Forecasting Onion Production in India Using Arima Models: A Research Study

Dr. P.SAMEERABANU
Assistant Professor in Mathematics
School of Engineering & Technology, Dhanalakshmi Srinivasan University
Trichy, Tamilnadu,India
Sameerabanup.set@dsuniversity.ac.in

Abstract

This study explores the application of ARIMA (AutoRegressive Integrated Moving Average) models for forecasting onion production in India. Accurate forecasting of agricultural production is essential for effective planning and decision-making in the agricultural sector. ARIMA models, which integrate autoregression, differencing, and moving average components, offer a robust methodology for time series forecasting. This study aims to explore the application of ARIMA models in forecasting onion production in India. By utilizing historical production data, the study will identify suitable ARIMA parameters and evaluate the model's effectiveness in predicting future production levels. Use the fitted model to forecast future onion production. Evaluate the model's performance using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE).The goal is to provide insights that can help stabilize onion supply and contribute to better planning and decision-making within the agricultural sector.

Keywords: ARIMA,MAE,RMSE,MSE,SPSS

Introduction

Onion production plays a crucial role in the agricultural sector of India, which is one of the largest producers of onions globally. The vegetable is an essential staple in Indian cuisine, contributing significantly to the country's food security and agricultural economy. However, onion production is subject to various factors such as climatic conditions, market demand, and agricultural practices, which can cause substantial fluctuations in yield.

Effective forecasting of onion production is vital for managing these fluctuations and ensuring a stable supply to meet consumer demand. Accurate forecasts can help farmers, policymakers, and stakeholders in making informed decisions related to production planning, pricing strategies, and supply chain management.

In recent years, time series forecasting models have gained prominence in predicting agricultural outputs. Among these, the ARIMA (AutoRegressive Integrated Moving Average) model is widely used due to its ability to handle various types of time series data by incorporating autoregressive, differencing, and moving average components. The ARIMA model is especially useful in analyzing and forecasting data that exhibit trends or seasonality.

In the following sections, we will outline the methodology for data collection, preprocessing, model identification, and forecasting, followed by an analysis of the results and their implications for stakeholders involved in onion production and distribution.

Understanding ARIMA Models

ARIMA models are used for analyzing and forecasting time series data. They consist of three main components:

1. **AutoRegression (AR):** Refers to the model that uses the relationship between an observation and a number of lagged observations.
2. **Integrated (I):** Refers to differencing of raw observations to make the time series stationary (i.e., removing trends and seasonality).
3. **Moving Average (MA):** Involves modeling the error of the forecast as a linear combination of error terms observed at previous time points.

The model is typically denoted as ARIMA(p, d, q):

- **p:** Number of lag observations (lag order).

- **d:** Degree of differencing (number of times the data have had past values subtracted).
- **q:** Size of the moving average window.

Build an ARIMA Model for Forecasting Onion Production:

1. **Data Collection:**

○ Gather historical data on onion production. This can include monthly or yearly production volumes over several years.

2. **Preprocessing:**

○ **Check for Stationarity:** The data should be stationary for ARIMA to work effectively. Use statistical tests like the Augmented Dickey-Fuller test.

○ **Differencing:** Apply differencing to make the series stationary if it is not already. This involves subtracting the previous observation from the current observation.

3. **Model Identification:**

○ Determine the values of p, d, and q using tools like the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots.

4. **Parameter Estimation:**

○ Use statistical software or programming languages like R or Python to estimate the parameters of the ARIMA model. Libraries such as statsmodels in Python can be very useful.

5. **Model Fitting:**

○ Fit the ARIMA model to your data and check the residuals to ensure they resemble white noise (i.e., they are uncorrelated and have a constant mean and variance).

6. **Forecasting:**

○ Use the fitted model to forecast future onion production. Evaluate the model's performance using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE).

7. **Validation:**

○ Validate the model by comparing the forecasted values with actual values using a test dataset.

Time Series Data

Time series data is defined as a collection of values of a variable that differs over time. The intervals between observations of a time series can vary. However, the range of the intervals should be consistent throughout the observed period e.g. daily, weekly, monthly etc. In general, the time series is assumed to be stationary in empirical work based on time series (Gujarati & Porter 2008).

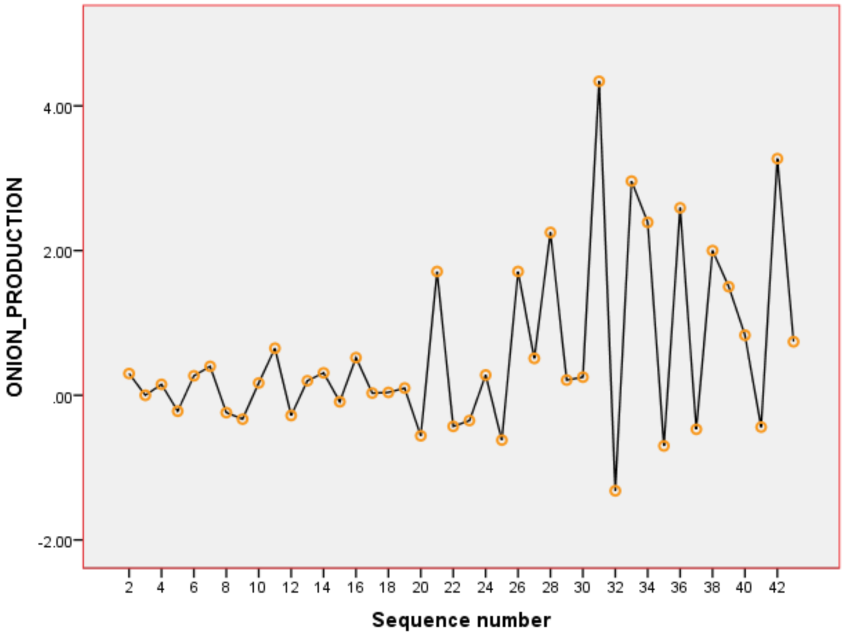
Literature Review:

Forecasting the Indian Stock Market As mentioned previously, the article “A Prediction Approach for Stock Market Volatility Based on Time Series Data” In the article, the logarithmic transformation is applied to the data and two ARIMA models are estimated to forecast the two indices. The best estimated models for the data were two ARIMA(0, 1, 0) with drift and the authors of the article conclude that a correctly chosen ARIMA model is sufficiently accurate in forecasting time series data. The conclusion is based on the fact that the predicted values of the used models in the article, on average, deviated by approximately 5% from the actual outcome, computed by the out-of-sample MPE (Idrees et al. 2019).

Comparison of Forecasting Models Accuracy In the article “ARIMA: An Applied Time Series Forecasting Model for the Bovespa Stock Index” the MAPE is used to determine which model, among several different forecasting models, is the most accurate in forecasting the Brazilian stock index Bovespa. Among the models, the authors compare an autoregressive model, two different exponential smoothing models, and an ARIMA(0, 2, 1). The Box-Jenkins methodology is followed when building the ARIMA model in the article. The authors conclude that according to the data, an AR(1) is the most accurate model since it has the lowest out-of-sample MAPE. The authors further conclude that an AR(1) for the Bovespa stock index is an adequate model to use as a tool to forecast the index (Rotela Junior et al. 2014).

Comparing the AIC of Different Models to Find the Best Fit By using Akaike’s information criterion, Snipes & Taylor (2014) performed a research to discover the best-fitted model to explain the relationship between the rating of wines and the respective price. In their research, they used what is known as the AICc which is a slightly modified AIC. Similar to AIC, the AICc penalizes the addition of unnecessary information to a statistical model and the model with the lowest AICc score, among different models, has the best fit based on the data.

Sequence Plot



Transforms: difference(1)

ONION_PRODUCTION

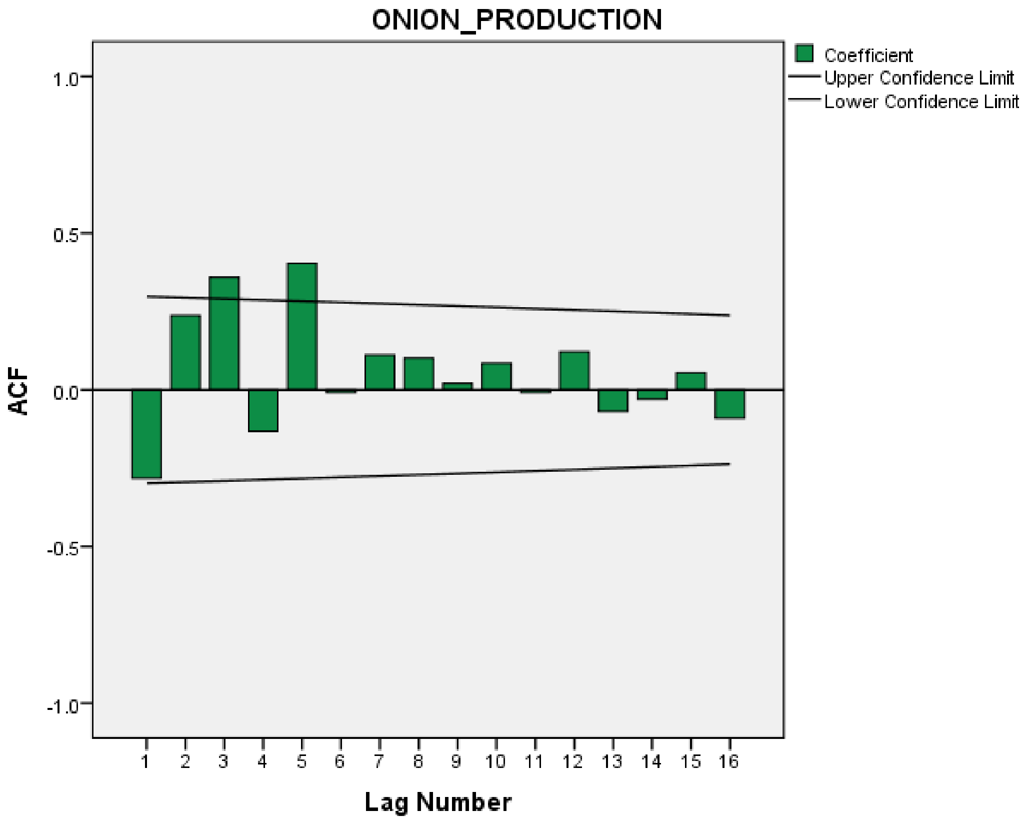
Autocorrelations & Partial Autocorrelation:

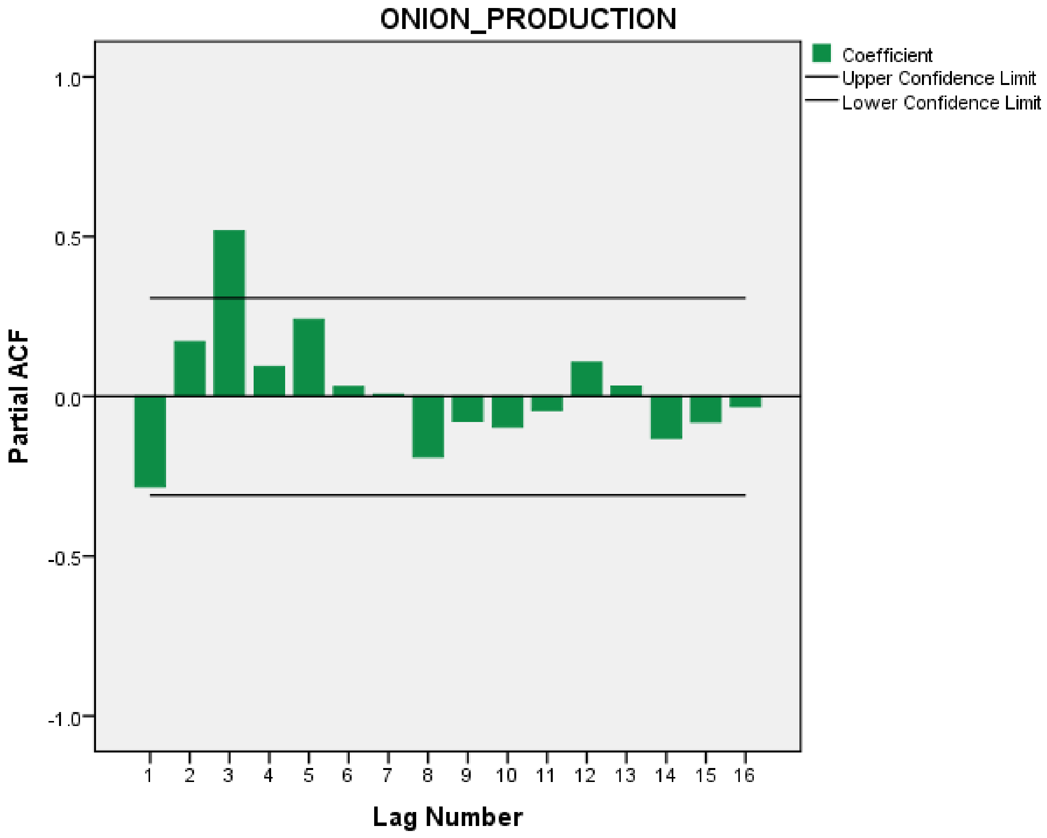
Series: ONION_PRODUCTION

Lag	Autocorrelation	Partial Autocorrelation	Box-Ljung Statistic		
			Value	df	Sig.[b]
1	-.282	-.282	3.585	1	.058
2	.236	.170	6.167	2	.046
3	.359	.518	12.282	3	.006
4	-.132	.093	13.132	4	.011
5	.403	.241	21.263	5	.001
6	-.009	.030	21.267	6	.002
7	.111	.005	21.919	7	.003
8	.101	-.190	22.476	8	.004
9	.020	-.078	22.499	9	.007
10	.084	-.095	22.907	10	.011
11	-.009	-.043	22.912	11	.018
12	.121	.106	23.814	12	.022
13	-.069	.031	24.117	13	.030
14	-.030	-.131	24.175	14	.044
15	.054	-.080	24.376	15	.059
16	-.091	-.031	24.965	16	.070

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

b. Based on the asymptotic chi-square approximation.


ONION_PRODUCTION
🟩 Coefficient
Upper Confidence Limit
Lower Confidence Limit
ACF
1.0
0.5
0.0
-0.5
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Lag Number


ONION_PRODUCTION
🟩 Coefficient
Upper Confidence Limit
Lower Confidence Limit
Partial ACF
1.0
0.5
0.0
-0.5
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Lag Number

The above graph indicates that ARIMA (1, 1, 1),(1,1,3),(1,1,5),(3,1,1),(3,1,3),(3,1,5) overall provides a good fit for model ARIMA(3,1,5) ,Compared with above ARIMA Models., it has a wider confidence interval. It can be seen that the actual data from 1978 is to 2020- actually touches the interval, however, there still is a gap between the blue line and the interval in the ARIMA (1, 1, 5) forecast.

I have evaluated all possible models and examined their error terms. Among the six models applied, that models are ARIMA (3,1,5) which error term reduction indicates the most appropriate fit for the models. Further I selected ARIMA (3,1,5) model.

Time Series Modeler

Model Description

			Model Type
Model ID	ONION_PRODUCTION	Model_1	ARIMA(3,1,5)

Model Summary

Fit Statistic	Mean
Stationary R-squared	.539
R-squared	.989
RMSE	.918
MAPE	10.535
MaxAPE	33.588
MAE	.627
MaxAE	2.489
Normalized BIC	.629

Hypotheses

The Ljung-Box test uses the following hypotheses:

H_0: The residuals are independently distributed.

H_A: The residuals are not independently distributed; they exhibit serial correlation.

Ideally, we would like to fail to reject the null hypothesis. That is, we would like to see the p-value of the test be greater than 0.05 because this means the residuals for our time series model are independent, which is often an assumption we make when creating a model.

Test Statistic

The test statistic for the Ljung-Box test is as follows:

Q = $n(n+2)\ \Sigma p_k^2 / (n-k)$

Statistics	DF	Sig.	Number of Outliers
4.991	10	.892	0

The test statistic of the test is Q = **4.991** and the p-value of the test is **0.892**, which is much larger than 0.05. Thus, we fail to reject the null hypothesis of the test and conclude that the data values are independent.

ARIMA Model Parameters

					Estimate	SE	t	Sig.
ONION_PRODUCTION-Model_1	ONION_PRODUCTION	No Transformation	Constant		.633	.317	1.995	.054
			AR	Lag 1	-.238	.478	-.498	.622
				Lag 2	.278	.311	.893	.378
				Lag 3	.306	.390	.784	.439
			Difference		1			
			MA	Lag 1	.207	10.957	.019	.985
				Lag 2	.197	13.378	.015	.988
				Lag 3	-.526	11.367	-.046	.963
				Lag 4	-.020	5.528	-.004	.997
				Lag 5	-.501	5.593	-.090	.929

For Model:

$Y_i = C_1 + \varphi\ (Y_{i-1}) + \theta\ \varepsilon_{1,\ i-1}$ ... (1)

Where C is constant, ε_i is white noise

$Y_i = Y_i - Y_{i-1}$ ... (2)

Combine (1) and (2), we have:

$Y_i - Y_{i-1} = C_1 + \varphi\ (Y_{i-1} - Y_{i-2}) + \varepsilon_1 + \theta_1\ \varepsilon_{1,\ i-1}$ ... (3)

Then,

$Y_i = 0.760324 - 0.191462\ Y_{i-1} + \varepsilon_1 + 0\ .207\varepsilon_{1,\ i-1} + 0\ .197\varepsilon_{2,\ i-2} - 0\ .526\varepsilon_{3,\ i-3} - 0\ .020\varepsilon_{4,\ i-4} - 0\ .501\varepsilon_{5,\ i-5}$ ....(4

Forecast:

Forecast

Model		44	45	46	47	48	49	50	51	52	53
ONION_PRODUCTION-Model_1	Forecast	27.21	29.92	29.41	31.60	32.28	32.99	34.09	34.65	35.45	36.17
	UCL	29.03	31.99	31.90	35.14	36.41	38.42	40.53	42.08	43.93	45.57
	LCL	25.40	27.85	26.92	28.07	28.15	27.55	27.66	27.21	26.98	26.76

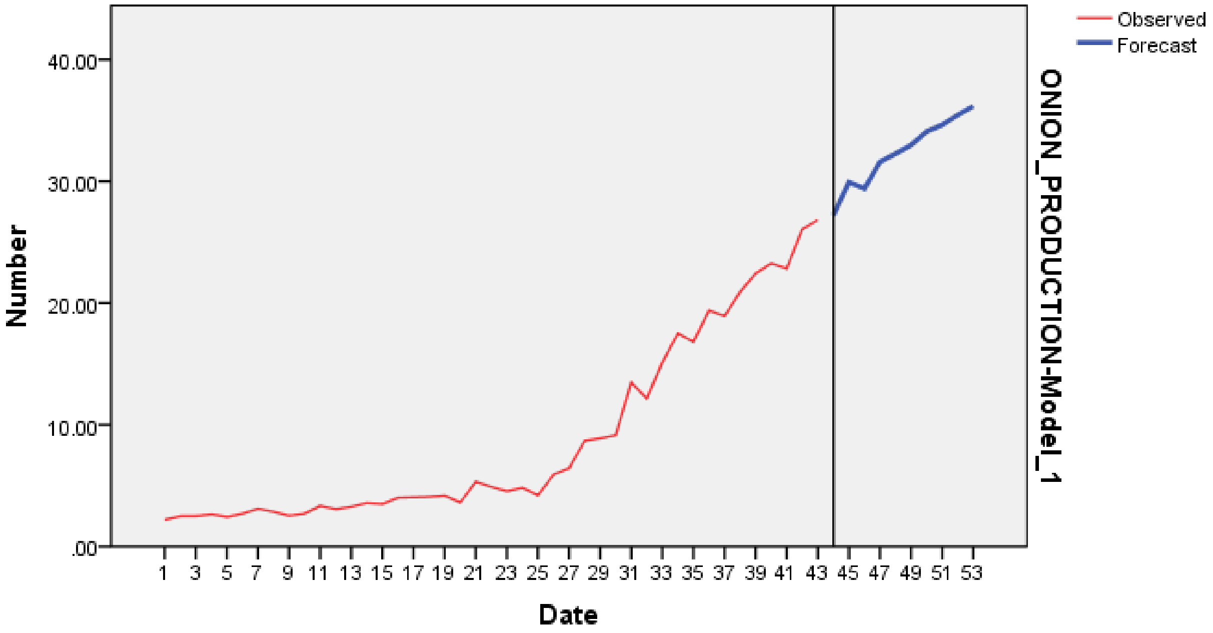
For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

Fit ARIMA Models

Statistical fit	ARIMA(1,1,1)	ARIMA(1,1,3)	ARIMA(1,1,5)	ARIMA(3,1,1)	ARIMA(3,1,3)	ARIMA(3,1,5)
RMSE	1.189	1.039	0.914	0.980	0.937	**0.913**
MAE	0.917	0.791	0.668	0.711	0.641	**0.627**

Result

The best forecast is obtained from ARIMA (3,1,5) become it has low RMSE, MAE value compared to other model.



Conclusion:

Overall, it is noted that ARIMA(3,1,5) provides a good fit for Onion production in India. Its gives a fairly accurate forecasting. However, although forecast from 1978-2030 are within the 95% interval, the graph shows that the green line of actual data has gradually moved out of the confidence interval.

References:

1. Amin, M., Amanullah, M., and Akbar, A. (2014). “Time Series Modeling for Forecasting Wheat Production of Pakistan”. Department of Statistics Bahauddin Zakariya University, Multan, Pakistan. The Journal of Animal & Plant Sciences, 24(5): 1444-1451.

2. Anand Kumar Shrivastav., Dr. Ekata. (June 2012). “Applicability of Box Jenkins ARIMA Model in Crime Forecasting: A case study of counterfeiting in Gujarat State”. ISSN: 2278-1323 *International Journal of Advanced Research in Computer Engineering & Technology* Volume 1, Issue 4.

3. Arnold Zellner, ed. (1979): “Seasonal Analysis of Economic Time Series”. http://www.nber.org/books/zell79-1.

4. Awogbemi, C. and Ajao, O. (2011). “Modelling Volatility in Financial time Series: Evidence from Nigerian inflation rates”, Working Paper, *Ozean Journal of Applied Sciences* 4(3).

5. Banhi Guha. and Gautam Bandyopadhyay. (2016). “Gold Price Forecasting Using ARIMA Model “, *Journal of Advanced Management Science* Vol. 4, No. 2, Department of Management Studies, National Institute of Technology, Durgapur, India.

6. Bartlett, M.S. (1964). “On The Theoretical Specification of Sampling Properties of Autocorrelated FTime Series”, J. Roy. Stat. Soc., B 8: 27–41.

7. Bernardo, J M., Smith, A F M. (1994). “Bayesian Theory”, John Wiley & Sons.

8. Blanchard, O.J, and Perotti, R. (2002). “An Empirical Characterization of the Dynamic Effects of Changes in Government Spending and Taxes on Output”, Quarterly *Journal of Economics*, 117, 1329-1368.

9. Blanchard, O.J, and Quah, D. (1989). “The Dynamic Effects of Aggregate Demand and Supply Disturbances”, American Economic Review 79, 655-673.

10. Bollerslev, T., (1986). “Generalized autoregressive conditional heteroskedasticity”, Working Paper, *Journal of Econometrics* 31 (307-327).