These are,

./ Over fitting

./ Use of residuals

./ Portmanteau test

./ Cumulative periodgram checks

For the purpose of this paper, however, we consider the second and third approaches, as they are more convenient to deal with. Of course, the use of residuals method is used in most of the literatures sited in the previous chapter. Moreover, Box and Jenkins concluded that there is no one best way of checking a models adequacy.

a) Diagnostic checks applied to residuals

In this approach the overall model diagnostic check is done through visual inspection of a plot of the residual autocorrelations. The autocorrelation gram of the residuals is plotted and we can then see how many coefficients are significantly different from zero and whether any further terms are indicated for ARIMA model. [Chatfield 1996:73] The basic premise is that the autocorrelation of the residuals can yield valuable evidence concerning lack of fit and the possible nature of the model adequacy.

b) Portmanteau lack of fit test

As compared to the previous approach the portmanteau approach IS somewhat objective. Suppose that we have the first m autocorrelations Yk 2(e) (k=1,2,3 ... ,m) from any ARIMA (p,d,q) model. Then it can be shown that the expression

$$Q = n \sum_{k=1}^{n} \quad Y\ k^{\ 2}(e) \text{ where } n = N - d$$

is approximately distributed as X^2 (m -p -q) if the fitted model is adequate On the other hand, the value of Q will be inflated if the fitted model is inadequate. Putting differently, portmanteau model diagnostic check we test the null hypothesis.

Ho: The model is adequate

Versus its alternative

HA: The model is not adequate

The test statistics used is

$Q = n\sum_{k=1}^{m} Yk^2 (e')$ where n = N –d

The rejection criterion is Q> X a^2 (m - p - q) for a selected level of significance. That means, reject the null Q> X a^2 (m - p - q) for a selected level of significance &.

Forecasting

The task of forecasting future values, which is the basic aim of time series analysis, is usually done by minimizing the mean square error (MSE). Consequently, the resulting forecast values are called minimum mean squared error forecasts. The minimum mean square error forecast is defined In terms of the conditional expectation

$E(X_{t+1}) = E (X_{t+1} / X1, X_{t-1}.....)$

That is, the minimum mean squared error forecast at origin t, for lead time I, is the conditional expectation of ... at time t. When XI (I) is regarded as a function of I for fixed t, it will be called the forecast function for origin t.

Accordingly, the general forecasting equation for ARIMA (p,d,O) model, which is the interest of this paper, is derived from the difference equation as:

$¢B (1 -B)^d X_t = e_t$ Setting e_t to zero, we get

$¢B (l-B)^d X_t=O$

Where $¢B=1-¢_1B-¢_2B^2-'''-¢_pB^p$

The specific form of the above forecasting equation depends on the value of d. For d= 1, the equation will have the form

$X_t = (1+ ¢1) X_{t-1}+ \sum_{i=1}^{p}(¢ - ¢ i - 1)X_{i-1} -¢_p x_{t-p=1}$

It can also be written as

$\hat{} X_t= \sum_{i=1}^{p} ¢ X t-1$

$\hat{}Xt = Xt-1 + \sum_{k=0}^{n} ¢ X t - 1$

For example, the forecasting equation of ARIMA (3,1,0) is:

$\hat{}X_t = (1+¢_1)X_{t-1}+ (¢_2-¢_1) x_{t-2} + (¢_3-¢_2) x_{t-3}- ¢_3 x_{t-4}$

Data Analysis and Discussion of Results

Test of Randomness

As it is stated in the review of literatures part any time series analysis is necessarily based on the assumption that the data is not random. In other words it is assumed that the data is dependent on the time factor and hence data recorded or observed at time period t will depend on the previous lags namely t-l, t-2, etc.

It is, therefore, necessary to check whether the sales data on hand meets the non-randomness assumption. Of course, there are a number of techniques that can be used to perform test of randomness. For the purpose of this study the Turning Point test is preferred as it is the most commonly used and also the simplest test.

In the turning point test of randomness we test the null hypothesis

Ho: The data is random.

Versus the alternative

HA: The data is not random

The test statistic to be used is,

$$Z = \frac{p - E(p)}{S.E(p)}$$

Where

P is the number of turning points in the plot of the data,

$E(p) = \frac{2(n-2)}{3}$ is the expected value of p, Value of P and

$S.E(p) = \frac{\sqrt{16n-29}}{90}$ is the standard error of P.

Moreover, the distribution of the random variable p tends to normality as n, which is the number of observations approaches to infinity. That is, for largy value of n (n>50) one can safely assume the distribution of p to be normal.

We reject the null hypothesis if and only if absolute value of the calculated value of Z is greater than its tabulated value at a give level of significance usually taken to be 5% or 10%.The plot of the sales data for wushwush tea plantation factory is displayed in the next page. From the plot it can be observed that the graph involves 36 turning points for the available 60 monthly sales. Therefore our data we get

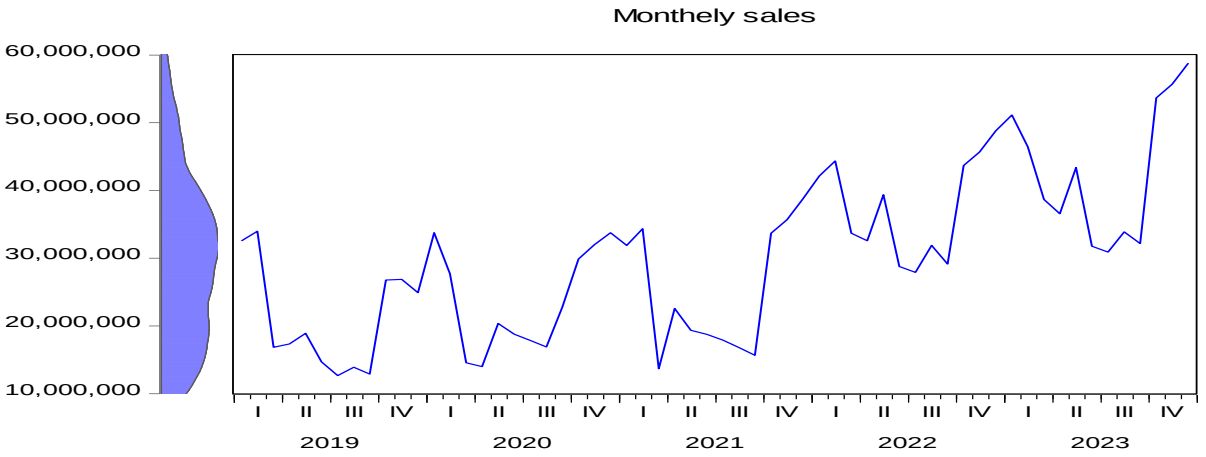
$E(p) = \frac{2(n-2)}{3} = \frac{2(60-2)}{3} = 38$ and $S.E(p) = \frac{\sqrt{16n-29}}{90} = \frac{\sqrt{16\,(60)-29}}{90} = \frac{\sqrt{10.344}}{3.216} = 3.216$

Accordingly the calculated Z value will be

$z = \frac{p - E(p)}{S.E(P)} = \frac{34 - 38}{3.216} = -1.243$

$|Z| = 1.243$

From the standardized normal distribution we refer that the tabulated values of Z for 5% and 10% levels of significant are respectively 1.96 and 1.644. Hence, since the calculated value is greater than the tabulated values at both 5% and 10% levels of significance, we reject the null hypothesis and accept the alternative, which asserts that the data is not random. Therefore, the data is identified to be time dependent and invites time series analysis.



Model Identification

It is already emphasized that the Box-Jenkins approach, which is the procedure that is followed in this study, requires the data to be stationary. Although it involves some sort of Subjectivity, the stationary of data is usually testified by observing the behaviour of the autocorrelation plots. The plot of the autocorrelation function is presented below.

Series Monthly sales and sales Data

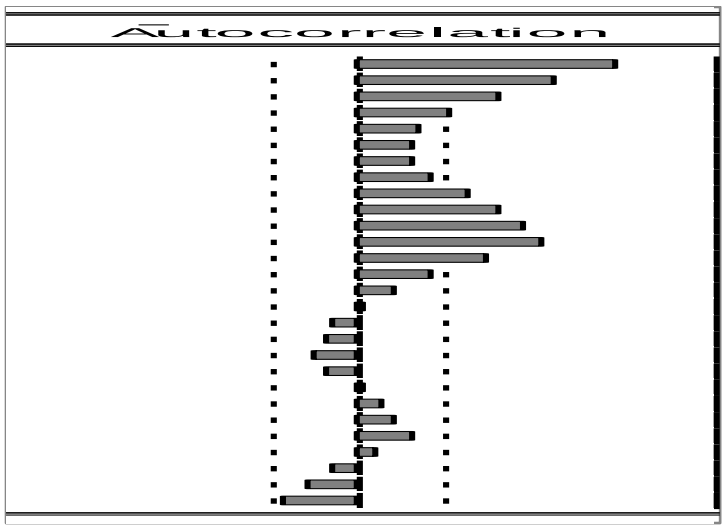


Figure 3 Plot of the Autocorrelation Function for the indifference series.

It can be seen from the above autocorrelation function plot that the graph does not die fairly quickly. Even for the lags around 17 the autocorrelation function is above the 20' -limit lines. This shows that the autocorrelation function assumes large values even at higher lags. It is known that one of the indicators of the non-stationary behavior of a given time series data is that the autocorrelation function does not die fairly quickly. Furthermore, as it can be clearly seen from the graph in Fig 2 the sales data has a decreasing or negative trend. Nevertheless, for stationary data there should be no trend in the data. In other words, the data should have a distribution with constant mean. Of course existence of significant increasing or decreasing trend confirms the non-stationary behavior of the data. Therefore, there is evidence to say that the data under consideration has some sort of non-stationary. If it had been for the autocorrelation function to die quickly after attaining large values in the first few lags, then it would have been said that the data depicts stationary behavior.

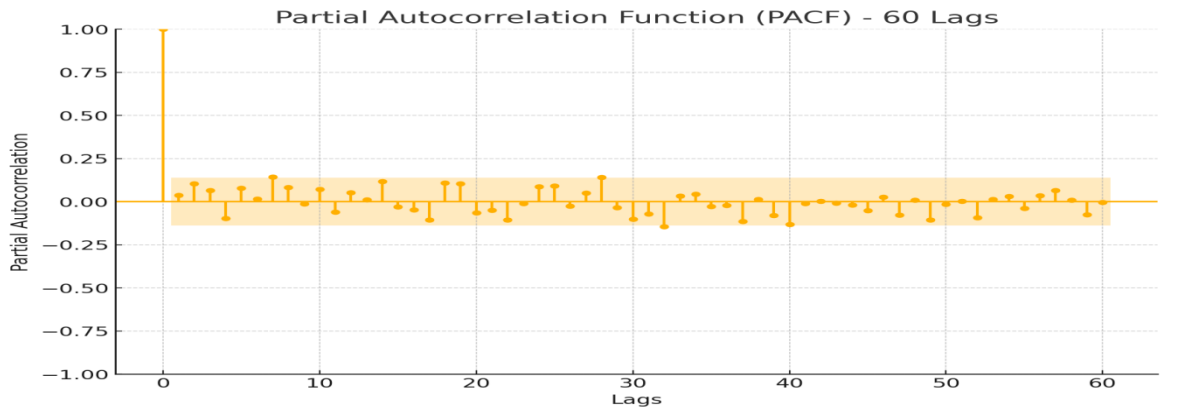


Fig 5.Partial Autocorrelation Function**(PACF) Plot for the first-degree**

The partial autocorrelation plot in Fig 5 has a cut off after lag three. In other words all the partial autocorrelation function values for the lags exceeding three are not significant as all of them fall within the control limit. Conversely, for all lags below the third the value of each of the partial autocorrelation functions is significant as it is out of the control limit. Moreover, it is known that whenever the partial autocorrelation cuts off after some lag p the implication is that the data under consideration pertains to the Autoregressive process of order p. Hence, the partial autocorrelation plot clearly reveals that the differenced sales data under consideration follows the autoregressive process of order three (p=3). Nevertheless, this identified value of p is only temporary or starting value. Its final value will be determined after performing the model diagnostic check.

On the other side, it is natural for the autocorrelation function to tail off given that the partial autocorrelation function showed a cut off after some lag. With this regard the plot displayed in Fig 4 seems to conform the expected behavior of the autocorrelation function. The autocorrelation function tails off rather than cutting off. In fact, the tailing off behavior is not clearly depicted in the plot. It IS, however, possible to rely on the partial autocorrelation function analysis and select tentative model(s).

It has to be emphasized, however, that it is possible to consider more than one temporary or candidate model. It is just as a matter of incidence that we obtained one single tentative models. Other alternative models could not be observed from the autocorrelation and partial correlation functions analysis. That is, the analysis does not suggest any other alternative model other than ARIMA (3,1,0).

Model Estimation

The purpose in the model estimation process is to determine the parameters or coefficients of the model identified in the previous section namely, ARIMA (3,1,0). In fact the model estimation process is very difficult and sophisticated process. But thanks for computers we can handle it using different statistical packages. Accordingly, the following parameter estimates are obtained using S-Plus 2000 Software Package.

X_t= -.282235X_{t_1} - .299125X_{t_2} -.447951X_{t-3}

Where Xt represents the sale at the I^{th} period.

This result tells that the values of ¢I' ¢2' and ¢3 are -0.282235, -0.299125 and -0.447951 respectively. This model is a tentative model that relates the sale at time t with the values at periods t-1, t-2, and t-3. Nevertheless, this model cannot be directly used for forecasting purpose at least for two reasons. First the model is only tentative. It needs further investigation for its adequacy. This is what we call model diagnostic check. Second, even then the model was adequate it is presented in terms of the differenced series and hence we need to rearrange terms and express it in terms of the original series in order to make forecasts.

Model Diagnostic Checking

In model diagnostic checking we testify the adequacy of the proposed ARIMA model as a representative of the data under consideration. As it is stated in the methodology, there are different techniques that can be employed to perform the adequacy of a tentative model. For the purpose of this study we use the Autocorrelation check and the Portmanteau lack of fit test. The plot of the auto correlations of the residuals, presented in fig 6 shows that all the autocorrelations are insignificant as they all are within the reference limit. Residuals are the difference or deviation of the observations from the fitted value. A satisfactory model is the one that minimizes the residuals. The absence of significant autocorrelation among the residuals indicates that there is no implied pattern in the error terms. The error terms are random in their nature and do not exhibit any common pattern that could result from the inadequacy of the model. Putting differently, whenever the model is inadequate the influence of other influential factors or variables that are not considered in the proposed model would be included in the error terms and tends to introduce a recognizable pattern in the error terms. This recognizable pattern would definitely result in high autocorrelation of residuals. This is due to the fact that high autocorrelation will occur if and only if the residuals have some common behavior.

ARIMA Model Diagnostics: Monthly Sales $ Sales Data

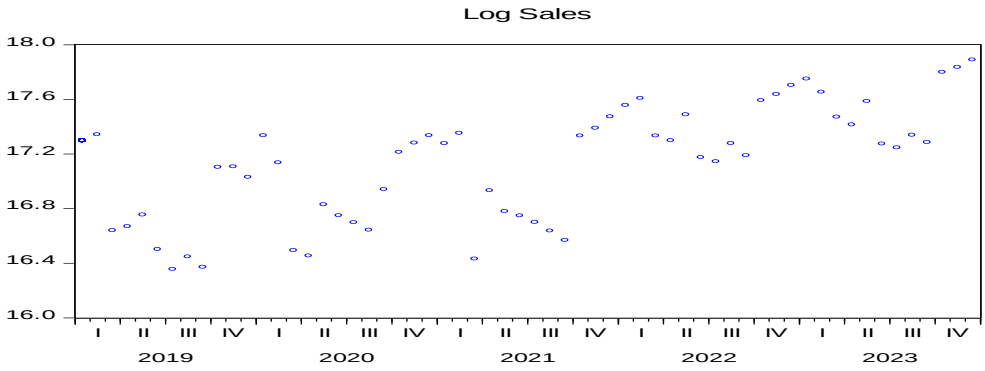


Fig 7P-Values of the Chi-Squared Statistics**in the Portmanteau test**

On the other hand, the portmanteau test also confirms that the tentative ARIMA (3,1,0) model is adequate. The detailed hypothesis testing for the portmanteau test is presented in the annex. Here we simply present the plot of the P-Values for different number of observation considered in the test.

The horizontal dotted line represents the S% level of significance (a=O.OS). In addition, each point in the plot corresponds to the P-Value of the calculated Chi-squared value. Obviously, the higher the P-value is the smaller the calculated Chi-Square value. In other words, in terms of the P-Value we reject the null hypothesis if and only if the P-Value of the calculated Chi-Square is less than a=O.OS. As it can be seen from the above plot, however, all the P-Values of the calculated Chi-squares are above the significance level (a=0.05) suggesting that the null hypotheses stating "the model is adequate" is acceptable. If at least one point in the plot were to fall below the limit then we would say there is some evidence as to the inadequacy of the model. For the sake of flexibility other similar models were tested (over fitting) for adequacy and all of them turns out to have less precise result than the ARIMA (3,1,0) model. This leads to the conclusion that the proposed ARIMA (3,1,0) is the most appropriate time series model that best fits the laundry soaps sales data collected from Gule1e Soap Factory.

Interpretation of the final Model

From the analysis performed in the previous sections of this chapter we have determined that the appropriate time series model that represents the laundry soap sales data is ARIMA (3,1,0). In its explicit form the model may be written as:

Xt= -.282235Xt_1 - .299125Xt_2 - .447951X t-3

Where X,=X,-X' I

Though the model seems to refer to only the previous three consecutive lags, it actually refers to four consecutive lags. This is due to the fact that the model is presented in the form of the differenced series. If we try to convert the above model to the one that involves the original monthly sales rather than the differenced ones, we get the following model representation.

$XI = 0.44325X_{t\_1} + 0.0\ 1595X_{t\_2} + 0.2465X_{t\_3} + 0.2943X_{t\_4}$

Where X_t is the estimated or forecasted value of the actual value X_t As it can be observed from the above representation the forecasted or the estimated value of current sales depends on the previous four lags. Meaning, in order to make sales forecasts of the monthly sales of Wushwush tea planation to take the previous four consecutive sales data in to consideration.

According to the model, the four most recent sales data determines the current sales data. In other words, it is found that the sales data for a given month is a function of the sales data for the proceeding four consecutive months. The above result, forecasting sales based on the previous four lags, seems to be in line with the implication of the autocorrelation values presented in Table 1. It can be seen that the degree of correlation between current sale and each one of the past sales data for the different lags is significant only for the first four lags. Except for the fifth lag all the remaining autocorrelation magnitudes are less than 0.5. This shows that the correlation is insignificant. In addition, the correlation between the fifth lag and the current sale is 0.53, which can also be regarded not significant as compared to the remaining lower lags. These significant autocorrelation values for the first four lags imply that current sales can be adequately forecasted from the previous four monthly sales data.

lags	1	2	3	4	5	6	7	8	9	10
Auto-corelation	0.7772	0.6682	**0.6706**	0.639	0.5374	0.4976	**0.5088**	0.4737	0.4628	**0.4725**
Lags	11	12	13	14	15	16	17	18	19	20
Auto-corelation	0.4487	0.3592	**0.379**	0.3726	0.3392	0.2543	**0.2623**	0.2258	0.1574	0.1034

Figure 1. Autocorrelation Values**for the first 20 lags**

It is also observed that all the coefficients in the model turn out to be positive. The implication being the sales volume at any given month is directly related to the previous sales records. In other words, if there is a general decreasing trend In sales volume of soaps In the past four months then the next month sales volume is also expected to decrease from its the current value, and vice versa. Clearly, this should be the case as we have observed positive autocorrelation values for different lags in Table 1. Of course, this interpretation is based on the assumption that the influence of other external factors remains to be constant. Otherwise, 20 " I I if there is an extraordinary event occurring in any single month that affects sales volume and that has not been observed in the other months, then it is possible to come up with a reversed trend than the past. Such an event can be, for example, new contracts, unusual advertisement campaigns, and other promotional activities.

Fortunately, however, the impact of such new actions or factors may not extend for more than four months. This is because the model will take the action made in to consideration through sales observed after the happening of the new event. Meaning the influence of the new happening will be reflected in the sales data recorded for four or more consecutive months so that the model will take it in to consideration. Hence, after that point the model may serve as a forecasting tool with no or little modification. Nevertheless, if the change is going to happen frequently then it will be necessary to involve the changing variable into the model, which leads us to formulating a multivariate model that involves both time series and econometric data. In addition, the sum of all the coefficients adds up to the value one. This result can be interpreted in such a way that if the sales for the preceding four months are identical, then the sales for the current period will remain to be the same. In this case we will have a constant sales forecast each month. As a result, the model will generate the same result as that of the naIve approach. Another interesting finding is that the sale at time t is influenced by or dependent much on

the previous month, i.e. t-l, sale than the other months. This is observed from the relative comparison of magnitude of the different coefficients. The value 0.44325, which is the coefficient of the most recent sales data relative to the current period, is the greatestas compared to other coefficients. Through this result the model tries to suggest that the most influential factor in determining current sale is the sale recorded for the previous month. The model attaches the highest weight to the most resent data, which is the observed sale at the period t-l. This finding is logical because for any two consecutive months the probability of having same pattern is high due to the lesser tendency of change to occur in such short period. It is to mean that recent sales data usually have more similar behavior with current or future sale as compared to other relatively older sales data. After all, in time series analysis the basic assumption is that current sales can be forecasted from previous sales. In other words, it is accepted that the major factor that affects sales is the time factor. Moreover, it is natural to expect that recent sales data will give more information than old data in forecasting future sales. This is for the reason that if time is assumed to be the determinant variable in forecasting sales then the shorter the time between an observed sale and current sale the higher will be their correlation. High correlation means that current sales can be forecasted well by using recent sales data that occur at law lags. With this regard it can be seen from table 1 that the highest correlation occurs for the first lag. Thus, it is convincing for the coefficient of the first lag to be highest as compared to the others.

Forecasting

The purpose of this chapter is to generate forecasts for future sales and examine the forecasting capability of the model. In the previous chapter we have developed an adequate ARIMA (3,1,0) model that can be used for forecasting purpose. The adequacy of a model, however, is not a guarantee for its capability of forecasting.

It is, therefore, essential to examine the forecasting power of the identified model. Strictly speaking there is no hard and fast rule for evaluating the forecasting power of a model. Usually we refer to the standard error of forecasts and comment on the forecasting power of the model, which is quite subjective and case dependent. To that end we recall that the identified model has the form:

$$X_t = 0.44325X_{t\_1} + 0.0\ 1595X_{t\_2} + 0.2465X_{t\_3} + 0.2943X_{t\_4}$$

Moreover, this model is obtained using the monthly sales data for the 60months ranging from January 2019 to August 2023. The remaining monthly sales data for the five months that range from January 2019to August 2023 are used here for forecasting purpose.

The following table summarizes the forecasts for the five lead times.

Years	Actyal sales	Forecasted Sales	Absolute Forecast Error
2019	55,955.00	51439	4,516.00
2020	55,616.00	60179	- 4,563.00
2021	45,105.00	60426	- 15,321.00
2022	50,830.00	62108	- 11,278.00
2023	54,010.00	47096	6,914.00

Table 2 Forecasted monthly Sales**of Tea measured in terms of Kuntal**

An overall look at the above result bears an impression that whenever the lead-time increases the accuracy of the forecasts tends to decline. For example, the forecast error for the first lead-time, which is 4516, is smaller when it is compared to the error for the forecast of 2020, which is 65.

This implies that the forecast done for the first month or the first lead time is more precise as compared to that of the second, and so on. In other words it is observed that whenever the lead-time increases the error also increases. The justification for this pattern is that whenever the lead-time gets larger and larger there will be increased probability for different changes to take place. It is evident that a number of factors may bring changes in the pattern of sales data whenever the forecast interval gets wider.

It is generally agreed that the time factor can be a determinant variable for forecasting sales data but not the only. There are other variables that readily affect the sales in a given month. With this regard the decline in the accuracy of the forecasts for the sales of the wushwush tea planation Factory can be attributed to the change that has been introduced in terms of advertisements or other promotional activities. It is true that the factory carry out advertisements in media like television and radio. Nonetheless, since the pattern of the advertisement carried out by the factory lacks consistency and planning it is hardly possible to take this factor into account. Obviously, unless there is consistent pattern of advertisement and other promotional activities, it would be very difficult to detect the impact of such activities on the sales volume.

Another consideration of the forecasting process is that of error analysis. Till now we were dealing with the absolute deviations of the forecasted values from the actual values. This approach, however, is more crude measure for evaluating the forecasting power of a model. Though thereare a number of other related error analysis techniques the most commonly used technique for evaluating the forecasting accuracy of a model is the U-Statistics. The interpretation of the U-statistics can be considered for three distinct cases. Although the interpretation is adapted for the purpose of this specific study, it can also be extended to the evaluation of any forecasting technique.

Case 1: U= 1 implies that the naive forecasting method is as good as the Box Jenkins approach.

Case 2: U< 1 suggests that the naive forecasting approach outperforms the Box-Jenkins approach.

Case 3: U> 1 suggests that the naive forecasting approach outperforms the Box-Jenkins approach.

For the Tea sales data on hand the U-statistics is computed to be 0.88. The interpretation of this figure is, therefore, that the derived Box Jenkins time series model outperforms the naive approach. In other words, the failure of the naive approach to outperform the Box-Jenkins approach leads to a conclusion that the latter should be preferred. Nevertheless, this conclusion applies only to the forecasts of the monthly sales of Wushwush tea planation factory ; it cannot be in general extended to other time series data.

Therefore, it can be said that the developed model can be used as a satisfactory tool for forecasting the monthly sales of Tea produced by wushwush factory. Particularly the forecasting model is found to be more powerful for short forecasting horizons than for long forecasting intervals.

The final point with regard to forecasting is the process of updating forecasts. The idea of updating is that whenever new sales data comes to existence then it should be used in forecasting other future values. For instance when we forecast the sales volume for the month October we have used the forecasted sales volume for September. Now in updating forecasts we use the actual value of the sales for September instead of the forecasted value. It is obvious that the updating process usually improves the forecasting process. Accordingly, the following updated forecasts are obtained for the five lead times considered previously.

Summary, Conclusion and Recommendation

Summary

Despite the indispensable role played by sales forecasting in decision making and strategy formulation, there has been no formal means of sales forecasting developed or used in wushwush tea planation factory that enables management forecast monthly sales of produced Tea. This situation has brought problem in the decision making process of the management of the factory. The management usually encounters problems in making decisions about future actions in connection with raw material requirement planning, operations planning and financing, human resource planning, marketing activities, etc.

~ It provides the management of wushwush tea planation factory with a more reliable and appropriate forecasts that obviously improve, or at least facilitate, the decision making task.

~ It serves as a model that greatly helps other similar business organizations in Ethiopia in developing their own forecasting models through time series analysis. In other words the paper clearly illustrates how time series analysis can be applied in sales forecasting provided all the required assumptions are satisfied.

~ It serves as a reference material for future researches, i.e., it adds something to the available literatures especially in developing countries like Ethiopia.

In order to achieve its basic objective the study applies the Box-Jenkins Time Series Analysis approach on the monthly sales of Tea produced and sold by wushwush Factor. A total of 60 monthly salesdata are collected from which the last five are reserved for forecasting purpose.

The analysis and discussion of results, which is the major team of any study, is presented in the fourth chapter. The analysis is performed using S-Plus 2000 statistical package and gives the result that the monthly sales pattern of Tea sold by wushwush Factory can be represented by Autoregressive Integrated Moving Average model of order 3,1,0 (ARIMA (3,1,0)). The general model in its explicit form is found tohave the form;.

$$X_t = 0.44325X_{t\_1} + 0.0\ 1595X_{t\_2} + 0.2465X_{t\_3} + 0.2943X_{t\_4}$$

Furthermore, the model is testified for its forecasting power or accuracy using the reserved five months sales data. As stipulated in chapter five, the model is found to be satisfactorily powerful for the intended forecasting purpose. Besides, the model is also compared with the naive forecasting method, which is assumed to be the simplest and also less costly method. With this regard the U-Statistics confirmed that the identified ARIMA (3,1,0) model performs better as compared to the naive forecasting method.

Conclusion

- The monthly sales of Tea can be forecasted using the ARIMA (3,1,0) model with a satisfactory degree of accuracy. Moreover, the sales data for the previous four lags is vital in making current sales forecast.
- The suggested model seems to serve better for short forecasting horizons. Whenever the forecasting interval gets larger the accuracy of forecasts are observed to reduce significantly.
- There is a seasonal component pertaining to the sales data considered that affects the monthly sales pattern. The periodicity of this seasonal factor is approximated to be between 3 and 4 months.
- The sales data recorded for the first lag (X_{t-1}) is the major determinant factor in forecasting current sales (X_t). That is, the time factor has a significant role in forecasting sales.

Recommendation

- ✓ Management of wushwush tea planation Factory may facilitate its decision making task by conducting sales forecasting based on the suggested ARIMA (3,1,0) model.
- ✓ The suggested forecasting model is most appropriate to be used for forecasting interval(s) preferably not more than three months.
- ✓ The model should be updated with some interval using the new sales record for the new months that has not been included within the model formulation process. The model should also be revised in cases where there is some significant change in the environment that is believed to have some sort of impact on sales of Tea produced by the factory.
- ✓ Although more than satisfactory results can be obtained, pure time series may not be sufficient in conducting sales forecasting. In addition to the time factor, it is extremely advantageous to incorporate other independent variables that readily affect sales volume. One such critical variable could be advertisement expense both by the factory and also by other competitors in the same industry.
- ✓ Other companies may also use the same approach or procedure for building their own forecasting model that satisfactorily forecast their future sales, and hence facilitating the overall effectiveness of the companies.

Bibliography

1. Box and Jenkins. "Time Series Analysis: Forecasting and Control," Revised Edition, California: Holden-Day, 1976.
2. Chatfield, "The Analysis of Time Series: An Introduction," Fifth Edition, UK: Chapman and Hall, 1996.
3. Geurts D. and Ibrahim B., "Comparing the Box-Jenkins Approach with the Exponential Smoothing Forecasting Model Application to Hawaii Tourists." Journal of Marketing Research, Vol. XII, May

4. 1975(182-187}.
5. Green and Tull. "Research for Marketing Decisions," Second Ed. USA: Prentice-Hall, 1970.
6. Heizer and Render."Principles of Operations Management." Third Ed, USA: Prentice-Hall, 1999.
7. Helmer M. and Johansson K., "An Exposition of the Box-Jenkins Transfer Function Analysis with an Application to the AdvertisingSales Relationship."Journal of Marketing Research, Vol. XIV, May 1977(227 -239}.
8. Kappoor S.G.,Madhok P., and WU, "Modeling and Forecasting Sales Data by Time Series Analysis." Journal of Marketing Research, Feb. 1981.
9. Keay F. "Marketing and Sales Forecasting." UK: Pergamon Press, 1972.
10. Kotlet P. "Marketing Management," Third Ed., USA: Prentice Hall Inc., 1967.
11. Lancaster and Reynolds. "Introduction to Marketing: A Step-By-Step Guide to all the Tools of Marketing." UK, 1999.
12. Lancaster and Reynolds."Marketing." UK: Reed Educational and Professional Pub., 1998.
13. Leone P. "Modeling Sales-Advertising Relationship: An Integrated Time Series Econometric Approach." Journal of Marketing Research, Vol. XX, Aug. 1983(291-295)
14. Moriarty and Adams, "Issues In Sales Territory Modeling and Forecasting Using Box-Jenkins Analysis." Journal of Marketing Research, Vol. XVI, May 1979(221-232)
15. Moriarty and Gerald Salamon."Estimating and Forecasting Performance of a Multivariate Time Series Model of Sales." Journal of Marketing Research, Vol. XVII Nov. 1980(558-564)
16. MurdickG., and Schaefer E. "Sales Forecasting for Lower Costs and Higher Profits." USA: Prentice-Hall Inc., 1967.
17. Parsons W. " Improving Marketing Performance." UK: Gower Publishing, 1987.
18. Smith and Shelby." A two Stage Sales Forecasting Procedure Using Discounted Least Squares." Journal of Marketing Research, Vol. XXXI, Feb. 1994(44-56)
19. Umashanhar and Johannes."Forecasting with Diagonal Multiple time Series Models: an Extension of Univariate Models." Journal of Marketing Research, Vol. XX, Feb 1983(58-63).