JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 3353-3361
https://publishoa.com
ISSN: 1309-3452

Design and Development of Stochastic Modelling for Solanum Tuberosum Production in India

T. Jai Sankar *and* **P. Pushpa**

Department of Statistics, Bharathidasan University, Tiruchirappalli, Tamilnadu, INDIA
tjaisankar@gmail.com *and* pushbkr@gmail.com

Received 2022 March 25; **Revised** 2022 April 28; **Accepted** 2022 May 15.

Abstract

This study aims at design and development of stochastic modelling for *Solanum tuberosum* production in India based on *S. tuberosum* production during the years from 1950 to 2018. The study considers Autoregressive (AR), Moving Average (MA) and ARIMA processes to select the appropriate ARIMA model for *S. tuberosum* production in India. Based on ARIMA (p,d,q) and its components Autocorrelation Function (ACF), Partial Autocorrelation Function(PACF), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Normalized BIC and Box-Ljung Q statistics estimated, ARIMA (1,1,0) was selected. Based on the chosen model, it could be predicted that *S. tuberosum* production would increase from 53.03 million tons in 2018 to 66.12 million tons in 2025 in India.

***Key Words*:** ARIMA, BIC, Forecasting, MAPE, Potato.

Introduction

Potato (*Solanum tuberosum*) familiarly known as 'The King of Vegetables', has emerged as fourth most important food crop in India after rice (*Oryza sativa*), wheat (*Triticum aestivum*) and maize(*Zea mays*). Indian vegetable basket is incomplete without *S. tuberosum*. Because, the dry matter, edible energy and edible protein content of potato makes it nutritionally superior vegetable as well as staple food not only in our country but also throughout the world.*S. tuberosum* is a temperate crop grown under subtropical conditions in India. Uttar Pradesh is the major *S. tuberosum* producing state with 31.26% of production share, followed by West Bengal, Bihar, Gujarat and Madhya Pradesh with 23.29%, 13.22%, 7.43% and 6.20% share respectively in India. Figure 1 shows that the health benefits of *S. tuberosum*. The Minerals and Vitamins as available in *S. tuberosum* is given in Figure 2.

Material and Methods

As the aim of the study was to design and development of stochastic modelling for *S. tuberosum* production in India, various forecasting techniques were considered for use. ARIMA model, introduced by Box and Jenkins (1976), was frequently used for discovering the pattern and predicting the future values of the time series data. Box and Pierce, (1970) considered the distribution of residual autocorrelations in ARIMA. Akaike (1970) discussed the stationary time series by an AR(p), where p is finite and bounded by the same integer. MA models were used by Slutzky (1973). Wankhade et al. (2010) forecasted pigeon pea production in India with annual data from 1950-51 to 2007-08. Singh et al. (2013) developed and fitted forecast ARIMA (2,1,0) model during 2011-12 to 2014-15 for paddy production in Bastar division of Chhattisgarh for the period from 1974-75 to 2010-11. The study of Debnath et al. (2013) revealed that area, production and yield of cotton in India would increase from 2016-17 to 2020-21. Moyazzem Hossain and Faruq Abdulla (2016) analysed and fitted ARIMA (0,2,1) model for yearly potato production in Bangladesh over the period 1971 to 2013. Dasyam Ramesh et al. (2016) identified to fit ARIMA (1,1,0) model for production of Potato in West Bengal for the period of 1963-2012 and forecasted up to 2020.Borkar et al. (2016) in their empirical study showed that ARIMA (2,1,1) is the appropriate model for forecasting the production of cotton in India. Vijaya Wali et al. (2017) analyzed ARIMA (1,1,1) model for forecasting production of cotton in India. Saleem Abid et al. (2018) showed that forecasting time series data of potato production in Pakistan from 1980-81 to 2012-13. Hemavathi and Prabakaran (2018) calculated rice production data for the period of 1990-91 to 2014-15 and forecasted ARIMA (0,1,1) model up to 2020. BholaNath et al. (2019) found ARIMA (1,1,0) model for wheat production in India during the period 1949-50 to 2016-17 and forecasted up to 2026-27.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 3353-3361
https://publishoa.com
ISSN: 1309-3452

Figure 1. Health Benefits of *S. tuberosum*

Stochastic time-series ARIMA models were widely used in time series data which are having the characteristics (Alan Pankratz, 1983) of parsimonious, stationary, invertible, significant estimated coefficients and statistically independent and normally distributed residuals. When a time series is non-stationary, it can be made stationary by taking first differences of the series i.e., creating a new time series of successive differences (Y_t-Y_{t-1}). If the first differences do not convert the series form to stationary form, then first differences can be created. This is called second order differencing. A distinction is made between a second differences (Y_t-Y_{t-2}).

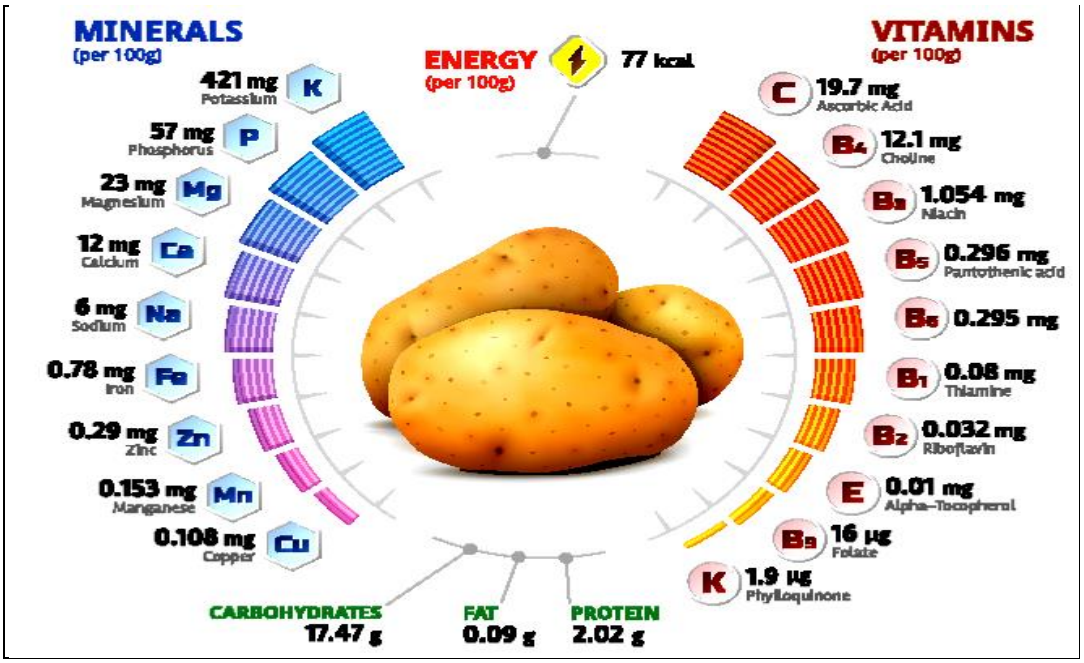

Figure 2. Minerals and Vitamins level of *S. tuberosum*

Source: Potato in India, Central Potato Research Institute (CPRI), Shimla.

The time series when differenced follows both AR and MA models and is known as ARIMA model. Hence, ARIMA model was used in this study, which required a sufficiently large data set and involved four steps: identification, estimation, diagnostic checking and forecasting. Model parameters were estimated to fit the ARIMA models.

Autoregressive process of order (p) is, $Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \varepsilon_t$;

Moving Average process of order (q) is, $Y_t = \mu - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - .... - \theta_q \varepsilon_{t-q} + \varepsilon_t$; and

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 3353-3361
https://publishoa.com
ISSN: 1309-3452

The general form of ARIMA model of order (p,d,q) is

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \mu - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - .... - \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where Y_t is *S .tuberosum* production, ε_t's are independently and normally distributed with zero mean and constant variance σ^2 for t = 1,2,..., n; d is the fraction differenced while interpreting AR and MA and ϕs and θs are coefficients to be estimated.

***Trend Fitting* :** The Box-Ljung Q statistics was used to transform the non-stationary data into stationarity data and alsoto check the adequacy for the residuals. For evaluating the adequacy of AR, MA and ARIMA processes, various reliability statistics like R^2, Stationary R^2, RMSE, MAPE, and BIC as suggested by Gideon Schwartz (1978) were used computed as follows:

$$RMSE = \left[\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \right]^{1/2} ; \; MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{(Y_i - \hat{Y}_i)}{Y_i} \right| \; and$$

$$BIC(p,q) = \ln v^*(p,q) + (p+q) [\ln(n) / n]$$

where p and q are the order of AR and MA processes respectively and n is the number of observations in the time series and v* is the estimate of white noise variance σ^2.

$$Q = \frac{n(n+2) \sum_{i=1}^{k} rk^2}{(n-k)}$$

where n is the number of residuals and rk is the residuals autocorrelation at lag k.

In this study, the data on *S. tuberosum* production in India were collected from the Annual Report (2018),Directorate of Economics and Statistics, Department of Agriculture, Cooperation and Farmers Welfare, Ministry of Agriculture and Farmers Welfare, Government of India for the period from 1950 to 2018 (Table 1) and were used to fit the ARIMA model to predict the future production.

Table 1. Actual *S. tuberosum* production (million tons) in India

Year	Production	*Year*	Production	*Year*	Production	*Year*	Production
1950	1.66	*1968*	4.73	*1986*	12.74	*2004*	23.63
1951	1.71	*1969*	3.91	*1987*	14.05	*2005*	23.91
1952	1.99	*1970*	4.81	*1988*	14.86	*2006*	22.18
1953	1.96	*1971*	4.83	*1989*	14.77	*2007*	28.47
1954	1.76	*1972*	4.45	*1990*	15.21	*2008*	34.39
1955	1.86	*1973*	4.86	*1991*	16.39	*2009*	36.58
1956	1.72	*1974*	6.23	*1992*	15.23	*2010*	42.34
1957	2.00	*1975*	7.31	*1993*	17.39	*2011*	41.48
1958	2.35	*1976*	7.17	*1994*	17.40	*2012*	45.34
1959	2.73	*1977*	8.14	*1995*	18.84	*2013*	41.56
1960	2.72	*1978*	10.13	*1996*	24.22	*2014*	48.01
1961	2.45	*1979*	8.33	*1997*	17.65	*2015*	43.42
1962	3.37	*1980*	9.67	*1998*	23.61	*2016*	48.6
1963	2.59	*1981*	9.91	*1999*	24.71	*2017*	51.31
1964	3.61	*1982*	9.96	*2000*	22.49	*2018*	53.03

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 3353-3361
https://publishoa.com
ISSN: 1309-3452

1965	4.08	*1983*	12.15	*2001*	23.92		
1966	3.52	*1984*	12.57	*2002*	23.27		
1967	4.23	*1985*	10.42	*2003*	23.06		

Results and Discussion

***Model Identification*:** ARIMA model was designed after assessing that transforming variable under forecasting was a stationary series. The stationary series was the set of values that is varied over time around a constant mean and constant variance. The most common method to check the stationarity is to explain the data through graph and hence is done in Figure 1.

Figure 1 reveals that the data used were non-stationary. Again, non-stationarity in mean was corrected through first differencing of the data. The newly constructed variable Y_t could now be examined for stationarity. Since, Y_t was stationary in mean, the next step was to identify the values of p and q. For this, the ACF and PACF of various orders of Y_t were computed and presented in Table 2 and Figure 2.

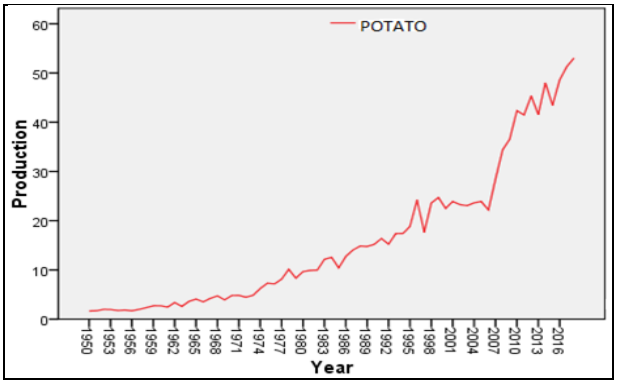

Figure 1. Time plot of *S. tuberosum* production

Table 2. ACF and PACF of *S. tuberosum* production

Lag	AC	Std. Error [a]	Box-Ljung Statistic	PAC	Std. Error	Lag	AC	Std. Error [a]	Box-Ljung Statistic	PAC	Std. Error
	Value	Df	Sig.[b]	Value	Df		Value	Df	Sig.[b]	Value	Df
1	0.931	0.118	62.502	0.931	0.120	***17***	0.231	0.103	429.660	-0.093	0.120
2	0.875	0.117	118.438	0.053	0.120	***18***	0.205	0.102	433.690	0.018	0.120
3	0.815	0.116	167.781	-0.042	0.120	***19***	0.174	0.101	436.654	-0.077	0.120
4	0.766	0.115	211.962	0.038	0.120	***20***	0.138	0.100	438.564	-0.061	0.120
5	0.702	0.114	249.728	-0.120	0.120	***21***	0.105	0.099	439.697	0.005	0.120
6	0.654	0.113	282.991	0.061	0.120	***22***	0.082	0.098	440.396	0.043	0.120
7	0.593	0.112	310.734	-0.110	0.120	***23***	0.042	0.097	440.579	-0.111	0.120
8	0.539	0.112	334.040	-0.004	0.120	***24***	0.014	0.096	440.599	0.058	0.120
9	0.482	0.111	353.034	-0.028	0.120	***25***	-0.012	0.095	440.614	0.016	0.120
10	0.432	0.110	368.563	-0.016	0.120	***26***	-0.036	0.094	440.765	-0.054	0.120
11	0.387	0.109	381.238	0.036	0.120	***27***	-	0.093	441.153	0.038	0.120

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 3353-3361
https://publishoa.com
ISSN: 1309-3452

							0.058				
12	0.357	0.108	392.164	0.060	0.120	***28***	-0.081	0.091	441.941	-0.096	0.120
13	0.337	0.107	402.127	0.100	0.120	***29***	-0.102	0.090	443.213	0.008	0.120
14	0.312	0.106	410.801	-0.063	0.120	***30***	-0.123	0.089	445.116	-0.040	0.120
15	0.287	0.105	418.251	-0.018	0.120	***31***	-0.146	0.088	447.846	-0.066	0.120
16	0.263	0.104	424.622	-0.015	0.120	***32***	-0.167	0.087	451.557	-0.008	0.120

[a] The underlying process assumed is independence (white noise).

[b] Based on the asymptotic chi-square approximation.

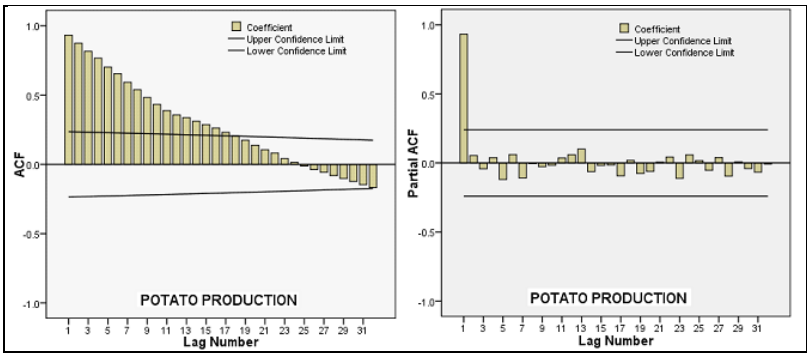

Figure 2. ACF and PACF of differenced data

The tentative ARIMA models are discussed with values differenced once (d=1) and the model which had the minimum normalized BIC was chosen. The various ARIMA models and the corresponding normalized BIC values are given in Table 3. The value of normalized BIC of the chosen ARIMA was 1.597.

Table 3. BIC values of various ARIMA (p,d,q)

ARIMA (p,d,q)	0,1,0	0,1,1	0,1,2	**1,1,0**	1,1,1	1,1,2	2,1,0	2,1,1	2,1,2	3,1,0	3,1,1	3,1,2
BIC Values	1.782	1.658	1.627	**1.597**	1.675	1.705	1.675	1.751	1.71	1.722	1.799	1.765

***Model Estimation*:** Model parameters and fit statistics were estimated and the results of estimation are presented in Tables 4 and 5. Hence, the most suitable model for *S. tuberosum*production was ARIMA (1,1,0), as this model had the lowest normalized BIC value, good R^2 and better model fit statistics RMSE and MAPE.

Table 4. Estimated ARIMA model

	Estimate	**SE**	**t**	**Sig.**
Constant	-58.191	17.094	-3.404	0.001
AR1	-0.473	0.109	-4.351	0.000

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 3353-3361
https://publishoa.com
ISSN: 1309-3452

Table 5. Estimated ARIMA model fit statistics

Performances of different ARIMA (p,d,q) models of *S. tuberosum* production in India								
ARIMA (p,d,q)	Stationary R^2	R^2	RMSE	MAPE	MaxAPE	MAE	MaxAE	Normalized BIC
0,1,0	0.062	0.976	2.291	11.065	43.604	1.471	7.696	1.782
0,1,1	0.234	0.98	2.087	10.679	34.009	1.376	6.003	1.658
0,1,2	0.312	0.982	1.992	10.165	32.628	1.29	7.438	1.627
1,1,0	**0.278**	**0.981**	**2.025**	**10.376**	**32.111**	**1.322**	**6.772**	**1.597**
1,1,1	0.279	0.981	2.041	10.4	32.255	1.322	6.775	1.675
1,1,2	0.312	0.982	2.008	10.176	32.721	1.291	7.446	1.705
2,1,0	0.279	0.981	2.041	10.37	32.289	1.321	6.781	1.675
2,1,1	0.28	0.981	2.055	10.06	32.55	1.311	6.895	1.751
2,1,2	0.361	0.983	1.952	10.49	33.729	1.327	5.953	1.71
3,1,0	0.3	0.982	2.026	10.169	32.539	1.31	7.198	1.722
3,1,1	0.301	0.982	2.041	10.062	31.528	1.309	7.192	1.799
3,1,2	0.375	0.984	1.945	10.392	35.899	1.267	6.336	1.765

***Diagnostic Checking*:** The model verification is concerned with checking the residuals of the model to see if they contained any systematic pattern which still could be removed to improve the chosen ARIMA, which has been done through examining the autocorrelations and partial autocorrelations of the residuals of various orders. For this purpose, various autocorrelations up to 32 lags were computed and the same along with their significance tested by Box-Ljung statistic are provided in Table 6. As the results indicate, none of these autocorrelations was significantly different from zero at any reasonable level. This proved that the selected ARIMA model was an appropriate model for forecasting *S. tuberosum* production in India.

The ACF and PACF of the residuals are given in Figure 5, which indicated the 'good fit' of the model. Hence, the fitted ARIMA model for *S. tuberosum* production data was

$$Y_t = \phi_1 Y_{t-1} + \mu + \varepsilon_t$$

$$Y_t = -58.191 - 0.473 Y_{t-1} + \varepsilon_t$$

Table 6: Residual of ACF and PACF

Lag	**ACF**		**PACF**		***Lag***	**ACF**		**PACF**	
	Mean	**SE**	**Mean**	**SE**		**Mean**	**SE**	**Mean**	**SE**
1	-0.001	0.121	-0.001	0.121	***17***	-0.240	0.148	-0.215	0.121
2	0.092	0.121	0.092	0.121	***18***	0.145	0.153	0.059	0.121
3	0.096	0.122	0.097	0.121	***19***	-0.118	0.155	-0.004	0.121
4	-0.202	0.123	-0.213	0.121	***20***	-0.073	0.157	0.046	0.121
5	-0.187	0.128	-0.218	0.121	***21***	0.194	0.157	0.051	0.121
6	-0.111	0.132	-0.092	0.121	***22***	-0.052	0.161	-0.116	0.121
7	-0.164	0.133	-0.094	0.121	***23***	-0.039	0.161	-0.043	0.121
8	-	0.136	-	0.121	***24***	0.023	0.161	-	0.121

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 3353-3361
https://publishoa.com
ISSN: 1309-3452

	0.090		0.088					0.037	
9	0.229	0.137	0.213	0.121	***25***	-0.030	0.161	-0.048	0.121
10	-0.042	0.143	-0.064	0.121	***26***	-0.021	0.161	0.126	0.121
11	-0.070	0.143	-0.228	0.121	***27***	-0.047	0.161	-0.124	0.121
12	0.184	0.143	0.059	0.121	***28***	-0.026	0.161	-0.107	0.121
13	0.048	0.147	0.160	0.121	***29***	-0.067	0.161	-0.036	0.121
14	-0.054	0.147	-0.040	0.121	***30***	0.031	0.162	-0.039	0.121
15	0.017	0.147	-0.103	0.121	***31***	0.038	0.162	0.077	0.121
16	0.045	0.147	0.116	0.121	***32***	-0.040	0.162	-0.010	0.121

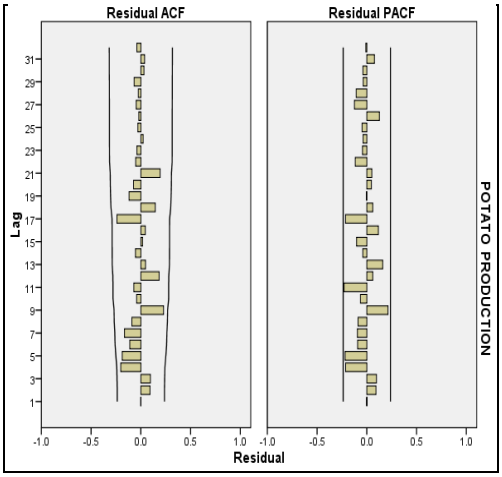

Figure 5. Residuals of ACF and PACF

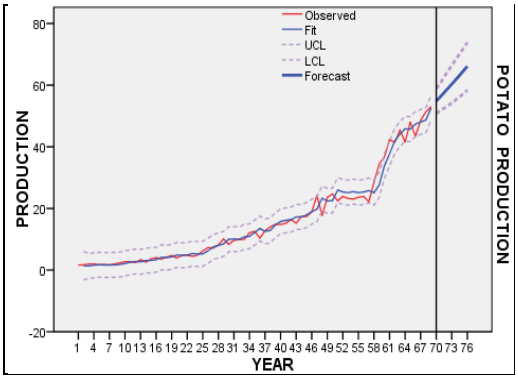

Figure 6. Actual and Estimate of Production

***Forecasting*:** Based on the model fitted, forecasted *S. tuberosum* production (in million tons) for the year 2019 through 2023 respectively given by 54.82, 56.62, 58.47, 60.33, 62.23, 64.16 and 66.12 are given in Table 7. To assess the forecasting ability of the fitted ARIMA model, the measures of the sample period forecasts' accuracy were also

computed. This measure also indicated that the forecasting inaccuracy was low. Figure 6 shows the actual and forecasted value of *S. tuberosum* production (with 95% confidence limit) in the country. The upper control limit (UCL) and lower control limit (LCL) values of the forecasted *S. tuberosum* production in India is provided in the same Table 7.

Table 7. Forecast of *S. tuberosum* production

Year	Predicted	UCL	LCL
2019	54.82	58.87	50.78
2020	56.62	61.20	52.05
2021	58.47	63.95	52.98
2022	60.33	66.41	54.26
2023	62.23	68.92	55.54
2024	64.16	71.38	56.94
2025	66.12	73.85	58.38

Conclusion

The most appropriate ARIMA model after design and development of stochastic modelling for *S. tuberosum* production forecasting of data was found to be ARIMA (1,1,0). From the time series data, it can be found that forecasted production would increase to 66.12 million tons in 2025 from 53.03 million tons in 2018 in India for using time series data from 1950 to 2018 on *S. tuberosum* production, this study provides an evidence on future *S. tuberosum* production in the country, which can be considered for future policy making and formulating strategies for augmenting and sustaining *S. tuberosum* production in India.

References

1. H. Akaike, Statistical Predictor Identification, *Annals of Institute of Statistical Mathematics*, (1970), no. **22**, 203-270.
2. Alan Pankratz, Forecasting with Univariate Box - Jenkins Models: Concepts and Cases, John Wiley & Sons, New York, 1983.
3. BholaNath, D.S. Dhakre and Debasis Bhattacharya, Forecasting wheat production in India: An ARIMA modelling approach, *Journal of Pharmacognosy and Phytochemistry*, (2019), no. **8(1)**, 2158-2165.
4. Borkar, Prema and P.M. Tayade, Forecasting of Cotton Production in India Using ARIMA Model, *International Journal of Research in Economics and Social Sciences*, (2016), no. **6(5)**, 1-7.
5. G.E.P. Box and G.M. Jenkins, Time Series Analysis: Forecasting and Control, San Francisco, Holden-Day, California, USA, 1976.
6. G.E.P. Box and D.A. Pierce, Distribution of Residual Autocorrelations in ARIMA Models, *J. American Stat. Assoc.*, (1970), no. **65**, 1509-1526.
7. Dasyam Ramesh, Bhattacharyya Banjul and P. Mishra, Statistical Modeling to area, production and yield of Potato in West Bengal, *International Journal of Agriculture Sciences*, (2016) no. **8(53)**, 2782-2787.
8. M.K. Debnath, Kartic Bera and P. Mishra, Forecasting Area, Production and Yield of Cotton in India using ARIMA Model, *Research and Reviews: Journal of Space & Technology*, (2013), no. **2(1)**, 16-20.
9. Gideon Schwarz, Estimating the dimension of a model, *Annals of Statistics*, (1978), no. **6(2)**, 461-464.
10. M. Hemavathi and K. Prabakaran, ARIMA Model for Forecasting of Area, Production and Productivity of Rice and Its Growth Status in Thanjavur District of Tamil Nadu, India, *Int. J. Curr. Microbiol. App. Sci.*, (2018), no. **7(2)**, 149-156.

11. Md. Moyazzem Hossain and Faruq Abdulla, Forecasting Potato Production in Bangladesh by ARIMA Model, *Journal of Advanced Statistics*, (2016), no. **1(4)**, 191-198.
12. Saleem Abid, Nasir Jamal, Muhammad Zubair Anwar and Saleem Zahid, Exponential growth model for forecasting of area and production of potato crop in Pakistan, *Pakistan Journal of Agricultural Research*, (2018), no. **31(1)**, 24-28.
13. D.P. Singh, Prafull Kumar and K. Prabakaran, Application of ARIMA model for forecasting Paddy production in Bastar division of Chhattisgarh, *American International Journal of Research in Science, Technology, Engineering & Mathematics*, (2013), no. **5(1)**, 82-87.
14. E. Slutzky, The summation of random causes as the source of cyclic processes, *Econometrica*, (1973), no. **5**, 105-146.
15. Vijaya B. Wali, Devendra Beeraladinni and H. Lokesh, Forecasting of Area and Production of Cotton in India: An Application of ARIMA Model, *Int. J. Pure App. Biosci*., (2017), no. **5(5)**, 341-347.
16. R. Wankhade, S. Mahalle, S. Gajbhiye and V.M. Bodade, Use of the ARIMA model for forecasting pigeon pea production in India, *Int. Rev. Bus. Finance*, (2010), no. **2**, 97-102.