



Seasonal Autoregressive Integrated Moving Average (SARIMA) for Melon (*Cucumis melo*) Yield from 2011 to 2020 Based on Planting Date Period

MOHAMMAD REZA NAROUI RAD^{1*}, SOMAYEH SOLTANI-GERDEFARAMARZI²,
HAMMED ADEBAYO AKANO³ and AMMARA CHEEMA⁴

¹Horticulture Crops Research Department, Sistan Agricultural and Natural Resources Research and Education Center, AREEO, Zabol, Iran.

²Department of Water Science and Engineering, Collage of Agriculture and Natural Resources , Ardakan University, Ardakan, Iran.

³Federal University of Technology Akure, Ondo State, Nigeria.

⁴Department of Mathematics, Air University, Islamabad, Pakistan.

Abstract

The main goal of time series modelling is to collect and analyze past values to develop appropriate models that describe the inherent structure and characteristics of the series about the planting date of melon. The data used in this study are the melon yield based on twenty-nine days of planting date in March from the first to the last planting date of March 2011 to 2020, to evaluate the prediction performances of the models, two indices: the root mean squared error (RMSE) and mean absolute percentage error (MAPE) were used to compare the forecasting capabilities of the SARIMA models, Excel was used first for preparing as CSV file and then changed to time series data to make the dataset of yield and R version 4.0.2 (the R Development Core Team) was used to perform ETS and SARIMA models For planting date time series based on yield, seasonal autoregressive integral moving average models (SARIMA) were constructed. The Student's t-test was one of the statistical tests used to evaluate the effectiveness and validity of the SARIMA models Through the use of SARIMA models, it is feasible to create synthetic records that maintain the statistical properties of the historical record. Finally, the results can be applied to different planting dates based on yield. Based on the Bayesian Information Criteria (BIC) values and the overall highest R^2 values of 0.94, the optimal SARIMA



Article History

Received: 13 August 2023

Accepted: 20 November 2023


Keywords

Forecasting;
Melon; Prediction;
Precision agriculture.

CONTACT Mohammad Reza Naroui Rad ✉ mr.narouirad@areeo.ac.ir 📍 Horticulture Crops Research Department, Sistan Agricultural and Natural Resources Research and Education Center, AREEO, Zabol, Iran.



© 2023 The Author(s). Published by Enviro Research Publishers.

This is an  Open Access article licensed under a Creative Commons license: Attribution 4.0 International (CC-BY).

Doi: <https://dx.doi.org/10.12944/CARJ.11.3.06>

(0,1,1) model was chosen. No limitations were found. It is the first time that this mathematical method was used to analyze time series data for 10 years of planting date of melon and it could be useful for agronomists and horticulturists to choose the best model.

Introduction

In tropical and subtropical regions, melons, *Cucumis melo* L., are also widely grown as significant horticulture crops, which are also grown extensively in temperate climates. With an annual production of 16,009,584 tonnes, China is the world's top producer of melons, followed by Turkey and Iran.¹ Cucurbits account for more than 50% of all vegetable production in Iran. Melon is the most significant crop among them. According to statistics from 2011, over 75,000 acres of melons were grown by Iranian farmers in 2018, yielding a total of 1.31 million tons.¹ Different types of melon cultigens are grown in Iran. The primary commercial types are melons of the Firoozi variety, grown in south-eastern Iran. Applying various modeling techniques to forecast the expected harvest is crucial for managing production and statistical estimations. It can only be done perfectly with sophisticated calculations of the production method and a perfect valuation of the various factors involved.² Due to urbanization and environmental damage, arable land has been shrinking, making it urgently necessary to find a solution to increase its productivity. Therefore, the study of potential yield has drawn the attention of researchers, and it is reasonable to explore grain production potential to the fullest extent to increase food security.^{3,4} In a time series analysis, a forecasted variable (in this case, yield) is modeled as a function of time, as in $Y_t = f(t) + \epsilon_t$, where Y_t is the yield for year t , $f(t)$ is a function of time t , and ϵ_t denotes error (i.e., the discrepancy between the observed yield and the forecasted yield for year t). Yield can be predicted for year $t+1$ once a functional relationship between yield and time (i.e., a time series model) has been developed. When creating this model, the first step is determining whether the time series being studied is stationary or nonstationary.⁵ Time series analysis has long been used in grain yield analysis.^{4,5} The advantage was that it only required a small amount of data, which could be easily accessed. A forecasting model was developed utilizing time series analysis and historical yield data based on the novel notion of agricultural yield. The most crucial method

and model parameters were thoroughly explained. It seeks to provide a more straightforward and accurate way of estimating the prospective yield.⁶

Materials and Method

Source of Materials

The data used in this study are the melon yield based on twenty-nine days of planting date in March from the first to the last planting date of March 2011 to 2020. The model was tested in Sistan and Baluchestan Province, data used in the paper was yield per unit(t/ha), which was obtained from 2011 to 2020, the yield of melon was collected from the Sistan Agricultural and Natural Resources Research and Education Center. From these, data from 2011 to 2020 were used to construct the ETS and SARIMA models. Data from all days of March 2011 to all days of March 2020 were used to evaluate the forecasting performances of these models.

Seasonal Autoregressive Integrated Moving Average (SARIMA) model

The SARIMA model defined constitutes a straightforward extension of the non-seasonal autoregressive-moving average (ARMA) and autoregressive integrated moving average (ARIMA) models presented, The generalized model is called an autoregressive moving average (ARMA) model and has both elements of autoregressive (AR) and moving average (MA) processes.¹⁴

SARIMA models are written as: ARIMA (p, d, q) (P, D, Q)_m, Where (p, d, q) and (P, D, Q) m are the non-seasonal and seasonal parts of the model, respectively. The parameter m is the number of periods per season. The seasonal part of the model is very similar to the non-seasonal part, but it is involved in backshifts of the seasonal period. Using the available dataset, the ARIMA model is finalized by changing the values of p, d and q. To determine the parameters of an ARIMA model, Akaike's Information Criterion (AIC) is widely used. It is given by $AIC(p) = n \ln(RSS/n) + 2K$. Where n is the number of data points and RSS is the residual sums of squares. The

model with the minimum AIC value will be selected as the best forecasting model. Another method to determine the appropriate parameters of an ARIMA model is to analyze autocorrelation function (ACF) and partial autocorrelation function (PACF) plots.⁷ The parameters of the SARIMA model were estimated based on the autocorrelation function (ACF) graph and partial autocorrelation (PACF) plots⁸ meanwhile, the (time series) function of R 3.4.2 software was used to select the best SARIMA model according to either the minimum BIC.

Exponential Smoothing (ETS) Model

In one study⁵ approaches are the foundation for the ETS technique, available in the R program environment through the forecast package. According to,⁹ it contains three primary parameters: the error, trend, and seasonal components, which might be additive (A), multiplicative (M), or none (N). To fit exponential models with multiplicative components, we used the automatic selection of the ETS models. We considered potential alternatives before choosing the best-performing model to simulate the data.¹⁰ According to the minimum of the Akaike information criterion (AIC), the corrected Akaike information criterion (AICc), or the Bayesian information criterion (BIC), the best model was selected.¹¹ The residual error sequence's status as a white-noise sequence was determined using the Ljung-Box Q test.³ Several theoretical and experimental considerations have been made to select an appropriate smoothing constant value. An estimate is only sometimes better if the smoothing constant is big. For instance, using a more incredible amount can result in significant forecast mistakes, while using a smaller value can result in a trend not responding as rapidly. Therefore, choosing the smoothing constant's value is crucial.¹² A forecasting model was developed utilizing time series analysis and historical yield data based on the novel notion of agricultural yield. It seeks to build a more straightforward and accurate way of forecasting the prospective yield and make it more useful.

Performance Statistic Index

To evaluate the prediction performances of the models, two indices: the root mean squared error (RMSE) and mean absolute percentage error

(MAPE) were used to compare the forecasting capabilities of the SARIMA models. The main drawback of the MAPE criteria is the adverse effect resulted from small actual values. If the actual value is small, the MAPE will be very large even if the difference between the predicted and actual value is small. For this reason, an adapted MAPE (AMAPE) is introduced for some cases, the formulas for calculation are as follows.

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2}$$

$$AMAPE = \left(\frac{1}{N} \sum_{t=1}^N \left(\frac{|\hat{y}_t - y_t|}{\frac{1}{N} \sum_{t=1}^N y_t} \right) \right) \cdot 100\%$$

Data Preparing and Evaluation

Excel was used first for preparing as csv file and then changed to time series data to make the dataset of yield and R version 4.0.2 (the R Development Core Team) was used to perform ETS and SARIMA models.¹⁴ Figures depicted in R environment and results below of 0.05 statistical level are considered significant.

Results and Discussion

In this study, a graph related to self-correlated data for planting date was drawn to determine the trend of changes, to determine the seasonality and identification of outflow data, and then using information related to the past years of planting date in the station of Zahak, In regions where meteorological data has a sufficient statistical duration, serial-time models are applied. Numerous time series models help analyze changes and simulate outcomes. We can precisely highlight the seasonal model of Sarima and the typical model of Sarima among these. The first stage should involve looking at the average stability and static data. It shrinks in on itself. Data were static by differentiation in the current investigation (Figures 1 and 2). station flow correlation diagrams and partial self-correlation diagrams. Figures 3 to 5 show the before and after of differentiation so that the proper coefficients for Extracted p, q can be applied.



Fig. 1: Location of Zahak in Sistan and Baloochestan Province

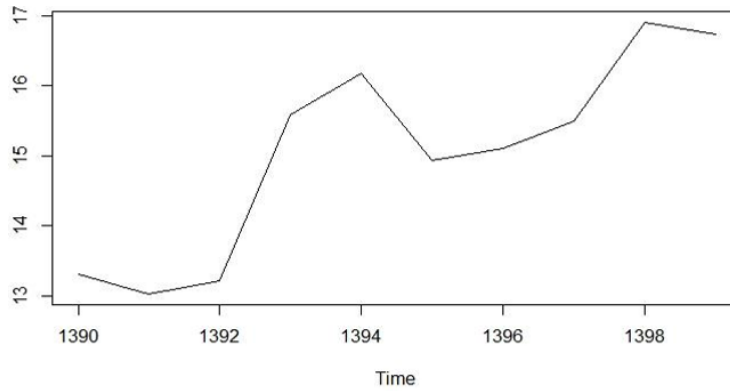


Fig. 2: Graph of changes in the planting date of the station zahak before differentiation

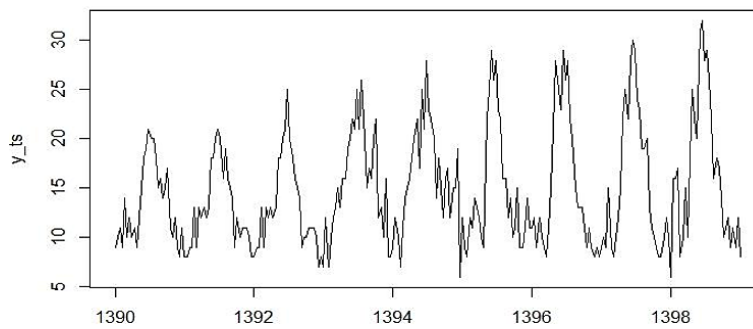


Fig. 3 : Graph of the trend of changes in the planting date of the station Zahak after differentiation

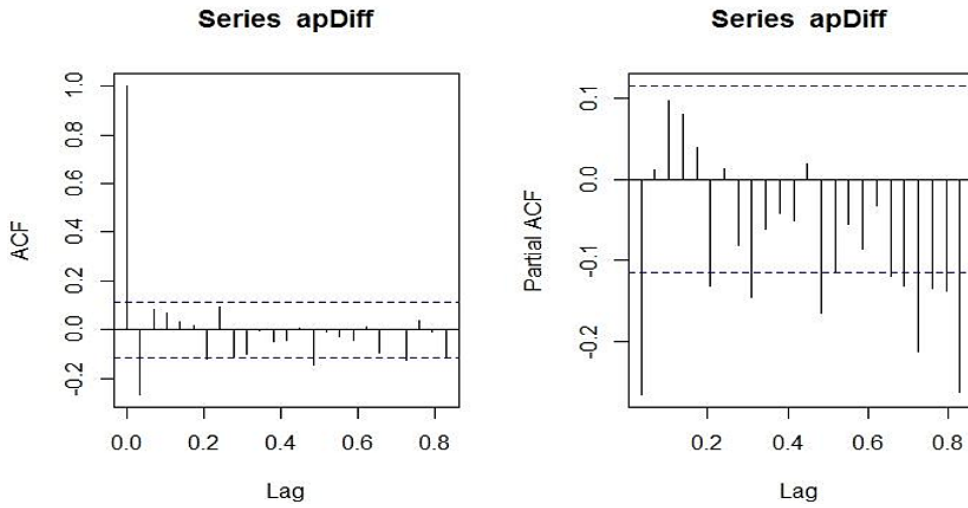


Fig. 4: Self-correlation diagram and partial self-correlation diagram before the annual flow differentiation at the station

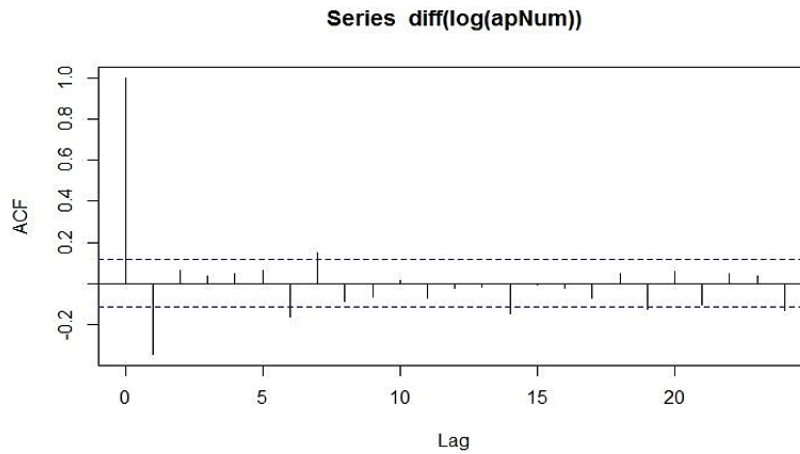


Fig. 5: Autocorrelation diagram after differentiation of planting date in the station zahak

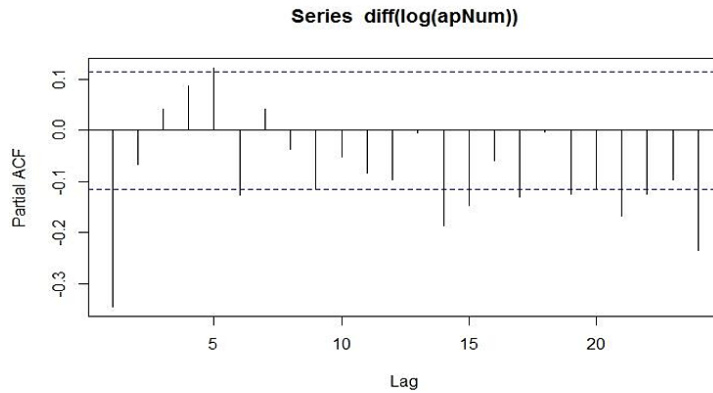


Fig. 6: Partial autocorrelation diagram after planting date differentiation at the station zahak.

The SARIMA family of models is one of the key models for predicting climatic variables. The planting date has been predicted using this model. The original model must be first identified and assessed, its parameters must be estimated, and the model's accuracy must then be assessed using a variety of criteria to characterize the planting date behavior of a year. In this case, we first determined the degrees of autocorrelation (p), difference (d), and moving average (q). According to the autocorrelation diagram in Figures (5 and 6), the trend in the relevant series causes the value of ACF to decline slowly and sinusoidally.¹³ Thus the first-degree differential conversion ($D = 1$) was performed to eliminate the trend from the data series. The date of planting seems appropriate. One of the branches in the

autocorrelation diagram emerges from the significant boundary following the first-degree differentiation and is hence ($q = 1$) (Figure 5). The second fork crossed the boundary after differentiation (Figure 6) and is ($p = 0$). Hence the initial model for model (0,1,1) is SARIMA, according to the partial autocorrelation itself. Based on the studies shown in Table (1), some models were fitted. The results are shown in this Table to help evaluate which model is best for predicting the planting date of parameters such as BIC (Bayesian information criterion), P.value, and T.value. Mvanga,¹⁶ 2017, reported SARIMA(2,1,2) (2,0,3)₄ was the best model which had the lowest AICc and therefore fit to the quarterly sugarcane yields data from 1973-2014.

Table 1: Modeling sufficiency evaluation of all the models

Adjusted R squared	AMAPE%	RMSE	BIC	T.value	P.value	Model
0.917	8.93	18.63	1259.3	1.34	0.54	SARIMA (0,0,0)
0.921	9.32	19.27	1364.1	3.89	0.0049	SARIMA (0.1.0)
0.934	9.54	19.98	45.5	1.23	0.89	SARIMA (3.1.0)
0.941	8.34	18.1	49.23	8.96	0.034	SARIMA (0.1.1)

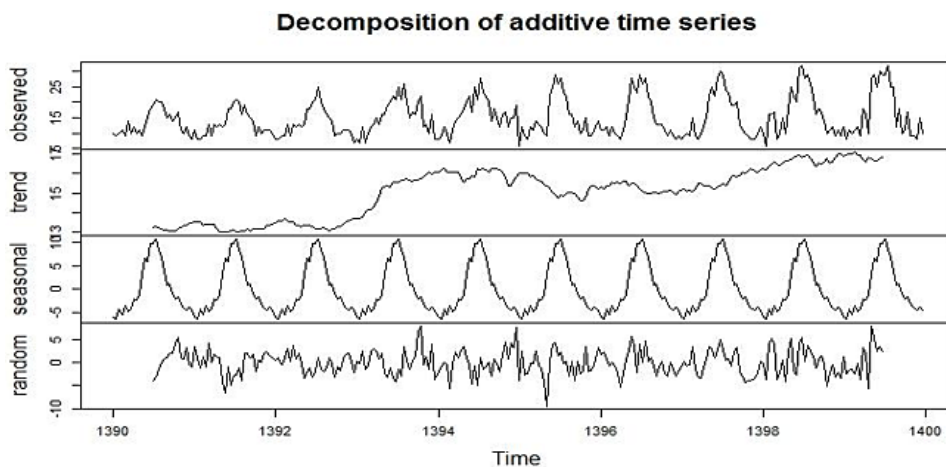


Fig. 7: Decomposition of the calibration vector of the transformed yield time series using Box-Cox (time series graphs with random, seasonal and trend components).

According to Table (1) only in SARIMA (0, 1, 1)(0, 1, 1) model, the absolute value of T statistic in all parameters is more than 2, P-value is less than 0.05 and is significant. Bayesian information criterion is the same or less in other models, but due to the inadequacy of other conditions (high P-VALUE and low T-value) were rejected, and finally the model (0,1,1) (0,1, 1) SARIMA was selected. Therefore, this model was selected as the most appropriate model to predict planting dates. Also According to the adjusted R squared value which indicates the goodness-of-fit, the (0,1, 1) SARIMA model with the highest adjusted R squared value 0.941 performs the best. The adjusted R squared value is not meaningful for time series models because increasing the number of predictors will weakly increase R squared, however fitting an SARIMA is to reproduce the underlying process with a model that is as simple as possible. According to the average of all the tested days, (0,1, 1) SARIMA gives the smallest root mean squared error as well as the adapted mean absolute percentage error. The target is to choose the model orders that result in minimum values of BIC, in results(table 1.) demonstrate to low

value but it is a little more than SARIMA (3,1,0) but based on p and t values accepted as the best model.

Figure 7 depicts the additive decomposition of the relevant series, which reveals that the series has a clear seasonality and primarily complies with the unimodal yield characteristic from the date, it is clear that a trend is not sustained throughout the series. This leads to the conclusion that $s = 29$, meaning that there are 29 periods per month (one period per year). Evident periodicity exists. Compared to other days, the incidence of yield was higher in the middle of the month and in the fall, and it has been rising recently. At this point, it is crucial to emphasize that although the examination of the residuals is considered crucial for analyzing the model's performance, there may be circumstances in which the Ljung-Box test finds that they are not random for a specific level of significance. Even so, it is decided that the chosen model is the best suited among all those examined and can be used to create forecasts if it can accurately depict the series' behavior and maintain the mean of the original data (Figure 8)

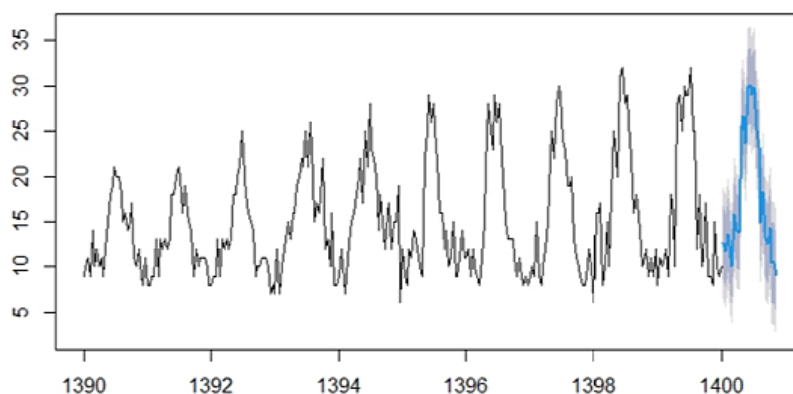


Fig. 8: The graph of the yield with the forecast values

Conclusion

The study can provide the reference basis to plan planting date for the government in Sistan region. Accurate precipitation forecasting is always a challenging problem, which is more attractive in many fields. SARIMA model is one of the most popular and models for precipitation forecasting. The SARIMA model (0,1,1) is the best model for predicting planting date data. Prediction results

show that the highest yield occurs from the end of February to the middle of March. This can maximize the yield of melon.

Acknowledgement

The authors would like to thank for the collaboration of the director of the Agricultural Research Station in Zahak.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest

The authors do not have any conflict of interest.

References

1. FAO. (2014). FAOSTAT. *Agriculture database*.
2. Rad M.R.N., Fanaei, H.R., & Rad, M.R.P. 2015. Application of Artificial Neural Networks to predict the final fruit weight and random forest to select important variables in native population of melon (*Cucumis melo* L.). *Scientia Horticulturae*, 181: 108-112. <http://dx.doi.org/10.1016/j.scienta.2014.10.025>.
3. Liu H., Li C., Shao Y., Zhang X., Zhai Z., Wang X., Qi X., Wang J., Hao Y., Wu Q.J.J.o.i., & Mingli J. 2020. Forecast of the trend in incidence of acute hemorrhagic conjunctivitis in China from 2011–2019 using the Seasonal Autoregressive Integrated Moving Average (SARIMA) and Exponential Smoothing (ETS) models. *Journal of Infection and Public Health*. 13(2): 287-294. <http://dx.doi.org/10.1016/j.jiph.2019.12.008>
4. Li L., & Xue J. 2009. The changing tendency and forecasting of world food security. *Journal of Shanghai University: Social Science Edition*, 3: 29-36.
5. Hyndman R.J., Koehler A.B., Snyder R.D., & Grose S. 2002. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of forecasting*, 18(3): 439-454. [http://dx.doi.org/10.1016/S0169-2070\(01\)00110-8](http://dx.doi.org/10.1016/S0169-2070(01)00110-8)
6. Hong-ying L., Yan-lin H., Yong-juan Z., & Hui-ming Z. 2012. Crop yield forecasted model based on time series techniques. *Journal of Northeast Agricultural University* 19(1): 73-77. [http://dx.doi.org/10.1016/S1006-8104\(12\)60042-7](http://dx.doi.org/10.1016/S1006-8104(12)60042-7)
7. Chen P., Niu A., Liu D., Jiang W., & Ma B. 2018. Time series forecasting of temperatures using SARIMA: An example from Nanjing. Paper presented at the IOP Conference Series: *Materials Science and Engineering*. 394:1-7. <http://dx.doi.org/10.1088/1757-899X/394/5/052024>
8. Wang Y., Shen Z., & Jiang Y.J.P.O. 2018. Comparison of ARIMA and GM (1, 1) models for prediction of hepatitis B in China. *Plos One* 13(9). <http://dx.doi.org/10.1371/journal.pone.0201987>
9. Hyndman R.J., & Khandakar Y. 2007. Automatic time series for forecasting: the forecast package for R. *Journal of Statistical Software*. 27 (3).1-23. <http://dx.doi.org/10.18637/jss.v000.i00>
10. Kabacoff R. (2015). R in action: *data analysis and graphics with R*: 1-579.
11. Zeng Q., Li D., Huang G., Xia J., Wang X., Zhang Y., Tang W., & Zhou H.J. 2016. Time series analysis of temporal trends in the pertussis incidence in Mainland China from 2005 to 2016. *Scientific Reports*. 6(1): 1-8. <http://dx.doi.org/10.1038/srep32367>
12. Guleryuz D.J.P.S., & Protection E. 2021. Forecasting outbreak of COVID-19 in Turkey; Comparison of Box–Jenkins, Brown’s exponential smoothing and long short-term memory models. *Process Safety and Environmental Protection*. 149: 927-935. <http://dx.doi.org/10.1016/j.psep.2021.03.032>
13. Boken V.K. 2000. Forecasting spring wheat yield using time series analysis: a case study for the Canadian Prairies. *Agronomy Journal*, 92(6): 1047-1053. <http://dx.doi.org/10.2134/agronj2000.9261047x>
14. Forte R.M. (2015). Mastering predictive analytics with R: *Packt Publishing Ltd*.
15. Yu Q., Wu W., Tang H., Chen Y., & Yang P.J.S.A.S. 2011. A food security assessment in APEC based on grain productivity. 2nd *IITA Conference on Geoscience and Remote Sensing*. (2): 2838-2848. <http://dx.doi.org/10.1109/IITA-GRS.2010.5603063>
16. Mvanga D., Ong’ala J., & Orwa G. 2017. Modeling sugarcane yields in the Kenya sugar industry: A SARIMA model forecasting approach. *International Journal of Statistics and Applications*. 7(6): 280-288.