Strathmore
UNIVERSITY

Forecasting Exotic Vegetable Wholesale Prices Using Time Series Analysis Methods

Kate Njoki Mbugua | 097791

Submitted in partial fulfilment of the requirements for the Degree of
Bachelor of Business Science Actuarial Science at Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

February, 2021

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the Research Project contains no material previously published or written by another person except where due reference is made in the Research Project itself.

Kate Njoki ..... [*Name of Candidate*]

..... [*Signature*]

10/02/2020 ..... [*Date*]

This Research Project has been submitted for examination with my approval as the Supervisor.

JOHN NGIDA ..... [*Name of Supervisor*]

..... [*Signature*]

10/02/2021 ..... [*Date*]

Strathmore Institute of Mathematical Sciences

Strathmore University

Abstract

Vegetables are known to be highly perishable and seasonal in nature. Forecasting future prices of vegetables is important to the Government of Kenya, the buyers and particularly to farmers as they can use this information to help them maximise their profits and minimise their losses. Therefore, accurate forecasting of their prices requires the use of models that take the seasonal nature of vegetables into account. In this research paper, the Seasonal Autoregressive Integrated Moving Average (SARIMA) and Holt-Winter's Exponential Smoothing (HWES) models were used to forecast the wholesale prices of kales and cabbages in Nairobi, Kenya using monthly price data from January 2012 to December 2019. The Augmented Dickey Fuller (ADF) test showed that both the kales and cabbages price data were stationary hence no need for differencing. SARIMA$(1,0,0)(1,1,1)_{12}$model was the best forecasting model for cabbages. This model was selected amongst other SARIMA models as it had the least Akaike Information Criterion (AIC) value. The Holt-Winter's Exponential Smoothing method was the best for kales. Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were the forecast performance measures used to select the best forecasting model for kales and cabbages.

Table of Contents

List of Tables

List of Figures

List Of Abbreviations

ACF	Autocorrelation Function
ADF	Augmented Dickey Fuller
AIC	Akaike Information Criterion
AR	Autoregressive
ARCH	Autoregressive Conditional Heteroskedasticity
ARMA	Autoregressive Moving Average
ARIMA	Autoregressive Integrated Moving Average
ARIMAx	Autoregressive Integrated Moving Average-Exogenous
ASTGS	Agricultural Sector Transformation and Growth Strategy
BIC	Bayesian Information Criterion
EAC	East African Community
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GDP	Gross Domestic Product
GoK	Government of Kenya
HWES	Holt-Winter's Exponential Smoothing
Kg	Kilogram
Ksh	Kenya Shilling
MA	Moving Average
MAD	Median Absolute Deviation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLE	Maximum Likelihood Estimation
MoAFL&I	Ministry of Agriculture, Fisheries, Livestock and Irrigation
MSD	Mean Squared Deviation
MSE	Mean Squared Error
NAFIS	National Farmers Information Services
PACF	Partial Autocorrelation Function
RMSE	Root Mean Squared Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SSE	Sum of Squares Error
VAR	Vector Autoregressive

1. Introduction

A time series is a sequential set of data points measured typically over successive times. The data points can be measured hourly, daily, monthly, yearly or at any other regular intervals and in some cases, irregular intervals. Time series modelling involves collecting and studying past observations to make forecasts. Therefore time series forecasting can be defined as the process of predicting future events by understanding the pattern in the past (Ratnadip et al., 2013).

1.1 Background Information

The Kenyan economy is mainly driven by agriculture, despite the fact that less than 8% of its land is used for farming. In 2019, agriculture contributed 34.5% to the country's GDP, second in lead after the service industry (GoK, 2019). About 56% of the country's total labour force is employed in the agricultural sector, which has contributed highly in the reduction of poverty. (World Bank, 2017).

Smallholder farmers make up at least 70% of Kenyan farmers and they produce about 63% of food in the country. However, more than half of their agricultural produce is non-marketed subsistence production, meaning it is retained for household consumption (Food and Agriculture Organisation, 2015). The ministry of agriculture came up with the Agricultural Sector Transformation and Growth Strategy (ASTGS, 2019-2029), which is a 10 year strategy that aims to aid in achieving Food Security, one of the government's Big Four agenda. In addition to that, it strives to improve farmers' income, lower cost of food and create employment. In order to achieve these set goals, the strategy has to combat price volatility, which is twice that of its neighbouring EAC countries. Hence, Kenyan smallholder farmers are highly disadvantaged. Price instability not only affects the farmers (producers), but also the consumers. The strategy also has to improve access to agricultural price information. The smallholder farmers will then be able to maximise their output by planning production properly while simultaneously being aware of resource limitations, environmental and soil conditions.

Therefore, the challenge is on how to keep the stable prices reasonable for farmers and simultaneously affordable to the consumers. (MoAFL&I, 2019)

To be able to tackle these challenges, it is important to determine and forecast prices of some crops in the market to aid in proper planning and decision making.

Horticulture is a key earner of foreign exchange to the country. In 2019, it contributed Ksh 153 billion to the GDP, a 33% improvement from 2018. The increase in earnings is mainly due to an increase in demand and favourable international prices, despite the shortage of fertilisers that was experienced countrywide. Vegetables, fruits and flowers are the main components of horticulture in Kenya. According to (PASGR, 2019), they contribute 44.6%, 29.6%, 20.3% respectively to the total horticulture produce, with the remainder being accounted for by nuts, medicinal and aromatic plants.

Consumption of vegetables has become more popular as a means to maintain good health, prevent chronic diseases, and achieve food security. Vegetables are classified as exotic vegetables, indigenous vegetables, aromatic plants and Asian vegetables. In 2018, exotic vegetables contributed 21.78% to the domestic value of horticulture. Some of the leading exotic vegetables in production were kales and cabbages respectively. They are the most widely consumed vegetables in the country. According to (Agriculture Food Authority, 2018), they make up 71.1% of exotic vegetables produced in the country. Therefore, by forecasting these 3 crops' prices, farmers will be able to properly plan their production.

1.2 Problem Statement

Poor accuracy and access of price market information weakens farmers' negotiating power. As they sell their produce to middlemen, they are at a disadvantage since the more informed middlemen tend to exploit the information asymmetry and as a result, the farmers are forced to take the low prices offered to them (MoAFL&I, 2019).

Vegetables are bought quite frequently by consumers. Their prices are very unstable as they fluctuate based on supply and demand mainly because they are seasonal and highly perishable in nature. Therefore, with improved accuracy of market information on the vegetables' prices, smallholder farmers will be able to capitalise on the best price and hence, be more encouraged to sell their produce, rather than retaining it for household consumption.

If the smallholder farmers consider the forecasted vegetable prices, they may have an idea on the direction of the anticipated price fluctuations and hence they will be able to plan production better. They will also be able to sell their produce at profitable prices.

1.2.1 Objectives Of The Study

This research aims at accurately forecasting the vegetable prices which will improve the smallholder farmers' decision making process when planning on what to grow, when to harvest and what markets to sell their produce, given the seasonality and high perishability of the exotic vegetables. The smallholder farmers will also be able to balance the markets supply and demand.

The government will also benefit since forecasting prices will reduce the common practice that involves farmers planting a particular crop when its prices skyrocket, just for them to experience losses in the following season.

1.2.2 Research Objectives

1. Evaluate and compare different time series models that account for seasonality.
2. Test the models' accuracy in prediction of the wholesale monthly vegetable prices.
3. Identify trend and seasonality patterns in the wholesale prices of the exotic vegetables.

1.2.3 Research Questions

1. Does seasonality affect the prices of the exotic vegetables?
2. Does the time series become stationary after differencing once?

2. Literature Review

In Kenya, the demand for natural foods, including local vegetables, has increased over the years as the consumers have become more health conscious. Vegetables are highly perishable and their prices are interlinked with seasonality such that the vegetables are cheaper when in season and more expensive when it is off season (Research Solutions Africa, 2015). Therefore, forecasting vegetable prices is essential as it will provide market price information to the smallholder farmers (producers), consumers, processors and intermediaries who as a result will be able to make better planning decisions.

Assis et. al (2010) forecasted prices of cocoa beans in Malaysia by comparing 4 time series models namely , autoregressive integrated moving average (ARIMA), generalised autoregressive conditional heteroskedasticity (GARCH), the mixed ARIMA/ GARCH models and exponential smoothing. A regression test was carried out and it showed that seasonality had no significant effect on the changes of the monthly cocoa bean prices. The time series was also stable which contradicted previous studies on cocoa bean prices in Malaysia which concluded that the domestic cocoa bean prices were volatile. The results of ex-post forecasting showed that the mixed ARIMA/GARCH model was the best model for forecasting cocoa bean prices.

Gathondu (2014) used three autoregressive models: Vector Autoregressive (VAR), Autoregressive Moving Average (ARMA), GARCH and the mixed ARMA /GARCH model to forecast and model the wholesale prices of selected vegetables for markets in Nairobi, Eldoret, Mombasa, Nakuru and Kisumu. The wholesale price data used was recorded weekly for four years, from 2010 to 2013. By plotting the data, it was observed that the prices were volatile and not stable. A regression model was used to determine the existence of linear trends in the data for each of the vegetables. The regression models were not the best fit for the data as they only explained a small portion of the data. Since the price data had presence of volatility, the mixed ARMA/ GARCH model was used for accurate measurement and reliable forecast and as a result, it was the best

model in forecasting. In the study, it was found that the best forecasting ARIMA models were ARIMA(1,1,0) for kales and ARIMA(2,1,2) for cabbages.

Mutwiri (2015) challenged the ARIMA(3,0,1) model for tomatoes as he claimed that the models used by Gathondu (2014) failed to capture the seasonality component of vegetable prices. For his research, the SARIMA model was used to forecast tomato wholesale prices in Nairobi, Kenya using monthly data collected from 1981 to 2013. The model's prediction and accuracy was tested using mean absolute error (MAE), root mean squared error (RMSE) and mean absolute percentage error (MAPE). In this research, SARIMA $(2, 1, 1)(1, 0, 1)_{12}$was the best forecasting model.

Boateng et.al (2016) came up with a model to analyse tomato prices in the Ashanti Region of Ghana. The price data which was recorded from 1994 to 2015 was analysed using the SARIMA model as they suspected a seasonal pattern in their data. The SARIMA$(0,1,1)(0,1,1)_{12}$was the best model. They reported that there was a price trend for the period used in the analysis apart from 2015, where all things being equal, prices rose in the second quarter and were expected to drop in both the third and fourth quarters.

In another study in Turkey, Adanacioglu & Yercanm (2012) used the SARIMA model to analyse the seasonality of the monthly wholesale prices of tomatoes. SARIMA $(1, 0, 0)(1, 1, 1)_{12}$ was the model that best fit the price data. According to their research, the predicted tomato prices were close to the actual prices apart from the months of March and April. The forecasts showed that uncertainty in the market leads the farmers into facing price risk. In addition to that, they concluded that it would be important to provide a reasonable price to the farmers and also to provide measures against risk factors of the price.

Vibas & Raqueño(2019) analysed monthly retail price movements of fruits and vegetables in the Philippines. The time series models used include ARIMA, SARIMA and Autoregressive Integrated Moving Average-Exogenous Variable (ARIMAx The best model for estimating cabbage prices was ARIMAx (3,2,1,x=pechay) and for estimating tomato prices was SARIMA $(2,1,1)(2,1,1)_{12}$. The monthly retail prices of the

fruits and vegetables had an increasing trend which was expected to continue increasing in subsequent years. They also concluded that the monthly retail prices for fruits not only depend on their previous prices but are also affected by the prices of other fruits. For vegetables, they noted that seasonality is a key factor in forecasting as their prices are influenced by the prices in the previous season of the same month.

Three univariate forecasting models are compared by Sukiyono et al., (2010) , namely ARIMA, Exponential Smoothing and Decomposition so as to predict cacao prices. The best cacao forecasting model was the ARIMA model. Accuracy indicators, namely MSD, MAD and MAPE were used in the selection of the best forecasting model. They reported that in addition to the ARIMA model, exponential smoothing and decomposition methods are also very useful in predicting agriculture prices to be used for decision making.

Dieng (2008) carried out a research aimed at comparing the forecasting performance of different time series models on vegetables including kales and cabbages. He stated that forecasting methods can be classified into parametric and non-parametric models. The parametric models included the Box and Jenkins, the ARIMA model, the naive model and exponential smoothing. The non-parametric model used the technique of spectral analysis. He reported that the findings of his study suggested that for the parametric models, the Box and Jenkins ARIMA model would be a good technique for forecasting selected vegetable prices for both producers and consumers in Senegal. The RMSE forecasting accuracy showed that the parametric models outperformed the non-parametric models.

Sukiyono & Janah(2019) carried out a study that aimed to choose the most accurate model for forecasting curly chili peppers. The data used showed the presence of trend and seasonal variations. The forecasting models used were Decomposition, Single Exponential Smoothing, Double Exponential Smoothing, Moving Average, and ARIMA. ARIMA (1,1, 9) was chosen as the most accurate model since selection was based on that which had the smallest MAPE, MSE and MAE values.

2.1 Conceptual Framework

Time series forecasting methods are either multivariate or univariate. In a univariate time series, only one variable varies over time i.e the single observations are recorded sequentially over time. Univariate time series models include exponential smoothing, Moving Average (MA), Autoregressive (AR), ARMA, ARIMA , ARCH and SARIMA models. A multivariate time series not only depends on past information but also incorporates the past of other input variables that vary over time. Multivariate models include ARIMAx model, Vector Autoregression (VAR) model and Vector ARMA.

In the case of forecasting exotic vegetable wholesale prices using a univariate time series model, the only key variable of interest is past price data.
However, if a multivariate time series model is to be used for this research, variables of interest would include the past price data of the specific exotic vegetable being forecasted and the price data of the other 2 vegetables.

3. Research Methodology

3. 1 Data Overview

The commodities used in this price forecasting research were kales and cabbages. They were chosen since they are one of the most widely consumed vegetables in the country. The wholesale price data used in this study is the average monthly prices of the aforementioned exotic vegetables in Nairobi County from January 2012 to November 2019 (95 observations). The wholesale price data was obtained from the National Farmers Information Services (NAFIS) website.

3.2 The ARIMA model

The Box-Jenkins (1970) analysis is a systematic methodology that involves identification, estimation and diagnostic checking.

3.2 The ARIMA model

The Box-Jenkins (1970) analysis is a systematic methodology that involves identifying, checking and using Autoregressive Integrated Moving Average (ARIMA) time series models.

3.2.1 The ARMA Model

The Autoregressive (AR(p)) and Moving Average (MA(q)) linear time series models are combined to form the ARMA (p, q) model. The AR (p) model aims to explain the future value of a random variable as a linear combination of past p values, a constant mean and a random error factor. The AR (p) can be mathematically expressed as:

$$X_t = \mu + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \ldots\ldots + \varphi_p X_{t-p} + \varepsilon_t \qquad (3.1)$$

X_t is the actual value and ε_t is the random error term at time t. $X_{t-1}, X_{t-2}, \ldots\ldots X_{t-3}$ are the variables at different time lags and $\phi_1, \varphi_2, \ldots\ldots, \varphi_p$ are the coefficients of the explanatory variables. The MA (q) model explains the future value of a random variable as a regression against past random errors.

The MA (q) model can be mathematically expressed as:

$$X_t = \mu + \vartheta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots\ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (3.2)$$

μ is the constant mean of the series, $\theta_1, \theta_2 \ldots\ldots, \theta_q$ are coefficients of the error terms and ε_t the random error term.

The ARMA model, with zero mean, is formed by combining the AR and MA linear models which can be mathematically expressed as:

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \ldots\ldots + \varphi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots\ldots + \theta_q \varepsilon_{t-q} \quad (3.3)$$

The backshift operator, B, is used by Box-Jenkins (1970) to make writing the equations easier. Therefore, $BX_t = X_{t-1}$, $B^2 X_t = X_{t-2}$ and the above expression can be written as:

$$(1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p) X_t = (1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q)\, \varepsilon_t \quad (3.4)$$

This can further be abbreviated by:

$$\varphi_p(B) X_t = \theta_q(B) \varepsilon_t \quad (3.5)$$

3.2.2 Stationarity Analysis

A statistical process is stationary if the mean and variance are constant for all t values. A time series whose values depend on t is not stationary. The Box-Jenkins methodology is for only stationary time series.

The ARIMA (p, d, q) model is a generalized ARMA model that makes a non-stationary time series model stationary by differencing the data points. The integer d represents the level of differencing required to make a time series stationary. Therefore, when d=0, the time series is already stationary without differencing meaning the ARIMA model is reduced to an ARMA (p, q) model.

3.3 The SARIMA model

Box and Jenkins (1970) proposed the Seasonal ARIMA (SARIMA) model to deal with non-stationary time series that exhibit seasonality. SARIMA models take the seasonality component of time series models into account. Seasonal differencing is used to remove

seasonality from a non-stationary time series. According to Ratnadip et al., (2013), a first order seasonal difference is the difference between an observation and the corresponding observation from the previous season, i.e.$Z_t = X_t - X_{t-s}$, where Z_t is the seasonally differenced series and s is the number of seasons per year. The SARIMA model incorporates both seasonal and non-seasonal factors and is denoted as SARIMA (p, d, q) x (P, D, Q) s. P is the seasonal AR order, Q the seasonal MA order and D the seasonal differencing

The model can be expressed as:

$$\varphi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D X_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad (3.6)$$

where:

$$(1-B^s)X_t = X_t - X_{t-s} \quad (3.7)$$

The non-seasonal components are:

$$\text{AR:} \quad \varphi_p(B) = (1 - \varphi_1 B - \ldots - \varphi_p B^p) \quad (3.8)$$

$$\text{MA:} \quad \theta_q(B) = (1 + \theta_1 B \ldots + \theta_q B^q \quad (3.9)$$

The seasonal components are:

$$\text{Seasonal AR:} \quad \Phi_P B^s = (1 - \Phi_1 B^s - \ldots - \Phi_Q B^{sP}) \quad (3.10)$$

$$\text{Seasonal MA:} \quad \Theta_Q(B^s) = (1 + \Theta_1 B^s + \ldots + \Theta_Q B^{sQ}) \quad (3.11)$$

According to the Box-Jenkins methodology, SARIMA modelling involves the following three-step iterative approach:

1. **Identification:** This model identification step first aims to identify the order of ARIMA(p, d, q) where p, d and q represent the orders of auto regression, differencing, and moving average respectively.

   This is done by first converting the data to a stationary time series, if data is non-stationary. The time series is differenced if it exhibits a trend component. Plots of the data, autocorrelation function and partial autocorrelation function are then observed.

The *Akaike Information Criterion (AIC)* and the *Bayesian Information Criterion (BIC)* can also be used for model identification. The model chosen is the one that minimizes either AIC or BIC.

2. **Estimation:** Once the order has been identified, we move forward with a time series (3.3) to estimate the parameters.
   The φ and θ parameters can be estimated using Maximum Likelihood Estimation (MLE). MLE finds the values of the parameters which maximise the likelihood of obtaining the observed data. For SARIMA models, this is similar to least squares estimation which is equivalent to MLE if the error terms can be assumed to be normally distributed.

3. **Diagnostic checking:** It is carried out by checking the ACF plots of the residuals and ensuring that they are patternless.
   The residuals are checked and the p and q values adjusted continuously until the residuals show no additional structure. If the model chosen is a good fit, then the estimates of the error terms $\hat{\varepsilon_t}$ are expected to be uncorrelated random variables with zero mean.

   The Ljung and Box 'portmanteau' test is essentially a test of lack of fit which can be carried out to determine whether the residual terms are uncorrelated. The null hypothesis is that the residuals are white noise meaning that the model does not show lack of fit. This can be mathematically expressed as:

$$n\,(n+2)\ \sum_{k=1}^{m} \frac{r_k^2}{n-k}\ \sim \chi_m^2 \qquad (3.12)$$

   Where n is the number of observations, m is the time lag and the Ljung and Box statistic follows the chi distribution with m degrees of freedom.

Once a suitable ARIMA model is chosen, it can then be used to forecast future values.

3.4. Exponential Smoothing

Brown(1959) defines exponential smoothing as a type of time series forecasting that uses past historical data to predict the future by giving more weight to recent observations than older observations.

3.4.1. Additive model vs. Multiplicative model

The additive model is used when the seasonal variations tend to be constant over time regardless of the overall level of the series while the multiplicative model is used mostly when the seasonality tends to increase over time.

In addition to that, seasonality may be additive if each season differs from the other by a specific amount or it may be multiplicative where each season differs from the other by a specific percentage.

For the additive trend, the trend grows or decays linearly while for the multiplicative trend, the trend grows or decays exponentially over time.

The additive model:

$$X_t = Level + Trend + Seasonality + Noise \tag{3.13}$$

The multiplicative model:

$$X_t = Level * Trend * Seasonality * Noise \tag{3.14}$$

3.4.2. Holt-Winters Exponential Smoothing (HWES)

Holt (1957) and Winters (1960) developed the Holt's model to incorporate seasonality component. This time series forecasting method is used when the data exhibits both seasonality and trend components. This model consists of one forecasting equation and three smoothing equations (one level equation, one trend equation and one seasonal equation).

The additive model:

The level equation:

$$L_t = \alpha(X_t - S_{t-M}) + (1-\alpha)(L_{t-1} + T_{t-1}) \tag{3.15}$$

The trend equation:

$$T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)\,(T_{t-1}) \tag{3.16}$$

The seasonal equation:

$$S_t = \delta(X_t - L_{t-1} + T_{t-1}) + (1 - \delta)\, S_{t-M} \quad (3.17)$$

Hence,

Forecast = estimated level + trend + seasonality at most recent point in time

$$F_{t+k} = L_t + kT_t + S_{t+k-M} \quad (3.18)$$

Where:

L_t: level at time t

T_t: trend at time t

S_t: seasonality at time t

α: smoothing parameter for the level

γ: smoothing parameter for the trend component

δ: smoothing parameter for the seasonality component

X_t: most recent data point

M: seasonal period

The multiplicative model:

The level component:

$$L_t = \alpha(X_t / S_{t-M}) + (1 - \alpha)\,(L_{t-1} + T_{t-1}) \quad (3.19)$$

The trend component:

$$T_t = \gamma(L_t - L_{t-1}) + (1 - \gamma)\,(T_{t-1}) \quad (3.20)$$

The seasonal component:

$$S_t = \delta(X_t/(L_{t-1} + T_{t-1}) + (1 - \delta)\, S_{t-M} \quad (3.21)$$

Forecast = most recent estimated (level + trend) x seasonality at most recent point in time

$$F_{t+k} = (L_t + kT_k)x\, S_{t+k-M} \quad (3.22)$$

3.5 Forecast Performance Measures

Performance measures are used to check and compare the performance and forecast accuracy of different models. In each of the performance measures below:

X_t= actual observation at time t

$\hat{X}_t$= forecasted value at time t

$e_t = X_t - \hat{X}_t$= forecast error at time t

n = size of observations

3.5.1 The Mean Absolute Error (MAE) & The Mean Absolute Percentage Error (MAPE)

It measures the average absolute deviation of forecasted values from the actual observed values and is given by:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|X_t - \hat{X}_t| = \frac{1}{n}\sum_{t=1}^{n}|e_t| \tag{3.23}$$

MAPE is the percentage of the average absolute deviation (MAE) that has been observed.

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{X_t - \hat{X_t}}{X_t}\right| x\ 100 = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{e_t}{X_t}\right| x\ 100 \tag{3.24}$$

3.5.2 The Mean Squared Error (MSE)

It measures the average squared deviations of the forecasted values from the actual observed values and emphasizes that deviations are more affected by large individual errors than small errors.

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n}(X_t - \hat{X_t})^2 = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \tag{3.25}$$

The Error Sum of Squares (SSE) measures the total squared deviations of the forecasted values from the actual observed values

$$\text{SSE} = \sum_{t=1}^{n} e_t^2 \tag{3.26}$$

3.5.3 The Root Mean Squared Error (RMSE)

This is the square root of MSE and is mathematically defined as:

$$\text{RMSE} = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \tag{3.27}$$

The smaller the obtained MAE and/or RMSE value, the better the forecast and the bigger the value the lower the forecasting power of the model.

4. Data Analysis

The commodities used in this price forecasting research were kales and cabbages. They were chosen since they are one of the most widely consumed exotic vegetables in the country. The wholesale price data used in this study is the average monthly prices of the aforementioned exotic vegetables in Nairobi County from January 2012 to December 2020 (96 observations). The wholesale price data (Ksh per Kg) was obtained from the National Farmers Information Services (NAFIS) website.

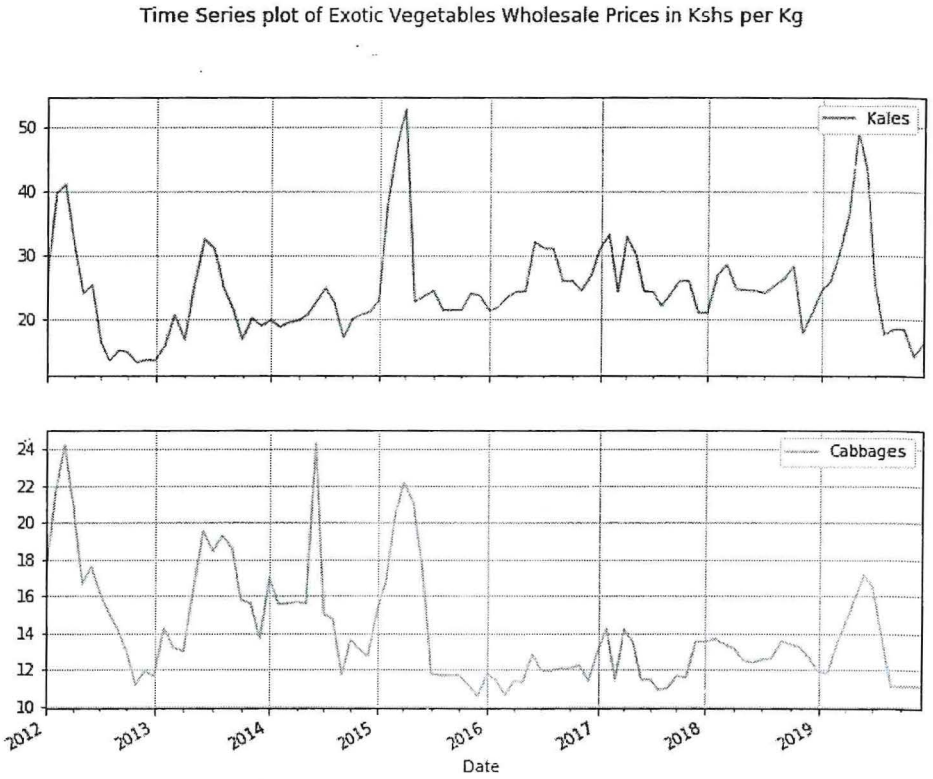

Figure 1. Wholesale prices of the exotic vegetables in Nairobi

Based on the historical data, the prices of the exotic vegetables are variable and unstable (Figure 1). Figure 2 provides a good visualization of the price instability. For each of the exotic vegetables, prices start declining from the month of June due to excess supply of the produce and poor weather. The prices start picking up from October until the first quarter of the year.

Hence, there is evidence of seasonality in the wholesale prices of the exotic vegetables. As a result, models such as SARIMA and Holt - Winters Exponential Smoothing are used to forecast price data that exhibit seasonal variations.

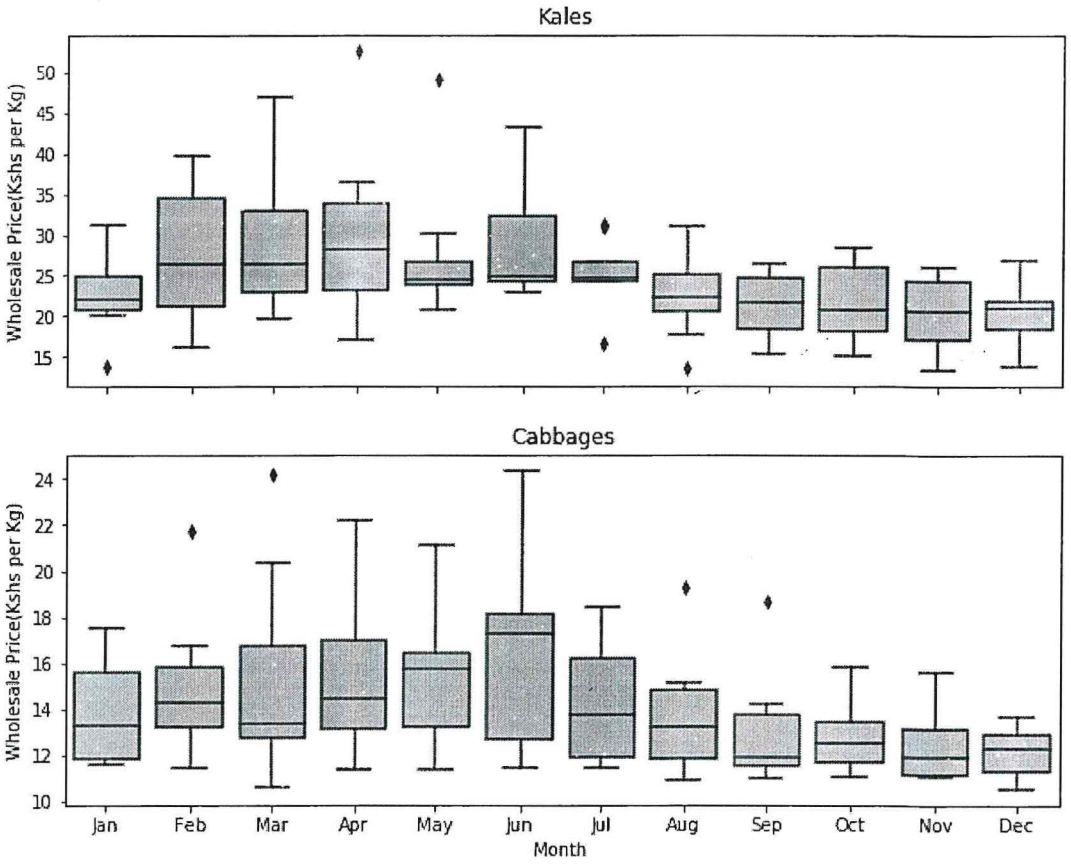

Figure 2. Monthly distribution of wholesale prices of the exotic vegetables in Nairobi

Kales are sold in 50kg bags and cabbages in 126kg bags. Tables 1 and 2 show the monthly summaries for the period between January 2012 and December 2019 for cabbages and kales respectively. The values are expressed in Ksh per Kg.

Table 1. Real Cabbage Prices by months (Ksh/Kg) (2012-2019)

Months	Mean	Minimum	Maximum	Std. Deviation
January	13.95	11.63	17.50	2.36
February	14.94	11.43	21.74	3.27
March	15.27	10.63	24.20	4.67
April	15.64	11.39	22.19	3.85
May	15.38	11.35	21.11	3.02
June	16.60	11.47	24.33	4.29
July	14.24	11.47	18.46	2.65
August	13.78	10.91	19.31	2.69
September	13.01	11.02	18.65	2.54
October	12.81	11.10	15.81	1.51
November	12.40	11.11	15.60	1.55
December	12.20	10.56	13.71	1.14

Table 2. Real Kale Prices by months (Ksh/Kg) (2012-2019)

Months	Mean	Minimum	Maximum	Std. Deviation
January	22.65	13.70	31.10	5.09
February	27.59	16.08	39.70	8.70
March	29.39	19.60	47.00	9.82
April	29.93	16.96	52.60	11.30
May	27.70	20.80	49.16	9.07
June	28.58	22.87	43.12	6.97
July	25.24	16.48	31.12	4.60
August	22.38	13.64	31.07	5.21
September	21.35	15.28	26.40	4.08
October	21.56	14.98	28.31	4.80
November	20.17	13.40	26.02	4.69
December	20.41	13.80	26.88	4.07

4.1 SARIMA

4.1.1 Stationarity test on the time series data

This step involves identifying whether the data is stationary. According to Figure 1, none of the vegetables show any significant trend which is an indication that the data may be stationary. To ascertain this, the Augmented Dickey Fuller (ADF) test was carried out for each of the vegetables to determine whether they are stationary or not.

Table 3. Testing for stationarity on the time series price data

Vegetable	ADF Test	p-value	Hypothesis (H_1)	Decision
Kales	-5.04284	1.82446e-05	Stationary	Reject H_0
Cabbages	-4.24851	0.000544	Stationary	Reject H_0

H_0: Time Series has a unit root

H_1: Time Series is stationary

A p-value of less than 5% means that the null hypothesis is rejected i.e. the series is stationary.

Since the vegetables' data is already stationary as per the ADF test there is no need for non-seasonal differencing. However, seasonal differencing was carried out since the data exhibits seasonality. Thereafter, ACF and PACF graphs of the seasonally differenced data were then plotted.

4.1.2 Cabbages

1. **Model Identification and Estimation**

In Figure 3, the ACF slowly decays to zero with the significant peak at lag 1 suggesting an MA(1). The spike at lag 1 is assumed to explain all other higher order autocorrelations. The PACF cuts off at lag 1 hence suggesting an AR (1). After seasonal differencing once, Figure 4 shows the ACF has significant spikes. The lag 12 spike represents the seasonality, suggesting a seasonal MA (1). The PACF of the seasonally differenced cabbage data has significant positive and negative spikes. Consequently, there is no clarity for the AR term of the seasonal part of the model. Two alternatives

for the seasonal part are the model which the seasonal AR term is added (1, 1, 1) and the model which the seasonal AR is not added (0, 1, 1).

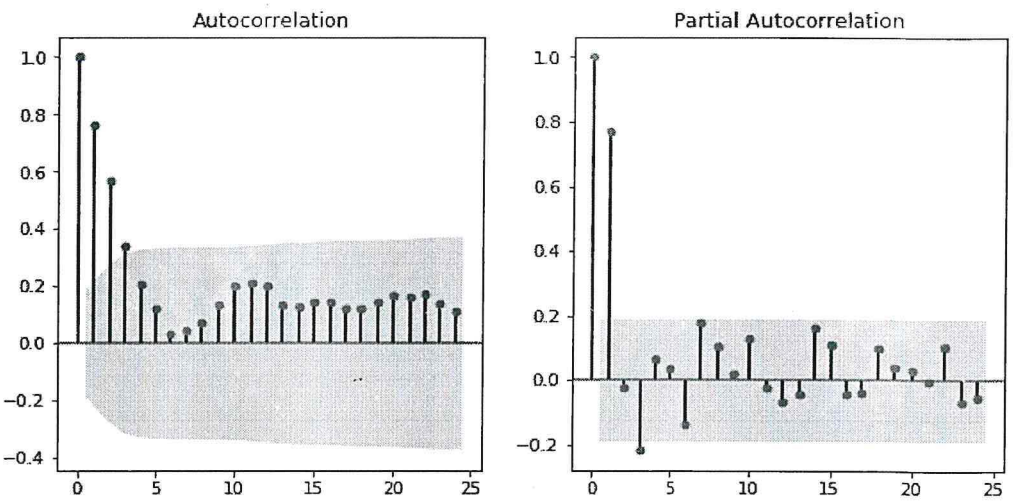

Figure 3. ACF and PACF of the monthly cabbage price data without differencing

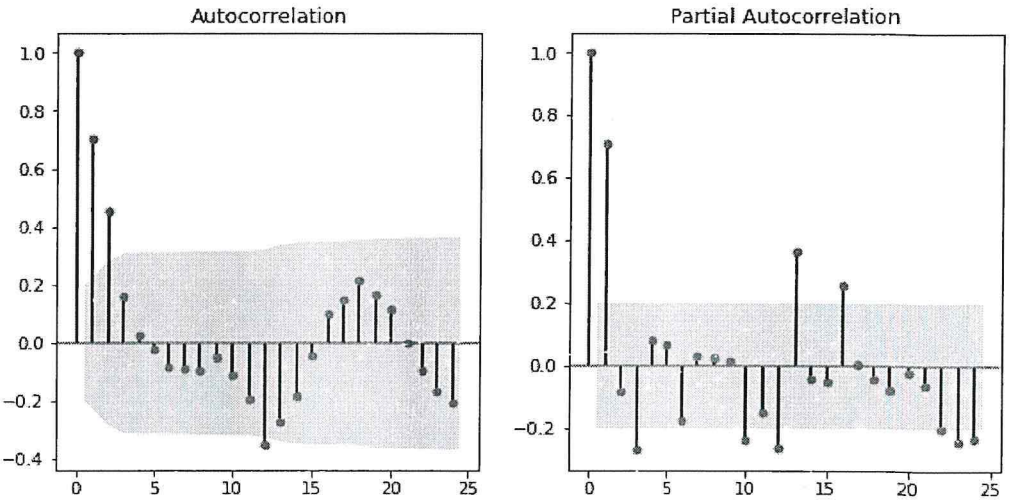

Figure 4. ACF and PACF of the monthly differenced cabbage price data

2. Model Selection

Selecting the best SARIMA model to be used in forecasting the monthly cabbage prices, four alternative models were analysed. The selected model should have the lowest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, and in this case the SARIMA$(1,0,0)(1,1,1)_{12}$model was selected as the model of best fit.

Table 4. (AIC), (BIC) and considering different SARIMA$(p,d,q)(P,D,Q)_{12}$models for the Cabbage price data

SARIMA Model	AIC	BIC
$(1,0,0)(0,1,1)_{12}$	378.663	385.955
$(1,0,0)(1,1,1)_{12}$	380.481	390.205
$(1,0,1)(0,1,1)_{12}$	380.488	390.211
$(1,0,1)(0,1,1)_{12}$	382.303	394.457

The selected SARIMA$(1,0,0)(1,1,1)_{12}$model was analysed to check whether the residuals are independent. According to the Normal Q-Q plot on Figure 5, the residuals almost have a straight line meaning there's no systematic departure from normality. The correlogram shows no autocorrelation in residuals. This therefore shows that the selected SARIMA model is the best model for forecasting monthly cabbage prices.

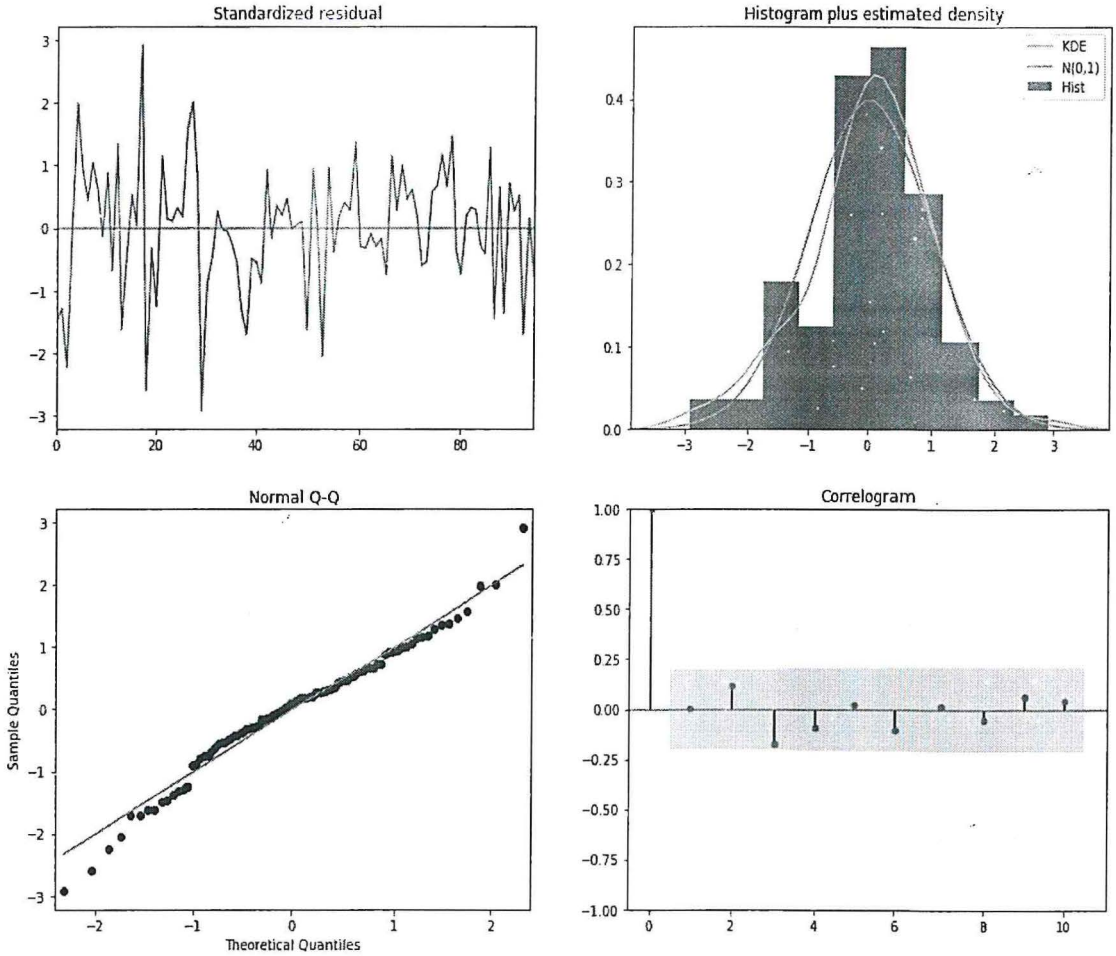

Figure 5. Graphical diagnostics for assessing the SARIMA$(1,0,0)(1,1,1)_{12}$ model fit

4.1.3 Kales

1. **Model Identification and Estimation**

In Figure 6, the ACF slowly declines to zero, with significant peaks at lag 1 and lag 2. The spike at lag 2 is effectively explained by the lag 1 autocorrelation hence suggesting MA(1) term. The PACF cuts off at lag 1 suggesting an AR(1) model.

The ACF and PACF plots after seasonal differencing the Kales price data are displayed on Figure 7. The ACF shows a cluster of negative seasonal spikes around lag 12.

The PACF has significant spikes near lags 12 and 24, suggesting a seasonal MA(1)component. Therefore, possible alternatives for the seasonal part of the SARIMA model are (0,1,1) and (1,1,1).

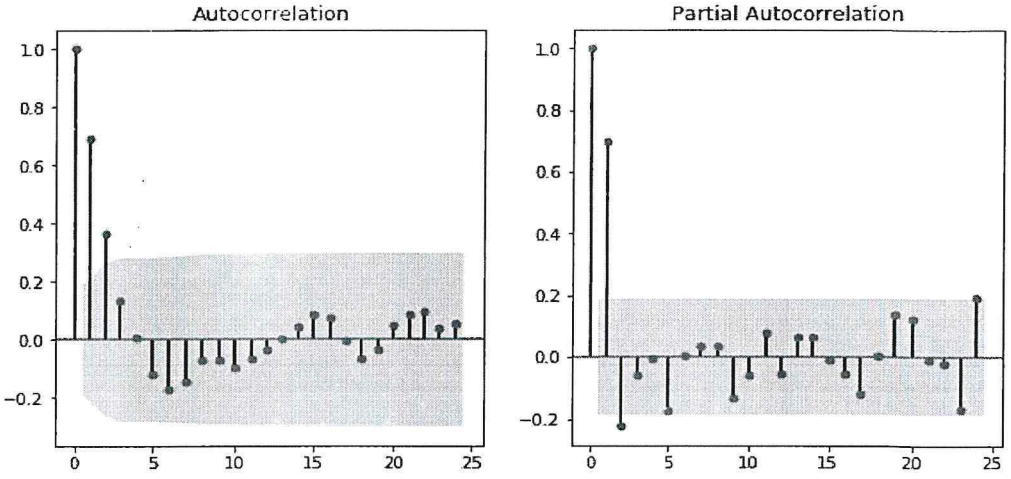

Figure 6. ACF and PACF of the monthly kale price data without differencing

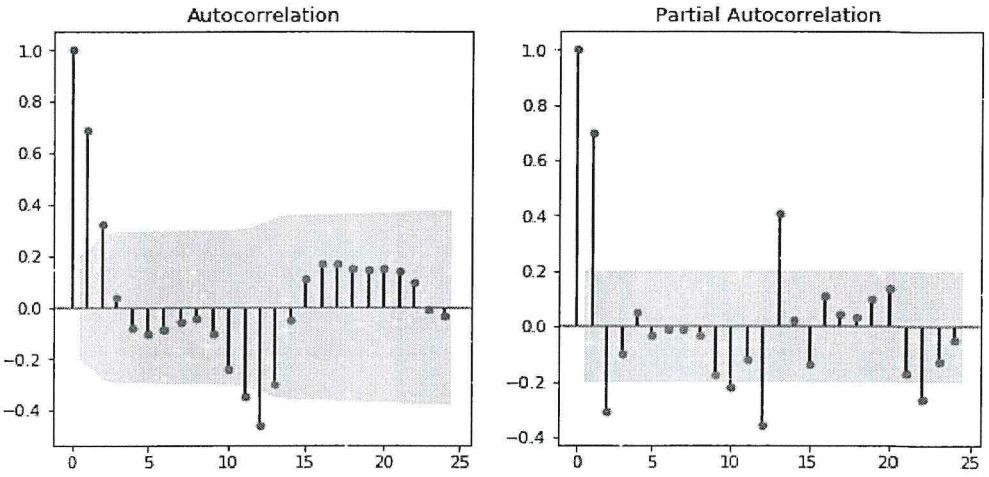

Figure 7. ACF and PACF of the monthly differenced kale price data

2. **Model Selection**

AIC and BIC were used to analyse four possible SARIMA models. The results of the analysis are shown in Table 5. The SARIMA $(1,0,1)(0,1,1)_{12}$has the lowest AIC and BIC values hence selected as the model of best fit.

Table 5. (AIC), (BIC) and considering different SARIMA$(p, d, q)(P, D, Q)_{12}$models for the Kale price data

SARIMA Model	AIC	BIC
$(1,0,1)(0,1,1)_{12}$	544.264	556.418
$(0,0,1)(0,1,1)_{12}$	552.180	559.472
$(0,0,1)(1,1,1)_{12}$	545.311	557.465
$(1,0,0)(1,1,1)_{12}$	553.875	563.598

The residuals of the SARIMA $(1,0,1)(0,1,1)_{12}$model were analysed in order to determine whether they are independent. Figure 8 shows that the residuals appear to be normally distributed and have no autocorrelation hence SARIMA $(1,0,1)(0,1,1)_{12}$is the best SARIMA model for forecasting monthly Kales'prices

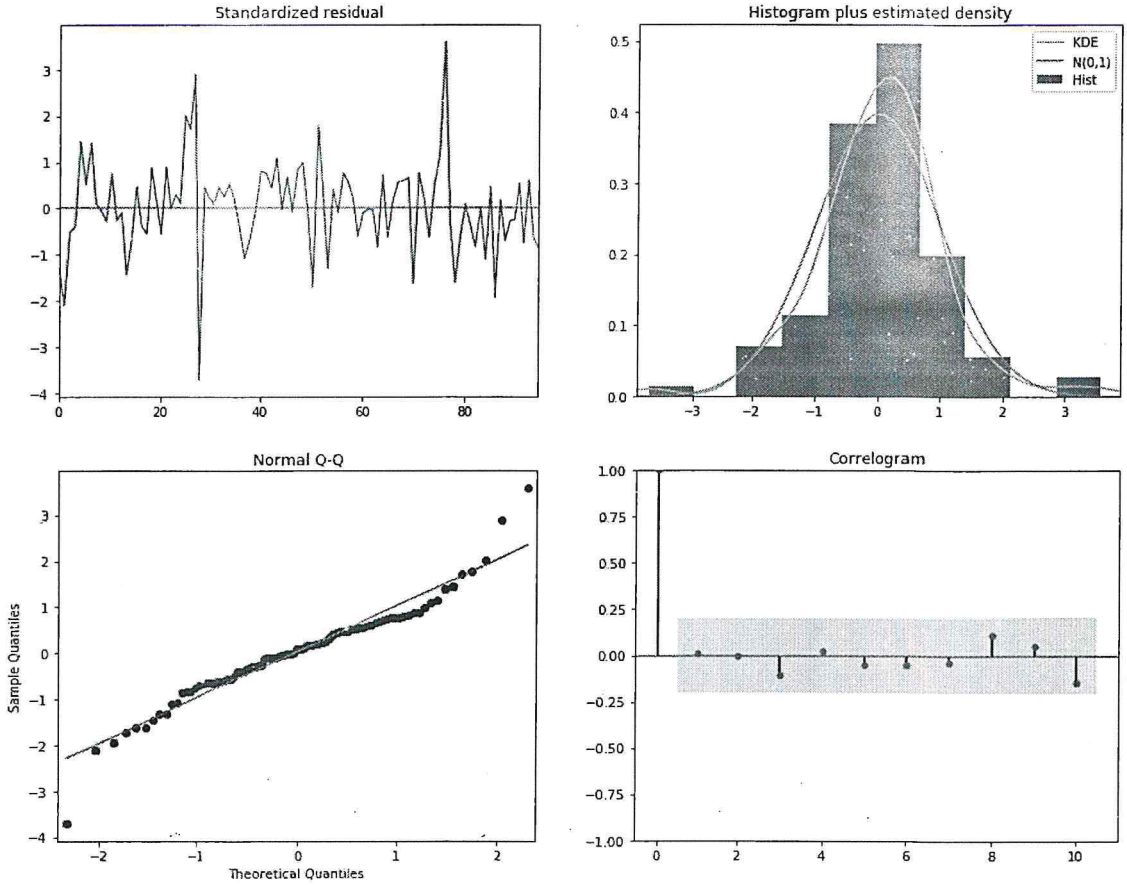

Figure 8. Graphical diagnostics for assessing the SARIMA $(1, 0, 1)(0, 1, 1)_{12}$ model fit

4.2 Holt-Winters Exponential Smoothing (HWES)

The data was split into two datasets: train and test. Data from January to December 2019 (96 observations) was used for the training phase and data from January to December 2020 (12 observations) was used for the testing phase.

The best parameter values for α (level), γ (trend) and δ (seasonal) were estimated using the train dataset and automatically generated by Python 3 software, prior to modelling using the Holt-Winters exponential smoothing method, as shown in Table 8.

The exotic vegetables do not have any significant trends as it was concluded that their price data were both stationary, hence γ=0.

Table 6. Best Parameter Values for the Exotic Vegetables

Vegetable	Best α	Best δ
Cabbages	0.6842105	0.3157895
Kales	0.9999957	4.2548e-06

4.3 Forecasting

The observed price data of the exotic vegetables from January 2012 to December 2019 was used to forecast the prices from January 2020 to December 2021, as shown in Figure 9 and Figure 10. Table 6 gives a more intricate depiction of the forecasts.

The predicted prices of the year 2020 were then compared to the actual observed prices of the same year. The forecasting performance measures i.e. MAE, MSE and RMSE were then used to select the best forecasting model for each exotic vegetable as shown in Table 7.

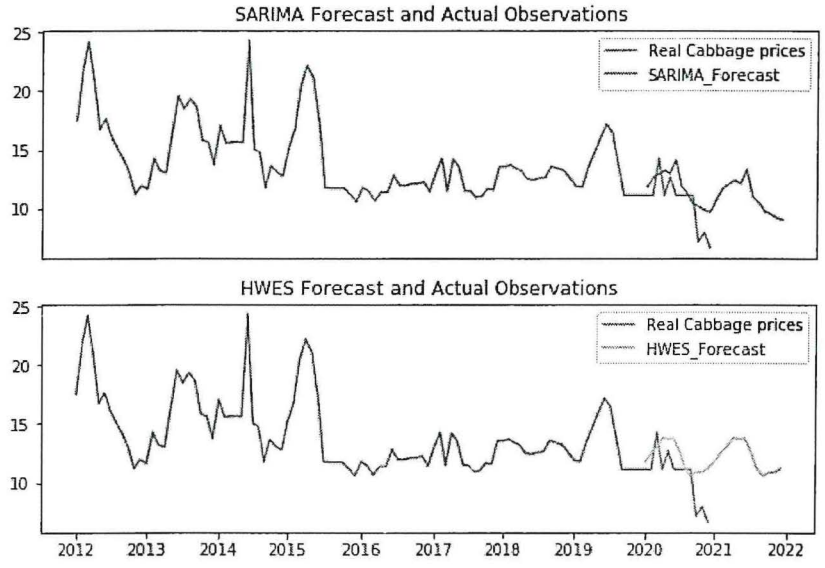

Figure 9. Actual Observations and Forecasts for the Real Cabbage Prices

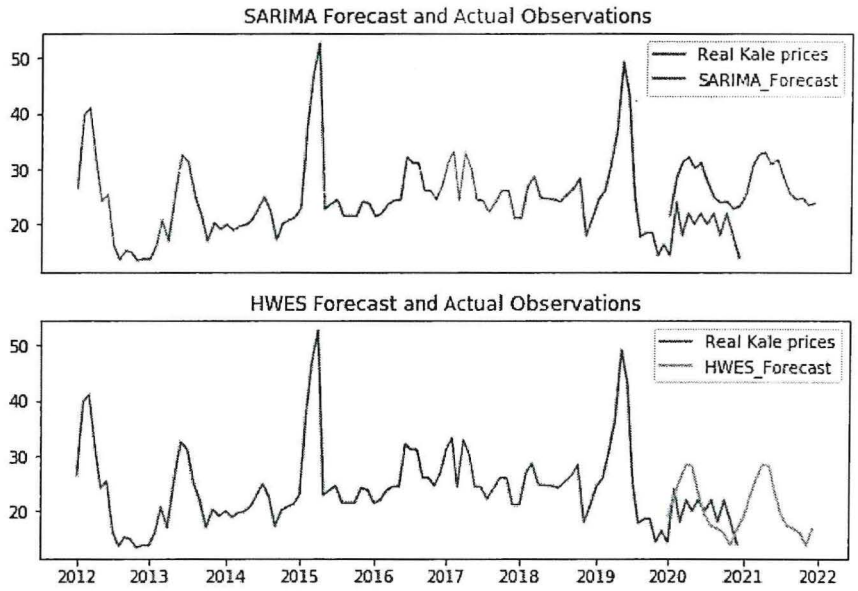

Figure 10. Actual Observations and Forecasts for the Real Kale Prices

Table 7. Real Vegetable Prices observed in 2020 and the forecasted prices obtained from the respective SARIMA models and HWES method (Ksh/Kg)

	Cabbage Prices			Kale Prices		
Date	Observed	Forecast		Observed	Forecast	
		SARIMA	HWES		SARIMA	HWES
2020-01-01	11.11	8.41	11.80	14.35	16.99	17.92
2020-02-01	11.11	9.80	12.53	24.00	24.66	22.24
2020-03-01	14.29	11.01	12.98	18.00	26.78	23.94
2020-04-01	11.11	10.98	13.77	22.00	28.35	24.72
2020-05-01	12.70	11.18	13.70	20.00	26.50	21.27
2020-06-01	11.11	12.02	13.73	22.00	27.69	21.58
2020-07-01	11.11	10.16	12.64	20.00	24.60	18.76
2020-08-01	11.11	9.84	11.21	22.00	22.35	17.07
2020-09-01	11.11	9.25	10.57	18.00	21.04	16.39
2020-10-01	7.14	8.41	10.85	22.00	21.72	16.90
2020-11-01	7.94	8.25	10.91	18.00	20.11	16.02
2020-12-01	6.67	7.91	11.20	14.00	20.00	16.46

Table 8: Performance measures of the selected SARIMA models and HWES method for forecasting the exotic vegetables for the period January - December 2020

Vegetable	Forecasting method	MAE	MSE	RMSE
Cabbage	SARIMA $(1,0,0)(1,1,1)_{12}$	1.5575	3.4666	1.8619
	HWES	1.9233	5.4081	2.3255
Kale	SARIMA$(1,0,1)(0,1,1)_{12}$	7.2650	62.658	7.9157
	HWES	2.75	10.394	3.2240

Based on the forecasting results shown on Table 6, the visualizations on Figure 9 and Figure 10, the best forecasting method for cabbages and kales are the SARIMA $(1,0,0)(1,1,1)_{12}$ model and Holt-Winters Exponential Smoothing method, respectively (Table 7).

5. Conclusion

The time series data for the exotic vegetables (starting from January 2012 to December 2020) do not have any increasing or decreasing trends and hence are concluded to be stationary. The actual observed prices were more or less lower than the forecasted prices from both the SARIMA and HWES methods. Despite the models taking seasonality into account, they still were not able to accurately predict the prices of the exotic vegetables, hence better seasonal models should be developed to improve the forecast accuracy.

The difference in the actual and forecasted prices could possibly be linked to the economic effects of the Covid-19 Global Pandemic. The pandemic has rendered many Kenyans jobless and consequently led to them turning to farming as a source of income. In addition to that, the abrupt and unanticipated closure of schools in which have an estimated 17 million learners 2020 meant farmers were stuck with more produce than they could sell. Therefore, the increased supply of vegetables in the market means that buyers had more bargaining power and hence would only offer a fraction of the normal price.

Forecasting future prices of vegetables is nevertheless important to the Government of Kenya, the buyers and particularly to farmers as they can use this information to help them maximise their profits and minimise their losses.

References

Ratnadip, A. & Agrawal, R.K. (2013). An Introductory Study on Time Series Modelling and Forecasting *Lap Lambert Academic Publ*

Food and Agriculture Organisation of the United Nations. (2015). The economic lives of smallholder farmers: An analysis based on household data from nine countries. http://www.fao.org/3/a-i5251e.pdf

Ministry of Agriculture, Livestock, Fisheries & Livestock. (2019). Agricultural Sector Transformation and Growth Strategy: Towards Sustainable Agricultural Transformation and Food Security In Kenya (2019- 2029). http://www.kilimo.go.ke/wp-content/uploads/2019/01/ASTGS-Full-Version-1.pdf

Partnership for African Social & Governance Research. (2019). Creating Employment in Horticulture Sector in Kenya: Productivity, Contracting and Marketing Policies. https://includeplatform.net/wp-content/uploads/2019/08/Creating-employment-in-horticulture-sector-in-Kenya-Productivity-contracting-and-marketing-policies.pdf

Agriculture Food Authority. (2018). Horticulture Validated Report (2017-2018). https://agricultureauthority.go.ke/?page_id=1331

Research Solutions Africa.(2015, December). Report on a Market Study on Fresh Vegetables in Kenya: Consumer's Survey. https://www.agroberichtenbuitenland.nl/binaries/agroberichtenbuitenland/documenten/rapporten/2015/12/part-1---report-of-a-study-on-fresh-vegetables-market-in-kenya/part-1--report-of-a-study-on-fresh-vegetables-market-in-kenya/Fresh+Vegetables+Market+in+Kenya+-+Part+1+-+Desk+Review.pdf

Assis, K., Amran, A. & Remali, Y. (2010). Forecasting Coco Bean Prices Using Univariate Time Series Models. *Journal of arts science & commerce*, Vol (1), 71-79.

Gathondu, E. K. (2014). Modelling of wholesale prices for selected vegetables using time series models in Kenya.

Robert Mathenge Mutwiri.(2019). Forecasting of Tomatoes Wholesale Prices of Nairobi in Kenya: Time Series Analysis Using Sarima Model. *International Journal of Statistical Distributions and Applications.* Vol. 5, No. 3, 2019, pp. 46-53. Doi: 10.11648/j.ijsd.20190503.11

Boateng, F. O., Amoah-Mensah, J., Anokye, M., Osei, L., & Dzebre, P. (2017). Modeling of tomato prices in Ashanti region, Ghana, using seasonal autoregressive integrated moving average model. *Journal of Advances in Mathematics and Computer Science,* 1-13.

Adanacioglu, H. & Yercanm, M. (2012). An analysis of tomato prices at wholesale level in Turkey: an application of SARIMA model

Ketut Sukiyono, Musriyadi Nabiu , Bambang Sumantri , R.R. Novanda , Nyayu Neti Arianti , Sriyoto , M. Zulkarnain Yuliarso , Redy Badrudin , M. Mustopa Romdhon , H. Mustamam. (2018). Selecting an accurate Cacao Price Forecasting model. *IOP Conf. Series: Journal of Physics*: Conf. Series 1114 (2018) 012116 doi :10.1088/1742-6596/1114/1/012116

Dieng, A. (2008). Alternative Forecasting Techniques for Vegetabie Prices in Senegal.Institut Sénégalais de recherches agricoles (ISRA), Vol (1), No.3, 5-9.

Ketut Sukiyono, Miftahul Janah (2019). Forecasting Model Selection of Curly Red Chili Price at Retail Level. *Indonesian Journal of Agricultural Research* Vol. 02, No. 01, 2019 | 01 - 12

G.E.P. Box, G. Jenkins, "Time Series Analysis, Forecasting and Control", Holden-Day, San Francisco, CA, 1970.

Brown, R. G. (1959). *Statistical forecasting for inventory control*. McGraw/Hill

C. C. Holt, Forecasting Seasonals and Trends by Exponentially Weighted Averages, O. N. R. Memorandum 52/1957, Carnegie Institute of Technology, 1957

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science, 6*, 324–342.