Development of Demand Forecasting Models for Improved Customer Service in Nigeria Soft Drink Industry_ Case of Coca-Cola Company Enugu

Godwin H. C.¹ and Fakiyesi O. B.²
Department of Industrial/Production Engineering, Nnamdi Azikiwe University Awka, Nigeria¹, ²
Email: hcgodwin@yahoo.com¹ and oladapo_fakiyesi@yahoo.com²

ABSTRACT
The study is targeted at developing appropriate forecasting models for the production of Coca-Cola products using Box-Jenkins method. This method is based on a group of stages which are (defining-estimating-diagnosing-and forecasting). The time series of five (5) products of the company were examined namely; Coke, Fanta, Sprite, Limca and Schweppes which were tagged ‘Product 1, 2, 3, 4 and 5’ respectively. Each of the products was also examined separately on both monthly and quarterly basis. The results of this study showed that the suitable and efficient model to represent the data of the time series according to AIC, BIC, MSE and RMSE criteria with the smallest values as well as the Box-Ljung test are the fitted models; SARIMA (0,1,0) (1,1,1) 4, SARIMA (1,1,0) (1,1,1) 4, SARIMA (1,1,1) (0,1,0) 4, SARIMA (1,1,0) (1,1,1) 12 & SARIMA (0,1,0) (1,1,1) 12. According to these results, the future demand of the products has been forecasted from 2015 to 2019 and those values gotten showed harmony with their counterparts in the original time series. It provided us with the image of the reality of the expected demand in future.

Keywords: Akaike and Bayesian Information Criteria (AIC & BIC), Demand, Forecasting, Seasonality, Time Series.

1. INTRODUCTION
Forecasts for the soft drink industry are made using Volume (in gallons) and revenue (in naira). Consumption from a volume perspective is expected to increase as a result of an anticipated increase in consumer spending as the recession ends, above-average expansion of the 55-and-older age groups, faster-paced lifestyles that demand convenience products, and rising demand for functional beverages. A number of factors determine demand for soft drinks; price, income, consumers’ lifestyles and tastes. The absence of effective and scientific demand forecasting methods in most Nigeria public organizations seems to be the main bane of shortages and excess in human resources resulting to unmanageable and experience imbalances in the number and quality of employees needed to optimally achieve organizational objectives and plans [1].

The study is significant in the following ways: The findings will guide the decision makers in coming up with economic growth Soft Drink sector policies that favor both individual and public owned investments. It will also enhance the productivity level of Nigerian Bottling Company as such reduces the bottle-neck in the production line of The Coca-Cola Company. This would eventually create efficient and effective system that will accommodate the future demand of their products.

Health issues are a hot topic with many consumers and, as a result, are driving demand in both directions. Soft drinks developed to be low-calorie, low-sugar, and preservative-free are in line with consumers’ health consciousness, and demand for these products is increasing. At the same time, the public debate about nutrition, and specifically about Sugar-Sweetened Beverages (SSBs), has reduced demand for non-diet Carbonated Soft Drinks (CSDs) or shifted demand to diet CSDs [2].

Several demand forecasting techniques currently exist. They vary from fairly simple qualitative methods based on individual or group judgments to highly complicated methods involving sophisticated statistical computerization. In this study, the problem is to analyze the demand-forecasting model using a Non-Seasonal and Seasonal Autoregressive Integrated Moving Average (SARIMA) model. The rationale for choosing this type of model is contingent on the behavior of the time series data. Also in the history of demand forecasting, this model has proved to perform better than other models because the model can replicate existing conditions, and therefore suitable to predict future demand [3].
Autoregressive Integrated Moving Average (ARIMA) is the method first introduced by Box and Jenkins (1976) and until now become the most popular models for forecasting univariate time series data. This model has been originated from the Autoregressive model (AR), the Moving Average model (MA) and the Combination of the AR and MA, the ARMA models. In the case where seasonal components are included in this model, then the model is referred to as Seasonal Autoregressive Integrated Moving Average (SARIMA) model. Box-Jenkins procedure comprises model identification, model estimation and model checking [4]. The scope of this research work is limited to development of demand forecasting models for Coca-Cola products. The application of Seasonal (ARIMA) models/methods were employed to develop the models that successfully forecast the future monthly and quarterly demand of these products. These models are capable of replicating the stochastic process that generated the time series. Forecasting is a scientifically calculated guess for estimating future event by casting forward past data. The past data is systematically combined in a predetermined way to obtain the estimate of the future [5]. Godwin and Igboanugo [6] analyzed time series of PHCN using ARMA and ARIMA models. According to their findings, the models predict fairly accurately in-sample and out-of-sample values. Ihueze and Okafor [7] developed predictive production rate model using general multiplicative regression equation form. Similarly, time series decomposition analysis was used to study seasonality and trend in some selected products [8]. Their findings shown reduction in production level and short period forecasting was also recommended.

2. METHODOLOGY AND MODELING
This study centered on month-to-month data for five years which were obtained from production/sales department of the case study. The model/method adopted in this study represents:

\[
SARIMA(p, d, q) \times (P, D, Q)s
\]  

(1)

Where \( p = \) non-seasonal AR order, \( P = \) seasonal AR order, \( d = \) non-seasonal differencing, \( D = \) seasonal differencing, \( q = \) non-seasonal MA order, \( Q = \) seasonal MA order. The entire procedure for the time series modeling selected for this study has been summarized in the flow chart shown in Fig 1.

![Flow Chart](image)

**FIGURE 1:** Schematic Representation of the Box-Jenkins Methodology for Time Series Modeling [9]
Following the method as proposed by Box and Jenkins [10], the components are then broken down as follows according to the order in Fig. (1).

2.1. Trend and Seasonal Differences: Based on the findings from the data used in this study, it was found that both trend and seasonality were present in the original time series data of the case study. Thus, there was need to detrend and also remove the presence of seasonality using the general stationarity transformation according to Box and Reinsel [4].

\[ Z_t = (1 - B^L)^D (1 - B)^d y_t \]  \( (2) \)

Where \( B \) = backshift operator, \( L \) = number of seasons, \( D \) = degree of seasonal differencing, \( d \) = degree of non-seasonal differencing. \( y_t \) = pre-differencing transformation.

2.2. Identification of Potential Models: After the time series data has been certified stationary (i.e. the mean, variance and autocorrelations are constant). Haven applied both seasonal and non-seasonal differences of order 1 to make the time series data stationary (i.e. \( d = 1, D = 1 \)). Therefore, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the stationary time series of the case study were then examined. Fig. 2 shows the basic characteristics for the identification of \( p, q, P, \text{and } Q \) in the form of Equation (1).

<table>
<thead>
<tr>
<th>Process</th>
<th>Autocorrelation Function (ACF)</th>
<th>Partial Autocorrelation Function (PACF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR (p)</td>
<td>Tails off towards zero (exponential decay or damped sine wave)</td>
<td>Cuts off to zero after lag p</td>
</tr>
<tr>
<td>MA (q)</td>
<td>Cuts off to zero after lag p</td>
<td>Tails off towards zero (exponential decay or damped sine wave)</td>
</tr>
<tr>
<td>ARMA (p, q)</td>
<td>Tails off towards zero (exponential decay or damped sine wave)</td>
<td>Tails off towards zero (exponential decay or damped sine wave)</td>
</tr>
</tbody>
</table>

**Figure 2:** Distinguishing Characteristics of Theoretical ACF and PACF

*Source:* Adopted from Chatfield [11]

The ACF and PACF of the stationary time series are computed and expressed as follows:

\[ ACF = \frac{Cov(x_t, x_{t-h})}{\text{Std.Dev.}(x_t)\text{Std.Dev.}(x_{t-h})} = \frac{Cov(x_t, x_{t-h})}{\text{Var}(x_t)} \]  \( (3) \)

\[ PACF = \frac{Cov(y_{x_i}, x_{1}, x_2)}{\sqrt{\text{Var}(y_{x_i}, x_1, x_2)\text{Var}(x_1, x_2)}} \]  \( (4) \)

Where \( t = \) time, \( h = \) number of lags, \( y = \) response variable, \( x = \) predictor variable

From the critically observation of the ACF and PACF plots of the time series data of the case study, after differencing the data to obtain stationarity. Considering also both quarterly and monthly plots of each product, the following models were tentatively identified as presented in Tables 1(a) and 1(b).

**Table 1(a): Quarterly Model Criteria**

<table>
<thead>
<tr>
<th>Product</th>
<th>SARIMA Models</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>( SARIMA (1,1,1)(1,1,1)_4 )</td>
<td>4.6987</td>
<td>5.1467</td>
</tr>
<tr>
<td></td>
<td>( SARIMA (2,1,1)(1,1,1)_4 )</td>
<td>4.7987</td>
<td>5.2965</td>
</tr>
<tr>
<td></td>
<td>( SARIMA (0,1,1)(1,1,1)_4 )</td>
<td>4.4987</td>
<td>4.8472</td>
</tr>
<tr>
<td></td>
<td>( SARIMA (0,1,1)(1,1,1)_4 )</td>
<td>4.5987</td>
<td>4.9981</td>
</tr>
<tr>
<td>P2</td>
<td>( SARIMA (1,1,1)(2,1,1)_4 )</td>
<td>4.7987</td>
<td>5.2965</td>
</tr>
<tr>
<td></td>
<td>( SARIMA (1,1,1)(1,1,1)_4 )</td>
<td>4.6987</td>
<td>5.1467</td>
</tr>
</tbody>
</table>

www.ijsret.org
Table 1(b): Monthly Model Criteria

<table>
<thead>
<tr>
<th>Product</th>
<th>SARIMA Models</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>SARIMA (1,1,2)(2,1,1)12</td>
<td>4.4653</td>
<td>5.1633</td>
</tr>
</tbody>
</table>

2.3. Model Selection: The tentative models selected in this study as shown in Tables 1(a) and 1(b) using a penalty functions statistics: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). These criteria were used to measure the goodness-of-fit of estimated statistical models and are calculated using the expressions below:

\[
AIC = 2k - 2\log(L) \quad \text{OR} \quad 2k + n \log(\frac{RSS}{n})
\]

\[
BIC = -2 \log(L) + k \log(n) \quad \text{OR} \quad \log(\sigma^2_e) + \frac{k}{n} \log(n)
\]

Where; \(k\) = The number of parameters in the statistical model, \(L\) = The maximized value of the likelihood function for the estimated model, \(RSS\) = The residual sum of squares of the estimated model, \(n\) = The number of observation, equivalently the sample size, \(\sigma^2_e\) = The error variance [12].

Finally, the models that has the least values of AIC and BIC for each product were then selected and summarized in Table 3.
2.4. Parameter Estimations: Immediately a suitable SARIMA \((p, d, q)(P, D, Q)_s\) structure is identified, the next step is the parameter estimation or fitting stage. The parameters of the models as given in Table 2 are estimated by the maximum likelihood method.

2.5. Diagnostic Checking and Model Validation: The selected models might appear to be the most appropriate among the series of models considered; it becomes necessary to do diagnostic checking to verify the adequacy of the models. This was done in this study by:

i. Verifying the ACF of the residuals.

ii. Verifying the normal probability plots of the residuals.

2.6. Normal Probability Plot of the Residuals

The normal probability plot is a graphical technique to identify substantive departures from normality. This includes identifying outliers, skewness, a need for transformations, and mixtures. Normal probability plots are made of raw data, residuals from model fits, and estimated parameters. In a normal probability plots (also called a “normal plot”), the sorted data are plotted versus values selected to make the resulting image look close to a straight line if the data are approximately normally distributed. Deviations from straight lines suggest departures from normality [11].

The normal probability plots of the selected models in this study are analyzed in Fig. 3-7.

It is critical for SARIMA model to examine the behavior of residual values to see whether they are normal, random and have constant variation. From Fig 3, the assumptions are met quite well, except there are non-constant variations in the Versus Fits Plot. This stems from the fact that the quality of the model’s fit is better for early data points than for more recent ones.

![Figure 3: Normal Probability Plot of Residual for MODEL (A)](image)

Fig. 4 shows the normal probability plot of the residuals for MODEL (B), non-constant variations were noticed in Versus Fits plot and also the Observation Order show reasonable fluctuations.

![Figure 4: Normal Probability Plot of Residual for MODEL (B)](image)

Looking at Fig. 5 which shows the normal probability plot of the residual for Model C, the residuals roughly form a “horizontal bond” around the 0 line at the versus
fits. This suggests that the variances of the error terms are equal.

Figure 6: Normal Probability Plot of Residual for MODEL (D)
Fig. 6 shows the normal probability plot of residual for Model D. Observing the plot critically shows that the residuals “bounces randomly” around the 0 line. This suggests that the assumption that the relationship is linear is reasonable. The points on this plot form a nearly linear pattern, which indicates that the normal distribution is a good model for this data set.

Figure 7: Normal Probability Plot of Residual for MODEL (E)
This type of residual test is carried out specifically to determine the true nature of the residuals. And it shown clearly from Fig. 3-7 that the nature of the residual followed a normal distribution (i.e. the residuals of the models are normally distributed). This phenomena added to the facts already proved in this study that these selected models are more adequate and can then be used to forecast the future demand of Coca-Cola products of the case study.

2.7. Forecasting: This is the last phase of Box and Jenkins methodology. The models that have been selected and tested as appropriate are used to make future predictions for demand of Coca-Cola products. Fig. 8-12 shows plot of 5-year (60 months) time series data for Products 1-5 with the forecast results for predicted demand for a future period of 5 years (60 months) therein (i.e. 2015-2019).
3. Discussion of Results
The time series of the differenced data shown the data is stable and yielded accurate values during the analysis. The first difference of the original data was taken to remove trend and the seasonal fluctuations at lag 12. This was done on both monthly and quarterly data series of each product. All these products of the differenced time series data judging from the ACF and PACF plots, the hypothesis of stationary dependency of the time series of both quarterly and monthly observations is not rejected and thereby made it possible to obtain stationarity in the series.

Two major goodness-of-fit were used for model selection; Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). These were determined based on likelihood function incorporating the sum of squared errors (SSE), the sum of parameters and the number of observation. Invariably, a single model for each product (monthly and quarterly) was selected with the least values of AIC and BIC. These selected SARIMA models were then simplified to five (5) models which expected to make forecast for Products 1-5 (monthly and quarterly) as shown in Table 2.

These models were properly checked and validated by verifying the ACF of the residuals as well as the normal probability plots of the residuals. The result showed that the residual of the models are normally distributed. This supports the fact that the models are adequate.

Using the general Box-Jenkins model of order (p, P, q, Q), the point forecast of the models for each product (quarterly and monthly) were made. Model-A (quarterly) was used to make forecast for Products 1, 3 and 5 after the model has been fitted as
\[ y_{t+1} = \gamma_t + \theta_4 e_{t-4} + \theta_1 \theta_4 e_{t-4} \]

The demand forecast of Period 21 was made
\[ P1 = 336173, P3 = 219657, P5 = 123624 \]

Model-B (quarterly) was used to make forecast for Products 1 after the model has been fitted as
\[ y_{t+1} = \gamma_t + \alpha y_{t-3} + \alpha y_{t-4} - \theta_1 e_t + \theta_1 \theta_4 e_{t-4} \]

The demand forecast of Period 21 was made
\[ P1 = 128751 \]

Model-C (quarterly) was used to make forecast for Products 1 and 2 after the model has been fitted as
\[ y_{t+1} = \gamma_{t-3} + \alpha y_{t-4} \]

The demand forecast of Period 21 was made
\[ P2 = 96040 \]
Model-D (monthly) was used to make forecast for Products 1 after the model has been fitted as

$$y_{t+1} = (1 + \alpha)y_t - \alpha y_{t-1} - \theta_1 e_{t+1} + \theta_1 \theta_{12} e_{t-12}$$

The demand forecast of Period 21 was made

$$P1 = 124994, P2 = 36546$$

Model-E (monthly) was used to make forecast for Products 1, 3 and 5 after the model has been fitted as

$$y_{t+1} = y_t + y_{t-11} - y_{t-12} - \theta_{12} e_{t-11} + \theta_{12} \theta_{12} e_{t-12}$$

The demand forecast of Period 21 was made

$$P3 = 68394, P4 = 41944, P5 = 34546$$

Note that all these values are in Centiliters ($10^2$CL).

Fig. 8-12 has shown clearly the variation between the actual data and forecast. The actual data covers a period of 60 month observations (i.e. 2010-2014) whereas the forecast was made from a period of 61-120 months (2015-2019).

3. Conclusion

The study mainly intended to develop forecasting models using Box-Jenkins Autoregressive Integrated Moving Average (SARIMA) to make future predictions. The statistical tests show that the time series of the monthly and quarterly Coca-Cola production of Nigerian Bottling Company (NBC), Enugu is not stable and has seasonal changes. To ensure stability in the series, firstly, the general trend was removed using differences of the first lag, second, the seasonal differences of lag 12. The most appropriate models were chosen using the balancing standards (the smallest value of each: AIC, BIC, MSE and RMSE as well as the Box-Ljung test). The models developed were able to replicate the stochastic process that generated the time series thereby eliminate other factors that might affect the demand of these products as demonstrated in Fig. 8-12. Coca-Cola products have become leading products in soft drink industry, adopting the results of these findings would help the company to meet future demand of their products and customers’ satisfaction. Effort should be made to have a centralized forecasting method to achieve even production and distribution of the products across the nation.

REFERENCES