Original Scientific Article Journal of Decision Analytics and Intelligent Computing
Vol. 3 issue 1, (2023) 105-112 https://doi.org/10.31181/jdaic10021072023k

Automated prediction of income from farming of a commodity: An ARIMA based framework

Soumyadipta Kar[1], Manas Kumar Mohanty[2,*] and Parag Kumar Guha Thakurta[2]

[1] Computer Science and Engineering, Haldia Institute of Technology, Haldia, West Bengal, India
[2] Computer Science and Engineering, National Institute of Technology Durgapur, Durgapur, West Bengal, India

* Correspondence: mkm19cs1102@phd.nitdgp.ac.in

Received 21 February 2023
Accepted for publication 16 July 2023
Published 21 July 2023

Abstract

In recent research, it has been found that an enormous amount of the population is involved in agriculture. Farmers are increasingly exposed to income risks from the effects of volatility in many factors directly or indirectly related to farming. Predicting farmers' income can be used to manage the income risks by assisting farmers. This paper proposes an ARIMA-based framework to forecast the income from a crop for the next consecutive years. A detailed analysis of the proposed work on the best-suited ARIMA framework is discussed. It is shown that the proposed work obtains a higher accuracy in predicting the income in the future than other alternative approaches.

Keywords: income, prediction, farmer, ARIMA.

1. Introduction

Most of the world's population is directly or indirectly involved in agriculture (Figueroa-Rodríguez et al., 2019). An improvement in the production from agriculture would be beneficial for those who are associated with farming. In such a context, to increase the farmer's earnings, the aim is to develop an efficient framework which can assist the farmers by forecasting the expected incomes from a crop for the next season. Several approaches are available to forecast such economic time series (Meyler et al., 1998). One such approach known as univariate forecasting includes only the time series being forecast. However, some of these studies have not taken care of the subject of rigorous forecast evaluation techniques. In such a scenario, the autoregressive integrated moving average (ARIMA) modelling can be effectively used for forecasting time series where it does not require any knowledge of an underlying economic model. Here, a time series is expressed in terms of past values of itself (the autoregressive (AR) component), in addition to the current and lagged values of a 'white noise' error term (the moving average (MA) component). Hence, an efficient ARIMA-based framework to predict the farmer's income is proposed in this paper as an assistance for their socio-economic benefit.

In this paper, the suitability of various types of ARIMA models, such as ARIMA (KumarMahto et al., 2019), seasonal autoregressive integrated moving average (SARIMA) (Dharavath and Khosla, 2019), vector autoregressive integrated moving average (VARIMA) (Rusyana et al., 2020) and autoregressive fractionally integrated moving average (FARIMA) (Wu et al., 2020), for our input dataset is estimated. In this context, the value of difference (d)

needed to make the data stationary is determined. Then, the input data is tested to detect the presence of white noise. Using the values of d, AR and MA components, the suitable ARIMA model among those alternatives is determined. By this best ARIMA model, the income of the farmers from a crop in the future years is efficiently predicted. In order to highlight the significance of the AR and MA components in the best ARIMA model used in the proposed work, the corresponding autocorrelation function (ACF), and partial autocorrelation function (PACF) plots are shown. The accuracy of the proposed approach is measured in terms of several performance metrics, such as R-squared error, root mean square error (RMSE), mean absolute percentage error (MAPE) and mean absolute error (MAE).

The rest of this paper is organized as: section 2 presents the literature review, the proposed methodology is described in section 3 and Section 4 shows the results and related discussions. Finally, section 5 completes the article with concluding remarks.

2. Literature review

In recent years, a lot of research works involves predict with using time-series based prediction models. However, research works related to predict farmer's income using ARIMA model are very few. The prediction of a future value using a model of time dependent causal relationship is relatively difficult (Mélard and Pasteels, 2000).

The time-series based method guarantees a satisfactory model through repetitive revisions to get the optimized prediction with minimum variance (Yamak et al., 2019). Detrending and seasonality testing methods are generally used for better outcomes (Newbold, 1983). The former makes the input data trend stationary, while the latter checks the effect of any seasonal parameter in the model.

Christias and Mocanu (2021) applies various machine learning algorithms for Olive farm profit prediction. However, it did not experiment with time-series based prediction. Another work forecasts profit of an enterprise using Long Short-Term Memory Neural Network (Qianyu et al., 2021). It does not assess the accuracy of any time series based prediction. Wang (2010) predicts farmer's income by ARIMA method using non-stationary process.

The main contribution of this article is that it comprehensively adopts more formal method as well as various feasibility tests before applying ARIMA model. As the ARIMA method works best for stationary data, we have applied the ARIMA method after making the data stationary. It finds the most suitable extension of ARIMA for a particular dataset. Then, the model applies correlogram based identification of ARIMA parameters. Finally, it shows a higher accuracy for predicting farmer's income for coming few years.

3. Proposed Methodology

It is known that presence of white noise in any data signifies that the data is random, which indicates an inability in prediction by any time series-based method. In the proposed work, firstly, the input data, 'D', is tested to detect the presence of white noise. It is done by Whites Lagrange Multiplier test (Hosking, 1980) returns one value, either zero or one. It is shown below.

$$WN = Whites\ Lagrange\ Multiplier\ Test\ (D) \tag{1}$$

If WN is one, then the D has white noise. So, the D can not be trained using any time-series based method. When the value of WN is zero, the proposed model can use the D to build a time-series based prediction model, as shown in equation (2).

$$Model\ Proposed\ (D) = \begin{cases} Infeasible, & \text{if}, WN = 1 \\ Model\ Proposed\ (D), & \text{if}, WN = 0 \end{cases} \tag{2}$$

Next, the D is made stationary by using the Augmented Dickey Fuller test (Baum, 2001) as ARIMA gives better result when applied on stationary data. *DStat* is the stationary version of D and d is the number of times for which the differencing method is needed by the Augmented Dickey Fuller test. The Augmented Dickey Fuller test returns two values, *DStat* and *d*. It is expressed by the following.

DStat, d = Dickey Fuller Stationary Test (*D*) (3)

When the value of d is greater than or equal to one, the D is non-stationary. So, the *DStat* is used as next input for the proposed model as shown in (4).

$$D = \begin{cases} DStat, & \text{if, } d \geq 1 \\ D, & Otherwise \end{cases} \tag{4}$$

In the next step, the remaining parameters of ARIMA are determined. The parameters, p and q are determined from the number of significant lags present in the correlograms, such as PACF and ACF plots, respectively, as shown in (5) and (6). The parameter, s, in SARIMA (p, d, q, s) is determined from Kruskal Wallis Seasonal test (Theobald and Price, 1984) as stated in equation (7).

p= Correlo. PACF (D) (5)
q= Correlo. ACF (D) (6)
s = Kruskal Wallis Seasonal. Component (*D*) (7)

As a next step, the suitability of different variations of ARIMA models, such as ARIMA (p, d, q), SARIMA (p, d, q, s), VARIMA (p, d, q) and FARIMA (p, d, q) are determined as per conditions shown in equation (8). The suitability of a variation of ARIMA method solely depends upon the nature of the input data.

$$\textit{Model Proposed}\ (D) = \begin{cases} SARIMA\ (p,d,q,s)(D), if\ s \geq 1 \\ VARIMA\ (p,d,q,s)(D), \text{if } Dimension(D) > 1 \\ FARIMA\ (p,d,q,s)(D), \text{if } 0 < d < 1 \\ ARIMA\ (p,d,q,s)(D), Otherwise \end{cases} \tag{8}$$

Finally, the best suitable selected variation of ARIMA method is applied on the 'D'. After applying the selected ARIMA model, the incomes of the farmers for a crop in the future years can be efficiently predicted. The flowchart of the proposed methodology is shown in Figure 1.

4. Results and Discussion

Various simulation results are shown to highlight the efficacy of the proposed work. In this context, the simulation setup and dataset description are shown next.

4.1. Simulation Setup

In order to simulate the behaviour of the proposed work, the dataset, D (Department for Environment, Food Rural Affairs, 2022), is used. The dataset contains the income of farmers, in million GBP (British Pound Sterling), at different years for a specific crop, i.e., wheat.

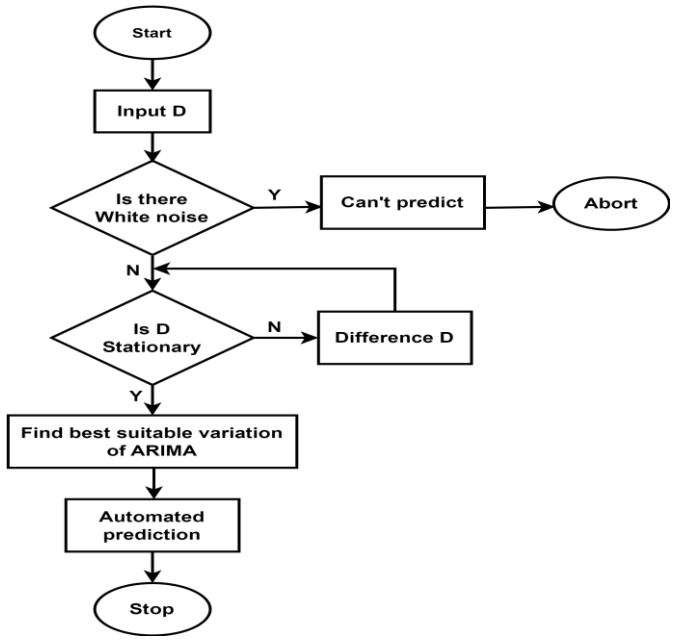


Figure 1. Flow chart of the proposed model

Various useful information regarding the income parameter of the dataset is shown in Table 1. The incomes of the future years are predicted using the best suitably parameterized ARIMA method. The proposed work is implemented using Python and Origin software in the machine having an Intel I7 processor and 16 GB RAM.

Table 1. Summary of Income in D

Test Parameter	Value
count	49
mean	1391.758958
std	597.033758
min	264.75
max	2704.680083

4.2. Simulation Results

It is known that if the data has white noise, then it can not be used for Time-Series based prediction. So, in the first step of of the proposed method, the D is tested for white noise. D is tested using Whites Lagrange multiplier test. It is observed from the Table 2 that the Test Statistic p-value is 0.0019715149386920127. As the p-value is very less than 0.05, the D does not contain white noise. It makes D suitable for time-series based prediction.

The sequence chart of D is shown in Figure 2. It is clearly seen that there is an uptrend in the farmers' income with respect to years. So, there is a higher chance that the D is non-stationary. It is known that the ARIMA is applied to a stationary data.

Table 2. Observations of White-Noise test

Test Parameter	Value
Test Statistic	12.457906054704887
Test Statistic p-value	0.0019715149386920127
F-Statistic	7.841144508225538
F-Test p-value	0.001174191803920481

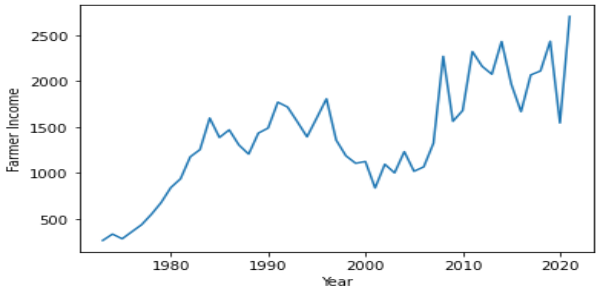


Figure 2. Sequence chart of D

Hence, as a next step of the proposed method, we have checked the the data for stationarity. We have used Augmented Dickey-Fuller test to check the stationarity of D. It is seen from the Table 3 that the Test Static value is higher than all the critical values and the p-value is higher. So, the D is detected as non-stationary data.

Table 3. Observations of Augmented Dickey-fuller test

Test Parameter	Value
Test Statistic	-1.474126
p-value	0.546270
lags used	2.000000
number of observations used	46.000000
critical value (1%)	-3.581258
critical value (5%)	-2.926785
critical value (10%)	-2.601541

As a next step of the proposed method, differencing method is applied on D. After that Augmented Dickey-fuller test is performed again to check stationarity of D after differencing once. Now, Table 4 contains result of this test. It is seen from the Table 4 that the Test Static value is lower than all the critical values and the p-value is very low. So, now the D is transformed into stationary after first differencing. So, the value of d of ARIMA (p,d,q) is successfully determined as 1.

Table 4. Observations of Augmented Dickey-fuller test after first differencing

Test Parameter	Value
Test Statistic	-6.078143
p-value	0.0000001
lags used	1
number of observations used	46
critical value (1%)	-3.581258
critical value (5%)	-2.926785
critical value (10%)	-2.601541

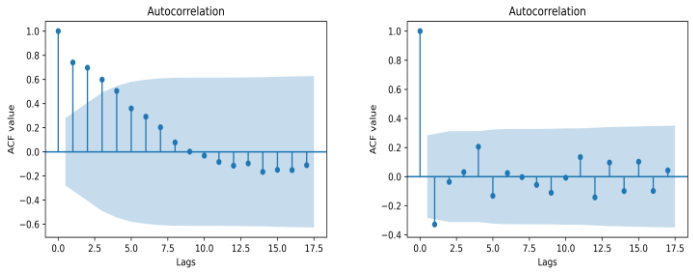
Next, the most suitable variation of ARIMA among ARIMA (p,d,q), VARIMA (p,d,q), SARIMA (p,d,q,s) and FARIMA (p,d,q), which will give the most accurate prediction on D is determined. The VARIMA is used when the predicted variable is multidimensional. However, the income field of D is unidimensional. SO, VARIMA is not selected for the current dataset. Next, SARIMA is used when the input data is seasonal. So, as a next step, D is tested for seasonal. The test statistics value and p-value for Kruskal-Wallis test for seasonality is found as 0 and 72.8179289 which is shown in Table 5.

Table 5. Observations of Kruskal-Wallis test

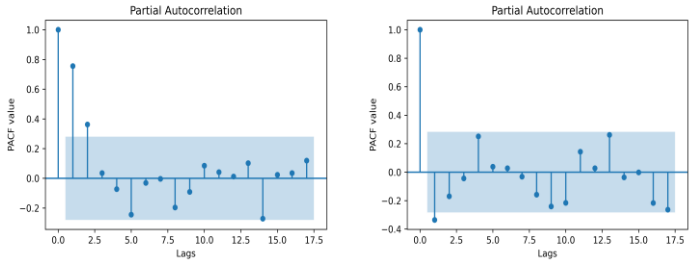
Test Parameter	Value
Test Statistic	72.8179289
Test Statistic p-value	0

This p-value is much less 0.05. It shows that the data is not seasonal. So, SARIMA is not applied here. Next, FARIMA is applied when the data is a long historical data and the correlation is there for a large number of lags. In this case, the value of d is found as a fraction. It is not the case for the proposed datset, D. So, finally, ARIMA is selected for predicting in case of the proposed dataset, D.

As a next step, it is necessary to find the p and q parameters of ARIMA (p,d,q). The values, p and q, are determined from ACF and PACF plots. The portion outside of the blue shaded area shows the significant values. Figure 3 (a) and (c) show that ACF values and PACF values are significant till lag 3 and lag 2 respectively. It indicates that ARIMA (2,0,3) needs to be determined when the data is not differenced. It can be seen from Figure 3 (b) and (d) that ACF and PACF values are significant for till lag 1, for both the plots. It shows that ARIMA (1,1,1) should be determined when the data is differenced once.



(a) ACF without differencing (b) ACF with first differencing



(c) PACF without differencing (d) PACF with first differencing

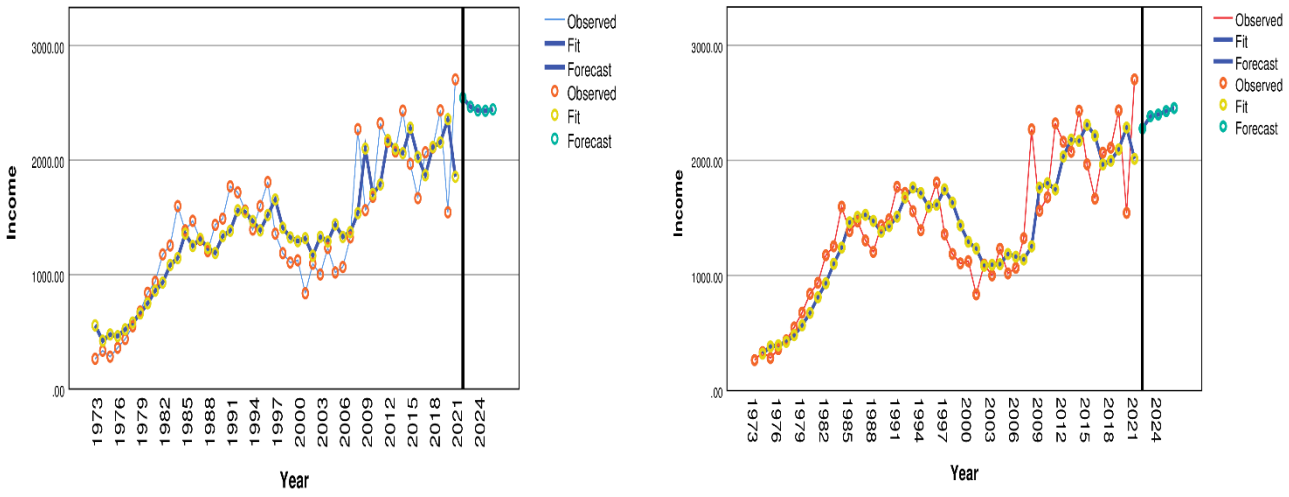
Figure 3. ACF and PACF plots without differencing and with first differencing

The performance measures of ARIMA (2,0,3) and ARIMA (1,1,1) are shown in Table 6. It shows that the performance of ARIMA (1,1,1) is better than ARIMA (2,0,3) in terms of all performance measures. The prediction of future incomes by both ARIMA (1,1,1) and ARIMA (2,0,3) are shown in Table 7 and in Figure 4. It can be seen that both the models have predicted differently.

Table 6. Performances of ARIMA (2,0,3)

Test Parameter	ARIMA (2,0,3)	ARIMA (1,1,1)
R-Squared (Chicco et al., 2021)	0.725	0.784
RMSE (Chicco et al., 2021)	319.561	310.368
MAPE (Chicco et al., 2021)	19.458	15.990
MAE (Chicco et al., 2021)	230.488	225.821

ARIMA (2,0,3) and ARIMA (1,1,1) have predicted farmer income as 2543.96 million GBP and 2277.33 million GBP for the year 2022, respectively.



(a) ARIMA (2,0,3) (b) ARIMA(1,1,1)

Figure 4. Plots of (a) ARIMA (2,0,3) and (b) ARIMA (1,1,1) models

Table 7. Comparisons of predictions by ARIMA models

Year	Predictions (million GBP)	
	ARIMA (2,0,3)	ARIMA (1,1,1)
2022	2543.96	2277.33
2023	2465.16	2384.05
2024	2434.00	2398.74
2025	2430.53	2428.52
2026	2443.16	2455.05

5. Conclusion

In this paper, an efficient ARIMA based framework is proposed for predicting farmer's income for future years. The pre-processing tasks are used to check the validity of input dataset. The major focus of the proposed work is to determine the best suitable ARIMA model in the prediction of the farmer's income for the next future years. Moreover, the ACF and PACF plots are shown to highlight the coefficients of ARIMA method. A higher accuracy by the proposed ARIMA method is obtained. The framework may assist the farmers in predicting the income of a crop in future which in turn can be beneficial for their economy. In this context, the prediction of incomes can be determined using other related parameters as a future enhancement of the proposed work.

References

Baum, C.F. (2001). Tests for stationarity of a time series. Stata Technical Bulletin, Stata Corp LP, 10(57).

Chicco, D., Warrens, M.J., & Giuseppe, J. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, e623.

Christias, P., & Mocanu, M. (2021). A Machine Learning Framework for Olive Farms Profit Prediction. Water, 13(23), 3461.

Department for Environment, Food Rural Affairs. Total income from farming in the UK. (2022). https://www.gov.uk/government/statistics/totalincome-from-farming-in-the-uk, Accessed 13 August 2022.

Dharavath, R., & Khosla, E. (2019). Seasonal ARIMA to forecast fruits and vegetable agricultural prices. 2019 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS) (pp. 47-52). Rourkela, India.

Figueroa-Rodríguez, K.A., Álvarez-Ávila, M.d.C., Hernández Castillo, F., Schwentesius Rindermann, R., & Figueroa-Sandoval, B. (2019). Farmers' Market Actors, Dynamics, and Attributes: A Bibliometric Study. Sustainability, 11(3), 745.

Hosking, J.R.M. (1980). Lagrange-multiplier tests of time-series models. Journal of the Royal Statistical Society: Series B (Methodological), 42(2), 170-181.

KumarMahto, A., Biswas, R., & Alam, M.A. (2019). Short Term Forecasting of Agriculture Commodity Price by Using ARIMA: Based on Indian Market. In: Singh, M., Gupta, P., Tyagi, V., Flusser, J., Ören, T., Kashyap, R. (eds) Advances in Computing and Data Sciences. ICACDS 2019. Communications in Computer and Information Science, vol 1045. Singapore: Springer.

Mélard, G., & Pasteels, J.-M. (2000). Automatic ARIMA modeling including interventions, using time series expert software. International Journal of Forecasting, 16(4), 497-508.

Meyler, A., Kenny, G., & Quinn, T. (1998). Forecasting Irish inflation using ARIMA models. Technical paper. Dablin: Economic Analysis, Research and Publications Department, Central Bank of Ireland.

Newbold, P. (1983). ARIMA model building and the time series analysis approach to forecasting. Journal of forecasting, 2(1), 23-55.

Qianyu, Z., Dongping, L., Xueying, Z., Huaisen, C., & Xiaozhou, Z. (2021). Enterprise Profit Forecast Model Based on Long Short-Term Memory Neural Network. 2021 International Conference on Big Data Analysis and Computer Science (BDACS) (pp. 62-65). Kunming, China.

Rusyana, A., Tatsara, N., Balqis, R., & Rahmi, S. (2020). Application of Clustering and VARIMA for Rainfall Prediction. OP Conference Series: Materials Science and Engineering, 796(1), 012063.

Theobald, M., & Price, V. (1984). Seasonality estimation in thin markets. The journal of finance, 39(2), 377-392.

Wang, H. (2010). Prediction of Farmers' Income and Selection of Model ARIMA. Asian Agricultural Research, USA-China Science and Culture Media Corporation, 2(11), 1-5.

Wu, F., Cattani, C., Song, W., & Zio, E. (2020). Fractional ARIMA with an improved cuckoo search optimization for the efficient Short-term power load forecasting. Alexandria Engineering Journal, 59(5), 3111-3118.

Yamak, P.T., & Yujian, L., & Gadosey, P.K. (2019), A Comparison between ARIMA, LSTM, and GRU for Time Series Forecasting. Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence (ACAI 19) (pp.49-55), New York.