SALES FORECASTING OF PERISHABLE PRODUCTS: A CASE STUDY OF A PERISHABLE ORANGE DRINK

T. Musora, Z. Chazuka, A. Jaison, J. Mapurisa, and J. Kamusha

School of Natural Sciences and Mathematics,Department of Mathematics,
Chinhoyi University of Technology ,Pivate Bag-7724, Chinhoyi, Zimbabwe

ABSTRACT

The primary goal of any organization involved in trading business is to maximize profits while keeping costs to a bare minimum. Sales forecasting is an inexpensive way to achieve the aforementioned goal. Sales forecasting frequently leads to improved customer service, lower product returns, lower deadstock, and efficient production planning. Because of short shelf life of food products and importance of product quality, which is of concern to human health, successful sales forecasting systems are critical for the food industry. The ARIMA model is used to forecast sales of a perishable orange drink in this paper. The methodology is applied successfully. ARIMA $(0,1,1)(0,1,1)_{12}$ was concluded as the appropriate model. Model diagnostics were done; results showed that no model assumption was violated. Fitted values were regressed against observed values. A very strong linear relationship was evident with an R^2 value of over 90% which is very plausible.

KEYWORDS

Sales forecasting, Orange Drink, ARIMA, Model Diagnostics, R^2- value.

1. INTRODUCTION

Forecasting plays a key role in decision-making and business planning. Probably, the most important function of business is forecasting. Demand forecasting in brief is an estimation of a supply chain constituent's expected sales in a specified future period [1] A forecast is a starting point in planning. The objective of forecasting is to minimize the risk in decision making. To a large extent, success depends on getting those forecasts rightly, [2] gives some important forecasting applications for the strategic areas in business. Also, [2] explains the types of forecasts and managerial planning. From this explanation, it can be concluded that a distribution company aims to determine the optimal supply of orange drinks that minimizes costs or maximizes profits in the face of uncertain demand. In the case of more shipping than demand the company has undue costs caused by stocking, high returns, transportation, and other operational issues, or in the case of less shipping than demand the company has sales lost. [3] used machine learning to forecast horticultural sales and concluded that machine learning outperforms classical forecasting on horticultural sales. Classical forecasting methods for example Autoregressive Integrated Moving Average and Exponential Smoothing are nevertheless widely used in research and industry. Regardless of their rather simple concept, they often show a competitive performance. ([4],[5] , [6]).

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT 2023
pp. 103-116, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130408

[7] states that reliable forecasts are essential for a company to survive and grow. In a manufacturing environment, management must forecast the future demands for its products and on this basis provide for the materials, labor, and capacity to fulfill these needs. These resources are planned and scheduled well before the demands for the products are placed on the firm. Forecasting is the heart and blood of any inventory control system. A firm with hundreds or thousands of items must anticipate in advance demands that will occur against these items. This is needed to have the proper inventory available to fill customers' demands as they come in. Management must plan several months for this inventory since procurement lead times from suppliers generally runs from one to six months. With each time, forecasts are needed for the months in the planning horizon. The forecasts are used to determine whether or not an order to the supplier is needed now and if so how large the order should be [7] explains that forecasting techniques can be categorized into three groups. The first is called qualitative, where all information and judgment relating to an item are used to forecast the item's demands. This technique is often used when little or no demand history is available. The forecasts may be based on marketing research studies, the Delphi method, or similar methods. The second group is called causal, where a cause-and-effect type of relation is sought. Here, the forecaster seeks a relation between an item's demands and other factors, such as business industrial, and national indices. The relationship is used to forecast the future demands of the item. The third group is called time series analysis, where a statistical analysis of past demands is used to generate the forecasts. A basic assumption is that the underlying trends of the past will continue into the future. This paper is primarily concerned with forecasting as it relates to time series analysis. In this context, the time series represents the demands recorded over past time intervals. The forecasts are estimates of the demands over future time intervals and are generated using the flow of demands from the past. This paper proceeds as follows. Section 2 gives the literature review, some theoretical structures for exponential smoothing models, and autoregressive integrated moving average (ARIMA) models. Section 3 includes comprehensive empirical results and analysis of orange drink circulation and results. Section 4 is the discussion and conclusion

2. TIME SERIES ANALYSIS AND MODELLING STRATEGY

The importance of predicting future values of a time series cuts across a range of disciplines. Economic and business time series are typically characterized by trend, cycle, seasonal, and random components. Powerful methods have been developed to capture these components by specifying and estimating statistical models. These methods comprise; log transformation, square root transformation exponential smoothing, and ARIMA, which are described by [9] and [10]. They reveal that ARIMA gives more accurate out-of-sample forecasts on average compared to other smoothing methods, although ARIMA requires much more effort. [11] states that exponential smoothing originated in Robert G. Brown's work as an OR analyst for the US Navy during World War II. [12] identify that the more sophisticated exponential smoothing methods seek to isolate trends or seasonality from irregular variation. Where such patterns are found, the more advanced methods identify and model these patterns. The models can then incorporate those patterns into the forecast. Exponential smoothing uses weighted averages of past observations for forecasting. The effect of past observations is expected to decline exponentially over time. [13] states that the exponential smoothing methods are relatively simple but robust approaches to forecasting. They are widely used in business for forecasting demand for inventories. Three basic variations of exponential smoothing are given simple exponential smoothing, trend-corrected exponential smoothing, and the Holt-Winters method. [14] states that the ARIMA method developed by [15] is one of the most noted models for time series data prediction and is often used in econometric research. The ARIMA method has been originated from the autoregressive (AR) model, the moving average (MA) model, and the combination of the AR and MA, the ARMA model. Compared with the early AR, MA, and ARMA models, the ARIMA model is more flexible in application and more accurate in the quality of the simulative or predictive results. [15] highlight that in the ARIMA analysis, an identified underlying process is generated based on observations to a time series for generating a good model which shows the process-generating mechanism precisely.

[17] and [18] have considered that the only problem with ARIMA appears that the modeling is mathematically sophisticated in theory and requires a deep knowledge of the method. Therefore, building an ARIMA model is often a difficult task for the user, requiring training in statistical analysis, a good knowledge of the field of application, and the availability of an easy-to-use but a versatile specialized computer program. The BoxJenkins approach to modeling and forecasting time series data is but one of a large family of quantitative forecasting methods which have been developed in the fields 12 of operations research, statistics, and management science. Box-Jenkins models are also known as ”ARIMA” models, the acronym standing for Autoregressive Integrated Moving Average. This terminology is made clear in the following sections. Exponential smoothing, linear regression, Bayesian forecasting, and generalized adaptive filtering are some of the other techniques which are termed ”extrapolative” forecasting [6]. Many of these methods have a common element; they utilize only the previous values of a series of numbers to forecast the future values of interest. Hence, they are referred to as univariate models, since the values from a single variable are used to predict the future values of the same variable. This is in contrast to multivariate models, where the variable of interest is also considered to depend on other variables

2.1. ARIMA Model

The ARIMA model is an extension of the ARMA modelling the sense that by including auto-regression and moving average it has an extra function for differencing the time series. If a dataset exhibits long-term variations such as trends, seasonality and cyclic components, differencing a dataset in ARIMA allows the model to deal with them. Two common processes of ARIMA for identifying patterns in time-series data and forecasting are auto-regression and moving average.

2.2. Autoregressive Process

Most time series consist of elements that are serially dependent in the sense that one can estimate a coefficient or a set of coefficients that describe consecutive elements of the series from specific, time-lagged (previous) elements. Each observation of the time series is made up of random error components (random shock; a_t) and a linear combination of prior observations.

2.3. Moving Average Process

Independent from the autoregressive process, each element in the series can also be affected by the past errors (or random shock) that cannot be accounted for by the autoregressive component. Each observation of the time series is made up of a random error component (random shock, ϵ) and a linear combination of prior random shocks.

2.4. Autoregressive Integrated Moving Average Process, ARIMA (p,d,q)

A series $X_t$is called an autoregressive integrated moving average process of orders p,d,q, ARIMA(p,d,q), if $W_t = \nabla^d X_t$, where $W_t$is the differenced time series.

We may define the difference operator ∇ as $\nabla X_t = X_t - X_{t-1}$. Differencing a time series $\{X_t\}$ of length n produces a new time series $\{W_t\} = \{\nabla^d X_t\}$ of length $n-d$. If $\{Z_t\}$ is a purely random process with mean zero and variance σ_z^2, the general autoregressive integrated moving average process is of the form

$$Wt = \phi 1Wt-1 + \phi 2Wt-2 + ... + \phi pWt-p + Zt + \theta 1Zt-1 + ... + \theta qZt-q$$

In terms of the backward shift operator, the ARIMA(p,d,q) process is

$$\Phi_p(B)W_t = \Theta_q(B)Z_t$$

Remark: The autoregressive integrated moving average process is specifically for non-stationary time series. The differencing transformation is useful in reducing a nonstationary time series to a stationary one.

2.5. Seasonal Auto-regressive Integrated Moving Average Process

Let s, be the number of observations per season. Then the time series, X_t, is called a seasonal autoregressive integrated moving average process of orders p,d,q, seasonal orders P,D,Q and seasonal period s, if it satisfies;

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D X_t = \theta_q(B)\Theta_Q(B^s)Z_t$$

Where $\nabla_s^D X_t = \sum_{j=0}^{D}\binom{D}{j}X_{t-js}$, and $\phi_p(B)$ and $\theta_q(B)$ are polynomials in B of order p and$q$, that is ;

$$\phi_q(B) = \left(1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q\right)$$
$$\theta_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$$

We identified the stationary component of a data set by performing the Ljung and Box test. We tested this hypothesis by choosing a level of significance for the model adequacy and compared the computed Chi-square (χ^2) values with the (χ^2) values obtained from the table. If the calculated value is less than the actual(χ^2) value, then the model is adequate, otherwise not. The $Q(r)$ statistic is calculated by thefollowing formula:

$$Q_((r)) = n(n+2)\sum\frac{r^2(j)}{n-j}$$

where$n$ is the number of observations in the series and $r(j)$ is the estimated correlation at lag j. Furthermore, we tested the data to specify the order of the regular and seasonal autoregressive and moving average polynomials necessary to adequately represent the time series model. For this purpose, model parameters were estimated using a maximum likelihood algorithm that minimized the sums of squared residuals and maximized the likelihood (probability) of the observed series. The maximum likelihood estimation is generally the preferred least square technique. The major tools used in the identification phase are plots of the series, correlograms (plots of autocorrelation and partial autocorrelation verses lag) of the autocorrelation function (ACF) and the partial autocorrelation function (PACF).The ACF and the PACF are the most important elements of time series analysis and forecasting. The ACF measures the amount of linear dependence between observations in a time series that are separated by a lag k. The PACF plot helps to determine how many autoregressive terms are necessary to reveal one or more of the following characteristics:time lags where high correlations appear, seasonality of the series, and trend either in the mean level or in the variance of the series. In diagnostic checking, the residuals from the fitted model were examined against their adequacy. This is usually done by correlation analysis through the residual ACF plots and by goodness-of-fit test using means of Chi-square statistics. At the forecasting stage, the estimated parameters were used to calculate new values of the time series with their confidence intervals for the predicted values.

2.6. Performance Valuation

To choose the best model among the class of plausible model, the estimated parameters were tested for their validity using, ACF , PACF, Probability Plot and Histogram of residuals, a time series plot of observed and fitted values and other error statistics such as coefficient of determination(R^2) were analysed.

2.7. Data Source

The data used in this research is historical data of monthly sales of cases of the perishable drink from a small drink manufacturing company in Harare, Zimbabwe which among other products manufactures the perishable orange drink. Each case contains 24 bottles of the drink. The company intends to minimise losses due to returns of the drinks as result of reduced shelf life.

3. RESULTS AND ANALYSIS

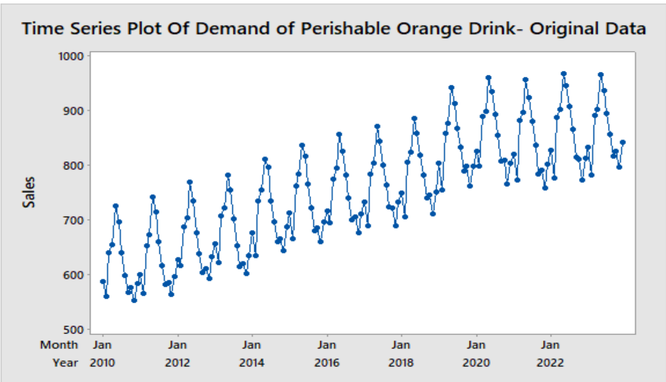

Fig.1: Time Series Plot Of Demand of Perishable Orange Drink- Original Data

Visual inspection of the plot shows that the series is dynamic. So need is there to transform the data so as to make it stationary.

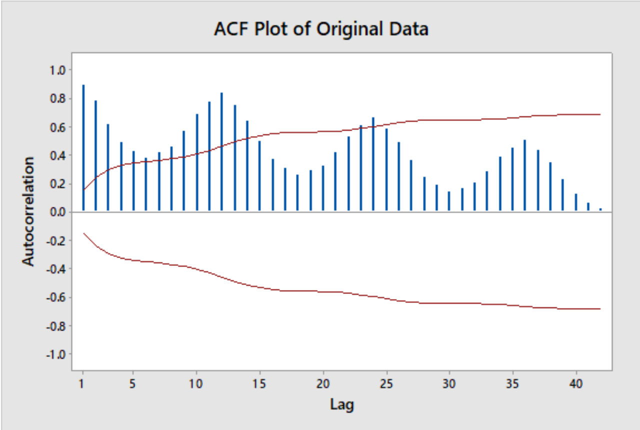

Fig.2: ACF Plot of Original Data

ACF of most lags are very high, there is evidence of positive and negative autocorrelation. This is a typical ACF plot of a non stationery time series. Thus a model cannot be fitted at this stage. This further affirms need to transform the data.

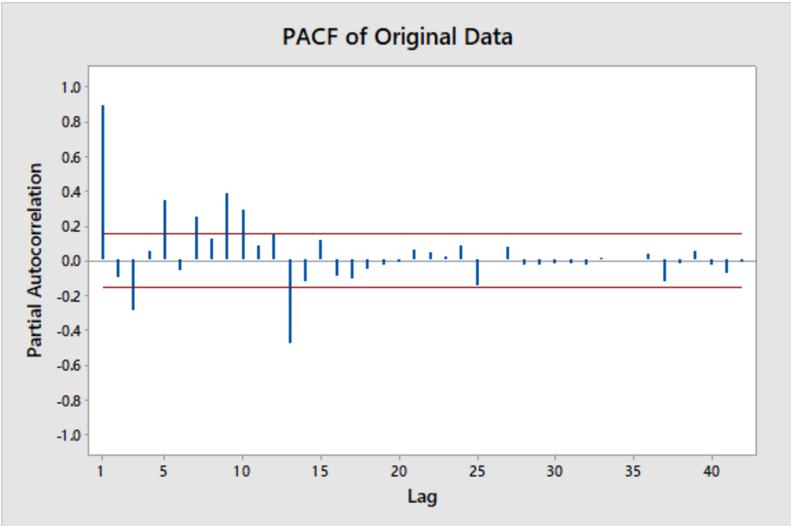

Fig.3: PACF plot of Original Data

The PACF plot shows a number of significant spikes, which is typical of a non stationary series. Thus we have to transform the data to make it stationary.

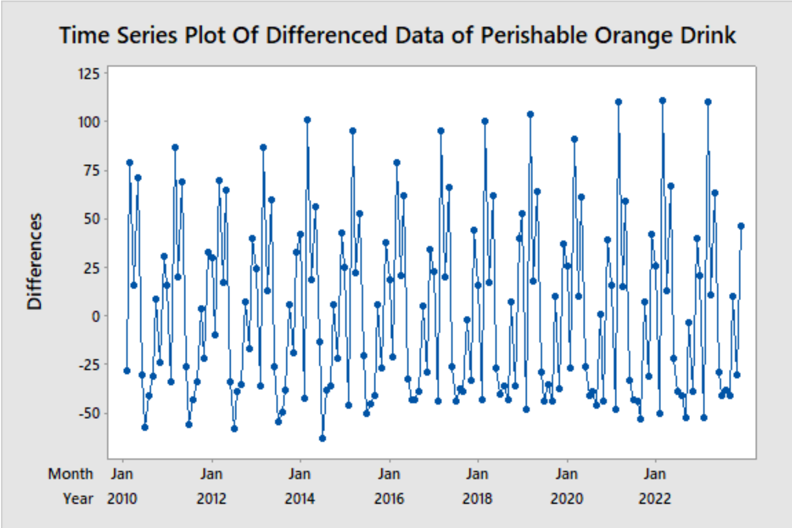

Fig.4: Time series plot of Differenced Data of Perishable Orange Drink

Visual inspection of the plot reveals that the differenced series fluctuates around zero, thus the data is now stationary

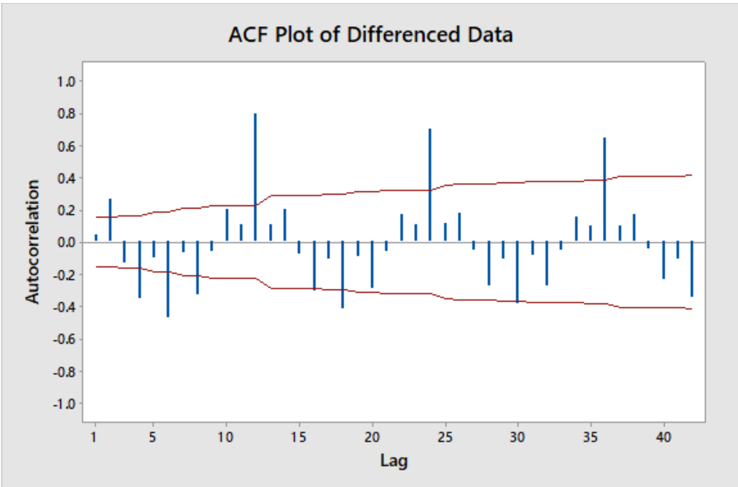

Fig.5: ACF plot of Differenced Data

The ACF shows a significant spike at lag 2 and there is evidence of negative dumped oscillations with the rest of the ACF's essentially zero, hence a seasonal ARIMA model is suggested

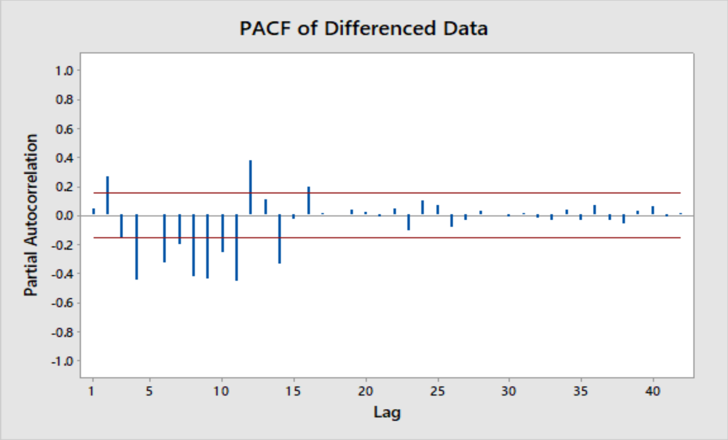

Fig.6: PACF plot of Differenced Data

PACF plot shows a significant spike at lag 2 which is seasonal and there is evidence of negative dumped oscillations with the rest of the PACFs essentially zero, hence a seasonal ARIMA model is also suggested.

3.1. Parameter Estimation

Final Estimates of Parameters

Type	Coef	SE Coef	T	P
MA 1	0.9707	0.0321	30.21	0.000
SMA 12	0.6533	0.0660	9.90	0.000
Constant	-0.00181	0.01501	-0.12	0.904

Differencing: 1 regular, 1 seasonal of order 12 Number of observations: Original series 167, after differencing 154 Residuals: *SS* = 8667.31 (back forecasts excluded) *MS* = 57.40
DF = 151 Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	99.2	232.2	326.0	405.3
DF	9	21	33	45
P-Value	0.000	0.000	0.000	0.000

Thus the fitted model is SARIMA(0,1,1)(0,1,1)$_{12}$

3.2. Model Diagnostics

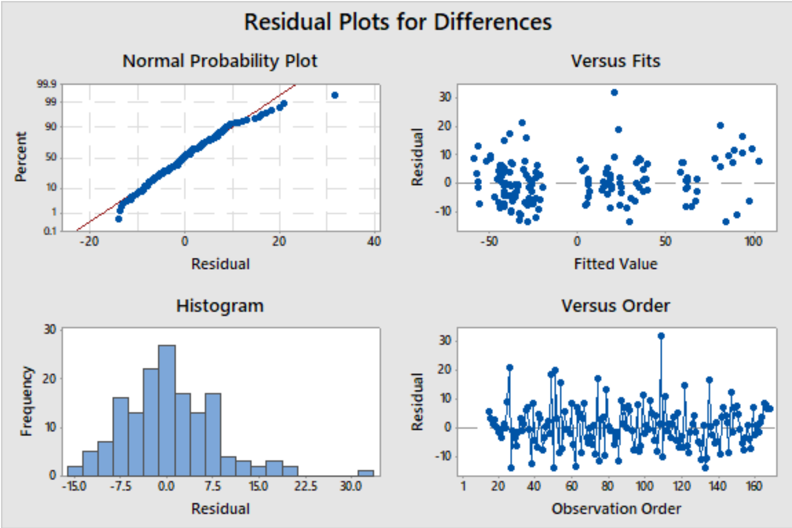

Fig.7: Residual Plot for Differences

The normal probability plot is almost a straight line, an indication that the normality assumption has not been violated. A plot of residuals against fitted values shows no pattern and the histogram of residuals also indicates that the normality assumption has not been violated. Hence the fitted model is good and thus can be used for forecasting.

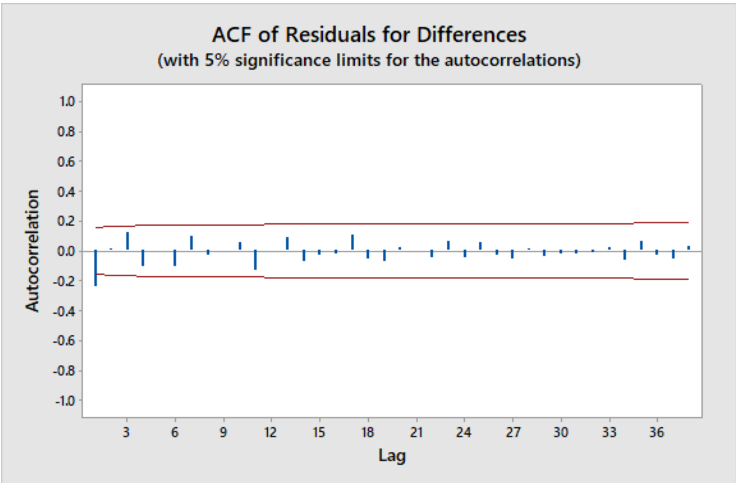

Fig.8: ACF of residuals for Differences

Figure 8 ACF plot has no significant spikes suggesting that there might be no possible additional parameters which may have been omitted in this model.

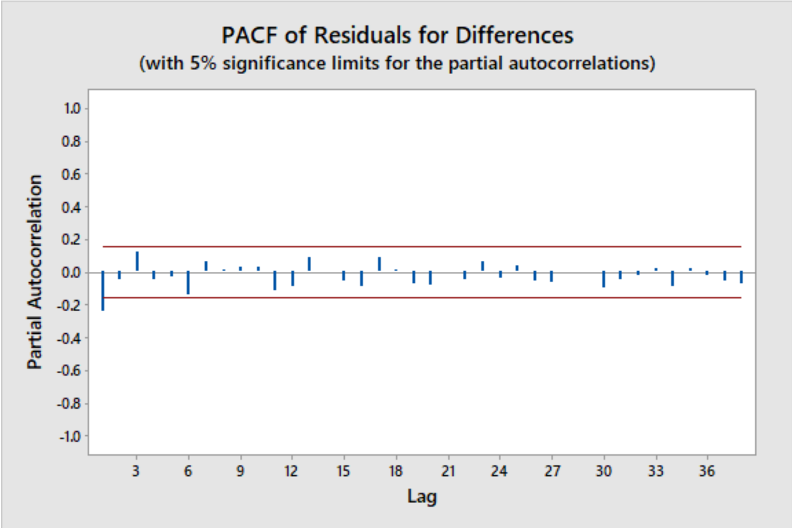

Fig.9: PACF of residuals for Differences

Figure 9, The PACF plot of residuals refuses any significant spikes suggesting that there might be no possible additional parameters that may have been omitted in this model. Since the fitted model appears good enough, it can be used for forecasting future demand of the perishable orange drink.

3.3. Inference Based on the Model

3.3.1. Forecasts from period 159

Period	Forecast	Lower	Upper	Actual
159	847.49	783.94	911.04	892.00
160	836.74	752.93	920.55	903.00
161	896.10	795.37	996.84	966.00
162	897.13	782.01	1012.25	937.00
163	881.36	753.45	1009.26	896.00
164	869.28	729.77	1008.80	858.00
165	831.16	683.39	983.87	817.00
166	808.89	670.91	991.40	827.00
167	826.39	639.32	978.55	797.00
168	828.86	639.23	1013.47	
169	830.30	609.70	1024.07	
170	831.37	591.24	1050.89	
171	833.43	573.95	1072.51	
172	834.99	557.99	1092.21	
173	836.55	543.07	1111.99	
174	838.11	529.03	1130.03	
175	839.67	515.73	1147.19	
176	841.23	503.09	1163.60	
177	842.79	491.03	1179.36	
178	842.79	479.47	1194.55	
179	844.35	487.93	1209.23	
180	845.91	468.36	1223.45	
181	847.47	457.67	1236.26	

The fitted values compares well with the observed values, thus the fitted model is reliable.

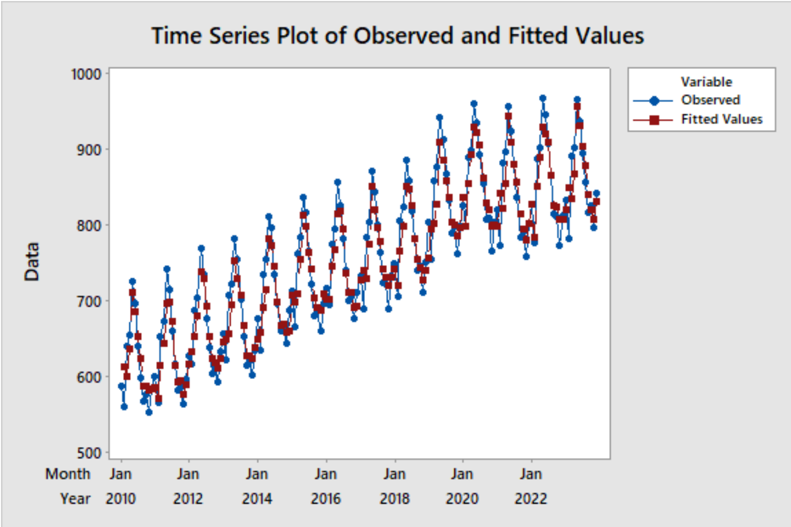

Fig.10: Time Series plot of Observed and Fitted Values

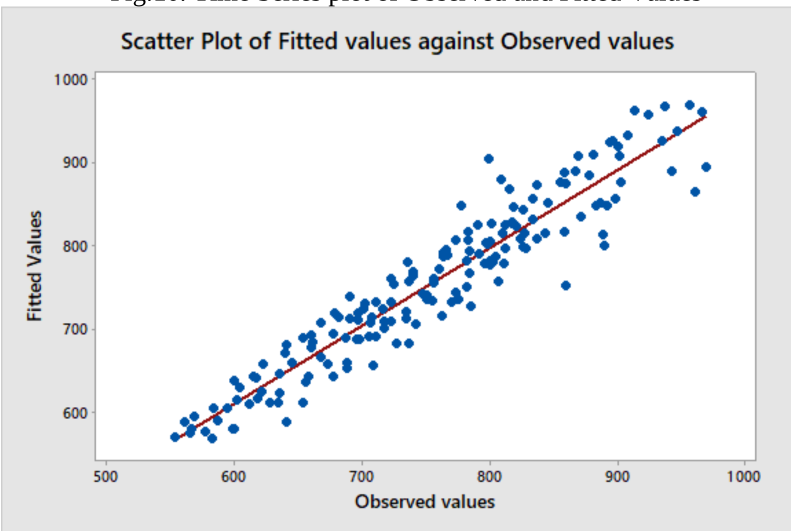

Fig.11: Scatter plot of Fitted values against Observed Values

3.4. Regression Analysis: Fitted Values versus Sales

The scatter plot of fitted values against observed values suggests a positive linear relationship.

Method

Rows unused 1

Analysis of variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	150339	153391	1500.63	0.000
Error	165	165303	1002		
Total	166	1668694			

Model Summary

R- sq	R-sq(adj)	R-sq(pred)
90.09%	90.03%	89.84%

Coefficients

Term	Sales	Se Coef	T-Value	
Constant	48.4	18.4	2.62	0.009
Sales	0.9360	0.0242	38.74	0.000

Regression Equation

$$\text{Fitted Values} = 48.4 + 0.9360 \times \text{Sales}$$

The coefficient of determination value is 90.09% indicates that the fitted model accounts for about 91% of the variation in the fitted values. Thus the fitted seasonal ARIMA model which generated the fitted values must me appropriate and hence can be used to forecast sales values.

4. Discussions and Conclusions

This study demonstrates how ARIMA time series and Regression models are useful to study and forecast sales for a particular company. This paper demonstrates also how the Time Series Forecasting System can be used to construct a model of forecasting. The ARIMA$(0,1,1)(0,1,1)_{12}$ predicted the data considerably well and gave reliable forecasts. According to the data presented, this model was best in forecasting the sales, but could not tell why the sales will contain outliers. The Time Series forecasting system helped construct a model, the ARIMA time series and the Regression, which is effective for forecasting and can be applied to other businesses in order to plan their sales. However, it would be interesting to do further research on the factors that influence the sales, such as the growth of the population of consumers, the industrial growth in the region, the immigration, and so on; this would consolidate better this company's planning.

References

[1] Islek, I.; Oguducu, S, 2017. A Decision Support System for Demand Forecasting based on ClassifierEnsemble. In: Communication Papers of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, 13: 2017, p. 35–41.

[2] Shim, J.K. (2009). Strategic Business Forecasting: Including Business Forecasting Tools and Applications, Global Professional Publishing.

[3] Haselbeck, F. Killinger, J, Menard, K, Hannus T and Grim , G.D.(2020) Machine Learning OutperformsClassical Forecasting on Horticultural Sales Predictions

[4] Kolassa, S. (2021). Commentary on the M5 forecasting competition. International Journal of Forecasting, http://dx.doi.org/10.1016/j.ijforecast.2021.08.006, Advance online publication

[5] Makridakis, S., Spiliotis, E., &Assimakopoulos, V. (2020). The M4 competition: 100, 000 time series and 61 forecasting methods. International Journal of Forecasting, 36(1), 54–74. http://dx.doi.org/10.1016/j.ijforecast.2019.04.014.

[6] Makridakis, S., Spiliotis, E., &Assimakopoulos, V. (2021). The M5 competition: Background, organization, and implementation. International Journal of Forecasting,

[7] http://dx.doi.org/10.1016/j.ijforecast.2021.07.007, Advance online publication

[8] Thomopoulos, N.T. (1980). Applied Forecasting Methods, Prentice Hall.

[9] Wei, W.W.S. (2006). Time Series Analysis: Univariate and Multivariate Methods, Second edition, Addison Wesley.

[10] Granger, C.W.J. and Newbold, P. (1974). Spurious regressions in econometrics, Journal of Econometrics,2, pp. 111–120.

[11] Reid, D.J. (1975). A review of short-term projection techniques, Practical Aspects of Forecasting, H.Gordon, ed., London: Operational Research Society.

[12] Gass, S.I. and Harris, C.M. (2000). Encyclopedia of operations research and management science,Centennial edition, Dordrecht, The Netherlands: Kluwer.

[13] Yafee, R. and McGee, M. (2000). Introduction to Time series Analysis and Forecasting with Applicationof SAS and SPSS, Academic Press.

[14] Gardner, E.S. (1985).Exponential smoothing: The state of the art, Journal of Forecasting, 4, pp. 1–28.

[15] Ediger, V.S., Akar, S. and Ugurlu, B. (2006). Forecasting production of fossil fuel sources in Turkeyusing a comparative regression and ARIMA model, Energy Policy, 34, pp. 3836–3846.

[16] Box, G.E.P. and Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control, Revised ed.Holden-Day, San Francisco, USA.

[17] Gardner, E.S. (2006). Exponential smoothing: The state of the art – Part II, International Journal ofForecasting, 22, pp. 637–666.

[18] Ho, S.L. and Xie, M. (1998). The use of ARIMA models for reliability forecasting and analysis, Computers Industrial Engineering, 35, pp. 213–216.

[19] Melard, G. and Pasteels, J.M. (2000). Automatic ARIMA modeling including interventions, using timeseries expert software, International Journal of Forecasting, 16, pp. 497–508.

[20] Billah, B., King, M.L., Snyder, R.D. and Koehler, A.D. (2006). Exponential smoothing model selectionfor forecasting, International Journal of Forecasting, 22, pp. 239–247.

[21] Cho, S.H. and Song, I. (1996). A Formula for Computing the Autocorrelations of the AR Process, TheJournal of the Acoustical Society of Korea, 15, 4–7.

[22] Diebold, F.X. (2001). Elements of Forecasting, Second edition, Thomson Learning.

[23] Enders, W. (1995). Applied Econometric Time Series, John Wiley and Sons, New York.

[24] Eshel, G. (2003). The Yule Walker Equations for the AR Coefficients, Citeulikearticle- id: 763363.

[25] Fuller, W.A. (1996). Introduction to Statistical Time Series, Second edition, John Wiley and Sons,New York.

AUTHORS

T Musora Is a PhD student at Chinhoyi University of Technology, He is currently being supervised by Dr. F Matarise from the university of Zimbabwe.He holds a MSc in Operations Research from NUST, Zimbabwe, B.Sc., (Special Hons)in Operation Research & Stats from NUST, Zimbabwe, BSc., Maths & Stats with Zimbabwe Open University, Zimbabwe and has a Dip. Ed.from Gweru Teachers College, Zimbabwe. Mr Musora is also a lecturer in the Department of Mathematics at Chinhoyi University of Technology.

Dr. Z Chazuka Holds a PhD in Mathematical Biology from University of South Africa. She also holds an Msc In operations Research from the National University of Science of Technology Zimbabwe, and a B.Sc Applied Mathematics from the National University of Science of Technology . Currently, she is lecturer at Chinhoyi University of Technology. Her research interests is in Mathematical Biology.

J Mapurisa received a MSc In Mathematics from University of Zimbabwe. He holds a B.Sc in Mathematics from the University of Zimbabwe . Currently, he is Lecturer at Chinhoyi University of Technology. His research interests is in Fluid Dynamics (flow in channels) .

A Jaison is a Lecturer in the Department of Mathematics at Chinhoyi University of Technology. He holds a MSc in Operations Research from NUST, Zimbabwe, B.Sc. in Operation Research from NUST, Zimbabwe. His research interest is in Multivariate Analysis, Financial and Statistical Modeling

J Kamusha received aMsc In Mathematical Sciences from Stellenbosch University. He holds a Bachelor’s degree in Mathematics majoring in Actuarial Science . Currently, he is Lecturer at Chinhoyi University of Technology. His research interests is in Graph Theory and Convolutional Neural Networks .