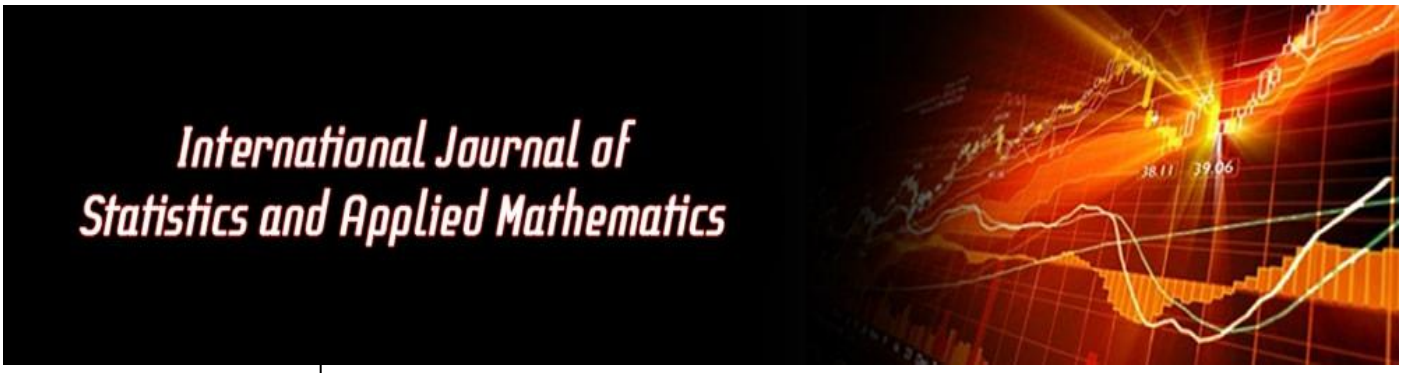
International Journal of Statistics and Applied Mathematics 2025; 10(8): 230-234



ISSN: 2456-1452
NAAS Rating (2025): 4.49
Maths 2025; 10(8): 230-234

https://www.mathsjournal.com
Received: 18-06-2025
Accepted: 27-07-2025

Ayyoob KC
Department of Agricultural Statistics, Palli Siksha Bhavana, Visva-Bharati University, Santiniketan, West Bengal, India

Debasis Bhattacharya
Department of Agricultural Statistics, Palli Siksha Bhavana, Visva-Bharati University, Santiniketan, West Bengal, India

Kader Ali Sarkar
Department of Agricultural Statistics, Palli Siksha Bhavana, Visva-Bharati University, Santiniketan, West Bengal, India

Digvijay Singh Dhakre
Department of Agricultural Statistics, Palli Siksha Bhavana, Visva-Bharati University, Santiniketan, West Bengal, India

Corresponding Author:
Ayyoob KC
Department of Agricultural Statistics, Palli Siksha Bhavana, Visva-Bharati University, Santiniketan, West Bengal, India

Evaluating statistical and machine learning models for paddy yield forecasting in Kerala

Ayyoob KC, Debasis Bhattacharya, Kader Ali Sarkar and Digvijay Singh Dhakre

DOI: https://www.doi.org/10.22271/maths.2025.v10.i8c.2147

Abstract
Paddy is the most important food crop in Kerala which plays a significant role in the food security of the state. The present study analyses paddy yield data from 1956-57 to 2022-23, along with meteorological variables such as rainfall and maximum temperature, to develop a forecasting model for paddy yield in Kerala. The study employed statistical and machine learning approaches for the model development. Various time series models have been developed using the ARIMA, ARIMAX, NNAR, and NNARX methods. The selected models are compared for the relative performance using the metrics like RMSE and MAPE. All the models have exhibited good performance in the model building phase, with minimal RMSE and MAPE values. The ARIMA (0,1,1) model with a constant and the ARIMAX (0,1,1) model with a constant exhibited comparable accuracy and model fit in both the training and testing phases. Considering both accuracy and simplicity, the ARIMA (0,1,1) model has been identified as the optimal model for forecasting paddy yield in Kerala.

Keywords: ARIMA, ARIMAX, forecasting, MAPE, NNAR, NNARX, RMSE

Introduction
Rice (*Oryza sativa* L.) is the most important food crop in Kerala, which provides food security and supports the rural economy. Paddy cultivation once occupied a major share of the state's cultivated area, but it has undergone significant changes over the years. The cultivated area under paddy in the state has increased from 7.6 lakh hectares in 1955-56 to a maximum of 8.82 lakh hectares in the mid-1970s. Thereafter, the area has exhibited a steady decline to 1.9 lakh hectares by 2022-23. The steady decline in crop area has been attributed to the rising labour cost, land usage changes, and shifting cropping patterns in the state (Joseph & Joseph, 2005) [1]. The implementation of the Kerala Conservation of Paddy Land and Wetland Act, 2008, to prevent the conversion of paddy fields for other purposes has effectively reduced land conversion (Rasheed *et al*., 2021) [5].

Time-series analysis deals with data recorded over time and looks for patterns or trends within it. Since each observation is often related to the previous ones, special models are used to capture this relationship. Such analysis is widely applied for forecasting, making it useful in many real-life situations.

Several studies have been made to predict the yield of agricultural crops using different statistical and machine learning methods. ARIMA and ARIMAX models have been used for the yield prediction of paddy in Telangana and reported that ARIMAX performed better (Mishra and Supriya, 2019) [7] when exogenous variables are included in the model. In a later study, Supriya (2021) [3] used a hybrid ARIMAX-ANN model, which improved the forecast accuracy over traditional time-series models. Similarly, Shafie *et al*. (2024) [6] made a comparison of ARIMA and NNAR models to forecast the production of paddy in Malaysia and found that NNAR provided more accurate predictions. Tyagi *et al*. (2023) [9] applied ARIMAX models incorporating monsoon rainfall to forecast sugarcane production. ARFIMA models are more effective than other time-series models when data show both short- and long-term dependencies (Muhammed Irshad *et al*., 2024) [4].

The Neural Network Auto-Regression (NNAR) model, uses lagged values as inputs to predict time-series outputs (Maleki *et al*., 2018) [2]. The NNAR models do not impose restrictions on parameters for stationarity (Thoplan, 2014) [8]. NNAR models capture complex non-linear patterns effectively and improves the predictive performance (Muhammed Irshad *et al*., 2024) [4].

Accurate prediction of paddy yield is significant for ensuring food security and extending support for farmers through timely interventions. Forecasting paddy yield enables farmers to optimize resources, stabilize income and adapt to market changes. This study aims to develop and compare different time series models to identify the most suitable model for forecasting paddy yield in Kerala.

Methodology

Paddy yield data from 1956-57 to 2022-23 were collected from the Department of Statistics, Government of Kerala. Meteorological data, including annual rainfall and mean annual maximum temperature, which strongly influence paddy productivity, were obtained from the India Meteorological Department. The time-series data were split, with 90% used as the training set for building the forecasting models. The remaining 10% of the data was reserved for validating the developed models.

To assess the stationarity of the data, Augmented Dicky-Fuller (ADF) test was applied on the series. Different models including statistical and machine learning approaches were developed for forecasting the yield of paddy. The models such as Auto Regressive Integrated Moving average (ARIMA), ARIMA with exogenous variables (ARIMAX), Neural Network Autoregression (NNAR) model, and NNAR with exogenous variables (NNARX) were applied to develop the forecasting models for the purpose of study.

Mathematically, ARIMA (*p.d,q*) can be expressed as

$y_t = \mu + \sum_{i=1}^{p}(\Phi_i y_{t-i}) + \sum_{j=1}^{q}(\theta_j \varepsilon_{t-j}) + \varepsilon_t$, where y_t is the observed value at time *t*, μ is the mean, $\Phi_i's$ (*i=1,2,...p*) are the autoregressive coefficients, $\theta_j's$ $(j = 1,2,\ldots q)$ are the coefficient of MA process and ε_t is the white noise at time *t*.

ARIMAX model is an extension of ARIMA incorporating exogenous variables that effects the forecasting variable with the ARIMA forecasting models. Mathematically,

$y_t = \beta_0 + \sum_{i=i}^{m}\beta_i x_{i,t} + \sum_{i=1}^{p}(\Phi_i y_{t-i}) + \sum_{j=1}^{q}(\theta_j \varepsilon_{t-j}) + \varepsilon_t$,

where β_i is the regression coefficient, $x_{i,t}$ is the exogenous variable. The best models are identified on the basis of on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). AIC and BIC values are computed as

$AIC=2k-2ln(L)$, and

$BIC=k\ ln(n)-2ln(L)$,

where *k* is the number of parameters and *L* is the likelihood of the model.

NNAR models the time series using past values as inputs and captures non-linear relationships within the data. Mathematically, for a time series X_t, an NNAR maps the lagged inputs $[X_{t-1}, X_{t-2}\ldots\ldots X_{t-p}]$ through a network of interconnected neurons to predict the output X_t.

$X_t = f(X_{t-1}, X_{t-2}, \ldots., X_{t-p}, w) + \varepsilon_t$,

where *w* is the weights of the lagged values, ε_t is the error term and *f* is the activation function.

The integration of exogenous variables to the input layer of this NNAR model makes the model NNARX.

Result and discussion

The annual productivity of paddy in Kerala has exhibited a steady increasing trend over decades (Fig 1). It indicates that the productivity of paddy has increased from around 1,200-1,500 kg/ha in the early 1960s to more than 3000 kg/ha in 2022-23.

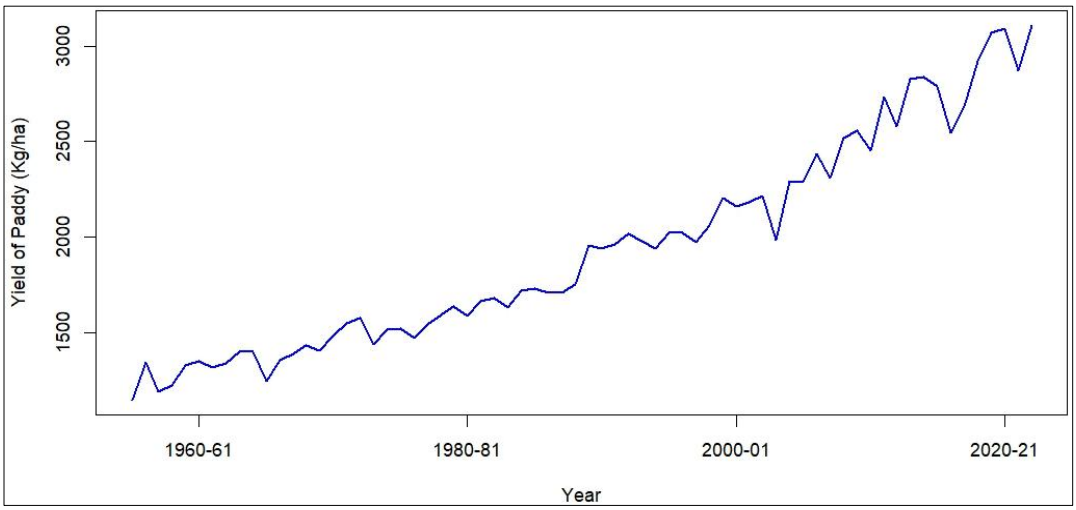


Fig 1: Time series Plot of Yield of Paddy

For developing ARIMA models for the data the stationarity of the series has to be assessed. The plot of the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the original series (Fig 2.) provides indications about the stationarity of the series. The ACF function has exhibited a slow decay and was significant up to 16 lags. This indicates the presence of non-stationarity in the data. The PACF function has shown significant spike at lag 1 and cuts off quickly.

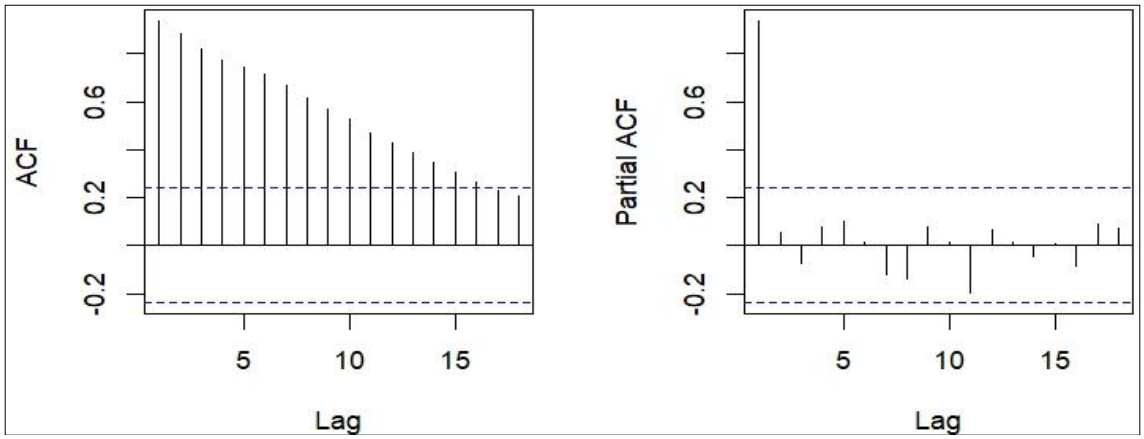


Fig 2: ACF and PACF plot of Paddy yield

To assess the stationarity of the series, ADF test has been performed and the results (Table 1.) revealed that the original series is found to be non-stationary. Hence the series has been differenced to achieve the stationarity. The first-differenced series achieved stationarity, as confirmed by the Augmented Dickey-Fuller (ADF) test. The test yielded a statistic of -4.87 with a p-value less than 0.01. This gives a strong statistical evidence to reject the null hypothesis of a unit root and confirms the stationarity of the series.

Table 1: Augmented Dicky-Fuller Test

Data	ADF statistic(d)	*p-value*
Original Series	-1.09	0.92
First Differenced Series	-5.76	< 0.01

The plot of ACF and PACF functions of the differenced series (Fig 3.) confirms the stationarity of the series. It is evident from the plot that after first lag, none of the lags are significant in the case of ACF function, whereas the PACF function has shown significant spikes at lag 1 and 3. This clearly indicates the stationarity of the differenced series as the ACF and PACF functions quickly drops off immediately after the initial lags.

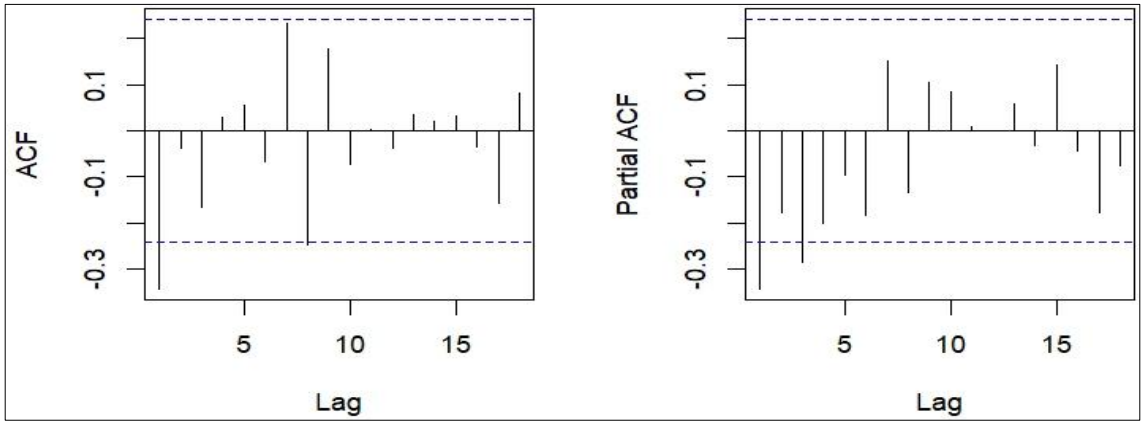


Fig 3: ACF and PACF plot of First differenced Paddy yield

To identify the best model for forecasting the yield of paddy in Kerala, various models are assessed and the models with performance metrics are given in Table 2. Different ARIMA models are explored based on the ACF and PACF plots and the model ARIMA (0,1,1) with constant has been identified as the best ARIMA model based on the performance metrics with lowest AIC and BIC criteria. To develop the ARIMAX model the meteorological variables which are highly correlated with the yield of paddy has been used as the exogenous variables in the ARIMA model. ARIMAX (0, 1, 1) with a constant has been selected as the best model among different ARIMAX models based on the minimum AIC and BIC criteria. The Neural Network Auto Regression (NNAR) model with 3 lagged inputs in the input layer and 2 nodes in the hidden layer has been chosen as the optimum model among various NNAR models and the models are selected based on the RMSE and MAPE values. The inclusion of exogenous variables in the NNAR models resulted in NNARX models with 3 lagged inputs, and rainfall and maximum temperature as the input layer, with 2 nodes in the hidden layer, forming an NNARX (3,2) model with a 5-2-1 network. The Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) have shown consistently lower values in all the selected models, which indicates that the models have captured the underlying patterns in the training data set very effectively. It is observed that the lowest RMSE and MAPE have been reported in the NNAR (3,2) model, followed by the NNARX (3,2), indicating that these models performed better on the training dataset compared with the ARIMA and ARIMAX models.

Table 2: Performance Metrics of Different Forecasting Models for Paddy Yield in Kerala

Model	Parameters				RMSE	MAPE
ARIMA (0,1,1) with constant	constant			MA (1)	86.27	3.47
	26.99*			-0.59*		
	AIC=700.83, BIC=707.06					
ARIMAX (0,1,1) with constant	constant	MA (1)	Rainfall	Max.Temp.	85.51	3.44
	26.65	-0.57*	-0.024	17.98		
	AIC=703.76, BIC=714.15					
NNAR (3,2)	Lag	Nodes in hidden layer		Network	67.64	2.64
	3	2		(3-2-1)		
NNARX (3,2)- NNAR with regressors	Lag	Nodes in hidden layer	Regressors	Network	62.55	2.51
	3	2	2	(5-2-1)		

Residual diagnostics have been performed on the residuals of the selected models using the Ljung-Box test and the results are presented in Table 3. The results revealed that the Ljung-Box statistics of all the selected models are nonsignificant ($p > 0.05$) eliminating the presence of autocorrelation among the residuals of the selected models.

Table 3: Residual Diagnostics Using Ljung-Box Test

Model	Ljung-Box (Q)	*p-value*
ARIMA (0,1,1) with constant	7.67	0.66
ARIMAX (0,1,1) with constant	7.14	0.71
NNAR (3,2)	10.28	0.42
NNARX (3,2)- NNAR with regressors	13.12	0.29

All the selected models are validated using their performance on the test data set. The results of the model validation of the identified models has been presented in Table 4. The higher RMSE and MAPE values of the NNAR and NNARX models on validation phase indicates that these models have failed to capture the fluctuations effectively in the yield of paddy happened in the recent years. The ARIMA (0,1,1) model with a constant has exhibited the lowest RMSE (156.97) and MAPE (4.71) during the validation phase. This indicates that ARIMA performed well on the training dataset and demonstrated better forecasting accuracy.

Table 4: Performance Metrics for Model Validation on Test Data.

Model	RMSE	MAPE
ARIMA (0,1,1) with constant	156.97	4.71
ARIMAX (0,1,1) with constant	161.61	4.92
NNAR (3,2)	214.96	6.49
NNARX (3,2)- NNAR with regressors	202.95	5.71

Although the ARIMAX model also showed comparable performance in both the model-building and validation phases, the ARIMA model is considered the best due to its simplicity. The fitted vs. actual plots using the ARIMA (0,1,1) model with a constant for the yield of paddy in Kerala has been illustrated in Fig. 4.

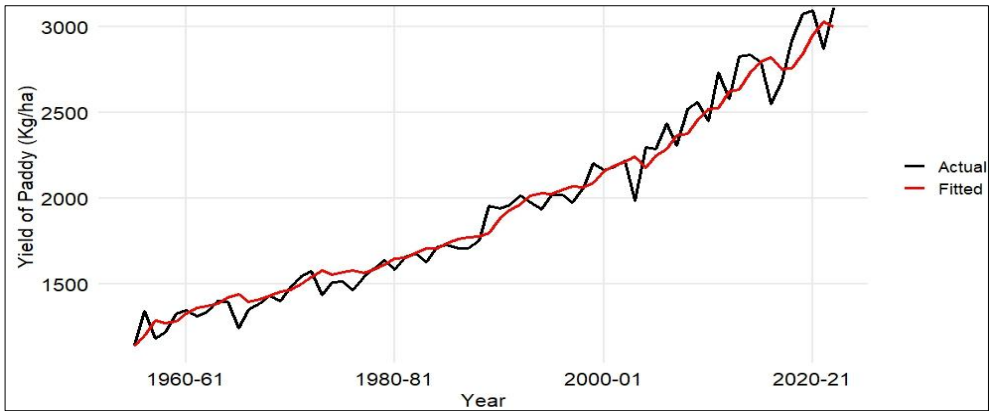


Fig 4: Actual vs fitted plot for the yield of paddy using ARIMA (0,1,1) model.

Conclusion

The study focused on building a robust time series model to forecast paddy yield in Kerala. Four different models-ARIMA, ARIMAX, NNAR, and NNARX-have been tried based on the inherent structure of the data. The comparison of models for relative performance revealed that both conventional linear models and machine learning models performed well in capturing the patterns in the data during model development. The performance of the selected models on validation indicated that the conventional linear models outperformed the machine learning models, as they effectively captured the fluctuations in recent years. Among the selected models, ARIMA (0,1,1) with a constant has been identified as the best model for forecasting paddy yield in Kerala.

References

1. Joseph B, Joseph KJ. Commercial agriculture in Kerala after WTO. South Asian Economic Journal. 2005;6(1).
2. Maleki A, Nasseri S, Aminabad MS, Hadi M. Comparison of ARIMA and NNAR models for forecasting water treatment plant's influent characteristics. KSCE Journal of Civil Engineering. 2018;22(9):3233-45. DOI: 10.1007/s12205-018-1195-z
3. Mishra GC, Supriya M. Forecasting rice yield using ARIMA and ARIMAX models. The Journal of Research, PJTSAU. 2019;47(1&2):68-72.
4. Irshad M, Sarkar KA, Dhakre DS, Bhattacharya D. Comparative Analysis of Statistical Model and Machine Learning Algorithms in Forecasting Black Pepper Price of Kerala. Biological Forum - An International Journal. 2024;16(8):63-8.

5. Rasheed S, Venkatesh P, Singh DR, Renjini VR, Jha GK, Sharma DK. Who cultivates traditional paddy varieties and why? Findings from Kerala, India. Current Science. 2021;121(9):1188. DOI: 10.18520/cs/v121/i9/1188-1193
6. Shafie SNM, Aziz NA, Nafi MNA, Malek SA@A, Amran A, Shafie SNA. Predicting Paddy Production in Malaysia: A Comparative Analysis between Arima and Neural Network Autoregressive (NNAR) Models. International Journal of Academic Research in Business and Social Sciences. 2024;14(12).
DOI: 10.6007/ijarbss/v14-i12/23351
7. Supriya K. A Study on the Performance of the ARIMAX-ANN Hybrid Forecasting Model over the other Time Series Forecasting Models ARIMAX and ANN in Forecasting the Rice Yield. International Journal of Current Microbiology and Applied Sciences. 2021;10(01):3421-8.
DOI: 10.20546/ijcmas.2021.1001.403
8. Thoplan R. Simple v/s Sophisticated Methods of Forecasting for Mauritius Monthly Tourist Arrival Data. International journal of statistics and applications. 2014;4:217-23.
9. Tyagi S, Chandra S, Tyagi G. Climate Change and its Impact on Sugarcane Production and Future Forecast in India: A Comparison Study of Univariate and Multivariate Time Series Models. Sugar Tech. 2023;25(5):1061-9. DOI: 10.1007/s12355-023-01271-2